

Language-Style Similarity and Social Networks

Balazs Kovacs¹  and Adam M. Kleinbaum²

¹School of Management, Yale University, and ²Tuck School of Business, Dartmouth College

Psychological Science
2020, Vol. 31(2) 202–213
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797619894557
www.psychologicalscience.org/PS



Abstract

This research demonstrates that linguistic similarity predicts network-tie formation and that friends exhibit linguistic convergence over time. In Study 1, we analyzed the linguistic styles and the emerging social network of a complete cohort of 285 students. In Study 2, we analyzed a large-scale data set of online reviews. In both studies, we collected data in two waves to examine changes in both social networks and linguistic styles. Using the Linguistic Inquiry and Word Count (LIWC) framework, we analyzed the text of students' essays and of 1.7 million reviews by 159,651 Yelp reviewers. Consistent with our theory, results showed that similarity in linguistic style corresponded to a higher likelihood of friendship formation and persistence and that friendship ties, in turn, corresponded to a convergence in linguistic style. We discuss the implications of the coevolution of linguistic styles and social networks, which contribute to the formation of relational echo chambers.

Keywords

linguistic style, social networks, big data, polarization

Received 7/17/18; Revision accepted 10/27/19

How do social networks form and evolve? A long line of social science research has documented that a key driver of social interaction is the principle of *homophily*: “birds of a feather flock together” (Mark, 1998; McPherson, Smith-Lovin, & Cook, 2001). Homophily has been demonstrated to exist along myriad dimensions, including race, gender, religion, nationality, and personality, and to act in such disparate social relations as friendship, marriage, hiring, entrepreneurship, business collaboration, and online interaction (Fowler & Christakis, 2008; Ibarra, 1992; Wimmer & Lewis, 2010).

Although most of the dimensions in which homophily occurs are readily apparent, such as gender or race, or are easily discovered, such as religion or nationality, we aimed to explore homophily in a different dimension: linguistic style. Recent research has argued that subtle cues in linguistic style can reveal a variety of underlying personality traits (Pennebaker, 2011). Some research has demonstrated the existence of homophily along specific personality traits, such as extraversion, conscientiousness, or agreeableness (Feiler & Kleinbaum, 2015; Noë, Whitaker, & Allen, 2016; Youyou, Stillwell, Schwartz, & Kosinski, 2017), and even homophily along

neural activity (Parkinson, Kleinbaum, & Wheatley, 2018). Our aim here was broader: to show that similar linguistic styles provide cues about underlying interpersonal similarity that will facilitate friendship formation. Beyond their indirect role in revealing underlying similarities in personality, linguistic similarities may also play a direct role in facilitating tie formation and persistence, perhaps allowing people with similar linguistic styles to communicate more easily. Indeed, sociolinguists studying cognitive style have long conjectured that this is the case (Eckert, 2012; Nguyen, Doğruöz, Rosé, & de Jong, 2016).

Of course, as prior work has shown (Aral, Muchnik, & Sundararajan, 2009), correlation does not imply causation, and we suggest that the causal arrow points in

Corresponding Authors:

Balazs Kovacs, Yale University, School of Management, 165 Whitney Ave., New Haven, CT 06511
E-mail: Balazs.Kovacs@yale.edu

Adam M. Kleinbaum, Dartmouth College, Tuck School of Business, 100 Tuck Hall, Hanover, NH 03755
E-mail: Adam.M.Kleinbaum@tuck.dartmouth.edu

the other direction as well: In addition to linguistic similarity driving tie formation, friendship ties will also induce increases in linguistic similarity. An individual's linguistic style may change fluidly over time and evolve in response to that person's interaction partners. Indeed, a long history of research in psychology shows that people are motivated to fit into their social worlds and, as a result, tend to mirror the behaviors in general—and the linguistic style in particular—of those around them (Chartrand & Bargh, 1999; Gonzales, Hancock, & Pennebaker, 2010; Niederhoffer & Pennebaker, 2002). Such language-style matching has been shown to improve the outcomes of romantic relationships, for example (Ireland et al., 2011). We argue that over and above the *ex ante* similarity that leads people to become friends, these tendencies will lead friends to converge linguistically over time.

The proposition that linguistic similarity coevolves with network formation remains untested, but the development of new techniques in computational linguistics and the recent emergence of large-scale text corpora with associated network data now make such analyses both possible and relevant. We studied these processes in two unique and complementary empirical settings. In the first study, we collected two waves of linguistic and social network data on the complete incoming class of students working on their masters of business administration at a private East Coast university. This setting allowed us to study a bounded population and, given the rich set of additional covariates available, also allowed us to disentangle the effect of linguistic-style similarity from other competing sources of homophily. In the second study, we used data from 1.7 million online reviews written by 159,651 reviewers on Yelp.com—the full set of reviews for businesses in seven metropolitan areas over more than a decade (2005–2016)—as well as the online social networks of all active reviewers. Although each of these observational data sets was limited in significant ways, each had strengths that matched the limitations of the other, and together they provide strong and compelling evidence for both selection and convergence effects of linguistic homophily.

Finally, we discuss the consequences of the coevolutionary dynamics of linguistic style and network formation. We suggest that in settings in which both of these mechanisms are present, their coevolutionary dynamics will drive the population toward greater fragmentation and more homogenous clusters. This idea is consistent with prior work (DellaPosta, Shi, & Macy, 2015; Kalish, Luria, Toker, & Westman, 2015) and with our own simple network-simulation model (see the Supplemental Material available online). We argue, further, that these mechanisms go beyond mere clustering

of political views (Boutyline & Willer, 2017) and give rise to more fundamental social “echo chambers” that insulate us from dissimilar others.

Study 1

Method

Data. Our first study used data from all 285 first-year students in the graduate management program at a U.S. university (44% women; 78% White; 67% U.S. citizens). To examine their linguistic styles, we collected two writing samples from each student: their application essays, written prior to matriculation (and, therefore, prior to social network formation), and essays written for an exam in October, 2 months after the start of the school year. The first text was relatively unstructured, leaving students with broad latitude to express their individual linguistic styles; the second was more structured but still contained significant variance (see Fig. S1 in the Supplemental Material). In both texts, students were writing to a generalized other person rather than addressing a specific audience directly, making these samples good measures of individuals' default linguistic style. In addition to the two text corpora, we collected two waves of social network data (details about the survey instrument, developed by Kleinbaum, Jordan, & Audia, 2015, appear in the Supplemental Material).

We also measured personality using the broad-based HEXACO personality inventory (Ashton & Lee, 2009) as part of the first survey. Finally, we collected demographic data from the registrar to account for demographic sources of homophily, including each student's gender, ethnicity, and nationality. Students' identities remained anonymous because the various data sets were linked by encrypted student identifiers. All data were collected for pedagogical or administrative purposes, and their subsequent use for research, in deidentified form, was approved by the university's institutional review board. We had complete data across all data sources for 247 students, comprising 87% of the population.

Linguistic Inquiry and Word Count (LIWC) dimensions and linguistic similarity. To assess the linguistic styles of students, we used the LIWC coding system in the main set of analyses. We note, however, that our findings were still robust when we controlled for a broad range of alternative linguistic measures, as documented in the Supplemental Material. LIWC was developed by Pennebaker and colleagues (Chung & Pennebaker, 2007; Pennebaker, Boyd, Jordan, & Blackburn, 2015; Pennebaker & King, 1999; Tausczik & Pennebaker, 2010), who argue that although content words (such as verbs or objects) are

crucial to communicate meaning, each speaker or writer also simultaneously communicates a linguistic style, which is best captured by his or her pronoun usage. Through decades of work (for a review, see Pennebaker, 2011), they have developed a coding dictionary that categorizes almost 6,400 words into 89 themes (Pennebaker et al., 2015), and across a series of studies, they have documented how these themes relate to the psychology of individuals (Chung & Pennebaker, 2007; Jordan & Pennebaker, 2017; Pennebaker et al., 2015; Pennebaker & King, 1999; Tausczik & Pennebaker, 2010). Of these 89 themes, 18 directly capture linguistic style, and in our analyses, we focused on these dimensions. For example, heavy use of first-person pronouns (“I,” “me”) is related to introversion and depression, but frequent use of third-person pronouns (“he,” “she,” “they”) indicates high levels of abstraction and cognitive processes (Pennebaker, 2011; Pennebaker & King, 1999; Tausczik & Pennebaker, 2010). See Table S1 in the Supplemental Material for the list of categories included in our analyses.

What was important for the current research is that usage of these linguistic cues indicates personal style, which is largely independent of the content of the communication. Even though these markers of linguistic style are unconscious, they reflect students’ psychology in ways that are observable to one another and that, consequently, affect their choices of whom to befriend. These styles are also susceptible to peer influence over time. To provide a clearer view of the differences at the heart of the quantitative analysis, we include an illustrative example of linguistic difference in Table S2 in the Supplemental Material.

In our quantitative analyses, we measured linguistic-style similarity as the aggregate similarity across 18 dimensions of word usage. We first calculated, for each text, the total number of words within each dimension. For example, the dimension “first-person singular” counts all instances of “I,” “me,” “myself,” and so on. Negations were intentionally included in these counts: Even if people write “not me,” they are still talking about themselves.

After determining the word count for each dimension in each text, we normalized these counts by the total number of words in the text. Because the dimensions vary in their global prevalence, we standardized each dimension separately, constructing the distribution of individuals’ language use along each dimension to have a mean of 0 and a standard deviation of 1.

Next, to create a composite linguistic-similarity measure between two individuals, we aggregated their linguistic similarity along the 18 dimensions by calculating the total variation distance as the average difference between person i and person j across those dimensions.

Finally, following Shepard (1987), we calculated dyadic linguistic similarity as the negative logarithm of the total variation distance:

$$\text{linguistic similarity}_{ijt} = -\log \left(\frac{\sum_d |NWC_{dit} - NWC_{djt}|}{D} \right),$$

where NWC_{dit} represents the normalized word count of linguistic dimension d in person i ’s time t text, NWC_{djt} represents the same for person j , and D is the total number of linguistic dimensions analyzed (18). This linguistic-similarity variable was standardized for greater comparability across samples. We constructed a data set of all possible pairwise combinations of students and calculated linguistic similarity for each dyad. Figure S1 plots the distribution of these pairwise similarities for Time 2.

Estimation procedures. We used dyad-level models (Kenny, Kashy, & Cook, 2006) to investigate friendship choice and linguistic-style convergence. In dyad-level models, the unit of analysis is not a person but a pair of persons. In these dyadic models, an observation is an ij undirected pair, and the dependent variable is an indicator of whether person i and person j both cited each other as a friend (0 = no, 1 = yes). Therefore, each individual appeared in the data not only as an i but also as a j for all others in the social environment, and the 247 students were entered into the analyses as 30,381 ($0.5 \times 247 \times 246$) undirected dyads. Further details on the dyad-level sample appear in the Supplemental Material.

Models predicting the existence of a dyadic friendship tie were estimated using logistic regression. As mentioned above, each possible pair of individuals was entered as an observation, and the dependent variable was the presence (1) or absence (0) of a friendship between the members of that pair. The main independent variable here was the similarity in linguistic style between the two individuals in the dyad in the prior time period. We controlled for the number of social relationships each dyad member, i and j , participated in (in network terminology, their *degree scores*) to account for both members’ base rates of tie formation. In addition, we controlled for person i and person j having the same class section, study group, gender, race, and nationality and for the similarity of i and j along the HEXACO dimensions (see Table S4 in the Supplemental Material for the full list of covariates included in the models). Formally, this equation would be written as follows:

$$E[\text{friendship}_{ijt_2}] = \beta_0 + \beta_1 \text{linguistic similarity}_{ijt_1} + \beta_2 X_{ij} + \varepsilon,$$

where T_1 and T_2 refer to the two waves of data collection (Time 1 and Time 2, respectively) and X_{ij} is a vector of dyadic control variables including measures of person i 's and person j 's baseline propensities to form network ties and the dyadic similarity between i and j along demographic and personality dimensions.

To capture linguistic convergence, we used ordinary least squares regression to model the dyadic change in linguistic-style similarity as a function of friendship and controlled for prior linguistic similarity:

$$\Delta \text{linguistic similarity}_{ij} = \beta_0 + \beta_1 \text{friendship}_{ij} + \beta_2 \text{linguistic similarity}_{ijt_1} + \beta_3 X_{ij} + \varepsilon,$$

where $\Delta \text{linguistic similarity}_{ij}$ is the change in linguistic similarity between person i and person j from Time 1 to Time 2, standardized across the population of dyads. Friendship_{ij} and $\text{linguistic similarity}_{ij}$ are binary indicators of whether (1) or not (0) a reciprocated friendship or linguistic similarity, respectively, existed between person i and person j at both Time 1 and Time 2 (see the Supplemental Material for additional details on the model specifications).

The dyadic data structure means that each person participates in many dyadic observations. This violation of the assumptions of regression would result in artificially small standard errors, yielding results that appear to be more precisely estimated than they actually are. Fortunately, such dyadic dependencies are easily accounted for in network data via the multiway-clustering approach (Cameron, Gelbach, & Miller, 2011; Kleinbaum, Stuart, & Tushman, 2013; Lindgren, 2010). Prior research in psychology (Feiler & Kleinbaum, 2015) has shown that clustering on both dyad members properly accounts for structural autocorrelation in dyad models. All standard errors reported in this article were estimated with the multiway-clustering approach; this is the most statistically conservative approach to calculating standard errors for such dyadic data structures, and all our results would hold with other error-clustering methods, such as robust or bootstrapped standard errors.

Results

Descriptive statistics for the sample used in Study 1 appear in Table S3 in the Supplemental Material; a histogram of dyadic linguistic similarity appears in Figure S1. The results of multivariate regressions appear in

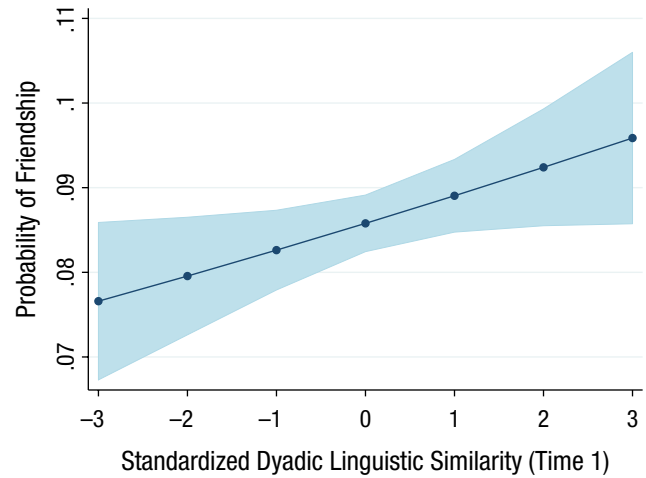


Fig. 1. Marginal-effects plot from Study 1 showing the probability of friendship as a function of linguistic similarity at Time 1, holding dyad members' degree centrality and similarity in demographic and personality variables at mean levels. This figure is based on results from a dyad-level logistic regression shown in Table S4, Model 2, in the Supplemental Material available online ($N = 30,381$ dyads). The shaded area represents 95% multiway cluster-robust confidence intervals.

Table S4. Model 1 showed that independently from endogenous network-structure controls, linguistic similarity was related to the probability of becoming friends, $b = 0.079$, 95% confidence interval (CI) = [0.0308, 0.1273], $z = 3.22$, $p = .001$, odds ratio (OR) = 1.0823. The magnitude of the effect is notable: A 1-standard-deviation increase in dyadic linguistic similarity increased the likelihood of friendship by 8.2% ($OR = 1.082$). The results are depicted in Figure 1 and Figure 2a. In Model 2, we added controls for shared demography and similar personality; the effect of linguistic similarity was, as expected, diminished somewhat but remained statistically significant, $b = 0.049$, 95% CI = [0.0025, 0.0963], $z = 2.06$, $p = .039$, $OR = 1.051$.

Linguistic similarity also acts on friend selection by reducing the rate of tie decay. In Models 3 and 4 (see Table S4), we modeled the presence of a friendship tie at Time 2 on the set of dyads with a reciprocal friendship tie at Time 1. There was a positive coefficient of linguistic similarity in Model 4, which indicates that, all else being equal, a 1-standard-deviation increase in linguistic similarity increases the likelihood of tie persistence (i.e., reduces the likelihood of tie decay) by 14%, $b = 0.1292$, 95% CI = [0.0040, 0.2544], $z = 2.02$, $p = .043$, $OR = 1.1379$.

Next, we examined the association between friendship ties and linguistic convergence. The covariate for prior linguistic similarity in Models 5 and 6 (see Table S4) indicates that previously similar dyads had less room for convergence. However, friendship was

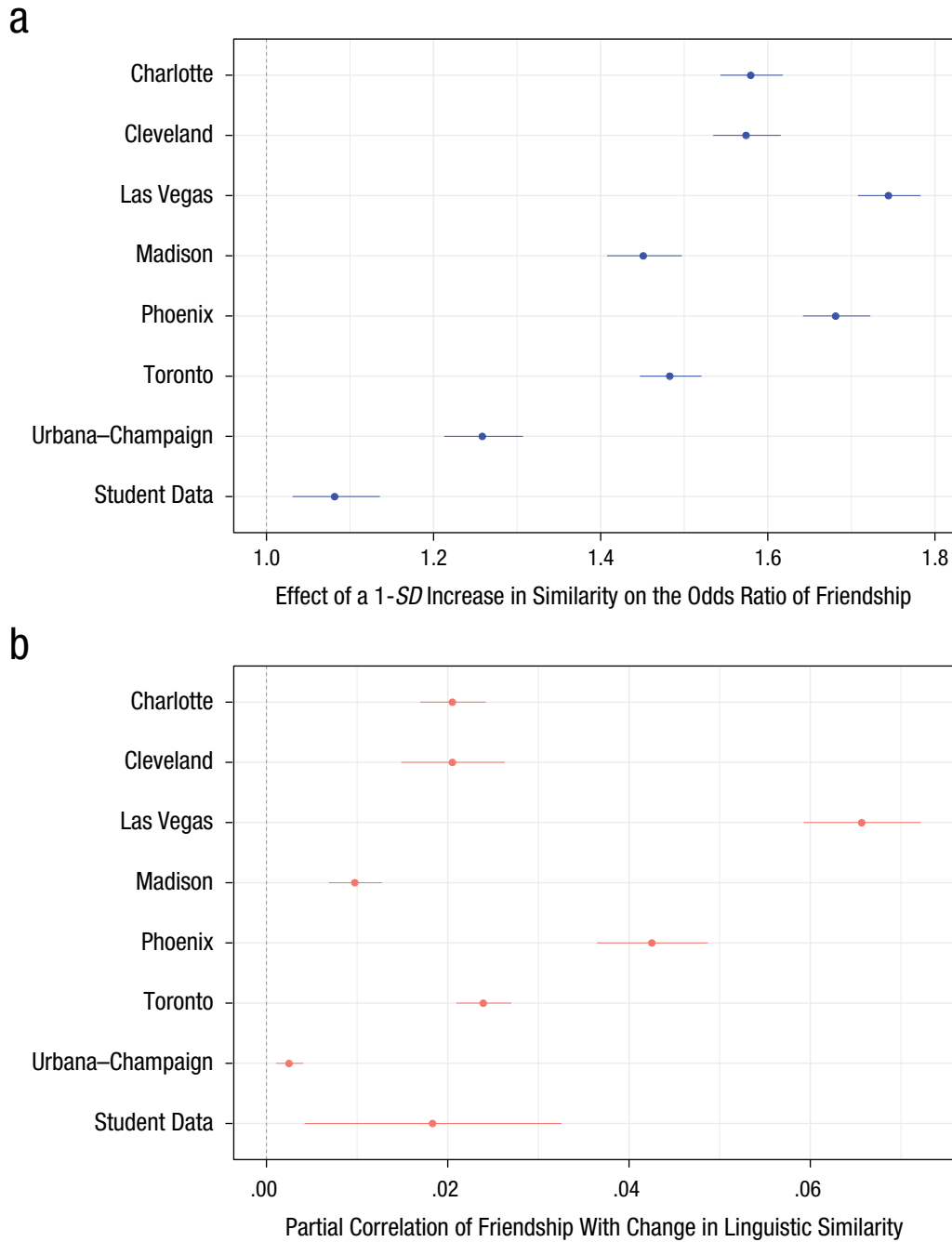


Fig. 2. Selection effects (a) and convergence effects (b) in Studies 1 and 2. The effect of a 1-standard-deviation increase in linguistic-style similarity on the likelihood of a friendship tie (a) is shown for the seven geographical locations analyzed in Study 2 and for student data from Study 1. For each study in this analysis, results are based on the dyad-level logistic regression estimates of Model 2 (see Tables S4 and S6 in the Supplemental Material available online). The effect of the existence of a friendship tie on the change in similarity of linguistic style within a dyad (b) is also shown for the seven geographical locations analyzed in Study 2 and for student data from Study 1. For each study in this analysis, results are based on the dyad-level linear regression estimates of Model 4 (see Tables S4 and S7 in the Supplemental Material). Error bars in both panels represent 95% confidence intervals based on standard errors clustered on both members of each dyad.

associated with increased linguistic similarity over time. As in Models 1 and 2, the effect was strongest in uncontrolled regressions because of shared variance, $b = 0.1292$, 95% CI = [0.0296, 0.2289], $t(245) = 2.55$, $p = .011$, partial $r = .0184$ (depicted in Fig. 2b), and persisted after demographic and personality controls were added, $b = 0.1078$, 95% CI = [0.0027, 0.2128], $t(245) = 2.02$, $p = .044$, $r = .0153$.

Discussion

Study 1 demonstrated that linguistic similarity in application essays predicts increased likelihood that students will become friends and stay friends and, furthermore, that students who became friends early in the program converged in their linguistic styles by the time of the exam. These findings held even after we controlled for other possible factors influencing network formation and linguistic-style change, such as gender, nationality, native language, race, and personality, though the effect sizes were small, particularly the convergence effects, perhaps because of the short study interval. This result motivated us to replicate the study in a larger sample and over a longer time frame, which we did in Study 2.

Study 2

Method

Data. Yelp.com is an online review platform in which users can post reviews of restaurants, museums, barber shops, or any other business, including star ratings and written comments. As of 2016, Yelp.com had more than 70 million registered users worldwide and more than 100 million reviews of 2 million establishments (Yelp.com/factsheet, accessed August 1, 2016). Like the writing samples used in Study 1, reviews are written to a generalized audience, not to a specific target, thus capturing the author's default linguistic style.

The data we analyzed came from two data sets made publicly available by Yelp.com to researchers as part of the Yelp Challenge (see <https://www.yelp.com/dataset/challenge>); our data came from Rounds 8 and 9 (we will refer to these as Waves 1 and 2, respectively). The data contain all reviews published in 10 metropolitan areas: 6 in the United States (Phoenix, Arizona; Las Vegas, Nevada; Pittsburgh, Pennsylvania; Urbana–Champaign, Illinois; Charlotte, North Carolina; and Madison, Wisconsin), 2 in Canada (Toronto and Montreal), 1 in the United Kingdom (Edinburgh), and 1 in Germany (Karlsruhe). Because we wanted to conduct our analyses on a comparable set of primarily English-speaking cities, we excluded the European cities and Montreal from our

analyses and focused on the 7 North American metropolitan areas in which English is the primary language.

Round 8 contained all reviews published in these metropolitan areas prior to August 3, 2016. The data set for Round 9 was released on January 21, 2017, and contained all reviews written in the same metropolitan areas as in Round 8. We matched these two waves of data to create a two-wave panel data set.

An important feature of Yelp.com is that it also has social networking functionality that allows people to tag their friends. These friendship relationships are symmetric by design: They must be approved by the receiving party (so they are not one-sided relationships in which only one person “follows” the other). No information is available about the strength of ties. As with most online social networks, the meaning of “friend” is somewhat different from that endorsed by the students in Study 1, but anecdotal evidence suggests that some of these reviewers also know each other in the off-line world. For example, Donna B. wrote in one review, “I went here for a quick snack before a Yelp event,” referring to an in-person event that Yelp organized to bring its reviewers together.

Of the 593,939 unique users in the data set, 27% (159,651) also used the social networking functionality of Yelp in both waves. On average, Yelp users in Wave 1 who both reviewed local businesses and used the social networking feature on the site had 14.0 friends; the median friend count was 3, indicating a highly skewed distribution. As is typical of large-scale social networks, the Yelp-reviewer friendship network is sparse (density $\ll 1\%$). By Wave 9, more friendship ties had formed for the same set of reviewers, averaging 71.7 per person. The serial autocorrelation in individuals' network scores was .895.

Because we wanted to analyze how friendship formation and linguistic style influence each other, we focused on the set of reviewers who contributed at least one review and had at least one friend in each wave. (See Table S5 in the Supplemental Material for descriptive statistics.) The data set we analyzed contained 1,749,470 reviews written by 159,651 reviewers. The average Yelp review is 115.8 words long and is addressed to a generalized audience, providing a suitable platform to assess the linguistic style of reviewers. For reviewers who contributed more than one review, we calculated the normalized word counts for each review and linguistic dimension separately, and then to measure the individual's overall linguistic profile for that period, we averaged these values for each dimension that appeared in posts by that reviewer in each observation period.

Estimation procedures. To assess the linguistic styles of reviewers, we used the LIWC coding system in the

main set of analyses, as in Study 1. We also analyzed the data in a dyadic format, exactly as in Study 1. We estimated logistic regressions on the sample of all possible friendship dyads; the dependent variable was a binary indicator of reciprocal friendship in 2016. Because geographical proximity is a major driver of friendship-tie formation (Marmaros & Sacerdote, 2006), we analyzed each metropolitan area separately; this approach ensured that all pairwise dyads in the analyses had at least a nonnegligible probability of forming a friendship tie. Because the network was large and sparse (< 1% of possible friendship ties were present), we used a case-cohort design (King & Zeng, 2001; Kleinbaum et al., 2013), sampling all observations with an observed tie but only a fraction of the nonpresent ties. Consequently, for each focal person, we sampled an average of 50 other persons who were not friends with the focal person. For example, for a person with 16 friends, we included 16 observations with 1 as an outcome variable and 50 observations with 0 as an outcome variable. To ensure that this estimation strategy was efficient, we reweighted all such zero observations so their weight would be representative of the whole sample. We viewed the choice of 50 matched counterfactuals as a reasonable compromise between including all zero observations and including only a few nonobserved ties. Including all zero observations could make the size of the emerging data set too large to handle; for example, if all pairwise combinations of 60,204 reviewers in Phoenix were to be included, the data set would contain 3.6 billion observations. In contrast, including only a few nonobserved ties could result in unstable estimates. This estimation strategy still yielded robust results when we used matched samples of other sizes (such as 20 or 100), which resulted in substantially similar patterns of findings.

Results

Linguistic similarity predicts network formation.

Figure 2a depicts the estimated coefficients for each metropolitan area (see also Table S6, Model 1, in the Supplemental Material for the dyad-level logistic regression results). We found that similarity in linguistic styles between two reviewers corresponds to a higher likelihood of a friendship tie between the reviewers—Charlotte: $b = 0.4576$, $SE = 0.0121$, 95% CI = [0.4339, 0.4812], $z = 37.9694$, $p < .001$, $OR = 1.5866$; Cleveland: $b = 0.4540$, $SE = 0.0131$, 95% CI = [0.4283, 0.4796], $z = 34.7344$, $p < .001$, $OR = 1.5794$; Las Vegas: $b = 0.5568$, $SE = 0.0110$, 95% CI = [0.5353, 0.5783], $z = 50.7573$, $p < .001$, $OR = 1.7623$; Madison: $b = 0.3727$, $SE = 0.0157$, 95% CI = [0.3419, 0.4036], $z = 23.6818$, $p < .001$, $OR = 1.4470$; Phoenix: $b = 0.5199$, $SE = 0.0122$, 95% CI = [0.4960, 0.5439],

$z = 42.5417$, $p < .001$, $OR = 1.6996$; Toronto: $b = 0.3943$, $SE = 0.0127$, 95% CI = [0.3695, 0.4191], $z = 31.1658$, $p < .001$, $OR = 1.4897$; Urbana–Champaign: $b = 0.2303$, $SE = 0.0191$, 95% CI = [0.1929, 0.2677], $z = 12.0621$, $p < .001$, $OR = 1.2610$. The effect size was quite substantial: A 1-standard-deviation increase in linguistic similarity between members of a dyad increased the odds of a friendship tie anywhere from 26% (in Urbana–Champaign) to 76% (in Las Vegas). As mentioned, these models were estimated with standard errors clustered on both members of each dyad. We also controlled for the baseline probability that these two reviewers became friends.

The results thus far were correlational, but with the help of two waves of network data, we were able to begin disentangling the dual causal mechanisms. To test whether similarity in linguistic style predicts increased probability of creating a friendship tie, we reestimated the dyadic logistic models of the previous analysis on the 2016 data but excluded the set of dyads who were already friends in the 2016 wave. In other words, we tested whether linguistic-style similarity in 2016 led to formation of new network ties. The test therefore was estimated on the same set of reviewer dyads minus the already existing friendship dyads, resulting in 4,175,668 observations. Out of these, 32,617 new friendships were born. We estimated a logistic regression at the dyad level, as before, with multiway-clustered standard errors in which the explanatory variable was the linguistic-style distance between the members of the dyad in 2016.

We found that linguistic similarity predicted the formation of new ties—Charlotte: $b = 0.4333$, $SE = 0.0142$, 95% CI = [0.4055, 0.4611], $z = 30.562$, $p < .001$, $OR = 1.5477$; Cleveland: $b = 0.4129$, $SE = 0.0177$, 95% CI = [0.3782, 0.4475], $z = 23.348$, $p < .001$, $OR = 1.5231$; Las Vegas: $b = 0.5002$, $SE = 0.0175$, 95% CI = [0.4660, 0.5344], $z = 28.6294$, $p < .001$, $OR = 1.6762$; Madison: $b = 0.3443$, $SE = 0.0213$, 95% CI = [0.3026, 0.3859], $z = 16.2004$, $p < .001$, $OR = 1.4144$; Phoenix: $b = 0.2671$, $SE = 0.0203$, 95% CI = [0.2273, 0.3069], $z = 13.151$, $p < .001$, $OR = 1.5262$; Toronto: $b = 0.3899$, $SE = 0.0206$, 95% CI = [0.3496, 0.4302], $z = 18.9572$, $p < .001$, $OR = 1.4863$; Urbana–Champaign: $b = 0.2796$, $SE = 0.0313$, 95% CI = [0.2182, 0.3411], $z = 8.9225$, $p < .001$, $OR = 1.3244$ (see Table S6, Model 2, for full results).

To further investigate the functional form of the selection effect, we reestimated Model 1 (see Table S6), but instead of assuming a linear functional form of the effect, we rounded the standardized similarity measure to the closest 0.2 resolution (i.e., to similarity z -score = $-3, -2.8, -2.6, \dots, 2.6, 2.8, 3$) and included an indicator variable in the regression for each of these levels. Figure 3 shows the marginal effect of similarity on the

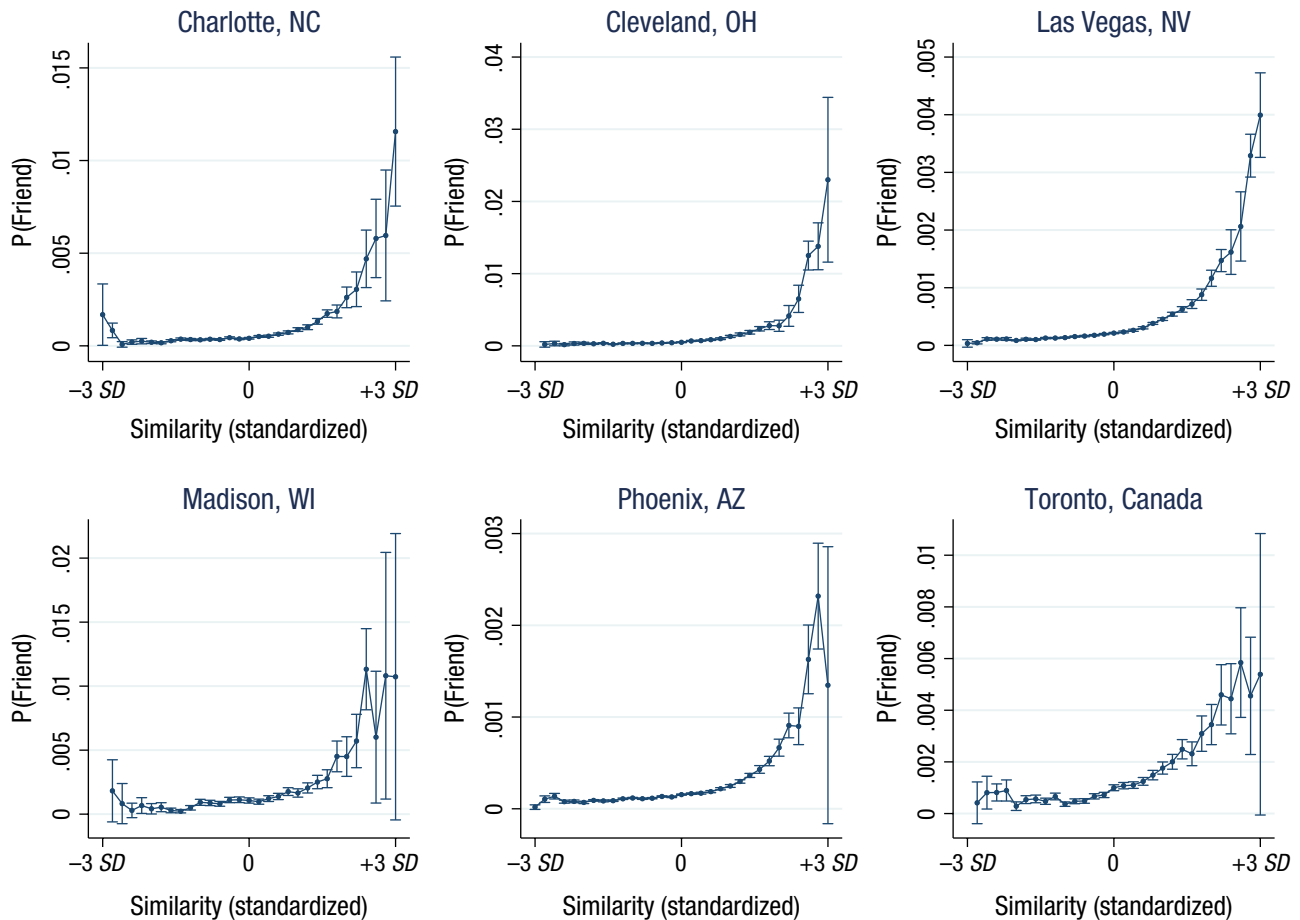


Fig. 3. Marginal effect in Study 2 of linguistic-style similarity on the probability of a friendship tie (dyad-level logistic regression with dummy variables at each 0.2 level of the standardized z-score dyadic linguistic-similarity measure). The data set contained 4,488,715 individuals and had 81,452 degrees of freedom. Error bars show 95% multiway-clustered confidence intervals.

likelihood of a friendship tie. In all of these models, a significant amount of the variation of interest lies in the tails of the distribution; but in the large number of observations in our dyadic analysis, this constituted meaningful and important variance. Our conclusions were still robust after we controlled for gender effects and idiosyncratic restaurant-level effects; see Tables S9 and S10 in the Supplemental Material.

Friends' linguistic styles become more similar over time. Next, we examined the reverse mechanism: the linguistic convergence between friends. To do this, we tested whether the linguistic similarity between members of a dyad increased more between 2016 and 2017 if they were friends in 2016 than if they were not friends at that time. On all possible dyads, we ran a linear regression in which the dependent variable was the change of linguistic similarity in the set of reviews that were written between the two waves. The independent variables were (a) the linguistic-style similarity at the time of the first wave and (b) whether the dyad members were friends at the time of the first wave. Results are depicted visually by

geographical location in Figure 2b (see also Table S7 in the Supplemental Material). We found that although linguistic similarity at Wave 1 strongly predicted linguistic similarity at Wave 2 ($r = .463, p < .0001$), linguistic similarity was greater between reviewers who were friends. That is, our finding is consistent with previous literature in that linguistic style is to a large extent stable within a person (across-time $r_s = .70-.85$) and that to the extent that it changes, friends converge in their linguistic styles. This pattern was evident across geographical areas, although its strength varied—Charlotte: $b = 0.2839, SE = 0.0323, 95\% CI = [0.2206, 0.3471], t(81452) = 8.7919, p < .001, r = .0206$; Cleveland: $b = 0.3112, SE = 0.0793, 95\% CI = [0.1558, 0.4667], t(81452) = 3.9234, p < .001, r = .0205$; Las Vegas: $b = 0.1676, SE = 0.0122, 95\% CI = [0.1437, 0.1915], t(81452) = 13.7507, p < .001, r = .0658$; Madison: $b = 0.3154, SE = 0.0513, 95\% CI = [0.2148, 0.4159], t(81452) = 6.148, p < .001, r = .0096$; Phoenix: $b = 0.1702, SE = 0.0192, 95\% CI = [0.1326, 0.2078], t(81452) = 8.8692, p < .001, r = .0426$; Toronto: $b = 0.5284, SE = 0.0406, 95\% CI = [0.4488, 0.6080], t(81452) = 13.0087, p < .001, r = .024$; Urbana-Champaign: $b = 0.1938,$

$SE = 0.0558$, 95% CI = [0.0845, 0.3031], $t(81452) = 3.4745$, $p < .001$, $r = .0026$.

Discussion

Study 2 demonstrates that linguistic similarity in Yelp reviewers' earlier reviews predicts subsequent friendship between them. Moreover, linguistic styles of reviewers who were friends during the time of the first data collection (August 2016) converged in later reviews (August 2016–January 2017). These findings held even after we controlled for other factors influencing network formation and linguistic-style change, such as gender and business fixed effects (see Tables S9 and S10 in the Supplemental Material).

The great virtues of Study 2 are, of course, its large sample size and multiple sites, but its major limitation is that online friendship ties may not represent off-line friendship ties. Some Yelp reviewers do have opportunities to meet in real life, but most of them interact only by reading each other's reviews online. Thus, the only basis they have on which to know one another is their writing. Indeed, prior evidence suggests that in online relationships, people put less emphasis on observable sociodemographic characteristics, such as gender, age, or physical attractiveness (Jacobson, 1999). Thus, it is not surprising that in Study 2, we found a much stronger effect of linguistic similarity in determining who was friends with whom on an online social network than we did in Study 1 (Study 1: $OR = 1.08$; Study 2: $ORs = 1.26$ – 1.76). Relatedly, although effect sizes varied somewhat across sites, they were statistically significant in all cities. Future research could investigate why the effect size of linguistic similarity may vary across cities.

Another limitation is that because these are online friendship data, people are much more likely to add friends than to (formally) drop friends. In off-line settings, friendships typically just fall dormant (Levin, Walter, & Murningham, 2011) when people meet and talk less often than they once did. In online social networks, by contrast, dissolving an online friendship tie requires deliberately "unfriending," an act seen by most people as openly hostile. Unfriending is therefore very rare; we observed only 22 cases in our whole sample. Taken together, these forces imply that a secular increase in network size is the norm in online social networks.

This leads to certain limitations of Study 2. First, the findings of Study 2 would be less likely to generalize to settings in which adding or dropping a network tie is equally easy or likely. Second, because dropping ties is very rare, we could not reliably measure tie-persistence effects in Study 2. Finally, our data speak more to properties of growing networks. Future research could test

whether stable, or even shrinking, networks would exhibit similar patterns.

Given, however, that the limitations of Study 2 are matched by the strengths of Study 1, the two studies together constitute robust evidence of the selection and convergence mechanisms that give rise to linguistic homophily. We believe that this second study substantially generalizes the findings of Study 1 not only to a different setting that is becoming ever more important but also to a much larger data set that covers multiple geographical locations and demographic backgrounds.

General Discussion

In this research, we demonstrated the dual mechanisms of linguistic homophily: that people with similar linguistic styles are more likely to form and maintain friendships and that friends experience linguistic convergence over time. While prior research has demonstrated homophily processes along social dimensions such as gender, age, personality, and national background, we show that even after analyses control for all these dimensions, linguistic-style similarity plays a role in explaining network formation. Finally, we suggest that these mechanisms give rise over time to fragmentation of the network, creating structural echo chambers, not only in partisan politics but also in the very structure of the social network itself.

We believe that our findings have ever-increasing relevance in the digital age. During most of the history of humankind, communication and tie-formation patterns were predominantly driven by face-to-face interactions, and thus attributes such as age, gender, or socioeconomic status were readily observable. In a world that is increasingly dominated by online communications, however, the role of such off-line cues will be diminishing, partly because they are not readily available or not highly salient. For example, it is much easier to forget about the gender of an interlocutor whom you cannot see. Therefore, we conjecture that linguistic similarity will be of increasing relevance on platforms dominated by textual communication, such as e-mail, chat rooms, or online reviews. Linguistic-style similarity, therefore, is an important factor in various social processes, including network formation, but also in other related phenomena, such as the flow of influence or information (Traud, Mucha, & Porter, 2012).

By studying two such markedly different empirical settings, we effectively counterbalanced the limitations of each setting against the strengths of the other. However, as in all research, limitations remain. First, as in any observational study, our ability to make causal inferences was limited; in this case, however, this limitation was counterbalanced by the benefits of studying

the coevolution of social networks and individual linguistic style in two field settings over substantial periods of time. Future research could examine these effects in the controlled setting of the lab, though it is unclear what treatment over what duration could induce such effects. Second, research on language-style matching posits that how we talk may depend on whom we are talking to (Nguyen et al., 2016). In our settings, texts are addressed to a generalized audience (an unknown admissions committee; users of Yelp), rather than to a specific other person. Future research could investigate how linguistic code switching may facilitate network formation. Third, our measure of linguistic convergence was based on a change score, which some researchers have criticized as unreliable and others have defended. Finally, the observed effect sizes are quite modest, especially for models of linguistic convergence. Such small effects are expected for two reasons. For one, substantive change in the use of subtle function and grammar words is likely to be a slow process; for another, our observation period was only a couple months. In other words, if we were to observe the evolution of the social networks for a longer time period, such as decades, we would probably see larger effects.

The findings are striking because many of these linguistic-style dimensions relate to psychological processes that are unconscious and deeply ingrained in human personality and thus are relatively stable over time (Pennebaker, 2011). This is important because the stability of linguistic-style patterns points to limits of the malleability of social networks and to the limits of social network mobility.

More generally, our evidence of linguistic selection and convergence suggests that over time, people will connect with increasingly similar others and become increasingly similar to their contacts. The implication—consistent with observations of society in recent years—is that networks will increase in fragmentation and polarization over time. Indeed, societal observers have pointed to an increase in the incidence of echo chambers worldwide, in which people interact with others like themselves and, as a result, hear messages that reaffirm their preexisting beliefs (Sunstein, 2002). Our findings shed light on these dynamics: We argue that the dual mechanisms of homophily—selection into friendship and subsequent convergence between friends—form the microfoundations of echo chambers, not only in our political views or our consumption of information (Boutyline & Willer, 2017) but in the very fabric of the social network itself. Our empirical work documents these dual mechanisms with respect to linguistic style, and both prior research (Kalish et al., 2015) and our own simulation model (see the

Supplemental Material) suggest that these processes lead to increasing fragmentation of the network.

However, echo chambers are something of a double-edged sword. While they tend to cut us off from distant information and dissimilar perspectives, they also enable coordination between like-minded people and, in doing so, may facilitate the performance of existing tasks. Indeed, in research literatures as diverse as organization design (Thompson, 1967) and entrepreneurship (Ruef, 2010), there is a well-known trade-off between efficiency and novelty (March, 1991). These functional benefits must be considered alongside the potential dangers of echo chambers.

In a world of dramatic and seemingly increasing polarization—in which we talk primarily to other people who share our views and utterly fail to comprehend those who do not—elucidating the mechanisms that bring about such fragmentation offers the possibility that we can begin to reintegrate our society and, in the process, promote civil discourse about politics and, more fundamentally, in all facets of social life.


Action Editor

Brent W. Roberts served as action editor for this article.

Author Contributions

B. Kovacs developed the original study concept. Both authors designed the studies. Data for Study 1 were collected and analyzed by A. M. Kleinbaum. Data for Study 2 were collected and analyzed by B. Kovacs. A. M. Kleinbaum and B. Kovacs wrote the article together. Both authors approved the final version of the manuscript for submission.

ORCID iD

Balazs Kovacs  <https://orcid.org/0000-0001-6916-6357>

Acknowledgments

We appreciate the feedback we received from seminar presentations at Carnegie Mellon University, Central European University, Dartmouth College, the European School of Management and Technology (ESMT) Berlin, Kibbutzim College, Massachusetts Institute of Technology, Northwestern University, Paris Institute of Political Studies, and Yale University and from conference participants at the Academy of Management and Northwestern University's 2018 ANN/SONIC/NICO Workshop.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/0956797619894557>

Open Practices

Data and materials for these studies have not been made publicly available, and the design and analysis plans were not preregistered.

References

- Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences, USA*, 106, 21544–21549.
- Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91, 340–345.
- Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, 38, 551–569.
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29, 238–249.
- Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76, 893–910.
- Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social communication* (pp. 343–359). New York, NY: Psychology Press.
- DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, 120, 1473–1511.
- Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41, 87–100.
- Feiler, D. C., & Kleinbaum, A. M. (2015). Popularity, similarity, and the network extraversion bias. *Psychological Science*, 26, 593–603.
- Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal*, 337, Article a2338. doi:10.1136/bmj.a2338
- Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37, 3–19.
- Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative Science Quarterly*, 37, 422–447.
- Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22, 39–44.
- Jacobson, D. (1999). Impression formation in cyberspace: Online expectations and offline experiences in text-based virtual communities. *Journal of Computer-Mediated Communication*, 5(1), Article JCMC511. doi:10.1111/j.1083-6101.1999.tb00333.x
- Jordan, K. N., & Pennebaker, J. W. (2017). The exception or the rule: Using words to assess analytic thinking, Donald Trump, and the American presidency. *Translational Issues in Psychological Science*, 3, 312–316.
- Kalish, Y., Luria, G., Toker, S., & Westman, M. (2015). Till stress do us part: On the interplay between perceived stress and communication network dynamics. *Journal of Applied Psychology*, 100, 1737–1751.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9, 137–163.
- Kleinbaum, A. M., Jordan, A. H., & Audia, P. G. (2015). An alter-centric perspective on the origins of brokerage in social networks: How perceived empathy moderates the self-monitoring effect. *Organization Science*, 26, 1226–1242.
- Kleinbaum, A. M., Stuart, T. E., & Tushman, M. L. (2013). Discretion within constraint: Homophily and structure in a formal organization. *Organization Science*, 24, 1316–1336.
- Levin, D. Z., Walter, J., & Murnighan, J. K. (2011). Dormant ties: The value of reconnecting. *Organization Science*, 22, 923–939.
- Lindgren, K. O. (2010). Dyadic regression in the presence of heteroscedasticity—an assessment of alternative approaches. *Social Networks*, 32, 279–289.
- March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2, 71–87.
- Mark, N. (1998). Birds of a feather sing together. *Social Forces*, 77, 453–485.
- Marmaros, D., & Sacerdote, B. (2006). How do friendships form? *The Quarterly Journal of Economics*, 121, 79–119.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–444.
- Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42, 537–593.
- Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21, 337–360.
- Noë, N., Whitaker, R. M., & Allen, S. M. (2016, August). *Personality homophily and the local network characteristics of Facebook*. Paper presented at the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA.
- Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications*, 9, Article 332. doi:10.1038/s41467-017-02722-7
- Pennebaker, J. W. (2011). *The secret life of pronouns*. New York, NY: Bloomsbury Press.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77, 1296–1312.
- Ruef, M. (2010). *The entrepreneurial group: Social identities, relations, and collective action*. Princeton, NJ: Princeton University Press.

- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317–1323.
- Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, 10, 175–195.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 24–54.
- Thompson, J. D. (1967). *Organizations in action: Social science bases of administrative theory*. New York, NY: McGraw-Hill.
- Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of Facebook networks. *Physica A: Statistical Mechanics and Its Applications*, 391, 4165–4180.
- Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *American Journal of Sociology*, 116, 583–642.
- Youyou, W., Stillwell, D., Schwartz, H. A., & Kosinski, M. (2017). Birds of a feather do flock together: Behavior and language-based personality assessment reveal personality homophily among couples and friends. *Psychological Science*, 28, 276–284.