On the Structure of Changes in Dynamic Contact Networks

Vincent Neiger[1], Christophe Crespelle[2] and Eric Fleury[3]

Abstract

We present a methodology to investigate the structure of dynamic networks in terms of concentration of changes in the network. We handle dynamic networks as series of graphs on a fixed set of nodes and consider the changes occurring between two consecutive graphs in the series. We apply our methodology to various dynamic contact networks coming from different contexts and we show that changes in these networks exhibit a non-trivial structure: they are not spread all over the network but are instead concentrated around a small fraction of nodes. We compare our observations on real-world networks to three classical dynamic network models and show that they do not capture this key property.

1 Introduction

During the last decade, the study of large scale networks has attracted a large amount of attention and works from several domains: sociology [12], biology [7], computer science [1], epidemiology [8]. Complex networks have become a new area of research. The availability of large data sources on human mobility [9] has opened new perspectives for investigating the interplay of social networks, human mobility and dynamical processes. Then, it becomes crucial to study the evolution in time of complex networks, i.e. their dynamical aspect. Indeed most complex networks change: new edges appear while some other disappear. In all the scientific domains cited above, the dynamics is an intrinsic property: people make new acquaintances, change their relations, communication links fail, etc.

A dynamic network is made of interactions between entities occuring at different times. In the networks considered here, interactions are contacts occuring between sensor devices. We refer to this type of networks as dynamic contact networks. And as usual, entities of the network will be referred to as nodes or vertices and interactions, or contacts, as links or edges. A very common way to describe such a dynamic network is to use a graph series: a series of network snapshots taken at different times. These

[1]ENS de Lyon, Université de Lyon – `vincent.neiger@ens-lyon.fr`

[2]Université Claude Bernard Lyon 1, DNET/INRIA, LIP UMR CNRS 5668, ENS de Lyon, Université de Lyon – `christophe.crespelle@inria.fr`

[3]ENS de Lyon, DNET/INRIA, LIP UMR CNRS 5668, Université de Lyon – `eric.fleury@inria.fr`

snapshots are obtained by aggregating all interactions on a given time period, that is, by forming the graph of the interactions which occurred during this period. Doing so for fixed-length and disjoint time periods covering the whole period of experiment, one obtains a series of graphs describing the dynamics of the network. Then, many works dedicated to analyse or model such dynamics focus on the structural properties of the obtained snapshots. This reveals very useful information, but also suffers from a clear limitation. Since all snapshots are considered independently, this approach is unable to capture a key information about the dynamics: the relationships that link consecutive graphs in the series. This information is essential to understand the very evolution of the network, which is more than a simple juxtaposition of its states at different times. This is why some works designed models that introduce some form of temporal correlations: for example, [5, 3] takes into account the probability of presence of an edge depending on whether this edge is present at the previous step, and [9] respects the distribution of contact times (number of consecutive snapshots containing a given edge) and the distribution of intercontact times (number of consecutive snapshots not containing a given edge). As pointed out by [9], models based on these time parameters and only simple graph parameters, such as number of edges and degrees of nodes, are unable to properly reproduce the structure of real-world graph series regarding many aspects. Thus, there is a strong need for introducing non-trivial structural graph parameters for analysing and modelling the evolution of real-world dynamic networks. Here, we present a novel approach for this purpose and prove its relevance for the analysis of real-world dynamic contact networks.

Our approach

In this paper, we do not consider the snapshots of the network but the changes between two consecutive snapshots, which we call the *difference graph*. Formally the difference graph ΔG of two consecutive graphs G_1 and G_2 in the series has the same set of nodes[4] as G_1 and G_2, and for two nodes a and b, ab is an edge in ΔG if and only if either ab is an edge in G_1 and not in G_2, or ab is an edge in G_2 but not in G_1. In other words, the difference graph is the graph whose edge set is formed by all the pairs of nodes whose adjacency relationship changes between G_1 and G_2.

Difference graphs have already shown their interest for graph visualisation [2] and for IP traffic analysis [6]. The reason why we focus on difference graphs instead of the graphs of the series themselves is that those difference graphs contain a key information on the evolution of the networks: the correlation between two consecutive graphs of the series. This constitutes

[4]In this paper, we consider networks on a fixed set of nodes. This is natural in many contexts and this is not a limitation in general, as one can consider the set of nodes encountered during the whole life of the network.

a basic but fundamental part of the time correlations in the series, which is precisely the part used in the well-known *Edge-Markovian Model* [5, 3]. The originality of our approach is that we do not consider time correlations independently edge by edge like in the Edge-Markovian Model. Instead, we consider all the changes occurring between two consecutive steps of the series as a whole and we fully tackle the graph structure of this object. Moreover, we do not investigate this structure only by classic (and limited) graph parameters, such as number of edges and degrees (as in [6] which stresses the need for more advanced metrics), but we use a non-trivial structural graph parameter, called *Minimum Vertex Cover*, which, applied to difference graphs, capture key properties on the structure of changes in a dynamic network. More precisely, it allows us to determine whether the changes in the network topology are well spread all over the network or rather concentrated in some restricted parts of it.

We emphasize that our approach is purely based on the contacts between nodes of the network and does not assume any additional information on the location of nodes. Here, we are not interested in the mobility of nodes in some metric space but only in the consequence of this mobility on the structure of contacts in the dynamic networks, which is the key information in many contexts, such as communication networks or the spread of epidemics. The fact that our approach is purely contact-based is a very important feature that makes it applicable to a broad class of dynamic networks independently from the context where they come from. Indeed, in many cases, the changes in the links of the network are not the consequence of any mobility. Let us cite for example e-mail exchanges, online social networks, disruption tolerant networks (where links fail). Moreover, even when the changes are the consequence of some geographic mobility, the information on the location of nodes is often not available in data footprint. This is in particular true for the 5 data sets we use, where sensors do not record any information on the location, but only contacts occurring with other sensors.

Our contribution

We design a general methodology to analyse the concentration of changes in dynamic networks, and we apply it to five real-world contact networks. Our approach reveals striking common properties for contact networks coming from various contexts (Sec. 5): changes affect only a fraction of nodes at each step (Sec. 2), and these changes are concentrated around a small fraction of those nodes (Sec. 3). In other words, at each time step, it is possible to find a very restricted set of nodes of the network that can be held responsible for changes, while other nodes do not change the adjacency relationships between them. A striking point of our results is that these observations hold regardless of the time scale (aggregation period) that has been chosen to study the network (Sec. 4). Finally, we show that these concentration properties are not captured by classical models of dynamic networks (Sec. 6), which emphasizes the need for models that take them into account.

Data sets

We use five different data sets to which we apply our methodology (see Sec. 5). All these data sets were obtained using the same technique: contacts between participants to an experiment were recorded using sensor devices carried by the participants and that send signals at fixed time intervals (sampling period). Those signals include the ID of their source device which is recorded together with a time stamp by devices that are close enough from this source. These experiments took place in various contexts (see Sec. 5), ranging from a hospital to a rollerblade tour in Paris, and presenting very different properties regarding the struture of the changes that occur in the network.

In what follows, except in Sec. 5 which is devoted to the comparison of the different data sets, we use a single data set called *Infocom'06* [10] in order to illustrate and discuss our methodology. This data set contains the contacts between participants to an Infocom'06 workshop. In the case of Infocom'06, the network contains 78 nodes and the sampling period is 120s for a three-day measurement. As the network is almost empty during night time, we removed nights from the data set. We chose an aggregation period of 900s (15mn), which we consider to be average-to-long: within 15mn, many changes can occur in such a human group. It follows that this choice should not be too favorable to our conclusions on the property of concentration of changes. Sec. 4 specifically studies the impact of the aggregation period on our observations.

2 Number of nodes involved in changes

The first quantity we examine in order to study the concentration of changes in the network is the number of nodes that are affected by changes between two consecutive graphs of the series, that is nodes whose neighbourhoods in the two graphs are different. This is exactly the number of non-isolated nodes in the difference graph (nodes having at least one neighbour). We compute this quantity for the difference series of the *Infocom'06* data set and depicts the results on Fig.1(top). First, we see that at any time in the series, some nodes are not concerned by changes: the number of non-isolated nodes in the difference series (dashed line) never reaches the total number of nodes in the network, its maximum value is 72 out of 78. Moreover, this number is usually much lower, it is less than 55 in average, i.e. 70% of the total number of nodes.

One may think that the fact that not all nodes are non-isolated simply comes from the number of edges in the difference graphs. Indeed, since we consider the difference between two consecutive graphs in the series, there may be only few changes in the adjacency relationships between nodes of the network. In this case, the number of edges of the difference graphs will be small and consequently some nodes will be isolated. In order to compare the result we obtain for the difference graphs of the real-world series to what can be expected for graphs of this density, we also compute, for each difference graph of the series, the number of non-isolated nodes in an *Erdös-Rényi random graph* (ER graph for short) having the same number of edges (see e.g. [4]). We will often use this series of ER graphs for comparison in the rest of the paper; we call it the *Diff-ER series*. The purpose of such comparisons is not to assess the fact that difference graphs of the Infocom'06 experiment are not random, which is clear. The rationale behind these comparisons is actually to determine whether the behaviour of the real graphs with regard to a given parameter (like the number of non-isolated nodes) is a consequence of its number of edges (i.e. whether graphs with this number of edges usually have a similar value for this parameter), or whether the real series reveal a particular behaviour with regard to this parameter. And in the latter case, it is still to know if the value of the parameter is higher or lower than what is expected and by how much. In all the article, because of space limitation, we present the comparisons with a single random choice. For each curve, we actually made ten random choices which all gave the same result.

Here, the comparison is striking: in the ER graphs having the same number of edges as the real difference graphs (plain line in Fig.1(top)), the number of non-isolated nodes is almost always 78 (the total number of nodes) and the average value is more than 75. This indicates that the result observed on the real difference graphs cannot simply be explained by their density. The edges of those graphs touch only a restricted part of the nodes (around 70% here) which is not at all what is expected for a graph having this number of edges. Thus, the difference graphs of the real series have a special structure in terms of concentration of their edges, which we investigate deeper in the next section thanks to Minimum Vertex Cover.

3 Number of nodes that concentrate changes

In this section, we push further the analysis of where changes are located in the topology of the network. For this purpose, we use a well-known graph parameter called *Minimum Vertex Cover* (MVC in the following).

3.0.1 Minimum Vertex Cover

In a graph, a vertex cover is a subset of nodes such that all edges of the graph are incident to at least one node in this subset. The Minimum Vertex Cover (MVC) is the minimum number of nodes in a vertex cover. We compute this parameter on the difference graphs of the Infocom'06 data set. The rationale

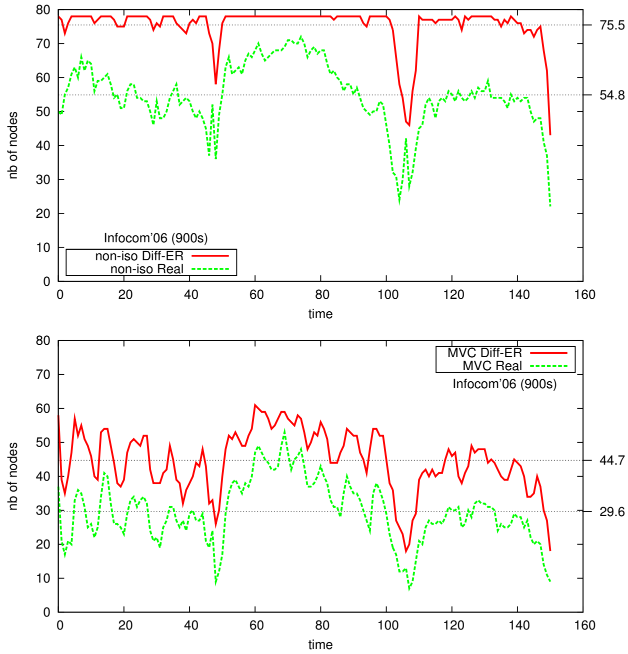

Figure 1: Top: number of non-isolated nodes in the difference graphs of the real series (dashed line) and in the corresponding Diff-ER series (plain line). Bottom: MVC value of the real difference series (dashed line) and of the corresponding Diff-ER series (plain line).

for doing so is that it gives us a minimal set of nodes that concentrate all the changes in the network. Consider the following situation: a contact network whose nodes are a group of people equipped with sensors (as in our data sets). Suppose that, within a time step of the series describing this dynamic network, all nodes stay still except one which moves and changes its adjacency relationships with all the other nodes (for example moving from a conversation group to another). In this toy example, the difference graph will be a star, made of all possible edges from the only moving node to all other nodes. Then, following the analysis of the previous section, all nodes of the network are affected by changes. But all these changes are due to the behaviour of a single node: the others stood still and did not change at all the adjacencies between them. This situation can be revealed by the MVC parameter: in this case the MVC of the difference graph is 1. Opposite to this example is the one where all nodes move in such a way that the adjacency between any pair of nodes is changed: in this case the MVC of the difference graph is $n - 1$, the number of nodes in the network minus 1. This is the meaning of the MVC of the difference graph: the minimum number of nodes that can be held responsible for the changes occurring in the network.

There is an important fact that one should keep in mind reading the subsequent analysis: if a graph has MVC equal to k then there exists a subset of $n - k$ nodes having no edges at all between them. In other words, if the MVC of some difference graph is 30% of the nodes in the network, it means that 70% of the nodes did not change at all the adjacency relationships between them.

Though computing the MVC of a graph is NP-complete, in practice, this does not constitute a real limitation to our approach. Indeed, the real-world graphs of the data sets we use do not present the pathologic cases that makes the problem difficult in theory. Therefore, thanks to the *leaf removal* technique, which is known to perform very well on sparse graphs [13] and which turned out to perform even better for the real-world graphs of our data sets, we could compute the exact value of the MVC, using a theoretically exponential algorithm, for all difference graphs in a quite reasonable time, even for graphs with a large number of nodes (more than 300 in the Mosar data set).

3.0.2 Results

We compute the MVC for the difference graphs of the real series and we compare it to the MVC of ER random graphs having the same number of edges (the Diff-ER series), see Fig.1(bottom). One can see that the MVC of real difference graphs is usually low (dashed line). The average value is less than 30 nodes, which corresponds to about 38% of the nodes in the network. This shows that only a relatively small number of the nodes concentrates all

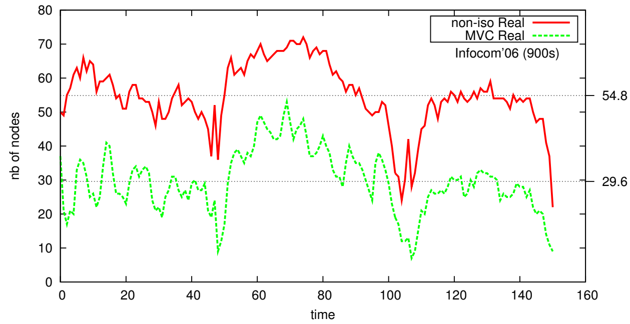

Figure 2: Number of non-isolated nodes (plain line) and value of the MVC (dashed line) for the difference graphs of the Infocom'06 data set.

changes that occur between a graph in the series and its successor. Again, this value has to be compared to what is expected for a graph with this number of edges. One can see that for the Diff-ER series (plain line) the value of the MVC is much higher: its average value is about 45, which is 50% higher than the value obtained for the real difference graphs. Thus, the property of concentration of changes around a low number of nodes in the real series is remarkable and denotes a very particular structure.

Now, we directly compare the value of the MVC and the number of non-isolated nodes, in order to determine, among nodes affected by changes, what is the fraction of nodes that can be held responsible for these changes and what is the fraction of nodes that simply undergo these changes. Fig.2 shows the evolution of these two quantities along the series. One can see that the value of the MVC (dashed line) is always much smaller than the number of non-isolated nodes. The average number of non-isolated nodes is around 55 while it is less than 30 for the MVC, i.e. almost twice less. This is a remarkable property in terms of concentration of changes: not only a relatively small number of nodes are involved in changes but also a much smaller part of them endorse the responsibility for those changes.

3.0.3 Further analysis

We saw that the number of non-isolated nodes and the MVC of the difference graphs are smaller than what is expected for graphs with this number of edges. A natural question is whether the latter property is a consequence of the former. To answer this question, we plotted on Fig.3 the MVC of the real difference graphs (dashed line) together with the MVC of ER graphs having the same number of edges and the same number of non-isolated nodes (plain line). That is, we generate an ER graph with the desired number of edges on a vertex set whose cardinality is precisely the number of non-isolated nodes in the corresponding difference graph. We refer to the obtained series as the

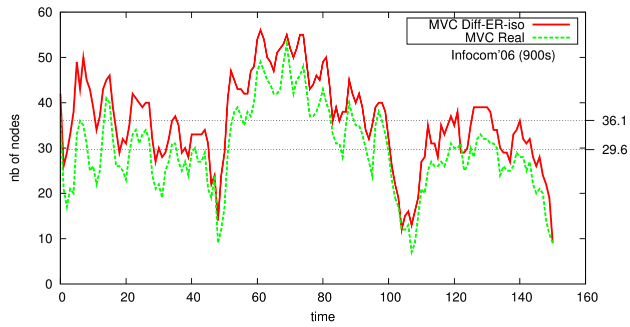

Figure 3: MVC of the Diff-ER-iso series (plain line) and MVC of the real difference series (dashed line).

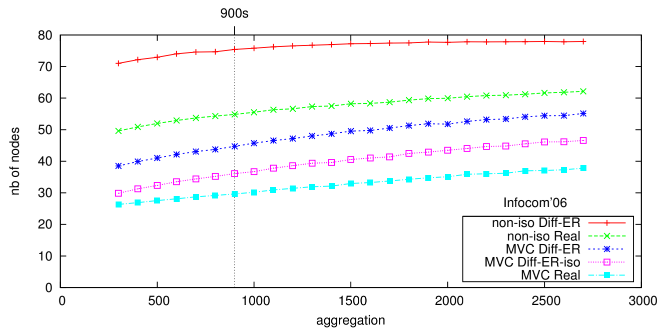

Figure 4: Average number of non-isolated nodes and average MVC value for the real difference series, the Diff-ER series and the Diff-ER-iso series.

Diff-ER-iso series.

One can see that, as expected, the difference between the MVC of the real difference graphs and the MVC of the ER graphs is smaller when the number of non-isolated nodes is respected (Fig.3) than when it is not (Fig.1(bottom)). But there is still a clear difference between the two series: the MVC of the real difference graph is always smaller than the MVC of the Diff-ER-iso graph, and the mean value of the MVC in the real series is less than 30 while it is more than 36 in the Diff-ER-iso series, that is about 20% higher. This shows that the special structure of the real series regarding the concentration of changes is not entirely captured by the sole number of nodes involved in changes. The MVC of the difference graphs are remarkable and are not a simple consequence of the number of non-isolated nodes, showing that this parameter should be taken into account for analysing and modelling these networks.

4 Influence of the aggregation period

At this point, one may have doubts about the generality of our results: do they hold for other dynamic contact networks? Aren't they particular to the aggregation period we chose? The former question is addressed in the next section. Here, we consider the latter and study the impact of the aggregation period on the concentration of changes in the network, both in terms of non-isolated nodes and in terms of MVC.

On Fig.4 we plotted the mean value of the number of non-isolated nodes and the MVC for the real difference series of the Infocom'06 data set, for different aggregation periods. For comparison, we also plotted the same curves for the corresponding Diff-ER series and Diff-ER-iso series. This gives a synthetic view of how the property of concentration of changes evolves when the aggregation period varies, from 300s (5mn) to 2700s (45mn). From these curves, one can retrieve the results presented in the previous sections for an aggregation of 900s (15mn). The plot shows that, as expected, when the aggregation period becomes smaller than 900s, the concentration of changes increases: there are less non-isolated nodes and the MVC becomes smaller. On the opposite, when it grows higher than 900s, the concentration of changes decreases. Nevertheless, it is striking to see that the concentration properties highlighted in Sections 2 and 3 for an aggregation of 900s still holds regardless of the aggregation period chosen between 300s and 2700s.

On Fig. 5, we plotted, for the extreme values of the aggregation periods (namely 300s and 2700s), the time evolution of the number of non-isolated nodes in the real series as well as the MVC in the real series, the Diff-ER series and the Diff-ER-iso series. One can observe that the qualitative conclusions of the previous sections remain valid for these extremal values. For 300s aggregation period, the average number of non-isolated nodes in the difference series is 49 (63%) and the average MVC is 26 (34%), while the average MVC for an ER random graph having the same number of edges (Diff-ER series) is 39, i.e. 46% higher. For 2700s aggregation period, the average number of non-isolated nodes and the average MVC are respectively 79% and 48% of the total number of nodes in the network while, again, the average MVC in the Diff-ER series is much higher than in the real series: 55 nodes instead of 38, i.e. 45% higher.

Note that studying such a contact network at a time scale of 45mn of aggregation is not completely natural: doing so, one loses a lot of information on the dynamics. Then, such an aggregation period should be considered more as a limit case to test our observations than as a suitable time scale to study the network. In particular, one could expect that for such an aggregation time, there are so many changes in the network that the properties of concentration highlighted at finer time scale may completely disappear. But Fig. 5(bottom) shows that this is not the case. It is striking to see that even in this range of time aggregation the topological correlations of changes in

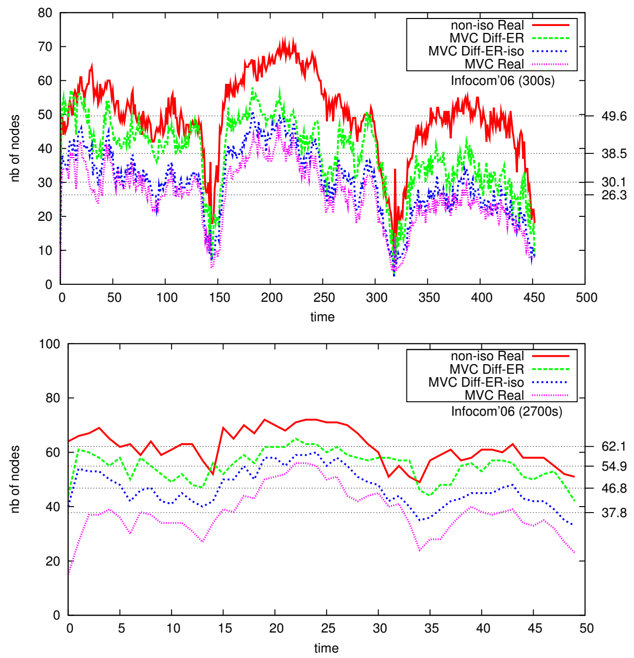

Figure 5: MVC value for the real difference series, the Diff-ER series and the Diff-ER-iso series, and number of non-isolated nodes for the real difference series, for aggregation periods of respectively 300s (top) and 2700s (bottom).

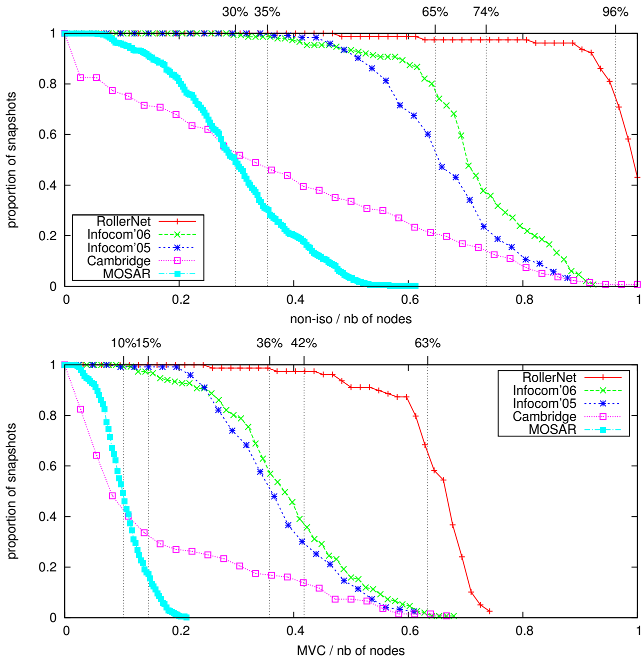

Figure 6: Normalized inverse cumulative distributions of the number of non-isolated nodes (top) and of the MVC (bottom) for the 5 datasets.

the network, i.e. their concentration around a small subset of nodes, are still clearly visible. This shows that this property should be taken into account whatever is the time scale used to study the network.

5 Results on different data sets

We consider five data sets, the Infocom'06 data set we used in the previous sections plus 4 new ones: *Infocom'05* [10] with 41 nodes, 120s sampling over three days (nights removed), on participants to a workshop of Infocom'05; *RollerNet* [11] with 62 nodes, 15s sampling over three hours, on participants to a rollerblade tour in Paris; *Cambridge* [10] with 36 nodes, 10mn sampling over two months, on students at the Cambridge University Computer Lab; and the *Mosar* project data set with 315 nodes, 30s sampling over two weeks, on staffs and patients in an hospital. Clearly, the mobility patterns of people in a conference, an hospital, a universitary campus or a rollerblade tour are very different, resulting in very different characteristics of the dynamics in the data sets we use.

Here we aim at comparing these data sets with regard to the concentration of changes. Does the property of concentration around a small number of nodes revealed for the Infocom'06 data set still hold in other contexts? Can our methodology be used to classify dynamic networks with regard to

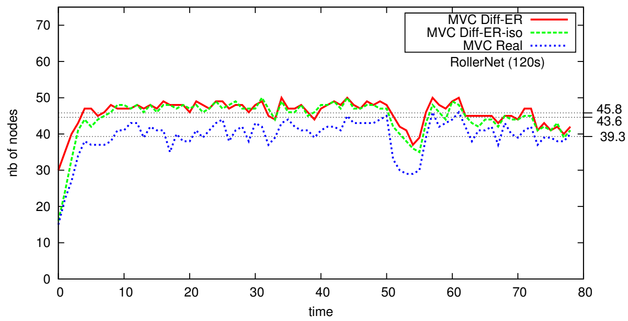

Figure 7: MVC values for the real difference series of the RollerNet data set, the corresponding Diff-ER series and the Diff-ER-iso series.

their properties of concentration of changes? In order to answer these questions, we compute the number of non-isolated nodes and the MVC for the difference series of the four new data sets. For each of them, we chose an aggregation period that we consider as average: 900s (15mn) for Infocom'05, 120s (2mn) for RollerNet, 2h for Cambridge and 40mn for Mosar. Remember that we saw in Section 4 that the choice of the aggregation period has a limited impact on the observations, and that the results are qualitatively preserved for a wide range of aggregation periods. Fig.6(top) shows the inverse cumulative distribution of the percentage[5] of isolated nodes for the difference series of the 5 data sets, while Fig.6(bottom) shows the inverse cumulative distribution of the MVC, expressed in percentage as well. Roughly speaking, we can classify the 5 data sets into three groups.

The first group is very homogeneous: the two Infocom data sets have very similar distributions. This is not surprising as they were collected in very similar contexts. More interestingly, it shows that the property of concentartion of changes still holds for a network with fewer nodes: Infocom'05 network has only 41 nodes while Infocom'06 has 78 nodes. This is worth of interest, as diminishing the number of nodes may make a small number of changes touch a wider part of the network and then threaten their concentration property. Fig. 6 shows that it does not happen (remember that curves are in percentage, not in number of nodes): the concentration of changes even in a network of limited size still appears as a key property.

The second group, containing only the RollerNet data set, is very different from the others: the number of non-isolated nodes is very high, since in average 96% of the nodes (i.e. almost 60 out of 62 nodes) are involved in changes from one snapshot to the following one. And the time series (not presented here) reveals that more than half of the time, all nodes of the network undergo changes in their neighbourhood. This is explained by

[5]Values are here expressed in percentage instead of number of nodes, in order to compare networks with different number of nodes.

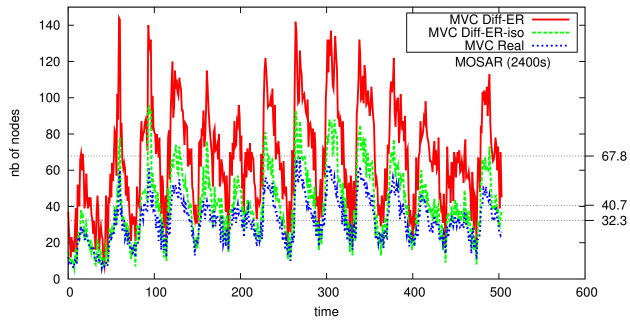

Figure 8: MVC values for the real difference series of the Mosar data set, the corresponding Diff-ER series and the Diff-ER-iso series.

the context where this data comes from: participants to the rollerblade tour move fast and constantly, and then are more likely to change their contacts very often. As a consequence, the MVC of the difference series is higher than in the four other data sets (see Fig. 6): it is around 63% of the nodes in average. Of course, this suggests that other parameters should be used to properly describe the structure of changes in this very specific type of dynamic contact networks, like in [11]. Nevertheless, even in this case, our approach reveals an interesting fact: even though almost all nodes are always affected by changes in their neighbourhood along time, the proportion of nodes that can be held responsible for those changes is significantly lower than what can be expected for a network with this number of changes. This is what is highlighted by Fig. 7 where one can see that the curve of MVC of the Diff-ER series is clearly distinct from the one of the real series (always above and about 10% higher in average). It must be clear that this sole fact is not sufficient to properly describe the structure of this specific dynamics. But on the other hand, this stresses the interest of the MVC parameter, which is able to reveal a special structure with regard to concentration of changes even in a context where changes affect the whole network at any time.

On the opposite, in the third group, the group of Mosar and Cambridge data sets, the proportion of non-isolated nodes and the MVC are almost always very low. The first reason may be that these data sets were collected in a wider-space environment and on a longer time period (2 weeks in a hospital for Mosar, 2 months in a universitary campus for Cambridge). This means that participants to these two experiments are more dispersed and have less frequent contacts, while in the Infocom experiments for example, the participants of the conference are gathered most of the time. This is the reason why we chose longer aggregation periods to study these data sets, namely 40mn for Mosar and 2h for Cambridge. Another difference is that we did not remove night periods in Mosar and Cambridge data, as it appears less relevant to do so for them. The Cambridge data set is very heterogeneous, the network is very sparse (or empty) at many times, and very often during daytime as well. For this data set, we pushed the aggregation period up to 8h, and the shape of the distributions remained very similar.

The time series for the Mosar data set are given on Fig. 8. Even though the day-night alternance is clearly visible for this data set, the network do not become empty during night, unlike in the Infocom experiments, resulting in a more homogeneous dynamics. This comes from the fact that the experiment was led in a specific service of the hospital which is active at night as well. The concentration of changes in this network is very sharp: in average, at each time step only 30% of nodes are concerned by changes and only 10% of them can be held responsible for those changes (see Fig. 6). This may be partly due to the large size of the network, which involves more than 300 nodes. But it should be clear that this strong concentration of the changes in the network is not a simple consequence of the amount of changes compared to the size of the network, but rather denotes a very particular topological structure. Indeed, as shown on Fig. 8, the MVC of the real difference series is always much smaller than what can be expected for graphs with this number of nodes and this number of edges (Diff-ER series): the MVC of the Diff-ER series is in average 110% higher than in the real difference series (67.8 nodes in average instead of 32.3), which has to be compared to 50% for the Infocom'06 data set (see Sec. 3 and Fig. 1(bottom)). Thus, the fact that the amount of changes at each time step in the Mosar network is very small compared to the size of the network is a critical property which results from the structure of changes itself. And it appears crucial to take this structure into account in order to properly describe the dynamics of such networks.

6 Comparison to classical models

We now compare the observations we made on real-world dynamic networks to classical models of graph series, in order to determine whether these models are able to reproduce the property of concentration of changes in the network. Let us emphasize again that here, for the sake of generality, we are interested in modellings of the dynamics that are based on contact information only and not on additional information concerning space location of nodes, as this information may not exist or may not be available in many contexts where dynamic networks appear (as it is the case for the data sets we studied). We use for comparison three different classical models of graph series. Each of these models takes as input some parameters of the real series and generate a random series having the same parameters. The first model, which we call the *ER series*, consists in generating a series of ER random graphs (see e.g. [4]), each of which having the same number of edges as the corresponding graph of the real primitive series (i.e. not the difference series). The ER series model is very simple but has a drawback: it produces uniform edge dynamics, in the sense that each pair u, v of nodes in the graph has the same expectation of its number of occurrences in the series.

The second classical model we use tackles this drawback: it preserves, in expectation, the number of occurrences in the series of each edge of the graph. We refer to it as the *Heterogeneous Edge series*, or HE series for short. More precisely, for any pair u, v of vertices which is an edge in N_{uv} graphs out of the N graphs of the real series, we put an edge between u and v with probability N_{uv}/N in each graph of the HE series. This does not result in a series where the density of each graph is the same as in the real series, but the average of densities over all graphs of the series is preserved, in expectation.

The last model we consider is the *Edge-Markovian Model* series [5, 3], or EM series for short. As the HE series does, it considers each pair of vertices independently. The first graph of the EM series is the same as the first one of the real series. Then, at each step, the model decides for each pair u, v of vertices to put an edge in the new graph depending on the stateof the pair (1 if there is an edge, 0 otherwise) in the previous graph in the following way: the probability to go from state s_1 to state s_2 ($s_1, s_2 \in \{0, 1\}$) in the EM model is set to be exactly the same as the probability observed on the real series. This model is very interesting for our purpose as it includes part of the correlation in time between the graphs of the series, namely the correlation of presence of an edge in two consecutive graphs of the series. As a consequence, it preserves, in expectation, the average density of the difference graphs over the whole series . And it also preserves, in expectation as well, the average density of graphs in the primitive series.

Number of non-isolated nodes

Fig.9(top) shows the inverse cumulative distribution of the number of non-isolated nodes for the real difference series of the Infocom'06 data set, the corresponding Diff-ER series, and the three models cited above (which were given as input the characteristics of the Infocom'06 series). Obviously, none of the three models reproduces correctly the number of non-isolated nodes in the real difference graphs. For the three of them, this number is very high, very close to the total number of nodes, for a vast majority of graphs in the series. While, on the contrary, this number is only 70% of nodes in average in the real series. Moreover, one can see that none of the model performs significantly better than the Diff-ER series to which we compare in Sec. 2. Thus, these models are not able to reproduce the fact that only a fraction of nodes are affected by changes at each step of the series.

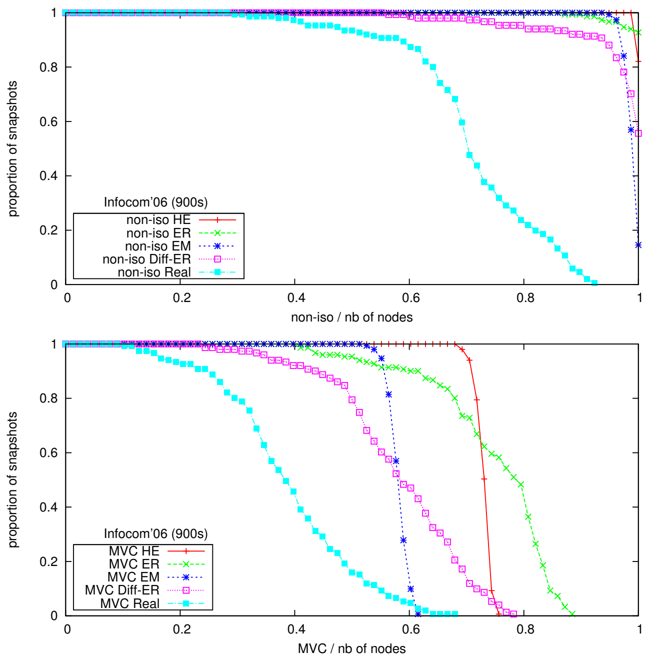

Figure 9: Inverse cumulative distributions of the number of non-isolated nodes (top) and of the MVC (bottom) for the real difference series of Infocom'06, the Diff-ER series and the difference series of the three models we consider.

Minimum Vertex Cover

Fig.9(bottom) shows the inverse cumulative distribution of the MVC of the difference graphs for the Infocom'06 data set, the corresponding Diff-ER series and the three models. One can see that for the MVC as well, the characteristics of the three models differ notably from those of the real series. The ER series and the HE series have a similar behaviour: for these two models the average value of the MVC of the difference graphs is around 57, instead of 30 in the real series, which is almost twice higher. The reason is that these models are based only on the probability of presence of edges in the primitive series: this sole property is obviously not sufficient to capture the structure of changes in the network.

The Edge Markovian Model gives better results. Indeed, the EM series has an average MVC around 45 nodes, which is 50% higher than the real series. Though it is far from reproducing the characteristics of the real series, it is a clear improvement on the two previous models. The reason is that the EM series respects the average density of the difference graphs, and not only of the primitive graphs. On the other hand, one can see on Fig.9(bottom) that the EM series does not perform significantly better, in terms of MVC, than the Diff-ER series, to which we compare in our previous analysis. This indicates, as one could expect, that, even though the Edge Markovian Model captures some of the time correlations of the edges in the series, the fact that it does not take into account the topological correlations of changes in the dynamic network does not allow it to properly capture the structure of the dynamics with regard to concentration of changes.

7 Conclusions and perspectives.

We designed a methodology to appreciate the concentration of changes in a dynamic network and we applied it to several contact networks, revealing their special structure with regard to this property: at each time step in the series describing the dynamics, the changes only affect a limited number of nodes and are concentrated around an even much smaller number of nodes. We showed that this property denotes a non-trivial structure by comparing to what can be expected for graphs with the same number of edges. Moreover, we showed that this holds in contact networks coming from contexts having very different spatial and time characteristics, and that this property appears independently from the time scale chosen to study the network. This shows that the concentration of changes is a non-trivial and fundamental charasteristics of these dynamics, which is not taken into account by most current modelling efforts, which concentrate mainly on time correlations of the dynamics.

The most immediate perspective of our work is to apply our methodology to other types of dynamic networks to determine whether they satisfy special properties with regard to concentration of changes. This is particularly appealing for networks that do not result from mobility in some metric spaces, like e.g. networks made of e-mail exchanges, telephone calls or online social networks.

Our work emphasizes the need for dynamic network models purely based on contacts that encompass the topological correlations of changes in the network. We argue that such models should be based on node characteristics, rather than only edge characteristics, in order to reproduce the concentration of changes around a restricted number of nodes of the network. The difficulty in doing so is that one must be able to select, at each step of the series, which nodes will change their neighbourhood and which adjacency relationships will be changed around them. We believe that this is a very promising research direction in which much remains to be done.

At last, let us mention that another key question arisen by our work is to study the impact of the concentration of changes on phenomena taking place over the network, such as diffusion of epidemics or information in human groups and routing in mobile networks.

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] Daniel Archambault, Helen Purchase, and Bruno Pinaud. Difference Map Readability for Dynamic Graphs. In *18th International Symposium on Graph Drawing*, volume 6502 of *LNCS*, pages 50–61, 2011.

[3] Hervé Baumann, Pierluigi Crescenzi, and Pierre Fraigniaud. Parsimonious flooding in dynamic graphs. *Distributed Computing*, 24(1):31–44, 2011.

[4] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

[5] Andrea E.F. Clementi, Claudio Macci, Angelo Monti, Francesco Pasquale, and Riccardo Silvestri. Flooding time in edge-markovian dynamic graphs. In *twenty-seventh ACM symposium on Principles of distributed computing*, PODC '08, pages 213–222. ACM, 2008.

[6] Marios Iliofotou, Prashanth Pappu, Michalis Faloutsos, Michael Mitzenmacher, Sumeet Singh, and George Varghese. Network monitoring using traffic dispersion graphs (tdgs). In *7th ACM SIGCOMM Internet Measurement Conference*, pages 315–320. ACM, 2007.

[7] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *ninth ACM international*

conference on knowledge discovery and data mining, KDD '03, pages 137–146. ACM, 2003.

[8] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physics Review Letters*, 86:3200–3203, 2001.

[9] Antoine Scherrer, Pierre Borgnat, Eric Fleury, Jean-Loup Guillaume, and Céline Robardet. Description and simulation of dynamic mobility networks. *Computer Networks*, 52(15):2842–2858, 2008.

[10] James Scott, Richard Gass, Jon Crowcroft, Pan Hui, Christophe Diot, and Augustin Chaintreau. CRAWDAD data set cambridge/haggle (v. 2009-05-29). Downloaded from http://crawdad.cs.dartmouth.edu/cambridge/haggle, May 2009.

[11] Pierre-Ugo Tournoux, Jeremie Leguay, Farid Benbadis, John Whitbeck, Vania Conan, and Marcelo Dias de Amorim. Density-aware routing in highly dynamic dtns: The rollernet case. *IEEE Trans. Mob. Comput.*, 10(12):1755–1768, 2011.

[12] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[13] Carsten Witt. Greedy local search and vertex cover in sparse random graphs. In *6th Annual Conference on Theory and Applications of Models of Computation*, TAMC '09, pages 410–419, 2009.