Self-management of routing on human proximity networks

Graham Williamson[1], Davide Cellai[1,2], Simon Dobson[1,2], and Paddy Nixon[1,2]

[1] Systems Research Group, School of Computer Science and Informatics, University College Dublin, Dublin, IE
[2] Lero, School of Computer Science and Informatics, University College Dublin, Dublin, IE

Abstract. Many modern network applications, including sensor networks and MANETs, have dynamic topologies that reflect processes occurring in the outside world. These dynamic processes are a challenge to traditional information dissemination techniques, as the appropriate strategy changes according to the changes in topology. We show how network dynamics can be exploited to design a self-organising data dissemination mechanism using only node-level (local) information, which detects and adapts to periodic patterns in the network topology. We demonstrate our approach against real-world human-proximity networks.

1 Introduction

Most networks in society and technology present a time dependent topology. Networks of friendships, phone calls, but also mobile devices, probes and satellites, are dynamic. Of course, networks often change slowly and can be effectively represented by a static model, or they change in a precise direction (e.g. they grow) and their dynamic behaviour is well understood. There are cases, however, of highly dynamic networks, where change consists in a complex rewiring of the edges. A good example is the network formed by mobile objects (such as mobile phones) with short range communications. This example is particularly interesting, because the fast growing complexity and diffusion of mobile devices is envisioned to lead to a scenario where a number of applications will be used and shared by people living in the same city (Pocket Switched Networks). If these devices can communicate with each other ad hoc, without infrastructure, the result is a proximity network. For these reasons an interesting question is to understand when such networks can support stable applications such as information dissemination. In recent years, there has been growing interest in information propagation on dynamic networks. In particular, delay tolerant networking (DTN) protocols are designed to work in challenged or sparse networks. Recent papers have approached the problem of formulating efficient protocols for these types of networks [1–3]. Hui et al. [4] have proposed an efficient algorithm named BUBBLE, which exploits the popularity of a node to provide a quick and parsimonious way for a message to reach its destination.

However, most of these methods rely on either the particular characteristics of the considered experiment, or global information which will not presumably be available in real applications. Given the nature of this type of network, it appears sensible to design protocols with significant self-management capabilities.

In this paper we explore different data sets and investigate the importance of local properties for data communication. Then we propose a dissemination protocol which is defined locally at node level and does not imply global knowledge of the network. Finally, we define a self-managing mechanism allowing the nodes to adapt their dissemination strategy based on the detection of periodic patterns.

2 Analysis of experimental data

2.1 Data sets

In this paper we take advantage of two very interesting data sets: the *Reality* experiment, performed at MIT [5, 6], and the *Cabspotting* data set [7, 8].

In the Reality experiment, 103 smart phones are assigned to 97 people (mostly among undergraduate and graduate students, but also staff at MIT), who carry them along every day. The smart phones detect other smart phones or any discoverable bluetooth device every 5 minutes. In this way, a network of proximity based encounters is built at any time. The experiment lasts 9 months, covering the terms of the academic year 2004-05.

In the Cabspotting experiment, the positions of 536 taxi cabs where tracked for about a month in the city of San Francisco. The positions were recorded as GPS coordinates at intervals of approximately 10 s. The proximity of cabs can be calculated on the basis of their movement patterns. The network formed by connecting cabs at a distance of $10m$ or less is already quite dense for the purpose of testing our protocol, and thus we will not consider larger communication ranges in this paper.

2.2 Network connectivity

For communication purposes, it is clear that some nodes may be more important than others. For example, in a static network nodes can be considered more important if many shortest paths between node pairs pass by them. This concept was rigorously formulated about 30 years ago with the definition of *betweenness centrality* [9]. Moreover, it is quite well known that social networks are characterized by a community structure, where nodes within a community are very well connected, whereas few edges link different communities [10]. Therefore, it is clear that nodes communicating among different communities have an important role in information dissemination.

In this paper, the *degree* of a node is defined as the number of distinct nodes encountered in a certain time interval. This quantity is therefore different from the number of encounters, because many edges can come and go between the

same pair of nodes within a given time. The importance of the time scale in calculating the degree is addressed in Section 5.

We now show and comment results from the analysis of the Reality data set. Fig. 1(a) shows the daily behaviour of the node degree (the number of distinct nodes encountered in a day) over the duration of the experiment. We observe a significant difference in the activity, with lower degrees during public holidays, as well as strong oscillations for different working days. Since the global activity

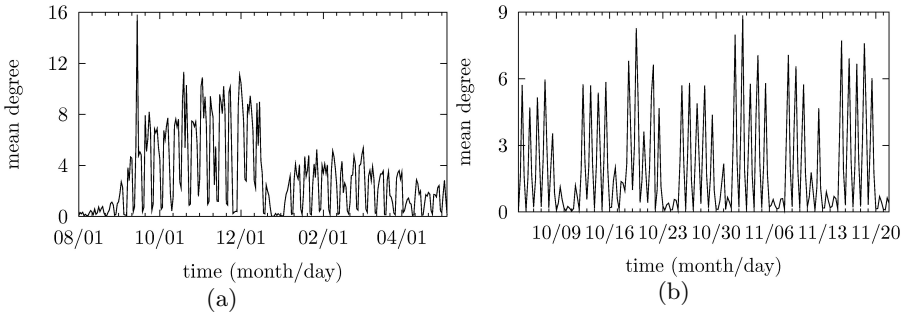

Fig. 1. (a) Daily degree vs time. (b) 6-hour degree vs time in a 7 week interval. The trace can be highly variable, but there are long (several weeks) periods with regular behaviour.

of the trace is so heterogeneous, it is useful to focus on some regular parts such as the one represented in Fig. 1(b), where node degree calculated in time slots of 6 hours is plotted vs time. As shown by the plot, week-ends are clearly recognizable as well as the alternation between day and night.

Most of the time the examined network is sparse, as different individuals are involved in different activities, usually far apart. It is then crucial to study the intrinsic capability of the network to support data communication. It is worth pointing out that it is possible that, within a given time, a message can be routed from node i to node j, but not *vice versa*. Hence, we must always consider ordered pairs of nodes. A preliminary question is to establish the maximum theoretical communications capabilities of the network, regardless of efficiency concerns. Thus, we define a quantity called *deliverability*: given a maximum delivery time, the ratio between all the ordered pairs of nodes for which message delivery is possible (i.e. it exists at least one time-dependent path connecting the source with the destination), divided by the total number of ordered pairs. Message delivery is established by unlimited flooding, which is the most effective way (but not the most efficient, of course). As we consider protocols where messages are not duplicated, but there is only a single copy in the network at any time, a useful quantity is the *delivery ratio*, defined as the fraction of distinct delivered messages over all the possible pairs source/destination within a given time. By definition, then, deliverability represents an upper bound on the delivery ratio of any routing protocol for a given network in a given time.

In Fig. 2 we show the comparison of the deliverability ratios between two different 3-week periods. It is interesting to note that week-ends constitute a real barrier for relatively fast deliveries: if the maximum delivery time is less than 4 days, there are days with very poor performance. On the whole, the deliverability is also quite low for higher waiting times, being around 0.5, but it stays constant over different days of the week. The comparison between the two periods also gives an interesting insight into the relationship between deliverability and short-time degree. The two periods only differ for the amount of activity (number of encounters, number of people involved in encounters, etc.), which is noticeably higher in the first set. We observe that this difference mostly affects the short delay deliveries (less than 3 days), whereas the deliverability after one week is more similar (the difference is about 0.1). This means that over time scales of about a week individuals get close to people from other communities and thus improve the deliverability even in periods of low activity. From the perspective of self-management, the important point is that the network possesses characteristic and recurrent dynamical features.

2.3 Network correlations

We examine the time dependence of some local properties of the nodes. To begin with, we divide the time into 6-hour slots, because, as observed by some authors [4], human daily routine can be divided into periods of activity which can be treated as roughly 6 hours long. We consider the aggregate graph of all the sightings happening in each time slot and calculate the betweenness centrality of each node, according to the Freeman algorithm [9]. In Fig. 3(a) we plot the time dependence of the Pearson correlation coefficient (pcc) of betweenness and node degree. We observe that there is a weak positive linear correlation between these two quantities: 0.653 ± 0.18. This means that, as an approximation, we can associate a high 6-hour degree to a high centrality of the node in the network, and *vice versa*. However, the value of the pcc has large oscillations with time with a standard deviation of 0.18.

This highly changing behaviour is not related to the incidence of holidays or singular events which may add noise to the measure. In fact, looking in Fig. 3(b) at the same curve restricted to the more regular 7-week period mentioned above, it emerges that the standard deviation of the pcc does not shrink at all, meaning that the oscillations are entirely due to the alternation of day/night and working days/week-ends. However, we observe a drop of the mean value of the correlation to 0.58, probably due to the occasional presence of starry structures during periods of low activity. (An explanation could be that, especially at the beginning of the experiment, some people were still not using the equipment properly, perhaps forgetting to enable the detection of other devices. If only one in a group is correctly recording neighbouring nodes, but all the others are not, the result would exactly be a star.)

The time dependence of the pcc between betweenness and number of connections in a time slot, shows a very similar behaviour, meaning that in a 6-hour

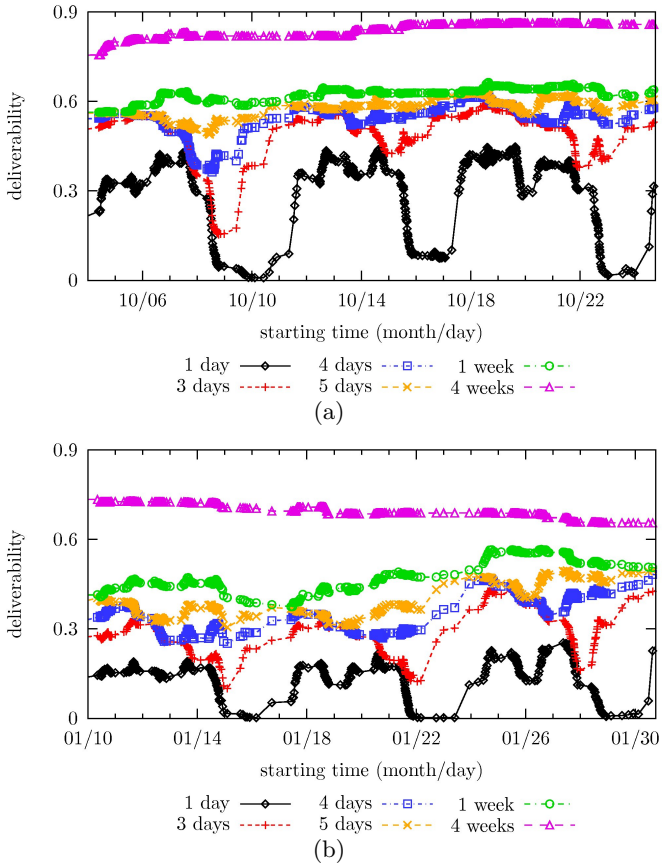

Fig. 2. Deliverability ratio based on flooding: comparison between two 3-week periods. It is necessary to allow at least 3 days for overcoming the drop of performance due to the small activity during week-ends. This comparison shows the similarity of the qualitative behaviour in two different periods of the data set.

interval most encounters are distributed among the encountered nodes, so that there is not so much difference between number of sightings and degree.

As in Hui et al. [4], we calculate the *centrality* of a node as the number of times the node is on the shortest path between every ordered pair of nodes. If there are multiple shortest paths between two nodes, we divide the contribution of this value by the number of the paths. We compare the correlation between this centrality and other significant quantities over the 3 weeks starting with October 4th, 2004. We find that the correlation with betweenness, degree, and number of encounters is 0.41, 0.57, and 0.65, respectively. This implies that if we assume this notion of centrality as the most expressive quantity of the importance of a node in data dissemination, either the degree or the number of

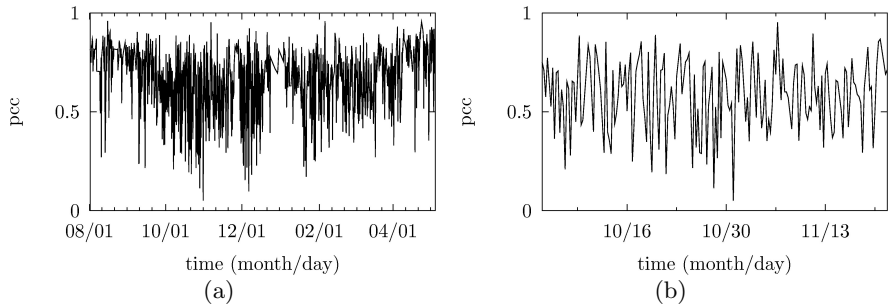

Fig. 3. (a) 6-hour pcc between betweenness centrality and node degree vs time. (b) 6-hour pcc between betweenness and degree vs time in a 7 week interval.

encounters seem to be the best candidates as local approximations. This means that a network protocol may, by making *local* observations, adapt its behaviour in a way that correlates strongly with an important *global* dynamic property of the network topology.

3 Dissemination algorithms

A number of algorithms have been formulated to efficiently disseminate information on a dynamic network. The specific problem we consider is to send a message from node i to node j, for arbitrary i and j. The goal is to achieve a high fraction of successful deliveries, as well as low overhead, within a given time.

The most effective method consists of flooding the network with an arbitrarily large number of message copies. Flooding is characterized by the following policy: when a node receives a message it immediately broadcasts copies to all current neighbours. It also forwards copies of the message to all nodes it comes into contact with for the remaining duration of the simulation. This has a huge cost, due to the high number of redundant transmissions. Multiple-Copy-Multiple-Hop (MCP) consists of a type of limited flooding, where it has been established a maximum number of message copies and number of hops a message is allowed. Choosing a suitable number of copies and hops allows us to tune the efficiency of the algorithm [11].

Many protocols deal with dissemination in ad hoc networks. We have, for example, PROPHET, which uses knowledge of the history of encounters of a node and the clustering coefficient to route a message based on the probability that a node will lead to the destination [12]. Directed Diffusion [13] tackles dissemination by setting up communication gradients over which information is routed towards interested nodes. Protocols such as Trickle [14] use epidemic based routing to provide practical dissemination protocols.

Finally, we focus on BUBBLE, which uses both a measure of centrality, and the community of the destination as a rationale which routing is based on [4].

Centrality is defined as the number of times a node appears on a shortest (in terms of number of hops) path between two other nodes, normalised to the highest score. As an approximation of this notion of centrality, the node degree calculated over a suitable time interval is more practical.

4 Metric-based routing and community structure

We now examine some important characteristics of the BUBBLE protocol, which we take as a benchmark. We use the *delivery ratio* as a measure of the performance of a dissemination process. We consider all the possible ordered pairs (i, j) of nodes and run for each of them a dissemination protocol designed to send a message from node i to node j. The delivery ratio at time t is defined as the ratio between the number of delivered messages after time t divided by the total number of messages (which equals the number of ordered pairs in the network). In Fig. 4(a) we plot the delivery ratio versus time in a 3-week period. The performance of the BUBBLE algorithm (with routing based on the pre-calculated centrality) is compared with routing without community knowledge. Quite remarkably, the behaviour of BUBBLE is very similar to an algorithm of routing based only on node degree (calculated over the previous 6 hour slot). This means that the community structure does not play a major role in improving the delivery ratio. It is interesting to note that applying the locally based degree routing to the algorithm, BUBBLE performs even better. Indeed, it emerges that the 6-hour degree, at least for this dataset, is the node property which best captures the importance of a node in data propagation.

In Fig. 4(b) we show the cost of the dissemination. Cost is defined as the total number of message hops from one node to another per message. We observe that centrality based BUBBLE is very efficient, with very low cost. However, centrality computation implies global knowledge of the network. Degree based BUBBLE is characterised by both a high delivery ratio and a significant cost, probably meaning that a portion of high degree nodes do not improve dissemination. The two plots show that centrality based routing leads to a large improvement in the cost, quite independently from the community structure. The cost advantage in introducing community routing into degree based BUBBLE is significant, but less important.

5 Self-management of routing

5.1 Definition of the self-management algorithm

Now we want to develop a self-managing mechanism able to choose the best strategy for routing optimization. In order to do that, we have to provide a way to automatically detect the time scale which allows the best routing performance. In fact, we have seen that routing based on the 6-hour degree achieves good performance both in terms of delivery ratio and cost. However, the choice

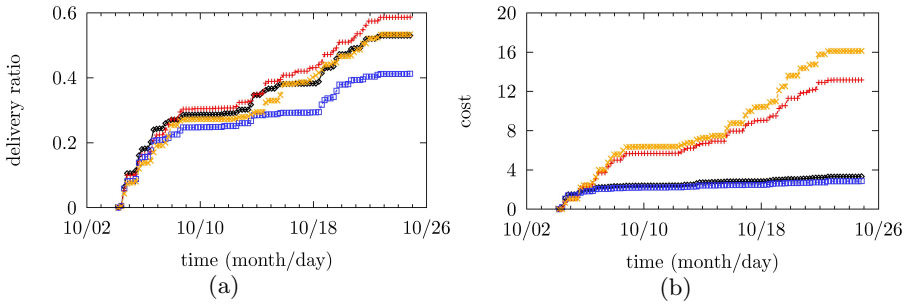

Fig. 4. Comparison of different implementations/modifications of the BUBBLE algorithm: delivery ratio (a) and cost (b) is plotted versus time over a 3-week period. Cost is defined as the total number of hops divided by the total number of messages sent between ordered pairs of nodes. Symbols refer to the same simulations in both figures. BUBBLE based on 3-day communities and centrality routing (+), is compared with the same algorithm based on 6-hour degree routing (×). The performance of a simple protocol based on centrality routing regardless community awareness (∗) leads to quite low delivery ratio, both with respect to 6-hour degree routing (□) – basically BUBBLE without community awareness – and the original BUBBLE. Cost is higher for degree than centrality routing.

of the 6-hour time scale seems quite arbitrary and relies on external considerations. We want instead to be able to find an intrinsic mechanism to detect the most appropriate time scale. As human mobility follows periodic patterns, we expect this to affect also the related proximity network. We also expect that the degree calculated over a characteristic periodicity of network dynamics can achieve better results in the dissemination protocol, because it captures better the repetitive behaviour of a node.

Therefore, we can state a self-managing policy of locally adapting the routing rule to achieve good efficiency. So, each node acts according to the following algorithm:

1. During a preliminary time interval Δt_0, each node calculates the main periodicity T in the number of contacts with other nodes.
2. The node calculates its degree over the time scale T obtained in the previous step.
3. Whenever a message reaches the node, the message is forwarded as soon as the node encounters a node with a higher degree metric. If instead the node encounters the destination, the message is delivered and the algorithm ends.
4. After n cycles, the node calculates again its periodicity T and goes on from point 2.

This algorithm only depends on two external parameters: the initial interval Δt_0 and the number of cycles n before calculating the dominant period again. Δt_0 can be based on some external considerations, including the requested maximum delivery time (e.g. a week may be a good choice for human activities). It

is important to underline, though, that the characteristic period is re-calculated every n cycles, so that the value can converge to an optimal period, hence the assigned value to Δt_0 is not particularly relevant, and could be set equal to the requested maximum delivery time. The number of cycles n should be at least about 10, so that the period can be calculated on a time scale an order of magnitude larger than the previous period T. The period can be practically determined by calculating the highest peak of the Fourier transform of the number of connections over time. In this way, each node may calculate a different period, and therefore base its routing according to a different time scale degree, the one which best captures the periodic activities of the node.

5.2 Algorithm evaluation

In order to evaluate the algorithm, we first have to investigate the importance of the degree criterion at different time scales. As a measure of efficiency, we calculate the ratio between delivery ratio and cost. This quantity can summarize the merit of a given protocol. In Fig. 5 and 6 we show the efficiency of routing based on node degree aggregated on different time scales for the Reality and the Cabspotting data sets. The plots show that the general behaviour is that higher efficiency corresponds to higher time scale. This can be explained by the fact that longer times allow to average over a larger number of events, and then to give a better estimate of the future importance of the considered node. However, Fig. 5 also shows that for the Reality data set this behaviour is not monotonic and there are time scales better than others. In particular, it appears that one day degree routing is more efficient than routing on a time scale of 2 or 3 days. 7 day routing obtains an even better performance than routing based on centrality (a property which implies global knowledge of network evolution). This behaviour is less important in the Cabspotting data set, but we can still notice that 1-day degree routing is sometimes more efficient than the 2-day degree routing.

In order to investigate the origin of this effect we look at the periodicities in node contacts. In fact, both the Reality and the Cabspotting datasets have periodicities in the hourly number of contacts. In Fig. 7 we show that by calculating the Fourier transform of the number of node encounters per hour. The observed peaks are due to the periodicity of human activities, and in fact the largest ones occur at 6 hours, 1 and 7 days for the Reality data, and 12 hours and 1 day for the cabspotting data set. Thus, our algorithm detects from the highest peak of the Fourier transform that the most important periodicity in the Reality data set is 1 day, and routes messages based on the 1-day degree, which we have seen being a particularly efficient metric. Similarly, in the Cabspotting case, the algorithm chooses 12-hour degree routing. It can also be shown that most nodes detect the same main periodicity.

Therefore, our interpretation is that the presence of periodicities in the patterns of activity of the nodes has an effect in improving corresponding time scales in degree routing. The difference between the two data sets shown in Figures 5 and 6 can be explained by the fact that this periodicity effect is more important in networks which are sparse most of the time (as the Reality network), where

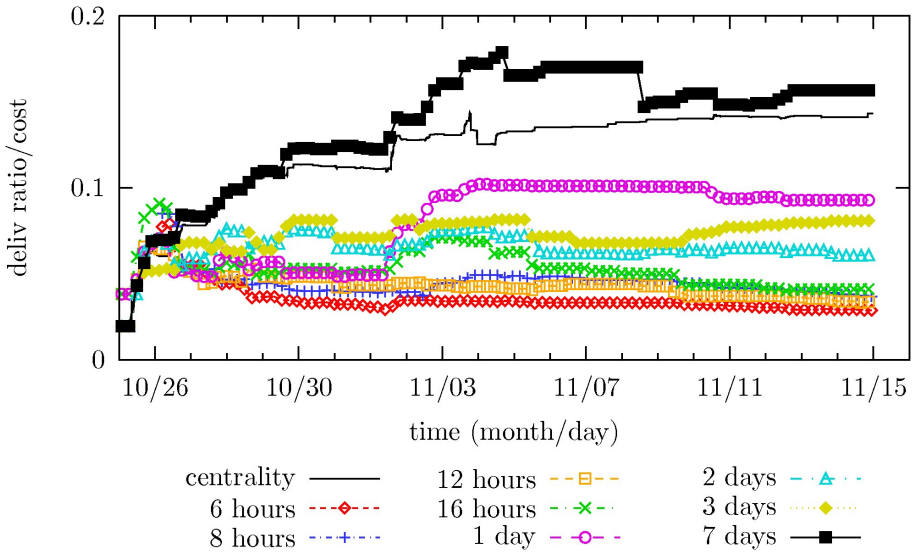

Fig. 5. Efficiency (defined as delivery ratio divided by cost) of routing based on the degree calculated over different time intervals for the Reality data set.

the choice of a smart routing policy is more critical, than in networks where the connectivity is generally good most of the time, as in the Cabspotting data set.

As we have seen, the proposed algorithm generates a routing protocol which is highly efficient and does not rely on external assumptions. The only parameter to fix is the duration of the preliminary interval, but it will change after n cycles if there is a better one in the system, or if the period of the node itself changes with time.

6 Conclusions

In this paper, we have investigated different approaches to data dissemination in dynamic human proximity networks. We have found that node degree is the best local property on which to base routing, and that there are time scales at which the protocol performs better. We have then formulated a self-management scheme where nodes automatically detect the best time scale and forward messages in an efficient way.

The significance of this approach is that it provides a mechanism by which to adapt data dissemination to the properties of the external processes affecting network dynamics, without having an explicit model of those dynamics embedded within the system. This makes the scheme purely topological and able to adapt autonomously to changing dynamics. Further work is needed to validate the approach against other kinds of dynamic networks (for example in environ-

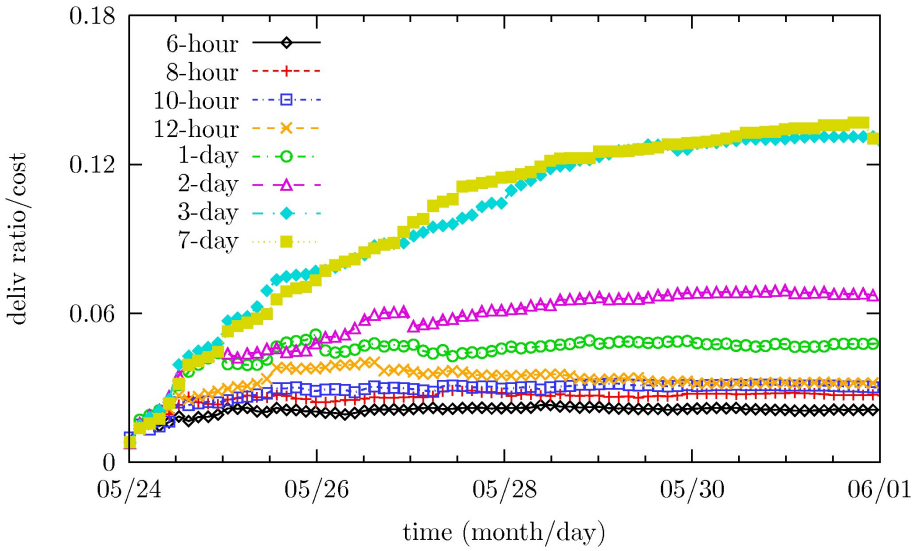

Fig. 6. Efficiency of routing based on the degree calculated over different time intervals for the the Cabspotting data set.

mental sensing), and to explore further local topological metrics that may be indicative of global properties useful for self-management.

Acknowledgements

This work is supported by Science Foundation Ireland under grant 07/CE/I1147, and 03/CE2/I303-1, "Lero: the Irish Software Engineering Research Centre". The authors thank the anonymous reviewers for insightful comments.

References

1. Fall, K.: A delay-tolerant network architecture for challenged internets. In: SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, ACM (2003) 27–34
2. Zhao, W., Ammar, M., Zegura, E.: A message ferrying approach for data delivery in sparse mobile ad hoc networks. In: MobiHoc '04: Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing, ACM Press (2004) 187–198
3. Pelusi, L., Passarella, A., Conti, M.: Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. Communications Magazine, IEEE **44**(11) (2006) 134–141
4. Hui, P., Crowcroft, J., Yoneki, E.: Bubble rap: social-based forwarding in delay tolerant networks. In: Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), ACM (2008) 241–250

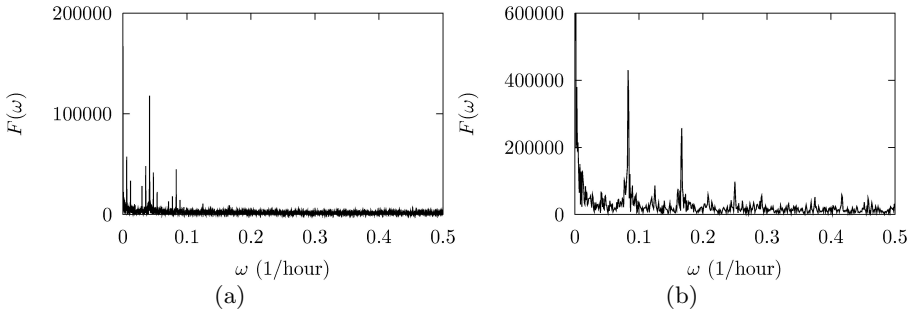

Fig. 7. Fourier transform of the time evolution of the number of connections, defined on time slots of 1 hour. The frequency ω has been normalized, so that a peak in $\bar{\omega}$ corresponds to a period $T = 1/\bar{\omega}$. The used data sets are Reality (a) and Cabspotting (b), respectively.

5. Eagle, N., Pentland, A.S.: CRAWDAD data set mit/reality (v. 2005-07-01). Downloaded from http://crawdad.cs.dartmouth.edu/mit/reality (July 2005)
6. Miklas, A., Gollu, K., Chan, K., Saroiu, S., Gummadi, K., de Lara, E.: Exploiting social interactions in mobile systems. In: UbiComp 2007: Ubiquitous Computing. (2007) 409–428
7. Piorkowski, M., Sarafijanovic-Djukic, N., Grossglauser, M.: CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility (February 2009)
8. Piorkowski, M., Sarafijanovic-Djukic, N., Grossglauser, M.: A parsimonious model of mobile partitioned networks with clustering. In: COMSNETS 2009: Communication Systems and Networks and Workshops. (2009) 1–10
9. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**(1) (1977) 35–41
10. Onnela, J.P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., Barabasi, A.L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences **104**(18) (May 2007) 7332–7336
11. Wang, Y., Jain, S., Martonosi, M., Fall, K.: Erasure-coding based routing for opportunistic networks. In: WDTN '05: Proceeding of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking, ACM Press (2005) 229–236
12. Lindgren, A., Doria, A., Schelén, O.: Probabilistic routing in intermittently connected networks. Service Assurance with Partial and Intermittent Resources (2004) 239–254
13. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking, ACM (2000) 56–67
14. Levis, P., Brewer, E., Culler, D., Gay, D., Madden, S., Patel, N., Polastre, J., Shenker, S., Szewczyk, R., Woo, A.: The emergence of a networking primitive in wireless sensor networks. Commun. ACM **51**(7) (2008) 99–106