Dynamics of Homophily: How Similarity in Enacted Identity Shapes Social Ties [*]

Lara Yang
Stanford University

September 1, 2023

This is a working paper. The most recent version can be found here.

Abstract

Homophily is a fundamental principle that orders and structures social ties. Existing work conceptualizes homophily as a static phenomenon. In the commonly studied case of gender homophily, for instance, two individuals either share the same gender or they do not. However, a core insight in the identity literature is that identities are enacted as a function of social contexts and interactions. Integrating this insight, I maintain that homophily can be a dynamic, contextualized, and interactional process. Building on prior work, I theorize that similarity in enacted identity predicts tie existence and strengthens existing ties. I deconstruct enacted identity similarity into its observable and unobservable components, contending that observability is an important mechanism that underlies homophily. Under the contextual view of identity, observable and unobservable enacted identities should diverge and only observable enacted identity similarity should strengthen social ties. Finally, I contend that the effect of observable enacted identity similarity is amplified when observed in private contexts, as privacy renders the enacted identity more authentic and intimate. By applying word embedding models to Slack communication records, I develop a novel approach to measuring enacted identity and its similarity. By analyzing channel membership on Slack, I estimate the extent to which such enacted identity similarity is observable. Combining this approach with responses from a network survey, I find consistent support for my hypotheses. Taken together, these findings demonstrate that homophily is an ongoing, dyadic process between the presentation of self and the observation of others.

[*] Direct all correspondence to Lara Yang: larayang@stanford.edu

Homophily, the tendency to associate with similar others, is a fundamental principle in social life. It is ubiquitous in the social world, and acts as a rule-like gravitational force that guides interpersonal relationships (McPherson, Smith-Lovin, and Cook 2001; DiMaggio and Garip 2012; Leszczensky and Pink 2019; Lawrence and Shah 2020). Homophily can operate along various dimensions and attributes, such as gender, race, attitudes, and experiences (see McPherson et al. 2001 for a review). At a macro scale, homophily shapes important social issues, such as segregation, polarization, and inequality (Ertug, Brennecke, Kovács, and Zou 2022; McPherson et al. 2001). It is particularly consequential in work organizations, given its implications for access to social capital, communication and coordination, and individual and organizational performance (Ertug et al. 2022).

In conceptualizing homophily, past work defines "individuals as similar to the extent that both hold some characteristic or attribute in common (Lawrence and Shah 2020, p. 523). This approach to conceptualizing and measuring similarity treats it as a stable and often binary attribute. For example, when studying racial homophily, two individuals are deemed racially similar if they are members of the same race, and not racially similar otherwise. Many commonly studied dimensions of homophily, such as gender, race, values (McPherson et al. 2001), are thought to be relatively stable over time and across situations. In these cases, this assumption is not incorrect. As one's membership to nominal categories does not vary across contexts, similarity in these categories can also be viewed as context-free. However, in other cases, this approach can also be a gross oversimplification.

Contrary to existing work, I argue that similarity, and, by extension, homophily, is constructed, dynamic, and contextual. I trace my argument to the long-established view of identity as a multifaceted product of intersecting social situations (Callero 2003; Cerulo 1997; Ramarajan 2014; Lahire 2011). This idea traces its origins to the foundational work done by Cooley (1902) and Mead (1934), and is later elaborated by Goffman (1959). This line of work argues that identity results from ongoing interactive performances put on for specific audiences in given contexts. As identity varies as a function of the audience and context, there is no single true self (Ramarajan 2014). The structure of the self is as "multidimensional and diverse as the social relationships" (Callero 2003, p. 127) in which one is surrounded. As such, similarity between two individuals should also vary by context as it depends on the identity each person presents.

Inspired by this line of work, this paper is centered on homophily in identity. Existing work distinguishes structural homophily, or similarity induced by constraints in opportunity structure, and choice homophily, or similarity induced by a preference for similar others. Identity and choice homophily are definitionally intertwined. As choice homophily is an attraction between similar individuals, identities of these individuals affect perceptions of similarity and are implicated in arguments on choice homophily. The role identity plays in homophily has also been explicitly explored (Ingram 2023; Leszczensky and Pink 2019; Mehra, Kilduff, and Brass 1998; Reagans 2005; Mollica, Gray, and Trevino 2003). Ingram (2023) focuses on self-reported identity elements. He finds that similarity in identity predicts tie formation, and mediates the link between identity multiplicity and the likelihood of professional tie formation. Other papers focus on the moderating role of identity strength, demonstrating that the strength of a given identity moderates the degree to which similarity in this nominal identity predicts relationship formation (Leszczensky and Pink 2019; Mehra et al. 1998; Reagans 2005; Mollica et al. 2003).

A key limitation in existing work is that it assumes identity as unitary and stable. Identity is treated as a set of fixed self-definitions that are applied across all situations and directly influence relational outcomes, and measures it using self-reported surveys. However, this assumption is at odds with research on identity. Sociological theories of identity contend that identities are enacted and performed, and these identity enactments vary from context to context and from interaction to interaction (Owens, Robinson, and Smith-Lovin 2010; Lahire 2011; Fine 2012; Stets and Burke 2003; Deaux 1993).

Integrating the literatures on identity and homophily, I develop a set of theoretical expectations on *enacted identity homophily.* First, I argue that enacted identity similarity predicts the existence of a tie even when controlling for homophily in nominal, ascriptive categories. Conditional on the existence of a tie, I predict that enacted identity similarity should positively relate to tie strength as it improves the quality of interactions between two individuals.

Subsequently, I unpack how enacted identity homophily operates by zooming in on how individuals learn that others are similar to them. To understand how similar others are to oneself, one needs to observe their enacted identities. One cannot directly access their cognitive states or their identity enactments outside of one's direct purview. Thus, only observable enacted identity similarity should be positively related to the strength of ties. These predictions are shown to be true, they would lend support to the idea that identity as enacted in social interactions, as opposed to identity as a stable set of self-referential meanings, drives homophily. If identity were instead stable as prior work has assumed, whether it were observed or not should not matter.

Finally, I examine how the effect of observable enacted identity similarity is moderated by the extent to which it is enacted in private. As seasoned audiences of others' identity performances, individuals may seek to evaluate the degree of performativity in others' enacted identities. Thus, observable enacted identity similarity should matter more when the identity is enacted in more private contexts, where identity enactments are assumed to be more authentic and intimate.

To test my theory, I need to measure enacted identity. As language is our primary mode of communication and expression, I develop a language-based model of enacted identity that draws from recent advances in natural language processing. I employ word embedding models (Mikolov, Sutskever, Chen, Corrado, and Dean 2013; Pennington, Socher, and Manning 2014) to represent words in a high-dimensional vector space. To measure enacted identity for each person, I train a company-wide word embedding model, then adapt it at the individual level using a finetuning algorithm. I measure enacted identity as the semantic meanings associated with the usage of the first-person singular pronoun "I." Subsequently, I operationalize enacted identity similarity between individuals as the semantic similarity between their "I"s, quantified as the cosine distance between word embeddings of the words "I." Conceptually, it represents the degree to which two individuals enact a similar set of self-referential meanings when referring to themselves.

I apply my language-based model of enacted identity to a corpus of anonymized Slack communications data collected from a non-profit firm based in the United States. In addition, I supplement this data with matched personnel records and responses from a roster-based network survey. To unpack observability, one needs to be able to track what each individual under study can and cannot observe, which is virtually impossible to do using self-reports. Disentangling observability thus requires detailed documentation of the social interactions in which all individuals are embedded. This organization is geographically distributed and operates virtually, where most social interactions take place on the widely used messaging application Slack. Thus, it is a perfect setting to test my theory. By examining the structure and content of digital workplace communications, I can pinpoint what is and is not observable for each individual in the study population.

After validating my measure of enacted identity, I test my hypotheses using linear regression models and find consistent support for my theory. Methodologically, this paper contributes to existing work on identity by developing a new measurement approach that examines enacted identity directly. This measure reflects the multifaceted, contextual nature of identity. It can also reveal facets of identity that people may be unable or unwilling to express in a survey, an approach that dominates existing work on identity. In doing so, I add to the growing field of work that uses computational tools to measure constructs of relevance to organizational scholars and social scientists writ large. Most importantly, by demonstrating how enacted identity impacts relationships, this work reconceptualizes homophily from a stable, static phenomenon to one that is dynamically, contextually enacted. Homophily is thus an ongoing interactional process that is influenced by the presentation of the self and the observation of the other.

Enacted Identity Homophily

Homophily operates in a variety of social settings (McPherson et al. 2001). Marriage-based ties demonstrate substantial homophily, especially in terms of race, ethnicity, and religion (McPherson et al. 2001). Friendships in schools also tend to be homophilous, especially along the dimensions of race and gender (Shrum, Cheek Jr, and MacD 1988; Joyner and Kao 2000; Stehlé, Charbonnier, Picard, Cattuto, and Barrat 2013). Furthermore, homophily influences social ties in voluntary associations (McPherson and Smith-Lovin 1987). Although the theory proposed in this paper is generalizable, I specifically examine homophily in work organizations, an especially important context for understanding homophily given that a large proportion of non-kinship ties of adults are formed at work (McPherson et al. 2001). Homophily is consequential for these organizations, as it can lead to both positive outcomes by improving coordinatio ysn and communication, and negative outcomes by reducing diversity in perspectives and knowledge (Ertug et al. 2022). Homophilous ties at work have also been found to form among many dimensions, such as gender (Ibarra 1992), tenure (Reagans 2005), and nationality (Rhee, Yang, and Yoo 2013).

Amongst the many dimensions along which homophily has been documented, research on homophily in ascriptive characteristics, especially gender and race, has dominated the field (McPherson et al. 2001; Lawrence and Shah 2020). These categories have been shown to structure social relationships across social contexts. The argument for why pervasive homophily exists along these dimensions is that people who share categorical membership tend to also share attitudes, values, and behaviors that are rooted in common language and experiences (Leszczensky and Pink 2019). Implicit in these arguments is the thread of identity, as the impact of these categorical memberships on attitudes, values, and behaviors depends on whether they are self-defining (Tajfel and Turner 1979; Stets and Burke 2000).

Some scholars have sought to make the role identity plays in homophilous processes more explicit. Bringing extant theories of identity into homophily research, several papers demonstrate that similarity in ascriptive categories affects relational outcomes to the extent that individuals identify with them (Leszczensky and Pink 2019; Reagans 2005; Mehra et al. 1998; Mollica et al. 2003). More recently, Ingram (2023) formalizes the tendency for individuals with similar identities to associate with each other as "identity homophily," and finds that it predicts the formation of new professional ties. This formulation extends the understanding of homophily from a few social categories to a holistic view of individuals by taking into account the unlimited number of elements that define individuals and influence their actions.

However, the way in which prior work conceptualizes identity belies the complexities of identity as a construct. Existing work implicitly assumes that identity, and by extension identity similarity, is stable and unitary. For example, Ingram (2023) conceptualizes identity similarity as overlap in a set of identity elements. To operationalize identity similarity, he asks individuals to write down a list of identity elements that define them, and then measures the degree of semantic distance in these elements between individuals.

This assumption is contrary to extensive work that maintains that identity is rarely fixed; instead, it is enacted and expressed as a function of social contexts and interactions (Callero 2003; Cerulo 1997; Ramarajan 2014; Lahire 2011). Erving Goffman's now classic work, *The Presentation of Self in Everyday Life*, is foundational to this idea. In his work, Goffman (1959) articulates his thesis that the self is a social process that emerges out of social interactions and exchanges (Goffman 1959; Lawler 2015). In his dramaturgical metaphor, identity rises out of the performances one orchestrates. These arguments imply that what leads to identity homophily is not necessarily identity in a vacuum, but its enactment in action.

In this paper, I thus focus on enacted identity homophily, or the tendency for two individuals with similar enacted identities to associate with one another. More concrete examples might help illustrate how it departs from prior work. In prior work, if two individuals both put down "Swiftie" (an ardent Taylor Swift fan) as an identity in their identity map, they would be seen as possessing similar identities and therefore more likely to form a strong relationship. My theory, however, argues that whether this similarity predicts tie formation depends on whether the "Swiftie" identity is enacted. In professional settings, for instance, one's commitment to Taylor Swift is unlikely to come up. Therefore, my theory would predict that two individuals are more likely to form a bond only if their "Swiftie" identities are enacted.

Conversely, certain identities are unlikely to be reported in self-reports. Consider smoking cigarettes as an example. Campaigns of the detrimental effects of tobacco consumption and its associated stigma render it less probable for individuals to include "smoker" as an identity when responding to a survey. Despite this unwillingness to self-label, this identity might still be enacted at social events or in watercooler conversations. If so, it likely will shape one's social relationships.

In line with theoretical arguments underlying existing work, similarity in enacted identity should increase the likelihood that a meaningful social tie exists. Similarly enacted identities promote interpersonal attraction as it provides a common ground and a common language (Byrne 1961; McPherson et al. 2001). This attraction serves as an incentive for relationships to form. Thus, I form the following expectation:

Hypothesis 1a: *Enacted identity similarity is positively associated with the likelihood of a tie nomination.*

Conditional on the existence of a tie, enacted identity similarity should also predict tie strength. The association between similarity and tie strength has been treated as a form of homophily by some (Reagans 2005, 2011; Marsden 1988), and as an outcome of homophily, defined as the association between similarity and tie existence, by others (Ertug et al. 2022). Following the former tradition, I treat the positive association between similarity and tie strength as a form of homophily. In organizational networks, whether two individuals are connected is often more a function of task- and role-based constraints than individual preferences (Lincoln and Miller 1979). In these settings, tie strength should be more reflective of individual compatibility and preferences, key to the "choice" in choice homophily. Thus, going forward, I only focus on the relationship between similarity and tie strength.

That tie strength is predicated on the similarity between two actors is central to Granovetter's strength-of-weak-ties thesis, where he argues that the redundancy that results from strong ties is due to the fact that those who are connected by strong ties, but not weak ties, tend to be more similar to one another (Granovetter 1973). Reagans, Argote, and

Brooks (2005) finds that those with similar tenure communicate more frequently. Looking at similarities in structural positions, Friedkin (1993) again finds homophily—structurally equivalent individuals communicate more frequently. More recently, Reagans (2011) finds that age similarity has a positive effect on tie strength, and this effect is intensified by physical proximity.

Enacted identity similarity should strengthen existing ties. Similarity breeds attraction and eases interpersonal communication (Byrne 1961; McPherson et al. 2001). As self-definitions are central to how individuals view and interact with others around them, shared identity is particularly valuable for building strong connections (Tajfel and Turner 1986; Reagans 2005; Ingram 2023). Enacted identity similarity should make it more rewarding for people to communicate and interact, as they are more likely to resonate and emotionally connect with each other. In particular, enacted identity similarity should also facilitate intimate exchanges, as it provides common ground and aids the development of trust over time. These intimate and repeated exchanges are the key to strengthening social bonds (Granovetter 1973). Thus, I predict:

Hypothesis 1b: *Enacted identity similarity is positively associated with tie strength.*

Separating Enacted Identity Similarity into its Observable and Unobservable Components

The insight that identity is dynamically enacted in social interactions and contexts reveals additional nuances of how enacted identity homophily may function. Insofar as interactions and contexts vary, the self will not be a singular, unified entity, and instead is multifaceted and contextual (Goffman 1959; Lawler 2015; Cerulo 1997). This idea is the central thesis put forward by Lahire in *The Plural Actor* (2011). As implied by the title, Lahire argues that each person intersects a plurality of identities, each instantiated through interactions. Integrating these insights, enacted identity homophily should not only consider similarity between two individuals' enacted identity overall, but also take into account the contextual nature of identity enactments.

Key to understanding the contextual nature of identity homophily might be a change in perspective, from that of the researcher to that of the individuals under study. Most studies on homophily conceptualize similarity from the perspective of the researcher, defining it based on whether the individuals in the dyadic relationship possess a characteristic in common (Lawrence and Shah 2020). This approach, however, is incomplete in that it misses the social reality individuals live in. Instead of only operating under a bird's-eye view of the dyads and their similarities, we should also seek to understand the perspective of the individuals under study and examine how they perceive and view their interaction partners. Adopting this perspective highlights the fact that the individuals under study can only observe the identities their partners enact in their presence. Given the contextual nature of identity enactments, these enacted identities can diverge from those that one does not observe, given that they are enacted in different contexts. Taken together, for the individuals under study, their impression of whether their identity is similar to that of another individual is necessarily constrained by observability.

Although the role it plays in homophily has yet to be examined, the construct of observability has piqued the interest of a few sociologists. Analyzing aristocracies, Simmel (1950) uses the term "surveyable" to describe the extent to which role performances of actors can be observed by others, which he argues is central to the effectiveness of aristocracy. Building on Simmel's writings, Merton (1968) elaborated extensively on this construct. Merton views observability as a property of social groups, referring to the extent to which role performances and norms are readily observable (Jaworski 1990; Merton 1968). It is largely determined by group structure and is of vital importance for the proper and effective functioning of groups. This construct has been applied in relation to understanding social control and power, where observability is seen as a predecessor to social surveillance (Warren 1968; Friedkin 1993). More recently, scholars have also attempted to generalize visibility into a field that stands on its own (Brighenti 2007).

Extending the construct to the dyadic level, observability has important implications for understanding identity homophily. When an actor attempts to ascertain, either explicitly or implicitly, how similar her identity is compared to that of her alters, she relies on the identity enacted by her alters within her purview. The rest of alters' identity enactments are invisible to the focal actor and therefore feature minimally in her perception of identity similarity with her alters. Thus, I compartmentalize enacted identity similarity into observable and unobservable enacted identity similarity. I define *observable enacted identity similarity* between two individuals as similarity between ego's enacted identity and alter's enacted identity that is observable to ego. I define *unobservable enacted identity similarity* between two individuals as similarity between ego's enacted identity and alter's enacted identity that is unobservable to ego. As such defined, within a given dyad, observable and unobservable enacted identity similarity are both asymmetric. In both of these definitions, the entirety of ego's identity is taken into account, as ego can observe all of her own enacted identity without constraints.

Observable enacted identity similarity should be positively associated with tie strength. In order for the relational benefits of enacted identity similarity to accrue to a relationship, similarity needs to be observed. Liking, trust, and intimacy that build from interacting with others with similar identities should hinge on one's ability to see these identities enacted. On the other hand, unobserved enacted identity similarity should have a much more subdued influence on tie strength. Under the assumption that identity enactments are varied and multifaceted, unobservable enacted identity similarity should diverge from observable enacted identity similarity. Without the channels of direct observation, it is difficult to infer similarity and for such similarity to influence tie strength. Thus, I hypothesize that only observable enacted identity similarity contributes to the development of strong ties.

Hypothesis 2: *Observable enacted identity similarity is positively associated with tie strength.*

This hypothesis differs meaningfully from predictions of extant theories of identity homophily. If identity were stable and unitary as assumed by prior work, there would be no difference between observed and unobserved enacted identity similarity. Continuing my earlier example, while prior work predicts that two individuals who both include "Swiftie" as an identity element will have a strong tie, I would predict that the strength of their tie depends on whether this identity has been enacted in front of each other.

Privacy of Context Moderates the Effect of Observable Enacted Identity Similarity

While similarity in observable enacted identity should on average positively influence tie strength, this effect is not necessarily universal. The context in which one's identity is enacted and observed should matter for its impact. A core distinction in sociological discourse might be particularly relevant and useful for the analysis here: the public versus private distinction (Bailey 2000; Brewer 2005; Slater 1998; Weintraub and Kumar 1997). Broadly speaking, private refers to "areas of social life which are protected from anything other than personal or domestic gaze" (p.384), separating a domain of experience from what is public and open to surveillance and control (Bailey 2000). Although this distinction is posed as a dichotomy, in reality, what is private and what is not is more ambiguous and often a question of degree (Brewer 2005).

This distinction is particularly illuminating in unpacking when the effect of observable enacted identity similarity is expected to be amplified or muted. Specifically, identities enacted in a relatively private realm should be perceived to be much more authentic, intimate, and convincing than those on display for all. Lay beliefs of identity contrast an authentic identity with a performative identity, with the latter seen as a case of pretension or acting (Lawler 2015). When an enacted identity is perceived as performative or insincere, it is likely to be discounted. In addition, private realms facilitate the development of intimacy (Bailey 2000). Therefore, identities enacted in a private context can more effectively create a sense of trust and intimacy. In other words, identity enactment is viewed as all the more convincing when it occurs privately.[1] This enactment should then have an amplified effect on tie strength when it serves as the basis from which observable enacted identity similarity is inferred.

Taken together, I thus predict that the effect of observable enacted identity similarity on tie strength should be moderated by the degree to which alters' identities are enacted privately. In other words:

Hypothesis 3: The effect of observed enacted identity similarity on tie strength is more positive when such similarity is inferred from private contexts.

METHODS

Empirical Setting and Data

To test my hypotheses, I employ Slack communication data, personnel records, and self-reported survey data collected from a midsized non-profit organization in North America. [2] The data used in this study span from July 2022 to July 2023. A noteworthy feature of

[1] Goffman (1959) argues that, when analyzing performances, instead of drawing the line between true or false performances, a more analytically useful distinction is between a convincing and an unconvincing performance (Lawler 2015).

[2] This data collection effort is achieved collaboratively by myself as well as several other colleagues as a part of a unrelated research project we have been working on. This paper is indebted to their collaboration in data collection, as well as the gracious support of the organization.

the organization under study is that it has transitioned to full distributed and remote work since the beginning of the COVID-19 pandemic. Slack is a popular messaging platform that serves as the primary mode of communication among employees in this organization. Thus, the vast majority of social interactions and exchanges that occurred during the study period are documented in the Slack data used for analysis.

The Slack communications data contains both the meta-data of Slack data (e.g., timestamp and sender ID) as well as the anonymized and de-identified content. To protect employee privacy and organizational confidentiality, I hashed or otherwise transformed raw message content and identifying information about employees. This data is used to measure enacted identity and infer enacted identity similarity, my main independent variable, which will be discussed in detail in the following section. In addition, I also use the frequency of Slack communications as a measure of tie strength in a robustness check. Between July 2021 and July 2022, the dataset contains 7.57 million messages and 148.94 million word tokens in total.

Measures of tie strength come from a network survey conducted in this organization in July 2022. This survey began by asking individuals to nominate alters with whom they have exchanged a meaningful interaction in the last three months. Typically, potential alters in network studies can be elicited using either network rosters (i.e. names of all potential alters are provided) or name generators (i.e., respondents are asked to generate the names of alters). The current network survey uses a network roster method, as past research has demonstrated that name-generator methods can suffer from issues of faulty recall. [3] (Agneessens and Labianca 2022)

Subsequently, respondents are shown several questions on the strength and content of the relationships they have with the alters they have previously nominated. However, given the size of this organization (around 1050 full-time employees when the survey was conducted), employees can have a large number of network alters, which can render responding to additional questions quite time-consuming and burdensome for respondents. Thus, based on the suggestions presented by Stadel and Stulp (2022), each respondent is asked to provide

[3] Given the size of the organization, the roster of the whole organization cannot be displayed at once. Specific survey procedures are as follows. Employees are first provided the names of all individuals in their own department from which they can make selections. Then, they are presented with a list of departments and are asked to select all departments with whom they have interacted. They are then presented with a list of individuals in each department and are asked to make tie nominations. Finally, employees have the opportunity to include additional colleagues in an open-ended response.

answers on tie strength for a subset of at most ten randomly selected alters. Tie strength is measured using a closeness scale, based on prior literature that suggests that affective closeness is the best conceptualization of tie strength (Marsden and Campbell 1984).

Finally, personnel records provided by the organization include both sociodemographic variables (i.e. gender and race) and job-relevant characteristics (i.e. tenure and department). These variables allow me to statistically control for endogeneity and structurally-induced homophily as much as possible. In addition, I use these variables to provide empirically estimates of and control for ascriptive homophily.

Tracing Enacted Identity in Language

To test my theory of enacted identity homophily, I need a new approach that can measure enacted identity directly. I propose to do so via language. As our primary mode of communication, language is a key channel through which identity is enacted and expressed. Much existing work has looked at tracking identity using language, either as a metaphor (Stets and Burke 2003; Callero 2003) or as a methodology (Ashokkumar and Pennebaker 2022; Pennebaker 2011; Yang, Goldberg, and Srivastava 2023). In particular, pronouns have served a unique role in the analysis of identity.

I propose to measure the enacted identity through the meanings people associate with their first-person singular pronoun "I." Self-referential meanings are the foundation of definitions of identity (Stets and Burke 2003; Ramarajan 2014; Stryker and Burke 2000). These meanings inhere in the use of "I," which spotlight the self as a meaningful, agentic, and individual entity. Prior work has highlighted that identity is reflected in the use of the word "I." In the early twentieth century, Mead (1934) used "I" and "me" as a discursive metaphor for identity. More recently, linguists have underscored the role of first-person pronouns in self-presentation in written texts (Ivanič 1998; Tang and John 1999). Furthermore, Yang et al. (2023) use first-person pronouns as a measure of group identification by tracking the semantic distance between self-identity represented by the first-person singular pronoun "I," and group identity, represented by the first-person plural pronoun, "we." Thus, in the current paper, I propose to measure enacted identity using the first-person pronoun "I." Accordingly, I measure enacted identity similarity between two individuals as the semantic similarity between their first-person singular pronouns.

Compared to existing approaches, this measure has several advantages. The most important advantage is that it explicitly captures identity as dynamically enacted in social contexts, allowing it to vary from one social interaction to another. This approach to measuring similarity also addresses recent critiques of measurement strategies that underlie existing homophily research (Lawrence and Shah 2020). Analyzing language produced by individuals in naturalistic settings provides a way to adopt a person-centric (instead of a researcher-centric) perspective to understanding and measuring homophily, as my measure of identity reflects social cues used by individuals in their everyday interactions when ascertaining the identity of others.

Furthermore, a language-based measure allows me to hone in on the key theoretical interest of the current paper—unpacking the role observability plays in homophily. This is nearly impossible to do with self-reports, as it would require participants to report what they can and cannot observe. Instead, analyzing observability necessitates a comprehensive and detailed understanding of the social interactions in which all individuals are embedded. Digital communications, particularly in virtual organizations, provide such a lens. By examining the structure of communications data and the content of language use, I can precisely identify what is observable and what is not for each individual.

Finally, the analysis of naturalistic archival data is less prone to the social desirability bias that plagues self-reports and can lower response burden for participants (Donaldson and Grant-Vallone 2002; Paulhus, Vazire, et al. 2007; Goldberg, Srivastava, Manian, Monroe, and Potts 2016). It also scales more easily to many employees and multiple organizations, which can be beneficial in establishing the generalizability of one's theory and findings.

Measuring Enacted Identity Similarity Using Word Embeddings

To measure the meanings people attach to their first-person singular pronouns, I use a class of machine learning models called word embeddings. Word embedding models quantify word meanings by representing each word as a continuous, multidimensional vector (also referred to as a word embedding) (Mikolov et al. 2013; Pennington et al. 2014). These vectors are generated based on the distribution of words. Words that have similar contexts, or similar neighboring words, are positioned close together in the vector space, and words with different contexts are positioned far apart. Distances between word vectors thus measure semantic similarities and differences between words. A common distance metric used to compare vectors is cosine distance, or cosine of the angle between the vectors.

To measure enacted identity similarity between two individuals, I compare the semantic meanings attached to their first-person singular pronouns. Formally, my operationalization of enacted identity similarity between two individuals is the cosine similarity between word embeddings of their respective "I." Intuitively speaking, my proposed measure of enacted identity similarity reflects the degree of convergence in the semantic meaning people attach to their first-person singular pronouns. Compared to using Large Language Models, such as BERT (Devlin, Chang, Lee, and Toutanova 2019) and GPT (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, et al. 2020), which are state-of-the-art in linguistic tasks but behave like a black box, this approach is guided by theoretical understandings of identity. In doing so, my approach melds theoretical insight with quantitative rigor.

An additional complication of the operationalization is observability. For each dyad, I am interested in measuring their enacted identity similarity but also their observable and unobservable enacted identity similarity with one another. To define (un)observable enacted identity similarity, I first need to identify which Slack messages are (un)observable. I do so by identifying the set of Slack channels in which both parties of the dyad are active over the observation window, where active is defined as having sent a message, replied to a message, or reacted to a message at least once. I use "channels" generically to refer to any Slack grouping where a set of individuals is intended to simultaneously receive all messages sent in that grouping, which can include public Slack channels, private Slack channels, and direct messages to one or multiple individuals. I assume that messages sent in any channel in which both parties of the dyad have been active are observable to both and that all messages sent outside of these channels are unobservable. I acknowledge that observable does not equate to actually observing and processing these messages, but believe this to be the best proxy of observability available.

As I define the construct of observability as dyadic, an observer and an observee are necessitated. The observee (henceforth referred to as u) produces a set of texts that is observable to the observer (referred to as v), on which observable enacted identity similarity can be deduced. For each dyad with members u and v, I define u's observable messages by v as the set of messages u has sent in all channels in which u and v are both active. Note that while observability is defined to be symmetric (if u can observe v, v can observe u), observable enacted identity similarity is asymmetric. From the observer v's perspective, their observable enacted identity similarity to the observee u should be based on the messages u sent that are observable to v and all the messages v has sent as they are by definition all observable to v. On the other hand, u's observable enacted identity similarity of v should be based on the messages v has sent that are observable to u and all of u's messages.

To compare word embeddings between individuals, one would typically need to train a set of word embeddings for each person on the texts she generates. However, most word embedding models, such as Word2Vec or GloVe, require large amounts of training data to generate high-quality word embedding models. To provide some context, the most compact version of Glove was trained on roughly a billion words (Pennington et al. 2014). The amount of data available per person in most datasets social scientists use, including the one in the current study, is decidedly thin in comparison.

To overcome the issue of small amounts of training data, I follow the word embedding finetuning approach taken by Yang et al. (2023). To develop individual-time-specific word embeddings and address the dataset size limitations of doing so, Yang et al. (2023) combines the GloVe word embedding model with a retrofitting-based finetuning technique called Mittens (Dingwall and Potts 2018). Specifically, the Mittens algorithm starts with pre-trained GloVe word embeddings and then finetune these embeddings on domain-specific data. In doing so, one can take advantage of high-quality word embeddings trained on thick, domain-free but also incorporate domain-specific information to develop domain-specific word embeddings.

Taken together, six sets of messages and their respective six sets of word embeddings are needed to compute enacted identity similarity, observable enacted identity similarity, and unobservable enacted identity similarity for each dyad. They are: 1) all messages sent by u, 2) all messages sent by v, 3) messages sent by u that are observable to v, 4) messages sent by v that are observable to u, 5) messages sent by u's messages that are unobservable to v, and 6) messages sent by v that are unobservable to u.

To train word embeddings on these sets of data, I implement the word embedding procedure as follows. I first train a set of word vectors on the full corpus of the Slack data sent by all employees during the study period using the GloVe algorithm, which I will refer to as $W_{company}$ (Pennington et al. 2014). [4][5] Subsequently, I finetune company-wide word embeddings on each of the six sets of data. Respectively, I arrive at six sets of

[4] Given the hashed nature of the message content (each unique word in the entire Slack corpus is converted to a unique 8-bit hash using the MD5 hashing algorithm), I cannot use pre-trained GloVe vectors as those vectors are defined for unhashed English words.

[5] The hyperparameters of the word embedding model I trained is as follows. I chose a window size of 10 and an embedding dimension of 50. 50 is chosen as the embedding dimension given the relative sparsity of the data. Mittens mincount parameter is set to 50. All other GloVe and Mittens hyperparameters are set to the default values.

word embeddings, 1) word embeddings of u, W_u, 2) word embeddings of v, W_v, 3) word embeddings of u's messages observable to v, $W_{u \to v}$, 4) word embeddings of v's messages observable to u, $W_{v \to u}$, 5) word embeddings of u's messages unobservable to v, $W_{u \not\to v}$, and 6) word embeddings of v's messages unobservable to u, $W_{v \not\to u}$.

To measure enacted identity similarity, I compare the word vector of "I" in all messages sent by u, referred to as $w_{I,u}$, to the word vector of "I" in all messages sent by v, referred to as $w_{I,v}$, and calculate the cosine distance between the two vectors. u's observable enacted identity similarity to v is defined as the cosine distance between the word vector of "I" in $W_{u \to v}$, referred to as $w_{I,u \to v}$, to the word vector of "I" in W_v, $w_{I,v}$, and vice versa for v's observable enacted identity similarity to u. Finally, I operationalize u's unobservable enacted identity similarity to v as the cosine similarity between $w_{I,u \not\to v}$ and $w_{I,v}$, and vice versa for v's unobservable enacted identity similarity to u.

Figure 1 provides a schematic overview of the word embedding training process.

[FIGURE 1 ABOUT HERE]

Variables

Using the dyad as the unit of analysis, the key independent variables are enacted identity similarity, observable enacted identity similarity, and unobservable enacted identity similarity. The main dependent variables of interest are tie nomination and tie strength. Importantly, similarity is computed on texts preceding the collection of the dependent variables. Privacy of context is used as a moderator to test H3. Gender, race, department, tenure, and number of tokens are included as control variables. Gender, race, department, and linguistic similarity are included to control for other forms of homophily.

Dependent Variable

Tie Nomination: The dependent variable used in testing Hypothesis 1a is tie nomination, a binary variable of whether u has nominated v as an alter. Provided a network roster of all possible alters, employees are asked to select names of other employees with whom they have interacted in a meaningful way. Tie nominations are unidirectional.

Tie Strength: The main dependent variable of interest is tie strength, operationalized as a subjective measure of closeness. Closeness is both the most frequently used measure of tie strength in previous work, and the most reliable and valid measure of tie strength available

Marsden and Campbell (1984). Employees are asked to indicate how close they are with each named alter using a 5-point Likert scale, ranging from really not close (1) to very close (5). As nominations are unidirectional and not necessarily reciprocated, the resulting ties are directed. As such, tie strength is a directed and asymmetrical measure.

Independent Variables

Enacted Identity Similarity: Enacted identity similarity is defined for each dyad symmetrically. For each dyad with members u and v, u's enacted identity similarity to v is defined as the semantic similarity between the word vector of "I" trained on u's messages and the word vector of "I" trained on v's messages. Semantic similarity is operationalized using cosine similarity, a commonly used distance metric for word embeddings.

$$\text{Enacted Identity Similarity}_{u,v} = cossim(w_{I,u}, w_{I,v}) \tag{1}$$

Observable Enacted Identity Similarity: Observable enacted identity similarity is defined for each dyad asymmetrically. For each dyad with members u and v, u's observable enacted identity similarity to v is defined as the semantic similarity between the word vector of "I" trained on u's messages observable to v and the word vector of "I" trained on all of v's messages. Semantic similarity is operationalized using cosine similarity, a commonly used distance metric for word embeddings.

$$\text{Observable Enacted Identity Similarity}_{u\to v} = cossim(w_{I,u\to v}, w_{I,v}) \tag{2}$$

Unobservable Enacted Identity Similarity: Unobservable enacted identity similarity is also defined for each dyad asymmetrically. For each dyad with members u and v, u's unobservable enacted identity similarity to v is defined as the semantic similarity between the word vector of "I" trained on u's messages unobservable to v and the word vector of "I" trained on all of v's messages.

$$\text{Unbservable Enacted Identity Similarity}_{u\not\to v} = cossim(w_{I,u\not\to v}, w_{I,v}) \tag{3}$$

Moderator

Privacy of Context: This variable measures the degree to which the context in which enacted identity similarity is observed is private. This moderator is used to test Hypothesis 3. I proxy privacy of context using a Slack-specific channel-level property. This property specifies whether messages sent in a channel can be browsed by all users of the organization (applies to public channels), or a pre-specified set of users (applies to private channels or direct messages). Specifically, I operationalize the degree to which Observable Enacted Identity $\text{Similarity}_{u \to v}$ is private as the proportion of word tokens u sent to v that took place in a private Slack channel or direct message.

Control Variables

I include demographics as controls, as they could simultaneously affect enacted identity similarity and tie strength. Demographic variables included are self-identified gender and ethnicity of both u and v. In addition, I also control for formal department and organizational tenure to account for unobserved heterogeneity across formal subunits and tenure that could affect both enacted identity similarity and tie strength. Tenure is logged given its right-skewed distribution. Finally, as enacted identity similarity is measured using language, I also control for the total number of tokens u and v have sent in the study period. The amount of text one sends could be related both to enacted identity similarity, as it dictates the volume of data available for training word embeddings, and tie strength, as those who send more texts could presumably have stronger relationships. This variable is also log-transformed.

In addition, several dimensions of homophily are included as controls at the dyadic level. I include gender and ethnicity similarity as indicators of ascriptive homophily, both of which could relate to enacted identity similarity and tie strength. Gender and ethnicity similarity are both dummy variables, taking on the value of one when there is an exact match between the gender (or ethnicity) of u and v. Furthermore, I also include a binary department similarity variable that takes on the value of one when u and v are in the same department. In this organization, department can be quite a granular variable—there are a total of 81 departments of various sizes. The inclusion of department similarity thus leads to a significant reduction in statistical power. However, including this variable is vital in accounting for induced homophily as much as possible. Individuals in the same department could develop stronger ties with one another given their shared focus (Feld 1981), and share similar identities that influenced their selection into certain departments and roles.

Finally, I control for similarity in language use. As identity is enacted and expressed through language, enacted identity similarity and linguistic similarity are inevitably interlaced. An individual who defines oneself through the lens of masculinity might be more likely to swear or curse. A new immigrant might use more sad and neurotic words due to a sense of displacement and alienation. Although I contend that linguistic similarity is reflective of enacted identity similarity, I try to isolate the effect of enacted identity similarity from that of linguistic similarity as the latter has been found to be a predictor of tie formation (Kovacs and Kleinbaum 2020). I compute a continuous measure of linguistic similarity using the widely established LIWC (Linguistic Inquiry Word Count) lexicon based on prior work on this topic (Kovacs and Kleinbaum 2020; Goldberg et al. 2016; Srivastava, Goldberg, Manian, and Potts 2018; Pennebaker, Francis, and Booth 2001).

Analytical Strategy

I estimate linear regression models of tie strength on the covariates described above. I measure enacted identity similarity using Slack data collected in the one-year period before the collection of survey responses on tie strength to try to account for reverse causality. While this structure can help address some potential endogeneity concerns, I emphasize that model estimates are not definitively causal. As the unit of analysis in these models is the dyad, observations are not independent, as the same individual can appear multiple times in different dyads. This can lead to correlated standard errors in the models. Thus, I correct this issue by simultaneously clustering the standard errors by both ego and alter.

RESULTS

Validation

Although my proposed measure of enacted identity and enacted identity similarity is informed by theoretical understandings of identity, validation of the measure is still needed. I do so in two complementary sets of analyses that together provide some validation to my linguistic measure of identity.

First, I demonstrate that the linguistic measure of enacted identity can predict one's gender and ethnic identity. As a measure of identity, the word embedding of "I" should be able to predict one's self-reported gender and ethnicity. Gender and ethnicity are both important components of one's identity; while gender identity and ethnic identity differ in salience across individuals, on average, they should at least partially inform the construction of one's identity. To test the relationship between my linguistic measure of enacted identity and gender and ethnic identity, I use random forest classification models. Random forest models are especially suitable for this task as they can model flexible, nonlinear relationships between predictors and outcome variables. To correct for data imbalance issues that can bias the results, I use the Balanced Random Forest algorithm.[6] Specifically, I binarized gender identity into Male versus Female and racial identity into White versus Non-White as the outcome variables, and used all 50 dimensions of the word vector of "I" as predictors. With cross-validation, the model predicts gender with an F1 score of 0.71, and race with an F1 score 0.65. [7] These results demonstrate that gender and race identity are encoded in the word embedding of "I" but not perfectly so, which is expected given the variance in the importance of race and gender identity to one's self-definitions across individuals.

In addition, at the dyadic level, I demonstrate that enacted identity similarity is associated with gender and ethnicity similarity. Using Spearman's correlation, I find that enacted identity similarity is correlated with both gender and ethnicity similarity at $\rho = 0.093$ ($p < 0.001$) and $\rho = 0.073$ ($p < 0.001$), respectively. To provide a baseline for comparison, I also selected the top ten most frequently used function words to test how similarity in the word vectors of these function words relates to gender and ethnicity similarity. Function words are selected as these words are as frequently used as first-person singular pronouns but are assumed to not carry any identity-relevant information. See Figure 2 for a correlation matrix between linguistic similarity of various words and sociodemographic similarity. As shown in this correlation matrix, similarity in "I" has a significant and positive association with gender and ethnic similarity. Similarity in function words, on the other hand, is weakly correlated with gender and ethnic similarity at best.

[FIGURE 2 ABOUT HERE]

Qualitative Analyses of the Word Embedding of "I"

In this section, I provide some qualitative understanding of the measure by unpacking the dimensions of the word embedding of "I." Dimensions of word embeddings are generally

[6] This algorithm was implemented in Python. See source package here.

[7] As a baseline, a random forest model predicting 1 randomly on a balanced dataset would generate a F1 score of 0.5

difficult to interpret; their positions are meaningful relative to one another but are not inherently meaningful in an absolute sense. Fortunately, recent advances have sought to overcome these issues and open up the black-box of word embedding models.

I employ a model called SPINE, or SParse Interpretable Neural Embeddings, (Subramanian, Pruthi, Jhamtani, Berg-Kirkpatrick, and Hovy 2018), which uses denoising k-sparse autoencoders to generate interpretable word embeddings from dense word embeddings like Word2Vec and GloVe. Using this method, one can investigate the top words from dimensions in which a given word is most active to get a sense of the different dimensions of meanings of this word. In Figure 3 below, I list some of the top dimensions in which "I" is most active and the top words associated with these dimensions. I labeled these dimensions with my interpretation of what each of these dimensions reflects for the readers' ease of understanding. Through this procedure, one can visually examine some of the dimensions encoded by the word embedding of "I."

An interesting dimension that emerged from this analysis is one that I am labeling as "Race and Racism." Words from this dimension seem to suggest that, in this organization, employees' racial identity is front and center in how they define themselves. But in addition, this dimension also encodes a salient antiracist identity. That is, embedded in how people think about and express themselves in this organization is both their racial identity as well as their stance towards racial issues. This dimension highlights why enacted identity homophily is an important construct. If one were to use existing categorical approaches to measuring homophily, antiracism is hardly a category one would come up with a priori. Furthermore, these words suggest that one's identity as an antiracist is enacted through discussing matters like systemic racism and anti-Asian attacks, highlighting the significance of identity enactment.

Other dimensions that this analysis highlights include a family-role dimension, a social-orientation dimension, and a place-based dimension. Some of these dimensions have been discussed in prior work as important dimensions of homophily (Lois and Becker 2023), while others are more ambiguous and amorphous. Collectively, these dimensions demonstrate the all-encompassing nature of identity and the abductive nature of this approach, as different forms of interpersonal similarity can all emerge in that of enacted identity similarity.

[FIGURE 3 ABOUT HERE]

Main Results

Table 1 reports descriptive statistics for key variables of interest. Table 2 reports Spearman's rank correlations among the main variables of interest. The correlation between observable and unobservable enacted identity similarity is quite weak, at $\rho = 0.162$. This suggests significant contextual variation in how individuals enact their identities, providing support for the interactional and contextual view of identity and highlighting the importance of studying enacted identity homophily.

[TABLE 1 ABOUT HERE]

[TABLE 2 ABOUT HERE]

The results of Hypothesis 1a are reported in Table 3. All models are logistic regression models, with coefficients exponentiated and standard errors clustered by both the ego (nominator) and the alter (nominee). In Model 1, I include only the key variable of interest, enacted identity similarity. In Model 2, I then add all individual-level control variables, including sociodemographic variables (gender and ethnicity), job characteristics (department and tenure), and linguistic control (logged number of tokens). In Model 3, I then add in ascriptive homophily, same gender and same ethnicity. Model 4 then adds in same department to control for structurally induced homophily. Finally, I add linguistic similarity in Model 5 to control for similarity in linguistic styles. Across all models, enacted identity similarity positively and significantly predicts tie nominations. A one-standard-deviation increase in enacted identity similarity leads to a 17% increase in the likelihood that ego will nominate alter as a meaningful interaction partner, holding constant similarities in gender, ethnicity, department, and language.

[TABLE 3 ABOUT HERE]

Hypothesis 1b is tested in Table 4 with linear regression models. Control variables are entered in the same order as Table 3. In Models 1 to 4, enacted identity similarity is positively and significantly related to tie strength. Therefore, H1b is only partially supported.

It is important to interpret the results of Model 5 carefully. In Model 5, when linguistic similarity is added as a control variable, the effect of enacted identity similarity on tie strength is still positive, but no longer statistically significant. This is perhaps unsurprising, as linguistic similarity is intertwined with enacted identity similarity. Theoretically, identity manifests in language. A person with a strong Christian identity, for instance, might refrain from vulgarities when they speak. The lack of profanities is then a linguistic signal of identity that individuals can match on when forming relationships. Empirically, in the current paper, enacted identity is measured via language. Thus, linguistic similarity also reflects similarities in response tendencies. In survey research, this is akin to similarities in baseline survey response preferences (e.g., a tendency to consistently respond neutrally to survey questions). Thus, controlling for linguistic similarity removes variation in both similarities in identity and response tendencies, an evidently stringent test.

[TABLE 4 ABOUT HERE]

Table 5 reports the models that test hypothesis 2. Model 1 includes the key variable of interest, observable enacted identity similarity, and all the control variables that are previously included in Model 5 of 4. Model 2 is set up in a similar fashion, replacing observable enacted identity similarity with unobservable enacted identity similarity as the key variable of interest. Finally, Model 3 includes both observable and unobservable enacted identity similarity alongside the controls. In all models, observable enacted identity similarity is positively and significantly associated with tie closeness, providing consistent support for H2. A one-standard-deviation-increase in observable enacted identity similarity is associated with a 0.219 standard-deviation increase in tie closeness ($p < 0.001$). The fact that observable enacted identity similarity predicts tie strength even when controlling for linguistic similarity, when enacted identity similarity does not, highlights that observability is a critical mechanism of homophily.

Across these models, there does not appear to be a statistically significant relationship between unobservable enacted identity similarity and tie closeness, lending credence to the idea that identity is contextual, and thus needs to be examined through the lens of observability. If identity were instead fixed as assumed in prior homophily research, we would anticipate that whether enacted identity similarity is observed or not should not affect its relationship with tie strength.

[TABLE 5 ABOUT HERE]

Finally, models in Table 6 test Hypothesis 3 and examine how the effect of observable enacted identity similarity varies as a function of the degree to which the context of one's identity enactment is private. These tables demonstrate that the effect of observable enacted identity similarity is amplified by the privacy of the context, lending support to Hypothesis 3. The interaction between observable enacted identity similarity and privacy of context, operationalized as the proportion of tokens sent in a private channel, is positive and significant across all four models. A one-standard-deviation increase in the proportion of private tokens increases the effect of observable enacted identity similarity by 0.089 standard deviations ($p < 0.001$).

[TABLE 6 ABOUT HERE]

Robustness Checks

I assess the robustness of my findings by replicating my findings using a different measure of tie strength. While tie closeness has been shown to be the best measure of tie strength, tie frequency and tie duration are also commonly used measures of tie strength (Friedkin 1993; Reagans 2005). In particular, these measures reflect time spent in relationship, and thus capture different aspects of tie strength than does tie closeness. Thus, I rerun all the analyses in Tables 4, 5 and 6 and use the number of Slack direct messages ego sent to alter during the study period as the dependent variable. [8]This measure is log-transformed given the right-skewed distribution of the variable.

As shown in Tables 7, 8, and 9, my results are robust to this alternate tie strength specification. The patterns in these tables are quite similar to those in the main analyses. This is remarkable given that these two tie strength measures are based on completely different forums of data (self-reported survey versus archival data) and types of response (Likert scale ratings versus count of naturally exchanged messages). These measures are correlated at $\rho = 0.52$ ($p < 0.001$), which is neither trivially small nor significantly collinear, pointing to substantial divergence in what these two measures of tie strength capture. Taken together, these tables demonstrate the consistency of my findings and provide additional support for Hypotheses 1, 2, and 3.

[TABLE 7 ABOUT HERE]

[8] The number of direct messages is used as it is difficult to infer to whom messages are sent in public and private channels. In analyses carried out in a different study, I found that direct message ties most strongly correspond to subjectively nominated ties in the network survey, lending support to the idea that direct message ties are most meaningful and substantive.

[TABLE 8 ABOUT HERE]

[TABLE 9 ABOUT HERE]

DISCUSSION

Synthesizing extant research on homophily and identity, this paper advances a contextual and dynamic theory of homophily. I first formalize an argument of how similarity in enacted identity predicts tie nomination and tie strength. Integrating the contention that identity is contextually enacted, I argue that enacted identity similarity should also vary by context, and only observable enacted identity similarity should strengthen relationships. Lastly, I further hypothesize that the impact of observable enacted identity similarity on tie strength is moderated by the extent to which it is enacted privately.

Adapting existing natural language processing techniques, I measure enacted identity similarity by comparing the semantic meanings people attach to their first-person singular pronoun “I.” I apply this measure to digital trace data, collected from a midsized American organization. Combined with responses from a network roster survey, I find support for my theory.

Contributions

This paper makes several theoretical and methodological contributions. By highlighting the enacted nature of identity, disentangling it into observable and unobservable components, and showing how privacy of context moderates homophily, I demonstrate the dynamic and contextual nature of homophily. Existing work on homophily treats similarity as a stable attribute: two individuals are either similar on a specific dimension or dissimilar. The current paper, however, illustrates that intricacies and complexities in similarity. Similarity is a dynamic interpersonal process, a tango between the enactment of self and the observation of others. In social interactions, individuals enact certain aspects of their identity in front of others, who in response, are processing these identity signals and evaluating the degree to which they are performative or authentic.

Relatedly, by viewing similarity through the lens of individuals under study, I demonstrate that observability is an important constraint to how homophily operates. I highlight that similarity needs to be observable for individuals to assess how similar others are to themselves. I demonstrate that observable enacted identity similarity predicts tie strength, while unobservable similarity does not. In doing so, I add a previously taken-for-granted and underexplored mechanism to homophily research, and address the call for more research on unpacking "how choice homophily operates in research" (DiMaggio and Garip 2012, p. 111).

Finally, this paper also makes a distinct methodological contribution. Building on Yang et al. (2023), I argue that enacted identity manifests in and can be measured via language. I use word embedding models to measure enacted identity through semantic meanings associated with one's first-person singular pronoun "I." Extending this measure dyadically, I operationalize enacted identity similarity as the degree of overlap in semantic meanings of "I." Compared to self-reports, which are commonly used in existing work, this approach of measuring identity more accurately reflects the contextual nature of identity and better captures the signals individuals rely on when drawing inferences of others' identity.

Future Directions

An immediate extension of this work is to investigate how convergence and divergence in identity enactment shape relational outcomes. This paper has demonstrated that observable enacted identity similarity shapes tie strength. Given the low correlation between observable and unobservable enacted identity similarity, one can explore how the difference between these two forms of similarity might relate to tie strength. For example, when observable and unobservable identity enactments of others diverge, will identity enactments lose their credibility and enacted identity similarity lose its impact? In addition, the degree of convergence and divergence in observable enacted identity can also vary between people. The degree of observable variance in how one expresses oneself could possibly affect the strength of relationships one forms.

Extending it beyond the dyadic level, the concept of enacted identity similarity may have interesting implications for group-level outcomes. Identity convergence among organizational members could make it easier for individuals to collaborate and coordinate. It may also adversely impact the organization's innovation capabilities. Lastly, tracking identities of group members could be an illuminating analysis in unpacking how organizations construct, manage, and uphold their own identities.

Conclusion

The breadth of the construct of homophily is astounding. It underlies a variety of structural phenomena, is documented along numerous dimensions, and characterizes relationships between entities as diverse as bees, chemical compounds, organizations, and humans (McPherson, Smith-Lovin, and Rawlings 2021; McPherson et al. 2001). This paper seeks to add depth to homophily by describing the complexities in how it operates. Through integrating research on homophily and identity, this paper shows that, insofar as identities are enacted and performed, homophily also occurs contextually and interactionally. Deviating from the portrayal of homophily as an objective description of persons, I reimagine homophily as a dynamic, intersubjective process between persons.

References

Agneessens, Filip and Giuseppe (Joe) Labianca. 2022. "Collecting survey-based social network information in work organizations." *Social Networks* 68:31–47.

Ashokkumar, Ashwini and James W Pennebaker. 2022. "Tracking Group Identity through Natural Language within Groups." *PNAS Nexus* 1:1–9.

Bailey, Joe. 2000. "Some Meanings of 'the Private' in Sociological Thought." *Sociology* 34:381–401. Publisher: Cambridge University Press.

Brewer, John D. 2005. "The Public and Private in C.Wright Mills's Life and Work." *Sociology* 39:661–677.

Brighenti, Andrea. 2007. "Visibility: A category for the social sciences." *Current sociology* 55:323–342.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language Models are Few-shot Learners." *Advances in Neural Information Processing Systems* 33:1877–1901.

Byrne, D. 1961. "Interpersonal attraction and attitude similarity." *The Journal of Abnormal and Social Psychology* 62:713–715.

Callero, Peter L. 2003. "The Sociology of the Self." *Annual Review of Sociology* 29:115–133.

Cerulo, Karen A. 1997. "Identity Construction: New Issues, New Directions." *Annual Review of Sociology* 23:385–409.

Cooley, Charles Horton. 1902. *Human Nature and the Social Order*. C. Scribner's Sons.

Deaux, Kay. 1993. "Reconstructing Social Identity." *Personality and Social Psychology Bulletin* 19:4–12. Publisher: SAGE Publications Inc.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805 [cs]* arXiv: 1810.04805.

DiMaggio, Paul and Filiz Garip. 2012. “Network effects and social inequality.” *Annual review of sociology* 38:93–118.

Dingwall, Nicholas and Christopher Potts. 2018. “Mittens: An Extension of GloVe for Learning Domain-Specialized Representations.” *arXiv:1803.09901 [cs]* arXiv: 1803.09901.

Donaldson, Stewart I and Elisa J Grant-Vallone. 2002. “Understanding self-report bias in organizational behavior research.” *Journal of business and Psychology* 17:245–260.

Ertug, Gokhan, Julia Brennecke, Balázs Kovács, and Tengjian Zou. 2022. “What Does Homophily Do? A Review of the Consequences of Homophily.” *Academy of Management Annals* 16:38–69. Publisher: Academy of Management.

Feld, Scott L. 1981. “The Focused Organization of Social Ties.” *American Journal of Sociology* 86:1015–1035.

Fine, Gary Alan. 2012. “Group Culture and the Interaction Order: Local Sociology on the Meso-Level.” *Annual Review of Sociology* 38:159–179.

Friedkin, Noah E. 1993. “Structural Bases of Interpersonal Influence in Groups: A Longitudinal Case Study.” *American Sociological Review* 58:861–872. Publisher: [American Sociological Association, Sage Publications, Inc.].

Goffman, Erving. 1959. *The presentation of self in everyday life*. The presentation of self in everyday life. Oxford, England: Doubleday.

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* 81:1190–1222. arXiv: 0903.3277 ISBN: 0029-6562 1538-9847.

Granovetter, Mark S. 1973. “The Strength of Weak Ties.” *American Journal of Sociology* 78:1360–1380.

Ibarra, Herminia. 1992. “Homophily and Differential Returns: Sex Differences in Network Structure and Access in an Advertising Firm.” *Administrative Science Quarterly* 37:422–447.

Ingram, Paul. 2023. “Identity multiplicity and the formation of professional network ties.” *Academy of Management Journal* 66:720–743.

Ivanič, Roz. 1998. *Writing and identity*, volume 10. John Benjamins Amsterdam.

Jaworski, Gary Dean. 1990. “Robert K. Merton’s Extension of Simmel’s Ubersehbar.” *Sociological Theory* 8:99–105.

Joyner, Kara and Grace Kao. 2000. “School racial composition and adolescent racial homophily.” *Social science quarterly* pp. 810–825.

Kovacs, Balazs and Adam M. Kleinbaum. 2020. “Language-Style Similarity and Social Networks.” *Psychological Science* 31:202–213. Publisher: SAGE Publications Inc.

Lahire, Bernard. 2011. *The plural actor*. Polity.

Lawler, Steph. 2015. *Identity: Sociological Perspectives*. Malden, MA: Polity Press.

Lawrence, Barbara S. and Neha Parikh Shah. 2020. “Homophily: Measures and Meaning.” *Academy of Management Annals* 14:513–597. Publisher: Academy of Management.

Leszczensky, Lars and Sebastian Pink. 2019. “What Drives Ethnic Homophily? A Relational Approach on How Ethnic Identification Moderates Preferences for Same-Ethnic Friends.” *American Sociological Review* 84:394–419.

Lincoln, James R. and Jon Miller. 1979. “Work and Friendship Ties in Organizations: A Comparative Analysis of Relation Networks.” *Administrative Science Quarterly* 24:181–199.

Lois, Daniel and Oliver Arránz Becker. 2023. “Parental status homogeneity in social networks.” *Demographic Research* 48:19–42.

Marsden, Peter V. 1988. “Homogeneity in confiding relations.” *Social networks* 10:57–76.

Marsden, Peter V. and Karen E. Campbell. 1984. “Measuring Tie Strength.” *Social Forces* 63:482–501.

McPherson, J. Miller and Lynn Smith-Lovin. 1987. “Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups.” *American Sociological Review* 52:370–379.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology; Palo Alto* 27:415–444.

McPherson, Miller, Lynn Smith-Lovin, and Craig Rawlings. 2021. "The Enormous Flock of Homophily Researchers: Assessing and Promoting a Research Agenda." In *Personal Networks*, edited by Mario L. Small and Brea L. Perry, pp. 459–470. Cambridge University Press, 1 edition.

Mead, George Herbert (ed.). 1934. *Mind, Self, and Society: From the Standpoint of a Social Behaviorist*. Chicago, IL: University of Chicago Press.

Mehra, Ajay, Martin Kilduff, and Daniel J Brass. 1998. "At the margins: A distinctiveness approach to the social identity and social networks of underrepresented groups." *Academy of Management Journal* 41:441–452.

Merton, Robert K. 1968. "Continuities in the Theory of Reference Groups and Social Structure." In *Social Theory and Social Structure*, pp. 344–449. New York, NY: Free Press.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." *Advances in neural information processing systems* 26:3111–3119. tex.ids: mikolov_distributed_2013-1.

Mollica, Kelly A, Barbara Gray, and Linda K Trevino. 2003. "Racial homophily and its persistence in newcomers' social networks." *Organization Science* 14:123–136.

Owens, Timothy J., Dawn T. Robinson, and Lynn Smith-Lovin. 2010. "Three Faces of Identity." *Annual Review of Sociology* 36:477–499. tex.ids= owens_three_2010-1.

Paulhus, Delroy L, Simine Vazire, et al. 2007. "The self-report method." *Handbook of research methods in personality psychology* 1:224–239.

Pennebaker, James W. 2011. *The Secret Life of Pronouns: What Our Words Say about Us*. New York, NY: Bloomsbury Press.

Pennebaker, James W, Martha E Francis, and Roger J Booth. 2001. "Linguistic inquiry and word count: LIWC 2001." *Mahway: Lawrence Erlbaum Associates* 71:2001.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. “Glove: Global Vectors for Word Representation.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics.

Ramarajan, Lakshmi. 2014. “Past, Present and Future Research on Multiple Identities: Toward an Intrapersonal Network Approach.” *Academy of Management Annals* 8:589–659.

Reagans, Ray. 2005. “Preferences, Identity, and Competition: Predicting Tie Strength from Demographic Data.” *Management Science* 51:1374–1383. Publisher: INFORMS.

Reagans, Ray. 2011. “Close Encounters: Analyzing How Social Similarity and Propinquity Contribute to Strong Network Connections.” *Organization Science* 22:835–849. Publisher: INFORMS.

Reagans, Ray, Linda Argote, and Daria Brooks. 2005. “Individual Experience and Experience Working Together: Predicting Learning Rates from Knowing Who Knows What and Knowing How to Work Together.” *Management Science* 51:869–881.

Rhee, Mooweon, Daegyu Yang, and Taeyoung Yoo. 2013. “National culture and friendship homophily in the multinational workplace.” *Asian Business & Management* 12:299–320.

Shrum, Wesley, Neil H Cheek Jr, and Saundra MacD. 1988. “Friendship in school: Gender and racial homophily.” *Sociology of Education* pp. 227–239.

Simmel, Georg. 1950. *The Sociology of Georg Simmel*. Simon and Schuster.

Slater, Don. 1998. “Public/Private.” In *Core Sociological Dichotomies*. London: SAGE Publications Ltd.

Srivastava, Sameer B., Amir Goldberg, V. Govind Manian, and Christopher Potts. 2018. “Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations.” *Management Science* 64:1348–1364.

Stadel, Marie and Gert Stulp. 2022. “Balancing bias and burden in personal network studies.” *Social Networks* 70:16–24.

Stehlé, Juliette, François Charbonnier, Tristan Picard, Ciro Cattuto, and Alain Barrat. 2013. "Gender homophily from spatial behavior in a primary school: A sociometric study." *Social Networks* 35:604–613.

Stets, Jan E. and Peter J. Burke. 2000. "Identity Theory and Social Identity Theory." *Social Psychology Quarterly* 63:224–237.

Stets, Jan E. and Peter J. Burke. 2003. "A Sociological Approach to Self and Identity Thoughts on Social Structure." In *Handbook of Self and Identity*, pp. 128–152. The Guilford Press.

Stryker, Sheldon and Peter J. Burke. 2000. "The Past, Present, and Future of an Identity Theory." *Social Psychology Quarterly* 63:284–297.

Subramanian, Anant, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. "SPINE: SParse Interpretable Neural Embeddings." *Proceedings of the AAAI Conference on Artificial Intelligence* 32.

Tajfel, Henri and John C. Turner. 1979. "An integrative theory of intergroup conflict." In *The social psychology of intergroup relations*, pp. 33–47. Monterey, CA: Brooks/Cole.

Tajfel, Henri and John C. Turner. 1986. "The Social Identity Theory of Intergroup Behavior." In *Psychology of Intergroup Relations*, Political psychology: Key readings, pp. 7–24. Nelson-Hall Publishers.

Tang, Ramona and Suganthi John. 1999. "The 'I' in identity: Exploring writer identity in student academic writing through the first person pronoun." *English for specific purposes* 18:S23–S39.

Warren, Donald I. 1968. "Power, visibility, and conformity in formal organizations." *American Sociological Review* pp. 951–970.

Weintraub, Jeff and Krishan Kumar. 1997. *Public and private in thought and practice: Perspectives on a grand dichotomy*. University of Chicago Press.

Yang, Lara, Amir Goldberg, and Sameer B. Srivastava. 2023. "Locally Ensconced and Globally Integrated: How Positions in Network Structure Relate to a Language-Based Model of Group Identification." *SocArXiv* .

Table 1: Descriptive Statistics

	Mean	SD	Min	Max
Enacted Identity Similarity	0.95	0.02	0.83	1.00
Observable Enacted Identity Similarity	0.94	0.02	0.88	1.00
Unobservable Enacted Identity Similarity	0.96	0.02	0.87	1.00
Proportion of Private Tokens	0.61	0.36	0.00	1.00
Linguistic Similarity	2.59	0.27	1.30	3.72
Tenure (Days)	2456.67	1703.68	40.00	11899.00
Number of Tokens	89133.50	86989.27	494.00	601978.00

Table 2: Correlation Matrix

	EIS	Observable EIS	Unobservable EIS	Proportion of Private Tokens	Linguistic Similarity	Tenure (Days)	Number of Tokens
Enacted Identity Similarity (EIS)	-						
Observable Enacted Identity Similarity	0.179^{***}	-					
Unobservable Enacted Identity Similarity	0.925^{***}	0.163^{***}	-				
Proportion of Private Tokens	0.055^{***}	0.086^{***}	0.036^{*}	-			
Linguistic Similarity	0.423^{***}	0.140^{***}	0.406^{***}	0.095^{***}	-		
Tenure (Days)	0.009^{***}	-0.120^{***}	0.028^{+}	0.102^{***}	0.086^{***}	-	
Number of Tokens	0.095^{***}	-0.262^{***}	0.182^{***}	0.125^{***}	0.269^{***}	0.342^{***}	-

$^{+}p<0.1$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Variable names abbreviated for readability

Table 3: Tie Nominations on Enacted Identity Similarity

	Model 1	Model 2	Model 3	Model 4	Model 5
Enacted Identity Similarity	1.530***	1.463***	1.455***	1.314***	1.177***
	(0.015)	(0.016)	(0.016)	(0.016)	(0.015)
Same Gender			1.066**	1.040+	1.025
			(0.023)	(0.024)	(0.024)
Same Ethnicity			1.330***	1.363***	1.340***
			(0.027)	(0.031)	(0.030)
Same Department				87.938***	84.626***
				(2.251)	(2.176)
Linguistic Similarity					1.392***
					(0.017)
Gender	-	Yes	Yes	Yes	Yes
Ethnicity	-	Yes	Yes	Yes	Yes
Department	-	Yes	Yes	Yes	Yes
Tenure	-	Yes	Yes	Yes	Yes
Number of Tokens	-	Yes	Yes	Yes	Yes
Num.Obs.	700074	700074	700074	700074	700074

Tenure and number of tokens are logged

Standard errors clustered by dyad

$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 4: Tie Closeness on Enacted Identity Similarity

	Model 1	Model 2	Model 3	Model 4	Model 5
Enacted Identity Similarity	0.092***	0.111***	0.107***	0.069***	0.013
	(0.024)	(0.020)	(0.020)	(0.019)	(0.020)
Same Gender			0.045	0.033	0.030
			(0.041)	(0.035)	(0.035)
Same Ethnicity			0.090**	0.106***	0.096**
			(0.033)	(0.031)	(0.030)
Same Department				0.802***	0.785***
				(0.045)	(0.045)
Linguistic Similarity					0.164***
					(0.020)
Gender	-	Yes	Yes	Yes	Yes
Ethnicity	-	Yes	Yes	Yes	Yes
Department	-	Yes	Yes	Yes	Yes
Tenure	-	Yes	Yes	Yes	Yes
Number of Tokens	-	Yes	Yes	Yes	Yes
Num.Obs.	4486	4486	4486	4486	4486

Tenure and number of tokens are logged

Standard errors clustered by dyad

$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 5: Tie Closeness on Observable and Unobservable Enacted Identity Similarity

	Model 1	Model 2	Model 3
Observable Enacted Identity Similarity	0.215***		0.219***
	(0.026)		(0.026)
Unobservable Enacted Identity Similarity		0.002	-0.029
		(0.024)	(0.024)
Same Gender	0.057	0.062	0.056
	(0.051)	(0.051)	(0.051)
Same Ethnicity	0.080*	0.094*	0.083*
	(0.038)	(0.038)	(0.038)
Same Department	0.311***	0.409***	0.310***
	(0.055)	(0.055)	(0.055)
Linguistic Similarity	0.128***	0.173***	0.137***
	(0.024)	(0.026)	(0.025)
Gender	Yes	Yes	Yes
Ethnicity	Yes	Yes	Yes
Department	Yes	Yes	Yes
Tenure	Yes	Yes	Yes
Number of Tokens	Yes	Yes	Yes
Num.Obs.	2602	2602	2602

Tenure and number of tokens are logged
Standard errors clustered by dyad
$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 6: Tie Closeness on Observable and Unobservable Enacted Identity Similarity by Private Communications

	Model 1	Model 2	Model 3	Model 4	Model 5
Observable Enacted Identity Similarity	0.184***	0.182***	0.180***	0.165***	0.148***
	(0.025)	(0.026)	(0.026)	(0.026)	(0.026)
Proportion of Private Tokens	0.221***	0.309***	0.310***	0.292***	0.284***
	(0.029)	(0.030)	(0.030)	(0.029)	(0.030)
Observable Similarity X Proportion of Private Tokens	0.056*	0.095***	0.093***	0.095***	0.089***
	(0.024)	(0.023)	(0.023)	(0.022)	(0.022)
Same Gender			0.055	0.054	0.052
			(0.052)	(0.051)	(0.051)
Same Ethnicity			0.088*	0.091*	0.086*
			(0.037)	(0.037)	(0.037)
Same Department				0.226***	0.231***
				(0.053)	(0.054)
Linguistic Similarity					0.108***
					(0.024)
Gender	-	Yes	Yes	Yes	Yes
Ethnicity	-	Yes	Yes	Yes	Yes
Department	-	Yes	Yes	Yes	Yes
Tenure	-	Yes	Yes	Yes	Yes
Number of Tokens	-	Yes	Yes	Yes	Yes
Num.Obs.	2602	2602	2602	2602	2602

Tenure and number of tokens are logged

Standard errors clustered by dyad

$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 7: Tie Frequency on Enacted Identity Similarity

	Model 1	Model 2	Model 3	Model 4	Model 5
Enacted Identity Similarity	0.152***	0.144***	0.143***	0.104***	0.068***
	(0.008)	(0.006)	(0.006)	(0.005)	(0.005)
Same Gender			0.033**	0.018+	0.014
			(0.012)	(0.010)	(0.010)
Same Ethnicity			0.050***	0.063***	0.057***
			(0.009)	(0.008)	(0.008)
Same Department				1.030***	1.026***
				(0.020)	(0.020)
Linguistic Similarity					0.107***
					(0.006)
Gender	-	Yes	Yes	Yes	Yes
Ethnicity	-	Yes	Yes	Yes	Yes
Department	-	Yes	Yes	Yes	Yes
Tenure	-	Yes	Yes	Yes	Yes
Number of Tokens	-	Yes	Yes	Yes	Yes
Num.Obs.	115857	115857	115857	115857	115857

Tenure and number of tokens are logged

Standard errors clustered by dyad

$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 8: Tie Frequency on Observable and Unobservable Enacted Identity Similarity

	Model 1	Model 2	Model 3
Observable Enacted Identity Similarity	0.456***		0.459***
	(0.025)		(0.025)
Unobservable Enacted Identity Similarity		0.047+	-0.022
		(0.026)	(0.024)
Same Gender	0.090+	0.100+	0.090+
	(0.051)	(0.053)	(0.051)
Same Ethnicity	-0.022	0.001	-0.020
	(0.038)	(0.042)	(0.038)
Same Department	0.438***	0.640***	0.438***
	(0.058)	(0.063)	(0.057)
Linguistic Similarity	0.129***	0.205***	0.136***
	(0.023)	(0.028)	(0.025)
Gender	Yes	Yes	Yes
Ethnicity	Yes	Yes	Yes
Department	Yes	Yes	Yes
Tenure	Yes	Yes	Yes
Number of Tokens	Yes	Yes	Yes
Num.Obs.	2538	2538	2538

Tenure and number of tokens are logged

Standard errors clustered by dyad

$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 9: Tie Frequency on Observable and Unobservable Enacted Identity Similarity by Private Communications

	Model 1	Model 2	Model 3	Model 4	Model 5
Observable Enacted Identity Similarity	0.295***	0.375***	0.375***	0.357***	0.343***
	(0.027)	(0.022)	(0.022)	(0.022)	(0.022)
Proportion of Private Tokens	0.511***	0.576***	0.576***	0.556***	0.547***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
Observable Similarity X Proportion of Private Tokens	0.147***	0.105***	0.104***	0.106***	0.101***
	(0.026)	(0.021)	(0.021)	(0.021)	(0.021)
Linguistic Similarity					0.093***
					(0.021)
Same Gender			0.089+	0.090+	0.089+
			(0.050)	(0.049)	(0.049)
Same Ethnicity			-0.005	-0.001	-0.006
			(0.034)	(0.033)	(0.033)
Same Department				0.287***	0.291***
				(0.052)	(0.052)
Gender	-	Yes	Yes	Yes	Yes
Ethnicity	-	Yes	Yes	Yes	Yes
Department	-	Yes	Yes	Yes	Yes
Tenure	-	Yes	Yes	Yes	Yes
Number of Tokens	-	Yes	Yes	Yes	Yes
Num.Obs.	2538	2538	2538	2538	2538

Tenure and number of tokens are logged

Standard errors clustered by dyad

$^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

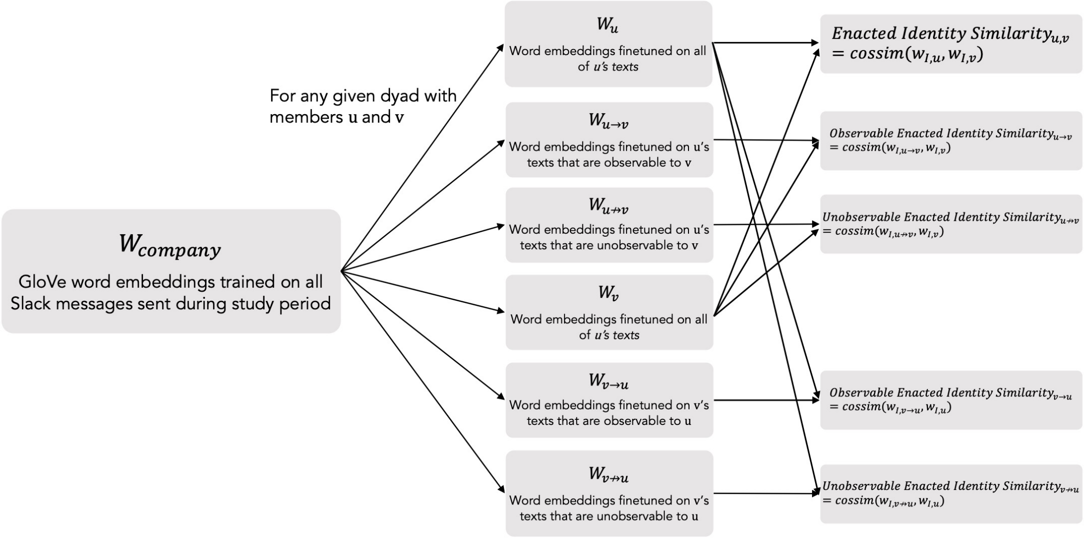
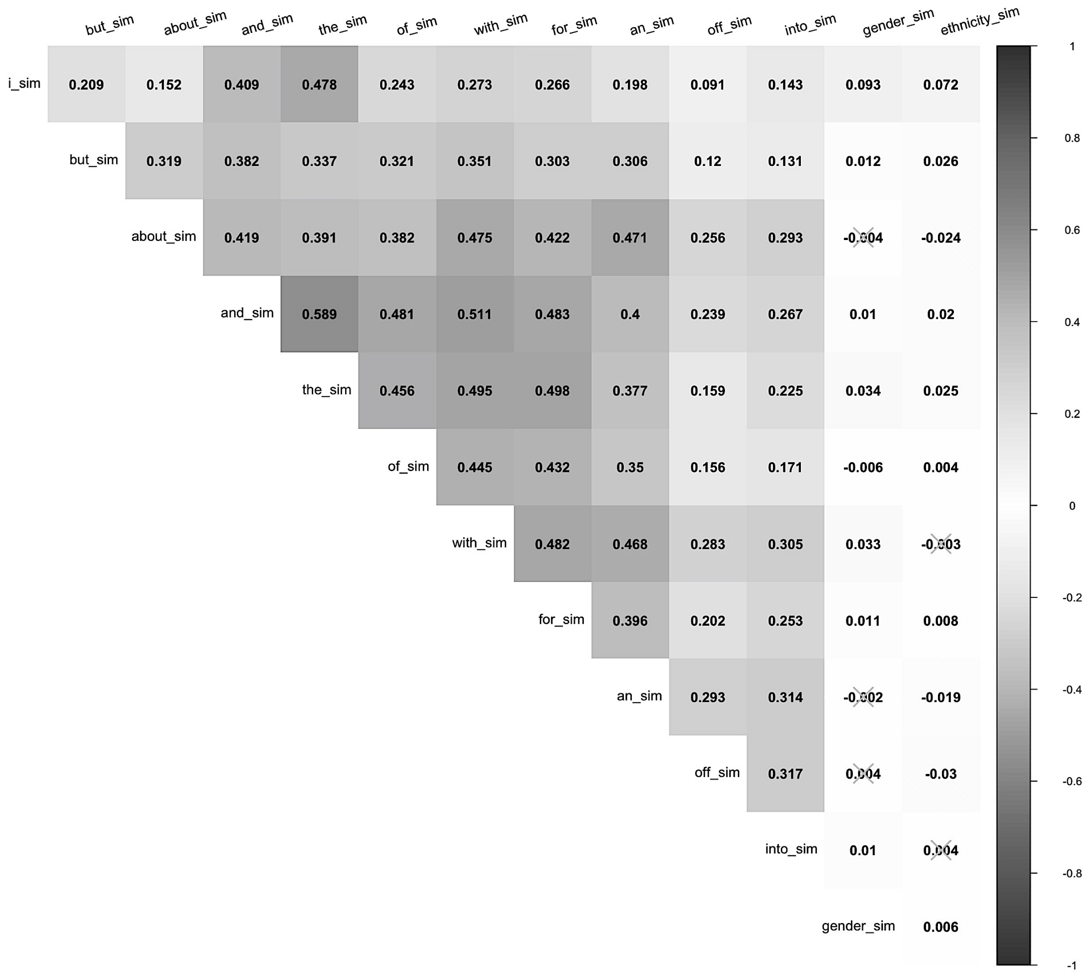
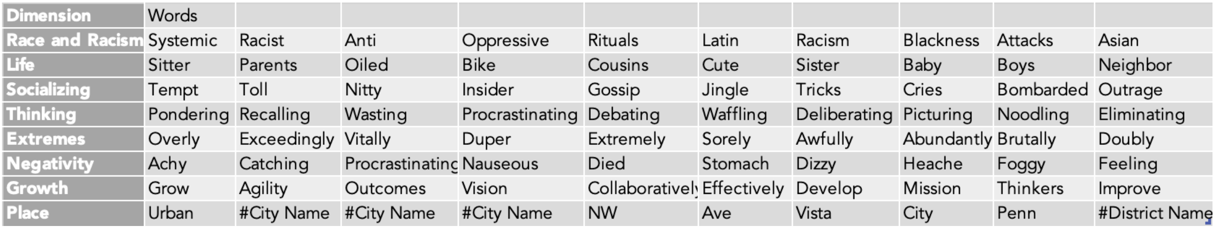

$W_{company}$
GloVe word embeddings trained on all Slack messages sent during study period
For any given dyad with members u and v
W_u
Word embeddings finetuned on all of *u's texts*
$W_{u \to v}$
Word embeddings finetuned on u's texts that are observable to v
$W_{u \nrightarrow v}$
Word embeddings finetuned on u's texts that are unobservable to v
W_v
Word embeddings finetuned on all of *u's texts*
$W_{v \to u}$
Word embeddings finetuned on v's texts that are observable to u
$W_{v \nrightarrow u}$
Word embeddings finetuned on v's texts that are unobservable to u
$Enacted\ Identity\ Similarity_{u,v} = cossim(w_{I,u}, w_{I,v})$
$Observable\ Enacted\ Identity\ Similarity_{u \to v} = cossim(w_{I,u \to v}, w_{I,v})$
$Unobservable\ Enacted\ Identity\ Similarity_{u \nrightarrow v} = cossim(w_{I,u \nrightarrow v}, w_{I,v})$
$Observable\ Enacted\ Identity\ Similarity_{v \to u} = cossim(w_{I,v \to u}, w_{I,u})$
$Unobservable\ Enacted\ Identity\ Similarity_{v \nrightarrow u} = cossim(w_{I,v \nrightarrow u}, w_{I,u})$

Figure 1: Measuring enacted identity similarity.
This figure provides a visual overview of the steps I took to compute enacted identity similarity, as well as observable and unobservable enacted identity similarity for a given dyad.



	but_sim	about_sim	and_sim	the_sim	of_sim	with_sim	for_sim	an_sim	off_sim	into_sim	gender_sim	ethnicity_sim
i_sim	0.209	0.152	0.409	0.478	0.243	0.273	0.266	0.198	0.091	0.143	0.093	0.072
but_sim		0.319	0.382	0.337	0.321	0.351	0.303	0.306	0.12	0.131	0.012	0.026
about_sim			0.419	0.391	0.382	0.475	0.422	0.471	0.256	0.293	~~-0.004~~	-0.024
and_sim				0.589	0.481	0.511	0.483	0.4	0.239	0.267	0.01	0.02
the_sim					0.456	0.495	0.498	0.377	0.159	0.225	0.034	0.025
of_sim						0.445	0.432	0.35	0.156	0.171	-0.006	0.004
with_sim							0.482	0.468	0.283	0.305	0.033	~~-0.003~~
for_sim								0.396	0.202	0.253	0.011	0.008
an_sim									0.293	0.314	~~-0.002~~	-0.019
off_sim										0.317	~~0.004~~	-0.03
into_sim											0.01	~~0.004~~
gender_sim												0.006

Figure 2: Correlation matrix of linguistic similarity and sociodemographic similarity. This figure shows the correlation matrix between similarity in "I," similarity in various function words, and similarity in gender and ethnicity. Correlation coefficients are spearman's ρ with $p = 0.001$. All coefficients that are insignificant are crossed out. This matrix demonstrates that similarity in "I" (enacted identity similarity) is positively and significantly associated with gender and ethnicity similarity, while similarities in function words are weakly correlated with gender and ethnicity similarity across the board.



Dimension	Words									
Race and Racism	Systemic	Racist	Anti	Oppressive	Rituals	Latin	Racism	Blackness	Attacks	Asian
Life	Sitter	Parents	Oiled	Bike	Cousins	Cute	Sister	Baby	Boys	Neighbor
Socializing	Tempt	Toll	Nitty	Insider	Gossip	Jingle	Tricks	Cries	Bombarded	Outrage
Thinking	Pondering	Recalling	Wasting	Procrastinating	Debating	Waffling	Deliberating	Picturing	Noodling	Eliminating
Extremes	Overly	Exceedingly	Vitally	Duper	Extremely	Sorely	Awfully	Abundantly	Brutally	Doubly
Negativity	Achy	Catching	Procrastinating	Nauseous	Died	Stomach	Dizzy	Heache	Foggy	Feeling
Growth	Grow	Agility	Outcomes	Vision	Collaborativel	Effectively	Develop	Mission	Thinkers	Improve
Place	Urban	#City Name	#City Name	#City Name	NW	Ave	Vista	City	Penn	#District Name

Figure 3: Top Dimensions of Word Embedding of “I”
This figure shows the top words associated with some of the dimensions in which “I” are most active. These dimensions are labeled by my own interpretation of what the dimension captures and represents.