THREE ESSAYS ON THE STRENGTH OF LONG-RANGE COMMUNICATION TIES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Patrick Park

May 2016

THREE ESSAYS ON THE STRENGTH OF LONG-RANGE COMMUNICATION TIES

Patrick Park, Ph. D.

Cornell University 2016

Granovetter's article on the strength of weak ties is one of the most widely cited in the social sciences in the past 40 years. Compared to a strong tie, a weak tie tends to span "long" network distances, thereby promoting access to otherwise unavailable information, greater social integration, and more rapid diffusion of innovations. However, the hypothesized length of weak ties has eluded empirical research, primarily due to the paucity of fine-grained network data at the population level. Using a bidirected phone call network (51.3M nodes) constructed from complete call logs during a one-month period in the entire United Kingdom, I confirm that the median tie strength, measured as call volume, initially declines as the length of the tie, measured as the second shortest path length, increases from 2 to 4 steps, as Granovetter predicts. However, I find that the opposite holds for ties of length greater than 4, such that ties of length 10, while relatively few in number, are nearly as strong as ties of length 3. Substantively similar patterns are found from the analysis of Twitter communication networks in eight countries that vary in national culture, suggesting that a common generative process may lie behind the increasing strength of long bridging ties. I examine three competing explanations: 1) nodes with few neighbors tend to invest heavily in their relations with one another but with a lower probability of having a neighbor in common; 2) the telephone is used both socially and instrumentally, such that the social use is consistent with Granovetter’s thesis while the instrumental use is not; 3) social and spatial mobility causes social ties to be

"stretched" across the network, with a probability of being broken that is greater for ties that are weak. I conclude that this selection effect is the explanation with the greatest empirical support.

BIOGRAPHICAL SKETCH

Patrick Park received his Bachelor of Arts in Sociology and Chinese Language and Literature and Masters of Arts in Sociology from Yonsei University in South Korea. He received his Ph.D. in Sociology from Cornell University. His research interests lie at the intersection of social networks, online communities, and computational social science.

For Michelle and Edie

ACKNOWLEDGMENTS

This dissertation would not have seen light without the enormous intellectual, social, emotional, and financial support generously offered by mentors, committee members, institutes, friends, and family throughout my graduate career.

I am deeply indebted to Dr. Michael Macy for his mentorship. His incessant curiosity for the swiftly transforming social world, coupled with creative and open-minded approaches to organizing and conducting collaborative research, is what I have come to identify and try to internalize as ideal qualities to which an academic ought to aspire throughout his career. I was blessed with the fortune of working with Dr. Douglas Heckathorn who showed me by example the importance of rigor and perseverance in building an entirely new subfield, as exemplified by his research on Respondent-Driven Sampling. Those qualities, I have tried to weave into my dissertation and other research. I am fortunate to have Dr. Matthew Brashears on my committee. His careful, systematic, and rigorous approach to research, from research design, grant proposal writing, to data analyses, formed a bundle of high standards that I tried to emulate in various phases of my own work.

Throughout my years in graduate school, fellow graduate students in the Social Dynamics Lab have offered invaluable help and advice at various stages. In particular, Chris Cameron, Milena Tsvetkova, Yongren Shi, and Shaomei Wu have been instrumental in the development of a number of research projects that I have undertaken and in shaping my perspective on sociology and the burgeoning field of computational social science. Michael Genkin who often toiled with me late into the night in Uris Hall made the rougher and isolating periods of the graduate school experience tolerable, if not enjoyable. The emotional support,

thoughtful advice, and insightful conversations from Youngjoo Cha, Wooseok Jung, Jeong-han Kang, Byungkyu Lee, Duk-Gyoo Kim, Chan Suh, and Yisook Lim have helped me renew my passion for research and kept me afloat during rough times.

I am also grateful to collaborators, Ryan Compton and Tsai-Ching Lu at the HRL Laboratories, for generously offering substantive comments on and resources towards parts of my dissertation research. Talented graduate and undergraduate students, Sean Ogden, Yingbin Zhao, Xiao Ma, Nikhil Bhat, and Christie Abel, offered amazing research assistance in data collection and in manual coding. I also acknowledge the generous grant support from the U.S. National Science Foundation (SES-1226483 and SES-1434164) that made data collection and analysis possible.

Finally, the completion of this dissertation would not have been possible without the inarticulable sacrifice and unwavering support of Ghalim Michelle Lee. This dissertation is evidence of her sacrifice and reminder for the future of my eternal indebtedness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THE STRENGTH OF LONG TIES: THE CASE OF A NATIONAL PHONE COMMUNICATION NETWORK

Abstract

Granovetter's article on the strength of weak ties is one of the most widely cited in the social sciences in the past 40 years. Compared to a strong tie, a weak tie tends to span "long" network distances, thereby promoting access to otherwise unavailable information, greater social integration, and more rapid diffusion of innovations. However, the hypothesized length of weak ties has eluded empirical research, primarily due to the paucity of fine-grained network data at the population level. Using a bidirected phone call network (51.3M nodes) constructed from complete call logs during a one-month period in the entire United Kingdom, I confirm that the median tie strength, measured as call volume, initially declines as the length of the tie, measured as the second shortest path length, increases from 2 to 4 steps, as Granovetter predicts. However, I find that the opposite holds for ties of length greater than 4, such that ties of length 10, while relatively few in number, are nearly as strong as ties of length 3. I examine three competing explanations: 1) nodes with few neighbors tend to invest heavily in their relations with one another but with a lower probability of having a neighbor in common; 2) the telephone is used both socially and instrumentally, such that the social use is consistent with Granovetter's thesis while the instrumental use is not; 3) social and spatial mobility causes social ties to be "stretched" across the network, with a probability of being broken that is greater for ties that are weak. I conclude that this selection effect is the explanation with the greatest empirical support.

Introduction

In the 40 years since Granovetter's (1973) pioneering work on the strength of weak ties, numerous studies have confirmed his counter-intuitive thesis that individuals are more likely to acquire novel information from an acquaintance than from a close friend (Campbell, Marsden, and Hurlbert 1986; Lin and Dumin 1986; Montgomery, 1994). The reason is straightforward: Information that spreads within a "small circle of friends" is more likely to be redundant than information acquired from an acquaintance in a distant region of a social network. In short, the strength of weak ties is that they facilitate access to novel information, social integration, and rapid diffusion of innovation.

On closer inspection, the theory can be decomposed into two claims – that bridge ties are important conduits of information and that these ties are weaker than those embedded in dense network clusters. Granovetter addresses both points. His primary emphasis is on showing that information about a job is more likely to be obtained from acquaintances. However, he also argues that these acquaintances are less likely to share common neighbors, stating for example that "no strong tie is a bridge (1973: 1364)" and "all bridges are weak ties (1973: 1364)."

While the importance of bridge ties for diffusion has been well documented (Watts and Strogatz, 1998), the strength of these ties has eluded research. In contrast to the effects on diffusion, which can be tested with computational models (Watts and Strogatz 1998), the strength of long-range ties is an empirical question that requires fine-grained network data at the population level which until recently have been exceedingly difficult to collect. Population-level network data has been obtained primarily through surveys that ask respondents about their "ego networks." The problem is that measuring the length of a tie is not possible with egocentric network data. Even if a respondent might be able to correctly guess that the tie she has with a

particular friend is a bridging tie, she cannot be reasonably expected to know how many steps that tie can bridge. A bridge that spans network clusters involves connections among third parties that exist outside the purview and control of a single respondent. Complete network data at the meso-scale (e.g. neighborhood and organization), which constitute the bulk of 20th-century social network research, overcome this shortcoming and allow the researcher to measure the network distance a tie bridges within a given social context. However, these meso-scale network data inevitably limit the observation of the full range of network distances that exist across contexts and in the wider population. As a result, four decades after Granovetter's famous theory first appeared, we know very little about how the relative frequency of communication, or the strength of social ties, covaries with their length in full variation. Empirical regularities ("no strong tie is a bridge") observed from such limited data can mislead one to draw untenable conclusions with misplaced confidence.

This empirical constraint is rapidly disappearing with the increasing availability of population-level social network data derived from the digital records of interpersonal communication. However, even Facebook is still a long way from providing a complete nationwide social network. Instead, I use data for close to 100% of landlines and over 90% of mobile phone subscribers in the UK over a one-month period to construct the most complete nationwide social network ever attempted. These data make it possible for the first time to test Granovetter's theory at population scale. The results were surprising. First, I find that long-range ties that span wide network distances are bimodally distributed and much more prevalent than expected, with peaks at tie ranges of 2 and 10 steps (where range is defined as the second shortest path length between adjacent nodes). Second, I find that the relationship between tie strength and length is nonlinear. Granovetter's thesis is confirmed for local bridges up to four

"jumps" but not for global bridges that span longer network distances. Above four, tie strength increases with tie range. I probe three alternative explanations for this highly non-linear pattern and find that the data are most consistent with the explanation that long-range ties form as a byproduct of focal tie stretching and third party tie decay.

Previous Work on Bridging Ties

Research spawned from Granovetter's pioneering work focuses primarily on testing the efficacy of weak ties across diverse social contexts from labor markets (Granovetter 1973; Yakubovich 2005), entrepreneurship (Singh 2000; Burt 1992), social movements (Oliver and Myers 2003), knowledge and information sharing (Bakshy et al. 2012; Hansen 1999), and new ideas (Burt 2004). Other studies building on this line of research focus on the different utility that strong and weak ties each provide in specific applications. The different uses of strong and weak ties are formulated, for example, as complimentary (Raegan and McEvily 2003; Tiwana 2008) or as a trade-off where the potential for information diversity through weak, bridging ties offsets the potential for information volume through strong, embedded ties (Aral and Alstyne 2011). However, the majority of these extensions seldom question Granovetter's basic insight that bridging ties are weak. Following Granovetter (1974), most of these studies identify bridge ties from the individual's perspective, such as Burt's (1992) measure of network constraint, which can at best identify local bridges. Granovetter defines a tie as a "'local bridge of degree *n*' if *n* represents the shortest path between its two points (other than itself), and $n > 2$ (1973: 1365)." That is, *n* is the number of intermediary ties required to reach from one node to the other if the tie that directly connects these two nodes were to be removed. Following recent studies, I refer to *n* as the range of the tie (Kossinets, Kleinberg, and Watts 2008; Watts 1999). According to this

definition, the minimum range is two (i.e. two nodes have a neighbor in common, forming a closed triad). However, a tie within a closed triad is not a bridge. The minimum range of a bridge is three since the bridge must span between two non-adjacent nodes. Building on Granovetter's definition, I distinguish between local and global bridges, where a local bridge has a minimum range of three and a maximum of four.

In addition to range, Granovetter also measures network distance as "embeddedness," defined as the number of common neighbors shared by two adjacent nodes. Whereas range measures the social distance that is spanned by an edge, embeddedness measures the "bandwidth" of an edge, defined as the number of alternative two-step paths by which information might pass between two adjacent nodes.

With few exceptions (e.g. Centola and Macy 2007; Kossinets et al. 2008; Onnela et al. 2007), the research literature is dominated by the dichotomy of embedded vs. unembedded ties at the individual unit of analysis, to the near exclusion of range, leaving a blind spot in the study of diffusion. The problem is not that researchers are uninterested in the "systematic investigation of the origin and development of those ties which bridge as compared to those which do not (Granovetter 1983: 229)." The problem is empirical. Embeddedness can be measured at the individual level, using standard survey methods for collecting ego-centric network data. In contrast, range requires data on the entire social network, which until recently was only possible for very small networks (Bott 1957; Zachary 1977). In small networks, the only bridges that can be detected will be local bridges. What is more, even with larger networks on the orders above 10^6, range cannot be accurately measured if the network does not include the vast majority of members within a set boundary (e.g. national population) due to hidden and unobserved paths from missing data. In short, without population-level network data, there is no way to know if

Granovetter's theory about the strength of local bridges generalizes to those that are global. In the absence of data, it is reasonable to assume that if local bridges are weaker than embedded ties, then surely global bridges are even weaker than local.

Reasonable but wrong. It turns out that it is the other way around: global bridges are nearly as strong as embedded ties, as I demonstrate below. With the readily available records of digitally mediated social interactions, from phone call and email logs to Facebook messages, it is now possible to follow up on Granovetter's call for the "systematic investigation of the origin and development of those ties which bridge as compared to those which do not" (Granovetter 1983: 229). I use nearly complete nationwide telephone call logs to measure the tie-strength of local and global bridges and to probe alternative explanations for the surprising strength of long ties.

Data

With the cooperation of British Telecom, I obtained nearly complete UK phone call records logged during August 2005. Each call record contains the caller, the receiver, the timestamp for the beginning of each call, and call duration in seconds. The raw call logs contain 90% of the mobile phones and 99% of residential and business landlines in the UK. From these raw data, I filter out calls that lasted less than five seconds in order to exclude mistaken calls. I also remove unreciprocated calls in which A calls B but B never calls A, which filters out automated calls, misplaced calls, and other relationships that are not socially meaningful. I further filter out call centers, business gateways, dorm phones, and other shared lines based on aggregate call volume and number of network neighbors that exceeds the physical limitations on a single person. Finally, I delete lines that deviate from the ordinary usage pattern of individually owned phones by frequently participating in calls with more than one other line simultaneously. For a more

detailed description of the data with slightly different filtering procedures, see Eagle et al. (2010). The resulting bidirected phone call network contains 51.3 million phone lines and 131.6 million calling pairs, in a population of 49.0 million adults in 2004 (Bates 2006) and 52.1 million in 2011 (Office for National Statistics 2012).

The main measure of tie strength is call volume, which is the aggregate call time (in seconds, mean=2582, s.d.=5190) between two phone lines during the one-month observation window. Although classic studies on the measurement of tie strength (e.g. Marsden and Campbell 1984) question the validity of behavioral measures of tie strength such as contact frequency, duration, or volume, more recent studies based on fine-grained behavioral data suggest that these behavioral measures are fairly accurate predictors of tie strength (Jones et al. 2013). I measure tie range using Granovetter's definition, which is also equivalent to the second shortest path length of two adjacent nodes (Kossinets et al. 2008). I use the term "length" and "range" interchangeably hereafter. Finally, I use the location of local telephone exchanges for residential landlines to calculate the geo-spatial distance spanned by a subset (66 million) of the network edges (Eagle et al. 2010).

The Distribution of Tie Range

Social networks tend to be highly clustered, with the implication that the majority of social ties will be short range whereas only a small proportion will be long range ties. Contrary to this intuition, Figure 1 shows that range has a bimodal distribution, with peaks at n=2 and n=7. The proportion of local bridges (n=3 or 4) is far less than the proportion of longer global bridges. A possible explanation is the small world structure of the BT network. Figure 2 demonstrates how the bimodal distribution can obtain on a simulated Watts-Strogatz small world network with

500,000 nodes with a uniform degree of 10 (Watts and Strogatz 1998). The first mode observed at n=2 reflects the high local clustering of the network (CC=0.15 for p=0.4), while the second mode at n=7 reflects the short characteristic path length (CPL = 6.56). This bimodal distribution disappears as local clustering decreases with larger randomness (p=0.8).

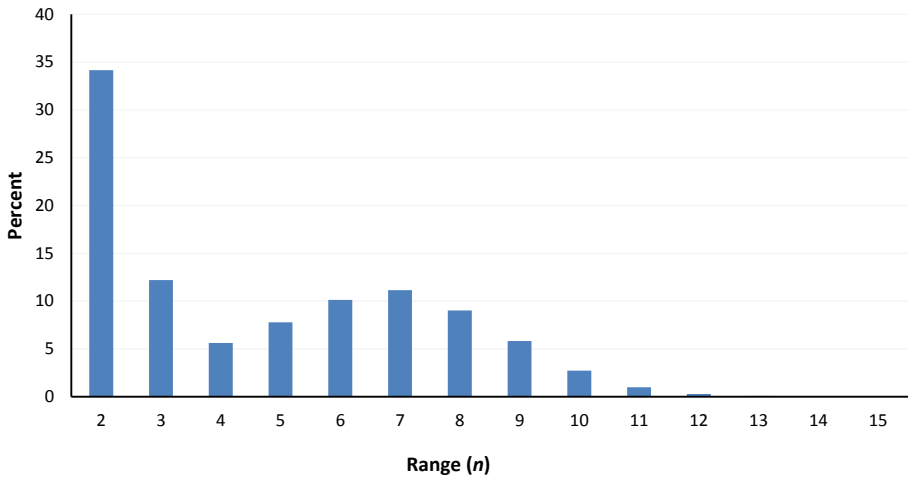
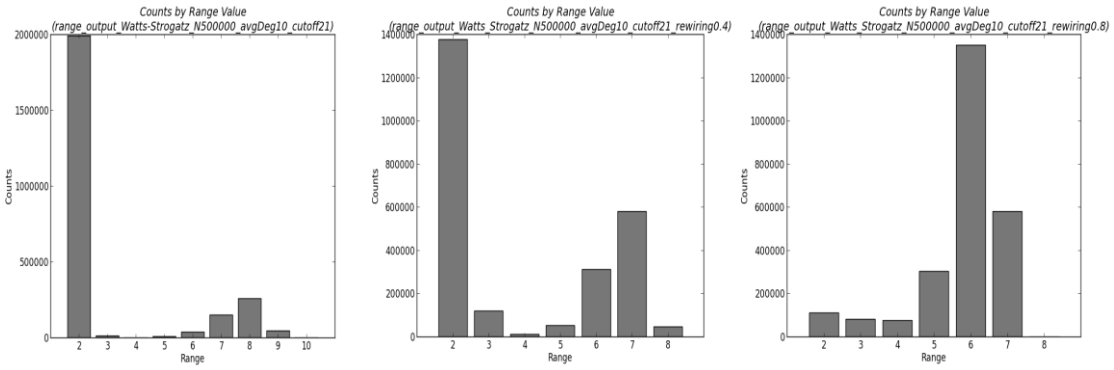


Figure 1: Tie Range Distribution of the Phone Network

To check for the generalizability of the phone network's bimodal range distribution, I reproduced the range distributions for a wide range of social networks, online and offline, large and small, declared (e.g. friendship networks) and interactional (e.g. communication networks). Specifically, I used intra-organizational employee networks on four types of ties (friendship, advice giving, and voicing problems and ideas) across nine U.S. credit unions (Detert et al. 2013), Facebook friendships across 100 US universities (Traud, Mucha, and Porter 2011), Live Journal friendship network (Yang and Leskovec 2012), an email exchange network among members of an EU research institute (Leskovec, Kleinberg, and Faloutsos 2007), and communications on Wikipedia (Leskovec, Huttenlocher, and Kleinberg 2010). In contrast to the BT phone network,

ties in these networks all exhibited skewed unimodal distributions with a maximum tie range of 7.



p = 0.2 *p* = 0.4 *p* = 0.8

Figure 2. Range Distributions in Simulated Small-World Networks. **500,000 nodes, uniform degree of 10, and varying levels of randomness.**

Strength and Range of Ties

Figure 3 reports changes in the tie strength distribution as range increases. Tie strength declines sharply as range increases from 2 (embedded ties) to 5 (local bridges), as Granovetter predicts. Surprisingly, however, for ties longer than 5, the pattern reverses, with tie strength increasing to a level at *n*=10 that approaches the strength of ties with *n*=2.

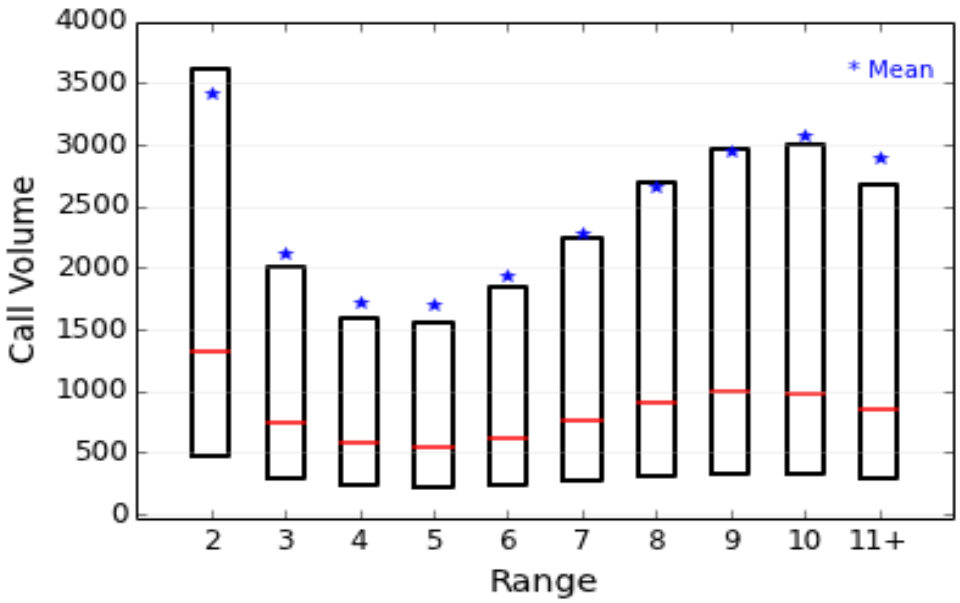


Figure 3: Call Volume by Tie Range. The black boxes represent interquartile range and the red horizontal line represents the median.

To ensure that the nonlinear pattern is not idiosyncratic to a particular measurement of tie strength, I present alternative measures of tie strength in Figure 4, which decomposes call volume into call frequency and mean call duration. Frequency and duration also exhibit nonlinear patterns similar to aggregate call volume as range increases from 2 to 11+. Mean call frequency declines as length increases from 2 to 6 but then increases thereafter. Call duration declines only from range 2 to 3, peaks at 9, and then declines again. Figures 4 then reveals an important difference between the strength of embedded and long-range ties. The strength of embedded ties consists of a relatively large number of short duration calls, while the strength of long-range ties comes from a relatively small number of long duration calls, suggesting that the phone is being used in very different ways – to "check in" and coordinate with embedded neighbors vs. to hold lengthy conversations with those at greater network distance.

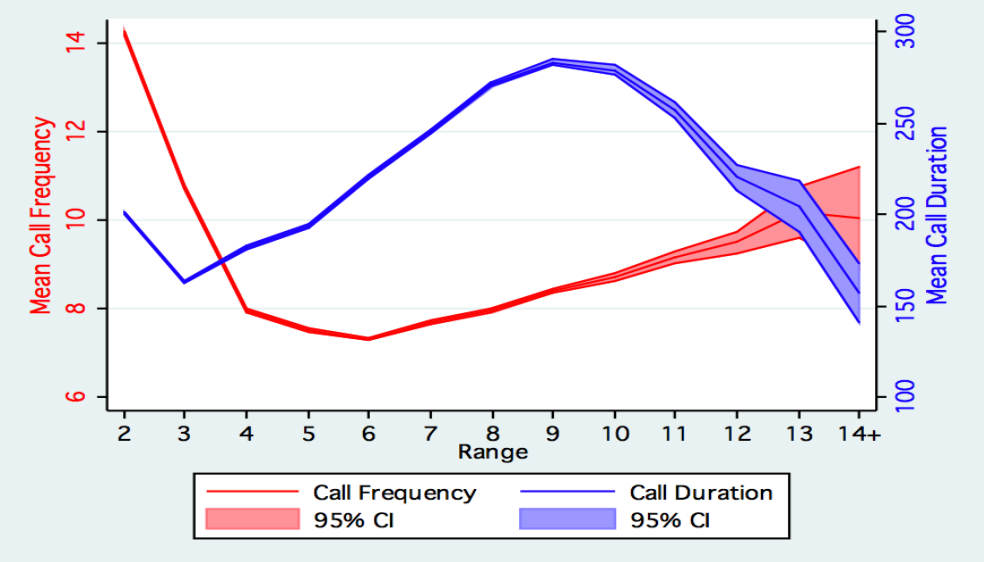


Figure 4: Mean Call Frequency and Call Duration Conditional on Tie Range.

The dyad level nonlinear relationship between tie strength and range could be confounded by node-level differences in baseline call volume. For example, a teenager who is on the phone talking to a tightly knit group of friends all the time will contribute to higher mean call volume at n = 2. I control for node-level baseline differences by standardizing call volume for each node. That is, I compute z_{ij}, the standardized call volume between i and j from i's perspective as,

$$z_{ij} = \frac{v_{ij} - \overline{v_i}}{\sigma_i}$$

where v_{ij} is the sum of raw call volumes from i to j and from j to i, $\overline{v_i}$ and σ_i are the mean and standard deviation of i's call volume, respectively. I then aggregate all z_{ij} values by their associated tie range values to get the within-range mean of the standardized call volumes. Note that z_{ij} and z_{ji} can assume different values and that both are included in the aggregation. Figure

5 shows that even after removing baseline differences in individual call volume, the nonlinear pattern persists.

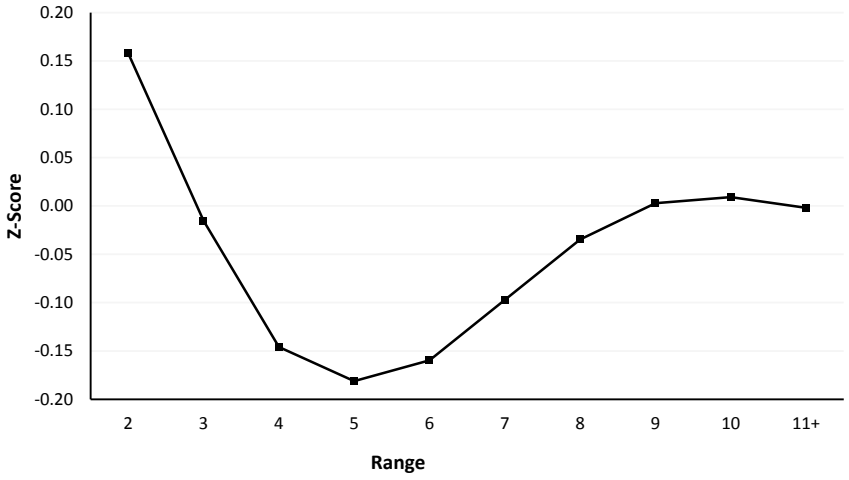


Figure 5: Mean Within-Individual Standardized Call Volume by Range.

Alternative Explanations

Degree and Time Constraint. Before turning to more substantive explanations, I first address a trivial explanation for the strength of ties that bridge large network distances – the limited time that is available to spend on the phone, regardless of the size of one's personal network. For example, a high-degree node with 15 network neighbors will have less time to communicate with each, compared to a node with only 1 neighbor, even if the more "popular" individual spends much more time each day talking on the phone (Miritello et al., 2013). Furthermore, the tie between the "isolate" and its single neighbor has indefinite length since there is no second shortest path between them. As the degree of a node and an adjacent node increases, so too does the number of indirect paths that connect them and therefore the probability of an outlier in the distribution. Since range is measured as the shortest of these indirect paths, high degree

neighbors have a statistical bias toward shorter ranges – and less time to spend on the phone with one another. Simply put, the nonlinear pattern observed in Figures 3 may be nothing more than a spurious artifact of the differences in the aggregate degree of two adjacent nodes.

I test this hypothesis by measuring the strength of ties as their range increases, holding degree constant. I first measure the degree of a tie three different ways: as the sum of the degree of the two adjacent nodes, the degree of the more popular node, and the degree of the less popular node. Since the results are qualitatively the same regardless of the measure used, Figure 6 reports median call volume broken down by degree quintiles based on aggregate degree. As expected, call volume declines with larger degree, as shown in the roughly downward shift from high to low degree quintile curves, but the nonlinear relationship with range is evident across all quintiles, especially the bottom quintile. This result suggests that the degree of the adjacent nodes do not account for the U-shape.

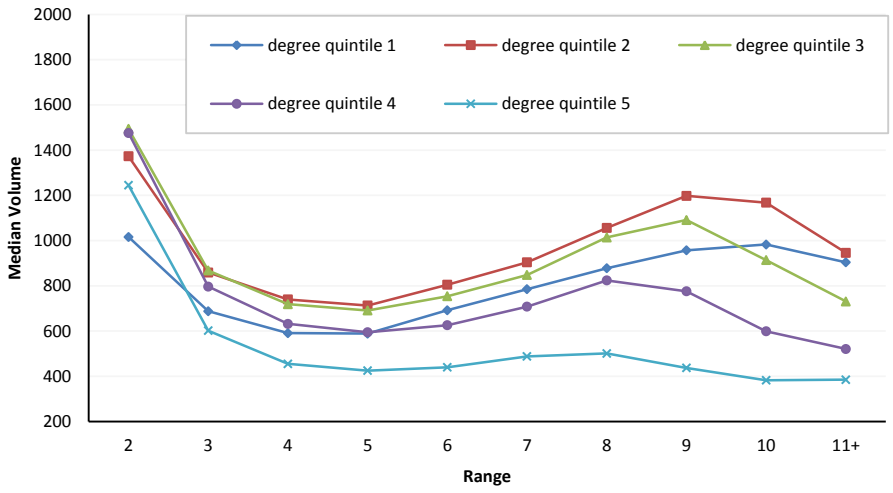


Figure 6: Median Call Volume Conditional on Range and the Sum of Node Degrees.

Affective vs. Instrumental Ties. The telephone is similar to face-to-face communication in

having both social and instrumental purposes. We interact socially as an end in itself, where the "medium is the message." We also interact instrumentally when we communicate to obtain some extrinsic end. The instrumental value inherent in bridging ties may be a reason why individuals consciously maintain long-range ties that are not reinforced or supported by mutual third parties (Burt 2002). If instrumental motives dictate certain relationships, those interacting for instrumental reasons would find bridges spanning longer network distances to be more valuable since novel information should flow through these longer bridges. It follows, then, that individuals would try to invest more time and care in fostering and sustaining longer bridges.

I test this dual-purpose hypothesis by distinguishing between calls made during business hours and those made on nights and weekends. A tie is classified as "work hour" (14.8%) if all phone calls between two parties were made exclusively during business hours, i.e. between 9 AM and 5 PM Monday to Friday, excluding national UK holidays. A tie is non-work hour (25.9%) if all calls were made exclusively during non-business hours. This distinction is based on the assumption that calls during work hours are more likely to be for business-related than for social purposes.[1] However there may be considerable asymmetry between the norm against calling business associates after hours and the norm against calling friends and family during work hours. I therefore included a hybrid category for ties where calls were made during both work hours and non-work hours (59.2%). These "anytime" ties are more likely to be affective/social than instrumental, as a stronger norm exists against calling business associates after work hours than calling an emotionally attached friend during work hours. If the strength of

[1] The work/non-work hour based categorization may be a crude and noisy proxy for the types of ties. Not all instrumentally oriented communications will occur explicitly during work hours and not all phone calls between people with affective relationships will be made during non-work hours. Nevertheless, the sheer size of the data should allow one to pick up sufficiently strong signals from this categorization despite sizeable noise. For a similar behavioral approach based on time segmentation for identifying friends vs. professional contacts, see Eagle, Pentland, and Lazer (2009).

long ties is due to the instrumental use of the phone, the high call volume of long-range ties should be most pronounced among work-hour ties.

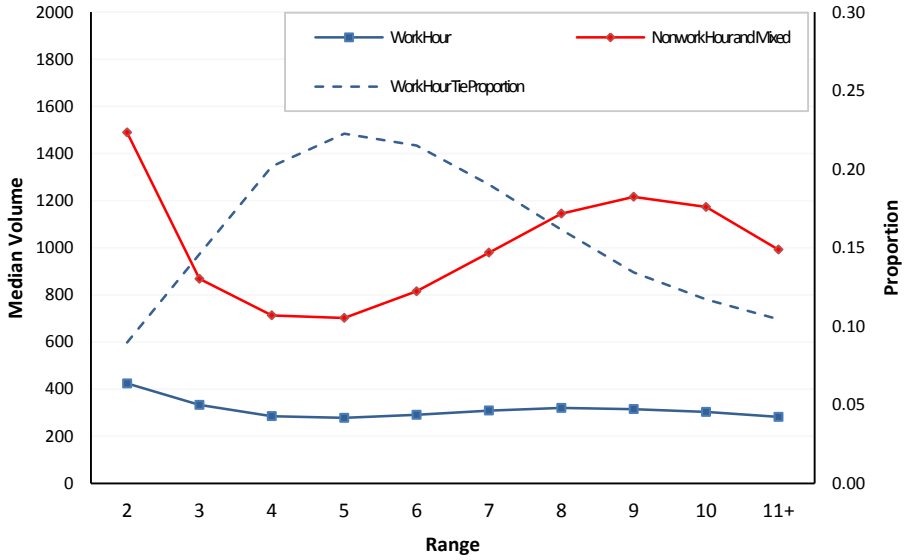


Figure 7: Median Call Volume by Types of Tie and the Distribution of Work Hour Ties within Range.

Figure 8 shows the median call volume conditional on tie range, broken down by the types of ties. I combine non-work hour ties and mixed ties since both categories are thought to be social. The results do not support the dual-purpose hypothesis that strong long-range ties reflect the use of the phone for business-related reasons. For non-work hour and mixed ties combined, the median call volume increases from a local minimum of 702 seconds at n=5 to a local maximum of 1216 at n=9 (73% increase) whereas, for work-hour ties, the median volume increases from 278 seconds at n=5 to 320 at n=8 (15% increase). In short, the surprising strength of long-range ties is more evident for relationships that include calls made after work than for those that occur exclusively during business hours, suggesting that the instrumental use of the phone is not the explanation.

Tie Stretching. When people experience social or spatial mobility throughout the life course (e.g. high school graduation), they become embedded in new social relationships, but they do not necessarily drop all their former ties. However, the ties that remain intact become "stretched" when the new and old networks do not overlap. Numerous studies have shown that strong social ties last longer and decay at a slower rate than weak ties across a wide range of time scales (Burt 2000; 2002; Dahlander and McFarland 2013; Feld, Suitor, and Hoegh 2007; Larner, 1990; Lubbers et al. 2010; Raeder et al. 2011; Wellman et al. 1997). This suggests the possibility that strong ties, such as those to close friends and kin, are more likely to survive the effects of social and spatial mobility, compared to weak ties, such as those to acquaintances.[2] When ties are stretched, it is the weak ones that are more likely to break. Viewed from within the group, the dispersion of each group member in different directions causes sparsification of ties within the group, which in turn, increases the range of those remaining ties. This selection process, propelled by mobility and diverging life courses, might explain the strength of ties that bridge the social and spatial distances.

I present suggestive evidence consistent with the above interpretation. First, the proportion of work hour ties conditional on range monotonically decreases from mid-range and onward (dashed line, Figure 7). Specifically, at n=5, work hour ties constitute 22% of all ties of this range whereas at n=10, work hour ties constitute only 11%. The relative increase of social ties (non-work hour and mixed ties) suggests the higher durability of social ties vs. instrumental ties with the passage of time.

[2] When the weaker ties break over time due to mobility, the remaining stretched ties, which become short-range bridges, also become subject to a higher probability of decay compared to embedded ties (Burt 2002). However, this probability decreases with time, such that the stretched ties that do survive the

Although the data do not include measures of either social or spatial mobility, one can use the spatial distance between adjacent nodes as an indirect test of the "tie stretching" hypothesis, based on the assumption that spatial distance may exert selection pressures similar to those that may be exerted by network distance. That assumption is supported by numerous studies showing sharp rates of tie decay (e.g. exponential and power) with respect to geographic distance across different types of offline and online social networks (Daraganova et al., 2012; Lambiotte et al. 2008; Liben-Nowell et al. 2005; Onnela et al. 2011; Preciado et al. 2012). Furthermore, there is evidence that the probability of observing embedded ties (Lambiotte et al. 2008) and the extent to which ties are embedded at a given geographic distance (Volkovich et al. 2012) also decrease as geographic distance increases up to a certain point. From these studies, it is clear that spatial distance is associated with the disembedding of ties and tie decay.

I tested two implications of the "tie stretching" hypothesis: 1) tie strength increases with spatial distance (due to selection pressures) and 2) a positive range and spatial distance interaction effect on tie strength. Tie strength measured in call volume may increase with spatial distance because maintaining a relationship is harder with less face-to-face contact. In the relative absence of face-to-face reinforcement, a relationship relying solely on phone communication is likely to dissolve or turn dormant unless the relationship is exceptionally strong. If a spatially long-distance relationship does not have third-party ties to support and sustain interaction, the relationship should be at even higher risk; hence the interaction effect.

I first show that social and spatial distances are distinct dimensions. In Figure 8, mean and median spatial distances do not increase linearly with range. The initial decrease in spatial distance from n=2 to n=5 suggests that increases in social distance are compensated by reduced

---

short term face marginally decreasing odds of breaking (Burt 2002; Lubbers et al. 2010; Martin and Yeung 2006).

spatial distance. However, beyond *n*=5, both spatial and social distance correlate positively.

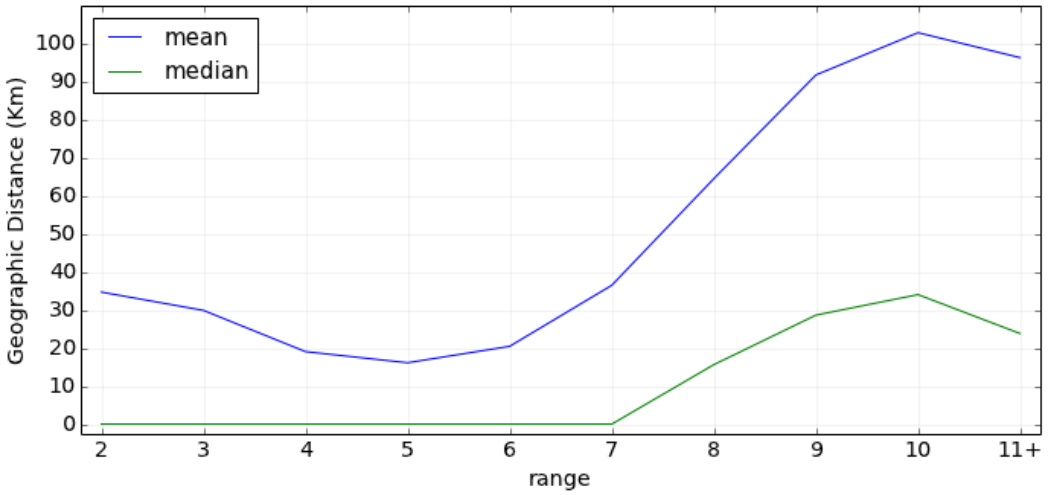


Figure 8: Geographic Distance by Tie Range

Figure 9 plots the median call volume, conditional on both range and geographic distance for phone calls made among geographically identifiable landlines. First, I find supporting evidence for the first test of the tie stretching hypothesis; call volume consistently increases with geographic distance across all tie range strata. As before, the median call volumes across tie ranges consistently draw the familiar U-shape across all spatial distance strata. This finding supports the idea that geographic distance and social (network) distance exert distinct selection pressures on the ties. Also note that the rate at which call volume increases with tie range beyond *n*=5 is higher (i.e. steeper slope) for ties that span longer geographic distances. I interpret this last observation as evidence for a compounding effect between network-distance-based and geographic-distance-based selection pressures.

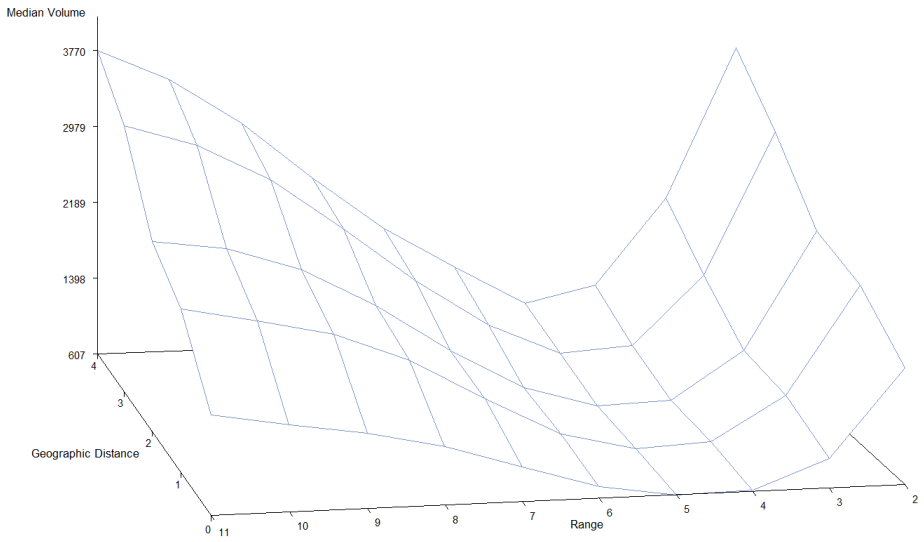


Figure 9: Median Call Volume (seconds), by Range and Geographic Distance

Discussion

Studies on the efficacy of bridging ties focused heavily on individual outcomes: an individual's probability of adoption, an individual's performance, and so on. Accordingly, researchers measured bridges primarily from the individual's perspective, from structural holes in an individual's personal network to the embeddedness of an individual's tie, proving to be sufficient for addressing research questions formulated at the individual level. Nonetheless, network research is quickly shifting towards questions that require a broader scope and that can be addressed with newly available large-scale network data. With this shift in approach and in data, viewing bridging ties in a new light that reveals their full length and strength can lead to new insights that deepen our understanding of both individual and system level outcomes.

As a first step, I used Granovetter's definition of bridging ties and showed that the lengths of those ties are bimodally distributed in a population scale communication network. I showed that the bimodality is expected in small-world networks with sufficient random rewiring and that

the strength of ties, measured by total call volume, diminishes with tie length to a certain point, but then starts increasing beyond that point. This surprising and hitherto unknown systemic behavior prompted three competing explanations: differential degree, instrumental vs. affective ties, and differential tie decay. The most plausible explanation, which agrees with results from preliminary analyses, is that ties decay at different rates depending on their strength and embeddedness.

A question remains regarding the social processes and their corresponding call patterns that produce the nonlinear strength and length relationship. That is, do old ties stay dormant for long stretches of time until a point at which the relationship revitalizes upon some occasion (McEvily, Jaffee and Tortoriello 2012; Levin, Walter, and Murnighan 2011) or do old ties that survive long-term decay maintain a steady communication volume throughout time? If the long bridges are in fact dormant ties that have been recently revitalized around the time of observation, it is possible that those previously dormant ties would exhibit bursts of phone communication during a relatively short time frame. The long and strong ties observed in the data, then, might reflect these dormant ties that engage in high volume communication, albeit for a relatively short period. Regardless of the actual process at play between the two possibilities, both explanations are built on the idea of third-party tie decay. I leave it to subsequent studies to adjudicate between the two possibilities: dormant tie revitalization and sustained contact.

This study is not without its limitations. First, the lack of qualitative dyadic information limits accurate classification of the nature of the ties (e.g. instrumental vs. affective). Although I believe that classifying types of ties based on call time is a satisficing proxy, further research based on rich qualitative information is needed for full confirmation. Furthermore, knowing the nature of each tie (e.g. when and in what context the ties were formed) would allow for testing

the selection-based mechanism of the positive relationship between tie strength and length beyond a certain range. Second, the relatively short observation window of the data may cause an inaccurate representation of the "true" underlying communication network. For example, the mean and variance of node degree could increase while the bimodality of the bridge length distribution could disappear since the network as a whole would have densified with a longer time window. However, the bimodal bridge length distribution is unlikely to disappear altogether because it is a characteristic generally observed in small-world networks with sufficient random rewiring. Also, capturing otherwise undetected communication ties in a longer observation window may not necessarily eliminate the non-linear pattern found between tie strength and length. Previously observed bridging ties might become shorter in a longer observation window, but with a longer observation window, it is also likely that previously unobserved bridging ties become observable. It is an empirical question whether the rate at which previously observed bridging ties shrink is faster or slower than the rate at which previously unobserved bridging ties become visible. To address this issue, I conducted additional tests by dividing the data into two consecutive two-week time windows and rerunning the analyses on each of them to mimic the situation where ties actually exist are not observed. Despite considerable number of missing links, I obtained qualitatively identical results in both.

References

Aral, Sinan and Marshall Van Alstyne. 2011. The Diversity-Bandwidth Trade-off. *American Journal of Sociology*, 117(1): 90-171.

Bakshy, Eytan, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. Role of Social Networks in Information Diffusion. *International World Wide Web Conference* (WWW).

Bates, Andy. 2006. Methodology Used for Producing ONS's Small Area Population Estimates. The Office of National Statistics. http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no--125--autumn-2006/population-trends-pt3.pdf.

Bott, Elizabeth. 1957. *Family and Social Networks: Roles, Norms, and External Relationships in Ordinary Urban Families*. London: Tavistock.

Burt, Ronald. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press

----------. 2000. Decay Functions. *Social Networks*, 22: 1-28.

----------. 2002. Bridge Decay. *Social Networks*, 24: 333-363.

----------. 2004. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2): 349-399.

Campbell, Karen, Peter Marsden, and Jeanne Hurlbert. 1986. Social Resources and Socioeconomic Status. *Social Networks*, 8(1): 97-117.

Dahlander, Linus and Daniel McFarland. 2013. Ties That Last: Tie Formation and Persistence in Research Collaborations over Time. *Administrative Science Quarterly*, 58(1): 69-110.

Darganova, Galina, Pip Pattison, Johan Koskinen, Bill Mitchell, Anthea Bill, Martin Watts, and Scott Baum. 2012. Networks and Geography: Modelling Community Network Structures as the Outcome of Both Spatial and Network Processes. *Social Networks*, 34(1): 6-17.

Detert, James, Ethan Burris, David Harrison, and Sean Martin. 2013. Voice Flows to and around Leaders: Understanding When Units are Helped or Hurt by Employee Voice. *Administrative Science Quarterly*, 58(4): 624-668.

Eagle, Nathan, Alex Pentland, and David Lazer. 2009. Inferring Social Network Structure using Mobile Phone Data. *Proceedings of the National Academy of Sciences*, 106(36): 15274-15278.

Feld, Scott, Jill Suitor, and Jordana Gartner Hoegh. 2007. Describing Changes in Personal Networks over Time. *Field Methods*, 19:218-236.

Granovetter, Mark. 1973. The Strength of Weak Ties. *American Journal of Sociology*, 78(6): 1360-1380.

----------. 1974. Getting a Job. Chicago, IL: University of Chicago.

----------. 1983. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1: 201-233.

Jones, Jason, Jaime Settle, Robert Bond, Christopher Fariss, Cameron Marlow, and James Fowler. 2013. Inferring Tie Strength from Online Directed Behavior. *PLoS One*, 8(1): e52168.

Kossinets, Gueorgi, Jon Kleinberg, and Duncan Watts. 2008. The structure of information pathways in a social communication network. *ACM SIGKDD*, 435-443.

Lambiotte, Renaud, Vincent Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. 2008. Geographical dispersal of mobile communication networks. *Physica A*, 387(21): 5317-5325.

Larner, Mary. 1990. Changes in Network Resources and Relationships over Time. Pp. 181-204 in *Extending Familes: The Social Networks of Parents and Their Children*. edited by Moncrieff Cochran, Mary Larner, David Riley, Lars Gunnarsson, and Charles Henderson. New York, NY: Cambridge University Press.

Leskovec, Jure, Dan Huttenlocher, and Jon Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks. *International World Wide Web Conference* (*WWW*).

Leskovec, Jure, Jon Kleinberg and Christos Faloutsos. 2007. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data* (ACM TKDD), 1(1).

Levin, Daniel, Jogre Walter, and J. Keith Murnighan. 2011. Dormant Ties: The Value of Reconnecting. *Organization Science*, 22(4): 923-939.

Liben-Nowell, David, Jasmine Novak, Ravi Kumar, Prabhakar Raghaven, and Andrew Tomkins. 2005. Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences*, 102(33): 11623-11628.

Lin, Nan and Mary Dumin. 1986. Access to Occupations through Social Ties. *Social Networks*, 8(4): 365-385.

Louch, Hugh. 2000. Personal Network Integration: Transitivity and Homophily in Strong-Tie Relations. *Social Networks*, 22: 45-64.

Lubbers, Miranda, Jose Luis Molina, Jurgen Lerner, Ulrik Brandes, Javier Avila, and Christopher McCarty. 2010. Longitudinal Analysis of Personal Networks: The Case of Argentinean Migrants in Spain. *Social Networks*, 32: 91-104.

Marsden, Peter and Karen Campbell. 1984. Measuring Tie Strength. *Social Forces*, 63(2): 482-501.

Martin, John Levi and King-To Yeung. 2006. Persistence of Close Personal Ties over a 12-Year Period. *Social Networks,* 28: 331-362.

McEvily, Bill, Jonathan Jaffee, and Marco Tortoriello. 2012. Not All Bridging Ties Are Equal: Network Imprinting and Firm Growth in the Nashville Legal Industry, 1933–1978. *Organization Science,* 23(2): 547-563.

Miritello, Giovanna, Esteban Moro, Ruben Lara, Rocio Martinez-Lopez, John Belchamber, Sam Roberts, and Robin Dunbar. 2013. Time as a Limited Resource: Communication Strategy in Mobile Phone Network. *Social Networks,* 35(1): 89-95.

Montgomery, James. 1994. Weak Ties, Employment, and Inequality: An Equilibrium Analysis. *American Journal of Sociology,* 99(5): 1212-1236.

Office for National Statistics. 2012. "2011 Census: Population and Household Estimates for the United Kingdom." UK Data Service Census Support. Retrieved Nov. 14, 2013 (http:// http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-the-united-kingdom/index.html).

Oliver, Pamela and Daniel Myers. 2003. Networks, Diffusion, and Cycles of Collective Action. Pp. 173-203 in Social Movements and Networks: Relational Approaches to Collective Action. edited by Mario Diani and Doug McAdam. New York, NY: Oxford University Press.

Onnela, Jukka-Pekka, Samuel Arbesman, Marta Gonzalez, Albert-Laszlo Barabasi, and Nicholas Christakis. 2011. Geographic Constraints on Social Network Groups. *PLoS One*, 6(4): e16939.

Onnela, Jukka-Pekka, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. -L. Barabasi. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18): 7332-7336.

Preciado, Paulina, Tom Snijders, William Burk, Hakan Stattin, and Margaret Kerr. 2012. Does Proximity Matter? Distance Dependence of Adolescent Friendships. *Social Networks*, 34: 18-31.

Raeder, Troy, Omar Lizardo, David Hachen, and Nitesh Chawla. 2011. Predictors of Short-Term Decay of Cell Phone Contacts in a Large Scale Communication Network. *Social Networks,* 33: 235-257.

Reagan, Ray and Bill McEvily. 2003. Network Structure and Knowledge Transfer: The Effects of Cohesion and Range. *Administrative Science Quarterly*, 48(2): 240-267.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman. 2012. Geography of Twitter Networks. *Social Networks*, 34(1): 73-81.

Tiwana, Amrit. 2008. Do Bridging Ties Complement Strong Ties?: An Empirical Examination of Alliance Ambidexterity. *Strategic Management Journal*, 29(3): 251-272.

Traud, Amanda, Peter Mucha, and Jason Porter. 2011. The Social Structure of Facebook Networks. *arXiv*:1102.2166v1 [cs.SI].

Uzzi, Brian and Jarrett Spiro. 2005. Collaboration and Creativity: The Small World Problem. *American Journal of Sociology*, 111(2): 447-504.

Volkovich, Yana, Salvatore Scellato, David Laniado, Cecilia Mascolo, and Andreas Kaltenbrunner. 2012. The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions. *ICWSM*, 346-353.

Watts, Duncan. 1999. Networks, Dynamics, and the Small-World Phenomenon. *American Journal of Sociology*, 105(2): 493-527.

Watts, Duncan and Steven Strogatz. 1998. Collective Dynamics of 'Small-World' Networks. *Nature*, 393: 440-442.

Wellman, Barry, Rnita Yuk-lin Wong, David Tindall, and Nancy Nazer. 1997. A Decade of Network Change: Turnover, Persistence, and Stability in Personal Communities. *Social Networks*, 19: 27-50.

Yakubovich, Valery. 2005. Weak Ties, Information, and Influence: How Workers Find Jobs in a Local Russian Labor Market. *American Sociological Review*, 70(3): 408-421.

Yang, Jaewon and Jure Leskovec. 2012. Defining and Evaluating Network Communities Based on Ground-Truth. *IEEE International Conference on Data Mining* (ICDM).

Zachary, Wayne. 1977. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research*, 33(4): 452-473.

CROSS-CULTURAL COMPARISON OF THE STRENGTH OF LONG TIES

Abstract

In an eight-country cross-cultural comparative analysis, I report the surprising discovery of "social wormholes" that allow information, beliefs, and opinions to travel between close friends across vast network distances. Previous research has hypothesized the existence of weak social ties between acquaintances that create shortcuts through social space, thereby promoting access to otherwise unavailable information, greater social integration, and more rapid diffusion of information. However, the global network data needed to identify and fully measure these shortcuts to test the weak tie hypothesis have not been available until recently. Furthermore, research in cross-cultural psychology which predicts systematic differences in the prevalence and relational strength of these shortcuts across cultures have not been verified empirically. Using communication network data for 56M Twitter users from eight countries with high Twitter penetration, I confirm the existence of long-range ties spanning far network distances, but also discover that these ties are much stronger than previously theorized. I examine three explanations: 1) users with few communication neighbors tend to invest heavily in their relations with one another but with a lower probability of having a neighbor in common; 2) users purposefully maintain long-range ties to gain instrumental benefits; 3) geographically long-distance network neighbors who do not share common third-party neighbors use Twitter as a complement to face-to-face interaction. Finally, I present a tentative fourth explanation based on differential rates of tie decay as a direction for future research.

Introduction

A century of research has shown how the structure of social networks constrains and enables individual life chances, shapes the timing and pattern of collective outcomes, and influences the emergence and diffusion of norms and institutions. At the individual level, research on social capital has demonstrated economic returns to network location (Burt 1995; Burt, Hogarth, and Michaud 2000; Coleman 1988). At the collective level, the spread of disease, beliefs, innovations, and social movements has been shown to depend on the topological characteristics of social networks (Christakis and Fowler 2009). The precise mathematical expressions of social network analysis have armed social scientists with a formalized vocabulary for describing and testing the influence of structural forces that have sparked the sociological imagination since Durkheim.

Until recently, however, the need to collect relational data through direct observation has limited studies of social networks to small groups such as clubs (Zachary 1977), towns (Entwisle et al. 2007), schools (Moody 2001), and organizations (Davis and Greve 1997). Network structure at the population level could only be measured and estimated from survey data on egocentric networks (a randomly chosen node and its network neighbors). This methodology can be useful for studying the attributes of network nodes (such as degree), and edges (such as tie strength). However, egocentric network data have serious limitations. First, the name generator (Marsden 1990) employed for data collection limits the number of alters one can specify, which has been shown to introduce measurement error in the data (Holland and Leinhardt 1973). Second, related to size constraints, the network ties are typically limited to strong ties such as kin or non-kin confidents. Third, the presence of a tie between two alters of a respondent is usually inferred from the respondent's perception, which has been shown to be biased by personality traits and the attribute similarities of the two alters in question (Flynn, Reagans, and Guillory

2010).

In short, it is far easier for researchers to observe friends than to observe friendship. Respondents can answer survey questions about their friends, but relationships exist independently of the opinions, beliefs, and knowledge of the participants. Population-level measures of social networks (such as clustering, connectivity, or polarization) are difficult to infer from individual perceptions of the relational patterns among one's closest friends and family. As a result, there are no comparative studies of the structure of social networks at the global level.

Our ability to measure network structure at the population level is finally catching up with the need, thanks to the rapid increase in the use of digital technologies that generate time-stamped records of social interactions, from cell phone calls to credit cards to email and social media. Initial studies of large online networks, composed of millions of nodes and billions of edges, have been conducted mainly by computer and information scientists, applied mathematicians, and socio-physicists with the technical skills to collect and analyze massive semi-structured datasets. Accordingly, these studies have focused on identifying law-like regularities, general mechanisms, and universal principles that apply to networks across diverse entities, species, and contexts. For example, the preferential attachment model (which explains widely-observed power-law degree distributions) and small world network models (that are highly clustered as well as closely connected) have been applied across radically different social and natural network domains, including C. Elegans, the World Wide Web, the food web, the power grid, transportation networks, scientific collaborations, and communication networks (Barabási 2009; Barabási and Albert 1999; Watts and Strogatz 1998).

In a similar universalist spirit, mathematicians, sociologists, and social psychologists

have also proposed universal principles of structural balance (Cartwright and Harary 1956) and homophily (McPherson and Smith-Lovin 1987; McPherson, Smith-Lovin, and Cook 2001) to explain the widely observed tendency for social networks to be composed of densely connected homogenous clusters. Structural balance is based in part on the human psychological need to minimize cognitive dissonance and explains why network relations tend to be transitive (i.e., if A is friends with B and B with C, then A and C will also be friends). Byrne's (1971) "law of attraction" is based on the greater ability and desire to interact with similar others. Theory and research on social capital have also focused on the generalizability of the social and economic effects of network structure, such as the utility of weak ties for getting a job, the information and bargaining advantages of bridging "structural holes" (or open triads), and the social support derived from cohesive subgroups.

Rather than looking for universals, an alternative approach has focused on documenting and explaining structural differences between diverse social systems. Sociologists have looked for personality traits that may explain differences in egocentric networks, and for ways that culture (Pachucki and Breiger 2010; Mehra et al. 2001; Yeung 2005; Emirbayer and Goodwin 1994; Lizardo 2006; Vaisey 2010), technology (Licoppe and Smoreda 2005), mobility (Macy and Sato 2002; South and Haynie 2004), and community context may be associated with variations in network structure.[3] For example, cross-cultural studies of trust and trustworthiness point to differences in the structure of social networks as an explanation (Allik and Realo 2004; Kuwabara et al. 2007; Yamagishi, Cook, and Watabe 1998; Yamagishi and Yamagishi 1994). Yamagishi explains lower levels of generalized trust in collectivist Japan compared to the individualist US by arguing that Japanese tend to have "networks of committed relations"

(Yamagishi and Yamagishi 1994:160). This cohesive network structure motivates trustworthy behavior by providing individuals with higher monitoring and sanctioning capabilities (Hechter 1987) that pre-empt the need for generalized trust.

Based on the current state of knowledge from comparative social network research, this study seeks to extend the horizons – from within-country to between, and from survey-based egocentric network data to behavioral data on social interactions on a global scale, with a focus on structural similarities as well as theoretically meaningful differences that exist across diverse populations. This comparative study of network topology across diverse cultures can contribute to a fundamental theoretical debate on structure vs. culture: Does the structure of social networks express universal, even mathematical, principles that are independent of cultural and economic differences between societies or are network structures shaped by the underlying culture that guides micro-level network formation processes? In particular, I use the digital traces of communication on the microblogging site, Twitter, to conduct a cross-national comparative analysis on the distribution and relational strength of bridging ties that span different network distances. In sharp contrast to the extensive literature on the bridging potential of relationally weak ties and on their associated benefits to individuals, research has neglected basic quantitative aspects of bridging ties such as the span-length and relational strength distributions in population scale networks. Twitter as a global communication platform that is used in over 200 countries not only allows for the analysis of bridging ties in a communication network within a country, but also enables one to do it comparatively.

[3] Other notable studies employing a comparative approach to social networks include Baldassarri and Diani (2007) and studies using the AddHealth data (Moody 2001; Currarini, Jackson, and Pin 2010;

The Strength of Long Ties

The "Strength of Weak Ties" (SWT) thesis, first proposed by Mark Granovetter in the early 1970's, is one of the most frequently and broadly cited papers in the social sciences to date. His counter-intuitive thesis is that relationally weak social ties, which are less prone to triadic closure, tend to bridge far away pockets in the overall social network. The counter-intuitive implication of these weak bridging ties is that novel, non-redundant information that tend to exist outside of one's local network neighborhood are likely to be transmitted through relationally weak ties (Granovetter 1973, 1983; Campbell, Marsden, and Hurlbert 1986; Lin and Dumin 1986; Montgomery 1994). The usefulness of weak ties have been tested in a number of different contexts, from labor markets (Granovetter 1973; Yakubovich 2005), entrepreneurship (Singh 2000; Burt 1995), social movements (Oliver and Myers 2003), knowledge and information sharing (Bakshy et al. 2012; Hansen 1999), to the generation of new ideas (Burt 2004). Recently, research on the strength of social ties (Reagan and McEvily 2003; Tiwana 2008) highlight complementary functions of strong ties (e.g. providing social support, exchanging complex, tacit knowledge) vs. weak ties (e.g. obtaining simple, novel information). Other research focusing on the information exchange aspect of the strength of ties proposes a trade-off in the volume of information that can flow through strong embedded ties and the non-redundancy of information that is provided by weak bridging ties (Aral and Alstyne 2011). The trade-off argument suggests that even though an individual tends to receive redundant information through strong embedded ties at a higher rate, the larger "bandwidth" of those strong ties could offset this rate such that the absolute volume of non-redundant information that reaches the individual through strong ties exceeds that of weak ties. In general, these recent developments in the literature maintain the dichotomous view of strong, embedded ties vs. weak, bridging ties as Granovetter formulated in

---

Allcott et al. 2007; Goodreau et al. 2009).

the 1970's.

Formally, I denote the network distance, or "range" that a tie bridges in a network as d_{ij}, of a tie between node i and node j and measure it as the second shortest path length of that edge (Granovetter 1973; Kossinets, Kleinberg, and Watts 2008; Watts 1999). It is the number of intermediary ties required to reach from one node to the other if the tie that directly connects these two nodes were to be removed. By this definition, the minimum possible range d_{min} is 2 (i.e. two nodes have at least one common neighbor, forming a closed triad) and has been referred to as "embedded" ties. Ties of $d > 2$ are network bridges since a bridge must span between more than two non-adjacent nodes. Note that d is defined only for pairs of nodes that share a tie whereas path lengths, which are commonly used to characterize networks, are defined for both connected and unconnected pairs of nodes.

With few exceptions (e.g. Centola and Macy 2007; Kossinets et al. 2008; Onnela et al. 2007), research has centered on the dichotomy of embedded vs. bridging ties at the individual unit of analysis, to the near exclusion of the full variations in range (d), leaving a blind spot in the study of diffusion. The problem is empirical. Embeddedness, or the lack thereof, can be measured at the local network level, using standard survey methods for collecting ego-centric network data. In contrast, observing the variation in the network distance that ties bridge require data for the entire social network, which until recently have been only possible for very small networks that rarely contain long bridging ties (Bott 1957; Zachary 1977). Based on this dichotomy, all bridging ties have been conceived as homogenous, even when the theory of bridging ties implies that the network distance spanned by a tie should be directly correlated with the novelty of the information that flows through them.

Compounding to our lack of knowledge regarding the range of bridging ties is the

variation in their relational strength. In the absence of data, it is reasonable to speculate that if short bridges are weaker than embedded ties, then surely, longer bridges of large d that span long network distances should be even weaker than the short ones. While this speculation seems reasonable, it is important to systematically assess the relationship between the strength and length of bridging ties due to its potential significance for understanding the dynamics of high-cost contagion (Centola and Macy 2007). Long-range bridging ties that may be relationally weak have the potential as network wormholes that accelerate low-cost diffusion (e.g. gossip) into far-away corners of the network. On the other hand, weak bridges may not be consequential for high-cost, high-risk diffusion (e.g. purchasing costly products, participating in risky social movements) because high-cost diffusion from one node to another typically requires reassurance from multiple network neighbors (Centola and Macy 2007) or from an entrusted strong-tie neighbor for adoption (McAdam 1986). However, if long-range bridging ties are in fact relationally strong, they could solve the reassurance problem and function as conduits through which high-cost diffusion occurring in one part of the network gets jump-started in other distant parts that would otherwise be affected later in the diffusion life-cycle.

Individualism, Collectivism, and Network Structure

In his seminal study, Hofstede (1991) defines individualism and collectivism as follows:

Individualism pertains to societies in which the ties between individuals are loose: everyone is expected to look after himself or herself and his or her immediate family. Collectivism as its opposite pertains to societies in which people from birth onwards are integrated into strong cohesive ingroups, which throughout people's lifetime continue to protect them in exchange for unquestioning loyalty (p. 51).

Individualist cultures, represented by North America and Western Europe,[4] emphasize the independence, self-sufficiency, and autonomy of the individual in one's perception of the self as well as of one's relationship with others. Collectivist cultures represented by East Asian countries emphasize social and emotional interdependence between people, group solidarity, the priority of group goals, and particularism (Kim et al. 1994). According to Markus and Kitayama (1991), an individualist sees the self as separate and independent. The distinction originates from an "internal repertoire of thoughts, feelings, and action, rather than by reference to the thoughts, feelings, and actions of others" (p. 226). On the other hand, a collectivist's sense of self derives from the interconnected web of social relationships in which one is embedded. Therefore, "one's behavior is determined, contingent on, and to a large extent organized by what the actor perceives to be the thoughts, feelings, and actions of others in the relationship" (p. 227).[5] Cross-cultural studies by social psychologists provide support for these distinctions by showing that consistency of the self is less important for East Asians than for Westerners. Rather than striving to maintain a consistent perception of self, irrespective of relational context, East Asians, who tend to have an interdependent self-construal, seek to resolve inconsistency and cognitive dissonance arising in the group and in the interpersonal relationships in which they are embedded (Hoshino-Browne et al. 2005; Spencer-Rodgers et al. 2007; Nisbett 2003).

Research also suggests that individualist and collectivist orientations may have implications for the formation of social ties. In particular, collectivists have been shown to have fewer but more durable and embedded relationships. Studies based on data collected at the individual level, either through surveys or small-group laboratory experiments, have found that,

[4] Although Western cultures generally exhibit stronger individualist cultural traits, there is also evidence of variability within North American and European countries (Oyserman et al. 2002).

compared to collectivists, individualists tend to affiliate with a larger number of in-groups (Markus and Kitayama 1991; Verma 1985), engage less in group interaction (Wheeler, Reis, and Bond 1989), have less emotional attachment or dependence to the in-group (Triandis, McCusker, and Hui 1990), allocate time more evenly between in-group and out-group members (Gudykunst et al. 1992)**,** and report less difficulty interacting with strangers (Gudykunst, Yang, and Nishida 1987).

Most comparative studies of cultural values and network structure have relied on surveys to collect data on egocentric network structure (Fischer and Shavit 1995; Grossetti 2007; Kalish and Robins 2006; Mehra et al. 2001; Oh and Kilduff 2008). The concomitant limitations of egocentric network designs may explain inconsistent results from study to study. For example, Igarashi et al. (2008) used surveys to obtain egocentric network data from England, Australia, Germany, Japan, and Korea. They found no significant differences in network density, but Japanese and Korean participants, who presumably embody collectivist cultures, exhibited higher betweenness centrality. In contrast, Kalish and Robins (2006) compared individuals who varied on individualist orientation and found that those with less individualist orientations were more likely to have strong social ties that are highly embedded within densely connected clusters.

In the organizational literature, studies have tried to demonstrate the importance of I/C as cultural contexts that structure individual incentives, with implications for network tie formation as well as the emergence of social capital (Burger and Buskens 2009). Xiao and Tsui (2007) provide one of the most striking examples of the importance of cultural context. While the organization literature has repeatedly documented the advantage of bridging "structural holes" (Burt 1995; Burt, Hogarth, and Michaud 2000) in the individualist context of U.S. organizations,

[5] I/C as cultural constructs closely relate to the personality constructs of I/C at the psychological level (Bond 1988; Schwartz 1994), also termed "idiocentrism" and "allocentrism" in order to differentiate the

Xiao and Tsui argue that those in bridge positions in Chinese companies are often viewed by their colleagues as untrustworthy. The authors also note that structural holes (or open triads) close rather quickly within Chinese firms characterized by a collectivist corporate culture. In line with these results, Chai and Rhee (2010), based on Podolny and Baron's (1997) argument that bridging positions can constrain ego with competing demands and norms, add that a brokering individual's tolerance for contradictory demands may differ by I/C. Together, the sensitivity to normative and role inconsistencies and the different level of tolerance for brokering behavior suggest that individuals in collectivist cultures will have a larger incentive to close open triads compared to those in individualist cultures.

From a cross-cultural comparative perspective, the above studies generally suggest a relative prevalence of bridging ties in individualist countries where social relationships are less bounded within groups. In addition, the ease of interacting with outgroup members and strangers for individuals in individualist cultures also implies that the bridging ties that exist in these cultures might be stronger in relational strength and could span longer network distances compared to those ties in collectivist cultures.

In sum, the characteristics of bridging ties, namely their range distribution and tie strength-range correlation, have not been examined systematically in a wide range of works whose social processes center on the STW thesis - social contagion, diffusion, collective action, and social integration. Furthermore, how the strength of bridging ties may vary with culture, despite arguments that would suggest that they may, have received even less attention among network scientists. To the extent that they do vary with culture, one may be able to gain a deeper understanding towards the cultural influence on the dynamics *of* networks (e.g. tie formation and decay processes) as well as of the cross-cultural differences in the dynamics *on* networks (e.g.

---

two levels of analyses (Triandis 1989).

high-cost diffusion). I address these gaps by analyzing Twitter bidirected @mention networks constructed from a large-scale data crawl of 158M user accounts worldwide. Using this corpus of data, the current study focuses on the analysis of eight countries (United States, United Kingdom, France, the Netherlands, Japan, Korea, Singapore, and Turkey) with relatively high Twitter penetration, but varying in cultural characteristics (see Materials and Methods).

Results

Characteristics of the Networks. As commonly observed in social networks, the @mention networks of the eight countries exhibit similarly skewed degree distributions (Figure 10A) and higher-than-random levels of local clustering (Figure 10B). Two of the three collectivist countries (i.e. South Korea and Singapore) show higher mean and variance in local clustering, which is consistent with the findings from previous studies in cross-cultural psychology. Nevertheless, Japan, the exemplary collectivist nation in the literature, exhibits one of the lowest mean and variance among the nations under study. These mixed results make it difficult to attribute differences in clustering solely on the basis of cultural differences.

The tie strengths, as measured by the bidirected mention frequencies of each tie, also exhibit a heavily skewed distribution (Figure 10C). reflecting the typically skewed communication volume distributions across alters at the individual level (e.g. Saramäki et al. 2014). Finally, as shown in Figure 10D, I find that the range distributions are heavily skewed in all countries with $d = 2$ constituting the vast majority. Excluding the embedded ties (i.e. $d = 2$), the range distributions are best fitted with gamma distributions that decline at a slower rate than exponential (See SI Appendix for details), suggesting the relative prevalence of extreme ranges in these networks. I find similar long-tailed range distributions in a number smaller social

and communication networks, which suggests that there may be a common generative process producing these distributions (See SI Appendix for details). Taken together, the eight within-country networks do not seem to exhibit stark differences in their basic characteristics of degree, clustering, tie strength, and range distributions.

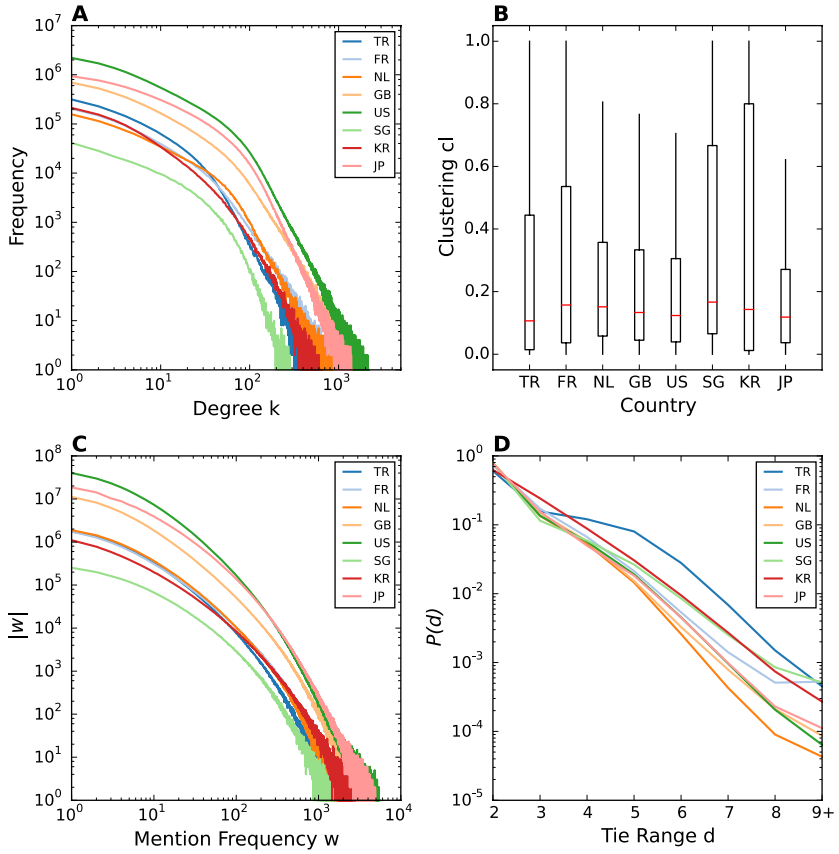


Figure 10. Description of the Bidirected Mention Networks of Twitter Users in Eight Countries (Turkey, France, Netherlands, United Kingdom, United States, Singapore, Korea, and Japan). (*A*) Degree distribution. (*B*) Distribution of local clustering coefficients. Asian countries tend to have higher mean and variance. (*C*) Distribution of tie strength. Tie strength, w_{ij} is measured by the sum of mentions in both directions $(w_{i\to j} + w_{j\to i})$ on each edge. (*D*) Distribution of range, *d*. The range distributions are best approximated by gamma distributions, $\Gamma(k, \theta)$ with $\bar{k} = 1.51$ (std=0.34) and $\bar{\theta} = 0.78$ (std=0.15) for the eight countries. $\bar{k} > 1$ suggests that the range of bridging ties decays slower than the reported exponential decay ($\bar{k} = 1$) of bridging ties with respect to time (Burt 2002).

Strength and Range of Ties. The primary measure of tie strength, $Ln(w_{ij})$ is the logarithm of the number of @mentions exchanged between two adjacent nodes, i and j in the bidirected @mention network. As expected and reported in previous studies, Figure 11A shows that ties that are embedded in fewer common neighbors tend to have lower mean $Ln(w)$. Surprisingly, however, Figure 11B shows the exact opposite pattern for the range of bridging ties where mean $Ln(w)$ tends to increase with d for $d > 2$. While all countries exhibit this pattern, they are much more pronounced in the three Asian countries (Singapore, Korea, and Japan). The striking increases in mean $Ln(w)$ in these countries may reflect genuine cultural differences, but other explanations are also possible. For example, the higher proportion of retweets among users in Western countries suggests that Twitter is used more as an impersonal news distribution channel (See SI Appendix Table 2). On the other hand, users in the East Asian countries show higher proportions of original tweets, which reflect the idiosyncratic uses of Twitter as a personal communication platform. The generally higher range-specific means for the three Asian countries in Figure 11B corroborates this point. The higher overall mean frequency of @mentions in these countries, in turn, means generally larger variations less affected by floor effects that may be masking true variability in the Western countries.

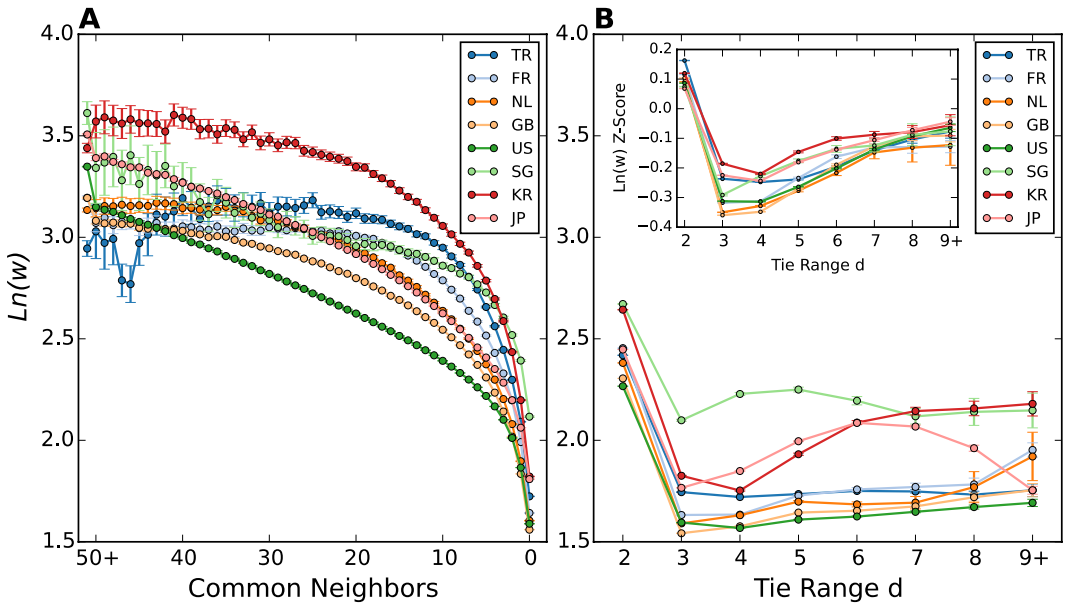


Figure 11 Mean Tie Strength and 99% CI by Network Distances for Eight Countries. Network distance is measured as (*A*) the number of common neighbors and (*B*) tie range, d for eight countries. Tie strength is measured as the log of bidirected mentions on each edge. The X axes of the two panels can be thought of as a continuum of social distance, where the far left (i.e. 50+ common neighbors) represents very short social distance while the far right (i.e. d=9+) represents extremely long social distance. In panel *B*, d=2 represents all embedded edges in panel A and its corresponding $Ln(w)$ is the grand mean tie strength of all those embedded edges. The inset in panel B plots the mean and 99% CI of the within-individual standardized tie strengths, $Ln(w)$. The increasing strength with increasing range persists, and even becomes more prominent, after accounting for individual differences in baseline mention frequency and degree.

Nodal Degree. A trivial explanation for the increasing $Ln(w)$ with d is that the network degrees of individual nodes cause a spurious correlation between tie strength and range. As the degrees of two adjacent nodes increase, so too do the number of indirect paths that connect them and, therefore, the probability of an outlier in the distribution. Since range is measured as the shortest of these indirect paths, low degree neighbors have a statistical bias toward longer ranges (Figure 12A) – and more time to message one another. Simply put, the increasing tie strength observed in Figure 11B might be nothing more than a spurious artifact of the differences in the aggregate degree of two adjacent nodes. I test this explanation by netting out individual-level

baseline differences in @mention frequencies. Specifically, the standardized mention frequency, z_{ij} between i and j is computed as, $z_{ij} = \frac{Ln(w_{ij}) - \overline{Ln(w_i)}}{\sigma_i}$, where w_{ij} is the sum of mentions from i to j and from j to i, $Ln(w_i)$ and σ_i are the mean and standard deviation of i's log-transformed bidirected mention frequencies across all her neighbors, respectively. I aggregate all z_{ij} values by their associated tie range values to get the within-range distribution of the standardized mention frequencies. Note that z_{ij} and z_{ji} can assume different values and that both are included in the aggregation. The inset of Figure 11B plots the mean and 99% confidence interval for z_{ij}, showing that the positive relationship between tie strength and range for $d > 2$ persists. Furthermore, unlike the weaker pattern observed for the non-East Asian countries in the unstandardized plot (Figure 11B), the standardized plot reveals unambiguous increases in tie strength across all countries.

Instrumental Ties. An alternative functionalist explanation for the strength of long-range ties is that individuals consciously foster long-range ties for the instrumental benefits (e.g. non-redundant and useful information) that they may provide (Burt et al 1998; Oh and Kilduff 2008; Sasovova 2010). That is, actors might invest more effort into maintaining and nurturing longer-range ties that structurally provide information and control advantages. If this assumption is true, the positive correlation between tie strength and range should be observed predominantly among instrumentally oriented ties rather than social or affective ties.

I test this explanation by comparing the relationship between tie strength and range by the likelihood of instrumental orientation of the ties. In order to classify the relational orientation of each tie, I adapt the technique introduced by Toole et al. (2015) to estimate the likelihood of face-to-face interaction during work-hours and off-work hours using GPS-tagged tweets. A pair

of individuals who form a purely instrumentally oriented relationship (e.g. coworkers and business ties) is less likely to meet face-to-face during off-work hours and weekends. Hence, if two Twitter users are observed in similar locations exclusively during work hours, they are potentially coworkers who do not socialize much out from work. For the subset of @mention ties where both nodes generated GPS-tagged tweets, I measure the likelihoods of face-to-face interaction during work-hours and off-work hours separately by computing the cosine similarity of tweet locations between the nodes. Specifically, for a tie, e_{ij}, I first create a $2 \times l$ time-location frequency matrix for user i, $\mathrm{A}_{2l}(i)$, where the two rows represent work hours (9AM to 5PM from Monday to Friday) and off-work hours (5PM to 9AM the following day during weekdays and any time during the weekends) respectively. The l columns represent the unique locations from which GPS-tagged tweets were created by either user i or j. A cell in $\mathrm{A}_{2l}(i)$ is then the frequency that user i tweeted at a specific location during a time interval l. I pair the corresponding row vectors of i and j and compute the cosine similarities in their tweet locations. This procedure gives two similarity score values, each of which corresponds to the similarities in i's and j's locations during work-hours and off-work hours, respectively. I interpret these similarity scores as a coarse measure of the likelihood of face-to-face interaction.

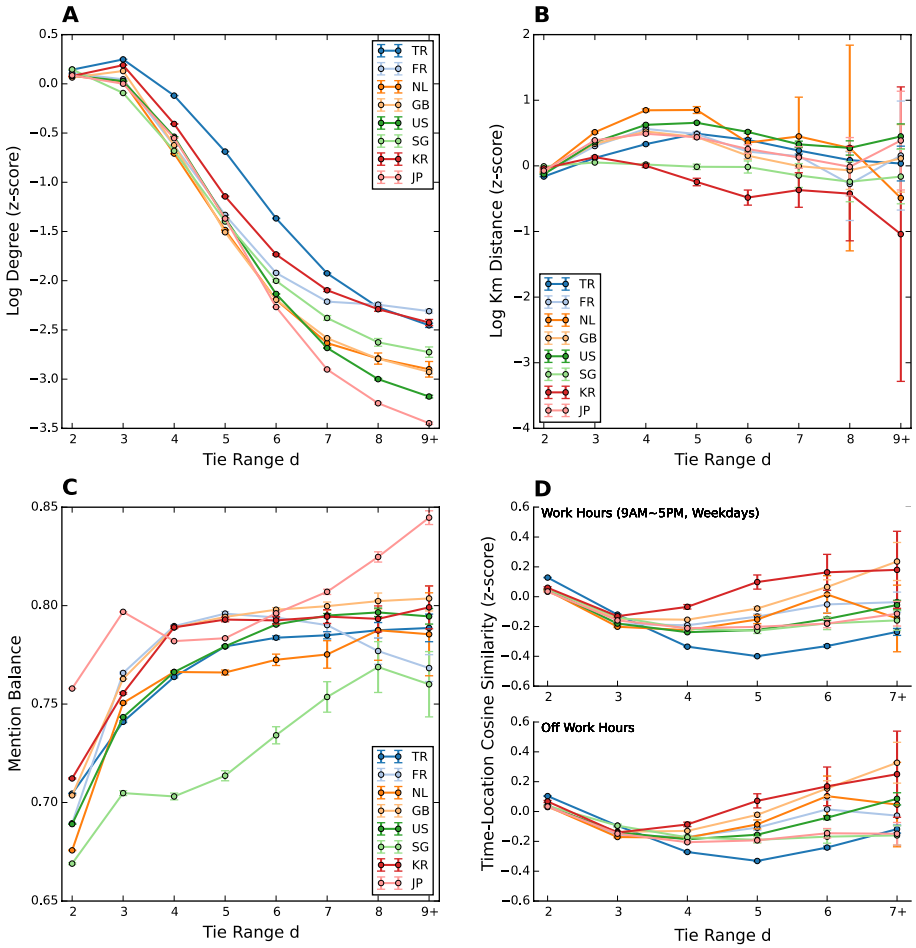


Figure 12 Bivariate Relationships Involving Range, d. Error bars indicate 99% CI of the means.

Figure 12D plots the mean and 99% confidence interval of the work hour and off-work hour cosine similarities aggregated within range. There is not a notable difference in the range-contingent likelihood of face-to-face interaction during work-hours vs. off-work hours. To identify the ties that have high likelihood of face-to-face interaction exclusively during work-hours, I use the two cosine similarity scores per tie for k-means clustering with four forced categories (k=4). The four categories correspond to different combinations of work-hour vs. off-work hour similarity scores. For convenience, I label these categories as "social-family" ties (high work-hour similarity and high off-work hour similarity), "social" ties (low work-hour

similarity and high off-work hour similarity), "coworker" (high work-hour similarity and low off-work hour similarity), and "acquaintance" (low work-hour and off-work hour similarity). Figure 13 shows that, for all four categories derived from the US network, tie strength tends to increase as bridge range increases. The same pattern is observed in the other seven country networks. Given that the positive correlation is not unique to coworker ties, the data do not lend support for the instrumental tie hypothesis.

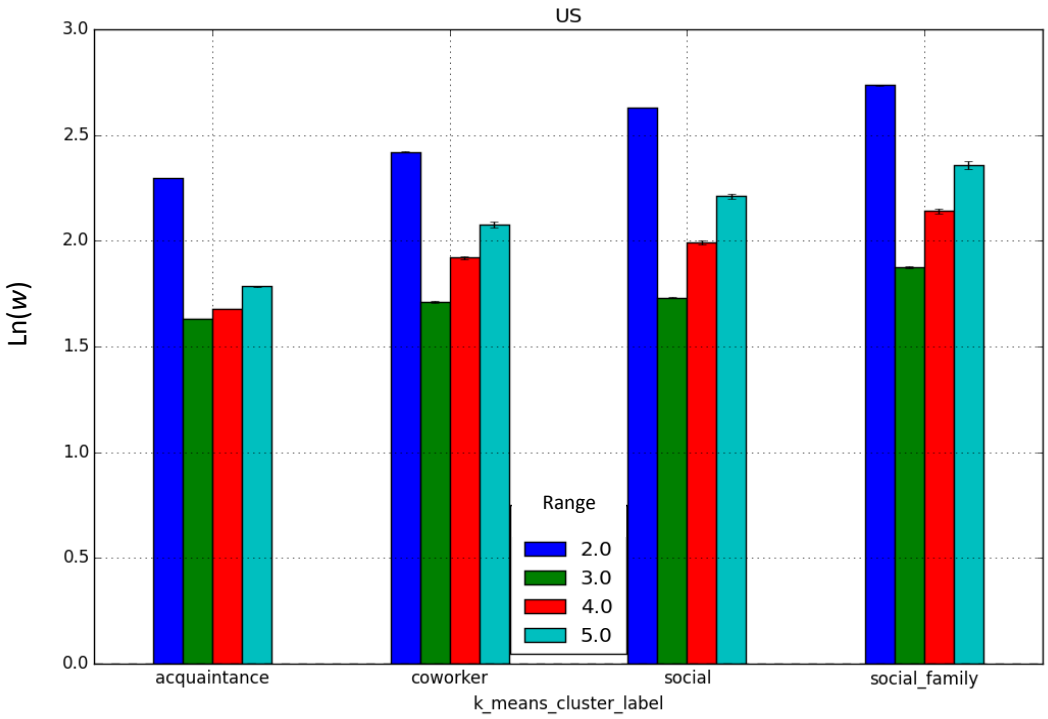


Figure 13. Mean and 99% CI of Tie Strength for Acquaintance, Coworker, Social, and Social-Family Ties for the United States. Social and social-family ties tend to exhibit slightly higher Ln(w) than acquaintance and coworker ties. The increases in Ln(w) for longer range ties is observed across all four categories.

Geographic Distance. Another possible explanation is that geographic distance causes a spurious correlation between the strength and range of bridging ties. Evidence from previous

research suggests that the probability of observing embedded ties decreases with geographic distance (Lambiotte et al. 2008) and the extent to which ties are embedded (i.e. number of common neighbors) decreases as geographic distance increases up to a certain point (Volkovich et al. 2012). If physical distance is related to tie embeddedness, which is a local measure of network distance, it is also possible that physical distance correlates with range, which is a global measure of network distance. Indeed, Figure 12B which plots the within-country standardized log-transformed geographic distance conditional on range shows that increasing tie range generally associates with increasing geographic distance for $d < 5$. However, the pattern reverses for $d > 5$, possibly reflecting the difficulty of maintaining a tie at an extreme network distance without the reinforcement of face-to-face interaction (i.e. short geographic distance).

Mediated communication technologies from phone to social media afford users with the possibility of maintaining contact with geographically distant network neighbors. Hence, friends who live geographically far apart may communicate more frequently via Twitter to compensate for the difficulty of face-to-face interaction. Figure 14 shows the resulting tie strength and tie range (d) correlation after controlling for geographic distance (g) for the eight country networks. For this analysis, I bin the geographic distances into quintiles rather than using the raw kilometer distance to allow for easier comparison of the eight countries that differ in geographical size. The first point to note is that communication frequency tends to decrease with geographic distance after controlling for range. This shows that, unlike phone and other mediated communication technologies, Twitter is used more often for short-distance communication. Second, across the eight countries, the pattern of increasing $Ln(w)$ for $d > 2$ is apparent across most geographic quintiles. Taken together, these observations do not lend support to the idea that geography causes a spurious correlation between the strength of ties with their ranges.

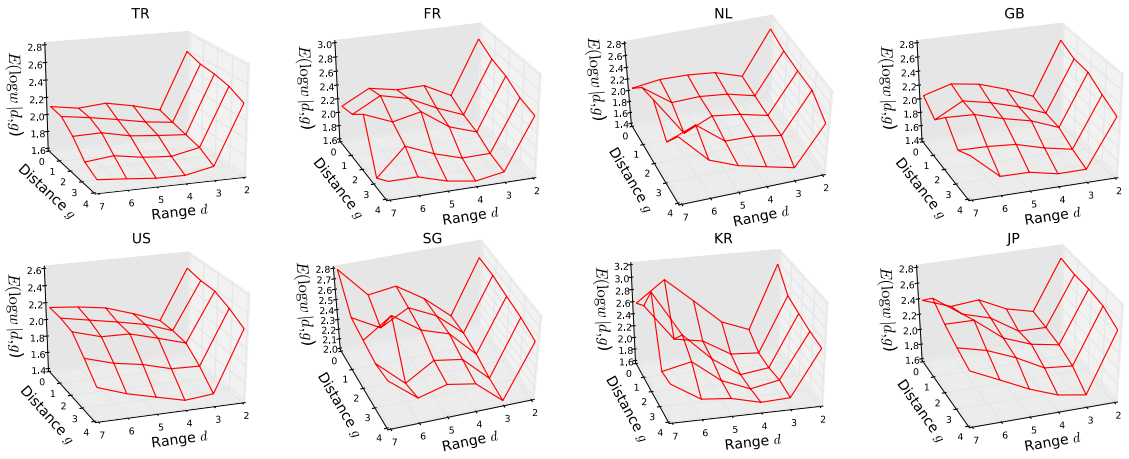


Figure 14. Expected Values of Ln(w)**, Conditional on Geographic Distance Quintiles (g) and Tie Range (d)**, generally increase with bridge range ($d > 2$) and decrease with geographic distance.

Tie Stretching. A plausible fourth hypothesis, which is difficult to directly test with the current observational data, is the fact that the decay probability of a tie over long periods of time is a function of tie strength. Numerous studies have shown that strong social ties last longer and decay at a slower rate than weak ties across a wide range of time scales (Burt 2000; 2002; Dahlander and McFarland 2013; Feld, Suitor, and Hoegh 2007; Larner, 1990; Lubbers et al. 2010; Raeder et al. 2011; Wellman et al. 1997). This suggests the possibility that strong ties, such as those to close friends and kin, are more likely to survive the effects of social and spatial mobility, compared to weak ties, such as those to acquaintances. When people experience social or spatial mobility throughout the life course (e.g. high school graduation), they tend to embed themselves in new social relationships, but not necessarily sever all previously formed ties. The ties that are maintained from previous social contexts become "stretched" when the new and old networks do not overlap. Certainly, when the weaker ties break over time due to social or

geographic mobility, the remaining stretched ties, which become short-range bridges, are also subject to a higher probability of decay compared to embedded ties (Burt 2002). However, this probability decreases with time, such that the stretched ties that do survive the short term tend to show marginally decreasing odds of decay that is even lower than embedded ties (Burt 2002; Lubbers et al. 2010; Martin and Yeung 2006). Viewed from within the group, the dispersion of each group member in different social and geographic directions causes sparsification of ties within the group, which in turn, increases the range of the remaining ties. This selection process on strong ties, propelled by mobility and diverging life courses, might explain the strength of long-range ties. Testing this explanation is beyond the current study since it requires data on the age of ties. Nevertheless, the current data provide some suggestive evidence for this explanation. That is, if long-range ties result from the survival of particularly strong ties and the decay of weaker ties over time, then the tie-decay rate with respect to range (Figure 10A) should be lower than the tie-decay rate with respect to time. This is precisely what I find upon closer inspection of the range distributions across the eight countries (See SI Appendix Figure 18).

Alternative Measures of Tie Strength. While $Ln(w)$ is a reasonable and widely used measure of tie strength, I further consider two related measures - reciprocity and affect in interpersonal communication. First, reciprocity measures the extent to which the frequencies of mentions between i and j are balanced (see Materials and Methods). Reciprocity has been shown to highly correlate with both contact frequency and the persistence of a tie (Hidalgo and Rodriguez-Sickert 2008; Martin and Yeung 2006). Figure 12C shows that the balance in the mention frequencies between i and j tends to increase with d for all countries. This pattern holds up even after controlling for possible individual-level confounds such as nodal degree (see SI

Appendix Figure 20). As a third measure of tie strength, I use the prevalence of positive and negative affect words used in the mention tweets exchanged between i and j in the three countries that predominantly use English on Twitter (Singapore, United Kingdom, and United States). The results reported in Figure 21 and Figure 22 show increasing prevalence of affect words as d increases for the United States and the United Kingdom (See SI Appendix for details).

Discussion

Studies on the efficacy of bridging ties focused heavily on outcomes at the individual level: an individual's probability of adoption, an individual's performance, and so on. Accordingly, researchers primarily measured aggregate bridging capacity from the individual's perspective, from structural holes in an individual's personal network to egonetwork clustering, proving to be sufficient for addressing research questions formulated at the individual level. Nonetheless, network research is quickly developing towards addressing questions that require a broader scope, often at the entire system level, and those can now be addressed with newly available large-scale network data. In line with those developments, I constructed communication networks from Twitter data that cover large portions of Twitter users within national populations to study the frequency and relational strength of bridging ties that span different network distances. The distribution of tie range is characterized by a gamma distribution that decays slower than the exponential or power-law distributions. Although long-range ties are far from prevalent, the extreme network distances they bridge may prove to be crucial for "seeding" and jump-starting simplex social cascades (e.g. emotional contagion, information diffusion) in far-reaches of the global network, thereby increasing the likelihood of social pandemics. I then

showed that the strength of long-range ties, measured by frequency of bidirected mentions, increases with tie range. The analysis of mention reciprocity and emotional affect, two alternative measures of tie strength, also corroborated this main finding. This surprising pattern, in light of Granovetter's long-standing insight that bridging ties are weak, implies that not only low-cost, simplex contagions, but even high-cost, high-threshold contagions (Centola and Macy 2007) that require reassurance from trusted, strong-tie neighbors for adoption (e.g. risky social movement participation) could potentially unfold through long-range bridging ties. Finally, I ruled out three possible explanations (nodal degree, instrumental ties, and geographic distance) for the increasing strength of longer-range bridging ties.

Despite differences in interpersonal communication patterns and tie formation principles discussed in cross-cultural psychology, the distributional characteristics of tie range and the positive correlation between the strength and range of bridging ties did not seem to differ significantly across individualist (United States, United Kingdom) and collectivist (Japan, South Korea, Singapore) countries. While these similarities do not necessarily obviate the general influence of culture on micro-level relational dynamics, they are suggestive of common constraints and rules that may be operating in most societies. One such universal constraint is the fact that human cognitive capacity to form and maintain social ties is finite (Brashears 2013; Miritello et al. 2013; Saramaki et al 2014). Given such cognitive constraint, an individual faces the pressure to reduce communication with certain preexisting ties as the need for forming new relationships arises within shifting social environments. As a result, the vast majority of one's social ties are bound to decay with time. Building on this idea of cognitive constraint, I proposed in this paper an evolutionary explanation that long-range ties result from the survival of strong ties that may have once been embedded in tightly knit clusters of relationships, but "stretched"

into long-range ties as the neighboring ties of the strong ties decay over time.

Although this paper demonstrated a novel and counter-intuitive aspect of bridging ties within population-scale communication networks, several questions immediately arise from the analyses. First, it is unclear whether the surprising strength of long-range ties is a general feature of human communication networks or a platform-specific characteristic of Twitter. Replication studies using different communication platforms (e.g. Facebook, phone, and email) are needed to address this question. Second, the question remains regarding the nature and typical role relations, if any, of long-range bridging ties. Although Granovetter's insights on the strength of weak bridging ties did not depend on the content of those ties, the types of relationships representative of long-range ties may have strong implications for what could or could not spread through them. In this study, the tie-stretching hypothesis suggests that long bridges are old friends. However, other possibilities are equally plausible. For example, the long-range bridges may be dissolved ties that rekindle with bursts of short-term communication upon random encounters or conscious efforts to reconnect (Boase et al. 2015; Levin, Walter, and Murnighan 2011). Another possibility is that these ties are between strangers who happen to engage in disputes and commiseration (i.e. high on negative affect word counts) or bond through similar opinions and views over common subjects of interest (i.e. high on positive affect word counts). A more detailed text analysis of the conversations in the tweets may be useful for starting to uncover the role relations of long-range ties.

Materials and Methods

Data Collection and Preprocessing. I built a distributed Twitter crawler that uses a snowball sampling approach to collect user tweets and metadata. The crawler retrieved 157.9M user

accounts worldwide among which 124.8M accounts had at least one bidirected mention tie with another account in the data (See SI Appendix for details). For these user accounts with bidirected mention ties, I performed a novel geographic location estimation method (Compton et al. 2014) based on a label-propagation algorithm on the global mention network to classify the country of each user. In essence, this algorithm leverages the known locations of the network neighbors of a target user i and assigns "representative" latitude, longitude coordinates as the best guess for i's unknown location. This assigned location is subsequently used in the next iteration to predict the unknown locations of i's neighbors (See SI Appendix for details). For the user accounts to which the classifier failed to assign country labels, I assigned the modal country label of the neighbors. Through this procedure, a total of 119.2M users were assigned 240 different country labels. Then, I grouped these user accounts by their inferred country and constructed within-country mention networks (excluding between-country mention ties) where two users from the same country share a tie if both sent at least one mention tweet or reply tweet to the other. Trivial one-way communication ties that are not based on acquaintance or friendship (e.g. fans mentioning Justin Bieber) are vastly reduced by only taking bilateral mentions and/or replies in the construction of the networks. For two reasons, I also do not consider retweet links as a communication tie for constructing the networks. First, the data do not readily reveal the retweet chain starting from the user who originally tweeted a message to the user who happened to retweet it. Second, it is not clear whether retweets are used for broadcasting information or for engaging in interpersonal communication. Using these criteria for tie construction, I analyze the largest connected components of eight within-country networks (United States, United Kingdom, France, Netherlands, Turkey, Singapore, Japan, and South Korea). The countries were selected on the bases of relatively high Twitter penetration and of variations in national culture and economic

development (See SI Appendix for details of the eight country networks).

Measures. Three behavioral measures of tie strength are considered in this paper. The first and main measure is the aggregate frequency of mentions exchanged between two users. Although classic studies on the measurement of tie strength question the validity of behavioral measures of tie strength such as contact frequency, duration, or volume (e.g. Marsden and Campbell 1984), more recent studies based on fine-grained behavioral data suggest that these behavioral measures are accurate predictors for subjective perceptions of tie strength (Jones et al. 2013). The second measure of tie strength is the reciprocity or balance in the dyadic exchange of mentions. I quantify mention balance as,

$$r_{ij} = 1 - 2\left|\frac{w_{i\to j}}{w_{i\to j} + w_{j\to i}} - 0.5\right|$$

where $r_{ij} = r_{ji}$, $w_{i\to j}$ is the number of tweets i mentioned j. Note that r_{ij} is scaled to vary between 0 and 1 and approaches 1 as the mention frequencies in each direction approach a perfect $1:1$ ratio. The third measure of tie strength counts the occurrence of affect-laden words in the tweets between i and j. To this end, I use the widely adopted Linguistic Inquiry and Word Count (Pennebaker et al. 2001) lexicon to identify the words that express affect (e.g. anxiety, anger, happiness) in the tweets exchanged between two neighbors and measure their prevalence relative to different baselines (See SI Appendix for details).

I use tie range as a measure of the global network distance that a tie spans, which is equivalent to the second shortest path length of two adjacent nodes in a network (Granovetter 1973; Kossinets et al. 2008). Range differs in scope from local distance measures of embeddedness (i.e.

number of common neighbors) and network constraint (Burt 1995) in that it is measured from a global breadth-first-search that transcends the local network structures of either node of a tie and traverses the entire network. Finally, for the measurement of geographic distance between two nodes, I compute the Vincenti distances for all ties whose nodes' geographic locations can be accurately inferred from the GPS coordinates contained in their tweets (see Supporting Information for details). By limiting the measurement of geographic distance to those ties with users whose GPS coordinates are available, I eliminate the otherwise spurious correlation between geographic distance and range caused by the label propagation algorithm.

Supporting Information (Appendix)

Twitter Data Collection. I built a distributed Twitter crawler that ran on Amazon Web Services from November 2013 to October 2014 for data collection. The crawler started from a set of 668K seed users from nine targeted countries (UK, US, France, the Netherlands, Turkey, Indonesia, Japan, South Korea, and Singapore) whose accounts were identified in a previous study (Golder and Macy 2011). The crawler collected the seed users' profile data and timelines that contain up to 3200 most recent tweets per user. Based on the time zone information in the user's profile data and the dominant language used in the collected tweets, the crawler makes a rough guess as to whether a given user is from one of the nine targeted countries. If the crawler guesses that the user is in a targeted country, it parses the tweets in the user's timeline to extract the IDs of the mentioned and retweeted users, cross-checks those IDs against a MongoDB database that records whether a particular user had already been crawled, and queues them for subsequent crawls if the database indicates that that user has not been crawled yet. This data collection cycle repeats itself, resulting in a snowball sample prioritized on the targeted countries, but not excluding other

countries.

Geolocation Inference. I build a user country classifier that implements a novel geographic location estimation algorithm based on label propagation (Compton et al. 2014). This algorithm leverages the fact that the vast majority of one's network neighbors on social media are geographically proximate (Takhteyev et al. 2012; Hawelkaab et al. 2014). By gauging the locations of one's network neighbors, therefore, it is possible to predict a focal user's location with high accuracy. Overall, the algorithm starts from a set of known user locations and uses them as data points for predicting the locations of network neighbors. Here, the network neighbors are defined as those who have exchanged at least one @mention tweet with a given Twitter user in each direction. The predicted locations of the neighbors are then used in the next iteration to predict the unknown locations of the neighbors' neighbors. This process can run until all the nodes in the network have been labeled, but usually cover a large proportion of users within a handful of iterations. Compton et al. (2014) report that this procedure yields 90% coverage (i.e. 100M Twitter users) with a median error distance of 6.8km.

Establishing "Ground Truth"

I set up the location prediction task by first establishing "ground truth" locations for the set of Twitter users who have at least three GPS tagged tweets in their timelines. For each of these users, the algorithm identifies the L1-median location of a user's GPS coordinates (i.e. the single GPS location among the set of GPS locations which minimizes the sum of distances) as a candidate "ground truth" location of the user. To enhance robustness, the algorithm employs a dispersion threshold as a criterion for accepting the candidate location as the final "ground truth."

That is, the algorithm designates the candidate location as the "ground truth" location only if the median of all the distances between the candidate location and each of the rest of the GPS locations in the set does not exceed the dispersion threshold. While Compton et al set this threshold at 15km, I set it at 30km in the current implementation to increase the number of unique ground-truth locations despite the tradeoff in accuracy. This decision is justified by the objective of the current classification task, which is focused on accurately labeling users at the country level.

Label Propagation

Based on the observation from previous studies that the majority of communication ties on Twitter span short geographic distances, the algorithm "propagates" the "ground truth" locations obtained from the previous step to the immediate network neighbors of those whose "ground truth" locations were estimated. For each of those neighbors whose location is unknown, the propagated ground truth locations form the set of possible representative locations. Then, the algorithm computes the L1-median location from this set and accepts it as the best-guess location for a given user if the median distance from the best-guess location to all the other locations in the set does not exceed a predetermined dispersion threshold, θ. Hence, the precision of the estimated location diminishes with larger θ, but the probability that the algorithm will accept the best-guess location increases. The estimated locations obtained through this process are then propagated and used in the next iteration to estimate the network neighbors of the users whose locations were estimated in the current iteration.

Number of Iterations

As the goal of user location estimation in the current application is to identify country locations, which does not require a high level of precision beyond the city level, I increase the θ in successive sets of iterations in the estimation process to maintain high estimation precision while also achieving high coverage. Specifically, I run three sets of iterations with five iterations per set where the θ is increased from 100km, 1000km, to 1500km, for the three successive sets, respectively. Hence, estimation precision is highest in the first five iterations and lowest in the last five iterations.

Results

To assess performance, I held out a random 20% (1.5M user accounts) of the ground truth labels and computed the distance between the ground-truth and the estimated locations (i.e. error distance). At the end of the first five iterations, the median error distance was 14.6km and approximately 80% of the predicted locations were within 100km of the ground truth location (Figure 15). Although the classifier achieved relatively low error overall, it tended to exhibit higher inaccuracy for high degree nodes as shown in Figure 16.

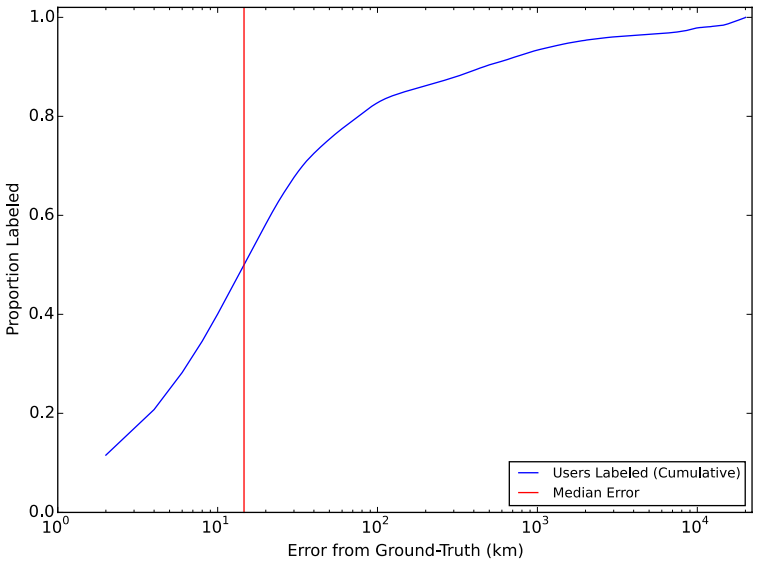


Figure 15 Precision of the Label Propagation Algorithm for the First Five Iterations. A random 20% (1.50M users accounts) of the ground-truth labels were held out for validation. Red vertical line indicates the median error. The cumulative error distance (blue curve) shows that approximately 80% of the predicted locations are within 100km from their corresponding ground truth locations.

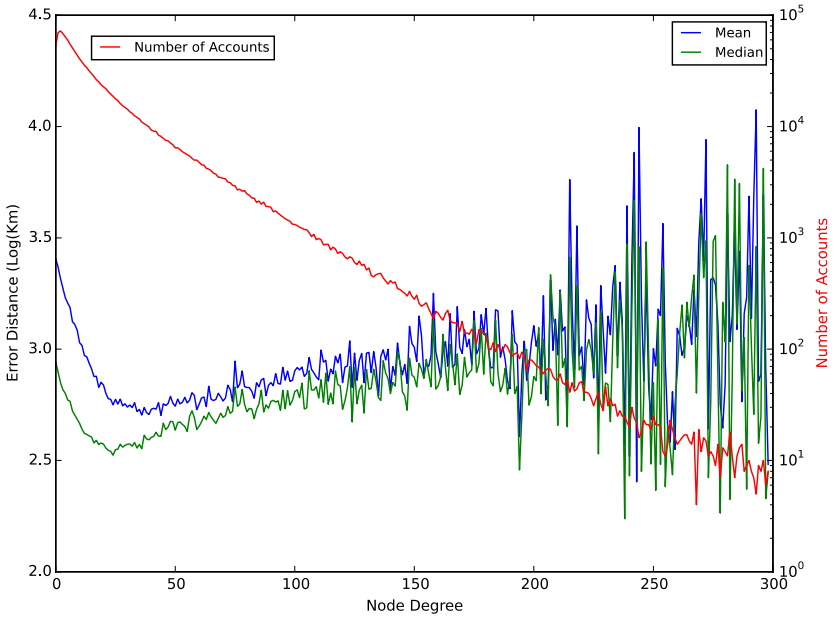


Figure 16 Precision of the Label Propagation Algorithm Conditional on Node Degree at the End of the First Five Iterations. Mean and median errors are U-shaped where the errors tend to decrease below degree = 25 and increase beyond that point.

At the end of the first five iterations with the dispersion threshold set to 100km, 80.1% of the 124.8M user accounts in the global bidirected mention network were assigned representative locations and by the end of the 15th iteration, the coverage marginally increased to 86.9%. This 86.9% of users consisted of 11.9M users identified initially through GPS tweets (i.e. "ground truth") and 96.6M users subsequently identified through the 15 iterations of label propagation. I subsequently reverse-geocoded the latitude-longitude coordinates of these geolocated users to obtain country labels for each user. For this task, I used the open-sourced Twofishes geocoder (twofishes.net) that is built primarily from the GeoNames data (geonames.org). Since the objective was to identify user locations at the country level, I additionally assigned the modal country label of the neighbors of 10.6M user accounts to which the label propagation algorithm failed to estimate. In the end, the entire procedure succeeded in assigning country labels for a total of 119.2M user accounts. The classification results by country are shown in Fig 17 and Table 1. The remaining 5.7M user accounts could not be assigned country labels due either to reverse-geocoding failure or to not having modal countries of the users' neighbors.

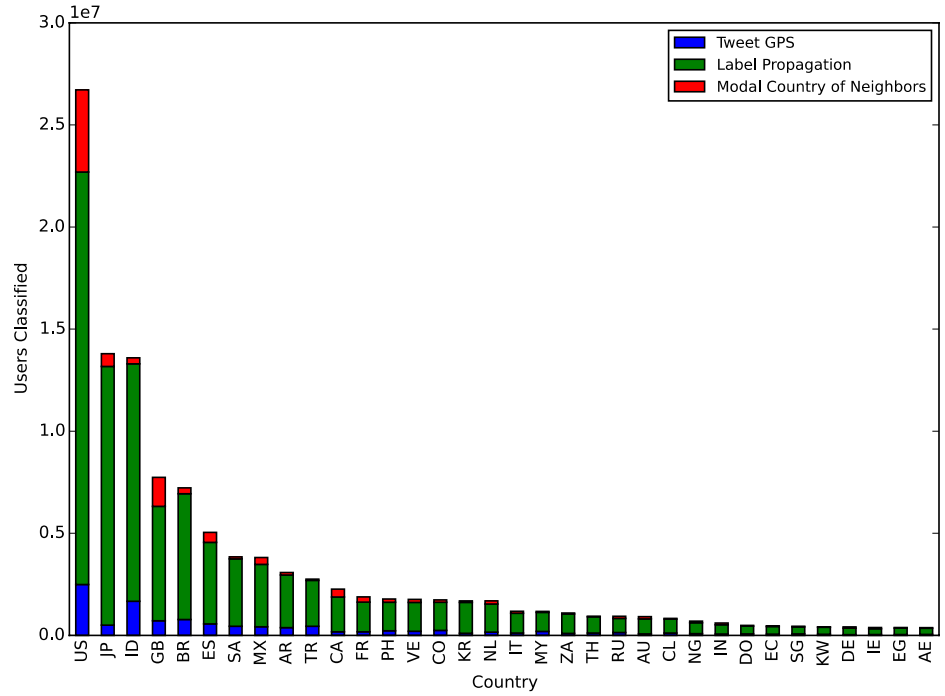


Figure 17. Number of User Accounts Classified Through GPS Tweets (blue), Label Propagation (green), and Modal Country of Neighbors for the top 34 countries in user accounts.

Table 1. Twitter User Country Identification in Eight Countries. The country of a user is identified either through GPS tweet locations, label propagation, or the modal country of a user's network neighbors. For the main analyses, user accounts are grouped by country label for constructing within-country communication networks. The largest connected component (LCC) of each country network constitutes over 94% of labeled users.

Country Label	Total Users Labeled	GPS Tweets	Label Propagation	Modal Country of Neighbors	% of Total Labeled Globally	LCC	LCC as % of Total Labeled	Mean Degree (LCC)
US	26,713,145	2,492,866	20,201,269	4,019,010	22.41%	26,352,018	98.6%	26.1
JP	13,795,718	499,351	12,670,125	626,242	11.57%	13,761,771	99.8%	27.6
GB	7,739,197	713,481	5,604,126	1,421,590	6.49%	7,652,737	98.9%	24.3
TR	2,754,121	443,617	2,249,003	61,501	2.31%	2,666,445	96.8%	11.5
FR	1,885,999	174,499	1,453,491	258,009	1.58%	1,842,853	97.7%	17.0
KR	1,689,004	103,473	1,513,793	71,738	1.42%	1,660,554	98.3%	12.7
NL	1,687,915	158,483	1,379,623	149,809	1.42%	1,668,599	98.9%	20.7
SG	445,226	71,346	344,220	29,660	0.37%	419,750	94.3%	14.4

Summary of User Activity Levels by Country. Previous Twitter research consistently demonstrates heavily skewed user-activity level distributions as measured by the number of tweets, followers, and friends. Accordingly Table 2 reports the medians of the metrics that inform the aggregate activity levels in the eight countries. For the users in the collected data, the median user account in the Netherlands was 901 days old whereas the median account in Japan was 443 days. These differences likely reflect the variation in the initial Twitter adoption periods across countries. The median "statuses count," which adds the number of tweets and retweets by a user since account creation, summarizes the overall Twitter activity level of each country. Here, the medians range between 400 and 900 for all countries except Singapore. Focusing on the tweets and retweets that were retrieved by the crawler, I find that Western countries generally exhibit higher ratios of retweets over user-generated tweets (US: 0.21, GB: 0.20, TR: 0.25, and FR: 0.31). This pattern is consistent with previous studies showing that Twitter is used more as a news distribution channel than an interpersonal communication platform (Garcia-Gavilanes et al. 2013; Poblete et al. 2011).

Table 2. Twitter User Activity Level. Median of activity levels, followers, and friends. Account Age is the difference between an account's creation date and the date of the last record in an account's timeline. Statuses Count aggregates all tweets and retweets of an account since account creation. Follower and Friend counts were obtained from user profiles.

	Account Age at Last Record (days)	Statuses Count	Tweets in Data	Retweets in Data	Followers	Friends
US	826	724	439	93	137	188
JP	443	793	527	32	92	109
GB	868	412	271	54	116	199
TR	643	598	331	82	145	157
FR	571	495	281	88	81	124
KR	446	229	167	5	27	44
NL	901	505	316	38	86	112
SG	854	2068	1051	150	129	162

Fitting Tie Range Distributions. The range distributions of the eight country networks are best described by Gamma functions with a shape parameter, k, larger than one (Figure 18). A Gamma distribution with k=1 reduces to the exponential distribution, while for k>1, the distribution decays at a slower rate than exponential. Hence, the tie-decay rate with respect to range is slower than the tie-decay rate with respect to time, which has been reported to follow a power-law distribution (Burt 2000). The slower decay rate of the range distribution is expected if the long-range ties result from the strong ties surviving over time.

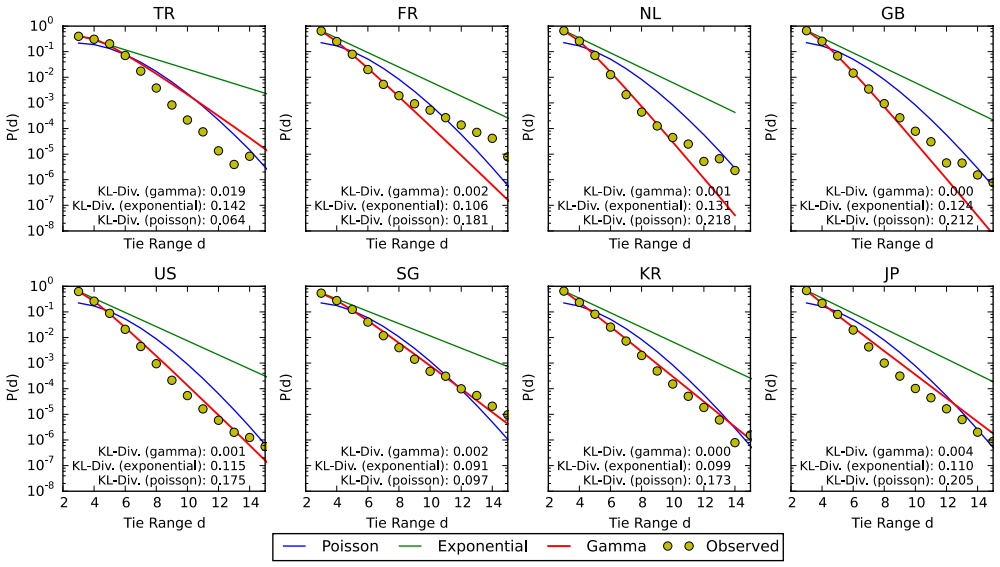


Figure 18. Best Fitting Poisson, Exponential, and Gamma PDFs of Eight Within-Country Twitter Networks. The empirically observed probabilities of tie range for d > 2 are displayed in blue, green, and red, respectively. Using Kullback-Leibler divergence to assess distributional fit, the range distributions across the eight countries are best approximated by gamma distributions with shape parameter, k (mean=1.51, std=0.34) and scale parameter theta (mean=0.78, std=0.15). With k > 1, range shows slower than power-law decay with respect to time.

To check if the Gamma distribution of range generalizes to other networks, I examined other social and communication networks that vary in size, data collection method, and type of

tie (Figure 19). These networks include intra-organizational employee networks on four types of ties (friendship, advice giving, and voicing problems and ideas) across nine U.S. credit unions (Detert et al. 2013), the LiveJournal friendship network (Yang and Leskovec 2012), an email exchange network among members of an EU research institute (Leskovec, Kleinberg, and Faloutsos 2007), and communications among Wikipedia authors and contributors (Leskovec, Huttenlocher, and Kleinberg 2010). Ties in these networks all exhibited skewed unimodal distributions with a maximum tie range of 7. For all networks, Gamma distributions ($k > 1$) seemed to fit best (lowest Kullback-Leibler divergence).

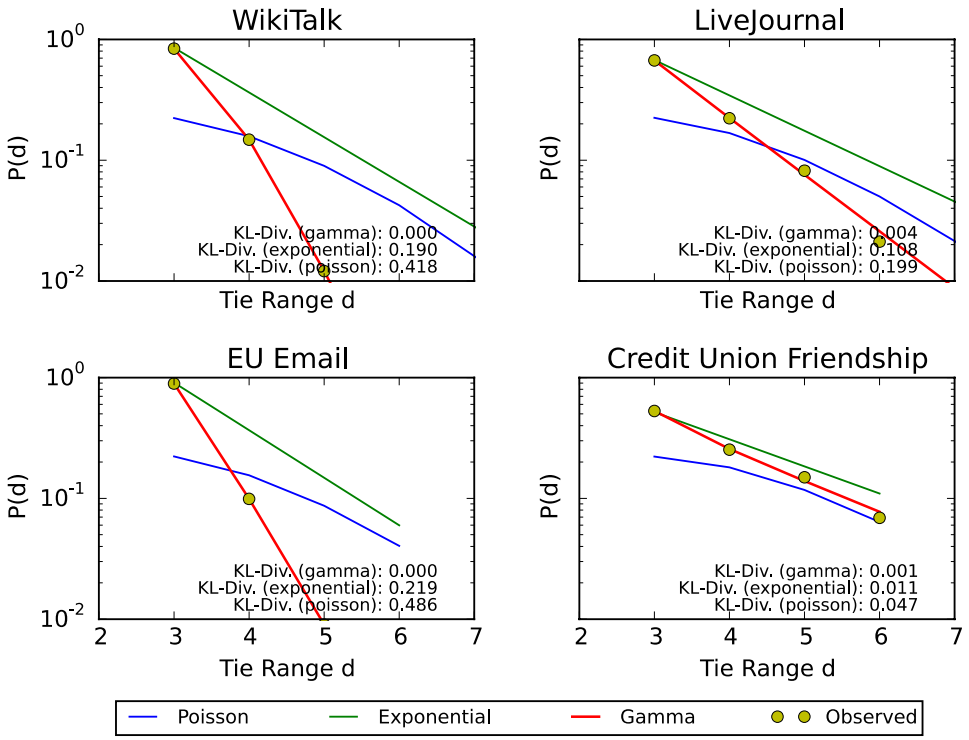


Figure 19. Best Fitting Poisson, Exponential, and Gamma PDFs on Range Distribution for Four Social and Communication Networks.

Within-Individual Standardized Mention Balance

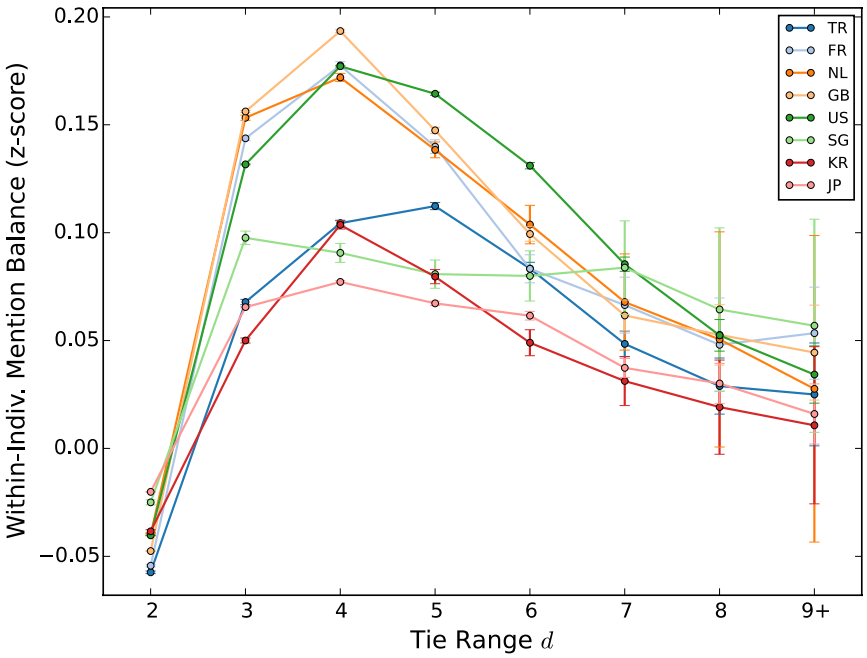


Figure 20. Within-Individual Standardized Mention Balance. Mention balance exhibits inverted U-shaped pattern with respect to range. The decrease in balance for longer range ties is largely driven by low mean degree, which limits the variation in z-scores.

Affect Word Count and Tie Range. I use the LIWC lexicon to compute the proportion of affect words in the @mention tweets exchanged between each pair of users in three Anglophone countries (US, UK, and Singapore). Excluding retweets, I compute the proportion of affect words with regard to three different baselines: within-range, within-dyad, and within-individual. For the within-range baseline, the proportion of affect words in range d is

$$A^d = \frac{1}{T^d} \sum_{k}^{N_{edge}^d} a_k^d$$

where a_k^d is the number affect words exchanged in mention tie, k of range d, T^d is the total number of words exchanged in all mention ties of range d, and N_{edge}^d is the number of mention ties of range d. This quantity allows us to examine the overall prevalence of affect words at varying range values. For the within-dyad baseline, I compute the mean within-edge affect as,

$$A_{edge}^d = \frac{1}{N_{T_k^d>200}^d} \sum_k^{N_{edge}^d} \frac{a_k^d}{T_k^d} \ , \ T_k^d > 200$$

where T_k^d is the total number of words exchanged in mention tie, k and $N_{T_k^d>200}^d$ is the number of edges with range d and $T_k^d > 200$. Hence, A_{edge}^d counts the affect words used between two users in the mention tweets directed to each other, relative to all the words that they jointly used in the mention tweets directed to the union of their neighbors. To ensure reliable results, I only include ties where $T_k^d > 200$. Finally, the within-individual baseline is computed as,

$$A_{indiv}^{d} = \frac{1}{N_{arc}^{d}} \sum_{i}^{N_{arc}^{d}} \frac{a_{ij}^{d}}{T_{ij}^{d}} / \frac{a_{i*}}{T_{i*}} \ , \ T_{ij}^{d} > 200$$

where a_{ij}^{d} and T_{ij}^{d} are the affect words and total words in the tweets of user i that mention user j, respectively, a_{i*} and T_{i*} are the affect words and total words in all of user i's @mention tweets, respectively, and N_{arc}^{d} is the number of arcs (i.e. directed ij pairs) of range d where $T_{ij}^{d} > 200$. In short, A_{indiv}^{d} is the mean ratio of the alter-specific proportion of affect words to the overall proportion of affect words of a user across all directed ij pairs.

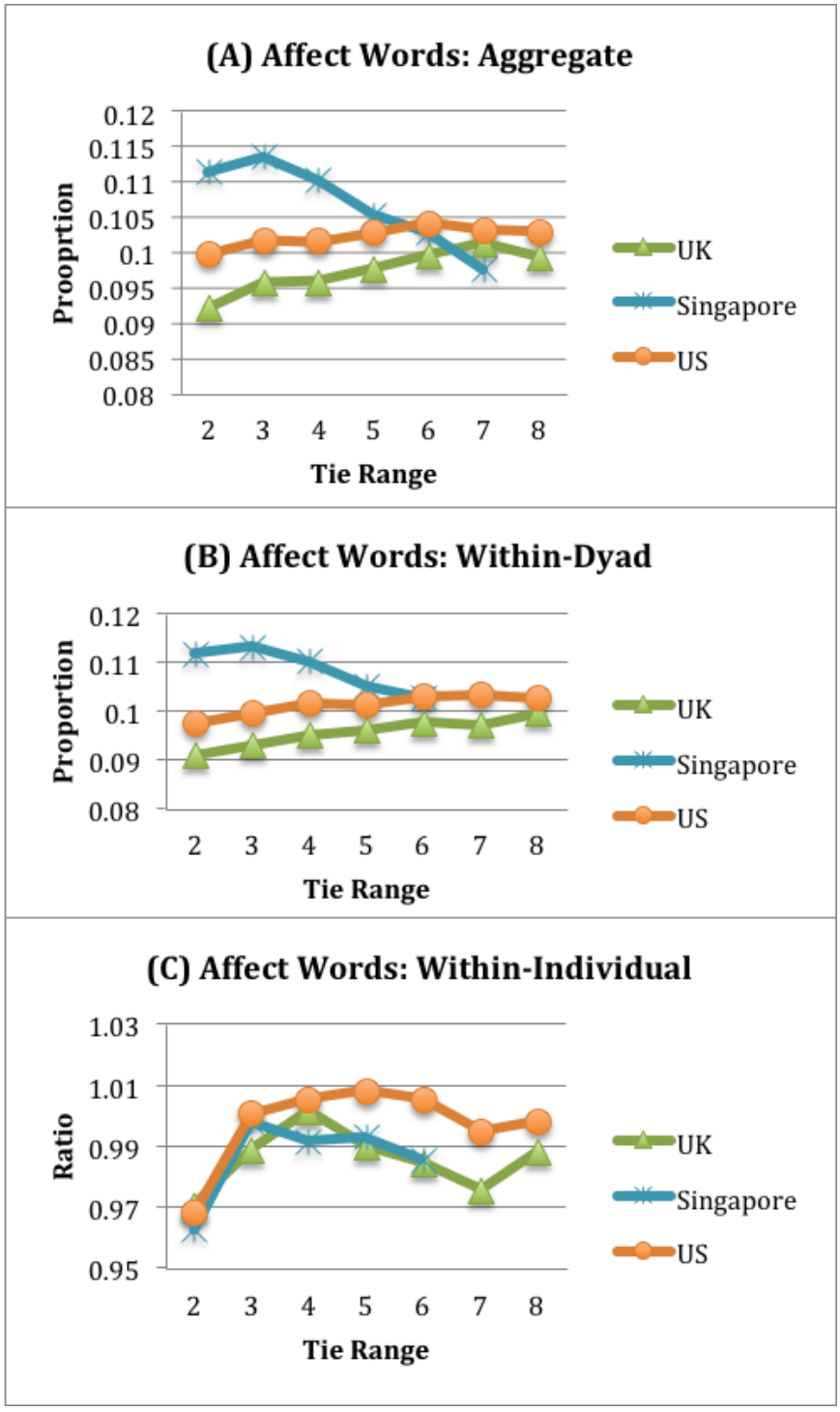


Figure 21. LIWC affect words tend to increase moderately with range across different baselines.

I find that the average proportion of affect words, relative to all words used within specific range values (Figure 21A) and relative to all words used within specific dyads (Figure 21B), increases moderately with tie range in the US and the UK. In contrast, Singapore shows an opposite pattern where affect is less expressed in long-range relationships. Results from other Anglophone countries that are not considered or presented in this study (Australia, New Zealand, and Canada) exhibit similar increasing patterns as the US and the UK. This curious difference between Singapore and the rest of the Anglophone countries is suggestive of either cultural differences or differences in how Twitter is used as a communication medium and calls for further qualitative investigation. Regarding the within-individual normed prevalence of affect words, all three countries exhibit a similar inverted U-shaped curve with increasing tie range.

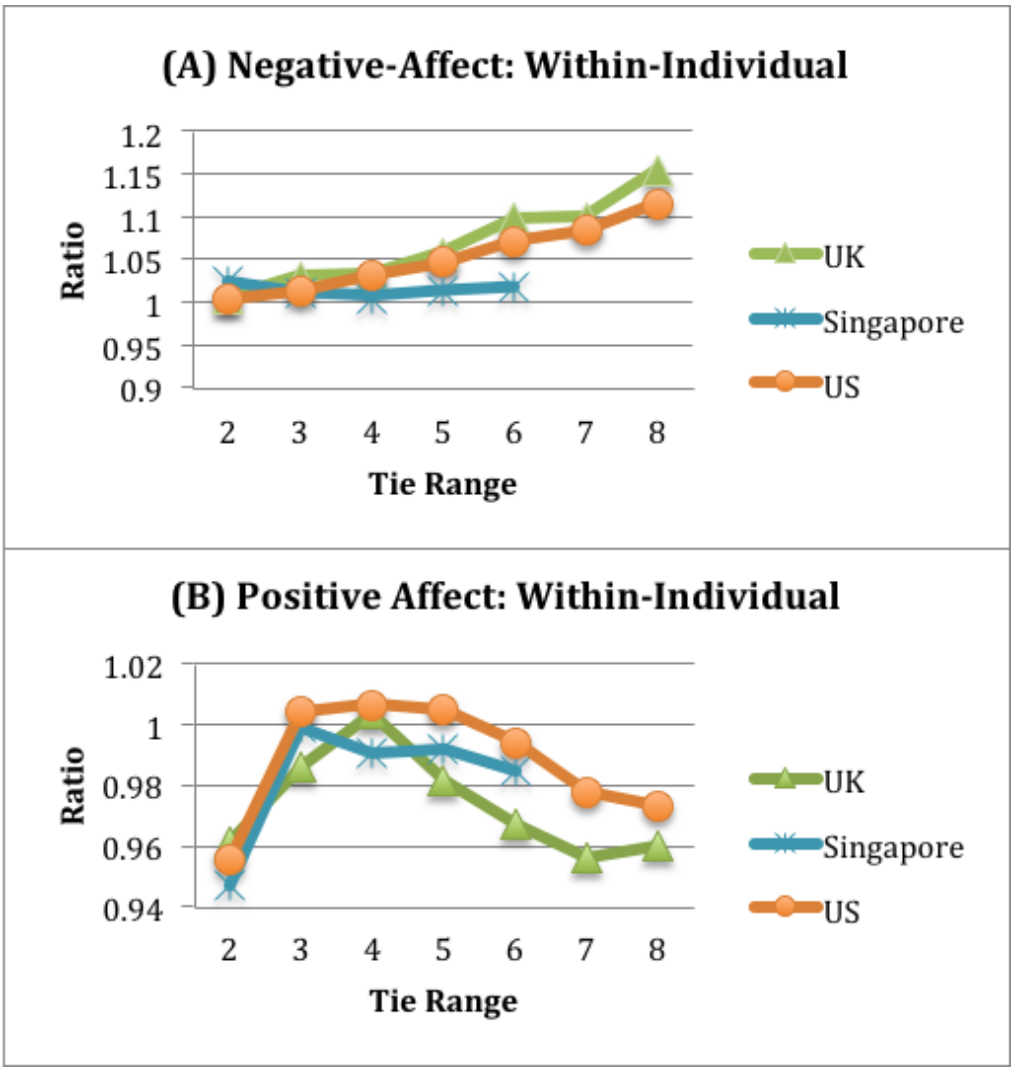


Figure 22. LIWC Affect Ratios for Singapore, UK, and US are Disaggregated into Positive (A) and Negative (B) Affect Ratios.

I further disaggregate the overall within-individual affect ratios in Figure 21C into positive affect and negative affect ratios to explore the driver of the the overall inverted U-shape in aggregate. Figure 22 suggests that the inverted U-shaped pattern results from the aggregate of a positive correlation of tie range and negative affect (Figure 22A) and a concave relationship between tie range and positive affect (Figure 22B). It is unclear what the increasing negative affect is signaling; it could be indicating adversary relationships (e.g. trolls), but it could also be interpreted as strong positive relationships (e.g. close friends) where negative sentiments (e.g.

commiseration) can be more freely expressed. I leave the interpretations of these findings for future research.

References

Allcott, Hunt, Dean Karlan, Markus M Möbius, Tanya S Rosenblat, and Adam Szeidl. 2007. "Community Size and Network Closure." *American Economic Review* 97:80-85.

Allik, Jüri and Anu Realo. 2004. "Individualism-Collectivism and Social Capital." *Journal of Cross-Cultural Psychology* 35:29-49.

Aral, Sinan and Marshall Van Alstyne. 2011. "The Diversity-Bandwidth Trade-off." *American Journal of Sociology* 117(1): 90-171.

Bakshy, Eytan, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. "Role of Social Networks in Information Diffusion." *WWW 2012*.

Baldassarri, Delia and Mario Diani. 2007. "The Integrative Power of Civic Networks." *American Journal of Sociology* 113:735-780.

Barabási, Albert-László. 2009. "Scale-Free Networks: A Decade and Beyond." *Science* 325:412-413.

Barabási, Albert-László and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286:509-512.

Boase, Jeffrey, Tetsuro Kobayashi, Andrew Schrock, Tsutomu Suzuki, and Takahisa Suzuki. 2015. Reconnecting Here and There: The Reactiviation of Dormant Ties in the United States and Japan. *American Behavioral Scientist* 59(8):931-945.

Bond, Michael Harris. 1988. "Finding Universal Dimensions of Individual Variation in Multicultural Studies of Values: The Rokeach and Chinese Value Surveys." *Journal of Personality and Social Psychology* 55:1009-1015.

Bott, Elizabeth. 1957. *Family and Social Networks: Roles, Norms, and External Relationships in Ordinary Urban Families*. London: Tavistock.

Brashears, Matthew E. 2013. "Humans Use Compression Heuristics to Improve the Recall of Social Networks." *Nature Scientific Reports* 3: 1513.

Burger, Martijn J., and Vincent Buskens. 2009. "Social Context and Network Formation: An Experimental Study." *Social Networks* 31:63-75.

Burt, Ronald. 1995. *Structural Holes: The Social Structure of Competition*. Harvard University Press.

Burt, Ronald. 2000. "Decay Functions." *Social Networks* 22: 1-28.

Burt, Ronald. 2002. "Bridge Decay." *Social Networks* 24: 333-363.

Burt, Ronald S., Robin M. Hogarth, and Claude Michaud. 2000. "The Social Capital of French and American Managers." *Organization Science* 11:123-147.

Burt, Ronald, Joseph Jannotta, and James Mahoney. 1998. " Personality correlates of structural holes." *Social Networks* 20(1):63-87.

Byrne, Donn. 1971. *The Attraction Paradigm*. 1st ed. New York: Academic Press.

Campbell, Karen, Peter Marsden, and Jeanne Hurlbert. 1986. "Social Resources and Socioeconomic Status." *Social Networks* 8(1): 97-117.

Cartwright, Dorwin and Frank Harary. 1956. "Structural Balance: a Generalization of Heider's theory." *Psychological Review* 63:277-293.

Centola, Damon and Michael Macy. 2007. "Complex Contagions and the Weakness of Long Ties." *American Journal of Sociology* 113(3): 702-734.

Chai, Sun-Ki and Mooweon Rhee. 2010. "Confucian Capitalism and the Paradox of Closure and Structural Holes in East Asian Firms." *Management and Organization Review* 6:5-29.

Christakis, Nicholas and James Fowler. 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York: Little, Brown, and Company.

Coleman, James S. 1988. "Social Capital in the Creation of Human Capital." *The American Journal of Sociology* 94: 95-120.

Compton, Ryan, David Jurgens, and David Allen. 2014. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. *IEEE International Conference on Big Data* 393-401.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2010. "Identifying the Roles of Race-based Choice and Chance in High School Friendship Network Formation." *Proceedings of the National Academy of Sciences* 107:4857-4861.

Dahlander, Linus and Daniel McFarland. 2013. "Ties That Last: Tie Formation and Persistence in Research Collaborations over Time." *Administrative Science Quarterly* 58(1): 69-110.

Davis, Gerald F. and Henrich R. Greve. 1997. "Corporate Elite Networks and Governance Changes in the 1980s." *American Journal of Sociology* 103:1-37.

Detert, James, Ethan Burris, David Harrison, and Sean Martin. 2013. Voice Flows to and around Leaders: Understanding When Units are Helped or Hurt by Employee Voice. *Administrative Science Quarterly*, 58(4): 624-668.

Emirbayer, Mustafa and Jeff Goodwin. 1994. "Network Analysis, Culture, and the Problem of Agency." *The American Journal of Sociology* 99:1411-1454.

Entwisle, Barbara, Katherine Faust, Ronald R. Rindfuss, and Toshiko Kaneda. 2007. "Networks and Contexts: Variation in the Structure of Social Ties." *The American Journal of Sociology* 112:1495-1533.

Feld, Scott, Jill Suitor, and Jordana Gartner Hoegh. 2007. "Describing Changes in Personal Networks over Time." *Field Methods* 19:218-236.

Fischer, Claude S. and Yossi Shavit. 1995. "National Differences in Network Density: Israel and the United States." *Social Networks* 17(2):129-145.

Flynn, Francis J., Ray E. Reagans, and Lucia Guillory. 2010. "Do You Two Know Each Other? Transitivity, Homophily, and the Need for (network) Closure." *Journal of Personality and Social Psychology* 99:855-869.

Goodreau, Steven, James Kitts, and Martina Morris. 2009. "Birds of a Feather, Or Friend of a Friend?: Using Exponential Random Graph Models to Investigate Adolescent Social Networks." *Demography* 46:103-125.

Golder, Scott and Michael Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333(6051): 1878-1881.

Granovetter, Mark S. 1973. "The Strength of Weak Ties." *The American Journal of Sociology* 78:1360-1380.

Grossetti, Michel. 2007. "Are French Networks Different?" *Social Networks* 29(3):391-404.

Gudykunst, William B., Seung-Mock Yang, and Tsukasa Nishida. 1987. "Cultural Differences in Self-Consciousness and Self-Monitoring." *Communication Research* 14:7-34.

Gudykunst, William B., Ge Gao, Karen L. Schmidt, Tsukasa Nishida, Michael H. Bond, Kwok Leung, Georgette Wang, and Robert A. Barraclough. 1992. "The Influence of Individualism Collectivism, Self-Monitoring, and Predicted-Outcome Value on Communication in Ingroup and Outgroup Relationships." *Journal of Cross-Cultural Psychology* 23:196-213.

Hansen, Morten. 1999. "The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits." *Administrative Science Quarterly* 44(1): 82–111.

Hawelkaab, Bartosz, Izabela Sitkoab, Euro Beinata, Stanislav Sobolevskyb, Pavlos Kazakopoulosa, and Carlo Rattib. 2014. “Geo-located Twitter as Proxy for Global Mobility Patterns.” *Cartography and Geographic Information Scienc*e 41(3): 260-271.

Hechter, Michael. 1987. *Principles of Group Solidarity.* Berkeley: University of California Press.

Hidalgo, Cesar and C. Rodriguez-Sickert. 2008. "The Dynamics of a Mobile Phone Network." *Physica A* 387:3017-3024.

Hofstede, Geert H. 1991. *Cultures and Organizations : Software of the Mind*. London; New York: McGraw-Hill.

Holland, Paul W., and Samuel Leinhardt. 1973. “The Structural Implications of Measurement Error in Sociometry.” *The Journal of Mathematical Sociology* 3:85-111.

Hoshino-Browne, Etsuko, Adam S. Zanna, Steven J. Spencer, Mark P. Zanna, Shinobu Kitayama, and Sandra Lackenbauer. 2005. “On the Cultural Guises of Cognitive Dissonance: the Case of Easterners and Westerners.” *Journal of Personality and Social Psychology* 89:294-310.

Igarashi, Tasuku, Yoshihisa Kashima, Emiko Kashima, Tomas Farsides, Uichol Kim, Fritz Strack, Lioba Werth and Masaki Yuki. 2008. “Culture, Trust, and Social Networks.” *Asian Journal of Social Psychology* 11:88-101.

Jones, Jason, Jaime Settle, Robert Bond, Christopher Fariss, Cameron Marlow, and James Fowler. 2013. Inferring Tie Strength from Online Directed Behavior. *PLoS One*, 8(1): e52168.

Kalish, Yuval and Garry Robins. 2006. “Psychological Predispositions and Network Structure: The Relationship between Individual Predispositions, Structural Holes and Network Closure.” *Social Networks* 28:56-84.

Kim, Uichol, Harry Triandis, Cigdem Kagitcibasi, Sang-Chin Choi, and Gene Yoon. 1994. *Individualism and Collectivism: Theory, Method, and Applications*. Sage Publications, Inc.

Kossinets, Gueorgi, Jon Kleinberg, and Duncan Watts. 2008. "The structure of information pathways in a social communication network." *ACM SIGKDD* 435-443.

Kuwabara, Ko, Robb Willer, Michael W. Macy, Rie Mashima, Shigeru Terai, and Toshio Yamagishi. 2007. “Culture, Identity, and Structure in Social Exchange: A Web-Based Trust Experiment in the United States and Japan.” *Social Psychology Quarterl*y 70:461-479.

Lambiotte, Renaud, Vincent Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. 2008. " Geographical dispersal of mobile communication networks." *Physica A* 387(21): 5317-5325.

Larner, Mary. 1990. "Changes in Network Resources and Relationships over Time." Pp. 181-204

in *Extending Familes: The Social Networks of Parents and Their Children*, edited by Moncrieff Cochran, Mary Larner, David Riley, Lars Gunnarsson, and Charles Henderson. New York, NY: Cambridge University Press.

Leskovec, Jure, Dan Huttenlocher, and Jon Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks. *International World Wide Web Conference* (WWW).

Leskovec, Jure, Jon Kleinberg and Christos Faloutsos. 2007. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data* (ACM TKDD), 1(1).

Levin, Daniel, Jogre Walter, and J. Keith Murnighan. 2011. Dormant Ties: The Value of Reconnecting. *Organization Science,* 22(4): 923-939.

Licoppe, Christian and Zbigniew Smoreda. 2005. "Are Social Networks Technologically Embedded?: How Networks are Changing Today with Changes in Communication Technology." *Social Networks* 27:317-335.

Lin, Nan and Mary Dumin. 1986. "Access to Occupations through Social Ties." *Social Networks* 8(4):365-385.

Lizardo, Omar. 2006. "How Cultural Tastes Shape Personal Networks." *American Sociological Review* 71:778 -807.

Lubbers, Miranda, Jose Luis Molina, Jurgen Lerner, Ulrik Brandes, Javier Avila, and Christopher McCarty. 2010. "Longitudinal Analysis of Personal Networks: The Case of Argentinean Migrants in Spain." *Social Networks* 32: 91-104.

Macy, Michael W. and Yoshimichi Sato. 2002. "Trust, cooperation, and market formation in the U.S. and Japan." *Proceedings of the National Academy of Sciences of the United States of America* 99:7214-7220.

Markus, Hazel R. and Shinobu Kitayama. 1991. "Culture and the Self: Implications for Cognition, Emotion, and Motivation." *Psychological Review* 98:224-253.

Marsden, Peter V. 1990. "Network Data and Measurement." *Annual Review of Sociology* 16:435-463.

Marsden, Peter and Karen Campbell. 1984. "Measuring Tie Strength." *Social Forces* 63(2): 482-501.

Martin, John Levi and King-To Yeung. 2006. "Persistence of Close Personal Ties over a 12-Year Period." *Social Networks* 28: 331-362.

McAdam, Doug. 1986. "Recruitment to High-Risk Activism: The Case of Freedom Summer."

American Journal of Sociology 92(1):54-90.

McPherson, Miller, Lynn Smith-Lovin, and James Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415-444.

Mehra, Ajay, Martin Kilduff, and Daniel Brass. 2001. "The Social Networks of High and Low Self-Monitors: Implications for Workplace Performance." *Administrative Science Quarterly* 46:121-146.

Miritello, Giovanna, Esteban Moro, Ruben Lara, Rocio Martinez-Lopez, John Belchamber, Sam Roberts, and Robin Dunbar. 2013. Time as a Limited Resource: Communication Strategy in Mobile Phone Network. *Social Networks*, 35(1): 89-95.

Montgomery, James. 1994. "Weak Ties, Employment and Inequality: An Equilibrium Analysis." *American Journal of Sociology* 99(5):1212– 1236.

Moody, James. 2001. "Race, School Integration, and Friendship Segregation in America." *The American Journal of Sociology* 107:679-716.

Nisbett, Richard. 2003. *The Geography of Thought: How Asians and Westerners Think Differently-- and why*. New York: Free Press.

Oh, Hongseok and Martin Kilduff. 2008. "The Ripple Effect of Personality on Social Structure: Self-Monitoring Origins of Network Brokerage." *Journal of Applied Psychology* 93:1155-1164.

Oliver, Pamela and Daniel Myers. 2003. "Networks, Diffusion, and Cycles of Collective Action." Pp. 173-203 in *Social Movements and Networks: Relational Approaches to Collective Action* edited by Mario Diani and Doug McAdam. New York, NY: Oxford University Press.

Onnela, Jukka-Pekka, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. -L. Barabasi. 2007. "Structure and tie strengths in mobile communication networks." *Proceedings of the National Academy of Sciences* 104(18): 7332-7336.

Oyserman, Daphna, Heather M. Coon, and Markus Kemmelmeier. 2002. "Rethinking Individualism and Collectivism: Evaluation of Theoretical Assumptions and Meta-analyses." *Psychological Bulletin* 128:3-72.

Pachucki, Mark A. and Ronald L. Breiger. 2010. "Cultural Holes: Beyond Relationality in Social Networks and Culture." *Annual Review of Sociology* 36:205–224.

Pennebaker, James W., M. E. Francis, and R. J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Erlbaum, Mahwah, NJ.

Poblete, Barbara, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. "Do All Birds Tweet the Same?: Characterizing Twitter around the World." *ACM CIKM* 1025-1030.

Podolny, Joel M. and James N. Baron. 1997. “Resources and Relationships: Social Networks and Mobility in the Workplace.” *American Sociological Review* 62:673-693.

Raeder, Troy, Omar Lizardo, David Hachen, and Nitesh Chawla. 2011. "Predictors of Short-Term Decay of Cell Phone Contacts in a Large Scale Communication Network." *Social Networks* 33: 235-257.

Reagan, Ray and Bill McEvily. 2003. " Network Structure and Knowledge Transfer: The Effects of Cohesion and Range." *Administrative Science Quarterly* 48(2): 240-267.

Saramaki, Jari, E. A. Leicht, Eduardo Lopez, Sam Roberts, Felix Reed-Tsochas, and Robin Dunbar. 2014. "Persistence of Social Signatures in Human Communication." *Proceedings of the National Academy of Sciences* 111(3):942-947.

Sasovova, Zuzana, Ajay Mehra, Stephen Borgatti, and Michaéla Schippers. 2010, "Network Churn: The Effects of Self-Monitoring Personality on Brokerage Dynamics." *Administrative Science Quarterly* 55:639–670.

Schwartz, Shalom H. 1994. “Beyond Individualism/Collectivism: New Cultural Dimensions of Values.” Pp. 85-119 in *Individualism and Collectivism: Theory, Method, and Applications*, edited by Uichol Kim, Harry C. Triandis, Cigdem Kagitcibasi, Sang-Chin Choi, and Gene Yoon. Sage Publications.

Singh, Robert. 2000. *Entrepreneurial Opportunity Recognition Through Social Networks*. London: Garland.

South, Scott and Dana Haynie. 2004. “Friendship Networks of Mobile Adolescents.” *Social Forces* 83:315-350.

Spencer-Rodgers, Julie, Melissa Williams, David Hamilton, Kaiping Peng, and Lei Wang. 2007. “Culture and Group Perception: Dispositional and Stereotypic Inferences about Novel and National groups.” *Journal of Personality and Social Psychology* 93:525-543.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman. 2012. "Geography of Twitter Networks." *Social Networks* 34(1): 73-81.

Tiwana, Amrit. 2008. "Do Bridging Ties Complement Strong Ties?: An Empirical Exami nation of Alliance Ambidexterity." *Strategic Management Journal* 29(3): 251-272.

Toole, Jameson, Carlos Herrera-Yaqüe, Christian M. Schneider, and Marta C. González. 2015. "Coupling Human Mobility and Social Ties." Journal of the Royal Society Interface 12: 20141128.

Triandis, Harry. 1989. “The Self and Social Behavior in Differing Cultural Contexts.” *Psychological Review* 96:506-520.

Triandis, Harry C., Christopher McCusker, and C. Harry Hui. 1990. “Multimethod Probes of Individualism and Collectivism.” *Journal of Personality and Social Psychology* 59:1006-1020.

Vaisey, Stephen. 2010. “Can Cultural Worldviews Influence Network Composition?” *Social Forces* 88:1595-1618.

Vandello, Joseph A. and Dov Cohen. 1999. “Patterns of Individualism and Collectivism across the United States.” *Journal of Personality and Social Psychology* 77:279-292.

Verma, Jyoti. 1985. “The Ingroup and Its Relevance to Individual Behavior: A Study of Collectivism and Individualism.” *Psychologica* 28: 173-181.

Volkovich, Yana, Salvatore Scellato, David Laniado, Cecilia Mascolo, and Andreas Kaltenbrunner. 2012. “The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions.” *ICWSM* 346-353.

Watts, Duncan. 1999. "Networks, Dynamics, and the Small-World Phenomenon." *America n Journal of Sociology* 105(2): 493-527.

Watts, Duncan and Steven Strogatz. 1998. “Collective dynamics of ‘small-world’ networks.” *Nature* 393:440-442.

Wellman, Barry, Rnita Yuk-lin Wong, David Tindall, and Nancy Nazer. 1997. "A Decad e of Network Change: Turnover, Persistence, and Stability in Personal Communitie s." *Social Networks* 19: 27-50.

Wheeler, Ladd, Harry T. Reis, and Michael H. Bond. 1989. “Collectivism-Individualism in Everyday Social Life: The Middle Kingdom and the Melting Pot.” *Journal of Personality and Social Psychology* 57:79-86.

Xiao, Zhixing and Anne S. Tsui. 2007. “When Brokers May Not Work: The Cultural Contingency of Social Capital in Chinese High-tech Firms.” *Administrative Science Quarterly* 52:1-31.

Yakubovich, Valery. 2005. "Weak ties, Information, and Influence: How Workers Find Jobs in Local Russian Labor Market." *American Sociological Review* 70(3):408-421.

Yamagishi, Toshio, Karen Cook, and Motoki Watabe. 1998. “Uncertainty, Trust, and Commitment Formation in the United States and Japan.” *The American Journal of Sociology* 104:165-194.

Yamagishi, Toshio and Midori Yamagishi. 1994. “Trust and Commitment in the United States and Japan.” *Motivation and emotion* 18:129–166.

Yang, Jaewon and Jure Leskovec. 2012. Defining and Evaluating Network Communities Based on Ground-Truth. *IEEE International Conference on Data Mining* (ICDM).

Yeung, King-To. 2005. "What Does Love Mean? Exploring Network Culture in Two Network Settings." *Social Forces* 84:391-420.

Zachary, Wayne W. 1977. "An Information Flow Model for Conflict and Fission in Small Groups." *Journal of Anthropological Research* 33:452-473.

NETWORK-BASED GROUP ACCOUNT CLASSIFICATION

Abstract

I propose a classification method for group vs. individual accounts on Twitter, based solely on communication network characteristics. While such a language-agnostic, network-based approach has been used in the past, this paper motivates the task from firmly established theories of human interactional constraints from cognitive science to sociology. Time, cognitive, and social role constraints limit the extent to which individuals can maintain social ties. These constraints are expressed in observable network metrics at the node (i.e. account) level, which I identify and exploit for inferring group accounts.

Introduction

User accounts in social media platforms (e.g. Twitter and Facebook) are constituted by individuals, celebrities, and groups from all over the globe, each of which appropriates social media in different ways with messy behavioral traces. This heterogeneity in the user population poses a considerable challenge to the classification of non-individual users as it is difficult to invoke "domain specific" ethnographic knowledge for all users across all countries. Accurately classifying individual vs. non-individual entities on a given social media platform is important for a range of applications. In social network research, for example, the ability to accurately filter out non-individual entities directly affects the network properties the researcher observes, on which her conclusions about the network rest. In online targeted marketing where the targets are, for example, small, local businesses, these small scale business entities often do not register their accounts on central online databases such as twellow.com. Campaigns targeting such small businesses would benefit from an effective classification tool.

A number of methods for identifying non-individual group accounts exist in the literature, from early attempts to identify organizations, celebrities, and media outlets using Twitter's "list" function (Wu et al. 2011; Sharma et al. 2012) to those leveraging the temporal signatures in the tweets of a given user account (Tavares and Faisal 2013) The state-of-the-art language-based classification approaches perform with high accuracy and variants of them which combine textual information with basic network measures (e.g. follower in-degree) have proven to be promising (De Choudhury, Diakopoulos, and Naaman 2012; De Silva and Riloff 2014). Nevertheless, the successful application of these language-based approaches depend on the NLP infrastructure developed for each language where considerable variation exists in the sophistication and reliability of tool-kits across languages. Furthermore, given that languages vary in the amount of information that can be packed into 140 characters per tweet (e.g. Chinese vs. English), the performance of a language-based classifier is prone to vary even more.

As a complementary approach to the language-based approach, I develop a purely network-based classifier to identify groups vs. individuals. The underlying cognitive and sociological theories (Saramaki et al. 2014; Brewer 1995; DeScioli and Kurzban 2009; Dunbar 1995; Goode 1960; Miller 1956; Roberts et al. 2009) for the proposed method postulate that due to time, cognitive, and resource constraints, (a) the number of interactions an individual can maintain at any given time period is typically limited to a few hundred alters and (b) an individual typically exhibits a high concentration of communication with only a handful of individuals while communicating intermittently with others. These constraints affect the ways in which individuals perceive of and construct social ties. I apply these broadly documented observations regarding limited attention and skewed distribution of social interactions to classify group accounts on Twitter. The advantage of such a language-agnostic, network-based classification strategy is that it can be

applied to any user regardless of language to the extent that users from different cultures, language traditions, and countries face similar resource and cognitive constraints in maintaining social ties.

Theoretical Basis

The proposed method focuses on the limitations in maintaining social relationships, which have been broadly identified across the cognitive and social sciences. At the cognitive level, research shows that the human capacity to remember social relationships around one's immediate social circles, process that information, and adjust behavior accordingly is biologically limited (Dunbar 1995; Miller 1956; Riberts et al. 2009; Powell et al. 2012). At the societal level, maintaining a large number of social relationships implies that an individual bears a heavier burden of maintaining diverse role relationships that often prescribe contradicting normative demands upon the individual, causing role strain (Goode 1960). A consequence of these cognitive and social costs is that the number of social relationships maintainable at any given period in time will be limited. Another postulated consequence is that humans develop a hierarchical perception of group structure of their affiliated groups from the emotionally closest to the most distant (Zhou 2005) and that such a hierarchical group structure is correlated with the heavy skew in the distribution of communication to an individual's alters (Saramaki 2014).

In contrast to individuals, human aggregates, from small groups to organizations, are not subject to the same level of cognitive and role constraints as individuals. With more than one person to manage communication ties, aggregate social entities can collectively wield more cognitive resources and overcome individual level time and cognitive constraints to expand the breadth of social relationships. In the context of Twitter, more than one group member can work on

maintaining a group account, which could increase the number and diversity of communication ties.

Methods

Data Source and Preprocessing. I use a 10% real-time tweet stream (a.k.a Decahose) collected from April 2012 to April 2014. From this corpus, I first geolocate the users by their country using a novel label-propagation method that performs with high coverage and accuracy: labeling over 100M users at a median error of 6.33 km (Jurgens 2013; Compton, Jurgens, and Allen 2014). Using the inferred country labels of the user accounts, I construct within-country @mention networks as described below. By constructing within-country @mention networks, I recognize the fact that the vast majority of @mentions occur among users in the same country and that the overall level of mentioning, which affects network structure, is partly determined by Twitter penetration and the baseline propensity of @mention tweets at the country level (Hong, Convertino, and Chi 2011).

Directed @mention Network. To leverage the widely observed constraint on human social relationships for the purpose of classifying group accounts, I extract information from the @mention tweets that capture social actions among Twitter users. Here, I adopt Max Weber's classical sociological definition of social action where an "action is 'social' if the acting individual takes account of the behavior of others and is thereby oriented in its course (Weber 1978)." Since @mention tweets are explicitly oriented to other users, often for conversational purposes (Honeycutt and Herring 2009; Sousa, Sarmento, and Rodrigues 2010), I maintain that @mentions constitute a less noisy and stronger indicator of interpersonal relationships than

follower ties that have been widely examined in previous Twitter studies (Java et al. 2007; Poblete et al. 2011; Kulshrestha et al. 2012).

Figure 23 shows the overall framework of the method. The procedure starts from extracting the @mentions from tweets in each country. These @mention records are then aggregated while preserving directionality (e.g. user A mentions user B X times) and are used to construct a directed, weighted @mention network for each country.

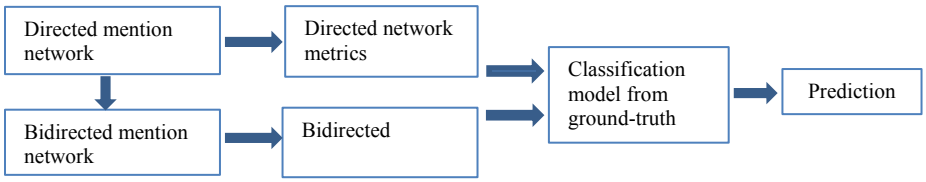


Figure 23. Overall Framework of the Network-Based Classification

Bidirected @mention Network. To incorporate more solid conversational ties, I construct the bidirected @mention network for each country which is a subset of the directed network where a tie exists by definition only if both user A and user B mutually @mention each other at least once. In addition, if account A has no two-way @mentions with any other account (i.e. network isolate), account A is removed from the bidirected network. The underlying assumption for this treatment is based on the observation that meaningful social ties are typically reciprocal (Blau 1964). The out-mentions and in-mentions within the bidirected network, then, indicate the breadth and depth of communication for each account with regard to its socially relevant communication partners. Imposing this bidirectionality constraint filters out unidirected @mention dyads where only one party recognizes the other (e.g. John Doe to Justin Bieber). Another advantage of the bidirectionality constraint is that it effectively filters out bots and spam

accounts which typically never form reciprocal relationships with ordinary users (Thomas et al. 2011; Quercia, Capra, and Crowcroft 2012). Note that it is possible that spam and bot accounts @mention one another to outsmart Twitter's spam filtering algorithm. To guard against such behavior, I further remove all accounts in the bidirected network that are not in the largest connected component. At the end of these preprocessing steps, there are 84.3M Twitter accounts in the largest bidirected network components across 218 countries.

Network Metrics. Using the directed and bidirected @mention networks, I derive a set of metrics to be used as predictors of individual vs. group labels. The directed network metrics represent the overall spectrum of communications involving a given user account, which shows its total communication capacity, irrespective of account type (bot, spam, individual, or group) or the account type of the @mentioned account. While these directed network metrics reveal the observed total amount of communications in the data and, therefore, better approximate the cognitive and social constraints as discussed above, they are impure in the sense that not all incoming or outgoing mentions are conversational or "social" in the Weberian sense; mere name dropping of a celebrity figure in one's tweet does not constitute a meaningful social discourse.

Directed Network Metrics

Log (in-degree). From the directed network, I measure the number of alters who @mentioned ego at least once and take the logarithm of that quantity to address the heavy skew in the distribution.

Log (out-degree to in-degree ratio). Individual users tend to have a balanced ratio of out-

degree to in-degree due to the combined effects of the norms of reciprocity in human interaction (Gouldner 1960) and the cognitive, time-bound, and social role constraints outlined above. Specifically, since an individual's out-degree has some limit due to cognitive and time constraints, her in-degree will also be somewhat limited by the norms of reciprocity. If an individual receives mentions from more than 100 alters, but can maintain communication with only 10 of them, the other 90 who could not engage in conversations with the focal individual would be more likely to reduce communication and, at some point, cease mentioning that individual altogether. On the other hand, groups and celebrities may be less subject to such norms such that those who mention group accounts simply may not hold the same expectations of reciprocity as they would toward individuals.

Log (out-mention to in-mention ratio). Similar to the out- to in-degree ratio, similar norms of reciprocity and constraints may apply at the mention level for individuals. Here again, I predict that a group account will exhibit lower levels of aggregate out- to in-mention ratios than an individual.

Gini coefficients of out- and in-mention signatures.. I extract the out-mention and in-mention distributions of each account from the directed network and measure their skew, which represents the cognitive, time, and social role constraints associated with the hierarchical structuring of one's ego network. Theory predicts that, other factors being equal, less skew should be observed in non-individual accounts as those accounts are less limited by cognitive and time constraints than individuals. The skew of an account's mention distribution can be summarized as the well-known Gini coefficient.

$$G = 1 - \frac{\sum_{i=1}^{n} f(y_i)(S_{i-1}+S_i)}{S_n} \quad (1)$$

In equation (1), $y_i$is the relative out- or in-mention frequency of alter i who is connected with ego through either out- or in-mentions ($y_i < y_{i+1}$), $f(y)$ is the probability mass function, and $S_i = \sum_{j=1}^{i} f(y_i)y_j$ and $S_0 = 0$.

Bidirected Network Metrics

Log (page rank). Previous work on spam detection reports the use of Infochimp's "trust quotient," which is a variant of page rank, for filtering spam accounts (Quercia, Capra, and Crowcroft 2012). I argue that the page rank, which is a measure of popularity, can also be a useful indicator of group accounts to the extent that groups are more systematically driven than ordinary individuals by the pursuit of exposure and influence on Twitter.

Log (alter's mean degree to ego's degree ratio). Human social networks are characterized by assortative mixing (Java et al. 2007) where connected individuals tend to have similar degrees (i.e. positive degree correlation). While evidence is sparse whether assortative mixing applies to group entities as well, there is reason to believe that disassortative mixing should be more prevalent for group accounts on Twitter. Since group accounts use Twitter primarily as a platform to engage with individual users (e.g. individual consumers) rather than with other group entities and since groups tend to have higher degree than individual users, the ratio of the alter's mean degree to ego's degree should be lower for group accounts compared to individual accounts.

Log (directed network in-mention to bidirected network in-mention ratio). The directed network in-mention to the bidirected network in-mention ratio captures the extent to which the account receives interactive or conversational @mentions relative to the total one-way @mentions. I also add a squared term in the logistic regression classifier to capture possible non-linear associations.

Ground-Truth Labels. "Group" accounts in the current context are synonymous to "managed" accounts where more than one individual manages a given Twitter account. By this definition, a celebrity account that is not likely to be managed solely by the individual celebrity is also labeled as a group account. For the ground-truth labels, I construct a stratified user account sample based on indegree to outdegree ratio from the largest bidirected @mention network components of the UK and South Korea (272 users). In addition, I merge these hand-labeled user data with ground-truth organization labels from a previous study that consists of user accounts whose tweets are predominantly in English (103831 users) and in Spanish (118367 users), respectively (De Silva and Riloff 2014). All spam and bot accounts are deleted in this process. In sum, the ground-truth data consist of 177 group accounts and 222,293 individual accounts.

Table 3. Descriptive Statistics of Network Metrics.

	Mean	St.D	Min	Max
Log(bidirected network pagerank)	-14.23	1.78	-18.61	-3.08
Log(directed network indegree)	4.03	1.16	0	13.96
Log(directed network out- to in- degree ratio)	0.15	0.44	-9.95	5.38
Log(directed network out- to in-mention ratio)	0.12	0.5	-10.37	7.36
Log(bidirected network neighbor degree to self degree ratio)	0.46	1.01	-5.43	8.88
Log(directed network inmention to bidirected network inmention ratio)	0.24	0.33	0	12.78
Log(directed network inmention to bidirected network inmention ratio) squared	0.16	1.09	0	163.23

Directed network outmention gini coefficient	0.54	0.14	0	0.96
Directed network inmention gini coefficient	0.54	0.14	0	0.96

Evaluation

Table 3 shows the descriptive statistics of the network metrics of the ground-truth dataset. I first performed a dimension reduction on the labeled data via t-SNE (van der Maaten and Hinton 2008) and plotted the results in Figure 24 to provide a graphical depiction of the discriminatory power of the network metrics. In brief, the t-SNE algorithm maps pairwise distances in the high-dimensional space to distances in a low-dimensional embedding by equating "distance" with a joint probability and learning low-dimensional joint probabilities which are close (in the sense of Kullback-Leibler divergence) to the high-dimensional joint probabilities. Since 99% of the training data are individuals, I randomly selected a subset of individuals (the same as the number of group accounts) before visualization. Concentration of group labels in the upper left side and individual labels in the lower right side suggest reasonable discriminatory power.

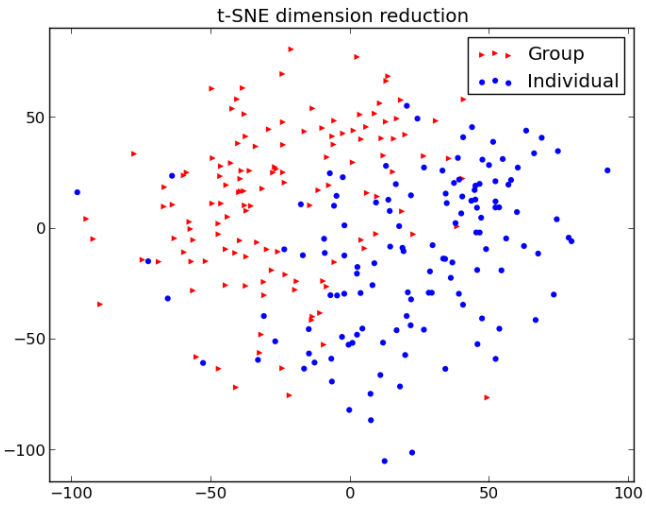


Figure 24. Sample of Ground-Truth Labels Plotted Using t-SNE Dimension Reduction

Next, using the network predictors with the ground-truth labels, I train K-nearest neighbor (KNN) and logistic regression classifiers. 80% of the labels are used for training and the other 20% for evaluation, reflecting the imbalance in composition between group and individual labels. Table 4 presents the aggregate performance of the two classifiers over 30 iterations for both KNN and logistic regression classifiers with the first set of rows reporting the result from models that exclude the Gini coefficients of in- and out-mention signatures and the second set of rows reporting the results from the models that include all metrics. All models achieve above 98% accuracy. However, the high accuracy is illusive since the ground-truth labels are not balanced (i.e. 222293 individuals vs. 177 groups). One would achieve a 99.92% mean accuracy by simply classifying all accounts as individual (222293/(222293+177) = 0.9992). For this reason, the more informative performance measures are precision (positive predictive value) and recall (true positive rate) shown in Table 4. Under fair random guessing ($\rho = 0.5$), the expected precision and recall are 0.000796 and 0.5, respectively. For the KNN classifier using all network predictors, the mean precision of 0.770 is 967 times higher and the mean recall of 0.568 is 1.13 times higher than random guessing, respectively. For the logistic regression classifier using all network predictors, the mean precision of 0.048 is 60 times higher and the mean recall of 0.976 is 1.95 times higher than random guessing, respectively.

I find stark differences between the KNN and logistic regression classifiers in terms of precision and recall where the former achieves higher precision (i.e. higher mean average precision (MAP)) than the latter while the latter shows higher recall than the former (i.e. higher area under the curve (AUC)). The models including the in- and out-mention Gini coefficients marginally outperform the models that exclude them, lending partial support for the cognitive constraint hypothesis.

Table 4. Overall performance of the K-nearest neighbor and logistic regression classifiers over 30 iterations

		KNN		Logistic Regression	
		mean	std	mean	std
No Gini	Precision	0.746	0.090	0.057	0.008
	Recall	0.567	0.082	0.976	0.025
	Accuracy	1.000	0.000	0.987	0.001
	MAP	0.657	0.070	0.516	0.012
	AUC	0.783	0.041	0.981	0.012
All Metrics	Precision	0.770	0.063	0.048	0.009
	Recall	0.568	0.046	0.976	0.026
	Accuracy	1.000	0.000	0.986	0.001
	MAP	0.669	0.041	0.512	0.014
	AUC	0.784	0.023	0.981	0.013

Discussion

Drawing from a number of theories in cognitive science and sociology, which identify sources of interactional constraints at different levels (i.e. time, cognitive, and social role), I introduced a set of language-agnostic network-based models for Twitter group account classification. The KNN classifier exhibited a high level of precision (small false negatives), but a mediocre level of recall (relatively large false positives). On the other hand, the logistic regression classifier identified nearly all group accounts, but at the expense of a high rate of false positives. Manual inspection of these false positive cases suggest that they tend to be active individuals who interact with a rather large number of other Twitter users or are local celebrities who come close to being classified as group account by my manual labeling criteria. Since the vast majority of Twitter users are individuals, the higher rate of false positives in the logistic regression classifier constitute only about 1.4% of all individual accounts.

I maintain that the focus of the specific application should dictate the choice between the KNN

and the logistic regression classifiers. For example, if the group account classification is used as a filtering step in a network analysis of individual Twitter users, the KNN classifier may be a more sensible choice, given that it is less prone to misclassifying high-degree, "influential" individuals whose presence or absence will affect the observed network structure significantly. On the other hand, if the objective is to discover as many potential group accounts as possible as an intermediate step, for example, in identifying target accounts for a B2B advertisement campaign, the logistic regression classification may prove to be more useful.

A potentially fruitful future direction would combine the network metrics I constructed in this paper with the temporal signatures inscribed in users' tweets (Tavares and Faisal 2013). Since temporal signatures and network signatures are both language-agnostic, the potential applicability of the combined classification could be extended to language communities, which do not yet possess reliable computational tools for accurate language-based classification.

References

Blau, Peter. 1964. Exchange and Power in Social Life. John Wiley and Sons: New York.

Brewer, Devon. 1995. "The social structural basis of the organization of persons in memory." *Human Nature* 6(4):379-403.

Compton, Ryan, David Jurgens, and David Allen. 2014. "Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization." *arXiv*. http://arxiv.org/abs/1404.7152.

De Choudhury, Mnmun, Nicholas Diakopoulos, and Mor Naaman. 2012. "Unfolding the event landscape of twitter: Classification and exploration of user categories." Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, *CSCW* pp. 241-244. ACM, New York.

Peter DeScioli, Robert Kurzban. 2009. "The alliance hypothesis for human friendship." *PLoS One*, 4(6): e5802.

De Silva, Lalindra, Ellen Riloff. 2014. "User type classification of tweets with implications for event recognition." *Proceedings of the Joint Workshop on Social Dynamics and Personal*

Attributes in Social Media, pp. 98-108.

Dunbar, Robin. 1995. "Neocortex size and group size in primates: A test of the hypothesis." *Journal of Human Evolution* 28:287-296.

Goode, William. 1960. "A theory of role strain." *American Sociological Review* 25(4):483-496.

Gouldner, Alvin. 1960. "The norm of reciprocity: A preliminary statement." *American Journal of Sociology* 25(2):161-178.

Honeycutt, Courtenay and Susan Herring. 2009. "Beyond microblogging: Conversation and collaboration via twitter." Proceedings of the 42nd Hawaii International Conference on System Sciences, *HICSS* '09 pp. 1-10.

Hong, Lichan, Gregorio Convertino, and Ed Chi. 2011. "Language Matters in Twitter." The 5th International AAAI Conference on Weblogs and Social Media, *ICWSM* '11, pp. 518-521.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. "Why we twitter: Understanding microblogging usage and communities." *Proceedings of the 9th WebKDD and 1st SNA-KDD* 2007 Workshop on Web Mining and Social Network Analysis pp. 56-65.

Jurgens, David. 2013. "That's what friends are for: Inferring location in online social media platforms based on social relationships." The 7th International AAAI Conference on Weblogs and Social Media, *ICWSM* '13, pp. 273-282.

Kulshrestha, Juhi, Farshad Kooti, Nikravesh Ashkan, and Krishna Gummadi. 2012. "Geographic dissection of the twitter network." The 6th International AAAI Conference on Weblogs and Social Media, *ICWSM* '12, pp. 202-209.

Miller, George. 1956. "The magical seven plus or minus two: Some limits on our capacity for processing information." *Psychological Review* 63:81-97.

Newman, Mark. 2002. "Assortative Mixing in Networks." *Physical Review Letters* 89(20):208701.

Poblete, Barbara, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. "Do All Birds Tweet the Same?: Characterizing Twitter around the World." ACM *CIKM* 1025-1030.

Powell, Joanne, Penelope Lewis, Neil Roberts, Marta Garcia-Finana, and Robin Dunbar. 2012. "Orbital prefrontal cortex volume predicts social network size: An imaging study of individual differences in humans." *Proceedings of the Royal Society B* 279(1736):2157-2162.

Quercia, Daniele, Licia Capra, and Jon Crowcroft. 2012. "The social world of twitter: Topics, geography, and emotions." The 6th International AAAI Conference on Weblogs and Social Media, *ICWSM* '12, pp. 298-305.

Roberts, Sam, Robin Dunbar, Thomas Pollet, and Toon Kuppens. 2009. "Exploring variation in active network size: Constraints and ego characteristics." *Social Networks* 31(2):138-146.

Saramaki, Jari, E. A. Leicht, Eduardo Lopez, Sam Roberts, Felix Reed-Tsochas, and Robin Dunbar. 2014. "Persistence of Social Signatures in Human Communication." *Proceedings of the National Academy of Sciences* 111(3):942-947.

Sharma Naveen, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012. "Inferring who-is-who in the twitter social network." Proceedings of the 2012 ACM Workshop on Online Social Networks, *WOSN* pp. 55-60.

Sousa, Daniel, Luis Sarmento, and Eduarda Rodrigues. 2010. "Characterization of the twitter @replies network: Are user ties social or topical?" Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, *SMUC* '10, pp. 63-70.

Tavares, Gabriela and Aldo Faisal. 2013. "Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users." *PLoS One* 8(7): e65774.

Thomas, Kurt, Chris Grier, Vern Paxson, and Dawn Song. 2011. "Suspended accounts in retrospect: An analysis of twitter spam." Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, *IMC* '11, pp. 243-258.

van der Maaten, Laurens, Geoffrey Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research* 9:2579-2605.

Weber, Max. 1978. *Economy and Society*. University of California Press.

Wu, Shaomei, Jake Hofman, Winter Mason, and Duncan Watts. 2011. "Who says what to whom on twitter." Proceedings of the 20th International Conference on World Wide Web, *WWW* 2011, pp. 705–714. ACM, New York (2011)

Zhou, W.X., D. Sornette, R. Hill, and R. Dunbar. 2005. "Discrete hierarchical organization of social group sizes." *Proceedings of the Royal Society B* 272(1561):439-444.