TRANSACTIONS ON EMERGING TELECOMMUNICATIONS TECHNOLOGIES
Trans. Emerging Tel. Tech. 2012; **00**:1–14
DOI: 10.1002/ett

RESEARCH ARTICLE

Interaction Predictability of Opportunistic Networks in Academic Environments

Radu-Ioan Ciobanu[1], Radu-Corneliu Marin[1] and Ciprian Dobre[1]*

University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers, 313 Splaiul Independentei, Sector 6, 060042, Bucharest, Romania

ABSTRACT

The ubiquitousness of mobile devices in recent years has led to an ever-growing interest in mobile networks. One such example is represented by opportunistic networks, which are composed of mobile devices that interact in a store-carry-and-forward fashion. A mobile node stores data and carries it around; when it encounters another node, it may decide to forward the data if the encountered node is the destination or has a better chance of bringing the data closer to the destination. If the encountered node is likely to have a contact with the destination sooner than the data carrier, the data should be forwarded, in order for it to reach its destination with a lower delivery latency. Since nodes in opportunistic networks are carried by humans, their mobility plays a very important role in the behavior of the network. Therefore, predicting a node's interactions and mobility patterns is paramount to the implementation of efficient routing algorithms. Thus, in this paper we present a mobile interaction trace collected at the University POLITEHNICA of Bucharest in the spring of 2012 and analyze it in terms of the predictability of encounters and contact durations. We show that there is a regular pattern in the contact history of a node and we prove that, by modeling the time series as a Poisson distribution (with the number of contacts being the Poisson events in a fixed one-hour interval), we can efficiently predict the number of contacts per time unit in the future. These assumptions are demonstrated both on the trace presented here, as well as on a trace recorded in a different environment, showing that predictability doesn't happen only in strict and controlled situations. Moreover, we analyze the predictability of an opportunistic node's sightings of wireless access points, which can help us discover a user's patterns and behavior.

*Correspondence

University POLITEHNICA of Bucharest, Faculty of Automatic Control and Computers, 313 Splaiul Independentei, Sector 6, 060042, Bucharest, Romania. Email: ciprian.dobre@cs.pub.ro

1. INTRODUCTION

In recent years, mobile devices have become more and more ubiquitous. Thus, ad-hoc wireless networks formed over these devices have been analyzed thoroughly by researchers. One type of such a network is an opportunistic network (ON), which consists of human-carried mobile devices that communicate with each other opportunistically. In ONs, disconnections and highly variable delays caused by human mobility are the norm. The solution consists of dynamically building routes, as each node acts according to a store-carry-and-forward paradigm, where contacts between nodes are viewed as an opportunity to move data closer to the destination. Such networks are therefore formed between nodes spread across the environment, without any knowledge of a network topology. The routes between nodes are dynamically created, and nodes can be opportunistically used as a next hop for bringing each message closer to the destination. Nodes may store a message, carry it around,

Prepared using ettauth.cls [Version: 2012/06/19 v2.10]

and forward it when they encounter the destination or a node that is more likely to reach the destination.

The challenge in opportunistic networks is knowing when and to whom should a node pass a message in order to obtain the best possible hit rate and latency. Thus, predicting the future encounters of a device is paramount to implementing a good opportunistic routing algorithm. In this paper, we propose a way to predict the future behavior of a node in the opportunistic network by analyzing its past encounters and approximating the time series as a Poisson distribution. We use a mobile trace gathered at the University POLITEHNICA of Bucharest in 2012 to prove our suppositions. Aside from the fact that it is a relatively small enclosed space with a lot of participants, the other great advantage of an opportunistic network implemented over an academic environment is that the contacts are inherently regular: students attend the same classes with the same professors every week. We thus show that the contact history of a node in terms of number of encounters may be approximated as a Poisson distribution. We then compute the Poisson probabilities for the final two weeks of the experiment and prove that they hold in practice.

Although ONs are formed mainly of mobile nodes, the participants in such a network may encounter static access points (APs) that can help relay their data. Furthermore, these APs can also act as a way to pinpoint the location of an opportunistic node and thus help discover its patterns and behavior. Therefore, in this paper we also analyze a node's interactions with wireless APs and show that they exhibit a certain degree of predictability. By using this information correctly, we can obtain a node's route.

Preliminary versions of our work were previously published in [1] and [2]. This paper adds more extensive work, presenting our collected trace in detail and describing a methodology for analyzing wireless access point interaction data in terms of predictability.

The rest of the paper is structured as follows. Section 2 presents recent work in the area of opportunistic networking and event detection and predictability. Section 3 offers information regarding the tracing experiment at the University POLITEHNICA of Bucharest. Next, Section 4 performs an analysis of the trace in terms of the predictability of node encounters, contact durations and wireless AP sightings. Section 5 shows that approximating the trace as a Poisson distribution can lead to good predictability results. Finally, Section 6 presents conclusions and future work.

2. RELATED WORK

A thorough review of opportunistic networking is presented in [3]. The analysis, developed in the context of the EU Haggle project, highlights the properties of main networking functions, including message forwarding, security, data dissemination and mobility models. The authors also propose various solutions for communication in opportunistic networks and introduce HCMM, a mobility model that merges the spatial and social dimensions. Moreover, several well-known opportunistic forwarding algorithms are also presented, such as BUBBLE Rap [4], PROPICMAN [5] and HIBOp [6]. Other dissemination techniques for opportunistic routing include Socio-Aware Overlay [7], Wireless Ad Hoc Podcasting [8] or ContentPlace [9]. Of these algorithms, only ContentPlace performs a prediction of the future encounters of a node, based on the history of previous contacts. A taxonomy for data dissemination algorithms is proposed in [10].

Jain et al. [11] propose a framework for evaluating routing algorithms for delay-tolerant networks and analyze the performance of several such algorithms in terms of the amount of knowledge about the network that they require. The authors state that the algorithms that use the least knowledge tend to perform poorly, but that efficient algorithms can be constructed using less than the complete global knowledge. A set of abstract knowledge oracles are proposed for the framework, that are able to answer questions about the environment (i.e. the network). There are four types of oracles: Contacts Summary Oracle (which provides the average waiting time until the next contact), Contacts Oracle (which answers questions regarding contacts between two nodes at any point in time), Queuing Oracle (which gives information about buffer queuing) and Traffic Demand Oracle (which can answer any question regarding the present or future traffic demand). We propose here a potential implementation of the first two oracles.

Hui and Crowcroft [12] study the impact of predictable human interactions on forwarding in pocket switched networks (PSNs). By applying vertex similarity on a dataset extracted from mobility traces, they observe that adaptive forwarding algorithms can be built by using the history of past encounters. Furthermore, the authors design a distributed forwarding algorithm based on node

Trans. Emerging Tel. Tech. 2012; **00**:1–14
DOI: 10.1002/ett
Prepared using ettauth.cls

centrality and show that it is efficient in terms of hit rate and delivery latency. We also show here that a node's contact history is a good starting point for predicting its future behavior.

Islam and Waldvogel [13] state that opportunistic routing protocols in the literature are dependent on the history of devices for extracting routing information, but the size of this routing information is limited, which may introduce inaccuracies and thus a weaker message delivery. The authors analyze the predictive qualities of history-based routing algorithms using extensive simulation on real CRAWDAD traces [14], and reach the conclusion that the repetitive nature of a path is proportional to the mobility extent of the devices and thus contact history obtained from dense opportunistic networks can be reliable. We extend these conclusions by showing that the location of a mobile user can be pinpointed with a certain degree of predictability, based on the user's history of AP sightings.

In [15], Song et al. analyze the predictability of human behavior and mobility on user traces obtained from mobile carriers. They use several variants of the entropy function as the most fundamental quantity that captures the degree of probability which characterizes a time series. In our paper, we try to enhance their analysis in order to map it onto wireless network traces. Ihler et al. [16] model a normal periodic behavior of a time series by a time-varying Poisson process model and then modulate it using a hidden Markov process, in order to account for bursty events. They show that using a Poisson model is significantly more accurate at detecting future behavior and known events than a traditional threshold-based technique. This model can also be used to investigate periodicity in the data, such as systematic weekday and time of day effects. Since contact information in an opportunistic network is also a time series, we believe that it can also be approximated as a Poisson distribution, as shown in Section 5.

Our initial premises for this paper are that synergic patterns in academic environments are subject to repeatability. As opposed to previous more generic studies [15, 17] which study mobility and interactions in wide geographical regions such as cities and metropolitan areas and which are generally more interested in the actual physical distances covered by users, we focus towards environments where human behavior can be predictable, and try to understand the laws governing the human processes. Our work is thus somewhat similar to [18] because we are addressing academic environments, such as university campuses. However, we cover a more generic space in determining the social and connectivity predictability patterns in case of an academic environment, instead of focusing on a particular use case (determining synthetic paths in [18]). In this sense, we study predictability and mobility of peers over multiple communication technologies (Bluetooth, Wi-Fi), as well as interactions with wireless access points.

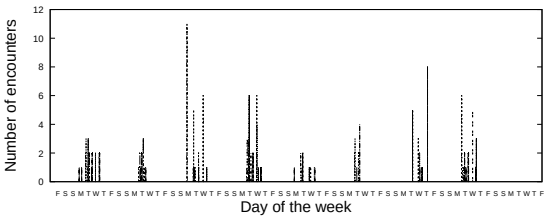
3. TRACING EXPERIMENT

For gathering traces of human mobility, an experiment was performed at the University POLITEHNICA of Bucharest in the spring of 2012 [2]. For this experiment, an application entitled HYCCUPS Tracer was implemented with the purpose of collecting contextual data from Android smartphones. It was ran in the background and collected availability and mobile interaction information such as usage statistics, user activity, battery statistics or sensor data, but what really interests us is the fact that it gathered information about a device's encounters with other nodes or with wireless access points. Encounter collection was performed in two ways: Bluetooth and AllJoyn [19]. Bluetooth interaction scanned for paired devices in the immediate vicinity and stored contact information such as the ID of the encountered device and the time and duration of contact. The information stored by AllJoyn tracing was similar, but was collected by constructing and deleting wireless sessions using the AllJoyn framework based on WiFi. The difference between Bluetooth and WiFi is that WiFi consumes more battery life, but is more stable [20]. We observed that AllJoyn interactions occurred much more often than those on Bluetooth. Thus, there were 20,658 WiFi encounters for a total of 66.27% of all the contacts, and only 6,969 Bluetooth contacts. We believe that such results were caused by the low range of Bluetooth. Tracing was executed periodically with a predefined timeout for Bluetooth, and asynchronously for AllJoyn interactions. More detailed information can be found in [2].

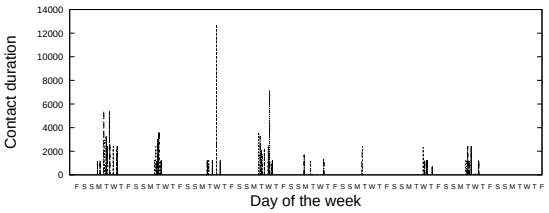
The duration of the tracing experiment was 64 days, between March and May 2012, and had 66 participants. They were chosen so that they covered as many years as

possible from both Bachelor and Master courses. Thus, there were: one first year Bachelor student, one third year Bachelor student, 53 fourth year Bachelor students (from five different study directions), three Master students, two faculty members and six external participants (from an office environment). We were interested only in the participants at the faculty, so we eliminated the external nodes. We also eliminated some nodes that had too little contact information, because they were irrelevant to our experiment. Such nodes belonged to students that did not keep their Android application on at all times when they were at the faculty as they were instructed, or who haven't attended many classes in the experiment period. In the end, we remained with 53 nodes in the experiment that had useful information.

An academic environment is a natural situation where an opportunistic network can bring benefits [21, 22, 23], therefore we consider the participants in this experiment as part of such a network. An academic environment, like the campus of a university, is a relatively small enclosed space with lots of potential device carriers (students and professors) and therefore many contact opportunities. Thus, opportunistic routing and dissemination are a natural fit for such an environment, since data can circulate quickly and reach all desired destinations with a good probability.



(a) Encounters



(b) Contact durations

Figure 1. Total encounters and contact durations per day for a random node from the POLITEHNICA trace.

4. PREDICTABILITY ANALYSIS

In order to verify if the behavior of nodes in an opportunistic network built in an academic environment is predictable, we have analyzed in detail the mobile trace presented in Section 3. This section presents an overview of this analysis, in regard to node encounters, contact duration and wireless access points sightings.

4.1. Encounters and Contact Duration

Because the nodes in the opportunistic network presented in this paper are students and teachers at the University POLITEHNICA of Bucharest, we believe that their behavior is predictable. This should happen because the participants have a fixed daily schedule and interact with each other at fixed times in a day. For example, a teacher and the students from a class interact when the students attend the teacher's class, which happens regularly each week. Likewise, two students in the same class would interact at almost all times when they are at the faculty.

We tried to prove that this supposition holds by analyzing the traces presented in Section 3. The first metric used was the total number of encounters between a node and all the other nodes. An encounter is considered to begin at the first moment when a node is in range of the current node and to end when the two nodes are not in range anymore for a certain period of time. The number of encounters specifies the popularity of a node (the more encounters it has, the more popular it is). The second metric we used was the contact duration of every encounter of a node in a given time interval. Similar to the number of encounters, it suggests the popularity of a node, but also its mobility. If a node has many encounters in a time interval, but all the encounters are short in terms of duration, it means that the node is highly mobile and it doesn't stay in the same place for long periods of time.

Figure 1(a) shows the total number of daily encounters of a randomly chosen node with each of the other nodes for the entire duration of the experiment. It can be seen that on Tuesdays, Wednesdays and Thursdays, the node has regular encounters with fairly the same nodes. The number of contacts per day sometimes differs, which probably happens because there were short periods of time when the nodes were not in contact (for example, a student went out of the faculty grounds), but the encountered nodes are basically the same every week. This shows (on a purely

Trans. Emerging Tel. Tech. 2012; **00**:1–14
DOI: 10.1002/ett
Prepared using ettauth.cls

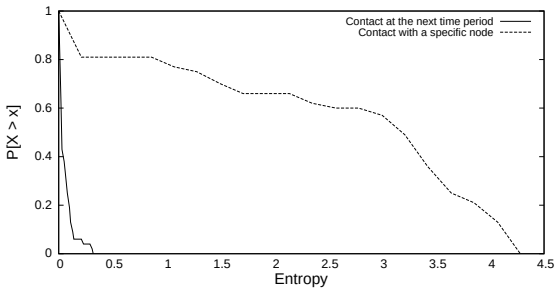


Figure 2. Entropy values for predicting the time of contact with any node and contacts with given nodes.

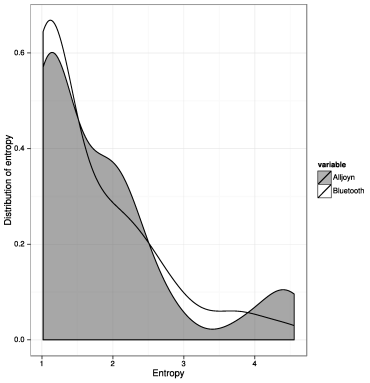


Figure 3. Distribution of node interaction entropy for Bluetooth and AllJoyn.

intuitive level for now) that there may be a certain amount of predictability in the behavior of our nodes (as we show later in this section, this is backed up by computing the entropy values).

Figure 1(b) presents contact durations per day for the same node as before. Just as in Figure 1(a), it can be seen that on Tuesdays, Wednesdays and Thursdays the contact durations are similar, the conclusion here being that contact durations are also seemingly predictable.

In order to verify if a node's behavior in the opportunistic network is predictable, we used Shannon's entropy, which is a measure of predictability (the lower the entropy, the higher the chances are of a prediction being successful). When the entropy is 0, it means that a node's behavior is 100% predictable. The general formula for entropy is:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i) \tag{1}$$

where X is a discrete random variable with possible values in the interval $x_1, ..., x_n$ and $p(X)$ is a probability mass function for X. We have split the entropy computation into two parts: predicting that the next encounter will be with a given node N, and predicting if there will be at least a contact at the next time interval. Combining these two values will result in a prediction of the time of an encounter with a given node.

The first step was to compute for each node the entropy for contacts with a given node N. The probability function of a node i in this case was computed as:

$$p_i(N) = \frac{enc_i(N)}{enc_i(*)} \tag{2}$$

where $enc_i(N)$ is the total number of encounters between i and N, and $enc_i(*)$ is the total number of encounters of i. The second probability function was computed as:

$$p_i = \frac{dur_i}{dur} \tag{3}$$

where dur_i is the total duration of node i's contacts, while dur is the total duration of the trace. Both these durations are measured in the same time units. We used different values for this unit (one second, one minute and one hour) and the results were similar. Figure 2 shows the cumulative distribution functions for entropies of the two probability functions, after computing the entropies for each node in the trace. Axis X shows the entropy value, while axis Y represents the amount of nodes that have a lower entropy than the value on axis X. It can be observed that having a contact at the next period of time is mostly predictable, because the entropy is always lower than 0.35. However, predicting the node that will be seen at the next encounter is not so easily done based solely on the history of encounters, and this is shown by the high entropy values (as high as 4.25, meaning that a node may encounter on average any one of $2^{4.25} \simeq 19$ nodes).

We also studied the hourly interactions of individuals on a daily basis and computed the probability that an individual interacts at least once each day at the same hour with any other peer. Figure 3 shows the density plot for the entropy of hourly interactions. As can be seen, AllJoyn hourly interactions peak almost as low as Bluetooth. The comparison between Bluetooth and WiFi comes down to a compromise between low range versus low power saving, since more powerful radios lead to much faster battery depletion.

In Section 5, we focus on attempting to predict the time of the next encounter with any node and the total number of encounters at that moment, and leave the prediction of contacts with a certain node for future work. As we have previously seen, further information is required in order to predict exactly what nodes will be encountered, such as knowledge about the social relationships between the participants in the opportunistic network. As shown in [24], there is a higher chance of a node interacting with nodes that are part of its social circle than with other nodes, and this information can be used in designing an efficient algorithm.

4.2. Wireless Access Point Interaction

This section presents an analysis regarding the predictability of a mobile node's interaction with fixed wireless access points. If it is high, it means we may be able to predict a node's daily route based on the locations of the access points it encounters.

In order to formalize the interactions between mobile nodes and wireless access points, we define a virtual location (VL) as the most relevant AP scanned during a predefined time interval. We empirically chose this interval to be an hour, while also taking into consideration the fact that, since we dealt with an academic environment, an hour was the usual unit of work. Because multiple access points can be scanned in an hour, it was necessary to use a heuristic that chose the most relevant one. Therefore, we propose two such heuristics, entitled FCFS (First Come First Served) and Alpha. FCFS chooses the first AP sighted during the time interval as the most relevant virtual location, whereas the Alpha heuristic weighs the number of sightings $count(VL_i)$ and average signal strength $avg(signal(VL_i))$ for each time interval, attempting to maximize the expression:

$$\alpha \times count(VL_i) + (1-\alpha) \times avg(signal(VL_i)) \quad (4)$$

The difference between the two is that FCFS describes a pseudo-random heuristic, while Alpha offers more control over the chosen VL, by tweaking α. A lower value means that signal strength is more important than the number of sightings, for situations where the mobility is reduced and the number of APs is low (such as closed spaces). On the other hand, a higher value for α is more suitable when we deal with scenarios where the nodes have high mobility. In such cases, signal strength is not as important as the number of sightings (since the signal strength is momentary, while the more encounters a node has with an access point, the more data it can send it).

Based on the two heuristics presented above, we define a VL sequence as the result obtained by splitting the collected mobility trace into one-hour intervals and generating an array of virtual locations for each interval. However, due to device malfunctions, low battery or lack of user conscientiousness, there were situations where a VL for an interval was not known. In such cases, we approximated it using a knowledge coefficient, similarly to the use of the q parameter by Song et al. [15], which characterizes the fraction of segments where the location is unknown. For our purposes, we chose a lower limit of 20% for the knowledge coefficient. Thus, for every node in the trace, 12 VL sequences were generated: one FCFS sequence and 11 Alpha sequences, which were obtained by varying the α parameter between 0 and 1 with a 0.1 step. We believe that these 12 sequences are sufficient for analyzing the predictability of wireless access point interaction.

In order to assess the predictability of AP interaction, we used three types of entropy for each node. The first one was the entropy of a node i traveling in random patterns and is defined as:

$$S_{rand}(i) = log_2(N_i) \quad (5)$$

where N_i is the total number of VLs that node i has sighted. Secondly, we analyzed the entropy of spatial traveling patterns without taking into account the temporal component of an interaction (also named temporally-uncorrelated entropy). It is defined as:

$$S_{unc}(i) = \sum_{j=1}^{N_i} -p_i(j) \times log_2(p_i(j)) \quad (6)$$

where $p_i(j)$ is the probability that node i interacts with a specific VL j. As opposed to S_{rand} in which we consider that each VL is visited with equal probability, S_{unc} also takes into consideration how often a user visits a specific VL. As such, $p_i(j)$ is the historical probability of $user_i$ visiting a virtual location VL_j. Finally, the third entropy value we used was the estimated entropy computed by means of a variant of the Lempel-Ziv algorithm [25],

Trans. Emerging Tel. Tech. 2012; **00**:1–14
DOI: 10.1002/ett
Prepared using ettauth.cls

which takes into account the history of a node's past encounters, thus correlating the temporal dimension with the VL interaction patterns. We constructed an estimator that computes this entropy as:

$$S_{est} = \left(\frac{1}{n} \sum_i \lambda_i \right)^{-1} \ln n, \tag{7}$$

where n is the length of the symbol sequence and λ_i is the shortest substring that appears starting from the index i, but which is not present for indexes lower than i. Kontoyiannis et al. [26] proved that S_{est} converges to the real entropy when n approaches infinity.

Song et al. [15] showed that there is a relationship between the three types of entropies:

$$S_{est}(i) \leq S_{unc}(i) \leq S_{rand}(i) < \infty \tag{8}$$

We believe that this is true for each node i in the trace presented in Section 3, as a participant taking random actions should be less predictable than another one frequenting VLs irregardless of time, and both are less invariable than a real mobile device owner taking logical decisions. Thus, we attempt to prove it in the rest of this section.

The first step in attempting to show that inequality 8 is valid for an academic environment was to analyze the distribution of distinct access points discovered for various weekly intervals.

Because we focus on interaction with wireless APs, we need to define a measure of sufficiency of the tracing data, namely observed interval sufficiency. This measure helps us determine if the tracing data has converged to a point where it is sufficiently informed in order to perform additional operations on it. Moreover, we define the observed interval sufficiency as the minimum interval in which the discovery of access points converges. Since we are dealing with WiFi networks in an academic environment, we can assess that the surroundings of such a tracing experiment are limited and patterns are visible sooner than in mobile networks (e.g. GSM), which ought to limit the tracing interval to several months or weeks, rather than a year. It can be seen in Figure 4 that 10 weeks of tracing were enough for the number of discovered APs to converge. Thus, when the convergence occurs, it is safe to say that the most encountered wireless network devices have already been sighted. It should also

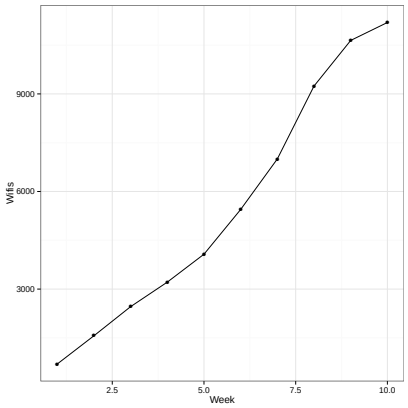
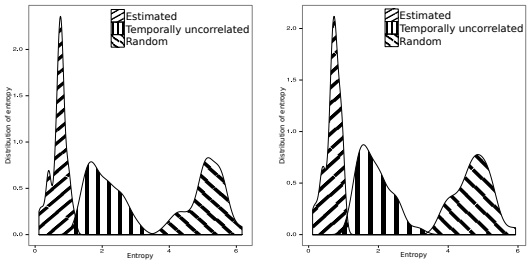


Figure 4. The number of discovered access points for each week in the POLITEHNICA trace.



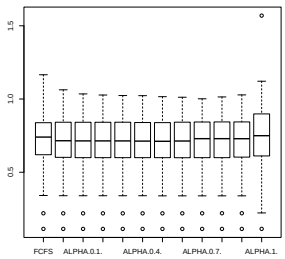
(a) FCFS (b) Alpha 0.7

Figure 5. The inequality of entropies for S_{rand} , S_{unc} , S_{est} for the POLITEHNICA trace.

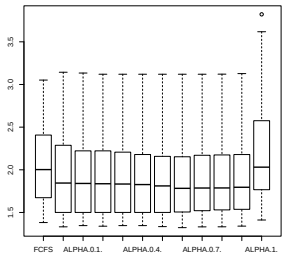
be noted that most of the participants appear to have limited mobility, since they encountered few access points. This was expected, because we collected the trace in an academic environment, which is a small closed space with many participants and interactions between them.

Due to the fact that for a period of 18 days at the beginning of the tracing experiment, the participants have not been very conscientious in regard to keeping the HYCCUPS Tracer on at all times whilst on faculty grounds, there was a lot of missing information. For this reason, we performed our wireless AP analysis using only information gathered in the last 46 days of the experiment, for 8 hours each day (the interval when the most participants were at the faculty). Therefore, we obtained VL sequences with 368 symbols each (8 hours $\times$ 46 days), where each symbol corresponded to an outstanding VL for a specific hour.

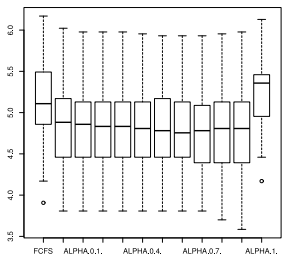
Figure 5 illustrates the density plots for the distributions of entropy $P(S_{rand})$, $P(S_{unc})$ and $P(S_{est})$ for all users



(a) Estimated entropy



(b) Temporally uncorrelated entropy



(c) Random entropy

Figure 6. Comparison of entropy distributions for all sequences for the POLITEHNICA trace.

considering the two VL-choosing heuristics: FCFS and Alpha (with a parameter of 0.7). As expected, it can easily be seen that inequality 8 holds for our experiment. Figure 6 shows a comparison between the distributions of the three proposed entropies on all VL sequences by using box-and-whisker diagrams. FCFS exhibits one of the most skewed distributions for all three types of entropy, which highlights the fact that pseudo-random simulations have a tendency to suffer from unrealistic traits. It can also be observed that Alpha with a parameter of 0.7 generates results close to normal distributions, which is why we chose to show it in comparison to FCFS in Figure 5.

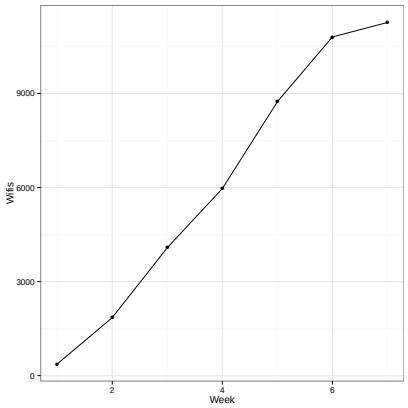
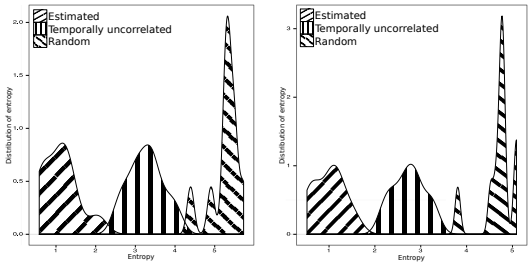


Figure 7. Distribution of the number of discovered access points for each week in the Rice trace.



(a) FCFS (b) Alpha 0.7

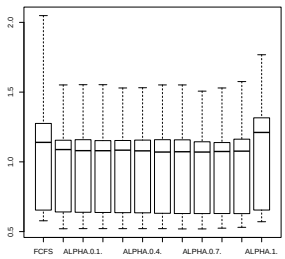
Figure 8. The inequality of entropies for S_{rand} , S_{unc} , S_{est} for the Rice trace.

Based on these results, it can be stated that the interactions of nodes from our collected mobility trace is subject to predictability. A user carrying a mobile device can thus be pinpointed to one of $2^{0.68} \simeq 1.6$ locations, whereas a user making random decisions can be found in one of $2^{4.94} \simeq 30.7$ locations.

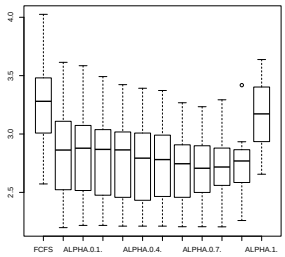
In order to show that our conclusions hold for other scenarios, we also performed a wireless AP interaction analysis on a different type of trace, namely Rice*. The trace set is composed of cellular and WiFi scan results from the Rice community in Houston, Texas. The tracing experiment was performed in 2007, between January 16 and February 28, totaling 6,055 wireless access points discovered by 10 participants. Figure 7 shows the distribution of discovering APs and it can be seen that the eight weeks of the experiment were almost sufficient for acquiring convergence, which means that we were able to

*http://crawdad.cs.dartmouth.edu/rice/context/

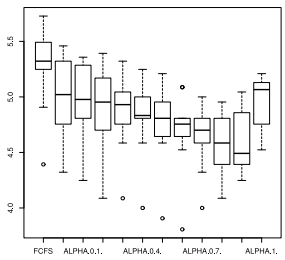
Trans. Emerging Tel. Tech. 2012; **00**:1–14
DOI: 10.1002/ett
Prepared using ettauth.cls



(a) Estimated entropy



(b) Temporally uncorrelated entropy



(c) Random entropy

Figure 9. Comparison of entropy distributions for all sequences for the Rice trace.

study the predictability of wireless AP interaction. Thus, Figure 8 shows the density plots for the distributions of the three entropies for FCFS and Alpha with a parameter of 0.7. The situation is similar to our trace, and inequality 8 holds for Rice as well.

Furthermore, Figure 9 illustrates the comparison between the distributions for the three measures of entropy on all generated sequences. As opposed to our trace, the distributions of the estimated entropy are heavily skewed, but consistent. Although the Rice trace set contains data from only 10 users, all of them had a high degree of collected knowledge. This increased informational gain also appears to have affected the Random entropy, as can be seen in Figure 9(c): each generated sequence is generally different from the others. This further proves that, in real life, random heuristics are not able simulate human behavior. Also, in both traces, FCFS seems to have the same behavior: the distributions are skewed and the peaks are higher, but they still do not reflect the worst case scenario.

These results show that, in a different trace such as Rice, human-carried nodes are still subject to repeatability. Furthermore, a node can be pinpointed to one of $2^{1.6} \simeq 3.03$ locations, whereas a random user can be found in one of $2^{4.9} \simeq 29.85$ locations.

5. PREDICTING FUTURE NODE ENCOUNTERS

As we have seen in Section 4.1, the entropy for predicting if a contact will take place at the next time interval is lower than 1, which means that the behavior of a node in terms of encounters with any other nodes is highly predictable. Since there are only two cases in this prediction (i.e. having a contact or not having a contact), we can model a node's behavior as a Bernoulli distribution, which is a particular case of a binomial distribution. However, simply knowing if there will be a contact at a given time may not be enough for a good opportunistic routing algorithm. Therefore, it would be good if we were able to know exactly how many contacts will there be in the specified time interval, and a binary distribution such as Bernoulli does not offer such information. Consequently, we believe that a Poisson distribution might be suitable for this situation, because it expresses the possibility of a number of events (in our case encounters with other nodes) to occur in a fixed time interval. We show in the rest of this section that the Poisson distribution applies to the trace presented in Section 3.

The probability mass function of a Poisson distribution is the following:

$$P(N, \lambda) = \frac{e^{-\lambda}\lambda^N}{N!} \tag{9}$$

where in our case $P(N, \lambda)$ represents the probability of a node having N contacts at a given time interval, and λ is the total number of events divided by the number of units in the data. In order to prove that a

Poisson distribution applies to our trace, we used Pearson's chi-squared test [27], which tests a null hypothesis stating that the frequency distribution of mutually exclusive events observed in a sample is consistent with a particular theoretical distribution (in our case Poisson). We applied the chi-squared test for every node in the network.

The time interval chosen for applying the Poisson distribution and the chi-squared test was one hour. We tried to choose this interval in order to obtain a fine-grained analysis of the data. Choosing a smaller interval (such as a minute) and estimating the next contact incorrectly may lead to missing it completely (in the sense of using it appropriately). When we have an interval such as an hour, we can predict that in the next hour there will be a certain number of contacts with a higher rate of success, and the opportunistic routing algorithm can be ready for those contacts in the respective hour.

The first step of the chi-squared test was to count the frequency distribution of contacts per hour for the entire duration of our trace. We then stated the null hypothesis, namely that the number of encounters a node has per hour follows a Poisson distribution. The λ parameter can either be included in the hypothesis or it can be estimated from the sample data (as it was in our case). We computed it using the max likelihood method by averaging the number of encounters per hour over the entire experiment. Knowing λ, we were then able to find out the probability for having N encounters at the next time interval according to the Poisson distribution. Using this probability, we finally performed the chi-squared test according to the formula:

$$\chi^2_{k-p-1} = \sum_k \frac{(f_o - f_e)^2}{f_e} \qquad (10)$$

where f_o is the observed frequency, f_e is the expected frequency (computed using the Poisson distribution), k is the number of classes and p is the number of parameters estimated from the data (in this case 1, the λ value).

We used the 0.05 level of significance for proving the hypothesis using a chi-squared table, and the results can be seen in Figure 10 (1). We also included the nodes that have not had any encounters in the "Accepted" category, since a distribution with only zeros is a valid Poisson distribution. As can be seen from Figure 10, only 20.75% of the hypotheses were accepted in this case. However, we have observed in Section 4 that a node's encounter

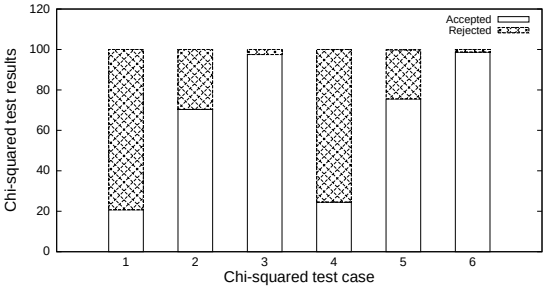


Figure 10. Results of chi-squared tests for various scenarios. Datasets 1, 2 and 3 are computed using the total number of encounters and varying the max likelihood (1 - for the entire experiment, 2 - per weekday, 3 - per hour of a day of the week). Datasets 4, 5 and 6 are computed using unique encounters.

history has a somewhat repetitive pattern for days of the week, so we then attempted to compute λ as the averaged number of contacts in the same day of the week. Therefore, we ended up with a larger number of chi-squared hypotheses to prove (53 nodes $\times$ 7 days in a week) for each node, but also with a much finer-grained approximation of the data. The results for this situation can be seen in Figure 10 (2), with only 29.65% of the hypotheses being rejected. Still we went one step further, knowing that students at a faculty generally follow a fixed schedule in given days of the week and thus we computed the max likelihood value as an average per hour per day of the week (consequently having 53 $\times$ 7 $\times$ 24 hypotheses per node). Thus, the results obtained were very good, with only 2.49% of all the hypotheses rejected.

The previous results were computed for the total number of encounters in an hour. However, if the Android tracer application misbehaved at some point in the experiment and instead of logging a long contact between two nodes, logged a large number of very short contacts, the results of applying a Poisson probability may be wrong. Because of this situation, we also applied the chi-squared tests described above using only unique contacts. Therefore, the number of contacts in an hour will be equal to the number of different nodes encountered in that hour. The results are shown in Figure 10 (4,5,6). For the first test case (with λ computed over the entire experiment), 75.47% of the hypotheses were rejected. In the test that uses the average per day of the week, 24.26% of all chi-squared hypotheses were rejected and finally just 1.31% of distributions were not Poisson according to the chi-squared test for computing the max likelihood value per hour.

Trans. Emerging Tel. Tech. 2012; **00**:1–14
DOI: 10.1002/ett
Prepared using ettauth.cls

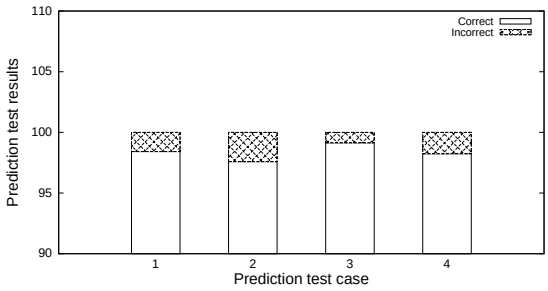


Figure 11. Prediction success of the Poisson distribution. Datasets 1 and 2 are computed using the total number of encounters (1 - the next to last week, 2 - the last week) and datasets 3 and 4 are computed using unique encounters.

Although these results look good, it might be that this situation only happens on this particular trace. Consequently, we ran all the tests presented in this section on another trace [24] which was also performed at the University POLITEHNICA of Bucharest, but for a shorter duration (35 days) and with fewer participants (22 students and teachers). However, the results for the λ-per-hour test with unique contacts are even better than for the current trace, since only 0.11% of the hypotheses were rejected.

However, the traces performed in the campus of our faculty may be a particular case, meaning that we cannot generalize our assumptions yet. Therefore, we tested our theory on a different type of trace, entitled St. Andrews [28], which was collected using a mobile sensor network with Tmote Invent devices carried by 27 members of the University of St. Andrews for a period of 79 days, in which the participants were asked to carry their devices and to keep them on at all times, whether in or out of the town of St. Andrews. The Invent devices were able to detect and store information about encounters between each other within a radius of 10 meters, and were programmed to send discovery beacons at every 6.67 seconds. This trace corresponds to a different situation that the traces performed at the University POLITEHNICA of Bucharest, since this is a larger and more open space, with less regularity, less contact opportunities and smaller contact durations. The results we obtained were good, with only 12.39% of the chi-squared hypotheses being rejected.

To further prove our assumptions, we have eliminated the last two weeks from the trace and computed the Poisson distribution probabilities for each hour per day of the week on the remaining series. We compared the value that had the highest Poisson probability (i.e. the most likely value according to the distribution) with the real values. If the Poisson predictions were to be correct, then the two values should be equal. The results of this test for both total and unique contacts are shown in Figure 11. It can be seen that for total encounters 97.59% of the Poisson-predicted values are correct for the next to last week, and 98.42% for the last. When taking into account individual encounters, the predictions are even better (98.24% and 99.14% respectively).

We have shown in this section that, by knowing the history of encounters between the nodes in an opportunistic network in every hour of every day of the week, we can predict the future behavior of a device in terms of number of contacts per time unit. Having this knowledge helps the opportunistic routing algorithm decide what to do with the data it carries (data that has either been generated locally, or that is currently only being carried around by the node). If the node knows that it will have few contacts this hour, but the number will increase for the next hour, it may choose to keep the data bundles until then, instead of forwarding and then deleting them. Likewise, if the next time intervals are known to have few encounters, the node might choose to forward data while it still has the chance, lest the messages are delayed for long periods of time. An entire set of heuristics can be applied in such a situation, and this is something we wish to study in detail as future work. We also wish to analyze a wider range of scenarios in the future, not only academic environments. We strongly believe that the results we presented here will still hold true, since humans have strong habits and generally follow repetitive mobility patterns every day.

6. CONCLUSIONS AND FUTURE WORK

We have presented here a mobile trace performed over an academic environment at the University POLITEHNICA of Bucharest for two months in the spring of 2012. We analyzed it from the point of view of node encounters, contact durations and wireless access point interactions, and we also showed that it exhibits predictability both from the standpoint of node contacts, as well as in regard to wireless AP sightings. This means that not only is an ON node's behavior in terms of encountered devices

predictable, but also its future position (with regard to the access point it will be in contact with at a given time).

We have shown that a node's history of contacts in terms of the number of encounters can be modeled as a Poisson distribution with very good predictability results. We have proven these assumptions to hold for two other traces [24, 28] that have different conditions than the one presented in this paper. Having the knowledge about future encounters is basically an implementation of the Contacts Summary Oracle, according to Jain et al. [11]. Because the network presented here offers a lot of contact opportunities, a good routing algorithm can easily be created with only this information. Thus, we believe that expanding routing algorithms with knowledge about future node encounters can lead to better performance.

However, adding even more knowledge in the shape of Jain's Contacts Oracle, i.e. knowing which nodes will be encountered in a given time interval, will improve the routing algorithm further. Therefore, we plan to attempt a prediction of the contact durations and particular nodes that a device will be in contact with in the future, using similar methods. It is also important to note that, since nodes in an opportunistic network are devices belonging to human carriers, the social aspect should be taken into account.

Since it has been shown that nodes tend to interact more often with other members of their social communities [24], if the predictions are made by combining the history of events with knowledge about social relationships and communities (instead of only using the history), a very efficient opportunistic routing algorithm can be implemented.

As we have shown that the location of a user carrying a mobile device can be predicted with a high chance of success, we also wish to attempt this prediction for various types of traces, starting with an academic environment.

Having the Contacts Summary Oracle and the Contacts Oracle, we plan to develop our own opportunistic routing algorithm and test its efficiency on the traces presented here. We wish to focus on the advantages of the social aspects of opportunistic networking, since we believe this is the future in pocket-switched networks.

ACKNOWLEDGEMENT

This work was partially supported by project "ERRIC - Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1", ID: 264207, and by national project "TRANSYS - Models and Techniques for Traffic Optimizing in Urban Environments", Contract No. 4/28.07.2010, Project CNCSIS-PN-II-RU-PD ID: 238. The work has been co-funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

REFERENCES

1. Ciobanu RI, Dobre C. Predicting encounters in opportunistic networks. *Proceedings of the 1st ACM workshop on High performance mobile opportunistic systems*, HP-MOSys '12, ACM: New York, NY, USA, 2012; 9–14, doi:10.1145/2386980.2386983. URL http://doi.acm.org/10.1145/2386980.2386983.
2. Marin RC, Dobre C, Xhafa F. Exploring Predictability in Mobile Interaction. *Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on*, IEEE, 2012; 133–139, doi:10.1109/eidwt.2012.29. URL http://dx.doi.org/10.1109/eidwt.2012.29.
3. Conti M, Giordano S, May M, Passarella A. From opportunistic networks to opportunistic computing. *Comm. Mag.* 2010; **48**:126–139. URL http://dl.acm.org/citation.cfm?id=1866991.1867009.
4. Hui P, Crowcroft J, Yoneki E. BUBBLE Rap: social-based forwarding in delay tolerant networks. *Proc. of the 9th ACM int. symp. on Mobile ad hoc networking and computing*, MobiHoc '08, ACM: New York, USA, 2008; 241–250, doi:http://doi.acm.org/10.1145/1374618.1374652. URL http://doi.acm.org/10.1145/1374618.1374652.
5. Nguyen HA, Giordano S, Puiatti A. Probabilistic routing protocol for intermittently connected mobile ad hoc network (propicman). *2007 IEEE Int. Symp. on a World of Wireless Mobile and Multimedia*

Networks 2007; :1–6.

6. Boldrini C, Conti M, Jacopini J, Passarella A. HiBOp: a History Based Routing Protocol for Opportunistic Networks. *World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE Int. Symp. on a*, 2007; 1–12, doi:10.1109/WOWMOM.2007.4351716. URL http://dx.doi.org/10.1109/WOWMOM.2007.4351716.
7. Yoneki E, Hui P, Chan S, Crowcroft J. A socio-aware overlay for publish/subscribe communication in delay tolerant networks. *Proc. of the 10th ACM Symp. on Modeling, analysis, and simulation of wireless and mobile systems*, MSWiM '07, ACM: New York, NY, USA, 2007; 225–234, doi:http://doi.acm.org/10.1145/1298126.1298166. URL http://doi.acm.org/10.1145/1298126.1298166.
8. Lenders V, May M, Karlsson G, Wacha C. Wireless ad hoc podcasting. *SIGMOBILE Mob. Comput. Commun. Rev.* 2008; **12**:65–67, doi:http://doi.acm.org/10.1145/1374512.1374535. URL http://doi.acm.org/10.1145/1374512.1374535.
9. Boldrini C, Conti M, Passarella A. Exploiting users' social relations to forward data in opportunistic networks: The HiBOp solution. *Pervasive Mob. Comput.* 2008; **4**:633–657, doi:10.1016/j.pmcj.2008.04.003. URL http://dl.acm.org/citation.cfm?id=1412744.1412768.
10. Ciobanu R, Dobre C. Data dissemination in opportunistic networks. *18th Int. Conf. on Control Systems and Computer Science*, CSCS-18, 2011; 529–536.
11. Jain S, Fall K, Patra R. Routing in a delay tolerant network. *SIGCOMM Comput. Commun. Rev.* Aug 2004; **34**(4):145–158, doi:10.1145/1030194.1015484. URL http://doi.acm.org/10.1145/1030194.1015484.
12. Hui P, Crowcroft J. Predictability of human mobility and its impact on forwarding. *2008 Third International Conference on Communications and Networking in China*, IEEE, 2008; 543–547, doi:10.1109/chinacom.2008.4685085. URL http://dx.doi.org/10.1109/chinacom.2008.4685085.
13. Islam MA, Waldvogel M. Prediction quality of contact history in opportunistic networks. *2011 IFIP Wireless Days (WD)*, IEEE, 2011; 1–3, doi:10.1109/WD.2011.6098154. URL http://dx.doi.org/10.1109/WD.2011.6098154.
14. CRAWDAD. http://crawdad.cs.dartmouth.edu/.
15. Song C, Qu Y, Blumm N, Barabsi AL. Limits of Predictability in Human Mobility. *Science* 2010; **327**:1018–1021, doi:10.1126/science.1177170.
16. Ihler A, Hutchins J, Smyth P. Learning to detect events with markov-modulated poisson processes. *ACM Trans. Knowl. Discov. Data* Dec 2007; **1**(3), doi:10.1145/1297332.1297337. URL http://doi.acm.org/10.1145/1297332.1297337.
17. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C. A tale of many cities: universal patterns in human urban mobility Oct 2011. URL http://arxiv.org/abs/1108.5355.
18. Kim M, Kotz D, Kim S. Extracting a Mobility Model from Real User Traces. *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, 2006; 1–13, doi:10.1109/infocom.2006.173. URL http://dx.doi.org/10.1109/infocom.2006.173.
19. Inc QIC. Introduction to AllJoyn. HT80-BA013-1 Rev. B 2011.
20. Ferro E, Potorti F. Bluetooth and Wi-Fi wireless protocols: a survey and a comparison. *IEEE Wireless Comm.* 2005; **12**(1):12–26, doi:10.1109/MWC.2005.1404569. URL http://dx.doi.org/10.1109/MWC.2005.1404569.
21. McNett M, Voelker GM. Access and mobility of wireless PDA users. *SIGMOBILE Mob. Comput. Commun. Rev.* 2003; **7**:55–57, doi:http://doi.acm.org/10.1145/965732.965744. URL http://doi.acm.org/10.1145/965732.965744.
22. Henderson T, Kotz D, Abyzov I. The changing usage of a mature campus-wide wireless network. *Proc. of the 10th annual int. conf. on Mobile computing and networking*, MobiCom '04, ACM: New York, USA, 2004; 187–201, doi:http://doi.acm.org/10.1145/1023720.1023739. URL http://doi.acm.org/10.1145/1023720.1023739.
23. Hui P, Chaintreau A, Scott J, Gass R, Crowcroft J, Diot C. Pocket switched networks and human mobility in conference environments. *Proc. of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, ACM: New

York, USA, 2005; 244–251, doi:http://doi.acm.org/10.1145/1080139.1080142. URL http://doi.acm.org/10.1145/1080139.1080142.

24. Ciobanu RI, Dobre C, Cristea V. Social aspects to support opportunistic networks in an academic environment. *Proceedings of the 11th international conference on Ad-hoc, Mobile, and Wireless Networks*, ADHOC-NOW'12, Springer-Verlag: Berlin, Heidelberg, 2012; 69–82, doi:10.1007/978-3-642-31638-8_6. URL http://dx.doi.org/10.1007/978-3-642-31638-8_6.
25. Ziv J, Lempel A. Compression of individual sequences via variable-rate coding. *Information Theory, IEEE Transactions on* Sep 1978; **24**(5):530–536, doi:10.1109/tit.1978.1055934. URL http://dx.doi.org/10.1109/tit.1978.1055934.
26. Kontoyiannis I, Algoet PH, Suhov Y, Wyner AJ. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *Information Theory, IEEE Transactions on* May 1998; **44**(3):1319–1327, doi:10.1109/18.669425. URL http://dx.doi.org/10.1109/18.669425.
27. Stuart A, Ord K, Arnold S. *Kendall's Advanced Theory of Statistics, Classical Inference and the Linear Model*, vol. Volume 2A (2007 reprint)). Sixth edn., Wiley, 1999. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0470689242.
28. Bigwood G, Rehunathan D, Bateman M, Henderson T, Bhatti S. Exploiting self-reported social networks for routing in ubiquitous computing environments. *Proceedings of the 2008 IEEE International Conference on Wireless & Mobile Computing, Networking & Communication*, IEEE Computer Society: Washington, DC, USA, 2008; 484–489, doi:10.1109/WiMob.2008.86. URL http://dl.acm.org/citation.cfm?id=1475703.1476536.