Finding remarkably dense sequences of contacts in link streams

Noe Gaumont, Clémence Magnien, Matthieu Latapy

Social Network Analysis and Mining manuscript No.
(will be inserted by the editor)

Finding remarkably dense sequences of contacts in link streams

Noé Gaumont · Clémence Magnien · Matthieu Latapy

Received: date / Accepted: date

Abstract A link stream is a set of quadruplets $(b,\ e,\ u,\ v)$ meaning that a link exists between u and v from time b to time e. Link streams model many real-world situations like contacts between individuals, connections between devices, and others. Much work is currently devoted to the generalization of classical graph and network concepts to link streams. We argue that the density is a valuable notion for understanding and characterizing links streams. We propose a method to capture specific groups of links that are structurally and temporally densely connected and show that they are meaningful for the description of link streams. To find such groups, we use classical graph community detection algorithms, and we assess obtained groups. We apply our method to several real-world contact traces (captured by sensors) and demonstrate the relevance of the obtained structures.

Keywords link stream · temporal network · density · face-to-face interaction · dense subgraphs

1 Introduction

In this paper, we deal with link streams, *i.e.* sequences of quadruplets (b,e,u,v) meaning that a link exists between u and v from time b to time e. Link streams model many real-world situations like contacts between individuals, connections between devices, and others [5,6,16,17,21]. An illustration is given in Figure 1.

This research was supported by a DGA-MRIS scholarship, by a grant from the French program "PIA – Usages, services et contenus innovants" under grant number 018062-44430 and by the CODDDE project ANR-13-CORD-0017-01.

N. Gaumont, C. Magnien and M. Latapy
Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris.
E-mail: noe.gaumont@lip6.fr

The problem of finding dense sub graphs has been extensively studied in the static case. Indeed, detecting cliques and dense groups [2,20] allows finding particularly important sets of nodes in graphs. Some community detection methods [1,10] also use density in order to obtain a high level description of a graph.

The notion of density has been extended to link streams and has been used to study complex networks [13]. It measures to which extent all pairs of nodes are connected all the time. We use this density measure to find relevant groups. This makes it possible to characterize partially the link stream by highlighting the most relevant groups of links.

Such groups should have a high density but this is not sufficient. Above all, they should have a density that is higher than their neighbour groups because, just like in graphs, the value of density in itself is not sufficient to evaluate a group. Therefore, a group is meaningful if it has a higher density than its neighbourhood, both structurally and temporally. For example, a group which has a low density may be considered relevant if the neighbouring link stream is empty. Conversely, a dense group may be considered irrelevant if it is included in a larger dense group. An example of which groups should be captured is given in Figure 1.

In order to find relevant groups, we proceed as follows:

- We build a projection of the link stream into a graph where each link is mapped to a node in the graph;
- We apply a community detection algorithm on the projection and obtain a partition of links in the link stream;
- From the resulting partition, we keep only the relevant groups, *i.e.* the ones which are denser than their neighbourhood and are large enough; we feel that very small groups have limited interest in terms of description of the link stream.

To prove the efficiency of the process, we apply our method on several real world networks and we argue that the groups

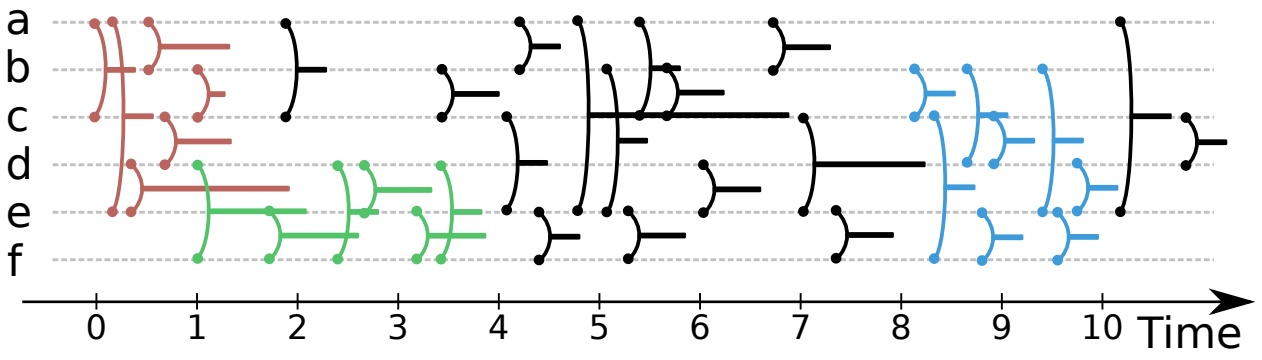


Fig. 1: Example of a link stream. The nodes are on the vertical axis and time is on the horizontal axis. In the example, there is a link between nodes d and f from time 1 to 2. Three dense groups of links are identified by colour (red, green or blue).

found are meaningful and could not have been retrieved by static approaches.

The remainder of this paper is organized as follows. The decomposition and validation method are explained in Section 2. The data sets we use and the associated results are presented in Section 3 and 4. In Section 5, the state-of-the-art in dense groups detection is described and then we conclude and draw some perspectives in Section 6.

2 Method

Our method uncovers meaningful groups of links in a link stream. To this end, we first compute a link partition of the link stream. Then, we propose several criteria to evaluate each group of the partition and retrieve the most meaningful ones.

2.1 Definitions and notations

A link stream is defined as a triplet: $\mathscr{L} = (T, V, E)$, where $T = [\alpha, \omega]$ is a time interval, V a set of nodes and $E \subseteq T \times T \times V \times V$ a set of links. Links of E are quadruplets (b, e, u, v), meaning that the pair (u, v) is continuously linked in the time interval $[b, e] \subseteq [\alpha, \omega]$.

If, for all $(b, e, u, v) \in E$, $u \neq v$ and for all $(b, e, u, v) \in E$ and $(b', e', u, v) \in E$, $[b, e] \cap [b', e'] = \emptyset$, then $\mathscr{L}$ is *simple*. In the following, $\mathscr{L}$ will always be considered to be simple. Also, we do not consider link orientation, *i.e.* (b, e, u, v) and (b, e, v, u) are equivalent.

An extension of the density notion to link streams has been proposed by Viard *et al.* [26], in the case where $b = e$. It measures to which extent all pairs of nodes are connected all the time. We adapt it to the general case of links with durations. The density of a set of nodes, $V' \subseteq V$, in a time interval $[t, t+\delta]$, is expressed as follows:

$$d(V', t, \delta) = \frac{1}{|V'| \cdot (|V'| - 1)} \sum_{(u,v) \in V' \times V', u \neq v} \frac{\tau_{t,\delta}(u, v)}{\delta}, \qquad (1)$$

where $\tau_{t,\delta}(u, v) = \sum_{(b,e,u,v) \in E} |[b, e] \cap [t, t+\delta]|$ is the sum of the durations of links between u and v in interval $[t, t+\delta]$. This includes links which are fully or partly included in the time interval $[t, t+\delta]$. This definition is valid only for simple link streams.

Notice that this is a natural extension of density because if all existing links in $V' \times V'$ last for the whole time interval $[t, t+\delta]$, *i.e.* $\forall u, v \in V'\ \tau_{t,\delta}(u, v) = \delta$, then static and temporal densities are equal.

In a graph, the density is the probability that two randomly chosen nodes are linked together. In a link stream, the density in a given time interval is the probability that two randomly chosen nodes are linked together at a randomly chosen time. Both densities have values in $[0, 1]$. For example in Figure 1, nodes a, b, c, d and e in the time interval $[6, 8]$ have a density of 0.05 because this node set is poorly connected in this time interval. On the other hand, the same node set in the time interval $[0, 1.5]$ has a much higher density because it is much more connected.

As in the case of graphs, density in link stream is very sensitive to the node set size. For example in Figure 1, $d(V \setminus \{a\}, 8, 2)$ is much denser than $d(V, 8, 2)$. Furthermore, the duration δ also has a huge impact on the density. If the duration is expanded and no additional link occurs, then the density decreases quickly, *e.g.* $d(\{e, f\}, 2, 2)$ is denser than $d(\{e, f\}, 2, 3)$ in Figure 1.

This definition holds for a group of nodes in a time interval. For a given group of links $E' \subset E$, let $\alpha(E') = min_{(b,e,u,v) \in E'}(b)$ and $\omega(E') = max_{(b,e,u,v) \in E'}(e)$ denote respectively the beginning and the end of E' and $\delta_{E'} = \omega(E') - \alpha(E')$ denote its duration. The density of E' is then the density of the induced nodes $V_{E'} = \{u, \exists (b, e, u, v) \in E'\}$ in the time interval $[\alpha(E'), \omega(E')]$:

$$d(E') = d(V_{E'}, \alpha(E'), \delta_{E'}). \qquad (2)$$

With this formulation, all links in E' are considered in the computation of the density. However, other links might also contribute to the density. For instance in Figure 1, if

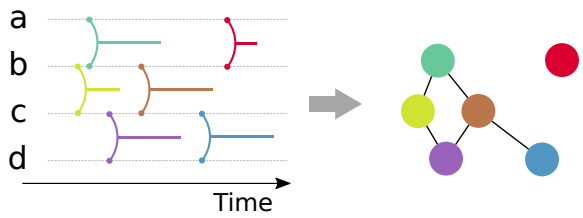


Fig. 2: Transformation of a simple link stream with 4 nodes (a-d) and 6 links into its link graph.

the density of plain black links is computed, all nodes are induced by those links, thus $V' = V$, the beginning equals 2 and the duration equals 9. Therefore in the computation of $d(V,2,9)$, all black links are considered but red and blue links are also taken into account.

2.2 Discovering candidate groups of links

We propose a transformation of a link stream into an unweighted undirected graph that we call the link graph. Each link in the link stream is represented by a node in the link graph. Two different links (b,e,u,v) and (b',e',u',v') are connected in the link graph if they share a node and if their time intervals are overlapping, *i.e.* $\{u,v\} \cap \{u',v'\} \neq \emptyset$ and $[b,e] \cap [b',e'] \neq \emptyset$, see Figure 2. Therefore, a link in the link graph represents both a temporal and structural connection between two links in the link stream. We tested a variation with a weighted link graph where the weight is equal to the duration of the time intersection. However, the results are very similar to the unweighted version and therefore omitted.

Dense groups in the link graph therefore represent groups of closely interconnected links in the link stream, both structurally and temporally. To detect those groups, we apply one of the most used methods to detect communities: the Louvain method [1] [3]. The output of the algorithm is a partition where each community is a candidate relevant group of links. However, there are several reasons why some candidates may not be relevant. First, the candidates uncovered by community detection method usually have heterogeneous sizes and some are very small. As we do not consider very small candidates as relevant, we discard them. Second, as the Louvain algorithm greedily optimizes the modularity in the link graph, the partition in the link stream might also contain candidates which are irrelevant in a link stream context. For example in Figure 1, the links in the time interval $[4,8]$ form a connected component in the link graph and are considered as a community by the Louvain method. However, this candidate should not be considered as relevant because it is less dense than the links in $[0,4]$ or even $[7,11]$. Therefore, we need a method to validate or discard each candidate.

[1] Other community detection methods can be applied.

2.3 Selecting relevant candidates

Density in itself is not sufficient to evaluate a candidate's relevance. Indeed, the optimum value is trivially obtained for a group of two nodes in the time interval when a link between them exists. This is why we consider as relevant the candidates which are denser than their neighbourhoods in the link stream. To define neighbourhoods, we observe that the density, defined by Eq. 1, depends on three aspects: the group of nodes V', the start time t and the duration δ. To take into account these three aspects, we consider groups which differ from the considered candidate in only one of these aspects, which we call neighbour groups. We propose to evaluate the relevance of a candidate by comparing its density to that of the corresponding three kinds of neighbours. The higher its density is compared to the one of its neighbour groups, the better the candidate is.

2.3.1 Neighbourhood definition

For the start time (resp. duration) aspect, we consider all possible values of start time (resp. duration) in a given interval. Each of these values defines neighbour groups with the same set of nodes, duration (resp. start time) as the candidate group and with the considered start time (resp. duration). We then compute the resulting density values of these neighbours by varying the start time (resp. duration) in a continuous fashion. Let $\mathbf{I}$ and $\mathbf{I}'$ be intervals and $L \subseteq E$ be a candidate. The density values of its start time neighbours are thus defined by $d(V_L, y, \delta_L)$ with $y \in \mathbf{I}$. The density values of its duration neighbours are defined by $d(V_L, \alpha(L), z)$ with $z \in \mathbf{I}'$.

The interval considered is $[\alpha, \omega - \delta_L]$ for the start time aspect. For the duration aspect, the interval considered is $[0.8\,\delta_{min}, 1.2\,\delta_{max}]$, where δ_{min} (resp. δ_{max}), is the smallest (resp. largest) candidate duration in the partition. We use this interval for two reasons. First, it contains all reasonable durations when applied to our data sets. Second, we also tested the interval $[1, \omega - \alpha]$ which is much larger. It changes the evaluation quantitatively but not qualitatively. Indeed, groups with a duration of $\omega - \alpha$ have a density close to zero and therefore are always less relevant than candidate groups.

For the nodes aspect, it is impossible to consider all possible node sets as there are too many of them: $2^{|V|}$. Moreover, most of the node sets will mostly be disconnected if the link stream is sparse. Therefore, if V_L contains k nodes, then we consider all the sets of nodes of size k which share $k-1$ nodes with V_L. Therefore, only similar groups are taken into account. This comparison is therefore stricter but also fairer

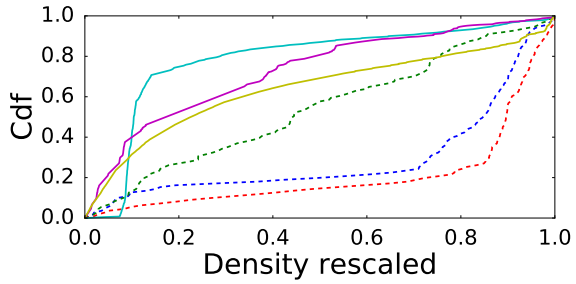


Fig. 3: Cumulative distribution of density rescaled to $[0,1]$ in the duration aspect for several candidates. Full (resp. dashed) lines are candidates from the Rollernet (resp. Socio Pattern) data set.

than the one to all sets of nodes or to randomly chosen set of nodes. A lower number of shared nodes could be used but was not considered because of computational tractability. Indeed, there are already $k(|V|-k)$ set of nodes considered for a single candidate with our current method.

We end up for each aspect with several density values. For the node aspect, we have $k(|V|-k)$ density values. For the start time (resp. duration) aspect, we have a function of the start time (resp. duration) to density.

2.3.2 Score definition

We evaluate a candidate by comparing its density to the density values in each aspect. If a candidate is truly meaningful, then it should have a higher density than most values in each aspect. From experiments, the density values do not seem to follow the same probability density function for all candidates. As an illustration, Figure 3 presents the cumulative distribution of the density for the duration aspect for 3 candidates in the Socio Pattern data set and 3 in the Rollernet data set (these data sets are described in Section 3). If they were following, for instance, a normal law, we could have computed the *z-score* to check if candidates have a density significantly larger than expected. Here the distributions do not seem to follow a common law. Hence, we use percentiles that quantify what fraction of the values are smaller than the considered candidate's density. For the node aspect which is represented by a set of density values S, the score of a given candidate L is:

$$p_{node}(L) = \frac{\sum_{d_i \in S} \mathbf{1}_{d_i < d(L)}}{|S|}\,, \tag{3}$$

where $\mathbf{1}$ is the indicator function. For the start time and duration aspects which are represented by functions, the scores p_t and p_δ of a given candidate L are:

$$p_t(L) = \frac{1}{\omega - \delta_L - \alpha} \int_{\alpha}^{\omega-\delta_L} \mathbf{1}_{d(V_L,z,\delta_L)<d(L)} dz\,, \tag{4}$$

$$p_\delta(L) = \frac{1}{1.2\delta_{max} - 0.8\delta_{min}} \int_{0.8\delta_{min}}^{1.2\delta_{max}} \mathbf{1}_{d(V_L,\alpha(L),z)<d(L)} dz\,. \tag{5}$$

A low score means that the candidate's density is smaller than most densities in the aspect and for this reason the candidate should be discarded.

To sum up, a candidate is evaluated by a triplet consisting of its score for each aspect, and should be discarded if one of them is low, *i.e* lower than a given threshold. The definition of what is a low value is non-trivial and depends on the purpose of the study and the characteristics of the link stream. For these reasons, we set no *a priori* thresholds and instead choose them *a posteriori* based on the observations made on the studied data set, as described in Section 4.

3 Data sets

We apply our method on four data sets. Table 1 lists the number of nodes ($|V|$), the number of links ($|E|$) and the duration ($\omega - \alpha$) of each data set.

Socio Pattern [12][2] contains the temporal network of contacts between students in a high school in Marseilles, France. It gives the contacts between 180 students of 5 classes during 9 days (from a Monday to the Tuesday of the following week) in Nov. 2012. The class of each student is known.

Rollernet [25] was collected during a rollerblade tour in Paris in August 2006. The tour is a weekly event and gathers approximately 2500 participants. Among these, 62 were equipped with wireless sensors recording when they are at a communication distance from one another. The data set therefore contains the proximity links between the persons carrying the sensors. We know the role of each person, *e.g.* staff member at the front or association member.

Reality Mining [8] contains the temporal network of contacts between 94 persons at the MIT Media Laboratory between September 2004 and June 2005. Of these 94 subjects, 68 were colleagues working in the same building on campus (90% graduate students, 10% staff) while the remaining 26 subjects were incoming students at the university's business school. These data set was recorded by Bluetooth on loaned mobile phones.

Baboon [7,24] contains the position of 28 wild olive baboons (*Papio anubis*) at Mpala Research Centre in Kenya between 5 a.m. and 5 p.m. during 2 weeks. These 28 baboons represent around 80% of a troop. Each baboon was fitted with a custom-designed GPS collar that recorded its location every second. We transform this data into a link stream by creating a link between two baboons when the

[2] `http://www.sociopatterns.org`

Data sets	$\|V\|$	$\|E\|$	$\omega - \alpha$
Socio Pattern	180	19774	9 days
Rollernet	62	15803	3 hours
Reality Mining	94	44975	9 month
Baboon	28	95616	14 days

Table 1: number of nodes $|V|$, number of links $|E|$ and duration $\omega - \alpha$ for each data set.

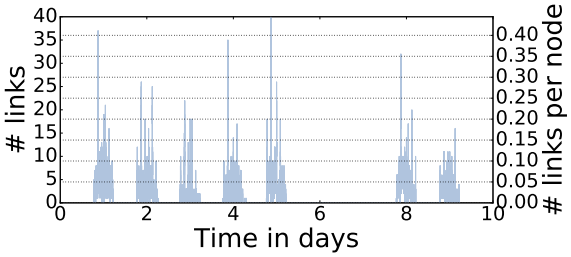


Fig. 4: Number of active links (left axis) and average number of active links per node (right axis) as a function of time on the Socio Pattern data set.

distance separating them is less than 1.5 meter and make each link last at least 10 seconds to smooth potential GPS inaccuracy.

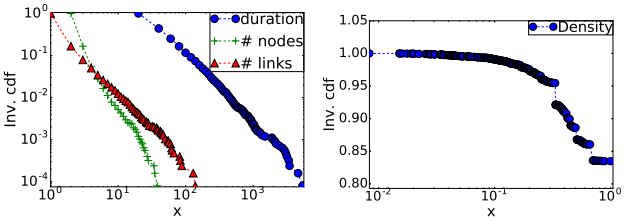
Even if all these data sets are face-to-face interactions networks, they are quite different in their dynamics. For example, the Rollernet data set has roughly the same number of links as Socio Pattern but it lasts only 3 hours compared to 9 days for Socio Pattern. To visualize the data sets' sparse temporal dimension, the number of active links in time are plotted in Figure 4 for the Socio Pattern data set. As there are at most 40 links present at the same time between the 180 students, a graph representing a network at a specific time is mostly empty. The figures for the other data sets are in appendix (Figures 12,17 and 22). They also present a very sparse temporal dimension.

4 Results

For each data set, we apply our method to uncover relevant groups of contacts. As presented in Section 2, the first step is to uncover a partition $\mathscr{P}$ of links. The basic statistics of each found partition such as the number of candidates $|\mathscr{P}|$, the median number of links $\langle|L|\rangle$, and the median node size $\langle|V|\rangle$ are listed in Table 2. The first striking fact in this table is that there are a lot of candidates in the partitions and most of them are very small in terms of number of links and nodes. There are however larger candidates in the Rollernet data set. This difference might be caused by the short link streams duration — 3 hours compared to several days — and the high number of links which makes the link stream denser than other data sets. In order to get a more precise picture of

Data sets	$\|\mathscr{P}\|$	$\langle\|V\|\rangle$	$\langle\|L\|\rangle$
Socio Pattern	12532 (155)	2 (9)	1 (15)
Rollernet	559 (75)	2 (31)	1 (194)
Reality Mining	5737 (474)	2 (12)	1 (36)
Baboon	37671 (1249)	2 (7)	1 (16)

Table 2: Median of some characteristics for each data set partition. The value in parentheses corresponds to the value for candidates with at least 10 links only.



(a) (b)

Fig. 5: Inverse cumulative distributions of the number of links, nodes and duration (a) and density (b) for the candidates found by the Louvain method on the Socio Pattern data set.

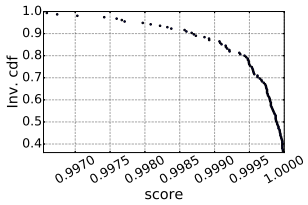
the Socio Pattern data set, the inverse cumulative distributions of the number of nodes, the number of links, durations and density are presented in Figure 5. The distributions are all heterogeneous, except for the density distribution. The steps observed for the density are caused by small candidates; typically a group with 2 links between 3 nodes will have a density of 0.33 and candidates with 1 link [3] will have a density of 1. The figures for the other data sets are in appendix (Figures 13, 18 and 23). For those data sets, we also observe heterogeneous distributions for the number of links, the number of nodes and the duration and steps in the density distribution.

As small candidates have a very limited interest in terms of description of link streams, we begin by discarding groups having less than 10 links. This leaves us with many candidates in all data sets, and we need to distinguish the relevant ones from the others.

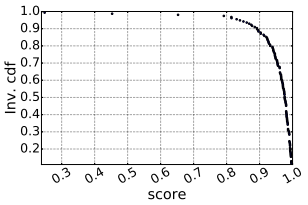
4.1 Groups validation

To separate relevant candidates from others in the partition, we set thresholds on the scores. A candidate is considered as relevant if all its scores are above the thresholds. Notice that decreasing the thresholds will increase the number of candidates kept without removing any candidates previously kept. Therefore, the thresholds can be chosen *a posteriori*. As we do not have any knowledge for the thresholds selection, we

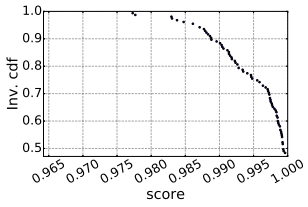
[3] Candidates with one link represent 83% of all candidates.



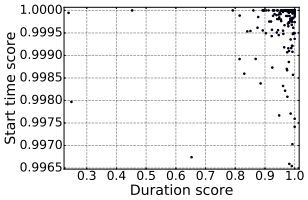
(a) Start time: p_t



(b) Duration: p_δ

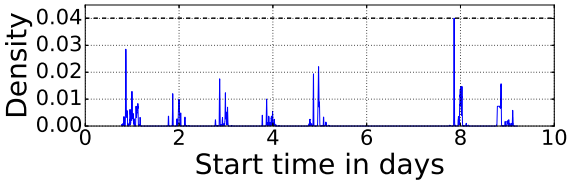


(c) Node set:p_{node}

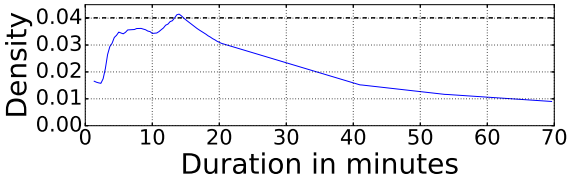


(d) Correlation between start time and duration aspects.

Fig. 6: Inverse cumulative distribution of scores for each aspect for the data set Socio Pattern.



(a)



(b)

Fig. 7: Plain lines: the function $d(V_L, z, \delta_L)$ in (a) and $d(V_L, \alpha(L), z)$ in (b) for a group L in the Socio Pattern data set. Dashed lines: density of L.

study the inverse cumulative distribution of scores for each aspect to fix these thresholds. They are presented for the Socio Pattern data set in Figure 6 (a-c). For this data set, the vast majority of scores are very high, *i.e.* close to 1, regardless of the aspect. We also observe for each aspect a sharp bend in the inverse cumulative distribution of score and use the corresponding value as threshold.

We use for the start time, duration and node aspect the following values: 0.998, 0.85 and 0.98. For other data sets see Figures 15, 20 and 25 in appendix; the scores are similar except for the Rollernet data set which has lower scores, which may be caused by the denser underlying link stream. Finally, Table 3 gives the threshold we used for each aspect and data set.

A candidate is discarded when at least one score is below the corresponding threshold. Therefore, we study if the aspects discard different candidates. The correlation between start time and duration scores is plotted in Figure 6 (d). Both aspects discard different candidates and therefore are not redundant. The same observation holds with the node aspect and for other data sets (see Figures 15 (d), 20 (d) and 25 (d)).

We present a manual analysis for two candidates captured in the Socio Pattern and Rollernet data sets respectively, thus illustrating the notions of scores defined in Section 2 and showing the relevance of the method. We chose these candidates for manual checking because their respective data sets contain information on either the participants or the usual schedule, which helps in evaluating the relevance of groups.

In the Socio Pattern data set, we know the participants' class; we also know when the lectures start in the morning and, when the breaks happen. The group we consider contains 50 links, 17 nodes and lasts for almost 15 minutes and is one of the largest candidates. As this group starts at 7:44 am, this is likely a group of friends gathering before the first lecture of the day which takes place at 8:00 am. Moreover, all of them except one are in the same class, thus making the possibility of a small class gathering even more likely.

This simple qualitative evaluation is a first step but we might have missed a bigger structure. The underlying gathering behind this group of links might have started earlier, lasted longer or even impact other persons. As for all candidates, we have computed its score for each aspect.

Node aspect. For this aspect, the group has a score of 0.98. This means that of all adjacent node sets, the group has almost the highest density and therefore we are sure that this specific node set is indeed important at this period.

Start time aspect. For this aspect, the function start time to density is presented in Figure 7 (a). The circadian cycle is clearly visible along with the week-end. Moreover the group, with a density of 0.04, is denser than all its neighbours for the start time aspect. Therefore the group has a score of 1 which indicates that the start time of the group is relevant.

Duration aspect. For this final aspect, the group has a score of 0.98 which means that the group has also almost the best duration. The function duration to density is presented in Figure 7 (b). The group density is a bit higher if a shorter duration is considered, which is not unexpected because the duration affects the density. Moreover, this is still a very high score because a broad interval of durations is considered in the duration aspect, including durations that do not correspond exactly to any group of links. Indeed, for a given node

set V', duration δ and start time t, some links may be only partially considered, *i.e.* $\exists u,v \in V',\ (b,e,u,v) \in E\ s.t.\ 0 < |[b,e] \cap [t,t+\delta]| \leq |[b,e]|$. In this case, the neighbour groups considered cannot be recovered by a group of links, and their density is not reachable by a group of links. Therefore, a group of links with a perfect score of 1 would be very surprising. Notice that even when increasing the interval for the considered values of duration and exploring $d(V_L,y,z)$ for $y \in [\alpha;\omega-\delta_L]$ and $z \in [0.8g_{min};1.2g_{max}]$, the candidate is still often the best. This emphasizes even more greatly the peculiarity of this candidate.

In the Rollernet data set, we know the role of some participants (staff member, a group of friends, ...). The group we consider contains 38 links, 9 nodes and lasts for almost 5 minutes. The start time is just at the beginning of the roller tour and 8 of the group members are labelled as staff members that should be at the rear of the tour. The last member is labelled as a front member. This group might indicate a quick talk before the beginning of the tour. Again we computed its score for each aspect.

Node aspect. For this aspect, the group has a score of 0.99 which highlights the relevance of the node set at this start time and for this duration.

Start time aspect. For this aspect, the group has a score of 0.85, see Figure 8 (a). As the Rollernet data set is denser than the other data sets, the function start time to density is more stable and other start times achieved higher density. However the group still captures a local maxima of density.

Duration aspect. For this aspect, the group has a score of 0.86 for the duration aspect, see Figure 8 (b).

Even if these scores are not optimal, they are above the thresholds and this is why the group is considered as relevant.

Altogether, there are 136 groups considered as relevant from the 12532 in the Socio Pattern data set. These groups are selected because they have a high score. Most of them have scores that are close but no equal to 1 meaning that some neighbours are still denser. Finally for each aspect, we check how far the group density is from the maximum density. The group density is, for 75% of all groups, on average only 20% smaller than the maximum density in the aspect. Therefore, groups are dense parts of the link stream and have a density close to the maximal density.

Finally, the execution time for each data set is presented in Table 3. The code has been run on a laptop with 8 *Go* of RAM and an *Intel core-i7* processor without any parallelization. The code in $C++$ to compute the scores is available online [4].

[4] `https://bitbucket.org/nGaumont/densityanalysis/`

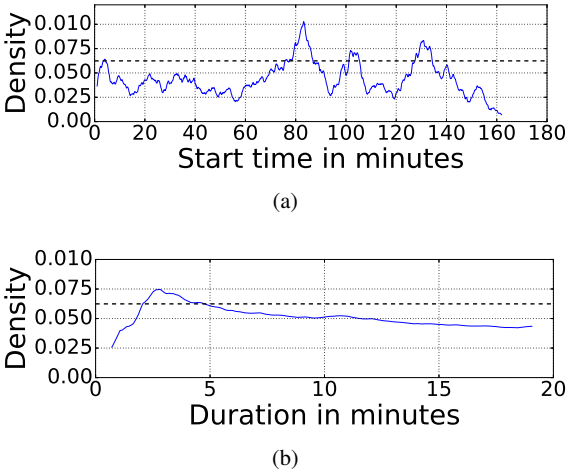


Fig. 8: Plain lines: the function $d(V_L,z,\delta_L)$ in (a) and $d(V_L,\alpha(L),z)$ in (b) for a group L in the Rollernet data set. Dashed lines: density of L.

Data sets	p_{node}	p_t	p_δ	N_c	execution
Socio Pattern	0.98	0.998	0.85	136	2 min
Rollernet	0.9	0.7	0.6	37	4 min
Reality Mining	0.97	0.98	0.8	394	1h
Baboon	0.95	0.99	0.85	1023	52 min

Table 3: Threshold used, number of groups captured and execution time for each data set. p_{node},p_t,p_δ and N_c are respectively the node aspect threshold, the start time aspect threshold, the duration aspect threshold and the number of group captured.

4.2 Group characteristics

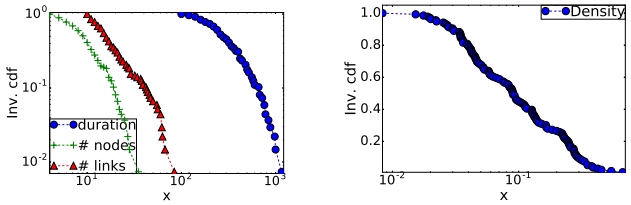
To get a more precise picture for the groups captured, the inverse cumulative distributions of the number of nodes, the number of links, durations and density are presented in Figure 9 for the Socio Pattern data set. The duration, node size and link size distributions are again heterogeneous. The figures for the other data sets are in appendix (Figures 16,21 and 26).

In order to study to which extent the detected groups offer a global description for the data set, we now study how they cover the nodes and the time interval of the link stream.

Concerning node coverage, the distribution of the number of groups per node for the Socio Pattern data set is shown in Figure 10. Notice that all the nodes are in at least one group while some nodes belong to more than 20 groups. With an average of 7.4 groups per node, the uncovered groups form a highly overlapping cover. For other data sets, see Figures 14,19 and 24. For all data sets, we found that the structure is highly overlapping, especially for the baboon data set; this is because a lot of groups are captured for only 28 nodes in the link stream.

	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$
Socio Pattern	0.30 (0.55)	0.30 (0.55)	0.30 (0.55)	0.30 (0.55)	0.29 (0.54)	0.33 (**0.85**)
Rollernet	0.40 (0.52)	0.39 (0.64)	0.41 (0.54)	0.37 (0.54)	0.31 (0.57)	0.26 (0.75)
Reality Mining	0.42 (0.76)	0.42 (0.77)	0.36 (**0.88**)	0.42 (**0.86**)	0.38 (**0.87**)	0.36 (**1**)
Baboons	0.33 (**0.95**)	0.38 (**1**)	0.40 (**0.84**)	0.46 (**0.94**)	0.46 (**0.93**)	0.44 (**0.85**)

Table 4: Median (and maximum) Jaccard index between the groups uncovered and the communities found by bigclam in the aggregated graph where a proportion of λ edges having the smallest weight have been removed. Values which are above 0.8 are in bold.



(a) (b)

Fig. 9: Inverse cumulative distributions of the number of links, nodes and duration in (a) and density in (b) for the candidates captured by our method on the Socio Pattern data set.

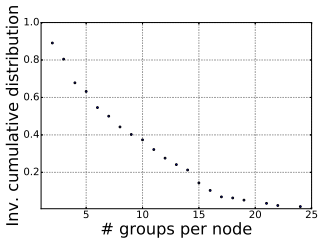


Fig. 10: Inverse cumulative distribution of the number of groups captured per node for the Socio Pattern data set.

One could argue that this structure could also have been obtained by a method computing overlapping communities in a graph. To check this, we construct the aggregated graph $G = (V, E')$ of $\mathscr{L} = (T, V, E)$ such that there is an edge $(u, v) \in E'$ in G if and only if $\exists (b, e, u, v) \in E$. The edge weight is the sum of the durations of each corresponding link in the link stream. We test the bigclam method proposed by Yang *et. al* [29] which captures overlapping groups of nodes in a graph. This method considers only unweighted static graphs. Since edge weights are heterogeneous, we need to find a way to take them into account. As no natural weight separation appears in the link's weight distribution, we use a simple rule such that an edge is present in the unweighted graph if it has a weight strictly greater than at least $\lambda\%$ links, λ being a given threshold. Thus even if the graph is unweighted, we still keep some weight information. We then use the Jaccard index to compare the nodes of a community found by bigclam and the nodes induced by a group uncovered by our method in the original link stream. For each group L captured and its induced nodes V_L, we compute $\mathbb{J}(L) = \max_{C \in \mathscr{C}} \frac{|C \cap V_L|}{|C \cup V_L|}$, the maximum Jaccard index among all communities $C \in \mathscr{C}$ found by bigclam. Table 4 presents the median and maximum Jaccard indexes of $\mathbb{J}(L)$ for different data sets and thresholds λ. According to this table, uncovered groups and communities found by bigclam are different (the median Jaccard index is low). Some groups are nevertheless present in both structures as reflected by a high maximum Jaccard index. This happens in the Rollernet and Baboons data sets for only very few groups and specific link removal thresholds. These high Jaccard indexes might by correlated to the high number of uncovered groups for these data sets because more uncovered groups induce a higher probability to find a matching community.

Overall, our method has highlighted groups of nodes which would not have been uncovered by this static method.

For the temporal aspect, we also look for the time overlap of groups. We compute the time proportion where no captured group is present. For all data sets except the Rollernet data set, there is no group present between 82% and 95% of the time. For Rollernet, this proportion drops to 15%. This difference is likely caused by the very short duration of the link stream. Indeed, the Socio Pattern, Reality Mining and Baboons data sets span nights, during which activity is lower.

While most of the time there is no active group in the link stream, at some times several groups are active at once. This result is very different from the structural aspect. Indeed, the groups form a cover of nodes but they are covering only a small fraction of the time. In the Socio Pattern data set, the groups are typically before the first lecture, during the breaks and lunch, and after the last lecture. Therefore, our method is able to recover important time intervals in link streams, which further show its relevance. Notice however that isolating important time intervals could be done in a simpler way, by observing the activity. To illustrate this, the number of active groups over time is presented in Figure 11 along with the number of links for the first day of the Socio Pattern data set.

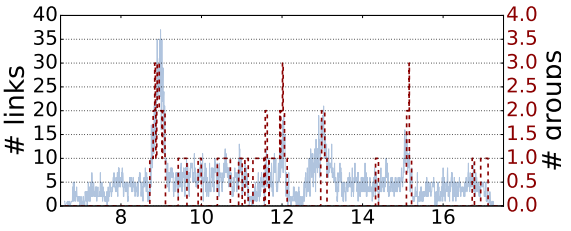


Fig. 11: Number of active links (plain line) and number of active groups (red dashed line) as a function of time for the first day of the Socio Pattern data set.

5 Related Work

In recent years, significant efforts have been devoted to finding subgraphs in a temporal context. Subgraphs have mainly been studied in two ways: either as communities in graph snapshots or as dense sub-parts in evolving networks. These approaches are related but not equivalent. The former finds a whole partition of a network while the latter captures a single group. As they optimize different metrics and respect different constraints, it is difficult to compare the two approaches. See for instance these surveys and the references therein [11, 15, 28] that describe community detection methods in graph snapshots. The results of these methods are partitions of nodes for each snapshot. In most of them, a group of nodes in a specific snapshot is tracked in the next snapshot to follow its evolution. These methods suffer from instability between two snapshots and from the construction of the snapshot in which some information is lost. For example, a group of nodes in a snapshot could be mistakenly captured because the aggregation method makes it seem like a relevant group. Symmetrically, a relevant group in the link stream could be ignored if it is split between two snapshots.

Several models keeping all the temporal information have been proposed and used in different contexts to gain meaningful insights. For example, the methods in [18, 23] aim at detecting communities in diachronic networks, *i.e.* links extremities are associated with possibly distinct timestamps. A diachronic network is modelled as a graph, $G = (V, E)$, defined upon a set of time-labelled nodes, $V = \{(u_i, t_i)\}$, and a set of links, $E = \{((u_i, t_i), (u_j, t_j))\}$. This accurately models citation networks where an author A publishing a paper at time t_1 is represented by the node (A, t_1). The link $((A, t_1), (B, t_2))$ exists if B has published a paper at time t_2 in which the paper from A published at time t_1 is cited. Thus, the link $((A, t_1), (B, t_2))$ in the diachronic network has a different meaning than the link (t_1, t_2, A, B) in a link stream and both models represent intrinsically different objects.

The problem of finding the densest sub-part in evolving networks and temporal networks has also been studied. The method in [4] considers graphs where the weights on the edges vary between -1 and 1, which is rather different from link streams. Also, it recovers a subset of nodes on a time interval such that it maximizes the sum of the weight. Thus, some temporal information is lost because of the time aggregation.

Epasto *et al.* [9] recover the densest sub-graph in evolving networks. An evolving network is a network that changes according to a sequence of updates, namely edge additions or removals. It maintains at each time a single group of nodes which maximizes the average degree at this specific time. Therefore, the method is relevant only when a relevant network structure exists after each update, which is not necessarily the case for link streams, as we have seen.

Rozenshtein *et al.* [19] designed a method to find the densest sub-graph in temporal networks. The method captures a group of nodes and potentially several time intervals. The maximum number of time intervals and the maximum sum of the duration of time intervals are parameters of the method, thus the results are very dependent on the parameter choice. Also, the notion of density used is the average degree in the graph aggregated over the chosen time intervals and therefore some temporal information is lost.

Viard *et al.* [27] consider link streams where links do not have durations and use a similar notion of density relying on a parameter, Δ, emulating link duration. With this notion of density, they define maximal Δ-cliques (in terms of nodes or time interval) as sets of nodes having a Δ-density of one in a given time interval. Δ-cliques are a great way to decompose a link stream. However, the Δ-cliques are typically very small in size and duration. On the the Socio Pattern data, the interactions are captured with a time precision of 20 seconds therefore, when Δ equals to 20 seconds, Δ-density and density are equal. With $\Delta = 20s$, the biggest Δ-cliques contains only 5 nodes and 10 links. As the relevant groups can have any density (see Figure 9(b)), we are able to detect bigger groups and build more coarse description of link streams.

Sekara *et al.* [22] study a rich data set of approximately 1000 students during 36 months at a large European university. They use the temporal network of face-to-face interactions measured via Bluetooth. They study gatherings which are represented by sets of nodes and time intervals. Gatherings are captured by first transforming the temporal network into snapshots of small duration and computing the connected components in each snapshot. A gathering is then a matching of connected components across snapshots using a hierarchical clustering based on the Jaccard index [14]. In the method, a gathering is considered only if it lasts at least 20 minutes. The definition of gathering relies on three intricate parameters which are set *a priori*: the snapshots duration, a threshold for connected component matching and the minimum duration of each gathering. Moreover, the notion of gathering is defined specifically to detect meeting of students, which explains the high minimum duration. Our

method is able to detect groups of diverse durations (see Figure 9(a)).

To sum up, some methods [4,9,19] seek the densest node set in evolving networks. However, the criteria they use do not take time into account as they rely on classical graph metrics which might not be relevant for temporally sparse data sets. Also, they recover only the densest node set while we uncover several groups of links.

Other methods [22,27] detect node sets during time intervals in link stream. However, the groups uncovered with these methods have strict characteristics tailored to the object of study: cliques or long duration. Our method is able to detect more diverse groups. Finally, our method can be tuned by parameters set *a posteriori*.

6 Conclusion

In this paper, we consider the problem of uncovering dense groups in link streams where each link has a duration. As opposed to some existing methods, we take full advantage of the temporal information thanks to the link stream formalism which does not use any aggregation. This shift is important as it eases the definition of algorithms and metrics that mix structural and temporal information. The extension of density for link streams that we propose is such an example. Moreover, we do not consider a group as a combination of a node set and a time interval but simply as a group of links existing for a given time interval.

To uncover dense groups of links, we propose a method that first transforms a link stream into a static, unweighted and undirected graph where links of the link stream become nodes in the graph. Two nodes in the graph are connected if there is a temporal and a structural connection between the corresponding links in the link stream. Thus looking for communities in the graph is a way to find a set of candidate groups. To do this, we use an existing community detection method, namely the Louvain method.

To assess the relevance of the candidates in the found partition of links, we propose to study several aspects. Each aspect defines and builds neighbour groups from a given candidate in the time or topological dimension. We use each aspect to compute a score that compares the candidate density to neighbourhood densities. Once the scores of each candidate are computed, the relevant ones are the one having scores above thresholds that are set *a posteriori* using the scores distributions. Changing the thresholds does not interfere in the set of candidate but on the candidates selection.

We apply our detection method to 4 real world data sets. All of them are face-to-face interaction networks. However, they come from drastically different contexts. Two of them are networks of students, one is a network of roller blade tour participants in Paris and the last one is a network of baboons in wild life. Also, the duration of each data set is drastically different: from a few hours to several months. In all data sets, we find groups having a very high density compared to their neighbour groups.

There is no known ground truth in the used data sets, however some metadata are known for two of them. In this way we are able to analyze manually some of the uncovered groups. We find that the groups correspond to relevant events in the data and they have empirical justification such as students gathering before the first class or for lunch. More generally, we also studied the distribution of groups in time and over the nodes. We find out that the groups are highly overlapping over the nodes but concern only a small fraction of time.

To test the novelty of our results, we apply a method that detect overlapping communities on static graphs: the bigclam method. We find that just very few groups were recovered by the bigclam method. Therefore by using time information, we are able to uncover new structures that might hardly be detected by a static method.

6.1 Perspectives

Our method relies on a static graph, a classical community detection algorithm, a minimum group size, and some thresholds set *a posteriori*. An automatic selection of the thresholds used is a promising direction, for example by detecting inflection points in the scores distributions. Classification methods could also be used on the groups based on their three scores. However, we need to apply this method in more diverse contexts to understand how reliable an automatic method could be.

It would also be interesting to detect relevant groups without relying on the transformation into a graph for two main reasons. First, the density in the link stream cannot be deduced from the static graph we created. Second, we might have missed groups by using the Louvain algorithm on the link graph.

Our method detects groups of links in link streams while the literature is mainly focused on community structure in graph snapshots. Communities in graph snapshots are related to dense groups, yet they are different. It would be interesting to study the interplay between the two objects. One way to do this could be to find appropriate time windows and build graph snapshots from link streams and then to use existing methods on the snapshots.

Finally, another promising direction is to build generative models of link streams with some constraints on density. This would be beneficial to test the accuracy of our method but also to design new algorithms applied to community detection in link streams.

References

1. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, aug 2010.
2. O. D. Balalau, F. Bonchi, T.-H. H. Chan, F. Gullo, and M. Sozio. Finding Subgraphs with Maximum Total Density and Limited Overlap. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pages 379–388, New York, New York, USA, feb 2015. ACM Press.
3. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.
4. P. Bogdanov, M. Mongiovì, and A. K. Singh. Mining Heavy Subgraphs in Time-Evolving Networks. In *2011 IEEE 11th International Conference on Data Mining*, pages 81–90. IEEE, dec 2011.
5. A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro. Time-varying graphs and dynamic networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6811 LNCS, pages 346–359, 2011.
6. R. Cazabet, F. Amblard, and C. Hanachi. Detection of Overlapping Communities in Dynamical Social Networks. *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, 2010.
7. M. C. Crofoot, R. W. Kays, and M. Wikelski. Data from: Shared decision-making drives collective movement in wild baboons, 2015.
8. N. Eagle, A. S. Pentland, and D. Lazer. Inferring Social Network Structure using Mobile Phone Data. *Pnas*, 106(usually 1):15274–15278, 2009.
9. A. Epasto, S. Lattanzi, and M. Sozio. Efficient Densest Subgraph Computation in Evolving Graphs. In *Proceedings of the 24th International Conference on World Wide Web*, pages 300–310. International World Wide Web Conferences Steering Committee, may 2015.
10. T. Falkowski, A. Barth, and M. Spiliopoulou. DENGRAPH: A density-based community detection algorithm. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007*, pages 112–115. IEEE, nov 2007.
11. T. Falkowski, M. Spiliopoulou, and J. Bartelheimer. Community dynamics mining. In *Proceedings of 14th European Conference on Information Systems (ECIS 2006)*. Citeseer, 2006.
12. J. Fournet and A. Barrat. Contact patterns among high school students. *PloS one*, 9(9):e107878, jan 2014.
13. N. Gaumont, T. Viard, R. Fournier-S'niehotta, Q. Wang, and M. Latapy. Analysis of the temporal and structural features of threads in a mailing-list. In *Complex Networks VII*, Studies in Computational Intelligence. Springer International Publishing, 2016.
14. D. Greene, D. Doyle, and P. Cunningham. Tracking the Evolution of Communities in Dynamic Social Networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, ASONAM '10, pages 176–183, Washington, DC, USA, 2010. IEEE Computer Society.
15. T. Hartmann, A. Kappes, and D. Wagner. Clustering evolving networks. *arXiv preprint arXiv:1401.3516*, 2014.
16. P. Holme. Modern temporal network theory: a colloquium. *Eur. Phys. J. B*, 88(9):234, 2015.
17. P. Holme and J. Saramäki, editors. *Temporal Networks*. Understanding Complex Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
18. B. Mitra, L. Tabourier, and C. Roth. Intrinsically dynamic network communities. *Computer Networks*, 56(3):1041–1053, feb 2012.
19. P. Rozenshtein, N. Tatti, and A. Gionis. Discovering Dynamic Communities in Interaction Networks. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 678–693. Springer Berlin Heidelberg, 2014.
20. R. Samudrala and J. Moult. A graph-theoretic algorithm for comparative modeling of protein structure. *Journal of Molecular Biology*, 279(1):287–302, 1998.
21. J. Saramäki and E. Moro. From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B*, 88(6):164, jun 2015.
22. V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences*, page 201602803, aug 2016.
23. L. Speidel, T. Takaguchi, and N. Masuda. Community detection in directed acyclic graphs. *The European Physical Journal B*, 88(8):203, aug 2015.
24. A. Strandburg-Peshkin, D. R. Farine, I. D. Couzin, and M. C. Crofoot. Shared decision-making drives collective movement in wild baboons. *Science*, 348(6241):1358–1361, jun 2015.
25. P.-U. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. Dias de Amorim, and J. Whitbeck. The Accordion Phenomenon: Analysis, Characterization, and Impact on DTN Routing. In *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*, pages 1116–1124. IEEE, apr 2009.
26. T. Viard and M. Latapy. Identifying roles in an IP network with temporal and structural density. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 801–806. IEEE, apr 2014.
27. T. Viard, M. Latapy, and C. Magnien. Computing maximal cliques in link streams. *Theoretical Computer Science*, 609:245–252, 2016.
28. Q. Wang, E. Fleury, T. Aynaud, and J.-L. Guillaume. Communities in evolving networks: definitions, detection and analysis techniques. *Dynamics of Time Varying Networks*, 2012.
29. J. Yang and J. Leskovec. Overlapping community detection at scale. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, page 587, New York, New York, USA, feb 2013. ACM Press.

A Rollernet data set

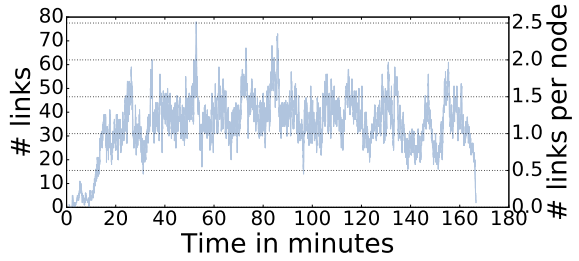
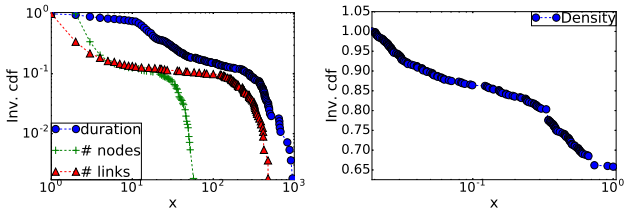


Fig. 12: Number of active links (left axis) and average number of active links per node (right axis) as a function of time on the Rollernet data set.



(a) (b)

Fig. 13: Inverse cumulative distributions of the number of links, nodes and duration (a) and density (b) for the candidates found by the Louvain method on the Rollernet data set.

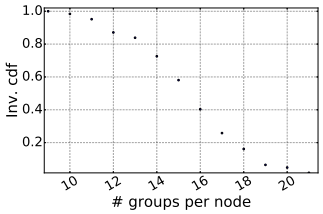
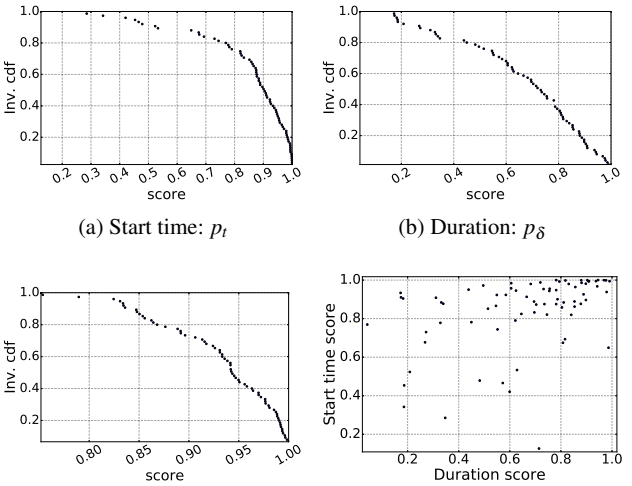


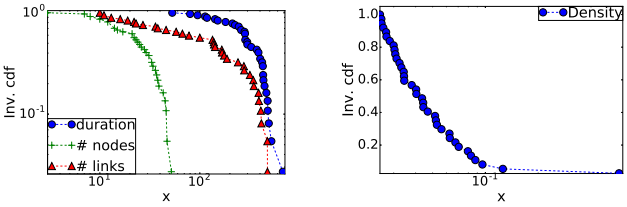
Fig. 14: Inverse cumulative distribution of the number of groups captured per node for the Rollernet data set.



(a) Start time: p_t (b) Duration: p_δ

(c) Node set: p_{node} (d) Correlation between start time and duration aspects.

Fig. 15: Inverse cumulative distribution of scores for each aspect for the data set Rollernet.



(a) (b)

Fig. 16: Inverse cumulative distributions of the number of links, nodes and duration in (a) and density in (b) for the candidates captured by our method on the Rollernet data set.

B Baboon data set

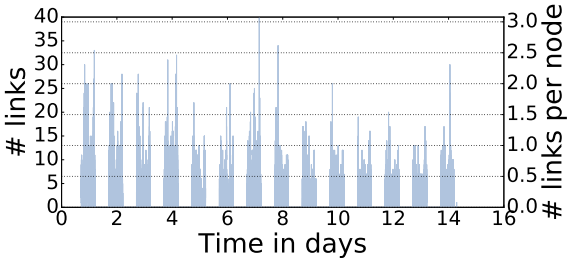
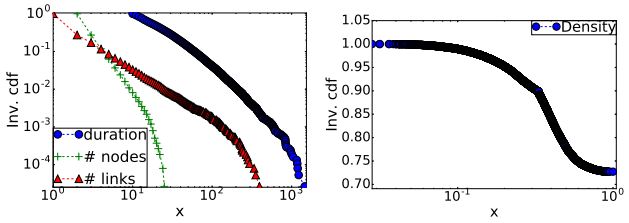


Fig. 17: Number of active links (left axis) and average number of active links per node (right axis) as a function of time on the Baboon data set.



(a) (b)

Fig. 18: Inverse cumulative distributions of the number of links, nodes and duration (a) and density (b) for the candidates found by the Louvain method on the Baboon data set.

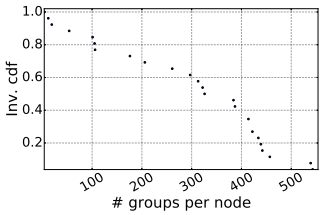
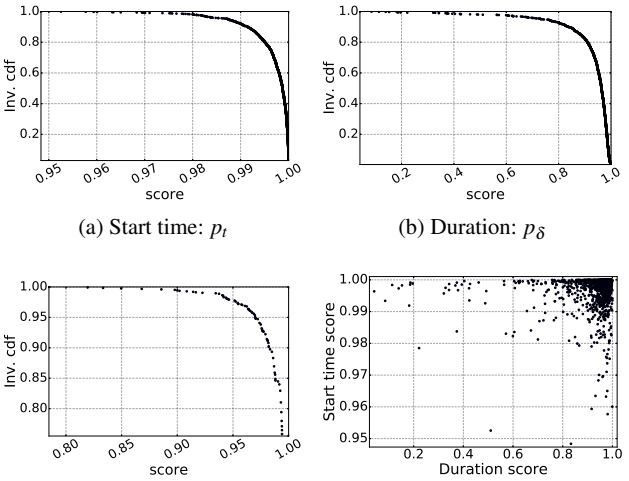


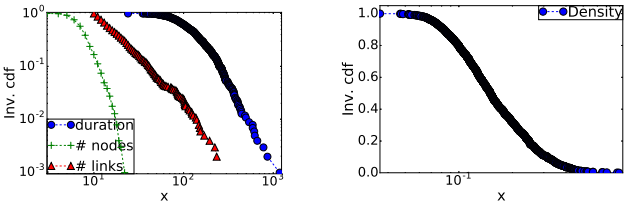
Fig. 19: Inverse cumulative distribution of the number of groups captured per node for the Baboon data set.



(a) Start time: p_t (b) Duration: p_δ

(c) Node set:p_{node} (d) Correlation between start time and duration aspects.

Fig. 20: Inverse cumulative distribution of scores for each aspect for the data set baboon.



(a) (b)

Fig. 21: Inverse cumulative distributions of the number of links, nodes and duration in (a) and density in (b) for the candidates captured by our method on the Baboon data set.

C Reality mining data set

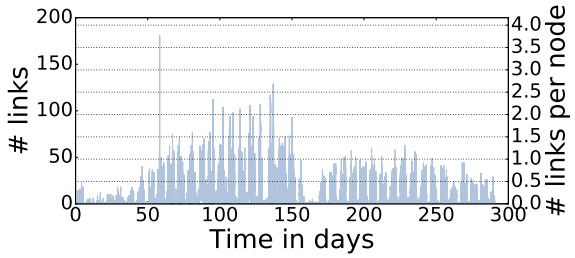
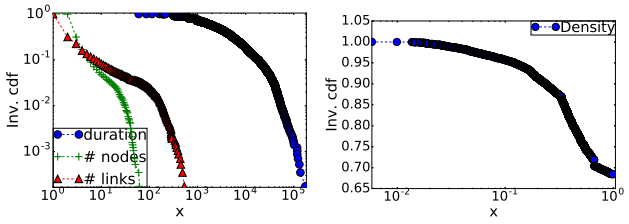


Fig. 22: Number of active links (left axis) and average number of active links per node (right axis) as a function of time on the Reality Mining data set.



(a) (b)

Fig. 23: Inverse cumulative distributions of the number of links, nodes and duration (a) and density (b) for the candidates found by the Louvain method on the Reality Mining data set.

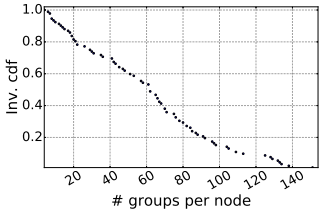
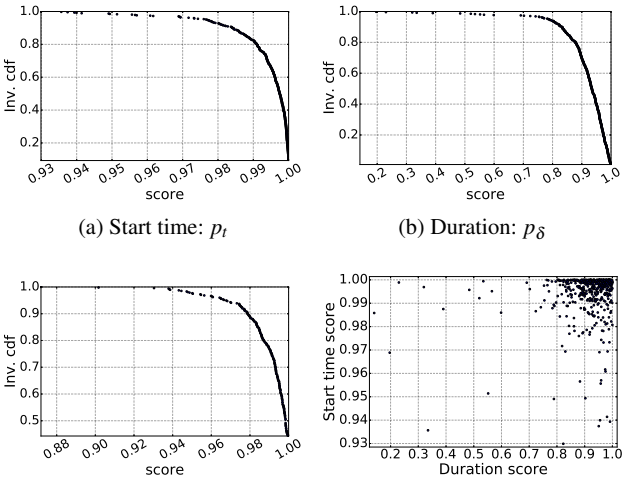


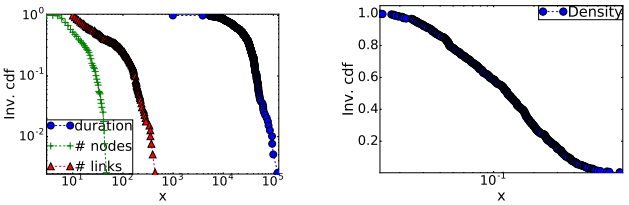
Fig. 24: Inverse cumulative distribution of the number of groups captured per node for the Reality Mining data set.



(a) Start time: p_t (b) Duration: p_δ

(c) Node set:p_{node} (d) Correlation between start time and duration aspects.

Fig. 25: Inverse cumulative distribution of scores for each aspect for the data set Reality Mining.



(a) (b)

Fig. 26: Inverse cumulative distributions of the number of links, nodes and duration in (a) and density in (b) for the candidates captured by our method on the Reality Mining data set.