PHILOSOPHICAL TRANSACTIONS
OF
THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2008) **366**, 2005–2016
doi:10.1098/rsta.2008.0010
Published online 6 March 2008

Human mobility models and opportunistic communications system design

By Pan Hui and Jon Crowcroft*

Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

In this paper, we seek to improve understanding of the structure of human mobility, with a view to using this for designing algorithms for the dissemination of data among mobile users. We analyse community structures and node centrality from the human mobility traces and use these two metrics to design efficient forwarding algorithms in terms of delivery ratio and delivery cost for mobile networks. This is the first empirical study of community and centrality using real human mobility datasets.

Keywords: human mobility; community; centrality; forwarding algorithm; mobile network; delay tolerant network

1. Introduction

A mobile network has a dual nature: it is both a physical network and, at the same time, a social network. A node in the network is a mobile device and is also associated with a mobile human. Mobility traces capture interactions between nodes in a real environment and are very important for the mobile network study. We have been actively gathering the human mobility traces covering a rich diversity of environments from a busy metropolitan city to a quiet university town in the past 3 years to facilitate the study of mobile networks, social networks and epidemiology.

Encounters from the mobility traces can be used to build relationship graphs in the social networks. The nodes of the graphs are the physical nodes from the traces, the edges are the contacts and the weights of the edges are the values based on the metrics specified such as the number of contacts and contact duration. We call these graphs 'contact graphs' in this paper. We can measure the relationship between two people by how many times they meet each other and how long they stay with each other. We naturally think that if two people spend more time together or see each other more often, they are in a closer relationship. We use these contact graphs as a way to represent the mobility traces and to choose a threshold for community detection.

In a real network, it is well known that some nodes may be more highly connected to each other than to the rest of the network. The set of such nodes are called clusters, communities, cohesive groups or modules. Many different

* Author for correspondence (jon.crowcroft@cl.cam.ac.uk).

One contribution of 16 to a Discussion Meeting Issue 'Networks: modelling and control'.

approaches to community detection in complex networks have been proposed such as *k*-clique (Palla *et al.* 2005), betweenness (Newman & Girvan 2004) and modularity (Newman 2006). In this paper, we apply the *k*-clique community detection algorithm on the contact graphs and discover rich community structures in all the experimental datasets. The presence of clusters means that nodes are not equal in terms of their ability to relay data to other parts of the network and helps us design good strategies for information dissemination. *Community* is one social context we want to explore in this paper.

Centrality is a good measure for path finding in the social networks. Freeman (1977) defined several centrality metrics to measure the importance of a node to the network. 'Betweenness' centrality measures the number of times a node falls on the shortest path between two other nodes. In data forwarding, we normally prefer nodes with higher centrality values to lower centrality nodes. In this paper, we want to empirically verify the heterogeneity of *centrality* from the mobility traces.

The mobile network scenario we consider in this paper is called pocket switched networks (PSNs; Hui *et al.* 2005) that are a category of delay-tolerant networks (DTNs; Fall 2003) aimed at supporting applications for human-to-human communications, through the so-called store-and-forward paradigm. Previous DTN-routing algorithms (Lindgren *et al.* 2004; Jones *et al.* 2005) provide forwarding by building and updating routing tables whenever mobility occurs. We believe this approach is not cost effective for a PSN, since mobility is often unpredictable and topology changes can be rapid. Rather than exchanging much control traffic to create unreliable routing structures, we prefer to search for some characteristics of the network, which are less volatile than mobility. A PSN is formed by people and their social relationships probably vary much more slowly than the topology and therefore can be used for better forwarding decisions.

The contribution of this paper is to explore human mobility and interaction from the mobility traces, to understand heterogeneity at multiple levels of detail and to improve forwarding in the PSN focusing on the two social contexts: community and centrality, learnt from the interaction analysis.

2. Evaluating node relationships

We use four experimental datasets gathered by us for a period of 2 years, referred to as *Hong Kong*, *Cambridge*, *Infocom05* and *Infocom06*, and one other dataset from the MIT Reality Mining Project (Eagle & Pentland 2006), referred to as *Reality*.

Details about the experimental environments can be found in our separate technical report (Hui & Crowcroft 2007). Here, we explore further properties of the experimental scenarios and present statistics concerning the contact graphs for each dataset.

(*a*) *Contact duration and frequency*

We assume that contact duration indicates familiarity. Two people sharing the same office might hate each other and not talk, but we will ignore this kind of extreme situation here. The number of times two people meet each other

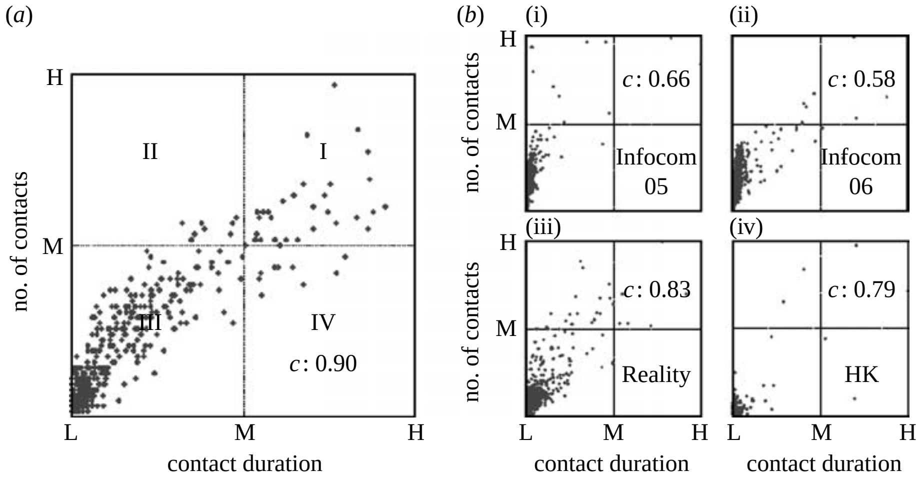

Figure 1. (*a*,*b*) Number of contacts versus the contact durations for all pairs in the four datasets (I, community; II, familiar strangers; III, strangers; IV, friends), with correlation coefficients.

implicitly reveals a periodicity of the relationship. In this work, we infer regularity of meetings from the number of contacts. Two people might meet a lot of times in a short period (e.g. a day) and then not at all. But, according to the datasets, there are very few such cases and here we will ignore these too as outliers.

Figure 1*a* shows the correlation between regularity and familiarity in the Cambridge dataset. Here, the regularity is positively correlated to the familiarity with a correlation coefficient of 0.9026. We define four kinds of relationships between a pair of nodes: Community; Familiar Strangers; Strangers; and Friends. A pair of nodes that has long contact duration (high familiarity) and large number of contacts (high regularity) belongs to the same community. A pair of nodes that meet regularly but do not spend time with each other are familiar strangers (Paulos & Goodman 2004) meeting everyday. People who do not meet regularly and do not spend time with each other are strangers. Finally, node pairs that do not meet very frequently but spend quite a lot of time together for each meeting are counted as friends. It is not necessary that the division of the four quarters is exactly at the middle; it acts only as a reference or example. A clear-cut division may need more empirical experimental results. But here we provide the methodology to classify these four kinds of relationship based on pure contact duration and frequency.

Figure 1*b* shows the correlation between the number of contacts and contact durations for the other four experiments. We can see that conference environments are quite similar, both with a narrow stripe in figure 1*b*. This stripe shows that people in the conference tend to meet each other more frequently, rather than spending a long time together, which is a typical conference scenario, since people may meet each other many times in coffee breaks, corridors, registration desk, etc. They may stand together and chat for a while and then shift to chat with others instead of spending all the time together. Infocom06 contains double the number of participants and hence more data points. The Reality set is similar to the Cambridge set, with most of the points lying on or above the diagonal line. However, it also seems that people have more contacts rather than spending time together. In the Hong Kong figure, we can find two pairs of friends, two pairs of close community members and two pairs of familiar strangers. All the other pairs lie in the strangers quarter. This is in line with our expectations for an experiment designed to contain little social correlation.

(*b*) *Node betweenness centrality*

In many mobility models such as the random waypoint, nodes are assumed, explicitly or implicitly, to have homogeneous speed distributions, importance and popularity. Our intuition is that the last two assumptions, at least, are not true. People have different levels of popularity: salespeople and politicians meet people frequently, whereas computer scientists may only meet a few of their colleagues once a year. Here, we want to employ heterogeneous popularity to help design more efficient forwarding strategies: we prefer to choose popular hubs as relays rather than unpopular ones.

Each mobility trace can be modelled as a temporal graph (network), a graph whose connectivity is time dependent. A temporal network is a kind of weighted network. The centrality measure in traditional weighted networks may not work here since the edges are not necessarily available concurrently. Hence, we need a different way to calculate the centrality of each node in the system. Our approach is as follows: first, we carried out a large number of emulations of unlimited flooding with different uniformly distributed traffic patterns created using the *HaggleSim* emulator (Hui & Crowcroft 2007).

Second, we count the number of times a node acts as a relay for other nodes on all the shortest delay deliveries. Here, the shortest delay delivery refers to the case when a single message is delivered to the destination through different paths, where we only count the delivery with the shortest delay. We call this number the ‘betweenness centrality’ of this node in this temporal graph. Of course, we can normalize it to the highest value found. Here, we use unlimited flooding since it can explore the largest range of delivery alternatives with the shortest delay. This captures the spirit of Freeman betweenness centrality (Freeman 1977).

Initially, we only consider a homogeneous communications pattern, in the sense that every destination is equally probable, and we do not weight the traffic matrix by locality. We then calculate the global centrality value for the whole system. We will analyse the heterogeneous system, once we have understood the community structure.

Figure 2 shows the number of times a node falls on the shortest paths between all other node pairs. We can simply treat this as the centrality of a node in the system. We observed a very wide heterogeneity in each experiment. This clearly shows that a small number of nodes have extreme centrality and thus high relaying ability and a large number of nodes have moderate or low centrality values, across all experiments. One interesting point from the Hong Kong data is that the node showing highest delivery power in the figure is actually an external node. This node could be some very popular hub for the whole city, e.g. a postman or a newspaper man in a popular underground station, which relayed a certain amount of cross-city traffic. The 30th and 70th percentiles and the means of normalized individual node centrality are shown in table 1.

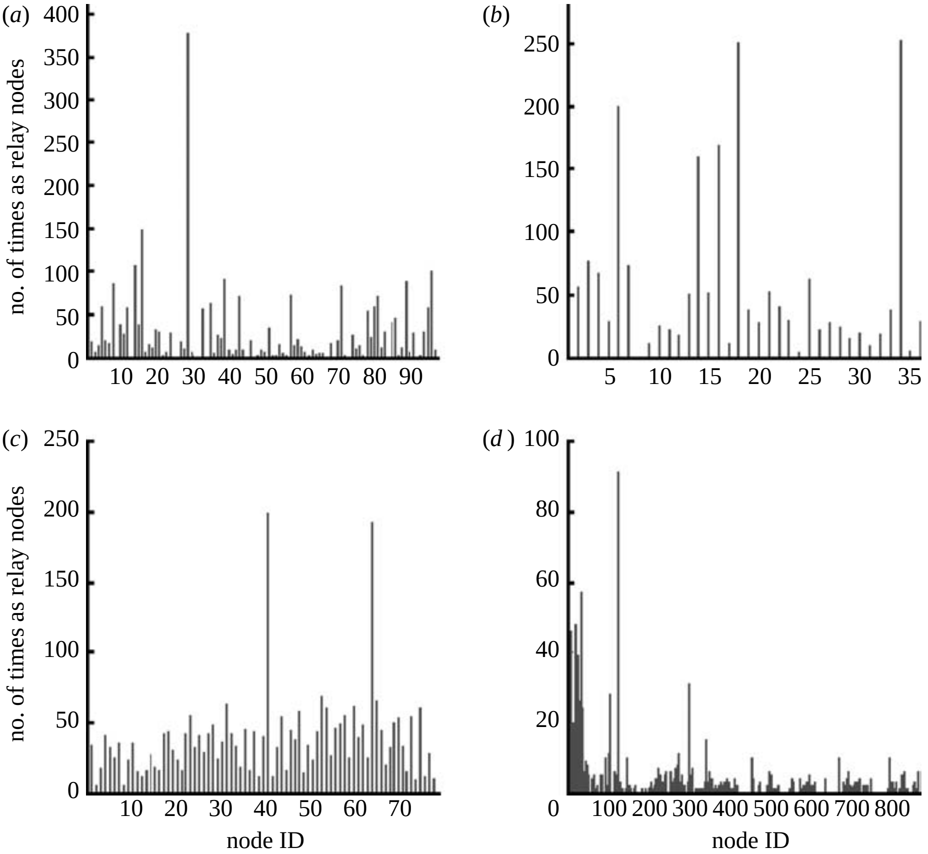

Figure 2. Number of times a node acts as a relay for others on four datasets. (*a*) Reality, (*b*) Cambridge, (*c*) Infocom06 and (*d*) Hong Kong.

Table 1. Statistics about normalized node centrality in four experiments.

experimental dataset	30th percentile	mean	70th percentile
Cambridge	0.052	0.220	0.194
Reality	0.005	0.070	0.050
Infocom06	0.121	0.188	0.221
Hong Kong	0	0.017	0

3. Apply node relationship graph to community detection

The aim of this section is to empirically prove the existence of community structures in the human mobility traces and show that these network structures correctly match the real world social structures. We apply k-clique community detection (Palla *et al.* 2005) to the node contact graphs that we introduced in the previous section. We have calculated all the results by using both contact duration and number of contacts on all the five experiments, but owing to space limitations we just show two cases here.

Palla *et al.* (2005) define a k-clique community as a union of all k-cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k-cliques, where two k-cliques are said to be adjacent if they share $k-1$ nodes. The value of k indicates the degree of mutual awareness of each node in a community. As k is increased, the k-clique communities shrink, but on the other hand become more cohesive since their member nodes have to be part of at least one k-clique.

(*a*) *k-clique communities in Reality mining*

This is a campus environment. Out of 100 participants, 75 are either students or faculty in the MIT Media Laboratory, while the remaining 25 are incoming students at the adjacent MIT Sloan Business School. Of the 75 users at the MIT Media Laboratory, 20 are incoming masters students and 5 are incoming MIT freshmen. First, we look at communities detected by using a contact threshold of 388 800 s or 4.5 days on the nine months Reality dataset. The threshold was obtained from assuming three lectures per week, four weeks per month and a total trace duration of nine months (2% of the total links are taken into consideration). Research students in the same office may stay together all day, so their contact duration threshold could be very large. For students attending lectures, this estimation should be reasonable. Using a looser threshold still detects the links with much stronger fit. We observe eight communities of size (16, 7, 7, 7, 6, 5, 4, 3) when $k=3$. When $k=4$, the 3-clique community is eliminated and other communities shrink or are eliminated, and only five communities of size (13, 7, 5, 5, 4) are left. All of these five communities are disjoint. When $k=5$, three communities of size (9, 6, 5) remain, the size-9 and the size-5 are split from the 13-sized in the 4-clique case. Moving to $k=6$ and 7, there are two communities and one community, respectively.

We are also interested in knowing about small groups that are tightly knit. We set a strict threshold of 648 000 s, that is 1 hour per weekday on average, four weeks per month and for a total of nine months. Approximately 1% of the links is taken into account for the community detection. When $k=3$, there are three disjoint communities of size (12, 7, 3). When $k=4$, there are only two communities left of size (8, 6). A single 7-clique community remains in $k=5$ and 6 cases; this 7-clique community is the same as in the 388 800 s case. These seven people could be people from a same research group, who know each other and spend long periods with each other.

(*b*) *k-clique conference communities*

In this section, we will show the community structures in a conference environment. Here we take Infocom06 as an example since it contains more participants than Infocom05 and more participant information. The total dataset covers only 3 days; hence we do not expect the threshold to be very big. People usually socialize during conferences in small groups so we expect clique sizes of 3, 4 or 5 to be reasonable. For Infocom06, the participants were specially selected so that 34 out of 80 form four subgroups according to academic affiliations. Out of these four groups, there were two groups from institutes in Paris with size of 4 and 10, respectively (named Paris group A and Paris group B), one group from Lausanne, Switzerland, of five people and one larger group with 15 people from

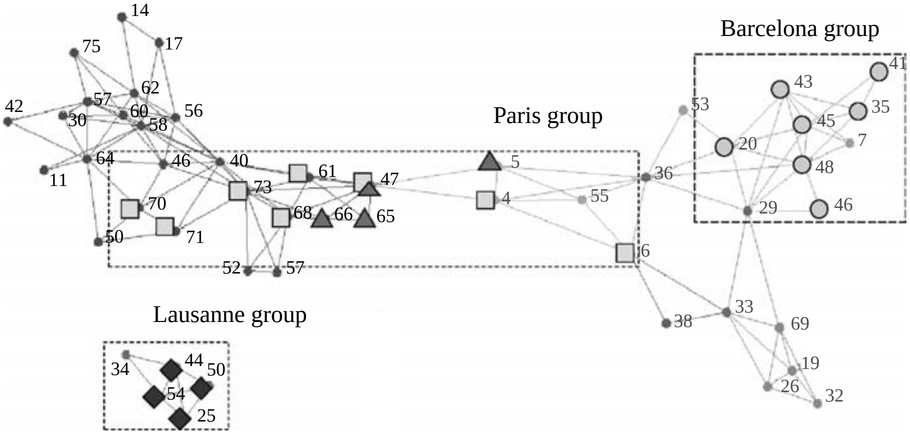

Figure 3. 3-clique communities based on contact durations with weight threshold that equals 20 000 s (Infocom06; circles, Barcelona group; squares, Paris group A; triangles, Paris group B; diamonds, Lausanne group).

the local organization in Barcelona. For this local organization group, the volunteers are from different local institutions and are responsible for different sessions in the conference, so we do not expect them to be all together. After collecting the data, all the personal information about the participants is deleted for privacy, except the node ID, the affiliation and the nationality.

Figure 3 shows the 3-clique communities with a threshold of 20 000 s, i.e. approximately 1.85 h d^{-1}. For the community calculation, 1.68% of all edges are taken into account. We observe six communities of size (25, 11, 6, 6, 5, 3) in this case. The size-25 overlaps at one node with size-6 that also overlaps with the size-11 community at one node and the size-3 at another node. The second size-6 community also overlaps the size-3 and size-11 at another two nodes. The size-5 community stands alone. Although we know that during a conference people from different sub-communities tend to mix together and hence the boundary of affiliation communities should become less clear, we still find hints of the original affiliation communities from the figure. The algorithm correctly classified the nodes belonging to the local organizers into a community (the Barcelona group at the r.h.s. of the figure) and the members of the Lausanne group into another community. There are several nodes that do not belong to these affiliations and are falsely classified into the same communities, but this also truly reflects the nature of a conference, to socialize with people in other institutions. The two Paris groups are also clearly identified as they tend to socialize with each other. Node 47 belongs to both groups and it is important to link these two groups together. There are many members in the size-25 group not belonging to a common institution but they are here linked together by different small groups mixing together in conference.

Rich community structures are also observed in the other experiments that also correctly match the real world social structures. This agrees with our assumption about the duality of the mobile networks (i.e. the physical network and the social network). We can infer social relationships from physical encounters and use this information to assist data forwarding.

4. Interaction and forwarding

(a) Introduction to algorithms

We have shown the existence of heterogeneity at the level of individuals and groups in all the mobility traces. This motivates us to consider a new heterogeneous model of human interaction and mobility.

- *Categories of human contact patterns.* Human relationships can be modelled by using the correlation of contact duration and the number of contacts. We define four types of human relationship based on the correlation of contact duration and number of contact.
- *Cliques and community.* We explored the community structures inside different social environments and found that these community structures match quite well to the real underlying social structures.
- *Popularity ranking.* We shall see that popular hubs are as useful in the PSN context as they are in the wireline Internet and in the Web.

We now conjecture how we can use this information to make smart forwarding decisions. The following three pre-existing schemes provide lower and upper bounds in terms of cost and delivery success.

- *WAIT.* Hold onto a message until the sender encounters the recipient directly.
- *FLOOD.* Messages are flooded throughout the entire system.
- *MCP* (Multiple-Copy-Multiple-Hop). Multiple copies are sent subject to a time-to-live hop count limit on the propagation of messages. By exhaustive emulation, the 4-copy-4-hop MCP scheme is found to be the most cost-effective scheme in terms of delivery ratio and cost for all naive schemes among all the datasets except the HK data. Hence, for fair comparison, we evaluate our algorithms against the 4-copy-4-hop MCP scheme.

All of these schemes are inefficient because they assume a homogeneous environment. If the environment is homogeneous, then every node is *statistically equivalent* and every node has the same likelihood of delivering the messages to the destination. As we showed in the first half of this paper, the environments and nodes are diverse, and hence here we want to design algorithms that make use of this rich heterogeneity.

Figure 4 shows the design space for the forwarding algorithms in this paper. The vertical axis represents the explicit social structure, that is, facets of nodes that can be specifically identified such as affiliation, organization or other social context. This is the social or human dimension. The two horizontal axes represent the network structural plane, which can be inferred purely from observed contact patterns. The Structure-in-Cohesive Group axis indicates the use of localized cohesive structure, and the Structure-in-Degree axis indicates the use of hub structure. These are observable physical characteristics. In our design framework, it is not necessary that physical dimensions are orthogonal to the social dimension, but since they represent two different design parameters we prefer to separate them. The design philosophy here is to consider both the social and physical aspects of mobility.

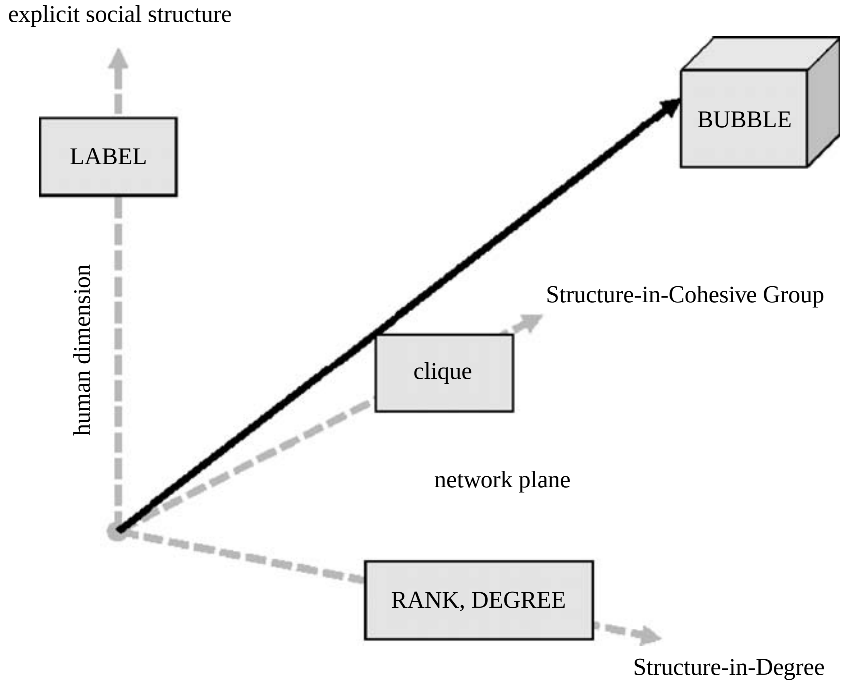

Figure 4. Design space for forwarding algorithms.

- *LABEL*. Explicit labels are used to identify forwarding nodes that belong to the same organization. Optimizations are examined by comparing the label of the potential relay nodes and the label of the destination node. This is in the human dimension, although an analogous version can be done by labelling a k-clique community in the physical domain.
- *RANK*. This is analogous to the degree of a node in a fixed network; we use a modified ranking scheme, namely the node betweenness centrality in a temporal network. A message is forwarded to nodes with higher centrality values than the current node. (This is a similar concept to the work by Adamic *et al.* (2001) in a fixed network.) It is based on observations in the network plane, although it also reflects the hub popularity in the human dimension.
- *DEGREE*. A heuristic based on the observed average of the degree of a node over some longer interval. Either the last interval window or a long-term cumulative estimate is used to provide a fully decentralized approximation for each node's centrality, and then it is used to select forwarding nodes. It is found that DEGREE can approximate RANK well in delivery ratio with slightly larger delivery cost.
- *BUBBLE*. The BUBBLE family of protocols use high centrality nodes to spread the messages out from the source and use community information to limit the scope of spreading. Messages will forward only to nodes with higher centrality values than the current node or in the same community as the destination. It is a combination of RANK and LABEL.

In the technical report (Hui & Crowcroft 2007), we show how we can make use of all these different metrics to improve forwarding performance in a heterogeneous system and also when they will fail. In this paper, we only show the BUBBLE algorithm that takes advantages of both community and centrality due to space constraints.

(*b*) *BUBBLE algorithm*

For the BUBBLE algorithm, we make two assumptions.

- Each node belongs to at least one community. Here we allow single node communities to exist.
- Each node has a global ranking (i.e. global centrality) across the whole system and also a local ranking within its local community. It may also belong to multiple communities and hence may have multiple local rankings.

Forwarding is carried out as follows. If a node has a message destined for another node, this node first *bubbles* the message up the hierarchical ranking tree using the global ranking, until it reaches a node that is in the same community as the destination node. Then, the local ranking system is used instead of the global ranking, and the message continues to bubble up through the local ranking tree until the destination is reached or the message expires. This method does not require every node to know the ranking of all other nodes in the system, but just to be able to compare ranking with the node encountered and to push the message using a greedy approach.

In order to evaluate different forwarding algorithms, we use the same *HaggleSim* emulator. For each emulation, 1000 messages are created, uniformly sourced between all node pairs. Each emulation is repeated 20 times with different random seeds for statistical confidence. For all the emulations conducted for this work, we have measured the following two metrics and compute the 95th percentile using *t*-distribution. In this paper, we show Reality experiment as an example, and the results about other experiments can be found in the technical report (Hui & Crowcroft 2007).

- *Delivery ratio*. The proportion of messages that have been delivered out of the total unique messages created.
- *Delivery cost*. The total number of messages (include duplicates) transmitted across the air. To normalize this, we divide it by the total number of unique messages created.

From figure 5*a*,*b*, we can see that of course FLOOD achieves the best for delivery ratio, but the cost is 2.5 times that of MCP and 5 times that of BUBBLE. On the other hand, WAIT has very low cost but it only has at most 10% delivery. BUBBLE is very close in performance to MCP in the multiple-group case as well, and even outperforms it when the time TTL of the messages is allowed to be larger than two weeks. However, the cost is only 50% compared with that of MCP.

In order to further justify the significance of social-based forwarding, we also compare BUBBLE with a benchmark 'non-oblivious' forwarding algorithm, PROPHET (Lindgren *et al.* 2004). The PROPHET uses the history of encounters

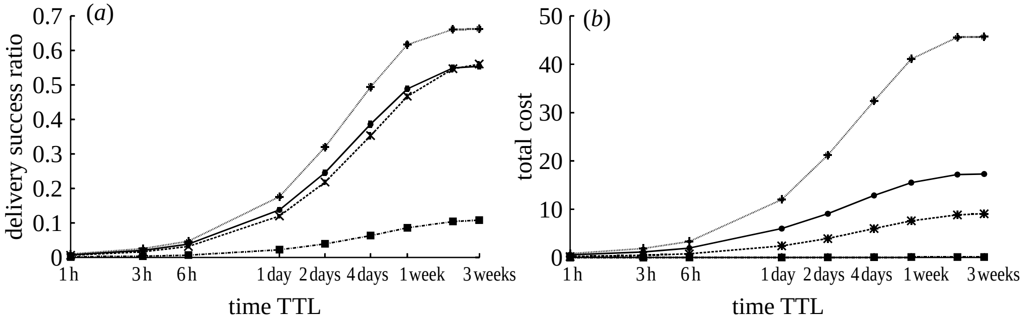

Figure 5. (*a*,*b*) Comparisons of several algorithms on Reality dataset (pluses, FLOOD; circles, MCP; crosses, BUBBLE; squares, WAIT).

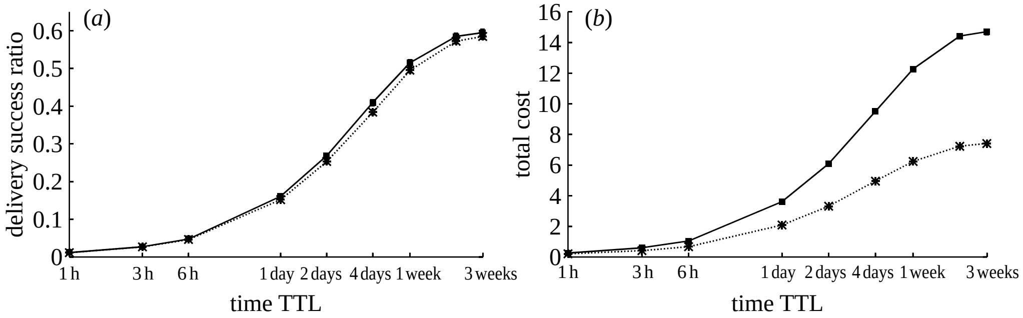

Figure 6. (*a*,*b*) Comparisons of BUBBLE (asterisks) and PROPHET (squares) on Reality dataset.

and transitivity to calculate the probability that a node can deliver a message to a particular destination. Since it has been evaluated against other algorithms before and has the same contact-based nature as BUBBLE (i.e. does not use location information), it is a good target to compare with BUBBLE.

Figure 6*a*,*b* shows the comparison of the delivery ratio and delivery cost of BUBBLE and PROPHET. Here, for the delivery cost, we only count the number of copies created in the system for each message as we have done before for the comparison with the oblivious algorithms. We did not count the control traffic created by PROPHET for exchanging routing table during each encounter, which can be huge if the system is large (PROPHET uses flat addressing for each node and its routing table contains an entry for each known node). We can see that most of the time BUBBLE achieves a similar delivery ratio to PROPHET with only half of the cost.

Considering that BUBBLE does not need to keep and update routing table for each node pair, the improvement is significant. Similar significant improvements by using BUBBLE are observed in other datasets, demonstrating the generality of the BUBBLE algorithm.

In the technical report (Hui & Crowcroft 2007), we show the limitations of the LABEL and the RANK algorithms and how BUBBLE was developed to improve on these limitations. We also propose two solutions to locally determine node centrality using approximation or knowledge of past centrality measure. We also survey work in the areas of forwarding algorithms and community detection in the same report.

5. Conclusion

We have explored the contact duration, contact frequency, node centrality and community structure of the human mobile networks using mobility traces. We have shown that these observations on the physical network correctly match the real world's social structure. We have proposed a family of forwarding algorithms that choose the next hop over which to relay packet, by making use of the *observed patterns* in encounters between nodes, as well as information known *a priori* about social relationships between the owners of nodes. We demonstrated that by incorporating this information, forwarding efficiency (in term of delivery ratio and delivery cost) can be significantly improved over the best oblivious forwarding strategy and the state-of-the-art history of encounter prediction algorithm.

This work was supported in part by the Haggle Project under the EU grant IST-4-027918. We would like also to acknowledge the comments from Steven Hand, Eiko Yoneki, Sid Chau, Andrea Passarella, Andrew Warfield, Richard Mortier and Pietro Lio.

References

Adamic, L. A., Huberman, B. A., Lukose, R. M. & Puniyani, A. R. 2001 Search in power law networks. *Phys. Rev. E* **64**, 46 135–46 143. (doi:10.1103/PhysRevE.64.046135)

Eagle, N. & Pentland, A. 2006 Reality mining: sensing complex social systems. *Pers. Ubiq. Comput.* **10**, 255–268. (doi:10.1007/s00779-005-0046-3)

Fall, K. 2003 A delay-tolerant network architecture for challenged internets. In *Proc. SIGCOMM*, pp. 27–34.

Freeman, L. C. 1977 A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41. (doi:10.2307/3033543)

Hui, P. & Crowcroft, J. 2007 Bubble rap: forwarding in small world DTNs in ever decreasing circles. UCAM-CL-TR-684, Computer Laboratory, University of Cambridge.

Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J. & Diot, C. 2005 Pocket switched networks and human mobility in conference environments. In *WDTN'05*, pp. 244–251.

Jones, E. P. C., Li, L. & Ward, P. A. S. 2005 Practical routing in delay-tolerant networks. In *SIGCOMM 2005. Workshop on delay tolerant networking and related topics*, pp. 237–243.

Lindgren, A., Doria, A. & Schelen, O. 2004 Probabilistic routing in intermittently connected networks. *SIGMOBILE Mob. Comput. Commun. Rev.* **7**, 19–20. (doi:10.1145/961268.961272)

Newman, M. E. J. 2006 Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577. (doi:10.1073/pnas.0601602103)

Newman, M. E. J. & Girvan, M. 2004 Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113. (doi:10.1103/PhysRevE.69.026113)

Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. 2005 Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818. (doi:10.1038/nature03607)

Paulos, E. & Goodman, E. 2004 The familiar stranger: anxiety comfort, and play in public places. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 223–230.