Finding Self-Similarities in Opportunistic People Networks

Ling-Jyh Chen[1], Yung-Chih Chen[1], Tony Sun[2], Paruvelli Sreedevi[1], Kuan-Ta Chen[1], Chen-Hung Yu[3], and Hao-hua Chu[3]

[1]Institute of Information Science, Academia Sinica
[2]Department of Computer Science, University of California at Los Angeles
[3]Department of Computer Science and Information Engineering, National Taiwan University

Abstract— **Opportunistic network is a type of Delay Tolerant Networks (DTN) where network communication opportunities appear opportunistic. In this study, we investigate opportunistic network scenarios based on public network traces, and 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 show 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.**

I. Introduction

Opportunistic network is a type of challenged networks, where network contacts are intermittent, an end-to-end path between the source and the destination may have never existed, disconnection/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 to networks in this category. With numerous emerging opportunistic networking applications, such as wireless sensor networks (WSN) [4][22], underwater sensor networks (UWSN) [12], transportation networks [3][7], pocket switched networks(PSN) [8][16], and people networks [20][21], it remains desirable to develop effective schemes that can better accommodate the characteristics of opportunistic networks.

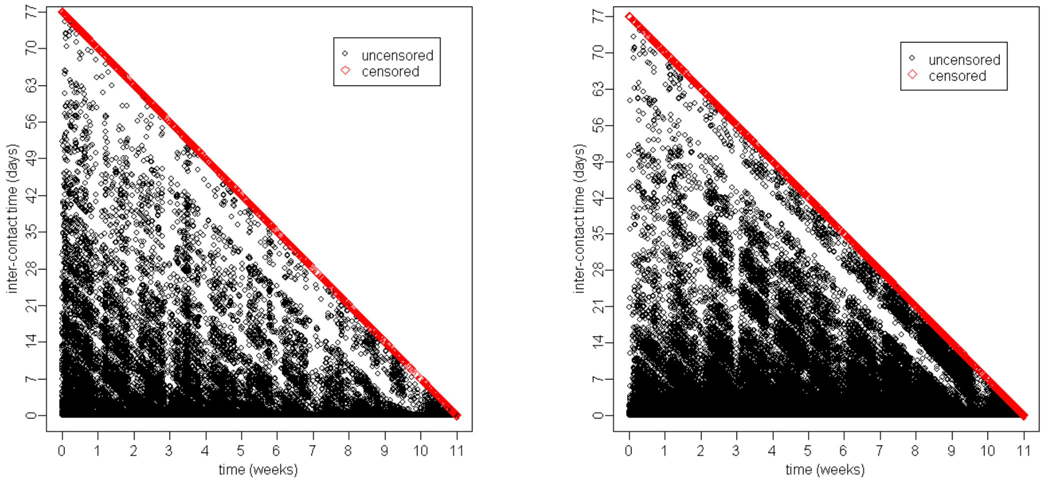
Knowing fundamental properties of opportunistic networks is the key for the design of effective routing protocols and/or applications. Among all, knowledge of *inter-contact time distribution* is particularly important, since this distribution provides a good description of network connectivity. By *inter-contact time*, we mean that the time duration between two contiguous network contacts (between a particular node pair). The more inter-contact time events in the network trace, the more reconnection/disconnection events have occurred during the network measurement period.

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 [2] 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 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 II, we summarize related work in this area. In section III, we briefly describe the basic properties of the opportunistic network traces examined. Section IV presents our survival analysis and the proposed censorship removal algorithm for the employed network traces. Section V performs self-similarity analysis on the recovered network traces. Finally, section VI concludes the paper.

II. Related Work

Statistical analysis of opportunistic network traces has been performed [8][16], 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 [10][14][18][19], 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.



(a) UCSD trace

(b) Dartmouth trace

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

TABLE I

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%

III. Description of Opportunistic Network Traces

In this paper, we select two publicly available network traces, namely UCSD [2] and Dartmouth [1] traces, due to their large number of participating nodes and sufficiently long measurement duration. Table I outlines the basic properties of the two network traces[1].

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 [8][16], 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 due to the costly overall computation overhead. 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.

Similar to [8][16], the goal of this study is to analyze the distribution of the *inter-contact time*, T_{i_c}, 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).

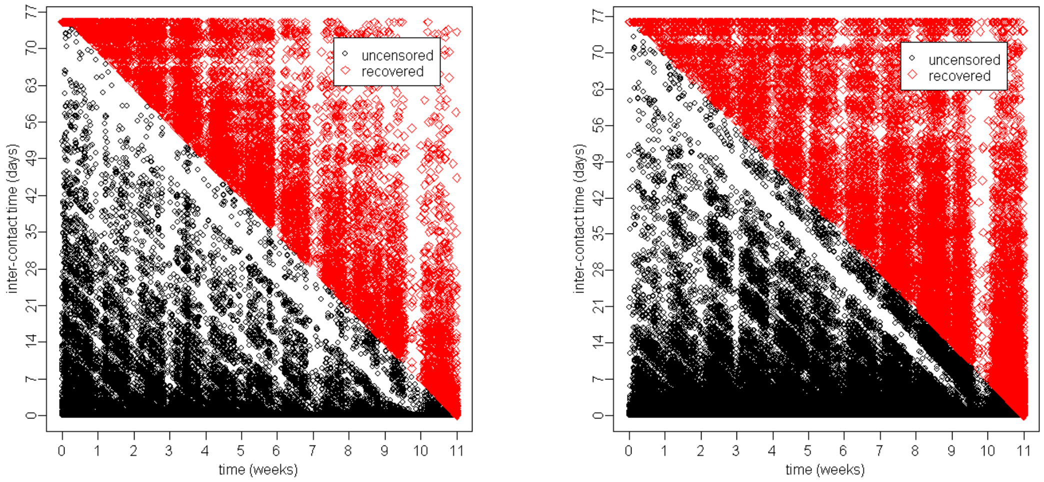
In Fig. 1, it is clear that the inter-contact time distribution is strongly skewed and upper-bound by a straight line (i.e., $T_{upper\_bound} = 11 - T_{cur}$, where T_{cur} is the starting day of the inter-contact time in the network trace and 11 is the trace length in weeks). Moreover, one can also 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.

IV. Calibrating Censored Measurements

As identified previously, the inter-contact time measurement is a kind of survival data (i.e., time to death or event) [13] 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

[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.

[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

Fig. 2. 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.

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

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 [9]. 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 Kaplan-Meier Estimator, in subsection IV-A to estimate the survivorship of the employed network traces. We present the Censorship Removal Algorithm (CRA) in subsection IV-B, and the evaluation of CRA in IV-C.

A. Kaplan-Meier Estimator

The Kaplan-Meier Estimator (K-M Estimator, a.k.a. Product Limit Estimator) [17] 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.

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

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.

B. Censorship Removal Algorithm (CRA)

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

For simplicity, we assume the decision process is uniformly distributed. Fig. 2 shows the results of inter-contact time distribution after censorship removal for both UCSD and Dartmouth traces.

C. Evaluation

Here we present evaluation showing the correctness of the proposed CRA technique. The shortened Dartmouth trace (77

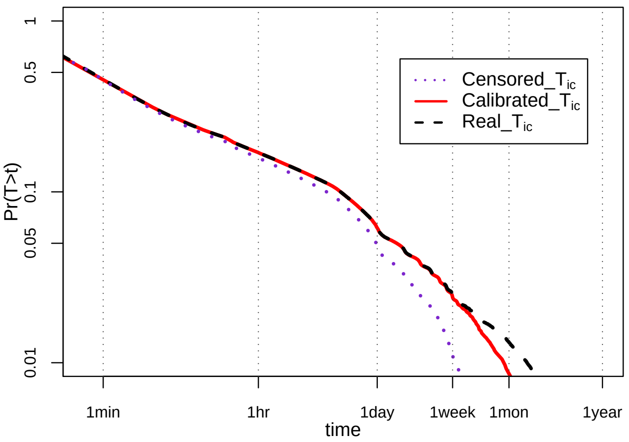


Fig. 3. Comparison of measured T_{i_c} distribution, calibrated T_{i_c} distribution, and real T_{i_c} distribution (the full version trace) of Dartmouth trace.

days long) is employed as the raw network trace, and the full trace (1177 days long) is used to provide complete T_{i_c} information censored in the shortened one. As we have discovered previously, about 1.3% events (i.e., T_{i_c} measurements) are censored in the shortened network trace, and 80.4% of them become uncensored when the 1177-day trace is employed (i.e., the T_{i_c} measurement ends after the 77th day but before the end of network measurements). Fig. 3 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.

V. Analysis of Self-Similarities Using Opportunistic Network Traces

In this section, we perform analysis of self-similarities on inter-contact time measurements of opportunistic people network traces that have been calibrated using the proposed CRA technique as presented. We firstly investigate the power-law property that shows heavy tails in the distribution in subsection V-A, recap the definition of self-similarity in subsection V-B and show the analysis of self-similarity in subsection V-C.

A. Heavy-Tailed Distribution

As mentioned previously, the inter-contact time distribution of opportunistic networks has been found power-law distributed, and thus heavy-tailed [8]. In this subsection, we first give an overview of the heavy-tailed distribution and then 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 $x \rightarrow \infty$, where c is a positive constant and α is the power-law exponent [11].

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

We find that the $alpha$ value for the tail (slope of the curve in log-log scale) is 0.26 for UCSD trace and 0.47 for Dartmouth trace. Therefore, we conclude that both UCSD and Dartmouth traces are heavy-tailed, which confirms the results of previous studies [8].

B. 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:

$$Z(t) = a^{-H} Z(at); \quad \forall t \geq 0, \forall a > 0, 0 < H < 1 \tag{3}$$

where the equality is in the sense of finite-dimensional distributions. The H is called *hurst* that expresses the the degree of self-similarity of the series. If the series is self-similar, $1/2 < H < 1$. Moreover, as H approaches 1, the degree of self-similarity increases. Note that a process satisfying Eq. 3 can never be stationary but Z is typically assumed to have stationary increments.

C. Graphical Methods and Statistical Analysis

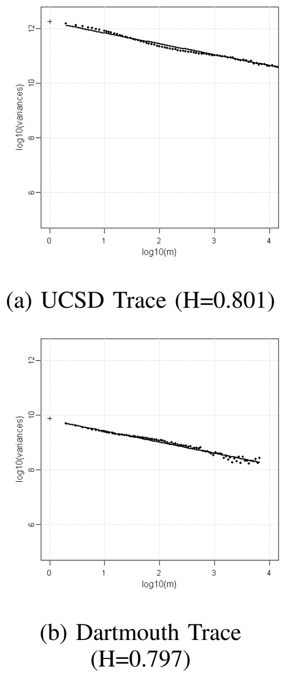
In this subsection, we apply four techniques (namely variance-time plot, R/S plot, periodogram plot, and Whittle estimator) [5][6][15] to investigate self-similarities within our network traces. We present the analysis in the followings.

1) Variance-Time Plot: The variance-time plot tests the property of the slowly decaying variance that exists in self-similar series. In Fig. 4-a, the slope of is estimated by regression as -0.4, and the hurst parameter, H, is estimated to be 0.8; whereas, in Fig. 4-b, the slope is about -0.405 and the hurst estimate is about 0.797.

2) Rescaled Adjusted Range Plot: The R/S method 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 [15]. Fig. 5 shows the R/S plot of the employed network traces, and the hurst parameter, H, is thus estimated by the regression slope. Specifically, the H estimate is 0.747 in UCSD trace and 0.749 in Dartmouth trace that indicates the inter-contact time measurements of both network traces are self-similar.

3) Periodogram Plot: A Periodogram Plot can be obtained by collecting multiple periodograms of various frequency values [6]. Fig. 6 illustrates the periodogram plots of UCSD and Dartmouth traces. The hurst estimate is about 0.78 in UCSD trace and 0.76 in Dartmouth trace that again confirm the inter-contact time measurements are self-similar in both traces.

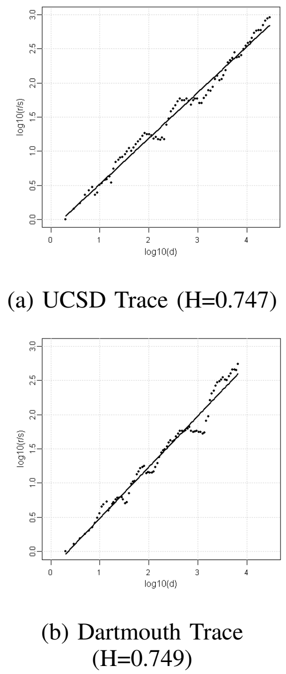
4) Whittle Estimator: Whittle estimator is usually regarded as the most robust indicator for self-similarity analysis in that it provides a confidence interval [11]. As shown in Fig. 7, the Whittle estimator is stabilized to about 0.8 for UCSD trace and 0.75 for Dartmouth trace while the comparison results of the 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 (H=0.801)

(b) Dartmouth Trace (H=0.797)

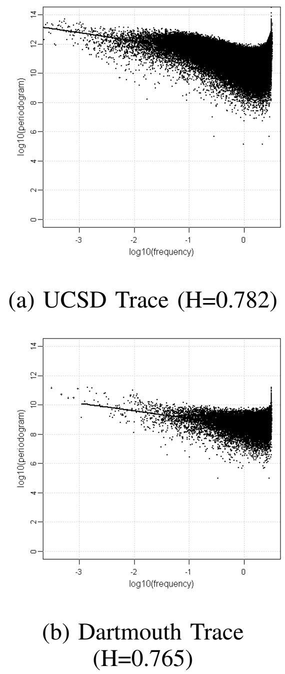
Fig. 4. Variance-Time Method.



(a) UCSD Trace (H=0.747)

(b) Dartmouth Trace (H=0.749)

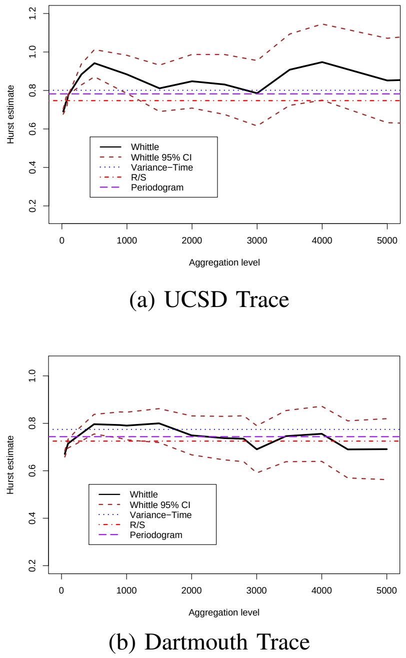
Fig. 5. R/S Method.



(a) UCSD Trace (H=0.782)

(b) Dartmouth Trace (H=0.765)

Fig. 6. Periodogram Method.



(a) UCSD Trace

(b) Dartmouth Trace

Fig. 7. Whittle Estimator.

VI. 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) 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] Ucsd wireless topology discovery project. http://sysnet.ucsd.edu/wtd/.

[3] Umass dome project. http://prisms.cs.umass.edu/diesel/.

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

[5] H. Abrahamsson. Traffic measurement and analysis. Technical report, Swedish Institute of Computer Science, 1999.

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

[7] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networking. In *IEEE Infocom*, 2006.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] M. Gospodinov and E. Gospodinova. The graphical methods for estimating hurst parameter of self-similar network traffic. In *ICCST*, 2005.

[16] 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.

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

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

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

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

[21] 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.

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