The elliptic model for social fluxes

C Herrera-Yagüe[1,2,3], CM Schneider[1], Z Smoreda[4], T Couronné[4], PJ Zufiria[2,3] and MC González[1]

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

[2]Depto. Matemática Aplicada a las Tecnologías de la Información, ETSI Telecomunicación, Universidad Politécnica de Madrid (UPM), Spain

[3]Cátedra Orange. Universidad Politécnica de Madrid (UPM), Spain

[4]Sociology and Economics of Networks and Services department, Orange Labs, Issy les Moulineaux, France

E-mail: carlos@hyague.es

Abstract. We analyze the anonymous communications patterns of 25 million users from 3 different countries. We assign a location to each user based on her most used phone tower or billing zip code. With this information, we build spatial social networks at three levels of resolution: tower, city and region for each of the three countries. We propose an elliptic model, which considers the number of relationships between two locations is reversely proportional to the population in the ellipse whose focuses are in such locations. We compare the performance of this model to analogous models of transportation fluxes and find that the elliptic model outperforms them in all scenarios. This shows that human relationships are at least as influenced by distance as human mobility is, thus spatial models can still be corrected when paying attention to some subtle differences to the geometry of the underlying processes.

1. Introduction

While social networks have been known for years to play a key role in various human phenomena [1, 2], only recently it was possible to map large social networks in order to explore how their structures influence processes occurring in the network. These large social network data sets, usually coming from telecommunication records originated from e-mail [3], phone [4] or online communication platforms [5], have been used to explore a wide range of topics such as adoption of innovation [6], social groups discovery [7, 8, 9], epidemic spreading [10, 11, 12], social mobilization [13] or sentiment spreading [14].

Despite the publication of such studies, network data is not widely available to the community due to either privacy or commercial issues. In addition, even with access to the electronic records, extracting a meaningful social network may be difficult at a large scale [4]. For these reasons, the creation of models that are able to mimic different social network properties have recently attracted a fair amount of research interest [15, 16, 17, 18, 19]. While most models try to generate synthetic networks with some desired characteristics (degree distribution and clustering coefficient among others), in this paper we will focus on reproducing a macroscopic feature of real social networks: friendship fluxes within different locations, this is, how many friendships exist between two cities, two regions or even two neighborhoods.

The effect of geography on social networks

During the end of the 19th century and the beginnings of the 20th, a considerable amount of effort was dedicated to the development of telecommunication systems. Such systems, whether they carried written messages (telegraph) or voice (telephone) were designed to achieve single goal: allowing people to communicate with those who are far away (indeed Greek prefix *tele-* means *distant*). Interestingly, massive records of such systems obtained a hundred years later show that people do not commonly use them to talk to those far away, but with people who are actually close by. Precisely, it has been found consistently across records from emails, phones and blogs that the probability of a communication to happen between two people who are r kilometers from each other follows a decay function, typically a power law [20, 21, 22].

Although geographic friendship fluxes have not been the focus of much research to date, this new evidence shows friendship fluxes behave similar to trip fluxes and other phenomena driven by the population density. In transportation research, flux prediction is a well-defined problem: given a set of locations $\{i, j\}$ whose coordinates and populations $\{n_i, n_j, ...\}$ are known, the goal is inferring the flux matrix T where each element T_{ij} represents the number of trips from location i to location j. The problem was traditionally approached using gravity models [23, 24, 25] which try to gather the effect of decaying probability with distance following equation

$$T_{ij} \propto \frac{n_i^{\alpha} n_j^{\gamma}}{f(r_{ij})}$$

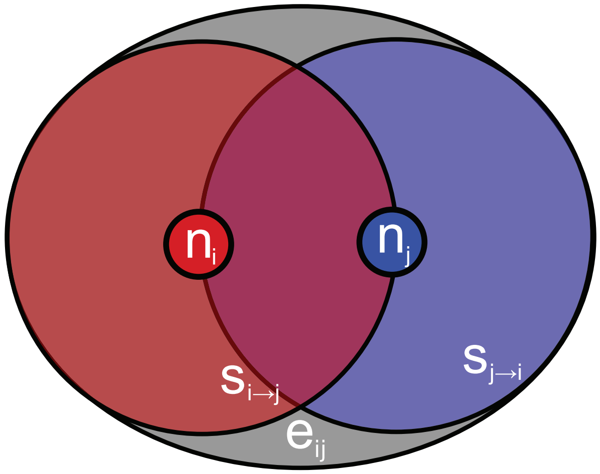

Figure 1. Model scenario. n_i represents the population of city i while $s_{i\to j}$ represents the population within the circle with center in i and radius up to j. As long as population is not homogeneously distributed $s_{i\to j} \neq s_{j\to i}$, the radiation model predictions will not be symmetrical. e_{ij} represents the population within the smallest ellipse whose focuses are in i and j and contains both previous circumferences, as well as n_i and n_j.

where α and β are fitting parameters usually got from training data, and $f(r_{ij})$ increases with distance, typically following an exponential or power-law function. A powerful idea was brought recently by the radiation model [26], which claims that is not the distance matters, but the amount of opportunities between i and j, that can be estimated by the population in the area. In short, the authors explain someone from rural Iowa is more likely to travel further to satisfy her needs than someone in New York City, given the latter has a handful of options just a few blocks from her. In this paper we will present a model derived from this radiation model which is able to predict friendship fluxes at least as good as current transportation models predict trips.

2. Model description

Formally, radiation model estimates fluxes T_{ij} between to locations i and j using population in both locations, and population within the circle whose center is i and radius up to j. Its formulation is

$$T^{rad}_{i,j} = K_i \frac{n_i n_j}{(n_i + s_{i\to j})(n_i + n_j + s_{i\to j})}$$

where n_i represents the population of location i, $s_{i\to j}$ the number of people who are not in i and closer to i than j and the normalization $K_i = n_i \frac{N_T}{N}$, where N_T is sum of all fluxes $N_T = \sum_i \sum_j T_{ij}$ and N the total population $N = \sum_i n_i$.

It is straightforward to check T^{rad} matrices will not be symmetrical in general, because $s_{i\to j} \neq s_{j\to i}$. While asymmetry is a desirable feature for mobility models (think of commuting's origin-destination matrix, with strongly asymmetric suburbs-downtown fluxes) it is not when dealing with friendship, which are commonly considered to be

Country	Users U	Links E	$\langle k \rangle$	$\langle c \rangle$	$\frac{N}{Total\ Population}(\%)$
France	$18.7 \cdot 10^6$	$81.3 \cdot 10^6$	8.73	0.16	30.21
Portugal	$1.21 \cdot 10^6$	$4.00 \cdot 10^6$	6.57	0.26	11.21
Spain	$5.92 \cdot 10^6$	$16.1 \cdot 10^6$	5.44	0.21	13.45

Table 1. Characteristic properties of the social networks in the studied countries: number of users (Nodes) and relationships (Links), average degree $\langle k \rangle$, average clustering coefficient $\langle c \rangle$ and relative sample size of the users in the data set.

mutual [27]. In fact, reciprocity became standard to filter out spurious links (marketing callers, misdialed numbers...) when extracting social networks from phone records [21, 4, 28].

Our model, which we will refer as elliptic model (EM), is oriented to model social relationships. Given those are usually understood as mutual, the predicted flux matrix T should be symmetrical. Therefore, EM considers the probability of someone living at location i having an acquittance at location j is reversely proportional to the population of the area where both could meet without their combined traveling exceeds a certain distance. This area forms an ellipse whose focuses are in locations i and j (see figure 1 for graphic explanation and comparison to the radiation model). Thus, EM formulation is

$$T_{ij}^{ellip} = K\frac{n_i n_j}{e_{ij}}$$

where e_{ij} is the population within the ellipse depicted in figure 1 (note e_{ij} includes n_i and n_j). Since $e_{ij} = e_{ji}$, $T_{ij} = T_{ji}$ and thus our model will always produce symmetrical matrices T. Note that the ellipse contains previous circles so $e_{ij} \geq (s_{i\to j} \cup s_{j\to i} + n_i + n_j)$ and it is smallest containing both, so usually $e_{ij} \simeq (s_{i\to j} \cup s_{j\to i} + n_i + n_j)$. During the discussion we will compare the proposed model to a trivial symmetrization of radiation, which we will denote *radBI* and whose formulation is

$$T_{ij}^{radBI} = \frac{1}{2}(T_{ij}^{rad} + T_{ji}^{rad}).$$

3. Data description

To evaluate the performance of EM, we test it versus a mobile phone data set containing records of a six month period in 3 different countries: France, Portugal, and Spain. In total, over 7 billion calls are considered to build the social graphs, whose links are included only if there is at least one call per direction during the observation period thus producing an undirected graph. In Table 3, some characteristics of the networks are presented. These characteristics, like high clustering and relatively low average degree, are expected from previous literature about mobile phone networks.

Additionally to the communication records, our data includes a location for each user: most used mobile phone tower in France and Portugal and billing zip code in Spain. In order to benchmark multi-scale performance of EM, in total 3 aggregation levels

Country	Locations	Cities	Regions
France	17475	3520	96
Portugal	2209	297	20
Spain	8928	5446	52

Table 2. Number of locations considered in different geographic aggregation levels for each country. Locations are mobile phone towers in France and Portugal and zip codes in Spain. Aggregation is based on administrative boundaries: cities are *cántons* in France, *concelhos* in Portugal and *municipios* in Spain while regions mean *départements* in France, and *provincias* in Portugal and Spain.

have been used: country-wide fluxes between cities and regions and on the other hand metropolitan fluxes within cities. Table 3 presents the number of locations considered in each aggregation level. When applying these spatial aggregations the center of mass of the population was used as the higher level location, instead of the centroid of the region polygon, in order to avoid undesirable effects in the fairly common case of a big city located in a corner of a polygon.

4. Friendship fluxes in country scale

To validate the predictions of EM at large scale, we attempt to reproduce flux matrices T in two aggregation levels. At the region level T has thousands of elements while at the city level there are tens of millions of fluxes to predict (see table 3). Input data for the predictions only consists of the location's coordinates and populations, and the total number of fluxes to predict N. Note that just like radiation model, EM keeps the advantage of being parameter-free, so no training data is needed.

In figure 2 we present a box-plot of the predictions from all the three models versus real data for city-to-city fluxes. Results prove consistently that EM outperforms both radiation model and its bilateral version. To present further evidence of the enhaced performance of EM, we include in table 3 the r-squared of the predictions in both aggregation levels. Results confirm EM overcomes previous models, and indeed reaches accuracy levels even higher than mobility models when predicting trips. This is an interesting finding, since in principle, while there is an increasing cost (like time and energy) associated to distance with traveling, there is not such cost when calling someone who is farther.

5. Friendship fluxes within cities

Next step is studying friendship within the finer spatial aggregation level available: phone towers or zip codes. While it would be possible predicting all possible tower to tower relationships within the country, such T matrix would have up to 300 million elements with only less than 1% of them not null. Thus, prediction accuracy would be severely biased by the huge amount of zero cells. Instead, we study the short range

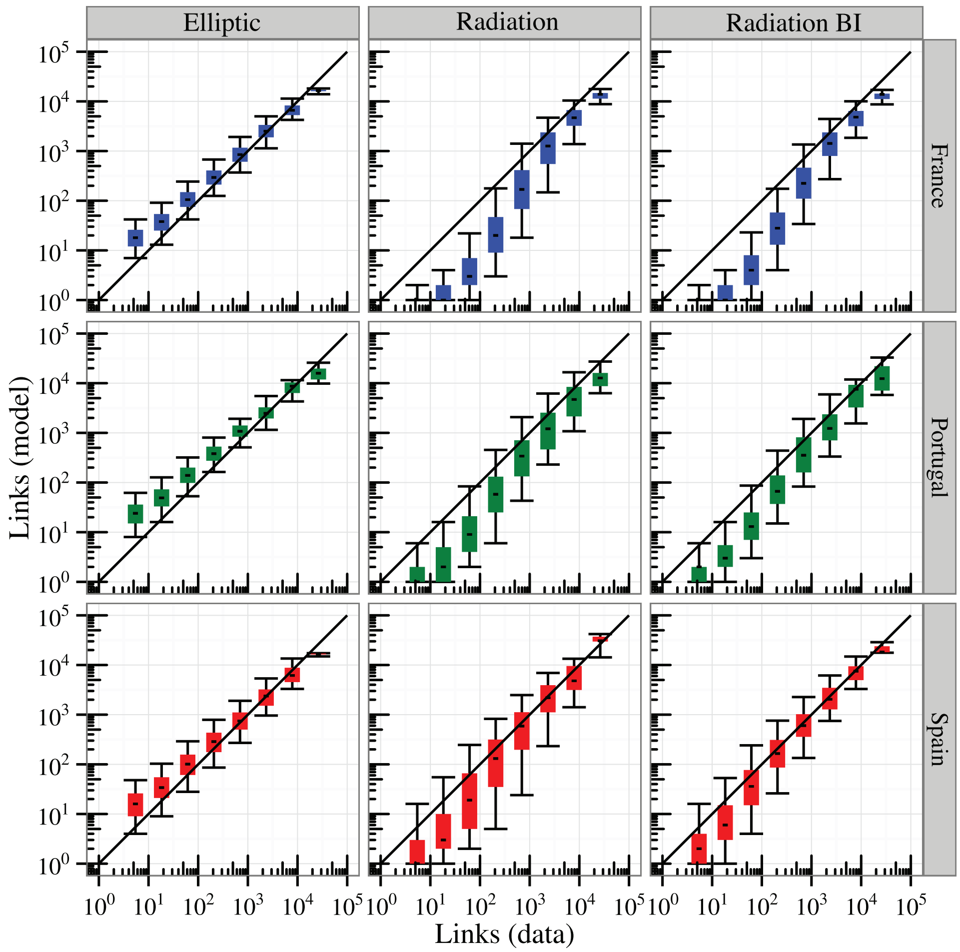

Figure 2. Predictions by different models versus real data. In this figure, fluxes between every city are presented. We consider 297 cities in Portugal, 5446 in Spain, and 3520 in France. Error bars plot 10%, 30%, 50%, 70% and 90% quantiles. Elliptic model overcomes both radiation and bilateral radiation in all three scenarios.

	France		Portugal		Spain	
	City	Province	City	Province	City	Province
Radiation	0.534	0.615	0.621	0.776	0.556	0.588
RadiationBI	0.626	0.723	0.730	0.847	0.676	0.668
Elliptic	0.723	0.790	0.816	0.891	0.693	0.748

Table 3. R^2 of the different countrywide predictions. Note that these R^2 are calculated without any logarithmic transformation on data or predictions. Number of provinces considered is 97, 20 and 52, respectively. Since number of cities is up to two orders of magnitude larger, flux matrix T is up to 4 orders of magnitude larger. While elliptic model is always more accurate than previous models predictions, the improvement is specially remarkable in city to city fluxes.

	France	Portugal	Spain
Radiation	0.377	0.527	0.434
RadiationBI	0.436	0.608	0.498
Elliptic	0.653	0.658	0.501

Table 4. Average R^2 for urban fluxes prediction for every city in the data set where we got at least 20 different locations (towers or zip codes). Number of locations range from this 20 up to 1000 in Paris. This sums up to 40 cities in France, 29 Spain and 20 in Portugal. Although elliptic again outperforms previous models, performance is small compared to country-wide scenarios.

accuracy of the model by applying it in every city where we got at least 20 different locations (upper limit being Paris, where we have 1000 mobile phone towers). This sums up to 40 cities in France, 29 Spain and 20 in Portugal.

Applying previous algorithms, the results in table 4 are reached, which presents them in terms of average R^2. Results confirm that EM outperforms the other models, while the overall prediction accuracy is smaller compared to country-wide experiment. This loss of accuracy within urban areas for any model purely based in distance is expected and observed in transportation field [29]. Among the reasons, the fact of the distance being a poorer proxy for travel time or cost in city arises. Within cities we tend to be within our daily radius of action and the decision of who we befriend depend on other metrics related to the different hierarchies that could define a social distance [21].

Flexibility ε as model improvement for urban areas

If one thinks of the model as presented in figure 1 when applied to urban relationships, one could realize a tower k whose distance to tower i is $r_{ik} = r_{ij} + \varepsilon$ where $\varepsilon \ll r_{ij}$ will not be taken into account in the prediction. Since within cities this is quite common, because towers tend to be closer to each other in urban areas, we propose the correction in figure 3 for urban environments. The variation consists of including a flexibility parameter ε so that the focal distance f_{ij} of the is $f_{ij} = r_{ij} + \varepsilon$. We have tried several values of ε in our data set and we found prediction accuracy peaks near $\varepsilon = 1$km for nearly all the cities (as shown in figure 3). The interpretation of such maximum may varies: one could argue it comes from the location error, known to be close to the average distance to neighborgs in the voronoi tessellation [30], which is around 1 km in average in our dataset. However we find even in those cities were the cells are really small, such is Paris, where $r_{voronoi} = 300$m, still the optimal occurs around 1km. So an alternative explanation arises: 1km, or 10 minutes walk, is the maximum distance people consider *roughly the same* when forming acquittance. Besides, when applied back to country-wide scenarios we found flexibility does not improve the predictions and no peak emerges near $\varepsilon \backsimeq r_{voronoi}$ (note that in this case, $r_{voronoi} \gg 1km$, being in the tens of kilometers for cities and in the hundreds for regions).

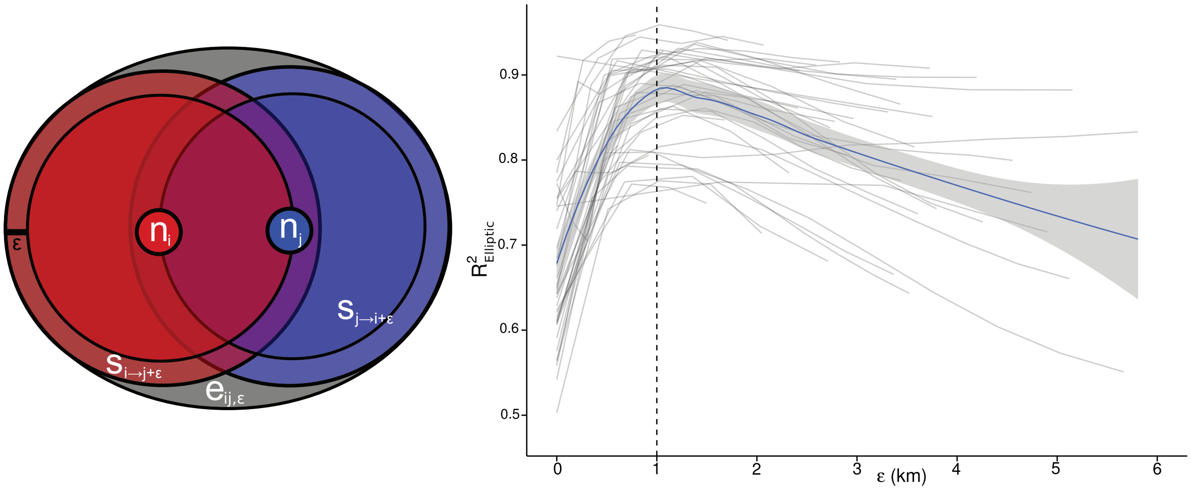

Figure 3. Model modified for intracity predictions, adding the flexiblity parameter ε. We find predictions improve when some flexibility is included, reaching a maximum around $\varepsilon = 1$km.

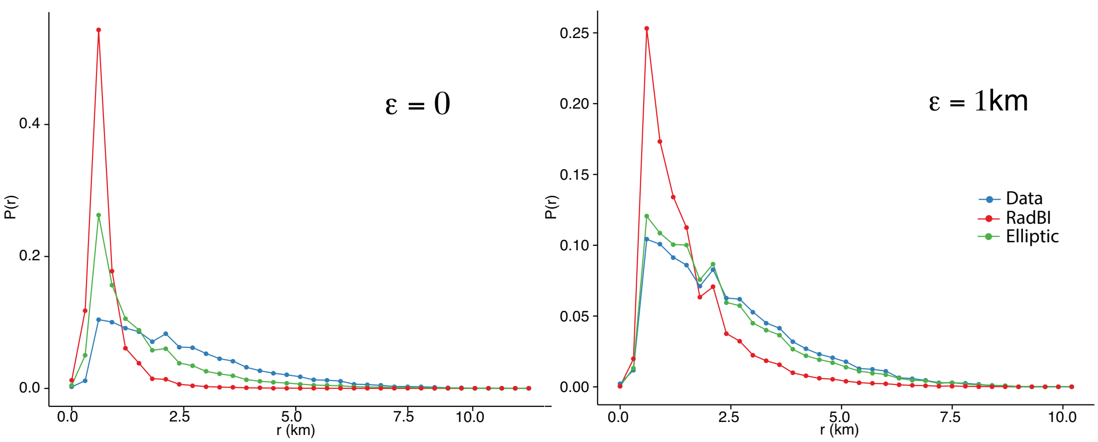

Figure 4. In the left we present the fraction of friendships $P(r)$ within distance r in the real dataset compared to predictions by both elliptic and bilateral radiation models where $\varepsilon = 0$ for Porto (Portugal). In the right we show how elliptic prediction becomes almost identical when using $\varepsilon = 1$km. Although radiation model predictions also improve, they still predict shorter fluxes than those observed in reality.

Additionally, figure 4 shows the improvement with $P(r)$. Without flexibility, short-range friendships are over represented, while EM with flexibility fits almost perfectly with the distribution obtained from the data. Note that although radiation model predictions also improve, it still predicts shorter fluxes than those observed in the data. Table 5 shows results of the corrected model for urban environments in terms of average R^2, which confirm a significant performance increase when including using $\varepsilon = 1$ km across all cities in the data set.

	France	Portugal	Spain
Elliptic $\varepsilon = 0$	0.670	0.645	0.494
Elliptic $\varepsilon = 1\,\mathrm{km}$	0.846	0.740	0.688

Table 5. In the table we present the average R^2 of the predictions when using the corrected model with $\varepsilon = 1$km compared to original model.

6. Conclusions

In short, we apply a transportation framework to the problem of predicting social fluxes between different locations. As a result, we propose a model to calculate the friendship fluxes using only population distribution data, which is available worldwide through projects such as Landscan, which provide population estimates for almost every square mile on earth.

We find that predicting the number of social links with our model works on different scales, from provinces to cities, and in 3 different reciprocal call graphs at national scale. Interestingly, we show that predicting social fluxes from population data is at least as easy as predicting trip fluxes, which is somehow unexpected; because while trips have obvious distance dependent costs (time, energy...), phone calls do not. The essence of the elliptic model is to take into account that the number of acquaintances is mutual, and the predicted flux matrix should be symmetrical, which is not the case for trips, which are based on the distribution of opportunities or facilities.

Taken together, our model is readily available to be used by researchers in different social sciences to study the various phenomena where human ties are known to be crucial like in information propagation or disease spreading. Overall, our model implies friendships are to a large extent driven by geographically factors, thus deviations from the model could point interesting facts in social networks: one could expect national or language borders (such as the example of Belgium presented in [31]) will produce large deviations from the model, so the model could become itself a useful tool to discover anomalous patterns of social connections between regions.

At last, to enhance the employment of EM by the research community, we have made available on our homepage [32] implementations in three widely used programming languages, as well as an interactive tool with France *départments* as an example scenario.

7. References

[1] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[2] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[3] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[4] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[5] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[6] Jameson L Toole, Meeyoung Cha, and Marta C González. Modeling the adoption of innovations in the presence of geographic and media influences. *PloS one*, 7(1):e29528, 2012.

[7] Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE, 2012.

[8] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[10] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[11] Pu Wang, Marta C González, Cesar A Hidalgo, and Albert-László Barabási. Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–1076, 2009.

[12] Christian M Schneider, Tamara Mihaljev, Shlomo Havlin, and Hans J Herrmann. Suppressing epidemics with a limited amount of immunization units. *Physical Review E*, 84(6):061911, 2011.

[13] Galen Pickard, Wei Pan, Iyad Rahwan, Manuel Cebrian, Riley Crane, Anmol Madan, and Alex Pentland. Time-critical social mobilization. *Science*, 334(6055):509–512, 2011.

[14] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ: British medical journal*, 337, 2008.

[15] Duncan J Watts and Steven H Strogatz. Collective dynamics of "small-world" networks. *nature*, 393(6684):440–442, 1998.

[16] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[17] Brian Karrer and MEJ Newman. Random graphs containing arbitrary distributions of subgraphs. *Physical Review E*, 82(6):066118, 2010.

[18] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2):026107, 2002.

[19] Carlos Herrera and Pedro J Zufiria. Generating scale-free networks with adjustable clustering coefficient via random walks. In *Network Science Workshop (NSW), 2011 IEEE*, pages 167–172. IEEE, 2011.

[20] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[21] Carlos Herrera-Yagüe, Christian M Schneider, Thomas Couronne, Zbigniew Somoreda, Rosa M Benito, Pedro J Zufiria, and Marta C González. Understanding social structures behind the six degrees of separation. *under review*, 2013.

[22] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[23] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[24] Rosalind M Eggo, Simon Cauchemez, and Neil M Ferguson. Spatial dynamics of the 1918 influenza pandemic in england, wales and the united states. *Journal of The Royal Society Interface*, 8(55):233–243, 2011.

[25] José J Ramasco and Alessandro Vespignani. Commuting and pandemic prediction. *PNAS*,

106(51):21459–21460, 2009.
[26] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.
[27] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.
[28] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. *PLoS one*, 6(4):e16939, 2011.
[29] Yingxiang Yang, Carlos Herrera-Yagüe, Nathan Eagle, and Marta C González. A multi-scale study of commuting patterns incorporating digital traces. *under review*, 2013.
[30] Michael Ulm and Peter Widhalm. Properties of the positioning error of cell phone trajectories. *under review*, 2013.
[31] Vincent Blondel, Gautier Krings, and Isabelle Thomas. Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. *Brussels Studies*, 42(4), 2010.
[32] http://humnetlab.mit.edu/elliptic. 2013.