Hub-Betweenness Analysis in Delay Tolerant Networks Inferred by Real Traces

Giuliano Grossi, Federico Pedersini

Hub-Betweenness Analysis in Delay Tolerant Networks Inferred by Real Traces

Giuliano Grossi and Federico Pedersini
Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Via Comelico 39, I-20135 Milano, Italy
{grossi,pedersini}@dsi.unimi.it

***Abstract*—In this paper we study the influence of using hub nodes to relay messages in human-based delay tolerant networks (DTNs), by analyzing empirical traces obtained by human mobility monitoring experiments. Four traces known in literature have been considered. We exploit a measure of centrality (or betweenness) over underlying graphs associated to temporal networks, in order to establish (in probability) the ability to forward information using a restricted number of active relay nodes (hubs). The proposed analyses are carried out by introducing time-dependent networks associated to real traces.**

The empirical cumulative distribution of the node betweenness and the shortest paths length (or geodetic) are derived and characterized. The analysis shows that the geodetic path length follows a lognormal (skewed) distribution. It is also observed that the measures of betweenness on the nodes, if ordered decreasingly and interpreted as probability distribution, exhibit an exponential-like decay, with very high betweenness for few nodes and much lower for all the others.

Based on this knowledge, we study the probability of successful delivery when a set of nodes with low betweenness are deactivated as forwarding nodes. Under these assumptions, we give the probability that a k-length path connecting an arbitrary source-destination pair belong to the set of the activated hub nodes. The results show how a trade-off can be found between the number of relay nodes (hubs) activated in a temporal network and the network's delivery rate, when message forwarding is allowed only for these hubs.

I. Introduction

We focus on opportunistic communication in DTNs where the contacts appear opportunistically without any prior information on future encounters. As the agents (or nodes) in a network communicate over time, information flows in complex ways. Gossip protocols in such networks, for example, are based on the dissemination of information through a network using node-to-node transmissions.

The task of understanding the temporal dynamics of human mobility is difficult and can be accomplished by capturing traces of human interaction in pervasive environments [1]. To this aim, we consider temporal networks in which nodes have been communicating with their neighbors for a fixed time. Referring to the graph representing the temporal network at a given time, each edge represents an active connection and is labeled by the time at which the involved nodes start to exchange information. Information flows along a path in this network only if the time labels on the path edges are monotonically non-decreasing; thus, such time-respecting paths are of crucial importance in understanding the way in which information flows through the network [2].

A key element to capture such paths is the concept of centrality [3]–[5]. The centrality of a node in a DTN is a useful measure for its potential capacity to reduce the path lengths that connect other members of the network. The tendency to restrict the forwarding activity to nodes with high centrality can reduce considerably the message complexity of the whole network. These findings are especially true on communities where communication occurs according to people's social relations [6], [7]. It has been shown that some nodes in a community are the common acquaintances of other nodes acting as communication hubs, and that social-based forwarding schemas outperform traditional approaches based on prediction [8]. Another study on wireless DTN communication [9] focuses on encounter patterns instead of hub centrality. This work develops an analysis of the properties of potential infra-structureless networks in environments like campuses and conferences. The authors tried to remove nodes from the forwarder set (making them inactive) starting from the nodes with most unique encounters. They discovered that the underlying encounter pattern is so rich that, even if 20-30% of nodes is removed from the forwarder set, the success ratio of message forwarding does not degrade significantly.

The main goal in this paper is to estimate the probability that, for a fixed subset of hub nodes with high centrality, a path connecting two arbitrary nodes entirely belongs to such special set. We denote as *successful delivery rate* the probability that all relays used as forwarders are hubs.

In order to achieve this result we first characterize the distribution of the shortest path length on the real traces. We find that, despite the fact that the traces are coming from different experiments, the shortest path length exhibits an empirical distribution that can be well fitted by a lognormal distribution. Since the computation of all shortest paths in a temporal network is too expensive, we sample a sufficiently large number of paths by randomly choosing the source-destination pairs and the initial delivery time. Based on this dataset, we estimate the node centrality and call it *hub-betweenness*. We combine the empirical distribution of the shortest path length with the hub-betweenness to derive the probability that, for a fixed subset of hubs, the nodes that form a path of fixed length belong to the subset. In other words,

we study the probability of successful delivery when a set of nodes with low betweenness is deactivated. This study can be useful for distributed routing strategies that tend to behave like the centralized algorithms, by collecting information about the nodes' neighborhood.

The analysis reported here can be useful in case of mobility-assisted routing, where each node independently makes forwarding decision at each encounter. In particular the presented results hold under the assumption of single-copy delivery. Multiple-copy schemes are not considered here because they make the computation of the successful delivery rate much more complex.

We used the dartmouth/campus data set [10] from CRAWDAD for the trace analysis in Section III. We have seen that for the datasets CAMBRIDGE and INFOCOM'05, the use of half of the nodes as hubs approximately guarantees the same delivery ratio as using all nodes, while for the other two datasets (MIT and PTR), at least 75% of nodes is needed. Selecting nodes with high betweenness as forwarders lead to improve the network traffic as the network prefers the shortest paths for delivery.

II. Connectivity properties of real traces

In the past few years many researchers have devoted significant resources and energy to collecting realistic network traces. We pursue the study of opportunistic network scenarios based on human mobility. Among many real traces reflecting human-to-human relations, we choose four datasets gathered respectively in experiments MIT Reality [11], CAMBRIDGE [1], INFOCOM05 [12] and PTR [13].

A common framework used to capture temporal dynamics in DTNs is a network with an explicit time-ordering on its edges, i.e., a temporal network [2]. Formally, a temporal network is an undirected graph $G = (V, E)$ in which each edge e is annotated with a time label $\lambda(e)$ specifying the time at which its two endpoints communicated. Thus, one can view a temporal network as the pair (G, λ), where λ is a function from the edge set to the real numbers; we refer to λ as a time labeling of G. A n-length path P in G, denoted by $P = v_1 \rightarrow \cdots \rightarrow v_n$, is called time-respecting if the labels on its edges are non-decreasing.

To understand the network structure from the traces, we use different tools like metrics to measure the centrality or the popularity of the nodes, and empirical distribution on the length of the the geodesic (shortest) paths connecting pairs of nodes.

A. Hub-betweenness measure

There are several ways to measure the centrality of nodes. One of the most used is the so called *betweenness centrality* given by Freeman [3], [4], usually called simply *betweenness*. In some sense this metric measures the information flowing over a node by giving the extent to which the node lies on the geodesic paths linking others nodes. Since transport is more efficient along shortest paths, nodes of high betweenness are important for transport.

It is normally calculated as the fraction of geodesic between node pairs that pass through the node of interest. More formally, let $\rho_j(s,t)$ be the number of geodesic paths form node s to node t that pass through j and $\rho(s,t)$ the total number of geodesic paths form node s to node t. Then the betweenness of node j is

$$b_j = \sum_{s<t} \frac{\rho_j(s,t)}{\rho(s,t)}. \quad (1)$$

This definition of betweenness, however, takes into account for all the shortest paths from any pair of nodes originated at every time unit. Therefore, the use of (1) for practical computations is unrealistic.

For this reason, we consider a different measure of betweenness which approximates quite well that in (1) and can be directly obtained from real traces. The idea is to exploit the fact that betweenness of a node is proportional to the number of shortest paths that go through it. We call this measure *hub-betweenness* because it consists of a rank-computation on hubs in geodetic paths randomly drawn. Indeed, it indicates how often each node is used to relay data to other nodes. We simulate flooding over the temporal network extracted from the trace and counted the number of times each node is used to relay the data. A practical way to achieve this measurement is to randomly draw many source-destination pairs and, for each one, to draw a random time to start the construction of the temporal network on which the geodetic paths are computed. More precisely, given a source-destination pair (s,t) and an initial delivery time d belonging to the simulation interval, we build a network $G = (V, E)$ (initially empty) adding nodes (and edges) when the flooding scheme infects new ones, until the destination t is found. If the destination is not found the network is discarded. At the end of the construction, a k-length geodesic path $P = s \rightarrow h^{(1)} \rightarrow \cdots \rightarrow h^{(k-1)} \rightarrow t$ is achieved and the label $\lambda(s, h^{(1)}) \geq d$ set as the infection time of the first hub.

Thus the hub-betweenness is defined as

$$b_j = \sum_{(s,t,d)\in\mathcal{S}} \rho_j(s,t,d),$$

where $\mathcal{S} \subseteq \langle S, T, D\rangle$ is a set of randomly chosen triples from the sets of sources S, destinations T and initial times of the network construction D respectively.

In Fig. 1 the empirical distribution of the node betweenness computed for the four traces is shown. The plot has been obtained over up to 250,000 source-destination-time randomly chosen triple for the trace with most nodes.

Note that these empirical curves decrease in an exponential way, suggesting the idea that only a subset of hubs play a significant role in data forwarding along geodetic paths. In the INFOCOM and Cambridge experiments the 90% of the geodetic paths passes trough the 20% of the nodes, while in MIT and PTR the 50% of nodes is required to gather the same percentage of passages.

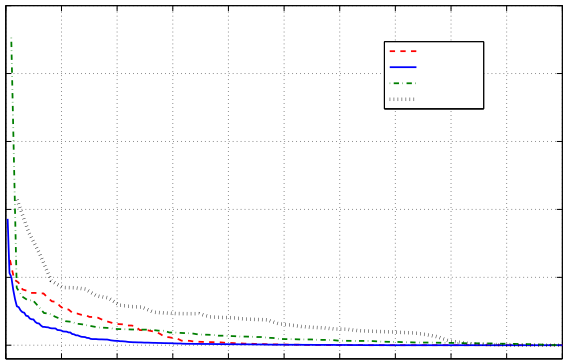

Fig. 1. Hub-betweenness index obtained by MIT, CAMBRIDGE, INFOCOM05 and PTR traces.

B. Shortest paths

We use the mentioned datasets to explore the geodesic path length in order to characterize the behavioral pattern of a wide class of DTNs based on human-to-human mobility. We found that this length follows a lognormal distribution with various values of the mean μ and standard deviation σ parameters obtained via the maximum likelihood method.

In Fig. 2, we show the fitting of the shortest path lengths with a lognormal distribution with parameters given in the previous table. These results stress the fact that despite their often large size, in most networks there is a relative short path between any two nodes. Once again the commonly believed small-world model for explaining human mobility is confirmed by experiments on real traces [9]. Table I reports the shortest path average values and parameter estimate of the lognormal distribution for each dataset.

TABLE I
SHORTEST PATH AVERAGE VALUES AND PARAMETER ESTIMATE OF THE LOGNORMAL DISTRIBUTION.

Trace	*average (std)*	μ	σ
MIT	3.11 (1.47)	0.79	0.53
CAMBRIDGE	4.02 (1.40)	1.19	0.37
INFOCOM	3.02 (1.22)	0.82	0.45
PTR	2.58 (1.02)	0.62	0.44

III. HUB IMPACT ON TRANSMISSION

We have given empirical evidence that hubs have a great impact in the delivery activity when the degree of success is provided by centrality measures like those discussed in the previous section.

The nodes betweenness plotted in Fig. 1 shows that, in all the experiments, the role of different nodes as relay nodes is very disuniform: most of the nodes participate as relay node in few paths, while few other nodes, characterized by the highest betweenness, are likely to participate to most message paths.

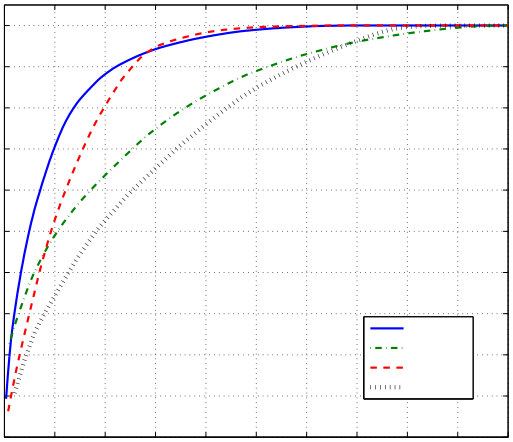

Fig. 3. Ecdf plotting of the hub-betweenness index, obtained respectively by MIT, CAMBRIDGE, INFOCOM05 and PTR traces.

This suggests the idea of allowing only the latter nodes to be active for message forwarding: this would greatly reduce the traffic on the network without impacting too much on the delivery rate, as these nodes would nevertheless allow for successful delivery of most messages along their geodesic path.

In this section we want to study the successful delivery rate of the network when only a subset of nodes (with high betweenness) is allowed to forward messages, thus working as hubs, while the other nodes are not active. The analysis can be useful in case of mobility-assisted routing, where each node independently makes forwarding decision at each encounter. These results are particularly significant when the message delivery scheme is single-copy.

When all network nodes are active, we can expect the probability to deliver a message along a geodesic path of length at most k follows a lognormal distribution:

$$F(k) = \Phi\left(\frac{\ln k - \mu}{\sigma}\right),$$

where Φ is the standard normal cumulative distribution. As said above, we estimated empirically this result by trace fitting (see Fig. 2).

For a given network $G = (V, E)$ of N nodes with $V = \{v_1, \ldots, v_N\}$, let $Q = \{h_1, \ldots, h_s\}$ be the subset of nodes (hubs) having highest hub-betweenness, with $s \leq N$. Let us also suppose that the hubs in Q are ordered by hub-betweenness in descending order, i.e., $b_{h_i} \geq b_{h_{i+1}}$ for $i = 1, \ldots, s-1$, as shown in the curves of Fig. 1.

For a given subset of s nodes Q, the plots in Fig. 3 represent the normalized cumulated hub-betweenness $H(s)$, given by:

$$H(s) = \frac{\sum_{i=1}^{s} b_i}{\sum_{i=1}^{N} b_i} = \sum_{i=1}^{s} h_i \ , \quad 1 \leq s \leq N \qquad (2)$$

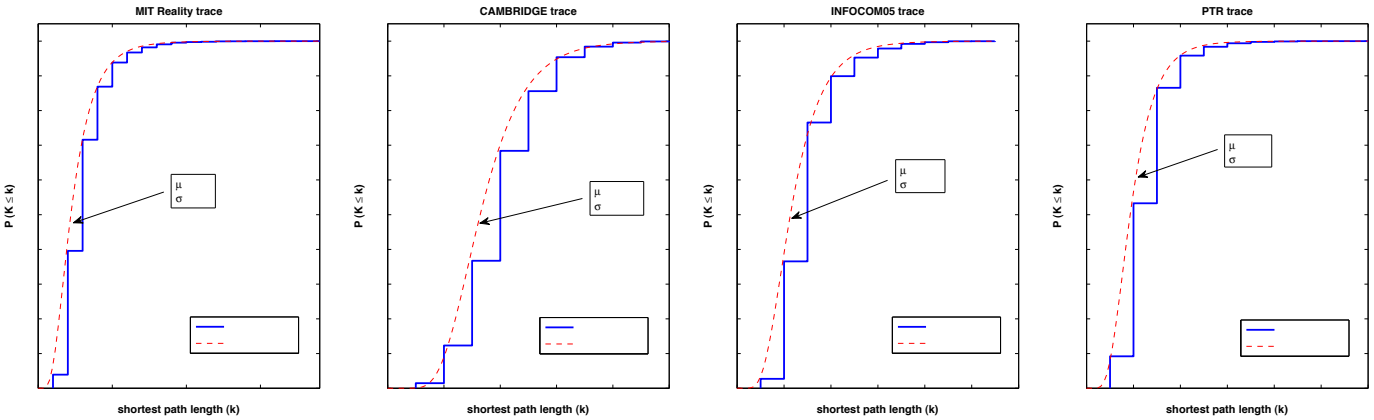

Fig. 2. Fitting the distribution of the shortest path lengths with a lognormal distribution, respectively for MIT, CAMBRIDGE, INFOCOM'05, and PTR traces.

being h_i the normalized hub-betweenness of node i:

$$h_i = \frac{b_i}{\sum_{j=1}^{N} b_j}.$$

The function $H(s)$ can be also interpreted as the probability that a message be forwarded by a relay node (a hub) that belongs to the subset Q. Following this interpretation, we derive the probability that an arbitrary k-length path belongs to the subset of selected hubs Q.

Let us first consider a path of length 2, that is, a path in which there is only one hub node forwarding the message from the source to the destination node. The probability for the hub node to belong to Q is then:

$$p_k(s) = H(s), \qquad k = 2.$$

For a generic path of length $k > 2$, the message is forwarded by $k-1$ relay nodes before reaching the destination. In this case, it is necessary to compute the probability that all these forwarders belong to the subset Q conditioned to the fact that they must be all different, since no repetitions are admitted. Such probability can be then expressed in terms of a composition of non-independent events, as follows:

$$p_k(s) = H(s) \prod_{i=2}^{k-1} \frac{H(s) - \sum_{j=1}^{i-1} h^{(j)}}{1 - \sum_{j=1}^{i-1} h^{(j)}} \qquad k > 2, \quad (3)$$

where $h^{(i)}$ denotes the betweenness of the node placed at the i-th position along the considered path, as defined in Section II. It is clear from equation (3) that this probability depends on the path, that is, on the specific position taken by relay node i along the path. Since this probability is not easy to compute because it is given by the the product of a non-polynomial number of terms, it would be of practical interest to provide a lower bound easy (linear number of terms in path length) to compute. This can be achieved by considering the worst case, corresponding to the choice of hub with the highest betweenness as first relay node, the second-highest betweenness for the second hop, and so on. This would mean, in our expressions, that $\hat{h}_i = h_i$ for all $i = 1, \ldots, k-1$. This provides a lower bound for $p_k(s)$:

$$\begin{aligned} p_k(s) &> H(s) \prod_{i=2}^{k-1} \frac{H(s) - \sum_{j=1}^{i-1} h_i}{1 - \sum_{j=1}^{i-1} h_i} \\ &= H(s) \prod_{i=2}^{k-1} \frac{H(s) - H(i-1)}{1 - H(i-1)}, \quad k > 2. \end{aligned} \quad (4)$$

Equation (4) expresses a lower bound for the probability of delivery along a minimum path of length k, in a network in which only the s nodes characterized by the highest betweenness are allowed to work as hub relaying messages: for s hub nodes, the probability of delivery (corresponding to the successful delivery rate) along a path of length at most k is:

$$\mathcal{D}_s(k) = P_k(s) F(k) \quad (5)$$

where

$$P_k(s) = \sum_{i=1}^{k} p_i(s).$$

We have employed equation (5) to estimate the successful delivery rate in the four considered experiments. For different percentages of the total amount of nodes, the cumulated betweenness $H(s)$ is extracted from the curves of Fig. 3, as reported in Table II. For each experiment (*MIT, INFOCOM'05, Cambridge, PTR*) the estimated delivery rate is compared to the lower bound of the delivery rate obtained from equation (5), for the different percentages of active hub nodes reported in Table II. The graphs of these comparisons are reported in Fig. 4.

In particular, it can be seen that for CAMBRIDGE and INFOCOM'05 the use of half of the nodes as hubs still guarantees the same delivery ratio as using all nodes. The same does not hold for MIT and PTR. It can be also noticed that, for values of k (geodesic length) greater than the average length (see Table I), the probability does not change significantly. This indicates that long paths do not significantly contribute to

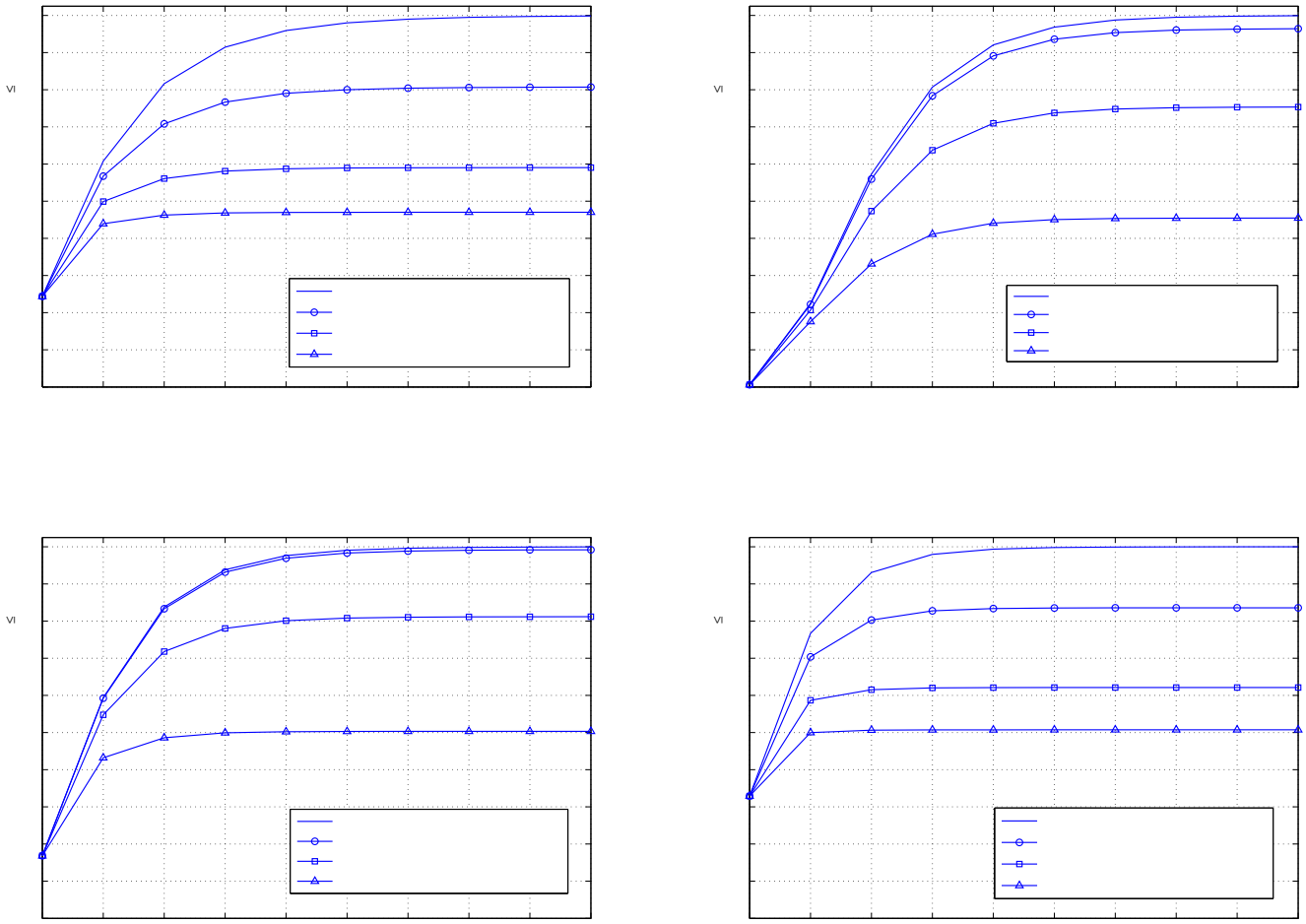

Fig. 4. Successful delivery rate along paths of length $l \leq k$, obtained using different percentages of nodes as relay nodes (hubs), for the four considered experiments (*MIT Reality, Cambridge, INFOCOM'05*, and *PTR*).

TABLE II
CUMULATED BETWEENNESS $H(s)$ OBTAINED FROM DIFFERENT PERCENTAGES OF RELAY NODES (SELECTING THOSE WITH THE HIGHEST BETWEENNESS), FOR THE DIFFERENT EXPERIMENTS.

Trace	**Percentage of nodes**		
	12.5 %	25 %	50 %
CAMBRIDGE	77.6 %	91.9 %	99.0%
INFOCOM	62.0 %	89.1 %	99.6%
MIT	53.7 %	70.1 %	88.9%
PTR	39.0 %	58.9 %	85.6%

successful delivery, suggesting the idea to take the path length into account in routing strategies. Finally, it is interesting to note that, for large k, the 12.5% of hubs provide a successful delivery rate of approximately 50% in all the considered experiments.

IV. CONCLUSION AND FUTURE WORK

In this work we have derived the probability of successful delivery in opportunistic networks in which only a subset of nodes works as message hub. We obtained these results starting by combining distribution of the shortest path length and measures of node centrality on human mobility traces. The results show the trade-off between the number of nodes used as relays during transmissions and the successful delivery rate.

This initial study shows that in such networks, when only a small portion of the population with high betweeness is enabled as forwarders, the success ratio of message forwarding remains high. However, we did not consider how the geodesic path length distribution actually changes when only a fixed subset of hubs is active. This further work can be done only experimentally, due to the difficulties to derive analytical results. Moreover, we are interested in extending this study to multiple-copy routing strategies.

REFERENCES

[1] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Pocket switched networks: Real-world mobility and its consequences for opportunistic forwarding," University of Cambridge, Tech. Rep. 617, 2005.

[2] D. Kempe, J. Kleinberg, and A. Kumar, "Connectivity and inference problems for temporal networks," *Comput. Syst. Sci.*, vol. 64, no. 4, pp. 820–842, 2002.

[3] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[4] ——, "Centrality in social networks: Conceptual clarification," *Social Networks*, vol. 1, pp. 215–239, 1979.

[5] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in *MobiHoc '07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2007, pp. 32–40.

[6] A. Chaintreau, P. Hui, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, 2007, fellow-Crowcroft, Jon.

[7] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: social-based forwarding in delay tolerant networks," in *MobiHoc '08: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2008, pp. 241–250.

[8] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *MobiHoc '09: Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2009, pp. 299–308.

[9] W. jen Hsu and A. Helmy, "On nodal encounter patterns in wireless lan traces," in *IEEE Int. Workshop on Wireless Network Measurement (WiNMee)*, 2006.

[10] D. Kotz, T. Henderson, and I. Abyzov, "CRAWDAD data set dartmouth/campus (v. 2004-12-18)," Downloaded from http://www.crawdad.org/dartmouth/campus, Dec. 2004.

[11] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255–268, 2006.

[12] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Pocket switched networks and human mobility in conference environments," in *Proceedings of ACM SIGCOMM'05 Workshops*, 2005, pp. 244–251.

[13] F. P. S. Gaito, G. Grossi and G. Rossi, "Experimental validation of a 2-level social mobility model in opportunistic networks," in *IFIP Wireless Day Conference (IFIP'08)*. IEEE press, 2008, pp. 1–5.