Computers, Environment and Urban Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Follow thy neighbor: Connecting the social and the spatial networks on Twitter

Monica Stephens [a,*], Ate Poorthuis [b]

[a] Department of Geography, University at Buffalo (SUNY), 105 Wilkeson Quadrangle, Buffalo, NY 14261, United States
[b] Department of Geography, University of Kentucky, 817 Patterson Office Tower, Lexington, KY 40506, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Social network analysis
Twitter
Geospatial analysis

ABSTRACT

This paper compares the social properties of Twitter users' networks with the spatial proximity of the networks. Using a comprehensive analysis of network density and network transitivity we found that the density of networks and the spatial clustering depends on the size of the network; smaller networks are more socially clustered and extend a smaller physical distance and larger networks are physically more dispersed with less social clustering. Additionally, Twitter networks are more effective at transmitting information at the local level. For example, local triadic connections are more than twice as likely to be transitive than those extending more than 500 km. This implies that not only is distance important to the communities developed in online social networks, but scale is extremely pertinent to the nature of these networks. Even as technologies such as Twitter enable a larger volume of interaction between spaces, these interactions do not invent completely new social and spatial patterns, but instead replicate existing arrangements.

1. Introduction

Internet Communication Technologies (ICTs) have reconfigured the role of distance in social relationships. Email, mobile phones, and online social networks allow relationships that would have previously been neglected or discontinued to be more easily maintained. These weak ties, or acquaintance based relationships that were often discontinued when an individual relocated or their interests evolved are now maintained as a digital community linking an individual to those they interacted with previously. Weak ties differ from strong ties such as familial or friendship relationships that an individual maintains throughout life. An individual generally maintains more weak ties in their personal network than strong ties as less time and energy are needed to maintain these connections. Both weak ties and strong ties often emerge out of spatial proximate social interaction, but can be maintained with online interactions. Twitter, a popular microblogging social network, is one example of a weak tie online social network that allows millions of users to establish digital communities that incorporate a combination of offline contacts and online contacts of interest and maintain relationships that otherwise would have faded.

This paper uses Twitter as an example of a weak tie network to understand how distance impacts social relationships and networks. Communities on Twitter (Gruzd, Wellman, & Takhteyev, 2011) form through users following other users they either know already or whose interests are relevant to them. As most Twitter users disclose their location and contacts, this provides scholars with a way to measure the geography of digital networks established by millions of users around the world. We use this established online social network to ask: what is the relationship between the physical distance of virtual connections established by one's Twitter contacts and the social connectivity among those contacts?

This paper contributes a new way of integrating physical and social distance online to understand the geography and transitivity of communities connected through Twitter. This paper will proceed as follows: Section 2 introduces scholarship integrating geography and the Internet; Section 3 introduces literature pertaining to Twitter and social capital as well as our data collection and geocoding procedures; Section 4 introduces terms and means of measuring spatial and social geographies of Twitter and compares these measures. In Section 5 we conclude that Twitter is reflective of "social neighborhoods" that exist offline through replicating existing social patterns. We determined that networks existing at less than 500 km are stronger and more effective in disseminating information.

* Corresponding author. Tel.: +1 7166450499; fax: +1 7166452329.
E-mail address: mgstephe@buffalo.edu (M. Stephens).

http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002
0198-9715/

2. Geography, the Internet and community

A large part of the early thinking on digital communities and its consequences can be labeled as 'naïve' with hindsight. Many thinkers – especially in the popular media – thought that the Internet would make geographical differences smaller and smaller. Even relatively recent, the oft cited Friedman (2007) defends this idea in his 'world is flat' thesis. A decade before Friedman, Cairncross (1997) made a similar claim already: not only does history end (cf. Fukuyama, 1992), the death of distance is near. After humankind gets rid of distance, telecommunications will help dissolve the differences between rich and poor; between small and large. This death of distance discourse in popular media is closely related to what Graham (1998) calls the substitution perspective on cyberspace. It argues that attachment to 'place' is replaced by new technologies: cyberspace is thus replacing 'human' space. This perspective is often used by those who feel technological change is endangering social functions, leading to placelessness (Leamer & Storper, 2001). Online communities are alleged to form a complete substitute for the sense of belonging that place offers (Crang, Crosbie, & Graham, 2007).

The distance destroying capability of technology has been disputed by other scholars who note the geographical dependence in the uses of the Internet (Adams & Ghose, 2003). Geography remains relevant to transport costs, the ongoing evolution of digital divide, borders, and cultural differences. Additionally, co-presence remains a key element in the development of social capital, and building new relationships is greatly aided by spatial proximity, while the social depth of knowledge exchange declines across distance (Leamer & Storper, 2001; Morgan, 2004). Thus, geography is very much alive in the digital cities of the 1990s, the place-based review sites of the early 2000s (e.g. Yelp, Google Maps), and the hyper-local social network sites of the late 2000s (Foursquare, Facebook Places).

In short, geographers insist that "The net cannot float free of conventional geography" (Hayes, 1997 in Zook, Dodge, Aoyama, & Townsend, 2004) and emphasize that it is impossible for Internet users to completely disconnect from the material world in which we are embedded. Thus the social networks represented through platforms like Twitter and Facebook have a geography that blends digital and material dimensions. Online networks can function as a hub of camaraderie among individuals with unique interests unrepresented in the material community around them (such as online support groups or sexual fetish sites); or as a precursor to interactions in the material world (such as online dating sites or job seeking sites). Although online, these social networks are still intrinsically connected to the offline world and subject to similar social, cultural, linguistic and economic constraints.

However, as computers are increasingly used for social interactions that connect people and organizations around ideas, the geography of social relationships becomes more complex. Wellman (2001) examined computer networks that function as social networks and found that email connections increase social capital as ties—the bonds between individuals, are maintained with respect to geography. When distance increases email replaces face to face communication for strong tie relationships. These individualized "fragmented community networks" are reinforced by email, which allow existing offline communities to "sustain interactions across vast distances" (Juris, 2004).

ICTs enable complex social geographies of use, with interactions in cyberspace simultaneously influenced by physical proximity as well as a network distance in cyberspace (Li, Whalley, & Williams, 2001). This network distance in cyberspace is not mutually exclusive from the distances traversed in the material world; network and physical distances are related, reflexive and co-constructive. And it is precisely this nexus that makes studying the geography of Twitter networks so relevant.

In March, 2012, Twitter was the second largest online social network in the world with 500 million registered users and 100 million active users (Twitter Blog, 2011). Twitter as a microblogging service allows users to set up profiles with a self description of 160 characters and select a group of individuals to 'follow'. When the user visits the site they can peruse through the 140 character updates, 'tweets,' that each of the users they follow has sent out. Users select those they follow, 'friends,' but they do not select those who follow them, 'followers.' The combination of 'followers' and 'friends' are considered the ties with whom a user communicates. These ties are the focus of this paper.

3. Understanding the geography of Twitter

Twitter, along with many of the ICT technologies that pre-date it, enables users to connect and communicate around mutual interests and needs rather than just spatial proximity (Civin, 2000; Zuckerman, 2008). Twitter users establish social ties "based on shared interests instead of shared place", especially for interests lacking a critical mass in material space (Hampton, 2004, p. 218). While this gives users the potential ability to bypass local constraints and connect to individuals in geographically distant spaces (Graham, 1998), it is doubtful that it renders geography meaningless in the constitution of social networks. This relation between online social network and the underlying 'real world' geography has been of interest to many geographers. For example, in a special issue of Cartography and Geographic Information Science on 'mapping cyberspace' (Tsou & Leitner, 2013) several authors explored this same relationship of new, digital data with 'real' world phenomena. Li, Goodchild, and Xu (2013) show that the digital data footprint is very much related to various variables derived from the US census. Kent and Capello (2013) show that, even with a small number of available tweets on a wildfire in Wyoming, careful handling of this data can result in useful, hyper-local, insights. Similarly, in their study of a Lexington, KY riot through Twitter data, Crampton and et al. (2013) show both the place-based nature as well as the scale-jumping that online social networks can exhibit.

Email networks are strong tie networks that increase social capital through a one-to-one relationship. Twitter-based relationships are less of a strong tie relationship, as the many-to-many structure of communication does not necessarily build social capital. Similar to email networks, Twitter does not require reciprocal ties. A user can follow a user without that user following them back. For example, a Twitter user may follow a public figure they admire and want updates from without ever meeting them. Unlike email, the effort required to establish a tie is much lower on Twitter (a matter of pushing a button), thus twitter networks contain several weak tie relationships.

Although requiring less effort to maintain, weak ties within a network can be more useful than strong ties. Granovetter (1973, 2005) suggests that weak ties provide 'bridges' to parts of a network that would otherwise not be connected. This provides new and novel information while strong ties only connect to well-known parts of the network (they are less likely to be a 'bridge') and thus often yield redundant information. Twitter networks are generally comprised of a combination of weak ties and strong ties. For a tie to develop between users, geographic proximity is not necessary per se. Although these connections may seem inconsequential, these weak ties generate a constant stream of information that can build social capital at both the local geographic level and within networks of interest at a variety of scales. Twitter as a social network gives us a unique opportunity to understand how people connect across space and build networks online.

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002

3.1. Earlier work on Twitter geography

As an online social network, Twitter gives us the opportunity to examine the relationships among individuals and makes it possible to understand how spatial distance influences weak tie networks. More specifically, researchers have analyzed the connection between the physical distance of a Twitter user (the ego) to their followers (their alters), and the social connectivity among the Twitter networks (the ego-alters) at the user level. However, an analysis of ties among millions of ego networks would be difficult to compute even in a "big data" framework (Manovich, 2011) and thus many scholars (e.g. Takhteyev, Gruzd, & Wellman, 2012) use small samples of geocodable users (based on user supplied location information).

Takhteyev, Wellman and Gruz (2012) found that 39% of Twitter ties are shorter than 100 km (roughly the size of a metropolitan area), and both national boundaries and language ties are significant in limiting the ties formed on Twitter. Often, the ties that do form across boundaries and long distances replicate airline ties between cities, economic ties, and migration patterns. Additionally, they discovered that ties at distances of up to 1000 km are more frequent than would be expected from random ties, while ties at distances larger than 5000 km are underrepresented. In short, Twitter networks are shaped in part by geographical constraints.

The Takhteyev et al. (2012) analysis of Twitter, while incorporating both geography and a large social network, aggregates the connections between individuals to regional clusters and uses cities (rather than individuals) as a unit of analysis. As a result, their work tells less about the individuals who inhabit the cities and the ties they form by using Twitter. Like Takhteyev et al., we analyze the user supplied location of Twitter *users* and specifically not the location of geotagged *tweets* (generally via a GPS enabled phone, which can increase geographic accuracy but represents only 1–3% of all tweets). We also limited our analysis to randomly sampled egos that originate in the United States for three reasons: the gazetteers in available geocoding services are not as complete for other parts of the world; there are fewer language barriers; and the expected geographic patterns are well known to the researchers. This eliminated the need to control for national borders or common language ties. Like Takhteyev et al., we determined social network analysis as the appropriate methodological approach for this study as it allows for graphing and analyzing the relationships among virtual entities, and provides tools for comparing aspects of networks.

3.2. Data collection

This paper examines the relationship between the spatial distance and social connectivity of ego-alter networks on Twitter. To study this, we calculated the locations of a sample of Twitter users (egos) and harvested the network of each ego's followers (those who follow the ego's updates) and friends (those whose updates the ego follows). We define the aggregate of both followers and friends as alters, as these are the nodes with ties to the subject, or ego, of our sample.

To collect the necessary Twitter data we made extensive use of the Twitter Application Programming Interface (API). The API allows developers and researchers to query parts of Twitter's database with HTTP requests. To do this, we wrote a series of custom scripts in the Ruby programming language using the 'Twitter' library/gem.[1] We mainly used the REST API in a multistep approach to assemble the dataset used in this paper. Since the basic unit of analysis in this paper is an ego network, the first step is constructing a random sample representative of Twitter users in the US. Previous authors (e.g. Takhteyev et al., 2012) have sampled Twitter users by 'listening' to Twitter's Streaming API. The Streaming API allows researchers to tap into a continuous stream of Twitter messages that, depending on one's access level, constitutes of a random sample of 1% or 10% of all tweets sent around the world. However, using sent tweets to sample Twitter users will oversample the most active users. In addition, if harvesting is performed during specific time intervals, users in time zones that are generally asleep will be under-sampled.

To circumvent this issue, we take a different approach. Twitter assigns each user an incremental identifying integer. Jack Dorsey – one of Twitter's founders – has user id 12 and by signing up for a new Twitter account we determined the last id at the time of data mining (id 388628058). Using those ids as the lower and upper bound, a sample of 100,000 random user ids was taken. For each of those ids we used the 'users/lookup' API call to download information for that user. Of the 100,000 initial users, only 86,313 are still active (the others have been deleted or suspended by Twitter due to a violation of the Terms of Service). The information available for each active user includes: name, location (user defined), number of friends, number of followers, number of tweets, and most recent tweet.

Since we are interested in spatial networks, the second step is to geocode all user defined locations to specific latitude–longitude pairs. This means converting "Chicago" into "41.89, −87.70". While it might be easy to geocode "Chicago", deciding whether "Paris" refers to "Paris, France," "Paris, Kentucky," or "Paris, Texas" is much more challenging. To reduce some of the messiness inherent in geocoding user defined locations (especially user defined locations such as "The Moon" or "Cloud 9" (see also Graham, Hale, & Gaffney (2014)), we used three geocoding services (Bing, Google, Yahoo) simultaneously. A geocoding result was considered reliable if all three services were in agreement (i.e. the distance between the coordinate pair for all three services was less than 10 miles) and the entry is at least at the city level (i.e. "Brooklyn" and "New York City" are included but "Kentucky" is not). From the 86,313 existing users in the sample, 21,518 (24.9%) satisfy these criteria. As a final step, and a starting point for the dataset used in this paper, we randomly selected 500 users whose location was within the contiguous United States and had tweeted more than once. We outline this sampling strategy in Table 1.

Table 1
Break down of sampling strategy.

Data Set	Size
Original sample	100,000
Existing users in sample	86,313
Geocodable users	21,518
Geocodable users within US	6943
Starting sample of geocodable users within US	500
Final sample after cleaning	400

For every user in the starting sample ($n = 500$), all friends and all followers (alters) were downloaded using Twitter's 'followers/ids' and 'friends/ids' methods. All alters were geocoded using the same process outlined in the previous section – and thus non-geocodable alters were removed. As a final step, we download all alter–alter connections for each ego's geocodable alters. This establishes the strength of the social relationship within each ego's network (how frequently do their friends/followers connect with each other). We consider each ego network to be independent and do not include alter–alter ties that connect the ego network to other ego networks. The final data set of 400 ego networks[2] has a total

[1] https://github.com/jnunemaker/twitter.

[2] 41 egos did not have any geocodable alters and were thus not included in the final dataset. 59 egos were geocoded to fictional locations and had to be cleaned from the dataset manually. Based on their alter-alter connections a user-generated location of "Earth"/"Mars" did not refer to Earth, TX or Mars, PA, but was entered flippantly, a common problem with user-generated data (Graham et al., 2014).

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002

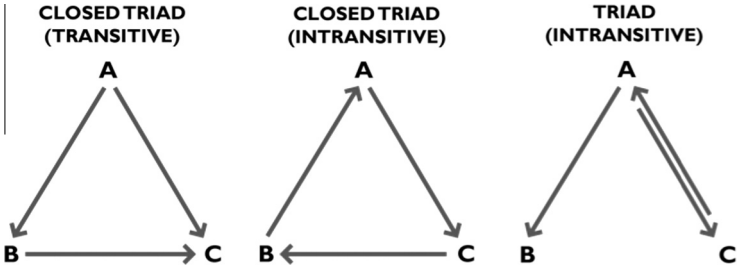

Fig. 1. Three of the potential network sub-structures for transitive and intransitive triads.

Fig. 2. Ties for 400 US-based ego networks.

55,880 unique alters and approximately 4.5 million ties. Of those ties 87,292 are between egos and alters, while the other approximately 4.4 million are between alters.

4. The spatial and social geography of Twitter

As we analyze ego networks, we compare social and spatial distance on the network level.[3] That means that in both dimensions, we need distance measures that reflect the average distance within that network. In spatial statistics, the standard distance is often used for this purpose. Analogous to how standard deviation is used to measure the dispersion or variation in a conventional statistical distribution, standard distance measures how clustered or dispersed points are around the mean (geographic) center. Standard distance (SD) is calculated in a similar fashion:

$$SD = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$$

where d is the distance to a given point i (x, y) from the mean center and n is the total number of features.

Analogous to spatial distance, the average distance within a social network can be approximated by network density:

$$Networkdensity = \frac{T}{n(n-1)}$$

where n is equal to the number of nodes (alters plus egos) and T is equal to the number of ties. Density is thus simply a ratio of the number of ties in the network to the number of possible ties (Faust & Wasserman, 1994). Since the networks in this paper are all non-reflexive ego-networks and thus always have ego-alter connections, density becomes an indicator for how often alters are connected among themselves. As such, density indicates how closely knit a specific ego network is. A density of 1 means that all nodes in the network are connected to every other node. A density of 0 means that there are no ties between any of the nodes.

In a similar fashion, we also look at triads. A triad is a set of three nodes, who may or may not share a social relation with each other (Kitts & Huang, 2010). Simmel (Simmel & Wolff, 1950) argued specifically that triads are a complex unit that cannot be understood simply through analyzing individual nodes or ties. For example, in a work relationship with manager A and employees B and C, ties exist between A–B and A–C. However, if no tie exists between employees B–C, manager A may exert specific power over the employees (e.g. paying one more than the other). Taking into account the direction of the ties, there are 16 possible triad configurations. For any network, these configurations can be counted, which is often referred to as a triad census (Davis & Leinhardt, 1972). Of specific interest for network analysis are transitive configurations. If a tie from A to B exists, and one from B to C exists, a triad is transitive if a tie from A to C also exists (if that tie does not exist, the triad is referred to as intransitive). See Fig. 1 for a demonstration of transitive and intransitive triads. Some social relations are transitive: if A has power over B and B has power over C, then it is likely that A also has power over C. On the other hand, romantic relationships tend to be intransitive (Kitts and Huang,

[3] All social network analyses were done using *igraph* (Csardi & Nepusz, 2006), spatial analysis using *aspace* (Bui, Buliung, & Remmel, 2011), graphs were made with *ggplot2* (Wickham, 2009) – all of which are packages for R (R Core Team, 2013).

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002

2010). In Twitter networks, we expect to find transitive triads much more often in networks that are spatially clustered: if user A follows users B and C, user B is much more likely to also follow user C if they are spatially co-located, as this implies they may have some relationship outside of Twitter. We calculate transitivity for network (G) as:

$$T(G) = \frac{3\lambda(G)}{\tau(G)}$$

where 3λ is equal to all transitive triads and τ represents the total number of triads.

4.1. Exploring spatial distance within Twitter networks

The abundance of Twitter ego-alter connections within the U.S. seen in Fig. 2 shows that although users can connect to people all over the world, the majority of ties from this sample of U.S. based egos are domestic. Indeed the fact that Twitter provides a location field and many Twitter users fill it out suggests that material location is an important part of how people are using Twitter. This supports the findings that Takhteyev et al. (2012) made based on Twitter ego-alter distances.

Looking at the distribution of standard distances (Fig. 3), it is clear that most networks (68%) have a standard distance radius of less than 3000 km (mean = 2541; median = 2322; standard deviation = 1464). As the distance from the East Coast to the West Coast of the United States is approximately 4800 km, this supports what is visually illustrated in Fig. 2. Twitter networks in the U.S. are spatially constrained and follow established network patters that are constrained by national borders and population density. Takhteyev et al. (2012) similarly found that the city pairs of Twitter users follow existing routes of connectivity throughout the world, such as airlines and migration routes.

We assume that Twitter connections are frequently within the same city. Users also follow a few national-level users, which extends ties to distant cities and thus increases the standard distance. Fig. 3 illustrates this relationship as some users connect to friends and followers on both coasts (extending the standard distance upward) but very few develop ties outside of the United States.

Moreover, standard distance is positively correlated with the size of a user's Twitter network. This suggests that as an ego-network expands there are more chances for outliers that increase the average spatial distance. Indeed, Fig. 4 shows the extent to

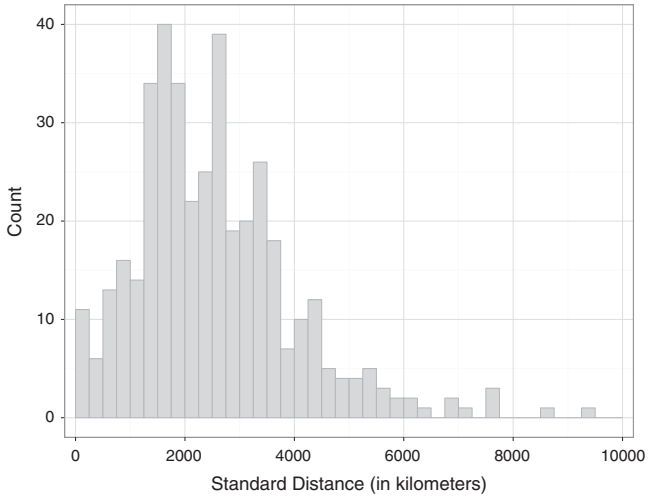

Fig. 3. Histogram of standard distance deviation for each ego network.

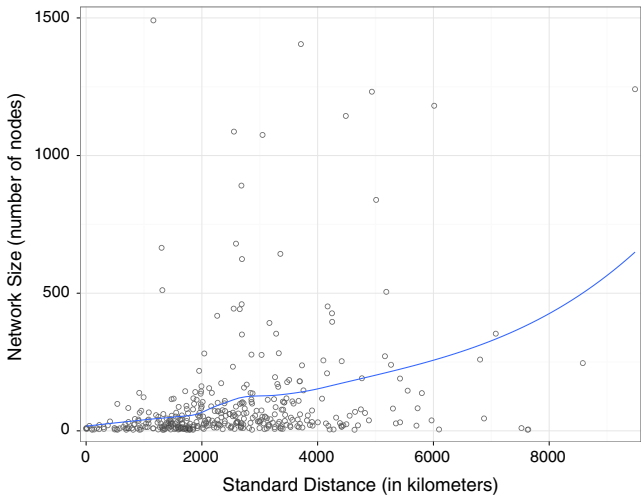

Fig. 4. The network size versus standard distance for each ego network.

which network size is correlated with standard distance (Spearman's rho[4] correlation coefficient = 0.404). Larger networks with more ties extend over a larger distance. Moreover, larger ego networks on Twitter are more likely to be Twitter accounts associated with corporations (e.g. The New York Times), which have followers throughout the world, or Twitter accounts used by marketers that adopt a strategy of randomly following accounts with the hope that it will be reciprocated.

In addition to the standard distance among ego networks, we also calculated the median distance for the same networks. However, as most networks are relatively small with few alters (mean network size is 150.3), the median distance is affected by outliers and thus not representative of the average distance in the total network. Therefore, we will use the standard distance for the remainder of this analysis. In the next section we expand our analysis of the spatial distance of network ties to consider how social distance is measured, and subsequently how social distance relates to these spatial extents of Twitter networks.

4.2. Social distance within Twitter

Although Twitter networks are shaped by geographic proximity, they are inherently social as well. Fig. 5 visualizes one of the 400 ego networks in this study. While some alters in this network do not connect to any other alters, it is evident there is a sub-community of alters who connect with each other as well as the ego. This is consistent with other studies on virtual communities (Wellman & Gulia, 1999) as well as previous work on communities on Twitter (van Meeteren, Poorthuis, & Dugundji, 2009), which shows that Twitter users form communities often centered around common interests. This clustering tendency is present in our data set as well. For an ego-network with no alter–alter connections, the network density multiplied by the number of nodes will always equal 1, which is only true for 9% of the networks in the sample. This is low, especially considering that all non-geocodable alters were excluded from this data set,[5] indicating a high degree of clustering.

The mean network density for each ego network across the data set is 0.17, which is a common network density threshold among communities comprised of friends and family in material space (Wellman, 1979). Similar to spatial distance, network density is

[4] Since the relationship in Fig. 4 is not linear but is monotonic, we use Spearman's rho instead of Pearson's r.

[5] 56% of all alters were filtered, bringing down the number of ties from 47 million to 4.5 million.

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002

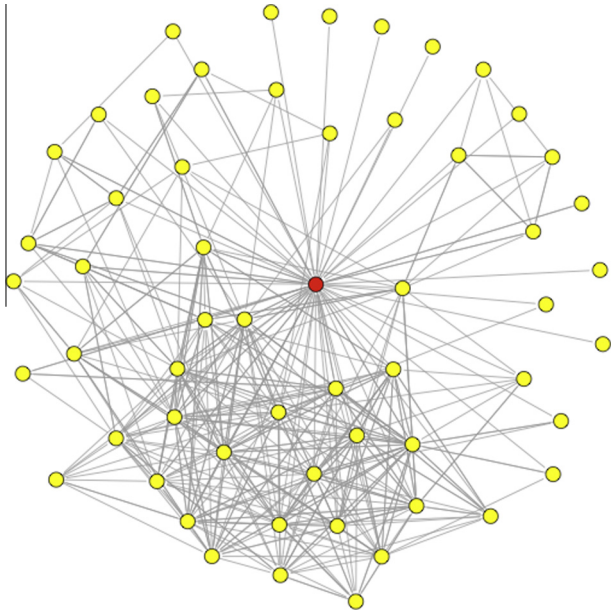

Fig. 5. Sample ego-network for user id 16850608. Ego in red. Alters in yellow.

correlated with the size of the network but the direction is reversed. Fig. 6 shows the frequency of network densities among ego-networks. Larger networks generally have lower network density and this pattern is repeated in this dataset (Spearman's rho = −0.78). Fig. 7 demonstrates this concept: as the size of the network increases, its density drops off dramatically. As the number of alters increases for an ego, it is unlikely that every alter will continue to develop a connection with every other alter. In smaller networks (with 3–5 alters), it is easy for the alters to all connect with each other.

4.3. Comparing social and spatial distance within Twitter

While most users have networks that combine local ties and long distance ties, this analysis shows that more spatially clustered networks are also more dense—friends and followers create triads by also following each other. This relationship is demonstrated in Fig. 8 where the standard distance and the network density are negatively correlated (Spearman's rho = −0.40). As standard

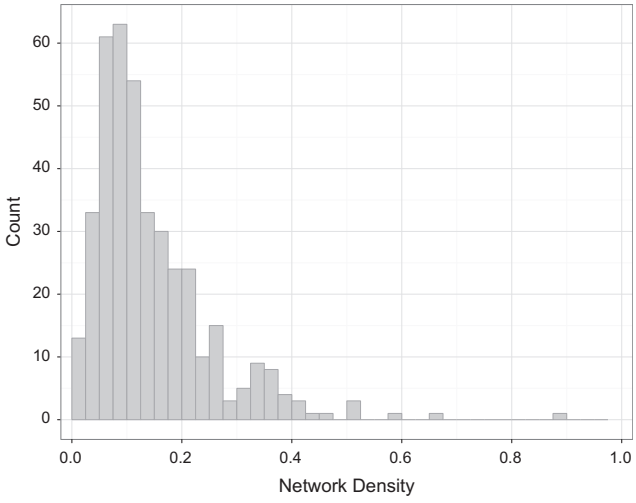

Fig. 6. Histogram of network density for each ego network.

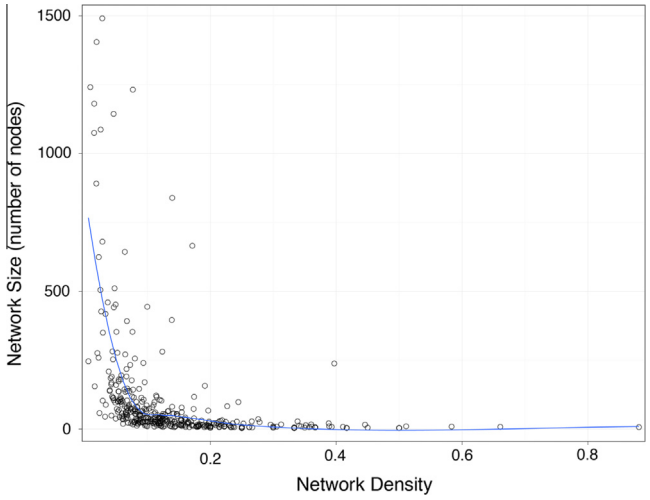

Fig. 7. Network size versus the density of each ego network.

distance among the ego and alters increases, the density of the network (relationships among the alters) decreases, implying that networks are less socially clustered as they extend over larger distances.

Granovetter's (1973) Strength of Weak Ties (SWT) theory argues that the stronger a tie between two people, the more their social networks will overlap. For example a marriage is a strong tie between two individuals, and as their tie is strong, it is likely their social networks will overlap, and individual contacts will become contacts of both people. This is especially relevant as Twitter ties are most frequently weak ties. Granovetter (1973) referred to this type of cluster as a "social neighborhood" with clear spatial implications as "neighborhood" implies spatial proximity. In the same time period, Waldo Tobler, famously inferred that spatial proximity (nearness) implies similarity (Tobler, 1970). We explore both SWT theory and Tobler's Law for Twitter, comparing the social and spatial distance of ego-alter relationships on Twitter.

We defined physical nearness using the standard distance deviation calculated in kilometers. Networks are more spatially clustered if they have a lower measure of standard distance. They are more dispersed if they have a higher standard distance. Fig. 8 identifies the relationship between this measure of physical distance and the network density of this study's ego networks (the clumpiness of a network, or social distance, described above). This relationship demonstrates a decay in network density as physical distance increases. Distance strongly impacts the social clustering of networks at shorter distances while leveling out after about 2000 km (roughly the distance from Boston, MA to Chicago, IL). We assume this relationship results from a bi-coastal effect with few connections between the densely populated cities on the coasts of the United States and the less populated hinterlands. This relationship of nodal connectivity among cities was also identified by Castells (1996) and by Takhteyev et al. (2012).

While there is a clear general trend in the relationship between physical distance and social distance, it is also evident that there are a number of ego-networks that do not match this pattern. Better understanding the reasons for this variation may provide important insight; outliers in Fig. 8 were individually identified for closer examination. Many outliers were not ego-networks of independent individual users, but networks comprised of (or heavily influenced by) the actions of corporate Twitter accounts, group users and bots whose specific patterns we describe below.

Chu, Gianvecchio, Wang, and Jajodia (2010) defined a classification system to separate bots, automated programs that use Twitter to exploit the platform, from human users. They found that

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002

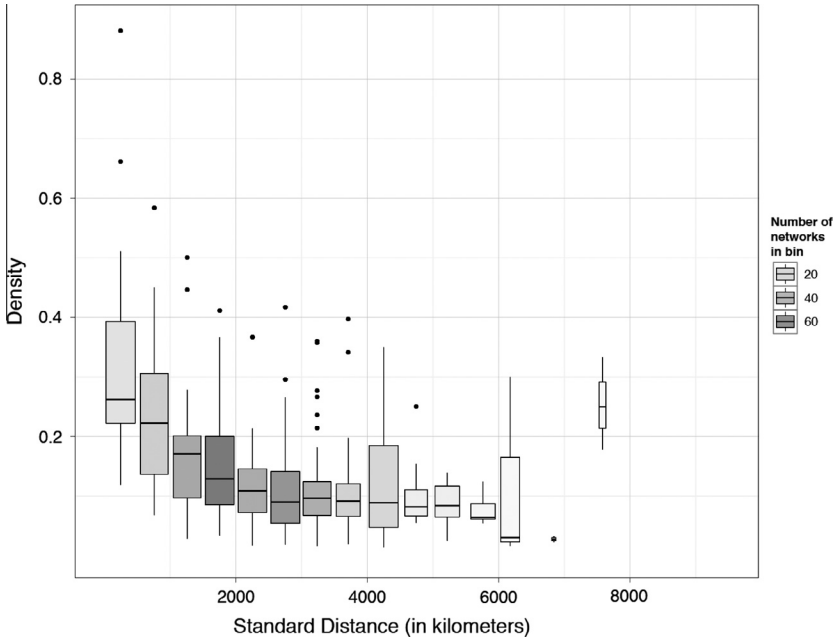

Fig. 8. Box plot of network density versus standard distance. The shade of the boxes in the plot indicate how many networks are represented by the box. The decline of density as distance increases is demonstrated by the positioning of the boxes, the whiskers represent the outliers.

automated bots comprise 13.8% of Twitter users, humans are 47.7% of the users, and a combination of humans and bots that they refer to as 'cyborgs' are 37.5% of users (Chu et al., 2010). In our study we qualitatively determined that some of the largest networks are not individual users but 'bots' or 'cyborgs'. For example, in Fig. 8, the network with a density of (roughly) 0.4 and a SD of 3800 km is a nursing jobs list for Miami, FL. This bot follows the other "tweet-my-jobs.com" Twitter bots, which also reciprocally follow it. It is additionally followed by an assortment of human Twitter users who are interested in nursing jobs in Miami, FL (career counselors and unemployed nurses). This arrangement of reciprocal connections with other tweet-my-job bots that are scattered around the US explains why it is such an outlier compared to the relationship between physical distance and social distance present in most ego-networks.

Another example of a large outlying ego-network is the Twitter network for a band, which has a standard distance of 6337 km and a density of 0.02999. This user represents not an individual, but a group that follows individuals in the cities they visit in hopes of a reciprocal following. This is different from an individual user who follows users to obtain updates from their tweets. This band was following users to enhance their own publicity.

In contrast, the network with the lowest standard distance (0 km) and highest network density (0.88), is associated with a church youth group in Noblesville, IN that follows and is followed by 7 individuals (not bots) within Noblesville, IN. This represents a near perfect conflation of space and sociability as this social group is using Twitter in a manner very similar to how they interact in 'material' space.

Another variation is a user located in Palmdale, CA that follows mostly high school friends in and around Palmdale, but is additionally followed by a spam bot in Saudi Arabia (whom he does not follow back). This increases the SD in his network (which is 7628 km) while lowering his network density to 0.333 (as none of his friends follow the bot). Other outliers in the sample appear distant from their alter as a result of locational mismatches. One Twitter user indicated his location as "San Francisco", but most likely lives in Seoul, Korea, tweets in Korean and all of his followers are in Seoul.

This suggests that the relationship between spatial and social distance in Twitter networks might be even stronger than suggested by Fig. 8 as the scores for many outliers are inflated or deflated by single alters (e.g., the Palmdale student and his spam-bot), unusual strategies not generally associated with real social networks (the nursing list or band), or by miscoding of the locational data (such as the user in Seoul/San Francisco).

4.4. Transitivity within Twitter networks

One advantage of this ego-level methodological design is that we can understand the geography of Twitter users at the ego level rather than an aggregate of their geography or connectivity at the city level. This also provides insight into how information travels through these social networks. Because network density is a rather crude measure of social distance, the following section uses transitivity as a more meaningful approximation of social closeness. Transitivity allows us to examine how effectively information can travel among alters. If a network has a high level of transitivity, alters are connected (as with network density), but they also have a directional connection that allows them to transmit information directly and efficiently without connecting to the ego as an intermediary.

In social relations, not all ties all equal. Network density crudely counts the number of ties in a network, treating them all equally. Transitivity compares the number of transitive triads (see Fig. 1) to the total number of triads in a network. Transitivity accounts for directional connectivity as it looks at ties within triads – a set of three nodes. As demonstrated in Fig. 1, three different network substructures all have a network density of 0.5 (calculated as three ties between nodes divided by six possible ties), but each has a different implication for how information travels within the network. In the first example (Fig. 1: left), there are three directed ties between users A–B, A–C and B–C (a transitive triad[6]) and any information (e.g. a tweet) that A sends reaches both B and C directly and

[6] All transitive triads are closed triads, but not all closed triads are transitive as transitivity accounts for the directional connectivity in the triad. In Fig. 1 the left and center triads are closed triads.

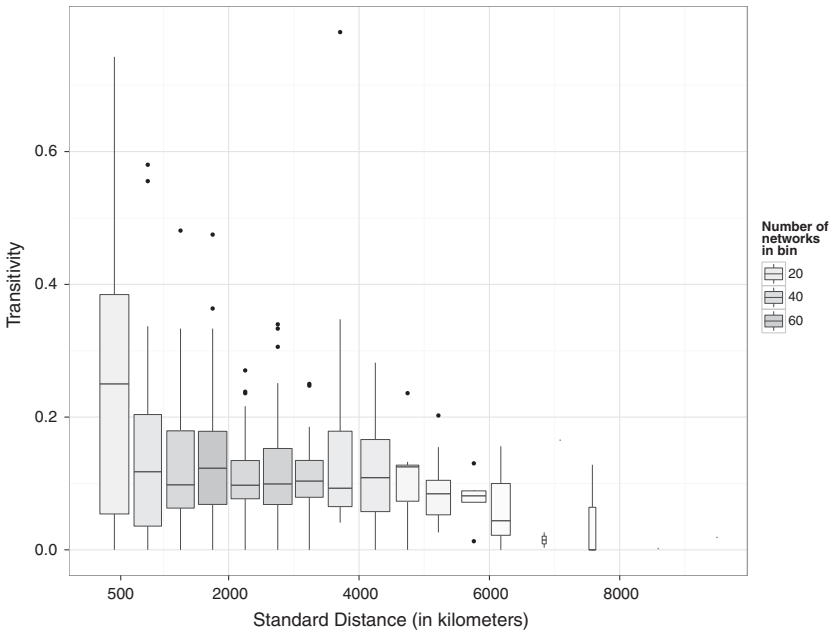

Fig. 9. Network transitivity versus standard distance.

has a chance to be amplified (by B's retweeting of the message to C). In the second example (Fig. 1: center) B–A, A–C, and C–B, A's tweet initially only reaches C and will only reach B if C decides to retweet. In the third example (Fig. 1: right) A–B, A–C, and C–A, a tweet from A will reach B and C but, since B and C are not connected, has no chance of being retweeted between the two. In short, a higher degree of transitivity in a Twitter network means that the echo chamber effect in that network will be larger and information will be more efficiently disseminated.

We examined the impact of transitivity on distance in Twitter ego networks and discovered a comparable relationship to Butts, Acton, Hipp, and Nagle's (2012) study[7] on travel patterns where distance increased as network transitivity decreased. This relationship is demonstrated in Fig. 9, (Spearman's rho = −0.13), where as distances increases, the ability for information to travel from the ego through the network decreases. We attribute the low Spearman's rho coefficient to the transitivity dropping sharply until the 500 km mark, after which it stabilizes around a transitivity of 0.1 for distances over 500 km (see Fig. 9).

Both network density and network transitivity decline as distance increases. However, there is an important difference in these measures at the local level. In ego-networks with a standard distance of less than 500-km, transitivity is high (0.25). After 500-km, transitivity appears to remain relatively constant (oscillating around 0.1) even as physical distance increases greatly. This implies that while there is a high-level of network density among ties with a standard distance less than 3000-km, the triads are not as transitive as they are at a distance of less than 500-km. This suggests that the Twitter network is structured in such a way that information flows most effectively among ties extending less than 500-km; most likely because those users have an offline relationship or share something in common.

[7] Butts et al. (2012) analyzed aggregated human travel networks to look at the connection between the social networks and spatial travel patterns for several Metropolitan Statistical Areas in the United States. While they used ties based on daily travel patterns and not node based individual data, they identified a decreasing relationship ($y \propto x^{-1.34}$) between distance (in decimal degree) and transitivity implying that the alters were less connected as distance increases.

5. Conclusions

To understand the complexity of relationships on Twitter one has to look at both the social network and the geographic network. Despite the ability of Twitter to transcend physical distance, it retains a strong local connectivity. Distance is not dead, nor is it completely dominant as Twitter functions similarly to older ICTs, such as email, by blending digital and off line relationships. The ties formed on Twitter do not rely on spatial contiguity, but are stronger within physical, cultural, and linguistic boundaries.

In many ways Twitter is reflective of the "social neighborhoods" existing offline. Twitter networks primarily replicate existing social and spatial patterns, such as flights and telecommunications patterns. As networks get larger in size they do not keep the same ratio of connectivity present in smaller communities: network density decreases over distance. Similarly, as a user adds more friends or gains more followers to their network, the network becomes less dense as those friends/followers are less likely to become friends with the existing followers. Additionally these followers are unlikely to absorb or rehash information to their followers as networks with ties formed beyond 500 km are less likely to be transitive and less likely to function as a cohesive community. For some, Twitter encompasses the small town of the Internet where everybody knows and connects to everybody in the community. For others, Twitter represents the global city with lots of migrants and strangers, extending over a vast distance.

Our analysis demonstrates that the social and spatial relationships on the Internet are not distinctly separate. Material and digital relationships are intertwined and co-constructive. Twitter, as a networked community has a role in mimicking the relationships we form in material space. These networks are stronger and increasingly effective in more local geographies. Takhteyev et al. (2012) identified a large number of ties between cities at less than 3000 km apart and a rapid decrease of tie strength as distance increases beyond that threshold. We determined that networks among users at a distance less than 500 km are stronger with connections that more effectively disseminate information than networks extending beyond 500 km. This finding has notable implications for the use of Twitter among local organizations,

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002

governments, and media sources to communicate pertinent information to a geographically proximate population. As such, Twitter could serve as an effective tool for disseminating information during crisis situations.

While Twitter is one online social media platform, these findings are likely indicative of patterns present in other online networks. This analysis demonstrates how locality and spatial proximity still matter in the 21st century. Although technology enables individuals to build ties far beyond their local community to communicate with individuals across the globe, these ties are less frequent and weaker than local ties mimicking existing relationships.

Acknowledgements

We would like to thank Matthew Zook for providing excellent suggestions and feedback that enriched this project. This paper would not have been possible without the generous support of the New Maps Collaboratory and the department of Geography at the University of Kentucky.

References

Adams, P. C., & Ghose, R. (2003). India. com: The construction of a space between. *Progress in Human Geography, 27*(4), 414–437.

Bui, R., Buliung, R., & Remmel, T. K. (2011). *A collection of functions for estimating centrographic statistics and computational geometries for spatial point patterns*. Package for R. <http://cran.r-project.org/web/packages/aspace/index.html>.

Butts, C. T., Acton, R. M., Hipp, J. R., & Nagle, N. N. (2012). Geographical variability and network structure. *Social Networks, 34*(1), 82–100. <http://igraph.org>.

Cairncross, F. (1997). *The death of distance: How the communications revolution will change our lives*. Harvard Business Press.

Castells, M. (1996). *The rise of the network society*. Oxford: Basil Blackwell.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*.

Civin, M. A. (2000). *Male, female, email: The struggle for relatedness in a paranoid society*. New York: Other Press.

Crampton, J. W. et al. (2013). Beyond the Geotag: Situating 'big Data' and leveraging the potential of the Geoweb. *Cartography and Geographic Information Science, 40*(2), 130–139.

Crang, M., Crosbie, T., & Graham, S. (2007). Technology, time space, and the remediation of neighbourhood life. *Environment and Planning A, 39*(10), 2405–2422.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Inter Journal, Complex Systems*, 1695. <http://igraph.org>.

Davis, J. A., & Leinhardt, S. (1972). *The structure of positive interpersonal relations in small groups* (Vol. 2). Boston: Houghton Mifflin.

Faust, K., & Wasserman, S. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Friedman, T. L. (2007). *The world is flat: A brief history of the twenty-first century*. Farrar Straus & Giroux.

Fukuyama, F. (1992). *The end of history and the last man*. New York: Free Press.

Graham, S. (1998). The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography, 22*(2), 165–185.

Graham, M., Hale, S., & Gaffney, D. (2014). *Where in the World Are You? Geolocation and Language Identification in Twitter*. The Professional Geographer. http://dx.doi.org/10.1080/00330124.2014.907699.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 1360–1380.

Granovetter, M. S. (2005). The impact of social structure on economic outcomes. *The Journal of Economic Perspectives, 19*(1), 33–50.

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining twitter as an imagined community. *American Behavioral Scientist, Special issue on Imagined Communities, 55*(10), 1294–1318.

Hampton, K. (2004). Networked sociability online, off-line. In M. Castells (Ed.), *The network society: A cross-cultural perspective* (pp. 218). Cheltenham, UK: Edward Elgar.

Juris, J. (2004). Networked social movements: Global movements for global justice. In M. Castells (Ed.), *The network society: A cross-cultural perspective* (pp. 347). Cheltenham, UK: Edward Elgar.

Kent, J. D., & Capello, H. T. (2013). Spatial patterns and demographic indicators of effective social media content during the horse thief canyon fire of 2012. *Cartography and Geographic Information Science, 40*(2), 78–89.

Kitts, J. A., & Huang, J. (2010). Triads. In G. Barnett (Ed.), *Encyclopedia of social networks*. New York: Sage Publications.

Leamer, E., & Storper, M. (2001). The economic geography of the Internet age. *Journal of International Business Studies, 32*(4), 641–665.

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science, 40*(2), 61–77.

Li, F., Whalley, J., & Williams, H. (2001). Between physical and electronic spaces: The implications for organisations in the networked economy. *Environment and Planning A, 33*(4), 699–716.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the Digital Humanities*.

Morgan, K. (2004). The exaggerated death of geography: Learning, proximity and territorial innovation systems. *Journal of Economic Geography, 4*(1), 3–21.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Simmel, G., & Wolff, K. H. (1950). *The sociology of Georg Simmel*. Glencoe: Free Press.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter Networks [Special issue on Space and Networks]. *Social Networks, 34*(1), 73–81. http://dx.doi.org/10.1016/j.socnet.2011.05.006.

Tobler, W. R. (1970). Computer movie simulating urban growth in detroit region. *Economic Geography, 46*(2), 234–240.

Tsou, M., & Leitner, M. (2013). Visualization of social media: Seeing a mirage or a message? *Cartography and Geographic Information Science, 40*(2), 55–60.

Twitter Blog (2011). *One hundred million voices*. <http://blog.twitter.com/2011/09/one-hundred-million-voices.html>.

van Meeteren, M., Poorthuis, A., & Dugundji, E. (2009). Mapping communities in large virtual social networks.

Wellman, B. (1979). The community question: The intimate networks of East Yorkers. *American Journal of Sociology*, 1201–1231.

Wellman, B. (2001). Computer networks as social networks. *Science, 293*(5537), 2031–2034.

Wellman, B., & Gulia, M. (1999). Virtual communities as communities. *Communities in Cyberspace*, 167–194.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer. p. 5. [ISBN 978-0-387-98140-6].

Zook, M., Dodge, M., Aoyama, Y., & Townsend, A. (2004). New digital geographies: Information, communication, and place. *Geography and Technology* (pp. 155-176). Netherlands: Springer.

Zuckerman, E. (2008). Meet the bridgebloggers. *Public Choice, 134*(1–2), 47–65. http://dx.doi.org/10.1007/S11127-007-9200-Y.

Please cite this article in press as: Stephens, M., & Poorthuis, A. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.002