On Nodal Encounter Patterns in Wireless LAN Traces

Wei-jen Hsu and Ahmed Helmy
Department of Computer and Information Science and Engineering, University of Florida
Email: {wjhsu, helmy}@ufl.edu

***Abstract*—In this paper we analyze multiple wireless LAN (WLAN) traces from university and corporate campuses. In particular, we consider an important event between mobile nodes in wireless networks –** ***encounters*. We seek to understand encounter patterns in the mobile network from a holistic view by a graph analysis approach. Such an analysis sheds light on the diverse, non-homogeneous nature of users in the given environments in terms of their encounter events with other nodes. Furthermore, we evaluate the feasibility of forming an infrastructure-less network to reach most of the nodes utilizing time-varying inter-node connectivity through encounters, and the robustness of such an ad hoc communication network.**

Our analysis shows that while the encounter events are "sparse" (i.e., any given node does not encounter with many other nodes), the connectivity of the whole network is well-maintained, and a Small World pattern of nodal encounter emerges for the observation periods longer than one day. More interestingly, the encounter events collectively form a robust communication network, in which store-carry-forward message dissemination can be mostly successful with at least 20% of non-cooperative nodes or removal of short-lived (up to minutes) encounter events.

I. Introduction

In recent years, due to the rapid increase in popularity of wireless network devices, there has been a pressing need to understand how these devices are used in realistic settings. Such investigations would help both the network administrators and the protocol or service designers to plan, design, and implement future mobile networks with the awareness of actual user behaviors and adequate solutions. To address this need, many researchers have collected detailed wireless network traces with different natures (e.g., those available at [36], [35]), with a majority of them being collected from the currently dominant wireless network technology: 802.11 wireless LANs (WLANs).

Most empirical analysis of wireless LAN (WLAN) users focuses on understanding users as individuals to unravel their behavioral patterns [3], [4], [2], [23], classify users [5], [31], or propose realistic mobility or traffic models [10], [11], [26]. The understanding of individual behavior is important in itself, but it does not reveal how the mobile nodes (MNs) in the WLAN could potentially interact with one another based on their realistic activities. In this paper, we seek an understanding beyond the level of individual users, and look into a simple yet important interaction event among MNs: *encounters*, defined as the event when two MNs move into the radio range and they are able to communicate with each other directly if the users choose to do so. Encounters are important events in wireless networks as they provide chances for MNs to directly communicate, even without an infrastructure. By studying the encounters between MNs in realistic settings, we develop an understanding of the properties of potential infrastructure-less networks in such environments. We analyze month-long WLAN traces from five university and corporate campuses, and compare our observations to distill and explain the commonalities and differences observed.

The goal of the research is to extend the WLAN trace-based analysis beyond the understanding of individual user behaviors, and seek to form a new line of analysis from inter-user relationship and network structure points of view. Specifically, this research attempts to empirically investigate and answer the following questions: (1) Given the current usage pattern of WLAN users, how do we characterize the opportunities of direct inter-user encounter patterns? Based on these encounter events, do we have sufficient opportunities to link the whole user population into one connected community? (2) How can we interpret the underlying reason of, if any, common patterns of the encounters across multiple data sets? (3) Based on encounter relationships between different user pairs, can we classify users and evaluate how such relationships affect the network-wide connectivity? and finally (4) How robust is potential encounter-based communication (when information is spread using mutual encounters between mobile nodes only) based on current encounter patterns?

We first aim to quantify the distribution of encounter events for a mobile node. From all the traces we studied, we find that the distribution of encounter events is highly asymmetric, indicating a heterogeneous user population. Surprisingly, on average, we find that a given user only encounters between 1.33% and 6.7% of the whole user population within a month. The total number of encounter events for each MN follows the BiPareto distribution and spans across several orders of magnitude. These findings suggest that the behavior of MNs is not *i.i.d.*, which is commonly assumed in many research work. Secondly, we utilize a graph analysis approach to understand the network pattern formed by encounter events among MNs. We utilize the Small World model [7] to understand the characteristics of the *encounter-relationship graphs (ER graphs)*, in which two nodes are connected by a link if they ever encounter. We find that although direct encounters of individual nodes happen only to a small portion of node pairs among the whole population, WLAN users form connected Small World graphs via encounters, and the metrics of these resulting Small World graphs (i.e., disconnected ratio, clustering coefficient, and path length) converge surprisingly quickly in about one day to its long-term steady values in most cases.

We look further into the asymmetry of inter-user relationships by quantifying the closeness (i.e., *potential friendship*) between node pairs with several metrics (e.g., based on time duration, frequency, etc. of a given user pair spends together). These friendship indexes capture the observed closeness between the involved MNs from the trace. The empirical distributions of the friendship indexes mostly follow the exponential distribution, with few node pairs showing strong relationships. Furthermore, we investigate the issue of how friendship influence the network connectivity (i.e., the graph properties of the *encounter-relationship (ER) graphs*). Our finding points out that, similar to social relationship networks [12], close friends in WLANs often form cliques and not-so-close friends are keys to widely-reached connectivity in a network.

Finally, we propose information diffusion experiments to understand how messages could be spread among users *without* the help of an infrastructure. We use a simple message spreading strategy (i.e., the epidemic routing [15]) to investigate whether it is possible to rely on mutual encounters to spread messages across the network. Surprisingly, given the seemingly very low ratio of the whole population a given node encounters with, the encounter events form a wide-reaching communication network, and the messages spread to most nodes in the population successfully. We further show that even with a relatively high percentage of users being selfish (i.e., not participating in information propagation), the information still spreads and reaches most of the population, indicating the richness of the encounter patterns in current WLAN users. Also, if encounters with short time duration are not exploited, the performances of information diffusion do not degrade significantly.

The remainder of the paper is organized as follows: We review related work in section II, and introduce the WLAN data sets we use with the basic definitions in III. We study the encounters between MNs in section IV and introduce the Small World approach to explain the encounter-relationship graph in section V. Then we discuss the findings in our effort to capture potential friendship between MNs in section VI. Finally, the information diffusion experiment is explained in section VII. We discuss the potential applications of our findings in section VIII and conclude in section IX.

II. Related Work

In this paper we utilize the WLAN traces available to the research community (e.g., [39], [40], [38], [37], downloaded from trace archives at [36] and [35]) collected from university and corporate campuses with different characteristics. Former studies on the traces focused mostly on analyzing the behavior of individual users, such as obtaining individual usage statistics [3], [4], quantifying user mobility related metrics [2], [23], classifying users based on behavioral characteristics [5], [31], or proposing mobility and traffic models for the users [10], [11], [26]. In this work we take one step further to study the *relationships* between users in the traces. We bring up new perspectives to study the WLAN traces by looking into encounter distributions and utilizing the Small World theory to describe the encounter relationship graph. Small World graph model is proposed in [7] and widely utilized to describe various networks in many areas, such as social networks, Internet topology, and electrical power networks [8]. In [9] the author applied the concept of Small World to devise a contact-based resource discovery scheme in wireless networks.

The understanding of inter-node relationships and encounter patterns in mobile networks can be useful and sometimes essential for classes of future mobile networking protocols. For example, encounter histories are used to discover routes in ad hoc network routing protocols (e.g. MAID [22], EASE [21]), and nodal encounter events are used directly in delay tolerant networks (DTNs [13]) to propagate packets. Bai et. al. find that, under mobility models with homogeneous, *i.i.d.* nodal behaviors (i.e., each node follows exactly the same model with some randomness), eventually each node encounters with all other nodes in the network (i.e., achieving 100% encounter ratio) [22]. However, the empirical observations from large WLAN traces show very different characteristics, with most nodes encountering only a very small portion of the whole population, during a time frame as long as a month. This observation indicates that the user populations in university and corporate campuses are actually *not* homogeneous.

In recent years, message forwarding in sparse, frequently disconnected mobile networks (generally known as delay tolerant networks, DTNs) receives increasing attention from the research community. Most of the previous work in this area focus on designing packet forwarding heuristics [14], [15], [17], [16]. In general, different degrees of knowledge of mobility pattern is assumed [14], or a homogeneous mobility model is used [17]. As nodal encounter events directly provide the communication opportunities in DTNs, an understanding of the nodal encounter patterns from realistic environments is important for designing DTN protocols. While designing a DTN routing protocol is not our goal in this particular study, our work complements the above studies from an empirical point of view, and investigates the issue of whether the store-and-forward model is potentially feasible under current usage pattern of wireless devices. Our findings, as shown in section VII, are encouraging. In [29], the authors also use derived encounter patterns from WLAN traces and develop a routing strategy based on the frequent association pattern of MNs, with the main focus on utilizing the encounter patterns. We, on the contrary, focus more on the understanding of encounter patterns itself.

There are also research works explicitly focused on collecting encounter traces using small mobile devices [18], [28], [27] with similar motivation as ours. We believe that the traces collected from these small-scale experiments and the large-scale WLAN traces complement each other and provide understanding of communication opportunities in different settings. We will discuss this point more in the next section when we introduce our data sets.

Similar graph analysis of potential communication opportunities has also been done based on student class schedules in a university in [32]. We must note, however, that the class registration information is an indirect indicator of the physical locations of the students, and hence does not directly translate into a graph of communication opportunities. In addition, the class schedule does not capture the mobility patterns outside of the classes. Using the traces for *actual* location information of the devices, by our discretion, seems to be a better information source to understand the communication opportunities.

III. Background

A. Wireless LAN Traces

In this study we mainly focus on WLAN traces collected from university campuses and corporations. We obtain WLAN traces from five different sources, including totally over 60,000 distinct users and over 2,000 access points (APs) in the traces. To our best knowledge this is the most extensive data set analyzed so far. Among the traces, the USC and UF traces are collected specifically for the purpose of our study, while Dartmouth [38], UCSD [40], and MIT [39] traces were collected by other research groups.

These five WLAN traces are chosen to represent different environments, user populations, location granularity, and trace-collection methods. The traces were collected with different methodologies, but we are able to derive the association history for each MN from all five traces. By association history, we mean the timestamps of events related to changes of user association, including a MN starts/ends association with an AP, or re-associates (roams) to another AP. We derive the location of a MN (i.e., the AP associated with) at any given time from these events. In these traces, each mobile node (MN) is represented by its unique MAC address. We assume that each MN is controlled by a unique user in this paper, and we use the terms MN, node, and user interchangeably.

In order to make the results we get below comparable between traces, we only analyze selected one-month chunks from the longer data sets. We summarize detailed information of these traces in Table I. Note that, we have chosen to analyze one month chunks from the traces only to keep the analysis manageable. We have taken other months from the traces to validate that the findings in the paper are not specific to the selected time frames.

Please note that the selection of data sets span across a quite long time frame, from year 2002 to 2008. While most of the traces were collected in an earlier stage of WLAN adoption, we have added a recent data set from University of Florida[1]. The importance of the Florida data set is in its recency and scale. The addition of the data set validates that our findings are valid not only in the early stage of WLAN introduction (around 2003 to 2005), but also after the technology has been widely accepted and become ubiquitous. This data set is collected from UF campus where most of the heavily populated area on campus has both indoor and outdoor WLAN coverage, and using WLAN outdoor is quite common – this justifies that the Small World encounter pattern we discover are not due to the limitation of the technology deployment (i.e., earlier WLAN deployments were more limited to indoor "hot spots" and the Small World could arise due to people clustering in these hot spots). The scale of the data set is the largest in terms of both number of access points and users.

[1]We are also in the process of anonymizing the data set and preparing it for public release on MobiLib [35].

B. Derived Encounter Events

In this paper, we focus on analyzing user encounter events based on the WLAN traces. The encounter events between users are derived based on their association history in the WLAN traces. If two users associate with the same location (i.e., switch port in the USC trace, access point (AP) for all other traces) for overlapped time intervals, they are assumed to encounter (i.e., being able to communicate) with each other. The approximation may be not completely accurate, as there can be nodes associated with the same AP but unable to communicate directly, nodes being able to communicate while associated with different APs, or nodes encountering outside the coverage of any AP. However, we believe that the encounter events derived from WLAN traces capture the major portion of MNs within direct communication range under current usage pattern.

Our approach of deriving encounter events from WLAN traces complements the experimental approach used in [18], [28], [27]. Utilizing always-on, easy-to-carry devices such as iMOTES or PDAs, in those works researchers are able to capture human encounters with high accuracy. However, encounters in human mobility do not always translate into encounters in computing devices and hence communication opportunities, as the natural usage patterns (such as devices turning on/off, preferred locations to operate the devices, etc.) are not captured by these always-on devices. The WLAN-trace-based analysis, on the other hand, captures the usage pattern truthfully, but may miss encounters out of the coverage region of access points. The two approaches have strengths in different aspects and can be complementary. In addition, deriving encounter events from WLAN traces has the advantage of obtaining a much larger data size (e.g., thousands of nodes) as compared to the experiment approach (e.g., tens to hundreds of nodes) taken in [18], [28], [27]. The WLAN traces have been collected for years in university campuses or corporations and they provide opportunities to understand encounter patterns across longer time period. We also believe that, the derived encounter events from WLAN traces occur mainly when users remain stationary and use their devices (as opposed to passing by each other while walking). These encounter events with longer durations are more useful for information exchange. In addition, many users would either turn off their devices or put them into sleep mode while moving (and not actively using the devices). In this aspect the derived encounter traces are suitable for investigation of information diffusion performances under realistic usage pattern.

IV. Encounters between Nodes

The distribution of the encounter events is the first step to understand the structure of inter-MN relationship in the traces. The direct questions to ask about the encounter events are: How many other MNs does a user meet? Do nodes meet with each other repeatedly or not? To answer these questions, we show in Fig. 1 the complementary CDF (CCDF) of the fraction of other MNs a given MN has encountered through the whole trace period (i.e., one month). From the figure we observe that all the nodes in WLAN traces encounter only at most about 50% of the user population within a month, with the UCSD trace being the only

TABLE I
STATISTICS OF STUDIED TRACES

Trace source	Unique users	Unique APs	Unique buildings	Trace duration	User type	Environment	Analyzed part in this paper	Users in analyzed part	Labels used in graphs
MIT [39]	1,366	173	3	Jul. 20 '02 to Aug. 17 '02	Generic	3 Engineer buildings	Whole trace	1,366	MIT
Dartmouth [38]	10,296	623	188	Apr. '01 to Jun. '04	Generic	University campus	Jul. 2003	2,518	Dart-03
							Apr. 2004	5,582	Dart-04
UCSD [40]	275	518	N/A	Sep. 22 '02 to Dec. 8 '02	PDA only	University campus	Sep. 22 '02 to Oct. 21 '02	275	UCSD
USC [37]	4,548	79 ports	73	Dec 03-Now (trap) Apr 20 05-Now (detail)	Generic	University campus	Apr. 20, '05 to May. 19 '05	4,528	USC
UF	44,751	728	N/A	August 2007 to Now	Generic	University campus	Jan. 14, '08 to Feb. 13, '08	32,695	UF

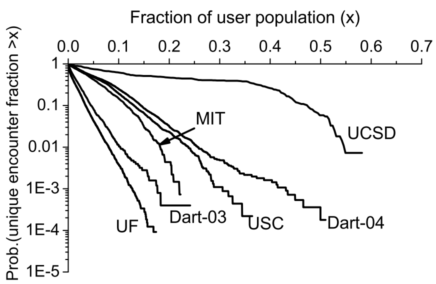


Fig. 1. CCDF of unique encounter fraction.

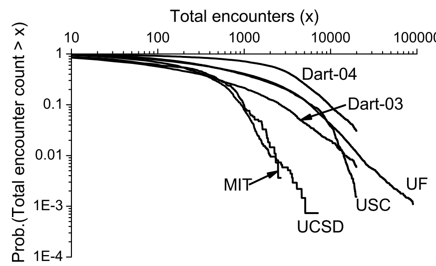


Fig. 2. CCDF of total encounter count.

exception. This may be partly due to the fact that the 275 PDA users in the UCSD trace were all selected from the freshman class, and they tend to stay in several common dorms [3] (in other words, the MNs in this trace are selected from a *correlated* sub-group of the whole population on campus). In all other traces, **on average a MN encounters with only** 1.33% **(UF) to** 6.70% **(Dart-04) of the whole user population** within the 30-day trace period. The small average encounter ratio is a combined result of several reasons: (1) most MNs are not always on, and (2) most MNs do not visit many locations [23], hence they can only meet with those who also visit this small set of locations.

Low encounter percentage as shown in the traces is not observed in typical simulation scenarios used for performance evaluation in the literature. In typical synthetic mobility scenarios [6], all nodes follow the same model to make *i.i.d.* movement decisions, and eventually encounter with all other nodes [22]. The encounter pattern from real wireless network traces reflects that university campus is a *heterogeneous* environment rather than a homogeneous one constructed by the synthetic models with statistically *i.i.d.* nodes. To better understand how protocols perform in such heterogeneous environments, using homogeneous synthetic models is not sufficient. This finding adds to the motivation of using a flexible mobility model, such as the TVC model [10], which is capable of describing nodes with diverse, heterogeneous behavior for future protocol evaluations.

We also show the CCDF of the total encounter events a MN has throughout the trace period in Fig. 2. We observe the **total encounter counts for MNs in each trace span across several orders of magnitude**. There are both MNs with extremely few or many encounters. This is another evidence of *heterogeneous behavior* among MNs. The actual number of total encounters depends on the environment and population size in the traces. However, regardless of the environment, **the curves for the total encounter count derived from WLAN traces seem to follow the BiPareto distribution**. We fit the BiPareto distribution curves to the empirical distribution curves, and use the Kolmogorov-Smirnov test [19] to examine the quality of fit. The resulting D-statistics for all traces are between 0.068 and 0.025, which indicates we have a reasonably good fit between the BiPareto distribution curves and the empirical distribution curves. Details of the Kolmogorov-Smirnov test and the parameters of the fitted BiPareto distribution curves are listed in Appendix A.

A closer investigation of the relationship between the unique encounter count and the total encounter count of the same MN reveals that **high unique encounter count does not always imply high total encounter count**. The correlation coefficients between the unique encounter count and the total encounter count for various traces range from 0.732 to 0.195. Except for the UCSD trace, all other traces have correlation coefficients below 0.6. In particular, we observe that some nodes have not many unique encounter counts, but high total encounter counts. This indicates that some node pairs may have a lot of repetitive encounters, suggesting their closer relationship than other pairs. This point warrants further study, and we will show some initial attempts on quantifying the *potential friendship* between MNs in section VI.

V. ENCOUNTER-RELATIONSHIP GRAPH

In section IV, we see that **MNs have low percentage of unique encounters among the whole population**. Given this fact, we raise a question regarding the possibility of establishing campus-wide communication among the majority of MNs via encounters alone. That is, do encounter events link MNs on the campus into one single community, or just many small cliques?

To investigate this question, we define a *static encounter-relationship graph (ER graph)* as follows: Each MN is represented by a node in the *ER graph*, and an edge is added between two nodes if the two corresponding MNs have encountered at least once during the studied trace period. By the construction of the *ER graph*, we collect all encounter events between MNs within a time period and collapse them on a static graph. The exact timing of encounters are ignored, and we focus on the structure of interconnections built between nodes by available encounter events during that period of time. The concept of *ER graph* is introduced to capture the potential of establishing a connected network among MNs based on direct encounters alone, and understand the structure of such a network.

We use three important metrics to describe the characteristics of the *encounter-relationship graphs*, defined as follows:

- **The clustering coefficient (CC)** is used to describe the tendency of nodes to form cliques in a graph. It is formally defined as [8]:

$$CC = \frac{\sum_{n=1}^{M} CC(n)}{M}, \tag{1}$$

where

$$CC(n) = \frac{\sum_{a,b \in N(n)} I(a \in N(b))}{|N(n)| \cdot (|N(n)| - 1)}. \tag{2}$$

$N(n)$ is the set of neighbors of node n in the ER graph and $|N(n)|$ is its cardinality. $I(\cdot)$ is the indicator function. M is the total number of nodes in the graph.
Intuitively, the clustering coefficient is the average ratio of neighbors of a given node that are also neighbors of one another. Higher *CC* indicates higher tendency that neighbors of a given node are also neighbors to each other, or heavy "cliquishness" in the relationship between MNs formed through encounters.

- **The disconnected ratio (DR)** is used to describe the connectivity of the ER graph. It is defined as:

$$DR = \frac{\sum_{a=1}^{M} (M - |C(a)|)}{M(M-1)}, \tag{3}$$

where $C(a)$ is the set of nodes that are in the same connected sub-graph with node a. *DR* indicates, on average, the percentage of unreachable node starting from a given node in the graph.

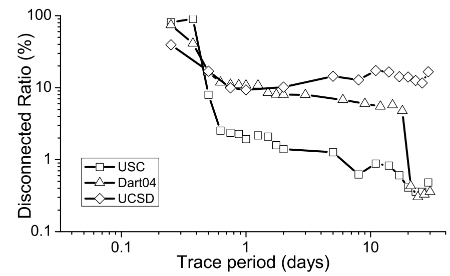
- **The average path length (PL)** is used to describe the degree of separation of nodes in the *ER graph*. It is defined as:

$$PL = (1 - DR) \cdot PL_{con} + DR \cdot PL_{disc}, \tag{4}$$

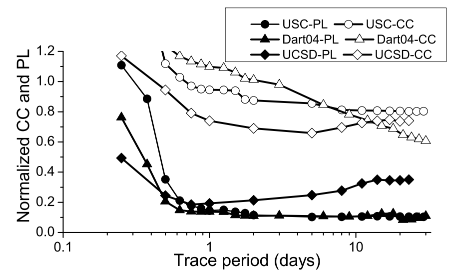
where PL_{con} is the average path length among the connected part of the *ER graph*, defined as:

$$PL_{con} = \frac{\sum_{a=1}^{M} \sum_{b \in C(a)} PL(a,b)}{\sum_{a=1}^{M} |C(a)|}. \tag{5}$$

$PL(a,b)$ is the hop count of the shortest path between node pair (a,b) in the *ER graph*[2]. PL_{disc} is the penalty on the average path length for *disconnected* node pairs in the *ER graph*. In the following we use the average path length of the regular graphs (defined later) with the same node number and average node degree for PL_{disc}.



(a)Disconnected ratio.



(b)Normalized clustering coefficient and average path length. The figure is cut from above to show the details between 0 and 1 on Y-axis.

Fig. 3. Change in the *ER graph* metrics with respect to trace period.

Taking the USC trace, the Dartmouth trace (Dart-04), and the UCSD trace as examples, we show the evolution of the three metrics with respect to various studied trace periods in Fig. 3. The graphs for other traces show very similar trends, and we leave them in Appendix B. From Fig. 3 (a) we note that given sufficient long trace durations, the *ER graphs* **have low *DR* (not larger than 10% for traces longer than one day in most cases), which implies that nodal encounters are sufficient to provide opportunities to connect almost all nodes**, even though each node encounters only a small subset of MNs directly. This is an encouraging result that points out the feasibility of building a large, widely-reach network relying only on direct encounters. Although the *DR* starts out very high with very short trace periods (i.e., for trace durations under one day) since MNs have not moved around to create encounters yet, it decreases rather quickly as the trace period increases. Within one day, the *disconnected ratios* reduce to around 10%. Although the numbers of MNs in the *ER graph* keep increasing as we look at longer trace periods, in most cases the *DR* does not change significantly after one day.

Another interesting finding is revealed by the other two graph metrics, the clustering coefficient (*CC*) and the average path length (*PL*). To highlight a unique property of these *ER graphs*,

[2]Note this path is not the same as the shortest spatial path between node pair (a,b), which may not even exist.

TABLE II
EQUATIONS FOR THE CC AND PL FOR THE REGULAR AND RANDOM GRAPHS WITH M NODES AND AVERAGE NODE DEGREE d [7], [8].

Graph type	Clustering coefficient	Average path length
Regular graph	$3(d-2)/4(d-1)$	$M/2d$
Random graph	d/M	$log(d)/log(M)$

we also calculate the *CC* and the *PL* for *regular graphs* and *random graphs* with the same corresponding total node number M and average node degree d. These quantities can be calculated according to equations in Table II. In the regular graphs, nodes are first arranged on a circle and each node is connected to d closest neighbors on the circle. In the random graphs, d randomly chosen nodes are assigned as neighbors for each node. Typically, *regular graphs* have high *CC* and *PL* while *random graphs* have low *CC* and *PL*. They are the two extreme cases on the spectrum. In Fig. 3 (b), we show the normalized *CC*'s and *PL*'s of the *ER graphs* for various trace periods. These normalized metrics represent, on the scale from 0 (corresponding to the random graph) to 1 (corresponding to the regular graph), where the metrics of the *ER graphs* fall. They are defined as:

$$CC_{norm} = \frac{CC - CC_{rand}}{CC_{reg} - CC_{rand}}, \quad (6)$$

$$PL_{norm} = \frac{PL - PL_{rand}}{PL_{reg} - PL_{rand}}, \quad (7)$$

where CC_{norm} and PL_{norm} represent the normalized CC and PL, respectively. The subscripts reg and $rand$ imply that the corresponding metric is obtained from the *regular graph* and the *random graph*, respectively, with the same total node number and average node degree.

We observe that *ER graphs* display **high normalized CC's** which are close to those of the corresponding *regular graphs* (i.e., normalized *CC*'s being close to 1, and in some cases even higher than 1), and **low normalized PL's** which are close to those of the corresponding *random graphs*. This highlights a special pattern of encounters in all WLAN traces: Nodes visiting similar sets of APs are highly likely to encounter with all others and introduce highly connected clusters among these nodes, leading to high *CC*. This phenomenon is especially obvious for very short traces, since most MNs do not change its association to the APs to create many encounters. The *ER graphs* for short trace periods feature many small disconnected cliques, each of them being a full-mesh formed by MNs associated with the same AP for that trace period. As we look at longer traces, some of the nodes in one cluster also have random encounters with nodes in other clusters, and these links serve as the "shortcuts" in the *ER graphs* that reduce the *PL*. In the literature, graphs with high *CC* close to the *regular graphs* and low *PL* close to the *random graphs* are referred to as the Small World graphs [7], [8]. By looking at various traces, we indicate that the *ER graphs* **formed by encounters among the mobile nodes appear to be Small World graphs**. We also observe that **both *PL* and *CC* converges to its final values rather quickly in about one day for most traces**, although

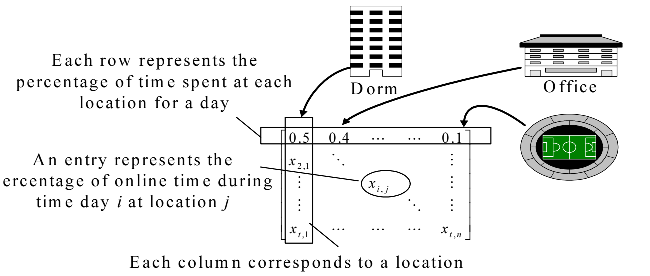


Fig. 4. Illustration of the association matrix to describe a given user's location visiting preference.

the size of *ER graphs* keeps increasing as more nodes appear in longer traces.

To further unravel the reasons of the emergence of Small World *ER graphs*, we follow up on the intuition briefly introduced in the last paragraph. We correlate the notion of similarity metric of nodal association pattern introduced in [5] and the Small World graphs to validate this intuition.

In this similarity metric, we compare the *eigen-behavior vectors* [5] of mobile nodes using the following formula:

$$Sim(User_A, User_B) = \sum_{i=1}^{rank(A)} \sum_{j=1}^{rank(B)} w_{a_i} w_{b_j} |a_i \cdot b_j|, \quad (8)$$

where vectors a_i's and b_j's are the *eigen-behavior vectors*; w_{a_i} and w_{b_j} are the corresponding weights. This is essentially the weighted cosine similarity between the two sets of *eigen-behavior vectors*. The *eigen-behavior vectors* are obtained by collecting user mobility preferences into an *association matrix*, as illustrated in Fig. 4. This *association matrix* lists in each row the percentage of online time a user spends at various locations each day, while each column in the matrix corresponds to a given location in the trace. The *eigen-behavior vectors* and its weights are obtained by applying singular value decomposition to the *association matrix*, and they serve as summaries of the major mobility trends of a given user. Essentially, this similarity metric compares how similar two users are in terms of their long-run mobility trend. We have demonstrated that this similarity metric can be used to classify users effectively into groups with similar mobility trends [5]. Here, we focus on the relationship between this similarity metric and the Small World encounter pattern we discovered.

We devise the following experiment to understand the effect of mutual similarities between users' mobility on the global encounter patterns. Using USC trace as an example, we categorize all user pairs into four zones, as illustrated in Fig. 5. Zone A consists of user pairs who are highly similar (with the similarity metric above 0.8), and zone B, C, and D consist of user pairs with less similarity in each zone. The boundaries between the zones are so chosen that, when we consider an average user, it has roughly similar number of encounters falling in each zone.

After designating user pairs into zones, we redraw the *ER graphs* to include only links between two nodes in the graph if the node pair belongs to a certain zone. This is an effort to evaluate how links among similar or dissimilar users play its roles in the resulting *ER graphs*. For *ER graphs* including links from various

TABLE III
THE GRAPH PROPERTIES OF THE ER GRAPHS WITH SELECTED LINKS (ONLY LINKS FALLING INTO CERTAIN SIMILARITY CATEGORIES (SEE FIG. 5 FOR THE BINS) ARE INCLUDED).

Links included from zone	A	B	C	D	AB	BC	CD	ABC	BCD	ABCD (all)
Average node degree	72.48	72.16	62.27	62.73	144.62	134.43	125.00	206.89	197.16	269.62
Disconnected Ratio (%)	96.85	8.98	11.35	7.25	6.36	4.22	4.26	2.40	1.49	0.53
Clustering Coefficient	0.7814	0.4568	0.1737	0.2968	0.6973	0.4896	0.3578	0.6339	0.5003	0.6117
Average Path Length	30.424	5.653	6.483	5.005	3.895	3.010	3.033	2.591	2.375	2.233

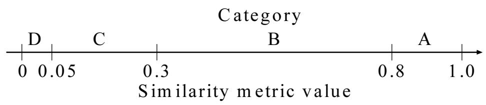


Fig. 5. Classification of node pairs into different categories based on their similarity metric range.

zones, we again obtain the three graph metrics introduced earlier this section, and summarize them in Table III.

We see from Table III that when the *ER graphs* include only edges from one zone, under similar average node degree in the *ER graph* (we have chosen the categorization bins carefully to ensure this), if the edges are formed between nodes with high similarity, it results in high disconnected ratio and clustering coefficient in general. This trend is especially pronounced for the *ER graph* including only edges in zone A, validating our intuition that extremely similar nodes (in terms of their *mobility preferences*) form disjoint clusters. The node pairs that are dissimilar to each other (e.g., node pairs in zone D) lead to an *ER graph* with low disconnected ratio, low clustering coefficient and low average path length. Similar trend is also observed when we include edges from two or three zones – indeed taking edges from only similar nodes increase the CC and PL, and the inclusion of edges between dissimilar nodes decrease DR, CC, and PL.

The above observations reveal that the heavy cliquishness in the ER graphs stems from groups of nodes visiting similar locations. Notice although it is not guaranteed that all of them end up encountering each other[3], in practice users do meet with other users with similar mobility preference with higher probability. On the other hand, we observe that as encounter events between dissimilar nodes are added into the *ER graph*, the DR, CC, and PL begin to fall, indicating the special role of "short-cuts between the cliques" played by these random links.

Finally, we acknowledge that the *ER graph* representation does not capture the time order of the encounter events. Nonetheless, it helps us to understand that different mobile users play different roles from the view point of the network-wide encounter pattern. One can leverage this understanding to incorporate network structure awareness into routing protocols. One such example is to push messages to nodes with higher centrality in the network to help message delivery [24]. Also, it has been observed that mobile nodes tend to display daily/weekly recurrent mobility pattern [23]. Thus, some of the encounter events could also potentially be recurrent. Given this, ignoring the ordering of encounter events in *ER graphs* should not diminish its usefulness completely.

In the next section, we further investigate the interplay of inter-node relationship and the *ER graph* structure, from a different perspective. We consider the notion of *potential friends* as people who I encounter repeatedly and frequently, and see how potential friendship changes the structure of the *ER graphs*.

VI. CAPTURING POTENTIAL USER FRIENDSHIP IN WLAN TRACES

In this section we quantify the *potential friendship* between MNs based on information available from the traces, and its influences on the *ER graphs*.

In our daily lives, we are bound to meet with colleagues and friends much more often than others. We investigate using the WLAN traces whether such an uneven distribution of closeness among MN pairs exists, and try to measure it using the concept of *friendship dimensions*. The likelihood or duration of encounters between two MNs captures the *closeness* between them. Although such closeness may or may not reflect *actual friendship* in a social context[4], it reveals the relationship between wireless devices as displayed in their association patterns. We propose to identify potential friendship between MN pairs based on three different dimensions – Encounter duration, encounter count, and encounter AP count, with the following definitions:

- **Friendship index based on encounter time** is defined as $Frd_t(a,b) = E_t(a,b)/OT(a)$, which is the ratio of the total encounter duration (i.e., the sum of the durations of all their encounter events) between node a and b, $E_t(a,b)$, to the total online time of node a, $OT(a)$. This is an indication of how close node b is to node a based on the duration of encounters. Note that in general $Frd_t(a,b) \neq Frd_t(b,a)$ and $0.0 \leq Frd_t(a,b) \leq 1.0$ for any node pair a and b.
- **Friendship index based on encounter count** is defined as $Frd_c(a,b) = E_c(a,b)/S(a)$, which is the ratio between the count of association sessions of node a that contain encounter events with node b, $E_c(a,b)$, to the total association session count of node a, $S(a)$.
- **Friendship index based on encounter AP count** is defined as $Frd_{AP}(a,b) = E_{AP}(a,b)/AP(a)$, which is the ratio between the number of APs at which node a has encounter events with b, $E_{AP}(a,b)$, to the total APs node a visits, $AP(a)$.

[3]One can construct a synthetic trace where a group of people visit several locations in a perfectly staggered cycle. Now while all these users are exactly the same in terms of the location visiting preferences, they never encounter each other.

[4]It is impossible to validate the actual relationship between users as the WLAN traces are anonymized. However, we borrow the term *friendship* to indicate the close relationship between MNs observed from the traces.

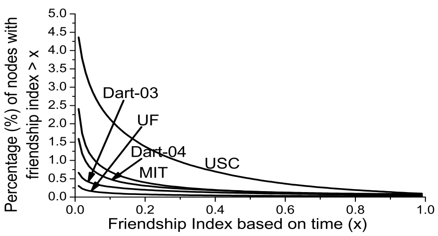


Fig. 6. CCDF of friendship index based on time.

TABLE IV
CORRELATION COEFFICIENT FOR FRIENDSHIP INDEXES $Frd(a, b)$ AND $Frd(b, a)$ FOR ALL TRACES

Trace name	Friendship index based on		
	encounter time	encounter count	AP count
MIT	0.415	0.327	0.186
UCSD	-0.024	-0.004	-0.003
USC	0.158	0.205	0.130
Dart-03	0.351	0.278	0.043
Dart-04	0.629	0.201	0.068
UF	0.190	0.091	0.036

We first observe how friendship indexes distribute among all node pairs in the traces. As shown in Fig. 6, the CCDF curves of friendship indexes based on encounter time seem to follow exponential distributions for all campuses. We again use the Kolmogorov-Smirnov test [19] to examine the quality of fit. The resulting D-statistics for all traces are between 0.0356 and 0.0052, which indicates we have a reasonably good fit between the exponential distribution curves and the empirical distribution curves. The actual parameters we use for the fitting are listed in Appendix A.

The exponential distribution of the friendship indexes is an indication that the majority of nodes do not have tight relationship with one another, even if they do encounter. In all the traces, only less than 5% of ordered node pairs (a, b) have friendship index $Frd_t(a, b)$ larger than 0.01. Among all node pairs with non-zero friendship index, only 4.47% of them have friendship index larger than 0.7, and another 11.85% of them with friendship index between 0.4 to 0.7. In other words, we can say that the friendship between the MNs is very "sparse" (i.e., only few pairs of nodes can be called "friends" based on the above definitions). Friendship indexes based on encounter frequency or encounter AP count also show similar exponential distributions.

We also look into the issue of whether the friendship index for an ordered node pair $Frd_t(a, b)$ and the reversed tuple $Frd_t(b, a)$ are symmetric. We calculated the correlation coefficients for all the traces for all three definitions of friendship indexes considered, as shown in Table IV. The resulting correlation coefficients between the friendship indexes of ordered node pair (a, b) and (b, a) are low in most cases (ranging from 0.415 to -0.024, the only exception being 0.629 for friendship index based on encounter time for Dartmouth 2004 trace), implying high asymmetry in friendship indexes.

After seeing the sparseness and high asymmetry of the friendship between the MNs, we ask the following question: if we consider friendship as a form of trust and establish inter-node communication selectively based on friendship indexes, how would that influence the structure of the *encounter-relationship graphs*? Typically, a MN may not maintain relationships with random MNs it encounters with for the first time, but is more likely to maintain connections selectively only with those MNs that are considered "trust-worthy". To better understand the interplay between the inter-node relationship and the resulting *ER graph* structure, we try to include friends with various degrees of closeness in the *ER graph*, and see how it influences the structure of the graph. We use the friendship index based on time as an example to show how different friendship levels of included encounter events can change the structure of the *ER graph* significantly.

We sort the list of nodes that node a has encountered according to friendship index, $Frd_t(a, b), \forall b \ni Frd_t(a, b) \neq 0$. After sorting, each node picks a certain percentage of nodes from the list with which to establish a link on the *ER graph*. This is equivalent to a scenario where not all encounter events are considered for inter-node communication. We choose nodes from the top, middle, or bottom of the list and with various percentages, and obtain the corresponding metrics for the new *ER graphs* that include only the links to the chosen nodes. Note that the links in these *ER graphs* are directed links when we consider friendship, as friendship is asymmetric between a given node pair. Therefore, we replace the definition of the clustering coefficient of a node in Eq. (2) by the following

$$CC(n) = \frac{\sum_{a \in F(n)} \sum_{b \in F(n)} I(a \in F(b))}{|F(n)| \cdot (|F(n)| - 1)}, \tag{9}$$

where $F(n)$ is the set of friends that node n chooses to maintain links with. Note that friendship is an asymmetric relationship, so $b \in F(a)$ does not imply $a \in F(b)$. Here the clustering coefficient is the average ratio of the included friends of a node that also include each other as a friend. When calculating the average path length and the disconnected ratio, we follow the same definitions as introduced in section V, but the paths must follow the direction of edges on the ER graph.

Following the above definitions, we obtain the metrics when including given percentages of all encountered nodes from the top, middle, or bottom of the sorted node list according to the friendship index based on time. The figures are shown in Fig. 7. We use the USC trace as an example, and similar results are also observed in other traces. The figures show a clear trend that if neighbors ranked high in the friendship index are included, the resultant ER graph shows stronger clustering, and the average path length and the disconnected ratio are much higher (i.e., inclined towards a *regular graph*). The result stems from the fact that top friends of a given node are also likely to be top friend among one another, forming small cliques in the graph. The clustering coefficient remains high due to these cliques. The disconnected ratio and the average path lengths are high due to the lack of links between different cliques. On the other hand, when low-ranked friends are included in the graph, the links included are distributed in a more random fashion, reflected by the low clustering coefficient and low average path length. Similar results are also observed in a social science study of friendship between pupils [12]. As a larger portion of friends are included in the graph, all three metrics converge to the values when all encounter events are included[5].

Therefore, although it is possible to create a campus-wide community based solely on nodal encounters, it is not sufficient to trust and utilize only top-ranked friends (or the MNs one encounters frequently), as this results in an *ER graph* with high clustering coefficient and average path length, and more importantly may lead to a disconnected network. In order to remain connected to a larger community, one should also use some randomly-chosen users with smaller friendship index as they are the key to reduce the degree of separation in the underlying *ER graph*.

VII. Information diffusion using encounters

In addition to establishing relationship between nodes, encounters can also be utilized to diffuse information throughout the network. In the network architecture generally known as delay tolerant networks (DTN), information is spread with nodal mobility and message exchanges at nodal encounters. The speed and reachability of information diffusion among the nodes are determined by the actual pattern and sequences of encounters. In this section we seek to answer the question of whether the *current* encounter patterns between MNs in wireless networks are rich enough to be utilized for information diffusion. If the answer is yes, what is the delay incurred in such a information diffusion scheme, and how robust is it?

In this section, we first understand the optimistic expectation of the potential performance of information diffusion under idealistic assumptions in subsection VII-A. We then remove some of the assumptions and evaluate the performance in more realistic settings in subsequent subsections.

A. Ideal Performance of Information Diffusion

As the first step to understand the potential of information diffusion under realistic encounter patterns, we make the following idealistic assumptions: (1) There are sufficient bandwidth and reliable communication between MNs, and sufficient storage space on all MNs. (2) MNs discover the communication opportunities immediately when they encounter other MNs, and (3) every MN in the network is willing to participate in forwarding information for others. In the experiment in this subsection, we focus mainly on analyzing how the encounter pattern itself influences the performance of information diffusion. The experiments in subsection VII-B and VII-C deal with more realistic scenarios when some of the above assumptions are removed.

The diffusion mechanism we use is the following: When a source node has some information to send, it simply transmits it to all nodes it encounters with if they have not received the information yet. All intermediate nodes cooperate in the information diffusion process, keeping a copy of received information and forwarding it the same way as the source node does. This simple approach is known as the epidemic routing in the literature [15]. Under perfect environment with sufficient resources, it achieves the lowest delay and the highest delivery rate possible. While there are many intelligent routing protocols proposed in the literature of message forwarding protocols in DTN [16], [17], [24], [29], [34], we choose to use the simple epidemic routing for the following reasons: (1) The focus of the analysis is to understand the reliability of the encounter patterns. We choose the simplest protocol to avoid complex interaction between the protocols chosen and the conclusion in our analysis. (2) Epidemic routing may represent a "last resort" of sending a message in DTN when other protocols fail. By observing when it fails in our analysis, we establish the conditions where message forwarding is not possible by *any* protocol.

In all the simulations (in this and the subsequent subsections), we use a traffic pattern in which the source node has some information to send to all other nodes. The source starts to "diffuse" the information when it is first online. As time evolves, nodes encounter with each other and an increasing portion of the whole population receive the information. We study the percentage of nodes that have not received the information within various trace periods (i.e., the unreachable ratio) and show the results in Fig. 8. Each point in the figures of this section is an average value of choosing 30% of the nodes that appear the earliest in the corresponding trace period as the sources.

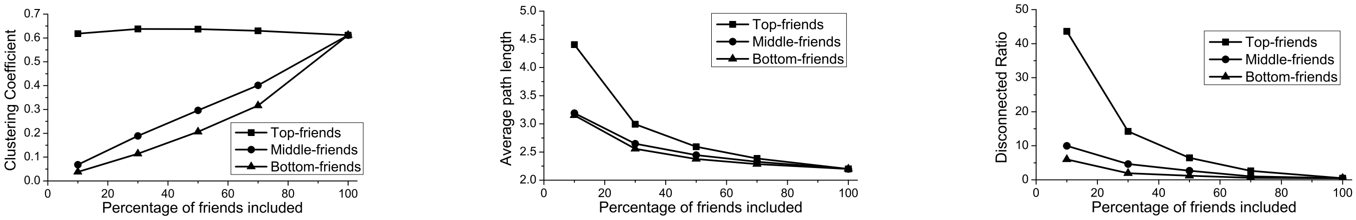
From Fig. 8 we observe that even within a short trace period (e.g., three days) the information can reach a moderate portion of the population as the unreachable ratio is less than 20% in all traces. As the trace period increases, reachability also improves. In all except the Dart-03 trace, **the unreachable ratios are less than 2% if we allow one month for the information diffusion**. Given that most nodes encounter with only a very small portion of the whole population, this result is perhaps beyond our original expectation. **It gives a positive confirmation that it is potentially possible to deliver information relying only on encounters**, in a campus environment with high success rate, under *current* user encounter pattern.

B. Performance of information diffusion with selfish users

After studying the ideal case, we consider a more realistic setup. We first relax the ideal assumption (3) above. In some cases, some of nodes may not be cooperative to propagate the information. To understand how uncooperative users potentially influence the feasibility of information diffusion, we carry out the following experiment – we make a portion of users *selfish* such that they never forward information for other sources, and we study the performance degradation under this setup. For each of the trace periods used, we increasingly make a certain percentage of nodes selfish, starting from those with the *highest unique encounter counts*. By making nodes with high unique encounter counts selfish first, we eliminate more transmission opportunities than if we pick selfish nodes randomly, hence we expect to observe a greater impact on the performance.

The relationship between the percentage of selfish node and the unreachable ratio for the USC trace is shown in Fig. 9. For

[5]Note that including 100% of friends means to include every MN encountered in the *ER graph*, hence the resulting *ER graph* is the same as the one defined earlier in section V.



(a) Clustering coefficient (b)Average path length (c)Disconnected ratio

Fig. 7. Metrics of encounter-relationship graph by taking various percentage of friends

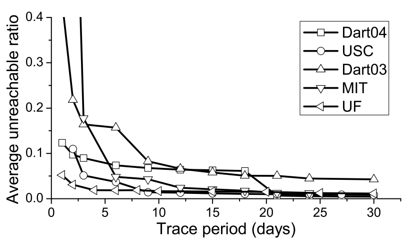


Fig. 8. Unreachable ratio of information diffusion using the epidemic routing.

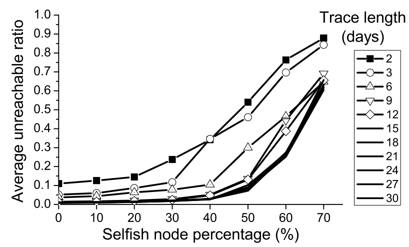


Fig. 9. The unreachable ratio with various selfish node percentage and trace period (USC trace).

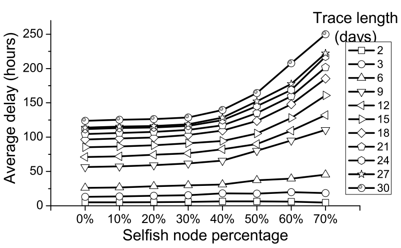


Fig. 10. The average message delay with various selfish node percentage and trace period (USC trace).

the sake of conciseness, the figures for other traces are shown in Appendix B, and they display similar trends. The result is very surprising – for all trace periods tested, the unreachable ratio does not increase significantly before at least 20% of nodes are selfish. The performance is even more robust if we take longer periods of trace. This implies that **even a significant portion of users are not willing to propagate information for others, the underlying nodal encounter pattern is rich enough for the information to find an alternative way through**. Hence the delivery rate is quite robust for up to an intermediate percentage of selfish nodes. Note that the performance of information diffusion is robust even if the nodes with the *most* chances to propagate the information are not cooperative. We show how the average delay of information diffusion changes with the increasing selfish node percentage in Fig. 10 for the USC trace. In the figure, the average delay increases for longer trace duration because information that is not deliverable in shorter trace periods becomes deliverable. More interestingly, for all tested trace durations, the average delay does not increase significantly before more than 40% of the nodes are selfish. This implies **the average delay is also robust against selfish user behavior up to an intermediate percentage**.

C. Performance of Information Diffusion with Long Encounters only

Another idealistic assumption we made is that the MNs can communicate with each other successfully regardless of the durations of encounter events. This may not be true in realistic scenario due to wireless bandwidth limitations or delay in discovering encounter events. To address this issue, we remove short-lived encounter events that do not permit prompt discovery and useful information exchange in the following experiment, and re-evaluate the performance of information diffusion with different minimum duration thresholds for an encounter event to be considered useable.

In Fig. 11, we show the relationship between the unreachable ratio versus the lower limit of encounter duration (i.e., we remove all encounter events that have shorter durations than the value), using the first 15-day traces from USC, Dartmouth, and UF as examples. From the graph we observe that, the unreachable ratio increases almost linearly as we increase the lower limit of usable encounter duration. There is no obvious point at which the performance suddenly degrades severely. We carry out the experiments up to the shortest usable encounter threshold set at *one hour*, a rather demanding scenario. Even in such cases, besides the UF trace which has a very low encounter ratio (see Fig. 1), the unreachable ratio is below 30%. This implies **removing encounters with short durations does not cause abrupt degradation in the performance of information diffusion**, in terms of both the reachability and the average delay (see Fig. 12). In other words, short encounters are not the key reason for the success of information diffusion. The encounter events with long durations are also rich enough to be utilized for message propagation in most cases.

VIII. Discussions and Future Work

In this section we discuss the lessons learned and potential applications of the findings in this paper.

In section IV we have seen very skewed encounter event distributions and low encounter percentage of a given node. This phenomenon is not observed in any of the synthetic mobility models used for performance evaluation in the literature, as most of them are simplistic *i.i.d.* models. Such a discrepancy

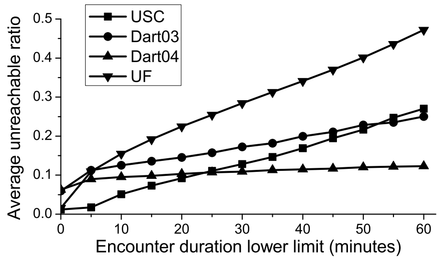


Fig. 11. The unreachable ratio after removing short encounters under the duration lower limit.

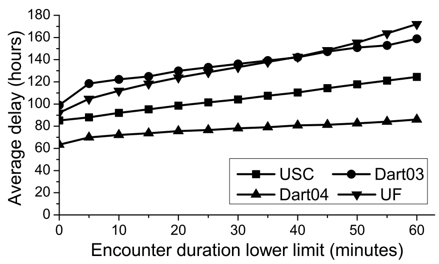


Fig. 12. The delay after removing short encounters under the duration lower limit.

calls for a realistic mobility model that not only captures the mobility features of *individual nodes*, but also retains the *inter-node relationships*. The time-variant community model [10] has the potential to construct a scenario with heterogeneous nodal behavior, given suitable parameters. In the future, we would like to construct a formal procedure to obtain from collected WLAN traces the right parameters to capture the mobility characteristics of individuals and encounter characteristics of the whole population.

The result of information diffusion experiments (section VII) highlights positive potential of building a campus-wide communication network without infrastructures. The robustness of information diffusion brings up two interesting points: (1) For message delivery, the delivery ratio and delay are not affected significantly, even if we can not choose the shortest paths due to non-cooperative users or unutilized short encounters. (2) On the other hand, it would be difficult to prevent harmful or malicious messages, such as computer worms or viruses, from propagating through encounters [33]. Both observations are due to the richness in the underlying encounter pattern providing abundant chances for message delivery. The dynamics in the occurrence of such message delivery paths also call for a more detailed investigation. Although this is partly done in [30] based on a small-scale encounter trace collected at a conference, we believe that the encounter patterns in a complex environment contain many interesting phenomenons (such as regular, predictable encounters and their role in message forwarding) to be further researched. Building on top of such an understanding, the performance of information diffusion under various information delivery schemes and potential methods to thwart malicious information from spreading are both directions for future work.

One promising direction we envision is to incorporate the understanding of encounter patterns (e.g., the Small World relationship graph) into message forwarding protocols. This approach has been taken, for example, in [24] where MNs exchange their past encountered nodes to learn the "structure" of the network, and nodes with high *centrality* measure serve as the nodes bridging the cliques in the social network. Our take on this task is slightly different – we use nodal mobility preferences as a way to *profile* the intrinsic behavior of MNs [5]. This can be done by each MN individually without any message exchange. As nodal mobility leads to encounters, our theory is that MNs with similar mobility *profile* are likely to form the cliques in the *ER graphs*. Leveraging this fact, we seek to propose a protocol that sends messages targeted at a specific *mobility profile* (e.g., those who prefer to use WLAN in the library) [34]. We would like to extend the usage of the concept of MN similarity based on the mobility profile to design efficient message dissemination strategy – leveraging the richness of random links between different cliques in the Small World *ER graph* to spread the messages fast and with low overhead.

IX. Conclusion

In this paper we investigate the network properties formed by mobile nodes via inter-node encounters, based on multiple empirical WLAN traces. We find that MNs encounter with only a small subset of other nodes (on average between 1.33% to 6.70%), and the total encounter counts follow the BiPareto distribution. In spite of low percentage of unique encounters, the *encounter relationship graph* connects most of the MNs. Furthermore, such *encounter relationship graphs* display Small World graph characteristics, and its graph metrics converge to its long-term value within short time periods. The relationship between different pairs of MNs, however, is very skewed and can be modeled by the exponential distribution. Establishing relationships only with high-ranked friends leads to a network with high clustering and disconnections, and using low-ranked friends is the key for good reachability in the encounter-relationship graphs. Finally, using simulation study with a simple protocol, we also display the potential for information diffusion without relying on the infrastructure, utilizing encounters and mobility of MNs alone.

The contributions of this work are two-folds: First, by investigating the inter-node encounters and utilizing the concept of Small World, we provide new methodologies to understand inter-user interactions in wireless networks. The understanding gained by studying distributions of encounter events and the *encounter relationship graphs* reveals the pattern of the network formed between MNs under their usage pattern in the studied environments. Second, by experimenting information diffusion with current WLAN traces, we display the potential for the success of information diffusion by the participation of only wireless users (i.e. without infrastructure). The findings could be utilized to design better user models, protocols or applications in the future, as outlined in section VIII.

References

[1] W. Hsu and A. Helmy, "On Nodal Encounter Patterns in Wireless LAN Traces," the Second International Workshop On Wireless Network Measurement (WiNMee 2006), April 2006.

[2] M. Balazinska and P. Castro, "Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network," In Proceedings of MobiSys 2003, pp. 303-316, May 2003.

[3] M. McNett and G. Voelker, "Access and mobility of wireless PDA users," ACM SIGMOBILE Mobile Computing and Communications Review, v.7 n.4, October 2003.

[4] T. Henderson, D. Kotz and I. Abyzov, "The Changing Usage of a Mature Campus-wide Wireless Network," in Proceedings of ACM MobiCom 2004, September 2004.

[5] W. Hsu, D. Dutta, and A. Helmy, "Extended Abstract: Mining Behavioral Groups in Large Wireless LANs" In Proceedings of ACM MOBICOM, Sep. 2007. Longer technical report available at http://arxiv.org/abs/cs/0606002

[6] T. Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," Wireless Communication & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, vol. 2, no. 5, pp. 483-502, 2002.

[7] D. J. Watts and S. H. Strogatz. "Collective Dynamics of 'Small-World' Networks," Nature, vol. 393, pp. 440-442, 1998.

[8] R. Albert and A. Barabasi, "Statistical mechanics of complex networks," Review of modern physics, vol. 74, no. 1, pp. 47-97 , Jan. 2002.

[9] A. Helmy, "Small Worlds in Wireless Networks," IEEE Communications Letters, pp. 490-492, Vol. 7, No. 10, October 2003.

[10] W. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, Modeling Time-variant User Mobility in Wireless Mobile Networks, In Proceedings of IEEE INFOCOM, Apr. 2007.

[11] C. Tuduce and T. Gross, "A Mobility Model Based on WLAN Traces and its Validation," in Proceedings of IEEE INFOCOM, March 2005.

[12] A. Rapoport and W. Horvath, "A Study of a Large Sociogram," Behavioral Science 6, 279-291, 1961.

[13] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets," In Proceedings of ACM SIGCOMM, August 2003.

[14] S. Jain, K. Fall, and R. Patra, "Routing in a delay tolerant network," In Proceedings of ACM SIGCOMM, Aug. 2004.

[15] A. Vahdat and D. Becker, "Epidemic Routing for Partially Connected Ad Hoc Networks," Technical Report CS-200006, Duke University, April 2000.

[16] A. Lindgren, A. Doria, and O. Scheln, "Probabilistic routing in intermittently connected networks," Lecture Notes in Computer Science, vol. 3126, pp. 239 -V 254, Sep. 2004.

[17] T. Spyropoulos, K. Psounis, and C. Raghavendra, Efficient Routing in Intermittently Connected Mobile Networks: The Single-copy Case, in ACM/IEEE journal of Transactions on Networking, Feb. 2008

[18] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, " Pocket Switched Networks and the Consequences of Human Mobility in Conference Environments," Workshop on Delay Tolerant Networking, Aug. 2005.

[19] R. Hogg and E. Tanis, "Probability and Statistical Inference," sixth edition, Prentice Hall, 2001.

[20] W. Hsu and A. Helmy, "IMPACT: Investigation of Mobile-user Patterns Across University Campuses using WLAN Trace Analysis," unpublished USC technical report. Available at http://nile.usc.edu/MobiLib/Trace_analysis_TR.pdf

[21] M. Grossglauser and M. Vetterli, "Locating Nodes with EASE: Mobility Diffusion of Last Encounters in Ad Hoc Networks," In Proceeding of IEEE INFOCOM, April 2003.

[22] F. Bai and A. Helmy, "Impact of Mobility on Last Encounter Routing Protocols," in Proceedings of the Fourth Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 2007), June 2007.

[23] W. Hsu and A. Helmy, "On Important Aspects of Modeling User Associations in Wireless LAN Traces," the Second International Workshop On Wireless Network Measurement (WiNMee 2006), April 2006.

[24] E. Daly and M. Haahr, "Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETs," In Proceedings of ACM MOBIHOC, Sep. 2007.

[25] C. Nuzman, I. Saniee, W. Sweldens, and A. Weiss, "A Compound Model for TCP Connection Arrivals for LAN and WAN Applications," Computer Networks, 40:319V337, October 2002.

[26] M. Papadopouli, H. Shen, and M. Spanakis, "Characterizing the Duration and Association Patterns of Wireless Access in a Campus," 11th European Wireless Conference 2005, Nicosia, Cyprus, April 10-13, 2005.

[27] J. Su, A. Chin, A. Popivanova, A. Goel, and E. de Lara, User mobility for opportunistic ad-hoc networking, in Proceedings of the 6th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA04), 2004.

[28] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," in Journal of Personal and Ubiquitous Computing, vol.10, no. 4, May 2006.

[29] J. Leguay, T. Friedman, and V. Conan, "Evaluating Mobility Pattern Space Routing for DTNs," in Proceedings of INFOCOM 2006.

[30] V. Erramilli, A. Chaintreau, M. Crovella, and C. Diot, "Diversity of forwarding paths in pocket switched networks," in Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC 2007), Oct. 2007.

[31] M. Kim and D. Kotz, "Periodic properties of user mobility and access-point popularity," in Journal of Personal and Ubiquitous Computing, 11(6), August, 2007.

[32] V. Srinivasan, M.Motani, and W. T. Ooi, "Analysis and Implications of Student Contact Patterns Derived from Campus Schedules," in Proceedings of MOBICOM 2006, Sep. 2006.

[33] S.Tanachaiwiwat and A. Helmy, "On the Performance Evaluation of Encounter-based Worm Interactions Based on Node Characteristics" ACM Mobicom 2007 Workshop on Challenged Networks (CHANTS 2007), Montreal, Quebec, Canada, Sep. 2007.

[34] W. Hsu, D. Dutta, and A. Helmy, " Profile-Cast: Behavior-Aware Mobile Networking," in Proceedings of IEEE WCNC, Apr 2008.

[35] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements. http://nile.cise.ufl.edu/MobiLib.

[36] CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. http://crawdad.cs.dartmouth.edu/index.php.

[37] W. Hsu and A. Helmy, MobiLib USC WLAN trace data set. Downloaded from http://nile.cise.ufl.edu/MobiLib/USC_trace/

[38] D. Kotz, T. Henderson and I. Abyzov, CRAWDAD data set dartmouth/campus/ movement/01_04 (v. 2005-03-08). Downloaded from http://crawdad.cs.dartmouth.edu/dartmouth/ campus/movement/01_04

[39] M. Balazinska and P. Castro, CRAWDAD data set ibm/watson (v. 2003-02-19). Download from http://crawdad.cs.dartmouth.edu/ibm/watson

[40] M. McNett and G. M. Voelker, Wireless Topology Discovery project data set. Download from http://sysnet.ucsd.edu/wtd/

Appendix A. BiPareto distribution and Kolmogorov-Smirnov test

In this section we first briefly introduce the Kolmogorov-Smirnov test and the BiPareto distribution, and then list the detail numerical results of fitting BiPareto and exponential distribution curves to total encounter (section IV) and friendship index (section VI) distributions, respectively.

The BiPareto distribution is used in [25] to fit the number of connections per user TCP session and mean connection inter-arrival time in a TCP session. Later, BiPareto distribution is again used in [26] to fit the distribution of association session length in wireless LAN. The CCDF of BiPareto distribution is as follows:

$$Prob(X > x) = 1,\ x \leq k, \tag{10}$$

$$Prob(X > x) = (\frac{x}{k})^{-\alpha}(\frac{x+c}{k+c})^{\alpha-\beta},\ x > k. \tag{11}$$

The left part of the CCDF curve of the BiPareto distribution on log-log scale is a straight line with slope $-\alpha$. As the x variable comes close to the turning point, c, the slope of the CCDF curve gradually changes from $-\alpha$ to $-\beta$. In our study of total encounter distributions, we choose $k = 1$ for all curves.

The Kolmogorov-Smirnov test is used to determine whether the hypothesized distribution (in our case, the BiPareto distribution) adequately fits the empirical distribution. The K-S test is not sensitive to the binning of data set, unlike the Chi-square test [19]. Therefore we choose the K-S test in our study.

TABLE V
BIPARETO DISTRIBUTION FITTING TO THE TOTAL ENCOUNTER CURVES AND THE D-STATISTICS FOR THE K-S TEST

Trace name	BiPareto parameters			D-statistics
	α	β	c	
MIT	0.027	9.8	4000	0.036
UCSD	0.062	16.3	9900	0.068
USC	0.019	0.83	550	0.049
Dart-03	0.0723	0.81	290	0.049
Dart-04	0.0285	4.43	11850	0.025
UF	0.1071	1.324	392	0.0066

TABLE VI
EXPONENTIAL DISTRIBUTION FITTING TO THE FRIENDSHIP INDEX BASED ON ENCOUNTER TIME CURVES AND THE D-STATISTICS FOR THE K-S TEST

Trace name	λ	D-statistics
MIT	369.19	0.0167
USC	305.3	0.0356
Dart-03	500.4	0.0052
Dart-04	411.81	0.0116
UF	579.06	0.0023

Referring to Fig. 13, in the K-S test the distances between the hypothesized distribution and the empirical distribution are measured throughout the range of random variable x, and the maximum of the measured distances is called the D-statistics. More formally, the D-statistics is defined as:

$$D_n = sup_x[\mid F_n(x) - F_0(x) \mid], \qquad (12)$$

where $F_n(x)$ and $F_0(x)$ are the empirical and hypothesized distributions, respectively. Intuitively, the D-statistic measures the maximum difference between the two distribution curves. A smaller D-statistic indicates a better fit of the hypothesized distribution to the empirical distribution.

We use the minimum squared error method to find the best fit of BiPareto distribution curves to the empirical total encounter distributions for various traces. The parameters are listed in Table V. From the table we observe that the D-statistics are no larger than 0.05 except for UCSD trace (0.07), indicating a reasonable fit of the BiPareto distribution.

We also list the λ parameters we obtained using the minimum squared error method to fit exponential distributions to the empirical distribution of friendship indexes based on encounter time in Table VI. The corresponding D-statistics are also listed.

APPENDIX B. ADDITIONAL GRAPHS FOR ENCOUNTER-RELATIONSHIP GRAPHS METRICS AND INFORMATION DIFFUSION EXPERIMENTS

In addition to the figures shown in section V, we also obtain the same graph metrics for MIT, Dart-03, and UF traces. The figures (Fig. 14) have similar trends as discussed in section V. One interesting observation here is that for the MIT trace, the disconnected ratio is very high until day 3 in the trace. A further investigation reveals that the MIT trace collection was started on a Saturday, and for a pure working environment (i.e., corporate buildings) Saturdays and Sundays are the least active days. The disconnected ratio is almost 100% until day 3 because the MNs that were on during the weekend are mostly stationary ones. We observe a jump of number of node in the trace, a sudden decrease in *DR*, and an abrupt change in both *CC* and *PL* on day three.

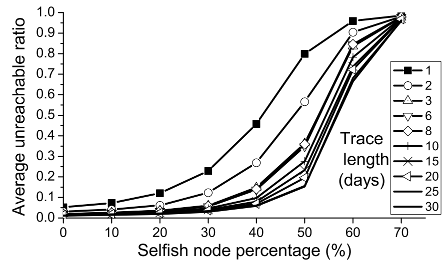


Fig. 21. Information delivery ratio with various selfish node percentage and trace period (UF trace).

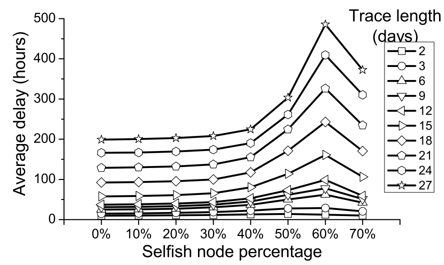


Fig. 22. Average message delay with various selfish node percentage and trace period (UF trace).

In addition to the USC trace, we further perform similar information diffusion experiments on adding selfish user behavior to the Dartmouth, MIT, and UF traces. The experiment setup is the same as described in subsection VII-B. The results for the average unreachable ratio are shown in Fig. 15, 17, 19, and 21 for the Dart-04, MIT, Dart-03, and UF traces, respectively. The trends for Dart-04, MIT, and UF traces are similar to those shown in subsection VII-B. For longer trace periods (above 9 days), the unreachable ratio does not change significantly for up to 20% of selfish nodes, and the robustness of performance increases if longer trace periods are used. This confirms that the robustness of information diffusion under *current* encounter patterns is not an artifact of coarse location granularity in the USC trace. In the Dart-03 trace, the performance of information diffusion is less robust than other traces, due to its relatively lower encounter ratio (cf. Fig. 1) and population among all the traces. The unreachable ratio for the Dart-03 trace increases faster as compared to other traces when we make users selfish. The results for the average delay are shown in Fig. 16, 18, 20, and 22 for the Dart-04, MIT, Dart-03, and UF traces, respectively. The results are similar to Fig. 10 in subsection VII-B. One noticeable difference is that, in some cases the average delay first increases as the selfish node percentage increases, but later it decreases. This is due to the low reachability (i.e., high unreachable ratio) – in this situation, only MNs that are easy to reach will be able to receive the message, leading to a decrease in the average delay (calculated from the small subgroup of still reachable MNs).

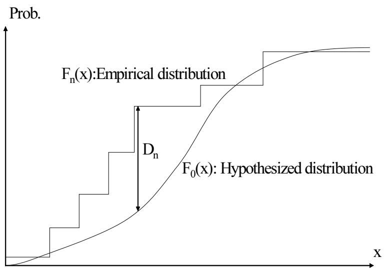
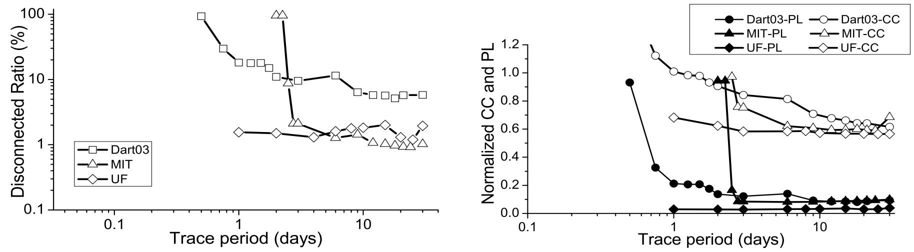


Fig. 13. Illustration of the D-statistics and the K-S test.



(a)Disconnected ratio

(b)Normalized clustering coefficient and average path length

Fig. 14. Change in the *ER graph* metrics with respect to trace period.

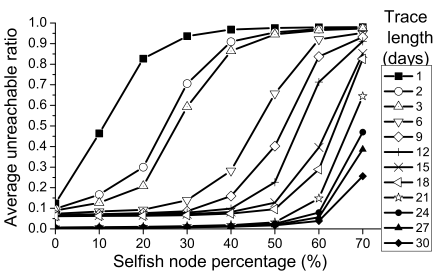


Fig. 15. Information delivery ratio with various selfish node percentage and trace period (Dart-04 trace).

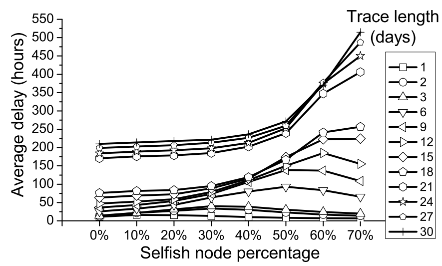


Fig. 16. Average message delay with various selfish node percentage and trace period (Dart-04 trace).

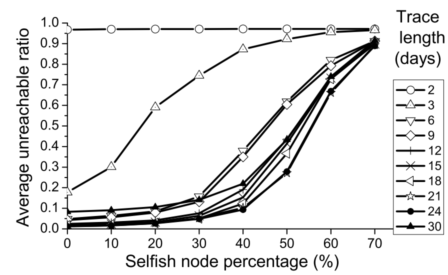


Fig. 17. Information delivery ratio with various selfish node percentage and trace period (MIT trace).

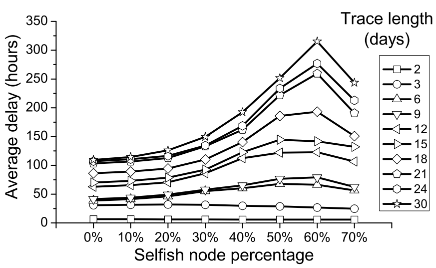


Fig. 18. Average message delay with various selfish node percentage and trace period (MIT trace).

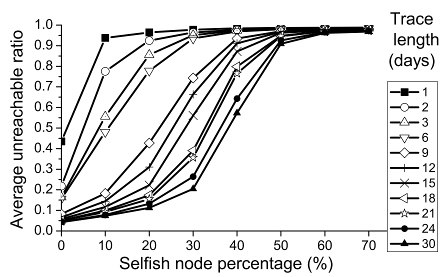


Fig. 19. Information delivery ratio with various selfish node percentage and trace period (Dart-03 trace).

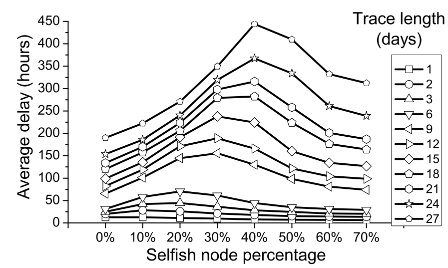


Fig. 20. Average message delay with various selfish node percentage and trace period (Dart-03 trace).

Wei-jen Hsu was born in Taipei, Taiwan, in March 1977. He received the B.S. degree in Electrical Engineering and the M.S. degree in Communication Engineering from National Taiwan University in June 1999 and June 2001, respectively. He received the Engineer degree in Electrical Engineering from University of Southern California in August 2006, and the Ph.D. degree in Computer Science from University of Florida in August 2008. His main research interest involves the utilization of realistic measurement data in various tasks in computer networks, including user modeling and behavior-aware protocol design.

Ahmed Helmy received Ph.D. in Computer Science '99 from the University of Southern California (USC), M.S. in Electrical Engineering (EE) '95 from USC, M.S. Eng. Math '94 & B.S. in EE '92 from Cairo University. He was a key researcher in the NS-2 and PIM projects at USC/ISI from '95-99. Starting '99 has been on the EE Dept faculty at USC. Since Fall '06, he became an associate professor and the director of the wireless networking laboratory in the CISE Dept at the University of Florida. He received the NSF CAREER Award in '02, the '00 Zumberge Research Award, and the best paper award at IEEE/IFIP MMNS '02. In '04 & '05, he got the best merit ranking in the EE-USC faculty. In '08 & '07, he was a finalist and winner in the ACM MobiCom SRC. His research interests include design, analysis & measurement of wireless adhoc, sensor & mobile social networks, mobility modeling, multicast protocols, IP mobility & network simulation.