<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Proceedings of the IEEE*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Holme, P. (2014)

Analyzing Temporal Networks in Social Media.

Proceedings of the IEEE, 102(12): 1922-1933

<http://dx.doi.org/10.1109/JPROC.2014.2361326>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-96754>

Analyzing temporal networks in social media

Petter Holme

Abstract—Many types of social-media metadata comes in forms of temporal networks—networks where we have information not only who that is in contact with whom but also when the contacts happen. We review methods to analyze temporal networks developed the last few years applied to social media data. These methods seek to identify important spreaders and, in more generality, how the temporal and topological structure of interaction affects spreading processes.

Index Terms—Social network services, Temporal networks, Network analysis

I. INTRODUCTION

MANY types of interactions in social media form temporal networks—collections of unique interaction events between pairs of individuals [1]. The structure of these temporal networks determines the dynamics of information spreading. To study such phenomena is of interest, not only from an academic standpoint, but also for word-of-mouth marketing and similar types of technologies that relies on information spreading between people.

To understand how a large, integrated system functions as a whole, one needs to zoom out and look at it from a distance. In other words, one needs a consistent way of simplifying and discarding irrelevant information. A solution that has gained a huge amount of interest the last decade goes under the names of complex networks, or network theory. In this paradigm, one only keeps the information about the interacting units (in social media, that would typically be individuals or advertisers) and who that has been in contact with whom. Usually, even if one is concerned only with metadata and not message content, one has more information than only this static network. The advantage with looking only at a static network (a simple graph in mathematics jargon) is that one has a huge toolbox for analyzing the data. If one also includes information about when interaction events occur—i.e. studies a temporal network rather than a static network—then the number of methods are much more restricted. The good news is that the study of temporal networks has been a very active field the last few years, so the number of methods that study synergetic temporal and topological effects is increasing fast.

In this paper, we will review temporal network methods and

discuss how they can be used to understand spreading events, identify influential spreaders and describe social organization as reflected in activities on social media.

II. TEMPORAL NETWORKS AS A MODELING FRAMEWORK

A. Representations and types of interactions

The two main classes of mathematical representations of temporal networks are *contact sequences* and *interval graphs*. These two concepts are more similar than their names suggest. In contact sequences, a contact is assumed to be instantaneous with respect to the temporal granularity of the data. They could thus be represented as triples (i,j,t) encoding an interaction between individual i and j at time t . In interval graphs one can picture an interaction event as being temporally extended, so it has a beginning and end time. Mathematically, a contact would be represented as a quadruplet rather than triple— (i,j,t,t') , meaning that i has a contact with j that lasts from time t to t' . In social media the interaction is usually directed, so that the order of the individuals i and j carry a meaning in the contact representations. In many cases the same methods can be used for contact sequences and interval graphs. The latter can always be reduced to the former at the expense of some information loss. In this paper, we focus on contact sequences. Fig. 1 illustrates some ways to draw a contact sequence that displays all its information. It is a system of six vertices, and already quite complicated. Unfortunately, displaying social media data like these figures would not be very informative. They serve to discuss our concepts.

Social media interaction that would suit the contact sequence format includes direct (e-mails or e-mail like) messages [2,3], Facebook wall posts [4] and friendships [5,6], Internet forum posts [7,8], blogs [9], micro-blogging [10], location sharing [11], etc. Usually, assigning a time of an event is not completely trivial. From a spreading point of view, the relevant time would be when the recipient is exposed to the information, but it is often the sending time that is stored in the logs of the media. Most studies use this reception time as the only time stamp of the contacts. This causes the short time order of event to be jumbled and alter the paths the information can flow, but probably not change the results for larger spreading events. (A thorough study of this bias would be interesting.)

Interaction that takes the form of a dialog, either over a voice or chat channel could be better modeled as an interval graph. This type of data is rather well studied when it comes to mobile phone communication [12-14] (perhaps a border case of social media). We are not aware of any study addressing voice call features embedded in social media platforms (per-

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2011947).

Petter Holme is with the Department of Energy Science, Sungkyunkwan University, 440-746 Suwon, Korea (e-mail: holme@skku.edu) and IceLab, Department of Physics, Umeå University, 90187 Umeå, Sweden.

Fig. 1. Graphical representations of temporal networks. This figure shows three ways of displaying a temporal network (more specifically a contact sequence). Panel A shows a plot of the network of aggregated contacts with the edges annotated with the times of contacts. Panel B shows a timelines of the vertices where a contact is indicated by a horizontal line. Panel C displays a timeline of edges where each contact between the vertex pair is indicated by a triangle. This temporal network illustrates the non-transitive nature of temporal networks. Information spreading over the contacts could reach from A to B (e.g. via the contact at time 17) and from B to D, but not from A to D via B, because by the time the information would have reached B, all contacts between B and D has already happened.

haps it makes sense not to classify these as social media anyway). In the rest of this paper, we will restrict our discussion to contact sequences. One reason is that one can, and often does, reduce interval graphs to contact sequences, either by discretizing time and logging one contact in every interval of the contact, or by taking the beginning (or middle) of the interval as the interaction time.

B. Models for spreading processes

There is a large and fast evolving field studying social influence and other forms of information spreading between people that are applicable to social media. There are two main classes of such models—*simple* and *complex contagion* models. The first class seeks to capture processes where only one individual affects another (relays a rumor, changes an opinion, etc.). Input from more than one person is negligible with respect to the system-wide dynamics. There is a popular idea that information (Internet memes as a notable example [15]) spreads

like infections (as reflected in expressions such as “go viral”). Also in the academic literature the idea goes back about half a century [16]. Even if this analogy can capture the fundamental features of the spreading, it can probably not hold in a strict sense—transmission probabilities in information spreading probably have much more individual variation than infectious disease spreading, and they can be related to ongoing events (elections in the examples of Ref. [15]) which is something you would never see for disease transmission. Nevertheless, if the fundamental properties of the contagion are the same as for infectious diseases, then one can of course borrow models from epidemiology to understand spreading in social media. The basic class of models for infectious disease spreading is called compartmental models [17]. These divide people into classes, or compartments, with respect to the disease—canonical classes include susceptible (S), infectious (I) and recovered (R). The latter class comprises individuals that have stopped being infectious and acquired immunity to the pathogen, or died. In a social-media context, R would correspond to individuals that have had the information but stopped being interested in spreading it further (like a once viral video that people have lost interest in). The second defining feature of compartmental models is that they assign transition rules between the classes. One of the rules that is (to the best of our knowledge) always included is that a susceptible, upon encountering an infectious, with some probability can become infectious.

Another type of simple contagion models is the ones derived from the *voter model* [18,19]. These are typically non-equilibrium models to show the spreading of an opinion in a population. In the standard setting, one lets the opinion of a person be an integer valued variable. Then one updates the system by iteratively picking a random node i , then picking a random neighbor j of i and finally copying j 's opinion to i . This model is analytically tractable, but other than that maybe something of an oversimplification as a model of real information spreading. Typically one is interested in the time it takes for the voter model to reach a state where every node has the same opinion as its neighbors.

In complex contagion, the spreading can be contingent on the interaction with more than one other person. A popular idea that opinion spreads following a threshold model [20]. The idea is that social contagion could happen if someone has been influenced by a certain fraction of others. In a temporal network one would also have to decide how to take time into account. For most cases, social influence that happened a long time ago will not matter for e.g. the adoption of a product or spreading of an opinion. Ref. [21] discusses this further and uses a sliding time window to represent the duration of possible influence. Ref. [22] uses an exponentially decaying weight for the same purpose in a bit more elaborate model of social contagion.

C. Time-respecting paths, components and distances

Any process that is confined to a temporal network has to follow sequences of contacts of increasing times. (In some

situations, depending on the spreading process in question, one would also allow non-decreasing times, so that the spreading could pass two contacts at the same time step.) There are a few different names for such a sequence of contacts—we will follow Kempe and call it a *time-respecting path* [23]. Although time-respecting paths, as a concept, is the counterpart of the paths in static graphs, it has some fundamentally different properties. To begin with, if, at time t , one can reach node j from node i by a time-respecting path that does not imply there is a time respecting path from j to i at time t . In other words, the property of i being connected to j is not commutative. This property is the same for directed graphs. As a consequence, just like for directed networks, one can talk about *strongly* and *weakly connected components* in temporal networks. A strongly connected component is a set of vertices where one can reach from any vertex to any other

vertex following time-respecting paths (e.g. B, C, D and E in Fig. 1). To define a weakly connected component, we first construct an undirected graph where edges represent pairs that have at least one contact throughout the sampling time. A connected component in this graph would then be a weakly connected component in the original temporal network.

There are, however, other features of temporal networks that set them aside from directed graphs. Most conspicuously, time-respecting paths are not transitive. This means that even if, at time t , there are time-respecting paths from i to j and from j to k , there are not necessarily any time-respecting path from i to k . This is the case if all time-respecting paths from j to k have already happened by the earliest time a time-respecting path from i reaches j . As a word of caution, since a time-respecting path is a collection of contacts that happens at certain times, a statement such as “there is a time-respecting

Fig. 2. Illustration of various randomization schemes. Each panel shows a possible instance of a randomization scheme applied to the network in Fig. 1. In panel A, the Random Permuted Times method (RP), the time stamps are randomly shuffled between the contacts. The resulting temporal network has the number of contact per time and the number of contacts per edge conserved. For example, at time 2 there are two contacts both before and after applying the randomization scheme (see the boxed number). Similarly, the edge (B,C) has three contacts before and after the randomization. In panel B, we show the Random Times (RT) scheme where time stamps are not swapped (like RP) but created. Thus this randomization does not conserve the number of contacts per time period (so, as illustrated, at time two the number of contacts changes from two to zero). In C, we illustrate the Random Contacts (RC) randomization, which does not conserve the number of contacts per edge. After the randomization the (B,C) edge has two contacts, rather than three. Panel D shows the Randomized Edges (RE) scheme that keeps the degrees in the network of accumulated contacts constant, but randomize the edges. Furthermore, the timelines of the edges are held fixed. Panel E shows the combination of RC and RE where the only thing conserved is the degrees in the network of accumulated contacts.

path between i and j is ambiguous. One always needs to specify a time t , meaning that at t you can reach j from i following a time-respecting path. In Fig. 1 there is a time-respecting path from F to A at time 5 but not at time 10. Also for components one needs to specify a time. An interesting extension of the component concept would be to consider sets of vertices that are, within a time window, transitive.

A useful concept related to time-respecting paths is the set of vertices that, at a certain time, can be reached by time-respecting paths from vertex i is called the *set of influence* of i . This is important for spreading processes, as it is the set of vertices that can eventually be influenced by i . Some studies have measured the *reachability ratio*—the average fraction of vertices in sets of influence averaged over all beginning times during the sampling time [24]. Similarly to the set of influence, one can also define the *source set* of i —the set of vertices that can reach i through time-respecting paths within an observation window. This set consists of all vertices that can have been the source of a spreading process influencing i . Moody [25] is the earliest we are aware of that studies the size of the source set (the *source count*).

Of course, since the source set and set of influence are time specific, one may also monitor the reachability ratio and source count functions of time, i.e. study how many other vertices may reach vertex i by time-respecting paths by time t' , when the paths begin no earlier than $t < t'$.

For static graphs, the *distance* between two vertices is defined as the length of the shortest path joining them (where path length being defined as the number of links forming a path). Short average distance is, of course, assumed to be a sign that a temporal network is efficient with respect to spreading. One of the main findings for static networks is that they are often remarkably compact. This “small-world phenomenon” exists in virtually all empirical networks except where links are physical objects (like roads, power-lines, etc.). Naturally, when the dimension of time is added to the picture, it is useful to define similar quantities characterizing how quickly vertices can reach each other through time-respecting paths. Cooke and Halsey studied this kind of quantity in the 1960’s [26]. The precise definitions of such quantities are not completely straightforward. How should one, for example, compare two if one exist half of the sampling time and another only 10%.

A time-respecting path is associated with a *duration*, measured as the time difference between the last and first contacts on the path. (Even though the dimension is time, some authors have called it the temporal path length [12].) So the duration of the path from F to A in Fig. 1 is 18 (the first contact that can be involved in spreading is between F and E at time 1, the last contact is between B and A at time 19). Analogously to the shortest paths of static graphs, that define the distance, one can study the fastest time-respecting path(s) between two nodes.

The concept of latency was originally introduced in the field of distributed computation [28]. A central problem in this area is to keep track of the age of information that a vertex has about other vertices. Then one commonly assumes that verti-

ces in contact update each other’s information so that, after the contact, both vertices share the most recent information that either of them had before the contact. This scenario is similar to the fastest-spreading limit of spreading processes. (For example, a disease-spreading model with 100% per-contact infection probability.)

We will sketch the framework of latency as introduced by Lamport [27] and further developed by Mattern [28]. Consider information spreading in a temporal network and, specifically, the vertex i . Then let $\phi_{i,t}(j)$ denote the latest time such that information from j could have reached i by time t . This quantity is called i ’s *view* of j ’s information at time t . Furthermore, $\lambda_{i,t}(j) = t - \phi_{i,t}(j)$ is called j ’s *information latency*, or just *latency*, with respect to i at time t , and is thus a measure of how old i ’s information coming from j is at time t . For example, the latency from F to A in Fig. 1 at $t = 17$ is 10. In other words, $\phi_{A,17}(F) = 7$, so $\lambda_{A,17}(F) = 17 - 7 = 10$. Finally, the vector $[\phi_{i,t}(1), \dots, \phi_{i,t}(N)]$ is called i ’s *vector clock*. A difference to other approaches is that this concept is looking backwards in time. Looking forward in time, one may define a quantity corresponding to latency (*temporal distance* [12]) $\tau_{i,t}(j)$ that measures how long it takes to reach j from i along the fastest path, starting the clock at time t . The expected temporal distance for a random starting time is called *reachability time* [24]. An adaptation of the vector-clock concept to social media data can be found in Ref. [29].

As we allude to above, latency and vector clocks form a basis for measuring times and optimal spreading speeds in temporal networks. However, taking an average over the sampling time to get a value for the entire graph—or even only for a pair of vertices—is not that straightforward. Problems are typically related to the finite time windows of empirical data sets. For example, as the time gets closer to the end of the sampling, the number of time-respecting paths between a pair of vertices decreases. If the sampling would be longer, one could presumably see more time-respecting paths starting even before the last time-respecting paths. One possible quantity for measuring the velocity of paths in general is to find all fastest time-respecting paths between vertices and then compute the average duration of such paths. This measure would, however, not reflect the frequency of the paths, and would not be affected by waiting times before the first contacts of such paths. For example, if one or ten time-respecting paths of one unit connect two vertices, this average duration would equal unity in both cases.

Measuring the average latency is also complicated by the fact that latency varies with time with a saw-tooth pattern [1]. The latency goes up linearly until there is contact that is the beginning of a new time-respecting path from the source node j , carrying newer information from i than i already has. Close to the beginning of the observation window the latency is infinite as no time-respecting path yet has reached its destination. In a steady state situation where all links are active throughout the sampling time, it would be a good approximation to assume the proceeding time interval of the sampling duration would be similar to the first. This suggests a boundary condition where one, repeat the entire temporal contact sequence,

Fig. 3. Reachability ratio and reachability time analysis for interactions at an Internet dating site. Panel A shows the reachability ratio for the empirical temporal network—the probability there is a time respecting path between a pair of vertices at a random time. Panel B shows the reachability time—given there is a time-respecting path from one vertex to another, then what is the expected time for the information to complete this path (assuming it spreads over a contact whenever it can). Averages over the randomizations are over 10^4 runs. Standard errors are not larger than 0.01 in A and 0.5 in B (and thus too small to be meaningful as error bars).

and thereby get around the problem that there are few paths in the beginning of the sampling time. However, this procedure may give rise to artifacts and connect pairs of vertices that are not connected at all within the observation window. Yet another option would be to average the time between the first and last contacts, which would underestimate any spreading processes especially for short paths.

For long enough periods of observation, another difficulty comes from the dynamics of vertices entering and leaving the system. For disease spreading in empirical data sets, this turnover of vertices and edges has been argued to be of a great importance [30,31]. In such a case questions about how large a spreading event can be is probably more important than the time it takes to reach between the vertices.

Finally, some words of caution about the terminology. Some authors use the terms as distance and length as measures of time—e.g. Kossinets *et al.* [32] define the “distance” between two vertices as the shortest duration of any time-respecting path between them. Tang *et al.* [33] calls the average time to reach vertices for time-respecting paths starting early in the data “temporal path length”. To be fair, we should point out that “distance”, as in the standard static graph definition is, being a dimensionless quantity, also a misnomer. Furthermore, statistics of latency have several names even though their definitions and purpose are very similar: “reachability time” in Ref. [24], “temporal proximity” in Kostakos [34], “characteristic temporal path length” in Tang *et al.* [33] and “temporal distance” in Ref. [12].

D. Randomization techniques

In static networks, network structures are keys to understanding both the forces that shape the network and how dynamic processes on the network behave. The loose definition of network structure is that it is the way the network differs from a random network. To make this definition operational, one commonly specifies a null model—a model for how to construct a network that is random except the most fundamental constraints. What is fundamental depends on the system one studies and the type of analysis one wants to do. Probably the most common null model for static networks is networks with

the same set of degrees (and consequently the same number of nodes and links) as the original network. One can easily sample such networks by randomizing edges—one picks random pairs of edges (ij) and $(i'j')$ and swap them to (ij') and $(i'j)$ or (i,i') and (j,j') (unless the swapping would introduce a self-link or multiple link). Then one could compare quantities, like the number of triangles, in the real network to the average number of triangles in the null model. If there is an over-representation of triangles in the real network compared to the null-model (which is commonly the case in social networks) one can conclude that the network was formed by process that has a bias for triangle formation (like people getting acquainted by being introduced by a common friend). One can, furthermore, study processes on the network (spreading phenomena for example) and comparing properties of the process in the real networks and the randomized networks.

Comparing the real data with a null-model based on randomizing the real data, becomes even more powerful in temporal networks—it is probably the only way to consistently analyze temporal and topological features of the data, and thus to discover synergetic features where time and topology together influence a spreading process. Since there are several possible temporal correlations and several time scales where the correlations can affect the system, there will not be any method as common as the random link shuffling procedure described above for static networks. Rather, by designing appropriate null models, one may switch off different types of correlations in order to understand their contribution to some quantity describing a process on the temporal network (such as a measure of the spreading speed). A typical use for such models in studies of spreading processes would be to apply them sequentially, and by monitoring how the dynamics of the process depends on these null models, pinpoint the role of different temporal and topological correlations on the process—if removing a certain type of correlations changes the dynamics more than another, then obviously the first played an more important role for the spreading process.

Below, we review temporal-network null models introduced in the literature (some of them illustrated in Fig. 2). The first paper using this type of methods was (to the best of our knowledge) Holme [24]. In the end of the section, we summarize and provide some guidelines for choosing reference models. A summary of the randomization techniques can be found in Table 1.

1) Randomized edges (RE)

This method is similar to the edge swapping for static graphs mentioned above, with the additional ingredient that contact sequences of edges follow the edges when these are rewired. In pseudo code, the method is defined as follows:

1. Go over all edges sequentially.
2. For every edge (ij) , pick another edge $(i'j')$.
3. With a probability 1/2 replace (ij) and $(i'j')$ by (ij') and $(i'j)$, otherwise replace them by (i,i') and (j,j') .
4. If the move in step 3 created a self-edge or a multiple edge, then undo it and start over from step 1.

The times of contact over an edge are kept constant. Note that the two alternatives in step 3 where one is randomly selected are needed to remove spurious correlations if (as is usu-

Fig. 4. Illustration of ways to simplify temporal network data. Panel A shows the reachability graph of the contact sequence displayed in Fig. 1. Panel B illustrates the link turnover picture as applied to the Fig. 1 network. In this picture one thinks of the first observed contact as the first of an edge and the last contact as the end of the edge. Panel C illustrates the ongoing link picture where one assumes that the temporal network can be well described as an underlying static network with contacts happening over the edges following some inter-event time distribution.

ally the case) the data structure stores edges returns the vertices of an edge in a specific order. Without this rule, one would keep the number of times a vertex appears in the first argument conserved, which could give very different results. This algorithm is usually not a bottleneck in a temporal-network analysis, but to speed up the process, one can skip edges that already have been rewired in 3.

Just as for static network, this null model can be used to study the effects of the network topology (apart from the degrees—the number of neighbors). The model implicitly also assumes that it is the edges rather than the vertices that govern the times of contacts. This is evident since after the randomization procedure, both the number and times of contacts for a vertex will probably have changed. However, their degrees in the aggregated network are the same. As the contact sequences follow their edges when rewiring, all temporal correlations and inhomogeneities associated with individual edges, such as burstiness and the distribution of inter-contact times of edges, are retained. Also the system-wide contact rates are unaltered.

2) Randomly permuted times (RP)

To understand the role of the order of the contacts one perform the RP randomization. In this procedure (used in e.g. Refs. [14] and [35]), one permutes the contact times randomly while keeping the network structure and the numbers of contacts between all pairs of vertices fixed. Technically, this is much simpler than applying the edge-rewiring scheme dis-

cussed above, as it only requires randomly swapping the time stamps of all contacts. No checks similar to step 4 of the RE rule need to be performed for contact sequences. Like RE, this scheme also retains the overall rate of events in the network at every point in time, such as daily or weekly patterns in communication networks.

3) Random times (RT)

The ensemble defined by RE and RP randomizations conserve the set of times of the original contact sequence. Hence, although it destroys time-structures of events related to individual vertices and edges, the rate of events in the entire temporal network is unchanged and will still follow the typical circadian and weekly patterns of human activity (see e.g. Refs. [36–39]). This type of randomization needs a process where the spreading depends on time, not only the order of contacts (such as the SIR model of disease spreading, where individuals would change state even though no contact happens). In maximal speed spreading such as the one behind the latency discussion above, this randomization would have exactly the same effect as RP. Note that an alternative to uniformly random contact times is to generate them from a specific distribution or process, such as the Poisson process, with parameters set up so that the numbers of contact per each edge are on average conserved.

4) Randomized contacts (RC)

Randomization method	Conserved structures
Randomized edges	Time sequence of edges. Degree distribution.
Randomly permuted times	Network of accumulated contacts. Number of contacts per edge. Times of contacts in general.
Random times	Network of accumulated contacts. Distribution of number of contacts per edge.
Randomized contracts	Network of accumulated contacts. Times of contacts in general.
Equal-weight edge randomization	Network of accumulated contacts. Number of contacts per edge.
Edge randomization	Network of accumulated contacts. Distribution of number of contacts per edge. Times of contacts in general.
Time reversal	Everything except the direction of time.

Table I. Summary of the randomization techniques and which structures they preserve (everything else is randomized).

For this randomization scheme, one keeps the graph topology fixed but redistributes the contacts randomly among the edges. After this randomization, the number of contacts per edge follows the binomial distribution. It is intended to test the effect of fat-tailed distributions of this quantity as typically seen, especially in social media and other forms of human generated communication data. If one would like to test the effect of the distribution of the number of contacts alone, keeping the structure of the temporal order of the real data, then one would need a different approach. For example, a vertex that is active primarily in the early stage of the data would be so in the randomized data too, one would need to compensate for such effects.

5) *Equal-weight edge randomization (EWER)*

Sometimes one would need to remove correlations between static network structure (the network of accumulated contacts), and at the same time retain the temporal structure of the edges (including the inter-event time distributions of edges). This is achieved by randomly swapping entire contact sequences of edges with the same number of contacts. I.e., all the contacts and their time stamps are randomly exchanged between edges that have the same number of contacts. Thus single-edge patterns, such as burstiness of the contacts between two individuals, are retained, together with other properties preserved by the RP model (like number of contacts of an edge, the system-wide contact frequency and the topology of the network of accumulated contacts). This null model requires a large enough system so that there are enough edges with the same number of events.

6) *Edge randomization (ER)*

This null model is similar to the EWER model with the exception that the sequences can be exchanged between edges that have any numbers of contacts. This corresponds to randomly exchanging the edge weights (measured as numbers of contacts) in the network of aggregated contacts. The correlation between weight and topology is destroyed in this null model. However, the inter-contact time distributions of contact sequences of edges are preserved—the sequences are just moved elsewhere in the network. Both EWER and ER were introduced in Ref. [12].

7) *Time reversal (TR)*

This null model is designed for assessing the frequency and importance of causal sequences [40] of contacts, where e.g. contacts trigger further contacts. It simply involves running

the original event sequence backwards in time. If sequences of consecutive contacts would be caused by temporal correlations alone, similar numbers of such sequences should be observed when time runs forwards and backwards. A lack of such chains in the time-reversed null model compared to the original sequence could be attributed to the arrow of time.

8) *Summary and guidelines*

The different randomized reference models discussed above retain and destroy specific kinds of topological and temporal correlations, and thus e.g. in studies of processes on the temporal network. They allow for pointing out the importance of various correlations—the most important correlations can be pinpointed by comparing the effects of different randomization models on the dynamics. The RE and RC models permute edges and contact times. Taken together they (RC+RE) destroy all correlations except for the degrees of the network of accumulated contacts—this provides a good starting point for the limiting case of uncorrelated temporal networks. If one randomizes the contact times (RT), the overall contact frequencies are also removed. When studying the roles of the exact contact timings on edges and the correlations between adjacent edges, comparing the EWER and ER models to the RT model should work, as the static network features are retained except for correlations between weight and topology that can be removed by ER. An illustration of some of these randomization schemes can be found in Fig. 2.

In Fig. 3A, we show an example of these techniques applied to a social media dataset. The temporal network in this case records the interaction (private messages, posts on home pages and friendship requests) at an Internet dating site [41]. The dataset records 536,276 contacts over 174,662 edges between 29,341 users for 512 days. First we note that there is a very strong effect of the temporal ordering. These randomization schemes all keep the network of accumulated contacts unchanged. Still the reachability ratio can differ more than a factor three. Even a very mild randomization such as RP—that tests the effect of the order of event—increases the reachability ratio from 29% to 47%. Comparing RP and RT, the reachability ratio increases further, meaning that the actual time things happen makes spreading reaching fewer people. Applying the RC scheme makes the spreading yet less efficient, so the heterogeneous distribution of contacts per edge [24] also limits the spreading. Randomizing the edges, on the other hand decreases the reachability ratio from its original

value, but not much. Changing the temporal properties (like the previously mentioned randomization schemes) is thus more influential than the topology of the network of accumulated contacts. Further proving this point, the combination of RE and RC gives the largest reachability ratio. In Fig. 3B, we show reachability times for the same dataset as in panel A. All the randomizations that preserves the network topology speeds up the spreading, meaning that all (considered) temporal structures slow down the spreading. This has been called the “slow-world effect” [14] (paraphrasing the “small-world effect”—that empirical static networks does often have very short average path length). The slow-world effect is not completely universal, Ref. [35] finds the reversed situation in a network of sexual contacts reported at a web forum.

E. Centrality and influential spreaders

There has been a tremendous interest in identifying important spreaders in the social media literature (see e.g. Refs. [42–44] and further references therein). Note that “influential” (someone who are likely to influence many others), is not necessarily the same as “important” (someone who is able to boost a spreading process). Important spreaders also need to be susceptible to influence [42]. Except obvious applications such as seeding word-of-mouth advertisement [42–44], finding important spreaders could e.g. be useful in the detection of disease outbreaks and other applications in public health (this is discussed further in the Conclusions section).

In static networks, concepts such as influential and important are often read synonymous to centrality. There are a number of centrality measures for static networks, each capturing a different facet of centrality. Many of these measures can be translated to temporal networks fairly straightforwardly. For example, betweenness—roughly how many shortest paths that passes through a vertex—can be adapted to temporal networks by changing shortest paths to shortest time-respecting paths. It gets a bit more complicated with measures based on distances [33]. Closeness is defined as the reciprocal average distance from a node to the other nodes of a static graph. A simple way of dealing with missing time-respecting paths is to measure the average reciprocal distance (rather than the reciprocal average distance) and let a missing time-respecting path make a zero contribution to the average [12,45]. Furthermore, as discussed in Ref. [46], measures like Katz centrality and PageRank can be generalized to temporal networks.

Another approach to identifying important spreaders is the vaccination problem—assume you could vaccinate a fraction of a population for a future disease outbreak, then how can this fraction be chosen. This differs from the centrality approaches of finding important spreaders as it also takes the chance of getting the disease (or information, if we think of social media) into account. A celebrated approach in static networks is the neighbor vaccination method—pick persons at random, ask them to name an acquaintance (or someone they met in a way that the disease in question can have spread), and then vaccinate the acquaintance. The probability the friend has a degree k is proportional to k ; so high-degree individuals have

a higher chance of being vaccinated. Lee *et al.* [47] proposed a version of this where one ask the randomly chosen person is to name the most recent, or most frequent (for some time into the past), acquaintance. Ref. [48] proposed further improved methods. Such methods can exploit temporal heterogeneities and improve the efficiency of neighborhood vaccination.

F. Network motifs

The idea behind network motifs comes from electronics. Small over-represented subgraphs, primarily in directed networks, could be interpreted as building blocks in the network like transistors in electronic circuitry [49]. In temporal networks, motifs have a bit different role. Rather, temporal network motifs are often thought of as common sequences of contacts among a small group of people. For example, Kovanen *et al.* [13] focus on contact sequences between vertices that are maximally separated by a time δt . Specifically, two contacts are said to be δt -adjacent if they share a vertex and are separated in time by δt or less. Pairs of contacts are then defined as δt -connected if there is a sequence of δt -adjacent events joining them. Then Kovanen *et al.* proceeds counting δt -connected subgraphs and comparing their frequency to those in randomized null models. They find an over-representation of subgraphs that seem to be causally connected (such as A contacts B who contacts C and D, as opposed to a non-causal sequence where B contacts C and D, and A only then contacts B). This work was motivated by cell-phone call data, but its methodology could be straightforwardly applied to social media data. Also not quite social media data, Jurgens and Lu [50] studied the evolution of Wikipedia by counting similar motifs. They associate common motifs with edit episodes such as “content reversion and anti vandalism” or “collaborative editing”.

G. Simplifying temporal networks

A final class of methods we will discuss is how to simplify temporal network data. As mentioned above, it is hard to visualize temporal networks to give a feeling of the structure of even mid-sized data sets. For this reason—and also the more fundamental purpose to understand the important structures for spreading processes—one would like to simplify temporal network data. One approach is to project a temporal network to a static network. The straightforward way—to include an edge between all pairs of individuals that has at least one contact during the sampling period is not always a good idea. If one is interesting in a spreading process, that kind of projection could include too many irrelevant edges [51]. Holme [52] proposes either a carefully selected time window (and make a network of aggregated contacts within that window), or an *exponential threshold representation* where each contacts contributes to an edge’s weight with a term that is decaying exponentially from the start of the spreading process. This representation was shown to perform well to make a static network where the static network predictors of node importance for disease spreading matches their actual importance in simulations directly on the temporal network.

Another type of projection to static graphs that could be

useful at least for very sparsely connected contact structures is *reachability graphs*, or “path graphs” [25], or “associated influence digraph” [53]. In these graphs, a directed edge from A to B means that (at the beginning of the data) there is a time respecting path from A to B. This type of graphs tends to be extremely connected for real-world networks and thus not well adapted to complex network methods. See Fig. 4A for an illustration.

Instead of simplifying temporal networks to static networks one could project them to simpler forms of temporal networks. Ref. [31] discusses “pictures” or components in models of temporal networks. In particular, two pictures are contrasted—a *link turnover picture* where one think of the first and last contact of an edge as its beginning and end and ignores the timing of the other contacts; and a *ongoing link picture* where one think of the links as continuously active and the contacts as drawn from a probability distribution (perhaps reflecting a bursty inter-event time statistics). Ref. [31] argues that, with respect to disease spreading and empirical data sets, the link turnover picture is a better way of simplifying temporal networks. Fig. 4B and C illustrate the link turnover and ongoing link pictures respectively. Ultimately, what determines which picture that is most relevant is if the relative time scales of the spreading processes and sampling time. If the sampling time is much shorter than spreading processes, then the ongoing link picture will be more relevant, and vice versa.

III. DISCUSSION AND CONCLUSIONS

Social media generates huge amount of metadata that could be used to understand social information flow, identify important spreaders, etc. Many kinds of such metadata could be represented as temporal networks—networks that records when contacts happen, in addition to who that has been in contact with whom. The field of temporal networks is still young and under rapid development. Probably, there are methods more specifically suited for social media data waiting to be developed. Nevertheless, there are a number of, primarily computational, techniques to analyze social media datasets that we discussed in this paper. These methods serve to identify important spreaders, characterize the behavior of the social media users and map out temporal and topological structures. There has recently been much advancement in the understanding of how such structures affect spreading phenomena [54–56]. It would be interesting to validate such theories using social media data. This would be going the other way than usual—inferring predictors of spreading cascades from data [42,57,58]. Both fields of temporal networks and theory of social media need each other—the former need real data to test and develop theories, the latter need a theoretical framework to handle large time-resolved data.

Word-by-mouth marketing and the detection and surveillance of infectious disease are two areas where temporal network methods have a great potential of improving existing methods [59]. These are areas where researchers already have applied static network methods to a great extent. At the same time, the underlying interaction has a strong temporal component meaning that temporal network techniques could poten-

tially be very fruitful. For example, Christakis and Fowler [60] propose to use neighbors of random individuals as sentinels for detection of epidemic outbreaks. Neighbors of average nodes are more central than average [61], which could be exploited. Christakis and Fowler do not discuss the use of social media data as such, but social networks in general. However, finding infected individuals from social media data is not difficult *per se* [62–65], so their method seems possible to straightforwardly extend to this type of data—and possible to improve by methods like the ones discussed in this paper (e.g. in Section II-E and G). Social media data could also be used to monitor the sentiments toward public health programs [66,67]. Temporal network methods could be used to analyze or model the opinion spreading behind changes in this type of sentiments.

Another interesting application of temporal networks to social media data could be network interventions the process of “using social network data to accelerate behavior change or improve organizational performance” [68]. A typical application would be to identify individuals or groups whose change of behavior can trigger a cascade of behavioral change. Another typical task is to find undesired grouping, splits or hierarchical dependencies in the social network of an organization, and then find a way to improve the situation by reorganizing. These social networks that network interventions rely on are not static and in the dynamic aspects of their nature lies much information that could be exploited by methods in this paper.

In this paper we have tried to argue that temporal network techniques are well suited for social media data. Compared to many other areas within the umbrella term of network science (including traditional social network analysis, focusing on [69,70]), social media data is accurately time tagged, and thus readily analyzable by temporal network methods. We have mentioned how temporal structures can change dynamics of information spreading events. Clearly without using temporal information prediction and modeling would be less precise, just as the static network structure can add precision compared to well-mixed models. Not only that, one could miss the most important ways to: estimate the spreading speed, decide who is the most important (or influential) individual, find efficient ways to mitigate or enhance spreading, etc. Temporal network approaches to social media are, we believe, an understudied area, so we expect much more research in this direction in the near future.

ACKNOWLEDGMENT

The author thanks Jari Saramäki and Naoki Masuda for discussions.

REFERENCES

- [1] P. Holme and J. Saramäki, “Temporal networks,” *Phys. Rep.*, vol. 529, pp. 97–125, 2012.
- [2] J.-P. Eckmann, E. Moses, and D. Sergi, “Entropy of dialogues creates coherent structures in e-mail traffic,” *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 14333–14337, 2004.
- [3] H. Ebel, H.-I. Mielsch, and S. Bornholdt, “Scale-free topology of e-mail networks,” *Phys. Rev. E* 66, art. no. 035103, 2002.

- [4] B. Viswanath, A. Mislove, M. Cha, and F. P. Gummadi, "On the evolution of user interaction in Facebook," *Proceedings of the 2nd ACM workshop on Online Social Networks*, pp. 37–42, 2009.
- [5] S. A. Golder, D. M. Wilkinson, and B. A. Huberman, "Rhythms of social interaction: Messaging within a massive online network," *Communities and Technologies 2007*, C. Steinfield, B. T. Pentland, M. Ackerman, N. Contractor (Eds.), Springer, London, pp. 41–66, 2007.
- [6] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, pp. 211–220, 2009.
- [7] P. Holme, C. R. Edling, and F. Liljeros, "Structure and time-evolution of an Internet dating community," *Soc. Networks*, vol. 26, pp. 155–174, 2004.
- [8] F. Karimi, V. C. Ramenzoni, and P. Holme, "Structural differences between open and direct communication in an online community," *Physica A*, vol. 414, pp. 263–273, 2014.
- [9] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, ACM, New York, NY, USA, pp. 1361–1370, 2010.
- [10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in: *Proc. 19th Int. WWW Conf.*, April 26–30, 2010, Raleigh, 2010.
- [11] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in Foursquare," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, Barcelona, 2011, pp. 570–573.
- [12] R. K. Pan, J. Saramäki, "Path lengths, correlations, and centrality in temporal networks," *Phys. Rev. E*, vol. 84, art. no. 016105, 2011.
- [13] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, "Temporal motifs in time-dependent networks," *J. Stat. Mech.*, art. no. P11005, 2011.
- [14] M. Karsai, M. Kivel, 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*, vol. 83, art. no. 025102, 2011.
- [15] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," *KDD '09 Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, pp. 497–506, 2009.
- [16] D. J. Bartholomew, *Stochastic Models for Social Processes*, Wiley, New York, 1967.
- [17] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford, 1991.
- [18] P. Clifford and A. W. Sudbury, "A model for spatial conflict," *Biometrika*, vol. 60, pp. 581–588, 1973.
- [19] T. M. Liggett, *Interacting Particle Systems*, New York, 1985.
- [20] M. Granovetter, "Threshold models of collective behavior," *Am. J. Sociol.*, vol. 83, pp. 1420–1443, 1978.
- [21] F. Karimi and P. Holme, "Threshold model of cascades in empirical temporal networks," *Physica A*, vol. 392, pp. 3476–3483, 2013.
- [22] T. Takaguchi, N. Masuda, and P. Holme, "Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics," *PLoS ONE*, vol. 8, art. no. e68629, 2013.
- [23] D. Kempe, J. Kleinberg, and A. Kumar, "Connectivity and inference problems for temporal networks," *Proc. STOC*, pp. 504–513, 2000.
- [24] P. Holme, "Network reachability of real-world contact sequences," *Phys. Rev. E*, vol. 71, art. no. 046119, 2005.
- [25] J. Moody, "The importance of relationship timing for diffusion," *Social Forces*, vol. 81, pp. 25–56, 2002.
- [26] K. L. Cooke and E. Halsey, "The shortest route through a network with time-dependent intermodal transit times," *J. Math. Anal. Appl.*, vol. 14, pp. 493–498, 1966.
- [27] L. Lamport, "Time, clocks, and the ordering of events in a distributed system," *Comm. ACM*, vol. 21, 558–565, 1978.
- [28] F. Mattern, "Virtual time and global states of distributed systems," in: *Proc. Worksh. Parallel Dist. Algo.*, 1989.
- [29] C. Lee, B. Nick, U. Brandes, and P. Cunningham, "Link prediction with social vector clocks," *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Disc. Data Mining*, pp. 784–792.
- [30] G. Miritello, R. Lara, M. Cebrian, and E. Moro, "Limited communication capacity unveils strategies for human interaction," *Sci. Rep.*, vol. 3, art. no. 1950, 2013.
- [31] P. Holme and F. Liljeros, "Birth and death of links control disease spreading in empirical contact networks," *Sci. Rep.*, vol. 4, art. no. 4999, 2014.
- [32] G. Kossinets, J. Kleinberg, and D. J. Watts, "The structure of information pathways in a social communication network," in *Proc. 14th ACM SIGKDD Intl. Conf. Knowledge Disc. Data Mining*, pp. 435–443, 2008.
- [33] J. Tang, M. Musolesi, C. Mascolo and V. Latora, "Temporal distance metrics for social network analysis," in *Proc. 2nd ACM SIGCOMM Worksh. Online Soc. Netw.*, 2009, p. 3.
- [34] V. Kostakos, "Temporal graphs," *Physica A*, vol. 388, pp. 1007–1023, 2009.
- [35] L. E. C. Rocha, F. Liljeros, and P. Holme, "Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts," *PLoS Comp. Biol.*, vol. 7, art. no. e1001109, 2011.
- [36] W. E. Leland, D. V. Wilson, "High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection," in *Proc. InfoCom91*, pp. 1360–1366, 1991.
- [37] P. Holme, "Network dynamics of ongoing social relationships," *Europhys. Lett.*, vol. 64, pp. 427–433, 2003.
- [38] H.-H. Jo, R. K. Pan, and K. Kaski, "Emergence of bursts and communities in evolving weighted networks," *PLoS ONE*, vol. 6, art. no. e22687, 2011.
- [39] T. Zhou, Z.-D. Zhao, Z. Yang, and C. Zhou, "Relative clock verifies endogenous bursts of human dynamics," *EPL*, vol. 97, art. no. 18006, 2012.
- [40] P. Bajardi, A. Barrat, F. Natale, L. Savini, and V. Colizza, "Dynamical patterns of cattle trade movements," *PLoS ONE*, vol. 6, e19869, 2011.
- [41] P. Holme, C.E. Edling, and F. Liljeros, "Structure and time-evolution of an Internet dating community," *Soc. Netw.*, vol. 26, pp. 155–174, 2004.
- [42] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, pp. 337–341, 2012.
- [43] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, art. no. 5, 2007.
- [44] P. Basaras, D. Katsaros, and L. Tassioulas, "Detecting influential spreaders in complex, dynamic networks," *Computer*, vol. 46, pp. 24–29, 2013.
- [45] N. Santoro, W. Quattrociocchi, P. Flocchini, A. Casteigts, and F. Amblard, "Time-varying graphs and social network analysis: temporal indicators and metric," in *Proc. 3rd AISB Soc. Netw. Multiagent Syst. Symp.*, pp. 32–38, 2011.
- [46] P. Grindrod, M. C. Parsons, D. J. Higham, and E. Estrada, "Communicability across evolving networks," *Phys. Rev. E*, vol. 81, art. no. 046120, 2011.
- [47] S. Lee, L. E. C. Rocha, F. Liljeros, and P. Holme P, "Exploiting temporal network structures of human interaction to effectively immunize populations," *PLoS ONE*, vol. 7, art. no. e36439, 2012.
- [48] M. Starnini, A. Machens, C. Cattuto, A. Barrat, and R. Pastor-Satorras, "Immunization strategies for epidemic processes in time-varying contact networks," *J. Theor. Biol.*, art. no. 337, pp. 89–100, 2013.
- [49] U. Alon, "Network motifs: Theory and experimental approaches," *Nature Rev. Genetics*, pp. 450–461, 2007.
- [50] D. Jurgens and T.-S. Lu, "Temporal Motifs Reveal the Dynamics of Editor Interactions in Wikipedia," *Proc. 6th Int. AAAI Conf. Weblogs Soc. Media*, 2012.
- [51] 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 Science*, vol. 1, art. no. 4, 2012.
- [52] P. Holme, "Epidemiologically optimal static networks from temporal network data," *PLoS Comput. Biol.*, vol. 9: art. no. e1003142, 2013.
- [53] E. Cheng, J. W. Grossman, M. J. Lipman, "Time-stamped graphs and their associated influence digraphs," *Disc. Appl. Math.*, vol. 128, pp. 317–335, 2013.
- [54] N. Masuda, K. Klemm, and V. M. Eguíluz, "Temporal networks: Slowing down diffusion by long lasting interactions," *Phys. Rev. Lett.* vol. 111, art. no. 188701, 2013.
- [55] R. Lambiotte, L. Tabourier, and J.-C. Delvenne, "Burstiness and spreading on temporal networks," *Eur. Phys. J. B*, vol. 86, art. no. 320, 2013.
- [56] H.-H. Jo, J. I. Perotti, K. Kaski, and J. Kertész, "Analytically solvable model of spreading dynamics with non-Poissonian processes", unpublished.
- [57] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality Prediction and Community Structure in Social Networks," *Sci. Rep.*, vol. 3, art. no. 2522, 2013.
- [58] J. Cheng, L. A. Adamic, P. A. Dow, J. Kleinberg, and J. Leskovec, "Can cascades be predicted?," *Proc. WWW Conf.*, Seoul, 2014.

- [59] P. Holme and M. Huss, "Understanding and exploiting information spreading and integrating information technologies," *J. Comput. Sci. Technol.*, vol. 26, pp. 829–836, 2011.
- [60] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PLoS ONE*, vol. 5, art. no. e12948, 2010.
- [61] M. E. J. Newman, "Ego-centered networks and the ripple effect," *Soc. Netw.*, vol. 25, pp. 83–95, 2003.
- [62] V. Lampsos, T. de Bie, and N. Cristianini, "Flu detector: Tracking epidemics on Twitter," *Proc. ECML PKDD*, pp. 599–602, 2010.
- [63] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic," *PLoS ONE*, vol. 6, art. no. e19467, 2011.
- [64] T. Bodnar and M. Salathé, "Validating models for disease detection using twitter," *Proceedings WWW '13*, pp. 699–702, 2013.
- [65] E. Coeira, "Social networks, social media, and social diseases," *Brit. Med. J.*, vol. 346, art. no. f3007, 2013.
- [66] M. Salathé and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," *PLoS Comput. Biol.*, vol. 7, art. no. e1002199, 2011.
- [67] M. Salathé, D. Q. Vu, S. Khandelwal, and D. R. Hunter, "The dynamics of health behavior sentiments on a large online social network," *EPJ Data Sci.*, vol. 2, art. no. 4, 2013.
- [68] T. W. Valente, "Network interventions," *Science*, vol. 337, pp. 49–53, 2012.
- [69] S. Wasserman and K. Faust, "Social Network Analysis: Methods and Applications," Cambridge University Press, Cambridge, 1994.
- [70] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323, pp. 892–895, 2009.