Abstract

We use a sample of publicly available data on Twitter to study networks of mostly weak asymmetric ties. We show that a substantial share of ties lie within the same metropolitan region. As we examine ties between regional clusters, we find that distance, national borders and the difference in languages all affect the pattern of ties. However, Twitter connections show the more substantial correlation with the network of airline flights, highlighting the importance of looking not just at distance but at pre-existing ties between places.

1. Introduction

Social contacts benefit from physical proximity. This fact of social life is so basic that for a long time proximity was often simply taken for granted. Social interaction was generally understood to mean face-to-face interaction, for which distance acts as a powerful barrier. In other words, the fact that being near each other facilitates the formation of social ties for the most part was not so much a finding of social research, but its basic assumption. With a few odd exceptions, communities could be safely presumed to be local (Keller 1963). Social network analysts were among the first to argue that community should not be assumed to be local (Webber 1968; Mitchell, 1969; Tilly 1974; Fischer et al., 1977; Fischer 1982; Wellman 1979; Wellman 1979; Wellman and Leighton 1979). They showed that the social network approach afforded the possibility of following social ties as they cross space, thus mapping the more distributed communities that were replacing the ones based on neighbourhoods.

The advent of the Internet created even greater possibilities for maintaining useful social ties over long distances, as well as greater awareness of such possibilities. Pundits proclaimed that distance was dead (Cairncross, 1997) and that the world was now “flat” (T. Friedman, 2005), assuming that low-cost, instantaneous, content-rich Internet communication was eliminating the need for proximity in maintaining contact. Many recent studies have shown that while social ties can operate over distance, proximity does make a difference (e.g., Butts, 2009). However, most such studies have looked at e-mail, which was shown to help extend and maintain existing strong ties (e.g., Wellman & Hogan, et al., 2006). In the recent years, new relational means have developed on the Internet, some of which seem less tied to strong ties or face-to-face-contact. Are those new forms of electronic interaction also affected by proximity?

We focus on one such Internet-based system, Twitter, a popular social networking and microblogging 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. (Twitter users can make their tweets private, but most do not.) While visitors can access each user’s Twitter page directly to read their tweets, the preferred method of using Twitter is to identify a set of users that you want to “follow.” Once you select accounts for following, you see recent tweets from those accounts whenever

§ The authors thank Lilia Makarova and MinKyu Kim for their help in preparation of this paper.
1 We have found, for example, that only 10 percent of the users followed by our original sample of “egos” protect their tweets.
you log on to Twitter. A users’ choice of whom to follow is public. Additionally, Twitter allows users to specify their location in their profile, and most of them do so. Twitter thus offers us a publicly available, spatially embedded network dataset, a rare luxury in network analysis (Butts and Acton, 2010).

In the next section we discuss two features of Twitter that make it an interesting case for analysis: its global popularity and the low cost of ties. We then look at some of the ways distance can affect tie formation, considering both the actual physical distance, as well as several closely related variables, such as ease of travel, national borders and differences in language. We then explain how we built our sample of Twitter ties and present an analysis of how distance and related variables affect tie formation in our sample.

Our analysis shows that distance matters on Twitter, both at short and longer ranges. We find that 39 percent of the ties are shorter than 100 km and that ties up to about 1000 km are over-represented. This result is interesting, considering the ease with which long-distance Twitter connections can be formed. After a closer investigation of other variables that can impede or facilitate ties while being closely intertwined with distance, we find that national boundaries and 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. The strength of prior ties between places matters more 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 international reach and popularity. In June 2009 alone, Twitter attracted more than 44.5 million unique visitors, according to comScore.com (Schonfeld, 2009). Surveys conducted in the United States, Brazil and other countries found a substantial extent of Twitter use (PEW 2009; Ribeiro 2009). Our own data suggests that over half of Twitter users were located outside the United States at the time when we collected our data, including many users in Brazil, the UK, Japan, Australia and Indonesia. This wide distribution of Twitter users makes it possible to explore the effects of distance at different scales, from fairly short to nearly antipodal.

The second aspect of Twitter that makes it an interesting case for analysis 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 interaction fails Granovetter’s definition on all counts. 140 character messages take little time to read and yet less time to ignore. (A user who follows a lot of accounts might spend substantial time on Twitter overall, but would not be spending much time per followed account.) The fact that Twitter messages are mostly public and are broadcast rather than directed at specific users reduces the level of intimacy and emotional intensity of such communication. Finally, Twitter ties are quite often asymmetric: if A follows B, B does not have to follow A. In our sample, 60 percent of the ties are unidirectional. In other words, the majority of Twitter ties are so weak that the followed users do not even bother to reciprocate the followers’ interest in their tweets, despite the rather low cost of doing so.

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.
Twitter also differs from other “social networking” sites, such as LinkedIn or Facebook, which often aim to capture pre-existing ties. (See boyd and Ellison 2007 for an overview of social-networking systems.) Twitter’s design encourages following strangers through the two aspects mentioned earlier: the public nature of communication and the lack of enforcement of bi-directionality. Lack of bi-directionality means that users can follow any of the public accounts. The user who is followed can actively “block” some of their followers, but they do not need to approve them. Followers therefore do not face the risk of rejection that is often inherent in initiating a contact with a stranger. (Since non-reciprocity of following is the norm, lack of reciprocal following would not imply a rejection, unless the users already have a prior connection.) The followers also do not need to worry about imposing a burden on a person they do not know (or do not know well), since their decision to follow merely indicates an interest in reading on a somewhat more regular basis messages that are already accessible to anyone with an Internet connection. Huberman et al. (2009) report a quantitative investigation on the weakness of Twitter ties, concluding that only a small share of such ties imply any meaningful interaction.

This easy formation of ties makes Twitter accounts similar to blogs — a term “micro-blogging” is often used. However, unlike the weak ties between bloggers and their readers, which most often stay invisible, Twitter ties can be easily observed and analyzed. 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.

The combination of weak, low cost ties and global popularity creates an opportunity for people to make links that transcend distance 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, not only to coordinate local protests against the national regime but also to inform the world about these protests (Cardwell, 2009). It is important, however, to avoid overstating the extent to which Twitter networks transcend space and political structures. This paper provides a quantitative investigation of the effect of distance on Twitter ties, considering several dimensions of distance that we discuss in the next section.

3. Ties, Distance and Related Variables

Distance has been shown to have an effect on social ties, including those based on electronic communication. One can therefore expect that it would affect formation of Twitter ties as well. 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 act on social ties directly. Even in its purest form, distance usually impedes formation of social ties by raising the cost of travel that is required for face-to-face interaction. Therefore, the effect of distance depends on the extent to which

---

3 Note that blocking a user cannot actually prevent them from reading the messages if the blocking account is public. An informal poll conducted by Twitercism.com (http://twittercism.com/poll-why-block/) suggests that blocking is reserved primarily for spammers (81 percent), accounts found to be offensive (38 percent), or those who aim to initiate further (and undesirable) contact with the followed accounts (“They kept bothering me,” 13 percent).

4 Liben-Nowell et al. 2005 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.

5 Many also changed their profiles to display their location as “Tehran,” to both express their solidarity with the protesters and to disrupt the ability of Iran’s police to identify those users of the system who were actually Iranian. Luckily for our study, it appears that they have changed their location back by August — we found only a relatively small number of users claiming to be in Tehran.
greater distance translates into higher cost of travel, which depends on the existing transportation infrastructure. Distance is also intertwined with many other factors, including national boundaries and differences in language. 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 proximity between the users, the frequency of flights between their cities, whether the users are in the same country, and the match in language.

3.1 Physical Proximity

In the pre-Internet days, Wellman’s second study of East Yorkers (a former borough of Toronto) found that only 22 percent of their socially-close friends and relatives were in East York and no East Yorkers had most of their active ties living within a mile’s walking distance (Wellman et al., 1988). The East Yorker’s longer ties were enabled by the growing use of cars and phones, which allowed people to coordinate meetings by phone and the drive to them. (See also Ito, Matsuda & Daisuke, 2005; McEwen, 2009.) East Yorker’s worlds were expanding, but were not indifferent to face-to-face interactions. Rather, social contacts were now primarily bounded by the scale of metropolitan area, reflecting increased mobility and flat-rate metropolitan phone calls (Wellman & Tindall, 1993). While the telephone was important, its use was complimentary rather than substitutive with in-person contact. While East Yorkers had substantially more distant ties, they rarely had frequent telephone contact with people whom they did not also often see in person (Wellman, 1979; Wellman and Tindall, 1993).

A newer study of East Yorkers conducted in 2005 discovered again that the number of social ties drops sharply as the distance increases between 1 and 20 miles (Mok et al., 2010). While modern East Yorkers maintain yet more distant connections, some reaching as far as Europe or Pakistan, their phone use is lower for distant connections. Email — a new medium not available to the East Yorkers of the 1980s — is less sensitive to distance, in the sense that the amount of email contact per tie is roughly the same regardless of the contact’s location. However, since East Yorkers’ ties are local, the overwhelming majority of their email is local as well. Other studies have similarly found that most email use is local, and that there is lower overall contact with strong ties who live further away (e.g., Boase, 2008). Studies of “friendship” ties on LiveJournal (e.g., Liben-Nowell et al., 2005) have similarly found an effect of distance.

We can expect that Twitter ties would also be influenced by geographic proximity. Like other forms of electronic communication, Twitter ties may be, to some extent, complementary to face-to-face interaction. Distance can also influence ties through users’ local interests. For instance, users may follow minor local celebrities or local people whom they hope to meet one day. At the same time, the low cost of Twitter ties can allow for longer connections with distant strangers. The essence of globalization is that events in one place are increasingly affected by occurrences elsewhere. We can expect that people would take interest in such events, and the substantial increase in the consumption of transnational media (e.g., Bruck et al., 2004) suggests that they in fact do. Ties created through cosmopolitan interests may be quite weak. This, however, can be their strength. Unlike strong ties that often require nurturing through in-person contact, with communication technologies serving as only an imperfect substitute, weak ties can traverse large distances, taking full advantage of the opportunities offered by new technologies.

In particular, it becomes important to ask whether distance matters at different scales. It is quite possible, for example, that distance plays an important role at short range, but makes no difference past a certain threshold.
3.2 Air Travel

Distance, however, does not affect social interaction directly. Perhaps the most immediate effect of distance is that it limits the opportunities for face-to-face interactions, which are important for social ties. People who live in different places are less likely to meet in the first place. They would be a lot less likely, for example, to take up jobs at the same company or to meet at a bar. In cases where they do meet (e.g., while traveling or due to moving), lack of face-to-face interaction can lead to a decline in the activeness of other ties, as for example found by Mok et al. (2010). The extent to which distance translates into a reduced likelihood of face-to-face interaction can vary, however. New York is equally far from London and from the town of Eirunepé in the Brazilian state of Amazonas. Getting from New York to London, however, is much much simpler than getting to Eirunepé.

For intercity ties, an important component of ease of travel is the availability of airline flights. Numerous flights connect New York and London, while a flight from New York to Eirunepé would require many stops. We must therefore ask whether the frequency of airline flights may be a better prediction of non-local ties than the physical proximity. (Note that frequency of flights is strongly correlated with proximity and may to some extent serve as a proxy for proximity the the shorter range, while being a better predictor ties between places that are further apart.)

Frequency of airline connections can also be interpreted as a proxy for the more general connectedness. In the literature on global cities, for example, airline data has become one of the popular ways of mapping the global structure of the city network. Such data has shown that cities that are most central in the network of airline connections are also important in the network of relationships between transnational accounting firms (Beaverstock et al., 1999), being also the same cities that were deemed important in the earlier theoretical literature. Measures such as the frequency of airline connections can thus be interpreted not just as a factor that reduces travel time (thus facilitating the formation of social ties through in-person contact), but also as a measure of broader economic connectivity between the world’s cities.

3.3 National Boundaries

Today’s world is organized as a system of nation-states (Meyer et al. 1997), that is, units that tie together territory, political power and identity. Nation-states are states, in the sense that they control a well-defined territory (which is usually contiguous), and their right to control that territory is recognized by other states. Unlike other historical forms of states, however, a nation-state is associated with a “nation” — a culturally defined group of people that it claims to represent. Citizens of nation-states do not merely submit to the authority of the state. Rather, they typically identify with the nation from which the state draws its legitimacy, that is, they see themselves as members of the nation and often understand this link as primordial (Cornell & Hartmann, 2006).

National boundaries affect long-distance 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 then. In addition to impeding travel of people and things, national boundaries often contain communities of interest. Some of the most important decisions affecting people’s lives are made within the countries in which they live. Consequently, people living within the borders of a single nation-state would have a higher degree of interest in such events than those living outside. Additionally, local events in one place sometimes have significance for other places in the country, but less so across the border. For example, residents of New York may have good reason to follow the San Francisco court case about the legality of gay marriage in California, since the decision may eventually affect people
throughout the United States. Residents of Toronto, just a few hundred kilometers away, however, may have a lot less interest in the issue, since decision of US courts have limited bearing on Canada. It is important, however, to avoid “methodological nationalism” (Wimmer & Schiller, 2002) or “implicit state-centrism” (Derudder & Witlox, 2005): that is, 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.

While for the purpose of this paper we expect national boundaries to limit Twitter ties, it is important to note that in the longer term, nations can themselves be affected by communication technologies. Anderson (1983/1991) links the emergence of national identity with the emergence of the printing press. Such technologies, he argues, enabled communication between people inhabiting the domain of the same ruler and speaking related languages (the many different “Frenches” and “Englishes”), while simultaneously setting them apart from those living under different monarchs and speaking more distant languages. They could increasingly imagine themselves as forming a single ethno-national community, rather than associating just with their narrow locality, clan, or status group. Today, modern communication technologies can be undoing modern nations by facilitating interaction over longer distances, thus helping the emergence of “global imaginaries” that contend with national understandings of the world (Steger, 2008). In this sense, Twitter ties that cross national boundaries help undermine those boundaries.

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 the countries (Frank, 1969; Smith & Timberlake, 1995b). People who live in large and powerful nations may have more opportunities for domestic connections, while also having less interest in foreign events. Resident of the smaller and less powerful nations, on the other hand may have a lot more interest in what happens abroad, since their lives are quite often affected by foreign events. They may also simply have fewer options for domestic ties. (Small but powerful nations and large powerless ones could fall somewhere in-between.)

3.4 Language

Social interaction often depends quite crucially on the two parties’ ability to communicate, which nearly always requires that they have competency in the same language or rely on a bilingual mediator. Consequently, language boundaries can structure social interactions in a variety of contexts. For example, the degree of dissimilarity the dominant languages of different countries affects international trade (Hutchinson, 2005). One would expect shared language competency to be even more important for use of ICTs, which are often used for practices that involve language as a sine qua non. Barnett and Choi (1995) report, for example, that language similarity explained 28% of the structure of the global tele-communication network in the 1980s. The use of Twitter, a purely textual medium, would similarly require at least some degree of competence in the language of the tweets.

Like national boundaries, linguistic differences are intertwined with distance. People living in the same place typically share a language (or several). Because of ancient settlement patterns, more recent patterns of colonization and today’s national boundaries, language similarity is also affected by distance at longer ranges. An individual located in New York, who is likely to speak English, is surrounded by a circle of around 3,000 km within which most people can be expected to speak English as well. Looking past 3,000 km and up to about 8,000, such an individual would encounter locations where people speak related languages, such as Spanish, Portuguese or German. The majority of people who live yet further away, speak languages that are quite dissimilar to English, for example Chinese and Japanese. The effects of language and physical distance can thus be closely intertwined.
On the other hand, it is important to note that 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 speak English in addition to their local and national languages. (See for example, Herring et al, 2007, on the role of English in LiveJournal networks.) The informal and asynchronous nature of Twitter can somewhat reduce the required degree of competence: reading (or even writing) 140 character messages can be less challenging than interacting by phone or email and it helps that one can simply ignore messages one does not understand, since no reply is usually expected. It is therefore important to consider not just the similarity of languages between places where the users are located, but also the actual language employed by the users.

4. Building a Sample of Twitter Ties

The primary data source used in this article is a sample of pairs of geocoded Twitter accounts connected by a “follow” relation. To assemble this set of pairs, 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.

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 2010). We queried public timeline every 25 seconds, using a Python script, for a period of seven days in August of 2009. To ensure that we would only collect tweets available to the public, the public timeline was retrieved without logging into Twitter. We collected a total of 481,248 messages. We do not know the precise total number of messages posted during the analyzed period, but we believe the number was around 100 million.

The tweets included in the public timeline do not necessarily represent a random sample of public tweets sent in the corresponding time period, since we do not know exactly what method Twitter uses for selecting tweets that go into the public timeline. Furthermore, we have received reports (citations to personal communication removed for review) that the mix of messages may vary somewhat depending on the account used to retrieve the public timeline. 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), due to the fact that other public methods require a substantially higher investment in hardware but similarly provide a sample of messages that could not be guaranteed to be random. (Twitter provides access to the full stream of messages — known as “the firehose” — only to specific partners, such as Google and Bing. At the time of our data collection, all other methods of obtaining Twitter messages provided a subset that, according to Twitter documentation, could not be guaranteed to be statistically representative.)

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 rate at which Twitter messages are produced varies through 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

6 According to Twitter documentation, the messages included in the public timeline are cached for 60 seconds, but we found that requests sent at 25 seconds intervals returned a different set of messages every time.

7 A graph released by Twitter in February of 2010 suggests that Twitter 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.
consequently under-sample the users who tweet on a New York schedule (which would likely include most 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 egos connections were sampled in a separate step.)

4.2 Geocoding and Subsampling

Most (75 percent) of the messages in this large sample had some location value associated with them. In other words, they were sent by users who either specified location in their Twitter profile or, in the minority of cases, used a Twitter client that automatically updated the location field in their profile. Our analysis of a sample of those location descriptions showed that most of such descriptions (about 85 percent) referred to a real place, at the level of precision anywhere between a country name and exact coordinates. For the purpose of our analysis we assume that such descriptions represent user’s actual location, either the place where the user tends to be in general or where they were at the time the message was posted, discounting the possibility that the users mischievously mis-identify their location. 8 We were careful, however, to properly classify location descriptions that suggest that they are wishful (“America i wish, England :()”) or outdated (“From Dallas but live in ATL”). We note, additionally, that to the extent that users may misidentify their location, doing so would introduce noise that would weaken the effects that we find. (In other words, we can expect that Twitter ties are in reality at least as localized as we would find if we take users’ self-reported locations at face-value.) A small number of users (less than 2 percent) specified multiple locations (e.g., “kuala lumpur/jakarta”) or somehow identified that the identified location is not the only one (e.g., “Iowa City, IA and global”), though in many cases such multiple locations were adjacent (“La Verne/Rancho Cucamonga, CA”). In all such cases we identified the user with the first location mentioned.

In 6 percent of the cases we found exact coordinates. Those 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 UberTwitter, a popular Twitter application for Blackberry. Location values set by geo-aware devices have the attraction of precision and ease of parsing. Selecting just the 6% of the users who specified their locations as coordinates, we could easily collect and geocode a very large sample of users, while also being certain that the geo-coordinates represent actual locations of where the users was when the message was posted. Users of Blackberries and iPhones, however, can be expected to be distributed quite differently from Twitter users in general. (For example, we found that while the United States accounted for only 49 percent of egos that we geocoded, its share among users who specified their locations in the form of coordinates was 72 percent.) Perhaps more importantly, such users can be expected to have a different pattern of ties. (In our sample, users whose locations were specified as coordinates had somewhat shorter ties, with a higher percentage of local and domestic ties.) Ties in which both parties specify their locations as coordinates can be even more atypical. For this reason we rely on locations specified either as coordinates or as textual descriptions. 9

8 Perhaps the most likely case of deliberate misidentification of location would be the well-publicized attempt by some American users to confuse Iranian police efforts to crack down on protests in Tehran by changing their location to “Tehran.” In our sample, however, we found that only about 0.5 percent of the egos identified their location as being in Tehran. Given the small number of such accounts, we believe it does not matter whether they represent users who are actually in Tehran or their American supporters.

9 At the time the data was collected Twitter did not yet offer an API for attaching location to individual messages. 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 feature. Again, we can expect this small minority to be different from the overall Twitter population.
While most of the provided location descriptions could eventually be mapped onto a rather specific location, we found that the users employed a wide range of conventions for describing where they are. While many of the locations in Los Angeles were identified as “Los Angeles, CA” and some even as “Los Angeles, CA, USA,” users also provided descriptions such as “LA,” “L.A.,” “Floss Angeles,” “Floss Town (LA), CA,” or “LosAngeles.” Locations outside the United States added yet more variation: table 1 shows a sample of the different ways in which the users have identified their location as being within the Greater Tokyo area in Japan, demonstrating, among other issues, the lack of word boundaries in written Japanese. We have found that automatic geocoding of such data was quite error prone, resulting in false positives or failure to correctly locate proper place descriptions. More importantly, since automatic coding performed worse for locations outside the United States (such as those shown in table 1), it introduced a substantial risk of a geographic bias. 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.

| Table 1 around here. (“Location descriptions corresponding to Greater Tokyo.”) |

The need for manual processing made it impossible to geocode location 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 the sample based on an earlier hour, we drew a replacement, resulting in a sample of 3,360 unique users. We refer to those users as egos.

We coded the location of the 3,360 egos by trying to identify the city and the country in the location description provided by each user. (In the case of the USA, Canada, Brazil and several other countries we also identified the state or province.) The locations were coded by hand, considering only the location value specified in the user profile and not any other information about the user or their posts. In particular, locations specified in foreign languages (e.g., Japanese) were coded based on the specified place rather than on the language in which the place was specified. (Some of the profiles used Japanese to refer to locations outside of Japan, while many others used English names for places in Japan.)

The locations specified by the users varied in their precision. Some provided exact geographic coordinates (see above) or addresses. Others (the majority) provided a name of the city (“Caxias do Sul”), a part of a city (“Bronx, Where da cute peeps @”), a metropolitan area (“DC Metro Area”), or a county (“UK, Bedfordshire”). Some identified just a state or province (“Ontario, Canada” or “四川”), a part of a country (“Midwest”) or a just a country (“USA”). Yet others (a small minority) referred to whole continents, or the planet as whole.

We summarize the distribution of location precision in table 2. We considered the location to be specific if it named a place with an area of up to 25,000 km², choosing this value since it 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.) This area also roughly corresponds to the upper limit on commuting distance. In most cases, the larger of the specific place names represented cities or metropolitan agglomerations. For consistency, however, we applied the same criteria to all named places (states, provinces, countries) that satisfied this area requirement. For example, locations
specified as “Israel”, “Wales”, “Malta”, or “Northern Ireland” were considered specific. Such specific locations accounted for 65% of the coded sample, which included the 7.5% who provided exact geographic coordinates. An additional 20% of the cases had locations that were sufficiently narrow to determine the country. We use those cases for our analysis of Twitter use by country. Finally, the remaining 15% of the cases were non-spatial, involving very broad, fictional, metaphorical, or ambiguous locations (“Global”, “3rd planet”, “The Moon”, “Hogwarts”), values that did not contain locations at all (e.g., “heading for promotion”) and a small number of values that we could not make decipher.

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. 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. For this reason, we did not attempt to analyze ties between sampled egos, and instead sampled an additional user — an alter — for each ego who was successfully assigned to a specific location or a country and who followed between one and 500 Twitter accounts, by randomly drawing an account from among those “followed” by each of those egos.10

When picking each alter, we immediately checked if the alter had provided a location description in their profile. In cases when the alter had not provided a description (about 25% of the time), we drew a replacement, until we found for each ego an alter with a location. 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.

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, and in 40 percent of those cases where both ego and alter had a specific location. When doing our analysis we found that mutual ties were more likely to be local and domestic, as one could expect. (Mutual following involves a

10 The reasons for discarding users who follow more than 500 accounts included concern about spam as well as the efficiency of data collection. Some of the accounts in our sample of egos followed tens of thousands other accounts. To pick alters for such egos we would need to first retrieve the information for each of the tens of thousands accounts that they followed. Since Twitter imposes limits on how many requests can be made per hour, collecting such accounts would lengthen the data collection process substantially. We were also concerned 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 his 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).
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

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 each specific location a pair of coordinates (obtained from Google Maps) and then merging nearby locations by replacing their coordinates with a weighted average of the coordinates of the locations that got merged. This reduced the 3134 location descriptions to a set of 386 regional clusters, which are comparable in size to metropolitan areas. (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% of the observations were assigned to a cluster with the center more than 100 km away.) We labeled each cluster with the most common place name associated with it in our sample. For example, the cluster centered on Manhattan is referred to as “New York.”

4.5 Language Coding

We coded the language of tweets from accounts that could be associated with a country or a specific location, using 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 just on the text of the messages.\footnote{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% 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% and 0.07%), Swedish and Norwegian (0.35% and 0.21%), as well as some even less common languages that jointly accounted for 0.81% 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.

5. Analyzing the Twitterverse

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 proximity, the frequency of air travel, national boundaries and language differences. In addition to descriptive statistics demonstrating the effects of each variable and investigating the nature of such effects, we analyze the effects of each variable using correlation with Quadratic Assignment Procedure (QAP, Krackhardt, 1987). QAP makes it possible to test the correlation between graphs, while compensating for auto-correlation inherent in network graphs and is a popular approach for graph comparison (Butts, 2007). In the last subsection we look at 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 the purpose of correlation and regression analysis we used networks with nodes representing 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 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, except for the network of airline connections. For this reason, we limited our correlation and regression analysis to the ties between just the 25 largest clusters, which allowed for a much less sparse Twitter network (an average of 0.76 ties per pair).

5.1 Physical Proximity

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 (see table 3). At the same time, over half of the egos are in other countries, as are four of the ten largest clusters, which include Tokyo, São Paulo, and two clusters in the United Kingdom. In this sense, Twitter users are distributed quite widely around the globe, with distance seemingly not serving as much of a barrier to Twitter use. At the same time, however, the users are highly concentrated in a relatively small number of locations. 25 clusters account for 54 and 61 percent of the egos and alters respectively. This level of concentration exceeds the general concentration of population in major urban agglomerations.

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 mentioned earlier: 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. Given the distribution of egos, only 2 percent of ties would be local if ties were formed randomly between the egos. (An average user’s cluster accounts for 2 percent of the total number of egos.)

Figure 1 shows the distribution tie lengths (the first graph, labeled “actual”), demonstrating that distance has an effect on non-local ties as well. 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 third graph of figure 1 (labeled “simulation 2”). Twitter users, however are not distributed evenly around the globe. (Neither is human population in general, of course.) The uneven distribution substantially skews the expected average length of randomly formed connections. The fact that users are concentrated in small number of clusters, as described above, would further lead us to expect the distribution to peak at values

Note that all networks other than the airline one were much less sparse than the Twitter network.

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.
corresponding to distances between major clusters. In our case, for example, we can expect a substantial number of connections with the length of 3,900 km – the distance between New York and Los Angeles, the two largest clusters. We can expect to find fewer of the shorter ties, due to the smaller number of users located in the range of 2,000 to 3,000 km from New York and Los Angeles. Similarly, we can expect to find very few ties approaching the length of 20,000 km — not just because of the prohibitive nature of such distance, but also because the antipodal points of all major clusters fall in the ocean. (The closest to an antipodal distance between a pair of major clusters would be the one between Sao Paulo and Japan, at around 18,600 km.)

The second graph of figure 1 shows the distribution of ties in a simulation in which egos, located where they are in our sample, form ties among each other at random. The graph shows the expected peak at 3,900 km (see above), followed by a valley corresponding to not-quite-transatlantic distances, followed by a rise as we reach Europe. A larger peak (at 7,600 km), corresponding in large part to the distance between New York and Sao Paulo but reinforced by a large number of other connections. The last two notable peaks (at 10,800 and 18,500 km) corresponds to the distance from Tokyo to New York and from Tokyo to Sao Paulo.

Compared with this baseline, the observed distribution of tie lengths shows a surplus of ties for distances to 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 Sao Paulo, New York and Tokyo, and Tokyo and Sao Paulo.

For network comparison we created a “proximity” network the weight of edges was set to a reciprocal of distance (20,000 km divided by the great-circle distance between the two clusters, calculated by haversine formula). The comparison of this network to the network of Twitter ties for the top 25 clusters shows a correlation of 0.33 for the top 25 clusters, significant at 0.005 level (see table 4). We note that this correlation represents the effect of proximity on ties between clusters, that is, after the shortest ties are already excluded.

Table 4 around here. (“QAP Correlations.”)

5.2 Air Travel

To investigate the effect of the ease of travel on Twitter ties we obtained a dataset showing a number of flights between pairs of 3,023 airports on five different days in 2008 and 2009. (The sample was kindly provided by a researcher at a different institution whose name is omitted for the review.) 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 the observed number of flights between the airports assigned to each of them.

The resulting network showed a correlation of 0.29 between Twitter ties and flights when looking at the full network of 386 clusters (the only case we found a meaningful correlation for networks between all 386 cluster) and a larger correlation of 0.42 for the top 25 clusters (see table 4). The network of flights thus appears to 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, is 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 note, though, that top Twitter clusters intersect only to an extent to 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 on both ends, 75 percent connect users in the same country. This prevalence of domestic ties is partly explained by the high frequency of local connections that we discussed above, since all local ties are domestic. Looking at just the non-local ties (i.e., ties between 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 fact that the users are concentrated 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 (26 percent). 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 that could 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 0.34, significant at <0.001 level (see table 4).

Table 5 around here. (“Top Countries.”)

The substantial share of the United States in the sample warrants a comparison of it 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 difference in the pattern of outgoing ties (following) and incoming ties (being followed). As table 5 shows (column 5), the majority of US 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 the sixth column 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. Japan closely follows the United States with 25 percent of location descriptions identifying the country. In 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 Japanese characters), which make them intelligible only to people who know Japanese and would therefore likely 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, though, have proportionally even fewer incoming foreign ties. This does not, however, stop them from identifying their country explicitly.) In case of the United States, the extremely low percentage of explicit country identification 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 the Twitterverse outside of the U.S.

5.4 Language

A large majority of egos (62 percent) and a yet larger majority of alters (68 percent) are located in countries where English is the dominant language. 96 percent of the egos 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 the ties are to English speaking countries. For non-English speaking countries, the share of same-language ties 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 build a language match network using a dataset of access to Wikipedias in different languages 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. The resulting language network shows correlation of 0.29 with the network of

14 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 – Japanese pair than to English – Dutch). We have found that the two language networks had a correlation of
twitter ties between the top 25 clusters (table 4).

It is important to note, however, that language stands in a more complicated relation with geography than, for example, the country effect. While a user located in the New York cluster is necessarily located in the United States by virtue of the fact that the New York cluster is in the United States, she may or may not tweet in English. It becomes important, therefore, to look at the language of individual tweets in addition to the language of the clusters.

Table 7 shows the most common languages used in the tweets. English is by far the dominant language, accounting for 73 percent of the messages. 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 8 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. 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.

The share of same language ties in other languages 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, only 87 percent of the tweets from users located in Brazil are in Portuguese, with another 8 percent being in English. (Informal analysis of the profiles suggests that many of the English-tweeting users located in Brazil are Brazilians rather than traveling English speakers.)

**5.5 Multivariate Analysis**

Having found that all four variables that we considered have an effect on Twitter ties, we used a regression analysis to see whether their effects are independent. The results of the regressions are presented in table 6. Comparing model 3 with models 1 and 2, we see that proximity and being in the same country have an independent and significant effect. A comparison of models 1, 4 and 5 shows that the same is true for proximity and language. (The effect of language, however, is not significant when we control for country.) Proximity and the number of flights similarly have an independent effect.

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.
(models 7 and 8). A model combining all four variables shows a significant effect of proximity, country match and flight number (though only at the 0.05 significance level). The number of flights appears to be overall a single best predictor of non-local ties and has the heaviest weight in the combined model. We note, however, the residual effect of simple physical distance.

### Table 6 around here. ("QAP Regressions.")

6. Conclusion

Looking at the network of ties in Twitter we found that distance and related variables (language, country, and the number of flights) all have an effect on Twitter ties despite the substantial 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 that this is not quite so. A very substantial portion of ties (39 percent) connect users within the same regional clusters, typically the size of a metropolitan area. (All such ties are also domestic and connect users in the same linguistic area. Most of them would fall within easy driving distance.) Even for the remaining longer range ties, connecting users in different clusters, distance appears to matter. Ties up 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 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 an independent analysis of such variables and a comparison between them. We find that proximity, country and the number of flight frequencies all have an independent effect. (The effect of language is no longer significant when country is included in the model.) The number of flights, however, appears to be the best predictor of non-local ties, highlighting the importance of the structural constraints on ties rather than simple physical distance.

7. References


PEW Internet and American Life Project, 2009. Twitter and status updating.


Ribeiro, E. 2009. Participação do Twitter no Brasil atinge 15% em junho, informa Ibope. IDG Now!, July 13,


8. Tables

8.1 Table 1: Location descriptions corresponding to Greater Tokyo.

<table>
<thead>
<tr>
<th>Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokyo</td>
<td>About 12 percent of the egos assigned to Greater Tokyo (counting alternative case)</td>
</tr>
<tr>
<td>Tokyo, Japan</td>
<td>About 7 percent of the egos assigned to Greater Tokyo (counting alternative case)</td>
</tr>
<tr>
<td>東京都</td>
<td>“Tokyo” (lit. “Tokyo Capital”), about 12 percent of egos assigned to Greater Tokyo</td>
</tr>
<tr>
<td>東京のベッドタウン: 埼玉</td>
<td>“Tokyo’s bedroom community: Saitama”</td>
</tr>
<tr>
<td>東京都中央区月島</td>
<td>“Tsukishima, Chūō ward, Tokyo”</td>
</tr>
<tr>
<td>東京都千代田区</td>
<td>“Chiyoda ward, Tokyo”</td>
</tr>
<tr>
<td>Shibuya, Tokyo</td>
<td>Shibuya is a ward of Tokyo</td>
</tr>
<tr>
<td>Shibuya</td>
<td>A ward of Tokyo.</td>
</tr>
<tr>
<td>渋谷</td>
<td>“Shibuya”</td>
</tr>
<tr>
<td>六本木とか千葉とか</td>
<td>“Some place like Roppongi or Chiba” (Roppongi is a neighborhood in Tokyo, Chiba is an adjacent city.)</td>
</tr>
<tr>
<td>六本木</td>
<td>“Roppongi” (a neighborhood of Tokyo)</td>
</tr>
<tr>
<td>小田急線沿線</td>
<td>“Along the Odakyū line” (a commuter rail line in Tokyo)</td>
</tr>
<tr>
<td>秋葉原の隣</td>
<td>“Next to Akihabara” (a neighborhood of Tokyo)</td>
</tr>
</tbody>
</table>

8.2 Table 2: Location precision in the sample of egos

<table>
<thead>
<tr>
<th>Location specification</th>
<th>Share of the sample</th>
<th>Used for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude and longitude</td>
<td>7.5%</td>
<td>cluster-level and country-level analysis</td>
</tr>
<tr>
<td>A named location with an area of &lt;25,000 km²</td>
<td>57.0%</td>
<td></td>
</tr>
<tr>
<td>Not specific, but enough to identify a country</td>
<td>20.4%</td>
<td>used for country-level</td>
</tr>
<tr>
<td>Very broad, non-spatial or undecipherable</td>
<td>15.1%</td>
<td>not used</td>
</tr>
</tbody>
</table>

8.3 Table 3: Top clusters

In this and other tables headings, “in” ties means being followed, “out” ties means following.

<table>
<thead>
<tr>
<th>cluster*</th>
<th>share of egos (%)</th>
<th>share of alters (%)</th>
<th>locality**</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>8.5</td>
<td>10.2</td>
<td>54.3</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>5.1</td>
<td>10.4</td>
<td>53.3</td>
</tr>
<tr>
<td>Tokyo</td>
<td>4.1</td>
<td>5.0</td>
<td>62.9</td>
</tr>
<tr>
<td>London</td>
<td>3.6</td>
<td>4.9</td>
<td>48.8</td>
</tr>
<tr>
<td>City</td>
<td>Q</td>
<td>R</td>
<td>Correlation</td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>------</td>
<td>-------------</td>
</tr>
<tr>
<td>São Paulo</td>
<td>3.5</td>
<td>3.6</td>
<td>78.4</td>
</tr>
<tr>
<td>San Francisco</td>
<td>2.8</td>
<td>4.1</td>
<td>41.2</td>
</tr>
<tr>
<td>New Jersey***</td>
<td>2.5</td>
<td>2.1</td>
<td>20.0</td>
</tr>
<tr>
<td>Chicago</td>
<td>2.2</td>
<td>1.7</td>
<td>32.0</td>
</tr>
<tr>
<td>Washington</td>
<td>2.1</td>
<td>2.6</td>
<td>34.3</td>
</tr>
<tr>
<td>Manchester</td>
<td>1.9</td>
<td>1.1</td>
<td>30.8</td>
</tr>
<tr>
<td>Atlanta</td>
<td>1.7</td>
<td>2.1</td>
<td>46.2</td>
</tr>
<tr>
<td>San Diego</td>
<td>1.5</td>
<td>1.1</td>
<td>26.3</td>
</tr>
<tr>
<td>Toronto</td>
<td>1.3</td>
<td>1.5</td>
<td>42.9</td>
</tr>
<tr>
<td>Seattle</td>
<td>1.3</td>
<td>1.2</td>
<td>58.8</td>
</tr>
<tr>
<td>Houston</td>
<td>1.2</td>
<td>1.0</td>
<td>40.0</td>
</tr>
<tr>
<td>Rio de Janeiro</td>
<td>1.2</td>
<td>1.1</td>
<td>30.8</td>
</tr>
<tr>
<td>Dallas</td>
<td>1.2</td>
<td>1.4</td>
<td>61.5</td>
</tr>
<tr>
<td>Boston</td>
<td>1.2</td>
<td>1.1</td>
<td>20.0</td>
</tr>
<tr>
<td>Amsterdam</td>
<td>1.1</td>
<td>0.9</td>
<td>50.0</td>
</tr>
<tr>
<td>Jakarta</td>
<td>1.1</td>
<td>0.3</td>
<td>42.9</td>
</tr>
<tr>
<td>Austin</td>
<td>1.0</td>
<td>1.3</td>
<td>50.0</td>
</tr>
<tr>
<td>Sydney</td>
<td>0.9</td>
<td>0.8</td>
<td>38.5</td>
</tr>
<tr>
<td>Orlando</td>
<td>0.9</td>
<td>0.6</td>
<td>16.7</td>
</tr>
<tr>
<td>Phoenix</td>
<td>0.8</td>
<td>0.6</td>
<td>11.1</td>
</tr>
<tr>
<td>York (UK)</td>
<td>0.8</td>
<td>0.5</td>
<td>25.0</td>
</tr>
<tr>
<td>Osaka</td>
<td>0.8</td>
<td>1.0</td>
<td>25.0</td>
</tr>
</tbody>
</table>

* identified by the most common place name
** share of local of ties among all ties for egos in this cluster
*** This cluster is centered between Philadelphia and Trenton, NJ and includes all locations identified as just “New Jersey.”

### 8.4 Table 4: QAP Correlations

<table>
<thead>
<tr>
<th></th>
<th>twitter</th>
<th>flights</th>
<th>language</th>
<th>domestic</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance</td>
<td>0.329</td>
<td>0.595</td>
<td>0.339</td>
<td>0.412</td>
</tr>
<tr>
<td>domestic</td>
<td>0.336</td>
<td>0.557</td>
<td>0.709</td>
<td></td>
</tr>
<tr>
<td>language</td>
<td>0.292</td>
<td>0.442</td>
<td></td>
<td></td>
</tr>
<tr>
<td>flights</td>
<td>0.420</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All p values are <= 0.005

### 8.5 Table 5: Top countries

<table>
<thead>
<tr>
<th>share of egos (%)</th>
<th>share of alters</th>
<th>percentage of</th>
<th>percentage of</th>
<th>following foreign</th>
<th>country named</th>
</tr>
</thead>
</table>

22
### Table 6: QAP Regressions

<table>
<thead>
<tr>
<th>Models</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>6.22</td>
<td>0.234</td>
<td>0.131</td>
<td>-0.200</td>
<td>-0.195</td>
<td>0.000</td>
<td>0.283</td>
<td>0.231</td>
<td>-0.0417</td>
</tr>
<tr>
<td>proximity</td>
<td>0.0390** (0.329)</td>
<td>0.0272** (0.230)</td>
<td>1.47*** (0.336)</td>
<td>1.54*** (0.292)</td>
<td>1.05*** (0.241)</td>
<td>1.08*** (0.203)</td>
<td>0.0309** (0.260)</td>
<td>0.0145* (0.123)</td>
<td>0.0123* (0.104)</td>
</tr>
<tr>
<td>domestic language</td>
<td>1.14** (0.260)</td>
<td>0.567 (0.107)</td>
<td>0.0035*** (0.420)</td>
<td>0.0029** (0.347)</td>
<td>0.0023* (0.278)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flights</td>
<td>0.385* (0.088)</td>
<td>0.375 (0.071)</td>
<td>0.0145* (0.123)</td>
<td>0.0123* (0.104)</td>
<td>0.0029** (0.347)</td>
<td>0.0023* (0.278)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.108</td>
<td>0.113</td>
<td>0.157</td>
<td>0.085</td>
<td>0.145</td>
<td>0.119</td>
<td>0.177</td>
<td>0.186</td>
<td>0.201</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.108</td>
<td>0.113</td>
<td>0.155</td>
<td>0.085</td>
<td>0.144</td>
<td>0.117</td>
<td>0.177</td>
<td>0.185</td>
<td>0.197</td>
</tr>
<tr>
<td>Number of observations</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
<td>600</td>
</tr>
</tbody>
</table>

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

### Table 7: Most common languages

<table>
<thead>
<tr>
<th>Language</th>
<th>% of ego tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>72.5</td>
</tr>
<tr>
<td>Portuguese</td>
<td>10.1</td>
</tr>
<tr>
<td>Japanese</td>
<td>5.4</td>
</tr>
<tr>
<td>Spanish</td>
<td>3.1</td>
</tr>
<tr>
<td>Indonesian</td>
<td>1.8</td>
</tr>
</tbody>
</table>
### 8.8 Table 8: Language combinations

<table>
<thead>
<tr>
<th>Language combinations</th>
<th>as a percentage of...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all ties</td>
</tr>
<tr>
<td>Same language (total)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>88.4</td>
</tr>
<tr>
<td></td>
<td>67.5</td>
</tr>
<tr>
<td></td>
<td>20.9</td>
</tr>
<tr>
<td>Cross-language</td>
<td></td>
</tr>
<tr>
<td>Other-English</td>
<td>7.4</td>
</tr>
<tr>
<td>English-Other</td>
<td>3.1</td>
</tr>
<tr>
<td>Different languages where neither is English*</td>
<td>1.1</td>
</tr>
</tbody>
</table>

* The most common combinations were Japanese-Chinese, Spanish-Italian, and Portuguese-Spanish.
9. Figures

9.1 Figure 1: Histogram of tie lengths

The graph omits ties shorter than 200 km, which number 543, since including such ties would dwarf the rest of the graph.