RECAST: Telling apart social and random relationships in dynamic networks

Pedro O.S. Vaz de Melo [a,*], Aline Carneiro Viana [b], Marco Fiore [c], Katia Jaffrès-Runser [d], Frédéric Le Mouël [e], Antonio A.F. Loureiro [a], Lavanya Addepalli [b,f], Chen Guangshuo [b]

[a] *Universidade Federal de Minas Gerais, Brazil*
[b] *INRIA, France*
[c] *CNR-IEIIT, Italy*
[d] *University of Toulouse, INPT-ENSEEIHT, IRIT, France*
[e] *University of Lyon, INSA Lyon, INRIA CITI Lab, France*
[f] *Polytechnic University of Valencia, Spain*

When constructing a social network from interactions among people (e.g., phone calls, encounters), a crucial task is to define the threshold that separates social from random (or casual) relationships. The ability to accurately identify social relationships becomes essential to applications that rely on a precise description of human routines, such as recommendation systems, forwarding strategies and opportunistic dissemination protocols. We thus propose a strategy to analyze users' interactions in dynamic networks where entities act according to their interests and activity dynamics. Our strategy, named *Random rElationship ClASsifier sTrategy (RECAST)*, allows classifying users interactions, separating random ties from social ones. To that end, RECAST observes how the real system differs from an equivalent one where entities' decisions are completely random. We evaluate the effectiveness of the RECAST classification on five real-world user contact datasets collected in diverse networking contexts. Our analysis unveils significant differences among the dynamics of users' wireless interactions in the datasets, which we leverage to unveil the impact of social ties on opportunistic routing. We show that, for such specific purpose, the relationships inferred by RECAST are more relevant than, e.g., self-declared friendships on Facebook.

1. Introduction

In network theory, edges represent some kind of relationship between the vertices they connect. In social networks, i.e., networks where the vertices map to individuals, edges may represent, among others, friendship, work interactions, similarity [1]. When building such a network, edges can be derived from explicit information, such as declared friendship on Facebook, or from implicit knowledge inferred from the reciprocate behavior of the vertices. The second type of interactions may be, e.g., phone calls [2], physical encounters [3], or chat messages [4], just to mention some examples from the literature. However, as in all inferences, there is always a risk of incurring in mistakes when drawing (or not) an edge between two

vertices. For instance, can we label two individuals as friends if they communicated via phone calls only once in the past? And twice? And ten times? What is, in this case, the threshold that separates, e.g., friends from acquaintances?

The difficulty in answering these questions comes from the fact that we generally do not know which interactions were driven from social reasons and which were not. For instance, some phone calls are made to invite friends for a dinner party, i.e., they are the result of socially-driven interactions, whereas other calls may be placed just to order food, i.e., they are not represent any stable relationship between the involved individuals. The ability to accurately identifying social relationships becomes then essential to applications that rely on a precise description of human routines, such as recommendation systems, forwarding strategies and opportunistic dissemination protocols. As an example, if an administrator of a mobile access network knows which customers share a social tie she/he selectively diffuse data among clients so as to offload the network infrastructure [5–7].

In this paper, we propose a *Random rElationship ClASsifier sTrategy (RECAST)* to classify relationships among users from their past interactions. RECAST can classify a relationship as *random*, i.e., as the result of an occasional encounter, or *social*, separating in the second case *Friends*, *Bridges*, or *Acquaintances* interactions based on the nature of the contacts. To that end, RECAST examines how the real system would evolve if users' decisions were completely random. More precisely, we use the temporal graphs originated from the real network dataset and a random counterpart of the same to tell apart edges representing random events from those created by actual social relationships, such as friendship or professional interactions. By comparing the two graphs in terms of metrics reflecting two major social features, i.e., frequent user interactions and shared acquaintances, RECAST provides a simple yet very effective way of classifying contacts. RECAST has a single, intuitive and easily configured parameter, which makes our technique preferable to conventional methods for filtering and cleaning social network data—which require instead arbitrary thresholds, involve many parameters, and often demand a deep knowledge of the system.

RECAST builds on the high predictability of human behaviors [8], known to be primarily driven by regular, routine activities. As a consequence, social connections among individuals nodes can be modeled by mechanisms such as preferential attachment [9,10] and triangle formation [11,12], which leverage the existence of communities or highly connected hubs [13] in the network of social contacts. This makes social networks different from random ones, such as the Erdős and Rényi network [14], where node connections are instead purely stochastic. However, we remark that random events are possible also in real social systems. These are hardly predictable situations that deviate from the regular patterns that dominate the network formation, and, for their own nature, are occasional and unlikely to arise repeatedly in time. Additionally, random events tend to veil social patterns by introducing a significant amount of noise, making the process of knowledge discovery in social datasets more complex.

When applied to five real-world datasets describing user activity in city and campus scenarios, the RECAST classification allows observing significant differences in the evolution of relationships in the different scenarios. Diversities are due to the intrinsic features of each environment. For instance, we show that the dataset describing the movement of cab drivers in San Francisco, CA, USA [15] has mostly non-social properties, which makes its graph representation similar to a random network. The same is not true for people moving in a campus: however, different campuses yield dissimilar interaction dynamics as well. As a consequence, we stress that conclusions drawn from evaluating a single dataset cannot be generalized, and that the validation of any networking protocol or service has to consider multiple heterogeneous scenarios. Along similar lines, we also show that the neat classification of user relationships provided by RECAST can be leveraged for networking purposes, i.e., to take opportunistic forwarding decisions when disseminating delay-tolerant contents within the social network.

In summary, the contribution of this paper is threefold:

- We introduce RECAST, a simple yet very effective way of classifying wireless contacts by leveraging metrics that reflect two major features of social networks: frequent user encounters and shared acquaintances (Section 5).
- We unveil the large differences among contact datasets, and claim that conclusions drawn from evaluating a single dataset should not be generalized; rather, the validation of networking protocols or services has to consider different types of datasets (Section 5.1).
- We show that the knowledge of the relationship given by RECAST can be leveraged for the design of opportunistic epidemic forwarding and is more useful for this purpose than self-declared friendships on online social networks (Section 6).

The paper is organized as follows. Section 2 discusses the related work. Section 3 details the design behind RECAST. Section 4 shows how we model real-world network traces into complex temporal networks. Section 5 details the RECAST implementation. Section 6 shows how RECAST can improve opportunistic routing solutions. Section 7 compares RECAST classes with friendships in online social networks. Finally, conclusions are drawn in Section 8.

2. Related work

Social network analysis builds on the high predictability of human behaviors [16], which are mostly driven by regular, routine activities. As a consequence, connections among social network nodes can be modeled by mechanisms such as preferential attachment [10] and triangle formation [12], that leverages the existence of communities or highly connected hubs [13] in the network. This makes social networks different from random ones, such as the Erdős and Rényi network [14], where node connections are purely stochastic, being determined by a constant probability.

Given this predictability, social ties have been widely exploited in opportunistic mobile networks so as to favor network services. The considered problems range from multi-hop message forwarding [17] and multicasting [3], to network security [18]. All studies above try to exploit the regularity and the repeated space–time patterns expected in the human behavior and tend to ignore the random or non-social connections among mobile users. Instead, we focus on the analysis of both random and social ties among users. To the best of our knowledge, the works of Miklas et al. [19] and Zyba et al. [20] are those the most closely related to ours. These studies differ in that they classify either *users* or their *interactions* (i.e., vertices or edges in the social graph), respectively. Zyba et al. distinguish social and vagabond users according to their social mobility behavior. They analyze regularity of appearance and duration of visits in a given area of traces to sort out users. Hence, the resulting classification only works on a per-individual per-area basis. Miklas et al. classify links between friends and strangers. They assume that frequent pairwise node encounters represent friendship interactions, and empirically decide that pairs of users meeting 10 days or more out of 101 days are friends, whereas others are strangers.

Overall, our work extends the investigations in [19,20] in the following ways. First, we propose a finer grained classifier, able not only to clearly characterize random interactions, but also to identify different kinds of social interactions: *Friends*, *Acquaintances* and *Bridges*. As such, we go a step further than [19] as we are able to identify edges corresponding to, e.g., *familiar strangers*, as defined by Milgram [21]: indeed, the users sharing a bridge interaction repeatedly encounter but may never experience an explicit social relationship. In addition, unlike the proposal by Zyba et al. [20], our strategy has no geographical dependency, i.e., it is of general validity.

3. Rationale

The techniques we show in this work can be applied to any dynamic social network. However, as a use case, from now on we will focus our analysis on social networks composed of individuals who are wirelessly connected over time via physical encounters. Encounters in these networks are driven by behaviors that tend: (i) to be regular and to repeat periodically; (ii) to build persistent communities of individuals or to generate commons acquaintances between them [22]. We refer to contact networks deriving from such systems as *Dynamic Complex Wireless Networks (DCWN)*. For instance, a contact network composed of wireless hand-held devices is a clear example of DCWN, since the user mobility creates neither purely regular nor purely random connections among the entities composing the network. The classification strategy we present in this paper leverages these two behaviors to efficiently distinguish social from random encounters in DCWNs. In the following, we detail our methodology and present the real world datasets considered in our analysis.

Social and random interactions: In DCWNs, interactions among the system entities are usually a consequence of semi-rational decisions. We say "usually" and "semi-rational" decisions because any system is subject to random events and irrational choices. Nevertheless, because most of the interactions still arise from conscious decisions made by their entities, the evolution of DCWNs is significantly different from the evolution of random networks, e.g., Erdős and Rényi networks [14]. Indeed, while in DCWNs the edges are created from semi-rational decisions, which tend to be regular and to repeat over time, in a *random network the edges are created independently of the attributes of the network entities, i.e., the probability of connecting any two entities is always the same*.

For instance, in a network of people, routines in everyday life correlate to individuals' interactions: if Smith and Johnson work at the same office, they are likely to meet at 9 AM during weekdays. This is because Smith and Johnson *decide* to go to work every day on time, as this is the rational decision to perform. However, their decision can be affected by random events, such as being stuck in traffic on a turnpike. Although rational decisions are regular, random events may occur as well.

In other words, an individual may take a *social decision*, or a *random decision*. Intuitively, if its probability of performing a social decision is greater than its probability of a random one, the network evolves to a well-structured social network. If the opposite it true, the network evolves as a random network, such as the Erdös and Rényi one.

Social communities: A major feature of DCWNs that we exploit in our study is the presence of communities, i.e., groups of individuals who are strongly connected to each other because they share the same interests or activity dynamics [12]. In contrast, communities cannot be found in random networks where, as previously stated, edges are created stochastically and independently of the attributes of each node.

The network clustering coefficient has been widely used to discriminate random from social networks. Given an undirected graph $G(V, E)$ (where V represents the set of network graph nodes, e.g., individuals, and E is the set of links describing relationships among entities, e.g., contacts among individuals), the clustering coefficient c_i of node i measures the probability that two of its neighbors to be also connected among them. Formally, it is calculated as $c_i = 2|E_i|/|N_i|(|N_i| - 1)$, where N_i is the set of neighbors of i, E_i is the set of edges between nodes in N_i and $|\cdot|$ is the cardinality of the included set. The clustering coefficient of the whole network is the average of all node clustering coefficients c_i, $\forall i \in V$.

By introducing the equivalent random network G^R as the random network constructed with the same number of nodes, edges and empirical degree distribution of its real world counterpart G, Watts and Strogatz [23] show that the clustering coefficient of a social network G is one order of magnitude higher than the clustering coefficient of G^R. Thus, when a given network G exhibits a clustering coefficient that is significantly (i.e., orders of magnitude) higher than that of its random equivalent G^R, then we can state that (part of) the decisions made by the entities that compose the network graph G are non-random.

Table 1
Datasets used in the presented investigations.

Dataset	Local	Number of entities	Duration	Entities type	Avg. # encounters/node/day
Dartmouth [24]	University campus	1156	2 months	Individuals	145.6
USC [25]	University campus	4558	2 months	Individuals	23.8
San Francisco [26]	City	551	1 month	Cabs	834.7
Sassy [28]	City	27	79 days	Individuals	69.4
UPB [29]	University campus	22	42 days	Individuals	6.56

Real-world DCWN datasets: Our evaluations are performed on five real-world datasets (also referred to as *traces* in the following) that describe movements of entities in campus and city scenarios. The *Dartmouth* dataset [24] is a mobility trace of more than 1000 individuals in the university campus, recorded over eight weeks using WiFi network access information. The *USC* dataset [25] is also a mobility trace in a campus scenario, comprising movement information of more than 4000 individuals over eight weeks, again collected through WiFi access. The *San Francisco* dataset [26] contains records of the mobility of 551 taxis in San Francisco, CA, USA, over one month, gathered through GPS logging at each cab in the urban area. For both the Dartmouth and USC traces, two individuals are assumed to generate a contact if they are using the same WiFi access point to connect to the wireless network on campus. In the San Francisco trace, two taxis are in contact if their distance is lower than 250 m. Extensive experimental analysis in [27] shows that a distance of 250 m grants a 50% packet delivery ratio in urban environments, under common power levels (15–20 dBm) and with robust modulations (3-Mbps BPSK and 6-Mbps QPSK). These first three datasets are be used in Sections 5 and 6 for RECAST assessment.

Moreover, we also evaluate RECAST on two other datasets, where we have information about the physical encounters among human users, as the Dartmouth and USC datasets, and also data about the self-reported social networks of these users. In both cases, the self-reported social network was collected directly from the Facebook pages of the users. This allows us to construct the social network of these users, that tells which pairs of users declared a friendship relationship between them on Facebook. The *Sassy* dataset [28] contains the information about 27 human users associated with University of St Andrews comprising 22 undergraduate students, 3 postgraduate students, and 2 members of staff. Each user carried a mobile IEEE 802.15.4 sensors (T-mote invent devices) that are able to detect each other within a radius of 10 m. At each detection an encounter is characterized and stored in the dataset. Participants were asked to carry the devices whenever possible over a period of 79 days starting on February, 15, 2008.

The *UPB* dataset [29] was also collected in an academic environment, at University Politehnica of Bucharest, from the mobility of 22 participants. The participants were twelve Bachelor students (one in the first year, nine in the third, and two in the fourth), seven Master students (four in the first year and three in the second) and three research assistants. The data was collected only inside the grounds of the faculty between 8 AM and 8 PM during week-days. Data was collected from November 11th, 2011 to December 22nd, 2011, using an Android application that registers contacts between mobile devices with Bluetooth. These two datasets are used in Section 7 to compare RECAST classes with friendships on online social networks.

In all cases, the contact events between two individuals are traced using start date of contact and its duration. Table 1 summarizes the features of the different datasets.

4. Modeling

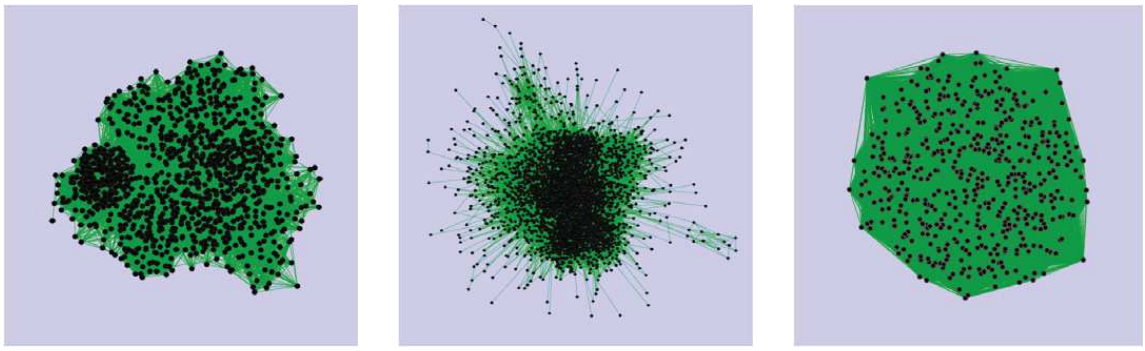
This section introduces the main properties of interest of the first three datasets, which will then be employed to evaluate the performance of RECAST in Sections 5 and 6. Section 4.1 shows how we model the dynamic DCWN as a temporal aggregation graph. Section 4.2 details the algorithm we employ to obtain the equivalent random graph of a given temporal aggregation graph, and discusses how these two graphs compare in terms of clustering coefficient.

4.1. Temporal aggregation graph

As previously discussed, the datasets we employ list contacts among individuals or vehicles, and associate to each encounter a start time and a duration. In order to generate the temporal graph, we discretize time into steps of duration δ,[1] and represent all the encounters occurring at time step k as a graph $\mathcal{G}_k(\mathcal{V}_k, \mathcal{E}_k)$. The set of vertices $\mathcal{V}_k$ is composed of all network nodes (i.e., individuals or vehicles) involved in a contact during the kth time step, while the edges in the set $\mathcal{E}_k$ represent the pairwise contacts during the same time step. Therefore, an edge between two nodes i and j, with $i, j \in \mathcal{V}_k$, exists in $\mathcal{E}_k$ if i and j have met during time step k.

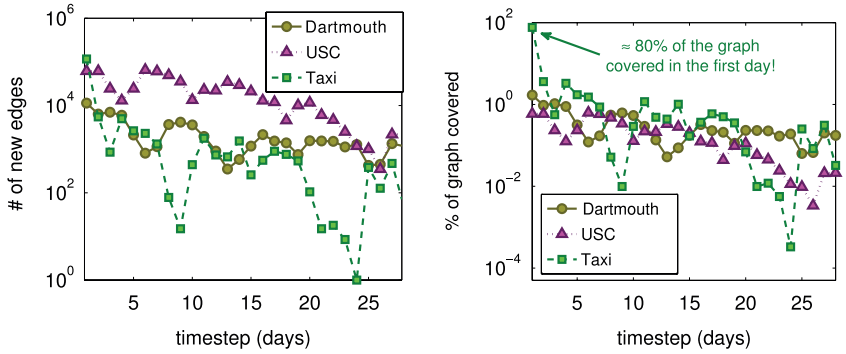
We can then define a time varying representation of the DCWN using a temporal accumulation graph $G_t = (V_t, E_t)$. Formally, $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup \cdots \cup \mathcal{G}_t\}$. As such, V_t (respectively E_t) is the set of all vertices (edges) that have appeared in the dataset between time 0 and time step t included. Note that G_t evolves over time and aggregates all the contacts in the dataset, thus comprising both social encounters and random encounters between network entities.

[1] In our study, we considered a duration of $\delta = 1$ day, since the datasets originate from human activities that feature daily routines.



(a) Dartmouth. (b) USC. (c) San Francisco.

Fig. 1. Snapshots of the temporal accumulation graph G_t after two weeks, for each examined dataset. In all cases, random encounters mix with non-random ones and hide the social structure of the contact network.



(a) Number of new encounters. (b) % of graph covered.

Fig. 2. The densification of the temporal graph $G_t(V_t, E_t)$ per day t. (a) The number of new edges added to $G_t(V_t, E_t)$ per day t. (b) The percentage of the fully connected graph that is covered by the new edges.

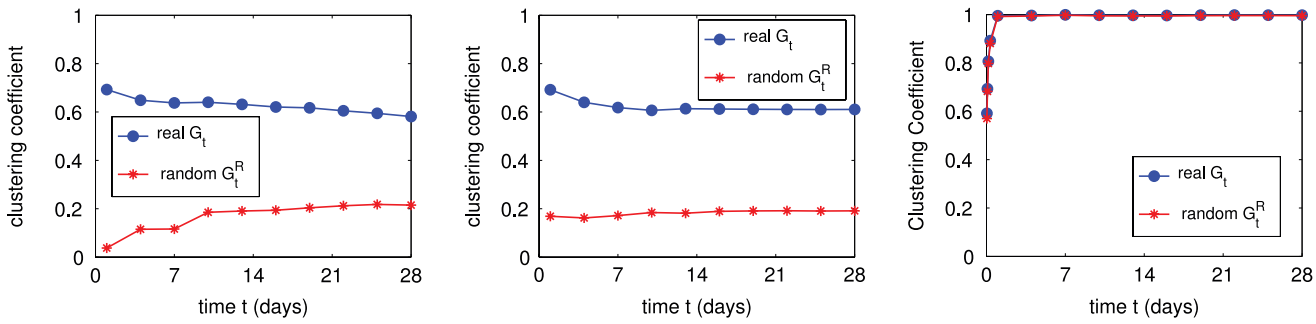
This aspect is clear in Fig. 1, showing the temporal accumulation graph G_t calculated when $t = 14$, i.e., for 2 weeks of data, drawn using a force-direct layout algorithm (FDLA) [30]. Although the FDLA draws the graph so there are as few crossing edges as possible, the presence of random contacts is mixed with the social structure of the network, making it difficult to extract useful knowledge from the temporal accumulative graph.

In Fig. 2, we show the densification of G_t. First, in Fig. 2(a), we show the number of new edges added by $\mathcal{G}_k$ to $\mathcal{G}_1 \cup \mathcal{G}_2 \cup \cdots \cup \mathcal{G}_{k-1}$ per each day k. Observe the oscillations in the curves, which are clearly a consequence of the presence of dynamics spanning over several days: namely, weekdays yield more activity, and thus more contacts, than weekends. Also, observe that the number of new edges added to the USC network is, in general, orders of magnitude higher than the other networks. However, when we normalize the number of new edges by the number of possible edges $\frac{|V_t| \times |V_t - 1|}{2}$, in Fig. 2(b), we see that the San Francisco dataset is the one with the highest densification, reaching almost 80% of the fully connected graph in the first day of the analysis. Note that the USC and the Dartmouth networks, both referring to campus environments, densify similarly.

4.2. Comparison with random graphs

The first step to analyze the mobility patterns of the temporal accumulation graph G_t is to build its random version G_t^R. The latter must feature similar topological characteristics as the original G_t graph, i.e., the same number of nodes, edges, and empirical degree distribution. That way, the only difference between G_t and G_t^R lies in the way nodes are connected to each other. While in G_t the nodes connect in a "semi-rational" way, in G_t^R the connections happen in a purely random fashion. As we will show later, this difference can be leveraged to accurately determine the extent of randomness in the mobility of individuals in DCWNs.

We use two algorithms to generate G_t^R from G_t. The first algorithm, which we will call RND, is well known in the network science community [31]. The algorithm $G^R = \texttt{RND}(G)$ receives a graph $G(V, E)$ as a parameter and returns a random graph $G^R(V, E^R)$ with the same topological characteristics as G, i.e., the same number of nodes, number of edges and degree distribution. In this way, we guarantee that the only difference between the real graph G and G^R is to whom each node connects, that is the focus of our analysis. Thus, given the degree distribution $D = (d_1, d_2, \ldots, d_n)$ of G with n nodes, this algorithm assigns an edge between nodes i and j with probability $p_{ij} = (d_i \times d_j) / \sum_{k=1}^{|V|} d_k$.



(a) Dartmouth. (b) USC. (c) San Francisco.

Fig. 3. Evolution of the clustering coefficient in the G_t of the three datasets, and in their random equivalents G_t^R.

The second algorithm, which we call T-RND, is an extension of RND and is able to generate random graphs from a temporal network G_t. As mentioned in Section 4.1, the temporal aggregation graph $G_t = \{\mathcal{G}_1 \cup \mathcal{G}_2 \cup \cdots \cup \mathcal{G}_t\}$ is the union of event graphs $\mathcal{G}_t$. Thus, the algorithm $G_t^R = \text{T-RND}(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_t)$ receives as parameters a set of consecutive event graphs $\mathcal{G}_t$ and returns a random temporal graph G_t^R. It constructs G_t^R by executing RND in each event graph $\mathcal{G}_t$ and then aggregating it in a way that $G_t^R = \{\text{RND}(\mathcal{G}_1) \cup \text{RND}(\mathcal{G}_2) \cup \cdots \cup \text{RND}(\mathcal{G}_t)\}$. In summary, both RND and T-RND randomly replicate the total number of contacts with distinct persons each individual had in a given time snapshot. In the following study, we generate multiple instances of T-RND and average the results over these instances.

We first demonstrate the potential of random comparison in Fig. 3, where we show the behavior of the clustering coefficient for graphs G_t and G_t^R over time for the three analyzed networks. As we have previously mentioned, the clustering coefficient is a good metric to differentiate social networks from random ones. As we observe in Fig. 3(a), for the Dartmouth dataset, the clustering coefficient of G_t and G_t^R are different in orders of magnitude over the first days. However, as time goes by, their values get closer, as random encounters grow in number and tend to veil the social network structure. On the other hand, as we see in Fig. 3(b), the clustering coefficients of G_t and G_t^R for the USC dataset are almost constant over time. However, the difference between them is not exceedingly high, since they have the same order of magnitude. We discuss these differences in detail later on.

Finally, as we observe in Fig. 3(c), the clustering coefficients of the San Francisco networks are practically the same, being close to 1. In fact, after a few hours, the network becomes similar to a clique, indicating a global high mobility, allowing each individual taxi encountering most of the other taxis at some point of the day. Formally, this indicates that G_t and G_t^R are very similar for the San Francisco dataset, i.e., the probability of random encounters is much higher than the probability of social contacts, what makes the San Francisco network similar to a random mobile network. This makes sense since taxis' decisions depend on occasional customers' requests rather than on routine mobility patterns of the driver.

5. Classifier

In this section, we describe the *Random rElationship ClASsifier sTrategy (RECAST)* we propose to differentiate relationships among individuals in a social network. More precisely, the purpose of RECAST is to tell apart random interactions from social-driven ones. To that end, Section 5.1 presents the DCWN features used by RECAST. Then, Section 5.2 introduces the RECAST algorithm and, Section 5.3 discusses the results obtained by applying RECAST to the previously introduced datasets.

5.1. Social networks features

In order to identify social relationships, we must point out which features distinguish a social relationship from a random one. Indeed, two characteristics are always present in social relationships [2,32]:

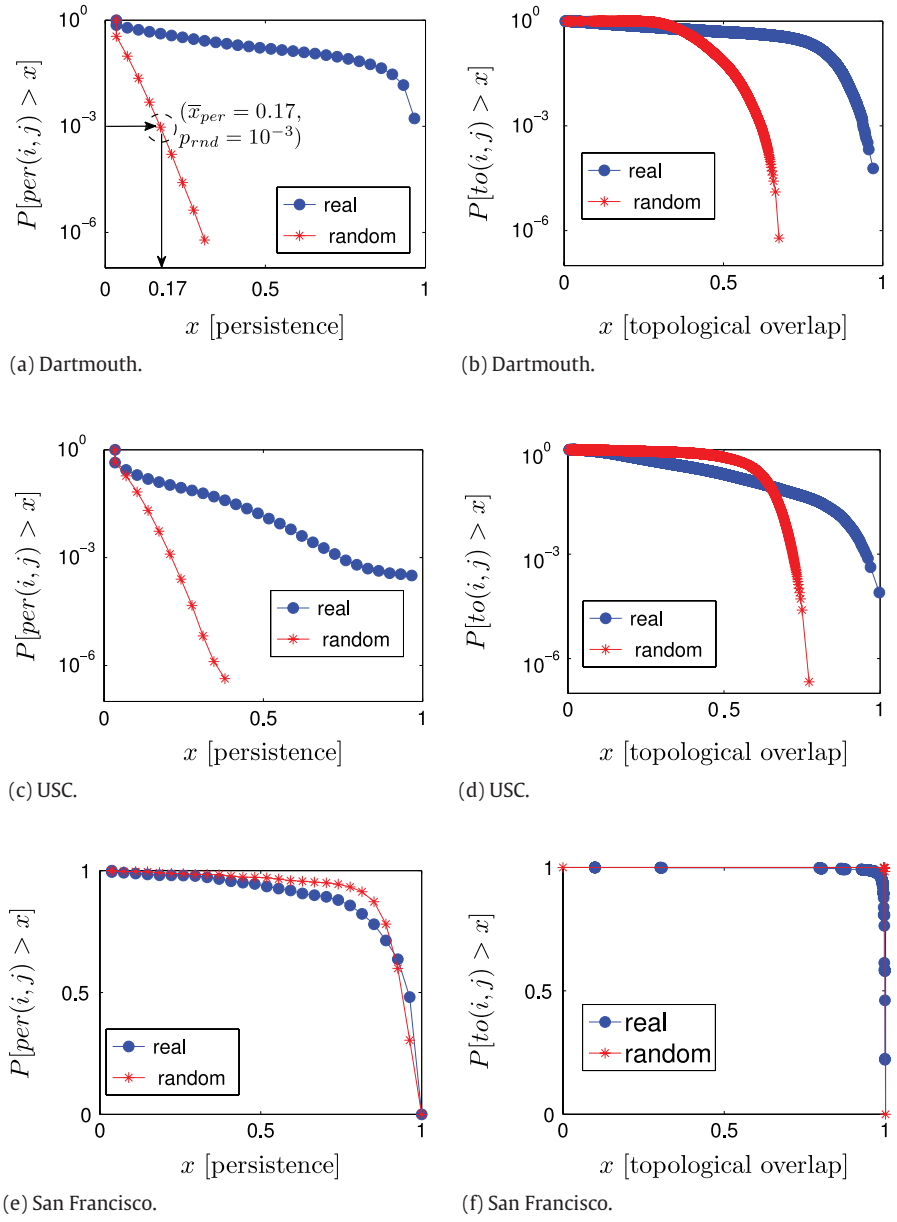
1. **Regularity**. It is well known that social relationships are regular, in that they repeat over time. If two individuals are, for example, friends, co-workers, or daily commuters, they see each other regularly.
2. **Similarity**. It is expected that two individuals who share a social relationship have common acquaintances between them. As an example, two individuals who share a large number of friends will most probably know each other as well.

Regularity and **Similarity** can be mapped into DCWN features that, in turn, can be computed from a contact dataset so as to identify what kind of relationship two individuals share. In the following, we discuss such features.

5.1.1. Edge persistence

A complex network metric mapping of the **Regularity** of a relationship is the edge persistence. Basically, considering the set of event graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$,[2] the edge persistence $per_t(i, j)$ measures the percentage of times the edge (i, j) occurred

[2] Note that edge persistence is computed over the set of graphs $\mathcal{G}_t$ and not over the temporal accumulation graph G_t.



(a) Dartmouth. (b) Dartmouth.

(c) USC. (d) USC.

(e) San Francisco. (f) San Francisco.

Fig. 4. The complementary cumulative distribution function of the edge persistence (a)(c)(e) and topological overlap (b)(d)(f) for the G_t of the three datasets and for their random correspondents G_t^R after four weeks.

over the past discrete time steps $1, 2, \ldots, t$. Formally, it is defined as $per_t(i,j) = \frac{1}{t}\sum_{k=1}^{t} \mathbb{1}_{[(i,j)\in\mathcal{E}_k]}$, where $\mathbb{1}_{[(i,j)\in\mathcal{E}_k]}$ is an indicator function that assumes value 1 if the edge (i,j) exists in $\mathcal{E}_k$ at time k, and 0 otherwise.

For instance, assuming that each day of the week is a time step, if Smith and Johnson met each other twice in a week, their edge persistence is the number of times they encountered, i.e., 2, divided by the total number of time steps, i.e., 7, or $per_{t=7}(Smith, Johnson) = 2/7$. The edge persistence allows spotting regular relationships between two entities. We again emphasize that in the aggregated graph G_t, only one edge exists between Smith and Johnson after this week.

We show in Fig. 4 (first column) the edge persistence as measured in the three datasets. Considering the set of event graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$ of all three real networks and their RND-generated random counterparts[3] $\{\mathtt{RND}(\mathcal{G}_1), \ldots, \mathtt{RND}(\mathcal{G}_t)\}$, we portray the complementary cumulative distribution function (CCDF) $\overline{F}_{per_t}(x) = P[per_t > x]$. There, the time step is one day and each curve is obtained by analyzing four weeks of contacts, since $t = 28$ corresponds to the length of the shortest considered dataset, i.e., the San Francisco one. From Fig. 4(a) and (c), we observe that the Dartmouth and USC networks have edge persistence distributions that significantly differ from those computed in their random equivalents. More precisely, while the CCDFs of random networks show an exponential decay, the individuals in the real network tend to see each other regularly, i.e., for reasons beyond pure randomness, leading to a heavy-tailed distribution. Conversely, as from Fig. 4(e), the encounters in the San Francisco dataset show an edge persistence similar to that obtained in the random equivalent graphs.

[3] We generated five instances for every random graph and the cumulative distribution considers all of them.

Table 2
RECAST relationships classes.

Class	Edge persistence	Topological overlap
Friends	Social	Social
Acquaintance	Random	Social
Bridges	Social	Random
Random	Random	Random

5.1.2. Topological overlap

The **Similarity** of contacts can be mapped to the topological overlap feature of a complex network. This metric is extracted from the aggregated temporal graph G_t. The topological overlap $to_t(i, j)$ of a pair of nodes i and j is defined as the ratio of neighbors shared by two nodes, or, formally,

$$to_t(i,j) = \frac{|\{k \mid (i,k) \in E_t\} \cap \{k \mid (j,k) \in E_t\}|}{|\{k \mid (i,k) \in E_t\} \cup \{k \mid (j,k) \in E_t\}|}.$$

In Fig. 4 (second column), we show the CCDF $\overline{F}_{to(i,j)}(x) = P[to_t(i,j) > x]$ of the topological overlap of the edges of the real networks G_t and their respective random networks G_t^R, generated by the T-RND mechanism.[4] Again, we pick one day as the time step and consider four weeks of contacts (i.e., $t = 28$ days). Similar to what occurred to the edge persistence, we note that the Dartmouth and USC network CCDFs significantly differ from their random counterparts, in Fig. 4(b) and (d). Indeed, pairs of individuals in these datasets share common neighbors in a way that could not happen randomly. Conversely, in Fig. 4(f), the San Francisco network again behaves like a random contact network. Since all results indicate that the San Francisco network is random by nature, in the remainder of this work we will focus on the Dartmouth and the USC datasets.

5.2. The RECAST algorithm

We have seen that both the edge persistence and the topological overlap behave differently in contact graphs generated from real-world social networks and in their random equivalent graphs. We exploit such a diversity to identify which edges are consequences of random or social events. In particular, we propose a classification of relationships among network entities into four categories, depending on whether the edge corresponding to the relationship features random-like persistence and topological overlap. The four classes of relationships are described in Table 2. A feature value is called "social" if there is an almost zero probability of this value being generated randomly. On the other hand, a feature value is called "random" if there is a significant probability of this value be generated randomly. In fact, as we explain in the following paragraphs, the unique parameter p_{rnd} of RECAST defines if a given feature value is social or random.

Relationships classified as *Friends* characterize pairs of individuals whose connection shows social edge persistence and topological overlap, i.e., who meet each other regularly and also tend to know the same set of people.[5] The *Acquaintance* class includes relationships among individuals sharing many common encounters, but not meeting often. As an example, friends of friends who see each other once in a while, in occasions such as birthday parties, graduation ceremonies or weddings, would be classified as *Acquaintance*. The last social class is that of *Bridges*, characterizing pairs of individuals who see each other regularly, but do not share a large number of common acquaintances. E.g., the so-called familiar strangers, people who meet every day but do not really know each other (e.g., because they just commute between common home and work areas) are very likely to be classified as *Bridges*. Finally, when an edge is neither persistent nor characterized by topological overlap, it is considered the result of a random contact, and we classify it as *Random*.

In order to distinguish "social" from "random" values of the DCWN's features, we resort to the distributions we previously discussed. More precisely, we define a value p_{rnd}, the *only* parameter in RECAST, and we identify the feature value $\bar{x}$ for which $\overline{F}(\bar{x}) = p_{rnd}$ for the random network G_t^R. The value $\bar{x}$ represents then a threshold, such that feature values higher than $\bar{x}$ occur with a probability lower than p_{rnd} in a random network. If we set p_{rnd} to some small value, we can finally state that feature values higher than $\bar{x}$ are very unlikely to occur in a random network, i.e., they are most probably due to actual social relationships. The parameter p_{rnd} can also be seen as the expected classification error percentage. For instance, a value of $p_{rnd} = 10^{-3}$ in Fig. 4(a) provides a threshold of $\bar{x} = 0.17$. In this case, all values of x higher than 0.17 can be classified as social edges in terms of edge persistence. Moreover, in this case there is a $p_{rnd} = 10^{-3}$ that RECAST misclassified random edges as social ones in terms of edge persistence. In other words, we expect that $10^{-3} = 0.1\%$ of the edges classified as social in terms of edge persistence to be, in fact, random. The full RECAST mechanism is described in Algorithm 1, where the criteria used in each classification are detailed. In this algorithm, index t is omitted for *per* and *to* metrics for clarity purposes.

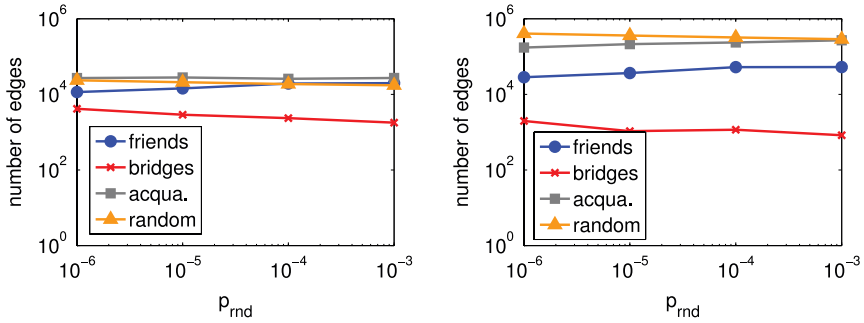
The complexity of RECAST is upper bounded by the construction of G_t^R using T-RND, which is $O(t \times (|\mathcal{V}_t| + |\mathcal{E}_t^R|))$, i.e., the minimum complexity for the generation of a degree sequence-based random graph available to date [31]. After the construction of G_t^R, the complexity of the classification mechanism is $O(|E_t^R| \times |V_t|)$, where $O(|V_t|)$ is the cost of computing the topological overlap of an edge.

[4] Again, we generated five instances for every random graph and the cumulative distribution considers all of them.

[5] It is worth mentioning that, although the *friend* terminology implies attachment among two individuals by affection or personal regard, we use it here to describe strong social ties in terms of regularity and similarity.

Algorithm 1 RECAST: classify edges of G_t

Require: $p_{rnd} \geq 0$
  return class(i, j) $\forall (i, j) \in \cup_t E_t$
  Construct G_t^R and set $\{\texttt{RND}(\mathcal{G}_1), \ldots, \texttt{RND}(\mathcal{G}_t)\}$ using T-RND
  Get $\overline{F}_{to}(x)$ and $\overline{F}_{per}(x)$ from G_t^R
  Get $\bar{x}_{to} \mid \overline{F}_{to}(\bar{x}_{to}) = p_{rnd}$ and $\bar{x}_{per} \mid \overline{F}_{per}(\bar{x}_{per}) = p_{rnd}$
  for all edges $(i, j) \in E_t$ **do**
    if $per(i, j) > \bar{x}_{per}$ and $to(i, j) > \bar{x}_{to}$ **then**
      class$(i, j) \leftarrow$ *Friends*
    else if $per(i, j) > \bar{x}_{per}$ and $to(i, j) \leq \bar{x}_{to}$ **then**
      class$(i, j) \leftarrow$ *Bridges*
    else if $per(i, j) \leq \bar{x}_{per}$ and $to(i, j) > \bar{x}_{to}$ **then**
      class$(i, j) \leftarrow$ *Acquaintance*
    else
      class$(i, j) \leftarrow$ *Random*
    end if
  end for



(a) Dartmouth. (b) USC.

Fig. 5. The number of edges of a given class that appears in the first four weeks of data versus p_{rnd}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

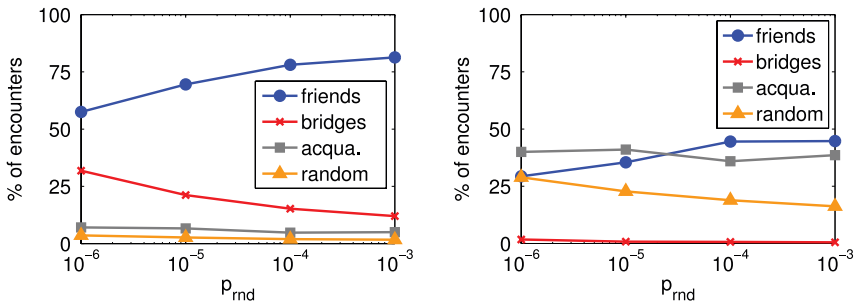
5.3. Classification results

We apply RECAST to the Dartmouth and the USC networks. We are omitting the results for the San Francisco dataset, since, as previously stated, the random-like nature of taxi routes makes the analysis uninteresting, with all edges classified as *Random*. In Fig. 5, we show the number of edges per class as a function of the p_{rnd} value. An initial and quite surprising observation is that, by varying p_{rnd} through four orders of magnitude, the number of edges per class stays in the same magnitude. This shows that RECAST is robust with respect to p_{rnd}, i.e., it does not need a fine calibration of the parameter to return a consistent edge classification.

Secondly, in both datasets, the number of *Bridges* is orders of magnitude lower than for the other classes, a clear indication that in the analyzed social networks regular connections among different communities are rare. Also the number of *Friends* edges is similar in the two networks, implying similar dynamics in tight relationships among individuals in the two campuses. This also agrees with the biological constraint on a social interaction that limits human social networks' size, i.e., the number of *Friends* relationships [33]. However, the two datasets differ when looking at the number of edges classified as *Acquaintances* and *Random*, that are one order of magnitude larger in USC than in Dartmouth. This is the result of the actual size of the two campuses, USC accounting for a population around ten times larger than that of Dartmouth. This aspect is also reflected by the size of the traces in Table 1 that clearly leads, in the USC network, to (i) many more *Random* contacts among individuals who do not actually know each other, but just happen to cross each other while strolling on campus, and (ii) an increased presence of strangers who happen to know the same people, leading to more *Acquaintances* edges.

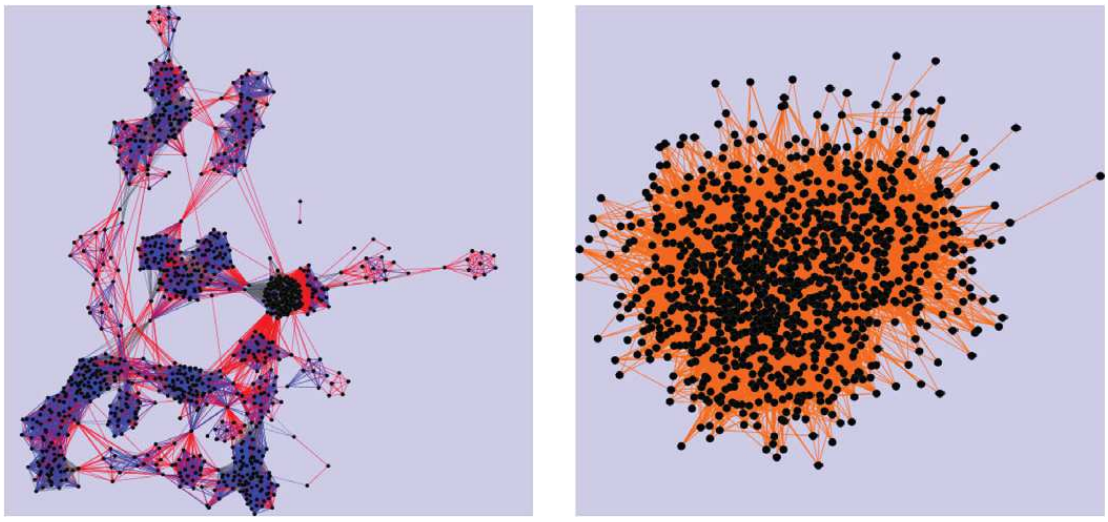
The observations above are even more evident when observing the percentage of individual encounters of each type in the Dartmouth and USC networks. In Fig. 6, we show the percentage of encounters of a given class that appear in the first four weeks of data for a given value of p_{rnd}. The percentage of *Random* encounters in the Dartmouth network is close to zero, varying from 1.7% to 3.6% as p_{rnd} decreases. On the other hand, in the USC network, this percentage varies from 16% to 29%. In fact, the proportion of *Random* encounters provides a good estimate of the probability of random decisions mentioned in Section 4. Thus, the USC network has a significantly higher tendency to evolve to a random topology than the Dartmouth network.

The analysis is confirmed by Fig. 7, portraying the snapshots of the Dartmouth and USC networks after two weeks of interactions, when considering only social edges (i.e., those classified as *Friends*, *Acquaintances* and *Bridges*) or only edges tagged as *Random*. Edges of the former networks in Fig. 7(a) and (c) are distinguished by colors, according to the same code used in Figs. 5 and 6 (*Friends* edges are in blue, *Bridges* in red, *Acquaintances* in gray, and *Random* in orange). The difference between the social-only networks and the random-only ones is striking. Social networks are characterized by a complex

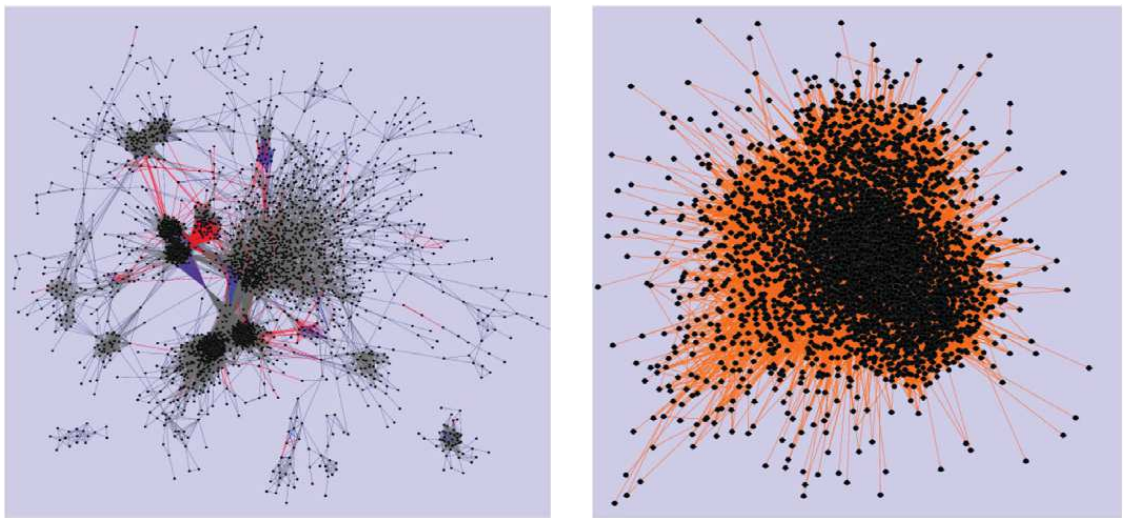


(a) Dartmouth. (b) USC.

Fig. 6. The percentage of encounters of a given class that appears in the first four weeks of data versus p_{rnd}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Dartmouth, social edges. (b) Dartmouth, random edges.

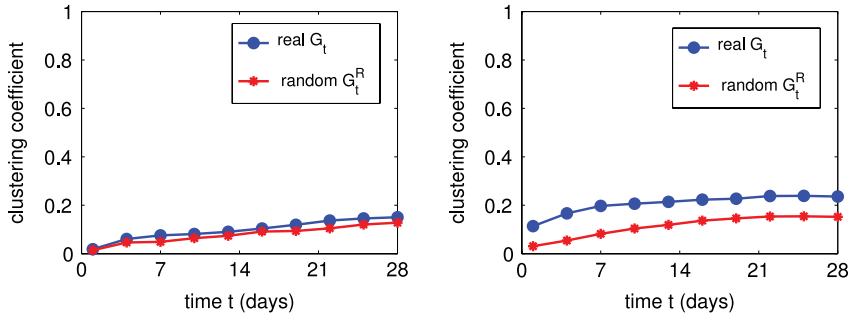


(c) USC, social edges. (d) USC, random edges.

Fig. 7. Snapshots of the Dartmouth and USC networks after two weeks of interactions, considering only the social edges and only the random edges. *Friends* edges are painted in blue, *Bridges* in red, *Acquaintances* in gray and *Random* in orange. This figure is best viewed in colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

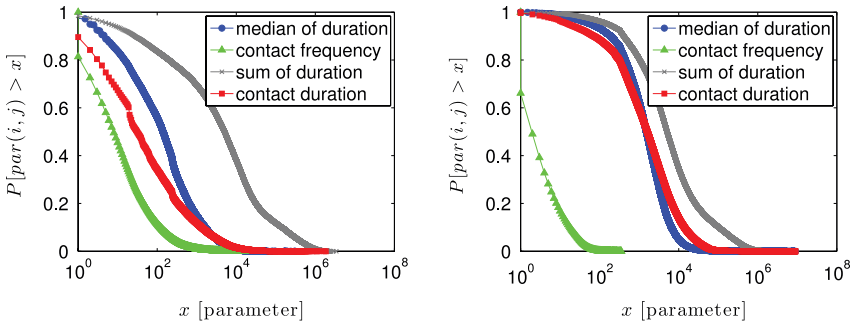
structure of *Friends* communities, linked to each other by *Bridges* and *Acquaintances*. More precisely, when comparing the Dartmouth and USC social networks, the former appears to be dominated by *Friends* interactions, while the sheer number of *Acquaintances* in the latter drives its graph structure.

Conversely, networks containing only *Random* edges do not show any structure and look like random graphs. A rigorous way to verify the randomness of such networks, and thus validate the efficiency of the RECAST classification, is to perform a



(a) Dartmouth. (b) USC.

Fig. 8. Evolution of the clustering coefficient of G_t when only *Random* edges are present, compared to their random correspondents G_t^R.



(a) Dartmouth, in 28 days. (b) USC, in 28 days.

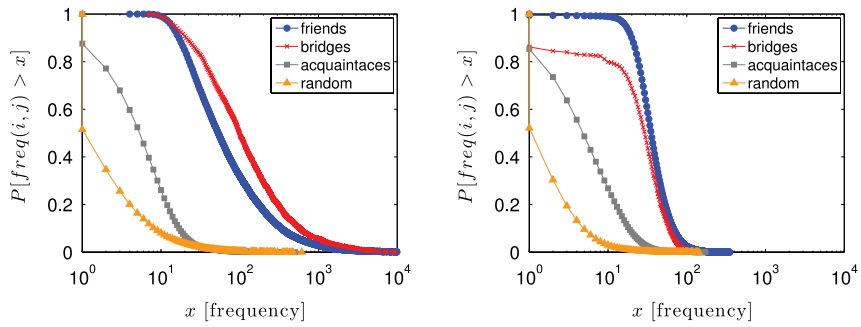
Fig. 9. The complementary cumulative distribution function of the: contact frequency, contact duration (in sec), sum and median of contact (in sec) of (a) Dartmouth and (b) USC networks after four weeks of data.

clustering coefficient analysis. Fig. 8 compares the clustering coefficients of the Dartmouth and USC networks G_t *when only Random edges are present* against the same metric computed in their random counterparts G_t^R. The clustering coefficient, commonly employed to determine the actual randomness of a network, has very similar evolutions in G_t and G_t^R: this proves that the network of contacts tagged as *Random* by RECAST is actually a random network. Therefore, RECAST is able to extract from a real-world contact dataset edges that correspond to random encounters.

The observations above are also confirmed when analysis contact duration and contact rate of the Dartmouth and USC networks The duration and frequency of contacts are two paramount factors that characterize the actual daily interaction among users. They are also critical for message transmission, routing decisions, or capacity planning in dynamic networks [20,34]. While the duration of contacts has a direct impact on the amount of data that can be transferred during one encounter, frequency of contacts may provide an idea of the number of possible retries a transmission may be given.

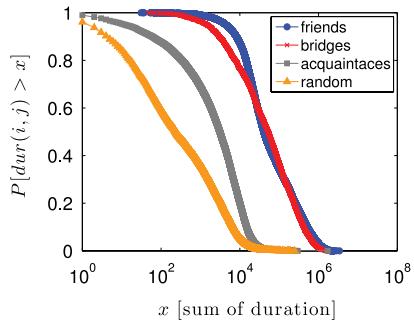
In Fig. 9, we provide a general overview, i.e., the CCDF of contact frequency, contact duration, sum and median of contact duration in the Dartmouth and USC datasets, for four weeks of data. Although these are both campus environments, we can clearly see differences in the way people are interacting: the distribution shape appears to be driven by the campus characteristics. USC contacts last longer and happen less frequently due to the higher stationary nature of devices (as shown in Table 1, the average number of encounters/node/day is 23.8 in USC against 145.6 in Dartmouth).

We expect the frequency and duration of contacts to be affected by social relations. For instance, *Friends* meet regularly and tend to spend more time together. In Fig. 10, we investigate whether such an intuition is validated, by separately analyzing the contact frequency and duration for each class of ties identified by RECAST in the two datasets. An initial observation is that, in both traces, users having edges classified as social spent more time together and have higher contact rate. In particular, *Friends* and *Bridges* totalize the longest time period spent together and the highest contact rate in the two networks besides exhibiting quite homogeneous behavior: The contact rate is higher in the Dartmouth than in USC network, probably due to its highest percentage of social edges (see Fig. 6). When considering edges classified as *Bridges*, this result is surprising, considering the lower number of such edges in the two networks (see Fig. 5). Finally, both networks present similar distributions of sum of contact duration and contact rates when it comes to edges classified as *Acquaintance* and *Random*: with nodes spending less time together or meeting less than in the other classes. Their distribution shape also appears to be driven by the population size of the networks: The tail of the contact rate and sum of contact duration distributions are longer in Dartmouth, while they decay faster in USC network, probably due to the fact that the USC population is around ten times larger than that of Dartmouth. The USC network features a higher percentage of *Random* encounters (see Fig. 6) and, consequently, has a higher tendency to evolve into a random topology. However, it is interesting to observe how the duration and frequency of contacts in this network are still dominated by social encounters (see Fig. 10(d)).

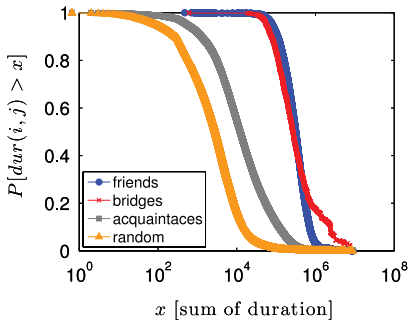


(a) Dartmouth, the frequency of contact.

(b) USC, the frequency of contact.



(c) Dartmouth, the sum of duration.



(d) USC, the sum of duration.

Fig. 10. The contact rate and the sum of contact duration (in s) of a given class after four weeks of data.

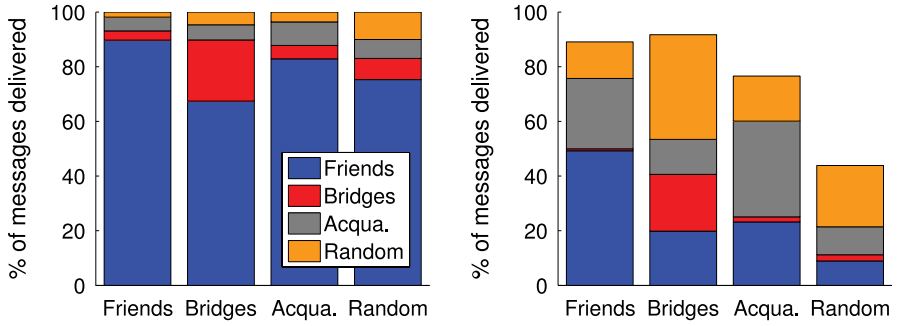
6. RECAST application

In this section, we use the Dartmouth and USC contact traces to simulate an epidemic dissemination. We consider that the users communicate with each other in an opportunistic fashion, i.e., without any infrastructure and exchanging messages only when they are within physical proximity. Thus, if user i wants to send a message to user j, he (she) has to deliver it personally to j or has to ask other users to relay it for him (her), through a multi-hop carry-and-forward path. Also, we consider the transfer to be epidemic, i.e., in order to reach j, user i sends a message to all other users he (she) is in contact with at a given time t. The latter forward it to all of the users they later meet and so on, until user j is reached. Such an epidemic approach allows us characterizing the lower bound on the delay required by the opportunistic transfer. As discussed hereafter, the two used traces present significant differences. In particular, more contacts exist in the Dartmouth trace, which explains the high delivery ratio of the evaluated opportunistic transfer. On the other hand, USC trace presents temporal disconnections, resulting in undelivered messages.

For both Dartmouth and USC contact traces, we use RECAST to classify the relationships between users over one month of contact data, which we refer to as the *classifying stage*. Then, we simulate the opportunistic transfer scenario above during the two following weeks, termed the *routing stage*, containing only future encounters, not known previously by RECAST. For each user i, we randomly pick a time $t_{0,i}$ within the first week of the *routing stage* for him/her to start the epidemic transfer process, and a destination user j. We leave one week for the message to reach its destination: if the message is not delivered by then, the transfer is deemed failed and the data lost. In this way, the following results consider all the routing attempts between every pair of users that have a class of relationship given by RECAST in the *classifying stage*. When a message is successfully delivered, we consider the path that delivered the message first. Moreover, to make the comparison fair, we remove from the traces all the users who did not appear in the *routing stage*, most of them sharing a *Random* relationship with their contacts. Our goal is to study how the different classes of relationships between the source i and the destination j affect the epidemic transfer of the message itself. We are also interested in understanding the nature of the contacts used by successfully delivered messages.

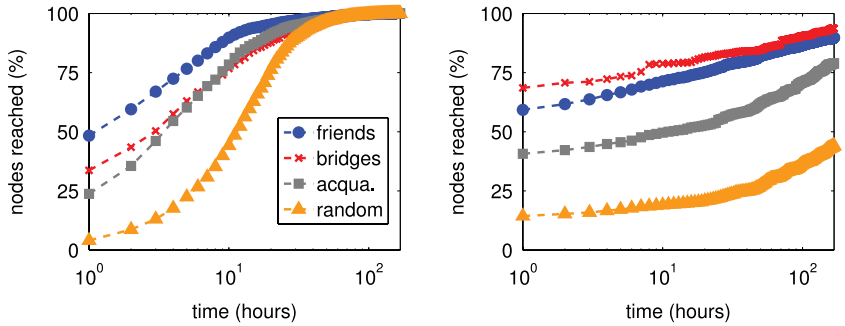
In Fig. 11, we show the overall forwarding efficiency. Fig. 11(a) and (b) show the percentage of messages that were successfully delivered to their destinations in the Dartmouth and in the USC scenarios, respectively. Each bar represents one relationship shared by the message source and destination, and within each bar we depict the fraction of edges of a given class of relationship that was used to deliver the message. First, we observe that all the messages reached their destinations in the Dartmouth scenario, a consequence of the limited network size. Moreover, the edges classified as *Friends* were the most used to deliver the messages and the *Random* ones the least used. In fact, considering all the paths directed to *Friends*, less than 2% of the hops in these paths were given by users who share a *Random* relationship.

Different from the Dartmouth scenario, not all messages are delivered in the USC scenario. The lower delivery ratio is partially due to the larger network that is harder to navigate. Moreover, we removed from the *routing stage* all encounters



(a) Dartmouth. (b) USC.

Fig. 11. The forwarding efficiency when user *i* sends a message to user *j* in the opportunistic network, and *i* and *j* share a specific RECAST relationship. Within each bar we also show the fraction of edges of a given class of relationship that was used to deliver the message.



(a) Dartmouth. (b) USC.

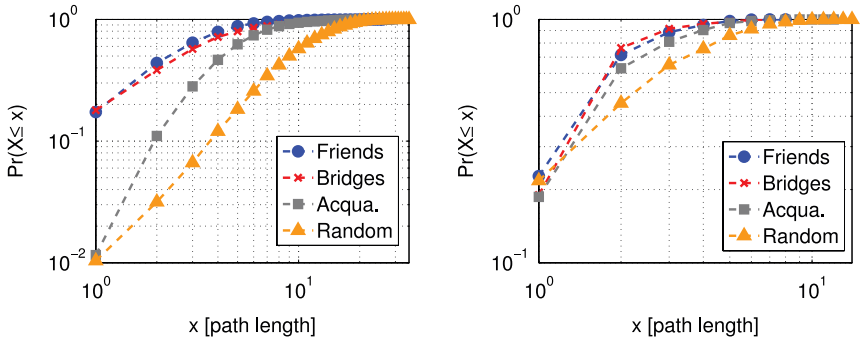
Fig. 12. The % of users who were reached over the time.

between users who do not share a RECAST class, or approximately 25% of all encounters. However, those are not the only reasons, and the social relationship between the source and destination significantly affects the probability of success. Socially connected pairs (tagged as *Friends*, *Bridges* and *Acquaintances*) can actually exchange data: 90% of the messages were successfully delivered to *Friends*, 92% to *Bridges* and 77% *Acquaintances*. If one wants to send a message to a *Random* contact in the USC scenario, there is only 44% of chances that this message will arrive successfully. Moreover, although the majority of the classified edges in the USC scenario are *Random* (see Fig. 5), the majority of the hops in the paths are between users who share a social relationship. Considering all the paths directed to *Friends*, less than 13% of the hops in these paths were given by users who share a *Random* relationship.

In Fig. 12, we show how much time it was necessary for the messages to reach their destinations. We grouped together all the routings from source user *i* to destination user *j* by the class of relationship $c \in \{Friend, Bridge, Acquaintance, Random\}$ that *i* and *j* share. Then, we cumulatively count how many destinations of the class *c* are reached at each hour, considering the total number of routings that were performed between sources and destinations of class *c*. Observe that the expected time to reach a *Random* contact is significantly higher than the time needed to reach a social contact. Moreover, observe that the majority of the messages sent to *Friends* arrive in the first hours for both scenarios. These results may serve to leverage the performance of various routing solutions for opportunistic networks. If we previously know the class of relationship the destination share with the source, we also know the chances and the probable time the message will take to arrive.

It is not only the time it takes for a message to arrive at its destination that is relevant to the design of forwarding solutions for opportunistic networks. Another fundamental aspect is the number of hops required to reach the destination. In Fig. 13, we show the cumulative distribution function (CDF) of the path lengths of messages between users *i* and *j* who share a determined class of relationship. Observe that the expected number of hops for a message to arrive at a *Random* contact is significantly higher than to arrive at a social contact. For the USC scenario, 89%, 92% and 81% of the routes to *Friend*, *Bridge* and *Acquaintance* destinations, respectively, have path lengths lower or equal to 3. In the meanwhile, only 65% of the routes to *Random* destinations have path lengths lower or equal to 3. This difference is even more striking for the Dartmouth scenario, where 65%, 57% and 28% of the routes to *Friend*, *Bridge* and *Acquaintance* destinations, respectively, have path lengths lower or equal to 3, and only 6% to *Random* destinations have path lengths lower o equal to 3.

Overall, our results show how the RECAST classification allows identifying those who share social relationships with the sender, whose opportunistic paths are usually short and reliable. In fact, such paths usually pass through a few number of hops, mostly using social ties among users, and rarely leverage random encounters. As an intuitive corollary, reaching users that share some social relationship is significantly easier than attaining users one does not know, especially in large systems. These results may serve to leverage the performance of various routing solutions for opportunistic networks. If we



(a) Dartmouth. (b) USC.

Fig. 13. The histogram of the path lengths of messages between users *i* and *j* who share a determined class of relationship.

Table 3
RECAST classes vs. Facebook friendships in the Sassy and UPB datasets.

	Friends	*Bridges*	*Acquaintances*	*Random*	Total
Sassy					
Friends on facebook	3	0	19	43	65
Non-friends on facebook	2	1	15	35	53
Total	5	1	34	78	118
UPB					
Friends on facebook	0	0	16	17	33
Non-friends on facebook	0	0	26	12	38
Total	0	0	42	29	71

previously know the class of relationship the destination share with the source, we also know the chances and the probable time the message will take to arrive.

7. RECAST vs. online friendship

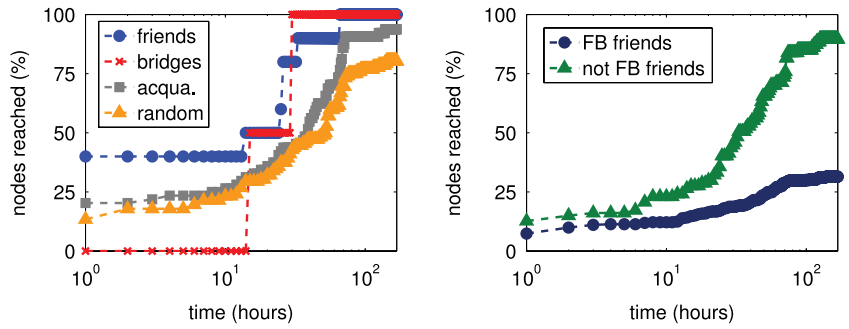
We are living in a time where online social networks (OSNs) such as Facebook and Google+ are present in the lives and routines of billions of people. In such systems people are able to maintain a personal list of other users whose they share a social relationship with, such as their friends and colleagues. In the context of this paper, questions that naturally arise are: what are the differences between the relationships labeled as *social* by RECAST and by OSNs? Considering the same set of people, their self-reported social relationships on OSNs have a significant intersection with their social relationships given by RECAST? If not, can we also use the social relationship given by OSNs to route messages in opportunistic scenarios?

In order to answer these questions, we refer to the two remaining real-world mobility datasets introduced in Section 3. In these two datasets, we have information about the physical encounters among human users, as the Dartmouth and USC datasets, and also data about the self-reported social networks of these users. In both cases, the self-reported social network was collected directly from the Facebook pages of the users. This allows us to construct the social network of these users, that tells which pairs of users declared a friendship relationship between them on Facebook.

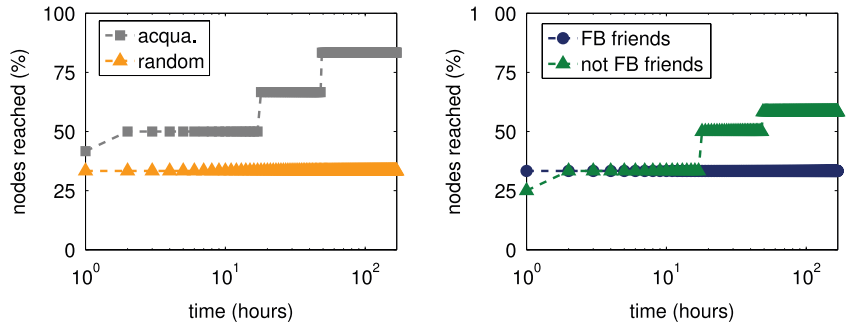
The first dataset, that from now on we refer as the *Sassy* dataset, contains the information about 27 human users associated with University of St Andrews comprising 22 undergraduate students, 3 postgraduate students, and 2 members of staff. Each user carried a mobile IEEE 802.15.4 sensors (T-mote invent devices) that are able to detect each other within a radius of 10 m. At each detection, an encounter is characterized and stored in the dataset. Participants were asked to carry the devices whenever possible over a period of 79 days starting on February, 15, 2008. For more details see [28].

The second dataset, that from now on we refer as the *UPB* dataset, was also collected in an academic environment, at University Politehnica of Bucharest, from the mobility of 22 participants. The participants were twelve Bachelor students (one in the first year, nine in the third and two in the fourth), seven Master students (four in the first year and three in the second) and three research assistants. The data was collected only inside the grounds of the faculty between 8 AM and 8 PM during week-days. Data was collected from November 11th, 2011 to December 22nd, 2011, using an Android application that registers contacts between mobile devices with Bluetooth. For more details see [29].

In order to verify if RECAST classes are able to indicate the presence of friendship in online social networks, we classified the relationships in both datasets using RECAST after 3 weeks of encounters. After that, we verified which of the classified relationships are and are not labeled as *friendship* on Facebook. Table 3 shows the number of friends (and non-friends) on Facebook (lines) who were classified in a given RECAST class (column) for the Sassy and the UPB datasets. Observe that there is not a RECAST class that clearly yields more (or less) friendships on Facebook. This strongly suggests that RECAST classes are *not* related with friendship in online social networks. This is not a surprising outcome: while RECAST aims at spotting



(a) Sassy: RECAST. (b) Sassy: Facebook.



(c) UPB: RECAST. (d) UPB: Facebook.

Fig. 14. The % of users who were reached over time grouped by RECAST classes and Facebook (FB) friendship.

social interactions induced by physical mobility, online social networks register social relationships that do not necessarily map to physical encounters. For instance, it is common to see Facebook friendships between people from different countries, ex-colleagues who still want to keep in touch, or even classmates who do not go along very well.

Since RECAST classes are not related with Facebook friendships, one could ask which classes are more valuable for opportunistic routing scenarios, i.e., is it better to know my RECAST relationships or my Facebook friends when I need to deliver a message in a opportunistic network? To address this question, we simulate an epidemic dissemination in the same way we described in Section 6. After classifying the relationships using three weeks of data, we simulate the opportunistic transfer scenario during the following week (*routing stage*), containing only future encounters. Again, for each user i, we randomly pick a time $t_{0,i}$ within the first day of the *routing stage* for him/her to start the epidemic transfer process, and a destination user j. If the message is not delivered within one week, the transfer is deemed failed and the data is lost. The following results consider all the routing attempts between every pair of users that have a class of relationship given by RECAST in the *classifying stage*. Moreover, together with the RECAST class, we also group the results by the friendship relationship on Facebook, i.e., we verified if Facebook friends are more likely to be reached than non Facebook friends.

In Fig. 14 we show how much time it was necessary for the messages to reach their destinations for the Sassy and UPB datasets. We grouped together all the routings from source user i to destination user j by the class of RECAST relationship $c_1 \in \{Friend, Bridge, Acquaintance, Random\}$ that i and j share (Fig. 14(a) and (c)) and also by their Facebook friendship status $c_2 \in$ {FB friends, non-FB friends} (Fig. 14(b) and (d)). Then, as done previously, we cumulatively count how many destinations of the class c_i are reached per each hour, considering the total number of routings that were performed between sources and destinations of class c_i. Again, observe that among the RECAST classes the expected time to reach a *Random* contact is significantly higher than the time needed to reach a social contact. Moreover, contrary to our intuition, observe that, for both datasets, non Facebook friends are more likely to be reached than Facebook friends. We argue that this is a clear proof of the significant abstraction of online social relationships from physical encounters driven by social mobility. Therefore, we conclude that Facebook relationships are of small use for decision making in opportunistic data transfers. Conversely, RECAST classes are useful to opportunistic routing. As a matter of fact, as already observed in the Dartmouth and USC datasets, and also in the Sassy and UPB scenarios, RECAST social classes are reached faster and more likely than both Facebook classes.

8. Conclusions

The contribution of this paper is threefold. First, we modeled five real-world mobile user encounter datasets as temporal contact graphs and we proposed the use of random equivalent graphs to outline their hidden social structure. Our original approach shows that different mobility traces can yield completely different social structures, determined by diverse

behaviors of the entities participating in the system. For instance, we showed a dataset describing the movement of cab drivers in San Francisco, CA, USA, has mostly random properties, making the relative contact network similar to a random network. Conversely, encounters in university campuses show strong social features: however, different campuses features diverse types of social ties. These results let us speculate that researchers should not generalize their results based on the analysis of a single dataset. Second, to perform this issue, we proposed and implemented the RECAST strategy that lazily classifies random and social relationships in dynamic social networks, and demonstrates its simplicity and effectiveness. Third, we employed the RECAST classification to the case of epidemic opportunistic transfers, and showed its relevance towards the identification of faster, reliable paths leveraging social ties among users. Moreover, RECAST benefits for opportunistic forwarding was highlighted when compared to the use of friendship status in online social networks.

Acknowledgments

The second, third and fourth authors were supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO. The first and sixth authors were partially supported by the authors individual grants from CNPq and FAPEMIG.

References

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan, Group formation in large social networks: membership, growth, and evolution, in: KDD'06: Proceedings of the 12th ACM SIGKDD, ACM, New York, NY, USA, 2006, pp. 44–54. http://doi.acm.org/10.1145/1150402.1150412.
[2] C.A. Hidalgo, C. Rodriguez-Sickert, The dynamics of a mobile phone network, Physica A 387 (12) (2008) 3017–3024.
[3] W. Gao, Q. Li, B. Zhao, G. Cao, Multicasting in delay tolerant networks: a social network perspective, in: ACM MobiHoc, 2009.
[4] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks, in: KDD'06: Proceedings of the 12th ACM SIGKDD, ACM, New York, NY, USA, 2006, pp. 611–617. http://doi.acm.org/10.1145/1150402.1150476.
[5] E.M. Daly, M. Haahr, Social network analysis for routing in disconnected delay-tolerant manets, in: Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc'07, ACM, New York, NY, USA, 2007, pp. 32–40.
[6] T. Hossmann, F. Legendre, T. Spyropoulos, From contacts to graphs: pitfalls in using complex network analysis for DTN routing, in: Proceedings of the 28th IEEE International Conference on Computer Communications Workshops, INFOCOM'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 260–265.
[7] D. Katsaros, N. Dimokas, L. Tassiulas, Social network analysis concepts in the design of wireless ad hoc network protocols, Netwrk. Mag. Global Internetwkg. 24 (2010) 23–29.
[8] M. Buchanan, Behavioural science: secret signals, Nature 457 (7229) (2009) 528–530.
[9] D.D. Price, A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inf. Sci. 27 (5) (1976) 292–306. http://dx.doi.org/10.1002/asi.4630270505.
[10] A. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509.
[11] A. Vázquez, Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations, Phys. Rev. E 67 (5) (2003) 056104+.
[12] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins, Microscopic evolution of social networks, in: KDD'08: Proceeding of the 14th ACM SIGKDD, ACM, New York, NY, USA, 2008, pp. 462–470. http://doi.acm.org/10.1145/1401890.1401948.
[13] R. Albert, H. Jeong, A.-L. Barabási, Diameter of the world wide web, Nature 401 (1999) 130–131.
[14] P. Erdös, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 7 (1960) 17.
[15] M. Piorkowski, N. Sarafijanovic-Djukic, M. Grossglauser, A parsimonious model of mobile partitioned networks with clustering, in: COMSNETS, 2009.
[16] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD, KDD'11, ACM, New York, NY, USA, 2011, pp. 1082–1090.
[17] A. Mtibaa, M. May, C. Diot, M. Ammar, PeopleRank: social opportunistic forwarding, in: IEEE INFOCOM, 2010.
[18] M. Conti, R. Di Pietro, A. Gabrielli, L.V. Mancini, A. Mei, The smallville effect: social ties make mobile networks more secure against the node capture attack, in: ACM MobiWac, 2010.
[19] A.G. Miklas, K.K. Gollu, K.K.W. Chan, S. Saroiu, K.P. Gummadi, E. De Lara, Exploiting social interactions in mobile systems, in: Proceedings of the UbiComp'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 409–428.
[20] G. Zyba, G.M. Voelker, S. Ioannidis, C. Diot, Dissemination in opportunistic mobile ad-hoc networks: the power of the crowd, in: Proceedings of IEEE INFOCOM 2011, IEEE, 2011, pp. 1179–1187.
[21] S. Milgram, The familiar stranger: an aspect of urban anonymity, in: The Individual in a Social World, Addison-Wesley, 1977, pp. 51–53. (Chapter).
[22] C.C. Aggarwal, Social Network Data Analytics, first ed., Springer Publishing Company, Incorporated, 2011.
[23] D.J. Watts, S.H. Strogatz, Collective dynamics of "small-world" networks, Nature 393 (1998) 440–442.
[24] T. Henderson, D. Kotz, I. Abyzov, The changing usage of a mature campus-wide wireless network, in: Proceedings of the 10th Annual International Conference on Mobile Computing and Networking, MobiCom'04, ACM, New York, NY, USA, 2004, pp. 187–201.
[25] W.jen Hsu, A. Helmy, IMPACT: investigation of mobile-user patterns across university campuses using wlan trace analysis, CoRR arxiv.org/pdf/cs/0508009.
[26] A. Rojas, P. Branch, G. Armitage, Experimental validation of the random waypoint mobility model through a real world mobility trace for large geographical areas, in: Proceedings of the 8th ACM MSWiM, MSWiM '05, ACM, New York, NY, USA, 2005, pp. 174–177.
[27] F. Bai, D. Stancil, H. Krishnan, Toward understanding characteristics of dedicated short range communications (DSRC) from a perspective of vehicular network engineers, in: ACM MobiCom, 2010.
[28] G. Bigwood, D. Rehunathan, M. Bateman, T. Henderson, S.N. Bhatti, Exploiting self-reported social networks for routing in ubiquitous computing environments, in: WiMob, IEEE, 2008, pp. 484–489.
[29] R. Ciobanu, C. Dobre, V. Cristea, Social aspects to support opportunistic networks in an academic environment, in: X.-Y. Li, S. Papavassiliou, S. Ruehrup (Eds.), Ad-hoc, Mobile, and Wireless Networks, in: Lecture Notes in Computer Science, vol. 7363, Springer, Berlin, Heidelberg, 2012, pp. 69–82.
[30] T.M.J. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Softw. - Pract. Exp. 21 (11) (1991) 1129–1164.
[31] F. Chung, L. Lu, Connected components in random graphs with given expected degree sequences, Ann. Comb. 6 (2) (2002) 125–145.
[32] J.P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.L. Barabási, Structure and tie strengths in mobile communication networks, Proc. Natl. Acad. Sci. 104 (18) (2007) 7332–7336.
[33] R.I.M. Dunbar, The social brain hypothesis, Evol. Anthropol. 6 (5) (1998) 178–190.
[34] W. Moreira, P. Mendes, S. Sargento, Opportunistic routing based on daily routines, in: 2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), IEEE, 2012, pp. 1–6.

Pedro O.S. Vaz de Melo is an assistant professor in the Computer Science Department (DCC) of Federal University of Minas Gerais (UFMG). He has degree (2003) and Masters (2007) in Computer Science from the Pontifical Catholic University of Minas Gerais (2003). He got his Ph.D. at Federal University of Minas Gerais (UFMG) with a one year period as a visiting researcher in Carnegie Mellon University and a five months period as a visiting researcher at INRIA Lyon. His research interest is mostly focused on knowledge discovery and data mining in complex and distributed systems.

Aline Carneiro Viana is a CR1 at INFINE research team of INRIA Saclay - Ile de France. She received her habilitation from Université Pierre et Marie Curie, Paris, France in 2011. From November 2009 to October 2010, Dr. Viana was in a sabbatical leave at the Telecommunication Networks Group (TKN) of the Technischen Universität Berlin (TU-Berlin), Germany. Dr. Viana got her Ph.D. in Computer Science from the University Pierre et Marie Curie -Paris VI in 2005. After having hold a postdoctoral position at IRISA/INRIA Rennes - Bretagne Atlantique in the PARIS research team, she obtained a permanent position at INRIA Saclay - Ile de France, in 2006. Dr. Viana's research addresses the design of solutions for self-organizing and dynamic networks with the focus on opportunistic networking, data offloading techniques, mobile social networking, and smart cities. She has published more than 70 research papers and is the scientific coordinator of two international research projects (EU CHIST-ERA and STIC AmSud). She has chaired several IEEE/ACM workshops, participated of the organizing committee of numerous conferences, and served on the technical program committee of several international conferences and workshops including IEEE SECON, ACM CoNEXT, IEEE PIMRC, IEEE MASS, and IEEE LCN. Aline has also served for three consecutive years as reviewer for the European Commission and is Associate Editor of ACM Computer Communication Review (ACM CCR).

Marco Fiore (S'05, M'09) holds Ph.D. and HDR degrees from Politecnico di Torino, and Université de Lyon, respectively. He is a researcher at CNR-IEIIT, Italy, and an associate researcher at INRIA UrbaNet, France. Previously, he has been a visiting researcher at Rice University and Universitat Politecnica de Catalunya, as well as an Associate Professor at INSA Lyon. His main research interests are on mobile data analysis and vehicular networking.

Katia Jaffrès-Runser received both a Dipl. Ing. (M.Sc.) in Telecommunications and a DEA (M.Sc) in Medical Imaging in 2002 and a Ph.D. in Computer Science in 2005 from the National Institute of Applied Sciences (INSA), Lyon, France. From 2002 to 2005 she was with Inria, participating in the ARES project while working towards her Ph.D. thesis. In 2006, she joined the Stevens Institute of Technology, Hoboken, NJ, USA, as a post-doctoral researcher. She is the recipient of a three-year Marie-Curie OIF fellowship from the European Union to pursue her work from 2007 to 2010 in both Stevens Institute of Technology and INSA Lyon on wireless networks modeling and multiobjective optimization. In 2011, she participated in the GreenTouch consortium as a delegate from INRIA. Since September 2011, she joined the University of Toulouse - ENSEEIHT as a Maître de Conférences (Associate Professor), working at IRIT laboratory on hybrid embedded networks optimization.

Frédéric Le Mouël is currently associate professor in INSA Lyon - a leading engineering school in France, part of the university of Lyon. He conducts his research in the INRIA CITI laboratory where he is leading the Dynamid team - Dynamic software and distributed systems for the internet of things. Frdric Le Mouël holds a master and a Ph.D. degree in computer science from the university of Rennes 1, France. His main interests are distributed systems, operating systems, middleware, virtual machines, programming languages and more specifically in dynamic adapting, self-coordinating and autonomic environments. Application fields concerned are ambient intelligence, internet of things, home automation, vehicular networks and service robotics.

Antonio A.F. Loureiro received his B.Sc. and M.Sc. degrees in computer science from the Federal University of Minas Gerais (UFMG), Brazil, and the Ph.D. degree in computer science from the University of British Columbia, Canada. Currently, he is a Full Professor of computer science at UFMG, where he leads the research group in wireless sensor networks and ubiquitous computing. His main research areas are wireless sensor networks, urban sensing, ubiquitous computing, and distributed algorithms. In the last 15 years he has published extensively in international conferences and journals related to those areas, and also presented tutorials and keynote talks at international conferences.

Lavanya Addepalli is a Ph.D. candidate registered since 2012 in Polytechnic University of Valencia, Spain under supervision of Dr. Jaime Lloret Mauri. Her research is about Architecture design and deployment of an online social network that uses data classification to provide more trustable relations. She is currently working on mobile data analysis for Kubhmela 2015 for The State Government of Maharashtra, India, in association with Kumbhathon Team headed by Dr. Ramesh Rasker from MIT Media Labs Boston, USA. She also spent six months in 2014 as an intern with Inria Saclay - Ile de France, under supervision of Dr. Aline Carneiro Viana, Understanding the worth of contact duration in wireless social networking. She received her Master's degree in Medical Software in 2011 from Manipal University; Bachelor's degree in Information Technology in 2009 under Jawaharlal Nehru Technological University, and Diploma in Information Technology in 2006 from Mumbai Technical Board, India. Lavanya has done her Master's thesis and project in Charles University Prague, Czech Republic in 2010–2011 under department of Mathematics and Physics, developed a software tool that allows generation of datasets for Inductive logic programming. She is having over 20 publications in various fields of research and also serves as reviewer for couple of journals.

Chen Guangshuo received the B.E and M.S. degree from Shanghai Jiao Tong University, China, in 2011 and 2014, respectively. Now he is a Ph.D. degree candidate of Ecole Polytechnique, France, and located in INRIA Saclay - Ile de France. He has published 3 papers about wireless sensor networks in IEEE WCNC and WiMob. At present, he is working on understanding the correlations between user mobility and content demand, under the supervision of Aline Carneiro Viana.