Identifying Connectors and Communities: Understanding Their Impacts on the Performance of A DTN Publish/Subscribe System

M. Chuah, Alexandra Coman
Department of Computer Science & Engineering
Lehigh University
chuah@cse.lehigh.edu, alc308@lehigh.edu

Abstract—**Mobile devices carried by people are dynamically networked. Understanding the social structures within the human mobility traces captured from the mobile devices help us design efficient message dissemination schemes. Furthermore, community is an important attribute of future human contact-based networks. People who are in multiple communities are good message carriers. Thus, a distributed community detection scheme that can identify different communities efficiently from the various communication traces e.g. users' emails, human mobility traces is very useful. In this paper, we first identify nodes that can play key roles from some real-world human mobility and email traces using the traditional social network metrics. Then, we investigate the usefulness of several community extraction schemes that can handle both email and contact traces. Last but not least, we demonstrate how the connector identification helps to improve the performance of a DTN publish/subscribe system.**

I. Introduction

With the advancement in technology, more powerful yet small wireless devices e.g. PDAs emerge. Many users carry small mobile devices and use them to access information anywhere anytime. Such mobile devices can form peer to peer networks, e.g., Pocket Switched Network (PSN) [1]. Several researchers have attempted to design new forwarding schemes for such human-based networks e.g.[1]. Their work indicates that identifying hidden communities from the human mobility traces helps us to design more efficient schemes. In addition, identifying key nodes that bridge between various partitioned groups of nodes is equally important for such nodes can be used as forwarding agents across different communities.

Besides contact traces from mobile devices, there are other types of communication-based traces which can be derived from social networking sites e.g. Facebook, MySpace or corporate email databases. Identifying community information in such traces can also helps in designing more efficient message delivery schemes since it is common belief that it is likely for an online user to interact more frequently with someone he/she knows rather than to a stranger. In addition, some people have better human networking skills than others. Identifying such key personnel helps an organization to disseminate important information faster to its employees.

In this paper, we are interested in exploring how the identification of key role nodes and communities within a disruption tolerant network help in the design of a publish/subscribe system. First, we describe how one can identify key role nodes from two human-based contact traces and an email trace using some of the proposed centrality measures in social network analysis. Then, we explore the differences in the communities identified by two community detection schemes. Next, we investigate how the "connectors" information can be used to improve the performance of a publish/subscribe system. In summary, our contributions in this paper are as follows: (i) we compare the top 5 key nodes (users) obtained using different centrality measures and explain the differences, (ii) we compare the communities results obtained using two different community detection schemes, (iii) we compare the successful retrieval rates in a publish/subscribe system with and without using the identified key role nodes information. Our preliminary results suggest that (1) position role centrality and betweenness centrality identify similar key role nodes which can be significantly different from those identified using closeness centrality, (2) the density-based community extraction scheme we propose identifies larger communities than the SIMPLE scheme [8], and (3) the successful retrieval rates improve by at least 10% to 27% when the "connectors" information is utilized in the retrieval scheme.

The rest of the paper is organized as follows: In Section II, we discuss related work, including some existing DTN forwarding schemes which are designed for human-based mobile networks, some community extraction schemes for email/blog datasets or contact traces of mobile devices. In Section III, we describe several social network related centrality measures that we use to identify key nodes from the experimental datasets that we study. In Section IV, we describe several community extraction schemes that we explore. In Section V, we present our results and discussion. We conclude by discussing some future work in Section VI.

II. Related Work

Researchers in [1] have studied the characteristics of the human mobility traces in terms of the devices' intercontact time, the contact durations etc. In addition, based on such characteristics, some researchers have proposed more efficient forwarding schemes [1],[8] for such networks.

However, not much has been done in studying the community structures of these traces that can help in designing more efficient one-to-many, and many-to-many message dissemination strategies.

Some work has been done in extracting interesting information from the ENRON email dataset[11]. The initial work focuses more on developing learning algorithms to categorize the various emails into predefined categories [13]. More recent work focuses on identifying communities based on probabilistic models [14]. The authors in [14] proposed a probabilistic model for community extraction which leverages both topic and link information from the social network. However, their approach is a centralized community extraction method and will not be applicable to human-based contact traces which do not contain any topic-related information.

Many centralized community detection methods have been proposed for social network data in the literature. Recent review papers in this area include [3]. Such centralized methods are useful for offline data analysis on collected mobility traces to explore structures in the data. However, for self-organizing networks, it will be useful to have a distributed community detection scheme which allows each mobile device to detect its own local community.

III. IDENTIFYING CONNECTORS

Social networking researchers often use the term "connectors" (defined by Gladwell [5]) to refer to individuals who are acquainted with many individuals from different circles, and can be the bridging nodes between disjoint or weakly-linked social groups. Different social-network based centrality metrics can be used to identify nodes that play key roles in their networks. Below, we discuss a few that we use in this paper:

1. Position Role Centrality, **C_{pr} [7]**, is computed based on the network efficiency concept. C_{pr} is computed by subtracting from the current network efficiency the new network efficiency obtained by removing the ego (the node being analyzed) from the current network as shown in the following equation

$$C_{pr} = E(G) - E(G\text{-}v_i) \quad \text{Eqn (1)}$$

where E(G) is the efficiency of the network represented by graph G, and (G-v_i) is the graph obtained by removing v_i (the node for which we are currently computing Position Role Centrality) from graph G. Network Efficiency is computed as follows: $E(G) = \frac{1}{N*(N-1)}\sum_{i \neq j}\frac{1}{d_{ij}}$, where N is the number of nodes in the network graph and d_{ij} is the length of the shortest path between nodes i and j.

As connectors are individuals who hold the network together, they are defined herein as being the nodes with the values of **Position Role Centrality** that exceed a certain threshold. The removal of connectors would affect greatly the information flow within the network.

2. Betweenness Centrality, $C_B(v)$, is computed as follows [6]: $C_B(v) = \sum_{s \neq v \neq t \in V}\frac{\sigma_{st}(v)}{\sigma_{st}}$, where σ_{st} is the number of the shortest paths from s to t, $\sigma_{ss} = 1$ and $\sigma_{st}(v)$ is the number of shortest paths from s to t that include node v. If the nodes with high betweenness values are removed, then a significant number of shortest paths would become unavailable.

3. Closeness Centrality, $C_C(v)$, measures how close a node is to all other nodes, based on the shortest-path length. The equation used for computing closeness centrality is as follows [6]:

$C_C(v) = \frac{1}{\sum_{t \in V} d_G(v,t)}$, where $d_G(v,t)$ the length of the shortest path between node v and node t.

4. **Degree Centrality, $C_D(v)$** is based on the number of out-degree or in-degree of a node.

There are similarities and differences between the first two centrality measures. Position Role Centrality uses the length of the shortest path with and without the considered node while betweenness centrality considers whether there are alternate shortest paths that do not go through the considered node. All shortest paths between two nodes need to be computed before the betweenness centrality can be evaluated. Thus, for large networks, using the position role centrality may be more efficient in terms of computations.

IV. COMMUNITY DETECTION SCHEMES

In this work, we consider two community detection schemes. The first scheme is the SIMPLE scheme proposed in [8]. To understand how SIMPLE works, we need the following definitions [8]:

Familiar set: each node, say *a,* keeps a list of nodes it has encountered with the corresponding cumulative contact durations. When the cumulative contact duration with a node, say *b,* exceeds a certain threshold, T_{th} , then node b is promoted to be included in node *a*'s familiar set, *F.*

Local Community: a node's local community, denoted by *C*, contains all the nodes in its familiar set (its direct neighbors) and also the nodes that are selected by the community detection algorithm.

The SIMPLE scheme works as follows [8]:

1. Each node, v_o, maintains the following information: a list of nodes it encounters (or communicates with), and the contact duration (or number of emails exchanged), its familiar set, F_o , its local community, C_o detected so far.
2. Initialization: $C_o = \{v_o\}$, $F_o = \Phi$ (empty set)
3. When v_o encounters another node v_i, they exchange local information, i.e., v_o will acquire C_i and F_i from v_i .

4. If v_i is not in F_o , v_o updates the total contact duration (or total number of emails exchanged) of v_i until v_i falls out of contact and meanwhile the algorithm forks and proceeds to Step 5. When the total contact duration (or total number of email exchanges) has exceeded a certain threshold, v_o will insert v_i in F_o and C_o .
5. If v_i is not in C_o , then add v_i to C_o if it satisfies the following criteria: $|F_i \cap C_o| / |F_i| > \lambda$
6. If v_i is added to C_o in the previous steps, the aggressive variants of the algorithm behaves as follows:

If $|C_o \cap C_i| > \gamma |C_o \cup C_i|$, then, the two communities are merged.

In our experiment, both λ and γ are set to 0.6. The original SIMPLE scheme is designed for the contact-based traces. We adapt it for the email dataset as follows: instead of measuring cumulative contact duration, we count the cumulative number of emails exchanged between any 2 nodes.

Another class of community extraction schemes is motivated by a common understanding that a social community is a set of members who communicate more often within the set than with those outside the set. The formal definition is as follows: given a graph G(V,E), a density function *D* is defined on the set of all subsets of V. Then, a set $C \subseteq V$ is called a cluster if it is locally maximal w.r.t *D* in the following sense: for every vertex $x \in C$, removing *x* from *C* creates a set whose density is smaller than *D(C).* A density function considered in [9] is as follows: $D(C) = \frac{w_{in}}{w_{in} + w_{out}}$ - Eqn(2), where w_{in} is the number of edges uv with u,v $\in C$ and w_{out} is the number of edges uv with either $u \in C \,\&\, v \notin C$ or $u \notin C \,\&\, v \in C$. We refer to this scheme as the density-based community extraction (DCE) scheme. The density function defined in Eqn(1) may result in a large cluster with loosely connected nodes. Thus, an alternative definition of D(C) has been proposed [12] as follows:

$D(C) = \frac{w_{in}}{w_{in} + w_{out}} + \lambda \frac{2w_{in}}{|C|(|C|-1)}$ - Eqn(3) which takes into consideration the edge probability in the cluster.

In this work, we want to give different weights to the relationship between 2 nodes based on either their contact duration or their communication frequency. Thus, we propose forming an edge between two nodes only if either their contact duration or the frequency of their communications exceeds a certain threshold. Using the modified graph, we then extract the communities using Eqn(2) or Eqn(3). Unfortunately, both approaches resulted in a large cluster with loosely connected nodes. Thus, we propose the following scheme:

DCE-Scheme: In this scheme, we define the density function D(C) as in Eqn (2). Then, each node, *u,* decides to add another node, v, into its community provided *D({C+v})> D(C) and (either* $(G_1(v) \geq \eta)$ *or*$(G_2(v) \geq \eta)$ where $G_1(v) = \frac{|F(v) \cap C|}{|F(v)|}$, $G_2(v) = \frac{|F(v) \cap C|}{|C|}$. η is chosen to be 0.35.

V. RESULTS AND EVALUATIONS

A. Experimental Datasets

We use three experimental datasets: a contact-based trace gathered by the Haggle Project [1], another contact-based trace gathered by the MIT Reality Mining Project [10], and one ENRON email dataset [11]. The dataset from the Haggle Project that we use is the Infocom 2005 trace which contains 268 active devices. In the MIT Reality Mining Project, 100 smart phones running Bluetooth device discover software were deployed to students and staff at MIT and contact traces of these devices are collected. There were only traces of 97 devices in the dataset. The characteristics of the first two datasets such as the inter-contact, and contact distribution have been explored in several studies [1]. The third dataset is the UC Berkeley Enron Email dataset [11]. We only use a subset of about 1700 labeled email messages. This email dataset has also been studied in various papers [13],[14].

B. Results for the Identification of Connectors

For the contact traces of human-based networks, e.g., the MIT Reality Mining trace, an edge between two nodes is formed if the corresponding cumulative contact duration exceeds the contact duration threshold. The resulting graph formed is then fed into UCINET, a popular social networking analysis tool [2] and our program to identify the "connectors" nodes of a network and to extract communities. The **Betweenness**, **Closeness** and **Degree Centrality** were computed using UCINET while the **Position Role Centrality** was computed using our own Java program.

The key role identification results for the MIT Reality network trace with different contact duration thresholds are shown in Tables 1(a) & 1(b). From the results, one can see that there is a large overlap (80-100%) between the top 5 nodes identified using position role centrality and betweenness centrality. As the contact duration threshold is increased, some edges between certain pairs of nodes no longer exist and hence there is a slight change in the top 5 list, e.g., node 29 disappears when the threshold reaches 200,000s.

Results for the Infocom 2005 trace can be found in [4]. Similar to the MIT Reality trace results, there is a 80%

overlap between the lists of top 5 nodes returned using the Positional Role and Betweenness Centrality.

Position Role Centrality	Betweenness Centrality	Closeness Centrality	Closeness & Degree Centrality
Node 14	Node 14	Node 29	Node 29
Node 29	Node 29	Node 57	Node 14
Node 86	Node 78	Node 91	Node 57
Node 83	Node 83	Node 86	Node 86
Node 57	Node 86	Node 83	Node 83

(a) Contact Duration Threshold: 10,000s

Position Role Centrality	Betweenness Centrality	Closeness Centrality	Closeness & Degree Centrality
Node 83	Node 83	Node 78	Node 14
Node 14	Node 14	Node 14	Node 78
Node 78	Node 20	Node 86	Node 83
Node 20	Node 29	Node 32	Node 20
Node 29	Node 78	Node 76	Node 76

(b) Contact Duration Threshold: 100,000s

Table 1: Key Role Identification for MIT Reality Mining Trace

Table 2 show the "connectors" identification results for the ENRON email dataset. Here, we set the threshold for the total number of emails exchanged between two nodes before an edge between them is created to 15. The list of top 5 connectors identified using the position role centrality overlaps 60% with the top 5 connectors identified using the betweenness centrality. However, the top 5 connectors identified using the closeness centrality is very different.

Position Role Centrality	Betweenness Centrality	Closeness Centrality
Miyung Buster	Steven Kean	Richard Shapiro
Steven Kean	James.Steffes	James Steffes
John Shelk	Richard.Shapiro	Linda Robertson
Elizabeth Linnell	Linda.Robertson	Mark Palmer
Linda Robertson	Miyung.Buster	Elizabeth Linnell

Email Threshold=15

Table 2: Key Role Identification for ENRON email dataset

By looking into the relevant email messages, we found out that Miyung Buster sent emails to many recipients who do not correspond with other ENRON personnels. Thus, it is not surprising to see his name in the top 5 for position role centrality (PRC) and betweenness centrality. Similarly, Steven Kean often sent many emails to a smaller group of people, and some of his email recipients do not correspond with others. Thus, Steven Kean too will emerge as the top 5 candidates when either the PRC or the betweenness centrality is used. John Shelk also sent many emails too but unlike Miyung, his emails often have smaller sets of recipients. Again, some of John's email recipients only correspond with John, thus, John Shelk also emerges as key players using the PRC and betweenness centrality.

Closeness centrality measures how close an actor is to all other actors. Richard Shapiro corresponded with Steven Kean, Miyung Buster and John Shelk. All these 3 personnels sent emails to many recipients. Thus, it is not surprising to see Richard emerged as the top candidate when closeness centrality is used as the metric. Similarly, James Steffes, and Linda Robertson also corresponded with Steven Kean, Miyung Buster and John Shelk. Hence, they (James and Linda) too emerge as top candidates when closeness centrality is used.

C. Results for Community Extraction

Recall that for the contact-based traces, we form an edge between two nodes only if their cumulative contact duration exceeds a certain threshold (refers to as CDT) e.g. 50,000s. Once we have the graph connectivity, we use either the Simple or DCE scheme to identify the local communities.

Results for MIT Reality Trace

With a CDT of 50000s, the SIMPLE scheme (threshold=0.6) returns two big communities: Community 1 has nodes {14,18,39,57,75,81,83, 86, 95,96} while Community 2 has nodes {32,37,76,78,80,91}. With a CDT of 100,000s, the two communities remain but the number of nodes in the first cluster is smaller because the contact duration of some node-pairs is smaller than 100,000s. Community 1 now has nodes {14,18,81,83, 86,96} while the nodes in Community 2 remain the same.

With the DCE scheme (threshold=0.35), there are two big communities too. There are 33 nodes in the community that includes node 18. The additional nodes included are {7,8,15,17,21,24,28,33,34,51,54,59,60,69,70,71,75,80,89, 91}. When the threshold is increased to 0.4, only 18 nodes remain. Note that the identified members are different from the SIMPLE scheme because the DCE scheme aggressively include nodes with a high percentage of their familiar set members within the community even though these nodes are not directly connected to node 18 (the seed node). The DCE scheme excludes direct neighbors which have a high percentage of their familiar set members that are outside the community. The results for Infocom 2005 are provided in [18].

Results for ENRON email database

Threshold=15: Applying the SIMPLE scheme over the ENRON email dataset, we obtain three communities centered around 3 individuals: Steven Kean, John Shelk, Miyung Buster. For example: C1={Steven Kean, Maureen Mcvicker, Bernadette Hawkins, Elizabeth Linnell, Karen Denne, Linda Robertson, Richard Shapiro, James Steffes, Jeff Dasovich, Jeff Skilling, Joe Hartsoe, Kenneth Lay, Mark Palmer, Marck Schroeber}. When we set the threshold to be 10 emails, a 4^{th} community centered around the individual Jean Munoz emerges. In Fig 1, we illustrate the extracted communities using the SIMPLE scheme when the threshold is set to 10 emails. We can easily see the 4 mostly star-based communities.

Applying the DCE-Scheme, we also obtain three communities. The community that includes Steven Kean is the same as C1 identified in the SIMPLE scheme. Even though John Shelk and Miyung Buster correspond frequently with Steven Kean, they are not included in C1 because of the large familiar set they each have and a high percentage of the members in their familiar sets are not found in C1.

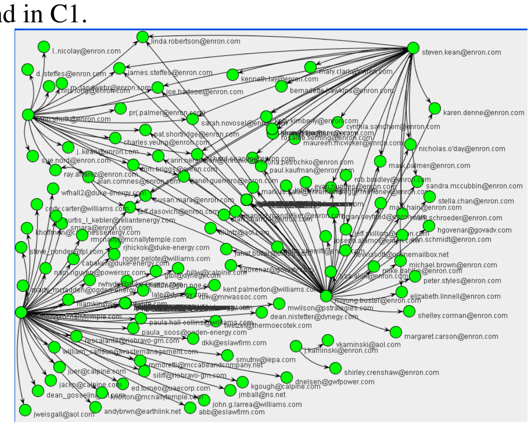


Fig 1: Identified Communities for ENRON dataset (threshold=10 emails)

VI. PERFORMANCE OF A PUBLISH/SUBSCRIBE SYSTEM

In this section, we investigate how to use the information related to the connectors to improve the performance of the data retrievals in a DTN publish/subscribe system. In the simulated DTN publish/subscribe system, there are five publishers and 10 subscribers. Each publisher publishes data items that belong to one of the 64 categories. Each data item has a certain data expiration time. When a data item expires, new one is created to replace it. The total number of data items that belong to a particular category is maintained to be about 3 throughout the whole simulation.

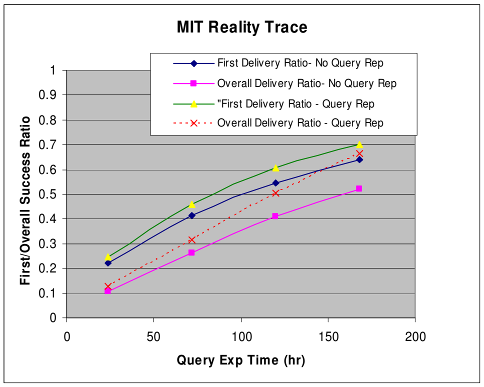
To increase the chances of a subscriber finding the data item, a publisher generates 4 copies of any data item it publishes, and spreads them to other nodes for storage. Each subscriber generates a query for data items of a particular category. Each query has a certain expiration time. When a query expires, it is removed from the querying node. Any intermediate node that has data items that belong to the requested query category will forward them when this intermediate node meets a querying node. A querying node can also choose to replicate its query to a "connector" node when the querying node encounters the "connector" node. In this case, the "connector" node can retrieve the requested data items from another node it encounters and forwards them to the querying node. We will refer to this scheme as the Connector Aided Query Replication (CAQR) Scheme. We are interested in determining how much improvement the CAQR scheme will provide when compared to the basic scheme with no query replication. In our plots below, the performance of the CAQR scheme is marked with a "Query Rep" legend.
We evaluated both schemes using all the three datasets. For the ENRON dataset, we assume that the email sent time is the encounter time between the sender and its recipients and that the contact duration at each encounter is long enough to transfer all relevant messages across. If an email has 10 recipients. we assume that there is a link between the sender of the email to each of the recipients but not links among the recipients. The metrics we used to compare the two schemes are: (i) first success rate - a query is considered successful as long as it succeeds in retrieving at least one data item, (ii) overall success rate – a query is considered successful only if it retrieves all data items of the requested category, (iii) the first retrieval delay which only considers the retrieval delay of the 1^{st} response to a query, and (iv) the overall retrieval delay which averages the retrieval delay of all unique responses.

For the MIT reality trace, we fix the data expiration time to be 7 days and vary the query expiration time from 1 day to 7 days. Figs 2 (a) & (b) plot the results with/without query replication using the MIT reality trace. It shows that with query replication, the first success rate improves by 10% (9.5%)and the overall success rate improves by 15.9% (28%) when the query expiration time is 1day (7 days). The overall delay improves by 8% when the query expiration time is 7 days.

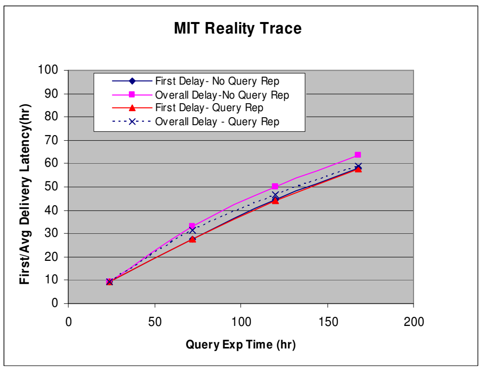
For the Infocom 2005 trace, we fix the data expiration time to 24 hours and vary the query expiration time from 3 to 24 hours. Figs 3 (a) & (b) plot the results with/without query replication using the MIT reality trace. It shows that with query replication, the first success rate improves by 8.3% (8.3%)and the overall success rate improves by 15.9% (27.5%) when the query expiration time is 1day (7 days). The overall delay improves by 9.8% when the query expiration time is 7 days.

The results for the ENRON trace are omitted due to space limitations. Our ENRON results indicate that with query replication, the first success rate improves by 5% (4.1%) and the overall success rate improves by 6.9% (5.9%) when the query expiration time is 1day (7 days). The overall delay improvement is small (0.5%). This may be due to the fact that many users are merely recipients of emails and hence are connected only to the senders. Thus, fewer alternative paths exist between different user pairs.

All our results indicate that replicating queries at connector nodes improves the retrieval performance. Note that in our design, we allow each publisher to replicate 4 copies of each data item it publishes. From our simulation traces, we found that many data items are stored in the "connector" nodes since they do meet the publishing nodes.



(a) First/Overall Success Rate vs QET

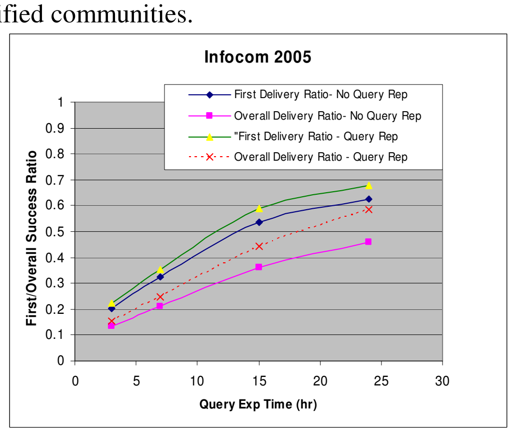


(b) FirstOverall Delay vs QET

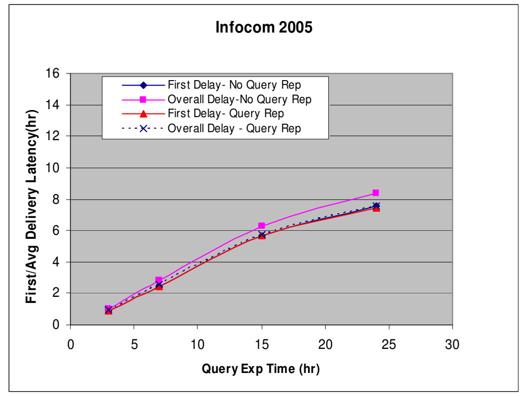
Fig 2: Retrieval Performance (MIT Reality Trace)

VII. CONCLUDING REMARKS

In this paper, we have compared the results obtained for identifying connectors from three contact-based or communication traces using various centrality measures. We also explored how two community extraction schemes behave when they are applied to these traces. Last but not least, we demonstrate that better retrieval performance can be achieved by replicating queries at the "connector" nodes. There are two directions for future work: (a) we would like to investigate how the community information can be used to further improve the performance of a DTN publish/subscribe system, and (b) we would like to investigate how we can design a multicast scheme using the combined knowledge of the connector nodes and the identified communities.



(a) First/Overall Success Rate vs QET



(b) First/Average Delivery Latency vs QET

Fig 3: Retrieval Performance (Infocom 2005)

ACKNOWLEDGMENT

The authors thank Dr. Pan Hui for sharing his Java simulator and the MIT Reality Mining and Haggle traces in a useful format.

REFERENCES

[1] A. Chaintreau et al, "Impact of Human Mobility on Opportunistic Forwarding Algorithms", Proceedings of IEEE Infocom, April, 2006
[2] UCINet:A Social Network Analysis tool www.analytictech.com/downloaduc6.htm
[3] M. Newman, "Detection Community Structure in Networks", European Physic Journal B 38:321-330, 2004.
[4] M. Chuah, A. Coman, "Understanding the impact of using connectors and communities in a DTN publish/subscribe system", CSE Technical Report, Lehigh University, May, 2009
[5] M. Gladwell, "The Tipping Point: How Little Things Can Make a Big Difference", 2000.
[6] U. Brandes, "A Faster Algorithm for Betweenness Centrality", 2001.
[7] D. Hicks, M. Nasrullah, "Detecting Key Players in 11-M Terrorist Network: A Case Study", 3rd IEEE National Conference on Availability, Reliability, and Security, pp 1254-1259, 2008.
[8] P. Hui et al, "Distributed Community Detection in Delay Tolerant Networks", Proceedings of ACM Sigcomm Workshop, MobiArch, 2007
[9] J. Baumes et al, "Dynamics of bridging and bonding in social groups, a multiagent model", Proceedings of 3rd Conference of the North American Association for Computational Social and Orgnizational Science (NAACSOS 05), Notre-Dame, Indiana, June, 2005
[10] N. Eagle, A. Pentland, "Reality mining: sensing complex social systems", Personal and Ubiquitous Computing, Vol 10(4):255-268, May, 2006
[11] UC Berkeley Enron Email Analysis, http://bailando.sims.berkeley.edu/enron_email.html
[12] M. Goldberg et al, "Communication Dynamics of Blog Networks", The 2nd SNA-KDD Workshop, Aug, 2008.
[13] R. Bekkerman et al, "Automatic Categorization of Email into Folders: Benachmark Experiments on Enron and SRI Corpora", UMASS CIIR Technical Report IR-418.
[14] N. Pathak et al, "Social Topic Models for Community Extraction", Proceedings of the 2nd SNA-KDD Workshop, Aug 2008.