Characterising inter and intra-community interactions in link streams using temporal motifs

Jean Creusefond[1] and Remy Cazabet[2]

[1] GREYC, Normandie Université,
[2] Sorbonne Universites, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, France

Abstract. The analysis of dynamic networks has received a lot of attention in recent years, thanks to the greater availability of suitable datasets. One way to analyse such dataset is to study temporal motifs in link streams , *i.e.* sequences of links for which we can assume causality. In this article, we study the relationship between temporal motifs and communities, another important topic of complex networks. Through experiments on several real-world networks, with synthetic and ground truth community partitions, we identify motifs that are overrepresented at the frontier –or inside of– communities.

1 Introduction

Communication networks represent human interactions that happen at certain times. The properties of these networks are often studied in order to have a better understanding of human dynamics [2,8].

The basic building blocks of networks are called motifs, small structures that appear multiple times in the network. This concept was originally formulated for static networks [12] and has been extended for temporal networks [19]. In the case of communication networks, these motifs are an indication of the nature of the communication [18]. For instance, a set of messages in a back-and-forth pattern between two individuals is probably a conversation.

It is a common assumption that the nature of the relationship of two individuals define the nature of the communities that they share [1]. If the motifs characterise the relationships between individuals, they may be related to the community structure.

The existing definitions of motifs describe messages that are received and sent in a short time-frame. Such motifs do not include causally-linked interactions that happen outside of the time-frame. These interactions could be due to an individual that is not active on the network at that time, and therefore unaware of the messages received.

In this paper, we first propose an adaptation of the definition of a motif that takes into account users' activity periods. We then study experimentally the frequency of motifs inside and outside communities in order to test the hypothesis that temporal motifs are linked to the community structure.

2 Related work

Zhao *et al.* [19] defined temporal motifs. They measured the frequency of the different motifs and characterised them by their shape (ping pong, star, chain). Kovanen *et al.* [10] extended the definition of motifs in order to take into account the order of communications. For instance, their definition differentiates a "AB-BA-AB" motif from a "AB-AB-BA" motif, which the previous definition does not.

Zhang *et al.* [18] considered the relative frequency of some 3-events motifs when increasing the time window. They observed that the dominant 3-event motifs were related to the dominant 4-event motifs in the 6 datasets that were used.

In order to decide of their significance, the frequency of the motifs in the dataset is often compared with null models [19,16]. These models describe a network that is identical to the data, except for one feature that is randomised. This methodology is used to evaluate the influence of the randomised feature on the measurements.

Zhao *et al.* [19] compared their results to the time-mixing model, a null model where all timestamps of the dataset are randomised. They observed that the time-mixing model created mainly isolated entries, which is an important difference with empirical observations. However, the time-mixing model deletes the phenomenon of burst in the activity of individuals, on top of deleting causality effects.

Tabourier *et al.* [16] presented a null model that conserves this feature, the correlation-mixing model. As for the time-mixing model, all source and destinations are kept and timestamps are randomised. However, this randomisation is carried out over the messages that were emitted by the same individual, and not over all messages. It implies that temporal features such as the burstiness of communications is conserved, but not the causal link between messages.

Several works have been done on the question of detecting communities on dynamic networks [4]. However, these approaches focus on slowly evolving networks, in which edges are persistent along time (relations, for instance friendship or colleague relation). On the contrary, this work focuses on networks which have a much faster temporality than communities, i.e interactions are short-lived (for instance messages, calls between friends or colleagues). We therefore assume a fixed community structure, and observe interactions over this structure.

3 Adapting motifs for communication networks

In this section, we will introduce a variation on motifs that take into account the activity periods of individuals. We call this variation an a-motif.

We model the communication network as links streams $G = (V, E)$. A link stream is composed of a set V of nodes and a set $E \subset V \times V \times \mathbb{R}^+$ of timestamped links between nodes. We note that multiple links may exist between the same pair of nodes.

A temporal motif describes the structure of a sequence of communications. Formally, a temporal motif is an equivalence class of a communication graph [19], that is defined as follows on link streams :

Definition 1 (communication graph). *A communication graph on a window of size $W \in \mathbb{R}^+$ is a link stream $G = (V, E)$ such that $\forall (u_i, v_i, t_i) \in E$, $\exists (u_j, v_j, t_j) \in E$ that respects $(u_i, v_i, t_i) \neq (u_j, v_j, t_j)$, $\{u_i, v_i\} \cap \{u_j, v_j\} \neq \emptyset$ and $0 < |t_i - t_j| < W$.*

Two communication graphs belong to the same equivalence class (*i.e.* motif) if the corresponding weighted graphs (a link is weighted by the number of communications) are isomorphic. Kovanen *et al.* [10] extend this equivalence relationship by taking into account the order of the links in the communication graphs. We call a communication graph that belong to such an equivalence class an *instance of a motif.*

This paper focuses on communication networks such as e-mails or answers in an online forum. In such networks, the individual receiving a message is not always aware of the message at the time of reception. Typically, receiving an e-mail does not mean that it is acknowledged. In that case, the causal link between two communications may not be directly related to the reaction time. We define the a-motif (for *activity motifs*) in order to take that phenomenon into account.

We first split the messages emitted by the individuals into activity periods. These periods are time intervals when an individual emits messages in a short burst.

Definition 2 (μ-activity period). *For each node $v \in V$ in a link stream $G = (V, E)$, we note E_v the set of messages emitted by v and $med(v \in V)$ the median of the time elapsed between two consecutive messages emitted by v. We also note $t((u, v, x) \in E) = x$ the date of an edge. A* ***μ-activity period*** *of an individual $v \in V$ is a time interval $[a; b]$ during which v emitted a set of messages $M(a, b) = \{e \in E_v \mid a \leq t(e) \leq b\}$, that respects the following properties :*

- *$\exists e_1 \in M(a, b)$, $t(e_1) = a$ and $\exists e_2 \in M(a, b)$, $t(e_2) = b$ and*
- *$\forall e_1 \in M(a, b)$, $t(e_1) \neq b \Rightarrow \exists e_2 \in M(a, b)$, $0 < t(e_2) - t(e_1) \leq \mu \cdot med(v)$ and*
- *$\forall e \in E_v$, $t(e) < a \Rightarrow t(e) < a - \mu \cdot med(v)$ and $t(e) > b \Rightarrow t(e) > b + \mu \cdot med(v)$.*

We then define the a-motifs as equivalence classes of *activity graphs*, formed as follows. If an edge (u_1, v_1, t_1) belongs to an activity graph, the edge (u_2, v_2, t_2) may also belong in that graph if $t_1 < t_2$ and :

- $u_1 = u_2$ and t_1 and t_2 belong to the same activity period of u_1. There might be a causal link between two messages emitted by an individual in the same activity period.
- $v_1 = u_2$ and t_2 belong in the next activity period of u_2 that happens after t_1. If t_1 is inside an activity period of u_2, then t_2 must belong to the same activity period. There might be a causal link between a message received and the next messages sent by the recipient during his/her next activity period.

We use the equivalence function introduced in Kovanen *et al.* [10] to define the a-motifs as equivalence classes of activity graphs. The detection of a-motifs instances is illustrated Fig. 1.

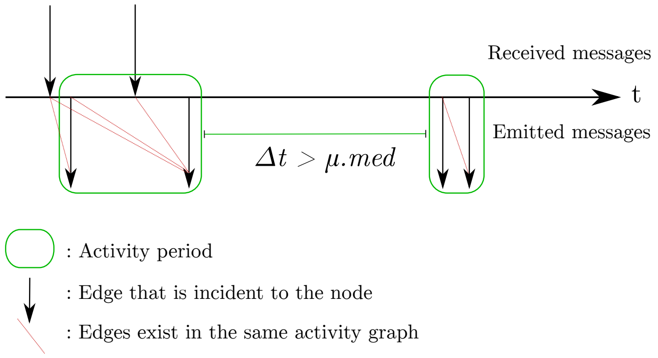
Received messages
t
Emitted messages
$\Delta t > \mu.med$
: Activity period
: Edge that is incident to the node
: Edges exist in the same activity graph

Fig. 1. For a node, the messages that are emitted are grouped into activity periods. The set of incident edges forms activity graphs.

For complexity reasons, we restrict our study to size 3 a-motifs, *i.e.* those that are made of three edges. This size is chosen as a compromise between the computation time needed for the detection of instances and the complexity of the structures that are observed.

We identify the motifs by letters that correspond to the nodes that are involved in the motif. For instance, the motifs illustrated Fig. **??** are, from left to right, "AB-BA", "AB-AB", "AB-AC" and "AB-BC".

Some size 3 a-motifs are geometrically similar, such as "AB-AC-BC" and "AB-AC-CB", or "AB-BC-CB" and "AB-BA-BC". In order to reduce the number of observations, we focus on four motifs that have been identified as important in the associated literature [19,16,18] and a fifth that we identified as interesting. Those are : the star "AB-AC-AD", the ping-pong "AB-BA-AB", the triangle "AB-BC-CA" and the chain "AB-BC-CD". We add the spam "AB-AB-AB" to that list because of its direct possible interpretation. Those motifs are illustrated Fig. 2.

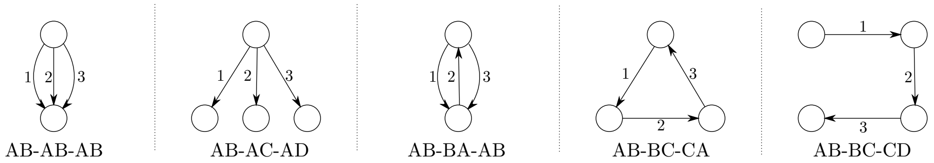

Fig. 2. The five studied motifs. Numbers indicate the order of the edges.

There may be activity periods including dozens of messages while others include only a few. If an activity period of a node v is made of k edges and if v received l messages before that period, then $k \cdot l$ instances of size 2 a-motifs are created. The impact of a message on a-motifs frequency is therefore dependent of the size of the activity periods of the receiver.

In this work, we consider that a message should not have more impact on the results than another because of the size of activity periods. To that purpose, we weight instances of a-motifs such that the weights of the set of instances that has the original edge sum to one. That weight is computed in the following manner: from an instance that has a weight w, if that instance is extended to generate k instances of bigger size, each of these instances has a w/k weight.

For instance, if an edge creates k_1 instances of size two, each of them has weight $1/k_1$. If the first of these instances generates k_2 instances of size three, each of them has weight $1/(k_1 \cdot k_2)$. If the second of these instances generates k'_2 instances of size three, each of them has a $1/(k_1 \cdot k'_2)$ weight, and so on. Each measure that is presented in following experiments is weighted accordingly.

4 Experiments

In this section, we present our study of the properties of a-motifs.

These experiments were implemented in Python. They were run in parallel on 40 AMD Opteron CPUs (2.6 GHz). Due to the size of the dataset and the number of null-model instances, the full run takes about a day.

4.1 Datasets

In order to carry out our experiments, we collected a dataset that includes messages between individuals and three ground-truth community partitions. This dataset is original since, to the best of our knowledge, no openly available dataset features both types of data.

Caen University dataset We obtained metadata for all emails transferring through servers of Caen University, France, for a period of three months. Available information include source, destination and timestamp. Individuals in this network are students and employees of the university.

Three kinds of partitions can be extracted from available data:

- For researchers, we know the **research laboratory** they belong to.
- For students and researchers, we also know their **CNU section** (CNU stands for Universities National Council), which indicates to which scientific field they belong to.
- For all users, we know to which **administrative entity** they belong to, typically their school.

This dataset includes 45 **research laboratories**, 146 **CNU sections** and 57 **administrative entities**.

The network has the following properties:

- It contains 7 688 665 messages sent between 210 085 addresses.
- 168 507 messages sent between 918 addresses with a **research laboratory**.
- 378 721 messages sent between 17 275 addresses with a **CNU section**.
- 1 275 662 messages sent between 26 177 addresses with a **administrative entity**.

We created three link streams, one for each partition, that includes only nodes corresponding to individuals present in the corresponding partition, and that includes communication between these nodes.

Other datasets

Name	n	m	nodes	edges
Enron [9]	86978	1134990	employees	e-mails
Facebook [17]	45813	855542	users	wall posts
UC Irvine [13]	1899	59835	students	messages
Radoslaw [11]	167	82876	employees	e-mails
Debian [6]	34648	316569	users	answers
Digg [5]	30360	86203	users	answers
Linux Kernel Mailing List (LKML)[3]	26885	1028233	users	answers
Slashdot [7]	51083	139789	users	answers

Table 1. Konect's networks

Besides the Caen university dataset, we analysed a set of communication networks available on the Konect[4] website (see Table 1). After filtering out self loops and nodes with no links, we considered them as link streams.

Because these datasets do not have a known ground truth partition, we used Louvain [3] and Infomap [15] community detection algorithms on the aggregated network to generate two reference partitions. The aggregated network contains an edge between a pair of nodes if there is at least one interaction at any point in time between these two nodes in the link stream. Since the results on the partitions of both algorithms are similar, we will only present the results on the partitions obtained with the Louvain algorithm.

4.2 Comparing with the correlation-mixing model

For each measure on the motifs, we compare the value on the original graph and the same value on graphs generated by the correlation-mixing model. We

[3] http://konect.uni-koblenz.de/networks/lkml-reply

[4] http://konect.uni-koblenz.de

consider statistically significant differences to be a consequence of causality, as described by Tabourier *et al.* [16].

In practice, we observe that these measures are normally distributed. In such a case, we can use the "66-95-99.7 rule" [14], that states that about 66% of normally distributed values are within one standard deviation of the mean, about 95% of them are within two standard deviations and about 99.7% of them are within three standard deviations. Therefore, a value that is further from the mean than three times the standard deviation would have less than 0.3% chance to be generated by the normal distribution. For each measure s on the data, we obtain the average μ_s and the standard deviation σ_s of s on the graphs generated by the null model. We then evaluate the difference between the data and the null model using the z-score:

$$z\text{-}score(s) = \frac{s - \mu_s}{\sigma_s} \tag{1}$$

If the z-score is more than three in absolute value, we conclude that the null model does not explain the value of the measure in the data. Since we use the correlation-mixing model, a significant difference would be caused by the removal of the correlation between messages in the null model.

4.3 Experimental properties of a-motifs

We start by studying the differences between motifs and a-motifs. In order to have enough messages during activity periods, we take $\mu = 2$. Indeed, $\mu = 1$ implies that half of edges finish an activity period since half of the edges are separated by more than the inter-edge time median. In the datasets, $\mu = 1$ implies that these periods include a small amount of edges.

Zhao et al. [19] observed that star and chain motifs are the most common ones. Analysis of the corresponding a-motifs on our datasets confirm this observation in average (Fig. 3), despite a few exceptions for some datasets. Overall, the chain motif represents 16% of all motifs, stars represent 6%, while ping-pong comes third at 3%.

We also study the z-score of the frequency of each motif Fig.4. We can observe significant tendencies at least for 4 of the 5 studied motifs: in most networks, stars and chains are less common in observed data than in the null model, while spam and ping-pong are more common.

In [16], a similar analysis was conduced on a phone call dataset, only for stars and chains. While their conclusion for stars was the same than ours, their conclusion for chains was the opposite. This difference might be due to the difference in nature of datasets, or to a difference in the method of analysis: they segmented time using fixed temporal windows, while we used activity periods.

4.4 A-motifs and communities

In this section, we study the relation between a-motifs and communities. In particular, we are interested to know if some a-motifs are more common inside or in-between communities.

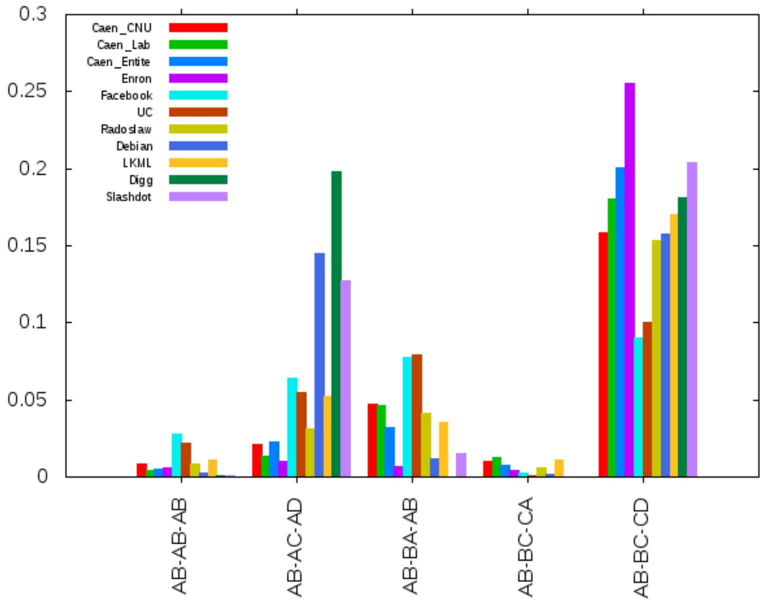

Fig. 3. a-motif frequencies for different networks.

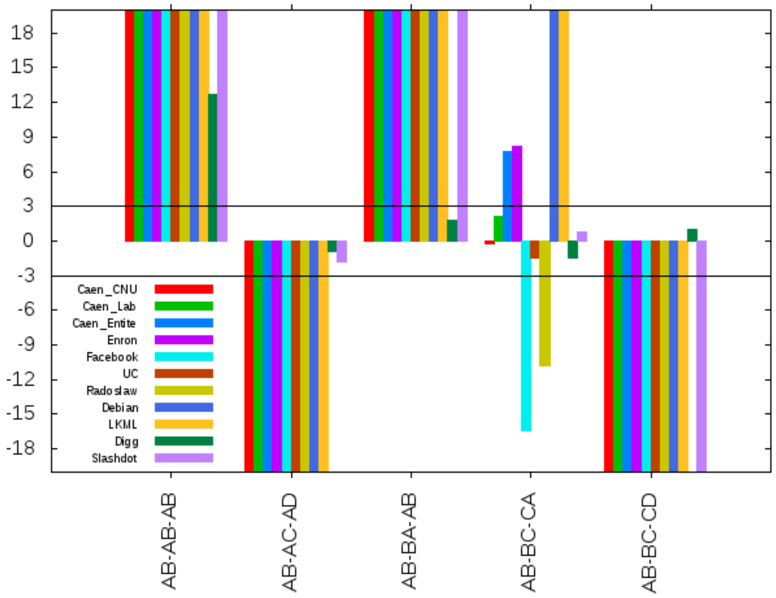

Fig. 4. z-score of a-motif frequency for different networks. Scores above 3 in absolute value are considered significant. Values beyond 20 are truncated.

We define the normalised internal weights of a-motifs of type m as:

$$w_{in}^{norm}(m) = \frac{w_{in}(m)}{\sum_{m\prime \in M} w_{in}(m\prime)}$$

with $w_{in}(m)$ the sum of weights of a-motifs of type m that have at least an edge inside a community. We similarly define the normalised external weights.

We now compute a normalised cross-community score for a-motifs of type m:

$$ccscore(m) = \frac{w_{ext}^{norm}(m) - w_{in}^{norm}(m)}{max(w_{ext}^{norm}(m), w_{in}^{norm}(m))}$$

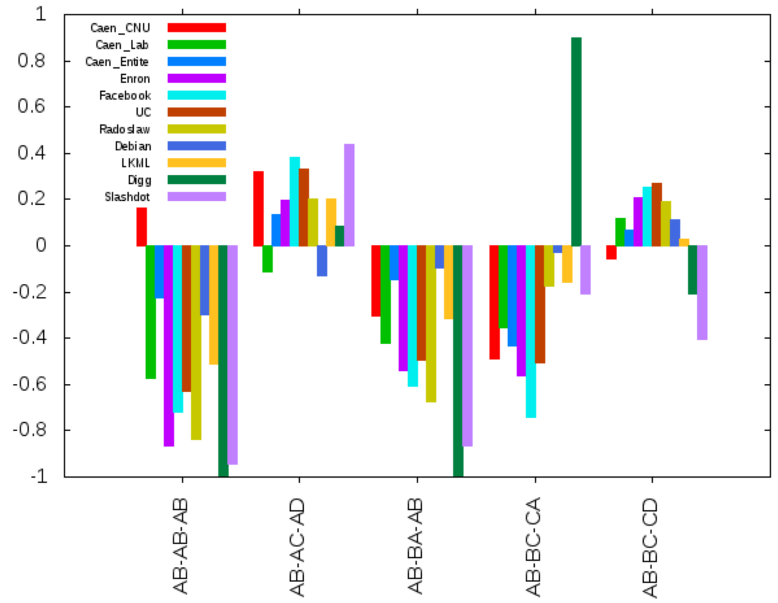

Fig. 5. Ratio between external and internal proportions of a-motifs

Interpretation of ccscore This score can vary between -1 and 1, with positive score indicating a higher relative prevalence of cross-community instances, while negative values indicate a-motifs more commonly found inside communities. Results are presented Fig. 5.

We observe that three a-motifs have negative scores in most datasets: spams, ping-pong and triangles. This means that, comparatively to others, these a-motifs tend to occur more inside communities than outside.

The two other a-motifs (star and chain) have less clear tendencies, but seem to occur slightly more often in-between communities.

It is nevertheless important to note that there are notable exceptions to these tendencies, in particular the Caen CNU dataset for spam and chains, or a divergent result for triangles on Digg.

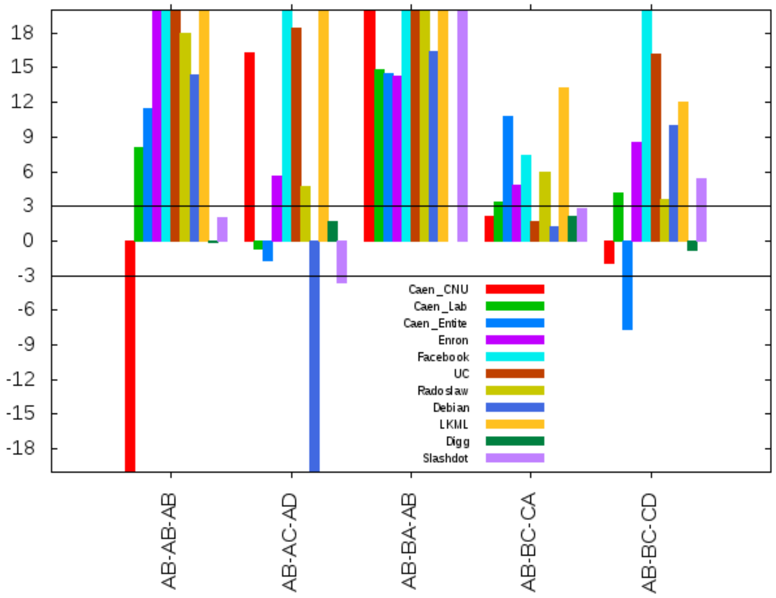

Fig. 6. z-score of ccscores

z-score of ccscore As previously, we compute the z-score of the ccscore in order to evaluate how significant are the tendencies (see Fig. 6). We observe that most values are significantly higher than those in the null-model, therefore that ccscores observed in the dataset are higher than those in the null model. We conclude that studied a-motifs appear more frequently between communities with respect to the the null-model.

4.5 Discussion

In previous sections, we have observed that some a-motifs are more likely to occur inside or outside communities, and that these patterns are significant. As a consequence, we propose that a-motifs could be used, given a temporal network dataset, to distinguish internal and external edges. Identifying such edges could be used to later identify communities.

Another observation is that a-motifs occurring more frequently inside communities seem to be different in nature from those occurring outside. On one hand, inter-community edges are marked by patterns of diffusion of information, including various, different actors: chains and stars. On the other hand, motifs observed inside communities are characterised by an information travelling inside a same set of actors, either several times the same pair of actors (spam, ping-pong), or a cycle coming back to its origin (triangle).

Finally, it is interesting to observe that results are coherent between datasets with ground truth communities (Caen-university) and those in which topological communities have been discovered using the Louvain algorithm. It implies that observed temporal properties are characteristics of structural communities.

5 Conclusion

In this paper, we present an alternate definition of temporal motifs that takes into account the activity periods in communication networks. We measure a large difference of the frequency of these motifs between the empirical data and a null model that ignores causality. This result suggests that our definition captures causally-linked communications.

We also studied the relationship between temporal motifs and community structure. We observed that the conversational motifs such as spam, ping-pong and triangle are generally more frequent inside communities than outside. The star motif, on the other hand, appears more frequently outside communities. The comparison with the null model shows that causally-linked motifs happen frequently outside communities.

These results open the way for future works: on the one hand, it could be possible to detect communities in link streams based on the frequency of a-motifs, taking advantage of our observations. On the other hand, a more detailed analysis of the nature of interactions occurring inside a-motifs could help us to understand better why some of them occur more often inside or outside communities, hence improving the global understanding of the structure of communications.

Acknowledgments

This work is funded in part by the European Commission H2020 FETPROACT 2016-2017 program under grant 732942 (ODYCCEUS), by the ANR (French National Agency of Research) under grants ANR-15-CE38-0001 (AlgoDiv) and ANR-13-CORD-0017-01 (CODDDE), by the French program "PIA - Usages, services et contenus innovants" under grant O18062-44430 (REQUEST), and by the Ile-de-France program FUI21 under grant 16010629 (iTRAC).

References

1. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.
2. A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

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, 2008.
4. R. Cazabet and F. Amblard. Dynamic community detection. In *Encyclopedia of Social Network Analysis and Mining*, pages 404–414. Springer New York, 2014.
5. M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Social synchrony: Predicting mimicry of user actions in online social media. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 151–158. IEEE, 2009.
6. N. Gaumont, T. Viard, R. Fournier-Sniehotta, Q. Wang, and M. Latapy. Analysis of the temporal and structural features of threads in a mailing-list. In *Complex Networks VII*, pages 107–118. Springer, 2016.
7. V. Gmez, A. Kaltenbrunner, and V. Lpez. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645–654. ACM, 2008.
8. M. Karsai, M. Kivel, R. K. Pan, K. Kaski, J. Kertsz, A.-L. Barabsi, and J. Saramki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2), Feb. 2011.
9. B. Klimt and Y. Yang. Introducing the enron corpus. In *CEAS*, 2004.
10. L. Kovanen, M. Karsai, K. Kaski, J. Kertsz, and J. Saramki. Temporal motifs. In *Temporal Networks*, Understanding Complex Systems. Springer, Berlin, Heidelberg, 2013.
11. R. Michalski, S. Palus, and P. Kazienko. Matching Organizational Structure and Social Network Extracted from Email Communication. In *Business Information Systems*, volume 87, pages 197–206. Springer Berlin Heidelberg, 2011.
12. R. Milo. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, Oct. 2002.
13. T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social Networks*, 31(2):155–163, May 2009.
14. F. Pukelsheim. The three sigma rule. *The American Statistician*, 48(2):88–91, 1994.
15. M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.
16. L. Tabourier, A. Stoica, and F. Peruani. How to detect causality effects on large dynamical communication networks: a case study. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference On*, pages 1–7. IEEE, 2012.
17. B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009.
18. Yi-Qing Zhang, Xiang Li, Jian Xu, and A. Vasilakos. Human Interactive Patterns in Temporal Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):214–222, Feb. 2015.
19. Q. Zhao, Y. Tian, Q. He, N. Oliver, R. Jin, and W.-C. Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1645–1648. ACM, 2010.