

Collaborative Data Collection in Global Sensing Systems

Greg Bigwood^{†,‡}, Aline Carneiro Viana[‡], Marcelo Dias de Amorim^{*}, and Mathias Boc[◇]

[†] University of St Andrews [‡] INRIA ^{*} UPMC Sorbonne [◇] CEA LIST Universités

gjb@cs.st-andrews.ac.uk, aline.viana@inria.fr, marcelo.amorim@lip6.fr, michael.boc@cea.fr

Abstract—Here, we first investigate data collection delegation strategies based on the relative importance of nodes in their social interactions. 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 using real mobility data sets.

I. CONTEXT AND MOTIVATION

We consider mobile nodes (i.e., producers) in a DTN deployment, carrying smart phones. In this context, we consider a collection system composed of specialized devices called *collectors* whose role is to move around and collect data when they enter within communication range with producers. We use direct wireless connectivity among nodes whenever available. There is one main issue that arises when designing such a system: The number of producers might be much larger than the number of collectors. To address these challenges, we need alternative solutions to overcome the insufficiency of the collection system. We leverage the contact opportunities that emerge due to the natural mobility of nodes to tackle this problem. Related works in the literature deal with collector placement or designing the trajectory to visit all producers in the network. We instead focus on detecting *which* subset of producers could help collectors in the data collection task, while maintaining their “normal” behavior within the network.

More specifically, 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 for collecting data from other producers and serve as intermediate relays between them and the collectors. 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 [8] and that encounters are sufficient to build a connected relationship graph [5]. The problem here is determining which producers should be promoted as delegates. To this end, we make the following contributions:

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

- We evaluate the strategies using trace-driven analysis obtained in real situations: i.e., Dartmouth College wireless traces [6] and San Francisco taxi traces [9] (Section V).

Full details on the delegation strategies and prediction system are available in [2], along with numerical analysis.

II. PROBLEM DEFINITION AND SKETCH OF THE SOLUTION

Our goal is to select a subset of producers that will be promoted as delegates to help the collectors obtain the data generated by each producer. In fact, producers that are promoted as delegates do not have to change their trajectories to meet collectors or other producers. They continue producing their own data while gathering data from producers they meet. Because of storage and communication capacity constraints, we wish to avoid relying on fully epidemic approaches. We adopt in this paper a two-level strategy where a node is either a delegate or a producer that meets a delegate. In this way, simple producers transfer their data to one or more delegates they meet and collectors have only to visit the latter. Note that *no forwarding is required*, since data is transmitted through direct contact opportunities. Our problem consists then in computing the delegate subset Δ : the smallest subset of producers whose movement guarantees the biggest achievable number of visited other producers within a certain time slot. To compute the set of delegates, we need to know in advance the contact patterns among producers. 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. We naturally adopt the latter alternative (Section III). Once encounters are predicted, we apply *social-inspired selection schemes* consisting of the quantification of the relative importance of producers, to compute the set of delegates.

III. PREDICTION STRATEGY

We propose to use a slotted prediction strategy with a history window of two days. 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 matrices $M(i, j) = [N \times N]$ 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. Note

that because matrix $M(i, j)$ is only available *at the end* of each slot period $c(i, j)$, i.e., after the real contacts are observed, it cannot be used to compute delegates for the time slot $c(i, j)$.

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. We empirically observe that this prediction procedure does indeed yields good results. More formally we have:

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

Note that prediction may change on a periodic basis, which also periodically affects the selection of delegates, which is also performed periodically (see Section IV).

IV. SOCIAL-ORIENTED DELEGATION

In this section, we investigate which nodes should be promoted as delegates. We combine socially-inspired metrics and different levels of topology-awareness to identify six strategies to determine the most appropriate set of delegates, based on their relative importance. To investigate the relative importance of producers, we exploit the well known centrality concept from graph theory and network analysis. In the literature, “importance” is described through three different metrics [3], [4]: closeness, degree, or betweenness. We use the degree and betweenness centrality metrics in our analysis and a combination of the two, described hereafter.

Degree centrality (DC) describes the number of direct connections that involve a given node.

Betweenness centrality (BC) measures the frequency that a node appears on the shortest paths linking any two nodes.

Betweenness and degree centrality (BDC). A metric utilizing both centralities seeks to gain the benefit of connecting to a large number of nodes, whilst simultaneously reaping the benefit of bridging groups of nodes within the network. Considering that we are interested in determining a subset of good delegates, the resultant set, BDC, consists of 50% top nodes from the betweenness centrality set, BC, and 50% high popular nodes from the degree centrality set, DC.

A. Topology-awareness

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

Complete and bounded network (C). We use this level of knowledge to get the upper-bounded results to be compared with the ones obtained when using ego-centric networks.

Ego-centric networks (E). An ego-centric network consists of a single actor (named *ego*), its 1-hop neighbors (named *alters*), and all the links among those alters. An ego network requires less state than a 2-hop neighborhood topology, such as the one required for the computation of dominating sets. Existing literature shows that degree and betweenness centrality, when

computed in ego networks, provide quite good results when compared with complete knowledge [7].

B. Reference strategy

We also consider a reference solution that leads to the best possible result. This corresponds to computing the minimum dominating set (MDS) on the encounter graph for each time slot. As this problem is known to be NP-hard, we consider an alternative solution borrowed from the computation of multi-point relays as our reference [1]. Although leading to better coverage, this reference approach requires a large number of delegates to perform the work.

V. EVALUATION

A. Data sets and methodology

We use the well-known data sets: Dartmouth College campus [6] and San Francisco Taxi cabs [9]. By nature, they show different mobility patterns, types of users, and environment conditions. We thus expect to observe different social behaviors, resulting in different usage models. This diversity will allow us to better understand the characteristics of the analyzed strategies. For both traces we assume that two nodes are connected if they are within in $250m$ range.

Prediction vs. measurement. The predicted traces result in smaller average numbers of contacts per node and larger average numbers of active nodes. There are two reasons for this. 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 hand, we assume 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 of presence.

B. Evaluation methodology

We focus on both the coverage, and size of the resulting delegation sets on a per-slot basis. We evaluate the coverage property 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. Note that no simulator is required for these computations as we use scripts operating directly over the measured and predicted traces.

C. Reference analysis

Table I summarizes the average amount of producers that become delegates (in both absolute and relative terms) per collection period. When we apply the reference strategy to this trace, the resultant delegates sets guarantee 100% coverage (i.e., 0 missed nodes).

D. Topology-awareness

1) *Influence of complete network view (C)*: Producers are first ordered based upon their social influence. To select the best delegates, we use as a reference *the same size of the delegate set obtained for the measured data set in the reference*

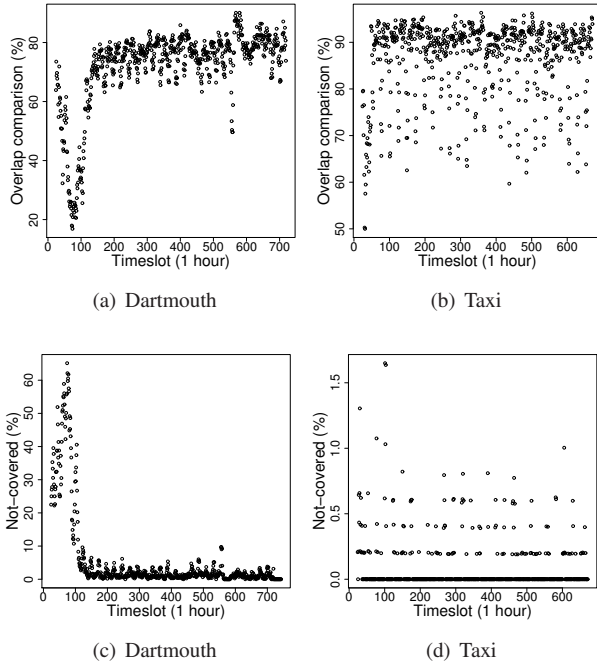


Fig. 1: Reference approach. (a)-(b) Percent of delegates per time slot obtained from the measured traces that are also delegates at the predicted traces. (c)-(d) Amount of missed nodes by the delegates obtained from the predicted traces.

TABLE I: Average results for the reference.

Data set	Traces	$ \Delta $	% delegates	% missed
Dartmouth	meas.	304.36	56.14	0
	pred.	508.43	70.97	0
Taxi	meas.	326.06	64.22	0
	pred.	419.33	81.38	0

approach (cf. 1st and 3rd lines of the Table I), and pick the top-rated producers until the delegate sets' sizes are the same.

DC-C. Although the identity of selected delegates may differ, the resulting average set size and percentage of missed nodes are equivalent to the ones in Table I. We see an average of 71.42% (resp. 86.77%) of delegates overlapping between the traces, while an average of 9.22% (resp. 0.59%) of nodes are not covered by predicted delegates.

BC-C. As with the previous strategy, the results of the betweenness centrality selection applied to the complete network of both traces are equivalent to the ones in Table I. We see an average of 86.76% (resp. 99.07% in the Taxi data set) of delegates' overlap. An average of 7.31% (resp. 0.13% in the Taxi data set) of nodes remain uncovered.

BDC-C. Results related to set size and percentage of delegates when applying BDC-C strategy are summarized in Table I. As in the previous two cases, we compare the overlaps between the measured and predicted 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. We see that the proportion of nodes not covered is 8.04% in the Dartmouth case and only 0.23% in the Taxi case.

2) *Influence of ego-centric network view (E):* We relax the assumption of the previous section by only requiring a local knowledge of the network. For this, we use the notion of ego-centric networks [7]. The goal is to verify if sets of delegates can be correctly computed based only on the local network view of nodes. As in the previous section, the sizes $|\Delta|$ of the measured sets in the reference case are used as reference sizes for the delegate sets. Firstly, all nodes in each ego-centric network of a node are ranked according to the social-inspired strategies. An ordered list containing the highest ranked nodes of each ego network is then generated and the highest $|\Delta|$ 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 reference approach.

DC-E. Table II summarizes the average percentage set sizes resulting from the degree centrality selection applied to an ego-centric network. We see that an average of 61.94% (resp. 71.68% in the Taxi data set) of delegates overlap between the traces. An average of 11.85% of nodes (resp. 1.88% in the Taxi data set) remain uncovered.

BC-E. We now evaluate the betweenness centrality selection applied to an ego-centric network. The results are also summarized in Table II. We see a delegate overlap of 69.78% on average (resp. 77.63% in the Taxi data set), while numbers of uncovered producers are 15.7% for the Dartmouth trace and 1.91% for the Taxi data set.

BDC-E. Table II also summarizes the results for the BDC-C strategy. We see 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 for the Dartmouth set, we see an average of 10.67%, while 1.48% of producers left uncovered in the Taxi data set.

It is worth noting that, though generating smaller delegate sets, selection strategies applied in ego-centric networks and measured traces result in non-zero missed nodes sets, in contrast to the reference strategy.

E. Discussion

To summarize our analysis, we present in Tables III and IV: (i) the average percentage of producers selected as delegates from the predicted sets, and (ii) 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 networks are equivalent [7], we are 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, due to its higher average contacts per nodes.

For the sake of clarity, we show in Fig. 2 the average percentage of delegates against the average percentage of uncovered producers, as described in Section V-D. Despite the differences in mobility, the two traces present similar

TABLE II: Average results for the different ego-centric combinations

Data set	Traces	DC-E		BC-E		BDC-E	
		Δ	% delegates	Δ	% delegates	Δ	% delegates
Dartmouth	measured	159.62	32.6	110.06	21.78	174.75	35.29
	predicted	255.22	38.9	190.88	28.34	292.54	44.02
Taxi	measured	33.15	6.57	31.67	6.27	37.71	7.48
	predicted	84.93	16.48	80.58	15.64	97.93	19.01

clustering properties for the social-inspired strategies. In addition, we observe that the results obtained through prediction are equivalent to the those 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. Finally, the BDC and BC strategies seem to show a compromise between coverage and uncovered producers.

TABLE III: Summary of Dartmouth results.

Strategy	% missed	% delegates
Reference	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

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

VI. SUMMARY AND OUTLOOK

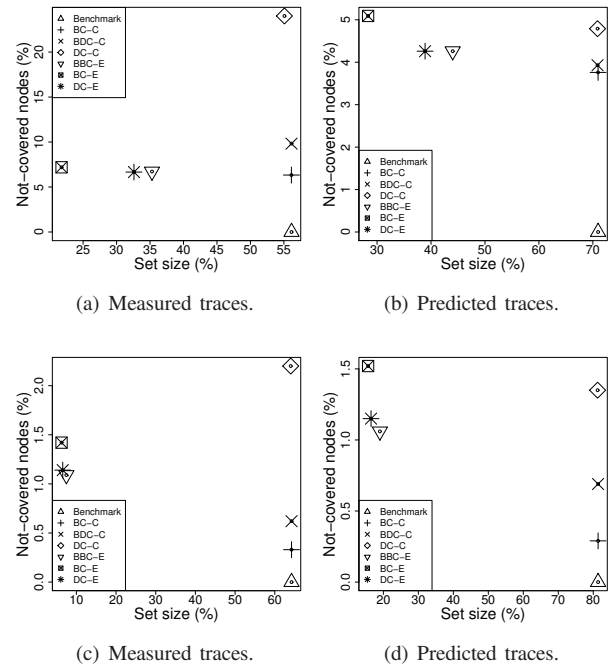
We have addressed the design of system support for robust data collection in global sensing pervasive applications. We have proposed a two-tiered approach where a subset of the producers are promoted to the rank of delegates, which are responsible for helping collectors gather data from the network. To decide which producers to promote, we have investigated several strategies based on the social interactions among producers. 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 are 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 synthetic mobility traces, and are working on trajectory design to decide the order in which collectors should visit delegates.

REFERENCES

[1] C. Adjih, P. Jacquet, and L. Viennot, "Computing connected dominated sets with multipoint relays," INRIA, Tech. Rep. 4597, 2002.
[2] G. Bigwood, A. C. Viana, M. Boc, and M. D. de Amorim, "Opportunistic data collection through delegation," INRIA, Research Report, 2010, rR-7361. [Online]. Available: <http://hal.inria.fr/inria-00508273/en>

TABLE IV: Summary of Taxi results.

Strategy	% missed	% delegates
Reference	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

**Fig. 2:** 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.

[3] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.
[4] —, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.
[5] W. Hsu and A. Helmy, "On nodal encounter patterns in wireless lan traces," in *Proc. of IEEE WinMee*, Boston, USA, Apr. 2006.
[6] D. Kotz, T. Henderson, and I. Abyzov, "CRAWDAD trace dartmouth/campus/syslog/01_04 (v. 2004-12-18)," December 2004. [Online]. Available: <http://crawdad.cs.dartmouth.edu/dartmouth/campus/syslog/0104>
[7] P. V. Marsden, "Egocentric and sociocentric measures of network centrality," *Social Networks*, vol. 24, no. 4, pp. 407–422, 2002.
[8] S. Milgram, "The small world problem," *Psychology Today* 1, pp. 60–67, May 1967.
[9] San Francisco Exploratorium's Invisible Dynamics initiative, "Cabspotting," 2005. [Online]. Available: <http://www.cabspotting.com/>