Expressions of Community

Connor Gilroy

05/28/2023

Table of contents

Preface

This page intentionally left blank.

1 Introduction

This dissertation is about how community happens. Structural features – the social density of interactions, the physical density of proximity, the presence and abundance of others with shared group characteristics – create the conditions for the existence of a community. These social features are so essential for community to be perceived and real and as a meso-level social fact that they can be considered part of the definition of the phenomenon.

How individuals outwardly express and subjectively experience community relates to those structural features, and depends on what groups and social identity categories they are members of at all. There are two hypotheses for how the relationship between those aspects of community could play out. First, structural and contextual characteristics could be co-constitutive with expressive and subjective ones, a mutually reinforcing feedback loop. Members of a group who are the most deeply structurally integrated into a group context simultaneously express the greatest subjective investment in it.Alternatively, expressions of community might substitute or stand in for the structural elements that promote togetherness and belonging. For the most embedded group members, community becomes relegated to the background and taken for granted (Zerubavel 2018), a phenomenon of "ambient community" (Brown-Saracino 2017).

This tension is especially relevant for communities organized around marginalized and minoritized identity categories, like LGBTQ identities. The experience of being

an outsider, of not being able to take acceptance and belonging for granted in communities of origin, pushes LGBTQ individuals to seek out community in subcultural, identity-based contexts. At the same time, the preestablished existence of queer spaces and queer collectivities affords opportunities to belong and serves as a positive draw, an attractive force. These push and pull factors do not necessarily coincide in the same places or at the same times, raising the question of which matters more for creating individual experiences of community. Of the two hypotheses, the latter implies heightened salience for community in contexts of stigma, marginalization, scarcity, and lack of access, with those leading to greater attachment to and expression of community. The former hypothesis implies that community is most relevant and most expressed for those who structurally have the most access to it.

1.1 Conceptual background

Community itself is a meso-level social entity that comes into existence through the overlap of social interactional density and shared cultural commonalities – where culture is a broad umbrella encompassing moral values (Tavory 2016; Vaisey 2007), group discursive styles (Eliasoph and Lichterman 2003), and embodied habitus and practices (Lizardo 2017; Orne 2017). What distinguishes community from other similar social phenomena (e.g., "group," "organization," or even "society") is a sense of unity or togetherness, of sharing something in common. Community in this specific sociological sense is also called *Gemeinschaft* (Brint 2001; Tönnies [1887] 2001), a term I will use to emphasize its distinctive social features.

One of those potential commonalities is a shared social identity. Identities are categorical divisions that can be cognitively and socially salient or can be relegated to the background (Blau 1977; White [1965] 2008; Zerubavel 2018). To be a focal point for

community-building, identities must become salient rather than ambient; this may be aided by expressive features that reflect identity such as language and group style (Bourdieu 1991; Eliasoph and Lichterman 2003). For identity to be available to individuals as an opportunity for community and belonging, a sufficient number of others belonging to the identity category must be present, visible, and recognized; this is where contextual features like demographics might come into play.

Identity is not the only organizing principle for community. Shared activities, especially when understood as ritual or other meaningful practices (Brint 2001; Collins 2004; Orne 2017), further develop community and belonging. This might occur in conjunction with shared identity or in place of it.

Shared place provides a third focal point for organizing community. (This remains true even though virtual communities also can exist (Rheingold 2000).) Place, as generally understood in the discipline of geography and in spatial sociology, is meaningful space; place is what happens when a physical and material geographic location is invested with meaning and value (Gieryn 2000). Togetherness in place provides another structural opportunity for interaction and for developing a shared sense of unity in the process; a space becomes place through that collective meaning-making process, and itself becomes an object of attachment. Each of these focal features becomes a vehicle for belonging when it is imbued with meaning.

The question, however, is meaning for whom? In a literal, mundane sense, every individual who shows up and participates is a member of a community. But community is not always internally recognized or conscious, nor always outwardly expressed. Structural integration could straightforwardly be reflected in the behavior and subjectivity of core community members. Alternatively, those who struggle to belong for structural reasons might invest the most in subjective or expressive dimensions of community,

while those who are well-integrated take community for granted.

1.2 The case

Lesbian, gay, bisexual, transgender, queer and other (LGBTQ) sexual and gender minority individuals experience a structural struggle for belonging. In general, they cannot presume acceptance of their identities in communities of origin (Orne 2011, 2013), and they do not automatically have access to queer spaces and communities. At the same time, their individual experiences are varied rather than uniform: greater or lesser acceptance, greater or lesser access. Because of this, recent historical and contemporary LGBTQ communities provide an interesting set of cases to draw on in order to investigate the tension in the expression of community that I outlined above.

The emergence of queer spaces and the communities that inhabit them is improbable and surprising; their continued existence is contingent and fragile (Bérubé 2011; D'Emilio 1992; Ghaziani 2014b). But enough people evidently desire those communities to bring them in into existence and allow them to persist – at first, in spite of stigma and oppression, and still, in spite of normative and assimilatory pressures. Even today, many LGBTQ people want distinctive spaces in which to build queer community and culture (Pew Research Center 2013).

LGBTQ identities are often strongly held and offer a strong basis for possible community formation. LGBTQ identity shapes attitudes toward sexuality, but also beliefs and values more broadly (Schnabel 2018). LGBTQ identities are potentially encompassing enough to constitute a subculture (Fischer 1975; Mattson 2015a) or a counterpublic (Berlant and Warner 1998), leading theorists to write about gay culture (Halperin 2012) or a way of life (Foucault 1998). Structurally speaking, the mere fact that gay/queer/LGBTQ community exists as a possible object of attachment and source

of belonging is itself noteworthy; the identity itself structurally and cognitively provides a potential axis of belonging that is not otherwise available. By contrast, the unmarked category of "heterosexual" is not available as a source of community in the same way – referring, for instance, to "members of the heterosexual community" is nonsensical (Zerubavel 2018), even if many communities do have heteronormativity and heterosexuality as defining or central traits (Eliasoph 1998).

One unique feature of LGBTQ communities is a lack of rootedness, which shapes queer life trajectories: for the most part, LGBTQ people are not born into and do not grow up in queer communities and queer spaces (Weston 1995). This is one reason that the existence and experience of queer community is not something LGBTQ people can necessarily take for granted. Because the potential need for community is visible and salient in LGBTQ contexts, surveys targeted toward LGBTQ people ask about community in detailed and explicit ways, where more general surveys often do not (Meyer 2020). Similarly, this is why some LGBTQ people have sought out virtual communities since those spaces first came into being (Auerbach 2014; Rheingold 2000). Of course, this does not mean that all LGBTQ people experience the salience of community to the same degree. But because queer community is only taken for granted in unusually accepting contexts (Brown-Saracino 2017), it makes an ideal case for observing explicit processes of community.

Queer communities are marked by their diversity and fluidity, by a proliferation of identities, expressions, practices, and ways of being. Due to processes of social sorting, this variation is not observable everywhere; some particular LGBTQ groups and spaces are more homogeneous and exclusionary, some more diverse and welcoming. Regardless, anarchic, chaotic variation is a defining feature of the overarching LGBTQ community as a whole (Brekhus 2003; Brown-Saracino 2017; Lichterman 1999; Mattson 2015b; Orne

2017). At the extreme, some queer theorists, like Berlant and Warner (1998), engage in a theoretical refusal of community and identity and deny that ephemeral and fluid "queer counterpublics" can entail community at all; some queer sociologists disagree and argue that in some ways these liminal queer spaces are the most powerful sites of community-building (Orne 2017). In this work, I will not presume but rather seriously investigate the possibility of community being central for LGBTQ people in their everyday lives.

1.3 Methodology

To study LGBTQ expressions of community, I adopt a quantitative and computational approach. This is an innovation and a departure from most research on the topic, which is qualitative and ethnographic (Baldor 2018; e.g., Brown-Saracino 2017; Orne 2017; Winer 2020). I complement and extend that deeply grounded work by offering breadth and scale instead. Statistical methods are well-suited for uncovering patterns in the relationships between structural features like place demographics and interaction network structures and expressed or reported community-oriented outcomes. Computational text analysis methods extend the reach of quantitative methods and bring them closer to the insights derived from qualitative work (Nelson 2017, 2021); they are especially well-suited for measuring and operationalizing culture and meaning (Arseniev-Koehler and Foster 2022; Mohr et al. 2020). One limitation of these methods is that they are better suited for studying identity and linguistic expressions, rather than activities and embodied practices. The latter are also an essential element of group culture that contributes to community formation (Lizardo 2017; Orne 2017), and this methodological limitation does not diminish their importance. To ensure that my results are robust, I leverage variation and comparison across contexts, triangulating across different cases to build a more complete picture of what features lead to strong experiences of LGBTQ

community.

1.4 Plan of the work

In Chapter 2, I examine the question of whether, when LGBTQ people talk about community, what they mean invokes a *gemeinschaftliche*, belonging- and social organization-oriented sense of the concept to a greater extent than in more general and generic contexts. As a site, I use the the first example of an LGBTQ virtual community, founded in 1983, a Usenet group called soc.motss; the data are the text of individual messages sent to the group, from the late 1990s through the 2000s. For a comparative baseline, I use a pre-trained model based on generic Internet text. In a self-selected virtual community based on LGBTQ identity, if community itself is salient as a topic, then members will visibly talk about community as Gemeinschaft. If community is backgrounded and ambient, other senses of the word will be more evident.

In Chapter 3, I ask whether dense places full of LGBTQ people (and institutions) facilitate a greater sense of connection to the LGBTQ community. Or, conversely, are those exactly the places where LGBTQ community fades into the background? I use a representative survey of cohorts of LGBQ people from the contemporary United States (2017-2018), the *Generations* study (Meyer 2020), with demographic information and survey question responses measuring community connectedness. I geographically link these responses to American Community Survey data on zip code and metropolitan characteristics. This chapter addresses the core question head-on through the association between spatial context and self-reported experiences of communtiy and belonging, asking whether those are positively associated and complementary or inversely associated and substitutive.

In chapter 4, I look at whether, in virtual communities organized around LGBTQ

identities, core or peripheral members engage in more talk about community. I analyze Reddit conversations from 11 LGBTQ-themed subreddits from their founding (earliest 2008) through to 2018 (Chang et al. 2020). Reddit is a topic and group based social media platform, so it emphasizes "community" to a greater degree than many (but not all) such contemporary platforms. This data structure has the text of top-level posts and all comments; importantly, it identifies which individuals are making those comments and who, exactly, they are replying to in the thread of a conversation. More central members of a group could either engage in a lot of expressive community talk, or they could take it for granted; whereas peripheral members' marginal status could be reflected in an absence of community-oriented language, or they could performatively create their own sense of group belonging through the language they use in conversation.

To conclude, in Chapter 5, I assess the joint contributions of these three empirical projects. I draw on my empirical findings to anticipate and imagine the possible future trajectories of LGBTQ communities.

2 Geography, Gemeinschaft, Identity

Disentangling the meanings of "community" through word embeddings

When LGBTQ people talk about community, does what they mean invoke a *gemeinschaftliche*, belonging- and social organization-oriented sense of the concept? Do they invoke *Gemeinschaft* to a greater extent than in more general and generic contexts? In answering these questions, this chapter responds, in part, to a call from Levine (2017) to investigate the power and ambiguity of "community" in contexts other than his own, where he shows how it operates as a rationale in the case of local governance; it also responds to a host of LGBTQ research that sometimes problematizes community (Orne 2017; Winer 2020) and sometimes takes it for granted (Frost and Meyer 2012). I demonstrate how the empirical everyday meaning of "community" in an LGBTQ social context invokes sociological understandings of *Gemeinschaft*, of community as a meso-level form of social organization (Brint 2001; Tönnies [1887] 2001) – albeit not exclusively. This evident salience of Gemeinschaft matters given the prevailing view of modern, urban society as instead promoting individualism and isolation (Putnam 2001; Simmel [1903] 1971), and how those individualistic trends have been intertwined with the historical development of LGBTQ identities (D'Emilio 1992).

To approach the question of how salient community as Gemeinschaft is for LGBTQ

people, I examine how the concept of "community" shows up in one LGBTQ-centered social group. Founded in 1983, a Usenet group called soc.motss ("soc" meaning one of the many groups for social discussion, and motss standing for "members of the same sex") is the earliest known virtual community built around a shared LGBTQ identity (Auerbach 2014). This makes it unique and important, because it sets a precedent and influences how later online groups become ubiquitous fora for discussing community, identity, and many other topics (Dym et al. 2019). The data I use from soc.motss are an archived corpus of individual messages sent to the group from the late 1990s through the 2000s. For a comparative baseline, I draw on a large volume of Internet text, not directly, but mediated through an off-the-shelf natural language processing (NLP) model. This baseline represents general or generic contemporary English-language discourse.

This chapter adopts innovative NLP methods to learn about community from how the word is used in naturally-occurring language. I systematically investigate the connotations of community across text using word embeddings (Mikolov et al. 2013) – a type of model that mathematically represents words based on the contexts in which they appear. Word embeddings are well-suited for investigating the semantic dimensions of social concepts and how they vary across contexts because – unlike other methods for computational text analysis – they move from surface-level words to underlying meanings and their relations (Arseniev-Koehler and Foster 2022; Stoltz and Taylor 2021). By first engaging in a close read of how "community" is represented in a general word embeddings model pretrained on large amounts of online text (approximating a generic social context), I lay the foundation for interrogating how the meaning of "community" in the context of soc.motss differs from that generic context, and how it remains similar. Specifically, I use dimensionality reduction and algebraic transformations to assess how related "community" is to three latent semantic dimensions – geography, Gemeinschaft,

identity – in each of those contexts.

I hypothesize that, in a self-selected virtual community based on LGBTQ identity, if community itself is salient as a topic, then members will visibly talk about community as Gemeinschaft. If, by contrast, community is backgrounded and ambient, other senses of the word will be more evident in the localized embedding of the term. What I find is that the discourse in the soc.motss LGBTQ Usenet group uses community in the sense of Gemeinschaft to at least an equal extent compared with general English-language text. The key semantic difference is that it also deemphasizes the geographic aspects of community and replaces that with connotations specific to LGBTQ identities. In other words, I show that "community" in the general sense brings together two semantic domains – geography and Gemeinschaft; in a queer context the geographic connotation recedes to the background and a third domain, identity, emerges to take precedence. Community in the sense of Gemeinschaft turns out to be the common bridge between the general and LGBTQ-specific contexts.

2.1 Background

2.1.1 "Community" and the meanings of concepts

Community has a power and ambiguity that render it suitable for strategic rhetorical uses in everyday discourse. Two qualitative examples illustrate the shades of meaning that "community" can take on; together, these illustrate the range of variation that I might expect to see in a computational investigation of what community means as a folk concept. First, in the context of local governance, Levine (2017) observes that "community" becomes a "floating signifier of the good," a halo of positivity to cover the real operation of local decision-making and to provide legitimacy for action. For Levine,

this is harmful; he notes the impossibility of "the community" wanting one single thing as a uniform entity, and the harms of ascribing collective representational authority to whoever can show up to participatory events (Levine 2021). In his case, the word is constantly used in a positive and justificatory light, but so flexibly as to lose coherent meaning. Second, and in contrast to that wholehearted positivity, Winer (2020) finds that his interviewees have an ambivalent and distancing relationship with the "imagined gay community," drawing a distinction between "the community" at large and their own social circles. Rather than pure vagueness, this points to another specific rhetorical use, to critique an in-group's flaws rather than to justify desired actions. In these accounts, "community" assumes differing valences, with slippery or counterintuitive referents, but in each case the concept does important discursive work. I do not aim to create a taxonomy of these rhetorical strategies; instead I will show how all of these uses together add up and contribute to the overall semantic resonance that "community" takes on. Language, after all, is social and shared (Saussure [1916] 1972); later I will show how that shared foundation can be a springboard for understanding local deviations.

For fully understanding the social life of a complex concept, academic definitions are insufficient on their own, but worth reviewing as an anchor for comparison. "Community" is a phenomenon sociologists have elaborated on since Tönnies ([1887] 2001); they have created taxonomies of different types of communities (Brint 2001); argued over what communities count as "real" (Driskell and Lyon 2002; Rheingold 2000); and debated about what features – shared social networks/interactions or shared cultural/moral traits – are most fundamental to the creation and experience of it (Boessen et al. 2014; Vaisey 2007). What emerges consistently is the metaphor of a tightly knit social fabric, a group of people bound together by shared ties, shared culture, and possibly shared place.

However, the issue with using academic definitions of "community" as a starting point is that a strict definitional logic of concepts and categories does not apply "out there" in the real social world; classical logic may be useful for technical jargon, but it is not how ordinary human concepts work. Instead, everyday concepts are fuzzy and prototype-based (Bowker and Star 2000; Lakoff [1987] 2008, [1987] 2008; Monk 2022; Rosch and Mervis 1975; Zerubavel 2002). One way to see this is to think about how people figure out that something *is* a community in the first place. As Bruckman (2022) argues, we might decide that a virtual community like Wikipedia is a community through mental comparison to prototypes of community like a small town. A given example of community does not have to have all of a specific set of features in order to fall under the concept; rather, concepts and categories are bundles of "intensions," inherently fuzzy constellations of characteristics and cues. "Community," as an everyday concept, bundles together a spread of connotations; as I will show, the most notable of these are a geographic sense of "local place" and a sociological sense of "social group" or "object of belonging." Because "community" encompasses both, even the most mundane use of community to refer to local place might still invoke the sentiment and connotation of *Gemeinschaft*. That is partly where the fuzziness and ambiguity of community as a concept could come from, and also part of the concept's discursive power. However, distinguishing these two senses is necessary in order to open up a window into where and how the underlying meaning of *Gemeinschaft* and belonging appears in different discursive contexts.

Alongside these two senses of geography and Gemeinschaft, I examine the potential overlap between community and identity. In general, shared social identity characteristics are one potential basis for communities (Brint 2001).At the same time, in the context of the cultural sociology of markedness, Zerubavel (2018) notes that this basis

is not necessarily available for unmarked identity categories: "the heterosexual community" is an empty, nonsensical statement, while "the LGBTQ community" is a common and sensical one. The open question, however, is about how community and identity are link when a virtual community like soc.motss is already centered around LGBTQ identities. Given the shared context, there is the possibility that those identities will be backgrounded rather than foregrounded in the context of discourse about community.

2.1.2 Using word embeddings to measure meaning

This project, then, uses an empirical, inductive, and computational approach to discover what community means and compare it to theoretical expectations derived from sociological literature. It integrates the sociological definition of Gemeinschaft in an iterative way, making this an abductive approach (Brandt and Timmermans 2021), rather than a purely grounded one (Nelson 2017). To examine the resonances and connotations of community in generic English discourse, word embeddings are my computational method of choice. Word embeddings are a relatively recent (Mikolov et al. 2013) computational operationalization of an old linguistic idea, called the *distributional hypothesis* (Sahlgren 2008). As Firth (1957) put it, "You shall know a word by the company it keeps." Accordingly, these models represent words as a function of all of their immediate contexts. To give one example:

> "The history of all hitherto existing **society** is the history of class struggles."
> (Marx [1848] 1972)

An embeddings model would take this sentence and learn about the semantic connotations of the the word "society" from its position near "history", "class", and "struggle"; it might also learn linguistic features common to nouns from its position in relation to

words like "is" and "of." Naturally, a model needs many such examples as training data, to produce a single overarching numeric representation for each word in a vocabulary.

There are two main commonly-used word embeddings models – word2vec, based on a shallow neural network (Mikolov et al. 2013), and GloVe, based on cooccurrence matrix factorization (Pennington, Socher, and Manning 2014). These approaches are mathematically related to each other and the substantive differences in the resultant embeddings are minor, and so I choose one or the other for practical reasons of convenience and convention Nelson (2021). Specifically, high-quality and widely-used pretrained models have been released based on the GloVe method (Pennington et al. 2014), while a high-quality and robustly-engineered Python software package, gensim, implements the word2vec method for training new models on particular corpora (Řehůřek and Sojka 2010). These basic word embeddings approaches are foundational for a host of subsequently developed NLP methods, from contextual word embeddings models like BERT (Devlin et al. 2019) all the way to large language models like the GPT family of generative models (Brown et al. 2020). The simpler word embedding models offer the most straightforward and interpretable entry point for addressing the question of what "community" means in everyday English-language contexts.

A generic model with a robust, comprehensive view of as many contexts as possible would approximate "the" meaning of every word in a language. To train general models for a given language, the most common corpora are large and publicly accessible texts from the Internet, e.g., Wikipedia pages, newswire articles, social media, or anything else that can be conveniently crawled from the web. (Historical embeddings use digitized book corpora.) These pretrained models can be used for a variety of questions and tasks, but they stand in contrast to locally trained models derived from specific corpora. Of course, meanings of words do vary – over time, over space, and by other so-

cial characteristics (Bamman, Dyer, and Smith 2014; Soni, Klein, and Eisenstein 2021). A general model trained on easily-accessible data works to the extent that meanings are common or shared; of course, this flattens variation. Given the social characteristics of the authors of formal online texts like Wikipedia or news corpora, in terms of gender, race, education, nationality, etc. (Hargittai and Shaw 2015; Vrana, Sengupta, and Bouterse 2020), models trained on those data necessarily overrepresent hegemonic cultural viewpoints. This overrepresentation is a form of bias, but also a matter of substantive interest (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2017; Jones et al. 2020).

Distinct from other text-as-data methods (e.g. keyword dictionaries or topic models), embeddings models create dense, distributed vector representations of words. In this way, word embeddings encode a relational model of meaning; they build up a system of signs (Saussure [1916] 1972), a vocabulary, in which distances (or their inverse, *similarity* measures) in a high-dimensional space can be calculated between every pair of words. This makes embeddings useful for social-science problems where meaning matters – especially where variations or changes in meaning are of interest. For instance, they have been applied fruitfully in cultural sociology to show how the distinct dimensions of class correlate and evolve over the course of the 20th century (Kozlowski, Taddy, and Evans 2019), and in political science for modeling ideology in parliamentary debates (Rheault and Cochrane 2020). An embeddings model can be thought of as distilling shared, declarative public culture (Lizardo 2017); Arseniev-Koehler and Foster (2022) go even further to argue that the training process is a reasonable heuristic model for actual cultural cognition.

But the dimensions the models learn are not themselves interpretable (nor are they consistent across models, meaning that different embedding matrices must be aligned for

comparison). To derive interpretable dimensions, social scientists use anchor words and simple algebra. One common approach is to construct new binary dimensions through subtraction (Kozlowski et al. 2019; Taylor and Stoltz 2020), opposing pairs of concepts that can be thought of as antonyms (e.g. rich - poor, woman - man). This idea springs from the algebraic analogy tasks that first made word embeddings notable in NLP (e.g., king - man + woman $\approx$ queen). While these binary oppositions have a clear basis in cultural sociology (Douglas 1966; Durkheim [1912] 2001; Saussure [1916] 1972), they are not the only possibility. They do not necessarily make sense for a concept that might bundle together multiple overlapping connotations or characteristics.

Instead, in this work, I deviate and borrow a different algebraic idea from NLP: "de-biasing" an embedding through orthogonal projection away from a target word vector (Gonen and Goldberg 2019). This approach originates in an attempt to mitigate gender bias in the words for different professions and occupations – which is undesirable for NLP tasks such as machine translation (Caliskan et al. 2017), even if it represents cultural associations or demographic facts about particular occupations that might be worth studying in themselves (Jones et al. 2020). That foundational work on this method makes it clear that it does not remove all the connotations of the undesired word – i.e., it does not fully succeed in de-biasing – but it is successful enough to use to disentangle the connotations of a concept like community. I can then compare how those connotations vary between a general context and the specifical local context of soc.motss.

2.1.3 An early LGBTQ virtual community: the soc.motss Usenet newsgroup

My case study is an early virtual community, an LGBTQ Usenet group called soc.motss. The soc.motss archive spans a key time period when the position of LGBTQ people in American society is shifting towards greater acceptance and equality, centered on the late 1990s and early 2000s (see Figure 2.1 below). It captures LGBTQ discourse on the heels of the Don't Ask, Don't Tell military policy (1993) the Defense of Marriage Act (1996), at the early end of a two-decade-long shift in public opinion in favor of LGBTQ equality (Rosenfeld 2017). Simultaneously, the legal landscape starts to change, from *Lawrence v. Texas* decriminalizing same-sex sexual relations in 2003 to *Obergefell v. Hodges* legalizing same-sex marriage in 2015.

Usenet is a distributed system for sharing electronic messages which predates the contemporary Internet (Rheingold 2000), organized into topical groups such as alt.atheism or rec.motorcycles (to take two examples from the "20 Newsgroups" dataset (Lang 1995)). Some of these groups are reported to have had a strong sense of community, while others were known for their hostility (Baym 1994; Dame-Griff 2019). Usenet is of interest because the moment of its heyday is when people were demonstrating that virtual community was, in fact, possible (Calhoun 1998; Driskell and Lyon 2002; Hampton and Wellman 2003; Rheingold 2000). Moreover, as boyd (2014) points out, even if particular platforms are passé, the social processes that unfold on them are not. Usenet groups are well-suited for studying the creation of community using computational text analysis methods because they are both conversation-oriented and text-based (McCulloch 2019); these kinds of virtual interactions can have similar feedback effects to those associated with face-to-face interaction rituals (DiMaggio et al. 2018).

Usenet preceded the period when the majority of people in the United States used

social media, and is instead from a time period when virtual spaces provided an outlet for outsiderness (boyd 2014). Soc.motss members are a self-selected group, both in their avant-garde usage of technology and in their desire to seek out LGBTQ community spaces. They are described as more introverted on the whole, and so these individuals might not have had as much access to or comfort in in-person LGBTQ spaces like gay bars (Auerbach 2014). As were other virtual communities of the time (Rheingold 2000), they are likely disproportionately concentrated in tech centers and centers of gay culture like the Bay Area. Despite that geographic distribution, however, soc.motss affords the opportunity to use digital technology to find LGBTQ community in a way that potentially transcends geography.

Because of stigma and isolation, LGBTQ people may be especially predisposed to seek out queer community in digital spaces. The internet has long been recognized for its potential for marginalized groups, given the possibility of anonymity and the ability to manage contexts to control the disclosure of identities in unwanted ways (boyd 2015; Mehra, Merkel, and Bishop 2004). As such, LGBTQ people can use virtual communities for connection and support that may be lacking in offline social spaces (Dym et al. 2019). Corroborating this potential, LGBTQ groups and interests have been present and visible since the earliest virtual spaces came into being(Auerbach 2014; Rheingold 2000). The continued utility of technology for LGBTQ community-building is evident in the way that LGBTQ people have continued to act as early adopters of new technologies for digital social life (e.g., mobile and location-based platforms (Orne 2017)) through to the present day.

Soc.motss – where "motss" stands for "members of the same sex" – was oldest and largest LGBTQ Usenet group. According to a history recounted by Auerbach (2014) in *Slate*, it was founded in 1983 (as net.motss), not long after Usenet came into existence

in 1980. In terms of case selection, soc.motss is worth studying not because it is typical or representative, but because it is unique and historically important. As "the first gay space on the Internet" (Auerbach 2014), it influenced the many queer spaces that would come after it; as one of the largest spaces of its time period, it provides a sufficient corpus for modeling with word embeddings.

An archived version of the soc.motss FAQ from 2001 describes the group as follows (in a section headed 'Our "we" '):

> Soc.motss serves non-heterosexual Internet communities. To signal inclusiveness, we use the acronym LGBTO, for Lesbian, Gay, Bisexual, Transgendered and Others, "others" meaning supportive straight people. The newsgroup is a predominantly non-heterocentric space where we can discuss issues of importance to our communities.

Elsewhere, the FAQ provides ample evidence that this *is* a cohesive group and a virtual community by any definition of the term. It has norms for participation, a group discursive style (Eliasoph and Lichterman 2003), and community-building events like in-person meetups ("motss.con") (Auerbach 2014; Rheingold 2000). Of course, there were other LGBTQ newsgroups as well, such as soc.support.youth.gay-lesbian-bi or the several trans Usenet groups Dame-Griff (2019) has already studied. Few of these are large enough to support training distinct models, and it would muddle the analysis to combine many potentially distinct group cultures and discursive styles together in one corpus. On its own, soc.motss offers a well-defined, prototypical early LGBTQ virtual space, in which to investigate how LGBTQ people construct the meaning of community together.

2.2 Data and methods

I use two word embedding models to compare how the term "community" between the Usenet soc.motss corpus and in contemporary English-language texts more generally. The general, generic model is a pretrained GloVe model, and the local, LGBTQ-specific model is a word2vec model I train on the soc.motss corpus. GloVe and word2vec are the simplest methods for creating word embeddings, and for each respective case I have chosen the highest-quality and most widely-used tool. I describe both models, the corpus, and the methods for comparison here, with the goal of determining how central the semantic dimension of *Gemeinschaft* is in the meaning of "community" in each context.

2.2.1 Pretrained model (GloVe, Wikipedia + newswire text)

I use pretrained GloVe embeddings (Pennington et al. 2014),originally trained on a full English Wikipedia corpus from 2014 and a newswire corpus called Gigaword 5. The rationale for combining those two corpora is that more data produces more stable and generalized embeddings. Prior social science researchers have found pretrained embeddings to be reasonably robust, stable, and generalizable (Rodriguez and Spirling 2020; Stoltz and Taylor 2020). They can be used to study, for instance, ideology in 20th-century politics (Rheault and Cochrane 2020; Rodriguez and Spirling 2020); but not race and gender in 19th-century literature, which is far enough removed in context that a locally-trained model would be more appropriate (Nelson 2021; Soni et al. 2021). Kozlowski et al. (2019), who use a similar set of historical embeddings, suggest thinking of the associations encoded in these embeddings as coming from a "literary public" with known and unknown biases compared to the general population. Those associations are shared to some degree with the meanings encoded in the soc.motss corpus; if that were

not true, it would not be reasonable to make a comparison at all.

The GloVe model uses co-occurrences in the Wikipedia and newswire corpus to "embed" words in a 200-dimensional space, resulting in a 400,000 by 200 matrix. That is, there are 400,000 words in the full vocabulary of the model, the 400,000 most prevalent words in the corpus, and each word is represented by a vector of 200 numbers. Many of the words in the 400,000-word vocabulary are extremely rare in ordinary English, including proper nouns and foreign-language terms. Rare words add noise and reduce interpretability; to mitigate this, I subset the model vocabulary to the most common words. In the first section of the results below, I do this by intersecting the vocabulary with a second GloVe model pretrained on a corpus of text drawn from Twitter, for 150,396 words in total. In the second section, I necessarily must subset to the vocabulary held in common with the soc.motss corpus for model comparison.

The latent dimensions represented by each of the 200 numbers in a single word embedding vector are arbitrary, without intrinsic meanings. To interrogate how potential meanings for my focal concept of "community" are encoded in this GloVe model, I construct interpretable semantic dimensions through the following inductive steps:

1. I select the 1,000 nearest neighbors to the word "community." Distance in the 200-dimensional embedding space is measured with cosine similarity between word vectors, which potentially ranges from -1 to 1. A word has a cosine similarity to itself of 1. (Cosine distance, defined as 1 - cosine similarity, is used in some of the derived metrics I describe below. A word has a cosine distance from itself of 0.)
2. I decompose this local neighborhood of 1,000 nearest neighbors, the words with the highest cosine similarities to "community," through principal components analysis (PCA).
3. I inspect the resulting PCA dimensions for the proportion of local variation they

explain and for any potential substantive interpretation. I use the whatlies Python package (Warmerdam, Kober, and Tatman 2020) for interactive visual exploration. Here, even though the embeddings in the neighborhood of "community" do not fall into discrete clusters, I cluster them with K-means clustering to aid in interpretation.

4. I average the vectors for extreme words (N = 10) along a PCA axis of substantive interest to "debias" (Gonen and Goldberg 2019) the focal word vector through orthogonal projection. This axis substantively contrasts geography-related words and Gemeinschaft-related words, with "community" squarely in the middle.

This process is interpretive and cannot be performed automatically. In principle, however, this method could be extended to other complex social-scientific concepts – for instance, "freedom" and "democracy," as political scientists have studied (Rodriguez and Spirling 2020), or social class and its dimensions (Kozlowski et al. 2019).

2.2.2 Local corpus and model (soc.motss, word2vec)

I build my local corpus and model as follows. I download the full set of available posts from the soc.motss newsgroup archived in the Usenet Historical Collection (UHC), hosted by the Internet Archive. This archive contains nearly 300,000 posts spanning the years 1999-2013, with a peak in the early 2000s and a continual decline in post volume thereafter (shown in Figure 2.1). According to Dame-Griff (2017), there were no systematic attempts to archive Usenet before 1995. While I investigate other potential archival sources, the UHC archive is so much larger that I do not attempt to merge sources together. (The second available archive, from Google Groups, is less comprehensive; it has only 1,074 posts from net.motss, the first iteration of the group, for 1983-1986. The Google Groups soc.motss archive contains only 60,400 posts from 1986-2022 (though

the most recent posts are entirely spam), with only 9,847 posts from before 1999-04-17, when the UHC archive begins.) Archives inherently risk being incomplete, but I believe this corpus is comprehensive enough to characterise the culture and language of the group in the early-2000s time period.

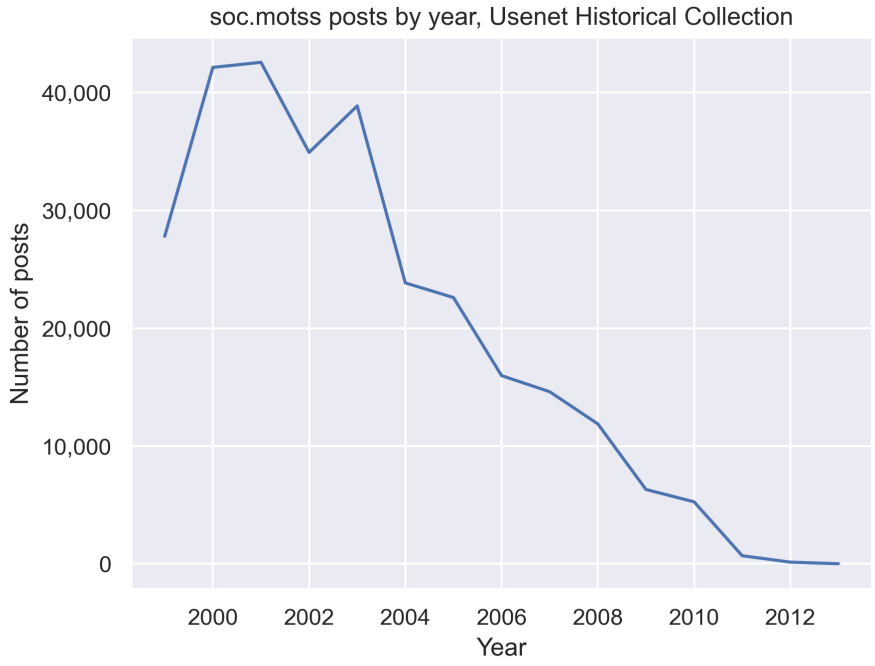

Figure 2.1: Posts from soc.motss archived in the Usenet Historical Collection by year.

I load and preprocess the soc.motss UHC archive using Dame-Griff's Python scripts, developed for analyzing transgender Usenet groups, strip quotes and footers from the text with scikit-learn (Pedregosa et al. 2011) tools intended for the 20 Newsgroups dataset, and lowercase and tokenize the text with the gensim package (Řehůřek and Sojka 2010). This preprocessing is relatively minimal, but I note that any computational text analysis can be sensitive to preprocessing choices (Denny and Spirling 2016; Nelson 2017).

I fit a word2vec model to the processed corpus using gensim (Řehůřek and Sojka

2010). Key word2vec model parameters are set at a vector size of 200, a context window of 6, and a minimum word occurrence threshold of 5. I use skip-gram with negative sampling as the model architecture and train for 10 epochs. The parameter choices I make are consistent with Nelson (2021), Rodriguez and Spirling (2020), and other social science research using word2vec models. Following Nelson (2021), I also bootstrap key estimates by fitting forty new models to create a 95% confidence interval. Because Usenet posts are relatively short documents, I bootstrap posts with replacement, not individual sentences.

After preprocessing and model fitting, there are 287,680 documents, 26,958,729 tokens, and 71,617 unique words in the final vocabulary (with the aforementioned minimum threshold for inclusion set to 5 occurrences). Of that vocabulary, 60,728 words also exist in the GloVe model's vocabulary. The words found only in the soc.motss vocabulary are largely misspellings, concatenations of words, encoding errors, foreign languages (especially Spanish), and colloquialisms (especially gay slang, Usenet slang, and emotive language).

Notably, this corpus is larger than all six trans usenet groups that Dame-Griff (2019) has archived put together, and much larger than conventional Usenet data sets like 20 Newsgroups (Lang 1995) (which uses a smaller temporal slice from more groups). The corpus size is still on the lower bound of what might be desirable for model quality, but I believe my results below show that the model produces embeddings that are stable and robust enough (Antoniak and Mimno 2018) to validate the broad, high-level substantive patterns I find.

To complement the semantic dimensions derived from the PCA decomposition of the GloVe model, I create a third semantic vector of words related to LGBTQ identity. Unlike the other two semantic dimensions, I choose these keywords by hand based on

domain knowledge and manual inspection of nearest neighbors. The exact words are therefore more ad hoc and less principled, but this is not an atypical approach; it is no different from previous studies that use keywords to create vectors for concepts like "power" (Nelson 2021) or "social class" (Kozlowski et al. 2019). To match the other dimensions, I pick 10 words: "lgbt," "lgbtq," "glbt," "gay," "lesbian," "bi," "bisexual," "transgender," "queer," and "homosexual." ("Gay" is the most common of these words in the soc.motss corpus, appearing 49,486 times; the rarest is "lgbtq," appearing only 18 times.) As with the geography and Gemeinschaft dimensions, by taking an average the vector becomes more robust to the inclusion or exclusion of any given word.

When comparing the local soc.motss model to the pretrained GloVe model, I address one technical question and two substantive questions. At a technical level, is the local word2vec model high-quality enough to have face validity? Finding that it is, I then address two substantive questions. First, *how different* is the term "community" between the models, in a relative sense compared to other words. The same three metrics that I use to assess model quality also address this question, and set me up to answer the second, more important, substantive question – not *how different* but rather *different how*?

The first metric, a "query rank" correlation measure, comes from Rodriguez and Spirling (2020). It is the Pearson correlation between models of within-model cosine similarities for a word and every other word in the vocabulary. This metric can be interpreted as a general measure of embedding stability and model quality. The second and third measures come from Hamilton, Leskovec, and Jurafsky (2016a), who use two distinct applications of cosine distance to study semantic change over time. Unlike the first measure, these two require mathematically aligning the embeddings matrices to the extnet possible, using a matrix alignment method called orthogonal Procrustes

(Hamilton, Leskovec, and Jurafsky 2016b). While originally temporal in nature, these measures are equally well-suited to comparing general and specific contexts. The second measure, called "linguistic drift," directly compares the cosine distance of a given word across model. This captures regular change due to linguistic processes. In a temporal context, languages are constantly evolving (Saussure [1916] 1972); because some linguistic innovations begin in or are confined to particular subcultures – like the LGBTQ community – this metric is also useful for my atemporal comparison. The third measure, called "cultural shift," is a second-order comparison metric, using the cosine distance of the cosine similarities of the nearest neighbors to the word. According to Hamilton et al., this second-order signal is more sensitive to "irregular" changes due to societal shifts, e.g., due to technological advances or social movements. For example, in their work, they show that the shift in meaning of the word "gay" – from happy to homosexual – is better captured by the "cultural shift" measure than the "linguistic drift" measure. I use all three metrics for a comprehensive picture of how much the term "community" differs between soc.motss and more general usage.

Only then do I have the confidence to tackle my core question: do the LGBTQ members of soc.motss invoke community as Gemeinschaft to a greater degree than in general English usage?

2.3 Results

2.3.1 Semantic dimensions of "community" in the general, pretrained model

I begin by examining the meanings of the term "community" as derived from a standard, generic set of word embeddings, a GloVe model pre-trained on Wikipedia and newswire

text and released for public use (Pennington et al. 2014). As a general set of embeddings, the model necessarily encodes the cultural biases of those that produce formal, written online English text. By turning words into numeric vectors, word embeddings encode a relational notion of similarity. Thus, the primary way to understand what a given word "means" in a model is to examine the words that are most closely related to it. These are the words that would appear in similar contexts to that given word; they could be synonyms or otherwise semantically or functionally similar.

As an initial pass, I examine the 10 most similar words to "community" in the GloVe model, measured by cosine similarity. These are "communities," "organizations," "society," "local," "established," "area," "part," "within," "public," and "council." The specific words range from the near-identical ("communities") to related but distinct concepts ("society") to words that would fit together into phrases ("local community"). Among these nearest neighbors, both spatial and social dimensions are evident. A cursory interpretation might stop there. However, this narrow view of the local neighborhood in the embedding space around community does not provide a sense of which semantic dimension might be more important or how they might be related.

For a fuller and more structured view of what community means, I expand the set of most similar words from 10 to 1,000. These 1,000 nearest neighbors constitute a local neighborhood of words in the GloVe vocabulary, with the word "community" as the focal word. This neighborhood is a subset of the overall embeddings space of the model. To understand the semantic structure of this neighborhood, I need to mathematically transform this set of embeddings (i.e., a 1,000 by 200 matrix), because the 200 dimensions of the original vectors are not interpretable and do not have any intrinsic meanings. I transform the embeddings by decomposing them with principal components analysis, a method I choose for its relative interpretability compared to other dimen-

sionality reduction methods (e.g., t-SNE). Table 2.1 shows the ends of the first six PCA dimensions from the nearest neighbors to "community" in the GloVe model. Unlike the mixed set of words that were the ten closest to "community," each set of 10 words in the table qualitatively shows a reasonable amount of semantic coherence, and in some cases the opposition of each end of a given dimension is also interpretable. Note, however, that the original 200 dimensions of the embeddings space encode a substantial amount of subtle information that is lost with dimensionality reduction, so the proportion of variance explained by the first several dimensions is relatively low. (Kozlowski et al. (2019) have shown similar results in an experiment with PCA and with explicit cultural dimensions, so this is unsurprising.)

Table 2.1: Top 10 and bottom 10 words for first 6 principal components, out of 1000 nearest neighbors to "community," from pre-trained GloVe model (Wikipedia + Gigaword 5)

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
n't	cooperation	research	teach	populations	historical
we	promote	management	kids	populated	contemporary
if	governance	library	music	farming	political
do	awareness	science	teaching	areas	history
get	stakeholders	facility	religion	sustainable	founding
could	initiatives	university	contemporary	infrastructure	influential
what	sustainable	provides	teacher	vast	described
know	understanding	institute	traditions	cultures	cultural
would	commitment	facilities	tradition	population	role
really	implement	program	feel	coastal	movement

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
…	…	…	…	…	…
baptist	district	minorities	government	alumni	families
african-american	county	refugees	infrastructure	meeting	parents
encompasses	situated	arab	aid	met	workers
nonprofit	township	settlers	region	member	employees
methodist	nearby	orthodox	regional	joined	teachers
interfaith	near	ethnic	economic	university	volunteers
lgbt	village	jews	summit	attending	residents
community-based	suburb	muslim	nations	invited	kids
non-profit	town	christians	security	attend	homes
not-for-profit	located	muslims	cooperation	attended	educate

Figure 2.2 focuses on the first two dimensions, effectively projecting the 200-dimensional vectors down into two-dimensional space. The first dimension (the x-axis in Figure 2.2) ranges from words like "if" and "we" and "not" to words like "not-for-profit," "community-based," "lgbt," and "interfaith." Based on the distributions of words along these dimensions, I label this first dimension as a *linguistic* dimension. It encodes distinction between common, functional words and words that are more complex and substantive. While important for structuring the overall space of meaning, this distinction is not relevant for my analysis, because it is linguistic and not substantive.

The second dimension (the y-axis in Figure 2.2), however, is more substantively salient. Ranging from words like "town" and "located" to words like "cooperation," "governance," "organizations," and "collective," it encodes what I interpret as a distinction between geography and Gemeinschaft. "Community" itself falls nearly in the middle between the two poles of this dimension. In the figure, I have highlighted three clusters derived from k-means clustering to aid in interpretation. The first cluster captures the functional linguistic words on dimension 1, which are not differentiated much on dimension 2; the second and third clusters separate out the more substantive words on dimension 1 into two semantic groupings: geography words on the high end of dimension 2 and Gemeinschaft words on the lower end.[1]

Because this second PCA dimension clearly reflects the contrast between community as space and community as Gemeinschaft, my analysis focuses on this dimension. Drawing on each end of this geography-Gemeinschaft continuum, I select the 10 words (from the 1000-word neighborhood) that are the most extreme on either end. Figure 2.3 displays these two sets of words again and shows that they do in fact fall into two distinct blocks – highly similar within each group, and highly distinct from the other group. By construction, the word "community" is highly similar to both groups – it quite literally bundles these two connotations together in a single concept. To produce a more robust vector measure for each underlying connotation of "geography" and "Gemeinschaft," I average the 10 individual word vectors, as is common practice (Kozlowski et al. 2019; Waller and Anderson 2021).

Finally, I project the vector for "community" away from the averaged geography vector. Using the linear algebra shown in Equation 2.1, the "community" embedding ($\vec{c}$) is transformed so that the dot product of the new vector ($\vec{c'}$) with the geography

[1]While there's no evidence clustering would have been a better approach than dimensionality reduction, the consistency is additional evidence that the principal components are robust.

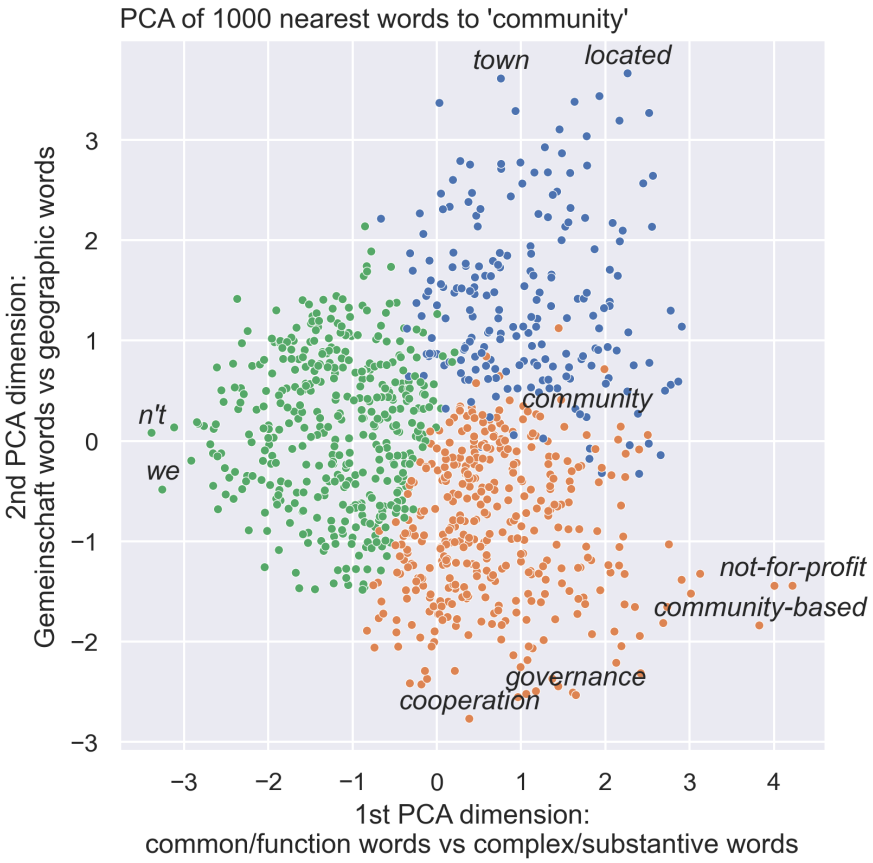

Figure 2.2: PCA decomposition of 1000 nearest word embeddings to "community", showing the first two dimensions. While the space is continuous, k-means clustering with k = 3 effectively divides it into functional words in green, geography words in blue, and Gemeinschaft words in red.

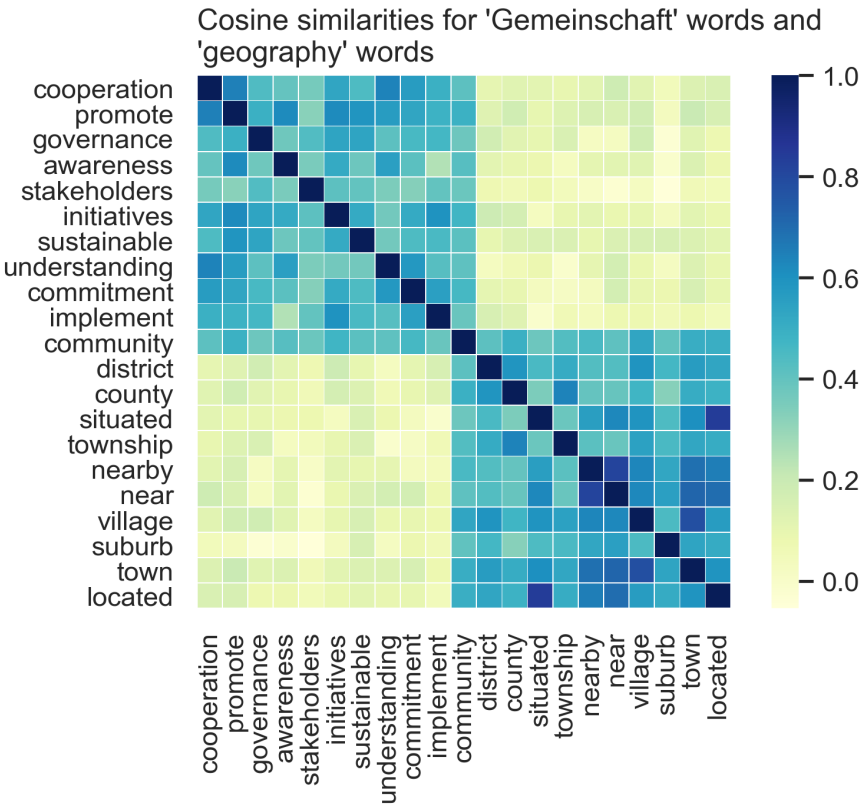

Figure 2.3: Similarities of the 10 highest and 10 lowest words along PCA dimension 2, which I have labeled geography – Gemeinschaft.

vector ($\vec{g}$) is 0 (Warmerdam et al. 2020).

$$\vec{c'} = \vec{c} - \frac{\vec{c} \cdot \vec{g}}{\vec{g}^2} * \vec{g} \tag{2.1}$$

By moving "community" away from the geography vector, I also move it toward the Gemeinschaft end of the continuum. The algebraic transformation "de-biases" (Gonen and Goldberg 2019) the new "community" vector of its spatial connotations. It creates a new concept vector – like the embedding "community," but purged of its geographic element. I label this new vector community-without-geography – or, alternatively, community-as-Gemeinschaft.

Figure 2.4 is a two-dimensional representation of this process, illustrating the effect of projecting the embedding for "community" to create a modified concept vector. In the figure, the x-axis represents similarity to the new community-as-Gemeinschaft vector; the y-axis represents similarity to the averaged geography words. By definition, each vector has a similarity to itself of 1. The result of orthogonal projection is that the community-as-Gemeinschaft vector has a similarity to the geography vector of exactly 0, again by definition. This has two consequences: community-as-Gemeinschaft remains very similar to the original community vector *and* to the averaged Gemeinschaft vector. An alternative approach – subtracting out the geography domain – does not result in a vector with the same properties. I argue that the projection approach produces an embedding that means community in a purely sociological sense, rather than a spatial one. An equivalent projection can strip out the social connotations of community, leaving a vector representing community as a spatial concept alone. Both of these derived concept vectors provide a comparative tool for analyzing the meaning of community in the context of the soc.motss corpus in the next section, as well as a more general measurement tool in ways that I will outline in the discussion.

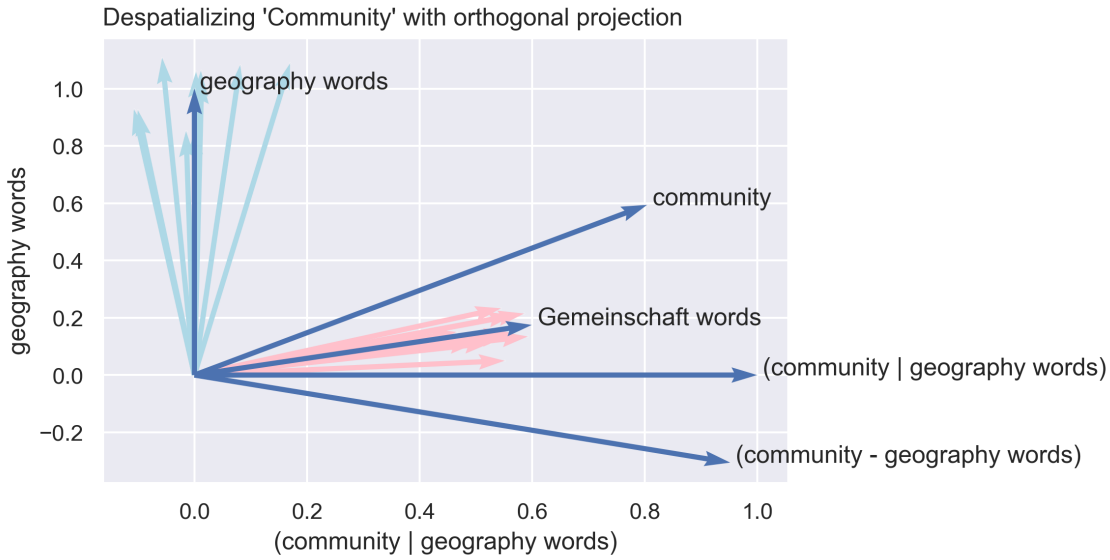

Figure 2.4: Orthogonal projection of the embedding for "community" away from the averaged vector of 10 geography-related words. The resultant embedding (community | geography) is highly similar to the averaged vector of 10 Gemeinschaft words. A binary opposition (community - geography) is shown for comparison, but is less similar to the Gemeinschaft words.

2.3.2 Semantic differences in an LGBTQ Usenet group

After fitting a word2vec model to the soc.motss corpus, I examine the neighborhood of words similar to "community" in this model, in order to determine what *community* "means" in this context. This allows me to investigate how LGBTQ usage of community in this corpus differs from usage in the more generic and general contexts that produced the GloVe model. The geographic and *gemeinschaftliche* semantic dimension vectors and the corresponding community concept vectors derived from them, alongside a third semantic dimension for LGBTQ identities arising from distinct patterns in the soc.motss corpus, are anchor points for comparison. These semantic anchors enable me to systematically compare of the connotations of the embeddings for "community" in each model.

In the soc.motss model, the 10 most similar words to "community" include words that are identical or thematically similar to words in the GloVe model ("communities," "organizations," "collectives"), words about queer identities ("glbt," "lgbt," "lgbtq"), as well as words related to religious entities: "keshet," an LGBTQ Jewish organization, and both "metropolitan" and "churches." (The 10th word is "webshots.") The similarity between "metropolitan" and "community" is *not* a geographic reference, but rather a reference to the Metropolitan Community Church (MCC), an LGBTQ-focused Protestant church. With that context, spatial terms are absent from these ten nearest neighbors; the words that overlap are *gemeinschaftlich* in nature, and the words unique to the soc.motss context relate to identity.

Examining only 10 words is insufficient for systematic conclusions. As with the GloVe model, I then expand the window of similarity to include the embeddings for the 1,000 nearest neighboring words to the word "community," and decompose this subset of embeddings with PCA. To exclude the rarest and most idiosyncratic words (e.g., un-

common proper nouns like personal usernames), I only include words shared between the GloVe and soc.motss vocabularies.[^With more – or fewer – words overall, the PCA dimensions do not appear to remain consistent. By contrast, the PCA dimensions for the GloVe model neighborhood are more robust to taking different subsets of the vocabulary.] The first six dimensions, shown in Table 2.2, are considerably less interpretable than those for the GloVe model. There are some thematic groupings; for instance, dimension 2 ranges from geographic words like 'neighborhood' and 'village' to religious words like 'clergy' and 'congregations'. This dimension comes closest to reproducing the geography-Gemeinschaft continuum of the GloVe model, although the semantic scope of the words at the latter end is much narrower. Many of the extreme words are duplicated across dimensions (e.g., there are similar religion-themed words at the bottom of dimension 3), making it difficult to label them distinctly, and many sets of words are semantically mixed or contain too many rare words or proper nouns to characterize. Qualitatively, this shows that while there is some overlap with the connotations of "community" found in the GloVe model above, the same structure of those meanings cannot be discerned here. Instead, more context-specific themes start to appear: for instance, "msm" (men who have sex with men) and "transgender" appearing alongside "outreach" in Dimension 3, and "bathhouses" occurring among "businesses" and other location-based terms in Dimension 4. Importantly, more words shared in common with the GloVe model's nearest neighbors to the term "community" are evident, including spatial terms. This is why I now move beyond qualitative interpretation and turn to mathematical comparisons of the two models.

Table 2.2: Top 10 and bottom 10 words for first 6 principal components, out of 1000 nearest neighbors to "community," from soc.motss word2vec model

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
subgroup	neighborhood	wellness	potpourri	officials	croome
deepen	scoured	stumbleupon	ahrens	centers	activists
subgroups	village	msm	irminsul	nonprofit	activist
disparities	thrift	transgender	troth	auspices	transgender
hindered	metreon	mobilization	weintraub	lockup	parade
inequalities	wildflower	nonprofit	yahad	disparities	quintero
personhood	mayfair	disparities	badb	epidemic	lesbian
institutionalize	etobicoke	outreach	gitlin	sectors	gay
salience	telmo	visibility	pittman	courtrooms	auckland
assimilate	area	linkedin	gajic	pediatric	staged
…	…	…	…	…	…
morristown	pastors	communion	clientele	gajic	universal
metropolitan	clergy	catholics	walkable	glbt	membership
sholom	tongzhi	parishes	influx	facilitator	haifa
alejandro	laity	congregations	communities	chatroom	stumbleupon
cla	churches	unitarian	bathhouses	queer	nypl
bradenton	advocacy	church	affluent	bisexual	affiliated
citywide	soulforce	congregation	areas	transgendered	metropolitan
auckland	rivalries	episcopal	businesses	newsgroup	user
rodeph	congregations	denomination	populations	poc	fellowship

Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
montclair	interfaith	presbyterian	neighborhoods	webshots	webshots

I use three quantitative measures – query rank correlation, linguistic drift, and cultural shift, defined in the methods section above – to systematically characterize differences between the soc.motss-derived and pretrained GloVe models. This is a necessary contextualing step before the direct semantic comparisons that follow. By situating the metric values for "community" in comparison to all other words in the vocabulary shared between the two models, I show, in a relative sense, how different "community" is overall between the general-English and LGBTQ-specific contexts.

This has the simultaneous benefit of evaluating the general quality of the soc.motss model, beyond its qualitative face validity. If every word embedding were to shift significantly, it would be as if LGBTQ people were speaking an entirely different language from the GloVe model's corpus. That would in turn cast doubt on the premise that the corpus provided sufficient data for locally training a stable and valid word2vec model. This is not the case; for the vocabulary as a whole, the three measures are largely consistent, with moderate-to-high correlations between metrics.

However, the metric values for the word embedding for "community" are *not* consistent in the amount of change they indicate:

- Figure 2.5 shows the distribution **query rank correlations**, for all 60,728 words shared between the GloVe model and the soc.motss model vocabularies. These values are the between-model correlations of within-model cosine similarities. Compared to the correlations that Rodriguez and Spirling (2020) report between a pretrained GloVe model and word2vec models trained on the *Congressional Record*

corpus, which range between 0.3-0.5 for randomly selected words and 0.5-0.7 for political concepts, the distribution of correlation values is somewhat lower on the whole, but in the same range. (Rodriguez and Spirling pick out only a handful of words for comparison rather than systematically comparing every word in the vocabulary.) "Community" has a between-model correlation of only 0.151 (95% confidence interval from bootstrapped models: 0.116-0.153). This falls at the low end of this distribution and indicates only a weak association of all cosine similarities across the models.

- Figure 2.6 shows **linguistic drift** (Hamilton et al. 2016a). These values are the between-model cosine distances for every word, after aligning the matrix of soc.motss embeddings to the matrix of GloVe embeddings. The cosine distance for "community" is 0.492 (95% confidence interval from bootstrapped models: 0.471-0.560). This is substantially below the average distance of 0.672, implying that by this semantic measure community differs less than the typical word. Subjectively, these distance values seem high in general – in a previous experiment using historical word embeddings from Hamilton et al. (2016b) to replicate the work of Kulkarni et al. (2015), I found that the word "community" shifts by a distance of only 0.403 from 1900 to 1990. By comparison, the word "gay," which undergoes a strong shift in meaning, changes by a distance of 0.822 over the course of the same century. The GloVe and soc.motss corpora are from similar time periods, so I surmise that these differences arise from distinct linguistic styles – formal newswire and Wikipedia articles versus less formal social text (McCulloch 2019). (Different model architectures are another possibility.)

- Figure 2.7 shows **cultural shift** (Hamilton et al. 2016a). Mathematically this is an intermediate measure between the other two – it is also a cosine distance, but

of within-model cosine similarities to a given focal word, of the local neighborhood of words around that word. Hamilton et al. (2016a) develop this "cultural shift" measure on the premise that these neighbors are semantically relevant in a way that more distance words (which were included in the first correlation measure) are not. On this metric, "community" shows a slightly above average shift of 0.303 (95% confidence interval from bootstrapped models: 0.177-0.396, median cultural shift for all words = 0.264). The wide bootstrapped interval, however, implies that this is the least stable of all three measures.

Taken together, these measures show that "community" differs in meaning to some degree between the generic GloVe model and the local soc.motss corpus, but they offer no definitive conclusion on the comparative magnitude of that change.

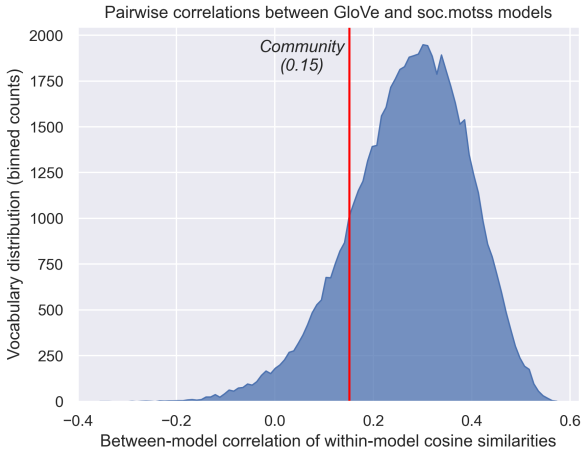

Figure 2.5: Correlation (between the two models) of (within-model) cosine similarities for every word and all other words.

Thus far, I have decomposed the structure of meanings for "community" in a general English language model and shown how it brings together geography and Gemeinschaft. I have fit a second word embeddings model on a corpus from the LGBTQ virtual com-

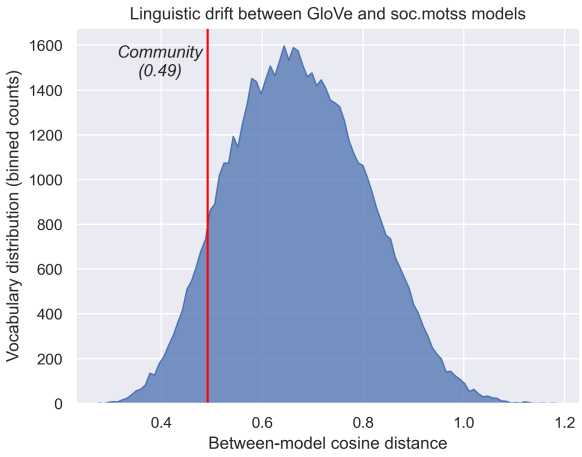

Figure 2.6: Between-model cosine distances for every word ("linguistic drift", Hamilton et al. (2016a)).

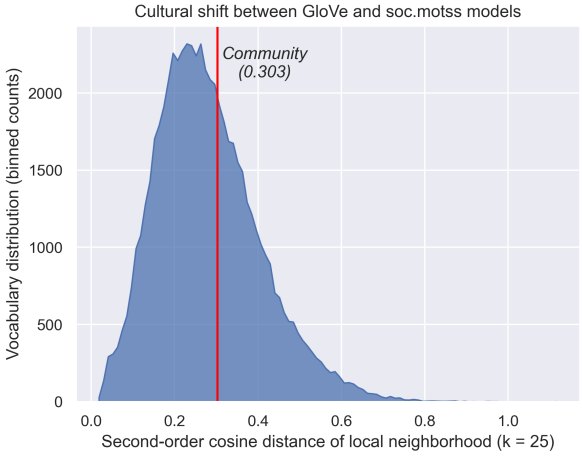

Figure 2.7: Nearest-neighbor distances, k = 25 neighbors ("cultural shift", Hamilton et al. (2016a)).

munity soc.motss, which shows some semantic overlap with general usage but not the same structure. I have compared the two models and shown a mixed picture of how distinctive this newsgroup's usage of "community" is. With this thorough foundation, I am able to address the core question: *how* does the meaning of community in the context of soc.motss differ from the generic context, and how does it remain similar? If community-as-Gemeinschaft is salient rather than ambient in a virtual community for LGBTQ people, then I would expect that semantic dimension to predominate, to a greater degree than in the general case.

Using the semantic dimensions I derived from the GloVe model in the previous section, alongside the additional identity-related dimension I curated based on the soc.motss corpus and model, I investigate what changes and why between the two different embeddings for "community." Because these averaged semantic vectors are intrinsically more likely to be closer to the "community" vector derived from the same model, I recreate each of them from the other model's embedding matrix, and I show both sets for consistency. The soc.motss model matrix is aligned to the GloVe model matrix, as are the 40 bootstrapped models for a 95% confidence interval (although the values from the true corpus are in at one case systematically higher than the bootstrapped values). In total, I make 12 comparisons (3 comparison vectors x 2 source models for those vectors x 2 source models for the "community" vector), shown in Figure 2.8.

This key figure shows how "community" in the context of soc.motss de-emphasizes geography but foregrounds identity instead. At the same time, it retains the Gemeinschaft connotations of community.

- When the averaged vector of geographic words is derived from the GloVe model, the gap in cosine similarities is the largest among any of the comparisons: the soc.motss embedding for "community" has a cosine similarity with the geography

vector of only 0.165 (0.108-0.185), compared to 0.594 for the GloVe "community" embedding. The gap narrows to near 0 when the geography vector is re-derived from the same 10 words in the soc.motss model, with a similarity of 0.398 (0.374-0.437) for "community" in soc.motss vs 0.404 for "community" in GloVe – but the following two semantic dimensions have wider gaps in the opposite direction when constructed in this way.

- The Gemeinschaft vector in the GloVe model is almost identical to the geography vector in its similarity to the GloVe embedding for "community" (0.586), but its similarity to the soc.motss "community" embedding increases to 0.398 (0.374-0.437) – still lower, but a narrower gap. Recreated with the same words from the soc.motss vectors, the similarities flip: the similarity to "community" from GloVe is 0.386 (almost the same as the comparable geography vector), but the similarity to soc.motss "community" is now higher, at 0.508 (0.440-0.511).
- The averaged vector for LGBTQ identity-related words is unambiguously more similar to "community" in soc.motss than in the GloVe model, no matter which model's word vectors are used to derive it. The cosine similarities to the soc.motss "community" embedding are 0.497 (0.429-0.485) when derived from the GloVe model and 0.627 (0.585-0.641) from the soc.motss model (the highest within both sets of comparisons). These compare to similarities of only 0.367 and 0.333 for the GloVe "community" embedding respectively (the lowest in both sets of comparisons).

This approach offers clear, stable rankings and comparisons, showing that "community" as used in the language as a whole is equally similar to both geography and Gemeinschaft, and less so to identity. By contrast, "community" as used in soc.motss carries stronger connotations of identity, roughly comparable connotations of Gemein-

schaft, and weaker connotations of geography.

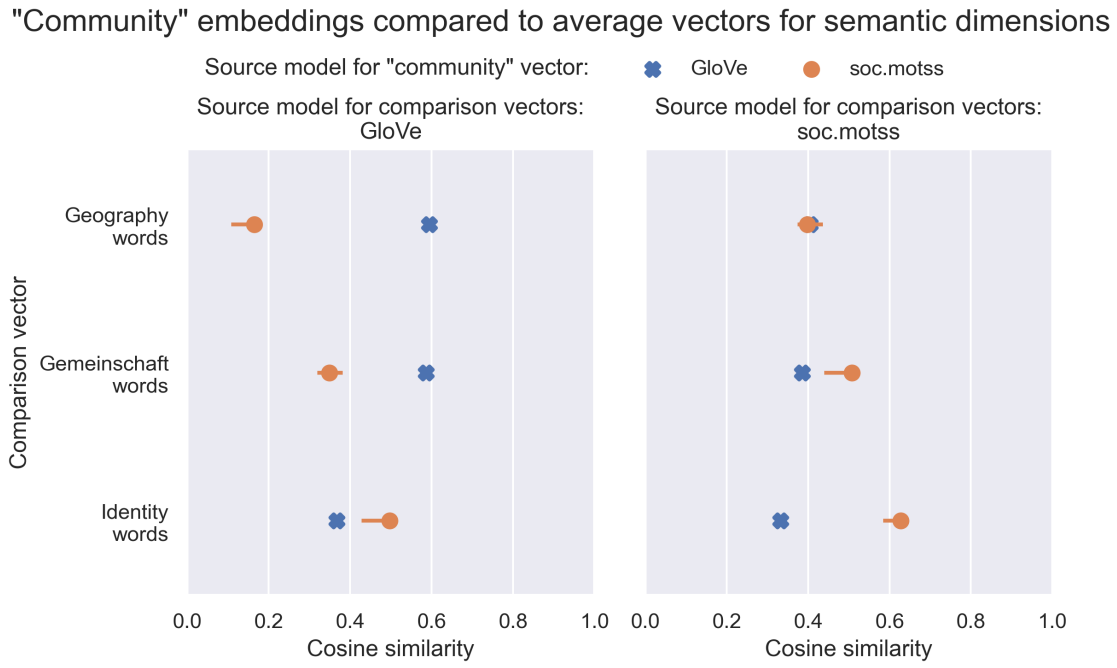

Figure 2.8: "Community" embeddings from GloVe (Wikipedia 2014 + Gigaword 5) model and word2vec soc.motss (Usenet) model compared to semantic dimension vectors based on averages of 10 words.

Finally, to provide even stronger evidence to confirm that conclusion, I apply the orthogonal projections from the end of the previous section. These come from the GloVe model alone, and I compare them to the "community" embeddings from both models in Figure 2.9. (For reference, I also compare the two "community" embeddings themselves in the third row of the figure.) Both orthogonal projections of the GloVe "community" embedding – away from the Gemeinschaft words and away from the geography words – retain high cosine similarities to the original vector, 0.810 and 0.804 respectively. However, these derived variations of the community concept are not equally similar to the soc.motss "community" embedding. Community, in the context of soc.motss,

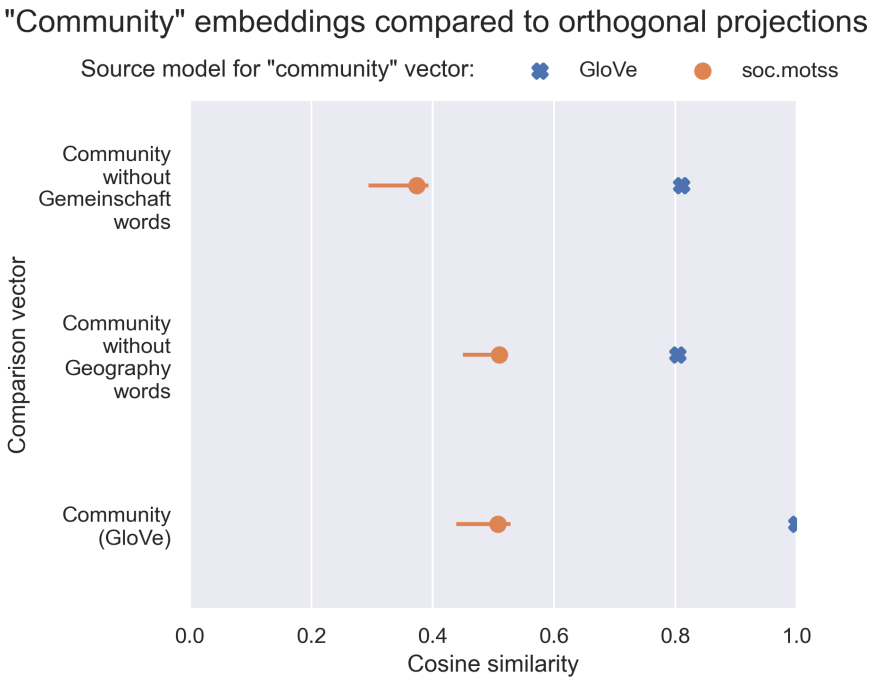

Figure 2.9: "Community" embeddings from GloVe (Wikipedia 2014 + Gigaword 5) model and word2vec soc.motss (Usenet) model compared to orthogonal projections of GloVe "community" embedding away from semantic dimensions.

is markedly less similar to the derived concept of community-without-Gemeinschaft (i.e., community-as-geography), with a similarity of only 0.374 (0294-0.393). By contrast, it remains just as similar to community-without-geography (or, community-as-Gemeinschaft) as it is to the GloVe "community" embedding overall, with similarities of 0.510 (0.451-0.523) and 0.508 (0.440-0.529) respectively. (The magnitude is lower than 0.804 because the comparison vectors are derived from the GloVe model originally.) This way of examining the semantic similarities and differences of "community" between soc.motss and general English usage confirms that Gemeinschaft is present in both.

2.4 Discussion

The structure of meaning around "community" in the general GloVe embeddings shows, mathematically, how the word bundles and links together two sets of connotations: on the one hand spatial or geographic, and on the other hand social or *gemeinschaftlich*. This dual structure might appear intuitive, but it was not predictable in advance.

This structure does not fully carry over into the local context of an LGBTQ virtual community. In soc.motss, the meaning of community is not totally alien or unrecognizable from the general meaning (and why would it be?), but it is clearly distinct. Across both contexts, the sociological element of community is present in everyday usage. In the soc.motss corpus, "community" retains the *Gemeinschaft* connotations of social organization and groupness, but incorporates markedly less of the geographic or spatial connotations of community. Instead, it substitutes language related to LGBTQ identities. Queer identities are salient and central to the way the members of soc.motss discuss community, at a time of significant political shifts toward greater societal acceptance and equality (Rosenfeld 2017); they are not absent, invisible, or ambient. Moreover, this

case affirms the presence and visibility of LGBTQ community in the process of creating virtual community in general (Auerbach 2014; boyd 2014; Rheingold 2000).

Potential limitations of this work relate to robustness, scope, and generalization. While I am convinced that the findings I present here are robust, a key limitation of any text analysis work centers around corpus and model quality. It would be possible to do even more to evaluate the stability and viability of the locally-trained embeddings (see Antoniak and Mimno 2018), or to incorporate corpora from additional groups – although there is little evidence from prior work to expect that the results would differ dramatically (Dame-Griff 2019). This work also demonstrates that there is some payoff to closely interrogating the embedding representation of a single word, provided that the associated concept has enough theoretical and empirical complexity to warrant a close read. The downside, then, is that generalizing to other complex concepts is inherently slow and requires interpretative work.

This chapter analyzed the meaning of "community" at a discursive and cultural level. While I qualitatively contextualized the social environment of Usenet and the soc.motss newsgroup, I did not explicitly connect any social characteristics to experiences of and expressions of community. In the following chapters, I will bring in those characteristics, by characterizing how features of places relate to individual experiences of community and how features interaction networks relate to group expressions of community.

2.5 Acknowledgments

I thank Avery Dame-Griff for sharing code, advice, and encouragement for working with Usenet archives, and Steve Goodreau for the initial suggestion to look more closely at soc.motss in particular. Previous versions of this chapter received generous feedback from the Community Data Science Collective and the Metaphors and Meaning

roundtable at ASA 2021.

3 Density and abundance

How place characteristics shape individual sense of community for LGBQ people

For a community to exist, people have to *feel* that one exists, and that they're a part of it. How do they come to feel that way? This chapter investigates one aspect of that process, by using nationally-representative survey data from LGBQ people in the United States to assess how contextual characteristics of places are associated with individual experiences of sense of community and belonging. I identify two key place-based elements that might contribute toward creating a sense of community around LGBTQ identity, namely density and minority abundance.

Community itself is a meso-level social entity that comes into existence through the overlap of social interactional density and shared cultural or moral density (i.e. group styles, embodied habitus, values). Geographical proximity – what Durkheim called physical density (Durkheim [1912] 2001; Tavory 2016) – often facilitates the emergence of this collective entity, but it isn't intrinsically essential. But if community is a collective and sociological construct, *sense of community* is more of a psychological and individual one – the personal experience of belonging to a larger collective entity. (Which, unless we want to contend that people are delusional, implies the existence of that social entity.)

The motivation for this work, then, is the fact that an individual's sense of community

is shaped by structural features of the contexts in which they are embedded (Boessen et al. 2014). Purely virtual and distributed communities (Driskell and Lyon 2002) aside, experiences of community are typically local and emplaced (Brint 2001). At the broadest level, then, place characteristics might shape how and whether individuals experience a strong sense of community. Place characteristics most obviously matter for place attachment; but they also matter for other objects of belonging, like identity-based groups.

LGBTQ people provide a particularly interesting case to study the strength of community. For marginalized and minority groups, the stakes of belonging are heightened in light of a history of stigma and exclusion; the place characteristics that matter might differ from those that facilitate community for the generic, unmarked majority (see Zerubavel 2018 for a discussion of markedness). For instance, whether homogeneity lead to a feeling of cohesion, or whether it is instead stifling, depends on who a person is. Because LGBTQ experiences of community have the potential to be unique, I'm interested in what those experiences might reveal about the relationship between place and community. Moreover, LGBTQ attachments to community are already known to be emplaced, both through the existence of archetypical institutions like gay bars (Mattson 2020) and gayborhoods (Ghaziani 2014b) as well as through broader and more diffuse constellations of significant places (Gieseking 2020). By linking features of places to individual LGBQ experiences, I aim to uncover the conditions that facilitate sense of community.

My broad research question, then, is **what features of individuals and places are associated with individual sense of community for LGBTQ people?** I focus on the role of two place characteristics in particular: overall population density, and the prevalence or abundance of LGBTQ people (using same-sex couples as a proxy

measure). Specifically, do dense places full of LGBTQ people (and institutions) facilitate a greater sense of connection to the LGBTQ community? Or, conversely, are those exactly the places where LGBTQ community fades into the background? To preview my results, I ultimately find more support for the former hypothesis, with moderate, mixed evidence for density mattering for community in a positive way, and strong evidence for the prevalence of LGBTQ people mattering in the same way. Of course, this analysis only uncovers statistical associations from observational data rather than causal relations, focusing on relatively durable place characteristics and controlling for relatively fixed or seldom-changing individual traits (neither of which are great candidates for observational causal inference).

A note on terminology before I proceed – the alphabet is complicated. The Generations survey (Meyer 2020) purports to be a study of *LGB* people, but it also includes people with other sexual minority identities. It asks these *LGBQ* respondents about *LGBT* community – i.e., trans people are included in the imagined community, even though they are not among those screened into this survey. It might be analytically convenient if the bounds of identity inclusion represented in these acronyms were consistent, but that's not how LGBTQIA+ identities, communities, and discourse work. I would interpret the survey questions as gesturing toward an expansive and inclusive imagined LGBTQ community, and I will write about LGBTQ community or queer community when I mean to speak generally rather than about the specific questions or respondents.

3.1 Background

My baseline expectation is that a dense place-based context with an abundance of queer people and institutions will facilitate a correspondingly strong sense of belong-

ing and connectedness to the LGBTQ community for individual survey respondents, with ancillary spillover benefits for belonging and wellbeing overall. By density, I simply mean physical density of people; since Durkheim, sociologists recognized that this physical density facilitates a social density of interactions (Durkheim [1912] 2001; Tavory 2016). Moreover, cities are historically entwined with the formation of collective LGBTQ identity (D'Emilio 1992), so much so that LGBTQ studies has been critiqued for its metronormative emphasis (Halberstam 2005). A main mechanism for density to contribute to a sense of community for LGBTQ people is *movement*; this stereotypical story of the attractive force of gay urban life (D'Emilio 1992) has also been called the "great gay migration" (Weston 1995). This means that the prevalence or abundance of LGBTQ pople is likely to be a specific draw that also positively contributes to experiences of community, especially given that most LGBTQ people do not necessarily grow up around many other queer people or with access to queer spaces. In any case, it's easy to imagine concentrations of LGBTQ life as an attractive force, leading queer people to concentrate around each other and to individually experience a greater sense of community as a result.

More broadly, however, the case for cities as sites of community is more ambivalent. Though some have advocated for cities as enablers of collective life (Jacobs 1961), classic sociological work instead sees urban life as facilitating either individualism (Simmel [1903] 1971) and the substitution of *Gemeinschaft* for *Gesellschaft* (Tönnies [1887] 2001). Queer urban migration may fit into the broader social process of what's been called the "big sort" (Bishop 2009), where people have self-selected into geographic regions according to values, lifestyles, and other factors. If those that remain in less-dense places with fewer LGBTQ people are also self-selected in this same way, then that might lessen the differences between rural and urban LGBTQ people according to geographic

context.

The qualitative literature on LGBTQ collectivities additionally hints at some alternative possibilities. Perhaps individuals' perceived sense of belonging and connectedness to the LGBTQ community does not correlate with – or even runs counter to – the abundance, density, and diversity of LGBTQ people and institutions in a place. If that were the case, that would suggest that the perceptual experience of community might stand in and symbolically substitute for structural and demographic "facts on the ground," rather than complementing or arising from them. In other words, maybe a place like San Francisco winds up being like Brown-Saracino (2017)'s Ithaca, and "ambient community" there takes the place of LGBTQ community specifically. Winer (2020)'s related key finding of "solidarity with disdain" might have place-based limitations as well – his interviewees, after all, come only from Southern California. In smaller and more scattered contexts LGBTQ people might not be able to afford to symbolically distance themselves from the imagined center of the gay community the way Winer's respondents frequently do. Finally, Forstie (2020)'s study of LGBTQ communities in small cities raises the possibility that those LGBTQ communities might in fact be less fragmented and more cohesive, especially across lines of difference, compared to communities in larger cities.

The data sources I'll combine allow me a unique opportunity to evaluate theories from this qualitative literature, by complementing their depth with breadth and scale. For instance, while Brown-Saracino (2017) gains analytic leverage by looking at places that are very similar to each other on the surface and exploring their differences, she can't explore the full space of places where LGBTQ people live and where they might experience community (or not). But the three kinds of place-based identities she uncovers can be mapped onto community connectedness measures like those in the Generations

study. In her language, "hybrid identities" and traditional lesbian communities would *both* be consistent with a strong sense of belonging to the LGBTQ community. By contrast, "ambient community" would be signaled by high belonging in general but low LGBTQ-connectedness specifically. Thus, those sense-of-community survey measures have the potential to be informative even though they can't distinguish what, exactly, "LGBTQ community" means in a given place or what precise forms it takes. (While the Generations survey asks about connections to "*the* LGBT community," actual LGBTQ communities are often multiple or fragmented.)

One of the challenges of studying experiences of community is the slipperiness of the referent – what, in a given case, is "the community"? A second important challenge is to distinguish community from related concepts like identity, even when a community might be based on a particular social identity. One of the unique aspects of the Generations survey is that it actually captures (some of) this complexity. Most importantly, it asks respondents about their connection and belonging to *any* community and to *LGBTQ community specifically*. It also includes separate questions about LGBTQ community and identity, which matters because community connectedness is not the same construct as identity salience, although the two are presumably related. I use the two distinct measures of community connectedness as my primary outcomes and include individual characteristics (including sexual orientation, gender identity, race, age) alongside my key place-based measures as covariates that capture potentially-salient sources of variation.

Of course, the two factors of density and abundance aren't exhaustive of spatial characteristics that might matter for community. Material resources and institutions – especially the presence and concentrations of third places like bars and coffee shops (Oldenburg 1998) – no doubt could positively contribute to a sense of community as

well. Queer anchor institutions (Ghaziani 2014a), however, are likely to coincide with the presence of same-sex couples. In addition, structural and demographic characteristics have the drawback of being overly broad, in that they don't capture the specificity of a particular place (Gieryn 2000) – which might be more or less conducive to community. This remainder can be conceptualized as *place-based culture* (Brown-Saracino 2017). Place narratives might even mediate the relation between structural features and individual experiences of community; indeed, ethnographers say these stories matter (Brown-Saracino 2017; Orne 2017), that they're part of how community plays out differently in practice even in places that appear similar on the surface. I'm unable to address those fundamental limitations in this chapter; the geographic characteristics I analyze here can only measure the background context in which individual social worlds take place.

3.2 Data and methods

The contribution of this chapter is to analyze the association between place characteristics from the ACS (at the ZCTA and MSA levels) with individual survey responses about community and belonging in wave 1 of the Generations study.

The key data source for this chapter is the *Generations* study (Meyer 2020), a three-wave representative panel survey of three age cohorts of cisgender Black, white, and Latinx lesbian, gay, bisexual, and queer and other nonheterosexual (LGBQ) people in the United States. The publicly-available version of the data set includes multiple measures of sense of community and belonging as parts of two composite scales: a generic *Social Wellbeing* scale and an LGBT-specific *Community Connectedness* scale. The public data, however, do not include geographic location beyond urban/rural and Census region. The restricted portion of the data set records respondent locations at

more granular geographies: state, metropolitan/micropolitan statistical area, and finally zip code. To maintain respondent privacy, these restricted data are held by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan and accessed through a secure virtual environment, with a data use agreement to maintain confidentiality. (One key provision of this agreement: I cannot disclose *which* geographic locations are actually observed in the Generations data, although I present summary statistics below.) The University of Washington IRB approved the use of the restricted data for this study.

I use four measures from wave 1 of the Generations study - both the two full scales and the first single item from each - as outcome variables to represent community and belonging. The multi-item scales are validated and robust. By contrast, single items from those scales are framed to more precisely target *belonging*. While it's important to justify deviations from standard, validated scales (Mustillo, Lizardo, and McVeigh 2018), it's also important not to over-project meaning onto latent constructs derived from survey measures or take those meanings for granted (Martin and Lembo 2020, 2021). As a compromise, I analyze both and discuss any noteworthy divergences below. These are the scales and items I analyze:

- The **LGBT Community Connectedness** scale, a 4-point scale constructed from an average of 7 items, previously validated as a cognitive/affective construct in Frost and Meyer (2012). Items touch on emotional attachment, participation, and political and collective action within "the" (singular) LGBT community.
- The **Social Wellbeing** scale, a 7-point scale constructed from an average of 15 items. This scale includes not only items that ask about respondents' relations to community, but also to society and the world more broadly.
- The first item on the Community Connectedness scale, "You feel you're a part of

the LGBT community," with four response options ranging from "Agree strongly" to "Disagree strongly."

- The first item on the Social Wellbeing scale, "I don't feel I belong to anything I'd call a community," with seven response options ranging from "Strongly disagree" to "Strongly agree." (*Strongly disagree* with this negative statement corresponds in direction to *agree strongly* on the LGBT Community Connnectedness question, so I align my model results and figures accordingly.)

For place-based data, I draw on the US Census Bureau's American Community Survey via the tidycensus package (Walker and Herman 2023). At the smallest levels of geography, only pooled 5-year ACS estimates are available. I use the 2019 estimates (from 2015-2019), which overlap with wave 1 of the Generations survey (2017-2018) and also have the highest-quality same-sex couples data. I uploaded relevant covariates, for all ZCTAs and MSAs, to ICPSR's virtual data enclave, and then joined and filtered them to only the zip codes and MSAs from which respondents in the Generations study were sampled.

In this paper, I consider two focal place-based covariates from the ACS: population density and prevalence of same-sex couples. I analyze these place variables at two geographic levels available in the restricted Generations data - ZCTA and MSA. At the MSA level, I use population-weighted densities aggregated up from the zip code level, to account for the fact that MSA boundaries - derived from counties - vary hugely across different regions of the country (Ottensmann 2018). I focus primarily on results at the zip-code level, which turn out to be more substantively and statistically significant.

In my statistical models, I test four different functional specifications of the association between population density and community/belonging:

- **Linear.** The most straightforward specification, allowing me to detect whether

denser places are associated with a stronger or weaker experience of community.

- **Quadratic.** A theoretically-informed extension of the linear specification. This allows for the possibility, discussed above, that both rural small towns and large urban environments are more conducive to community and belonging than suburban sprawl.
- **Logarithmic.** A data-driven transformation, because the distribution of ZCTA population densities is right-skewed. Taking the (base-10) log produces a more normal distribution of the variable. (Because of this same skew, I present plots involving population density on a log scale below.) Like the linear specification, the log transformation can only be monotonically increasing or decreasing.
- **Spline.** Agnostic and flexible, using a generalized additive model (GAM) to learn a potentially nonlinear functional form from the data.

As I'll show below, while the more complex specifications sound plausible, there's no evidence to support their use.

I use only one specification for the prevalence/abundance of LGBTQ+ people: the proportion of households that are same-sex couples. There are no fine-grained geographic estimates of LGBTQ+ identity (and based on Census Bureau trends, there never will be), so partnered households are the best proxy. I considered instead using *counts* of same-sex couples, as well as total population counts, rather than constructing two variables that are essentially *rates*. (Note: Because I pull same-sex couple counts and household counts from separate ACS tables, a few zip codes have nonsensical or extreme values for percent same-sex couples, which I drop from summary statistics. These are zip codes with small populations and/or high proportions of residents in group quarters.) It's an open question whether the *proportion* or *count* of same-sex couples is theoretically more important, but this alternate (and potentially simpler) specification did not turn

out be especially promising, so I did not explore it systematically.

I control for a set of individual demographic, socioeconomic, and other identity-based characteristics from the Generations data: gender (cisgender women, cisgender men, and [some] nonbinary/genderqueer people; transgender potential respondents took the parallel TransPop survey, not the Generations survey), sexual orientation (lesbian/gay, bisexual, queer/other), age cohort (younger, middle, older), race (white, Black/African American, Latino/Hispanic), political affiliation (Republican, Democrat, independent/other), and education (high school or less, more than high school). Some of these traits are associated with community and belonging in interesting ways I won't discuss in this paper. Individual covariates can be analyzed with the publicly available version of the Generations data set, and so they're peripheral to my central aim of analyzing the restricted geographic data. (Most notably - and unsurprisingly - the small minority of LGBQ Republicans report low LGBT community connectness and belonging. This finding was previously reported in a bivariate analysis (Meyer and Choi 2020), but my results show that it holds up in the presence of other controls.) Others of these controls aren't associated with one or another of the outcomes, but I retain them for consistency across models.

Beyond potentially being associated with community/belonging, these individual traits vary spatially. Race is of course central to the spatial demography of the United States, and political affiliation is markedly geographically structured as well (both regional and urban/rural divides). In terms of gender and sexuality, gay men are more likely than lesbian women to concentrate in urban centers (Black et al. 2000). Education level affects opportunities to relocate, and respondents of different ages may be at different points in their life course that affect where they choose to live (e.g., younger respondents in cities, older respondents in suburbs).

The marginal effects plots I show below present predicted values and predicted probabilities with the individual controls set at their reference categories: young white lesbian Republican women with a high school education or less. Of course, reference categories are political and theoretically important (Johfre and Freese 2021), but the choice of reference categories does not affect my main results. (Most notably, "white" and "Republican" both shift the level of LGBT belonging and connectedness downward compared to Black/Latinx or Democrat/independent, but they do not alter the interpretations of the place covariates.)

With 4 outcomes, 2 geographic levels, and 4 functional specifications of density, my main results consist of a series of 32 statistical models – although many of these models turn out to be uninformative. I model the multi-item scale outcomes (which can take fractional values) with linear regressions, and the single-item outcomes with ordered logistic regressions. For the spline functional specification of population density, I use GAM extensions of both model types (from the mgcv package (Wood 2011)). As a robustness check for the multi-item scale outcomes, I test two additional varying-intercept multilevel models grouped at the MSA level. This is not a fruitful approach, but if it had been, it would have been appropriate to recast many of the other models as multilevel models.

3.3 Results

First, I'll describe the distributions of the relevant variables, to contextualize the main statistical models that follow. Table 3.1 shows that the typical Generations respondent (47.2%) agrees that they feel a part of the LGBT community, and this corresponds closely to the overall 7-item average of 2.97 on the Community Connectedness scale. The scale is flipped, so that higher numbers represent greater connectedness, and a

respondent who agrees with every item would receive a score of 3. Consistent with the LGBT-specific trend, respondents tend on average to disagree with the notion that they don't belong to any community, and this aligns with the 4.67 average response for the 15-item Social Wellbeing scale (again, aligned so that higher values represent more positive outcomes). All four measures point to a moderately positive sense of community and belonging on average, but with enough variability to attempt to model systematic differences among respondents and the places in which they live.

Table 3.4 shows the distributions of the two key place covariates for the zip codes and metropolitan areas represented in the *Generations* study. Population densities vary widely across zip codes (sample SD = 5,200 individuals per sq. km), and a strong right skew is evident where a tail of zip codes are especially densely populated. The mean respondent lives in a zip code with 2,500 individuals/km^2, while the median respondent lives in a zip code with only 1,000 individuals/km^2. The population-weighted mean density for all populated zip codes is 1,560 individuals per sq. km, meaning that LGBQ individuals in this representative sample live in zip codes that are on average substantially denser than the American population at large. Same-sex couples are around 1.1% of the households in the average zip code represented in the study, with a slightly lower median (0.8%) and a reasonable amount of variation (SD = 1.1%). As with density, the average is higher than the population-weighted mean for all populated zip codes of roughly 0.75%. (This is shaped by the fact that 54% of populated ZCTAs recorded 0 same-sex couples in the 2015-2019 ACS time period. Because the ACS is not a census and same-sex couples are rare, many of these are not likely to be true zeroes.) On both distributions, especially percentage of same-sex couples, note that MSAs show much less variability than zip codes.

These covariate distributions shape how I model, present, and interpret my results.

Table 3.1: Outcomes

(a) Individual item outcomes

	N	%
You feel you're a part of the LGBT community.		
Agree strongly	239	17.4%
Agree	649	47.2%
Disagree	389	28.3%
Disagree strongly	99	7.2%
I don't feel I belong to anything I'd call a community.		
Strongly disagree	252	18.3%
Moderately disagree	297	21.6%
Slightly disagree	227	16.5%
Neither agree nor disagree	142	10.3%
Slightly agree	198	14.4%
Moderately agree	168	12.2%
Strongly agree	92	6.7%

Source: *Generations* study (Meyer 2020)

(a) Scale outcomes

	Mean	Std. dev.
LGBT Community Connectedness scale (1-4)	2.97	0.56
Social Wellbeing scale (1-7)	4.67	0.91

Source: *Generations* study (Meyer 2020)

The distribution of population densities, in particular, informs my decision to display predicted values on a log scale. For interpreting magnitudes, you can anchor on the idea that around half of zip codes are above and below 1,000 individuals per sq. km, and around half are above and below 1% same-sex couple households

Table 3.4: Place characteristics

	Median	Mean	Std. dev.
Zip codes in the Generations study (N = 1,238)			
Population density (individuals/sq. km)	1,000	2,500	5,200
Percent same-sex couple households	0.8%	1.1%	1.1%
MSAs in the Generations study (N = 217)			
Weighted population density (individuals/sq. km)	1,200	2,050	2,540
Percent same-sex couple households	0.9%	0.88%	0.23%

Source: 5-year American Community Survey, 2019

Note: Characteristics for zip codes and metropolitan areas represented in the Generations study. Values are rounded to maintain privacy. Values are weighted by number of respondents, meaning that these are the values *experienced* by the average respondent.

Two examples drawn from the full set of zip codes will provide anchor values for those distributions and help contextualize the model results that follow. These are shown in Figure 3.1. (Remember, I can't discuss which zip codes and MSAs are actually included in the Generations study.) I've chosen these examples as quantitative outliers corresponding to culturally significant places, illustrating what a place where 10% or more of households are same-sex couples actually look like.

- Zip code 94114 encompasses the Castro, San Francisco's gayborhood and one of

the most prominent gay neighborhoods in the country. 12% of households in the area are same-sex couples. The Castro is in a densely-populated residential part of the city (9,500 individuals per sq. km), and adjacent to the extremely dense downtown core of San Francisco (with densities reaching 20,000 individuals per sq. km). San Francisco is is in turn the densest part of the wider Bay Area, and one of the densest major cities in the country.

- Zip codes 92262 and 92264 coincide closely with Palm Springs, a gay resort town in Southern California. 12% and 14% of households are same-sex couples, respectively. However, as the map shows, Palm Springs is far from the most densely-populated part of the Riverside, CA, metropolitan area (population densities are 300 and 150 individuals per sq. km, respectively). In other words, it's only extreme on one place characteristic, not both dimensions. Overall densities in the area are much lower compared to the Bay Area, and more typical of the country as a whole.

These real places illustrate the plausible upper end of the range for the prevalence and abundance of LGBTQ people, proxied through same-sex couples, and two distinct points on the spectrum of population densities. In the subsequent results, moving from 1% same-sex couples and 1,000 individuals per sq. km to 10% and 10,000 individuals per sq. km is like moving from a typical zip code in the sample to a place like the Castro.

I will now describe those model results, organized as follows. I primarily discuss zip code results, and then briefly touch on metropolitan area results. Within each geographic scale, I describe first the patterns for population density, and second the patterns for percentage same-sex couple households. For each place covariate, I cover the four outcomes – first the combined scales, and second the individual questions. I close by

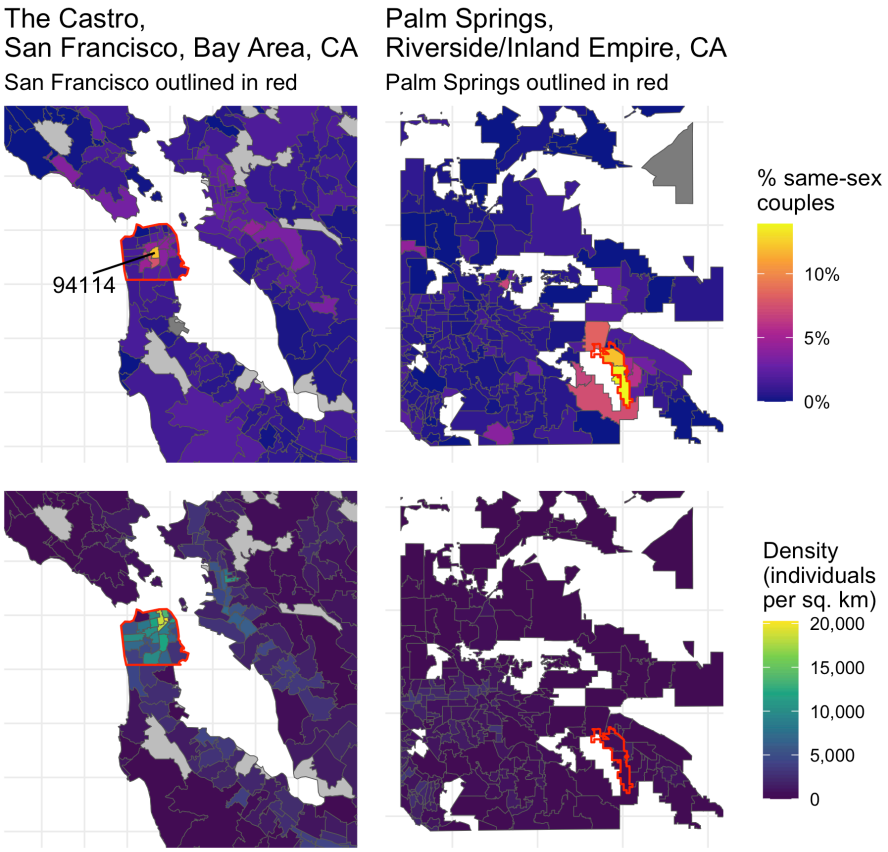

Figure 3.1: Example zip codes with high prevalence of same-sex couples: the Castro (94114) and Palm Springs (92262, 92264).

highlighting some individual-level trends for notable sociodemographic characteristics, which are independent of the place-level findings.

At the zip code level, population density shows an association with both scale outcomes, but not with the specific questions about feelings of belonging. Figure 3.2 compares predicted values for four different functional specifications for zip code population density, controlling for percent same-sex couples and individual characteristics and holding these constant at their mean and reference categories respectively. Functional form turns out to matter quite a lot; with ill-fitting specifications, it becomes impossible to estimate any clear association at all. The best-fitting models for LGBT community connectedness and generic social wellbeing have different functional forms from each other. In neither case is there any evidence for non-monotonic relationships (quadratic, spline) between density and community belonging.

Population density has a linear association with LGBT community connectedness ($\beta = 0.034$, $SE = 0.015$). In terms of predicted values (see top-left panel of Figure 3.2), this means there is little movement in community connectedness at the bottom half of the density distribution (below 1,000 individuals per sq. km), but substantial increases at very high densities. The association between population density and generic social wellbeing is better modeled as logarithmic (the third panel of the bottom row of Figure 3.2) ($\beta = 0.087$, $SE = 0.032$). An increase from very low densities (< 1 individual per sq. km) to the midpoint is associated with an increase in social wellbeing larger than from 1,000 individuals per sq. km to 10,000 or beyond, meaning that, descriptively, there are diminishing returns at higher densities. Despite being positively associated with both multi-item scales, population density is not associated with responses to either of the two single items alone, no matter the functional specification ($p = 0.16$ and $p = 0.29$ respectively). I interpret this contrast and speculate on methodological and substantive

reasons for it in the discussion.

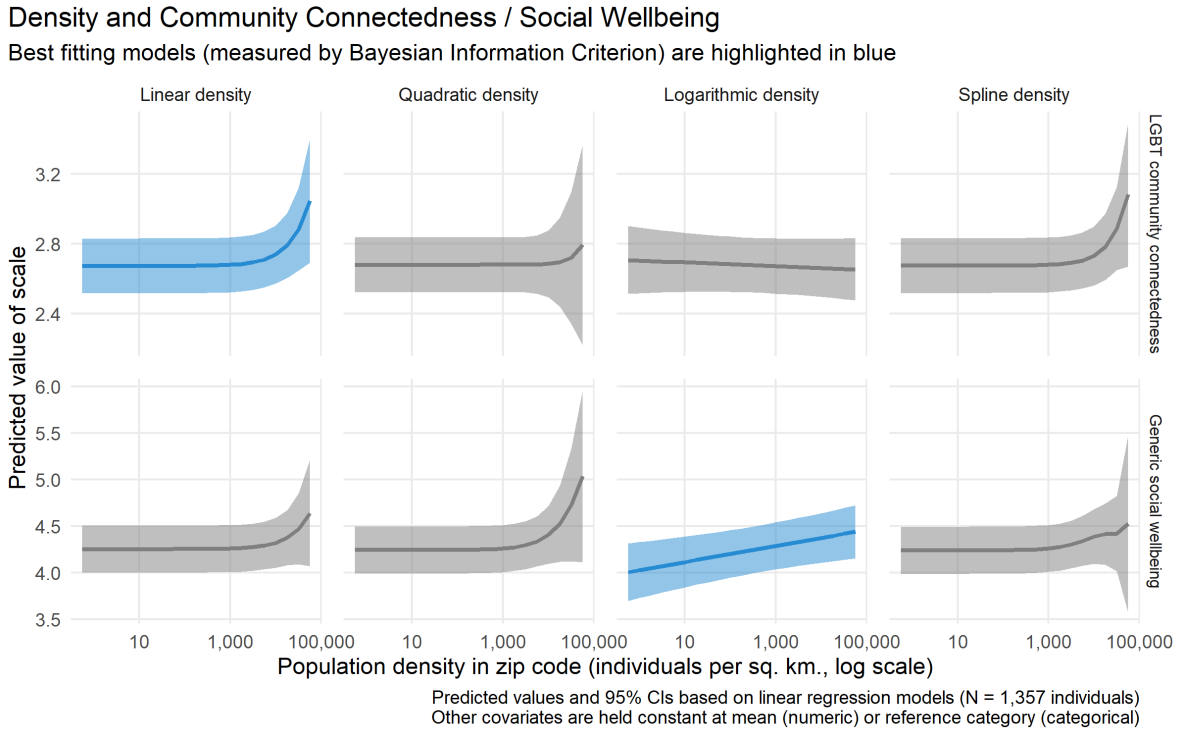

Figure 3.2: Predicted associations of zip code population density with LGBT Community Connectedness and generic Social Wellbeing scales, from multiple linear regression models with four distinct functional transformations. Best models are highlighted.

Across all four outcomes, a higher zip-code-level prevalence of same-sex couples is consistently associated with a greater sense of community and belonging for LGBQ people. As Figure 3.3 shows, a shift from 1% to 10% of households in a zip code being same-sex couples is associated with a quarter-point increase in community connectedness ($\beta = 2.32$, $SE = 1.41$) and a half-point increase in social wellbeing ($\beta = 5.88$, $SE = 2.34$). (The two scales have different ranges, so the units are not comparable. All models discussed here use the best-fitting specifications for density and control for individual

characteristics.) These strong associations are replicated in the individual question responses. With other characteristics held constant, the ordinal model in Figure 3.4 shows that as the percentage same-sex couples moves from 1% to 10%, the predicted probability of a respondent *agreeing strongly* with feeling a part of the LGBT community more than doubles from around 15% to 40%. The probability of disagreeing at all falls from 30% to near 10%. Similarly, Figure 3.5 shows that as the percentage moves from 1% to 10%, the predicted probability of a respondent *strongly disagreeing* with the statement that they do not feel they belong to any community also nearly doubles, to almost 40%. Given that a sense of belonging to LGBT community specifically ought to logically entail belonging to at least one community in general (although individual responses aren't always consistent), this consistency is to be expected. The overall signal that more same-sex couples in a zip code area are associated with more subjective community for LGBQ people is robust.

By contrast with those clear associations at the zip code level, there is less clear evidence to report at the metropolitan level. Among the combinations of the two covariates and four outcomes, only one association can be estimated sufficiently precisely to be distinguishable from zero. Even weighted for zip code population, MSA-level population density does not have a discernable association with any of the four community-related outcomes, with any functional form. The MSA-level percentage same-sex couple households is only discernably associated with the social wellbeing scale (shown in Figure 3.6), with a 1 percentage point increase associated with around a 0.3 point increase on the scale ($\beta = 27.63$, $SE = 11.65$). That change is of a similar magnitude to the level of change associated with a larger shift in same-sex couples described for zip codes above, but MSAs show much less variation in same-sex couple prevalence overall. The apparent direction of the coefficient for the generic community belonging item is consistent with

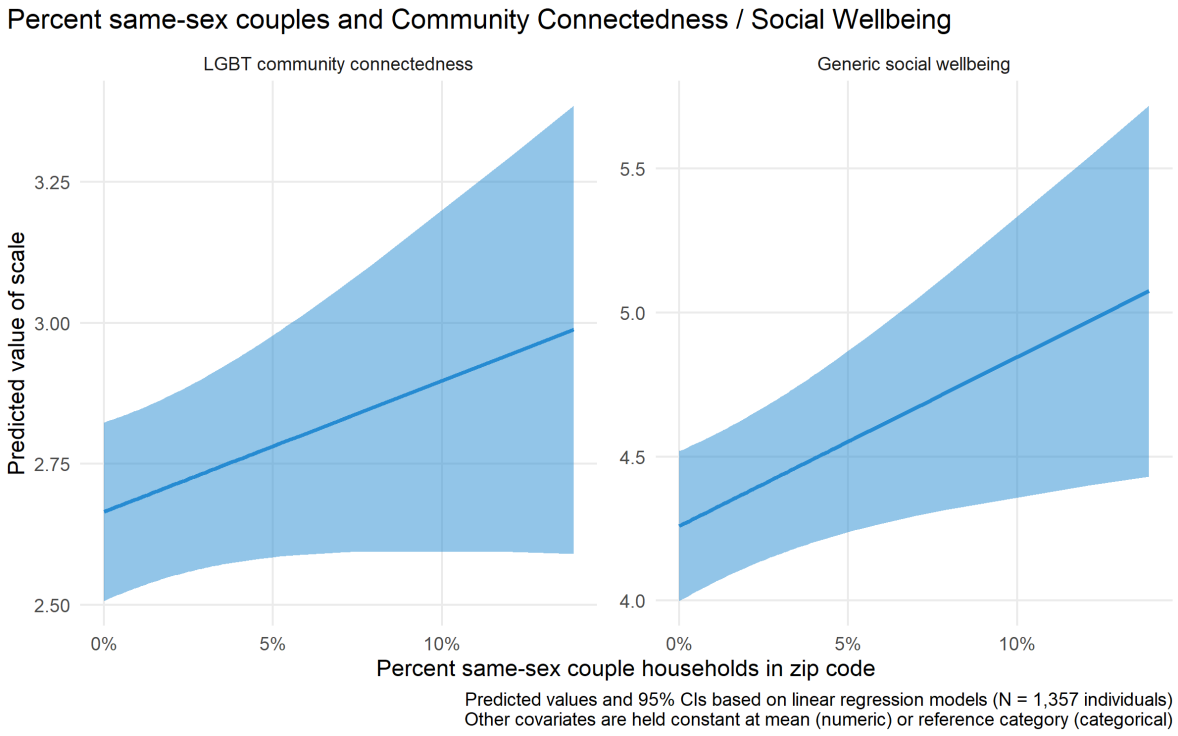

Figure 3.3: Predicted associations of zip code percentage same-sex couple households with LGBT Community Connectedness and generic Social Wellbeing scales from multiple linear regression models.

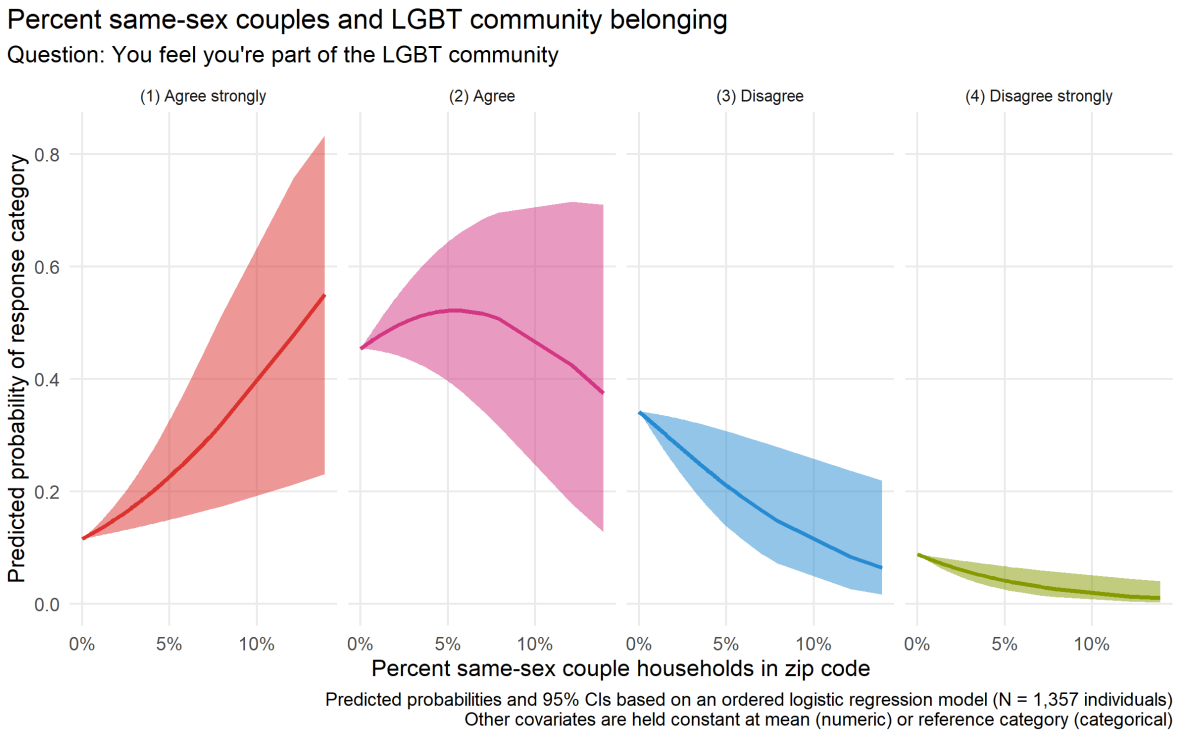

Figure 3.4: Predicted probabilities of agreement/disagreement with LGBT community belonging item by zip code percentage same-sex couple households, from ordinal regression model.

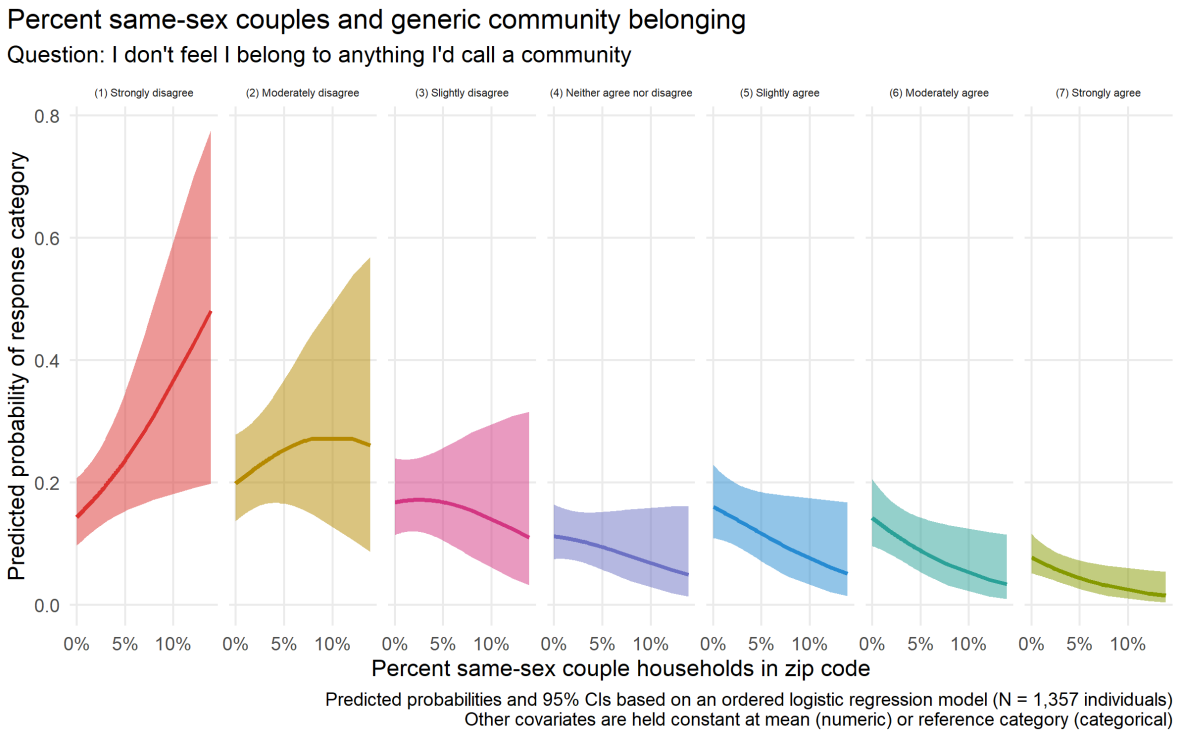

Figure 3.5: Predicted probabilities of agreement/disagreement with generic community belonging item by zip code percentage same-sex couple households, from ordinal regression model.

the social wellbeing result, but the level of uncertainty is too large to reliably distinguish it from zero. The prevalence of same-sex couples in a metropolitan area is not at all associated with either measure of LGBT-specific community connectedness. As a robustness check, even a completely different approach to capturing differences between MSAs – a simple varying-intercepts model with random effects – doesn't reveal notable between-MSA variation.

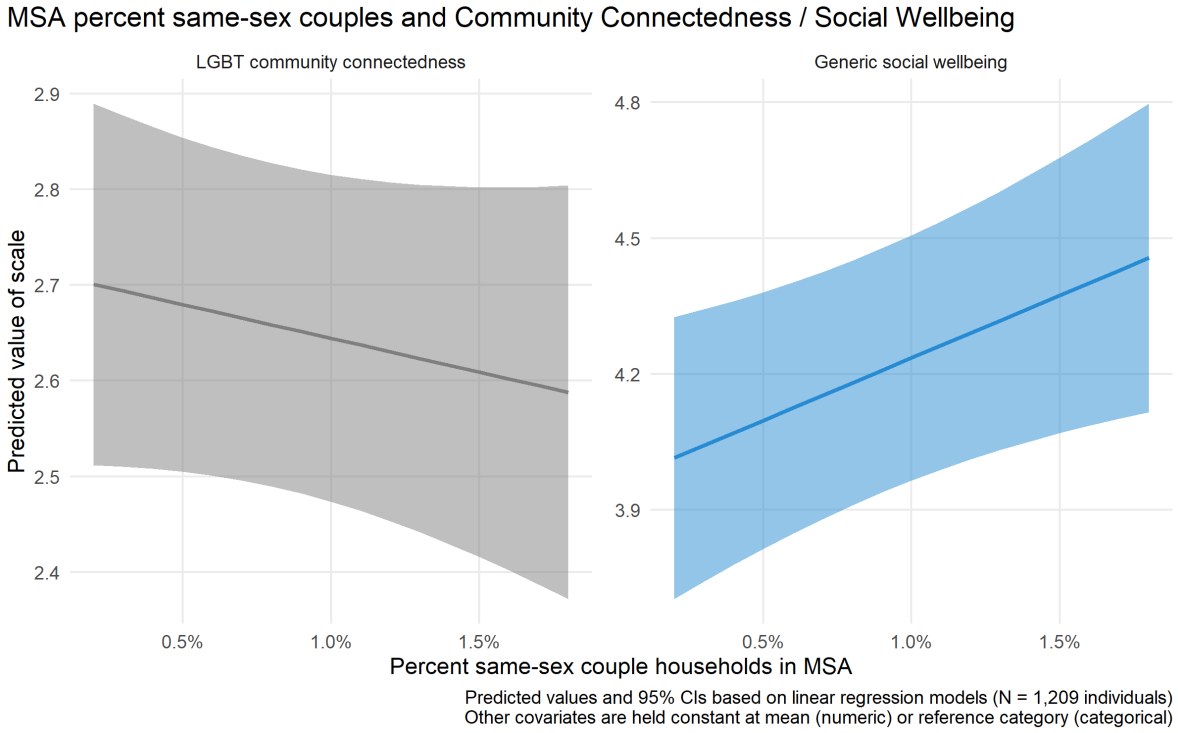

Figure 3.6: Predicted associations of MSA percentage same-sex couple households with LGBT Community Connectedness and generic Social Wellbeing scales from multiple linear regression models. LGBT Community Connectedness model is non-significant.

Finally, while not the focus of this chapter, the trends for the individual-level sociodemographic controls are worth noting. The inclusion of the place characteristics in the

models doesn't seem to affect these associations, which makes analyzing these trends a possible avenue for future work using the public version of the Generations data set.

- **Gender** has no association with feeling a part of the LGBT community or with the LGBT community connectedness scale. However, non-binary respondents are significantly less likely to report belonging to any community and report lower social wellbeing than men or women, all else equal.
- In terms of **age**, middle and older cohorts both report feeling less a part of the LGBT community and lower community connectedness belonging than the youngest cohort, all else equal. All else equal, there are no cohort differences in generic community belonging or in social wellbeing.
- In terms of **race**, Black and Latinx respondents report higher LGBT community connectness, including feeling a part of the LGBT community, than white respondents, all else equal. For the generic questions, the direction is reversed: Black respondents are less likely to report belonging to any community than white respondents, and both Black and Latinx respondents report lower social wellbeing overall.
- Among specific **sexual minority identities**, bisexual respondents report lower feelings of LGBT belonging and connectedness than lesbian/gay respondents, as well as lower social wellbeing overall; differences between lesbian/gay respondents and other sexual minority identities are not detectable.
- As previously noted, the starkest difference in feeling a part of the LGBT community and in LGBT community connectedness overall is for **political affiliation**. Republicans report much lower LGBT belonging and connectedness than Democrats or independents, all else equal. In terms of generic belonging and wellbeing, the only statistically detectable difference is that LGBQ Democrats report

higher overall wellbeing than LGBQ Republicans. Less than 5% of respondents are Republicans, which makes the clear observed differences especially notable.

- Finally, **education** presents associations in two opposing directions. Respondents with higher levels of education report *less* LGBT belonging and connectedness but *more* generic belonging and social wellbeing, all else equal.

3.4 Discussion

Taken holistically, these results are consistent with expectations. The findings for density are weaker and more ambiguous than I might have expected, while the results for same-sex couple households come through strongly across the board – not only for LGBT community connectedness, but for general social well being as well. This speaks to the fundamental importance of LGBTQ abundance for LGBTQ people.

The evidence for population density being positively associated with sense of community for LGBQ people is moderate and mixed, and sensitive to the chosen functional specification. The best statistical models at the zip code level show either *no association* or a *monotonically increasing* association between density and community, not any more complex or inverse relationship. The lack of association for the single-item responses may be substantive – the scale differences are driven by other items on the scales – or simply due to the fact that the ordinal models have less information from which to estimate an association. Regardless, there's no evidence that small towns are good for queer community or wellbeing. This finding is consistent with a "metronormative" narrative of the historical development of LGBTQ identities, subcultures, and communities. At least for sexual minorities, cities are more of a site for Durkheimian collective coming-together than Simmelian individualism.

The prevalence of same-sex couples in a zip code has a clear, positive association with

sense of community for LGBQ people. Places with truly high proportions of same-sex couples are rare, but the models all predict that in these places respondents will report a heightened experience of belonging. By extension, I would interpret this to mean that an abundance and concentration of LGBTQ people in general would facilitate both identity-specific community connectedness and overall social wellbeing. This might seem obvious, but affective attachment to LGBT community might have been independent of or even inversely related to the actual presence of other LGBTQ people; it could have been more of a symbolic than a material phenomenon. Prior research on LGBTQ people in small cities (Brown-Saracino 2017; Forstie 2020), or on solidarity with disdain in large metropolitan areas (Winer 2020), might have predicted a different result than what I actually found. There is no sign of symbolic substitution; instead, material contextual conditions of LGBTQ abundance matter.

The example places I introduced earlier can be viewed in light of the model predictions based on their characteristics. The set of models would consistently predict the greatest sense of community in places like the Castro in San Francisco, with a high abundance of LGBTQ people in a densely concentrated space overall. Those places would be followed by places like Palm Springs, with a high prevalence of LGBTQ people in a space that isn't so dense in general. Ordinary and below-average places on both characteristics would be expected to rank below that. The importance of these durable place characteristics suggests that declaring queer spaces to be passé or viewing them as increasingly ephemeral (Stillwagon and Ghaziani 2019) may be premature. In the context of those countercurrents, such seemingly unsurprising findings might, in fact, be somewhat surprising.

Metropolitan-area characteristics appear to matter less. Inter-regional variation can't be entirely dismissed – higher MSA-level prevalence of same-sex couples *is* associ-

ated with higher general social wellbeing, after all. But there's no sign that either metropolitan-level characteristic is associated with a sense of connection to LGBTQ community. One interpretation of this is substantive; the scales that matter for queer experiences of community are not counties or MSAs but zip codes and neighborhoods. A second interpretation would interrogate the implications for measurement. Twin, interrelated features make estimation at larger spatial scales difficult. There are inherently fewer units, which reduces statistical power; and, because the units are larger aggregates, there is inherently less variability. This means any conclusions drawn from an absence of evidence at the MSA level must be taken with caution. Either way, these findings pose a challenge for future spatial work on small social groups like LGBTQ people. Even zip codes are still quite large, but it's hard to measure anything precisely at smaller geographies. And trends run in the opposite direction: the Census Bureau plans to withhold same-sex couple data below the county level in the public data released from the 2020 Census, rendering even the research I've done here impossible without access to a federal Research Data Center.

The individual-level findings have interpretations ranging from the puzzling to the expected, with some suggesting potential spatial connections for future work. In terms of gender and sexual minority identities, the most surprising findings are the lack of difference between LGBQ women and men, and between gay or lesbian respondents and other sexual minority identities. But the differences that do emerge align with intuition and prior research. Namely, non-binary respondents may find queer communities to be more gender inclusive than society at large (even if queer spaces are themselves often structured by gender). By contrast, bisexual respondents may experience biphobia and bi erasure in queer communities, while also experiencing marginalization in a heteronormative society. This is, regrettably, consistent with more negative outcomes

for bi individuals observed across other domains (e.g., Mize 2016).

The differences by both race and education present a paradox. The lower outcomes for Black and Latinx respondents compared to white respondents in terms of generic belonging and social wellbeing are consistent with an impact of structural racism and oppression; the similar trends for respondents with lower versus higher levels of education are consistent with marginalization by social class. However, in both cases, the opposite is true for the LGBT-specific measures of community connectedness. Given the realities of sexual racism and other forms of exclusion in queer communities (Held 2017; Orne 2017; Stacey and Forbes 2021), this is surprising. Altogether, there is no indication that those at the "imagined center" of the "imagined gay community" (Winer 2020) – educated white gay men – feel a greater sense of LGBT community connectedness in line with their relative privilege.

Finally, the fact that the handful of Republican respondents feel markedly less a part of the LGBT community is a strong indicator that even an identity-based community that is in many ways heterogeneous and inclusive has clear boundaries around moral values (Vaisey 2007). Of course, part of the motivation for including race, education, and political affiliation in models focused on place characteristics is that these individual traits are deeply spatially structured in their distributions, by segregation and social sorting. While the Generations study sample size is too small to uncover meaningful trends with any precision, in principle it would be generative to explore the impact of these three factors measured at the place level as well.

Returning to theories grounded in queer qualitative research, what can my results say about ambient community in a place like Ithaca, New York? Ithaca is its own micropolitan area and coincides mostly with zip code 14850, with around 2% same-sex couple households – high, but not extremely high. One possible response to Brown-

Saracino's argument in *How Places Make Us* might be that she's right – different places *are* different, with consequences for identity and community. But some of those differences are structural and can be quantified, rather than being ineffable qualitative differences in culture and identity. Those just mostly play out at finer spatial scales than cities or metropolitan areas. At broader levels, those differences wash out and become invisible.

In that vein, an especially interesting null finding is that there isn't huge variation across MSAs in a multilevel model. At the outset, I had thought I might see meaningful differences in sense of community between places if I used a better operationalization of differences in place-based cultures. For instance, future work might explore using computational text methods like topic modeling to measure local place-based cultures and narratives (Mohr et al. 2020), but my findings here suggest that city or metropolitan-level measurement might not be sufficiently granular in many cases.

Another unresolved conceptual issue I leave unaddressed here: place-based factors that are difficult to disentangle from density and LGBTQ abundance. In particular, one way future work might explore both the conditions for a strong sense of community and the form that community might take would be to measure heterogeneity and diversity within a place. Any measure of heterogeneity, however, is highly correlated with the two factors I centered in this study. Moreover, diversity is conceptually difficult to disentangle from minority abundance, as previous work by Abascal, Xu, and Baldassarri (2021) has shown. To address those challenges with any precision, a larger or more spatially clustered data set than the Generations study would be necessary.

This chapter has examined the consistencies and discrepancies between multiple measures of community and belonging, at multiple geographic scales, for a unique marginalized identity group. The representative sample of the Generations survey offers system-

aticity and breadth, even though my models leave considerable variation unexplained. These models draw attention to consistent patterns, which means they don't afford the same opportunity for nuance found in the qualitative work I've engaged with to frame my work. Future work in this vein might use these data to explore how responses related to community, identity, and other experiences fit together into overarching frames and schemas of beliefs. However, I've reached the limit of what the Generations data set can tell me about the strength, meaning, and expression of community, and so I turn to different, less conventional sources in the following chapters to explore variations in what community means and how it is expressed.

3.5 Appendix: Statistical tables

TODO: format two tables with four statistical models for zip codes and four models for MSAs. The zip code table is done but only displays in the html version.

4 Community talk

Interactional contexts and consequences

"Community" is something that people do – they build community, they create it. This chapter aims to study how the process of creating and invoking community plays out in social networks and social interactions. Social density of interactions and relations, not mere physical proximity, produces the social reality and the individual subjective experience of community, as well as observable expressions of the same. But then in turn expressions of community help create the conditions for group-based social life. In this chapter, I use a set of virtual communities, LGBTQ-centered groups on the social platform Reddit, as a site to study the relation between social density and expressions of community.

In the previous two chapters, I first explored the conditions under which LGBQ people experience a sense of community in relation to their identity group, finding that they do so in dense, abundant contexts. I then showed that when people talk about community, they often mean something analogous to the sociological concept of Gemeinschaft, of a feeling of "we"-ness and belonging, and that this is especially true in LGBTQ virtual communities, where that sense of Gemeinschaft again intersects with identity. All of this provides evidence of the reality and salience of this sociological thing

called "community" for LGBTQ people. Here, I examine what happens when LGBTQ people seek community out, and express community in the process, by using LGBTQ virtual communities to ask who produces community talk. This combines the tools and insights from the previous two chapters, pairing text analysis methods with an attention to structural context.

As in the chapter on place characteristics, a primary question is whether pro-community factors reinforce each other, or whether community becomes ambient and backgrounded once it has been created. In this chapter, the question is whether community talk is most prevalent among those who are most embedded or central in a group, perhaps performatively creating community for downstream and peripheral group members; or, is it the peripheral members of a group that engage in the most community talk in an effort to create their own sense of belonging? Those are two distinct possibilities for the sources and the potential impacts of expressions of community. Ultimately, the evidence I find is mixed, with some group contexts eliciting community talk most often from the most central group members.

4.1 Background

I focus on linguistic or textual expressions of community, the kind of community talk or vocabulary of belonging that I explored in the previous chapter. This could be explicit invocations of community and belonging that keyword methods can surface, or broader language that more generally resonates with those concepts, of the kind that embeddings models can uncover (Stoltz and Taylor 2021). Text is key partly for measurement and analysis reasons; while image, voice, and video data are all also important aspects of the online social experiences that can create or sustain communities, the computational toolkit for using those types of data are less robust and more challenging to apply. This

methodological constraint shapes my ultimate choice of Reddit as a platform and data source (as opposed to, e.g., Discord).

In addition to the methodological motivation, there are also important theoretical reasons to focus on text and language, especially in the context of virtual communities that produce informal written language (McCulloch 2019). There are two ways to think about text: as something that matters in itself, or as a proxy for something else. What comes out of people's mouths (or keyboards) isn't the same thing as what's in their heads. While one view might be that community-oriented language is only interesting as a proxy for the sense of community that individuals feel and perceive internally, in this case I'd argue for the importance of language on its own. This is because I think community talk can be performative, not merely expressive; what's in people's heads can't do anything in the social world unless expressed in some way, and those expressions might have consequences regardless of what individuals really feel.

A couple analytic considerations shape the research questions I pose. First, do I focus on the causes or consequences of community talk? The structural/interactional formation of a community and the expressions of community that emerge from and in turn reinforce that social reality are a reciprocal, self-reinforcing process that unfolds over time. Because of that, disentangling cause and consequence might be difficult. For simplicity, I treat community talk as the outcome of structural and interactional features of a group. This is largely because the most straightforward treatment of the network involves taking it as static rather than dynamic.

Second, there is the matter of scale. There are two temporal scales or levels to consider when thinking about social density. Social density could be structural and relational, adhering in durable, culturally-recognized ties, like a friendship. Or it could be interactional, ephemeral, and activity-based, like a conversation. In a sense, of course, the

latter coalesce into and constitute the former, or the former are an emergent, culturally-perceived (White [1965] 2008) property of the latter. The key analytic question is what each time scale lets me observe; it is easier to observe the outcome of a conversation than to track the evolution of a friendship. I elect to focus on networks constructed out of conversation-based interactions, which aligns well with using textual measures of community constructed out of those same conversations. Of course, these interactions often unfold in the context of longstanding ties and durable groups – in the case of the virtual communities I analyze, groups lasting for many years.

With those analytic issues in mind, I can sketch out a few possible scenarios from which I will derive more abstract research questions. For instance, on the interactional level of a conversation, how might participants bring in or invoke community, whether implicitly or explicitly? People might be chatting back and forth in an amiable way that steadily builds positive emotional energy (Collins 2004), and that leads to more effusive or expressive community oriented language. Or, community might be invoked in contentious situations (perhaps by a third party), in order to manage contention and steer people toward interactions with more positive emotional energy. In either case, what precedes those moments of heightened community-building? What's the outcome? Backing out further, to the relational level of a group, who are the people who consistently use high amounts of community-oriented language? Are they the most embedded? Accordingly, is combining network and discursive measures a way of identifying people who structurally *and* culturally play a key role in a group (Goldberg et al. 2016)? Or, are individuals on the margins the most effusive, to compensate for a lack of structural belonging and performatively create their own sense that they belong?

Virtual communities provide a key site for observing how community emerge in inter-

action. Virtual communities are real sites for building community (Baym 1994; Driskell and Lyon 2002; Hampton and Wellman 2003; Rheingold 2000), often intertwined with individuals' offline lives as a kind of "augmented reality" (Jurgenson 2011; Orne 2017) rather than being completely distinct and separate (i.e., "digital dualism"). Moreover, the interactions that constitute virtual community building often happen *through* text (McCulloch 2019). By contrast, it's difficult to collect detailed social density data from offline interactions or ties alone (But not impossible, see Boessen et al. 2014). Another noteworthy aspect of digital spaces is that LGBTQ virtual communities are and have long been highly visible; queer people have been using digital technology to to form connections, to "find" or "build" community with each other, since the virtual communities of the 1990s, like Usenet or the WELL (boyd 2014; Dame-Griff 2019; Rheingold 2000).

Different digital platforms have distinct structures and affordances, which shape the ways people stage social interactions and the kinds of communities they build. The one that provides a logical option for studying virtual communities is Reddit, because the entire platform is structured around public groups (called *subreddits*). In other words, the key affordance of Reddit is the existence of groups in which to participate; these groups are the focal points for almost all interaction, which occurs in threads of posts and comments within subreddits. Posts and comments can be rated and ranked through upvotes and downvotes, providing a crowdsourced measure of quality (Medvedev, Lambiotte, and Delvenne 2018). Direct structural ties between people are deemphasized; analyses of networks of Reddit users instead focus on the web of group affiliations (Olson and Neal 2015; Simmel 1971; Waller and Anderson 2019, 2021) created by ties of subreddit co-membership. In one sense, this group-based structure means that Reddit takes "groupness" for granted – i.e., it can appear to presume the existence of real

community rather than showing how strong or cohesive a given community really is. Everyone who participates to any degree could be said to be a "member" of a "community", which potentially drains those words of any deeper meaning.

However, subreddits vary immensely, in size and activity level, but also in moderation efforts, adherence to local rules and norms, and other signals of distinctive group styles and subcultures. Users vary as well, in their levels of participation, how specialized and selective they are in engagement across different groups, and how embedded they are in within-group conversations. Existing research leverages some of those variations. For instance, Zhang et al. (2017) characterize user engagement across a typology of community-level linguistic features; Lucy and Bamman (2021) use contextual word embeddings to study linguistic variation and conformity and identify semantically unique communities; and Waller and Anderson (2021) study polarization of subreddits through "community embeddings" based on co-membership. In the LGBTQ context, Reddit affords opportunities to observe community-oriented language across both large and generic (e.g. r/lgbt) and small and niche (e.g. r/LesbianGamers/) groups. Here, I ask how the embeddedness of an individual within a subreddit's interaction network relates to explicit and implicit instances of community talk.

4.2 Data and methods

I examine the relation between interaction networks and comment text across 11 LGBTQ-centered online groups ("subreddits") on Reddit. These subreddits come from an internal taxonomy of subreddits (r/ListOfSubreddits/), referenced in Lucy and Bamman (2021). They are likely to be the among the largest, most well-known, and most general LGBTQ-themed subreddits, although there are many more. There, the groups are categorized specifically under "Communities" alongside other recognizable

groups (e.g., parents, teachers, vegans, people with beards), and as opposed categories like "Discussion" and "Entertainment," suggesting that community-building will be an intentional focus in a way that is not necessarily true for all online groups. Notably, Reddit (the company and platform) itself also recognizes LGBTQ-themed groups as a paradigmatic case for on-platform community-building – their 2020 comments to the FCC in defense of volunteer moderation under Section 230 specifically choose to highlight and give voice to community moderators from r/lgbt.

I construct measures of community talk from the text of the Reddit comments. Due to their relational nature, word embeddings are well-suited for uncovering implicit references in a text. But to move from individual words to longer texts – sentences or paragraphs or entire documents – some kind of aggregation method is necessary. Multiple such methods for summing or averaging word embeddings exists; one normalized or weighted method with advocates in sociology is called Word Mover's Distance (WMD) (Kusner et al. 2015) or Concept Mover's Distance (CMD) (Stoltz and Taylor 2019). CMD calculates a distance metric between any given text and a target vector – in this case, the vector for "community". In disciplines other than sociology, simple averages of embeddings are used instead (Kennedy et al. 2021), but my brief investigations show little difference from WMD. Because I use the Python implementation of Word Mover's Distance in the gensim package, there are minor differences in the distance metric and algorithm from Stoltz and Taylor's implementation of CMD; I expect the results reported below to be robust to such minor variations. I also do not standardize or invert the values I report.

Rather than construct a latent continuous measure of distance from the concept of "community," represented through its GloVe word embedding (Pennington et al. 2014), for each comment, a simpler approach might measure the comments that explicit

reference "community" (or "communities"). To contextualize this analytic choice, there is a methodological debate within the literature on computation text analysis for social science, about whether to measure surface-level, explicit keywords or latent, implicit meanings, and about the tradeoffs between each family of approaches (Stoltz and Taylor 2021). In the context of morality, Kennedy et al. (2021) use averages of word embeddings instead of WMD, but also provide evidence in favor of embeddings approaches over explicit lexicon-based approaches. For many instances where measuring culture through the meanings of text (Mohr et al. 2020) is the desired aim, I tend to agree that the latter is more appropriate. I described and used such an approach with word-level embeddings in the previous chapter. However, as I will show, issues arise when moving from single-word embeddings to overarching document-level measures. Because of this, I choose to compare both explicit and implicit measurement approaches here.

I obtain subreddit data through ConvoKit (Chang et al. 2020), a Python-based toolkit for retrieving and analyzing conversation-based data sets which archives subreddits through 2018. While I do not use many of the conversation-specific features of the toolkit, the format is ideal for extracting interaction and their metadata, making it easy to retrieve both textual data and the interaction network structure. (A limitation of this choice is that it would be hard to automatically generalize my data processing code to Reddit data stored in a more typical database format.) I calculate Word Mover's Distance using gensim (Řehůřek and Sojka 2010). I transform interaction pairs into a network and calculate network statistics using igraph (Csardi and Nepusz 2006), dropping deleted users and the AutoModerator (a bot). Following Lucy and Bamman (2021), I use closeness centrality as my key measure of user embeddedness in the subreddit conversation networks. (By contrast, Foote, Shaw, and Hill (2023) look at inequality in *betweenness* centrality. These two measures, plus eigenvector centrality,

do not give consistent results.) Unlike Lucy and Bamman (2021), I do not restrict the user networks to the 20% of users who are most active; instead, I subset to users within the largest connected component of each of the networks. (This drops the number of users in each subreddit considerably, but users outside the largest component are almost all singletons with minimal activity in the group.)

One key measurement issue makes it difficult to use Word Mover's Distance from "community" as the sole outcome measure. When calculated at the comment level, Word Mover's Distance exhibits a systematic variation in a way that appears to make it ill-suited for statistically modeling with other variables. Specifically, short texts, especially one- or two-word texts, exhibit very high variability; longer texts converge to roughly the average value overall. This makes theoretical sense when WMD is used to compare a variable-length document to a one-word concept, as Stoltz and Taylor (2019) do with their Concept Mover's Distance adaptation of the measure: WMD measures the effort it would take to transform one text into another, normalized for document length; in this view, single words can be quite far apart. Longer texts are a sort of average of all the words they can contain; the longer the document, the closer it becomes to what might be an average distance from, perhaps, the corpus overall. Variation washes out. I suspect this issue has not been previously reported because prior work has compared texts of roughly similar sizes, e.g., sentences, speeches, or books (Stoltz and Taylor 2019), and has compared relatively few documents. Reddit comments are highly variable in length – many are only a few tokens long, while some stretch to thousands of words, and there are millions to compare. This makes the problem quite visible, as shown in Figure 4.1. A simple correction might be to use weighted least squares and weight longer comments more highly; more advanced approaches might incorporate and estimate that heteroskedasticity within the model itself. Still, this ignores something fundamental

about the measure itself, with no obvious solution.

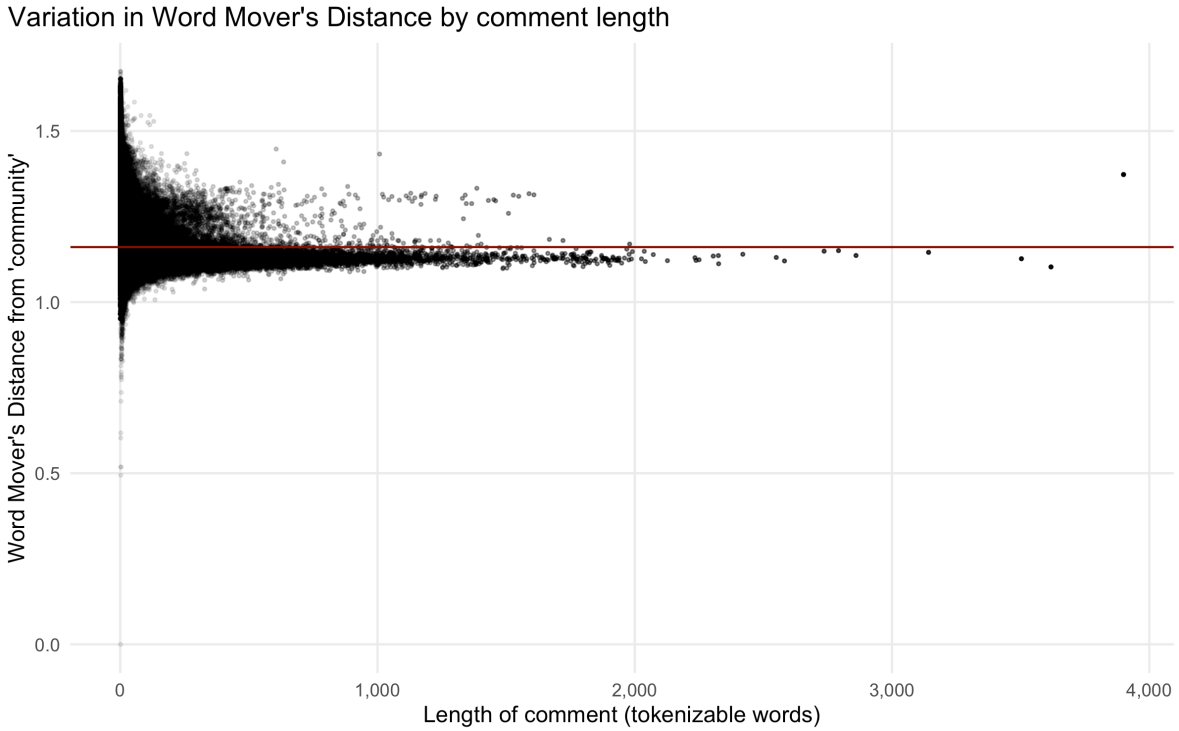

Figure 4.1: Variation in Word Mover's Distance from 'community' converges to a value below the average with increasing comment length (comments from r/gaybros subreddit)

Similarly, when I average WMD values across a user's comments, the same phenomenon occurs; users with more comments have converging average WMD values. In this case, I use weighted least squares weighted by the number of comments per user. (In modeling the explicit use of "community," I control for the number of comments instead; I explore controling for average comment length.) Moreover, the network measures are calculated at the user level, meaning that differences (especially for explicit mentions of community) are diminished at the comment level. Even for users who talk about community frequently, most of their comments will not include the word.

With all of these considerations in mind, I present user-level models below. In addition to using weighted least squares to model WMD by user closeness centrality and subreddit, I use logistic regression to model the probability that a user ever writes "community" in any comment and negative binomial regression to model the number of times a user writes comments with the word "community." (Given that the majority of users do not write about community explicitly, a zero-inflated model would be theoretically justified, but model fit metrics do not support this additional complexity.) While initially I fit separate models for each subreddit, in the final results I present a pooled model with centrality by subreddit interaction terms, which are generally warranted by measures of model fit.

4.3 Results

Table 4.1 reports descriptive statistics for the largest connected component of each subreddit's network, and Table 4.2 reports descriptive statistics calculated on the text of the comments on each network. A large minority of users, typically those who post only once, are dropped when the ConvoKit data are subset to the connected users, but the vast majority of comments are retained. The few users with no tokenizable text are dropped from the second table and the models. Subreddits vary in activity levels, from 140,000 comments (r/gay_irl) to 2.7 million (r/askgaybros). From 6% to 24% of users mention "community" at least once, but only 1% - 3.5% of comments do.

Table 4.1: Subreddit network statistics for largest connected component

Subreddit	N. nodes	N. edges	Density	Mean closeness centrality
r/lgbt	81,535	638,067	0.0001	0.22

Subreddit	N. nodes	N. edges	Density	Mean closeness centrality
r/gaybros	55,784	1,088,938	0.0003	0.28
r/actuallesbians	52,077	952,966	0.0004	0.28
r/gaymers	31,980	462,184	0.0005	0.27
r/bisexual	34,328	299,363	0.0003	0.24
r/askgaybros	63,404	1,573,925	0.0004	0.30
r/ainbow	25,330	297,793	0.0005	0.28
r/gay	23,823	182,482	0.0003	0.24
r/gay_irl	12,243	107,815	0.0007	0.26
r/asktransgender	58,175	1,384,561	0.0004	0.30
r/transgender	11,151	102,337	0.0008	0.26

Table 4.2: Subreddit text statistics

Subreddit	N. users	N. comments	Median comments per user	Mean comments per user	Pct. 'community', users	Pct. 'community', comments	Mean user WMD
r/lgbt	81,436	1,047,003	4	12.9	0.19	0.035	1.15
r/gaybros	55,727	1,671,630	6	30.0	0.16	0.015	1.16
r/actuallesbians	52,037	1,492,782	7	28.7	0.18	0.015	1.16
r/gaymers	31,938	703,158	6	22.0	0.14	0.014	1.17
r/bisexual	34,286	467,598	5	13.6	0.18	0.030	1.15
r/askgaybros	63,366	2,777,469	7	43.8	0.17	0.013	1.15
r/ainbow	25,309	530,889	4	21.0	0.24	0.037	1.15
r/gay	23,792	305,432	5	12.8	0.15	0.023	1.15
r/gay_irl	12,189	140,147	3	11.5	0.06	0.012	1.18

Subreddit	N. users	N. comments	Median comments per user	Mean comments per user	Pct. 'community', users	Pct. 'community', comments	Mean user WMD
r/asktransgender	58,162	2,430,525	7	41.8	0.23	0.019	1.14
r/transgender	111,144	173,941	4	15.6	0.21	0.034	1.15

I first present model results for one large subreddit as an example, and then show the full range. While r/lgbt has more users and is the most general, r/gaybros has more comments and more edges, and thus more overall activity. Only the r/ask* subreddits (r/askgaybros and r/asktransgender) are more active, but these are intrinsically oriented toward question-answering and discussion rather than community-building per se, so it makes sense to treat them as distinct rather than representative. In any case, r/lgbt and r/gaybros show generally similar trends – which is not true for all of the remaining nine. In another signal of the robustness of my findings, the two measures of community talk, explicit and implicit, are often consistent. As with the other trends, this does not necessarily generalize to the full set of subreddits.

The models control for the number of comments per user, logged. In the context of the negative binomial model, this means that the outcome can be considered approximately as rates rather than counts. (Fixing the coefficient for log(number of comments) at 1 as an offset would be a true rate, but estimating the coefficient results in a better model fit.) This control is essential because a user's closeness centrality and number of comments are highly correlated. As Figure 4.2 shows, there is still sufficient conditional variation to model. (However, bear in mind the constrained conditional range when viewing the model prediction figures below, which means that some predictions are beyond the scale observed in the actual data.) Other controls I considered, average comment length and membership duration, did not affect estimates of the closeness centrality coefficient, and

so I exclude them here.

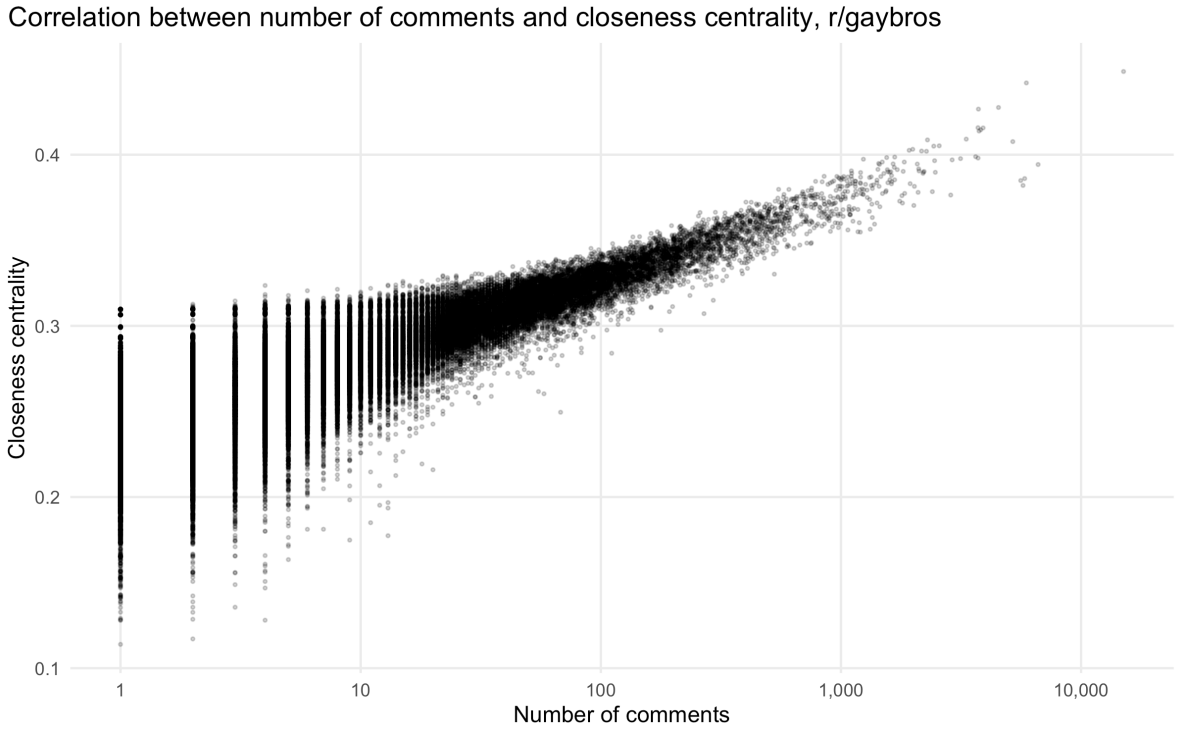

Figure 4.2: User closeness centrality by number of comments, r/gaybros

As is typical for online communities, the distribution of user activity is highly skewed, with a handful of users providing the bulk of interactions and many contributing only once. Accordingly, I illustrate predicted values by closeness centrality with the number of comments held at the median (6), mean (30), and 95th percentile (113) values for members of r/gaybros. Figure 4.3 shows the results for the three models: one of implicit Word Mover's Distance from "community," and two predicting whether a user explicitly mentions community. At n = 30 comments, an increase in closeness centrality from 0.2 to 0.3 increases a user's probability of ever mentioning "community" from 25% to 31%. At n = 113, a user is predicted to mention community 1.17 times at a closeness centrality of 0.2, and 1.52 times at a closeness centrality of 0.3. The implicit community results

are consistent; the average WMD of a user's comments from "community" declines as the user's centrality increases, although the meaning of the decrease is not as concrete to interpret.

Modeled at the comment level, results are much less interpretable. There is no association with closeness centrality in a logistic regression of explicit community mentions – which makes sense, given that users who mention community will also have comments where they do not. There appears to be a small *positive* association between closeness centrality and Word Mover's Distance from community (β = 0.002), in a linear model weighted by comment length; this is an order of magnitude smaller than the user-level coefficient (β = -0.029), and so it is substantively unimportant. That the result is not consistent with the user-averaged finding suggests that weighting and aggregating may not have overcome the mathematical issues inherent in the funnel-shaped distribution of WMD values. Because these results are difficult to interpret, and because there are millions of comments across all of these subreddits, I do not pursue a pooled model at the comment level.

I next present results from combined models of all 11 subreddits, allowing coefficients and levels to vary by subreddit with interaction terms. Here I show predicted values by subreddit and closeness centrality with the number of comments fixed at 30, the mean for r/gaybros (which is, again, one of the most active subreddits). While users may actually overlap between subreddits, centralities and text measures are calculated on a per-group basis, meaning that observations are actually user by subreddit. Figure 4.4 shows predicted probabilities from a logistic regression of whether a user ever mentions "community" in a comment. Figure 4.5 shows predicted counts of the number of comments mentioning community that a user makes (again, effectively rates, given that a user's total number of comments is controlled for). Finally, Figure 4.6 shows predicted

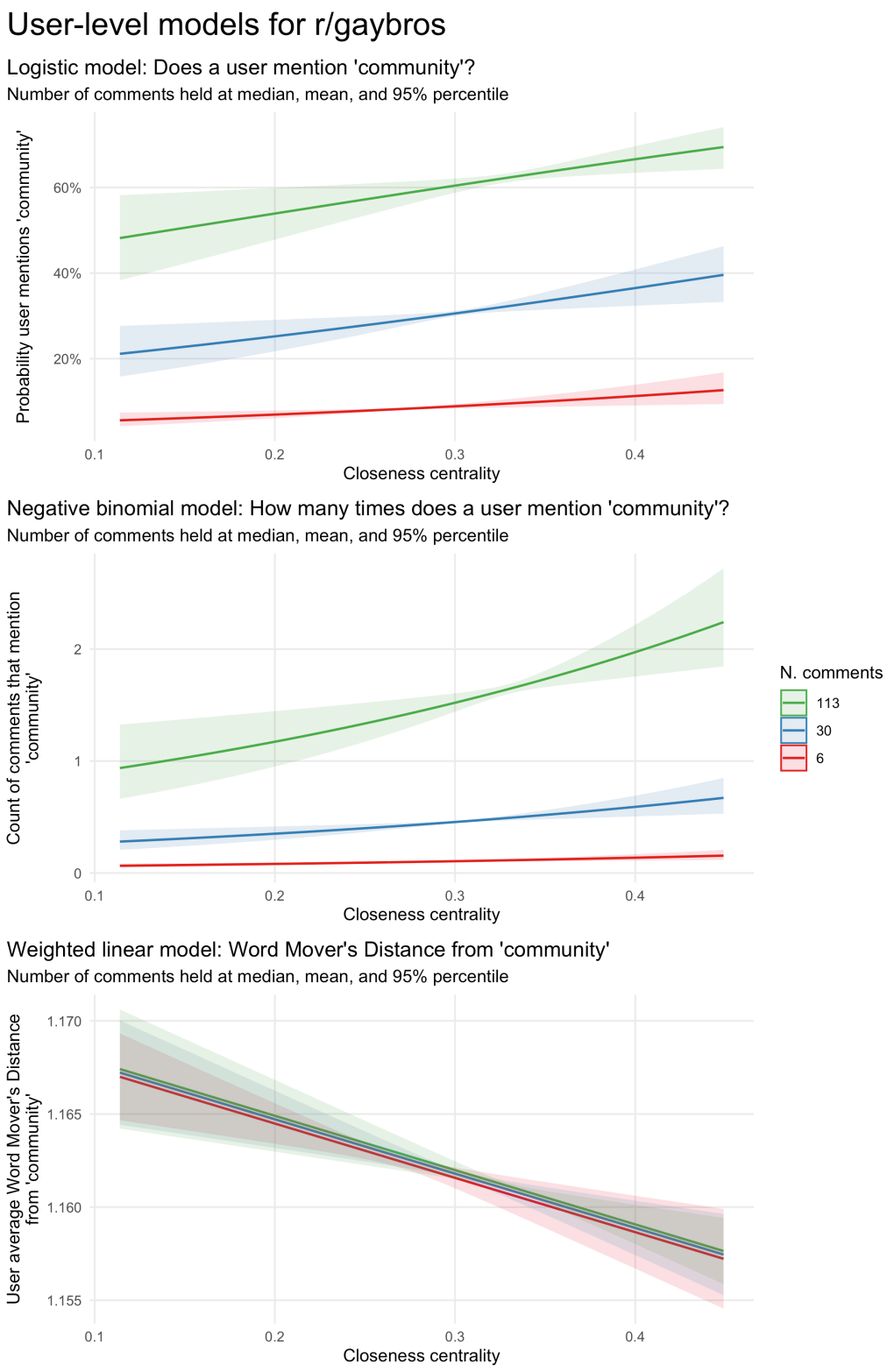

Figure 4.3: Modeling textual measures of community by member centrality in one subreddit

WMD from "community" in a linear regression weighted by number of comments.

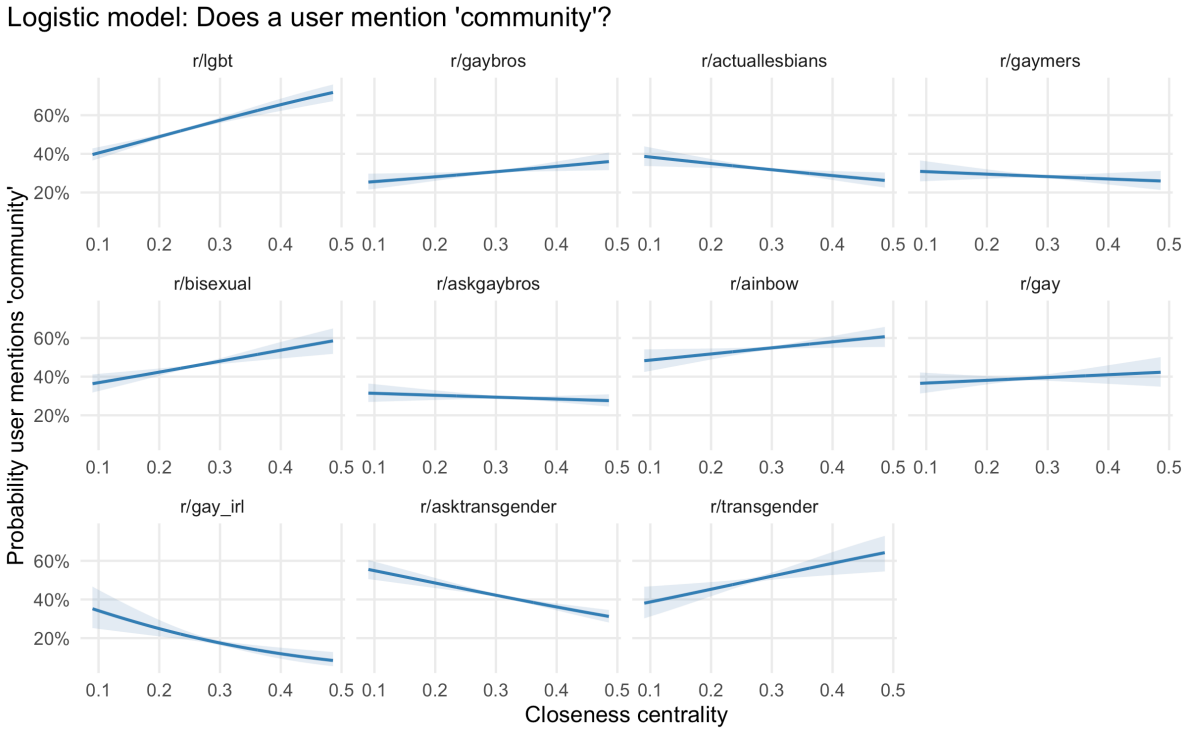

Figure 4.4: Predicting whether a member ever mentions 'community' across 11 subreddits. Number of comments held constant at 30.

Rather than consistency, these results show considerable variation in levels and trends. r/lgbt aligns in direction with the previously described results for r/gaybros, although it tends to have higher predicted values of community overall (consistent with the much higher proportion of comments explicitly mentioning community, 3.5% compared to 1.5%). r/ainbow, a split from r/lgbt with a looser moderation policy, does not show consistent trends, neither across the three outcomes nor with the other subreddits. The r/gaymers group aligns with r/lgbt and r/gaybros on the implicit community measure, but shows little association on the explicit measures. There is no particular pattern among the groups for specific identities under the LGBT umbrella (r/actuallesbians,

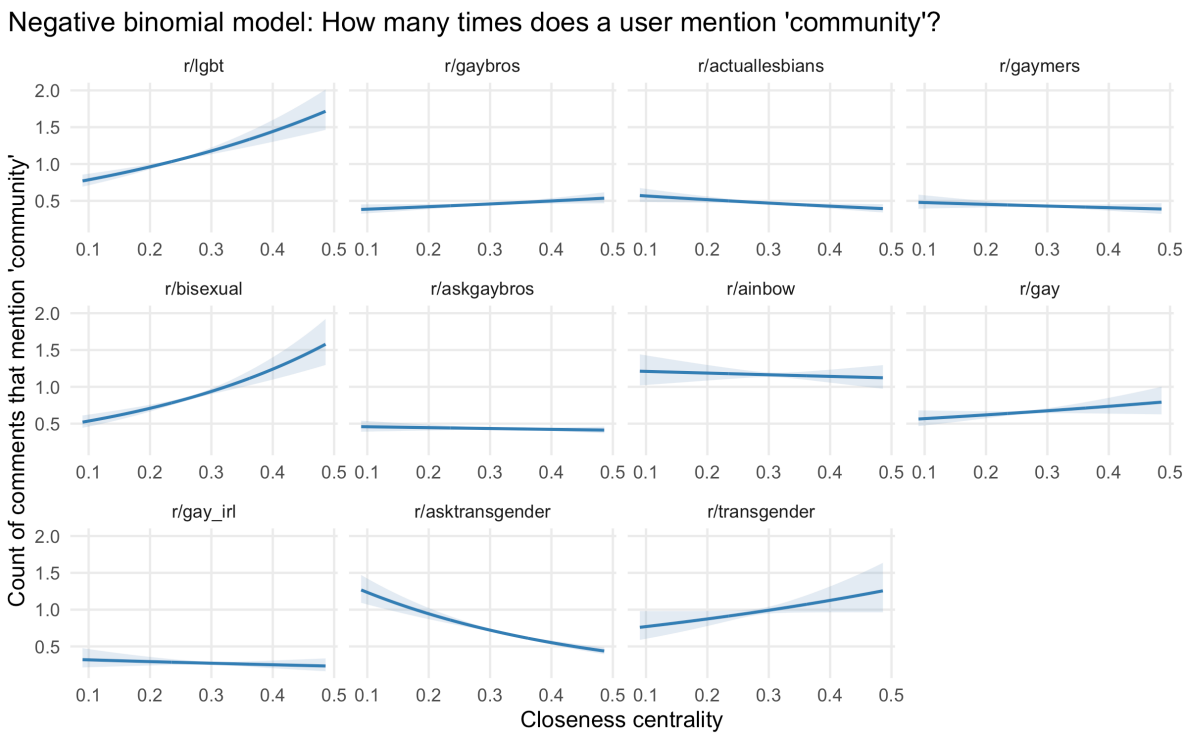

Figure 4.5: Predicting how many times a member mentions 'community' across 11 subreddits. Number of comments held constant at 30.

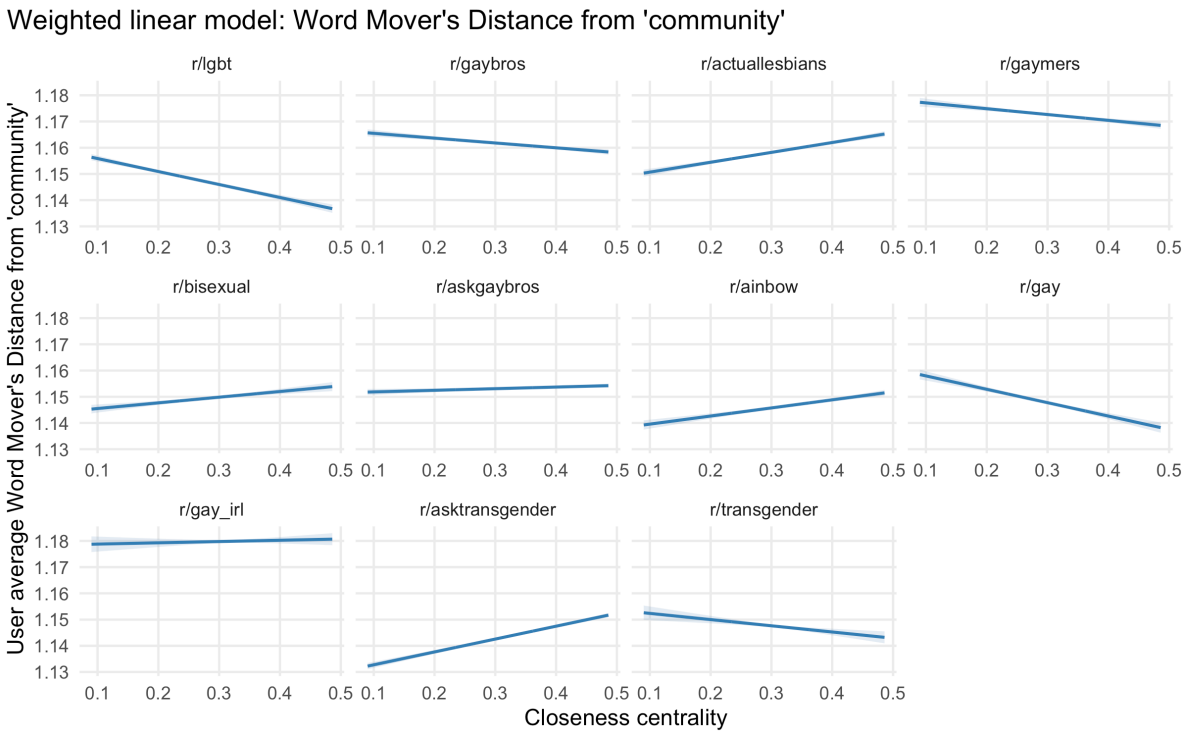

Figure 4.6: Predicting a member's average Word Mover's Distance from 'commmunity' across 11 subreddits. Number of comments held constant at 30.

r/bisexual, r/gay, r/transgender); all except r/actuallesbians show increasing probabilities and rates of explicit community talk with increasing closeness centrality, but implicit community WMD values go in both directions. The two question-and-discussion groups, r/askgaybros and r/asktransgender, show inverted or no associations between community talk and closeness centrality. Finally, the subreddit with the fewest mentions of community and highest WMD, as well as the fewest tokenizable comments overall, is r/gay_irl, which shows a declining pattern of explicit mentions of community with closeness centrality. This subreddit is for memes, which may provoke different types of discussion and which also heavily incorporate images, a medium not accounted for by the text-centric methods I use.

4.4 Discussion

What seemed like a potentially coherent empirical signal in the context of a single subreddit loses narrative coherence across the other ten. What might it mean that the signals are so mixed? The fact that I observe so much variation – even among groups for similar purposes, even among similar identities – leads me to expect to see even more variation if I were to widen my lens to other popular and active groups on Reddit.

One interpretation is that the structure of interactions in online groups does not have much bearing on the production of community-oriented talk. This negative finding might be surprising, but it would align with other recent work on the network structure of virtual communities. First, Lucy and Bamman (2021) find, also using subreddits as data, that use of distinctive in-group language – which they take to indicate belonging – is not predicted by user centrality. Second, Foote et al. (2023) find that multiple network measures do not predict successful outcomes in peer production communities; because those groups are organized around work and information, rather than community-building per

se, it remains surprising evidence for generalization that I observe such heterogeneous and unclear results in a more purely social context.

Another way to view these results is to cast doubt on how I've operationalized "community" from text-based signals. Text analysis involves an overwhelming number of analytic choices, and researcher degrees of freedom are known to be a serious problem for reproducibility and validity (Lucas et al. 2015; Nelson 2019; Wilkerson and Casas 2017). In particular, I have shown reasons for concern when moving from word-level analyses using word embeddings to aggregating to document-level analyses using methods like Word Mover's Distance, and I believe the mathematical properties of these aggregate measures need to be better understood before they are more widely adopted. This methodological work might look similar to prior work assessing the validity of word embeddings at the micro scale (e.g., Antoniak and Mimno 2018; van Loon et al. 2022). To mitigate those concerns, I used a very simple keyword-based approach (Stoltz and Taylor 2021), and am most confident in my findings where those two measures align with each other. Still, my results were not terribly robust when I extended them to a wider range of subreddits for evaluation.

Is talking about community the same as creating it? Not necessarily. To complement this cursory investigation, I have a few suggestions that might be taken up in future work. My initial qualitative explorations showed that even this small collection of outwardly similar subreddits varies substantially in their orientations (e.g. in attitudes toward community moderation, types of content that it is normative to share, etc.). Qualitatively examining conversations that appear to be about community (implicitly or explicitly) and the trajectories of users engaged in those conversations might uncover potential dynamics or mechanism at play. Rather than broadening the scope of analyses, more granular and contextualized quantitative analyses might be worth

pursuing. If the micro-, user-level findings are not consistent, it strikes me unlikely that aggregating to macro-, group-level outcomes would reveal anything systematic and meaningful. I skirted over the temporality of the network structure and treated it as static, but interaction networks are dynamic and both individual and group trajectories change over time; incorporating that information might sort out the circumstances under which community talk might be either a cause or a consequence of structural features of groups.

5 Conclusion

5.1 Summary

This dissertation set out to investigate how structural features – social density of interactions, physical density of proximity, abundance of others with shared group characteristics – create the conditions for the existence of a community. It connects those contextual features to how LGBTQ individuals outwardly express and subjectively experience community, through their language, group participation, and self-reports of belonging.

Chapter 2 showed that not only does the discourse in the soc.motss Usenet group, an early virtual community, use "community" in the sense of *Gemeinschaft* to at least an equal extent as generic English-language text, it also deemphasizes the geographic aspects of community and replaces that with connotations specific to LGBTQ identities. This strongly implies that LGBTQ community is foregrounded, not backgrounded, in a self-selected virtual community that is densely organized around that identity category.

Chapter 3 showed that small-area abundance of same-sex couples, and to a lesser extent overall density, is associated with greater sense of community connectedness for LGBQ people. Associations weaken or disappear at a larger spatial scale. The key finding is consistent with the idea that contextual features that make access to

community easier heighten the experience of it rather than leading it to be taken for granted, and not consistent with the idea that places with less access (and presumably more stigma and marginalization) lead people to seek out and accordingly report more of a connection to community. These results imply that community does not generally recede into the background in the places where access to LGBTQ people and institutions would be most abundant, nor does it take on extra salience in more peripheral places.

Chapter 4 showed significant heterogeneity in the relationship between interaction network centrality and expressions of community talk in a set of 11 LGBTQ groups on Reddit, a virtual platform. In the largest and most general subreddits, centrality is associated with explicit mentions of community, and to a lesser extent with the implicit distance measure as well. In other cases the association is reversed or absent. The cases with positive associations support the idea that central actors' embeddedness is reflected in their language as well. This work implies that, if these 11 ostensibly similar groups differ in their patterns, there is no universal relationship between community-oriented language and embeddedness in group interactions. A second, methodological implication is that applying a text analysis tool to a substantive use case with a wide variety of text can reveal shortcomings of the method that might not otherwise be apparent; this weighs into a debate between simpler and more complex approaches to text analysis in computational sociology.

5.2 Contributions

Overall, this work demonstrates how community can be built around a particular marginalized identity. Each chapter investigates the relationship between structural features that facilitate social proximity and how LGBTQ people outwardly express and subjectively experience community and belonging. It could have been the case that

expressions of community stood in for structural elements that promote belonging, that community would be ambient and relegated to the background in those contexts. That would have gone hand-in-hand with a sense of outsiderness pushing LGBTQ individuals to seek out community most saliently in contexts of stigma, marginalization, and scarcity.

Instead, across this work, I find that experiences of community are most intense and expressions of community most frequent, not for peripheral members of LGBTQ communities but for the most central. This core finding might appear relatively conventional, not counterintuitive. But queer communities are an unusual phenomenon, both built around a core social identity, but also something almost no LGBTQ individual is born or raised with intrinsic access to. That lack of rootedness might have translated into LGBTQ community arising mainly from transient sites like ephemeral queer pop-ups (Stillwagon and Ghaziani 2019) or fleeting queer counterpublics (Berlant and Warner 1998), rather than the durable contextual characteristics I observe. Nor do I find signs of rejection or ambivalence toward LGBTQ community in the contexts where access is most abundant, as some qualitative researchers have (Brown-Saracino 2017; Winer 2020). Where community for LGBTQ people is most readily available, it turns out to be highly valued rather than passé. The attractive force of community arises from something more than stigma.

Attaining these substantive insights required methodological innovations, a contribution on their own. Most studies of community, especially of LGBTQ communities, are qualitative (Brown-Saracino 2017; Forstie 2020; e.g., Orne 2017); I chose to complement these detailed studies with the breadth afforded by quantitative methods. Linking structural features to expressive and subjective outcomes required stitching together diverse data sources and triangulating across contexts, which gives my results spatial

representativeness and temporal breadth. Most notably, I adapted natural language process techniques for computational text analysis to delve into the specific meaning of "community" as a complex, ambiguous concept with a rich social life beyond the academy. To lay the groundwork for interpreting my subsequent, more straightforward studies, linking community both to the social organization of *Gemeinschaft* and to LGBTQ identities was a necessary innovation.

5.3 Limitations

This dissertation is not without its limitations. The empirical results, while largely pointing toward community as foregrounded and central rather than backgrounded and ambient, do contain mixed signals. This is most apparent in the varied results from different LGBTQ groups on Reddit in Chapter 4. In addition, the spatial patterns in Chapter 3 are not entirely consistent with qualitative work finding between-place heterogeneity, most notably ambient community in places like Ithaca, New York (Brown-Saracino 2017). This points toward a fundamental limitation of a partial and triangulated approach – I am unable to observe any one context fully. A comprehensive view would instead measure locations, interactions, expressions, and subjective experiences for the same people in the same groups. With that level of detail, heterogeneity between communities, rather than common patterns among them, might come into view.

This work is a broad examination of fundamentally dynamic processes that are hard to model. If structural density and embeddedness leads to subjective and expressive forms of community belonging, and these in turn promote proximity, connection, and interaction, then this feedback loop is difficult to disentangle. Some of this is micro-scale, unfolding at the level of repeated participation in group conversations and activities; future work might investigate the fine-grained temporality that forms and sustains

LGBTQ spaces and groups. Some of this unfolds at the temporal scale of entire lifetimes, as LGBTQ people with the motivation and means self-select into places where they hope to find belonging; following LGBTQ individuals across the life course as they migrate and make other key life choices might show how central LGBTQ community can be for life trajectories.

The entwined nature of the factors influencing community limits the concepts I am able to operationalize. While I operationalize signs of density, abundance, and embeddedness, those factors obscure the role of heterogeneity, diversity, and inclusivity. Empirically, I found these two sets of attributes to be too correlated to disentangle. Theoretically, I continue to believe inclusivity is important as one of the key features of some queer spaces, and absolutely essential for understanding the moral boundaries of LGBTQ community (Meyer and Choi 2020; Vaisey 2007).

While my data span two decades and multiple sites, they do not automatically generalize to all times and places. The survey data in Chapter 3 are only from the contemporary United States. While virtual communities like those I analyze in Chapter 2 and Chapter 4 have the potential for international reach, they are still U.S.-centric (Rheingold 2000). As in much NLP-based work, the textual data I rely on are English-only (Bender 2011); analogues to "community" might not have the same resonances in other languages. Indeed, in a language as closely related to English as French, "community" is hard to translate (Anderson [1983] 2016). Fundamentally, it is strange to study place attachment in the context of a settler-colonial society, or to study community talk in a colonially imposed and globally hegemonic language; I do not adequately address that dissonance in this work.

5.4 Future directions

I close by considering the future of studying LGBTQ communities, and the future of those queer communities themselves. First, new empirical findings might arise from engaging innovative data sources. Digitized archives like the *Mapping the Gay Guides* project (Regan and Gonzaba 2019) increasing offer the opportunity to link historical trajectories of LGBTQ spaces with the present. More recent virtual platforms than Reddit, like Discord (Jiang et al. 2019), offer new sites of queer community building for study. Second, key research questions can be drawn from what this dissertation leaves unanswered. Novel quantitative research methods are need to understand how not just shared identity and discourse, but shared activities and practices (Brint 2001; Orne 2017), contribute to community. Finally, as I alluded to above, I hypothesize that inclusivity plays an important role in queer spaces as a value and a practice. Anarchic and accepting does not mean incohesive; instead, it might be a positive, normative vision of the world, a foundation for collective action rather than a demobilizing force. As they say, Stonewall – the central event of queer collective memory (Armstrong and Crage 2006) – was a riot.

Since I began this project, times have changed for digital and computational research methods. We are at the end of an era of relatively open data access, in the "post-API age" (Freelon 2018). Twitter, once the "model organism" of social media research (Tufekci 2014), has closed off academic access, and even access to the Reddit API has come into question. It is not clear where future data to study virtual social interactions will come from. At the same time, natural language processing methods have advanced significantly in the past few years, well beyond the simple word embeddings I used in this work to more complex large language models (LLMs), such as the series of GPT models and their competitors. LLMs are too data- and computation-intensive to train

locally on niche data sources like queer Usenet groups, but will no doubt find other uses in academic research. In all, there are now greater limits to what an independent graduate student can do on their own. This is a loss because it is hard to imagine large research groups, much less private organizations, paying much attention to queer lives and experiences.

Outside of research, what kind of world makes for strong LGBTQ communities? This dissertation shows that experiences of community are not divorced from material reality. In terms of place-based community, availability of housing, transit, and public spaces would allow people to make the choice to come together. Without attention to queer spaces and queer visibility specifically in the realm of policy and planning, I would be concerned about a widening bifurcation in access among LGBTQ people. For creating a sense of community virtually, "augmented reality" – an overlay of digital and offline lives (Jurgenson 2011; Orne 2017) – seems the most promising way forward. Soc.motss had its meetups, and I suspect the most cohesive subreddits do the same. Again, without deliberate thought, digital spaces too can be hostile or unwelcoming for LGBTQ people. Rather than a world of diverging or shrinking access to collective queer life, LGBTQ community can and should be available to everyone who wants it.

This work was meant to be something of a love letter to queer spaces. They can feel precarious and fragile; that they can exist at all can feel like a small miracle. In the present work, I have shown the conditions under which LGBTQ communities continue to thrive now, but I cannot predict their future. Over the course of the years I have spent on this dissertation, there has been a marked shift from increasing formal legal equality nationwide for LGBTQ people in the United States to growing backlash in the form of state-level legislation – especially, in this moment, backlash against transgender people and against any forms of gendered existence and gender expression that challenge norms.

Living queer lives, in public, is an expression of community. While LGBTQ people may be somewhat ambivalent about the continued salience of distinctive queer spaces, we are not that divided on the issue. On the whole, participating in queer community makes us feel more connected. May it continue to be so.

References

Abascal, Maria, Janet Xu, and Delia Baldassarri. 2021. "People Use Both Heterogeneity and Minority Representation to Evaluate Diversity." *Science Advances* 7(11):eabf2507. doi: 10.1126/sciadv.abf2507.

Anderson, Benedict. [1983] 2016. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.

Antoniak, Maria, and David Mimno. 2018. "Evaluating the Stability of Embedding-based Word Similarities." *Transactions of the Association for Computational Linguistics* 6:107–19. doi: 10.1162/tacl_a_00008.

Armstrong, Elizabeth A., and Suzanna M. Crage. 2006. "Movements and Memory: The Making of the Stonewall Myth." *American Sociological Review* 71(5):724–51.

Arseniev-Koehler, Alina, and Jacob G. Foster. 2022. "Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What It Means to Be Fat." *Sociological Methods & Research* 51(4):1484–1539. doi: 10.1177/00491241221122603.

Auerbach, David. 2014. "The First Gay Space on the Internet." *Slate*, August 20. Retrieved April 21, 2023 (https://slate.com/technology/2014/08/online-gay-culture-and-soc-motss-how-a-usenet-group-anticipated-how-we-use-facebook-and-twitter-today.html).

Baldor, Tyler. 2018. "No Girls Allowed?: Fluctuating Boundaries Between Gay Men and Straight Women in Gay Public Space." *Ethnography* 1466138118758112. doi:

10.1177/1466138118758112.

Bamman, David, Chris Dyer, and Noah A. Smith. 2014. "Distributed Representations of Geographically Situated Language." Pp. 828–34 in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics. doi: 10.3115/v1/P14-2134.

Baym, Nancy K. 1994. "From Practice to Culture on Usenet." *The Sociological Review* 42:29–52. doi: 10.1111/j.1467-954X.1994.tb03408.x.

Bender, Emily M. 2011. "On Achieving and Evaluating Language-Independence in NLP." *Linguistic Issues in Language Technology* 6. doi: 10.33011/lilt.v6i.1239.

Berlant, Lauren, and Michael Warner. 1998. "Sex in Public." *Critical Inquiry* 24(2):547–66.

Bérubé, Allan. 2011. *My Desire for History: Essays in Gay, Community, and Labor History*. edited by J. D'Emilio and E. B. Freedman. University of North Carolina Press.

Bishop, Bill. 2009. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Houghton Mifflin Harcourt.

Black, Dan, Gary Gates, Seth Sanders, and Lowell Taylor. 2000. "Demographics of the Gay and Lesbian Population in the United States: Evidence from Available Systematic Data Sources." *Demography* 37(2):139–54. doi: 10.2307/2648117.

Blau, Peter M. 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. 1st edition. New York: Free Pr.

Boessen, Adam, John R. Hipp, Emily J. Smith, Carter T. Butts, Nicholas N. Nagle, and Zack Almquist. 2014. "Networks, Space, and Residents' Perception of Cohesion." *American Journal of Community Psychology* 53(3-4):447–61. doi: 10.1007/s10464-

014-9639-1.

Bourdieu, Pierre. 1991. *Language and Symbolic Power*. Harvard University Press.

Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. n edition edition. Cambridge, Massachusetts London, England: The MIT Press.

boyd, danah. 2014. *It's Complicated: The Social Lives of Networked Teens*. New Haven: Yale University Press.

boyd, danah. 2015."What World Are We Building?" Presented at the Everett C Parker Lecture, October 20, Washington, DC.

Brandt, Philipp, and Stefan Timmermans. 2021. "Abductive Logic of Inquiry for Quantitative Research in the Digital Age." *Sociological Science* 8:191–210. doi: 10.15195/v8.a10.

Brekhus, Wayne. 2003. *Peacocks, Chameleons, Centaurs: Gay Suburbia and the Grammar of Social Identity*. University of Chicago Press.

Brint, Steven. 2001. "Gemeinschaft Revisited: A Critique and Reconstruction of the Community Concept." *Sociological Theory* 19(1):1–23.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models Are Few-Shot Learners." Retrieved May 27, 2023 (https://arxiv.org/abs/2005.14165).

Brown-Saracino, Japonica. 2017. *How Places Make Us: Novel LBQ Identities in Four*

Small Cities. 1st ed. Chicago ; London: University of Chicago Press.

Bruckman, Amy S. 2022. *Should You Believe Wikipedia?: Online Communities and the Construction of Knowledge*. Cambridge: Cambridge University Press.

Calhoun, Craig. 1998. "Community Without Propinquity Revisited: Communications Technology and the Transformation of the Urban Public Sphere." *Sociological Inquiry* 68(3):373–97. doi: 10.1111/j.1475-682X.1998.tb00474.x.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356(6334):183–86. doi: 10.1126/science.aal4230.

Chang, Jonathan P., Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. "ConvoKit: A Toolkit for the Analysis of Conversations." *Proceedings of SIGDIAL* 4.

Collins, Randall. 2004. *Interaction Ritual Chains*. Princeton, NJ: Princeton University Press.

Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Systems:1695.

D'Emilio, John. 1992. "Capitalism and Gay Identity." in *Making Trouble: Essays on Gay History, Politics, and the University*. New York: Routledge.

Dame-Griff, Avery. 2017. "Archiving Usenet: Adopting an Ethics of Care." Retrieved April 20, 2023 (https://mith.umd.edu/news/archiving-usenet-adopting-ethics-care/).

Dame-Griff, Avery. 2019. "Herding the 'Performing Elephants:' Using Computational Methods to Study Usenet." *Internet Histories* 3(3-4):223–44. doi: 10.1080/24701475.2019.1652456.

Denny, Matthew James, and Arthur Spirling. 2016. *Assessing the Consequences of Text*

Preprocessing Decisions. *SSRN Scholarly Paper*. ID 2849145. Rochester, NY: Social Science Research Network.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Retrieved May 27, 2023 (https://arxiv.org/abs/1810.04805).

DiMaggio, Paul, Clark Bernier, Charles Heckscher, and David Mimno. 2018. "Interaction Ritual Threads: Does IRC Theory Apply Online?" P. 50 in *Ritual, emotion, violence: Studies on the micro-sociology of Randall Collins*.

Douglas, Mary. 1966. *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge & K. Paul.

Driskell, Robyn Bateman, and Larry Lyon. 2002. "Are Virtual Communities True Communities? Examining the Environments and Elements of Community." *City & Community* 1(4):373–90. doi: 10.1111/1540-6040.00031.

Durkheim, Émile. [1912] 2001. *The Elementary Forms of Religious Life*. Oxford University Press.

Dym, Brianna, Jed R. Brubaker, Casey Fiesler, and Bryan Semaan. 2019. ""Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work." *Proceedings of the ACM on Human-Computer Interaction* 3:1–28. doi: 10.1145/3359256.

Eliasoph, Nina. 1998. *Avoiding Politics: How Americans Produce Apathy in Everyday Life*. Cambridge University Press.

Eliasoph, Nina, and Paul Lichterman. 2003. "Culture in Interaction." *American Journal of Sociology* 108(4):735–94. doi: 10.1086/367920.

Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930-1955." *Studies in Linguistic Analysis*.

Fischer, Claude S. 1975. "Toward a Subcultural Theory of Urbanism." *American Jour-*

nal of Sociology 80(6):1319–41.

Foote, Jeremy, Aaron Shaw, and Benjamin Mako Hill. 2023. "Communication Networks Do Not Predict Success in Attempts at Peer Production." *Journal of Computer-Mediated Communication* 28(3):zmad002. doi: 10.1093/jcmc/zmad002.

Forstie, Clare. 2020. "Theory Making from the Middle: Researching LGBTQ Communities in Small Cities." *City & Community* 19(1):153–68. doi: 10.1111/cico.12446.

Foucault, Michel. 1998. "Friendship as a Way of Life." Pp. 135–40 in *Ethics: Subjectivity and Truth*, edited by P. Rabinow. New York: The New Press.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35(4):665–68. doi: 10.1080/10584609.2018.1477506.

Frost, David M., and Ilan H. Meyer. 2012. "Measuring Community Connectedness Among Diverse Sexual Minority Populations." *Journal of Sex Research* 49(1):36–49. doi: 10.1080/00224499.2011.565427.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes."

Ghaziani, Amin. 2014a. "Measuring Urban Sexual Cultures." *Theory and Society* 43(3-4):371–93. doi: 10.1007/s11186-014-9225-4.

Ghaziani, Amin. 2014b. *There Goes the Gayborhood?* Princeton University Press.

Gieryn, Thomas F. 2000. "A Space for Place in Sociology." *Annual Review of Sociology* 26:463–96.

Gieseking, Jen Jack. 2020. "Mapping Lesbian and Queer Lines of Desire: Constellations of Queer Urban Space." *Environment and Planning D: Society and Space* 0263775820926513. doi: 10.1177/0263775820926513.

Goldberg, Amir, Sameer B. Srivastava, V. Govind Manian, William Monroe, and Christopher Potts. 2016. "Fitting In or Standing Out? The Tradeoffs of Structural

and Cultural Embeddedness." *American Sociological Review* 81(6):1190–1222. doi: 10.1177/0003122416671873.

Gonen, Hila, and Yoav Goldberg. 2019. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them."

Halberstam, Jack. 2005. *In a Queer Time and Place: Transgender Bodies, Subcultural Lives*. NYU Press.

Halperin, David M. 2012. *How To Be Gay*. Sew edition. Cambridge, Mass: Belknap Press.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016a. "Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change."

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016b. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change."

Hampton, Keith, and Barry Wellman. 2003. "Neighboring in Netville: How the Internet Supports Community and Social Capital in a Wired Suburb." *City & Community* 2(4):277–311. doi: 10.1046/j.1535-6841.2003.00057.x.

Hargittai, Eszter, and Aaron Shaw. 2015. "Mind the Skills Gap: The Role of Internet Know-How and Gender in Differentiated Contributions to Wikipedia." *Information, Communication & Society* 18(4):424–42. doi: 10.1080/1369118X.2014.957711.

Held, Nina. 2017. "'They Look at You Like an Insect That Wants to Be Squashed': An Ethnographic Account of the Racialized Sexual Spaces of Manchester's Gay Village." *Sexualities* 20(5-6):535–57. doi: 10.1177/1363460716676988.

Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. Random House.

Jiang, Jialun Aaron, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. "Moderation Challenges in Voice-based Online Communities on Discord." *Proceedings of the ACM on Human-Computer Interaction* 3:1–23. doi:

10.1145/3359157.

Johfre, Sasha Shen, and Jeremy Freese. 2021. "Reconsidering the Reference Category." *Sociological Methodology* 0081175020982632. doi: 10.1177/0081175020982632.

Jones, Jason, Mohammad Amin, Jessica Kim, and Steven Skiena. 2020. "Stereotypical Gender Associations in Language Have Decreased Over Time." *Sociological Science* 7:1–35. doi: 10.15195/v7.a1.

Jurgenson, Nathan. 2011. "Digital Dualism Versus Augmented Reality." Retrieved April 29, 2021 (https://thesocietypages.org/cyborgology/2011/02/24/digital-dualism-versus-augmented-reality/).

Kennedy, Brendan, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. 2021. "Moral Concerns Are Differentially Observable in Language." *Cognition* 212:104696. doi: 10.1016/j.cognition.2021.104696.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 0003122419877135. doi: 10.1177/0003122419877135.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. "Statistically Significant Detection of Linguistic Change." Pp. 625–35 in *Proceedings of the 24th International Conference on World Wide Web - WWW '15*. Florence, Italy: ACM Press. doi: 10.1145/2736277.2741627.

Kusner, Matt J., Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. "From Word Embeddings To Document Distances." 10.

Lakoff, George. [1987] 2008. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Lang, Ken. 1995. "Newsweeder: Learning to Filter Netnews." Pp. 331–39 in *Proceedings*

of the twelfth international conference on machine learning.

Levine, Jeremy R. 2017. "The Paradox of Community Power: Cultural Processes and Elite Authority in Participatory Governance." *Social Forces* sf;sow098v1. doi: 10.1093/sf/sow098.

Levine, Jeremy R. 2021. *Constructing Community: Urban Governance, Development, and Inequality in Boston*. Princeton University Press.

Lichterman, Paul. 1999. "Talking Identity in the Public Sphere: Broad Visions and Small Spaces in Sexual Identity Politics." *Theory and Society* 28(1):101–41.

Lizardo, Omar. 2017. "Improving Cultural Analysis: Considering Personal Culture in Its Declarative and Nondeclarative Modes." *American Sociological Review* 82(1):88–115. doi: 10.1177/0003122416675175.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(02):254–77. doi: 10.1093/pan/mpu019.

Lucy, Li, and David Bamman. 2021. "Characterizing English Variation Across Social Media Communities with BERT." *Transactions of the Association for Computational Linguistics* 9:538–56. doi: 10.1162/tacl_a_00383.

Martin, John Levi, and Alessandra Lembo. 2020. "On the Other Side of Values." *American Journal of Sociology* 126(1):52–98. doi: 10.1086/709778.

Martin, John Levi, and Alessandra Lembo. 2021. "Response to Vaisey." *Sociological Forum* n/a(n/a). doi: 10.1111/socf.12790.

Marx, Karl. [1848] 1972. "The Communist Manifesto." in *The Marx-Engels Reader*, edited by R. C. Tucker. New York, Norton.

Mattson, Greggor. 2015a. "Bar Districts as Subcultural Amenities." *City, Culture and Society* 6(1):1–8. doi: 10.1016/j.ccs.2015.01.001.

Mattson, Greggor. 2015b. "Style and the Value of Gay Nightlife: Homonormative Placemaking in San Francisco." *Urban Studies* 52(16):3144–59. doi: 10.1177/0042098014555630.

Mattson, Greggor. 2020. "Small-City Gay Bars, Big-City Urbanism." *City & Community* 19(1):76–97. doi: 10.1111/cico.12443.

McCulloch, Gretchen. 2019. *Because Internet: Understanding the New Rules of Language*. Penguin.

Medvedev, Alexey N., Renaud Lambiotte, and Jean-Charles Delvenne. 2018. "The Anatomy of Reddit: An Overview of Academic Research."

Mehra, Bharat, Cecelia Merkel, and Ann Peterson Bishop. 2004. "The Internet for Empowerment of Minority and Marginalized Users." *New Media & Society* 6(6):781–802. doi: 10.1177/146144804047513.

Meyer, Ilan H. 2020. "Generations: A Study of the Life and Health of LGB People in a Changing Society, United States, 2016-2019: Version 1."

Meyer, Ilan H., and Soon Kyu Choi. 2020. *Differences Between LGB Democrats and Republicans in Identity and Community Connectedness*. Williams Institute, UCLA School of Law.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space."

Mize, Trenton D. 2016. "Sexual Orientation in the Labor Market." *American Sociological Review* 0003122416674025. doi: 10.1177/0003122416674025.

Mohr, John W., Christopher A. Bail, Margaret Frye, Jennifer C. Lena, Omar Lizardo, Terence E. McDonnell, Ann Mische, Iddo Tavory, and Frederick F. Wherry. 2020. *Measuring Culture*. Columbia University Press.

Monk, Ellis P. 2022. "Inequality Without Groups: Contemporary Theories of Cate-

gories, Intersectional Typicality, and the Disaggregation of Difference." *Sociological Theory* 07352751221076863. doi: 10.1177/07352751221076863.

Mustillo, Sarah A., Omar A. Lizardo, and Rory M. McVeigh. 2018. "Editors' Comment: A Few Guidelines for Quantitative Submissions." *American Sociological Review* 83(6):1281–83. doi: 10.1177/0003122418806282.

Nelson, Laura K. 2017. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research* 0049124117729703. doi: 10.1177/0049124117729703.

Nelson, Laura K. 2019. "To Measure Meaning in Big Data, Don't Give Me a Map, Give Me Transparency and Reproducibility." *Sociological Methodology* 49(1):139–43. doi: 10.1177/0081175019863783.

Nelson, Laura K. 2021. "Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South." *Poetics* 101539. doi: 10.1016/j.poetic.2021.101539.

Oldenburg, Ray. 1998. *The Great Good Place: Cafés, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community*. Marlowe.

Olson, Randal S., and Zachary P. Neal. 2015. "Navigating the Massive World of Reddit: Using Backbone Networks to Map User Interests in Social Media." *PeerJ Computer Science* 1:e4. doi: 10.7717/peerj-cs.4.

Orne, Jason. 2011. "'You Will Always Have to "Out" Yourself': Reconsidering Coming Out Through Strategic Outness." *Sexualities* 14(6):681–703. doi: 10.1177/1363460711420462.

Orne, Jason. 2013. "Queers in the Line of Fire: Goffman's Stigma Revisited." *The Sociological Quarterly* 54(2):229–53. doi: 10.1111/tsq.12001.

Orne, Jason. 2017. *Boystown: Sex and Community in Chicago*. Chicago ; London: University of Chicago Press.

Ottensmann, John R. 2018. "On Population-Weighted Density." Retrieved April 1, 2023.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–30.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." Pp. 1532–43 in *Empirical methods in natural language processing (EMNLP)*.

Pew Research Center. 2013. "A Survey of LGBT Americans." Retrieved April 29, 2021 (https://www.pewresearch.org/social-trends/2013/06/13/a-survey-of-lgbt-americans/).

Putnam, Robert D. 2001. *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster.

Regan, Amanda, and Eric Gonzaba. 2019. "Mapping the Gay Guides." Retrieved (http://www.mappingthegayguides.org).

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." Pp. 45–50 in *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*. Valletta, Malta: ELRA.

Rheault, Ludovic, and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28(1):112–33. doi: 10.1017/pan.2019.26.

Rheingold, Howard. 2000. *The Virtual Community: Homesteading on the Electronic*

Frontier. 2 edition. The MIT Press.

Rodriguez, Pedro L., and Arthur Spirling. 2020. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *Journal of Politics* 54.

Rosch, Eleanor, and Carolyn B. Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7(4):573–605. doi: 10.1016/0010-0285(75)90024-9.

Rosenfeld, Michael J. 2017. "Moving a Mountain: The Extraordinary Trajectory of Same-Sex Marriage Approval in the United States." *Socius* 3:2378023117727658. doi: 10.1177/2378023117727658.

Sahlgren, Magnus. 2008. "The Distributional Hypothesis." *Italian Journal of Linguistics* 20(1):33–53.

Saussure, Ferdinand de. [1916] 1972. *Cours de linguistique générale*. Paris: Payot.

Schnabel, Landon. 2018. "Sexual Orientation and Social Attitudes." *Socius* 4:1–18. doi: 10.1177/2378023118769550.

Simmel, Georg. 1971. *Georg Simmel on Individuality and Social Forms*. edited by D. Levine. Chicago, University of Chicago Press.

Simmel, Georg. [1903] 1971. "The Metropolis and Mental Life." in *Georg Simmel on Individuality and Social Forms*, *Heritage of sociology*, edited by D. Levine. Chicago, University of Chicago Press.

Soni, Sandeep, Lauren F. Klein, and Jacob Eisenstein. 2021. "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers." *Journal of Cultural Analytics* 18841. doi: 10.22148/001c.18841.

Stacey, Lawrence, and TehQuin D. Forbes. 2021. "Feeling Like a Fetish: Racialized Feelings, Fetishization, and the Contours of Sexual Racism on Gay." *The Journal of*

Sex Research 14. doi: 10.1080/00224499.2021.1979455.

Stillwagon, Ryan, and Amin Ghaziani. 2019. "Queer Pop–Ups: A Cultural Innovation in Urban Life." *City & Community* 18(3):874–95. doi: 10.1111/cico.12434.

Stoltz, Dustin S., and Marshall A. Taylor. 2019. "Concept Mover's Distance: Measuring Concept Engagement via Word Embeddings in Texts." *Journal of Computational Social Science* 2(2):293–313. doi: 10.1007/s42001-019-00048-6.

Stoltz, Dustin S., and Marshall A. Taylor. 2020. *Cultural Cartography with Word Embeddings*. SocArXiv.

Stoltz, Dustin S., and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics* 88:101567. doi: 10.1016/j.poetic.2021.101567.

Tavory, Iddo. 2016. *Summoned: Identification and Religious Life in a Jewish Neighborhood*. University of Chicago Press.

Taylor, Marshall A., and Dustin S. Stoltz. 2020. "Integrating Semantic Directions with Concept Mover's Distance to Measure Binary Concept Engagement." 18.

Tönnies, Ferdinand. [1887] 2001. *Community and Civil Society*. edited by J. Harris. Cambridge ; New York: Cambridge University Press.

Tufekci, Zeynep. 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." *ICWSM* 14:505–14.

Vaisey, Stephen. 2007. "Structure, Culture, and Community: The Search for Belonging in 50 Urban Communes." *American Sociological Review* 72(6):851–73.

van Loon, Austin, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. 2022. "Regional Negative Bias in Word Embeddings Predicts Racial Animus–but Only via Name Frequency."

Vrana, Adele Godoy, Anasuya Sengupta, and Siko Bouterse. 2020. "16 Toward a Wikipedia For and From Us All." in *::wikipedia @ 20*.

Walker, Kyle, and Matt Herman. 2023. *Tidycensus: Load US Census Boundary and Attribute Data as 'Tidyverse' and 'Sf'-Ready Data Frames*.

Waller, Isaac, and Ashton Anderson. 2019. "Generalists and Specialists: Using Community Embeddings to Quantify Activity Diversity in Online Platforms." Pp. 1954–64 in *The World Wide Web Conference on - WWW '19*. San Francisco, CA, USA: ACM Press. doi: 10.1145/3308558.3313729.

Waller, Isaac, and Ashton Anderson. 2021. "Quantifying Social Organization and Political Polarization in Online Platforms." *Nature* 1–5. doi: 10.1038/s41586-021-04167-x.

Warmerdam, Vincent, Thomas Kober, and Rachael Tatman. 2020. "Going Beyond T-SNE: Exposing Whatlies in Text Embeddings." Pp. 52–60 in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. Online: Association for Computational Linguistics.

Weston, Kath. 1995. "Sexual Imaginary and the Great Gay Migration." *GLQ: A Journal of Lesbian and Gay Studies* 2(3):253–77.

White, Harrison C. [1965] 2008. "Notes on the Constituents of Social Structure. Soc. Rel. 10 - Spring '65." *Sociologica* (1):0–0. doi: 10.2383/26576.

Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529–44.

Winer, Canton. 2020. "Solidarity, Disdain, and the Imagined Center of the Gay Imagined Community." *Sociological Inquiry* n/a(n/a). doi: 10.1111/soin.12403.

Wood, S. N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society (B)* 73(1):3–36.

Zerubavel, Eviatar. 2002. "The Fine Line: Making Distinctions in Everyday Life." Pp. 223–32 in *Cultural Sociology*, edited by L. Spillman. Malden, MA: Wiley-Blackwell.

Zerubavel, Eviatar. 2018. *Taken for Granted: The Remarkable Power of the Unremarkable*. Princeton ; Oxford: Princeton University Press.

Zhang, Justine, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. "Community Identity and User Engagement in a Multi-Community Landscape." 10.