Entangling Mobility and Interactions in Social Media

Przemyslaw A. Grabowicz, José Ramasco, Bruno Goncalves, Victor Eguiluz

Entangling Mobility and Interactions in Social Media

Przemyslaw A. Grabowicz[1,2]*, **José J. Ramasco**[1], **Bruno Gonçalves**[3,4], **Víctor M. Eguíluz**[1]

[1] Instituto de Fisica Interdisciplinar y Sistemas Complejos, CSIC–UIB, Palma de Mallorca, Spain, [2] Max Planck Institute for Software Systems, MPG, Saarbrücken, Germany, [3] Le Centre de Physique Théorique, CNRS, Aix-Marseille Université, Marseille, France, [4] Le Centre de Physique Théorique, CNRS, Université de Toulon, La Garde, France

Abstract

Daily interactions naturally define social circles. Individuals tend to be friends with the people they spend time with and they choose to spend time with their friends, inextricably entangling physical location and social relationships. As a result, it is possible to predict not only someone's location from their friends' locations but also friendship from spatial and temporal co-occurrence. While several models have been developed to separately describe mobility and the evolution of social networks, there is a lack of studies coupling social interactions and mobility. In this work, we introduce a model that bridges this gap by explicitly considering the feedback of mobility on the formation of social ties. Data coming from three online social networks (Twitter, Gowalla and Brightkite) is used for validation. Our model reproduces various topological and physical properties of the networks not captured by models uncoupling mobility and social interactions such as: i) the total size of the connected components, ii) the distance distribution between connected users, iii) the dependence of the reciprocity on the distance, iv) the variation of the social overlap and the clustering with the distance. Besides numerical simulations, a mean-field approach is also used to study analytically the main statistical features of the networks generated by a simplified version of our model. The robustness of the results to changes in the model parameters is explored, finding that a balance between friend visits and long-range random connections is essential to reproduce the geographical features of the empirical networks.

Citation: Grabowicz PA, Ramasco JJ, Gonçalves B, Eguíluz VM (2014) Entangling Mobility and Interactions in Social Media. PLoS ONE 9(3): e92196. doi:10.1371/journal.pone.0092196

Editor: Matjaž Perc, University of Maribor, Slovenia

Received November 29, 2013; **Accepted** February 19, 2014; **Published** March 20, 2014

Funding: We have received partial financial support from the Spanish Ministry of Economy (MINECO) and FEDER (EU) under project MODASS (FIS2011-24785), and from the EU Commission through projects EUNOIA, INSIGHT and LASAGNE. PAG acknowledges funding from the JAE-Predoc program of Consejo Superior de Investigaciones Científicas and JJR from the Ramón y Cajal program of MINECO. BG was partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pms@mpi-sws.org

Introduction

The advent of the big data revolution has opened the door to the analysis of massive datasets on all aspects of society. New technologies have made possible the access to unprecedented amount of information on human behavior generated unobtrusively whenever people interact with or through modern technologies such as cell phones, online services, mobile applications, etc. This fact is facilitating the pursuit of a computational approach to the study of problems traditionally associated with social sciences [1]. Over the course of the last few years, it has allowed for the development of greater insights, for instance, into human mobility [2–4], structure of online social networks [5,6], cognitive limitations [7,8], information diffusion and social contagion [9–13], the importance of social groups [13–15] or even how political movements raise and develop [16–18].

The relation between physical location and social interactions can be also explored with the new available data. In general, people tend to interact and maintain relations with geographically close peers. A tendency that gets reflected in a decay of the social interaction probability with the physical distance. This effect has been observed, for example, in phone call records [19–21] and in online friendships [22]. Furthermore, it has been shown that online [23] social links can be inferred from user co-occurrences in space and time and, likewise, that the location of a person can be predicted from the geographic positions of his or her online friends [25]. Some further aspects of the relation between geography and online social contacts have been studied such as the probability that a link at a given distance closes a triangle [19,22,26], the connections between users in different countries [27], the social interactions and mobility in emergency situations [28] or the overlap between users' ego networks and how it decays with the distance [29]. Multi-parametric inference methods have been applied to empirical data with the aim of predicting link presence and users' locations [30–32]. These works show that the accuracy of link prediction is considerably improved by taking into account the geographical information, and that the accuracy of location prediction is enhanced when the online social links are provided.

The availability of geo-localized information has also allowed for a detailed exploration of human mobility [2,3,21,33–36]. The length of displacements between locations was found to follow a broad distribution, well fitted by a power-law decaying function [2,3]. The asymmetry of the travels was studied by considering ellipsoidal boundaries to the average individual displacements and analyzing the scaling of the radius of gyration. Memory effects in the individual displacements was also analyzed, finding that individuals' home and workplace have a considerable impact on their mobility patterns [4]. These results motivated the introduction of several mobility models with the aim of explaining the features observed in the data [4,36–39]. Despite the supporting evidence [21], most of these models lack a connection between mobility and social interactions [40].

In this work, we lay a bridge between these two worlds by introducing a model coupling social tie formation and spatial mobility. Preceding models considering network structure and geography are uncoupled [24,41]. Our model simulates the movement of individuals and creates links between them when they are physically close mimicking the effect of face-to-face interactions. We study the model both numerically and analytically and confront its results with empirical data obtained from three online social networks. We show that the model generates more realistic networks than uncoupled models.

Materials and Methods

The Datasets

We have collected data from online social networks containing both social links and information about the users' physical positions. The first dataset was obtained from Twitter by means of its API [42]. We identify over 714,000 single users, who tweeted using a GPS enabled mobile device during the month of August 2011 [43]. If those users reported various locations in different tweets, the most recent one is taken for the purpose of the study. The other two datasets contain information referring to the users' location check-ins and the social networks of Gowalla and Brightkite [31]. Both were location-based online social networks, in which users can check-in at their current locations and receive information about services in the area as well as about their friends' positions. Gowalla and Brightkite are no longer active but their data is available online [44]. The main statistical features of our three datasets are displayed in Table 1.

Social interactions across country borders have particular properties and are affected by political, linguistic or cultural factors. We overcome this difficulty by restricting our analysis to the networks within each country. Intra-country mobility and social contacts account for the large majority of a user activity [45,46]. For simplicity, we focus on the three major countries with more than one thousand users in each of our datasets: the United States (US), the United Kingdom (UK) and Germany (DE). We have analyzed and modeled other countries and found similar results to the ones presented in this manuscript.

The Travel and Friend (TF) Model

The model structure is illustrated in Figure 1. The initial condition is a set of individuals located in the last known positions of the online network users as extracted from the data. At each step of the model, a randomly chosen agent performs actions in two stages:

1. Travel

(a) Visit a randomly selected friend at his current location with probability p_v.

(b) Otherwise, travel to a new location. The distance of travel is obtained from a distribution of jump lengths, while the direction is chosen proportionally to the population density at the target distance.

2. Friendship

(a) With probability p, create directed links to agents within a neighborhood of size $\delta \times \delta$.

(b) With probability p_c, create a directed connection to a randomly chosen agent anywhere in the system.

The acronym of the TF model comes from the initials of these two stages. The model is iterated until the number of created connections is equal to the number of links measured in the empirical networks. Despite its simplicity, the model incorporates several major features of human behavior. The *Travel* stage accounts for both recurring visits to the same location and exploration of new places and the *Friendship* component generates both face-to-face contacts and online acquaintances independent of the geography. Note that in the Friendship phase both of two possible actions happen concurrently with the respective probabilities. The effect of each of the underlying assumptions is systematically explored through analysis of model variants in Appendix S1.

The model has four input parameters: p_v, p, p_c and δ, besides the distribution of jump lengths. Following the empirical findings of Ref. [4], we take a power-law distribution for the jump lengths with an exponent -1.55 for the main simulations shown in this work. Still, other functional shapes for the jump distribution are also discussed in the following section. The direction of the jump is chosen proportionally to the population density at the target distance using the gridded population estimates of the world for 2005 with the cell size $2.5'$ [47]. The values of the probability $p=0.1$ and the box size $\delta=0.001^\circ$ are fixed to match the relation between the probability of friendship and the number of daily spatiotemporal coincidences measured in [23]. (To this end, we assume that one time step of the model corresponds roughly to one day. Most of our simulations finish in less than a 1,000 time steps, corresponding to a few years, which is of the order of magnitude of users' lifetime, given that Twitter was founded in 2005 and our dataset is from 2011.) Furthermore, we tested different values of δ and p and did not observe strong deviations in the model results. This leaves us only with p_v and p_c as free model parameters, we will systematically explore in the coming sections the impact of these parameters on the model results, since, as it will be shown, they are

Table 1. Datasets.

	TOTAL($\times 10^3$)		US($\times 10^3$)		UK($\times 10^3$)		DE($\times 10^3$)	
	N	L	N	L	N	L	N	L
Twitter	714	15000	132	1100	28	117	3.8	8.5
Gowalla	196	950	46	350	5.2	20	5.2	30
Brightkite	58	214	27	167	3.1	10	1.3	7.2

Number of users (nodes) *N* and of links *L* of the networks obtained from the different geo-localized datasets for the United States (US), the United Kingdom (UK) and Germany (DE).

doi:10.1371/journal.pone.0092196.t001

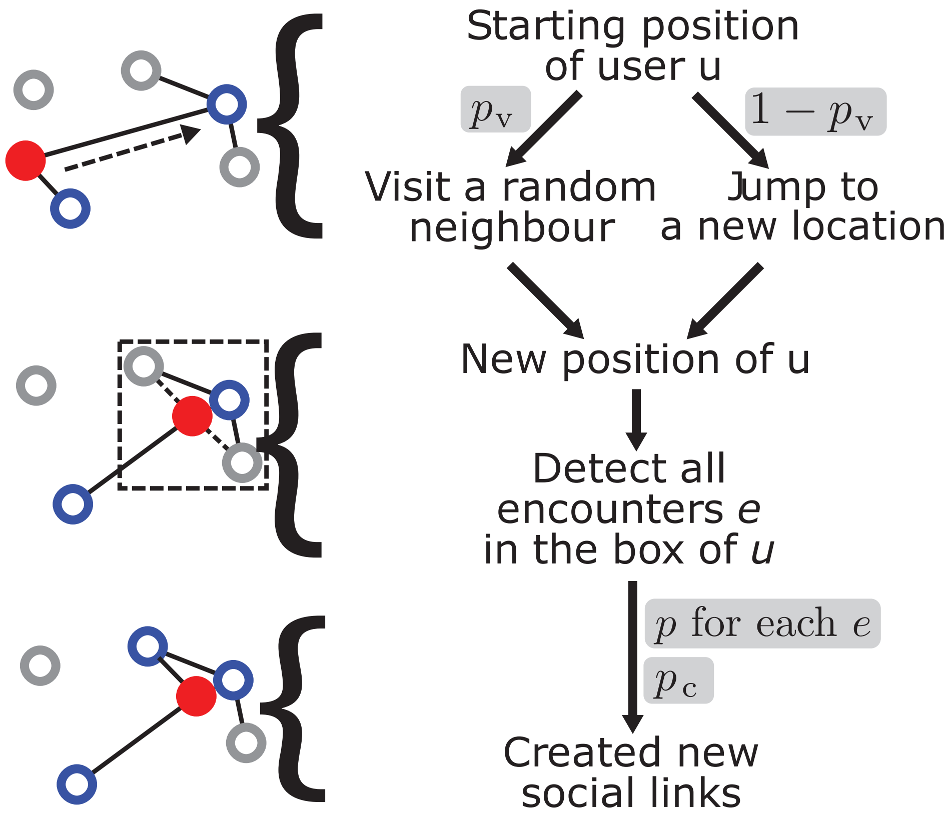

Figure 1. Schematic of the TF model. The central node is the filled red circle and its neighbors are marked in blue. Directionality of links is neglected in this schematic to maintain simplicity.
doi:10.1371/journal.pone.0092196.g001

essential for generating network comparable with the empirical ones.

Geo-social Properties of the Networks

We start by establishing a set of metrics in order to characterize networks structure and its relation to geography. First, we measure the probability of two users to have a link at a certain distance $P_l(d)$. It is defined as the ratio between the number of existing links at distance d and the total number of users pairs separated by d, and thus it is constrained to the interval $[0,1]$. $P_l(d)$ decays slowly with the distance for empirical networks, essentially as a power-law with exponent -0.7, which is followed by a plateau for very large distances (see Figure 2A). This functional shape remains identical for all the countries and all the datasets considered. It matches, besides, the behavior reported in the literature for online social systems [22,26].

A second metric that we consider is the degree distribution of the social networks (see Figure 2B for the empirical networks). For Twitter, which has a directed social network, we consider the degrees of its symmetrized version. The distribution $P(k)$ displays heavy tail in all the datasets, even though there are slight differences between them.

Connections in Twitter are directed: one user follows the messages emitted by another. Reciprocated connections indicate mutual interest between the two users and a closer type of social relation [8,14]. To assess how geography and reciprocity correlate, we measure the probability $R(d)$ of reciprocation conditional on a link at a distance d (Figure 2C). We find that the reciprocity decreases with the distance in all the countries analyzed. This trend is consistent with the idea that stronger relations occur close to where users spend most of their time, with some longer connections composed of friends who moved, former residences, online acquaintances, etc. Furthermore, long not-reciprocated connections may include users following public figures or celebrities.

With the aim of quantifying social closeness between users, we define the social overlap J_{f} of two connected users i and j as

$$J_{\mathrm{f}} = \frac{|\mathcal{K}_i \cap \mathcal{K}_j|}{|\mathcal{K}_i \cup \mathcal{K}_j| - 2} \qquad (1)$$

where $\mathcal{K}_i$ represents the set of friends of user i. J_{f} is inspired by the Jaccard index but is modified to ensure that it takes a value of 1 if i and j share all their friends, and 0 if they have no common friends. In Figure 2D, the average of the social overlap $J_{\mathrm{f}}(d)$ over all pairs of connected users is plotted as a function of the distance between them. The social overlap decreases with the distance. The functional shape of the curves is similar for all the datasets, even though the overlap level is different for each of them. For Twitter, we use the symmetrized version of the network to study social overlap and clustering.

Another well known phenomenon in social networks is triadic closure. As one individual has a close relation with other two persons, there are high chances that these two individuals end up creating a social relation between themselves. In network analysis, a magnitude that quantifies this effect is the average clustering coefficient C. It is defined as the ratio between the number of closed triads and the total number of triads in the network. A triad is a sequence of 3 nodes i,j,k such that the central node j is connected to both extreme nodes i and k. A closed triad is a triad

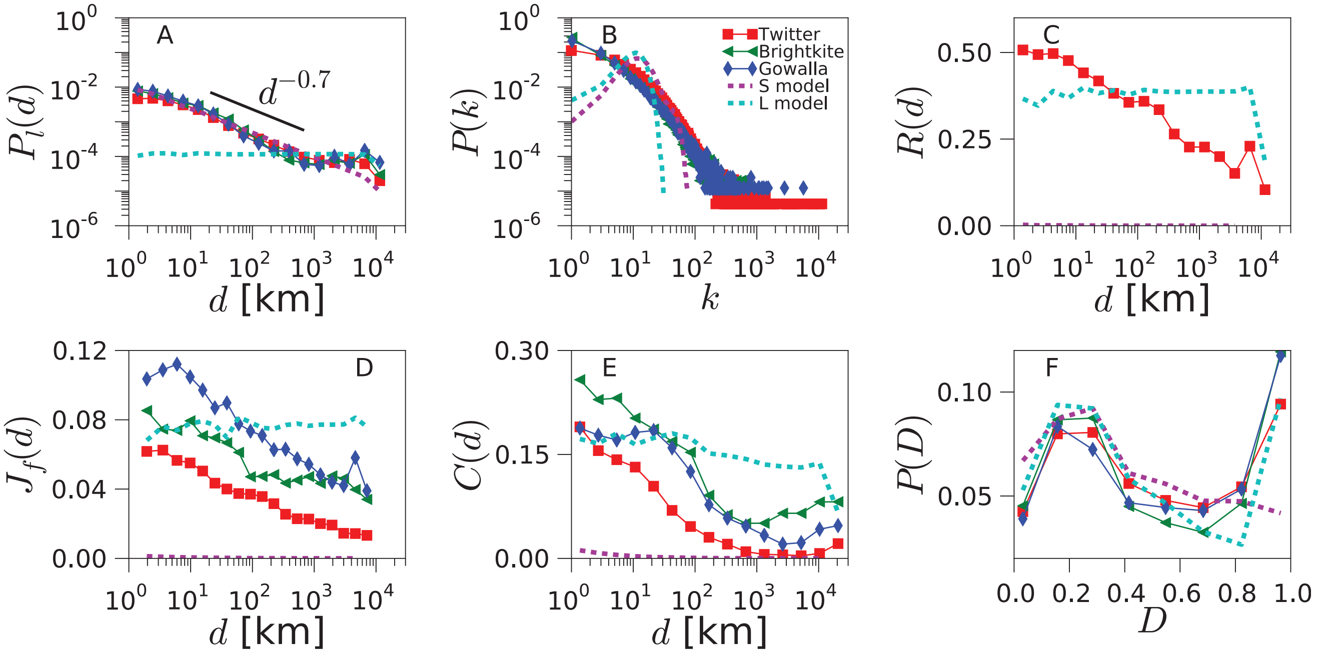

Figure 2. Network geo-social properties. Various statistical network properties are plotted for the data obtained from Twitter (red squares), Gowalla (blue diamonds), Brightkite (green triangles) and the null models (dashed lines), for the US (for the UK and Germany, see Figures S1 and S2). The spatial model (magenta), based on geography, matches well the data in $P_l(d)$, but yields near-zero values for $R(d)$, $J_f(d)$ and $C(d)$. The linking model (cyan), based on triadic closure, produces enough clustering, but it does not reproduce the distance dependencies of $P_l(d)$, $R(d)$, $J_f(d)$ and $C(d)$.
doi:10.1371/journal.pone.0092196.g002

that has also an edge between i and k, forming a triangle. Note that a triangle consists of 3 triads centered on different nodes. The effect of the distance on the clustering coefficient can be incorporated by measuring the distances from each central node j to two neighbors i and k forming a triad, $d = d_{ij} + d_{jk}$, and calculating the network clustering restricted to triads with distance d. This new function $C(d)$ is the probability of closing a triangle given the distance d in a triad

$$C(d) = \frac{\Delta(d)}{\Lambda(d)}, \tag{2}$$

where $\Lambda(d)$ and $\Delta(d)$ are the numbers of triads and closed triads for the distance d, respectively. The value of the global clustering coefficient C can be recovered by averaging $C(d)$ over d. In the datasets, we observe a drop in $C(d)$ followed by a plateau, which is best visible for the US networks (Figure 2E).

Given a triangle, several configurations are possible if there is diversity in the edge lengths. The triangle can be equilateral if all the edges have the same length, isosceles if two have the same length and the other is smaller, etc. We estimate the dominant shapes of the triangles in the network by measuring the disparity D defined as:

$$D = 6\left[\frac{d_1^2 + d_2^2 + d_3^2}{(d_1 + d_2 + d_3)^2} - \frac{1}{3}\right], \tag{3}$$

where d_1, d_2 and d_3 are the geographical distances between the locations of the users forming the triangle. The disparity takes values between 0 and 1 as the shape of the triangle passes from equilateral to isosceles, where one edge is much smaller than the other two. D shows a distribution with two maxima in the online social networks (Figure 2F), for low and high values. The two most common geometries of the triangles are: i) all 3 users are at a similar distance, ii) 2 users are close to each other, while the third one is distant. Since most edges correspond to small distances, this means that most triangles are constituted by three users that are all close to each other geographically. However, the stretched isosceles configuration is also relatively common.

Summarizing, we have defined the following metrics in order to characterize the networks structure and its relation to geographical distance:

- $P_l(d)$: Probability of linking at a distance d (Figure 2A).
- $P(k)$: Degree distribution (Figure 2B).
- $R(d)$: The probability of reciprocation conditional on a link at a distance (Figure 2C).
- $J_f(d)$: Average overlap as a function of the distance (Figure 2D).
- $C(d)$: Clustering coefficient as a function of the triad distance (Figure 2E).
- $P(D)$: Distribution of distance disparity for the triangles' edges (Figure 2F).

We will use these metrics in the coming sections to estimate the ability of model to produce social networks comparable with those obtained from the empirical datasets.

Model Calibration

Next, we will find a compromise between the different metrics and search for the parameter values for which a given model best fits simultaneously the various statistical properties. To do so, we define an overall error Err to quantify the difference between the networks generated with the model and the empirical ones. The parameters of the model are then explored to find the values that minimize Err. We measure the error $\mathrm{Err}[X]$ for each property X and take the average over all the properties

$$\mathrm{Err} = \frac{1}{8}\{\mathrm{Err}[P_l(d)] + \mathrm{Err}[P(k)] + \mathrm{Err}[R(d)] + \mathrm{Err}[J_f(d)] + \mathrm{Err}[C(d)] + \mathrm{Err}[P(D)] + \mathrm{Err}[N_c] + \mathrm{Err}[C_{avg}]\}, \quad (4)$$

where N_c is the total number of nodes in connected components of the network and C_{avg} is the undirected local clustering coefficient averaged over the N_c connected nodes. The local clustering coefficient of a node i is defined as the ratio between number of closed triads centered on node i and the total number of triads centered on that node.

The properties X integrating Err can be scalars, functions or distributions and encompass different orders of magnitude. We define the error of a property X as

$$\mathrm{Err}[X] = \frac{\sum_{i=1}^{n} |y_i^X - f_i^X|}{\sum_{i=1}^{n} |y_i^X|}, \quad (5)$$

where y_i^X is the i-th observed value of the property X, f_i^X is the corresponding i-th value of the property obtained by the model. In the case of a distribution, i runs over the n measured bins, while for a scalar (such as the number of nodes or the clustering coefficient) the sum has only one term.

We perform a Latin square sampling of the parameter space of p_v and p_c as shown in Figure 3 in order to find the minimum value of Err. The parameter space is covered uniformly in a linear scale for p_v and in a logarithmic one for p_c. For all the countries, the minimum value of the error is obtained for p_v in the interval $(0.05, 0.3)$ and p_c in the range $(5\cdot10^{-3}, 5\cdot10^{-2})$. The values of Err found at the minimum are 0.30 for the US, 0.18 for the UK and 0.39 for Germany. For simplicity, we focus on the Twitter networks only, although similar results are obtained for the other datasets.

Results

Simulations for the Optimal Parameters

An example with the displacements between the consecutive locations and the ego networks for a sample of individuals, as generated by the TF model, are displayed in Figure 4. The parameters of the model are set to the ones that correspond to the minimum of the error Err. As shown, the agents tend to stay close to their original positions. Occasional long jumps occur due to friend visits that live far apart. In this range of parameters and simulation times, the main mechanism for generating long distance connections is random linking (controlled by p_c). Agents typically return back to their original positions because this is where most of their contacts live. The frequency of the long distance jumps and connections varies for the three countries due to the different spatial distribution of the user populations. In the ego networks, the presence of multiple triangles with long distance edges can be observed.

The geo-social properties of the networks generated by the TF model are shown in Figure 5 for the US and in Figures S3 and S4 for the UK and Germany, respectively. Additionally, we show how each of the introduced properties contributes to the total error of the model in Table S1. The model is able to reproduce the trends in the probability $P_l(d)$, the reciprocity $R(d)$, the social overlap $J_f(d)$ and the disparity distribution $P(D)$ with good accuracy. The difficulties encountered with the degree distribution $P(k)$ and the clustering as a function of the distance $C(d)$ are not unexpected since the model does not incorporate mechanisms to explicitly enhance the heterogeneity in the agents' contacts nor favor any specific dependence of the clustering on the distance. We have tested variants of the TF model in which connections are created using the preferential attachment rule. The overall fitting error for these variants of the model is not lower than for the basic TF model, as we show in Appendix S1.

Insights of the TF Model

In this section we explore two null models uncoupling mobility and social interactions to help us interpret the mechanisms acting in the TF model. The first null model, the spatial model (S model), is based solely on the geography and consists of randomly connecting pair of users with a probability depending on the distance, but does not take network structure into account. The second null model, the linking model (L model), in contrast, is based only on random linking and triadic closure, and it is equivalent to the TF model without the mobility. We consider the two uncoupled null models and compare their results with those of the TF model. In this way, we demonstrate the importance of the coupling through a realistic mobility mechanism to reproduce the empirical networks.

The spatial model (S model) consists of randomly connecting pair of users with a probability that decays as power-law of the distance between them (suggested in [41]). The exponent of the power-law is fixed at -0.7 following Figure 2A. The results of the S model are shown in the panels of Figure 2. While it is set to match $P_l(d)$, other properties such as $P(k)$, $R(d)$, $J_f(d)$, $C(d)$ or $P(D)$ are not well reproduced. The S model fails to account for the high level of clustering and reciprocity in the empirical networks

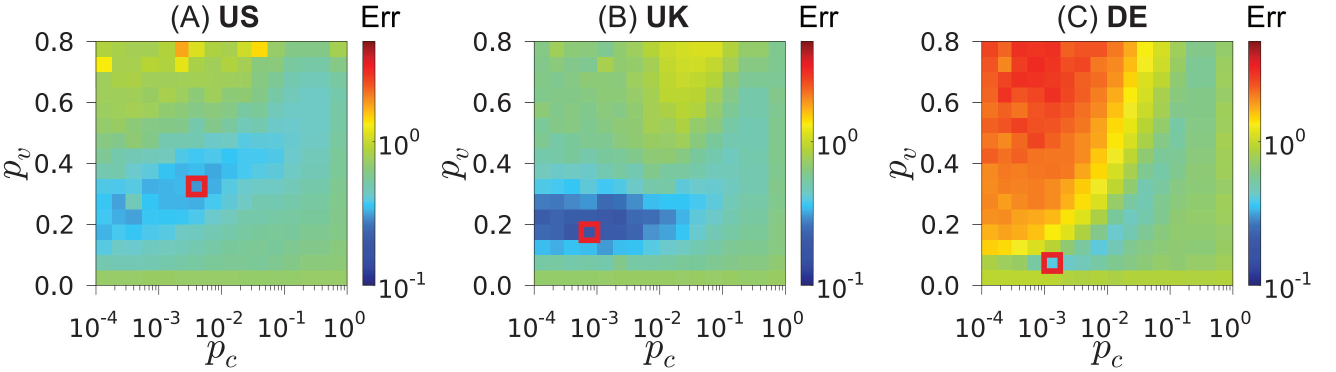

Figure 3. Fitting the TF model. Values of the error Err when p_v and p_c are changed. The minimum error for each of the plots is marked with a red rectangle.
doi:10.1371/journal.pone.0092196.g003

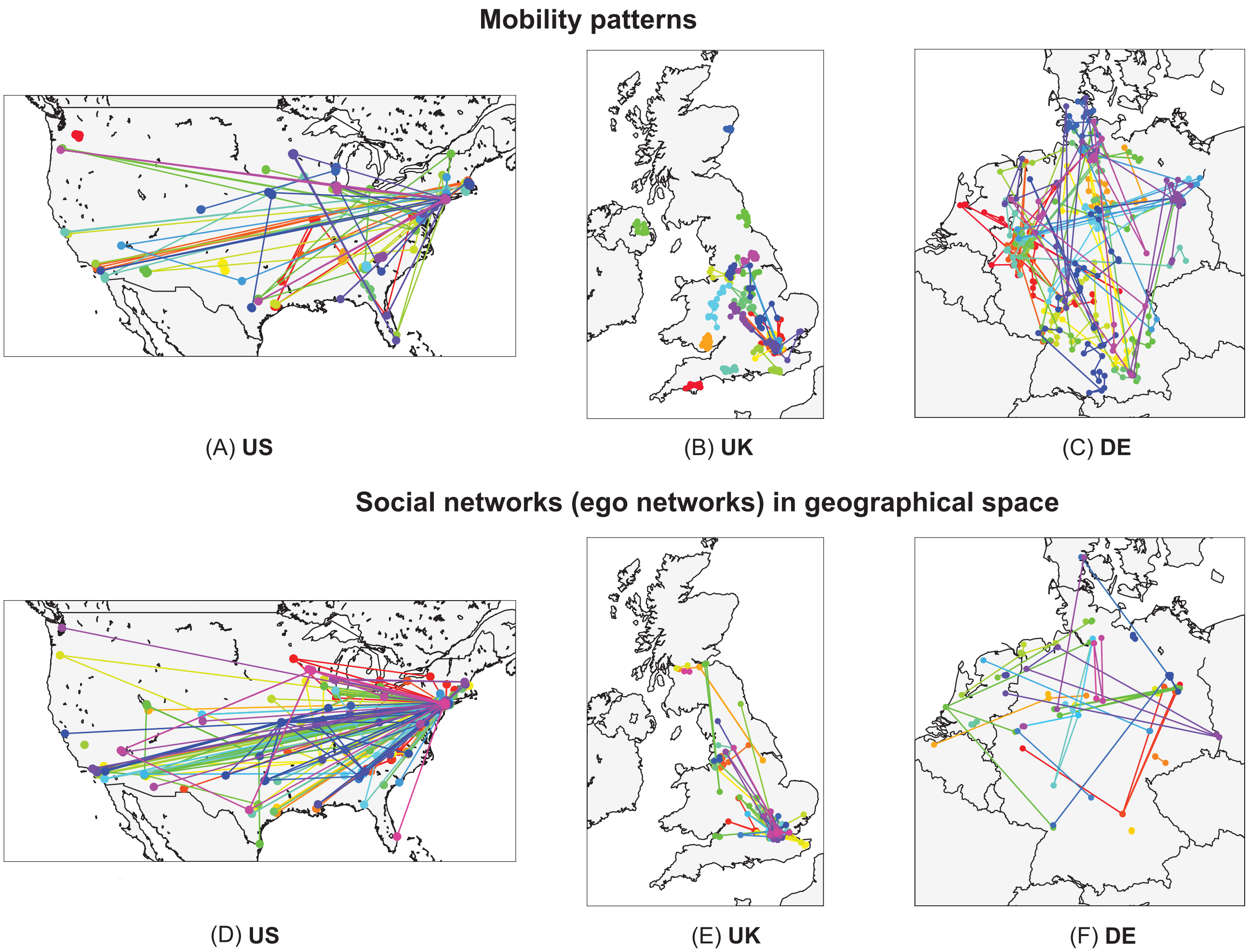

Figure 4. Simulation results: mobility and social networks. Mobility (upper row) and ego networks (lower row) of 20 random users (different colors) for the instances of the TF model yielding the lowest error Err (see Figure 3). Mobility network shows mobility patterns of individual users throughout entire simulation. Ego network shows the social connections at the end of the simulation.
doi:10.1371/journal.pone.0092196.g004

and for their dependence on the distance. The error Err of this null model is between 0.66–0.76 for the three countries, around twice the error of the TF model (see Figure 6).

The linking model (L model) is a simplified version of the TF model, without random mobility and the box size $\delta \to 0$. Agents move to visit their contacts with probability p_v, whereas with probability $1-p_v$ they do not perform any action. In this version of the model, users can connect only by random connections or when two of them coincide, visiting a common friend, which leads to triadic closure. These two processes do not depend on the distances between the users. A thorough description can be obtained with a mean-field approach (see the corresponding section). The results of the L model are shown in Figure 2. Due to the triangle closing mechanism, this null model creates networks with a considerable level of clustering. However, it does not reproduce the distance dependencies of $P_l(d)$, $R(d)$, $J_f(d)$ and $C(d)$. The error Err of the L model is also around twice higher than the error of the TF model (see Figure 6).

The geography and the structure are coupled in the TF model through the random mobility. Changes in the underlying mobility mechanism affect the quality of the results. The lowest Err values are obtained with the power-law distribution in the jump lengths, while normal or uniformly distributed jumps yield worse results (e.g., for the US the TF model has Err lower by 0.5 and 1.5 than the TF-normal and the TF-uniform models, respectively, as shown in Figure 6).

Simplified models that neglect either geography or network structure perform considerably worse than the TF model in reproducing the properties of real networks. Likewise, non-realistic assumptions on human mobility mechanism yield worse results than the default TF model. To conclude, the coupling of geography and structure through a realistic mobility mechanism produces networks with significantly more realistic geographic and structural properties.

Sensitivity of the TF Model to the Parameters and its Modifications

The results presented so far have been obtained at the optimal values of p_v and p_c. The question remains, however, of how robust these results are to changes in the values of the parameters. In Figure 7, we report the effect of varying p_v while p_c is maintained constant in its optimal value. The linking probability $P_l(d)$ loses its power-law shape for very low values of p_v, marking the limit in which random mobility is the main mechanism for the agents' traveling in detriment of friend visits. In this case, most of the links are created due to encounters occurring in nearby locations or are

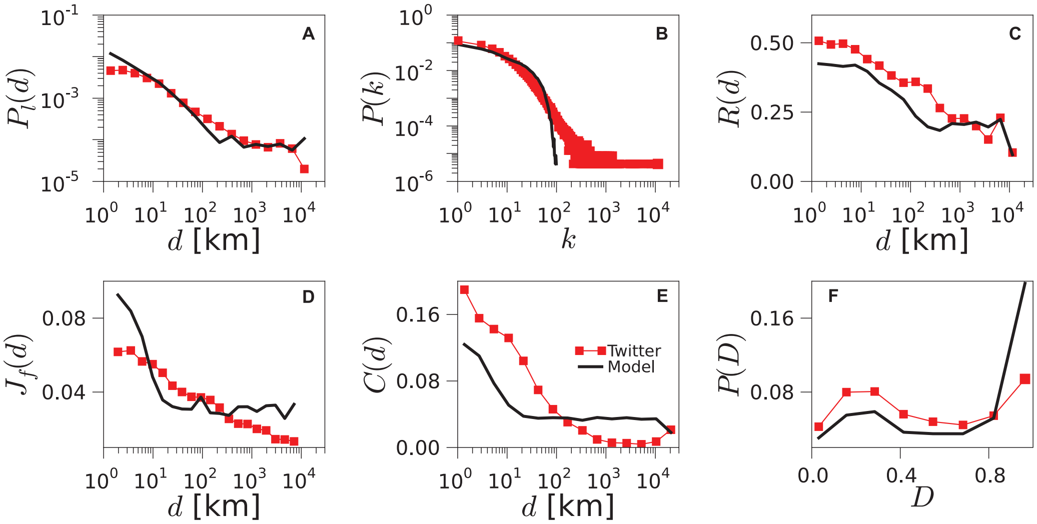

Figure 5. Geo-social properties of the model networks. Various statistical properties are plotted for the networks obtained from Twitter data (red squares) and from simulation of the TF model (black line) for the US. Corresponding results for the UK and Germany can be found in Figures S3 and S4.
doi:10.1371/journal.pone.0092196.g005

random connections, and so the distribution of triangles disparity $P(D)$ loses its bimodal shape. Furthermore, the friend visits provide opportunities to reciprocate the connections. This is why for extremely low values of p_v, the reciprocity $R(d)$ is close to zero. Towards the other limit, *i.e.*, $p_v \rightarrow 1$ the social overlap $J_f(d)$ and the triangle-closing probability $C(d)$ steadily increase. In this limit, the linking probability $P_l(d)$, the reciprocity $R(d)$ and the distribution of triangles disparity $P(D)$ recuperate their shapes of the optimum.

Figure 8 explores the impact of varying p_c while p_v is fixed to its optimal value. The effect of p_c on $J_f(d)$ and $C(d)$ is the opposite to that of p_v: these metrics decrease at all distances with increasing p_c. The reason for this is that visits to friends are the main forces behind the creation of new triads and the subsequent closure of triangles. Note that the more connections are created randomly (higher p_c), the less links will be a result of friend visits. We will expose and describe in detail the interplay between these two mechanisms in the mean-field calculations.

A possible variation of the TF model consists of eliminating friend visits or random connections (*i.e.*, setting p_v or p_c to 0). This prevents the model from producing networks with characteristics comparable to the real ones in all the cases, leading to increase in Err of around 0.5. Interestingly, the model results are quite robust to variations in the update rules, the random connection mechanism, the connecting rules in each agent neighborhood and the variants in the way users visit friends. These variations lead to changes in Err smaller than 0.1. A detailed discussion of the results with different model variants is included in Appendix S1.

Mean-field Approach

In this section, we consider the L model, introduced earlier in this section, to gain some analytical insights on the mechanisms ruling the final network structure. Although this model is a simplified version of the TF model, the results of the simulations yield a relatively low value of Err (Figures 6, and Figures S9 and S10 in Appendix S1). We write the equations for the time evolution of the properties of the network and solve them numerically. Among all the properties, we focus on the average

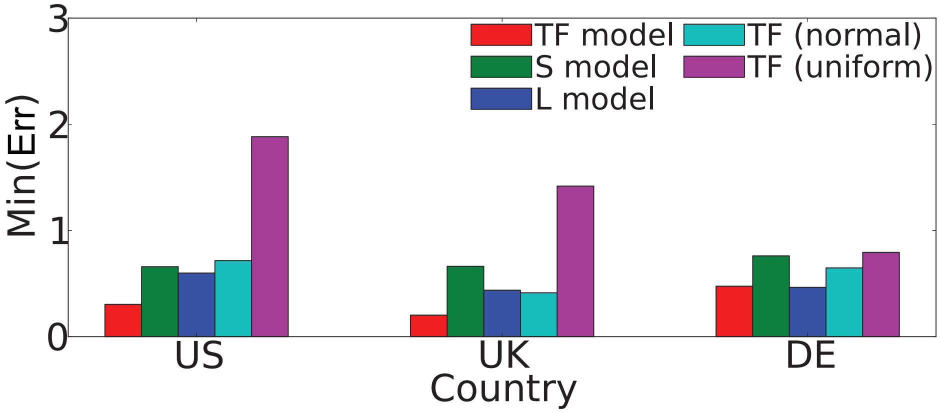

Figure 6. Comparison of different models. The minimal values of the error Err for the TF model, the two null models: spatial (S model) or linking (L model), and the TF model with normally or uniformly distributed travel distances.
doi:10.1371/journal.pone.0092196.g006

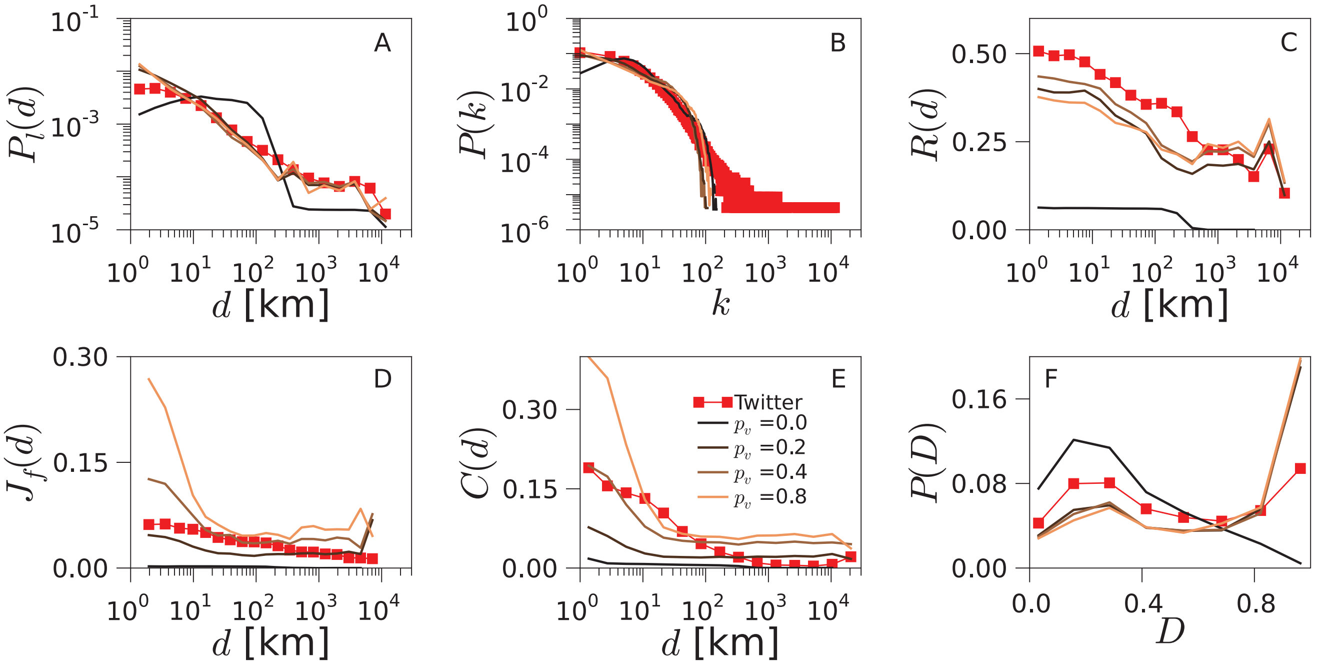

Figure 7. Impact of p_v on the TF model. We change the value of p_v while keeping p_c fixed to the optimal value. Note that this corresponds to an exploration of the parameter space along the vertical line crossing the minimum of Err as plotted in Figure 3 for the US. Corresponding results for the UK and Germany can be found in Figures S5 and S6.
doi:10.1371/journal.pone.0092196.g007

clustering coefficient C, the overall reciprocity R and the degree distribution $P(k)$.

The clustering coefficient is defined as a ratio of all the closed triads to all triads existing in the network, *i.e.*, $C = \Delta/\Lambda$. The number of triads Λ can be calculated knowing the degree distribution. The number of closed triads Δ in the L model grows with time mostly due to the friend visits mechanism. A triangle is formed every time two friends of the same hosting agent meet in the host's place and decide to connect. Note that an undirected triangle corresponds to 3 undirected closed triads. Assuming that

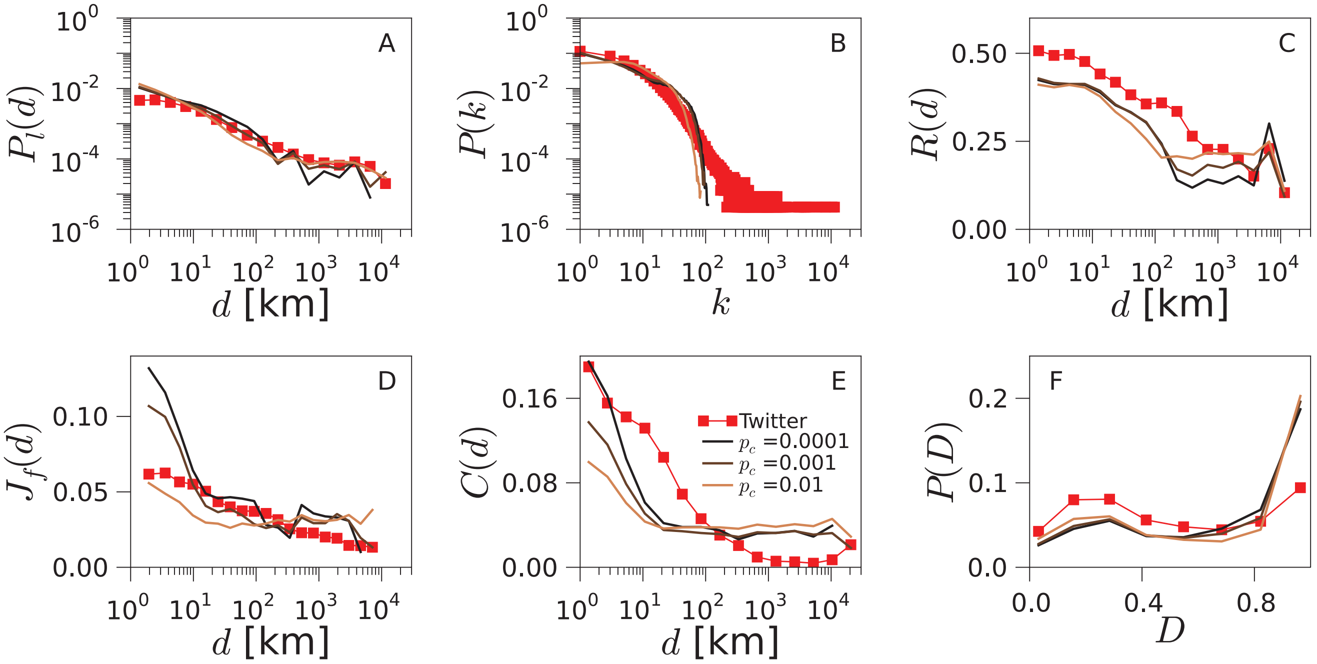

Figure 8. Impact of p_c on the TF model. We change the value of p_c while keeping p_v fixed to its optimal value. Note that this corresponds to an exploration of the parameter space along the horizontal line crossing the minimum of Err as plotted in Figure 3 for the US. Corresponding results for the UK and Germany can be found in Figure S7 and S8.
doi:10.1371/journal.pone.0092196.g008

the contribution of random links is negligible, the evolution of the number of closed triads is described by

$$\frac{d\Delta}{dt} = 3N(k>0)\left(1-(1-p)^2\right)(1-C)MS, \tag{6}$$

where $k=\left(k^{\mathrm{in}}+k^{\mathrm{out}}\right)/2$, meaning that we do not distinguish between in-degree and out-degree; $N(k>0)$ represents the number of nodes with the degree higher than 0, *i.e.*, the number of potential hosts, M is an estimate of the lower bound for the number of triangles closed by one closing link $M=1+C^2\left(\frac{2}{1+R}k-2\right)$. Finally, S is the expected number of encounters per host, which can be calculated as

$$S=\sum_{k=2}^{\infty}\frac{N(k)}{N} \times \sum_{i=2}^{k}\left(\frac{p_{\mathrm{v}}}{\langle k\rangle}\right)^i\left(1-\frac{p_{\mathrm{v}}}{\langle k\rangle}\right)^{k-i}\binom{k}{i}\binom{i}{2}, \tag{7}$$

where $N(k)$ is the number of nodes with a given degree k in the network. Finally, note that the above definition of degree and the one obtained from symmetrizing directed networks (used in previous sections) are related by a proportionality factor $k=k_{\mathrm{sym}}(1+R)/2$.

The reciprocity of connections R can be expressed as $R=L_{\mathrm{p}}/(L_{\mathrm{p}}+2L_{\mathrm{s}})$, where L_{p} is the number of reciprocated links, L_{s} is the number of non-reciprocated links and the total number of links $L=L_{\mathrm{s}}+L_{\mathrm{p}}$. The numbers of links evolve as

$$\frac{dL_{\mathrm{p}}}{dt}=2N(k>0)\{p_{\mathrm{rec}}+p^2(1-C)S+p(1-R)CS\}, \tag{8}$$

$$\frac{dL_{\mathrm{s}}}{dt}=p_{\mathrm{c}}N+\frac{1}{3M}\frac{d\Delta}{dt}-\frac{1}{2}\frac{dL_{\mathrm{p}}}{dt}, \tag{9}$$

where $p_{\mathrm{rec}}=pp_{\mathrm{v}}(1-p_{\mathrm{v}})(1-R)$ corresponds to the probability that an agent visiting a neighbor gets her connection reciprocated (their connection is initially single directional). As can be seen, Δ, L_{p} and L_{s} are mutually dependent.

To calculate the degree distribution $P(k)$, we estimate the probability p_{con} of a node to increase its degree by one unit in the current time step due to multiple encounters with friends of her friends

$$p_{\mathrm{con}}=\sum_{k'=2}^{\infty}\frac{k'N(k')}{\langle k\rangle N}\binom{k'-1}{2}p_{\mathrm{c}}^2(1-p_{\mathrm{c}})^{k'-2}, \tag{10}$$

where $p_{\mathrm{c}}=pp_{\mathrm{v}}/\langle k\rangle(1-(1+R)/2C)$. In the L model, however, every node can increase its degree by multiple links in each time step. For simplicity, we neglect higher order terms induced by the possibility of creating multiple links. Moreover, we note that Equation (10) is a good estimate if there is not a strong correlation between node degrees. The number of nodes of certain degree k is given by

$$\begin{aligned} k>1: \quad \frac{dN(k)}{dt} &= p_{\mathrm{inc}}(N(k-1)-N(k)), \\ \frac{dN(1)}{dt} &= p_{\mathrm{c}}N(0)-p_{\mathrm{inc}}N(1)+p_{\mathrm{rec}}N_{\mathrm{s}}(0), \\ \frac{dN(0)}{dt} &= -p_{\mathrm{c}}N(0)-p_{\mathrm{rec}}N_{\mathrm{s}}(0), \end{aligned} \tag{11}$$

where $p_{\mathrm{inc}}=p_{\mathrm{c}}+p_{\mathrm{rec}}/2+p_{\mathrm{v}}\,p_{\mathrm{con}}$ is an estimate of the probability that the node degree increases, $N_{\mathrm{s}}(0)$ is the number of nodes with 0 out-degree and non-zero in-degree. Such nodes are important because their connection can be easily reciprocated as a result of a friend visit. However, these nodes are not counted directly into $N(1)$, and so a correction is needed to account for them explicitly, as in Equation (11). The number of such nodes can be calculated as

$$\frac{dN_{\mathrm{s}}(0)}{dt}=p_{\mathrm{c}}N(0)-p_{\mathrm{rec}}N_{\mathrm{s}}(0). \tag{12}$$

The numerical solution of this set of equations describing the evolution of the L model is shown in Figure 9. The equations accurately predict the dynamics of the clustering coefficient C, the reciprocity R and the degree distribution $P(k)$ for certain values of the parameters (*i.e.*, for medium and high values of p_{c}, as in the lower plots of Figure 9). The approximation yields slightly worse results when the number of random connections is small in comparison with the number of connections created due to friend visits (*i.e.*, for low values of p_{c}, as in the upper plots of Figure 9B). In the latter case, neither the degree distribution is well approximated, probably due to the degree-degree correlations introduced through the friend visit mechanism.

The mean-field analysis of the L model shows that the friend visiting mechanism is a direct cause of triangle closure and link reciprocity. Equation 6, which estimates the growth of the number of triangles in the network, accounts only for the friend visiting mechanism; yet it approximates closely the value of the clustering coefficient, also when p_{c}, which controls the mechanism of random connections, is high. Similarly, Equation 8, which estimates the growth of the number of reciprocated connections, accounts for the friend visiting mechanism and approximates well the value of reciprocity.

Discussion

We introduce a model that couples human mobility and link creation in social networks. The aim is to characterize the relation between network topology and geography observed in empirical online networks. The model has two free parameters p_{c} and p_{v} but, despite its simplicity, it is able to reproduce a good number of geo-social features observed in real data at a country level. Comparing the TF model with simplified null models, we find that the coupling of geography and structure through a realistic mobility mechanism produces significantly more realistic social networks than the uncoupled models.

Social links in our model are formed mostly with relational (due to triadic closure), and proximity (through spatio-temporal coincidences) mechanisms [48]. Visiting friends helps to reinforce the existing relations and favors the closure of triads with particular properties regarding the distance balance of their edges. Random link creation accounts for online acquaintances or for historical face-to-face encounters as individuals move their

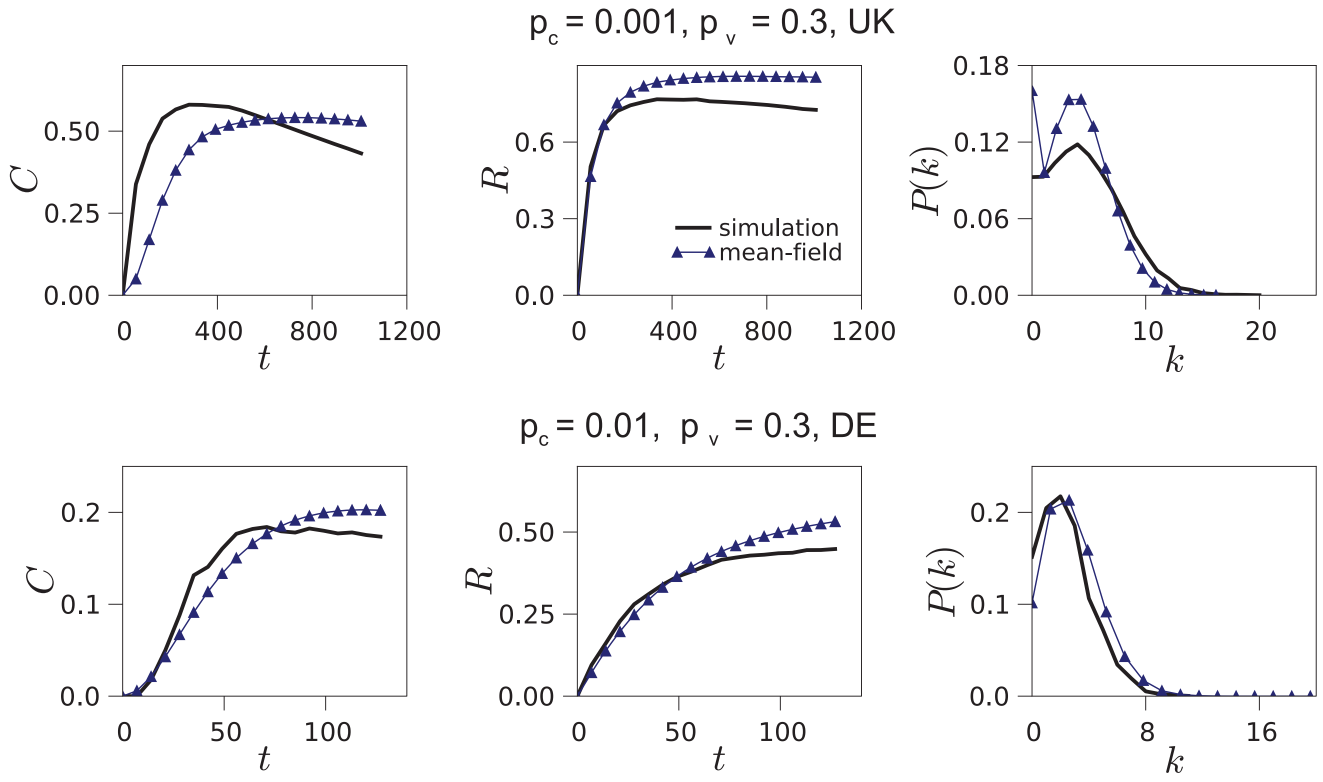

Figure 9. Mean-field approximation. Predictions of the analysis versus results of the simulation of the L model for the clustering coefficient *C*, the reciprocity *R* and the degree distribution $P(k)$. In this case, we are taking the users from the UK and Germany because their lower numbers facilitate the numerical integration of the Equations 6, 8, 9, 11 and 12.
doi:10.1371/journal.pone.0092196.g009

residence from one city to another. Finally, individual random mobility allows the agents to explore new locations. We expect that in our model the number of unique locations visited over time grow linearly in time, due to the fact that with constant probability $(1-p_v)$ an agent jumps to a new location, as opposed to slower growth reported in [4]. We leave the exploration of temporal aspects of our model for the future research. Our results show that by establishing an appropriate balance between friend visits and random link creation, the model can reproduce the main features of online social networks, e.g., we show that $10\%-30\%$ of the mobility has to be directed towards existing friends. We demonstrate that these are the fundamental mechanisms at play in the model.

The TF model is generic and functional for different datasets. Human mobility driven by social ties has impact on the modeling of disease spreading, and may improve its predictions. This model can also be used in simulations of processes that involve social networks and geography, e.g., simulations of opinion formation, language evolution, or responses of a population to extreme events. Moreover, it can also be helpful to design network benchmarks with realistic geo-social properties to test, for instance, the scalability of technical solutions in social online networks related to geography of its physical infrastructure.

Ethics Statement

The data analyzed are publicly available as they come from public online social sites or data repositories (Twitter, Gowalla and Brightkite). Since our analysis relies on statistical features and not on single cases, any private information about users had been removed and the analysis was performed on anonymized datasets.

Supporting Information

Figure S1 Network geo-social properties. Various statistical network properties are plotted for the data obtained from Twitter (red squares), Gowalla (blue diamonds), Brightkite (green triangles) and the null models (dashed lines), for the UK.
(EPS)

Figure S2 Network geo-social properties. Various statistical network properties are plotted for the data obtained from Twitter (red squares), Gowalla (blue diamonds), Brightkite (green triangles) and the null models (dashed lines), for Germany.
(EPS)

Figure S3 Geo-social properties of the model networks. Various statistical properties are plotted for the networks obtained from Twitter data (red squares) and from simulation of the TF model (black line) for the UK.
(EPS)

Figure S4 Geo-social properties of the model networks. Various statistical properties are plotted for the networks obtained from Twitter data (red squares) and from simulation of the TF model (black line) for Germany.
(EPS)

Figure S5 Impact of p_v on the TF model. We change the value of p_v while keeping p_c fixed to the optimal value. Note that this corresponds to an exploration of the parameter space along the vertical line crossing the minimum of Err as plotted in Figure 3 for the UK.
(EPS)

Figure S6 Impact of p_v on the TF model. We change the value of p_v while keeping p_c fixed to the optimal value. Note that this corresponds to an exploration of the parameter space along the vertical line crossing the minimum of Err as plotted in Figure 3 for Germany.
(EPS)

Figure S7 Impact of p_c on the TF model. We change the value of p_c while keeping p_v fixed to its optimal value. Note that this corresponds to an exploration of the parameter space along the horizontal line crossing the minimum of Err as plotted in Figure 3 for the UK.
(EPS)

Figure S8 Impact of p_c on the TF model. We change the value of p_c while keeping p_v fixed to its optimal value. Note that this corresponds to an exploration of the parameter space along the horizontal line crossing the minimum of Err as plotted in Figure 3 for Germany.
(EPS)

Table S1 The contribution of each of the properties to the total error of the TF model. Value of the error $\mathrm{Err}[X]$ per property X at the minimum of the total error Err for Twitter for the three considered countries.
(PDF)

Appendix S1 Variants of the TF model. In this appendix, we consider several variants of the TF model and the L model and evaluate their results. Appendix S1 contains Figures S9 and S10.
(PDF)

Acknowledgments

We would like to warmly thank Luis F. Lafuerza for helpful discussions on the analytical treatment of the model.

Author Contributions

Conceived and designed the experiments: PAG JJR BG VME. Performed the experiments: PAG. Analyzed the data: PAG. Contributed reagents/materials/analysis tools: PAG JJR BG VME. Wrote the paper: PAG JJR BG VME.

References

1. Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, et al. (2009) Computational social science. Science 323: 721.
2. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439: 462–5.
3. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453: 779–82.
4. Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. Nature Physics 6: 818–823.
5. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World Wide Web - WWW '10, p.591–600.
6. Mislove A, Koppula HS, Gummadi KP, Druschel P, Bhattacharjee B (2008) Growth of the flickr social network. Proceedings of the first workshop on Online Social Networks - WOSP '08. 25–30.
7. Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, et al. (2013) Time as a limited resource: Communication strategy in mobile phone networks. Social Networks 35: 89–95.
8. Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on twitter networks: Validation of Dunbar's number. PLoS ONE 6: e22656.
9. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. Proceedings of the 21st international conference on World Wide Web - WWW '12, 519–528.
10. Ugander J, Backstrom L, Marlow C, Kleinberg J (2012) Structural diversity in social contagion. Proc Natl Acad Sci (USA) 109: 5962–5966.
11. Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '09, p.497–506.
12. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C (2012) Dynamical classes of collective attention in Twitter. Proceedings of the 21st international conference on World Wide Web - WWW '12. p.251–260.
13. Grabowicz PA, Ramasco JJ, Moro E, Pujol JM, Eguíluz VM (2012) Social features of online networks: the strength of intermediary ties in online social media. PLoS ONE 7: e29358.
14. Grabowicz PA, Aiello LM, Eguíluz VM, Jaimes A (2013) Distinguishing topical and social groups based on common identity and bond theory. Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13, p.627–636.
15. Ferrara E (2012) A large-scale community structure analysis in Facebook. EPJ Data Science 1: 9.
16. Borge-Holthoefer J, Rivero A, García I, Cauhé E, Ferrer A, et al. (2011) Structural and dynamical patterns on online social networks: The Spanish may 15th movement as a case study. PLoS ONE 6: e23883.
17. González-Bailón M, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Scientific Reports 1: 197.
18. Conover MD, Davis C, Ferrara E, McKelvey K, Menczer F, et al. (2013) The geospatial character-istics of a social movement communication network. PLoS ONE 8: e55957.
19. Lambiotte R, Blondel V, Dekerchove C, Huens E, Prieur C, et al. (2008) Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and its Applications 387: 5317–5325.
20. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: A model for inter-city telecom-munication flows. Journal of Statistical Mechanics: Theory and Experiment 2009: L07003.
21. Phithakkitnukoon S, Smoreda Z, Olivier P (2012) Socio-geography of human mobility: A study using longitudinal mobile phone data. PLoS ONE 7: e39253.
22. 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.
23. Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, et al. (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci (USA) 107: 22436–22441.
24. González MC, Lind PG, Herrmann HJ (2006) System of mobile agents to model social networks. Phys Rev Lett 96: 088072.
25. Backstrom L, Sun E, Marlow C (2010) Find me if you can: Improving geographical prediction with social and spatial proximity. Proceedings of the 19th international conference on World Wide Web - WWW '10, p.61–70.
26. Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. Proceedings of the 5th international AAAI conference on weblogs and social media - ICWSM'11, p.329–336.
27. Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. Social Networks 34: 73–81.
28. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci (USA) 109: 11576–11581.
29. Volkovich Y, Scellato S, Laniado D, Mascolo C, Kaltenbrunner A (2012) The length of bridge ties: Structural and geographic properties of online social interactions. Proceedings of the 6th international AAAI conference on weblogs and social media - ICWSM '12, p.346–353.
30. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '11, p.1100–1108.
31. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: User movement in location-based social networks. Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining - KDD '11, p.1082–1090.
32. Sadilek A, Kautz H, Bigham JP (2012) Finding your friends and following them to where you are. Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12, p.723–732.
33. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, et al. (2009) Multiscale mobility networks and the large scale spreading of infectious diseases. Proc Natl Acad Sci (USA) 106: 21484–21489.
34. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. Science 324: 1071–1076.
35. Brockmann D (2010) The physics of where to go. Nature Physics 6: 720.
36. Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. Nature 484: 96–100.
37. Jia T, Jiang B, Carling K, Bolin M, Ban Y (2012) An empirical study on human mobility and its agent-based modeling. Journal of Statistical Mechanics: Theory and Experiment 2012: P11024.
38. Szell M, Sinatra R, Petri G, Thurner S, Latora V (2012) Understanding mobility in a social petri dish. Scientific Reports 2: 457.
39. Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. Journal of Statistical Physics 151: 304–318.
40. Giannotti F, Pappalardo L, Pedreschi D, Wang D (2013) Mobility data: Modeling, management, and understanding In: Renso C, Spaccapietra S, Zimańyi E, editors. Cambridge University Press.

41. Butts CT, Acton RM, Hipp JR, Nagle NN (2012) Geographical variability and network structure. Social Networks 34: 82–100.
42. Twitter API, section for developers of Twitter Web page. Available: https://dev.twitter.com. Accessed 2014 Feb 6.
43. Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, et al. (2011) Truthy: Mapping the spread of astroturf in microblog streams. Proceedings of the 20th international conference companion on World Wide Web - WWW '11, p.249–252.
44. Data available at the Stanford large network dataset collection. Available: http://snap. stanford.edu/data. Accessed 2014 Feb 6.
45. Ugander J, Karrer B, Backstrom L, Marlow C (2011) The anatomy of the Facebook social graph. arXiv:1111.4503.
46. State B, Weber I, Zagheni E (2013) Studying inter-national mobility through IP geolocation. Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13, p.265–274.
47. Gridded Population of the World, Version 3 (GPWv3): Population Count Grid, Future Estimates. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Available: http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count-future-estimates. Ac-cessed 2014 Feb 6.
48. Rivera MT, Soderstrom SB, Uzzi B (2010) Social networks: Assortative, relational, and proximity mechanisms. Annu Rev Sociol 36: 91–115.