Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs

Elizabeth M. Daly and Mads Haahr

***Abstract*—Message delivery in sparse Mobile Ad hoc Networks (MANETs) is difficult due to the fact that the network graph is rarely (if ever) connected. A key challenge is to find a route that can provide good delivery performance and low end-to-end delay in a disconnected network graph where nodes may move freely. We cast this challenge as an information flow problem in a social network. This paper presents social network analysis metrics that may be used to support a novel and practical forwarding solution to provide efficient message delivery in disconnected delay-tolerant MANETs. These metrics are based on social analysis of a node's past interactions and consists of three locally evaluated components: a node's 'betweenness' centrality (calculated using ego networks) and a node's social 'similarity' to the destination node and a node's tie strength relationship with the destination node. We present simulations using three real trace data sets to demonstrate that by combining these metrics delivery performance may be achieved close to Epidemic Routing but with significantly reduced overhead. Additionally, we show improved performance when compared to PRoPHET Routing.**

***Index Terms*—Delay & Disruption Tolerant Networks, MANETs, Sparse Networks, Ego Networks, Social Network Analysis**

I. Introduction

A Mobile Ad hoc Network (MANET) is a dynamic wireless network with or without fixed infrastructure. Nodes may move freely and organise themselves arbitrarily [9]. Sparse Mobile Ad hoc Networks are a class of ad hoc networks in which the node population is sparse, and the contacts between the nodes in the network are infrequent. As a result, the network graph is rarely, if ever, connected and message delivery must be delay-tolerant. Traditional MANET routing protocols such as AODV [45], DSR [27], DSDV [46] and LAR [29] make the assumption that the network graph is fully connected and fail to route messages if there is not a complete route from source to destination at the time of sending. One solution to overcome this issue is to exploit node mobility in order to carry messages physically between disconnected parts of the network. These schemes are sometimes referred to as *mobility-assisted routing* that employ the *store-carry-and-forward* model. Mobility-assisted routing consists of each node independently making forwarding decisions that take place when two nodes meet. A message gets forwarded to encountered nodes until it reaches its destination.

Current research supports the observation that encounters between nodes in real environments do not occur randomly [24] and that nodes do not have an equal probability of encountering a set of nodes. In fact, one study by Hsu and Helmy observed that nodes never encountered more than 50% of the overall population [23]. As a consequence, not all nodes are equally likely to encounter each other, and nodes need to assess the probability that they will encounter the destination node. Additionally, Hsu and Helmy performed an analysis on real world encounters based on network traffic traces of different university campus wireless networks [22]. Their analysis found that node encounters are sufficient to build a connected relationship graph, which is a small world graph. Therefore, social analysis techniques are promising for estimating the social structure of node encounters in a number of classes of disconnected delay-tolerant MANETs (DDTMs).

Social networks exhibit the small world phenomenon which comes from the observation that individuals are often linked by a short chain of acquaintances. The classic example is Milgrams' 1967 experiment, where 60 letters were sent to various people located in Nebraska to be delivered to a stockbroker located in Boston [40]. The letters could only be forwarded to someone whom the current letter holder knew by first name and who was assumed to be more likely than the current holder to know the person to whom the letters were addressed. The results showed that the median chain length of intermediate letter holders was approximately 6, giving rise to the notion of 'six degrees of separation'. Milgram's experiment showed that the characteristic path length in the real world can be short. Of particular interest, however, is that the participants did not send on the letters to the next participant randomly, but sent the letter to a person they perceived might be a good carrier for the message based on their own local information. In order to harness the benefits of small world networks for the purposes of message delivery, a mechanism for intelligently selecting good carriers based on local information must be explored. In this paper we propose the use of social network analysis techniques in order to exploit the underlying social structure in order to provide information flow from source to destination in a DDTM which extends on the authors' previous work [10].

The remainder of this paper is organized as follows: Section II reviews related work in the area of message delivery in disconnected networks. Section III examines network theory that may be applied to social networks along with social network analysis techniques. Section IV discusses SimBetTS, an example routing protocol which applies these techniques for routing in DDTMs. Section V evaluates the performance of the protocol along with a performance comparison between SimBetTS Routing and Epidemic Routing [51] and the PRoPHET Routing protocol [36] using three real trace data sets from the Haggle project [8], [24]. We conclude in Section VI.

Distributed Systems Group, Trinity College Dublin, Ireland

II. Related Work

A number of projects attempt to enable message delivery by using a virtual backbone with nodes carrying the data through disconnected parts of the network [15], [47]. The Data MULE project uses mobile nodes to collect data from sensors which is then delivered to a base station [47]. The Data MULEs are assumed to have sufficient buffer space to hold all data until they pass a base station. The approach is similar to the technique used in [2], [15], [17]. These projects study opportunistic forwarding of information from mobile nodes to a fixed destination. However, they do not consider opportunistic forwarding between the mobile nodes.

'Active' schemes go further in using nodes to deliver data by assuming control or influence over node movements. Li et al. [32] explore message delivery where nodes can be instructed to move in order to transmit messages in the most efficient manner. The message ferrying project [54] proposes proactively changing the motion of nodes in order to meet a known 'message ferry' to help deliver data. Both assume control over node movements and in the case of message ferries, knowledge of the paths to be taken by these message ferry nodes.

Other work utilises a time-dependent network graph in order to efficiently route messages. Jain et al. [26] assume knowledge of connectivity patterns where exact timing information of contacts is known, and then modifies Dijkstra's algorithm to compute the cost edges and routes accordingly. Merugu et al. [39] and Handorean et al. [21] likewise make the assumption of detailed knowledge of node future movements. This information is time-dependent and routes are computed over the time-varying paths available. However, if nodes do not move in a predictable manner, or are delayed, then the path is broken. Additionally, if a path to the destination is not available using the time-dependent graph, the message is flooded.

Epidemic Routing [51] provides message delivery in disconnected environments where no assumptions are made in regards to control over node movements or knowledge of the network's future topology. Each host maintains a buffer containing messages. Upon meeting, the two nodes exchange summary vectors to determine which messages held by the other have not been seen before. They then initiate a transfer of new messages. In this way, messages are propagated throughout the network. This method guarantees delivery if a route is available but is expensive in terms of resources since the network is essentially flooded. Attempts to reduce the number of copies of the message are explored in [44] and [49]. Ni et al. [44] take a simple approach to reduce the overhead of flooding by only forwarding a copy with some probability $p < 1$, which is essentially randomized flooding. The Spray-and-Wait solution presented by Spyropoulos et al. [49] assigns a replication number to a message and distributes message copies to a number carrying nodes and then waits until a carrying node meets the destination.

A number of solutions employ some form of 'probability to deliver' metric in order to further reduce the overhead associated with Epidemic Routing by preferentially routing to nodes deemed most likely to deliver. These metrics are based on either contact history, location information or utility metrics. Burgess et al. [7] transmit messages to encountered nodes in the order of probability for delivery, which is based on contact information. However, if the connection lasts long enough, all messages are transmitted, thus turning into standard Epidemic Routing. PRoPHET Routing [36] is also probability-based, using past encounters to predict the probability of meeting a node again, nodes that are encountered frequently have an increased probability whereas older contacts are degraded over time. Additionally, the transitive nature of encounters is exploited where nodes exchange encounter probabilities and the probability of indirectly encountering the destination node is evaluated. Similarly [28] and [50] define probability based on node encounters in order to calculate the cost of the route. In other work, [11] and [20] use the so-called 'time elapsed since last encounter' or the 'last encounter age' to route messages to destinations. In order to route a message to a destination, the message is forwarded to the neighbour who encountered the destination more recently than the source and other neighbours.

Lebrun et al. [30] propose a location-based routing scheme that uses the trajectories of mobile nodes to predict their future distance to the destination and passes messages to nodes that are moving in the direction of the destination. Leguay et al. [31] present a virtual coordinate system where the node coordinates are composed of a set of probabilities, each representing the chance that a node will be found in a specific location. This information is then used to compute the best available route. Similarly, Ghosh et al. propose exploiting the fact that nodes tend to move between a small set of locations, which they refer to as 'hubs' [16]. A list of 'hubs' specific to each users movement profile is assumed to be available to each node on the network in the form of a 'probabilistic orbit' which defines the probability with which a given node will visit a given hub. Messages destined for a specific node are routed towards one of these user specific 'hubs'.

Musolesi et al. [41] introduce a generic method that uses Kalman filters to combine multiple dimensions of a node's connectivity context in order to make routing decisions. Messages are passed from one node to a node with a higher 'delivery metric'. The messages for unknown destinations are forwarded to the 'most mobile' node available. Spyropoulos et al. [48] use a combination of random walk and utility-based forwarding. Random walk is used until a node with a sufficiently high utility metric is found after which the utility metric is used to route to the destination node. More recently, Hui et al. [25] investigated assigning labels to nodes identifying group membership. Messages are only forwarded to nodes in the same group as the destination node.

Our work is distinct in that the SimBetTS Routing metric is comprised of both a node's centrality and its social similarity. Consequently, if the destination node is unknown to the sending node or its contacts, the message is routed to a structurally more central node where the potential of finding a suitable carrier is dramatically increased. We will show that SimBetTS Routing improves upon encounter-based strategies where direct or indirect encounters may not be available.

III. Social Networks for information flow

In a disconnected environment, data must be forwarded using node encounters in order to deliver data to a destination. The problem of message delivery in disconnected delay-tolerant networks can be modeled as the flow of information over a dynamic network graph with time-varying links. This section reviews network theory that may be applied to social networks along with social network analysis techniques. These techniques have yet to be applied to the context of routing in disconnected delay-tolerant MANETs. Social network analysis is the study of relationships between entities, and on the patterns and implications of these relationships. Graphs may be used to represent the relational structure of social networks in a natural manner. Each of the nodes may be represented by a vertex of a graph. Relationships between nodes may be represented as edges of the graph.

A. Network Centrality for Information Flow

Centrality in graph theory and network analysis is a quantification of the relative importance of a vertex within the graph (for example, how important a person is within a social network). The centrality of a node in a network is a measure of the structural importance of the node, typically, a central node has a stronger capability of connecting other network members. There are several ways to measure centrality. Three widely used centrality measures are Freeman's degree, closeness, and betweenness measures [13], [14].

'Degree' centrality is measured as the number of direct ties that involve a given node [14]. A node with high degree centrality maintains contacts with numerous other network nodes. Such nodes can be seen as popular nodes with large numbers of links to others. As such, a central node occupies a structural position (network location) that may act as a conduit for information exchange. In contrast, peripheral nodes maintain few or no relations and thus are located at the margins of the network. Degree centrality for a given node p_i where $a(p_i, p_k) = 1$, if a direct link exists between p_i and p_k is calculated as:

$$C_D(p_i) = \sum_{k=1}^{N} a(p_i, p_k) \tag{1}$$

'Closeness' centrality measures the reciprocal of the mean geodesic distance $d(p_i, p_k)$, which is the shortest path between a node p_i and all other reachable nodes [14]. Closeness centrality can be regarded as a measure of how long it will take information to spread from a given node to other nodes in the network [43]. Closeness centrality for a given node, where N is the number of reachable nodes in the network, is calculated as:

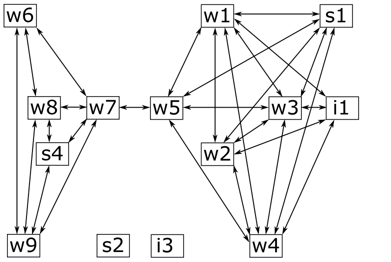
$$C_C(p_i) = \frac{N-1}{\sum_{k=1}^{N} d(p_i, p_k)} \tag{2}$$

'Betweenness' centrality measures the extent to which a node lies on the geodesic paths linking other nodes [13], [14]. Betweenness centrality can be regarded as a measure of the extent to which a node has control over information flowing between others [43]. A node with a high betweenness centrality has a capacity to facilitate interactions between nodes it links. In our case, it can be regarded as how much a node can facilitate communication to other nodes in the network. Betweenness centrality, where g_{jk} is the total number of geodesic paths linking p_j and p_k, and $g_{jk}(p_i)$ is the number of those geodesic paths that include p_i is calculated as:

$$C_B(p_i) = \sum_{j=1}^{N} \sum_{k=1}^{j-1} \frac{g_{jk}(p_i)}{g_{jk}} \tag{3}$$

Borgatti analyses centrality measures for flow processes in network graphs [5]. A number of different flow processes are considered, such as package delivery, gossip and infection. He then analyses each centrality measure in order to evaluate the appropriateness of each measure for different flow processes. His analysis showed that betweenness centrality and closeness centrality were the most appropriate metrics for message transfer that can be modeled as a package delivery.

Freeman's centrality metrics are based on analysis of a complete and bounded network, which is sometimes referred to as a sociocentric network. These metrics become difficult to evaluate in networks with a large node population as they require complete knowledge of the network topology. For this reason the concept of 'ego networks' has been introduced. Ego networks can be defined as a network consisting of a single actor (ego) together with the actors they are connected to (alters) and all the links among those alters. Consequently, ego network analysis can be performed locally by individual nodes without complete knowledge of the entire network. Marsden introduces centrality measures calculated using ego networks and compares these to Freeman's centrality measures of a sociocentric network [37]. Degree centrality can easily be measured for an ego network where it is a simple count of the number of contacts. Closeness centrality is uninformative in an ego network, since by definition an ego network only considers nodes directly related to the ego node, consequently by definition the hop distance from the ego node to all other nodes in the ego network is 1. On the other hand, betweenness centrality in ego networks has shown to be quite a good measure when compared to that of the sociocentric measure. Marsden calculates the egocentric and the sociocentric betweenness centrality for the network shown in figure 1.



Node	Sociocentric betweenness	Egocentric betweenness
w1	3.75	0.83
w2	0.25	0.25
w3	3.75	0.83
w4	3.75	0.83
w5	30	4
w6	0	0
w7	28.33	4.33
w8	0.33	0.33
w9	0.33	0.33
s1	1.5	0.25
s2	0	0
s4	0	0
i1	0	0
i3	0	0

Fig. 1. Bank Wiring Room network [25]

The betweenness centrality $C_B(p_i)$ based on the egocentric measures does not correspond perfectly to that based on sociocentric measures. However, it can be seen that the ranking of nodes based on the two types of betweenness are identical in this network. This means that two nodes may compare their own locally calculated betweenness value, and the node with the higher betweenness value can be determined. In effect, the betweenness value captures the extent to which a node connects nodes that are themselves not directly connected. For example, in the network shown in figure 1, *w9* has no connection with *w4*. The node with the highest betweenness value connected to *w9* is *w7*, so if a message is forwarded to *w7*, the message can then be forwarded to *w5* which has a direct connection with *w4*. In this way, betweenness centrality may be used to forward messages in a network. Marsden compared sociocentric and egocentric betweenness for 15 other sample networks and found that the two values correlate well in all scenarios. This correlation is also supported by Everett and Borgatti [12].

Routing based on betweenness centrality provides a mechanism for information to flow from source to destination in a social network. However, routing based on centrality alone presents a number of drawbacks. Yan et al. analysed routing in complex networks and found that routing based on centrality alone causes central nodes to suffer severe traffic congestion as the number of accumulated packets increases with time, because the capacities of the nodes for delivering packets are limited [53]. Additionally, centrality does not take into account the time-varying nature of the links in the network and the availability of a link. In terms of information flow, a link that is available is one that is 'activated' for information flow.

B. Strong Ties for Information Flow

The previous section's discussion of information flow based on centrality measures does not take into account the strength of the links between nodes. In terms of graph theory, where the links in the network are time-varying, a link to a central node may not be highly available. Brown and Reingen explored information flow in word-of-mouth networks and observe that it is unlikely that each contact representing potential sources of information have an equal probability of being activated for the flow of information [6]. They hypothesize that tie strength is a good measure of whether a tie will be activated, since strong ties are typically more readily available and result in more frequent interactions through which the transfer of information may arise. In a network where a person's contacts consisted of both strong and weak tie contacts, Brown and Reingen found that strong ties were more likely to be activated for information flow when compared to weak ties.

Tie strength is a quantifiable property that characterises the link between two nodes. The notion of tie strength was first introduced by Granovetter in 1973. Granovetter suggested that the strength of a relationship is dependent on four components: the frequency of contact, the length or history of the relationship, contact duration, and the number of transactions. Granovetter defined tie strength as: 'the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterise a tie' [18]. Marsden and Campbell extended upon these measures and also proposed a measure based on the depth of a relationship referred to as the 'multiple social context' indicator. Lin et al. proposed using the recency of a contact to measure tie strength [34]. The tie strength indicators are defined as follows:

Frequency - Granovetter observes that 'the more frequently persons interact with one another, the stronger their sentiments of friendship for one another are apt to be' [18]. This metric was also explored in [3], [4], [18], [35], [38].

Intimacy/Closeness - This metric corresponds to Granovetter's definition of the time invested into a social contact as a measure for a social tie [4], [18], [38]. A contact with which a great deal of time has been spent can be deemed an important contact.

Long Period of Time (Longevity) - This metric corresponds to Granovetter's definition of the time commitment into a social contact as a measure for a social tie [4], [18], [38]. A contact with which a person has interacted over a longer period of time may be more important than a newly formed contact.

Reciprocity - Reciprocity is based on the notion that a valuable contact is one that is reciprocated and seen by both members of the relationship to exist. Granovetter discusses the social example with the absence of a substantial relationship, for example a 'nodding' relationship between people living on the same street [4], [18]. He observes that this sort of relationship may be useful to distinguish from the absence of any relationship.

Recency - Important contacts should have interacted with a user recently [34]. This relates to Granovetter's amount of time component and investing in the relationship, where a strong relationship needs investment of time to maintain the intimacy.

Multiple Social Context - Marsden and Campbell discuss using the breadth of topics discussed by friends as a measure to represent the intimacy of a contact [4], [38].

Mutual Confiding (Trust) - This indicator can be used as a measure of trust in a contact [18], [38].

Routing based on tie strength in network terms is routing based on the most available links. A combination of the tie strength indicators can be used for information flow to determine which contact has the strongest social relationship to a destination. In this manner, messages can be forwarded through links possessing the strongest relationship, as a link representing a strong relationship more likely will be activated for information flow than a weak link with no relationship with the destination. These social measures lend themselves well to a disconnected network by providing a local view of the network graph as they are based solely on observed link events and require no global knowledge of the network.

However, Granovetter argued the utility of using weak ties for information flow in social networks [18]. He emphasised that weak ties lead to information dissemination *between* groups. He introduced the concept of 'bridges', observing that

> information can reach a larger number of people, and traverse a greater social distance when passed through weak ties rather than strong ties ... those who are weakly tied are more likely to move in circles different from our own and will thus have access to information different from that which we receive [18].

Consequently, it is important to identify contacts that may act as potential bridges. Betweenness centrality is a mechanism for identifying such bridges. Granovetter differentiates between the usefulness of weak and strong ties, 'weak ties provide people with access to information and resources beyond those available in their own social circle; but strong ties have greater motivation to be of assistance and are typically more easily available.' As a result, routing based on a combination of strong ties and identified bridges is a promising trade-off between the two solutions.

C. Tie Predictors

Marsden and Campbell distinguished between indicators and predictors [38]. Tie strength evaluates already existing connections whereas predictors use information from the past to predict likely future connections. Granovetter argues that strong tie networks exhibit a tendency towards transitivity, meaning that there is a heightened probability of two people being acquainted, if they have one or more other acquaintances in common [18]. In literature this phenomenon is called 'clustering'. Watts and Strogatz showed that real-world networks exhibit strong clustering or network transitivity [52]. A network is said to show 'clustering' if the probability of two nodes being connected by a link is higher when the nodes in question have a common neighbour.

Newman demonstrated this by analysing the time evolution of scientific collaborations and observing that the use of examining neighbours, in this case co-authors of authors, could help predict future collaborations [42]. From this analysis, Newman determined that the probability of two individuals collaborating increases as the number m of their previous mutual co-authors increases. A pair of scientists who have five mutual previous collaborators, for instance, are about twice as likely to collaborate as a pair with only two, and about 200 times as likely as a pair with none. Additionally Newman, determined that the probability of collaboration increases with the number of times one has collaborated before, which shows that past collaborations are a good indicator of future ones.

Liben-Nowell and Kleinberg explored this theory by the following common neighbour metric in order to predict future collaborations on an author database by assigning a score to the possible collaboration [33]. The score of a future collaboration $score(x, y)$ between authors x and y, where $N(x)$ and $N(y)$ are the set of neighbours of author x and y respectively, is calculated by:

$$score(x, y) = |N(x) \cap N(y)| \tag{4}$$

Their results strongly supported this argument and showed that links were predicted by a factor of up to 47 improvement compared to that of random prediction. The *common neighbour* measure in equation 4 measures purely the similarity between two entities. Liben-Nowell also explored using Jaccard's coefficient which attempts to take into account not just similarity but also dissimilarity. The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets:

$$score(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|} \tag{5}$$

Adamic and Adar performed an analysis to predict relationships between individuals by analysing user home pages on the world wide web (WWW) [1]. The authors computed features of the pages and also took into account the incoming and outgoing links of the page, and defined the similarity between two pages by counting the number of common features, assigning greater importance to rare features than frequently seen features. In the case of neighbours, Liben-Nowell utilised this metric which refines a simple count of neighbours by weighting rarer neighbours more heavily than common neighbours. The probability, where $N(z)$ is the number of neighbours held by z, is then given by:

$$P(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{log|N(z)|} \tag{6}$$

All three metrics performed well compared with random prediction. Liben-Nowell explored a number of different ranking techniques based on information retrieval research, but generally found that common neighbour, the Jaccard's coefficient and the Adamic and Adar technique were sufficient, and performed equally as well, if not better than the other techniques.

Centrality and tie strength are based on the analysis of a static network graph whose link availability are time-varying. However, in the case of DDTMs the network graph is not static; it evolves over time. Tie predictors can be used in order to predict the evolution of the graph and evaluate the probability of future links occurring. Tie predictors may be used not only to reinforce already existing contacts but to anticipate contacts that may evolve over time.

IV. Routing based on Social Metrics

We propose that information flow in a network graph whose links are time-varying can be achieved using a combination of centrality, strong ties and tie prediction. Social networks may consist of a number of highly disconnected cliques where neither the source node nor any of its contacts has any direct relationship with the destination. In this case, relying on strong ties would prove futile, and therefore weak ties to more connected nodes may be exploited.

Centrality has shown to be useful for path finding in static networks, however, the limitation of link capacities causes congestion. Additionally, centrality does not account for the time varying nature of link availability. Tie strength may be used to overcome this problem by identifying links that have a higher probability of availability. Tie strength evaluates existing links in a time varying network but does not account for the dynamic evolution of the network over time. Tie predictors may be used to aid in predicting future links that may arise. As such, we propose that the combination of centrality, tie strengths and tie predictors are highly useful in routing based on local information when the underlying network exhibits a social structure. This combined metric will be referred to as the SimBetTS Utiltity. When two nodes meet they exchange a list of encountered nodes, this list is used to locally calculate the betweenness utility, the similarity utility and the tie strength utility. Each node then examines the messages it is carrying and computes the SimBetTS Utility of each message destination. Messages are then exchanged where the message is forwarded to the node holding the highest SimBetTS Utility for the message destination node. The remainder of this section describes the calculation of the betweenness utility, the similarity utility and the tie strength utility and how these metrics are combined to calculate the SimBetTS utility.

A. Betweenness Calculation

Betweenness centrality is calculated using an ego network representation of the nodes with which the ego node has come into contact. When two nodes meet, they exchange a list of contacts. A contact is defined as a node that has been directly encountered by the node. The received contact list is used to update each node's local ego network. Mathematically, node contacts can be represented by an adjacency matrix A, which is a $n \times n$ symmetric matrix, where n is the number of contacts a given node has encountered, (for a worked example please see [10]). The adjacency matrix has elements:

$$A_{i,j} = \begin{cases} 1 & \text{if there is a contact between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Contacts are considered to be bidirectional, so if a contact exists between i and j then there is also a contact between j and i. The betweenness centrality is calculated by computing the number of nodes that are indirectly connected through the ego node. The betweenness centrality of the ego node is the sum of the reciprocals of the entries of A' where A' is equal to $A^2\left[1-A\right]_{i,j}$ [12] where i,j are the row,column matrix entries respectively. A node's betweenness utility is given by:

$$Bet = \sum \frac{1}{A'_{i,j}} \tag{8}$$

Since the matrix is symmetric, only the non-zero entries above the diagonal need to be considered. When a new node is encountered, the new node sends a list of nodes it has encountered. The ego node makes a new entry in the $n \times n$ matrix. As an ego network only considers the contacts between nodes that the ego has directly encountered, only the entries for contacts in common between the ego node and the newly encountered node are inserted into the matrix.

B. Similarity Calculation

Node similarity is calculated using the same $n \times n$ matrix discussed in section IV-A. The number of common neighbours between the current node i and destination node j can be calculated as the sum of the total overlapping contacts as represented in the $n \times n$ matrix (for a worked example please see [10]). This only allows for the calculation of similarity for nodes that have been met directly, but during the exchange of the node's contact list, information can be obtained in regards to nodes that have yet to be encountered. As discussed in section III-C the number of common neighbours may be used for ranking known contacts but also for predicting future contacts. Hence a list of indirect encounters is maintained in a separate $n \times m$ matrix where n is the number of nodes that have been met directly and m is the number of nodes that have not directly been encountered, but may be indirectly accessible through a direct contact. The similarity calculation, where N_n and N_e are the set of contacts held by node n and e respectively and is given as follows:

$$Sim(n,e) = \left| N_n \bigcap N_e \right| \tag{9}$$

C. Tie Strength Calculation

Measuring tie strength will be an aggregation of a selection of indicators based on those discussed in section III-B. An evidence-based strategy is used to evaluate whether each measure supports or contradicts the presence of a strong tie. The evidence is represented as a tuple (s, c) where s is the supporting evidence and c is the contradicting evidence. The trust in a piece of evidence is measured as a ratio of supporting and contradicting evidence [19]. Below elaborates on specific tie strength indicators, representing them as a ratio of supporting and contradicting evidence and bring them into the context of DDTMs.

Frequency - The frequency indicator may be based on the frequency with which a node is encountered. The supporting evidence of a strong tie strength is defined as the total number of times node n has encountered node m. The contradicting evidence is defined as the amount of encounters node n has observed where node m was not the encountered node. The frequency indicator, where $f(m)$ is the number of times node n encountered node m and $F(n)$ is the total number of encounters node n has observed, is given by:

$$FI_n(m) = \frac{f(m)}{F(n) - f(m)} \tag{10}$$

Intimacy/Closeness - The duration indicator can be based on the amount of time the node has spent connected to a given node. The supporting evidence of intimacy/closeness is defined as a measure of how much time node n has been connected to node m. The contradicting evidence is defined as the total amount of time node n has been connected to nodes in the network where node m was not the encountered node. If $d(m)$ is the total amount of time node n has been connected to node m and $D(n)$ is the total amount of time node n has been connected across all encountered nodes, then the intimacy/closeness indicator can be expressed as:

$$ICI_n(m) = \frac{d(m)}{D(n) - d(m)} \tag{11}$$

Recency - The recency indicator is based on how recently node n has encountered node m. According to social network analysis theory a strong tie must be maintained in order to stay strong which is captured by the recency indicator. The supporting evidence $rec(m)$, is how recently node n last encountered node m. This is defined as the length of time between node n encountered node m and the time node n

has been on the network. The contradicting evidence is the amount of time node n has been on the network since it has last seen node m. The recency indicator, where $L(n)$ is the total amount of time node n has been a part of the network, is given as:

$$RecI_n(m) = \frac{rec(m)}{L(n) - rec(m)} \tag{12}$$

Tie strength is a combination of a selection of indicators and the above metrics are all combined in order to evaluate an overall single tie strength measure. A strong tie should, ideally, be high in all measures. Consequently, the tie strength of node n for an encountered node m, where T is the set of social indicators, is given by:

$$TieStrength_n(m) = \sum^{I \in T} I_n(m) \tag{13}$$

D. Node Utility Calculation

The SimBetTS utility captures the overall improvement a node represents when compared to an encountered node across all measures presented in sections IV-A, IV-B and IV-C.

Selecting which node represents the best carrier for the message becomes a multiple attribute decision problem across all measures, where the aim is to select the node that provides the maximum utility for carrying the message. This is achieved using a pairwise comparison matrix on the normalised relative weights of the attributes.

The tie strength utility $TSUtil_n$, the betweenness utility $BetUtil_n$ and the the similarity utility $SimUtil_n$ of node n for delivering a message to destination node d compared to node m are given by:

$$SimUtil_n(d) = \frac{Sim_n(d)}{Sim_n(d) + Sim_m(d)} \tag{14}$$

$$BetUtil_n = \frac{Bet_n}{Bet_n + Bet_m} \tag{15}$$

$$TSUtil_n(d) = \frac{TieStrength_n(d)}{TieStrength_n(d) + TieStrength_m(d)} \tag{16}$$

The $SimBetTSUtil_n(d)$ is given by combining the normalised relative weights of the attributes, where $U = \{TSUtil, SimUtil, BetUtil\}$:

$$SimBetTSUtil_n(d) = \sum^{u \in U} u_n(d) \tag{17}$$

All utility values are considered of equal importance and the SimBetTSUtil is the sum of the contributing utility values.

Replication may be used in order to increase the probability of message delivery. Messages are assigned a replication value R. When two nodes meet, if the replication value $R > 1$ then a message copy is made. The value of R is divided between the two copies of the message. This division is dependent on the SimBetTS utility value of each node therefore the division of the replication number for destination d between node n and node m is given by:

$$R_n = \left\lfloor R_{cur} \times \frac{SimBetTSUtil_n(d)}{SimBetTSUtil_n(d) + SimBetTSUtil_m(d)} \right\rfloor$$

$$R_m = R_{cur} - R_n \tag{18}$$

Consequently the node with the higher utility value receives a higher replication value. If $R = 1$ then the forwarding becomes a single-copy strategy. For evaluation purposes, a replication value of $R = 4$ is used. Replication improves the probability of message delivery, however comes at the cost of increased resource consumption. If replication is used, in order to avoid sending messages to nodes that are already carrying the message the Summary Vector must include a list of message identifiers it is currently carrying for each destination node. However a benefit of this replication methodology is that the message is only split if a carrying node encounters a node with a greater SimBetTSUtil value, as a result if the message reaches the most appropriate node before the replication has occurred, then the maximum number of replicas will not be used.

V. Simulation Results

In this section we describe the simulation setup used to evaluate SimBetTS Routing. The first experiment evaluates the utility of each metric discussed in section IV in terms of overall message delivery and the resulting distribution of the delivery load across the nodes on the network. The second experiment illustrates the routing protocol behaviour and utility values that become important in the decision to transfer messages between nodes. The third experiment demonstrates the delivery performance of SimBetTS Routing compared to that of Epidemic Routing and PRoPHET Routing.

A. Simulation Setup

Cambridge University and Intel Research conducted three experiments of Bluetooth encounters using participants carrying iMote devices as part of the Haggle project [8], [24]. The experiments lasted between three to five days and only included a small number of participants. Encounters between Bluetooth devices were logged. Log entries include: the encounter start time, the deviceID, the deviceID of the encountering and encountered nodes. Logs are only available for the participating nodes, but the data set also includes encounters between participants carrying the iMote devices and external Bluetooth contacts. These include anyone who had an active Bluetooth device in the vicinity of the iMote carriers. The trace file of these sightings was used in order to generate an event-based simulation where a Bluetooth sighting in the trace is assumed to be a contact where nodes can exchange information.

Table I summarises the experiment details for each data set.

In order to evaluate the delivery performance of SimBetTS Routing the event-based simulations are used to explore routing performance. In order to provide an opportunity to gather information about the nodes within the network, 15% of the

	Intel	Cambridge	Infocom
Participants	8	12	41
Total Devices	128	223	264
Total Encounters	2,264	6,736	28,250
Average Encounter Duration	885 secs	572 secs	231 secs
Duration	3 days	5 days	3 days

TABLE I
HAGGLE DATA SET

simulation duration is used as a warm up phase. After the warm up phase, each node on the network generates messages to uniformly randomly selected destination nodes at a rate of 20 messages per hour from the start of the message sending phase when the sending node first reappears on the network until the last time the sending node appears on the network during the sending phase. The sending phase lasts for 70% of the simulation duration, allowing an additional 15% at the end in order to allow messages that have been generated time to be delivered. The simulations were run for all three protocols 10 times with different random number seeds. In this case, it is assumed that each Bluetooth encounter provides an opportunity to exchange all routing data and all messages.

B. Evaluating Social Metrics for Information Flow

The aim of the first experiment is to evaluate the utility of each SimBetTS Routing utility component for routing and the benefit of combining these three metrics in order to improve delivery performance. We evaluate the delivery performance based on the following criteria.

Total Number of Messages Delivered: This metric captures the total number of messages delivered by routing based on the utility metrics of betweenness, similarity, tie strength and finally SimBetTS Routing which combines all three utilities.

Distribution of Delivery Load: This metric is the percentage of messages delivered per node, which signifies the distribution of load across the network. If a high proportion of the messages are delivered by a small subset of nodes, then these nodes may become points of congestion. The aim of SimBetTS routing is to maximise delivery performance while avoiding heavy congestion on central nodes in the network.

1) Intel Data Set: The first experiment included eight researchers and interns working at Intel Research in Cambridge and lasted three days. A total of 128 nodes, including the eight participants, were encountered for the duration of the experiment.

Figure 2 shows the total number of messages delivered and the load distribution for the Intel Data Set. As can be seen from figure 2 b) when evaluating betweenness, similarity and tie strength utility; betweenness achieves the highest delivery performance. Figure 2 a) however shows the disadvantage of using betweenness utility in routing where over 50% of the total messages delivered in the network were delivered by a single node. Routing based on similarity and tie strength show a better distribution of load but this comes at the price of reduced overall delivery performance. SimBetTS Routing achieves the highest delivery performance by combining the three metrics. Additionally, the load distribution shows that routing based on the combined metrics reduces congestion on highly central nodes.

2) Cambridge Data Set: The second data set included twelve doctoral students and faculty from Cambridge University Computer Lab [8]. The experiment lasted 5 days, and a total of 223 devices, including the participants, were encountered during the experiment.

Figure 3 shows a similar trend in terms of delivery performance and distribution. SimBetTS Routing achieves the highest overall delivery performance. The load distribution of SimBetTS Routing when compared to Betweenness utility is improved but to a lesser extent than the Intel data set. Similarity and tie strength both show similar behaviour where a single node delivers a high proportion of the messages. This illustrates that in this particular social network, a single node was highly important in terms of message delivery. A large proportion of the extra messages delivered by combining the utility metrics are delivered by the node that shows a sharp increase in figure 3. As a consequence, this result is not directly due to a reduction in load distribution, but rather that these message were not delivered by using tie strength alone.

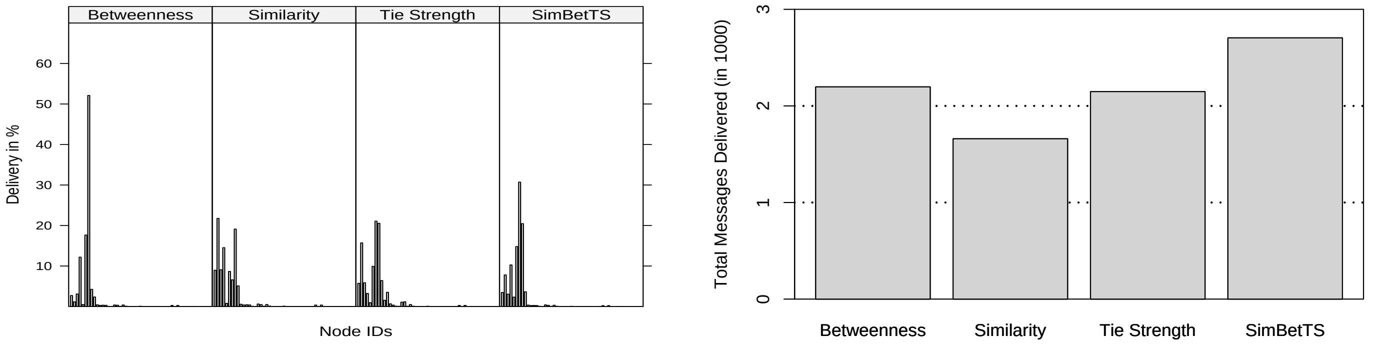
3) InfoCom Data Set: The third experiment collected data of encounters using iMotes equipped with Bluetooth distributed among conference attendees at the IEEE 2005 Infocom conference [24]. The participants were chosen in order to represent a range of different groups belonging to different organisations. Participants were asked to carry the devices with them for the duration of the conference. The experiment lasted 3 days and encounters between 264 devices were recorded.

Figure 4 shows that routing based on betweenness utility results in a single node delivering as much as 65% of the total messages delivered. SimBetTS Routing reduces this load significantly where less than 30% of the messages are delivered by this central node. As with the previous data sets, SimBetTS Routing achieves the highest overall delivery.

This experiment has demonstrated that SimBetTS Routing achieves superior delivery performance by combining the three utility metrics of betweenness, similarity and tie strength when compared to routing based on the individual metrics alone. Additionally, we have shown that although betweenness centrality achieves high delivery performance, routing based on centrality alone results in significant load on a small subset of the nodes on the network. Combining the utility metrics shows a better load distribution engaging more nodes in message delivery.

C. Demonstrating Routing Protocol Behaviour

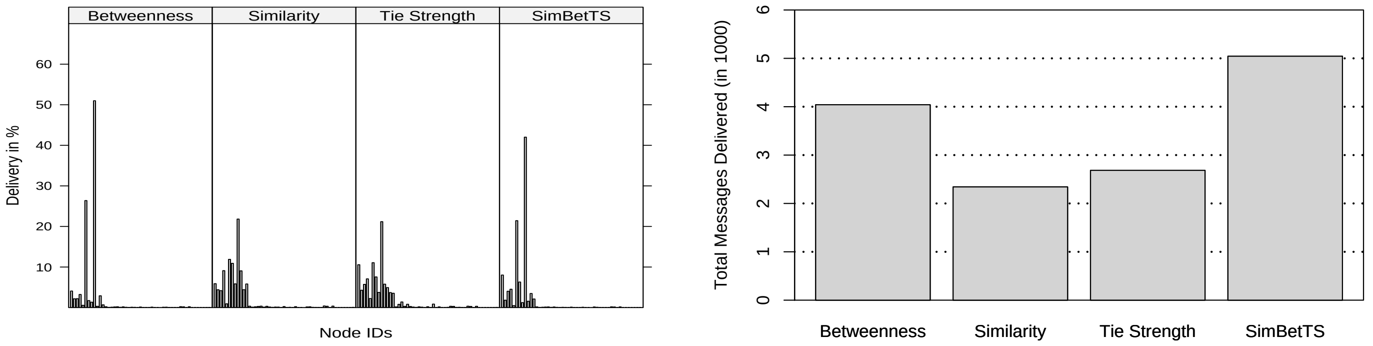
The goal of this experiment is to illustrate the benefit of utilising a multi-criteria decision method for combining the values in order to identity the node that represents the best trade off across these three metrics. The SimBetTS routing protocol is based on the premise that when forwarding a message for a given destination, the message is forwarded to a node with a higher probability of encountering the destination node. A node with a high tie strength has a higher probability of encountering the destination node. If a node with a high tie



(a) Distribution of Message Delivery

(b) Total Message Delivery

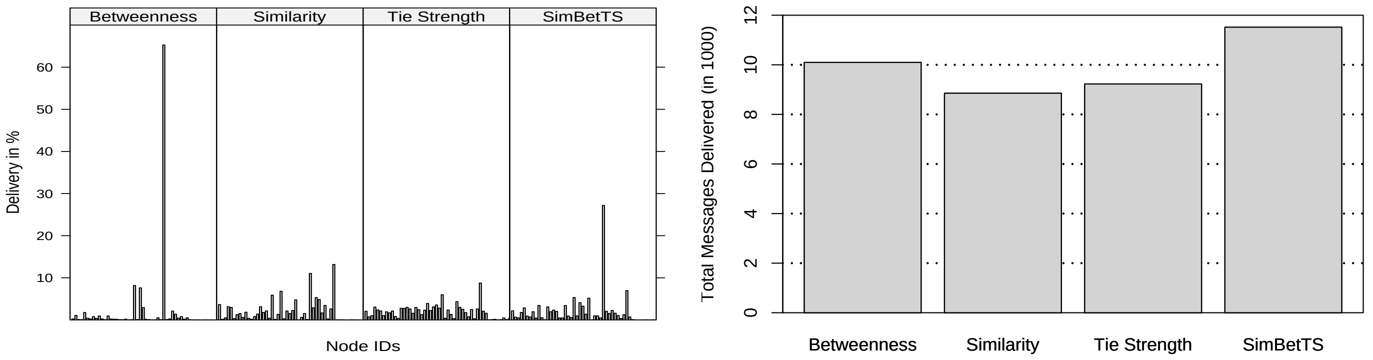
Fig. 2. Performance of Protocol Utility Components Intel Data



(a) Distribution of Message Delivery

(b) Total Message Delivery

Fig. 3. Performance of Protocol Utility Components Cambridge Data



(a) Distribution of Message Delivery

(b) Total Message Delivery

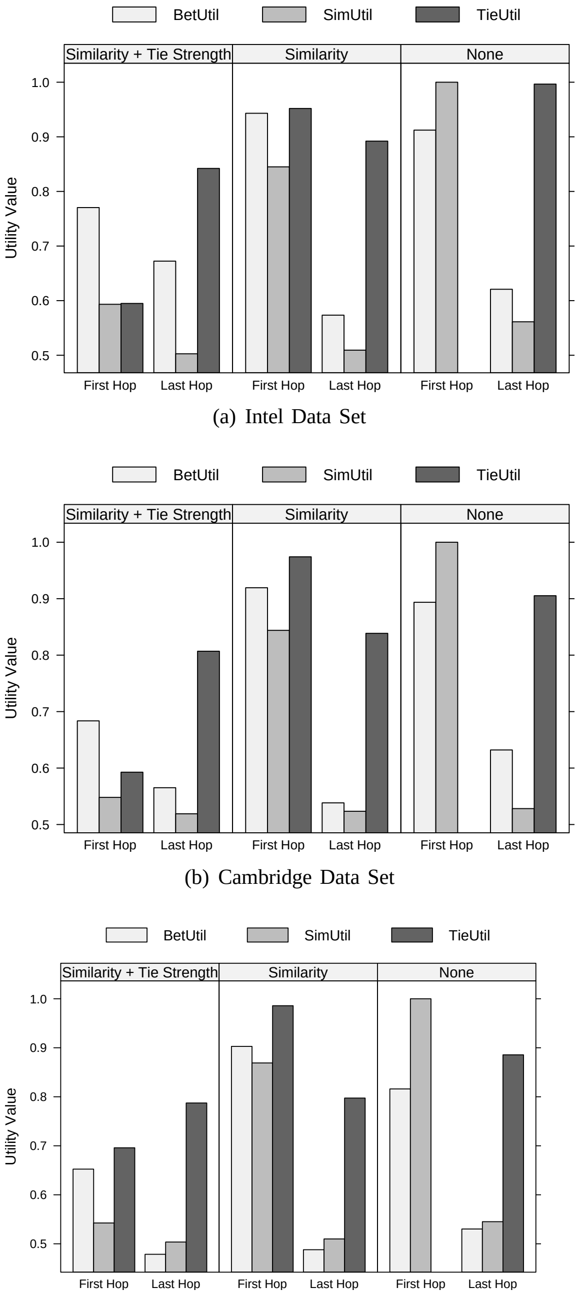
Fig. 4. Performance of Protocol Utility Components InfoCom Data

strength cannot be found, the utility metrics of social similarity and betweenness centrality are used. To demonstrate that the SimBetTS routing protocol achieves this behaviour, we divide the paths taken by message from source to destination into three categories:

- **Similarity + Tie Strength**: the sending node has a non-zero tie strength value for the destination node and has a non-zero similarity to the destination node.
- **Similarity**: the sending node has never encountered the destination node resulting in a zero tie strength value, however the sending node has a non-zero similarity value.
- **None**: the sending node has never encountered the destination node, and has no common neighbours, hence a zero similarity value.

Figure 5 shows the utility values of the first hop node on the message delivery path and the utility values of the final delivery node along the message path for each of these categories. It is important to note that these values are the relative utility values of the node selected to forward the message compared to the currently carrying node. As a result a high utility value at the first hop and a lower utility value at the final hop does not mean a lower absolute value, but lower when compared to the currently carrying node.

1) Similarity and Tie Strength: In the category of Similarity and Tie strength a combination of tie strength and betweenness utilities is the deciding factor in forwarding to the next hop.



(a) Intel Data Set

(b) Cambridge Data Set

(c) InfoCom Data Set

Fig. 5. Utility values at First hop and Last hop along the message path divided into the categories of Similarity + Tie Strength, Similarity and None

Similarity has a lower impact, most likely due to the fact that if the sending node has a social similarity and a tie strength value for the destination node, then the nodes encountered by the sending node most likely also have a social similarity. In all data sets, it can be seen that the tie strength utility is the highest contributing utility when forwarding to the last hop delivering node.

2) Similarity: In the category of Similarity where the sending node has a non-zero similarity value to the destination node, the highest contributing utility value is consistently the tie strength utility in both the first hop and the last hop delivering node. It can also be seen that the betweenness utility is also a contributing utility on the first hop, however for the final delivering node betweenness utility is much reduced. This is the desired behaviour of the protocol, because betweenness utility should only have a high impact in finding a node with a high tie strength to the destination node and then should have a lower impact on the forwarding decision.

3) None: In the category where the sending node has no social similarity to the destination node, and has also never encountered the destination node, we can see the benefit of the routing protocol. In all data sets a combination betweenness and similarity are the most contributing utility values. It was found that in this category, tie strength has a zero utility value in all cases at the first hop, however it can be consistently seen the final delivering node had a high tie strength utility. As a result, we can conclude that the routing protocol functions as intended. Whenever the destination node is unknown, a combination of betweenness utility and similarity utility will navigate the message to a node in the network that has a higher tie strength for the destination node.

This experiment has demonstrated the navigation behaviour of the SimBetTS routing protocol. The message is forwarded to nodes with a high betweenness and social similarity, until a node with a high tie strength for the destination node is found. In all case the tie strength utility for the final hop is the highest contributing utility value, and in all cases the betweenness utility value is much reduced in its influence of the forwarding decision as the message is routed closer to the destination.

D. Delivery Performance of Combined Metrics

The goal of the third experiment is to evaluate the delivery performance of SimBetTS routing protocol compared to two protocols: Epidemic Routing [51] and PRoPHET Routing [36]. The default parameters for PRoPHET Routing were used as defined in [36].

Two versions of SimBetTS Routing and PRoPHET are evaluated: a single-copy version and a multi-copy version. In the single-copy strategy, when two nodes meet, messages are exchanged between nodes where message are forwarded to the node with the highest utility. The node that has forwarded the message must then delete the message from the message queue. In the multi-copy strategy replication is used where messages are assigned a replication value R. For evaluation purposes, a replication value of $R = 4$ is used.

Total Number of Messages Delivered: The ultimate goal of the SimBetTS Routing design is to achieve delivery performance as close to Epidemic Routing as possible. This is because Epidemic Routing always finds the best possible path to the destination and therefore represents the baseline for the best possible delivery performance.

Average End-to-End Delay: End-to-End delay is an important concern in SimBetTS Routing design. Long end-to-end delays means the message must occupy valuable buffer space for longer, and consequently a low end-to-end delay is desirable. Again Epidemic Routing presents a good baseline for the minimum end-to-end delay possible.

Average Number of Hops per Message: It is desirable to minimise the number of hops a message must take in order to reach the destination. Wireless communication is costly in terms of battery power and as a result minimising the number of hops also minimises the battery power expended in forwarding the message.

Total Number of Forwards: This value represents the network overhead in terms of how many times a message forward occurs. PRoPHET and SimBetTS are expected to perform similarly in this respect, as both only assume the existence of one copy of the message on the network. Epidemic Routing, however, assumes the existence of multiple copies and continues forwarding a given message until each node is carrying a copy. This means Epidemic Routing is costly in terms of the number of transmissions required along with the amount of buffer space required on each node.

1) Intel Data Set: Figure 6 graphs the results for this simulation. As shown in figure 6 a) it is clear that Epidemic Routing outperforms both SimBetTS Routing and PRoPHET. Single-copy SimBetTS Routing delivers fewer messages than Epidemic but shows improvement when compared to Single-copy PRoPHET. Multi-copy SimBetTS Routing shows significant improvement with the addition of replication resulting in an improvement of nearly 50% than the single-copy strategy. Multi-copy PRoPHET shows much less improvement with the addition of replication, achieving results comparable to single-copy SimBetTS Routing. All protocols show a number of plateaus where no messages are delivered before the 150 mark and at the 200 mark. These plateaus represent night time when the devices are not within range of other devices.

Figure 6 b) shows the average number of hops taken by the delivered messages. All protocols result in a small number of hops of around 3. Single-copy and multi-copy PRoPHET show the largest and smallest number of average hops respectively. Single-copy and multi-copy SimBetTS Routing show very similar average hop values, meaning the path lengths found by single-copy SimBetTS Routing were close to that achieved by multi-copy SimBetTS. Epidemic achieves short paths from the source to destination of just below an average of 3. SimBetTS Routing shows a distinct increase in the average number of hops after the 300 mark, which coincides with a large number of messages being delivered. This indicates that these messages required a larger number of hops in order to be delivered, thus raising the average.

Figure 6 c) shows the average message end-to-end delay. As expected, Epidemic Routing shows the lowest average end-to-end delay. Single-copy PRoPHET shows the highest overall delay. Single-copy SimBetTS Routing shows similar delays. Both multi-copy SimBetTS Routing and PRoPHET show reduced delays with multi-copy SimBetTS Routing resulting in low delays near those achieved by Epidemic Routing. All protocols show a significant increase in delay around the 250 mark, which coincides with a steep increase in the number of messages delivered, meaning these delivered messages required a longer amount of time to be delivered, thus increasing the average end-to-end delay.

The true disadvantage of Epidemic Routing becomes clear when examining the total number of forwards. The overhead associated with Epidemic Routing is so great that it has been omitted from figure 6 d) in order for the differentiation of SimBetTS and PRoPHET to be seen. Epidemic Routing continues to forward messages throughout the network and as a result incurs a great deal of overhead. Single-copy SimBetTS and PRoPHET only have a single copy of each message on the network, resulting in a significantly lower number of forwards. Multi-copy SimBetTS Routing and PRoPHET as expected show an increase in the number of forwards with the use of replication. However, when compared to that of Epidemic Routing, multi-copy SimBetTS Routing results in approximately 98% reduction in number of forwards. SimBetTS results in a higher number of forwards when compared to PRoPHET however, this result should be viewed in the context that PRoPHET delivered significantly less messages overall.

Figure 7 a) shows the protocol control data overhead. Epidemic Routing overhead is so large it makes it difficult to differentiate between the overhead generated by single-copy SimBetTS Routing and PRoPHET. This is because the summary vector must contain a list of all messages the node is currently carrying, and as the number of messages on the network increases, so does the control data. Single-copy SimBetTS Routing and PRoPHET generate significantly lower overhead due to the fact that nodes exchange information about message destination rather than explicit message identifiers. As a result, there is an upper limit on the amount of bytes required. This upper bound depends on the node population. Multi-copy SimBetTS Routing and PRoPHET show an increase in control data due to the necessary exchange of message identifiers, thus reducing the benefit of exchanging only routing data, as is the case for the single-copy strategy. However, the overhead is still significantly lower than that of Epidemic Routing due to the fact that nodes do not carry a copy of every message in the network.

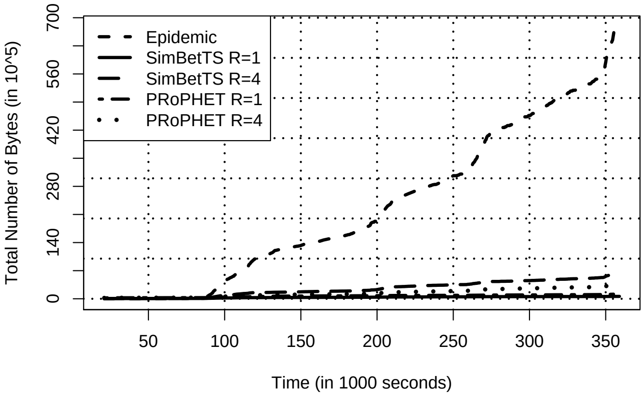
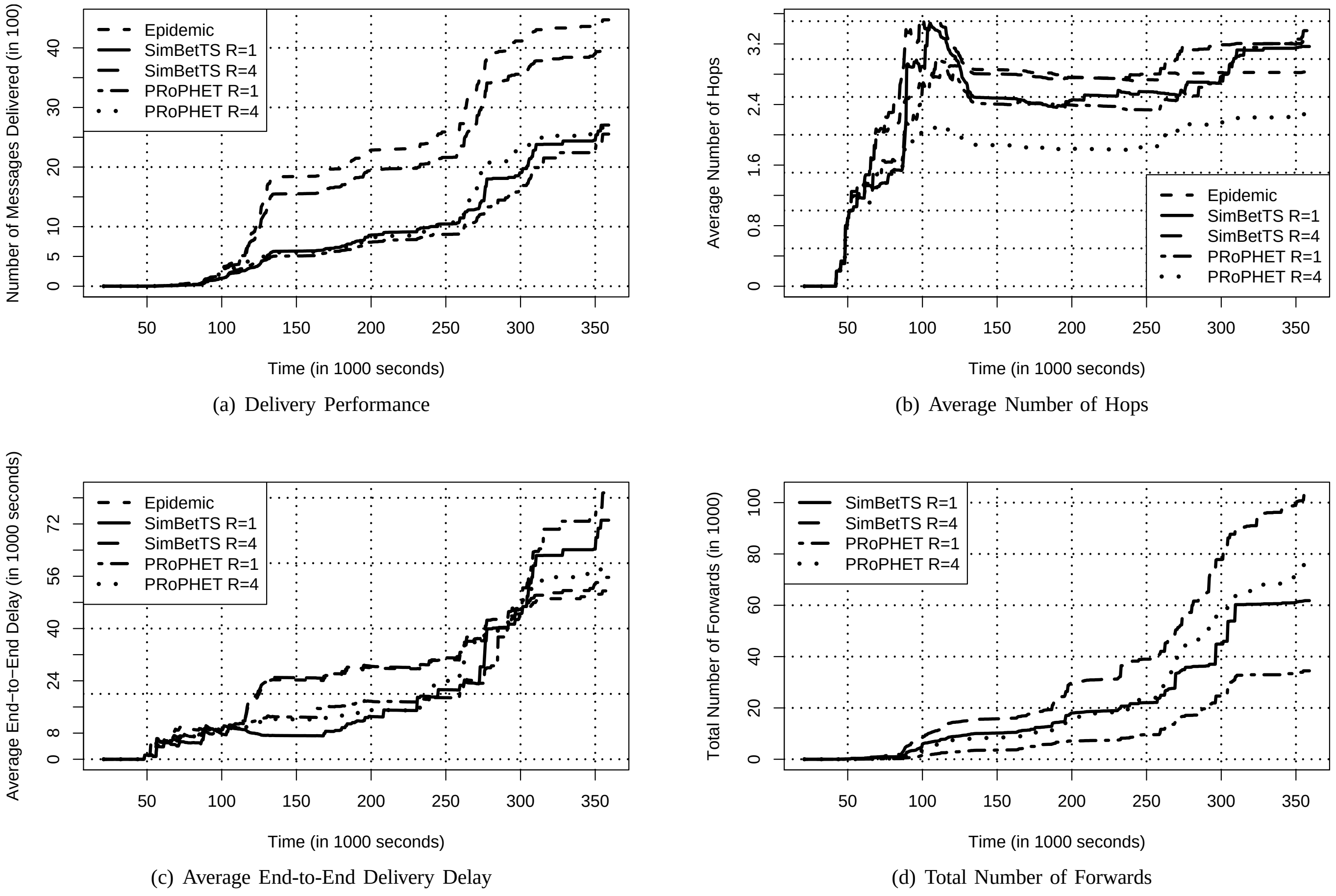


Fig. 7. Control Data Overhead for Intel Data

2) Cambridge Data Set: Figure 8 graphs the results from the Cambridge Data set performance tests. Plateaus are again evident where message delivery remains level. There are three such plateaus around the 200, 300 and 400 marks are again night time where devices are inactive. The first night occurred before the message sending phase commenced and the 200, 300 and 400 represent the 2nd, 3rd and 4th night of the experiment respectively. As expected, Epidemic Routing shows superior delivery performance compared to that of



(a) Delivery Performance

(b) Average Number of Hops

(c) Average End-to-End Delivery Delay

(d) Total Number of Forwards

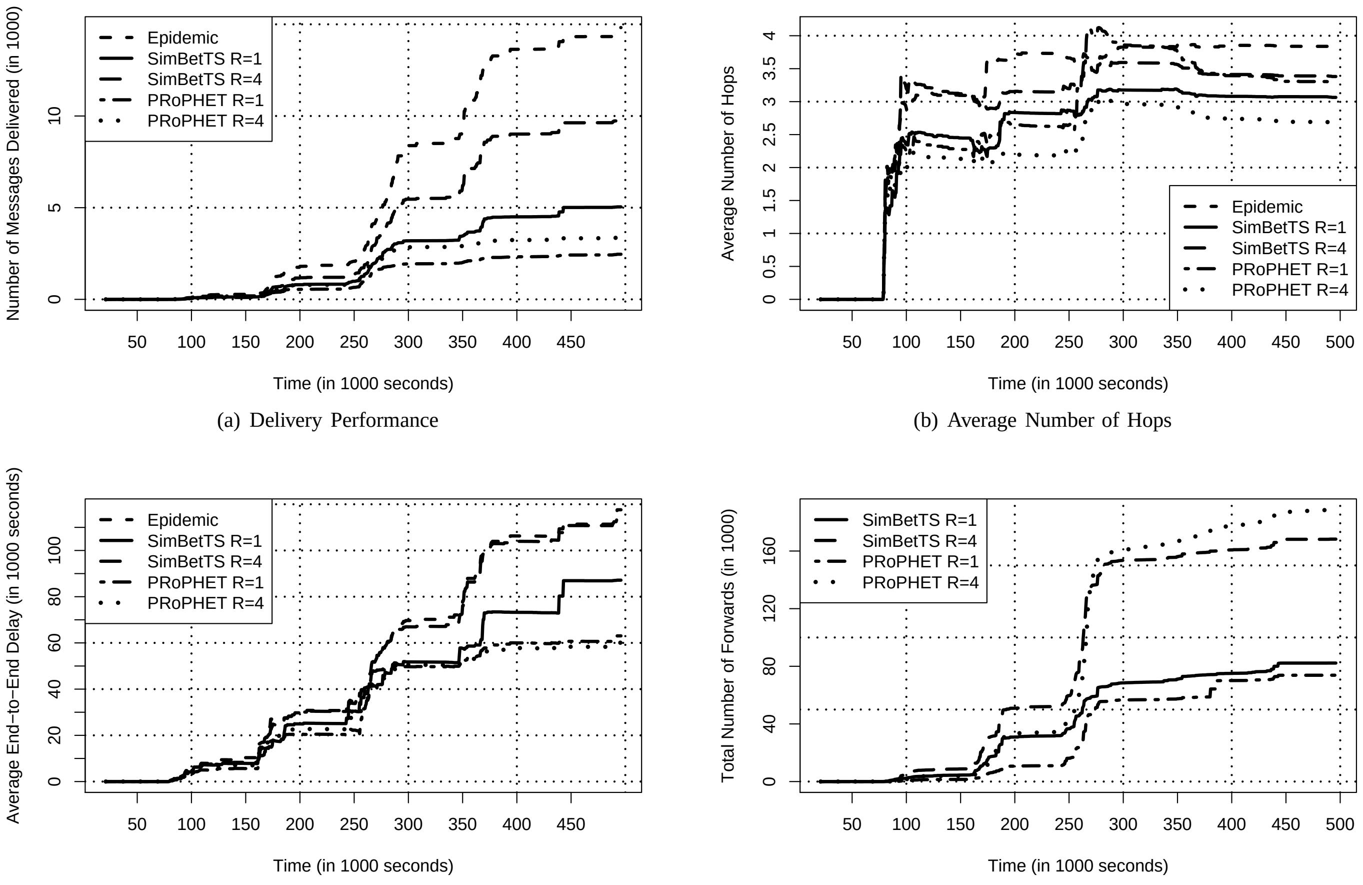
Fig. 6. Protocol Performance for Intel Data

PRoPHET and SimBetTS Routing as shown in figure 8 a). Single-copy SimBetTS Routing outperforms both single-copy and multi-copy PRoPHET. Single-copy SimBetTS Routing achieves improved delivery performance of 100% when compared to single-copy PRoPHET. More dramatically, multi-copy SimBetTS Routing achieved improved delivery performance of 300% when compared to that of multi-copy PRoPHET.

Figure 8 b) shows the average number of hops. As with the previous experiment, all protocols achieve message delivery in a relatively small number of hops of around 3-4. Epidemic Routing results in the highest average. It could be assumed that this is due to the higher number of messages delivered by Epidemic Routing resulting in a number of messages that required a larger number of hops in order to reach the destination. However, upon inspection of the message delivery graph, the average number of hops remains level for Epidemic Routing even after a large proportion of the messages are delivered. This can be explained by the fact that Epidemic Routing finds the optimum path in terms of delivery delay rather than the shortest path. Multi-copy PRoPHET results in the least number of hops. Interestingly, multi-copy SimBetTS Routing results in a higher average when compared to single-copy SimBetTS Routing, but this increase around the 250 mark coincides with a sharp increase in message delivery. After this time, the average number of hops starts to decrease.

Figure 8 c) shows the average message delay. Epidemic Routing shows the highest overall delay, but it can be noted that the delay increases most sharply around the 350 mark. Upon inspection of the delivery graph 8 a) this increase can be explained by the increase in total messages delivered, representing the fact that the messages delivered during this time phase were unable to be delivered in a shorter time. Approximately 30% of the encountered nodes do not appear in the network until after the 300 mark. PRoPHET shows the lowest overall delivery delay due to the fact that it delivered a smaller proportion of the total messages. SimBetTS Routing results in an average end-to-end delay between that of PRoPHET and Epidemic Routing. Similar to the average hop count, multi-copy SimBetTS Routing shows an increase in average end-to-end delay when compared to single-copy SimBetTS Routing. The increase in delay is clearly related to the increase in message delivery, and follows a trend almost identical to that of Epidemic Routing.

Figure 8 d) shows the total number of message forwards throughout the simulation. As with the previous data set, Epidemic Routing results in a large number of forwards, which is to be expected with a flooding protocol. Single-copy SimBetTS Routing and PRoPHET result in a relatively low number of forwards. This value increases for multi-copy SimBetTS Routing and PRoPHET. Unlike the previous data set, multi-copy PRoPHET results in a higher number of forwards than that of multi-copy SimBetTS Routing. As a



(a) Delivery Performance

(b) Average Number of Hops

(c) Average End-to-End Delivery Delay

(d) Total Number of Forwards

Fig. 8. Protocol Performance for Cambridge Data

result, it can be determined that this value is dependent on the dynamics of the underlying network rather than a deterministic side effect of the protocol.

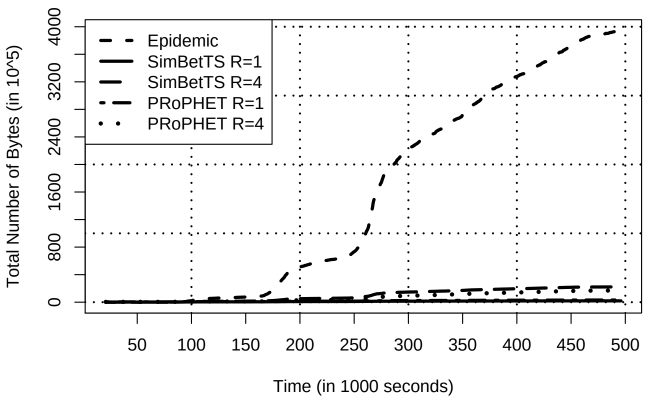
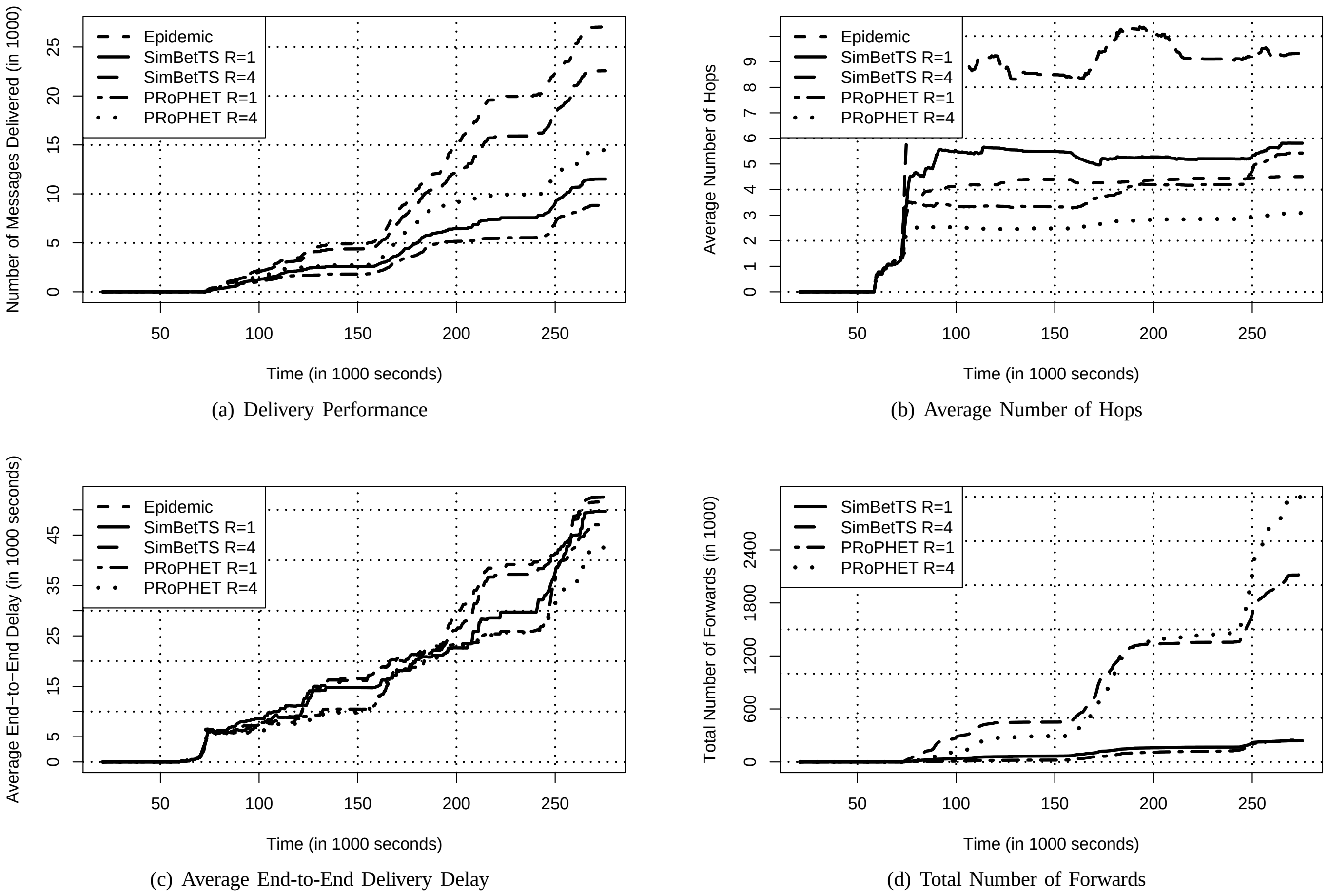


Fig. 9. Control Data Overhead for Cambridge Data

Figure 9 shows the total overhead associated with the control data for each protocol. Epidemic Routing as expected increases dramatically as the number of messages on the network increases. Both single-copy SimBetTS Routing and PRoPHET protocols increase at approximately the same rate, however, single-copy SimBetTS Routing results in the lowest number of bytes. Multi-copy SimBetTS Routing results in a larger amount of control data when compared to multi-copy PRoPHET, however, both protocols follow a similar trend.

3) Infocom Data Set: The overall delivery performance in figure 10 a) shows that Epidemic Routing achieves significant improvement after the 150 mark. This can be explained by the fact that approximately 34% of the node population first appear after this time frame. At this point, SimBetTS Routing and PRoPHET start to build up encounter information in regards to these nodes. This explains why SimBetTS Routing starts to outperform PRoPHET after the 150 mark, because messages destined for nodes that have yet to be encountered are already routed to more central nodes which are more likely to gather encounter information about the previously unseen nodes. Multi-copy SimBetTS Routing achieves message delivery close to Epidemic and outperforms PRoPHET.

Figure 10 b) shows the average number of hops of delivered messages. Epidemic Routing results in hop lengths of approximately 4. Single-copy SimBetTS Routing and PRoPHET achieve similar results of between 5 and 6 hops. Multi-copy PRoPHET results in the lowest number of hops. In contrast, multi-copy SimBetTS Routing shows significant increase, resulting in hops of approximately 9. This increase is due to an increased path length of the additional messages delivered by multi-copy SimBetTS Routing that remained undelivered for single-copy SimBetTS Routing. Figure 10 c) shows the average end-to-end delay. Even with the increased path lengths



(a) Delivery Performance

(b) Average Number of Hops

(c) Average End-to-End Delivery Delay

(d) Total Number of Forwards

Fig. 10. Protocol Performance for Infocom Data

used by multi-copy SimBetTS Routing, the average end-to-end delay is similar to that of Epidemic Routing. The sharpest increase occurs for Epidemic after the 150 mark which is most likely when messages are delivered to nodes previously unseen on the network. SimBetTS Routing shows a similar increase. PRoPHET shows the lowest average delay. However, the fact that it shows delays lower than that of Epidemic Routing illustrates that the increase in message delay shown by Epidemic is caused by the additional messages delivered, which required a greater amount of time before it was possible to deliver these messages.

Figure 10 c) shows the average message delay. Similarly to the Intel and Cambridge data sets, Epidemic Routing shows the greatest average delay due the higher proportion of messages delivered. The sharpest increase occurs after for Epidemic after the 150 mark which is most likely due to message delivery of messages destined for the previously unseen on the network. SimBetTS Routing shows a similar increase just before the 250 mark which coincides with an increase in message delivery. Consequently, this increase also coincides with the message delivery to nodes that have yet to be encountered.

Figure 10 d) shows the total number of messages forwarded in the network. As expected, Epidemic Routing results in the highest number of total forwards and has been omitted. SimBetTS routing and PRoPHET result in a similar number of forwards when comparing the single-copy strategy, however multi-copy PRoPHET results in a higher number of forwards than multi-copy SimBetTS.

Figure 11 shows the total control data. The overhead of Epidemic Routing increases dramatically as the number of messages on the network also increases. As seen with the other data sets, single-copy SimBetTS Routing and PRoPHET result in a similar amount of overhead. Multi-copy SimBetTS Routing results in a larger amount of control data when compared to multi-copy PRoPHET. However this result should be viewed in the context that SimBetTS results in an improvement of message delivery of nearly 50% compared to multi-copy PRoPHET.

From this experiment, we can determine that single-copy and multi-copy SimBetTS Routing show higher message delivery when compared to that of PRoPHET. Multi-copy SimBetTS Routing achieves delivery performance similar to that of Epidemic Routing with short path lengths and low end-to-end delay. The use of replication comes at the cost of an increased number of forwards and an increase in control data. However, when compared to that of Epidemic Routing, these metrics are still relatively low in terms of overhead.

VI. Conclusions

We have presented social network analysis techniques and shown how they may be applied to routing in disconnected

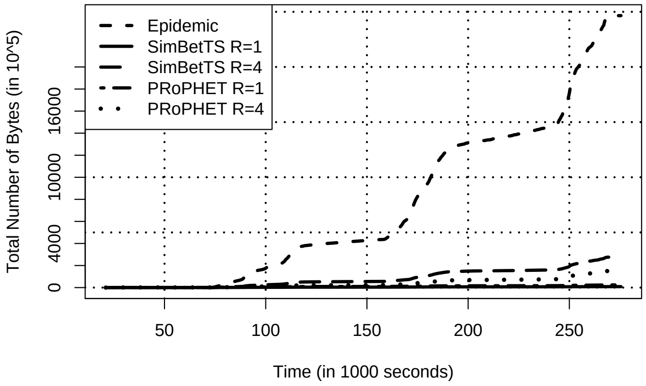


Fig. 11. Baseline Control Data Overhead for Infocom Data

delay-tolerant MANETs. Three metrics have been defined: betweenness utility, similarity utility and tie strength utility. Each of these metrics have been evaluated individually, and the results show that betweenness utility results in the best overall delivery performance. However it results in congestion on highly central nodes. SimBetTS Routing combines these three metrics which results in improved overall delivery performance with the additional advantage that the load on central nodes is reduced and better distributed across the network.

We have evaluated SimBetTS Routing compared to two other protocols, Epidemic Routing and PRoPHET. SimBetTS Routing achieves delivery performance comparable to Epidemic Routing, without the additional overhead. We have also demonstrated that SimBetTS Routing outperforms the PRoPHET routing protocol in terms of overall delivery performance. The evaluation consisted of three separate real world trace experiments. Although the first two datasets are relatively small networks, the consistency of the results across all three scenarios shows that, given the existence of an underlying social structure, SimBetTS Routing provides message delivery with greatly reduced overhead when compared to Epidemic Routing. Similar results can be seen when analysing the larger MIT Reality Mining dataset covered in our previous work [10]. Additionally, these utility metrics may be applied in other distributed systems where global topology information is unavailable, especially where the underlying networks exhibit small-world characteristics.

References

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, July 2003.

[2] Allan Beaufour, Martin Leopold, and Philippe Bonnet. Smart-tag based data dissemination. In *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, pages 68–77, New York, NY, USA, September 2002. ACM Press.

[3] Mario Benassi, Arent Greve, and Julia Harkola. Looking for a network organization: The case of GESTO. *Journal of Market-Focused Management*, 4(3):205–229, October 1999.

[4] Philip Blumstein and Peter Kollock. Personal relationships. *Annual Review of Sociology*, 14:467–490, 1988.

[5] Stephen P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, January 2005.

[6] Jacqueline J. Brown and Peter H. Reingen. Social ties and word-of-mouth referral behavior. *The Journal of Consumer Research*, 14(3):350–362, December 1987.

[7] John Burgess, Brian Gallagher, David Jensen, and B. N. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networking. In *Proceedings of the IEEE INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE Press, March 2006.

[8] Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. In *Proceedings of IEEE INFOCOM 2006*, 2006.

[9] S. Corson and J. Macker. Mobile ad hoc networking (MANET): Routing protocol performance issues and evaluation considerations, rfc 2501, January 1999.

[10] Elizabeth M. Daly and Mads Haahr. Social network analysis for routing in disconnected delay-tolerant MANETs. In *Proceedings of the 8th ACM Interational Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2007, MontrÃl’al, QuÃl’bec, Canada, September 9-14, 2007*, September 2007.

[11] Henri Dubois-Ferriere, Matthias Grossglauser, and Martin Vetterli. Age matters: efficient route discovery in mobile ad hoc networks using encounter ages. In *MobiHoc ’03: Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*, pages 257–266, New York, NY, USA, 2003. ACM Press.

[12] Martin Everett and Stephen P. Borgatti. Ego network betweenness. *Social Networks*, 27(1):31–38, January 2005.

[13] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, March 1977.

[14] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978-1979.

[15] R. H. Frenkiel, B. R. Badrinath, J. Borres, and R. D. Yates. The infostations challenge: balancing cost and ubiquity in delivering wireless data. *Personal Communications, IEEE [see also IEEE Wireless Communications]*, 7(2):66–71, 2000.

[16] Joy Ghosh, Hung Q. Ngo, and Chunming Qiao. Mobility profile based routing within intermittently connected mobile ad hoc networks (ICMAN). In *IWCMC ’06: Proceeding of the 2006 international conference on Communications and mobile computing*, pages 551–556, New York, NY, USA, 2006. ACM Press.

[17] Natalie Glance, Dave Snowdon, and Jean-Luc Meunier. Pollen: using people as a communication medium. *Computer Networks*, 35(4):429–442, March 2001.

[18] Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, May 1973.

[19] Elizabeth L. Gray. *A Trust-Based Reputation Management System*. PhD thesis, April 2006.

[20] Matthias Grossglauser and Martin Vetterli. Locating nodes with ease: last encounter routing in ad hoc networks through mobility diffusion. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE*, volume 3, pages 1954–1964 vol.3, 2003.

[21] Radu Handorean, Christopher Gill, and Gruia-Catalin Roman. Accommodating transient connectivity in ad hoc and mobile settings. In Alois Ferscha and Friedemann Mattern, editors, *Proceedings of the Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004*, volume 3001 of *Lecture Notes in Computer Science*, pages 305–322. Springer-Verlag, March 2004.

[22] W. Hsu and A. Helmy. On nodal encounter patterns in wireless LAN traces. In *Proceedings of the 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006*, pages 1–10. IEEE Press, April 2006.

[23] Wei-Jen Hsu and Ahmed Helmy. Impact: Investigation of mobile-user patterns across university campuses using wlan trace analysis, August 2005.

[24] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. Pocket switched networks and human mobility in conference environments. In *Proceeding of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251, New York, NY, USA, August 2005. ACM Press.

[25] Pan Hui and Jon Crowcroft. How small labels create big improvements. In *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops ’07. Fifth Annual IEEE International Conference on*, pages 65–70, 2007.

[26] Sushant Jain, Kevin Fall, and Rabin Patra. Routing in a delay tolerant network. *SIGCOMM Comput. Commun. Rev.*, 34(4):145–158, October 2004.

[27] David B. Johnson and David A. Maltz. *Dynamic source routing in ad-hoc wireless networks*, volume 353 of *The Springer International*

Series in Engineering and Computer Science, pages 153–181. Kluwer Academic Publishers, 1996.

[28] A. Khelil, P. J. Marron, and K. Rothermel. Contact-based mobility metrics for delay-tolerant ad hoc networking. pages 435–444, 2005.

[29] Young-Bae Ko and Nitin H. Vaidya. Location-aided routing (LAR) in mobile ad hoc networks. *Wireless Networks*, 6(4):307–321, September 2000.

[30] J. Lebrun, Chen-Nee Chuah, D. Ghosal, and Michael Zhang. Knowledge-based opportunistic forwarding in vehicular wireless ad hoc networks. In *Proceedings of the IEEE 61st Conference on Vehicular Technology, 2005. VTC 2005-Spring*, volume 4, pages 2289–2293. IEEE Press, May 2005.

[31] JÃl’rÃl’mie Leguay, Timur Friedman, and Vania Conan. Evaluating mobility pattern space routing for DTNs. In *INFOCOM 2005. Proceedings of the 24th IEEE Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE Press, April 2006.

[32] Qun Li and Daniela Rus. Sending messages to mobile users in disconnected ad-hoc wireless networks. In *MobiCom ’00: Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 44–55, New York, NY, USA, August 2000. ACM Press.

[33] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM ’03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, November 2003. ACM Press.

[34] Nan Lin, Paul Dayton, and Peter Reenwald. Analyzing the instrumental use of relations in the context of social structure. *Sociological Methods and Research*, 7(2):149–166, 1978.

[35] Nan Lin, John C. Vaughn, and Walter M. Ensel. Social resources and occupational status attainment. *Social Forces*, 59(4):1163–1181, June 1981.

[36] Anders Lindgren, Avri Doria, and Olov SchelÃl’n. Probabilistic routing in intermittently connected networks. In *Proceedings of the First International Workshop, SAPIR 2004, Fortaleza, Brazil, August 1-6, 2004*, volume 3126 of *Lecture Notes in Computer Science*, pages 239–254. Springer-Verlag GmbH, August 2004.

[37] P. V. Marsden. Egocentric and sociocentric measures of network centrality. *Social networks*, 24(4):407–422, October 2002.

[38] Peter V. Marsden and Karen E. Campbell. Measuring tie strength. *Social Forces*, 63(2):482–501, December 1984.

[39] Shashidhar Merugu, Mostafa Ammar, and Ellen Zegura. Routing in space and time in networks with predictable mobility. Technical report, July 2004.

[40] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, May 1967.

[41] M. Musolesi, S. Hailes, and C. Mascolo. Adaptive routing for intermittently connected mobile ad hoc networks. pages 183–189, 2005.

[42] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102+, July 2001.

[43] M. E. J. Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, January 2005.

[44] Sze-Yao Ni, Yu-Chee Tseng, Yuh-Shyan Chen, and Jang-Ping Sheu. The broadcast storm problem in a mobile ad hoc network. In *MobiCom ’99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 151–162, New York, NY, USA, August 1999. ACM Press.

[45] C. E. Perkins and E. M. Royer. Ad-hoc on-demand distance vector routing. In *WMCSA ’99: Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications*, pages 90–100. IEEE Press, February 1999.

[46] Charles E. Perkins and Pravin Bhagwat. Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. In *SIGCOMM ’94: Proceedings of the conference on Communications architectures, protocols and applications*, volume 24, pages 234–244. ACM Press, October 1994.

[47] R. C. Shah, S. Roy, S. Jain, and W. Brunette. Data mules: modeling a three-tier architecture for sparse sensor networks. In *Proceedings of the First IEEE International Workshop on Sensor Network Protocols and Applications, 2003*, pages 30–41. IEEE Press, May 2003.

[48] T. Spyropoulos, K. Psounis, and C. S. Raghavendra. Single-copy routing in intermittently connected mobile networks. In *Proceedings of the First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004*, pages 235–244. IEEE Press, October 2004.

[49] Thrasyvoulos Spyropoulos, Konstantinos Psounis, and Cauligi S. Raghavendra. Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In *WDTN ’05: Proceeding of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 252–259, New York, NY, USA, August 2005. ACM Press.

[50] Kun Tan, Qian Zhang, and Wenwu Zhu. Shortest path routing in partially connected ad hoc networks. In *Proceedings of the IEEE Global Telecommunications Conference, 2003. GLOBECOM ’03*, volume 2, pages 1038–1042. IEEE Press, December 2003.

[51] A. Vahdat and D. Becker. Epidemic routing for partially connected ad hoc networks. Technical report, April 2000.

[52] D. J. Watts and S. H. Strogatz. Collective dynamics of ’small-world’ networks. *Nature*, 393(6684):440–442, June 1998.

[53] Gang Yan, Tao Zhou, Bo Hu, Zhong Q. Fu, and Bing H. Wang. Efficient routing on complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(4), April 2006.

[54] W. Zhao, M. Ammar, and E. Zegura. 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*, pages 187–198, New York, NY, USA, May 2004. ACM Press.

Dr Elizabeth M. Daly received the B.A.B.A.I. (2001) degree in Computer Engineering and an MSc. degree in Networks and Distributed Systems (2002) from Trinity College Dublin. She received her PhD in November 2007. Dr Daly currently works in the IBM Dublin Software Lab with the Lotus Connections team and is also a part-time lecturer in Trinity College Dublin. Her research interests include social networks, peer-to-peer, distributed search and trust.

Dr Mads Haahr received BSc (1996) and MSc (1999) degrees from the University of Copenhagen and a PhD (2004) from Trinity College Dublin. He currently works as a Lecturer at Trinity College Dublin. Dr Haahr is Editor-in-Chief of Crossings: Electronic Journal of Art and Technology and also built and operates RANDOM.ORG. His current research interests are in large-scale self-organising distributed and mobile systems and in sensor-augmented artefacts.