On The Impact of Users Availability In OSNs

Antoine Boutet, Anne-Marie Kermarrec, Erwan Le Merrer, Alexandre Van Kempen

On The Impact of Users Availability In OSNs

Antoine Boutet
INRIA Rennes
antoine.boutet@inria.fr

Anne-Marie Kermarrec
INRIA Rennes
anne-marie.kermarrec@inria.fr

Erwan Le Merrer
Technicolor
erwan.lemerrer@technicolor.com

Alexandre Van Kempen
Technicolor
alexandre.vankempen@technicolor.com

Abstract

Availability of computing resources has been extensively studied in literature with respect to uptime, session lengths and inter-arrival times of hardware devices or software applications. Interestingly enough, information related to the presence of users in online applications has attracted less attention. Consequently, only a few attempts have been made to leverage user availability pattern to improve such applications. Based on an availability trace collected from MySpace, we show in this paper that the online presence of users tends to be correlated to those of their friends. We then show that user availability plays an important role in some algorithms and focus on information spreading. In fact, identifying central users *i.e.* those located in central positions in a network, is key to achieve a fast dissemination and the importance of users in a social graph precisely vary depending on their availability.

Categories and Subject Descriptors C 2.1 [*Network Architecture and Design*]: Network Topology; C 2.0 [*General*]: Data Communications

Keywords Availability, Time Varying Graphs, Information Spreading, Centrality

1. Introduction

Online Social Networks (OSNs) now attract hundreds of millions of users around the world and represent a crucial place to gather data with countless possibilities of analysis, ranging from extracting user habits to influencing recommendation systems. Information on the availability of resources are usually of crucial importance to build reliable applications in computer systems. While this availability has been extensively measured, studied and leveraged for computers and application runtimes [9, 11, 13, 15], we know little about the online presence patterns of users and their impact on specific features in the case of OSNs [2, 7]. In this paper, we show that online presence is very heterogeneous among users, and that complex interactions may influence it. This constitutes a significant difference with studies about computing devices. More specifically, observing device uptime patterns only allows coarse grain observations such as extracting the correlation between day and night patterns or between the presence of users and the occurrences of significant worldwide events (sport event, holidays, etc). On the opposite, the online presence of users on a given OSN enables to extract finer characteristics and correlations such as the simultaneous presence of users with respect to their friends.

This availability information can be of particular interest to various practical functionalities such as efficient information spreading and influence in OSN [16], load prediction on the service provider's platform [12], or resource management in distributed social networks. In addition, while the 'who-knows-who" relations of OSN structures are inherently *time varying graphs* [4], extracting at a fine granularity the user presence is interesting. Indeed, some recent works emphasize the fact that graph analysis aggregating the interactions usually miss the time dimension [4, 8, 16].

In this paper, we investigate the dynamic nature of OSNs, based on a trace we have extracted from Myspace[1]. This trace contains both the social network structure (friendship relations), as well as availability information in each timeslot. While a majority of related works [2, 7] focus on models that fit session lengths, inter-arrival times or rely on browsing habits of users, here we are interested in studying the correlations between users availability and those of their friends

SNS'12 April 10, 2012, Bern, Switzerland

[1] http://www.myspace.com

and its impact on information spreading. From our analysis, we conclude that the availability of a user is highly correlated with that of her friends: on average a user is 42% more likely to be online when 50% of her friends are connected. In addition, leveraging this availability information to spread information in the network leads to an 8.5% increase in the number of nodes reached by the information during the dissemination, while speeding up the whole process.

2. User availability in MySpace

Accurate information on user activity on a given social platform can only be obtained by the service operator or the ISP. Given the fact that we did not have direct access to this information, we used the *online status* information displayed on the user profile in MySpace.

2.1 Measurement and methodology

We first crawled MySpace for a few weeks and collected for each user, her profile and the one of her explicit friends in the OSN. From this first dataset, we only kept the users who were online at least once during the crawling period. There are two kinds of profile in the MySpace OSN: regular and music band. We conducted an analysis on standard users, by removing the band profiles. In addition, in order to obtain a connected component (friendship relations are non symmetrical in MySpace) and to observe the availability of users over a subset of the MySpace network, we selected the most active French users in the same weakly connected component during this first period. This represents 833 users. In MySpace, the list of online friends of a specific user is available on a dedicated webpage (possibly including band profiles). We collected that information for each of these 833 preselected users from that webpage, yielding about $180,000$ tracked MySpace users. The period of observation was 17 days, in May 2011. For each user, we collected availability information every 10 minutes.

Surprisingly, although consistent with observations made in [7] on the loss of interest of users after an initial adoption period, only $13,286$ among those $180,000$ users had been indicated as online at least once. Yet, a potential second reason for the unavailability may also come from the fact that a user can hide her availability information for privacy purposes.

The resulting graph used in this study contains $13,286$ users, along with $68,064$ friendship relations[2]. This graph can be observed, as traditionally, in a static aggregated view, or in a more realistic time varying manner with nodes and edges appearing and disappearing at a 10-minute granularity.

2.2 Static vs time varying MySpace graph

In contrast to a static graph which aggregates the users and relations observed at least once during the whole crawling period, the time varying graph reflects actual availability for each of the $2,500$ timeslots of 10 minutes. Table 1 and Figure 1 show the structural properties of both directed graphs, as well as the associated degree distribution (the SNAP library[3] has been used for the graph statistics presented in this paper).

We observe that the 90-percentile effective diameter [6] for the time varying graph is significantly spread around its average. In addition, the average degree for the time varying graph is reduced by almost 40% when compared to the value depicted for the static graph. As shown in Figure 1, the distribution degrees for the static graph and the aggregate of all time varying graphs are also notably different. These observations confirm the relevance of considering a time varying graph in the particular context of OSNs, thus providing more information.

Characteristic	Static	Time Varying		
		min	avg	max
Nodes	13,286	2214	3668	3997
Edges	68,064	5764	11706	13858
Avg degree	9.60	5.20	6.30	7
90-percentile effective diameter	5.80	5.50	6.15	8.21
Average Clustering Coefficient	0.131	0.117	0.130	0.1482

Table 1: *Structural properties: static vs time varying graph.*

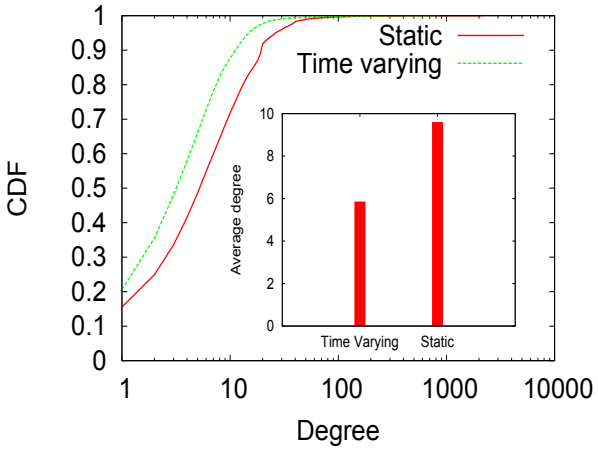


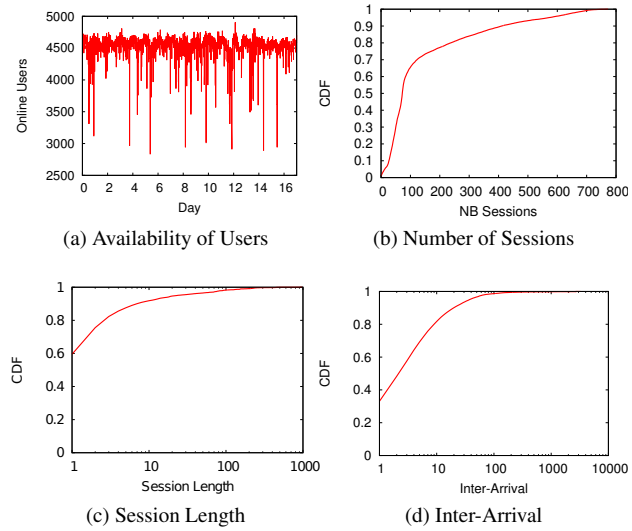
Figure 1: *Degree distribution: static vs the aggregate of time varying graphs.*

2.3 User availability on the platform

This section presents statistics on users availability during the observed period: (i) the number of users connected over time, (ii) the number of sessions, (iii) the session length and (iv) the inter-arrival times. The number of sessions shows the number of connections for each user to her account. The session lengths captures how long users remain online each time they get connected and inter-arrival times illustrates the interval between the moment a user departs and her next

[2] We make this graph publicly available on request.

[3] http://snap.stanford.edu/



(a) Availability of Users (b) Number of Sessions

(c) Session Length (d) Inter-Arrival

Figure 2: *Information about the availability of users (*1 *unit equals* 10 *minutes).*

arrival. Figure 2 depicts the distribution of those availability statistics.

We observe that the availability patterns of users connected at the same time vary according to the instant in the day and week. The number of users connected simultaneously does not exceed $4,800$ and is relatively stable over time. This can be explained by the fact that selected users are from the same country (France) but their friends may come from various countries on different time zones. However, daily patterns are clearly observable on the figure. The Cumulative Distribution Functions (CDF) show that 50% of users have less than 50 different sessions during the observed period (an average of 1.6 sessions per day) while most of the sessions are short. In fact 60% of the sessions last less than 10 minutes. In addition, more than 80% of inter-arrival times are shorter than 2 hours. Note that these observations are consistent with measures previously made in [2].

3. Correlation between friends

Social networks like MySpace enable users to interact with each other. These interactions may influence the behavior of users such that they do not behave independently. For example, an application such as the instant messaging service offered by MySpace may push users to share some connection patterns (by chatting together). The goal of this section is to evaluate to what extent the uptime patterns of users are correlated with those of their friends on our dataset. The underlying question is to evaluate if friends are connected at the same time and thus characterize how *active* are the links between them.

We define the notion of availability correlation between two users as a measure on how similar their availability patterns are. This correlation is evaluated using a *cosine*

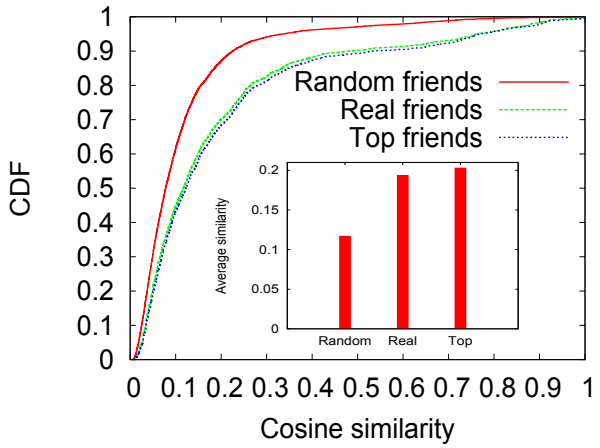


Figure 3: *Correlation between the user's availability and the one of their friends.*

similarity metric, which is frequently used when trying to determine similarity between documents, user profiles, or in our case availability patterns. The availability of each user is represented as a vector of a predefined size. For each timeslot, the corresponding vector entry is set to 1 if the particular user was online at that time and 0 otherwise. As uptime sessions are relatively short, we choose a ten-minute timeslot in order to capture all the connection events.

More formally, let $\overrightarrow{A_x}$ and $\overrightarrow{A_y}$ be the availability vector of respectively user x and y. In practice, the cosine similarity $S_{x,y}$ is modelled as the cosine of the angle between the two vectors $\overrightarrow{A_x}$ and $\overrightarrow{A_y}$.

$$S_{x,y} = \left(\frac{\overrightarrow{A_x} \cdot \overrightarrow{A_y}}{\|\overrightarrow{A_x}\| \cdot \|\overrightarrow{A_y}\|}\right)$$

- $S_{x,y} = 1$: Perfect uptime correlation between x & y
- $S_{x,y} = 0$: No uptime correlation between x & y

Note that a correlation implicitly exists at the dataset level as users of the same country tend to connect during daytime for example. However in order to check if explicitly declaring friends had an impact on this correlation, we computed the correlation between each user and her friends. For comparison purposes, we also computed the correlation between a user and a given number of friends, chosen randomly in the dataset; this dissociates the behavior of friends, allowing us to measure the impact of friendship on uptime correlation. As MySpace offers the possibility to declare *top friends*, we finally computed the correlation for each user with her top friends.

The cumulative distribution functions of correlation values are plotted in Figure 3. The first observation is that correlation values are rather low with random friends as well as with friends and top friends. Even if the correlation with top friends is the best one, we note that there is almost no differ-

ence between the correlation with friends and top friends. However we clearly see that the correlation with explicit friends is higher than with random ones (twice as much), thus showing the impact of friendship on availability correlation.

In the previous experiment we studied the correlation at the user level i.e. between pairs of user. To go further we also evaluated how this correlation between a user and her friends impacts the probability for this user to be connected by taking into account the aggregated behavior of her friends. In other words, we wanted to answer the following question: is the probability of a user to be connected at a given time impacted by the fact that her friends are also connected at this particular time or are these two independent events? For example if 50% of her friends are connected, what is the statistical probability that she is also connected?

In order to estimate this impact, we computed for each user the probability to be connected depending on the proportion of her friends being effectively connected. For the sake of clarity we distinguished two cases, one when less than 50% of her friends are connected, and the other when more than 50% are online. This probability is simply evaluated using conditional probabilities. More formally, let UP be the event "User is UP", and $More50$ be the event "More than 50% of her friend are UP" then:

$$\Pr(UP \mid More50) = \frac{\Pr(UP \cap More50)}{\Pr(More50)}$$

where $\Pr(UP \cap More50)$ is the number of timeslots when the user and more than 50% of her friends are connected, normalized by the total number of timeslots. $\Pr(More50)$ is the number of timeslots when 50% of her friends are connected, normalized by the total number of timeslots.

Results are plotted on Figure 4. The cumulative distribution function of all *a priori* probabilities to be connected is also plotted as a means of comparison. In absolute term, all these probabilities are rather low. However, results clearly show that users are more likely to be online when their friends are, and on the contrary this probability decreases when the majority of their friends are not connected.

This correlation between friends' behavior may have a direct impact on applications requiring the computation of metrics like *centrality* (*i.e.* reflecting the importance of a given user in the graph) for example. In fact even if a user is qualified as central according to the static graph, her centrality may not be representative if she is never connected when her friends are. In other words the availability knowledge of users is an important aspect when looking for central nodes in a social graph.

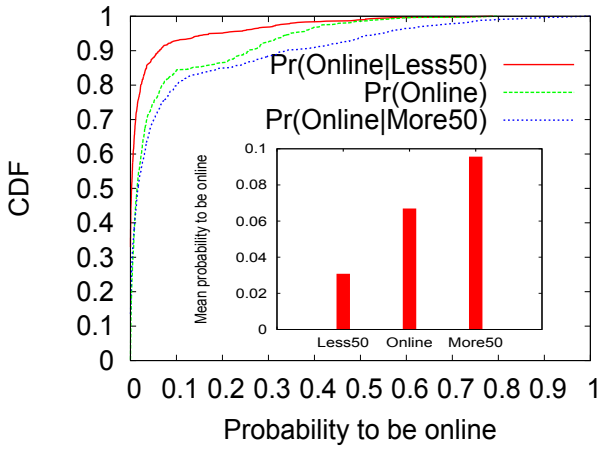


Figure 4: *Users are more likely to be online if their friends are present.*

4. Impact of Availability on Information Spreading

Users which are able to spread information quickly to a large number of users are commonly identified using *centrality measures* [10, 14, 16] (including degree, betweenness or eigenvector centralities). In this section we evaluate the impact of the availability of users on information spreading. To this end, we compared the spreading process depending on the (importance of the) user that initiates the dissemination (seed user). We ranked seeds based on their centrality extracted from both the static and the time varying graphs. Note that in this section, measures and analysis are conducted over the undirected version of the graph. This undirected graph models a bidirectional communication channel between users as a chat application for instance.

Every second timeslot in the interval $[0 : 500]$, a seed s was elected to spread a message in the network. A user u was infected by a message m once she was connected at one timeslot to a user who get already infected by the message m. A total of 250 messages were thus injected in the system offering about $3,250,000$ of possible infections if each user received each message.

We measured both the size and the spreading time by observing the number of users reached by a new message from every user. We also looked at the time needed by an injected message at a seed to propagate and to be received by reachable users in the time varying graph. For instance, assuming the seed s injected the message m at the timeslot t, and a user u connected to s in the time varying graph and online for the first time at the timeslot $t + \alpha$, the reception time of m for this user is equal to α. We believe this simple model fits the example of live news propagation.

We then compared different seed selections. The seed was selected among the online users at each timeslot, based on its centrality measure on the actual graph (we use the

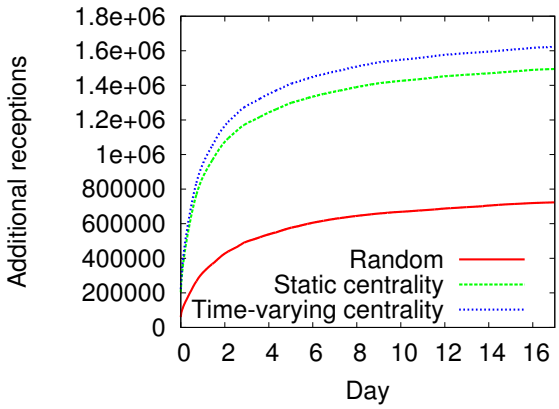


Figure 5: *Number of infected nodes, depending on the seed selection method (centrality computed on the static graph, on the time varying graph, or randomly on the time varying graph).*

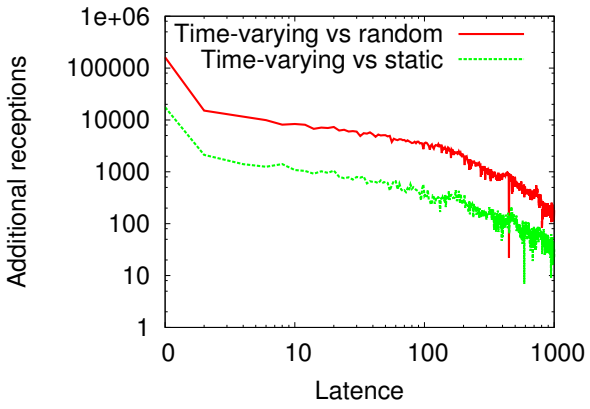


Figure 6: *Seed selection using the centrality metric, computed on the time varying graph, speeds up the spreading (normalized comparison).*

betweenness centrality [3]), or simply randomly chosen. In order to evaluate the impact of the availability of users, we compared this approach against the one selecting the most central node in the static graph as an initial seed. In addition, in order to model multiple channels of communications, in case of the static graph, the seed was chosen randomly among the 10 users with the higher centrality measure.

Figure 5 depicts the number of infections over time according to the seed selection. Results show that seeds selected using the time varying centrality measures increase the dissemination process by 8.5% at the end of the experiment and by 7% at the timeslot 1000, when compared to a selection based on the static graph. It is interesting to note that only about half of the users were infected at the end of the experimentation ($1,600,000$ against $3,250,000$). In addition, we show that random selection reduces by more than 50% the spreading as compared to the selection based on the time varying graph.

We now consider the infection speed. Figure 6 depicts the additional infections of the time varying centrality based selection and the one based on the static centrality and randomly according to the latency of the infection. Results show that the highest number of additional receptions is mainly achieved quickly after the injection from the seeds. This phenomenon highlights the importance of selecting initial seeds according to the current topology of the social network taking into account the availability of users for a fast and efficient dissemination. Indeed, seeds connected to more people or better placed in the time varying graph, infect more users directly and meanwhile reduce the infection latency. Almost $20,000$ direct additional infections are achieved thanks to seeds selected according to their time varying centrality measure, as compared to the static one. Only $2,000$ additional infections are made possible with a latency of 1 (infection 10 minutes after the injection). In addition, latency is drastically reduced as compared to random seeds with almost $200,000$ additional direct infections using seed selection in the time varying graph.

5. Related Works

There have been several studies on availability of computing resources, mainly in the field of distributed and peer-to-peer systems. When resources availability patterns (mostly on/off patterns) are well understood, prediction systems can bring significant gains on reliability, load balancing, resources savings in terms of bandwidth and storage space [9, 11, 13, 15].

Interestingly enough, availability patterns of users on online applications have been far less considered, while we believe such studies can yield interesting and new insights. In [7], the time spent online by users of Bebo, Myspace, Netlog and Tagged has been studied through statistical analysis; authors discovered that a Weibull distribution accurately models the behavior of 80% of users. Session lengths, as well as the number of sessions, follow power law distributions. Notably, authors note a clear loss of interest for part of the users over a short time period (they do not reconnect frequently afterwards, or simply leave), after a few initial days of activity when registering to the service. This remark underlines the fact that analysis should not be conduced on a basic aggregated friendship, as user access to the service is fairly variable. Paper [2] studies Orkut, MySpace, LinkedIn and Hi5 OSNs. Authors concluded that session lengths follow a heavy-tailed distribution, while inter-arrival times a Lognormal distribution. It also describes browsing habits of users on their friends' pages. Our work is complementary to those findings, as we address the impact of users sessions over information spreading and study the correlation between uptime periods of friends.

The general Time Varying Graph notion captures in a framework the dynamics of such a complex system [4, 14]. While dynamic objects that are currently considered are the relations between people or computers (i.e. the graph edges),

node dynamics can be trivially added to fit the framework. This would allow traditional algorithms to be expressed or redesigned in the context of inherently non static networks, as for instance dissemination algorithms [5]. As we provide the dataset along with this paper, it could be leveraged for future research on time varying graphs. Paper [16] studies the varying importance of persons in a network (with the help of the *centrality* toolbox). The Enron dataset (mail exchanged within the company) is used as a basis for analysis, and to point out that temporal metrics should be designed instead of the static known ones. Two new centralities are defined, in order to better reflect the varying importance of individuals as a function of their actual interactions over time in the network. We choose to study in this paper if similar conclusions could be highlighted when considering an explicit online social network.

6. Conclusion

It has been well understood that the availability of resources that a computer system relies on is critical for both reliability and performance. In this paper, we have shown that the information of the availability of users in online social networks has also an important impact. More specifically, results demonstrate that users availability may also be leveraged in order to provide a more efficient information dissemination. This confirms the need to take into account this crucial time varying metric when developing online applications. Secondly, we have shown that users availability is not only dependent of the daytime, but also correlated to the presence of their friends on the platform. This observation may serve in future works for the development of more accurate and fine-grained models of users' behavior on specific online applications.

Acknowledgments

This work was partially funded by the ERC Starting Grant Gossple number 204742 [1].

References

[1] Gossple project: `http://www.gossple.fr`.

[2] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *IMC*, 2009.

[3] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[4] A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro. Time-varying graphs and dynamic networks. *CoRR*, abs/1012.0009, 2010.

[5] K. Censor-Hillel and H. Shachnai. Fast information spreading in graphs with large weak conductance. In *SODA*, 2011.

[6] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, 1999.

[7] L. Gyarmati and T. A. Trinh. Measuring user behavior in online social networks. *IEEE Network*, 24(5):26 –31, 2010.

[8] H. Kim and R. Anderson. Temporal node centrality in complex networks. *Phys. Rev. E*, 85:026107, 2012.

[9] S. Le Blond, F. Le Fessant, and E. Le Merrer. Finding good partners in availability-aware p2p networks. In *SSS*, 2009.

[10] E. Le Merrer and G. Trédan. Centralities: capturing the fuzzy notion of importance in social graphs. In *SNS*, 2009.

[11] J. W. Mickens and B. D. Noble. Exploiting availability prediction in distributed systems. In *NSDI*, 2006.

[12] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez. The little engine(s) that could: scaling online social networks. In *SIGCOMM*, 2010.

[13] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz. Handling churn in a dht. In *ATEC*, 2004.

[14] N. Santoro, W. Quattrociocchi, P. Flocchini, A. Casteigts, and F. Amblard. Time-varying graphs and social network analysis: Temporal indicators and metrics. In *SNAMAS*, 2011.

[15] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A measurement study of peer-to-peer file sharing systems. In *MMCN*, 2002.

[16] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia. Analysing information flows and key mediators through temporal centrality metrics. In *SNS*, 2010.