Ad Hoc Networks 19 (2014) 92–110

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

Data offloading in social mobile networks through VIP delegation

Marco Valerio Barbera [a], Aline Carneiro Viana [b], Marcelo Dias de Amorim [c], Julinda Stefa [a,*]

[a] Sapienza University of Rome, Via Salaria 113, 00198 Rome, Italy
[b] INRIA, Route de Saclay, 91128 Palaiseau Cedex Inria Saclay, Ile de France, France
[c] CNRS/LIP6, UPMC Sorbonne Universités, LIP6 Boîte courrier 169 Couloir 26-00, Étage 1, Bureau 109 4 Place Jussieu, 75252 Paris Cedex 05, France

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form 9 November 2013
Accepted 30 January 2014
Available online 5 March 2014

Keywords:
Centrality-based metrics
Coverage strategy
Delay tolerant networks

ABSTRACT

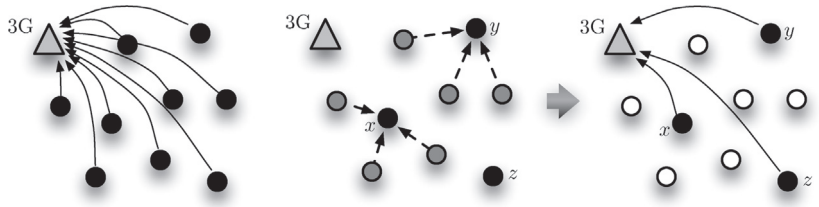
The recent boost up of mobile data consumption is straining cellular networks in metropolitan areas and is the main reason for the ending of unlimited data plans by many providers. To address this problem, we propose the use of series opportunistic delegation as a data traffic offload solution by investigating two main questions: (i) "Can we characterize a given social mobile scenario by observing only a small portion of it?". (ii) "How to exploit this characterization so to design solutions that alleviate overloaded cellular networks?". In our solution we build a social-graph of the given scenario by observing it for a period as short as 1-week, and then leverage a few, socially important users in the social-graph—the VIPs—to offload the network. The proposed VIP selection strategies are based on social network properties and are compared to the optimal (offline) solution. Through extensive evaluations with real and synthetic traces we show the effectiveness of VIP delegation both in terms of coverage and required number of VIPs – down to 7% in average of VIPs are needed in campus-like scenarios to offload about 90% of the traffic.

1. Introduction

Since the modern smartphones have been introduced worldwide, more and more users have become eager to engage with mobile applications and connected services. This eagerness has boosted up sales in the market – more than 64% up annually worldwide in Q2 2010 [1]. Simultaneously, smartphone owners are using an increasing number of applications requiring the transfer of large amounts of data to/from mobile devices. Opportunistic applications [2], crowd-source based ones [3], global sensing [4,5], and content distribution [6] are just a few of the examples. As a consequence, the traffic generated by such devices has caused many problems to cellular network providers. AT&T's subscribers in USA were getting extremely slow or no service at all because of network straining to meet iPhone users' demand [7]. The company switched from unlimited traffic plans to tiered pricing for cellular data users in summer 2010. Similarly, Dutch T-Mobile's infrastructure has not been able to cope with intense mobile traffic, by thus forcing the company to issue refunds for affected users [8]. All these issues are bringing new technical challenges to the networking and telecommunication community. In fact, finding *new ways to manage* such increased data usage is essential to improve the level of service required by the new wave of smartphones applications. One of the most promising solutions to avoid overwhelming the cellular network infrastructure is to *offload part of the traffic onto* direct communication between wireless devices whenever possible. The offloading process targets at the part of the data that tolerates some delay before delivery. This means that data is stored and transferred

* Corresponding author. Tel.: +39 06 4925 5164.
E-mail address: stefa@di.uniroma1.it (J. Stefa).

http://dx.doi.org/10.1016/j.adhoc.2014.01.012
1570-8705/



(a) Pure infrastructure mode. In this case, all nodes upload their data in an individual basis.

(b) By using opportunistic offloading, it becomes possible to reduce redundant traffic sent over the cellular network infrastructure.

Fig. 1. Network without and with opportunistic offloading.

directly between nodes in an ad hoc fashion until the deadline arrives and, if necessary, nodes access the cellular infrastructure to download or upload the data. Note that in the case of redundant data (e.g., measurement of pollution levels), the traffic on the cellular infrastructure can be drastically reduced through opportunistic offloading (see Fig. 1(a and b) for an example in the data gathering case).[1]

In this paper, we propose *VIP delegation*, a solution to this problem based solely on the inherent social aspects of user mobility. Our idea is to exploit a *few, important* subscribers (users) that with their movement and interactions are able to possibly contact regularly all the rest of the network users. These VIP devices would act as a bridge between the network infrastructure and the remaining of the network, each time large amount of data has to be transferred. Because nodes move and meet each other, it suffices to contact a subset of them to reach a large number of nodes (possibly all of them). In the example of the figure, nodes $\{x, y\}$ can serve as VIPs on behalf of the others, as they meet by themselves the rest of the network.

VIP delegation can help alleviate the network traffic in different delay-tolerant scenarios. Distribution of content to users by service providers, free updates of mobile software, system patches and so on, are just some examples that show how the network can exploit VIPs. However, their role for upload traffic (from the nodes to the network) is even more prominent: VIPs drastically reduce the number of competing upload network flows. As a direct consequence, also the number of collisions, and most importantly, of retransmissions decreases drastically. This becomes very valuable for users in overloaded networks. That said, we believe that VIPs can be even more valuable when involved in collection of urban-sensing related data [4,5]. In these cases data can be aggregated under the control of a central entity (e.g., government institutions) so to give feedback to the users on the environment they are living in (e.g. average noise pollution, smog, and so on). This aggregation could happen at the VIPs, as they collect the data sensed by the users. At the end of the coverage period, the VIPs would only send the aggregated value through the cellular network, by thus actually reducing the traffic flowing through the cellular network.

[1] There are several other offloading alternatives (through Wi-Fi access points for example, as discussed in Section 2.1). In this paper, we focus only on the case of opportunistic offloading.

The question here is *how to compute an appropriate VIP set given some requirements*. For this, we present, formalize, and evaluate two methods of VIPs selection: *global* and *neighborhood* VIP delegation (see Section 3). While the former focuses on users that are *globally* important in the network (namely, *global VIPs*), the latter selects users that are important within their *social communities*. The importance of a user within the network is given in terms of well-known attributes such as centrality (betweenness, degree, and closeness) and PageRank. In both cases, we observe that a short observation period (one week) is enough to detect users that keep their importance during long periods (several months). Selected nodes are then used to cover the network during a certain time window, through solely direct wireless contacts with the remaining nodes (see Section 4). In this paper, we provide significant extension over a companion work by deeply investigating coverage aspects [9].

Through extensive evaluations on real-life and synthetic traces, we evaluate the performance of the global and neighborhood VIP delegation methods in terms of network coverage, by varying the number of VIPs chosen (see Section 6). We compare our solution with an optimal benchmark computed from the full knowledge of the system. The results reveal that our strategies get very close to the performance of the benchmark VIPs: Only 5.93% page-rank VIPs against almost 4% of the benchmark VIPs are required to offload about 90% of the network in campus-like scenarios. Additionally we discuss on possible VIP incentives, the way VIPs offload the traffic accumulated to the network, and leveraged applications in Section 8. Finally, we conclude with Section 9.

2. Related works

We go through the related work in the area, discussing the most representative results on both data offloading and user-aided networking services.

2.1. Data offloading

Consumption of mobile data by the pervasive usage of smartphones is forcing carriers to find ways to offload the network. So far the most reasonable solution to the problem is offloading to alternate networks, such as femtocells and Wi-Fi. Femtocells exploit broadband connection to the service provider's network and leverage the licensed

spectrum of cellular macro-cells to offer better indoor coverage to subscribers [10]. As a side effect, automatic switching of devices from cellular network to femtocells reduces the load of the network. Besides from being localized (indoors only), such solution suffers from the non-proliferation of femtocells to subscribers' homes. Moreover, charging users for the necessary equipment as the network providers are currently doing (150 USD for AT&T's microcell) will not help in this direction. On the other hand, the proliferation of modern Wi-Fi enabled smartphones, together with the network providers' tendency towards already existing technologies has turn Wi-Fi offloading into a reality. More and more carriers worldwide are investing in this direction [11], by installing access points and hot-spots close to overloaded cellular towers, and by providing to clients Wi-Fi access within tiered monthly subscription. In this direction, Balasubramanian et al. propose a system to augment access to cellular network through Wi-Fi offloading [12]. This system, called Wiffler focuses only on Internet access from *moving vehicles*. It leverages delay tolerance and fast switching of devices to overcome the poor availability and performance of Wi-Fi.

Even though offloading to Wi-Fi seems to be the best solution so far to cellular network overloading [13], the continuous increasing of mobile data-traffic demand suggests for integration of Wi-Fi with other offloading methods. Indeed, according to a CISCO report,[2] the mobile data traffic will increase 18-fold within 2016, in front of a mere 9-fold increase in connection speeds. This huge traffic increase is very likely to pose problems also to public Wi-Fi access points. Most importantly, this forecast in mobile traffic demand outgrows the capabilities of planned cellular technology advances like 4G and LTE. According to Ericsson's CTO, there is strong scepticism about possible further improvements brought by 5G technology.[3] As we will show in this paper, our solution is essentially different from Wi-Fi and femtocell-based offloading; nevertheless, it can be integrated to these methods to further help alleviate mobile data overloads.

2.2. User-aided networking

Polat et al. suggest some sort of network members' promotion to enhance network functionalities [14]. They focus on providing multi-hop connectivity in a mobile ad hoc network. Their solution makes use of the concept of connected message ferry dominating set (CMFDS), where ferry-members of the network are connected over space–time paths. Besides from the difference in both problematic and application scenario with our work, no social aspect/importance of the network members is considered in promotion.

Many research works targeting social mobile networks make use of social ties between users to leverage network services. To the best of our knowledge, Han et al. were the first to exploit opportunistic communication to alleviate data traffic in cellular networks [15]. Later on, they extended their work in [16]. However, conversely from ours, their solutions only apply to information dissemination problems such as broadcasting. They focus on selecting k target users to which the information is first sent through cellular network. These target users will then, through *multi-hop* opportunistic forwarding, disseminate the information to all users in the network. We believe that multi-hop forwarding can be applicable to broadcasting, since users might willingly volunteer to share with others the same data they are interested in (data that they would anyway pull from other people). But, if the user is not directly interested in the data, collecting (disseminating) and eventually multi-hop forwarding data generated from other users becomes a burden without any gain for her (e.g., scenarios where different data is being distributed for different users, or collection and aggregation of sensing data [4,5]). In these cases, it would be very costly to stimulate all the users in the network to cooperate. Rather, our solution relies on upgrading a crucial small set of users' devices (down to 5.93% according to experiments with real campus-like data traces) that, through direct contact with network members, help alleviate the data traffic in both upload and download directions, assuring that no packet is lost. That said, while we believe that multi-hop forwarding is not applicable in our setting, for the sake of completeness we have compared the coverage performance of the sets obtained by the Heuristic strategy [16] with our VIP sets on the Dartmouth real trace.

Finally, also Push-and-Track, presented in [17] is relevant, though different to our work in various aspects: Firstly, it focuses on the dissemination scenario only. Second, in Push-and-Track the infrastructure relies on some performance targets to determine how many copies should be injected. Thirdly, and most importantly, Push-and-Track only makes use of the time a node enters the network, its geographic position, or its connectivity degree. It does not investigate the node interactions so to derive reliable future communication possibilities, as we do in this work.

3. VIP delegation in a nutshell

In view of the scenario presented in the introduction, we propose an offload method based solely on the social aspects of user mobility. Our idea is to detect subscribers (users) that, with their inherent mobility, are able to *encounter* a large number of users (possibly all them) in a regular fashion. These VIPs would act as a bridge between the network infrastructure and the remaining of the network, each time large amount of data has to be transferred.

The movement of smartphone users is not random; rather, it is a manifestation of their social behavior [18–21]. This movement, along with contact-based interactions among users, generates a social mobile network. The analysis of these mobility patterns and the understanding of how mobile users meet play a critical role at the design of solutions/services for such kind of networks. In particular, though the number of network users can be very high, just a few of them have an "important" role within the social graph induced by the encounters. The natural behavior

[2] Cisco visual networking index: Global mobile data traffic forecast, 2011–2016.

[3] http://www.dnaindia.com/money/interview_there-will-be-no-5g-we-have-reached-the-channel-limits-ericsson-cto_1546408.

of these VIP nodes, which are considerably fewer than the rest of the network, can be a valuable resource in both information dissemination and collection to/from the rest of the network. Motivated by the fact that opportunities for users to exchange data depend on their habits and mobility patterns, our idea is the following: turn those *few* VIP nodes into bridges between regular users and the Internet, each time large amount of data is to be uploaded/downloaded by these latter ones. In a word, VIPs would act as delegates of the network infrastructure builder. As a side effect, this would immediately drop down the cellular network usage. That said, in the following of this manuscript we will denote as *covered* a user that is visited by a VIP node. Similarly, we will denote as a *covered network* a network whose every user has been covered.

In our scenario, we assume that users download/upload large amount of data. This would demand a lot of networking resources and thus, makes the use of multi-hop protocols unfeasible. Indeed, it is quite hard to convince the average user to act as a relay for others, even though to the closest access point, of such an overloading traffic. Rather, our solution relies on the upgrade of the devices, or payment, of a small, crucial set of VIP nodes that regularly visit network users and collect (disseminate) data to them on behalf of the network infrastructure. Such upgrade would serve as incentive to users to play the role of VIPs (see Section 8 for discussions on possible VIP incentives).

Now the problem becomes: *How to choose the smallest VIP set that with their natural movement in the network cover all users in a certain time window?* More formally, the problem is defined as follows: Let $N = \{n_1, \ldots, n_n\}$ be the network nodes and let $G_i = (V_i, E_i)$ be the graph whose set of vertexes $V_i \subset N$ are the network nodes that have at least a meeting during the time window i, whereas the set of edges E_i represents those meetings, i.e. $\{u_1, u_2\} \in E_i$ iff u_1 and u_2 meet in the time window i. We are looking for $S \subset N$ such that $S = argmin_{S \subset N} \mid \cup_i [(V_i \setminus (S \cup \{u | \{u, s\} \in E_i, s \in S\})] \mid$. Note that set S can be seen as the smallest set of vertexes from N that dominates the nodes in each G_i according to the respective E_i. Though it has a dominating set flavor, this problem is different from it: Indeed, here we deal with a series of graphs instead of a singular one.

As previously mentioned, we solve this problem by presenting two VIP selection methods that rely on either a global or a local view of the network (the methods are detailed in Section 4). We also present a benchmark solution for VIP delegation. The benchmark provides an optimal selection method that (i) requires the complete pre-knowledge of users' behavior and (ii) is based on an adaptation of the well known NP-hard problem of the Minimum Dominating Set [22]. Such a method is clearly not feasible in real-life applications, but useful to evaluate the performance of our social-based VIP selection methods.

4. VIP selection methods

The selection of VIPs in a social mobile network is based on the ranking of nodes according to their social structural attributes and requires knowledge on their mobility. For this, a social graph describing the tightness of links in the network has to be designed. As the authors of [23] show, the performance of network protocols is strictly related to the accuracy of the mapping between the mobility process and the network social graph. They propose an online algorithm that uses concepts from unsupervised learning and spectral graph theory to infer the "correct" graph structure. However, this approach is not applicable in our case, where VIPs are to be predicted. We thus decide to follow a simpler method in detecting the network's social graph, based on the following intuition: The movement of users guided by their interests generates repeatability in their behaviors (e.g., go to work/school every day, hang out with the same group of friends) [18–21]. Intuitively, by observing meeting patterns for a certain monitoring period reveals enough information to characterize the tightness of the social links in the network graph.

In a real-life application, we could imagine the network infrastructure builder asking participating users to log their meetings for a certain time, called here as *monitoring period*. These logs serve then to build the networks' social graph on which the VIPs selection is made. More specifically, from mobility patterns and wireless interactions of users in a network, we establish a *social* undirected graph $G(V; E)$, where V is the set of users and E is the set of social ties (encounters) among them. Note that such social graph is different from each of the $G_i = (V_i, E_i)$ we mentioned in the previous section. Indeed, G_i represents *exactly* who meets who in a certain time window i; whereas G is only a representation of the tightest *friendship* relations among network nodes that appears during the monitoring period, which is composed of a set of time windows i. Here, by friendship we mean some sort of mobility tie among users in the network. The details of the construction of the social graph will be given in Section 6.4. However, we anticipate that social ties (edges) in the graph $G(V; E)$ are strictly related to users' contact frequency: A link exists between two users if the number of times they meet is larger than a certain threshold, which depends on the considered networking scenario.

At the following, we present our global and local VIPs selection methods as well as the social structural attributes used at nodes ranking.

4.1. VIPs selection methods

4.1.1. Global VIP selection

In the global selection, all network nodes are first ordered according to their importance in the network, determined by their social structural attributes (see Section 4.3). Afterwards, the smallest VIP set over the *global social graph* that covers the network during a certain time window through direct contacts is chosen, by applying one of the following VIP promotion methods:

- *Blind global promotion.* It selects the top-ranked nodes not yet promoted, until the network is covered.
- *Greedy global promotion.* This is a set-cover flavored solution. In particular, it starts with promoting to VIP the top-ranked node. After this promotion, the nodes covered by this VIP are dumped and ranking on the remaining nodes are re-computed. Again, the procedure is repeated until the network is covered.

4.2. Hood VIP selection

The second strategy, neighborhood VIP delegation, is based on the intuition that repetitive meetings among people happen usually in the same places. The mobile social network generated by this behavior encompasses, besides contact locality, well tight social-community sub-structures. With this in mind, the hood strategy aims to cover each community at a time, independently from other communities. It selects *hood VIPs* by their importance within the communities they belong to. Before doing so, we first detect social-communities using the *k-clique* community algorithm [18]. The reason behind this choice is that the *k*-clique algorithm can detect *overlapping* communities, i.e., nodes may appear to belong to more than one community. This characteristic of the *k*-clique algorithm makes it well suited for a scenario like ours in which people belong to more than one social community (e.g. gym members that are also computer science students, etc.). This is also the reason why the *k*-clique algorithm is widely used in the area of social mobile networking [19,24,25,27]. Finally, to study the overlapping between communities, we use the Jaccard similarity index [26]. For two sets A and B is computed as: $J_{A,B} = \frac{|A \cap B|}{|A \cup B|}$.

Afterwards, we rank members of each community according to their importance in the network (see Section 4.3). Then, we start covering each community by promoting its members to VIPs similarly to the global VIPs methods:

- *Blind hood promotion.* It continuously selects the top-ranked nodes not yet promoted in the community, until the network is covered.
- *Greedy hood promotion.* The highest-ranked member in the community is promoted, nodes it covers within the community are dropped, and rankings are computed again in the remaining graph.

In both promoting ways, when the whole community is covered, the procedure continues with another one, until all the communities are covered.

4.3. Social structural attributes of nodes

We define the importance of a node in the network by applying to the network social graph several structural *attributes*: betweenness centrality, closeness centrality, degree centrality, and PageRank. All these are well-known attributes in social network theory [28,29]:

4.4. Betweenness centrality

Measures the number of occurrences of a node in the shortest-path between pairs of others nodes. It thus determines "bridge nodes" that, with their movement, act as connectors between node groups (communities). For a given node k it is calculated as:

$$C_B(k) = \sum_{\substack{j=1 \\ j \neq k}}^{N} \sum_{\substack{i=1 \\ i \neq k}}^{N} \frac{g_{i,j}(k)}{g_{i,j}},$$

where N is the number of nodes in the network, $g_{i,j}$ is the total number of shortest paths linking i and j, and $g_{i,j}(k)$ is the number of those shortest paths that include k.

4.5. Degree centrality

Ranks nodes based on the number of their direct ties (i.e., neighbors) in the graph. It identifies the most popular nodes, also called *hubs* in social network theory, possible conduits for information exchange. Degree centrality is calculated as: $C_D(k) = \sum_{i=1}^{N} a(k,i)$, where $a(k,i) = 1$ if k and i are linked, and $a(k,i) = 0$ otherwise.

4.6. Closeness centrality

Ranks higher nodes with lower multi-hop distance to other nodes of the graph. It describes "independent nodes" that do not dependent upon others as intermediaries or relayers of messages due to their closeness to other nodes. The closeness centrality for a node k is calculated as $C_C(k) = \frac{N-1}{\sum_{i=1}^{N} d(k,i)}$, where $d(k,i)$ is the length of the shortest path between nodes k and i. To deal with disconnections it is computed within the subgraph induced by the elements of the connected component to which k belongs.

4.7. PageRank

The well known Google's ranking algorithm, measures the likelihood of nodes in having important friends in a social graph [29]. In particular, PageRank of a node i in the social graph is given by the equation $PR(k) = \frac{1-d}{N} + d \sum_{i \in F(k)} \frac{PR(i)}{|F(i)|}$, where d $(0 \leqslant d \leqslant 1)$ is the damping factor and $F(k)$ is the set of neighbors of k in the social graph (the graph is undirected). The damping factor d controls the amount of randomness in page ranking: Values close to 1 will give high PageRank to socially best-connected nodes.

Finally, note that, though betweenness centrality and closeness centrality are metrics defined for multi-hop applications, they do capture somehow the node popularity within the network: E.g., a very high-degree node is very likely to have also a high closeness and a high betweenness. This is why we considered also these metrics in our strategies.

5. Benchmark approach

To evaluate the efficiency of our strategies, we propose a benchmark approach that gives the optimal solution: 100% of user coverage in a time window of one day, with minimum number of VIPs. It is important to underline that the benchmark serves only for comparison purposes, as it requires knowing the future to compute the exact set of VIPs.

5.1. Application scenario's abstraction

Suppose the network has to be covered daily (i.e., the time window i is of one day) by VIP delegates, for a period P during which the activity of all network users is known.

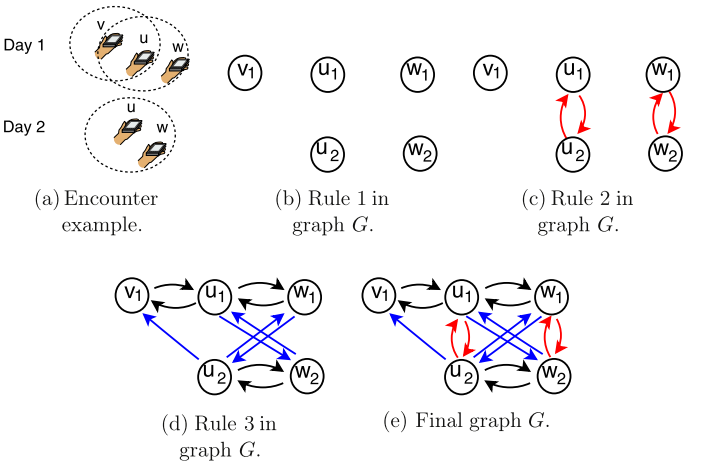


Fig. 2. (a) Meeting between u, v, and w during days 1 and 2. (b–d) Rules for the construction of graph G. (e) Final representation of graph G.

Let also P be n days long. We construct a directed graph $G = (V, E)$ through the following rules (a step-by-step generation of graph G is illustrated in Fig. 2):

Rule 1. Graph G has a vertex u_i for each day i in which user u is active (i.e., u has at least one contact during the day). This vertex impersonates u during day i in G and is referred to as the *clone* of u during that day (see Fig. 2(a and b)). The set of clones representing user u in G is denoted as C_u.

Rule 2. For every user u, C_u forms a clique in G, i.e., each pair of clones u_i, u_j of user u in G is connected by two directed edges, namely, (u_i, u_j) and (u_j, u_i) (see Fig. 2(c)).

Rule 3. If users u and v meet on day i, then every member u_t of C_u is connected to v_i through an edge (u_t, v_i). Similarly, every member v_t of C_v is connected to u_i through an edge (v_t, u_i). In particular, G also contains edges (v_i, u_i) and (u_i, v_i) representing that u and v met on day i (see Fig. 2(c)).

The graph G constructed with the above rules represents users' behavior in the network during the whole period P. Take for example a certain user u. According to Rules 1 and 2, user u is "expanded" in G into a clique C_u, containing clones of u only for the days u is active (see Fig. 2(b and c)). Moreover, if u meets v in day i, Rule 3 guarantees that all members of C_u point to v_i (see Fig. 2(d)). The intuition behind this rule is that outgoing edges from u's clique indicate that "u can be a delegate for v on day i".

Rule 3 is applied to every day on which user u is active. As a consequence, all members of the u's clique in G point to the same members of *other users' cliques*. Thus, any clone of user u in G (any member of C_u) is enough to determine the set of users v for which u can be a delegate, and on which days.

5.2. Benchmark delegates selection

Intuitively, in order to cover all the network day by day, it is enough to select as delegates the members of a minimum out-dominating set of graph G. Moreover, such a set of delegates is the smallest set that can achieve full coverage. The following theorems prove such intuition.

Theorem 1. *Let MDS be a minimum out-dominating set of G. The set MDS can cover 100% of the active users for each day $i \in P$.*

Proof. First recall that according to Rule 2, the set C_u of clones of a same user u form a clique in G. Since *MDS* is minimum, it contains *at most* one clone for every user u. When members of such a set are promoted to delegates, we get at most one delegate-instance per user.

Suppose, without loss of generality, that user v is active during day i, i.e., $v_i \in G$. As *MDS* is a dominating set, either of the following cases might happen: (i) some clone v_t of v is in *MDS* or (ii) there is at least one other node u's clone $u_l \in MDS$ such that the edge (u_l, v_i) is in G. In case (i), since $v_t \in MDS$, v is promoted to a delegate and is covered by itself. Case (ii) can only happen if edge (u_l, v_i) was added by Rule 3, i.e., u and v met during day i. Given that $u_l \in MDS$, u is promoted to a delegate. Thus, v is necessarily covered on day i. □

Theorem 2. *Let MDS be a minimum out-dominating set of G. Let also S be the smallest set of VIP delegates able to cover, for every day $i \in P$, 100% of the active network users on day i. Then, $|MDS| \leqslant |S|$.*

Proof. Suppose, on the contrary, that $|S| < |MDS|$. By construction, and with a similar reasoning used in the proof of Theorem 1 it is easy to see that S is an out-dominating set of G. Then, by the minimum cardinality of *MDS*, we are done. □

The above theorems indicate how to proceed to find the best possible solution to our problem: after constructing graph G according to Rules 1–3, find a minimum out-dominating set of G and use the members of such set as benchmark VIP delegates.

The minimum dominating set is notably a NP-hard problem. Thus, to individuate our benchmark VIP delegates, we reduce our problem to Set Cover (equivalent to *MDS* under L-reductions [22]) for which a simple greedy algorithm is known to be only a logarithmic approximation factor away from the optimum [22]. Moreover, the inapproximability result for this problem shows that no polynomial algorithm can approximate better than $(1 - o(1)) \log n$ unless NP has quasi-polynomial algorithms. Thus, there is no polynomial-time algorithm with a smaller approximation factor. The delegates obtained by this heuristic are then used as benchmark VIPs in our experiments.

6. Experimental setting: from data-sets to social graphs

We now give detailed information on the data-sets (real and synthetic) we use in the evaluation of our strategies.

6.1. Real data-sets

Two real data-sets are used: Dartmouth [30] (movement of students and staff in a college campus) and Taxis [31] (movement of taxi cabs in San Francisco). The vehicular mobility of the cabs is different from human mobility (Dartmouth). However, the purpose of using the taxis trace is to test our solution's extendibility to different contexts.

6.2. Dartmouth

Dartmouth includes SNMP logs from the access points across the Dartmouth College campus from April 2001 to June 2004. To generate user-to-user contacts from the dataset, we follow the popular consideration in the literature that devices associated to the same AP at the same time are assumed to be in contact [32]. We consider activities from the 5th of January to the 6th of March 2004, corresponding to a 2-month period during which the academic campus life is reasonably consistent. We chose to work with the set of nodes that have a fairly stable activity in time: at least 500 contacts per week with *any* other device. This results in a set of 1146 nodes with an average of 1060 daily active devices and 292 daily contacts in average per device.

6.2.1. Taxis

The Taxi dataset contains GPS coordinates of 536 cabs collected over 24 days in San Francisco. Here, we assume that two cabs are in contact when their geographical distance is smaller than 250 m (following suggestions of Piorkowski et al. [31]). This yields an average of 491 active nodes per day and 7804 daily contacts per node.

6.3. Synthetic data-sets

Synthetic traces are generated using the SWIM model [33–35], shown to simulate well human mobility from statistical and social points of view. We use SWIM [35] to simulate a 500-node version of the Cambridge Campus dataset (of only 36 Bluetooth enabled iMotes, 11 days long) according to the Phoenix Model (scaling by keeping the density constant). We call this trace SWIM-500. It simulates user activity during 2 months, yielding 128 daily contacts per node in average. Then, we scale up to 1500 nodes in two ways: (i) by keeping density constant (D-SWIM-1500) and (2) by keeping the area constant (A-SWIM-1500). The purpose of the two different scalings is to study the behavior of our strategies in two scenarios: D-SWIM-1500 simulates an urban growing in both area and population and A-SWIM-1500 refers to a sudden over-population of a given city with people that are there to stay for a long time-period (e.g. students returning to the campus after summer holidays).

Table 1 summarizes the details of the data-sets. Note that, although both data-sets represent campus scenarios, they yield different activity per node per day as they used distinct technologies (Wi-Fi capable APs in Dartmouth and Bluetooth-like characteristics in SWIM) in the two data-sets.

6.4. Monitoring period and social graph

As we have already discussed in Section 4, we observe nodes encounters during an observation/monitoring period and exploit repeatability of users' movement patterns and recurrence of contacts among them. The used length of the monitoring period is not casual: *It is as short as 1 week* and divided in time windows of *1 day*. Usually, our life and the activities we conduct are organized on a week-base, mostly having a common routine repeated day by day (e.g., go to work/school or have lunch in the same place). Such repetition also infers the common meetings generated by those activities.

In the case of the Taxi dataset, the repeatability of contacts is due to several factors including the popularity of geographical zones in the city (e.g., center, stations, and airports), the fixed tracks leading to such zones, and the common city areas covered by groups of taxis. As shown by Piorkowski et al., popularity of areas generates clusters of connectivity among cabs [31]. Taxis' movements are

Table 1
Details on the real datasets and respective monitoring period. The term monitoring indicates that the parameters are shown for the monitoring-period only, whereas trace, for the whole trace duration.

Data set	Taxi	Dartmouth	SWIM-500	D-SWIM-1500	A-SWIM-1500
Total nodes	536	1142	500	1500	1500
AVG active nodes/day (trace)	491	1060	499.98	1500	1500
AVG active nodes/day (monitoring)	429	1061.5	500	1500	1500
AVG contacts/node/day (trace)	7804	292	128	130	380
AVG contacts/node/day (monitoring)	7656	284	131	129	378

guided by clients' (humans) necessity to reach a specific geographic location. Thus, one week observation is again enough to predict future meetings.

Our intuition on the length of the monitoring period is also confirmed by the results shown in Table 1. Indeed, the properties of the monitoring period are very close to the whole trace, for each considered scenario. Hence, this makes prediction of future meetings easy: the monitoring period we have chosen allows us characterizing social relationships. We are then able to generate a *social graph*, where two users are connected only if they have met with a certain frequency – that we call *social connectivity threshold* – during the monitoring period. The social connectivity threshold depends on the scenario considered:

- In the Dartmouth dataset, social connectivity in mostly due to the frequentation of the same classes, or studying in the same library, or living in the same dormitory. All these activities generate lots of meetings among people. We thus set the social connectivity threshold in this case to be at least 1 contact per day, for at least 5 days during a week.
- The social connectivity threshold in the Taxi dataset is higher due to higher speeds: at least 8 contacts per day during the monitoring period were considered.
- As the SWIM-500 trace also represents a University campus, we use the same social connectivity threshold of the Dartmouth trace: at least 1 contact per day for at least 5 days of the monitored week. This leads to a set of 498 nodes. When scaling up with constant density (D-SWIM-1500) the social connectivity threshold remains constant. It increases to at least 8 contacts per day for at least 5 days of the monitored week when scaling up with constant area (A-SWIM-1500).

The social graphs generated by the social connectivity thresholds are then used to individuate the VIP delegates, according to each of the strategies of Section 4.

6.5. Community detection

Our *hood VIPs* selection strategies operate on a community basis and aim at covering single communities by selecting members that are important within each community. After applying the *k*-clique algorithm [18] to determine the communities and with respect to the campus-like scenarios, we have the following parameters: the Dartmouth dataset has 24 communities of 41 members in average, the SWIM-500 trace has 16 communities with 32 members in average, the D-SWIM-1500 trace has 39 communities with 39.6 members in average, and the A-SWIM-1500 has 35 communities with 44 members in average. Note that constant-area scaling yields less, bigger communities.

The communities are well-knit and do not show much intersection between them. Indeed, the average Jaccard similarity index [26] between intersecting communities is 0.038 in the Dartmouth case and about 0.025 in SWIM-500 and D-SWIM-1500 case. This result supports recent findings on universities' communities detected with the *k*-clique algorithm [19]. Conversely, in the constant-area scaling of A-SWIM-1500, the communities have a higher overlapping: the Jaccard similarity index in this case is 0.045.

The Taxi dataset, due to the large number of contacts and the high mobility of nodes, does not present any community sub-structuring. When applying the *k*-clique algorithm, we observe one huge community containing almost 80% of the nodes, whereas the remaining 20% do not belong to any community. Thus, we decided to apply only the global VIP selection strategies to this trace.

7. Experimental results

We analyze the performance of all our strategies in terms of coverage when applied to real and synthetic traces. For better understanding the quality of the VIPs selected by each strategy, we investigate the coverage trend with regard to an increasing number of the VIPs. The set used for coverage is updated from time to time following the order in which each strategy selects VIPs. For the sake of comparison, the results for the benchmark ("Bn") are included in the plots. We use the same technique as above to build the benchmark's trend: updating the VIPs set and the corresponding network coverage, following the order in which the benchmark promotes nodes to VIPs. For the PageRank attribute, we noted that, varying the damping factor in the interval $[0.51; 0.99]$ does not change the performance of PageRank VIPs with respect to the VIPs selected according to other centralities. However, we decided to use $d = 0.85$, since, for PageRank, it results in the best performance in terms of network coverage.

7.1. Results with real data-sets: Dartmouth case

7.1.1. Blind promotion

We show in Fig. 3 the coverage obtained by each of the promotion strategies. The *blind* promotion in the global and hood VIP selection strategies yields the results presented in Fig. 3(a and b). Notice that there is a coverage efficiency gap between PageRank VIPs (referred as "PR" in the figure) and those of other centralities (referred as "BW", "DG", and "CL"). In addition, PageRank is very close to the benchmark, even for small percentages of delegates considered. For instance, in the *global blind strategy*, to get to 90% of coverage, PageRank only requires the promotion of 5.93% of nodes as delegates against 3.92% with the benchmark approach (see Table 2). Another consideration to be made is that hood selection is more effective than global selection. Hence, aiming to cover the network by forcing VIP selection within different communities seems to be a very good strategy. Nevertheless, there exist social attributes such as PageRank that do not gain much from the hood selection. Indeed, global and hood PageRank VIPs perform very similarly in both data-sets. This is because, on the one hand, PageRank VIPs already target different communities, even in the global case. On the other hand, betweenness, degree, and closeness centrality tend to over-select VIPs from a few network communities, and consequently, leave uncovered many marginal ones. The

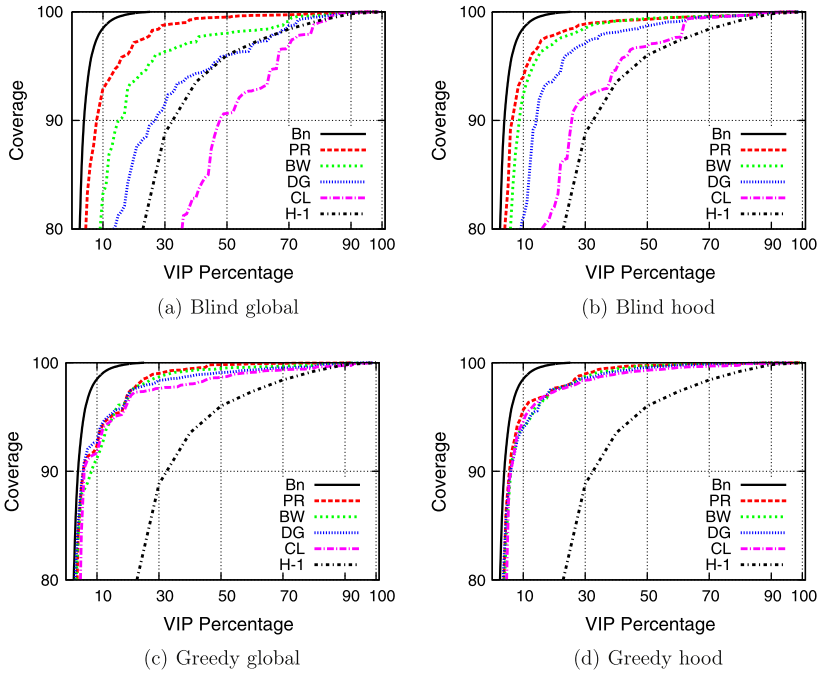


Fig. 3. Performance of the selection strategies on the Dartmouth dataset. "Bn" refers to the benchmark, "PR" to the page-rank, "BW" to betweenness centrality, "DG" to degree centrality, and "CL" to closeness centrality. "H–1" refers to the Heuristic strategy of [16].

Table 2
VIP sets cardinality to get 90% coverage on Dartmouth. The benchmark needs 3.92% of nodes.

	G-blind (%)	H-blind (%)	G-greedy (%)	H-greedy (%)
PR	8.98	6.89	5.93	6.19
BW	15.96	9.16	8.98	6.19
DG	26.96	15.09	5.93	6.19
CL	47.993	26.0035	5.93	6.19

tendency of these social attributes to target only a few communities is attenuated with the hood selection that boosts their efficiency in covering the network. In Fig. 4, we show how the global strategy distributes VIPs among communities of different centralities.

7.1.2. Greedy promotion

When applying the *greedy promotion*, the performance of all strategies improves considerably (see Fig. 3(c and d)). In addition, VIPs obtained with each social attribute perform very similarly to each other, in both hood and global selections. This is due to the capacity of the greedy approach to *not* promote as VIPs nodes that are too close to each other in the social graph. Indeed, after every node's promotion to VIP, all its neighbors in the social graph and their links are removed. Since communities are very well tight, the promotion of one member can remove a large part of the community (if not all of it). Thus, attributes such as betweenness and closeness *do not* concentrate their selection on a few communities as in the global selection. This is also confirmed by Fig. 5, where we show how the greedy strategy distributes VIPs among communities for different social attributes.

Finally, we have compared our strategies with the target-sets selected by the best-performing strategy in [16] which does not rely on knowledge on the future: the Heuristic strategy. As suggested in [16], for Heuristic we exploit a History period of 1-day before the day to be covered. The results are included in Fig. 3. We notice that, besides Closeness in the Global Blind case (Fig. 3(a)), all our other strategies outperform Heuristic independently on the selection methodology (either Blind/Greedy, or Global/Hood). In addition, Heuristic is very close to the performance of Global Blind Degree for VIP sets larger than 40%, though never outperforming it. These results are somewhat expected: Heuristic target-sets are not to be used in a scenario like ours, but they are to be exploited in scenarios with multi-hop forwarding. Heuristic selects as target-set members nodes that have a peak of popularity during some previous day, but that, on average, are not the most popular ones. While this is more than enough if other network nodes help with multi-hop forwarding, in scenarios like ours where network coverage relies on the VIPs only, the performance is much more sensible to the behavioral changes of the nodes. Here we believe our training-period comes to help: it gives the strategies more information on the behavior of the nodes in the network. In addition, it simplifies the system (the set of VIPs is fixed and not re-computed daily as in Heuristic) and helps filtering-out potentially misleading perturbations in the data. Secondly, even when the percentage of VIPs (Heuristic target-sets) is high enough to smooth out the first inefficiency, Heuristic is not able to distribute very well the VIPs across the different communities, which we have shown to be the reason why the Pagerank and Betweenness Centrality based

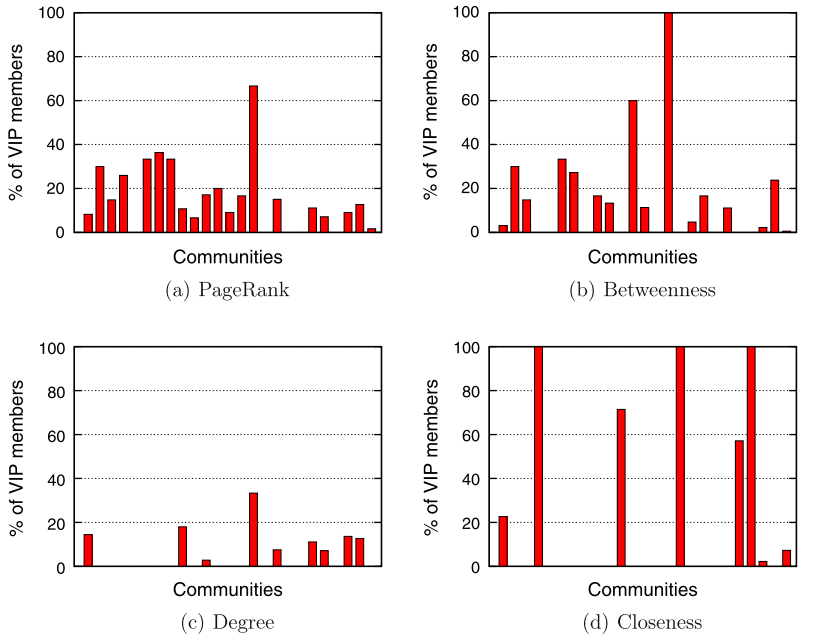


Fig. 4. Distribution of VIPs per social attribute on the Dartmouth dataset with the *blind global* promotion strategy. The *x*-axis represents different communities detected.

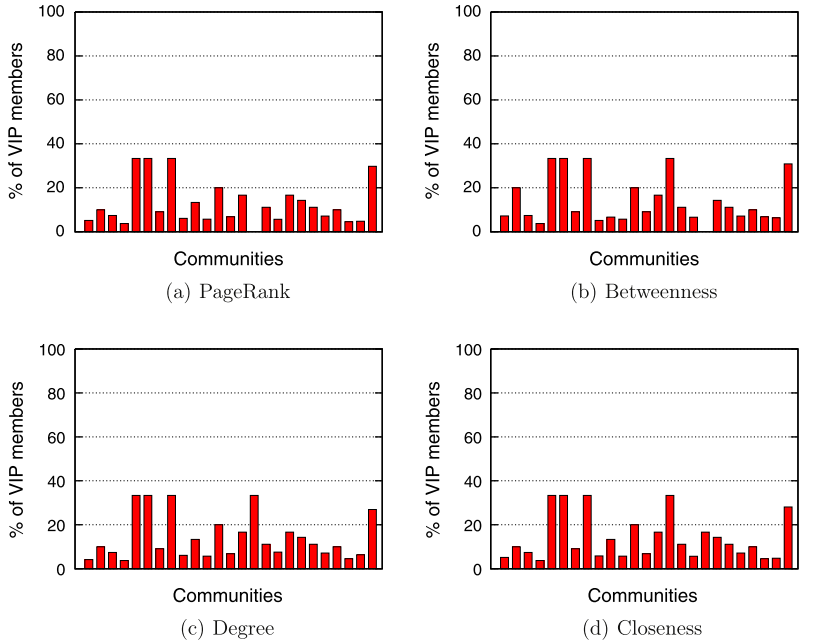
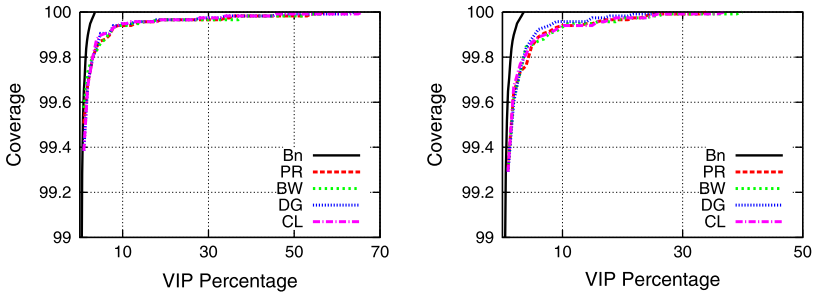


Fig. 5. Distribution of VIPs per social attribute on the Dartmouth dataset with the *greedy global* promotion strategy. The *x*-axis represents different communities detected.

selection strategies significantly outperform all the other ones in the Global-Blind and Global-Greedy case.

7.2. Results with real data-sets: Taxi case

As discussed in Section 6.5, the community sub-structuring of the Taxi dataset is flat. This means that, due to the high mobility of nodes, a huge unique community containing 80% of nodes is detected and the 20% remaining nodes do not belong to any community. Thus, only the global selection strategy is applicable to this dataset. Fig. 6 shows the performance of blind and greedy global selection strategies in terms of coverage for the Taxi dataset. As we can see, due to the high mobility of nodes, all



(a) Blind global (b) Greedy global

Fig. 6. Performance of blind global and greedy global selection strategies on the Taxi dataset. "Bn" refers to the benchmark, "PR" to the page-rank, "BW" to betweenness centrality,"DG" to degree centrality, and "CL" to closeness centrality.

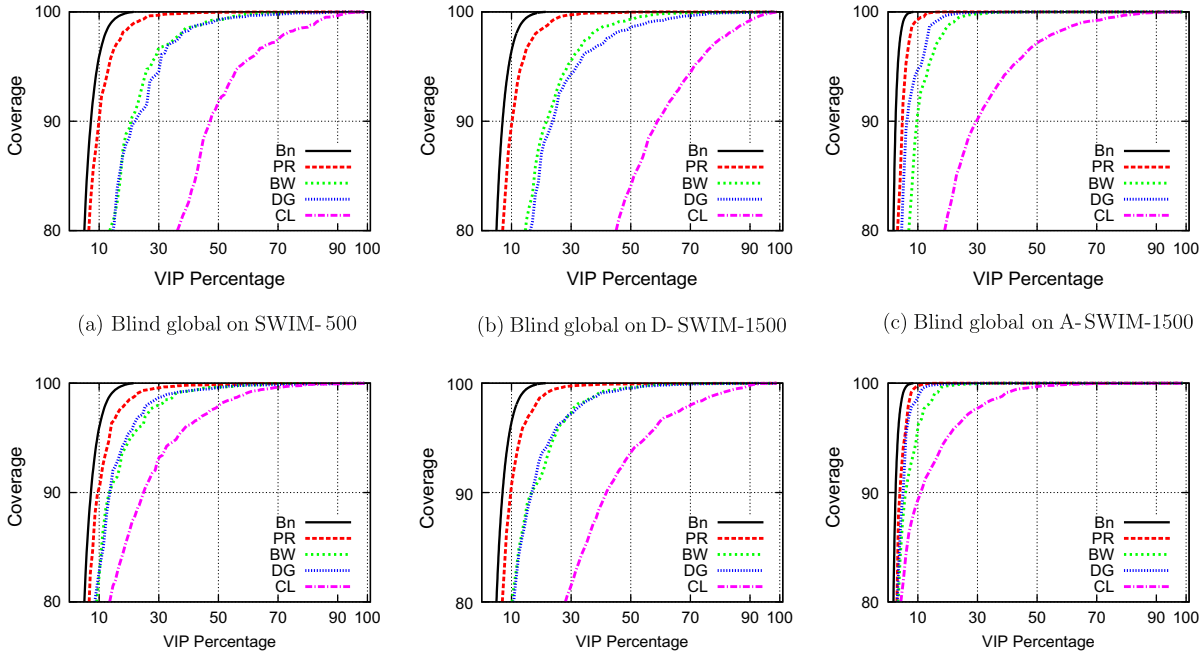
strategies perform very well in this scenario. Moreover, the sets selected by each strategy to guarantee up to 90% of coverage are exactly of the same (small) size: Only 0.93% of network nodes. The benchmark guarantees the same coverage with 0.2% of network nodes selected.

7.3. Results with synthetic data-sets: SWIM

As discussed in Section 6.3, starting from SWIM-500 (a 500-node simulation of a University scenario [36]), we generate two scaled versions with 1500 nodes: D-SWIM-1500 (constant density scaling) and A-SWIM-1500 (constant area scaling). Our purpose is to study the reaction of the different strategies in two cases: an urban growing in both area and population (constant density) and a sudden over-population of a city (constant area).

We start from *blind promotion* (see Fig. 7). Again, like in the Dartmouth scenario, PageRank VIPs are more efficient than VIPs of other centralities. The reason is the same as discussed in the previous section, i.e., PageRank global VIPs are better distributed within communities with respect to VIPs obtained with other centralities. This is also confirmed by Fig. 9 where such distribution is shown for the trace D-SWIM-1500 (the relative figures for traces SWIM-500 and A-SWIM-1500 are omitted due to space constraints). Once again, aiming to cover the network by forcing VIPs to fall in different communities (hood selection) is a winning strategy.

Results related to *greedy promotion* are presented in Fig. 8. As in the Dartmouth case, the performance of all strategies is boosted up by the greedy selection of VIPs, yielding a better distribution of delegates within communities (see Fig. 10) and thus, much better coverage results



(a) Blind global on SWIM-500 (b) Blind global on D-SWIM-1500 (c) Blind global on A-SWIM-1500

(d) Blind hood on SWIM-500 (e) Blind hood on D-SWIM-1500 (f) Blind hood on A-SWIM-1500

Fig. 7. Performance of (a–c) blind global and (d–f) blind hood selection on SWIM. "Bn" refers to the benchmark, "PR" to the page-rank, "BW" to betweenness centrality, "DG" to degree centrality, and "CL" to closeness centrality.

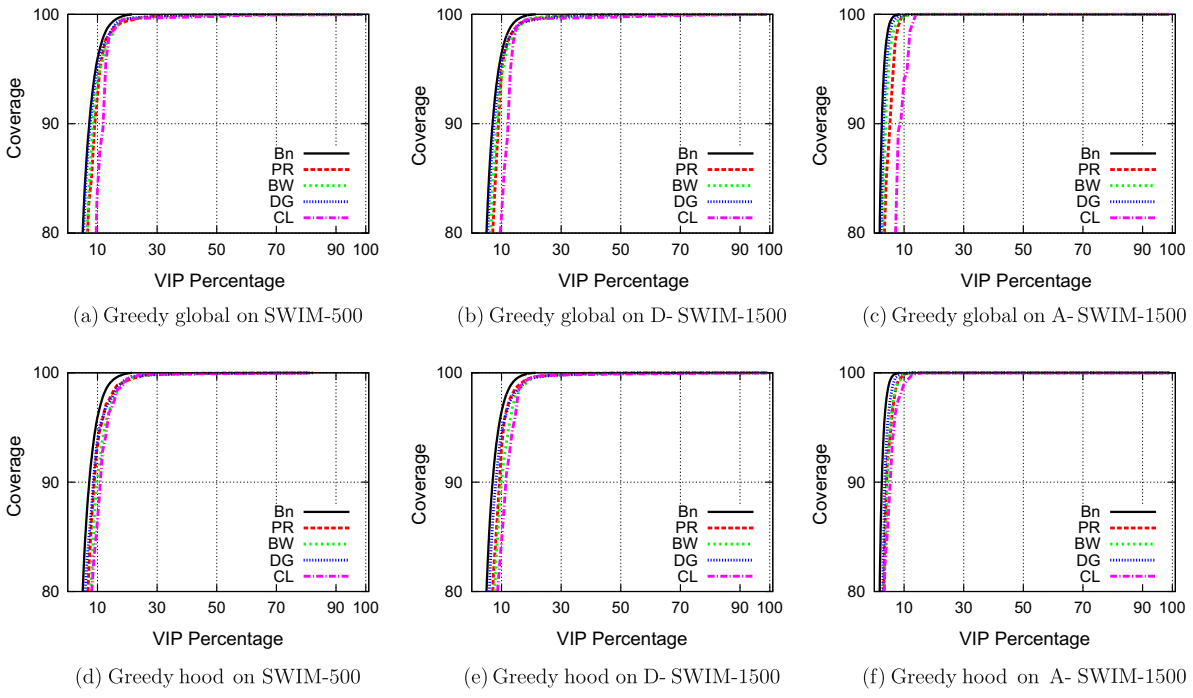


Fig. 8. Performance of (a–c) greedy global and (d–f) greedy hood selection on SWIM. "Bn" refers to the benchmark, "PR" to the page-rank, "BW" to betweenness centrality, "DG" to degree centrality, and "CL" to closeness centrality.

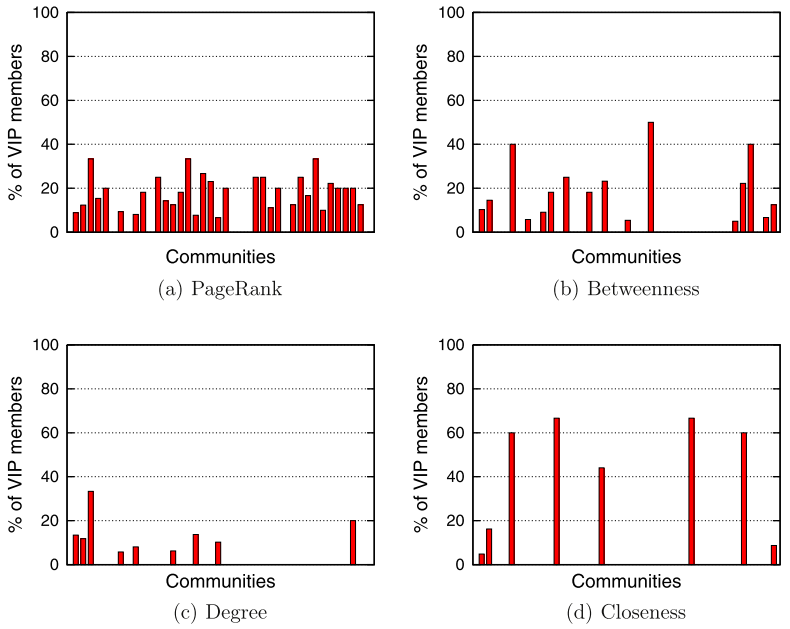


Fig. 9. Distribution of VIPs per social attribute on the D-SWIM-1500 dataset with the *blind global* promotion strategy. The *x*-axis represents different communities detected.

with respect to the *blind promotion*. What is interesting to notice here is the impact of the way of scaling in our strategies. When passing from SWIM-500 to D-SWIM-1500 (constant density), all strategies perform very similarly in both blind and greedy promotions. Conversely, in an emergency situation where the network is suddenly much more overloaded as a result of the over-population of the network area (A-SWIM-1500), our strategies perform even

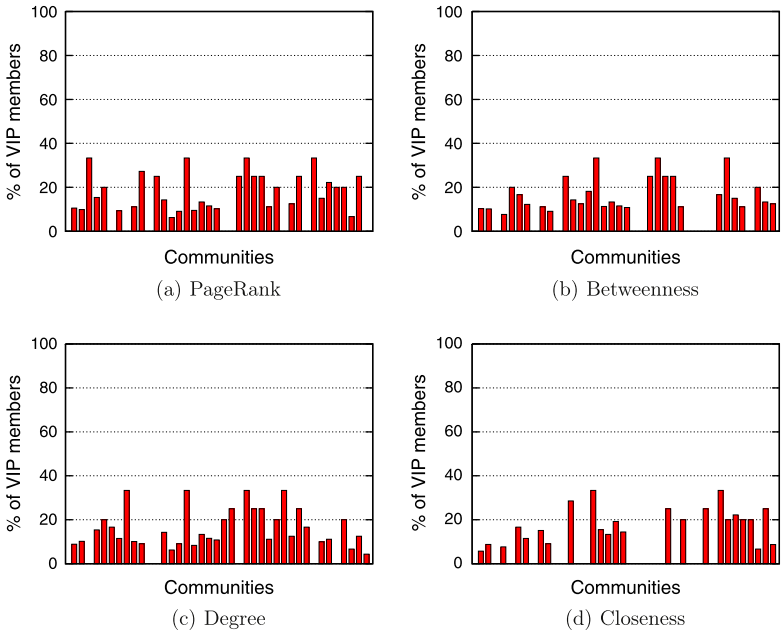


Fig. 10. Distribution of VIPs per social attribute on the D-SWIM-1500 dataset with the *greedy global* promotion strategy. The *x*-axis represents different communities detected.

Table 3
Delegates given by each strategy to get 90% coverage on SWIM-500. The benchmark approach needs 7.4%.

	G-blind (%)	H-blind (%)	G-greedy (%)	H-greedy (%)
BW	21	14	9	10.6
CL	48	25.4	12.8	11.6
DG	23	13.6	8	8.8
PR	10.8	10.2	9.8	9.6

Table 4
VIP sets cardinality to get 90% coverage on D-SWIM-1500. The benchmark approach needs 7.06%.

	G-blind (%)	H-blind (%)	G-greedy (%)	H-greedy (%)
BW	22	17.26	9	9.93
CL	59	42.06	12.93	11.6
DG	24	17.2	8	9.06
PR	10.9333	10.06	9	9.93

Table 5
VIP sets cardinality to get 90% coverage on A-SWIM-1500. The benchmark approach needs 2.53%.

	G-blind (%)	H-blind (%)	G-greedy (%)	H-greedy (%)
BW	10	6.73	4	4.4
CL	30	11.6	9	6.06
DG	7	5.13	4	3.53
PR	5	4.33	6	4.4

better (see Figs. 7(c), (f), 8(c), and (f)). This is also confirmed by the results shown in Tables 3–5 that contain, for each dataset, the percentage of delegates needed by the different strategies to cover 90% of the network.

Table 6
Coverage potential for each strategy on Dartmouth. The benchmark's potential is 0.91.

	G-blind	H-blind	G-greedy	H-greedy
PR	0.830	0.069	1.0	0.831
BW	0.657	0.067	1.0	0.969
DG	0.530	0.061	1.0	0.890
CL	0.144	0.059	1.0	0.886

Table 7
Coverage potential for each strategy on TAXI. The benchmark's potential is 0.96.

	G-blind	G-greedy
PR	0.760	1.0
BW	0.758	1.0
DG	0.742	1.0
CL	0.745	1.0

Table 8
Coverage potential for each strategy on D-SWIM-1500. The benchmark's potential is 0.99.

	G-blind	H-blind	G-greedy	H-greedy
PR	0.888	0.558	1.0	0.992
BW	0.688	0.451	1.0	0.993
DG	0.607	0.288	1.0	0.994
CL	0.180	0.094	1.0	0.993

7.4. Coverage potential

To complete our study, we investigate the *coverage potential* of the first 10% of nodes promoted to delegates according to each strategy. To this end, we measure, for

each delegate, the ratio of non-delegates neighbors on the social graph (i.e., the number of non-delegates neighbors of delegate i over the total number of neighbors in the social graph). We average then the result over the set of all delegates chosen by the corresponding strategy. Intuitively, the bigger this value, the larger the coverage potential of the strategy, and vice versa. In Tables 6–8 we present the results for every strategy/social attribute for respectively the Dartmouth, Taxi and the D-SWIM-1500 trace. Because of space constraints we omit the tables related to SWIM-500 and A-SWIM-1500.

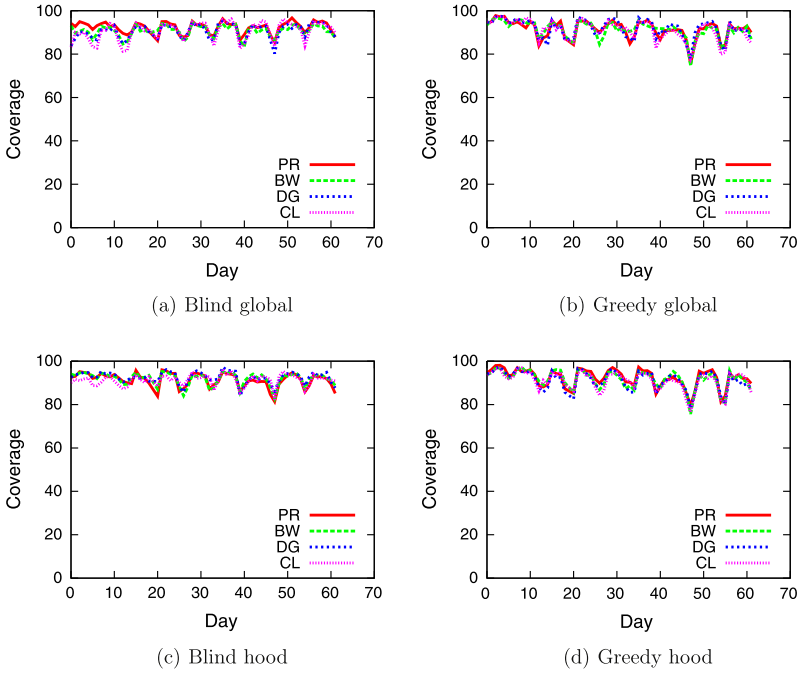
Note that in the *global blind* selection, page-rank is the one with the highest value, followed by betweenness, degree, and finally closeness. This again supports the results of Fig. 3(a). The potential falls drastically when considering the *hood blind* selection (second column of Table 6): delegates are forced to be in the same community, in a blind way. Thus, with high probability, they are socially connected with each other. However, PageRank remains the attribute with the highest value, supporting the results of Fig. 3(b).

The *global greedy* selection naturally yields the highest coverage potential for every attribute: after each node promotion, its neighbors are eliminated from the graph; thus, the ratio of non-delegate neighbors of a node is 1. The *hood greedy* selection (fourth column of Tables 6–8) leads to smaller values. This is because selection is done on a community basis and *only* community neighbors are eliminated after promotion. Since communities are not totally distinct, it might happen that two VIP neighbors in the social graph belong to different communities and, consequently, are eliminated after the promotion, decreasing thus the coverage potential of the strategy. This effect is smaller for high betweenness nodes: they tend to belong to the same group of communities (the ones that they connect). Closeness/degree attributes suffer less from this effect, as they select nodes that are central to communities. Finally, PageRank is the one that suffers most: high PageRank nodes are well distributed within the community to which they belong (being communities well-knit). Thus, they are more likely to have high PageRank neighbors belonging to other communities (that the hood greedy selection does not eliminate).

It is worth to note that the coverage potential just gives a hint on the real coverage power of a method: It does not affect the real ability of the selection method/attribute in covering the network. Indeed, for all traces (see Tables 6–8) the coverage power of the benchmark in all traces is less than all the values related to the global greedy selection. Regardless of the coverage potential, the benchmark performs better with respect to every strategy. In addition, the results with 90% coverage presented in Tables 2–5, confirm page ranks's high performance ability when combined with every strategy.

7.5. Coverage stability

Finally, we investigate the stability of coverage of our strategies in time. We focus on the delegates set needed to reach, in average, 90% coverage for each strategy on all traces. In Figs. 11 and 12, we plot the coverage per day. Due to the lack of space, we only present results related to the Dartmouth and D-SWIM-1500 data sets. We stress however that the results are similar also for the omitted traces. We observe that coverage is quite constant in time for every strategy. This reinforces our intuition on both the monitoring period and the way the social graph is generated. With minimal information on the scenario and a very



(a) Blind global (b) Greedy global

(c) Blind hood (d) Greedy hood

Fig. 11. Coverage stability in time for all strategies (Dartmouth dataset).

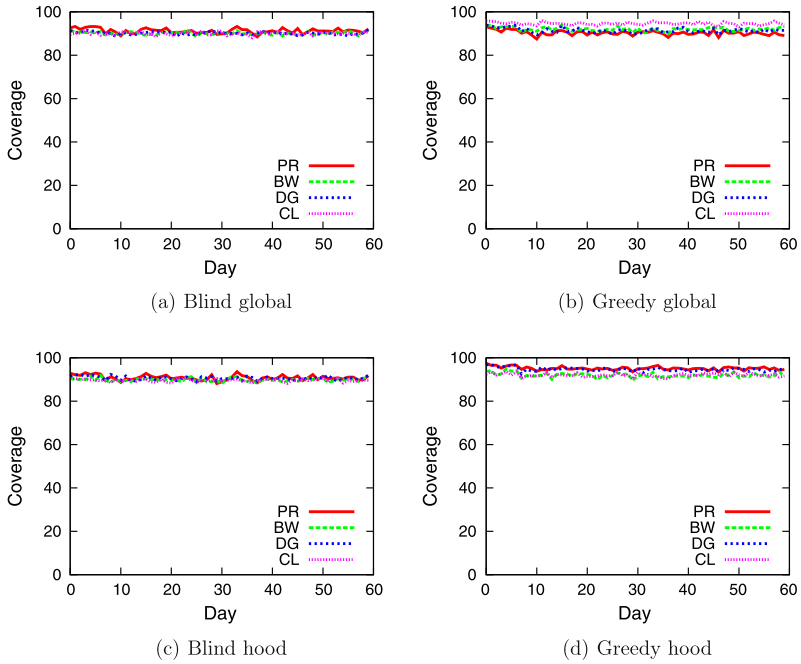


Fig. 12. Coverage stability in time for all strategies (D-SWIM-1500 dataset).

short observation of the network, our strategies are able to compute VIP sets that are small, efficient, and stable in time.

7.6. Coverage intervals vs. VIPs

The VIPs selected by our strategies are expected to cover all nodes every day of the data-trace by carrying data traffic from/to the users. All our target applications (e.g., urban-sensing related data, software updates) are delay-tolerant and would not suffer from the 1-day latency of the daily coverage of VIPs. But what happens for applications that require coverage intervals different from 1-day? How does the length of the coverage interval impact the selection of VIPs? Clearly, if the coverage interval is longer, the 1-day coverage VIPs are a superset of the required number of delegates. Indeed, for a coverage interval long e.g. 2 days, the 1-day coverage VIPs would perform as good as in the 2-day coverage case. However, if the coverage interval required is smaller, the VIP set required to cover the network is likely to change. To quantify such change we have also studied the half-day network coverage for all the traces. For both the Taxi and the A-SWIM-1500 traces we noted absolutely no difference from the 1-day coverage case. We believe this is due to the high mixing and speed of movement of cabs in the Taxi case, and due to the high node density in the A-SWIM-1500 trace. Recall that A-SWIM-1500 is obtained scaling SWIM-500 with constant area.

In Dartmouth, SWIM-500 and D-SWIM-1500 we observed a growth in the VIPs number required to assure 90% of network coverage (due to lack of space here we only show results related to the largest trace: D-SWIM-1500. However, we stress that both the Dartmouth and SWIM-500 traces yield similar results). Intuitively, this is because the meeting patterns of the first half of the day are different from those of the second half. The growth on the number of delegates required to cover 90% of the network in the half-day coverage case is also reflected in the benchmark's VIPs, which are almost doubled with respect to the 1-day coverage interval case (see Table 9). Hence, one would expect that the same should happen also with the VIP sets selected by our strategies. However, from the comparison of the 1-day coverage interval results of Table 4 with the half-day coverage interval results of Table 9 we note that the VIP sets have increased of about 60%. This again means that selecting VIPs according to their "importance" in the network is a good strategy: Indeed, most of the important people during the morning remain so also during evening. However, the coverage interval length indeed does impact the cardinality of VIPs. This suggests for application developers or network infrastructure builders to trade off between data transfer frequency and number of VIPs.

Table 9
VIP sets cardinality to get 90% coverage on D-SWIM-1500. The benchmark approach needs 13%. Half-day coverage interval.

	G-blind (%)	H-blind (%)	G-greedy (%)	H-greedy (%)
BW	30	25.53	15	15.63
CL	70	56.33	15	15.69
DG	34	28.66	15	15.65
PR	18	17.46	15	15.6

To conclude, Figs. 13 and 14 show respectively, the coverage trend, and the performance of the different selection strategies for the case of half-day interval coverage. Again

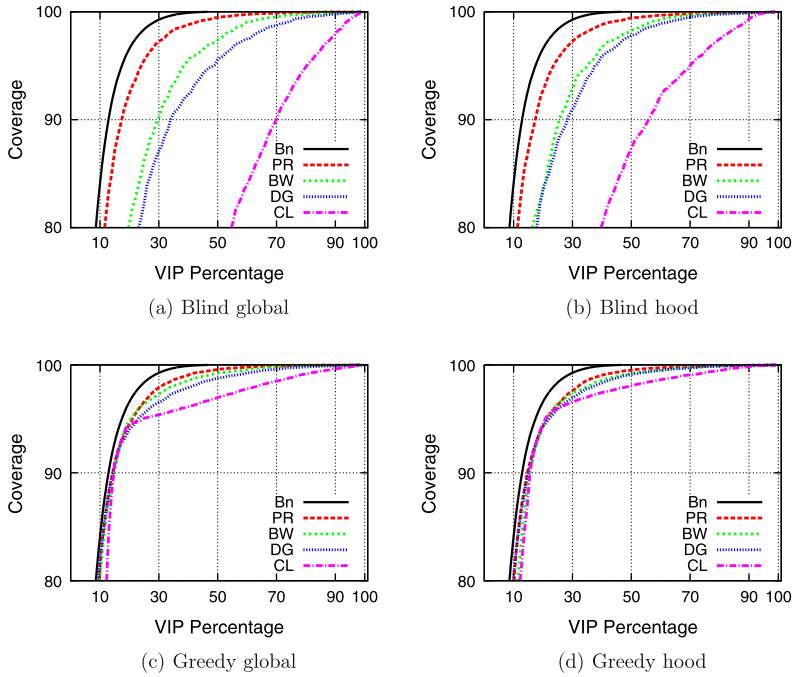


Fig. 13. Performance of the selection strategies on the D-SWIM-1500 dataset. "Bn" refers to the benchmark, "PR" to the page-rank, "BW" to betweenness centrality, "DG" to degree centrality, and "CL" to closeness centrality. Half-day coverage interval.

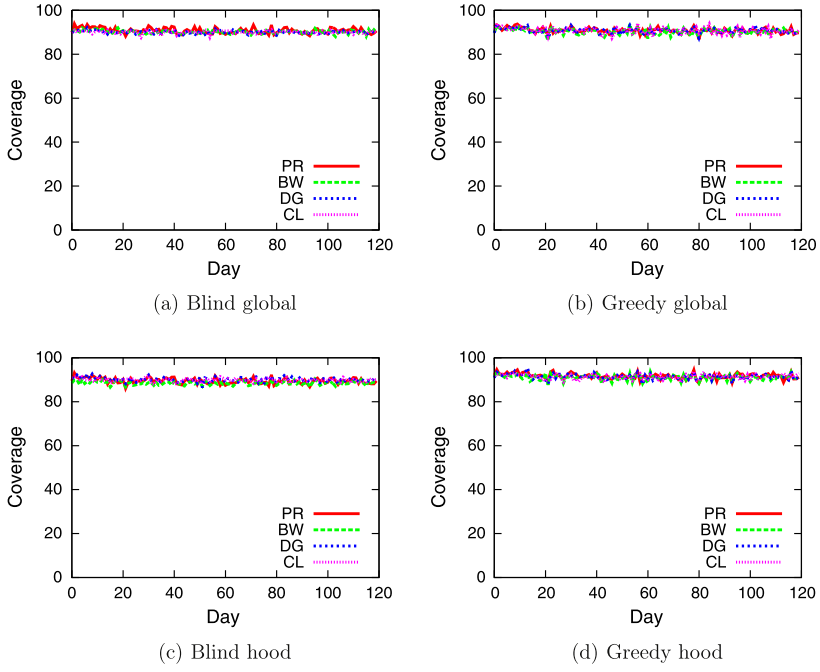


Fig. 14. Coverage stability in time for all strategies (D-SWIM-1500 dataset). Half-day coverage interval.

we note that page-rank wins over the other centralities, and the hood selection strategy wins over the global selection one. Finally, Fig. 14 confirms the stability in time of our VIPs, regardless of the length of the coverage interval.

8. Incentivizing VIPs, data transfer and monitoring period

The VIP selection strategies that we presented in this work achieve very good results in terms of network

coverage. Nonetheless, there are aspects of the VIP approach that we deem important to discuss, starting with user cooperation: The human nature is undoubtedly inherently selfish. So, it is very likely that no user, without being stimulated somehow, would accept the promotion to VIP. Luckily, the number of VIPs selected by our strategies to guarantee 90% coverage is quite low (8% in SWIM, 5.93% in Dartmouth, less than 1% in Taxi). In view of this, it becomes reasonable and convenient, from a network operator point of view, to either upgrade the devices of VIP user to more fancy, recent ones, or to actually pay VIP nodes for their service. Considering the amount of funding that Governments worldwide are putting into global-sensing research [37–39], these incentives become feasible. Another possibility involves considering users' traffic load at the delegates selection and use it to establish a maximum load threshold per delegate. Accordingly, combine it with the social attributes for delegate selection considering fairness and resource constraint among delegates.

Another issue to be considered is that network operators cannot handle the data-traffic from/to VIPs in the classic way (whenever VIPs like), as it would not be of any benefit to offloading. A potential solution to this issue is to make VIPs use the cellular network for the data transfer in different moments of the day. This way, the network load would result distributed in time rather than concentrated in highly congested hours. Another possibility is to transfer the data through wired networks, whenever a VIPs device gets connected to a broadband network during the day. After all, if VIPs are being paid to perform such task, this becomes a reasonable assumption.

A further aspect to be taken into account is the following: "How to handle the traffic of 10% of network users that remain uncovered by our delegates?". As discussed above, the coverage of 90% of nodes requires the promotion as delegates of very few and constant in time network members. This confirms the advantage of our opportunistic delegation approach for covering a high percentage of nodes in a daily basis. Additionally, we claim that the impact of the few 10% non-covered nodes on the cellular network will be small and typically generated by nodes that are marginal to the network (e.g., people frequenting peripheral areas of a city). Usually, nodes having a high activity or mostly visiting central areas in the network will be represented in the constructed social graph, stressing their frequent encounters. In this way, we believe that such more active nodes will be mostly responsible for the traffic overloading previously mentioned and will be covered by the selected delegates, with high probability. Therefore, to answer the previous consideration, the few remaining uncovered nodes could directly transfer their data using cellular networks, at the end of the day, once no delegate visit was detected.

Additionally, an important question to be asked is how does the monitoring period impact the VIP performance. As we already discussed, we believe that the monitoring period length should be defined on a week-base—the week is intuitively the smallest amount of time that regularly generates recurrent patterns in our lives. That said, it is important to point out that the length of the monitoring period might change from scenario to scenario. For the scenarios considered in this work, a 1-week long monitoring period turned out to work well. But, we acknowledge that for other scenarios might either be sufficient shorter monitoring periods, or might require longer monitoring periods. Nonetheless, the important take-away is that, social-related solutions for networking can safely rely on short observations of the social-properties of the society, as the people do preserve their movement patters in time.

That said, we acknowledge that if sudden (but permanent) changes happen in the network, e.g. students leaving the campus for summer holidays, the VIPs selected a priory probably will lose their efficiency. In these cases, the monitoring period should start again to let the VIP sets adapt to the new network conditions. Differently, we believe that this is not true for short-lasting (say, one day) sudden changes in the network, like, e.g., public holidays. Intuitively, the behavior of the people during a public holiday in the middle of the week is likely to be very similar to their behavior during weekends: stay with family, see friends, and so on. Because the monitoring period includes also weekends, it is thus very likely that the VIPs selected a priory perform as good as during any weekend days in our traces. Nonetheless, we believe that, introducing more complex and adaptive techniques that monitor recurrently the network dynamics, like for example machine learning mechanisms, could impact positively the performance of VIP sets. It would allow a smoother and faster adaptation of the VIPs to the changing conditions of the network itself, and to the new opportunities of communication that it brings. Such adaptive learning approach will be considered in a future work.

9. Conclusions

Dense metropolitan areas are suffering network overloading due to the data-traffic generated by the proliferation of smartphone devices. In this paper, we describe VIP delegation, a mechanism to alleviate such traffic based on opportunistic contacts. Our solution relies on the upgrade of a small, crucial set of VIP nodes that regularly visit network users and collect (disseminate) data to them on behalf of the network infrastructure.

VIPs are defined according to well known social network attributes (betweenness, closeness, degree centrality and page-rank), and are selected according to two methods: global (network-based) and hood (community-based) selection. Our observations reveals that 1 week of monitoring period is enough to characterize the tightness of the social links in the network graph. Hence, all methods rely on this network monitoring period and select VIP sets that result small, efficient, and stable in time. Extensive experiments with several real and synthetic data-sets show the effectiveness of our methods in offloading: VIP sets of about 7% and 1% of network nodes in respectively campus-like and vehicular mobility scenarios are enough to guarantee about 90% of network offload. Additionally, the performance of the VIPs selected by our methods is very close to an optimal benchmark VIPs set computed from the full knowledge of the system (i.e., based on past, present, and future contacts among nodes).

References

[1] Canalys, 2010. <http://www.canalys.com/pr/2010/r2010081.html>.
[2] N. Eagle, A. Pentland, Social serendipity: mobilizing social software, IEEE Pervasive Comput. 4 (2) (2005) 28–34.
[3] S. Gaonkar, J. Li, R.R. Choudhury, L. Cox, A. Schmidt, Micro-blog: sharing and querying content through mobile phones and social participation, in: ACM MobiSys, 2008.
[4] E. Miluzzo, N.D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S.B. Eisenman, X. Zheng, A.T. Campbell, Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application, in: ACM SenSys, 2008.
[5] S.B. Eisenman, N.D. Lane, E. Miluzzo, R.A. Peterson, G. Ahn, A.T. Campbell, Metrosense project: people-centric sensisng at scale, in: ACM SenSys, 2006.
[6] S. Ioannidis, A. Chaintreau, L. Massoulie, Optimal and scalable distribution of content updates over a mobile social network, in: IEEE Infocom, 2009.
[7] Customers Angered as iPhones Overload AT&T, New York Times, September 2009. <http://www.nytimes.com/2009/09/03/technology/companies/03att.html>.
[8] iPhone Overload: Dutch T-Mobile Issues Refund after 3G Issues, Ars Technica, July 2010. <http://arstechnica.com/tech-policy/news/2010/06/dutch-t-mobile-gives-some-cash-back-because-of-3g-issues.ars>.
[9] M.V. Barbera, J. Stefa, A.C. Viana, M.D. de Amorim, M. Boc, Vip delegation: enabling vips to offload data in wireless social mobile networks, in: IEEE DCOSS, 2011.
[10] V. Chandrasekhar, J. Andrews, A. Gatherer, Femtocell networks: a survey, IEEE Commun. Mag. 46 (9) (2008) 59–67.
[11] AT&T, Verizon Wireless Join Wi-Fi Interoperability Group, June 2010. <http://news.cnet.com/8301-30686_3-20008476-266.html>.
[12] A. Balasubramanian, R. Mahajan, A. Venkataramani, Augmenting mobile 3g using wifi, in: ACM MobiSys, 2010.
[13] K. Lee, I. Rhee, J. Lee, Y. Yi, S. Chong, Mobile data offloading: how much can wifi deliver?, in: ACM SIGCOMM 2010, 2010.
[14] B.K. Polat, P. Sachdeva, M.H. Ammar, E.W. Zegura, Message ferries as generalized dominating sets in intermittently connected mobile networks, in: ACM MobiOpp, 2010.
[15] B. Han, P. Hui, V.S.A. Kumar, V.M. Marathe, G. Peig, A. Srinivasan, Cellular traffic offloading through opportunistic communications: a case study, in: ACM CHANTS, 2010.
[16] B. Han, P. Hui, V.S.A. Kumar, V.M. Marathe, J. Shao, A. Srinivasan, Mobile data offloading through opportunistic communications and social participation, IEEE Trans. Mob. Comput. (2012).
[17] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, M.D. de Amorim, Push-and-track: saving infrastructure bandwidth through opportunistic forwarding, Pervasive Mob. Comput. (2012).
[18] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[19] P. Hui, People are the Network: Experimental Design and Evaluation of Social-based Forwarding Algorithms, Ph.D. Thesis, UCAM-CL-TR-713. University of Cambridge, Computer Laboratory, 2008.
[20] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (2008) 779–782.
[21] C. Boldrini, M. Conti, A. Passarella, The sociable traveller: human traveling patterns in social-based mobility, in: ACM MobiWac, 2009.
[22] V. Kann, On the Approximability of NP-Complete Optimization Problems, Ph.D. Thesis, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, 1992.
[23] T. Hossmann, T. Spyropoulos, F. Legendre, Know thy neighbor: towards optimal mapping of contacts to social graphs for dtn routing, in: IEEE INFOCOM'10, 2010.
[24] P. Hui, J. Crowcroft, E. Yoneki, BUBBLE Rap: social-based forwarding in delay tolerant networks, in: ACM MobiHoc, 2008.
[25] A. Mei, J. Stefa, Give2Get: forwarding in social mobile wireless networks of selfish individuals, in: IEEE ICDCS, 2010.
[26] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.
[27] A. Mei, J. Stefa, Give2Get: forwarding in social mobile wireless networks of selfish individuals, in: IEEE Transactions on Dependable and Secure Computing, 2012.
[28] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Netw. 1 (3) (1979) 215–239.
[29] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1998) 107–117.
[30] T. Henderson, D. Kotz, I. Abyzov, J. Yeo, CRAWDAD trace dartmouth/campus/movement/01_04 (v. 2005–03–08). <http://crawdad.cs.dartmouth.edu/dartmouth/campus/movement/01_04>.
[31] M. Piorkowski, N.S.-Djukic, M. Grossglauser, A parsimonious model of mobile partitioned networks with clustering, in: COMSNETS, 2009.
[32] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on the design of opportunistic forwarding algorithms, IEEE Trans. Mob. Comput. 6 (6) (2007) 600–620.
[33] A. Mei, J. Stefa, SWIM: a simple model to generate small mobile worlds, in: IEEE Infocom, 2009.
[34] S. Kosta, A. Mei, J. Stefa, Small world in motion (SWIM): modeling communities in ad-hoc mobile networking, in: IEEE SECON, 2010.
[35] S. Kosta, A. Mei, J. Stefa, Large-scale social mobile synthetic networks with SWIM, IEEE Trans. Mob. Comput. (2014).
[36] J. Leguay, A. Lindgren, J. Scott, T. Riedman, J. Crowcroft, P. Hui, CRAWDAD trace upmc/content/imote/cambridge (v. 2006–11–17). <http://crawdad.cs.dartmouth.edu/upmc/content/imote/cambridge>.
[37] CENS Urban Sensing. <http://urban.cens.ucla.edu/projects>.
[38] MIT Senseable City Lab. <http://senseable.mit.edu>.
[39] EARSeL. <http://www.earsel.org/?target=SIGs>.

Marco Valerio Barbera is a PhD student at the Computer Science Department of Sapienza University of Rome, Italy. He received the Laurea degree in Computer Science, summa cum laude in 2009. From March 2011 to December 2011 he was a visiting fellow at the Network Security Lab of the Columbia University, NY, USA. His research interests include mobile cloud computing, distributed systems, and analysis and modeling of social mobile wireless networks.

Aline Carneiro Viana is a Senior Researcher at INRIA Saclay. Dr. Viana received her MSc degree in Electrical Engineering in 2001 from the Federal University of Goi·s, after spending one year at the Polytechnic School of the Federal University of Rio de Janeiro, Brazil. She got her PhD in Computer Science from the University Pierre et Marie Curie in 2005. After holding a postdoctoral position at IRISA/INRIA Rennes, she joined INRIA Saclay in 2006. She was an Invited Researcher at the TKN Group of the Technical University of Berlin in 2010. Her research is primarily in data management and routing in wireless self-organized networks.

Marcelo Dias de Amorim is a CNRS permanent researcher at the computer science laboratory (LIP6) of UPMC Sorbonne Universites, France. His research interests focus on the design and evaluation of dynamic networks as well as service-oriented architectures. For more information, visit http://www-npa.lip6.fr/~amorim.

Julinda Stefa is a Post-Doc at the Computer Science Department of Sapienza University of Rome, Italy. She received the Laurea degree in Computer Science, summa cum laude, and the PhD in Computer Science from Sapienza University of Rome respectively in July 2006 and February 2010. In 2005 she joined Google Zurich for 3 months as an engineering intern. She was a visiting scholar at the CS Dept. of UNC-Chapel Hill, USA, from From November 2008 to April 2009, and a Research Intern at Microsoft Research, Cambridge, UK, from January to April in 2011. Her research interests include computer systems and network security, parallel and distributed systems, and analysis and modeling of social mobile wireless networks. She has published in some of the topmost conferences and journals like IEEE INFOCOM, ACM MobiHoc, IEEE ICDCS, and IEEE Transaction on Computers, and is involved as reviewer and in technical program committees of several conferences and workshops in the field. She is winner of a research grant from Working Capital PNI and offered by Telecom Italia (30 winners out of 2138).