Role of persistent cascades in diffusion

Steven Morse
Operations Research Center, Massachusetts Institute of Technology and The Charles Stark Draper Laboratory, Cambridge, Massachusetts 02139, USA

Marta C. González
College of Environmental Design, University of California, Berkeley and Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

Natasha Markuzon[*]
The Charles Stark Draper Laboratory, Cambridge, Massachusetts 02139, USA

(Received 30 May 2018; published 28 January 2019)

We define a structural property of real-world large-scale communication networks consisting of the recurring patterns of communication among individuals, which we term persistent cascades. Using methods of inexact tree matching and agglomerative clustering, we group these patterns into classes which we claim represent some underlying way in which individuals tend to disseminate information. We extend methods from epidemic modeling to offer a way to analytically model this recurring structure in a random network, and comparing to the data, we find that the real cascading structure is significantly larger and more recurrent than the random model. We find that the cascades reveal a habitual hierarchy of spreading, alternative roles in weekday vs weekend spreading, and the existence of hidden spreaders. Finally, we show that cascade membership increases the likelihood of receiving information spreading through the network through simulation on the real order of communication events.

DOI: 10.1103/PhysRevE.99.012323

I. INTRODUCTION

The increasing availability of large-scale communication data allows us to study the dynamics of human information spreading patterns at an unprecedented depth. As an example, consider cell-phone data, which provides long-term, second-by-second, unfiltered communication records of individuals among their social contacts (and we note are more precisely termed metadata, in that the records contain no information about the content of each event).

Consider the following two essential questions in the study of human communication dynamics: (i) Do individuals (or groups of individuals) exhibit patterns of communication? (ii) If so, what are the effects of these patterns on information spread? To answer these questions with the breadth of data that resources like cell-phone records provide, researchers often of necessity use simplifying assumptions. For example, one may aggregate the available interpersonal events over some time window and investigate the resulting static network. Many studies of diffusion processes, or measures of centrality (i.e., the identification of individuals' roles, or importance, within the populations), take place on such static networks (see, e.g., [1–3]). More recently, however, the importance of incorporating temporal knowledge has gained attention (see, e.g., [4–10]), which is the approach we take here.

Specifically, we approach the first question, of identifying patterns of communication, by searching for recurring temporal patterns of cascading communication that indicate persistent group conversations engaged in information spread. We define a structural property of large-scale communication networks consisting of the recurring patterns of communication among individuals, which we term persistent cascades (reported in [11]). This approach takes advantage of a recurring observation in the study of human communication dynamics that events are bursty, or temporally clustered [12,13], and moreover that the mechanism for this phenomenon is information spread and its inherently rapid, cascading structure [14–19].

We approach the second question, of measuring the effect of these patterns on diffusion dynamics in the temporal network, by simulating the spread of information over the real order of communication events and comparing the susceptibility of the individuals engaged in persistent communication activity against those not. We demonstrate that the diffusion accelerates under certain conditions and that the persistent cascades appear to be the key mechanism in this regime.

A. Background

1. Identifying information spread

We first review previous work in identifying meaningful structure in large-scale metadata and then turn our attention to how this affects our understanding of diffusion dynamics.

[*] nmarkuzon@gmail.com

One approach to the problem of filtering out meaningful communication events from large-scale data would be to simply threshold the number of observed events. For example, require the call event $a \to b$ to occur at least k times in some time window or require the call to be reciprocated, i.e., to also observe $a \leftarrow b$ [3,20,21]. This certainly achieves some denoising of the data (e.g., it reassures us that the call was not accidental), but gives us little idea of the purpose or information-carrying potential of the event.

Some studies use exogenous anomalous events to filter for meaningful or information-carrying events. For example, it seems safe to assume the calls immediately following an earthquake, robbery [18], rocket attack [17], or major sporting event [17] will tend to carry more weight than calls selected from an arbitrary period of time. Indeed, these studies find the heavy-tailed interevent times and cascading patterns associated with information spread [13]. Yet this approach restricts our analysis to a very small portion of the available data and moreover restricts the roles and patterns we might observe.

Another approach is to interpret the problem probabilistically and ask what the conditional probability is that b calls c given that a first called b. Often in this context we model networked communication events as realizations of some underlying multidimensional stochastic point process [22–24]. This statistical learning approach introduces a large parameter space to reckon with, which we may tame by constraining our patterns of interest to have cascading structure, as in [25], or by introducing regularization through priors [22] or explicit constraints on the network structure [23].

We may instead consider searching for temporally recurring patterns (i.e., time-ordered patterns of calls) [26–28], which is the approach we take in this paper. This extends the earlier threshold idea to persistent temporal structures. For example, in [29] the authors study social structure in large-scale data by looking for persistent membership in certain core groups and activities. In [8,9,30] the authors identify temporal motifs that occur more frequently in the data than expected under a null model by looking for isomorphic patterns of communication over time. This focus on motifs gives us an abstract picture of meaningful structure at a population level, although in these studies is not carried back to analysis of individuals.

2. *Effect on diffusion*

Once the information cascades are identified, we may turn to the question of their effect on, and role in, diffusion dynamics [14–16,31]. Previous research has demonstrated a surprising slowing effect [7], due to the long periods of inactivity between bursts (or cascades). Compare this with classical models where we assume complete mixing of the population, which is closer to random activity. However, it is not immediately clear why the bursts themselves would not also accelerate the spreading enough to offset the long tails of inactivity.

An explanation for this seeming paradox, proposed in [32], is that it depends on the probability of information passing (similar to the infection rate of a disease). We provide a summary of their argument here, since it gives mathematical motivation to our results in the final section.

We will first borrow terminology from epidemic spread to say that for a pair of individuals i and j connected via an edge (ij), i infects j if j receives information from i. Let the probability of this infection event occurring be the transmissibility $\mathcal{T}_{ij}$, and we can then relate the problem of infection (i.e., information) passing to a bond percolation model [33].

Let i become infected at time t_α and let $n_{ij}(t_\alpha)$ be the number of contacts i has with j after this event before i is no longer transmitting the information (i.e., recovers). We denote this interval by $[t_\alpha, t_\alpha + \tau]$ and note that the transmissibility in this interval, for some infection probability λ, is $1 - (1-\lambda)^{n_{ij}(t_\alpha)}$.

Now if the $i \to j$ interactions do not depend on i's initial time of infection t_α, then we can apply the total probability to calculate transmissibility. For example, in the case where the $i \to j$ interactions come from a memoryless Poisson process over the time interval $[0, T_0]$ with rate r_{ij}, this leads to the standard result

$$\begin{aligned} \mathcal{T}_{ij} &= \sum_{n=0}^{\infty} \text{Prob}(n_{ij} = n)[1 - (1-\lambda)^n] \\ &= 1 - e^{-\lambda r_{ij}}, \end{aligned} \tag{1}$$

with, for example, rate $r_{ij} = w_{ij}\tau / T_0$, where w_{ij} is the total number of interactions between i and j over $[0, T_0]$ [33].

However, let us posit that the $i \to j$ events are highly dependent on i's initial infection at t_α, since i will tend to pass on information soon after it is received. So we assume only that the $i \to j$ infection events themselves are independent and average over them:

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

When $\lambda \leqslant 1$, $1 - (1-\lambda)^n \approx \lambda n$, and when $\lambda \approx 1$, $1 - (1-\lambda)^n \approx 1$, for $n > 0$. So substituting these approximations into Eq. (2) [32] defines two regimes of transmissibility depending on the infectivity λ,

$$\mathcal{T}_{ij} = \begin{cases} \lambda \langle n_{ij} \rangle_{t_\alpha}, & \lambda \leqslant 1 \\ 1 - P^0_{ij}, & \lambda \approx 1, \end{cases} \tag{3}$$

where P^0_{ij} is the probability of *zero* contacts with j after the event t_α.

Therefore, in group conversations, $\langle n_{ij} \rangle_{t_\alpha}$ is higher than random calls and so for low λ explains the increased spreading. By contrast, the long periods of inactivity observed in group conservations lead to higher P^0_{ij} than in random calls and so for high λ explains the decreased spreading.

The authors postulate in [32] there are information cascades doing the heavy lifting of spreading information under this low-λ regime and whose effects are only masked under large infectivity λ. We would like to actually identify the conversations displaying these cascading properties, measure their role in information spread, and test this claim.

B. Contributions

In this work we first identify information spreading patterns in large-scale communication metadata by extracting recurrent cascading patterns of communication, which we

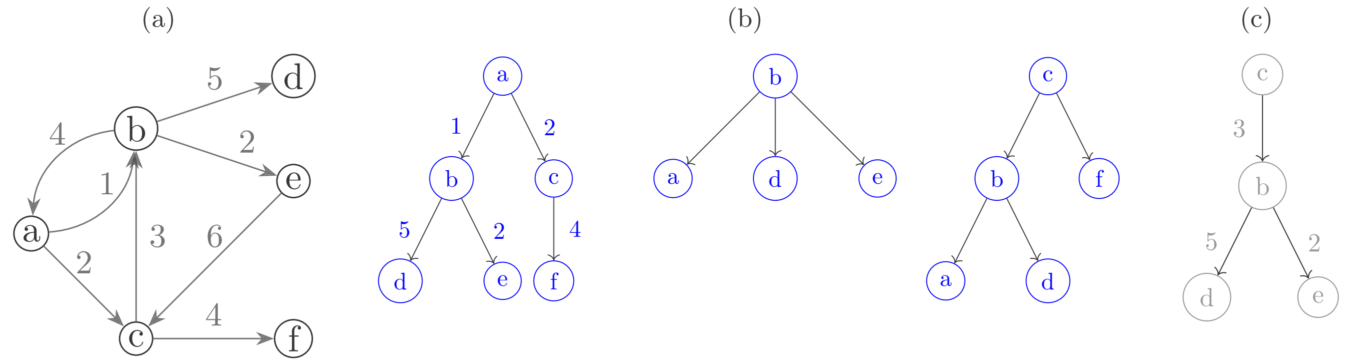


FIG. 1. Simplified illustration of cascade extraction from a temporal graph. (a) Full temporal information ($\Delta t = 6$ units, times depicted on edges). (b) Three valid cascades given this temporal snapshot. Note that there is no time ordering of children within a cascade. (c) Invalid cascade because c-b-e is not a time-connected path and it is missing the edge c-f.

term persistent cascades, using a methodology which is not typically used in a combination of inexact tree matching and agglomerative clustering. We show that the resulting patterns have long-term persistence (multiple months to a year), exhibit a habitual hierarchy of individual roles in spreading, reveal new roles, such as exclusivity to weekends or weekdays, and reveal hidden spreaders not evident in the static network. We also demonstrate that the persistent cascades we identify under this methodology are significantly larger and more persistent than what we would observe under a random network with similar structure and rates of interaction, using an extension of epidemic modeling. Finally, we show that participation in persistent cascades increases the individual probability of receiving information spreading through the network (even after controlling for overall activity) when the probability of information passing is low.

II. METHODS

A. Identifying persistent cascading patterns

1. Extracting cascades

We now propose a methodology for understanding information spreading patterns in large-scale communication metadata, which we will validate using cell-phone records. We make an assumption that a person receives information upon first exposure, or in other words, at the earliest possible time. This implies that information passing exhibits a treelike network structure where there is a single in-edge to each person, spanning some interval of time Δt (e.g., a few hours or a week). Formally, this assumption leads to the construction of a rooted, directed, Δt-connected tree which we term a cascade, following previous work [17–19].

Denote a cascade with root r by C^r, denote the set of all cascades for root r with maximum time interval Δt and total time period T by $\mathcal{C}^r(T, \Delta t)$, and use subscripts as necessary to distinguish multiple cascades with the same root. For example, we might have the set of all cascades for some root a:

$$\mathcal{C}^a(T = 1 \text{ month},\ \Delta t = 24\text{ h}) = \{C_1^a, C_2^a, C_3^a\}. \tag{4}$$

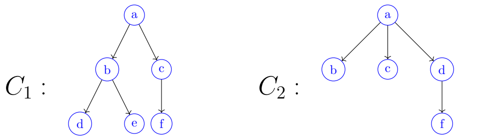
Note that we require that the intervals not overlap, i.e., no calls from C_1^a can also be in C_2^a, etc. An example of cascade construction from a network with all temporal information is shown in Fig. 1.

2. Measuring persistence

One objective of our study is to identify persistent cascading activity over time. Similar cascading patterns over long periods of time are more indicative of meaningful communication and more particularly information spread [8,25]. However, we wish to relax any requirement that the patterns be identical (or isomorphic) due to the inherently inconsistent nature of human activity.

Therefore, to search for persistent patterns, we employ inexact tree matching measures of the form $s_* : C \times C \rightarrow [0, 1]$, where C is a cascade. We consider two metrics: tree edit distance, which gives a sense of structural similarity, and reach set similarity, which gives a sense of membership similarity (i.e., how similar the members of the cascade are).

Tree edit distance. Edit distance is the process of counting the minimum number of insertions, deletions, or mutations required to transform one string into another. One can extend this concept to trees. Denote the tree edit distance between two trees (or cascades) C_1 and C_2 by $\delta(C_1, C_2)$, which maps two cascades to a non-negative integer. As an example, consider the following two trees:



(5)

To change C_1 into C_2, we can delete d and e, mutate c into d, and add c again, giving $\delta(C_1, C_2) = 4$ (note that this is the same to change C_2 into C_1).

A canonical algorithm for computing this distance is due to Zhang and Shasha [34], which we implement with [35]. We can now define a similarity measure using this distance as follows.

Definition 1 (tree edit distance similarity). Define the normalized tree edit similarity as

$$s_\delta(C_1, C_2) \overset{\text{def}}{=} 1 - \frac{2\delta(C_1, C_2)}{|C_1| + |C_2| + \delta(C_1, C_2)} \tag{6}$$

and note that s_δ lies on [0,1].

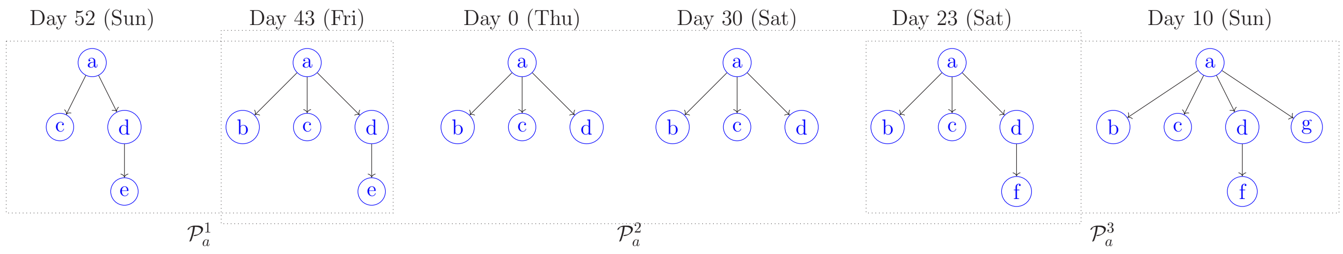


FIG. 2. Actual set of cascades for a root a over a 60-day period. Six persistent cascades are shown, each from temporal subgraphs with $\Delta t = 24$ h. Dotted rectangles depict the persistence class groupings. We see a clear set of core friends (nodes b, c, and d) and slight variations incorporating other groups. We also see the overlap that occurs when a cascade appears to fit in multiple classes. Labeled above each cascade is the day of the week.

This definition is due to Li and Zhang [36], who also prove that the corresponding distance metric $1 - s_\delta$ meets the triangle inequality. Note that we make every edit operation unit cost. Using the example trees above in (5), we now compute $s_\delta = 1 - \frac{2\times 4}{6+5+4} = \frac{7}{15} \approx 0.47$.

Reach set. Consider the unordered set of all nodes in a tree. For a cascade, this corresponds to all users who potentially received some information during the time period Δt. We term this the reach set of a cascade (similar to concepts in [37]).

A simple first approximation of the similarity of two cascades is by comparing their reach sets. Let $R(C_i)$ denote the reach set of a cascade C_i. Now, given two cascades C_1 and C_2, we define the similarity measure s_ρ as the Jaccard index of the two reach sets as follows.

Definition 2 (Reach set similarity). Given two cascades C_1 and C_2 and their reach sets $R(C_1)$ and $R(C_2)$, define

$$s_\rho(C_1, C_2) \overset{\text{def}}{=} \frac{|R(C_1) \cap R(C_2)|}{|R(C_1) \cup R(C_2)|} \tag{7}$$

and note that s_ρ lies on [0,1].

Continuing with the previous example in (5), we have $s_\rho(C_1, C_2) = \frac{5}{6} \approx 0.83$.

3. *Finding persistence*

We may now cluster similar cascades originating from a given root r into a group we term a persistence class, denoted by $\mathcal{P}_r$, such that each cascade in the class is at least ℓ-similar to each other. This group now represents various incarnations of some underlying, implicit communication structure.

Definition 3 (persistence class). Define the ith persistence class of root r, similarity threshold ℓ in time period T over intervals Δt, as the set

$$\mathcal{P}_i^r(\ell, T, \Delta t) \overset{\text{def}}{=} \left\{C_1^r, C_2^r \in \mathcal{C}^r(T, \Delta t) : s_*\left(C_1^r, C_2^r\right) \geqslant \ell\right\}. \tag{8}$$

Definition 4 (persistent cascade). Define a persistent cascade as any cascade C_i^r such that $C_i^r \in \mathcal{P}_i^r(\cdot)$, for at least one i.

Note we may also choose to ignore any persistence classes below a certain size. The minimum size is 2 by construction, but we may decide based on the parameters T and Δt that a minimum size of 3 or more is appropriate.

To find these classes, given our definition and Eq. (8), we take a clustering approach which allows for overlapping clusters. Specifically, represent each data point (cascade) as a vertex in a graph $H(\ell)$ such that each any two vertices u, v with similarity $s_*(u, v) \geqslant \ell$ are connected. (Note a single vertex in this construction represents an entire cascade and the similarity between two vertices is one of the previously defined similarity measures.) Then the persistence classes are the maximal completely connected subgraphs in H, also known as the maximal cliques.

Let the reader note this is closely related to an agglomerative clustering approach with complete linkage. That is, define the similarity between two clusters U and V as $s(U, V) = \min s_*(u_i, v_j) \forall i \in U,\ j \in V$, where u_i, v_j represent cascades within U and V. Then the clusters at iteration k, such that every pairwise similarity within the cluster is greater than or equal to s_k, represent persistence classes with $\ell = s_k$.

The agglomerative clustering approach, however, assumes that each cascade falls uniquely into one class, which we can imagine is not always true: A spreading pattern among work friends may overlap with the pattern among social friends and there may be cascades that are not clearly in one class or the other. By taking the graph-theoretic maximal clique approach, we avoid this limitation.

Figure 2 gives an example of the result of this clustering methodology on several cascades in the data for a particular root a. We see a core pattern consisting of root a calling b, c, and d captured in $\mathcal{P}_a^2$. Then we see two variations on this core structure: $\mathcal{P}_a^1$, which incorporates e, and $\mathcal{P}_a^3$, which incorporates f and g. Since they are mostly weekend calls, we might easily imagine this being a core group of social friends, with variations possibly for family or work acquaintances. Note the clusters overlap.

B. Random network model

To quantify the significance of the observed patterns in the data (i.e., the patterns' nonrandomness), we compare the distribution of their occurrence against a null model. Specifically, given a random network with a degree distribution matching the real network, and with average interindividual call event rates also matching the real data, but without any of the temporal clustering or mutually influencing effects that we hypothesize are present in the data, what is the probability of a cascade of s users occurring n times in a month? That is, how likely is it to observe persistence classes of size $|\mathcal{P}^r| = n$ such that $C^r \in \mathcal{P}^r$ have $|C^r| = s$?

To this end we will adapt techniques from the rich fields of percolation theory and epidemic spreading [33]. We are concerned with the dynamics of some contagion through a population with network structure. In our application, the contagion is information, the initial infected population are the cascade roots, and the outbreak is the cascade itself. Further, we assume that the probability of infection is only dependent on the rate of interaction between individuals.

We first outline a straightforward way to generate this null model through simulation and then extend the methods in [33,38] to precisely describe the probability of a particular outbreak (i.e., cascade) and subsequently the probability of its recurrence (i.e., persistence).

1. Simulation model

We need to generate both the network structure and the interpersonal call activity. This will output a sequence of events, or time series, to which we can then apply the temporal graph-mining algorithm outlined before to find recurring similar patterns (persistent cascades).

Network structure. We choose a configuration model with a degree distribution p_k that matches the observed $\hat{p}_k$ in the data. [39] This will give us a closer representation of network structure than, say, an Erdős-Renyí model, which can only match the expected degree. We do not consider other random network models and leave this to future work.

Edge interactions. We will assume that each pair (ij) can be modeled by a Poisson process with rate parameter r_{ij}. We estimate this rate for every edge as w_{ij}/T_0, where again w_{ij} is the total calls between (ij) and T_0 is the total time period observed (e.g., 2 months). This gives us a real distribution of rates $\hat{P}(r)$, which we approximate as $P(r)$ with a Γ distribution $P(r) \propto r^{\alpha-1}e^{-\beta r}$. We choose the Γ distribution because it is non-negative, conjugate with the exponential family, and qualitatively provides a good form for $\hat{P}(r)$ and leave consideration of other model distributions for future work.

2. Analytical model

We can instead represent this entire framework analytically. We will make the same assumptions as in the preceding section (i.e., network structure and average rates of interaction fit to the data, but all events independent and identically distributed). At a high level, we will (i) derive a probability distribution of a cascade (outbreak) of size s happening after n steps, denoted by $P_s^{(n)}$, (ii) use this to upper bound the probability of a particular cascade occurring among a particular set of users, and (iii) use this in a binomial distribution to describe the probability of this cascade occurring multiple times (i.e., persistence).

Deriving a stepwise distribution of cascade size. Denote the probability distribution of a cascade of size s after n steps as $P_s^{(n)}$. To derive $P_s^{(n)}$ we follow techniques in Refs. [33,38], which we very briefly summarize here.

We can first derive a recursive expression for the probability generating function (PGF) of our desired distribution $P_s^{(n)}$, denoted by $H^{(n)}(x)$, in terms of the PGF for the excess degree of our degree distribution p_k,

$$H^{(n)} = H^{(n-1)}(xG_1(x)). \tag{9}$$

We can then extract $P_s^{(n)}$ by taking the appropriate derivative of H. However, we need to resort to numerical differentiation here, and in practice, the recursive definition of H and inherent small values leads to machine precision errors beyond the first ten or so values of s. Instead, Newman [33] recommends applying the Cauchy integral formula to instead derive

$$P_s^{(n)} = \frac{1}{s!}\frac{d^s H}{dx^s} = \frac{1}{2\pi i}\oint_\gamma \frac{H^{(n)}(z)}{z^{s+1}}dz, \tag{10}$$

with γ the unit circle (in the complex plane) $|z| = 1$.

Following Marder [38], we can evaluate this integral at some large number of points M around the unit circle m/M for $m = 0, 1, \dots, M-1$ and apply an inverse discrete Fourier transform, that is,

$$P_s^{(n)} = \frac{1}{M}\sum_{m=0}^{M-1} e^{-2\pi i s m/M} H_m^{(n)} = \frac{1}{M}\mathcal{F}_{\text{DFT}}(H, -1)[s], \tag{11}$$

where $H_m^{(n)} = H(e^{2\pi i m/M})$ and the notation $[s]$ denotes retrieving the sth element from the returned spectra of the transform.

Transmissibility. We assumed the probability of information transmission (or infection) was 1 at each step in the previous derivation, which we do not want to assume. Instead, for an infected individual i interacting with a susceptible contact j, the probability of infection should be governed by the average rate of contact (i.e., the average rate of call activity) over some interval $[0, T_0]$, which we define $r_{ij} = w_{ij}/T_0$, where w_{ij} is the total observed calls on the edge (ij) and varies from pair to pair. (Again, for purposes of our application, this has only to do with the average rate of interaction and nothing to do with a notion of the infectivity of the information or disease itself.)

Denote the probability of transmission (transmissibility) from i to j as $\mathcal{T}_{ij}$. The probability there is *not* infection is then

$$1 - \mathcal{T}_{ij} = \lim_{\delta t \to 0}(1 - r_{ij}\delta t)^{\tau/\delta t} = e^{-r_{ij}\tau} \tag{12}$$

for recovery period τ and therefore $\mathcal{T}_{ij} = 1 - e^{-r_{ij}\tau}$ [compare with Eq. (1)].

However, since we are assuming r_{ij} is independent and identically distributed for each pair in the network, then on a population level it is sufficient [33] to consider the average transmissibility $\mathcal{T} = \langle \mathcal{T}_{ij} \rangle$, which we can recover by averaging over all possible values of r. For $P(r) \sim \Gamma(\alpha, \beta)$ this gives

$$\begin{aligned}\mathcal{T} = \langle \mathcal{T}_{ij} \rangle &= 1 - \int_0^\infty e^{-r\tau}P(r)dr \\ &= 1 - \frac{\beta^\alpha}{(\beta+\tau)^\alpha}\int_0^\infty \frac{(\beta+\tau)^\alpha}{\Gamma(\alpha)} r^{\alpha-1}e^{-r(\beta+\tau)}dr \\ &= 1 - \frac{\beta^\alpha}{(\beta+\tau)^\alpha}. \end{aligned} \tag{13}$$

This shows that when τ gets larger (i.e., our period of interest gets longer), the transmissibility goes toward one, and as the rates skew smaller with larger β, the transmissibility

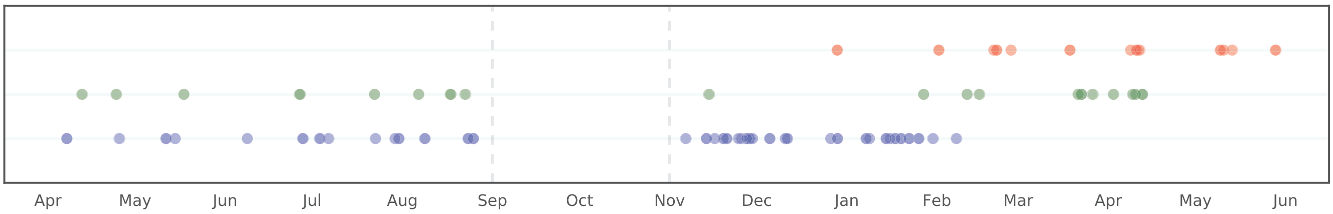


FIG. 3. Long-term persistent cascade activity, for three sample persistence classes. Each line of dots represents a single persistence class and each dot represents an entire cascade. We note temporal clustering and long-term persistence. (Dashed lines indicate a period of no data.)

goes toward zero. We can then simply express the generating function for the degree and excess degree distributions now as

$$G_0(x) = G_0(1 + \mathcal{T}(x - 1)), \tag{14}$$

$$G_1(x) = G_1(1 + \mathcal{T}(x - 1)) \tag{15}$$

to capture this effect [33].

Persistence. Finally, to capture the probability that a particular outbreak (i.e., cascade) happened between the same set of users multiple times (i.e., was persistent), we can take advantage of the fact that, given the cascade center and size, any set of users is equally likely. We will also discard any notion of approximate similarity and only consider outbreaks of exactly equal size.

The probability of a specific set χ of users being in an outbreak rooted at r, for any particular seed or root node r, is

$$q_\chi^{(n)} = \frac{1}{\text{no. of ways to make } \chi} P_{|\chi|}^{(n)} \tag{16}$$

and so we note that $q_\chi^{(n)} \leqslant P_{|\chi|}^{(n)}$. Qualitatively speaking, this upper bound is reasonably tight since most of the population has only two to four persistent contacts and we consider only $|\chi| = 3$ or 4.

Denote the probability of a particular pattern χ happening k times over the course of D disjoint periods (for example, $D = 60$ days) by Q_χ^D; we can approximate it with a binomial distribution with parameters D and $P_{|\chi|}^{(n)}$,

$$\hat{Q}_\chi^D \sim \text{Binom}\big(D, P_{|\chi|}^{(n)}\big). \tag{17}$$

We may now compare the distribution of $\hat{Q}_\chi^D$ to the observed distribution of the size of persistence classes with the same size cascades, that is,

$$\big\{\big|\mathcal{P}_i^r(1, D, \Delta t = 1)\big| \forall r, \forall i : |C^r| = |\chi|\big\}. \tag{18}$$

III. RESULTS

A. Data

We perform the analysis on call detail records from two midsize European cities and their greater metropolitan areas (cities A and B), over a 15-month period. Both cities contain residential, commercial, and industrial areas. The data consist of caller, callee, and time stamp for each phone call or SMS event recorded by the carrier. (Location information is also recorded, but not used in this study.)

In city A there are approximately 648 100 unique individuals generating a total of over 82.3×10^6 calls in the period observed. In city B there are approximately 523 500 unique individuals generating a total of over 55.7×10^6 calls in the period observed. Unless stated otherwise in the following analysis, we perform no preliminary filtering on this data (e.g., requiring reciprocated calls).

Three examples of persistent cascade classes in the data are shown in Fig. 3, where each line of dots represents a single persistence class and each dot represents an entire cascade. We note temporal clustering and long-term persistence.

B. Properties of persistent cascades

1. Habitual hierarchy

We may expect that the two similarity measures (one comparing structure, the other membership) would give us entirely different groupings of cascades. In fact, the opposite appears to be true. We compared the similarity of 10^5 pairs of cascades using a Jaccard index over the sets of constituent nodes (i.e., the fraction of individuals common to both cascades out of the total individuals in both) and (normalized) tree edit distance, which judges both membership and structural similarity. Figure 4 shows a heatmap of the result.

The Pearson correlation coefficient between the measures is $\rho = 0.91$, which is perhaps surprisingly high. In particular, consider the pairs of cascades with reach set similarity $s_\rho = 1$

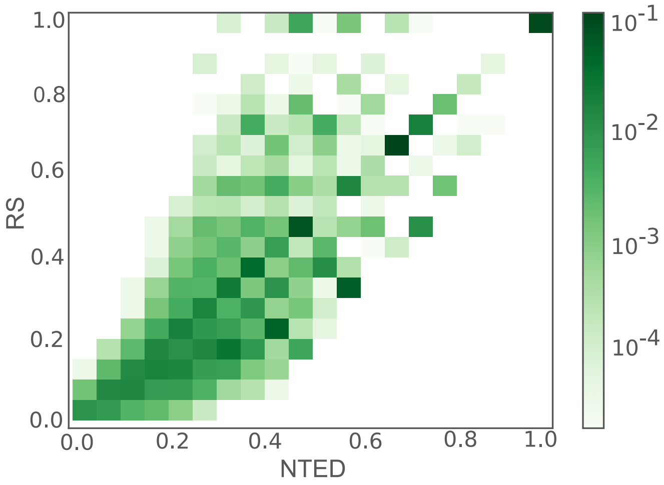


FIG. 4. Comparing similarity of pairs of cascades using reach set (RS) similarity (y axis) and normalized tree edit distance (NTED) similarity (x axis) results in correlation of $\rho = 0.91$. Since the RS measure captures similar membership while NTED captures similar structure, their correlation indicates a habitual hierarchy. The result is for 10^5 pairs of cascades in the data.

TABLE I. Distribution of root nodes by time of cascade. Persistent cascades reveal groups that have a tendency for exclusively weekend or weekday information spreading. ("Only weekend or weekday" signifies at least 90% of events. Fridays are designated as the weekend.

Cascade type	Data set	Only weekend	Mix	Only weekday
all	city A	<1%	99.2%	<1%
all	city B	<1%	99.4%	<1%
persistent	city A	1.8%	82.5%	15.6%
persistent	city B	2.6%	84.5%	12.9%

but tree edit distance similarity $s_\delta < 1$; these are conversations among the same individuals, just in different order. Remarkably, this type of cascade pair makes up less than 1% of the sample. This result implies that when the same individuals pass information, they tend to do so in the same order. We may consider this a realization of some habitual hierarchy in group conversations.

2. Weekend vs weekday roles

In Table I we examine all cascade initiators with at least one persistent class and at least three persistent cascades. If we consider all cascades of this group (not just persistent ones), we see that there is an even mix throughout the week, as expected: Nearly all users are generating cascades (that is, making calls to multiple people) on some mix of both weekend and weekdays. Very few users (less than 1%) are active exclusively on weekdays and/or weekends.

However, if we examine only persistent cascades, two new groups emerge: a large portion of root users who only initiate persistent cascades on weekdays and a slightly smaller portion who only initiate on weekends. These two extremes constitute over 15% of all root users, while the same extremes measured in all cascades are less than 1%. This is a complement to the observation that people have different mobility similarities than weekend and weekday contacts [40].

In other words, for these two groups, although they make calls throughout the week, their role in spreading information appears to be specialized: Their only persistent patterns of information spread happen during either weekday (i.e., work week) hours or weekend hours, but not both. Their other communication is sporadic, or random, and one might conclude, not meaningful.

3. Comparison to a null model

We compared the resulting distributions of size and frequency of cascades in the real data, simulation model, and analytical model over an arbitrary 2-month period of data. Results are shown in Fig. 5 for both the 3-person and 4-person cases (i.e., persistent cascades among 3 and 4 people, respectively). We see that the results of the synthetic data and analytical model are similar, as designed, with slight deviations due to approximations in the analytical methodology.

Importantly, in both cases, the real data exhibit a heavy tail of cascades occurring five or more times, in contrast with the null model, which exhibit near-zero probability of recurrence past four to five. This lack of long-term persistence in the model is due to the exponential decay of the binomial distribution. This allows us to reject the hypothesis that degree distribution and call volume (i.e., the distribution of average rates of activity) are enough to explain the patterns in the data. In the Conclusion we mention possible improvements to this model (e.g., would a nonhomogeneous edge process be enough to explain the recurrence?).

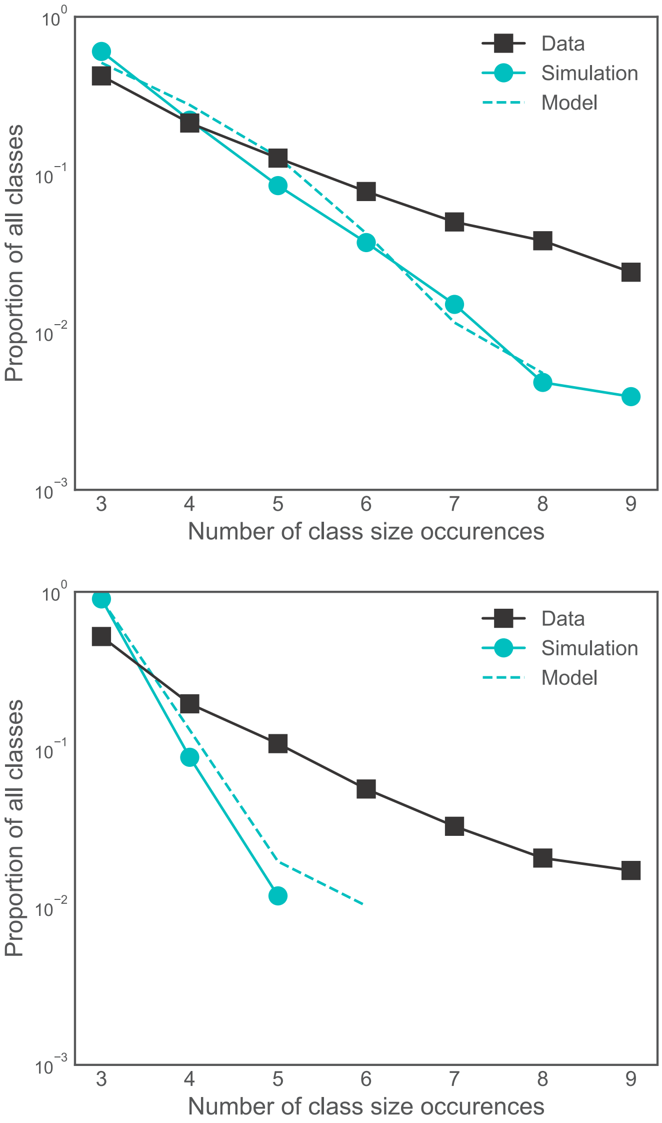


FIG. 5. Data exhibit significantly more recurring patterns than we would observe under Poissonian communication in a network with similar degree distribution. Shown are distributions of $\hat{Q}_\chi$ for $(|\chi| = 3)$-person (top) and 4-person (bottom) cascades.

4. Effect on connectedness and centrality

Now consider applying this knowledge of persistent structure back to a static structure and observing the effect on, in particular, centrality. Specifically, consider a static network $G = (V, E)$ consisting of all individuals observed making calls over a time period T, placing an edge (ij) whenever $i, j \in V$ make at least m calls.

We weight the subset of edges E_C that are present in at least one persistent cascade with $w_c = \alpha \in [0.5, 1]$ and all

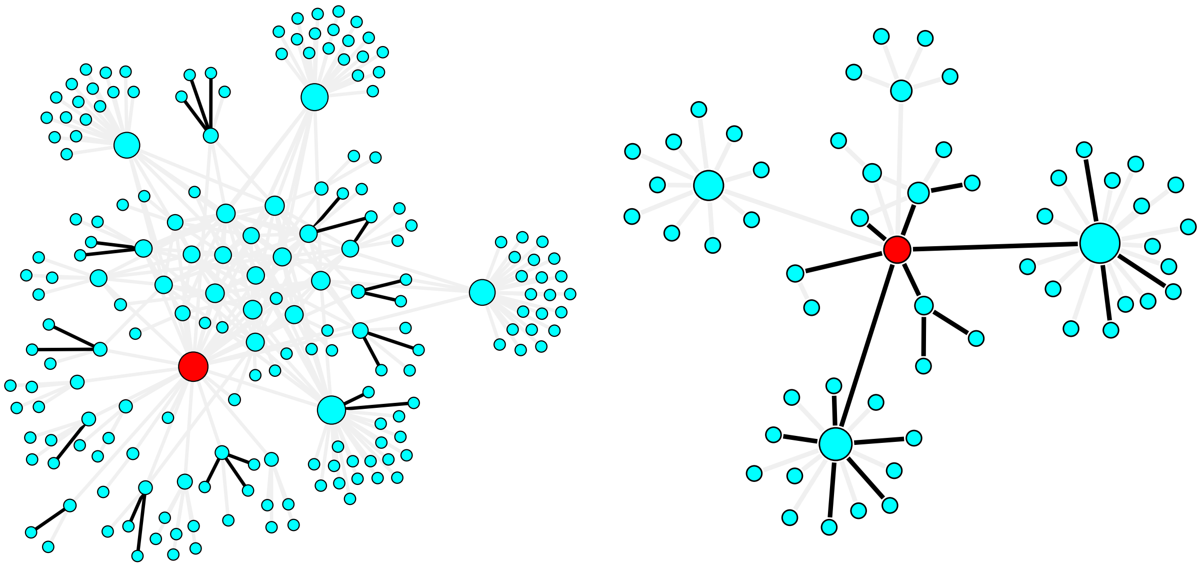


FIG. 6. Two example individuals' ego networks, with persistent edges in thick black. On the left is a high-degree node which participates in no persistent information cascades and on the right is a hidden spreader with an unremarkable number of connections but a large proportion of which are persistent.

$e_n \in E \setminus E_C$ with $w_n = 1 - \alpha$. With $\alpha = 0.5$ we recover the standard aggregated network and with $\alpha > 0.5$ we are putting extra weight on the persistent edges which we claim carry more meaning.

For city A over a $T = 1$ month period, this results in a network of approximately 278 000 nodes and 505 000 edges, with about 45 000 users having at least one persistence class of 2 or more cascades. (Results are similar month to month.) Setting α in [0.5,1), we find a large connected component comprising 80%–85% of the total network for all three data sets (cf. [3]). With $\alpha = 1$, the large connected component splits into several thousand smaller subgraphs, the largest being approximately 2000 nodes. This echoes previous results that show the inability of information to reach any sort of macroscopic diffusion when traveling solely through information cascades [19].

Consider the weighted degree (sometimes node strength) of a user i, defined $k_i = \sum_j A_{ij}$, where A is the adjacency matrix of G and $A_{ij} = w_c$ if $(i, j) \in E_C$, w_n if $(i, j) \in E \setminus E_C$, and 0 otherwise. (We use the tree edit distance similarity measure s_δ for this analysis, with $\ell = 0.8$.) Considering again a 1-month time period in city A, for both the unweighted (i.e., $\alpha = 0.5$) and cascade-weighted ($\alpha > 0.5$) networks, we compare the individuals identified as high centrality (top 10% of users) in both groups in Table II.

We note several groups that emerge. First is the large group of people (about 6% of the total population) that are only central in the cascade-weighted network. This suggests a group of individuals with unremarkable importance as measured in a naive way by counting calls, but who play a pivotal role in the persistent communication patterns of their social network. Similarly, a large group of influential users in the standard unweighted network disappears when we begin weighting cascades, implying their centrality was only due to a web of edges corresponding to mostly random calls. Finally, we note that a large portion of the network has their status essentially unchanged.

C. Measuring the role of persistent cascades in diffusion

The analysis in [32], outlined in the Introduction, postulates that there are information cascades doing the heavy lifting of spreading information and whose effects are only masked under large infectivity λ. We claim to have actually identified conversations displaying these cascading properties, the so-called persistent cascades of this paper. Our hypothesis is then that when λ is small and we follow the real order of interactions, the persistent cascades will play a significant role in spreading and cascade membership will contribute to higher probability of infection (i.e., receiving information). On the other hand, when λ is large, the cascades' importance will be masked by the high volume of random calls and we will see no significant difference between cascade membership or not. In another sense, we observe that there is both random and cascading activity occurring simultaneously in the real data and we have identified the individuals constituting both groups, and so by tracking the epidemic spread separately for both, we should see the contrast in infective

TABLE II. Persistent cascade effect on notions of centrality. When edges are weighted by persistent cascade activity, a group of hidden spreaders emerges (the top right group is 6.8% of the population in this sample). Bottom ranked refers to the bottom 90% of users and top ranked to the top 10% of users.

		Weighted	
	k_i degree rank	Bottom ranked	Top ranked
Unweighted	Bottom ranked	192 626 (83.2%)	15 771 (6.8%)
	Top ranked	15 771 (6.8%)	7385 (3.2%)

dynamics between regimes of λ without even randomizing the order of calls.

Note that in this section we simulate the spread of information but use the *real* order and timing of interactions observed in the data. We of course do not have access to second-by-second tracking of the spread of some real piece of information or news through the network (as we might in a Twitter or internet blog or email data set), but we are instead claiming that if such a spreading process were occurring, where the probability of the news being passed was λ, our simulations reveal the dynamics of what that spread would be.

Therefore, we simulate the susceptible-infected-recovered model in the temporal social network resulting from one month of cell-phone data. We start each simulation by choosing at random 1000 nodes and considering all other nodes as susceptible. We ensure there is an equal probability of cascade members or nonmembers chosen as seeds in each simulation. We then step through the call data in order and in each call letting the caller infect the callee with probability λ. Infected nodes recover after a period τ, and cannot be infected again. We continue until all nodes are susceptible or recovered. We repeat this for 100 simulations of 1000 seeds spread across the network.

Throughout the simulation we monitor two populations: those individuals involved in a persistent cascade and those not (see Fig. 6). We consider two regimes of infectivity, $\lambda = 0.05$ and $\lambda = 0.3$, with the recovery period $\tau = 3$ days. We measure the probability that a node is infected by counting the fraction of times it is infected over all simulations and average this across all nodes in a particular type of cascade membership and range of call activity. We control the number of total calls since we want to separate out any increase in probability of infection from simply having more exposure in general. (Note that this gives a series of conditional probabilities, not a distribution.)

The specific values of λ are chosen to be comfortably far away from the transition point (i.e., when the spreading process tends to become population scale) on each side and follows [32]. This transition point is determined empirically to be approximately $\lambda = 0.15$ for these data. Our results are in Fig. 7, which plots the conditional probability of being infected given some range of total call activity and population membership.

We find that the argument from the preceding section is borne out: When λ is low, the group conversations are doing the majority of work of spreading and so we see an increased probability of infection for these individuals. By contrast, when λ is high, the long tails of inactivity in these conversations allow them to be overpowered by the more random calls going on in the rest of the population.

We find that this result remarkable, as it implies that there is a special subgroup of the population who are more likely to receive information when receiving information is difficult, which in some sense is the more interesting case. Moreover, we are able to identify this subgroup using the methodology outlined in Sec. II, and we know their membership in the group is by definition persistent over long periods of time.

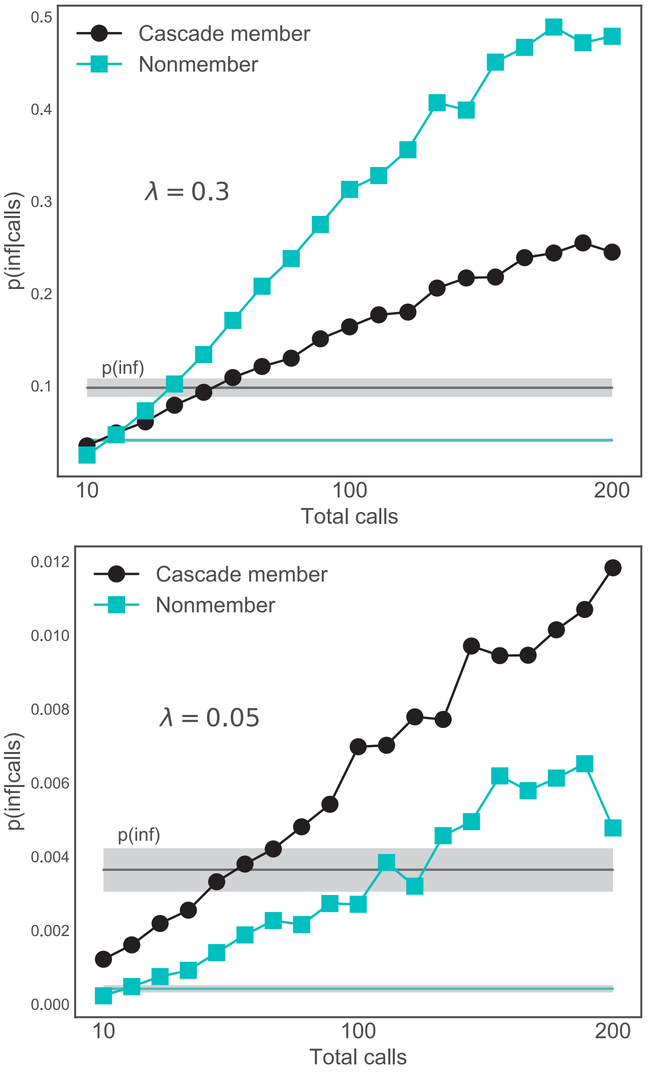


FIG. 7. Probability of receiving information given a particular range of total call activity for high (top) and low (bottom) infectivity. Persistent cascade members (black circles) are more likely to receive information in a low-infectivity regime than nonmembers (cyan squares). The population average across all simulations (not conditioned on call activity) is shown as a solid line, with two standard deviations above and below shown as a shaded rectangle.

IV. CONCLUSION

We introduced a method for extracting temporal patterns of information spread from large-scale communication metadata, using methods of inexact tree matching and hierarchical clustering. We showed that analysis of these so-called persistent cascades reveals different properties of information spread, such as weekday-weekend roles, a habitual hierarchy of spreading, and long-term persistence on the scale of months and years. The analysis also leads to an understanding of centrality and revealed a population of superspreaders who were otherwise unremarkable under an aggregated approach. We then showed that these patterns are significant by comparing them to both analytical and simulated models of the network, indicating that the temporal clustering inherent in real communication patterns is critical to producing the persistent cascading patterns we observe in the real data.

Finally, we also demonstrated that these persistent cascades play a crucial role in information spreading through simulation of diffusion processes on the temporal network. Specifically, members of a persistent cascade are more likely to receive information spreading through the network under conditions where the probability of transmission is low.

A. Future work

This analysis imposes a specific ideal on the structure of information spread, that is, cascading structure; it may be revealing to examine instead a more general model of spread (such as the temporal subgraphs examined in [8]) while still incorporating inexact graph matching.

The null model we proposed uses a homogeneous Poisson process as the underlying model of interindividual communication, and we show that this is not enough to explain the persistence observed in the data. However, it may be possible to explain the persistence using a nonhomogeneous Poisson process that varies with time, such as demonstrated in [41].

We also hope to validate methods proposed in this paper, and others, by using large-scale communication data where the content *is* known (e.g., email data). In this way, we can apply the method on the anonymized metadata and then reveal the content to validate our claims.

ACKNOWLEDGMENT

This work was partially supported by the Draper internal grant.

[1] A. V. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, Gossip: Identifying central individuals in a social network, National Bureau of Economic Research Report No. 20422, 2014 (unpublished).

[2] M. E. J. Newman, The structure and function of complex networks, SIAM Rev. **45**, 167 (2003).

[3] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, Structure and tie strengths in mobile communication networks, Proc. Natl. Acad. Sci. USA **104**, 7332 (2007).

[4] P. Holme and J. Saramäki, Temporal networks, Phys. Rep. **519**, 97 (2012).

[5] S. Huang, A. W.-C. Fu, and R. Liu, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne* (ACM, New York, 2015), pp. 419–430.

[6] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, Small but slow world: How network topology and burstiness slow down spreading, Phys. Rev. E **83**, 025102(R) (2011).

[7] M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, and M. Karsai, Multiscale analysis of spreading in a large communication network, J. Stat. Mech. (2012) P03005.

[8] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, Temporal motifs in time-dependent networks, J. Stat. Mech. (2011) P11005.

[9] L. Kovanen, K. Kaski, J. Kertész, and J. Saramäki, Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences, Proc. Natl. Acad. Sci. USA **110**, 18070 (2014).

[10] R. Lambiotte, L. Tabourier, and J.-C. Delvenne, Burstiness and spreading on temporal networks, Eur. Phys. J. B **86**, 320 (2013).

[11] S. Morse, M. C. González, and N. Markuzon, in *Proceedings of the 2016 IEEE International Conference on Big Data, Washington, DC, 2015*, edited by J. Joshi *et al.* (IEEE, Piscataway, 2016), pp. 969–975.

[12] A. L. Barabási, The origin of bursts and heavy tails in human dynamics, Nature (London) **435**, 207 (2005).

[13] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási, Uncovering individual and collective human dynamics from mobile phone records, J. Phys. A: Math. Theor. **41**, 224015 (2008).

[14] J. L. Iribarren and E. Moro, Impact of Human Activity Patterns on the Dynamics of Information Diffusion, Phys. Rev. Lett. **103**, 038702 (2009).

[15] J. L. Iribarren and E. Moro, Branching dynamics of viral information spreading, Phys. Rev. E **84**, 046116 (2011).

[16] A. Vazquez, B. Rácz, A. Lukács, and A.-L. Barabási, Impact of Non-Poissonian Activity Patterns on Spreading Processes, Phys. Rev. Lett. **98**, 158702 (2007).

[17] J. P. Bagrow, D. Wang, A.-L. Barabási, and Y. Moreno, Collective response of human populations to large-scale emergencies, PLoS One **6**, e17680 (2011).

[18] C. Hui, Y. Tyshchuk, W. A. Wallace, M. Magdon-Ismail, and M. Goldberg, *Proceedings of the 21st International Conference on the World Wide Web, Lyon, 2012* (ACM, New York, 2012), pp. 653–656.

[19] F. Peruani and L. Tabourier, Directedness of information flow in mobile phone communication networks, PLoS One **6**, e28860 (2011).

[20] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. De Menezes, K. Kaski, A. L. Barabási, and J. Kertész, Analysis of a large-scale weighted network of one-to-one human communication, New J. Phys. **9**, 179 (2007).

[21] G. Krings, M. Karsai, S. Bernhardsson, V. D. Blondel, and J. Saramäki, Effects of time window size and placement on the structure of an aggregated communication network, EPJ Data Sci. **1**, 4 (2012).

[22] S. W. Linderman and R. P. Adams, in *Proceedings of the International Conference on Machine Learning, Beijing, 2014*, edited by E. P. Xing and T. Jebara (PMLR, 2014), Vol. 32, pp. 1413–1421.

[23] K. Zhou, H. Zha, and L. Song, in *Proceedings of the International Conference on Machine Learning, Atlanta, 2013*, edited by S. Dasgupta and D. McAllester (PMLR, 2013), Vol. 28, pp. 1301–1309.

[24] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song, in *Proceedings of the Conference on Advances in Neural Information Processing Systems, Montreal, 2015*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran, Red Hook, 2015), Vol. 28, pp. 1954–1962.

[25] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, Inferring networks of diffusion and influence, ACM Trans. Knowl. Discov. D **5**, 1 (2012).

[26] P. Bogdanov, M. Mongiov, and A. K. Singh, *Proceedings of the IEEE International Conference on Data Mining (ICDM), Vancouver, 2011* (IEEE, Piscataway, 2011), Vol. 1 pp. 81–90.

[27] Y. Hulovatyy, H. Chen, and T. Milenković, Exploring the structure and function of temporal networks with dynamic graphlets, Bioinformatics **31**, i171 (2015).

[28] L. Tabourier, A. Stoica, and F. Peruani, *Proceedings of the International Conference on Communication Systems and Networks (COMSNETS), Bangalore, 2012* (IEEE, Piscataway, 2012), pp. 1–7.

[29] V. Sekara, A. Stopczynski, and S. Lehmann, Fundamental structures of dynamic social networks, Proc. Natl. Acad. Sci. USA **113**, 9977 (2016).

[30] Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, and W.-C. Lee, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, 2010* (ACM, New York, 2010), pp. 1645–1648.

[31] A. Anderson, D. Huttenlocher, J. Kleinberg, J. Leskovec, and M. Tiwari, *Proceedings of the 24th International Conference on the World Wide Web, Florence, 2015* (ACM, New York, 2015), pp. 66–76.

[32] G. Miritello, E. Moro, and R. Lara, Dynamical strength of social ties in information spreading, Phys. Rev. E **83**, 045102(R) (2011).

[33] M. E. J. Newman, Spread of epidemic disease on networks, Phys. Rev. E **66**, 016128 (2002).

[34] K. Zhang and D. Shasha, Simple fast algorithm for the editing distance between two trees and related problems, SIAM J. Comput. **18**, 1245 (1989).

[35] T. Henderson and S. Johnson, Zhang-Shasha: Tree edit distance in Python, available at https://github.com/timtadh/zhang-shasha

[36] Y. Li and C. Zhang, A metric normalization of tree edit distance, Front. Comput. Sci. **5**, 119 (2011).

[37] R. K. Pan and J. Saramäki, Path lengths, correlations, and centrality in temporal networks, Phys. Rev. E **84**, 016105 (2011).

[38] M. Marder, Dynamics of epidemics on random networks, Phys. Rev. E **75**, 066103 (2007).

[39] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications.Phys. Rev. E **64**, 026118 (2001).

[40] J. L. Toole, C. Herrera-Yaque, C. M. Schneider, and M. C. González, Coupling human mobility and social ties, J. R. Soc. Interface **12**, 20141128 (2015).

[41] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral, A Poissonian explanation for heavy tails in e-mail communication, Proc. Natl. Acad. Sci. USA **105**, 18153 (2008).