Techniques and Applications to Analyze Mobility Data

Radu-Corneliu Marin, Radu-Ioan Ciobanu, Ciprian Dobre, Fatos Xhafa

Abstract Mobility is intrinsic to human behavior and influences the dynamics of all social phenomena. As such, technology has not remained indifferent to the imprint of mobility. Today we are seeing a shift in tides as the focus is turning towards portability, as well as performance; mobile devices and wireless technologies have become ubiquitous in order to fulfil the needs of modern society. Today the need for mobility management is gradually becoming one of the most important and challenging problems in pervasive computing. In this chapter, we present an analysis of research activities targeting mobility. We present the challenges of analyzing and understanding the mobility (is mobility something that is inherently predictable? are humans socially inclined to follow certain paths?), to techniques that use mobility results to facilitate the interaction between peers in mobile networks, or detect the popularity of certain locations. Our studies are based on the analysis of real user traces extracted from volunteers. We emphasize the entire process of studying the dynamics of mobile users, from collecting the user data, to modelling mobility and interactions, and finally to exploring the predictability of human behavior. We point out the challenges and the limitations of such an endeavour. Furthermore, we propose techniques and methodologies to study the mobility and synergy of mobile users and we show their applicability on two case studies.

Radu-Corneliu Marin
University POLITEHNICA of Bucharest, Splaiul Independentei 313, Bucharest, Romania, e-mail: radu.marin@cti.pub.ro

Radu-Ioan Ciobanu
University POLITEHNICA of Bucharest, Splaiul Independentei 313, Bucharest, Romania, e-mail: radu.ciobanu@cti.pub.ro

Ciprian Dobre
University POLITEHNICA of Bucharest, Splaiul Independentei 313, Bucharest, Romania, e-mail: ciprian.dobre@cs.pub.ro

Fatos Xhafa
Universitat Politecnica de Catalunya, Girona Salgado, 1–3, 08034, Barcelona, Spain, e-mail: fatos@lsi.upc.edu

1 Introduction

In recent years, the ubiquitousness of mobile devices has led to the advent of various types of mobile networks. Such is the case of opportunistic networks (ONs), which are based on the store-carry-and-forward paradigm: a node stores a message, carries it until it encounters the destination or a node that is more suitable to bring the message closer to the destination, and then finally forwards it. Thus, an efficient routing algorithm for ONs should be able to decide if an encountered node is suitable for the transport of a given message with a high probability. It should also be able to decide whether the message should be copied to the encountered node, or moved altogether.

Since opportunistic networks are composed of human-carried mobile devices, routing and dissemination algorithms deployed in such networks should take advantage of the properties of human mobility, in order to be effective. Although for a time people have used mathematical models to simulate human mobility, it has been shown in recent years that the properties that were believed to be true regarding human mobility are actually incorrect. For example, contrary to what was believed, it was shown that human interactions follow a power law function, while the times between two successive contacts are described by a heavy-tailed distribution [4]. These results have shifted the focus from synthetic mobility models to real life traces that can offer a far better view of human interaction.

In this chapter, we present an analysis of current research activities on mobility. We highlight research efforts designed towards the collection and analysis of mobility traces. We present two case studies, on traces collected with the purpose of acquiring human interaction data in an academic environment. Both traces (entitled UPB 2011 and UPB 2012) were collected at the University POLITEHNICA of Bucharest, their participants being students and professors at the faculty.

Knowledge about the distributions of encounters in a trace, or the dependence on the time of day, is necessary in designing routing and forwarding algorithms. Therefore, our second contribution is to analyze the collected traces in terms of contact times distribution and highlight each trace's properties. Moreover, since we are dealing with academic environments where the participants' fixed schedules make them interact with a certain regularity, we attempt to prove the predictability of a node's future encounters. We propose doing this by approximating a node's behavior as a Poisson distribution and using the chi-squared test to demonstrate our assumption. Finally, we also analyze the predictability of an ON node's contacts with fixed wireless access points (APs), while also proposing a methodology for studying traces which involve the scanning of APs.

Preliminary versions of our work were previously published in [25] and [6]. In this chapter we present more extensive results, describing the proposed traces in detail and analyzing them in regard to various ON-specific metrics (such as contact and inter-contact times) and the impact that these metrics have on the outcome of the trace. Furthermore, we describe the two tracing applications that were used to collect the data, and highlight a series of benefits and limitations of using mobility traces instead of mathematical models.

The chapter is organized as follows. Section 2 emphasizes on the process of collecting tracing data from real mobile users, discussing both the advantages and pitfalls of conducting such an endeavour. Sections 3 and 4 focus on the techniques for exploring mobility and interactional patterns applied on two case studies, as well as the resulted experimental data. Section 5 concludes our chapter with the implications of our work and thoughts for future development.

2 Collecting Inter-Cooperative Mobility Data

The challenge in working with mobility arises from two difficult problems: formalizing mobility features and extracting mobility models. Currently, there are two types of mobility models in use: real mobile user traces and synthetic models [3]. Basically, traces are the results of experiments recording the mobility features of users (location, connectivity), while synthetic models are pure mathematical models which attempt to express the movement of devices.

Although they have been regarded as suspect models due to the limitations in mapping over reality [26], synthetic models have been largely used in the past, the two most popular models being random walks on graphs [13], which are similar to a Brownian motion, and the random waypoint mobility model [21], in which pauses are introduced between changes in direction or speed.

In 2005, Barabási [2] introduced a queueing model which disproved the claims of synthetic models based on random walks on graphs. Furthermore, Barabási's model showed that the distributions of inter-event times in human activity are far from being normal as they present bursts and heavy tails. This happens because people do not move randomly, but their behavior is activity-oriented [10, 11, 19]. This endeavour has paved the way for researchers in human dynamics, as the Barabási model [2] is continuously being developed [36, 28, 37, 32] and experiments with it are using a variety of new interesting sources: web server logs [9, 14, 15], cell phone records [30, 29, 33] and wireless network user traces [24, 25].

The remainder of this section briefly describes two important sources for mobility analysis and modelling: the Haggle project and the CRAWDAD archives. Furthermore, it presents two mobility tracing applications developed and deployed at the University POLITEHNICA of Bucharest, and finally it highlights the pros and cons of such applications and the lessons learned from using them.

2.1 The Haggle Project

Haggle[1] is a European Commission-funded project that designs and develops solutions for opportunistic networks communication, by analyzing all aspects of the

[1] http://www.haggleproject.org/

main networking functions, such as routing and forwarding, security, data dissemination and (most importantly for the work we present here) mobility traces and models [35]. The results proposed in Haggle were soon followed by a series of subsequent other research projects targeting similar interests: SCAMPI [31], SOCIALNETS [1], etc. Haggle is today seen by many as the project that created the premises for the advancements on human mobility for information and communications technology-related aspects.

In order to obtain mobility models, Haggle deals with the analysis and modelling of contact patterns between devices, introducing notions such as contact duration and inter-contact time. The contact duration, or contact time (CT), is the time when two devices are in range of each other, while the inter-contact time (ICT) is the period between two successive contacts of the same two devices. The contact duration influences the capacity of the network, while the inter-contact time affects the feasability and latency of the network.

Several mobility traces have been performed in the context of Haggle, mostly using Bluetooth-enabled devices such as iMotes. These are mobile devices created by Intel, based on the Zeevo TC2001P SoC, with an ARMv7 CPU and Bluetooth support. Two iMote traces, called Intel and Cambridge, have been presented and analyzed in [4]. The Intel trace was recorded for three days in the Intel Research Cambridge Laboratory, having 17 participants from among the researchers and students at the lab. The Cambridge trace was taken for five days, at the Computer Lab of the University of Cambridge, having as participants 18 doctoral students from the System Research Group. For both traces, the iMotes performed five-second scans at every two minutes, and searched for in-range Bluetooth devices. Each contact was represented by a tuple (MAC address, start time, end time). Both internal as well as external contacts were analyzed, where encounters between two devices participating in the experiment were considered internal contacts, while encounters with other devices were external contacts. The authors analyzed the distribution of CT and ICT, as well as the influence of the time of day on encounter opportunities. Regarding inter-contact time, the traces showed that it exhibits an approximate power law shape, which means that inter-contact distribution is heavy-tailed. The authors showed this observation to hold regardless of the time of day, by splitting a day into three-hour time intervals and noticing that the resulting distributions still maintained power law shapes. Contact durations were also noticed to follow power laws, but with much narrower value ranges and higher coefficients.

In addition to the Intel and Cambridge traces, another trace entitled Infocom was presented and analyzed in [5]. It was conducted during the IEEE INFOCOM conference in Miami in 2005, and had 41 conference attendees as participants, for a total duration of four days. The conclusions were similar to the ones above, namely that the distribution of the inter-contact times between two nodes in an opportunistic network is heavy-tailed over a large range of values, and that it can be approximated to a power law with a less than one coefficient. The authors showed that certain mobility models in effect at the time the paper was written (such as the random waypoint model) did not approximate the real life traces correctly. Similar to the Infocom trace, another trace was performed the following year at the same conference, but

on a larger scale. There were 80 participants, chosen so that 34 of them formed four groups based on their academic affiliations. Apart from the 80 mobile devices, 20 other long-range iMotes were also deployed at strategic positions around the conference site [17]. Moreover, a trace was also performed in Hong Kong, where 37 people from a bar were given iMotes and were asked to return after five days [17].

2.2 CRAWDAD

The Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD)[2] represents the effort of the National Science Foundation (NSF) to create an archive which stores wireless tracing data from many international contributors. The need for such a high capacity datastore has spawned from the data starvation that plagued research in wireless networks, as well as the limitations of synthetic models which were used as a replacement for real life user traces. CRAWDAD comes to the aid of researchers by hosting the contributed traces and developing better tools for collecting and post-processing data (e.g. anonymizing user traces).

Based on the fact that tracing experiments and studies related to them are extremely difficult to set up (as shown in Sect. 2.4), CRAWDAD is aimed at solving problems that are automatized: anonymizing the captured data in order to preserve privacy, or creating development tools for traces such as parsers.

The CRAWDAD initiative supports the human dynamics community as it understands the importance of data captured from live wireless networks in identifying and understanding the real problems, in evaluating possible solutions for said problems and also in evaluating new applications and services.

2.3 Social Tracer and HYCCUPS

In order to have mobility traces for an academic environment in certain conditions and containing a specific set of features, we performed two tracing experiments at our faculty. For each of these traces, we implemented an Android application that was deployed on the participants' smartphones for the duration of the experiment. This section presents the two applications and the tracing experiments performed.

2.3.1 Social Tracer

For our initial trace (which we called UPB 2011), we implemented an application entitled Social Tracer[3] in Android [7]. The participants in the tracing experiment

[2] http://crawdad.cs.dartmouth.edu/

[3] http://code.google.com/p/social-tracer/

were asked to run the application whenever they were in the faculty grounds, as we were interested in collecting data about the mobility and social traces in an academic environment. Social Tracer sent regular Bluetooth discovery messages at certain intervals, looking for any type of device that had its Bluetooth on. These included the other participants in the experiment, as well as phones, laptops or other types of mobile devices in range. The reason Bluetooth was preferred to WiFi was mainly the battery use [12]. For example, in four hours of running the application on a Samsung I9000 Galaxy S with discovery messages sent every five minutes, the application used approximately 10% of the battery's energy. The period between two successive Bluetooth discovery invocations could be set from the application, ranging from 1 to 30 minutes (the participants were asked to keep it as low as possible, in order to have a more fine-grained view of the encounters).

When encountering another Bluetooth device, the Social Tracer application logged data containing its address, name and timestamp. The address and name were used to uniquely identify devices, and the timestamp was used for gathering contact data. Data logged was stored in the device's memory, therefore every once in a while participants were asked to upload the data collected thus far to a central server located within the faculty premises. All gathered traces were then parsed and merged to obtain a log file with a format similar to the ones from Haggle. Successive encounters between the same pair of devices within a certain time interval were considered as continuous contacts, also taking into account possible loss of packets due to network congestion or low range of Bluetooth.

The UPB 2011 tracing experiment was performed for a period of 35 days at the University POLITEHNICA of Bucharest in 2011, between November 18 and December 22. There were a total of 22 participants, chosen to be as varied as possible in terms of study year, in order to obtain a better approximation of mobility in a real academic environment. Thus, there were twelve Bachelor students (one in the first year, nine in the third and two in the fourth), seven Master students (four in the first year and three in the second) and three research assistants.

2.3.2 HYCCUPS

In order to get more relevant data regarding a mobile device user's behavior, we implemented a new tracer, called HYCCUPS, which is an Android application designed to collect contextual data from smartphones. The application runs in the background and can collect traces for multiple features. These features can be classified by the temporality of acquisition into static or dynamic, or by the semantic interpretation into availability or mobility features.

Moreover, static properties can be determined at application startup and are comprised of the device's traits, while dynamic features are momentary values acquired on demand. On the other hand, availability features represent values pertaining to the overall computing system state, while mobility features describe the interaction of the device with the outside world. We chose to collective an extensive dataset for future use.

As such, the features that the HYCCUPS Tracer can collect are as follows:

- **Minimum and maximum frequency**: static properties describing the bounds for Dynamic Voltage/Frequency Scaling (DVFS).
- **Current frequency**: momentary value of the frequency according to DVFS.
- **Load**: the current CPU load computed from */proc/stat*.
- **Total memory**: static property of the device describing the total amount of memory available.
- **Available memory**: momentary value which represents the amount of free memory on the device (bear in mind that, in Android, free memory is wasted memory).
- **Out of memory**: asynchronous event notifying that the available memory has reached the minimal threshold and, in consequence, the Out Of Memory (OOM) Killer will stop applications.
- **Memory threshold**: the minimal memory threshold that, when reached, triggers the OOM events.
- **Sensor availability**: static property which conveys the presence of certain sensors (e.g. accelerometer, proximity).
- **Accelerometer**: the accelerometer modulus is a mobility feature which characterizes fine grain movement (if available).
- **Proximity**: proximity sensor readings (if available).
- **Battery state**: the current charging level (expressed in %) and also the current charge state.
- **User activity**: availability events representing user actions that trigger opening/closing application activities.
- **Bluetooth interactions**: momentary beacons received from nearby paired devices (similar to what Social Tracer does).
- **AllJoyn interactions**: interactions over WiFi modelled using the AllJoyn framework [20]. AllJoyn is an open-source peer-to-peer software development framework which offers the means to create ad hoc, proximity-based, opportunistic inter-device communication. The true impact of AllJoyn is expressed through the ease of development of peer-to-peer networking applications provided by: common APIs for transparency over multiple operating systems; automatic management of connectivity, networking and security and, last but not least, optimization for embedded devices.
- **WiFi scan results**: temporized wireless access point scan results.

Tracing is executed both periodically, with a predefined timeout, as well as asynchronously on certain events such as AllJoyn interactions or user events. This chapter concentrates on dynamic mobility features represented by the last three tracing features from the above list.

Therefore, the second tracing experiment, entitled UPB 2012, lasted for 65 days, in the spring of 2012 and also took place at the University POLITEHNICA of Bucharest. A total of 66 volunteers participated varying in terms of year and specialization: one first year Bachelor student, one third year Bachelor student, 53 fourth year Bachelor students, three Master students, two faculty members and six external participants (only from office environments). The experiment implied an initial

startup phase, also called pairing session, when all attendants were asked to meet and pair all devices for Bluetooth interactions.

The participants were asked to start the HYCCUPS Tracer each weekday between 10 AM and 6 PM as we assumed this was the interval in which most participating members were attending classes or work. As expected, the volunteers in our experiment did not always respect the instructions with conscientiousness, so on occasion they didn't turn the application on. Nonetheless, the results proved that the captured tracing data was sufficient for our needs.

2.4 Benefits and Limitations

The main reason for developing and using tracing applications such as the ones previously presented instead of synthetic mobility models spawns from the need for better mapping onto real life situations. As previously stated, trace models follow a heavy-tailed distribution with spikes and bursts, making the random walks on graphs model and other such models obsolete.

The major benefit of tracing applications is the use of a custom data model in order to relate to real situations, real problems and optimized solutions for said issues. However, this can also lead to a pitfall: if the data model is not correctly designed at the start of the experiment, the entire outcome of the analysis can be biased.

Among the potential challenges of setting up our tracing experiments, we dealt with the following:

- Finding volunteers representative to our goals was not such an easy task as it may seem. For example, if we would have chosen all participants from the same class, then our results would have been biased because we would have been limiting our targeted scope to a partition of our community graph instead of reaching the entire collective. Moreover, all of the candidates for the experiment needed to have Android devices capable of tracing our data model: Bluetooth connectivity, WiFi connectivity, sensors etc.
- The design and development of the tracing application needed to take into account compatibility with multiple types of viable Android devices of variate versions. Furthermore, when developing the tracers, we were obliged to take into account the additional overhead of our applications, as most participants complained about the supplementary power consumption.
- The installation effort of the tracer was tremendous due to issues such as Bluetooth pairing: all of the participants' devices needed to pair to each other in order for us to be able to trace their interactions.
- Last, but not least, we were confronted with the human factor of such experiments: the lack of conscientiousness of our volunteers. Due to the participants not running the tracing application as instructed, the collected data was incomplete. Furthermore, this affected the analysis of said results, as we needed to deploy measures to deal with uncertainty.

3 Techniques for Data Analysis

This section presents several techniques for analyzing the data collected in a mobility trace. We begin by looking at the distribution of encounters and contact times, and then go on to present how we can verify if a trace exhibits contact predictibility in the shape of a Poisson distribution. Finally, we study the limits of predictability of the mobility and behavior of mobile users with regard to wireless access points.

3.1 Contact Times Distribution and Time of Day Dependence

The first step in analyzing a mobility trace should be looking at the distribution of contact and inter-contact times. Approximating these distributions using heavy-tailed or power law functions (as shown in Sect. 2.1) can help in the creation of synthetic mobility models, but more importantly it can aid in the development of routing and forwarding algorithms suitable for the specific network that is being analyzed. Apart from the contact and inter-contact times described in Sect. 2.1, two other metrics defined in [16] are the any-contact time (ACT) and inter-any-contact time (IACT). These are similar to the previous metrics, except that they are computed with regard to any node in the trace, not per pairs of nodes. Thus, the ACT is the time in which any internal or external node is in range with the current observer, while the IACT is the period when the current device does not see anyone in range. The former metric specifies the time window in which a node can forward messages to other participants in the network, whereas the latter is the opposite: the period when a node doesn't have any contacts at all.

The distribution of contacts with internal or external devices also highlights various characteristics of a mobility trace, such as the contact opportunities with given nodes, or the possibilities of using external devices for transporting data to a certain destination. In addition to analyzing the contact times and distribution, one should take into account the dependence of a trace on the time of day. Knowledge about this dependence can prove to be truly useful, especially in situations like the ones we analyzed and presented in Sect. 2.3, where there is generally little activity in the network during the times of day when the students aren't at the faculty. This means that a routing or forwarding algorithm should take advantage of those times of day when there are many contacts, in order to reduce the effect of time periods with few encounters. We analyze the UPB 2011 and UPB 2012 traces in terms of contact times, contact distribution and time of day dependence in Sect. 4.1.

3.2 Contact Predictability

An important challenge in mobile networks is knowing when and to which node should a message be passed, in order for it to reach its destination as fast as possible.

Therefore, it would be important if we were able to predict the future behavior of a node in such a network, in regard to its encounters and contact durations. We propose a way to predict this behavior by analyzing a node's past encounters and approximating the time series as a Poisson distribution.

Since the nodes in an academic trace are students and professors from the faculty, 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 professor and his students interact when the students attend the professor's class, which happens regularly each week. Likewise, two students from the same class would interact at almost all times when they are at the faculty.

Thus, we attempt to prove this supposition by analyzing the traces in terms of predictability. The first metric we use is the total number of encounters between a node and the other nodes. The number of encounters mainly specifies the popularity of a node, since the more encounters a node has, the more popular it is. The second metric is 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 very mobile and it doesn't stay in the same place for long periods of time.

In order to verify if a node's behavior in the opportunistic network is predictable, we use Shannon's entropy, which is basically 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 formula for entropy is $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$, 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. A first possibility would be to compute $p(X)$ as the probability of encountering node N at the next time interval. However, the sum of probabilities in this case would not be 1, because a node might be in contact with more than one other node at a given time. Thus, we split the entropy computation into two parts: predicting that the next encounter will (or won't) be with node N, and predicting if a contact will take place at the next time interval. In theory, combining these two values will result in a prediction of the time of an encounter with a given node. Moreover, they also have a use on their own and not necessarily together.

We start by computing the entropies for contacts with a given node N. The probability function in this case is given by the ratio between the total number of times node N was encountered and the total number of contacts with any other nodes during the experiment. The second entropy function is given by the ratio between the number of time units a node was in contact with another node for, and the entire duration of the experiment. Based on these two entropies, we can decide whether it is possible to predict the node that the next encounter will be with, and whether there will be a contact at the next time interval. However, for now we will only focus on the second scenario.

Because there are only two values for this prediction (having a contact or not having a contact at the next time interval), a node's behavior regarding future contacts can be modelled as a Bernoulli distribution, which is a particular case of a binomial

distribution. However, simply knowing if there will be a contact at the next time interval is not enough for a good opportunistic routing algorithm, since we need to know how many contacts there will be, in order to decide if the data packet should be forwarded, kept, or forwarded but also kept. The Bernoulli distribution does not offer such information, so we propose using the Poisson distribution, because it expresses the possibility of a number of events (in our case encounters with other nodes) to occur in a fixed time interval.

The probability mass function of a Poisson distribution is $P(N,\lambda) = \frac{e^{-\lambda}\lambda^N}{N!}$, where in our case $P(N,\lambda)$ represents the probability of a node having N contacts at a given time interval. In order to prove that a Poisson distribution applies to our traces, we use Pearson's chi-squared test [34], 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 apply the chi-squared test for every node in the network individually and present the results in Sect. 4.2.

3.3 Predictability of Interacting with Wireless Access Points

This section presents a proposal for a methodology not just as a basis for studying already existing mobile traces which involve scanning of nearby wireless APs, but also as a set of guidelines for future tracing application developers to take into consideration when designing and developing tracers.

The basic principles for the methodology are inspired from the analysis conducted by Song et al. in [33]. In their paper, the authors study the limits of predictability in the mobility and behavior of mobile users over Cell towers. Here we try to formalize, adapt and enhance their analysis in order to map it onto wireless network traces. As such, we need to fill the gap between their analyzed context and ours, namely bridging the difference between the range of Cell towers and wireless access points.

Seeing that the main focus of this methodology is interaction with wireless APs we need to define a measure of sufficiency of the tracing data, namely observed interval sufficiency. This measure determines 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. Recalling that we are dealing with WiFi networks in academic and office environments, we can assess that the surroundings of such a tracing experiment are limited and, as such, patterns are visible sooner than in mobile networks (e.g. GSM). This should limit the tracing interval to several months or weeks, rather than a year.

Also an important factor of our analysis is the number of subjects involved in the experiment, as well as their conscientiousness (or control over the tracing application; we will emphasize more on conscientiousness later in this section). As such,

we have empirically discovered that a minimum of 10–20 users are necessary in order to provide statistical correctness of the analysis.

Naturally, the next step in such an analysis is to formalize the interactions between users and wireless APs. The reader should be aware that the tracing experiments that this methodology is aimed at must contain the temporized results of wireless AP scans. As such, we define a virtual location (VL) as the most relevant access point scanned by a user during an hour. Taking into consideration that multiple APs can be scanned by a user during an hour, we need to define a heuristic of choosing the most relevant one. We represent VLs as Basic Service Set Identifiers (BSSIDs), since Service Set Identifiers (SSIDs) are prone to name clashes, fact which may influence our study.

Although it may seem unintuitive at first, choosing hours to be the analysis' temporal step has many reasonable explanations. First of all, tracing applications which gather wireless scan results might use different timing intervals and we considered an hour should be viewed as a maximum value. Furthermore, the object of our study refers to academic and office environments in which an hour is usually the unit of work. We propose the use of two VL-choosing heuristics:

1. **First Come First Served (FCFS)**: choose the first sighting of an access point as the most relevant VL. The purpose of this heuristic is to mimic a pseudo-random algorithm of picking VLs.
2. **Alpha**: choose the most outstanding virtual location by weighing both the number of sightings during an hour, as well as the average wireless signal strength. Basically, we choose a VL as the access point that maximizes the following expression:

$$\alpha \times count(VL_i) + (1-\alpha) \times average(signalStrength(VL_i)) \tag{1}$$

If FCFS describes a pseudo-random heuristic, Alpha offers more control over choosing access points; by tweaking the α factor we can guide the algorithm towards more realistic situations: when α is lower, signal strength is more important than the number of sightings, thus better mapping on a situation with reduced mobility (closed surroundings) where the are few access points and the signal strength is the most valuable feature. On the other hand, when increasing α we turn our attention towards situations with a high range of mobility; signal strength is only momentary, whereas sighting an access point multiple times shows a certain level of predictability.

Based on these two heuristics, we define a VL sequence as the result of splitting up the entire tracing interval into hourly intervals and generating a chain of VL symbols for each hour of the monitored period. Whenever the VL of a user is unknown for a segment, it is marked with a special symbol (e.g. '?'). These shortcomings in tracing data, also known as lack of conscientiousness, are approximated by means of the knowledge coefficient similar to the q parameter used by Song et al. [33], which characterizes the fraction of segments in which the location is unknown. Also similar to [33], we chose a lower limit of 20% for our knowledge coefficient as we found it sufficient for our needs.

In total, a set of 12 sequences are to be generated for each user: 1 FCFS and 11 Alpha sequences (by sweeping the α value from 0 to 1 with a 0.1 step value). Based on the VL sequences, three measures of entropy for each user should be computed:

- S_{rand} is the entropy of a user i travelling in random patterns and is defined as:

$$S_{rand}(i) = \log N_i \tag{2}$$

  where N_i is the total number of VLs that user i has discovered.
- S_{unc} is the entropy of spatial travelling patterns without taking into account the temporal component of an interaction (also named temporally uncorrelated entropy). It is defined as:

$$S_{unc}(i) = \sum_{j=1}^{N_i} -p_i(j) \times \log p_i(j) \tag{3}$$

  where $p_i(j)$ is the probability of user i to interact with a specific VL_j.
- S_{est} is the estimated entropy computed by means of a variant of the Lempel-Ziv algorithm [38], which takes into consideration the history of passed encounters. By so doing, we correlate the temporal dimension with the VL interaction patterns. We have constructed an estimator which computes entropy as:

$$S_{est} = \left(\frac{1}{n} \sum_i \lambda_i \right)^{-1} \log n, \tag{4}$$

  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. S_{est} converges to the real entropy when $n \to \infty$ [23].

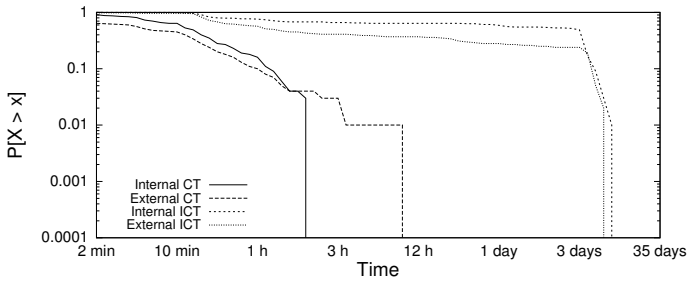
We consider that $S_{est}(i) \leq S_{unc}(i) \leq S_{rand}(i) < \infty$ [33] to be a reasonable assumption for each user i, as a participant taking random actions will be less predictable than another one frequenting VLs iregardless of time, and both are less invariable than a real user taking logical decisions.

4 Experimental Results

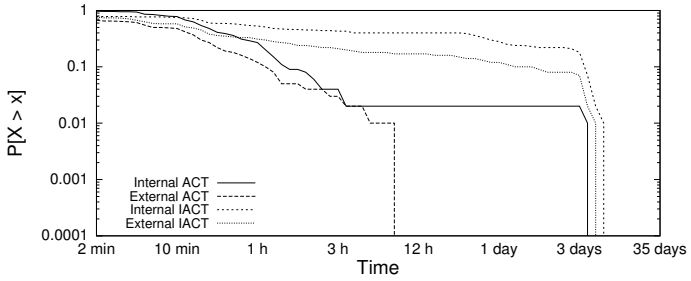
This section present the results obtained when applying the data analysis techniques presented in Sect. 3 on our two traces.

4.1 Contact Times Distribution and Time of Day Dependence

We begin our analysis with the contact times distribution and time of day dependence for the two traces presented in Sect. 2.3.



(a) Contact (CT) and inter-contact (ICT) times



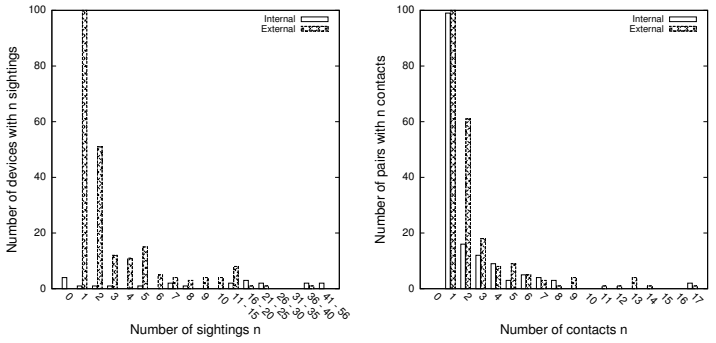
(b) Any-contact (ACT) and inter-any-contact (IACT) times

Fig. 1 Probability distributions of contact and inter-contact times (UPB 2011)

4.1.1 UPB 2011

In the UPB 2011 trace, there were 22 internal devices numbered from 0 to 21. The total number of contacts between two internal devices (i.e. internal contacts) was 341, while the number of external contacts was 1,127. There were 655 different external devices sighted during the course of the experiment, which means that in average each different external device has been seen about 2 times. External devices may be mobile phones carried by other students or laptops and notebooks found in the laboratories at the faculty. Some of these external devices have high contact times because they may belong to the owner of the internal device that does the discovery, therefore being in its proximity for large periods of time. However, external contacts are in general relatively short.

Figure 1(a) shows the distribution of contact and inter-contact times for the entire duration of the experiment for all internal devices. As shown in [16], the distribution of contact times follows an approximate power law for both types of devices, as well as contact time and inter-contact time. The contact time data series is relevant when discussing the bandwidth required to send data packets between the nodes in an opportunistic network, because it shows the time in which a device can communicate with other devices. As stated before, the number of internal contacts is 341, with the average contact duration being 30 minutes, which means that internal contacts have generally been recorded between devices belonging to students at-



(a) Sightings of a device (b) Contacts between pairs of devices

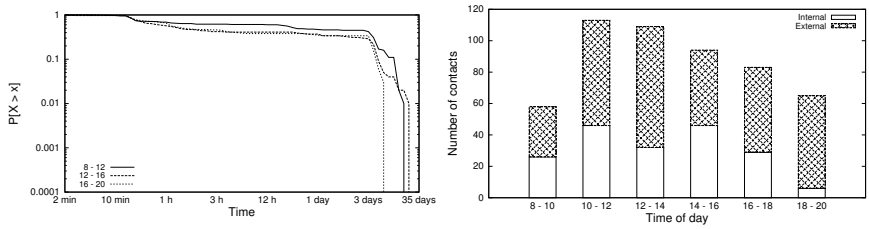
Fig. 2 Distribution of the number of sightings of a device and the number of contacts between pairs of devices (UPB 2011)

tending the same courses or lecturers and research assistants teaching those courses. External contacts also follow an approximate power law, with an average duration of 27 minutes. However, in this case there are certain external contacts that have a duration of several hours. This situation is similar to the one previously described, where these devices belong to the same person carrying the internal device. The inter-contact time distribution shows a heavy tail property, meaning that the tail distribution function decreases slowly. The impact of such a function in opportunistic networking has been studied in more detail in [4] for four different traces. The authors conclude that the probability of a packet being blocked in an inter-contact period grows with time and that there is no stateless opportunistic algorithm that can guarantee a transmission delay with a finite expectation.

Figure 1(b) shows any-contact and inter-any-contact times. As can be seen from the figure, they are greater than regular contact times, but the shape of the distribution is also a power law function. A conclusion that can be drawn from these charts is, as observed in [16], that contact times are bigger and intervals between contacts are smaller, so if a node wants to perform a multicast or to publish an object in a publish/subscribe environment it has a great chance of being able to do so.

Figure 2(a) shows the distribution of the number of times an internal or external node was sighted by other devices participating in the experiment. It can be seen that the maximum number of encounters of an internal device is 55 during the course of the 35 days of the experiment, whereas some internal nodes have never been seen. Most internal devices have been seen from 16 to 20 times. As for external devices, the majority of them have been encountered less than 5 times, with 534 of them having been sighted only once. There are few exceptions, as three external devices have been encountered more than 16 times. The conclusion is that there is a large number of nodes available in such an environment that can be used to relay a message, meaning that there is a lower chance of traffic congestion.

Figure 2(b) presents the number of times specific pairs of devices saw each other. It shows that the maximum number of contacts between two internal nodes or an in-



(a) Inter-contact times for three time periods (b) Contacts for six time periods

Fig. 3 Time of day dependence (UPB 2011)

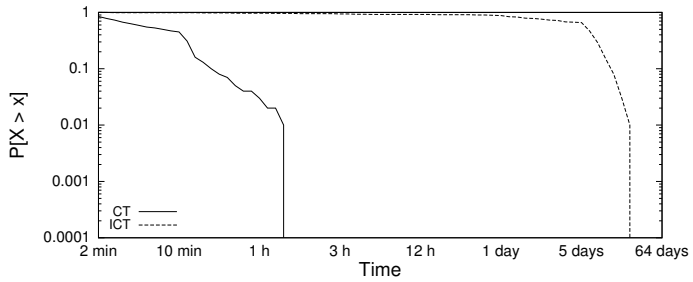
ternal and an external node is 17. Generally the number of contacts with external devices is larger than the number of contacts with internal devices. The maximum number of internal devices spotted by a participant is 17, whereas some participants have only encountered external nodes. Most of the internal devices have been in contact with between 10 and 15 other internal devices. As shown previously, the total number of external devices encountered during the 35 days of the experiment is far greater than that of the internal devices. There are six participants that have encountered between 15 and 50 external devices and five that have been in contact with more than 50 external nodes. The maximum number of different external devices spotted by a single participant is 197.

Figure 3(a) shows the distribution of inter-contact times for both types of devices for three time intervals. They are chosen between 8 AM and 8 PM because that is when students or teachers are at the faculty, and this experiment in not concerned with what happens in the rest of the day. The 8 AM–8 PM interval has been split into three parts, corresponding to three main time periods of the working day: morning (8 AM–12 PM), noon (12–4 PM) and afternoon (4–8 PM). As we can see from the figure, the three plots are very similar, following the same approximate power law function. This is different from [16], where daytime periods have a greater power law coefficient than night periods. This happens because we are only interested in periods when there are classes.

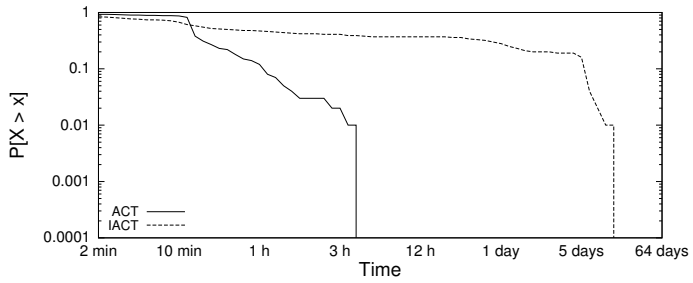
Figure 3(b) shows the percentage of contacts that take place in six two-hour intervals between 8 AM and 8 PM. It can be seen that most contacts (113) happen between 10 AM and 12 PM and the smallest number of contacts in a two-hour interval (58) is recorded between 8 AM and 10 AM. External contacts have a distribution similar to the one for all contacts, which shows that the faculty is populated the most between 10 AM and 2 PM. This can also be explained by the fact that at noon students usually have lunch at the cafeteria, so they meet in a common place.

4.1.2 UPB 2012

The total number of internal contacts for the UPB 2012 trace was 12,003, which is far greater than for UPB 2011, showing not only that there were more partici-



(a) Contact (CT) and inter-contact (ICT) times

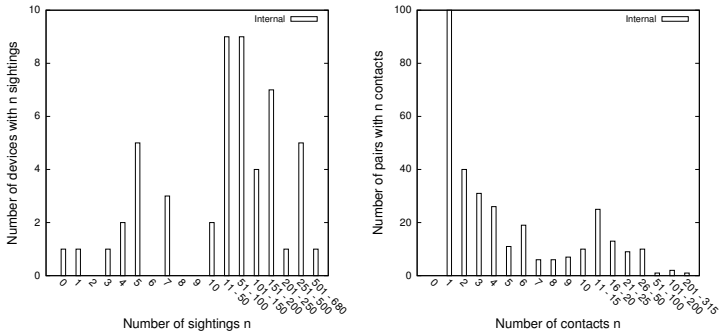


(b) Any-contact (ACT) and inter-any-contact (IACT) times

Fig. 4 Probability distributions of contact and inter-contact times (UPB 2012)

pants in this experiment, but that they were more conscientious in terms of turning their tracing application on when they were at the faculty. Unfortunately, the tracing application did not export information regarding external contacts, so in this section we will only analyze encounters with internal devices. Out of all the internal contacts, 13% of them were registered on Bluetooth and 87% on AllJoyn.

We analyzed the distribution of contact and inter-contact times for this trace as well, and the results are shown in Fig. 4(a). It can be seen that the curve of the distribution of contact times for internal devices is very similar to the corresponding one from the first trace. However, since there are more devices in this situation, the average contact duration decreases to about 14 minutes. The inter-contact time distribution, also shown in Fig. 4(a), exhibits a heavy tail property just like the one from UPB 2011. The average duration between two contacts for this trace is 4.5 hours, but it has to be taken into account that this value is computed simply by subtracting contact start and finish times. This means that if a contact takes place one day, and another happens the next day, the inter-contact time will be more than 8 hours. Both the contact and the inter-contact time distributions follow an approximate power law. Figure 4(b) shows any-contact and inter-any-contact times as well, which also follow an approximate power law, with the times higher than for regular contacts. The inter-any-contact times are very high in this case because there are no external nodes in this trace, which means that the time between any contacts is only computed using internal nodes. In the UPB 2011 trace, the inter-any-contact time



(a) Sightings of a device (b) Contacts between pairs of devices

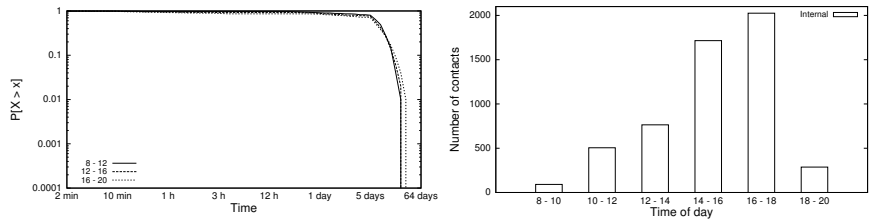
Fig. 5 Distribution of the number of sightings of a device and the number of contacts between pairs of devices (UPB 2012)

computation used the external nodes as well, which appeared more often, even if only for short periods of time.

Figure 5(a) shows the distribution of the number of times a device participating in the experiment was seen by the other participants. The maximum number of encounters of a certain node is 680, which is a lot higher than for the UPB 2011 trace. This means that some nodes in this trace are more popular and that there are more encounters and thus more possibilities of exchanging information. Out of the chosen nodes, only one has not been encountered at all during the experiment, and the majority of devices have been sighted between 51 and 100 times, which is a clear improvement over the previous trace. The average number of times a device has been encountered by other nodes in the experiment is 106, which means almost twice a day.

Figure 5(b) outlines the number of times specific pairs of nodes have encountered each other. Most of them have met each other only once (100), but there are plenty of pairs of devices that have been in contact more than once (going to as much as 315 encounters between a pair of devices). The average number of encounters between the same two devices is 50 (much higher than for the UPB 2011 trace), which is a sign that devices met relatively often. This conclusion is useful in implementing a routing algorithm for opportunistic networks, because it means that this trace is closer to a real life situation than UPB 2011.

Figure 6(a) shows the distribution of inter-contact times per time interval. The three time intervals were the same as the ones chosen in Sect. 4.1.1: 8 AM–12 PM, 12–4 PM, 4–8 PM. Again, the three plots are very similar to each other, following the same power law function. Figure 6(b) shows how many contacts happened in six two-hour intervals from 8 AM to 8 PM. Unlike the UPB 2011 trace, the most contacts happened between 4 PM and 6 PM. This is an indicative of the fact that most participants in the experiment had classes in that time interval. Many contacts were also recorded between 12 PM and 2 PM, which was lunch time, when students may meet in the cafeteria.



(a) Inter-contact times for three time periods (b) Contacts for six time periods

Fig. 6 Time of day dependence (UPB 2012)

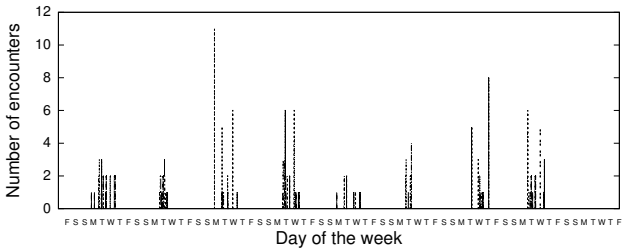
4.2 Contact Predictability

As we have shown in Sect. 3.2, we propose approximating a node's future behavior as a Poisson distribution, and we show the results in this section. We performed our analysis on the UPB 2012 trace presented in Sect. 2.3 because it has a larger number of participants and encounters for a longer period of time than UPB 2011.

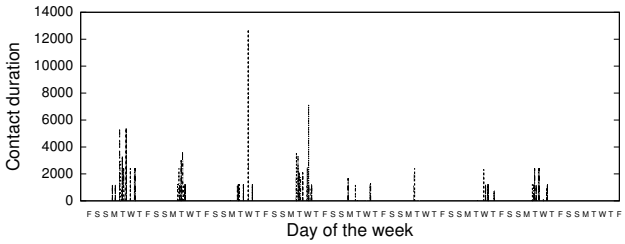
Figure 7(a) shows an example of a random node's behavior in terms of encounters with other nodes. It can clearly be seen that there is a weekly pattern, i.e. that on Tuesdays, Wednesdays and Thursdays the node has regular encounters with roughly the same nodes. The number of contacts in a day may differ, but this generally happens because there are short periods of time when the nodes weren't in contact or because of the unreliability of the Bluetooth protocol (thus yielding bursty contacts). The figure shows (for now just on a purely intuitive level) that there is a certain amount of regularity (and thus, predictability) in the behavior of the participants in the UPB 2012 experiment. Figure 7(b) presents contact durations per day for the same node as before. Just as in Fig. 7(a), it can be seen that on Tuesdays, Wednesdays and Thursdays the contact durations are similar.

As we previously said in Sect. 2.3, we use two entropy functions: one for predicting that the next encounter will (or won't) be with node N, and the other for predicting if a contact will take place at the next time interval. Figure 8 shows the cumulative distribution functions for the two entropies. 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 $2^{4.25} \approx 19$ nodes).

Since the entropy for predicting if a contact will take place at the next time interval is always lower than 1, the behavior of a node in terms of encounters with other nodes is highly predictable. This is the reason why we proposed using a Poisson distribution. 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. When



(a) Encounters



(b) Contact durations

Fig. 7 Total encounters and contact durations per day for a random node

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 traces. 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 maximum 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 for the time interval according to the formula $\chi^2_{k-p-1} = \sum_k \frac{(f_o - f_e)^2}{f_e}$, where f_o is the observed frequency, f_e is the expected frequency (computed using the Poisson distribution), k is the number of classes (which depends on the way the number of encounters is distributed for each node) 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 by using a chi-squared table, and the results can be seen in Fig. 9(a) (chart 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 Fig. 9(a), only 20.75% of the hypotheses were accepted in this case. However, we have observed previously that a node's encounter 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

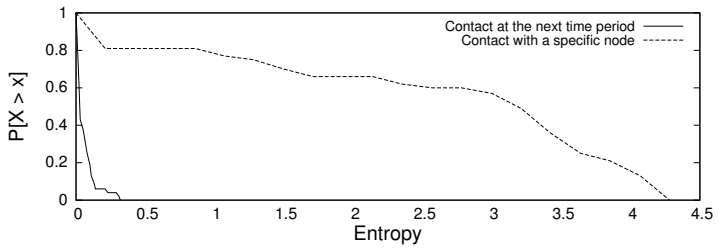


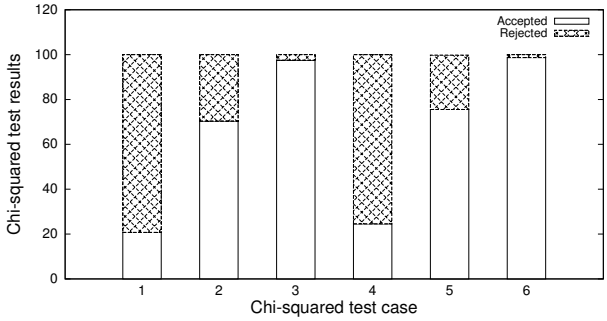
Fig. 8 Entropy values for predicting contacts and the time of contact

of chi-squared hypotheses to prove, but also with a much finer-grained approximation of the data. The results for this situation can be seen in Fig. 9(a) in chart 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 maximum likelihood value as an average per hour per day of the week. Thus, the results obtained were very good, with only 2.49% of all the hypotheses rejected, as shown by chart 3 in Fig. 9(a).

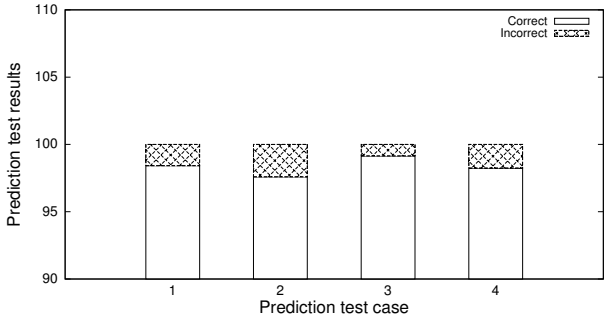
The results from charts 1–3 in Fig. 9(a) are computed for the total number of encounters in an hour. However, if the HYCCUPS application misbehaved at some point in the experiment and instead of logging a long contact between two nodes, logged a large number of very short contacts, the results of applying 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 is equal to the number of different nodes encountered in that hour. The results are shown in the final three charts in Fig. 9(a). For the first test case (with λ computed over the entire experiment, as seen in chart 4 from Fig. 9(a)), 75.47% of the hypotheses were rejected. In the test that uses the average per day of the week (chart 5), 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 maximum likelihood value per hour of a weekday (chart 6).

In order to prove that these results aren't valid for the UPB 2012 trace only, we also ran them on UPB 2011. 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.

To further prove our assumptions, we eliminated the last two weeks from the UPB 2012 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 Fig. 9(b). It can be seen that for total encounters 97.59% of the Poisson-predicted values are correct for the next to last week (chart 1), and 98.42% for the last (chart 2). When taking into account individual encounters, the predictions are even better: 98.24% for next to last week (chart 3) and 99.14% for the last week (chart 4).



(a) Chi-squared



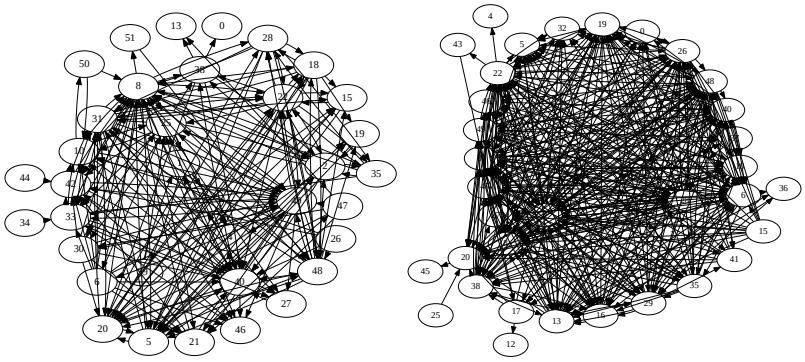
(b) Prediction success

Fig. 9 Chi-squared test results and prediction success of the Poisson distribution; for the chi-squared tests, 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), while datasets 4, 5 and 6 are computed using unique encounters; for the prediction success, 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

We have shown in this section that, by knowing the history of encounters between the nodes in a mobile network in every hour of every day of the week, we can successfully predict the future behavior of a device in terms of number of contacts per time unit.

4.3 Predictability of Interacting with Wireless Access Points

Our initial premises for the following experiment are that synergic patterns in academic and office environments are subject to repeatability. As opposed to previous more generic studies [33, 27], we focus towards environments where human behavior can be predictable, and try to understand the physical laws governing the human processes. Our work, from this perspective, is somewhat similar to [22].



(a) AllJoyn (b) Bluetooth

Fig. 10 Detected communities with a contact threshold of 1,200 s and a community threshold of 6

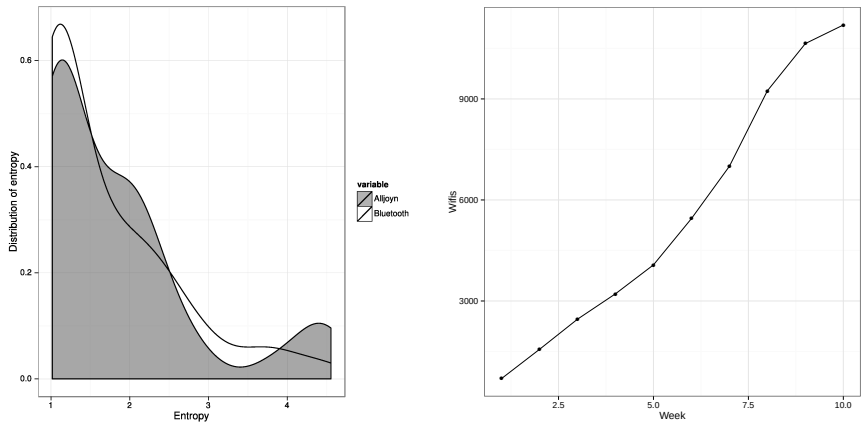
However, we cover a more generic space in determining the social and connectivity predictability patterns in case of academic and office environments.

As expected, the social aspects of mobile interactions have influenced tracing data, as participants tend to interact more with users in their community or social circle. Academic and office environments are naturally grouped into social communities, in our case groups of students, faculty members and office colleagues. The Faculty of Automatic Control and Computer Science at the University POLITEHNICA of Bucharest is structured as follows: there are four years for Bachelor students split up into four groups of about 30 persons each and ten Masters directions with about 20 students each. By running the MobEmu emulator [8] with *k*-CLIQUE [18] on our tracing data, we have computed the UPB 2012 communities which are illustrated in Fig. 10(a) for AllJoyn interactions, respectively in Fig. 10(b) for Bluetooth contacts.

In computing the communities we have varied the two *k*-CLIQUE parameters, namely the contact threshold and the community threshold, as follows:

- **Contact threshold = 3,600 s, community threshold = 8**: this configuration proved to be too restrictive as we ended up ignoring interactions and even omitting nodes from the communities.
- **Contact threshold = 600 s, community threshold = 4**: as opposed to the previous configuration, the current one is placed at the other extreme being too permissive as we obtained an almost full-mesh community.
- **Contact threshold = 1,200 s, community threshold = 6**: this is the appropriate balance between the previous two configurations as can be observed in Figs. 10(a) and 10(b).

As expected, there is a high degree of connectivity considering we usually obtain one large community. This is easily explained by the spatial restraint, as almost all participants are students of the same school and therefore interact on the grounds of the university. However there is a slight difference, as interacting over Bluetooth tends to isolate stray mini-communities as can be seen in Fig. 10(b). We believe that



(a) Entropy (b) Discovered APs

Fig. 11 Distribution of entropy of interacting for Bluetooth and AllJoyn, and of number of discovered access points for each week in the UPB 2012 trace

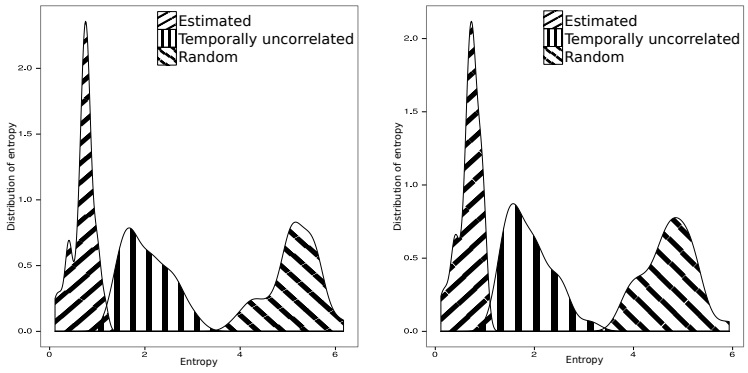
the key factor in this separation is range, as Bluetooth is designed for shorter ranges (about 5–10 meters) while WiFi APs have ranges up to 30–40 meters.

After ascertaining the social structures in our experiment, it is high time we explored the predictable behavior of participants while interacting with peers. We take into consideration both Bluetooth and AllJoyn interactions.

As such, we analyze and compare the tracing data for both types of synergy. We observe that AllJoyn interactions occur much more often than those on Bluetooth, respectively WiFi encounters cumulate up to 20,658, while Bluetooth sums up only 6,969 which amounts to only 33.73% of the latter. We believe that such results are reflected by the low range of Bluetooth which was also observed in the community analysis.

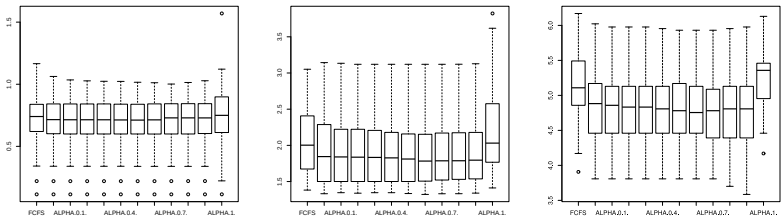
We study the hourly interactions of individuals on a daily basis and as such we compute the probability that an individual interacts at least once each day at the same hour with any other peer. Figure 11(a) shows the entropy of hourly interactions. As can be seen, AllJoyn hourly interactions peak almost as low as Bluetooth. We point out that the comparison between the two peer-to-peer solutions actually comes down to a compromise between low range versus low power saving as more powerful radios lead to much faster battery depletion. In this experiment, we choose to further analyze WiFi interactions as their exceeding rate of interactions offers higher statistical confidence.

During the UPB 2012 experiment, a total of 6,650 access points were discovered; Fig. 11(b) shows the distribution of distinct APs discovered for various weekly intervals. As can be observed, 10 weeks are sufficient for the number of discovered APs to converge and, as such, we can state that the most frequented wireless network devices have already been detected. Also noteworthy, most participants have limited mobility as they meet few access points; these restricted travel patterns favour inter-



(a) FCFS (b) Alpha 0.7

Fig. 12 The inequality of entropies for S_{rand} , S_{unc} , S_{est} for the UPB 2012 trace



(a) Estimated entropy (b) Temporally uncorrelated entropy (c) Random entropy

Fig. 13 Comparison of entropy distributions for all sequences for the UPB 2012 trace

acting, as individuals are clustered into communities situated in closed surroundings in the range of a few preferential wireless APs.

By applying the proposed methodology on the UPB 2012 trace, we obtain VL sequences with 368 symbols (8 hours x 46 weekdays), each symbol corresponding to an outstanding VL for a specific hour. Unfortunately, the lack of conscientiousness of volunteer participants has left its imprint on the tracing data, as by applying such a knowledge coefficient we trimmed down more than half of the participants.

Figure 12 illustrates the distributions of entropy P(S_{rand}), P(S_{unc}), respectively P(S_{est}) for FCFS and Alpha(0.7) and, as expected, we found that the inequality $S_{est} \leq S_{unc} \leq S_{rand}$ holds for our experiment as well.

Figure 13 illustrates a comparison of the distributions for the three proposed entropies on all VL sequences. As expected, FCFS presents one of the most skewed distributions for each of the proposed entropies; this proves that pseudo-random simulations tend to suffer from unrealistic traits. Most surprising, the cases for Alpha(1) also show that in the UPB 2012 trace signal strength was a tie-breaker for

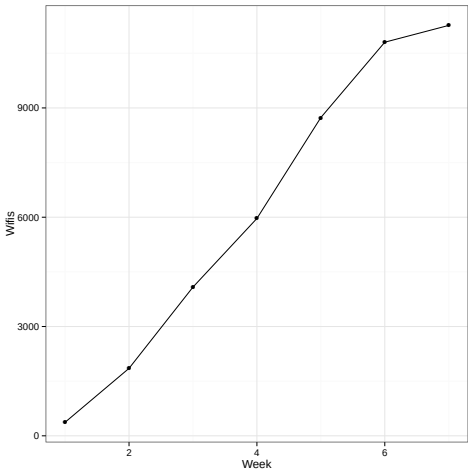


Fig. 14 Distribution of number of discovered access points for each week in the Rice trace

choosing VLs; one interpretation could be that the users involved in the experiment have a low range of mobility and travel in surroundings limited to school and offices. What can also be observed is that Alpha around 0.7 generates results close to normal distributions.

Based on the results obtained by applying the proposed methodology, we can state that the wireless behavior of users in the UPB 2012 trace is subject to predictability, as a real user can be pinpointed to one of $2^{0.68} \approx 1.6$ locations, whereas a user taking random decisions will be found in one of $2^{4.94} \approx 30.7$ locations.

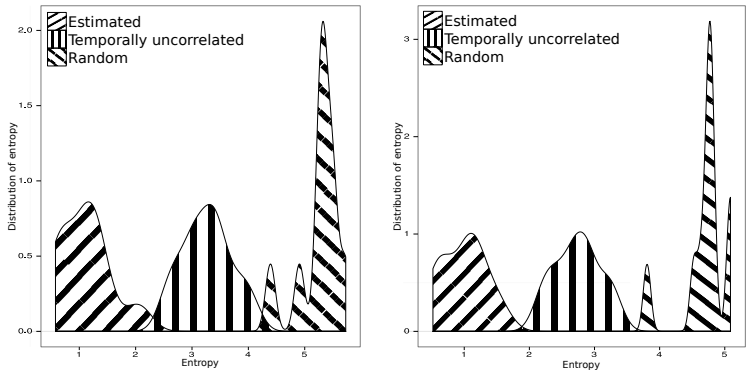
The remainder of this section presents the application of the proposed methodology on two external traces accessed from CRAWDAD, namely Rice and Nodobo. These external traces have also been studied in order to show the applicability of our guidelines and methodology.

4.3.1 Rice

The Rice[4] trace set is composed of cellular and WiFi scan results from the Rice community in Houston, Texas; 10 subjects have participated in the tracing experiment that lasted for 44 days, from 16 January 2007 to 28 February 2007. During the experiment, a total of 6,055 wireless access points have been discovered and Fig. 14 shows the distribution of discovering APs and, as can be seen, the 8 weeks are almost sufficient for convergence.

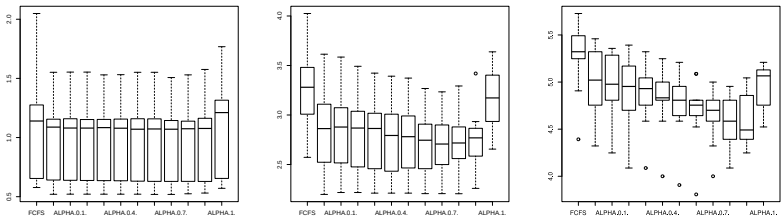
In consequence, the proposed methodology can be applied in order to study the predictability of interacting with wireless APs. The distributions of entropy $P(S_{rand})$, $P(S_{unc})$, respectively $P(S_{est})$ for FCFS and Alpha(0.7) are illustrated in Fig. 15.

[4] http://crawdad.cs.dartmouth.edu/rice/context/



(a) FCFS (b) Alpha 0.7

Fig. 15 The inequality of entropies for S_{rand} , S_{unc} , S_{est} for the Rice trace



(a) Estimated entropy (b) Temporally uncorrelated entropy (c) Random entropy

Fig. 16 Comparison of entropy distributions for all sequences for the Rice trace

Furthermore, Fig. 16 illustrates all of the distributions for the three measures of entropy on all generated sequences. As opposed to UPB 2012, the distributions of the estimated entropy are heavily skewed, but consistent; this may be a consequence of the knowledge factor. Most surprisingly, although the Rice trace set contains data from only 10 users, all of them have a high degree of collected knowledge; all users have a knowledge factor of over 60%. This increased informational gain may also affect the Random entropy as can be seen in Fig. 16(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.

As a resemblance with the previously presented UPB 2012 analysis, signal strength also is a tie-breaker in choosing the outstanding VL; as such, it seems that while tracing wireless access points the quality of an AP is more important than the number of sightings. 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.

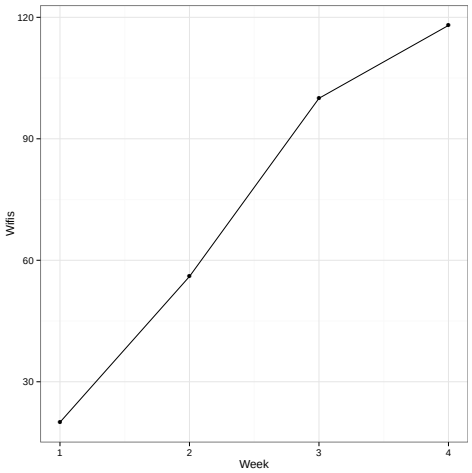


Fig. 17 Distribution of number of discovered access points for each week in the Nodobo trace

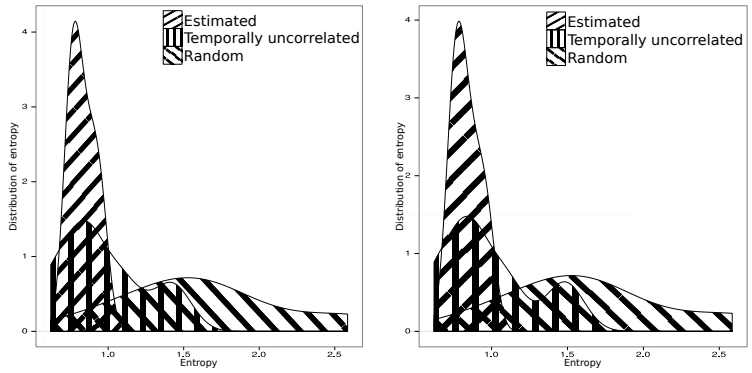
By applying the proposed methodology, we show that the users which collected the traces in the Rice experiment are subject to repeatability. Furthermore, a real user can be pinpointed to one of $2^{1.6} \approx 3.03$ locations, whereas a random user can be found in one of $2^{4.9} \approx 29.85$ locations. The estimated entropy seems quite higher than that of the UPB 2012 trace which also could be explained by the higher knowledge factor.

4.3.2 Nodobo

The Nodobo[5] trace set was collected by means of a social sensor software suite for Android devices (also dubbed Nodobo); the experiment involved 21 subjects and lasted for 23 days (it actually lasted for a longer period, but we chose this subset for it was longest contiguous interval) from 9 September 2010 to 1 November 2010.

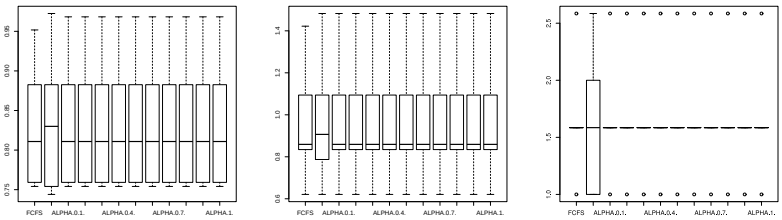
In applying the methodology, the Nodobo trace has triggered more than one warning as, not only is the tracing period insufficient, but there aren't sufficient users with a knowledge factor over 20%. As the guidelines from the methodology point out, such a trace cannot be studied for predictability; as can be seen in Fig. 17, the distribution of the discovered nodes does not converge. As a comparison with the previous two traces which accumulated up to more then 6,000 discovered wireless APs, the Nodobo trace discovered only 153. The reader should bear in mind that the low knowledge factor is not necessarily influenced by the lack of wireless APs in the vicinity of mobile users, but more by the lack of conscientiousness of the volunteers involved in the experiment.

[5] http://crawdad.cs.dartmouth.edu/strath/nodobo



(a) FCFS (b) Alpha 0.7

Fig. 18 The inequality of entropies for S_{rand} , S_{unc} , S_{est} for the Nodobo trace



(a) Estimated entropy (b) Temporally uncorrelated entropy (c) Random entropy

Fig. 19 Comparison of entropy distributions for all sequences for the Nodobo trace

To further validate the methodology and guidelines, we attempted to analyze the predictability of interacting with wireless access points in Nodobo. Figure 18 illustrates the distributions of entropy P(S_{rand}), P(S_{unc}), respectively P(S_{est}) for FCFS and Alpha(0.7); we can state that the goal inequality of predictability does not hold. Furthermore, when attempting to compare the distributions for all sequences (as illustrated in Fig. 19), the insufficiency of both the tracing interval and of the sample size impact heavily on the statistical analysis, the results being inconclusive.

The proposed methodology applied on Nodobo has shown that this trace set is insufficient to determine any measure of predictability of wireless behavior. The guidelines prove to be efficient in filtering the trace set before the statistical analysis is performed in positive cases (like UPB 2012 and Rice), but also in negative cases (i.e. Nodobo).

We have also shown in this section that, after eliminating the traces that don't have sufficient information, we are left with data that confirms Song's inequation $S_{est} \leq S_{unc} \leq S_{rand}$.

5 Conclusions and Future Work

In this chapter, we presented techniques and applications to analyze mobility data. We began by highlighting existing projects and frameworks that have gathered mobility traces, and then we described two tracing applications that we implemented ourselves and the resulting mobility experiments performed at the University POLITEHNICA of Bucharest in 2011 and 2012. Afterwards, we extracted the benefits of performing such experiments rather than using mathematical mobility models, as well as several limitations and challenges brought by this approach.

We then presented ways in which mobility data can be analyzed in terms of contact distribution, as well as the predictability of encounters and interactions with access points. We showed that the future behavior of a node in an opportunistic network in terms of the number of contacts in the next time interval can be approximated as a Poisson distribution, with high levels of predictability. This happens because contacts in an academic environment are highly regular, since the participants have fixed daily schedules.

We also analyzed the repeatability and predictability of access interactions in academic and office environments based on two separate points of view: the group view and the individual view. Furthermore, we applied a distributed community detection algorithm and found that confined surroundings lead to the creation of large highly-adhesive communities. However, the wireless communication media can have an important influence as low ranged solutions, such as Bluetooth, tend to isolate loosely-coupled micro-communities. As for the individual's perspective over interactions, we focused more on AllJoyn since Bluetooth interactions occurred three times more rarely than the latter. Furthermore, we proposed a methodology and a set of guidelines to be used in analyzing the predictability of interaction between mobile users and wireless access points based on the study of Song et al. [33]. By applying the methodology on three cases (UPB 2012, Rice and Nodobo), we proved that mobile users have a predictable wireless behaviour if the trace sets are complete, correct and sufficiently informed.

For future work, we plan to explore availability and usage patterns and further correlate them with the current mobility and interaction patterns. By doing so, we will be able to build an accurate detector for smart mobile collaborations based on machine learning techniques trained with the tracing data. We believe that studying the predictability of human behavior based on real mobile user traces can prove to be the key to intelligent mobile collaboration in opportunistic networks comprised of smartphones, that will eventually lead to less power consumption and which will be able to harness the full potential of contextual data by distributed context aggregation and detection.

References

1. Stuart M. Allen, Marco Conti, Jon Crowcroft, Robin Dunbar, Pietrò P. Lió, José Fernando Mendes, Refik Molva, Andrea Passarella, Ioannis Stavrakakis, and Roger M. Whitaker. Social Networking for Pervasive Adaptation. In *Self-Adaptive and Self-Organizing Systems Workshops, 2008. SASOW 2008. Second IEEE International Conference on*, pages 49 –54, oct. 2008.
2. Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, May 2005.
3. Tracy Camp, Jeff Boleng, and Vanessa Davies. A Survey of Mobility Models for Ad Hoc Network Research. *Wireless Communications & Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications*, 2(5):483–502, 2002.
4. Augustin Chaintreau and Pan Hui. Pocket Switched Networks: Real-world mobility and its consequences for opportunistic forwarding. Technical report, 2006 Computer Laboratory, University of Cambridge, February 2005.
5. Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, June 2007.
6. Radu-Ioan Ciobanu and Ciprian Dobre. Predicting Encounters in Opportunistic Networks. In *Proceedings of the 1st ACM workshop on High performance mobile opportunistic systems*, HP-MOSys '12, pages 9–14, New York, NY, USA, 2012. ACM.
7. Radu-Ioan Ciobanu, Ciprian Dobre, and Valentin Cristea. Social Aspects to Support Opportunistic Networks in an Academic Environment. In Xiang-Yang Li, Symeon Papavassiliou, and Stefan Ruehrup, editors, *Ad-hoc, Mobile, and Wireless Networks*, volume 7363 of *Lecture Notes in Computer Science*, chapter 6, pages 69–82. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2012.
8. Radu-Ioan Ciobanu, Ciprian Dobre, Valentin Cristea, and Dhiya Al-Jumeily. Social Aspects for Opportunistic Communication. In *Parallel and Distributed Computing (ISPDC), 2012 11th International Symposium on*, pages 251 –258, june 2012.
9. Zoltan Dezsö, Eivind Almaas, András Lukács, Balázs Rácz, István Szakadát, and Albert-László Barabási. Dynamics of information access on the web. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(6):066132+, 2006.
10. Arta Doci, Leonard Barolli, and Fatos Xhafa. Recent Advances on the Simulation Models for Ad Hoc Networks: Real Traffic and Mobility Models. *Scalable Computing: Practice and Experience*, 10(1), 2009.
11. Arta Doci, William Springer, and Fatos Xhafa. Impact of the Dynamic Membership in the Connectivity Graph of the Wireless Ad hoc Networks. *Scalable Computing: Practice and Experience*, 10(1), 2009.
12. Erina Ferro and Francesco Potorti. Bluetooth and Wi-Fi wireless protocols: a survey and a comparison. *Wireless Communications, IEEE*, 12(1):12–26, Feb.
13. Peter Gerl. Random Walks on Graphs. In Herbert Heyer, editor, *Probability Measures on Groups VIII*, volume 1210 of *Lecture Notes in Mathematics*, pages 285–303. Springer Berlin Heidelberg, 1986.
14. Bruno Gonçalves and José Javier Ramasco. Human dynamics revealed through Web analytics. *CoRR*, abs/0803.4018, 2008.
15. Bruno Gonçalves and José Javier Ramasco. Towards the Characterization of Individual Users through Web Analytics. In Jie Zhou, editor, *Complex (2)*, volume 5 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 2247–2254. Springer, 2009.
16. Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. Pocket Switched Networks and Human Mobility in Conference Environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, WDTN '05, pages 244–251, New York, NY, USA, 2005. ACM.

17. Pan Hui, Jon Crowcroft, and Eiko Yoneki. BUBBLE Rap: Social-Based Forwarding in Delay-Tolerant Networks. *Mobile Computing, IEEE Transactions on*, 10(11):1576–1589, Nov.
18. Pan Hui, Eiko Yoneki, Shu-Yan Chan, and Jon Crowcroft. Distributed Community Detection in Delay Tolerant Networks. In *Proceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture*, MobiArch '07, pages 1–8, New York, NY, USA, 2007. ACM.
19. Karin Anna Hummel and Andrea Hess. Movement activity estimation for opportunistic networking based on urban mobility traces. In *Wireless Days (WD), 2010 IFIP*, pages 1–5, Oct. 2010.
20. Qualcomm Innovation Center Inc. Introduction to AllJoyn. HT80-BA013-1 Rev. B, 2011.
21. David B. Johnson and David A. Maltz. Dynamic Source Routing in Ad Hoc Wireless Networks. In Imielinski and Korth, editors, *Mobile Computing*, volume 353. Kluwer Academic Publishers, 1996.
22. Minkyong Kim, David Kotz, and Songkuk Kim. Extracting a Mobility Model from Real User Traces. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, pages 1–13, 2006.
23. Ioannis Kontoyiannis, Paul H. Algoet, Yuri M. Suhov, and Abraham J. Wyner. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text. *Information Theory, IEEE Transactions on*, 44(3):1319–1327, May 1998.
24. Radu-Corneliu Marin, Ciprian Dobre, and Fatos Xhafa. Exploring Predictability in Mobile Interaction. In *Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on*, pages 133–139. IEEE, 2012.
25. Radu-Corneliu Marin, Ciprian Dobre, and Fatos Xhafa. A Methodology for Assessing the Predictable Behaviour of Mobile Users in Wireless Networks. Submmited at INCoS-2012 Special Issue "Concurrency and Control" Wiley, 2013.
26. Mirco Musolesi and Cecilia Mascolo. Mobility Models for Systems Evaluation. In Benoît Garbinato, Hugo Miranda, and Luís Rodrigues, editors, *Middleware for Network Eccentric and Mobile Applications*, chapter 3, pages 43–62. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
27. Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A Tale of Many Cities: Universal Patterns in Human Urban Mobility, October 2011.
28. João Gama Oliveira and Albert-László Barabási. Human dynamics: Darwin and Einstein correspondence patterns. *Nature*, 437(7063):1251, October 2005.
29. Jukka-Pekka Onnela, Jari Saramäki, Jari Hyvönen, Gábor Szabó, Marcio Argollo de Menezes, Kimmo Kaski, Albert-László Barabási, and János Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007.
30. Jukka-Pekka Onnela, Jari Saramäki, Jari Hyvönen, Gábor Szabó, David Lazer, Kimmo Kaski, János Kertész, and Albert-László Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.
31. Mikko Pitkänen, Teemu Kärkkäinen, Jörg Ott, Marco Conti, Andrea Passarella, Silvia Giordano, Daniele Puccinelli, Franck Legendre, Sacha Trifunovic, Karin Hummel, Martin May, Nidhi Hegde, and Thrasyvoulos Spyropoulos. SCAMPI: Service Platform for Social Aware Mobile and Pervasive Computing. *SIGCOMM Comput. Commun. Rev.*, 42(4):503–508, September 2012.
32. César A. Hidalgo R. Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems. *Physica A: Statistical Mechanics and its Applications*, 369(2):877–883, September 2006.
33. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010.
34. Alan Stuart, Keith Ord, and Steven Arnold. *Kendall's Advanced Theory of Statistics, Classical Inference and the Linear Model*, volume Volume 2A (2007 reprint)). Wiley, sixth edition, 1999.

35. Jing Su, James Scott, Pan Hui, Jon Crowcroft, Eyal De Lara, Christophe Diot, Ashvin Goel, Meng How Lim, and Eben Upton. Haggle: Seamless Networking for Mobile Applications. In *Proceedings of the 9th International Conference on Ubiquitous Computing*, UbiComp '07, pages 391–408, Berlin, Heidelberg, 2007. Springer-Verlag.
36. Alexei Vázquez. Exact Results for the Barabási Model of Human Dynamics. *Physical Review Letters*, 95(24):248701+, December 2005.
37. Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127+, March 2006.
38. Jacob Ziv and Abraham Lempel. Compression of Individual Sequences Via Variable-Rate Coding. *Information Theory, IEEE Transactions on*, 24(5):530–536, September 1978.

Index

Glossary of Terms and Accronyms

ACT	Any-Contact Time
AP	Access Point
BSSID	Basic Service Set Identifier
CT	Contact Time
FCFS	First Come First Served
IACT	Inter-Any-Contact Time
ICT	Inter-Contact Time
ON	Opportunistic Network
SSID	Service Set Identifier
VL	Virtual Location