Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–12
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797619894557
www.psychologicalscience.org/PS
SAGE

Language-Style Similarity and Social Networks

Balazs Kovacs[1] and Adam M. Kleinbaum[2]
[1]School of Management, Yale University, and [2]Tuck School of Business, Dartmouth College

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:

$$\text{linguistic similarity}_{ijt} = -\log\left(\frac{\sum_{d}^{D}\left|\text{NWC}_{dit} - \text{NWC}_{djt}\right|}{D}\right),$$

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 × 247 × 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:

$$E\left[\text{friendship}_{ijT_2}\right] = \beta_0 + \beta_1 \text{linguistic similarity}_{ijT_1} + \beta_2 X_{ij} + \varepsilon,$$

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:

$$\Delta\text{linguistic similarity}_{ij} = \beta_0 + \beta_1 \text{friendship}_{ij} + \beta_2 \text{linguistic similarity}_{ijT_1} + \beta_3 \text{X}_{ij} + \varepsilon,$$

where Δlinguistic 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 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

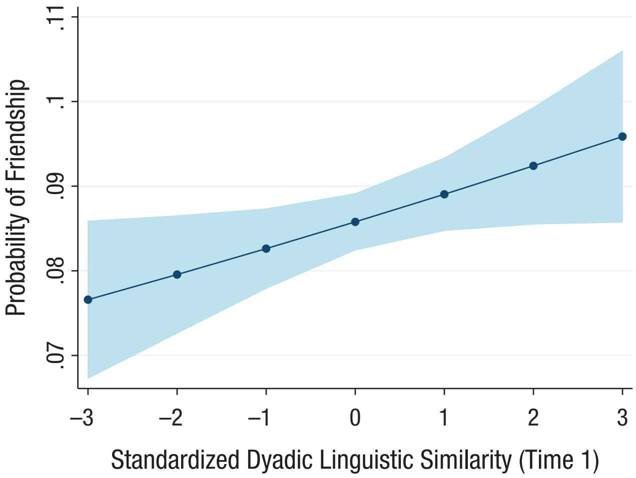


Fig. 1. Marginal-effects plot from Study 1 showing the probability of friendship as a function of linguistic similarity at Time 1, holding dyad members' degree centrality and similarity in demographic and personality variables at mean levels. This figure is based on results from a dyad-level logistic 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

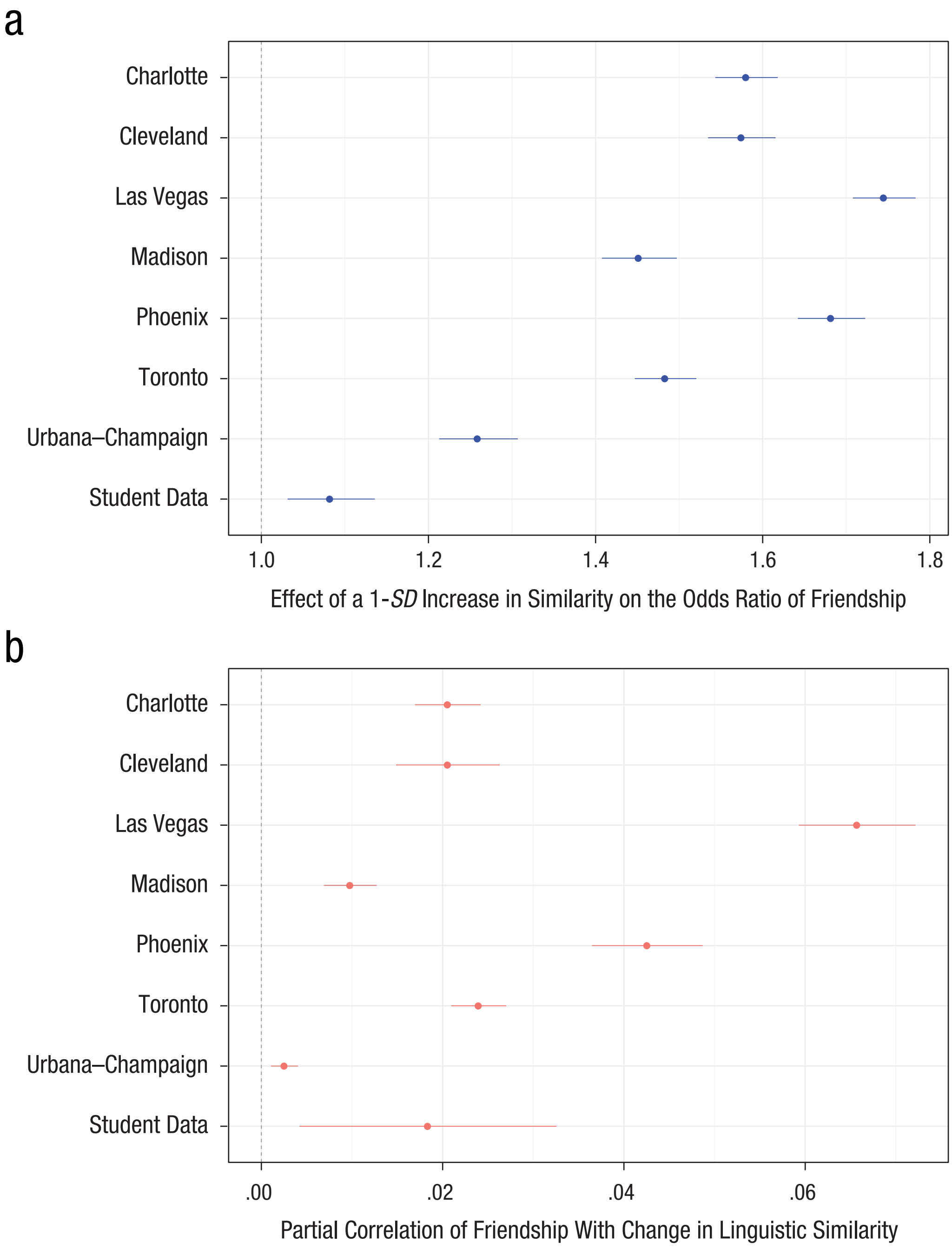


Fig. 2. Selection effects (a) and convergence effects (b) in Studies 1 and 2. The effect of a 1-standard-deviation increase in linguistic-style similarity on the likelihood of a friendship tie (a) is shown for the seven geographical locations analyzed in Study 2 and for student data from Study 1. For 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 << 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, . . . 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

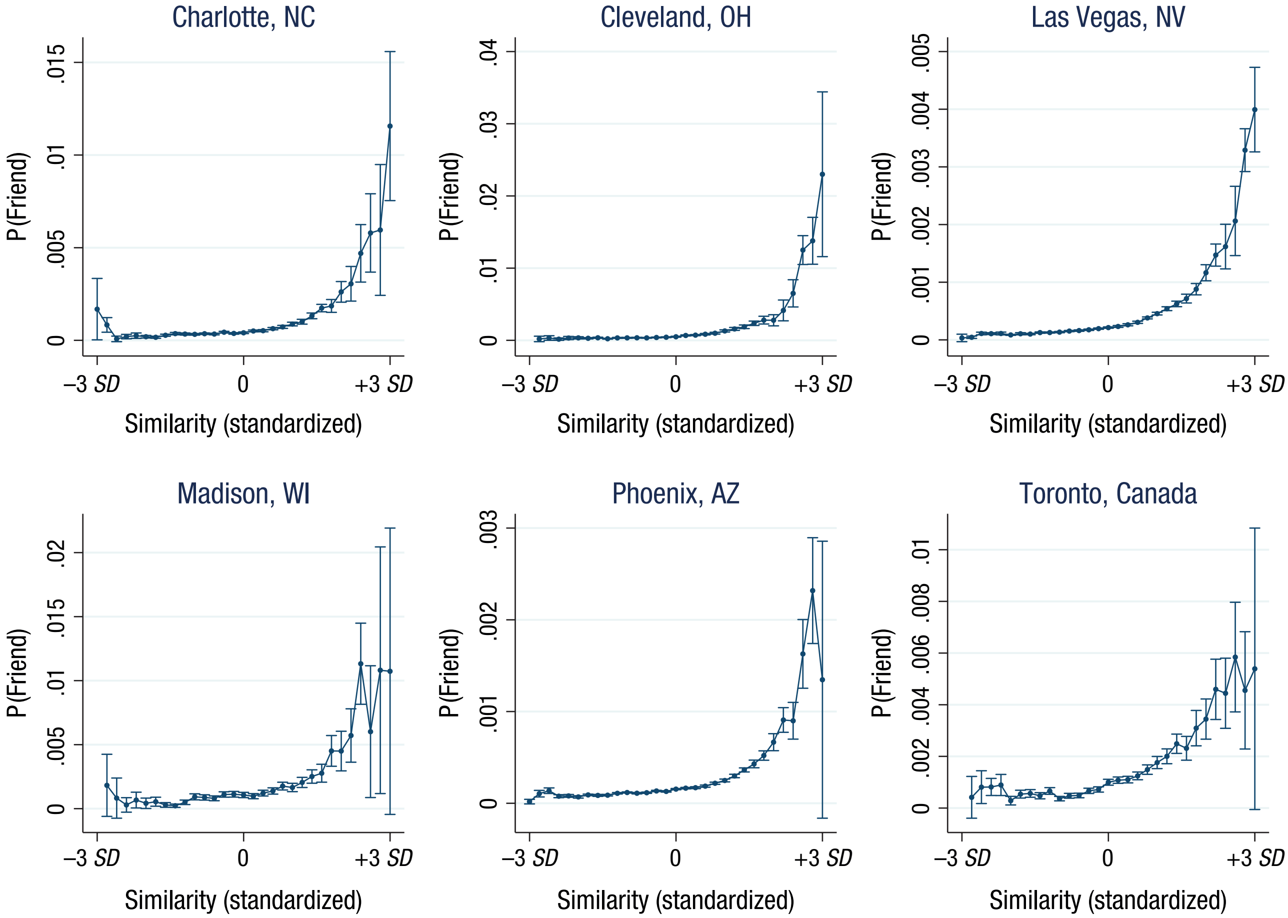


Fig. 3. Marginal effect in Study 2 of linguistic-style similarity on the probability of a friendship tie (dyad-level logistic regression with dummy variables at each 0.2 level of the standardized *z*-score dyadic linguistic-similarity measure). The data set contained 4,488,715 individuals and had 81,452 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: *OR*s = 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, & Murninghan, 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 altercentric 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.

SUPPORTING ONLINE MATERIAL

TABLE OF CONTENTS

General Supporting Material

Table S1. Linguistic dimensions corresponding to pronoun and grammar used in the LIWC analysis
A note on dyadic data structure and multi-way clustering

Study 1

Study 1 – General notes
Table S2. Examples of linguistic similarities (Study 1)
Table S3. Descriptive statistics for Study 1
Figure S1. Distribution of the standardized dyadic linguistic similarity in the student exams
Table S4. Ex ante linguistic similarity predicts tie formation (Models 1 and 2); and friendship predicts further increases over time in linguistic similarity (Models 3 and 4)
Syntax used to estimate Study 1 models

Study 2

Table S5. Descriptive statistics for Study 2
Table S6. Ex ante linguistic similarity predicts tie formation in Study 2
Table S7. Dyad-level linguistic convergence model in Yelp data
Figure S2. Exploratory analysis of linguistic cues
Table S8. Exploratory regressions of linguistic convergence in the 18 individual dimensions of language use
Table S9. Yelp analyses with inferred gender controls
Table S10. Robustness check: Including Yelp business fixed effects
Syntax used to estimate Study 2 models

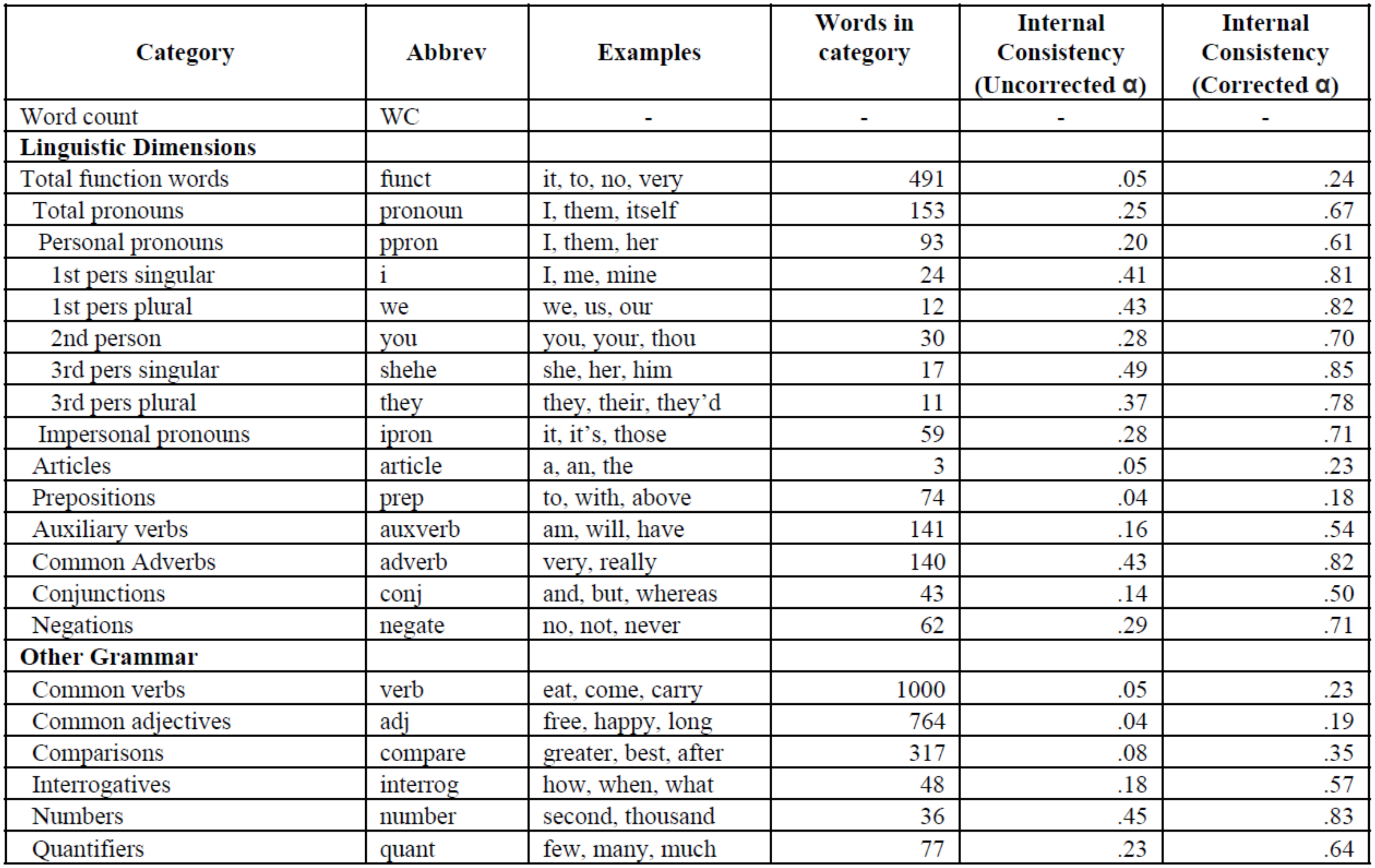
Additional analyses (Studies 1 and 2)

Figure S3: The robustness of the results to alternative measures of linguistic similarity (LSA, Stylo, LIWC-LSM)
Additional analyses: Empirical analysis of network fragmentation (Studies 1 and 2)

Extension: Simulated Consequences

References cited in the Supporting Material

Table S1. Linguistic dimensions corresponding to pronoun and grammar used in the main set of LIWC analysis. From page 3 of Pennebaker, Boyd, Jordan, and Blackburn (2015).



Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word count	WC	-	-	-	-
Linguistic Dimensions					
Total function words	funct	it, to, no, very	491	.05	.24
Total pronouns	pronoun	I, them, itself	153	.25	.67
Personal pronouns	ppron	I, them, her	93	.20	.61
1st pers singular	i	I, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
Impersonal pronouns	ipron	it, it's, those	59	.28	.71
Articles	article	a, an, the	3	.05	.23
Prepositions	prep	to, with, above	74	.04	.18
Auxiliary verbs	auxverb	am, will, have	141	.16	.54
Common Adverbs	adverb	very, really	140	.43	.82
Conjunctions	conj	and, but, whereas	43	.14	.50
Negations	negate	no, not, never	62	.29	.71
Other Grammar					
Common verbs	verb	eat, come, carry	1000	.05	.23
Common adjectives	adj	free, happy, long	764	.04	.19
Comparisons	compare	greater, best, after	317	.08	.35
Interrogatives	interrog	how, when, what	48	.18	.57
Numbers	number	second, thousand	36	.45	.83
Quantifiers	quant	few, many, much	77	.23	.64

Note that the categories "Total function words", "Total pronouns" and "Personal pronouns" represent aggregations of sub-categories; we therefore did not use them in our analyses.

A note on dyadic data structure and multiway clustering

Our data are dyadic, meaning that they are structured to include one observation for each pairwise combination of people in the sample. In the student sample, this means that the 247 students about whom we have complete data comprise 30,381 (= ½ × 247 × 246) undirected dyadic observations. This means that each of the 247 students for whom we have complete data participates in 246 dyadic observations – one with each of her classmates. (The one-half in the equation acknowledges that our dyads are undirected, so the *i-j* tie is identical to the *j-i* tie and we do not include both.) These are the observations in our regressions.

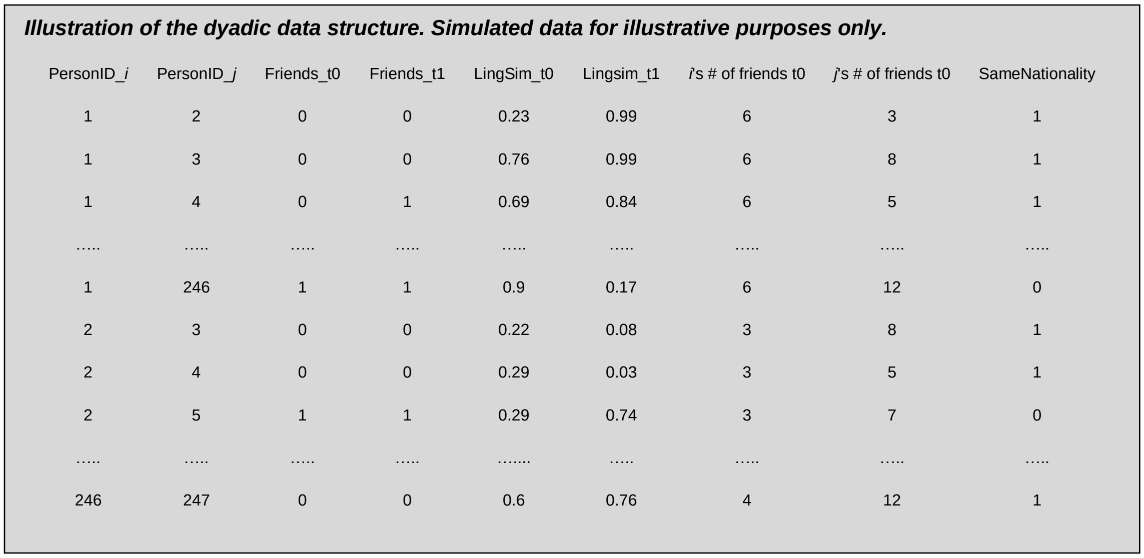


Illustration of the dyadic data structure. Simulated data for illustrative purposes only.

PersonID_*i*	PersonID_*j*	Friends_t0	Friends_t1	LingSim_t0	Lingsim_t1	*i*'s # of friends t0	*j*'s # of friends t0	SameNationality
1	2	0	0	0.23	0.99	6	3	1
1	3	0	0	0.76	0.99	6	8	1
1	4	0	1	0.69	0.84	6	5	1
…..	…..	…..	…..	…..	…..	…..	…..	…..
1	246	1	1	0.9	0.17	6	12	0
2	3	0	0	0.22	0.08	3	8	1
2	4	0	0	0.29	0.03	3	5	1
2	5	1	1	0.29	0.74	3	7	0
…..	…..	…..	…..	…….	…..	…..	…..	…..
246	247	0	0	0.6	0.76	4	12	1

Models predicting the existence of a dyadic friendship tie were estimated using logistic regression. As mentioned above, each possible pair of individuals is entered as an observation, and the dependent variable is the presence (1) or absence (0) of a social network tie between the pair of people. The main independent variable here is the similarity in linguistic style between the two individuals in the dyad in the prior time period. We controlled for *i* and *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 Model for the full list of covariates). Formally,

$$\boldsymbol{E}\left[friendship_{ijt_2}\right] = \beta_0 + \beta_1 linguistic\ similarity_{\mathrm{ijt}_1} + \beta_2 \mathrm{X}_{ij} + \varepsilon$$

where X$_{ij}$ is a vector of dyadic control variables including measures of i's and j's baseline propensities to form network ties and the dyadic similarity between *i* and *j* along demographic and personality dimensions. We control for i's and j's baseline propensities to form network ties by including covariates for their degree scores (i.e., their total number of friends) in the relevant time period. For example, when modeling friendship ties at Time 2, we control for Ego's Degree (time 2) and Alter's Degree (time 2); when modeling the change in linguistic similarity as a function of ongoing friendship (i.e., friendship at both time 1 and time 2), we control for each actor's degree in each time period.

To capture the alternative causal mechanism, we model the dyadic change in linguistic style similarity as a function of friendship and controlling for prior linguistic similarity, using ordinary least squares:

$$\mathit{\Delta LS}_{ij} = \beta_0 + \beta_1 tie_{ij} + \beta_2 linguistic\ similarity_{\mathrm{ijt_1}} + \beta_3 \mathrm{X}_{ij} + \varepsilon$$

where $\mathit{\Delta LS}_{ij}$ is the change in linguistic similarity between *i* and *j* from time 1 to time 2, standardized across the population of dyads; tie_{ij} is a binary indicator for whether (1) or not (0) a reciprocated network tie existed between *i* and *j* at both time 1 and time 2.

A regression in which 247 observations contain the same person has serious non-nested structural autocorrelation that must be accounted for in order to estimate consistent standard errors; failing to account for such dependencies will result in erroneous standard errors that appear to be smaller than the correct standard errors, creating the false impression of statistical precision. We correct for these dependencies using multiway clustering, a simple extension of conventional clustered standard errors. One-way clustering is a standard extension to multiple regression in which the assumption of independent observations is relaxed to permit observations that are correlated within explicitly specified groups, but are independent across groups (see, for example, Milligan 1980). When the clustering variable is well-specified and properly accounted for, regression analysis can yield consistent standard errors in the presence of structural dependencies. Multi-way clustering extends this logic to situations in which there are multiple, non-nested sources of dependence between observations (Cameron, Gelbach, & Miller, 2011; Egger & Tarlea, 2015; Kleinbaum, Stuart, & Tushman, 2013; Lindgren, 2010). In our data, dyadic observations exhibit non-independence related to common person effects (Kenny, Kashy, & Cook, 2006) across both members of the dyad. Because dyads are not nested, we cannot use hierarchical linear models, but two-way clustering offers an effective and parsimonious solution to the network autocorrelation problem in dyadic data that is well-established in the econometrics (Angrist & Pischke, 2008; Cameron et al., 2011) and social networks (Kleinbaum, Stuart & Tushman 2013; Lindgren 2010) literatures.

The simple (i.e., one-way) clustered covariance matrix is:

$$\hat{\Omega}_{cl} = (X'X)^{-1}\left(\sum_g X_g \hat{\Psi}_g X_g\right)(X'X)^{-1}$$

where

$$\hat{\Psi}_g = a\hat{e}_g\hat{e}'_g$$

$$= a\begin{bmatrix} \hat{e}_{1g}^2 & \cdots & \hat{e}_{1g}\hat{e}_{n_g g} \\ \vdots & \ddots & \vdots \\ \hat{e}_{1g}\hat{e}_{n_g g} & \cdots & \hat{e}_{n_g g}^2 \end{bmatrix}$$

Here, X_g is a matrix of regressors for group *g* (i.e., the common-person effect across dyadic observations) and *a* is a degrees of freedom adjustment factor. The degrees of freedom in a multiway clustered model is the number of grouping variables (i.e., the number of students) minus one (Angrist and Pischke 2009, chapter 8).

Two-way clustering repeats this procedure three separate times to create three separate cluster-robust variance matrices: one that clusters on person *i*, one that clusters on person *j*, and one that clusters on the intersection of matrices *i* and *j*. The two-way cluster-robust variance matrix used to estimate our standard errors is then calculated, very simply, as the sum of the first two matrices minus the third matrix. Cameron et al. (2011) show and Lindgren (2010) independently validates that by using this approach, we can account for the common-person dependencies across dyadic observations and achieve consistent standard errors in analyses of the non-nested dyadic data used to model social networks. Since its development, this method has been widely used in research on social networks from numerous disciplines, including sociology (Greenberg & Fernandez, 2016; Liu & Srivastava, 2015), organization theory (Dahlander & McFarland, 2013), economics (Andersen, 2018), and psychology (Feiler & Kleinbaum, 2015).

The need to account for dyadic autocorrelation also complicates the estimation of confidence intervals around our effect size estimates. The effect sizes in our selection models are odds-ratios, calculated by exponentiating the coefficients from a logistic regression with multi-way clustering. Similarly, the correct point estimates of the upper and lower bounds of the 95% confidence interval around the odds ratio can be obtained by exponentiating the upper and lower bounds of the confidence interval around the coefficient (but not by exponentiating the standard error).

The effect size estimates in our convergence models are partial correlations (r). Partial correlations are easily obtained in most statistical packages, including Stata. However, to correct the confidence intervals around these partial correlations to properly account for dyadic structural autocorrelation, we had to manually re-estimate the partial correlations. To do this, we regressed our dependent variable (change in linguistic similarity) on a vector of control variables and stored the standardized residuals; regressed our covariate of interest (network tie status) on the vector of control variables and stored the standardized residuals; then regressed – using multiway clustering – the first set of residuals on the second. The result was the same partial correlation value reported by Stata, but with a 95% confidence interval that correctly accounts for dyadic autocorrelation.

Study 1 – General notes

Illustrative Examples of Texts. Table S2 shows the first ~60 de-identified words of two student responses to the same essay question, along with selected LIWC analyses. A quick read of these texts reveals striking differences between their authors: whereas the first person seems to be very oriented to his own, internal vision for his future, the second is much more other-directed, basing her vision on interactions and external technology trends. These differences, which are readily apparent to the human reader, are equally apparent in the quantitative analysis: in these brief excerpts, Person 1 used more than twice as many "I-words," whereas Person 2 used more than three times as many "social words" and much more "analytic" language. Consistent with these differences, the analysis of their full texts reveals that the linguistic distance between these students is more than two standard deviations larger than the mean linguistic distance between dyads.

Network Survey Instrument. In each survey, we asked: "Consider the people with whom you like to spend your free time. Since you arrived at [university name], who are the classmates you have been with most often for informal social activities, such as going out to lunch, dinner, drinks, films, visiting one another's homes, exercising together, and so on?" The first network survey was conducted in late September, one month after the start of classes (and one week prior to the exam that provided our second corpus of textual data); the second was conducted in February. We define a binary network tie to exist when both parties to a relation report its existence (i.e., network ties are reciprocal).

A Note on the Excluded Observations. Out of the 285 students in the cohort, five students declined to complete one or more surveys. Thirteen students applied to dual-degree programs while already matriculated in other graduate schools and were not required to write the standard application essays. Twenty-two students (two of whom were previously excluded from the sample) applied through an underrepresented minority applicant program that used application essays different from the school's standard essays. Textual data at time 1 were not available for these students, who were therefore omitted from the sample. Our final sample included all of the remaining 247 students, comprising 30,381 (= ½ × 247 × 246) complete dyadic observations. In some analyses that required fewer variables, fewer observations had to be dropped.

Demographic Refinements. For our student sample, we have extensive demographic data, allowing us to construct and control for a broad range of dyadic attributes. Some of these dyadic attributes are nested – for example, we can control for "Same Gender" and can split out the same gender effect by gender by adding a control variable for "Both Men" (with "Both Women" as the comparison group). Because the sample is roughly equally divided between men and women, the correlation between "Same Gender" and "Both Men" is moderate. However, there are relatively few students from any single non-U.S. country and, as a result, the "Same Citizenship" variable is highly correlated with the "Both U.S. Citizens" variable; only 361 dyads share the same non-U.S. citizenship.

Because of this high correlation, we re-estimated all models without the "Both U.S. Citizens" variable and found results that were substantially identical. Because we are sometimes able to estimate precise effects of both "Same Citizenship" and "Both U.S. Citizens" and because including both covariates does not seem to make our estimates of interest less stable, our primary results include this variable.

Table S2. Examples of linguistic similarities (Study 1)

	Person 1	**Person 2**
Initial lines	"Ten years from now, I envision myself leading an operations team through a production floor. I'll be wearing my steel toe boots and doing work on the ground with the team, because that's the type of operations head I plan to be. What motivates me most is knowing that my work directly adds value by streamlining processes and increasing efficiency to maximize outputs."	"As I started learning about [the school], I was disappointed that I had missed [someone]'s visit in 2013. During his talk, he equated the coming advances in information technology to the Enlightenment and highlighted the opportunities that lie before our generation to unlock Big Data's potential and fundamentally change the way information will be shared for generations to come."
Word Count	63	59
I-words	11.1	5
Social words	3.2	11.7
Analytic words	74.9	89

Table S3. Descriptive statistics for Study 1. All of these values are based on dyads. N = 30,381 dyadic observations.

	Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
(1) Recip Tie (time 1)	0.03	0.17	1																				
(2) Recip Tie (time 2)	0.08	0.27	0.345	1																			
(3) Linguistic Similarity (time 1)	0	1	0.016	0.00	1																		
(4) Linguistic Similarity (time 2)	0	1	0.005	0.00	0.098	1																	
(5) Ego's Degree (time 1)	12.7	9.8	0.133	0.13	0.004	-0.042	1																
(6) Alter's Degree (time 1)	11.8	8.6	0.142	0.11	0.016	-0.047	0.000	1															
(7) Ego's Degree (time 2)	22.5	14.9	0.095	0.20	-0.042	-0.028	0.661	0.001	1														
(8) Alter's Degree (time 2)	22.8	14.0	0.073	0.19	-0.006	-0.057	-0.008	0.535	-0.004	1													
(9) Same Gender	0.51	0.50	0.051	0.08	0.005	0.000	-0.034	0.000	-0.015	0.018	1												
(10) Both Male	0.32	0.47	-0.014	0.04	0.010	-0.010	-0.111	-0.048	-0.052	0.061	0.677	1											
(11) Same Ethnicity	0.36	0.48	0.078	0.11	0.010	0.052	-0.030	0.051	0.012	0.043	0.019	-0.020	1										
(12) Both White	0.24	0.43	0.076	0.10	0.016	0.104	0.081	0.135	0.111	0.117	-0.013	-0.110	0.737	1									
(13) Same Citizenship	0.44	0.50	0.107	0.13	0.059	0.099	0.184	0.151	0.144	0.079	-0.015	-0.098	0.434	0.669	1								
(14) Both U.S. Citizens	0.43	0.50	0.085	0.10	0.054	0.100	0.194	0.158	0.152	0.086	-0.016	-0.098	0.415	0.686	0.977	1							
(15) Same Class Section	0.25	0.43	0.046	0.072	-0.003	-0.001	0.001	-0.002	0.001	-0.001	-0.003	-0.002	-0.006	-0.002	-0.005	-0.003	1						
(16) Same Study Group	0.02	0.13	0.119	0.128	-0.001	-0.007	0.000	-0.002	0.004	-0.004	-0.022	-0.011	-0.020	-0.009	-0.016	-0.013	0.231	1					
(17) Honesty/Humility Similarity	0	1	0.006	0.01	0.017	0.057	-0.001	-0.005	-0.011	-0.012	-0.002	0.003	0.064	0.081	0.017	0.015	0.006	0.008	1				
(18) Emotionality Similarity	0	1	0.016	0.04	-0.046	0.005	0.034	0.006	0.065	0.033	0.124	0.065	-0.019	-0.036	-0.018	-0.019	0.002	-0.001	0.090	1			
(19) eXtraversion Similarity	0	1	0.034	0.06	-0.014	0.006	0.018	0.042	0.030	0.038	-0.008	-0.037	0.045	0.069	0.058	0.057	-0.007	-0.002	0.025	-0.009	1		
(20) Agreeableness Similarity	0	1	0.027	0.02	-0.006	0.010	0.040	0.016	0.078	0.059	-0.007	-0.023	0.001	-0.029	-0.041	-0.043	0.002	-0.006	0.061	0.002	0.030	1	
(21) Conscientiousness Similarity	0	1	0.004	0.01	-0.021	0.019	-0.048	0.045	-0.036	0.010	-0.008	-0.021	0.047	0.054	0.049	0.045	-0.003	0.002	0.032	0.027	0.015	0.042	1
(22) Openness Similarity	0	1	0.002	0.00	0.035	-0.003	-0.019	-0.030	-0.042	-0.001	-0.002	0.006	-0.030	-0.032	-0.007	-0.008	-0.004	-0.010	0.025	-0.022	0.008	0.048	-0.002

Figure S1. Distribution of the standardized dyadic linguistic similarity in the student exams (Study 1).

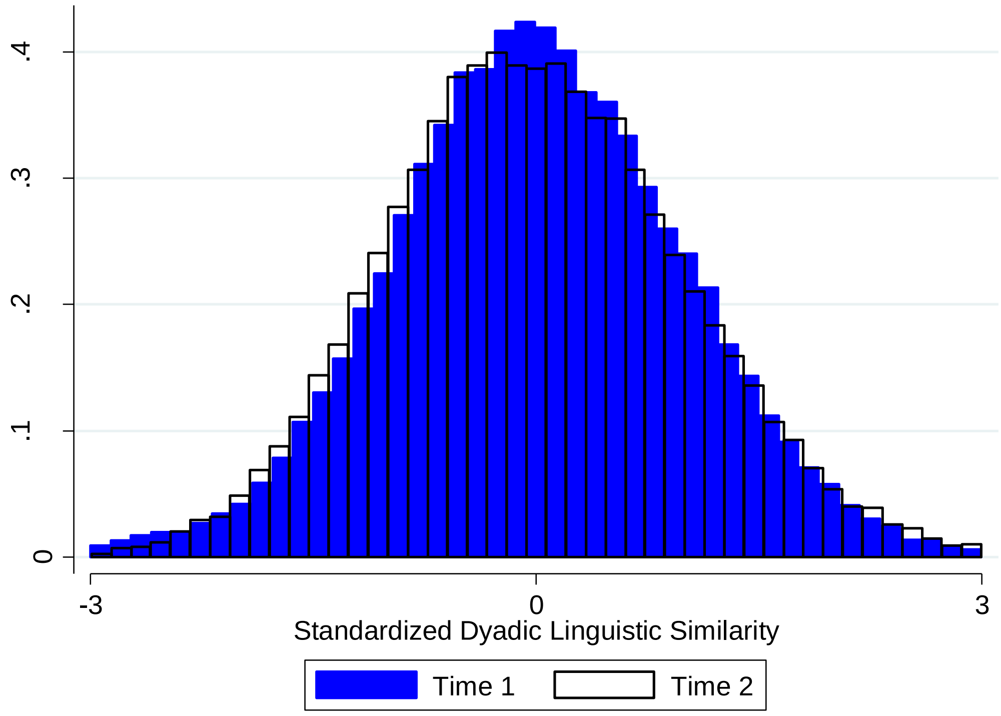


Table S4. Ex ante linguistic similarity predicts tie formation (Models 1 and 2); and friendship predicts further increases over time in linguistic similarity (Models 3 and 4).

Model 1: Parameter estimates from dyad-level selection model on student data with minimal control variables. Logistic regression, where the dependent variable is a network tie in time 2 (values of 0 indicate “not friends” and values of 1 indicate “friends”). N = 30,381 dyads (245 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Linguistic Similarity (time 1)	0.0791	0.0246	[0.0309, 0.1273]	3.2153	1.0823
Ego’s Degree (time 2)	0.0471	0.0030	[0.0412, 0.0530]	15.7131	1.0482
Alter’s Degree (time 2)	0.0488	0.0033	[0.0423, 0.0553]	14.6985	1.0500
Intercept	-4.9831	0.1329	[-5.2436, -4.7226]	-37.4914	0.0069

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 2: Parameter estimates from dyad-level selection model on student data with full control variables. Logistic regression, where the dependent variable is a network tie in time 2 (values of 0 indicate “not friends” and values of 1 indicate “friends”). N = 30,381 dyads (245 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Linguistic Similarity (time 1)	0.0499	0.0242	[0.0025, 0.0973]	2.0636	1.0512
Ego’s Degree (time 2)	0.0524	0.0030	[0.0465, 0.0582]	17.6004	1.0537
Alter’s Degree (time 2)	0.0531	0.0033	[0.0468, 0.0595]	16.3031	1.0546
Same Gender	0.7511	0.0903	[0.5742, 0.9281]	8.3208	2.1194
Both Male	-0.1016	0.0806	[-0.2596, 0.0564]	-1.2600	0.9034
Same Race/Ethnicity	0.8656	0.1453	[0.5808, 1.1505]	5.9566	2.3765
Both White	-0.4547	0.1583	[-0.7650, -0.1445]	-2.8730	0.6346
Same Citizenship	3.1784	0.2072	[2.7722, 3.5845]	15.3376	24.007
Both U.S. Citizens	-2.7083	0.2112	[-3.1223, -2.2944]	-12.8235	0.0666
Same Class Section	0.4497	0.0625	[0.3272, 0.5722]	7.1951	1.5678
Same Study Group	2.1738	0.1488	[1.8822, 2.4655]	14.6082	8.7917
Honesty/Humility Similarity	0.0306	0.0280	[-0.0242, 0.0854]	1.0929	1.0310
Emotionality Similarity	0.0558	0.0289	[-0.0008, 0.1125]	1.9312	1.0574
eXtraversion Similarity	0.1486	0.0348	[0.0804, 0.2169]	4.2700	1.1602
Agreeableness Similarity	-0.0226	0.0232	[-0.0682, 0.0229]	-0.9739	0.9776
Conscientiousness Similarity	-0.0193	0.0238	[-0.0659, 0.0274]	-0.8094	0.9809
Openness Similarity	0.0946	0.0219	[0.0517, 0.1375]	4.3229	1.0992
Intercept	-6.4533	0.1661	[-6.7788, -6.1278]	-38.8582	0.0016

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 3: Parameter estimates from dyad-level tie persistence model on student data with minimal control variables. The model is conditioned on the presence of a network tie at Time 1; the DV is a network tie at Time 2 (where 0="not friends" and 1="friends"). N = 1,232 dyads (230 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Linguistic Similarity (time 2)	0.1007	0.0597	[-0.0163, 0.2177]	1.6876	1.1060
Ego's Degree (time 2)	0.0230	0.0061	[0.0111, 0.0349]	3.7815	1.0232
Alter's Degree (time 2)	0.0300	0.0063	[0.0177, 0.0424]	4.7643	1.0305
Intercept	-1.1608	0.2292	[-1.6100, -0.7115]	-5.0638	0.3133

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 4: Parameter estimates from dyad-level tie persistence model on student data with full control variables. The model is estimated on the subsample of dyads that were friends in Time 1; the DV is a network tie at Time 2 (where 0="not friends" and 1="friends"). N = 1,232 dyads (230 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Linguistic Similarity (time 2)	0.1292	0.0639	[0.0040, 0.2544]	2.0232	1.1379
Ego's Degree (time 2)	0.0277	0.0060	[0.0158, 0.0395]	4.5786	1.0281
Alter's Degree (time 2)	0.0322	0.0065	[0.0194, 0.0449]	4.9402	1.0327
Same Gender	0.3516	0.1678	[0.0228, 0.6805]	2.0958	1.4213
Both Male	0.3215	0.2195	[-0.1088, 0.7517]	1.4645	1.3792
Same Race/Ethnicity	0.5139	0.2586	[0.0069, 1.0208]	1.9868	1.6718
Both White	-0.3534	0.3003	[-0.9420, 0.2353]	-1.1765	0.7023
Same Citizenship	0.3902	0.3661	[-0.3273, 1.1078]	1.0659	1.4773
Both U.S. Citizens	-0.2834	0.3938	[-1.0553, 0.4884]	-0.7197	0.7532
Same Class Section	0.2653	0.1450	[-0.0188, 0.5495]	1.8300	1.3038
Same Study Group	0.5673	0.6086	[-0.6256, 1.7602]	0.9321	1.7635
Honesty/Humility Similarity	-0.0374	0.0727	[-0.1799, 0.1051]	-0.5144	0.9633
Emotionality Similarity	0.1055	0.0724	[-0.0364, 0.2474]	1.4577	1.1113
eXtraversion Similarity	0.0660	0.0803	[-0.0914, 0.2233]	0.8215	1.0682
Agreeableness Similarity	-0.0780	0.0720	[-0.2192, 0.0631]	-1.0838	0.925
Conscientiousness Similarity	0.0302	0.0830	[-0.1324, 0.1928]	0.3641	1.0307
Openness Similarity	0.1762	0.0592	[0.0602, 0.2923]	2.9765	1.1927
Intercept	-1.9909	0.2975	[-2.5739, -1.4079]	-6.6932	0.1366

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 5: Parameter estimates from dyad-level convergence model on student data with minimal control variables. N = 30,381 dyads (245 df).

Predictor	*b*	*SE*	95% CI	*t(245)*	*r*
Network Tie (time 1 & 2)	0.1292	0.0506	[0.0296, 0.2289]	2.5547	0.0184
Linguistic Similarity (time 1)	-0.9022	0.0266	[-0.9547, -0.8497]	-33.8579	-0.6733
Ego’s Degree (time 1)	-0.0049	0.0049	[-0.0145, 0.0047]	-1.0032	-0.0373
Alter’s Degree (time 1)	-0.0031	0.0046	[-0.0122, 0.0060]	-0.6673	-0.0233
Ego’s Degree (time 2)	0.0005	0.0036	[-0.0066, 0.0076]	0.1378	0.0055
Alter’s Degree (time 2)	-0.0030	0.0027	[-0.0084, 0.0023]	-1.1134	-0.0368
Intercept	0.1465	0.1019	[-0.0542, 0.3472]	1.4380	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; r = partial correlation.

Model 6: Parameter estimates from dyad-level convergence model on student data with full control variables. N = 30,381 dyads (245 df).

Predictor	*b*	*SE*	95% CI	*t(245)*	*r*
Network Tie (time 1 & 2)	0.1078	0.0533	[0.0027, 0.2128]	2.0211	0.0153
Linguistic Similarity (time 1)	-0.9067	0.0257	[-0.9574, -0.8560]	-35.2540	-0.6774
Ego’s Degree (time 1)	-0.0064	0.0047	[-0.0157, 0.0029]	-1.3565	-0.0483
Alter’s Degree (time 1)	-0.0053	0.0048	[-0.0147, 0.0041]	-1.1080	-0.0395
Ego’s Degree (time 2)	-0.0004	0.0036	[-0.0075, 0.0067]	-0.1229	-0.005
Alter’s Degree (time 2)	-0.0036	0.0027	[-0.0090, 0.0018]	-1.3223	-0.0432
Same Gender	-0.0040	0.0513	[-0.1050, 0.0970]	-0.0775	-0.0015
Both Male	0.0050	0.1003	[-0.1924, 0.2025]	0.0502	0.0017
Same Race/Ethnicity	-0.1420	0.0528	[-0.2461, -0.0380]	-2.6891	-0.0457
Both White	0.2815	0.1094	[0.0660, 0.4970]	2.5732	0.0656
Same Citizenship	0.0459	0.0926	[-0.1366, 0.2284]	0.4957	0.0048
Both U.S. Citizens	0.0809	0.1232	[-0.1618, 0.3235]	0.6564	0.0082
Same Class Section	0.0007	0.0088	[-0.0166, 0.0180]	0.0790	0.0003
Same Study Group	-0.0624	0.0381	[-0.1374, 0.0127]	-1.6373	-0.008
Honesty/Humility Similarity	0.0440	0.0231	[-0.0014, 0.0895]	1.9075	0.045
Emotionality Similarity	0.0133	0.0199	[-0.0258, 0.0524]	0.6691	0.0132
eXtraversion Similarity	0.0014	0.0201	[-0.0382, 0.0409]	0.0680	0.0014
Agreeableness Similarity	0.0202	0.0207	[-0.0205, 0.0609]	0.9775	0.0203
Conscientiousness Similarity	0.0114	0.0231	[-0.0341, 0.0570]	0.4946	0.0116
Openness Similarity	-0.0086	0.0237	[-0.0553, 0.0381]	-0.3613	-0.0088
Intercept	0.1600	0.1016	[-0.0402, 0.3601]	1.5745	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; r = partial correlation.

Syntax used to estimate Study 1 models

The commands were estimated using Stata 15.1

```
* Tie Formation Models

clus_nway logit reciptie_2 LingSim_T1_z ego_recipdeg_2
alter_recipdeg_2, vce(cluster i_int j_int)

clus_nway logit reciptie_2 LingSim_T1_z ego_recipdeg_2
alter_recipdeg_2 samegender bothmale sameethnicity bothwhite
samecitizenship bothus samefallsection samefallsg sim_H_z
sim_E_z sim_X_z sim_A_z sim_C_z sim_O_z, vce(cluster i_int
j_int)

clus_nway logit reciptie_2 LingSim_T2_z ego_recipdeg_2
alter_recipdeg_2 if reciptie_1==0, vce(cluster i_int j_int)

clus_nway logit reciptie_2 LingSim_T2_z ego_recipdeg_2
alter_recipdeg_2 samegender bothmale sameethnicity bothwhite
samecitizenship bothus samefallsection samefallsg sim_H_z
sim_E_z sim_X_z sim_A_z sim_C_z sim_O_z if reciptie_1==0,
vce(cluster i_int j_int)

* Tie Decay Models

clus_nway logit reciptie_2 LingSim_T2_z ego_recipdeg_2
alter_recipdeg_2 if reciptie_1, vce(cluster i_int j_int)

clus_nway logit reciptie_2 LingSim_T2_z ego_recipdeg_2
alter_recipdeg_2 samegender bothmale sameethnicity bothwhite
samecitizenship bothus samefallsection samefallsg sim_H_z
sim_E_z sim_X_z sim_A_z sim_C_z sim_O_z if reciptie_1,
vce(cluster i_int j_int)

* Linguistic Convergence Models

clus_nway reg dLingSim_z reciptie LingSim_T1_z ego_recipdeg_1
alter_recipdeg_1 ego_recipdeg_2 alter_recipdeg_2 , vce(cluster
i_int j_int)

clus_nway reg dLingSim_z reciptie LingSim_T1_z ego_recipdeg_1
alter_recipdeg_1 ego_recipdeg_2 alter_recipdeg_2 samegender
bothmale sameethnicity bothwhite samecitizenship bothus
samefallsection samefallsg sim_H_z sim_E_z sim_X_z sim_A_z
sim_C_z sim_O_z, vce(cluster i_int j_int)
```

Table S5. Descriptive statistics for Study 2. We note that the average Degree increased significantly from time 1 to time 2 in all Yelp metro areas, as in the student data of Study 1.

Metro area	**Reviews**	**Reviewers**	**Businesses**	**Avg. friend# time 1**	**Avg. friend# time 2**
Charlotte	113,837	11,696	12,694	13.33	64.01
Cleveland	85,204	9,371	9,072	16.51	72.54
Las Vegas	769,890	68,312	39,072	28.36	114.36
Madison	41,965	3,738	4,787	14.10	63.23
Phoenix	633,031	60,204	55,908	15.98	60.26
Toronto	92,541	4,457	17,409	30.39	139.33
Urbana Champaign	13,002	1,873	1,687	10.32	60.09
Total	1,749,470	159,651	140,629	14.02	71.71

Table S6. Ex ante linguistic similarity predicts tie formation in Study 2. Dyad-level logistic regression models were estimated on binary indicators of a network tie. The key covariates were interactions between metro area dummies and the linguistic similarity measure at Time 1.

Model 1: Dyad-level logistic regression, DV: Network tie (time 2) (where 0="not friends" and 1="friends"). Sample: N= 4,488,715 (81,452 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Charlotte	(baseline)				
Cleveland	0.0374	0.1315	[-0.2204, 0.2951]	0.2842	1.0333
Las Vegas	-1.2304	0.0788	[-1.3849, -1.0758]	-15.6041	0.3374
Madison	0.6315	0.1417	[0.3538, 0.9091]	4.4577	1.8856
Phoenix	-1.3013	0.078	[-1.4541, -1.1485]	-16.691	0.2772
Toronto	0.5416	0.1078	[0.3304, 0.7528]	5.0251	1.6939
Urbana Champaign	0.8102	0.1239	[0.5673, 1.0531]	6.538	2.2037
Charlotte × Linguistic Similarity (time 1)	0.4576	0.0121	[0.4339, 0.4812]	37.9694	1.5866
Cleveland × Linguistic Similarity (time 1)	0.4540	0.0131	[0.4283, 0.4796]	34.7344	1.5794
Las Vegas × Linguistic Similarity (time 1)	0.5568	0.0110	[0.5353, 0.5783]	50.7573	1.7623
Madison × Linguistic Similarity (time 1)	0.3727	0.0157	[0.3419, 0.4036]	23.6818	1.4470
Phoenix × Linguistic Similarity (time 1)	0.5199	0.0122	[0.4960, 0.5439]	42.5417	1.6996
Toronto × Linguistic Similarity (time 1)	0.3943	0.0127	[0.3695, 0.4191]	31.1658	1.4897
Urbana Champaign × Linguistic Similarity (time 1)	0.2303	0.0191	[0.1929, 0.2677]	12.0621	1.2610
Ego DegreeDegree (time 2)	0.0076	0.0012	[0.0052, 0.0101]	6.1579	1.0076
Alter DegreeDegree (time 2)	0.006	0.0011	[0.0039, 0.0082]	5.5149	1.0002
Intercept	-7.3166	0.0724	[-7.4585, -7.1748]	-101.1213	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 2: Dyad-level logistic regression, DV: Network tie (time 2) (where 0="not friends" and 1="friends"). Sample: Dyads that are not friend in time 1. N= 4,099,004 (81,419 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Charlotte	(baseline)				
Cleveland	-0.3086	0.2113	[-0.7229, 0.1056]	-1.4604	0.7477
Las Vegas	-1.1897	0.1569	[-1.4973, -0.8822]	-7.5823	0.3382
Madison	0.175	0.2642	[-0.3429, 0.6929]	0.6622	1.1643
Phoenix	-1.4884	0.1596	[-1.8011, -1.1757]	-9.3283	0.2237
Toronto	0.5152	0.2078	[0.1079, 0.9225]	2.479	1.6873
Urbana Champaign	-0.2371	0.2728	[-0.7717, 0.2976]	-0.869	0.7609
Charlotte × Linguistic Similarity (time 1)	0.4333	0.0142	[0.4055, 0.4611]	30.562	1.5477
Cleveland × Linguistic Similarity (time 1)	0.4129	0.0177	[0.3782, 0.4475]	23.348	1.5231
Las Vegas × Linguistic Similarity (time 1)	0.5002	0.0175	[0.4660, 0.5344]	28.6294	1.6762
Madison × Linguistic Similarity (time 1)	0.3443	0.0213	[0.3026, 0.3859]	16.2004	1.4144
Phoenix × Linguistic Similarity (time 1)	0.2671	0.0203	[0.2273, 0.3069]	13.1510	1.5262
Toronto × Linguistic Similarity (time 1)	0.3899	0.0206	[0.3496, 0.4302]	18.9572	1.4863
Urbana Champaign × Linguistic Similarity (time 1)	0.2796	0.0313	[0.2182, 0.3411]	8.9225	1.3244
Ego DegreeDegree (time 2)	0.007	0.0009	[0.0054, 0.0087]	8.2655	1.0070
Alter DegreeDegree (time 2)	0.0052	0.0009	[0.0035, 0.0070]	5.7836	1.0002
Intercept	-9.5523	0.1495	[-9.8453, -9.2593]	-63.8944	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Table S7. Dyad-level linguistic convergence model in Yelp data. Linear regression, the dependent variable is change in linguistic similarity from Time 1 to Time 2. N=4,408,493 (81,452 df).

Predictor	b	SE	95% CI	t(81,452)	r
Linguistic Similarity (time 1)	-0.5982	0.0029	[-0.6038, -0.5926]	-208.1832	-0.5328
Cleveland	0.1136	0.0161	[0.0821, 0.1451]	7.0693	0.0215
Las Vegas	-0.0377	0.0116	[-0.0604, -0.0149]	-3.246	-0.0088
Madison	0.0722	0.0237	[0.0258, 0.1185]	3.0503	0.0085
Phoenix	-0.0782	0.0117	[-0.1011, -0.0554]	-6.7058	-0.0212
Toronto	0.4384	0.0183	[0.4024, 0.4743]	23.9103	0.0766
Urbana Champaign	0.0615	0.031	[0.0007, 0.1223]	1.9827	0.0052
Charlotte × Network Tie (time 1&2)	0.2839	0.0323	[0.2206, 0.3471]	8.7919	0.0206
Cleveland × Network Tie (time 1&2)	0.3112	0.0793	[0.1558, 0.4667]	3.9234	0.0205
Las Vegas × Network Tie (time 1&2)	0.1676	0.0122	[0.1437, 0.1915]	13.7507	0.0658
Madison × Network Tie (time 1&2)	0.3154	0.0513	[0.2148, 0.4159]	6.1480	0.0096
Phoenix × Network Tie (time 1&2)	0.1702	0.0192	[0.1326, 0.2078]	8.8692	0.0426
Toronto × Network Tie (time 1&2)	0.5284	0.0406	[0.4488, 0.6080]	13.0087	0.024
Urbana Champaign × Network Tie (time 1&2)	0.1938	0.0558	[0.0845, 0.3031]	3.4745	0.0026
Ego Degree (time 1)	0.0001	0.0007	[-0.0012, 0.0015]	0.171	-0.0076
Alter Degree (time 1)	-0.0024	0.001	[-0.0043, -0.0004]	-2.3386	0.0017
Ego Degree (time 2)	0.0017	0.0007	[0.0004, 0.0029]	2.5418	0.0173
Alter Degree (time 2)	0.0037	0.0009	[0.0019, 0.0055]	4.0019	-0.0016
Intercept	-0.0335	0.0106	[-0.0543, -0.0126]	-3.1516	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; r = partial correlation

Study 2: Exploratory Analysis of Linguistic Cues

In additional analyses, we exploit the large size of the Yelp data to investigate which dimensions of linguistic similarity are most predictive of a friendship tie. To do this, we disaggregate the linguistic similarity measure into measures of similarity along each of the 18 LIWC dimensions, calculated analogously as the multiplicative inverse of the logged absolute dyadic difference. We then regressed these 18 linguistic similarity variables on the probability of having a tie (logistic regressions, with baseline tie probability as a control, standard errors clustered on each member of the dyad). Table S8 show the regression results and Figure S2 plots the effects, i.e., for each of the dimensions on the log odds of two members of a dyad becoming friends. To explore the robustness of the patterns, we estimated the effects independently for each metropolitan region in our dataset (as before, we do not plot the effects for Urbana Champaign because the estimates are noisy due to the small sample size). We find that similarity along linguistic dimensions "I" ("I", "me", "mine", etc.) and "we" ("us", "ours", etc.) appear to be the strongest, most consistent predictors of friendship. "Negations" (e.g., "not", "mustn't", "cannot"), "interrogatives" (e.g., "how", "what") and numbers ("seven", "ten", etc.) are also consistently positive and strong predictors. The homophily along the first-person singular pronoun usage is especially interesting: reviewers who like to talk about themselves are likely to be friends with others who also like to talk about themselves.

Figure S2. The effects of individual linguistic dimensions on tie formation. The dots represent point estimates and the lines represent multi-way cluster-robust 95% confidence intervals

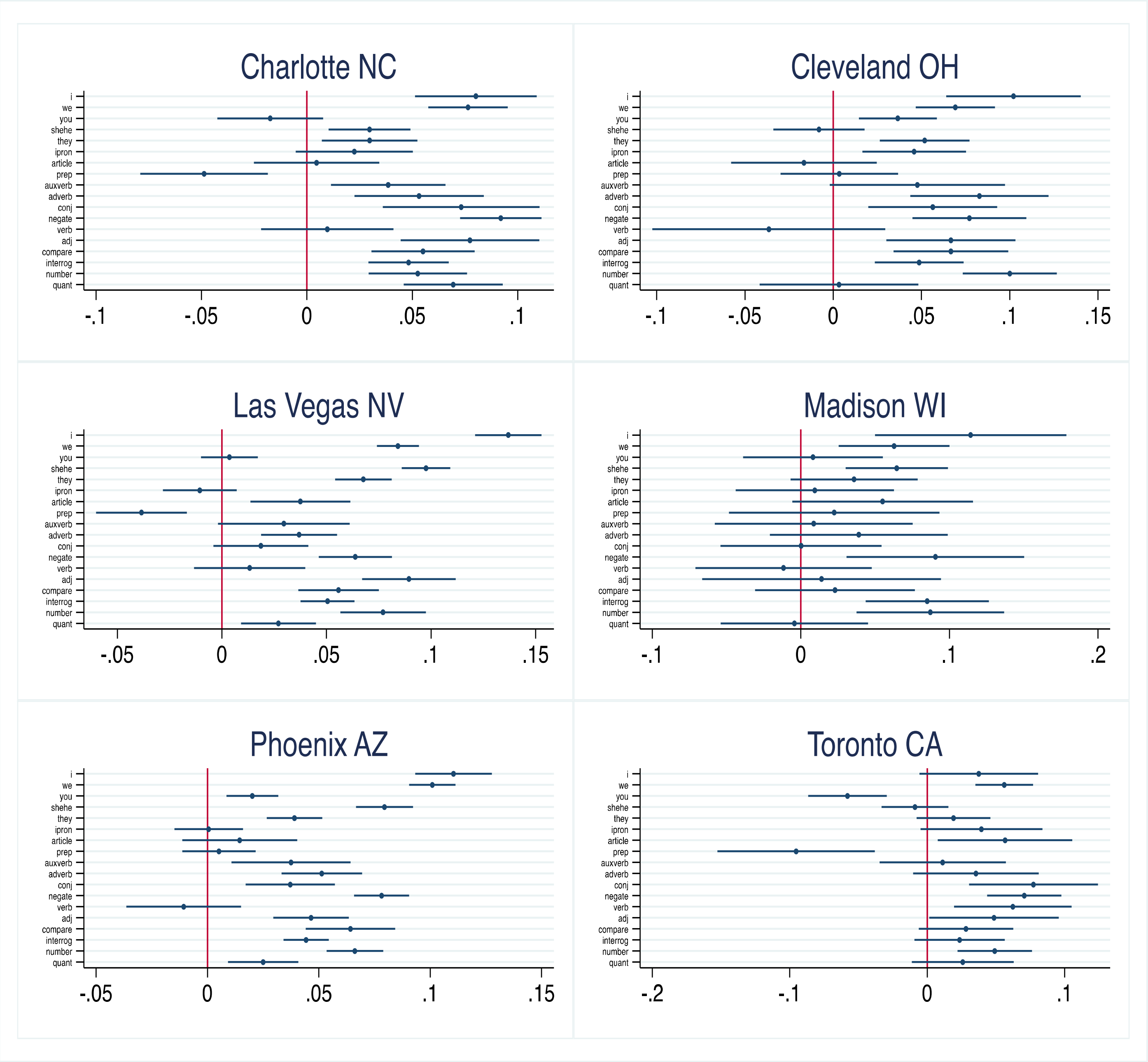


Table S8. Exploratory regressions of linguistic convergence in the 18 individual dimensions of language use. Separate models were estimated for each dimension in each metro area. Dyad-level linear regressions, the dependent variable was change in similarity in the use of one specific language category at Time 2; the key explanatory variable is a network tie; controls for Time 1 similarity and intercepts were estimated, but not shown.

		(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
				Las				**Urbana**
Metro Area		**Charlotte**	**Cleveland**	**Vegas**	**Madison**	**Phoenix**	**Toronto**	**Champaign**
i	b	0.3205	0.4443	0.371	0.3309	0.3048	0.4698	0.2362
	95% CI	[0.2693, 0.3717]	[0.3950, 0.4936]	[0.3410, 0.4009]	[0.2341, 0.4278]	[0.2751, 0.3345]	[0.4219, 0.5178]	[0.1437, 0.3288]
we	b	0.0502	0.049	0.0439	-0.0003	0.0041	0.3744	-0.0685
	95% CI	[0.0194, 0.0809]	[0.0147, 0.0832]	[0.0258, 0.0620]	[-0.0729, 0.0722]	[-0.0121, 0.0204]	[0.3262, 0.4227]	[-0.1649, 0.0278]
you	b	0.0215	0.0917	0.0245	-0.0269	-0.0201	0.3021	-0.039
	95% CI	[-0.0133, 0.0562]	[0.0512, 0.1322]	[0.0042, 0.0447]	[-0.1009, 0.0470]	[-0.0421, 0.0019]	[0.2478, 0.3563]	[-0.1259, 0.0479]
shehe	b	-0.0696	-0.0510	0.0942	-0.0770	0.0067	0.1210	-0.0747
	95% CI	[-0.1040, -0.0352]	[-0.0849, -0.0170]	[0.0703, 0.1181]	[-0.1577, 0.0037]	[-0.0116, 0.0250]	[0.0780, 0.1641]	[-0.1544, 0.0050]
they	b	0.1724	0.2342	0.1803	0.1659	0.1573	0.3749	0.0132
	95% CI	[0.1345, 0.2103]	[0.1911, 0.2773]	[0.1594, 0.2012]	[0.0985, 0.2332]	[0.1356, 0.1791]	[0.3235, 0.4263]	[-0.0628, 0.0891]
ipron	b	0.3345	0.4274	0.343	0.3025	0.2967	0.4639	0.1401
	95% CI	[0.2784, 0.3906]	[0.3749, 0.4798]	[0.3140, 0.3721]	[0.2161, 0.3889]	[0.2662, 0.3273]	[0.4104, 0.5175]	[0.0362, 0.2441]
article	b	0.3623	0.3974	0.3540	0.2553	0.3101	0.4606	0.158
	95% CI	[0.3102, 0.4144]	[0.3421, 0.4528]	[0.3246, 0.3833]	[0.1608, 0.3497]	[0.2786, 0.3416]	[0.4041, 0.5171]	[0.0551, 0.2609]
prep	b	0.38	0.4015	0.3613	0.2964	0.3134	0.4317	0.1662
	95% CI	[0.3237, 0.4363]	[0.3500, 0.4530]	[0.3321, 0.3905]	[0.2148, 0.3780]	[0.2800, 0.3467]	[0.3777, 0.4858]	[0.0516, 0.2808]
auxverb	b	0.3658	0.4191	0.4087	0.3460	0.3296	0.4824	0.2027
	95% CI	[0.3167, 0.4148]	[0.3708, 0.4674]	[0.3790, 0.4385]	[0.2505, 0.4415]	[0.2977, 0.3615]	[0.4313, 0.5334]	[0.1206, 0.2849]
adverb	b	0.4123	0.4459	0.4051	0.3244	0.3214	0.4793	0.176
	95% CI	[0.3580, 0.4666]	[0.3911, 0.5008]	[0.3772, 0.4330]	[0.2055, 0.4433]	[0.2873, 0.3555]	[0.4259, 0.5327]	[0.0788, 0.2731]
conj	b	0.3851	0.4352	0.4046	0.3391	0.3235	0.4718	0.1431
	95% CI	[0.3324, 0.4378]	[0.3851, 0.4854]	[0.3754, 0.4338]	[0.2496, 0.4286]	[0.2903, 0.3567]	[0.4152, 0.5285]	[0.0502, 0.2360]
negate	b	0.3029	0.433	0.2859	0.2197	0.2443	0.5324	0.1005
	95% CI	[0.2571, 0.3486]	[0.3852, 0.4809]	[0.2610, 0.3108]	[0.1163, 0.3231]	[0.2192, 0.2694]	[0.4774, 0.5875]	[-0.0140, 0.2151]
verb	b	0.3626	0.4176	0.4160	0.3543	0.3324	0.4625	0.1818
	95% CI	[0.3123, 0.4129]	[0.3687, 0.4665]	[0.3870, 0.4450]	[0.2623, 0.4463]	[0.3003, 0.3644]	[0.4097, 0.5153]	[0.0775, 0.2860]
adj	b	0.4075	0.4326	0.3944	0.3444	0.3218	0.5157	0.1086
	95% CI	[0.3535, 0.4615]	[0.3816, 0.4836]	[0.3653, 0.4234]	[0.2596, 0.4293]	[0.2892, 0.3545]	[0.4592, 0.5722]	[-0.0110, 0.2281]
compare	b	0.3494	0.4353	0.2981	0.2657	0.2534	0.5068	0.0682
	95% CI	[0.2989, 0.3998]	[0.3847, 0.4858]	[0.2713, 0.3248]	[0.1664, 0.3651]	[0.2249, 0.2820]	[0.4499, 0.5637]	[-0.0312, 0.1676]

interrog	b	0.1517	0.2599	0.1463	0.1300	0.1256	0.4544	0.0368
	95% CI	[0.1152, 0.1882]	[0.2144, 0.3055]	[0.1231, 0.1695]	[0.0484, 0.2116]	[0.1021, 0.1492]	[0.3986, 0.5102]	[-0.0373, 0.1108]
number	b	0.2395	0.3182	0.2102	0.2520	0.1828	0.4601	0.1029
	95% CI	[0.1897, 0.2894]	[0.2692, 0.3671]	[0.1834, 0.2369]	[0.1885, 0.3155]	[0.1601, 0.2055]	[0.4043, 0.5159]	[0.0135, 0.1922]
quant	b	0.3701	0.4000	0.2961	0.2317	0.2333	0.4793	0.1762
	95% CI	[0.3221, 0.4182]	[0.3444, 0.4556]	[0.2695, 0.3227]	[0.1520, 0.3115]	[0.2014, 0.2653]	[0.4262, 0.5325]	[0.0598, 0.2927]

Note: CI = multiway-clustered confidence interval

Study 2 Robustness check: Controlling for the gender of the reviewers

Coding the Gender of Reviewers. Although the large scale of the Yelp data is appealing, it suffers for the lack of any information about the individual reviewers. Although we compensate for this limitation by pairing the Yelp data with the smaller, richer student data set, we nevertheless tried to infer individual attributes about Yelp reviewers. Specifically, by matching the profile names of the reviewers to the Social Security Gender Database. The Social Security Database contains the gender distribution of babies born in the U.S. since 1880. For each year, the dataset contains the number of boys and the number of girls born with that first name (if fewer than five people are born with a given first name in a given year, the name is omitted for privacy reasons). For example, in 1977 there were 5,818 male baby named Peter, while only 33 female babies called Peter. We therefore assign a 99.4% probability that a user with a first name Peter is male (5,815/(5,815+33)). There are 16,542 unique first names in the dataset. We matched these first names to the Yelp dataset, and we assign a gender to a reviewer if their gender can be inferred with at least 95% precision. With this procedure, we could infer the gender of 79.2% of the reviewers. As in Study 1, we added the gender of the reviewers as covariates in the models, yielding the set of results in Table S9. All our results still hold.

Table S9. Yelp analyses with inferred gender controls. Note that the number of observations is lower in these models because for 18% of the reviewers we could not unambiguously infer gender.

Model 1. Dyad-level logistic regression, DV: Network Tie (time 2). Sample: All dyads. N=3,011,271 (67,109 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Charlotte	(baseline)				
Cleveland	-0.0142	0.1417	[-0.2918, 0.2635]	-0.1000	0.9751
Las Vegas	-1.2904	0.0847	[-1.4564, -1.1244]	-15.2372	0.3182
Madison	0.5447	0.1597	[0.2316, 0.8578]	3.4099	1.7282
Phoenix	-1.3434	0.0847	[-1.5094, -1.1773]	-15.8562	0.2651
Toronto	0.5031	0.1198	[0.2683, 0.7379]	4.1991	1.6208
Urbana Champaign	0.8147	0.1358	[0.5487, 1.0808]	6.0013	2.1881
Charlotte × Linguistic Similarity (time 1)	0.4408	0.0125	[0.4163, 0.4652]	35.3216	1.5610
Cleveland × Linguistic Similarity (time 1)	0.4464	0.0142	[0.4185, 0.4742]	31.3834	1.5711
Las Vegas × Linguistic Similarity (time 1)	0.5521	0.0116	[0.5294, 0.5749]	47.6000	1.7607
Madison × Linguistic Similarity (time 1)	0.3587	0.0171	[0.3251, 0.3923]	20.9465	1.4278
Phoenix × Linguistic Similarity (time 1)	0.5123	0.0134	[0.4861, 0.5386]	38.2641	1.6925
Toronto × Linguistic Similarity (time 1)	0.3740	0.0138	[0.3470, 0.4009]	27.1820	1.4637
Urbana Champaign × Linguistic Similarity (time 1)	0.2272	0.0189	[0.1901, 0.2643]	12.0156	1.2568
Ego Degree (time 2)	0.0075	0.0013	[0.0050, 0.0100]	5.8847	1.0074
Alter Degree (time 2)	0.0058	0.0012	[0.0034, 0.0082]	4.6545	1.0002
Same Gender	-0.2207	0.0323	[-0.2840, -0.1574]	-6.8366	0.7557
Both Male	0.0889	0.0414	[0.0078, 0.1701]	2.1488	1.1757
Intercept	-7.1428	0.0786	[-7.2969, -6.9887]	-90.8573	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 2. Dyad-level logistic regression, DV: Friend (time 2). Sample: Dyads that are not friend in t1. N=2,796,814 (67,078 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Charlotte	(baseline)				
Cleveland	-0.4088	0.2206	[-0.8412, 0.0236]	-1.8529	0.6757
Las Vegas	-1.2268	0.1641	[-1.5484, -0.9051]	-7.4744	0.3252
Madison	0.1479	0.2994	[-0.4389, 0.7347]	0.4941	1.1312
Phoenix	-1.4866	0.1675	[-1.8149, -1.1583]	-8.8746	0.2240
Toronto	0.5324	0.2246	[0.0923, 0.9726]	2.3709	1.7100
Urbana Champaign	-0.1296	0.3012	[-0.7199, 0.4607]	-0.4302	0.8575
Charlotte × Linguistic Similarity (time 1)	0.4184	0.0153	[0.3884, 0.4484]	27.3477	1.5275
Cleveland × Linguistic Similarity (time 1)	0.4067	0.0192	[0.3690, 0.4443]	21.1919	1.5179
Las Vegas × Linguistic Similarity (time 1)	0.4843	0.0208	[0.4435, 0.5252]	23.2532	1.6611
Madison × Linguistic Similarity (time 1)	0.336	0.0241	[0.2888, 0.3832]	13.9596	1.4050
Phoenix × Linguistic Similarity (time 1)	0.264	0.0273	[0.2105, 0.3174]	9.6801	1.5464
Toronto × Linguistic Similarity (time 1)	0.3648	0.0246	[0.3165, 0.4131]	14.8048	1.4526
Urbana Champaign × Linguistic Similarity (time 1)	0.2688	0.0293	[0.2113, 0.3263]	9.164	1.3114
Ego Degree (time 2)	0.0069	0.0009	[0.0052, 0.0086]	7.9718	1.0069
Alter Degree (time 2)	0.0051	0.001	[0.0032, 0.0070]	5.1729	1.0002
Same Gender	-0.165	0.0452	[-0.2537, -0.0764]	-3.6482	0.8074
Both Male	-0.0273	0.0723	[-0.1690, 0.1145]	-0.377	1.0209
Intercept	-9.4009	0.1556	[-9.7057, -9.0960]	-60.4356	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 3. Dyad-level linguistic convergence model in Yelp data. Linear regression, the dependent variable is change in linguistic similarity from Time 1 to Time 2. N= 3,011,271 (67,109df).

Predictor	b	SE	95% CI	t(67,109)	r
Linguistic Similarity (time 1)	-0.5982	0.0032	[-0.6044, -0.5919]	-188.0666	-0.5312
Cleveland	0.1222	0.0175	[0.0879, 0.1566]	6.9752	0.0235
Las Vegas	-0.0327	0.0127	[-0.0576, -0.0077]	-2.5682	-0.0074
Madison	0.1014	0.0259	[0.0506, 0.1522]	3.9119	0.0129
Phoenix	-0.0766	0.0128	[-0.1016, -0.0516]	-5.9991	-0.0206
Toronto	0.458	0.0204	[0.4181, 0.4979]	22.4929	0.0795
Urbana Champaign	0.0838	0.0339	[0.0174, 0.1502]	2.4734	0.0073
Charlotte × Network Tie (time 1&2)	0.2767	0.0354	[0.2073, 0.3462]	7.8098	0.0215
Cleveland × Network Tie (time 1&2)	0.3014	0.0936	[0.1179, 0.4850]	3.2197	0.0216
Las Vegas × Network Tie (time 1&2)	0.1714	0.0141	[0.1438, 0.1990]	12.1864	0.0672
Madison × Network Tie (time 1&2)	0.2899	0.0576	[0.1770, 0.4028]	5.0326	0.0091
Phoenix × Network Tie (time 1&2)	0.1675	0.0209	[0.1266, 0.2084]	8.0292	0.0446
Toronto × Network Tie (time 1&2)	0.5202	0.0474	[0.4273, 0.6130]	10.9809	0.0239
Urbana Champaign × Network Tie (time 1&2)	0.184	0.0549	[0.0764, 0.2915]	3.3524	0.0025
Ego Degree (time 1)	0.0004	0.0008	[-0.0013, 0.0020]	0.4424	-0.0068
Alter Degree (time 1)	-0.0024	0.001	[-0.0044, -0.0003]	-2.2816	0.0002
Ego Degree (time 2)	0.0015	0.0008	[-0.0001, 0.0030]	1.8779	0.0158
Alter Degree (time 2)	0.0036	0.0009	[0.0018, 0.0055]	3.8214	-0.0001
Same Gender	0.0271	0.004	[0.0192, 0.0351]	6.7094	0.015
Both Male	-0.0522	0.0074	[-0.0667, -0.0377]	-7.0666	-0.0237
Intercept	-0.0428	0.0116	[-0.0656, -0.0200]	-3.6781	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; r = partial correlation

Study 2 Robustness Check: Including Yelp Business Fixed Effects.

There may be systematic associations between linguistic style and idiosyncrasies of types of businesses or even the particular business that reviewers write about. For example, prepositions (e.g., “on”) might be more common with pizza (a food known for its many and diverse toppings) than with Chinese food. Or for restaurants whose owner/manager actively engages with customers, reviewers might be more likely to use third person pronouns referring to him or her. If so, and if unobservable demographic variables – such as ethnic heritage or race – make some types of people both more likely to review pizza joints than Chinese restaurants and also more likely to be friends, the results presented above could obtain spuriously.

To rule out this possibility, we estimated models with restaurant fixed effects. Specifically, before calculating the aggregate linguistic style of each reviewer, we normalize the linguistic style of each review, along each linguistic dimension, by the average of the linguistic styles of all the reviews submitted about the focal business. For example, if 4% of the words in the focal review relates to the “we” dimension, and on average the reviews submitted for the focal business include 3% of “we” words, then the linguistic style of the reviewer will be recoded to 1%. With these recoded values, we proceed to calculate each reviewer’s linguistic style, which we then use to estimate the models. Table S10 below shows the results. After controlling for business fixed effects, the effect sizes get larger, indicating that selection into reviewing specific types of restaurants could not account for our findings regarding linguistic similarity.

Table S10. Ex ante linguistic similarity predicts tie formation in Study 2. Dyad-level logistic regression models were estimated on binary indicators of friendship network tie. The key covariates were interactions between metro area dummies and the linguistic similarity measure at Time 1. This table compares with Table S6, the difference being that these analyses are based on business-normalized language styles, i.e., before calculating the linguistic style of reviewers, each of their reviews were normalized with the average tendency of the reviews for the focal business.

Model 1. Dyad-level logistic regression. DV: Network Tie (time 2). Sample: All dyads. N=4,440,227 (81,792 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Cleveland	0.073	0.0969	[-0.1168, 0.2628]	0.7537	1.0757
Las Vegas	-1.2006	0.0704	[-1.3386, -1.0625]	-17.0449	0.3010
Madison	0.4027	0.1134	[0.1805, 0.6249]	3.5523	1.4959
Phoenix	-1.4341	0.0786	[-1.5881, -1.2801]	-18.2539	0.2383
Toronto	0.1022	0.0928	[-0.0796, 0.2840]	1.1017	1.1076
Urbana Champaign	0.6664	0.1083	[0.4540, 0.8787]	6.1511	1.9471
Charlotte × Linguistic Similarity (time 1)	0.7656	0.0296	[0.7076, 0.8235]	25.9063	2.1502
Cleveland × Linguistic Similarity (time 1)	0.8489	0.0277	[0.7945, 0.9032]	30.6316	2.3370
Las Vegas × Linguistic Similarity (time 1)	0.7335	0.027	[0.6805, 0.7865]	27.1217	2.0824
Madison × Linguistic Similarity (time 1)	0.6743	0.0432	[0.5895, 0.7591]	15.5907	1.9626
Phoenix × Linguistic Similarity (time 1)	0.7666	0.0366	[0.6950, 0.8383]	20.9643	2.1525
Toronto × Linguistic Similarity (time 1)	0.6736	0.0294	[0.6160, 0.7313]	22.9116	1.9614
Urbana Champaign × Linguistic Similarity (time 1)	0.35	0.0443	[0.2632, 0.4368]	7.9066	1.4191
Ego Degree (time 2)	0.008	0.0006	[0.0068, 0.0092]	12.9103	1.0081
Alter Degree (time 2)	0.0049	0.0008	[0.0033, 0.0064]	6.2713	1.0049
Intercept	-7.3776	0.0631	[-7.5013, -7.2540]	-116.9589	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 2. Dyad-level logistic regression. DV: Network Tie (time 2). Sample: Dyads that are not friend in time 1. N=4,142,140 (81,766 df).

Predictor	*b*	*SE*	95% CI	*z*	*OR*
Cleveland	-0.341	0.1993	[-0.7316, 0.0495]	-1.7113	0.7110
Las Vegas	-1.1841	0.158	[-1.4938, -0.8745]	-7.4948	0.3060
Madison	-0.2153	0.2168	[-0.6401, 0.2095]	-0.9933	0.8063
Phoenix	-1.6258	0.1628	[-1.9449, -1.3067]	-9.9866	0.1968
Toronto	-0.0303	0.189	[-0.4008, 0.3401]	-0.1604	0.9701
Urbana Champaign	-0.4268	0.242	[-0.9010, 0.0475]	-1.7637	0.6526
Charlotte × Linguistic Similarity (time 1)	0.7208	0.0419	[0.6386, 0.8030]	17.1865	2.0560
Cleveland × Linguistic Similarity (time 1)	0.7736	0.0444	[0.6865, 0.8607]	17.4081	2.1676
Las Vegas × Linguistic Similarity (time 1)	0.6562	0.0458	[0.5665, 0.7459]	14.3422	1.9274
Madison × Linguistic Similarity (time 1)	0.6559	0.0644	[0.5297, 0.7821]	10.1878	1.9269
Phoenix × Linguistic Similarity (time 1)	0.4784	0.0606	[0.3597, 0.5971]	7.8974	1.6135
Toronto × Linguistic Similarity (time 1)	0.7068	0.0476	[0.6136, 0.8000]	14.8637	2.0276
Urbana Champaign × Linguistic Similarity (time 1)	0.5303	0.051	[0.4303, 0.6303]	10.3922	1.6994
Ego Degree (time 2)	0.0077	0.0005	[0.0067, 0.0087]	14.9528	1.0077
Alter Degree (time 2)	0.0045	0.0005	[0.0034, 0.0056]	8.2533	1.0045
Intercept	-9.5622	0.1484	[-9.8530, -9.2714]	-64.4513	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; OR = odds ratio.

Model 3. Dyad-level linguistic convergence model in Yelp data. Linear regression, the dependent variable is change in linguistic similarity from Time 1 to Time 2. N= 4,440,227 (81,792 df).

Predictor	b	SE	95% CI	t(81,792)	r
Linguistic Similarity (time 1)	-0.0413	0.0122	[-0.0651, -0.0175]	-3.3972	-0.5306
Cleveland	-0.5895	0.0034	[-0.5961, -0.5829]	-175.2443	0.019
Las Vegas	0.1014	0.0181	[0.0659, 0.1369]	5.5945	-0.0055
Madison	-0.0245	0.0133	[-0.0506, 0.0015]	-1.8486	0.0168
Phoenix	0.1251	0.0254	[0.0754, 0.1748]	4.9308	-0.011
Toronto	-0.0435	0.0132	[-0.0694, -0.0176]	-3.2956	0.0516
Urbana Champaign	0.2763	0.0184	[0.2401, 0.3124]	14.9814	0.0034
Charlotte × Network Tie (time 1&2)	0.038	0.0303	[-0.0214, 0.0974]	1.2545	0.0132
Cleveland × Network Tie (time 1&2)	0.1969	0.0423	[0.1141, 0.2797]	4.6587	0.019
Las Vegas × Network Tie (time 1&2)	0.291	0.0458	[0.2012, 0.3807]	6.3559	0.040
Madison × Network Tie (time 1&2)	0.1238	0.0166	[0.0914, 0.1563]	7.4783	0.0048
Phoenix × Network Tie (time 1&2)	0.1758	0.0505	[0.0767, 0.2748]	3.4774	0.0279
Toronto × Network Tie (time 1&2)	0.1085	0.0215	[0.0664, 0.1507]	5.0517	0.0283
Urbana Champaign × Network Tie (time 1&2)	0.558	0.0331	[0.4931, 0.6229]	16.8592	0.0015
Ego Degree (time 1)	0.1054	0.0486	[0.0102, 0.2005]	2.1701	-0.0199
Alter Degree (time 1)	-0.0025	0.0009	[-0.0044, -0.0007]	-2.6603	-0.0107
Ego Degree (time 2)	-0.0003	0.001	[-0.0023, 0.0017]	-0.2861	0.026
Alter Degree (time 2)	0.0039	0.0008	[0.0023, 0.0055]	4.7027	0.0162
Intercept	0.0018	0.0009	[-0.0000, 0.0037]	1.924	N/A

Note: SE = multi-way cluster-robust standard error; CI = confidence interval; r = partial correlation

Syntax used to estimate Study 2 models

The commands were estimated using Stata 15.1

```
* Tie Formation Models
clus_nway logit friend9 metro#c.sim8_std friendcountA_9
friendcountB_9 [pweight=weight], vce(cluster userid1_num
userid2_num)
clus_nway logit friend9 metro#c.sim8_std friendcountA_9
friendcountB_9 if friend8==0 [pweight=weight], vce(cluster
userid1_num userid2_num))
clus_nway logit friend8 metro#c.sim8_std friendcountA_8
friendcountB_8 samegender bothmale [pweight=weight], vce(cluster
userid1_num userid2_num)
clus_nway logit friend9 metro#c.sim8_std friendcountA_8
friendcountB_8 samegender bothmale if friend8==0
[pweight=weight], vce(cluster userid1_num userid2_num)
clus_nway logit friend9 i.metro metro#c.sim8_std friendcountA_9
friendcountB_9 samegender bothmale [pweight=weight]
clus_nway logit friend9 i.metro metro#c.sim8_std friendcountA_9
friendcountB_9 samegender bothmale [pweight=weight] friend8==0

* Tie Decay Models
clus_nway logit friend9 metro#c.sim9_std friendcountA_9
friendcountB_9 if friend8==1 [pweight=weight], vce(cluster
userid1_num userid2_num)
clus_nway logit friend9 metro#c.sim8_std friendcountA_8
friendcountB_8 samegender bothmale if friend8==1
[pweight=weight], vce(cluster userid1_num userid2_num)

* Linguistic Convergence Models
clus_nway reg change_sim sim8_std i.metro metro#friendboth89
friendcountA_8 friendcountB_8 friendcountA_9 friendcountB_9
[pweight=weight], vce(cluster userid1_num userid2_num)

*comparison of estimates with alter distance measures

logit friend9 i.metro metro#c.sim8_std friendcountA_9
friendcountB_9 [pweight=weight],or
logit friend9 i.metro metro#c.sim8_90dim_std friendcountA_9
friendcountB_9 [pweight=weight],or
logit friend9 i.metro metro#c.sim8_JS_std friendcountA_9
friendcountB_9 [pweight=weight],or
logit friend9 i.metro metro#c.sim8_lsm9_std friendcountA_9
friendcountB_9 [pweight=weight],or
logit friend9 i.metro metro#c.sim8_lsm18_std friendcountA_9
friendcountB_9 [pweight=weight],or
```

```
logit friend9 i.metro metro#c.sim8_9dim_std friendcountA_9
friendcountB_9 [pweight=weight],or
logit friend9 i.metro metro#c.sim8_lsa_std friendcountA_9
friendcountB_9 [pweight=weight] if metro!=3&metro!=5,or
logit friend9 i.metro metro#c.sim8_stylo_std2 friendcountA_9
friendcountB_9 [pweight=weight] if metro!=3&metro!=5,or

pcorr change_sim sim8_std i.metro i.metro#friendboth89
friendcountA_8 friendcountB_8 friendcountA_9 friendcountB_9
pcorr change_sim_90dim sim8_90dim_std i.metro
i.metro#friendboth89 friendcountA_8 friendcountB_8
friendcountA_9 friendcountB_9
pcorr change_sim_JS sim8_JS_std i.metro i.metro#friendboth89
friendcountA_8 friendcountB_8 friendcountA_9 friendcountB_9
pcorr change_sim_lsm9 sim8_lsm9_std i.metro i.metro#friendboth89
friendcountA_8 friendcountB_8 friendcountA_9 friendcountB_9
pcorr change_sim_lsm18 sim8_lsm18_std i.metro
i.metro#friendboth89 friendcountA_8 friendcountB_8
friendcountA_9 friendcountB_9
pcorr change_sim_9dim sim8_9dim_std i.metro i.metro#friendboth89
friendcountA_8 friendcountB_8 friendcountA_9 friendcountB_9
pcorr change_sim_lsa sim8_lsa_std i.metro i.metro#friendboth89
friendcountA_8 friendcountB_8 friendcountA_9 friendcountB_9 if
metro!=3&metro!=5
pcorr change_sim_stylo2 sim8_stylo_std2 i.metro
i.metro#friendboth89 friendcountA_8 friendcountB_8
friendcountA_9 friendcountB_9 if metro!=3&metro!=5
```

Alternative Measures of Linguistic Similarity

A key assumption of this paper lies in our approach to measuring linguistic similarity. If the results were contingent on this particular approach, their generalizability would be deeply suspect, so we explored numerous alternative measures of linguistic style similarity.

Pennebaker, the father of LIWC analysis, has done some work comparing the linguistic styles of different people. His approach, called Linguistic Style Matching (LSM) is quite similar to ours, based on the same underlying LIWC dictionary, but uses a slightly more complex method of aggregating across the individual dimensions (Gonzales, Hancock, & Pennebaker, 2010; Ireland et al., 2011). Specifically, he calculates the dimension-level distance between person *i* and person *j* as the absolute difference between their LIWC scores on the given dimension, divided by the sum of their scores plus *epsilon*. The *epsilon* ensures that the dimension-level distance score will be defined, even if both actors have scores of zero; the denominator effectively re-scales each dimension by its collective frequency of use. We prefer to re-scale each dimension by each individual's frequency of use by standardizing each person's dimension-level distance score prior to calculating the absolute difference, but besides this nuance, their approach is extremely similar to ours. Next, both approaches average the LIWC dimensions of interest. The LSM approach uses the similarity=(1-distance) formula to transform similarities to distances, but given the fact that the psychometrics literature has long established that such a transformation should be done with a log transformation (see the seminal paper of Shepard (1987) in *Science*), we decided to use a log transformation. Finally, Pennebaker and colleagues focus on 9 of the LIWC dimensions, while our main measure include 9 additional dimensions for a total of 18. Therefore, for comparability, we also calculated measures that apply our aggregation method to the 9 LIWC dimensions studied by Pennebaker (**LingSim-9**) and to all 89 LIWC dimensions (**LingSim-89**), and one that applies Pennebaker's aggregation method to our 18 LIWC dimensions of interest (**LIWC-LSM-18**) – Having said all of this, as Figures S3(a) and S3(b) show, the main results are substantially the same with Pennebaker and colleagues' measure and with ours.

We also measured linguistic similarity using approaches completely different from Linguistic Inquiry and Word Count. The first was latent semantic analysis (**LSA**; Landauer and Dumais, 1997, an unsupervised natural language processing technology that measures similarities between two documents (in our case, the set of documents produced by two authors) based on co-occurrence of similar terms. Thus, unlike LIWC, LSA measures content similarity without relying on human-created dictionaries. Nevertheless, we found that our core results hold up in LSA: that people who write about similar things are more likely to become friends and that friends write about things that are more similar compared to non-friends, controlling for their earlier similarity.

For maximum divergence in approach (i.e., to minimize the assumption about what kind of linguistic similarity is important), we also examined linguistic similarity with **Stylo**, a suite of tools from computational stylistics. Computational stylistics compares the linguistic styles across documents, often for purposes of adjudicating unknown or disputed authorship. Notable examples have examined the possibility of multiple authorship of the works of Shakespeare or of the Bible (Craig & Kinney, 2009). We apply this approach to examine the similarity across known different authors to determine whether people with similar linguistic styles are more likely to be friends and whether friends show greater subsequent similarity than non-friends, controlling for their earlier similarity. And as with LSA, our results hold up with this stylistic measure of linguistic similarity.

Overall, we find that across several, diverse measures of linguistic similarity, our results support the dual mechanisms of linguistic homophily. We present the results of these analyses of selection effects in Figure S3(a) and of convergence effects in Figure S3(b).

Figure S3(a): Comparison of friend selection models across diverse measures of linguistic similarity. LSA and Stylo analyses were computationally unfeasible for the large cities of Las Vegas and Phoenix and are omitted. Based on dyad-level logistic regressions that compare to Table S4's Model 1 and Table S6's Model 1. The dots show the effect sizes (odds ratios), the lines represent the 95% confidence intervals around them. For cross-sample comparability, these results are based on models with controls for endogenous network structures (actors' degree centrality) but not individual attributes.

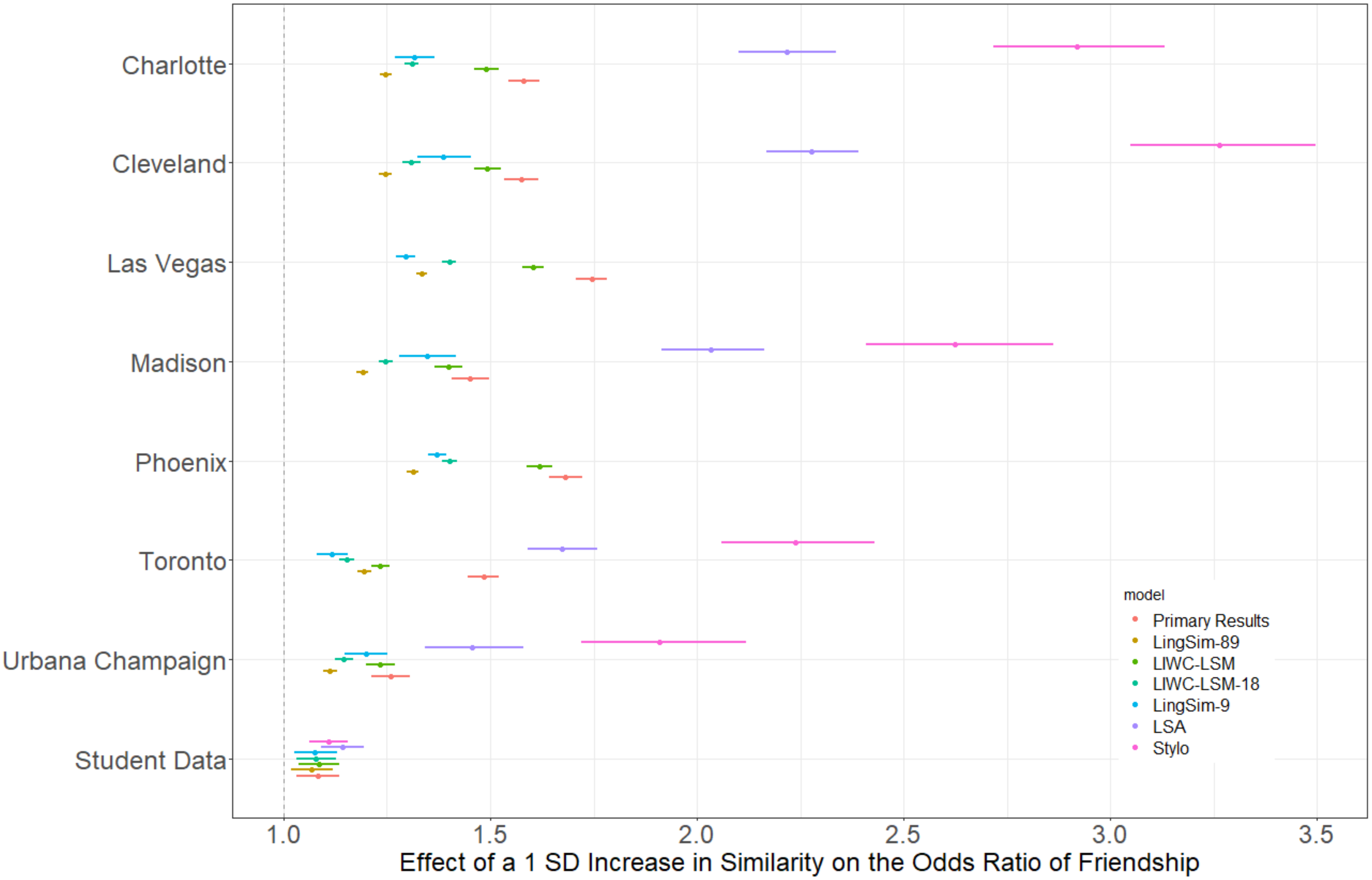
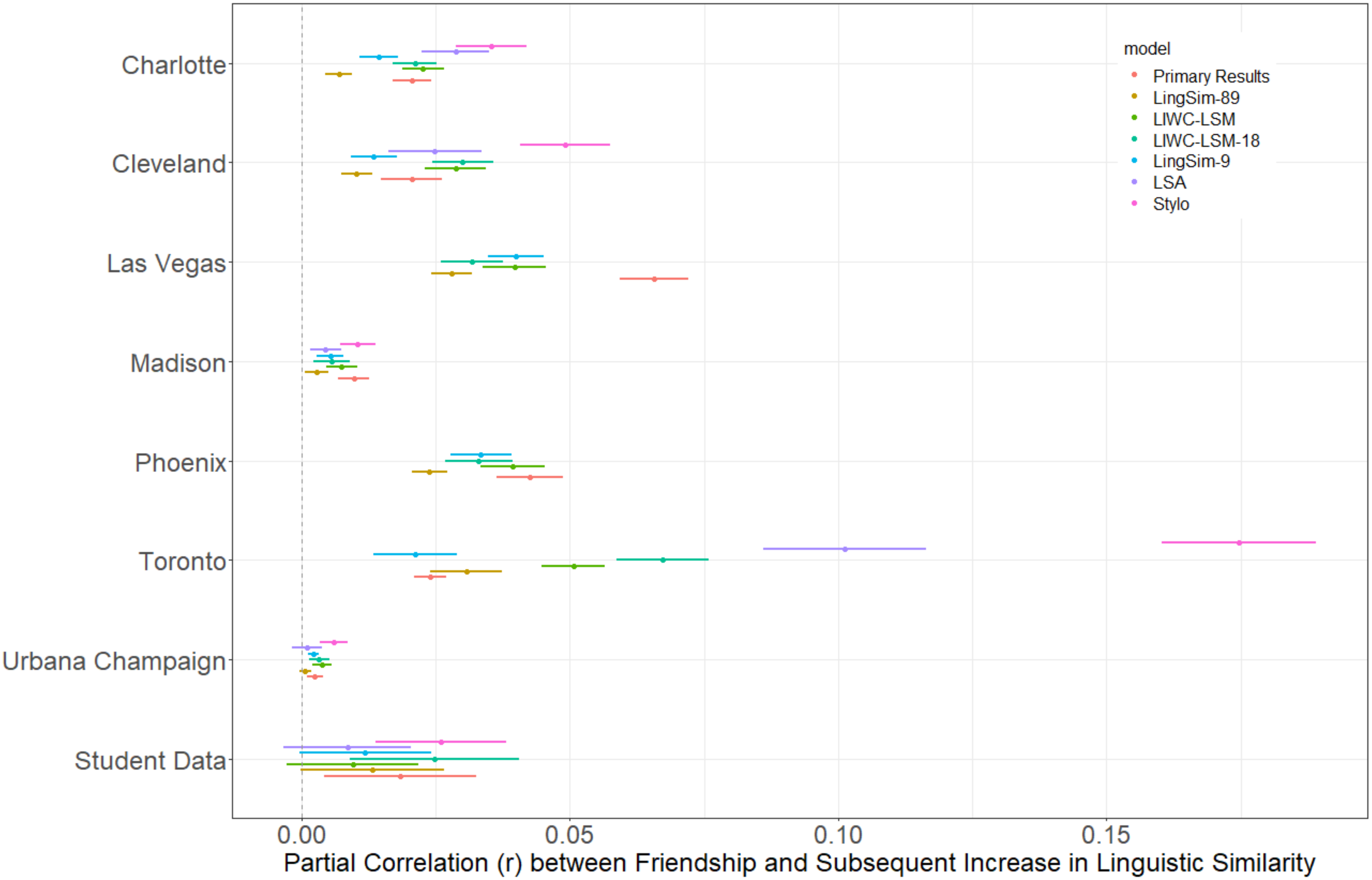


Figure S3(b): Comparison of linguistic convergence models across diverse measures of linguistic similarity. LSA and Stylo analyses were computationally unfeasible for the large cities of Las Vegas and Phoenix and are omitted. Based on dyad-level linear regressions that compare to Table S4's Model 3 and Table S7. The dots show the effect sizes (the partial correlation r), the lines represent the 95% confidence intervals around them. For comparability, these results are based on models with controls for endogenous network structures (actors' degree centrality) but not individual attributes.



Additional Analyses: Empirical Analysis of Network Fragmentation.

To assess the extent of fragmentation in the observed network data, we compared each observed network (two from the student data set and two from the Yelp data set) with a simulated population of networks with the same size, density, and degree distribution as the original network. To do this, we generated a population of 1,000 “randomly rewired” networks (Watts & Strogatz, 1998) for each observed network. Each rewired network was generated by randomly choosing two edges from the network, swapping their heads – thus preserving the degree distribution, while introducing randomness to the pattern of connections – and repeating this swap N_{edges} times, where N_{edges} is the number of edges in the network. We then calculated the modularity of the observed network and compared it against the distribution of modularity scores of randomly-simulated networks.

In the student data, we found that the observed time 1 network is more modular than all 1,000 random networks and more than 20 standard deviations greater than their mean modularity ($z = 20.72$; $p < 0.001$). At time 2, the modularity of the observed network was greater than that of all simulated networks and more than 24 standard deviations greater than their mean ($z = 24.02$; $p < 0.001$). Similar results obtained in the Yelp data: $z(t1) = 44.850$; $p(t1) < 0.001$; $z(t2) = 46.689$; $p(t2) < 0.001$. The increase in z-scores from time 1 to time 2 in both the student data and the Yelp data suggest that these networks grow increasingly fragmented over time. These empirical results lend credence to the findings of the simulation study.

Extension: Simulated Consequences

In Studies 1 and 2, we demonstrate empirically that linguistic homophily occurs through two distinct mechanisms: selection and convergence. In this extension, we offer a simple, stylized simulation of the consequences of these dual effects for the overall topology of the network. We argue that when people connect with similar others, then become increasingly similar with their contacts, the consequence of such co-evolution of social networks and linguistic styles is the fragmentation of the network.

Much has been made recently of “echo chambers” in the political sphere (Boutyline & Willer, 2017) and the provision of news (Jacobson, Myung, & Johnson, 2016). Through this simulation, we argue that, more fundamentally, the dual mechanisms of selection and convergence shape the social networks that we inhabit to be increasingly populated with similar others and disconnect us from dissimilar others. In doing so, we build on prior work (e.g., Kalish, Luria, Toker, and Westman 2015) that has demonstrated how homophilous selection and convergence in non-linguistic attributes contribute to network fragmentation.

Methods

The simulation is similar in spirit to that of Carley (1991), who studied preferential attachment and group stability in cultural processes. The simulation setup also builds on insights in (Baldassarri & Gelman, 2008) and (DellaPosta, Shi, & Macy, 2015), but while those authors focus on the correlation of different attribute dimensions, we focus on fragmentation. The simulation model starts with N agents, each of whom has an initial linguistic style which can be characterized along M dimensions (akin to the dimensions of the LIWC model, such as “pronoun usage” and “article usage”). Each agent has a randomly selected set of starting values such that their style along each dimension is drawn from a uniform random distribution. The initial network structure is Eroös-Rényi random.

To simulate the co-evolution of network and linguistic style, we build on the SIENA framework for stochastic actor-oriented network models (Snijders, 2005). SIENA allows for creation of a new tie, maintenance of a tie, or dissolution of a tie, and for social influence. The simulation proceeds as a series of microsteps, in each of which, two changes take place simultaneously. In the network updating step, the network of friendships is updated such that a random pair of agents is selected and the state of the friendship tie is determined as a (variable) function of the linguistic similarity of the two agents at that time. In the influence step, two agents are chosen randomly, and if they are friends at that time point, their linguistic styles are updated such that the value of their linguistic styles along each dimension become (variably) more similar. See the RSiena Manual (Ripley, Snijders, Boda, Voros, & Preciado, 2017) for further details on the simulation framework. We run the simulations in R, using the RSiena package.

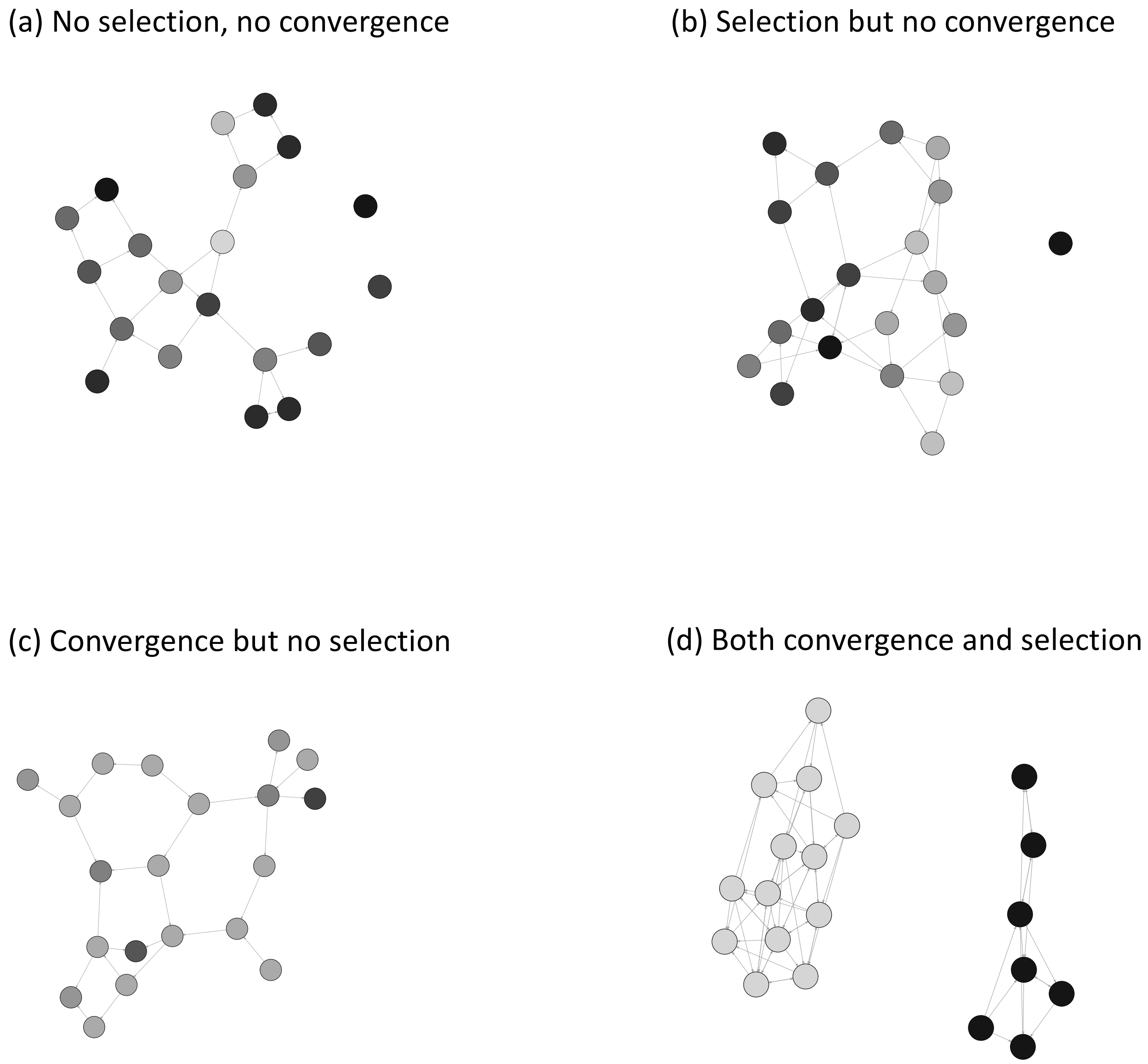
Our simulations differ from the *SimulateNetworksBehavior* function in that it incorporates different network evolution parameters. First, we include in our models “density” to regulate the overall network density of the simulated models. Second, we include the parameter “simX” which regulates the selection processes: the higher the value set for this parameter, the more strongly similarity along linguistic styles will regulate tie formation processes. Third, and finally, we include in the simulation model the *avAlt* parameter, which regulates the convergence process: the larger the parameter, the faster the linguistic style of a focal agent will converge to the average of the linguistic styles of the agents she is connected to. See the RSiena manual at www.stats.ox.ac.uk/~snijders/siena/RSiena_Manual.pdf for the exact specifications in the effect.

In the simulation results provided in Figs. S4 and S5, we present results of network simulations with N=20 agents, M=1 one dimension of linguistic style, and dens=0.08 network density. In additional analyses we conducted a series of robustness checks with alternative specifications, such as larger or smaller networks, more than one linguistic dimensions, and alternative density values, but as the main findings did not change qualitatively, we do not present them here.

Results

As an illustration of the simulated networks, Figure S4 shows examples under four parameter conditions. The top left panel (4a) shows a network simulated with the selection and convergence parameters both set to zero. The simulated network is a random graph and does not show clustering. The top right panel (4b) shows a case in which the convergence effect is set to zero but the selection effect is set to be strong (γ_s=3; see e.g., Steglich, Snijders, and West (2006)). While we see some evidence for clustering here, the overall network is mostly connected. The bottom left panel (4c) shows a simulated network with no selection effect and with γ_c=3, a strong convergence effect. This network exhibits homogenous values along the linguistic dimension, but no clustering. Finally, (4d) shows a network that was simulated with both convergence and selection parameters set to 3. When both selection and convergence effects are strong, the network breaks down into two separate components, and each component is fully homogenous and starkly different in linguistic style from the other component. This case illustrates that when both selection and convergence are strong, we can see the emergence of echo chambers, in which people are densely connected with similar others, but disconnected from those who are different. A limitation of this extension is that, for technical reasons related to the structure of stochastic actor-oriented models, we were unable to condition the simulation model on empirical parameter estimates from Studies 1 and 2.

Figure S4. Simulated networks under four parameter conditions. The shade of the nodes denotes the agents' position on the behavioral dimension, ranging from 1 (white) to 10 (black)



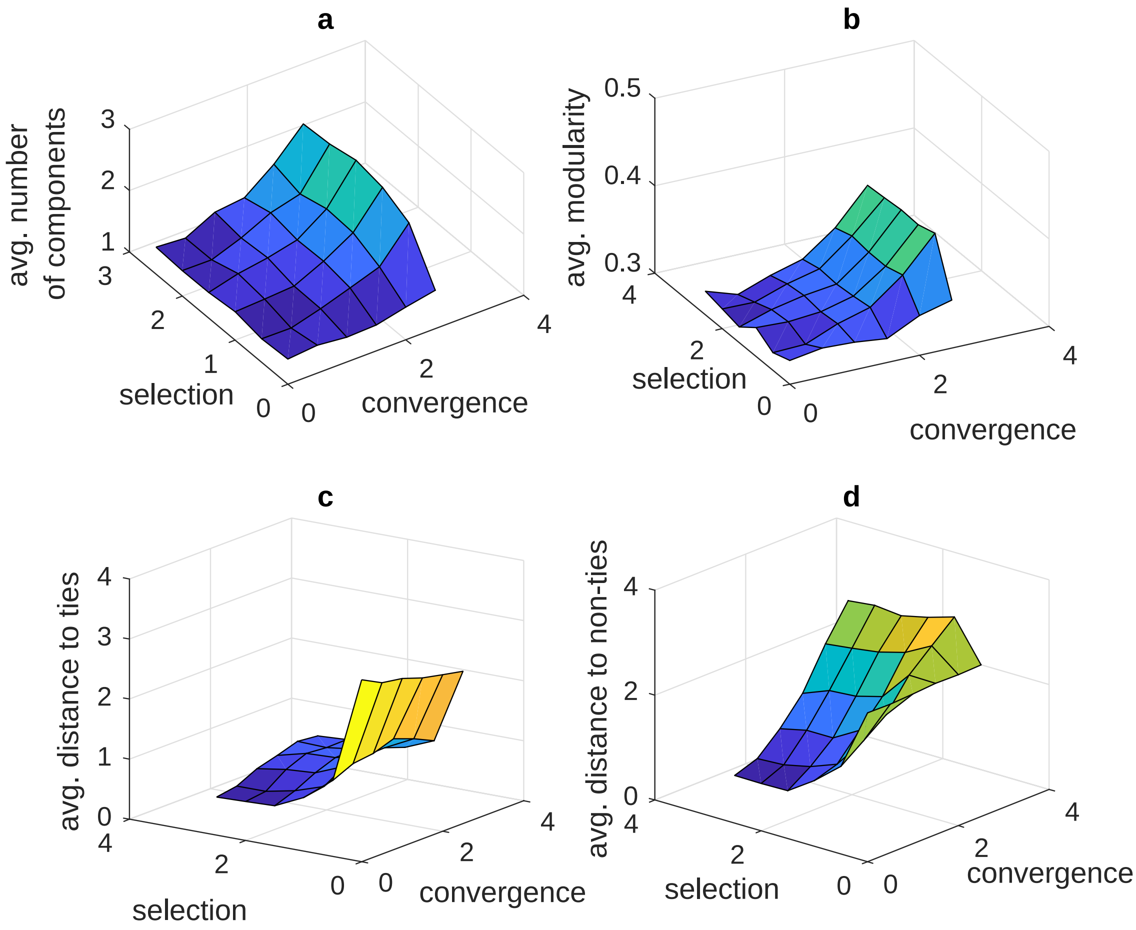
(a) No selection, no convergence

(b) Selection but no convergence

(c) Convergence but no selection

(d) Both convergence and selection

Figure S5. Simulation results. Each value is calculated as the average of outcomes of 500 independent simulation



To move beyond illustrations, we conducted a large number of simulations across various parameter configurations. Figure S5 shows the results of our simulation of network and linguistic style change. Each value on the figure is the average outcome of 500 independent simulations. For each network we calculated the modularity scores (Clauset, Newman, & Moore, 2004) based on a community detection algorithm (Pons & Latapy, 2006). In Figure S5a we show the number of components in the simulated networks, and Figure 5b we show the average modularity score resulting from that specific parameter combination. The results show that, controlling for network size and density, the average modularity and the number of components is the largest when both selection and convergence are strong. These patterns reinforce our empirical findings. Note that while the component count and modularity score gradually increase with the convergence parameter, the consequences of an increase in selection are strong in the lower ranges but then flatten out. This indicates that even a moderately strong selection effect is enough to result in networks with high modularity and disconnected components. Finally, Figs S5c and 5d show the average linguistic distance between dyads that are connected (Fig 5c) and not connected (5d). Interestingly, the selection parameter mostly regulates the linguistic distance between ties, while the convergence parameter mostly influences the distance between non-ties.

Combining the patterns from these two figures we arrive at a complete characterization of the joint effect of selection and convergence: when both selection and convergence are weak, the distances between members of connected and also members of unconnected dyads will be high. But when both selection and convergence are strong, the distances between members of connected dyads will be low while and members of unconnected dyads will be high – this is the echo chamber effect of highly cohesive groups with strong disagreements across groups.

Although our two empirical data sets lack the "continuous time" longitudinality of our simulation, we nevertheless wondered whether these simulated consequences of our observed mechanisms played out empirically. We find that across both data sets, and at both points in time, the observed network is dramatically more modular (i.e., fragmented) than randomly simulated networks of the same size, density, and degree distribution ($p < 0.001$). Further, in both data sets, fragmentation appears to be increasing from time 1 to time 2.

Discussion of the Simulation

In a set of computational simulations, we have demonstrated that if both mechanisms of linguistic homophily – selection and convergence – are present, they will lead to overall fragmentation of the network. In doing so, we build on and extend (DellaPosta et al., 2015), who showed that lifestyle and ideology tend to cluster together; we argue that this self-reinforcing dynamic may serve to structure the network itself. Indeed, the simulation builds on our empirical analysis of linguistic styles, but may generalize to any setting in which individual attributes are time-varying and subject to peer influence through the network. This finding has important consequences and may shed light on the increase in fragmentation and polarization observed in modern societies and the retreat of individuals (political or corporate leaders, for example) into informational echo chambers.

Finally, we note that ideally our simulations should be calibrated with empirical data, i.e., the parameters driving the simulations should be estimates from empirical models. Unfortunately, we were unable to do so with our current dataset because the simultaneous estimation of selection and convergence coefficients in the SIENA framework require three waves of data, while we only have two waves. We encourage future researchers to extend our endeavors into this direction.

References cited in the Supplemental Online Material

Andersen, M. S. (2018). Effects of Medicare coverage for the chronically ill on health insurance, utilization, and mortality: Evidence from coverage expansions affecting people with end-stage renal disease. *Journal of Health Economics, 60*, 75-89.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*: Princeton university press.

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology, 114*(2), 408-446.

Boutyline, A., & Willer, R. (2017). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology, 38*(3), 551-569.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics, 29*(2), 238-249.

Carley, K. (1991). A theory of group stability. *American Sociological Review*, 331-354.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*(6), 066111.

Craig, H., & Kinney, A. F. (2009). *Shakespeare, computers, and the mystery of authorship*: Cambridge University Press.

Dahlander, L., & McFarland, D. A. (2013). Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly, 58*(1), 69-110.

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology, 120*(5), 1473-1511.

Egger, P. H., & Tarlea, F. (2015). Multi-way clustering estimation of standard errors in gravity models. *Economics Letters, 134*, 144-147.

Feiler, D. C., & Kleinbaum, A. M. (2015). Popularity, similarity, and the network extraversion bias. *Psychological Science, 26*(5), 593-603.

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*(1), 3-19.

Greenberg, J., & Fernandez, R. M. (2016). The Strength of Weak Ties inMBA Job Search: A Within–Person Test. *Sociological Science, 3*, 296-316.

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*(1), 39-44.

Jacobson, S., Myung, E., & Johnson, S. L. (2016). Open media or echo chamber: The use of links in audience discussions on the Facebook pages of partisan news organizations. *Information, Communication & Society, 19*(7), 875-891.

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*(6), 1737.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*: Guilford press.

Kleinbaum, A. M., Stuart, T. E., & Tushman, M. L. (2013). Discretion within constraint: Homophily and structure in a formal organization. *Organization Science, 24*(5), 1316-1336.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Lindgren, K.-O. (2010). Dyadic regression in the presence of heteroscedasticity—An assessment of alternative approaches. *Social networks, 32*(4), 279-289.

Liu, C. C., & Srivastava, S. B. (2015). Pulling closer and moving apart: Interaction, identity, and influence in the US Senate, 1973 to 2009. *American Sociological Review, 80*(1), 192-217.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Retrieved from Austin, TX:

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications, 10*(2), 191-218.

Ripley, R. M., Snijders, T. A. B., Boda, Z., Voros, A., & Preciado, P. (2017). Manual for RSiena.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317-1323.

Snijders, T. A. (2005). Models for longitudinal network data. *Models and methods in social network analysis, 1*, 215-247.

Steglich, C., Snijders, T. A., & West, P. (2006). Applying SIENA: An illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2*(1), 48-–56.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature, 393*(6684), 440.