Geography of Twitter Networks

Yuri Takhteyev
University of Toronto
Faculty of Information
yuri.takhteyev@utoronto.ca

Anatoliy Gruzd
Dalhousie University
School of Information Management
gruzd@dal.ca

Barry Wellman
University of Toronto
Department of Sociology
wellman@chass.utoronto.ca

Revised on February 16, 2011.

Abstract

The paper examines the influence of geographic distance, national boundaries, language and frequency of air travel on the formation of social ties on Twitter, a popular micro-blogging website. Based on a large sample of publicly available Twitter data, our study shows that a substantial share of ties lies within the same metropolitan region, and that for ties between regional clusters, distance, national borders and language differences all predict Twitter ties. We find that the frequency of airline flights between the two parties is the best predictor of Twitter ties. This highlights the importance of looking at pre-existing ties between places and people.

1. Introduction

Social contact benefits from physical proximity. This fact of social life is so basic, that for a long time proximity was often taken for granted: social interaction was understood to mean face-to-face interaction, for which distance acts as a powerful barrier. The fact that being near each other facilitates the formation of social ties was for the most part not so much a finding of social research, but its basic assumption. Social network analysts were among the first to challenge this assumption, showing that the social network approach allowed following social ties as they cross space, mapping the more distributed communities that were replacing those based on neighborhoods (Webber, 1963; Wellman, 1979; Fischer, 1982). A few decades later, the Internet brought new opportunities for maintaining social ties over distance, as well as greater awareness of such possibilities. Pundits proclaimed that distance was dead (Cairncross, 1997). However, the evidence challenged this assertion, showing that proximity still makes a difference. Yet, most such studies have looked at e-mail, which was shown to help extend and maintain existing strong ties (e.g., Mok et al, 2010). Recent years have brought new ways of interacting over the Internet, some of which seem less tied to strong ties or face-to-face contacts. Are these new forms of electronic interaction also affected by proximity?

We focus on one such Internet-based system, Twitter, a popular social networking and micro-blogging service that allows users to post and read short messages, limited to 140 characters. Such messages — called "tweets" — are usually public, visible to anyone on the Internet. (Users can make their tweets private, but most do not. Our own sample suggests that only 10 percent of the users protect their tweets.) While tweets can be read anonymously, the preferred method is to create an account and select a set of users that you want to "follow," so that you would see recent tweets from those accounts whenever you log on to Twitter. A users' choice of whom to follow is public. Additionally, Twitter users usually specify their geographic location in their profile. Twitter thus offers us a publicly available, spatially embedded network dataset, a rare luxury in network analysis (Butts and Acton, 2010).

Our analysis shows that distance matters on Twitter, both at short and longer ranges: 39 percent of the ties are shorter than 100 km and ties up to about 1000 km are substantially more common than

§ The authors thank Lilia Smale, MinKyu Kim, Andrew Hilts, Annie Shi, and the anonymous reviewers for their help in preparation of this paper.

we could expect if they were formed at random. This result is interesting, considering the ease with which long-distance Twitter connections can be formed. We also look at several other variables that can impede or facilitate ties while being closely intertwined with distance. We find that national boundaries and a shared language both affect ties but do not explain away the effect of physical proximity. Frequency of airline connections, on the other hand, predicts non-local Twitter ties better than proximity, with the latter adding relatively little to a model that already includes flight frequency. Thus, the strength of prior ties between places matters more than the simple distance between them.

2. Twitter: Global Reach and Weak Ties

Several aspects of Twitter make it a particularly valuable case for analysis. First is Twitter's popularity and international reach. When we collected our data in the summer of 2009, Twitter (founded in 2006) was already attracting tens of millions of unique visitors per month (Schonfeld, 2009) who were posting and reading millions of messages every day. Our data suggests that over half of the service's users were located outside the United States at the time, which included many users in Brazil, the UK, Japan, Australia and Indonesia.[1] This wide distribution of users allows us to explore the effects of distance at different scales, from fairly short to nearly antipodal.

The second relevant aspect is the relative weakness of Twitter ties. Granovetter (1973) defines the strength of a tie as "a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (p. 1361). Even compared to other forms of electronic communication, Twitter-based interaction fails Granovetter's definition on all counts. 140–character messages take little time to read and even less time to ignore. The fact that messages are publicly broadcast reduces the level of intimacy and emotional intensity of such communication. Finally, Twitter ties are often asymmetric: if A follows B, B does not have to follow A. In our sample, 60 percent of the ties are unidirectional.[2] This aspect of Twitter contrasts with other "social networking" sites, such as LinkedIn or Facebook that often aim to capture pre-existing ties and enforce bi-directionality. Instead, Twitter may be better compared to blogs, where there is similarly low cost (technically or socially) to establishing a unidirectional tie by becoming a reader. However, unlike the weak ties between bloggers and their readers, which most often stay invisible, Twitter ties can be easily observed and analyzed.[3]

The combination of weak, low-cost ties and global popularity creates an opportunity for people to make links that transcend distances and national borders. Twitter's ability to support such ties became the subject of many news articles in 2009 when the service was actively used by residents of Tehran, Iran, and most recently in 2011 by people in Egypt and Tunisia, and not only to coordinate

1 Some of the other reports, e.g., one by Sysomos (2010) produced around the same time suggested that American users accounted for slightly more than half of the users, while earlier reports by Java et al. (2007) and Krishnamurty et al. (2008) show a lower share for North American users. Unfortunately, none of such reports describe the data collection and geo-coding process in sufficient detail for us to investigate the possible sources of discrepancies.

2 In other words, the majority of Twitter ties are so weak that the followed users do not bother to reciprocate the followers' interest in their tweets, despite the rather low cost of doing so. Compare this with 18-20 percent rate of unreciprocating for a sample of LiveJournal users in Gaudeul and Peroni (2010). See also Huberman et al. (2009) on the weakness of Twitter ties.

3 Similar to Twitter ties, LiveJournal "friendship" ties can also be traced, and have in fact been studied, for example, by Liben-Nowell et al. (2005) who found that distance has an effect. LiveJournal "friendship" ties, however, are stronger ties, as suggested by the much lower average number of connections. Liben-Nowell et al. report that an average user in their sample has eight "friends." For Twitter, we found that users on average had around 400 outgoing ties overall, with around 100 ties per user after we excluded those with over 500 ties. (See also the previous footnote on the higher rate of reciprocation in LiveJournal.)

local protests against the national regime but also to inform the world about these protests (Cardwell, 2009). We must ask, however, whether such cases are typical or exceptional. This paper provides a quantitative investigation of the effect of distance on Twitter ties.

3. Ties, Distance and Related Variables

As discussed below, distance has been shown to have an effect on social ties, including those based on electronic communication. The "weak" nature of Twitter ties may reduce the effect of distance, but would be unlikely to eliminate it altogether. It is important to ask, however, not only *whether* distance matters, but also the *mechanisms* through which distance and ties relate.

It is clear that distance does not usually influence social ties directly. Even in its purest form, distance usually impedes the formation of social ties by raising the cost of travel that is required for face-to-face interaction. Consequently, the effect of distance would likely be mediated by the existing transportation infrastructure. Distance is also intertwined with other factors. We focus on two: national boundaries and language differences. Both limit interaction while being correlated with distance, and their effects can reduce the average length of ties. We thus focus on four variables in our investigation: the physical distance between the users, the ease of travel measured by the frequency of flights between their cities, whether the users are in the same country, and the match in language.

3.1 Physical Distance

In the late 1970s, Wellman's second study of East York, a Toronto neighborhood, found expanding social ties: only 22 percent of East Yorkers' close friends and relatives were in East York and none had most of their active ties living within a mile's walking distance (Wellman, 1979; Wellman et al., 1988). Yet, East Yorkers' ties still depended principally on face-to-face interaction, which could now happen at the scale of a metropolitan area thanks to the expanded use of cars. The telephone was important, but its use was complementary rather than substituting for it (Wellman and Tindall, 1993). A more recent study by Mok et al (2010) found that today's East Yorkers maintain more distant connections, some reaching as far as Europe or Pakistan. Technologies such as email and Skype help maintain such ties (see also Boase, 2008). Nonetheless, the number of social ties, drops sharply even as the distance increases between 1 and 20 miles. Most of East Yorkers' email use is also local. Other studies have found similar results (e.g., Wellman & Hogan, 2006; Wang & Wellman, 2010). Studies of "friendship" ties on LiveJournal (e.g., Liben-Nowell et al., 2005) also found an effect of distance.

The low cost of Twitter connections can make it less sensitive to distance, giving people an opportunity to "follow" others around the world, without being constrained by the spatial extent of their face-to-face networks. Nonetheless, we can expect that like other forms of electronic communication, Twitter ties may be complementary to face-to-face interaction. We can also expect people to follow friends-of-friends and people they have heard of, all of whom are more likely to be nearer than randomly chosen alters. Users may also select Twitter accounts that distribute information about topics they find relevant — again likely displaying a bias towards nearer sources. We thus expect that Twitter ties will be influenced by geographic distance. We do not have, however, an *a priori* hypothesis as to how strong that effect is likely to be and whether distance is only important at a short range (local vs. non-local) or at different geographic scales.

3.2 Air Travel

Perhaps the most important reason why distance limits the formation of social is because it reduces the opportunities for face-to-face interactions. The strength of this effect, however, depends on the ease of travel between the places. For longer distances, one of the important components of travel is the availability of airline flights (e.g., Zook & Brunn, 2006). Additionally, frequency of airline connections can be interpreted as a proxy for more general connectedness. Research on global cities has shown, for example, that the cities that are most central in the network of airline connections are also important in the network of relationships among transnational accounting firms (Beaverstock et al., 1999). In other words, frequent flights between New York and London not only facilitate travel but also indicate that New York residents have many reasons to travel to London. The frequency of airline flights may thus be a better predictor of non-local ties than physical proximity.[4]

3.3 National Boundaries

Today's world is organized as a system of nation-states: that is, units that tie together territory, political power and identity. National boundaries inhibit social ties in multiple ways. Most trivially, they affect mobility since people usually can move freely within their states, but often need visas to move between them. National boundaries often also define communities of interest: people usually care about domestic events more than comparable events abroad. This is in part because some of the most important decisions affecting their lives are made at the national level, but also because access to mass media is often shaped by national boundaries. At the same time we must avoid "methodological nationalism" (Wimmer & Schiller, 2002) or "implicit state-centrism" (Derudder & Witlox, 2005) and avoid taking nations for granted as a unit of analysis. Instead, the extent to which nations matter must be treated as a question to be addressed empirically. In our case, we expect national boundaries to reduce the likelihood of Twitter ties, separately from the effect of physical distance.

When looking at the effect of national boundaries we must consider the fact that not all nations are created equal. National populations affect the likelihood of a tie between two nations, and core-periphery structures affect differential attention between countries (Smith & Timberlake, 1995). People who live in large and powerful nations may have more opportunities for domestic connections while also having less interest in foreign events. We can expect Twitter users in such countries to have a disproportionate number of domestic ties. Residents of smaller and less powerful nations, on the other hand, may have a greater interest in what happens abroad, since their lives are quite often affected by foreign events, and they may have fewer opportunities for connecting to like-minded individuals locally. We expect people living in such countries to be more active in following those who tweet from abroad.

3.4 Language

Social interaction depends on the two parties' ability to communicate, which nearly always requires that they are competent in the same language or rely on a bilingual mediator. Consequently, language differences can structure social interactions (e.g., Hutchinson, 2005; Barnett & Choi, 1995). Like national boundaries, linguistic differences are intertwined with distance. People living in the same place typically share a language (or several). Additionally, because of ancient settlement patterns and more recent patterns of colonization, people in nearby places are more likely to speak the same or similar languages than people who are far apart. We may expect, however, that shared language

4 Note that the frequency of flights is strongly correlated with distance (in our data we find a correlation of -0.82 between logs of distance and the number of flights), and may serve as a proxy for it, especially in the shorter range.

competency would have an effect that is separate from the effect of distance.

Yet language stands in a more complicated relationship to distance than national boundaries. While people usually speak the dominant language of the city or country where they live, they can *also* speak other languages. In particular, many people around the world today learn English in addition to their local and national languages. (See for example, Herring et al, 2007, on the role of English in LiveJournal networks.) We can thus hypothesize two separate effects: ties may be more likely between users in places with a strong linguistic connection and between pairs of users who tweet in the same language.

4. Building a Sample of Twitter Ties

The primary data source used in this article is a sample of dyads of geocoded Twitter accounts connected by a "follow" relation. To assemble this set of dyads, we first collected a sample of ego accounts, then sampled one alter from among the accounts followed by each ego, resulting in a set of ego-alter pairs in which the ego subscribes to (or "follows") the Twitter messages authored by the alter. (Additional details about our method are provided in the Appendix.)

4.1 Collecting the Sample of Egos

To build our sample of egos, we first collected a large number of Twitter messages by querying Twitter's "public timeline," a public interface provided by Twitter that returns twenty of the most recent public messages (Twitter, 2011). We used a Python script to query the public timeline every 25 seconds for a period of seven days in August of 2009, collecting a total of 481,248 messages (Table 1). The tweets included in the public timeline represent a small subset of the messages posted that week. It may not necessarily constitute a *random* sample, since we do not know exactly what method Twitter uses for selecting tweets that go into the public timeline. Furthermore, we have received reports (*citation to personal communication removed for review*) that the mix of messages may be somewhat different depending on which account (if any) is used to request the tweets. Despite those problems, the public timeline is commonly used to sample Twitter messages (e.g., Java et al., 2007; Naaman et al., 2010; Golder & Yardi, 2010).

Table 1 around here. ("Sample size at different steps.")

Our sample includes an equal number of messages for each 25-second period of the week, without accounting for diurnal and weekly cycles in Twitter use. Some sources suggest that the rate at which Twitter messages are produced varies throughout the day with twice as many messages produced at the peak time (1:00 pm in New York) than at the quietest time (5:00 am in New York). Our dataset may consequently under-sample the users who tweet on a New York schedule (which would likely include many of those in North and South America) and oversample those who tweet when New York sleeps. (This distortion, however, would only affect the original sample of egos and not the length of the ties, since the ego's connections were sampled in a separate step.)

4.2 Geocoding and Subsampling

Most of the messages in the original sample (75 percent) had some location value associated with them. They were sent by users who either specified a location in their Twitter profile or, in the minority of cases, used a Twitter client that automatically updated the location field in their profile. The format of those descriptions varied. Some of them provided specific addresses or even coordinates. Many identified cities. Some named a country, a continent, or the planet. Some referred to fictional locations or did not appear to refer to any locations at all. We found, however, that about 85 percent of such descriptions referred to a real place, at the level of a country or smaller, while 65 percent referred to a geographic unit the size of a major metropolitan area or smaller (Table 2). We assume that descriptions referring to real places represent users' actual locations, that is, either places where they tend to be in general or where they were at the time the message was posted. However, we were careful to properly classify location descriptions that suggested that they were wishful or outdated, for example, “America i wish, England :(,” or “From Dallas but live in ATL.”

A small minority of location descriptions (6 percent) provided exact geographic coordinates, which were nearly always prefixed either with “iPhone:” (about 39 percent of locations with coordinates) or “ÜT:” (57 percent), suggesting that they were submitted either by a Twitter application running on the iPhone or by ÜberTwitter, a popular Twitter application for BlackBerry. The rest described their location using a variety of ways. While US locations often followed the city-comma-state pattern, many alternative conventions were employed. For example, locations in Brazil often included dashes instead of commas (“Maceió-AL”) or put the name of the state before the name of the city (“RJ - Petrópolis”). Many locations were identified with just the name of a city. While most descriptions used Roman letters, some used other scripts, including those, like Japanese, that do not separate words with spaces (Table 2). The language in which the location was named did not always match the language of the location. We also encountered wide use of nicknames, such as “L.A.”, “Floss Angeles,” “Floss Town (LA), CA,” as well as idiosyncratic spelling (“LosAngeles”). We found that automatic geocoding of such data was quite error prone; it resulted in either false positives or a failure to correctly locate place descriptions. Furthermore, this appeared to introduce a geographic bias, as locations outside the United States were more likely to be misidentified. For this reason, we decided to code the locations by hand, using a variety of reference materials, including Google Maps, to resolve place names that we were not familiar with, but avoiding applying any of them blindly.

The need for manual processing made it impossible to geocode locations for all collected tweets. Instead, we took a sub-sample of users who provided a location description, drawing twenty users from each one-hour segment of the seven day period. In the small number of cases where we drew a user who had already been included in a sample based on an earlier hour, we drew a replacement, resulting in a sample of 3,360 unique users. We refer to these users as *egos*.

Table 2 summarizes the precision with which we were able to identify the location of the 3,360 egos. We divide location descriptions into three classes. We considered a description to be *specific* if it identified a place with an area of up to 25,000 square km.[5] Egos who provided such specific locations or actual coordinates comprised 65 percent of the sample of 3,360 egos and were used to investigate the extent of spatial clustering. An additional 20 percent of the egos provided locations sufficiently narrow to determine the country. We use those cases for our analysis of Twitter use by country.

5 This value roughly represents the size of the largest of the metropolitan regions frequently named in our sample. (For example, the San Francisco Bay Area has 22,000 km².) It also roughly corresponds to the upper limit on commuting distance. In most cases, such descriptions included names of cities or metropolitan agglomerations. For consistency, however, we applied the same criteria to all named places (states, provinces, and countries) that fell under the 25,000 km² threshold, for example “Wales” and “Jamaica”.

Our success rate in identifying users' location translates into an overall response rate of 48 percent at the level of metropolitan area and 63 percent at the level of a country. Both numbers account for the cases where no location was specified at all. It is possible that the users' location influences the likelihood of them reporting the location (or the precision with which they report it), potentially creating a non-response bias. Such a bias, however, is unlikely to affect the main conclusions of the paper, which focus on the effect of distance and related variables on Twitter ties rather than on the geographic distribution of Twitter users *per se*. When we analyze the effect of distance, we do so in relation to the observed distribution. The only kind of non-response bias that would undermine our conclusions would be if users were substantially less likely to specify their location when they have long-range ties. We cannot rule out this possibility, but we do not see reasons to expect such an effect on a substantial scale.

Table 2 around here. ("Location precision in the sample of egos.")

4.3 Sampling the Alters

Since our sample of egos included a relatively small number of users picked from among hundreds of millions of user accounts, the sampled egos were predominantly connected to users outside our sample of egos. For this reason, we did not attempt to analyze ties between sampled egos, and instead sampled an additional user — an *alter* — for each ego whose location was identified at the level of country or better and who followed between one and 500 Twitter accounts, by randomly drawing an account from among those "followed" by each of those egos. We then coded the locations of the alters using the same procedure as we did for the egos, removing those pairs where the alter could not be assigned to a country. In the end, we obtained a sample of 1,953 ego-alter pairs with both the ego and the alter could be assigned to a country, including 1,259 pairs with a "specific" location for both parties (Table 1).

4.4 Aggregating Nearby Locations

Since specific locations vary substantially in precision and since users can often choose between a range of specific names for the same place (e.g., "Palo Alto" vs. "Silicon Valley" vs. "SF Bay"), we aggregated nearby locations within each country, by assigning a set of coordinates (obtained from Google Maps) to each location smaller than 25,000 km² and then merging nearby locations within each country by replacing their coordinates with a weighted average of the coordinates of the merged locations. This reduced our location descriptions to a set of 386 *regional clusters,* which are comparable in size to metropolitan areas. We labeled each cluster with the most common name associated with it in our sample. For example, the cluster centered on Manhattan is referred to as "New York."

5. Analysis

In this section we analyze the factors affecting the formation of Twitter ties. We first look at the effect of each variable identified earlier based on theoretical considerations: the actual physical distance, the frequency of air travel, national boundaries and language differences. In addition to presenting the descriptive statistics demonstrating the effects of each variable and investigating the nature of such effects, we correlated the effects using the Quadratic Assignment Procedure (QAP, Krackhardt, 1987; Butts, 2007). In the last subsection we also examined the relationship between the

variables using QAP regression (Double Dekker Semi-Partialling MRQAP). All statistical calculations were done using UCINet 6.277 (Borgatti, et al., 2002).

For correlation and regression analysis we used networks with nodes representing the 25 largest regional clusters of users (see previous section). The edges of each network were then assigned weights based on an operationalization of the corresponding variable. For the dependent variable network the weight of the edges represented the natural logarithm of the number of Twitter ties between users in the two clusters. The weights for the edges in the independent variable networks are described below, when we discuss each variable. We have found that the network of 386 Twitter clusters was extremely sparse, since the number of ties in the sample was small relative to the number of nodes. As a result, more than 99 percent of cluster pairs had zero Twitter connections between them, leading to low correlation (between 0.05 and 0.1) with the comparison networks, with the only exception of the network of airline connections.[6] For this reason, we limited our correlation and regression analysis to the ties between just the 25 largest clusters, which allowed for a much denser Twitter network (an average of 0.76 ties per pair).

5.1 Physical Distance

The use of Twitter is concentrated in the United States, which accounts for 49 percent of our sample of egos, 54 percent of the alters, and six of the ten largest clusters (Table 3). At the same time, over half of the egos are in other countries, as are four of the ten largest clusters: Tokyo, São Paulo, and two clusters in the United Kingdom. In this sense, Twitter users are distributed quite widely around the globe. In addition to the relative concentration of users in certain countries, however, we also observe a very a substantial concentration a relatively small number of specific local clusters. 25 clusters account for 54 and 61 percent of the egos and alters respectively. This level of concentration exceeds the general concentration of the population in major urban agglomerations.[7]

Table 3 around here. (“Top clusters.”)

Being in the same cluster also has a strong effect on the formation of ties: 39 percent of the ties between egos and alters fall within the same regional cluster. The large share of in-cluster ties can be partly explained by the substantial degree of clustering: when users are concentrated in a handful of places, a large share of ties would be local even if ties were formed randomly, with full disregard for location. The share of local (in-cluster) ties, however, is substantially higher than what we would expect just due to clustering. Considering the distribution of egos in our sample, only 2 percent of the ties would be local if the ties were formed randomly. (An average user’s cluster accounts for 2 percent of the total number of egos.)

Figure 1 shows the distribution distances between egos and their alters, comparing them to two

6 Note that all the airline network was very sparse, much like the Twitter networks. The other networks, by comparison, had non-zero values for all pairs.

7 For example, the New York cluster in our sample accounts for 17 percent of US-based egos, while the New York Metropolitan Area (which exceeds the size of our “New York” cluster) accounts for only 6 percent of the United States population. For the two main clusters located outside North America and Europe, the degree of concentration is even more substantial: the São Paulo cluster accounts for 37 percent of egos located in Brazil, while Tokyo accounts for 64 percent of those located in Japan.

simulated baselines and showing that distance also has an effect on non-local ties. The observed distribution is shown as the thick solid line. When analyzing the distribution of tie lengths, it is again important to consider the uneven distribution of users' location around the globe. If ties were formed by picking random points on the surface of the planet (with full disregard for uneven distribution of land mass and population), we would expect a symmetric distribution on the range from 0 to 20,000 km, with a peak at 10,000 km, represented by the smooth thin line in Figure 1 (labeled "simulation 1"). Twitter users, however are not distributed evenly around the globe. (Nor is human population in general.) This uneven distribution substantially skews the expected distributions of distances between egos and alters towards shorter ties. Further, since the users are concentrated in a few clusters, we can expect the distribution to peak at values corresponding to distances between major clusters.

This distribution is demonstrated by the second simulation (Figure 1, medium line, labeled "simulation 2"), in which egos, located where they are in our sample, form ties among each other at random. The graph shows a substantial number in the very first bin (0-200 km), followed by a decline in bins representing longer distances. The count goes up, however, as we approach bins that include distances that span the two coasts of the United States, with a particularly sharp peak for the 3,800-4,000 km bin, which catches the distance between New York and Los Angeles. This peak is followed by another valley, corresponding to not-quite-transatlantic distances, and then a rise as we reach Europe. The simulation shows another large peak corresponding to the distance between New York and São Paulo, followed by one matching the distance between New York and Tokyo. We see relatively few ties longer than 12,000 km, since the antipodal points of all major clusters fall in the ocean.

Compared with this baseline, the observed distribution of tie lengths shows a clear surplus of ties for distances to up 1,000 km, a somewhat mixed record from there to 5,000 km and a consistent deficit of ties at greater distances. We note, though, that the peak in the number of ties at the New York – Los Angeles distance is actually *higher* than we would expect if ties were formed randomly. On the other hand, several other expected peaks remain unrealized. In particular, we observe no peaks at the values corresponding to the distances between New York and São Paulo, New York and Tokyo, and Tokyo and São Paulo.

For network comparison we created a "distance" network in which the weight of edges was set to a natural logarithm of the great-circle distance between the two clusters, calculated using the standard haversine formula. The comparison of this network to the network of Twitter ties for the top 25 clusters shows a correlation of -0.45 for the top 25 clusters, with $p<0.001$ (Table 4). We note that the our dependent network ("Twitter") is based only on ties that connect users in *different* clusters, omitting the 39 percent of the ties that fall within clusters. The correlation with the distance network, therefore, cannot be explained simply by the large number of local ties, but rather, shows the effect of distance on *non-local* ties.

Table 4 around here. ("QAP Correlations")

5.2 Air Travel

To investigate the effect of the ease of travel on Twitter ties we obtained from [name suppressed] a dataset showing a number of direct flights between pairs of 3,023 airports on five

different days in 2008 and 2009. We assigned those flights to pairs of clusters by matching each cluster to the airports located within 100 km from its center. We then constructed a network by giving each pair of clusters a weight based on natural logarithm of the observed number of flights between the airports assigned to each of them.

Comparing the air travel network with the network of Twitter ties shows a correlation of 0.51 for the top 25 clusters, with $p<0.001$ (Table 4). The network of flights thus appears to actually be a better predictor of non-local Twitter ties than the physical distance. One interpretation of the predictive power of flight frequency is that frequent flights facilitate travel, which allows for formation of face-to-face ties and increases the likelihood of Twitter connections. (This may, for example, include fact that when people travel or move they may continue to follow people back home.) Another interpretation suggests that flight connections themselves reflect the structure of the world city system, and that Twitter ties are influenced by this structure. Our data does not allow us to disambiguate between those two interpretations. We also note that top Twitter clusters intersect only to an extent with the Alderson and Beckfield's (2004) ranking of world cities based on multinational corporations' branch headquarters. (Of Alderson and Beckfield's top 25 cities by in-degree or "prestige," 13 appear in the top 25 Twitter clusters ranked by in-degree centrality, with another 6 appearing in top 100.)

5.3 National Borders

Of the ties that were matched to countries, 75 percent connect users in the same country. This prevalence of domestic ties is partly explained by the high frequency of *local* connections, since all local ties are domestic. Looking at just the non-local ties (i.e., ties between users in different clusters), we find that the share of domestic ties is lower but still substantial: 63 percent.

As with distance, the high frequency of domestic ties can be partly explained by the concentration of users in a small number of countries, with nearly half of them in the United States. The share of domestic ties, however, substantially exceeds what we would expect if users formed connections randomly while being distributed as they are now, which result in only 26 percent of the ties being domestic. Further, the surplus of domestic connections holds for all major countries, including those that account for just a small fraction of the egos, as shown in Table 5. (The effect is somewhat reduced for countries that have only one major cluster, since in those cases removing local ties means removing the majority of the domestic ties.) The table also shows that the share of domestic ties is generally higher for non-English speaking countries (as long as they have several clusters), yet even the English-speaking countries show a higher share of domestic ties than would be expected from their share of egos. A comparison between the network of Twitter ties between the top 25 clusters and "domestic" network (where edges were set to 1 for domestic ties and 0 for international) shows a correlation of 0.44, with $p<0.001$ (Table 4).

Table 5 around here. ("Top Countries.")

The substantial share of the United States in the sample warrants a comparison with other countries. We find that the share of domestic ties is lower for egos located outside the United States: 62 percent of all ties and 42 percent of non-local ties. However, the share of domestic ties is *higher* for pairs where both parties are located outside the United States: 80 percent of all ties and 65 percent of

non-local ones. In other words, Twitter users outside of the U.S. can be said to have a somewhat more international orientation than American users, but only in the sense that they tend to follow users in the U.S.

It is also important to note the differences in the pattern of outgoing ties (following) and incoming ties (being followed). As column 5 in Table 5 shows the majority of U.S. international ties are incoming: Twitter users in the United States are often followed from abroad, with over 3 incoming ties for each outgoing tie. For some of the other countries, on the other hand, international ties are overwhelmingly outgoing. For users in Brazil, for example, the ratio of incoming ties to outgoing is nearly 1:5. Brazilian users of Twitter actively follow foreign accounts, but receive little attention in return.

The more domestic orientation of the American users also reflects itself in how they describe their locations. When coding the locations we noted whether the country was stated explicitly or implied (e.g., "São Paulo, Brasil" vs. just "São Paulo"). As shown in column 6 of Table 5, only 8 percent of U.S. location descriptions explicitly name the country, compared to, for example, 55 percent of locations in Brazil. This may suggest that American users of Twitter either see their audience as exclusively domestic (even though it is not), expect foreign users to know the names of American cities, or simply do not think about Twitter users abroad. The United States is closely followed by Japan, where only 25 percent of location descriptions identify the country explicitly. However, in the case of Japan, this may be explained by the fact that in the overwhelming majority of cases, locations in Japan were identified in Japanese (using *kanji* or *kana*), which makes them intelligible only to people who know Japanese and would be familiar with Japanese cities. Additionally, Japanese users are followed almost exclusively by others in Japan: ties from foreign egos account for a relatively small fraction - 10 percent - of the ties received by Japanese alters. The Brazilian users have proportionally even fewer incoming foreign ties. This does not, however, stop them from identifying their country explicitly.

5.4 Language

A large majority of egos (62 percent) and alters (68 percent) are located in countries where English is the dominant language. Almost all egos (96 percent) located in the English-speaking countries follow alters who are also located in English-speaking countries. This number, of course, reflects in part the large share of domestic ties within the United States. However, even for egos located in English-speaking countries other than the United States, 91 percent of ties are to English speaking countries. For non-English speaking countries, the share of ties to users in countries with the same dominant language is lower but still significant: 69 percent. (For the most part, however, this latter number simply represents the share of domestic ties for the non-English speaking countries.)

For the purposes of correlation analysis we built a language match network using a dataset of access to language-specific versions of Wikipedia from each country. For example, the dataset indicated that that requests for English Wikipedia accounted for 94 percent of all requests coming from the United States and for 15 percent of requests coming from Brazil, while requests for the Portuguese Wikipedia accounted for 83 percent of requests coming from Brazil and 0.16 percent of requests coming from the United States. To get a measure of proximity between a pair of clusters we summed the products of the two countries' preferences for languages. For example, New York – São Paulo pair received a weight of 0.14, reflecting the match in English (0.94*0.15 = 0.14), together with negligible terms for other languages (0.001 for the match in the preference for Portuguese and about the same amount for the match in preference for Spanish). New York – Tokyo pair received 0.03, while New

York – Amsterdam pair received a weight of 0.39, reflecting primarily the much lower preference for the English Wikipedia among the requests from Japan and the much higher preference among requests coming from the Netherlands.[8] The resulting language network shows a correlation of 0.42 with the 25-cluster Twitter network, which is slightly smaller than our results for the country match network (Table 4).

It is important to consider, however, that users do not necessarily tweet in the language that is dominant in their location. For this reason, we coded the language of messages from egos and alters who could be located at least at the level of a country to see whether egos and alters use the same language. Table 6 shows the most common languages used in the tweets. English is by far the dominant language, accounting for 73 percent of egos' messages — higher than the percentage of egos located in English-speaking countries. Portuguese is the only other language accounting for more than 10 percent. Japanese, Spanish, Indonesian and German each account for 1–10 percent, with all other languages being under 1 percent. Table 7 shows the most common combinations of languages between egos and alters. In 88 percent of the cases, the ego and the alter tweet in the same language — slightly higher than the 86 percent of ties that connect users located in the same country or countries with the same dominant language. Over three quarters of those (68 percent of all ties) are cases where both are using English, with slightly over one quarter being cases where both use a different language, most often Portuguese or Japanese. Cross-language ties are relatively rare.

Table 6 around here. ("Most common languages.")

Table 7 around here. ("Language combinations.")

The share of same-language ties in languages other than English is substantially higher for local ties (28 percent) and substantially lower for ties between clusters (14 percent). It falls even further if only international ties are considered (5 percent). The total share of same-language ties drops somewhat as well: from 92 percent for local ties, to 88 for ties between clusters, to 76 percent for international ties. This loss is made up almost exclusively by the share of ties in which an English-tweeting alter is followed by a non-English-tweeting ego.

Looking at the languages used by egos in each cluster or country, we found a somewhat imperfect match between the language used by individual users and the dominant language of the cluster. For example, while Portuguese is unambiguously the dominant language of Brazil, 13 percent of the tweets from users located in Brazil are in other languages, including 8 percent in English. An informal analysis of the profiles suggests that many of the English-twittering users located in Brazil are Brazilians rather than traveling English speakers.

8 We also constructed an alternative network based on the languages spoken in each clusters and the proximity between the languages in the hierarchical classification of languages (for example, assigning a higher degree of similarity to English – Dutch pair than to English – Japanese). We have found that the two language networks had a correlation of 0.95 and produced nearly identical results. For this reason we avoid the discussion of the alternative language metric, focusing just on the network produced from the Wikipedia dataset.

5.5 Multivariate Analysis

Having found that all four variables that we considered have an effect on Twitter ties, we used regression analysis to see whether their effects are independent. The results of the regressions are presented in Table 8. Comparing model 3 with models 1 and 2, we see that distance and being in the same country have independent and significant effects. A comparison of models 1, 4 and 5 shows that the same is true for distance and language. The effect of language is not significant when we control for country. Similarly, the effect of distance is no longer significant when we control for flights (model 8). In a model combining all four variables (model 9), only the effect of flights remains significant.

Table 8 around here. ("QAP Regressions.")

6. Conclusion

Looking at the network of ties in Twitter we find that distance and related variables (language, country, and the number of flights) all have an effect on Twitter ties despite the seeming ease with which long range ties can be formed. As a lightweight system that takes little effort to set up and can be used from either personal computers or mobile devices, Twitter offers a promise of transcending distance, connecting everyone with anyone. Our analysis shows, however, that distance considerably constrains ties. Two fifth of ties (39 percent) connect users within the same regional cluster, typically the size of a metropolitan area. All such ties are also domestic and connect users in the same linguistic area. Most of them fall within easy driving distance. Even for the remaining longer range ties between different clusters, distance matters. Ties at distances of up to 1,000 km are more frequent than what we would expect if the ties were formed randomly, while ties at longer than 5,000 km are underrepresented.

For such longer ties, distance, language differences, country boundaries, and ease of travel can vary independently, even as they remain strongly correlated. This warrants a comparison of such variables. We find that country and the frequency of flights have independent effects in pair-wise comparisons. The effect of language is no longer significant when country is included in the model. A closer look at language suggests that the language effect might is weakened by the wide use of English as a *lingua franca*. The effect of distance is no longer significant when the frequency of airline travel is included.

The number of airline flights is the best predictor of non-local Twitter ties. This likely reflects the role of air travel in facilitating long-distance face-to-face interaction, which in turn influences the formation of electronic ties. Air travel can also stand as a proxy for other kinds of pre-existing connections between places, which in turn influence formation of electronic ties. The finding highlights the importance of considering structural constraints on ties rather than simple physical distance.

7. References

Alderson, A., Beckfield, J., 2004. Power and position in the world city system, American Journal of Sociology 109, 811–851.

Barnett, G. A., Choi, Y., 1995. Physical distance and language as determinants of the international telecommunication network. International Political Science Review 16, 249–265.

Beaverstock, J.V., Smith, R.G., Taylor, P.J., 1999. A roster of world cities. Cities 16, 445–458.

Boase, J., 2008. Personal networks and the personal communication system: Using multiple media to connect. Information, Communication and Society 11, 4, 490–508.

Borgatti, S.P., Everett, M.G., Freeman, L.C. 2002. UCINet for Windows: Software for Social Network Analysis. Analytic Technologies, Cambridge, MA.

boyd, d., Ellison, N., 2007. Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13(1).

Butts, T.C., 2007. Permutation Models for Relational Data. Sociological Methodology 37, 257–281.

Butts, T.C., Acton, R., 2010. Spatial Modeling of Social Networks. In: Nyerges, T., Couclelis, H., McMaster, R. (Eds.), Sage Handbook of GIS and Society Research, forthcoming. Sage, Thousand Oaks, CA.

Cairncross, F., 1997. The Death of Distance: How the Communications Revolution Is Changing Our Lives. Harvard Business School Press, Cambridge, MA.

Cardwell, S., 2009. A Twitter Timeline of the Iran Election. Newsweek, http://www.newsweek.com/id/203953 (visited: 09-15-09).

Derudder, B., Witlox, F., 2005. An Appraisal of the Use of Airline Data in Assessing the World City Network: A Research Note on Data. Urban Studies 42, 2371–2388.

Fischer, C., 1982. To Dwell Among Friends. University of California Press, Berkeley, CA.

Gaudel, A., Peroni, C. 2010. Reciprocal Attention and Norm of Reciprocity in Blogging Networks. Jena Economic Papers #2010-020, Friedrich Schiller University Jena, Jena, Germany.

Golder, S., Yardi, S., 2010. Structural Predictors of Local Tie Formation in a Conversational Network. Unpublished manuscript.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78 (6): 1360–1380.

Herring, C.S., Paolillo, C.J., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, A.L., Clark, B., 2007. Language Networks on LiveJournal. In: Proceedings of the Hawaii International Conference on System Sciences. IEEE Press, Los Alamitos, CA.

Honeycutt, C., Herring, S. C. 2009. Beyond microblogging: Conversation and collaboration via Twitter. In: Proceedings of the Forty-Second Hawai'i International Conference on System Sciences. IEEE Press, Los Alamitos, CA.

Huberman, B., Romero, D., Wu, F., 2009. Social networks that matter: Twitter under the microscope. First Monday 14 (1).

Hutchinson, K.W., 2005. Linguistic Distance as a Determinant of Bilateral Trade. Southern Economic Journal 72, 1–15.

Java, A., Song, X., Finin, T., Tseng, B., 2007. Why We Twitter: Understanding the Microblogging Effect in User Intentions and Communities. In: Paper presented at the KDD Workshop on Web Mining and Web Usage Analysis (WebKDD), San Jose, CA.

Krackhardt, D., 1987. QAP Partialling as a Test of Spuriousness. Social Networks 9, 171–186.

Krackhardt, D., 1992. A Caveat on the Use of the Quadratic Assignment Procedure. Journal of QuaWive Anthropology 3, 279–296.

Krishnamurthy, B., Gill, P., Arlitt, M. 2008. A few chirps about Twitter. In: Proceedings of the first workshop on Online social networks, Seattle, WA, USA.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A. 2005. Geographic routing in social networks. PNAS, 102 (33), 11623–11628 .

Mok, D., Wellman, B., Carrasco, J.A., 2010. Does distance still matter in the age of the internet? Urban Studies 46(13), 2743-83.

Naaman, M., Boase, J., Lai, C., 2010. Is it Really About Me? Message Content in Social Awareness Streams. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10). ACM, New York, NY.

Schonfeld, E., 2009. Twitter Reaches 44.5 Million People Worldwide In June (comScore), TechCrunch. Available at http://www.techcrunch.com/2009/08/03/twitter-reaches-445-million-people-worldwide-in-june-comscore/

Smith, D., Timberlake, M., 1995. Conceptualizing and mapping the structure of the world systems city system. Urban

Studies 32, 287–302.

Sysomos, 2010. Exploring the Use of Twitter Around the World. Available at http://www.sysomos.com/insidetwitter/geography

Twitter, 2011. Twitter API Documentation. http://apiwiki.twitter.com/w/page/22554679/Twitter-API-Documentation. Accessed Feb. 15, 2011.

Wang, H., Wellman, B., 2010. Social Connectivity in America: Changes in Adult Friendship Network Size From 2002 to 2007. American Behavioral Scientist, 53, 1148-1169.

Webber, M., 1963. Order in Diversity: Community without Propinquity. In: Wingo, L. (Eds.), Cities and Space: The Future Use of Urban Land. Johns Hopkins Press, Baltimore, MD.

Wellman, B., 1979. The Community Question: The Intimate Networks of East Yorkers. American Journal of Sociology 84 (5), 1201–1231.

Wellman, B, Hogan, B. with Berg, K., Boase, J. Carrasco, J-A., Côté, R., Kayahara, J., Kennedy, T.L.M., Tran, P., 2006. Connected Lives: The project. In: Purcell, P. (Ed.), Networked Neighbourhoods: The Online Community in Context. Guildford, UK, Springer,. pp. 157–211

Wellman B., Leighton, B., 1979. Networks, Neighborhoods and Communities. Urban Affairs Quarterly 14, 363–90.

Wellman, B, Tindall, D., 1993. How Telephone Networks Connect Social Networks. Progress in Communication Science 12, 63–94.

Wellman, B, Carrington, P., Hall. A., 1988. In: Wellman, B. and Berkowitz, S.D. (Eds.), Social Structures: A Network Approach. Cambridge University Press, Cambridge, MA, pp. 130–84

Wimmer, A., Schiller, N.G., 2002. Methodological nationalism and beyond: nation-state building, migration and the social sciences. Global Networks 2 (4), 301–334.

Zook, M., Brunn, S.D., 2006. From podes to antipodes: positionalities and global airline geographies. Annals of the Association of America Geographers 96 (3), 471–490.

8. Tables

Table 1: Sample size at different steps.

Step	Sample size
Initial collection	481,248 tweets (7 weeks * 20 messages every 25 seconds)
Subsampling	3360 egos (20 per 1 hour period)
Geocoding at the level of country or smaller	2852 egos (including 2167 egos with precision of < 25,000 km²)
Picking alters	2423 dyads
Geocoding the alters at the level of country or smaller	1953 dyads
Picking pairs where both accounts have a location with precision of < 25,000 km²	1259 dyads
Spatial clustering	386 clusters
Selecting the top 25 clusters for regression analysis	25 clusters

Table 2: Location precision in the sample of egos

Location specification	Examples	Share of the sample	Used for:
Latitude and longitude	ÜT: 34.246769,-118.394672 iPhone: 35. 498447,-97.477180	7.5%	cluster-level and country-level analysis
A named location with an area of <25,000 km²	Los Angeles, CA Floss Angeles (LA) São Jorge do Ivaí – PR indonesia-bali-depok south wales in the uk Ραφήνα 六本木とか千葉とか [“Some place like Roppongi or Chiba”] モスクワ [“Moscow”]	57.0%	cluster-level and country-level analysis
Not specific, but enough to identify a country	U.S.A. indonesia body=midwest, mind=elsewhere	20.4%	country-level analysis
Very broad, non-spatial, humorous, or undecipherable	where trouble is forks 127.0.0.1 Hogwarts	15.10%	not used

Percentages are based on the 3360 egos, sub-sampled from the original sample.

Table 3: Top clusters

rank	cluster	share of egos (%)	share of egos (%) for egos included in dyads	share of alters (%)	locality
1	New York	8.5	8.3	10.2	54.3
2	Los Angeles	5.1	5.6	10.4	53.3
3	Tokyo	4.1	4.8	5.0	62.9
4	London	3.6	3.3	4.9	48.8
5	São Paulo	3.5	3.0	3.6	78.4
6	San Francisco	2.8	2.7	4.1	41.2
7	New Jersey	2.5	2.8	2.1	20.0
8	Chicago	2.2	2.0	1.7	32.0
9	Washington	2.1	2.8	2.6	34.3
10	Manchester	1.9	2.0	1.1	30.8
11	Atlanta	1.7	2.1	2.1	46.2
12	San Diego	1.5	1.5	1.1	26.3
13	Toronto	1.3	1.1	1.5	42.9
14	Seattle	1.3	1.4	1.2	58.8
15	Houston	1.2	1.2	1.0	40
16	Dallas	1.2	1.0	1.4	61.5
17	Rio de Janeiro	1.2	1.0	1.1	30.8
18	Boston	1.2	1.2	1.1	20.0
19	Amsterdam	1.1	1.1	0.9	50.0
20	Jakarta	1.1	0.6	0.3	42.9
21	Austin	1.0	1.0	1.3	50.0
22	Sydney	0.9	1.0	0.8	38.5
23	Orlando	0.9	1.0	0.6	16.7
24	Phoenix	0.8	0.7	0.6	11.1
25	Osaka	0.8	1.0	1.0	25.0

Notes:

1. Cluster name is based on the label most commonly used for locations assigned to the cluster.
2. Ego percentages are based on the 2167 egos located with precision of < 25,000 km².
3. The fourth column shows percentages out of the 1259 egos included in dyads with both parties located with precision of < 25,000 km².
4. Alter percentages are based on the 1259 alters included in dyads with both parties located with precision of < 25,000 km².
5. Locality is defined as the share of local of ties among all ties for egos in a cluster.
6. The cluster labeled “New Jersey” is centered between Philadelphia and Trenton, NJ and includes all locations identified as just “New Jersey.”

Table 4: QAP correlations, top 25 clusters

	twitter	flight	language	domestic
distance	-0.448	-0.817	-0.617	-0.720
domestic	0.440	0.723	0.709	
language	0.418	0.637		
flights	0.510			

Notes:

1. Distance, the number of twitter ties and the number of flights are logged
2. All p values are <= 0.005

Table 5: Top countries

	share of egos (%)	share of egos (%) for egos included in dyads	share of alters (%)	percentage of domestic ties	percentage of domestic ties among non-local ties	following foreign alters / being followed from abroad	country named explicitly (% of egos)
USA	48.5	45.7	54.5	91.6	89.3	0.3	8.1
Brazil	10.6	12.1	10.5	83.5	72.5	4.9	55.4
UK	7.6	8.3	7.6	50.6	33.3	1.2	45.3
Japan	5.5	6.5	6.3	92.1	86.0	1.4	25.0
Canada	3.7	3.8	2.9	33.3	23.1	1.6	58.5
Australia	2.7	2.7	1.9	50.0	32.0	2.2	69.7
Indonesia	2.6	1.8	1.2	60.0	25.0	7.0	83.3
Germany	2.1	1.8	1.3	62.9	58.8	3.2	58.6
Netherlands	1.4	1.4	1.2	66.7	22.2	1.5	54.3
Mexico	1.2	1.3	0.7	44.0	8.3	7.0	56.7

Notes:

1. Ego percentages are based on the 2852 egos located at the level of country or better.
2. Column 2 is based on the egos included in 1953 dyads with both parties located at the level of country or better.
3. Alter percentages are based on the 1953 alters located at the level of country or better.
4. Percentage of domestic ties refers to the number of ties with the ego and the alter in the given country as a share of all ties for egos in that country.

Table 6: QAP Regressions

	Models								
	1	2	3	4	5	6	7	8	9
intercept	2.35	0.14	1.45	-0.05	1.50	0.01	-0.01	0.45	0.07
distance	-0.240*** (-0.448)		-0.146** (-0.272)		-0.164*** (-0.307)			-0.049 (-0.092)	-0.014 (-0.027)
domestic		0.522*** (0.440)	0.290* (0.244)			0.342** (0.288)			0.093 (0.079)
language				0.600*** (0.418)	0.329* (0.229)	0.307 (0.214)			0.171 (0.119)
flights							0.118*** (0.510)	0.101** (0.435)	0.082* (0.356)
R^2	0.201	0.193	0.229	0.175	0.233	0.216	0.261	0.263	0.278
Adj. R^2	0.201	0.193	0.228	0.175	0.232	0.215	0.261	0.262	0.274
Number of observations	600 combinations of 25 nodes								

Significance level: * = 5%, ** = 1%, *** = 0.1%
Standardized coefficients are shown in parentheses.

Table 7: Most common languages

Language	% of egos
English	72.5
Portuguese	10.1
Japanese	5.4
Spanish	3.1
Indonesian	1.8
German	1.7
Dutch	1.0
Chinese	0.9
Korean	0.4
Swedish	0.4
Russian	0.4

Based on 2852 egos.

Table 8: Language combinations

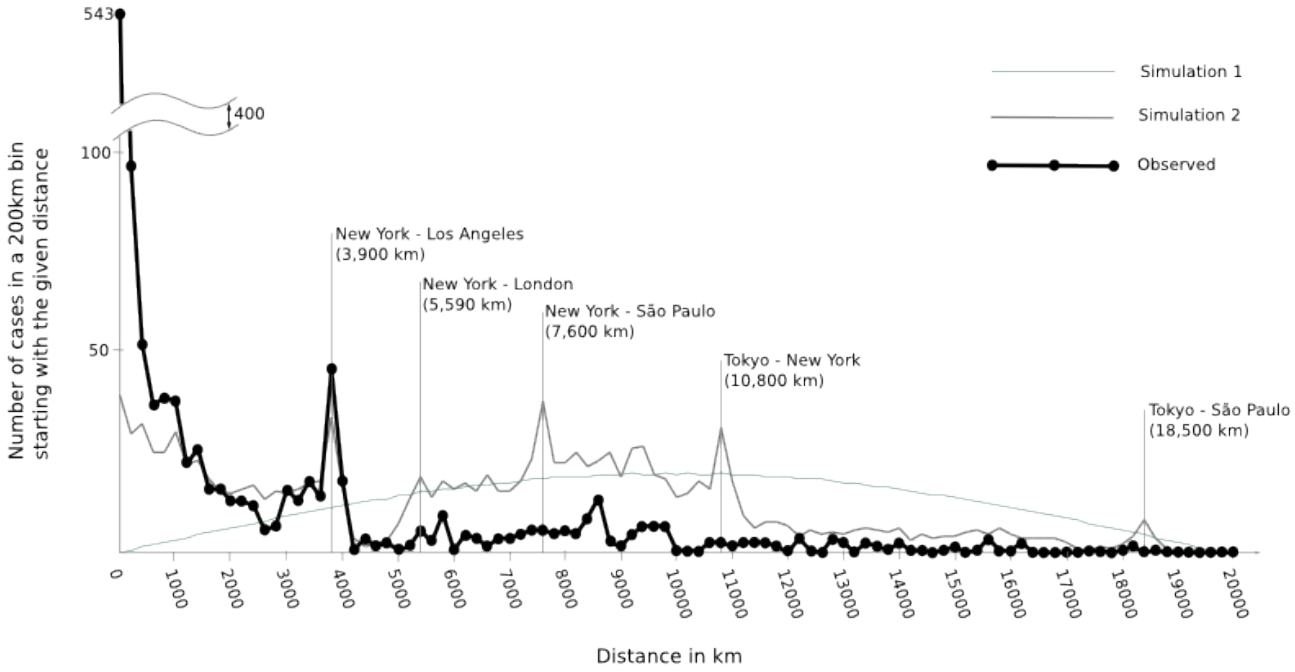
Language combinations	as a percentage of...			
	all ties	in-cluster ties	x-cluster ties	int'l ties
Same language (total)	88.4	91.6	88.4	75.5
English-English	67.5	63.3	74.4	70.1
Same language, non-English	20.9	28.3	14.0	5.4
Cross-language				
Other-English	7.4	2.6	8.5	20.5
English-Other	3.1	4.5	2.1	3.1
Different languages where neither is English*	1.1	1.2	1.0	0.9

Based on 1768 dyads in which both ego's and alter's language was identified.
* The most common combinations were Japanese-Chinese, Spanish-Italian, and Portuguese-Spanish.

9. Figures

Figure 1: Histogram of physical distances between egos and alters

The graph shows the number of ties by distance, in 200 km bins. (For example, New York – London ties, at 5,590 km, are counted towards the bin 5,400 km bin.) The total number of ties in each of the two simulations is the same as in the observed data.



Based on 1259 dyads.

Methodological Appendix

This Methodological Appendix includes some of the points that we found too detailed for the paper itself but that we would like to share with the readers. The sections numbers refer the section numbers in the paper.

4.1 Collecting the Sampling Egos

We do not know the exact number of messages sent in the period during which we collected our original sample. We believe, however, that it likely numbered in the range of 100 million. A graph released by Twitter in February of 2010 suggests there may have been between 10 and 15 million Twitter messages posted per day in August of 2009. Our analysis of message IDs suggests 150 million as the upper limit.

Access to the full stream of public messages (known as "the firehose") has been reserved for specific corporate partners of Twitter, such as Google and Bing. All other methods produce a selection of messages. At the time of our data collection, none of those methods were guaranteed to produce a representative sample.

4.2 Geocoding and Subsampling

We assume in our paper that locations provided by the users represent their actual locations. We realize that it is possible that some of the users mischievously mis-identified their location. The most likely effect of such misidentification, however, is that it would introduce noise and *weaken* the effects that we observe. In other words, we can expect that Twitter ties are in reality *at least* as localized as we find when taking users' self-reported locations at face-value. (Perhaps the most likely case of deliberate misidentification of location would be the well-publicized attempt by some American users in June of 2009 to confuse Iranian police by changing their location to "Tehran." In our sample, however, we found that only about 0.5 percent of egos specified Tehran as their location.)

We note that a number of users provide their precise coordinates via iPhone or ÜberTwitter (a BlackBerry application). Analyzing the messages just from such accounts would have made it possible to easily identify the location of a very large sample of users, while also being certain that the geo-coordinates represent the users' actual locations at the time of posting. Users of BlackBerries and iPhones, however, can be expected to be distributed quite differently from other Twitter users and to have a different pattern of ties. We later found, for example, that they are much more likely to be located in the United States and that their ties tend to be more local. We therefore decided to avoid limiting ourselves to such users.

At the time the data was collected Twitter did not yet offer an API for attaching location to individual messages. (The Twitter application for iPhone and ÜberTwitter both specified user's location by changing user's profile.) More recently, users have gained more options for posting their locations. In particular, users of the most recent browsers can now allow Twitter to identify their location from metadata sent by the browser. A look at Twitter messages posted in June of 2010, however, showed that only a small minority of users make use of such features.

4.3 Sampling the Alters

The main reason for sampling alters in a separate step was the extreme sparseness of our original sample of egos. Only about 2 percent of the egos' "following" ties pointed back to users already included in our sample of egos, and the overwhelming majority (89 percent) of the egos were not connected to any of the users in our original sample. This was due to the fact that our sample of egos represented just a tiny fraction of all users.

When picking alters, we avoided doing this for egos following more than 500 accounts. Our reasons for doing this included a concern that ties of users who followed very large number of accounts did not represent meaningful relationships, but were either a result of Twitter "spam" (following other users at random in the hope that they would look at the account that follows them and visit their links) or a way for highly popular accounts to acknowledge their fans by reciprocating the following. (E.g., Barack Obama's account currently "follows" 716,453 Twitter accounts.) Ties created by spammers would add random noise to our data, while the reciprocal following by highly popular accounts would essentially invert the direction of ties. The choice of 500 as the cut-off was influenced by the fact that this number had been cited as unrealistic by many observers, including a post on Twitter's blog noted in 2008 that said that "[m]ost users may have a hard time finding 500 accounts they are interested in" (http://blog.twitter.com/2008/08/making-progress-on-spam.html). We note, though, that Twitter spam was not nearly as prevalent at the time when we collected our data than it is today.

When picking each alter, we immediately checked if the alter had provided a location description in their profile. In cases when the alter had provided no description at all (about 25 percent of the time), we drew a replacement, until we found for each ego an alter with a location.

For all pairs, we recorded whether the relationship was mutual, that is, whether the alter also followed the ego. We found this to be the case in 41 percent of the pairs. During our analysis, we found that mutual ties were more likely to be local and domestic, as one might expect. (Mutual following involves a degree of reciprocity and is therefore more likely to represent a somewhat stronger tie.) To simplify the presentation, however, we do not discuss our comparison of mutual and non-mutual ties.

4.4 Aggregating Nearby Locations

The average distance between an observation and the geographic center of the cluster to which it was assigned was 22 km, with a standard deviation of 28 km and the maximum distance of 172 km. Less than 3 percent of the observations were assigned to a cluster centered more than 100 km away.

5.4 Language

When coding the language of individual messages we used Google's language classification API and then checking and correcting the results manually. The manual correction was done without looking at the geographic coding, relying solely on the message text. The author who performed the manual checking was sufficiently familiar with the twelve most common languages in our sample to confidently tell them apart from each other. Those twelve languages accounted for 98.7 percent of our sample. The only cases where we had doubts involved pairs of languages where at least one language was rarely present in our sample, e.g., Indonesian vs. Malay (1.75 percent and 0.07 percent), Swedish and Norwegian (0.35 percent and 0.21 percent), as well as some even less common languages that jointly accounted for 0.81 percent of our sample. If the account contributed multiple tweets, we used the first tweet in our sample for coding. In cases where a message mixed multiple languages, we coded what appeared to be the main language of the message.