

Dynamical strength of social ties in information spreading

Giovanna Miritello,^{1,2} Esteban Moro,^{1,3,4} and Rubén Lara²

¹*Grupo Interdisciplinar de Sistemas Complejos, Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Spain*

²*Telefónica Research, Madrid, Spain*

³*Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM, 28049 Madrid, Spain*

⁴*Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, 28049 Madrid, Spain*

(Received 22 November 2010; revised manuscript received 17 March 2011; published 27 April 2011)

We investigate the temporal patterns of human communication and its influence on the spreading of information in social networks. The analysis of mobile phone calls of 20 million people in one country shows that human communication is bursty and happens in group conversations. These features have the opposite effects on the reach of the information: while bursts hinder propagation at large scales, conversations favor local rapid cascades. To explain these phenomena we define the dynamical strength of social ties, a quantity that encompasses both the topological and the temporal patterns of human communication.

DOI: [10.1103/PhysRevE.83.045102](https://doi.org/10.1103/PhysRevE.83.045102)

PACS number(s): 89.65.–s, 89.75.–k, 05.10.–a

A quantitative understanding of human communication patterns is of paramount importance in the explanation of the dynamics of many social, technological, and economic phenomena [1–4]. Most studies have focused on the complex topological patterns of the underlying contact network (whom we talk to) and its influence on the properties of spreading phenomena in social networks such as diffusion of information, innovations, computer viruses, and opinions [2]. Paradoxically, most of these studies of dynamical phenomena on social networks neglect the temporal patterns of human communication: humans act in bursts or cascades of events [5–8], most ties are not persistent [9,10], and communication happens mostly in the form of group conversations [8,11–14]. However, since information transmission and human communication are concurrent, the temporal structure of communication must influence the properties of information spreading. Indeed, recent experiments of electronic recommendation forwarding [15] and simulations of epidemic models on email and mobile databases [6,16] found that the asymptotic speed of information spreading is controlled by the bursty nature of human communication, which leads to a slowing down of the diffusion. However, even though the asymptotic speed is an important property of the propagation of information in social networks, there is still no general understanding of how and what temporal properties of human communication influence spreading processes and how they affect the very definition of social interaction.

The answer to these questions can be framed in the more general problem of how to model dynamical social networks [9,17]. In most studies, real temporal activity is aggregated over time, thus giving a static snapshot of the social interaction where ties are described by static strengths that do not include information about the temporal aspects of how humans interact. Temporal and topological aspects are therefore disentangled in the analysis. In this Rapid Communication we merge both aspects in the case of information diffusion by adopting a functional definition of the social ties using the well-known map between dynamical epidemic models and static percolation [18]. The network is still described by a static graph, but the interaction strength between individuals

now incorporates the causal and temporal patterns of their communications and not only the intensity [19].

To this end we study the mobile communication patterns from a European operator in a single country over a period of 11 months. The data consist of 2×10^7 phone numbers and 7×10^8 communication ties for a total of 9 billion calls between users. The call detail record (CDR) contains the hashed number of the caller and the receiver, the time when the call was initiated, and the duration of the call. We consider only events in which the caller and the receiver belong to the operator under consideration because of the partial access to the records of other operators. Our data for the connectivity of the social network, the duration of the calls, etc., are very similar to the results reported in previous studies [19].

First we investigate the communication temporal patterns that might affect information diffusion. The spreading from user i to user j ($i \rightarrow j$) happens at the relay time intervals τ_{ij} (also called intercontact time [8]), i.e., the time interval it takes for i to pass on to j any information received from any another person $* \rightarrow i$ (where $* \neq j$; see Fig. 1). Information spreading is thus determined by the interplay between τ_{ij} and the intrinsic time scale of the infection process. Note that τ_{ij} depends on the correlated and causal way in which group conversations happen since it depends on the interevent intervals δt_{ij} in the $i \rightarrow j$ communication as well as on the possible temporal correlation with the $* \rightarrow i$ events [18]. By ignoring this correlation it is possible to approximate the probability distribution function (PDF) for τ_{ij} by the waiting-time density for δt_{ij} [6,16],

$$P(\tau_{ij}) = \frac{1}{\overline{\delta t_{ij}}} \int_{\tau_{ij}}^{\infty} P(\delta t_{ij}) d\delta t_{ij}, \quad (1)$$

where $\overline{\delta t_{ij}}$ is the average interevent time. In this approximation the dynamics of the transmission process depends only on the dyadic $i \rightarrow j$ sequence of communication events and, in particular, the possible heavy-tail properties of $P(\delta t_{ij})$ are directly inherited by $P(\tau_{ij})$. Figure 2 shows our (rescaled) results for $P(\delta t_{ij})$ and $P(\tau_{ij})$. For comparison we also show the results obtained (i) when the time stamps of the $* \rightarrow i$ events are randomly selected from the complete CDR, thus

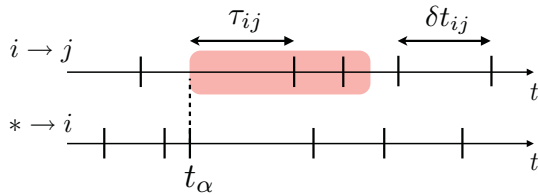


FIG. 1. (Color online) Schematic view of communication events around an individual i : Each vertical segment indicates an event between $i \rightarrow j$ (top) and $* \rightarrow i$ (bottom). At each t_α in the $* \rightarrow i$ time series, τ_{ij} is the time elapsed to the next $i \rightarrow j$ event, which is different from the interevent time δt_{ij} in the $i \rightarrow j$ time series. The shaded area represents the recovery time window T_i after t_α .

destroying any possible temporal correlation with $i \rightarrow j$ and effectively mimicking Eq. (1), and (ii) when the whole CDR time stamps are shuffled, thus destroying both tie temporal patterns and the correlation between ties. Both shufflings preserve the tie intensity w_{ij} [19], i.e., the number of calls and their duration and also the circadian rhythms of human communication [16]. The result for $P(\delta t_{ij})$ shows that small and large interevent times are more probable for the real-time series than for the shuffled-time series, where the PDF is almost exponential as in a Poissonian process, apart from a small deviation due to the circadian rhythms. This bursty pattern of activity has been found in numerous examples of human behavior [6] and seems to be universal in the way a single individual schedules tasks. Here we see that it also happens at the level of two individuals interacting, thus confirming recent results in mobile [16] and online community [7] dynamics. The PDF for τ_{ij} is also heavy tailed, but displays a larger number of short τ_{ij} compared to the shuffled series of events. The abundance of short τ_{ij} suggests that receiving information ($* \rightarrow i$) triggers communication with other people ($i \rightarrow j$), a manifestation of group conversations [11,12,14]. While the

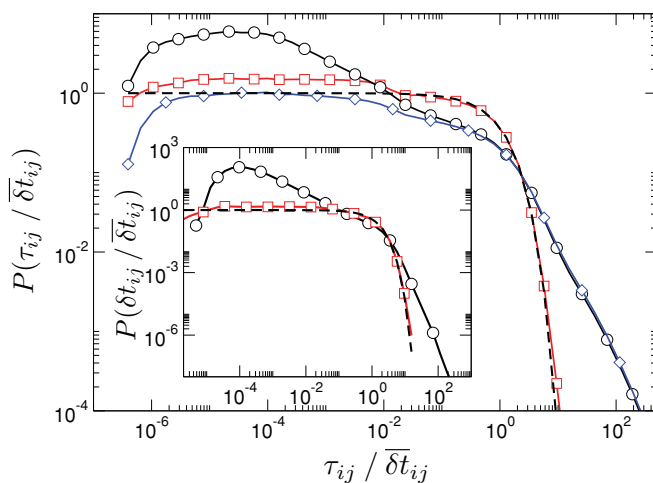


FIG. 2. (Color online) Distribution of the relay time intervals τ_{ij} (main part) and of the interevent times δt_{ij} (inset) in the $i \rightarrow j$ tie rescaled by δt_{ij} . The open circles correspond to the real data, while the open squares are the overall shuffled results. Open diamonds correspond to the case in which only the $* \rightarrow i$ sequence is randomized. Only ties with $w_{ij} \geq 10$ are considered. In both graphs the dashed line corresponds to the e^{-x} function.

heavy tail of $P(\tau_{ij})$ is accurately described by Eq. (1), i.e., large transmission intervals τ_{ij} are mostly due to large interevent communication times in the $i \rightarrow j$ tie, the behavior of $P(\tau_{ij})$ is due not only to the bursty patterns of δt_{ij} , but also to the temporal correlation between the $i \rightarrow j$ and the $* \rightarrow i$ series is destroyed, the probability of short-time intervals decreases and approaches the Poissonian case (Fig. 2). In summary, relay times depend on two main properties of human communication that compete with one another. While the bursty nature of human activity yields large transmission times, thus hindering any possible infection, group conversations translate into an unexpected abundance of short relay times, favoring the probability of propagation.

To investigate the effect of these two conflicting properties of human communication on information spreading, we simulate the epidemic susceptible-infectious-recovered (SIR) model in our social network considering the real-time sequence of communication events [14,16] and compare the results with the shuffled-time data. We start the model by infecting a node at a random instant and considering all other nodes as susceptible. In each call an infected node can infect a susceptible node with probability λ . Due to the synchronous nature of the phone communication, this happens regardless of who initiates the call. However, since the same results are obtained by considering directionality in the calls, for computational reasons we consider the latter case. Nodes remain infected during a time T_i until they decay into the recovered state. For the sake of simplicity we simulate the simplest model in which the recovery time T_i is deterministic and homogeneous, $T_i = T$ and $T = 2$ days, although different and/or stochastic T_i can be studied within the same model. The spreading dynamics generates a viral cascade that grows until there are no more nodes in the infected state. We repeat the spreading process for 3×10^4 randomly chosen seeds. Note that our model includes the SI model simulations in Ref. [16], where $\lambda = 1$ and $T = T_0$, with T_0 being the total duration of the dataset.

By looking at the size of the largest cascade s_{\max} (over all realizations) at each value of λ , we first ensure the existence of a percolation transition [4] (see Fig. 3), which is confirmed by a change in the behavior of s_{\max} from small to large cascades at a given value of λ (tipping point). The same behavior is observed for the shuffled-time data where the transition seems to happen almost at the same value of λ , although an accurate analysis of the percolation point is beyond the scope of this Rapid Communication. In contrast, there is a significant difference in the behavior of the asymptotic average size s_∞ between the real-time data and the shuffled-time data for different regimes of λ : when λ is small, s_∞ is larger for the real-time data than for the shuffled-time data, while the opposite behavior is observed for large λ . This difference, which can be very large for moderate values of λ , shows the impact of the real-time dynamics of communication on the influence of information in society. Specifically, if information propagates easily (large λ), the average extent in social networks is narrower than the one expected when a Poissonian dynamics is considered. In this sense, temporal patterns make social networks bigger than expected at large scales. However, in most real situations λ is very small [15] and in this case the observed behavior is the

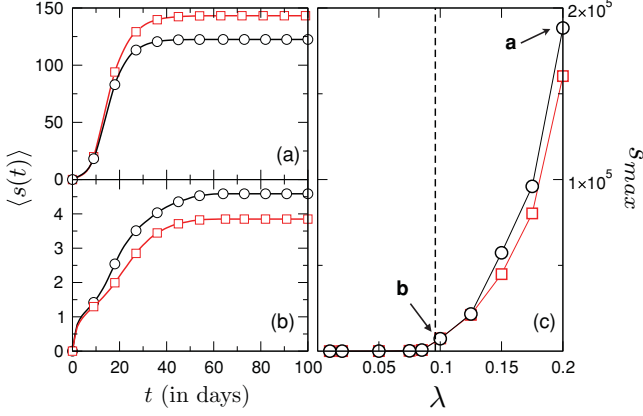


FIG. 3. (Color online) Average size dynamics for (a) a large and (b) a small value of λ (left) and the maximum size (right) of the infection outbreaks (over 10^4 realizations) for the real-time data (lines with open circles) and shuffled-time data (lines with open squares) for $T = 2$ days. The dashed line shows the critical point estimation of the percolation transition given by $R_1[\lambda, T] = 1$, with R_1 calculated using Eq. (6).

opposite: despite the low propagation, information cascades are larger in real data than in the Poisson case, which suggests that information spreading is more efficient at small (local) scales.

To understand this behavior, we follow the approach of Ref. [18] by mapping the dynamical SIR model to a static edge percolation model where each tie is described by the transmissibility \mathcal{T}_{ij} , which represents the probability that the information is transmitted from i to j and is a function of λ and T . If user i becomes infected at time t_α and the number of communication events $i \rightarrow j$ in the interval $[t_\alpha, t_\alpha + T]$ is $n_{ij}(t_\alpha)$, then the transmissibility in that interval is $\mathcal{T}_{ij} = 1 - (1 - \lambda)^{n_{ij}(t_\alpha)}$ (see Fig. 1). User i may become infected at any $* \rightarrow i$ communication event. If we assume that these events are independent and equally probable, we can average \mathcal{T}_{ij} over all the t_α events to get

$$\mathcal{T}_{ij}[\lambda, T] = \langle 1 - (1 - \lambda)^{n_{ij}(t_\alpha)} \rangle_\alpha. \quad (2)$$

If the number of $* \rightarrow i$ events is large enough we could use a probabilistic description of Eq. (2) in terms of the probability $P(n_{ij} = n; T)$ that the number of communication events between i and j in a given time interval T is n . Thus

$$\mathcal{T}_{ij}[\lambda, T] = \sum_{n=0}^{\infty} P(n_{ij} = n; T) [1 - (1 - \lambda)^n], \quad (3)$$

which, in principle, can be nonsymmetric ($\mathcal{T}_{ij} \neq \mathcal{T}_{ji}$). This quantity represents the real probability of infection from i to j and defines the dynamical strength of the tie. Note that \mathcal{T}_{ij} depends on the series of communication events between i and j , but also on the time series of calls received by i . In Ref. [18] Newman studied the case in which both time series are given by independent Poisson processes in the whole observation interval $[0, T_0]$. Thus $P(n_{ij} = n; T)$ is the Poisson distribution with rate $\rho_{ij} = w_{ij}T/T_0$, where w_{ij} is the total number of calls from i to j in $[0, T_0]$, and so

$$\tilde{\mathcal{T}}_{ij}[\lambda, T] = 1 - e^{-\lambda\rho} = 1 - e^{-\lambda w_{ij}T/T_0}, \quad (4)$$

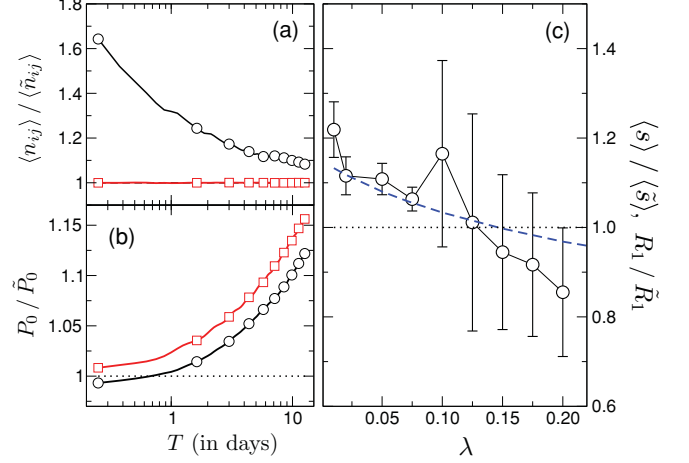


FIG. 4. (Color online) (a) Ratio of the number of events and (b) probability of no events as a function of the recovery time T for the real-time (open circles) and shuffled-time $* \rightarrow i$ (open squares) data with respect to the overall shuffled data. (c) Ratio of the average size of the outbreaks (open circles) and of R_1 calculated using Eq. (6) (dashed line).

which shows the one-to-one relationship between the intensity w_{ij} and the transmissibility \mathcal{T}_{ij} in the Poissonian case: the more intense the communication is, the larger the probability of infection. However, as we have seen in Fig. 2, the real $i \rightarrow j$ and $* \rightarrow i$ series are far from being independent and Poissonian and in order to investigate the effect of real patterns of communication on the transmissibility we approximate Eq. (2). For small values of λ we have $1 - (1 - \lambda)^n \simeq \lambda n$, while for $\lambda \simeq 1$ we get that $1 - (1 - \lambda)^n \simeq 1$ for $n > 0$. Thus the transmissibility for the two regimes is given by

$$\mathcal{T}_{ij}[\lambda, T] = \begin{cases} \lambda \langle n_{ij} \rangle_{t_\alpha} & \text{for } \lambda \ll 1 \\ 1 - P_{ij}^0 & \text{for } \lambda \simeq 1, \end{cases} \quad (5)$$

where $P_{ij}^0 = P(n_{ij} = 0; T)$. Specifically, P_{ij}^0 can be estimated directly from Eq. (1) for each link $P_{ij}^0 = \int_T^\infty P(\tau_{ij}) d\tau_{ij}$ since it measures the probability of finding a relay time bigger than T . Figure 4 shows the comparison of n_{ij} and P_{ij}^0 (averaged over all links) for different values of T for the real- and shuffled-time data (denoted by a tilde). On one side, due to the correlation between the $* \rightarrow i$ and $i \rightarrow j$ time series, the number of events in a tie following an incoming call is always larger for the real-time data than for the shuffled-time data. This is the reason why, for small λ , the average transmissibility (and thus the size of the epidemic cascades) is always higher in real communication patterns [14]. In contrast, the bursty nature of the $i \rightarrow j$ communication makes the tail for the real-time $P(\tau_{ij})$ heavier than the exponential distribution found in the shuffled-time data. Thus, if T is large enough, P_{ij}^0 is larger in the real-time data than in the shuffled-time data and this is why we observe smaller cascades in that region. Note, however, that this does not apply for very small values of T ($T \lesssim 1$ day), where the causality between the $* \rightarrow i$ and $i \rightarrow j$ time series can make P_0 even smaller in the real-time case.

To give a more quantitative analysis of the observed behavior we investigate the percolation process in a social

network in which links have transmissibility \mathcal{T}_{ij} . The important quantity is the secondary reproductive number R_1 , which is the average number of secondary infections produced by an infectious individual. R_1 gives information about percolation transition in the SIR process (which happens at $R_1 = 1$ [18]), but also about the speed of diffusion (which is proportional to R_1 [20]) and the size of the cascades (which is an increasing function of R_1 [18]). If we assume that the \mathcal{T}_{ij} are given and that the social network is random in any other respect, R_1 can be approximated as

$$R_1[\lambda, T] = \frac{\langle (\sum_j \mathcal{T}_{ij})^2 \rangle_i - \langle \sum_j \mathcal{T}_{ij}^2 \rangle_i}{\langle \sum_j \mathcal{T}_{ij} \rangle_i}. \quad (6)$$

Note that in the homogeneous case in which $\mathcal{T}_{ij} = \mathcal{T}$ we recover the common result in random networks $R_1 = \mathcal{T}(\langle k_i^2 \rangle / \langle k_i \rangle - 1)$ [18]. Figures 3 and 4 show the accuracy of the approximations used to get Eq. (6) to predict the tipping point in the SIR process and the change in the average size of the cascades in the two regimes. This suggests that the dynamical strength of the ties \mathcal{T}_{ij} , defined in Eq. (2), can be effectively used to model the real strength of human interactions in social networks.

In conclusion, we have seen that both the bursty nature of human communications and the existence of group conversations are the two main dynamical ingredients in the understanding of the spread of information in social networks. These two effects compete in the spreading of information by promoting and hindering the reach of the information compared with the homogeneous case. Our results indicate the necessity to incorporate temporal patterns of communication in the description and modeling of human interaction. Actually, we have proved an effective way of mapping the dynamics of human interactions onto a static representation of the social network through the concept of the dynamical strength of ties. We believe its success in explaining information diffusion would encourage the use of this dynamical strength in other areas of network research that are based on information spreading such as the determination of influence (or centrality) or popularity [21,22], community finding [23], and viral marketing [14,15].

We thank J.L. Iribarren and R. Cuerno for discussions and Telefónica for access to anonymized data. G.M. and E.M. acknowledge support from Ministerio de Educación y Ciencia (Spain) through Ingenio Mathematica projects i-MATH and MOSAICO.

-
- [1] D. Lazer *et al.*, *Science* **323**, 721 (2009).
 - [2] A. Barrat, M. Barthélémy, and A. Vespignani, *Dynamical Process on Complex Networks* (Cambridge University Press, Cambridge, 2008).
 - [3] C. Castellano, S. Fortunato, and V. Loreto, *Rev. Mod. Phys.* **81**, 591 (2009).
 - [4] M. E. J. Newman, *SIAM (Soc. Ind. Appl. Math.) Rev.* **45**, 167 (2003).
 - [5] A.-L. Barabási, *Nature (London)* **435**, 207 (2005).
 - [6] A. Vázquez *et al.*, *Phys. Rev. Lett.* **98**, 158702 (2007).
 - [7] D. Rybski *et al.*, *Proc. Natl. Acad. Sci. USA* **106**, 12640 (2009).
 - [8] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, *Journal of Theoretical Biology* **271**, 166 (2011).
 - [9] G. Kossinets and D. J. Watts, *Science* **311**, 88 (2006).
 - [10] C. A. Hidalgo and C. Rodriguez-Sickert, *Physica A* **387**, 3017 (2008).
 - [11] J.-P. Eckmann, E. Moses, and D. Sergi, *Proc. Natl. Acad. Sci. USA* **101**, 14333 (2004).
 - [12] Y. Wu *et al.*, *Proc. Natl. Acad. Sci. USA* **107**, 18803 (2010).
 - [13] C. Cattuto *et al.*, *PLoS ONE* **5**, e11596 (2010).
 - [14] Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, and W.-C. Lee *Proceeding CIKM '10 Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (ACM, New York, NY, USA, 2010).
 - [15] J. L. Iribarren and E. Moro, *Phys. Rev. Lett.* **103**, 038702 (2009).
 - [16] M. Karsai *et al.*, *Phys. Rev. E* **83**, 025102(R) (2011).
 - [17] A. Gautreau, A. Barrat, and M. Barthélémy, *Proc. Natl. Acad. Sci. USA* **106**, 8847 (2009).
 - [18] M. E. J. Newman, *Phys. Rev. E* **66**, 16128 (2002); E. Kenah and J. M. Robins, *ibid.* **76**, 036113 (2007).
 - [19] J.-P. Onnela *et al.*, *Proc. Natl. Acad. Sci. USA* **104**, 7332 (2007).
 - [20] M. Barthelemy *et al.*, *Phys. Rev. Lett.* **92**, 178701 (2004).
 - [21] M. E. J. Newman, *Soc. Networks* **27**, 39 (2005).
 - [22] J. Ratkiewicz *et al.*, *Phys. Rev. Lett.* **105**, 158701 (2010).
 - [23] S. Fortunato, *Phys. Rep.* **486**, 75 (2010).