



**HAL**  
open science

## Opportunistic data collection through delegation

Greg Bigwood, Aline Carneiro Viana, Mathias Boc, Amorin Marcelo Dias De

► **To cite this version:**

Greg Bigwood, Aline Carneiro Viana, Mathias Boc, Amorin Marcelo Dias De. Opportunistic data collection through delegation. [Research Report] RR-7361, 2010, pp.27. inria-00508273v1

**HAL Id: inria-00508273**

**<https://inria.hal.science/inria-00508273v1>**

Submitted on 2 Aug 2010 (v1), last revised 9 Mar 2011 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

## *Opportunistic data collection through delegation*

Greg Bigwood — Aline Carneiro Viana — Mathias Boc — Marcelo Dias de Amorim

**N° 7361**

August 2010

Thème COM

*R*apport  
de recherche



## Opportunistic data collection through delegation

Greg Bigwood , Aline Carneiro Viana , Mathias Boc , Marcelo Dias de Amorim

Thème COM — Systèmes communicants  
Équipes-Projets Asap

Rapport de recherche n° 7361 — August 2010 — 23 pages

### Abstract:

We consider a collection system where collectors move around gathering information generated by data producers. In such a system, data may remain uncollected when the number of collectors is insufficient to cover the whole population of producers. Motivated by the observation that node encounters are sufficient to build a connected relationship graph, we propose to take advantage of the inherent interactions among nodes and transform some producers into *delegates*. With such an approach, collectors only need to meet delegates that, in turn, are responsible for gathering data from a subset of standard producers. We achieve this goal through two contributions. First, we investigate several *delegation strategies* based on the relative importance of nodes in their social interactions (i.e., the node centrality). Second, by considering a *prediction* strategy that estimates the likelihood of two nodes meeting each other, we investigate how the delegation strategies perform on predicted traces. We evaluate the delegation strategies both in terms of coverage and size of the delegation existing real mobility data sets. We observe that delegation strategies that rely on localized information perform as well as the ones that consider a complete view of the topology.

**Key-words:** collaborative data collection; centrality; sparse networks; social interactions

## Opportunistic data collection through delegation

**Résumé :** Nous considérons un système de collecte où les collectionneurs se déplacent et collectent les informations générées par les producteurs de données. Dans un tel système, les données peuvent ne pas être collectées lorsque le nombre de collectionneurs est insuffisant pour couvrir l'ensemble de la population des producteurs. Motivé par le fait que les rencontres de nIJud sont suffisantes pour construire un graphe connecté, nous proposons de profiter des interactions inhérentes entre les nIJud et transformer certains producteurs en *délégués*. Avec une telle approche, les collectionneurs ont seulement besoin de rencontrer les délégués que, à leur tour, sont responsables de la collecte de données d'un sous-ensemble des producteurs. Nous atteignons cet objectif grâce à deux contributions. Tout d'abord, nous étudions *plusieurs stratégies de délégation* basée sur l'importance relative des nIJud dans leurs interactions sociales (par exemple, la centralité du nIJud). Deuxièmement, en considérant une *stratégie de prédiction* qui donne les estimations de la probabilité d'une rencontre de deux nIJud, nous étudions les stratégies de délégation avec les traces prédit. Nous évaluons les stratégies de délégation à la fois en termes de couverture et de la taille du groupe de délégation en utilisant des traces de mobilité réelles. Nous n'observons que les stratégies de délégation qui se basent sur des informations localisées fournis aussi des bons résultats comparés aux résultats considérant une vue complète de la topologie.

**Mots-clés :** collecte collaborative de données, centralité, les réseaux de faible densité, les interactions sociales

## 1 Context and motivation

We consider a class of self-organizing dynamic networks where a number of mobile *producers* generate data that must be transferred to a collection system composed of specialized nodes called *collectors*. Collectors move around and collect data when they enter within communication range with producers. The problem of deciding which collectors “visit” which producers is by itself a difficult issue. But the problem becomes even more challenging when the number of producers is much larger than the number of collectors. In such a case, we need alternative solutions to overcome the insufficiency of the collection system.

To the best of our knowledge, this is the first time that this specific problem is addressed. In fact, a number of works in the literature also propose to benefit from mobility to perform some task, but under a different perspective [2, 11, 14, 26]. In particular, data mules [14] and message ferries [26] are mobility-assisted strategies that aim at providing “connectivity opportunities” among nodes. These approaches are concerned with controlling the trajectories of ferries or mules to meet static or mobile nodes with the goal of minimizing message drops. Trajectories are adjusted based on requests from nodes, or the nodes adjust themselves their trajectories to meet the ferries. Although these approaches solve only partially the problem we focus on, as we consider the problem of determining *which* subset of nodes should be visited, instead of trying to compute the best trajectory or positioning for collectors.

We propose to take advantage of the inherent mobility of producers and transform some of them into *delegates* of the collection system. Delegates are responsible to collect data from other producers and serve as intermediate relays between them and the collectors (cf., Fig. 1). This is motivated by the fact that social networks exhibit the small world phenomenon which comes from the observation that individuals are often linked by a short chain of acquaintances [21] and that encounters are sufficient to build a connected relationship graph [12]. The problem here is to determine which producers should be promoted as delegates. To this end, we make the following contributions:

- Because contacts are not deterministic, we propose a *prediction* strategy to estimate the likelihood of two producers meeting each other (cf. Section 3).
- Based on such a prediction system, we investigate different social-inspired strategies to select which producers should become delegates (cf., Section 4).

In all of the social-based strategies, the selection of delegates is based on the quantification of the relative importance of a producer in the network. Consider the example of data collection in a city, it consists of investigating how important a person is within a social network. To this end, we study two types of centrality approaches, namely degree and betweenness, combined with different network knowledge, namely global and ego networks (cf., Section 4). The analysis are performed on two different data sets available in the literature (Dartmouth College wireless traces [17] and San Francisco taxi traces [22]). By using measured and predicted traces generated from these data sets, we investigate the properties of the social-based strategies in terms of coverage and size of the resulting delegation set. Finally, we make recommendations for the use of specific social-based strategies and show that the use of predicted traces is effective for the selection of good delegates (cf., Section 5).

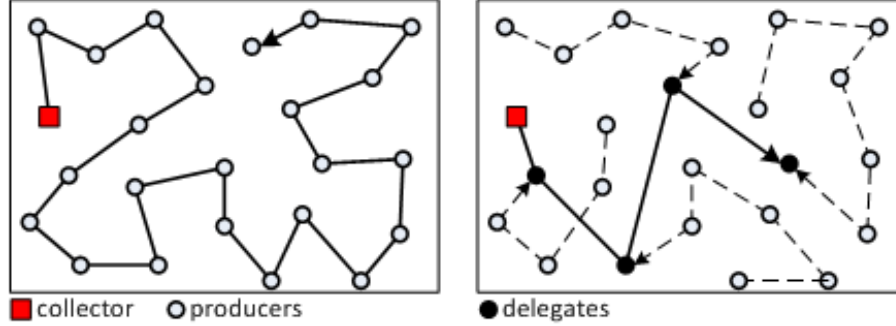


Figure 1: In the leftmost picture, the collector has to visit all producers. In the second illustration, delegates gather information on behalf of the collector.

## 2 Problem definition

**Data collection.** In intermittently connected mobile networks, high degree of end-user mobility and unreliable links raise different communication challenges since forwarding paths may be unstable and device reachability may be highly variable. In this paper, we focus on the problem of *collaborative data collection*, where a number of *producers* generate data that must be transferred to a collection system composed of *collectors*. Contrary to approaches found in the literature that use randomly moving humans and animals as data mules to opportunistically collect or disseminate data [2, 11, 14], we benefit from both the controllable mobility of collector nodes and the inherent social interactions among producers to opportunistically perform data collection in a collaborative fashion. This is orthogonal to other approaches such as message ferrying where adaptation would be performed at the producer level [26].

**Problem formulation.** Let  $\Pi = \{p_1, p_2, \dots, p_P\}$  be the set of producers in the network. We define  $\mathbf{N}_{p_i}$  as the set of producers that  $p_i$  encounters at least once during a certain time slot (to be defined in Section 3). Our problem consists in computing a subset  $\Delta \subset \Pi$  of producers such that:

$$\begin{cases} |\Delta| \leq |\Delta'|, \forall \Delta' \subset \Pi, \\ \Delta \cup \mathbf{N}_\Delta = \Pi, \end{cases} \quad (1)$$

where  $\mathbf{N}_\Delta = \bigcup_i \mathbf{N}_{p_i}, \forall p_i \in \Delta$ . Note that this formulation can be seen as a reduction problem, where collectors have to visit a smaller number of nodes and visit the whole population of producers.

Note that the formulation above reflects the ideal case, where we can determine in advance the exact set of delegates that will cover all producers. In practice, we cannot rely on such an assumption. Indeed, since our system is causal, we cannot know in advance how encounters will take place. We have two alternatives to solve this problem. Either we promote all producers as delegates, which would bring our system to the original problem (i.e., the collectors visit all producers), or we try to predict future encounters. Once these encounters predicted, we apply a social-inspired selection scheme to determine the proper set of delegates (cf., Section 4).

Finally, in order to guarantee the traceability of the message passing scheme, we focus in this paper on *single delegation* schemes, where a producer transmits its data to either a collector or a single delegate.

### 3 Sketch of the solution

In this section, we present a high-level view of our approach. By applying different social-inspired strategies, we investigate the relative importance of producers in their social interactions or in the global network and select a set of potential good relay nodes (i.e., delegates) to help collectors perform their task. In this scheme, data is first collected from producers by delegates and then sent by these latter to the collectors.

There are many ways to predict future contacts based on the history of encounters [8, 19]. In this paper, we propose to use a slotted prediction strategy with a history window of two days (see explanation below). We will see in Section 5 that, although simple, this strategy leads to good results.

Let  $C(i)$  be the *collection period*  $i$  of duration  $|C|$ . We divide this period into  $J$  slots of fixed duration so that  $C(i) = \{c(i, 1), c(i, 2), \dots, c(i, J)\}$ . To each  $c(i, j)$  we associate two matrixes  $M(i, j) = [N \times N]$  and  $P(i, j) = [N \times N]$ , where  $N$  is the total number of producers.

The first matrix  $M(i, j)$  is called the *measurement matrix*. In this matrix, element  $m_{x,y}(i, j)$  is 1 if node  $x$  meets node  $y$  at least once during the time slot  $c(i, j)$ , and 0 otherwise. As stated before, the problem is that we cannot compute the delegates based on the measurement matrix for a given time slot because this matrix is only available at the end of the slot period (after the real contacts are observed).

The second matrix  $P(i, j)$  is the *prediction matrix*. Each element of this matrix is based on the history of observations previously made. We propose to rely on the observations made during the same slot of the previous collection periods. We predict that a contact will happen if it already happened during the same slot of the previous two measurement periods. This procedure is illustrated in Fig. 2. The reason for this value is that we empirically observed that this leads to better results. More formally, we have:

$$p_{x,y}(i, j) = \begin{cases} 1, & \text{if } [m_{x,y}(i-1, j) \wedge m_{x,y}(i-2, j)] = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

## 4 Social-oriented delegation

In the previous section, we proposed a prediction approach to determine the encounters that are likely to happen during a given time slot. We have now to determine, based on these expected encounters, which producers should be promoted to delegates. In the following, we investigate several strategies to this end. We classify them based both on the social metric and on the awareness a producer has about the topology.

### 4.1 Social-inspired metrics

To investigate the relative importance of producers, we exploit the well known centrality concept from graph theory and network analysis. Centrality will allow us designing how important a producer is within a social sphere, in order to identify potential delegates. In a social network, an individual is considered as important if, for instance,



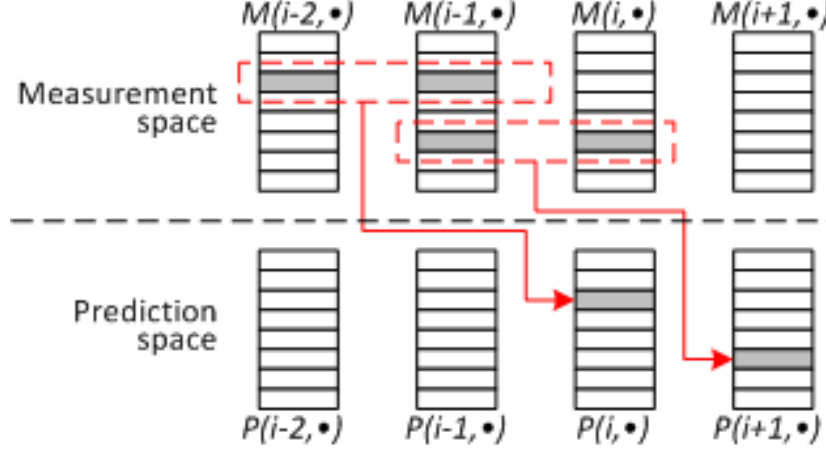


Figure 2: Observed encounters are used to feed the prediction matrix.

she/he has a strong capability of meeting others or connecting disjoint tightly connected groups. In the literature, “importance” is described through three different metrics [9, 10]: closeness, degree, or betweenness. We use the last two centrality metrics in our analysis and a combination between them, described hereafter.

**Degree centrality (DC).** This metric describes the number of direct connections that involve a given node. In this way, a node with high degree centrality can be seen as a popular node. This also means it can be used as a good conduit for information exchange, since it maintains contact with numerous other network nodes. Thus, by exploiting degree centrality we aim to see if node degree can be used to find a near to perfect level of coverage, without having to select large numbers of nodes. The degree centrality of a given producer  $p_i$  is calculated by simply counting the number of contacts this producer has during a time slot with other producers in the network:  $D(p_i) = |N_{p_i}|$ .

**Betweenness centrality (BC).** This metric measures the frequency a node lies on the paths linking other nodes. In this way, a node with a high betweenness centrality is on more paths than the average and, therefore, has the capacity to facilitate interaction between the nodes that it links. The betweenness of producers is thus computed by analyzing all the paths between all nodes in the network, and then scoring a producer based on the amount of times it appears on the paths of other nodes. Note, however, that, nodes with high betweenness do not frequently have a large number of connections, since they are usually the only route to a place and are frequently “bridging nodes”. By harvesting these nodes, we attempt to see if they are enough to cover the rest of the network.

**Betweenness and degree centrality (BDC).** By combining popularity and betweenness, we hope to play off the strengths of both of these metrics, whilst overcoming their shortcomings. A metric utilizing both degree and betweenness centrality seeks to gain the benefit of connecting to a large number of nodes, whilst simultaneously reaping the benefit of selecting nodes that bridge groups of nodes within the network. This would then ensure capturing nodes that bridge distinct cliques of nodes as well as

network locations that would expect to see the most traffic. Considering we are interested in determining a subset of size  $|\Delta|$  of good delegates, the resulting set BDC is half made up of top nodes in the betweenness centrality set BC, and half made up of high popular nodes in the degree centrality set DC.

## 4.2 Topology-awareness

We compute the three metrics presented above to each node by considering nodes have access to the following levels of network knowledge:

**Complete and bounded network (C).** Also referred to as socio-centric network, it requires the complete knowledge of the network topology. A large node population may thus make difficult the analysis of centrality metrics in this kind of network topology. Nevertheless, we use this socio-centric network to get the upper-bounded results to be compared to the ones obtained when using knowledge-limited network topologies.

**Ego-centric networks (E).** This network topology represents the network viewed from the perspective of a single node and can be locally computed without the complete knowledge of the entire network. An ego-centric network consists of a single actor (named *ego*), its 1-hop neighbor (named *alters*), and all the links among those alters. This means the ego node itself can work out its ego network, after exchanging its neighborhood list with each new encounter, which allows distributing computation. An Ego network requires less state than a 2-hop neighborhood topology such as the one required for the computation of dominating sets. It has been shown in the literature that degree and betweenness centrality, when computed in ego networks, allow quite good results when compared to the socio-centric networks [20].

We combine the previous described centrality metrics and types of networks and originate six strategies, which we investigate in the rest of this paper.

## 4.3 Benchmark strategy

For the sake of comparison, we also consider a benchmark solution that leads to the best possible result. More specifically, the ideal case would be the smallest set of delegates that guarantee 100% coverage (i.e., during the collection period, all producers could deliver their data to either a collector or a delegate). This corresponds to computing the minimum dominating set (MDS) on the encounter graph for each time slot. Because this problem is known to be NP-hard, we consider an alternative solution borrowed from the computation of multi-point relays as our benchmark [1].

A dominating set DS of a graph  $G = (V, E)$  is a subset  $V' \subseteq V$  such that every vertex not in  $V'$  (i.e. for all  $u \in V - V'$ ) is adjacent to at least one vertex of  $V'$  by some edge (i.e., there is a  $v \in V'$  for which  $(u, v) \in E$ ). We have used the following greedy algorithm for computing dominating sets, which takes  $O(m^2)$  time for a maximum connectivity degree  $m$  [1]: (i) begin with an empty set; (ii) select the nodes that are the only ones neighbor of some two-hop neighbors of  $i$ , and add them to the DS set; (iii) add in the DS the neighbor node of  $i$  that covers the largest number of two-hop neighbors of  $i$  that are not yet covered by the current DS. Repeat this step until all two-hop neighbors are covered.

The results obtained with the proposed delegation strategies are then compared to the benchmark. Note that, unlike [23], connection among delegates is not required here, since collectors will later visit them. The reason for not adopting the benchmark

Table 1: Acronyms for the different strategies.

<i>Acronym</i>	<i>Strategy</i>
DC-C	Degree centrality with complete view
DC-E	Degree centrality with egocentric view
BC-C	Betweenness centrality with complete view
BC-E	Betweenness centrality with egocentric view
BDC-C	Betw. and degr. centrality with complete view
BDC-E	Betw. and degr. centrality with egocentric view
Benchmark	Dominating set

as the delegation strategy is that, as we will see in Section 5, although leading to better coverage, it requires a large number of delegates to perform the work.

For the sake of clarity, we present in Table 1 the acronyms used in the remainder on this paper.

## 5 Evaluation

We evaluate the performance of the social-inspired delegation strategies described in Section 4 using measured and predicted traces out of two data sets. We first describe the data sets used and the evaluation methodology in Sections 5.1 and 5.2, respectively. Secondly, in Section 5.3, we investigate the performance of the benchmark approach when applied to the different traces. Thirdly, by considering nodes with different network topology awareness (cf., Section 4), we evaluate the performance obtained with each social metric (cf. Section 5.4). Finally, we provide some discussion on the results in Section 5.5.

### 5.1 Data sets and methodology

We decided to use two data sets that are well-known by the research community: Dartmouth College campus and San Francisco Taxi cabs. By nature, they show different mobility patterns, types of users, and environment conditions. We expect then to observe different social behaviors, resulting in different usage models. This diversity will allow us better understanding the characteristics of the analyzed strategies. Although well-known, we briefly describe them in the following.

*Dartmouth College campus* [16]. This data set shows associations and disassociations of wireless devices with 566 wireless access points in the Dartmouth College campus. For the purposes of this work, this data set has to be translated into a contact graph. In the literature, authors generally assume that two nodes are in contact with each other if they are associated with the same access point at the same time [5]. We decided instead to use the geographic coordinates of the nodes and consider that two nodes are in contact if their distance is below 250 meters. To obtain their geographic positions, we adopt the filtering approach proposed by Kim et al. [15]. We consider mobility information over one month and focus our analysis on the 1,000 most active nodes (in terms of presence). Note that nodes are not active at the same time, leading to an average number of 577 active nodes and 6 contacts per node per time slot of one hour, as shown in Figs. 3(a) and 3(c). The corresponding *predicted traces*, generated following the prediction strategy described in Section 3, lead to an average number of 737 active nodes and 4 contacts per node, as shown in Figs. 3(b) and 3(d).

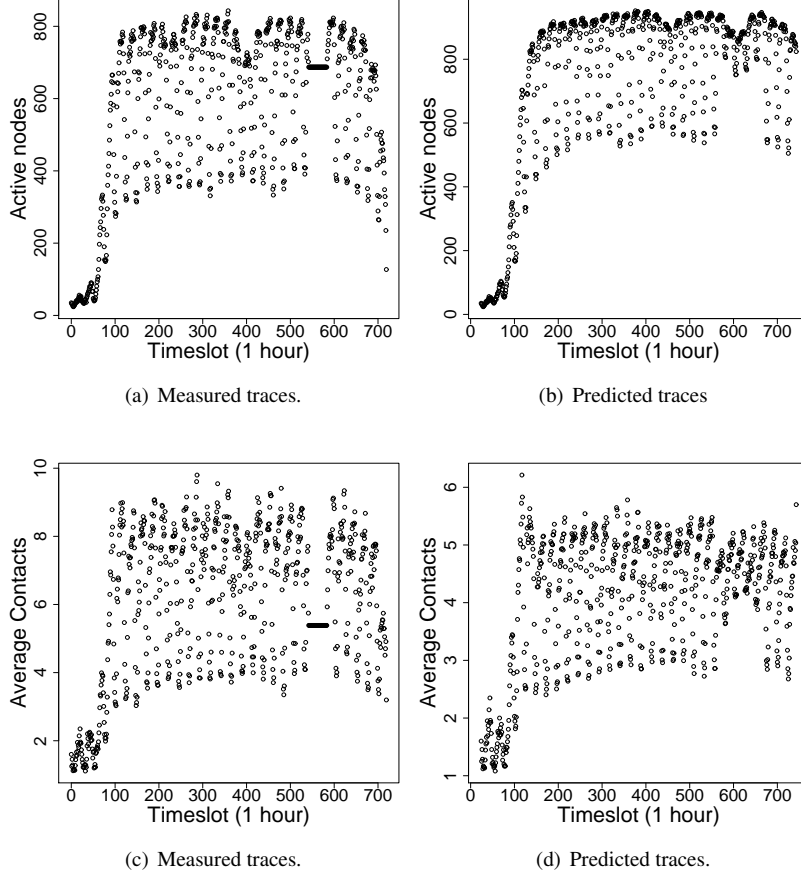


Figure 3: Dartmouth measured and predicted traces: (a)-(b) active nodes and (c)-(d) average number of contacts over 695 hours, using time slots of 1 hour.

*Taxi cabs in San Francisco* [25]. This data set describes the movement of taxicabs in San Francisco. GPS devices installed in the cabs are polled every minute or so whether they are free, inactive, or on route. We consider movements of the entire population of taxicabs that were active during a period of 28 days, resulting in a total of 542 nodes. We have, per time slot, an average number of 506 active nodes and of 130 contacts per node, as shown in Figs. 4(a) and 4(c). As in the Dartmouth trace, we consider that two taxis are in contact if their distance is below 250 meters. The predicted results are shown in Figs. 4(b) and 4(d), with an average number of 515 active nodes and 52 contacts per node.

**Prediction vs. measurement.** Note that for both data sets, the predicted traces resulted in lower average numbers of contacts per node and higher average numbers of active nodes. The reasons are as follows. On the one hand, we predict that a contact will happen at a given time slot if the same contact happened in the equivalent slot in *both* the two previous days. This explains why the expected degree is lower. On the other

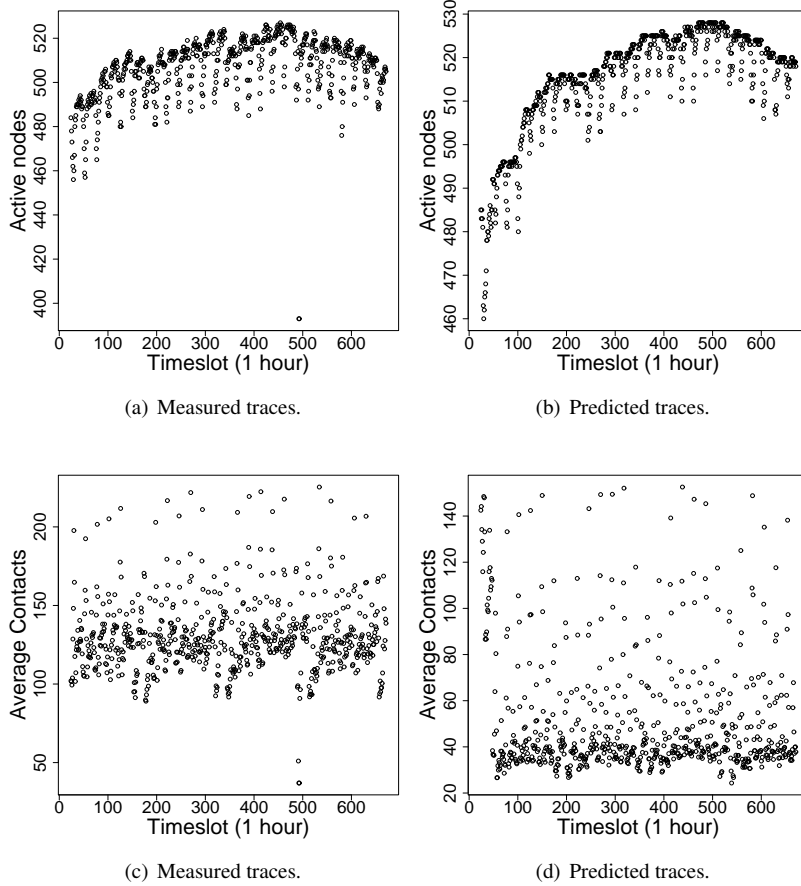


Figure 4: Taxi measured and predicted traces: (a)-(b) active nodes and (c)-(d) average number of contacts over 672 hours, using time slots of 1 hour.

hand, we assume the a node will be active in a given time slot if it was active in *either* of the two previous days. This results in a higher expectation.

## 5.2 Evaluation methodology

We are interested in selecting a limited set of producers  $\Delta$  as delegates to opportunistically collect data generated by producers. In order to evaluate the performance of each delegation strategy, we focus on both coverage and size of the resulting  $\Delta$  sets, on a per-slot basis. We evaluate the coverage property of the strategies by determining the percentage of producers not met by any of the delegates – we refer to this parameter as *missed nodes*. We also evaluate the size of the  $|\Delta|$  set as an indicator of the efficiency of the delegation system.

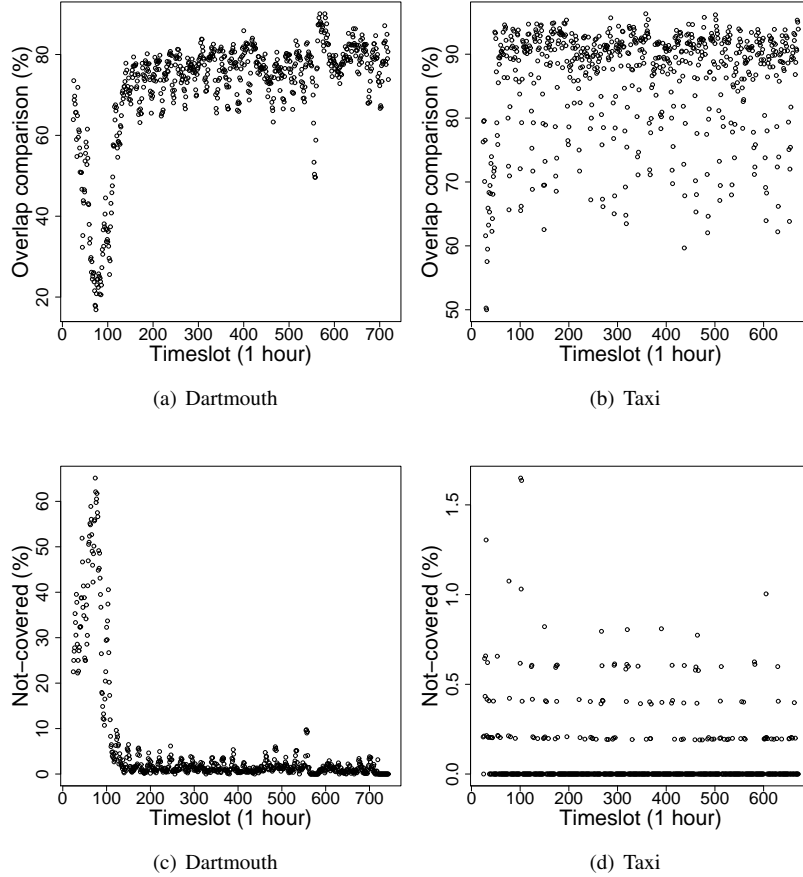


Figure 5: Benchmark approach. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: (a) Dartmouth: 71.47%, (b) Taxi: 86.57%. (c)-(d) Amount of not-covered nodes by the  $\Delta$  sets gotten from the predicted traces: (c) Dartmouth: 5.7% (d) Taxi: 0.08%.

### 5.3 Benchmark analysis

We first evaluate the performance of the benchmark approach when applied to the measured traces and the corresponding predicted traces, of both Dartmouth and Taxi data sets. Table 2 summarizes the average amount of producers that become delegates (in both absolute and relative terms). As previously discussed, we observe that the sizes of the prediction sets are larger than the measured ones (cf., Section 5.1).

In order to evaluate the precision of results obtained with the predicted traces, we investigate (i) the delegates that appear in both measured and predicted traces and (ii) the amount of nodes not covered when we use the predicted delegates. Fig. 5(a) (resp. Fig. 5(b)) shows the overlap comparison for the Dartmouth data set (resp. Taxi data set). An average of 71.47% (resp. 86.57% in the Taxi data set) of delegates found using the measured traces are also present in the corresponding predicted sets. Additionally,

Table 2: Avg. results for the benchmark.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	304.36	56.14
	predicted	508.43	70.97
Taxi	measured	326.06	64.22
	predicted	419.33	81.38

Table 3: Avg. results for DC-C.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	304.36	55.08
	predicted	508.43	70.87
Taxi	measured	325.09	64.02
	predicted	418.59	81.24

Fig. 5(c) shows for the Dartmouth data set (resp. Fig. 5(d) for the Taxi data set) that an average of 5.7% (resp. 0.08%) of nodes are not covered by the predicted delegates. These results are encouraging and confirm that the prediction scheme leads to good performance.

## 5.4 Topology-awareness

We investigate now the performance of the social-inspired metrics for delegate selection according to the topology-awareness of nodes. For every strategy discussed hereafter, we provide: (i) tables summarizing the average percentage of set sizes, for both the measured and predicted traces, (ii) figures showing the overlap between the measured and predicted delegate sets, and (iii) figures showing the amount of uncovered producers if we use the predicted delegates.

### 5.4.1 Influence of complete network view (C)

We consider here that nodes have a complete view of the network at each time slot. Producers are first ordered based on their social influence. In order to select the best delegates, we use as reference the same size of the delegate set obtained for the measured data set in the benchmark approach, and pick the top-rated producers until the delegate sets' sizes are the same.

**DC-C.** Table 3 summarizes the average percentage of missed nodes and the set size resulting from the degree centrality selection applied to the complete network. Fig. 6(a) shows that an average of 71.42% (resp. 86.77% in Fig. 6(b)) of delegates overlapping between the traces, while Fig. 6(c) shows that an average of 9.22% (resp. 0.59% in Fig. 6(d)) of nodes are not be covered predicted delegates.

**BC-C.** As with the previous strategy, Table 4 shows the results of the betweenness centrality selection applied to the complete network. Fig. 7(a) shows that an average of 86.76% (resp. 99.07% in the Taxi data set in Fig. 7(b)) of delegates' overlap. Fig. 7(c) shows that an average of 7.31% (resp. 0.13% in the Taxi data set) of nodes remain uncovered.

**BDC-C.** Table 5 summarizes the results for the BDC-C strategy. As in the previous two cases, we show in Figs. 8(a) and 8(b) the overlaps between the measured and predicted

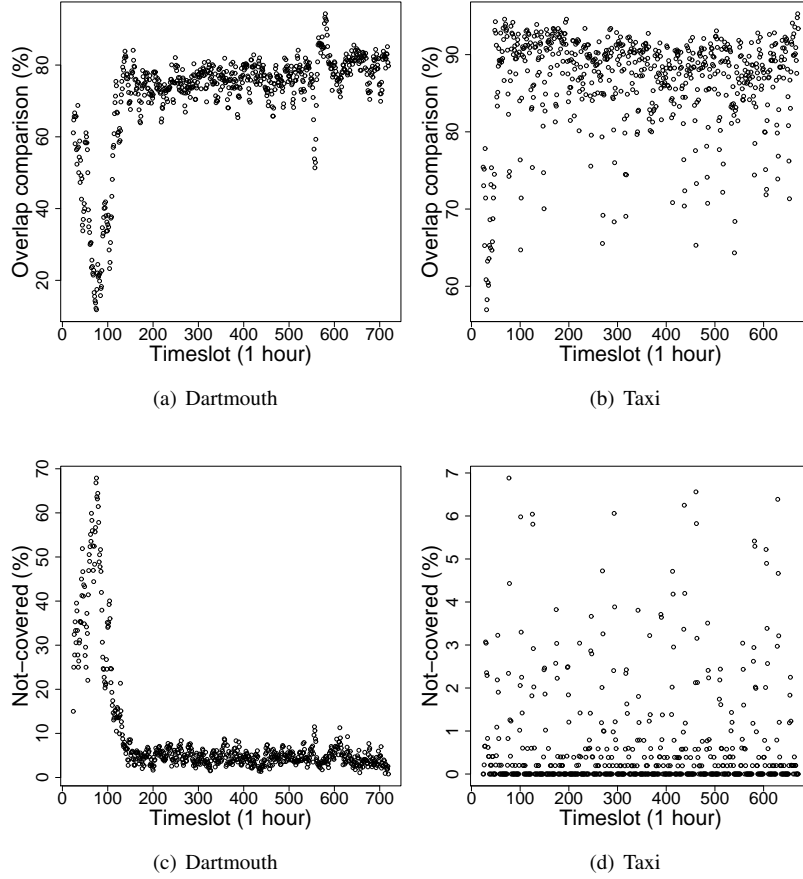


Figure 6: DC-C strategy. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: (a) Dartmouth: 71.42% (b) Taxi: 86.77% . (c)-(d) Amount of not-covered nodes by the  $\Delta$  sets gotten from the predicted traces: (c) Dartmouth: 9.22% (d) Taxi: 0.59%.

Table 4: Avg. results for BC-C.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	304.30	56.12
	predicted	507.54	70.97
Taxi	measured	326.06	64.22
	predicted	419.33	81.38

traces for both data sets. We observe an average overlap of 74.86% for the Dartmouth data set and 90.64% for the Taxi data set. Figs. 8(c) and 8(d) show that the proportion of nodes not covered is 8.04% in the Dartmouth case and 0.23% in the Taxi case.



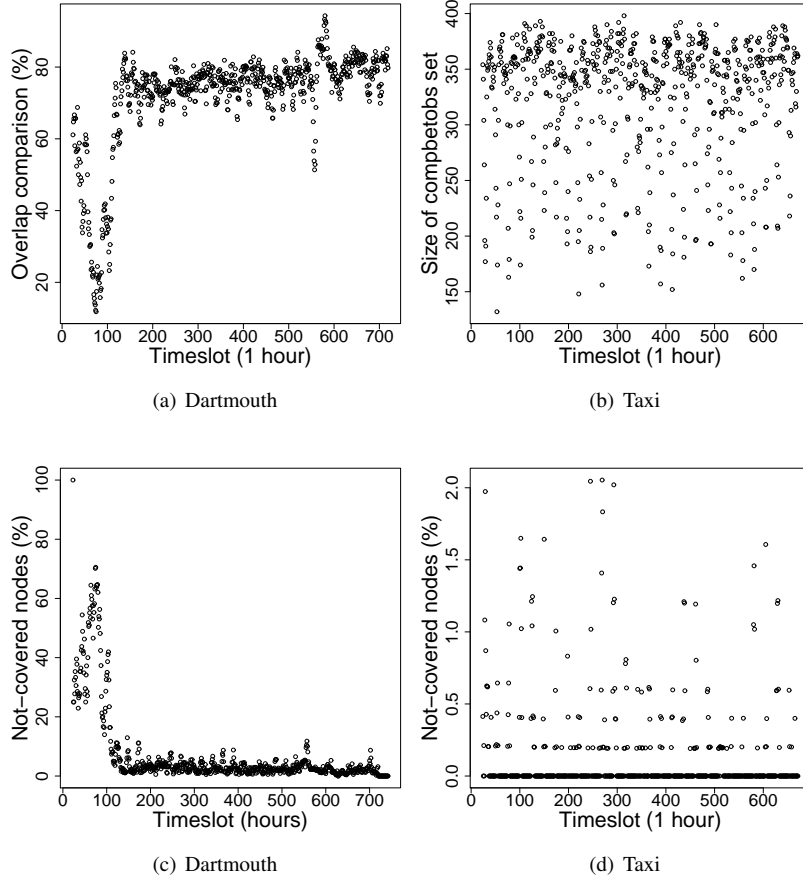


Figure 7: BC-C strategy. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: (a) Dartmouth: 86.76%, (b) Taxi: 99.07%. (c)-(d) Amount of not-covered nodes by the  $\Delta$  sets gotten from the predicted traces: (c) Dartmouth: 7.31% (d) Taxi: 0.13%.

Table 5: Avg. results for BDC-C.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	304.36	56.14
	predicted	507.33	70.87
Taxi	measured	326.06	64.22
	predicted	419.33	81.38

#### 5.4.2 Influence of ego-centric network view (E)

Nodes are now considered to have an ego-centric view of the network. The goal is to verify if nodes can locally select good delegate sets. As in the previous section, the sizes  $|\Delta|$  of the measured sets in the benchmark case are used as reference sizes for the delegate sets. The difference here is that producers use the ego-centric view instead of

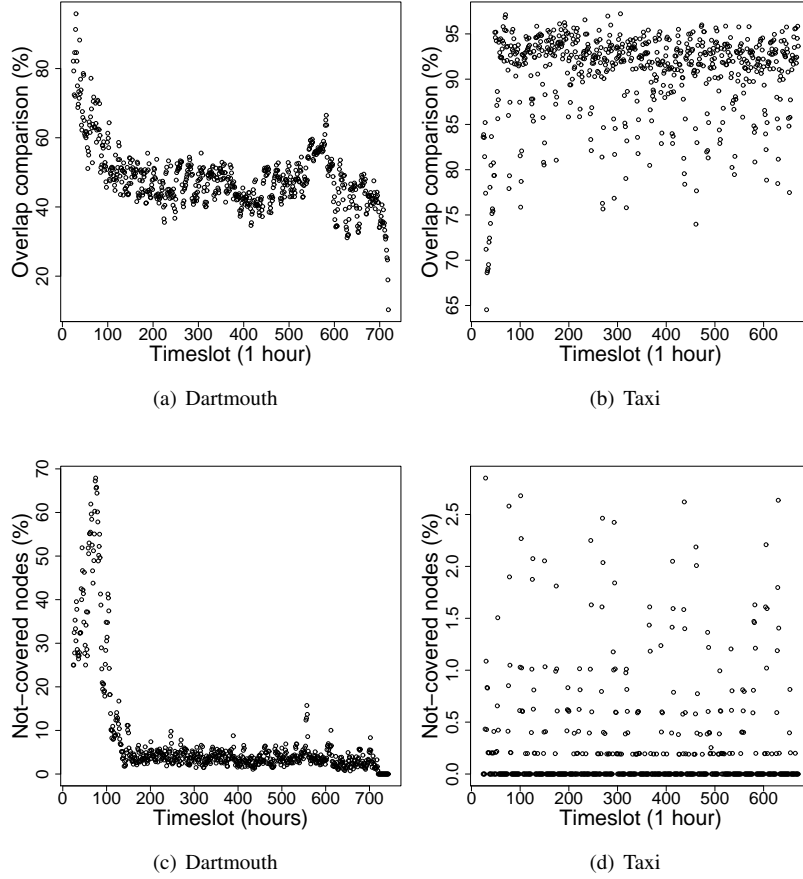


Figure 8: BDC-C strategy. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: (a) Dartmouth: 74.86%, (b) Taxi: 90.64%. (c)-(d) Amount of not-covered nodes by the  $\Delta$  sets gotten from the predicted traces: (c) Dartmouth: 8.04% (d) Taxi: 0.23%.

the complete view of the network, which gives a different network view to each node. Thus, each node firstly ranks all the nodes it sees in its ego-centric network according to the considered social-inspired strategy. An ordered list containing the best nodes of each ego network is then generated and the best  $|\Delta|$  nodes are selected to compose the delegate sets. Since duplications can happen (i.e. a best node in the ego network of node  $p_i$  can also be the best node in the ego network of node  $p_j$ ), they are removed from the final selected delegate set. This may result in smaller sets compared to the ones provided by the benchmark approach.

**DC-E.** Table 6 summarizes the average percentage set sizes resulted from the degree centrality selection applied to an ego-centric network. Details are presented in Fig. 9(a), which shows that an average of 61.94% (resp. 71.68% in the Taxi data set in Fig. 9(b))

Table 6: Avg. results for DC-E.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	159.62	32.6
	predicted	255.22	38.9
Taxi	measured	33.15	6.57
	predicted	84.93	16.48

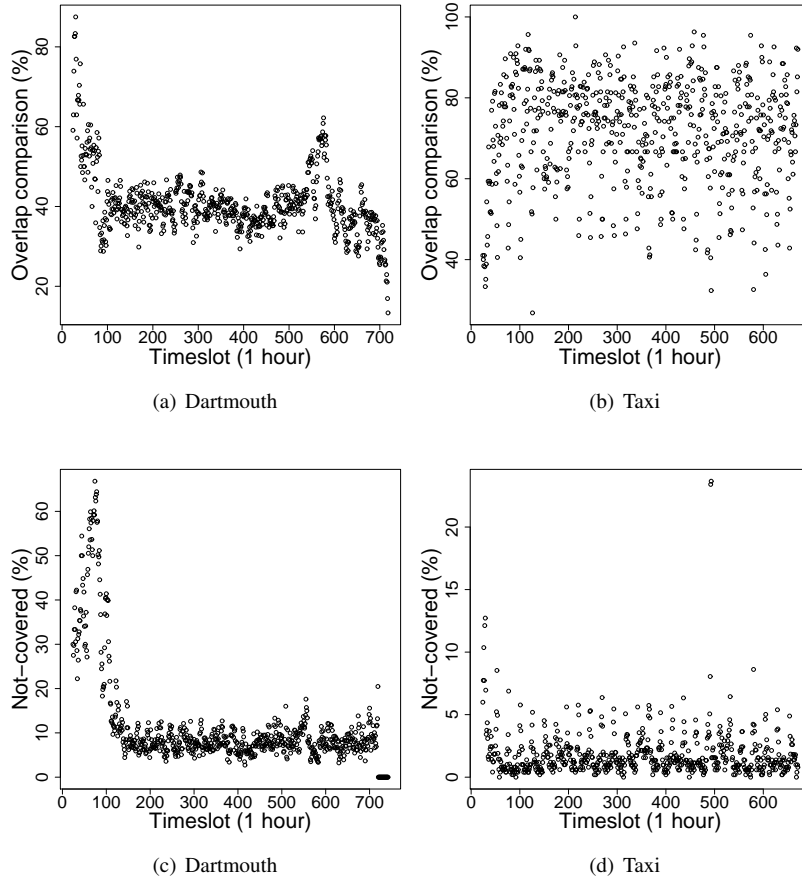


Figure 9: DC-E strategy. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: 11.85% (d) Taxi: 1.88%.

of delegates overlap between the traces. Fig. 9(c) shows that an average of 11.85% of nodes (resp. 1.88% in the Taxi data set in Fig. 9(d)) remain uncovered.

**BC-E.** We now evaluate the betweenness centrality selection applied to an ego-centric network. The results are summarized in Table 7. Fig. 10(a) shows an delegate overlap of 69.78% on average (resp. 77.63% in the Taxi data set in Fig. 10(b)), while numbers of uncovered producers are presented in Figs. 10(c) and 10(d). The average values are 15.7% for the Dartmouth trace and 1.91% for the Taxi data set.

Table 7: Avg. results for BC-E.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	110.06	21.78
	predicted	190.88	28.34
Taxi	measured	31.67	6.27
	predicted	80.58	15.64

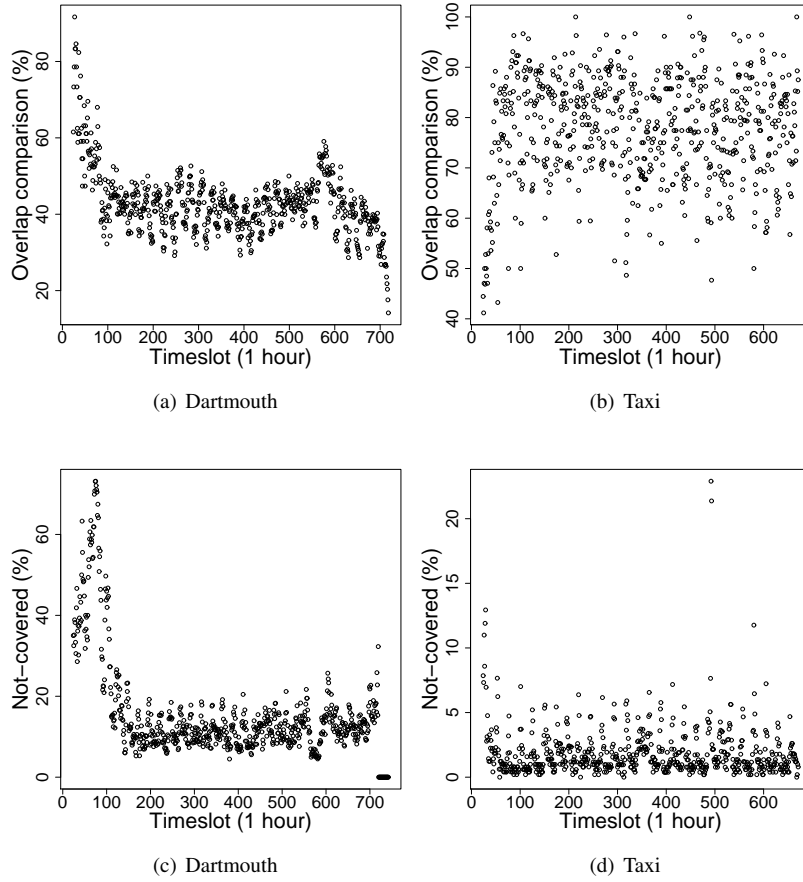


Figure 10: BC-E strategy. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: (a) Dartmouth: 69.78%, (b) Taxi: 77.63%. (c)-(d) Amount of not-covered nodes by the  $\Delta$  sets gotten from the predicted traces: (c) Dartmouth: 15.7% (d) Taxi: 1.91%.

**BDC-E.** Table 8 summarizes the results for the BDC-C strategy. Fig. 11(a) shows that, on average, 69.24% (resp. 73.33% in the Taxi data set) of the delegates overlap between the traces. In terms of producers that remain uncovered, Fig. 11(c) details the results for the Dartmouth set, with an average of 10.67%, while Fig. 11(d) shows 1.48% of producers left uncovered in the Taxi data set.

Table 8: Avg. results for BDC-E.

Data set	Traces	$ \Delta $	% delegates
Dartmouth	measured	174.75	35.29
	predicted	292.54	44.02
Taxi	measured	37.71	7.48
	predicted	97.93	19.01

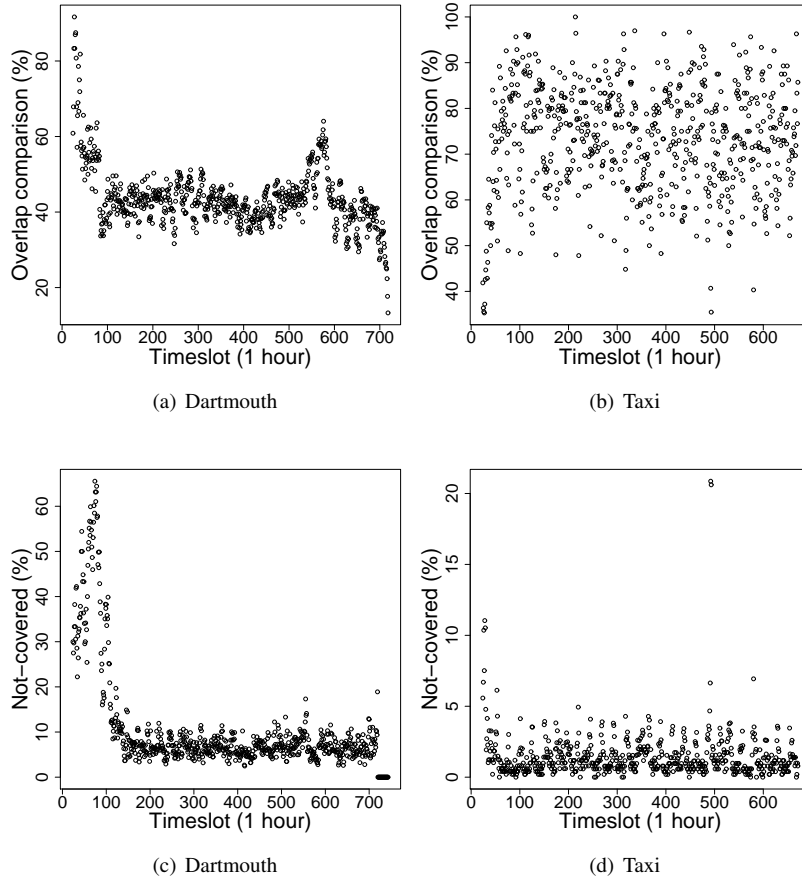


Figure 11: BDC-E strategy. (a)-(b) Percent of nodes per timeslot in the  $\Delta$  set obtained from the measured traces that are also found in the  $\Delta$  sets obtained from the predicted traces: (a) Dartmouth: 69.24%, (b) Taxi: 73.33%. (c)-(d) Amount of not-covered nodes by the  $\Delta$  sets gotten from the predicted traces: (c) Dartmouth: 10.67% (d) Taxi: 1.48%.

## 5.5 Discussion

To summarize our analysis, we present in Tables 9 and 10 the average percentage of producers selected as delegates, from the predicted sets and the average percentage of missed producers in the measured traces when those delegates are used. Although we know that degree and betweenness measures on both ego-centric and complete

Table 9: Summary of Dartmouth results.

Strategy	% missed	% delegates
Benchmark	5.7	70.97
DC-C	9.22	70.87
BC-C	7.31	70.97
BDC-C	8.04	70.87
DC-E	11.85	38.9
BC-E	15.7	28.34
BDC-E	10.67	44.02

Table 10: Summary of Taxi results.

Strategy	% missed	% delegates
Benchmark	0.08	81.38
DC-C	0.59	81.24
BC-C	0.13	81.38
BDC-C	0.23	81.38
DC-E	1.88	16.48
BC-E	1.91	15.64
BDC-E	1.48	19.01

networks are equivalent [20], we still were surprised to see that, in both data sets and for equivalent coverage, ego-centric solutions require a much smaller number of delegates. Additionally, we observe that better coverage and set sizes are obtained for the Taxi data set. This is expected because of its much higher average contacts per nodes (cf. Fig. 4).

For the sake of clarity, we show in Fig. 12 the average percentage of delegates against the average percentage of uncovered producers, as described in Section 5.4. We can observe that the results obtained through prediction are equivalent to the ones obtained through measurements. This confirms our expectation that: (i) a window of two days is enough to obtain good prediction and (ii) our prediction strategy is a reliable basis for good delegate set selection.

We have also investigated time slots of 24 hours, which resulted, as expected, in smaller delegate sets (we do not show the results here due to the lack of space). It is worth noting however that the time slot impacts the speed and trajectory design of the collectors. As for the design of collectors' trajectories, this impact study is left for future work.

## 6 Related work

Reactive and proactive schemes have been proposed in the domain of data delivery in sparsely connected networks and mobility-assisted schemes. Reactive schemes rely on movement that is inherent of the devices themselves to help deliver messages. When disconnected, nodes passively wait for their own mobility to allow them to reconnect [3, 18, 24]. More closely related to our work, proactive approaches make nodes modify their trajectories for communication purposes, making use of mobility to improve capacity and connectivity [2, 14, 26]. In particular, data mules [14] and mes-

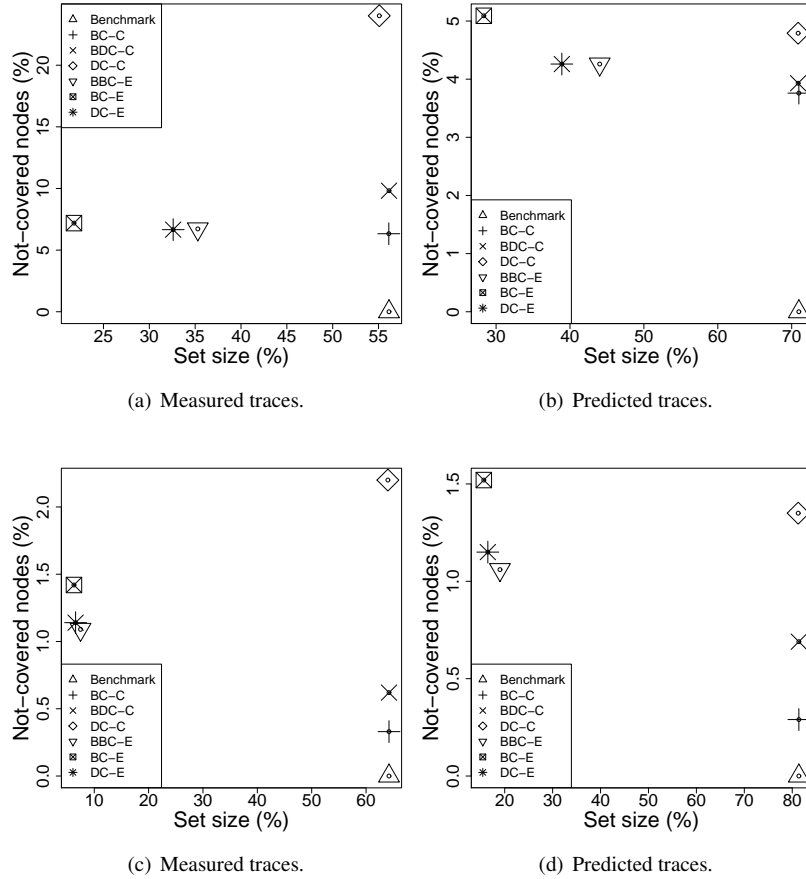


Figure 12: Average percentage of delegates as a function of the average percentage of not-covered nodes. (a)-(b) Dartmouth data sets. (c)-(d) Taxi data sets.

sage ferrying [26] are mobile nodes that move around the deployment area and take responsibility for carrying data between nodes. Smart-Tag [2] proposes to use mobile individuals to carry messages between disconnected devices and physical places, but no social interaction between nodes is considered.

Other works investigate human mobility in terms of pairwise contact and inter-contact times [4–6]. Recently, underlying mobility patterns have been explored for social-based routing [7, 13, 23]. In particular, Plat et al. propose to use connected dominating sets as message ferries and routing relays [23]. Nevertheless, neither social behavior nor relative importance of mobile nodes are explored, since message ferries are selected based only on neighborhood analysis.

The originality of our work is that we investigate established social-inspired techniques and evaluate the relative importance of mobile nodes in the specific case of data collection. By taking advantage of inherent social cyclicity, we do not need to enforce mobility to help the collection system.

## 7 Summary and outlook

In this paper, we have addressed the design of system support for robust data collection in wireless networks that face problems as sparse connectivity, high degree of mobility, and unreliable links. We have focused on the case of a data collection system where the number of collectors is much larger than the number of producers. To solve this problem, we have proposed a two-tiered approach where a subset of the producers are promoted to the rank of delegates, which are responsible for helping collectors gather data from the network. By relying on the social behavior of nodes, we moved beyond the current state-of-the-art that introduces particular entities such as data mules or message ferries. To decide which producers to promote, we have investigated several strategies based on the social interactions producers have among them. In building our system, we observe that much can be extracted from the inherent mobility of the nodes and that our prediction strategy is effective when used for delegate selection. Among the many results using existing real mobility traces, we were very surprised to note that local knowledge of the network is more than enough to achieve high collection ratios, with values that are close to those obtained with full knowledge of the topology. As part of ongoing work, we are applying the same analysis to other mobility traces and are working on trajectory design to decide the order in which delegates should be visited by collectors.

## References

- [1] C. Adjih, P. Jacquet, and L. Viennot. Computing connected dominated sets with multipoint relays. Technical Report 4597, INRIA, 2002.
- [2] A. Beaufour, M. Leopold, and P. Bonnet. Smart-tag based data dissemination. In *ACM International Workshop on Wireless Sensor Networks and Applications*, 2002.
- [3] J. Burgess, B. Gallaghern, D. Jensen, and B. N. Levine. MaxProp: routing for vehicle-based disruption-tolerant networks. In *Proc. of IEEE Infocom*, Apr. 2006.
- [4] H. Cai and D. Y. Eun. Toward stochastic anatomy of inter-meeting time distribution under general mobility models. In *Proc. of ACM MobiHoc*, May 2008.
- [5] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):600–620, June 2007.
- [6] V. Conan, J. Leguay, and T. Friedman. Characterizing pairwise inter-contact patterns in delay tolerant networks. Oct. 2007.
- [7] E. Daly and M. Haahr. Social network analysis for routing in disconnected delay-tolerant manets. In *Proc. of ACM MobiHoc*, Sept. 2007.
- [8] J.-M. François and G. Leduc. Delivery guarantees in predictable disruption tolerant networks. In *IFIP Networking*, May 2007.
- [9] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.



- [10] L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1979.
- [11] N. Glance and D. Snowdon. Pollen: using people as a communication medium. *Elsevier Computer Networks*, 35(4):429–442, March 2001.
- [12] W. Hsu and A. Helmy. On nodal encounter patterns in wireless lan traces. In *Proc. of IEEE WinMee*, Boston, USA, April 2006.
- [13] P. Hui, J. Crowcroft, and E. Yoneki. BUBBLE Rap: Social-based forwarding in delay tolerant networks. In *Proc. of ACM MobiHoc*, May 2008.
- [14] D. Jea, A. Somasundara, and M. Srivastava. Multiple controlled mobile elements (data mules) for data collection in sensor networks. In *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2005.
- [15] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *Proc. of IEEE Infocom*, Barcelona, Spain, April 2006.
- [16] D. Kotz, T. Henderson, and I. Abyzov. CRAWDAD trace dartmouth/campus/syslog/01\_04 (v. 2004-12-18), December 2004.
- [17] David Kotz, Tristan Henderson, and Ilya Abyzov. CRAWDAD data set dartmouth/campus (v. 2007-02-08). From <http://crawdad.cs.dartmouth.edu/dartmouth/campus>, February 2007.
- [18] A. Lindgren, A. Doria, and O. Schelen. Probabilistic routing in intermittently connected networks. *ACM SIGMOBILE MC2R*, 7(3):19–20, July 2003.
- [19] C. Liu and J. Wu. Routing in a cyclic mobispace. In *ACM Mobihoc*, May 2008.
- [20] P. V. Marsden. Egocentric and sociocentric measures of network centrality. *Social Networks*, 24(4):407–422, 2002.
- [21] S. Milgram. The small world problem. *Psychology Today* 1, pages 60–67, May 1967.
- [22] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from <http://crawdad.cs.dartmouth.edu/epfl/mobility>, February 2009.
- [23] B. K. Polat, P. Sachdeva, M. H. Ammar, and E. W. Zegura. Message ferries as generalized dominating sets in intermittently connected mobile networks. In *ACM MobiOpp*, Pisa, Italy, 2010.
- [24] R. Ramanathan, R. Hansen, P. Basu, R. Rosales-Hain, and R. Krishnan. Prioritized epidemic routing for opportunistic networks. In *ACM MobiOpp*, June 2007.
- [25] San Francisco Exploratorium’s Invisible Dynamics initiative. Cabspotting, 2005.
- [26] W. Zhao, M. Ammar, and E. Zegura. A message ferrying approach for data delivery in sparse mobile ad hoc networks. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2004.

---

## **Contents**

<b>1</b>	<b>Context and motivation</b>	<b>3</b>
<b>2</b>	<b>Problem definition</b>	<b>4</b>
<b>3</b>	<b>Sketch of the solution</b>	<b>5</b>
<b>4</b>	<b>Social-oriented delegation</b>	<b>5</b>
4.1	Social-inspired metrics . . . . .	5
4.2	Topology-awareness . . . . .	7
4.3	Benchmark strategy . . . . .	7
<b>5</b>	<b>Evaluation</b>	<b>8</b>
5.1	Data sets and methodology . . . . .	8
5.2	Evaluation methodology . . . . .	10
5.3	Benchmark analysis . . . . .	11
5.4	Topology-awareness . . . . .	12
5.4.1	Influence of complete network view (C) . . . . .	12
5.4.2	Influence of ego-centric network view (E) . . . . .	14
5.5	Discussion . . . . .	18
<b>6</b>	<b>Related work</b>	<b>19</b>
<b>7</b>	<b>Summary and outlook</b>	<b>21</b>



---

Centre de recherche INRIA Saclay – Île-de-France  
Parc Orsay Université - ZAC des Vignes  
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex  
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier  
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq  
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique  
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex  
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex  
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex  
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

---

Éditeur  
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)  
<http://www.inria.fr>  
ISSN 0249-6399