# How to detect causality effects on large dynamical communication networks: a case study

Lionel Tabourier
LIP6, Université Pierre et Marie Curie,
Paris, France
lionel.tabourier@lip6.fr

Alina Stoica
Orange Labs R&D,
Issy-les-Moulineaux, France

Fernando Peruani
Laboratoire J.A. Dieudonné,
Université de Nice Sophia-Antipolis,
Nice, France
peruani@unice.fr

***Abstract*—Here we propose a set of dynamical measures to detect causality effects on communication datasets. Using appropriate comparison models, we are able to enumerate patterns containing causality relationships. This approach is illustrated on a large cellphone call dataset: we show that specific patterns such as short chain-like trees and directed loops are more frequent in real networks than in comparison models at short time scales. We argue that these patterns - which involve a node and its close neighborhood - constitute indirect evidence of active spreading of information only at a local level. This suggests that mobile phone networks are used almost exclusively to communicate information to a closed group of individuals. Furthermore, our study reveals that the bursty activity of the callers promotes larger patterns at small time scales.**

## I. Introduction

The analysis of dynamical features in large interaction networks has recently focused much activity [Bar05], [OS$^+$07], [MS$^+$08], [IM09], [ZT$^+$10], [KK$^+$11], [SQ$^+$10], [MML10], [MGV11] to overcome the limits of usual static representations. For example, in a social network, if A discusses with B and then B with C, that means that information can flow from A to C but not from C to A, so that a classical static description of data cannot account for the possible chronological constraints on the information spreading. In addition, the data is very often such that its representation would require the use of directed edges. For instance in phone communication data, each event involves a caller and a callee whose roles are asymmetric as the first one intends to call the second, which means that we should distinguish $A \to B$ from $B \to A$.

A wide-spread point of view consists in processing dynamical network data as if it were a succession of static pictures. Yet this does not give a comprehensive understanding of the features of the dynamical network, as it misses the chronological order of the interactions occurring between two successive snapshots. In other cases, the dynamical aspect is taken into account but the directedness of the data is neglected — even in communication datasets — making the analysis blind to the intention of the agents. Therefore, there is a great need for intrinsically dynamical measures that take into account features which cannot be seen using a sequence of static and/or undirected network representation.

The limitation discussed above are particularly relevant when it comes to the study of diffusion of information, such as rumor spreading. Information spreading is strongly affected by the directedness of the underlying network and the temporal ordering of the communications. Despite the ubiquity of the problem, practical tools to observe and measure such phenomena remain scarce. Since tracking an item that propagates over a real communication network is a very difficult task, there are only few works providing measures or estimations of this kind [PV01], [GG$^+$04], [LM$^+$07], [CR09], [SGL10], [FCL11]. Facing the absence of information on the spreading items, some studies use simulations and models to get clues about the influence of the structure and dynamics of interactions on the spreading processes [OS$^+$07], [IM09], [KK$^+$11], [MGV11].

We propose in this paper a set of dynamical tools that take into account the timescales and the directedness associated to the emerging patterns. We implement these tools in the context of a mobile phone call dynamical dataset, as these mobile networks provide an adequate ground for such studies [OS$^+$07], [SQ$^+$10], [ZT$^+$10], [KK$^+$11]. Our goal is to detect traces indicating events causality-correlated. *Causality* refers here to the fact that the existence of an event (a phone call) triggers the existence of another one at a certain point in time. We assume that such correlations are due to an exchange of an information item among the users. So the detection of correlated events inform us about the characteristics of the spreading process. This paper is organized as follows: section II consists in a description of the dataset and the models which will be used throughout this paper. Section III is devoted to the description of the measured patterns whose appearance frequency informs us about how information spreads across the communication dynamical network. Finally Section IV proposes some promising perspectives to this work, regarding in particular how these tools can be generalized.

## II. Database and comparison models

The measures we propose can be applied in any context where the data can be represented as “temporal” and directed events of the form:

*source (s) - destination (d) - timestamp (t)*

That is usually the case for dynamical communication networks such as phone calls, instant messaging, e-mail ex-

978-1-4673-0298-2/12/$31.00 © 2012 IEEE

changes etc. In the following, we will call "static network" the picture obtained when considering all the nodes and links appearing at least once during the whole recording duration, the neighbors of a node being users calling or called by the considered user throughout the record.

### A. Dataset

We apply our tools on a cellphone call record. Nodes are anonymized phone numbers of a European mobile phone provider. Some aspects of this dataset are described in [LB$^+$08], and a statistical analysis of the underlying static network obtained is proposed in [SP09]. To remain as general as possible, our study will be constrained to the simplest attainable information in directed dynamical networks: an event will be described by the triplet $\{s, d, t\}$.

For confidentiality reasons, we were given a connected subset of around a million individuals selected randomly among the users of the provider. Moreover, as we are interested in information transmission, we constrain ourselves to the study of "successful" phone calls — i.e., those where the receiver answers the phone call. Finally, we have a collection of around 14 million phone calls over a period of 1 month.

### B. Comparison models

Our approach consists in comparing the features of the real dynamical network to randomized datasets, which we refer here as comparison models. These comparison models will allow us to identify the features of real data that drive the information spreading process, by comparing real data statistics to statistics obtained from randomized data that lacks correlations, bursting activity, etc. We provide below a short description of the comparison models.

*1) Time-mixing model (tmm):* The phone calls timestamps are randomly mixed on the whole database, source and destination remain identical. With $\mathcal{T}$, the set of timestamps of the whole dataset, an event $\{s, d, t\}$ of the original dataset is described in this model as

$$\{s, d, t'\},\ t' \in \mathcal{T}$$

This trivial model keeps the global activity rate unchanged (such as daily or weekly periodicities), making it very close to models existing in the literature [MS$^+$08], [ZT$^+$10]. On the other hand it breaks the activity rates of each node taken individually, which is known to be non-poissonian in human communication datasets. Some authors described it as *bursty* in phone networks, and this is supposed to play an important role in the spreading phenomena [VR$^+$07]. However, we show in the following that it yields results sufficiently close to the real data to be used as a baseline for comparing them to the other model that we define.

*2) Correlation-mixing model (cmm):* Source and timestamps are kept identical, but destinations are shuffled within the set of destinations which are reached by this particular source during the whole record. If $\mathcal{D}_s$ is the set of the destinations reached by $s$ (if $d$ is called $x$ times by $s$, he will be present $x$ times), then any event $\{s, d, t\}$ becomes

$$\{s, d', t\},\ d' \in \mathcal{D}_s$$

This model is specifically designed to detect traces of diffusion phenomena. In both its purposes and features, it has to our knowledge no equivalent in the literature: it is supposed to keep all the characteristics of the original dataset except for the causality link that may exist between a received phone call and subsequent phone calls. In other words, we wipe out the receiver-sender correlations. The calling activity of each individual remains indeed unchanged as well as the destinations of the calls but the correlation possibly existing between the phone calls given by two different users are broken.

## III. Measurements and results

We describe in this section the statistical measurements that we use for tracking diffusion traces. They consist in enumerating dynamical patterns making use of a tunable time scale denoted $\tau$. This approach may be related to the one proposed by Zhao *et al.* in a recent article [ZT$^+$10]; they also acknowledged the importance of such patterns but with different purposes, as identifying patterns characterizing a specific communication network. Here the patterns detected are chosen to be compared with *cmm* results, as we believe that they may support spreading processes.

Let us stress that we aim at characterizing the spreading behavior of the agents, we are not interested in identifying particular events, or knowing the exact and complete route of an information item.

From an algorithmic point of view, as the measures are designed to characterize large datasets, they must be efficient in terms of time complexity. The ones we propose in the following may be applied to a part of the dynamical dataset, so that it is possible to trade precision for speed.

### A. Causality cascades

Let us consider the following situation: user A calls user B intending to give him a piece of information. If information spreading does occur in such database, we expect that B will call C within a rather short time span to relay this information. There is a causality link between these two events and we expect it to be detectable by observing shorter time elapsed between the call received and the next call given, when compared to the normal activity. Such phenomenon may involve more than two events, so that we can imagine some tree-like patterns whose abundance would be affected by the existence of a diffusion process. We explore this possibility in the following.

*1) Definition:* A cascade is a tree whose nodes are users and links are phone calls. The event which sets the starting point is chosen randomly, its destination will be the first node of the cascade — i.e. the root of this tree. A new user is included in the cascade when he is called by an "active" user, that is to say a user who is already in the cascade for less than a time $\tau$; the corresponding phone call is included in the tree as a link. When all nodes of the cascade are not active anymore, it gets extinguished. This definition is implemented in Algorithm 1.

An example of cascade is represented on Figure 1: the root calls three different nodes within a duration of $\tau$, they will themselves call respectively 3, 0 and 1 node and so on, until nobody in the cascade calls anyone within a period of $\tau$ after being called. We call the total number of nodes in the cascade its size and denote it by $\sigma$, the number of levels from the root to the leaves will be named its depth $\delta$.

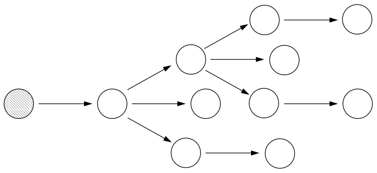


Fig. 1. An example of cascade. The first node is not included in the cascade, so that the root is the first destination. Its size (number of nodes) $\sigma = 11$, its depth (number of levels from the root to the leaves) $\delta = 5$.

Such cascade is supposed to be sufficiently representative of a possible spreading process to provide a good proxy of it. A description close to the one we are using can be found in the literature [MML10], here it is especially suited to record traces of diffusion processes using the timescale $\tau$. Let us notice that our cascade definition can be thought of as a deterministic SIR model, as described for example in [New02], with the difference that we consider directed networks. However, this aspect is out of the scope of this paper and all the questions that can arise from this comparison are the focus of another work [PT11].

For efficiency reasons, we do not use every event as a seed of a cascade, but we made samples with $10^2$ different realizations of the models and measured cascades using $10^5$ starting points for each of them.

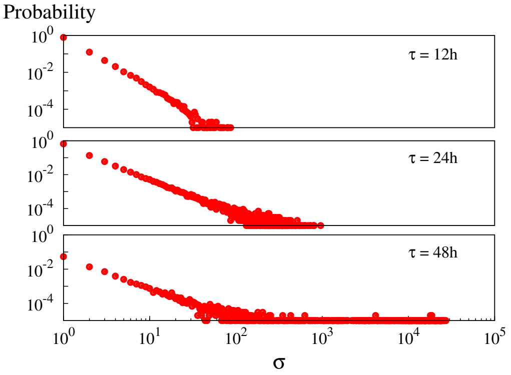


Fig. 2. Probability distribution of sizes of the cascades (real data), for $\tau =$ 12h, 24h, 48h.

*2) Size and depth distributions:* Figure 2 shows the distribution of sizes of cascades for three values of $\tau$: 12h, 24h and 48h. We observe that above a certain threshold (comprised between 24h and 48h), very large cascades appear, with sizes comparable to the size of the static network (which is of $10^6$ nodes, according to our former definition). Moreover, at large values of $\tau$ the distribution of sizes and depths obtained from real data, *tmm*, and *cmm* are virtually indistinguishable. From this fact, we conclude that the average information spreading process can be described at this scale by a random calling behavior, which only takes into account the activity patterns over the whole dataset (daily and weekly cycles for instance), but is not affected by either correlations among phone calls or individual bursty activity [1]. In other words, diffusion traces in these giant cascades are not detectable, so from now on, we will consider smaller time scales.

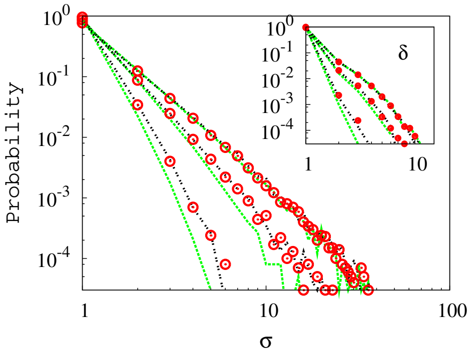


Fig. 3. Probability distribution of sizes (main) and depths (inset) of the cascades. Red circles: real data, green line: *time-mixing model*, dashed black line: *correlation-mixing model*. For each picture, left sheaf of curves: $\tau =$ 30min, middle: $\tau = 3$h, right: $\tau = 12$h.

We present in Figure 3 the distribution of sizes (and depths) of the cascades obtained from the real data and both models, for different values of $\tau =$ 30min, 3h and 12h. At small $\tau$ values, *time-mixing model* produces smaller cascades than real data. Users are known to call at a very heterogeneous rate in real world: high activity periods may be followed by long rests — what is referred to as a *bursty behavior*. But *tmm* breaks down this individual activity rate: for each event, it associates a random timestamp of the dataset, therefore this effect can be explained by the fact that the individual bursty activity increases the probability of calling a new destination after being called within a $\tau$ period. Note that in the literature, it is suggested that bursty individual activity patterns slow down the spreading of information [VR$^+$07]. However, the observation of larger (in both size and depth) cascades in the real data than in the *tmm* indicates rather that the bursty behavior of individuals tends to promote the spreading. This is due to the fact that in [VR$^+$07], the authors study the decay time in a SI model, which corresponds to a SIR model with $\tau \to \infty$, while our study deals with small time scales. In short, a spreading process occurring on a social network should be promoted by the bursty behavior at a local scale while it is slowed down at a macroscopic level.

[1] A deeper analysis at these time scales can be found in [PT11].

**Algorithm 1:** Definition of a cascade of parameter $\tau$.

```
input  : $\tau$ ; $\mathcal{E} = \{e_i\}_{i \in I}$ : sequence of events ordered by increasing timestamp;
output: tree $\mathsf{T} = \mathsf{N} \times \mathsf{L}$;
// Initialization:
draw randomly $e_i = \{s_i, d_i, t_i\} \in \mathcal{E}$ ;                 // random selection of the root
$\mathcal{L} \leftarrow (e_i)$ ;                               // ordered list of events with active nodes
$\mathsf{T} \leftarrow \{d_i\} \times \varnothing$ ;                      // initialization of the cascade
// walk through the data as long as a node is active:
while $\mathcal{L} \neq ()$ and $i \neq max\ I$ do
    $i \leftarrow i + 1$;
    $e_i \leftarrow \{s_i, d_i, t_i\}$;
    $\mathcal{L}' \leftarrow \mathcal{L}$;
    if $d_i \notin \mathsf{N}$ then
        // search for a father in the cascade:
        while $\mathcal{L}' \neq ()$ do
            $e_f = \{s_f, d_f, t_f\} \leftarrow$ head $\mathcal{L}'$ ;
            // test if $d_f$ is still active:
            if $t_f + \tau < t_i$ then
                $\mathcal{L} \leftarrow$ tail $\mathcal{L}$;
                $\mathcal{L}' \leftarrow$ tail $\mathcal{L}'$;
            else
                // test of inclusion in the cascade:
                if $d_f = s_i$ then
                    $\mathsf{N} \leftarrow \mathsf{N} \cup \{d_i\}$ ; $\mathsf{L} \leftarrow \mathsf{L} \cup \{(s_i, d_i)\}$;
                    $\mathcal{L} \leftarrow$ append $\mathcal{L}$ $(e_i)$;
                else
                    $\mathcal{L}' \leftarrow$ tail $\mathcal{L}'$
```

Besides, the *correlation-mixing model* curves fit real data values for any $\tau$, implying that correlations between the activity of two communicating users — and consequently causality effects — do not impact the statistic of size and depth distributions. In the following we perform more refined statistical analysis and show that comparing them to the *cmm* provides detectable traces of the causality effects.

*3) Shape of the cascades:* We focus on short time scales where correlations may have a stronger impact on the statistics, according to Fig. 3. One of the simplest ways of describing the relative abundance of cascade types consists in enumerating cascades with both size and depth fixed. We collect in Table I the probability $\mathcal{P}_{\sigma,\delta}(\tau)$ of observing a cascade of size $\sigma$ and depth $\delta$, for different (but low) $\tau$. *Correlation mixing model* is supposed to be in all points identical to the real data except for the existence of correlation between events. As we are looking for the amount of patterns corresponding to causality-related events, we focus in this analysis on *cmm* and real data. In addition, we only keep low size ($\sigma$) values (which implies low depths $\delta$, as $\delta \leq \sigma$) to have sufficiently large amounts of cascades.

As can be seen on these examples, the global trend is the same for various $\tau$ values: the *correlation-mixing model* overestimates the probability of low-depth cascades while it underestimates high-depth ones. In other words, at this scale, real-world promotes more "chain-like" cascades and less "star-like" than in the *cmm* case, where causality links between events have been broken.

We can perform a quantitative comparison by measuring the ratio:

$$R_{\sigma,\delta}(\tau) = \frac{\mathcal{P}^{real}_{\sigma,\delta}(\tau)}{\mathcal{P}^{cmm}_{\sigma,\delta}(\tau)}$$

For example, in the case of ($\sigma$=5, $\delta$=5) cascades, corresponding to the chain-like pattern: ○→○→○→○→○ , $R_{5,5}(12\text{h}) = 1.72$, indicating that the *cmm* model can only account on average for 58% of the cascades observed using real data. We can thus conclude that the remaining 42% of these cascades contain causality-correlated events, which suggests that some information is propagating through these cascades. Along similar lines, the same kind of estimates can be done using any other measurement.

Then, we plot on Figure 4 $R_{\sigma,\delta}$ as a function of $\tau$ for several values of ($\sigma$,$\delta$), to get some insights about the time scales involved in the information spreading process. This measure is meaningful for patterns which are sufficiently numerous: large causality-correlated cascades are too rare to be observed through a statistical detection method. In other words, large-scale spreading processes does not occur frequently in such

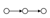
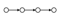
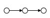
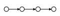
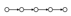
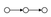
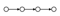
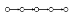
| | $\sigma$ | $\delta$ | type | $\mathcal{P}^{real}_{\sigma,\delta}(\tau)$ | $\mathcal{P}^{cmm}_{\sigma,\delta}(\tau)$ |
|---|---|---|---|---|---|
| $\tau$ = 30min | 3 | 2 | | 2.34 $\cdot 10^{-4}$ | 2.88 $\cdot 10^{-4}$ |
| | 3 | 3 |  | 1.64 $\cdot 10^{-4}$ | 1.37 $\cdot 10^{-4}$ |
| | 4 | 2 | | 0.22 $\cdot 10^{-4}$ | 0.33 $\cdot 10^{-4}$ |
| | 4 | 3 | | 0.42 $\cdot 10^{-4}$ | 0.42 $\cdot 10^{-4}$ |
| | 4 | 4 |  | 0.10 $\cdot 10^{-4}$ | 0.07 $\cdot 10^{-4}$ |
| $\tau$ = 3h | 3 | 2 | | 1.31 $\cdot 10^{-3}$ | 1.48 $\cdot 10^{-3}$ |
| | 3 | 3 |  | 1.06 $\cdot 10^{-3}$ | 0.93 $\cdot 10^{-3}$ |
| | 4 | 2 | | 2.34 $\cdot 10^{-4}$ | 3.06 $\cdot 10^{-4}$ |
| | 4 | 3 | | 5.39 $\cdot 10^{-4}$ | 5.37 $\cdot 10^{-4}$ |
| | 4 | 4 |  | 1.36 $\cdot 10^{-4}$ | 1.06 $\cdot 10^{-4}$ |
| | 5 | 2 | | 0.51 $\cdot 10^{-4}$ | 0.77 $\cdot 10^{-4}$ |
| | 5 | 3 | | 2.39 $\cdot 10^{-4}$ | 2.47 $\cdot 10^{-4}$ |
| | 5 | 4 | | 1.16 $\cdot 10^{-4}$ | 1.09 $\cdot 10^{-4}$ |
| | 5 | 5 |  | 0.19 $\cdot 10^{-4}$ | 0.11 $\cdot 10^{-4}$ |
| $\tau$ = 12h | 3 | 2 | | 2.52 $\cdot 10^{-3}$ | 2.79 $\cdot 10^{-3}$ |
| | 3 | 3 |  | 1.85 $\cdot 10^{-3}$ | 1.69 $\cdot 10^{-3}$ |
| | 4 | 2 | | 6.18 $\cdot 10^{-4}$ | 7.80 $\cdot 10^{-4}$ |
| | 4 | 3 | | 12.13 $\cdot 10^{-4}$ | 11.66 $\cdot 10^{-4}$ |
| | 4 | 4 |  | 2.30 $\cdot 10^{-4}$ | 2.33 $\cdot 10^{-4}$ |
| | 5 | 2 | | 1.65 $\cdot 10^{-4}$ | 2.16 $\cdot 10^{-4}$ |
| | 5 | 3 | | 5.95 $\cdot 10^{-4}$ | 6.75 $\cdot 10^{-4}$ |
| | 5 | 4 | | 2.77 $\cdot 10^{-4}$ | 2.54 $\cdot 10^{-4}$ |
| | 5 | 5 |  | 0.36 $\cdot 10^{-4}$ | 0.21 $\cdot 10^{-4}$ |

TABLE I

PROBABILITY OF HAVING A CASCADE OF A FIXED SIZE $\sigma$ AND DEPTH $\delta$.

cell phone dataset. If it does exist, it might be observed through a method to detect outliers, but this is out of the scope of this work.

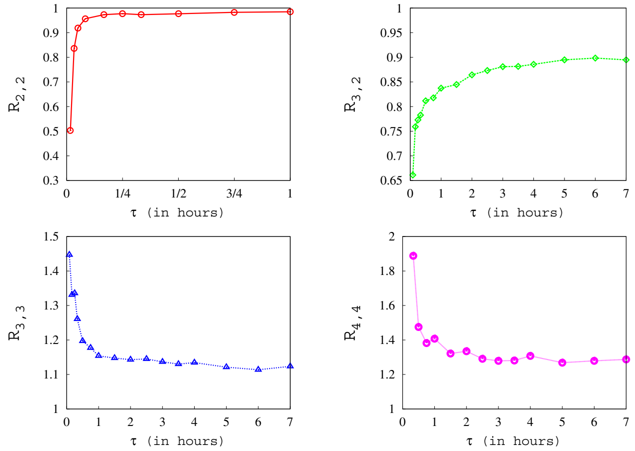


Fig. 4. $R_{\sigma,\delta}(\tau) = \mathcal{P}^{real}_{\sigma,\delta}/\mathcal{P}^{cmm}_{\sigma,\delta}$ as a function of $\tau$. Red circles: $R_{2,2}(\tau)$, green diamonds: $R_{3,2}(\tau)$, blue triangles: $R_{3,3}(\tau)$, pink circles: $R_{4,4}(\tau)$.

For the patterns under consideration, we observe a transient whose length varies from 10 minutes ($R_{2,2}$) to 5 hours ($R_{3,2}$), then the ratio is steady. Afterwards, the abundance of cascades does not bring any new information: both $\mathcal{P}^{real}_{\sigma,\delta}$ and $\mathcal{P}^{cmm}_{\sigma,\delta}$ can change, but their ratio remains constant.

Furthermore, during the first 5 hours, a large majority of the patterns containing correlations involve 2 or 3 nodes, meaning that the causality effects at short time scales impact significantly only the statistics of small cascades in this mobile phone dataset. That does not mean that these patterns are the only one which may support information spreading, nor that all the events of these patterns spread exactly the same information, but it means that at least one event as a causality bond with another: a call has been given *because* of a former call.

### B. Information loops

The study of cascades has helped us to understand the features of the causality relationships between events possibly relaying an information in the network. Information loops do not have an impact on a large-scale spreading dynamics, since they do not involve new informed users. Nevertheless, it may occur that there is an active flow of information in close loops which is not visible using cascades — where nodes appear only once. On the other hand, we expect information loops to be strongly affected by causality effects. In this section, we focus on the possibility of such information loops.

*1) Reciprocal patterns:* More precisely, after a certain call $\{A, B, t_1\}$, we record the calls $\{B, C, t_2\}$ given by $B$ during a period of $\tau$, i.e. $t_2 - t_1 \leq \tau$, which leads to 3-nodes paths $A \to B \to C$. In the case where $C = A$, we will denote the pattern obtained a *reciprocal pattern*, a schematic representation is given in Figure 5, and a simple algorithm to enumerate them in Algo. 2.

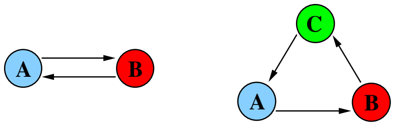


Fig. 5. Schematic representation of dynamical reciprocal pattern (left) and directed triangular patterns (right).

Let us first count the number of such motifs for a few values of $\tau$:

| $\tau$ | 5min | 1h | 3h | 10h |
|---|---|---|---|---|
| $\mathcal{P}^{real}_r$ | 0.018 | 0.076 | 0.120 | 0.166 |
| $\mathcal{P}^{cmm}_r$ | 0.008 | 0.046 | 0.084 | 0.131 |

We can see that *cmm* clearly underestimates these motifs for all these $\tau$, meaning that the causality effects existing in real data but not in *cmm* tends to create more reciprocal motifs — what was expected.

To obtain quantitative estimates of the motifs containing a causality bond, we plot the ratio $R_r(\tau) = \mathcal{P}^{real}_r/\mathcal{P}^{cmm}_r$ of probabilities measured in the one hand for the real data and in the other for *cmm* — in the same way as what we did formerly in the case of cascading motifs. The corresponding results are shown on Figure 6. We also represented in the inset $\mathcal{P}^{cmm}_r/\mathcal{P}^{tmm}_r$ to show that there are more reciprocal motifs in the *cmm* than in the *tmm* during the first 8 hours. It suggests that the activity of neighbors in the static network — which stands in *cmm* but not in *tmm* — are correlated at short time scales.

**Algorithm 2:** Algorithm to enumerate reciprocal motifs and directed triangular motifs.

```
input : $\tau$ ; $\mathcal{E} = \{e_i\}_{i \in I}$ : sequence of events ordered by increasing timestamp;
output: Number of reciprocal motifs $N_r$ and directed triangular motifs $N_t$ ;
draw randomly $e_m = \{s_m, d_m, t_m\} \in \mathcal{E}$ ;                // memorization of the seed event
$e_i = \{s_i, d_i, t_i\} \leftarrow e_m$ ;
$N_r \leftarrow 0$ ; $N_t \leftarrow 0$ ;                                 // initialization of the output
while $t_i < t_m + \tau$ and $i \neq max\ I$ do
    $i \leftarrow i + 1$;
    $e_i \leftarrow \{s_i, d_i, t_i\}$;
    if $s_i = d_m$ then
        // reciprocal motifs enumeration:
        if $d_i = s_m$ then
            $N_r \leftarrow N_r + 1$
        else
            $j \leftarrow i$;
            // triangular motifs enumeration:
            while $t_j < t_i + \tau$ and $j \neq max\ I$ do
                $j \leftarrow j + 1$;
                $e_j \leftarrow \{s_j, d_j, t_j\}$;
                if $s_j = d_i$ and $d_j = s_m$ then
                    $N_t \leftarrow N_t + 1$;
```

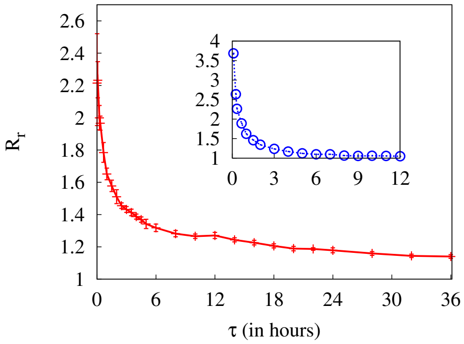


Fig. 6. Main: $R_r(\tau) = \frac{\mathcal{P}_r^{real}}{\mathcal{P}_r^{cmm}}$ as a function of $\tau$. Inset: $\frac{\mathcal{P}_r^{cmm}}{\mathcal{P}_r^{tmm}}$.

We consider the main plot which compares the real data to *cmm* making use of the ratio $\mathcal{P}_r^{cmm}/\mathcal{P}_r^{real}$, the behavior can be roughly described as follows: during the first 8 hours the ratio drops sharply, then $R_r$ decreases much more slowly.

We can make quantitative estimates of the amount of correlated events: for instance, there are 2.3 times more reciprocal motifs in the real data than in the *cmm*, for $\tau$ =5min, we thus state that in 57% of these reciprocal motifs observed contains some causality relationship. In other words, the fact that B calls back A within 5 minutes is directly related to the fact that A called B just before in at least 57% of the cases.

*2) Directed triangular motifs:* Following our reasoning, reciprocal calls are not the only communication motifs missed by the cascading behavior: it is blind to longer cycles of communication too.

We define a *directed triangular motif* in the same way as what we did for a reciprocal motif: after a certain call $\{A, B, t_1\}$, we record the calls given by $B$ during a period of $\tau$: $\{B, C, t_2\}$, and then the calls given by $C$ within $\tau$ after receiving this phone call: $\{C, D, t_3\}$, thus creating a 4-node path of the form: $A \rightarrow B \rightarrow C \rightarrow D$. If $D = A$ the 3 nodes motif formed is denoted a (dynamical) directed triangular motif, a schematic representation of it is given in Fig. 5, and an enumeration method is included in Algo. 2.

Along similar lines as before, we define $\mathcal{P}_t(\tau)$ as the probability to observe a directed triangular motif with parameter $\tau$, and then plot $R_t(\tau) = \mathcal{P}_t^{real}/\mathcal{P}_t^{cmm}$ of motifs obtained for real data and *cmm* on Figure 7. We can observe qualitatively the same kind of behaviors as above but at a smaller scale: there is a fast decay of the triangular motifs ratio during a transient regime of around 3 hours and then a steady state is reached. So the triangular motifs carry indeed detectable traces of correlation effects: for example, observing such a triangle 30 minutes after the root phone call is around 2.4 times more frequent in real data than in *cmm*, meaning that 68% of these motifs contain a causality relationship.

Such measures can be easily generalized to motifs involving more nodes. Yet when we consider motifs larger than or equal to 4 nodes on this specific dataset, the difference of behaviors between real data and *cmm* is not conclusive as the measures are too noisy. In other words, the causality-correlated behaviors that we measured in this mobile phone dataset rarely involve more than 3 nodes and they are observed at short time scales only — except for the reciprocal patterns which may be detected during more than a day. It suggests

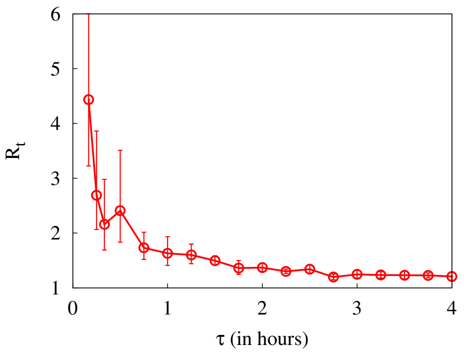


Fig. 7. Main: $R_t(\tau) = \mathcal{P}_t^{real}/\mathcal{P}_t^{cmm}$ as a function of $\tau$.

that such communication network is not used to support large scale diffusion processes. This observation is consistent with our daily experience: most of us use the telephone to talk with friends, colleagues, and relatives without the intention of passing along information we received from other phone call.

## IV. Conclusion

We showed in this article that using the appropriate models and statistical tools, it is possible to count dynamical motifs of a large communication dataset which contain events causality-related. We saw that the main features of cascades — such as their size and depth distributions — could be essentially explained by a random process, without any correlation between events. It is however important to take into account the bursty activity behavior of the users, which seems to promote the spreading processes at short time scales. This analysis reveals a remarkably low amount of causality bonds on mobile phone database, probably because the spreading of information over a large part of such network simply does not happen. On the contrary we observe causality effects at the level of short chain-like patterns as well as closed loops. It indicates the existence of information flows at a local scale — e.g. among group of friends. Such bond can be understood as the trace of a diffusion process in the cell phone dataset. We understand the absence of large-scale spreading as an indication that people usually do not use their mobile phone as a means to relay general news but rather personal information.

The procedure here described to detect causality-related events can be tested on other kind of dynamical networks, where the information exchanged is available and the flow is directly observable, e.g. Twitter. It is likely that other dynamical networks exhibit different information spreading properties. For instance, in Twitter, we can expect a usage of the network intended to spread news to a wider audience. Along the same lines, peer-to-peer datasets should exhibit interactions involving correlations at larger scales (both in time and depth), because packets are circulating between total strangers, yet the formalism needs to be adapted to this context. More generally, these versatile tools can be specialized to richer datasets, where information such as ties strengths or node features can give new insights on the interaction motifs and especially on diffusive behaviors.

## Acknowledgments

We would like to thank Clémence Magnien, Matthieu Latapy, Christian Borghesi, Fabien Tarissan and Eduardo Altmann for their useful comments. This work was partly funded by a grant from the *Agence Nationale de la Recherche*, with reference ANR-10-JCJC-0202 (*DynGraph* project).

## References

[Bar05] Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. Nature **435** (2005) 207–211.
[CG+08] Candia, J. et al.: Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A **41** (2008).
[CR09] Cointet, J.P. and Roth, C.: Socio-semantic dynamics in a blog network. Proc. of ICCSE'09 (2009) 114–121.
[DA+06] Dezsö, Z. et al.: Dynamics of information access on the web. Physical Review E **73** (2006).
[FCL11] Friggeri, A. and Cointet, J.P. and Latapy, M.: A real-world spreading experiment in the blogosphere. To appear in Complex Systems.
[GG+04] Gruhl, D. et al.: Information diffusion through blogspace. Proc. of the $13^{th}$ Int. Conf. on the World Wide Web (2004) 491–501.
[IM09] Iribarren, J.L. and Moro, E.: Impact of human activity patterns on the dynamics of information diffusion. Physical Review Letters **103** (2009).
[KK+11] Karsai, M. et al.: Small but slow world: how network topology and burstiness slow down spreading. Physical Review E **83** (2011).
[LB+08] Lambiotte, R. et al.: Geographical dispersal of mobile communication networks. Physica A **387** (2008) 5317–5325.
[LM+07] Leskovec, J. et al.: Cascading behavior in large blog graphs patterns and a model. Proceedings of SIAM International Conference on Data Mining (2007).
[MS+08] Malmgren, R.D. et al.: A poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences (2008).
[MGV11] Min, B. and Goh, K.-I. and Vazquez, A.: Spreading dynamics following bursty human activity patterns. Physical Review E **83** (2011).
[MML10] Miritello, G. and Moro, E. and Lara, R.: The dynamical strength of social ties in information spreading. Arxiv preprint 1011.5367 (2010).
[New02] Newman, M.E.J.: Spread of epidemic disease on networks. Physical Review E **66** (2002).
[OS+07] Onnela, J.P. et al.: Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics **9** (2007).
[PV01] Pastor-Satorras, R. and Vespignani, A.: Epidemic spreading in scale-free networks. Physical Review Letters **86** (2001) 3200–3203.
[PT11] Peruani, F., Tabourier, L.: Directedness of information flow in mobile phone communication networks. To be published.
[SGL10] Salahbrahim, A. and Le Grand, B. and Latapy, M.: Some insight on dynamics of posts and citations in different blog communities Proc. of ICC Workshops (2010).
[SQ+10] Song, C. et al.: Limits of predictability in human mobility. Science **327** (2010) 1018–1021.
[SP09] Stoica, A. and Prieur, C.: Structure of neighborhoods in a large social network. Proc. of ICCSE'09 **4** (2009) 26–33.
[VO+06] Vazquez, A. et al.: Modeling bursts and heavy tails in human dynamics. Physical Review E **73** (2006).
[VR+07] Vazquez, A. et al.: Impact of non-Poissonian activity patterns on spreading processes. Physical Review Letters **98** (2007).
[ZT+10] Zhao, Q. et al.: Communication motifs: a tool to characterize social communications. Proc. of CIKM'10 (2010) 1645–1648.