An Analytical Study of People Mobility in Opportunistic Networks*

Ling-Jyh Chen[1] and Yung-Chih Chen[2]

[1]Academia Sinica

[2]University of Massachusetts Amherst

Abstract

An opportunistic network is a type of Delay Tolerant Network (DTN) in which communication opportunities are intermittent. Moreover, an end-to-end path between the source and the destination may never have existed, disconnection and reconnection are common occurrences, and link performance is highly variable or extreme. With numerous emerging opportunistic networking applications, strategies that can facilitate effective data communication in such challenging environments have become increasingly desirable. In particular, knowing the fundamental properties of opportunistic networks will soon be the key to the proper design of

opportunistic routing schemes and applications. In this study, we investigate opportunistic network scenarios based on two public network traces, namely, the UCSD and Dartmouth traces. Our contribution is twofold. First, we identify the censorship issue in network traces that usually leads to a strongly skewed distribution of the measurements. Based on this knowledge, we then apply the Kaplan-Meier Estimator to calculate the survivorship of network measurements. The survivorship feature is used to design our proposed censorship removal algorithm (CRA) for recovering censored data. Second, we perform an in-depth analysis of the UCSD and Dartmouth network traces. We show that they exhibit strong self-similarity, and can be modeled as such. We believe these newly revealed characteristics will be important in the future development and evaluation of opportunistic networks.

1 Introduction

An opportunistic network is a type of challenged network in which contacts (i.e., communication opportunities) are intermittent. Moreover, an end-to-end path between the source and the destination may never have existed, disconnection and reconnection are common occurrences, and link performance is highly variable or extreme. Therefore, traditional Internet and Mobile Ad-hoc NETwork (MANET) routing techniques can not be applied directly to opportunistic networks. However, given the numerous emerging opportunistic networking applications, such as wireless sensor networks (WSN) [11, 30, 43], underwater sensor networks (UWSN) [23], transportation networks [5, 14, 34], pocket switched networks (PSN) [16, 17, 31], and people networks [38, 41], there is an urgent need to develop effective schemes that can better accommodate the characteristics of opportunistic networks.

Clearly, knowledge of the fundamental properties of opportunistic networks

is the key to designing effective routing protocols and applications. Among the properties, knowledge of the *inter-contact time distribution* is particularly important, since it is a good indicator of network connectivity. Even so, this key property has not been studied extensively. The *inter-contact time* is the period of time between two contiguous network contacts (between a particular node pair); the longer the inter-contact time, the longer the two nodes will be disconnected. Moreover, the number of inter-contact time events indicates how many disconnection/reconnection events have occurred during the network measurement period.

Statistical analyses of opportunistic network traces are reported in [16, 17, 31]. Generally, these studies suggest using a power-law model (with heavy tails) to approximate the inter-contact time distribution of opportunistic networks. Other studies (e.g., [20, 42]) have employed a simple power-law distributed random number generator to create opportunistic network scenarios for developing and evaluating various routing (data forwarding) schemes. However, the power-law model can only fit a portion of the realistic inter-contact time distribution (i.e., the portion with a shorter inter-contact time); in other words, the heavy-tailed portion can not be approximated successfully. As a result, it is questionable whether existing schemes can maintain their performance in realistic opportunistic networks.

In this paper, we analyze opportunistic network scenarios based on real-world opportunistic people network traces. Using publicly available network traces from UCSD [4] and Dartmouth College [1], we first propose a survival analysis-based approach for handling the censorship issue in network traces. The issue arises in most network measurements, since it is inevitable that events lasting longer than the measurement period will be included in the calculations. Unlike previous studies, we consider censored measurements. Our contributions are as

follows. First, we identify the censorship issue in network measurement traces, and propose a simple yet effective censorship removal algorithm (called CRA) for recovering censored measurements. Second, using recovered network measurements, we perform a set of analysis to show the existence of self-similarity in opportunistic people networks.

The remainder of this paper is organized as follows. In Section 2, we review related works. In Section 3, we describe the basic properties of the UCSD and Dartmouth traces. Section 4 presents our survival analysis and our proposed censorship removal algorithm for the employed network traces. Section 5 details the self-similarity analysis of the recovered network traces, and Section 6 contains some concluding remarks.

2 Related Work

Recent studies of opportunistic networks have noted that the mobility model of such networks is very different to traditional models. Unfortunately, a generic model of this type of challenged network is lacking, even though it would be very useful. Various opportunistic network traces have been developed (e.g., wildlife network traces [11], people network traces [1, 4, 6, 7], pocket switched network traces [2, 9, 10], and vehicular network traces [3, 5, 8]), and several studies have been carried out using network scenarios based on these traces [14, 17, 31, 42].

Statistical analysis of opportunistic network traces are reported in [16, 17, 31]. These studies propose using the power-law distribution (with heavy tails) to model the distribution of *inter-contact time* and *contact duration* in opportunistic networks. Conan et al. [21] question the widely-held assumption that every pair of nodes in an opportunistic network has the same inter-contact time distribution. Hence, they point out the existence of heterogeneous inter-contact time distributions and show that pairs of nodes with a sufficient level of con-

4

nectivity are exponentially distributed rather than power law distributed. They conclude that one should consider the heterogeneous distribution when modeling inter-contact time distributions. Meanwhile, to develop better routing strategies, some researchers have focused on mapping wireless networks to social networks [28, 39] to gather information about node clustering and the hop distance between specific node pairs.

However, the above studies ignore the *censorship* issue, which is common in network measurements, and concentrate on fitting the distribution curve instead. In other words, a thorough statistical analysis of other fundamental network properties of opportunistic networks is lacking, even though it is widely recognized that Internet traffic is self-similar [22, 26, 35, 37]. In an attempt to address this research gap, we investigate whether the self-similarity property holds in opportunistic networks. In the following sections, we present a detailed analysis and discussion of self-similarity.

3   Opportunistic Network Traces

In recent years, researchers of DTN, MANET, and opportunistic networks have devoted significant resources to collecting realistic network traces. They realize that wireless and in-motion networks are usually much more dynamic and more unpredictable than mobility models (e.g., the Pursue Mobility Model [15], the Reference Point Group Mobility Model [29]) used in traditional mobile and wireless network studies. Network traces of opportunistic networks can be classified into several groups, such as those for wildlife networks [11], people networks [1, 4, 6, 7], pocket switched networks [2, 9, 10], and vehicular networks [3, 5, 8].

Table 1: Comparison of opportunistic network traces.

Trace Name	UCSD	Dartmouth
Device	PDA	WiFi Adapter
Network Type	WiFi	WiFi
Duration (days)	77	1,177
Granularity (sec)	120	300
No. of participating devices	273	5,148
Number of contacts	195,364	172,308,320
Avg # Contacts/pair/day	0.06834	0.01105
% of censored measurements	7%	1.3%

3.1 Trace Description

Since wildlife network traces usually have a rather small number of participants (e.g., 34 nodes in [11]), and the majority of nodes in vehicular networks usually have regular mobility patterns (buses with scheduled routes), we investigate opportunistic networks based on human mobility.

We select two publicly available opportunistic network traces, namely the UCSD [4] and Dartmouth [1] traces, because they have a large number of participating nodes and the measurement periods are sufficiently long. Similar to [17, 20, 31], we assume that two participating devices encounter a communication opportunity (i.e., a network contact) if they are both associated with the same AP at the same time. Table 1 lists the basic properties of the two traces[1].

The UCSD trace is a client-based trace that recorded the availability of WiFi-based access points (APs) for each participating portable device (e.g., PDAs and laptops) on the UCSD campus. The duration of the trace was approximately two-and-a-half months, and 273 devices participated. The Dartmouth trace, on the other hand, is an interface-based trace that recorded the APs associated

---

[1]In the Dartmouth trace, there were 13,888 devices in the network, but only 5,148 had contact with other devices. In the shortened version of the Dartmouth trace, only 3,953 nodes appeared in the 77-day period.

with a particular wireless interface over a three-year period[2]. In the following analysis, we only use a subset of the Dartmouth trace from 09/22/2002 to 12/08/2002, which is the same period as the UCSD trace[3]. We then evaluate the proposed censorship removal scheme by comparing it with the full trace, which is detailed in the next section.

3.2 Trace Properties

We focus on the network connectivity of opportunistic networks. In this section, we first describe several basic properties of the UCSD and Dartmouth traces, and then construct a connectivity graph to investigate the connectivity and mobility behavior of the nodes involved in the traces [39]. Suppose a opportunistic network is represented as an undirected graph in which the vertices represent the nodes, and two nodes are connected by an edge degree if and only if they have had contact for a specific period[4] (i.e., they have associated with the same AP). Fig. 1 shows the distribution of the degrees of nodes in the UCSD and Dartmouth traces. In the UCSD trace, on average, the degree of all nodes is 100.67 (36.6% of the nodes) and the most connected node is 210 degrees (76%). Meanwhile, the average degree of all nodes in the Dartmouth trace is 137.67 (3.48% of the nodes), and the most connected node is 770 degrees (19.5%).

There are two types of nodes in the traces. One type moves in clusters, and nodes in the same cluster contact each other frequently and have similar mobility behavior. The other type has contact with different groups of nodes

---

[2]Note that, in the Dartmouth trace, wireless interfaces can be used by different devices at different times, and each device can use multiple wireless interfaces. For simplicity, we assume each network interface represents a single mobile user in the network.

[3]Although the Dartmouth trace is longer and has more participating mobile nodes than the UCSD trace, its network connectivity is actually very poor. This is because network contacts (for each source-destination pair) occur much less frequently in the network (approximately one sixth that of the UCSD scenario).

[4]Note that, if two nodes B and C are both neighbors of the node A, it is not necessarily true that the node B is a neighbor of the node C (i.e., A may have never met B and C coincidently).

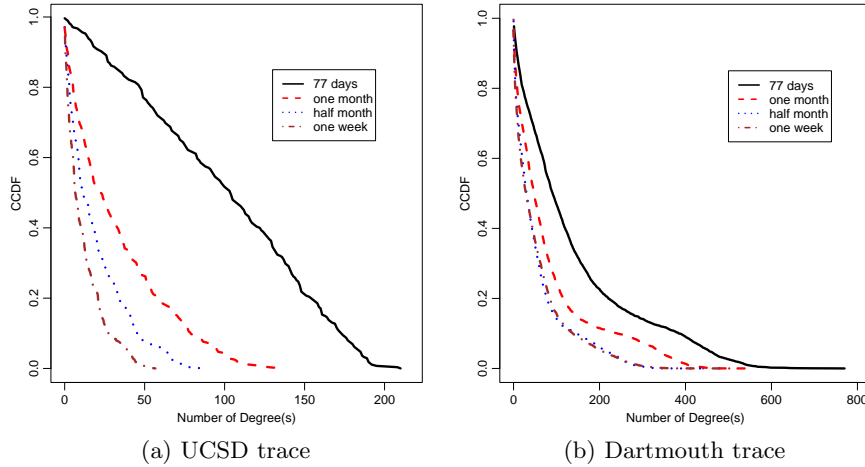(a) UCSD trace (b) Dartmouth trace

Figure 1: The CCDF of the Node Degree

and therefore has higher average degrees (i.e., they tend to meet more distinct nodes in the network). Nodes with higher degrees are able to forward data to groups farther away from the source, while those with more frequent contacts are more likely to deliver packets successfully to specific destinations. Moreover, we find that most contact opportunities are dominated by nodes that have frequent contacts. Fig. 2 shows that the most active 17% of nodes in the UCSD trace are responsible for more than 50% of the contacts, while only 7% of the nodes in the Dartmouth trace dominate more than half of the contact opportunities.

Since we know the node degree and contact frequency, it would be interesting to investigate the *degrees of separation* of node pairs. The concept of degrees of separation was first proposed in the field of sociology [36]. Unlike conventional network analysis, in opportunistic networks, we consider the number of hops from the source node to the destination node (called the hop distance) as well as the transmission delay. In this study, we compute the smallest number of hops within the minimum data forwarding time for each node pair, instead of the shortest path for each node pair. Fig. 3 shows the hop distance distribution
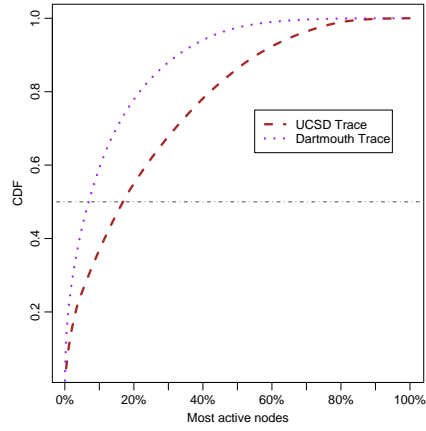
8

Figure 2: CDF of the total network contacts for the most active nodes

(a) UCSD trace
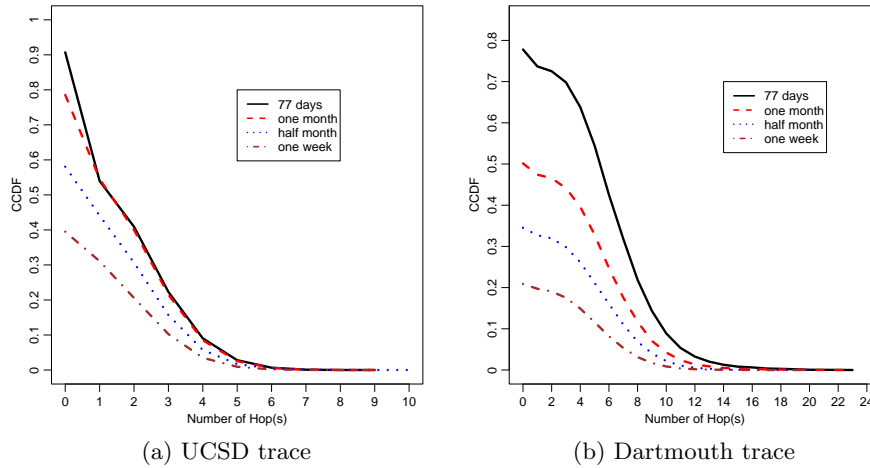

(b) Dartmouth trace

Figure 3: CCDF of the Hop Distance

for the UCSD and Dartmouth traces. In Fig. 3-a, 90% of the node pairs in the UCSD trace are reachable and the diameter of the graph is 9. This means a source node can forward data to the destination node within 9 hops, with an average hop distance of 2.43. Similarly, in Fig.3-b, 80% of the node pairs in the Dartmouth trace are reachable in the same period with a maximum hop distance of 23 and an average of 7.01.
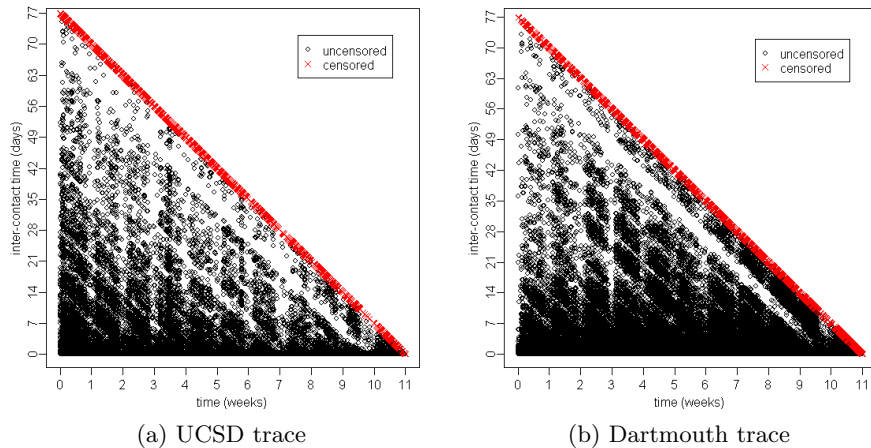
9

Figure 4: The inter-contact time distribution of the UCSD and Dartmouth traces.

As with the approaches in [16, 17, 31], our goal is to analyze the distribution of the *inter-contact time*, T_{i_c}, because this property reflects the network connectivity of the traces. We assume that a transmission contact exists between two participating devices (in ad hoc mode) if they are both associated with the same AP during some time period. Fig. 4 shows the distribution of the inter-contact time for the two opportunistic network traces. Each point in the figures represents an inter-contact time measurement that starts at the time point shown on the horizontal axis.

From Fig. 4, we observe that the inter-contact time distribution is strongly skewed and the upper-bound is a straight line (i.e., $T_{upper\_bound} = 11 - T_{cur}$, where T_{cur} is the first day of the inter-contact time in the network trace and 11 is the trace length in weeks). The data points can be classified into two groups: uncensored inter-contact time and censored inter-contact time[5]. Specifically, 7% of the inter-contact time measurements in the UCSD trace are censored, while 1.3% of the measurements in the Dartmouth trace are censored. All the

---

[5]An inter-contact time period is called censored if it starts during the measurement period, but terminates after the end of the measurement.

censored data lies on the upper bound straight line, while the uncensored data is below the line. As censorship leads to strongly skewed inter-contact time measurements, it is necessary to *recover* those censored measurements in order to analyze the opportunistic traces precisely.

4 Calibrating Censored Measurements

The inter-contact time measurement is a type of survival data (i.e., time to death or an event) [24], since an inter-contact period is likely to start while the measurement is in progress, but end after the measurement period expires. Survival analysis has been used extensively in many disciplines, such as biostatistics, bioinformatics, and life sciences, as well as in the analysis of online game traffic [18]. However, survival analysis has not been applied to mobility traces, even though censored data is prevalent and the distribution of the measurements is strongly skewed.

To address this issue, in Sub-section 4.1, we use the Kaplan-Meier Estimator, to estimate the survivorship of network traces. The proposed Censorship Removal Algorithm (CRA) is described in Sub-section 4.2, and evaluated in Sub-section 4.3.

4.1 Kaplan-Meier Estimator

The basic idea of the Kaplan-Meier Estimator (K-M Estimator, a.k.a. the Product Limit Estimator) [32] is that censored measurements have the same distribution likelihood as uncensored measurements, as long as the amount of uncensored data is sufficiently large. Specifically, we define the survival function (a.k.a. the survivorship or reliability function), $S(t)$, as the probability that an inter-contact time measurement of the network trace is larger than t; that is, $S(t) = \Pr[T_{i\_c} > t]$.

Suppose there are N distinct T_{i_c} observations in the network trace (i.e., $t_1, t_2, ..., t_N$ in ascending order such that $t_1 < t_2 < ... < t_N$), n_i events (i.e., T_{i_c} measurements) have $T_{i\_c} \geq t_i$, and d_i events are uncensored with $T_{i\_c} = t_i$. Then the K-M Estimator is a nonparametric maximum likelihood estimate of $S(t)$ defined as follows:

$$\widehat{S}(t) = \prod_{t_i \leq t} \Pr\left[t > t_i \,|\, t \geq t_i\right]$$
$$= \begin{cases} 1 \;;\; t_1 > t \\ \prod_{t_i \leq t \leq t_N} \left[\dfrac{n_i - d_i}{n_i}\right] \;;\; t_1 \leq t \end{cases} \tag{1}$$

Note that, since the calculation of the K-M Estimator is based on the likelihood of uncensored data, there is no survivorship value when $t > t_N$, which is the maximum inter-contact time measurement in the trace. Fig. 5 illustrates the complementary cumulative distribution function (CCDF) of uncensored Ti_c measurements, as well as the survival curves, using the UCSD and Dartmouth traces in log-log scale.

4.2 Censorship Removal Algorithm (CRA)

A censorship removal scheme that could recover censored measurements would be highly desirable for the analysis of inter-contact time measurements of mobility traces. Based on the K-M Estimator results, we propose a censorship removal algorithm (CRA) that calibrates the censorship based on the $\widehat{S}(t)$ estimates as follows. Let C_i/D_i denote the sets of censored/uncensored inter-contact time measurements with $T_{i\_c} = t_i$. The CRA iteratively moves a portion of censored data (based on the probability, $\frac{\widehat{S}(t_{i-1}) - \widehat{S}(t_i)}{\widehat{S}(t_i)}$) from C_i to D_i, and then moves the remaining entities of C_i to C_{i+1}, as shown in Algorithm 1.

More precisely, in each iteration, each entity of C_i will be moved to D_i with
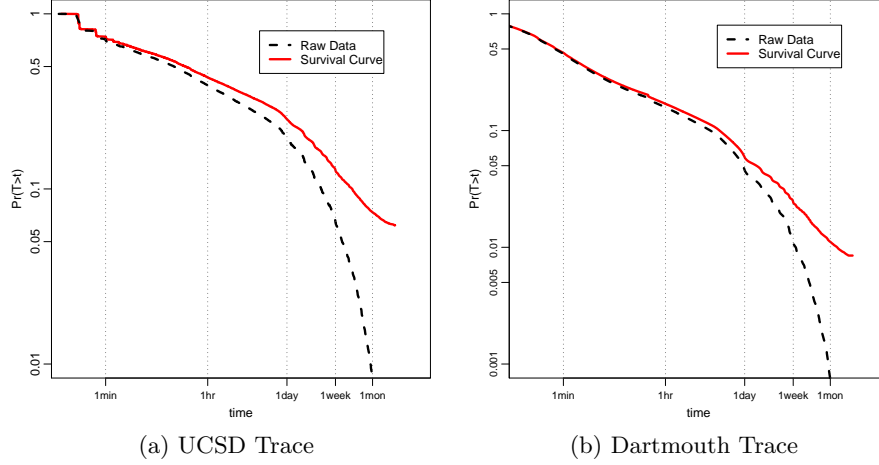
(a) UCSD Trace   (b) Dartmouth Trace

Figure 5: Comparison of CCDF and survival curves (using the K-M Estimator) of the UCSD and Dartmouth traces in logarithmic scale.

Algorithm 1 The CRA algorithm for calibrating the censorship of inter-contact time measurements in network traces.

1: **for** $i = 1$ to $N - 1$ **do**
2:      Randomly select $\frac{\widehat{S}(t_{i-1}) - \widehat{S}(t_i)}{\widehat{S}(t_i)}$ of C_i and move them to D_i
3:      Move remaining entities of C_i to C_{i+1}
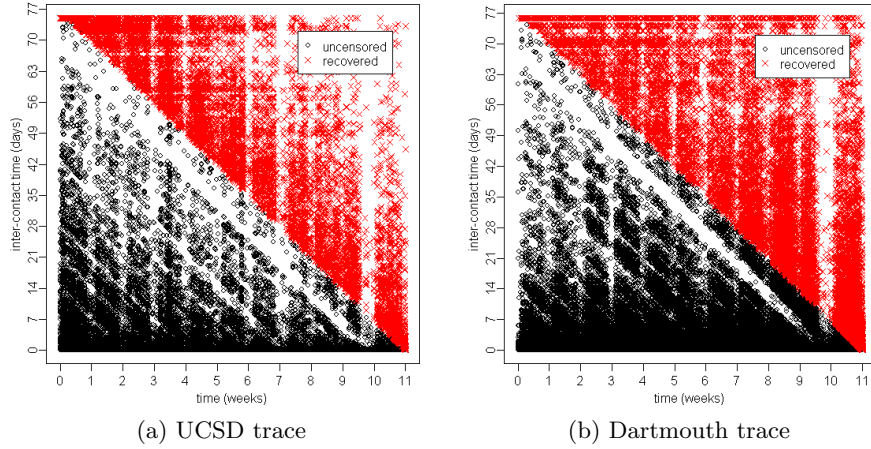4: **end for**
5: Move C_N into D_N

Figure 6: The inter-contact time distribution of the UCSD and Dartmouth traces after calibration.

probability $\frac{\widehat{S}(t_{i-1})-\widehat{S}(t_i)}{\widehat{S}(t_i)}$; otherwise, it will be moved to C_{i+1}. For simplicity, we assume the decision process is uniformly distributed. Note that the probability $\frac{\widehat{S}(t_{i-1})-\widehat{S}(t_i)}{\widehat{S}(t_i)}$ can be interpreted as the *death ratio* between t_i and t_{i+1} (i.e., $\widehat{S}(t_{i-1}) - \widehat{S}(t_i)$) with respect to (normalized by) the survivorship at t_i; that is, $\widehat{S}(t_i)$. Fig. 6 shows the inter-contact time distribution of the UCSD and Dartmouth traces after calibration.

4.3 Evaluation

We now evaluate the efficiency of the proposed CRA technique. A shortened Dartmouth trace (77 days) is employed as the raw network trace, and the full trace (1,177 days) is used to provide complete information about T_{i_c} which was censored in the shortened trace. As noted earlier, about 1.3% of the events (i.e., T_{i_c} measurements) are censored in the shortened network trace, but 80.4% of them are uncensored in the 1,177-day trace (i.e., the T_{i_c} measurement ends after the 77th day, but before the end of the network measurements). Fig. 7 compares the CCDF of the measured T_{i_c} (using the extracted trace), the
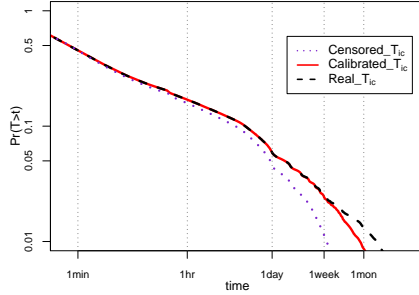
14

Figure 7: Comparison of measured T_{i_c} distribution, calibrated T_{i_c} distribution, and real T_{i_c} distribution (i.e., the full version) of the Dartmouth trace.

recovered T_{i_c}, and the real T_{i_c} (using the 1,177-day trace) of the Dartmouth trace. The results clearly show that, after applying CRA, the recovered T_{i_c} has an almost identical distribution to the real one. This demonstrates that the CRA can correctly calibrate censorship in time-limited network traces.

5 Analysis of Self-Similarity Using Opportunistic Network Traces

In this section, we analyze the self-similarity of opportunistic people network traces after calibration by the proposed CRA technique. Similar to the approaches in [16, 17], we focus on the inter-contact time measurements of the network traces because this property is an indicator of the network connectivity of an opportunistic network. In subsection 5.1, we investigate the power-law property, which exhibits a heavy-tailed distribution. Subsection 5.2 contains the definition of self-similarity. We present an analysis of self-similarity in subsection 5.3, and discuss the analysis results in subsection 5.4.

15

5.1 Heavy-Tailed Distribution

As noted earlier, it has been shown that the inter-contact time distribution of an opportunistic network follows a power-law distribution; thus, it is heavy-tailed [16, 17]. We begin with an overview of the heavy-tailed distribution, and then show that the UCSD and Dartmouth traces are both heavy-tailed. The distribution of a random variable X is called heavy-tailed if Eq. 2 is satisfied with $0 < \alpha < 2$ as $x \to \infty$, where c is a positive constant and α is the power-law exponent [22].

$$P[X > x] \sim cx^{-\alpha} \tag{2}$$

The simplest way to determine whether a distribution is heavy-tailed is to plot the complementary cumulative distribution function, $F(x)$, of the data set in log-log scale. The heavy tail index, α, can then be approximated by calculating the slope of the curve, as shown in Eq. 3.

$$\frac{d \log F(x)}{dx} \sim -\alpha \tag{3}$$

Applying Eq. 3, we find that the *alpha* value for the tail is 0.26 for UCSD trace and 0.47 for Dartmouth trace. Therefore, we conclude that both traces are heavy-tailed, which confirms the results reported in [16][17].

5.2 Self-Similarity Definition

The standard definition of a continuous-time process states that $Z = \{Z(t), t \geq 0\}$ is self-similar if it satisfies the following condition:

$$Z(t) = a^{-H} Z(at); \;\; \forall t \geq 0, \forall a > 0, 0 < H < 1 \tag{4}$$

where the equality is a finite-dimensional distribution and H is an indicator

16

of self-similarity called *hurst parameter*, which describes the self-similarity. Note that a process that satisfies Eq. 4 can never be stationary, but Z is typically assumed to have stationary increments.

A second definition of self-similarity that is more appropriate in the context of standard time series involves a stationary sequence $X = \{X(i), i \geq 0\}$. The corresponding aggregated sequence with aggregation level m can be obtained by dividing the original series X into non-overlapping blocks of size m. Each block is then averaged, as shown in Eq. 5, where k indexes the block number.

$$X_k^{(m)} = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i), \quad k = 1, 2, ..., \frac{N}{m} \tag{5}$$

If X is an incremental process of Z, i.e., $X(i) = Z(i+1) - Z(i)$, then, for all integers of m, we can obtain

$$X = m^{1-H} X^{(m)} \tag{6}$$

Based on Eq. 6, we define self-similarity as follows:

1. A stationary sequence $X = \{X(i), i \geq 0\}$ is called exactly self-similar if it satisfies Eq. 6 for all m aggregated levels.

2. Stationary sequence $X(i)$, where $i \geq 1$, is said to be asymptotically self-similar if Eq. 6 holds as $m \to \infty$.

3. A covariance-stationary sequence $X(i)$, where $i \geq 1$, is exactly second-order self-similar or asymptotically second order self-similar if $m^{1-H} X^{(m)}$ has the same variance and autocorrelation as X for all m, or as $m \to \infty$.

The degree of self-similarity of a series can be expressed by a single parameter, called the hurst parameter H, which indicates the speed of decay of the

autocorrelation function of the series. If a series is self-similar, $1/2 < H < 1$. Moreover, as H approaches 1, the degree of self-similarity increases.

5.3 Graphical Methods and Statistical Analysis

As mentioned above, the most attractive property of the self-similar process is that the degree of self-similarity is expressed by the magnitude of the hurst parameter H. In other words, the statistical properties of a self-similar process do not change with different aggregation levels. In this subsection, we apply four techniques (a variance-time plot, a rescaled adjusted range plot, a periodogram plot, and the Whittle estimator) [12, 13, 27] to investigate the self-similarity of network traces. We present the analysis in the following subsections.

5.3.1 Variance-Time Plot

The variance-time plot tests the property of slowly decaying variance that exists in self-similar series. The variance of the process $X^{(m)}$ is plotted against the aggregated level m in a log-log plot. The m-aggregated process $X^{(m)} = (X_1^{(m)}, X_2^{(m)}, X_3^{(m)}, ...)$ is defined by Eq. 7, where m and j are positive integers.

$$X_j^{(m)} = \frac{1}{m} \sum_{i=m(j-1)+1}^{jm} X_i \; ; j = 1, 2, 3, ..., \frac{N}{m} \tag{7}$$

The variance of the process can thus be obtained by:

$$Var[X^{(m)}] = \frac{1}{\frac{N}{m} - 1} \sum_{j=1}^{N/m} (X_j^{(m)} - \bar{X})^2 \tag{8}$$

Fig. 8 depicts the variance-time plot with various aggregation levels for the UCSD and Dartmouth traces. From the figures, we observe that the aggregated variances of the inter-contact time measurements in both traces are nearly linear

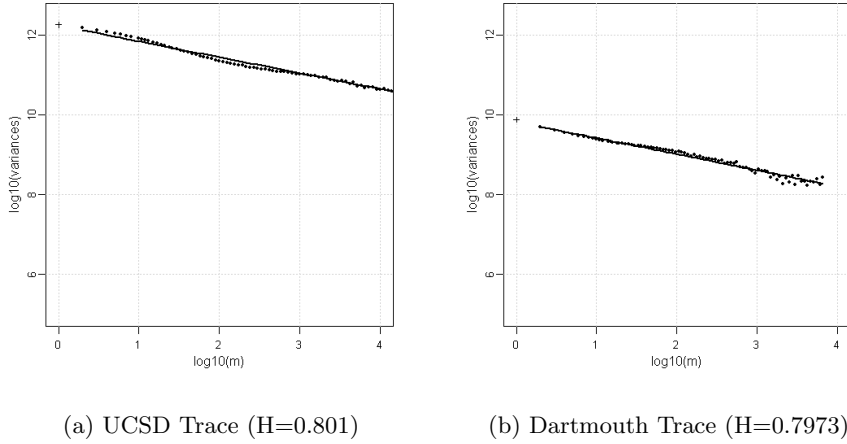(a) UCSD Trace (H=0.801)   (b) Dartmouth Trace (H=0.7973)

Figure 8: Graphical Analysis of the Aggregated Variance Method.

and could be fitted by a simple least squares line with a slope smaller than -1, which is an indicator of self-similarity. Moreover, the hurst parameter H can also be derived from the absolute value of slope β by $H = 1 - |\beta|/2$. For instance, the slope of the fitting curve in Fig. 8-a is estimated by regression as -0.4; therefore, the hurst parameter, H, is estimated to be 0.8. In Fig. 8-b, however, the slope is estimated to be about -0.405 and H is estimated to be about 0.7973.

5.3.2 Rescaled Adjusted Range Plot

One key property of self-similarity is that a dataset keeps the same statistical properties no matter how many sub-datasets it is divided into [27]. The second analysis technique, called the R/S method, sequentially divides a dataset by dichotomy to calculate the rescaled adjusted range for each sub-dataset, and then takes the average of all the calculated values. The R/S method is subject to the exponent H of the power law, which acts as a function related to the number of points involved.

For instance, given a dataset $X_1, X_2, X_3, ... X_n$ with sample mean $\mu = E[X_i]$,

an adjusted partial sum is defined as

$$W_k = (X_1 + X_2 + X_3 + ... + X_n) - k\overline{X}(n) \tag{9}$$

where $k = 1, 2, 3, ..., n$ and $\overline{X}(n)$ is the arithmetic mean of the first n obser-vations. The range $R(n)$ is also defined by

$$R(n) = \max(0, W_1, W_2, ..., W_n) - \min(0, W_1, W_2, ..., W_n) \tag{10}$$

Suppose $S(n)$ denotes the standard deviation of the sample size n, then the R/S value of the dataset can be calculated by Eq. 11.

$$E\left[\frac{R(n)}{S(n)}\right] \to cn^H \text{,as } n \to \infty \tag{11}$$

Fig. 9 shows the R/S plot (i.e., the R/S values against n in a log-log plot) of the network traces, and the hurst parameter, H, is estimated by the regres-sion slope. Specifically, the estimate of H is 0.7472 in the UCSD trace and 0.7493 in the Dartmouth trace. The values indicate that the inter-contact time measurements of both traces are self-similar.

5.3.3 Periodogram Plot

The periodogram is defined by Eq. 12, where ν is a frequency, N is the series length, X is the series, and $I(\nu)$ is the estimator in the spectrum field [13].

$$I(\upsilon) = \frac{1}{2\pi N}\left|\sum_{j=1}^{N} X(j)e^{ij\nu}\right|^2 \tag{12}$$

A periodogram plot can be obtained by collecting multiple periodograms of various frequency values. The plot can be fitted using a straight line in log-log scale, and the slope, β, of the fitting line can be approximated by $1 - 2H$. Note

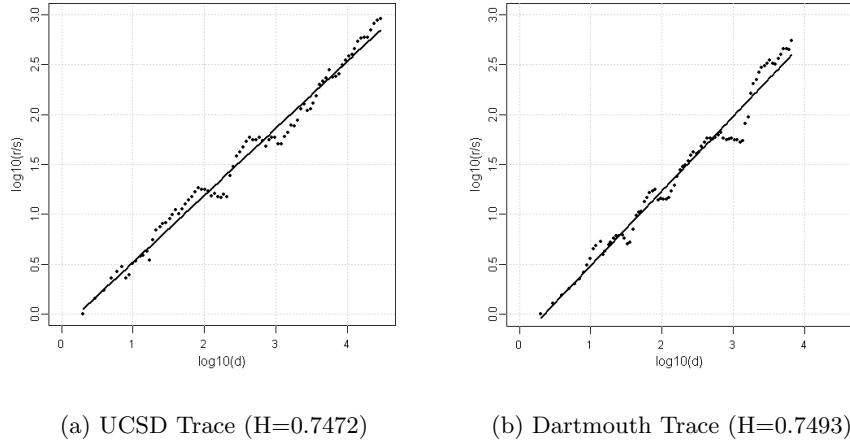(a) UCSD Trace (H=0.7472)   (b) Dartmouth Trace (H=0.7493)

Figure 9: Graphical Analysis of the R/S Method.

that, in practice, researchers usually use the lowest 10% of frequencies [40] to make a periodogram plot, since the power law only holds for frequencies close to zero. Fig. 10 illustrates the periodogram plots of the UCSD and Dartmouth traces. The slope of the fitting line is approximately -0.56 in the UCSD trace and -0.53 in Dartmouth trace; therefore, the hurst estimate is about 0.78 in UCSD trace and 0.76 in the Dartmouth trace. Once again, the values confirm that the inter-contact time measurements are self-similar in both network traces.

5.3.4   The Whittle Estimator

The Whittle estimator is generally regarded as the most robust indicator of self-similarity because it provides a confidence interval for the whole estimation procedure. There are two types of Whittle estimator models, namely, the Fractional Gaussian Noise (FGN) model with $1/2 < H < 1$ and the Fractional ARIMA (p,d,q) model with $0 < d < 1/2$ [22]. The main difference between the two methods is that ARIMA assumes the existence of short-range dependency, but FGN does not. Since we are interested in the long-range dependency of the

21

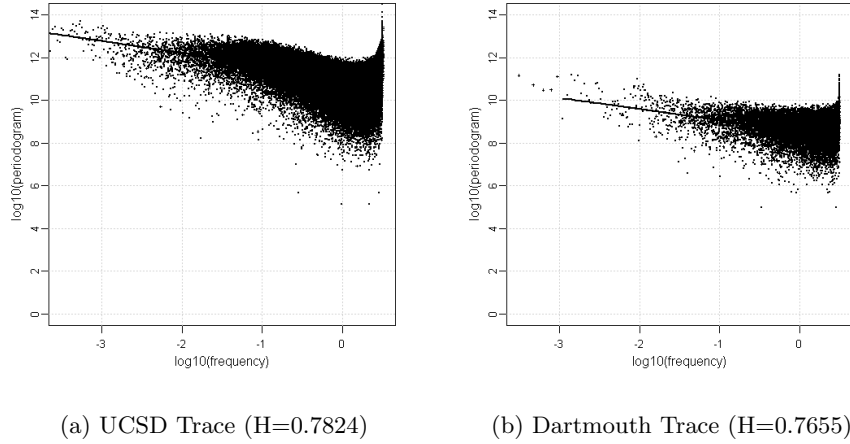(a) UCSD Trace (H=0.7824)   (b) Dartmouth Trace (H=0.7655)

Figure 10: Graphical Analysis of the Periodogram Method.

two network traces, we apply the Whittle estimator with FGN by aggregating the datasets into different levels. As shown in Fig. 11, the Whittle estimator is stabilized at about 0.8 for the UCSD trace, and the results of the three graphical methods are all within the 95% confidence interval (when the aggregation level is greater than 1000). The results also show that the Whittle estimator is stabilized at about 0.75 for the Dartmouth trace, and the results of the three graphical methods are all within the 95% confidence interval when the aggregation level is greater than 1000. Therefore, we conclude that the inter-contact time measurements of the UCSD and Dartmouth traces are self-similar.

5.4   Discussion

Recall that Fig. 4 and Fig. 6 show that there exist some regular patterns in the distribution of inter-contact time measurements of the UCSD and Dartmouth traces. More specifically, from the figures, one can easily observe that "*inter-contact time gaps*" appear about once a week, and each gap lasts about one day. The phenomenon can be explained by the fact that students tend to have contact
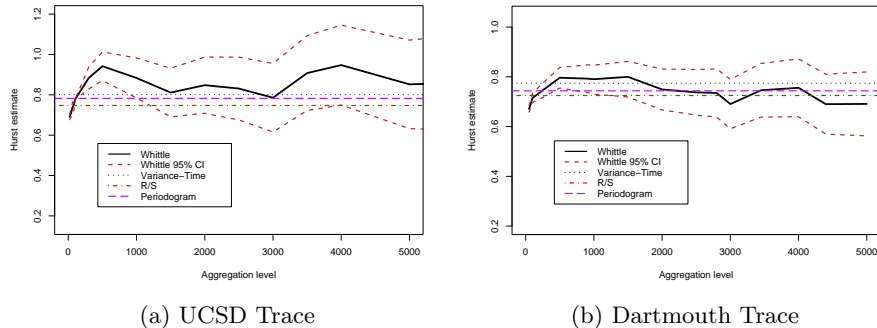
(a) UCSD Trace (b) Dartmouth Trace

Figure 11: Graphical Analysis of the Whittle Estimator.

with each other on weekdays only. As they do not usually visit the campus on weekends, there are very few wireless network associations with campus access points in the network traces.

The estimation of the hurst parameter has been shown to be unreliable in the presence of strong periodic components, trends, and non-stationarity [33]. In the inter-contact time process, we can easily remove trends and non-stationarity through elaborately designed sub-sampling and short sampling interval when dealing with the traces. However, periodicity can not be easily removed without loss of sampling accuracy, and we do not consider people's behavior as an artificial mechanism or that it is characterized by periodicity [25].

In Fig. 4, most uncensored measurements are clustered at the bottom of the figure, which means their inter-contact times are small. The results indicate that once a pair of nodes make contact, they are more likely to meet each other again in the very near future. For instance, in campus scenarios, most mobile devices are carried by students, and they tend to team up and move around the campus in clusters (e.g., attending classes, visiting the library, and going to restaurants on the campus). Based on the mathematical analysis and our explanation, we conclude that the process of opportunistic people networks under such scenarios

is self-similar.

6   Conclusion

We have investigated the fundamental properties of opportunistic people networks. Using public network traces from UCSD and Dartmouth College, we identify the censorship issue in the traces that usually leads to strongly skewed distribution of the measurements. Based on this knowledge, we apply the Kaplan-Meier Estimator to calculate the survivorship of network measurements, which we then use to design our proposed censorship removal algorithm (CRA) for recovering censored data. We show that, after applying CRA, the inter-contact time distribution of the recovered network trace is almost identical to that of the real trace. Additionally, we conduct a detailed analysis of the UCSD and Dartmouth traces and show that they exhibit strong self-similarity. We also note the importance of these newly revealed characteristics for future research into opportunistic people networks. We believe the results of this study are significant and should be taken into consideration in the design, evaluation, and deployment of future opportunistic network applications.

Acknowledgements

References

[1] Crawdad project. http://crawdad.cs.dartmouth.edu/.

[2] Haggle project. http://www.haggleproject.org/.

[3] Monarch project. http://www.monarch.cs.rice.edu/papers.html.

[4] Ucsd wireless topology discovery project. http://sysnet.ucsd.edu/wtd/.

[5] Umass diverse outdoor mobile environment - umass dieselnet. http://prisms.cs.umass.edu/dome/index.php?page=umassdieselnet.

[6] Unc/forth archive of wireless traces, models, and tools. http://www.cs.unc.edu/Research/mobile/datatraces.htm.

[7] Usc wireless lan traces. http://nile.cise.ufl.edu/MobiLib/USC_trace_intro.html.

[8] Vehicular networks - georgia institute of technology. http://www-static.cc.gatech.edu/~mpalekar/Vehicular_Networks.htm.

[9] Wireless lan traces from acm sigcomm 2001. http://sysnet.ucsd.edu/pawn/sigcomm-trace/.

[10] Wireless lan traces from acm sigcomm 2004. http://www.cs.washington.edu/research/networking/wireless/.

[11] The zebranet wildlife tracker. http://www.princeton.edu/~mrm/zebranet.html.

[12] H. Abrahamsson. Traffic measurement and analysis. Technical Report T99:05, Swedish Institute of Computer Science, September 1999.

[13] J. Beran. *Statistics for Long-Memory Processes*. Chapman & Hall/CRC, 1 edition, October 1994. ISBN: 0412049015.

[14] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networks. In *IEEE Infocom*, pages 1688–1698, 2006.

[15] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communication and Mobile Computing Journal*, 2(5):483–502, 2002.

[16] A. Chaintreau, P. Hui, J. Crowcroft, C. D. andRichard Gass, and J. Scott. Pocket switched networks: Real-world mobility and its consequences for opportunisticforwarding. Technical Report UCAM-CL-TR-617, University of Cambridge, Computer Laboratory, February 2005.

[17] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, June 2007.

[18] K.-T. Chen, P. Huang, G.-S. Wang, and C.-Y. H. andChin Laung Lei. On the sensitivity of online game playing time to network qos. In *IEEE Infocom*, pages 2387–2398, 2006.

[19] L.-J. Chen, Y.-C. Chen, T. Sun, P. S. andKuan Ta Chen, C.-H. Yu, and H.-H. Chu. Finding self-similarities in opportunistic people networks. In *IEEE Infocom*, pages 2286–2290, 2007.

[20] L.-J. Chen, C.-H. Yu, T. Sun, Y.-C. Chen, and Hao-huaChu. A hybrid routing approach for opportunistic networks. In *ACM SIGCOMM Workshop on Challenged Networks*, pages 213–220, 2006.

[21] V. Conan, J. Leguay, and T. Friedman. The heterogeneity of inter-contact time distributions: its importance for routing in delay tolerant networks. http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0609068, 2006.

[22] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE /ACM Transactions on Networking*, 5(6):835–846, December 1997.

26

[23] J.-H. Cui, J. Kong, M. Gerla, and S. Zhou. The challenges of building mobile underwater wireless networks for aquatic applications. *IEEE Network*, 20(3):12–18, May-June 2006.

[24] R. C. Elandt-Johnson and N. L. Johnson. *Survival Models and Data Analysis*. Wiley-Interscience, September 1980. ISBN: 0471031747.

[25] D. Figueiredo, B. Liu, A. Feldmann, V. Misra, and D. T. andW. Willinger. On tcp and self-similar traffic. *Performance Evaluation*, 61:129–141, July 2005.

[26] M. Garrett and W. Willinger. Analysis, modeling and generation of self-similar vbr video traffic. *ACM SIGCOMM Computer Communication Review*, 24(4):269–280, October 1994.

[27] M. Gospodinov and E. Gospodinova. The graphical methods for estimating hurst parameter of self-similar networktraffic. In *International Conference on Computer Systems and Technologies*, 2005.

[28] A. Helmy. Small worlds in wireless networks. *IEEE Communications Letters*, 7(10):490–492, October 2003.

[29] X. Hong, M. Gerla, R. Bagrodia, and G. Pei. A group mobility model for ad hoc wireless networks. In *ACM International Workshop on Modeling, Analysis and Simulation of Wirelessand Mobile Systems*, pages 53–60, 1999.

[30] J.-H. Huang, S. Amjad, and S. Mishra. Cenwits: A sensor-based loosely coupled search and rescue system using witnesses. In *ACM International conference on Embedded networked sensor systems*, pages 180–191, 2005.

[31] P. Hui, A. Chaintreau, J. Scott, R. Gass, JonCrowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In

ACM SIGCOMM workshop on Delay-tolerant networking, pages 244–251, 2005.

[32] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observation. *Journal of the American Statistical Associatio*, 53(282):437–481, June 1958.

[33] T. Karagiannis, M. Faloutsos, and R. Riedi. Long-range dependence: now you see it, now you don't! In *IEEE Global Telecommunications Conference*, pages 2165–2169, 2002.

[34] J. LeBrun, C.-N. Chuah, and D. Ghosal. Knowledge based opportunistic forwarding in vehicular wireless ad hoc networks. In *IEEE Vehicular Technology Conference (VTC-Spring)*, pages 2289–2293, 2005.

[35] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[36] S. Milgram. The small world problem. *Psychology Today*, 1:61–67, 1967.

[37] V. Paxson and S. Floyd. Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

[38] D. Snowdon, N. Glance, and J.-L. Meunier. Pollen: using people as a communication medium. *Elsevier Computer Networks*, 35(4):429–442, February 2001.

[39] V. Srinivasan, M. Motani, and W. T. Ooi. Analysis and implications of student contact patterns derived from campus schedules. In *ACM International Conference on Mobile Computing and Networking*, pages 86 – 97, 2006.

[40] M. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: an empirical study. http://math.bu.edu/people/murad/pub/estimators-posted.ps, 1995. preprint.

[41] R. Y. Wang, S. Sobti, N. Garg, E. Z. andJunwen Lai, and A. Krishnamurthy. Turning the postal system into a generic digital communication mechanism. *ACM SIGCOMM Computer Communication Review*, 34(4):159–166, October 2004.

[42] Y. Wang, S. Jain, M. Martonosi, and K. Fall. Erasure coding based routing for opportunistic networks. In *ACM SIGCOMM Workshop on Delay Tolerant Networks*, pages 229–236, 2005.

[43] Y. Wang and H. Wu. Delay/fault-tolerant mobile sensor network (dft-msn): A new paradigm for pervasive information gathering. *IEEE Transactions on Mobile Computing*, 6(9):1021–1034, September 2007.