中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-06-009

Analysis of Opportunistic Networks based on Realistic Network Traces

Yung-Chih Chen, Paruvelli Sreedevi, Kuan-Ta Chen, and Ling-Jyh Chen

August 8, 2006 || Technical Report No. TR-IIS-06-009

http://www.iis.sinica.edu.tw/LIB/TechReport/tr2006/tr06.html

Analysis of Opportunistic Networks based on Realistic Network Traces

Yung-Chih Chen Paruvelli Sreedevi Kuan-Ta Chen
Ling-Jyh Chen

Institute of Information Science
Academia Sinica
Taiwan
{ycchen, devirao, ktchen, cclljj}@iis.sinica.edu.tw

3 August 2006

Abstract

Opportunistic network is a type of Delay Tolerant Networks (DTN) where network communication opportunities appear opportunistic, an end-to-end path between source and destination may have never existed, and disconnection and reconnection is common in the network. With numerous emerging opportunistic networking applications, strategies that can enable effective data communication in such challenged networking environments have become increasingly desirable. In particular, knowing fundamental properties of opportunistic networks will soon become the key for the proper design of oppor-

tunistic routing schemes and applications. In this study, we investigate opportunistic network scenarios based on two public network traces, namely UCSD and Dartmouth network traces. In this paper, our contributions are the following: First, we identify the censorship issue in network traces that usually leads to strongly skewed distribution of the measurements. Based on this knowledge, we then apply the Kaplan-Meier Estimator to calculate the survivorship of network measurements, which is used in designing our proposed censorship removal algorithm (CRA) that is used to recover censored data. Second, we perform a rich set of analysis illustrating that UCSD and Dartmouth network traces shows strong self-similarity, and can be modeled as such. Third, we pointed out the importance of these newly revealed characteristics in future development and evaluation of opportunistic networks.

1 Introduction

Opportunistic network is a type of challenged networks, where network contacts (i.e., communication opportunities) are intermittent, an end-to-end path between the source and the destination may have never existed, disconnection and reconnection is common, and/or link performance is highly variable or extreme. Therefore, traditional Internet and Mobile Ad-hoc NETwork (MANET) routing techniques can not be directly applied towards networks in this category. With numerous emerging opportunistic networking

applications, such as wireless sensor networks (WSN) [11][28][42], underwater sensor networks (UWSN)[22], transportation networks [2][14][31], pocket switched networks(PSN) [17][16][29], people networks [37][40], and etc., it remains desirable/necessary to develop effective schemes that can better accommodate the characteristics of opportunistic networks.

Knowing fundamental properties of opportunistic networks is the key for the design of effective routing protocols and/or applications. Among all, knowledge of *inter-contact time distribution* is particularly important, since this distribution provides a good description of network connectivity. Yet,this important fundamental property has not been extensively studied in the literature. By *inter-contact time*, we mean that the time duration between two contiguous network contacts (between a particular node pair). The larger the inter-contact time is, the longer the two nodes are disconnected. Moreover, the more inter-contact time events in the network trace, the more reconnection/disconnection events have occurred during the network measurement period.

Statistical analysis of opportunistic network traces has recently been undertaken by [16][17][29]. These studies generally suggest a power-law model (with heavy tails) to approximate the inter-contact time distribution of opportunistic networks, and several studies (e.g., [19] [41]) have employed a simple power-law distributed random number generator to create an opportunistic network scenario 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 smaller inter-contact time), whereas the heavy-tailed portion can not be successfully approximated. As a result, it is still questionable if previously proposed schemes can remain their performance in realistic opportunistic network scenarios.

It is the interest of this study to further analyze opportunistic network scenarios based on realistic opportunistic people network traces. Using publicly available network traces from UCSD [5] and Dartmouth college [1], we first propose a survival analysis based approach to cope with censorship among network traces. The censorship issue commonly exists in most network measurements since it is inevitable to have measured events lasting longer then the measurement period. While previous studies simply ignore censored measurement data, our contributions are the following: First, we identify the censorship issue in network measurement traces, and propose a simple yet effective algorithm (called CRA) to recover censored measurements. Second, using recovered network measurements, we perform a set of analysis showing the existence of self-similarities in opportunistic people networks. Lastly, we pointed out the importance of these characteristics in future development and evaluation of opportunistic networks.

The rest of the paper is organized as follows. In section 2, we summarize related work in this area. In section 3, we briefly describe the basic properties of the opportunistic network traces examined. Section 4 presents our survival analysis and our proposed censorship removal algorithm for the employed

network traces. Section 5 performs self-similarity analysis on the recovered network traces. Finally, section 6 concludes the paper.

2 Related Work

Research activities of opportunistic networks have been carried out with focus mostly on the design of effective data forwarding schemes that can provide high performance data delivery and remain resilient against extremely poor network connectivity. For instance, replication based routing schemes have been proposed to inject multiple identical copies of data into the network, and rely on node mobility to disseminate the data toward the destination [39]. However, the main drawback of this type of schemes is the tremendous traffic overhead associated with flooding data replicates. As a result, when network resources (e.g., buffer space or network bandwidth) are limited, replication based schemes tend to degrade performance of reliability (i.e, the delivery ratio) unless additional overhead reduction techniques are implemented (e.g., [14][26][32][35]).

Additionally, coding based routing schemes have been proposed recently to transform a message into another format prior to transmission [41] [43]. The design principle of coding based schemes is to embed additional information (e.g., redundancy [41] or decoding algorithm [43]) within the coded blocks such that the original message can be successfully reconstructed with only a certain number of the coded blocks. The performance of coding based

routing schemes can be further improved by dynamically adapting encoding levels [34] and/or combining replication based techniques [19].

However, while most previous work simply employ some well-known random way-point mobility model (e.g., the Pursue Mobility Model [15] and the Reference Point Group Mobility Model [27]) in the performance evaluation of opportunistic network scenarios, recent studies have noticed that the mobility model of opportunistic networks is far different from traditional ones, and a generic model for this type of challenged networks is still lacking and highly desired. Various opportunistic network traces have been contributed to the research community (e.g., wildlife network trace [11], people network traces [1][5][6][7], pocket switched network traces [3][9][10], and vehicular network traces [2][4][8]), and several studies have been carried out using network scenarios based on these network traces [14][17][29][41].

Statistical analysis of opportunistic network traces has been performed [16][17][29], and the power-law distribution (with heavy tails) has been proposed to model the distribution of *inter-contact time* and *contact duration* in opportunistic networks. However, as we will elaborate later in this paper, these studies simply ignore the presence of *censorship* that are common in network measurements, and they only concentrate on fitting the distribution curve whereas thorough statistical analysis of other fundamental network properties are still lacking. Particularly, while Internet traffic has been well-recognized to be self-similar [20][24][33][36], it is one of our interests to investigate whether the same property holds in opportunistic networks. We

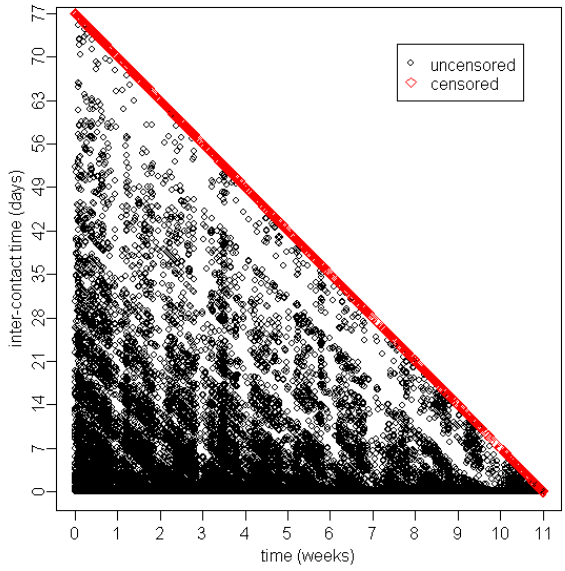
present detailed analysis and discussion in the followings.

3 Description of Opportunistic Network Traces

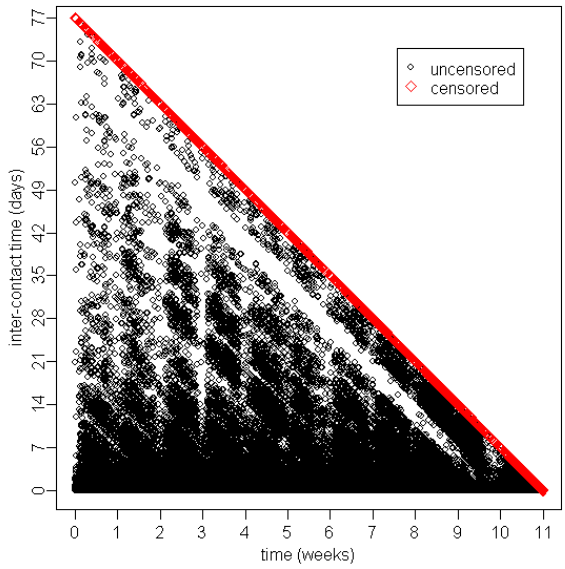
In the past few years, the research communities of DTN, MANET, and opportunistic networks have devoted significant resources and energy to collecting realistic network traces. Researchers have realized that realistic wireless and in-motion networks are usually much more dynamic and more unpredictable than the existing mobility models (e.g., the Pursue Mobility Model [15], the Reference Point Group Mobility Model [27]) used in traditional mobile and wireless network studies. From our target network scenarios, network traces can be classified into several groups, such as wildlife networks [11], people networks [1][5][6][7], pocket switched networks [3][9][10], and vehicular networks [2][4][8].

Since wildlife network traces usually have 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 pursue the study of opportunistic network scenarios based on human mobility. We select two publicly available network traces, namely UCSD [5] and Dartmouth [1] traces, due to their large number of participating nodes and sufficiently long measurement duration. Table 1 outlines the basic properties of the two network traces[1].

[1]In Dartmouth trace, there were a total of 13,888 devices in the network, but only 5,148 of them have contact experience with other devices.



(a) UCSD trace



(b) Dartmouth trace

Figure 1: Illustration of inter-contact time distribution of UCSD and Dartmouth traces.

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
Devices participating	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%

More specifically, the UCSD trace is a client-based trace that records the visibility of WiFi based access points (APs) with each participating portable device (e.g., PDAs and laptops) on UCSD campus. The network trace is about two and half months long, and there are 273 devices participated. Similar to [16][17][29], we make the assumption that a communication opportunity (i.e., network contact) is encountered between two participating devices (in ad hoc mode) if and only if both of them are associated to the same AP during some time period.

Similarly, the Dartmouth trace is an interface-based trace that records the APs that have been associated with a particular wireless interface during a three year (1177 days) period. However, we do not intend to use the full length trace in the following analysis since the overall computation overhead would become too costly. We will use only a subset of the trace, which is with the same period (77 days, from 09/22/02 to 12/08/02) as the UCSD

trace, for analysis purpose, and use the full trace to verify the correctness of our censorship removal algorithm that we will detail in the next section.

It should also be noted that, in Dartmouth trace, wireless interfaces can be used by different devices at different times, and each device may use multiple wireless interfaces. For simplicity, we assume each network interface represents a single mobile user in the network. Moreover, like in the UCSD scenario, a network contact is encountered when two mobile users are associated to the same access point. Note that, although the Dartmouth trace is a lengthier trace with a greater number of participating mobile nodes, the network connectivity is actually very poor in the Dartmouth scenario, since network contacts (for each source-destination pair) occur much more infrequently in the network (nearly one sixth of the UCSD scenario).

Similar to [16][17][29], it is the goal of this study to analyze the distribution of the *inter-contact time*, T_{i_c}, (i.e., the time period between two consecutive *contacts* of a given node pair) in that this property reflects the network connectivity of the network. Fig. 1 depicts the inter-contact time distribution of the two employed network traces, and each point on the figure represents one inter-contact time measurement that starts at the corresponding time point (horizontal axis). In Fig. 1, it is clear that the inter-contact time distribution is strongly skewed and upper-bound by a straight line (i.e., $T_{upper\_bound} = 11 - T_{cur}$, where T_{cur} is the starting day of the inter-contact time in the network trace and 11 is the trace length in weeks). Moreover, one can also find that the data points can be classified into two groups:

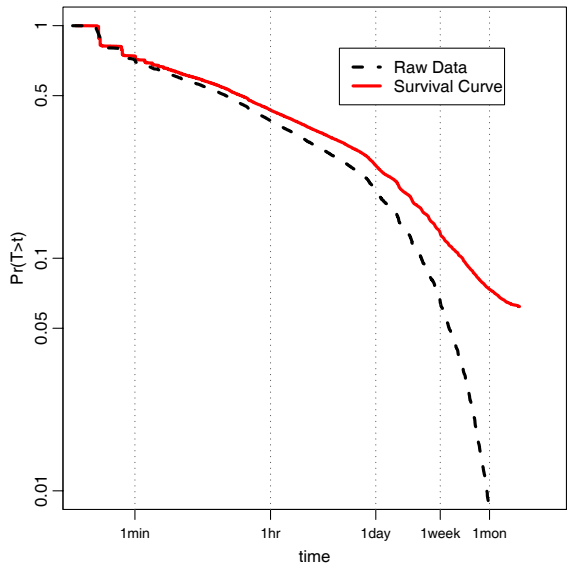
one is uncensored inter-contact time, and the other is censored inter-contact time[2]. More precisely, 7% of inter-contact time measurements are censored in UCSD trace, and 1.3% are censored in Dartmouth trace. In addition, all censored data lie on the upper bound straight line, whereas uncensored data are located in the lower region of the straight line. It turns out that the censorship leads to strongly skewed inter-contact time measurements, and it is necessary to *recover* those censored measurements in order to have more precise analysis for opportunistic networks.

4 Calibrating Censored Measurements

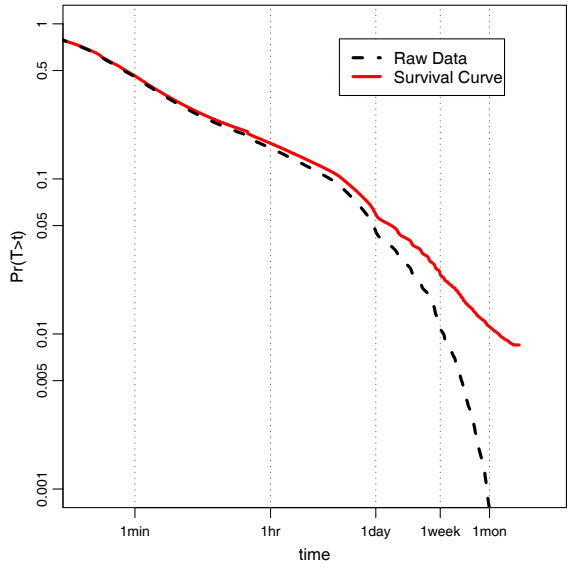
In this section, we present algorithms that can calibrate censored measurements in opportunistic traces. As identified previously, the inter-contact time measurement is a kind of survival data (i.e., time to death or event) [23] by nature, since an inter-contact time is likely to start when the measurement is going on, but stop after the end of the measurement. Analysis of survival data has been extensively studied in many disciplines, such as biostatistics, bioinformatics, life science, and etc., and it has been applied to the subject of network analysis for online gaming traffic recently [18]. However, survival analysis has not yet been applied to opportunistic network traces, even though censored data are prevalent and measurements are strongly skewed.

Targeting this issue, we present one survival analysis technique, called

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



(a) UCSD Trace



(b) Dartmouth Trace

Figure 2: Comparison of CCDF and survival curves (using K-M Estimator) of UCSD and Dartmouth traces in log-log scale.

Kaplan-Meier Estimator, in subsection 4.1 to estimate the survivorship of the employed network traces. We present the Censorship Removal Algorithm (CRA) in subsection 4.2. In subsection 4.3, we show that the proposed CRA technique can effectively calibrate skewed inter-contact time measurements.

4.1 Kaplan-Meier Estimator

The Kaplan-Meier Estimator (K-M Estimator, a.k.a. Product Limit Estimator) [30] has been proposed by Kaplan and Meier in 1958. The basic idea of K-M estimator is that, given survival data as an independent random variable, the censored measurements shall have the same likelihood of distribution as the uncensored ones as long as the number of uncensored measurements is sufficiently large. More specifically, we define a survival function (a.k.a. survivorship function or reliability function), $S(t)$, as the probability that an inter-contact time measurement from the given network trace is larger than t, i.e., $S(t) = \Pr[T_{i\_c} > t]$.

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 $T_{i\_c} = t_i$, the K-M Estimator is a nonparametric maximum likelihood estimate of $S(t)$ as defined by Eq. 1.

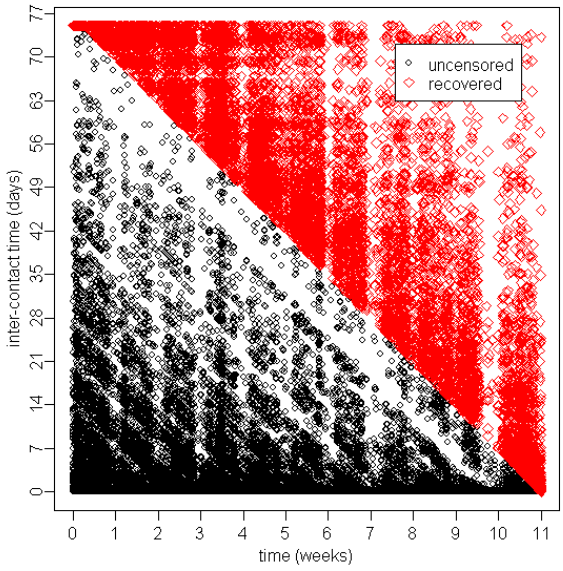
$$\widehat{S}(t) = \prod_{t_i \leq t} \Pr\left[t > t_i \,|t \geq t_i\right]$$
$$= \begin{cases} 1 \; ; \, t_1 > t \\ \prod_{t_i \leq t \leq t_N} \left[\dfrac{n_i - d_i}{n_i}\right] \; ; \, t_1 \leq t \end{cases} \quad (1)$$

Note that, since the calculation of K-M Estimator is based on the likelihood of uncensored data, the survivorship does not exist when $t > t_N$, that is the maximum inter-contact time measurement in the trace. We illustrate the complementary cumulative distribution function (CCDF) of uncensored T_{i_c} measurements, as well as the survival curves, using UCSD and Dartmouth traces in log-log scale in Fig. 2.

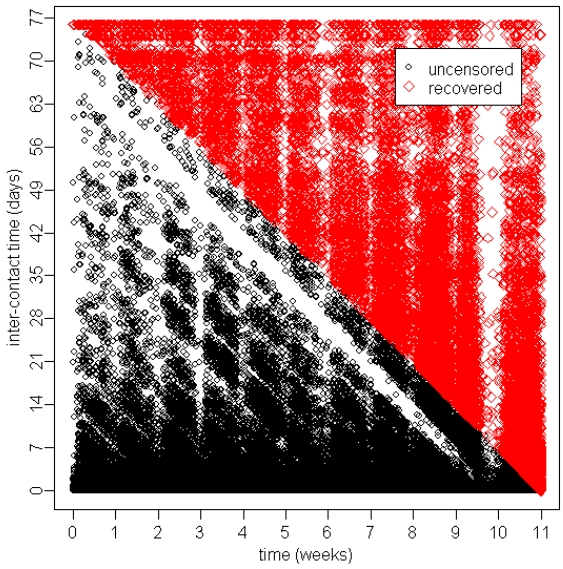
4.2 Censorship Removal Algorithm (CRA)

In Fig. 2, it is clear that the survival curve consistently has higher population probability than the CCDF one. The reason is because the CCDF curve is calculated simply using complete (uncensored) T_{i_c} measurements, whereas the survival curve takes the censorship into account and presents the more approximate distribution of the T_{i_c} distribution of realistic opportunistic networks. Intuitively, the area between the two curves represents the amount of population that is recovered (calibrated) by survival analysis, and that is indeed unnegligible.

It turns out that a censorship removal scheme that can recover censored



(a) UCSD trace



(b) Dartmouth trace

Figure 3: Illustration of inter-contact time distribution of UCSD and Dartmouth traces after calibration.

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

1: **for** $i = 1$ to $N - 1$ **do**
2: Randomly select $\frac{\widehat{S}(t_i)-\widehat{S}(t_{i-1})}{\widehat{S}(t_i)}$ 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

measurements is still highly desired for further analysis of inter-contact time measurements in opportunistic networks. Based on the K-M Estimator results, we propose a censorship removal algorithm (CRA) to calibrate the censorship based on $\widehat{S}(t)$ estimates. More specifically, suppose C_i/D_i denotes the set of censored/uncensored inter-contact time measurements with $T_{i\_c} = t_i$, the censorship removal algorithm iteratively moves a portion of censored data (based on the probability, $\frac{\widehat{S}(t_i)-\widehat{S}(t_{i-1})}{\widehat{S}(t_i)}$) from C_i to D_i and moves the remaining entities of C_i to C_{i+1} afterward. Alg. 1 shows the algorithm.

More precisely, in each iteration, each entity of C_i is going to be moved to D_i with probability $\frac{\widehat{S}(t_i)-\widehat{S}(t_{i-1})}{\widehat{S}(t_i)}$ or moved to C_{i+1} otherwise. For simplicity, we assume the decision process is uniformly distributed. Note that, the probability, $\frac{\widehat{S}(t_i)-\widehat{S}(t_{i-1})}{\widehat{S}(t_i)}$, can be interpreted as the *death ratio* between t_i and t_{i+1} (i.e., $\widehat{S}(t_i) - \widehat{S}(t_{i-1})$) with respect to (normalized by) the survivorship at t_i, i.e., $\widehat{S}(t_i)$. Fig. 3 shows the results of inter-contact time distribution after censorship removal for both UCSD and Dartmouth traces.

4.3 Evaluation

In this subsection, we present evaluation showing the correctness of the proposed CRA technique. The shortened Dartmouth trace (77 days long, from 09/22/02 to 12/08/02) is employed as the raw network trace, and the full trace (i.e., 1177 days long) is used to provide complete T_{i_c} information that are however censored in the shortened one. As we have discovered previously, there are about 1.3% events (i.e., T_{i_c} measurements) censored in the shortened network trace, and 80.4% of them become uncensored when the 1177-day long trace is employed (i.e., the T_{i_c} measurement ends after 12/08/02 but before the end of network measurements). Fig. 4 compares the CCDF of the measured T_{i_c} (using the shortened trace), recovered T_{i_c}, and real T_{i_c} (using the 1177-day long trace) of Dartmouth trace. The results clearly show that, after applying CRA, the recovered T_{i_c} has nearly identical distribution as the real one. This clearly shows that the proposed CRA algorithm can correctly calibrate censorship in time-limited network traces.

5 Analysis of Self-Similarities Using Opportunistic Network Traces

In this section, we perform analysis of self-similarities on the opportunistic people network traces that have been calibrated using the proposed CRA technique as presented. Similar to previous studies [17][16], we focus on

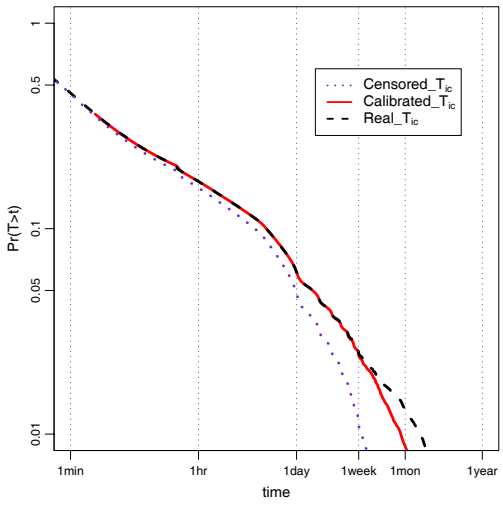


Figure 4: Comparison of measured T_{i_c} distribution, calibrated T_{i_c} distribution, and real T_{i_c} distribution (referring to the full version trace) of Dartmouth trace.

inter-contact time measurements of the network traces in that this property can be regarded as an indicator of network connectivity of an opportunistic network. We firstly investigate the power-law property that shows heavy tails in the distribution in subsection 5.1, and recap the definition of self-similarity in subsection 5.2. We present analysis of self-similarity in subsection 5.3, and discuss the analysis results in subsection 5.4.

5.1 Heavy-Tailed Distribution

As we have mentioned previously, researchers have found that the inter-contact time distribution of an opportunistic network is power-law distributed,

and thus heavy-tailed [17][16]. In this subsection, we first give an overview of the heavy-tailed distribution, and then we show that both UCSD and Dartmouth traces are 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 [21].

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

The simplest way to tell 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 thus be approximated by calculating the slope of the curve as shown in Eq. 3.

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

Using 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 UCSD and Dartmouth traces are heavy-tailed, and the conclusion confirms the results of previous studies [17][16].

5.2 Self-Similarity Definition

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

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

where the equality is in the sense of finite-dimensional distributions and H is an important effect to describe self-similarity called *hurst*. Note that a process satisfying 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 thus be obtained by dividing the original series X into non-overlapping blocks of size m and averaging each block as shown in Eq. 5, where k indexes the block number.

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

If X is the increment process of Z, i.e., $X(i) = Z(i+1) - Z(i)$, then, for all integers of m, one can obtained

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

Based on Eq. 6, the self-similarity is thus defined 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 sequences $X(i), i \geq 1$, is said to be asymptotically self-similar if Eq. 6 holds as $m \rightarrow \infty$.

3. A covariance-stationary sequence $X(i), i \geq 1$, 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 $m \rightarrow \infty$.

The degree of self-similarity of a series is expressed using only a single parameter called hurst parameter, H, that expresses the speed of decay of the autocorrelation function of the series. If the 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 self-similar process is that the degree of of self-similarity is expressed by the extent of hurst parameter, H. In other words, the statistical properties of a self-similar process shall not change for different aggregation levels. In this subsection, we apply four techniques (namely variance-time plot, rescaled adjusted range plot, periodogram plot, and Whittle estimator) [12][25][13] to investigate self-similarities within our network traces. We present the analysis in the followings.

5.3.1 Variance-Time Plot

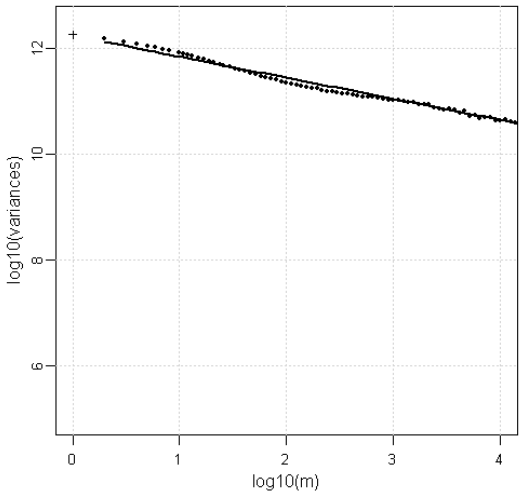
The variance-time plot tests the property of the slowly decaying variance that exists in self-similar series. The variance of the process $X^{(m)}$ is plotted against the aggregated level m on log-log plot. The m-aggregated process $X^{(m)} = (X_1^{(m)}, X_2^{(m)}, X_3^{(m)}, ...)$ is defined as Eq.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

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

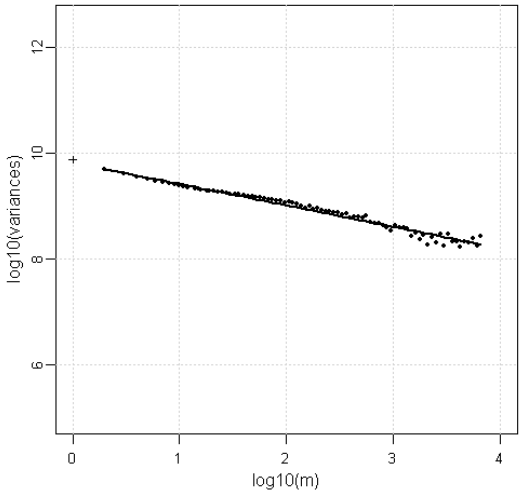
The variance of the process can thus be obtained by:

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

Fig. 5 depicts the variance-time plot with various aggregation levels of UCSD and Dartmouth traces. From the figures, it is observed that the aggregated variances of the inter-contact time measurements in our network traces are nearly linear and could be fitted by a simple least squares line with the 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 slop β by $H = 1 - \beta/2$. For instance, in Fig. 5-a, the slope is estimated by regression as -0.4, and the hurst parameter, H, is therefore estimated to be



(a) UCSD Trace (H=0.801)



(b) Dartmouth Trace (H=0.7973)

Figure 5: Graphical Analysis of Aggregated Variance Method.

0.8; whereas, in Fig. 5-b, the slope is about -0.405 and the hurst estimate is about 0.7973.

5.3.2 Rescaled Adjusted Range Plot

One important property of self-similarity is that the dataset shall keep the same statistical properties no matter how it is divided into several sub-datasets [25]. The second analysis technique, called R/S method, thus sequentially divides the dataset in dichotomy to calculate the rescaled adjusted range for each sub-dataset and then takes the average of all 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.

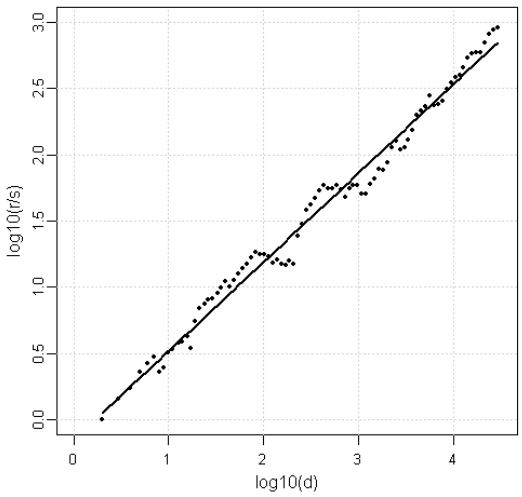
For instance, given a dataset : $X_1, X_2, X_3, ...X_n$ with sample mean $\mu = E[X_i]$, an adjusted partial sum is defined as

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

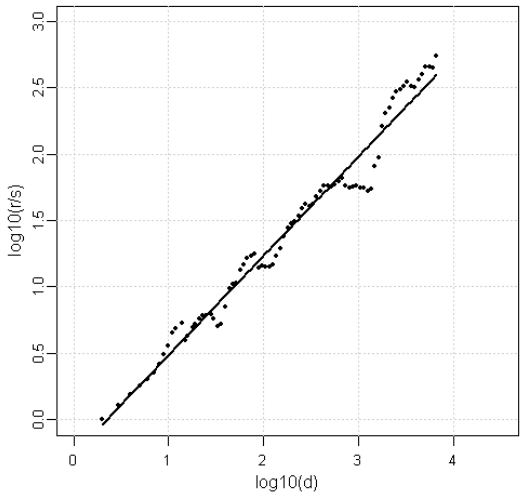
where $k = 1, 2, 3, ..., n$ and $\overline{X}(n)$ is the arithmetic mean of the first n observations. The range $R(n)$ is also defined by:

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

Suppose $S(n)$ denotes the standard deviation of sample size n, the R/S value of the dataset is thus defined by Eq. 11.



(a) UCSD Trace (H=0.7472)



(b) Dartmouth Trace (H=0.7973)

Figure 6: Graphical Analysis of R/S Method.

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

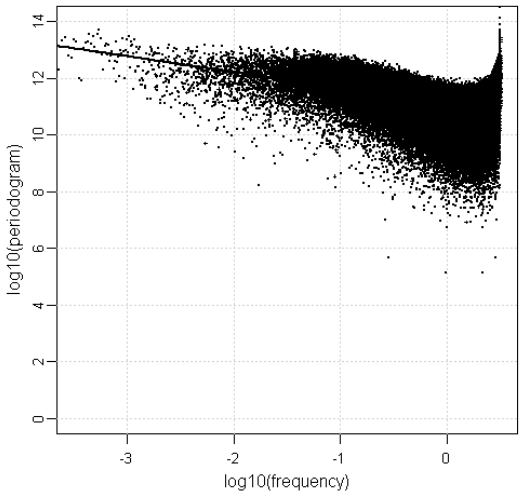
Fig. 6 shows the R/S plot (i.e., R/S values against n on a log-log plot) of the employed network traces, and the hurst parameter, H, is thus estimated by the regression slope. Specifically, the H estimate is 0.7472 in UCSD trace and 0.7493 in Dartmouth trace that indicates the inter-contact time measurements of both network traces are self-similar.

5.3.3 Periodogram Plot

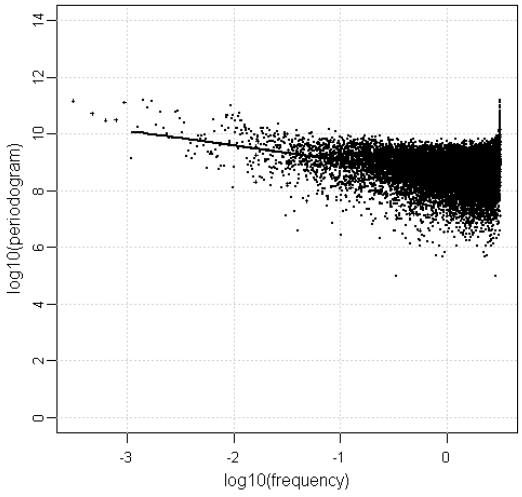
The periodogram is defined as Eq. 12, where ν is a frequency, N is the series length, X is the series, and $I(\nu)$ is the estimator on the spectrum field [13].

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

A Periodogram Plot can be therefore obtained by collecting multiple periodograms of various frequency values. The plot can be fitted using a straight line in the log-log scale, and the slope, β, of the fitting line can be approximated by $1-2H$. Note that, in practice, people usually use the lowest 10% of the frequencies [38] to make the periodogram plot, since the power law behavior only holds for frequencies close to zero. Fig. 7 illustrates the periodogram plots of UCSD and Dartmouth traces. The slope of the fitting line is aound -0.56 in UCSD trace and -0.53 in Dartmouth trace, and, therefore, the hurst estimate is about 0.78 in UCSD trace and 0.76 in Dartmouth trace



(a) UCSD Trace (H=0.7824)



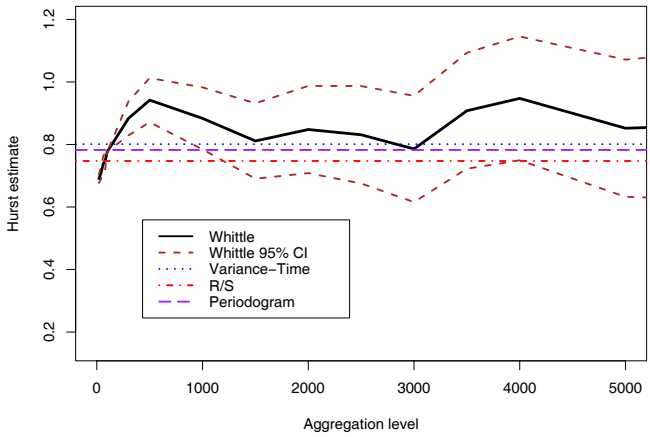
(b) Dartmouth Trace (H=0.7655)

Figure 7: Graphical Analysis of Periodogram Method.

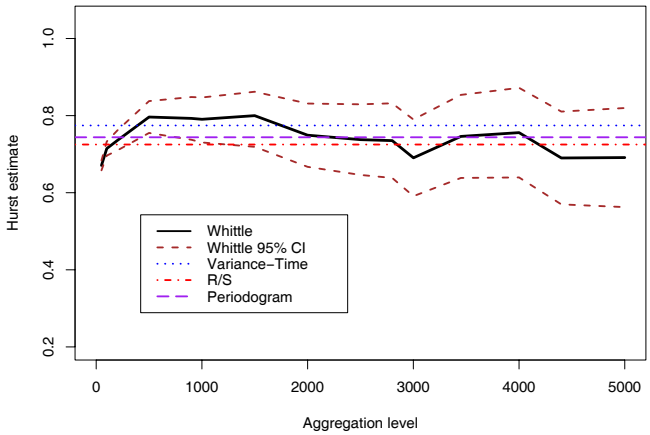
that again confirm the inter-contact time measurements are self-similar in both network traces.

5.3.4 Whittle Estimator

Whittle estimator is usually regarded as the most robust indicator for self-similarity analysis in that 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) with $0 < d < 1/2$ [21]. The main difference between these two methods is that ARIMA assumes the existence of short-range dependency while FGN does not. Since we in this study focus on the long-range dependency of the employed network traces, we apply Whittle estimator with FGN by aggregating the datasets into different levels. As shown in Fig. 8, the Whittle estimator is stabilized to about 0.8 for UCSD network trace while the comparison results of the three graphical methods are all within 95% confidence interval (when the aggregation level is greater than 1000). Moreover, the results also show that the Whittle estimator is stabilized to about 0.75 for Dartmouth network trace, and the comparison with previous three graphical methods are all within 95% confidence interval when the aggregation level is greater than 1000. We conclude here again the inter-contact time measurements of UCSD and Dartmouth traces are both self-similar.



(a) UCSD Trace



(b) Dartmouth Trace

Figure 8: Graphical Analysis of Whittle Estimator.

5.4 Discussion

Recall that, in Fig. 1 (and Fig. 3), it is apparent that there do exist some regular patterns in the distribution of inter-contact time measurements for both UCSD and Dartmouth traces. More specifically, from the figures, one can easily observe that "*periodical gaps*" appear about once a week, and each gap is about one-day long in the figures. The phenomenon can be explained that people in the campus tend to have *contacts* with others only during weekdays (i.e., students always escape from the campus in weekends, and thus there are very few wireless network associations with campus access points present in both network traces). Moreover, similar phenomena are also observed when we zoom in/out the inter-contact time distribution, i.e., periodical gaps are also present in daily, quarterly, and yearly basis, that represents sparse network connectivity during off office hours, quarter breaks[3], and summer vacations respectively.

In addition, Fig. 1 also shows that most uncensored measurements are clustered at the bottom area of the figure, i.e., the inter-contact time of these clustered measurements are small. The results indicate that once a network contact is encountered for a given pair of nodes, they are more likely to meet each other again in the very near future. For instance, in the employed campus scenarios, most mobile devices are carried by students, and they tend to team up and move as clusters in the campus-life (e.g., attending classes,

[3]Both of UCSD and Dartmouth College are on a quarter-based system, rather than a semester system.

dining in campus restaurants, and studying in campus libraries). Based on the mathematical analysis and our explanation, we conclude that the process of opportunistic people networks is self-similar.

6 Conclusion

In this study, we investigate fundamental properties of opportunistic people networks. Using public network traces from UCSD and Dartmouth college, we identify the censorship issue in network traces that usually leads to strongly skewed distribution of the measurements. Based on this knowledge, we then apply the Kaplan-Meier Estimator to calculate the survivorship of network measurements, which is used in designing our proposed censorship removal algorithm (CRA) that is used to recover censored data. We show that, after applying CRA, the recovered network trace has nearly identical inter-contact time distribution as the real one. Additionally, we perform a rich set of analysis illustrating that UCSD and Dartmouth network traces shows strong self-similarity, and we pointed out the importance of these newly revealed characteristics to the future of opportunistic people network research. The results of this study is indeed influential and should be taken into consideration in the design, evaluation, and deployment of future opportunistic network applications.

References

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

[2] The disruption tolerant networking project at umass. http://prisms.cs.umass.edu/diesel/.

[3] Haggle project. http://www.cambridge.intel-research.net/haggle/.

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

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

[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.usc.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, Swedish Institute of Computer Science, 1999.

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

[14] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networking. In *IEEE Infocom*, 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. Diot, R. Gass, and J. Scott. Pocket switched networks: Real-world mobility and its consequences for opportunistic forwarding. 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. In *IEEE Infocom*, 2006.

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

[19] L.-J. Chen, C.-H. Yu, T. Sun, Y.-C. Chen, and H.-H. Chu. A hybrid routing approach for opportunistic networks. In *ACM SIGCOMM Workshop on Challenged Networks*, 2006.

[20] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. In *ACM SIGMETRICS*, 1996.

[21] 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, 1997.

[22] J.-H. Cui, J. Kong, M. Gerla, and S. Zhou. Challenges: Building scalable mobile underwater wireless sensor networks for aquatic applications. *IEEE Network, Special Issue on Wireless Sensor Networking*, May 2006.

[23] R. C. Elandt-Johnson and N. L. Johnson. *Survival Models and Data Analysis*. Wiley, September 1980.

[24] M. Garrett and W. Willinger. Analysis, modeling and generation of self-similar vbr video traffic. In *ACM SIGCOMM*, 1994.

[25] M. Gospodinov and E. Gospodinova. The graphical methods for estimating hurst parameter of self-similar network traffic. In *CompSys-Tech'2005*, 2005. International Conference on Computer Systems and Technologies-CompSysTech'2005.

[26] K. A. Harras, K. C. Almeroth, and E. M. Belding-Royer. Delay tolerant mobile networks (dtmns): Controlled flooding in sparse mobile networks. In *IFIP Networking*, 2005.

[27] 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 Wireless and Mobile Systems*, 1999.

[28] J.-H. Huang, S. Amjad, and S. Mishra. Cenwits: A sensor based loosely coupled search and rescue system using witnesses. In *ACM SenSys*, 2005.

[29] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *ACM SIGCOMM Workshop on DTN*, 2005.

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

[31] J. LeBrun, C.-N. Chuah, and D. Ghosal. Knowledge based opportunistic forwarding in vehicular wireless ad hoc networks. In *IEEE VTC Spring*, 2005.

[32] J. Leguay, T. Friedman, and V. Conan. Dtn routing in a mobility pattern space. In *ACM SIGCOMM Workshop on Delay Tolerant Networks*, 2005.

[33] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. In *ACM SIGCOMM*, 1993.

[34] Y. Liao, K. Tan, Z. Zhang, and L. Gao. Estimation based erasure-coding routing in delay tolerant networks. In *International Wireless Communications and Mobile Computing Conference*, 2006.

[35] A. Lindgren, A. Doria, and O. Schelen. Probabilistic routing in intermittently connected networks. *ACM SIGMOBILE Mobile Computing and Communications Review*, 7(3):19–20, July 2003.

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

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

[38] M. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: an empirical study, 1995. preprint.

[39] A. Vahdat and D. Becker. Epidemic routing for partially-connected ad hoc networks. Technical Report CS-2000-06, Duke University, 2000.

[40] R. Y. Wang, S. Sobti, N. Garg, E. Ziskind, J. Lai, and A. Krishnamurthy. Turning the postal system into a generic digital communication mechanism. In *ACM SIGCOMM*, 2004.

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

[42] Y. Wang and H. Wu. Dft-msn: The delay fault tolerant mobile sensor network for pervasive information gathering. In *IEEE Infocom*, 2006.

[43] J. Widmer and J.-Y. L. Boudec. Network coding for efficient communication in extreme networks. In *ACM SIGCOMM Workshop on Delay Tolerant Networks*, 2005.