Endogenous Measures for Contextualising Large-Scale Social Phenomena:

A Corpus-Based Method for Mediated Public Discourse

J. Clark Powers, MS, DEA

Thesis submitted for the award of PhD, December 2022
School of Communications, Dublin City University

Supervisor: Dr. Dónal Mulligan

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Jason Clark [illegible]

(Candidate) ID No.: 17213171

Date: 15 December 2022

Contents

List of Figures

Abstract

Endogenous Measures for Contextualising Large-Scale Social Phenomena:
A Corpus-Based Method for Mediated Public Discourse

J. Clark Powers

This work presents an interdisciplinary methodology for developing endogenous measures of group membership through analysis of pervasive linguistic patterns in public discourse. Focusing on political discourse, this work critiques the conventional approach to the study of political participation, which is premised on decontextualised, exogenous measures to characterise groups. Considering the theoretical and empirical weaknesses of decontextualised approaches to large-scale social phenomena, this work suggests that contextualisation using endogenous measures might provide a complementary perspective to mitigate such weaknesses.

This work develops a sociomaterial perspective on political participation in mediated discourse as affiliatory action performed through language. While the affiliatory function of language is often performed consciously (such as statements of identity), this work is concerned with unconscious features (such as patterns in lexis and grammar). This work argues that pervasive patterns in such features that emerge through socialisation are resistant to change and manipulation, and thus might serve as endogenous measures of sociopolitical contexts, and thus of groups.

In terms of method, the work takes a corpus-based approach to the analysis of data from the Twitter messaging service whereby patterns in users' speech are examined statistically in order to trace potential community membership. The method is applied in the US state of Michigan during the second half of 2018—6 November having been the date of midterm (i.e. non-Presidential) elections in the United States. The corpus is assembled from the original posts of 5,889 users, who are nominally geolocalised to 417 municipalities. These users are clustered according to pervasive language features. Comparing the linguistic clusters according to the municipalities they represent finds that there are regular sociodemographic differentials across clusters. This is understood as an indication of social structure, suggesting that endogenous measures derived from pervasive patterns in language may indeed offer a complementary, contextualised perspective on large-scale social phenomena.

Note

This work is a decidedly interdisciplinary work. That is to say, it is not interdisciplinary by happenstance, but by choice. For that reason, it is written with fewer assumptions about the reader's knowledge of terms and topics compared to disciplinary works. Such an explicit approach to social science can seem cumbersome, verging on pedantic, but it is taken to increase accessibility for readers across disciplines. Moreover, such an approach serves a valuable pedagogical end. The aim is for this monograph to have continued value beyond its immediate singular purpose—may it be an aid and guide for those coming after.

Further to the ends of interdisciplinarity and pedagogy, the text is heavily annotated. In the main annotations offer context, often historical or disciplinary, or justifications of certain choices made in this work and the document itself. Notes that are marked with a black star (★), predominantly found in **chp. 5** describing the method, are intended for students interested in computational methods for social inquiry. Those notes strive to give an honest account of techniques and situations, and to offer guidance for learning more. The notes are found in **annex E**.

All paragraphs are sequentially numbered. Cross-references to paragraphs are indicated with a bold pilcrow and the relevant paragraph number (e.g. **¶ 123**). Cross-references to sections are also in bold, giving the chapter and section (e.g. **chp. 1 § 1.2.3.4**). Section references without a chapter prefix refer to the current chapter. Note cross-references are indicated by the abbreviation **n.** followed by the relevant note number.

Once upon a time there was an inventive fish, who discovered water.

Some day, perhaps, an inventive man may discover love, the atmosphere our souls breathe. And other men will tell him, “How you’ve changed!”

Roger Pocock, *The Cheerful Blackguard*, 1915, p. 335

Chapter 1

Introduction

> The production of ideas, of conceptions, of consciousness, is at first directly interwoven with the material activity and the material intercourse of [people], the language of real life.
>
> Karl Marx and Friedrich Engels, *The German Ideology*, 1846, p. 47

1.1. Hybrid Methodology for Hybrid Society

1. This work presents an interdisciplinary methodology for macro-level research of large-scale social phenomena grounded in communication studies. The title of this work, *Endogenous Measures for Contextualising Large-Scale Social Phenomena,* does not mislead, but rather encapsulates. A corpus-based method for mediated public discourse is developed, and the site of research to which it is applied is the US state of Michigan during the election period of 2018 (specifically the last six months of that year). The empirical results were not sought to learn about the site of research specifically, however, but rather to seek warrant for the operationalisation proposed and so for the methodology overall. Such warrant is found. While the methodology is the larger purpose of this work, the method is the crucial test of it, and so gets the title.

2. The concern with interdisciplinary methodology for macro-level research is motivated by the challenges witnessed across fields and disciplines in grappling with the emergence of 'hybrid' society. Beyond being networked by information and communication technologies (Castells, 1996 ; van Dijk, 1999), hybrid society is characterised by deep mediatisation (Chadwick, 2013 ; Couldry and Hepp, 2017) and the ever-present blending of the offline and online, the physical and virtual (Blommaert, 2019). A key challenge is that structural conceptualisations of large-scale phenomena must give way to socially communicative conceptualisations if we wish to see hybrid society as it is, rather than as some might wish it to be (Boulianne, 2020). On the one hand, this is a pickle. Disciplinary social science is rooted in structural understandings, and serves to reproduce them. On the other hand, this is impetus to move beyond convention. Methodology suited to hybrid society will need to blend micro-level theory with macro-level method—thus the development of hybrid methodology will be interdisciplinary. Given its thematic orientations and diaphanous boundaries, communication studies is well positioned for such an undertaking. This work demonstrates that.

3. The fundamental argument is that the move from structural conceptualisations to socially communicative conceptualisations is ontological. Any such move in regard to large-scale phenomena thus faces a hurdle in that we are socialised into structural understandings by life in the modern State—arguably the preeminent structural phenomenon. Society writ large is understood, measured, and evaluated in terms that serve the State. Crucially, such service to the State enabled and shaped the rise of large-scale social research[1]★ during the twentieth century, which continues to rely upon conceptualisations and measures of society suiting structural logics. For example, the study of political participation—the large-scale phenomenon that this work will reconceptualise—could be understood as the study of electoral measures in relation to socioeconomic and demographic measures. This work terms such measures 'exogenous' in that they are not derived from social characteristics, but rather from structural attributes. While exogenous measures have their utility and place, their use in social inquiry is indicative of a certain subject model—the 'fungible individual'. With exogenous measures, as far as individuals share structural attributes they are effectively interchangeable. Society is reduced from the broad amalgam of people and their groupings in relation to "nothing but a numerical aggregate, a conglomeration of units" (Dewey, 1888, p. 4). While this could be taken as an epistemological framing, it will be shown that this subject model is so deeply ingrained that it forms the ontological core of structural conceptualisations.

4. This work argues that reconceptualising large-scale phenomena for hybrid society necessitates an ontological shift. The subject model of the fungible individual has no place for socially communicative—that is, intersubjective—phenomena of any scale, be they person or nation. And so that subject model renders structural conceptualisations effectively blind to socially communicative phenomena ; they cannot account for them, and so do not see them. Thus the move from structural conceptualisations to socially conceptualisations entails a shift in subject model from the fungible individual to the fully social person.

5. For this reason, the evaluation of this work focuses on the operationalisation. Reconceptualisation of phenomena is the bread and butter of knowledge work—it is the routine and the fundamental purpose. But ontological shifts in theory implicate consequent shifts in method (cf. Kuhn, 1970). This work argues for the development of 'endogenous' measures that are grounded in and derived directly from the social characteristics of the phenomena under study. To that end it explores the operationalisation of language to derive such measures. The most expeditious way to see if the methodology presented here is worth its salt is to evaluate the empirical results of the method implementing that operationalisation. If those results satisfy the requirements specified in the research questions, we have warrant for the methodology overall.

6. As a contribution to the development of hybrid methodology, this work reconceptualises and operationalises a topic of inquiry that traditionally falls outside of communication studies, that has long been characterised by structural conceptualisations, and that is known to be challenged by the emergence of hybrid society—the study of political participation. That study concerns the actions of individuals and groups in regard to political systems ; given the place of communication and media in political phenomena, the topic is a sound choice for exploration of hybrid methodology for macro-level research.

7. A conceptual framework is developed for reconceptualising political participation in socially communicative terms, and an operationalisation based on endogenous measures is proposed. A method is assembled to implement the operationalisation, and the empirical results allow us to evaluate the conceptual framework as a potential response to the initial problem—both in light of the original discipline's understandings and in terms of the need to adapt methodologically to hybrid society. To reiterate, the method is not intended for study of the topic itself, but rather to provide a test of the proposed operationalisation of the reconceptualised topic.

8. The operationalisation draws on language to provide endogenous measures of sociation, and the method is directed towards the investigation of discourse at scale as attested in social media. Political activity—both the political participation of everyday people and the party political efforts of officeholders, candidates, and professionals—has moved markedly into social media in recent years (Bennett, 2012 ; Boulianne, 2015 ; Nwafor *et al.*, 2013 ; Valeriani and Vaccari, 2016). That move is part and parcel of the emergence of hybrid society, although social media serve only to accelerate deeper changes that were already underway (e.g. Bimber, 2003). At the same time, political participation by such means facilitates non-institutional activities (Lee, 2017 ; Theocharis and van Deth, 2015). While the conventional study of political participation takes an institutional perspective, looking to the voting booth and party membership, the people themselves may be otherwise occupied (Margetts *et al.*, 2016). The need to move from structural conceptualisations to socially communicative conceptualisations is clear to see. But moreover, if we wish to see political participation as it is in hybrid society, we are obliged to operationalise in a manner that can engage with mediated public discourse.

9. Language provides the most ready resource for developing endogenous measures for the purposes of this work. The measures that will be derived are proxies of sociation. Quite simply, we are looking for groupings of people that have similar patterns of speech. We do not seek to understand what people are saying or what issues they discuss ; rather, the method uses language-derived measures to seek out social groupings as people themselves might perceive them, instead of as they might be categorised in a government census, for example. Importantly, the method does not seek to replace conventional methods based on exogenous measures, but rather to complement them by way of contextualisation—the development of hybrid methodology will proceed more smoothly as dialogue than as diatribe.

10. The remainder of this chapter is as follows. The topic to be reconceptualised, political participation, is introduced ; this is followed by a description of the problem (i.e. the challenges to the study of the topic in hybrid society) and the proposed remedy (§§ **1.2–1.3**). These sections are relatively lengthy, but they serve to introduce themes and points that will anchor the overall discussion. The purpose of approaching the topic with a hybrid methodology is stated (§ **1.4**), and the research questions are presented (§ **1.5**). The conceptual framework and method are then described, with a brief description of the approach to analysis and evaluation (§§ **1.6–1.7**). The critical character of this work is then addressed, specifying the main assumptions driving critique (§ **1.8**). The chapter concludes with an overview of the document (§ **1.9**).

1.2. The Study of Political Participation

NOTE: This section, and much of this document, is heavily annotated. The annotations in the main provide context and elaborate rationale, are in place to facilitate interdisciplinary understanding, and are intended to give this document lasting pedagogical value beyond its immediate purpose.

11. The topic of political participation in itself is generally framed as concerning how non-establishment members of a polity (i.e., 'everyday citizens')[2] engage with the structures of governance of that polity. Understood broadly, the topic is one of the central questions at the foundation of modern social inquiry, present for example in the cultural analysis of Tocqueville (1835, p. 14), the empirical philosophy of Comte (1851), the economic theory of Marx (1867), and the civic epistemology of Durkheim (1893). The scientific study of the topic is roughly 100 years old, most commonly traced back to the work of Charles Merriam and Harold Gosnell on electoral behaviour in Chicago during the 1920s (Gosnell, 1926 ; cf. Merriam, 1922 ; Merriam and Gosnell, 1924). Large-scale empirical work on the topic began to emerge mid-century (Berelson, Lazarsfeld and McPhee, 1954 ; Campbell, Gurin and Miller, 1954 ; Lazarsfeld, Berelson and Gaudet, 1944), with national meta-analyses and cross-national comparative studies appearing the 1960s (Almond and Verba, 1963 ; Campbell *et al.*, 1960 ; Conway, 1991, pp. 33–34 ; Milbrath, 1965).

12. In a general (i.e., non-operational) sense, political participation can be understood to indicate "citizens' activities affecting politics" (van Deth, 2014, p. 351). In an operational sense, there is no single definition, and thus there has been a variety of approaches to study of the topic. This should not be understood as problematic, but rather as a reflection of the contextual and contingent nature of the political and its study. As such, it is important to bear in mind that political participation is an "umbrella concept" (Huntington and Nelson, 1976, p. 14) encompassing a huge range of evolving behaviours and thus possible measures. However, a review of measures of the topic from the turn of the century—that is, at a point in time just before the effects of social media and networking began to be broadly acknowledged (Boulianne, 2020)—noted that "almost all" operational definitions could be distilled to four key elements (Brady, 1999, p. 737):

- **action**, that is, observable (and thus measurable) activities, as opposed to thoughts, attitudes, interests, or intentions ;
- **ordinary citizens**, as opposed to members of the political establishment ;
- **influence**, that is, intentional behaviours (i.e. actions) seeking to effect outcomes, as opposed to more passive behaviours such as information-seeking or topical discussion ; and

- **political outcomes**, that is, behaviours must be targeted at "government policy or activity", as opposed to mundane or quotidian behaviours, regardless of the political implications that such behaviours might have (such as purchasing patterns, community engagement, etc.) (1999, pp. 737–738).

Considering that definitions of political participation are understood to be fitted to the specific research at hand (Verba, Nie and Kim, 1978, pp. 46–48) in a manner reflective of the researcher's understanding of the purpose or nature of the topic (Fox, 2014, p. 496), the observed ubiquity of these elements is interpreted here as reflecting the conventional approach to operationalising political participation, and thus to its conceptualisation and study generally, prior to the emergence of hybrid society.

13. The conventional approach to the study of political participation is premised on reductionism through the abstraction of human attributes and behaviour (Brady, Verba and Schlozman, 1995 ; McClurg, 2003 ; e.g. Quintelier and van Deth, 2014), and through foregrounding the structure of the 'political system' (e.g. Easton, 1953 ; see also Gunnell, 2004, 2013).[3] This premise reflects the development of a field,[4] and subsequently a discipline, driven by the pursuit of an objective science of the political through abstraction, quantification, large-scale data collection, and statistical analysis (Almond, 1998, pp. 64–73 ; Bulmer, 1981 ; Gow, 1985 ; Monroe, 2004). The term 'conventional' is used here as the approach in question:

- was characteristic of the field during its establishment and initial expansion (Almond, 2004 ; Campbell, 2013 ; Dahl, 1961) ; and
- was integral to the development of US political science generally in the twentieth century, during which the US academy was overwhelmingly dominant internationally (King, 1990 ; Monroe, 2004, p. 95 ; Sartori, 2004 ; Sigelman, 2006).

These points are deeply intertwined for reasons beyond the scope of this discussion. However, of immediate importance to this work is that the conventional approach to the study of political participation gave rise to the behavioural movement in political science (Almond, 2004 ; Bulmer, 1981 ; Conway, 1991 ; Dahl, 1961 ; Monroe, 2004).[5] That movement has had a significant and continuing disciplinary impact, shaping political science in the United States (Berkenpas, 2016 ; Gunnell, 2013 ; Sigelman, 2006) and internationally (Boncourt, 2015 ; Cairns, 1975 ; Lenine and Mörschbächer, 2020 ; Valles and Newton, 1991). While the behavioural movement lost prominence in the second half of the twentieth century, the post-behavioural period saw some of the movement's methodological tenets[6] institutionalised in the disciplinary study of political science internationally (Gunnell, 2002)—notably the abstraction and quantification of human attributes and behaviour. In this manner, the term 'conventional' indicates not only the accepted manner in which political participation research specifically was pursued prior to the emergence of hybrid society, but more deeply the institutionalised approach to empirical political research generally that remains commonplace today (e.g. Brady, Collier and Box-Steffensmeier, 2011 ; Feezell, 2016 ; Gerber *et al.*, 2011 ; McClurg, 2003).

14. The study of political participation, while originating and initially developed in the United States context, has grown far from those roots, providing a broad temporal and cultural perspective on political sociation and action around the world (e.g. Dalton and Klingemann, 2011 ; Marien, Hooghe and Quintelier, 2010 ; Teorell, Torcal and Montero, 2006). Similarly, from an initial institutional focus, the study has also shifted in levels of analysis, from the macro-level perspective on mass collective behaviour, to the micro-level perspective on individual behaviour, and now increasingly towards meso-level perspectives on collectives as emergent phenomena (Boulianne, 2020 ; Fox, 2014 ; Gibson and Cantijoch, 2013 ; Gil de Zúñiga *et al.*, 2010 ; Hooghe, Hosch-Dayican and van Deth, 2014 ; Theocharis and van Deth, 2018).

15. Nevertheless, the pressures that produced the conventional approach persist.[7] From among them, it is important to note that the conventional approach to political participation, as well as the institutionalisation of abstraction and quantification, can be interpreted as the outcome of efforts to define and sustain a discipline, firstly in relation to other disciplines, and secondly in relation to 'dissidents' within the discipline—that is, to establish a disciplinary status quo and subsequently to preserve it.[8] The conventional approach to the study of political participation evolved, as did political science generally, so as to be distinct from approaches to the political taken in other disciplines, such as economics, history, psychology and sociology (Almond *et al.*, 1962, cited in Kaufman-Osborn, 2006, fn. 4). "Pure science" was to be found in the regularity of numbers (Easton, 1965, p. 7). In that light, the argument here is that the conventional approach persists because it is deeply rooted in institutional disciplinary identity and power structures (cf. Dryzek, 2006).[9]

1.3. Adapting the Study to Hybrid Society

16. The challenge faced by the conventional approach to the study of political participation is not in the character of the approach as such ; the discipline of political science is better placed to debate its own methodological approaches. Rather, the challenge identified in this work—and a fundamental motivation for addressing a political topic from a perspective of communication studies—concerns the application of the conventional approach in contexts of deep mediatisation (Couldry and Hepp, 2017, pp. 34–56) and media hybridity (Chadwick, 2013, pp. 23–59)—that is, in hybrid society. The approach was not developed to address such contexts, and thus it is poorly suited to the task in terms of theory. Communication studies, on the other hand, is directed towards understanding such contexts. Yet while we may have theory for the task, we are not well equipped in terms of method. Blending approaches drawn from communication studies with the conventional approach of political science could yield a hybrid approach that is better adapted to the study of political participation in hybrid society.

17. This work thus proposes an approach to the study of political participation that 1) is more suited to contemporary contexts, yet remains compatible with and complementary to the conventional approach as observed in political science, and 2) provides one possible pathway towards harnessing the advantages of computational social science (Edelmann *et al.*, 2020) for researchers regardless of discipline that privilege—or seek to incorporate—more grounded, contextualised understandings (Hall, 2007 ; cf. Törnberg and Törnberg, 2018).

18. As stated, the conventional (i.e., reductive and structural) approach to the study of political participation has become institutionalised in contemporary political science—the early positivist work at the University of Chicago during the 1920s laid the foundation for the development of the behavioural movement after the war (Monroe, 2004), and one can discern "an almost straight line" from those approaches through to the formal approaches of the late twentieth century "and beyond" (Barrow, 2011, p. 82 ; see also Gunnell, 2004). That process of institutionalisation unfolded alongside significant changes in the social, political, and technological domains that have led to a vast expansion in the scope of what has been studied as political participation, and to contentions over the appropriate methods with which to study it (Gibson and Cantijoch, 2013 ; Theocharis and van Deth, 2018 ; van Deth, 2014). It is these external factors that are evidence of the challenges of social research in hybrid society, and that motivate this work.

19. Among these external factors are developments in information and communication technologies during the second half of the twentieth century (Castells, 1996 ; Preston, 2001 ; Thompson, 1995), most notably the advent in the twenty-first century of 'New Media' and their array of artefacts and practices (Jenkins and Deuze, 2008 ; Lievrouw and Livingstone, 2006 ; van Dijck, 2013). These developments—all of which impact processes of communication and sociation—have revealed shortcomings in accepted understandings of political participation and have presented complications to the conventional study of the topic (Farrell, 2012 ; Fox, 2014 ; Gibson and Cantijoch, 2013 ; Theocharis and van Deth, 2018). Of central concern to this work is the growing body of literature that points to the need to reconceptualise political participation from a socially communicative perspective (for an overview, see Boulianne, 2020).

20. The conventional approach has little place for such a perspective. In terms of theory, structural explanations are blinded to the complex of situated and emergent social phenomena underlying the political (e.g. Almond, 2004 ; Monroe, 2004 ; cf. Rorty, 1980a). In terms of method, the tools and techniques for operationalisation, data collection, and analysis common to the conventional approach (e.g. Esarey, 2018 ; King, 1990 ; e.g. King, 1998 ; King, Keohane and Verba, 1994 ; Mahoney, 2010) were developed to respond to conceptualisations of phenomena that effectively 'black box' the structures and collectivities of interest by, at best, mistaking their social nature and, at worst, mistaking the nature of the social. This is the effect of an inappropriate subject model. Thus, even if these tools and techniques are sound in and of themselves, their normative and objectifying character is not appropriate for the study of phenomena that emerge from the contingent action of social, communicative people.

21. This work offers a reconceptualisation of political participation from a socially communicative perspective. The reconceptualisation is derived from theoretical and empirical work common to communication studies, and translated in a manner that addresses the conceptual and methodological needs of the conventional approach to political participation. The general argument can be broken into four parts:

- political participation is increasingly understood to be a complex of communicative phenomena (Boulianne, 2015 ; Carpentier, 2016 ; Dahlgren and Alvares, 2014 ; Gil de Zúñiga *et al.*, 2010 ; Loader and Mercea, 2011 ; Milner, 2013 ; Ohme, de Vreese and Albæk, 2017 ; e.g. Valenzuela *et al.*, 2019 ; van Dijk and Hacker, 2018) ;
- such communicative phenomena are fundamentally social (Burke, 1966 ; Habermas, 1984 ; Knoblauch, 2013) as opposed to information or communication theoretic (e.g. Berlo, 1974 ; McGuire, 1961 ; Schramm, 1954 ; Shannon and Weaver, 1949) ;
- the focus on quantification and on structure of the conventional approach to the study of the topic has blinded it to socially communicative phenomena by effectively 'black boxing' them (Hustinx and Denk, 2009 ; Pinch and Bijker, 1984)—that is, such phenomena have been rendered theoretically uninterpretable to it (Feyerabend, 1962 ; cf. Kuhn, 1970, pp. 198–200) ; and thus
- the remedy is to reconceptualise these phenomena in a manner that 1) permits operationalisation that can be demonstrated as compatible with the conventional approach, while still 2) attending to their intersubjective, contextual, and thus social character. In that manner the remedy respects the operational preferences of the conventional approach while making the phenomena in question amenable to interpretation.

22. The remedy just described can be understood as a process of situating knowledges (Haraway, 1988) of communicative action and political participation, among others. By attending to the situatedness of social phenomena and of various approaches to their study, bridges can be built for translating across theoretical understandings, methodologies, disciplines, and paradigms (Callon, 1986 ; Clarke, 2003 ; Star and Griesemer, 1989 ; for interdisciplinary examples see Buller, 2009 ; Kalenda, 2016 ; Nightingale, 2016).[10] Here, we shall speak simply of contextualising structural understandings.

1.4. Purpose of the Methodology

23. Key to developing a reconceptualisation of political participation suitable for both communication studies and political science is to contextualise the political through language. This work engages in such contextualisation by understanding political participation as a mass phenomenon of collective action shaped—and effected—by the language of participants (cf. Austin, 1955 ; Searle, 1969). Such a perspective on the political has received sustained disciplinary attention only in recent decades (Chilton and Schäffner, 2011 ; Joseph, 2006), yet recognition of the critical intersection of language and politics has much deeper roots. Such can be traced back most prominently to Aristotle, who addressed the topic in a passage in the *Politics*. The adage 'Man is a political animal' comes from the passage in question ; what is generally elided is the subsequent explicit recognition that the political nature of humankind is bound to our singular capacity for language. Through language, we are able to develop shared understandings and values, and it is these shared meanings that produce and characterise collectivities of all size, from the household to the *polis* (Rackham, 1932, p. 11 ; *Politics*, 1.1, 1253a).

24. This venerable perspective on the intersection of language and politics is foregrounded in rather more recent scholarly work under rubrics such as 'digital citizenship' (Couldry *et al.*, 2014 ; Isin and Ruppert, 2020), 'discursive participation' and 'deliberative democracy' (Delli Carpini, Cook and Jacobs, 2004 ; Neblo *et al.*, 2010), and the 'mediatisation' of politics (Dahlgren and Alvares, 2014 ; Marcinkowski, 2014)—whether natural language and its role are treated explicitly or implicitly, each of these rubrics hinges upon the symbolic construction of the political and upon the discursive negotiation and application of symbols by individuals in the doing of politics (e.g. Edelman, 1977 ; Krotz, 2017 ; also cf. Murru, 2018 ; Silverstone, 1999, chp. 2).

25. In that light, this work engages with more conventional (i.e., structural, behavioural, or positivist) conceptualisations of political participation (Brady, Verba and Schlozman, 1995 ; Milbrath, 1965 ; Parry, Moyser and Day, 1992 ; Pattie, Seyd and Whiteley, 2004 ; Verba and Nie, 1972 ; Verba, Nie and Kim, 1978 ; e.g. Whiteley, 2012) with the purpose to demonstrate that intersubjective understandings of the sociopolitical (provided by attention to language, in the case of this work) can be conceptually and methodologically compatible with, and thus complement, more objective—and thus generally more quantified and decontextualised—understandings of conventional approaches (King, Keohane and Verba, 1994 ; cf. Mahoney, 2010). In this manner, the hybrid methodology proposed in this work can be considered as a 'bridging' methodology, serving to contextualise the study of political participation, and thereby to build links to, and for, common knowledge and understanding across the divides of disciplines and academies (cf. Star and Griesemer, 1989). Oftentimes interdisciplinarity is viewed *à la carte*, whereby studies are understood to be built up from specific components taken from various disciplines. For example, this work could be viewed as taking its overall theoretical framing from communication and media studies ; its large-scale, macro-level focus on sociopolitical phenomena from political science ; its method and analytical lens from linguistics ; and so forth. However, such a perspective serves to firm up the perceived boundaries between disciplines, while obscuring the presence of consonant understandings and approaches among them, and at the same time oversimplifying the work being undertaken within them. It is for this very reason that the hybrid methodology developed in the following chapters pursues explicitly cross-boundary work—to highlight where commonalities and thus links exist, so that interdisciplinarity can be viewed as common effort towards common concerns, rather than as supplementary piecework within a more rigid disciplinary context.

1.5. The Research Questions

26. The research questions to be answered by this work are as follows:

RQ_1 – How can political participation as reconceptualised in hybrid society be operationalised for computational and statistical analysis? and

RQ_2 – Can the results of such operationalisation remain interpretable from a structural perspective?

The rationale supporting these questions as valid tests of the work is now given.

27. The hypothesis motivating this work is:

H_M – The empirical study of language can help to ground political inquiry by contextualising socially communicative phenomena.

On the one hand, it is generally acknowledged that language and politics are thoroughly intertwined (Edelman, 1985). And while political phenomena do not always manifest through language, there is a growing acceptance that language should be conceptualised as "political from top to bottom, in its structure as well as its use" (Joseph, 2006, p. *ix*). On the other hand, the translation of that awareness into theory and method is contingent on disciplinary context—in certain contexts, the statement might not be axiomatic, so much as trivial or antithetical.[11]

28. Among those social sciences that foreground the relations of power and meaning—that is, a broad conception of the political—there is little need to argue the point (e.g. Austin, 1955 ; Bourdieu, 1991 ; Duranti, 1997 ; Fairclough, 1989 ; Lukes, 1974 ; Mead, 1934 ; Wittgenstein, 1953) ; in such contexts, the hypothesis can be considered axiomatic, in the sense of postulating a foundation from which to base further argument. In the case of 'conventional' political science, where a structural, systemic perspective on the political is characteristic (e.g. March and Olsen, 1984),[12] such an understanding may or may not obtain (Rhodes, 2017)—nevertheless the role of language as an important tool of research is recognised (Axelrod, 1976 ; Lasswell and Leites, 1949 ; Laver, Benoit and Garry, 2003 ; Monroe and Schrodt, 2008 ; Wilkerson and Casas, 2017). However, the affordances observed in that tool will vary according to the assumptions and norms of the scholarly context in which it is employed (cf. Yanow, 2003). This work argues that a socially communicative reconceptualisation of political participation can be made compatible with the conventional approach by treating 'text as data'. The 'text-as-data' methodological approach has expanded rapidly in political science since the turn of the century,[13] and is characterised by "feature abstraction" as opposed to treating text as text (Benoit, 2020)—that is to say, treating text itself as the bearer of meaning through language, and not as a resource for indicators of concepts to which text-external meaning is attributed. Whereas text-as-text allows for direct interpretation,[14] text-as-data is an empirical move that produces further text (i.e., data) amenable to computational and statistical analysis for subsequent interpretation.

29. Following this rationale, the first research question is:

RQ_1 – How can political participation as reconceptualised in hybrid society be operationalised for computational and statistical analysis?

30. This work seeks to demonstrate that such operationalisation is indeed possible by way of quantification. Rather than avoiding the quantification of the conventional approach, it is embraced so as to facilitate the study of communicative political phenomena at scale.[15] It should be noted that demonstrating the amenability of the reconceptualisation of political participation to text-as-data (i.e., computational and statistical) analysis is not trivial. For reasons beyond the scope of this discussion, common computational and statistical approaches often hinge on certain assumptions of randomness in the production and analysis of data (cf. **chp 6. § 6.2.1**). Language, however, is never random (Kilgarriff, 2005). No social phenomenon can be truly random, that is, independent of all other phenomena. This work thus seeks to avoid the use of techniques that depend on assumptions of randomness, and as such **RQ$_1$** is contingent on what sort of analysis the data should be amenable to. As observed in **§ 1.4**, the more immediate concern of this work is to demonstrate the complementarity of the reconceptualisation to the conventional conceptualisation. In doing so, **RQ$_1$** and its answer both are transposed from the realm of the contingent to that of the pragmatic.

31. The second research question is thus:

> **RQ$_2$ – Can the results of such operationalisation remain interpretable from a structural perspective?**

32. The logic of evaluating complementarity by means of **RQ$_2$** rests on a central assumption of this work, which is that the blindness of structural perspectives to socially communicative phenomena is not symmetric. That is, socially communicative perspectives are not themselves blinded to structural phenomena. Rather, it is understood that structural phenomena will shape socially communicative processes (McLuhan, 1964).[16] Given that understanding, a socially communicative perspective expects to see patterns indicative of social structures, but without reifying such through assigning them agency or causality.

33. **RQ$_1$** is an open question addressing the operationalisation of the reconceptualisation of political participation. An answer is developed in **chp. 4** and implemented by the method as described in **chp. 5**. Nevertheless, without empirical warrant, the answer remains provisional. **RQ$_2$** is a closed question addressing the evaluation of the empirical results of the method implementing the operationalisation. Answering **RQ$_2$** in the affirmative indicates that the provisional answer to **RQ$_1$** is warranted and that the operationalisation serves the purpose of the methodology as laid out in the previous section. Answering **RQ$_2$** in the negative indicates that the whole shebang best be returned to the drawing board. As will be demonstrated in **chp. 6**, **RQ$_2$** is answered in the affirmative, thus warranting the methodology.

1.6. The Conceptual Framework

34. To guide the reconceptualisation of political participation in a manner that responds to the research questions just given, this work develops a conceptual framework that is fully specified (i.e. explicitly declared) from ontology to operationalisation. In part, such a framework serves an important pedagogical purpose for both author and reader. But foremost such a framework is necessary because of the argument that will undergird the answer to **RQ$_2$**—that a quantified operationalisation can avoid the blindness of structural orientations to socially communicative phenomena through a reconceptualisation of the 'subject model' of political participation.[17] Specifically, this work proposes an **ontological shift**:

1. **from the structural 'fungible individual'**—where collectives are identified by their exogenously presumed place, function, and intention within the structures of society and governance (cf. Fox, 1996) ;
2. **to the communicative, intersubjective 'social person'**—where collectives are seen, but not defined, by means of their endogenously produced descriptions of context.

Such an ontological shift allows for an epistemology that is sensitive to intersubjectivity (Duranti, 2010) and manifold contexts (Emirbayer, 1997), and thus one that enables the interpretation of phenomena that are socially communicative—and thereby contextual—in character. However, as noted in the opening section, changes in ontology have cascading effects all the way through to method. As we intend an ontological shift, we must account for its effects.

35. In practical terms we are further obliged to develop a fully specified framework given the central role of language in the reconceptualisation. As language is a functional social phenomenon (Östman and Simon-Vandenbergen, 2009), it readily serves as both object and method of social inquiry. Nevertheless, to serve as object it must be conceptualised, and to serve as method it must be operationalised. Yet as language is the essential stuff of both society and its study, we risk a conceptual muddle with no clear path through. But we must pick a path, at each and every step, and thus we depend on fully specified framework (Kaplan, 1964, §§ 34–35). For a given subject of social inquiry, a conceptual framework is comprised of:

- **ontology**, which declares the objects implicated ;
- **epistemology**, which specifies how those objects are conceptualised ;
- **theory**, which suggests how those objects are understood to relate ; and further
- **operationalisation**, which suggests the phenomena in which we might perceive indications of those expected relations.

Note that in disciplinary work, frameworks in general are well established and many of these elements are assumed or have an accepted set of options—such delimitation and facilitation is the functional purpose of disciplines (cf. Kuhn, 1970, pp. 181–187). In interdisciplinary work this is not the case, and so a fully specified framework is obligatory.

36. Following the terms just described, the general subject of the reconceptualisation is the role of language in shaping social relations and collectivities. More specifically, the subject is how observation of language in public discourse can inform social inquiry by providing indications of shared meanings, and thus indications of possible contextual similarities, among the various discussants. From that subject of inquiry, the following conceptual framework is elaborated:

- the **ontological footing** (**chp. 2 § 2.3**) declares the fundamental objects of concern to this work (the <u>social</u> domain of meaning, the <u>material</u> domain of mediation, and the <u>technical</u> domain of affordance[18]) and describes how they are understood in themselves ;
- the **epistemological stance** (**chp. 3 § 3.3**) is derived from a reconceptualisation of political participation in hybrid society, and outlines the expected character of phenomena fitting that rubric ;
- the **theoretical approach** (**chp. 4 § 4.1.1**) draws on the preceding components to suggest a conceptual model of the subject of inquiry—in which the semiotic

affordances of language serve to bind the social and the material in a manner that is modulated by context—in a manner that is fitted to a site of research ; and

- the **operational step** (**chp. 4 § 4.1.2**) suggests specific phenomena of language—specifically pervasive but differential patterns of lexis and grammar—that might yield empirical evidence with which to evaluate the research questions.

Each component of the conceptual framework supports the components that follow, and the framework as a whole undergirds the method (**chp. 5**).

1.7. The Method

37. The method developed in this work gathers empirical evidence with which to evaluate the methodology itself—the topic of political participation itself is set aside. Recall that the purpose of the methodology is to reconceptualise a structural conceptualisation of a sociopolitical topic **1)** in a socially communicative manner that is suited to hybrid society, and **2)** in manner that remains interpretable from a structural perspective. Thus the method collects discursive and sociodemographic data of a common population within the site of research (the US state of Michigan), and subsequently processes those data so that the two categories may be directly compared. As will be explained in **chp. 6**, analysis of the processed data seeks evidence of a relationship between these two categories. A relationship is in fact observed, thus answering the research questions given in **§ 1.5** in a manner that warrants the methodology. In this light, note that there is nothing special about Michigan as the site of research—as will be explained in **chp. 4**, Michigan was chosen for practical reasons that made it a suitable and tractable case with which to address the research questions.

38. Given that the operationalisation hinges on language, and that **RQ$_1$** requires it to be amenable to computational and statistical analysis, the method is corpus-based.[19★] Such approaches to the study of language seek statistical associations in large quantities of text to reveal patterns that might otherwise be missed in close reading.[20★] While corpus-based approaches are in the main associated with the study of language itself, such approaches are well suited to the study of socially communicative phenomena (e.g. Friginal, 2018). Such approaches have great potential to further large-scale research as the social sciences come to grips with hybrid society, as this method aims to demonstrate. The method was inspired by, and is largely based on, a method developed by Douglas Biber (1988, 1992) to study linguistic variation across genres of text. Called Multi-Dimensional Analysis (MDA), that method is used to identify patterns of variation that are characteristic of genres and thus distinguish them. This method adapts that approach for the purpose of tracing social groupings. As our habits of language use are learned through socialisation and are themselves a central mechanism of socialisation (Ochs and Schieffelin, 2017 ; Schieffelin and Ochs, 1986), such habits are a function of context. It is thus reasonable to expect social groupings to exhibit characteristic patterns of variation.[21] In adapting Biber's method to the purposes of this work, this method effectively sees people as writing the 'document' of themselves by way of their contributions to public discourse. By identifying those human documents of similar 'genre' (i.e. having common patterns of variation), we have good reason to suspect that the people so represented share similar contexts.

39. The basic thrust of the method is this:

- within the site of research, collect a sample of public discourse that is attributable to <u>people</u> of known location (municipalities, in this implementation), as well as sociodemographic statistics characterising each location ;
- analyse the documents representing each person to produce a linguistic 'signature', then group together those people with similar signatures ; and finally
- compare the sociodemographic statistics across the linguistic groups.

40. Sociodemographic differentials are examined as a structural proxy for context. As will be discussed in **chp. 3**, the conventional study of political participation has long relied on measures of 'resources' and 'socioeconomic status' to categorise social groupings. We look for similar signals, except such exogenous measures are not our starting point. Rather, the method begins by deriving endogenous measures based on language with which to categorise social groupings. If we can observe regular differentials in sociodemographic measures across <u>linguistic</u> groupings, then we have reasonable evidence with which to answer the **RQ$_2$** in the affirmative, thus warranting **RQ$_1$**. <u>Such differentials are observed</u>.

41. The method has five stages:

1) **Data collection**. A broad swatch of public discourse is collected through the Twitter API (application programming interface) using a small set of politically and geographically oriented keywords. This collection was performed throughout the second half of 2018, yielding a primary collection of approximately 350 million tweets (comprising the textual component and metadata). During that time, sociodemographic and geographic data was collected for all officially recognised municipalities in the site of research from the US Census Bureau and the US Geological survey, respectively. These data were used jointly to prepare a geolocalised gazetteer of the site of research.
2) **Corpus preparation and enrichment**. The primary collection, having been sourced from the Twitter 'stream' (which is more or less global), is processed to yield a much smaller set of accounts (hereafter 'users') meeting a set of criteria, including creation data, localisation to Michigan, not being algorithmic or corporate, etc. This set of users was subject to a process of 'nominal localisation' (described in **chp. 5**). The process as applied here used account metadata to select those users that could be associated with a single known location (per the gazetteer) for the duration of the primary collection period. That set of localised users was passed to the secondary collection, which used the Twitter API to collect all publicly available tweets posted by those users. The secondary collection was trimmed according to the window of analysis (also the second half of 2018), resulting in a dataset of 37 million words across 2.6 million tweets associated with 5,889 users representing 417 known locations. The tweets of each user were compiled, cleaned, and normalised to produce 'user–documents', thus yielding the corpus itself. The corpus was subsequently enriched by linking the sociodemographic and geographic data from the gazetteer to each user–document according to location.
3) **Linguistic analysis**. Having prepared the corpus, the method proceeds in the manner of Biber's MDA—the constituent documents are analysed to assess the frequency of certain linguistic features.[22] The grammatical analysis and feature counting was performed using an application developed by Andrea Nini (2019) to replicate the process of Biber's original analysis. A lexical analysis was also performed. That process, to be described in **chp. 5**, identifies pervasive words in the corpus—that is, these words are both frequent and dispersed (i.e. not occurring all

in a clump). From the pervasive words are generated two lists, List 1 for pervasive words that are key among certain user–documents, and List 2 for words that are equally pervasive across all user–documents. These lists were prepared for single words and for bigrams. No tagging is required for the lexical analysis as the documents are composed of the tokens in question ; document feature counts are generated directly with these feature lists.

4) **Factor analysis and scoring**. Once all grammatical and lexical features have been tallied across the user–documents, the compiled scores are subjected to factor analysis. This step is a process of data reduction that serves to make the final clustering step feasible, and moreover to facilitate eventual interpretation. Factor analysis essentially calculates a lower-dimensional representation of the information contained in the feature scores ; the 'factors' in question are groups of features with similar variation that have been lumped together. The grammatical analysis yielded scores across 67 features, and the lexical analysis yielded scores across a total of 1,054 features. Factor analysis reduced these features to 8 factors and 32 factors, respectively. The grammatical features are treated using Principal Axis Factoring. This was the procedure used by Biber (1988, p. 82) and it remains the recommended procedure for MDA (Cantos-Gomez, 2019, pp. 99–106). The lexical features, for reasons to be explained in **chp. 5**, are treated using a technique introduced by Frank H. Walkey (1997) called Composite Variable Analysis. Following the results of the factoring, the individual feature counts are then compiled into 'factor scores'. The corpus is further enriched by linking these scores to each user–document.

5) **User–Document clustering**. Having produced linguistic 'signatures' for each user–document in the previous stages, they are now sorted into groups having similar signatures. This sorting is done using the *k*-medoids algorithm (Kaufman and Rousseeuw, 1990, chp. 2) at *k* 2, 3, and 5. The clustering itself is calculated on several 'variable sets', each representing a collection of factor scores. The base variable sets are the set of factor scores for the grammatical analysis, and the four sets of factor scores for the List 1 and List 2 words and bigrams. Further variable sets were produced from various combinations of these base sets. The final enrichment of the corpus is performed by linking the resultant cluster assignments to each user–document.

42. The enriched corpus itself comprises the results of the method. It gathers together the Twitter metadata of each user, the document comprising their publicly available tweets posted during the widow of analysis, sociodemographic and geographic data of their associated location, their linguistic signatures, and their resultant linguistic cluster assignments.[23] As described in **¶ 39**, the basic approach to analysis of these results is the comparison of compiled sociodemographic data across linguistic clusters. As the corpus has been appropriately enriched, it is the only data object needed for final analysis.

1.8. The Critical Character of this Work

43. There are portions of this work that are strongly critical of certain strands of work, disciplinary tendencies, and scientific philosophies. In the main this is critique, aiming to further the work of knowledge. But there is also outright criticism—it is reasoned, and it is sound both theoretically and empirically, but it is nonetheless rooted in anger. The reason for this is that the author believes—I believe—that social science matters, or at least should. It should matter to science, certainly, but foremost to society. Understanding society in terms of numerical aggregates, as John Dewey put it, does demonstrable harm to others thus eventually to ourselves. Our collective humanity depends upon resisting that tendency. Thus the critical portions of this work are not intended to offend or belittle. Rather they are motivated by a need to point out thinking that diminishes the humanity in the phenomena we study, and thus in ourselves. Following are the two underlying assumptions of this work, from which flow the bulk of critique and criticism.

44. It needs to be understood clearly by the reader that the first underlying assumption of this work is that researchers engaged in social inquiry—regardless of philosophical, theoretical, methodical, or practical differences—are motivated fundamentally by the desire to extend and enrich our understandings of the world in which we live. That is not to say that there are not countless other motivations and pressures that shape us, our disciplines, and our institutions. Indeed there are, and some of these serve to form the base of the arguments presented in this work. However, acknowledgement of these pressures should be interpreted with detachment. They are not intended as attacks or slanders, though certainly as critiques—but, foremost, these acknowledgements are intended to contextualise. That contextualisation can be viewed in two ways: analytically, in that this works seeks to provide context for better understanding the subjects addressed ; and reflexively, in that these descriptions (whether critical or not) also serve to contextualise this work and researcher for the reader (cf. Neumann and Neumann, 2015).

45. The second underlying assumption is that the social must be considered relationally. That is nevertheless a stance—not a claim of truth, but rather of perception (cf. Hick, 1983 ; Kant, 1781). A relational perspective on the social is no better or worse than a structural perspective. However, we see what we expect to see, and thus relational perspectives see relations and structural perspectives see structure. In part this tendency is a variety of the age-old scientific (and very human) problem of confirmation bias (Bacon, 1620, pp. 82–85). But, it should also be understood as a variety of availability bias (cf. Tversky and Kahneman, 1973). That is, we develop our contemplational and computational tools from a certain perspective, and thus that perspective is attached to and embedded in them (Law, 1992 ; Lievrouw, 2014 ; MacKenzie and Wajcman, 1999 ; Winner, 1980). Thus they are best suited to produce results fitting that perspective. Above it is observed that otherwise sound tools and techniques may not be fit for purpose if applied to the study of phenomena for which they were not conceived (**¶ 20**). For these reasons, this work foregrounds the relational perspective not because it is the preferred stance of the author (although it certainly is), but because that stance seems suited for studying sociopolitical phenomena in contemporary, hybrid contexts. For that very reason, the approach of this work is firmly rooted in the study of communication, as its various traditions and bodies of literature are well positioned, in terms theory, to address the changing character and underlying nature of the phenomena concerned and, in terms of method, to suggest productive paths to operationalisation in such contexts.

1.9. Document Overview

46. The remainder of the document is structured as follows. Each chapter addresses specific components of the methodology as outlined in **§§ 1.6–1.7** above. **Chapter 2**, "Theorising the Social–Technological Question", engages with perspectives on the relationship between society and its technologies in order to develop an ontological footing suited to the study of large-scale phenomena in hybrid society. **Chapter 3**, "Reconceptualising Political Participation in Hybrid Society", reviews the development of the scientific study of political participation and, considering emerging thinking and evidence, draws upon the ontological footing to propose an epistemological stance appropriate for hybrid society—this is the core reconceptualisation of the methodology. **Chapter 4**, "Operationalising Language in Mediated Public Discourse", draws on the previous components to elaborate the theoretical approach, which serves as the conceptual model of large-scale sociopolitical phenomena mediated by language. From that model the operational step is made, which suggests how exogenous measures of socially communicative phenomena might be derived from language. **Chapter 5**, "The Method", details how the suggested operationalisation is implemented for this work. **Chapter 6**, "Analysis and Evaluation", explains the approach to analysing the results of the method, performs that analysis, and concludes with an evaluation of the analysis. The evaluation finds that **RQ$_2$** is answered in the affirmative and thus that the provisional answer to **RQ$_1$**—the operationalisation suggested in **chp. 4**—is warranted, and by extension the methodology overall. **Chapter 7**, "Conclusion", provides a review of the methodology, a brief discussion of the contributions to knowledge of this work as well as its limitations, and concludes with suggestions of further work.

Chapter 2

Theorising the Social–Technological Question

The two phenomenal realms that we inhabit, as human beings, are the realm of matter and the realm of meaning. Human history is the unfolding of a constant interplay, and a constant tension, between these two.

M.A.K. Halliday, "On Matter and Meaning", 2005, p. 61

One may without exaggeration now speak of technological compulsiveness: a condition under which society meekly submits to every new technological demand and utilizes without question every new product, whether it is an actual improvement or not ; since under this dispensation the fact that the proffered product is the result of a new scientific discovery or a new technological process, or offers new opportunities for investment, constitutes the sole proof required of its value.

Lewis Mumford, *The Myth of the Machine*, vol. 2, 1970, p. 186

47. This chapter develops the ontological footing on which the conceptual framework of the methodology is built. It addresses the relationship between the social and the technological, reviews the development of conceptualizations and theories of the social–technological question, and provides the framing of that relationship which guides this work in relation to hybrid society. The chapter opens with an overview of thinking on the social and the technological, and then addresses milestones in the historical and modern development of the social–technological question. The current state of theory and methodology is compared in communication studies and science and technology studies. A possible integration of these multifaceted disciplinary approaches is proposed by way of a post-material framing of the social–technological question. This is followed by a brief discussion of key concepts from communication studies relevant to the post-material framing of technology proposed in this work. A brief description of this framing is then given ; thereafter is an elaboration of key terms to highlight conceptual complications, and to make explicit the terminological understandings employed here. Current approaches to the social–technological question face an essential challenge, in that deeply held notions of human and non-human do not map well to hybrid society. The chapter concludes with a formal description of the ontological footing that is proposed as a resolution to this challenge, and which supports the conceptual framework to be elaborated in subsequent chapters.

2.1. Framing the Social–Technological Relationship

48. The question of how the social and the technological should be understood in relation to each other is an enduring theme in the social sciences. For some disciplines it is a fundamental question, such as organisation studies (Kallinikos, Leonardi and Nardi, 2012) and science and technology studies (Sismondo, 2007). For others it is a question that grows more central given the increasing prevalence of technology in modern life, such as communication studies (Lievrouw and Livingstone, 2006). However, the changing nature of the social–technological question—by which is meant the shared understandings of the domains in question, and assumptions of agency, causality, epistemology, methodology, and so forth (Boczkowski and Lievrouw, 2007)—make the question one of perennial debate.

49. In many fields, the question is relatively recent, emerging in the mid-twentieth century in the context of increasingly technological environments, and has focused on the causal aspect as epitomised in the debate of technological determinism versus social constructivism. While 'hard' technological determinism is now readily dismissed in scholarship, if not in society generally (Marx and Smith, 1994), and the more extreme and reflexive forms of social constructivism are seen as revealing but relatively unproductive (cf. Latour, 2004 ; Winner, 1993), there remains an ontological tension between understandings of the social and the technological. Currently, the debate tends towards a middle ground, where the social and the technological are understood as inextricable and co-constitutive (Orlikowski and Scott, 2008). In that middle ground, causality is necessarily indeterminate (Pentland and Singh, 2012). To overcome this oppositional dilemma (Boczkowski and Lievrouw, 2007), theory has moved towards a conception of dynamic causality amidst shifting sociomaterial configurations (Jasanoff, 2004 ; Lievrouw, 2002). Such an approach is dialectical—the question of the social and the technological is seen as irresolvable, yet nevertheless an essential site of inquiry. These framings of the social–technological question thus often sustain the social and technological as distinct for reasons not only of analytical utility but also of disciplinary identity (Wyatt, 2007). Note that this work does not hold with this perspective—as elaborated in this chapter, the ontological tension is resolved, and effectively disappears, by taking an affordance perspective on the social–technological question.

50. It is important to note that the social–technological question was posed as the social sciences emerged in their modern disciplinary forms. Indeed, different approaches to the question were characteristic of emerging disciplinarity in the study of the species and its habits and habitats. What had been broad questions and suppositions of human lifeways began to be more specifically framed in the nineteenth century, while in the twentieth century specific framings and objects of inquiry multiplied rapidly as did disciplinary subfields and programmes of work (Wagner, Whitley and Wittrock, 1991). It should furthermore be noted that disciplinary consolidation occurred in a context of sweeping technological change and growth, which shaped the perspectives of scholars and laypersons alike. However, it was not always the case that the question was considered implicitly in terms of artefacts and materiality. That preoccupation (in a positive and negative sense) could be understood as a consequence of disciplinarity (Hickman, 1990, chp. 1 ; Scharff and Dusek, 2014 ; Wyatt, 2007). Before addressing modern (i.e. disciplinary) approaches, the following section provides a brief historical overview of the development of understandings of the social–technological question.

2.1.1. Historical Framings

51. The framing of the social–technological question in causal theory is a development of the last hundred years (Gunderson, 2016 ; Volti, 2004). Yet if we relax the causal and material aspects of the question, conceiving of change and technology in a broader manner, we see that the question has a lengthy history that permeates many of the fundamental sociocultural taxonomies through which Western thought has developed (Lenski, 1994).

52. Lewis Mumford (1961) argued that our understanding of the role of technology—from a historical, developmental perspective—has been too readily biased by the relative durability of material culture (meaning artefacts) given the ephemerality of non-material culture. In this regard he noted that the first scholars to attend to technology in light of sociocultural changes were archaeologists and anthropologists studying pre-literate societies, and thus by necessity privileging the artefact as a source of knowledge. While Mumford hoped that scholars had outgrown the fixation on technology as a distinct component of history—along the lines of Marx (Mumford, 1967, p. 4)—he warned that, "we must still be particularly on guard against residual tendencies to this kind of distortion of the evidence, in both early and contemporary cultures, because our own society is in fact over-determined by its excessive and almost exclusive preoccupation with technology" (1961, p. 232). This is not a constructivist statement, but rather a warning not to presuppose or project the centrality of technology as material artefact (2014, p. 382). Mumford argued that the fundamental technologies harnessed by our species depend not on artefacts, but on symbols. Such technologies include culture, language, and social structures for control and mobilisation, together enabling the assembly of 'human machines' that derive power not through artefacts but through social coordination (1967, chp. 1). This idea is strongly reminiscent of Lev Vygotsky's 'psychological tools'—symbolic artefacts such as signs, symbols, mnemonics, diagrams, maps, counting systems, and languages that serve to shape behavioural and cognitive processes individually and collectively (Kallinikos, Leonardi and Nardi, 2012, p. 10 ; Kozulin, 1998, p. 13). Such a perspective on the development of culture, society, and language finds a solid foundation in a variety of disciplines, such as anthropology, biology, ethology, linguistics, neuroscience, and psychology (Deacon, 1998 ; Fitch, 2005 ; Fitch, Huber and Bugnyar, 2010 ; Gibson, 1993 ; Hauser, Chomsky and Fitch, 2002 ; Humphrey, 1976).

53. This perspective does not, and is not intended to, diminish the social–technological question. Rather, it deepens it by allowing for a longer-term and more transdisciplinary view, which helps to be better situate current approaches to the social–technological question. While proper theory addressing the question is relatively recent (Erickson and Webster, 2012, p. 610), such developments are the outgrowth of thousands of years of thinking of the matter. This earliest thinking, however, is beyond the scope of this work.

2.1.2. Modern Framings

54. Two threads running through the historical development of thinking on the social–technological question were the predominance of taxonomic framings and an essential concern with social technologies. However, engagement with the question shifted as disciplines emerged and began to take their modern forms from the middle of the nineteenth century to the middle of the of the twentieth century. Disciplinarity reframed the social–technological question from one needing a descriptive answer to one needing an explanatory answer. As such, the technological would become more than simply the material artefacts that remained to be unearthed and ordered (Daniel, 1943, p. 8), and thus to be used as proxy for delimiting stages of cultural development (Morgan, 1877, p. 12). Rather, the technological would become a fundamental analytical category, reliance on which is characteristic of certain disciplines, such as science and technology studies, organisation studies, technical communication, communication studies, and information studies. The development of theory addressing the social–technological question has revolved around the causal balance between the social and the technological, and the conceptualisation and role of materiality. The following review is constrained to addressing the development of theory most relevant to the work at hand.

2.1.2.1. Technological Determinism

55. In the historical development of the social–technological question, there were some theoretical suggestions that technology was responsible for social change, for example Lewis Morgan (1877) and William Fielding Ogburn (1922). This is the basic idea of technological determinism. The most common understanding of the term, from both scholarly and public perspectives, is what Bruce Bimber called a 'nomological' account, meaning that it is culturally independent and based on assumed natural laws—technology stands apart from society and develops of its own accord, yet these developments force adaptation and change in society (1994, pp. 81–84). This would be called 'hard' determinism, in which agency to effect social change rested with technology in and of itself (Marx and Smith, 1994).

56. The technological determinist perspective on the social–technological question is often attributed to the work of Harold Innis and Marshall McLuhan in the 1950s and 1960s. Their work has come to be called 'medium theory', as distinct from media theory, in that it investigates the specific attributes of any given medium and how those attributes set the medium apart physically, psychologically, and socially from other media and other forms of communication (Meyrowitz, 1994, p. 50).

57. An early example is Innis's description of media as biased in terms of space and time (1950). He considered the structuring effects on societies of the physical properties of media, in that some were durable though less portable (such as those using stone), while others were highly portable but ephemeral (such as those using paper). The former case, Innis claimed, encouraged the development and persistence of conservative, authoritarian structures, while the latter encouraged adaptive, democratic structures (1950). Such development was the result of the nature of the media themselves—for media (and thus any message) to be temporally durable, it was necessary to compromise in terms of territory covered by that media ; to be expansive in terms of territory, it was necessary to compromise on the durability of media (and thus the persistence of any message). While Innis's idea of media bias is much deeper than a superficial appreciation of materiality (Comor, 2001), the material perspective was foundational to medium theory (Lievrouw, 2014, p. 39).

58. While Innis originated medium theory, his colleague Marshall McLuhan is most strongly associated with it. Innis had focussed on large-scale and long-term social effects of communication technology ; McLuhan, on the other hand, was concerned with effects on and affordances in terms of human agency and perception (Lievrouw, 2014, p. 40). This concern is represented by McLuhan's idea of media as 'extensions of man'. By this he meant that media do not endow us with heretofore unknown capabilities, but rather only increase the natural capabilities of our bodies, senses, and minds. The idea was not new ; consider, for example, the opening sentences of Emerson's *Works and Days*:

> Our nineteenth century is the age of tools. They grow out of our structure. "Man is the metre of all things," said Aristotle ; "the hand is the instrument of instruments, and the mind is the form of forms." The human body is the magazine of inventions, the patent-office, where are the models from which every hint was taken. All the tools and engines on earth are only extensions of its limbs and senses. (Emerson, 1870, p. 141)

59. McLuhan's concept, however, is much more expansive, as seen in his (and Innis's) use of 'medium'. The public (and oftentimes scholarly) understanding is that McLuhan uses 'medium' in the sense of communication technologies, such as print, radio, television, and so forth. While this is certainly the case—he addresses such media in *Understanding Media*—he also included as media light bulbs, roads, money, clocks, bicycles, and firearms, for example (1964). McLuhan stated that by 'medium' he indicates "any extension of ourselves", and that 'extension' and 'technology' are synonymous (1964, p. 7).

60. But what is conveyed by these media? A common, Lasswellian understanding would be that media (such as print, radio, television, and so forth) convey 'content'. But McLuhan asked, what of the light bulb? It conveys no content, only information (i.e. the binary on or off). However, light bulbs could be said to convey content if grouped together in a sign. McLuhan countered that idea by pointing out that media do not convey content, but only other media. That is, the lighted sign could be shaped into words, but written language itself is a medium. That medium coveys spoken language, which itself conveys thought (1964, pp. 7–9). This regress (going ever deeper if we interrogate thought as a medium) is the reason that McLuhan sought to decouple media from the idea of content (which itself is media).[24] How then should we understand the famous dictum "the medium is the message"? Having already clarified his use of 'medium', McLuhan explained that “the ‘message’ of any medium or technology is the change of scale or pace or pattern that it introduces into human affairs” (1964, p. 9). Framed in this manner, “the medium is the message” is best understood as the essence of medium theory—to study the material nature of media, as that is what dictates its social impact.

61. Medium theory, then and now, has kept the question of material technology at the heart of its investigations. However, it is this aspect of the approach that finds ideas in the vein of McLuhan and Innis accused of technological determinism, which is generally, though not exclusively, meant in a pejorative sense (Lievrouw, 2014 ; Wyatt, 2007). Nevertheless, medium theory had substantial impact, and a general sense of technological determinism held sway in academia until the 1980s and remains in society and mass media to the present day (Lister *et al.*, 2009 ; MacKenzie and Wajcman, 1999 ; Marx and Smith, 1994 ; Winner, 1977).

2.1.2.2. Social Constructionism

62. In the 1970s, resistance was growing to technologically deterministic approaches to the social–technological question. Raymond Williams was an early voice speaking out against implicit assumptions of technological determinacy. A common perception is that the arrival of major technologies brings about new historical, sociocultural periods (consider the so-called ‘ages’ of stone, bronze, iron, industry, information, etc., and the use of the material to delimit cultural stages, as noted in § **2.1.2.1**). In his study of *Television* (1974), Williams rejected such folk notions of cause, insisting that we must interrogate what it means to attribute cause to technology, whether that is reasonable and thus what that even would mean, and how such causes stand in relation to each other. As to questions of effect, any answer would be of token value without an appreciation of the web of cause and effect that binds technology and culture (1974, pp. 1–2). Williams held that technology is always a product of a specific sociocultural context and that pathways of social–technological change are never determinant but contingent (Freedman, 2002). McLuhan and Williams debated these issues for some years ; in the end, it was Williams’s contextual approach that would come to characterise the study of the social–technological question (Lister *et al.*, 2009, pp. 14–15).

63. The 1980s saw a significant move away from technological determinism towards social constructionism.[25] The idea of social construction is that phenomena have no reality external to the sustained, discursive practices of human behaviour (Berger and Luckmann, 1966 ; Sismondo, 1993). An important milestone in this move is 'social shaping', introduced by Donald MacKenzie and Judy Wajcman in the collection *The Social Shaping of Technology: How the Refrigerator Got its Hum* (1985). Social shaping is not a single coherent theory, but rather an approach that emerged in a certain context that aimed towards building a "broad church" of fields and methods to motivate and assist in the move away from technological determinism (Williams and Edge, 1996, p. 892). There is no agreed definition of what constitutes this approach, besides a rejection of technical determinism, although a convergence of methods has been noted (Russell and Williams, 2002, p. 37). Fundamentally, the social shaping approach aims to expose and analyse the "socio-economic patterns embedded in both the *content of technologies* and the *processes of innovation*"—a central focus of the approach is to highlight that content and process are contingent on choices that, while technical, are always socially conditioned (Williams and Edge, 1996, p. 866, emphasis original). This distinction between the social and the technological shows that while this approach is highly constructionist, it is not purely so nor is it meant to be. MacKenzie and Wajcman decided on the term 'shaping' specifically to avoid the term 'construction'. They wanted to emphasise that the approach stressed the independent (i.e. material) reality of the technological, and thus was compatible with realist approaches. Furthermore, they wanted to emphasise that technology can be a social product without necessarily being impacted by broader social relations and concerns—for MacKenzie and Wajcman, social construction implied external (i.e. societal) factors, whereas social shaping implied more 'local' factors (MacKenzie and Wajcman, 1999, pp. 18–19). It should be noted that the various strands of study of technology have never embraced fully constructionist approaches despite the social turn in the 1980s and 1990s. Whereas the social shaping approach, for example, is something of an umbrella term for approaches to the technological that posit a central role for social factors (Williams and Edge, 1996), such approaches have not superseded the concern with materiality that has influenced theory and method (Lievrouw, 2014 ; Wyatt, 2007).

2.1.2.3. Materialism

64. While constructionist approaches to the social–technological gained significant momentum in the 1970s and 1980s, other approaches were developing that sought a middle ground between determinist and constructionist views. Indeed there was relatively sudden scholarly migration towards the technological (Forman, 2007). This 'new sociology of technology' remained fundamentally constructionist, but sought to foreground attention to technology as artefact in theory and empirical work (Bijker and Pinch, 2012, p. *xvi*).

65. Among these developing approaches it is important to note the social construction of technology. As introduced by Trevor Pinch and Wiebe Bijker (Bijker, Hughes and Pinch, 1987 ; Pinch and Bijker, 1984), the approach has three main stages of inquiry into the study of an artefact. The first is that of *interpretive flexibility*—to consider the development of the artefact as a process of variation and selection, resulting in a model of development that is multidirectional rather than linear. This hinges on a symmetrical approach to success and failure, attending not only to instances of uptake and further development, but also instances of neglect and failure. In addition, the flexibility emerges from the perspective that no artefact has fixed meaning or utility ; rather this is determined by relevant social groups, identified by shared interpretations of the artefact. The process of variation and selection in the development of an artefact is the 'negotiation' between competing interpretations of social groups. The second stage is that of *stabilisation and closure*. Variation in an artefact reduces, or stabilises, as the interpretations of social groups shift and merge. In this process, it should be noted that not only do social groups influence the development of artefacts, but the artefacts likewise influence the social groups. Closure is that point where sufficient stability is reached that development essentially ceases, as no significant competing interpretations remain among the social groups (in that they are identified by their interpretations, essentially only one relevant group remains). The third stage is relating the whole of the evidence to the *wider context*, that is, back to established social–technological theory.

66. Another approach that sought to reconcile the realist–constructionist tension inherent in the social–technological question is actor–network theory. As associated with

Michael Callon, Bruno Latour, and John Law, actor–network theory is not a theory in the explanatory sense, but rather is a collection of "material–semiotic tools" and methods to analyse all phenomena as relational constructs (Law, 2009, p. 141). By all phenomena is meant just that—the material, accounting for artefacts, nature, and physical reality ; and the semiotic, accounting for the relational, ideational, and immaterial. Although actor–network theory has taken many forms in its development, there are four tenets that characterise the approach. The first is *semiotic relationality*—phenomena are conceived as networks of actors that define and influence each other through communicative interaction. The second is *heterogeneity*—network actors are of different types, human and non-human. The third is *materiality*—a distinguishing feature of actor–network theory is the concept of "generalised symmetry", that no distinction is made between human and non-human actors in attributing agency in the network (Callon, 1986, pp. 4, 17). The fourth is *precarity*—actor–network theory does not view phenomena as permanent or even static, but rather as emergent effects sustained and shaped by relations in a network.

67. The renewed focus on technology as artefact—or the "turn to technology", as Woolgar termed the shift in the context of science and technology studies (Woolgar, 1991)—should not be understood as a re-emergence of technologically deterministic thinking, despite recent work that suggests a 'rehabilitation' of the concept (Dafoe, 2015 ; de la Cruz Paragas, Fernando and Lin, 2014 ; McCarthy, 2013 ; Wyatt, 2007). Rather, the perspective seeks to situate the technological as both socially emergent and embedded. In this, the materiality of concern was gradually expanded from the artefact in and of itself as the unit of analysis. The social construction of technology approach just discussed is a case in point, with a shift in focus from the artefact itself to the artefact and its supporting social structures (Bijker and Law, 1992) and in level of analysis from the artefact to 'sociotechnical ensembles' (Bijker, 1995). Bijker (2010) provides a succinct account of this shift in approach, noting how it developed from a focus on specific artefacts to a focus on technological culture, and from a social construction perspective to a more material co-production perspective. This expanded perspective, as represented here by the social construction of technology and actor–network theory, is now well recognised and utilised by researchers addressing the social–technological question (Sovacool and Hess, 2017).

2.1.3. From the Technological to the Post-Material

68. For this work, the social–technological question must be adapted. Of primary concern is specifying the technological. Technology as an analytical construct is certainly productive—be it as artefact or knowledge (Kline, 1985), process (Orlikowski, 2007), institution (Bijker and Law, 1992), system (Hughes, 1987), or culture (Jasanoff, 2004). However, such conceptualisations are biased towards operationalising technology in a manner that is isolating or compartmentalising the technological from its context (Kallinikos, 2004 ; Orlikowski, 2007 ; Suchman, 2007). For some fields, such as science and technology studies, this bias can be viewed as necessary from a disciplinary perspective and thus characteristic (Wyatt, 2007). However, Thomas Misa (1994) maintains that the degree to which technology is conceptualised as a thing unto itself is a function of level of analysis. In the context of scholarship on the history of technology, Misa observes that macro-level analyses tend to impose order through abstraction from specific cases and the assumption of rationality of actors and functionality of their actions, thus leading to deterministic accounts. Micro-level analyses, on the other hand, in seeking to examine contingency in the variety of specific cases, naturally lead to non-deterministic accounts. The trajectory of approaches in organisation studies provides a compact example. Early studies, such as those of Joan Woodward (1958) in manufacturing and of Charles Perrow (1967) in hospitals, took a rather deterministic stance towards overall organisational structure as an outcome of technology. Later work descended from the macro-level perspective of the organisation as unit of analysis and looked inside the organisation to study how structure and practice were negotiated in the context of technology. The result was a shift towards understanding outcomes as contingent on social factors (Barley, 1986 ; Hatch, 1993 ; Orlikowski, 1992 ; Orlikowski and Robey, 1991).

69. The conceptual shift towards the material has been taken up in organisation studies. There were concerns that understandings of contingency that too heavily privilege the social could not make room for material phenomena for which social effects are not an adequate explanation (Hutchby, 2001 ; Kallinikos, 2004 ; Pickering, 1995, 2001). Such observations are consonant with recent 'ontological' concerns in science and technology studies—that approaches privileging the social risk misinterpreting a world that is materially situated (van Heur, Leydesdorff and Wyatt, 2013 ; Woolgar and Lezaun, 2013).

In addressing this issue, organisation studies has drawn heavily on science and technology studies in integrating social and material contingencies, with theoretical extensions suited to the scrutiny of the nexus of the social–technological question—situated at Misa's "middle-level" between technological determinism and social construction (1994). Fundamental to these extensions is the repurposing of the term 'material'. In contrast to the study of technology in which materiality is more readily intuited in physical reality (e.g. the material effects of a hammer can be understood as stemming from the weight and hardness of its head, the length of its handle, etc.), organisation studies, especially research focusing on information technology, has grappled with aspects of technology that have organisational effect but that exist outside of social practice (in terms of application, not development) and that have no distinct physical reality. Software, for example, embodies features, such as interfaces, algorithms, or tracking functions, that do organisational work ; these aspects are persistent and effect organisational structure and social practice even without a physical reality unto themselves (Leonardi, 2007 ; Orlikowski, 2000 ; Volkoff, Strong and Elmes, 2007).

70. The use of the term 'material' to refer to such aspects underlines that their organisational effects are intrinsic to the technology—while these aspects must be at some point enacted or utilised through the medium of physical "bearers" such as a computer monitor or keyboard (Faulkner and Runde, 2011), they have no independent existence outside of that manifested in socially situated use. This framing of materiality has no regard for physicality. Rather, the crucial notion in this understanding of materiality is that intrinsic, persistent aspects are manifested in context, but do not depend on context. While effects necessarily depend on context, the material aspects do not change ; they are stable across space and time, regardless of physical reality (Leonardi, 2012). The material aspects of a given accounting software remain the same in a home office or government agency, just as the material aspects of a given hammer remain the same whether used as a tool one day and a weapon the next. This framing allows for an essential constructionist perspective, in that effects are dependent on the social context in which they are made manifest, yet furthermore enables aspects of effect to be detached conceptually from both physicality and social context. Materiality then can be understood as potentiality that manifests differentially across space and time.

71. The differential manifestation of potentiality is subsumed in what Wanda Orlikowski (2007) terms "sociomateriality". She holds that a perspective of distinct phenomena complicates the incorporation of materiality into organisational research. To move beyond perspectives on the social–technological that privilege either side requires the relationship to be understood not as causally unidirectional, as in extreme determinist or constructionist stances, or even as reciprocal (Barad, 2003 ; Jasanoff, 2004), but rather as fundamentally inextricable. Orlikowski stressed that "there is no social that is not also material, and no material that is not also social"—they are "constitutively entangled" (2007, p. 1437). This perspective was framed as an extension and generalisation of preceding work—including on actor–networks (Callon, 1986 ; Latour, 2005), sociotechnical ensembles (Bijker, 1995), and relational materiality (Law, 2004)—which Orlikowksi described as 'post-humanist' in that it sought to decentre the social from the study of the empirical world (2007, pp. 1437–1438). Similarly, such work can be viewed as 'post-material' if we again consider the navigation of levels of analysis across fields as discussed just above.[26] This latter characterisation coheres with the sociomaterial stance against a discretised (or dualistic), rather than relational, ontology of the social–technological. The sociomaterial approach has become a significant and growing strand of research in organisation studies and information studies during the last decade (Jones, 2014).

2.1.4. Framing Information and Communication Technologies

72. For this work, sociomateriality and its precedents are significant in that they all take a fundamentally relational perspective—what Law refers to as the "semiotic insight" that all things are relational and thus are produced in relations (Law, 1999, p. 4). In seeking to situate social phenomena (political participation, in this work) in the context of mediated public discourse (enabled by and enacted in a matrix of information and communication technologies), both subject and object of study are inherently relational in concept and in practice. From a conceptual perspective, the relationality of information and communication technologies runs deeper than simply enabling ties among entities. As Leah Lievrouw states, "Communication technologies—at once resources for and manifestations of communication, meaning, and culture—[seem] to epitomize the articulation between the technical and the social" (2014, p. 22).[27] Such technologies and the practices that emerge in their use blur analytical distinctions such as the human, technological, or material, in part due to their ubiquity—not only are information and communication technologies extensive sociotechnical systems in their own right, they are embedded in almost all other sociotechnical systems and thus pervade modern life (Boczkowski and Lievrouw, 2007), and thus enable the emergence of hybrid society.

73. The semioticity of information and communication technologies provides a lens through which to examine the social–technological question in communicative dimensions, and likewise a bridge into the literature of communication studies. Roger Silverstone's concept of 'double articulation' is illustrative in this regard: it refers to "the ways in which information and communication technologies, uniquely, are the means (the media) whereby public and private meanings are mutually negotiated ; as well as being the products themselves, through consumption, of such negotiations of meaning" (Silverstone, Hirsch and Morley, 1992, p. 28). Recalling McLuhan: "through its double articulation, the medium does become the message, though that message is not pre-given by the technology" (Silverstone, 1994, p. 83). This framing of information and communication technologies was central to early work in domestication theory, which sought to examine the everyday mediations in the articulation of the material and the symbolic (Silverstone and Haddon, 1996). While theoretically sound, double articulation was difficult to operationalise, and revealing such mediations proved difficult (Livingstone, 2007). In part this could be understood as a difficulty of the level of analysis, where the specificity of articulated contexts requires increasingly ethnographic approaches. This move can be seen in the extension to 'triple articulation' where, in addition to attending to the meaning of objects and symbolic environments, one further addresses individual instances of message and meaning-making (Courtois *et al.*, 2011 ; Hartmann, 2006 ; Silverstone and Haddon, 1996, p. 74).

74. Approaches that engage with content rather than form or materiality are characteristic of communication studies, although there are calls for the integration of more material approaches (Baldwin-Philippi, 2011). There is specific interest in the potential productivity of dialogue between communication studies and science and technology studies. In the study of information and communication technologies, Pablo Boczkowski and Leah Lievrouw (2007) identified three "bridges" between the fields, these being shared notions of or concern with 1) causality in the social–technological relationship, 2) processes of technological development, and 3) social consequences of technological change. These bridges have allowed for communication studies to gain from conceptual language and methods with which to engage the material, and for science and technology studies to benefit from an extensive body of scholarship at more situated levels of analysis. Nevertheless, deepening and expanding these links is essential given the deepening ubiquity of sociotechnical mediation blurring the lines between traditionally distinct domains of study (2007, p. 965). Such 'cross-pollination' appears to have taken place among researchers working in 'New Media', which Lievrouw and Sonia Livingstone have characterised as information and communication technologies and their social contexts—comprising of artefacts, activities and practices—and the social and organisational patterns that emerge around artefacts and practice (2006). Considering this definition, it is no surprise to see convergence in understandings of social–material interrelations and articulations (Lievrouw, 2014).

75. Nevertheless, while a sociomaterial, or at least reciprocal, understanding of the social–technological question is now well established in science and technology studies in regard to new media, communication studies in the main retains a perspective that—even when attendant on technological or material concerns—tends towards sociocultural understandings. The reasons for this latter perspective are varied and are beyond the scope of this discussion. However, the ontological concerns in science and technology studies mentioned in **§ 2.1.3**— that approaches privileging the social risk misinterpreting a world that is materially situated (van Heur, Leydesdorff and Wyatt, 2013 ; Woolgar and Lezaun, 2013)—are equally valid for communication studies. Although a focus on ontological concerns is itself a constructionist exercise (Aspers, 2015 ; Sismondo, 2015), a deeper integration of the material in the study of communicative relationality—in the manner of a 'balanced' sociomaterial understanding of the social–technological question—could be advantageous for both fields. A possible pathway to such integration through an affordance perspective is discussed in the following section.

2.2. Framing this Work

76. While both communication studies and science and technology studies are increasingly engaged with 'New Media' (Boulianne, 2015 ; Cihon and Yasseri, 2016 ; Schultze, 2014), this work is not concerned specifically with information and communication technologies and their associated contexts and practices (cf. Castells, 1996 ; van Dijk, 1999). Rather, this work casts a wider net, attending to emergent communicative, and thus social, structures (Lomborg, 2017) in contexts understood to be thoroughly technologised and mediatised (Chadwick, 2013 ; Couldry and Hepp, 2013, 2017) such that the physical and virtual are constantly blended (Blommaert, 2019).[28] This work denotes such contexts with the term hybrid society.

77. Thus there is advantage in an integration of communication studies and science and technology studies—the former to provide a critical and thick sociocultural perspective (Geertz, 1973) to the study of empirical life among ubiquitous technology (Srivastava, 2004), and the latter to provide a material and embodied perspective to the study of activities and practices in environments that are essentially, and empirically, disembodied (Turkle, 1995).

78. From this perspective of pervasive mediation and embodiment, such integration—that is, transdisciplinarity—can provide conceptual space and structure for the incorporation of knowledge from other fields necessary to this work. The relational–material understanding suggested here, achieved through the integration of communication studies and science and technology studies:

- provides a discursive space for linking bodies of scholarship while minimising difficulties that might arise from competing disciplinary matrices (Craig, 1999 ; Kuhn, 1970) ; and furthermore
- allows topics hinging on the symbolic to be more readily articulated at greater degrees of materiality (i.e. complexity) and thus higher levels of analysis (Leonardi and Barley, 2010 ; McLeod, Kosicki and McLeod, 2010 ; Misa, 1994).

79. Nevertheless, given the tendency of communication studies to privilege social understandings, and the tendency of media studies and science and technology studies to privilege material understandings, an ontological tension remains between the fields that must be addressed. That said, Bijker and Pinch (2012) note that an argument can be made that constructionism is unavoidable, and thus the ontological issue becomes an issue of epistemology. From the perspective of this work, an epistemological interpretation is fitting, so long as it rests upon an ontological foundation amenable to the fields concerned. As noted at the end of the preceding section, the concept of affordance is a potential pathway towards integration.

2.2.1. An Affordance Perspective

80. The concept of affordance originated with the work of James Gibson (1979) on animal perception. Affordance describes the qualities of an environment as perceived by an animal in that environment. Specifically, the concern is what the environment is perceived to offer to the animal, whether or not beneficial. Thus, open ground offers a place to rest, a tree offers shelter, and water offers refreshment. At the same time, open ground offers exposure, a tree offers concealed predators, and water offers drowning. A certain fruit might be camouflage to one animal, nourishment to another, and poison to yet another. These are all affordances of the environment, contingent on the nature of the perceiving animal, on what is or is not perceived, and on what has been learned or understood about what is perceived. They are not properties of either the animal or the environment. Rather, they emerge in contextual relation of the two, which Gibson framed as “complementarity” (1979, p. 127). However, Gibson did consider that the affordances offered by a thing remained unchanged regardless of the needs of the observer (such as the edibility of a certain fruit by certain animal remaining constant) ; this appears to weaken the relational perspective in that it rests on essential characteristics of the object and observer (1982, p. 409).

81. Ian Hutchby (2001), in proposing the concept of affordance as an approach to the study of technology that reconciles constructionist and determinist (or realist) positions, countered this essentialist understanding. He elaborated four aspects of affordances that are key to extending the concept to technology, while avoiding the attribution of essential characteristics. Each aspect builds from the one before. First, affordances are *manifold*. Gibson described affordances deriving from the environment in a natural sense ;

considering the environment broadly as the empirical world, affordances derive from many sources: from the natural environment, from animals different from the observer, from animals like the observer, from the self, from events, and so on. Events as a source of affordances are important to consider more closely. While this could be understood as strokes of good luck (pot of gold) or bad luck (asteroid impact), it is more the case to consider the spatiotemporal arrangement of the phenomenal world and observers therein at any given moment—sources of affordances are interrelated, interacting, and contextually specific.

82. Second, affordances are *functional* and *relational*. They are functional in that they enable or constrain possible pathways of action of an observer. They enable activities that could not readily be accomplished otherwise. For example, to ascend a cliff might not be possible for a given observer, with the cliff offering a negative affordance of movement. However, there could be a pathway up, a tree nearby, a rope, a ladder, an elevator, and so forth. These entities would offer a positive affordance in terms of ascending the cliff. There are constraints as well, in that the intended activity might still be enabled, but the affordances constrain, or channel, how the activity is accomplished. In their functionality, affordances are relational in that they differ according to observer. The cliff would offer little negative affordance to a bird or mountain goat, and indeed might offer positive affordances. A tree might be easily scaled by a monkey, but not by an elephant. A rope or ladder might readily afford use by an animal with appropriate appendages, such as a monkey, but might offer no affordance at all to a turtle or snake. The elevator is a more complicated case. A variety of observers could potentially make use of its affordances, provided they interacted with it in the appropriate manner either by luck or through prior knowledge.

83. Third, affordances are *learned*. That is, although they emerge in being perceived, they are only actualised in the performance of activities. The sense of learning is rudimentary in the case of many observers. Consider the case of fruit that is potentially nourishing or poisonous. An animal of one species that eats it and lives might eat it again in future. An animal of another species that eats it and dies will not have that opportunity. From an evolutionary perspective, the members of a species that correctly perceive an affordance (positive or negative) are more likely to reproduce, while members that incorrectly perceive, or do not perceive, an affordance will be less likely. Over generations (assuming survival), the correct perception of an affordance (the edibility or inedibility of a fruit, for instance) will predominate, by virtue of the inheritance of whatever mechanism enabled that perception. In this sense, a species can be said to have 'learned' the affordance. Active learning is of course observed in animals as well, for example in finding and returning to sources of water and food. In the case of human activity, the process of learning—and teaching—is more clear. The elevator just mentioned provides certain positive or negative affordances given patterns of interaction with it. Step in, press the correct button, and one ascends the cliff. Press the incorrect button and the door will remain open, or the elevator will not move, or an alarm will sound. Put a finger, arm, or head in the wrong place at the wrong time, and they might be injured or lost. Attempt to ascend simultaneously with too many other observers and the system might be blocked from functioning or be damaged. These various affordances can be learned from experimentation by an individual observer, or learned from others who have themselves experimented, or have themselves learned from others, and so on.

84. Fourth, affordances must be *interpreted*. This aspect is related to perception of the affordance. In Gibson's formulation, an affordance inhered to a thing, remaining potential until it was observed and acted upon by an observer. While this understanding might be problematic if considered as an essentialist framing, it should be remembered that Gibson was concerned primarily with animal behaviour. Considering more intentional observation, such as in human activity, the question of perception goes beyond the sensory or the context of stimulus–response. Donald Norman stressed the perception of affordances in the sense of knowing, intuiting, or comprehending them. From his perspective, if an affordance went unknown or unrecognised, it essentially did not exist. This understanding highlights that affordances do not depend on the physical nature of that which affords, but rather on their semiotic nature. Affordances, according to Norman, need to "communicate" (2007, p. 68). This is an important point in terms of human artefacts. Humans build affordances into objects, and we come to recognise such affordances in these objects. But further, the experience of affordances (consider the other aspects above—the learning of their functionality, relationality, and sources) provides crucial heuristics for humans in their world of artefacts. Our lives depend in myriad ways, large and small, on the appropriate interpretation of artefacts and their affordances.

Norman’s crucial argument, however, is not that we all must correctly interpret affordances. Rather, he argued that—as designers of artefacts—we interpret *into* artefacts that which we wish them to afford ; the materiality of created artefacts (again, not necessarily physical) reflects their intended affordances. This reflection, of course, can be more or less intentional. Furthermore, the reflection is not necessarily interpreted in the intended manner by other observers. This is key in applying affordances to the question of technology (Hutchby, 2001, 7ff): regardless of the affordances interpreted into an artefact in its design and manufacture, or the affordances that are communicated to others in regard to that artefact, observers nevertheless interpret affordances in the artefact according to their context and needs. The affordances offered by an artefact, and realised in use, are a product of negotiation between the material (physical, factual) and the social (intentional, symbolic) aspects of a given context.

85. The affordance approach has gained broad currency as an analytical concept in regard to human interaction with technology (Gaver, 1996 ; Norman, 1988, 1993), including in the study of organisations (Fayard and Weeks, 2007 ; Leonardi and Barley, 2010) and information systems (Leonardi, 2011 ; Volkoff and Strong, 2013). Sociomaterial approaches to the social–technological can also be understood from an affordance perspective (Faraj and Azad, 2012), especially when considering the experimentation and adaptation that shape, but do not determine, practices emerging around technology in use (Gaver, 1991 ; Leonardi, 2011). In the study of communication, the application of an affordance perspective is readily found in the subfields of technical and computer-mediated communication (boyd and Ellison, 2007 ; Leonardi, Huysman and Steinfield, 2013 ; Majchrzak *et al.*, 2013).

2.2.1.1. Affordance as a Relational Ontology

86. The application of affordance to the study of technology is varied as are the understandings of affordance. This is partly because there is no unified theory of affordance, “as they emerge in the mutuality between those using technologies, the material features of those technologies, and the situated nature of use” (Evans *et al.*, 2017, p. 36). The concept of affordance alone does not provide a foundation for theory, in that it is situated as opposed to universal, and furthermore that it is fundamentally acausal

(Craig, 2013 ; Sutton and Staw, 1995). At root, an affordance perspective is an ontological perspective that balances the potential functionality of the material with the organisational and agentive interpretivity of the social. This is the perspective taken in this work. However, it is necessary to articulate this perspective in order to undergird the epistemological stance to be developed in **chp. 3**. In that the affordance ontology is relational, those relations will now be specified.

87. A common thread running through understandings of the social–technological question is the issue of change and the attribution of causality. How does change in the social or the technological come about, and what are its drivers? At one extreme there are the technological determinist perspectives. In the 'hard' versions, technology is a separate and self-directing *primum movens* that is both impetus and channel for social change. In the 'soft' versions technology is acknowledged to be subject to an interplay with societal forces, but nevertheless exhibits a certain degree of separateness and even autonomy (Marx and Smith, 1994 ; Winner, 1977). At the other extreme are social constructionist perspectives, that can be categorised as 'radical' in that they deny external material reality outside of discursive, social relations, or as 'mild' in that they allow for technological influence but privilege the role of sociopolitical forces in shaping outcomes (Marx and Smith, 1994 ; Pinch and Bijker, 1987 ; Sismondo, 1993 ; Winner, 1977). Between these poles there are those perspectives that aim to overcome the theoretical limitations and empirical challenges to dualistic formulations of the social–technological question—be they society–technology, social–material, subject–object, structure–agency, determinism–voluntarism, constructionist–realist, etc. (Faraj and Azad, 2012 ; Orlikowski, 2000 ; Smith and Marx, 1994).

88. At lower levels of analysis—in spatiotemporal terms, material terms (system–artefact–component), or social terms (culture–collective–individual)—causality can be understood and theorised in terms of agency. Moreover, causality is a theoretical construct that seeks to abstract situated agency. As discussed in **§§ 2.1.2.2–2.1.2.3**, the attribution or proper localisation of agency has long been key to the debate surrounding the social–technological question. In many ways, understandings of agency and its manifestations underlie the material and post-material approaches to technology (Kallinikos, Leonardi and Nardi, 2012 ; Kaptelinin and Nardi, 2006 ; Leonardi, 2011 ;

Pickering, 1995). Current debates in this regard pivot around what is often portrayed as an essential tension between technological/material agency and social/human agency. However, as Paul Leonardi (2012) points out, the relational perspective of affordance allows us to sidestep such sticking points, as affordances are understood to arise at the intersection of social and material influences. But such an interstitial framing needs to be more than a conceptual gloss. To state that affordances arise in between the social and material does little to illuminate the relation between the social and the material other than to posit that a relation exists. Thus, there is no analytical utility to consider that an affordance might arise *ex nihilo*. We must specify what a relation is considered to be. In terms of causality and agency, relations of affordance will be understood as functional. In this sense, affordances are more than simply potentials of action (Hutchby, 2001) ; to consider the potentials of action provided by an environment (that is, the phenomenal environment and all within it) is an exercise in the hypothetical. Rather, affordances exist in the "domain of the real" (Volkoff and Strong, 2013, p. 822). The reality of affordances hinges on capability and intent—that is, in a given environment there must exist an actor that is able to perceive and to perform the realisation of an affordance (Chemero, 2003) and, crucially, such a capable actor must have an intent or goal that is materialised through that affordance (Stoffregen, 2003). In this sense, affordances are realised and materialised in the world through the invitation and gratification of agency (Withagen *et al.*, 2012 ; Withagen, Araújo and Poel, 2017). Going a step further, for the purpose of this work, it is suggested that affordance is the <u>mechanism</u> of agency.

2.2.1.2. Affordance as the Mechanism of Agency

89. To consider affordance as the mechanism of agency, the latter term must be deconstructed. In the context of discussing sociomaterial understandings of the social–technological question, Leonardi (2011) distinguishes between human agency and material agency. The two are intertwined (as Leonardi aimed to describe sociomaterial entanglement) yet distinct. Human agency is defined as the ability to form and realise goals, often in the navigation of the benefits and constraints of material agency. Material agency is defined as the ability to act absent human intervention. Leonardi (2012), citing Andrew Pickering (2001) and James Taylor *et al.* (2001), draws the distinction between the two that human agency is intentional, whereas material agency is devoid of intention.

Humans, individually and collectively, through the impetus of biology or culture, formulate goals and plans to achieve those goals. They then seek to realise these, often through material agency. Material agency, while it can act in the real world, has no inherent impetus to formulate goals other than which might be given to it.

90. In distinguishing between the human and material, the question remains of where to draw the line. Some might argue from the standpoint of technology of increasing complexity and capability that might exhibit unexplained, even seemingly intentional, behaviour. A more salient interrogation is on the human side—to what degree are we as entities human versus material? If one uses a lever to accomplish a task, the lever is certainly a material component. But is not the hand a more proximate material agent? Did the hand itself decide to manipulate the lever, or was the hand the specific tool chosen by the mind to manipulate the lever? A similar question could be posed for the voice, or for vision, or for any embodied interactant.[29] Furthermore, what of systems over which we have no direct control that comprise the human body, such as those governed by the autonomic nervous system (cf. Maturana and Varela, 1980)? What if an individual acts automatically because another individual has instructed them to do so? What if an individual is used bodily by another too accomplish some end? Are they human or material in that instance? This is not to engage in a discussion of the mind–body problem (Ryle, 1949), although the issue is certainly germane from the perspective of distributed or social cognition (Fiske and Taylor, 1984 ; Nardi, 1996 ; Resnick, Levine and Teasley, 1991 ; Zhang and Patel, 2006). The point to be made is that the distinction between human and material agency is indeterminate, and thus it is ontologically problematic.[30] Thus that distinction is not made in this work.

91. The ontological perspective taken in this work is to acknowledge a fundamental distinction between intention and agency. Intention is understood as a representation of effecting consequence in the world, and agency as the consequence of such representations effected. Affordance is understood as the point where intent and agency are bound together and made real. However, as discussed above, affordances obtain in pre-existing reality and materiality apart from the intentional actor, and thus they both enable and constrain. Consequently, here affordance is understood as <u>mediating</u> intention and agency—representations materialised in the world are never perfect, but rather are shaped through the affordances concerned. Furthermore, this relation of intention and agency is understood to be embedded in a reality built of three essential, distinct yet interdependent domains—the social, which comprises meaning and representations ; the material, which comprises consequence (that is, the effects of being mediated) and that which is subject to direct consequence, regardless of physicality ; and the technical, which comprises that which mediates the social and material, binding them into the real. These domains are defined in the following section.

2.2.2. Definition of Ontological Terms

2.2.2.1. The Social

92. The concept of the social is readily understood in a general sense, yet difficult to delimit. The appearance in common terms such as social science, social structure, and social context essentially references the shared experience of humans living amongst other humans. However, as Émile Durkheim noted, such scope is problematic from the perspective of investigation as no human experience would fall outside of the rubric of the social, thus providing no distinct field for social inquiry. He constrained the concept of the social to collective human understandings of proper behaviour, thinking, and affect, and the actions resulting from those understandings—these understandings and actions together producing the "facts" to which the social should refer (Durkheim, 1895, pp. 50–59). While this approach is intuitive, and furthermore allows for our experiences to be socially constructed (cf. Berger and Luckmann, 1966), it suffers from the conceptual "vagueness" that Herbert Blumer (1954, p. 8) warned hinders empirical research and theory development. Of concern here is the specification of the mechanisms and materiality of the social.

93. For this work, the understanding of the mechanisms of the social is premised on symbolic interactionism (Carter and Fuller, 2016), which posits that our actions towards things are contingent on the meanings that things hold for use, that these meanings arise from social interaction, and that we negotiate these meanings through an interpretive process through which we order and structure our worlds (Blumer, 1969, p. 2). To focus the structuring aspect (Hall, 2003), this work recognises "communicative action as the basic process in the social construction of reality” (Knoblauch, 2013, p. 297). Thus, while meaning gives force to social action, it is in the continual negotiation of meaning that social structures emerge and are sustained. This perspective recalls Niklas Luhmann’s systems theory (1995), importantly that constituent elements of social structures are not, in fact, individuals—rather, the constituent elements are the communicative interactions of actors, regardless of the nature of the actor (Stichweh, 2016, pp. 9–10), and that from the communicative processes of negotiating meaning emerge the structures that shape those processes, so suggesting autopoiesis (Cadenas and Arnold, 2015 ; Luhmann, 1986).[31] This work sees such actors as meaningful, in that their constitutive role hinges on their capacity to negotiate and sustain meaning. In accepting the primacy of communicative interaction in the structuration of the negotiation of meaning, the social is understood to comprise the intersubjective processes among meaningful actors through which meanings are negotiated, are sought to be made material, and are sustained. The social thus denotes the domain of meaning.

2.2.2.2. The Material

94. The term material, though widely encountered, is often ill defined and variously used (Kallinikos, Leonardi and Nardi, 2012). A common, intuitive understanding of materiality is physicality or corporeality. Extending this basic notion of matter, we consider its persistence—that which is material does not simply exist unto itself, but rather endures as itself through space and time (Faulkner and Runde, 2011). Given the understanding of the social used here, this persistence arises and is sustained relationally (Law and Mol, 1995), which is to say that the material only has reality in continued relation to that which is also material. These relations are spatial and temporal, as described, but also—arising from and transcending these—semiotic. Such meaningful relations inhere among the material because to be material is to be consequential (Pentland and Singh, 2012, p. 292): “Materiality is not about artifacts, people, ideas, or any *thing*. Or rather, it’s about all of them, but they only become *material* when they influence a particular course of actions or events that we value” (p. 294, emphasis in original). As the essence of the material is consequence and, given relationality, to be subject to consequence, its domain cannot be restricted to the physical or corporeal. Thus, the material is understood to comprise all that has consequence and is subject to consequence through spatial, temporal, and semiotic relations (i.e. extent, duration, and meaning). In that materiality is not a function of physicality but rather of relational influence, the material thus denotes the domain of mediation (not that which mediates, but the consequences thereof).

2.2.2.3. The Technical

95. The concept of the technical isolates function. A more common denotation would likely be technological, another seemingly intuitive concept that is difficult to delimit. As discussed in this chapter, technology does not exist unto itself but rather is a complex of society and artefact and practice. This complexity is evident in the description provided by Stephen J. Kline (1985) of the most common denotations of the term—the manufactured article ; the sociotechnical system of manufacture ; knowledge, technique, or methodology ; and the sociotechnical system of use. However, while there can be analytical utility in a separation between the technological and the social (Leonardi, 2012 ; Mutch, 2013), this work considers the two inextricable (Barad, 2003). The "constitutive entanglement" of sociomateriality (Orlikowski and Scott, 2008) points to an understanding that does not privilege physicality or embodiment, but rather function (Kallinikos, Leonardi and Nardi, 2012). The ontological perspective presented here is that the social and the material are indeed bound by function, but only because function and physicality are equivalent.[32] The social and the material are bound and brought into the empirical world by the mediating fact of function—the technical denotes the physical manifestation of the union of meaning and consequence. As the character of that union will necessarily shape the character of the function, the technical is the domain of the real, and thus of affordance.[33]

2.3. The Ontological Footing

96. This chapter has discussed historical perspectives and contemporary theories conceptualising the social–technological question. This question concerns the nature of change in society and technology, the mechanisms through which they are understood to relate and to influence each other, the localisation and direction of causality in change, and even whether the question itself is a valid framing of the empirical world. Current understandings of the question are characterised by a tension between the social and material—concepts that are used variously to indicate notions of nature (frequently physical reality is at issue), origin (such as whether an entity can be said to be constructed or manufactured), and agency (often devolving to claims or attributions of intentionality). To provide for productive analytical engagement with an experiential world of ubiquitous technology, disembodied social relations, and inseparability of the social and technical—that is, hybrid society—it is necessary to reconcile this tension. Such categorisation at root is a confluence of positivist and humanist traditions, and the potential of empirical anachronism weakens analytical claims arising from this dualism. To surpass, rather than bypass, the social–material tension in current understandings, this chapter argues for an affordance perspective that rests on three ontological categories—the social, the material, and the technical. The definitions put forth here, which depart from those common in the literature, are intended to isolate 1) the social as the domain of meaning and intention, 2) the material as the domain of consequence and that which is subject to consequence, and 3) the technical as the domain of affordance, of that which translates intention into consequence, thus bridging, mediating, and binding the social and the material and so making them real. By extension, only by means of the technical are the social and material observable.

97. This work is concerned with contextualising large-scale social phenomena as observed in public discourse. In hybrid society, such contextualisation necessarily hinges on modes of communication that are highly mediated and disembodied. Thus, there is utility to argue for a perspective that supersedes a focus on conventional notions of the social or material, of their physicality, and instead focuses on intention and consequence. But, as will be discussed in the following chapter, we cannot directly observe intention, and consequence is as slippery as causation. We can, however, look for them indirectly. We do so with a focus on affordance—that we can observe. This perspective is not concerned with the nature of actors but only with their relations—essentially with meaning (the social) and changes in meaning (the material). Affordance provides the locus of empirical observation. This is the ontological footing of this work, developed specifically to address large-scale social phenomena in hybrid society. While this discussion is rather baroque and perhaps lacking in articulation, it nevertheless provides a sufficient basis from which to develop the epistemological stance. That is developed in the following chapter, addressing the reconceptualisation of political participation in hybrid society.

Chapter 3

Reconceptualising Political Participation in Hybrid Society

> [W]hat [people] say about themselves and others represents an infinitely rich source of information about behavior. And the meanings that people give to politics are appropriate data for scientific analysis because people behave in terms of these meanings. … But, whatever definition of politics the political scientist adopts, it cannot be altogether arbitrary. It must itself be "meaningful" in terms of the meanings that [people] give to their political behavior.
>
> Heinz Eulau, *The Behavioral Persuasion in Politics*, 1963, p. 6

98. This chapter addresses the reconceptualisation of political participation for hybrid society, and from that derives the epistemological stance of this work. It begins with an overview of the scientific study of participation, elaborating how the empirical study of civic engagement has developed and expanded in response to theoretical development, growing bodies of evidence, and sociotechnological[34] change. The empirical study of political participation, while originating and initially developed in the US context, has grown far from those roots, providing a broad temporal and cultural perspective on civic action in democratic societies around the world. Likewise, from an initial institutional focus, the study of participation has also shifted in levels of analysis, from the macro-level perspective on large-scale collective action, to the micro-level perspective on individual action, and now increasingly towards meso-level perspectives on emergent action of collectivities. This overview is followed by a discussion of the conceptual dimensions of political participation, how these are challenged by evidence from the study of networked participation, and how we might reconceptualise political participation in hybrid society. Guided by the ontological footing developed in the previous chapter, this chapter sets out the epistemological stance that is derived from the reconceptualisation of political participation, and that will inform the development of the specific theoretical approach to its study in hybrid society.

3.1. The Scientific Study of Political Participation

99. The scientific study of political participation dates back roughly 100 years. During the 1920s, among students of the political, there was a growing recognition of the need for direct observation and quantification of phenomena. This drive to apply scientific rigour to a field that so often found its work to be archival can be seen as a manifestation of the scientific and technological fervour of that era. In the post-war context, such disciplinary development was seen as an urgent corrective:

> In these days of the improvement of means of communication and of efficient organization of means of collecting facts, we have fallen behind the possibilities of our times, and that by a very long interval. … While scientific expeditions are being equipped to cover all parts of the world and for all sorts of objects, the tremendous human experiment of democracy going on before our very eyes is not subjected to any process of scientific observation at all adequate to the needs of the occasion, and to the scientific possibilities in the case. … When something like exact measurement of recurring processes

> begins, we are on the way to exact knowledge, to scientific verifiable inference. It is natural to inquire to what extent the process has been applied to the study of political behavior. (Merriam, 1922, pp. 315–319)

100. It was no passing fancy ; this drive would culminate during the 1950s and 1960s as the behavioural approach (Dahl, 1961), later recognised as the "Behavioural Revolution" (Berkenpas, 2016). This shift in disciplinary conduct of the study of the political was characterised by an effort towards systematisation and formalisation, set on a foundation of empiricism (Dalton and Klingemann, 2011 ; Eldersveld *et al.*, 1952).[35] While the early decades of behaviouralism were predominantly a US development (Mandler, 2002), the approach eventually took hold broadly in Western political research (Berndtson, 1975 ; Frognier, 2002 ; Gamble, 1990, pp. 411–412 ; Shiviah, 1969 ; von Beyme, 2000, pp. 111–121). Behaviouralism shares a lineage with John B. Watson's 'behaviourist manifesto' (Kline, 1985 ; 1913), which insisted that the study of psychological processes could only be studied properly by way of observable phenomena, and with the later radical elaborations of B.F. Skinner (1938). However, despite conceptual overlap on the part of some scholars—for example, Floyd H. Allport's curious characterisation of political behaviour as that subset of social behaviour which delivers and responds to "political stimuli" (1927, p. 612)—it is important to distinguish between the 'behaviourism' of Watson and especially Skinner, which is fundamentally epistemological, and 'behaviouralism' in the study of the political, which is a methodological focus with a certain ontological core.

101. While dissatisfaction with a perceived lack of relevance and the inherent conservatism of behaviouralism quickly led to a "Post-Behavioural Revolution" (Easton, 1969), the behavioural approach has had lasting influence in the study of the political, and significantly so in the study of political 'behaviour'. Considering that life is a negotiated and social undertaking, political behaviour as a subject is ill defined and theoretically unbounded—it risks becoming the "study of everything" (cf. van Deth, 2001). The term and concept of 'political participation' is itself indicative of the move, necessitated by empiricism, to demarcate the subject of study. Early in the behavioural period, it was recognised that the antecedents of political behaviour were a valid subject of inquiry ; yet lacking reliable measures, focus was given to more tractable elements—votes, for example, as the "most tangible and measurable units of political behaviour" (Rice, 1925, p. 60). Understandings of what constitutes political participation reflect the tension between what is possible to study and what is needful to study. The following section briefly reviews these changing understandings of political participation and advances in its study.

3.1.1. The Behavioural Approach

102. The earliest studies of political participation demonstrating an empirical emphasis examined voting behaviour and party political activities. Among these, Charles Merriam and Herbert Gosnell (1924) investigated the phenomenon of non-voting. Working with a team of assistants, they interviewed 6,000 citizens who were eligible to vote in the 1923 Chicago municipal election, yet did not vote. The researchers sought to determine the causes of non-voting, and compared these data with explanations gathered from the political establishment, such as office holders, and party officials and workers. Fieldwork to obtain primary evidence, versus archival work to obtain secondary evidence, was seen as a progressive and welcome departure.[36] In a continuation of that study, Gosnell (1926, 1927) sought to test the causes of non-voting. A similar survey of non-voters in the 1924 US Presidential election had been conducted, which roughly confirmed the causes of non-voting identified previously by Merriam and Gosnell (1924). To test the causes, Gosnell selected a sample of 6,000 eligible voters, which was divided into two experimental groups. One group served as the control, whereas among the other group a non-partisan voter drive was conducted by post. The drive appeal was designed in such a manner as to address the identified causes of non-voting. If the treatment group registered to vote and voted at a higher rate than the control group, this would provide support for the validity of the identified causes. Gosnell indeed found that the voter drive had significant effect ; across the voting districts studied, the treatment groups registered and voted at a rate between 15–20% higher than the control groups (1926). During the 1930s, Herbert Tingsten of Stockholm University provided an analysis of electoral participation and political attitudes, according to various categoric groups, across a number of European countries (1937). His work was influential not only in the Nordic academy, but likewise in the European and US academies. The parameters and groupings that he applied in his work were taken forward by the 'Columbia school', which would be the locus of empirical political participation research in the coming decades (von Schoultz, 2015, p. 343).

103. The Columbia school broadly refers to the model of and research on electoral behaviour developed at Columbia University's Bureau of Applied Social Research. An early landmark in this work came with the publication by Paul Lazarsfeld, the first director of the Bureau, and co-authors Bernard Berelson and Hazel Gaudet of *The People's Choice* (1944). This study analysed data collected in one county in the US state of Ohio during the 1940 presidential election period. Six hundred subjects were interviewed monthly for a period of seven months leading up to the election. Of interest were the subjects' political opinions, reasons for those opinions and any changes they experienced, social networks, and media exposure. A similar study was published 10 years later by Berelson, Lazarsfeld, and William McPhee ; *Voting* (1954) also employed a panel approach, investigating similar themes, in the town of Elmira, New York, during the 1948 presidential election period. Both works were highly influential. Importantly to the development of conceptualisations of political participation, while these studies were designed along the lines of previous work that had a party political focus (although the results of the former study softened this focus in the latter), both found that the conduct of electoral campaigns had relatively low influence on voter choice—stances often seemed predetermined and stable. These results were not in line with the assumptions of rational choice theory, and pointed rather to the importance of social context. The authors noted this explicitly:

> the usual analogy between the voting 'decision' and the more or less carefully calculated decisions of consumers or businessmen or courts … may be quite incorrect. For many voters political preferences may better be considered analogous to cultural tastes … . Both have their origin in ethnic, sectional, class, and family traditions. Both exhibit stability and resistance to change for individuals but flexibility and adjustment over generations for the society as a whole. Both seem to be matters of sentiment and disposition rather 'reasoned preferences'. While both are responsive to changed conditions and unusual stimuli, they are relatively invulnerable to direct argumentation and vulnerable to indirect social consequences. … In short, it appears that a sense of fitness is a more striking feature of political preference than reason and calculation. (Berelson, Lazarsfeld and McPhee, 1954, pp. 310–311)

What might constitute this 'sense of fitness' would begin to emerge in the second half of the twentieth century as studies of political participation shifted from a focus on aggregate, institutional behaviour to an exploration of individualised, non-institutional action.

104. Scientific studies of political participation proliferated during the first half of the twentieth century. However, there was no unifying approach or theory beyond an insistence on empiricism. Efforts to survey and synthesise this broad body of work included Robert Lane's *Political Life* (1959), Angus Campbell and colleagues' *The American Voter* (1960), and Lester Milbrath's *Political Participation* (1965). Such studies are indicative of, and perhaps partly responsible for, a hardening of the idea of 'conventional' political participation, that being behaviours that centre on campaigning, party membership and structures, and contact between the political establishment and the public (van Deth, 2001, p. 5). This notion of what is 'conventional' emerged from the collation of earlier work that was guided by methods available at the time and the ease of collecting readily quantifiable data, and by notions of how political participation should manifest that often lacked empirical support. Thus, during a time of great social change and turmoil in the United States, the notion of 'conventional' political participation arguably took hold in part due to a perception of relative clarity and correctness—despite indications (such as those of Lazarsfeld and Berelson) that such notions did not well describe political participation as observed. In this sense, 'conventional' political participation is a normative construct, and it remains in tension with empirically based understandings of what constitutes participation and civic engagement, as will be discussed below. This tension underlies many of the conceptual and methodological debates around political participation of the past 50 years.

105. Nevertheless, there was growing awareness of the central importance of social context to understanding political identification, behaviour, and participation. Angus Campbell *et al.* (1960) put forward a model of party affiliation that stressed the fundamental role of socialisation. (It should be noted that such a model is highly specific to the US context of two dominant parties, and does not map to contexts with greater variety of potential party affiliations such as those with proportional representation.) Milbrath (1965) examined participation as a function not only of 'political stimuli' (cf. Allport's use of political stimuli above), but also of personal and environmental factors, and of social position. Gabriel Almond and Sidney Verba (1963) looked at the structures supporting democracy across several countries using an approach that was informed by theories of culture and personality. Lending yet more support to the importance of context, they found that political attitudes and patterns of participation are more a function of prior socialisation than of any contemporaneous stimuli. While the 'political culture' approach was in fashion for a time, difficulties in measurement and weakness of the underlying concept led to it falling out of favour. However, the focus on political attitudes—and focusing on micro-level phenomena to explain macro-level phenomena—continues to have a significant role in the study of participation (Conway, 2000, chp. 3). Such changing emphases, as well as increasing disquiet among political researchers that a focus on 'conventional' political science was an inherently conservative stance that reinforced problematic sociopolitical structures and the entrenchment of establishment members, led to the decline of behaviouralism by the late 1960s. The "Post-Behavioural Revolution" was not a rejection of the techniques and methods of rigorous science that behaviouralism stressed, but rather a redirection towards concern and engagement with social questions and issues (Easton, 1969, pp. 1051–1052).

3.1.2. The Post-Behavioural Approach

106. During the 1970s, political participation research shifted towards an examination of unconventional forms of participation—that is, actions that may be directed towards political concerns and outcomes, but that emerge and are conducted outside of institutional structures—as well as an emphasis on non-institutional antecedents of political participation. The seminal work in this regard is Verba and Nie's *Participation in America* (1972). While their basic approach stressed categoric socioeconomic indicators, and the research did not attend to unconventional participation, the investigation of factors of social context (such as ethnicity, age, community and organisational engagement) and partisanship was a significant advance in developing a more contextualised understanding of participation. A later iteration of the study (using the 1987 General Social Survey,[37] which replicated Verba and Nie's survey work from 1967), in conjunction with the first, provided a longitudinal, cross-sectional perspective on how demographic and structural changes had impacted participation (Nie *et al.*, 1988).

107. Samuel Barnes and Max Kaase, in their cross-national study *Political Action* (1979), looked specifically at unconventional participation. They put forward a model where social context and individual actors' values and ideology were independent variables, with participation and manner of participation as dependent variables. The question of manner of participation is significant, in that earlier work on participation was generally based on a unidimensional, hierarchical concept of participation, with the manner of participation being a function of the citizen’s degree of engagement (Brady, 1999, p. 741 ; Lane, 1959 ; Milbrath, 1965). Verba and Nie had already offered a critique of this approach, acknowledging the utility of a unidimensional indicator of engagement, but stressing that such could not describe the range of patterns of participation (Verba and Nie, 1972, pp. 43–44, 61–63). While Milbrath (1965) took a unidimensional approach, the second edition of that work (Milbrath and Goel, 1977) recognised that participation is complex and thus that simple measures are insufficient to capture its nature. Underlining the conceptual shift to multidimensionality, Barnes and Kaase introduced the concept of the “political action repertory”, now generally referred to as the ‘repertoire’, which they defined as “the sum of all political skills an individual has acquired through vicarious reinforcement and imitative learning” (1979, p. 39). The function of the repertoire is to provide a means through which the public can express its demands. While institutional practices certainly fall into the repertoire, as such practices are nominally in place to the end of public expression, the repertoire is in no way limited to institutional practices—where these are seen as ineffectual, the public develops and engages in new means to accomplish sociopolitical ends.

108. Related to the recognition of the multiplicity of forms of participation was a reconceptualisation of and increased granularity in the study of actors and their groups. The study of groups can be traced back to Gosnell, Merriam, and Tingsten—categoric groups (e.g. age, sex, ethnicity, location) are relatively tractable indicators that were increasingly utilised by researchers to order their data and analysis. However, while useful, such groupings lack theoretical richness (Conway, 1991, p. 37). In part this is due to the fact they are essentially exogenous—they are mapped onto subject populations, not derived from those populations themselves. A more endogenous, and thus contextualised, approach would be to study groups that emerge organically in populations. Verba and Nie (1972), for example, considered the factor of voluntary affiliations. Demographic subgroups also have been studied productively (Gurin, Miller and Gurin, 1980). Of interest is not simply how researchers can categorise and describe groups, but rather how people identify groups to which they belong and where they perceive those groups to be socially located, and—crucially to the study of participation—

whether people are conscious of their groups as political entities, that is, having "a political awareness or ideology regarding the group's relative position in society along with a commitment to collective action aimed at realizing the group's interests" (Miller *et al.*, 1981, p. 495). This move from exogenous categorisations to endogenous understandings has been critical to the maturation of the study of political participation. M. Margaret Conway noted that the great variety of theoretical approaches to the study of participation needed integration ; her assessment was that socialisation theory is the most appropriate method to integrate the study of participation, given that—regardless of conceptualisation or definition thereof—all of these behaviours are learned (1991, p. 41).

109. Despite the growing appreciation of the social complexity of participation, most empirical studies in the 1970s and 1980s relied upon the categoric 'socio-economic standard model',[38] or SES model, which focuses on participation as driven by actors' resources (e.g. income or education) and civic attitudes. The model was the foundation of numerous influential works, including Almond and Verba (1963), Milbrath (1965) and most importantly Verba and Nie (1972). There is substantial evidence to support the model, and it has been used productively to investigate a wide variety of forms of participation (see Leighley, 1995 for a thorough overview). However, it should be noted that the uptake of the model was partly due to the fact that indicators that could be coded but not meaningfully quantified (such as gender, race, and ethnicity) were more difficult to parse and thus resulted in a large body of conflicting research. The relative ease of the SES model encouraged its uptake, but resulted in a focus on who participates rather than on why (Leighley, 1995, pp.183 fn. 2, 184). It should be noted that socioeconomic indicators can only provide an exogenous, categoric understanding of 'who'—that is, the fungible individual—thus contributing to the theoretical (that is, explanatory) weakness noted by Conway (1991, p. 37, see above).

110. Beginning in the 1990s, there was a significant expansion in the domain of participation research, as researchers began to look more deeply into the social antecedents of participation, and at the varieties of social engagement and non-institutional civic participation (van Deth, 2001, pp. 7–13). While many observers worried about declining rates of voter turn-out, which had been observed across decades (Brody, 1978 ; Rosenstone and Hansen, 1993) and countries (Blais, 2000 ; Franklin, 2004 ;

Wattenberg, 2002), and the perceived decline of social and civic engagement (popularised by Putnam, 2000), it should be understood that such concerns were couched in understandings of 'conventional' participation. Other observers saw a move—also cross-national—towards different understandings and thus manifestations of civic engagement in contexts where the distinction between the political and social was increasing blurred (Cain, Dalton and Scarrow, 2003 ; Pattie, Seyd and Whiteley, 2004 ; Zukin *et al.*, 2006). Rather than a general decline in participation, what seemed to be the case was a weakening of institutional, establishment-led participation in favour of spontaneous, public-led engagement (Dalton, 2008 ; Norris, 2002). Furthermore, there were indications from a cross-national perspective that civic and social engagement were on the rise (Putnam, 2002 ; Stolle and Howard, 2008) ; thus the case of perceived decline in the United States (which had received so much attention) was either atypical and misleading, or misinterpreted. It became increasingly apparent that understandings and manifestations of citizenship were changing significantly as societies, and the world at large, moved into the Internet age, and researchers began to engage directly with emerging forms and norms of participation (Bimber, 1999, 2001 ; Blumler and Gurevitch, 2001 ; Castells, 1996 ; Dahlgren, 2005). Hybrid society was looming.

3.2. Participation in Hybrid Society

111. In the current day of ubiquitous communication technologies and mediatisation, reconceptualising political participation to account for significant, and continuous, sociotechnological change has become a subject of pressing concern and attention (Castells, 1996 ; Dahlgren and Alvares, 2014 ; Ekman and Amnå, 2012a ; Fox, 2014 ; Gibson and Cantijoch, 2013 ; Hooghe, Hosch-Dayican and van Deth, 2014 ; Ohme, 2018 ; Theocharis and van Deth, 2018). The expansion of the domain of participation research during the 1990s is indicative of the overall expansion during the twentieth century ; from an initial narrow focus on voting and closely related behaviours, sociopolitical shifts and technological advances (primarily in information and communications technologies) led researchers to consider a vast array of actions that could be understood as political participation (van Deth, 2001, pp. 3–6). As Russell Dalton and Hans-Dieter Klingemann have noted, the empirical and cross-national knowledge of political participation has increased by a huge amount in a generation, and that increase has come at a time of great change that, by its nature, may limit the applicability of past theories and models (2011, pp. 336–337)—given how information and communication technologies both challenge and enable the empirical study of participation, these two points are not unrelated. (This is the core challenge of hybrid society to the study of sociopolitical phenomena.) While the expanding domain has led to a proliferation, and even confusion, of theoretical and methodological approaches (Conway, 1991, p. 45), it is important to note that the expansion of the domain itself has been driven by sociotechnological changes (van Deth, 2001, p. 4), while changes in the academy derive from responsiveness to the results of empirical investigation. This is a crucial aspect of the scientific study of participation—to see participation as it is, rather than as it is assumed or desired to be.

112. The trail of empirical evidence has also led to research at various levels of analysis, as investigators traced the varieties of participation and their antecedents. Through the 1950s the emphasis was on macro-level phenomena such as voting and campaigning, that is, aggregate and institutionally shaped phenomena. The following decades saw a move towards the micro level, that is, individualised non-institutional phenomena. In the network age, there has been a steady move in the direction of the meso level, that is, emergent group-level phenomena. Such meso-level understandings (e.g. Ohme, 2018) have emerged from micro-level understandings in that they see collectives as affiliatory and endogenous, in contrast to earlier work (e.g. Verba and Nie, 1972) based on categoric groups (defined by age, income, gender, etc.) that are arbitrary and exogenous (Conway, 1991, p. 37).

113. Reconceptualising political participation in hybrid society depends on attending to the variety of actions emerging from meso-level civic phenomena, as well as to the variety and social nature of actors, rather than fitting things into preordained categories. Such attention entails important taxonomic work. As Henry E. Brady notes, such work is in no way mundane or pedestrian ; rather it undergirds our understandings (1999, pp. 739–740). Indeed, classification is an essential precursor for theory building, yet classification of social phenomena must be descriptive, not prescriptive.[39] To this end, Jan W. van Deth stresses that we do not need a comprehensive definition of participation, but rather a functional definition that allows for objective, unambiguous classification (2014, p. 353 citing Hempel, 1965). Without delimitation of the subject, empirical study cannot be systematic, and thus would be undermined (Theocharis, 2015, p. 4). The challenge for the study of participation, also noted by Brady, is that researchers must attend not only to scientific classification, but also to natural categories, that is, "how ordinary people name and classify things and with how they understand the world" (Brady, 1999, pp. 739–740). This concern speaks directly to the second aspect of reconceptualisation, that of varieties of actors ; we must note that Brady's "ordinary people" are themselves reflexive—they also name and classify themselves and are conscious of this (cf. Miller *et al.*, 1981, p. 495). M. Margaret Conway acknowledged this factor, urging a move away from "categoric groups" (i.e. macro- and aggregate micro-level analysis) towards the examination of "primary groups" (i.e. meso-level analysis), that is, those collectivities shaping the social contexts of participation (1991, p. 45). Engaging with the meso level of analysis is a critical step for the study of political participation in hybrid society. By integrating interpersonal and mass communication, while incorporating users as participatory means of production, hybrid society fuels and is fuelled by identity (Benkler, 2006 ; Dahlgren, 2009 ; Jenkins and Deuze, 2008 ; Papacharissi, 2010 ; van Dijk, 1999)—identity being, to adapt Brady, how people name and classify themselves and how together they understand their world. The following section reviews past and contemporary definitions of participation, and presents how participation will be conceptualised in this work.

3.2.1. Definitions of Participation

114. The range of conceptualisations of political participation is broad, and definitions are many. Among these definitions, the most commonly accepted formulations are those with a 'conventional' (i.e. institutional) orientation. Arguably the most influential definition has been that of Verba and Nie (1972, p. 2): "Political participation refers to those activities by private citizens that are more or less directly aimed at influencing the selection of governmental personnel and/or the actions they take". Many subsequent definitions have echoed this institutional, instrumental framing (Kaase and Marsh, 1979, p. 42 ; Nagel, 1987, p. 1 ; Parry, Moyser and Day, 1992, p. 16 ; Verba, Schlozman and Brady, 1995, p. 37). Brady, in his review of measures of participation, provides a minimalist definition, where political participation "requires *action* by *ordinary citizens* directed toward *influencing* some *political outcomes*" (1999, p. 737, emphasis in original). He found the four italicised elements to be common among most definitions of participation.

Furthermore, he considered these elements to be necessary to a functional definition. His assessment was that a definition of political participation must, at base, be concerned with 1) observable (thus measurable) activities, as opposed to thoughts, attitudes, interests, or intentions ; 2) ordinary citizens—that is, people[40]—as opposed to political professionals or those elected to office ; 3) influence, that is, intentional behaviours seeking to affect outcomes, as opposed to more passive behaviours such as information-seeking or topical discussion ; and 4) political outcomes, that is, behaviours must be targeted at "government policy or activity", as opposed to mundane behaviours, regardless of the political implications that many mundane behaviours might have (such as purchasing patterns, community engagement, etc.) (1999, pp. 737–738).

115. It should be noted that instrumental definitions such as these are characteristic of 'conventional' macro- and micro-level participation research. While such definitions provide tight summations of the subject of research, they do not describe the criteria by which researchers arrived at these definitions nor, by extension, do they specify what researchers exclude from the subject. Articulating these decision rules, and making them explicit, is an important aspect of systematic investigation. Stuart Fox examined definitions of participation according to the underlying assumptions (2014 ; cf. Conge, 1988). Recalling Verba's discussion of the topic, Fox stresses that there is no one correct definition of participation ; rather, definitions are adopted or crafted to suit the research at hand (Fox, 2014, p. 496 ; Verba, Nie and Kim, 1978, pp. 46–48). Specifically, he suggests that a given definition is "dependent upon the scholar's assessment of the purpose of political participation" (2014, p. 496), putting forward a series of criteria—derived from the literature—by which definitions, and thus the underlying conceptualisations, can be compared. This suggestion is significant for two reasons: firstly, the criteria presented encapsulate many of the debates during the last 20 years about the nature of political participation and how to study it,[41] especially in the context of the emergence of hybrid society (addressed by Fox, pp. 500–501) ; secondly, the statement and the criteria highlight the need for reflexivity, that is, understanding how participation is structured and understood by the participants themselves.

3.2.2. Reconceptualising political participation

116. Fox's criteria for comparing definitions of political participation are whether it is conceptualised as:

> 1) an active or passive behaviour ;
> 2) an individual or group activity ;
> 3) an instrumental or symbolic activity ;
> 4) a voluntary or mobilised activity ;
> 5) necessarily having deliberate aims or allowing for unintended consequences ;
> 6) a conventional or unconventional activity ;
> 7) necessarily having tangible influence or accepting as sufficient the intent to influence ;
> 8) having a governmental target or non-governmental target ; and
> 9) necessarily achieving an intended aim or allowing for failed attempts. (2014, pp. 497–498)

117. Fox certainly did not intend these criteria as binaries ; they are better understood as conceptual dimensions with which to frame a research approach to the study of participation. However, as Fox discusses at length, the ground has been shifting under our feet—these nine dimensions, which might usefully characterise older or backward-looking research approaches, cannot so usefully be applied for the conceptualisation of participation in hybrid society. While Fox notes that the effects of sociotechnological changes are unclear and contested, he acknowledges a growing consensus in support of Pippa Norris's assertion that civic engagement (i.e. participation in a broad sense) is evolving and transforming rather than declining (cited in ¶ **110**), and that studies using "conventional indicators" (and thus conventional conceptualisations) risk seriously misinterpreting evidence of participation (Fox, 2014, p. 502 ; Norris, 2002, p. 4).

118. Fox's criteria are not the only manner to conceive of the differences in conceptualisations of participation ; however, they provide a ready gauge of important differences between earlier and emerging conceptualisations of participation. As indicated, these criteria are not binaries—in any case such binaries would be false. Verba (as Fox notes) addressed the ninth criterion, that of considering successful or also failed attempts at participation, and states specifically that it should be understood as a continuum (1967, p. 59). Yet even understanding all of these as continua, growing evidence suggests that those understandings would likewise be false, or at the least logically problematic, especially in the context of networked participation in hybrid society.

119. The reconceptualisation of political participation that this work develops is pursued through a step-wise engagement with each of Fox's criteria, which provide reference points for comparing emerging understandings with older conceptualisations characteristic of the conventional study of the topic. That analysis—which grounds this reconceptualisation in both contemporary and earlier empirical work—is lengthy and is presented in **appendix A**. The conclusions of that analysis, taken together, offer a reconceptualisation of political participation in hybrid society. In summary, the reconceptualisation hinges on five major adaptations (thorough overviews of current understandings of participation can be found in Theocharis, 2015 and Gibson and Cantijoch, 2013). **First**, the distinction between active and passive forms of participation is outmoded, as all acts are potentially consequential (criterion one). **Second**, the distinction between individual and collective (i.e. micro- and macro-level) behaviour is difficult to make online, as the environment simultaneously enables and comingles these levels of behaviour. For this reason we are obliged to continue the move towards meso-level understandings of participation (criterion two). **Third**, mobilisation is not a subset of participation. Rather, they are coextensive concepts that describe different aspects of the same underlying behaviours. Furthermore, mobilisation is the manifestation of participation in the communicative—but otherwise disembodied—realm of online behaviour, and in such context the concepts are functionally equivalent (criterion four). **Fourth**, the distinction between conventional and unconventional forms of participation, which has been shifting and weakening throughout the history of the study of participation, has lost relevance. While it is logically sound to distinguish between institutional and non-institutional participation (which from the perspective of operationalisation can be used to distinguish roughly between offline and online participation, if need be), the notion of 'conventional' participation rests not on an empirical foundation, but a normative one, and is misleading to the conduct of research and the interpretation of evidence ; as such it should be abandoned (criterion six). **Fifth**, distinctions of participation that require an assessment and observation of intent or consequence are fundamentally problematic, and should be avoided for both conceptualising and measuring participation. However, intent and consequence can be productively bracketed by shifting to a focus on context of action (criteria three, five, seven, eight, and nine).

3.3. The Epistemological Stance

120. The reconceptualisation serves to inform the epistemological stance of the work, in that it suggests how we might understand the constituent elements of this phenomenon and their potential relationships, and further where and how we might observe it. While the reconceptualisation is not the stance itself, the stance is derived from it. The ontological footing guides the derivation of the empirical stance—each adaptation of the reconceptualisation is considered in light of the ontological components. The adaptations and their ontological interpretations are as follows:

1) there is no purpose to distinguish between active and passive forms of participation, as all acts are potentially consequential (i.e., material) ;
2) it is difficult to distinguish between micro-level (individual) and macro-level (collective) behaviour, so we are obliged to seek understandings that are meso-level (i.e., social) ;
3) in network terms, participation and mobilisation can be understood as equivalent terms that hinge on behaviour that is communicative (i.e., technical) ;
4) the distinction between conventional and unconventional participation is no longer relevant (and in any case is a normative, disciplinary projection without sound empirical justification), so the social–technical–material configurations[42] of 'networked participation' (e.g., Theocharis, Moor and van Deth, 2021) are valid objects of inquiry ; and
5) the intent of participants, or the consequence of their behaviours, cannot be directly observed (as above, empirical observation occurs only in the technical domain), however both can be productively bracketed by a functional shift of focus to the context of action (i.e., a specific social–technical–material configuration, of which the technical can be observed).

121. It can be seen how the domains declared in the ontology map onto the reconceptualisation. While this may seem slight, we nonetheless can glean important information with which to take our empirical stance. It is not intended to be a laundry list ; rather we just need to know what we are about as we elaborate our theoretical approach in the next chapter. Working top down, we have the material and social domains declared (items 1 and 2). They are as yet unspecified—such specification is the function of the theoretical approach. Following, the technical domain is declared and specified to implicate communicative action (item 3) ; thus we see that the social and material are bound by communication of some sort, which gives us implicit information about those otherwise unspecified domains. As discussed in the previous section, some contest the validity of studying political participation from a network (and thus hybrid) perspective, insisting instead on a conventional (i.e., offline) perspective. In either case, the chosen perspective represents a decision that constrains the eventual specification of the social and material domains—a specification made to take a conventional perspective would run counter to the aims of this work. Luckily, there is no need to differentiate (item 4), so we are free to specify these domains in the theoretical approach in a manner that supports the work. Finally, to compensate for our inability to observe intentions and consequences, we can attend to a specific context of action. In other words, we can specify the social and material domains in whatever manner we see fit, so long as they are understood to be functionally (i.e. contextually) political (item 5). This is the epistemological stance.

122. The reconceptualisation of political participation in hybrid society presented here has served to inform the epistemological stance of this work. While it may not seem like much at first glance, the stance should be thought of as a prototypical theorisation of the phenomenon we seek to study. It is prototypical in that the core conceptual pieces are in place, but as of yet are ill-defined. That is because it is only a template for the theoretical approach. By fully specifying its components, as will be done in the next chapter, the epistemological stance is shaped into a theoretical approach suited to a specific inquiry and site of research. This will be done in the following chapter, which further will suggest the operational step with which to implement the theoretical approach for empirical investigation.

Chapter 4

Operationalising Language in Mediated Public Discourse

> An emergent form of political economy, facilitated by information and communication technologies (ICTs), is widely propagated as the apotheosis of unmitigated social, economic, and technological progress. Meanwhile, throughout the world, social degradation and economic inequality are increasing logarithmically. Valued categories of thought are, axiomatically, the basic commodities of the knowledge economy. Language is its means of exchange.
>
> Philip Graham, “Critical Systems Theory”, 1999, p. 482

NOTE: This chapter is lengthy. Primarily this is due to the need to draw together theoretical strands from across a range of fields and disciplines. As noted elsewhere, interdisciplinary work cannot call upon a common body of knowledge in the manner of disciplinary work and so must be relatively explicit in its assumptions and arguments. The chapter is also heavily annotated to provide further context and rationale for the discussion. In the main the annotations serve the interdisciplinary and pedagogical aims of this work beyond its immediate purpose, and can be skipped if the reader is so inclined.

123. This chapter integrates the conceptual framework of this methodology, in preparation for addressing the operationalisation of language for the study of large-scale social phenomena. Building on the ontological footing given in **chp. 2 § 2.3**, and extending the epistemological stance presented in **chp. 3 § 3.3**, this chapter specifies the theoretical approach that this work takes in seeking warrant for the operationalisation of the reconceptualisation of political participation in hybrid society. As the work is concerned with large-scale phenomena, the approach is directed specifically towards mediated public discourse. Having detailed the integrated conceptual framework, the operational step is presented—that is, the argument for the interrelation of the conceptual and the phenomenal.[43] That argument enables an appropriate method to be assembled, to be detailed in the following chapter.[44]

4.1. The Conceptual Framework

124. Language is a functional social phenomenon (Östman and Simon-Vandenbergen, 2009). As such, it readily serves as both object and method of social inquiry. Nevertheless, to serve as object it must be conceptualised, and to serve as method it must be operationalised. In this we face a curse of plenty: language is so much the fabric of our lives—so much the water in which we swim, to riff on Marshall McLuhan's often misunderstood observation[45]—that we risk becoming overwhelmed with possibilities in this regard, with no clear path through the muddle. Yet we must pick a path, at each and every step, and thus we depend on a fully specified framework (Kaplan, 1964, §§ 34–35). For a given subject of social inquiry, a conceptual framework is comprised of:

- **ontology**, which declares the objects implicated ;
- **epistemology**, which specifies how those objects are conceptualised ;

- **theory**, which suggests how those objects are understood to relate ; and further
- **operationalisation**, which suggests the phenomena in which we might perceive indications of those expected relations.

In disciplinary work, frameworks in general are well established and many of these elements are assumed or have an accepted set of options—such delimitation and facilitation is the functional purpose of disciplines (cf. Kuhn, 1970, pp. 181–187). In interdisciplinary work this is not the case, and so a fully specified framework is obligatory.

125. An overarching subject of this work is the role of language in shaping and effecting social relations and collectivities. More specifically, the subject is how observation of language in public discourse can inform social inquiry by providing indications of shared meanings, and thus indications of possible contextual similarities, among the various discussants (see **chp. 1 § 1.5**). From that subject of inquiry, the following conceptual framework has been elaborated:

- the **ontological footing** (**chp. 2 § 2.3**) declares the fundamental objects of concern to this work (the social domain of meaning, the material domain of mediation, and the technical domain of affordance) and describes how they are understood in themselves ;
- the **epistemological stance** (**chp. 3 § 3.3**) is derived from a reconceptualisation of political participation in hybrid society, and outlines the expected character of phenomena fitting that rubric ;
- the **theoretical approach** (**§ 4.1.1**, below) draws on the preceding components to suggest a conceptual model of the subject of inquiry—in which the semiotic affordances of language serve to bind the social and the material in a manner that is modulated by context—that is fitted to a site of research ; and
- the **operational step** (**§ 4.1.2**, below) suggests specific phenomena of language—specifically pervasive but differential patterns of lexis and grammar—that might yield empirical evidence from which to evaluate our research questions (see **chp. 1 § 1.5**).

Each component of the conceptual framework supports the components that follow, and the framework as a whole undergirds the method (presented in the following chapter).[46]

Before describing the theoretical approach, which serves to integrate the conceptual framework, we will recap the supporting elements.

126. The **ontological footing** of this work, presented in **chp. 2 § 2.3**, comprises three fundamental domains: the social, the material, and the technical.

- The **social** is the domain of meaning. Meaning is understood as the representation of consequence, and intention is a subset of meaning, understood as the representation of effecting consequence. This domain is immaterial (and thus incorporeal, following the understanding of material given below), and processual (i.e., it has no fixed state ; it is always becoming).
- The **material** is the domain of mediation, where materiality is not a function of physicality, corporeality, or any persistence in duration or extent, but rather is a function of relational influence—be that relation spatial, temporal, or semiotic. Put more simply, the material is consequential. If any thing influences or is influenced by any other thing, that is a consequential (i.e., material) relation and thus the things in question are material. This domain is asocial (thus devoid of any inherent meaning) and relational.
- The **technical** is the domain of affordance, the interface between the social and the material, where technicality has no relation to physicality or embodiment, but only to function. Here affordance is a synthesis of pre-existing meaning and consequence into new meaning and consequence—they are bound together and made real in the technical. The technical exists, as it were, between the social and the material, which themselves cannot directly interact. This domain is contingent (on the coexistence of the social and the material) and functional, and is the only domain that is real and thus observable.

To sum: the social is the domain of meaning, the material the domain of mediation, with the technical domain of affordance betwixt and between.

127. This is a practical ontology, not one elaborated for focused philosophical inquiry. It is a heuristic constructed specifically for this work, to help with thinking about sociomateriality, technology, mediation, communication, semiosis, etc., across fields, disciplines, strands of work, and periods. Key to the use of the heuristic is to understand that the social and the material are thoroughly intermixed in human experience. Those domains are nonetheless distinct, the former being noumenal, and the latter phenomenal—like east and west, they shall never meet. Yet, also like east and west they are useful relative descriptors with which to make sense of experience. The synthesis of meaning and mediation in the domain of the technical can be thought of as experience itself. By that token, empirical investigation is instantiated only within the technical domain.

128. The **epistemological stance** of this work, presented in **chp. 3 § 3.3**, is thankfully more straightforward than the heuristic of the ontological footing. As with the footing, it is not an articulation of a general epistemological stance, but rather a specific articulation for the purposes of this work. That specific stance was derived from an assessment of changing conceptions of political participation in relation to hybrid society, which was then interpreted in terms of the ontological footing.

129. That assessment drew five conclusions, each of which provides a component of the epistemological stance:

1) there is no purpose to distinguish between active and passive forms of participation, as all acts are potentially consequential (i.e., material) ;
2) it is difficult to distinguish between micro-level (individual) and macro-level (collective) behaviour, so we are obliged to seek understandings that are meso-level (i.e., social) ;
3) in network terms, participation and mobilisation can be understood as equivalent terms that hinge on behaviour that is communicative (i.e., technical) ;
4) the distinction between conventional and unconventional participation is no longer relevant (and in any case is a normative, disciplinary projection without sound empirical justification), so the social–technical–material configurations of 'networked participation' (e.g., Theocharis, Moor and van Deth, 2021) are valid objects of inquiry ; and
5) the intent of participants, or the consequence of their behaviours, cannot be directly observed (as above, empirical observation occurs only in the technical domain), however both can be productively bracketed by a functional shift of focus to the context of action (i.e., a specific social–technical–material configuration, of which the technical can be observed).

130. To reiterate the explanation provided at the end of the last chapter, we can see how the elements declared in the ontology map onto the assessment. We have the material and social domains declared, though as of yet unspecified, in items and 2. Following in item 3, the technical domain is declared and specified to implicate communicative action ; thus we see that the social and material are bound by communication of some sort, which gives us implicit information about those otherwise unspecified domains. As observed in **chp. 3 § 3.2** and **appendix A**, some contest the validity of studying political participation from a network (and thus hybrid) perspective, insisting instead on a conventional (i.e., offline) perspective. However, item 4 indicates that there is no need to differentiate, so in the theoretical approach we may specify these domains in a manner supporting the work. Finally, as we cannot observe intentions or consequences, we attend to a specific context of action. Thus we can specify the social and material domains as best suited to the theoretical approach, so long as they are understood to be functionally (i.e. contextually) political.

131. As it stands, the epistemic stance is relatively unspecified, but it has been structured (the social and material mediated by communicative action as the technical), and its specification has been partly bounded (by the allowance of a hybrid perspective in item 4, and by the bracketing of intent and consequence with a focus on political context of action in item 5). As noted in **chp. 3 § 3.3**, the stance provides a template for further theoretical elaboration suited to a specific study and site of research. We now do just that—with the stance in place, we proceed to fully specify its components and thereby to develop the theoretical approach.

4.1.1. The Theoretical Approach

132. As discussed in the previous section, the theoretical approach of this work is arrived at by specifying the components of the epistemological stance. Thus, whereas the stance comprises a set of somewhat bounded concepts, the approach develops more specific conceptualisations—that is, specific formulations of a more general concept—needed for the work at hand. Moreover, as the theoretical approach provides the basis for the operational step (presented in the following section), and thus the selection of a specific method, the approach also must specify how these conceptualisations are understood to relate. The needed conceptualisations, following from **¶ 130**, are:

- a bounding of the social–technical–material configuration in question (i.e., the prototypical site of research), to be given in **§ 4.1.1.1** ;
- a specification of the social, material, and technical domains in question, to be given in **§ 4.1.1.2** ; and
- a proposition of the functions of interest in the technical domain, and the expected effects in the social and material domains, to be given in **§ 1.2.1**. (NB: as the social and the material cannot be observed directly, the operational step puts forward the suggestion of how indications of effect in those domains might be observed in the technical domain.)

These items will be addressed in order. The following should be understood as conceptualisations suited to the needs of the work at hand, and not as broader claims of any sort. In developing these conceptualisations, recall that this work's motivating hypothesis and research questions are borne in mind (**chp. 1 § 5**). Also recall that the theoretical approach is not a fully specified and articulated theory in itself, but rather a 'sketch' suggesting the nature and interactions of our working conceptualisations.

4.1.1.1. Bounding the Site of Research

133. Given the interweaving nature of the social and the material (see **chp. 2 § 2.1.3**), it is logical to begin specification of the theoretical approach with a bounding of the social-technical-material configuration in question (cf. Boczkowski and Lievrouw, 2007, p. 957 ; and Schultze, 2014, p. 87) before looking at the individual elements. Foremost, this work is concerned with how to study large-scale social phenomena from an endogenous perspective (i.e., as they are perceived by those involved) as opposed to an exogenous perspective (i.e., as they are conceived by those observing). The range of social scientific methods developed to serve the former perspective are in general highly interpretive and time- and labour-intensive, and thus are appropriate only to smaller-scale inquiry. Methods developed to serve the latter perspective tend towards the highly reductive and positive, substituting statistical and computational techniques in order to reduce dependence on time and labour, thus facilitating large-scale application. The latter approach, while demonstrably effective for demographic understanding (for example, in the taking and tallying of national censuses), comes at significant cost to social understanding. Such a methodological split is evidenced in the history of the study of political participation, as discussed in **chp. 3 § 3.1**. However, developments in information and communication technologies, the concomitant expansion of access to such technologies at decreasing cost, and the resultant profusion of new media, artefacts, and practices (**chp. 2 § 2.1.4** ; cf. **chp. 3 § 3.2**) are enabling methodological possibilities that seek a middle path. This work explores such a middle path by focusing on the discursive aspect of large-scale social phenomena, specifically how discourse is instantiated through text shared via social media platforms. We have shown above in the epistemological stance that such a general configuration is a valid object of inquiry for the study of networked political participation (**¶ 129 no. 4**).

134. Furthermore, in light of 'deep' mediatisation—wherein media is not simply pervasive, but pervasive to a degree that decentres the primacy of face-to-face interaction (Couldry and Hepp, 2017, chp. 2)—it is reasonable to argue a step further: that discourse in highly mediated configurations is nevertheless broadly representative of the communities and societies that sustain that discourse. To borrow the affordance and parlance of Twitter, discourse *#onhere* is not so different that discourse *#outthere*. Why might that be the case?—because contemporary mediatisation weaves together the offline and online, each experienced, understood, and enacted in relation to the other (Couldry and Hepp, 2017, p. 33). While highly mediated discourse might seem or look substantially different from 'conventional' discourse,[47] the difference is superficial, as each type of discourse nevertheless implicates the same people, from the same communities, and from the same societies—that is, it implicates the same contexts. Some would immediately object, as will be discussed in a moment, given that we know online discourse does not in fact reflect the broader population. But reflect in what way? That question returns us to the demographic versus social perspective mentioned in the previous paragraph—the former perspective sees aggregations of isolated individuals, whereas the latter sees members of communities and of society.[48] Understood in this manner, deep mediatisation could be expected to reduce presumed discursive differentials that 'digital divides' or any form of technological inequality (cf. van Dijk, 2020) might introduce. The point being made is not that such inequalities are irrelevant or do not exist—they certainly do—but rather the point is that, in contemporary contexts of deep mediatisation, the concepts of 'digital divide' and 'social structure' approach a functional equivalence (cf. Treem *et al.*, 2016, p. 778).[49] Such a claim is difficult to warrant, not because social inquiry into mediated society and its evidence are of questionable validity or worth, but rather because notions of society and its character as enacted primarily through 'conventional', face-to-face discourse are anachronistic and essentially imaginary (Blommaert, 2017 ; Blommaert, Smits and Yacoubi, 2018). In that we have no empirical example of <u>un</u>mediatised public discourse to provide 'ground truth',[50] we should set aside any wariness towards the study of highly mediated (thus mediatised) discourse. For large-scale study, it is effectively the only game in town.

135. In this light, developing a large-scale view of society by studying its language and discourse in highly mediated contexts (e.g., social media platforms) is not only valid, but necessary. In that rich ethnographic accounts of society are not feasible at scale, the middle path of seeking to contextualise mediated discourse offers perhaps our best approximation of a large-scale social 'reality'—that is, an endogenous view of who we are, who we believe we are, and who we imagine others to be. To be absolutely clear: viewing society through its social media discourse should in no way be confused with efforts to reveal, represent, or evaluate public opinion (Gayo-Avello, 2013 ; cf. Jungherr *et al.*, 2017). Bear in mind, public opinion is a central construct in how modern polities are conceived, measured, studied, and reported (Manza and Brooks, 2012 ; McCombs, 2004 ; McGregor, 2019, 2020). And while of central importance in many strands of work and discourse, it is nevertheless a fiction, the nature and tenor of which is shaped to suit the motivations of the observer (Herbst, 1998). In an operational sense, these efforts are functionally similar to electoral polling and survey approaches (e.g. Amador Diaz Lopez *et al.*, 2017), thus suffering from problems similar to those observed in the study of political participation (cf. **chp. 3**), as well as being subject to long-held epistemological criticism that polling and survey approaches obfuscate the social component of public processes (McGregor, 2019, pp. 1071–1073). It is perhaps an encouraging sign to have empirical evidence that social media does not reflect a reductive, motivated notion of an overall societal 'opinion' (e.g., Barberá and Rivero, 2015 ; Mellon and Prosser, 2017 ; Mislove *et al.*, 2011)—that is exactly the type of approach to social inquiry that this work argues is outmoded and misleading, so perhaps we are on a good path.

136. The progressive mediatisation of contemporary societies has had significant impact on language in society, and thus we are obliged to reconsider how we view society through language (be that view ordinary or scientific). We must adapt our philosophy, theory, and method to a world of hybridity (cf. Chadwick, 2013), where the offline and online, physical and virtual, are constantly and unavoidably alloyed (Blommaert, 2019, p. 486).[51] Given the above, for a social science that means to be relevant and forward-looking, highly mediated configurations are not simply 'valid' objects of study, they are essential and fundamental objects of study. This is certainly the case for this work—as discussed in **chp. 2 § 2.1** regarding the evolution of thinking on the social–technological question, and in **chp. 3 § 3.2** on the study of political participation in hybrid society. The theoretical approach of this work is thus motivated and guided by the need to adapt to communicative hybridity. For that reason, the prototypical site of research of this work will be bounded to a general configuration that is arguably emblematic of mediated discourse—public microblogging platforms (Honeycutt and Herring, 2009 ; van Dijck, 2011).[52]

4.1.1.2. Specifying the Site of Research

137. Having bounded the site of research to the general rubric of public microblogging platforms, we now turn to specification of the social, material, and technical domains. Roughly speaking, this is equivalent to scoping potential sites of research. This scoping will be approached from an ethnographic perspective. The rationale follows.

138. The scoping stage of work may seem rather 'loose' to some observers, with certain decisions, including the ultimate selection of a site of research, seeming almost arbitrary. However, we should not overlook the role that disciplinary norms and trends play in guiding such preparatory work in dialogue with the researcher's domain and tacit knowledge, and understandings and assumptions of their research goals.[53] The work at hand is intentionally interdisciplinary and as such runs somewhat counter to the bulk of disciplinary norms and trends. To achieve its purposes, interdisciplinary work is necessarily less beholden to such boundaries and pressures. However, the overall project of knowledge remains an undertaking of cumulation and integration. The power of interdisciplinary work is not in blazing new paths into uncertain terrain, but rather in exploring interconnections between extant paths so as to see familiar terrain in new ways. As discussed at the beginning of this chapter, such work cannot rely on assumed practices and understandings. Rather, it must be an 'explicit' social science that articulates its decisions and steps, thus blazing the interconnections among differing paths to social knowledge.

139. An unavoidable side-effect of studying society at higher levels of analysis is that it yields lower levels of social information—as with physical lenses, there is an inverse relationship between field of view and magnification. However, what is avoidable is treating individuals and their communities as fungible. While abstraction of the attributes of individuals and communities is a necessary step in viewing society in broader scope (cf. **chp. 1 § 1.2**), to maintain a social perspective we must resist viewing society as "nothing but a numerical aggregate, a conglomeration of units" (Dewey, 1888, p. 4).[54] That is, as this work argues, we must set aside overly reductive perspectives and the reliance on exogenous measures—which remain dominant within and without the academy—and in their place seek more human understandings by way of relational perspectives that may facilitate the development of endogenous measures. In that spirit, the specification of the site of research will proceed in a manner informed by principles of ethnographic method (e.g., Gold, 1997).[55]

140. These ethnographic principles will be narrowly construed to suit the work at hand, as they should be. Contemporary ethnography manifests in wild variety. The more common aspects are broadly recognised and accepted as a certain 'qualitative' mode of building knowledge, for example extended social interaction and observation and narrative approaches to deriving meaning from social experience. In more extreme aspects that have appeared in the past decades, ethnographic thinking can seem incomprehensible and off-putting to those in less interpretivist fields, despite recognition of an underlying utility (cf. Bayard de Volo and Schatz, 2004).[56] This work hopes to proceed in the former manner by stressing a focus on three central concepts—interaction, the mundane, and place. The first two will be addressed in the following section. For now, we turn our attention to the concept of place.

141. Place is central to the ethnographic tradition. In part this has to do with ethnography emerging from the pre-anthropological[57] tradition of ethnology (cf. Voget, 1975), which itself can be understood as an evolution of the travel narrative (Thornton, 1985). Place is indelibly associated with observers and observations of this mode, such as Bronisław Malinowski in the Trobriand islands (1922) and Margaret Mead in the Samoan islands (1928). In broad strokes, the ethnographic emphasis on place came from a need for credence. The "exotic exemplar" of time lived in subjectively far-away places amongst subjectively little-known peoples gave such accounts facticity and authority (Clifford, 1997, p. 192 ; cf. Clifford, 1983). Place became a crucial identifier of 'real' ethnographic work, where fieldwork provided exposure to societies in their 'pure' and 'natural' state (Gupta and Ferguson, 1997b, pp. 12–15). Needless to say, the roots and implications of such thinking are problematic. However, justifications of 'the field' providing a sharper picture of things are put on firmer ground, so to speak, if we consider language-oriented work, such as that of Franz Boas and those who developed that tradition (such as his student Edward Sapir, and Sapir's student Benjamin Lee Whorf, to name only a few).[58] Boas was convinced that language was key to understanding cultures and societies (e.g., 1906) and by that was driven to document the disappearing languages of North America, in hope to preserve the social understandings embedded in them (Darnell, 1990 ; e.g., Boas, 1900). This notion of 'the field' as the place of fieldwork where one encounters social phenomena not present elsewhere, or perhaps to disappear, seems sadly well warranted.

142. Jumping ahead to the second half of the twentieth century (thus eliding an awkward and unfortunate period for many disciplines, a sliver of which was discussed in **chp. 3 § 3.1**), the concept of place has become yet more complicated. An increasingly critical perspective on ethnography was looking inward, grappling no longer with the field or fieldwork as such, but rather with the underlying need for facticity of account. The critical motive understood that 'place' did not and could not exist unto itself, but rather was situated within some larger whole (i.e., the State, capitalism, the world system, and the other usual suspects) ; the role of ethnographer then was to interrogate a place so as to critique the whole, however arbitrarily that might be conceptualised. The urge to holism simultaneously marks a deeper understanding that social life does not proceed in a clearly-bounded locale, but spreads out through time and space. We thus arrive, from say the 1980s, to the notion of 'multi-sited' ethnography (Marcus, 1989, 1995). To be fair, this is no new phenomenon, nor even a shift in practice—George Marcus (1995, p. 106) points to Malinowski's work in the Trobriand islands (which is, to be clear, a chain of islands) as the "archetypal account" of such multi-sited ethnography. Listen: place has come unstuck. In truth it never was stuck, only our collective fixation made it seem in anyway fixed.[59]

143. For those who conduct work using ethnomethodological approaches, changing understandings of place are challenging. The ethnographic concept of place—while it has been deeply historicised and criticised (e.g. Gupta and Ferguson, 1997a)—has been relatively undertheorised until recent decades. In the main, at least for those working in the more traditional mode of direct interaction with a community in a physical location,[60] such theorisation could be set aside, and thus place has been approached primarily in pragmatic, practical terms (Gupta and Ferguson, 1997b). More immediate than theoretical concerns in such cases are questions related to feasibility and quality of the study—Can one access the location in question? Why choose one location over others? How will the period and schedule of access impact social sampling? In that traditional mode, theoretical complications to the concept of place present a significant hinderance to achieving desired research goals. However, for work that seeks to apply ethnographic principles to large-scale social inquiry—which is decidedly non-traditional—avoiding the theorisation of place is a recipe for muddle and wasted effort. And so we are obliged to work it through.

144. Across the social sciences proper,[61] we grapple with the growing realisation that place not only is unstuck, but also extends beyond the corporeal and even the physical. In the fields of mass communication and media studies, among others, this notion has been embraced for a generation at least (Appadurai, 1996 ; Castells, 1996 ; Chadwick, 2013). However, it is essentially a lateral move to incorporate the notion into structural or systemic analysis. Such analysis is inherently multifarious, and it is 'fixed' by its conceptual boundings more so than any given location or area. But it is an altogether different undertaking—and one that often gives us pause—to reconceptualise our lives, both those we observe as well as those we ourselves live, as not defined by (and so privileged by) a singular point of awareness or of being (cf. Derrida and Ferraris, 2001, pp. 40–41).[62,] Such 'unstuck' thinking, that is, thinking in terms of communicative hybridity and its methodological implications (as noted at the beginning of this section), has thus taken hold more slowly in sociological work that directly engages with meaning in society and the constitution of society thereby.[63] For a variety of reasons, that slowness (and even resistance) in certain disciplines is understandable. It is nevertheless unfortunate (and even maladaptive), in face of the overwhelming hybridity evinced by even the most would-be pastoral of everyday lives. And for empirical disciplines whose stock in trade is communication, there is no warranted option but to embrace hybridity across all actors, modalities, and levels of analysis.[64]

145. Beyond the philosophical, existential, and disciplinary challenges posed by 'lived hybridity' (cf. **n. 63**), there is a slew of practical challenges posed to social inquiry, especially at scale (i.e., performed in or through highly mediated environments). Jan Blommaert and Dong Jie (2019) highlight three key concerns:

- <u>What</u> we observe is constrained and shaped in ways that not only are out of the hands of researchers, but moreover are configured algorithmically by software, artefacts, and data derived from habits of interaction. While we know this to be the case, we are not permitted to know the nature of that bounding and shaping—as technology and data are generally proprietary (cf. boyd and Crawford, 2012, p. 674)—and neither can we escape from it (Blommaert and Dong, 2019, pp. 2–3).
- <u>Who</u> we observe is likewise out of our hands. Media platforms and the interests underpinning them control access as they see fit, also in a manner mostly hidden from the researcher. Moreover, in such mediated spaces we do not engage with known individuals, but only with aliases (or other proxies, such as an email address), and there are innumerable reasons benign and malign that identities might be further obfuscated. Whatever the case may be, in these environments identity is fundamentally indeterminate (pp. 3–6).
- <u>Where</u> we are observing is likewise similarly unknown, as it is untethered from ordinary physical place and experience. This concern is perhaps the crux of hybridity. From the former perspective of ethnography, the 'where' of social inquiry—be it the place, the field, or the site of research—would naturally have bounded and helped to reveal the 'what' and 'who' (pp. 7–10). But without that anchor of a space in time, how are we as researchers to conceive of the social phenomena that we seek to study and thus go looking for them?

146. George Marcus (1995), in his discussion and unteasing of multi-sited ethnography, offers a way forward. He suggests 'modes of construction', wherein certain loci of meaning are seized upon by the ethnographer and traced in order to reveal the terrain of study and its whos and whats (although the loci of meaning in question might themselves well be whos and whats). His suggestions for tracing meaning include "Follow the People" (cf. Malinowski), "Follow the Thing", "Follow the Metaphor", "Follow the Allegory", "Follow the Conflict", and so on. These suggestions could be admonitions to a first-year creative writing class, and rightly so. We talk ourselves into being. That is the essence of human experience, as strange as it may seem. Learning to tell meaningful stories of our world is little different from learning to trace meaning in our world—each one face of an old coin (recall the mention of narrative construal of meaning in ¶ **140**). This is not a dismissive observation. Marcus has reason. Compare his suggestions to that of Yannis Theocharis in regard to the study of political participation as noted in **chp. 3 § 3.2**. To recap, in the study of political participation, two frequent components of scholarly conceptualisations of that complex of phenomena are the intention of participants and the consequence of their actions. For many, these are critical considerations. But, as noted in **chp. 3 § 3.2.2**, neither intention nor consequence are directly observable—thus they cannot be measured.[65] For that reason alone they are excluded from the reconceptualisation of political participation in hybrid society in that they cannot be operationalised. Yet they remain a critical consideration nonetheless. The solution proposed by Theocharis (2015) and developed by Theocharis and van Deth (2018) productively brackets these considerations, subsuming them into a focus on context of action[66]—that is, they suggest that we 'Follow the Action', to riff on the phrasing of Marcus. This is what we will do here. Having worked through retheorisations of place in terms of hybridity, we will reformulate those concerns in terms of context. Having firm footing from which to work, we can now specify the site of research as a pre-operational step. In light of the above discussion, there are two needs sufficient to get the ball rolling:

- a superficial situating of the site of research, supported by a pragmatic, practical rationale ; and
- an argument for how that superficially situated site will be understood as a political context of action.

Specification of the site of research is termed 'pre'-operational as it is a declarative, rather than metric, move yet nevertheless provides the situating context necessary to take the operational step. The operational step proper, presented in **§ 4.1.2**, then makes the metric move by presenting what is to be measured and how.

4.1.1.2.1. Superficial Situation

147. To proceed, the site of research must first be situated in a superficial manner, with 'superficial' denoting both incomplete and shallow, as well as its localisation in the world. It is for this reason that this contextualising step is termed as 'situating', as the concept of situatedness speaks not simply to context, but to context that is lived and experienced 'in place' (Haraway, 1988 ; cf. **¶ 22**). As noted at the beginning of **§ 4.1.1.2**, this work intends to be explicit at each methodological step. However, this step will be familiar to any empirical social researcher, so this step will be presented in an abbreviated manner.

148. From the start, the site of research was to be located within the United States. There are two primary motivations for this choice: first, the author is from the United States, and thus has tacit knowledge of the sociopolitical and sociolinguistic landscape ; and two, given the timing of this research (commencing September 2017), the author wanted to take advantage of the 2018 midterm elections for the investigation of public discourse in an environment of heightened political attention.[67] The electoral cycle also allows us to specify a temporal bounding. Federal law sets Election Day (which is an annual occurrence ; see **n. 67**) as the first Tuesday after 1 November (thus for 2018, Election Day was 6 November). It was decided to set the window of analysis to the second half of the year, thus the six-month period starting on 1 July and ending with 31 December. The rationale was to investigate the sociopolitical dynamics of public discourse in the run-up to Election Day (at which point public engagement would likely be at its peak), while also incorporating discourse in the post-election period in which *omne animal triste est* except for the elected candidates and their supporters. (Note that the temporal bounding thus also serves a dual purpose, in that it provides an initial implicit political contextualisation.)

149. It was decided to narrow the location to single state, given the size and variety of the total US population.[68] The aim was to identify a state that mostly closely reflected national proportions in terms of various demographic measures.[69] Furthermore it was decided to exclude any states with an incumbent governor running for re-election, in order to minimize situations where the party of the head of state government potentially had undue influence in the media and thus on public discourse (incumbency effects are wide-ranging and extremely strong in the United States—e.g. Ansolabehere and Snyder, 2002). There were 16 states (out of 50) with an open race (i.e., no incumbent) for the governorship in 2018.[70] Some of these were discounted out of hand—California, for example, with a population of nearly 40 million is still much too large and its population too diverse ; Maine is comparatively old and extremely white even amongst extremely white states, while Georgia is the opposite. All would be interesting cases, but again the goal was to find the 'median' state, and the selected state was Michigan.[71] The demographics of that state are comparable to the country overall, in addition to providing a wide range of analytically interesting cleavages in terms of party identification and electoral results, ethnicity, age, urban–rural setting, economy and education, and so forth. At a population just shy of 10 million, and a population that is well mixed in terms of the two major political parties, Michigan was a sound choice.

150. As the site of research was already bounded to public microblogging platforms, situating the site in the United States obliges us to further that bounding to the Twitter messaging service (e.g. Weller *et al.*, 2013). A general motivation of this choice is that platform is a key object and site of research for computationally oriented social inquiry (Bruns, 2018). More specific motivations come from the situating of the site to the United States. That service was created in the United States, first available publicly in 2006, and has since become a regular fixture in academic literature and news media. In the US environment, Twitter is arguably the recognised default among social media platforms for politically oriented discussion, news, outreach, grandstanding, and complaint. While this was perhaps already the case some years ago, due to differences across platforms in affordances and user base, it certainly has been the case since the 2016 presidential campaign and election (e.g. Ott, 2017). Moreover, in practical terms Twitter has maintained a certain popularity among researchers because of the continued relative accessibility of its public application programming interface (API), through which users can readily search for and collect a wide range of data from the platform. This popularity possibly has increased given that Meta (née Facebook) restricted or closed much of its public data access in the aftermath of the Cambridge Analytica scandal (Freelon, 2018).[72]

151. In sum, the site of research has been superficially situated to discourse on the Twitter messaging service during the second six months of 2018 in the US state of Michigan. How this superficial situation will be understood as a context of action, specifically as a political context of action (following Theocharis and van Deth, 2018), will now be discussed.

4.1.1.2.2. Political Contextualisation

152. The argument for understanding the superficially situated site of research as a political context of action is straightforward: it is political because it is discursive. To elaborate on that argument, we must begin with the underlying assumptions. Following that, we address why this understanding is essential to the study of political participation in hybrid society, that is, as a socially communicative phenomenon (**chp. 1 § 1.1** ; e.g. Boulianne, 2020).

153. In part the argument here rests on this work's epistemological stance, as encapsulated in the reconceptualisation of political participation in hybrid society, discussed in **chp. 3 § 3.2.2** and presented in **chp. 3 § 3.3**. To review, the reconceptualisation makes five adaptations to previous conceptualisations:

1) no distinction is to be made between passive and active participation, as all acts are potentially consequential[73] ;
2) the distinction between individual and collective action is abandoned, as hybrid environments comingle them, thus obliging a move to meso-level understandings[74] ;
3) mobilisation is no longer seen as a subset of participation, but rather as its communicative manifestation[75] ;
4) the distinction between conventional and non-conventional participation is abandoned[76] ; and
5) evaluations of whether an act is participation according to intent or consequence are productively bracketed by shifting focus to context of action.[77]

The last adaptation is key: following Theocharis and van Deth (2018), identifying what is 'political participation' is best approached by way of taxonomic decision rules, rather than blanket definition.[78] We can see that the first four adaptations throw open the doors to public discourse, and moreover to the potential for politicised public discourse. In a sense those adaptations could be thought to describe a public sphere where conventions and norms have been weakened or removed—thus a public sphere potentially disrupted by unconventional modes of discourse (cf. Bennett and Pfetsch, 2018), which will be an important consideration as the argument progresses.[79] But the last adaptation sets the defining measure of political participation: the action must occur in a political context.[80] But then, if we hold that discourse is essentially political, and if the reconceptualisation of political participation is practically a template for public discourse in hybrid society, how then does not all public discourse fit the bill of political participation? In short, it all does. All of it. Let us examine why.

154. The argument begins with a fundamental assumption—that discourse is action. Students of political participation, among many others, continue to debate how to understand, perhaps to integrate, online (rather, hybrid) behaviour in that broad, contentious space of study, with much wrangling about what and which behaviours might be deemed as intentional, consequential, etc.—thus fitting various conventional conceptualisations of things political—and this is especially the case for phenomena of language (which includes platform affordances such as hashtags) (**chp. 1 § 1.3** ; **chp. 3 § 3.2.2** ; cf. **n. 79**). Students of language and of language in society,[81] however, have a head start with such understandings, in that they generally agree on a fundamental premise that discourse itself is action, embodying both intent and consequence (e.g. Austin, 1955 ; Searle, 1969).[82,83] However, it is a further step to frame discourse as political.

155. To understand discourse as inherently political first requires consideration of what 'discourse' denotes. Across the social sciences, the term is used and understood in countless ways.[84] In the studies of communication and media, the most common uses are perhaps as a general in-group term for 'talk' or 'topic', as a Foucauldian shorthand for referencing negotiations of power, and occasionally in a more rigorously methodological sense (e.g. Wodak and Meyer, 2001). All of these uses are valid in themselves, but also taken together, in that the bright thread running through is a concern with language as a social practice, and thus with meaning.[85] However, for the purposes of this work, that is, for the purposes of scaling up social inquiry from a communicative perspective, we need a more functional understanding of discourse.

156. G. Thomas Goodnight's (1982) ideal typology of 'spheres of argument' is well suited to this need.[86] And although ideal, it is nonetheless functional. The typology describes three primary spheres of argument, or rather, deliberation—the personal, the technical, and the public. Each sphere is discursive, that is, concerned with negotiating and sustaining meaning and acting upon it (cf. **chp 2. § 2.2.2** and **n. 75**). The personal sphere is the most familiar, represented by conversations and the like, where interpersonal relationships begin, grow, and decline. At the start of relationships discourse is guided by social norms, but grows more distinct and personal as the relationships evolve. The technical sphere is a space of discourse suited to endeavours that would be hindered by the varied and individual nature of interpersonal discourse. Here discourse is by necessity highly coded, constrained, and bounded in ways that suit the needs of a given endeavour. It can be thought of as discourse where meaning and action is determined through expertise, as in trials, operations, experiments, and so forth. The public sphere is where the community wrangles to establish its meanings held in common and how jointly to act upon them. In this 'plenary' state (that is, according to how the community determines its entirety), interpersonal relationships and expertise have much less hold on the discourse, as these discourses are now many and disparate. The personal and technical spheres may be drawn upon here—indeed they are crucial elements allowing the public sphere to emerge and be sustained—but they are drawn upon only to the degree that they are understood to cohere with the needs and practices of the community. Thus the three spheres are interdependent, but that interrelationship is contingent.

157. Because of this contingent interrelationship among the spheres, public discourse depends foremost upon persuasion for the shaping of communal meaning and intent. Goodnight (1987)[87] observes that such persuasion can manifest as a recontextualisation of the community's attention both in terms of its own space (that is, questioning the bounds of the community and who then is a part of it) and its own time (that is, for sake of its own weal and continuity, arguing which meanings and intents must be placed before or after others). Thus public discourse is the community continually reckoning (one would hope) with its own space and time—that is, place (cf. **§ 4.1.1.2** esp. **¶ 142**). From a place public discourse emerges, and public discourse shapes what that place is and yet might be. Goodnight makes two observations crucial to the purposes of this work and the current argument. First, public discourse is always contentious because on the one hand it seeks to find agreement on the membership of the community and its priorities, and on the other hand that process engages the whole community and its full range of preferences as to who belongs and what is to be done (1987, pp. 430–431). Second (again, in the ideal), public discourse is open to inspection—that is, the decisions of the community regarding its place are not final, nor are the processes by which they are reached. Thus even in periods of seeming accord, contention over legitimacy writ large is always a possibility, for better and for worse. And the reckoning of public discourse is continuous and ever unsettled—as it is necessarily subject to modes of discourse that the self-recognised community sees as unconventional (cf. ¶ **153**). So, while not all discourse is political, public discourse is always and essentially political.

158. Having established that public discourse is inherently political, the question remains: why should this idealised understanding be applied to the study of political participation in hybrid society? As noted above, unconventional discourses—such as those widely observed in contemporary discourse—are not only disruptive to public spheres, there are also highly disruptive to the study of them (Bennett and Pfetsch, 2018 ; Boulianne, 2020). At a theoretical level (for the moment setting questions of method to the side), this politicised understanding of public discourse presents potentially significant problems to conceptual (and disciplinary) boundings. Jan van Deth observed already more than 20 years ago—in light of the expansion of government activities since the middle of the twentieth century,[88] combined with an expansion of the 'repertoire and domain' of political participation—that "the study of political participation has become the *study of everything*" (2001, pp. 3–4 ; emphasis original). Indeed, it may well should be. By comparison, Émile Durkheim observed already nearly 130 years ago—in light of trying to develop a social science where every single action, even the most negligible, ultimately has some follow-on effect on someone or something (cf. adaptation 1, ¶ **153**)—that "there is … no human occurrence that cannot be called social" (1895, p. 50). One can see the difficulty is persistent. The point is raised not to discount van Deth's observation, but rather to highlight its disciplinary purpose. Neither scholar was trying to bound the world of human action simply so that their thoughts might be more orderly. Quite the opposite: both sought to chart a clearer conceptual path to systematic and disciplined empirical study despite the phenomenal tumult of our lives together.[89]

159. In the study of political participation, as introduced in **chp. 1 § 1.2–1.3** and discussed in **chp. 3 § 3.1**, the conventional approach to the topic has been grounded in institutional understandings of the political for much the same reason. That is to say, conceptualisations (and thus operationalisations and analyses) have given pride of place to party political understandings and the occasional role of the public in electoral choice, as then the field of inquiry is bounded with relative clarity. In part for such disciplinary reasons, even as researchers across academies have been working to adapt to the drastic shifts in sociopolitical and informational contexts of the last several decades—with 'digital' participation now firmly ensconced among the modes of recognised and theorised participation (Boulianne, 2015, 2020 ; Theocharis, Moor and van Deth, 2021)[90]—the predominant, underlying conceptual theme remains an essential privileging of institutionally sanctioned forms of participation (e.g. Boulianne, 2020 ; Ohme, 2018 ; Ohme, de Vreese and Albæk, 2017 ; Stolle, Hooghe and Micheletti, 2005 ; Zukin *et al.*, 2006). And for these same disciplinary reasons, whereas the decline in 'conventional' (i.e. institutional) political participation has been observed and bemoaned across decades and countries (see **chp. 3 § 3.1.2**), the concomitant growth of, or at least attention paid to, more localised energy and action has frequently been classified as 'civic' engagement[91] ; while lauded it is nevertheless treated as distinct from 'political' engagement or participation (e.g. Ekman and Amnå, 2012b).[92] The implicit logic seems to be: if one does not act as institutions ordain, how can that be politics?[93]

160. The institutional approach is problematic from the perspective of reorienting the study of political participation as a socially communicative phenomenon (**chp. 1 § 1.3**). The core problem that this work seeks to address is that any institutionally oriented approach to the political will by definition view the myriad social phenomena of society through the lens of those same institutions. That is, an institutional perspective will see well only those groups that a government serves well. Mechanisms of governance are designed for and managed by those social groups that a government serves. In the ideal, that group is the community entire, and furthermore a community that is responsive to renegotiations of its space (**¶ 157**). In reality this is rarely the case—as issues of community membership and priorities are contested negotiations of power—and therein lies the oldest tension in large-scale social life.[94] Institutional perspectives, such as the conventional approach to political participation as described in this work, and ideal perspectives, such as Goodnight's notion of the spheres of deliberation,[95] are necessarily naive in regard to such contestations. By definition and by purpose, such idealised and institutional framings do not properly consider actors that work to reduce the space of the community and the systems that might thereby emerge, to which we will return shortly.[96] Rather, the institutional ideal assumes the maintenance or expansion of community membership and personal rights therewith.

161. At the heart of the conventional study of political participation is a belief that democratic institutions do right by changing spaces of community:

> Throughout history, most polities—including the United States—have not adhered, even in rhetoric, to principles of political equality. In most modern democracies, however, overt barriers to universal political rights have fallen. And, at least since the voting rights reforms of the 1960s, political rights have been universalized in the United States. With relatively insignificant exceptions, all adult citizens have the full complement of political rights. (Verba, Schlozman and Brady, 1995, p. 11)[97,98]

That belief extends, moreover, to understandings of differentials in participation. The bright outlook just quoted continues, noting that equal rights do not produce "substantive equality in their effective use. Individuals and groups differ significantly in terms of whether they take part at all and, if so, how much and in what ways" (ibid.). The tenor of this observation, which is characteristic of the conventional study of political participation, is an exonerative formulation that acknowledges differential participation without naming names or pointing fingers. In that the institutions have been doing their bit, the implicit onus of efficacy is placed on the individual. And so, for a generation and more, the more conventional literature has worried over various purported 'crises' observed in the data—crises of democracy (Huntington, 1975), participation itself (Curtice and Seyd, 2003), communication (Dryzek *et al.*, 2019), and so forth.[99]

162. Where did all these crises come from? What even does crisis mean? That notion can be traced back most clearly to the subfield of comparative politics, specifically to the work of the Committee on Comparative Politics of the Social Science Research Council.[100] Lucian Pye, in his work *Aspects of Political Development*, explains that there are "six crises [in political development] all of which must be successfully dealt with [regardless of sequence] for a society to become a modern nation-state" (1966, p. 63). These crises are of identity, legitimacy, penetration (of the State into society),[101] participation, integration, and distribution. Pye notes that the ‘participation crisis’ occurs when there is uncertainty over the expansion of popular participation and the influx of new participants strains existing institutions ; further, these new interests and issues disrupt the previous polity and it must then be refashioned—”In a sense the participation crisis arises out of the emergence of interest groups and the formation of a party system” (p. 65). Is crisis of participation then characteristic of the State? Can crisis ever truly be absent? (Note also how Pye’s idea is echoed in Goodnight’s later description of the public sphere’s perpetual renegotiation of community.) That discussion of crises is brief, a mere five pages, but Pye notes (p. 63, fn. 13) that his analysis is informed by a forthcoming study sponsored by the Committee on Comparative Politics (of which Pye was a member). That study is *Crises and Sequences in Political Development* (Binder *et al.*, 1971). In that work, Myron Weiner (in his chapter “Political Participation: Crisis of the Political Process”) offers to the reader a rare gift—a clear definition of a contested term:

> A participation crisis can be defined as a conflict that occurs when the governing elite views the demands or behaviour of individuals and groups seeking to participate in the political system as illegitimate. (Weiner, 1971, p. 187)[102]

163. Weiner’s framing of crisis—and explanation of its source—is a distinct shift from Pye’s brief description. It does in fact stand apart from most conventional discussions of crisis in participation. Consider the following from Sidney Verba, writing nearly 30 years before writing the congratulatory passage quoted in ¶ **161**:

> [T]he problem of participation concerns both the participants and the decision-makers. It is as important that decision-makers respond as it is that the participants participate. But what makes a decision-maker—and remember we are talking not only of congressmen or mayors or other elected officials, but of government employees of all sorts from postmen to policemen to planners who can make decisions affecting individuals or groups—listen to participants and modify [their] behavior to suit them? (1967, pp. 75–76)[103]

That is a fair question. From an institutional perspective, Verba observes that motivations, such as beliefs and values, are central, but they cannot be relied upon. And so, he suggested that motivation must be built into the system in a manner so that for decisionmakers there are benefits to promoting participation and costs to blocking it, and furthermore that decisionmakers must necessarily foster participation among supporting groups to offset opposing groups (p. 76). That is an apt description of the institutional landscape, at least in the United States. But we must be clear about the poison pill contained in such prescriptions—recalling that Verba is perhaps the foundational scholar to the study of political participation in the post-behavioural period (see **chp. 3 § 3.1.2**). That is, costs may be avoided and coalitions may be built if there is agreement among a subset of the population that they may empower themselves through the exclusion of others.[104] The group motivation can become exclusion itself, though veiled in legitimising words and actions (e.g. Arendt, 1951 ; cf. Butler and Spivak, 2007, pp. 14–16). Regardless of individual motivations, which we cannot know, the outcome for the remaining polity is the same—it is intentionally reduced, lessened, and thus is a distorted subset of a potentially greater society. The institution only sees those who it wishes to see.[105]

164. That is the core of the problem with the conventional, that is, institutional approach to political participation—it discounts those groups that the system has been shaped to exclude. So on the one hand the institutional approach blinkers itself, and that makes it bad science. But in continuing steadfastly forward in this manner, institutional approaches thereby contribute to an entrenchment of that same exclusionary, reductive system. So on the other hand the institutional approach contributes to continued processes of exclusion, and that makes it something for each researcher to ponder for themselves.

165. In the United States, in which the site of research is located, it has been the case throughout the country's history—and is patently resurgent in the current moment—that various social groupings have been systematically excluded, both by law and in practice, from institutional processes of governance, and even from society itself—the space of the community has been reduced (Foner, 1988 ; Keyssar, 2000 ; Manza and Uggen, 2006 ; Rogers, 1992 ; Shklar, 1991 ; Wiebe, 1995).[106] The boundaries of these groupings implicate ethnicity, sex, sexual orientation, political orientation, religious identification, relative wealth and education, place of birth, personal histories, and on and on and on.[107] Not only is there a broad body of literature addressing these various topics,[108] but moreover we have corroborating evidence, albeit circumstantial yet nonetheless empirical, of the fact of such exclusion and its effects. That evidence, from which the fact can be inferred, is to be found in the recent decades of literature on political participation (**chp. 3**). The continued hand-wringing over various crises is perhaps indicative of yet another—a crisis of normativity in the study of political participation.[109]

166. To conclude the overall argument, we recognise that an 'approach' is an orientation towards a field of study, a more or less implicit methodology. As such it comprises theory underpinning method. The preceding discussion demonstrates that institutional approaches are unsuitable for critical, quality examination of political participation in contemporary society (cf. Schwartz, 1984). We have addressed the issue of institutional conceptualisations, and now we must turn that same eye to method. Unfortunately, the same must then be said for institution-oriented methods which rely predominantly on quantificationism (cf. **n. 117**)—on the fungible individual—as their inherent decontextualisation and atomisation removes <u>communities</u> from view and thus further distorts the study of political phenomena, while enabling social distortions within the phenomena themselves—that is, society itself. To address political participation in hybrid society with any fidelity—which, as we have shown here, is not simply an academic exercise but rather an ethical and pragmatic obligation towards the community entire—we need to adapt our methods.

4.1.2. The Operational Step

167. The conventional study of political participation has long been dominated by models basing explanation on sociodemographic variables of one sort or another (**chp. 3 § 3.1.2**).[110] The operationalisation of such models has produced methods that are consistently effective for their intended purpose (cf. **n. 120**). However, those models are based on a subject model that is not appropriate for the study of political participation in hybrid society, that is, political participation reconceived as a socially communicative phenomenon. Therefore this work proposes a shift in methods from those based on sociodemographic variables to those based on variables rooted in the social practices of language. Before specifying the exact method applied in this work, we first address the ontological impetus for this change, the subject model in question, and how theoretical adaptation to large-scale social subjects can be accommodated by method.

168. Changes in methods can be understood as resulting from a technical, epistemological, or ontological impetus.[111] A technical impetus is the simplest, wherein developments in tools and their production prompt concomitant developments in the methods that apply them to research. A deeper epistemological impetus can then result from technical developments, in that tools and techniques are not simply better, but rather can now be applied to new questions and to new phenomena. A thoroughgoing ontological impetus for change in methods can result when accumulated technical and epistemological developments reveal the limits of our current practices—thus stimulating the re-examination of basic understandings (cf. Kuhn, 1970).[112] Note that while technical issues may or may not prompt questions of method,[113] epistemological and ontological issues necessarily prompt methodological problems, that is, those implicating theory and method jointly. Having completed the conceptual framework of this work with the specification of the theoretical approach in the preceding section, we now proceed with a specific focus on method by taking the operational step. First we ground the step taken here, which will be a focus on meaning as negotiated through public discourse.

169. In regard to the study of political participation, we have seen in **chp. 3** how adaptations in method with technical impetus—specifically, the incorporation of survey methods to aid ethnographic work—soon developed into methodological adaptations with epistemological impetus in large-scale survey research (§ **3.1.1**). The post-behavioural period spurred yet further epistemological adaptations (§ **3.1.2**), as did the sociotechnological shifts of the arrival of hybrid society and new media (§ **3.2**). Nothing has for a moment stood still. For some time now, many have noted the need for methodological adaptations with ontological impetus (§§ **3.2.1–3.2.2**). That is to say, preceding adaptations have resulted in a weight of evidence that—in its divergence from the expectations of accepted models—demands a re-evaluation of underlying theory. Calls to reconceptualise political participation, such as that of Stuart Fox (2014) and that made in this work (§ **3.3**), are not simply tweaks of the knobs. Rather, they implicate the subject model that underpins the conventional approach, and that is an ontological problem.[114]

170. The concept of 'subject model' describes how researchers conceive of the subject of their research. In the context of the social sciences, setting the science aside, how is the 'social' conceived? That model determines what phenomena are theorised, how theory is operationalised, and how data are analysed and interpreted. In the words of Herbert Simon, a twentieth-century paragon of the conventional approach in political science generally, "Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying" (1985, p. 303). Indeed. However, to be clear, Turing- and Nobel-recipient Simon modelled social processes by means of linear algebra (cf. Simon, 1957, part II).[115] Yet, while quantification in political inquiry is often the target of critique from those of a more interpretivist mindset (e.g. Monroe, 2005), it is not the root of the problem.[116] Quantification is essential for large-scale inquiry (and thus large-scale data). Rather the root of the problem is an asocial subject model.[117] The model sees the individual subject as isolate—not simply distinct in awareness in a Cartesian sense, but distinct unto itself. That conception was naturally strongest during the behavioural period,[118] but thereafter the isolate individual was subsumed into the axiomatic concept of 'the citizen' (as similarly the State would be subsumed into 'the system'). As observed in **chp. 3 § 3.2.1**, the actual participants—be they 'citizens', 'members', 'people', what have you—are seemingly without fail undefined (cf. **n. 40**), the core concept of the 'citizen' as participant unspecified (beyond prescribing how they might properly participate). In the most influential definition of political participation (**chp. 3 § 3.2.1**), Sidney Verba and Norman Nie (1972) state in their opening definition of the topic of inquiry that their focus is on "those actions of private citizens" (p. 2), with 'private' opposing the implicit 'public', that is, politically professional or elected citizens. Nevertheless, they are quite clear about their purpose shortly thereafter:

> Our main concern is with participation as an instrumental act by which citizens influence the government. Further, we wish to consider participation from the perspective of the functioning of the United States as a democratic polity, rather than from the point of view of the individual citizen as a participant. We are more interested in politics than in political psychology, more in the ways participation by citizens conditions the way political decisions are made than with the social and psychological reasons for individual participation. (p. 5)

And that is the core of the institutional focus, as has been demonstrated. Thus the isolate individual is replaced in the post-behavioural period with the fungible individual, now with even more quantifiability. In the analysis leading to the reconceptualisation of political participation, it is observed that "The State is the manifestation of mobilised citizens" (**appendix A, ¶ 5**)—we must be clear, the field of political participation does not study citizens to know how they might be served, but rather to know how the State might be sustained and nurtured.[119]

171. So then, one finds the subject model of the conventional approach to be an atomistic, economic, and rationalist conception of the individual (Lindenberg, 1990 ; cf. **n. 62**)—each interchangeable with the other, so long as they fit a bracket. The subject model proposed in this work however, to suit the study of political participation as a socially communicative phenomenon, does view individuals from an embodied perspective, but conceives of us not as isolate or fungible, but rather as acting jointly and intersubjectively through the warp and weft of meaning (Emirbayer, 1997 ; Emirbayer and Mische, 1998 ; Grossberg, 1982). One might say we are legion.

172. As noted in **chp. 1 § 1.3**, institutional approaches to social phenomena are effectively blind to communicative aspects beyond a narrow 'political' understanding of them. While this blindness is perhaps exacerbated by highly quantified methods, the underlying cause is theory, as demonstrated in the preceding section. Thus conventional methods are unsuitable for observation of political participation in hybrid society.[120] We have shown that the need for adaptation has an ontological impetus, and so requires an ontological shift. This work argues that quantified operationalisation—essential for large-scale inquiry—can avoid the blindness of institutional orientations to socially communicative phenomena through a reconceptualisation of the subject model of political participation. The **ontological shift** is:

1. **from the structural 'fungible individual'**—where collectives are identified by their exogenously presumed place, function, and intention within the structures of society and governance (cf. Fox, 1996) ;
2. **to the communicative, intersubjective 'social person'**—where collectives are seen, but not defined, by means of their endogenously produced descriptions of context.

173. A relational ontology of political participation allows for an epistemology that is sensitive to intersubjectivity (Duranti, 2010) and manifold contexts (Emirbayer, 1997), and thus one that enables the interpretation of phenomena that are socially communicative in character. As described in the first section of this chapter, the ontological footing of this work is relational in nature, and the epistemological stance is socially sensitive as it is directed towards issues of context. Building from this base, the considered theoretical approach is directed towards the examination of public discourse as an overarching context that is inherently political. Having demonstrated that the progressive weakness observed in the conventional study of political participation derives not from the methods themselves (cf. **ns 110**, **120**) but rather from a failure in the subject model, we note that the conceptual framework employed here has been designed to implement a social subject model as just described. The operational step that follows declares the empirical link between the conceptual framework and the phenomena of interest. That is, the step declares what exactly will be observed and measured at the site of research. Recall that the purpose here, as described in **chp. 1 § 1.3**, is to demonstrate an alternative approach to data collection and analysis for the study of political participation. The goal is to emulate the delimitation of social groupings by means exogenous sociodemographic variables, except to do so by means of endogenous linguistic variables.

174. The **operational step** is to observe the language used by individuals in terms of their various vocabularies and grammars, and to measure the relative frequencies of certain linguistic elements. The goal of such measurements is to look for pervasive and persistent differentials in patterns of language use in the population at the site of research. That is to say, we are looking for subsets of the studied population that share certain habits of speech. Such subsets will not be mutually exclusive (quite the opposite is expected, as will be discussed in the following section). Rather they are additive, in a sense, in that we expect to observe various reoccurring complexes of habits—think of these, if you like, as something of a 'style' or 'genre' of language, like an accent but in text. The supposition is that these complexes might serve as a proxy for social contexts, as will be discussed. Before moving to the method proper, first a brief justification of why the operational step is towards language, rather than something less complex and likely more tractable.

4.1.2.1. The Step towards Language

175. In the context of the conceptual framework, specifically the ontological footing, the operational step of this work can be understood as a technical view of language as mediating and thus bringing into being the sociomaterial configurations of society. An alternative, and understandable, assumption might be that language would naturally fall in the domain of the social, given the dominant roles of language in social life, especially in terms of expressing meaning. But that assumption would obscure the functional operation of language. Language is not some inherent capacity or faculty that we use simply to broadcast and decode received meanings (e.g. Reddy, 1993 ; cf. e.g. Chomsky, 1975 ; Jackendoff, 2002). Rather, language is the primary means through which meanings are negotiated[121] and with which we do things together (Austin, 1955 ; Grice, 1957 ; Searle, 1969)—it is the medium of sociation through meaning, of cooperative instantiation of intent (Hasan, 1995, 2005 ; McLuhan, 1964). Thus language is not meaning itself, but rather the mechanism of meaning—to put things again in terms of the ontological footing, the defining affordance of language is to negotiate meaning. Seeing language in this light promotes it from a second-order phenomenon that marks social structures and understandings (e.g. Labov, 1966) to a first-order phenomenon that is essentially the social hermeneutic from which structures and understandings emerge (cf. Labov, 2002). Society embeds itself in technology, as demonstrated in **chp. 2,** and language is the prototype.

176. The history of language as object, subject and method of inquiry is ancient.[122] It is axiomatic that observation of language can provide insight into social phenomena. With a sufficiently broad conception of language, such an axiom verges on tautology, restrained from self-reference solely by the fact that both observation and insight are socially contingent (cf. Rorty, 1980b, part I). From a social scientific perspective, the study of language presents a risk similar to that noted in **chp. 3 § 3.1** with regard to the study of political of participation—a risk of the "study of everything" (van Deth, 2001).[123] Thus in the contemporary social sciences there is great variety of delimitation in approaches to language as a tool of knowledge work, underpinned by a similarly great variety of theory and method.

177. The use of language for qualified, interpretive social analysis has a long tradition predating disciplinary efforts. Such uses abound, having been taken up across the fields and disciplines as the groundwork of much contemporary theory. The use of language for more quantified, positive social analysis is however a more recent development. While significant theoretical and practical advances have been made in computational and statistical approaches to linguistic analysis since their beginnings in the mid-twentieth century (Ferrari, 2004), the application to social analysis has only become prominent recently,[124] due in part to widespread availability and affordability of computing resources and to the proliferation of data accessible through new media and associated technologies. Such applications to social analysis have since rapidly taken hold in various fields.[125] However, while both interpretive and positive approaches might aim towards social analysis, the former approach generally draws on social theory, whereas the latter often draws on information theory or perpetuates questionable social assumptions (e.g. Nguyen *et al.*, 2016 ; cf. Coupland, 2007).[126]

178. Although it might seem an obvious critique to fault any one approach, or specific method, that is applied without sufficient domain knowledge, such criticism forgets the potential stimulation of new social theory—more quantified, positive approaches at the remove of scale may well reveal patterns in language that more qualified, interpretive approaches in proximity could not detect, and thus that current social theory might not account for. In that manner such approaches to language can provide evidence to inform the development of new social theory (e.g., Evans and Aceves, 2016). Why should that be?—for the simple reason that language at scale encodes collective meanings (Teubert, 2005, pp. 2–3 ; cf. Nartey and Mwinlaaru, 2019). That is the basic affordance of language as already stated: the negotiation of meaning. For the purposes of this work, the question then is how to reveal meaning in language at scale.

179. The conventional study of political participation has pursued statistical analysis of sociodemographic variables in order to identify and to understand the various groupings that it expects to see in society. As the level and scope of analysis is generally high, the statistical approach is logical, likely necessary. And as theory in that field has given an explanatory role to resources, the use of sociodemographic variables is a natural choice (and data is plentiful). However, purported social categorisations based on resources are synthetic, imagined from the outside. To identify social groupings by income, education, status, and so forth is to categorise by exogenous variables. That is, the variables do not emerge from any sociality of the grouping itself, but rather have been chosen by some other rationale (a different subject model is prime example). While fit for their appropriate uses, exogenous variables (and the resultant categories) are not suited for social phenomena—even when they seem 'good enough', they nevertheless mislead and obscure. The use of exogenous variables and categories has contributed in part to the observed weakness of empirical studies of participation in recent decades. As discussed in **chp. 3 § 3.1.2**, the problem at hand was that studies based on well-established models of participation (e.g. the arguably dominant 'socio-economic standard model') and their indicators (i.e. the various socio-economic resources in question) produced empirical results that seemed to suggest a decline in political participation and civic engagement. Other lines of inquiry, however, saw with the emergence of hybrid society a flourishing of non-institutional, non-establishment participation (¶ **110**). The problem as it came to be understood was not a weakness in the empiricism of these studies, but rather in the analytical lens that resulted from privileging exogenous variables, such as socio-economic indicators, and thus a focus on "categoric groups" (Conway, 1991, p. 45 ; see **chp. 3 § 3.2**). That is, the problem was in the subject model employed in studies relying on exogenous variables, as discussed in the previous section.

180. This work intends to maintain the use of statistical analysis. As noted above, the quantification of social phenomena is not itself the root problem—that being the atheoretical[127] application of quantification. Rather than relying on exogenous variables, such as the sociodemographic, this work will use endogenous variables that are derived from social collectivities themselves.[128] Given that the theoretical approach assumes a focus on public discourse and the site of research is partly specified as the Twitter microblogging service, drawing on language to supply endogenous variables is an evident choice. However, operationalising language is not straightforward, given the complexities of the social practices involved. Recalling the motivating hypothesis of this work that the empirical study of language can help to contextualise socially communicative phenomena, it must be stressed that context does not stand apart from collectivity—the concepts are coextensive.[129] Thus to contextualise is to distinguish among collectivities and to witness their interactions (cf. Crenshaw, 2019 ; and Haraway, 1988). Thus we focus on an operationalisation that highlights social boundaries.

181. Stepping back into the conceptual framework, we are interested in the functions served by language in the context of the site of research. That is, if we intend to trace the boundaries of collectivities and contexts as continually emergent sociomaterial configurations (which we are, in terms of the framework presented in **§ 4.1**), what are the functions (i.e. technical affordances) that mediate between meaning (i.e. the social) and consequence (i.e. the material)? And more specifically, what are the specific functions that are drawn upon differentially by subsets of the population that could be understood as proxy markers for social contexts, and thus as endogenous markers of social collectivities?[130] The following section addresses these functions of interest.

4.1.2.2. The Functions of Interest

182. The functions of interest that are operationalised in this work are certain lexicogrammatical features[131] that are expected to be markers of social context. Approaches with similar impetus are increasingly common in large-scale studies these days (as will be discussed briefly) but in light of the preceding discussion in this chapter, this work will take a largely different approach to the selection and measurement of the features concerned. This work will measure two broad categories of lexicogrammatical features: generic features that have been found appropriate for the study of genre (hence 'generic') and register in text[132]* ; and also structural features, those particles and wee clumps that glue everything together (such as the article, the conjunction, and the woolly preposition).[133] Before proceeding to the method in **§ 4.2**, we first note some current trends in quantitative operationalisation of language for social inquiry, and justify why a different approach is taken here.

183. As demonstrated above, that language has great utility for studying social phenomena is given, as its primary affordance is the negotiation of meaning, so enabling collective action and organisation.[134] Thus across the social sciences proper—and certainly including the humanities—language in its countless modes is variously object, subject, and method of study.[135] What is not immediately given is the utility of language as a source of empirical data for studying phenomena at scale.[136] Language, as both cause and effect of human sociation, is copious, contextual, and controversial. That it is copious obliges the use of quantitative techniques, and that it is contextual challenges such techniques as they often fail in accounting for the social practices of language. That it is controversial is to say that language is a bearer and object of ideologies, be they national–colonial (Anderson, 1983 ; Haugen, 1966), sociopolitical (Bernstein, 1971 ; Milroy, 2000 ; Trudgill, 1974 ; Wolfram, 2007), or disciplinary (Gal and Irvine, 1995 ; Milroy, 2001). This last aspect is key for how this work will approach the challenge of contextuality: language is not simply a bearer and object of ideologies but rather, as noted above, it encodes such collective meanings—including situated, motivated understandings of language and its use (Errington, 1999). In that a given polity will be a shifting assemblage of contexts, and thus ideologies (cf. Bourdieu, 1993), we are rightly justified to understand public discourse as inherently political (§ **4.1.1.2.2**), and furthermore we are rightly justified to expect differentials in use to mark those sociopoliticised contexts (cf. Eckert, 2008).

184. The study of such differentials, or variation, is the stock in trade for certain fields,[137] the most prominent being sociolinguistics.[138] That field, and the study of sociolinguistic variation generally, has emerged from several traditions of work, including the sociological (Bernstein, 1971 ; e.g. Fishman, 1972) and the ethnographic–interactionist (e.g. Gumperz, 1971 ; Hymes, 1974), although the emergence of the variationist tradition with William Labov's (1966) work on social stratification subsequently gained the most attention and disciplinary influence. Penelope Eckert (2012, 2018) describes the evolution of the study of sociolinguistic variation as proceeding through three phases: the variationist, which correlated differences in 'linguistic variables' (which has somewhat specific meaning in this tradition, which will be addressed) with socioeconomic and demographic categories ; the ethnographic, which sought more local understandings of variation ; and now an emerging phase that seeks to understand variation in context as emerging from meaning. This third phase understands that variation is not simply a side effect of pre-existing social groupings, but rather is itself constitutive of meanings and groupings. The relevance of such a perspective to this work is evident. Also evident is the broad similarity to the evolution of the study of political participation. Consider: Eckert observes that the first two phases of sociolinguistics were strongly bound to assumptions of relatively static social categories and hierarchies. While the variationist tradition began in an ethnographic mode (e.g. Labov, 1963), "subsequent studies came to focus on … macrosociological categories. In this way, speakers emerged as human tokens—bundles of demographic characteristics" (Eckert, 2012, p. 88). To this issue we will return in a moment.

185. Despite its methodological evolution, disciplinary work on sociolinguistic variation is still very much limited in scope of application. Regardless of specific tradition of study, all of which are active in one way or another, such research is nevertheless time- and labour-intensive. While in recent years computational approaches to variation have gained momentum (e.g. Nguyen *et al.*, 2016),[139] we have already noted that these approaches in the main tend to proceed in a socially atheoretic manner, drawing more on the 'computational' and 'linguistic'[140] than the 'social'. In terms of Eckert's phases, computational sociolinguistics seems to inhabit the first phase of categorical variationist study. There we should not and will not follow, for the same slew of reasons that motivates this work to adapt the study of political participation to hybrid society. Yet, it is important to note that the author's position of resisting socially atheoretic methods has only come near the end of this process, whereas in the first four years this was not the case. It was only in coming to understand the ontological problems in the conventional study of political participation that the author realised that a similar dynamic was at play (and arguably with similar drivers) with 'state of the art' approaches to the large-scale study of meaning in language.[141] Thus, roughly a year before this writing, a new approach was desperately needed (cf. **n. 155**).

186. The solution was found in approaches to linguistic variation in text—grounded in understandings that variation is a mode of social action (Miller, 1984)[142]—and specifically in corpus-based approaches to such variation at scale. That solution is based in the concept of genre,[143] understood here as complexes of linguistic features that correlate with recurrent social contexts and functions. In that this work is to operationalise by measuring endogenously produced linguistic variables, and in so doing detect potential collectivities, the solution is to conceive of such collectivities as genres.[144] Understood as an extension of the ideological approach to the challenge of contextuality noted in **¶ 183**, this is a potentially productive approach to the study of meaning at scale, and thus of society.

187. The specific method chosen for this work is a repurposing of Douglas Biber's multi-dimensional approach to the analysis of genre (Berber Sardinha and Veirano Pinto, 2019 ; also Biber, 1988, 1992, 2019 ; and Biber and Jones, 2005).[145,146] The basis of analysis is a theoretical foundation that addresses the interrelationship of social context, communicative function, and linguistic form (Biber and Conrad, 2019, chp. 1.3). Genre is seen as manifesting in pervasive patterns of lexicogrammatical features (recall that this means simply words and grammar) serving joint function in a given context. The approach is multi-dimensional in that it resolves genre according to certain categories of variation (e.g. those marking text as 'abstract vs situated' or 'reported vs immediate'). Moreover the approach uses multi-variate statistics (as genre implicates complexes of features) that are relatively simple compared to the 'state of the art'—this is a key point in that it makes Biber's approach readily accessible to a broad range of researchers, and readily applicable at scale without need for intensive computing resources.

188. The repurposing of the multi-dimensional approach that this work will implement is a reversal of focus. Whereas the multi-dimensional approach would ideally proceed by analysing context and function[147] in order in order to reveal, statistically, lexicogrammatical dimensions of variation characteristic of recognised textual genres, in this work specific contexts are unknown and will remain so. However, the overarching context is demonstrated to be contentious, politicised public discourse. If we work from the assumption that, in such overarching context, the population at the site of research will use linguistic affordances to index ideology, and thus identity (Gal and Irvine, 2019), then observed differentials in dimensions of variation can be used to impute the presence of social collectivities (cf. Biber, Egbert and Keller, 2020).

189. At least that is the plan. The plan is sound, in that the multi-dimensional approach has strong record and clear method of application. The method used here will follow the template for applying multi-dimensional analysis (Berber Sardinha and Veirano Pinto, 2019),[148★] which includes the use of Biber's (1988) original set of linguistic features as well as a recreation of the part-of-speech tagging algorithm used to annotate those features in text (Nini, 2019).[149] This is a challenge in two regards. First, in relation to the relatively 'non-standard' language variants encountered online, and especially in microblogging contexts such as the site of research, common taggers developed on 'standard' variants such as newsfeeds and such are not worth a damn (Derczynski and Maynard *et al.*, 2013 ; Finin *et al.*, 2010).[150,151] Second, and in a sense extending the first regard, statistical approaches to language (i.e. to language at scale) that have root in anachronistic social categorisations (cf. Coupland, 2007, 104)—which most taggers implicitly do[152]—or that rely on socialised expectations of how language (of whatever type) should look (e.g. Vidgen *et al.*, 2019 ; Warner and Hirschberg, 2012), will be confounded by common community practices (e.g. Nilep, 2006) and the sheer variety and messiness of real language as compared to training data.[153]

190. For these reasons, in part, the operational step supplements the multi-dimensional approach (which is necessary to this work as an important comparative baseline) with the inclusion of structural features. Such features, often referred to as 'function' words, are those bits and bobs that string together 'content' words such as nouns, verbs, adjective, and adverbs (Fries, 1952, chp. 6). These structural features will thus include the articles, those demonstratives, some quantifiers, modal markers (as they should), and conjunctions, among others (such as prepositions). The reason to include these is, first of all, such features comprise a relatively 'closed class', which is to say it is a limited feature set (perhaps one or two hundred commonly recognised in 'standard' English) and is also extremely resistant to change—shifts in such linguistic classes occur over historical time periods, not over brief spans such as observed with slang. Furthermore, evidence suggests that use of structural elements is fairly stable across the lifespan compared to other linguistic elements, that they are extremely resistant to conscious manipulation and even attention, and that their occurrence is relatively independent of topic but dependent on speaker ; such characteristics make structural elements of great interest to researchers in author attribution, for example (cf. Juola, 2006 ; Kestemont, 2014 ; Koppel, Schler and Argamon, 2009). Such characteristics are certainly of interest for this method. However, of greater interest is that structural elements are difficult for us to learn and to master (Jolly, 1981), influence how we learn and produce other languages (Angelis, 2005), and yet seem to provide a deep structure through which grammatical language is acquired (Dye, Kedar and Lust, 2019). Such evidence suggests that structural elements are perhaps a good place to go looking for markers of socialisation—that is, endogenous markers of context. Even as we all are multiple selves that communicate in shifting codes according to context, day after day, there are still some sticky bits that resist these shifts, and that are resistant even to our conscious attention and manipulation. Notably these are also features that computational approaches to language have tended to remove, ignore, elide, or subsume into other features. Any approach informed by traditions or procedures developed in information retrieval (such as the removal of so-called stopwords, or common methods of term weighting) will likely obliterate the wealth of social information that structural elements might carry. In any case, the potential efficacy of such words as endogenous measures is a hunch, and one that this method has been partly designed to investigate.[154] With the functions of interest described and justified, the method is now presented.

NOTE: The operationalisation of function words proposed here is expanded in the method to the operationalisation of pervasive features (**chp. 5 § 5.1.3.2**). The rationale for the change is that any set of pre-selected function words implicates an exogenous perspective of standardised language, whereas evaluating the pervasiveness of features in the corpus can be pursued in an endogenous manner. The discussion above is left in place because it is nonetheless a valid observation of the potential value of function words as markers of socialisation, because function words are subject to increasing attention in some computational strands of work, and because the rationale for expanding the operationalisation so as to take a more endogenous approach underscores the social reasoning that this work promotes.

4.2. The Methodology

191. This chapter has detailed the theoretical approach of this work and the operational step that will be taken in the method. The approach was developed by elaborating the epistemological stance, specifying it in a manner that maps the relations between the phenomenal (i.e. the site of research, the US state of Michigan) and the conceptual (i.e. the theoretical approach). The epistemological stance itself had been derived in **chp. 3** by interpreting the hybrid reconceptualisation of the topic of study (political participation, in this work) in light of the ontological footing. The footing itself was derived in **chp. 2** from a synthesis of thinking on the relationship between society and its technology, and a subsequent reframing in light of the nature of hybrid society. Taken together, these components comprise the conceptual framework of this work.

192. Much ground has been covered in this chapter and those preceding. This is due to the interdisciplinary character of this work ; to pursue macro-level social inquiry in a manner that departs from structural, institutional understandings so as to privilege social, communicative, and thus human understandings is no mean feat. To do so must necessarily weave together strands of work and understanding from across fields and disciplines, times and cultures. Yet the result is that we have arrived at a clear understanding of and argument for how the conceptual framework can be implemented in method. Taking the fully specified conceptual framework and that method together, we will arrive at the full methodology of the work. The method is now presented in the following chapter.

Chapter 5

The Method

> Yet it is clear that if the process of unification advances beyond a certain point, the city will not be a city at all ; for a state essentially consists of a multitude of persons, and if its unification is carried beyond a certain point, city will be reduced to family and family to individual … . [N]ot only does a city consist of a multitude of human beings, it consists of human beings differing in kind. A collection of persons all alike does not constitute a state.
>
> Harris Rackham, *Aristotle: Politics*, 1932, pp. 71–73 ; 2.1, 1261a

193. In terms of methodology, this work argues that the conventional approach to the study of political participation is misled by an asocial subject model, and so blinds itself to social phenomena in hybrid society. This work proposes a shift in subject model from the fungible individual to the fully social person. In terms of method for large-scale social inquiry, that is, for societal inquiry, such an ontological shift urges a move away from techniques relying on exogenous variables (such as the sociodemographic variables used in resource models of political participation), towards techniques that seek endogenous variables—thus moving away from assigning societal phenomena to expected categoric groupings, to instead seek groupings that emerge from societal phenomena themselves. The method described here derives such endogenous variables from discourse, and seeks to trace potential groupings by means of pervasive patterns in these variables. The goal, as noted in **chp. 1 § 1.1**, is not a wholesale refashioning of method for the study of political phenomena, specifically political participation, but rather an investigation into the effects of the shift in subject model on how groupings in society might be seen. The goal is that these methodological adaptations can produce analysis that is contextually sensitive, yet nevertheless remains interpretable from a structural perspective—as discourse is understood to be the root of social structure.[155] That is, the method aims to produce an alternative, more contextualised means of viewing societal groupings at scale.

5.1. Overview[156]

194. As a test of the research questions (**chp. 1 § 1.5**), the method presented here was used to prepare a dataset that combines discursive, sociodemographic, and geographic data of the same population sample. Such a composite dataset—here referred to as an 'enriched corpus'—allows us to identify conventional groupings according to socially exogenous variables (e.g. age, income, and education), but furthermore allows us to identify groupings in the same population by means of socially endogenous linguistic variables. Both sets of groupings in the sample population are then compared in terms of composition and geographic distribution. The expectation is that these sets will show broad similarity, with greater convergence among groupings associated with greater resources (in terms of the sociodemographic variables), and greater divergence among groupings associated with lesser resources. To be clear, it is well-established that language and social structure are interrelated, maintained, and reproduced through processes of socialisation (Ochs and Schieffelin, 2017 ; Schieffelin and Ochs, 1986). We are aware of this as researchers, but moreover as individuals perfused with language in society. In this manner, that awareness is built up from the cumulation of micro-level studies and experiences. This method is intended to expand that awareness by demonstrating the presence of the interrelationship of language and society empirically at scale. Furthermore, it should be noted that it is established practice for corpora to be supplemented with paratextual metadata (e.g. annotations for part of speech or segmentation) and contextual metadata (e.g. the incorporation of metadata to characterise the setting and collection of a text or to offer a sociodemographic profile of the people that produced the text ; cf. Love *et al.*, 2017, esp. § 3.3). Such supplementation serves to thicken the observations that can be gleaned from a corpus. The reason for the use here of the distinguishing neologism 'enriched corpus' is because this approach is somewhat different—the extratextual data (i.e. the composited sociodemographic and geographic data) is not metadata. That is, the extratextual data in this corpus does not characterise any specific passage or user–document. The discursive, sociodemographic, and geographic subsets of the enriched corpus perhaps are better understood as separate datasets, in that they each address different domains of data. The only functional linkage between them in this corpus is the textual–spatial correlation for each user–document established in the preparatory step of nominal localisation (to be described in **§ 5.1.2.1**). Thus one dataset does not serve to thicken the observation of another ; rather, the goal of the enriched corpus is to enable the datasets to be approached in a cross-analytical manner to reveal what joint patterns might exist.

195. The method was developed specifically for the purposes of this work ; to the author's knowledge, there is no established method for such cross-analytical work.[157] Whereas among the social sciences it is common to understand and to analyse the actions of social groupings from the starting point of patterns of sociodemographic variables,[158] it is arguably less common to do such from the starting point of patterns of linguistic variation—despite the fact that we all do this constantly throughout our daily lives. While variationist approaches to sociolinguistics have done much work in regard to studying variation across social contexts,[159] adapting such approaches to large-scale research is problematic (**chp. 4 § 1.1.2.2**). However, a solution is to be found in the large-scale analysis of variation across genres of text. As discussed in **chp. 1 § 1.7**, the method developed for this work is an adaption of Douglas Biber's multi-dimensional analysis (MDA) approach to linguistic variation, which is a well-developed and step-wise approach to the study of multi-feature variation in text (Berber Sardinha and Veirano Pinto, 2019). The MDA approach was introduced to study genres of text at scale in order to identify groups of features ('dimensions') that distinguish genres from one another.[160] As genres are understood as text types[161] associated with recurrent contexts and functions within contexts, it was reasoned that the social groupings this work seeks out could be understood in a similar manner. Moreover, as the large-scale study of variation in genre aims to differentiate among them according to patterns of variation—the central idea and empirical support being that recurrent contexts produce stable patterns of variation (cf. Passonneau *et al.*, 2014, p. 572)—it was decided to use the approach as a template for developing the method for this work.[162]*

196. Following Biber (1988, p. 64), the steps of the MDA approach are as follows:

1. Preliminary Analyses
   a. Identify features of interest
   b. Collect a sufficiently broad sample of genre-labelled texts,[163] and convert to a suitable format
   c. Count the features of interest in each text
2. Factor Analysis
   a. Cluster the features that co-occur with high frequency into groups
   b. Interpret these 'dimensions' according to the functions of grouped features
3. Factor Scoring
   a. For each factor, compute a 'factor score' for each text
   b. Calculate an average factor score for texts comprising each genre
   c. Compare genres by their factor-score profiles
   d. Interpret the dimensions further in light of that comparison

197. In that this work seeks to identify potential social groupings by recurrent linguistic patterns in discourse, the MDA approach—with necessary modifications—is a good template for the method. That Biber's 1988 MDA approach must be adapted is due to the nature of the social phenomena in question (i.e. social discourse observed 'in the wild' at the site of research, that being social media posts) and the overall goals of the work (i.e. tracing social groupings by linguistic variables). That second point is important—the 'necessary modifications' stem from this method being something of an inversion of MDA. Rather than working from certain recognised groupings (i.e. genres) of text to identify linguistic differentials that can distinguish among them, this method works from observed differentials in linguistic patterns to trace potential (social) groupings. In short, MDA works from genres to patterns of variation, and this method works from patterns of variation to social groups. To be clear, this is not to say that Biber's 1988 method is some sort of classifying algorithm—far from that, it identifies grouped patterns of variation that cut across the 'semi-exogenous' labels of genre. Furthermore, the MDA method generally has since been extended and adapted productively into social inquiry (Berber Sardinha, 2019). The characterisation presented here is made solely in relation to Biber's 1988 method, and to stress that this work could not rely on any initial categorisation or labelling of the 'groups' to be expected in the data.

198. Before detailing the procedures used in the method here, it is important to note where this method departs from MDA. Most importantly, this work is not concerned with 'dimensions' of variation. Biber's dimensional approach was developed with an eye for "particular situational or functional parameters" (1988, p. 9) used to study the relations among texts. While a dimensional framing of variation is conceptually appropriate for the study of genre and register coming from known or assumed contexts, there is little foundation for it when studying patterns of variation observed in unknown or uncertain contexts.[164] Nevertheless, we do expect endogenous markers of a given context to cluster together, although not with a regularity across contexts that would suggest dimensions of any sort. That this work samples its data from the 'uncertain contexts' comprising the site of research, rather than from corpora representative of certain contexts (cf. **n. 163**), necessitates further departures from that initial template. The modifications to MDA made here are as follows (referencing the outline above):

- Step 1a—Features of Interest (cf. **chp. 4 § 4.1.2.2**). The MDA approach focuses on grammatical features observed in more standardised forms of English. The method here also applies that focus, but supplements it with a further focus on high-frequency, low-salience lexical features—that is, function words and similar.[165]
- Step 1b—Sample of Texts. Whereas MDA operates on a sample drawn primarily from broadly representative corpora comprising relatively standard texts and contexts,[166] the method here works from a convenience sample of public discourse within the site of research comprising predominantly non-standard texts and uncertain contexts (beyond the overarching context of the site of research itself).

- Stage 2—Factor Analysis. In rejecting a dimensional framing of social variation, the method must also reject the idea of a simple structure underlying the feature data. Thus factoring is approached in a different manner, as will be explained.
- Step 3c—Compare Genres. The method does this, but not by factor scores directly. Rather, comparison of 'genres' (i.e. potential social groupings) is by means of a clustering on factor scores.
- Step 3d—Interpret Dimensions Further. As we are not concerned with dimensions, but rather with clusters themselves, the further interpretation comes from evaluating cluster assignments within a given clustering in light of the sociodemographic and geographic data that is linked to each user–document in the enriched corpus.

199. The method can be presented, with broad brush, in a format following the above:

1. Preliminary Analyses
   a. Identify features of interest (addressed in **chp. 4 § 4.1.2.2**)
   b. Collect a sufficiently broad sample texts (§ **5.1.1**), and convert to a suitable format (§ **5.1.2**)
   c. Count the features of interest in each text ; reducing the set of counted function words (§ **5.1.3**)
2. Factor Analysis
   a. Factor the feature counts and calculate factor scores (§ **5.1.4**)
   b. Understand these scores as linguistic profiles of potential[167] social groupings
3. User Clustering
   a. Cluster the user–documents according to their profiles (§ **5.1.5**)[168]
   b. Compare these linguistic clusters according to aggregate sociodemographic profiles (**chp. 6**)

This is simply a reworking of the template, presented in similar spirit—to put logical order to the procedures conducted in this work so that the reader may understand and, if they so desire, emulate. From this point, however, the template is set aside. The remainder of the chapter addresses the method in itself.

200. The method has five stages, each addressed in a subsequent section:

1. collection of discursive, sociodemographic, and geographic data (§ **5.1.1**) ;
2. preparation of the discursive corpus, and enrichment with sociodemographic and geographic data (§ **5.1.2**) ;
3. linguistic analysis of the enriched corpus (§ **5.1.3**) ;

4. factor analysis of those results and scoring of user–documents (§ **5.1.4**) ; and finally
5. clustering of the user–documents, by linguistic measures and by sociodemographic measures (§ **5.1.5**).

201. The general process is as follows. First, a broad swath of public discourse is collected from Twitter by matching posts against a small set of politically and geographically oriented keywords. In addition, sociodemographic and geographic data of officially recognised municipalities[169] in Michigan are collected from the US Census Bureau and the US Geological Survey, respectively. Second, using Twitter metadata, the discursive dataset is reduced so that collected posts fall within the time period in question (i.e. the second half of 2018) and so that the accounts present in the dataset are nominally localised to a single, recognised municipality within the state of Michigan. The resultant dataset is not a random collection of posts but rather can be thought of as a type of corpus, wherein the constituent 'documents' are comprised of text attributable to a single account associated with a single place within the site of research. That discursive corpus is enriched by linking relevant sociodemographic and geographic data to the constituent account documents (i.e. 'users') according to municipality. Third, lexicogrammatical analysis is performed on the textual component of the user–documents. The features of interest in each user–document are identified and tallied by frequency of occurrence. Fourth, the raw frequency counts are normalised by user–document length. For each set of features of interest, these scores are factored and a score is calculated for each factor, for each user–document. The corpus is enriched further by linking these factor scores to each user–document. Fifth, users are clustered according to the sociodemographic data of their municipalities, and according to their lexicogrammatical factor scores. The final enrichment of the corpus links the resultant cluster assignments to the user–documents, replacing the factor score data which are no longer needed. This final enriched corpus is then passed on for evaluation (**chp. 6**), where clusterings and clusters are compared in terms of overall composition and in terms of geographic distribution, thus providing evidence with which to answer the research questions. It should be noted that there is other work engaging with variation observed in Twitter discourses at a various scales that draws explicitly on MDA (Clarke, 2019 ; Clarke and Grieve, 2019) or that can be understood as in that tradition (e.g. Grieve, Nini and Guo, 2018 ; Louf et al., 2022). At the same time, it should also be noted that those works remain relatively focused on variation as the pivot of analysis or inference, whereas this work attends to variation as one component of a cross-analytical approach as described above in ¶ **194**.

202. Two key aspects of the enriched corpus are the structure and the character of the data links. First of all, while the discursive component of this work is sourced from Twitter, the basic 'unit' of discourse is not a single tweet as one might assume. Rather the corpus is structured around the users represented in it (that is, the users are the constituent documents) with the sum of their publicly available tweets providing the flow of text understood as reflecting their contexts. As noted, each 'user–document' is linked to a single place within the site of research by means of a technique referred to here as nominal localisation—in essence, people state where they are from, and we take them at their word. The sociodemographic and geographic data are linked to users with this technique, but first they must be linked to each other. That is a more straightforward task, in that these data can be linked by means of unique identifiers intended for that purpose. The localisation and linking techniques are described in **§ 5.1.2.1**.

203. The final enriched corpus prepared for this work comprises 37 million words across 2.6 million tweets associated with 5,889 users across 417 distinct places. Each user is associated with only one place. Each place is linked to an area geolocation (that being its statutory boundaries) and a sociodemographic profile including population percentage distributions across brackets of age, income, and education, and proxies of ethnicity. The stages of the method are now addressed.

5.1.1. Data Collection

204. The method requires a composite dataset, linking discursive, sociodemographic, and geographic data. The production of linked data is resource-intensive, and thus such data is often proprietary and costly. That situation is a challenge for social research on large-scale phenomena, such as this work. In response to the situation, it was decided that the method would rely on publicly (i.e. freely) accessible data and that the necessary linking would be integrated into the method. Furthermore, as this method implicates large quantities of data, that data would be accessed programmatically via publicly accessible APIs.[170]★ Aside from the cost advantage, the aim was to demonstrate a lexical–

geographical linking technique for large-scale data that could be adapted for use by other researchers. The remainder of this section describes the data sources, the data to be collected, and the collection process ; the linking technique is described in **§ 5.1.2**.

5.1.1.1. Discursive Data

205. The Twitter messaging service (Twitter hereafter) was the source of all discursive data. During the period of data collection in 2018 and early 2019, Twitter provided free access to a selection of various endpoints of its API v1.1 (hereafter v1.1).[171] This method used the 'streaming' and 'timeline' endpoints, both of which return a small sample or portion of the real-time stream of messages (i.e. 'tweets') on Twitter. The streaming endpoint affords the filtering of the sample by keywords ; when such a call is made repeatedly to the API, this affordance allows the tracking of specified keywords through time.[172]★ The timeline endpoint returns a specific portion of the stream, that being the 3,200 most recent tweets of the users specified in the API call.[173] These two endpoints were used in conjunction, as will be explained, and both were accessed using DMI-TCAT (Borra and Rieder, 2014).[174]

206. The collection of discursive data was a three-step process: first collect a broad swath of discourse by means of a small set of general keywords relevant to the political context, within the posts so collected identify accounts meeting certain criteria (most importantly of timestamp and location), and then collect all available posts from the accounts so identified. The rationale of this three-step process was to minimise researcher bias in selection (of keywords, accounts, etc.), to avoid potential yet significant problems with direct selection of discourse beyond researcher bias,[175] and also to obtain discourse of politically engaged persons as opposed to manifestly (i.e. lexically) political discourse in itself (recall that we are interested in social contexts, not party political sparring). A further consideration supporting this 'trawling' technique of the primary collection in Step 1 is that, at the level of free access, there is no ready way to bound a search or stream in terms of location. In v1.1, it is possible to specify a 'bounding box' using longitude and latitude to filter the returned sample, but a bounding does not filter other parameters (e.g. keywords) and would only return geocoded posts.[176] As location is critical to the method, this posed a significant problem. In the end it was decided that the best way forward was to incorporate geographical terms into the set of keywords, as will be explained in the following section. Each step will now briefly be described.

207. Step 1, the primary collection, was a broad 'trawling' of the streaming endpoint of the Twitter API using a small set of politically and geographically oriented keywords. As noted, the use of a small set of general keywords was chosen to minimise researcher bias in the primary collection, but also it was to maximise the quantity of data collected. Not that bigger is necessarily better—rather it was assumed that the bulk of the primary collection would be irrelevant to the project, and quantity is, unfortunately yet practically, a ready means of offsetting a high proportion of waste. The politically oriented keywords were chosen to reflect the fact of an upcoming election, the major offices up for election and the common abbreviations of title, and the site of research (i.e. a specific state, rather than the entire United States). The political keywords were: `district`, `elect`, `election`, `gov`, `governor`, `rep`, `representative`, `sen`, `senate`, `senator`, and `vote`. The geographic keyword was: `michigan`. (Note that the keyword filtering affordance of the streaming endpoint in v1.1 matches whole words only, is case insensitive, and ignores punctuation.) Abbreviations of the three major office titles (`gov`, `rep`, and `sen`) were included as this would be the expected form used to refer to an office holder by name (just as it is expected to see Mr. Spock and Mrs. Robinson, and not otherwise). Although the abbreviations as whole words have uses in other contexts, they would not be expected to appear with high frequency in comparison to the use as form of address. It was decided that the term `michigan` would suffice as the sole geographical term. The common abbreviations 'mi' and 'mich' are high-frequency words in other languages, and common state nicknames—such as the Great Lakes State, the Wolverine State, or the Mitten State (NB: Michigan looks like a mitten)—were judged unworkable.[177] Primary collection commenced on 1 July 2018 and proceeded until 1 December 2018 exclusive. The streaming endpoint was queried on an ongoing basis, filtered by the specified keywords. The political keywords were queried with a single API call, and the geographic keyword was simultaneously queried in a separate call.[178] The separate queries allowed the resultant matches to be sorted easily into separate data tables.[179] Over that time period,[180] 2–3 million tweets were collected per day. The primary collection took in a total of 154 GB of data, comprising 300–400 million tweets and their metadata.[181]

208. Note that any given post that one sees on Twitter is a 'tweet object' having a range of attributes. At the time of collection, calls to the streaming endpoint of v1.1 returned data that DMI-TCAT represented as an object having 36 attributes, combining elements of Twitter's 'tweet object' and 'user object'.[182] The most important attributes to this method are:[183]

- `text`, which is the content of the tweet, represented by a UTF-8 string, and which can include URLs linking to other content ;
- `id`, which is the unique identifier of the tweet, represented by a large integer ;[184]*
- `created_at`, which is the UTC timestamp of the tweet ;
- `from_user_id`, which is the unique identifier of the user account, represented also by a large integer ;
- `retweet_id`, which is the tweet identifier of a retweeted post ; and
- `location`, which is a string of up to 64 characters in which the user may opt to provide a nominal location (or whatever they wish, really).[185]

Note that `from_user_id` and `location` are user attributes, not tweet attributes.

209. Step 2, account selection, relied on the assumption that there would be accounts (identified by `from_user_id`) that were represented in both the political and geographic results.[186] First of all, the primary collection (as dumped in stages from the DMI-TCAT backend) was reassembled in a clean database.[187]* The primary collection was filtered by its metadata, specifically the fields relating to location (as specified by the user) and to the language setting of the operating system from which the tweet was posted. In the first case, the method relies on the very American habit of presenting locations as postal addresses, that is, as [Place], [State].[188] Working directly with the data (i.e. by means of a database client) using a combination of SQL (structured query language) and regular expressions,[189]* the primary collection was filtered to exclude tweets with `location` metadata that did not include `Michigan` or `MI` preceded by a comma and space as would be expected in a postal address.[190] In the second case, reasoning that Michiganders are more likely to have the operating systems of their computers and gadgets set to English, the primary collection was filtered to exclude tweets not having `en` (for English) specified in the system-language field.[191]* At this point, it would have been wise to filter accounts by the date of their creation (excluding, for example, accounts created only in 2018), but that was not done until after the secondary collection. The filtering resulted in a working dataset of just under 2 million tweets.

210. As noted, the primary collection was assembled in two data tables: one for the geographical keyword, and one for the political keywords. Using the `created_at` timestamp for each tweet, these tables were split by calendar week of the primary collection period, yielding 23 splits in each table.[192] The geographic and political data tables were then joined week by week. Users in <u>both</u> tables in a given week were added to a target list for secondary collection. The joining process yielded 55,630 distinct users. However, it was observed in the returned tweets that a large proportion was retweets, and that this proportion was dominated by retweets of users with a high follower count.[193] As this method seeks to target <u>discourse</u>, it was decided to exclude accounts that predominantly posted either retweets or original tweets. The logic is that neither group is properly part of public discourse. Moreover, the former group is not contributing its own linguistic signatures but rather reproducing those of others, and the latter group on inspection seemed to tend towards corporate or media-facing accounts (cf. **n. 193**). The balance of original tweets (including quote tweets) and retweets is an important consideration—close inspection of samples of the data during pilot testing suggested that filtering in this regard could focus the dataset towards accounts that had characteristics indicative of the average person, as opposed to corporate, collective, or patently fictitious accounts. To focus on those accounts that were 'part of the conversation', an 'RT/OT' ratio of **R**e**T**weets to **O**riginal **T**weets (including quote tweets) was calculated for each user. Those in the top and bottom quintiles (i.e. those having an RT/OT ratio greater than 80% or lower than 20%) were excluded, resulting in a target list of 20,598 accounts for secondary collection.[194]*

211. Secondary collection was performed in January 2019 using a DMI-TCAT script for querying the timeline endpoint of v1.1.[195]* The result was a working dataset of some 46 million tweets, which was filtered to removed accounts created after 1 January 2018, and to keep only those tweets with a timestamp between 1 July 2018 and 31 December 2018 inclusive. Tweets with a non-`NULL` `retweet_id` (i.e. retweets) were also removed. The resultant dataset, which is refined in the following stage of the method, comprised nearly 17 million tweets by 20,000 users across some 3,400 locations. Beyond this point, DMI-TCAT was no longer used.

5.1.1.2. Other Data

212. The discursive corpus is enriched in the next stage with sociodemographic and geographic data. The collection of that data was a straightforward matter, as will be described in a moment. However, as noted at the beginning of this chapter, the sociodemographic and geographic data are to be linked to the users represented in the corpus by means of 'nominal localisation', and furthermore must be linked to each other. Nominal localisation hinges on a simple thing: where people say that they live. Where is referred to by the term 'place', which is a formal term in the context of this data,[196] but which can be understood to denote a concentration of population with which people associate both a fixed name <u>and</u> their daily lives—more or less, it means home. However, not any statement of place can be accepted. All <u>officially recognised</u> places have a standardised name that is used across all federal bodies (thanks to the US Board on Geographic Names) and are assigned a unique geographic identifier (GEOID). Unfortunately, GEOIDs are not standardised, as there are a number of systems used for this purpose. The sociodemographic data used in the method is represented by FIPS codes and the geographic data by GNIS codes.[197] However, in that GEOIDs are nevertheless unique, it is possible to translate between systems and so <u>link</u> otherwise distinct datasets. To perform these steps, three types of data are required: 1) sociodemographic data profiling the places within the site of research, 2) geographic data detailing the location and extent of those places, 3) translation tables for converting between code sets.

213. Note that "places within the site of research" means <u>all of them</u>, not simply those represented by the users in the corpus. The rationale is that nominal localisation seeks to link personal statements of place with codified representations of officially recognised places. Collecting data for all official places—including official names and identifiers—functionally provides a gazetteer of the places within the site of research, thus a list of which statements of place to accept as valid ; statements of place not found on the list are rejected. In this manner, the final enriched corpus only represents users associated with officially recognised places.

214. Sociodemographic data was sourced from the 2018 5-year estimates of the American Community Survey (ACS), conducted by the US Census Bureau.[198] This data was obtained by querying the ACS API by means of an R script,[199]* requesting the following categories of data for places in Michigan. For each place, data was collected on distributions of age, income, education, ethnicity, and origin, as well as total population.[200] Queries to the API automatically return the relevant FIPS codes.

215. Geographic data was primarily sourced from the US Geological Survey. No API was needed for this collection, as the US Board on Geographic Names provides publicly available files for download.[201] The file listing all official places of Michigan was obtained. The fields of interest were the place name, the coordinates (specified point-wise by longitude and latitude), and both the GNIS and FIPS codes ; the set of places was extensive,[202] but it provided the baseline gazetteer to support the localisation process in the next stage.[203] Note that the downloaded dataset contains GEOIDs represented in both needed codesets (i.e. GNIS and FIPS), so a translation table was easily constructed from this one dataset.

5.1.2. Corpus Preparation and Enrichment

216. The second stage of the method involves the preparation of the discursive corpus, and enrichment of that corpus with sociodemographic and geographic data. The various procedures of this stage are concerned with structure—of the overall dataset, and of the data itself.[204]* While the previous stage was crucial for obtaining data that could satisfy the temporal bounding and political contextualisation of the site of research, the collected data would nevertheless be useless unless attributable to the site of research with confidence and formatted in a manner to allow computational processing. The first requirement concerns validity and the second practicality, as discussed in **chp. 1 § 1.7**. The validity requirement is addressed using the 'nominal localisation' process, which associates each user with a distinct place in the site of research. The warranted localisation of users serves as the foundation of the overall structure of the dataset, in that the subsets of data (discursive, sociodemographic, and geographic) are all interlinked by means of user locations. The practicality requirement is addressed by structuring those subsets by:

- ‘cleaning’[205] the discursive data to prepare for linguistic analysis using statistical methods ;
- assembling those millions of tweets and metadata into separate documents, each associated with a single user[206] ;
- linking the sociodemographic and geographic subsets[207⋆] of data to each other to prepare for incorporation into the discursive corpus ; and
- ‘enriching’ the discursive corpus by linking its constituent user–documents to the appropriate subset of linked sociodemographic and geographic data, thus yielding a coherent dataset (i.e. the enriched corpus) for the next stages of the method.

217. Note that to ‘enrich’ has a double sense in this method. In the first sense, the discursive corpus is enriched by concentrating its constituent data—the localisation process removes data of dubious validity and the structuring process removes or remediates computationally problematic data,[208⋆] leaving data in which we have increased confidence in both validity and computability. In the second sense, the discursive corpus is enriched by expanding its constituent data—both localisation and structuring processes allow sociodemographic and geographic data to be associated with the users represented in the corpus, thus enlarging the scope of possible analysis. Moreover, the associated data is incorporated directly into the corpus itself, making such analysis relatively straightforward. The localisation and structuring are now described in turn.

5.1.2.1. Nominal Localisation

218. As introduced in **chp. 1 § 1.7**, nominal localisation is a process for associating a person with a location according to their statements. In the case of this work, the method matches users represented in the discursive corpus with officially codified places within the site of research. At the point that this process was performed, users present in the discursive corpus had already been roughly filtered according to location by means of a simple regular expression. However, that filtering was imperfect,[209⋆] and moreover the method requires localisation to the level of municipality.

219. The process in general has three necessary components: 1) readily available and comparable statements of location, 2) a list of valid locations, and 3) a method for matching statements of location to valid locations. For this work, the components are as follows. The first component is provided by the `location` metadata of the Twitter user object, which is embedded in every status update on the platform (i.e. every tweet). The second component is provided by the gazetteer of codified places in Michigan derived from US Census Bureau and US Geological Survey data, as noted in the previous section. The third component was a set of criteria that statements had to meet, firstly in relation to themselves, and secondly in relation to the list of valid locations. The collection of the first two components has been described. The third component is now described.

220. The decision rules that comprise the geocoding of the nominal localisation process are extensive.[210] A full description would not advance the discussion here, but are few points need to be mentioned:

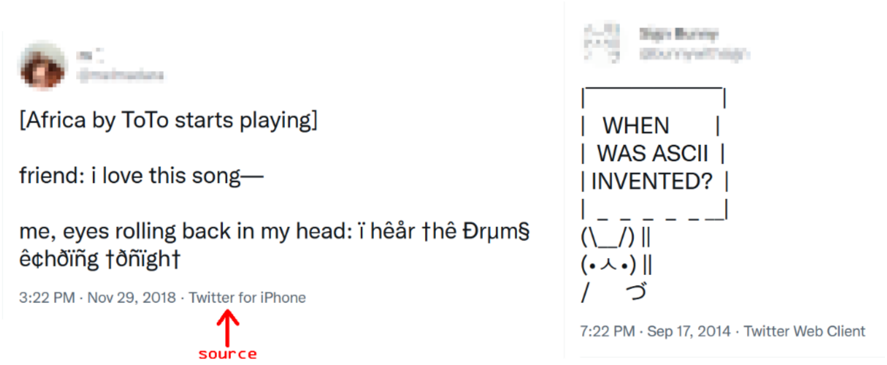
- Component 1, statements of place, was sourced from the second collection. However, it was decided first to filter that dataset further so that it comprised only those users that had a unique value in the location metadata, that is, it was not modified at any point in time during the period covered by the second collection (i.e. the second half of 2018).[211★] Regular expressions were used to isolate the `[Place], [State]` cluster of interest, and to extract `[Place]`. In cases with more than one such cluster (i.e. two or more locations stated), assignment was made according to the first cluster.[212]
- Component 2, the list of valid locations, was sourced as noted above, filtering for those locations with a unique codified name.
- Components 1 and 2 were modified using regular expressions to update the respective data tables so that common variants (e.g. Sainte as Ste, Mount as Mt, and of course Michigan as MI) were put in full-length form.
- Component 1 was cleaned, also by means of regular expressions, in a manner common to text pre-processing (e.g. lower casing ; replacing double, leading, and trailing spaces ; removing non-ASCII characters ; stripping digits and punctuation, etc.).[213★]

The above points describe much of Component 3, which in the main was applied programmatically to Components 1 and 2. Once the commonality of the location data in the source components was assured, the final step of Component 3 was to join the data tables storing the source components on the `location` field, and to insert the output into a new data table associating users and their stated locations with a codified place name and GEOID (at this point, a FIPS code). Of the approximately 20,000 users in the second collection, some 17,000 were matched with a determinate, codified place name and GEOID.[214,215★]

5.1.2.2. Data Structuring

221. As discussed, nominal localisation provides the overall structuring of the dataset by means of associating each user in the discursive corpus with a single codified place within the site of research. Thus the sociodemographic and geographic subsets of data, which are themselves structured by codified place, can be linked determinately to each user. That is the basic structure that will be applied to the enriched corpus. But first it must be enriched, and to do so the subsets of data must themselves be structured. As noted at the beginning of this section, that structuring is performed by 1) cleaning of the discursive data, 2) assembly of the separate tweets and metadata into separate 'user–documents' thus yielding a discursive corpus, 3) linking of the sociodemographic and geographic data subsets, and 4) enrichment of the discursive corpus by incorporation of the sociodemographic and geographic data according to the codified place assigned to each user–document. It should be noted that these are not truly distinct procedures, in that each has overlapping elements of cleaning and linking—the presentation here is rather a logical presentation of the process for reasons of clarity (cf. **n. 207**). Each logical step is now briefly addressed.

222. Cleaning of the discursive data was performed before assembly of the user–documents, as users had to be represented by a sufficient quantity of both tweets and total text to be included in the corpus and a high degree of loss was expected at this stage. Twitter content is highly multimodal. Even if multimedia content is not considered,[216] users find a variety of affordances in the interface (such as its dimensionality and Unicode-compliance) that allow expression beyond plain text (see **fig. 1**).[217★] Textual multimodality poses significant challenges to computational work, such as in the isolation of 'target' text and the removal of potentially problematic characters such as control



[Africa by ToTo starts playing]

friend: i love this song—

me, eyes rolling back in my head: ï hêår †hê Ðrµm§ ê¢hðïñg †ðñïgh†

3:22 PM · Nov 29, 2018 · Twitter for iPhone

source

```
|‾‾‾‾‾‾‾‾‾‾‾|
|  WHEN      |
|  WAS ASCII |
| INVENTED?  |
|_ _ _ _ _ __|
(\__/) ||
(•ㅅ•) ||
/    づ
```

7:22 PM · Sep 17, 2014 · Twitter Web Client

Figure 1. Multimodality of Twitter Posts

codes. Some of these (e.g. the line feed control code) can be modified in place, but much has to be removed. In this method, the default position was to err on the side of caution and remove the entirety of tweets and their metadata. (A benefit of large-scale work is that selective data removal has negligible impact on the overall dataset.) Key initial cleaning steps included:

- Removal or substitution of all control codes, whitespace characters, joiners, etc.[218]★ ;
- Removal of all tweets containing URLs[219] ;
- Removal of non-ASCII text[220]★ ;
- Standardisation of certain typographical variants (e.g. slant quotes replaced with straight quotes) ;
- Removal of all tweets with text suggesting a retweet (e.g. `RT`)[221]★ ;
- Management of remaining whitespace via substitution and trimming ; and
- Removal of extended quotes (i.e. reported speech).[222]

223. As expected, there was significant reduction from the initial size of the secondary collection (cf. **n. 219**). Numerous other cleaning steps were taken as needed, and were repeated at various stages of the work.[223]★ It was all rather routine and contingent on the content of the dataset. The central motivation of this stage was normalising the text as much as possible at character and tweet level, but not at word level (except as noted in regard to slugging ; see § **5.1.3.2**). This is a departure from common practice in computational approaches to text in the social sciences, which tend to follow procedures developed in information retrieval (cf. Nguyen *et al.*, 2016 ; also **chp. 4 § 4.1.2**).[224]★ Procedures of this sort include lowercasing, lemmatisation, punctuation removal, segmentation of contractions,[225]★ stemming, and stoplists.[226]★ None of these were undertaken. Such practices emerged, on the one hand, to enable text processing at a time when processing and storage resources were more rare and expensive and, on the other hand, to encourage a focus on textual salience (i.e. the information to be retrieved). Such procedures are anathema to this method, as they erase contextualising features of language—in the case of stoplists, the procedure would obliterate a class of words that this method seeks to measure.

224. Also, it should be mentioned that at this stage there was further filtering of the secondary collection according to metadata.[227] Three fields ended up being key to further refining the secondary collection in terms of validity and quality: `source`, `lang`, and of course `location`. `Source` indicates the application used to post a given tweet, in the form of an HTML `<a>` tag—the display text of that tag appears in every tweet (see **fig. 1**).[228] It was found that there were 799 distinct sources in the secondary collection. The bulk of these sources were not Twitter interfaces, but were other social media platforms, websites, businesses, marketing applications, etc. That is to say, `source` provided a spam filter. It was decided to keep only those tweets originating from official Twitter sources,[229] except for those that were known to be explicitly for marketing (Twitter Media Studio and Twitter Ads Composer) and also TweetDeck, just to be sure. Out of 799 distinct sources, only 11 were retained. Tweets from all other sources were deleted. This was a significant step in assembling a <u>discursive</u> corpus—as was the application of the RT/OT ratio (**¶ 210**)—rather than simply a large sample of the cacophony of Twitter.

225. The `lang` data required closer inspection, as the automatic tagging performed by Twitter is erratic and unreliable (see **n. 191**). There were 43 distinct codes in the `lang` metadata (42 languages plus 'undetermined'). On inspection, it was observed that the automatic tagger reliably identified non-Latin scripts as 'not English' (whether the actual language assignments were correct was not verified), but that otherwise it was misled by simple things such as surnames, exclamations, city names, etc. It is thick as a brick. Of the 43 distinct `lang` codes, tweets matching 11 non-Latin-script languages were removed. Beyond that, no further effort was made to refine the languages present in the dataset, as non-English discourse would effectively be ignored during the analytical procedures given its relatively minimal presence.[230]

226. Refining by `location` was very much work 'in the weeds'. The nominal localisation process was apparently effective, but not perfect, yielding many false positives and false negatives. On inspection, the failings of the procedure were due largely to the endless variety and messiness encountered in `location` metadata. It was decided that the best way to handle the situation was to iteratively filter distinct locations in the dataset manually using regular expressions. Direct engagement with the data allowed that filtering to be guided by observation. Eventually the number of distinct locations was whittled down to valid single locations expressed in an expected format. (For this remainder, the localisation procedure was found to have performed well.[231]) Those users and their tweets were retained, and the rest discarded.

227. There is little need to dwell on assembly of the corpus. In that each tweet in the dataset is associated with a single user by means of a unique identifier, it was a straightforward matter to output the text of each user's tweets, concatenate them, and store them in an individual text file. Tweets were output in order from oldest to newest, and concatenation used line breaks to separate the text of tweets. Thus each user–document is multi-paragraph document of varying size. As noted above, there were several criteria that a user (i.e. a user–document) had to meet to warrant inclusion. Most of these were operant at the tweet level, and were just addressed. The final criterion was that user–documents had to comprise at least 100 words total[232]—in order to be suitable for the analytical procedures (Biber and Jones, 2005, p. 158)—extracted from at least 10 tweets, to promote contextualisation over time rather than of a moment in time.[233★,234]

228. There is also little reason to discuss data linking of the socioeconomic and geographic data. Effectively all linking was performed at the moment that the localisation script was run in order to assign a codified location and GEOID. The fact of having a FIPS–GNIS translation table derived at the stage of data collection meant that the data were already linkable as needed.[235★]

229. That said, it is important to note the broader linking move, which is the enrichment of the discursive corpus. All the data needed for the method is already available, except it is distributed across an array of databases, tables within those databases, and thousands of text files. Work can nevertheless be done with such distributed data, as any given process can call upon those resources and cobble them together as needed. This is essentially how organisational data infrastructures operate in order to cope with their scale and distributed nature. Such a mode of operation is, however, inconvenient. Moreover, scale introduces new sources of potential error, increasing complexity and risk, whether at the scale of a multi-national or of an individual research project. It is a fraught situation. For the purposes of this work and method, complexity and risk are ameliorated by integrating these data sources. We have already linked subsets of the data (e.g. tweets and users assembled into user–documents, nominal localisation of user–documents, geolocalisation of sociodemographic profiles, and derivation of the GEOID translation table). Now, we are able to assemble those subsets into a larger resource that can be used by itself for data exploration and analysis. This is the enrichment of the discursive corpus with the geolocalised sociodemographic data.

230. In that both users and sociodemographic profiles share a common determinate attribute in the FIPS code, these separate datasets are now combined into a common resource where rows (i.e. the constituent data elements, in this case the user–document) comprise users and their metadata, their discursive document, their location within the site of research and its coordinates, and a sociodemographic profile of that location. Having this common resource greatly simplifies the more complicated steps of the remainder of the method, and reduces sources of potential error (i.e. the points at which the researcher must intervene in a given procedure). The larger goal is to continue with enrichment at each remaining stage of the method. That is, the results of the linguistic analysis of the enriched corpus will themselves be incorporated, thus enriching it further. The same will be done with the results of the clustering procedure. The end goal is to have a single resource to which the means of evaluation will be applied. As noted at the beginning of the chapter, the enriched corpus comprises 37,134,978 words associated with 5,889 users.[236]

231. How representative is the corpus is difficult to gauge, but representativity was encouraged through the use of a small set of keywords in the 'trawling' procedure of the primary collection, resulting in the sheer scale of that procedure (**§ 5.1.1.1**). Thus, although the corpus is composed predominantly of non-standard texts due to the nature of its source, it is expected to have very broad coverage in terms of contexts within the overarching context of the site of research. Analysis of results (**chp. 6**) support that expectation.

5.1.3. Linguistic Analysis

232. Whereas the first two stages of the method are somewhat ad hoc, for the simple reason that (to the author's knowledge) there is no established process for assembling and localising an enriched discursive corpus from social media content, the same cannot be said for the analytical stage. This stage adheres closely to the MDA approach, described in Step 1c (¶ **196**) as 'count the features of interest'. This is good salesmanship, as counting is the trivial aspect of the task. First you have to find what you wish to count! The reader likely recalls from school that breaking down a given sentence into its grammatical components takes some time and thought, and is not always a clear-cut task. The original approach taken by Biber measured 67 grammatical features (1988, p. 72). Some of these are simple (e.g. 'pronoun'), some are essentially compound (e.g. 'attributive adjective'), and some are rather more complicated (e.g. 'present participial clause'). That work relied on an extensive list of decision rules, based on parts of speech, to identify the features in question (1988, app. II). As the corpora sampled were untagged for parts of speech, it was first necessary to tag them. Given the quantity of text, this was performed programmatically.[237]* Implementation of the decision rules—that is, feature tagging—was partly programmatic, and partly by hand. Once that tagging was complete, all that remained was to take the frequency counts, being sure the normalise to a given length (1,000 words in the 1988 study) so that texts of varying length were comparable. After that you were off to the factor analyses.

233. The method here calls for a replication of those procedures on the enriched corpus, using the same set of linguistics features and the same set of decision rules, as will be described in § **5.1.3.1**. In addition, the method calls for similar procedures adapted to lexical features, specifically function words and similar, as will be described in § **5.1.3.2**. The sets of procedures were kept separate, as integrating them would have been problematic in theoretical and technical terms. The results of both sets of procedures are then passed on to the factoring stage. First, details of the feature tagging are given.

5.1.3.1. Grammatical Features

234. As noted above, this stage of the method seeks to reproduce the analytical procedures used in Biber (1988). While the decision rules for identifying features are laid out in extensive detail in that work, the tagger that Biber developed and used is not publicly available.[238]★ However, Andrea Nini (2019) has developed a reproduction of Biber's tagger. It is based on the Stanford tagger (Toutanova *et al.*, 2003) for initial tagging, and then implements the decision rules to tag all 67 original features. Nini (2019) demonstrates that the Multidimensional Analysis Tagger (MAT) can replicate the analyses in Biber (1988). Moreover, MAT has been packaged into a freely available application that, given a text file as input (or batched by directory), will handle all feature tagging and produce a report of normalised counts.[239]★ The method here used MAT, which provided a CSV output report that was easily incorporated into the working database. This data was then passed on to the factor analysis stage.

235. Note that MAT was not expected to work well on the discursive corpus because of its use of the Stanford tagger, which was trained on data from the *Wall Street Journal* (Toutanova *et al.*, 2003, p. 176). In a shocking turn, there is evidence that the Stanford tagger works poorly with Twitter data (Finin *et al.*, 2010). Nevertheless, Twitter is not comprised of solely non-standard speech. Direct inspection of the discursive corpus revealed a goodly portion of 'standard' forms. Moreover pilot testing with MAT (which can also produce evaluations of Biber's dimensions) showed solid indications of reportage and similar text present in the corpus.[240] The hope was that MAT would handle such user-documents with aplomb, but would also mishandle everything else in an equitably poor manner—thus leaving a strong signal for certain types of text, and noise for the remainder.[241]★ Whether that hope was borne out is a topic for later investigation.[242] Furthermore, it was important to use MAT as it aims to replicate the 'canonical form' of MDA—Nini notes, and demonstrates, that application of canonical MDA to new datasets allows comparison to a baseline (2019, p. 70). And so that is what was done.

5.1.3.2. Lexical Features

NOTE: The following describes a generalisation of the operationalisation of function words as proposed in **chp. 4 § 4.1.2.2** to pervasive features. The rationale for the change is that any set of pre-selected function words implicates an exogenous perspective of standardised language, whereas evaluating the pervasiveness of features could be pursued in an endogenous manner. Thus it was decided to focus on pervasive features, as will be described. Pervasiveness was selected as a metric given the author's suspicion that the characteristics of function words that make them potentially valuable as markers of context are due more to their high frequency and low salience than to any specific grammatical roles, strictly speaking. They are the water in which we swim. The author as of yet has no empirical evidence to support that suspicion, beyond several decades as a human language user. Note that investigation of pervasive features is item 3 under suggestions for further work (**chp. 7 § 7.4**). The brief discussion of function words in **chp. 4 § 4.1.2.2** has been left in place because it is nonetheless valid and instructive.

236. In addition to reproducing the grammatical analysis performed in Biber (1988), the method also performs a lexical analysis of the discursive corpus. The aim is to examine high-frequency, low-salience lexical items as carriers of social information, as there is reason to suspect that such items might serve well as endogenous markers of context. Whereas Biber's grammatical analysis relies on part-of-speech tagging of individual words (i.e. tokens), the lexical analysis of this method relies of measuring frequency and salience of word forms (i.e. types).[243*] Many approaches to the study of language, especially statistical approaches, are concerned with sussing out 'keywords', that is, those words that stand out as key to the meaning of a text. Such salience, or 'keyness', is a driving concern especially in information retrieval-oriented approaches to language[244]—hence we search and categorise by keyword. However, keyness as it is often understood (cf. Stubbs, 2010) is not our concern here. We are very much interested in frequent words, but not in the ones that stand out. The method seeks pervasiveness, not salience. Before describing the method for analysing lexical features, we first consider how this method measures pervasiveness, and how this measure provides a specific understanding of keyness for the purposes of this work.

237. The pervasiveness of a word is a function of its frequency and dispersion. A raw frequency count does not account for dispersion, and so raw counts can be misleading if a word occurs in bursts or clumps. The issue grows more complicated if one considers comparisons across corpora (cf. Gries, 2008). Thus the method uses a measure of frequency that does take account of dispersion—'average reduced frequency', or ARF (Savický and Hlaváčová, 2002).[245]⋆ The actual procedure for calculating ARF is a bit involved, but the basic idea is that the frequency of a given word is adjusted by cutting up a text into slices, measuring the distances between occurrences of the word in each slice, and accounting for the average distance.[246]⋆ For an evenly dispersed word, ARF will more or less equal the raw frequency count ; ARF grows smaller, however, the clumpier a word is. That adjusted measure is useful, but the method still requires a <u>measure</u> of pervasiveness. That is provided by *Gamma*, a measure developed in this work for that purpose.[247] *Gamma* is composed of two terms: a dispersion factor and a relative prevalence factor. The dispersion factor is simply ARF divided by raw token frequency (so ranging from ~1 to ~0). The relative prevalence factor is ARF divided by the corpus token count (i.e. the word count).[248]⋆ *Gamma* is the common logarithm of the product of the dispersion factor and the relative prevalence factor. The equation is given below in **eq. 1**, where **ARF** is the average reduced frequency of the word in question, **f** is the absolute frequency of that word, and $\mathbf{C_t}$ is the token count of the corpus.[249]

$$\Gamma_{\text{word}} = \log_{10}\left(\frac{\text{ARF}}{\text{f}} \times \frac{\text{ARF}}{\text{C}_\text{t}}\right)$$

$$\Gamma_{\text{word}} = \log_{10}\left(\frac{\text{ARF}^2}{\text{fC}_\text{t}}\right)$$

Equation 1. *Gamma*, a Measure of Pervasiveness

238. The understanding of keyness used here is not based in salience, but rather in <u>distinctive</u> pervasiveness, which *Gamma* allows us to calculate. For any given word in the corpus, *Gamma* is calculated at the corpus and subcorpus levels. The interest is keyness at the subcorpus level, in that the method is concerned with identifying words that are distinctively pervasive in a given user–document compared to the corpus overall. Calculating keyness of a word as understood here is a relatively simple matter of taking the ratio of subcorpus *Gamma* to corpus *Gamma* (cf. Kilgarriff, 2009). How this measure is applied in the method is now described.

239. The preparatory steps of the lexical analysis are as follows:

- First the corpus is part-of-speech tagged using a tagger developed specifically to address the challenges of short-form messages (Owoputi *et al.*, 2013 ; extending Gimpel *et al.*, 2011).[250★] While the analysis is lexically oriented, the tagging provides a helpful method for filtering the corpus.[251]
- The corpus is filtered to remove all tokens tagged as punctuation, discourse marker, URL or email, and emoticon or emoji. The lexical analysis is not concerned with word order,[252] so such piecemeal removal has no impact on the analysis.[253]
- The corpus is then slugged. In this context, 'slug' denotes a placeholder.[254] For example, all numbers were replaced with `[NUM]` and all hashtags were replaced with `[HASH]`.[255] The purpose of slugging is to preserve the place and function of the tokens in question, while substantially reducing the overall type count in the corpus.[256] The following tags were replaced by respective slugs: numerals, proper nouns, hashtags, mentions, and 'junk'.[257]
- The corpus is then analysed to compute the corpus-level *Gamma* for each token.[258★]
- The user–documents are then analysed in the same manner to compute the subcorpus-level *Gamma* for each token, for each user–document. In addition, the keyness of each word (as described above) is calculated.
- The corpus is processed to prepare a bigrammatic representation of it,[259] and the corpus- and subcorpus-level calculations are repeated for bigrams. Again, keyness is calculated for all subcorpus bigrams.

240. Having obtained a calculation of keyness for all words and bigrams in the user–documents, two lists are prepared from both lexical sets.[260] List 1 comprises those items that are distinctively pervasive in user–documents relative to the entire corpus, as measured by keyness (i.e. the *Gamma* ratio) ; the list was based on the top 10 items by keyness from each user–document. List 2 comprises items that are not distinctive in user–documents ; the list was based on those items from each user document have a keyness measure between 0.9 and 1.1. However, in that we are interested in potential groupings of features (rather than individual patterns), and that the lists as just described would number in the tens of thousands of items, there is initial data reduction in the assembly of the lists. This is a separate procedure from the factor analysis that follows ; rather, it is the simple application of thresholds for inclusion. The assembly of the lists is now specified (note that the process is the same for both words and bigrams).

241. List 1 was assembled from those user–documents containing at least 1,000 items after filtering (slugging did not alter item count). That cut-off was based on a concern with keyness being 'misrepresented' by user–documents of too short a length.[261] There were 1,514 user–documents below this threshold, leaving the list to be compiled from the remaining 4,430. The top 10 items by keyness were taken from each of those user–documents (thus yielding a list of 44,300 items). That list was aggregated by item to give a frequency for each. Thus if an item appeared once in the list, its frequency was 1, if it appeared twice, then 2, and so. Note that this number indicates how many times the item appeared in a 'Top 10' list of user–document. For inclusion in the final list, an item had to be attested in the Top 10 list of at least 0.5% of those 4,430 user–documents, that is, in at least 23 user–documents. This procedure yielded a word list of 171 items, and a bigram list of 308 items.[262]

242. List 2 was assembled in nearly an identical manner to List 1. There main difference, as noted above, is that this procedure took all items with a 'normal' keyness range (0.9 to 1.1) relative to the overall corpus—that is, non-descript words. The other difference is that the inclusion threshold, which was attestation at 0.5% of non-small user–documents for List 1, was set at 1 in 16 (or 6.25%).[263]⋆ This procedure yielded a word list of 459 items, and a bigram list of 116 items. Having these lists, the counts for each user–document were tallied up and normalised for document length. This data was then passed on to the factoring stage.

5.1.4. Factor Analysis and Scoring

243. Following linguistic analysis, the MDA approach factors the resulting data. Factor analysis comprises a broad family of statistical techniques for identifying correlations among variables in multivariate data so as to represent that data using fewer variables.[264]⋆ Not only does such 'data reduction' make analysis easier in general, simply because there are fewer variables to consider, a more significant analytical advantage is that the grouping (or clustering, same thing) of correlated variables can give indications of a deeper structure in the data. Such latent (i.e. unobserved, perhaps unobservable) or simple structures can go by many names, depending on the field of work—in MDA the simpler structure underlying the linguistic data are the dimensions of variation.

244. As noted at the beginning of the chapter, this work does not understand social variation from a dimensional perspective. While there are sensible reasons to argue for a dimensional perspective, this work resists on principle in that dimensional framings of social phenomena hew uncomfortably close to structural, exogenous approaches. Nevertheless, the perspective is not needed ; one may still avail of the practical benefits of data reduction in terms of easing the analytical (and likewise computational) burden.[265★] For that reason, and as well for reasons of comparability with MDA work proper, this method adheres to the MDA template in conducting a factor analysis of the linguistic data. The data from the grammatical analysis was subjected to Principle Axis Factoring, in much the same way as in Biber (1988), with a few modifications to suit this work. The data from the lexical analysis was handled in a different manner, as the items were assumed to be highly correlated. That data was subjected to a procedure introduced by Frank H. Walkey (1997) called Composite Variable Analysis, which was developed for highly correlated data. Following these procedures, the scoring procedure tallied up factor (and composite variable) scores for each user–document. The lexical factors themselves are presented in **appendix B**. Scores were linked to each user–document, thus further enriching the corpus.[266★] The enriched corpus was then passed to the clustering stage. The specifics of the analytical procedures are now described.

5.1.4.1. Principal Axis Factoring

245. The results of the grammatical analysis were factored in the manner used by Biber (1988, p. 82)—principal factor analysis. Also called Principal Axis Factoring (PAF), it remains the recommended procedure for MDA (Cantos-Gomez, 2019, pp. 99–106).[267★] This method modified Biber's procedure in three ways. First, whereas Biber prepared the necessary correlation matrix using Pearson's *r* (Biber, 1988, app. IV),[268★] this method used Kendall's *tau*. Initially the decision was not to use Pearson's *r* given that it is a parametric technique, and no assumption of normality seemed warranted for the dataset, regardless of how it might be sampled.[269]* Spearman's ρ was then considered, as it is a rank-order coefficient and thus non-parametric. However, it was decided in the end to use Kendall's *tau* for reasons of simplicity and computational ease.[270,271] Second, whereas Biber used the promax rotation on the extracted factors (1988, p. 85), this method used the equamax rotation.[272]* While the common assumption might be to continue with promax or some other oblique rotation given that we expect the factors to show some correlation (Costello and Osborne, 2005, p. 3 ; Gaskin and Happell, 2014, p. 517 ; Goretzko, Pham and Bühner, 2021, p. 3517), such assessments appear to be rooted in an expectation of relatively simple structure underlying data. This work expects rather the beautiful, unavoidable mess of society to underlie the data. For this reason, equamax was chosen—despite the fact that it is in the family of orthogonal rotations—because it has been found to deal well with complex structure (Sass and Schmitt, 2010 ; Schmitt and Sass, 2011).[273] Third, and extending from the second, while the notion of a simple structure underlying the data is rejected, the author is perfectly open to the possibility of multiple simpler structures. We just have no good idea what they might be. Thus factoring was performed at multiple degrees (2, 4 and 8 factors) in order to see what we might see.[274,275]* Those results were passed on to the scoring procedure.

5.1.4.2. Composite Variable Analysis

246. While correlation between variables, and thus factors, is an important consideration for the grammatical analysis, it is an unavoidable complication for the lexical analysis. In that the method is seeking out high-frequency items of relatively simple compositionality (i.e. words and bigrams, with the compositional unit being the word), strong correlations among the lexical variables were expected. In that light, it was decided to forgo exploratory factor analysis regardless of rotation and to seek another technique. The technique selected is called Composite Variable Analysis (CVA). This decision was made, on the one hand, to avoid difficulties posed by likely collinearity amongst groups of variables (Egbert and Staples, 2019)[276,277] and, on the other hand, to avail of a mathematically simple approach to factoring compared to that of exploratory factor analysis. Introduced by Frank H. Walkey (1997), CVA was designed to aid the development of psychological instruments that assess single characteristics by means of multiple items.[278]* Given the high correlations between such items, the results of factor analysis can prove difficult to interpret in terms of factor assignment. Walkey proposed a simple alternative whereby latent structure (i.e. the single characteristic just mentioned) is best revealed by compositing those variables (i.e. responses to the multiple items) that are most strongly associated (1997, p. 759). These composite variables are understood as factors, even if they are of a different character from those produced by principal factor analysis as previously described.

247. To apply CVA to a set of variables, the general steps are as follows (1997, p. 760):

1. A correlation matrix is prepared.
2. A composite variable is made by aggregating the two variables with the strongest correlation above a given threshold.[279]
3. The composite variable is set aside. Its components are removed from further consideration. Any remaining variable having its strongest correlation to either removed variable is temporarily set aside.
4. Steps 2 and 3 are repeated until all correlations above a given threshold have been accounted for by aggregation and removal of the variables.
5. A correlation matrix for the composite variables and any remaining simple variables (including those temporarily set aside) is prepared.
6. Steps 2 through 5 are repeated until no correlations above a given threshold remain, or the desired number of composite variables has been reached.

This algorithm was selected as it seemed a logical choice for the lexical data at hand, given the similarities that could be drawn between the needs of this analysis and the purpose for which CVA was designed. Moreover, it is a simple and transparent process, which can be worked through with pen and paper (given a toy dataset).[280]* Furthermore, pilot testing of the algorithm (using the implementation described below) indicated that it could compose variables differentiated by evident social cleavages.[281]

248. The algorithm above was implemented in R. As with the grammatical analysis, Kendall's *tau* was used to prepare the correlation matrices.[282] The script was then run against the word and bigram lists compiled in the previous stage. The initial correlation matrix gives the correlations between all types in the input list.[283] Subsequent matrices give the correlations between the composite variables (i.e. the previously aggregated types) and the remaining unaggregated variables. The input lists were prepared in order of descending *Gamma*, and thus the resultant correlation matrices are likewise ordered.

249. Aggregation proceeds by joining the type pair with the highest positive correlation.[284] Ties are not allowed. In the case of ties in maximum pair-wise correlation (thus a potential triplet, quadruplet, etc.), ties are broken by selecting for the first match, that being the token with the lower column number. As columns are ordered by descending *Gamma*, the script effectively breaks the tie by selecting for the higher gamma. Ties in gamma are frequent, however, due to rounding in calculations. Ties in gamma are broken by selecting for alphabetical order,[285★] in which case ties are not possible (in that a different spelling or capitalisation would be a different type). Note that aggregation is cumulative, and any degree of composite variable can be aggregated. Thus a 2-member composite can join with an unaggregated item to become a 3-member composite, just as a 20-member composite can join a 10-member composite to become a 30-member composite. In this regard, note that the correlations of composite variables were scaled according to the common logarithm of their degree.[286★]

250. As with the factoring of the grammatical data, CVA was performed on the words and bigrams of Lists One and Two to obtain 2, 4, and 8 composite variables (i.e. factors). Obtaining a targeted number of factors is not straightforward. While the algorithm begins with a number of factors equal to the number of input items, and while the algorithm will eventually produce one single factor if unchecked (hence the threshold of correlation), it was observed that the decrease in factors from *n* to 1 is not monotonic—that is, the factor count (of composite variables plus all remaining unaggregated variables) might increase or decrease during an iteration. For that reason, a simple cut-off at the desired factor count was insufficient. The solution employed here was to add in a 'loss threshold' whereby aggregation of variables was not forced once a certain proportion of the input data had been aggregated.[287]

251. The procedure proved to be extremely fast (see **n. 282**). Although obtaining a specified number of composite variables for a given set of input data requires some tweaking of parameters, the procedure is so fast that experimentation is relatively painless. Moreover, there is no probabilistic component in the algorithm or its implementation here, so results are stable—with no change to script or input, the output will not change. As the author was confident in the choice of algorithm and the implementation, and as the pilot testing had produced sensible results, the CVA results were passed on to the scoring procedure.

5.1.5. User–Document Clustering

252. The clustering of user–documents is a further step beyond MDA. In that approach, the clustering that is done is by means of factor analysis, and seeks to cluster observed features. This method also clusters features by means of Principal Axis Factoring and Composite Variable Analysis, as described in the preceding section. However, this method clusters yet again, this time grouping the user–documents, which are clustered on the scores linked to them in the factoring stage, and also on the sociodemographic data that was linked to each user–document in the corpus preparation stage.

253. In both cases (i.e. for the linguistic data and for the sociodemographic data), clustering was done using the *k*-medoids algorithm (Kaufman and Rousseeuw, 1990, chp. 2). Like the venerable *k*-means algorithm (Hartigan and Wong, 1979),[288★] *k*-medoids generates clusters by partitioning data on similarity of features. However, unlike *k*-means which seeks to cluster around a calculated 'centroid', *k*-medoids seeks to cluster around actual data points that are central to clusters (these are the medoids in question).[289★] The impetus for using *k*-medoids over *k*-means is that the former handles noise and outliers better, and has been found to work well with mixed data (i.e. different variable types), specifically sociodemographic data (Hennig and Liao, 2013).[290★] Thus *k*-medoids was a sound choice for this method.[291]

254. The user–documents were clustered using *k*-medoids on the results of the factoring stage, setting *k* as appropriate for the degree of factoring (2, 4, and 8 factors). Note that the sociodemographic data as collected was rather granular (see **n. 200**). Age, income, and education were divide up at the source according to narrow brackets. These data were not factored, but rather were reduced by collapsing the brackets into only three per category (thus roughly, 'low', 'medium', and 'high'). As the brackets were expressed in 'percentage population' terms (e.g. population 65 years of age or older: 12%), this was a simple additive process. These data were clustered at the same levels of *k* as the linguistic data. Sociodemographic and linguistic cluster assignments for each factoring level, are linked to each user–document, thus providing the final enrichment of the corpus.

5.1.6. Results

255. The final output of the method is the enriched corpus, comprising:

- user–documents ;
- their associated places and geographic information ;
- the sociodemographic profiles and the cluster assignments of those places ;
- the linguistic factor scores of users and their cluster assignments.

To facilitate analysis, the results of the method are extracted from the enriched corpus. At this stage, the user subcorpora themselves are no longer needed—we have already extracted the needed information. The corpus itself can now be temporarily set aside for safe-keeping, and for eventual disposal according to this project's data retention plan as specified during ethics clearance (cf. **n. 236**). The results thus differ little from the enriched corpus: both are essentially large data tables, however the results table is much more compact and thus easier to process.

256. Recall that the research questions posed in this work (**chp. 1 § 1.5**):

RQ$_1$ – How can political participation as reconceptualised in hybrid society be operationalised for computational and statistical analysis? and

RQ$_2$ – Can the results of such operationalisation remain interpretable from a structural perspective?

Chapter 4, on the operationalisation of language, responds to **RQ$_1$** by laying the groundwork for how that question might be answered. This chapter has explained the method by which that operationalisation is implemented. In so doing, these chapters have provided a provisional answer to **RQ$_1$**. However, the test of that answer lies in **RQ$_2$**—the first question is open-ended, whereas the second is not. Thus to answer **RQ$_2$**, and thereby assess whether the provisional answer to **RQ$_1$** is warranted, we now turn to the analysis of the results and the evaluation of the final question.

Chapter 6

Analysis and Evaluation

11. Facing uncertainty. The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise." Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate. It would be a mistake not to face up to this fact, for by denying it, we would deny ourselves the use of a great body of approximate knowledge, as well as failing to maintain alertness to the possible importance in each particular instance of particular ways in which our knowledge is incomplete.

John W. Tukey, "The Future of Data Analysis", 1962, pp. 13–14

257. This chapter evaluates the results of the method so as to address the research questions put forward in this work. These research questions, presented in **chp. 1 § 1.5**, are:

RQ_1 – How can political participation as reconceptualised in hybrid society be operationalised for computational and statistical analysis? and

RQ_2 – Can the results of such operationalisation remain interpretable from a structural perspective?

The first research question is addressed by the proposed operationalisation developed in in **chp. 4**. That operationalisation as implemented in the method was laid out in **chp. 5** to give a provisional answer to **RQ_1**. The evaluation of the results of the method will answer **RQ_2** and allow us to gauge if the provisional answer to **RQ_1** is warranted. The short answer to that question is that political participation, and any other mass political phenomenon, as reconceptualised in hybrid society is naturally interpretable from a structural perspective because it is a socially communicative phenomenon, and the modes and manners of human communication are inextricably bound to social structure—as argued in **chp. 4**, communication is the root and impetus of social structure. The long answer is of course a bit trickier, because interpretation hinges on perception. Thus, empirically speaking, the answer to the second question depends on how the first question has been answered. For that reason **RQ_2** must be answered in the affirmative for the provisional answer to **RQ_1** to be warranted.

258. As this work has chosen to investigate pervasive features of language on the grounds that they might provide relatively stable and durable endogenous markers of sociation, we are faced with the challenge of a 'weak signal'. In linguistic terms, salient features (in terms of information) and marked features (in terms of society) provide a strong signal—in a general sense these three terms are synonymous. However, strong signals can be misleading in social contexts. Unlike the constituent components of natural phenomena, the constituent components of social phenomena (i.e. the communicative actions of people ; cf. **chp. 2 § 2.2.2.1**) consciously manipulate strong signals for a variety of ends, in good faith and bad. Weak signals are another matter: we often are not consciously aware of them, and even when we are we nevertheless have difficulty modifying our production and understanding of them (**chp 4. § 4.1.2.2**). Thus weakness cuts both ways, in that weak-signal linguistic features might provide a more robust measure from a social perspective, but at the same time they are difficult to reveal from an informational perspective. In that light, this evaluation cannot provide any clear, single measure of relation between language in society and structure in society. Throughout, this work has argued against such measures in social inquiry—following the words of John Tukey (**p. 205**), it is better to have an approximate answer to the right question than an exact answer to the wrong question. And so this chapter will step through the results of the method in order to argue that we do in fact see this weak signal. While endogenous variables cannot replace exogenous variables—because they are addressing different questions—we nevertheless can see that they do echo each other, even if faintly.[292] As noted at the beginning of **chp. 1 § 1.7**, the essence of the analysis that reveals this echo is quite simple. As the method has collected sociodemographic and linguistic data of a common population and subsequently grouped that population according to linguistic similarities, one can compare those linguistic groups in terms of their aggregate sociodemographic data. Regular differentials in the sociodemographic data of linguistics groups are the echoes we will reveal. How that is to be done is now explained.

6.1. Approach to Analysis

259. Analysis and evaluation of the results of the method face a challenge, in that the method produces a complex dataset. It is certainly complicated, simply due to the number of components and elements thereof. Complication, however, is not necessarily a challenge to analysis and evaluation, as we can see and say clearly how elements and components relate ; the primary role of computational approaches to social research is to make complicated and extensive datasets more tractable. However, this dataset is complex in that it combines data types from different domains (in this case the sociodemographic and linguistic) that we know—both intuitively and through the work of a variety of research traditions—are deeply interrelated, but in a manner that is not deterministic and that is difficult to operationalise for study. The challenge of complexity is ever-present in social inquiry, and thus we rely heavily on established approaches to certain kinds of work. That is certainly the case with the method used here. However, in regard to analysis and evaluation, to the author's knowledge there is no established cross-analytical approach (cf. ¶¶ **194**, **201**) for the kind of complex dataset that the method has produced. The approach to analysis taken here is therefore exploratory. The three analytical steps to be taken thus are not definitive, but rather first cuts at seeking to reveal sociodemographic–linguistic relations in this complex dataset.

260. The results of the method will be addressed in a manner suitable to answer the second research question: Can political phenomena, operationalised as socially communicative phenomena, nevertheless be understood from a structural perspective? In that light, we are looking for evidence of sociodemographic differentials across groups exhibiting linguistic similarities. Thus we have assembled the dataset where the user–documents are each associated with a single place of known sociodemographic characteristics. These user–documents have been clustered variously according to a range of linguistic measures. A given cluster will thus comprise user–documents assessed to be more similar by the given measure. As user–documents also carry the sociodemographic profile of the place to which they are associated, we can compare the linguistic clusters in terms of their aggregate sociodemographic profiles. Regular differentials in these aggregate profiles across linguistic clusters are the evidence that we seek. If such differentials are found—and they are—then the research questions are answered in the affirmative. That is the general outline of analysis and evaluation.

261. The following describes the portion of the 'enriched' data of the user–documents that is considered in the analysis, that is, their sociodemographic and linguistic profiles. As explained in **chp. 5**, the sociodemographic profile is composed of a statistical description of the place to which the user–document is associated, and the linguistic profile is composed of factor scores representing measures of lexicogrammatical features attested in the document. These profiles were ordered into 'variable sets' on which clustering of the user–documents (**chp. 5 § 5.1.5**) was performed. The sociodemographic variable sets (denoted by V_{SD}) comprise four primary sets, representing US Census statistical profiles in terms of age, income, education, and ethnicity (sets `age`, `inc`, `edu`, and `eth`, respectively),[293] and a composite set (set `base`) produced by combining the age, income, and education sets, for a total of 5 V_{SD} sets. Note that no clustering was done on the ethnicity data, which was included only to allow for further analysis if warranted. The linguistic variable sets (denoted by V_L) comprise five primary sets, representing the factor scores for the grammatical analysis (set `m`), and the four sets of factor scores for the List 1 and List 2 words and bigrams (sets `w1`, `w2`, `b1`, and `b2`, respectively). Further variable sets were produced from various combinations of these primary sets, for a total of 13 V_L sets. As noted, the clustering stage of the method used these 18 variable sets for clustering, and thus each user–document has a resultant cluster assignment for each set, at *k* 2, 3 and 5. Of this 'portion' of the enriched data, only a selection of these items are explicitly addressed. The preceding description is offered to facilitate comprehension, and for reference.

262. As noted, the analysis proceeds in three exploratory steps. **Step 1** considers the cluster assignments of user–documents on the V_{SD} and V_L sets, evaluating those assignments to gauge a baseline of correlation. This is done by means of calculating the pair-wise percent agreement (α_0) between V_{SD} and V_L assignments. Thus each of the 5 V_{SD} sets will be compared to each of the 13 V_L sets, for each clustering at *k* 2, 3, and 5, and for each cluster within those clusterings.[294]★ This step aims simply to gauge the relatedness of the V_{SD} and V_L in a rough manner—this analysis is exploratory, and so we explore. In the case of step 1, such exploration is a failure ; in hindsight, for reasons to be explained, this specific approach cannot work on this data. The essential problem, as will be discussed, is that of skew in cluster assignments (i.e. clusters are not balanced). However, step 1 is nevertheless presented as the process of arriving to that hindsight is instructive. Specifically, the discussion takes issue with assumptions of 'randomness' underlying measures of agreement commonly encountered in the social sciences. More in **§ 6.2.1**.

263. **Step 2** compares the V_{SD} profiles within each clustering. Whereas the examination of α_0 concerns the cluster assignments according to variable sets, the examination of V_{SD} profiles looks at V_L clusterings (i.e. clusterings on linguistic factor scores) directly in order to see if they are differentiated in terms of their aggregate V_{SD} profiles. This is done by means of box plots, and evaluation is by eye. Recall that the V_{SD} sets represent statistical profiles of age, income, education, and ethnicity, each split into a number of brackets, and reported in terms of percentage population. These V_{SD} sets are presented (partially aggregated) on the horizontal axis of the plot ; thus it is categoric and has no dimension as such. Population percentage is represented on the vertical axis, and thus the axis is dimensional ranging from 0% and up. In functional terms, the plot functions similarly to a population pyramid turned on its side. The individual clusters themselves are presented by boxes in the basic style of Tukey (McGill, Tukey and Larsen, 1978), where the bounding box shows the interquartile range with a line representing the median value (i.e. median population percentage), and whiskers indicate the rough extent of the data. Plotting the data in this manner, it is possible to see the clusters differentiate in terms of their V_{SD} profiles. However, that differentiation is not immediately evident, as will be explained. The basic difficulty is that generally all V_{L} cluster assignments will be attested in all places. That is addressed by considering the skew observed in step 1, and by extension considering the proportions of cluster assignments across places. More in **§ 6.2.2**.

264. **Step 3** investigates the overall relationship between V_{SD} profiles and V_{L} clusterings. While step 2, as will be shown, is helpful in viewing profiles directly (again, in a manner similar to a pyramid plot), such an approach limits how we might view a more general relationship between V_{SD} profiles and V_{L} clusters. It is that general relationship that the **RQs** hinge upon. The approach taken here is again visual. As this analysis is exploratory we are not immediately concerned with the value of any specific data point ; rather, we aim to get a general sense of what is happening in the data so as to inform further investigation (cf. Tukey, 1977). The analysis of step 3 is presented in what here are called 'constellation charts'.[295] These charts are comprised of panels plotting the relationship between cluster proportion (on the horizontal axis) and a V_{SD} index (on the vertical axis). The index will be explained in step 3. There is one panel per cluster, so for example at *k* 3 there will be 3 panels. The points that are plotted in each panel are the places in the dataset, and each panel plots all places. The panels are not identical, however—the horizontal axis represents the proportion of population in a place assigned to a given cluster. As the dots of the places shift and rearrange horizontally according to V_{L} cluster proportions (but not vertically as the V_{SD} index of places is constant), one compares the panels to get a sense of the overall tendencies of the clusters. To the author's knowledge this is not a standard type of chart, so this description is perhaps confusing. Things will be made clear in **§ 6.2.3**. The thing to keep foremost in mind is this—constellation charts are not meant to be read in quantified terms, but rather impressionistic terms (pointillist, if you like). This is exploratory analysis, and we aim to sense the flow of the data, not to measure it at some arbitrary precision. In the constellation charts it is observed, for all V_{L} set clusterings *k*3 and *k*5, that there are certain clusters the prevalence of which correlates positively with V_{SD} index scores, certain clusters that show a roughly neutral correlation (rather, shallow convex, i.e. a hump), and certain clusters that correlate negatively. In short, it is found that user–documents clustered solely on V_L sets will demonstrate regular patterns when evaluated according to V_{SD} sets. That is, groupings of language similarities exhibit structure in sociodemographic terms. These clusters here are labelled 'proper', 'standard', and 'non-standard',[296] respectively, denoting the presumed structural–linguistic relationship observed in the data—that some modes of speech are favoured and thus rewarded (in V_{SD} terms), some are tolerated if in their proper place (hence the shallow convex relationship), and some modes are disfavoured and rewards are withheld.

265. The question of relative proportions of V_L clusters in a given place plays an important role in steps two and three. Without considering proportions, the dataset cannot serve our purposes. On the one hand, the method is intrinsically limited in its geographic resolution to named places ; with this method we are unable to see neighbourhoods, blocks, or streets where one might expect to find increased linguistic similarities. On the other hand, contemporary society is increasingly mobile and mixed—while we are all familiar with areas that are characterised by their constituent communities, it is uncommon to find municipalities that are not highly mixed when viewed in the round. In terms of this dataset, all V_L clusters (that is, the linguistic cluster assignments) are likely to be found in any given place. Generally, all V_L are found in all V_{SD}. This reflects the social intermixture of most places, but moreover the operationalisation of pervasive features—it is in the name. Thus steps two and three hinge upon proportions to intuit the social composition of places. Note that the proportion of V_L cluster assignments is in terms of user–documents in the dataset. For that reason, the dataset has been filtered further prior to analysis to remove all places represented by fewer than 10 user–documents. Furthermore, the bulk of the dataset comprises places in the southern region of Michigan, and so to facilitate the following analysis the dataset was bounded to places in the lower 33 counties of Michigan (out of a total of 87). This bounding was done as an expedient, simply for reasons of visualisation in the face of localised data sparsity (cf. **fig. 2**). These final filtering steps reduced the total user–document count from 5,872 to 4,637, and the total place count from 417 to 84. The distribution of users in both the full dataset and final dataset is shown in **fig. 2**. However, step 3 makes evident that the picture is not so clear on the ground, given the social mixing present in municipalities of all types and sizes, and given the limitations of the method in terms of geographic resolution. At the end of step 3, a small selection of maps is presented that show V_{SD} data and V_L clustering proportions across places. The V_{SD} profile data is mapped to the site of research to help contextualise the discussion in this chapter, following which the V_L data are mapped in a manner showing the relative cluster proportions of each place. The lack of clarity on the ground, however, is a question of the visualisation technique, not of the data. The results of the analysis are clear: **RQ$_2$** is answered in the affirmative. The analysis of the final dataset now follows.

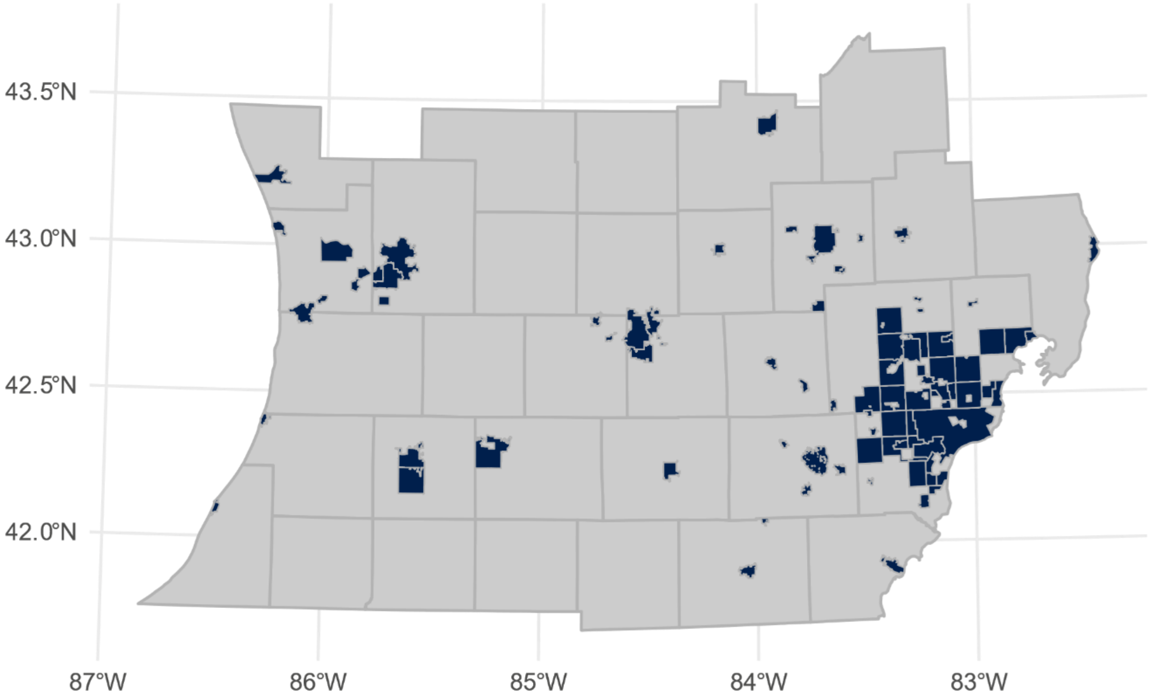


Figure 2. User Distribution in Final Dataset

NOTE: During preparations for the analysis of results, it was observed that clusterings produced using the Principal Axis Factoring (PAF) factor scores (specifically the clusterings on V_L set `m`, the grammatical factor scores) were problematic from a social perspective. These clusters were heavily skewed, even more so than is observed with the Composite Variable Analysis (CVA) factor scores in (**§ 6.2.3 ¶ 295**). This is likely the result of the nature of the PAF factoring technique in that most variance is 'front loaded' onto the first few factors. The rationale for using CVA, that it would produce more balanced factors, is thus supported. Despite the fact that PAF was performed in order to emulate Biber's MDA method (1988), concerns about its use for producing social clusterings resulted in a decision to abandon the PAF factor scores, and to use only the CVA factor scores throughout the analysis. Note that refinement of CVA for social research is item 4 under suggestions for further work (**chp. 7 § 7.4**).

6.2. Analysis of Results

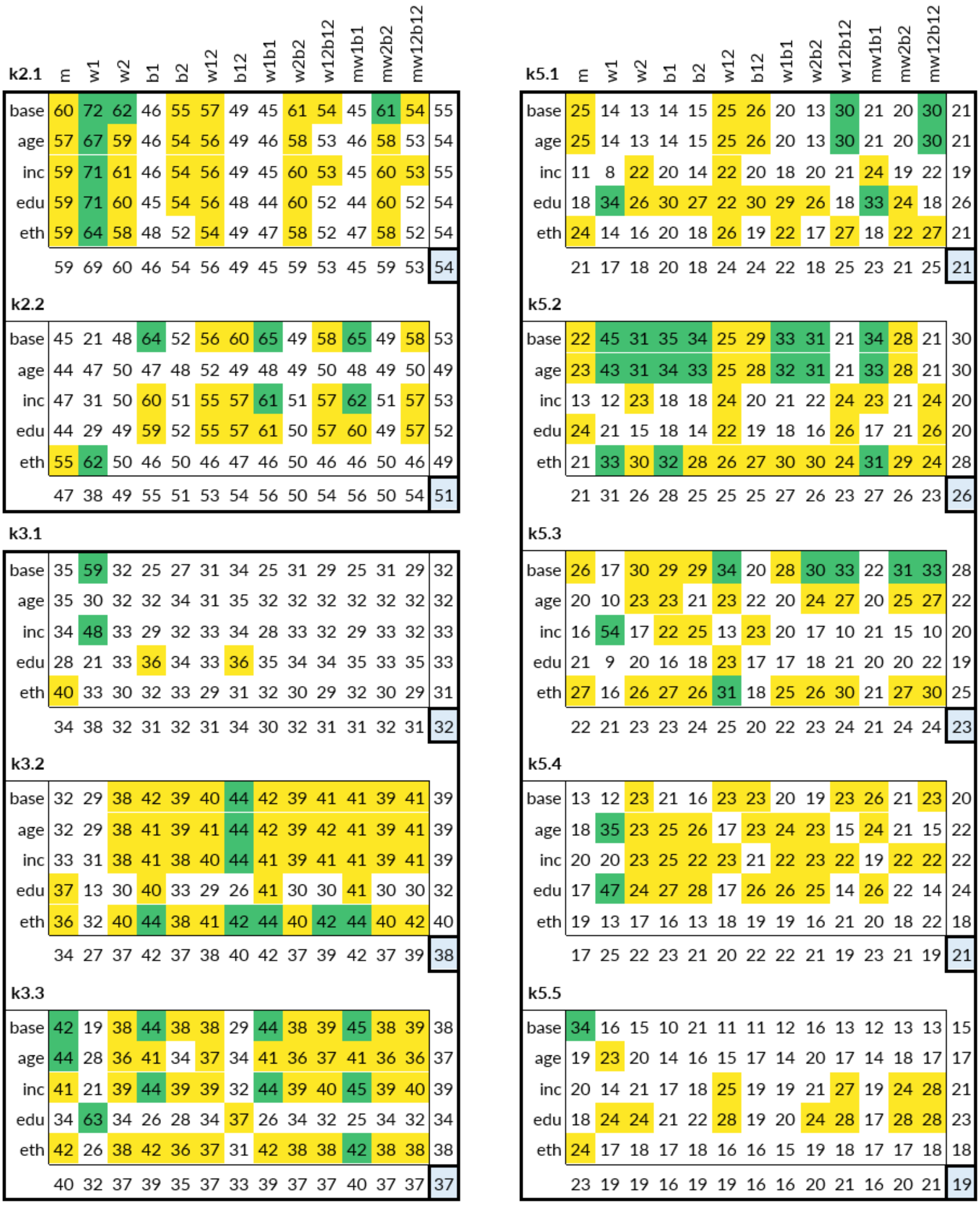
6.2.1. Step 1 – Cluster Assignments

266. As described in **chp. 5 § 5.1.6**, each user–document is assigned to clusters within clusterings of *k* 2, 3, and 5 according to the V_L scores derived from their compiled text. Places are clustered separately according to their V_{SD} profiles, and those cluster assignments are added to the user–documents according to their associated place. The resultant V_{SD} and V_L cluster assignments for all user–documents can then be compared pair-wise using common measures for assessing inter-coder agreement. Percent agreement (α_o) is used here, given concerns about the logic and applicability of more 'powerful' measures (Feng, 2014 ; Zhao *et al.*, 2018 ; Zhao, Liu and Deng, 2013),[297] the substantial increase in computing resources and time observed in applying such measures to this dataset, and moreover given the relative simplicity and intuitiveness of α_o.[298] That measure is amenable to the use case, which is a two-way comparison of nominal data that has been coded programmatically and which may or may not be highly skewed (Feng, 2014, pp. 1812–1813). (Note that the failure of this step due to skewness does not reflect on the recommendations of Guangchao Charles Feng, but rather on the repurposing of α_o to evaluate cluster assignments.)

267. There is certainly a limitation in applying any such intercoder agreement measure to evaluate cluster assignments: such labels are not 'codes' ; they have no inherent meaning or relation to the objects being clustered. There is no coding protocol to guide a clustering algorithm in its assignments of labels. In the case of α_0, this is a pickle. Consider two cluster assignments `A` and `B`. If `A` and `B` are both {1, 2, 3, 1, 2, 3}, then α_0 is 100%. Well done! But if the `B` labels were {3, 1, 2, 3, 1, 2} then α_0 would 0%, even though the underlying clustering is not changed at all. Here the arbitrariness of cluster labels is an advantage, as we can permute the labels without affecting the underlying data. Thus we can keep `A` fixed and calculate α_0 for all permutations of B (a total of *k*!—indicating factorial, not excitement), then take the highest score.[299]* This is not cheating in the least, but rather putting the data in proper alignment. As the underlying data are not affected, alignment is comparable to the process of rotation in factor analysis (**chp. 5 § 5.1.4**), serving the purpose of increasing clarity of results without altering the fundamental analysis. The final dataset was so aligned prior to calculating α_0 scores. All V_{SD} clusters were aligned to the composite V_{SD} set `base`. In terms of the hypothetical cluster assignments `A` and `B` above, `base` provided the fixed `A` assignments, and all other V_{SD} sets provided `B` in their turn. The appropriate permutation for a given V_{SD} set was that which yielded the highest α_0 with `base`. The most basic V_L set `w1` (the list of pervasive words that are key for certain user–documents ; § **6.1**) was also aligned to `base`, and all other V_L sets were then aligned to `w1`. The calculation of pair-wise α_0 scores for all V_{SD} and V_L scores was then performed without further permutation. The arbitrariness of cluster assignments poses a further limitation to the use of α_0 for evaluating cluster assignments: since there is no standard reference (e.g. a coding protocol) for cluster assignments, the range of α_0 (0–100%) cannot be understood as a linear scale. This will be discussed as the analysis proceeds. Nevertheless, as the purpose of the first step of analysis is simply to baseline relatedness in the V_{SD} and V_L cluster assignments, α_0 is deemed a reasonable measure.

268. The α_0 scores for all clusterings are now presented in tables with a common format. They are read in the following manner (see **fig. 3**). The tabular field itself presents the pair-wise α_0 scores. Cells are shaded according to how a given score compares to others, as specified in each case. Comparisons are either within or across clusterings, also as specified. The *k*2 clustering appears in the upper left corner, the *k*3 clustering in the lower left, and the *k*5 clustering makes up the right hand side. Each clustering is surrounded by a bounding box. The tabular field is divided into 10 subfields, and each clustering is composed of *k* subfields (labelled *k*2.*x*, *k*3.*x*, and *k*5.*x*, where *x* indicates the specific cluster). The V_{SD} sets are arranged row-wise, and the V_L sets are arranged column-wise. Note that each cluster subfield is anchored to the cluster assignment for the V_L sets—thus in *k*2.1, for example, all V_L sets have been assigned to cluster 1, in *k*2.2 to 2, etc. The marginal values are row and column averages. The box at the intersection of the marginal values in the lower right of each subfield indicates the average of the marginal values, giving a summary score for the entire cluster. In all cases, scores were calculated for the entire dataset, not a subsample.

269. **Figure 3** presents the raw α_0 scores for all clusterings. The scores are compared within each clustering, thus <u>within each bounding box</u>. Cells with a lighter shading



k2.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	60	72	62	46	55	57	49	45	61	54	45	61	54	55
age	57	67	59	46	54	56	49	46	58	53	46	58	53	54
inc	59	71	61	46	54	56	49	45	60	53	45	60	53	55
edu	59	71	60	45	54	56	48	44	60	52	44	60	52	54
eth	59	64	58	48	52	54	49	47	58	52	47	58	52	54
	59	69	60	46	54	56	49	45	59	53	45	59	53	54

k2.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	45	21	48	64	52	56	60	65	49	58	65	49	58	53
age	44	47	50	47	48	52	49	48	49	50	48	49	50	49
inc	47	31	50	60	51	55	57	61	51	57	62	51	57	53
edu	44	29	49	59	52	55	57	61	50	57	60	49	57	52
eth	55	62	50	46	50	46	47	46	50	46	46	50	46	49
	47	38	49	55	51	53	54	56	50	54	56	50	54	51

k3.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	35	59	32	25	27	31	34	25	31	29	25	31	29	32
age	35	30	32	32	34	31	35	32	32	32	32	32	32	32
inc	34	48	33	29	32	33	34	28	33	32	29	33	32	33
edu	28	21	33	36	34	33	36	35	34	34	35	33	35	33
eth	40	33	30	32	33	29	31	32	30	29	32	30	29	31
	34	38	32	31	32	31	34	30	32	31	31	32	31	32

k3.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	32	29	38	42	39	40	44	42	39	41	41	39	41	39
age	32	29	38	41	39	41	44	42	39	42	41	39	41	39
inc	33	31	38	41	38	40	44	41	39	41	41	39	41	39
edu	37	13	30	40	33	29	26	41	30	30	41	30	30	32
eth	36	32	40	44	38	41	42	44	40	42	44	40	42	40
	34	27	37	42	37	38	40	42	37	39	42	37	39	38

k3.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	42	19	38	44	38	38	29	44	38	39	45	38	39	38
age	44	28	36	41	34	37	34	41	36	37	41	36	36	37
inc	41	21	39	44	39	39	32	44	39	40	45	39	40	39
edu	34	63	34	26	28	34	37	26	34	32	25	34	32	34
eth	42	26	38	42	36	37	31	42	38	38	42	38	38	38
	40	32	37	39	35	37	33	39	37	37	40	37	37	37

k5.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	25	14	13	14	15	25	26	20	13	30	21	20	30	21
age	25	14	13	14	15	25	26	20	13	30	21	20	30	21
inc	11	8	22	20	14	22	20	18	20	21	24	19	22	19
edu	18	34	26	30	27	22	30	29	26	18	33	24	18	26
eth	24	14	16	20	18	26	19	22	17	27	18	22	27	21
	21	17	18	20	18	24	24	22	18	25	23	21	25	21

k5.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	22	45	31	35	34	25	29	33	31	21	34	28	21	30
age	23	43	31	34	33	25	28	32	31	21	33	28	21	30
inc	13	12	23	18	18	24	20	21	22	24	23	21	24	20
edu	24	21	15	18	14	22	19	18	16	26	17	21	26	20
eth	21	33	30	32	28	26	27	30	30	24	31	29	24	28
	21	31	26	28	25	25	25	27	26	23	27	26	23	26

k5.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	26	17	30	29	29	34	20	28	30	33	22	31	33	28
age	20	10	23	23	21	23	22	20	24	27	20	25	27	22
inc	16	54	17	22	25	13	23	20	17	10	21	15	10	20
edu	21	9	20	16	18	23	17	17	18	21	20	20	22	19
eth	27	16	26	27	26	31	18	25	26	30	21	27	30	25
	22	21	23	23	24	25	20	22	23	24	21	24	24	23

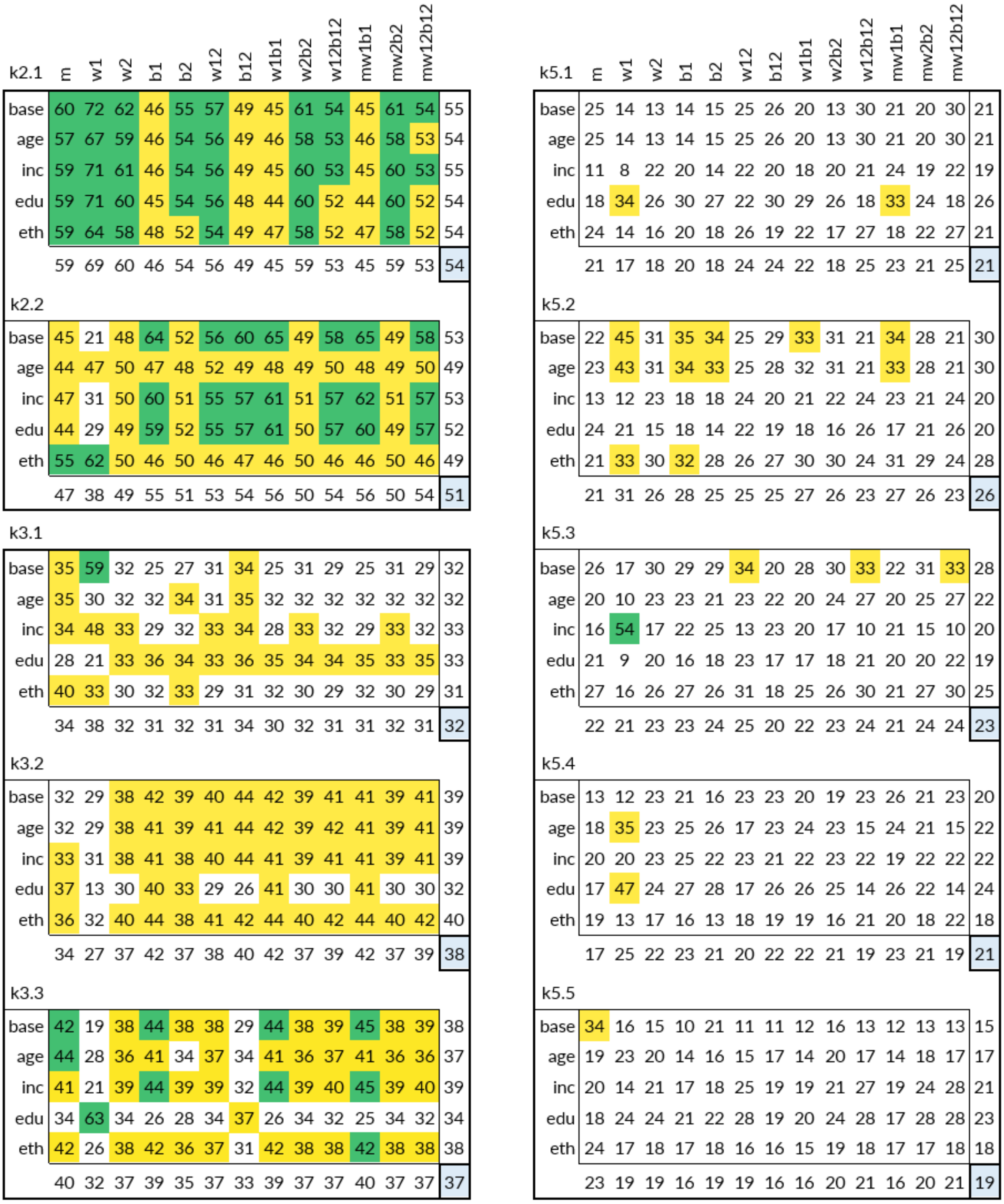
k5.4	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	13	12	23	21	16	23	23	20	19	23	26	21	23	20
age	18	35	23	25	26	17	23	24	23	15	24	21	15	22
inc	20	20	23	25	22	23	21	22	23	22	19	22	22	22
edu	17	47	24	27	28	17	26	26	25	14	26	22	14	24
eth	19	13	17	16	13	18	19	19	16	21	20	18	22	18
	17	25	22	23	21	20	22	22	21	19	23	21	19	21

k5.5	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	34	16	15	10	21	11	11	12	16	13	12	13	13	15
age	19	23	20	14	16	15	17	14	20	17	14	18	17	17
inc	20	14	21	17	18	25	19	19	21	27	19	24	28	21
edu	18	24	24	21	22	28	19	20	24	28	17	28	28	23
eth	24	17	18	17	18	16	16	15	19	18	17	17	18	18
	23	19	19	16	19	19	16	16	20	21	16	20	21	19

Figure 3. Raw Percent Agreement Scores, by Clustering

(presenting in yellow) are scores above the clustering average score, Cells with a darker shading (presenting in green) are in the top 10% of clustering scores. Cells with scores below the clustering average are unshaded. Note that α_o does not account for chance agreement. The cut-off for chance agreement depends on the number of possible cluster assignments ; thus for *k*2 the level is 50%, for *k*3 the level is 33%, and for *k*5 the level is 20%. Overall, the α_o scores are low in terms of what would be considered acceptable for content analytical work, for example Riffe, Lacy and Fico (1998, pp. 127–128). However, this is not content analytical work. We are not comparing the agreement of different coders on a dataset, but rather the agreement of different models of a dataset. Thus in the swaths of negative space in **fig. 3**, those being the uncoloured cells where agreement is below the clustering average (and which are generally near or below the threshold of chance agreement), we could perhaps surmise in those specific cases that the models in question (a given V_{SD} model versus a given V_L model) are unrelated. However, scores in any of the three clusterings that fall below the threshold of chance agreement could be understood as a divergence between models, and thus carry important information. Moreover, while there are sizable pockets of uncoloured below-threshold cells, there are likewise sizeable pockets of elevated V_{SD} possibly suggesting model agreement. In each clustering, there is a fair amount of model agreement among some of the variable sets. The vertical striping that can be observed indicates that such agreement is due to the character of the V_L sets more so than the V_{SD} sets (and the repeated striping from left to right is due to the composite nature of the six right-most sets). The strongest agreement is seen in the V_L set `w1` (e.g. in *k*2.1), which as noted is the set of single words assembled from the most key words in each user–document (see **chp. 5 § 5.1.3.2**).

270. Figure 4 shows the raw scores compared across all clusterings, that is, across bounding boxes (note that the scores remain the same as in **fig. 3** ; only shading has changed). The core of agreement, or rather model convergence, is to be found in the *k*2 clustering. Clusterings *k*3 and *k*5 show some level of model convergence, but it is minor compared to that observed in *k*2. It is all smoke, however. While the agreement is encouraging as an initial indication that there is some connection between the 'models' represented by the V_{SD} and V_L sets, the situation is obscured by not accounting for chance agreement. Models with fewer classes (in this case, a lower *k*) will have their α_o scores elevated compared to those with more classes. In reality we are not concerned with chance. Chance is a statistician's conceit that does not fit well with social phenomena, especially the assumption of absolute randomness that is made to adjust certain measures of agreement (cf. Zhao *et al.*, 2018). We are not concerned with chance agreement inflating scores or chance disagreement deflating scores. The cluster assignments obtained for both V_{SD} and V_L sets are made in a relatively deterministic manner, as intended and designed. With the same input documents, the same clusterings will be generated again and again.



k2.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	60	72	62	46	55	57	49	45	61	54	45	61	54	55
age	57	67	59	46	54	56	49	46	58	53	46	58	53	54
inc	59	71	61	46	54	56	49	45	60	53	45	60	53	55
edu	59	71	60	45	54	56	48	44	60	52	44	60	52	54
eth	59	64	58	48	52	54	49	47	58	52	47	58	52	54
	59	69	60	46	54	56	49	45	59	53	45	59	53	54

k2.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	45	21	48	64	52	56	60	65	49	58	65	49	58	53
age	44	47	50	47	48	52	49	48	49	50	48	49	50	49
inc	47	31	50	60	51	55	57	61	51	57	62	51	57	53
edu	44	29	49	59	52	55	57	61	50	57	60	49	57	52
eth	55	62	50	46	50	46	47	46	50	46	46	50	46	49
	47	38	49	55	51	53	54	56	50	54	56	50	54	51

k3.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	35	59	32	25	27	31	34	25	31	29	25	31	29	32
age	35	30	32	32	34	31	35	32	32	32	32	32	32	32
inc	34	48	33	29	32	33	34	28	33	32	29	33	32	33
edu	28	21	33	36	34	33	36	35	34	34	35	33	35	33
eth	40	33	30	32	33	29	31	32	30	29	32	30	29	31
	34	38	32	31	32	31	34	30	32	31	31	32	31	32

k3.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	32	29	38	42	39	40	44	42	39	41	41	39	41	39
age	32	29	38	41	39	41	44	42	39	42	41	39	41	39
inc	33	31	38	41	38	40	44	41	39	41	41	39	41	39
edu	37	13	30	40	33	29	26	41	30	30	41	30	30	32
eth	36	32	40	44	38	41	42	44	40	42	44	40	42	40
	34	27	37	42	37	38	40	42	37	39	42	37	39	38

k3.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	42	19	38	44	38	38	29	44	38	39	45	38	39	38
age	44	28	36	41	34	37	34	41	36	37	41	36	36	37
inc	41	21	39	44	39	39	32	44	39	40	45	39	40	39
edu	34	63	34	26	28	34	37	26	34	32	25	34	32	34
eth	42	26	38	42	36	37	31	42	38	38	42	38	38	38
	40	32	37	39	35	37	33	39	37	37	40	37	37	37

k5.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	25	14	13	14	15	25	26	20	13	30	21	20	30	21
age	25	14	13	14	15	25	26	20	13	30	21	20	30	21
inc	11	8	22	20	14	22	20	18	20	21	24	19	22	19
edu	18	34	26	30	27	22	30	29	26	18	33	24	18	26
eth	24	14	16	20	18	26	19	22	17	27	18	22	27	21
	21	17	18	20	18	24	24	22	18	25	23	21	25	21

k5.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	22	45	31	35	34	25	29	33	31	21	34	28	21	30
age	23	43	31	34	33	25	28	32	31	21	33	28	21	30
inc	13	12	23	18	18	24	20	21	22	24	23	21	24	20
edu	24	21	15	18	14	22	19	18	16	26	17	21	26	20
eth	21	33	30	32	28	26	27	30	30	24	31	29	24	28
	21	31	26	28	25	25	25	27	26	23	27	26	23	26

k5.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	26	17	30	29	29	34	20	28	30	33	22	31	33	28
age	20	10	23	23	21	23	22	20	24	27	20	25	27	22
inc	16	54	17	22	25	13	23	20	17	10	21	15	10	20
edu	21	9	20	16	18	23	17	17	18	21	20	20	22	19
eth	27	16	26	27	26	31	18	25	26	30	21	27	30	25
	22	21	23	23	24	25	20	22	23	24	21	24	24	23

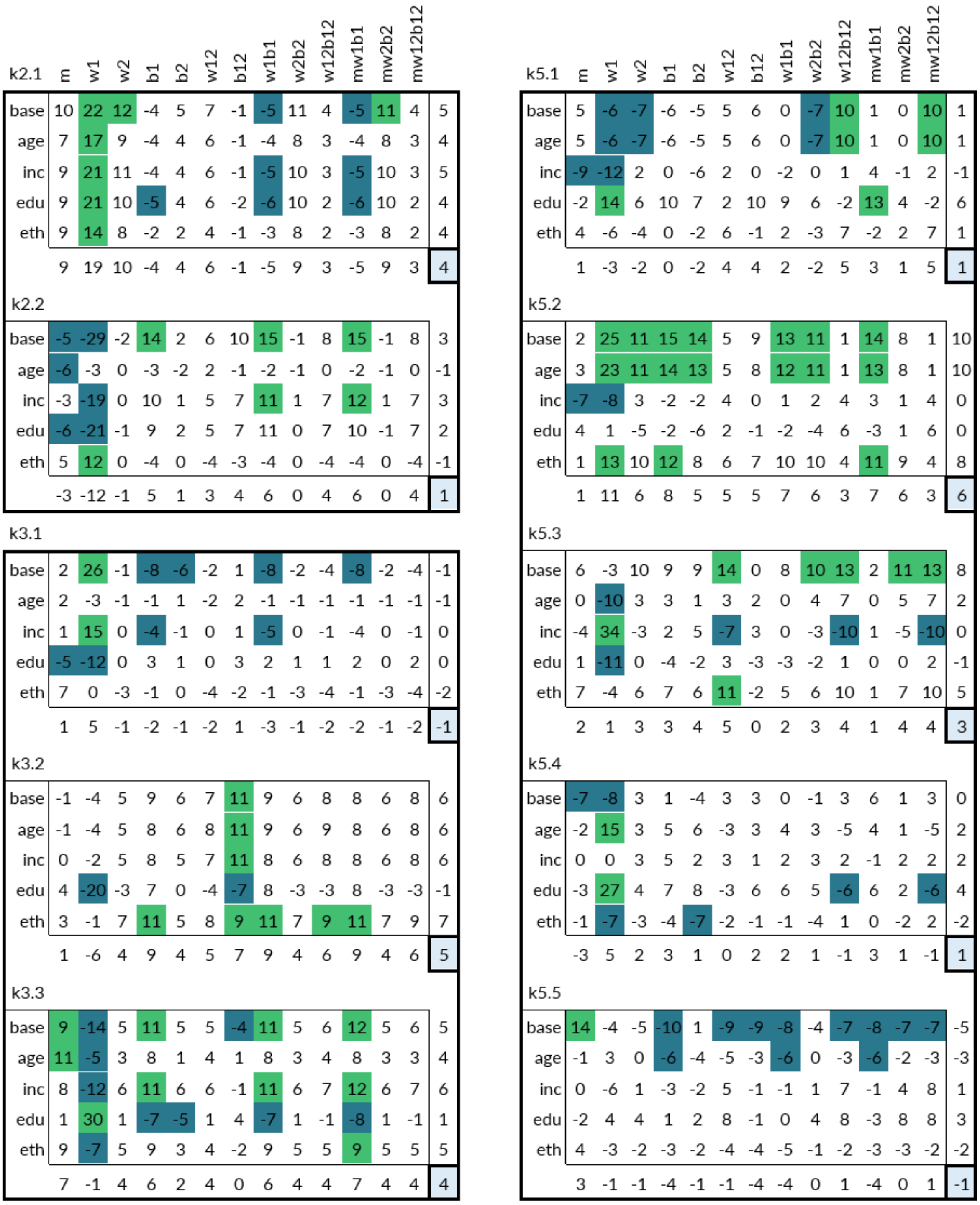
k5.4	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	13	12	23	21	16	23	23	20	19	23	26	21	23	20
age	18	35	23	25	26	17	23	24	23	15	24	21	15	22
inc	20	20	23	25	22	23	21	22	23	22	19	22	22	22
edu	17	47	24	27	28	17	26	26	25	14	26	22	14	24
eth	19	13	17	16	13	18	19	19	16	21	20	18	22	18
	17	25	22	23	21	20	22	22	21	19	23	21	19	21

k5.5	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	34	16	15	10	21	11	11	12	16	13	12	13	13	15
age	19	23	20	14	16	15	17	14	20	17	14	18	17	17
inc	20	14	21	17	18	25	19	19	21	27	19	24	28	21
edu	18	24	24	21	22	28	19	20	24	28	17	28	28	23
eth	24	17	18	17	18	16	16	15	19	18	17	17	18	18
	23	19	19	16	19	19	16	16	20	21	16	20	21	19

Figure 4. Raw Percent Agreement, Overall

A model is what a model is—it does not make mistakes, as such (those are left to the author), so chance is neither here nor there. Rather, the threshold of chance is better thought of as a point of uncertainty. That is, assuming that the models perform flawlessly in terms of measuring what they were intended to measure, there will nevertheless remain a limit to how much certainty we can attach to a measure of agreement. The point of uncertainty persists because there is always a certain range of data in which we cannot tell the difference between modelling decisions and noise. Here that range falls around the point of uncertainty—divergence from the threshold indicates some possible relationship in the models at hand, but the closer a score to the point, the less we can trust it. There is no shortage of measures that seek to rectify this situation, nor of debate on those measures. In the main these debates hinge on the statistical characteristics of measures ; the grounding statistical assumptions of these measures however do not map clearly to observed social phenomena. So perhaps there is an easier solution for applications to social inquiry: don't worry about it. This is what we choose to do (or not to do) here. To account for the point of uncertainty in the raw α_o scores, we simply subtract the expected level of uncertainty according to the degree of *k*—scores consistently below the point of uncertainty nonetheless indicate some relationship between the models at hand, it is simply unclear what it is. As noted above, we cannot consider α_0 as a linear measure when applied to cases with arbitrary class labels. Increasing α_0 (i.e. above the point of uncertainty) suggests model convergence, and we know that convergence is in the direction of the reference assignment. Recall that these tables anchor the cluster assignments to the V_{SD} sets. Thus in cluster *k*X.1, all V_{SD} sets (`base`, `age`, `inc`, etc.) are set to 1, in *k*X.2 all are set to 2, and so on. But for decreasing α_0, while we might surmise model divergence, we do not know its direction and the resolution of that end of the scale is a fraction of the convergence end—quite literally a fraction, as convergence is attributable to 1 label out of *k*, and divergence to *k*-1 labels. But, as already noted, measures such as α_o were not developed for this use case, so we must make do. In any case, α_o provides a valuable heuristic for the purpose here.

271. The adjusted α_o scores are presented in **fig. 5**. Comparison is within clusterings, rather than overall as in **fig. 4**. The only differences from **fig. 3** is that scores within each cluster have been down-weighted according to the appropriate point of uncertainty, and the approach to score comparison (as shown with shading) has been altered. Scores can now be negative—as discussed, α_0 below the point of uncertainty carries useful information. The darker shading (presenting in green) used in the preceding figures again indicates those cells in the top 10% of scores in a clustering, while the darkest shading (presenting in blue) shows those cells in the bottom 10% of scores. Having adjusted the α_o scores in this manner, we get a better idea of how the V_{SD} and V_L models compare. However keep in mind that these scores are not directly comparable, as each clustering has a different zero point. Scores in *k*2 have a possible range of 50 to -50, in *k*3 77 to -33,



k2.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	10	22	12	-4	5	7	-1	-5	11	4	-5	11	4	5
age	7	17	9	-4	4	6	-1	-4	8	3	-4	8	3	4
inc	9	21	11	-4	4	6	-1	-5	10	3	-5	10	3	5
edu	9	21	10	-5	4	6	-2	-6	10	2	-6	10	2	4
eth	9	14	8	-2	2	4	-1	-3	8	2	-3	8	2	4
	9	19	10	-4	4	6	-1	-5	9	3	-5	9	3	4

k2.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-5	-29	-2	14	2	6	10	15	-1	8	15	-1	8	3
age	-6	-3	0	-3	-2	2	-1	-2	-1	0	-2	-1	0	-1
inc	-3	-19	0	10	1	5	7	11	1	7	12	1	7	3
edu	-6	-21	-1	9	2	5	7	11	0	7	10	-1	7	2
eth	5	12	0	-4	0	-4	-3	-4	0	-4	-4	0	-4	-1
	-3	-12	-1	5	1	3	4	6	0	4	6	0	4	1

k3.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	2	26	-1	-8	-6	-2	1	-8	-2	-4	-8	-2	-4	-1
age	2	-3	-1	-1	1	-2	2	-1	-1	-1	-1	-1	-1	-1
inc	1	15	0	-4	-1	0	1	-5	0	-1	-4	0	-1	0
edu	-5	-12	0	3	1	0	3	2	1	1	2	0	2	0
eth	7	0	-3	-1	0	-4	-2	-1	-3	-4	-1	-3	-4	-2
	1	5	-1	-2	-1	-2	1	-3	-1	-2	-2	-1	-2	-1

k3.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-1	-4	5	9	6	7	11	9	6	8	8	6	8	6
age	-1	-4	5	8	6	8	11	9	6	9	8	6	8	6
inc	0	-2	5	8	5	7	11	8	6	8	8	6	8	6
edu	4	-20	-3	7	0	-4	-7	8	-3	-3	8	-3	-3	-1
eth	3	-1	7	11	5	8	9	11	7	9	11	7	9	7
	1	-6	4	9	4	5	7	9	4	6	9	4	6	5

k3.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	9	-14	5	11	5	5	-4	11	5	6	12	5	6	5
age	11	-5	3	8	1	4	1	8	3	4	8	3	3	4
inc	8	-12	6	11	6	6	-1	11	6	7	12	6	7	6
edu	1	30	1	-7	-5	1	4	-7	1	-1	-8	1	-1	1
eth	9	-7	5	9	3	4	-2	9	5	5	9	5	5	5
	7	-1	4	6	2	4	0	6	4	4	7	4	4	4

k5.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	5	-6	-7	-6	-5	5	6	0	-7	10	1	0	10	1
age	5	-6	-7	-6	-5	5	6	0	-7	10	1	0	10	1
inc	-9	-12	2	0	-6	2	0	-2	0	1	4	-1	2	-1
edu	-2	14	6	10	7	2	10	9	6	-2	13	4	-2	6
eth	4	-6	-4	0	-2	6	-1	2	-3	7	-2	2	7	1
	1	-3	-2	0	-2	4	4	2	-2	5	3	1	5	1

k5.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	2	25	11	15	14	5	9	13	11	1	14	8	1	10
age	3	23	11	14	13	5	8	12	11	1	13	8	1	10
inc	-7	-8	3	-2	-2	4	0	1	2	4	3	1	4	0
edu	4	1	-5	-2	-6	2	-1	-2	-4	6	-3	1	6	0
eth	1	13	10	12	8	6	7	10	10	4	11	9	4	8
	1	11	6	8	5	5	5	7	6	3	7	6	3	6

k5.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	6	-3	10	9	9	14	0	8	10	13	2	11	13	8
age	0	-10	3	3	1	3	2	0	4	7	0	5	7	2
inc	-4	34	-3	2	5	-7	3	0	-3	-10	1	-5	-10	0
edu	1	-11	0	-4	-2	3	-3	-3	-2	1	0	0	2	-1
eth	7	-4	6	7	6	11	-2	5	6	10	1	7	10	5
	2	1	3	3	4	5	0	2	3	4	1	4	4	3

k5.4	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-7	-8	3	1	-4	3	3	0	-1	3	6	1	3	0
age	-2	15	3	5	6	-3	3	4	3	-5	4	1	-5	2
inc	0	0	3	5	2	3	1	2	3	2	-1	2	2	2
edu	-3	27	4	7	8	-3	6	6	5	-6	6	2	-6	4
eth	-1	-7	-3	-4	-7	-2	-1	-1	-4	1	0	-2	2	-2
	-3	5	2	3	1	0	2	2	1	-1	3	1	-1	1

k5.5	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	14	-4	-5	-10	1	-9	-9	-8	-4	-7	-8	-7	-7	-5
age	-1	3	0	-6	-4	-5	-3	-6	0	-3	-6	-2	-3	-3
inc	0	-6	1	-3	-2	5	-1	-1	1	7	-1	4	8	1
edu	-2	4	4	1	2	8	-1	0	4	8	-3	8	8	3
eth	4	-3	-2	-3	-2	-4	-4	-5	-1	-2	-3	-3	-2	-2
	3	-1	-1	-4	-1	-1	-4	-4	0	1	-4	0	1	-1

Figure 5. Adjusted Percent Agreement, by Clustering

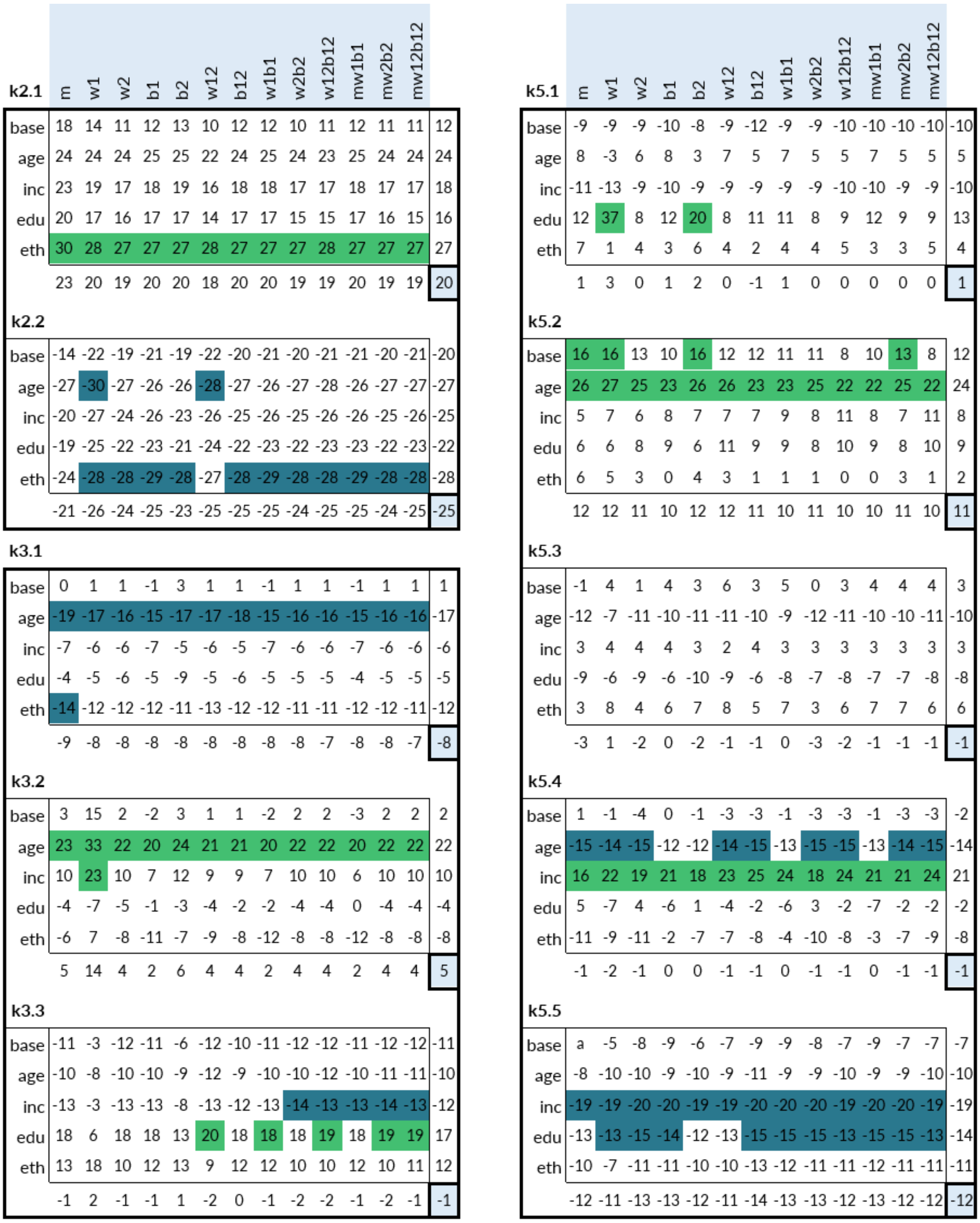
and in *k*5 80 to -20. This is not intended to mislead—the scale offsets are due to the relevant points of uncertainty. Adjusting these scales to a common zero-point is not an option—as noted, because of arbitrary class labels, the scales are not linear, and the point of uncertainty changes with *k*. If one were to normalise these scores (e.g. dividing each by its possible range, depending on sign, then multiplying by 100) this would put the clusterings on a common numerical scale, but the scales would differ in terms of sensitivity according to k and to sign. That is, in the case of normalisation on the positive components, each point increase in a raw score would result in a k2 increase of 2, a k3 increase of 1.3, and a k5 increase of 1.25. Such normalisation and the resultant differences in scale sensitivities would be the more misleading case, as it would obscure the inherent limitations of uncertainty present in a given k clustering. So we do not do that.

272. Individual pair-wise scores will not be discussed, as they do not mean much in themselves. Rather, what is important to observe in **fig. 5** foremost is the continued indications of model convergence and divergence. As discussed, the nature of these adjusted scales make any quantitative comparison of the clusterings problematic—these figures are essentially visualisations. Yet that is important in itself: the import is that we do see some degree of connectedness between the V_{SD} and V_L sets, that is between structural models and linguistic models of the user–documents. However, these tables have used the V_{SD} sets as the cluster anchors ; recall that places were clustered according to their V_{SD} profiles separately from users, and those cluster assignments were then assigned to the user–documents according to their single associated place. In that manner, the V_{SD} sets characterise places and the V_L sets characterise users. As noted in **¶ 265**, in a given place we expect to find all V_L cluster assignments. And thus these tables are very messy.

273. Places are associated with only one V_{SD} profile, but with many possible users and thus V_L profiles. This is important to keep in mind in the following steps—at a given level of k, each place is composed of a varying proportion of all k. Also recall that these tables present α_o scores according to the cluster assignments of places, as the V_{SD} cluster assignments are derived from the sociodemographic profiles of the places in the dataset. Thus each cell gives a summary α_o score across all users assigned to a subset of places regardless of their V_L set cluster assignments. This approach is taken due to an essential limitation of the method—its spatial resolution, and thus V_{SD} set resolution, is limited to codified places ; we cannot peer into neighbourhoods and communities. So places, and thus V_{SD} set `base` during alignment, provide the anchor, and we intuit the composition of a given place by its proportions of V_L set assignments (i.e. linguistic cluster assignments). So, where we observe clusters with high internal α_o, we might surmise that a place—having a fixed V_{SD} score—has an increased incidence of users sharing similar V_L scores.

High α_o places have decreased V_L variation, and low α_o places have increased V_L variation. That is, α_o will increase as commonality in speech increases. The rationale is that each for each cluster, the α_o score of each cell is calculated against a stable assignment (i.e. the cluster V_{SD} assignments). Thus, a higher α_o possibly indicates a more stable V_L assignment, with increasing V_L variety pulling the α_o score towards the point of uncertainty (which in **fig. 5** is zero).

274. To ameliorate the messy situation in these tables, **fig. 6** presents the α_0 scores recalculated by transposing the cluster assignment anchors. Whereas **fig. 3–5** anchored cluster assignments to the V_{SD} sets, as just noted, **fig. 6** anchors them to the V_L sets. Thus within each cluster, a given V_L column will represent a subset of the user–documents sharing the label of the cluster degree. So in `k5.3` (i.e. cluster 3 within clustering *k*5), column `b1` (containing the key bigrams from across all user–documents ; see **chp. 5 § 5.1.3.2**) is a subset of user–documents with the `b1` cluster label `3`. Those assignments are then compared to the V_{SD} assignments of that subset. For example, in `k5.3` the V_L set `b1` and the V_{SD} set `age` show an adjusted α_0 of -10, thus a raw α_0 of 10 at *k*5. If the age cluster assignments had tended towards the label 3, we would show an adjusted α_0 above 0. An adjusted α_0 of -10 shows a tendency away from label 3.[300] The horizontal striping that can be seen in **fig. 6** is perhaps an indication of some relationship between the V_{SD} and V_L sets. However, as noted in **n. 300**, a more immediate explanation is found in the skew of cluster assignments, and that issue brings this step to a halt. Both variable sets are highly skewed ; in V_{SD} at *k*5 the ratio of cluster assignment prevalence is roughly 5:3:3:2:1, and in V_L it is roughly 5:2:2:2:1. As the calculation of α_0 does not account for label prevalence (cf. Feng, 2013), and considering both the weak signal coming from pervasive features and that generally all V_L will be found in all V_{SD}, it is safest to assume that these tables predominantly reflect relative label prevalence, with higher α_0 marking more prevalent labels and lower α_0 marking less prevalent labels. In hindsight, it is of no use trying to visualise connectedness of the variable sets in this manner, as we will only see skew in label prevalence (which is useful, but was not our purpose). In terms of its intended purpose, this step must be deemed a failure. But it is not a complete failure—the tendency of α_0 to respond to skewness suggests that, all things equal, the alignment process (which hinges on α_0, as explained) would tend to match clusters of similar prevalence. This tendency is observed in the dataset when comparing the permutations of aligned and



k2.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	18	14	11	12	13	10	12	12	10	11	12	11	11	12
age	24	24	24	25	25	22	24	25	24	23	25	24	24	24
inc	23	19	17	18	19	16	18	18	17	17	18	17	17	18
edu	20	17	16	17	17	14	17	17	15	15	17	16	15	16
eth	30	28	27	27	27	28	27	27	27	28	27	27	27	27
	23	20	19	20	20	18	20	20	19	19	20	19	19	20

k2.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-14	-22	-19	-21	-19	-22	-20	-21	-20	-21	-21	-20	-21	-20
age	-27	-30	-27	-26	-26	-28	-27	-26	-27	-28	-26	-27	-27	-27
inc	-20	-27	-24	-26	-23	-26	-25	-26	-25	-26	-26	-25	-26	-25
edu	-19	-25	-22	-23	-21	-24	-22	-23	-22	-23	-23	-22	-23	-22
eth	-24	-28	-28	-29	-28	-27	-28	-29	-28	-28	-29	-28	-28	-28
	-21	-26	-24	-25	-23	-25	-25	-25	-24	-25	-25	-24	-25	-25

k3.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	0	1	1	-1	3	1	1	-1	1	1	-1	1	1	1
age	-19	-17	-16	-15	-17	-17	-18	-15	-16	-16	-15	-16	-16	-17
inc	-7	-6	-6	-7	-5	-6	-5	-7	-6	-6	-7	-6	-6	-6
edu	-4	-5	-6	-5	-9	-5	-6	-5	-5	-5	-4	-5	-5	-5
eth	-14	-12	-12	-12	-11	-13	-12	-12	-11	-11	-12	-12	-11	-12
	-9	-8	-8	-8	-8	-8	-8	-8	-8	-7	-8	-8	-7	-8

k3.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	3	15	2	-2	3	1	1	-2	2	2	-3	2	2	2
age	23	33	22	20	24	21	21	20	22	22	20	22	22	22
inc	10	23	10	7	12	9	9	7	10	10	6	10	10	10
edu	-4	-7	-5	-1	-3	-4	-2	-2	-4	-4	0	-4	-4	-4
eth	-6	7	-8	-11	-7	-9	-8	-12	-8	-8	-12	-8	-8	-8
	5	14	4	2	6	4	4	2	4	4	2	4	4	5

k3.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-11	-3	-12	-11	-6	-12	-10	-11	-12	-12	-11	-12	-12	-11
age	-10	-8	-10	-10	-9	-12	-9	-10	-10	-12	-10	-11	-11	-10
inc	-13	-3	-13	-13	-8	-13	-12	-13	-14	-13	-13	-14	-13	-12
edu	18	6	18	18	13	20	18	18	18	19	18	19	19	17
eth	13	18	10	12	13	9	12	12	10	10	12	10	11	12
	-1	2	-1	-1	1	-2	0	-1	-2	-2	-1	-2	-1	-1

k5.1	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-9	-9	-9	-10	-8	-9	-12	-9	-9	-10	-10	-10	-10	-10
age	8	-3	6	8	3	7	5	7	5	5	7	5	5	5
inc	-11	-13	-9	-10	-9	-9	-9	-9	-9	-10	-10	-9	-9	-10
edu	12	37	8	12	20	8	11	11	8	9	12	9	9	13
eth	7	1	4	3	6	4	2	4	4	5	3	3	5	4
	1	3	0	1	2	0	-1	1	0	0	0	0	0	1

k5.2	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	16	16	13	10	16	12	12	11	11	8	10	13	8	12
age	26	27	25	23	26	26	23	23	25	22	22	25	22	24
inc	5	7	6	8	7	7	7	9	8	11	8	7	11	8
edu	6	6	8	9	6	11	9	9	8	10	9	8	10	9
eth	6	5	3	0	4	3	1	1	1	0	0	3	1	2
	12	12	11	10	12	12	11	10	11	10	10	11	10	11

k5.3	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	-1	4	1	4	3	6	3	5	0	3	4	4	4	3
age	-12	-7	-11	-10	-11	-11	-10	-9	-12	-11	-10	-10	-11	-10
inc	3	4	4	4	3	2	4	3	3	3	3	3	3	3
edu	-9	-6	-9	-6	-10	-9	-6	-8	-7	-8	-7	-7	-8	-8
eth	3	8	4	6	7	8	5	7	3	6	7	7	6	6
	-3	1	-2	0	-2	-1	-1	0	-3	-2	-1	-1	-1	-1

k5.4	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	1	-1	-4	0	-1	-3	-3	-1	-3	-3	-1	-3	-3	-2
age	-15	-14	-15	-12	-12	-14	-15	-13	-15	-15	-13	-14	-15	-14
inc	16	22	19	21	18	23	25	24	18	24	21	21	24	21
edu	5	-7	4	-6	1	-4	-2	-6	3	-2	-7	-2	-2	-2
eth	-11	-9	-11	-2	-7	-7	-8	-4	-10	-8	-3	-7	-9	-8
	-1	-2	-1	0	0	-1	-1	0	-1	-1	0	-1	-1	-1

k5.5	m	w1	w2	b1	b2	w12	b12	w1b1	w2b2	w12b12	mw1b1	mw2b2	mw12b12	
base	a	-5	-8	-9	-6	-7	-9	-9	-8	-7	-9	-7	-7	-7
age	-8	-10	-10	-9	-10	-9	-11	-9	-9	-10	-9	-9	-10	-10
inc	-19	-19	-20	-20	-19	-19	-20	-20	-20	-19	-20	-20	-19	-19
edu	-13	-13	-15	-14	-12	-13	-15	-15	-15	-13	-15	-15	-13	-14
eth	-10	-7	-11	-11	-10	-10	-13	-12	-11	-11	-12	-11	-11	-11
	-12	-11	-13	-13	-12	-11	-14	-13	-13	-12	-13	-12	-12	-12

Figure 6. Adjusted Percent Agreement, Transposed

unaligned clusterings. Furthermore, the essential learning of the step is it is essential to account for the relative prevalence of clusters in this analysis. The following steps will do that.

275. In the end, step 1—gauging connectedness of the V_{SD} and V_L sets by means of α_0—cannot serve its intended purpose. Without accounting for skewness in cluster label prevalence, we cannot say with any certainty what these tables tell us about the possible connectedness of the variable sets. It seems we were a bit cavalier to say 'don't worry' about chance, as that issue is intimately connected with skew. Nevertheless the step is not in vain. Steps 2 and 3 attend to cluster proportions within places ; the critical influence of cluster label prevalence observed in step 1 highlights the need to account for proportions. Furthermore, the use of α_0 for aligning cluster labels according to prevalence is of potential value. In any case, a clearer picture of the situation can be provided by examining the V_{SD} profiles of the V_L clusters directly. We turn to that now.

6.2.2. Step 2 – Sociodemographic Profiles

276. In order to answer **RQ**$_2$ in the affirmative, we must be able to view linguistic groupings in terms of their sociodemographic profiles. For that reason, each user–document is associated to a single place, the sociodemographic profile of which is known. As each user–document also has an individual linguistic profile, clustering the members of the dataset by linguistic features produces a certain set of places. A composite sociodemographic profile of this set of places can then be produced. By comparing these sociodemographic profiles across linguistic clusters, we can assess the potential connectedness between language and social structure at scale, and thus provide an initial answer to the second research question.

277. The analysis will be presented in the following manner. The charts to follow have a common format. The horizontal axis is categorical, composed of V_{SD} set brackets, as will be described in a moment. The vertical axis displays the percentage of population, each data point in the sociodemographic profiles being summarised in this manner. Consider a cluster associated with one single place. If 25% of the population of that place were recorded by the US Census as middle income, then that cluster would show a mark at 25% on the vertical axis in the middle income column. However, the actual clusters will each comprise most of the remaining places in the dataset (84, after the final filtering). A summary value for any of the V_{SD} brackets would obscure much information, so to facilitate comparison of V_L clusters in V_{SD} terms, we will make use of box plots (McGill, Tukey and Larsen, 1978) to give a better sense of the profile of each cluster. To further facilitate comparison, this step will focus on clustering *k*3 , with a few examples included from clustering *k*5.

278. The V_{SD} brackets are derived from the data comprised by the V_{SD} sets. There are four primary sets: age, income, education, and ethnicity (the composite set `base`, comprising age, income, and education, was used only for clustering). As noted above, all V_{SD} data points characterise a single named place, and are expressed in terms of population percentage estimates (see **chp. 5 § 5.1.1.2**). The age data (V_{SD} set `age`) as retrieved from the US Census comprises 11 brackets, these are collapsed here to 4. The income data (V_{SD} set `inc`) comprises 10 brackets, collapsed to 5 ; note that income is reported by household. The education data (V_{SD} set `edu`) comprises 7 brackets, collapsed to 4. The ethnicity data (V_{SD} set `eth`) comprises 4 separate measures, and thus is not collapsed. Those measures are the portions of the population identifying as White and identifying as Black, the proportion born in Michigan, and the proportion of households speaking only English at home ; note that the two largest primary ethnic groups in Michigan are White (c.74%) and Black (c.14%), thus these are the only two included in the variable set. The V_{SD} brackets comprise the horizontal axis of the following charts. Their labels and denotations are as follows, recalling that each is expressed in terms of percentage of the population.

Age - Minor	19 years and younger
Age - YA	20–34 years
Age - Adult	35–54 years
Age - Senior	55 years and over
Inc - Poverty	$25k p.a. and below ; 2018 US Poverty Guideline, 4 persons
Inc - Low	$25–50k p.a.
Inc - Middle	$50–100k p.a. ; 2018 MI median household income was $57k p.a.
Inc - High	$100–200k p.a.
Inc - Wealthy	$200k p.a. and above
Edu - < 2nd	up to 12th grade education, no graduation
Edu - >= 2nd	secondary school graduate, up to associate degree (2 yr)
Edu - Grad	bachelor degree (4 yr)
Edu - Postgrad	postgraduate or professional degree
Eth - W	identifying as White
Eth - B	identifying as Black
Eth - MI born	born in Michigan
Eth - Eng. only	households speaking only English at home

279. There are challenges to the comparison being made in this step. As has been discussed, the use of pervasive linguistic features in this work generally results in all *V_L* clusters being represented in any given place. For that reason, this step must take the proportion of representation into account. Comparing the V_{SD} profiles of V_L clusters in the manner described at the beginning of this section—without accounting for the relative presence of clusters—results in clusters that are practically indistinguishable. To illustrate the situation, we shall compare the age and education profiles of *k*3 clusterings across all places on the V_{SD} `base` assignment versus on the V_L `w2` assignment (see **fig. 7** and **fig. 8**, respectively). We will also take the opportunity to suggest how these charts can be understood.

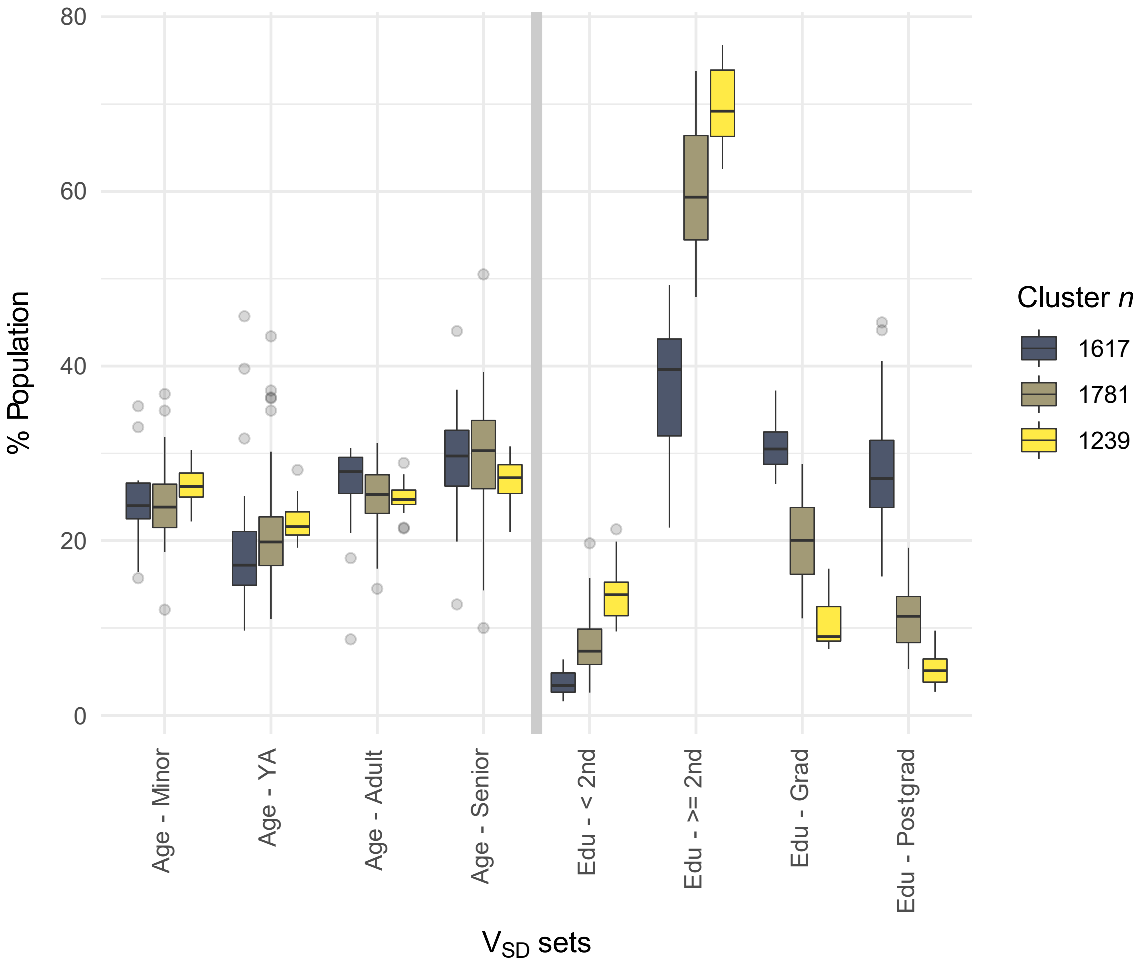


Figure 7. V_{SD} Aggregate Profiles – *k*3, clustered by V_{SD} set `base`

280. **Figure 7** shows the *k*3 clustering on the V_{SD} `base` assignment. As explained, cluster labels are arbitrary. Clusters are distinguished by shading, and are arranged in each V_{SD} column in the same order. These clusters are aligned on the population percentage of secondary school graduates. The legend (labelled Cluster *n*) shows the membership count of the cluster (in this case, 1,617, 1,781, and 1,239).[301] The V_{SD} profiles of each cluster are represented by standard box plots. Thus in each case the bounding box shows the interquartile range, the central horizontal bar marks the median value, the whiskers indicate the extent of data points within 1.5 times the interquartile range, and any outliers beyond that extent are marked by dots.

281. Examining the age profiles, it can be seen that the first cluster (in the darkest shading) has the lowest proportion of young adults and the highest of adults. The second cluster (with the medium shading) has a greater proportion of young adults, lesser of adults, and the largest of seniors. The third cluster (with the lightest shading) has the highest proportion of minors and young adults, and the lowest proportion of adults and seniors. Note that all of the V_{SD} sets in these charts can be summed within a given set, as they indicate proportions of the whole, and the set brackets cover all ranges (i.e. the brackets within a given V_{SD} set cover the entire population of the places they profile). Examining the education profiles, the first cluster has the lowest proportion of population having—at most—a secondary education or equivalent, and the highest proportion of college graduates and those with postgraduate or professional degrees. The third cluster is the opposite case, with the highest proportion of secondary school graduates and below, and the lowest proportion of college graduates and postgraduates. Note again that brackets with a V_{SD} set are zero-sum—an increase in one bracket must offset by an equivalent decrease elsewhere, and vice versa. Taking these two V_{SD} sets together, we can begin to envision a profile of three types of places. Cluster 1, representing 1,617 user–documents from 46 places (see **n. 301**) with a total population of 1,600,000 people, comprises what is likely the suburban and exurban areas of the lower 33 counties. The dip in proportion of young adults (aged 20–34) and the peak in adults (aged 35–44), as well as the far higher proportion of college graduates and postgraduates, suggests established families outside of the main cities. Places with a higher concentration of families with young children are suggested by the outliers observed in the Minor bracket, while the outliers in the Young Adult bracket are liable to be among Michigan’s college towns. Cluster 2, representing 1,781 user–documents from 19 places with a total population of 1,500,000 people, likely comprises these main cities. We see a decided jump in the proportion of young adults, as well as the highest proportion of seniors—this could be understood as the result of youth moving to the cities from outlying areas, and people originally from the cities remaining there throughout their lives. However, the profiles seen here could also be understood as balancing out relatively advantaged and disadvantaged built-up regions. Likely it is a mixture of these cases, that would need further data (e.g. the income and ethnicity sets) and mapping to resolve. Education follows a pattern that could be fit to youth leaving home for studies, or a mixture of advantaged and disadvantaged areas. Regardless, odds are good that we are seeing the main cities in this cluster. Cluster 3, representing 1,239 user–documents from 19 places with a total population of 811,730, most likely comprises disadvantage urban and exurban areas. It has the highest proportion of minors and young adults, and the lowest of adults and a sharp drop in seniors—these are young people and families with children, those who can move away, and lifespans are likely to be depressed compared to the other clusters. That interpretation is supported by this cluster having the lowest level of education among the clusters.

282. All this is supposition, of course. The suggested character of the places represented by these clusters would need to be warranted by further data, and as well as mapping to literally ground these interpretations. Nevertheless, we can intuit these places from sociodemographic data. In the broad context of Western society, we are socialised into this manner of thinking—not only are social phenomena discussed constantly in terms of sociodemographic measures, but they are also quite real and we witness these relationships with our own eyes. Yet, as this work has argued, this is perhaps not the best or only way to see the social world. The argument here is that we can construct a different, yet nonetheless coherent, image of society by means of endogenous measures derived from language. Here we have interpreted the clustering of sociodemographic data in sociodemographic terms—not a horribly daunting task, but useful in terms of showing how the remainder of this step will proceed. We now take this same approach to interpretation of profiles produced solely from the clustering of linguistic variables.

283. **Figure** 8 shows the *k*3 clustering on the V_L `w2` assignment, across <u>all</u> places. Note that whereas the `w1` clustering (described in the previous subsection) is derived from a list of the compiled key words of each user–document, the `w2` clustering is derived from a list of the compiled <u>non</u>-key (i.e. non-descript) words of each user–document—that is, these are words that do not stand out in terms of frequency (either high or low) compared to the overall corpus, but rather just blend in with it. These would be our habitual, unconscious words. Non-key linguistic features of this sort are more rightly the pervasive features that this work is interested in, compared to the key pervasive features in the List 1 group of words and bigrams (see **chp. 5 § 5.1.3.2**).

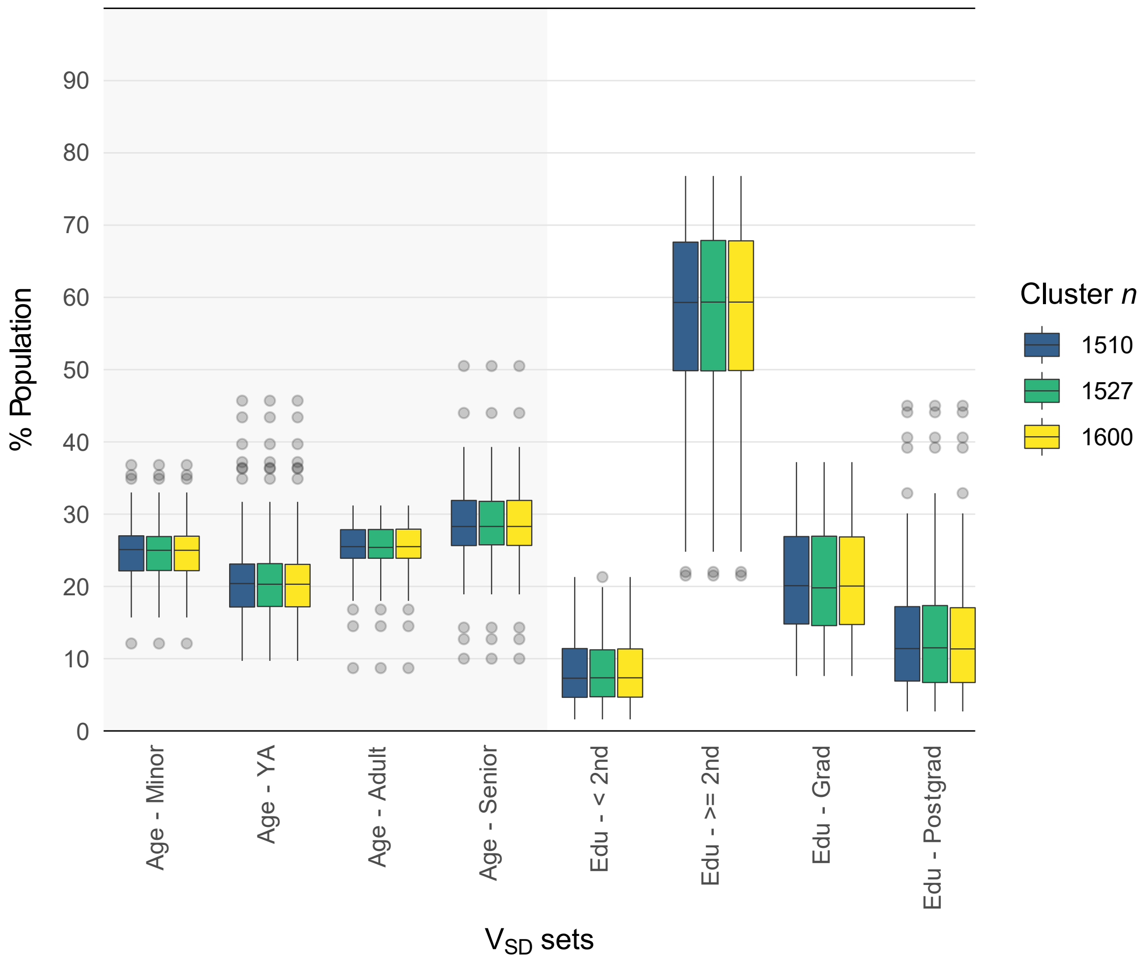


Figure 8. V_{SD} Aggregate Profiles – *k*3, clustered by V_L set w2 (all places)

284. Recall that **fig. 7** plotted 84 places—this chart plots 4,637 user–documents. Moreover, it plots the profiles of the clusters without accounting for cluster proportion in a given place. The most important thing to note is that, whereas the clusters in **fig. 7** were naturally distinct, the clusters in **fig. 8** are practically identical. This is the challenge of pervasive features, of all V_L cluster assignments being observed in all places—we can see nothing distinctly beyond the relative prevalence of cluster labels. As we are seeing V_{SD} data for nearly all places in each cluster, the age and education profile(s) that we see here can be understood as approximating a summary of the final dataset. The clusters assignments are well balanced. Cluster 1 represents 1,510 user–documents from 82 places, cluster 2 represents 1,527 user–documents from 83 places, and cluster 3 represents 1,600 user–documents also from 84 places. Cluster 3 is coextensive with the final dataset, representing the full set of places, which has a total population of 4,000,000 (a total shared by clusters 1 and 2 within about 1%), so it is not an approximation but rather an actual summary of the final dataset. Note that V_L set `w1` produces comparable clusterings in terms of number of places and total population, but with a highly skewed cluster prevalence (having a ratio of roughly 7:3:1).

285. The manner in which we will account for cluster assignment proportions across places is simple—we will limit our focus to subsets of places having proportions of a given cluster assignment above a certain threshold. That is, for the entire dataset, we give each place a score that is equivalent to the percentage of user–documents associated with that place that have been given a certain V_L cluster assignment. For inclusion in analysis, any given place must exceed the threshold for a given cluster (thus a single place can be represented in any or none of the place sets associated with each cluster). **Figure 8** can be understood as presenting the data with a proportion threshold of 0%. However, by increasing that threshold, and so removing the most 'unlike' places in a given cluster, we can begin to develop a sense of how sociodemographic and linguistic data might interact across the user–documents. Note that if we approached this problem by taking only the most representative places, so hoping to assemble profiles of the most linguistically 'typical' places for each cluster, we would tend towards a very small *n* for each cluster if we interpret 'most representative' in a strict manner (say, the top 10 places in each cluster, or the top 20% in each cluster).[302] Also, a given V_L set is likely to have skewed cluster prevalence, making the determination of 'the top' places a bit tricky, given that some of the top places in a given cluster might nevertheless have a low proportion score for that cluster label. Moreover, from a social perspective, it is logical to work from the bottom up—that is, by means of applying this threshold. Focusing only on the most linguistically typical is misleading. We are influenced by all language around us, regardless of how similar it is to our own. The threshold allows us to remove only those places where a given cluster is in a distinct minority, so that we may have a better sense of those places where the cluster is more present.

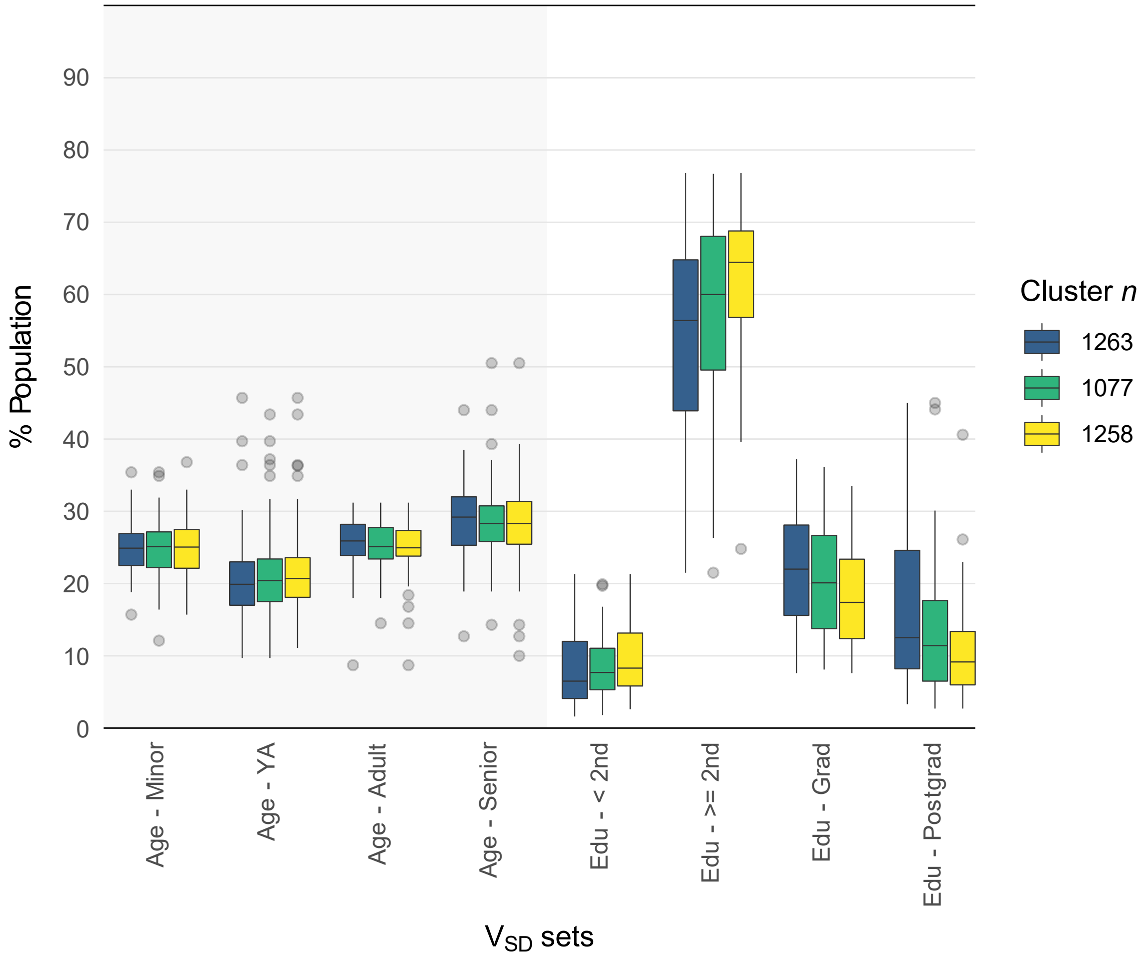


Figure 9. V_{SD} Aggregate Profiles – *k*3, clustered on V_L set `w2` (with threshold)

286. **Figure 9** reproduces the comparison in **fig. 8**, except now applying a cluster proportion threshold of 30% ; that threshold was selected as it is a round number close to the point of uncertainty, discussed in step 1. That is, for a place to be included in the V_{SD} profile of a given cluster, it must have at least 30% of its associated user–documents assigned to that cluster. All clusters taken together still represent all 84 places in the final dataset (this would not always be the case, depending on the threshold and the dataset), but now they are represented by a different distribution of user–documents compared to **fig. 8**. Cluster 1 represents 1,263 user–documents from 47 places having a total population of 2,700,00, cluster 2 represents 1,077 user–documents from 53 places having a total population of 2,300,000, and cluster 3 represents 1,258 user–documents from 54 places having a total population of 2,900,000.[303] Skew in cluster prevalence has increased slightly, with cluster 2 showing a relative reduction compared to the others. Of interest is the range of cluster proportions: while the threshold is set at 30%, the range of proportions differs among the clusters. Cluster 1 ranges up to 67% with a median of 39%, cluster 2 up to 82% with a median of 41%, and cluster 3 up to 75% with a median of 40%.

287. Overall, the clustering is fairly well balanced in terms of assignment prevalence amongst user–documents and in terms of the count and population of the places represented. At the same time, these relatively balanced clusters differ in terms of their V_{SD} profiles. On the one hand, the age profiles are relatively undifferentiated. This is encouraging—recall that one of the reasons that pervasive features were selected for this work is that they tend to be more stable through the lifecourse. Thus V_L clusters having similar age profiles is what we would expect (or at least hope) to see here. On the other hand, the education profiles show a pattern broadly similar to the V_{SD} clustering shown in **fig. 7**. Cluster 1 shows a higher level of educational attainment, cluster 3 the lowest, and cluster 2 in between (recall that these clusters are aligned on the secondary education column). The differentiation of the education profiles of the clusters is not as strong as in **fig. 7**, but then that plot represented clustering on the V_{SD} set `base`, which is a composite variable including education data. **Figure 9**, however, shows clusters based solely on frequency of use of non-descript words.

288. We will not continue with an interpretation of this chart, as was done with **fig. 7**. **Figures 8** and **9** were presented only to show that V_L clusters can be distinguished in V_{SD} terms, but only if we account for the relative proportions of cluster assignments across places. Having established the basic process of this step, we now consider a clustering profiled across all V_{SD} sets. This will provide a much richer case for interpretation, compared to the two-set comparison just performed. We will continue with the *k*3 clustering on the V_L set `w2`.

289. **Figure 10** presents the same clustering as **fig. 9**. The only difference being that the V_{SD} sets on income and ethnicity have been added for further contextualisation. The age and education profiles remain as before, and the cluster *n* distribution remains the same. We can see further differentials in the income profiles. Cluster 1 not only shows a profile with a higher relative percentage of people with higher educational attainment, but also of people with higher incomes. Cluster 3 is the inverse and cluster 2 sits in between again. Thus we see educational attainment positively associated with income, broadly speaking.

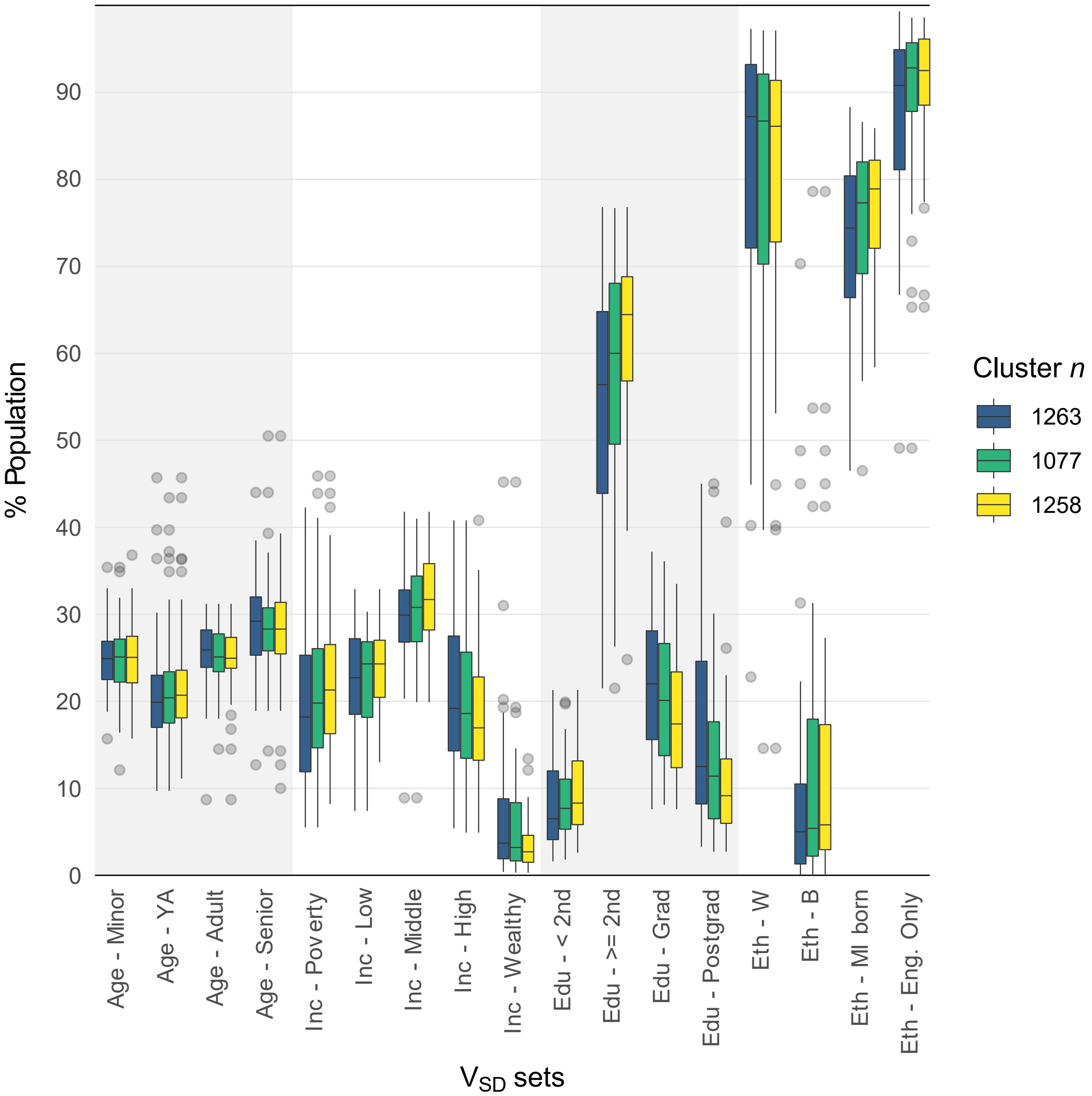


Figure 10. All V_{SD} Aggregate Profiles – *k*3, clustered on V_L set `w2`

There are outliers, of course, but recall that these charts are plotting the profiles of places—some places, especially in certain built-up environments, will show higher degrees of sociodemographic disparity. The cluster differentials observed in the income brackets are not so great, but they are nonetheless evident. More interesting differentials can be observed in the ethnicity profiles. In terms of general ethnic composition, that is, in terms of the relative percentages of the population identifying as White or Black, we can see the cluster medians are more or less the same. Michigan is predominantly White, so in aggregating data this would be expected. However, note the differences in the box plots in column `Eth-B` (percentage population identifying as Black)—the interquartile range in clusters 2 and 3 is much larger than in cluster 1, and there are quite a few outliers (those being predominantly Black communities). Note from the whiskers in the first two `eth` columns that there are no places that are completely White, but there are places where none of the population is Black. Notably, and unsurprisingly given the US context, the clusters (i.e. sets of places) having the profiles indicating a lower educational attainment and lower incomes are those showing a higher proportion of Black people. Also of interest is that cluster 1 shows a relative dip in the third `eth` column, showing the population percentage of people born in Michigan, and in the fourth column, showing the population percentage reporting that they speak only English at home. From these two components we might surmise that cluster 1 more likely includes people that have moved in from out of state for work or perhaps retirement (in both cases individuals likely to be of higher income and education), as well as educated immigrants coming for work (likely in academia and industry).[304]

290. The pictures becomes slightly more complicated if we look at the *k*5 clustering on V_L set `w2` (**fig. 11**). Here the proportion threshold has been reduced to 20% to account for the increase in possible cluster assignments from 3 to 5. Again, these clusters are aligned on secondary education (column Edu - >= 2nd). As before, all 84 places in the final dataset are represented, however only 3,041 user–documents are represented (as opposed to 3,598 in **fig. 9** ; recall that a given user–document will not be included if assigned to a cluster that is below threshold in their associated place). We have three large clusters (1,3, and 5), and two very small (2 and 4). Clusters 2 and 4 are unusual beyond their smallness: they have the lowest level of maximum cluster proportions (50% and 33%, respectively) ; their income profiles show a different education–income relationship compared to clusters 1 and 3 ; and notably they show the lowest proportions of people born in Michigan. The two clusters are also relatively young, having among the highest proportions of young adults and minors. (It is also worth noting that we can see increasing differentials in the age brackets at higher *k*, although these remain relatively small compared to other V_{SD} sets.) In the discussion of **fig. 10**, the presence of immigrants could be imputed from the data, and these were supposed to be those coming specifically to take up jobs (that being

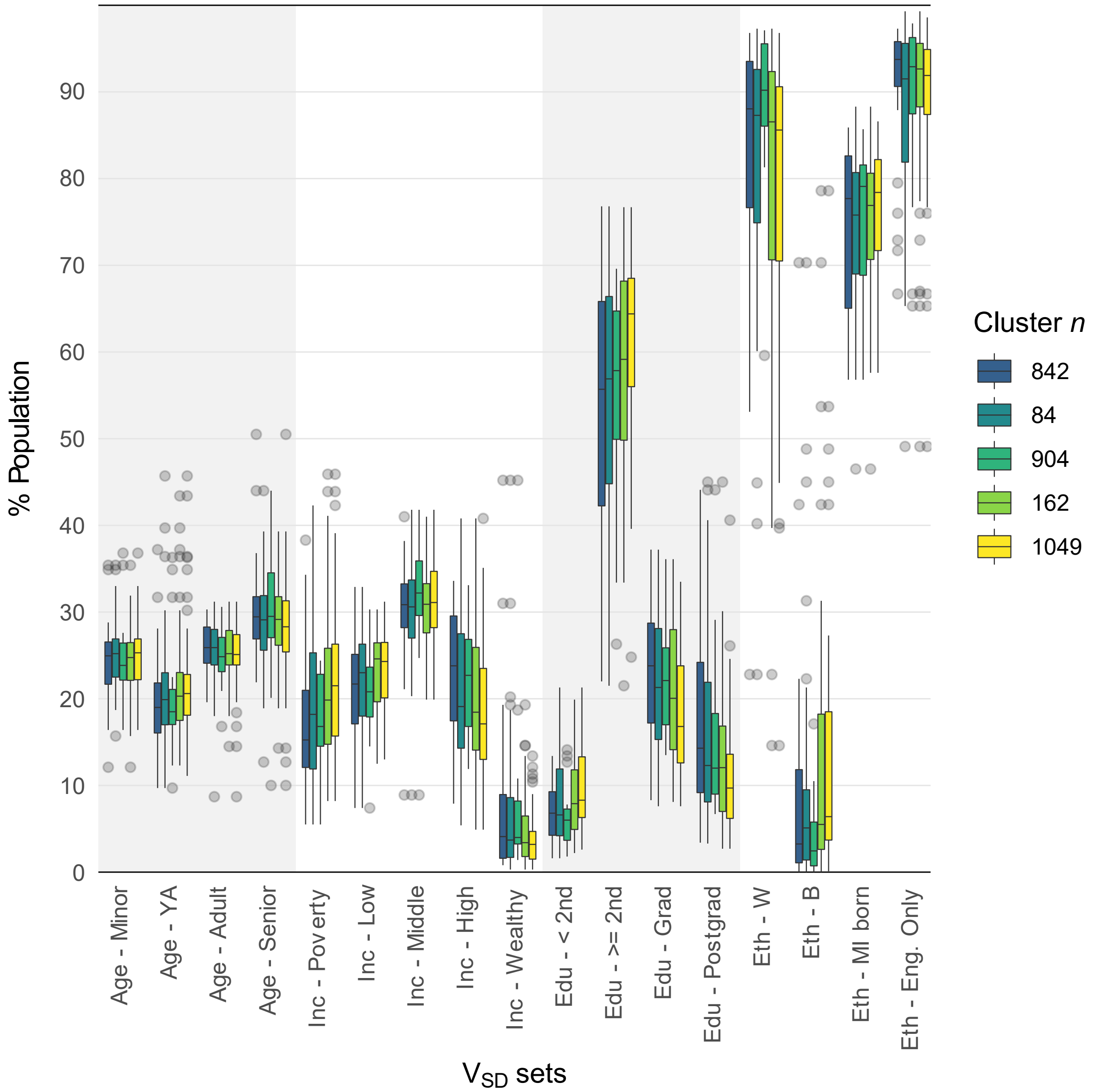


Figure 11. All V_{SD} Aggregate Profiles – *k*5, clustered on V_L set `w2`

simply one possible interpretation). Perhaps here we see a different sort of immigrant presence: once that is younger, less well off, and that (thus far) has less higher education.[305] These clusters might well be localised in urban centres or in university areas (although all clusters here are near the average population density of the places in the final dataset, and clusters 2 and 4 show the lower densities among the clusters). However, as with the discussion of **fig.** 7, all this is supposition. A closer inspection of the constituent places of each cluster would help to firm things up. Mapping of these clusters perhaps could help to firm up these suppositions and to make sense of things. But nevertheless we are limited by the method—the geographical resolution is limited to places as a whole, with summary measures of their V_{SD} profiles. While we can perhaps intuit something about the makeup of each place according to its proportions of V_L

clusters, that would be a tenuous and questionable step. Thankfully, that is not our goal, which simply is to demonstrate that endogenous measures based on language can provide an alternative view of society that nevertheless remains understandable from a structural perspective.

291. Step 2 allows us to say with some confidence that linguistic clusterings can be understood from a structural perspective. V_{SD} set `w2` was given here as the exemplar, as it seems to provide the clearest profile differentials amongst clusters in this step (the other sets are addressed in the following section). Charts similar to **fig. 10**, showing cluster profiles across all V_{SD} sets at *k*3, are provided for all V_L sets in **appendix C**. Now, in order to reveal more clearly the V_{SD} differentiation between V_L clusters observed in this step, we now proceed to step 3 which directly compares V_{SD} profiles to cluster proportions.

6.2.3. Step 3 – Sociodemographic Index

292. In this step, the V_{SD} profiles of each place are collapsed in order to obtain a sociodemographic index for each place.[306] The index has no real meaning in itself, except that higher wealth places (wealth here being a function of age, income, and education) will have a higher score than lower wealth places. As this is an index, the most wealthy place in a given dataset would have a score of 1, and the least wealthy a score of 0. The index is calculated and normalised against the final dataset only. In addition, the proportion of cluster assignments of each place is calculated. This allows us to plot the relationship between the V_{SD} index and V_L cluster proportions. As introduced in **§ 6.1**, these plots are presented in constellation charts. Each chart that follows comprises one plot (or panel) per cluster, and each panel shows all places in the dataset. The vertical axis shows the V_{SD} index, and so a given place will appear at the same height in each graph. The horizontal axis shows the relative proportion of the dataset population assigned to a given cluster. To help reveal patterns in the data, a local smoothing line is projected onto each panel ; the shaded range of error is a decent proxy for the density of data points—it expands as the data grows sparse. The smoothing line is a repurposing of a local regression technique for means of visualisation (see also **n. 310**). As noted in **§ 6.1**, the analysis that follows will show regular V_{SD} differentials between V_L clusters. These clusters are labelled ‘proper’, ‘standard’, and ‘non-standard’, referring to assumed social perceptions of language varieties in the clusters concerned. For a given place, an increasing proportion of 'proper' user–documents correlates positively with the V_{SD} index, an increasing proportion of 'standard' user–documents shows a flattened convex correlation, and an increasing proportion of 'non-standard' user–documents correlates negatively with the V_{SD} index. While the social perceptions of language varieties is assumed, the cluster labels are not ad hoc, but rather were chosen on consideration of the ethnicity component of the cluster V_{SD} profiles, the geographic distribution of the clusters, and the sociodemographic context and history of the site of research.

293. **Figure 12** shows the first of the constellation charts. It is an analogue to **fig. 10** in presenting the clustering on V_L set `w2`, except no threshold is applied—all places and user–documents in the final dataset are represented.[307] Unlike **fig. 10**, the cluster labels have not been permuted in any manner specific to this step. Note that each panel header indicates the cluster and the mean (and thus roughly overall) proportion of users assigned to that cluster (marked by a vertical line in the plot). Also note that while each place is a dot, they differ. To facilitate interpretation, the size of the dot is scaled according to the number of users in each place assigned to the cluster in question.[308] Places that have no representation of a given cluster are indicated by a hollow semicircle on the vertical axis (thus 0%) at the appropriate height per V_{SD} score. The lateral extent of the smoothing line indicates the range of prevalence in each cluster panel. In addition, places are shaded according to the proportion of anglophone Michiganders (i.e. the 'mainstream' of people born in Michigan, as opposed to anywhere else, and speaking only English at home)—a brighter shade indicates relatively more of them.[309]

294. While **fig. 12** is information dense, the broad outlines are clear. Attend to the smoothing lines—they mark a local regression through the data giving a suggestion of its centrality as read along the horizontal axis. Roughly speaking, the smoothing line gives a suggestion of the 'centre of gravity' in V_{SD} index terms at a given cluster proportion.[310★] Note that, reading from left to right in each panel, thus from lower proportion to higher, cluster 1 shows a rather flat arc, cluster 2 trends upwards, and cluster 3 generally trends downwards. What we are interested in here is that skew in the plot (not the skew discussed in previous sections)—that is, as the proportion of a given V_L assignment increases, what happens to the general level of the V_{SD} index? Cluster 1 shows a tendency

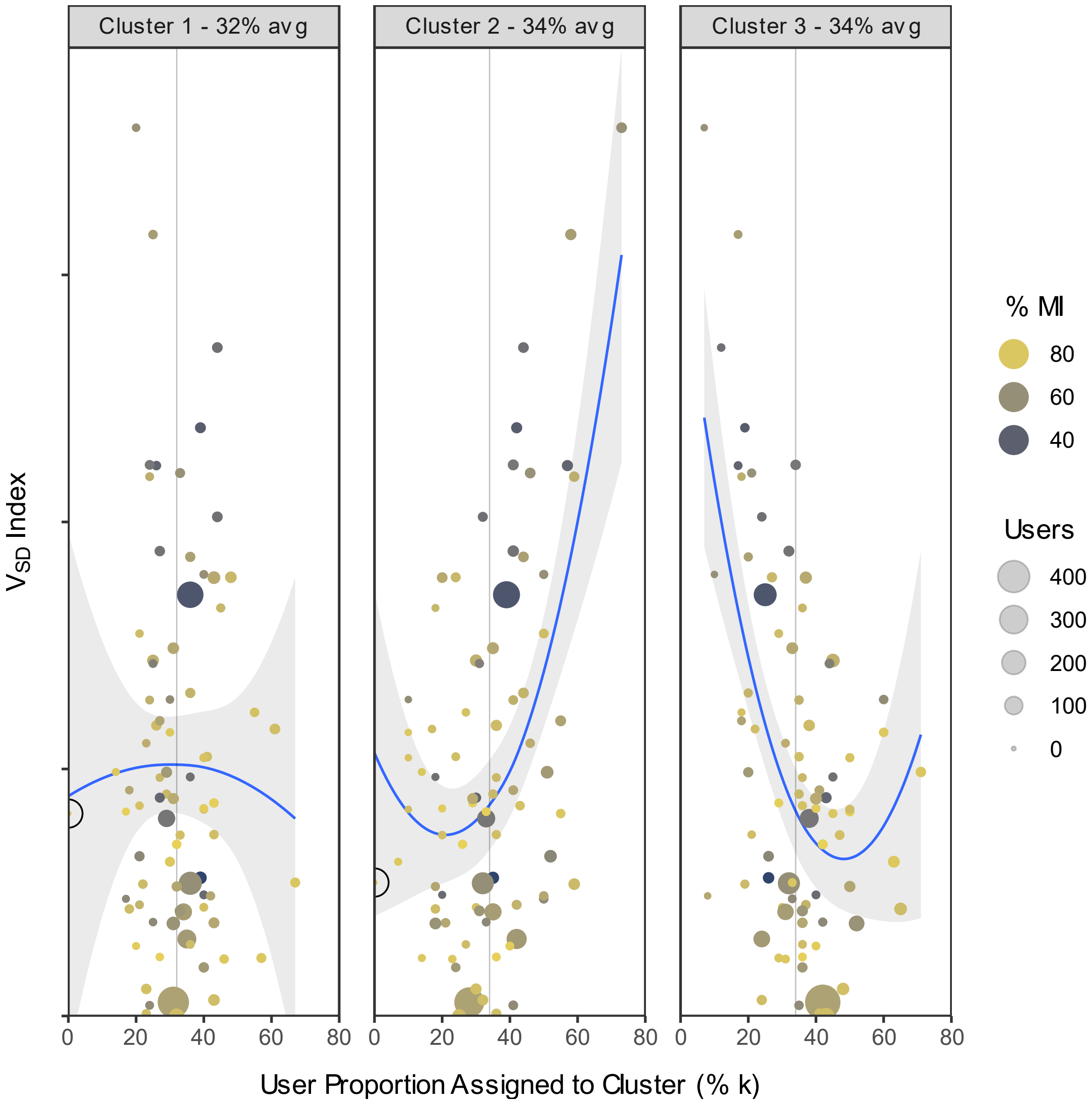


Figure 12. Clustering *k*3 on V_L set `w2`

towards the middle ground, as suggested by the distinct pinch in the range of error ; the *V_{SD}* index is fairly stable across proportions. Extrapolating for the purposes of discussion, let us call this cluster the 'standard' cluster. Cluster 2 shows a marked rise in *V_{SD}* score as its relative proportion increases. That is, places with an increased incidence of user–documents classified as cluster 2 tend towards increased *V_{SD}* indicators. This cluster we will consider as the 'proper' cluster. A place with a relatively high proportion of 'proper' user–documents is generally better off all around in terms of our sociodemographic measures. Cluster 3 is the opposite situation. Here we see increased proportion correlate negatively with *V_{SD}* scores. This cluster we will call the 'non-standard' cluster. (Note that cluster 3 is the only cluster that is represented in <u>all</u> places in the dataset.) The impetus behind these suggested cluster names is that perhaps the broad 'genres' of speech taken in

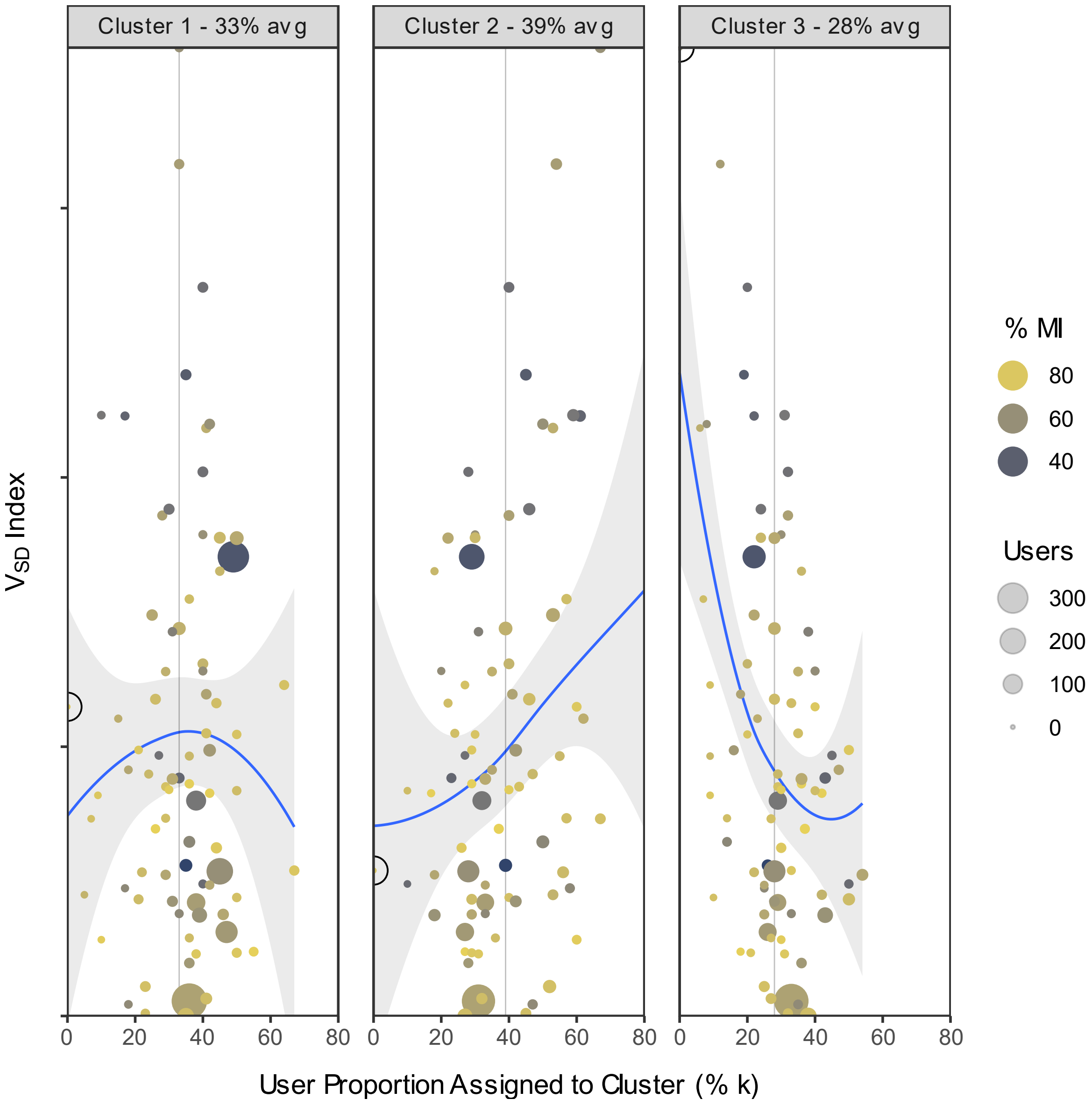


Figure 13. Clustering *k*3 on V_L set `m`

by these clusters can be thought of as representing ‘speaking’ (cluster 1), ‘speaking well’ (cluster 2), and ‘speaking poorly’ (cluster 3). Bear in mind that such framing is in common terms of sociodemographic socialisation. The better way to think about it is that these groups simply have something in common in their speech. All we know for certain is that these clusters represent user–documents that have been scored according to pervasive linguistic features (in this case according to the `w2` set of non-descript words) and found to be more similar compared to others in terms of that scoring. That we can observe a broad relationship between the prevalence of these linguistic clusters and the general wealth of a place proves nothing, but certainly suggests something. Let us look at another V_L set.

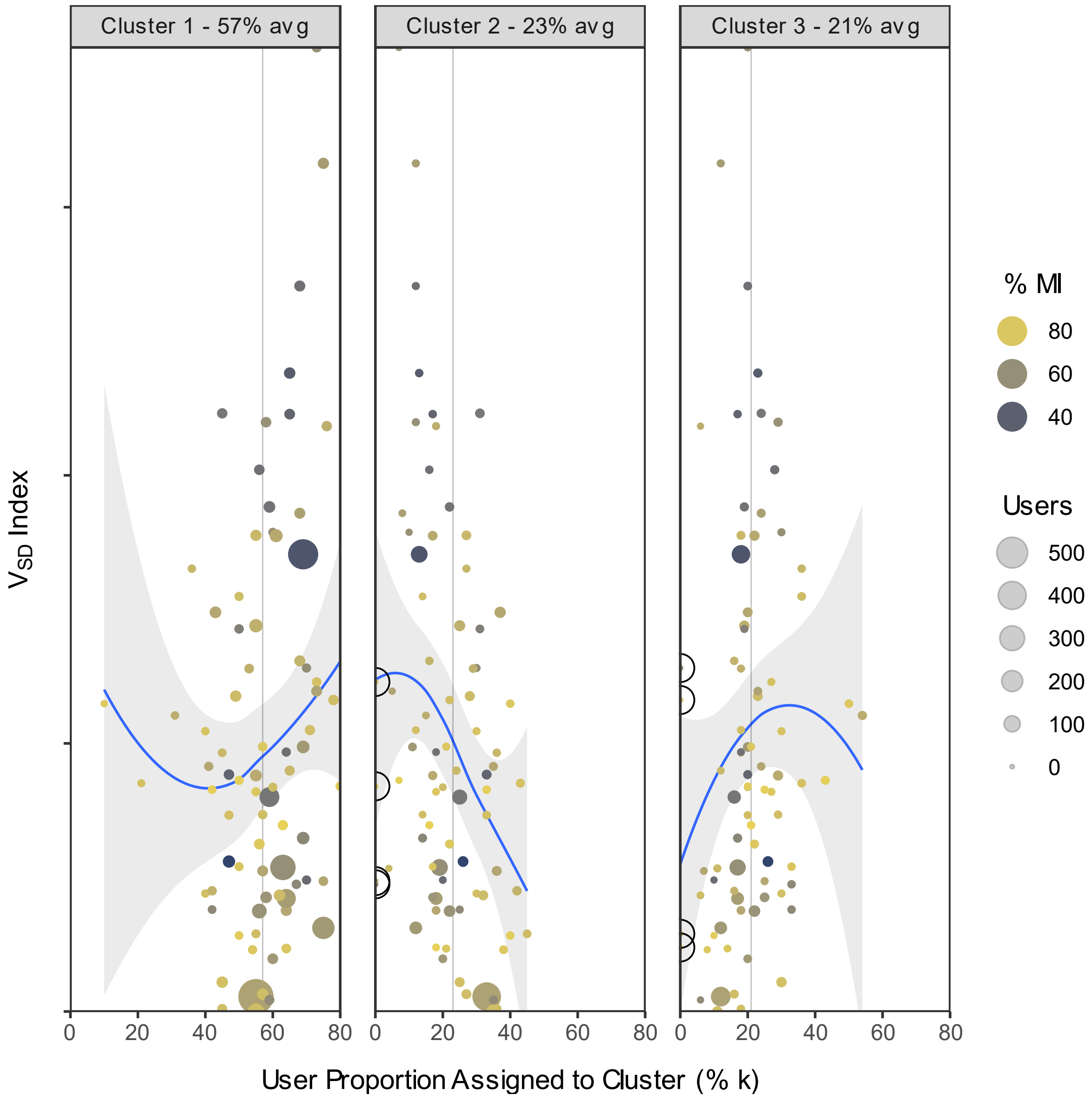


Figure 14. Clustering *k*3 on V_L set `w1`

295. **Figure 13** shows the clustering on *VL* set `m`, that being the factor scores produced by Andrea Nini's (2019) MAT application (**chp 5 § 5.3.3.1**). First of all, be aware that while these clusters represent all user–documents, the constituent users in each clusters have shifted as we are clustering on a different variable set. There is nevertheless substantial overlap—in terms of the intersection between `w2` and `m` clusterings, cluster 1 is 80% the same, cluster 2 69%, and cluster 3 61%. Even as we recalculate cluster membership (i.e. assignment), we see the same broad pattern in *VL* set `m` as we do with *VL* set `w2`. Cluster 1 shows again a relatively symmetrical arc and a distribution with little skew. Cluster 2 shows a similar positive correlation, and Cluster 3 the opposite. Note that cluster 2 shows the highest overall share among the clusters (39% average over places), if just by a bit.[311] These panels from left to right could also reasonably be framed in terms of 'standard',

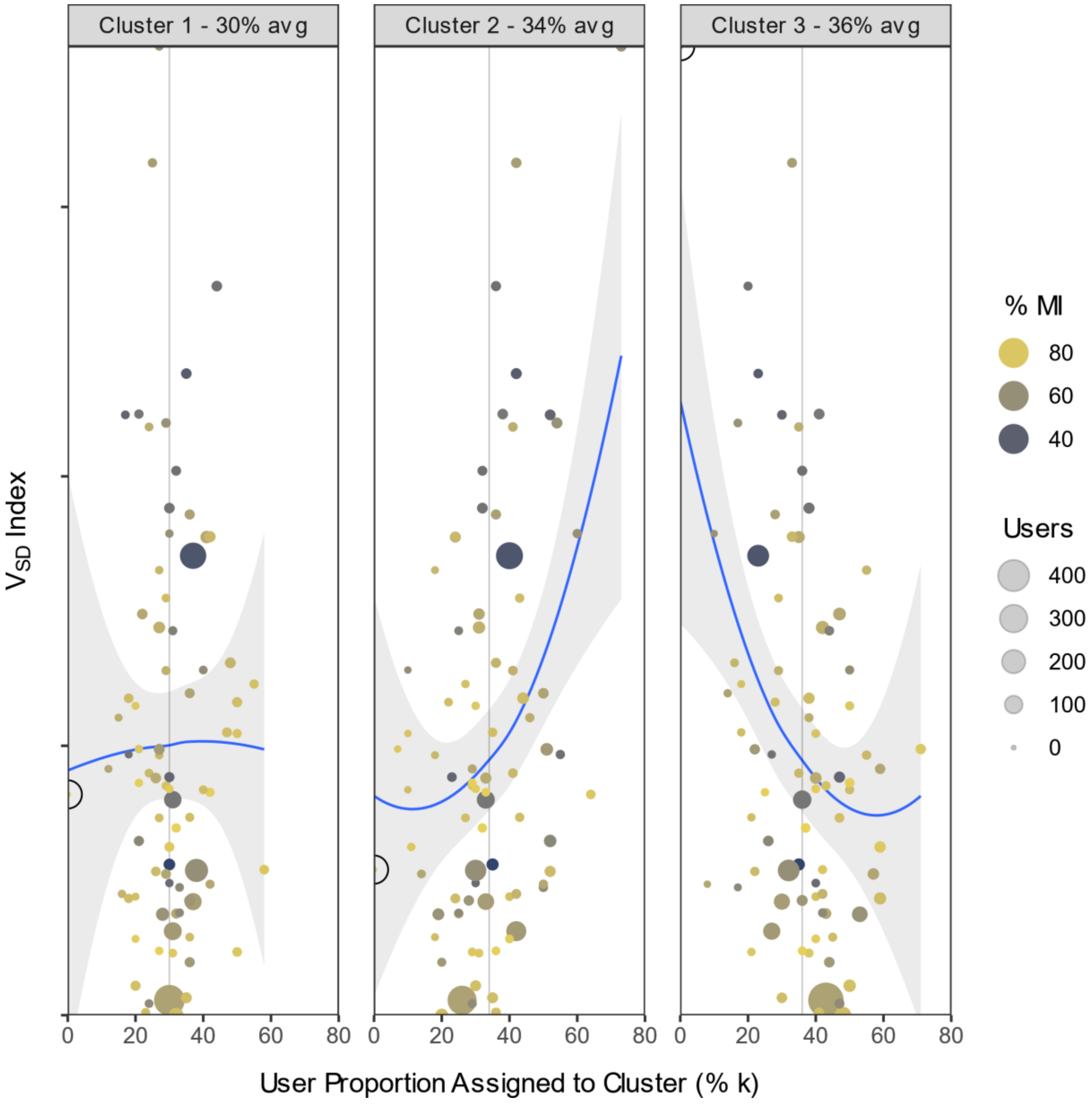


Figure 15. Clustering *k*3 on V_L set `w12`

‘proper’, and ‘non-standard’, with higher proportions of ‘proper’ grammar associating with higher V_{SD} scores, and higher proportions of ‘non-standard’ grammar showing a noticeable drop (also notice the lower prevalence overall of that cluster). Again, we do not actually know what sort of tendencies in grammar these clusters exhibit—that would require scrutiny of the MAT scorings in close consultation of the user–documents—only that the users in these groups are similar to each other in those terms. The labels of proper, standard, and non-standard are simply a sociodemographic framing, given the observed association between V_L clusters and V_{SD} scores.

296. **Figure 14** shows the cluster on V_L set `w1` (the set of pervasive words that are distinctively pervasive for certain user–documents compared to the corpus). Recall that while set `w2` is pervasive non-key words, set `w1` is pervasive key words, in terms of keyness as described in **chp. 5 § 5.3.3.2**. It can be seen in these panels that keyness exhibits a kind of gravity—whereas the last two sets showed relatively balanced clustering, the clustering on set `w1` is loaded heavily on cluster 1. Taking the `w2` clustering as a reference, we can get a sense of how this clustering is differently sorted. In terms of user–document composition, `w1` cluster 1 (that shown in this chart) intersects `w2` cluster 1 at 76%, `w2` cluster 2 at 68%, and `w2` cluster 3 at 36%. The other two `w1` clusters make up the remainder. In the case of the `w2` clusters 1 and 2, these remainders are likely marginal cases near cluster boundaries. That is not the case with `w2` cluster 3 (the 'non-standard' panel in **fig. 13**)—`w1` cluster 2 (that shown here) intersects with `w2` cluster 3 at 50%. That is unlikely to be marginal cases. While the smoothing lines in these panels show broad similarity with those in **fig. 6** and **fig. 7**, it would be a mistake to label these panels here from left to right as 'proper', 'non-standard', and 'standard'.[312]* Given the predominance of cluster 1 and the apparently 'non-marginal' membership in cluster 2, we have a situation of 'standard' and 'non-standard', respectively. Note that standard cluster 1 is found in some degree in all places (with the lowest prevalence of about 10% and the highest off the chart at about 85%) while non-standard cluster 2 is unattested in at least 4 places (shown by the hollow semi-circles on the vertical axis). Again a positive correlation can be observed between the prevalence of the standard cluster and V_{SD} scores, and a negative correlation the prevalence of non-standard, which is substantial in this case. Cluster 3 likely has a substantial component of marginal cases—again those being user–documents the V_L scores of which put them near cluster boundaries. Using V_L set `w2` again as a reference, `w1` cluster 3 (that shown here) intersects with `w2` cluster 1 at 15%, cluster 2 at 21%, and cluster 3 at 14%. These degrees of intersection do not point to this as a 'Destination Cluster', so to speak. Membership of the cluster 3 shown here is more likely associated with the character of V_L set `w1`. In representing a (small) selection of top key words found in the user–documents, `w1` scores can be swayed strongly by local events and topics (cf. **appendix B**). Moreover, bear in mind that the smallest dots indicate only a handful of user–documents—in cluster 3, all places above 30% prevalence are attested by an average of only 5 user–documents, meaning that their prevalence and attestation in the other clusters are even smaller. Such small-*n* places will have unstable scores in any case, showing exaggerated response to small changes in the processing pipeline (e.g. changes in thresholds or features). The situation could be ameliorated by raising the threshold of

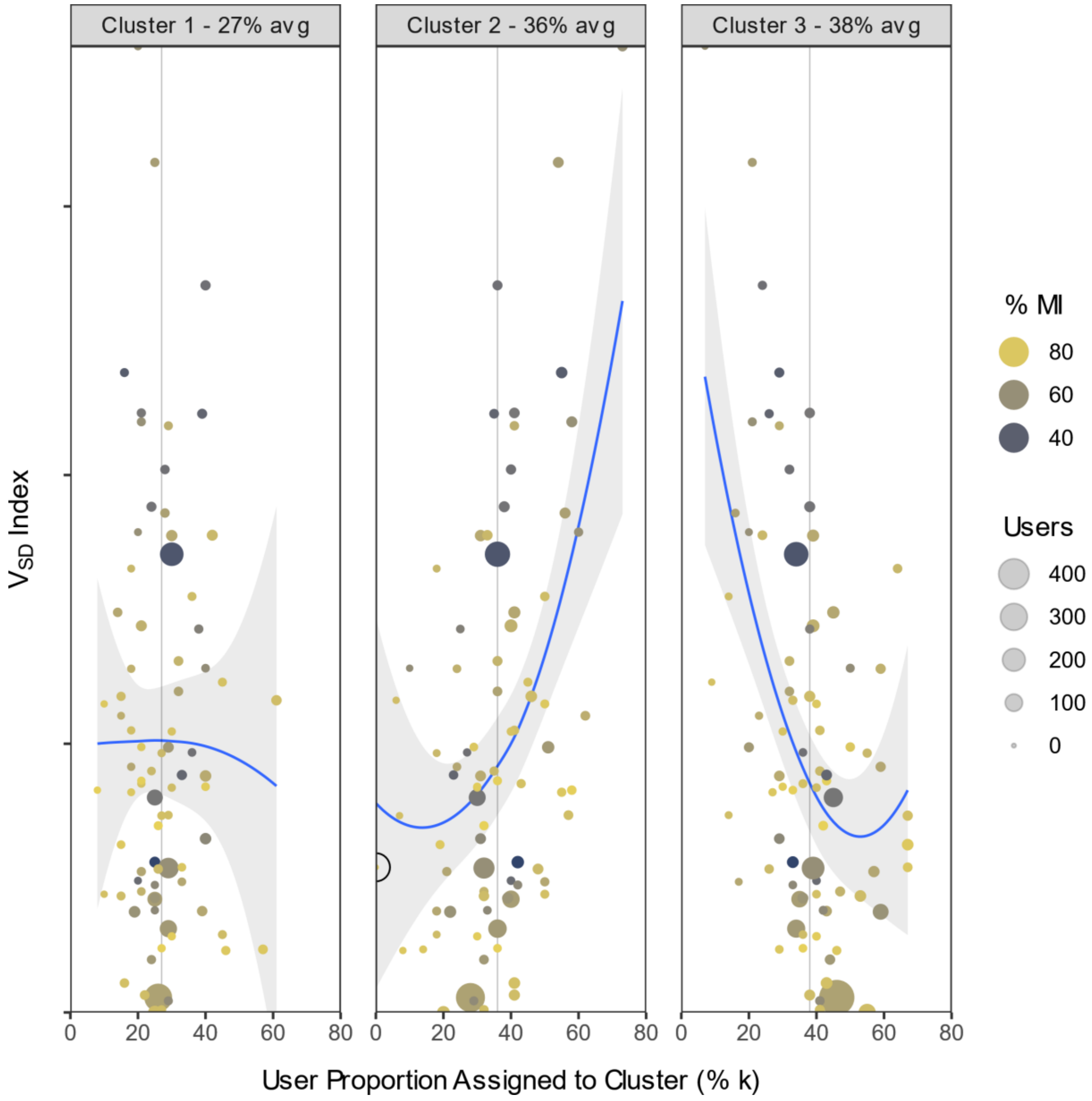


Figure 16. Clustering *k*3 on V$_L$ set b12

inclusion in the final dataset, but at the cost of a substantially reduced final dataset. Recall that the inclusion threshold was set at a minimum of 10 user–documents per place. That threshold was a considered choice to provide a good balance between score stability and coverage. While *V$_L$* set `w1` is a strange beast because of how it was assembled, it is nevertheless derived from the data.[313] And while it comprises key words that thus render scorings subject to local effects, these are nevertheless pervasive key words. It is perhaps a more natural case to consider *V$_L$* set `w12`, which is a clustering on `w1` and `w2` factor scores jointly (i.e. on both key and non-key pervasive words). This is presented in **fig. 15**. The pattern observed here is broadly comparable to that in **fig. 13** of the `w2` clustering. It is however softened somewhat—all three smoothing lines have been flattened, yet nevertheless show a similar flatness in 'standard' cluster 1, a positive correlation in

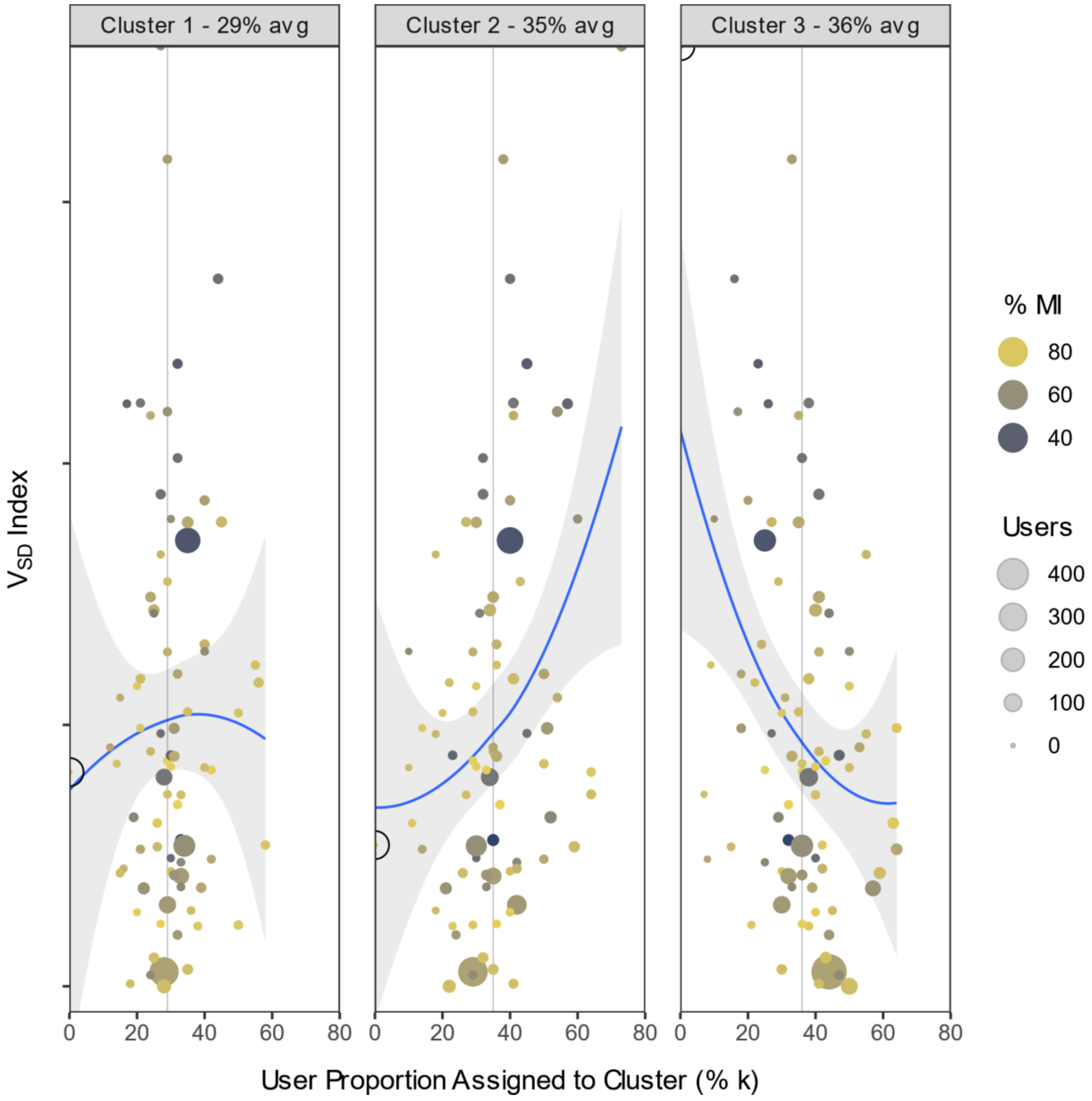


Figure 17. Clustering *k*3 on V_L set `mw12b12`

'proper' cluster 2, and a negative correlation in 'non-standard' cluster 3. There is a slight shift in cluster prevalence, but the cluster memberships remain largely the same—with `w1` as reference, `w12` cluster 1 intersects at 84%, cluster 2 at 87%, and cluster 3 at 87%.

297. Clusterings that can be understood as non-standard, standard, and proper (again, from a sociodemographic framing) are observed in the other V_L sets as well. Constellation charts for all V_L sets at *k*3 are presented in **appendix D**. With the key and non-key bigram sets (`b1` and `b2`, respectively) the pattern is harder to discern ; however, set `b12` (the joint clustering on key and non-key bigrams) produces a constellation chart remarkably similar to that of set `w12`, as shown in **fig. 16**. The joint clustering on all feature sets, that is on grammar (set `m`), key and non-key words (set `w12`), and on key and non-key bigrams (set

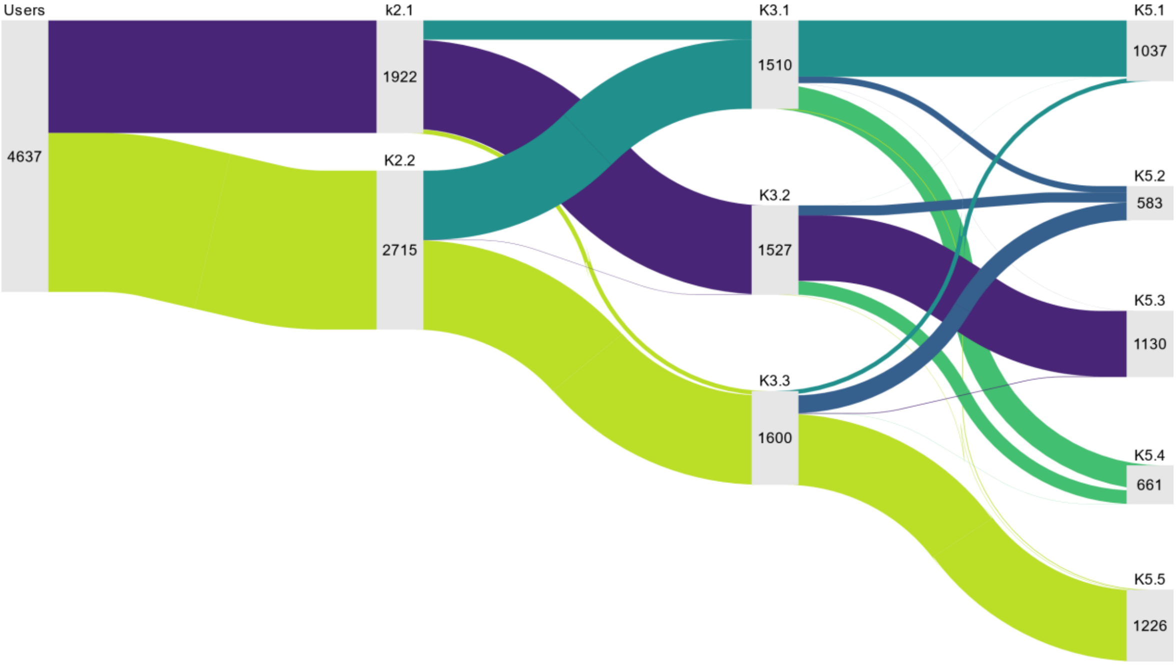


Figure 18. Flow of User Cluster Assignments on V_L set `w2`

`b12`), is shown in **fig. 17** [NB: the vertical axis begins slightly below zero in this figure for technical reasons only]. Once again, we can see fairly balanced cluster prevalence, a largely neutral 'standard' cluster (panel 1), a positively correlated 'proper' cluster (panel 2), and a negatively correlated 'non-standard' cluster (panel 3). Cluster membership remains largely unchanged compared to `w2`, with cluster 1 intersecting at 80%, cluster 2 at 87%, and cluster 3 at 86%.

298. The overall pattern that has been observed at *k*3 holds as well at *k*5. The *k*5 constellation charts are not presented for reasons of space, but we can demonstrate why the pattern can be expected to hold. It is largely due to the fact that clustering at slightly higher dimensions allows marginal cases to be sorted better, while leaving the main clusters observed at *k*3 intact. This sorting can be seen clearly in **fig. 18**, which gives a Sankey diagram of the flow of user–document cluster assignments on V_L set `w2` from the corpus, to *k2*, *k3*, and finally to *k5*. Each node shows the user–document count. The flow colours are assigned at *k*5, thus there are five colours, and all inflows to a node share that colour. At the preceding nodes, all inflows are coloured according to the colour of the largest outflow colour. This allows us to trace the common assignment flows. For example, consider the darkest shade of the chart, which belongs to cluster 3 at *k*5 (node K5.3). The bulk of its inflow was the largest outflow of cluster 2 at *k3* (node K3.2), so the inflows to that node are assigned the same colour. The same situation obtains with cluster 1 at *k*2 (node K2.1), and so those inflows are assigned the same colour. It is plain to see that, at least for V_L set `w2` (which has been providing our reference set, as it provides the most balanced cluster prevalence), the *k*3 clusters persist into *k5*, with *k5* cluster 2 emerging mostly from *k*3 cluster 3, and *k*5 cluster 4 emerging mostly from *k*3 cluster 1. That is, the largest clusters present at *k*5 are the standard (c.1), proper (c.3), and non-standard (c.5) clusters observed at *k*3. Recall that the smaller 'interstitial' clusters were also observed in the discussion of **fig. 11** showing the sociodemographic profiles of the *k*5 clustering on V_L set `w2`.

299. In sum, by plotting the V_{SD} index of places against their relative proportion of V_L cluster assignments—presented here by means of constellation charts—step 3 reveals regular V_{SD} differentials across V_L clusters. That is, there are regular sociodemographic differentials across linguistic groups. As noted at the beginning of this chapter, the situation is not so clear on the ground. As has been shown in the constellation charts here and in **appendix D**, and noted elsewhere in this chapter, places in contemporary societies are highly mixed in social terms. The figures on the following page show the reality on the ground. Both figures zoom in on the region around Detroit, Michigan. It is the largest concentration of population (and of everything, really) in Michigan, and provides a helpful tableau of the sociodemographic and linguistic clusters that we have been examining. **Figure 19** maps the V_{SD} index, and as well its disaggregated components. Each coloured shape is a municipality represented in the dataset (note that cities are irregularly shaped, and townships are mostly square ; cf. **n. 196**). **Figure 20** maps the same region and municipalities. Each map in that figure represents a single cluster according to its prevalence, in a fashion similar to the constellation charts. Note that this is not a neighbourhood map—these maps each cover some 250–300 square miles and millions of people. These maps are not simply eye-candy. Two key points are important to observe: 1) it is possible to suss out patterns in comparing the sociodemographic mappings to the linguistic mappings ; but 2) it is extremely difficult, and questionable in any case. The method is not suited to be visualised in such a manner, and standard cartographic maps themselves are not suited to this subject matter. Social phenomena at whatever scale are simply too mixed up and messy—that is life. Clear-cut maps are suited to exogenous data ; social research in hybrid society needs other ways to see.

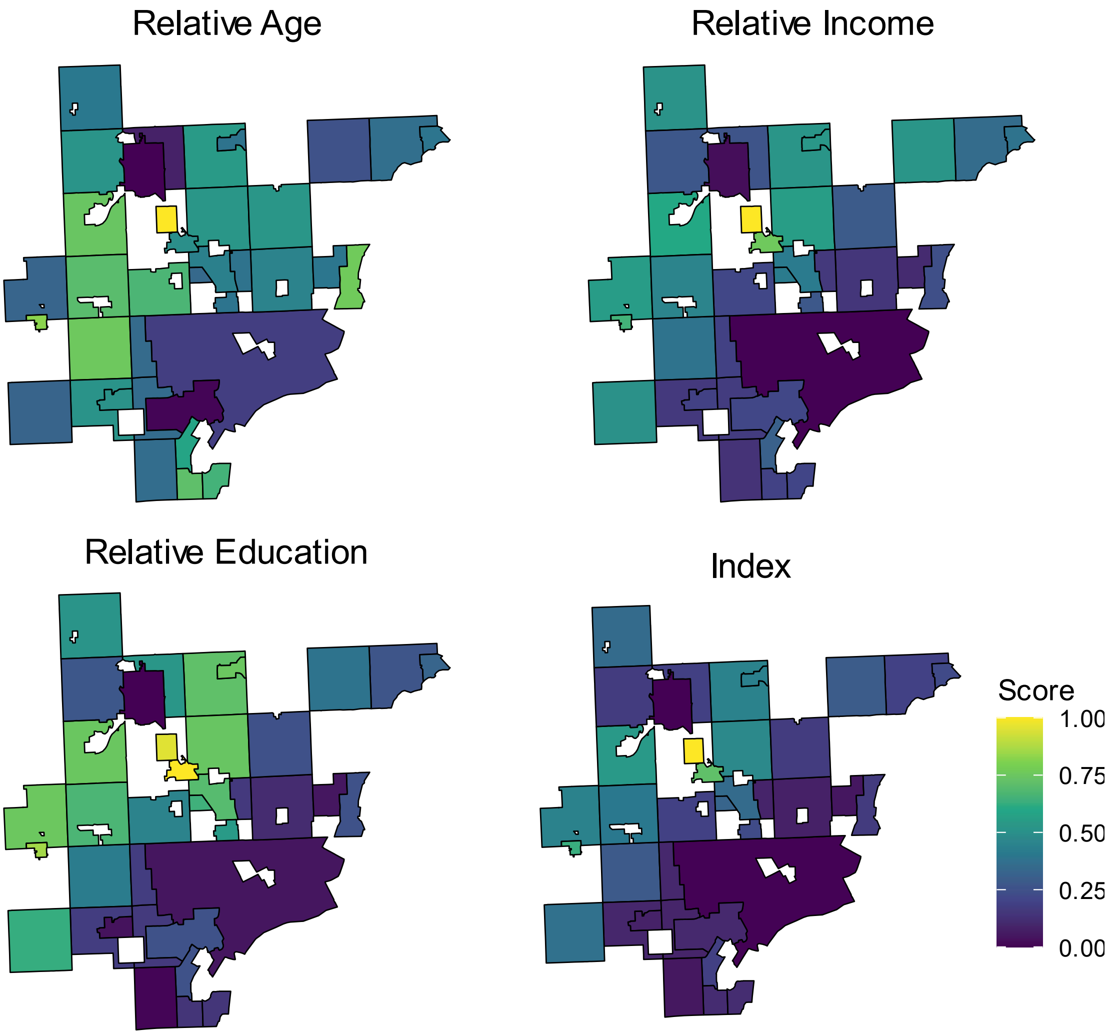


Figure 19. Mapping of the V_{SD} Index and Components (Greater Detroit Region)

6.3. Evaluation

300. Overall, the constellation charts presented here and in **appendix D** reveal more clearly the dynamic suggested in the profile plots of step 2. The dynamic is that, at a given *k*, certain V_L set clusters will show a positive correlation between their prevalence and the V_{SD} index scores of the places where they are attested, certain clusters will show a negative correlation in that regard, and certain clusters will show a shallow convex correlation. These clusters have been described here using terms reflecting a sociodemographic framing—the positively correlated clusters have been termed ‘proper’ (i.e. ‘proper’ speech, from the social perspective), the neutral clusters as ‘standard’, and the negatively correlated clusters as ‘non-standard’. These are not simply labels of convenience, but rather point towards what we are taught through socialisation, and

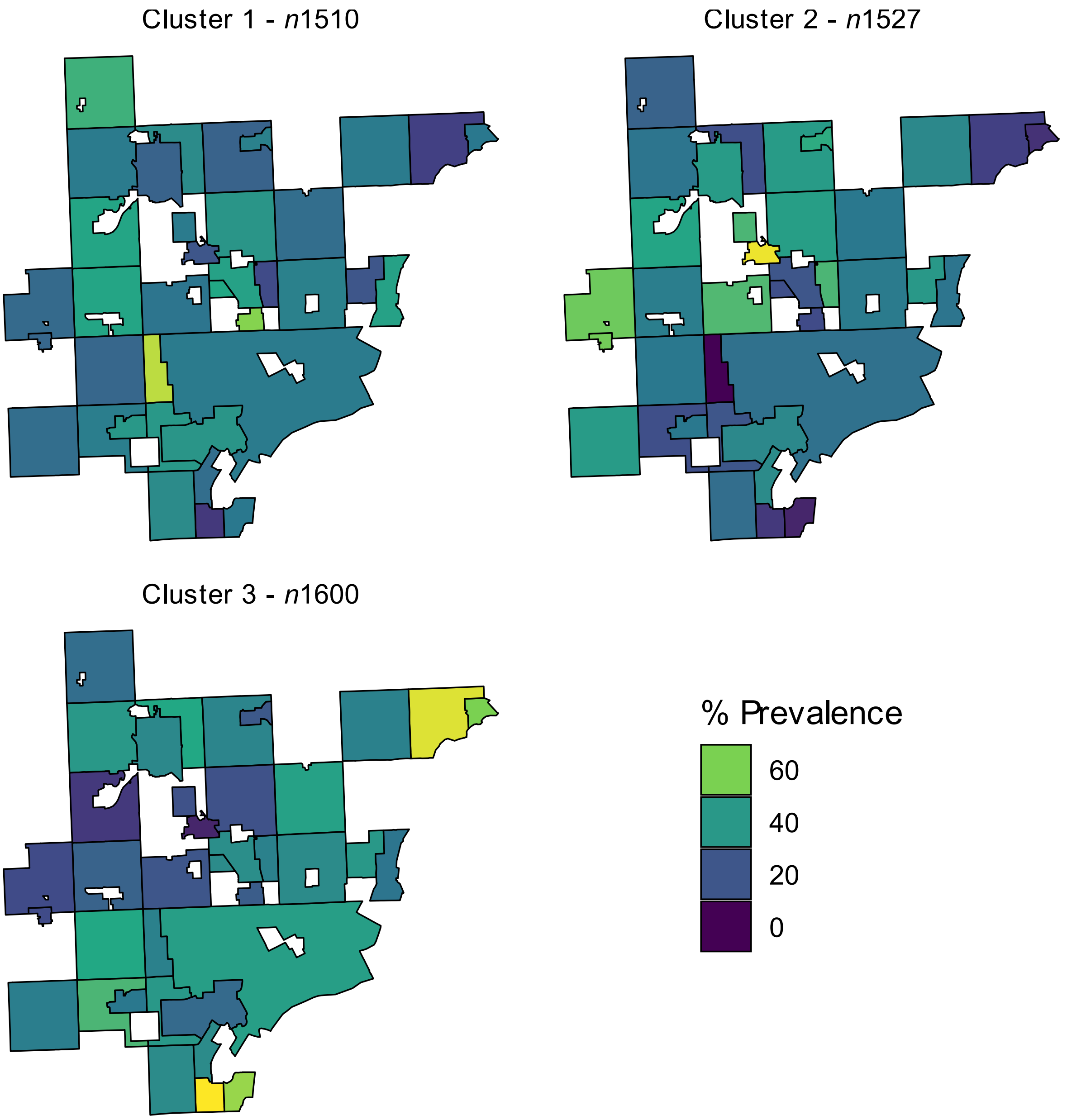


Figure 20. Mapping of *k*3 Cluster Prevalence (V_{SD} set `w2`)

moreover towards what has been well-documented at lower levels of analysis—linguistic features correlate with sociodemographic contexts. The preceding discussion has demonstrated this effect at scale, and in a manner that is clear. It was found that user–documents clustered solely on V_L sets will demonstrate clear patterns when evaluated according to V_{SD} sets. That is, groupings of language similarities exhibit structure in sociodemographic terms, and thus **RQ$_2$ is answered in the affirmative**, and by extension **the provisional answer to RQ$_1$ is warranted**. We now proceed to the conclusion of the work.

Chapter 7

Conclusion

> Tools and techniques do not exist in a conceptual void. Methods are linked to methodologies, which themselves are understandings of [our] stances concerning the reality … of what those methods allow us to study and the knowability that we presume about that world. … Deconstructing the qual–quant taxonomy and raising the visibility of constructivist–interpretive methods within political science research practices takes us further toward the conceptual complexity that marks the human sciences.
>
> Dvora Yanow, "Empirical Political Science", 2003, p. 12

301. This work has presented an interdisciplinary methodology for macro-level research of large-scale social phenomena grounded in communication studies. That methodology was developed through a reconceptualisation of the topic of political participation in hybrid society. The operationalisation based on that reconceptualisation was implemented by means of a corpus-based method developed for empirical study of mediated public discourse. Analysis of the results of the method indicates that the operationalisation is warranted. By that token, **the methodology is warranted**.

302. This methodology is not a 'plug and play' tool for other researchers, and was not intended to be. Rather, by developing the methodology in a manner of explicit social science, whereby all conceptual and operational elements are specified, the methodology thus provides a step-wise template that might guide other researchers intent on pursuing interdisciplinary work. In that light, the methodology itself is the fundamental contribution to knowledge of this work.

303. This concluding chapter is organised as follows. Section 1 reviews each step taken in this work, recapping each chapter and their specific contributions to the methodology overall. Section 2 addresses specific contributions to knowledge within the methodology. Section 3 notes certain limitations to the work. Section 4 concludes the document with suggestions for further work.

7.1. Review of the Work

304. Chapter 1, "Introduction", laid out the overarching goal of this work: to develop an interdisciplinary methodology for the macro-level study of large-scale social phenomena. The motivation for such a work is the challenge to social inquiry, witnessed across fields and disciplines in recent decades, of the emergence of hybrid society. Beyond being networked by information and communication technologies, hybrid society is characterised by deep mediatisation and the ever-present blending of the offline and online, the physical and virtual. A fundamental move necessitated by the character of hybrid society is that, in order to study it with fidelity—that is, to see society as it is, rather than as it is expected or desired to be—we must move away from structural conceptualisations of phenomena towards socially communicative conceptualisations.

Conceptualisations of the latter sort are common place in micro- and meso-level work. However, the former sort obtain in macro-level work, and thus generally in large-scale work as well. This is because we are socialised into a structural view of society—such a view serves the institutions of society, and thus it is predominant both within and without the academy. Nevertheless the conceptual move must be made if we wish to see ourselves clearly.

305. The conceptual move from the structural to the socially communicative implicates a shift in subject model. That term refers to how researchers conceptualise their fundamental unit of study. As demonstrated in this work, structural conceptualisations are based on a subject model of the 'fungible individual'. That subject model results from the dependence on exogenous measures and categoric variables for characterising society at scale. We each and all are slotted into various brackets of age, income, educational attainment, and so forth. In that manner, our collective humanity is ignored in favour of "nothing but a numerical aggregate, a conglomeration of units" (Dewey, 1888, p. 4). Thus this work argues for the development of endogenous, contextual measures—based in a subject model of the 'social person'—so as to recover the humanity inherent in the phenomena we study. If social science is not foremost and fundamentally social, then it can never be science for it has mistaken its subject.

306. The fungible individual could be glossed as an epistemological stance towards society, employed so as to make use of available data. However, as this work demonstrates, the structural perspective is so deeply ingrained in some fields and disciplines that it functions as the ontological footing upon which their work is built. In that the move from fungible individual to social person as subject model implicates an ontological shift, the effects of the move have thorough-going impacts on theory. In that the study of hybrid society necessitates re-evaluation of theory and the development of suitable method, it thus obliges a total reconsideration of methodology. It is certainly not the case that we must reinvent the wheel. The knowledge and practice appropriate to hybrid society is there, but scattered across fields and disciplines. As we attend to each component of hybrid methodology, we must weave together that scattered learning. Thus by reason of practicality alone, methodology for hybrid society will necessarily be interdisciplinary. As this work demonstrates, communication studies is well suited to this endeavour given its theoretical and subject orientations as well as its relatively diffuse disciplinary boundaries.

307. As a contribution to the development of 'hybrid methodology', this work demonstrates an example of such through the reconceptualisation and subsequent operationalisation of a topic that is indicative of the challenge to research posed by hybrid society—the study of political participation. The purpose in selecting such a topic is to demonstrate that a hybrid methodology can operate as a bridge across disciplinary understandings and modes of work so that they might work together productively in the face of societal change. Hybrid methodology does not come to destroy disciplines, but to fulfil them. To prompt the development of the methodology, this work posed two research questions:

RQ_1 – How can political participation as reconceptualised in hybrid society be operationalised for computational and statistical analysis? and

RQ_2 – Can the results of such operationalisation remain interpretable from a structural perspective?

To answers these questions empirically, a conceptual framework was elaborated and subsequently implemented by a method.

308. The components on the conceptual framework are as follows:

- the **ontological footing**, which declares the fundamental objects of concern to this work (the <u>social</u> domain of meaning, the <u>material</u> domain of mediation, and the <u>technical</u> domain of affordance) ;
- the **epistemological stance**, derived from a reconceptualisation of political participation focused on communication and context, which suggests how we might understand the elements of that phenomenon and their relations, and how and where we might observe it ;
- the **theoretical approach**, which elaborates the epistemological stance, fully specifying its elements in order to map the conceptual framework onto a specific study and site of research—in this case the US state of Michigan during the 2018 election ; and

- the **operational step**, which suggests the specific phenomena—in this case pervasive linguistic patterns attested in public discourse—that might yield empirical evidence with which to evaluate the research questions motivating the methodology.

The method implemented the operational step, in the manner specified in the theoretical approach, by means of a corpus-based approach public discourse that developed an 'enriched corpus' comprising the discursive, sociodemographic, and geographic profiles of a common population in the site of research. Evaluation of the results, and thus answers to the research questions, was done by way of analysing the categories of profiles in concert. The evaluation of the analysis of the results of the methodology—that is, the framework and method as a coherent whole—found warrant for the operationalisation and by extension the reconceptualisation. By that token, the hybrid methodology presented in this work is warranted.

309. The following sections review the conceptual framework, method, and analysis and evaluation, proceeding step-wise through the chapters of this document.

7.1.1. Ontological Footing

310. Chapter 2, "Theorising the Social–Technological Question", served to develop the ontological footing. It did so through an examination of the relationship between society and its technologies. The history of thinking on this topic was reviewed, and the contemporary trend towards sociomaterial understandings was highlighted. As it is essential to consider information and communication technologies in the study of hybrid society, they were considered from a sociomaterial perspective. It was observed that in most understandings the social and the material exist in ontological tension, which presents a significant complication to understanding hybrid society. This tension was resolved by means of elaborating an affordance perspective on the relationship between the social and the material. This perspective was then further elaborated to develop the ontological footing of the conceptual framework. It understands three fundamental domains: the social, the material, and the technical.

- The **social** is the domain of meaning, where meaning is understood as the representation of consequence, and intention is a subset of meaning, understood as the representation of effecting consequence. This domain is incorporeal and processual (i.e., it has no fixed state ; it is always becoming).

- The **material** is the domain of mediation, where materiality is not a function of physicality, corporeality, or any persistence in duration or extent, but rather is a function of relational influence, be that relation spatial, temporal, or semiotic. Note that the material does not mediate, but is mediation ; the material does not cause, but is consequential—cause is a meaning, and thus belongs to the social. If any thing influences or is influenced by any other thing, that is a consequential relation and thus the things in question are material. This domain is incorporeal, asocial (thus devoid of any inherent meaning) and relational.
- The **technical** is the domain of affordance, the interface between the social and the material, where technicality is a matter of function. Here affordance is a synthesis of meaning and consequence—they are bound together and made real in the technical. The technical exists between the social and the material, which themselves cannot directly interact. This domain is contingent (on the coexistence of the social and the material) and functional, and is the only domain that is real in a physical sense.

The ontological footing, though seemingly arcane, serves a basic heuristic purpose: to allow social processes and material relations to be, without unnecessary assumptions of their nature, thus focusing attention on the functional reality of phenomena, likewise without unnecessary assumptions of their nature. The ontological footing provides the basis with which to interpret the reconceptualisation of political participation in a manner whereby an epistemological stance can be derived that is suitable for the phenomenon as it manifests in hybrid society.

7.1.2. Epistemological Stance

311. Chapter 3, “Reconceptualising Political Participation in Hybrid Society”, serves to develop the epistemological stance. That stance tells us how we might see the elements of the ontological footing to be related in political participation in hybrid society, and thus how we might expect the phenomenon to manifest in the world. Simply put, the epistemological stance tells us what we are looking for. The stance is derived from the topic of inquiry. For this work that topic is political participation. The history of the scientific study of that topic is reviewed to contextualise it, and to demonstrate that the conventional study of the topic is ontologically structural. The challenges of hybrid society to its conventional study were discussed, and these challenges were attributed to a subject model of the 'fungible individual'. Effectively, the structural orientation of the conventional study of political participation is blinded to the fundamentally social character of participation in hybrid society. For that reason, an ontological shift is required—that being the move from the 'fungible individual' to the 'social person'. That the reconceptualisation of the topic requires an ontological shift obliges attention to all components of a methodology, as has been given in this work. To derive the epistemological stance, political participation was reconceptualised in hybrid society by way of a set of objective criteria in order to highlight where emerging understandings depart from conventional understandings. The reconceptualisation was then interpreted in terms of the ontological footing, thus yielding the epistemological stance.

312. The stance is unassuming, being little more than a declaration of the character of the ontological domains and how they are expected to relate. In this case, the material and social domains were unspecified, meaning that they are unbounded. In the conventional study of political participation, there is much effort put in to arguing over what actions, and performed by whom, count as participation. Not so in hybrid society—all actions are potentially consequential, and all actors potentially can act—and thus the domains are unbounded. The technical domain is specified to implicate communicative action. Moreover, as hybrid participation as a phenomenon can be fitted into this epistemological box, we know that the general social–technical–material configuration of hybrid participation is a valid object of study. Finally, in that the reconceptualisation of political participation replaces questions of intent and consequence with a focus on politicised contexts of action, we know that the stance will need to be elaborated in a suitable manner in the theoretical approach. The epistemological stance is a prototypical description of the phenomenon under study. As such, it provides a template. By fully specifying the elements of the template in consideration of a site of research, we make our theoretical approach.

7.1.3. Theoretical Approach and Operational Step

313. Chapter 4, “Operationalising Language in Mediated Public Discourse”, serves to specify the theoretical approach and to suggest the operational step based in language. Here the framework begins to move from abstraction towards application. As noted, the theoretical approach is arrived at by fully specifying the epistemological stance in relation to a specific site of research (the stance already providing the topic of research). Following from the relative simplicity of the epistemological stance, three components needed specific conceptualisation:

- a bounding of the social–technical–material configuration in question (i.e., the prototypical site of research) ;
- a specification of the social, material, and technical domains in question ; and
- a proposition of the functions of interest in the technical domain, and the expected effects in the social and material domains (recall that only the technical is observable).

314. The bounding was straightforward. In light of the deep mediatisation of hybrid society and the orientation of this work towards large-scale phenomena, public microblogging platforms were chosen as the prototypical site in that they are arguably emblematic of both mediatisation and scale. From that bounding the specification was undertaken. It was observed that the process of specification is roughly equivalent to a process of scoping potential sites of research. The specification was undertaken from an ethnographic perspective in order to highlight the difficulty of understanding place (that is, the confluence of time and space) in human terms. A discussion of conceptions and approaches to place led to the notion of ‘lived hybridity’ and an enumeration of the challenges to studying hybrid lifeways—the ‘what’, ‘who’, and ‘where’ of what we study is essentially out of our hands. That sobering fact was considered in terms of how multi-sited ethnography is approached, which was found to have a solution consonant with emerging (though still relatively conventional) approaches to hybrid participation—attend to the context of action. Given that solution, the specification of the site of research was determined to have two aspects: a superficial situation, and a political contextualisation. The former tells us more or less where in the world the site is, and the latter tell us how the site is to be understood as a valid context of study (recalling that the epistemological stance requires a configuration that is a politicised context of action). The rationale for selecting the US state of Michigan during the 2018 election was then given, as was the rationale for selecting Twitter as the specific bounding platform.

315. The political contextualisation was more involved. The essential argument is that the bounded site of research is inherently political because it is discursive. That argument was elaborated through a discussion of rhetorical models of public discourse that implicate negotiations of power relations, and thus social structure, among members of a polity. In that light, it was argued that structural understandings cannot serve the study of socially communicative phenomena, because effectively they can only see the institutional, establishmentarian side of the story. To undergird that argument, the notion of crisis in political participation was addressed. That notion, long established, distils to the case where the establishment views the demands of participants as illegitimate. From the rhetorical perspective on public discourse, the negotiation of who belongs to a polity and who does not is constant, but moreover is determined by the powerful. Thus crisis is a constant. By that token, it is recognised that the study of hybrid society must necessarily be critical if it is to have any fidelity. Institutional approaches (such as the conventional study of political participation) cannot be critical, as they are framed in terms of the status quo, and thus will never see hybrid society with fidelity, only normativity.

316. Having made the theoretical approach, the operational step is taken. The need for exogenous measures was reiterated, and focus was placed on language. The argument for language is that it is the root and impetus of social structure. It serves to negotiate meanings and understandings of consequence, and further to disseminate and interpret those meanings and understandings. From the ontological perspective of this work, language is the fundamental, final affordance that binds meaning and consequence as so creates our worlds. Such a philosophical gloss is valuable to consider, but the brute fact of the matter is that language is the basic tool with which we all understand the world and each other, and through which we negotiate such understandings. We all understand this intimately, and it is perhaps both the first and last thing that each of us rightly know. So let us attend to it, and thereby to each other.

317. To that end, the operational step proposed pervasive linguistic features expected to be markers of context as the functions of interest, that is, the specific functions in the technical domain that we seek to observe. Recall that this is roughly equivalent to 'phenomena of interest'. Such pervasive features have substantial potential in terms of developing exogenous measures for critical social inquiry. The study of linguistic variation across social groups was reviewed, and its limited scope was noted. However, corpus-based approaches—which by their nature are oriented towards large-scale study—have been used productively for the study of linguistic variation. Specifically, the method developed by Douglas Biber to study linguistic variation across textual genres was proposed as a model that could be adapted to the study of social groups as social 'genres' (recalling that the research questions speak to social structure, and thus the method must attend to social groups). It was decided to emulate Biber's method, and to supplement its grammatical focus (such features serving as his functions of interest) with attention to pervasive lexis such as function words (such features serving as our functions of interest). The concern with such words is due to the fact that there is reason to suspect that they might be more reliable markers of context than more socially marked (i.e. salient) words, and thus might work well as endogenous measures.

7.1.4. The Method

318. Chapter 5, "The Method", details the method used to implement the operational step in the manner specified in the theoretical approach. As noted, it is a corpus-based method that achieves its purpose by means of an 'enriched' corpus. The corpus of discursive content, comprised of 'user–documents' representing the compiled discourse of a certain population in the site of research, was enriched by linking to those user–documents further data characterising the places to which the user–documents are associated. The final enriched corpus comprises discursive, sociodemographic, and geographic data linked to a common population of persons. As chapter 5 is itself a step-wise account of the method, it will not be recounted further here.

319. The enriched corpus is the result of the method, and its enriched (i.e. complex) data allowed for the analysis comparing aggregate sociodemographic profiles of groups sharing similar lexicogrammatical patterns. The method presented in chapter 5, as the implementation of the operationalisation (i.e. the theoretical approach and its operational step) proposed in chapter 4, provided a provisional answer to **RQ**$_1$—How can political participation as reconceptualised in hybrid society be operationalised for computational and statistical analysis?

7.1.5. Analysis and Evaluation

320. Chapter 6, "Analysis and Evaluation", provides the analysis of the results of the method, and evaluates the analysis according to the research questions that prompted the methodology. As noted, a provisional answer to **RQ**$_1$ was given by the proposed operationalisation and the method that implemented it. To determine if that answer is warranted, the analysis focused on **RQ**$_2$—Can the results of such operationalisation [satisfying **RQ**$_1$] remain interpretable from a structural perspective?

321. The analysis was performed in three steps. All steps sought indications of regular sociodemographic differentials across linguistic groups in the population represented in the enriched corpus. Step 1 examined the relationships between sociodemographic and linguistic cluster assignments given to each user–document. It did so by means of evaluating the overall pair-wise agreement between these assignments. This step was found to be a failure, and for an important reason—social groupings can exhibit substantial skew, that is, group membership of any sort is unlikely to be evenly proportioned across society. Common measures of agreement do not handle skewed categories well, and those that are designed to handle skew are based in statistical assumptions of randomness that are of questionable applicability to social phenomena. Step 1 nevertheless provided valuable insight about the character of the groupings in the dataset (i.e. that cluster membership was not evenly distributed) that was then accounted for in the subsequent steps.

322. Step 2 examined the sociodemographic profiles of linguistic clusters directly, plotting the disaggregated profile brackets against the population percentage recorded for each bracket ; such a plotting is similar to that commonly seen in a population pyramid. Box plots were used to present the aggregated sociodemographic data in each bracket of the places represented in each cluster. In this step it was observed that initially no differentials could be observed between clusters ; this was attributed to the fact that generally all linguistic clusters are found across all places. Recalling from step 1 that cluster membership is skewed, the application of a minimum threshold of cluster representation to filter the data under analysis subsequently revealed regular sociodemographic differentials across linguistic clusters.

323. Step 3 approached the issue of cluster proportions directly. A sociodemographic index was calculated for each place in the dataset (and thus for each user–document, in that each are associated to a place). The index integrated data on age, income, and education ; places with higher indices were considered to be generally more 'wealthy' and otherwise privileged, and places with lower indices were considered to be less 'wealthy' and otherwise less privileged. The collected indices of all places in the dataset were plotted against the relative proportion of the population assigned to each cluster. These plots were presented in 'constellation charts' where each cluster is plotted in a panel and the panels set side-by-side for comparison. Across all linguistic variable sets, at *k*2 and *k*5, a regular sociodemographic differential was observed. One linguistic cluster would exhibit a positive correlation between the sociodemographic index and its relative proportion in a given place. Another cluster would exhibit a slightly convex but generally flattened correlation. And yet another cluster would exhibit a negative correlation. That is, more speakers of the first group associated with more wealth, for the second group sociodemographic outcomes were generally neutral, and for the third group increasing representation associated with decreased wealth. These groups were labelled 'proper', 'standard', and 'non-standard', respectively, on the assumption that the first group represented a favoured (and structurally rewarded) variety, the second group represented an accepted variety though not particularly favoured or disfavoured, and the third group represented a socially disfavoured variety (from which structural rewards were withheld).

324. The regular pattern of sociodemographic differentials across groups clustered solely on linguistic variables was evaluated to be a clear indication that social structure can be observed in linguistic features, and thus that the results of the method are interpretable from a structural perspective. This evaluation answered **RQ$_2$** in the affirmative, thus warranting the answer to **RQ**$_1$. By extension, the methodology overall is found to be warranted.

7.2. Contributions to Knowledge

325. This work makes four main contributions to knowledge in social inquiry. It demonstrates:

1) a bridging methodology for interdisciplinary research in hybrid society ;
2) an adaptation of corpus-based linguistic methods suited to the study of public discourse, and social topics generally ;
3) the potential role of communication theory, and of communications studies, in adapting the social sciences as a whole to the study of hybrid society ; and
4) the limitations in hybrid society of empiricism that is not guided and checked by theory with an appropriate subject model.

326. As noted at the beginning of the chapter, the methodology that has been presented here is the fundamental contribution of this work, by virtue of its interdisciplinary and pedagogical functions. The development of the methodology at every step has been intentionally and explicitly interdisciplinary, on the one hand highlighting valuable knowledge across the social sciences, while on the other hand observing that disciplinary boundaries can be stultifying in the face of societal change. For that reason, this work set out to develop a bridging methodology. By way of integrating thinking on phenomena of common interest, and demonstrating that the results of the methodology could be understood from multiple disciplinary perspectives, the methodology is an example of a conceptual and practical framework for potentially coordinating work across disciplines. The complexities of studying hybrid society oblige that we work together, and the methodology presents one possible roadmap. There is also a fundamentally important pedagogical role served by this methodology. In proceeding step-wise and explicitly through issues of theory and method that persistently dog students of society (and we all are students), the methodology provides not only a roadmap, but also driving instructions. For that reason the development of the methodology is heavily annotated throughout to contextualise and to explain, so that this work might have lasting value beyond its immediate purpose.

327. In terms of method, this work demonstrates that corpus-based approaches to language can be adapted productively, and in a theoretically sound manner, to social topics. It has done so specifically by adapting an established method used for the study of linguistic variation across textual genres to the study of such variation across social groupings. The emergence of hybrid society has been accompanied by a vast expansion in the ready availability of linguistic data. That expansion has not gone unnoticed across the disciplines, and much effort has been put towards utilising such data. However, we all tend to reach for the most proximate tool (be it theory or method), and in many cases language is approached as some isolated code to be broken or riddle to be solved, fundamentally mistaking its emergent social nature. This work demonstrates the use of tools that were developed to attend to language as meaning, as <u>human</u>, and shows that corpus-based approaches—developed to study patterns in language at scale—are well suited to studying patterns in society at scale in a manner that preserves the humanity of the phenomena we study, rather than obliterating it.

328. Central to this work have been theory and perspectives drawn from communication studies. This field of study has a potentially important role to play in the study of hybrid society, and moreover in helping the social sciences overall to adapt to hybrid society. This work has drawn on a number of traditions in communication studies—notably the studies of mass media, organisational communication, and rhetoric—to show that in concert they can put conceptual order and sense to hybrid society. Social theories of communication will necessarily prove valuable across the disciplines as we navigate changing social phenomena in hybrid society. The study of communication may likewise prove itself of similar value, if it first integrates its own disparate traditions that now stand divided across academies. This work demonstrates that these traditions are not disparate in themselves, but rather are inherently complementary when applied to large-scale social phenomena.

329. Finally, this work shows that empiricism leads the social sciences astray. To consider that empirical evidence is the measure of 'truth' in a social world is to fundamentally misunderstand the social. Tangible phenomena, from which we obtain our empirical evidence, will never be more than an echo of deeper intangible social realities. This work shows that no quantity or type of data will contain the truth, and no method or technique can squeeze truth from it. Theory must be the foundation of work, and to it must be harnessed appropriate method. This work demonstrates that methodology for hybrid society—and for social science writ large—must foremost draw on theory that is rooted in awareness of the ineffable and inescapable humanity of social phenomena. Social theory, that is, properly human social theory, serves to guide and check empirical work ; when empiricism leads, then the social is mistaken and obscured. While this notion is not new by any means, this work demonstrates practically how we might go about repair—how we might take an empiricist approach to a social phenomenon, and remake it in our image, so that we might again see ourselves clearly.

7.3. Limitations of the Work

330. This work faced two main limitations:

1) beyond its interdisciplinary and thus heterogenous character, it is quite frankly challenged by a hodge-podge of thinking and technique ; and
2) it is simultaneously burdened by having too large a scope.

331. The interdisciplinary path, whether taken by happenstance or intention, is not easy. In the case of this work, there were many false starts and dead ends. While the whole is theoretically sound, the elaboration of it—that is the development of suitable method, and the integrative step toward methodology—suffered. The end result was effectively cobbled together and jury-rigged from separate elements that, given the timing and history of the project, could not be redone and had to suffice. In this aspect, the work is not unusual, and is perhaps representative of most scholarly work being constrained by quotidian realities. Nevertheless, in the final analysis the work overall, and importantly the methodology that it presents, is not a velveteen rabbit pulled from a crumpled hat. To the contrary, this work shows that a dedication to step-wise theory allows us to productively navigate the challenges of practical research in hybrid society.

332. At the same time, such dedication to step-wise theory was enormously consuming in terms of time and attention. Once again, the elaboration of theory into method, that is, moving from the conceptual towards tools for research in the world, suffered. There are simply too many issues to address , too many questions to consider, and that necessarily takes time away from the actual study of society. That inescapable fact underlies the value of the disciplines. Disciplinary work rests on established approaches, on accepted and tested understandings, thus freeing the student to push on rather than to recreate the wheel. Thus the methodology of this work is <u>inter</u>disciplinary—it does not call for a new field of hybrid society studies or anything of the sort. The knowledge and methods applicable to hybrid society are there, but we must gather them together. That gathering must be based on an acknowledgement of the core roles played by the various disciplines to inform our work, while recognising that the work must be collective, and that none will go far on their ownsome.

7.4. Further Work

333. In reflecting on the development of the methodology, and on the conceptual and practical components of the work generally, six areas to be taken forward are noted:

1) the integration of traditions in communications studies to further its engagement with sociopolitical phenomena ;
2) the elaboration of corpus-based approaches to public discourse, and to social topics generally ;
3) the investigation of pervasive linguistic features as potential endogenous measures for the study of social groups in hybrid society ;
4) the refinement of Composite Variable Analysis for use in social research, in conjunction with the development of clustering methods adapted specifically to social phenomena ;
5) deeper consideration of assumptions of randomness in certain measures, in light of the non-randomness of social phenomena ; and
6) the development of a general methodological framework for interdisciplinary pedagogy.

334. This work has demonstrated the value of approaching hybrid social phenomena through a combination of various traditions of communication studies. Specifically this work has relied upon traditions of mass media, organisational communication, and rhetoric. These traditions are naturally complementary, and there is strength in their union far beyond that of division, yet nevertheless these traditions are divided by academy, faculty, school, and so on. Concerted effort should be made to integrate these traditions so that communication studies may more productively engage with pragmatic, sociopolitical topics in hybrid society, so that communication studies may be made stronger and deeper in itself, and so that the study of communication may better support other disciplines in their engagement with hybrid society.

335. It has been observed that "corpus linguistics has the potential to reorient our entire approach to the study of language" (McEnery and Hardie, 2012, p. 1). This work, which has drawn much from that broad approach, argues from the study of communication that, as language structures give rise to social structures, corpus approaches to language in society have the potential to reorient our entire approach to the study of society. This work has taken a cut at that, but much work is yet to be done. It should be done—but moreover, given the challenge of hybrid society, it must be done.

336. Pervasive linguistic features show promise for the study of social groups, especially in public discourse and in hybrid society. As has been observed in this work, many approaches to language in the social sciences are informed by information- and communication-theoretic understandings, and rely on salience and markedness. Practically speaking, this is understandable, given the weak signal produced by pervasive features. However, if we can learn to see the patterns in the noise, we will have a powerful tool with which to view society at scale, literally on its own terms.

337. The use of Composite Variable Analysis (CVA) proved an invaluable aid during the factor analysis stage of the method. In comparison to Principal Axis Factoring (PAF), the other technique employed in the method, CVA showed distinct advantages. By design, the CVA algorithm does not front-load as heavily as PAF, essentially factor weightings are more evenly distributed when applied to highly correlated items. This was advantageous for subsequent clustering, as the more even distribution allowed for clusterings that were more appropriate for social phenomena. In addition, the CVA algorithm is far simpler in terms of structure and mathematics, and thus simpler to teach, to learn, and to implement. Finally, it proved to be approximately 100 times faster than the CVA algorithm on this dataset. The CVA technique should be refined for use in social research (as opposed to behavioural, whence it comes), ideally in conjunction with the development of clustering approaches that are sensitive to the nature of social phenomena.

338. Deeper consideration is needed of assumptions of randomness in certain measures common in social science. Assumptions of randomness do not map well to social phenomena, which fundamentally are not random. Thus while such assumptions are intended as a statistical fix to certain problems, they may well create others depending on the domain in which they are applied. This issue is not unrelated to the development of socially sensitive approaches to clustering just mentioned. In adapting method to hybrid society, the question of assumptions of randomness is an underlying issue that will need attention.

339. Finally, this work is an example of interdisciplinary methodology with a pedagogical focus. However, it is only an example, and furthermore it is brought to bear at a certain scale and at a certain level of analysis. Beyond the traditions of disciplinary training, the broad challenges of hybrid society to social research will require students that additionally are trained to understand and to work across disciplinary boundaries. Methodologies such as that presented here should be developed—more specifically, generalised—into full, top-to-tail methodological frameworks for the purposes of teaching and training. Knowledge of theory is often a distinct and persistent challenge for students of all levels, consequently it is often a mechanism of gatekeeping and sorting ; better we banish that challenge and the shadows it casts by laying bare the mechanisms through which we work. In that manner, we may all understand better and more clearly, so that in the study of hybrid society we may likewise see ourselves better and more clearly.

Appendices

Appendix A. Conceptual Dimensions of Political Participation

1. As presented in **chp. 3 § 3.2.2**, Stuart Fox (2014, pp. 497–498) lays out nine criteria, or rather dimensions, with which definitions of political participation can be compared in terms of the underlying conceptualisations. Those criteria are whether political participation is conceptualised as:

1) an active or passive behaviour ;
2) an individual or group activity ;
3) an instrumental or symbolic activity ;
4) a voluntary or mobilised activity ;
5) necessarily having deliberate aims or allowing for unintended consequences ;
6) a conventional or unconventional activity ;
7) necessarily having tangible influence or accepting as sufficient the intent to influence ;
8) having a governmental target or non-governmental target ; and
9) necessarily achieving an intended aim or allowing for failed attempts.

The following is a step-wise engagement with each criterion, evaluating their continued validity in light of emerging understandings and evidence of political participation in hybrid society.

2. **Criterion one**, conceptualising participation as active or passive, derives from a long-standing notion of the study of political participation: "we are concerned with doing politics, rather than with being attentive to politics" (Verba, Schlozman and Brady, 1995, p. 39). Yet participation as an action or activity, while acknowledged as a crucial component of the concept of participation (Brady, 1999, p. 737 ; van Deth, 2016, p. 3), has not been an uncontested idea. The crucial question is where to draw the line between activity and passivity (Brady, 1999, pp. 738–739, provides a helpful overview). The more 'conventional' approach considers as passive information gathering, news watching, and political discussion (Verba, Schlozman and Brady, 1995, p. 39) ; while these are activities, they are seen as insufficiently instrumental to be considered active (here criteria one and three overlap). Other researchers have rejected this approach as too reductive, and have argued for the inclusion of 'passive' activities such as attendance at rallies, expressions of political interest, and so forth (Conway, 2000, pp. 3–4). There has also been a growth in support for the inclusion of communicative acts such as discussion (Bennet, Flickinger and Rhine, 2000 ; Delli Carpini, Cook and Jacobs, 2004, pp. 318–319) and self-expression

(Scheufele and Eveland Jr., 2001 ; Stanyer, 2005). From the perspective of operationalisation, a well-defined (though expansive) distinction between active and passive could be drawn at the point of observability—thus including behaviours undertaken in political context, but excluding interior states such as motivations or attitudes (Brady, 1999, p. 737 ; van Deth, 2014, pp. 355–356). However, in the context of the Internet, due to the nature of online infrastructures and business models, the distinction between manifest and latent behaviours is not straightforward. Brian Krueger, in noting the blurring line between passive and active participation online, makes a significant observation: "Candidates, political parties, interest groups, and political news outlets can elicit information from passive seekers of information on their Web sites without the individuals necessarily knowing they are relaying political inputs" (2002, p. 483).[1] Furthermore, there are suggestions that even sub- or pre-discursive practices, such as information-seeking, are an important aspect of online participation (Gil de Zúñiga *et al.*, 2010 ; Linaa Jensen, 2013). Gibson and Cantijoch note that the online context can "elevate" passive behaviours into active participation (2013, pp. 704, 714). Yet it is important to note that even 'passive' online behaviours are consequential in that they can result in the political mobilisation of individuals or networks of individuals (Theocharis, 2015, pp. 7–9 ; cf. Tufekci, 2014b, pp. 204–205). Thus, the active–passive distinction does not provide a useful conceptual building block for the online context in that all behaviours and practices are potentially consequential.

3. Criterion two, conceptualising participation as an individual or group activity, is something of a throw-away—it is generally not discussed ; rather researchers usually indicate that participation is both an individual and group activity. Verba et al., however, indicate the importance of the question for shaping inquiry. The antecedents of participation (e.g. resources, inequalities, opportunities) are not distributed equally among groups and individuals. Moreover, they are not distributed equally among individuals within groups. As such, the motivations and targets of participation are varied, as are manifestations of participation (Verba, Nie and Kim, 1978, pp. 10–12). This variation partly explains the theoretically blunt character of categoric approaches to the

[1] The use and abuse of user data collected online is a topic of growing concern among the public, decision-makers, and the academy, and one that speaks directly to the question of active versus passive participation.

study of participation (see **chp. 3 § 3.1.2 ¶ 108**). Jakob Ohme notes that the changing nature of social ties in hybrid society seems to have the curious effect of strengthening the perception of individual citizenship and collective citizenship simultaneously (2018, p. 6 ; cf. Bennett, 2012, p. 22). How the individual–group (i.e. micro–macro level) distinction can be understood and studied in the online context—and how we can move toward meso-level approaches—is an open question ; addressing this question is a primary goal of this work.

4. Criterion three, conceptualising participation as an instrumental or symbolic activity, poses a significant logical problem. M. Margaret Conway notes that such a distinction is difficult to make in reality, as various forms of participation can be one, the other, or both. However, she does indicate the effect that the distinction has on the framing of study of the subject—an instrumental conceptualisation should lead to an examination of the relationship between patterns of participation and the distribution of goods in the system ; a symbolic conceptualisation should lead to an examination of how, why and by whom such modes are used, and with what consequences (Conway, 2000, pp. 12–13). Yet the distinction is not made by the researcher alone. The actor will have their own understanding of the nature of their actions, as will targets or observers of those actions. These understandings are neither fixed nor isolate—the foundation of the study of political participation is a concern with the less powerful (participants) communicating desires to the more powerful (decision-makers), and so influencing behaviour. The process is not one way, as behaviours of the less powerful are conditioned by behaviours of the more powerful, and vice versa (cf. Verba, 1967, pp. 55–56). Furthermore, for all of these entities, there is a difference between the manifest purposes of an act and its latent consequences (Merton, [1949] 1968, p. 114). To attempt the articulation of these understandings, across entities and through time, leads us down the rabbit hole. Nevertheless, the conceptual distinction, or at least instrumental focus, has persisted in empirical work (Conway, 1991, p. 32 ; Gibson and Cantijoch, 2013, pp. 703–704). This is likely related to practical concerns with observability and the relative conceptual ease of classifying more 'conventional' forms of participation (this is at the root of the behavioural approach, as discussed in **chp. 3 § 3.1.1**). In addition, the focus on instrumentality is similar to the concern with the "most effective modes" of participation

(Verba, 1967, pp. 56–59),[2] which at base is rooted in normative framings of enhancing participation to the end of bolstering democratic systems (Dalton and Klingemann, 2011, p. 331 ; van Deth, 2001, p. 2 ; Wimmer *et al.*, 2018, pp. 2–7). As the study of participation has adapted to an expanding domain, multiplying forms, and the impacts and character of hybrid society, as discussed throughout **chp. 3**, the focus on instrumentality and symbolicity has faded.[3] Henry Brady's survey of empirical approaches to studying participation does not specify this concern as a core concept (1999), and current taxonomies of participation do not address the issue at all (Ekman and Amnå, 2012a ; Theocharis and van Deth, 2018). It would seem rather that the question of instrumentality (and thus symbolicity) has been subsumed into the expanding definition of what is considered to be participation and what are its proper targets—Jan Leighley noted nearly 30 years ago that a shift in focus away from voting was likely to lead researchers to behaviours that were seen as more instrumental to participants in terms of arriving at preferred sociopolitical outcomes (1995, p. 182). While institutionally structured participation has been extensively studied, less attention has been giving to the building of communities to the end of promoting the "public good" ; this is a lack to be corrected in the conceptualisation and design of research (Conway, 1991, p. 32). Current approaches thus stress the importance of context in delimiting and understanding participation (Theocharis, 2015, pp. 9–10 ; Theocharis and van Deth, 2018, p. 144 ; Uldam and Kaun, 2018 ; Wimmer *et al.*, 2018, pp. 6–7). This shift in focus is indicative of the move towards meso-level understandings of participation (see **chp. 3 § 3.2**). (The question of instrumentality, or any discussion of motive, also hinges on conceptual issues of intent and consequence, addressed below under criterion five.)

5. Criterion four, conceptualising participation as a voluntary or mobilised activity, provides a more straightforward case. There is a clear consensus that the study of political participation is concerned with voluntary participation, that is, acts that are not coerced (Parry, Moyser and Day, 1992, p. 3). Such acts are defining, or at least characteristic, of democratic systems (Brady, 1999, p. 737). This does not mean that participation does not exist in non-democratic systems—all polities incorporate public involvement in one

[2] Note that this is a concern with objective efficacy, not participant perceptions of efficacy, which are irrelevant to functional definitions of participation (Theocharis, 2015, p. 9).
[3] Consider this in light of the discussion of sociomateriality in **chp. 2 § 2.1.3**).

manner or another (Dalton and Klingemann, 2011, p. 331). Rather, coerced participation does not fit within the Western academic tradition of participation. While it might very well be active, it is not undertaken by ordinary citizens to influence governing structures ; rather it is undertaken by the establishment, using ordinary citizens to exert influence and to socialise the public.[4] However, Fox's formulation of this criterion seems rather clumsy: "Must political participation be voluntary, or can it be mobilised or forced by institutions and/or other people?" (2014, p. 497). We have clearly excluded the consideration of forced behaviour from political participation. But to group mobilised and forced behaviours in opposition to voluntary behaviour is not in keeping with how participation is conceptualised. The fundamental concept of civic life is grounded in mass mobilisation ; as John Dewey notes, to understand the emergence of publics and the State, we must begin with "the objective fact that human acts have consequences upon others, that some of these consequences are perceived, and that their perception leads to subsequent effort to control so as to secure some consequences and avoid others" (Dewey, 1927, p. 12). The State is the manifestation of mobilised citizens (compare Mumford's idea of 'human machines' ; **chp. 2 § 2.1.1**).[5] In terms of institutional political participation, the State mobilises citizens for legitimacy, information-seeking, decision-making, etc. (Dalton and Klingemann, 2011, p. 331 ; Verba, 1967, pp. 57–58). Parties mobilise citizens for similar reasons, and to gain influence in (and constitute) the machinery of State (Diamond and Gunther, 2001, pp. 7–9). Mobilisation, be it institutional (e.g. through parties or campaigns), social (e.g. through informal discussion or information exposure), or internal (i.e. arising from individual motivations or attitudes), is understood as both a central driver and consequence of political participation (Leighley, 1995, pp. 188–191). In that participation is also understood as a function of resources and opportunity (e.g. Verba, Nie and Kim, 1978, chp. 3), the relative ease, affordability, and rapidity of networked technologies and platforms is an enabling context, which can be understood as a form of

[4] There has been work done on such phenomena in authoritarian constructs, but this tends to be approached from a perspective of political culture (e.g. Mauk, 2017).

[5] Note that Gabriel Almond and Sidney Verba (1963, p. 4) observed that "Lucian Pye refers to modern social organization as being based on an organizational technology", for which they cite an internal document of the Committee on Comparative Politics, of the Social Science Research Council. The document is entitled 'Memorandum on the Concept of Modernization' and is dated 1961. The author has yet been unable to obtain a copy of the memorandum. On Pye, the Committee, and the Council, see **chp. 4 § 4.1.1.2.2 ¶ 162ff**.

external (i.e. non-internal) mobilisation (Bennett, 2012, p. 30). What has been observed is that political participation in this context, due to the networked nature of technologies and platforms, is an essentially mobilising act (Theocharis, 2015, 5). The former concern with whether participation was "spontaneous" (i.e. internally mobilized) or externally mobilized (Verba, Schlozman and Brady, 1995, p. 136)—which is essentially what this criterion addresses—is perhaps a hold-over from approaches to participation that were based on categoric understandings of citizens (e.g. Verba and Nie ; see **chp. 3 § 3.2**) and thus exaggerated and distorted attitudinal drivers of participation, while they obscured variation (temporal, spatial, and social) in participation within categories and throughout the population (Leighley, 1995, pp. 186–188). Following current understandings of participation in hybrid society, and building upon the foundation of seminal understandings of the nature of polities, it is reasonable to suggest that political participation and political mobilisation are coextensive concepts—the former focused on what is done, and the latter focused on why and how. The concepts will be considered synonymous here.

6. Criterion five, conceptualising participation as necessarily having deliberate aims or allowing for unintended consequences, hinges on intent, observability, and interpretation. (Criteria seven and nine likewise hinge on the same, and thus are conceptually equivalent to criterion five for the purposes of the following discussion. Also, as mentioned, criterion three is related to this discussion, as motive implicates intent and consequence.) Sidney Verba engages with the issue, framing a definition of participation that stresses the intention of influencing decision-makers (Verba, 1967, p. 55). He notes that this is somewhat problematic. First, the intentions of citizens may not be well-defined or stable. He suggests that we may safely ignore this problem, so long as we are satisfied that participants intend for decision-makers to "get the message" (ibid., fn. 6). Second, we must attend to the behaviour of decision-makers in order to gauge if the message was received, whether it was communicated with sufficient fidelity ("information content") to allow interpretation, how it was interpreted, and whether or not decision-makers interpolated other issues into the message (ibid., pp. 60–61). Verba rightly acknowledges problems of definition, stability, and interpretation on the part of participants and decision-makers. However, of much greater concern is the part of researchers—how are we to mark and measure something so ill-defined, fluid, and subjective? Patrick Conge notes that this is extremely problematic, and thus dismisses intent as a conceptual component of participation:

> [W]e should detach intentions and outcomes from the definition of political participation. Political participation should be restricted to the acts themselves ; it should not encompass the intentions of individual participants or the outcomes of their actions. Intentions may explain why people participate (without accounting for what political participation is), while outcomes (whether intended or unintended) explain the consequences of political participation (again without accounting for its nature). The aims of individual participants and the consequences of their actions are empirical questions and should not be defined away by including them in a definition of the concept. (1988, p. 247)

7. The last point is crucial: our conceptual definitions serve to restrict and refine inquiry. By positing the character of intent and consequence, we choose to ignore that they are situated and contingent. While such a stance was understandable in the early, behavioural days of the study of participation, to perpetuate such positivist framings contrary to generations of evidence would undermine such study as a scientific endeavour. Yet intent and consequence must be addressed ; as they are situated and contingent, they help to characterise the entities and behaviours of concern. However, these concepts are little elaborated in the literature, with the implication that their meaning is taken as evident. But their meaning is not evident. They are mental representations (Searle, 1983, pp. 11–13), and thus are fundamentally indeterminate from an empirical perspective. Neither can be observed, measured, and causally related with the certainty and reliability necessary for the empirical endeavour, and thus for the development of 'strong' social science theory (Merton, 1967, p. 68 ; Sutton and Staw, 1995, pp. 378–380).[6] As such, not only should they be excluded from definitions of participation, neither should they be indicators of it. Given the problematic nature of the concepts in terms of the study of participation, they arguably are irrelevant to it (Hooghe, Hosch-Dayican and van Deth, 2014, pp. 339–340). Yannis Theocharis provides an alternative approach that brackets the problematic concepts of intent and consequence, yet still addresses the situatedness and contingency of participation—he proposes to look instead at the context in which

[6] Drawing on Abraham Kaplan (1964, pp. 296–298), Robert Merton (1967, p. 68), and Talcott Parsons and Edward Shils (1951), 'strong' social science theory can be understood as a network of assertions and assumptions that predicts, explains, and reveals empirical phenomena across the range of social behaviours and structures, that is coherent across the range of social behaviours and structures, and that integrates levels of analysis.

behaviours are performed (2015, pp. 9–11). The question of context is essential to current approaches to the study of online political participation, whether context is established according to platform (Barberá *et al.*, 2015 ; Bond and Messing, 2015), technology (Margetts *et al.*, 2016 ; Woolley and Howard, 2016), discursive markers (e.g. hashtags, Bruns and Burgess, 2011) or artefacts (e.g. memes, Milner, 2013), or social networks (Maes and Bischofberger, 2015 ; Yardi and boyd, 2010). The awareness of the necessity to attend to context undergirds this work.

8. **Criterion six**, conceptualising participation as a conventional or unconventional activity, has two important aspects—that of 'conventional' or institutional participation versus non-institutional participation, and that of legal versus illegal participation. Questions of whether any new form of participation should be considered as a 'valid' act of participation should be put to bed. Shifts in society and technology alter the pathways of possible civic engagement ; not adjusting conceptually and practically to changing circumstance hinders the study of participation (Norris, 2002). The emergence of novel and increasingly widespread modes of participation online should no longer be doubted, and neither should their importance (Fox, 2014 ; Gibson and Cantijoch, 2013 ; Theocharis, 2015). The question of legality versus illegality is crucial, though beyond the scope of this discussion.[7]

9. As stated, **criterion seven**, conceptualising participation as necessarily having tangible influence or accepting as sufficient the intent to influence, is conceptually equivalent to criterion five for the purposes of this discussion.

10. **Criterion eight**, conceptualising participation as having a governmental target or non-governmental target, is most relevant to the delimitation of the field of study. Political participation, in a broad and unelaborated sense, refers to acts by those without decisional power that influence the acts of those with decisional power (Verba, 1967, p. 55). In such a minimalist framing, it is evident how 'greedy' under-determined definitions can be—they risk encompassing all relational human activity. Over the history of the scientific study of political participation, the subject has been bounded in part by

[7] Note that this question also implicates the issue of covert elicitation and collection of information for political ends noted by Kreuger (see the discussion of criterion one).

specification of the target of participation, that is, the entity to be influenced by a given action. Most approaches have held that political participation targets the government (Almond and Verba, 1963, pp. 117–120)—this framing is based on the idea of actors as citizens, not simply members of society (Almond and Verba, 1963, p. 120). However, even among researchers who stress this specification, it is understood that a governmental target comprises the State broadly in its various organs and services (Pattie, Seyd and Whiteley, 2004, chp. 4), and furthermore that the government can be targeted indirectly (Parry, Moyser and Day, 1992, p. 40). The evolution of this specification relates to the overall expansion of the domain of participation—the steady expansion of governmental influence and involvement throughout society increases the sites, occasions, and drivers of civic engagement (van Deth, 2001, pp. 8–11). At the same time, this process increases the complexity and consequences of the behaviour of all entities (Dewey, 1927, p. 126 ; e.g., Pierson and Skocpol, 2007, pp. 1–5 ; Verba, 1967, p. 58), leading to an expansion of the repertoire of participation as people engage with emerging issues (van Deth, 2001, pp. 3–8).[8] This could be understood as the fundamental process of political participation. The process also extends the civic sphere far beyond the boundaries of government proper—political participation as experienced and understood by people themselves often depends on non-governmental organisations and non-institutional practices, which in addition provide the socialising context for participation regardless of target (Verba, 1967, pp. 56–58). As described in the historical overview in **chp. 3 § 3.1**, research into political participation has not been unresponsive to societal changes.

[8] This refers to the organic repertoire not the conceptual repertoire, that is, those actions which individuals undertake in a civic context and thus as members of a polity, as opposed to those actions fitting a given definition of participation. Furthermore, note that the expansion of the repertoire is equated with emerging issues (an issue, broadly, being the conceptualisation of the circumstances that lead an individual to civic action). It is suggested that a given form of participation motivated by a given issue is specific to a context. The form of engagement and the motivating issue represent a specific sociotechnological configuration. A different form of engagement alters that configuration, and thus changing forms of engagement mark changing issues. For example, tax reform sought at the voting booth should not be considered the same issue as tax reform sought through violent protest ; voter mobilisation among a church congregation should not be considered the same issue as voter mobilisation through social media platforms. To equate nominally equivalent issues is to ignore the differences in context indicated by form of participation.

11. However, sociotechnological changes—especially in information and communication—have proven more difficult to grapple with (Fox, 2014, pp. 500–503 ; van Deth, 2016, pp. 5–6). In studying online participation, the specification of target depends on the role that online behaviours are considered to have in relation to political participation broadly. Often, online behaviours have been treated as the independent variable, generally speaking, with offline behaviours being the dependent variable (e.g. Boulianne, 2015) ; framed in this manner, there is no real need to revisit conceptual specifications of target, in that 'real' participation is implicitly considered to be offline. However, the situation is altered when considering online activity in itself. For some, their concern is that online relational behaviours are predominantly comprised of self-expression or undirected signalling. The underlying concern is that behaviours that might seem political are considered by some to be of dubious political value, stemming from the nearly zero cost and thus nearly zero impact of such behaviours (as observed by Theocharis, 2015, pp. 4–9). This position is weak.[9] The concern with expression and signalling ignores that such behaviours are accepted as political participation in offline contexts, and that indeed such behaviours have become crucial aspects in the expanding repertoire of participation (see the discussions under criteria one, three, and four). The assumption that a low-cost act is necessarily low impact does not hold with the nature of networked communication. A single act may have little impact on its own, but it may activate a network of entities that likewise engage in a single act of little impact, but that may activate a further network. To deny impact to a single online act is little different than denying the impact of a single vote ; more to the point, it is logically no different than denying the impact of the fission of a single atom—the concern is not with individual events, but with the effect of their exponential growth. As mentioned under criterion four, Theocharis defines online participation as a "mobilizing act" that is inherently expressive (Theocharis, 2015, p. 5). How to understand the target of participation online is a complex question ; however, in the context of hybrid society as addressed in this work, the target is assumed to be the public at large. As for this criterion, an insistence on a governmental target to identify participation is to posit a distinction that does not hold in the observed performance of civic life (van Deth, 2014, p. 357). Furthermore, the specification of target

[9] The suggestion of 'zero impact, zero cost' is fallacious considering the interwoven whole of embodied and disembodied lives.

of participation to delimit the scope of inquiry does not hold for online behaviours, because we simply are unable to determine the intended target with any certainty. This challenge is similar to that discussed under criterion five, discussing the indeterminacy of intention and consequence. The solution is also the same—we can bracket the specification of target by focusing instead on the context in which behaviours occur. If we identify a context as political, then we should accept that behaviours in that context are political acts (Theocharis, 2015, pp. 9–11 ; Theocharis and van Deth, 2018, p. 144).

12. As stated, **criterion 9**, conceptualising participation as necessarily achieving an intended aim or allowing for failed attempts, is conceptually equivalent to criterion 5 for the purposes of this discussion.

Appendix B. Lexical Factors

The following are the lexical factors produced in the factor analysis stage of the method (**chp. 5 § 5.1.4**). Specifically, these are the 8-degree factors produced by Composite Variable Analysis. There are four sets of lexical factors presented in two groups. The first group is the List 1 words and bigrams, those being distinctive pervasive items. The second group is the List 2 words and bigrams, those being non-distinctive pervasive items. Note that distinctiveness was calculated at the user–document level, and non-distinctiveness at the corpus level ; thus a given lexical item can be both distinctive of a given user–document, while simultaneously non-distinctive in the corpus overall. The items in each factor are ordered according to the order of variable composition (i.e. the most correlated items appear first). As explained at the end of **chp. 6 § 6.1**, the factors produced by Principal Axis Factoring were abandoned for the analysis of results, and thus are not included here. **NOTE:** The following items are presented as they were attested in the corpus. The lists contain terms that may be found offensive.

List 1. Distinctive Pervasive Words

Factor 1

Lmaoooo Lmaooooo deadass lmaooo Yall youre
doetryinIk ik woahFr

Factor 2

innings inning pts rebounds TDs TD
Slay LB

Factor 3

Thx thx ty Ty faculty eg
patients prof

Factor 4

UM B1G LET'SEND BREAKING UPDATE
Tune ya'll

Factor 5

illegals treasonous Liar scumbag Church council
blah comics

Factor 6

y de gotchu bass luv ily
aw UGH

Factor 7

trails underway volleyball Freshman Outstanding Wrestling

Factor 8

Sisbraids Ppl Plz BRO HMU
Awe S/O

List 1. Distinctive Pervasive Bigrams

Factor 1

think.deserves Click.the [PN].ClickI.cast player.you cast.my
ballot.for [PN].ballot the.lions [PN].state [NUM].p.m. birthday.bro
our.country You're.a the.leftyou.think

Factor 2

[VOC].& &.[VOC] &.I bc.I Thanks.for for.sharing
Thanks.[PN] you.Thank you.for thank.you am.so [NUM].am

Factor 3

vs.[PN] [PN].vs vs.[VOC] [NUM].pts [PN].leads[PN].lead
[NUM].left [NUM].lead [NUM].yards [NUM].points[NUM].games [PN].team
this.team the.league Big.[NUM] Red.Wings

Factor 4

[VOC].happy happy.birthday Happy.birthday birthday.[VOC] love.you Love.you
the.gym my.hair i'm.notand.i'm [VOC].i'm [VOC].oh
[JUNK].[JUNK] [JUNK].[VOC] [VOC].omg [VOC].lmao

Factor 5

i.can so.ii.was when.i[VOC].i i.love
i.am i.miss i.just but.i and.i i.have
i.need i.got i.can't i.don't

Factor 6

lol.[PN] [PN].lol [VOC].lol Lol.I I.dont i.dont
bout.tois.gonna gone.be a.nigga I.ain't I.been
I.gotta out.here I.be as.hell

Factor 7

[HASH].[NUM] [NUM].[HASH] [JUNK].[HASH] [HASH].[JUNK] [VOC].[VOC]
[NUM].[VOC] at.[VOC] Hey.[VOC]

Factor 8

Go.Blue GO.BLUE Good.luckLet's.go Happy.Birthday
Good.morning it.Don't my.dog

List 2. Non-Distinctive Pervasive Words

Factor 1

shit ass whole hell hate myself
everything wish someone anyone else different
house buy haven't almost friends friend
sorry sad love happy home miss
face leave together stay live world
change mind

Factor 2

he's He's wasn't wouldn't there's they're
I'd You're pretty probably guess maybe
Yeah Like though stuff went wanted
told knew started found took came
few ago heard saw might remember
name tweet

Factor 3

win team top season play playing
second half gets goes end left
run line far close guy guys
man gonna head hit big move
looks nice fun favorite show watching
yet times

Factor 4

wait Can't hope soon Don't Who
job Good very kind Yes yes
find part hear such Maybe Or
Did Are Is Do When Can
side comes once turn Now At
Well course

Factor 5

money pay won't wants She white
care yourself talk talking ask question
try thinking couldn't gave lost lose
needs against happen happened true matter
understand sense either agree aren't fact
reason problem

Factor 6

We our us & family kids
help support Thank Thanks All Love
please Please [JUNK] [HASH]

Factor 7

school high tonight tomorrow morninghours
week days looking seeing having break
taking workingcoming bring early late
checkseteach able past during
call called free open until place
making its

Factor 8

which means betweenAlso read idea
use used under says instead deal
without less full rest news country
must Your ThereAsmay story
In For state vote

List 2. Non-Distinctive Pervasive Bigrams

Factor 1

[PN].[NUM] [NUM].[PN] [NUM].[NUM] by.[PN] a.[NUM] [NUM].and
for.[NUM] the.first

Factor 2

[VOC].[VOC] [PN].[VOC] for.[PN] Thank.you will.be at.[PN]
for.the [HASH].[HASH]

Factor 3

I.am that.I it.is I.would

Factor 4

The.[PN] [PN].are like.[PN] [PN].but

Factor 5

I.love so.much a.good to.a

Factor 6

have.been has.been be.a should.be

Factor 7

you.have If.you to.have you.are

Factor 8

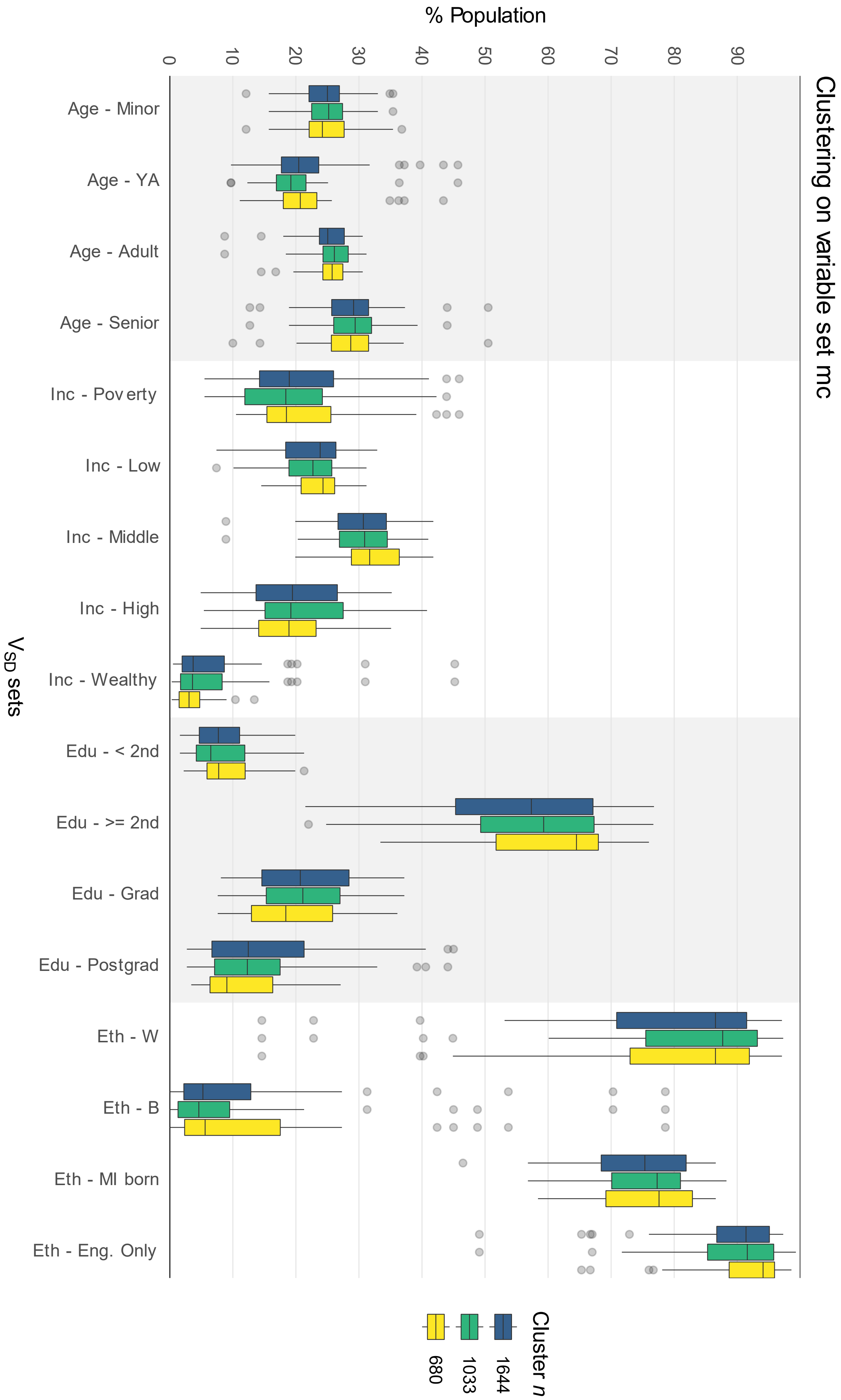
to.me you.can [JUNK].[JUNK]

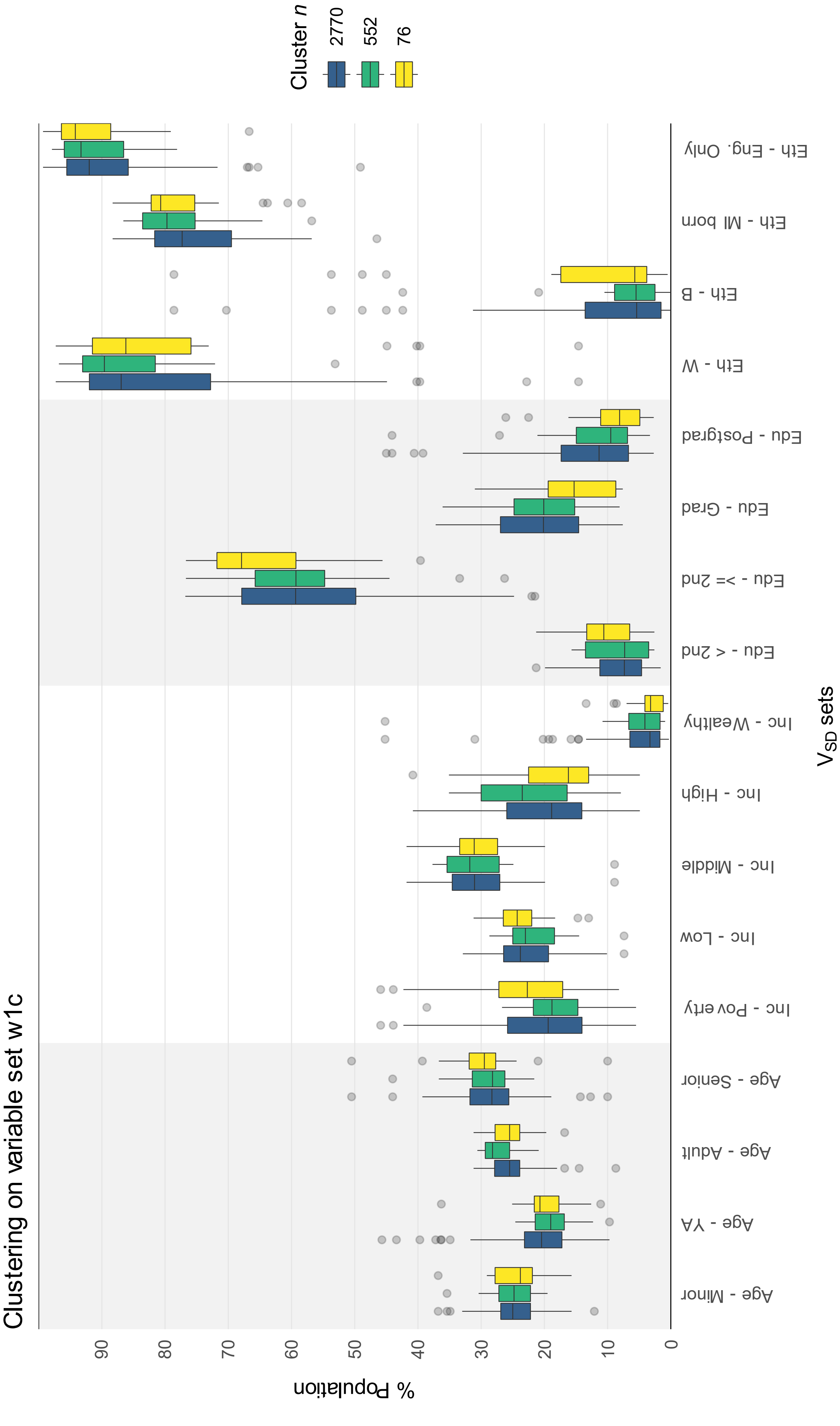
Appendix C. Sociodemographic Profiles

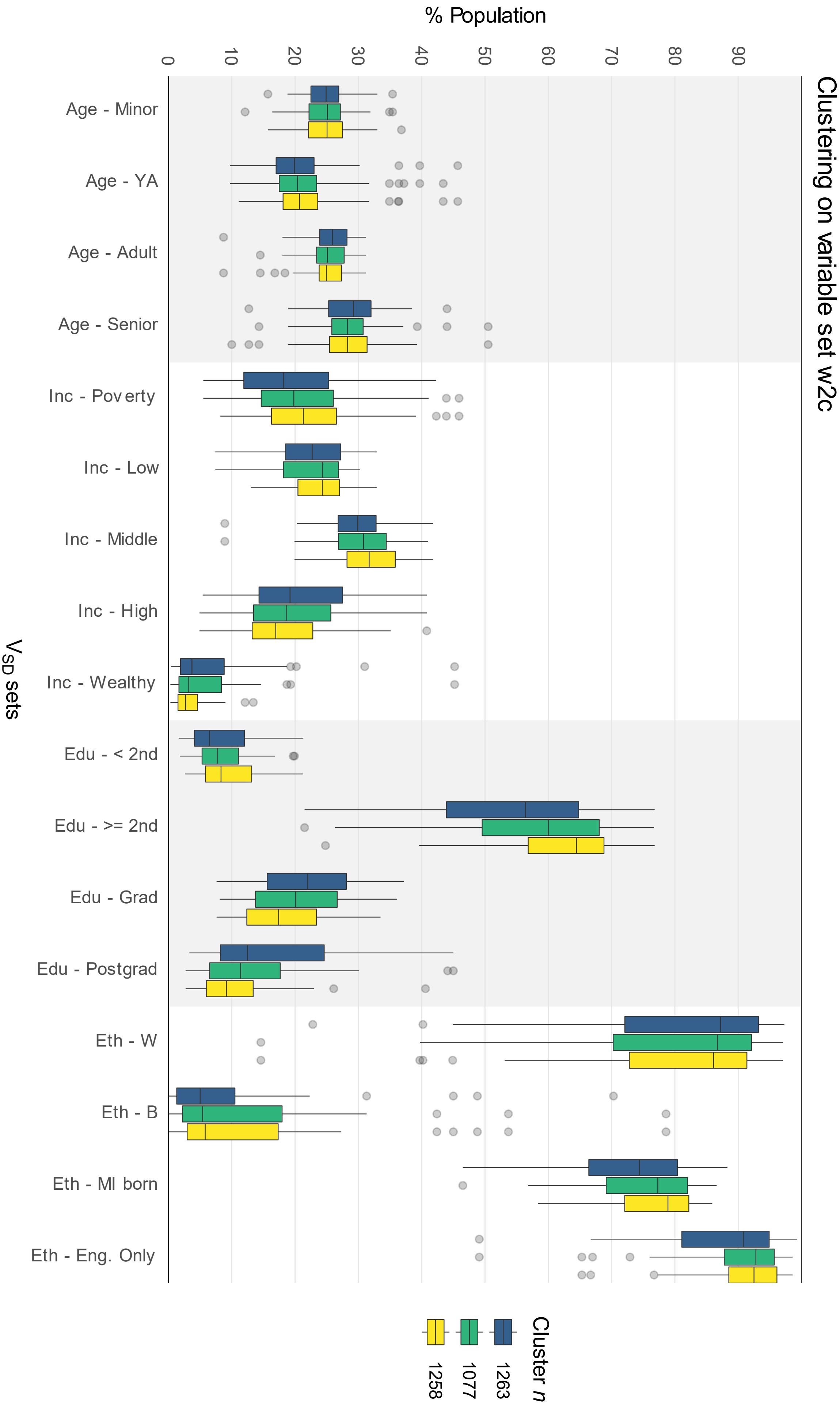
The following are sociodemographic profiles of linguistic clusterings as presented in step 2 of the analysis (**chp. 6 § 6.2.2**). Presented here are profiles at *k* 3 for all linguistic variable sets. Those set labels and the factor scores they denotate are:

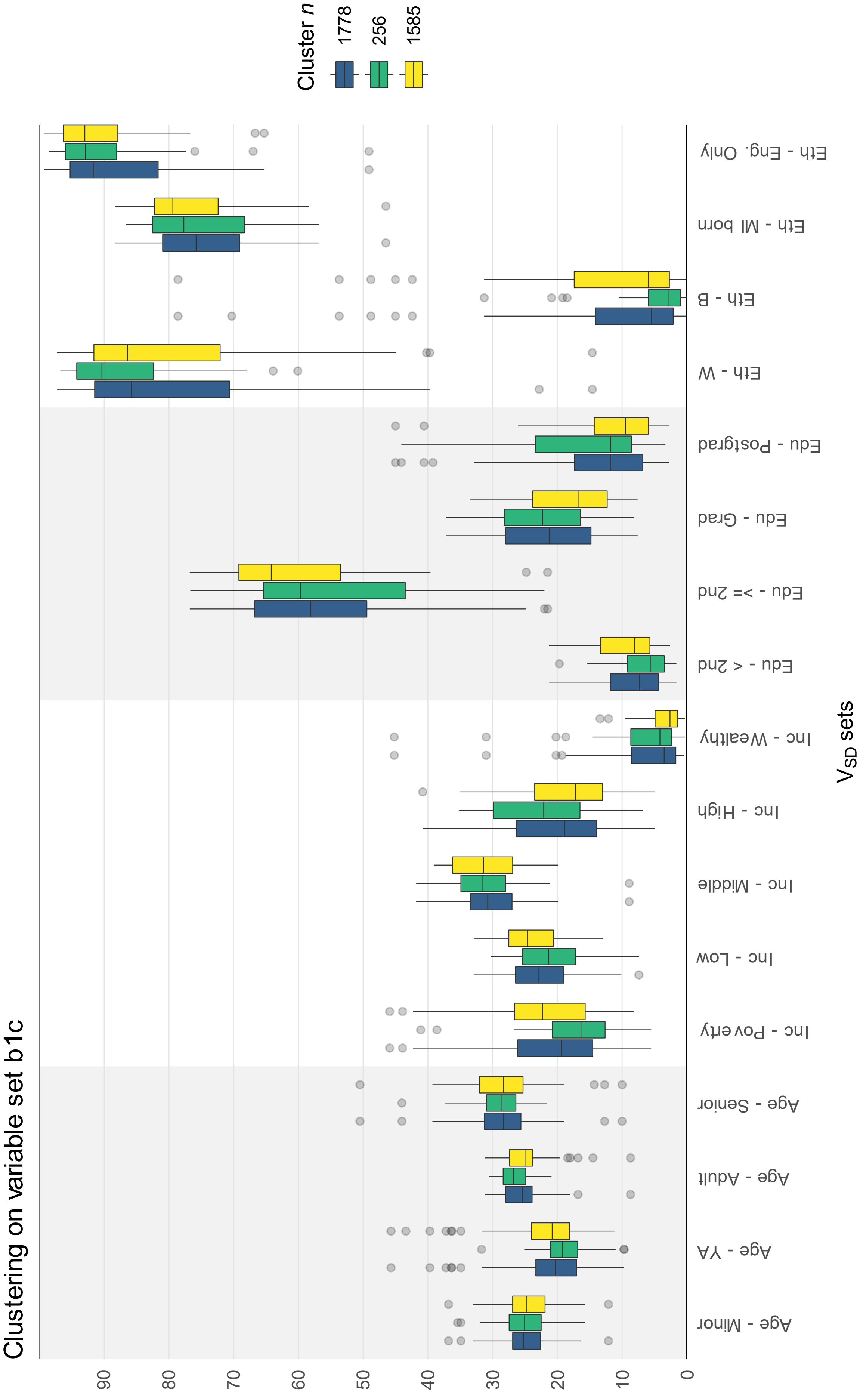
mc	grammar (m is for MAT ; see **chp. 5 § 5.1.3.2**)
w1c	List 1 distinctive pervasive words
w2c	List 2 non-distinctive pervasive words
b1c	List 1 distinctive pervasive bigrams
b2c	List 2 non-distinctive pervasive bigrams
w12c	Composite set of List 1 and List 2 words
b12c	Composite set of List 1 and List 2 bigrams
w1b1c	Composite set of List 1 words and bigrams
w2b2c	Composite set of List 2 words and bigrams
mw1b1c	Composite set of grammar and List 1 words and bigrams
mw2b2c	Composite set of grammar and List 2 words and bigrams
mw12b12c	Composite set of grammar and all word and bigram lists

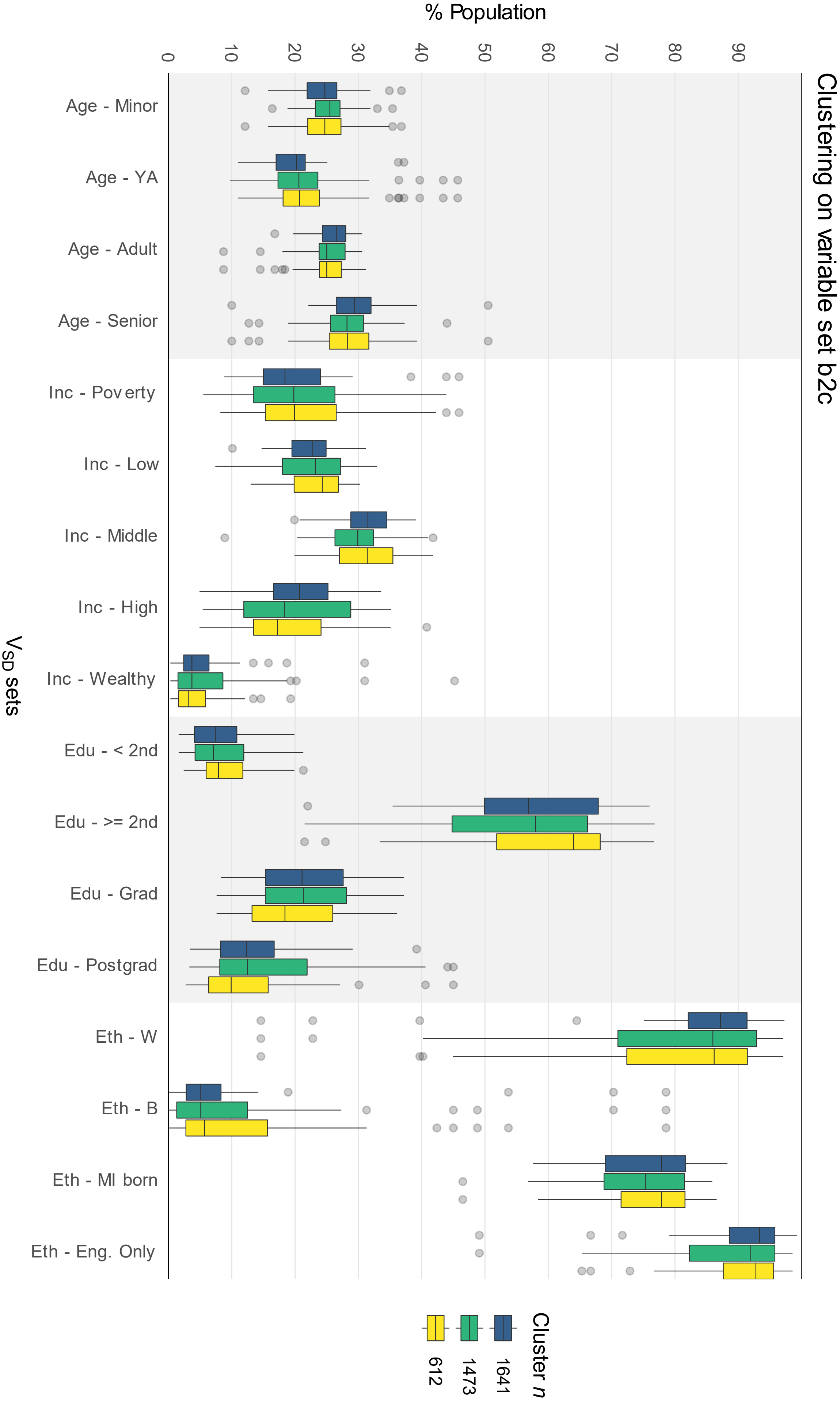
Note that these labels differ slightly for the variable set labels in **chp. 6 § 6.2.2** ; the appended 'c' on each label indicates that the variable set was produced using the Composite Variable Analysis factoring technique. The profiles are presented in the order of the variable sets as listed above. Recall that all clusterings have been aligned according to the education bracket representing prevalence of baccalaureate or associate degrees as the highest level of attainment (Edu - >= 2nd), as discussed in step 2 of the analysis.

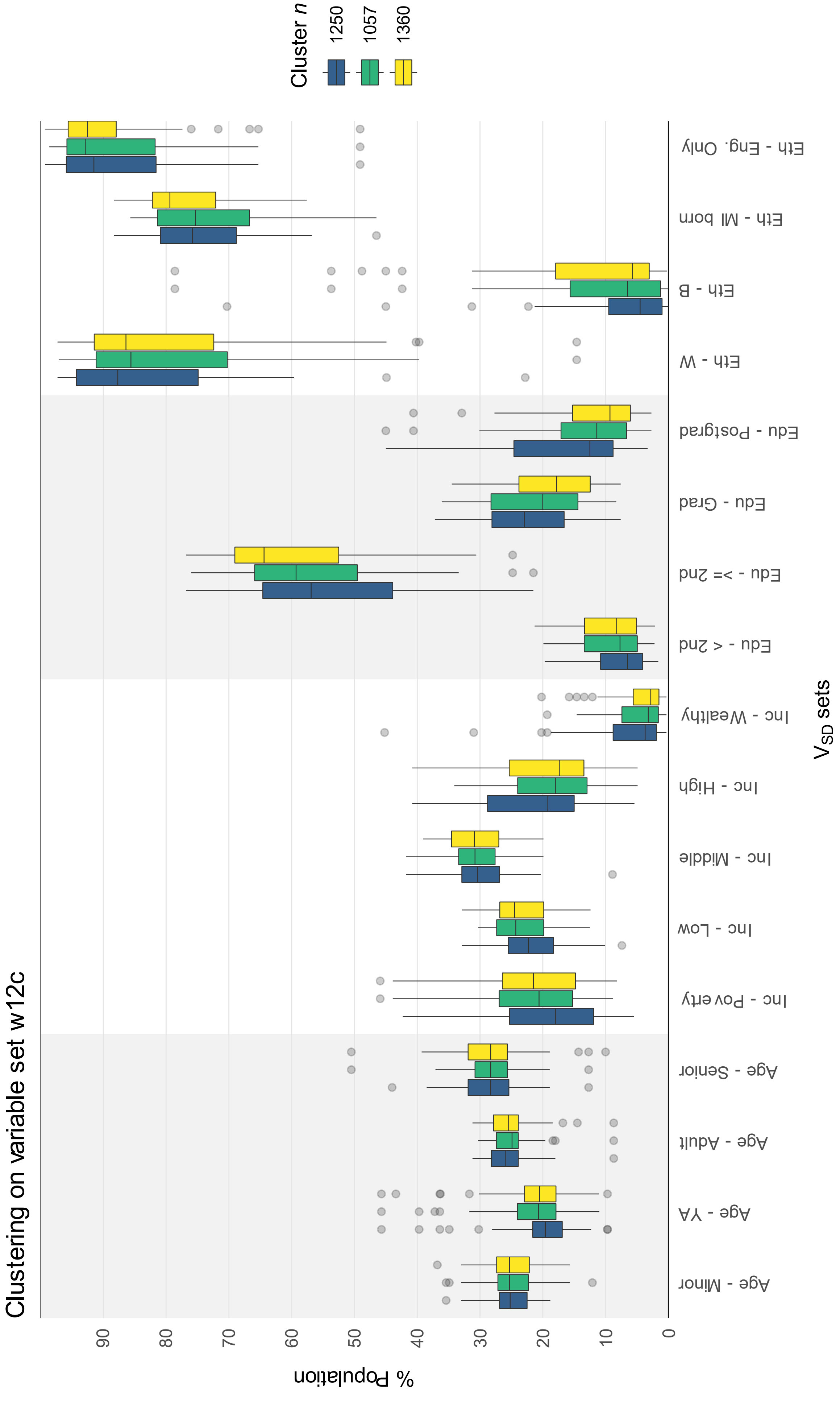

Clustering on variable set mc
% Population
0
10
20
30
40
50
60
70
80
90
VSD sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster n
1644
1033
680

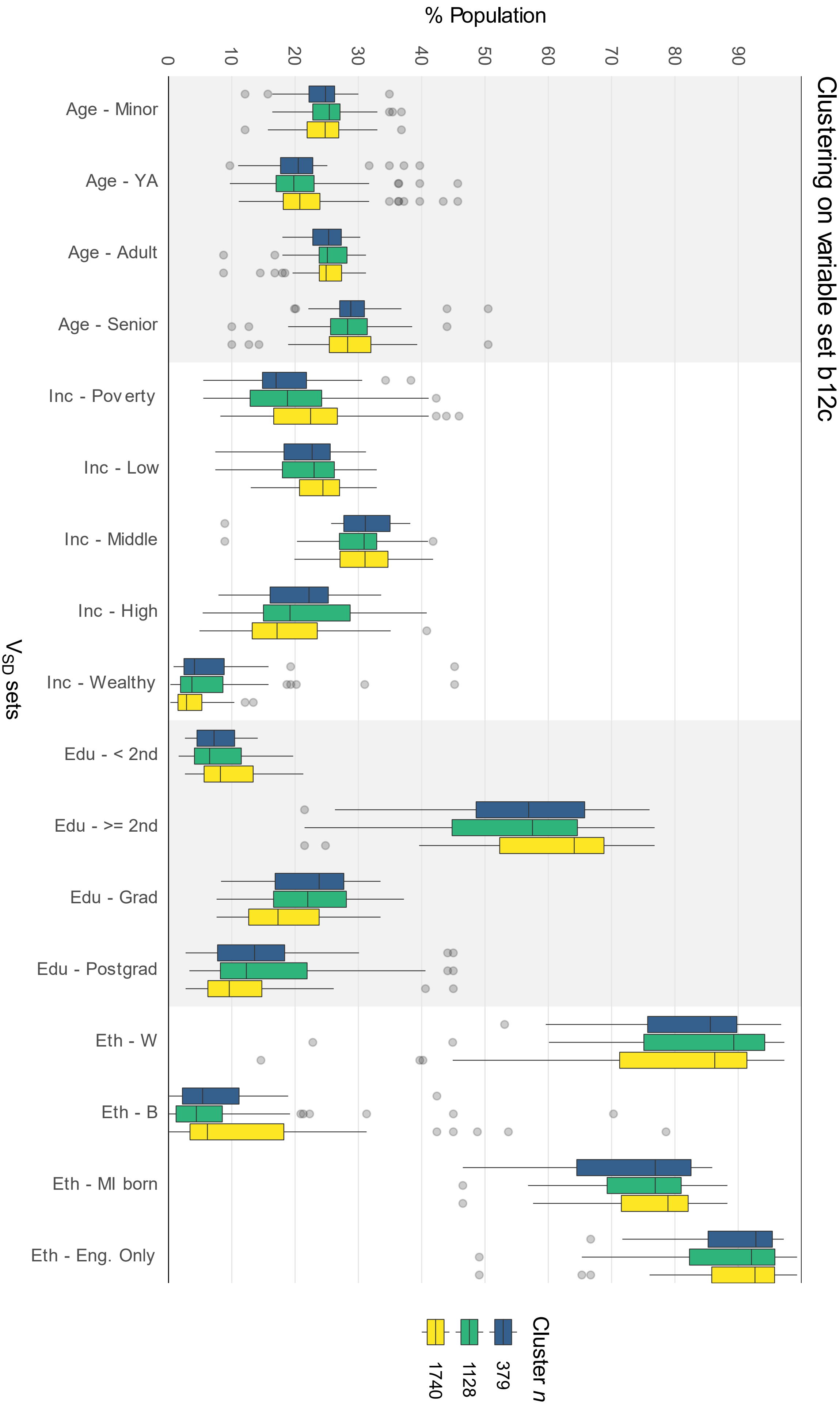

Clustering on variable set w1c
Cluster *n*
2770
552
76
% Population
0
10
20
30
40
50
60
70
80
90
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
V_{SD} sets

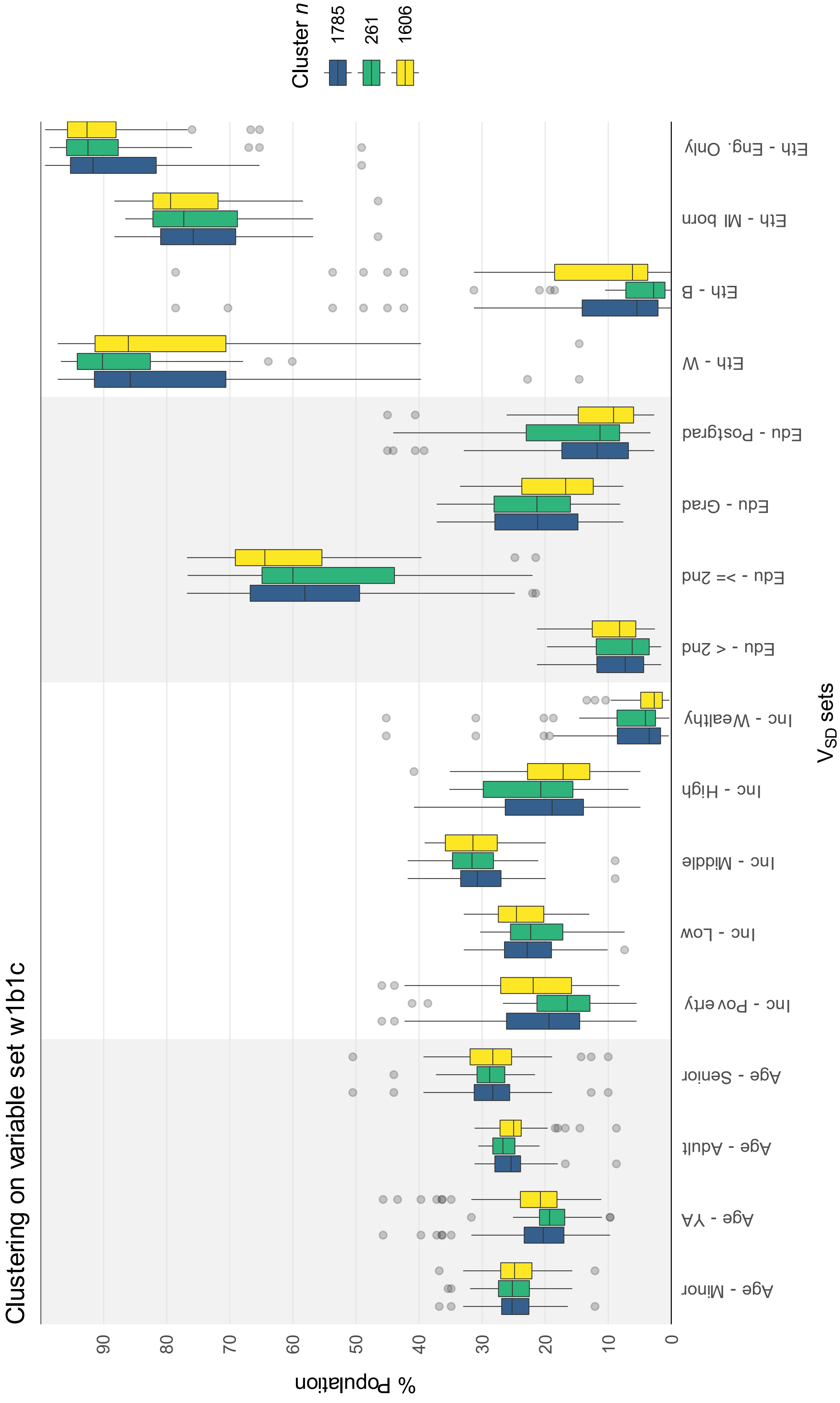

Clustering on variable set w2c
% Population
0
10
20
30
40
50
60
70
80
90
VSD sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster n
1263
1077
1258


Clustering on variable set b1c
% Population
0
10
20
30
40
50
60
70
80
90
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
V_{SD} sets
Cluster *n*
1778
256
1585

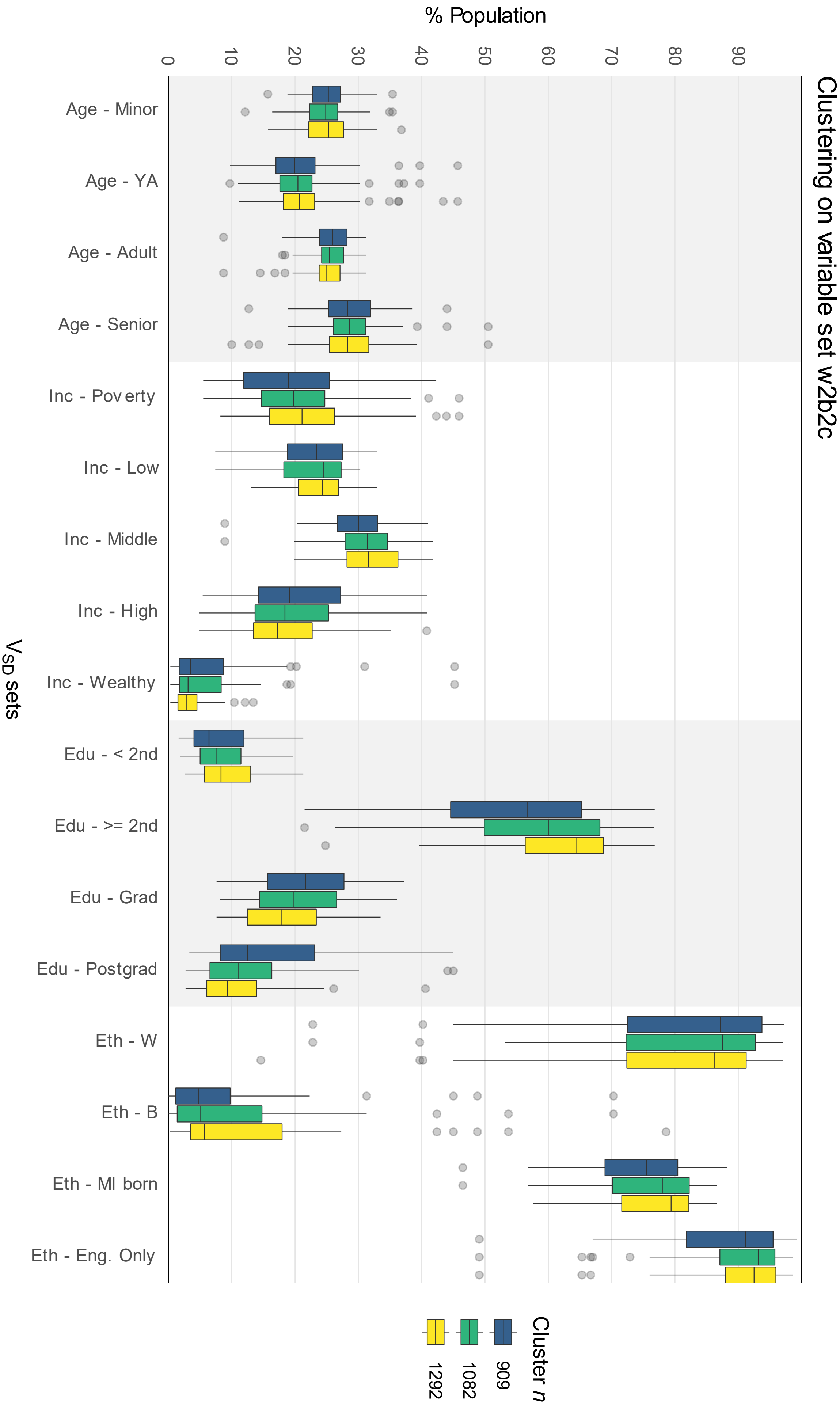

Clustering on variable set b2c
% Population
0
10
20
30
40
50
60
70
80
90
VSD sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Pov erty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster n
1641
1473
612

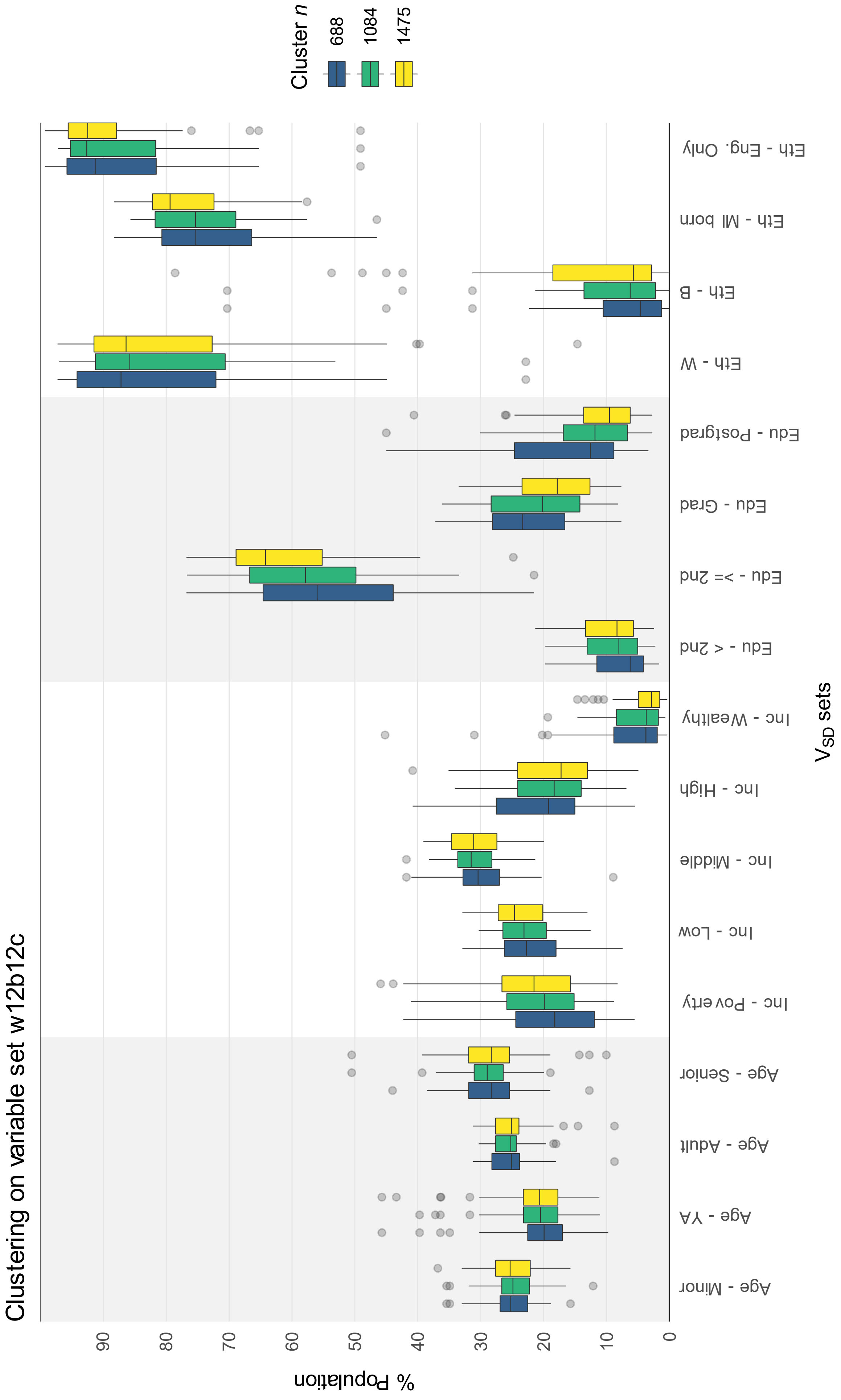

Clustering on variable set w12c
Cluster n
1250
1057
1360
% Population
0
10
20
30
40
50
60
70
80
90
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
V_{SD} sets

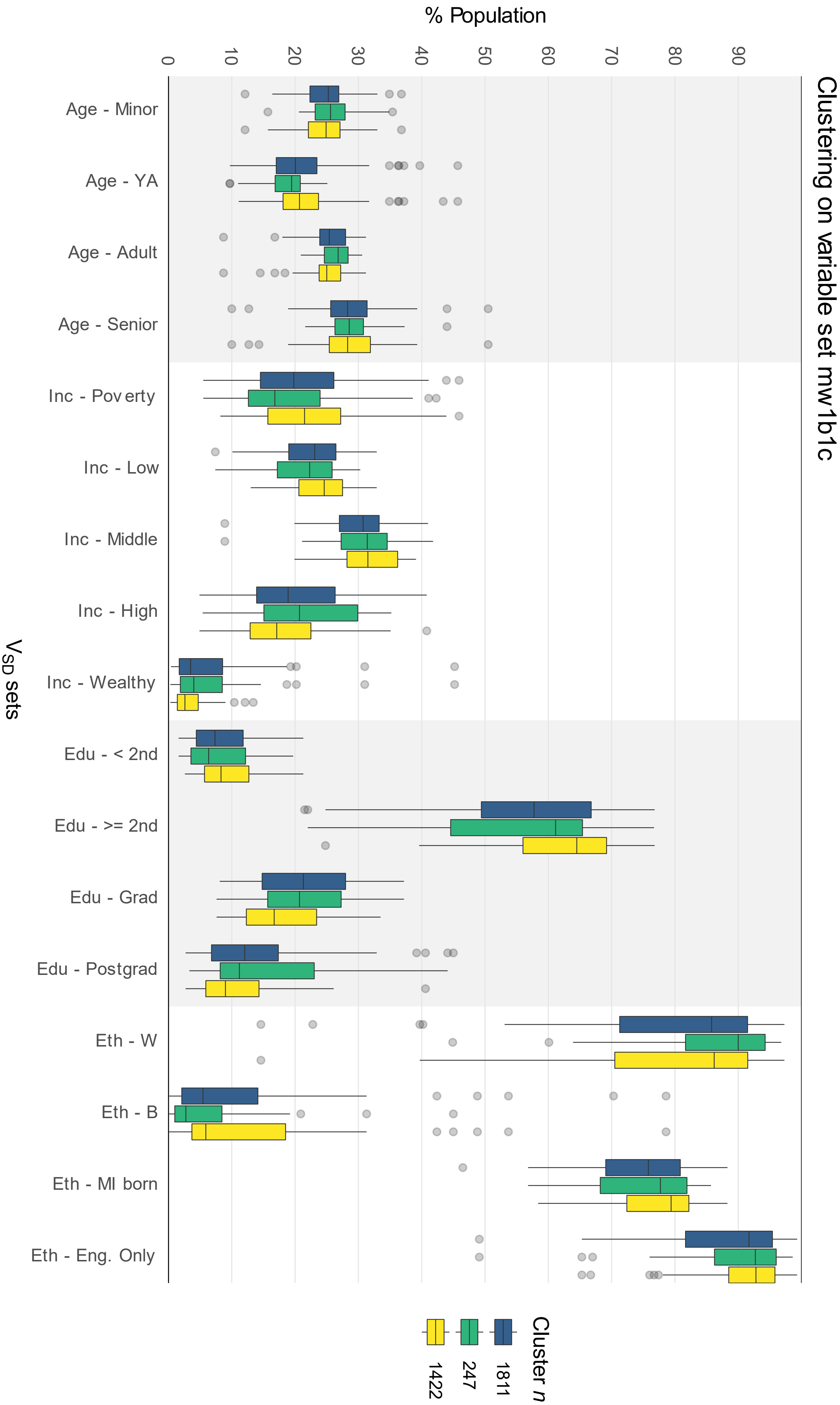

Clustering on variable set b12c
% Population
0
10
20
30
40
50
60
70
80
90
V_{SD} sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Pov erty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster *n*
379
1128
1740

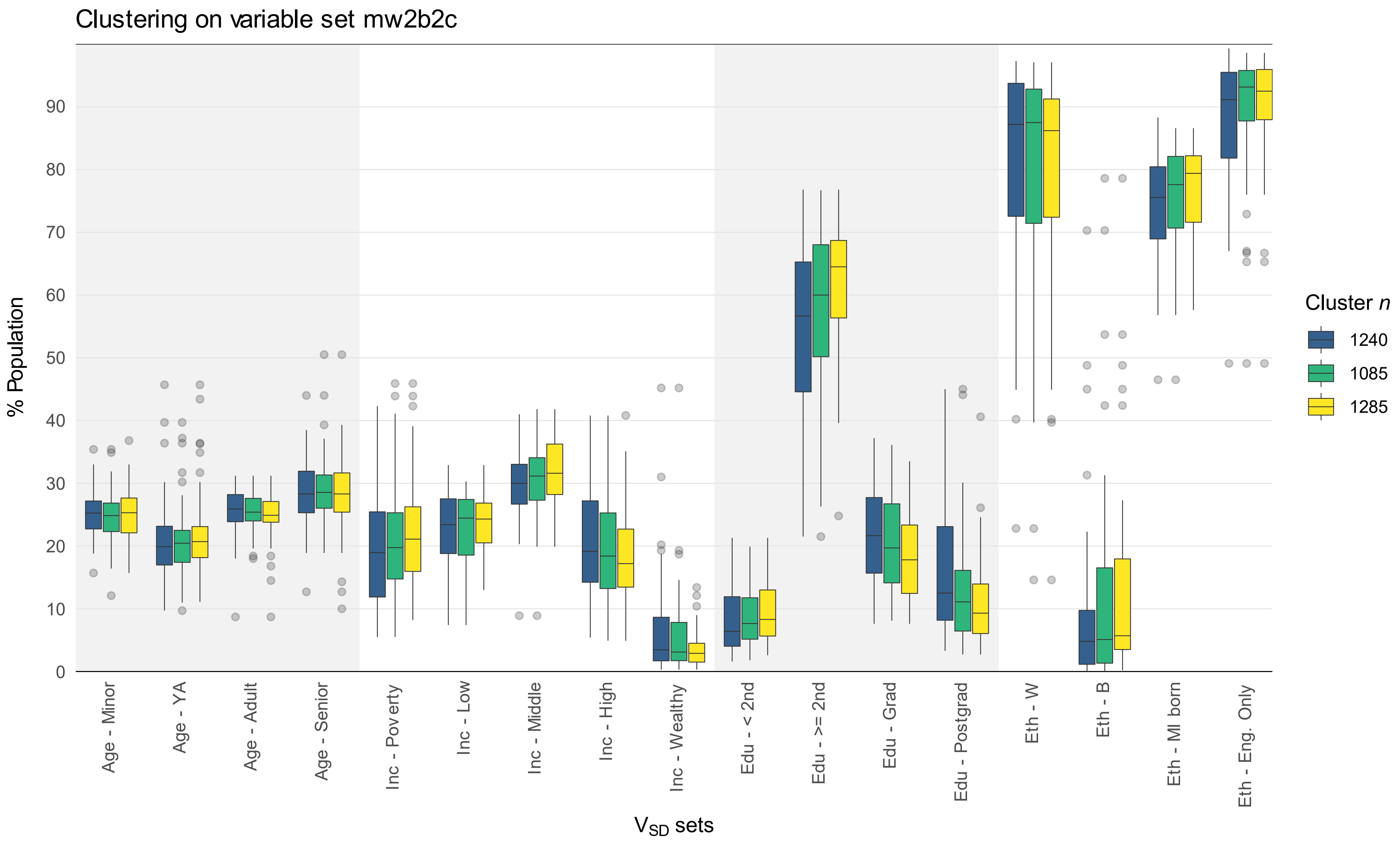


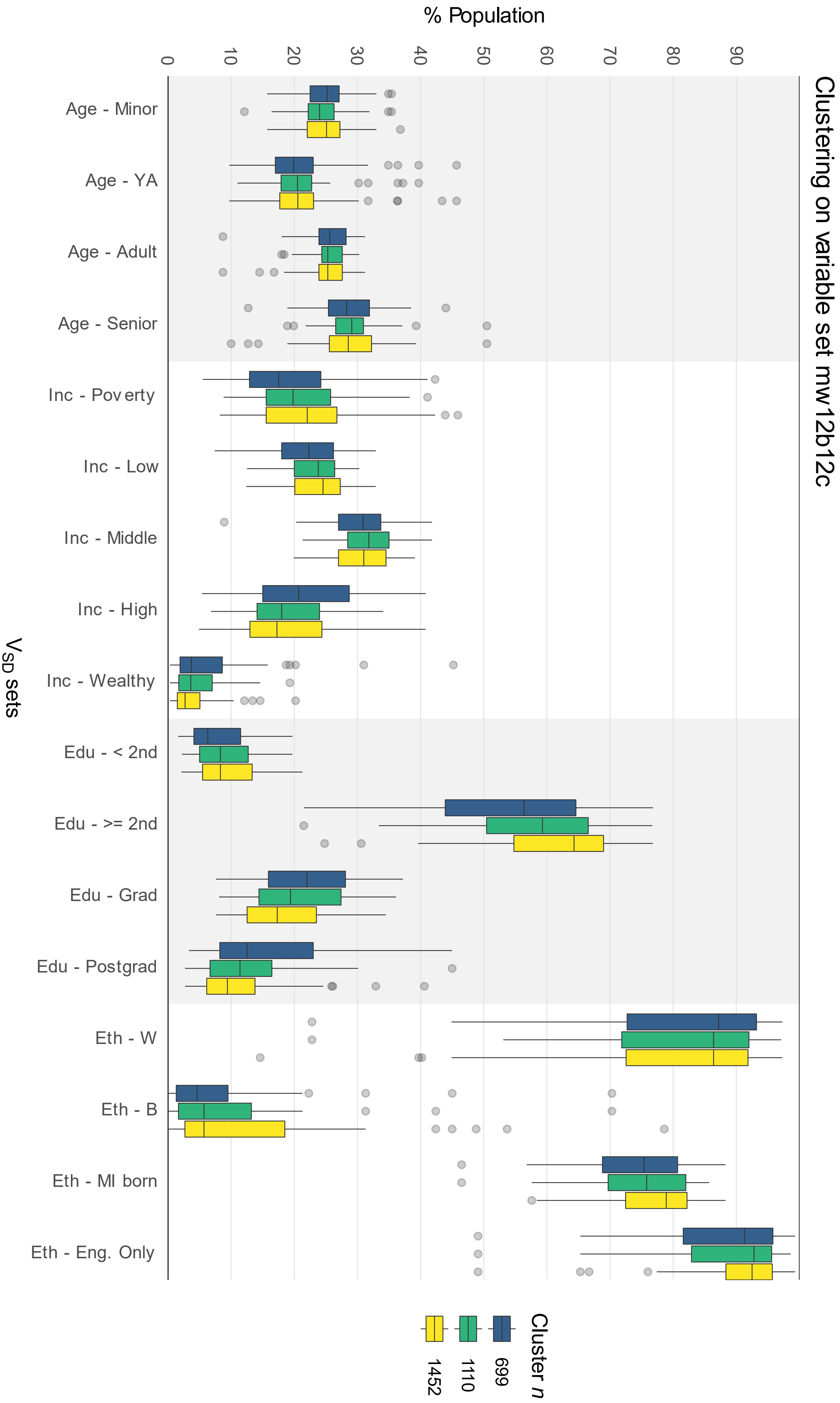
Clustering on variable set w1b1c


Clustering on variable set w2b2c
% Population
0
10
20
30
40
50
60
70
80
90
V_SD sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Pov erty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster n
909
1082
1292


Clustering on variable set w12b12c
Cluster *n*
688
1084
1475
% Population
0
10
20
30
40
50
60
70
80
90
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Pov erty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
V_{SD} sets


Clustering on variable set mw1b1c
% Population
0
10
20
30
40
50
60
70
80
90
VSD sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster n
1811
247
1422


Clustering on variable set mw2b2c
% Population
0
10
20
30
40
50
60
70
80
90
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
V_{SD} sets
Cluster *n*
1240
1085
1285


Clustering on variable set mw12b12c
% Population
0
10
20
30
40
50
60
70
80
90
VSD sets
Age - Minor
Age - YA
Age - Adult
Age - Senior
Inc - Poverty
Inc - Low
Inc - Middle
Inc - High
Inc - Wealthy
Edu - < 2nd
Edu - >= 2nd
Edu - Grad
Edu - Postgrad
Eth - W
Eth - B
Eth - MI born
Eth - Eng. Only
Cluster n
699
1110
1452

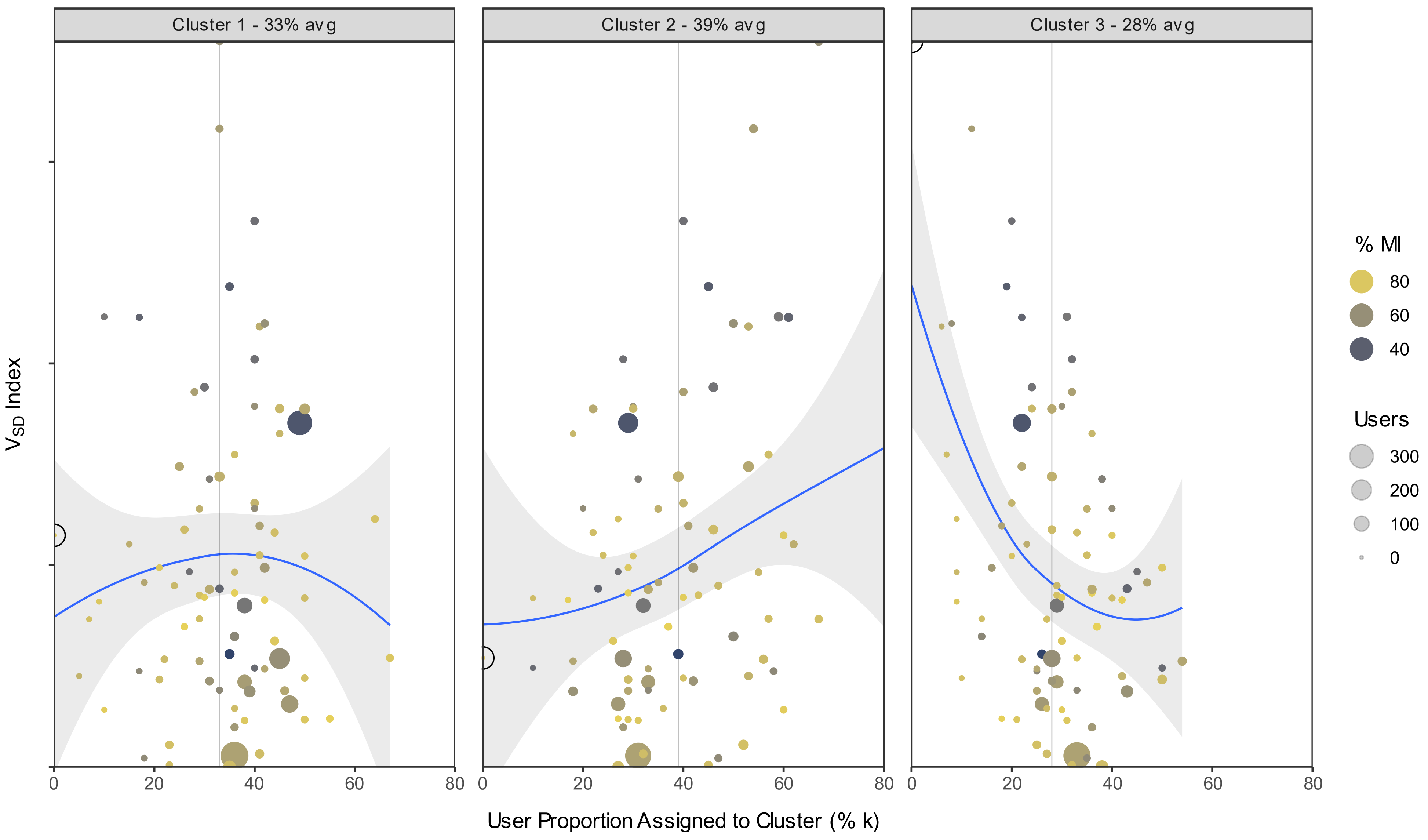
Appendix D. Constellation Charts

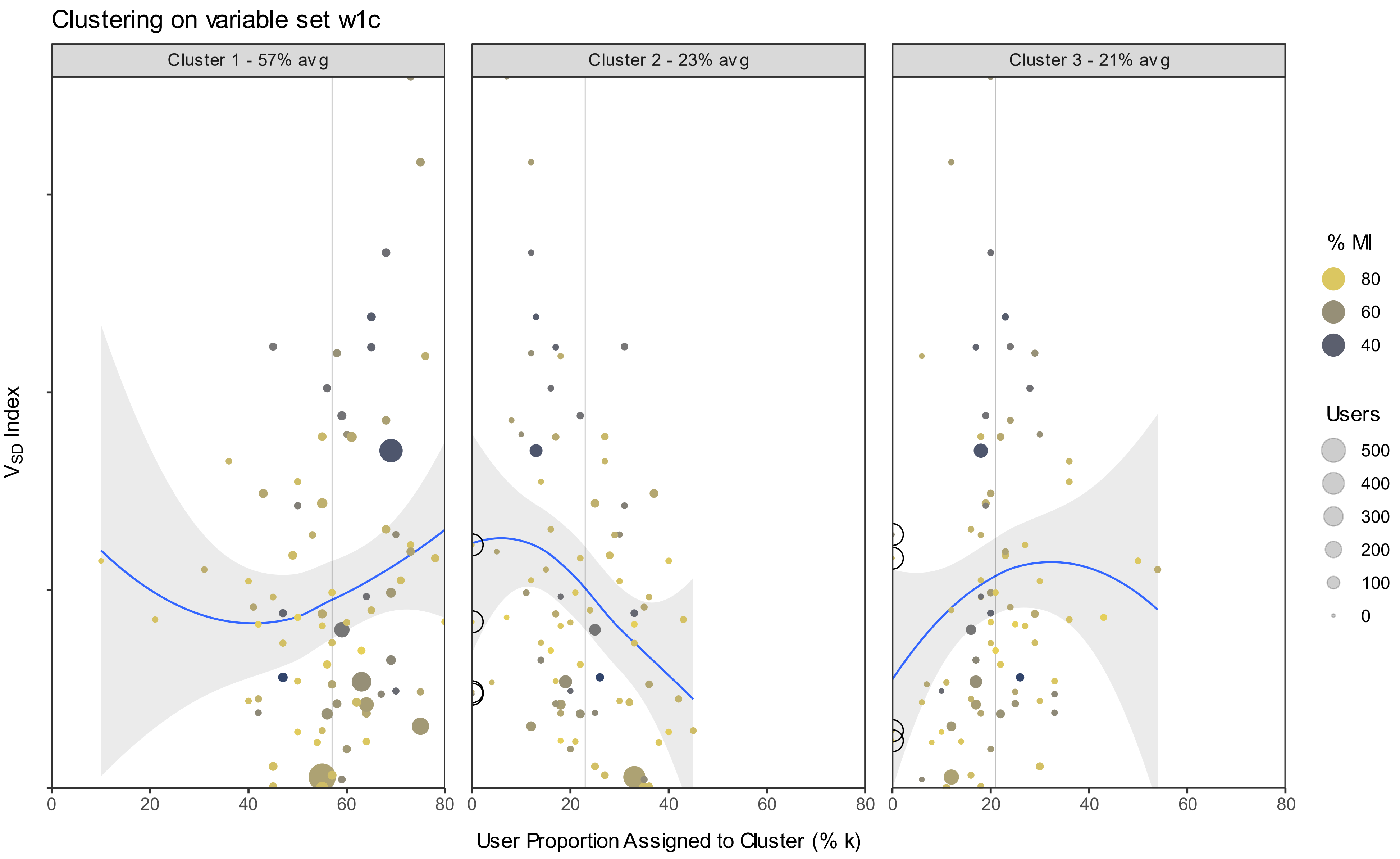
The following are constellation charts plotting the relationship between the sociodemographic index and the relative prevalence of linguistic clusters as presented in step 3 of the analysis (**chp. 6 § 6.2.3**). Presented here are profiles at *k* 3 for all linguistic variable sets. Those set labels and the factor scores they denotate are:

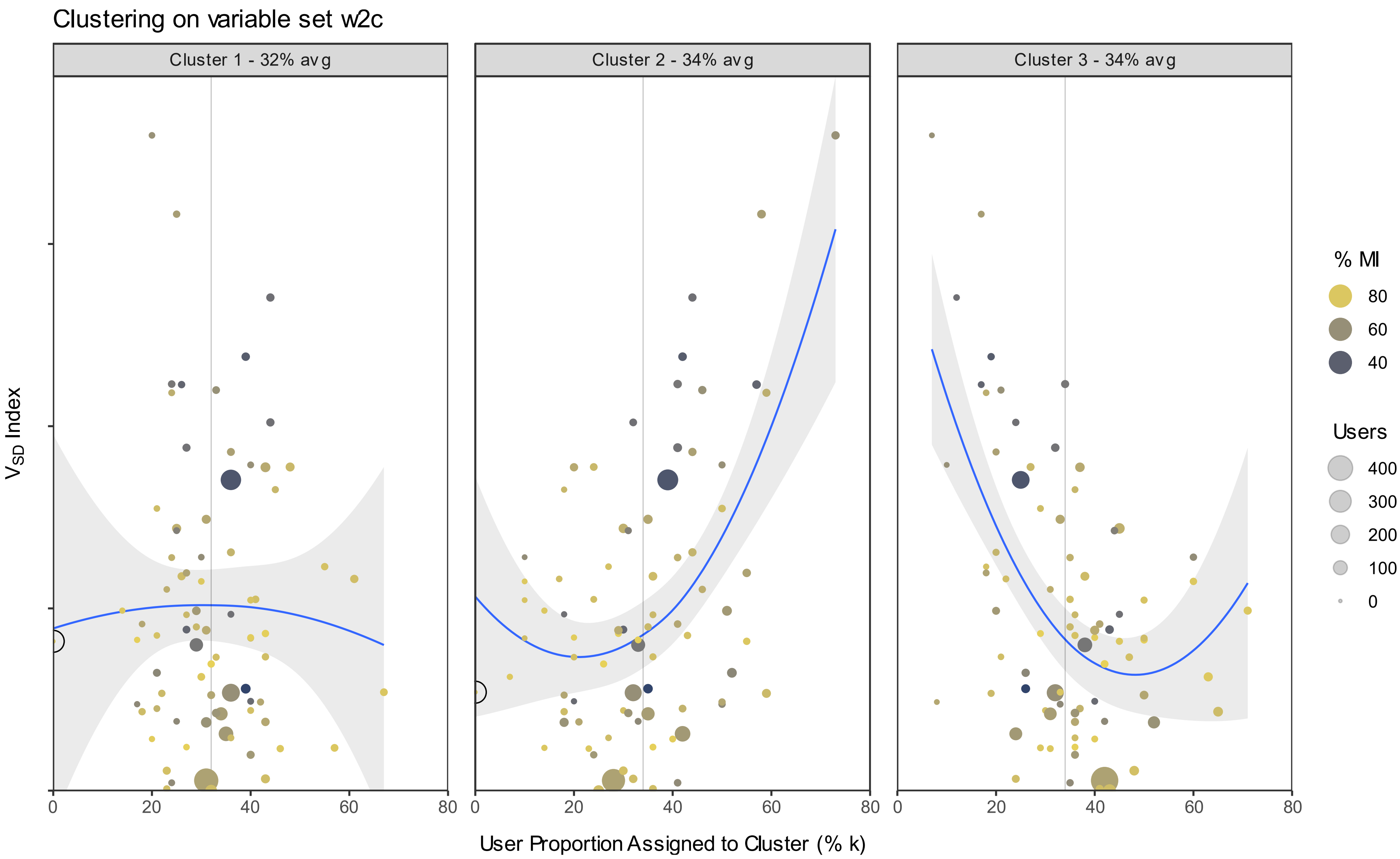
`mc`	grammar (`m` is for MAT ; see **chp. 5 § 5.1.3.2**)
`w1c`	List 1 distinctive pervasive words
`w2c`	List 2 non-distinctive pervasive words
`b1c`	List 1 distinctive pervasive bigrams
`b2c`	List 2 non-distinctive pervasive bigrams
`w12c`	Composite set of List 1 and List 2 words
`b12c`	Composite set of List 1 and List 2 bigrams
`w1b1c`	Composite set of List 1 words and bigrams
`w2b2c`	Composite set of List 2 words and bigrams
`mw1b1c`	Composite set of grammar and List 1 words and bigrams
`mw2b2c`	Composite set of grammar and List 2 words and bigrams
`mw12b12c`	Composite set of grammar and all word and bigram lists

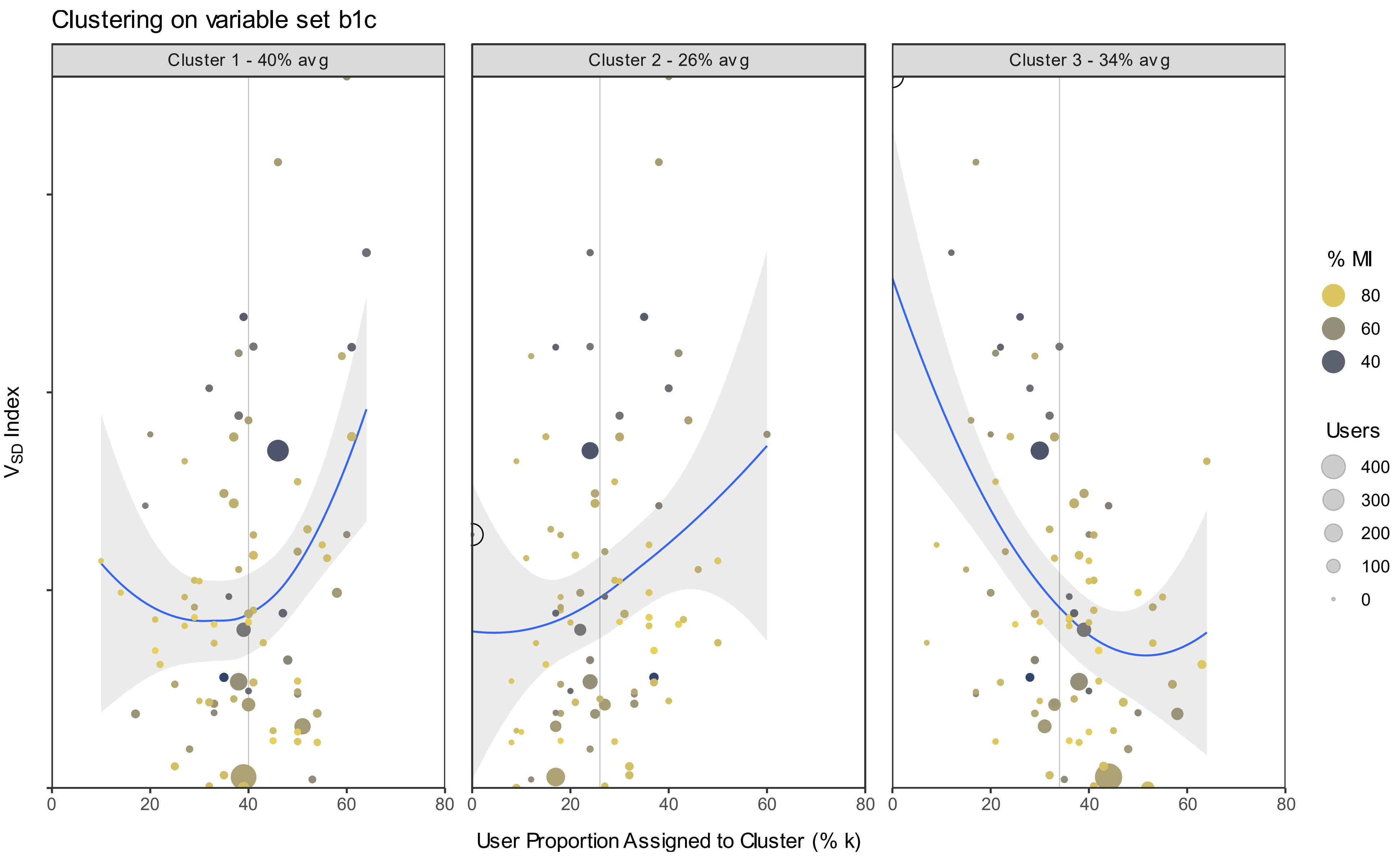
Note that these labels differ slightly for the variable set labels in **chp. 6 § 6.2.2** ; the appended '`c`' on each label indicates that the variable set was produced using the Composite Variable Analysis factoring technique. The profiles are presented in the order of the variable sets as listed above.

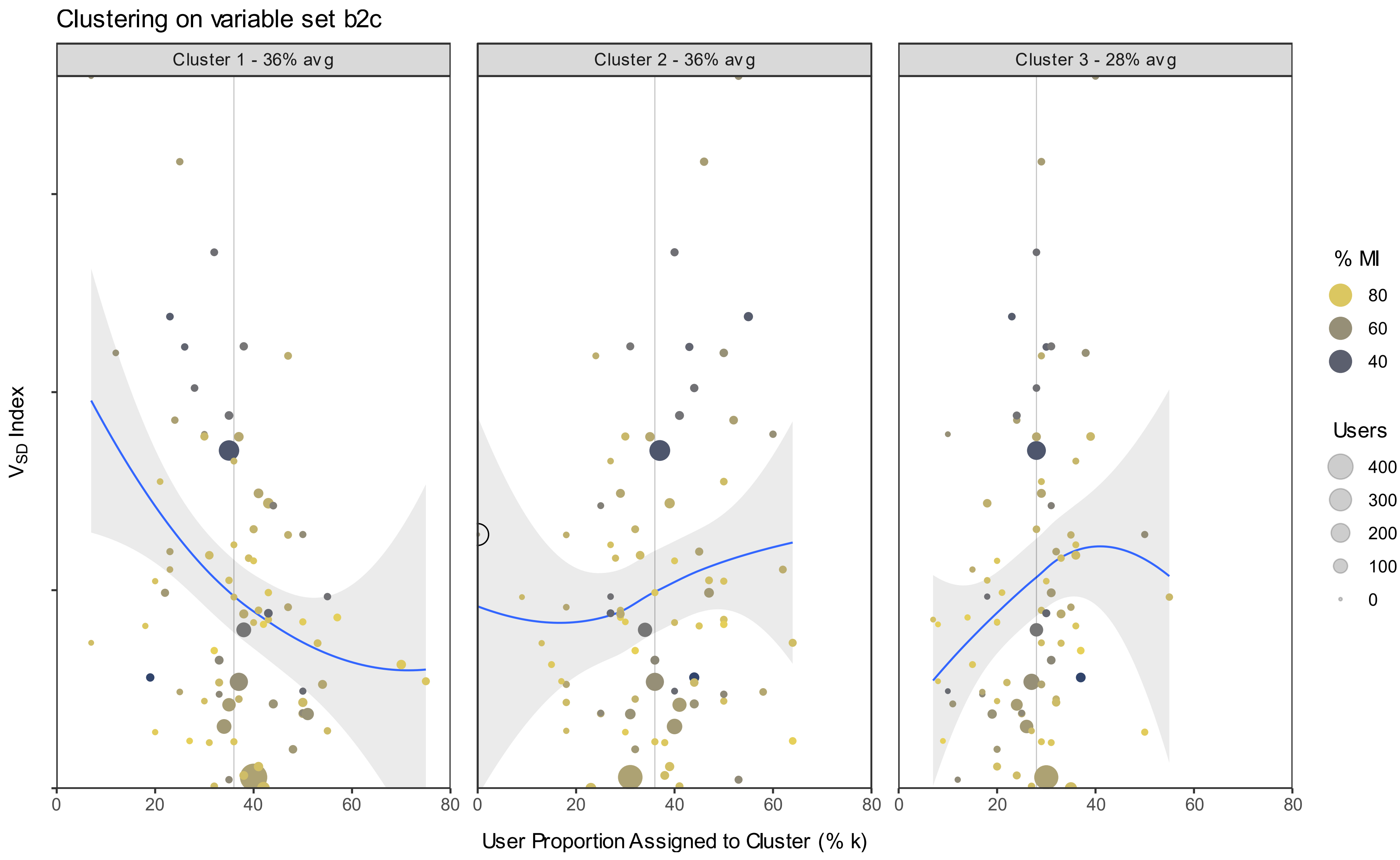
Clustering on variable set mc

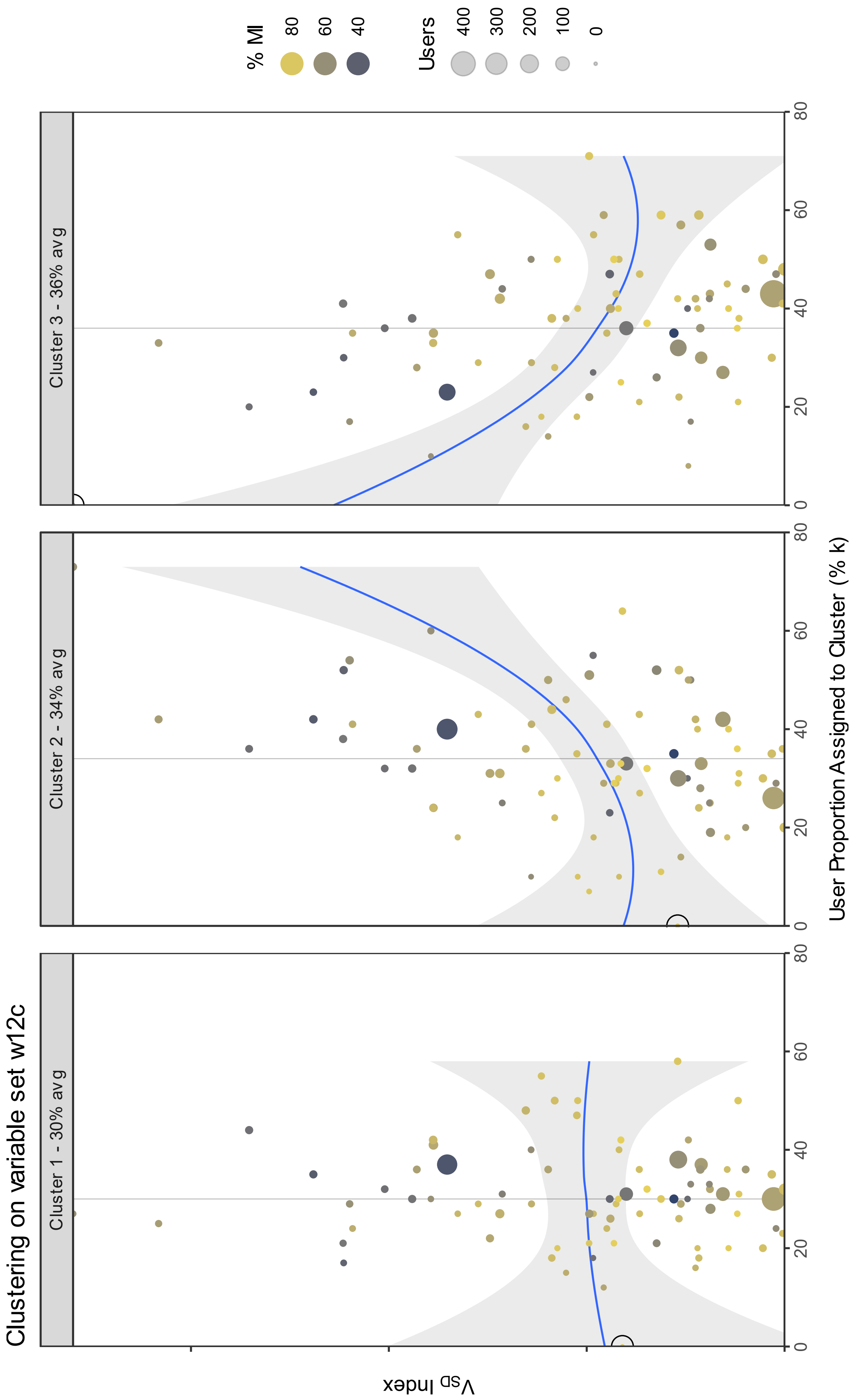


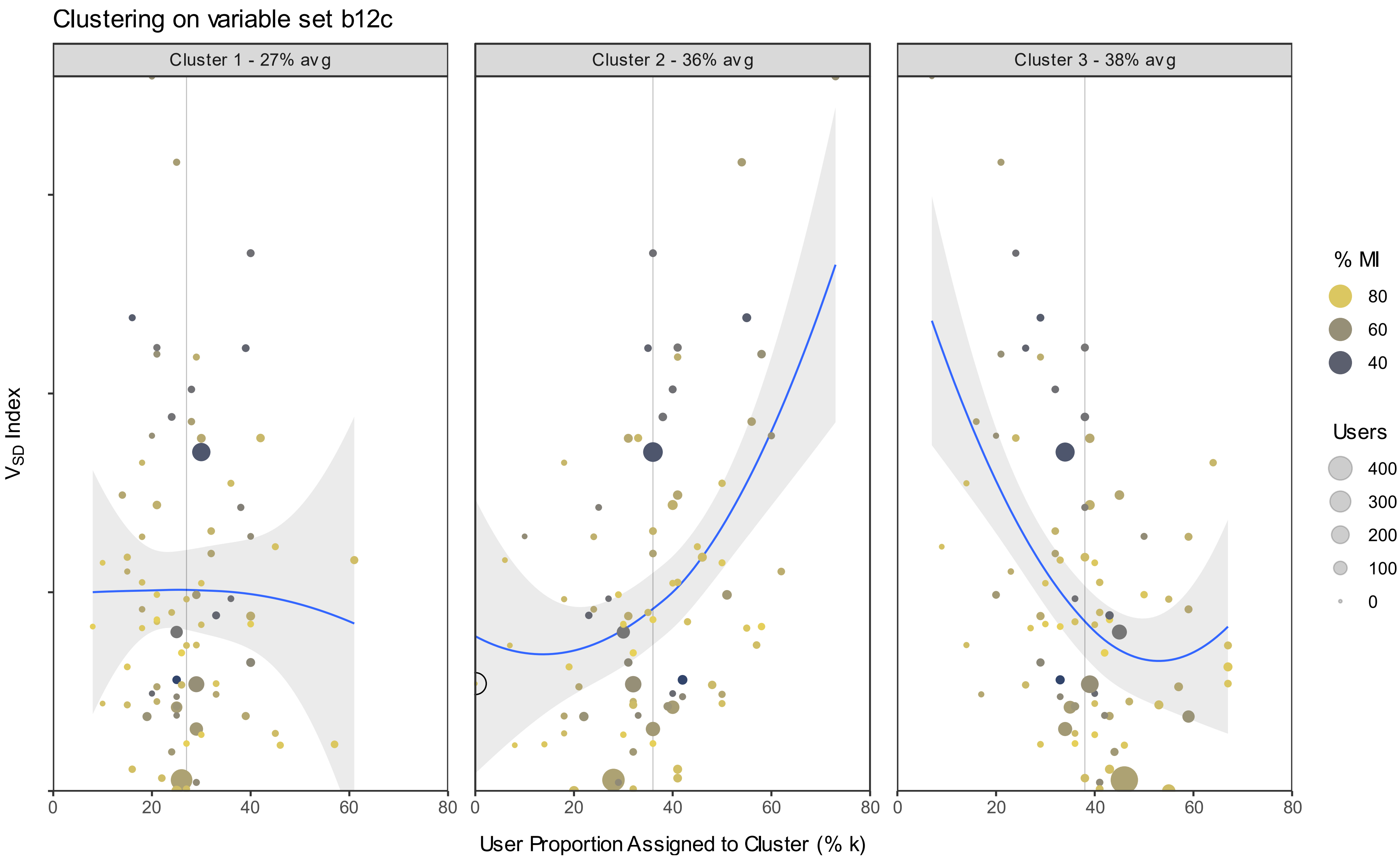

Clustering on variable set w1c
Cluster 1 - 57% avg
Cluster 2 - 23% avg
Cluster 3 - 21% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
500
400
300
200
100
0

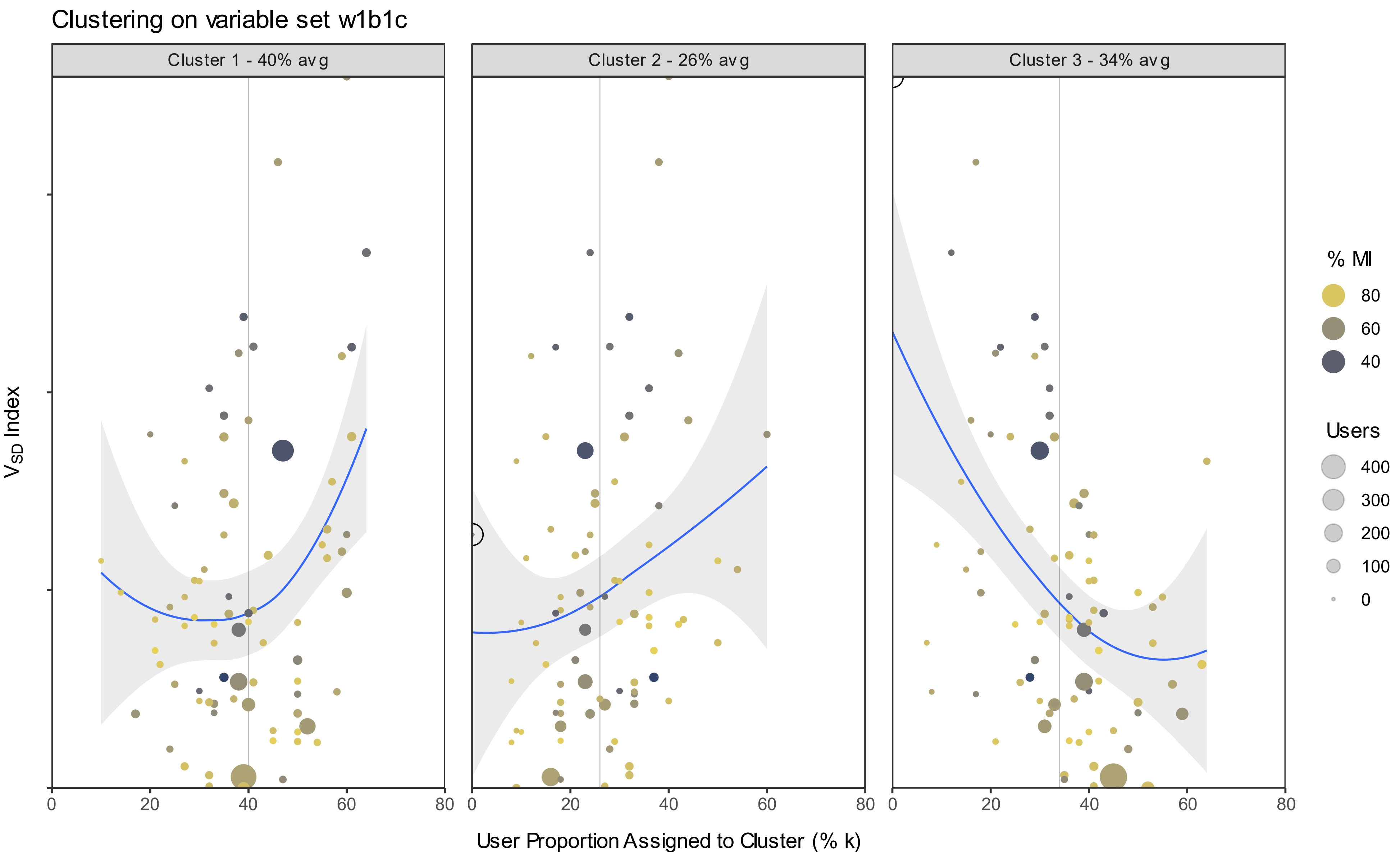

Clustering on variable set w2c
Cluster 1 - 32% avg
Cluster 2 - 34% avg
Cluster 3 - 34% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0

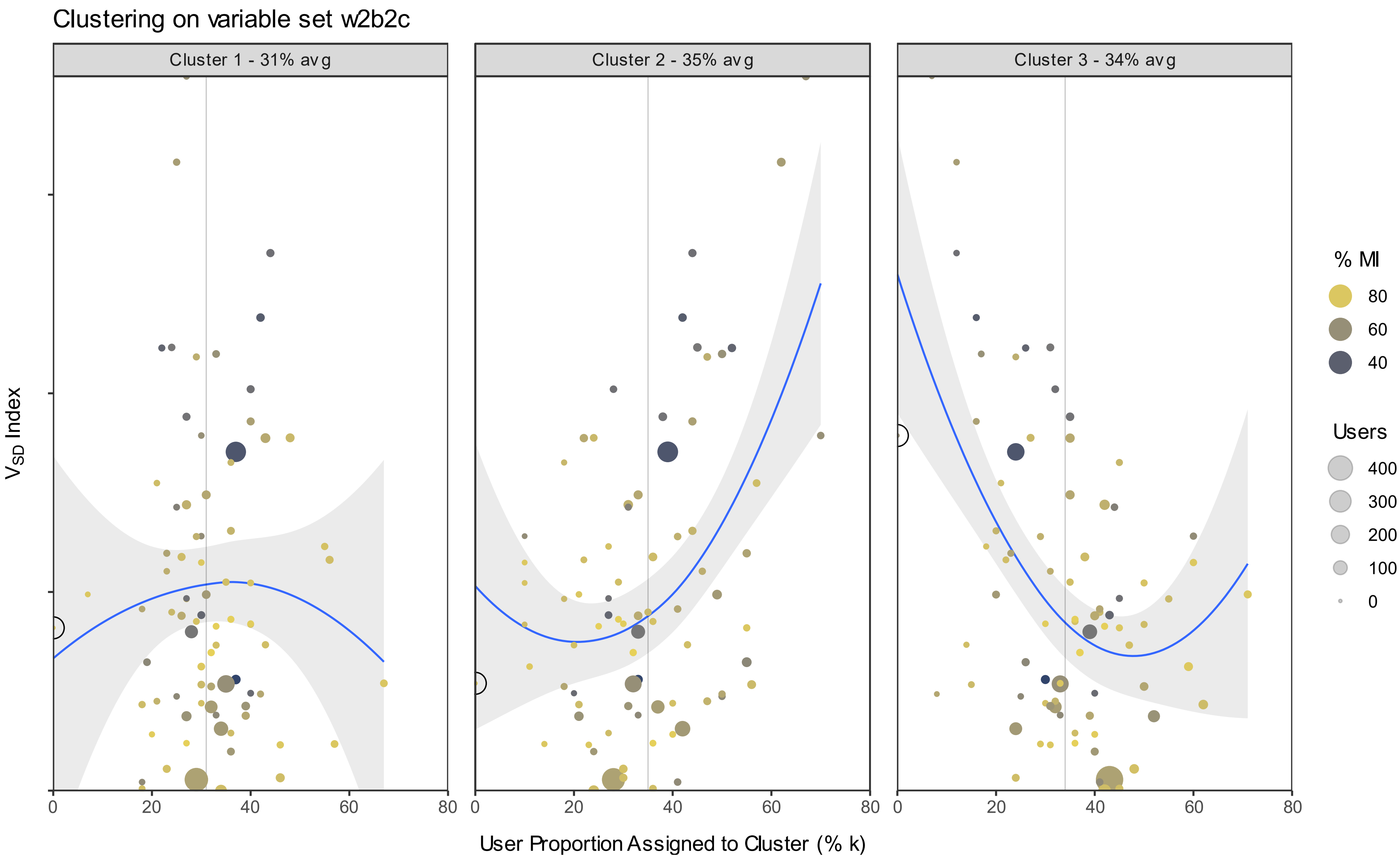

Clustering on variable set b1c
Cluster 1 - 40% avg
Cluster 2 - 26% avg
Cluster 3 - 34% avg
V_{SD} Index
0
20
40
60
80
User Proportion Assigned to Cluster (% k)
% MI
80
60
40
Users
400
300
200
100
0

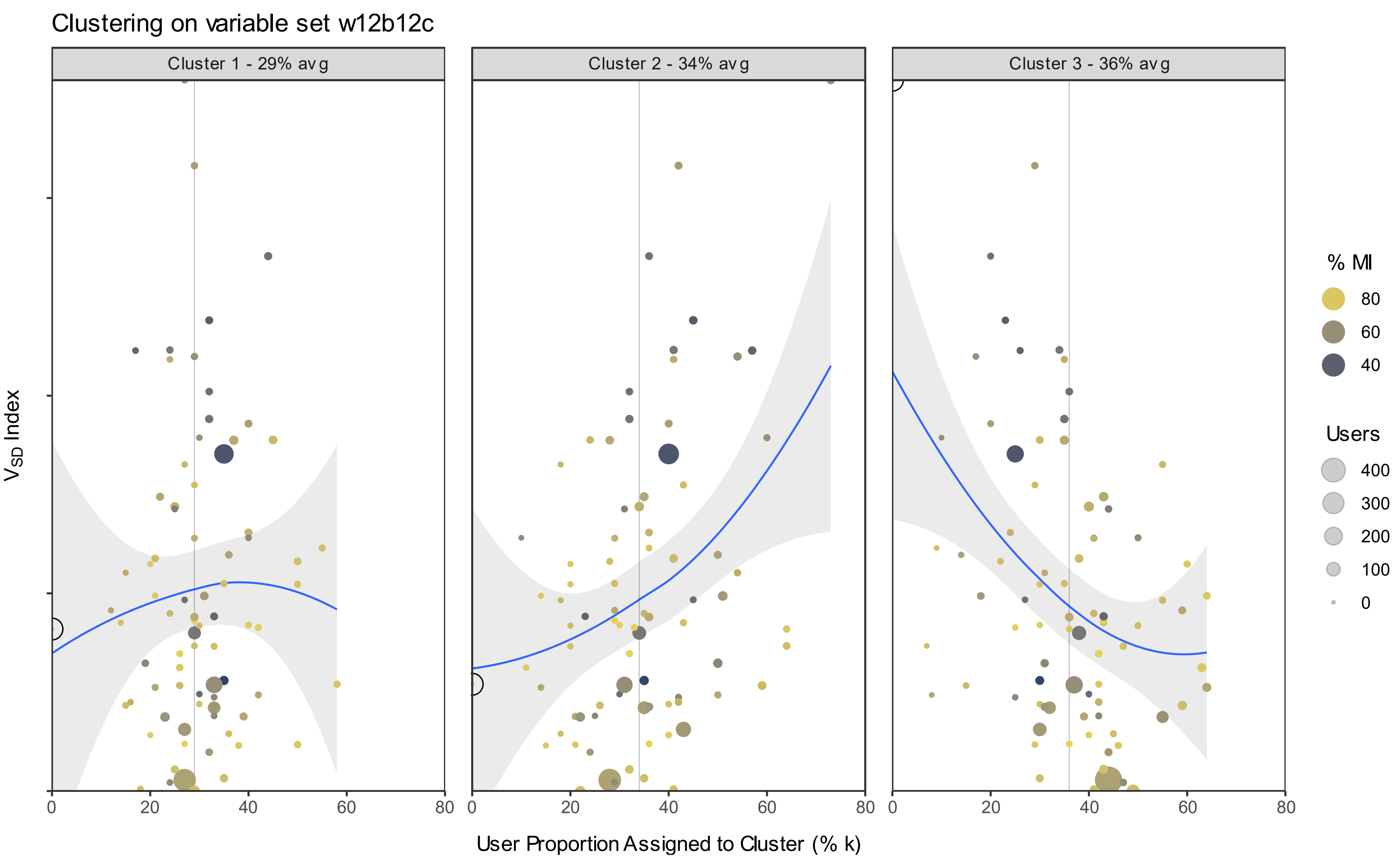

Clustering on variable set b2c
Cluster 1 - 36% avg
Cluster 2 - 36% avg
Cluster 3 - 28% avg
V_{SD} Index
0
20
40
60
80
User Proportion Assigned to Cluster (% k)
% MI
80
60
40
Users
400
300
200
100
0

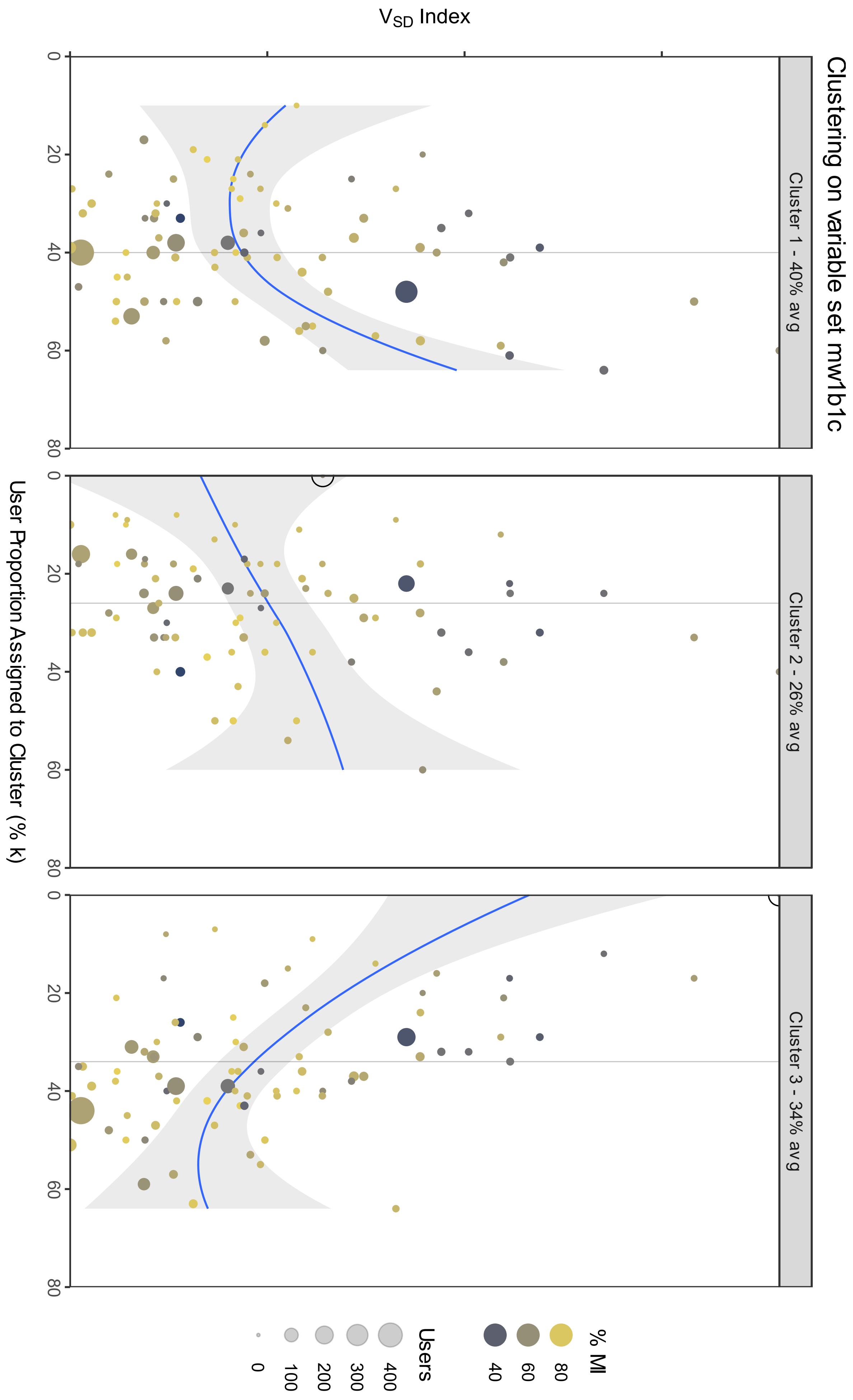

Clustering on variable set w12c
Cluster 1 - 30% avg
Cluster 2 - 34% avg
Cluster 3 - 36% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0

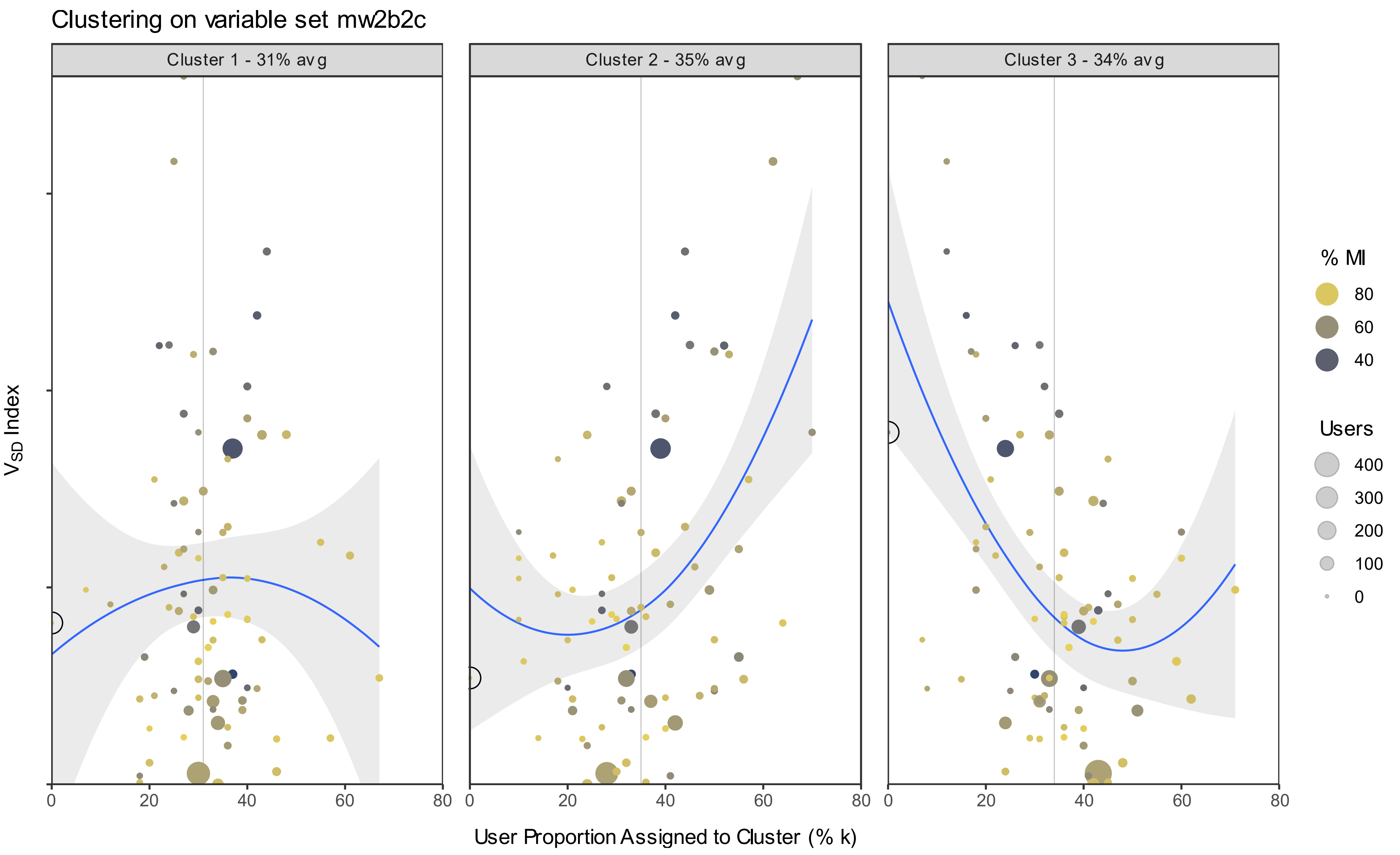

Clustering on variable set b12c
Cluster 1 - 27% avg
Cluster 2 - 36% avg
Cluster 3 - 38% avg
V_SD Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0

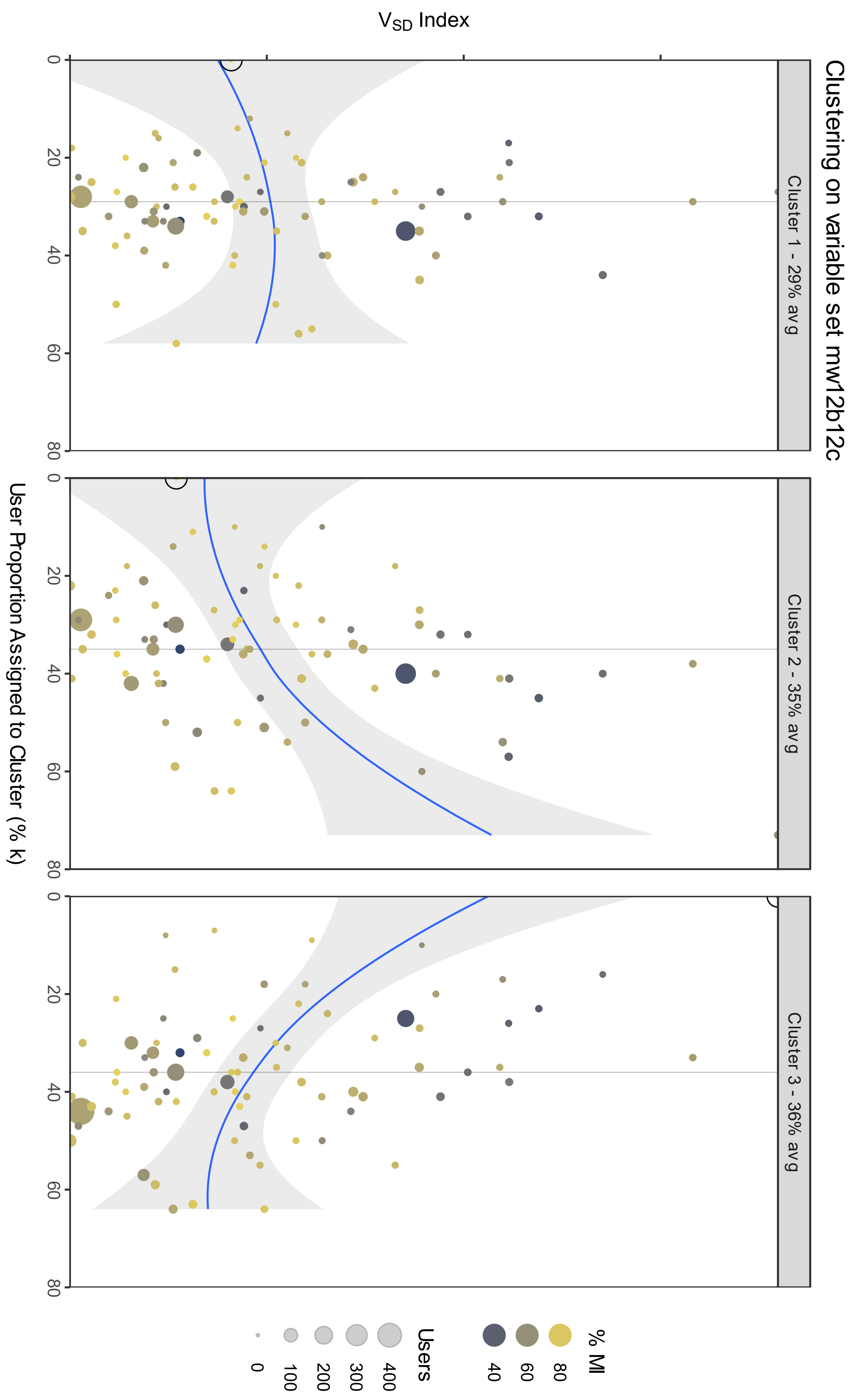

Clustering on variable set w1b1c
Cluster 1 - 40% avg
Cluster 2 - 26% avg
Cluster 3 - 34% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0


Clustering on variable set w2b2c
Cluster 1 - 31% avg
Cluster 2 - 35% avg
Cluster 3 - 34% avg
V_SD Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0


Clustering on variable set w12b12c
Cluster 1 - 29% avg
Cluster 2 - 34% avg
Cluster 3 - 36% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0


Clustering on variable set mw1b1c
Cluster 1 - 40% avg
Cluster 2 - 26% avg
Cluster 3 - 34% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
80
60
40
Users
400
300
200
100
0


Clustering on variable set mw2b2c
Cluster 1 - 31% avg
Cluster 2 - 35% avg
Cluster 3 - 34% avg
V_{SD} Index
0
20
40
60
80
User Proportion Assigned to Cluster (% k)
% MI
80
60
40
Users
400
300
200
100
0


Clustering on variable set mw12b12c
Cluster 1 - 29% avg
Cluster 2 - 35% avg
Cluster 3 - 36% avg
V_{SD} Index
User Proportion Assigned to Cluster (% k)
0
20
40
60
80
% MI
40
60
80
Users
0
100
200
300
400

Appendix E. Endnotes

1. ★ Note that large-scale research is not necessarily macro-level research. Scale refers to scope, and level refers to analysis. Thus one could have large-scale micro-level work, for example household interviews conducted as a subset of the decennial US census, or small-scale macro-level work, such as the work here which investigates linguistic and social structures in a subset of the population of Michigan. Similarly, scale applied to phenomena (e.g. the title of this work refers 'large-scale social phenomena') also refers to scope.

2. Throughout the disciplines and in lay discourse, it is common to encounter the term 'elite' used to indicate members of a polity that are directly involved in the structures of governance. This long-establish usage stems from the root meaning of 'to choose', which the term shares with 'elect'. However, the use of 'elite' to indicate 'exclusive' or 'superior' is also long established. Arguably, the common understanding of the term today is value-laden with a sense of 'better than'. In the light, to speak of a 'political elite' does not imply the existence of 'everyday citizens', so much as it does 'the rabble' (consider the similar case of how the Greek *hoi polloi*, meaning simply 'the many', has taken on a pejorative sense in English). The use of the term in question is reflective of—and possibly contributes to perpetuating—a structuralist orientation in the study of the political, and thus may hinder the conceptualisation of political life outside of establishment structures. For this reason, this work does not use the term in question, using instead 'members of the establishment' or similar.

3. Note that a theoretical focus on political structures (Easton, 1953) and a methodological focus on abstracted 'Models of Man' (Simon, 1957) are faces of the same State-issued coin (cf. Mannheim, 1929, pp. 146–153 ; for a contemporary, disciplinary interpretation, see Lowi, 1992).

4. Throughout this document, the term 'field' is used loosely to indicate a generally bounded area of scholarly work. The study of political participation would likely be termed by US political scientists as a specialty within the subfield of comparative politics (Conge, 1988). In the context of US political science as a discipline (which shapes much of the discussion in this work), 'field' is often synonymous with 'discipline', as implied by a persistent concern with the proper 'subfields' of political science (Almond, 1990 ; Kaufman-Osborn, 2006). The nature and implications of these debates (Reiter, 2015 ; Reid and Curry, 2019 ; Graham, Shipan and Volden, 2014 ; Almond, 1990 ; Goodin and Klingemann, 1998) are certainly germane to this discussion, but far beyond its scope. It should be noted that such debates appear to be characteristic of the political as subject matter, rather than of the US academy (Norris, 1997 ; Rosamond, 2007 ; Easton, Gunnel and Graziano, 1991 ; e.g. Boncourt, 2008)

5. For the purposes of the present discussion, it is not necessary to be familiar with the nature and details of the behavioural movement ; for readers not familiar with the movement, it is sufficient to bear in mind that it was driven by a focus on behaviour (whence the name) ; the privileging of statistical and quantitative methods ; and an insistence on 'pure' (as opposed to applied) science (Farr, 1995), underlain by a Statist, systemic perspective on the political (Gunnell, 2013) ; see also the following note. For a historically contextualised overview, see John Dryzek (2006) ; for a somewhat hagiographic overview by a central figure in the movement, see Gabriel Almond (1998, pp. 68–75).

6. David Easton, one of the central figures of the behavioural movement (Gunnell, 2013), articulated the premises and goals of behaviouralism—the so-called "credo" (Easton, 1965, p. 7)—including "1. Regularities" that are discoverable, universal, and generalisable ; "2. Verification" and testability of such regularities ; "3. Techniques", i.e. the assumption that methods must be rigorous and validated ; "4. Quantification", provided by techniques and which is essential to verification of regularities ; "5. Values", i.e. values are to be kept distinct and apart from empirical analysis ; "6. Systematization", i.e. theory and empirical research must work hand-in-hand ; "7. Pure Science", i.e. the focus is on theoretical understanding, not practical application ; and "8. Integration", i.e. that political science ignores findings from other disciplines 'at its own peril'.

7. For a broad contemporary perspective on the tensions arising from these pressures, see Kristen Renwick Monroe (2005). Historical perspectives are perhaps more enlightening in regard to understanding <u>how</u> such pressures both arise and persist. Among these pressures are those common to academia, such as the demands of research funding and publication (Pooley, 2016), the reconciliation of differing worldviews and sociopolitical philosophies (Gunnell, 2004, pp. 48–49), and the role of interpersonal rivalries (Karl, 1974, pp. 155–168 ; Almond, 2004). However, there are also less common pressures arising from the relation of political science to the State (Lowi, 1992), such as the direct influence of the government, the military, and major foundations (Berndtson, 1987 ; Ahmad, 1991 ; Hauptmann, 2012, 2006 ; Seidelman, 1985) ; and the hunger of the bureaucratic State for data and 'evidence-based' decisions (Lee, 1995 provides an excellent and thorough overview ; also Smith, 1997, pp. 255–258).

8. This statement should not be understood to suggest that 'the discipline' is an entity capable of judgement, decision, or action. Such an essential, reified sense is not in keeping with the philosophical perspective of this work. The statement is presented in this manner simply to facilitate the discussion. A more apt presentation of this statement would need to engage in a political economic discussion of disciplines and the role of their establishment (in terms of individuals and organisations) in shaping their development (Barrow, 2011 ; Waismel-Manor and Lowi, 2011 ; Monroe, 2007 ; cf. Sigelman, 2006). This is perhaps more the case for political science, given its symbiotic relationship with the State (Lowi, 1992 ; see also the ensuing discussion in Simon, 1993 and Lowi, 1993). In this light, the statement should be taken to suggest that the drive for scientific disciplinarity was a theme present over time in the literature, interpreted here as reflecting a perspective present among the establishment (Kaufman-Osborn, 2006, esp. fn. 4 ; Eulau, 1997, p. 583). In the first instance, that drive emerged from a concern that political science was not 'science' enough compared to other disciplines (Farr, 1988, pp. 1177–1178)—a concern that has marked political science from its early days (e.g. Merriam, 1922, pp. 315–319). Once that was concern was allayed, however, the drive was perpetuated as the establishment grappled with dissenting voices that sought not only methodological pluralism, but also a place at the high table. This contention in the late 1990s and early 2000s is referred to as the 'Perestroika Movement', which strove for methodological pluralism and an end to the entrenched control of a restricted set (in terms of demographics, *almae matres*, and methodology) of conventional scholars in the disciplinary institutions (namely the American Political Science Association and its journal the American Political Science Review) (Monroe, 2005 ; Barrow, 2017). While that movement made some notable gains, it eventually failed to achieve its overall aims. Aside from institutional momentum, the movement faced an issue of poor historical timing—its first public action arrived on 15 October 2000 (Monroe, 2005, pp. 9–11). Within 11 months, US political science would find itself in yet another flurry of structural activity spurred by war. Harold Lasswell (1968) observed a similar pattern with the Cold War, the First and Second World Wars, and the Spanish-American War. When your business is the State, war is good for business (cf. **n. 99**).

9. While a historiographic presentation of this argument is instructive, the development and persistence of both discipline and approach are interpreted throughout this work from a perspective of boundary work, that is, "… the discursive attribution of selected qualities to scientists, scientific methods, and scientific claims for the purpose of drawing a rhetorical boundary between science and some less authoritative residual non-science" (Gieryn, 1999, pp. 4–5, 1983).

10. Note how the concept of translation through 'boundary objects' (Star and Griesemer, 1989) is conceptually tied to the boundary work (Gieryn, 1983) of rhetorical delimitation of science and the disciplines (see also **n. 9**). Note further that boundary objects, in serving as shared markers of distinction, thus simultaneously function as shared spaces of possible understanding and action (Star, 2010, pp. 602–603).

11. The presumed objection being to contextualisation, for those contexts in which abstraction is the clarion call, regardless of purpose. For an illuminating discussion that moves beyond discipline, see Alberto Toscano (2008).

12. For a contemporary and much less conventional, yet nevertheless structural and systemic perspective, see Vivien Schmidt (2017, 2010).

13. The desire to study text in this manner, and theoretical and statistical understandings of how to do so, is much older (e.g. Stone *et al.*, 1962). However, the growth in text-as-data coincides with the social media era and a staggering increase in the availability of digitised text, alongside the increasing availability and decreasing costs of computing capacity and resources.

14. Text-as-text is no less empirical than text-as-data ; between the two approaches there is simply a different relationship with text as a tool for research. For an example of this difference in relationship that helps to contextualise the development of the conventional approach in political science, see David Sylvan (1991, esp. 279–283).

15. In certain domains, such as information technology, 'at scale' is often synonymous with 'large-scale'. However, a more appropriate understanding, especially in the domain of social research, is 'at the appropriate scale' (i.e., appropriate to the scale of the phenomena in question).

16. Arguably, this assumption is an essential description of medium theory ; see the discussion of Harold Innis and Marshall McLuhan in **chp. 2 § 2.1.2.1**.

17. The reconceptualisation of the subject model is an aspect of the overall reconceptualisation of political participation proposed in this work. The concept of 'subject model' itself describes how researchers conceive of the subject of their research. In the context of the social sciences, how is it that we conceive of our species and of individuals, and of society itself? In the words of Herbert Simon, a researcher that could be firmly placed within the conventional approach in political science, "Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying" (1985, p. 303). The subject model of the conventional approach is an atomistic, economic, and rationalist conception of the individual (Simon, 1957 ; Lindenberg, 1990). The subject model proposed here recognises individuals from an embodied perspective, but conceives of them not in isolation, but rather as existing intersubjectively through matrices of relations and meanings (Emirbayer, 1997 ; Emirbayer and Mische, 1998 ; Grossberg, 1982).

18. Note that these ontological concepts are developed in this work specifically for the purpose of theorising social phenomena in hybrid society. **Chapter 2** will explain.

19. ★ For a helpful (and succinct) introduction to corpus-based approaches to language, see Tony McEnery, Richard Xiao, and Yukio Tono (2006, unit 1). It is freely available online at https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm.

20. ★ It is a valid observation that corpus-based approaches are thus 'text-as-data' approaches as described in **§ 1.5**. This much is true. However, note that the labels are generally indicative of different disciplinary and philosophical contexts working with disparate subject models. In that vein, consider the prior assumptions required to prompt the label 'text-as-data'.

21. The study of social variation in language is the stock in trade for some fields, and it has a long history. For an overview of more than 50 years of sociolinguistics, see Penelope Eckert (2012).

22. ★ In broad strokes, this is done by passing each document to an application that parses the text and assigns each word a 'tag' according to part of speech. Such an application is called a 'part of speech tagger'. Biber's original method included an extensive set of decision rules used to identify more complicated grammatical structures composed of linked parts of speech. Once the tagging is done and the decision rules are applied, the total counts of each linguistic feature (a total of 67) are easily tallied up.

23. Note that all 417 locations represented in the corpus were separately clustered on sociodemographic data ; those cluster assignments were also linked to the user–documents according to associated location.

24. This touches on a deeper point, to which we will return in **chp. 4**.

25. The terms 'constructionism' and 'constructivism' are often used interchangeably. However, as terms, constructionism is more strongly associated with sociology, whereas constructivism is more strongly associated with epistemology. This work uses the term 'constructionism'.

26. Throughout this work, the term 'post-material' indicates emphasis on the material itself, as opposed to 'post-materialist', which suggests emphasis on the theory of the material.

27. Articulation as used here indicates context and associated behaviours emerging from contingent, complex, and relational entanglement. Note that this concept is fundamentally similar to actor–networks, discussed in **§ 2.1.2.3**.

28. Note that attending to emergent social structures is not indicative of a structural perspective ; rather, attention to emergence relies upon a relational perspective. As noted in **chp. 1 § 1.5**, a perspective on socially communicative phenomena—that is, a *relational* perspective—would expect to see structure emerge from the dynamics of sustained, intersubjective relations.

29. Consider this as a lower-level application of McLuhan's concept of the 'extensions of man'. The original concept implicitly suggests embodiment ; the question here challenges that assumption.

30. This distinction is not only indeterminate, but moreover it is ill-conceived—'intentional agency' (i.e. human agency) is a redundancy and agency devoid of intention (i.e. material agency) is a convoluted term for mere occurrence. Thus there is little surprise that agency has been a contested concept in the social sciences (Emirbayer and Mische, 1998). Note that contestations over agency—that is, the so-called 'agency–structure debate'—are essentially ontological disagreements about what it is that shapes the social world.

31. Consider this in terms of the 'social person' subject model introduced in **chp. 1 § 1.6**.

32. Function is understood here in an absolute sense, not a situated sense. In the domain of the technical, function is simply an essential quality, like mass or energy. Intention belongs to the social, and consequence to the material, thus in the technical function and being are equivalent, and in the technical being and physicality are equivalent.

33. In a Kantian framing, the technical could be understood as the phenomenal, and the social as the noumenal. Such a dualistic framing, however, holds the two in isolation. The material could thus be understood as bridging that isolation, in comprising the relations between phenomenal and noumenal. It is thus neither thought nor thing, but simply the accumulated happenstance of being—thus it is characterised above as the domain of mediation, which seems more straightforward.

34. In its general sense, 'sociotechnological' denotes the complex of society and technology. In the sense of the ontological footing developed in the previous chapter, the term can be understood to denote a given configuration of the social, material, and technical.

35. Given that this work draws deeply on linguistics, it is worth noting that discipline underwent a similar shift during this same general time period stemming from the work of Noam Chomsky (e.g. 1957). The 'Chomskyan revolution', however, proceeded in a manner at epistemological odds with the behavioural revolution—while both movements sought "systematisation and formalisation" in their disciplines, the latter was firmly empiricist while the formal emerged as firmly rationalist (i.e. anti-empirical). For linguistics, it was perhaps something of a bait and switch: Chomsky's early work stressed the systematic and formal aspects, ideas which were welcomed by the generally empiricist linguists of the time. But as Chomsky's work gained support and influence, he grew insistent on the necessity of a rationalist approach, which set the course of the discipline for decades to come (e.g. Harris, 1994a, 1994b). Further comparison of the behavioural and Chomskyan revolutions are beyond the scope and purpose of this discussion, but do note that both movements embraced an abstract subject model ; in Chomskyan linguistics, abstraction was an essential component of the work (cf. McEnery and Wilson, 2001, chp. 1).

36. One reviewer stated, "This book summarizes the results of one of the few pieces of genuine political research ever undertaken in this country. For many years, members of that group of scholars who choose to style themselves political scientists have been prating of research, but for them research has consisted in the main of fishing in a boundless sea of books and documents for citations and quotations to support preconceived theories. By their footnotes you shall know them" (Maxey, 1925, p. 369).

37. The General Social Survey is a national survey of US adults that is intended "to monitor and explain trends in opinions, attitudes and behaviors". It has been conducted regularly by the National Opinion Research Center at the University of Chicago since 1972 (NORC, no date).

38. Also called the 'socio-economic status' model.

39. To that end, this work operationalises language to develop endogenous measures for social classification, as will be addressed in **chp. 4**.

40. In the political participation literature, especially in the US academy, 'citizen' is generally used without definition or specification, its evident function being to denote members of a polity, or just 'people'. The lack of specification makes its use pointed, however, in that 'citizen' frequently takes an attributive adjective. Looking just at the foundational Verba and Nie (1972), for example one can find "individual" (p. 5), "ordinary" (p. 29), "average" (p. 89), "passive" (p. 97), "lower-status" and "upper-status" (p. 203), "black" and "white" (p. 206, fn. 9), and "active" and "inactive" (p. 317). Implicitly the semiotic domain of 'citizen' is one of distinction, hierarchy, and evaluation, yet nevertheless the concept remains unspecified (such concerns being left to political theorists, mainly ; see below). In the field there have been some empirical studies of notions of citizenship in a populace (cf. Theiss-Morse, 1993)—notably by Gabriel Almond and Sidney Verba (1963)—but the author considers such studies highly problematic, for reasons beyond the scope and purpose of this work. More on the role of the concept of 'citizen' in **chp. 4**.

In regard to theories of citizenship, for example, see Herman van Gunsteren (1998) in the (neo)republican vein, and Henry Tam (2019) in the communitarian vein. For a critical consideration of ‘deliberative’ citizenship, see Peter Dahlgren (2006). For an older, though nonetheless timely and trenchant critique of thinking on citizenship generally, see George Armstrong Kelly (1979).

41. Discussion of these debates is beyond the scope and purposes of this work. For detailed overviews of these debates see Ohme (2018), van Deth (2016), Fox (2014), Hooghe, Hosch-Dayican and van Deth (2014), and Gibson and Cantijoch (2013).

42. The notion of social–technical–material configuration is conceptually comparable to the established notion of sociomaterial configuration (see **chp. 2 § 2.1**).

43. For ‘phenomenal’ one might instead say ‘empirical’. However, the social sciences should in the main resist the urge towards ‘empiricism’, wherein the phenomenal/empirical is privileged above the experiential—which here is understood to denote our mediated interpretations of the phenomenal. Empiricism is rightly seen as the gold standard in the natural sciences. But the social sciences must account for the nature of their own subject, not ignore or obscure it. Empiricism is not always suited for making that account (compare the discussion of behaviouralism in **chp. 3 § 3.1.1**).

44. Recall that this work takes an explicit approach to presenting its structure because the work is intentionally, and necessarily, interdisciplinary. In that light, there is no single disciplinary matrix underpinning it to inform researcher and reader alike (Kuhn, 1970, pp. 181–187). For this reason the assumption of tacit knowledge is a pitfall in interdisciplinary work, and thus such work must be explicit at each turn. This is not pedantry, but rather clarity and courtesy.

45. There is a saying often attributed to McLuhan that goes something like “We don’t know who discovered water, but it certainly wasn’t a fish”. McLuhan did make statements of this sort, generally in speeches or interviews (2003, p. 106). The purpose of such jocular, aphoristic expressions was to anchor a deeper point: in the mediatised world, we often become so fixated on messages that we lose sight of the media that bear them (2003, p. 150). The semiotic and material modulations of media fade into an ‘environment’ to which we no longer properly attend—familiarity does not breed contempt, so much as inattention. For a helpful philosophical discussion along these lines, and which also engages with McLuhan’s thinking, see Timo P. Kylmälä (2012).

46. ★ Taken together, the conceptual framework and method comprise the methodology of this work. Whereas in disciplinary work a significant portion of methodology might well be subsumed in domain and tacit knowledge, interdisciplinary work cannot rely on a common matrix and thus these elements are made explicit (see **n. 43**). It is also worth noting that the ‘theory’ presented here, to again use the term in a general sense, is putative, and rightly so. This work is essentially exploratory, charting an explicitly interdisciplinary tack through complex social phenomena, with the relatively disciplinary study of communication as a point of reference to guide our way. And while the specific theory presented here might be putative, the theories that inform it are not. The goal of this work is to seek warrant for this specific theory by way of empirical test. If such warrant can be demonstrated, then let the theory be developed further ; if not, then the theory can be rethought, reworked, or discarded. In either case, the primary goal is to contribute to the development of interdisciplinary social inquiry, and thus to expand the overall body of knowledge.

47. We can only compare the character of any mediation, not the degree. While the mediating effect of any mediant or mediator can be quantified in information-theoretic terms (e.g., Shannon and Weaver, 1949), it cannot be measured in any meaningful way (i.e., its impact of shared meaning). Thus, until we can measure such effects on meaning, we are obliged to consider all communication (e.g. face-to-face versus remote) as thoroughly, and equally, mediated. Our characterisations of it, for example as 'deep', are perhaps more indicative of our awareness of the water than its depth (cf. **n. 44**). The world is mediation, all the way down.

48. Thus the hypothetical objection would reveal a fundamental difference in subject model compared to that used in this work.

49. Consider also in Jeffrey Treem *et al.* how the variety of concerns about digital inequalities reflect normative assumptions of how and why social media is or should be used. Compare such assumptions with prescriptive versus descriptive approaches to conceptualising political participation as discussed in **chp. 3 § 3.2**.

50. Consider this point in light of the discussion in the following section of the outmoded (and extremely problematic) notion of 'exotic' locales providing access to 'pure' and 'natural' societies for ethnographic study (¶ **141**).

51. Regarding the needed adaptation to hybridity, Jan Blommaert (who sadly passed away last year at a relatively young age) observed: "It is a shift from a scholarly universe almost entirely dominated by theoretical and methodological preferences for offline spoken discourse in fixed and clearly definable timespace, sociocultural and interpersonal contexts and identities, to one in which the world of communication is—at the most basic level—seen as an online–offline nexus [i.e. a state of hybridity] in which much of what we assumed to be natural, primordial and commonsense about language-in-society needs to be revised, rethought and redeveloped" (2019, p. 486).

52. This work is concerned with public microblogging platforms. Microblogging, alongside various other types of social media, has been put to extensive internal use by organisations (e.g., Riemer and Richter, 2010 ; Treem and Leonardi, 2013). Such internal implementations, however, are typically 'walled gardens' for approved and authenticated users, and those restrictions are set by the organisations themselves. Organisations operating microblogs for external use (i.e., public microblogs, such as Twitter) are motivated to increase their user base, and so can be expected to set the minimal restrictions on access possible within a given legal or political jurisdiction.

53. In social terms, nothing is ever arbitrary. Pierre Bourdieu noted that one of his students (Soulié, 1995) found that "research topics (masters theses and subjects of doctoral dissertations) in philosophy and sociology … are statistically linked to social origins and trajectory, gender, and above all to educational trajectory" (2003, p. 283).

54. The use of this quote and its context should be explained. The quote is taken from *The Ethics of Democracy*, written by a young John Dewey in response to and critique of Henry Sumner Maine's *Popular Government* (1885). In that work, Maine observed that "the dispassionate student of politics [understands] that Democracy is only a form of government" (p. 64)—such perspective was anathema to Dewey, who throughout his life argued that democracy must be thoroughly embodied in society and enacted through it. One of the main elements of Dewey's critique (1888) is that a strictly functional view of democracy as a mode of governance leads to a quantified, atomised view of society, wherein individuals and communities are treated as units of political power to be gathered up by those who would govern (cf. Maine, p. 29). Thus whereas Maine nominally restricts his discussion to forms of government, Dewey is reacting to the resultant effects in and on society. He repudiates Maine's way of thinking and its reliance on the "the idea of men as a mere mass" (which, Dewey noted, underpins 'social contract' understandings of political legitimacy), saying "The fact is, however, that the theory of the 'social organism,' that theory that men are not isolated non-social atoms, but are men only when in intrinsic relations to men, has wholly superseded the theory of men as an aggregate, as a heap of grains of sand needing some factitious mortar to put them into semblance of order" (Dewey, p. 6). His wroth and righteousness notwithstanding, in the decades to come the quantified, atomised view of society (if there were such a thing) would largely supplant the relational, social perspective, as discussed in **chp. 2**.

55. It is the author's understanding (not interpretation, mind you, but belief) that the central philosophical, and practical, motivation of ethnography—which underpins this entire work—is to perceive people as they perceive themselves, so that we all might see ourselves as we are. That is one individual understanding among countless others. In terms of methods, however, there is thankfully much more coherence.

Ethnographic approaches are well established in the social sciences, tracing back to the descriptivist approach to language, meaning, and culture of Franz Boas and his students in the first half of the twentieth century (Darnell, 1990), and rooted in Max Weber's practical imperative of an empirical *verstehende soziologie* (Herva, 1988). Ethnographic approaches, if not ethnography strictly speaking, have become mainstreamed in the last 50 years. For that reason, the author assumes sufficient familiarity on the part of the reader to proceed without undue elaboration on the character of ethnographic method and methodology. That said, those decades have produced a proliferation of more critical and diverse literatures—each with their palette of underpinnings, terminologies, and purposes—that is at once both heady and befuddling. For those who would appreciate a brief, straightforward account of some central tenets of ethnography presented in concrete, practical terms, see Raymond L. Gold (1997). Note that Gold was trained at the University of Chicago (cf. **chp. 3 § 3.1**) under Everett Hughes (Chapoulie, 1996, p. 17), receiving his PhD in 1954 for the dissertation "Toward a Social Interaction Methodology for Sociological Field Observation". Thus his perspective (1997) can be confidently viewed as rooted in mid-century ethnographic method within the US academy (cf. Driessen, 1997), although Gold helped to lay the groundwork for the critical turn to come in later decades. For an extensive account of sociological training in Chicago at that time, focusing on the roles played by Gold, Everett Hughes, and Buford Junker, see Daniel Cefaï (2000).

56. Lorraine Bayard de Volo and Edward Schatz (2004), in arguing for the utility of ethnography in political research, nevertheless feel obliged to begin by dispelling certain misconceptions about the approach. An example of this tension:

> [A] fallacy about ethnographic methods is to equate them with a more recent trend in cultural anthropology. In parts of that sub-discipline, the ethnographic method has achieved star status. Allowing for ongoing dialogue ("reflexivity") between subject and object, the method has become the defining feature of entire bodies of research. In its worst forms, this reflexivity can become a sort of transcendental principle that rivals any methodologically narrow navel-gazing practiced in other disciplines. [A footnote here directs the reader to a critical article entitled "Should We Make Political Science More of a Science or More about Politics?".] These excesses, however, are the exception. The norm is for ethnographic work to offer potentially profound contributions to the body of knowledge about social and political life—no matter the intellectual tradition of the researcher. (p. 268)

The mention of 'navel-gazing' is not cheap interdisciplinary sniping, but rather comes from valid, pragmatic concerns about the role of reflexivities in social inquiry. Consider, for example, Pierre Bourdieu's contrasting of "scientific reflexivity" versus the "narcissistic reflexivity of postmodern anthropology", which was the bright thread in his 2000 Huxley Memorial Lecture at the Royal Anthropological Institute of Great Britain and Ireland (2003).

57. Much of the history of ethnography is rooted in disciplinary anthropology and its antecedents. Note that the disciplinary identity of anthropology, specifically in contraposition to sociology, was long rooted in fieldwork (Gupta and Ferguson, 1997b, p. 1), moreover in a certain curious view of fieldwork. For example, in 1969 Donald MacRae declared that the mark of anthropologists is that "They have, in principle, all undergone the ordeals of a common *rite de passage*, i.e. they have all undergone at least a year of field work in some exotic area" (1974, p. 4). Such a framing of the field is increasingly *passée*, i.e. recognised as problematic and inappropriate.

58. Margaret Mead was also a student of Franz Boas, but much later, and her own focus was not linguistic.

59. The former fixity of place in ethnography was perhaps in part due to the reifying eye of savvy observers and eager audiences (cf. Stocking, 1992, pp. 40–59). In the author's view, fixity of place (i.e. a space in time) is a basic human process of sense-making ; however, that fixity might be pushed toward fixation when experience of certain places (e.g. the 'exotic') confers social benefit.

60. For an introduction to the terminology, basic methods, and core issues of this more traditional mode of ethnographic work, see *In the Field* by the recently departed Robert Burgess (1984). This book is part of a series edited by Martin Bulmer that aimed to present concise introductions to core methodological topics in social research. For those interested, it is strongly recommended.

61. That is, those sciences firmly grounded in social subject models.

62. Individual attempts at such reconceptualisation can be unsettling, and even rejected outright. Consider how this mode of thinking quickly runs headlong into the essential dualism that characterises the Western tradition, manifesting variously as divine–profane, man–woman, knowledge–innocence, purity–sinfulness, metaphysical–physical, mind–body, immediate–mediate, noumenal–phenomenal, self–other, sameness–alterity, quantitative–qualitative, right–left, and so on. While binary typologies are certainly useful for separating wheat from chaff or sheep from goats—and for basic methodological thinking, for better or worse—when applied to the wild variety observed in social (and natural) phenomena they will invariably confound and harm. Note that this is not a binary good–bad judgment on the part of the author, but rather recognition of the empirical effects of extreme reductionism (cf. **chp. 1 § 1.3**).

63. In the author's experience, introducing a notion of 'lived hybridity' (a nonce use here, extending the notion of 'lived experience'—of which more at the end of this note) to students (and colleagues) can on occasion elicit a strong negative reaction, partly conceptual and partly emotional, to the decentring of 'humanness' (however that may be conceived). The reaction is similar to humanist responses to the notion of radical symmetry (cf. Vandenberghe, 2002).

Regarding the notion of 'lived experience': this term is frequently encountered in the literature of a variety of fields. It is rarely specified in place, and adjoining citations often lead to the same or similar term, likewise unspecified. What it purports to denote may seem familiar, likely due to its constituent elements—we are unlikely to pause at 'lived' or 'experience' used by themselves—but nevertheless stands at the edge of intuition. Unanchored as it often is, the term seems to serve as a watchword (or shibboleth, if you prefer a meta-watchword) for certain philosophical and methodological stances.

That said, the term does indeed have a history of specifications and precise denotations. For an compact and step-wise account of the term's origin (from the German *Erlebnis*), development, and adaptation into English—as well as an explication of the concept itself—see Robert Burch (1990, pp. 132–137).

64. The obvious, likely, yet unwarranted alternative would be to carry on as before, and settle in for further decades of sporadic handwringing about disciplinary identity, purpose, and direction (e.g. Gerbner, 1983 ; Fuchs and Qiu, 2018 ; among many others, and across academies). While many are eager to see 'ferment in the field' in a positive light, as reflective of diversity and energy, the persistence of indisciplinarity renders schools, associations, and academies manipulable, thus ceding initiative to external (often not-so-external) forces and interests to shape a discipline rather than the scholars themselves (cf. Carrasco-Campos and Saperas, 2020). Unchecked fermentation leads, after all, to souring and spoilage.

65. Measurement by indirect observation would depend on causal argumentation, which in the author's view is questionable in proper social inquiry (cf. Arjas, 2001).

66. The solution proposed by Theocharis was an amendment to a functional typology of political participation developed by van Deth (2014). That typology included intention and consequence, which Theocharis rightly suggested to bracket by way of focusing on context of action.

Note that 'context of action' is this author's term, a shorthand intended to convey the intentionality and consequentiality that Theocharis and van Deth seek to subsume in a focus on political context. They say simply 'context'.

67. The cycle of US elections are commonly framed around federal (versus state-level) elections. Presidential elections occur in years evenly divisible by four (thus the most recent was 2020), and midterm elections occur at the midpoint of the four-year Presidential term. The latter elections select the entirety of the House of Representatives (the lower house) and roughly a third of the Senate (the upper house). While elections at state, county, and municipal levels take place annually, the bulk of participation (in terms of media coverage, voting, mobilisation, etc.) occurs in Presidential years, followed by midterm years.

68. For a country of over 330 million persons across six domestic time zones, the notion of a national-level discourse is a problematic fiction driven by the electoral politics of a two-party system.

There are nine time zones if one includes all US territories (i.e., the 50 states, plus American Samoa, Guam, the Northern Marian Islands, Puerto Rico, and the US Virgin Islands). However, the 50 states cover six zones, with the eastern most (the aptly named Eastern Time Zone) at UTC-05:00, and the western most (the Hawaii-Aleutian Time Zone) at UTC-10:00.

69. These proportional measures (to allow comparison across differing populations) were obtained from the 2016 American Community Survey conducted by the US Census Bureau, and included age brackets, gender, ethnicity, and education. Percentage change was also calculated against the 2010 American Community Survey, but all considered states showed similar percentage changes (except for Florida, which was above average in growth in the over 65 demographic ; Florida is rightly known as a retirement state).

70. These were California, Colorado, Connecticut, Florida, Georgia, Idaho, Maine, Michigan, Minnesota, Nevada, New Mexico, Ohio, Oklahoma, South Dakota, Tennessee, and Wyoming.

71. The original goal was to select two states that not only were demographically similar to the demographics of the entire country, but also to each other, so as to provide a comparative perspective in this work. Michigan and Ohio were selected, as they are very similar demographically and in population, and also border each other. However, for practicality and reasons of time, Ohio was eventually dropped as a case.

72. Twitter has likewise substantially tightened access in recent years (Bruns, 2018, p. 65). Changes to the API during the course of data collection for this work forced a number of changes to the method. For any researcher making active use of APIs, or interested to explore such digital methods, Deen Freelon's article "Computational Research in the Post-API Age" is an obligatory and cautionary read.

73. In this light, the interpretive distinction between behaviour and action (i.e. intentional behaviour) is weakened (see the following note), and consequence is implicitly presumed. Thus this adaptation can be understood as providing further support to the fifth adaptation, which brackets intent and consequence with a focus on context of action.

74. This is fundamentally a social perspective, in that distinctions between individual and collective are an analytical heuristic rather than a typology of observed phenomena. Some fields do support and study the latter typology, as those fields build upon a different subject model than the social sciences proper (and thus they are behavioural sciences ; cf. **n. 88**).

75. From a rhetorical perspective, this adaptation places political participation in hybrid society firmly in the sphere of public discourse. (Perhaps not coincidentally, the author's earliest formal training in the study of communication was in the subfield of 'technical communication', which is deeply influenced by the rhetorical tradition. Note that this subfield is to be found primarily in the US academy.)

76. That is a recognition that normative framings of political participation, which are a choice, lead to poor understandings in hybrid society.

77. They are bracketed because they are unobservable, but nevertheless can be interpretively proxied (as opposed to directly proxied) by context—hence 'productive' bracketing. Even if a researcher proposed a clever way to observe intent or consequence, they remain indeterminate concepts. That is to say, these concepts lack direct empirical grounding. As such, any analysis based on indeterminate concepts is necessarily interpretive—a post hoc rationalisation of observed events.

To elaborate briefly, and to indulge in a helpful bit of methodological dualism (cf. **n. 61**), we can conceive of the world as composed of the phenomenal (the purely physical) and of the noumenal (the purely mental). A concept can be understood as a construct assembled from our phenomenal sensory responses (percepts) and our noumenal thoughts and notions (incepts) so as to integrate experience and understanding. A concept is determinate if the noumenal and phenomenal aspects are consistent, that is, incepts are directly informed by percepts. 'Fire can be harmful' is such a determinate concept. An indeterminate concept does not have that direct percept–incept link. This is the case with intent and consequence. These concepts have no fixed phenomenal component, as intent and consequence do not themselves exist in the physical world ; we only name, describe and understand them through effects observed in (conceptually) related phenomena.

Note that indeterminate concepts are in no way 'lesser' than determinate concepts. Rather, they are just more difficult to work with, requiring informed interpretation to understand them in a systematic manner. That is the essence of social inquiry.

Note further that the above is reflective of the author's (nascent) efforts to repackage certain long-standing philosophical ideas into forms more readily digestible in methodological and pedagogical contexts. The reader's patience is appreciated.

78. Definitions, understood as explicit and static conceptualisations of a thing, reflect decisions already taken as to the nature or structure of the thing in question. This can lead to difficulties in adjudging whether a thing does or does not fit a definition. The advantage of the approach of Theocharis and van Deth is to replace blanket definitions with a step-wise approach to constituent elements of a possible overarching definition [NB: These elements are simply smaller definitions, as is clear in Theocharis and van Deth (2018, p. 144)]. As their approach does not require that all phenomena fit all elements, the approach then becomes taxonomic and decision-oriented. See **chp. 3 § 3.3** for a description of this approach.

79. It is important to note that W. Lance Bennett and Barbara Pfetsch (2018), speaking from the subfield of political communication, describe a situation much like that described in the subfield of political participation in **chp. 3**—that is, a subfield that had gone through various stages of evolution during the twentieth century, but that now finds itself challenged by wide-ranging sociotechnological shifts. This should be expected, as both subfields investigate similar social phenomena. They stress the urgent need for a reassessment of core concepts and methods in their subfield (pp. 246–247), in a manner quite similar to the suggestion of Stuart Fox (2014) in relation to political participation (**chp. 3 § 3.2.2**). It is worth reading Bennett and Pfetsch, bearing political participation in mind, as their suggestion is much more thoroughly developed and argued.

This is no criticism of Fox, mind you—the article of Bennet and Pfetsch was a contribution to a special issue of the *Journal of Communication* (68:2) addressing 'ferments in the field' (cf. **n. 63**), whereas Fox nonchalantly put forward his suggestion, while still a PhD candidate, in a review of Paul Whiteley's (2012) *Political Participation in Britain: The Decline and Revival of Civic Culture*. In the review Fox notes that, due to the almost exclusive focus of Whiteley on institutional forms of participation, "He cannot be sure that he has captured a realistic and valid assessment of the political participation of British citizens, nor can he identify whether political participation is actually in decline, or if it is actually evolving" (p. 502). Good on you, Stuart.

80. While Theocharis and van Deth (2018) incorporate the modification suggested in Theocharis (2015) to bracket questions of intent and consequence to a focus on context, the former piece nevertheless maintains that one must "look at the *political context or the motives* of the participant"(p. 144 ; emphasis original). However, as noted in **n. 76**, intent is fundamentally unknowable—arguably even to ourselves—and can only be approached in an interpretive mode. Furthermore, any explicit statements of motive (e.g. 'I am motivated by concern for the safety of children') should not then be viewed as statements of motive, but rather as statements of a situated social calculus, that is, of context. In this manner, a perspective on political context is sufficient in itself and, moreover, practicable (whereas a perspective on motives cannot be sensibly operationalised).

81. Here the author would greatly prefer to say 'communication', implicitly indicative of a coherent discipline. However, as of yet there has never been a coherent discipline, in any academy, that might properly claim the name. However, the author firmly believes that such a discipline will emerge in time, and that belief is among the central motivations behind their study and labour. Members of that future discipline perhaps could be thought of as 'students of language in structural mediation'.

In this regard, note that the reference to 'students of language and of language in society' does not intend 'language' to denote any specific mode (e.g. speech or writing) or ideology (e.g. English or Academian) of language, or even necessarily human language. Rather, the author conceptualises language—in an admittedly expansive manner—as the intersubjective semiotic relations whereby meaning is negotiated and manipulated. In terms of the ontological footing of this work (**chp. 2 § 2.3**), language would then be the prime technical affordance that interfaces between the social and material domains. In that light, the 'students' to which the author refers are left unspecified for practical reasons (and on the assumption that if you are one, then you know it). Note that the remainder of the body paragraph should be understood in this manner.

82. We are thus better positioned to recognise not only that 'we do things with words' (following Austin), but moreover that our words and things are jointly meaningful. In that way, speech, as a mode of language, is at the very core of all social phenomena and human experience.

83. Note that J.L. Austin and John Searle are given here as exemplars likely to be recognised within the fields of communication and media studies. Note further the distinction between behaviour and action being that the latter is intentional, and thus agentive. On intention see Searle (1983) and Stephan Fuchs (2007) ; cf. Duranti (2006) ; note that intention is defined in this work as "a representation of effecting consequence in the world" (**chp. 2 § 2.2.2.1**).

84. Ruth Wodak (2014) provides an impressively condensed tour of the range of use of 'discourse' in the social sciences:

> Almost no paper or article is to be found which does not revisit these notions, quoting Michel Foucault, Jürgen Habermas, Chantal Mouffe, Ernesto Laclau, Niklas Luhmann, or many others. Thus, "discourse" means anything from a historical monument, a *lieu de mémoire*, a policy, a political strategy, narratives in a restricted or broad sense of the term, text, talk, a speech, or topic-related conversations to language per se. We find notions such as racist discourse, gendered discourse, discourses on un/employment, media discourse, populist discourse, discourses of the past, and many more—thus stretching the meaning of "discourse" from a genre to a register or style, from a building to a political programme. (p. 302)

85. As an immediate example of this concern in the study of communication, consider how the ontological footing of this work could be understood as an ontological analysis of discourse (cf. **n. 80**). That is not a coincidence.

86. Goodnight's expertise is in the rhetoric of argumentation. The rhetorical tradition, while relatively commonplace in communication studies (though not so in media studies) in the US academy, seems to be less common in other Anglophone academies. While Goodnight's typology has been influential primarily (to the author's knowledge) in argumentation studies, naturally, its potential utility for the studies of communication and media should be considered. For more on Goodnight's typology and its influence, see Robert Rowland (2012) [NB: While Goodnight's 1982 article cited in the body might not be readily accessible—by grace of for-profit academic publishing—it is reprinted in Rowland.]

While we are on the subject: for students of communication (and as well students of language in society generally) who are unfamiliar with the rhetorical tradition, there is much to recommend. The main thrust, in terms of analytical and pedagogical utility to the field, is not a concern with the content of communication, but rather with how it comes to be and is conducted—Goodnight (1987) notes this clearly. From a perspective of technological mediation, consider that main thrust of rhetoric in relation to Marshall McLuhan's dictum 'the medium is the message' (see **chp. 2 § 2.1.2.1**).

87. This piece, snappily titled "Public Discourse", is a brief (barely four pages) discussion of Goodnight's typology, focusing specifically on the public sphere. It is a succinct and engaging introduction, and would be an excellent addition to post-graduate syllabi on communication theory.

88. Added to this, in the case of the United States, is an expansion of the franchise, despite ongoing and concerted efforts to the contrary.

89. Durkheim makes this quite clear, continuing to say that—if we did indeed attend to all possible social phenomena—then social science "would possess no subject matter peculiarly its own, and its domain would be [confounded] with that of biology and psychology" (1895, p. 50). The horror!

It is curious to consider this statement in light of cognitivist approaches in various fields, that in fact purport to blend these disciplines ; the sleight of hand there is an untoward tendency for individualist explanations of phenomena, thus remaining firmly in the domain of behavioural analysis, rather than social.

90. In fields that address questions of technology and political participation in terms of how it is done rather than what is being done, the incorporation of hybrid concepts is well underway. Consider, for example, the strand of work that engages with 'digital democratic affordance' (Dahlberg, 2011 ; Deseriis, 2020) Consider also the 'how versus what' approach in light of **n. 80**.

91. An instructive discussion in this regard is Ben Berger (2009), which discusses the emergence in the US academy (and broader public) over the last 30 years of 'civic engagement' as a concept, and trope, noting how it has come to distort the study of polities and their functioning.

92. The long-standing US commonplace 'All politics is local' evidently only applies to those who play the game properly, that is, according to institutionally sanctioned forms of participation and by way of conventional modes of discourse. Note that this sanctification of the political as a relatively rarefied and specific domain is long-standing in the US academy, as this chapter discusses further.

93. In regard to potential drivers of shifting patterns of participation, consider Jennifer Oser and Marc Hooghe's (2018) application of latent class analysis to the European Social Survey 2012 where they found that individuals with a social rights conception of citizenship (Marshall, 1950, pp. 46–74) have higher levels of non-institutionalised political participation, but lower levels of participation overall (among five latent classes).

94. Consider the *Epic of Gilgamesh*, among the oldest of our recorded narratives. The broad arc of the story is that Gilgamesh, ruler of the city of Uruk, begins as a harrier and abuser of his people, and ends as a builder, instructor, and exemplar—but only after the people's cries to the gods for pity leads that man to conflict, loss and thereby, in acceptance, to wisdom. Already then, from the context of one of our earliest cities, come accounts of how rulers are known to be, and dreams of how they might be made better. This is simply the author's own inexpert interpretation. For a properly informed discussion, which also touches on the topic of rulership, see Tzvi Abusch (2001). For the tale itself and a critical examination of its history, see Andrew R. George (2003).

95. To be clear and fair, Goodnight's typology is not intended to describe patterns of governance. Its application here is simply the author's suggestion of a heuristic for examining public discourse.

96. "By definition and by purpose" is not an off-hand comment. Those working in an institutional vein are neither ignorant nor themselves naive. Institutional framings of political participation exist amongst a wide range of other framings, some of which are expansive and critical. Thus institutional framings are a considered choice (Bennett and Bennett, 1986, pp. 160–162).

97. Such exceptions are not insignificant to the people thus deprived ; they are insignificant, however, if your concern is the institutions of the State.

98. Note that this is a significant work in the literature of political participation, and it is cited a number of times throughout **chp. 3**. It can be taken as a representative of the conventional conceptual view, and is quoted at length for that reason. Note also that such belief and trust in institutions is neither universally shared nor well supported by historical analysis (cf. **¶ 163** below). Note further that these same authors published another work around the same time that can be seen as highly representative of the conventional methodological approach (Brady, Verba and Schlozman, 1995). That paper addresses the 'socioeconomic status' model of political participation (see **chp. 3 § 3.1.2 ¶ 109**)—which seeks to explain political phenomena (at a given level of analysis) by way of three components: education, income, and occupation—and suggests to expand that model with the incorporation of resources: time, money, and 'civic skills' (more below). That paper has been cited with increasing frequency since its publication, and increasingly beyond the discipline of political science.

As a further comment, note that 'civic skills' are understood to be communicational and organisational skills ; a portion of the operationalisation of this component was a measure of years of education, as well as a vocabulary test. Consider how operationalising political participation using measures that in quite recent past featured prominently in de facto electoral exclusion (literacy tests were abolished by the Voting Rights Act of 1965 ; see **n. 102**) underscores the overall point being made in the argument here.

99. The continuing crisis in the study of political participation (that of hybrid society) is introduced in **chp. 1 § 1.3**, and is a basic motivation of this work. See also **n. 78** for a description of the state of affairs in the study of political communication.

100. The Social Science Research Council is an independent research organisation founded in 1923 through an initiative of the American Political Science Association. The Council's first president was Charles E. Merriam of the University of Chicago who, as noted in **chp. 3 § 3.1.1**, was among the central figures in the first days of the scientific study of political participation, and who also was a driving force behind the initiative to create the Council. The Committee on Comparative Politics was in operation from 1954 to 1972. It is worth noting, in the context of this work and its overall argument, that the Council is not a foundation with its own endowment, but rather relies on support of private organisations. During its first 50 years, thus including the period in which the Committee was active, the bulk of the Council's funding came from major foundations including the Russel Sage, the

Rockefeller, the Ford, and the Carnegie Corporation (cf. **n. 7** ; Rockefeller Archive Center, n.d.).

The Committee on Comparative Politics produced the influential series 'Studies in Political Development', which is notable because it marked a shift in comparative work from an internal focus (i.e. on Western states) to an external focus (i.e. non-Western) (Mello, 2011). As we have already touched on the topic of the funding driving this work, note that the series was initiated by the Committee under a grant from the Ford Foundation (Pye, 1963, p. *vii*). That series can still be found on its publisher's website as of March 2022: https://press.princeton.edu/series/studies-in-political-development. There were nine volumes in total, although only eight are listed by the publisher. The missing volume is no. 8, Charles Tilly (ed.) *The Formation of National States in Western Europe*. Why that volume is missing is unclear, although perhaps Tilly's thesis that "War made the state, and the state made war" (1975, p. 42) was not considered a good fit with the emphasis in the rest of the series on the 'proper and civilised' development of society and institutions—Tilly said the quiet part loud. Note that Tilly would eventually be part of the 'Perestroika Movement' of the late 1990s and early 2000s in US political science (cf. **n. 8** ; Monroe, 2005). He was in fact mentioned by name in the open letter that first brought this movement into public (pp. 9–11).

101. Consider van Deth's observation of the expansion of government roles in society (**¶ 158**), as well as the similar comment by George Armstrong Kelly in **n. 108**.

102. Weiner buried the lede—this definition arrives on the twenty-eighth page of the chapter.

103. It is sad to reflect on Verba's examples of non-elected decisionmakers, considering that the US Postal Service in recent years has been financially hamstrung thus opening the door for privatisation ; the police have assault weapons, body armour, tanks, and a tendency to murder and let murder happen ; and planners build bridges too short for public transport to pass and route highways through minority neighbourhoods. That is the state of the United States in 2022, and 1967 was little different, except the military equipment was mostly abroad and the Postal Service had not yet been impoverished (cf. **¶ 165** and **n. 97**).

It should be noted that Verba was writing in the wake of the 'Long, Hot Summer' of that year, in which there were more than 150 riots across US cities, the causes of which are generally understood to be 'racial tensions' (cf. Lemberg Center, 1966-1967). In relation to the discussion at hand, note that such tensions follow the 1965 passage in the United States of the Voting Rights Act, intended to prohibit racial discrimination in voting (not to claim that the Act was the cause ; the past is never so simple). The protections of the Act were extended by Congress a number of times in the years following its passage, although it has been significantly weakened ('gutted' is a term commonly heard) by more recent Supreme Court rulings (Shelby County v. Holder, 2013 ; Brnovich v. Democratic National Committee, 2021). That is a dispiriting example of Goodnight's public sphere in operation at a generational scale (cf. **¶ 157**) and also highlights what his ideal typology cannot address—and what the conventional approach to political participation chooses not to address—that is, members and groups acting to reduce the community (**¶ 160** ; see also the following note).

104. As noted in **n. 95**, institutional framings (and the naivety upon which they are necessarily premised) are a considered choice. It is not as if trained political observers do not see what a system does ; the question is how and what they choose to tell others about it (cf. **n. 93**). Consider Benjamin Ginsberg's critical perspective in *The Consequences of Consent: Elections, Citizen Control and Popular Acquiescence* (1982). Note that Ginsberg, along with the late Theodore Lowi, is co-author of the textbook *American Government*, now in its 16th edition (2021). Lowi is mentioned in **n. 8** for a fascinating cross-journal exchange between himself and Herbert Simon (Lowi, 1992, 1993 ; Simon, 1993), who will appear in the following section. Lowi's thesis (in "How We Became What We Study") was that US political science is in thrall to the State, to which the "diabolical mind" (1992, p. 4) of Simon naturally took exception. All three pieces are strongly recommended. Bear in mind that Lowi's piece was published the year following his tenure as president of the American Political Science Association, the preeminent association within that discipline (certainly in the US academy, but arguable also in the Western academy during the twentieth century). These trivia are more than simply contextual colour—the author's intent is to underline that institutional framings, with their consequent befuddlement in the face of systemic failures in regard of the general welfare, are a choice.

105. Consider the 2010 ruling by the US Supreme Court in Citizens United v. Federal Election Commission that found independent political expenditures by corporations to be protected speech under the First Amendment, thus effectively removing all caps on corporate electioneering and political advocacy.

106. On this topic: for a historically contextualising view of public discourse in the United States (prior to the period discussed in **chp. 3 § 3.1**), see Daria Frezza (2007) *The Leader and the Crowd: Democracy in American Public Discourse, 1880–1941*. For example, "it must be stressed that American national identity was defined according to the well-known paradigm of exclusion–inclusion" (p. 15).

107. These boundaries are set, of course, by those with the power to define the place of the community in terms of who is, can be, and can remain a 'legitimate citizen' (cf. Frezza, 2007, p. 15 ; also **n. 111**).

108. There is sadly little purpose to offer specific citations here ; it is perhaps more revealing to scan the results of a literature search performed on the terms 'United States' and 'exclusion'.

109. Given the overall discussion in this work of the conventional approach to political participation and its institutional fixation, it is perhaps enlightening to note a comment made in a 1966 review of books, of which one was Lucian Pye's *Aspects of Political Development* (**¶ 172**). The reviewer, David K. Marvin, an expert on African affairs in that post-colonial period, observed:

> By a kind of Parkinson's law, it is "only when the old is about to disappear that we make it the subject of self-conscious study as an eternal norm of human existence." Nationalism is enshrined conceptually by scholars, for example, just as the conditions favorable to nationalistic behavior disappear. (1966, p. 508) [NB: Parkinson's law is the

commonplace that “Work expands so as to fill the time available for its completion” (Parkinson, 1957, p. 2).]

Compare the observation of George Kelly Armstrong: “Today’s problematic nature of citizenship … is in part linked to the demise of the concept of the state in the twentieth century, the very time when the powers of the empirical state were growing inordinately” (1979, p. 21).

110. While there are of course alternative models, sociodemographic models have been dominant. The aptly name ‘resource model’ of political participation has been, and remains, one of the central models of the conventional mode. It was introduced by Sidney Verba (to whom we have been introduced), Kay Lehman Schlozman, and Henry E. Brady almost 30 years ago (1995), and expanded on in several books significant in the field (Verba, Schlozman and Brady, 1995 ; Schlozman, Brady and Verba, 2018 ; Schlozman, Verba and Brady, 2012). The model was an extension to the long-established ‘socioeconomic status’ model (cf. **n. 97**), which seeks to explain political participation by way of three components: education, income, and occupation (jobs being associated with status, you see). “No other social factor has been as consistently linked to differential rates of political participation as socio-economic status” (Bennett and Bennett, 1986, p. 183). For demonstrations of this observed linkage, see ibid. (pp. 183–186), Conway (2000, pp. 25–30), and Milbrath and Goel (1977, p. 92). Note that these sources are suggested by Brady, Verba and Schlozman themselves (1995, p. 290, n. 4), to which they add their own. Of the literature cited in **chp. 3 § 3.1**, much, if not most, will appear in the sources just given. Bear in mind that these sources related to the socioeconomic status model of participation ; after 1995, the literature begins to shift towards the resource model—Verba, Schlozman, and Brady are significant figures in the US academy, and their work has carried weight.

111. In truth it is impossible to untangle the three, as should be clear to the reader—each will impinge upon on the others ; nevertheless it is instructive from an analytical point of view to consider them separately.

112. Recall that, as discussed in **chp. 2**, there is no need for a tool to be a physical artefact. Consider the topically relevant example of survey method, which has played such an important role for social research—not only have those methods been adapted to developments in statistics, but there have been steady improvements in practice and adaptation to new contexts of application (Alwin and Campbell, 1987 ; Lupu and Michelitch, 2018).

If talk of ‘tools’ nevertheless seems strange to the reader, consider the case of the telescope. Its application to astronomy prompted a raft of new discoveries, and improvements to the instrument itself furthered research—for a dated though nonetheless delightfully detailed look at technical developments, see Henry C. King (1955). Those developments led to epistemological shifts, for example Herschel’s discovery of Uranus leading to new approaches and successes in the search for asteroids, touched on by Thomas Kuhn (1970, pp. 129–130). Other developments with an epistemological impetus, such as the development of stellar spectroscopy prompted by observation of absorption lines in light from the sun at the turn of the nineteenth century (see King, chp. 14), contributed to Edwin Hubble’s proof that there are galaxies beyond our own and the general recognition that the universe is expanding. Methodological changes with an ontological impetus resulted from those preceding epistemological shifts which produced growing evidence for the 'Big Bang' and thus a thoroughgoing reformulation, not simply of astronomy, but of cosmology itself (Osterbrock, Gwinn and Brashear, 1993). And if 'tools' still seems strange, consider Lev Vygotsky's concept of 'psychological tools', briefly described in chp. **2 § 2.1.1**.

113. Technical innovations in themselves must first be encountered, recognised, adopted, and shared (Rogers, 1962 ; NB: the term 'early adopter' originates in that work ; see p. 315). Consider how this contingent uptake relates to the concept of affordance (**chp. 2 § 2.2.1**), wherein an affordance only 'exists' if perceived as such. To continue the example of the telescope from the previous note: the properties of lenses (of liquid, rock crystal, glass) were known in ancient times, but they were originally used for decoration or for burning by focusing the sun's rays. Only later were they noted for their effects on vision itself, and the mounting of dual lenses in a frame (i.e. spectacles) was not done until the thirteenth century. It would be some time yet before lenses were used in a compound fashion for magnification, and still yet more before someone popped them into a tube.

114. To be clear, the explicit call of Fox (2014) is for a reconceptualisation of political participation. However, the call for a reconsideration of subject model is implicit, as consequence of the logical extension of such a reconceptualisation to a complex of phenomena increasingly recognised as social in nature. Thus the call in this work is explicit.

115. Simon's influence extends far beyond political science. Note that he was yet another product of the University of Chicago, having studied there as an undergraduate, and pursuing his PhD (1942, "Administrative Behavior", later published under the same title) under Charles Merriam and Harold Lasswell, among others. Simon went on to be a recipient of both the Turing Award (1975) and the Nobel Memorial Prize in Economics (1978), among numerous other awards. Simon was a major proponent of the rational actor model of behaviour, and his studies of decision-making (especially under uncertainty, for which he won the Nobel) have had enormous influence within and without the academy—if you have ever used the term 'satisficing', then you can thank Simon. Much of his work (for example in operations research and game theory) played an unnervingly central role in the Cold War. The author notes that his first postgraduate training was in political and international relations, 20 years ago across the street from the United Nations Secretariat in Geneva—in an environment suffused with conflict and 'Great Power' thinking, Simon and his writings were nigh unavoidable. In many ways they continue so. For more on Simon's life and work, see respectively Augier and March (2001) and Crowther-Heyck (2006). Also be sure to see **ns 8** and **103**.

116. Linear equations, for example, are merely an alternate grammatical mode of language writ large—such a symbolic calculus is functionally little different from the extended argument presented in this section.

117. In the domain of policy, Ralph E. Strauch (1976) addressed a prominent (though often implicit) approach that he termed 'quantificationism' (pp. 133–136). Such an approach holds that quantitative methods are a good in themselves, in that they will always yield results superior to other methods. The approach is furthermore characterised by a belief that such methods are appropriate even to complex phenomena ("squishy problems", p. 134). The rationale, Strauch observes, is based in the assumption that the application of such methods captures the scientific method itself because "it appears to emulate the reductionism inherent in the physical sciences" (ibid.). The key point he makes here is that reductionism is not the actual core of the physical sciences in terms of those processes of knowing, but rather is simply an effect of the types of problems (and thus subject models) that those sciences address.

While Strauch's discussion is nearly 50 years old, the issue is current. Consider a recent discussion observing that political science as a discipline uses the term 'science' because it is based in scientific method. Science being, in its modern sense, "a method of learning based on systematic observation using the scientific method [sic]" (Bond, 2007, p. 897), where the key aspect of scientific method is theory building through quantification and hypothesis testing (p. 899).

The author observes that, in the domain of the social and behavioural sciences generally, quantificationism can be understood as a mode of subject model failure driven by a range of disciplinary pressures, not simply a credulous approach to method.

118. "Therefore, the human organism, rather than groups or the political system, usually is taken as the unit of analysis" (Milbrath, 1965, p. 3).

119. From that perspective, the fungible individual as subject model is appropriate. After all, the concern is not individuals, but rather with collective actions in relation to the State. To the extent that models hinge on sociodemographic variables, there is little need to know about who acts beyond their brackets.

120. It should be noted that conventional methods are ruthlessly adapted to their purpose, much like sharks. Resource-based models of political participation find with "monotonous regularity" (Nagel, 1987, p. 59) that individuals with greater resources are more frequent and more successful participants (see **n. 109**). Similarly, Lester Milbrath and Madan Lal Goel charmingly observe that, in numerous countries, "No matter how class is measured, studies consistently show that **higher-class persons are more likely to participate in politics than lower-class persons**" (1977, p. 92 ; emphasis original). Do tell! As a further note in that regard, Milbrath and Goel also mention how "paradoxically, the very absence of class ideology and class parties in the United States makes it more likely that higher-class persons will participate in politics" with the consequence of giving further advantage to already advantaged groups (pp. 92–93). This observation has been made by others as well (e.g. Verba and Nie, 1972, p. 340). Consider, in light of the discussion in the preceding section, how that paradox might be explained. Consider further Milbrath's earlier insistence that "no political activity could be considered normal or routine for everyday existence. Persons turn to politics only when their basic physical needs, such as food, sex, sleep, safety, affection, have been met" (1965, p. 18, fn. 6). A paradox, indeed.

121. Note that throughout this work, there is no mention of meaning being 'made' or 'created', but rather of being 'negotiated'. It is the author's position, in this work and generally, that there is no new thing under the sun, and that our human variety and creativity stems simply from recombination (be it of genes or memes). The fixation on debating and seeking the origins of things seems characteristic of disciplines and other institutions that privilege non-social explanations of phenomena. However, as Frank

Kermode tenderly observed, we live out our lives constantly rushing "into the middest" (2000, p. 7)—we are born into a world long since begun, and die in a world unfinished. From a social perspective, aside from the finite corporeality of our lives, notions of origin and end are essentially sense-giving fictions used to buttress structures of power.

122. Arguably, it is the origin of inquiry.

123. The risk is in fact more similar (arguably identical) to Durkheim's concern with bounding the 'social' as a field of inquiry, noted in **chp. 2 § 2.2.2.1 ¶ 92**.

124. In regard to language as tool for studying language versus as a tool for studying society, consider Ruqaiya Hasan (2005) on endotropic versus exotropic theory.

125. The drivers of this phenomenon are often external to the academy—compare the discussion in **chp. 1 § 1.1**, especially **ns 7, 8**.

126. Parallels could be drawn to the study of political participation by ethnographic versus survey methods, as discussed in **chp. 3 § 3.1**. Consider as well how the survey by Nguyen *et al.* (2016) frames language as a social <u>resource</u>.

Also, there is much that could be said about the great range of <u>computational</u> (meaning algorithmic, beyond the statistical) approaches to language these recent years. However, in the context this discussion, we observe simply that they are predominantly information-theoretic or otherwise socially atheoretic. To return briefly to conventional approaches to things political, consider this observation by Jeffrey Friedman (1995):

> The problem of complexity encourages the use of large sets of quantified data ; these can be manipulated statistically … . [Such] practices are legitimate as long as they are checked by theoretical realism. If one aims at both accurate predictions and realistic assumptions, one will not pursue the former by mindlessly privileging quantifiable data over data that gives less accurate predictions but a more accurate grasp of reality. (p. 14)

127. Atheoretical in social terms. Theory is not readily portable across subject models when there is a basic ontological disconnect, as we have shown.

128. The terms 'exogenous' and 'endogenous' are borrowed from econometrics (e.g. Engle, Hendry and Richard, 1983). In broad strokes, an exogenous variable is one that is determined by factors (i.e. other variables) external to the system or model in question. Endogenous thus describes a variable that is determined by the system or model itself. Thus the borrowing is applied to <u>social</u> phenomena, where endogenous denotes that which emerges from the phenomena in question, and where exogenous denotes that which does not. In the study of language in society, comparable (though not synonymous) terms would be 'emic' for endogenous, and 'etic' for exogenous (cf. Pike, 1954, chp. 2).

129. It is worth noting a critique, now nearly 50 years old, of organisational studies for its <u>institutional</u> orientation (i.e. the assumption that organisations are things that exist unto themselves in and through which other things happen). Leonard C. Hawes (1974) countered that view, saying:

> In short, a social collectivity *is* patterned communicative behavior ; communicative behavior does not occur *within* a network of relationships but *is* that network.
>
> If we further assume, along with ethnomethodologists and social interactionists, that social collectivities are not *reflected in* communicative behavior but rather *are* communicative behavior, then we have no alternative but to treat observable communicative behavior as our primary data. (p. 500, emphasis original)

For a discussion of the evolution of this perspective, see Gail T. Fairhurst and Linda Putnam (2004). For yet further evolution see Nick Couldry and Andreas Hepp (2017, chp. 10 § 3). Note well: all studies of communication are studies of organisational communication. That dog will hunt, as we say.

Note that 'collectivity' denotes human sociation generally, without stipulation of extent or character. The term is used here to avoid implications of external reality that more common terms, such as 'group' or 'collective', might suggest.

Also note that Hawes was writing in the *Quarterly Journal of Speech*. In the US academy, the speech tradition is deeply intertwined with the rhetorical tradition and thus communicative approaches to understanding society, organisation, and technology (cf. Craig, 1990 ; see **ns 74, 85**).

130. This could be read to suggest that social context and social collectivity are coextensive terms. Correct. These are two conceptual aspects of the underlying sociomaterial configuration. Neither exists without the other.

131. The term 'lexicogrammar' is often encountered in the study of language and refers to the complex of vocabulary (i.e. lexis) and how vocabulary is structured together (via grammar) to negotiate meaning. Thus the operational step previously described is a lexicogrammatical step.

132. Genre and register are somewhat fuzzy, interrelated terms. While there is a great variety of approaches to them, for our purposes here they can understood in a straightforward manner. Genre refers to modes of language that are associated with certain contexts, such as poems, lectures, recipes, news reports, business letters, work emails, etc. Register refers to manners of language found within genre—consider the differences one might observe in a lecture delivered to first year students, versus a lecture to peers, versus a keynote lecture by a senior scholar. Consider similar differences in work emails. Across fields of study, genre is perhaps the more common term where text is concerned, and register the more common where speech is concerned. For more on these concepts, see John Swales (1990, chp. 3).

★ If the reader is by chance a graduate student, see also "Research articles in English" (ibid., chp. 7) as well as *Academic Writing for Graduate Students* (Swales and Feak, 2012). At the very least, search online for 'creating a research space'. Your time will be well repaid.

133. Such features are commonly called 'function' words, as opposed to 'content' words like nouns and such. To avoid overlapping uses of 'function', the term 'structural' has been used instead. This does change the sense, but the discussion at hand will clarify.

134. Note that organization can be understood as sustained collective action (see **n. 128**).

135. There is an argument to be made that language is the sole subject and mechanism of study across all practices of knowledge (cf. **n. 115**), but this is not the place for that argument. Nevertheless, consider Marshall McLuhan's reflections on language, Henri Bergson, and computerised technology (1964, pp. 79–80), bearing in mind that McLuhan uses the term 'language' in a common, unnecessarily restricted sense. From a more pragmatic, pedagogical perspective, consider the necessity of literacy to science (Norris and Phillips, 2003 ; Halliday and Martin, 1993).

136. This is not to say that language is not already used at scale—survey methodologies study social phenomena through spans of space, and historiographic accounts study social phenomena through spans of time. The author only intends to observe that the use of language as tool for the study of large-scale phenomena is often contested, with exception taken, for example, to the propriety of the scale of application or the propriety of language itself at scale.

137. Such differentials are implicated substantively in all domains of knowledge, but only a few fields take them consciously as the object of study.

138. Related and not necessarily distinct fields are sociocultural linguistics and linguistic anthropology. On the confluence of such work see Mary Bucholtz and Kira Hall (2008).

139. The attempt is not new, of course (e.g. Sedelow, 1967). But, as noted a number of times in this document, computational approaches to language have until fairly recent years faced significant challenges in terms of the availability of data and the availability and capacity of computing resources.

140. Bear in mind that from the mid-twentieth century, linguistics as a discipline (at least within the US academy) was largely dominated by formal, asocial approaches such as those in the mould of Noam Chomsky (e.g. 1957), the Herbert Simon of language (cf. **n. 114**). Such approaches have been perpetuated quite strongly in positive computational approaches to language (e.g. Manning and Schütze, 1999, p. 660 ; Jurafsky and Martin, 2009, chp. 1).

141. The 'state of the art', as it were, is difficult to pinpoint because of the seemingly endless expansion of conferences, journals, and fields that are engaging with computational approaches to language. Some of the methods that purport to engage with meaning in language, regardless of underlying theory, and which have been taken up broadly in the gold rush to 'computational social science', include various forms of semantic modelling (Hofmann, 1999 ; Landauer, Foltz and Laham, 1998 ; Pennington, Socher and Manning, 2014 ; Mikolov *et al.*, 2013), topic modelling (Wilson *et al.*, 2016 ; Blei, Ng and Jordan, 2003), and deep this, and neural that. There is rarely a lick of social theory underpinning the application of these techniques. Even relatively straightforward and seemingly benign lexical approaches (e.g. Tausczik and Pennebaker, 2010) are often rooted in a fundamentally problematic subject model.

142. Note that this an understanding rooted in the rhetorical tradition of communication (see **ns 74, 85, 128**).

143. Or register, or style, or whatever term you prefer (cf. Biber and Conrad, 2019). The distinction is not important for the purposes of this work.

144. This is not a huge remove from the 'community' concepts of discourse (Borg, 2003), practice (Holmes and Meyerhoff, 1999), and speech (Gumperz, 1968). Some might resist the comparison in terms of the shift in mode from speech to text, but that would be a tenuous complaint given the thoroughly multimodal nature of language. It's apples and oranges.

145. The author was obliged to reimagine and implement the methodical component of this work in the span of a year. There is no need to describe the stress and worry of such a situation. The way forward was inspired by a chance encounter with certain works by Adam Kilgarriff (e.g. 2001, 2005) ; consideration of observations on corpus linguistics, the philosophy of science, and social reality made by Tony McEnery at the Lancaster Symposium on Innovation in Corpus Linguistics 2021 ; and reflection on previous training and study in technical communication at North Carolina State University where the presence of genre, and Carolyn Miller (e.g. 1984), is strong. Tony Berber Sardinha and Marcia Veirano Pinto (2019) made the way clear and proved a lifeline in the long last stretch. The author's gratitude is deep.

146. In his work, Biber moves between genre and register (and beyond). As observed in **n. 131**, the concepts are fuzzy and interrelated. Here we will say 'genre', because that is how we have been taught. Regardless of the term used or preferred, the importance is to understand that by 'genre' we denote complexes of linguistic features that correlate with recurrent social contexts and purposes. This is fitting with Biber and Susan Conrad's (2019, chp. 1) position that genre and register should be understood not as indicating different modes or varieties of language, but rather as indicating different perspectives for linguistic analysis. Yet even that position is beyond our purpose, so we will stick with 'genre'.

147. As stated, this is an idealisation. Recognised genres (such as news articles, meeting minutes, business contracts, and so on) can be understood thankfully to have already done the work of contextual–functional analysis for us. That is genre, in essence—it is the observable result of social 'data reduction', all packaged up in a easily recognisable form.

148. ★ For any reader interested in applying multi-dimensional analysis to the work, Berber Sardinha and Veirano Pinto (2019) is intended to be a complete guide. That work is unusual in this regard amongst other methodological works, and the efforts of the contributors and editors are to be commended.

149. Biber's original tagger is not publicly available. Andrea Nini's (2019, p. 71) recreation uses the Stanford tagger (Toutanova *et al.*, 2003) for initial part-of-speech tagging, then applies Biber's rules (1988, appendix II) for identifying more complex features. More on this in the following chapter.

150. On the flip side, such taggers seem to be consistent and useful for speakers (accounts, whatever) that engage in reportage-style speech, which are not uncommon in the author's experience. In the context of this work, the hope was that, for non-standard language, conventional taggers would be randomly inaccurate and thus generate only background noise, statistically speaking. Non-random inaccuracies are bias, so that would be the problem. In a like manner, consistency with relatively constrained modes of speech could potential be useful for isolating automated accounts (which are effectively constrained in compositional terms).

151. In this regard, following the method detailed in the following chapter, the author prepared datasets using taggers developed specifically for application to Twitter—the GATE Twitter tagger (Derczynski and Ritter *et al.*, 2013) and the ARK TweetNLP tagger (Gimpel *et al.*, 2011 ; Owoputi *et al.*, 2013). Unfortunately, time constraints did not allow for the analysis of these datasets. The author's initial observations, however, are that the GATE tagger was simply a pain, and seems oriented primarily to be a business product, and that the ARK tagger—for which the author had high hopes in that it incorporates training based on Brown clustering (Brown *et al.*, 1992) of a significant quantity of messages—was nevertheless consistently inconsistent with certain features having evident social import. The ARK tagger did have utility for lexical analysis, however, as addressed in **chp. 5 § 5.1.3.2**.

152. Consider the degree to which taggers have been trained on newswire, newspaper, or similar data, and consider who may have been writing that source material.

153. For such reasons, some fields are eager to pursue neural approaches (e.g. Goldberg, 2017). If they so please, then let them, but such approaches are poster children for socially atheoretic approaches to social phenomena via language.

154. The author considers the statistical analysis of structural features to be a puzzle and challenge similar to that faced in relation to the cosmic background radiation. It was predicted in advance, though when first observed it was but noise. Only in studying the patterns in that noise did we learn what it encodes about the structure and history of our universe. It might be overly grand thinking, but nevertheless the author enjoys the comparison. Language is the human universe, after all.

155. Structural perspectives grounded in a social subject model are a valid analytical choice in terms of <u>level</u> of analysis. However, if lacking a social subject model, such perspectives are essentially normative statements of expectation mapped onto circumstantial data.

156. In contrast to the previous chapter, which was intentionally explicit in gathering together topics and concepts across a number of fields, this chapter assumes a threshold familiarity with the services, technologies, and techniques underlying the method described (comprising topics related to networks, databases, scripting/programming, statistics, etc.). For those who continue despite and regardless, the footnotes offer some suggestions for building familiarity—look for the black star ★ that marks comments targeted at researchers interested in pursuing computational approaches. The goal and burden are your own, so may you work well. Also, do not hesitate to contact the author.

157. Certain approaches in sociolinguistics (**chp. 4 § 4.1.2.2**) work at the intersection of exogenous variables of social structure and endogenous variables of linguistic variation. However, in the main it seems that exogenous variables (e.g. 'status' or 'class' or what have you) serve an ordering function for such approaches, rather than serving as a target of analysis themselves. Thus they should not be considered 'cross-analytical'.

158. It is common to do this because we are socialised to it, not because it makes the most analytical sense from an social perspective (cf. **chp. 3 § 3.1**).

159. Arguably the most foundational work in this regard has been done in philology (cf. Lass, 1997, chp. 1) , but that is beyond the scope of this discussion.

160. While the MDA approach is multi-variate, it is multi-dimensional in that it is concerned with identifying recurrent groups of functionally (thus contextually) related features in text. These groups of features are the eponymous dimensions. Identifying such groups is done statistically, which will be addressed here. For a discussion of their interpretation, from which this work departs, see Biber (1988, part III).

161. Biber (1988, p. 170) distinguishes between genre and text type, with genre determined by purpose of use (e.g. a dissertation, a software manual) and type determined by linguistic form (e.g. narrative, expository). That distinction is not useful to this work ; genre is understood to combine form and function, and type is used in a general sense.

162. ★ Berber Sardinha and Veirano Pinto (2019), cited just above, is intended by its editors to be sufficiently detailed so as to enable others researchers to undertake MDA work. Furthermore, as demonstrated in part 3 of that work, MDA is adaptable to a range of research, including on social topics. The author highly recommends this work to any reader interested in computational approaches to language in society, and is grateful the editors and contributors (which include Douglas Biber himself).

163. Biber (1988) sourced its sample of texts primarily from the Lancaster–Oslo–Bergen Corpus of British English and the London—Lund Corpus of Spoken English. These corpora are structured by genre, and thus each text in the sample corpus has an associated genre (pp. 65–67).

Note that 'sufficiently broad sample' is a phrasing of convenience. Biber actually says to "review previous research to insure that all important situational distinctions are included in the text sample". The point was that genres (and registers) are understood to emerge from, and likewise to foster, recurrent social contexts (i.e. situational distinctions). The further point being that to analyse the potential breadth of dimensional variation requires the potential breadth of context. Biber supplied that context by sampling and supplementing the above corpora ; this work supplies that context in the manner described in **chp. 4 § 4.1.1.2.2**.

164. For this work to take a dimensional approach would be flowing Lethe-wards into exogenous interpretation.

165. The words of this feature set were identified by frequency and pervasiveness, as will be explained. Thus, while this set does comprise those words that would be generally recognised as function words (membership of that set is debated), it also includes a wide range of everything else that walked and talked like a duck. We say 'function words' simply for convenience. While from a grammatical perspective a characteristic of function words is that they form a 'closed class' in terms of limited and stable membership, from a lexicogrammatical perspective on public discourse the notion of 'closure' is perhaps relative.

Note that this work hinges on a certain degree of closure within social groups, not necessarily across society.

166. The London–Lund Corpus of Spoken English naturally contains less standard English in terms of grammar, but nevertheless it is structured around relatively standard contexts, such as phone calls, broadcasts, debates, and speeches (cf. Biber, 1988, pp. 66–67).

167. A better term here might be 'latent' in that, while we expect coherent social groupings to share patterns of variation, we expect the variation to arise from the contexts themselves not necessarily from a given grouping being aware of itself. However, we do not use 'latent' as the groupings in reality might be quite manifest—nevertheless we are not in the position to know given this dataset.

168. Note that this could be understood as a step in the direction of dimensions. It is however an a posteriori step, rather than a priori as cautioned against in **n. 163**. There is nothing inherently wrong with dimensional interpretations—they are essentially typological in nature, and are thus central to knowledge work. However, we must be on guard against assuming their presence in phenomena, and thus dancing too close to reification.

Regardless, while there may well be some dimensional aspects to social variation (e.g. I say potato, you say potahto), the fundamental impulses of social differentiation would unerringly sniff these out and effectively render them exogenous. Consider the notion of recursiveness, whereby categories of identity are turned in on themselves, at both large scale (Irvine and Gal, 2000) and small (Eckert, 2008).

169. This paragraph refers to 'municipalities' as a term of convenience. In reality, this method engaged with a defined hierarchy of political and administrative divisions within the site of research. The term 'municipality' is used here as a short-hand in hope of implicitly conveying the idea of such divisions with a single term. As the discussion proceeds, we will use the term 'place', in the spirit and sense of the discussion in **chp. 4 § 4.1.1.2.2**.

170. ★ API stands for application programming interface. Whereas user interfaces—such as the 'desktop' of an operating system, a web browser window, or telephone keypad (be it touchscreen, push button, rotary, etc.)—enable a system to be connected to a person, an API is a system interface serving to connect one system to another. The user interfaces just described all rely on APIs working in the background to translate human instructions into calls that can be interpreted by the target system. While the complexity of information and communication technologies necessitates APIs in bewildering variety, the use of the term in this work has a limited sense, indicating the public interface through which a service makes its data available. In the case of Twitter, the API belongs to an architecture underlying the 'World Wide Web' called Representational State Transfer (REST). For a description of that architecture, see Roy Fielding and Richard Taylor (2002).

171. The API v2 was announced in late 2020, and has since been introduced. However, this method did not use that API, so bear in mind that the following steps must be viewed historically. Also bear in mind that all descriptions of the affordances of the API are in terms of what was available at the level of free access at the time of use. In addition, there were changes to the API in 2018 to make it compliant with GDPR ; those changes arrived early enough in the research project that they did not significantly impact data collection. In regard to APIs for research purposes, see Deen Freelon (2018).

172. ★ Platforms often have a ‘rate limit’ for their APIs, meaning the frequency of queries or quantity of data returned cannot exceed a given threshold with a given period of time. Such thresholds are arbitrary and changing, being a function of monetisation. Exceeding these arbitrary thresholds can result in throttling of throughput or a block to access, so stay within stated limits and plan ahead to adapt to unforeseen changes in terms of use.

173. In practice, the amount of posts returned is often less than 3,200, as posts can be deleted or otherwise removed, or a given user might not have that many posts in total. The latter seems to have been the more common case in the secondary collection.

174. At the time of collection, the use of DMI-TCAT was a sound choice of tool. However, having used that tool extensively for a number of years, the author counsels against its use for anything beyond exploratory work (cf. **n. 178**).

175. It is fairly common in political science literature to see discussion of selection bias (e.g. Geddes, 1990). Often such worries will be addressed under the rubric of ‘selecting on the dependent variable’. In a general sense this is unavoidable—bounding a field of study is such a move writ large. However, for those who privilege questions of inference and causality, selection effects are a problem that must be considered thoroughly. Similar problems can manifest in how we pursue large-scale social inquiry using programmatic means, as Zeynep Tufecki (2014a) observes in relation to studies that fixate on hashtags and similar affordances of social media—what Tufecki calls a potentially misleading ‘epidemiological’ approach to social networks. Consider how a focus on similarly salient yet relatively low-frequency lexical items in a political context, such as the names of candidates (e.g. Barberá and Rivero, 2015), might skew a dataset. Such an approach to selection lowers our guard to discourse that is highly coloured by potentially bad-faith actors, both human and automated (i.e. ‘bots’). In that regard, consider that the field of authorship attribution acknowledges such high-salience, low-frequency features as prime targets for imitators or forgers (Love, 2002, pp. 185–193 ; cf. Kestemont, 2014).

176. Posts are geocoded if the user permits. The location is taken from the device (likely a phone) at the time that is used to post the message. Thus location is associated with the post, not with the user, thus making geocoding useless for our purposes. Furthermore, geocoding is rare. In the primary collection, geocoding was present in less than 1% of tweets.

During pilot testing before primary collection began, an approach to rough bounding was attempted using time zone metadata. Michigan lies within UTC -4 and -5, thus filtering by time zone would have provided a significant reduction of junk in the primary collection (in that the streaming endpoint sample is global). However, that field was made private across all access levels for reasons of GDPR compliance ; see “May 23rd, 2018” at https://developer.twitter.com/en/updates/changelog.

177. Note also that v1.1 stream filtering does not support compound terms. According to the documentation, querying a compound term such as `great lakes state` via v1.1 is equivalent to `great OR lakes OR state`. DMI-TCAT does support compound queries, but due to the behaviour of the underlying API it is obliged to collect all possible matches then join the resultant data tables, discarding those matches not fitting the query. In practical terms, such an approach would have been prohibitive in terms of the demands on computational resources, storage, and bandwidth. Thus only single terms were used. For current information on stream filtering by keyword in v1.1, see https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters#track. The URL will likely change, in which case one may consult a cached copy in the Internet Archive (https://web.archive.org).

178. In DMI-TCAT terms, these were separate 'query bins'. However, the author regrets the use of TCAT and would discourage others from using it, and thus the discussion is not framed in terms of the tool. It is the author's considered opinion that methods generally should never be framed in terms of a specific tool, as no tool lasts. Learn to fish, teach to fish.

179. The DMI-TCAT backend at that time was MySQL. These data tables, represented in the front end as 'query bins' as mentioned in the previous note, were actually broken up across a variety of tables according to the TCAT architecture. For a given query, TCAT produces a range of tables to suit its in-built analytical functions. This proved to be a constant source of difficulty and frustration for the author. Full datasets had to be assembled (both manually and programmatically) outside of TCAT. For this reason alone, although among others, TCAT is not recommended for crucial work. Note that this is no fault of the backend—interacting directly with the data via SQL eventually proved invaluable to this work.

180. In reality, collection proceeded before and after the those dates. However, data timestamped before 2018-07-01 00:00:00 UTC and after 23:59:59 30-11-2018 was discarded. Also, while the collection period was 153 days in total, the collection processes was not constant. There was a dedicated machine performing the collection, but there was occasional downtime for necessary local maintenance and archiving of collected data. There was of course also unexpected downtime due to accident, error, and system-wide maintenance.

181. For technical reasons related to the DMI-TCAT backend architecture and to the author's approach to archiving and safeguarding data, a full count of tweets in the primary collection was not undertaken, nor was it needed. The count of 300–400 million is an extrapolation from 154 GB of data total, based on full counts of subsets of the primary collection.

Also note that these figures include the collection of a second geographic keyword: `ohio`. As noted previously, that state was originally intended as second case. The political keywords, having no geographic association in themselves, were unaffected.

182. See https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet.

183. Note that these are the attribute labels used by DMI-TCAT. For some attributes they differ from the labels used in Twitter objects. Possible label conflicts (e.g. Twitter uses the label id for both tweet and user objects) are avoided by the data structure of Twitter objects, but that is not the case for TCAT hence the modification of certain labels. This discussion will use the TCAT labels, as those are the labels that were used throughout the method.

184. ★ Twitter identifiers for tweets and users are often very long. Documentation for v1.1 states that it is 'safe' to store this identifier as a signed (i.e. positive or negative) 64-bit integer. Integers of that size can be up to 24 digits in length. Such numbers can cause unexpected havoc in your work. For example, if you import a sample of your data into Excel for inspection, the default behaviour of that application is to convert numbers over a certain length to scientific notation. Thus if you have a tweet with the identifier `9223372036854775807` (which, by the way, is the largest positive value of a signed 64-bit integer), then Excel will helpfully convert that identifier to `9.22337E+18`. It saves screen space and memory. But if you go back to standard notation, your identifier is now `9223372036854770000`. Oh no! That is a different tweet, and Excel will assign that same identifier to 9,999 other tweets if given the chance. Excel is not alone in this behaviour, so be on guard. For this general reason the v1.1 documentation suggests using the field `id_str`, which is a string representation of the identifier (this is not always a workable solution in large-scale work, as string representations of digits require more memory than binary representations ; DMI-TCAT uses binary representation of long integers). Also, avoid Excel for large-scale data work whenever possible ; a good way to begin weaning yourself off of spreadsheets is to experiment with R in the environment of RStudio, which is freely available in an open-source version. The investment of time will be rewarded. See https://www.rstudio.com/products/rstudio/.

185. The documentation notes that the data in this field is not necessarily a location (and as it is optional the data can be `NULL`), even though it is labelled as such in data objects and presented as such in user interfaces. Thus users sometimes have a bit of fun with this field, perhaps putting something like 'In the wind' or 'State of Grace'. However, users generally adhere to the intended use, supplying a description of location with varying degrees of formality and exactitude. This method relies on users taking the formal approach, such as 'Grand Rapids, MI'.

186. In terms of Twitter metadata, all users and all tweets have unique identifiers. The user identifier is distinct from the username (or 'handle'). Both nevertheless will point to the same user. For example, the handle of the President of Ireland is @PresidentIRL, but the user identifier is 569892832. You can test their equivalence by comparing https://twitter.com/PresidentIRL and https://twitter.com/i/user/569892832—both are resolved to the same resource. However, the former is chosen by the user and can be changed, whereas the latter is assigned at account creation and is fixed. This method relies on the latter.

187. In the course of this work, the author primarily used MySQL, later moving to MariaDB. For an extended period (prior to the overhaul of method noted in the first section of this chapter) there was also use of MonetDB, which is a column-oriented database (see Idreos *et al.*, 2012). That system is shockingly fast for the types of data one might encounter by way of a social media APIs. However, it also has some of the poorest and most confounding documentation that the author has ever seen (and the author has worked with both IBM and the United Nations). Use it for your work only if absolutely necessary. However, if that is in fact the case, perhaps first reconsider your approach. (NB: at the time the author made use of it, the methods he was seeking to apply had been taking days at a time to process, and speed was a priority ; the approach was eventually reconsidered.)

★ Effects of scale in computation can significantly impact your schedule of work—learn some technique for estimating, at least roughly, the time that will be required for your operations. Thereafter judge accordingly.

188. The author acknowledges that there will likely be some generational and social skewing in regard to the technique of 'nominal localisation' that appears throughout this method. It relies on an ingrained practice, one often taught in schools, of how to 'properly' address an envelope. Such practice predates the widespread use of email, and it could well be that the practice is fading. However, the [Place], [State] paradigm persists outside of email, and Americans still seem to reproduce that manner of place reference. In any case, if there is generational and social skewing incurred by the technique, odds are good that it would be in the opposite direction of generational and social skewing observed in social media generally. So, the author hoped for a neutral effect, but failing that, a corrective effect. In any case, we cannot afford location information in the United States, and it would be prohibited in the European Union, so nominal localisation it is! (All jokes aside, the technique as will be explained is non-invasive and transparent, which is a good sight better that unethical precision.)

189. ★ If the reader is unfamiliar with structured query language, yet is interested in large-scale social research of any kind, the author strongly recommends that they acquaint themselves with relational databases and SQL. *The Language of SQL* by Larry Rockoff (2017) offers a good start ; avoid reading anything too dated, because databases systems and their language implementations evolve. If the reader is unfamiliar with regular expressions and is interested in any sort of programmatic manipulation of text—be it in files, scripts, or in research data—then read whatever you can find on regular expressions (commonly REGEX). If you will, start with *Mastering Regular Expressions* by Jeffrey Friedl (2006).

190. The regular expression used was thus quite simple: '`, MI|, Michigan`'. The end-of-line anchor `$` was not used in order to avoid false negatives, which pilot testing showed would have been numerous. It is not only text that is messy in social media—metadata is a rarely seen quagmire. Be on guard.

191. ★ The author also hard-coded a filter into the DMI-TCAT scripts that control queries. With that tool, certain modifications to the PHP configuration files are necessary, such as entering access tokens and specifying other local system parameters, so changes are routine. The author decided to save himself some work in the future by restricting query results to those tweets (i.e. the content) marked as English by Twitter, as it was a simple change in the query passed to the API by TCAT. This worked like a charm, until a system update was posted. It seems that the author's alteration caused the update to fail (because of an out-of-sync file), and that failure left the software in a irreparable state. That is, it had to be wiped and reinstalled. The author cursed TCAT, and continued collecting tweets in all languages. This was a good thing, as only later did the author realise that Twitter's automatic language tagger is horribly bad at its job—for example, "Hahahahaha" tagged as Tagalog, "Merry Christmas" tagged as Estonian, and "Touchdown Georgia!" tagged as Indonesian (those are whole tweets from the dataset, by the way). Tweets would be filtered by language during corpus assembly (§ **5.1.2.2**).

192. The count is 23 as weeks were considered to start on Monday. Counting weeks on Sunday would yield 22 splits. There are possible selection effects in this step resulting from the use of calendar weeks, and from the choice of definition of calendar week, but the author felt that such effects could be safely ignored.

193. For example, there was large chunk of retweets of one account that, at one point, had a follower count of 3,266,352. The `from_user_id` resolved to `@cavs`, which is the account of the Cleveland Cavaliers (a professional basketball team in Ohio ; as noted in **n. 180**, the geographic keyword `ohio` was present in the primary collection).

194. ★ The use of the RT/OT ratio to encourage a focus on discourse and to privilege individuals or business or bots is certainly not foolproof. It is in fact rather crude. But it is exceptionally simple to calculate and to implement. And, as mentioned, close inspection of the data showed the measure to have good 'instincts' for differentiating between types of accounts. While the author was working on a hunch rather than from a body of evidence, the apparent utility to this work of such a basic step shows the potential value of incorporating social thinking into the design decisions of your computational approaches. Let the social subject model be a guide.

195. ★ As Twitter has rate limits for API queries, the list of accounts could not be queried all at once. The list was split into 12 blocks. The TCAT query script was in PHP, and requested accounts had to be inserted into the script manually. So, for each block a script was prepared. Using Linux (which is practically required for certain types of computational work), a one-line bash script was used to run these scripts four at a time (using `screen` for concurrence) in a three waves. That automated process took three days to complete. At the time it was possible to have multiple access tokens for v1.1 ; four scripts were run concurrently as the author had four tokens. API v2 does not allow for this, so such a process would take longer.

196. The formal definition of 'place', according to the US Census Bureau, is "A concentration of population either legally bounded as an incorporated place, or identified by the Census Bureau as a [Census Designated Place]. Incorporated places have political/statistical descriptions of borough …, city, town … or village" (1994, p. G-38).

Note that this work also included other populated areas that do not fall under this rubric. In the case of Michigan, these areas are 'townships'. The entire state is divided into townships, except for areas within cities, and there are some 1,200 of them. That fact significantly complicated this stage of work. However, it was noticed in the secondary collection that many users specified locations that were determined to be townships. It was decided that an a priori exclusion of this feature type would significantly bias the dataset towards the urban population. Thus townships were considered as places.

197. There is good reason for having different GEOID systems. For example, this work relies on data from the US Census Bureau and from the US Geological Survey. The former uses Federal Information Processing Series (FIPS) codes, which allow a hierarchical representation of political and statistical entities. The latter uses Geographical Names Information System (GNIS) codes, which are not nested as they are strictly geographic, and which also identify features without political or statistical standing, such as airports, beaches, churches, hospitals, schools, rivers, and so on. But, while GEOIDs are not standardised, official places are, and official places are also geographic. Thus, every FIPS code will have a matching GNIS code (although the reverse is not true).

198. The American Community Survey (ACS) is an annual survey of social, economic, housing, and demographic data. The survey uses a sample size of around 3.5 million addresses. Unlike the decennial census, which calculates its data for a point in time, the ACS calculates its data for a span of time. The 1-year ACS estimates are made available for areas with a population of at least 65,000. The 5-year estimates, however, have a resolution down to Census 'block groups', which represent areas having a population of 600–3,000. For more information on the ACS specifically in the context of geography see US Census Bureau (2020).

199. ★ APIs are not some scary, arcane thing. A query to the US Census, for example, looks much like a regular URL. The following query https://api.census.gov/data/2018/acs/acs5?get=B01001_001E&for=state:26 contains the URL of the API itself (as it is a resource), specifying the ACS year and type. The element `get=B01001_001E` requests table code B01001_001E which is the field for 'total population', and `for=state:26` requests that data specifically for the state coded 26, which is Michigan. You can try this yourself in a web browser, however to use the API programmatically (i.e. via scripting or other automation) requires an access token. For more on the ACS API, see https://www.census.gov/programs-surveys/acs/data/data-via-api.html.

200. The ACS data 'table codes' requested through the API were in terms of 'percentage population' to allow comparison across places. (The only exception to this was 'total population' ; the comparable measure of population density is calculated in the next stage). Also, the ACS data can be fine grained. The age distribution data was retrieved across 11 table codes, with 10 bracketing ages below 65 and the 11th bracketing those 65 and older. Income was retrieved across 10 table codes, each a progressively larger bracket with the 10th bracketing incomes of $200,000 p.a. and up. Education was bracketed across 7 table codes, roughly equating with levels of completion. The ethnicity and origin indicators were included to aid the interpretation of results. Those indicators are very general, with ethnicity represented simply by the percentage population indicating white or black (the two main ethnic groups in Michigan, per Census categories), and origin indicated by percentage population born in Michigan, and roughly proxied by percentage population (in terms of households) speaking only English at home.

Note that the ACS data tables are expansive, with more than 20,000 variables in the most detailed tables (of which this method uses 32). Many are variations on a theme, and many are derived from combinations of others. While gigantic and messy, such data—if protected—is a heritage greater than any museum. For the curious, a hyperlinked and annotated listing is available at https://api.census.gov/data/2018/acs/ acs5/variables.html.

201. These files are pipe-delimited, so the data can be imported into databases or spreadsheets with relative ease ; see https://www.usgs.gov/u.s.-board-on-geographic-names/download-gnis-data.

202. This datafile had 4,889 rows. It was a datafile intended specifically for translation between GNIS and FIPS codes, thus it is a subset of all GNIS features. Before the author was aware that such a datafile existed, work started from the full list of GNIS features (more than 62,000 in Michigan) which was filtered according to relevant feature classes ('census', 'civil', and 'place'), obtaining a feature count of 5,683. Time was not taken to investigate the disparity (likely to be found within the 'place' category). In either case, the end result of the localisation process in the next stage would have been much the same.

203. This basic geographic data was later supplemented (once the final list of places was known) with data specifying the land area of places, so that population density (a comparable measure) could be calculated using the population totals in the sociodemographic data. The land area data was obtained in a roundabout manner, as it was strangely hard to come by in itself. However, the Census provides certain ACS data prejoined to GIS shapefiles. Such data for places in Michigan was obtained. These were loaded into QGIS and an SQL spatial layer was created. The computed land area data was then subsequently joined with the working dataset with ease. See https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-data.html.

204. ★ In computational spaces, you may often hear language referred to as 'unstructured' data. Nuts. The basic affordance of recorded symbolic language is the manipulation of meaning by the serial, grammatical ordering of individually meaningful units—for example these sentences, or 2 + 2 = 4, or ··· - - - ···, or even ♩♫♩♩. However, the combinatorial infinitude of language (you can always add on one more bit, so it is limitless) makes it always a woolly mess. So semi-structured, yes. But unstructured? Never. See also Adam Kilgarriff (2005).

205. In computational approaches to text, cleaning often falls under the broader rubric of 'pre-processing'. While that term certainly has a computational air, it does not tell us much. The term 'cleaning' is preferred here, as the process focused specifically on tossing out the garbage and bringing order to the data—and to the text itself.

206. This is the point at which the discursive data begins to be shaped into a corpus.

207. ★ Note that this is a logical, rather than real, presentation of the process (as is the entire chapter). That is, the actual doing of this work was not performed exactly as described or necessarily in the steps given. The real process was relatively messy (as the author had few ready guides for the work), requiring much experimentation, repetition, and rethinking. The convenient fictions of logical presentation are for reasons of clarity. The subsets of sociodemographic and geographic data are a good example. As noted, the key to interlinking all of the data is the process of nominal localisation, which assigns a unique location to each user. That assignment is by codified name and a FIPS code. That data is thus nominally localised, as is the sociodemographic data. The geographic data, while also nominally localised by codified name and GNIS code, is also geolocalised by longitude and latitude. Recall that in the data collection stage, a FIPS–GNIS translation table was already derived from the collected data. In fact, all data linking was accomplished in the localisation step as a result of the procedure used (see also **n. 188** re databases and SQL). However, the actual assembly of the data into a single resource (e.g. a file) was performed later ; as that is an important conceptual step, that is the logic of the process presented here.

208. Recall that the secondary collection comprised 17 million tweets associated with some 20,000 users across some 3,400 locations, whereas the enriched corpus comprises 2.6 million tweets associated with 5,889 users across 417 locations.

★ Working with large-scale data means becoming comfortable with throwing away data of dubious quality. This is especially the case with textual data, and unavoidable with the 'live' sort of text found in social media. While it is possible to work closely with data to verify it, correct it, or inspect it in some way, one reaches a point where it is no longer reasonable to do this manually. Such manipulations can perhaps be done programmatically, but that move has its own challenges (cf. Find/Replace All). In working at scale, you must remember that individual data points lose value—the worth is in the overall picture—so do not hesitate to throw away troublesome or worrisome data. It is simply neither possible nor desirable to pore through millions of rows of data. (Note how this bit of advice parallels the 'fungible individual' subject model that is argued against in this work. Truly social research at scale is staring into the abyss, so let us find new ways to do it.)

209. ★ For example, the use of '`, MI`' as an acceptable match was perhaps a poor choice, given that this is a high-frequency word in certain languages and Twitter's automatic language taggers are very poor (except where a given language uses a specific alphabet). Also, the author did not bear in mind that `location` metadata is readily changeable, and thus not stable for a given user. That fact was incorporated into the localisation process, but was not considered during data collection (inspection of the secondary collection would reveal the variability of the `location` metadata).

210. In that the third component matches 'natural' language elements of the first component to 'codified' language elements of the second component, the nominal localisation process is, in a sense, a microcosm of the overall corpus preparation stage of the method as the process has its own needs for data cleaning and data linking. Initial cleaning, such as identifying the determinate locations (more below) in the second component, was performed manually working directly in the database using SQL. The results of the initial cleaning were stored in separate data tables, to be used as resources for this process. Subsequent cleaning and linking was performed with an array of bash and SQL scripts working directly on the database.

Regarding the determinacy of location names: At this point in the work, there were still two cases: Michigan and Ohio. The second component comprised nearly 14,000 codified locations across both states. Codified location names were considered determinate if they appeared only once (and thus in only state), semi-determinate if they appeared more than once total but only once in a given state, and indeterminate if they appeared more than once in both states. Michigan had 2,788 determinate locations.

211. ★ Such filtering was accomplished easily with a single SQL statement.

212. At a later point in corpus preparation, it was decided to remove accounts stating more than one location, reasoning that such might represent intra- and inter-state transplants, long-distance commuters, retirees and seasonal relocators (i.e. so-called 'Snowbirds' that move south in Winter), or some other type of peripatetic. That exclusion would substantially reduce the total number of users represented in the corpus, which is good. In that the method hinges on tracing commonalities in speech patterns as proxies of social contexts, a significant presence of extra-contextual (i.e. itinerant) speakers would have diluted those commonalities, thus impairing the method.

213. ★ This is not straightforward. Punctuation can be integral to proper names (e.g. O'Brien) and digits can appear in common informal presentations of local terms (e.g. `A2` for Ann Arbor, Michigan). Also, inspect your data closely for characters that you likely could not expect. Keep the multi-modality of text in mind. For example, the character 〽, which comes from Japanese musical notation, has been taken up by many Michiganders to present 〽ichigan (the character presents in yellow on Twitter ; the main element of the University of Michigan logo is a yellow block M). Thus always check through your data for anomalies before you begin cleaning, but also be prepared to expand and redo your processes as the unexpected will surely arise.

214. Bear in mind, once again, that this figure reflects a point where the data users and places were in both Michigan and Ohio. While the inclusion of a second state substantially complicated this process, the eventual removal of that second state had no negative effect on the localisation of accounts within the site of research, beyond increasing the number of semi-determinate (and thus excluded) place names. As the localisation process was labour intensive, it was not repeated after the removal of the second case.

215. ★ As noted in **n. 206**, full geolocalisation of users is already accomplished in this procedure, by way of the FIPS–GNIS translation table derived during data collection. The procedure is nevertheless termed 'nominal' because it hinges on statements of place.

216. In fact no multimedia content is embedded in Twitter objects. Those objects are purely character based (i.e. textual), with non-textual content represented by links to such resources. The various Twitter user interfaces parse these links in most cases before displaying posts, thus the user sees multimedia content immediately (and failing that, the relevant links).

217. ★ The textual creativity in in social media is a delight. However, it can also be a constant challenge for computational work. For example, there was one point in this work where a certain computational process kept failing. The problem was partly attributed a single tweet. The process failed on one token that looked like this in the terminal:
̬̌̆̑̈d̛̟̈̈̑̐̆̌̍̉͂i̦̮̭̼̍̓̈́n̨̮̈̆̌̆́ő̱̭̼̜̃̆ş̰̣̭̬̂̃̉ā̱̰͕̲̃̈̊ṵ̧̦̻̍̍̋̂̆̈̈r̳̃̈̀̃̐̆

For record-keeping, the author pasted this strange glob into a word processor (which, unlike the terminal, was Unicode-compliant). The result was this: d̛̟̈̈̑̐̆̌̍̉͂ḭ̦̮̼̍̓̈́n̨̮̈̆̌̆́ő̱̭̼̜̃̆ş̰̣̭̬̂̃̉ā̱̰͕̲̃̈̊ṵ̧̦̻̍̍̋̂̆̈̈r̳̃̈̀̃̐̆, that is, the word 'dinosaur'. The word is clearly present in the string above, only each letter is festooned with various diacritics. If you paste the expanded string above into the search box of Twitter, it will be read as 'dinosaur' and will return appropriate results. A database or spreadsheet, which will not see joined diacritics but rather separate characters, will balk.

218. ★ This is a difficult but crucial process. Control codes, especially line breaks, will wreak havoc in databases and spreadsheets. Whitespace and joiners are also persistent gremlins. Note that joiners are common in emojis—many emojis are actually multiple emojis joined together. Such compounds are called an 'emoji modifier sequence' and are part of the Unicode Technical Standard (see https://www.unicode.org/reports/tr51/tr51-21.html § 1.4.4). Users will not see the individual components, but database and scripts certainly will.

Anyone trying corpus-based approaches to social media text must plan for the ubiquity of the 'glue' that sticks emojis together—the Zero Width Joiner, code point U+200D. It is non-printing, meaning in most cases it is invisible to the eye (though not to your software). Embarrassingly, it took the author several years to grasp that much of the 'noise' in the discursive data (which was rarely viewed outside of databases, given its size) was in fact atomised emojis. Realisation only came when procedures began to fail, producing error messages such as `Untokenizable: (U+200D, decimal: 8205)`.

Also, it was observed in the dataset that many unexpected control codes and whitespace types frequently appeared. The author's suspicion is that such characters sneak in by way of desktop users pasting content into the Twitter interface. For this reason, all tweets with control codes present were removed, except where carriage return and line feed occurred separately (these codes as a cluster mark line breaks in Windows files).

219. This single procedure removed more than a third of the dataset.

220. ★ This is rather tricky in practice, but essentially the procedure relied on the fact that English text is represented with a relatively small range of characters compared to possible range of characters available in the UTF-8 encoding (i.e. Unicode, the dominant character encoding on the Web). For example, the first c.2,000 code points of Unicode can encode most Latin scripts. ASCII, your basic alphanumeric character set with some punctuation, is represented by the first 128 code points (the first 32 being the pesky control codes). In short, you can use certain regular expressions to isolate ASCII text.

221. ★ Removal by metadata, as described in the previous section, is insufficient given the way that Twitter objects (or rather Twitter users) work. It is quite possible for retweets (judged by their content) to slip into the stream with a `NULL` retweet attribute. There might be a simple explanation, and most likely the author made a basic error somewhere. The time was not taken to determine the case—all tweets suspected of being retweets were deleted.

222. Extended quotes were defined as more than three consecutive words enclosed by quotes (single or double). While it was considered to slug such quotes (see § **5.1.3.2**), for simplicity's sake the tweets were removed. These had comprised slightly more than 3% of the corpus prior to removal. Not that extended quotes were removed as they were assumed to mark a significant departure from normal speech patterns, regardless of their specific purpose.

223. ★ Note that the corpus was assembled and cleaned a number of times in the course of this work. In part this has to do with the author learning while doing. There were moments of realisation that previous choices (which could not be rolled back, and perhaps were weeks distant) were misjudged or poorly executed. But also in part this has to do with technical failures, such as database corruption (which happened) or the operating system updating to a new release, and in so doing deleting the codebase and databases (which happened). Make backups frequently, and archive your work locally on internal and external media, and remotely (preferably on more than one service if possible). Be prepared for catastrophe. However, if and when the worst happens, remember that you will know more after than you did before, and the work you redo will be better for it.

224. ★ For those of us who seek to adapt computational approaches to properly social inquiry, and especially those of us working in the fields of communication and media studies, it is essential to have an understanding of the history of information retrieval. This is not a trivial or academic exercise—the vast array of information and communication technologies that shape contemporary hybrid societies are built upon information retrieval. The Web is information retrieval. A helpful and brief history of the technology and theory of information retrieval is Sanderson and Croft (2012). Also see Lesk (1996) which addresses the topic from the perspective of library science ; the final sections of that paper discuss possible developments during the first 20 years of this century, and are worth pondering with the benefit of hindsight. And, when you have a spare minute, look up the Lesk algorithm.

225. ★ Segmentation of strings into constituent parts, especially into words in the case of tokenisation, is an essential procedure in computational work with text. However, some approaches will split contractions, such as `isn't` or `can't`, or other compounded word forms into constituent tokens (in the simplest case splitting on the apostrophe). One can see with these two words how the situation is not straightforward. Whatever the case, this work considered contractions and other compounds to carry contextual information—consider the contextual distances between `is not`, `isn't`, and `ain't`. The method here tokenised by splitting on whitespaces only, thus reflecting the implicit tokenisation given by the user in the first instance.

226. Strictly speaking, this method does employ stoplists of a sort, as can be seen in the previous paragraph. However, it does not employ stoplists based on the frequency of words. Similarly, slugging could be understood as an extreme form of lemmatisation. It is not. Slugging reduces variation unimportant to the analysis while preserving information about the functional categories of slugged words, whereas lemmatisation simply reduces variation.

★ Two foundational texts for contemporary work in this regard are Christopher D. Manning and Hinrich Schütze's *Foundations of Statistical Natural Language Processing* (1999, see chp. 4) and Daniel Jurafsky and James H. Martin's *Speech and Language Processing* (2009, see chp. 1 & part 1). For some, these books have near biblical status so it is worth being familiar with them even if they are not your bag of words. Focus on Manning and Schütze, as it is the more focused of the two, and will be of more ready use. Jurafsky and Martin will likely be encountered in its 2nd edition and a 3rd is in the works. It is expansive (and expensive) and will be of little immediate use ; however, it offers a deep dive into computational approaches to text and speech. Have a look at both, always keeping in mind the question of subject model.

227. "This stage" being roughly 2.5 years after the secondary collection, during which time the author had worked almost daily with the dataset, and had learned of its complications.

228. A common `source` value was `<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>`. Compare the display text of that tag to the left-hand tweet in **fig. 1**.

229. Helpfully, all but one official source contained the cluster `twitter.com` ; the exception being Twitter for Mac which had an Apple domain.

230. Furthermore, the structural analysis has no reliance on any specific language or encoding, and non-English discourse would be extremely valuable in terms of contextualisation (cf. **appendix B**).

231. On the one hand, this is a trivial success, in that the localisation procedure was able to match the lowest hanging fruit. One the other hand, it is exactly that, and there were thousands of low-hanging fruits. The author considers that the application of a relatively crude nominal localisation procedure, in combination with some hours of elbow grease in SQL, yielding a decently localised dataset to be a definite success and would recommend the combination to others.

232. The smallest user–document is 102 words and the largest is 95,397 words. Note that the largest user–document is unusual, though not an outlier strictly speaking. That word count comes from 2,105 tweets. With the average tweet being slight more than 45 words in length, and working from a publisher's rule of thumb that average word length in common text is 6 characters (i.e. five letters and a space), that tells us that all of those tweets were at the 280-character limit (the limit was raised from 140 in 2017). On closer inspection, the user–document was comprised of musings and exhortations of a religious nature. The text, however, was not addled or worrisome in any way, and given that the user–document met all criteria for inclusion in the corpus no further action was taken.

There were three 102-word documents. Strangely, two of those were localised to Mount Pleasant, a small city in the centre of the state. Also strangely, Mount Pleasant and its surroundings are predominantly flat.

233. ★ These two perspectives are central in the study of linguistic variation, which is studied both 'diachronically' (i.e. through time) and 'synchronically' (i.e. the same time, think 'synchronous'). The former perspective, for example, might consider variation in individuals through the lifecourse, while the latter might consider variation within an age group at a moment in time. This work would fall into the synchronic category, despite how the sentence above might be interpreted, in that its unit of temporal analysis is a single undifferentiated six-month period—the user–documents collapse all utterances into that unit. Note that these perspectives are not mutually exclusive, often being combined in study designs. Moreover note that these perspectives are central to all inquiry, not just the study of linguistic variation. Time provides a fundamental comparative reference point.

234. As has been noted a number of times, this chapter gives a logical presentation of the method. The actual procedures were often overlapping or intermixed, and work often had to be redone for a range of reasons. Thus exact quantification of the dataset at interim points of the work can be misleading. Nevertheless, the reader is likely curious. Roughly speaking, the reduction of the dataset proceeded as follows: 1) the secondary collection yielded 46 million tweets, filtered by timestamp and retweet status down to 17 million ; 2) filtering by `source` and an initial screening for valid `location` reduced the dataset to just shy of 12 million ; 3) on 29 July 2021, Ohio was removed as a case, immediately reducing the dataset to 4.2 million ; 4) textual cleaning and final screening of location reduced the dataset to 2.6 million, which is the quantity of tweets present in the final enriched corpus. During enrichment there was ad hoc removal of users for reasons of suspect data, reducing the enriched corpus by 58 users (1 per cent) and 22,522 tweets (0.9 per cent) to the final tally of 5,889 user–documents comprising 2,569,762 tweets.

At each stage there was significant loss, which was expected in that large-scale work with discursive social media data is like refining ore—most data will not be relevant to a given method, and thus is waste to be removed.

235. ★ This is how relational databases work, by means of various 'keys'—such as `from_user_id` or `location` metadata, or FIPS and GNIS codes—data spread across a variety of tables and databases can be linked together on command. However, many of us are not used to thinking in that manner, and that is why in **n. 206** it was noted that the enrichment step, wherein everything is combined into one resource, is an important logical step in terms of communicating the work.

236. The description in this section is partly misleading. The actual corpus (i.e. the texts themselves) remains separate from the other data during the remainder of this method. This is unavoidable for two reasons. First, the technical requirements of the tools used to conduct the analytical procedures are such that the corpus has to be maintained in separate text files for each user in the corpus. Thus, as enrichment proceeds, while results of analysis will be incorporated into the enriched 'corpus', strictly speaking it is empty—no corpus at all, only a specific data profile of it, and means to link to it. Second, Twitter terms of service are strict in terms of how content is used and managed, and redistribution of large-scale data is practically out of the question. For that reason, the actual content of the corpus is essentially quarantined.

Nevertheless, the method is described in ideal terms that could be followed to the letter in other circumstances to develop an enriched corpus that could be examined from many analytical perspectives beyond language and would include the source texts themselves for further examination. Although that is not the case here for reasons beyond the author's control, it is in fact what the method describes and intends.

The author considers that this approach is for the best from an ethical perspective. While Twitter posts are public, they nevertheless carry people's thoughts and feelings. Positive consent was not obtained by the author from these thousands of real individuals, and their assumed agreement to Twitter's terms of use has no bearing on the issue. Once this work is completed, all instances of the corpus and source data will be purged. Measurements and derived data will be preserved, but user identifiers will be hashed to prevent traceability on Twitter while preserving data linkages.

NB: The author did assemble something similar to a unitary enriched corpus to serve as a reference during final stages of work. That resource is a single database with only three tables, one for users, one for documents (i.e. the texts), and one for locations. These data were stored in three tables to simplify various query operations involving joins. The resource could not be used for procedures, however, for the technical reasons mentioned. For the actual procedures in the final stages, the author used a unitary enriched 'empty' corpus in R, which was generated using the reference database. The enriched empty corpus nonetheless proved invaluable during final analysis.

237. ★ In computational terms, assigning parts of speech to the individual words of an unmarked text is called 'tagging'. Computer programmes that automatically tag text for parts of speech (often through a combination of dictionary look-up and algorithm) are thus called 'taggers'.

238. ★ Douglas Biber is not being stingy. The algorithm was written in PL/I (1988, p. 211). That is Programming Language One, which was developed by IBM in the 1960s and is still in use. However, PL/I was designed for enterprise-level hardware (i.e. mainframes and such). It is a language of the old school, like COBOL and Fortran. Best to leave it be (cf. http://cs.ecs.baylor.edu/~maurer/SieveE/pl1.htm). However, Nini's replication (to be discussed) is written in Perl, which you should investigate if interested in text processing.

239. ★ The counts are normalised to 100 words, rather than the 1,000 used by Biber. The normalisation baseline will make no difference in most cases, so long as it is strictly adhered to throughout across procedures. Note that normalised counts in MAT are rounded to 2 decimal places

The application runs in Windows, which will be an advantage for most. For those running pipelines in other operating systems, it can be a complication but it is not insuperable. [Note, however, that the author attempted to run MAT using Wine on an Ubuntu system (21.04 Hirsute) ; whatever the details of the case, the end result was boot failure, loss of the GUI environment requiring reinstallation of GDM3, and the overall loss of a day of work. Use MAT in Windows, then port the results elsewhere.]

The application cannot be used programmatically, but it will process a batch of files at a go by inputting a directory of files. It has a graphical user interface, so it is quite easy to use. It is available at https://sites.google.com/site/multidimensionaltagger/ and https://github.com/andreanini/multidimensionalanalysistagger.

Note that Nini (2019) is a contribution in Berber Sardinha and Veirano Pinto (2019).

240. If the reader is on Twitter and follows any number of academics or journalists, this should come as no surprise at all.

241. ★ In this regard, use of MAT to evaluate Biber's dimensions could prove useful for exploring a dataset, as it is possible that reportage-style accounts would tend to be marked as 'written' texts, while more casual accounts would be marked as 'spoken' texts.
A very small scale test on a selection of 'formal' vs 'normal' accounts was done out of curiosity. The former group was marked as 'general exposition' (a subcategory of 'reportage'), and the latter was marked as 'involved persuasion' (a subcategory of 'spontaneous speech'). That warrants further testing, but such was not done for this work.

242. As noted in **n. 150**, other taggers developed specifically for Twitter were investigated and datasets prepared, but time constraints did not allow them to be incorporated into this work. Evaluation of the taggers mentioned in that note is also a topic held for later. Unfortunately, it is the author's impression that work on such social media-specific taggers has slowed, and that attention and effort has been redirected (cf. Rogers, Kovaleva and Rumshisky, 2020).

243. ★ For example, while 'word' could function in a variety of grammatical roles, each appearance would be a token of the type `word`. Thus a phrase `a a b b c` would be composed of five tokens (giving the full phrase) of three types (`a`, `b`, and `c`).

244. Note that this is not simply a reference to computational methods ; tables of contents and indices are also oriented to information retrieval.

245. ★ ARF is used in Sketch Engine, which can be a helpful tool for exploring corpora and corpus-based approaches (see https://www.sketchengine.eu/). However, if the reader is interested in exploring ARF for research, do not use Sketch Engine. It has questionable behaviour in certain cases. Rather, investigate Nilo Pedrazzini's R script for calculating ARF (https://github.com/npedrazzini/averageReducedFrequency). That script was used as a template for the implementation of ARF in this method.

246. ★ ARF was introduced by Petr Savický and Jaroslava Hlaváčová (2002). Full details of the technique can be found there. However, an accessible and brief explanation of ARF is provided at https://www.sketchengine.eu/documentation/average-reduced-frequency/.

247. There may well be an existing measure of pervasiveness in the literature. If so, the author was evidently unaware of it, and eventually had to stop looking for one. Furthermore, related or relevant measures (e.g. *DP* in Gries, 2008) were overly complicated in light of this work's emphasis on relative simplicity in maths (**chp. 1 § 1.7**). The name 'Gamma' was chosen as it is not used for any measure common to the fields that the author usually works in, it was the term used by Alan Turing for the 'input tape' of his machine, and the Greek letter Γ reminds the author of the word game 'Hangman' that he loved to play as a child. While there is a measure *gamma* (γ) used for rank correlation (a measure similar to Kendall's *tau*, used in this work), it is generally presented as Goodman and Kruskal's (1954) *gamma*, so this was not considered a conflicting use.

248. ★ Both terms have a theoretical maximum value of 1 (the actual calculations of ARF are a bit messier than that ; it is a theoretical maximum). While the logarithm is taken for its scaling effect, the equation above would return a maximum possible value of 0 (ie. the common logarithm of 1) with minimum values heading into negative territory. The lower bound is constrained by the logarithm. For example, a perfectly dispersed word (thus having a dispersion factor of 1) occurring 1,000 times amongst 1 billion tokens (thus having a relative prevalence factor of 0.000001) would have a *Gamma* of -6.

To make work a little bit easier, the method as actually performed for this work cheated a little by weighting the relative prevalence factor. It was multiplied by 1 million ; again, as we are taking the common logarithm, this put the maximum theoretical value at 6. *Gamma* as observed in the corpus ranged from roughly +4 to -3. As *Gamma* will be used to calculate a keyness measure of words that is a ratio of *Gamma* in a given subcorpus to *Gamma* in the full corpus, the weighting in any case cancels out. In the meantime, we had easier numbers to work with.

249. For all the author knows, *Gamma* could be a load of non-sense in theoretical or statistical terms. However, in practical terms it proved to be of great utility for discerning different categories of high-frequency words. It is obviously not a stable value, given that the relative prevalence factor is a function of corpus size. But, such questions are far outside of the author's training ; there is likely a better (and proven) measure available. If the reader has any suggestions, the author would be delighted to hear from you.

250. ★ An implementation of the tagger, as well as documentation, is available at https://www.cs.cmu.edu/~ark/TweetNLP. Note that, unlike Nini's MAT, this tagger does not run under Windows but rather Linux. This is a great advantage. If you are unfamiliar with Linux and have a personal Windows machine, consider exploring the Windows Subsystem for Linux (see https://docs.microsoft.com/en-us/windows/wsl/). While it is possible to install Linux alongside Windows, there are risks. Better to explore a bit first.

251. This tagger has a reduced tagset compared to more standard taggers, such as those based on the Penn Treebank. The default model has only 25 tags (a Treeback-style model is available if you so please) ; importantly, 5 of those tags are specific to Twitter—they are hashtag, mention, discourse marker, URL/email, and emoticon. The author found this simplified tagset to be advantageous for workflow, but moreover sensible for application to short-form social media texts, which tend towards the non-standard.

252. ★ In language processing, it is common to hear of 'bag-of-words' approaches. This is what that means: having no concern for word order. Just tossing all the words into a bag, as it were.

253. While the lexical analysis does touch on word order in that it incorporates bigrams, after consideration it was judged that the removed categories would not significantly impact perceived word order. Punctuation was removed without concern, as sentence-boundary bigrams were expected to be numerous in type (given the preservation of source capitalisation) and thus ignored in the procedure. Discourse markers and URLs were observed to be generally initial or final features, thus leaving little evidence if removed. And emoticons, while often taking grammatical function, are more often agrammatical (though nonetheless meaningful!) and thus removal was expected to have minimal impact.

254. For example, numerals were slugged ; thus any token made up of digits was replaced with the common token [NUM]. In the corpus there were some 650,000 numeral tokens. While the bulk of these were probably year references, the potential for variation is enormous.

255. The tagger, following Kevin Gimpel *et al.* (2011), treats phrase-internal hashtags as grammatical elements and tags them accordingly. After slugging, there remained nearly 8,000 unslugged hashtag types (out of some 290,000 types) ; however, only 180 of these were attested by 10 or more tokens. None met the threshold frequency criteria (to be explained) for inclusion in this stage of analysis.

256. The five slugged types accounted for roughly 9% of tokens (3.5 million slugged tokens out of 37 million). However, that is 5 types out of some 290,000 types. Without slugging, the following steps of the procedure might not have been successful within the timeframe given the equipment used (cf. **n. 257**).

Slugs were observed to have a frequency and dispersion similar to function words. This is unsurprising, as slugging effectively creates a high-frequency, closed-class lexis.

257. 'Junk' is a catchall category that the documentation describes as "other abbreviations, foreign words, possessive endings, symbols, garbage" (https://www.cs.cmu.edu/~ark/TweetNLP/annot_guidelines.pdf, p. 1). On inspection, this tag seemed to be given frequently to exclamations and vocatives (the author assumes that the tagger was befuddled by them).

258. ★ This was a time-consuming process. After filtering and slugging, there were 287,474 types present in the corpus, attested by 37,146,764 tokens. Using a 6-core, 32 GB RAM desktop running parallelised code, the computation of *Gamma* at corpus level took several hours (the calculation of ARF is time-consuming). The following process took longer still, and the bigram process took roughly 24 hours. On an average desktop machine the process—if successful—could take days on end. It is imperative to plan for such aspects of large-scale work, and to plan your methods according to what is feasible with your equipment. See **n. 202**.

259. As the calculation of ARF concerns the location of tokens within a corpus, the corpus was converted to bigrams in place. In this manner the phrase 'The quick brown fox', having four tokens, would be converted to 'The.quick quick.brown brown.fox', having three tokens. This technique preserves the relative location of bigrams in the corpus, and is extremely simple to perform. Note that while the token count of the bigram corpus is necessarily N-1 of the plain corpus, the type count was nearly 20 times greater.

260. Thus four lists: two of words and two of bigrams. A third list of words was prepared separately. That list focused on distinctive African American English orthography observed on Twitter, and was based primarily on a list developed by Taylor Jones (2015). Note that while some African American orthography is taken up broadly in appropriated terms (e.g. 'Yas Qween' and the like), Jones' list is of terms that are demonstrably resistant to such uptake, and moreover are deeply rooted—Jones found that observed orthographic variants mapped to demographic flows of the Great Migration. That list was adapted by the author to use as a 'sanity test' of the clustering and mapping stages of the method. Pilot testing showed that this list resolved to expected locations within the site of research. However, the words in this short list (only 31 items) are extremely salient in the broader US context. And, as observed in the user–documents, they appear with relatively lower frequency than more common variants. This is not the category of lexis that the method seeks, and so the list is not included in the final analysis.

261. Note that this is not a problem that can be addressed by normalisation, as bringing a small document up to an arbitrary normal size would simply 'magnify' the role played by any individual item. If keyness were misrepresented in a small corpus that was subsequently normalised, it would still be misrepresented just at a new scale. Considering that these lists would eventually be used to cluster the user–documents, it was decided to follow the lead of Hermann Moisl (2011)—who considers the clustering of documents of drastically varying size—and let go of the really small ones. The cut-off at 1,000 items seemed reasonable to the author as erring on the side of caution.

262. A higher cut-off was intended, but the resultant lists were so small that there were concerns about the subsequent factor stage. The cut-off was reduced stepwise until an acceptable list length was obtained at a minimum of 1 attestation in 200 user–documents (i.e. 0.5%). It was considered that this threshold was too low given the emphasis of this work on tracing social groupings (at a hopefully larger scale). However, smaller lists would prove difficult in the factoring stage for technical reasons. Furthermore, the concern was allayed by the much higher threshold arrived at in the preparation of List 2 (1 in 16).

263. ★ As with List 1, the procedure began with a higher threshold that was lowered stepwise until a list of reasonable and sufficient size resulted. Not too big, not too small, but just right. As the author was flying by the seat of his pants with the method for lexical analysis, that Goldilocks size was arrived at in an iterative back-and-forth between the lexical analysis and the factor analysis. Too big a list, and the factor analysis process might fail (or worse: never conclude) ; too small a list, and the process might not have enough material to work with. Work of this sort takes patience and involves seemingly endless stretches of mistake and error, especially if one (such as the author) is not properly trained in the maths and the techniques. Missteps are normal, so do not fear them—learn from them.

264. ★ An extended description of factor analysis will not be provided here. However, a helpful introduction to factor analysis (and multivariate analysis generally) with specific regard to MDA is Pascual Cantos-Gomez (2019). Note that this is another contribution in Berber Sardinha and Veirano Pinto (2019). For a deeper dive into factor analysis (specifically exploratory factor analysis, which will be appropriate for most social research) see Cadeyrn J. Gaskin and Brenda Happell (2014). Note that Gaskin and Happell is found in a nursing journal. The author has found nursing journals to be an excellent source of methods-oriented research that is theoretically solid but moreover pragmatic. Given the field of work, that is perhaps unsurprising. Let us learn then!

265. ★ Seeking to cluster the user–documents in the next stage of work without first reducing the data would be a fool's errand. Even if the equipment were up for the task (and up for it within a feasible timeframe!), seeking to interpret clusterings on hundreds, even dozens, of variables misses the point of analysis.

266. ★ Strictly speaking such enrichment is not necessary, as the information (i.e. feature counts) is contained in the discursive corpus itself. However, recalculation is a terrible chore, so corpus enrichment is not a matter of showing off, but a matter of practicality. Furthermore, if you get it right once and integrate the results into the corpus, you minimise potential sources of error down the road.

267. ★ In terms of factor analysis, you will most likely encounter two techniques: principal factor analysis, and principal component analysis. Cantos-Gomez gives a quick demonstration of the latter (p. 101), but then quickly notes that the former is the preferred method (p. 106). Both are useful techniques for certain purposes. However, it is important to keep in mind that factoring techniques rely on variance in datasets to do their job. The key distinction between them, and the reason that Cantos-Gomez stresses the preference for factor analysis, is how the techniques handle variance. Component analysis seeks to account for all variance in terms of how variables are grouped (i.e. factored). It is thus well-suited to data reduction (also frequently called 'dimension' reduction). Factor analysis does not necessarily seek to account for all variance, in that residual variance is allowed. This is because factor analysis models a dataset by assuming the presence of latent variables, which component analysis does not do. Thus factor analysis is preferred for factoring datasets that you suspect might have some deeper structure. Note that, for very large datasets, it is not uncommon to see component analysis and factor analysis used back to back—first to reduce the dataset to a more manageable size, and then to factor for a latent structure.

Also, the demonstration that Cantos-Gomez gives of component analysis is done in SPSS. That software and similar (such as Stata and SAS) are proprietary products frequently encountered in industry and in various disciplines with money to spend. By all means familiarise yourself with them if you have (free) access. However, for research work in the social sciences, it is best to avoid such proprietary, unwieldy software. You will be better off in the long term learning to use open-source statistical tools such as R. Doing so will benefit your understanding of the maths and techniques you use, as well as your pocketbook.

268. ★ If you refer to Cantos-Gomez (2019), you will find mention of the correlation matrix as the foundation of factoring techniques (other measures of association are possible, but mostly you will see correlation). That matrix is simply a representation of how each variable in a dataset correlates with every other variable (hence the matrix format). Most descriptions of factoring, including Cantos-Gomez, will not mention the exact technique for the calculation of correlations. That is because Pearson's *r* is the de facto standard, in that it is statistically 'powerful'. That is, you can calculate all sorts of other measures, such as 'significance' (or *p*-values). In some fields significance is an overriding concern due to a focus on inference and hypothesis testing, despite notable problems (e.g. Wasserstein and Lazar, 2016), and so Pearson's *r* is the thing. However, it is not the only option for calculating correlation, as will be discussed.

269. ★ Note that 'parametric' statistical techniques are built upon known probability distributions. Most frequently you will encounter the 'normal' (i.e. Gaussian) distribution. The <u>parameters</u> in question will be familiar: the mean, standard deviation, skew, etc. With these parameters, a known probability distribution can be described completely. This is not the case with unknown distributions. Such parameters cannot characterise them properly: the data must be measured, it cannot rightly be inferred. Thus, such 'non-parametric' techniques do not require the assumption of normality (or other distribution).

270. The package used for factor extraction implemented *tau-b* in cases of tied values (cf. Kendall, 1945).

271. The simplicity of Kendall's *tau* is not simply for the author's benefit. Rather, in light of the interdisciplinary and pedagogical motivations of this work (**chp. 1 § 1.7**), *tau* was the better choice in terms of developing a method that is more readily accessible to a range of researchers (cf. Noether, 1981).

272. ★ Rotation in factor analysis can be confusing. First of all, rotation of what? Basically, rotation is a mathematical manipulation that helps to sort variables more clearly into factors without changing the fundamental analysis. Think of it like turning a strange image one way or the other until it makes sense. It is a simple procedure in the end, although still confusing because there are many possible ways to do it. For a helpful introduction to rotation in the context of exploratory factor analysis, see Jason Osborne (2015).

273. Exactly why this is the case the author is not equipped to discuss, much less explain. However, his suspicion is that equamax plays well with complexity because it seeks to distribute variance equally across factors (Gorsuch, 1974, p. 195). That behaviour was considered desirable for this work—accounting for correlated factors was seen as less worrisome than variables overloading one or two factors, thus defeating the purpose of the exercise. Compare the decision in the analysis of results (**chp. 6**) to abandon the PAF factorings in favour of the CFA factorings.

274. Factoring was also done at 16 levels, but that proved cumbersome in later analysis and so was dropped from consideration. As with soil, finer granularity can get you stuck.

275. ★ There are endless textbooks and websites that will tell you the proper method(s) for determining how many factors to extract from a dataset. The fact of the matter is that there is no surefire way. With empirical social data, there may well be a strong latent structure underlying your data. But odds are good that it is not what you seek, and the first question should be what further latent structures might underlie that.

276. The main concern was the potential for arbitrary or shifting factor assignments to the variables in question.

277. ★ Collinearity refers to variables that are related in such a way that they mimic each other in the data. Thus perfectly collinear variables essentially would be clones of one another. In studies using exploratory factor analysis, it is common to see mention of 'checks for collinearity' (more often multicollinearity, meaning a gaggle of such related variables) or similar. Usually this is stated in a way indicating that such a check is required and important. Rarely does anybody note why. Jesse Egbert and Shelley Staples (2019, p. 127) give as clear a reason as you could ask for: if you have multicollinear variables, you are basically measuring single variables multiple times. That naturally has a negative impact on your analysis, so best avoid it. Many texts will suggest excluding variables that are too highly correlated (or not correlated at all) with other variables, or suggest certain tests (commonly 'sampling adequacy' or 'sphericity') to gauge the factorability of the data. Try not to worry about all that if working with large-scale linguistic data. Rather, follow the advice of Egbert and Staples (ibid.) and focus instead on how you sample. Such rules and tests are unlikely to be applicable to social linguistic data, but your sampling regime can make all the difference. Note that Egbert and Staples is yet another contribution in Berber Sardinha and Veirano Pinto (2019).

Also note that, in regard to determining thresholds of correlation in factor analysis (as in, what is 'highly correlated' anyway?), Richard Gorsuch (1974, pp. 30–31) notes that the question can only be answered properly by considering the effect of such correlations on the end purpose of your analysis. Such answers can be discouraging when learning a method (at least for this author), but they should not be—the key idea is to be sceptical of simple 'prefab' answers, and to think your own work through.

278. ★ For example, such instruments might use a battery of questions to assess if you are anxious or introverted. Note that the author generally avoids citation of literature from behavioural fields, including psychology. Such avoidance is primarily due to a fundamental disagreement with the subject model of such fields. Nevertheless, the psychology literature is often a useful resource in terms of method. Just bear in mind that research design hinges on method<u>ology</u>, thus method and theory jointly, so the question of subject model is never far away.

279. Note that Walkey's description notes a threshold of correlation, but also significance of correlation. This method uses Kendall's *tau* throughout for calculating coefficients, so here the description only mentions threshold.

280. ★ Being able to understand the nuts and bolts of any procedure or method is a clear advantage when dealing with the empirically messy data of social research. This is not always the case, of course, and often we apply techniques without a solid grasp of what happens 'inside the box'. That is okay! Learning always begins at that point. However, techniques that you can interrogate—that is, you can pull them apart in order to understand them—are essential to theoretically grounded social research, and likewise to personal understanding. The use of tools that are effectively 'black boxes', such as proprietary software or certain computational techniques, always yields the role of theory to technology in some degree. Thus large-scale social research, especially when it involves social media platforms, is a kind of balancing act. This work, for example, sourced its data from a black box. However, while hybrid society often puts us at the mercy of black boxes for social data, we have more leeway in our choice of methods. Those that are simple and clear will not get the most fanfare, but they serve better the ends of understanding.

281. The most immediately evident cleavages were political and ethnic. Amusingly, with certain parameters, the algorithm would quickly sort between lexical sets used by fans of different sports. None of this is surprising in the least ; nonetheless it is a joy to observe.

282. The calculation of Kendall's *tau* can be a lengthy process for larger datasets. The grammatical analysis relied on a package (`EFAtools`) that uses the base R implementation of *tau*, which scales in time to the square of items. This proved a problem during pilot testing of the lexical analysis, as only a few hundred items quickly led to unacceptable processing times. Another implementation of *tau* was identified (`cor.fk` in `pcaPP`), which scales to the log of items. The difference is significant, given that the CVA procedure requires repeated recalculation of a correlation matrix (the script was hard-coded to terminate after 200 iterations, but the actual count was generally much less). Using the `pcaPP` implementation, the CVA procedure for all datasets prepared for this work took about 20 minutes, whereas the PAF procedure required about a day and a half—roughly a 1:100 ratio.

283. A given bigram was taken as a single type (cf. **n. 258**).

284. Initially, it was decided only to compose positive correlations, with the rationale that we were concerned with commonalities of speech rather than markers of difference. The initial approach to negative correlations was to zero them out in the matrix. However, the script was eventually modified to handle negative correlations (though not composite them) ; the result was that the script ran 10 times as fast on a given dataset. There is probably a simple procedural explanation for that result, but the author is not savvy in that regard so cannot say. The improvement was a lucky outcome of experimentation. On consideration, it was eventually decided to incorporate positive and negative correlations into composites, and to account for them during factor scoring. This was seen to be the appropriate choice from a social perspective.

285. ★ As the R script was in charge of deciding alphabetical order, this means that capitals come before lower-case letters (i.e. `Apple` comes before `apple`). This is due to the fact that case is treated with different 'code points' by computers, and in UTF-8 (cf. **n. 219**) the codes of English capital letters precede those of lower-case letters.

286. ★ This is necessary because otherwise a composite variable has the correlational 'gravity' of all its constituent members. Without weighting, all items end up glommed onto the first or second aggregation, like Cheerios in milk. The log weighting means that for a given composite variable to have twice the pull of another, all else held equal, it would have to be ten times the size. The author experimented with other weightings, but log (i.e. the common logarithm) was found to be an easy way to get balanced distribution across factors (cf. **n. 272**).

287. The 'lost' remainder would be those items that were the least correlated with other items, so it is sensible to remove them.

288. ★ The *k*-means algorithm is among the most commonly encountered clustering algorithms. Partly this is because it does the job well enough, but also because of its age. The algorithm itself predates the Hartigan and Wong paper by 10 or 20 years. However, the 'Hartigan-Wong' take on the algorithm is commonly encountered.

289. ★ Regarding the name '*k*-means': '*k*' is the variable commonly used to indicate the number of clusters, and 'means' refers to the centre points of clusters, which are found by taking the <u>means</u> of the respective cluster members.

290. ★ In the computational literature of most any field, you will find some 12 bazillion different clustering algorithms. Every day or so a new one pops up, laying claim to the 'state of the art' because it out-performed the previous algorithm by a fraction of a percent on a synthetic dataset. Most of them, however, are variations on a few older themes (for an overview, see Jain, 2010). Better to stick with the tried and true (e.g. Kaufman and Rousseeuw, 1990)—you are more likely to find relevant implementations in the software and packages that you use, and you will have a much deeper literature in which to seek guidance.

291. While the user–documents are clustered solely on continuous data, not mixed data, that will not always be the case in other research undertakings.

292. Note that the faintness of the echo rests first of all on the exploratory nature of the author's method, and second of all on the non-invasive character of the method and its reliance on open-source data.

293. These are bracketed profiles of <u>places</u> in population percentage terms, that is, 10% under 5 years of age, 20% secondary school graduates, etc. See **chp. 5 § 5.1.1.2 n. 199** for more information on these profiles.

294. ★ Note that cluster refers to a specific group, and clustering refers to a set of groups. Thus a *k* 5 <u>clustering</u> produces a set of 5 <u>clusters</u>.

295. Constellation diagrams are already thing in signal analysis. These charts are different, but we borrow the name as it is apt, and is unlikely to cause confusion for signal analysts.

296. The choice of labels is not ad hoc, but rather based on the author's knowledge of the site of research in terms of its social context and history. Recall that ethnicity data is included in the V_{SD} profiles. When accounting for that component of V_{SD} profiles in the constellation charts, it is clear that the labels are broadly accurate for the site of research.

297. Many of the author's observations on inter-coder measures are informed and inspired by these publications. They are strongly recommended for consideration, and the author is thankful to the researchers for their continued work.

298. Recall that α_o is simply the number of cases of agreement divided by the total number of cases.

299. ★ As the number of label permutations is *k*!, this approach is only sensible for small *k*. This step must return 65 α_0 scores per cluster (5 V_{SD} sets * 13 V_L sets). At *k*2 that requires 260 calculations with permutation, and *k*3 requires 1,170 calculations. But *k*5 already requires 39,000 calculations—65 α_0 scores * 5! (120) permutations * 5 clusters. Calculating *k*5 on the author's machine took around 15 minutes ; assuming calculation time increases linearly with calculations, *k*6 (which was not attempted) would require nearly 2 hours and *k*7 more than half a day. Higher *k* clusterings would be an analytical nightmare in any case, so all is well.

300. This is an oversimplification of the case. No relationship, meaning a relatively random assignment of age labels, would yield an adjusted α_0 around 0, that being the point of uncertainty. But the question of randomness raises perhaps the biggest difficulty with applying α_0 in the manner that it is used here—there is every reason to expect that clustering results on socially derived data will be skewed, because society is skewed. As randomness hinges on equal likelihood, which we are unlikely to find amongst any of the variable sets here, we cannot specify a stable zero-point without accounting for the distributions of cluster labels. All things equal, an overlarge cluster would elevate α_0, while a narrow cluster would depress it. In part this is what we are seeing in table 4: V_{SD} set `age`, for example, has an overlarge cluster in k5.2 (nearly twice the size of the next largest) which shows elevated scores in that row, and has its narrowest cluster (only 6% of labels) in k5.4 which shows depressed scores. It seems reasonable to imagine that skewing in cluster assignments could be accounted for in a relatively straightforward manner. However, this realisation came rather late in the game, so the scores stand as they are.

301. This is a small fib. This chart plots the V_{SD} of places, of which there are only 84 in the final dataset. The legend shows the number of user–documents associated with each place cluster. The actual cluster membership in this chart is 46, 19, and 19 again. The user count is shown here as that is how the charts clustering on V_L sets will be presented.

302. In the final dataset, very few places have a dominant cluster assignment in any V_L set. Those that do tend to be represented by a small number of user–documents and to have a small population (both of which go hand in hand).

303. Note that these population figures are for the places in their entirety and not of the proportion of those places sharing linguistic similarities. The final dataset samples slightly more than 0.10% of the actual population, so we would have no way to estimate that figure. One could estimate a total figure by summing the population of places as a function of their cluster proportion, but this would be a rather dubious exercise.

304. Note that the American Community Survey, from which these data were sourced, is structured in terms of households. It is unclear to the author how people coming from out of state or country for study would recorded and reported. More specifically, the author is not clear on the methodology used by the US Census for collecting data in university contexts.

305. To be clear, the reason that immigrants (both from out of state and out of country) are mentioned repeatedly is because they exist across Michigan, and because it is reasonable to expect that they would have linguistic profiles that generally differ from Michiganders. Thus, it is reasonable to expect that they would appear in clusterings. While at lower levels of *k*, such as 3, such groups would be folded into larger groups (hence the supposition of educated out-of-state and out-of-country immigrants in cluster 1 in **fig. 9**), at higher *k*, such as 5, such individuals might be represented sufficiently in the dataset to begin clustering unto themselves (hence the supposition just made regarding clusters 2 and 4 in **fig. 11**). But, as noted previously, such interpretations are merely supposition until grounded in deeper analysis of the data.

306. The index is computed from the raw age, income, and education brackets retrieved from the US Census. As each is expressed in terms of population percentage, each bracket is multiplied by a weighting to give a bracket score. Age brackets are weighted according to the midpoint age they represent (i.e. the 5–9 years bracket is weighted at 7), with the top bracket (65+) given an upper bound of 77 (the average life expectancy in Michigan in 2018). Income brackets are weighted according to the lower bound of each bracket, except for the lowest bracket which was weighted at 1. Education was weighted to reflect the relative sociodemographic ‘value’ of each level of attainment. For example, the lowest bracket (education less than 9th grade) was weighted at 1, high school graduation and equivalent was weighted at 4, an AB at 8, a BA at 12, and a postgraduate or professional degree at 16. (Various weightings for education were tested, and all returned similar results.) Within each category, bracket scores are summed. The index is the min–max normalized product of the category scores.

307. Except for four outliers. The places removed were Bloomfield Hills (one of the wealthiest towns in the United States, pop. 4,000), Davison (hometown of film maker Michael Moore, pop. 5,000), Romeo (hometown of erstwhile musician Kid Rock, pop. 4,000), and Southgate (named for the Southgate Shopping Center, pop. 30,000). Note that, while the V_{SD} index ranges from 0 to 1, the removal of Bloomfield Hills (index score 1) results in the vertical axis in these panels ranging from 0 to approximately 0.7.

308. It is not the purpose of this chart to distinguish individual places. However, it can help to give more sense to what these charts show. For example, the largest city, Detroit, is the largest dot in the lower extremity of each panel ; Lansing, Grand Rapids, and Kalamazoo orbit just above. Ann Arbor is the dominant dot towards the centre in each panel. Each panel presents the site of research from a different 'angle', that being the relative proportion of *V_L* assignment for the cluster in question.

309. Note that this is a rough proxy measure, being the product of the US Census statistics on the percentage population born in-state, and speaking only English at home. That data is by household. It is an unfortunate measure in many ways, given the richness of US society. However, as we are concerned with social structure and sociodemographics, it is pertinent. Also, the question of ethnicity is ever-present in terms of language in society. But, as that is not the question at hand, it will not enter into the discussion in this section. Nevertheless, the information supplements the charts themselves, because it is socially important and revealing.

310. ★ The smoothing line is accomplished using a form of local regression called locally estimated scatterplot smoothing (LOESS, commonly pronounced like Lois Lane). The degree of smoothing (i.e. the span of values that comprise the local estimate) is an arbitrary parameter of the process. At low levels, the line is fit tightly to the data and is extremely janky. As we are not concerned with specific values at any point in the line, the level set here is just past the janky point ; thus the big swoops are more for clarity than aesthetics, though they do look nice.

311. Also note that the actual cluster 2 range for *V_L* set m extends a bit further than shown. While panel views of the full possible range of clusters (i.e. 0–100%) offer a better picture of actual cluster proportions, the panels presented here are right-bounded at 80% for reasons of space.

312. ★ Recall that cluster labels are arbitrary, and thus their order of presentation in these charts. The thing to consider is the character and relations of the data in clusters, regardless of label. And as labels are arbitrary, that is why they can be endlessly permuted without changing the underlying analysis.

313. And thus the first four components of `w1` factor 1 are `Lmaoooo`, `Lmaooooo`, `deadass`, and `lmaooo`. There is great variety in lmaos, as there is in abbreviations for 'thank you' in factor 3—the nerd factor—which also includes `faculty` and `e.g.`. See **appendix B**.

Bibliography

Abusch, T. (2001) 'The Development and Meaning of the Epic of Gilgamesh: An Interpretive Essay', *Journal of the American Oriental Society*, 121(4), p. 614.

Ahmad, S. (1991) 'American Foundations and the Development of the Social Sciences between the Wars: Comment on the Debate between Martin Bulmer and Donald Fisher', *Sociology*, 25(3), pp. 511–520.

Allport, F.H. (1927) 'The Psychological Nature of Political Structure', *American Political Science Review*, 21(3), pp. 611–618.

Almond, G.A. *et al.* (1962) 'Political Science as a Discipline: A Statement by the Committee on Standards of Instruction of the American Political Science Association', *American Political Science Review*, 56(2), pp. 417–421.

Almond, G.A. (1990) *A discipline divided: Schools and sects in political science*. Newbury Park: SAGE.

Almond, G.A. (1998) 'Political Science: The History of the Discipline', in Goodin, R.E. and Klingemann, H.-D. (eds.) *A New Handbook of Political Science.* Oxford: Oxford University Press, pp. 50–96.

Almond, G.A. (2004) 'Who Lost the Chicago School of Political Science?' *Perspectives on Politics*, 2(1), pp. 91–93.

Almond, G.A. and Verba, S. (1963) *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton, N.J.: Princeton University Press.

Alwin, D.F. and Campbell, R.T. (1987) 'Continuity and Change in Methods of Survey Data Analysis', *The Public Opinion Quarterly*, 51, S139-S155.

Amador Diaz Lopez, J.C. *et al.* (2017) 'Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data', *Statistics, Politics and Policy*, 8(1).

Anderson, B.R.O. (1983) *Imagined communities: Reflections on the origin and spread of nationalism*. Reprint, London: Verso, 2006.

Angelis, G. de (2005) 'Interlanguage Transfer of Function Words', *Language Learning*, 55(3), pp. 379–414.

Ansolabehere, S. and Snyder, J.M. (2002) 'The Incumbency Advantage in U.S. Elections: An Analysis of State and Federal Offices, 1942–2000', *Election Law Journal: Rules, Politics, and Policy*, 1(3), pp. 315–338.

Appadurai, A. (1996) *Modernity at large: Cultural dimensions of globalization*. (Public worlds, v. 1). Minneapolis, Minn.: University of Minnesota Press.

Arendt, H. (1951) *The Origins of Totalitarianism*. Reprint, New York: Harcourt, Brace & World, 1966.

Arjas, E. (2001) 'Causal Analysis and Statistics: A Social Sciences Perspective', *European Sociological Review*, 17(1), pp. 59–64.

Aspers, P. (2015) 'Performing ontology', *Social Studies of Science*, 45(3), pp. 449–453.

Augier, M. and March, J.G. (2001) 'Remembering Herbert A. Simon (1916-2001)', *Public Administration Review*, 61(4), pp. 396–402.

Austin, J.L. (1955) *How to Do Things with Words: The William James Lectures delivered at Harvard in 1955*. Reprint, London: Oxford University Press, 1962.

Axelrod, R. (ed.) (1976) *Structure of Decision: The Cognitive Maps of Political Elites*: Princeton University Press.

Bacon, F. (1620) *The Instauratio magna, part II: Novum organum and associated texts*. (Oxford Francis Bacon, 11). Reprint, Oxford.

Baldwin-Philippi, J. (2011) 'Bringing Science and Technology Studies to bear on Communication Studies research', *Communication Research Trends*, 30(2), pp. 4–20.

Barad, K. (2003) 'Posthumanist performativity: toward an understanding of how matter comes to matter', *Signs: Journal of Women in Culture and Society*, 28(3), pp. 801–831.

Barberá, P. *et al.* (2015) 'Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?' *Psychological Science*, 26(10), pp. 1531–1542.

Barberá, P. and Rivero, G. (2015) 'Understanding the Political Representativeness of Twitter Users', *Social Science Computer Review*, 33(6), pp. 712–729.

Barley, S.R. (1986) 'Technology as an occasion for structuring: evidence from observations of CT scanners and the social order of radiology departments', *Administrative Science Quarterly*, 31(1), pp. 78–108.

Barnes, S.H., Kaase, M. and et al. (eds.) (1979) *Political Action: Mass Participation in Five Western Democracies*. Beverly Hills, California: SAGE Publications.

Barrow, C.W. (2011) 'Politics Denied: Comments on Waismel-Manor and Lowi's "Politics in Motion"', *New Political Science*, 33(1), pp. 79–86.

Barrow, C.W. (2017) 'The Political and Intellectual Origins of New Political Science', in Boryczka, J.M. and Disney, J.L. (eds.) *50th Anniversary of the Caucus for a New Political Science 1967-2017: What is 'New' About New Political Science?* (39), pp. 437–472.

Bayard de Volo, L. and Schatz, E. (2004) 'From the Inside Out: Ethnographic Methods in Political Research', *PS: Political Science and Politics*, 37(2), pp. 267–271.

Benkler, Y. (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, Conn.: Yale University Press.

Bennet, S.E., Flickinger, R.S. and Rhine, S.L. (2000) 'Political Talk Over Here, Over There, Over Time', *British Journal of Political Science*, 30(1), pp. 99–119.

Bennett, S.E. and Bennett, L.L. (1986) 'Political Participation', in Long, S. (ed.) *Annual review of political science.* (1). Norwood, N.J: Ablex, pp. 157–204.

Bennett, W.L. (2012) 'The Personalization of Politics: Political Identity, Social Media, and Changing Patterns of Participation', *The Annals of the American Academy of Political and Social Science*, 644(1), pp. 20–39.

Bennett, W.L. and Pfetsch, B. (2018) 'Rethinking Political Communication in a Time of Disrupted Public Spheres', in Fuchs, C. and Qiu, J.L. (eds.) *Ferments in the Field: Introductory Reflections on the Past, Present and Future of Communication Studies.* (68), pp. 243–253.

Benoit, K. (2020) 'Text as Data: An Overview', in Curini, L. and Franzese, R.J. (eds.) *The SAGE handbook of research methods in political science and international relations.* Los Angeles: SAGE, pp. 461–496.

Berber Sardinha, T. (2019) 'Using Multi-Dimensional Analysis to Detect Representations of National Cultures', in Berber Sardinha, T. and Veirano Pinto, M. (eds.) *Multi-dimensional analysis: Research methods and current issues*. London, UK: Bloomsbury Academic, pp. 231–258.

Berber Sardinha, T. and Veirano Pinto, M. (eds.) (2019) *Multi-dimensional analysis: Research methods and current issues*. London, UK: Bloomsbury Academic.

Berelson, B.R., Lazarsfeld, P.F. and McPhee, W.N. (1954) *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: The University of Chicago Press.

Berger, B. (2009) 'Political Theory, Political Science and the End of Civic Engagement', *Perspectives on Politics*, 7(2), pp. 335–350.

Berger, P.L. and Luckmann, T. (1966) *The social construction of reality: a treatise in the sociology of knowledge*. New York: Anchor Books.

Berkenpas, J. (2016) ''The Behavioural Revolution'? A Genealogy of a Concept', *European Political Science*, 15(2), pp. 233–250.

Berlo, D.K. (1974) *The process of communication: An introduction to theory and practice*. 17th edn. New York, N.Y., [etc.]: Holt, Rinehart and Winston.

Berndtson, E. (1975) 'Political Science in the Era of Post-Behavioralism: The Need for Self-Reflection', *Scandinavian Political Studies*, 10(A10), pp. 173–184.

Berndtson, E. (1987) 'The Rise and Fall of American Political Science: Personalities, Quotations, Speculations', *International Political Science Review*, 8(1), pp. 85–100.

Bernstein, B. (1971) *Class, codes and control: Theoretical Studies towards a Sociology of Language* (4 vols). (1). Reprint, London: Routledge, 2003.

Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1992) 'The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings', *Computers and the Humanities*, 26(5/6), pp. 331–345.

Biber, D. (2019) 'Text-linguistic approaches to register variation', *Register Studies*, 1(1), pp. 42–75.

Biber, D. and Conrad, S. (2019) *Register, genre, and style*. 2nd edn. Cambridge: Cambridge University Press (Cambridge textbooks in linguistics).

Biber, D., Egbert, J. and Keller, D. (2020) 'Reconceptualizing register in a continuous situational space', *Corpus Linguistics and Linguistic Theory*, 16(3), pp. 581–616.

Biber, D. and Jones, J.K. (2005) 'Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles', *Corpus Linguistics and Linguistic Theory*, 1(2), pp. 151–182.

Bijker, W.E. (1995) *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change*. (Inside technology). Cambridge, MA: MIT Press.

Bijker, W.E. (2010) 'How is technology made? --That is the question!', *Cambridge Journal of Economics*, 34(1), pp. 63–76.

Bijker, W.E., Hughes, T.P. and Pinch, T.J. (eds.) (1987) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, Massachusetts: MIT Press.

Bijker, W.E. and Law, J. (eds.) (1992) *Shaping Technology / Building Society: Studies in Sociotechnical Change*. Cambridge, Mass.: MIT Press (Inside technology).

Bijker, W.E. and Pinch, T.J. (2012) 'Preface to the Anniversary Edition', in Bijker, W.E., Hughes, T.P. and Pinch, T.J. (eds.) *The social construction of technological systems: New directions in the sociology and history of technology.* Cambridge, Mass.: MIT Press, pp. xi–xxxiv.

Bimber, B. (1994) 'Three Faces of Technological Determinism', in Smith, M.R. and Marx, L. (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism.* Cambridge, MA: MIT Press, pp. 79–100.

Bimber, B. (1999) ‘The Internet and Citizen Communication with Government: Does the Medium Matter?’ *Political Communication*, 16(4), pp. 409–428.

Bimber, B. (2001) ‘Information and Political Engagement in America: The Search for Effects of Information Technology at the Individual Level’, *Political Research Quarterly*, 54(1), pp. 53–67.

Bimber, B. (2003) *Information and American Democracy: Technology in the Evolution of Political Power*. Cambridge: Cambridge University Press.

Binder, L. *et al.* (1971) *Crises and Sequences in Political Development*. Princeton: Princeton University Press (Studies in Political Development, 7).

Blais, A. (2000) *To Vote or Not to Vote: The Merits and Limits of Rational Choice Theory*. Pittsburgh, Pennsylvania: University of Pittsburgh Press.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) ‘Latent Dirichlet Allocation’, *Journal of Machine Learning Research*, 3(4/5), pp. 993–1022.

Blommaert, J. (2017) *Society through the lens of language* (Tilburg Papers in Culture Studies 178). Available at: https://pure.uvt.nl/ws/portalfiles/portal/32303872/TPCS_178_Blommaert.pdf.

Blommaert, J. (2019) ‘From groups to actions and back in online-offline sociolinguistics’, in Al Zidjaly, N. (ed.) *Society in Digital Contexts: New Modes of Identity and Community Construction.* (38), pp. 485–493.

Blommaert, J. and Dong, J. (2019) *When Your Field Goes Online* (Tilburg Papers in Culture Studies 227).

Blommaert, J., Smits, L. and Yacoubi, N. (2018) *Context and its Complications* (Tilburg Papers in Culture Studies 208). Available at: https://pure.uvt.nl/ws/portalfiles/portal/48993768/TPCS_208_Blommaert_Smits_Yacoubi.pdf.

Blumer, H. (1954) ‘What is Wrong with Social Theory?’ *American Sociological Review*, 19(1), pp. 3–10.

Blumer, H. (1969) *Symbolic Interactionism: Perspective and Method*. Englewood Cliffs, NJ: Prentice Hall.

Blumler, J.G. and Gurevitch, M. (2001) ‘The New Media and Our Political Communication Discontents: Democratizing Cyberspace’, *Information, Communication & Society*, 4(1), pp. 1–13.

Boas, F. (1900) ‘Sketch of the Kwakiutl Language’, *American Anthropologist*, 2(4), pp. 708–721.

Boas, F. (1906) ‘Some Philological Aspects of Anthropological Research’, *Science*, 23(591), pp. 641–645.

Boczkowski, P.J. and Lievrouw, L.A. (2007) ‘Bridging STS and Communication Studies: Scholarship on Media and Information Technologies’, in Hackett, E.J. (ed.) *The Handbook of Science and Technology Studies*, 3rd edn. Cambridge, Mass.: MIT, pp. 949–977.

Boncourt, T. (2008) ‘Is European Political Science different from European Political Sciences? a comparative study of the European Journal of Political Research, Political Studies and the Revue Française de Science Politique 1973-2002’, *European Political Science*, 7(3), pp. 366–381.

Boncourt, T. (2015) ‘The Transnational Circulation of Scientific Ideas: Importing Behavioralism in European Political Science (1950-1970)’, *Journal of the History of the Behavioral Sciences*, 51(2), pp. 195–215.

Bond, J.R. (2007) ‘The Scientification of the Study of Politics: Some Observations on the Behavioral Evolution in Political Science’, *Journal of Politics*, 69(4), pp. 897–907.

Bond, R.M. and Messing, S. (2015) ‘Quantifying Social Media’s Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook’, *American Political Science Review*, 109(1), pp. 62–78.

Borg, E. (2003) ‘Discourse community’, *ELT Journal*, 57(4), pp. 398–400.

Borra, E. and Rieder, B. (2014) 'Programmed method: Developing a toolset for capturing and analyzing tweets', *Aslib Journal of Information Management*, 66(3), pp. 262–278.

Boulianne, S. (2015) 'Social media use and participation: A meta-analysis of current research', *Information, Communication & Society*, 18(5), pp. 524–538.

Boulianne, S. (2020) 'Twenty Years of Digital Media Effects on Civic and Political Participation', *Communication Research*, 47(7), pp. 947–966.

Bourdieu, P. (1991) *Language and symbolic power*. Cambridge: Polity.

Bourdieu, P. (1993) *The field of cultural production: Essays on art and literature*.

Bourdieu, P. (2003) 'Participant Objectivation*', *Journal of the Royal Anthropological Institute*, 9(2), pp. 281–294.

boyd, d. and Crawford, K. (2012) 'Critical Questions for Big Data', *Information, Communication & Society*, 15(5), pp. 662–679.

boyd, d.m. and Ellison, N.B. (2007) 'Social Network Sites: Definition, History, and Scholarship', *Journal of Computer-Mediated Communication*, 13(1), pp. 210–230.

Brady, H.E. (1999) 'Political Participation', in Robinson, J.P., Shaver, P.R. and Wrightsman, L.S. (eds.) *Measures of Political Attitudes.* (Measures of social psychological attitudes, v. 2). San Diego: Academic, pp. 737–800.

Brady, H.E., Collier, D. and Box-Steffensmeier, J.M. (2011) 'Overview Of Political Methodology: Post-Behavioral Movements and Trends', in Goodin, R.E. (ed.) *The Oxford handbook of political science.* (The Oxford handbooks of political science). Oxford: Oxford University Press.

Brady, H.E., Verba, S. and Schlozman, K.L. (1995) 'Beyond SES: A Resource Model of Political Participation', *American Political Science Review*, 89(02), pp. 271–294.

Brody, R. (1978) 'The puzzle of political participation in America', in Beer, S. (ed.) *The New American Political System,.* Washington, D.C.: American Enterprise Institute.

Brown, P.F. *et al.* (1992) 'Class-based n-gram models of natural language', *Computational Linguistics*, 18(4), pp. 467–479.

Bruns, A. (2018) 'Big Social Data Approaches in Internet Studies: The Case of Twitter', in Hunsinger, J., Allen, M.M. and Klastrup, L. (eds.) *Second International Handbook of Internet Research.* (Springer eBook Collection). Dordrecht: Springer, pp. 65–81.

Bruns, A. and Burgess, J.E. (2011) 'The use of Twitter hashtags in the formation of ad hoc publics', *6th European Consortium for Political Research General Conference.* European Consortium for Political Research. University of Iceland, Reykjavik.

Bucholtz, M. and Hall, K. (2008) 'All of the above: New coalitions in sociocultural linguistics', *Journal of Sociolinguistics*, 12(4), pp. 401–431.

Buller, H. (2009) 'The lively process of interdisciplinarity', *Area*, 41(4), pp. 395–403.

Bulmer, M. (1981) 'Quantification and Chicago Social Science in the 1920s: A neglected tradition', *Journal of the History of the Behavioral Sciences*, 17(3), pp. 312–331.

Burch, R. (1990) 'Phenomenology, Lived Experience: Taking a Measure of the Topic', *Phenomenology + Pedagogy*, pp. 130–160.

Burgess, R.G. (1984) *In the Field*. (Contemporary Social Research Series, 8): Unwin Hyman Ltd.

Burke, K. (1966) *Language as Symbolic Action: Essays on Life, Literature, and Method*. 2019th edn. Berkeley, CA: University of California Press.

Butler, J. and Spivak, G.C. (2007) *Who sings the nation-state? Language, politics, belonging*. London: Seagull Books.

Cadenas, H. and Arnold, M. (2015) ‘The Autopoiesis of Social Systems and its Criticisms’, in Riegler, A. (ed.) *Constructivist Foundations 10:2.* (10). Brussels.

Cain, B., Dalton, R. and Scarrow, S. (eds.) (2003) *Democracy Transformed? Expanding Political Opportunities in Advanced Industrial Democracies*. Oxford: Oxford University Press.

Cairns, A.C. (1975) ‘Political Science in Canada and the Americanization Issue’, *Canadian Journal of Political Science*, 8(2), pp. 191–234.

Callon, M. (1986) ‘Some Elements of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay’, in Law, J. (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* London: Routledge, pp. 196–223.

Campbell, A. *et al.* (1960) *The American Voter*. New York: John Wiley.

Campbell, A., Gurin, G. and Miller, W.E. (1954) *The Voter Decides*: Row, Peterson, and Co.

Campbell, D.E. (2013) ‘Social Networks and Political Participation’, *Annual Review of Political Science*, 16(1), pp. 33–48.

Cantos-Gomez, P. (2019) ‘Multivariate Statistics Commonly Used in Multi-Dimensional Analysis’, in Berber Sardinha, T. and Veirano Pinto, M. (eds.) *Multi-dimensional analysis: Research methods and current issues.* London, UK: Bloomsbury Academic, pp. 97–124.

Carpentier, N. (2016) ‘Beyond the Ladder of Participation: An Analytical Toolkit for the Critical Analysis of Participatory Media Processes’, *Javnost - The Public*, 23(1), pp. 70–88.

Carrasco-Campos, Á. and Saperas, E. (2020) ‘Neoliberalism and Academia in Communication and Media Studies: A New Institutional Framework’, *Communication, Capitalism & Critique*, 19(1), pp. 195–211.

Carter, M.J. and Fuller, C. (2016) ‘Symbols, meaning, and action: The past, present, and future of symbolic interactionism’, *Current Sociology*, 64(6), pp. 931–961.

Castells, M. (1996) *The rise of the network society*. 2nd edn. (The information age : economy, society, and culture, v. 1). Reprint, Oxford: Wiley-Blackwell, 2010.

Cefaï, D. (2000) ‘The Field Training Project: A Pioneer Experiment in Fieldwork Methods: Everett C. Hughes, Buford H. Junker and Raymond Gold’s Re-invention of Chicago Field Studies in the 1950’s’, *Antropolítica*, 9, pp. 25–76.

Chadwick, A. (2013) *The Hybrid Media System: Politics and power*. (Oxford studies in digital politics).

Chapoulie, J.-M. (1996) ‘Everett Hughes and the Chicago Tradition’, *Sociological Theory*, 14(1), p. 3.

Chemero, A. (2003) ‘An Outline of a Theory of Affordances’, *Ecological Psychology*, 15(2), pp. 181–195.

Chilton, P. and Schäffner, C. (2011) ‘Discourse and Politics’, in van Dijk, T.A. (ed.) *Discourse studies: A multidisciplinary introduction,* 2nd edn. London: SAGE Publications Ltd, pp. 303–330.

Chomsky, N. (1957) *Syntactic Structures*. London: Mouton & Co.

Chomsky, N. (1975) *Reflections on language*. New York: Pantheon Books.

Cihon, P. and Yasseri, T. (2016) ‘A Biased Review of Biases in Twitter Studies on Political Collective Action’, *Frontiers in Physics*, 4.

Clarke, A.E. (2003) ‘Situational Analyses: Grounded Theory Mapping After the Postmodern Turn’, *Symbolic Interaction*, 26(4), pp. 553–576.

Clarke, I. (2019) ‘Functional linguistic variation in Twitter trolling’, *International Journal of Speech Language and the Law*, 26(1), pp. 57–84.

Clarke, I. and Grieve, J. (2019) ‘Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018’, *PLoS ONE*, 14(9), e0222062.

Clifford, J. (1983) 'On Ethnographic Authority', *Representations*, 2, pp. 118–146.

Clifford, J. (1997) 'Spatial Practices: Fieldwork, Travel, and the Disciplining of Anthropology', in Gupta, A. and Ferguson, J. (eds.) *Anthropological Locations: Boundaries and Grounds of a Field Science.* Berkeley: University of California Press, pp. 185–222.

Comor, E. (2001) 'Harold Innis and 'The Bias of Communication'', *Information, Communication & Society*, 4(2), pp. 274–294.

Comte, A. (1851) *System of Positive Polity*. (1).

Conge, P.J. (1988) 'The Concept of Political Participation: Toward a Definition', *Comparative Politics*, 20(2), pp. 241–249.

Conway, M.M. (1991) 'The Study of Political Participation: Past, Present, and Future', in Crotty, W.J. (ed.) *Political Science: Looking to the Future.* Political Behavior. (3). Evanston, Ill.: Northwestern University Press, pp. 31–50.

Conway, M.M. (2000) *Political participation in the United States*. 3rd edn. Washington D.C.: CQ Press.

Costello, A.B. and Osborne, J. (2005) 'Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis', *Practical Assessment, Research, and Evaluation*, 10.

Couldry, N. *et al.* (2014) 'Digital citizenship? Narrative exchange and the changing terms of civic culture', *Citizenship Studies*, 18(6-7), pp. 615–629.

Couldry, N. and Hepp, A. (2013) 'Conceptualizing Mediatization: Contexts, Traditions, Arguments', *Communication Theory*, 23(3), pp. 191–202.

Couldry, N. and Hepp, A. (2017) *The mediated construction of reality*. Cambridge: Polity Press.

Coupland, N. (2007) *Style: Language variation and identity*. (Key topics in sociolinguistics). Cambridge: Cambridge University Press.

Courtois, C. *et al.* (2011) 'The triple articulation of media technologies in teenage media consumption', *New Media & Society*, 14(3), pp. 401–420.

Craig, R.T. (1990) 'The Speech Tradition: Chautauqua: Are rhetoric and science incompatible?' *Communication Monographs*, 57(4), pp. 309–314.

Craig, R.T. (1999) 'Communication Theory as a Field', *Communication Theory*, 9(2), pp. 119–161.

Craig, R.T. (2013) 'Constructing Theories in Communication Research', in Cobley, P. and Schulz, P.J. (eds.) *Theories and Models of Communication.* Berlin: De Gruyter, pp. 39–57.

Crenshaw, K. (2019) *On intersectionality: Essential writings*. New York: New Press.

Crowther-Heyck, H. (2006) 'Herbert Simon and the GSIA: building an interdisciplinary community', *Journal of the History of the Behavioral Sciences*, 42(4), pp. 311–334.

Curtice, J. and Seyd, B. (2003) 'Is There a Crisis of Political Participation?' in Park, A. *et al.* (eds.) *British Social Attitudes, The 20th Report: Continuity and Change over Two Decades:* SAGE, pp. 93–107.

Dafoe, A. (2015) 'On Technological Determinism', *Science, Technology, & Human Values*, 40(6), pp. 1047–1076.

Dahl, R.A. (1961) 'The Behavioral Approach in Political Science: Epitaph for a Monument to a Successful Protest', *American Political Science Review*, 55(04), pp. 763–772.

Dahlberg, L. (2011) 'Re-constructing digital democracy: An outline of four 'positions'', *New Media & Society*, 13(6), pp. 855–872.

Dahlgren, P. (2005) 'The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation', *Political Communication*, 22(2), pp. 147–162.

Dahlgren, P. (2006) ‘Doing citizenship’, *European Journal of Cultural Studies*, 9(3), pp. 267–286.

Dahlgren, P. (2009) *Media and Political Engagement: Citizens, Communication, and Democracy*: Cambridge University Press.

Dahlgren, P. and Alvares, C. (2014) ‘Political Participation in an Age of Mediatisation’, *Javnost–The Public*, 20(2), pp. 47–65.

Dalton, R.J. (2008) ‘Citizenship Norms and the Expansion of Political Participation’, *Political Studies*, 56(1), pp. 76–98.

Dalton, R.J. and Klingemann, H.-D. (2011) ‘Overview of Political Behavior: Political Behavior and Citizen Politics’, in Goodin, R.E. (ed.) *The Oxford handbook of political science.* (The Oxford handbooks of political science). Oxford: Oxford University Press.

Daniel, G.E. (1943) *The Three Ages: An Essay on Archaeological Method*. Cambridge: Cambridge University Press.

Darnell, R. (1990) ‘Franz Boas, Edward Sapir, and the Americanist Text Tradition’, in Dinneen, F.P. and Koerner, E.F.K. (eds.) *North American contributions to the history of linguistics.* Amsterdam: J. Benjamins Pub. Co, pp. 129–144.

de la Cruz Paragas, Fernando and Lin, T.T.C. (2014) ‘Organizing and reframing technological determinism’, *New Media & Society*, 18(8), pp. 1528–1546.

Deacon, T. (1998) *The symbolic species: The co-evolution of language and the brain*. New York: W.W. Norton.

Delli Carpini, M.X., Cook, F.L. and Jacobs, L.R. (2004) ‘Public Deliberation, Discursive Participation, and Citizen Engagement: A Review of the Empirical Literature’, *Annual Review of Political Science*, 7(1), pp. 315–344.

Derczynski, L. *et al.* (2013) ‘Microblog-genre noise and impact on semantic annotation accuracy’, *Proceedings of the 24th ACM Conference on Hypertext and Social Media, 24th ACM Conference on Hypertext and Social Media.* Association for Computing Machinery (ACM), Paris, France, 1-3 May 2013. New York, NY: ACM, pp. 21–30.

Derczynski, L. *et al.* (2013) ‘Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data’, *Proceedings of the International Conference Recent Advances in Natural Language Processing, International Conference Recent Advances in Natural Language Processing (RANLP).* Association for Computational Linguistics (ACL), Hissar, Bulgaria, 9-11 September 2013, pp. 198–206.

Derrida, J. and Ferraris, M. (2001) *A Taste for the Secret*. Malden, MA: Polity.

Deseriis, M. (2020) ‘Rethinking the digital democratic affordance and its impact on political representation: Toward a new framework’, *New Media & Society*, 1-22.

Dewey, J. (1888) *The Ethics of Democracy*. Ann Arbor (Philosophical Papers 1). Available at: https://babel.hathitrust.org/cgi/pt?id=uc1.a0009195611.

Dewey, J. (1927) *The Public and Its Problems*. Denver: Alan Swallow.

Diamond, L.J. and Gunther, R. (eds.) (2001) *Political Parties and Democracy*. Baltimore MD: Johns Hopkins University Press (Journal of democracy book).

Driessen, J. (1997) ‘Raymond L. Gold’s Ethnographic Method in Sociology’, *Qualitative Inquiry*, 3(4), p. 387.

Dryzek, J.S. (2006) ‘Revolutions without Enemies: Key Transformations in Political Science’, *The American Political Science Review*, 100(4), pp. 487–492.

Dryzek, J.S. *et al.* (2019) ‘The crisis of democracy and the science of deliberation’, *Science (New York, N.Y.)*, 363(6432), pp. 1144–1146.

Duranti, A. (1997) *Linguistic anthropology*. (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press.

Duranti, A. (2006) 'The social ontology of intentions', *Discourse Studies*, 8(1), pp. 31–40.

Duranti, A. (2010) 'Husserl, intersubjectivity and anthropology', *Anthropological Theory*, 10(1-2), pp. 16–35.

Durkheim, E. (1893) *De la division du travail social: étude sur l'organisation des sociétés supérieures*. Paris: Alcan.

Durkheim, E. (1895) *The Rules of Sociological Method*. New York: Free Press.

Dye, C., Kedar, Y. and Lust, B. (2019) 'From lexical to functional categories: New foundations for the study of language development', *First Language*, 39(1), pp. 9–32.

Easton, D. (1953) *The political system: An inquriy into the state of political science*. New York: Alfred Knopf.

Easton, D. (1965) *A Framework for Political Analysis*. (& Contemporary political theory series): Prentice-Hall.

Easton, D. (1969) 'The New Revolution in Political Science', *The American Political Science Review*, 63(4), p. 1051.

Easton, D., Gunnel, J.G. and Graziano, L. (1991) *The Development of Political Science: A comparative survey*. London: Routledge.

Eckert, P. (2008) 'Variation and the indexical field', *Journal of Sociolinguistics*, 12(4), pp. 453–476.

Eckert, P. (2012) 'Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation', *Annual Review of Anthropology*, 41(1), pp. 87–100.

Eckert, P. (2018) *Meaning and linguistic variation: The third wave in sociolinguistics*. Cambridge: Cambridge University Press.

Edelman, M. (1977) *Political language: Words that succeed and policies that fail*. (Monograph series / University of Wisconsin. Institute for Research on Poverty). New York: Academic Press.

Edelman, M. (1985) 'Political Language and Political Reality', *PS: Political Science and Politics*, 18(1), p. 10.

Edelmann, A. *et al.* (2020) 'Computational Social Science and Sociology', *Annual Review of Sociology*, 46(1), pp. 61–81.

Egbert, J. and Staples, S. (2019) 'Doing Multi-Dimensional Analysis in SPSS, SAS, and R', in Berber Sardinha, T. and Veirano Pinto, M. (eds.) *Multi-dimensional analysis: Research methods and current issues.* London, UK: Bloomsbury Academic, pp. 125–144.

Ekman, J. and Amnå, E. (2012a) 'Political participation and civic engagement: Towards a new typology', *Human Affairs*, 22(3), p. 236.

Ekman, J. and Amnå, E. (2012b) 'Political participation and civic engagement: Towards a new typology', *Human Affairs*, 22(3), pp. 283–300.

Eldersveld, S.J. *et al.* (1952) 'Research in Political Behavior', *The American Political Science Review*, 46(4), pp. 1003–1045.

Emerson, R.W. (1870) *Society and Solitude: Twelve Chapters*. Boston: Fields, Osgood & Co.

Emirbayer, M. (1997) 'Manifesto for a Relational Sociology'. American Journal of Sociology, 103(2), 281-317, *American Journal of Sociology*, 103(2), pp. 281–317.

Emirbayer, M. and Mische, A. (1998) 'What Is Agency?'. American Journal of Sociology, 103(4), 962-1023, *American Journal of Sociology*, 103(4), pp. 962–1023.

Engle, R.F., Hendry, D.F. and Richard, J.-F. (1983) 'Exogeneity', *Econometrica*, 51(2), pp. 277–304.

Erickson, M. and Webster, F. (2012) 'Science and Technology', in Ritzer, G. (ed.) *The Wiley-Blackwell Companion to Sociology.* Oxford: Wiley-Blackwell, pp. 609–625.

Errington, J. (1999) 'Ideology', *Journal of Linguistic Anthropology*, 9(1-2), pp. 115–117.

Esarey, J. (2018) 'What Makes Someone a Political Methodologist?' *PS: Political Science and Politics*, 51(03), pp. 588–596.

Eulau, H. (1963) *The Behavioral Persuasion in Politics*: Random House.

Eulau, H. (1997) 'Book Review: Political Science in History: Research Programs and Political Traditions. Edited by James Farr, John S. Dryzek, and Stephen T. Leonard.', *Journal of Politics*, 59(02), p. 583.

Evans, J.A. and Aceves, P. (2016) 'Machine Translation: Mining Text for Social Theory', *Annual Review of Sociology*, 42(1), pp. 21–50.

Evans, S.K. *et al.* (2017) 'Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research', *Journal of Computer-Mediated Communication*, 22(1), pp. 35–52.

Fairclough, N. (1989) *Language and power*. (Language in social life series). Harlow: Longman.

Fairhurst, G.T. and Putnam, L. (2004) 'Organizations as Discursive Constructions', *Communication Theory*, 14(1), pp. 5–26.

Faraj, S. and Azad, B. (2012) 'The Materiality of Technology: And Affordance Perspective', in Leonardi, P.M., Nardi, B.A. and Kallinikos, J. (eds.) *Materiality and Organizing: Social Interaction in a Technological World.* Oxford: Oxford University Press, pp. 237–258.

Farr, J. (1988) 'The History of Political Science', *American Journal of Political Science*, 32(4), p. 1175.

Farr, J. (1995) 'Remembering the Revolution: Behavioralism in American Political Science', in Leonard, S.T., Farr, J. and Dryzek, J. (eds.) *Political science in history: Research programs and political traditions.* Cambridge: Cambridge University Press, pp. 198–224.

Farrell, H. (2012) 'The Consequences of the Internet for Politics', *Annual Review of Political Science*, 15(1), pp. 35–52.

Faulkner, P. and Runde, J. (2011) 'The Social, the Material, and the Ontology of Non-Material Technological Objects'. paper presented at the European Group for Organizational Studies (EGOS) Colloquium. Gothenburg.

Fayard, A.-L. and Weeks, J. (2007) 'Photocopiers and Water-coolers: The Affordances of Informal Interaction', *Organization Studies*, 28(5), pp. 605–634.

Feezell, J.T. (2016) 'Predicting Online Political Participation: The Importance of Selection Bias and Selective Exposure in the Online Setting', *Political Research Quarterly*, 69(3), pp. 495–509.

Feng, G.C. (2013) 'Underlying determinants driving agreement among coders', *Quality & Quantity*, 47(5), pp. 2983–2997.

Feng, G.C. (2014) 'Intercoder reliability indices: disuse, misuse, and abuse', *Quality & Quantity*, 48(3), pp. 1803–1815.

Ferrari, G. (2004) 'State of the Art in Computational Linguistics', in van Sterkenburg, P.G.J. (ed.) *Linguistics today: Facing a greater challenge.* Amsterdam: John Benjamins Pub, pp. 163–186.

Feyerabend, P.K. (1962) 'Explanation, reduction, and empiricism', in Feigl, H. and Maxwell, G. (eds.) *Scientific explanation, space, and time.* (Minnesota Studies in the Philosophy of Science, 3): University of Minnesota Press, Minneapolis.

Fielding, R.T. and Taylor, R.N. (2002) 'Principled design of the modern Web architecture', *ACM Transactions on Internet Technology*, 2(2), pp. 115–150.

Finin, T. *et al.* (2010) 'Annotating Named Entities in Twitter Data with Crowdsourcing', *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.* NAACL-HLT, Los Angeles, California, USA, 6 June, pp. 80–88.

Fishman, J.A. (1972) *Language in sociocultural change*. (Language science and national development). Stanford Calif.: Stanford University Press.

Fiske, S.T. and Taylor, S.E. (1984) *Social Cognition*. Reading, Mass.: Addison-Wesley Pub. Co.

Fitch, W.T. (2005) 'The Evolution of Language: A Comparative Review', *Biology and Philosophy*, 20(2), pp. 193–203.

Fitch, W.T., Huber, L. and Bugnyar, T. (2010) 'Social cognition and the evolution of language: Constructing cognitive phylogenies', *Neuron*, 65(6), pp. 795–814.

Foner, E. (1988) *Reconstruction: America's unfinished revolution, 1863-1877*. (The New American nation). New York: Harper & Row.

Forman, P. (2007) 'The Primacy of Science in Modernity, of Technology in Postmodernity, and of Ideology in the History of Technology', *History and Technology*, 23(1-2), pp. 1–152.

Fox, C.J. (1996) 'Reinventing Government as Postmodern Symbolic Politics', *Public Administration Review*, 56(3), pp. 256–262.

Fox, S. (2014) 'Is it Time to Update the Definition of Political Participation?' *Parliamentary Affairs*, 67(2), pp. 495–505.

Franklin, M. (2004) *Voter Turnout and the Dynamics of Electoral Competition in Established Democracies since 1945.* New York: Cambridge University Press.

Freedman, D. (2002) 'A 'Technological Idiot'? Raymond Williams and Communications Technology', *Information, Communication & Society*, 5(3), pp. 425–442.

Freelon, D. (2018) 'Computational research in the post-API age', *Political Communication*, pp. 1–4.

Frezza, D. (2007) *The leader and the crowd: Democracy in American public discourse, 1880-1941*. Athens, Ga.: University of Georgia Press ; London : Eurospan [distributor].

Friedl, J.E.F. (2006) *Mastering regular expressions*. 3rd edn. Farnham: O'Reilly.

Friedman, J. (1995) 'Economic approaches to politics', *Critical Review*, 9(1-2), pp. 1–24.

Fries, C.C. (1952) *The Structure of English: An introduction to the construction of English sentences*. New York: Harcourt, Brace & World.

Friginal, E. (ed.) (2018) *Studies in corpus-based sociolinguistics*. New York, NY: Routledge.

Frognier, A.-P. (2002) 'Une vue européenne sur la science politique française', *Revue française de science politique*, 52(5), p. 641.

Fuchs, C. and Qiu, J.L. (eds.) (2018) *Ferments in the Field: Introductory Reflections on the Past, Present and Future of Communication Studies* (68).

Fuchs, S. (2007) 'Agency (and Intention)', in Ritzer, G. (ed.) *The Blackwell Encyclopedia of Sociology.* Malden, Mass.: Blackwell, pp. 60–62.

Gal, S. and Irvine, J.T. (1995) 'The Boundaries of Languages and Disciplines: How Ideologies Construct Difference', in Mack, A. (ed.) *Defining the Boundaries of Social Inquiry.* (62): Johns Hopkins University Press, pp. 967–1001.

Gal, S. and Irvine, J.T. (2019) *Signs of Difference*: Cambridge University Press.

Gamble, A. (1990) ‘Theories of British Politics’, *Political Studies*, 38(3), pp. 404–420.

Gaskin, C.J. and Happell, B. (2014) ‘On exploratory factor analysis: a review of recent evidence, an assessment of current practice, and recommendations for future use’, *International Journal of Nursing Studies*, 51(3), pp. 511–521.

Gaver, W.W. (1991) ‘Technology Affordances’, *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI ‘91, the SIGCHI conference*, New Orleans, Louisiana, United States, 27 April-2 May 1991. New York: ACM Press, pp. 79–84.

Gaver, W.W. (1996) ‘Affordances for Interaction: The Social Is Material for Design’, *Ecological Psychology*, 8(2), pp. 111–130.

Gayo-Avello, D. (2013) ‘A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data’, *Social Science Computer Review*, 31(6), pp. 649–679.

Geddes, B. (1990) ‘How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics’, *Political Analysis*, 2, pp. 131–150.

Geertz, C. (1973) *The Interpretation of Cultures: Selected Essays*: Basic Books.

Geographic Areas Reference Manual (1994).

Geography and the American Community Survey: What Data Users Need to Know (2020). Washington, DC: US Government Printing Office.

George, A.R. (2003) *The Babylonian Gilgamesh Epic: Introduction, Critical Edition and Cuneiform Texts*. Oxford: Oxford University Press.

Gerber, A.S. *et al.* (2011) ‘Personality Traits and Participation in Political Processes’, *The Journal of Politics*, 73(3), pp. 692–706.

Gerbner, G. (ed.) (1983) *Ferment in the Field* (33).

Gibson, J.J. (1979) *The ecological approach to visual perception*. Dallas: Houghton Mifflin.

Gibson, J.J. (1982) ‘Notes on Affordances’, in Reed, E. and Jones, R. (eds.) *Reasons for Realism: Selected Essays of James J. Gibson.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 401–418.

Gibson, K.R. (1993) ‘Tool Use, Language and Social Behavior in Relationship to Information Processing Capacities’, in Gibson, K.R. and Ingold, T. (eds.) *Tools, language, and cognition in human evolution*, 251–270.

Gibson, R. and Cantijoch, M. (2013) ‘Conceptualizing and Measuring Participation in the Age of the Internet: Is Online Political Engagement Really Different to Offline?’ *Journal of Politics*, 75(3), pp. 701–716.

Gieryn, T.F. (1983) ‘Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists’, *American Sociological Review*, 48(6), p. 781.

Gieryn, T.F. (1999) *Cultural boundaries of science: Credibility on the line*. Chicago: University of Chicago Press.

Gil de Zúñiga, H. *et al.* (2010) ‘Digital Democracy: Reimagining Pathways to Political Participation’, *Journal of Information Technology & Politics*, 7(1), pp. 36–51.

Gimpel, K. *et al.* (2011) ‘Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments’, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics (ACL), Portland, Oregon, 19-24 June, pp. 42–47.

Ginsberg, B. (1982) *The Consequences of Consent: Elections, Citizen Control and Popular Acquiescence*. Reading, MA: Addison-Wesley Publishing.

Gold, R.L. (1997) 'The Ethnographic Method in Sociology', *Qualitative Inquiry*, 3(4), pp. 388–402.

Goldberg, Y. (2017) *Neural network methods for natural language processing*. (Synthesis lectures on human language technologies, 37). San Rafael: Morgan & Claypool Publishers.

Goodin, R.E. and Klingemann, H.-D. (eds.) (1998) *A New Handbook of Political Science*. Oxford: Oxford University Press.

Goodman, L.A. and Kruskal, W.H. (1954) 'Measures of Association for Cross Classifications', *Journal of the American Statistical Association*, 49(268), p. 732.

Goodnight, G.T. (1982) 'The Personal, Technical, and Public Spheres of Argument: A Speculative Inquiry into the Art of Public Deliberation', *Argumentation and Advocacy*, 18(4), pp. 214–227.

Goodnight, G.T. (1987) 'Public discourse', *Critical Studies in Mass Communication*, 4(4), pp. 428–432.

Goretzko, D., Pham, T.T.H. and Bühner, M. (2021) 'Exploratory factor analysis: Current use, methodological developments and recommendations for good practice', *Current Psychology*, 40(7), pp. 3510–3521.

Gorsuch, R.L. (1974) *Factor Analysis*. Philadelphia: Saunders.

Gosnell, H.F. (1926) 'An Experiment in the Stimulation of Voting', *The American Political Science Review*, 20(4), pp. 869–874.

Gosnell, H.F. (1927) *Getting Out the Vote: An Experiment in the Stimulation of Voting*. Chicago: University of Chicago Press.

Gow, D.J. (1985) 'Quantification and Statistics in the Early Years of American Political Science, 1880-1922', *Political Methodology*, 11(1/2), pp. 1–18.

Graham, E.R., Shipan, C.R. and Volden, C. (2014) 'The Communication of Ideas across Subfields in Political Science', *PS: Political Science and Politics*, 47(02), pp. 468–476.

Graham, P. (1999) 'Critical Systems Theory: A Political Economy of Language, Thought, and Technology', *Communication Research*, 26(4), pp. 482–507.

Grice, H.P. (1957) 'Meaning', *The Philosophical Review*, 66(3), pp. 377–388.

Gries, S.T. (2008) 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics*, 13(4), pp. 403–437.

Grieve, J., Nini, A. and Guo, D. (2018) 'Mapping Lexical Innovation on American Social Media', *Journal of English Linguistics*, 46(4), pp. 293–319.

Grossberg, L. (1982) 'Intersubjectivity and the Conceptualization of Communication', *Human Studies*, 5(3), pp. 213–235.

Gumperz, J.J. (1968) 'The Speech Community', in Sills, D.L. and Merton, R.K. (eds.) *International Encyclopedia of the Social Sciences* (19 vols). New York: Macmillan Publishing Company, pp. 381–386.

Gumperz, J.J. (1971) *Language in social groups*. (Language science and national development, vol. 3). Stanford: Stanford University Press.

Gunderson, R. (2016) 'The sociology of technology before the turn to technology', *Technology in Society*, 47, pp. 40–48.

Gunnell, J.G. (2002) 'Handbooks and History: Is It Still the American Science of Politics?' *International Political Science Review*, 23(4), pp. 339–354.

Gunnell, J.G. (2004) 'The Real Revolution in Political Science', *PS: Political Science and Politics*, 37(1), pp. 47–50.

Gunnell, J.G. (2013) 'The reconstitution of political theory: David Easton, behavioralism, and the long road to system', *Journal of the History of the Behavioral Sciences*, 49(2), pp. 190–210.

Gupta, A. and Ferguson, J. (eds.) (1997a) *Anthropological Locations: Boundaries and Grounds of a Field Science*. Berkeley: University of California Press.

Gupta, A. and Ferguson, J. (1997b) 'Discipline and Practice: "The Field" as Site, Method, and Location in Anthropology', in Gupta, A. and Ferguson, J. (eds.) *Anthropological Locations: Boundaries and Grounds of a Field Science.* Berkeley: University of California Press, pp. 1–46.

Gurin, P., Miller, A.H. and Gurin, G. (1980) 'Stratum Identification and Consciousness', *Social Psychology Quarterly*, 43(1), pp. 30–47.

Habermas, J. (1984) *The theory of communicative action*. Reprint, Cambridge: Polity, 1997.

Hall, P.A. (2007) 'The Dilemmas of Contemporary Social Science', *boundary 2*, 34(3), pp. 121–141.

Hall, P.M. (2003) 'Interactionism, Social Organization, and Social Processes: Looking Back and Moving Ahead', *Symbolic Interaction*, 26(1), pp. 33–55.

Halliday, M. and Martin, J. (1993) *Writing science: Literacy and discursive power*. (Critical perspectives on literacy and education). London: Falmer Press.

Halliday, M.A.K. (2005) 'On Matter and Meaning: The Two Realms of Human Experience', *Linguistics and the Human Sciences*, 1(1), pp. 59–82.

Haraway, D. (1988) 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies*, 14(3), p. 575.

Harris, R.A. (1994a) 'The Chomskyan Revolution I: Syntax, Semantics, and Science', *Perspectives on Science*, 2(1), pp. 38–75.

Harris, R.A. (1994b) 'The Chomskyan Revolution II: Sturm und Drang', *Perspectives on Science*, 2(2), pp. 176–230.

Hartigan, J.A. and Wong, M.A. (1979) 'Algorithm AS 136: A K-Means Clustering Algorithm', *Applied Statistics*, 28(1), p. 100.

Hartmann, M. (2006) 'The triple articulation of ICTs: Media as technological objects, symbolic environments and individual texts', in Berker, T. *et al.* (eds.) *Domestication of Media and Technology.* Maidenhead: Open University Press, pp. 80–102.

Hasan, R. (1995) 'On Social Conditions for Semiotic Mediation: The Genesis of Mind in Society', in Sadovnik, A.R. (ed.) *Knowledge and Pedagogy: The Sociology of Basil Bernstein.* (The David C. Anchin series in social and policy issues in education). Norwood, N.J.: Ablex Publishing, pp. 171–196.

Hasan, R. (2005) 'Semiotic Mediation, Language and Society: Three Exotropic Theories—Vygotsky, Halliday and Bernstein', in Hasan, R. *Language, Society and Consciousness,* edited by Jonathan J. Webster (7 vols). London: Equinox (The Collected Works of Ruqaiya Hasan, 1), pp. 130–156.

Hatch, M.J. (1993) 'The Dynamics of Organizational Culture', *Academy of Management Review*, 18(4), pp. 657–693.

Haugen, E. (1966) 'Dialect, Language, Nation', *American Anthropologist*, 68(4), pp. 922–935.

Hauptmann, E. (2006) 'From Opposition to Accommodation: How Rockefeller Foundation Grants Redefined Relations between Political Theory and Social Science in the 1950s', *The American Political Science Review*, 100(4), pp. 643–649.

Hauptmann, E. (2012) 'The Ford Foundation and the rise of behavioralism in political science', *Journal of the History of the Behavioral Sciences*, 48(2), pp. 154–173.

Hauser, M.D., Chomsky, N. and Fitch, W.T. (2002) 'The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?' *Science, Technology and Society*, 298(5598), pp. 1569–1579.

Hawes, L.C. (1974) 'Social collectivities as communication: Perspective on organizational behavior', *Quarterly Journal of Speech*, 60(4), pp. 497–502.

Hempel, C.G. (1965) *Scientific Explanation: Essays in the Philosophy of Science*. New York: Free Press.

Hennig, C. and Liao, T.F. (2013) 'How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification', *Applied Statistics*, 62(3), pp. 309–369.

Herbst, S. (1998) *Reading public opinion: How political actors view the democratic process*. (Studies in communication, media and public opinion). Chicago: University of Chicago Press.

Herva, S. (1988) 'The Genesis of Max Weber's "Verstehende Soziologie"', *Acta Sociologica*, 31(2), pp. 143–156.

Hick, J. (1983) 'On conflicting religious truth–claims', *Religious Studies*, 19(4), pp. 485–491.

Hickman, L.A. (1990) *John Dewey's Pragmatic Technology.* (The Indiana series in the philosophy of technology). Bloomington: Indiana University Press.

Hofmann, T. (1999) 'Probabilistic Latent Semantic Indexing', *Proceedings of SIGIR '99: 22nd international conference on research and development in information retrieval, the 22nd annual international ACM SIGIR conference.* ACM. Special Interest Group on Information Retrieval, Berkeley, California, United States, 15-19 August 1999. New York: ACM Press, pp. 50–57.

Holmes, J. and Meyerhoff, M. (1999) 'The Community of Practice: Theories and methodologies in language and gender research', *Language in Society,* 28(2), pp. 173–183.

Honeycutt, C. and Herring, S.C. (2009) 'Beyond Microblogging: Conversation and Collaboration via Twitter', *2009 42nd Hawaii International Conference on System Sciences, 2009 42nd Hawaii International Conference on System Sciences*, Waikoloa, Hawaii, USA, 5-8 January 2009: IEEE, pp. 1–10.

Hooghe, M., Hosch-Dayican, B. and van Deth, J.W. (2014) 'Conceptualizing political participation', *Acta Politica*, 49(3), pp. 337–348.

Hughes, T.P. (1987) 'The Evolution of Large Technological Systems', in Bijker, W.E., Hughes, T.P. and Pinch, T.J. (eds.) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology.* Cambridge, Massachusetts: MIT Press, pp. 45–76.

Humphrey, N.K. (1976) 'The Social Function of Intellect', in Bateson, P.P.G. and Hinde, R.A. (eds.) *Growing points in ethology.* Cambridge: Cambridge University Press, pp. 303–317.

Huntington, S.P. (1975) 'The United States', in Crozier, M., Huntington, S.P. and Watanuki, J. *The Crisis of Democracy: Report on the Governability of Democracies to the Trilateral Commission.* New York: New York University Press (The Triangle papers, no. 8), pp. 59–118.

Huntington, S.P. and Nelson, J.M. (1976) *No easy choice: Political participation in developing countries*. Cambridge, Mass.: Harvard University Press.

Hustinx, L. and Denk, T. (2009) 'The 'Black Box' Problem in the Study of Participation', *Journal of Civil Society*, 5(3), pp. 209–226.

Hutchby, I. (2001) 'Technologies, Texts and Affordances', *Sociology*, 35(2), pp. 441–456.

Hymes, D. (1974) *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Idreos, S. *et al.* (2012) 'MonetDB: Two Decades of Research in Column-oriented Database Architectures', *Database Engineering Bulletin*, 35(1), pp. 40–45.

Innis, H. (1950) *Empire and Communications*. Oxford: Oxford University Press.

Irvine, J.T. and Gal, S. (2000) 'Language Ideology and Linguistic Differentiation', in Kroskrity, P.V. (ed.) *Regimes of language: Ideologies, polities, and identities.* (School of American Research advanced seminar series). Santa Fe NM: School of American Research Press ; J. Currey.

Isin, E.F. and Ruppert, E.S. (2020) *Being digital citizens*. London: Rowman & Littlefield.

Jackendoff, R. (2002) *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jain, A.K. (2010) 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters*, 31(8), pp. 651–666.

Jasanoff, S. (ed.) (2004) *States of Knowledge: The Co-Production of Science and Social Order*. London: Routledge.

Jenkins, H. and Deuze, M. (2008) 'Convergence Culture', *Convergence: The International Journal of Research into New Media Technologies*, 14(1), pp. 5–12.

Jolly, H.B. (1981) 'Teaching Basic Function Words', *The Reading Teacher*, 35(2), pp. 136–140.

Jones, M.R. (2014) 'A Matter of Life and Death: Exploring Conceptualizations of Sociomateriality in the Context of Critical Care', *MIS Quarterly*, 38(3), pp. 895–925.

Jones, T. (2015) 'Toward a Description of African American Vernacular English Dialect Regions Using "Black Twitter"', *American Speech*, 90(4), pp. 403–440.

Joseph, J.E. (2006) *Language and Politics*. (Edinburgh Textbooks in Applied Linguistics). United Kingdom: Edinburgh University Press.

Jungherr, A. *et al.* (2017) 'Digital Trace Data in the Study of Public Opinion', *Social Science Computer Review*, 35(3), pp. 336–356.

Juola, P. (2006) 'Authorship Attribution', *Foundations and Trends in Information Retrieval*, 1(3), pp. 233–334.

Jurafsky, D. and Martin, J.H. (2009) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd edn. (Prentice Hall series in artificial intelligence). Upper Saddle River, N.J.: Prentice Hall.

Kaase, M. and Marsh, A. (1979) 'Political Action: A Theoretical Perspective', in Barnes, S.H., Kaase, M. and et al. (eds.) *Political Action: Mass Participation in Five Western Democracies.* Beverly Hills, California: SAGE Publications, pp. 27–56.

Kalenda, J. (2016) 'Situational analysis as a framework for interdisciplinary research in the social sciences', *Human Affairs*, 26(3).

Kallinikos, J. (2004) 'Farewell to constructivism: technology and context-embedded action', in Avgerou, C., Ciborra, C. and Land, F. (eds.) *The social study of information and communication technology: Innovation, actors, and contexts.* Oxford: Oxford University Press, pp. 140–161.

Kallinikos, J., Leonardi, P.M. and Nardi, B.A. (2012) 'The Challenge of Materiality: Origins, Scope, and Prospects', in Leonardi, P.M., Nardi, B.A. and Kallinikos, J. (eds.) *Materiality and Organizing: Social Interaction in a Technological World.* Oxford: Oxford University Press, pp. 3–22.

Kant, I. (1781) *Critique of Pure Reason*: Macmillian & Co., 1929.

Kaplan, A. (1964) *The Conduct of Inquiry: Methodology for Behavioural Science*. Somerset: Routledge.

Kaptelinin, V. and Nardi, B.A. (2006) *Acting with technology: Activity theory and interaction design*. (Acting with technology). Cambridge, Mass.: MIT.

Karl, B.D. (1974) *Charles E. Merriam and the study of politics*. Chicago, etc.: University of Chicago Press.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding groups in data: An introduction to cluster analysis*. (Wiley-Interscience paperback series). Reprint, Hoboken, N.J.: Wiley-Interscience, 2005.

Kaufman-Osborn, T.V. (2006) 'Dividing the Domain of Political Science: On the Fetishism of Subfields', *Polity*, 38(1), pp. 41–71.

Kelly, G.A. (1979) 'Who Needs a Theory of Citizenship?' *Daedalus*, 108(4), pp. 21–36.

Kendall, M. (1945) 'The treatment of ties in ranking problems', *Biometrika*, 33, pp. 239–251.

Kermode, F. (2000) *The sense of an ending: Studies in the theory of fiction*. with a new epilogue. Oxford: Oxford University Press.

Kestemont, M. (2014) 'Function Words in Authorship Attribution: From Black Magic to Theory?' *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), 3rd Workshop on Computational Linguistics for Literature (CLFL).* Association for Computational Linguistics (ACL), Gothenburg, Sweden, April, pp. 59–66.

Keyssar, A. (2000) *The right to vote: The contested history of democracy in the United States*. New York: Basic Books.

Kilgarriff, A. (2001) 'Comparing Corpora', *International Journal of Corpus Linguistics*, 6(1), pp. 97–133.

Kilgarriff, A. (2005) 'Language is never, ever, ever, random', *Corpus Linguistics and Linguistic Theory*, 1(2).

Kilgarriff, A. (2009) 'Simple Maths for Keywords', *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK, 20-23 July.

King, G. (1990) 'On Political Methodology', *Political Analysis*, 2, pp. 1–29.

King, G. (1998) *Unifying political methodology: The likelihood theory of statistical inference*. (Techniques in political analysis). Ann Arbor: University of Michigan Press.

King, G., Keohane, R.O. and Verba, S. (1994) *Designing social inquiry: scientific inference in qualitative research*. Princeton, N.J.: Princeton University Press,

King, H.C. (1955) *The History of the Telescope*. Reprint, London: Charles Griffin & Co. Ltd., 1979.

Kline, S.J. (1985) 'What Is Technology?' *Bulletin of Science, Technology & Society*, 5(3), pp. 215–218.

Knoblauch, H. (2013) 'Communicative Constructivism and Mediatization', *Communication Theory*, 23(3), pp. 297–315.

Koppel, M., Schler, J. and Argamon, S. (2009) 'Computational methods in authorship attribution', *Journal of the American Society for Information Science and Technology*, 60(1), pp. 9–26.

Kozulin, A. (1998) *Psychological Tools: A Sociocultural Approach To Education*. Cambridge: Harvard University Press.

Krotz, F. (2017) 'Explaining the Mediatisation Approach', *Javnost - The Public*, 24(2), pp. 103–118.

Krueger, B.S. (2002) 'Assessing the Potential of Internet Political Participation in the United States: A Resource Approach', *American Politics Research*, 30(5), pp. 476–498.

Kuhn, T.S. (1970) *The Structure of Scientific Revolutions*. 2nd edn. Chicago: University of Chicago Press.

Kylmälä, T.P. (2012) 'Medium, the human condition and beyond', *Empedocles: European Journal for the Philosophy of Communication*, 4(2), pp. 133–151.

Labov, W. (1963) 'The Social Motivation of a Sound Change', *Word*, 19(3), pp. 273–309.

Labov, W. (1966) *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, W. (2002) 'Review: Penelope Eckert, Linguistic Variation as Social Practice: Oxford: Blackwell, 2000. Pp. xvi, 240. Hb $62.95, pb $28.95', *Language in Society*, 31(02), pp. 277–284.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998) 'An Introduction to Latent Semantic Analysis', *Discourse Processes*, 25(2–3), pp. 259–284.

Lane, R.E. (1959) *Political Life: Why and How People Get Involved in Politics*. New York: Free Press.

Lass, R. (1997) *Historical linguistics and language change*. (Cambridge studies in linguistics, 81). Cambridge: Cambridge University Press.

Lasswell, H.D. (1968) 'The Future of the Comparative Method', *Comparative Politics*, 1(1), p. 3.

Lasswell, H.D. and Leites, N. (1949) *Language of politics: Studies in quantitative semantics*. (Library of policy sciences). New York: G.W. Stewart.

Latour, B. (2004) 'Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern', *Critical Inquiry*, 30(2), pp. 225–248.

Latour, B. (2005) *Reassembling the social: An introduction to actor-network-theory*. (Clarendon lectures in management studies). Oxford: Clarendon.

Laver, M., Benoit, K. and Garry, J. (2003) 'Extracting Policy Positions from Political Texts Using Words as Data', *The American Political Science Review*, 97(2), pp. 311–331.

Law, J. (1992) 'Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity', *Systems Practice*, 5(4), pp. 379–393.

Law, J. (1999) 'After Ant: Complexity, Naming and Topology', *The Sociological Review*, 47(1_suppl), pp. 1–14.

Law, J. (2004) *After Method: Mess in Social Science Research*. (International library of sociology). London: Routledge.

Law, J. (2009) 'Actor Network Theory and Material Semiotics', in Turner, B.S. (ed.) *The New Blackwell Companion to Social Theory.* (Blackwell companions to sociology). Oxford: Wiley-Blackwell, pp. 141–158.

Law, J. and Mol, A. (1995) 'Notes on Materiality and Sociality', *The Sociological Review*, 43(2), pp. 274–294.

Lazarsfeld, P.F., Berelson, B.R. and Gaudet, H. (1944) *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Duell, Sloan and Pearce.

Lee, E.W. (1995) 'Political Science, Public Administration, and the Rise of the American Administrative State', *Public Administration Review*, 55(6), p. 538.

Lee, S.H. (2017) 'Digital democracy in Asia: The impact of the Asian internet on political participation', *Journal of Information Technology & Politics*, 14(1), pp. 62–82.

Leighley, J.E. (1995) 'Attitudes, Opportunities and Incentives: A Field Essay on Political Participation', *Political Research Quarterly*, 48(1), pp. 181–209.

Lemberg Center (1966-1967) *Lemberg Center Poll: Six City Study on Violence*.

Lenine, E. and Mörschbächer, M. (2020) 'Pesquisa bibliométrica e hierarquias do conhecimento em Ciência Política', *Revista Brasileira de Ciência Política*, (31), pp. 123–160.

Lenski, G. (1994) 'Societal Taxonomies: Mapping the Social Universe', *Annual Review of Sociology*, 20(1), pp. 1–26.

Leonardi, P.M. (2007) 'Activating the Informational Capabilities of Information Technology for Organizational Change', *Organization Science*, 18(5), pp. 813–831.

Leonardi, P.M. (2011) 'When Flexible Routines Meet Flexible Technologies: Affordance, Constraint, and the Imbrication of Human and Material Agencies', *MIS Quarterly*, 35(1), p. 147.

Leonardi, P.M. (2012) 'Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How are They Related? Do We Need Them?' in Leonardi, P.M., Nardi, B.A. and Kallinikos, J. (eds.) *Materiality and Organizing: Social Interaction in a Technological World.* Oxford: Oxford University Press.

Leonardi, P.M. and Barley, S.R. (2010) 'What's Under Construction Here? Social Action, Materiality, and Power in Constructivist Studies of Technology and Organizing', *Academy of Management Annals*, 4(1), pp. 1–51.

Leonardi, P.M., Huysman, M. and Steinfield, C. (2013) 'Enterprise Social Media: Definition, History, and Prospects for the Study of Social Technologies in Organizations', *Journal of Computer-Mediated Communication*, 19(1), pp. 1–19.

Lesk, M. (1996) *The Seven Ages of Information Retrieval*. Ottawa (Occasional Paper 5). Available at: https://archive.ifla.org/VI/5/op/udtop5/udt-op5.pdf.

Lievrouw, L.A. (2002) 'Determination and Contingency in New Media Development: Diffusion of Innovations and Social Shaping of Technology Perspectives', in Lievrouw, L.A. and Livingstone, S.M. (eds.) *Handbook of New Media: Social Shaping and Consequences of ICTs.* London: SAGE, pp. 183–199.

Lievrouw, L.A. (2014) 'Materiality and Media in Communication and Technology Studies: An Unfinished Project', in Gillespie, T., Boczkowski, P.J. and Foot, K.A. (eds.) *Media technologies: Essays on communication, materiality, and society.* (Inside technology). Cambridge, Massachusetts: The MIT Press, pp. 21–51.

Lievrouw, L.A. and Livingstone, S.M. (2006) 'Introduction to the Updated Student Edition', in Lievrouw, L.A. and Livingstone, S.M. (eds.) *Handbook of New Media: Social Shaping and Social Consequences of ICTs.* London: SAGE Publications, pp. 1–14.

Linaa Jensen, J. (2013) 'Political Participation Online: The Replacement and the Mobilisation Hypotheses Revisited', *Scandinavian Political Studies*, 36(4), pp. 347–364.

Lindenberg, S. (1990) 'Homo Socio-oeconomicus: The Emergence of a General Model of Man in the Social Sciences', *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft*, 146(4), pp. 727–748.

Lister, M. *et al.* (eds.) (2009) *New Media: A Critical Introduction*. 2nd edn. London: Routledge.

Livingstone, S. (2007) 'On the material and the symbolic: Silverstone's double articulation of research traditions in new media studies', *New Media & Society*, 9(1), pp. 16–24.

Loader, B.D. and Mercea, D. (2011) 'Networking Democracy?' *Information, Communication & Society*, 14(6), pp. 757–769.

Lomborg, S. (2017) 'A state of flux: Histories of social media research', *European Journal of Communication*, 32(1), pp. 6–15.

Louf, T. *et al.* (2022) *American cultural regions mapped through the lexical analysis of social media*. Available at: https://arxiv.org/pdf/2208.07649.

Love, H. (2002) *Attributing authorship: An introduction*. Cambridge: Cambridge University Press.

Love, R. et al. (2017) 'The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations', *International Journal of Corpus Linguistics*, 22(3), pp. 319–344.

Lowi, T.J. (1992) 'The State in Political Science: How We Became What We Study', *American Political Science Review*, 86(01), pp. 1–7.

Lowi, T.J. (1993) 'A Review of Herbert Simon's Review of My View of the Discipline', *PS: Political Science and Politics*, 26(1), p. 51.

Lowi, T.J. *et al.* (2021) *American Government: Power and purpose*: W.W. Norton.

Luhmann, N. (1986) 'The Autopoiesis of Social Systems', in Geyer, R.F. and van der Zouwen, J. (eds.) *Sociocybernetic Paradoxes: Observation, Control and Evolution of Self-Steering Systems.* London: SAGE, pp. 172–192.

Luhmann, N. (1995) *Social Systems*. Stanford, Calif.: Stanford University Press.

Lukes, S. (1974) *Power: A radical view*. S.l.: Macmillan.

Lupu, N. and Michelitch, K. (2018) 'Advances in Survey Methods for the Developing World', *Annual Review of Political Science*, 21(1), pp. 195–214.

MacKenzie, D.A. and Wajcman, J. (1985) *The Social Shaping of Technology: How the Refrigerator Got its Hum*. Milton Keynes: Open University Press.

MacKenzie, D.A. and Wajcman, J. (1999) *The social shaping of technology*. 2nd edn. Milton Keynes, Eng.: Open University Press.

MacRae, D. (1974) 'When is an Anthropologist?' *RAIN*, (2), p. 4.

Maes, M. and Bischofberger, L. (2015) 'Will the Personalization of Online Social Networks Foster Opinion Polarization?' *SSRN Electronic Journal*.

Mahoney, J. (2010) 'After KKV: The New Methodology of Qualitative Research', *World Politics*, 62(1), pp. 120–147.

Maine, H.S. (1885) *Popular Government: Four Essays*. London: John Murray.

Majchrzak, A. *et al.* (2013) 'The Contradictory Influence of Social Media Affordances on Online Communal Knowledge Sharing', *Journal of Computer-Mediated Communication*, 19(1), pp. 38–55.

Malinowski, B. (1922) *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. London: Routledge & Kegan Paul.

Mandler, G. (2002) 'Origins of the Cognitive ®evolution', *Journal of the History of the Behavioral Sciences*, 38(4), pp. 339–353.

Mannheim, K. (1929) *Ideology and Utopia: An introduction to the sociology of knowledge*. Reprint, S.l.: Routledge & Kegan Paul, 1936 (1972).

Manning, C.D. and Schütze, H. (1999) *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Manza, J. and Brooks, C. (2012) 'How Sociology Lost Public Opinion', *Sociological Theory*, 30(2), pp. 89–113.

Manza, J. and Uggen, C. (2006) *Locked out: Felon disenfranchisement and American democracy*. (Studies in crime and public policy). Oxford: Oxford University Press.

March, J.G. and Olsen, J.P. (1984) 'The New Institutionalism: Organizational Factors in Political Life', *American Political Science Review*, 78(3), pp. 734–749.

Marcinkowski, F. (2014) 'Mediatisation of Politics: Reflections on the State of The Concept', *Javnost - The Public*, 21(2), pp. 5–22.

Marcus, G.E. (1989) 'Imagining the Whole: Ethnography's Contemporary Efforts to Situate Itself', *Critique of Anthropology*, 9(3), pp. 7–30.

Marcus, G.E. (1995) 'Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography', *Annual Review of Anthropology*, 24, pp. 95–117.

Margetts, H.Z. *et al.* (2016) *Political Turbulence: How Social Media Shape Collective Action*: Princeton University Press.

Marien, S., Hooghe, M. and Quintelier, E. (2010) 'Inequalities in Non-Institutionalised forms of Political Participation: A Multi-Level Analysis of 25 Countries', *Political Studies*, 58(1), pp. 187–213.

Marshall, T.H. (1950) *Citizenship and Social Class: and other essays*. Cambridge: Cambridge University Press.

Marvin, D.K. (1966) 'Book Reviews', *Midwest Journal of Political Science*, 10(4), pp. 507–510.

Marx, K. (1867) *Capital: A Critique of Political Economy*. (1).

Marx, K. and Engels, F. (1846) *The German Ideology*. New York: International Publishers.

Marx, L. and Smith, M.R. (1994) 'Introduction', in Smith, M.R. and Marx, L. (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism.* Cambridge, MA: MIT Press, pp. ix–xv.

Maturana, H.R. and Varela, F.J. (1980) *Autopoiesis and cognition: The realization of the living*. Dordrecht: Reidel.

Mauk, M. (2017) *The Political Culture of Authoritarian Regimes*. paper prepared for the workshop "The Legacy of Authoritarian Regimes – Political Culture, Institutions, and Democratisation", ECPR Joint Sessions of Workshops, Nottingham, 25-30 April. Nottingham.

Maxey, C.C. (1925) 'Non-Voting: Causes and Methods of Control by Charles E. Merriam and Harold F. Gosnell.', *Social Forces*, 3(2), pp. 369–370.

McCarthy, D.R. (2013) 'Technology and 'the International' or, How I Learned to Stop Worrying and Love Determinism', *Millennium: Journal of International Studies*, 41(3), pp. 470–490.

McClurg, S.D. (2003) 'Social Networks and Political Participation: The Role of Social Interaction in Explaining Political Participation', *Political Research Quarterly*, 56(4), pp. 449–464.

McCombs, M. (2004) *Setting the Agenda: The Mass Media and Public Opinion*: Blackwell Publishing Inc.

McEnery, T. and Hardie, A. (2012) *Corpus linguistics: Method, theory and practice*. (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press.

McEnery, T. and Wilson, A. (2001) *Corpus linguistics: An introduction.* 2nd edn. (Edinburgh textbooks in empirical linguistics). Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based language studies: An advanced resource book*. London: Routledge.

McGill, R., Tukey, J.W. and Larsen, W.A. (1978) 'Variations of Box Plots', *The American Statistician*, 32(1), p. 12.

McGregor, S.C. (2019) 'Social media as public opinion: How journalists use social media to represent public opinion', *Journalism*, 20(8), pp. 1070–1086.

McGregor, S.C. (2020) '"Taking the Temperature of the Room"', *Public Opinion Quarterly*, 84(S1), pp. 236–256.

McGuire, W.J. (1961) 'Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments', *The Journal of Abnormal and Social Psychology*, 63(2), pp. 326–332.

McLeod, J.M., Kosicki, G.M. and McLeod, D.M. (2010) 'Levels of Analysis and Communication Science', in Berger, C.R., Roloff, M.E. and Roskos-Ewoldsen, D.R. (eds.) *The handbook of communication science,* 2nd edn. Los Angeles: SAGE, pp. 183–202.

McLuhan, M. (1964) *Understanding Media: The Extensions of Man*. New York: McGraw-Hill.

McLuhan, M. (2003) *Understanding Me: Lectures and interviews*. Cambridge MA: MIT Press.

Mead, G.H. (1934) *Mind, Self and Society: from the Standpoint of a Social Behaviorist*. Edited by Charles W. Morris. Chicago: University of Chicago Press.

Mead, M. (1928) *Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilisation*. New York: William Morrow & Co.

Mello, N.N. de (2011) 'A teoria do desenvolvimento político e a questão da ordem e da estabilidade', *Revista de Sociologia e Política*, 19(39), pp. 139–152.

Mellon, J. and Prosser, C. (2017) 'Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users', *Research & Politics*, 4(3), 1-9.

Merriam, C.E. (1922) 'Political Research', *American Political Science Review*, 16(2), pp. 315–321.

Merriam, C.E. and Gosnell, H.F. (1924) *Non-Voting: Causes and Methods of Control*. Chicago: University of Chicago Press.

Merton, R.K. ([1949] 1968) *Social Theory and Social Structure*. New York: The Free Press.

Merton, R.K. (1967) *On Theoretical Sociology: Five Essays, Old and New*. New York: The Free Press.

Meyrowitz, J. (1994) 'Medium Theory', in Crowley, D.J. and Mitchell, D. (eds.) *Communication Theory Today.* Stanford Calif.: Stanford University Press, pp. 50–77.

Mikolov, T. *et al.* (2013) *Efficient Estimation of Word Representations in Vector Space*.

Milbrath, L.W. (1965) *Political Participation*. Chicago: Rand McNally.

Milbrath, L.W. and Goel, M.L. (1977) *Political Participation: How and Why People Get Involved in Politics*. Chicago: Rand McNally.

Miller, A.H. *et al.* (1981) 'Group Consciousness and Political Participation', *American Journal of Political Science*, 25(3), pp. 495–511.

Miller, C.R. (1984) 'Genre as Social Action', *Quarterly Journal of Speech*, 70(2), pp. 151–167.

Milner, R.M. (2013) 'Pop Polyvocality: Internet Memes, Public Participation, and the Occupy Wall Street Movement', *International Journal of Communication*, 7, pp. 2357–2390.

Milroy, J. (2001) 'Language ideologies and the consequences of standardization', *Journal of Sociolinguistics*, 5(4), pp. 530–555.

Milroy, L. (2000) 'Britain and the United States: Two Nations Divided by the Same Language (and Different Language Ideologies)', *Journal of Linguistic Anthropology*, 10(1), pp. 56–89.

Misa, T.J. (1994) 'Retrieving Sociotechnical Change from Technological Determinism', in Smith, M.R. and Marx, L. (eds.) *Does Technology Drive History? The Dilemma of Technological Determinism.* Cambridge, MA: MIT Press, pp. 115–142.

Mislove, A. *et al.* (2011) 'Understanding the Demographics of Twitter Users', *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), pp. 554–557.

Moisl, H. (2011) 'Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora', *Journal of Quantitative Linguistics*, 18(1), pp. 23–52.

Monroe, B.L. and Schrodt, P.A. (2008) 'Introduction to the Special Issue: The Statistical Analysis of Political Text', *Political Analysis*, 16(4), pp. 351–355.

Monroe, K.R. (2004) 'The Chicago School: Forgotten but Not Gone', *Perspectives on Politics*, 2(1), pp. 95–98.

Monroe, K.R. (ed.) (2005) *Perestroika!: The Raucous Rebellion in Political Science*: Yale University Press.

Monroe, K.R. (2007) 'The Perestroika Movement, Its Methodological Concerns, And The Professional Implications Of These Methodological Issues'.

Morgan, L.H. (1877) *Ancient Society: Researches in the Lines of Human Progress from Savagery through Barbarism to Civilization*. Chicago: Charles H. Kerr and Co.

Mumford, L. (1961) 'History: Neglected Clue to Technological Change', *Technology and Culture*, 2(3), p. 230.

Mumford, L. (1967) *The Myth of the Machine: Technics and Human Development*. New York: Harcourt, Brace & World.

Mumford, L. (1970) *The Myth of the Machine: The Pentagon of Power* (2 vols). (2). New York: Harcourt Brace Jovanovich.

Mumford, L. (2014) 'Tool Users vs. Homo Sapiens and the Megamachine', in Scharff, R.C. and Dusek, V. (eds.) *Philosophy of technology: The technological condition : an anthology,* 2nd edn. (Blackwell philosophy anthologies, 33). Malden, MA: Wiley Blackwell.

Murru, M.F. (2018) 'Enhanced Inter-visibility: The Experience of Civic Engagement in Social Media', in Wimmer, J. *et al.* (eds.) *(Mis)understanding Political Participation: Digital Practices, New Forms of Participation and the Renewal of Democracy.* (Routledge studies in European communication research and education). London: Routledge, pp. 57–77.

Mutch, A. (2013) 'Sociomateriality — Taking the wrong turning?' *Information and Organization*, 23(1), pp. 28–40.

Nagel, J.H. (1987) *Participation*. (Prentice-Hall foundations of modern political science series). Englewood Cliffs: Prentice-Hall ; London : Prentice-Hall International.

Nardi, B.A. (1996) *Context and consciousness: Activity theory and human-computer interaction*. Cambridge, Mass.: MIT Press.

Nartey, M. and Mwinlaaru, I.N. (2019) 'Towards a decade of synergising corpus linguistics and critical discourse analysis: a meta-analysis', *Corpora*, 14(2), pp. 203–235.

National Opinion Research Center (NORC) (no date) *About the GSS*. Available at: http://gss.norc.org/About-The-GSS.

Neblo, M.A. *et al.* (2010) 'Who Wants To Deliberate—And Why?' *The American Political Science Review*, 104(3), pp. 566–583.

Neumann, C.B. and Neumann, I.B. (2015) 'Uses of the Self: Two Ways of Thinking about Scholarly Situatedness and Method', *Millennium: Journal of International Studies*, 43(3), pp. 798–819.

Nguyen, D. *et al.* (2016) 'Computational Sociolinguistics: A Survey', *Computational Linguistics*, 42(3), pp. 537–593.

Nie, N.H. *et al.* (1988) 'Participation in America: Continuity and Change'. Paper presented at the annual meeting of the Midwest Political Science Association.

Nightingale, A.J. (2016) 'Adaptive scholarship and situated knowledges? Hybrid methodologies and plural epistemologies in climate change adaptation research', *Area*, 48(1), pp. 41–47.

Nilep, C. (2006) *"Code Switching" in Sociocultural Linguistics* (Colorado Research in Linguistics vol. 19).

Nini, A. (2019) 'The Multi-Dimensional Analysis Tagger', in Berber Sardinha, T. and Veirano Pinto, M. (eds.) *Multi-dimensional analysis: Research methods and current issues.* London, UK: Bloomsbury Academic, pp. 67–94.

Noether, G.E. (1981) 'Why Kendall Tau?' *Teaching Statistics*, 3(2), pp. 41–43.

Norman, D.A. (1988) *The Design of Everyday Things*. Cambridge, Massachusetts: MIT Press.

Norman, D.A. (1993) *Things that make us smart: Defending human attributes in the age of the machine*. Reading, Mass.: Addison-Wesley Pub. Co.

Norman, D.A. (2007) *The Design of Future Things*. New York: Basic Books.

Norris, P. (1997) 'Towards a more cosmopolitan political science?' *European Journal of Political Research*, 31(1), pp. 17–34.

Norris, P. (2002) *Democratic Phoenix: Reinventing Political Activism*. Cambridge: Cambridge University Press.

Norris, S.P. and Phillips, L.M. (2003) 'How literacy in its fundamental sense is central to scientific literacy', *Science Education*, 87(2), pp. 224–240.

Nwafor, K.A. *et al.* (2013) 'Social Media and Political Participation in Africa: Issues, Challenges and Prospects', in Des Wilson (ed.) *Communication and the new media in Nigeria: Social engagements, political development and public discourse.* Nigeria: African Council on Communication Education, Nigerian Chapter, pp. 64–84.

Ochs, E. and Schieffelin, B.B. (2017) 'Language Socialization: An Historical Overview', in Duff, P.A. and May, S. (eds.) *Language Socialization.* Cham: Springer International Publishing, pp. 1–14.

Ogburn, W.F. (1922) *Social Change with Respect to Culture and Original Nature*. New York: B.W. Huebsch, Inc.

Ohme, J. (2018) 'Updating citizenship? The effects of digital media use on citizenship understanding and political participation', *Information, Communication & Society*, 2, pp. 1–26.

Ohme, J., de Vreese, C.H. and Albæk, E. (2017) 'From theory to practice: how to apply van Deth's conceptual map in empirical political participation research', *Acta Politica*.

Orlikowski, W.J. (1992) 'The Duality of Technology: Rethinking the Concept of Technology in Organizations', *Organization Science*, 3(3), pp. 398–427.

Orlikowski, W.J. (2000) 'Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations', *Organization Science*, 11(4), pp. 404–428.

Orlikowski, W.J. (2007) 'Sociomaterial Practices: Exploring Technology at Work', *Organization Studies*, 28(9), pp. 1435–1448.

Orlikowski, W.J. and Robey, D. (1991) 'Information Technology and the Structuring of Organizations', *Information Systems Research*, 2(2), pp. 143–169.

Orlikowski, W.J. and Scott, S.V. (2008) 'Sociomateriality: Challenging the Separation of Technology, Work and Organization', *The Academy of Management Annals*, 2(1), pp. 433–474.

Osborne, J.W. (2015) 'What is Rotating in Exploratory Factor Analysis?' *Practical Assessment, Research, and Evaluation*, 20.

Oser, J. and Hooghe, M. (2018) 'Democratic ideals and levels of political participation: The role of political and social conceptualisations of democracy', *The British Journal of Politics and International Relations*, 20(3), pp. 711–730.

Osterbrock, D.E., Gwinn, J.A. and Brashear, R.S. (1993) 'Edwin Hubble and the Expanding Universe', *Scientific American*, 269(1), pp. 84–89.

Östman, J.-O. and Simon-Vandenbergen, A.-M. (2009) 'Firthian Linguistics', in Senft, G., Östman, J.-O. and Verschueren, J. (eds.) *Culture and Language Use.* (Handbook of pragmatics highlights, 1877-654X, v. 2). Amsterdam: John Benjamins Pub. Co, pp. 140–145.

Ott, B.L. (2017) 'The age of Twitter: Donald J. Trump and the politics of debasement', *Critical Studies in Media Communication*, 34(1), pp. 59–68.

Owoputi, O. *et al.* (2013) 'Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters', *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Conference of the North American Chapter of the Association for Computational Linguistics.* NAACL-HLT, Atlanta, Georgia, USA, 9-14 June: North American Chapter of the Association for Computational Linguistics (NAACL), pp. 380–390.

Papacharissi, Z. (2010) *A Private Sphere: Democracy in a Digital Age.* (Digital media and society). Cambridge UK: Polity.

Parkinson, C.N. (1957) *Parkinson's Law: and other studies in administration*. Cambridge, MA: Riverside Press.

Parry, G., Moyser, G. and Day, N. (1992) *Political participation and democracy in Britain*. Cambridge: Cambridge University Press.

Parsons, T. and Shils, E.A. (eds.) (1951) *Toward a General Theory of Action*. Cambridge, Massachusetts: Harvard University Press.

Passonneau, R.J. *et al.* (2014) 'Biber Redux: Reconsidering Dimensions of Variation in American English', *The 25th International Conference on Computational Linguistics: Proceedings of COLING 2014, International Conference on Computational Linguistics.* COLING, Dublin, Ireland, 23-29 August, pp. 565–576.

Pattie, C., Seyd, P. and Whiteley, P. (2004) *Citizenship in Britain: Values, Participation and Democracy*. New York: Cambridge University Press.

Pennington, J., Socher, R. and Manning, C.D. (2014) 'GloVe: Global Vectors for Word Representation', *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Conference on Empirical Methods in Natural Language Processing (EMNLP).* EMNLP, Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), pp. 1532–1543.

Pentland, B.T. and Singh, H. (2012) 'Materiality: What Are the Consequences', in Leonardi, P.M., Nardi, B.A. and Kallinikos, J. (eds.) *Materiality and Organizing: Social Interaction in a Technological World.* Oxford: Oxford University Press, pp. 287–295.

Perrow, C. (1967) 'A Framework for the Comparative Analysis of Organizations', *American Sociological Review*, 32(2), pp. 194–208.

Pickering, A. (1995) *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Pickering, A. (2001) 'Practice and Posthumanism: Social Theory and a History of Agency', in Knorr-Cetina, K., Savigny, E.v. and Schatzki, T.R. (eds.) *The practice turn in contemporary theory.* London: Routledge,

Pierson, P. and Skocpol, T. (eds.) (2007) *The transformation of American politics: Activist government and the rise of conservatism*. Princeton, N.J.: Princeton University Press (Princeton studies in American politics).

Pike, K.L. (1954) *Language: in Relation to a Unified Theory of the Structure of Human Behavior.* (Part I - Preliminary Edition). Ann Arbor, MI: Summer Institute of Linguistics.

Pinch, T.J. and Bijker, W.E. (1984) 'The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other', *Social Studies of Science*, 14(3), pp. 399–441.

Pinch, T.J. and Bijker, W.E. (1987) 'The Social Construction of Facts and Artifacts: Or How the Sociology ofScience and the Sociology of Technology Might Benefit Each Other', in Bijker, W.E., Hughes, T.P. and Pinch, T.J. (eds.) *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology.* Cambridge, Massachusetts: MIT Press, pp. 17–50.

Pocock, R. (1915) *The Cheerful Blackguard*. Indianapolis: Bobbs-Merrill.

Pooley, J.D. (2016) 'A "Not Particularly Felicitous" Phrase: A History of the "Behavioral Sciences" Label', *Serendipities. Journal for the Sociology and History of the Social Sciences*, 1(1).

Preston, P. (2001) *Reshaping communications: Technology, information and social change*. London: SAGE.

Putnam, R. (2000) *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Putnam, R.D. (ed.) (2002) *Democracies in flux: The evolution of social capital in contemporary society*. Oxford: Oxford University Press.

Pye, L.W. (ed.) (1963) *Communications and Political Development*. Princeton: Princeton University Press (Studies in Political Development, 1).

Pye, L.W. (1966) *Aspects of Political Development: An Analytical Study*. Boston: Little, Brown and Co.

Quintelier, E. and van Deth, J.W. (2014) 'Supporting Democracy: Political Participation and Political Attitudes. Exploring Causality using Panel Data', *Political Studies*, 62(1_suppl), pp. 153–171.

Rackham, H. (1932) *Aristotle: Politics*. (Loeb Classical Library). London: William Heinemann Ltd.

Reddy, M.J. (1993) 'The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language', in Ortony, A. (ed.) *Metaphor and thought*, 2nd edn. Cambridge: Cambridge University Press, pp. 164–201.

Reid, R.A. and Curry, T.A. (2019) 'Are We There Yet? Addressing Diversity in Political Science Subfields', *PS: Political Science and Politics*, 52(2), pp. 281–286.

Reiter, D. (2015) 'Should We Leave Behind the Subfield of International Relations?' *Annual Review of Political Science*, 18(1), pp. 481–499.

Resnick, L.B., Levine, J.M. and Teasley, S.D. (eds.) (1991) *Perspectives on socially shared cognition: Conference entitled "Socially shared cognition" : Revised papers*. Arlington, VA: American Psychological Association.

Rhodes, R.A.W. (2017) 'What is New About the 'Interpretive Turn' and Why Does it Matter?' in Rhodes, R.A.W. (ed.) *Interpretive political science.* (Selected essays, Volume II). Oxford: Oxford University Press, pp. 207–226.

Rice, S.A. (1925) 'The Behavior of Legislative Groups: A Method of Measurement', *Political Science Quarterly*, 40(1), p. 60.

Riemer, K. and Richter, A. (2010) 'Tweet Inside: Microblogging in a Corporate Context', *BLED 2010 Proceedings*.

Riffe, D., Lacy, S. and Fico, F. (1998) *Analyzing media messages: Using quantitative content analysis in research*. (LEA's communications series). Mahwah, N.J.: Lawrence Erlbaum.

Rockefeller Archive Center (n.d.) *Social Science Research Council records, 1923-1998: Summary*. Available at: https://dimes.rockarch.org/collections/iNo7dbyWw2GwSwKsC3nDj3.

Rockoff, L. (2017) *The language of SQL*. Hoboken NJ: Addison-Wesley.

Rogers, A., Kovaleva, O. and Rumshisky, A. (2020) 'A Primer in BERTology: What We Know About How BERT Works', *Transactions of the Association for Computational Linguistics*, 8, pp. 842–866.

Rogers, D.W. (ed.) (1992) *Voting and the spirit of American democracy: Essays on the history of voting and voting rights in America*. Urbana: University of Illinois Press.

Rogers, E.M. (1962) *Diffusion of Innovations*. 5th edn. Reprint, New York, N.Y.: Simon & Schuster, 2003.

Rorty, R. (1980a) *Philosophy and the mirror of nature*. 2nd edn. (Princeton paperbacks). Princeton, N.J.: Princeton Univ. P.

Rorty, R. (1980b) 'Pragmatism, Relativism, and Irrationalism', *Proceedings and Address of the American Philosophical Association*, 53(6), pp. 719–738.

Rosamond, B. (2007) 'European integration and the social science of EU studies: the disciplinary politics of a subfield', *International Affairs*, 83(2), pp. 231–252.

Rosenstone, S. and Hansen, J. (1993) *Mobilization, Participation, and American Democracy*. New York: Macmillan.

Rowland, R.C. (ed.) (2012) *Spheres of Argument: 30 Years of Goodnight's Influence* (48).

Russell, S. and Williams, R. (2002) 'Social Shaping of Technology Concepts: Frameworks, Findings and Implications for Policy'. with Glossary of Social Shaping, in Sørensen, K. and Williams, R. (eds.) *Shaping Technology, Guiding Policy: Concepts, Spaces and Tools.* Cheltenham, UK: Edward Elgar, pp. 35–111.

Ryle, G. (1949) *The Concept of Mind*. London: Hutchinson.

Sanderson, M. and Croft, W.B. (2012) 'The History of Information Retrieval Research', *Proceedings of the IEEE*, 100(Special Centennial Issue), pp. 1444–1451.

Sartori, G. (2004) 'Where Is Political Science Going?' *PS: Political Science and Politics*, 37(4), pp. 785–787.

Sass, D.A. and Schmitt, T.A. (2010) 'A Comparative Investigation of Rotation Criteria Within Exploratory Factor Analysis', *Multivariate Behavioral Research*, 45(1), pp. 73–103.

Savický, P. and Hlaváčová, J. (2002) 'Measures of Word Commonness', *Journal of Quantitative Linguistics*, 9(3), pp. 215–231.

Scharff, R.C. and Dusek, V. (2014) 'Human Beings as "Makers" or "Tool-Users"?' in Scharff, R.C. and Dusek, V. (eds.) *Philosophy of technology: The technological condition : an anthology,* 2nd edn. (Blackwell philosophy anthologies, 33). Malden, MA: Wiley Blackwell, pp. 377–380.

Scheufele, D.A. and Eveland Jr., W.P. (2001) 'Perceptions of 'Public Opinion' and 'Public' Opinion Expression', *International Journal of Public Opinion Research*, 13(1), pp. 25–44.

Schieffelin, B.B. and Ochs, E. (1986) 'Language Socialization', *Annual Review of Anthropology*, 15(1), pp. 163–191.

Schlozman, K.L., Brady, H.E. and Verba, S. (2018) *Unequal and unrepresented: Political inequality and the people's voice in the new gilded age*. Princeton: Princeton University Press.

Schlozman, K.L., Verba, S. and Brady, H.E. (2012) *The unheavenly chorus: Unequal political voice and the broken promise of American democracy*. Princeton: Princeton University Press.

Schmidt, V.A. (2010) 'Taking ideas and discourse seriously: explaining change through discursive institutionalism as the fourth 'new institutionalism'', *European Political Science Review*, 2(01), p. 1.

Schmidt, V.A. (2017) 'Theorizing Ideas and Discourse in Political Science: Intersubjectivity, Neo-Institutionalisms, and the Power of Ideas', *Critical Review*, 29(2), pp. 248–263.

Schmitt, T.A. and Sass, D.A. (2011) 'Rotation Criteria and Hypothesis Testing for Exploratory Factor Analysis: Implications for Factor Pattern Loadings and Interfactor Correlations', *Educational and Psychological Measurement*, 71(1), pp. 95–113.

Schramm, W. (1954) *Process and effects of mass communication*. Urbana: University of Illinois Press.

Schultze, U. (2014) 'Performing embodied identity in virtual worlds', *European Journal of Information Systems*, 23(1), pp. 84–95.

Schwartz, J.D. (1984) 'Participation and Multisubjective Understanding: An Interpretivist Approach to the Study of Political Participation', *Journal of Politics*, 46(4), pp. 1117–1141.

Searle, J.R. (1969) *Speech Acts: An essay in the philosophy of language*. Cambridge: University Press.

Searle, J.R. (1983) *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Sedelow, W.A. (1967) *Computational Sociolinguistics*. Chapel Hill, NC (Research Previews vol. 14, no. 2). Available at: http://files.eric.ed.gov/fulltext/ED019651.pdf.

Seidelman, R. (1985) *Disenchanted realists: Political science and the American crisis, 1884-1984*. (SUNY series in political theory. Contemporary issues). Albany: State University of New York Press.

Shannon, C.E. and Weaver, W. (1949) *A Mathematical Theory of Communication*. 10th edn. (27): University of Illinois Press, 1964.

Shiviah (1969) 'The Behavioural Approach in Political Science: An Essay on the Meaning and Orientation of a Movement', *The Indian Journal of Political Science*, 30(1), pp. 50–67.

Shklar, J.N. (1991) *American citizenship: The quest for inclusion*. Cambridge, MA: Harvard University Press.

Sigelman, L. (2006) 'The Coevolution of American Political Science and the "American Political Science Review"', *The American Political Science Review*, 100(4), pp. 463–478.

Silverstone, R. (1994) *Television and everyday life*. London: Routledge.

Silverstone, R. (1999) *Why study the media?* London: SAGE Publications Ltd.

Silverstone, R. and Haddon, L. (1996) 'Design and the domestication of information and communication technologies: technical change and everyday life', in Mansell, R.E. and Silverstone, R. (eds.) *Communication by design: The politics of communication technologies.* Oxford: Oxford University Press, pp. 44–74.

Silverstone, R., Hirsch, E. and Morley, D. (1992) 'Information and Communication Technologies and the Moral Economy of the Household', in Silverstone, R. and Hirsch, E. (eds.) *Consuming technologies: Media and information in domestic spaces.* London: Routledge, pp. 15–31.

Simon, H.A. (1957) *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*: Wiley.

Simon, H.A. (1985) 'Human Nature in Politics: The Dialogue of Psychology with Political Science', *American Political Science Review*, 79(2), pp. 293–304.

Simon, H.A. (1993) 'The State of American Political Science: Professor Lowi's View of Our Discipline', *PS: Political Science and Politics*, 26(1), pp. 49–51.

Sismondo, S. (1993) 'Some Social Constructions', *Social Studies of Science*, 23(3), pp. 515–553.

Sismondo, S. (2007) 'Science and Technology Studies and an Engaged Program', in Hackett, E.J. (ed.) *The Handbook of Science and Technology Studies,* 3rd edn. Cambridge, Mass.: MIT, pp. 13–31.

Sismondo, S. (2015) 'Ontological turns, turnoffs and roundabouts', *Social Studies of Science*, 45(3), pp. 441–448.

Skinner, B. (1938) *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century-Crofts.

Smith, M.R. and Marx, L. (eds.) (1994) *Does Technology Drive History? The Dilemma of Technological Determinism*. Cambridge, MA: MIT Press.

Smith, R.M. (1997) 'Still Blowing in the Wind: The American Quest for a Democratic, Scientific Political Science', *Daedalus*, 126(1), pp. 253–287.

Soulié, C. (1995) 'Anatomie du goût philosophique', *Actes de la Recherche en Sciences Sociales*, 109(1), pp. 3–28.

Sovacool, B.K. and Hess, D.J. (2017) 'Ordering theories: Typologies and conceptual frameworks for sociotechnical change', *Social Studies of Science*, 47(5), pp. 703–750.

Srivastava, L. (2004) 'Japan's ubiquitous mobile information society', *info*, 6(4), pp. 234–251.

Stanyer, J. (2005) 'The British public and political attitude expression: the emergence of a self-expressive political culture?' *Contemporary Politics*, 11(1), pp. 19–32.

Star, S.L. (2010) 'This is Not a Boundary Object: Reflections on the Origin of a Concept', *Science, Technology, & Human Values*, 35(5), pp. 601–617.

Star, S.L. and Griesemer, J.R. (1989) 'Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39', *Social Studies of Science*, 19(3), pp. 387–420.

Stichweh, R. (2016) 'Systems Theory as an Alternative to Action Theory? The Rise of 'Communication' as a Theoretical Option', *Acta Sociologica*, 43(1), pp. 5–13.

Stocking, G.W., Jr. (1992) *The Ethnographer's Magic and Other Essays in the History of Anthropology*. Madison, Wisconsin: University of Wisconsin Press.

Stoffregen, T.A. (2003) 'Affordances as Properties of the Animal-Environment System', *Ecological Psychology*, 15(2), pp. 115–134.

Stolle, D., Hooghe, M. and Micheletti, M. (2005) 'Politics in the Supermarket: Political Consumerism as a Form of Political Participation', *International Political Science Review*, 26(3), pp. 245–269.

Stolle, D. and Howard, M.M. (2008) 'Civic Engagement and Civic Attitudes in Cross-National Perspective: Introduction to the Symposium', *Political Studies*, 56(1), pp. 1–11.

Stone, P.J. *et al.* (1962) 'The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information', *Behavioral Science*, 7(4), pp. 484–498.

Strauch, R.E. (1976) 'Critical Look at Quantitative Methodology', *Policy Analysis*, 2(1), pp. 121–144.

Stubbs, M. (2010) 'Three concepts of keywords', in Bondi, M. and Scott, M. (eds.) *Keyness in texts.* (Studies in corpus linguistics, v. 41). Amsterdam: John Benjamins Pub. Co, pp. 21–42.

Suchman, L.A. (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd edn. Cambridge: Cambridge University Press.

Sutton, R.I. and Staw, B.M. (1995) 'What Theory is Not', *Administrative Science Quarterly*, 40(3), pp. 371–384.

Swales, J.M. (1990) *Genre analysis: English in academic and research settings*. (The Cambridge applied linguistics series). Cambridge: Cambridge University Press.

Swales, J.M. and Feak, C.B. (2012) *Academic writing for graduate students: Essential skills and tasks*. 3rd edn. (Michigan series in English for academic & professional purposes). Ann Arbor, Mich.: University of Michigan Press.

Sylvan, D.J. (1991) 'The Qualitative-Quantitative Distinction in Political Science', *Poetics Today*, 12(2), p. 267.

Tam, H.B. (2019) *The evolution of communitarian ideas: History, theory and practice*. Cham: Palgrave Macmillan.

Tausczik, Y.R. and Pennebaker, J.W. (2010) 'The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods', *Journal of Language and Social Psychology*, 29(1), pp. 24–54.

Taylor, J.R. *et al.* (2001) *The Computerization of Work: A Communication Perspective*. Thousand Oaks, Calif.: SAGE.

Teorell, J., Torcal, M. and Montero, J.R. (2006) 'Political Participation: Mapping the Terrain', in van Deth, J.W., Montero, J.R. and Westholm, A. (eds.) *Citizenship and involvement in European democracies: A comparative analysis.* (Routledge research in comparative politics, 17). London: Routledge.

Teubert, W. (2005) 'My version of corpus linguistics', *International Journal of Corpus Linguistics*, 10(1), pp. 1–13.

Theiss-Morse, E. (1993) 'Conceptualizations of good citizenship and political participation', *Political Behavior*, 15(4), pp. 355–380.

Theocharis, Y. (2015) 'The Conceptualization of Digitally Networked Participation', *Social Media + Society*, 1(2), 1-14.

Theocharis, Y., Moor, J. and van Deth, J.W. (2021) 'Digitally Networked Participation and Lifestyle Politics as New Modes of Political Participation', *Policy & Internet*, 13(1), pp. 30–53.

Theocharis, Y. and van Deth, J.W. (2015) 'The Curious Case of Digitally Networked Participation: Conceptualizing and Measuring Digitally Enabled Political Participation', *SSRN Electronic Journal*.

Theocharis, Y. and van Deth, J.W. (2018) 'The continuous expansion of citizen participation: a new taxonomy', *European Political Science Review*, 10(01), pp. 139–163.

Thompson, J.B. (1995) *The Media and Modernity: A Social Theory of the Media*. Cambridge: Polity.

Thornton, R.J. (1985) ''Imagine Yourself Set Down…': Mach, Frazer, Conrad, Malinowski and the Role of Imagination in Ethnography', *Anthropology Today*, 1(5), pp. 7–14.

Tilly, C. (1975) 'Reflections on the History of European State-Making', in Tilly, C. (ed.) *The Formation of National States in Western Europe.* (Studies in Political Development, 8). Princeton: Princeton University Press, pp. 3–83.

Tingsten, H. (1937) *Political Behavior: Studies in Election Statistics*. London: P.S. King.

Tocqueville, A. de (1835) *Democracy in America*. (1).

Törnberg, P. and Törnberg, A. (2018) 'The Limits of Computation: A Philosophical Critique of Contemporary Big Data Research', *Big Data & Society*, 5(2), 1-12.

Toscano, A. (2008) 'The Culture of Abstraction', *Theory, Culture & Society*, 25(4), pp. 57–75.

Toutanova, K. *et al.* (2003) 'Feature-rich part-of-speech tagging with a cyclic dependency network', *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, Morristown, NJ, USA. Morristown, NJ, USA: Association for Computational Linguistics.

Treem, J.W. *et al.* (2016) 'What We Are Talking About When We Talk About Social Media: A Framework for Study', *Sociology Compass*, 10(9), pp. 768–784.

Treem, J.W. and Leonardi, P.M. (2013) 'Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association', *Annals of the International Communication Association*, 36(1), pp. 143–189.

Trudgill, P. (1974) *The social differentiation of English in Norwich*. (Cambridge studies in linguistics, 0068-6794, 13). London: Cambridge University Press.

Tufekci, Z. (2014a) 'Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls', *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14).* International Conference on Weblogs and Social Media (ICWSM), Ann Arbor, Michigan, 1-4 June.

Tufekci, Z. (2014b) 'The Medium and the Movement: Digital Tools, Social Movement Politics, and the End of the Free Rider Problem', *Policy & Internet*, 6(2), pp. 202–208.

Tukey, J.W. (1962) 'The Future of Data Analysis', *The Annals of Mathematical Statistics*, 33(1), pp. 1–67.

Tukey, J.W. (1977) *Exploratory Data Analysis*. S.l.: Addison Wesley.

Turkle, S. (1995) *Life on the screen: Identity in the age of the Internet*. New York: Simon and Schuster.

Tversky, A. and Kahneman, D. (1973) 'Availability: A heuristic for judging frequency and probability', *Cognitive Psychology*, 5(2), pp. 207–232.

Uldam, J. and Kaun, A. (2018) 'Towards a Framework for Studying Political Participation in Social Media', in Wimmer, J. *et al.* (eds.) *(Mis)understanding Political Participation: Digital Practices, New Forms of Participation and the Renewal of Democracy.* (Routledge studies in European communication research and education). London: Routledge, pp. 181–195.

Valenzuela, S. *et al.* (2019) 'The Paradox of Participation Versus Misinformation: Social Media, Political Engagement, and the Spread of Misinformation', *Digital Journalism*, 7(6), pp. 802–823.

Valeriani, A. and Vaccari, C. (2016) 'Accidental exposure to politics on social media as online participation equalizer in Germany, Italy, and the United Kingdom', *New Media & Society*, 18(9), pp. 1857–1874.

Valles, J.M. and Newton, K. (eds.) (1991) *Political Science in Western Europe, 1960-1990*: Kluwer Academic Publishers (20).

van Deth, J.W. (2001) *Studying Political Participation: Towards a Theory of Everything?* (Joint Sessions of Workshops of the European Consortium for Political Research).

van Deth, J.W. (2014) 'A conceptual map of political participation', *Acta Politica*, 49(3), pp. 349–367.

van Deth, J.W. (2016) 'What is Political Participation?'. *Oxford Research Encyclopedia of Politics*.

van Dijck, J. (2011) 'Tracing Twitter: The rise of a microblogging platform', *International Journal of Media & Cultural Politics*, 7(3), pp. 333–348.

van Dijck, J. (2013) *The Culture of Connectivity: A Critical History of Social Media*: Oxford University Press.

van Dijk, J. (2020) *The Digital Divide*. Cambridge: Polity.

van Dijk, J.A. (1999) *The Network Society*. 3rd edn.: SAGE, 2012.

van Dijk, J.A. and Hacker, K.L. (2018) *Internet and democracy in the network society*. (Routledge studies in global information, politics and society). London: Routledge.

van Gunsteren, H.R. (1998) *A theory of citizenship: Organizing plurality in contemporary democracies*. Boulder, Colo.: Westview.

van Heur, B., Leydesdorff, L. and Wyatt, S. (2013) 'Turning to ontology in STS? Turning to STS through 'ontology'', *Social Studies of Science*, 43(3), pp. 341–362.

Vandenberghe, F. (2002) 'Reconstructing Humants: A Humanist Critique of Actant- Network Theory', *Theory, Culture & Society*, 19(5/6), pp. 51–67.

Verba, S. (1967) 'Democratic Participation', *The Annals of the American Academy of Political and Social Science*, 373(1), pp. 53–78.

Verba, S. and Nie, N.H. (1972) *Participation in America: Political democracy and social equality*. New York: Harper and Row.

Verba, S., Nie, N.H. and Kim, J.-O. (1978) *Participation and Political Equality: A Seven-Nation Comparison*: Cambridge University Press.

Verba, S., Schlozman, K.L. and Brady, H.E. (1995) *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge, Massachusetts: Harvard University Press.

Vidgen, B. *et al.* (2019) 'Challenges and frontiers in abusive content detection', *Proceedings of the Third Workshop on Abusive Language Online, Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 80–93.

Voget, F.W. (1975) *A history of ethnology*. New York: Holt.

Volkoff, O. and Strong, D.M. (2013) 'Critical Realism and Affordances: Theorizing IT-Associated Organizational Change Processes', *MIS Quarterly*, 37(3), pp. 819–834.

Volkoff, O., Strong, D.M. and Elmes, M.B. (2007) 'Technological Embeddedness and Organizational Change', *Organization Science*, 18(5), pp. 832–848.

Volti, R. (2004) 'William F. Ogburn, Social Change with Respect to Culture and Original Nature', *Technology and Culture*, 45(2), pp. 396–405.

von Beyme, K. (2000) *Die politischen Theorien der Gegenwart: Eine Einführung*. 8th edn. Wiesbaden: Westdeutschen Verlag.

von Schoultz, Å. (2015) 'Nordic Research on Political Behaviour', *Scandinavian Political Studies*, 38(4), pp. 342–368.

Wagner, P., Whitley, R. and Wittrock, B. (eds.) (1991) *Discourses on society: the shaping of the social science disciplines*. Dordrecht: Kluwer Academic Publishers (Sociology of the sciences, 15).

Waismel-Manor, I. and Lowi, T.J. (2011) 'Politics in Motion: A Personal History of Political Science', *New Political Science*, 33(1), pp. 59–78.

Walkey, F.H. (1997) 'Composite variable analysis: A simple and transparent alternative to factor analysis', *Personality and Individual Differences*, 22(5), pp. 757–767.

Warner, W. and Hirschberg, J. (2012) 'Detecting hate speech on the world wide web', *LSM '12: Proceedings of the Second Workshop on Language in Social Media.* Association for Computational Linguistics (ACL), pp. 19–26.

Wasserstein, R.L. and Lazar, N.A. (2016) 'The ASA Statement on p -Values: Context, Process, and Purpose', *The American Statistician*, 70(2), pp. 129–133.

Watson, J.B. (1913) 'Psychology as the Behaviorist Views It', *Psychological Review*, 20(2), pp. 158–177.

Wattenberg, M. (2002) *Where Have All the Voters Gone?* Cambridge Massachusetts: Harvard University Press.

Weiner, M. (1971) 'Political Participation: Crisis of the Political Process', in Binder, L., Coleman, J.S., LaPalombara, J., Pye, L.W., Verba, S. and Weiner, M. *Crises and Sequences in Political Development.* Princeton: Princeton University Press (Studies in Political Development, 7), pp. 159–204.

Weller, K. *et al.* (eds.) (2013) *Twitter and society*. New York: Peter Lang (Digital formations 1526-3169, vol. 89).

Whiteley, P. (2012) *Political participation in Britain: The decline and revival of civic culture*. (Contemporary political studies series). New York: Palgrave Macmillan.

Wiebe, R.H. (1995) *Self-rule: A cultural history of American democracy*. Chicago, Ill.: University of Chicago Press.

Wilkerson, J. and Casas, A. (2017) 'Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges', *Annual Review of Political Science*, 20(1), pp. 529–544.

Williams, R. (1974) *Television: Technology and Cultural Form*. London: Fontana.

Williams, R. and Edge, D. (1996) 'The social shaping of technology', *Research Policy*, 25(6), pp. 865–899.

Wilson, S. *et al.* (2016) 'Disentangling Topic Models: A Cross-cultural Analysis of Personal Values through Words', *Proceedings of the First Workshop on NLP and Computational Social Science, Workshop on NLP and Computational Social Science.* Association for Computational Linguistics (ACL), Austin, Texas, 5 November 2016. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 143–152.

Wimmer, J. *et al.* (2018) 'Introduction', in Wimmer, J. *et al.* (eds.) *(Mis)understanding Political Participation: Digital Practices, New Forms of Participation and the Renewal of Democracy.* (Routledge studies in European communication research and education). London: Routledge, pp. 1–13.

Winner, L. (1977) *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, Mass.: M.I.T. Press.

Winner, L. (1980) 'Do Artifacts Have Politics?' *Daedalus*, 109(1), pp. 121–136.

Winner, L. (1993) 'Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology', *Science, Technology, & Human Values*, 18(3), pp. 362–378.

Withagen, R. *et al.* (2012) 'Affordances can invite behavior: Reconsidering the relationship between affordances and agency', *New Ideas in Psychology*, 30(2), pp. 250–258.

Withagen, R., Araújo, D. and Poel, H.J. de (2017) 'Inviting affordances and agency', *New Ideas in Psychology*, 45, pp. 11–18.

Wittgenstein, L. (1953) *Philosophical Investigations*. New York: McMillan.

Wodak, R. (2014) 'Critical Discourse Analysis', in Leung, C. and Street, B.V. (eds.) *The Routledge companion to English studies.* London: Routledge, pp. 302–316.

Wodak, R. and Meyer, M. (eds.) (2001) *Methods of critical discourse analysis*. London: SAGE (Introducing qualitative methods).

Wolfram, W. (2007) 'Sociolinguistic Folklore in the Study of African American English', *Language and Linguistics Compass*, 1(4), pp. 292–313.

Woodward, J. (1958) *Management and Technology*. London: Her Majesty's Stationery Office.

Woolgar, S. (1991) 'The Turn to Technology in Social Studies of Science', *Science, Technology, & Human Values*, 16(1), pp. 20–50.

Woolgar, S. and Lezaun, J. (2013) 'The wrong bin bag: A turn to ontology in science and technology studies?' *Social Studies of Science*, 43(3), pp. 321–340.

Woolley, S.C. and Howard, P.N. (2016) 'Political Communication, Computational Propaganda, and Autonomous Agents — Introduction', *International Journal of Communication*, 10(0), pp. 4882–4890.

Wyatt, S. (2007) 'Technological Determinism is Dead ; Long Live Technological Determinism', in Hackett, E.J. (ed.) *The Handbook of Science and Technology Studies,* 3rd edn. Cambridge, Mass.: MIT, pp. 165–180.

Yanow, D. (2003) 'Interpretive Empirical Political Science: What Makes This Not A Subfield Of Qualitative Methods', *Qualitative & Multi-Method Research*, 1(2).

Yardi, S. and boyd, d. (2010) 'Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter', *Bulletin of Science, Technology & Society*, 30(5), pp. 316–327.

Zhang, J. and Patel, V.L. (2006) 'Distributed cognition, representation, and affordance', *Pragmatics & Cognition*, 14(2), pp. 333–341.

Zhao, X. *et al.* (2018) 'We Agreed to Measure Agreement—Redefining Reliability De-justifies Krippendorff s Alpha', *China Media Research*, 14(2).

Zhao, X., Liu, J.S. and Deng, K. (2013) 'Assumptions behind Intercoder Reliability Indices', *Annals of the International Communication Association*, 36(1), pp. 419–480.

Zukin, C. *et al.* (2006) *A New Engagement? Political Participation, Civic Life, and the Changing American Citizen*. New York: Oxford University Press.