Impact of Human Activity Patterns on the Dynamics of Information Diffusion

José Luis Iribarren
IBM Corporation, ibm.com e-Relationship Marketing Europe, 28002 Madrid, Spain

Esteban Moro
Instituto de Ciencias Matemáticas CSIC-UAM-UC3M-UCM,
Departamento de Matemáticas & GISC, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), Spain

(Received 27 January 2009; published 14 July 2009)

We study the impact of human activity patterns on information diffusion. To this end we ran a viral email experiment involving 31 183 individuals in which we were able to track a specific piece of information through the social network. We found that, contrary to traditional models, information travels at an unexpectedly slow pace. By using a branching model which accurately describes the experiment, we show that the large heterogeneity found in the response time is responsible for the slow dynamics of information at the collective level. Given the generality of our result, we discuss the important implications of this finding while modeling human dynamical collective phenomena.

DOI: 10.1103/PhysRevLett.103.038702 PACS numbers: 89.75.Hc, 05.10.−a

Modeling social dynamic phenomena as emerging from the interaction of individuals has recently attracted a lot of activity by statistical physicists [1]. Examples include epidemics spreading [2–4], cooperation, opinion formation, cultural dynamics, diffusion of innovations [5], etc. All of them are determined by the way humans spread or share information and thus depend on the rhythms and activity patterns of humans [6–9]. Despite its importance, detailed empirical data on how humans disseminate a specific piece of information are scarce or indirect [10–13]. Most understanding comes from epidemiological models run on empirical or synthetic social networks [2–4,14]. These models usually neglect human activity patterns, assuming that the response time τ_R, i.e., the time it takes for an individual to resend the information, is homogeneous or described by an exponential distribution which leads to a Poissonian description for human activity patterns. The main justification for this approximation is that it allows theoretical and computational descriptions through simpler Markovian (nonmemory) models. However, recent research shows that human activity is much more heterogeneous than considered in stark contradiction with the Poissonian approximation. For example, email activity [6–8], market trading frequencies, Web page visits [9], or activity in online social spaces [13,15] are all described by heavy-tailed or power-law distributions. We show how the large heterogeneity of human activity rhythms controls the information diffusion dynamics and question the validity of current models to describe it.

The same issue was considered in Ref. [16], where the authors investigate the effect of heavy-tailed distributions observed for τ_E, the time between consecutive emails (the interevent time), in email communication [6–8,17]. The authors proposed a relationship between τ_R and τ_E: The response time can be approximated (from below) by the time elapsed between receiving and sending emails τ_{RS}. If incoming emails arrive at random times, average τ_R is given by the solution of the waiting time paradox in renewal processes $\bar{\tau}_R \simeq \bar{\tau}_{\mathrm{RS}} = \bar{\tau}_E/2(1 + \sigma_E^2/\bar{\tau}_E^2)$ [18]. The database [17] used in Ref. [16] gives $\bar{\tau}_E \simeq 1$ day and $\sigma_E \simeq 3.3$ days, and then $\bar{\tau}_R \simeq 6$ days. Thus while the average interevent time is around one day, its heterogeneity makes the response time much bigger and information travels slower than expected. However, this is a poor approximation for τ_R since receiving an email may trigger action on it which correlates the reception and forwarding events. In fact, the same database [17] used in Ref. [16] contains data on incoming and outgoing emails from whence one can obtain that $\bar{\tau}_{\mathrm{RS}} \simeq 2.5$ hours, and then the approximation $\tau_R \simeq \tau_{\mathrm{RS}}$ used in Ref. [16] yields the opposite conclusion: Information should travel much faster than expected.

This example highlights an important shortcoming of currently available data: the inability to resolve the dynamics of a specific content item at the individuals' level (see, however, [10]). This forces one to make inferences about the dynamics from online communication data [12,13], or without knowledge of the nature of information transmitted [16,17], or from population averaged results [11]. In order to overcome this limitation, we conducted a large scale experiment to measure the influence of human activity on the diffusion of a specific piece of information. Subscribers to an online newsletter in 11 European countries were rewarded for recommending it. The offering email spread through viral propagation [19] tracked at every step. Web advertising spurred 7154 individuals to start recommendation cascades (Fig. 1) that, driven by 2111 *secondary spreaders*, grew until a total of 31 183 people received the message, 77% of them through recommendation emails. The propagation graph contains 7188 cascades of sizes between 2 and 146 nodes and diameters of up to 8 propagation steps [20].

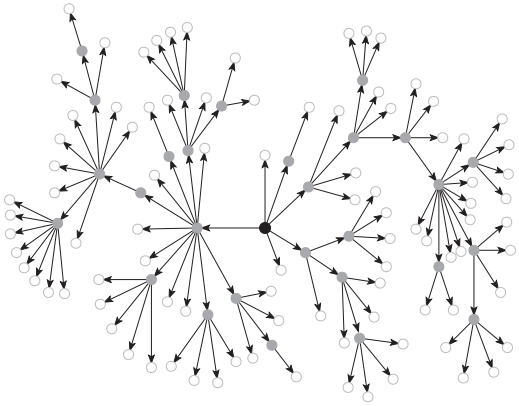

FIG. 1. Cascade with 122 nodes and 6 propagation steps found in the experiments. It starts out of a seed in the center (black) and grows by propagation through viral nodes (gray).

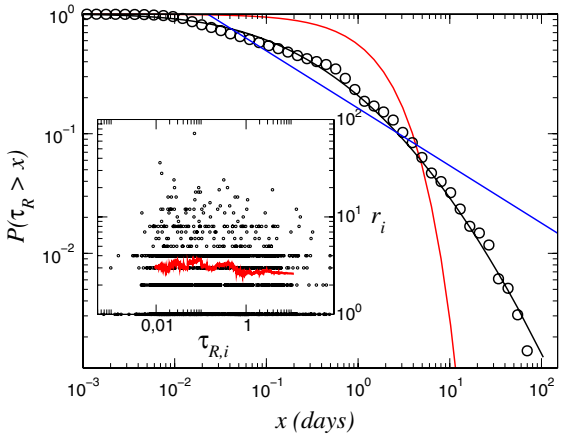

FIG. 2 (color online). Complementary cumulative distribution of the response time τ_R in the viral marketing campaigns (circles). The solid line shows fits to a log-normal distribution with $\hat{\mu} = 5.547$ and $\hat{\sigma}^2 = 4.519$ (black) and to an exponential distribution (red) and a power-law distribution (blue) with exponent -1.48. Inset: Scatter plot of the number of recommendations sent by each participant r_i vs her response time $\tau_{R,i}$ (dots). The red solid line is a running average of r as a function of τ_R.

Although message infection and propagation can be quite involved processes, population-level analyses describe viral propagation as a function of the basic reproductive number R_0, i.e., the average number of secondary cases generated by each informed individual [11,14]. If $R_0 > 1$, propagation reaches the *tipping point* where information reaches a significant fraction of the target population, but if $R_0 < 1$, propagation dies quickly. Secondary spreaders passed the message to $\bar{r} = 2.96$ individuals on average. Just a fraction $\lambda = 0.0879$ of those receiving it were *infected* by the viral process and forwarded the message again. Thus, the average of secondary cases per infected individual $R_0 = \lambda\bar{r} \simeq 0.26$ is below the tipping point. While R_0 is small, the large heterogeneity in the individual values of r and λ (see Fig. 2) led to a big variation in the cascade sizes found in our experiments [20]. Such heterogeneity was not enough to sustain the spreading, and all cascades stopped propagating after a finite number of recommendation steps. The campaign was a viral success [11], nevertheless, as the number of individuals virally reached was 4 times that of seeds.

A striking feature of the viral cascades found was the scarcity of loops, triangles, or closed paths (Fig. 1). Email redundancy (i.e., the fraction of emails sent to already informed individuals) was just 0.74% as cascades were mostly treelike shaped. This fact has also been found in recommendation cascades of online retailers [21] or information cascades in the blogosphere [13]. However, social networks are locally dense, and a large fraction of links connects members of communities or groups internally [22] anticipating a larger email redundancy. A possible explanation is that viral spreaders assume that the group from whence a message came knows it already and avoid that community, as suggested in Ref. [10] for chain letters. This self-avoiding feature may reduce the impact on information diffusion of the social network local structure [2] in favor of midrange to global topology properties.

In line with the studies mentioned in the introduction, our viral marketing campaigns show also high heterogeneity in the response time τ_R at the individual level: Participants forward the message after $\bar{\tau}_R = 1.5$ days on average, with a large standard deviation of $\sigma_{\tau_R} = 5.5$ days. Some participants resend the invitation email as much as $\tau_R = 69$ days after receiving it. Our data are fully consistent with a log-normal distribution for the distribution of response times $P(\tau_R)$. Power-law distributions or exponential distributions systematically over- and underestimate (respectively) the frequency of large response times (see Fig. 2). Moreover, response time does not show statistical correlation with the number of recommendations made by participants (see Fig. 2). Thus, the delay τ_R in forwarding a message and the number of recommendations sent r result from seemingly independent decisions.

At the collective level, we find an extraordinary behavior of information diffusion: If $i(t)$ is the average fraction (over all cascades) of informed individuals forwarding the message at time t, its dynamics follows an unexpectedly slow pace (see Fig. 3). This is in striking contrast with traditional epidemic models [14] where the dynamics of $i(t)$ in the cascades is modeled by the growth equation

$$\frac{di}{dt} = \alpha_0 i(t) \Rightarrow i(t) \sim i(0)e^{\alpha_0 t}, \tag{1}$$

where $\alpha_0 = (R_0 - 1)/\bar{\tau}_R$ is the naive approximation to the Malthusian rate parameter of the population. Equation (1) is the simplest version of more complicated models such as the Bass model of innovations diffusion [5] or the susceptible-infected-removed (SIR) epidemic model [14] used to model information propagation in social networks [1,3]. Equation (1) is based on the assumption that infection, or information diffusion, happens mostly around time $\tau_R \simeq \bar{\tau}_R$ and that new infections by individuals that have already infected others are very unlikely for $\tau_R \gg \bar{\tau}_R$.

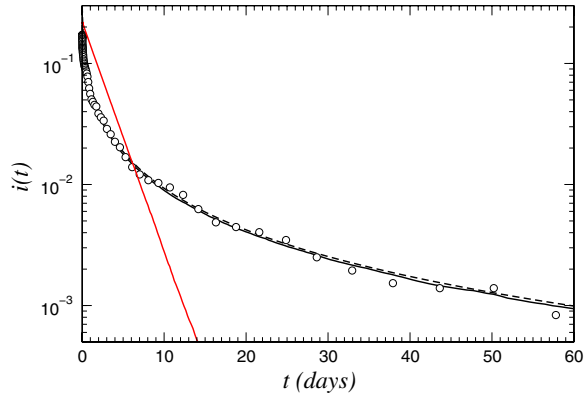

FIG. 3 (color online). Average fraction of new participants as a function of the cascade start time in our campaigns (circles) compared with the prediction of the Bellman-Harris model with $P(t)$ the log-normal distribution (black line) of Fig. 2 and with $P(t)$ exponential of the same mean (red). The dashed line is the asymptotic approximation (5) of the Bellman-Harris model with $P(t)$ log-normal. Inset: Time evolution of the cascade average size (circles) accurately predicted by the model for $G(t)$ log-normal. In red is the prediction for $G(t)$ exponential.

Since $\bar{\tau}_R \simeq 1.5$ days, Eq. (1) implies that most infections (new informed individuals) should happen during the first few days. However, we observe a significant fraction of new infections even at the month time scale as shown in Fig. 3. Moreover, the functional decay of new infected individuals cannot be explained by an exponential decay as (1) predicts. Thus, traditional epidemic models fail to predict information speed and also the functional form of its dynamics.

To explain our results, we model them with a branching process that considers activity heterogeneity. Given the low email redundancy, we consider only the growth of treelike cascades. Nevertheless, this approximation captures the main features of the spreading dynamics on social networks [2,22]. Each cascade starts from a seed that initiates propagation with a random number of recommendations whose average is $\bar{r}$. Touched individuals become secondary spreaders with probability λ and propagate the message further. Information forwarding happens after time τ, independent of r and distributed by $P(\tau)$. This process is the well known Bellman-Harris branching model [23] where the average fraction (over all cascades) of active individuals at time t is given by

$$i(t) = 1 - G(t) + R_0 \int_0^t i(t-\tau)P(\tau)d\tau, \qquad (2)$$

where $G(t) = \int_0^t P(\tau)d\tau$ is the cumulative distribution function (CDF) of $P(\tau)$. Equation (2) is non-Markovian, since the average number of new infections $i(t)$ at time t depends on the history of infections in the past $0 < \tau < t$. Explicit solutions of (2) do not exist for general $P(\tau)$ and R_0, but if there is a solution α of the implicit equation

$$R_0 \int_0^\infty e^{-\alpha\tau}P(\tau)d\tau = 1, \qquad (3)$$

then the asymptotic behavior of (2) is given by [23]

$$i(t) \sim Ce^{\alpha t}, \qquad C = \frac{R_0 - 1}{\alpha R_0^2 \int_0^\infty \tau e^{-\alpha\tau}P(\tau)d\tau}. \qquad (4)$$

Thus, although Eq. (2) is non-Markovian, it behaves asymptotically as the solutions of the simple Markovian model (1) with α given by the solution of (3). This approximation for general $P(t)$ is exact in the case of the exponential distribution of memoryless Poissonian statistics: If $P(\tau) = e^{-\tau/\bar{\tau}_R}/\bar{\tau}_R$, the solution of (3) is $\alpha = \alpha_0$ and Eq. (2) can be written in differential form as Eq. (1).

However, for $R_0 < 1$ (i.e., $\alpha < 0$), Eq. (3) has a solution for α only if $P(t)$ decays fast enough, specifically, faster than the exponential distribution in the limit $t \to \infty$. Thus, growth models like (1) or approximations like (4) are not valid for a large family of distributions $P(t)$ known as *subexponential distributions*, i.e., those decaying slower than exponential when $t \to \infty$. This family includes important cases like the log-normal, power-law, or stretched exponential distributions. In that case, the general asymptotic behavior of Eq. (2) is controlled instead by the tail of the CDF distribution [24]

$$i(t) \sim \frac{1}{1 - R_0}[1 - G(t)], \qquad (5)$$

which highlights the non-Markovian character of the solutions of Eq. (2), since they depend on those individuals whose response time is the longest. The distributions used to model the large heterogeneity of human response times (power-law [7] or log-normal [8]) are members of this class of distributions, and Eq. (5) shows the profound impact of large heterogeneity in response times: The very functional form of the time dependence is changed, and the dynamics of information does not depend on the mean value of the response time but on the tail of the distribution, thus drastically slowing down the propagation of information. Figure 3 shows the striking agreement of the approximation (5) with the data obtained in our campaigns assuming that $P(\tau)$ is given by the log-normal distribution in Fig. 2.

The slowing down of information diffusion due to the subexponential nature of human response times can explain the prevalence of some rumors, viral campaigns, chain letters, or computer viruses as suggested in Ref. [16]. For example, if we assume N_s seeds are initially infected and set the end of diffusion when the fraction of infected individuals decays to $i(t_f) \sim 1/N_s$, then the Poissonian approximation (1) gives $t_f \simeq \alpha_0^{-1} \ln N_s$, while in the log-normal case [Eq. (2)] we get $t_f \sim e^{\sqrt{b \ln N_s}}$, where b is independent of R_0. For large enough N_s, there is a huge difference between both estimations. For example, if $N_s = 10^4$ individuals (a large but moderate value), $t_f = 17$ days (with $R_0 = 0.26$) for Poissonian models while $t_f \simeq 1$ year if $P(\tau)$ is log-normal.

Interestingly, the large heterogeneity found in human response time has the opposite effect above the epidemic threshold ($R_0 > 1$) where Eq. (3) has a solution α much

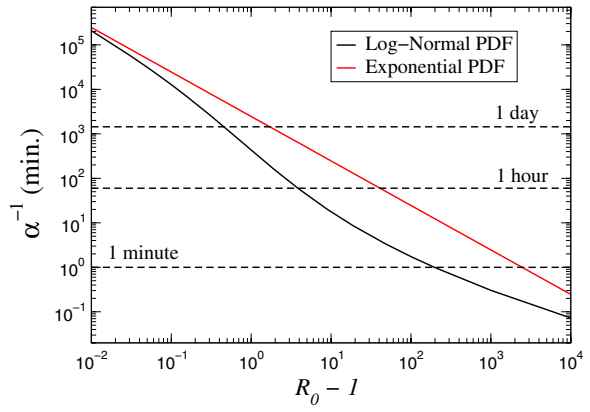

FIG. 4 (color online). Value of the inverse of the Malthusian parameter α (in minutes) as a function of R_0 for $R_0 \geq 1$. The black line below is the result given by Eq. (3), where $G(t)$ is the fitted log-normal distribution and the straight line in red corresponds to $G(t)$ an exponential distribution with the same $\bar{\tau}_R =$ 1.5 days.

bigger than α_0 (Fig. 4). In this case the Bellman-Harris model describes the important initial exponential growth of the epidemic spreading, and we see that information spread faster than expected with α being even 1 order of magnitude greater than α_0. For example, although the average response time is of the order of days in typical email exchange, if successful, spreading of information through email will occur in a matter of hours. The change of behavior above and below the tipping point stems from the different impact on the dynamics of individuals with small or large values of τ_R: While for $R_0 < 1$ the number of infected individuals decays in time up to a point where a sole individual can halt or resume a viral cascade growth, for $R_0 > 1$ the dynamics is governed by individuals with a small value of τ_R, more abundant than those with $\tau_R \simeq \bar{\tau}_R$, which speeds up diffusion.

In summary, we have shown that the large heterogeneity in human activity controls information spreading. Its impact is not merely quantitative (a mere renormalization of information speed); rather, it changes qualitatively the dynamics of information diffusion at the collective level. Specifically, below the tipping point, the very concept of information speed becomes ill defined as information progresses in logarithmic time. This effect is universal because it does not depend on the specific details of human activity patterns but on the subexponential character of their distribution. Since most information transmission and sharing in social networks has limited reach, thus occurring below the tipping point, our findings are bound to affect the way we understand and model social phenomena like rumor spreading, cooperation, opinion formation, cultural dynamics, diffusion of innovations, etc. Actually, we have shown that the most common and simple equation for epidemic dynamics [the growth equation (1)] cannot be used to model the information diffusion given the relevance of heterogeneity in this dynamical process. Since this equation is usually the first building block in more complicated and popular models (e.g., the Bass model [5], the SIR/susceptible-infected-susceptible models [2,3,14], etc.), we expect those models to suffer from the same problems found here. Finally, since non-Poissonian (subexponential) distributions also characterize the individual rhythms and activity patterns in other human actions [7–9,13,15], we expect a similar influence of heterogeneity on the corresponding dynamical collective behaviors. We hope our work will trigger further research on the impact of that heterogeneity in the way we model and understand human dynamics.

We thank R. Cuerno for discussions, IBM for the access to anonymized data of its viral marketing campaigns, and Dr. J.-P. Eckmann for sharing the database of [17]. E. M. acknowledges partial support from MEC (Spain) through grants Ingenio-MATHEMATICA and MOSAICO and Comunidad de Madrid through grant SIMUMAT-CM.

[1] C. Castellano *et al.*, Rev. Mod. Phys. **81**, 591 (2009).
[2] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[3] Y. Moreno *et al.*, Phys. Rev. E **69**, 066130 (2004).
[4] D. H. Zanette, Phys. Rev. E **64**, 050901(R) (2001).
[5] E. Rogers, *Diffusion of Innovations* (Free Press, New York, 1995).
[6] R. D. Malmgrem *et al.*, Proc. Natl. Acad. Sci. U.S.A. **105**, 18 153 (2008).
[7] A.-L. Barabási, Nature (London) **435**, 207 (2005).
[8] D. B. Stouffer *et al.*, arXiv:physics/0510216.
[9] A. Vázquez *et al.*, Phys. Rev. E **73**, 036127 (2006).
[10] D. Liben-Nowell and J. Kleinberg, Proc. Natl. Acad. Sci. U.S.A. **105**, 4633 (2008).
[11] D. J. Watts and J. Peretti, HBR Online, F0705A (2007).
[12] D. Gruhl *et al.*, in *Proceedings of the 13th International Conference on WWW* (ACM, New York, 2004).
[13] J. Leskovec *et al.*, in *Proceedings of the SIAM International Conference on Data Mining (SDM07)* (SIAM, Philadelphia, 2007).
[14] R. M. Anderson and R. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, New York, 1991).
[15] A. Kaltenbrunner *et al.*, in *Proceedings of the BIS 2007 Workshop on Social Aspects of the Web* (CEUR-WS.org, 2007).
[16] A. Vázquez *et al.*, Phys. Rev. Lett. **98**, 158702 (2007).
[17] J.-P. Eckmann, E. Moses, and D. Sergi, Proc. Natl. Acad. Sci. U.S.A. **101**, 14 333 (2004).
[18] L. Breuer and D. Baum, *An Introduction to Queueing Theory* (Springer, New York, 2005).
[19] S. Jurvetson and R. Draper, Viral Marketing, Netscape M-Files, 1997.
[20] J. L. Iribarren and E. Moro (to be published).
[21] J. Leskovec *et al.*, ACM Trans. Web **1**, 5 (2007).
[22] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[23] T. E. Harris, *The Theory of Branching Processes* (Springer-Verlag, Berlin, 2002).
[24] K. Athreya and P. Ney, *Branching Processes* (Springer-Verlag, Berlin 1972).