Aalto University
School of Science

Ville-Pekka Backlund

Threshold models of information spreading in empirical temporal networks

3.5.2013

Mat-2.4108 Independent research project in applied mathematics

Supervisor: Prof. Ahti Salo
Department of Mathematics and Systems Analysis

Instructor: Ph.D. Raj Kumar Pan
Department of Biomedical Engineering and Computational Science

Contents

1 Introduction

Humans encounter situations on a daily basis where the spreading of information plays a major role. A few examples are transfer of news, propagation of gossip and spreading of fads. Communication devices and the Internet are playing a growing role in the transfer of information. These media can coarsely be divided in two categories: to global broadcast media such as online news agencies which provide information to everyone interested, and to social media that leads to localized information spreading processes. The local processes take place within social networks and the interactions between individuals can occur e.g. via mobile phones or WWW based social media applications. For instance, significant part of the information leading to the revolution on 25 January 2011 in Egypt spread quickly via social media [1]. Information spreading in social networks can also be utilized knowingly for instance in word of mouth marketing [2]. Thus, the ability to understand information diffusion processes between individuals can provide warning signals of upcoming protests or help designing an efficient marketing plan for a new product.

There exists a wide theoretical background studying spreading processes in human societies. These diffusion models are often mean field models, i.e. the individual effects from multiple members of the system are approximated by a single averaged effect. Epidemic spreading models, such as the Susceptible-Infected-Recovered (SIR) -model [3], are an elementary example of them. The epidemic spreading models are applicable for studying diffusion processes where the participants are passive in the sense that they are incapable of determining their attendance to the process. For instance, individuals are usually not able to decide whether they become ill and consequently act as a spreader for the disease. Despite the simplicity of the SIR-models, they have been successfully utilized for example in vaccination planning. However, there exist a variety of spreading processes where humans can not be considered as passive. They either consciously or unconsciously assess the costs or benefits of acting as an adopter. For instance, it is beneficial to buy some communications device only if sufficiently many friends already have it, or a piece of news is a worth spreading if one hears it from many enough sources. A straightforward way to enable this kind of consciousness to the participants of the spreading processes leads to threshold models of collective behavior [4]. In threshold models stimuli from more than one source is required in order for advancement in the spreading.

All the spreading processes fundamentally depend on the framework they

function on. Consequently, the mean field approach provided by the simple diffusion models does not explain spreading in the real world but we need to take the underlying network of connections explicitly into account. The outcomes of the diffusion processes vary greatly if the underlying network is switched from e.g. regular lattices to random networks. Networks in the real world, whereas, are not completely regular nor random but have peculiar topological features and are called *complex networks*. For example, social networks include features such as broad degree distribution, short path lengths and community structure [5, 6]. These features have a distinctive effect on the dynamical processes occurring on the networks [7].

In addition to the topology of the underlying network for the diffusion process, we need to consider the possible dynamical properties of the network per se. This includes both the fact that the underlying network is changing over time and that the single links are active only at specific points of time. Rapid development of electronic means of communications seen during the last decades enables us to gather large amounts of data representing human-to-human interactions. This data can be represented as a *temporal network* in which the links can transmit something only when they are active, i.e. there is a call between two people. Studying the activity patterns of temporal networks reveal the bursty nature of human behavior, which means that there is plenty of activity in a short period of time and large gaps of low activity between them [8]. The temporality on the topology of the network creates a new kind of environment for all dynamical processes because they can advance only through active links. Hence, in order to study information propagation among humans as realistically as currently possible, we need to consider the temporality of the underlying complex network, as well as a threshold based approach.

Much work has already been carried out to study the two aforementioned factors separately, and here, the aim is to study them together. Previous work include examining the elementary dynamical processes in temporal networks, such as random walks [9] and SI-spreading. For instance, it has been shown that even though the topology of human communication networks advances spreading [10], the temporality of the system makes the SI-spreading slow [11]. Studies of spreading processes in social networks have also been conducted and they have been exploited for instance in finding the crucial nodes for the diffusion [12, 13]. Classification of the specific features enabled by the temporality that either hamper or advance the speed or reachability of the dissemination [14] is also possible via diffusion studies. Few empirical studies of information spreading in real Internet based social networks have also been conducted [15, 16]. Threshold model of spreading is examined ana-

lytically in random networks in [17] and a generalization of spreading models is presented and in simple cases solved in [18].

The aim of this special assignment is to study the threshold models of information spreading in empirical temporal networks, a question with recent activity [19,20]. Via our analysis we try to find common features of temporal networks that influence the spreading, and also to find differences between and characteristics within our four datasets describing human interactions. This work is organized as follows. In section 2 we introduce the crucial background information and definitions, and present the datasets used in the study. Next, we introduce the models used in this study and the results obtained from the simulations of the models in sections 3 and 4, respectively. The final section 5 discusses the results and their interpretation.

2 Background and data

In this section we introduce background information on temporal networks, diffusion processes on them and present the datasets used in this study. For the fundamental knowledge about complex networks one can refer to Newman's extensive introduction book [5] and Holme and Saramäki's review about temporal networks [21].

2.1 Temporal networks

In general, temporal networks can be represented as a set of events that define which edges of the network are momentarily active. Each single event e is a set of four members, $e \equiv (u, v, t, \delta t)$, where the members represent the transmitter, receiver, occurrence time and duration of the event, respectively. For example, in the case of a mobile phone call network, the network is presented as a list of calls (events) and each call contains the information about the caller, callee, the starting time and duration of the call. Often, for simplicity, the events are considered instantaneous, i.e. δt is ignored. Figure 1 presents an example of a temporal network.

The standard form of the representation of the temporal networks has the information about the direction of the events, thus the links of the network are *directed*. However, in the setting of information spreading, we need to reason whether to account the directionality. For instance in phone calls, once the link is active, the information can flow through it in both ways and thus the link is *bidirectional* or *undirected*, whereas, e.g. in email network one single event represents one email where the flow of the information is explicitly from the sender to the receiver and thus directed.

We are often interested in the total network structure that the temporal links span. In this *aggregated* version the temporality of the links is discarded by considering them to be constantly active.

2.2 Diffusion processes in static and temporal networks

In addition to the static topological properties of networks, we are often interested in some dynamic function taking place in them. A wide variety of dynamical processes has been studied in static networks, including important elementary processes such as *synchronization*, *percolation* and *random walks*.

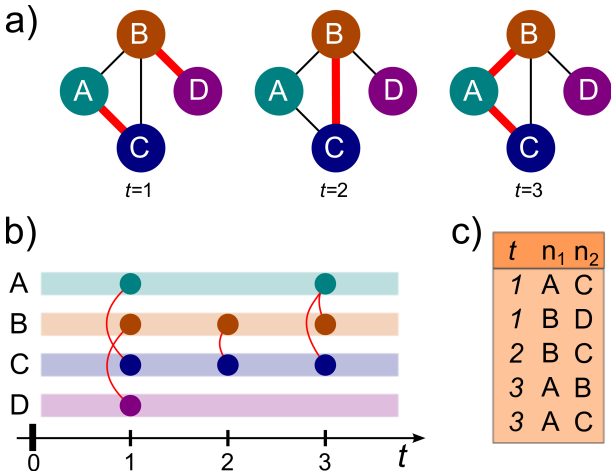

Figure 1: Three different representations of the same undirected temporal network with instantaneous events. In a) the red links are currently active, whereas the black links represent the links of the aggregated network. In b) and c) the network is presented as eventlist or event sequence. The columns in c) are respectively from left to right: time t, label of the sender node n_1 and label of the receiver node n_2. The format in c) is an efficient way of representing temporal networks for computers.

The goal in synchronization is to study the common dynamical evolution of coupled oscillators that interact via a network. A famous example is the synchronization of cricket chirping. Percolation studies the robustness of the networks, i.e. how the network is affected when nodes are removed from it. Percolation theory has important applications in studying for example power grid blackouts or Internet resiliency. Random walks are closely connected to searching in networks, and are the simplest example of a diffusion process in a network. In a random walk the network is explored by starting from one node, then advancing to one of its neighbors chosen at random and finally repeating the procedure. The PageRank algorithm developed by Google can be considered as an application of random walks. [7]

The next step in diffusion studies is modeling the epidemic spreading, which has applications for instance in epidemiology and in computer science when studying the spreading of computer viruses. The epidemic models are usually mean field models, i.e. the exact effect that each system's member has on a single specific member is replaced with a single representative average effect. They also assume a certain level of homogeneity in the interaction network, meaning that the egocentric network built for every system member looks exactly the same. Thus, they do not give correct answers when applied to

real life social networks with heterogeneous topological properties.

Even though the mean field models of epidemic spreading may not give correct results with real world networks, we can still utilize the concept of the epidemic dynamics on a node by node basis. This approach leads to an agent based model, and we can use them for instance with empirical networks. Moreover, the simple epidemic dynamics are usually the basis for studying complex collective behavior in social networks and they are used in simulating processes such as opinion formation or, especially, information spreading.

It is clear that the temporality in the underlying network creates fundamentally different environment for all dynamical processes, because they can advance only through active links. For example, it is shown that even with proper scaling of the time, the random walk exploration is slower in temporal networks than in the corresponding aggregated network [9]. On a more general level, empirical temporal networks include many properties that make the analytical treatment very complicated. Such features include the correlation between the event sequences of different links and temporal heterogeneities such as bursty behavior and daily patterns. Therefore dynamical processes in temporal networks are usually modeled using real empirical networks and agent based dynamics. Next we take a closer look at the compartmental model of epidemic spreading and see how it can be utilized as a basis for diffusion processes in temporal networks.

2.2.1 SIR-models of epidemic spreading

Though ultimately developed to describe epidemiological spreading, the Susceptible-Infected-Recovered (SIR) -models are significant and can be used as a base for more elaborate models [3]. The simplest model from the SIR-family has only two states, susceptible and infected, and is therefore called SI-model. In this model, when an individual becomes infected, it stays infected forever. The dynamics of the SI-model can be fully described with the help of few factors. Lets assume that the proportion of susceptible agents in the system at time t is s and the equivalent proportion for infected is i. Then we define transition rate β, which means that each individual has, on average, β contacts with other members of the system per unit time. At time t, the probability that each contact is between susceptible and infected individual and not between pairs S-S or I-I, is si. Now we can write differential equations for the rate of change between these groups:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\beta si \tag{1}$$

and

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \beta si. \tag{2}$$

Actually, since the expression $s + i = 1$ holds, the equations (1) and (2) are the same. This equation can then be solved analytically. More complicated models where the individuals can recover or become susceptible again, such as SIR and SIS, require the introduction of recovery rate. The rate equations for these cases can be derived following the same principles as above.

It is easy to see that the spreading processes from the SIR-family are crude simplifications of the reality and not very well suitable for information spreading. First, they assume that every individual of the system is able to contact directly with every other member of the system, which is clearly not true in real social networks. They also neglect the heterogeneous and correlated event sequences of real communication networks. Nevertheless, SIR-models are popular because of their simplicity and analytical solvability.

As stated, one possible way of utilizing SIR-dynamics and empirical underlying temporal network of connections is to make each node obey the dynamics individually. Thus, when simulating the spreading process, at each time step the nodes with an active event are checked in case of possible infection. New infection at a given time is possible only if susceptible node is in a contact with at least one infected neighbor. While the exact analytical manageability is lost, this kind of approach gives valuable information about the spreading processes in empirical network. For instance, the deterministic SI-model where the transition rate of the infection is 1 whenever there's a contact between infected and susceptible individual, gives the upper bound for all the spreading processes in empirical temporal networks.

An important step towards a more realistic information diffusion model in human societies was taken by Watts in [17]. He studies the occurrence of global cascades in random networks by first infecting a small initial population and then repeatedly applying the SI-model on a randomly chosen node. Instead of a transition rate, the new infection is governed by the fraction of infected neighbors around the specific node. If the fraction exceeds a predefined node specific constant, the node becomes infected. The Watts' model thus introduces local dependencies and also takes into account the underlying network of connections. These result in a great improvement compared to the simple epidemic models. The heterogeneities in the thresholds and in the degree distribution of the underlying network were seen to affect the possibility of a global cascade.

2.3 Datasets and data pre-processing

In our study, we use four different datasets describing human communication networks. The datasets and sources are:

CALL

The call network is constructed from the mobile phone call records of an European carrier. The network consists of ~ 5.8 million nodes and ~ 620 million calls. The data is gathered between 30.9.2009-31.3.2010 and the resolution is 1 second.

SMS

As above, but instead of phone calls, the network consists of SMS messages from the same time frame excluding Christmas Eve and New Year's Eve since the behavior on these days differs radically from all the rest (see appendix A). The sms network has ~ 3.1 million nodes and ~ 180 million sent text messages.

EMAIL

The email dataset is constructed from the logs of a given university's mail server. Only inter-institute emails are considered and some mass mailers are discarded. The duration of the dataset is $\sim$82 days with a resolution of 1 second. The real occurrence time of the emails concealed, yet the sequence of all the emails is kept unaltered. There are 2993 nodes in the email network that send in total $\sim 2 \cdot 10^5$ emails. [22]

CONFERENCE

The conference dataset is a collection of ongoing face-to-face conversations of the participants of the ACM Hypertext 2009 conference. The 113 tracked participants of the conference wore radio badges that monitored their proximity to other attendees. Thus, the data is a set of timestamps and IDs of pairs of participants who are having a conversation during instants t and $t + \Delta t$. The time resolution Δt of the the data is 20 seconds. Start of the recording took place at 8am on 29 of June and the overall duration of the eventlist is $\sim$2.5 days. [23, 24]

There are some distinctive features combining and separating the datasets. In the first three networks an electronic communication device is used and the sender and the receiver of the event can be physically far away from each other, whereas in the conference network, the participants are in the same physical space and the activation of new links (rewiring of the network) occurs through people leaving from old and joining to new conversation groups. Another difference separating the datasets is that the call and conference

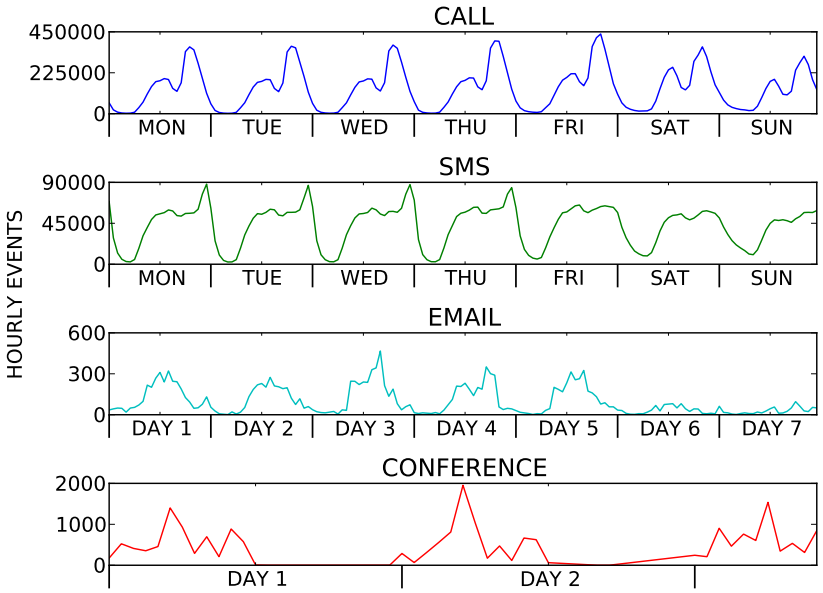

Figure 2: The number of hourly events during a certain time interval for all the four datasets. The conference dataset is represented as a whole whereas a one week representative sample is shown for the other datasets. In the call and sms data the day border is at 00:00, in the conference data at 08:00 and the real event times in the email data are unknown.

networks require activity from both the sender and the receiver in order for an event to happen (the call is answered) but in sms and email only the sender's activity matters. Once a link is active, the information in call and conference networks can flow to both directions, thus the links are considered bidirectional. In sms and email networks the direction of information flow is fundamentally from sender to receiver and thus the links are directed.

The call and sms networks are filtered in such a manner that only links which have at least one reciprocal event are considered. This is done to ensure that the networks represent better interactions between equal humans and not for example calls made in marketing purposes.

Since we are interested in the overall spreading within the networks and thus want to ensure that the spreading has the theoretical possibility of reaching every node of the network regardless of the starting node, we extract the largest connected component (LCC) from the raw data. The nodes which belong to the LCC are reclaimed from the aggregated version of the corresponding temporal network. The links in the aggregated network are considered bidirectional. Thus, all the nodes in the LCC are connected via a

(static) path, but because of the temporality and possible directionality of the links, some nodes might still be unreachable for the diffusion.

The events in all the datasets are assumed to be instantaneous, i.e. the durations of the events are neglected. For example, in reality each phone call event has a lifetime of δt, but we assume that the information can be passed anytime during the call, hence δt is ignored. This approximation does not have an effect on the overall spreading on the call network since multiple simultaneous events for a specific node are impossible. In sms and email networks, the events are instantaneous by nature. In conference dataset the possible overlapping nature of the events and their durations is built into the data.

The four datasets have different statistical properties. For example, in the sense of both topological and temporal characteristics, the conference network is very dense while the sms network is very sparse. Figure 2 shows the number of hourly events for each of the datasets for a certain period of time and reveals some characteristic properties of temporal networks. For instance, the call and the sms networks have clear daily and weekly patterns: the highest activity is reached late in the evening and there's less activity on the weekends. Siesta spent during the early afternoon explains the peculiar drop in the number of hourly calls. Naturally, the activity in the conference data happens during daytime. The basic properties of the networks are presented in Table 1.

Table 1: Statistical properties of empirical networks used in the study. Figures from the aggregated counterparts of the temporal networks are flagged with (**a**). The properties are: number of nodes N, number of events $N(e)$, number of links m, average degree $\langle k \rangle$, average clustering coefficient $\langle c \rangle$, data period length T, data time resolution Δt and whether the links are directed ($\rightarrow$) or bidirectional ($\leftrightarrow$).

	N	$N(e)$	m (**a**)	$\langle k \rangle$ (**a**)	$\langle c \rangle$ (**a**)	T	Δt	link
CALL	$5.8{\cdot}10^6$	$620{\cdot}10^6$	$15{\cdot}10^6$	5.0	0.23	180 d	1 s	$\leftrightarrow$
SMS	$3.1{\cdot}10^6$	$180{\cdot}10^6$	$5.7{\cdot}10^6$	3.7	0.10	176 d	1 s	$\rightarrow$
EMAIL	2993	$2.0{\cdot}10^5$	21736	14.5	0.21	82 d	1 s	$\rightarrow$
CONFERENCE	113	20818	2196	38.9	0.54	2.5 d	20 s	$\leftrightarrow$

3 Methods

3.1 Threshold spreading models

The fundamental idea in our models is to include the concept of thresholds to temporal networks, thus providing a good framework for studying information spreading in real human interaction networks. Our model can be seen as a temporal extension to the Watts' cascade model [17]. From now on, the lexicons of information and infection spreading are used interchangeably and thus a node becomes infected when it receives a piece of information. At the initial state every member of the network is susceptible and then one initial spreader gets chosen at random. The information has then the possibility to advance to other nodes via the events the initial spreader takes part in. Whenever a susceptible node is in contact with an infected one it calculates the fraction of infected neighbors around itself. This fraction is then compared with a predefined threshold which defines what level of neighborhood infection is needed in order for the information to diffuse. In a sense, each node must make a *binary decision with externalities*, i.e. whether to accept an information or not, according to the knowledge about nodes neighbors' state [25].

Each node obeys SI-dynamics where the infection rate is replaced with the fraction. Thus, if the predefined threshold is exceeded, the diffusion occurs with certainty. Whenever a new node accepts the information (becomes infected), it acts as a possible spreader in the events taking place after the infection. Once infected, the node stays infected forever. Nevertheless, we want to include a component in the model which decreases the importance of old events. In the context of information spreading between humans, it is natural to assume that eventually people forget the interactions. This behavior is achieved by defining a time range τ which rules how far back in history the events are noteworthy. In other words, the time range τ represents the memory of the node. Thus, the fraction of infected neighbors around one node at time t is calculated as the sum of infected neighbors which have been in contact with the specific node during interval $[t - \tau, t]$ divided by the degree of the node. Note that the degree is the static value calculated from the aggregated network and not a sum of all the neighbors, infected or susceptible, that have been in contact with a node during given interval. Since the underlying networks do not evolve by definition, i.e. no new nodes or links enter the system, the static degree is the correct measure describing the connectance of a node in the context of information diffusion between

humans.

Model I: More formally, the diffusion is controlled by the fraction $\phi(i, t|\tau)$ which is the measure of node specific infectious events for node i divided by the degree d_i of the node. An event is called infectious for node i if it is a contact between infected node and node i, and the possible directionality of the event is towards node i. The node specificity means that possible multiple contacts coming from one node are counted as single. Thus, the fraction can be defined as

$$\phi(i, t|\tau) := \frac{1}{d_i} \sum_{j(\neq i)} \chi(j, t'), \qquad t' \in [t - \tau,\ t], \tag{I}$$

where the indicator function χ is 1 if node j has had at least one infectious contact with node i during the interval $[t - \tau,\ t]$ and 0 otherwise. The summation is over all the other nodes of the network except node i. The diffusion spreads to node i if $\phi(i, t) \geq f$, where f is a predefined constant and common for all the nodes.

Model II: A simple variation to the first model can be achieved by leaving out the scaling with the degree. The result is a hard threshold model which highlights the absolute connectivity of a network. In this case the value to study, $k(i, t|\tau)$, is the number of node specific infectious events for node i. Formally it can be written as

$$k(i, t|\tau) := \sum_{j(\neq i)} \chi(j, t'), \qquad t' \in [t - \tau,\ t]. \tag{II}$$

The diffusion spreads if $k(i, t) \geq K$, where K is again a predefined constant and common for all nodes. In practice, only small values of K can enable cascades of notable sizes, because already with $K = 2$ diffusion can advance only via mutually overlapping triangles. Higher values of K require even denser networks. With hard threshold model the value of K must of be taken into account when choosing the initial spreader(s). For example, with $K = 2$ we must infect the both participants of a one random event.

There is a connection between these models and elementary SI-spreading. Model (II) with parameter $K = 1$ equals exactly to SI-model with infection rate 1. Also, model (I) with a small enough threshold values equals to model (II) with $K = 1$. The fractional threshold can thus be interpreted as a limiter which governs the type of spreading process the nodes obey according to their degree: they can either obey deterministic SI-dynamics or threshold dynamics. For example, nodes with small degree always have very high values of

$\phi(i, t|\tau)$, even with short memory and few infectious hits, thus they get easily infected and obey SI-model with infection rate 1. The threshold dynamics explicitly activate for nodes with degrees $d_i > f^{-1}$, because then they need infectious hits from more than one source in order to become infected.

Model III: The first two models are deterministic and only dependent on the initial spreader and the constant. We define also a stochastic variant of the first. Instead of comparing the figure $\phi(i, t|\tau)$ to the predefined constant, we use the value $\phi(i, t|\tau)$ directly as the probability of infection. Thus a node i gets infected if

$$\phi(i, t|\tau) \geq U, \tag{III}$$

where U is an uniform random number from interval $[0, 1]$.

As in the first two models, in the third model a node is given the opportunity of becoming infected after each received infectious event. A significant difference of this stochastic model and the first two is, that in the stochastic model, the bursts of a single link are meaningful. In the first two models only the latest hit from each of node's links is counted, assuming it fits within the memory, and therefore the bursts in a single link do not increase the likelihood of infection. In the stochastic model the bursts have a positive impact on the likelihood of infection through the increased number of opportunities of becoming infected.

Figure 3 shows an example of the spreading process for all the three models.

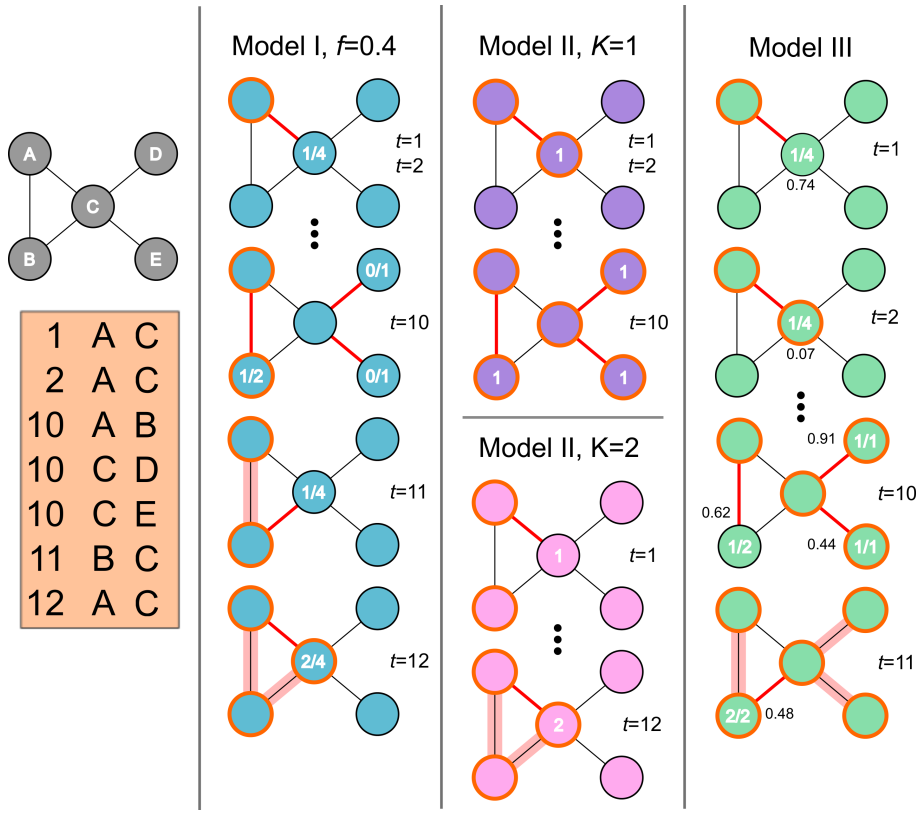

Figure 3: An example of the diffusion spreading for all the three models in a temporal network with bidirectional links and memory $\tau = 5$. Infected nodes are encircled with orange. The red links are currently active and infectious hits that are still within the memory of a node are emphasized with light red background. The calculated values of ϕ or k are seen within the corresponding node. In model II with $K = 2$ the nodes A and B are initially infected. With model III the random numbers used in the checking of the rule are seen close to the nodes in question.

3.2 Shuffling methods

The effect of structural properties on the spreading is known. Since the aim is to study how the temporal correlations affect the system, one needs a methodology to selectively destroy a few specific correlations while keeping all other temporal and structural characteristics intact. This is achieved by comparing the diffusion with the original event sequence to different null models. The null models destroy some and preserve other features of the networks, thus enabling us to deduce the important ones. The null models used and their effect to the network are introduced underneath. An illustration presenting the effects of the different shuffling schemes is seen in Figure 4.

RANDOM TIME SHUFFLE (RT)

The time stamps of all the events are randomly shuffled. This destroys all the temporal correlations of events between links and nodes but preserves the global activity such as daily patterns. Also, all the static properties of the nodes and network, such as degrees, are preserved.

EQUAL WEIGHT LINK SEQUENCE BIN SHUFFLE (EQW)

The links are sorted according to the total number of events and binned such that each bin has at least two links. Then, entire event sequences are randomly shuffled between the links in a certain bin. The shuffling destroys the link-link correlations of events, but preserves the burstiness of each link and static properties of the network. Event sequences of links connected to a same node have the possibility to be exchanged with each other. [14]

INTER LINK EVENT SEQUENCE SWITCH (IL)

The event sequence of each link is switched onwards with periodic boundary condition by a random integer in the interval $[0, T]$, where T is the occurrence time of the last event in the sequence. Hence we can destroy the link-link correlations while keeping the event sequences of the links as original as possible.

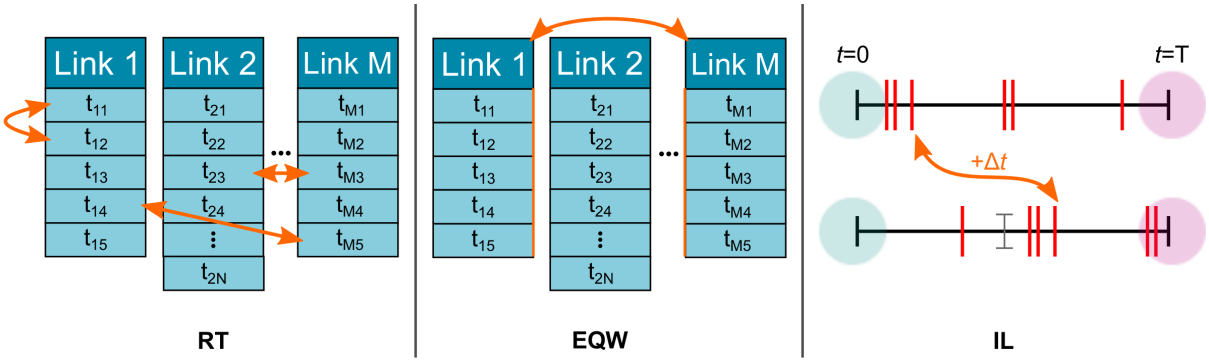

Figure 4: Illustration of the effects of the different shuffling methods. The first two panels list the events of a specific link ordered by time. The third panel illustrates the effect of the IL-shuffle on one link. The red bars are single events on a link connecting two nodes and at the same time representing time interval $[0, T]$.

4 Results

The models are simulated numerically by running the spreading process with the original event list and all of the null models. The infection is started from an event randomly chosen from the first p percents of the events and ran till the end of the event list. The program stores the fraction of the size of the infected cluster versus the number of active nodes in the specific run. Number of simulations is 10^3 for the call and sms datasets and 10^4 for the smaller email and conference datasets.

The initial infected node and the starting event for models (I) and (III) are chosen so that we first check which nodes are active during the first p percents of events. Then we choose a random node from the set of active nodes and finally a random event where this chosen node is active. This results in equal probability for a node and an event that are active within the first p percents of the events to be chosen as the initial spreader. For the case (II), we must infect both of the participants of the chosen event in order to be able to get any spreading with $K = 2$. The value $p = 0.15$ was chosen because large enough percentage of nodes are active within the first p percents of events and the infection has still at least $1 - p = 0.85$ percents of events left to spread. In Figure 5 we see the fraction of active nodes during the first p percents of events versus the total number of nodes in the network for all the four datasets.

In Figures 6, 7 and 8 we see the average fraction of infected nodes at the end of the eventlist ($\langle I(t_{\text{end}})/N \rangle$) as a function of memory τ for the models (I), (II) and (III), respectively. The $I(t)$ represents the number of infected nodes at time t. The possible threshold values in the figures are chosen so that they reveal the most interesting phenomena. The standard error for each datapoint is less than $1.1 \cdot 10^{-2}$ and thus confidence intervals are not drawn for clarity.

For the call network we see that the event list with random shuffling infects the greatest part of the network (6a). The original event sequence infects considerably lower amount of nodes. The explanation for the performance of the RT-shuffle is evident when we study the effects the shuffling has on the events: it spreads the events of the links more evenly by destroying all the correlation between events and links, therefore increasing the likelihood of a node obtaining infectious hits from multiple unique nodes. On the other hand, the destruction of link-link correlations in the null model has a negative effect on the spreading, but clearly the great positive effect of the even spreading of the events overcomes this issue. The EQW and IL -shuffles behave in

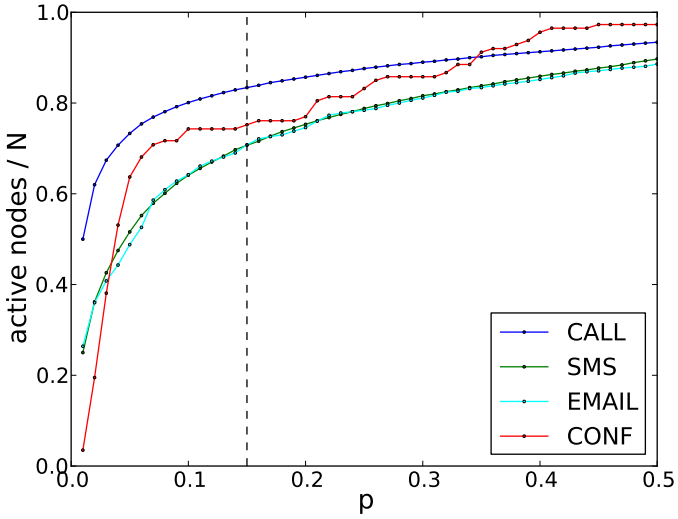

Figure 5: The fraction of active nodes during the first p percents of events. The chosen value of $p = 0.15$ is emphasized with the dashed line. In the call and sms networks the value of $p = 0.15$ corresponds to $\sim$27 days, in the email network $\sim$12 days and in the conference network $\sim$9 hours.

quite similar fashion, both infecting less than the original. This is due to the destruction of link-link correlations these shuffling methods have. Thus our results demonstrate that link-link correlations exist in call network and they enhance the diffusion. The behavior as a function of memory τ corresponds well to what we would expect from reality: the memory starts to have a significant increase in the spreading when it last longer than $\sim$0.5 hours and the increase in the memory becomes negligible after approximately one week.

The call network with second model (7a) and with $K = 1$ reaches almost a full infection. As stated, with $K = 1$ this model corresponds to the deterministic SI-model, and thus τ is meaningless. The reason why the RT-shuffle infects a bit more than the other shuffling methods, even though the behavior is now independent of memory, is that it enables small subcomponents which are active only in a reasonably short time interval compared to the total duration of the event sequence get more easily infected. With $K = 2$, the spreading is zero as we might expect, since spreading can not proceed to nodes which are connected to the infected component of the network through just one link. The results of the stochastic model (III) (8a) agree well with the results of the model (I). The dramatic drop in the absolute value of the size of the infected component between the stochastic and the fractional models results

from the fact that high degree nodes are very unlikely to get infected, and especially that the low degree nodes do not get infected with certainty.

The temporal and topological sparsity of the sms network hampers the spreading of infection, best seen in the zero infection in the stochastic model (8b) and in the τ-independence for other models (6b,7b). With small enough values of f the infection spreads (6b), but then the model corresponds to the model (II) with $K = 1$ (7b) and to the deterministic SI. Therefore, if the network is sparse in the sense of topology and the correlation between links is low, it is very unlikely to have a sufficiently infected neighborhood and the whole concept of threshold spreading becomes arguable.

The results of the conference network with $\tau > 20$s and with the models (I) and (III) resemble in outline to those of the call network. A notable occurrence with these models is the behavior of the diffusion with memory less than the resolution of the data, $\tau < 20$s. In this region the memory of the nodes is erased during each regrouping of the people (rewiring of the network). It is seen that the original order of events is the best for the spreading, which makes intuitively sense, because breaking the natural order of conversation groups is likely to hamper the information flow. The average degree of nodes in the conference network is high, resulting in a network dense enough for the model (II) with $K = 2$ (7c) to have non-zero outcome. In this curve we also see the strong phenomenon of the original event sequence performing best for the memory lengths less than 20 seconds. For the stochastic model (8c), due to increased amount of possibilities of becoming infected, the RT-shuffled data is slightly better than the original even in the short memory region but because of the small size of the network, the absolute difference is only a few infected people. Again, the behavior of the spreading as a function of the memory of the nodes τ agrees well with the real world.

The structure of the email network causes the EQW-shuffle behave in a manner that produces anomalies in the results. This is seen with the model (I), where the EQW-shuffle dominates the RT-shuffle with small threshold values but succumbs to it with high values (6d). This phenomenon is emphasized in Figure 9, where the difference between EQW and RT -shuffles is plotted against the threshold f with a few different values of the memory τ. With small threshold values the difference is positive, declining to negative values as the threshold is increased and finally settling to zero at the region where there is no spreading. This behavior can be explained by studying the exact links that are switched in the EQW-shuffle. In the email network, the links that have high number of events are likely to be connected to the same node. This results in that the EQW-shuffling preserves the temporal sequences of

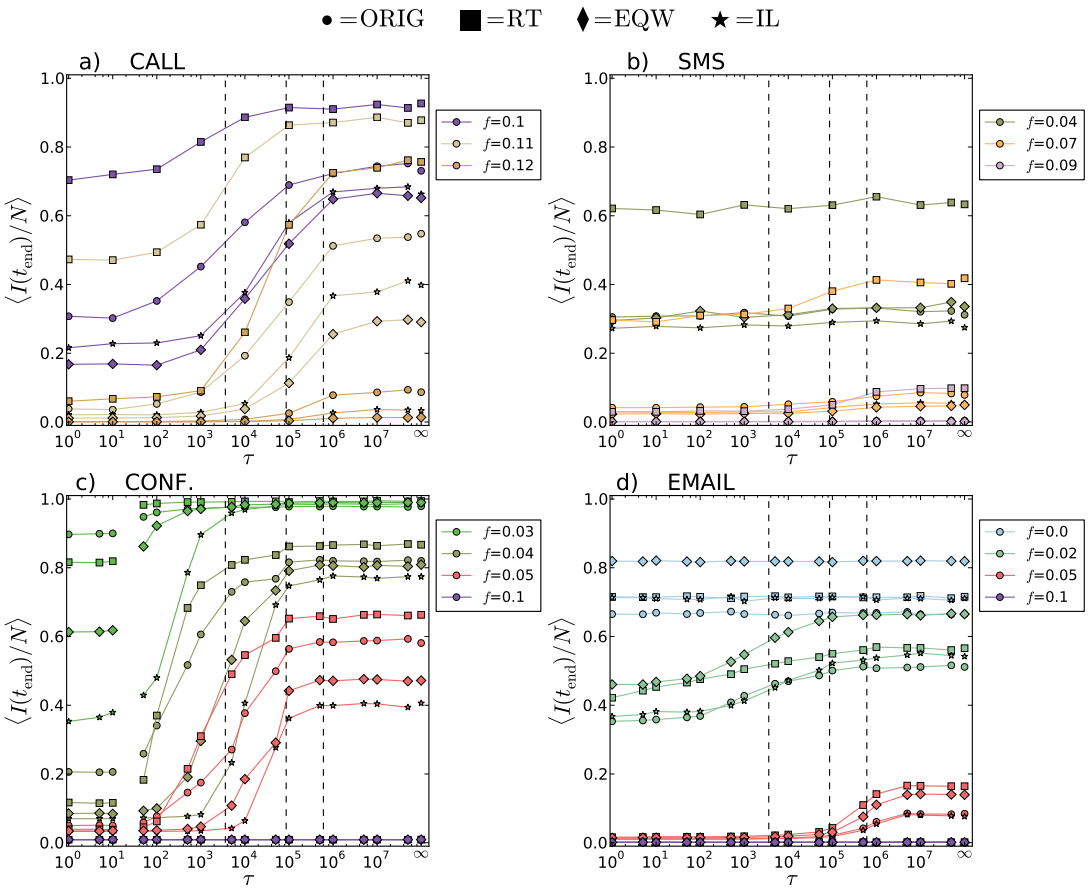

Figure 6: Model (I): numerical results of the average fraction of infected nodes at the end of the eventlist as a function of memory τ. Number of simulations is 10^3 for call and sms datasets and 10^4 for conference and email datasets. The gap in the curves of the conference dataset indicates the point where the data resolution (20s) is crossed. The slight slope in the conference dataset with IL-shuffle and memory less than 20s is due to the fact that the resolution of the switch in the IL-shuffle is 1 second.

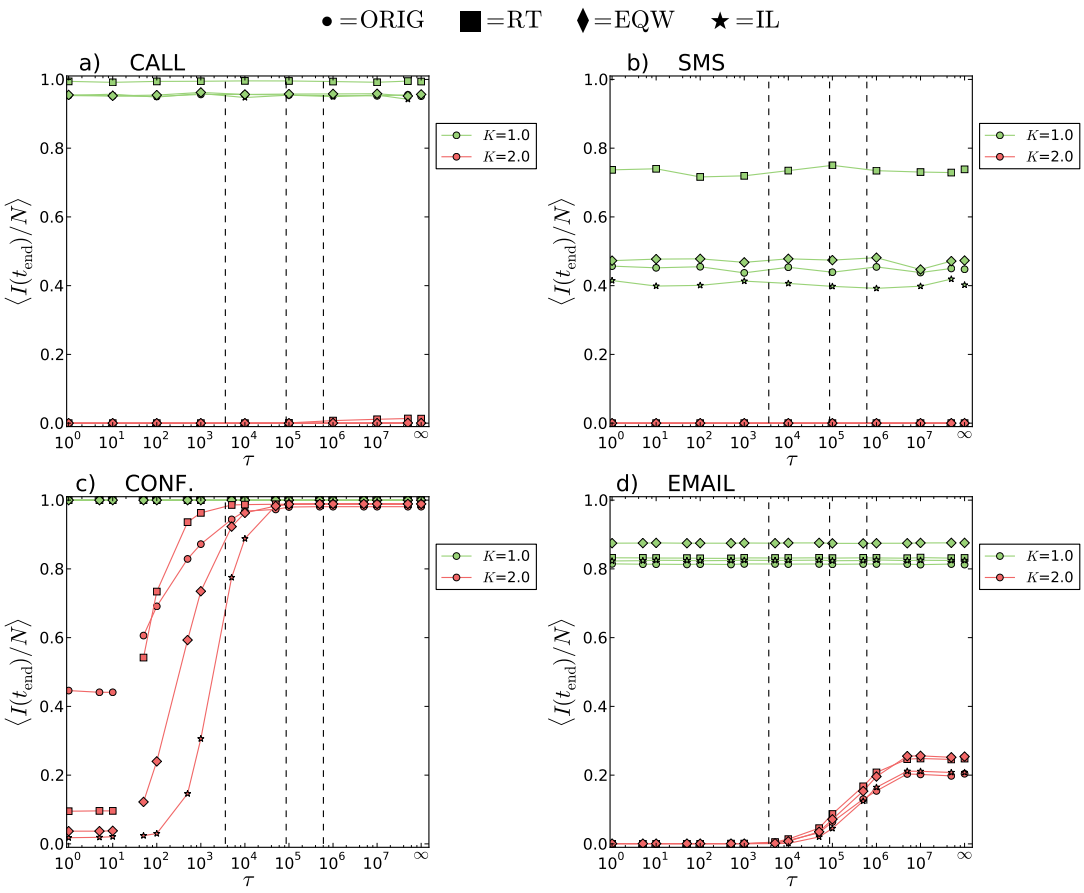

Figure 7: Model (II): numerical results of the average fraction of infected nodes at the end of the eventlist as a function of memory τ. Number of simulations is 10^3 for call and sms datasets and 10^4 for conference and email datasets. The gap in the curves of the conference dataset indicates the point where the data resolution (20s) is crossed.

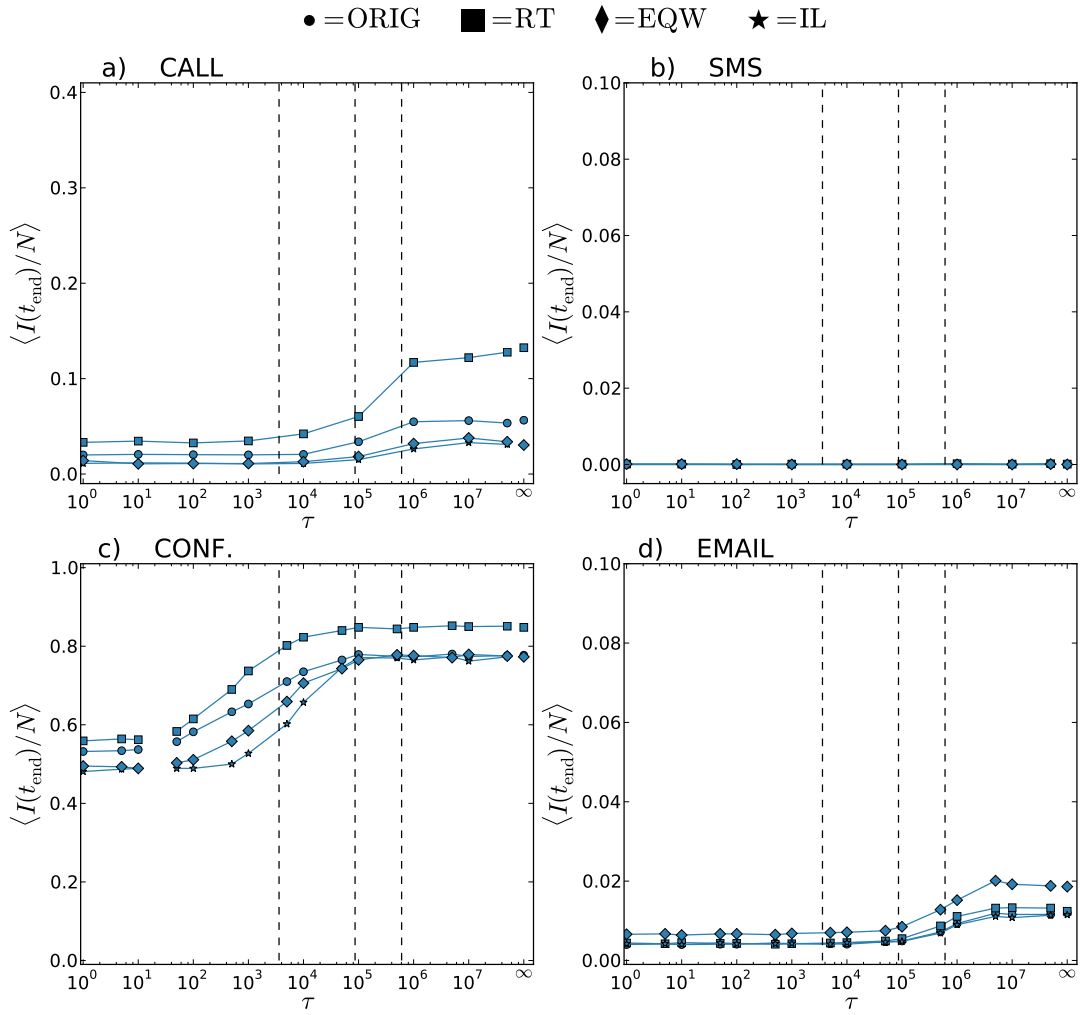

Figure 8: Model (III): numerical results of the average fraction of infected nodes at the end of the eventlist as a function of memory τ. Number of simulations is 10^3 for call and sms datasets and 10^4 for conference and email datasets. The gap in the curves of the conference dataset indicates the point where the data resolution (20s) is crossed. Note the different scales of y-axis.

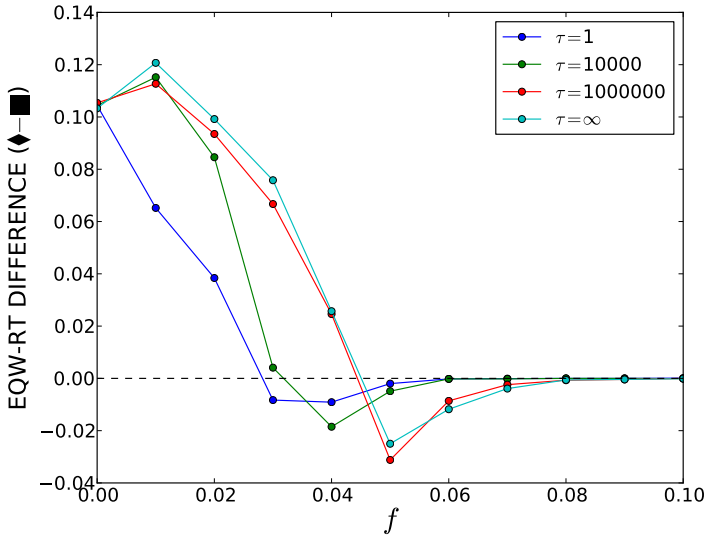

Figure 9: Difference between the EQW and RT -shuffles in the email data with model (I) as a function of f.

the backbone of the network, i.e. the nodes with links with high number of events, almost intact. In contrast, in the call network the EQW-shuffle switches the event sequences between totally separate nodes. In addition of preserving the backbone of the network, the EQW-shuffle functions essentially like the random time shuffle for the nodes with links that have small number of events, thus enhancing the diffusion. When the threshold is increased and thus the degree of a node needed for threshold dynamics decreases, the larger probability of hits from multiple sources provided by the RT-shuffle results in a larger infected component. Hypothesis of the preserved backbone indicates that even with large thresholds there should be a difference in what kind of nodes and components get infected with the RT and EQW -shuffles. For instance, we might expect that there is a difference in the diameter and in the degree distribution of the infected component, but to be able to explicitly state this, we would need to do further studies.

Another interesting outcome in the email dataset is that the IL-shuffle is never worse than the original sequence. Initially one could reason that because of email forwarding, there is a high amount of link-link correlation in the network. However, these kind of directed trains of correlated events do not result in better diffusion for the threshold models because the directionality of the correlated events do not point towards the same node. In the results obtained from simulations using undirected links we saw that the

original temporal network performs significantly better than the IL-shuffled version. This is due to message forwarding and especially multi-messaging (multiple recipients for a single message) which, if bidirectional information flow is allowed, are perfect for threshold spreading. In addition, the email network is dense enough for the model (II) to spread with $K = 2$ (7d) and again, the time scales the models output are reasonable and nicely in relation with the time scales of the other datasets.

5 Discussion

The aim of this special assignment was to study the threshold models of information spreading in empirical temporal networks describing human interactions. We defined three different threshold models and studied how social behavior has an effect on the spreading dynamics with four different datasets through computer simulations. The first two datasets were constructed from mobile phone calls and mobile phone short text messages from one European carrier. The third dataset was built from inter-institutional emails within a given university. The fourth dataset was gathered during a conference with the help of ID badges that recorded the proximity of other such badges. We used different shuffling methods on the original event sequences of the datasets to reveal which temporal characteristics have an effect on the diffusion.

From the results obtained, we want to point out three key findings. First, when combining the results from all the datasets and all the models, we can state that the hypothesis of bursty event patterns slowing down spreading [11] holds even for our threshold based models. This is seen in the better performance of the random time shuffled event sequence compared to the original sequence (excluding the anomalies in email data). Karimi et al., in addition of seeing the burstiness slowing down the spreading with their version of fractional threshold model, state that in a version of the hard threshold model the burstiness actually facilitates spreading [19]. The latter phenomenon was also seen by Takaguchi et al. in [20]. The difference to our results ensues from the fact that their models count all the hits regardless of the source whereas our models require the node specificity of them. This difference in the models results in different significance for the bursts in a one link. Note especially that our finding of the bursty patterns hampering the diffusion holds also in our stochastic version of the fractional threshold model, which values the bursts in a one link.

The second key finding was that the temporal correlations in the multiple links of a node was seen to facilitate the diffusion. This phenomenon is seen in the worse performance of the IL-shuffled event sequence than the original event sequence in suitable datasets (call, conference). The IL-shuffle destroys the link-link correlations by switching the event sequence of every link onwards with periodic boundary conditions by a random number.

The third important outcome from our threshold models was that the behavior as a function of the memory τ of the nodes match reality well. For example, memory more than one week does not promote the diffusion of in-

formation in the call network. Also, some features unique to specific datasets were revealed. Maybe the most interesting one of these was that the original event sequence, corresponding to the natural regrouping of participants, was best for information flow in the conference data with memory less than the data resolution. With so short memory, the participants don't remember the discussion that took place in the old conversation group when entering a new one. Thus they need to either carry the information with them to the new conversations or "seek" a conversation group where the members are already infected.

Further work with our threshold models of information spreading might include deeper examining of the infected clusters and the properties of the infected nodes. This approach might help us to see if there are some specific nodes that are important or even crucial for the information diffusion. We might also want to define heterogeneous thresholds for the nodes based on their degree, so that the effect that small degree nodes always obey deterministic SI-dynamics could be circumvented.

Acknowledgments

The numerical calculations presented in this assignment were performed using computer resources within the Aalto University School of Science "Science-IT" project.

References

[1] A. Kavanaugh, S. Yang, S. Sheetz, L. T. Li, and E. Fox, "Microblogging in crisis situations: Mass protests in Iran, Tunisia, Egypt," in *Workshop on Transnational Human-Computer Interaction, CHI*, 2011.

[2] J. Brown, A. J. Broderick, and N. Lee, "Word of mouth communication within online communities: Conceptualizing the online social network," *Journal of interactive marketing*, vol. 21, no. 3, pp. 2–20, 2007.

[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1st ed., 1992.

[4] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[5] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[6] M. E. J. Newman, "The Structure and Function of Complex Networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[7] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. New York: Cambridge University Press, 1st ed., 2008.

[8] A.-L. Barabasi, *Bursts: The Hidden Pattern Behind Everything We Do*. Dutton Books, 1st ed., 2010.

[9] M. Starnini, A. Baronchelli, A. Barrat, and R. Pastor-Satorras, "Random walks on temporal networks," *Physical Review E*, vol. 85, p. 056115, 2012.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[11] 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," *Physical Review E*, vol. 83, no. 2, p. 025102, 2011.

[12] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, no. 11, pp. 888–893, 2010.

[13] J. B. Holthoefer and Y. Moreno, "Absence of influential spreaders in rumor dynamics," *Physical Review E*, vol. 85, no. 2, p. 026116, 2012.

[14] M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, and M. Karsai, "Multiscale analysis of spreading in a large communication network," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2012, no. 03, 2012.

[15] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.

[16] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, pp. 519–528, ACM, 2012.

[17] D. J. Watts, "A simple model of global cascades on random networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 5766–5771, 2002.

[18] P. S. Dodds and D. J. Watts, "Universal Behavior in a Generalized Model of Contagion," *Physical Review Letters*, vol. 92, no. 21, p. 218701, 2004.

[19] F. Karimi and P. Holme, "Threshold model of cascades in temporal networks," *arXiv preprint arXiv:1207.1206*, 2012.

[20] T. Takaguchi, N. Masuda, and P. Holme, "Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics," *arXiv preprint arXiv:1206.2097*, 2012.

[21] P. Holme and J. Saramäki, "Temporal networks," *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[22] J.-P. Eckmann, E. Moses, and D. Sergi, "Entropy of dialogues creates coherent structures in e-mail traffic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 40, pp. 14333–14337, 2004.

[23] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *Journal of Theoretical Biology*, vol. 271, no. 1, pp. 166–180, 2011.

[24] SocioPatterns, "http://www.sociopatterns.org/."

[25] T. C. Schelling, "Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices with Externalities," *The Journal of Conflict Resolution*, vol. 17, no. 3, pp. 381–428, 1973.

A Preprocessing of the SMS data

When we looked at the number of daily events from the sms data, we noticed that the number of events is significantly different on four specific days than on all the rest. This phenomenon can be seen in Figure A1. These four days are actually around Christmas Eve and New Year's Eve when people send a large amount of short text messages to their acquaintances.

Since we are interested in studying the spreading processes under normal conditions, we remove these fours days from the dataset. The number of daily events from the processed data is also seen in the Figure A1.

We did not see this kind of drastic change in activity during the holiday season in the call data and therefore no days were removed from it.

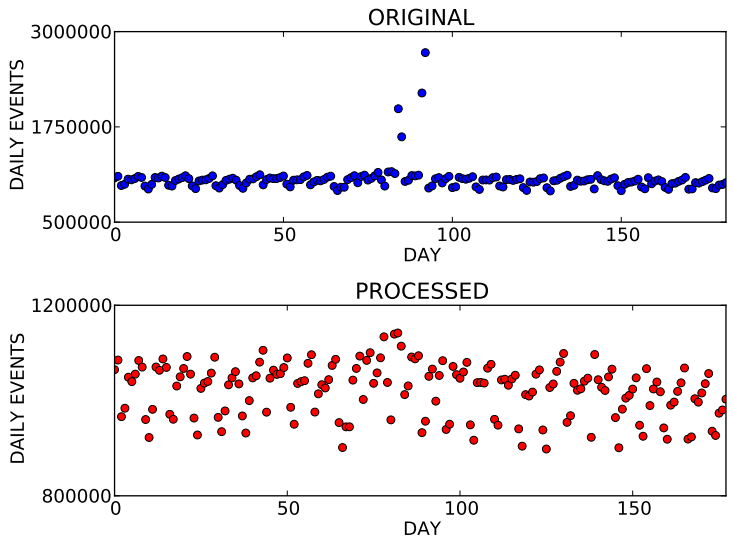

Figure A1: The number of daily events of the original and processed sms data.