Proximity-Driven Motives in the Evolution of an Online Social Network

Ákos Jakobi

Abstract Although early theoretical works dealing with the effects of information communication technologies on people's relation to spatiality claim that distance is no longer important in the information age, there is a growing number of empirical results stressing on the contrary the importance of geographical factors. In the era of big data now we have the chance to give more insights on the geography of the internet-related social processes, since there are unprecedentedly large enough samples to analyse social behaviour as well as to understand the changing role of geography. Accordingly, the following paper is focusing on the geographical analysis of a nowadays very popular topic: the online social networks. Examples of iWiW, the largest Hungarian social media site, are applied to show that such networks are evolving and are structured not independently of spatial constraints. The paper attempts to present many proximity-driven characteristics of this network starting from distance-based examples of space-time evolution and with examples of proximity-focused statistical analysis of the spatial structure. The calculations highlighted that—although it was changing in time—proximity-driven processes have been predominant in city-to-city diffusion, especially when dealing with intracity spreading, but also at cases of short distance neighbourhood diffusion. Calculations by comparing factors like the average strength of connectivity, population size and average relative distance rates of cities also confirmed that proximity had an influence on the network structure.

Keywords Online social networks · Proximity · Distance · Network geography · Hungary

Á. Jakobi (✉)
Department of Regional Science, Faculty of Sciences, Eötvös Loránd University,
Pázmány Péter Sétány 1/c, Budapest 1117, Hungary
e-mail: jakobi@caesar.elte.hu

I. Ivan et al. (eds.), *The Rise of Big Spatial Data*, Lecture Notes
in Geoinformation and Cartography, DOI 10.1007/978-3-319-45123-7_15

1 Introduction

In the internet era, online social networks (OSN) are one of the major platforms of communication (see Lazer et al. 2009), supporting place-independent social life; however, recent findings suggest that geographical location of users strongly affect network topology (Takhteyev et al. 2012). Although on the one hand cyberspace is clearly present in the vanishing distance dependent costs of online telecommunication, leading to the claim of the "Death of Distance" thesis (Cairncross 1997), on the other hand the role of geographical location and distance is not clear at all regarding online communication and online involvement itself, because internet seems to stimulate local offline communication (Storper and Venables 2004) and users mostly interact with their strongly connected cliques but are also able to extend their interactions to more distant places than ever before (Wellman 2002). It seems that physical place and distance has a determining power on online communities (Liben-Nowell et al. 2005), and internet infrastructure (Tranos and Nijkamp 2012).

The above mentioned debate raised the question that to what extent an online social network is spatially bounded. Does proximity matters or OSNs are realized forms of absolute spatial independency? To see it clearer we should note that social network sites are supplemental forms of communication between people who have known each other primarily in real life (Ellison et al. 2006) and OSNs are "biased versions of real-life networks" (Ugander et al. 2011). We claim that virtual space and physical world are strongly interrelated, since it is assumed that flesh and blood users document their offline friendships in the online environment. According to this statement online social networks should be geographically determined, but it is still unclear what spatial motives are decisive in the formation of an OSN. Therefore, the following paper has the aim to give evidences how geography influences OSNs by analysing one of the most important spatial factors of network evolution: the proximity.

2 The Dateset

The following analysis was made by the application of data of the once largest Hungarian online social network site, named iWiW (International Who Is Who). The iWiW was launched in April 2002 and the service became highly popular and reached a few hundred thousands of people by 2005. By the introduction of new functions in 2005–2006 the number of registered users grew rapidly from 1.5 to more than 4 million until December 2008. Later, the website could not meet the challenges of competing with market-leading OSNs (namely Facebook) and, after a long declining period, the service was shut down on June 30, 2014.

The examinations were based on a data collection for January 2013 (and provided for research purposes by the data owner company). Location of users was defined by profile information, which is occasionally considered to be problematic in papers focusing on OSN user and social media content localization (Hecht et al. 2011). In iWiW, however, it was compulsory to choose a town of residence from a scroll-down menu when registering as user; thus, location is documented in every profile. This place could be easily changed afterwards and certainly there was no eligibility check. One might consider our location indicator based on user profiles a biased and occasionally updated census-type data. The geolocated individual user data have been summed up to the level of cities, and because user profile information also contained friendship data (data of people a user is in connection with), we could draw the connectivity network of the cities as well. This was registered in the database in forms of settlement pairs.

Altogether 2562 cities had active user data with a sum of 4,058,505 users. The users have established 785,841,313 friendship ties in the website, out of which 369,789,373 ties remained within settlement borders (considered as intra-city loops) and 415,653,749 ties were established between users from two distinct settlements. Concerning the city-level aggregated data, the network database covered 1,369,978 settlement-to-settlement pairs.

Additionally, data were appropriate to trace the evolution of the network in time. Until the very late periods of iWiW life-cycle, new users could register a profile only after an invitation had been sent from a member. The ID of the inviter was involved in each user's profile. Therefore we were able to trace the diffusion of iWiW across time and space because we know not only the location of each new user, but also the location from where the invitation was sent to each new user, and the timestamp of the acceptance of the invitation. In that way, we could investigate more than 2.7 million geo-located invitations between April 2002 and June 2012. Since our data was collected in 2013, inviter ID was missing in cases when the profile of the inviter was already deleted or the profile was registered after June 2012 when invitation wasn't needed anymore for registration, therefore the following analysis was performed for data between 2002 and 2012. Concerning the number of invitations from 2002 until the end of 2005 only 1–2 thousand invitations were sent by members per month. Then, the number of invitations jumped to more than 50,000 monthly and increased to a peak about 90,000 invitations per month until the middle of 2007. After that period invitations per month started to decrease rapidly. Although free registration was also introduced in 2012, this previously only mode of diffusion remained the major means of spreading.

3 Space-Time Evolution of the Network: The Importance of Proximity

According to the main findings of the international literature (Oh et al. 2008; Lan et al. 2011; Takhteyev et al. 2012) we assume that the evolution of our OSN has also followed certain geographical characteristics. Our general assumption is that the formation of new connections between old and new members is largely depending on distance between them. Although it is possible to get in connection with anyone in networks of the cyberspace, we still believe that the majority of new friendships are evolving among closely located people. Naturally, there are exceptions, but the share of random or not proximity-driven new connections is expected to be small.

On the other hand it is also presumed that the importance of proximity-driven formation of new connections could be different in certain time periods in the life of the OSN. We expect that at the beginning, when only few people are involved, the importance of close friendship is supposed to be larger, than at later periods, when the number of users is much higher and the chance of getting in connection with new distant acquaintances is also larger. All in all, we nevertheless assume the dominance of proximity-driven invitation processes throughout the whole examined period.

In our examination we followed the main concepts of spatial innovation diffusion theories (Hägerstrand 1967; Gould 1975), which highlight that many of the new things are spreading in space not randomly, but often as determined by geographical or other proximity factors (Boschma 2010). In that sense, new things appear first close to its origin, then in the next phase it appears at the closest neighbouring areas, while at a later period also distant places adopt the new thing. In other words, neighbourhood diffusion refers to the spreading when an innovation will likely be adopted first close to its source and later at greater distances following a distance-decay pattern. Concepts of this approach are widely applied in the literature (Cliff 1968; Johnston and Pattie 2011) confirming that there are many spatial evolutionary processes, which follow proximity rules.

To test our assumption we analysed the geo-located invitation data for each months. Since the location of both the inviter and the invitee were known in the dataset, it was easy to determine the geographical distance between the two people, or at least their two cities. The calculation of the distance between the city of the sender and the city of the receiver has been done for all invitation cases. Based on that, at first the average distance of invitations was calculated for each month (Fig. 1). Consequently, the smaller the average distances of invitations the higher the probability that neighbourhood diffusion has a large proportion within all diffusion cases.

Concerning the average distance of the invitations we found evidence of the assumption that proximity factors were changing in time. According to the results a continuous increase of distance values were observable until 2006 and a decrease afterwards. At the beginning of the examined period the average distance between the city of the inviter and the city of the invitee was slightly larger than 15 km. It reached as high as approximately 45 km in April 2006 and started to decrease and to stabilise later at around 20–25 km. We suppose that the noteworthy break of the trend after 2006 was in relation with the increase of the total number of new invitations.

In order to test the importance of proximity-driven diffusion forms, then we classified the invitation cases into proximity based (or short distance) and

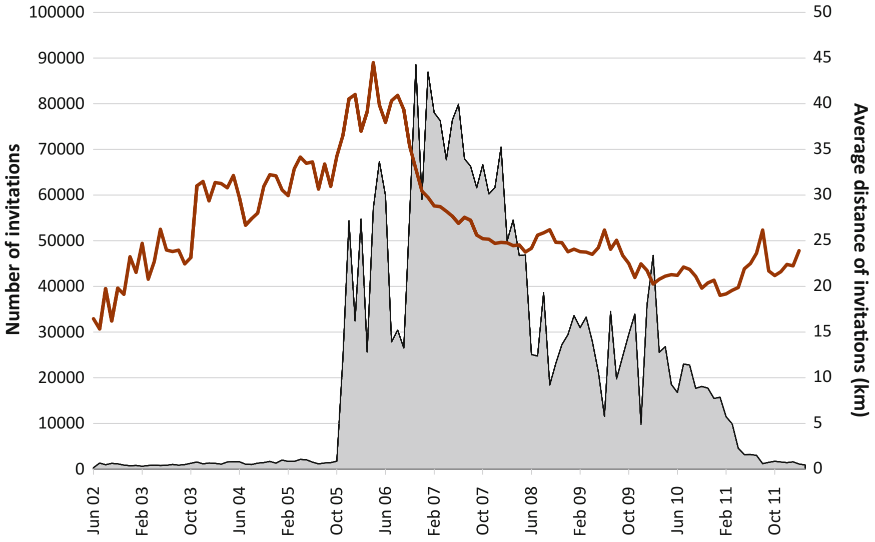

Fig. 1 The total number of iWiW invitations (*grey area*) and the average distance of iWiW invitations (*line*) by months (June 2002–March 2012)

non-proximity based (or long distance) categories. Since iWiW data were only possible to be geo-located on the level of cities (and not on the level of addresses or streets, etc.), zero distances were registered for invitations, where the inviter and the invitee were located in the same city. We consider these cases as intracity diffusion (or loop diffusion), which are in fact also proximity-based, since distances between inviter and invitee within a settlement are surely not zero. We know that there could be a distance-decay also within a city, however, our dataset was not able to detect intracity location differences. By the way, we still reckon loop diffusion as spreading to the closest distance, while the rest remained as cases of simple short distance or long distance diffusion.

It can be declared that intracity loop invitations happened to be the most prevalent category every time during the observed years. Invitations within the same city had always a dominant role, since the share of loops within all connections was permanently above 50 % (Fig. 2). Proportional values started from as high as 80 % and decreased to 52 % until 2006, when turned to rise again roughly until the end of the examined period. The curve more or less inversely followed the change of the total number of invitations.

Concerning the rest of the invitation data, we distinguished the group of short distance (but not loop) diffusion to the neighbouring zones and the group of long distance (not proximity-driven) spreading cases. In order to find short distance (neighbourhood) diffusion cases we had chosen threshold distance metrics instead of topological adjacency of cities, since it was assumed that virtual space connections took administrative topology less into account, distance on the other hand

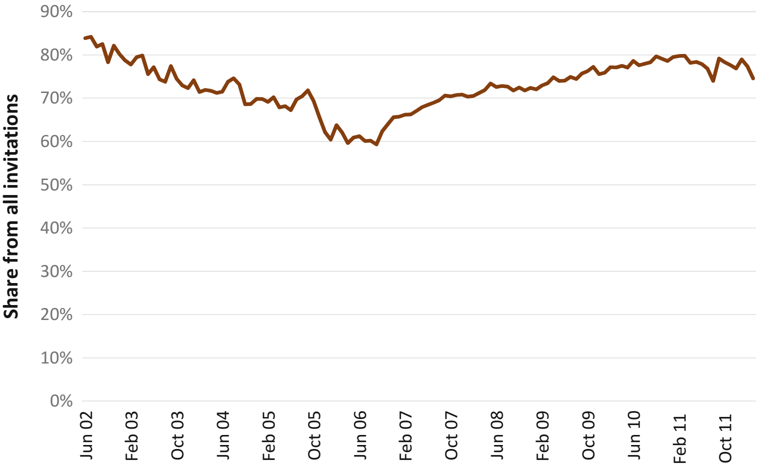

Fig. 2 The share of invitations sent within the same city (June 2002–March 2012)

seemed to play notable role. The threshold value was set to 15 km in line with results of calculations, where the probability of links as a function of distance had been calculated and the values returned a slight break of probability at around 15 km (Lengyel et al. 2015). This threshold possibly well separated neighbourhood diffusion zones from other spreading areas.

The share of invitations that were sent to a maximum of 15 km covered approximately 2–5 % of all cases in the early years (Fig. 3). Then, after 2006, the proportion of short distance invitations was doubled and stayed relatively high almost until the end. Consequently, the share of neighbourhood diffusion cases has

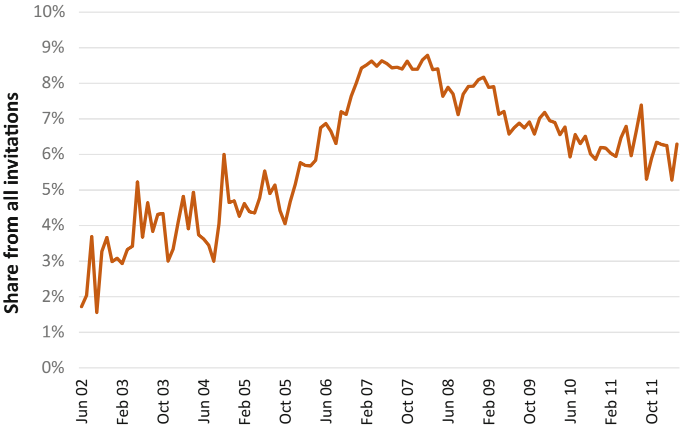

Fig. 3 The share of invitations sent to less than 15 km, intracity loops are excluded (June 2002–March 2012)

been considerably increased in the second half of the time. Apparently the proximity dimension was strengthened and became more important after 2006 resulting some decrease in the average distance of the invitations.

After taking all proximity-driven (loop and neighbourhood) diffusion cases away, the remaining city-to-city invitations were possible to be declared as independent of geographical distance. The overall share of such cases were approximately 15 % at the beginning, then increase to a top about 33 % around the year of 2006, while continuously decreased thereafter to a level of 15 % until the end of the examined period. Accordingly, the share of distance-independent cases has never been predominant during the years, and it seems that it could reach somewhat larger proportion only around 2006, when the annual number of new invitations was extraordinary high.

4 Proximity-Driven Motives in the Spatial Network Structure

Previously we have seen that proximity played an important role in the development of our online social network. It is a question, on the other hand, whether proximity has also a notable effect on user and connectivity patterns of the network structure. We assume that the locations, where iWiW was adapted earlier, would have higher rates of user penetration compared to those applying OSN services later. And because iWiW was firstly and primarily used by people in Budapest, we presumed higher user rates in the capital city (claimed as the city of origin) and smaller rates in farther distances from the city. In order to reveal such proximity-driven motives, we compared the rate of iWiW users and the distance from Budapest for each settlements.

According to the results a negative relationship could be found between the rate of users among the local population and the distance from Budapest, in which the departure from the experienced maximum level is, in fact, growing in negative terms (Fig. 4). Although the fitted linear regression model had not so large R-square results, the outcomes still reflect that proximity matters in OSN presence: as the distance increases, the probability of a lower user rate increases.

Although data of the rate of users revealed that this online social network can not be considered as aspatial, naturally, there are other geographical motives possible to be explored in a network structure, especially in connection with network ties. Since this OSN dataset is a typical example of big data, we may assume that the entities or nodes (in our case the localities, namely the cities) have the chance to get network connections with almost all the others. It is also known, however, that some of the cities have large number of connections with high variety, while others are rather connected to only few cities. The possible determining factor behind is

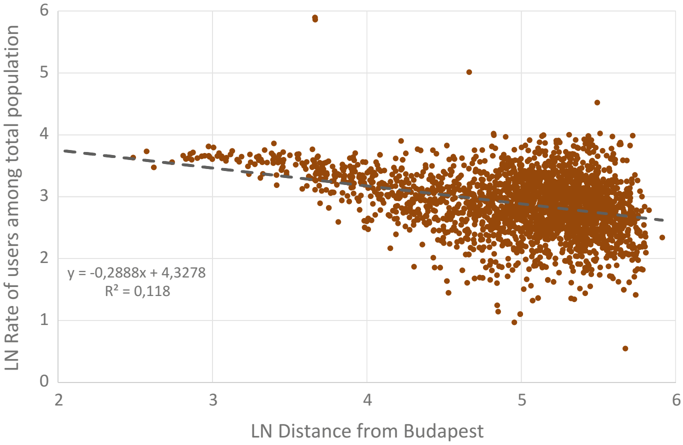

Fig. 4 The connection between the rate of iWiW users and the distance from Budapest. All variables are transformed to natural logarithm values

the size of the city, since large cities could have more connections and more divers network structure than small villages with only a couple of users, who are connected with users from only a few other places. On the other hand it is assumed, that cities having connections with not many other cities are possibly tied to others stronger than those having connections with large pool of cities.

In order to deal with different strength of network connections between cities we compared the ratio of the observed and randomly expected city-to-city connection weights for each pair of cities. The observed or raw weights have been calculated as the sum of connections between users of the two cities (Eq. 1):

$$w_{ij} = c_{ij} + c_{ji} \tag{1}$$

where w_{ij} is the observed (or raw) weight of connections between city i and j, c_{ij} is the number of connections between users, who are located in city i and have friends in city j, while c_{ji} is the number of connections between users, who are located in city j and have friends in city i.

The expected city-to-city connection weights have been calculated as follows (Eq. 2):

$$e_{ij} = \frac{s_i s_j}{\sum_{i=1,j=1}^{n} w_{ij}} \tag{2}$$

Here $s_i = \sum_j^n w_{ij}$ is the strength of node i, namely the total number of connections in the city, and e_{ij} is the expected number of links between cities i and j based purely on the total number of links at those cities assuming random tie formation.

Finally we calculated the log likelihood ratios of the above detailed observed (or raw) and randomly expected components (Eq. 3):

$$LLR_{ij} = Log\left(\frac{w_{ij}}{e_{ij}}\right) = Log\left(w_{ij} / \frac{s_i s_j}{\sum_{i=1,j=1}^{n} w_{ij}}\right) \quad (3)$$

in which LLR_{ij} refers to the log-likelihood ratio between settlement i and j. Note that LLR_{ij} can be negative or positive depending on the ratio of the measured weight and the expected one. The higher positive LLR refers to strong city-to-city ties, while negative LLR represents weak intercity connections.

As mentioned above, we assume that geography plays a role in connectivity strength between settlements. When looking at the scatter plot of average strength of connectivity (average LLR score by settlements) against the size of the cities (natural logarithm of population), it could be definitely noticed that the smallest settlements have on average the strongest connections with others (top left on Fig. 5). Such cities are typically connected to only few number of other cities but with strong relations.

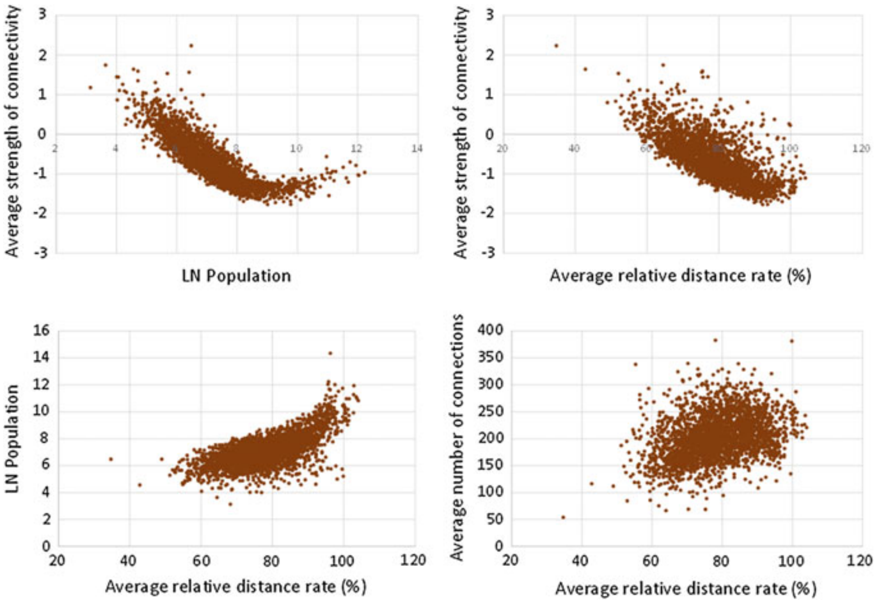

Fig. 5 *Top left* connection between the average strength of connectivity (average LLR score) and size (natural logarithm of population). *Top right* connection between the average strength of connectivity (average LLR score) and average relative distance rate. *Bottom left* connection between size (natural logarithm of population) and average relative distance rate. *Bottom right* connection between average number of connections and average relative distance rate

For ourselves it is nevertheless more interesting that also distance has a possible influence on the average strength values of cities. But we would get a biased picture if absolute distances between connected cities would have been applied, since absolute distance of connections largely depends on central or peripheral geoposition of a settlement, instead we suggest to use relative distance rates for non-biased network proximity. Average relative distance rates were calculated by the comparison of the observed and expected distance averages (Eq. 4):

$$ARD_i = \left(\frac{\sum_{i=k}^{k} d_{ij}}{k} \Bigg/ \frac{\sum_{i=1}^{n} d_{ij}}{n} \right) \times 100 \tag{4}$$

in which ARD_i refers to the average relative distance rate of settlement i, d_{ij} is the distance between settlement i and j, k is the observed number of connected settlements and n is the total (or expected) number of settlements.

By the comparison of average strength of connectivity (average LLR score) and average relative distance rates of cities (ARD), we should declare that the closer associates a city generally owns, the stronger the connections it has on average (top right on Fig. 5). The fitted linear regression model resulted a significant R-square above 0.5. Consequently it seems that the tightest and strongest network connections do not stretch too far. Additionally, it is also observable that the larger the city, the farther its connections are reaching on average (bottom left on Fig. 5). Finally, by the comparison of the average number of connections and average relative distance rates (ARD) we could more or less notice that the cities, which have more distant connections in general, are having users typically with larger number of friendships (bottom right on Fig. 5). This relationship is, however, less significant than the previously detailed ones, since the scatter plot reflects evidently larger standard deviation and also linear R-square results happened to be small (0.1).

Examples of the analysis of individual cities also confirmed that tighter connections in virtual space are falling in line with short distances in real physical geography, even though it is in principle the same simple to access any points in cyberspace. The example of Herend, a middle-sized city, well reflects that cyberspace is not independent from constraints of real physical space, although it is also observable on the picture that there is the chance to have connections with distant cities as well (Fig. 6). This city has basically strong connections with close cities (generally less than 40 km), but some strong connections are from larger distances. Based on that the relationship between distance and connectivity weight (strength) is not deterministic rather stochastic.

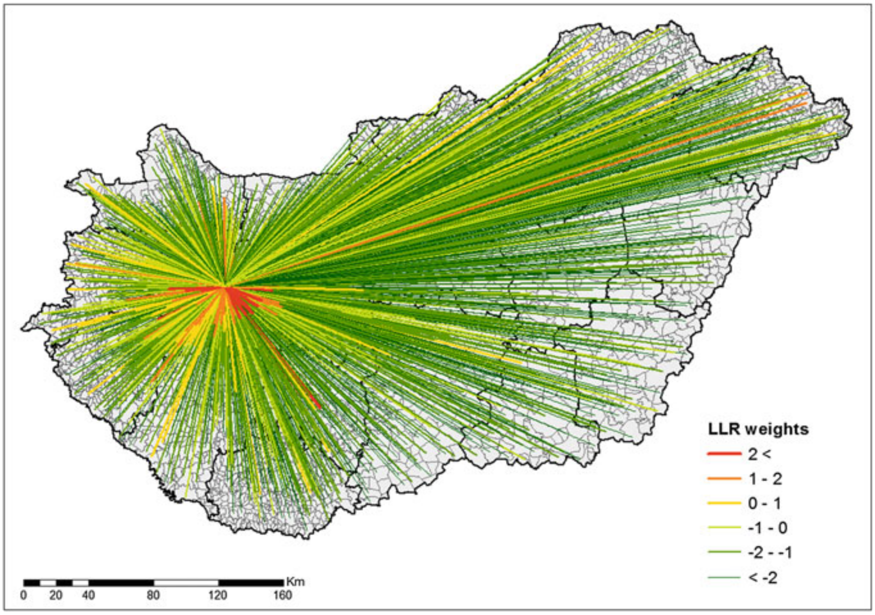

Fig. 6 The connectivity network or Herend according to the strong and weak connections (based on LLR scores)

5 Summary and Conclusions

OSNs are large-scale networks in which users are the nodes and their connections with other users are the edges. They are also defined as web-based services that 'enable users to articulate and make visible their social networks' (Boyd and Ellison 2007, p. 212.). The definition claims that OSNs are supplemental forms of communication between people who have known one another primarily in real life. In other words, major OSNs are not used to meet new people but rather to articulate relationships with people in their existing offline network. Furthermore, the degree distribution of online social networks, like Facebook, is very close to the degree distribution of real-life social networks (Ugander et al. 2011), in other words, OSNs clearly differ from other web-based networks, such as internet infrastructure. The latter are led by power law tie-distribution: a small share of webpages accounts for an outstandingly high number of links (Barabási and Albert 1999). In our understanding, OSNs are showing strong geography-related network characteristics and not a typical power law pattern.

The paper demonstrated that OSNs are definitely place-dependent, because many aspects of network connectivity happened to show geographical relatedness as well. Despite the fact that online social networks are virtual creations it was found that diffusion processes related to network evolution are not independent of spatiality. It was pointed out that proximity-driven diffusion processes are predominant, especially when dealing with intracity spreading, but also at cases of short distance neighbourhood diffusion. Although cyberspace allows creating connections independently of distance, the majority of new registrations arose in the geographic vicinity of earlier ones.

Also by determining the strongest and most important city-to-city connections or the average relative distance rates it turned out that distance does significantly matter in network formation. Many of the network structure characteristics were happened to be proximity-driven. As a combined result we should claim that big datasets of online social networks now has a good chance to give evidence on why geography matters (de Blij 2012) in the information age.

References

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Boschma R, Frenken K (2010) The spatial evolution of innovation networks. A proximity perspective. In: Boschma R, Martin R (eds) The handbook of evolutionary economic geography. Edward Elgar, Cheltenham, pp 120–135

Boyd D, Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput Mediat Commun 13:210–230

Cairncross F (1997) The death of distance. How the communication revolution will change our lives. Harvard Business School Press, Boston

Cliff AD (1968) The neighbourhood effect in the diffusion of innovations. Trans Inst Br Geogr 44:75–84

de Blij H (2012) Why geography matters: more than ever, 2nd edn. Oxford, New York

Ellison N, Steinfeld C, Lampe C (2006) Spatially bounded online social networks and social capital: the role of Facebook. Paper presented at the annual conference of the international communication association, Dresden, 19–23 June. http://www.ucalgary.ca/files/stas341/Facebook_ICA_2006.pdf. Accessed 30 Nov 2015

Gould PR (1975) Spatial diffusion: the spread of ideas and innovations in geographic space. Learning Resources in International Studies, New York

Hägerstand T (1967) Innovation diffusion as a spatial process. University of Chicago Press, Chicago

Hecht B, Hong L, Suh B, Chi EH (2011) Tweets from Justin Bieber's heart: the dynamics of the "location" field in user profiles. In Proceedings of the ACM conference on human factors in computing systems. ACM Press, New York, pp 237–246

Johnston RJ, Pattie C (2011) Social networks, geography and neighbourhood effects. In: Scott J, Carrington P (eds) The SAGE handbook of social network analysis. SAGE Publications Ltd, London, pp 301–311

Lan T, Lan C, Tserendondog O (2011) Analysis of social network sites diffusion in Mongolia. Afr J Bus Manag 5(23):9889–9895

Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Computational social science. Science 323(5915):721–723. doi:10.1126/science.1167742

Lengyel B, Varga A, Ságvári B, Jakobi Á, Kertész J (2015) Geographies of an online social network. PLoS One 10(9):e0137248. doi:10.1371/journal.pone.0137248

Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102:11623–11628

Oh J, Susarla A, Tan Y (2008) Examining the diffusion of user-generated content in online social networks. SSRN eLibrary. doi:10.2139/ssrn.1182631. Accessed 27 Nov 2015

Storper M, Venables A (2004) Buzz: face-to-face contact and the urban economy. J Econ Geogr 4:351–370

Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. Soc Netw 34:73–81

Tranos E, Nijkamp P (2012) The death of distance revisited: cyberplace, physical and relational proximities. Working Paper Tinbergen Institute, TI 2012-066/3. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2103024. Accessed 20 Nov 2015

Ugander J, Karrer B, Backstrom L, Marlow C (2011) The anatomy of the Facebook social graph. http://arxiv.org/abs/1111.4503. Accessed 30 Nov 2013

Wellmann B (2002) Little boxes, glocalization, and networked individualism. In: Tanabe M, van den Besselaar P, Ishida T (eds) Digital cities II: computational and sociological approaches. Springer, Berlin, pp 10–25