Analysis of dynamic sensor networks: power law then what?

(Invited Paper)

Éric Fleury, Jean-Loup Guillaume
CITI / ARES – INRIA
INSA de Lyon
F-69621 Villeurbanne FRANCE

Céline Robardet
LIRIS / CNRS UMR 5205
INSA de Lyon
F-69621 Villeurbanne FRANCE

Antoine Scherrer
CITI / COMPSYS – INRIA
INSA de Lyon
F-69621 Villeurbanne FRANCE

***Abstract*— Recent studies on wireless sensor networks (WSN) have shown that the duration of contacts and inter-contacts are power law distributed. While this is a strong property of these networks, we will show that this is not sufficient to describe properly the dynamics of sensor networks. We will present some coupled arguments from data mining, random processes and graph theory to describe more accurately the dynamics with the use of a random model to show the limits of an approach limited to power law contact durations.**

I. Introduction

Mobile devices equipped with wireless capabilities, which enable new communication services, have encountered a fantastic growth in the last few years. Advances and miniaturization of micro-electronic devices have pushed the development of new fields of applications for wireless networks. The increasing popularity of a wide range of wireless devices allows new paradigms and application classes.

Scientific research has massed worldwide around these new challenges presented by multi hop wireless networks. A broad spectrum of topics is under investigation. It covers research in pure ad hoc networks, in hybrid wireless ad hoc networks, in wireless sensor networks (WSN), in SISes (spontaneous information systems) and in DTNs (delay-tolerant networks). A common characteristic of this large range of wireless networks is that intermittent connectivity is the norm due to environmental dynamics and the intentional duty cycling of wireless nodes. However, despite several years of intensive research, a gap remains in terms of experimental aspects of in situ dynamic networks.

In such ambient context where nodes will be spread around in the environment and/or on each user, it becomes possible to route data on such network based on pairwise contacts between devices/users. Communication services based on such network, called also Delay Tolerant Networks (DTN), will deeply rely on the mobility and on the characteristics of the underlying networks. It appears crucial to better understand the intrinsic characteristics of such dynamic radio networks; to be able to analyze and model interactions between individuals/devices, in order to propose secure methods and protocols suitable for this context.

Recently, Chaintrau *et al.* provide data traces of pairwise contacts collected during the Infocom 2005 conference. From this experimental approach their analysis shows long tailed distribution for the inter contact time (time between two transfer opportunities for the same pair of devices) and they claim that "*inter contact time distribution can be compared to the one of power law*". As stated in [1], one can mark the popularity of power law in 1999. Several papers [2]–[5] appear in leading journals (Nature, Science) and report independently that the distribution of degrees in "real" graph (WWW, Internet Topology) seems to follow a power law: the frequency of nodes/web pages with connectivity k falls off as $k^{-\alpha}$.

If we agree on the fact that studying the dynamics and evolution of large-scale dynamic networks is a fundamental and difficult, but promising problem, one may have some doubts about the generality of scale free networks and that "*[...] nature has some universal organizational principles that might finally allow us to formulate a general theory of complex systems*" [6]. More precisely, finding a power law for some network characteristic distribution is not so surprising and what do power law distribution really signify? One should keep in mind that high variability does not necessarily imply power law. Moreover, characterizing a power law behavior is done by inferring a fitting curve on a log-log plot. It appears that various kind of data can be approximated by drawing straight lines on such log-log scale plots.

The main purpose of this paper is to demonstrate that the so called power-law argument is not the ultimate one and that it is worthy to study and analyze dynamic networks under several points of view in order to extract their characteristics and behaviors. As stated above, power-law are "quite" easy to generate, that's why we can find them "everywhere" and finding such scale free networks does not imply any deep or fundamental knowledge on the intrinsic structure of the network. A main contribution of this paper is that in order to extract knowledge on dynamic networks, we introduce and present some coupled arguments from data mining, random processes and graph theory to describe more accurately the dynamics with the use of a random model. We show the limits of an approach limited to a power-law contact duration.

We also emphasis the need of addressing interdisciplinary issues since dynamic networks are becoming a central point of interest, not only for engineers, computer scientists but also for other domains, such as sociology, epidemiology, and statistical physics. While far from complete, our results stay consistent with two complementary goals: a crucial need of real data gathered from in situ test beds and fostering the development of precise tools in order to analyze data to enable theoretical models to validate the ongoing research conducted in the various domains that touch on dynamic networks.

The remainder of this article is organized as follows. Section II presents the three main approaches that we used in this article: data mining, random processes and graph theory and provides the basic background, including mathematical definitions. Sections III and IV are dedicated to network dynamics. In section III we study the evolution of the network by presenting the inter contact approach but we also apply basic metrics from graph theory and random process. Then, in section IV we introduce more complex metrics based on *Maximal Connected Subgraph (*MCS*)* and on their evolutions. This analysis is based on graph theory tools and on a data mining approach that reveals to be very efficient for this kind of computation and analysis. In the section V we introduce a very simple random dynamical model for dynamic networks having a power law in their inter contact sequence. This models highlights the diversity of properties that are needed to characterize dynamic networks. Our model provides insight into existing notion of dynamic networks and demonstrates that the structure and the dynamics are not a direct consequence of the intra and inter contact durations. We conclude in section VI that many open problems and works remain, including the three fields considered in this article (data mining, random processes and graph theory). We present also future works related to gathering in situ data on a larger scale in terms of the number of sensors deployed, the diversity of the populations and the duration of the experiments.

A. Data

During the Infocom 2005 conference, Bluetooth sensors have been distributed to a small set of participants which where asked to keep the sensors with them continuously. These sensors where able to detect and record the presence of others Bluetooth devices in their radio range neighborhood and, even if they succeed to detect various kind of Bluetooth devices like laptop and others, the main objective was to detect the proximity between sensors. These data and the way they have been obtained is precisely described in [7]. Note also that the sensors had no localization capability, therefore we cannot have any information on the actual movements of individuals carrying the sensors or on the proximity of two given sensors: either they are near enough to "*see*" each other (from a pure radio/communication layer point of view) or not.

The available data concern 41 sensors over a period of nearly 3 days[1]. The data are precise at the second and, for each second, a set of existing links is given. Note first that it may happen that no link exist at a given time step, which means that all 41 sensors are far from each others. In the available data it may happen that a given sensor u had seen another sensor v and that v missed u. The data we are going to use hereafter are similar to the original one except that as soon as one link exists, the symmetric link also exists.

[1]254 151 seconds to be precise.

II. Approaches

A. Data mining

A branch of data mining research area concerns the computation of set patterns by means of complete solvers. Following the Agrawal et al. [8] seminal paper, it consists in defining the shape of a priori interesting patterns by means of constraints, some of them being sufficiently tight to drastically reduce the search space and turn the computation to be feasible. Mannila et al. [9] formally defined this task as computing the subset of a language $\mathcal{L}$ that satisfies a predicate q. $\mathcal{L}$ is a class of sentences that expresses properties or defines subgroups of the data $\mathbf{r}$, and q is used for evaluating whether a sentence $\phi \in \mathcal{L}$ defines a potentially interesting subclass of the data $\mathbf{r}$. The task thus consists in computing the theory

$$\mathcal{T}h(\mathcal{L}, \mathbf{r}, q) = \{\phi \in \mathcal{L} \text{ such that } q(\mathbf{r}, \phi) \text{ is true}\}$$

Most frequently used patterns to built theory are frequent itemsets (which are the first step to compute association rules) and closed sets, also called formal concepts. Other pattern types have been more recently defined like frequent trees, frequent graphs or fault-tolerant frequent sets, but their definitions are less established and give still rise to discussions.

In this paper, we will try to capture some properties of the dynamic of the studied sensor network thanks to formal concept computation. The data $\mathbf{r}$ are defined by a binary relation between a set of objects $\mathcal{O}$ (*e.g.* all possible point to point links between sensors) and a set of properties $\mathcal{P}$ (*e.g.* experiment time steps). This binary relation defines in the present case at which time steps the links between sensors are detected. Extracting formal concepts from this relation consists in finding all "natural" groups of properties and objects. Such "natural" groups contain all objects that share a common subset of properties, or all properties shared by a subset of objects. A nice property on these groups is that there exists a bijective function between the "natural" groups of objects and those of properties. This function and its inverse are forming a Galois connection. We first briefly recall the Galois connection definition, before explaining why such a property is useful as well for pattern interpretation as for computation facilities.

Definition 1 (Galois connection): Let $(A, \subseteq)$ and $(B, \subseteq)$ be two partially ordered sets. A Galois connection between these partially ordered sets consists of two antitone (*i.e.* order-reversing) functions, $F : A \rightarrow B$ and $G : B \rightarrow A$ such that for all $a \subseteq A$ and $b \subseteq B$, we have: $b \subseteq F(a)$ if and only if $a \subseteq G(b)$

For the Galois connection associated to formal concepts, A is the power set of objects ($2^{\mathcal{O}}$) ordered by set inclusion, and B is the power set of properties ($2^{\mathcal{P}}$) also ordered by

set inclusion. F is defined by $F(X) = \{y \in \mathcal{P} : (x, y) \in \mathbf{r} \text{ for all } x \in X\}$. Similarly, for any subset Y of $\mathcal{P}$, define $G(Y) = \{x \in \mathcal{O} : (x, y) \in \mathbf{r} \text{ for all } y \in Y\}$. The codomain of F contains only all the "natural" groups of properties, whereas the codomain of G contains only all the "natural" groups of objects.

Property 1: A Galois connection satisfies the following properties:

1) $G \circ F$ is extensive *(i.e.* $\forall X \subseteq A,\ X \subseteq G(F(X))$)
2) F and G are antitone (*i.e.* $\forall X \subseteq A,\ X' \subseteq A$, if $X \subseteq X'$ then $F(X') \subseteq F(X)$)
3) $G \circ F$ is monotone.
4) $F \circ G \circ F = F$. This property implies that $G \circ F$ is idempotent (*i.e.* $\forall X \subseteq B,\ G(F(G(F(X)))) = G(F(X))$)

Thus, the composite $G \circ F$ is monotone, extensive and idempotent. This states that $G \circ F$ is in fact a closure operator on A. Dually, $F \circ G$ is also a closure operator on B.
Thus, a formal concept is a couple (X, Y) containing both a closed set of properties ($Y = F \circ G(Y)$) and its corresponding natural object group ($X = G(Y)$). If the binary relation is represented by a boolean matrix, a formal concept is thus a maximal rectangle of true values, up to row and column permutations.

Thanks to the Galois connection, formal concepts are embedded by generalization/specialization relation: when the set of objects (resp. the set of properties) increases, then its associated set of properties (resp. set of objects) decreases. As such it provides powerful characterization mechanisms, useful during the interpretation phase of the patterns.

On the other hand, formal concepts can be computed by enumerating sets of objects or sets of properties (we choose the smallest dimension) with respect to the inclusion order, starting from the empty set to the whole set. The functions of the Galois connection are thus used to compute the closure of the enumerated set and the associated component on the other dimension. Several techniques have been proposed to guarantee that each formal concept is uniquely generated [10].

Depending on the input relation $\mathbf{r}$, the size of the formal concept collection can be really huge, and thus it can be useful to use minimal size constraints on both formal concepts components to select the largest ones. These constraints are actively pushed inside D-MINER [11], the formal concept solver we will use in the following. To summarize, large formal concept theory $\mathcal{T}h(\mathcal{L}, \mathbf{r}, q)$ is defined by

$$\begin{aligned} \mathbf{r} &\subseteq \mathcal{O} \times \mathcal{P} \\ \mathcal{L} &= \{(X, Y) \text{ such that } X \subseteq \mathcal{O} \text{ and } Y \subseteq \mathcal{P}\} \\ q &\equiv Y = F \circ G(Y) \text{ and } X = G(Y) \text{ and } |X| \geq \sigma_1 \\ &\quad \text{and } |Y| \geq \sigma_2 \end{aligned}$$

σ_1 and σ_2 being two integer thresholds used to select largest formal concepts[2].

[2] $|X|$ denotes the size of the set X.

B. Random processes

Another approach for the analysis of evolving sensor networks is to use stochastic (or random) processes theory. The basic idea is to consider that some quantities of a dynamic graph (degree evolutions of a given node, contact and inter-contact durations, number of connected components, etc.) are produced by a non-deterministic process. This is obviously not the case for the data we are analyzing, however we consider, as is common usage, that the system is so complex that deterministic chaos arise so that it is very hard to relate the raw data to the physics of the system. Furthermore in the data that we have used in the experiments, the physics of the system is not fully known because the location of the nodes at each time is not recorded. Stochastic analysis and modeling can help to characterize the variability of some quantities extracted from the evolving sensor network. This characterization can be used for building dynamic graphs models and generating dynamic graphs with similar stochastic properties for the evaluation of the performance of various communication protocols for instance.

Random (or stochastic) processes theory is well-established [12]. Let $S[k]$ be a sequence of data describing the evolution of a quantity as a function of the time k (note that k is not always corresponding to a physical time, it can be simply an index), the objective is, from the detailed analysis of the sequence, to find a random process model accurately describing its statistics. To this end, one must choose (manually or using model selection techniques) a parametric stochastic process model and state-of the art parameter estimation techniques can then be used in order to find the best parameter values given the data sequence $S[k]$.

1) Power-law and scaling: Recent studies have shown that the contact and inter-contact duration in dynamic sensor networks are well modeled by a power-law [7], [13]. Such a behavior is an important property since it results in a high variability of durations which can have a strong impact on the performances of communication protocols. Power-laws have also been used by Barabasi *et al.* [3] for the definition *scale-free* graphs. A (static) graph is said to be scale-free if the number of nodes having degree k falls off as $k^{-\alpha}$, thus exhibiting a power-law behavior.

One should however pay attention to the definition of a power-law behavior. There exists two definitions in the literature, a stochastic and a non-stochastic one [14].

The non-stochastic definition of a power-law is as follows. Let $y = (y_1, y_2, \ldots, y_n)$ be a finite sequence ordered such that $y_1 \geq y_2 \ldots \geq y_n$, y is said to follow a *power law* or *scaling* relationship with *scaling index* α if: $k = cy_k^{-\alpha}$

This definition implies that the rank k versus y appears as a line of slope $-\alpha$ when plotted in a log-log scale.

The stochastic definition of a power law behavior for a random variable X is related to the *tail* of its distribution function, it is actually also referred to as an *heavy tailed* distribution. It is usually defined by the shape of the *complementary*

cumulative distribution function (CCDF) as follows:

$$P[X > x] \underset{x\to\infty}{\sim} cx^{-\alpha}$$

Note that this implies that the *probability density function* (PDF) of the random variable X follows:

$$f(x) \underset{x\to\infty}{\sim} cx^{-(\alpha+1)}$$

For $\alpha > 2$, X has finite mean and variance and is not considered as *heavy tailed*; for $1 < \alpha < 2$, X has finite mean but infinite variance and for $0 < \alpha < 1$, X has both infinite mean and variance. This distribution thus characterizes random variable with high variability, as opposed to low variability which appears for instance with exponentially decaying tails.

Power-law distribution is also called *scaling distribution* because the conditional distribution $P[X > x|X > w]$ is the same as the original one $P[X > x]$, except for a change in scale:

$$P[X > x|X > w] \underset{x\to\infty}{\sim} \frac{c}{w^{-\alpha}} x^{-\alpha}$$

This is opposed to exponential distribution, for which conditioning implies a change of location rather than scale [14].

It appears that either the CCDF or the PDF can be used in order to check for power-law in actual data. However, as pointed out in [1], the CCDF provides a more discriminant way of deciding whether some data follow or not a power law and estimate the scaling index α. Moreover, using the PDF can lead to a misinterpretation of the results [14].

2) Dynamic graph analysis: In this work we are focusing on characterizing the *dynamics* of an evolving graph. Recent works have considered the distribution of contact and inter-contact durations [7] as a sufficient statistic to describe these dynamics, claiming for a power-law behavior of both contact and inter-contact durations. This is an interesting initial step, however we argue this is not sufficient to fully describe the evolution of a dynamic graph. One can for instance have a look at the evolution of other quantities in time, such as the degree evolution of each node, the count and size of connected components and the degree distribution among nodes for instance.

For the analysis of such evolutions, not only should be considered the distribution of values (first order statistics), but also the *covariance* (second order statistics), describing how values of the process are correlated at each possible time lag. It is interesting to note that a power-law behavior can be identified in the covariance as well, and is referred to as *long-range dependence* [15]. We want here to emphasize the fact that however different statistics can be modeled using the same family of function (the power-law function family for instance), the meaning and the interpretation of the results may be much different.

C. Graph theory

A lot of recent studies in the field of complex networks have been focused on the topological structure of real networks, and have defined a lot of properties to describe properly these networks, most of these properties being far beyond the scope of this paper. Despite these advances, very few is known on their dynamical properties, which would include both the study of changes of state for nodes and links but also changes of the underlying structure. The main reason to explain this lack of studies comes from the absence of easily observable dynamical networks. The data provided by [7], [13] which we are going to study here, and other similar datasets, are therefore interesting in themselves but could also lead to the definition of new dynamical properties.

Among the very few studies centered on the study of the dynamics of networks, let us cite [16], [17] which give some insight on the evolution of regulatory networks by considering different organisms. In such networks, typical subgraphs related to biological processes, for instance cycle of length 2 or feedforward loops, appear more often and can be tracked in various species. The same kind of problems have been studied in [18] to show the increase of typical subgraphs in networks like the internet or semantic networks over time.

However, these studies are based on nearly static networks where the number of time steps is extremely small, no more than a few dozens. On the contrary, the evolution of imote data is much more complex with thousands of modifications per day.

Let us first recall basic properties of graphs that are going to be used hereafter. Given a graph $G = (V, E)$, let us denote by $n = |V|$ the number of nodes in the graph and by $m = |E|$ the number of links. The density of G is defined as the number of existing links over the number of possible links, that is $d = \frac{2.m}{n(n-1)}$. The degree of a node $u \in V$ is the number of neighbors it has in the graph.

A maximal connected subgraph of G is a maximal subgraph (in terms of nodes and links) of G such that there exists a path between any two nodes. Using this definition, two connected sets of nodes but with different sets of links are assumed to be different maximal connected subgraphs. In a connected subgraph, the length of a path between two nodes u and v is the number of links used to go from u to v following this path and the distance between u and v is the length of a shortest path between these nodes. The diameter (resp. average distance) of a connected subgraph is the maximal (resp. average) distance between any two nodes of the subgraph. These notions related to distance are defined only for connected graphs.

From a wireless networking point of view, these properties have a direct influence on the ability of the network to transmit information from one node to another or to implement algorithms such as flooding. The degree or number of 1-neighbors is the number of nodes which can be reached with only one radio message. The density can give some clues on the number of conflicts one might face when sending a message. Finally the diameter or the average distance is the number of intermediaries that have to be used to send a message in the worst case or on average.

In the following we will use these kind of properties, first in the Section III where we will study the evolution of basic properties such as the number of nodes, links or the average

over time. Second, in Section IV, we will study the connected subgraphs mainly to evaluate the stability of the network over time.

III. Evolution of the network

In this section, we show experimental results obtained with the data described in Section I-A with tools from theories of graph and random processes.

A. Network evolution

The first basic dynamical properties concern the evolution of classical static properties. Fig. 1(a) and 1(b) display the evolution of the number of connected nodes and links over time. First of all one can notice a typical night and day effect which is clearly visible on the evolution of links. During the night, the number of nodes and links are more stable, but there is a high number of connected nodes: the network is basically a set of disjoint links which certainly correspond to roommates. Even if daytime exhibits more variations, one can note that there is no timestep during which the network is a single connected component and there is also no timestep where all nodes are at least connected, even if they are not part of a single component. The maximal value obtained is 34 nodes connected simultaneously, and it happens that all these nodes belong to a same component, the remaining nodes being isolated. The fact that there is always some isolated nodes, around $1/2$ on average for daytime and around $3/4$ for nighttime, implies that information can only be transmitted to these nodes by taking the evolution of the network into account.

Fig. 2(a) displays the same link values for the first day of the conference only. Note that even during the day the evolution of the number of links is not as flat as one could have guessed for that kind of event. While specific periods, like coffee breaks or lunches, can be identified by strong peaks in the number of links - around time 55000, 70000 and 90000 for the lunch breaks, and two smaller ones in the middle of the morning and afternoon sessions - the network is far from being static elsewhere. The main point to observe is that the human behavior has a strong impact on the observed data and that such results can hardly be generalised. For this kind of data, random processes approaches give more precise insight on the phenomenon and are discussed later on.

Fig. 2(b) is a scatter plot of the number of connected nodes versus the number of links over the whole period: for each time step where the network has k nodes and l links, a point of coordinates (k, l) is placed. As one could have guessed, this plot exhibit a positive correlation between nodes and links, that is to say that the more nodes are connected, the more links are present, since the minimum number of links is one half the number of connected nodes and the maximal number is $k(k-1)/2$. However the main point is to notice that the variation of the number of links is nonconstant over the number of nodes[3] with a variation which is approximatively quadratic in the number of nodes. This means that, for a given number of nodes, the network can have a large number of possible configurations, some of which are very sparse and some are more dense, up to around $1/8$ for the densier existing configuration.

[3]This nonconstant variation is named heteroscedasticity.

B. Random process modeling

Numerous of the graph properties evolutions shown in the previous section (degree, connected nodes and links, etc.) are good candidates for a random process based modeling. This modeling can be used to extract information about the data, and to enable the generation of artificial sequences with statistical properties close to the ones of the original data. For instance, as it will be explained in Section V, one can build a dynamic graph generator from the modeling of a quantity (the contact and inter-contact duration in our case).

A common feature of the evolutions shown in previous section is that they exhibit a strong non stationarity, the main cause for that being the different behavior between day and night periods (see Fig. 1(b) for instance). This is bad for (stationary) stochastic process fitting, and we should pay attention to split the data into reasonably stationary parts before fitting a stochastic process model to it. A simple *on/off* process can be used to model day and night periods alternation.

Another common and important characteristic of these evolutions is that they exhibit high variability (see Fig. 2(b) for instance). One way of describing this high variability is to plot the CCDF in a log-log plot and check for a power law behavior as explained in Section II-B. One can for instance use this method to model the contact and inter-contact duration distribution as reported in Fig. 3, which shows the contact and inter-contact duration CCDF in a log-log scale. The straight line behavior over a wide range of duration ($[300\ 20000]$ seconds for inter-contact duration and $[100\ 30000]$ seconds for inter-contact duration) indicates a power law behavior of which we can estimate the scaling index α. The value of α is reported on Fig. 3. These observations are in agreement with the ones reported in [7].

For the analysis of the evolution of a quantity (degree, number of connected nodes and links, etc.), it might be interesting to see the data as an accumulation process, and try to analyze the differential sequence rather than the original one. Classically, if $S[k]$ is the original data sequence, then the (first order) differential sequence is defined as: $D_S[k] = S[k+1] - S[k]$.

As an example, Fig. 4(a), 4(b) and 4(c) show respectively a part of the degree evolution (during a day period) of a particular node, the CCDF and the covariance. The distribution of the differential process is in this case a better information for modeling, the major reason being that the degree gets increased and decreased as links are created and destroyed. It is therefore natural to model the differential sequence rather than the original sequence itself.

The covariance is estimated using a wavelet-based tool [19] and the plot is a spectral log-log representation of the covariance, in the wavelet domain (j is the scale and S_j is roughly

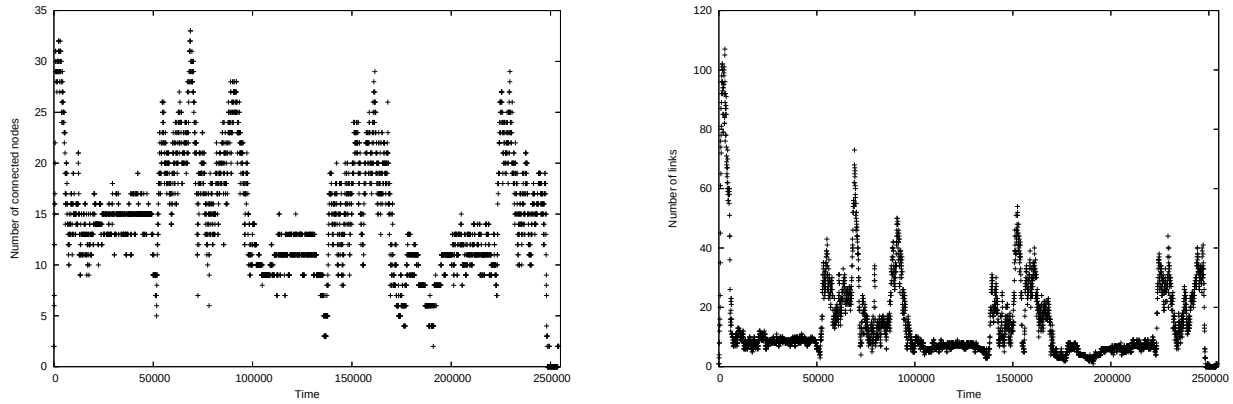


Fig. 1. Evolution of the number of connected nodes (left, a) and links (right, b).

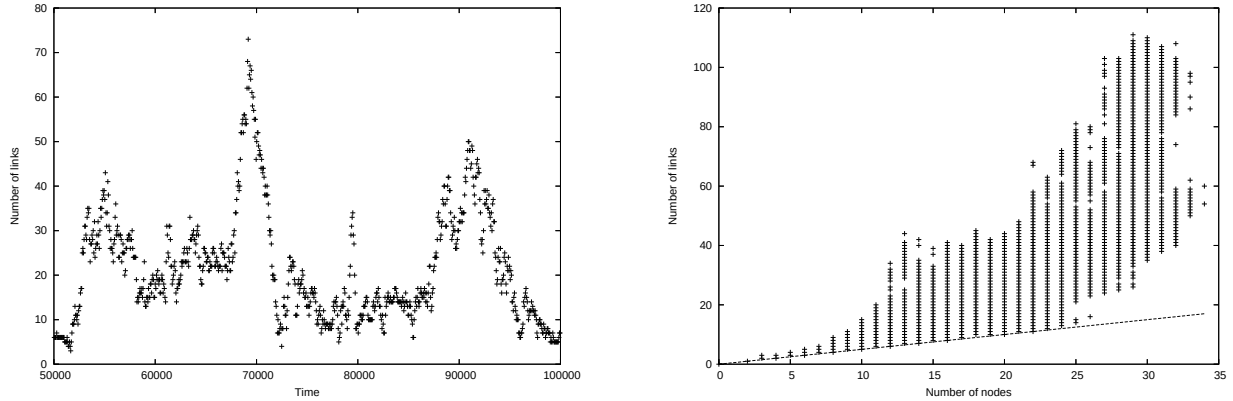
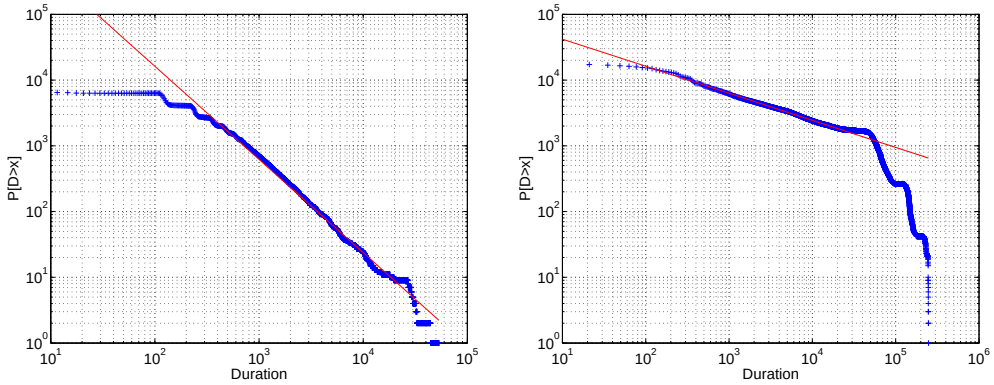


Fig. 2. Evolution of the number of connected links (left, a). Scatter plot of the number of connected nodes versus the number of links (right, b). The straight line correspond to the minimum number of links given a number of connected nodes.



$\alpha = 1.4$ $\alpha = 0.41$

Fig. 3. Contact (left, a) and inter-contact (right, b) duration distribution (CCDF).

the average of wavelet coefficients at scale j). The interested reader is referred [19] for details. A power law behavior in such a plot results in a straight line with slope directly linked to the Hurst parameter H. Note that a power law behavior in the covariance is much different than the one of a distribution, it implies *long range dependence* rather than high variability.

From the plots in Fig. 4, we can build a simple model for the degree evolution of each node. The estimated exponent for the covariance (Fig. 4(c)) implies that the Hurst exponent is close to the special value 0.5, meaning that there is no long range dependence. In this case a good approximation is to consider that the differential sequence of the degree evolution is an IID (Independent Identically Distributed) stochastic process, with distribution function plotted in Fig. 4(b).

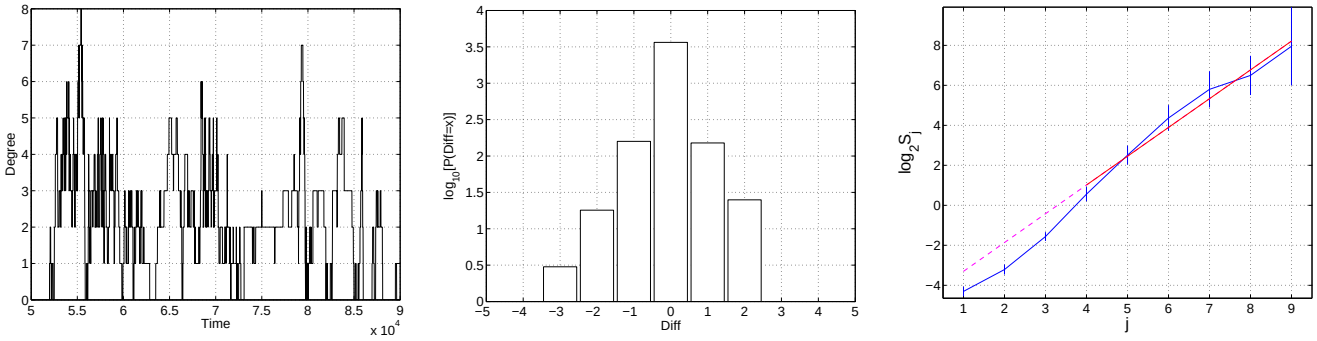


Fig. 4. Degree evolution of node 6. Original evolution (left, a). PDF of the differential sequence (middle, b). Covariance of the differential sequence in the wavelet domain (right, c).

These results are very much preliminary, and this method should be extended to other quantities of the graph (connected nodes and links, etc.).

IV. Components

In this section we are going to deal with components which are defined as sets of links. In the first part we will only consider connected sets of links to see how connected groups of individuals evolve over time, while the second part will be dedicated to sets of links which appear frequently.

A. Maximal connected subgraphs

Recall that a maximal connected subgraph (MCS in the following) is a maximal subgraph (in terms of nodes and links) such that a path exists between every pair of nodes. Since it is going to be widely used in the following, note that a set of connected nodes does not define properly a connected subgraph. The set of links is important too, therefore two similar sets of nodes with different set of links are assumed to be different subgraphs.

One cannot expect the network to be fully connected all the time and results from Section III have exhibited differents behaviors for daytime, nighttime, breaks, lunches, etc. Fig. 5(a) and 5(b) display respectively the number of MCS at each time step and the number of nodes of the biggest one. These figures show that most of the time there are many MCS which in this case might correspond to different sessions of the conference, and that there is nearly no time step during which there is only one MCS.

To go deeper in the study of these subgraphs, we have computed all the MCS to obtain a set of $14\,696$ MCS which exist during at least one time step. Note first that if a MCS c exists at time t and that a link is added to it at time $t+1$, then the MCS c ceases to exist at time $t+1$. Therefore we are going to consider in the following that MCS appear or disappear even if there is only small modifications of the underlying topology. A MCS is stable if and only if no node and no link appears or disappears.

Fig. 6(a) displays the distribution of the size of the MCS for, respectively, the number of nodes and links. On these figures, one can note that there is a lot of MCS of all size in terms of number of nodes while MCS with many links are fewer. This can be understood together with the Fig. 6(b) which displays a scatter plot of the size of the MCS. As for the Fig. 2(b) one can observe a positive correlation between the number of nodes and links in a MCS with a nonconstant variation of the number of links. In this case, the variation factor is around 4.5 which means that for a given number of nodes, one can expect a variation of density of the same factor.

The stability of these MCS is another crucial point: it is important to know whether these MCS stay alive for a long time and if some of them can disappear and reappear in the future. In the following, the total lifetime of a MCS c is defined as the number of timesteps for which c exists, and the number of apparitions of c is the number of times this MCS is nonpresent at a given time and present at the next timestep. To give more precise results, one could have looked at the distribution of presence durations rather than these two aggregated parameters. However, this would give a distribution for each MCS which is harder to study.

Fig. 7(a) and 7(b) display the distribution of the total lifetime and the number of apparitions of all MCS. The main result from these curves is that there is a strong heterogeneity for both parameters: while more than one half of the MCS exist only during one time step, some of them exist during nearly half of the whole time. Notice that the most frequent MCS is just a couple of nodes which are certainly roommates since they are connected and isolated from the rest of the network every night. Again the MCS which appear and disappear frequently are very small: couples or triples of nodes.

To give a better insight on this last remark, Fig. 8(a) and 8(b) are scatter plots for the total lifetime and number of apparitions as a function of the number of nodes. While not displayed, the same figures for the number of links give very similar results. On both figures, the main results concern the absence of stability of large MCS: there is no MCS with more than 12 nodes which have a lifetime greater than 100 seconds and there is only 17 MCS of size greater than 8 with such a lifetime. The more links a MCS contains the more potential modifications may happen, which explains in part the curves. The reason is that the probability of link creation in a MCS is as greater as the number of nodes in that same MCS.

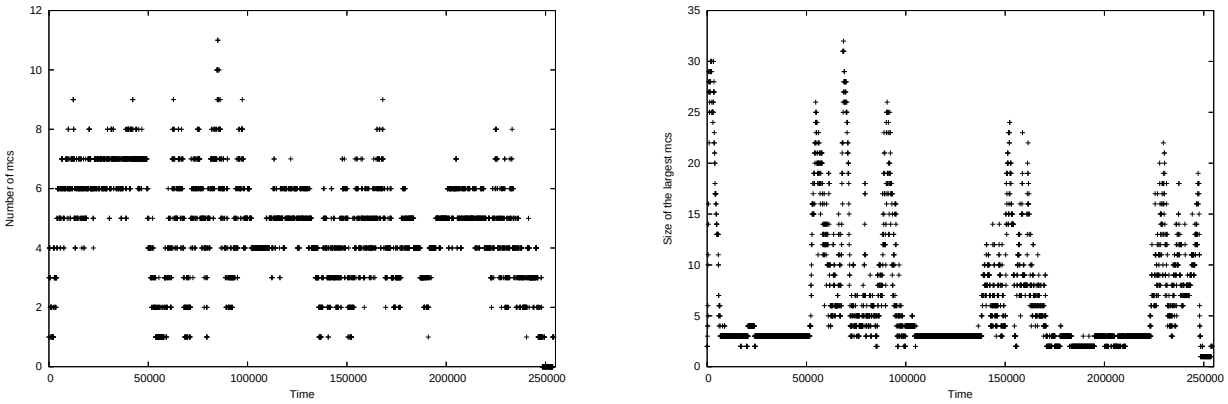


Fig. 5. Evolution of the number of MCS (left) and evolution of the size of the biggest MCS (right).

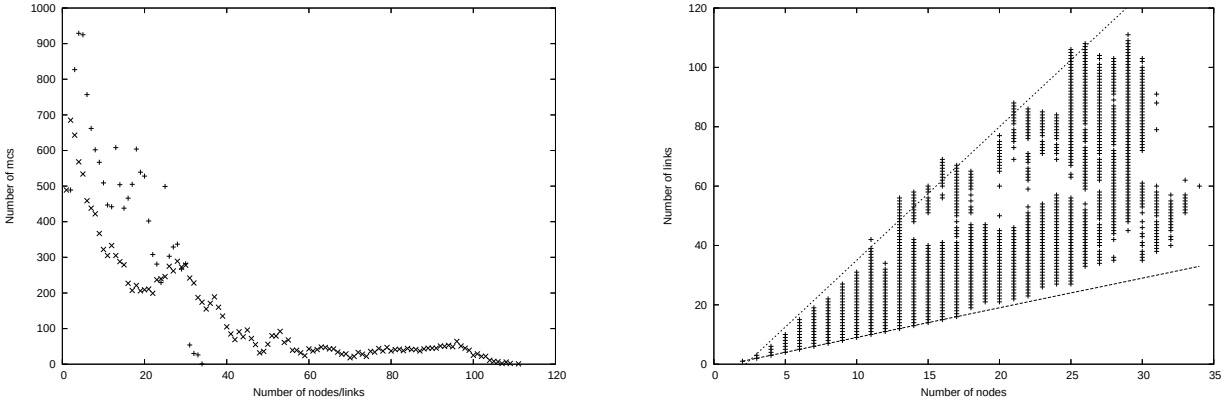


Fig. 6. Distribution of the number of nodes (+) and links (x) (left, a) and scatter plot of nodes versus links (right, b) of all MCS. For the right figure, $y = x - 1$ and $y = 4.5x - 10$ are drawn.

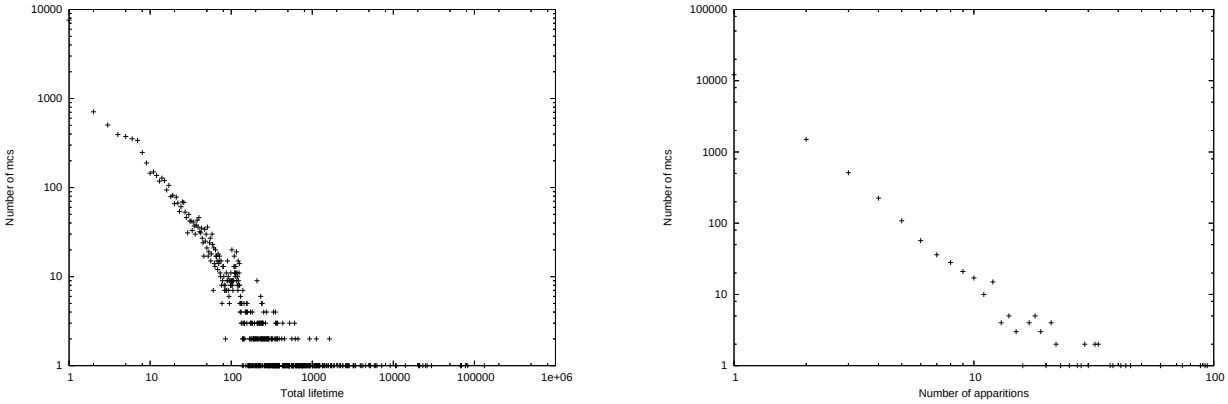


Fig. 7. Distribution of the total lifetime (left, a) and of the number of apparitions (right, b) of MCS.

Fig. 8(b) might let us think that some big MCS appear regularly more than once. However, if we define that a MCS reappears only if it has disappeared for more than five minutes (resp. more than one hour), then only MCS of size strictly lower than 8 (resp. 5) appear more than once. This means that most MCS reappear very soon after they have disappeared. The main reason is the following: suppose that a MCS c exists at time t, then at time $t+\delta$ a link is added between two nodes of c which creates MCS c' and this link disappears at time $t+\gamma$. Therefore the MCS c will be seen as absent for some time. The flickering of this link might be due to the movement of a node or to a failure in the measurement. Either ways, even if both nodes are near, the protocol ignore this link which cannot be used to transmit information.

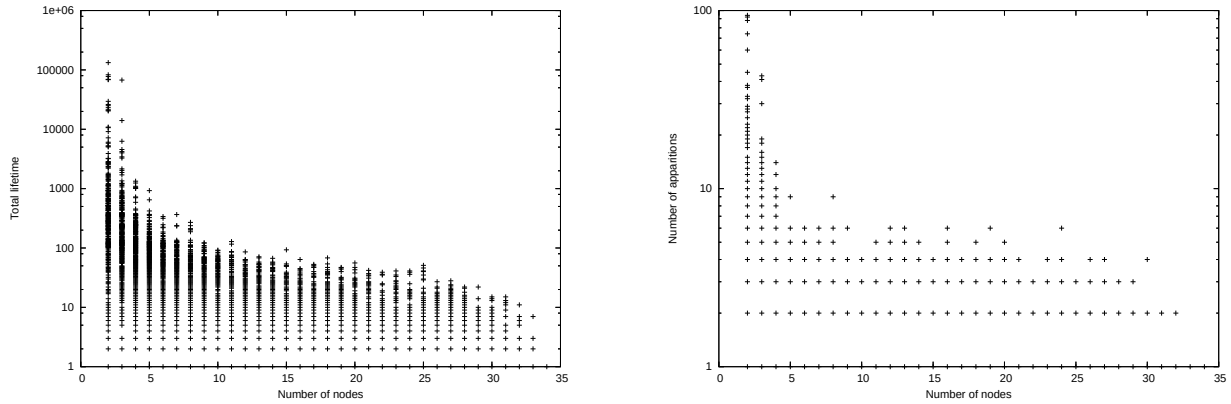


Fig. 8. Total lifetime (left, a) and of the number of apparitions (right, b) of MCS as a function of the number of nodes.

One may object that two MCS c and c' which are the same but for some links, *i.e.* which share the same nodes, might be assumed as equal. This would lower the number of apparitions of large MCS and increase their lifetime. However we believe that every change in the topology might induce a change in a given transmission protocol and therefore the notion of MCS is more precise. Notice also that given a set of nodes there is generally few, rarely more than 10, MCS constructed on this set of nodes.

The main conclusion of this subsection if that the dynamical effects observed at a global scale are also present in large MCS. These MCS have a very short lifespan and one cannot expect that they might reappear in the future. On the contrary, small MCS are generally more stable and have a probability of reappearance much higher.

B. Data mining

In this section we apply Data Mining techniques that have filtering capabilities on the collected data in order to identify social groups and describe the dynamic of individuals among these groups. The process we use is composed of two main steps. We first gather information on groups of links over time using connected and frequent subgraphs on the data. We then filter the obtained subgraphs using a density criteria to smooth the data and leverage most important and established subgraphs. By doing this we are able to take into account the time variability of the information gathered by the sensors. The second step goes back to individuals that are present in the resulting subgraphs. We merge the groups that concern similar individuals and time steps to obtain social groups. Finally, we built the dynamic trajectories of individuals by considering for each individual the social groups he/she belongs to and order them with respect to time to obtain the trajectories.

Frequent connected subgraphs as representation of groups: The imote data are the result of experimental measurements that leads to erroneous or incomplete data. The main problem is that some edges flicker. To cope with this problem we reduce the time dimension by considering that an edge exists if it appears at least ones during a 240 seconds period[4]. Such period corresponds to a time step in the following.

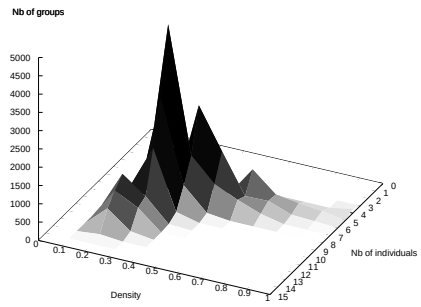


Fig. 9. Number of frequent connected subgraphs w.r.t. their density and their number of vertices (*i.e.*, number of individuals).

We want to compute on these data all maximal connected subgraphs that are sufficiently frequent (*i.e.* exist during at least 10 time steps) and sufficiently significant (*i.e.* contain at least 5 edges). We use the D-MINER solver [11] on the resulting matrix that associates to each edge the time steps for which it exists. D-MINER computes the whole collection of formal concepts (maximal rectangles of 1 values up to row and column permutation) in this boolean matrix. Frequency and significance are represented using minimal size constraints of 5 on the edge set and 10 on the time step set. The resulting 66 328 subgraphs are then processed to extract their connected component that are included in the formal concepts. We obtain this way a set of 23 316 frequent connected subgraphs having in average 7.66 vertices, 8.59 edges and appearing in 12.41 time steps.

Fig. 9 shows the distribution of the frequent connected graphs with respect to their density and the number of vertices (individuals) they cover. Most of the graphs only cover a

[4]This period corresponds to the sleep period between two successive hello packets in the neighborhood discovery protocol.

small set of individuals with a low edge density. These two criteria, when combined, show that the relationship between individuals in those groups might be understood as a relation chain instead of a strongly connected group.

For the next step of our study we only keep graphs that are dense enough to be considered as social groups. The filtering we apply using this simple criterion allows us to only keep 281 graphs using a 0.8 density threshold. This size reduction is mandatory if we want to consider relationships among individuals.

Individual trajectories among the social groups: The first step gives us 281 dense connected subgraphs, but they actually cover the same sets of vertices many times. Some of these groups are similar to each other in that they differ by only a few individuals, time steps, or are subsets of bigger groups. Going back to the formal concepts theory described in section II-A we can use the Galois connection principles to detect inclusions of groups within the set we obtained. Indeed, the Galois connection states that if two concepts (A, B) and (C, D) are such that $A \subseteq C$ then $D \subseteq B$.

As the connected graphs are dense, it is meaningful to associate to each graph the set of vertices (individuals) it covers. In the following we consider groups of individuals associated to a time step set. We merge two such formal concepts (A, B) and (C, D) if the set of vertices of A is included in the one of C and the time steps of $B \setminus D$ are close to a time step of D. In our experiment the allowed time distance used to merge two formal concepts is set to one time unit (240s).

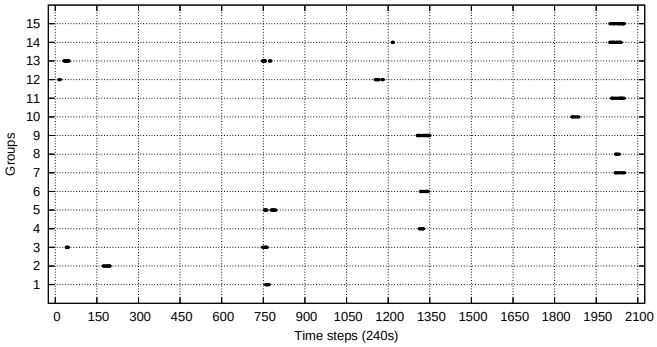


Fig. 10. Formal Concepts with respect to Time.

Fig. 10 represents the 15 groups of vertices with respect to time that have been obtained by applying the formal concept merging procedure. The group number identifies a formal concept while points on the time direction represents its time step set. Fig. 11 represents the individual set component (*i.e.* vertices in the graphs) of the formal concepts. These groups can be considered as social groups.

The last step of our procedure is to go from formal concepts back to individuals. The combination of Fig. 10 and 11 allows us to derive trajectories of individuals among groups during the experiment. We can follow the trajectory of individuals in groups as presented on Fig. 12. Dashed boxes are individuals

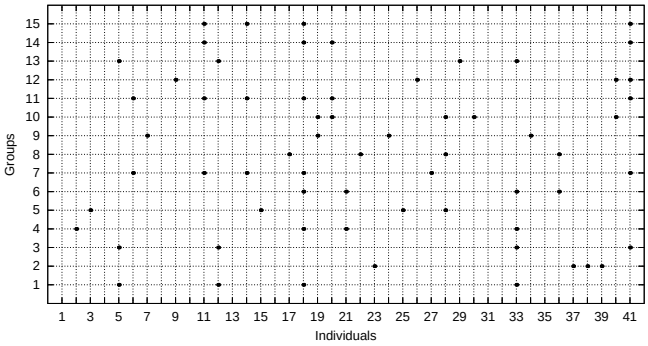


Fig. 11. Formal Concepts with respect to Individuals.

entering a group and edges are labeled by the individual number when he/she goes from one group to another. For example, individual 19 which enters group 13 at time step 1215 (given by Fig. 10) goes to group 9 before entering group 10.

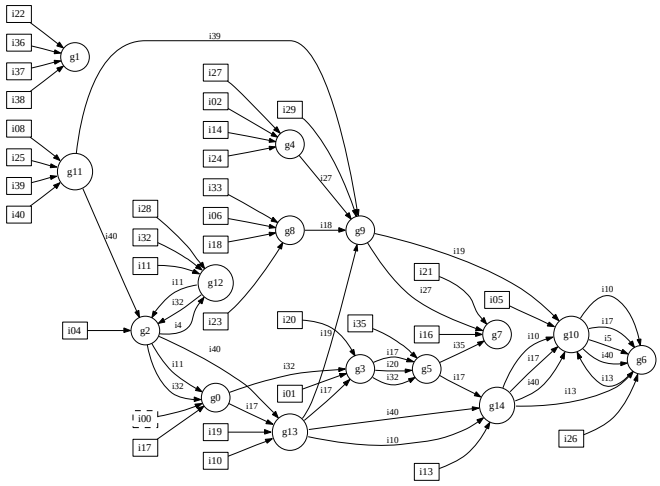


Fig. 12. Individual trajectories in groups ordered by time. ixx are individuals while gxx denotes social groups.

We have proposed in this section a way to use data mining techniques to analyze dynamic graphs. These techniques use exhaustive methods and algorithms but still require a supervisor to fix several thresholds and parameters to drive the graph structural exploration. Despite these manual interventions, the proposed methods are used within a formal framework that structures the data and offers some guarantees on the output.

V. Random model

In this section, we introduce a simple random dynamical model which is aimed at showing the limits of considering only inter and intra duration time. Our aim here is not to use a complex or realistic model but mainly to give insights on the structure and the dynamics of a contact network and to show that the structure and the dynamics are not a direct consequence of the intra and inter contact durations. Note also that real contacts are constrained by the bidimensional aspect

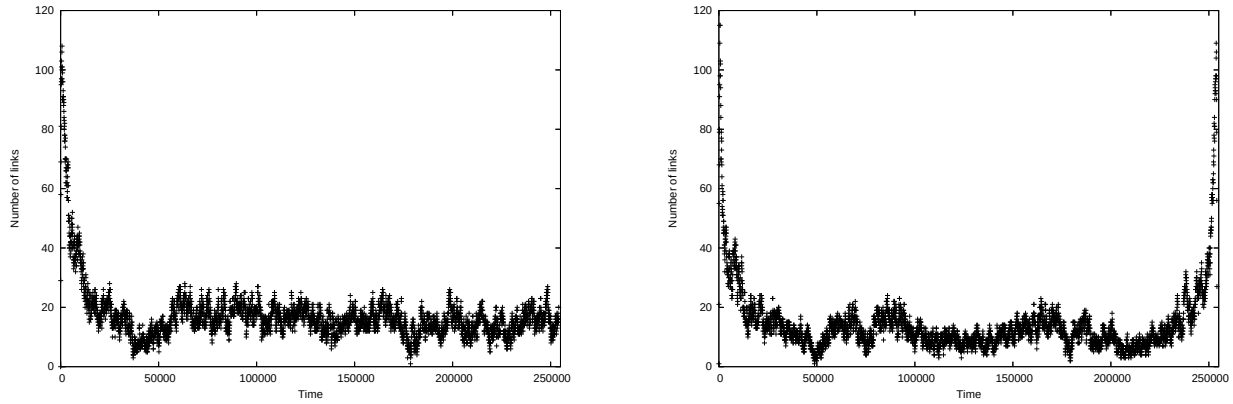


Fig. 13. Evolution of the number of links for the global model (left, a) and the link model (right, a).

which prevents some graph structures to appear, the model used in the following is not: any subgraph can potentially appear.

Following classical random models for static graph, such as the configuration model [20], we propose the simplest dynamical model possible with respect to the inter and intra duration time distributions. These distributions are extracted from the real data to obtain two distributions of durations. Then, we will consider each link in turn, and for each of these links we will generate a on/off sequence using the real intra and inter contact duration with a total duration time similar to the real one. Every duration is drawn independently from previous durations and from other links.

Note that we mentioned the distributions used by this model. There is at least two simple ways to generate a random dynamic graph, each relying on a distribution for both intra and inter contact durations. Let $d^+_{(u,v)}$ be the distribution of contact duration for the link (u, v), that is to say that $d^+_{(u,v)}(k)$ is the probability that a duration time last for k time units. Let also d^+ be the distribution for all links in the graph, which is in some way an union of $d^+_{(u,v)}$ for all links (u, v). In a similar way, $d^-_{(u,v)}$ and d^- are the distributions of inter contact durations for links or the whole graph.

Using these definitions, the on/off sequence for a given link (u, v) generated by the model can follow either a combination of $d^+_{(u,v)}$ and $d^-_{(u,v)}$, or a combination of d^+ and d^-. In the first case, the intrinsic nature of each link is kept by the model. For instance if one link is always off in reality, it will always be off with the model. Therefore a random graph generated by the model is just, for each link, a permutation of on and off periods of the original periods for this link. In the last case the intrinsic nature of each link is lost since one uses only one global distribution for the whole network: every link has a on/off sequence drawn from the same sequence. In the following we will refer to the first model as the link one, while the second will be referred as the global one.

A trade-off, but more complex, approach would be to consider distribution for a given node. One idea would be to define d^+_u to be the distribution of contact duration for all links which are ending at u. However the on/off sequence for a link (u, v) would be constructed from a combination of d^+_u, d^+_v, d^-_u and d^-_v in way to be defined.

For both link and global models, a random dynamic graph generated using these model gives, for each time step, a vision of the contact graph. Note also that even if we will use this model with the experimental parameters (number of nodes, distributions of durations and total duration) there is no restriction on these parameters which means that this model can be used as a basic comparison tool for any kind of experimentation.

In the following, we will compare the results from the previous sections with similar experiments made on random dynamical graphs with the same size, first for basic properties and then using connected components obtained with both link and global models.

A. Basic properties

Fig. 13(a) and 13(b) detail the evolution of the number of links for both models. These models present a strong peak of links at the beginning only for the global model and at both ends for the link model. This comes from the random nature of these models. While both models and the real data have nearly the same inter contacts average duration, the repartition is clearly not the same over the whole interval, see Fig. 16. In real data, the average inter contact duration at the beginning or at the end of the experiment is three to five times greater than the average over the whole period. On the contrary, for the link model for instance, the first and last values are much lower than on the real data and not so far from the average value, therefore there is a lot of links present at the beginning or at the end in the random networks until some of them disappear for a long time.

This could be seen as a drawback, however it reveals the nature of simple random models which are not able to capture such specific behaviors such a daytime/nighttime. In the following this specific structure will have some effects which can clearly be identified. However, note that our goal is to show that much work have to be done on the analysis of the

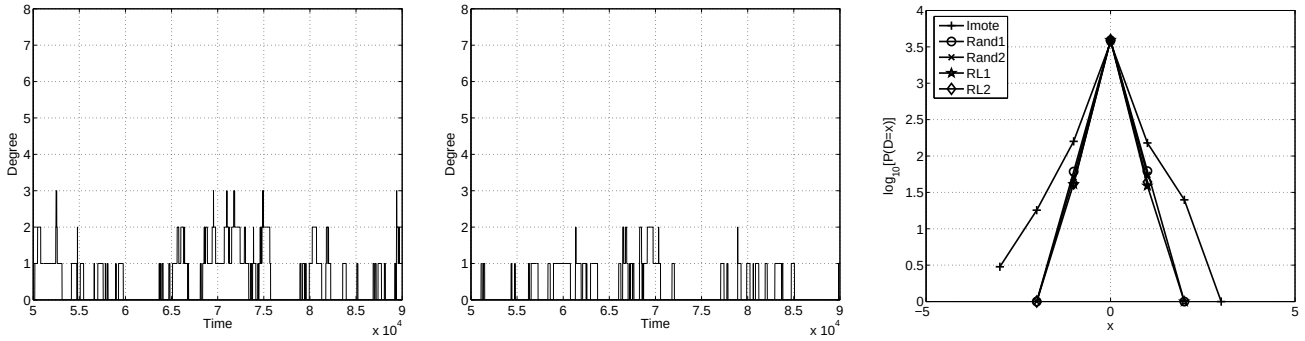


Fig. 14. Comparison between model and original data (for a particular node). Global model, degree evolution (left, a). Link model, degree evolution (middle, b). Probability distribution of various differential sequences (right, c).

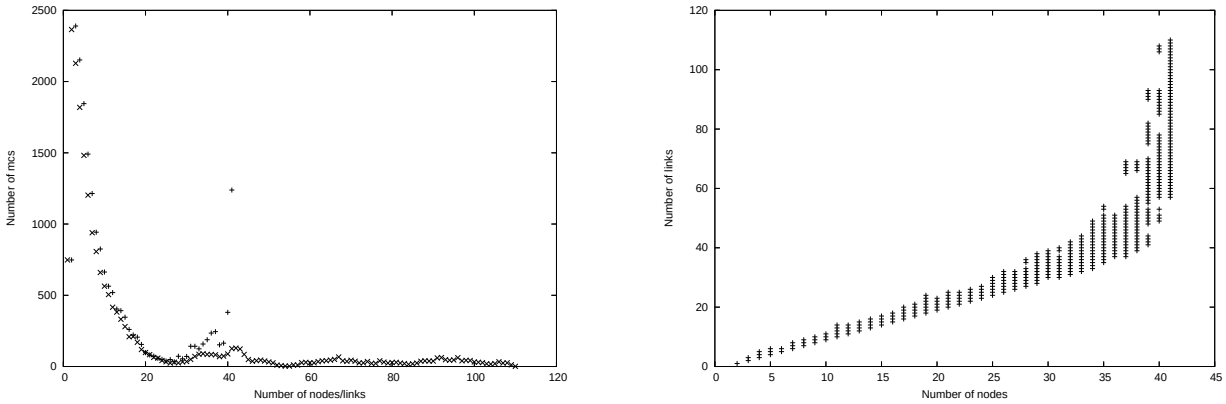


Fig. 15. Distribution of the number of nodes and links (left, a) and scatter plot of nodes versus links (right, b) of all MCS.

	real	global	link
first	27818	10886	16634
last	50479	46338	17219
average	10712	9035	10712

Fig. 16. Inter contact durations of the real data compared to global and link models. The first line display the average value of the first inter-contact duration for all links, the second line the average value for the last inter contact, the last line the average inter contact value on the whole period.

dynamics of the such networks to have a better understanding of these dynamics and maybe propose better models.

B. Degrees

Fig. 14(a) and 14(b) show respectively the degree evolution of a particular node in a simulation of each type of dynamic graph introduced in Section V (global model and link model).

When compared to Fig. 4(a), we can see that the variability of the degree evolution seems to be much lower in the simulation than in the original data. In order to show evidence for that, Fig. 14(a) shows the probability distribution function of the differential sequences for different simulations and the original data. The distribution for all random dynamic graph simulations is similar, but rather different than that of the original data.

This shows that a random graph model based on contact and inter-contact durations do not manage to reproduce accurately the variability of the degree evolution.

C. Connected components

In this section, we are going to concentrate on the global model only since both models present very similar behaviors. In order to avoid displaying more figures, a number of experiments have been made to show that the models behave in many ways like the original data:

- the number of distinct MCS for the real data and both models are very near;
- the distribution of the lifetime and number of apparitions of MCS also have a very similar shape as what can be observed on Fig. 7(a) and 7(b);
- the distribution of the lifetime and number of apparitions of MCS as a function of their number of nodes (see Fig. 8(a) and 8(b)) are also quite similar in the sense that no large MCS have a long lifetime or a large number of apparitions.

Concerning the last point, notice that the distributions of the sizes of MCS are not similar for real data and models. As it can be seen on Fig. 15(a) and 15(b), first of all there is some MCS containing all the links while real data do not contain any MCS of size greater than 34, and there is quite a large number of such complete MCS. However the most important point concern the Fig. 15(b) which is a scatter plot of the number of nodes versus the number of links. Compared

to Fig. 6(b) for the real data, it seems that, except for the few largest MCS, all the other MCS have a number of links which is nearly linear in the number of nodes. The few largest MCS are generated at the beginning when there is many links in the network.

The fundamental structure of the MCS obtained by the models is therefore completely different in the sense that the density is nearly always low in such MCS. Therefore at a given time the real interaction network might be seen as a set of very heterogeneous subnetworks in the sense of the size but also of the density, while data obtained by the model only capture the heterogeneity of size: at a given time, a random network is almost a set of disjoint trees. This has in particular a strong impact on the diameter (the maximal distance between two nodes of a MCS) since random MCS have in average a higher diameter and there exists some MCS with a much higher diameter as what is encountered in real data.

VI. Conclusions

We do believe that in order to be able to derive efficient algorithms and protocols for dynamic wireless networks, it is mandatory to know the underlying networks, their characteristics and how they evolve in time. Such basic knowledge is fragile nowadays but we did learn thought preliminary empirical studies and analytic approaches that real networks are far from being purely random. First empirical studies on real data where a first step but we show in this paper that the intrinsic characteristics of dynamic wireless networks cannot be totally captured by modelling only the inter and intra contacts through a simple power law.

One main conclusion of our preliminary results shows the intrinsically heterogeneous nature of maximal connected subgraphs in real networks. It appears also that components of individuals are playing a key role in the dynamics of the underlying topology of the networks observed.

The second major result is that the combination of approaches used reveals to be promising and we will pursue in this direction since both data mining and random processes are a great help. We can push even further this argument in the way that we need to address interdisciplinary issues, both in the computer science domain but also in others branches: sociology, epidemiology, and statistical physics.

Nevertheless, some points remains for future extension of this works. Based on the dynamic structure discovered like MCS, we are currently considering dissemination services for dynamic WSN environments. The dynamic of the network is clearly an issue that could affect the choice of the dissemination heuristic. The study of the latency and the way of limiting message duplications should be studied in forthcoming works.

Finally, to complete our work and in order to take advantage of the three powerful methods used, we are launching larger scale experiments in terms of the numbers of individuals (200 mobile data loggers wore continuously by students on a campus), the number of communities engaged (students are dispatched in two engineer departments and over 3 class levels), and the duration of the experimentations (1 month long). We hope that gathering further and larger in situ results will allow to deeply extend our understanding of dynamical networks. By combining several approaches we have proposed and experimented some tools in order to characterize generic topological and dynamical principles, but we hope that new tools will emerge in the near future in order to undertake studies that cross disciplinary boundaries.

References

[1] K. E. Fox, "Revisiting "scale-free" networks," *BioEssays*, vol. 27, no. 10, pp. 1060–1068, 2005.

[2] R. Albert, H. Jeong, and A. Barabasi, "The diameter of the World Wide Web," *Nature*, no. 401, pp. 130–131, 1999.

[3] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, no. 286, pp. 509–512, 1999.

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power law relationships of the internet topology," *Computer Communication Review*, no. 29, pp. 251–262, 1999.

[5] A. Broder, S. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks*, vol. 33, no. 1-6, pp. 309–320, 2000.

[6] G. Johnson, "First cells, then species, now the web," *NY Times*, pp. Section F, Page 1, Column 5, December 26 2000.

[7] A. Chaintreau, J. Crowcroft, C. Diot, R. Gass, P. Hui, and J. Scott, "Pocket switched networks and the consequences of human mobility in conference environments," in *ACM SIGCOMM 1st workshop on delay tolerant networking and applications (WDTN 2005)*, 2005.

[8] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings ACM SIGMOD'93*. ACM Press, 1993, pp. 207–216.

[9] H. Mannila and H. Toivonen, "Levelwise search and borders of theories in knowledge discovery," *Data Mining and Knowledge Discovery journal*, vol. 1, no. 3, pp. 241–258, 1997.

[10] C. Lucchese, S. Orlando, and R. Perego, "Dci_closed: A fast and memory efficient algorithm to mine frequent closed itemsets," in *Proceedings of the Workshop on Frequent Itemset Mining Implementations*, 2004.

[11] J. Besson, C. Robardet, J.-F. Boulicaut, and S. Rome, "Constraint-based concept mining and its application to microarray data analysis," *Intelligent Data Analysis*, vol. 9, no. 1, pp. 59–82, 2005.

[12] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes, 3rd edition*. Oxford University Press, 2001.

[13] A. Chaintreau, J. Crowcroft, C. Diot, R. Gass, P. Hui, and J. Scott, "Impact of human mobility on the design of opportunistic forwarding algorithms," in *INFOCOM 2006*, 2006.

[14] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, and W. Willinger, "Towards a theory of scale-free graphs: Definition, properties, and implications," *Internet Mathematics*, vol. 2, no. 4, pp. 431–523, 2005.

[15] P. Doukhan, G. Oppenheim, and M. Taqqu, Eds., *Theory and Applications of Long-Range Dependence*. Birkhäuser, 2002.

[16] M. Babu, N. Luscombe, L. Aravind, M. Gerstein, and S. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Curr Opin Struct Biol.*, vol. 14, no. 3, pp. 283–291, 2004.

[17] M. Babu, L. Aravind, and S. Teichmann, "Evolutionary dynamics of prokaryotic transcriptional regulatory networks," *J Mol Biol.*, vol. 358, no. 2, pp. 614–633, 2006.

[18] A. Vazquez, J. Oliveira, and A.-L. Barabasi, "The inhomogeneous evolution of subgraphs and cycles in complex networks," *Physical Review E*, vol. 71, p. 025103, 2005.

[19] P. Abry and D. Veitch, "Wavelet analysis of long-range dependent traffic," *IEEE Trans. on Info. Theory*, vol. 44, no. 1, pp. 2–15, Jan. 1998.

[20] B. Bollobas, *Random graphs*. Academic Press, 1985.