J. Stat. Mech. (2014) P04022

The elliptic model for communication fluxes

C Herrera-Yagüe[1,2,3], C M Schneider[1], Z Smoreda[4], T Couronné[4], P J Zufiria[2,3] and M C González[1]

[1] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] Departamento 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, cschneider84@gmx.de, zbigniew.smoreda@orange.com, thomas.couronne@telenor.com, pedro.zufiria@upm.es and martag@mit.edu

Abstract. In this paper, a model (called the elliptic model) is proposed to estimate the number of social ties between two locations using population data in a similar manner to how transportation research deals with trips. To overcome the asymmetry of transportation models, the new model considers that the number of relationships between two locations is inversely proportional to the population in the ellipse whose foci are in these two locations. The elliptic model is evaluated by considering the anonymous communications patterns of 25 million users from three different countries, where a location has been assigned to each user based on their most used phone tower or billing zip code. With this information, spatial social networks are built at three levels of resolution: tower, city and region for each of the three countries. The elliptic model achieves a similar performance when predicting communication fluxes as transportation models do when predicting trips. This shows that human relationships are influenced at least as much by geography as is human mobility.

Contents

1. Introduction

Although social networks have been known for years to play a key role in various human phenomena [1, 2], for decades their study has been limited to certain kinds of social relationships for which interaction records were available, such as authorship and cooperation in science [3, 4]. Only recently has it been possible to map large social networks representing a broader range of interactions in order to explore how their structures influence processes occurring in these networks. The required large social network data sets, usually coming from telecommunication records originating in e-mail [5], phone [6] or online communication platforms [7], have been used to explore a wide range of topics, such as adoption of innovation [8], social groups discovery [9]–[11], epidemic spreading [12]–[14], social mobilization [15] or sentiment spreading [16].

Despite the publication of such studies, network data is not widely available to the community due to legal, 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 [6]. For these reasons, creating models that are able to mimic different social network properties have recently attracted a fair amount of research interest [17]–[21]. While most models try to generate synthetic networks with some desired characteristics (such as degree distribution and clustering coefficients), we will focus here on reproducing a macroscopic feature of real social networks: the number of social ties between different locations, i.e. how many relationships exist between two cities, two regions or even two neighborhoods. Throughout the paper, we will employ the term *location* to generically denote any of these three spatial aggregation levels. The creation of social connectivity maps between locations from widely accessible data, such as the population geographic distribution (which is universally accessible for almost any region of the world through tools such as Landscan), will prove useful for the study of information [22] or behavior spreading in social as well as other networks[23].

1.1. The effect of geography on social networks

Towards the end of the 19th century and the beginning of the 20th century, 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 a single goal: allowing people to communicate with those who are far away (indeed the Greek prefix *tele-* means *distant*). Interestingly (and contrary to some predictions from the beginning of the Internet era [4]) recent analyses of records from such systems show that people do not commonly use them to talk to those far away, but rather with people who are actually close by. In fact, it has been consistently found across records from emails, phones and blogs that the probability of a communication occurring between two people who are r kilometers apart from each other follows a decay function, typically a power law [24]–[27].

Although the communication fluxes between regions have not been the focus of much research yet, the above-mentioned new evidence shows that communication fluxes behave in a similar way to trip fluxes and other phenomena driven by the distribution of population across the geographical space. In transportation research, flux prediction is a well-defined problem: given a set of locations $\{i, j, \ldots\}$ whose coordinates and populations $\{n_i, n_j, \ldots\}$ 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 [28]–[30], which try to gather the effect of decaying probability with distance $r_{i,j}$ following the equation

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

where α and β are fitting parameters usually estimated from training data, and $f(r_{ij})$ increases with distance, typically following an exponential or power-law function. A powerful idea was introduced recently by the radiation model [31], which claims that it is not the distance that matters, but the number of opportunities between i and j, which can be estimated by the population in the area. In short, the authors explain that someone from rural Iowa is more likely to travel further to satisfy their needs than someone in New York City, given the latter has a handful of options within a few blocks. While in its original publication the radiation model included testing against a phone call dataset (see figure 3 in [31]), the problem of predicting communication fluxes has not been the main focus of any model to date. In this paper we will present a new model inspired by this radiation model which is able to predict communication fluxes surprisingly well. Actually, the accuracy is similar to that reached by current transportation models when predicting trips.

2. Model description

Formally, the radiation model, when applied to social relationships, estimates the communication flux T_{ij} between two locations i and j using the population in both

locations, and the population within the circle whose center is i and whose radius is equal to the distance between i and j. Its formulation is

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

where n_i represents the population of location i, $s_{i \to j}$ the number of people who are not in i but closer to i than j and the normalization $K_i = n_i(N_{\mathrm{T}}/N)$, where N_{T} is the total number of relationships to predict (which in general is considered to be available) and $N = \sum_i n_i$ the total population.

It is straightforward to verify that T^{rad} matrices are not symmetric in general. While asymmetry is a desirable feature for mobility models (commuting origin–destination matrices have strongly asymmetric suburbs–downtown fluxes) it is not when dealing with communication fluxes, because the number of relationships people from location i have with people from location j must be the same as the number of relationships people from j have with people from i.

A natural modification of the radiation model to deal with communication fluxes could be a simple symmetrization of the model, which we will denote radBI and whose formulation is

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

As shown below, this model has a limited performance. This fact motivated us to develop the new model presented in this paper. Our model, which we will refer to as the elliptic model (EM), is oriented to deal with social relationships. The EM considers that the probability of someone living at location i having an acquaintance at location j is inversely proportional to the population of the area where both could meet, provided their combined travel distance does not exceed a certain threshold. This area forms an ellipse whose foci are in locations i and j. Among all possible ellipses the model selects the smallest one containing the two r_{ij} radius circles whose centers are in i and j respectively (see figure 1 for a graphical explanation and comparison to the radiation model). Thus, the EM formulation is

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

where e_{ij} is the population within the ellipse depicted in figure 1 (note that e_{ij} includes n_i and n_j) and K is a normalization parameter obtained from the total number of relationships to predict N_{T} $(\sum_i \sum_j T_{ij}^{\mathrm{ellip}} = N_{\mathrm{T}})$. Since $e_{ij} = e_{ji}$, $T_{ij}^{\mathrm{ellip}} = T_{ji}^{\mathrm{ellip}}$ and thus our model produces symmetrical matrices T.

In order to compare the quantities involved in the model, one needs to consider the sets $S_{i \to j}$ and $S_{j \to i}$, such as $\#S_{i \to j} = s_{i \to j}$ and $\#S_{j \to i} = s_{j \to i}$. Let us address the case of a very large city $C \subset S_{i \to j}$ whose population $n_C \approx e_{ij}$. While the radiation model predicts different fluxes depending on whether $C \subset (S_{i \to j} \cap S_{j \to i})$ or not (smaller when C belongs to the intersection) the EM will provide the same prediction for both cases. In fact, since $e_{ij} \geq \#(S_{i \to j} \cup S_{j \to i}) + n_i + n_j$ (and usually $e_{ij} \approx \#(S_{i \to j} \cup S_{j \to i}) + n_i + n_j$) the role of the union set is the main contribution of the model.
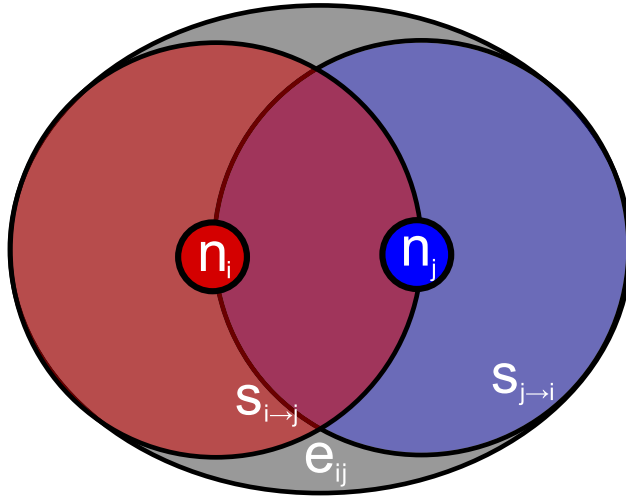
Figure 1. Model scenario: n_i represents the population of location i while $s_{i \to j}$ represents the population within the circle with its center in i with a radius up to j. As long as the population is not homogeneously distributed $s_{i \to j} \neq s_{j \to i}$, the radiation model predictions will not be symmetrical. e_{ij} represents the population within the smallest ellipse whose foci are in i and j and contains both previous circles, as well as n_i and n_j.

Table 1. Characteristic properties of the social networks in the countries studied: 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.

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

3. Data description

To evaluate the performance of the EM, we compare it with a mobile phone data set containing call detail records (CDRs) of a six-month period in three different countries: France, Portugal and Spain. In total, over seven billion calls are considered to build the social graph for each country, whose links are included only if there is at least one call in both directions during the observation period. The result is an undirected graph (this is a common technique in the literature to avoid both marketing callers and misdialed calls [6]). In table 1, some characteristics of the networks are presented, such as high clustering and relatively low average degree, which are expected from previous literature concerning mobile phone networks.

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

Country	Towers/zip codes	Cities	Regions
France	17 475	3520	96
Portugal	2 209	297	20
Spain	8 928	5446	52

In addition to the communication records, our data include a location for each user: the most used mobile phone tower in France and Portugal and the billing zip code in Spain. In order to benchmark the multi-scale performance of the EM, three aggregation levels have been used: country-wide fluxes, fluxes between cities and regions, and metropolitan fluxes within cities. Table 2 presents the number of locations considered in each aggregation level. When applying these spatial aggregations, the center of mass of the population is used as the higher level location, instead of the centroid of the region polygon (defined by administrative boundaries), in order to avoid undesirable effects in the fairly common case of a big city located in a corner of a polygon.

4. Communication fluxes on a country scale

To validate the predictions of the EM at large scale, we consider connectivity matrices T at two aggregation levels. At the region level, matrix T has thousands of elements while at the city level there are tens of millions of fluxes to predict (see table 2). Input data for the predictions consists only of the location's coordinates and populations, and the total number of relationships to predict N_T. Like the radiation model, the EM retains the advantage of being free of parameters, 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 fluxes between cities. The results prove consistently that the EM outperforms both the radiation model and its bilateral version. To present further evidence of the performance of the EM, we include in table 3 the R^2 of the predictions at both aggregation levels. The results confirm that the EM outperforms previous models.

Overall, the accuracy of the predictions is similar to that obtained when applying transportation models to trip prediction [31, 32]. This is an unexpected finding, since in principle, while there is an increasing cost (like time or energy) associated with distance when traveling, there is no such cost when making a phone call. While, as stated in the introduction, there were previous reports illustrating that social ties depend on physical distance, the capability of reproducing a significant portion of the distribution of social ties between locations just by employing a map placing them and their populations, highlights even more the importance of the geographical space when forming ties.
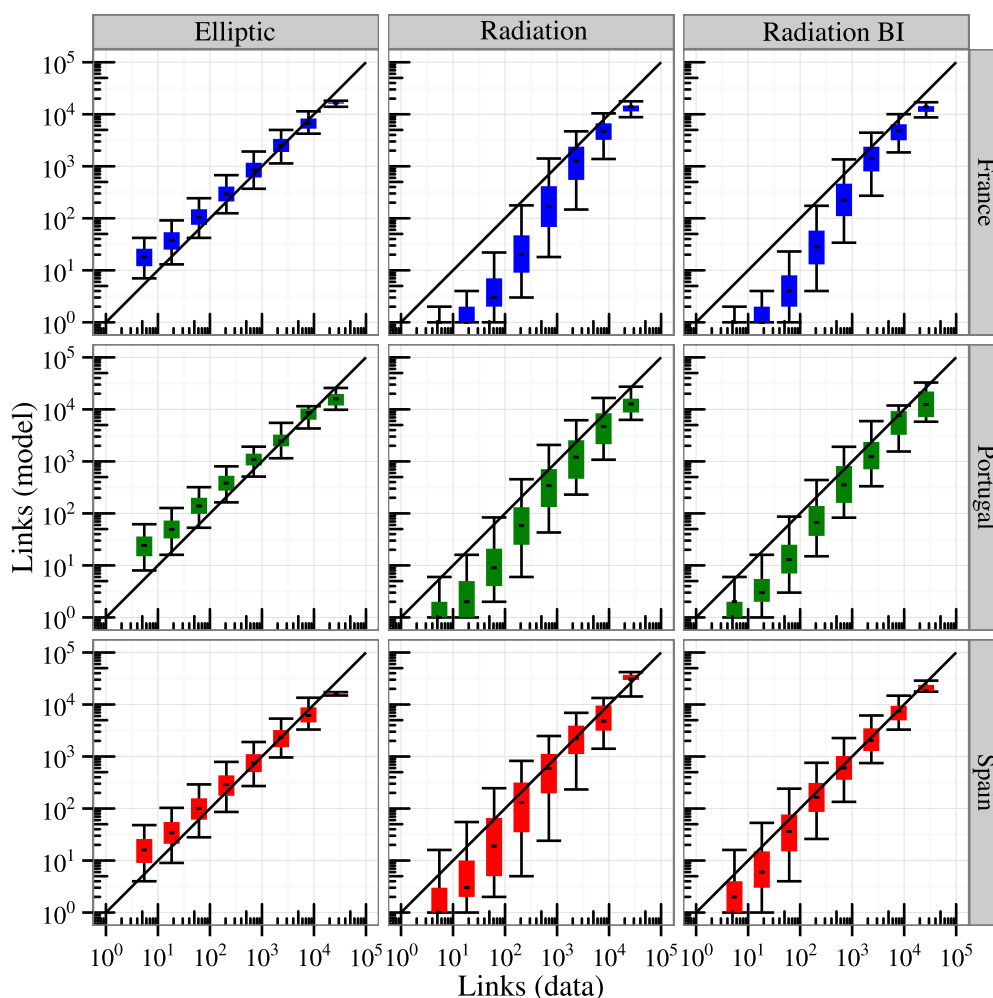
Figure 2. Predictions by different models versus real data. Fluxes between every city are presented, considering 297 cities in Portugal, 5446 in Spain, and 3520 in France. Error bars plot 10%, 30%, 50%, 70% and 90% quantiles. The elliptic model outperforms both the radiation and bilateral radiation models in all three scenarios.

5. Communication fluxes within cities

While previous literature has already stated that distance influences the creation of social ties between cities, our dataset allows us to study also urban relationships by using the finer spatial aggregation level available: phone towers or zip codes. Predicting all possible tower to tower relationships within the country would imply dealing with a T matrix with up to 300 million elements, with only less than 1% of them being not null. Thus, the prediction accuracy would be severely biased by the huge amount of zero cells. Instead, we study the short-range accuracy of the model by applying it in each city where we have at least 20 different locations (the upper limit being Paris, where we have over 1000 mobile phone towers). In total, the analysis includes 40 cities in France, 29 Spain and 20 in Portugal.

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

	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 4. Average R^2 for urban fluxes prediction for every city in the data set where there are at least 20 different locations (towers or zip codes). The number of locations range from this minimum of 20 up to 1000 in Paris. This amounts to 40 cities in France, 29 Spain and 20 in Portugal. Although the EM again outperforms previous models, each performance is low when compared to country-wide scenarios.

	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 presents the results for the three algorithms in terms of average R^2. These results confirm that the EM still performs better, while the overall prediction accuracy is smaller compared to the country-wide experiment. The loss of accuracy within urban areas for any model purely based on distance is expected and observed in the transportation field [32]. One of the main reasons for this loss of accuracy is the fact that the distance is a poorer proxy for travel time or cost in cities. People in cities tend to be within a daily radius of action and the decision of others they communicate with depends on other metrics related to the different hierarchies that could define a social distance (e.g. ethnicity, occupation, etc) [25].

5.1. Correction ε as a model improvement for urban areas

When applying the EM based on figure 1 to urban relationships one should be aware that 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 when predicting T_{ij}. Since 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 correction parameter ε so that the ellipse is now the smallest one containing the two circles of radius $r_{ij} + \varepsilon$ centered in i and j. After studying several values of ε, we
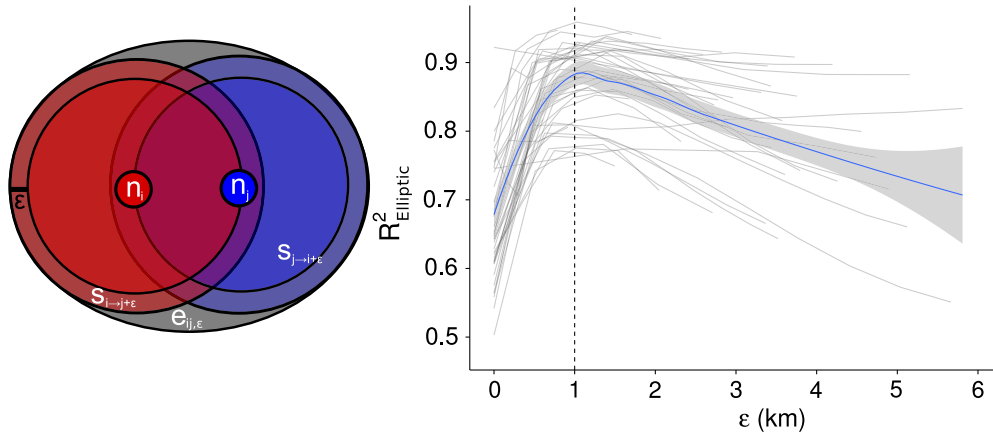
Figure 3. Model modified for intracity predictions, adding the correction term ε. Each gray line represents a certain city in the dataset with the blue line and the shadow representing the general trend. We find predictions improve when some correction term is included, reaching a maximum around $\varepsilon = 1$ km.

found that the prediction accuracy peaks near $\varepsilon = 1$ km for nearly all the cities (as shown in figure 3).

There may be several interpretations for such a maximum: one could argue that it comes from the location error, known to be close to the average distance to neighbors from the Voronoi tessellation [33], which is around 1 km on average for our dataset. This agrees with the fact that the optimal ε is a fixed value and does not depend on the distance r between i and j. On the other hand, when applied back to country-wide scenarios we found the correction term does not improve the predictions and no peak emerges near $\varepsilon \simeq r_{\text{Voronoi}}$ or elsewhere, reinforcing the hypothesis that within cities we are reaching the boundaries of the resolution of our location data.

Another way to evaluate the performance of the different models is to compare them against empirical data in terms of the link-distance distribution $P(r)$, which represents the probability of observing a relationship between two people living r kilometers from each other. Figure 4 shows the improvement in $P(r)$ when applying the $\varepsilon = 1$ km correction. Without the correction term, short-range relationships are over represented, while the EM with the correction fits almost perfectly with the distribution obtained from the data. Note that although radiation model predictions also improve, it still predicts shorter relationships 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 applying the corrected model with $\varepsilon = 1$ km across all cities in the data set.

6. Conclusions and further research

In this paper the problem of predicting communication fluxes between different locations has been successfully addressed. A new model has been proposed to calculate the communication fluxes using only population distribution data. This data is publicly
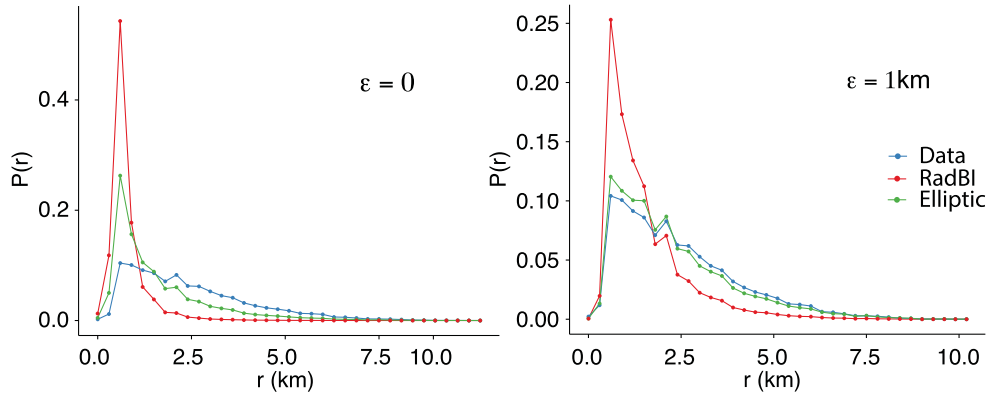
Figure 4. Left: fraction of relationships $P(r)$ within distance r in the real dataset compared to predictions by both elliptic and bilateral radiation models where $\varepsilon = 0$ for Porto (Portugal). Right: the elliptic prediction gets very close to the data when using $\varepsilon = 1$ km. Although the radiation model predictions also improve, they still predict shorter relationships than those observed in reality.

Table 5. Average R^2 of the predictions for the corrected model with $\varepsilon = 1$ km, compared to the original (non-corrected) model.

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

available worldwide through projects which provide population estimates for almost every square mile on earth.

The presented model successfully takes into account the symmetry of the communication fluxes, in order to predict the number of social ties between geographical locations at different scales, ranging from neighborhoods to regions. Interestingly, we have shown that geolocated population data is as useful to predict communication fluxes as it is to predict trip fluxes.

The proposed model is readily available to be used by researchers in different social sciences studying various phenomena where human ties are known to be crucial, such as information propagation or disease spreading. Overall, our model implies social ties are to a large extent driven by geographical factors. While there may be other factors influencing very long distance relationships (e.g. time zones, or natural, national [34] and language borders [35]) the available data did not allow us to check them, so that further research would be needed along this line.

In order to enhance the employment of the EM by the research community, implementations in three widely used programming languages have been made available on our homepage [36], together with an interactive tool using *départments* in France as an example scenario.

References

[1] Granovetter M S, *The strength of weak ties*, 1973 *Am. J. Sociol.* **78** 1360
[2] Milgram S, *The small world problem*, 1967 *Psychol. Today* **2** 60
[3] Jones B F, Wuchty S and Uzzi B, *Multi-university research teams: shifting impact, geography, and stratification in science*, 2008 *Science* **322** 1259
[4] Cairncross F, 2001 *The Death of Distance: How the Communications Revolution Will Change Our Lives* (Boston, MA: Harvard Business Press)
[5] Kossinets G and Watts D J, *Empirical analysis of an evolving social network*, 2006 *Science* **311** 88
[6] Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J and Barabási A-L, *Structure and tie strengths in mobile communication networks*, 2007 *Proc. Nat. Acad. Sci.* **104** 7332
[7] Mislove A, Marcon M, Gummadi K P, Druschel P and Bhattacharjee B, *Measurement and analysis of online social networks*, 2007 *Proc. 7th ACM SIGCOMM Conf. on Internet Measurement* (New York: ACM) p 29
[8] Toole J L, Cha M and González M C, *Modeling the adoption of innovations in the presence of geographic and media influences*, 2012 *PloS One* **7** e29528
[9] Yang J and Leskovec J, *Community-affiliation graph model for overlapping network community detection*, 2012 *ICDM: 2012 IEEE 12th Int. Conf. on Data Mining* (Piscataway, NJ: IEEE) p 1170
[10] Ahn Y-Y, Bagrow J P and Lehmann S, *Link communities reveal multiscale complexity in networks*, 2010 *Nature* **466** 761
[11] Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E, *Fast unfolding of communities in large networks*, 2008 *J. Stat. Mech.* P10008
[12] Pastor-Satorras R and Vespignani A, *Epidemic spreading in scale-free networks*, 2001 *Phys. Rev. Lett.* **86** 3200
[13] Wang P, González M C, Hidalgo C A and Barabási A-L, *Understanding the spreading patterns of mobile phone viruses*, 2009 *Science* **324** 1071
[14] Schneider C M, Mihaljev T, Havlin S and Herrmann H J, *Suppressing epidemics with a limited amount of immunization units*, 2011 *Phys. Rev.* E **84** 061911
[15] Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A and Pentland A, *Time-critical social mobilization*, 2011 *Science* **334** 509
[16] Fowler J H and Christakis N A, *Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study*, 2008 *BMJ: British Med. J.* **337** 23
[17] Watts D J and Strogatz S H, *Collective dynamics of 'small-world' networks*, 1998 *Nature* **393** 440
[18] Barabási A-L and Albert R, *Emergence of scaling in random networks*, 1999 *Science* **286** 509
[19] Karrer B and Newman M E J, *Random graphs containing arbitrary distributions of subgraphs*, 2010 *Phys. Rev.* E **82** 066118
[20] Holme P and Kim B J, *Growing scale-free networks with tunable clustering*, 2002 *Phys. Rev.* E **65** 026107
[21] Herrera C and Zufiria P J, *Generating scale-free networks with adjustable clustering coefficient via random walks*, 2011 *NSW: 2011 IEEE Network Science Workshop* (Piscataway, NJ: IEEE) pp 167–72
[22] Allen T J, 1984 *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization* vol 1 (Cambridge, MA: MIT Press Books)
[23] Fernández-Gracia J, Suchecki K, Ramasco J J, Miguel M S and Eguíluz V M, Is the voter model a model for voters? 2013 arXiv:1309.1131
[24] Krings G, Calabrese F, Ratti C and Blondel V D, *Urban gravity: a model for inter-city telecommunication flows*, 2009 *J. Stat. Mech.* L07003
[25] Herrera-Yagüe C, Schneider C M, Couronne T, Somoreda Z, Benito R M, Zufiria P J and González M C, *On the origins of searchability in social networks* 2013, under review
[26] Liben-Nowell D, Novak J, Kumar R, Raghavan P and Tomkins A, *Geographic routing in social networks*, 2005 *Proc. Nat. Acad. Sci. USA* **102** 11623
[27] Phithakkitnukoon S, Smoreda Z and Olivier P, *Socio-geography of human mobility: a study using longitudinal mobile phone data*, 2012 *PloS One* **7** e39253
[28] Erlander S and Stewart N F, 1990 *The Gravity Model in Transportation Analysis: Theory and Extensions* vol 3 (Utrecht: VSP)
[29] Eggo R M, Cauchemez S and Ferguson N M, *Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States*, 2011 *J. R. Soc. Interface* **8** 233
[30] Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco J J and Vespignani A, *Multiscale mobility networks and the spatial spreading of infectious diseases*, 2009 *Proc. Nat. Acad. Sci.* **106** 21484

[31] Simini F, González M C, Maritan A and Barabási A-L, *A universal model for mobility and migration patterns*, 2012 *Nature* **484** 96

[32] Yang Y, Herrera-Yagüe C, Eagle N and González M C, A multi-scale study of commuting patterns incorporating digital traces, 2013, under review

[33] Ulm M and Widhalm P, Properties of the positioning error of cell phone trajectories, 2013, under review

[34] Ugander J, Karrer B, Backstrom L and Marlow C, The anatomy of the Facebook social graph, 2011 arXiv:1111.4503

[35] Blondel V, Krings G and Thomas I, *Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone*, 2010 *Brussels Studies* **42**

[36] http://humnetlab.mit.edu/elliptic 2013