Recurring contact opportunities within groups of devices

Nuno Cruz[*,+]
ncruz@deetc.isel.ipl.pt
[*]Instituto Superior de Engenharia de Lisboa
Instituto Politécnico de Lisboa
ADEETC

Hugo Miranda[+]
hamiranda@ciencias.ulisboa.pt
[+]Faculdade de Ciências
Universidade de Lisboa
LaSIGE

ABSTRACT

The capability to anticipate a contact with another device can greatly improve the performance and user satisfaction not only of mobile social network applications but of any other relying on some form of data harvesting or hoarding. One of the most promising approaches for contact prediction is to extrapolate from past experiences. This paper investigates the recurring contact patterns observed between groups of devices using an 8-year dataset of wireless access logs produced by more than 70000 devices. This effort permitted to model the probabilities of occurrence of a contact at a predefined date between groups of devices using a power law distribution that varies according to neighbourhood size and recurrence period.

In the general case, the model can be used by applications that need to disseminate large datasets by groups of devices. As an example, the paper presents and evaluates an algorithm that provides daily contact predictions, based on the history of past pairwise contacts and their duration.

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Model Development; C.2.1 [**Network Architecture and Design**]: Wireless Communication

Keywords

Mobility, Wireless, Communities, Temporal Communities, Contact Prediction

General Terms

Measurement, Performance, Algorithms

1. INTRODUCTION

The knowledge on human mobility is used on pervasive computing environments to model applications [18] and routing protocols [12], to harvest computing resources [6] or provide network connectivity [20] to a group of mobile devices. Groups of devices facilitate for example the creation of distributed data stores [16], message passing in delay tolerant networks [20] and leverage middleware to efficiently find useful devices for resource sharing [6].

Contact patterns are usually estimated from observations of multiple metrics on a population of individuals by applying a statistical fitting on the data. The goal is to find a distribution that provides a good approximation to the metrics of interest and that can be evaluated in run-time. One of the most frequently cited metrics is the inter-contact time (ICT), which represents the time interval between two consecutive contacts of the same two peers. ICT is used in multiple applications and modelled by several mobility models using a power law distribution [2, 13, 15].

However, considering only ICTs limits optimization strategies for some classes of applications. A new metric is needed to effectively model the neighbourhood size and the recurrence of contacts of group members. Such a metric would pave the way to optimize applications that require cooperation of multiple devices, for example to distribute a large dataset while minimizing data redundancy and increasing data availability.

Unfortunately, the design of such a metric is severely constrained by the large amounts of mobility data required to give statistical relevance to any modelling effort. This paper gives a step in this direction by analysing a dataset of the eduroam wireless network site on the Polytechnic Institute of Lisbon, originally presented in [9]. The dataset contains all the records produced between 2005 and 2013 by the 76479 devices that accessed at least one of the network's 239 access points (APs).

The paper evaluates the dimension and the patterns of repetition of meetings between groups of devices of any size and presents statistical distributions that can be used to model the group contact probabilities. It shows that the extracted statistical distribution fits a *Pareto* distribution for most of the data, with different parameters according to neighbour size and recurrence period.

The paper also presents a ranking algorithm that uses the knowledge obtained to predict future contacts between pairs of devices. We applied the algorithm to a mobility scenario extracted from the same data but with a different methodology and year, and also to a trace of GPS positions of Taxis

MOBIQUITOUS 2015, July 22-24, Coimbra, Portugal

DOI 10.4108/eai.22-7-2015.2260048

in Rome, and found that the different environments share the same statistical properties, allowing the ranking algorithm to improve the odds of knowing which devices will be in range in a future day.

The paper is organised as follows. Section 2 makes a brief survey of the related work and discusses possible applications of this effort. The characterisation of the dataset and the methodology used for extracting and analysing the data is presented in Sec. 3 and 4. Section 5 addresses our efforts in modelling contacts recurrence into statistical distributions. Experiments on predicting contacts using data from the previous sections is detailed in Sec. 6. The conclusions and the directions of the future work are the focus of Sec. 7.

2. RELATED WORK

Applications of research on human mobility for mobile computing have been mostly evolving around the opportunities for data dissemination and opportunistic routing. Haggle [23] is a good example of a project that addressed applications for an opportunistic environment supported on the study of human mobility. The Haggle project characterized human mobility on two dimensions: inter-contact time (ICT) and contact duration and showed that ICTs tend to follow a Power Law with a Exponential Decay, something also supported by other studies (e.g. [14]). Bubble Rap [12] is a socially influenced routing protocol that leveraged on the mobility traces of the Haggle project to infer communities using K-Clique [19] and weighted network analysis algorithms [17].

The work described in [4] studies ICTs in two distinct datasets. One is based on records produced by an external observer. In particular, data collected from the access logs of WiFi networks. The second, named direct contact, contains records captured directly by the devices. These are either produced by devices designed specifically to be carried by users or by exploiting the Bluetooth connectivity of mobile devices. Authors observed that, in contrast with previous works, the distribution of inter-contact times follows a power law but only until 1 day of duration. As a follow up, the paper shows how this result impacts current forwarding algorithms and makes suggestions for improvements.

In contrast with the proposals above, which follow the limiting approach of using ICTs as the preferred metric, [21] addresses temporal communities and their relations. Authors extracted temporal communities from four distinct datasets, the largest of which considering the observation of 97 nodes over 9 months. In spite of the small scale and duration of the study, authors presented two interesting conclusions. On one side, that the establishment of social communities has direct implications on temporal communities. On the other, authors identified one particular class of devices, those with a high contact rate that are rarely seen in temporal communities, and show that they contribute significantly for the efficient content dissemination in opportunistic social networks. Social communities are equally the focus of SocialCast [7], which exploits the knowledge that humans tend to share interests and locations to develop an efficient routing protocol for publish-subscribe on Delay-Tolerant Networks. The authors use Kalman filters for forecasting future contacts, based on previous observations of being co-located with a subscriber. SocialCast was one of the firsts protocols supporting one-to-many communication for the Haggle framework.

Figure 1: Location of IPL sites

An innovative approach for detecting communities is presented in [18]. Authors added a duration variable to communities detection, thus creating spatio-temporal communities. The community relevance is increased proportionally to its duration. It was shown that spatio-temporal communities can contribute to improve the efficiency of information dissemination in opportunistic networks. Simulation experiments were conducted in the same datasets used in Bubble Rap.

PreKR [11] is a framework that improves the forwarding on opportunistic networks by using a kernel regression based estimation for link pattern prediction. Using historical observations of network maps on three datasets, one of which being Bubble Rap, PreKR determines what is the probability of a recurrence of a link between two devices. Authors show that PreKR outperforms all other prediction methods, including Prophet. The distinguishing factor was the use of kernel regression, that allowed PreKR to achieve an accuracy of more than 90%.

3. METHODOLOGY

The dataset used in this study aggregates the log records produced by all Access Points (APs) of the eduroam Wi-Fi network of the Lisbon Polytechnic Institute (IPL) generated between January 1, 2005 and December 31, 2013.

IPL is the 7^{th} largest high education institution in Portugal with approximately 1300 teachers and 15000 students, distributed by 10 distinct sites around the Lisbon metropolitan area (see Fig. 1). The Eduroam Wi-Fi network results from an international effort aiming to transparently provide wireless Internet connectivity to its members in the campus of all adhering institutions. The IPL's site of the eduroam network is supported by approximately 200 Cisco Systems APs, covering a total of 26 buildings and inter-building areas. Records are originated from all the users accessing the network, including visitors.

Figure 2 shows a continuous growth of the number of users and devices although at distinct rates, specially since 2010. This is coincidental with an increase in the sales of smart-

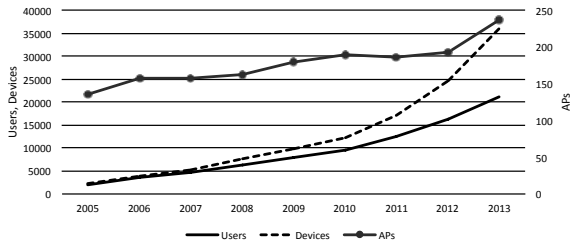


Figure 2: Evolution of devices, users and access points

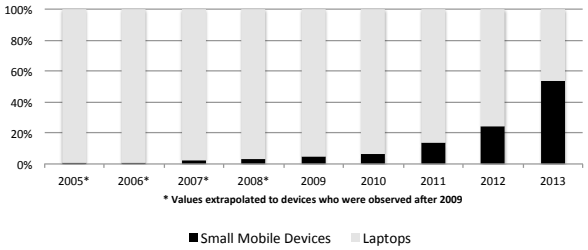


Figure 3: Proportion of Laptops to Small Mobile Devices observed

phones observed at the national level and suggests that the number of users accessing the network with more than one device has been increasing. Figure 3 compares the proportion of Small Mobile Devices and Laptops accessing the network in each year, determined from the *vendor*, *parameter request list* and *hostname* fields of the DHCP messages exchanged between the devices and the supporting infrastructure. The distinction is relevant mostly due to the observation that small mobile devices tend to reproduce more accurately the users' movement patterns [9].

Contacts data evaluated in this paper are extracted from the RADIUS protocol [22] session logs, which considers the association of each user to a single AP. Records contain the device MAC address, AP id, user name, session start and stop dates. Prior to the analysis, logs have been purged from inconsistencies that can be attributed to problems with the wireless network card drivers:

- Consecutive sessions between the same device and AP with an interval of less than 5 seconds have been merged in a single session;
- Overlapping sessions S1 and S2 of the same device to distinct APs have been serialized by setting the stop time of S1 to occur at the moment immediately before the start time of S2. Given that network cards cannot be concurrently associated to more than one AP, this impossibility can only be explained if the device did not disassociate correctly from one AP before associating to the next with the former artificially establishing the session stop time by timeout;
- Sessions with the same start and stop time were removed. Sessions with these characteristics are created when a user has some issue while connecting to the network, although the network considers the user authenticated (thus creating the RADIUS record).

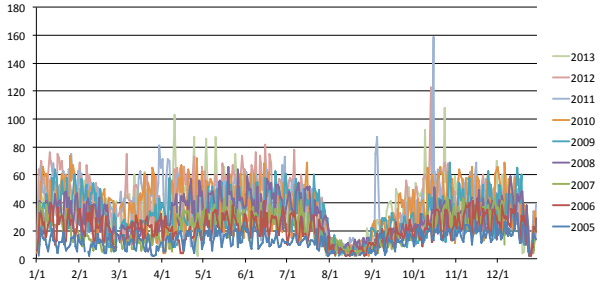


Figure 4: Max Community Size Per Day

The interested reader is referred to [9] for a more in depth analysis of the dataset and to [8] for its applications on the generation of mobility scenarios.

3.1 Temporal Communities

In this paper, a temporal community (TC) is defined as the set of devices connected simultaneously to the same AP. A TC exists as long as its membership does not change. The addition and/or removal of any member results in the creation of a new temporal community. The approach is oblivious to associations of devices to distinct APs with overlapping coverage and to the repetition of TCs. TCs with the same membership are considered distinct if: *i*) they occur in a distinct AP; or *ii*) there is some interval between the two occurrences where an exactly equal TC did not exist. Table 1 summarizes the TCs counted using this approach.

Each TC with size n implicitly defines $\sum_{i=1}^{n} \binom{n}{i}$ *Temporal Sub-Communities* (TSCs), that result from the combinations of the members of the TC. It should be noted that TSCs include the special case where all the members of the TC are represented (i.e., the TC itself). Relevant for this paper is the evaluation of the repetitive occurrence of groups of devices, independently of its members being or not part of larger groups. Therefore, the paper will focus mostly on the study of TSCs.

A multi-year analysis shows a non-negligible variation of the number and size of the communities. Part of this variation can be attributed to the addition of Access Points (APs) to the network (cf. Fig 2), mostly motivated by the need to resolve localised network performance issues. Such addition contributes to a decrease in the dimension of temporal communities as devices have more alternative APs for association on the most frequently accessed locations.

Figure 4, which depicts the size of the biggest TC observed each day, clearly shows the impact of the academic environment on the network. In the figure it is possible to observe the reduced activity during the Winter (end of December), Summer (August) and Easter (March) breaks. The irregularity of the plots can also be attributed to weekends and to the organization of conferences.

3.2 Temporal Patterns

In addition to presenting the number and size of all TSCs, the paper evaluates the probability of recurrence of two and three consecutive hits of the same TSCs on 4 distinct temporal patterns. The **Consecutive Days (CD)** and the

Table 1: Temporal communities observed in the dataset

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Max. TC Size	33	43	44	66	69	74	159	121	108
TCs	370245	1113630	1309098	1682684	1700471	1935039	2775835	5633825	11366159
Average TC Size	6.45	8.16	8.32	9.37	9.96	11.08	10.77	10.08	11.5

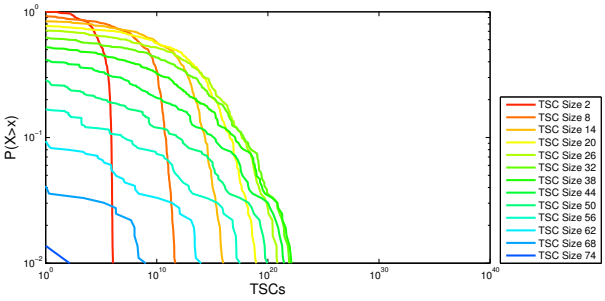


Figure 5: TSCs per day

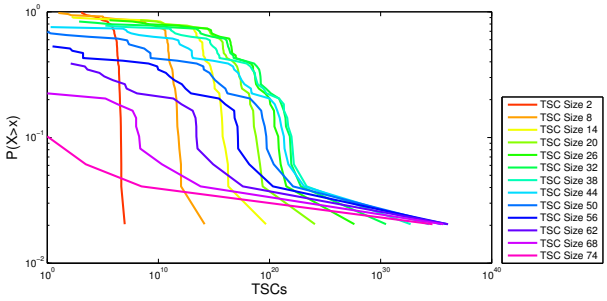


Figure 6: TSCs per week

Consecutive Week Day (CWD) patterns use intervals of respectively 1 and 7 days. These patterns serve to investigate repetitions inspired by common student activities, like the daily attendance to school and the weekly attendance to classes.

The **Consecutive Month Day (CMD)** and **Consecutive Week (CW)** use more irregular patterns. CMD seeks for repetitions in the same day number of consecutive months. CW in turns seeks repetitions in any weekday of consecutive weeks.

For clarity, and as an example, consider the observation of a TSC on July 18th, 2012 (Wed). A hit will be found if the same TSC is observed on July 19th for CD, July 25th for CWD, August 18th for CMD and on any day between the 22nd and the 28th of July (Sun-Sat) for CW.

These temporal patterns are negatively affected by calendar irregularities. No attempt to attenuate the effects of public holidays, weekends or school breaks has been made. This option was chosen to approximate the results from those found by some application using past experiences to estimate the probability of contact repetition.

4. GENERIC DATA ANALYSIS

The accumulated number of Temporal Sub-Communities (TSCs) found in every day of 2012 is depicted in Fig. 5 as a Complementary Cumulative Distribution Function (CCDF). For clarity, the figure presents TSC sizes in steps of 6. It was observed that the lines of the TSC sizes that were omitted evolve similarly to those that are represented.

A first surprising effect observed in Fig. 5 is the peak of the number of TSCs at size 38. However, this is due to the methodology used for determining TSCs. Recall from Sec. 3, that the paper considers all possible combinations of elements of any TC as TSCs. In this case, each of the observed TCs of size 74 produces, by itself, $\binom{74}{38} \approx 1.7 \times 10^{21}$ TSCs with size 38. Still, the figure is illustrative of the potential number and size of the groups of devices within transmission range that can be found in academia. Notice for example that in 5% or more of the days of 2012 it is possible to find at least 10^{10} communities of size 62 and that 80% of the days had more than 100 TSCs with 14 devices.

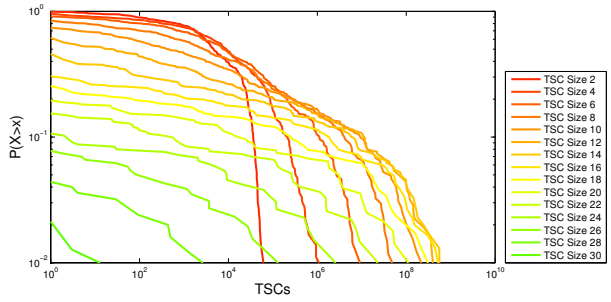
Figure 6 revisits these results after grouping the TSCs in weeks, what removes the spurious effects of occasional TCs of very large size. Still, the plot denotes an interesting regularity, suggesting that TSCs of sizes up to 23 tend to occur in a large proportion of more than 90% of the weeks.

5. RECURRENCE OF CONTACTS

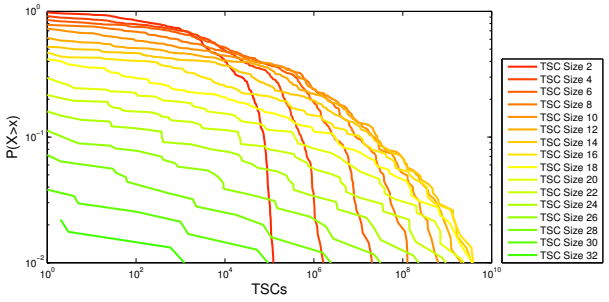
Consider some TSC t observed in some instant i. The study on t's recurrence will proceed in two steps. First, it will measure the frequency with which t is observed a second time, respecting one of the temporal patterns defined in Sec 3. I.e., we will look for occurrences of t in instant i', knowing that the relationship between i and i' must necessarily respect one of the temporal patterns. The second step will evaluate the persistence of these occurrences. It estimates the probability of observing t in a third instant i'', with the time interval between i' and i'' respecting the same temporal pattern that was found between i and i'. The analysis will focus in 2012, which presents a good trade-off between the manageability of the size of the dataset and its recency.

Figure 7 depicts the CCDF of the TSCs that were observed a second time satisfying each of the Temporal Patterns defined above. The figure clearly demonstrates that the selection of the temporal pattern has a strong impact on the results. The Consecutive Month Days (CMD) is the temporal pattern that performs poorly. This should be expected as it is hard to find any routines depending on the day of the month in the academic environment. In contrast, the CD and CWD temporal patterns, which reflect better the typical student schedule, perform reasonably well, specially for TSCs of size of 6 or less. In these cases, more than 90% of the days presented 100 or more TSCs which were equally observed in the previous instant of the temporal pattern.

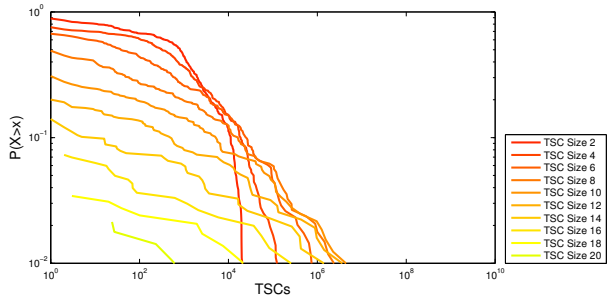
The best results are presented by the CW temporal pattern,



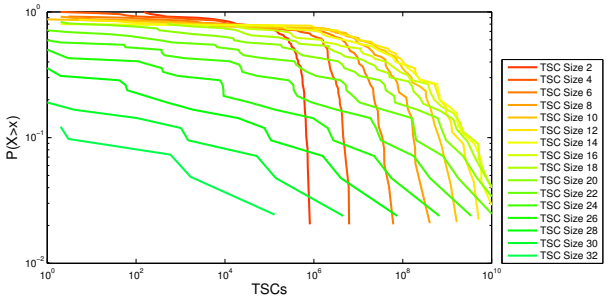
(a) TSCs on two consecutive days (CDs)



(b) TSCs on two consecutive week days (CWDs)



(c) TSCs on two consecutive month days (CMDs)



(d) TSCs on any day of two consecutive weeks (CWs)

Figure 7: Temporal patterns for two consecutive periods

where it was not hard to find 10000 communities of sizes 6 or less in 90% of the days. This is not a surprising result considering the great flexibility of the constraints imposed by CW in comparison with CD and CWD.

5.1 Predictability of Multiple Contacts

Fig. 8 shows the average and standard deviation of the proportion of TSCs that repeated in a third consecutive instant from those that were observed twice. Results contribute to decrease the relevance of the observations of large TSCs in two consecutive periods. Although it is frequent to find large TSCs, their membership tends to vary with time. Therefore, the occurrence of large TSCs can only be used by applications depending on ad hoc concentrations of users.

Concerning TSCs with small number of members, the results permit to separate the CD and CWD temporal patterns, with the daily one showing to be more predictable. CD and the more relaxed CW temporal patterns are the only able to obtain probabilities of repetition above 10% for TSCs of up to 4 members and to show a surprisingly 30% for TSCs of size 2. This can be considered as a non-negligible probability of finding the same device on, respectively, the next two days and weeks. CW outperforms CD in the stability of the predictions, as it presents a smaller standard variation of the sample.

Figure 9 depicts these results using CCDFs of the absolute number of occurrences observed. As suggested by Fig. 8, TSCs with significant results are those with small membership sizes. It is also interesting to notice the distinct pattern exhibited by each temporal pattern. Still, it should be noticed that it is not hard to find a considerable number of TSCs satisfying the CD and CW criteria for 3 consecutive intervals. In the case of CW, in 90% of the days of 2012 it is

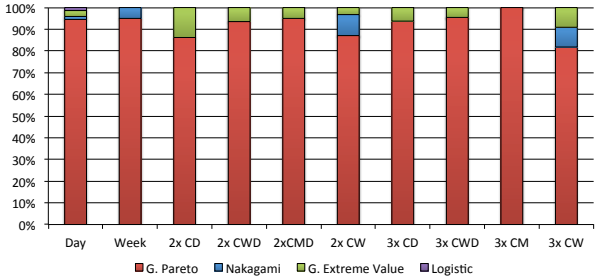


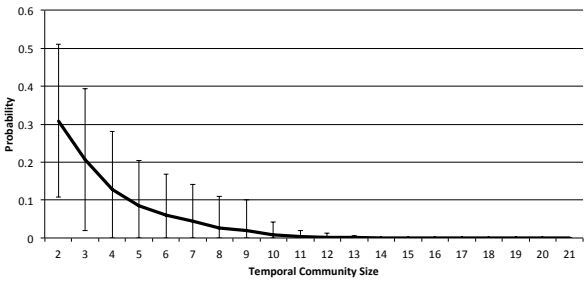
Figure 10: Distribution Fitting

possible to find 100 TSCs of 6 members that were observed on 3 consecutive weeks. In line with what was observed before, the CWD and CMD temporal patterns show disappointing results, in both the number of TSCs found and on the probability of their recurrence.

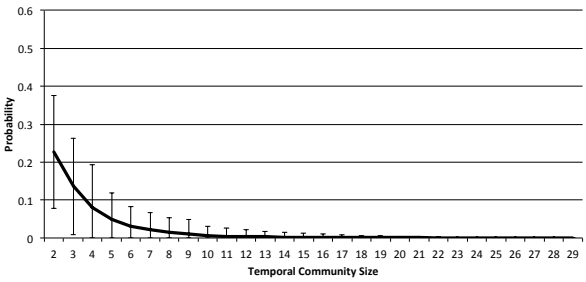
5.2 Probabilistic Model

To model the temporal patterns observed in TSCs, the results discussed in Sec. 5 were fitted to statistical distributions using the Akaike information criterion on Matlab. Figure 10, which aggregates TSC sizes by distributions, shows that the Generalized Pareto distribution is the most adequate to model the behaviours observed in the paper. The use of the Pareto distribution is consistent with results found for modelling other aspects of human mobility, for example those detailed in [5, 10, 14].

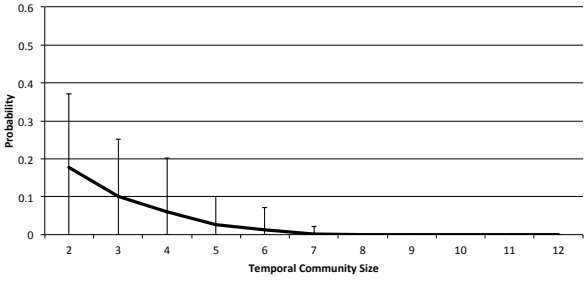
Interestingly, the CD, CWD and CMD temporal patterns exhibit an exception where only a single size of the TSCs is better represented by Generalized Extreme Value. As shown in Table 2, the sizes of the TSCs with an abnormal behaviour are distinct for each temporal pattern and no re-



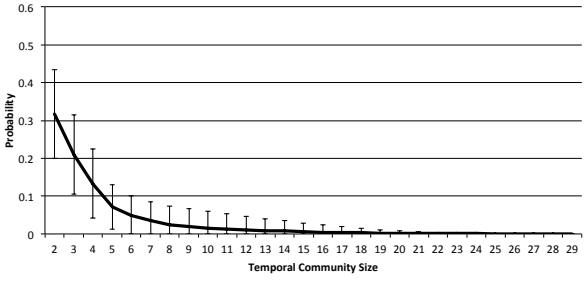
(a) Probability of finding the same TSC on three CDs



(b) Probability of finding the same TSC on three CWDs



(c) Probability of finding the same TSC on three CMDs



(d) Probability of finding the same TSC on three CWs

Figure 8: Probabilities for three consecutive periods

Table 2: TSC size with exceptional distribution

Temporal Pattern	TSC Size
CD	6
CWD	3
CMD	2

Table 3: Probability Function Parameters

Temporal Pattern (tp)	Probability a	Probability b	Std. dev. a	Std. dev. b
CD	0.7167	-0.4204	0.3335	-0.205
CWD	0.6081	-0.4971	0.2375	-0.2216
CMD	0.5947	-0.601	0.4295	-0.357
CW	0.7701	-0.4431	0.1499	-0.1391

lation between the values could be found. Therefore, these cases are considered as an anomaly and the reminder of the text handles them indifferently from the remaining.

A practical application of these results can be obtained by reproducing the approach discussed in Sec. 5.1 to create a fitted function that returns the probability of occurrence of a group of size n. Results are approximated by a function PC given by Eq. 1. Values for constants a and b depend of the temporal pattern and are given in Tab. 3.

$$PC_{tp}(n) = a_{tp} \times e^{b_{tp}n} \quad (1)$$

Standard deviation can be approximated by a function $SC_{tp}(n) = a_{tp} \times e^{b_{tp}n}$, which is similar to function PC although with a distinct set of constants a and b, equally depicted in Tab. 3.

6. CONTACT PREDICTION ALGORITHM

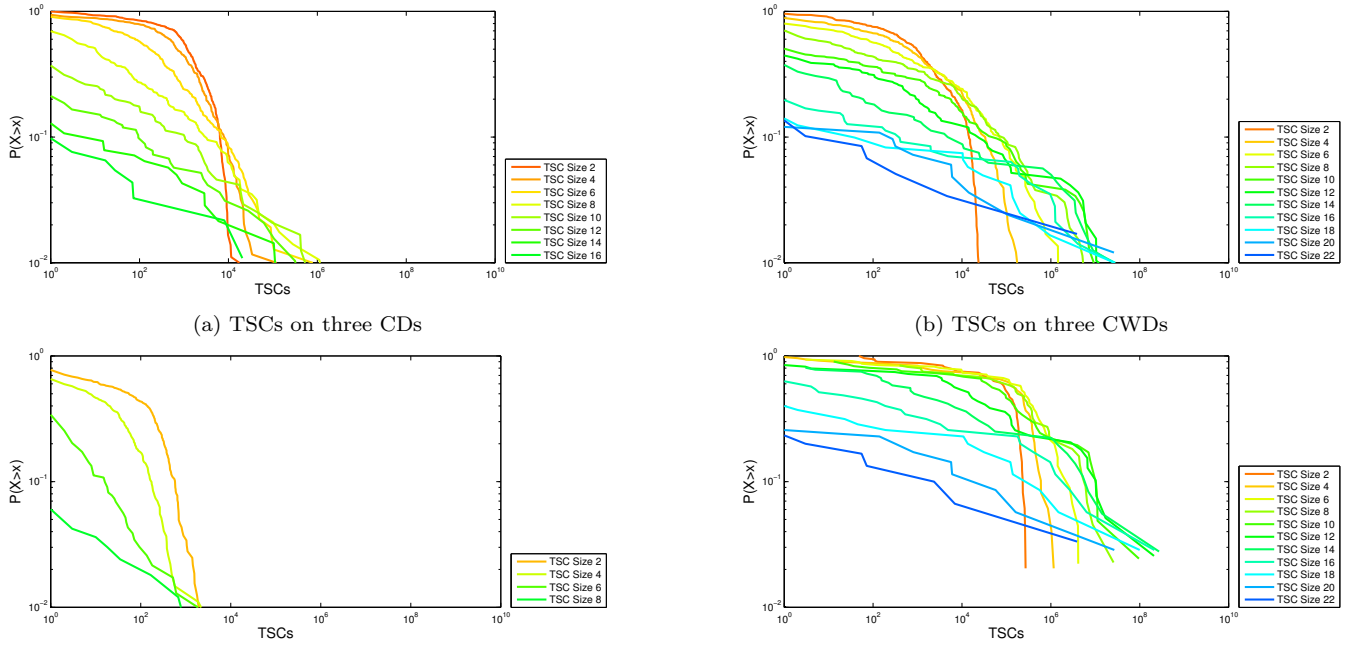
The capability to anticipate user contacts is valuable to a multitude of applications, which can be arranged according to the observer, in 3 distinct categories. In the *omniscient observer* category, a centralised server has access to the list of all contacts that have occurred in the past. An example is a reputation server that combines the contacts of all the service members to anticipate those that will occur in the future. The omniscient observer perspective is the one supporting the theoretical analysis of the previous section. In the *localised server* category, some external observer, for example an access point, creates a local perspective, that results from his limited observation point. Finally, in the *peer view* category, each device anticipates future contacts exclusively from those where he has participated in the past. The peer view is the one where the information is more limited and therefore, where predictions are more challenging. This section reports on our efforts to create an algorithm capable to anticipate future contacts with a reasonable accuracy in the peer view perspective.

The algorithm was designed to integrate with any application. It accepts a target date and a list of the contacts observed in the past. Each contact is tagged with the date and duration of the contact and the peer ID. The algorithm outputs a list of peers, ordered by the likelihood of finding it on the target day. Members of the output list are all the peers from the input contact list that satisfy any of the CD, CWD or CMD temporal patterns (described in Sec. 3.2) for the two previous consecutive instances and which, if observed on the target date, will result in a third consecutive occurrence of the pattern.

Peers are ranked according to the scoring function *score*, depicted in Eq. 2.

$$score = f_{CD}(d_{CD}) + f_{CWD}(d_{CWD}) + f_{CMD}(d_{CMD}) \quad (2)$$

where d_{CD}, d_{CWD}, d_{CMD} are the duration (in seconds) of the contact between the two nodes in the last event of the



(a) TSCs on three CDs

(b) TSCs on three CWDs

(c) TSCs on three CMDs

(d) TSCs on any day of three CWs

Figure 9: Temporal patterns for three consecutive periods

corresponding pattern. The *score* for each node is therefore dictated by an accumulation of partials from each temporal pattern where it was observed and given by:

$$f_{tp}(d) = w_{tp} \times PC_{tp}(2) \times CD_{tp}(d), \forall tp \in \{CD, CWD, CMD\} \quad (3)$$

where w_{tp} is the weight attributed to the temporal pattern, PC_{tp} is the probability function presented in Eq. 1.

The role of the CD_{tp} function is to convert the duration of the last contact in a weight. The function privileges longer contacts, mapping them on proportionally heavier weights. The two classes of functions experimented are depicted in Eq. 4 and 5. Both assume that the contact duration is bounded between $60s$ and $86400s$ (one day), considered to be the interval representing social interactions between peers.

$$CD(d) = \frac{kd}{86400 + (k-1)d}, k \geq 1, d \geq 60 \quad (4)$$

$$CD(d) = \left(\frac{d}{86400}\right)^2, d \geq 60 \quad (5)$$

Functions differentiate by the direction of their curve. The family of functions of Eq. 4 increases the weight linearly with the duration when $k = 1$, increasing the weight of shorter contacts as k increases. In contrast, Eq. 5 tends to decrease the relevance of shorter contacts, increasing the weight more rapidly as the duration approaches 86400.

Overall, the algorithm leverages from the probabilities of occurrence of a third repetition of a contact between a pair of nodes, which were found in the analytic study presented in the previous section, to derive a ranking algorithm. The algorithm uses the duration of the contacts and multiple temporal patterns between the same pair of nodes as tie breakers. Expectations are that these tie breakers correctly identify the contacts that are more likely to occur, providing to applications accurate estimates of upcoming contacts with other nodes.

Evaluation was performed by running contact datasets against the algorithm and comparing its ordering with the contacts that have been actually observed. The capability of the algorithm to correctly order the expectations of contacts was evaluated using two metrics. Both metrics measure the number of hits (defined as a prediction of contact that has effectively occurred), although using different perspectives:

The *Rank of the First Miss* (RFM) returns the rank in the list of the first failed prediction. RFM is useful for application programmers as it indicates the number of highly reliable predictions of the list.

The second metric compares the proportion of hits across the percentiles 10, 25, 50, 75 and 100 of the list. Percentiles permit to evaluate the quality of the ranking. Expectations are that the 100 percentile mirrors the analytic results discussed in Sec. 5. Therefore, the quality of the ranking will be evaluated by the increase in the proportion of hits in the lowest percentiles, which will confirm the capability of the algorithm to put hits at higher ranks.

6.1 Evaluation in MobIPLity

The ranking algorithm was experimented using a mobility scenario generated by MobIPLity [8] with all devices that connected to the eduroam network on IPL during the year of 2013. MobIPLity is a framework that produces mobility scenarios in bonnmotion [1] format. Contact data was produced by configuring the *LinkDump* application of bonnmotion to extract the periods in which two peers were within a 50m range from each other for a minimum of $60s$. To prevent disturbance on the results due to the distinct patterns

Table 4: Evaluation of contact duration functions

CD_{CD}	CD_{CWD}	rfm	p10
$k = 4$	$k = 5$	3.61	0.40
$k = 4$	$k = 6$	3.60	0.40
$k = 2$	$k = 2$	3.58	0.40
$k = 3$	$k = 6$	3.57	0.40
Eq. 5	Eq. 5	3.16	0.39

found on weekends, the original dataset was purged from the events occurring on Saturdays and Sundays.

It should be noted that the dataset used in the evaluation of the ranking algorithm is considerably distinct from the one used in Sec. 5 for the analytic evaluation of contacts. The later evaluated the 2012 dataset and defined a contact as the simultaneous association of two or more devices to the same access point, using RADIUS records. In this section, the 2013 dataset and a distinct methodology for defining contacts are used. In addition to exposing the ranking algorithm to a considerably distinct dataset from the one that inspired it, this approach permits to verify if the properties observed during the year 2012 are reproducible on a different year.

Parameters for the algorithm where experimentally tuned in order to obtain the best metrics. Table 4 depicts the experimental results for multiple variations of the CD functions with equal weights for w_{CD} and w_{CWD}. Results evidence a minimal impact of the k constant when the function of Eq. 4 is used, in contrast with the results exhibited by Eq. 5. In practice, this result evidences a preference of the algorithm for a fast growing of the weight of the contact duration in the ranking. As a result, the reminder of the text presents results using Eq. 4 with $k = 4$ for all the CD_{tp} functions.

Table 5 shows the average and standard deviation of the metrics when different weights are used. These results average the rankings produced for all devices and days, provided that the ranking contained 20 or more devices. The table shows some encouraging results. In particular, that the algorithm can correctly rank on average the first 3.6 devices and that 40% of the highest 10% ranked devices have been found as predicted. The contribution of the algorithm becomes more evident by noticing that a random sort of the list would equally distribute the 31% of the devices on the list that were effectively observed (p100) by all the percentiles.

Figure 11 further emphasis this result by evidencing the 28% performance gain of p10 over p100. A combined analysis of the figure and of the Table 5 highlights the distinct contribution of each of the temporal patterns to the algorithm, with the participation of the CMD or the use of individual patterns consistently presenting worst results than a 50%,50% combination of weights of CD and CWD.

In the elaboration of the results above, a ranking list is always prepared, independently of the connectivity of the device. However, a large number of cases were found where some devices did not connected to any another device in one complete day, although the algorithm predicted some connections. As can be confirmed by Table 6, this cannot be

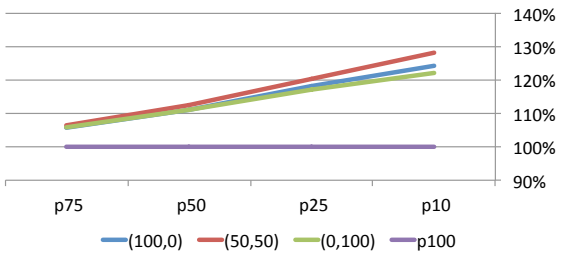


Figure 11: Improvement observed

considered a negligible aspect. The table presents the same metrics after excluding these lists. Not surprisingly, lists become much more accurate, with 100 percentile approaching an average of 50%. I.e., on average 50% of the devices predicted by the algorithm are effectively found. More demanding metrics, in particular RFM and p10 are in line with the improvement of p100: on average, the first 5 devices of each list are effectively found as predicted as well as more than 63% of the percentile 10 of each ranking list.

Table 7 discloses a final analysis of these results by presenting the metrics per day of the week. It is interesting to notice that the performance of the algorithm is not uniform across all the weekdays. The ranking algorithm presents better results for Tuesday, Wednesday and Thursday. The worst results of Mondays can be attributed to the weekend discontinuity impact on the CD temporal pattern. Surprisingly, Fridays present the worst performing results, although no evident explanation could be found.

6.2 Evaluation with Taxi Traces

To understand the applicability of the algorithm in a broad range of scenarios, the algorithm was experimented in a dataset containing 1 month GPS traces of 320 taxis in Rome [3]. The dataset was sanitized to include only positions in the metropolitan area of Rome, and to mark as off-line the taxis not reporting their position for an interval above $120s$.

Table 8 shows the metrics presented by the algorithm. Unfortunately, the smaller and shorter trace prevented experiments with the CMD temporal pattern and forced to a reduction of the minimum size of the rankings from 20 to 5. Surprisingly, p10 metric shows values comparable to the ones obtained from MobIPLity, for the same configuration parameters. The differences in the RFM can be attributed to the smaller dimension of the dataset, which necessarily reduces the ranking list and, proportionally impacts RFM. The difference between p10 and p100 loses significance, with a gain of 11%.

Table 9 show the outcome of the per day of the week analysis. One can observe that in this dataset, Monday is the worst performing day of the ranking algorithm. This result is attributed to the discarding of weekends that was kept from the MobIPLity analysis in an attempt to keep the comparison fair. However, the social constraints that encouraged the introduction of the exception for MobIPLity have no significance in a taxis scenario, where devices are expected to operate on all days of the week. To the extent of our knowledge, this was the unique characteristic of the

Table 5: Results (Average per day and Standard Deviation)

w_{CD}	w_{CWD}	w_{CMD}	Rank Totals	rfm (σ)	p10 (σ)	p25 (σ)	p50 (σ)	p75 (σ)	p100 (σ)
50	50	0	56892626	3.609 (6.75)	0.402 (0.38)	0.377 (0.35)	0.352 (0.32)	0.333 (0.30)	0.313 (0.29)
50	0	50	38423918	3.352 (7.13)	0.361 (0.37)	0.340 (0.34)	0.317 (0.32)	0.301 (0.30)	0.283 (0.28)
33	33	33	65604286	3.326 (6.30)	0.363 (0.37)	0.340 (0.34)	0.316 (0.32)	0.297 (0.30)	0.279 (0.28)
0	50	50	40827514	2.690 (3.81)	0.360 (0.37)	0.342 (0.35)	0.322 (0.33)	0.306 (0.31)	0.288 (0.29)

Table 6: Results excluding not connected days (Average per day and Standard Deviation)

w_{CD}	w_{CWD}	w_{CMD}	Rank Totals	rfm (σ)	p10 (σ)	p25 (σ)	p50 (σ)	p75 (σ)	p100 (σ)
50	50	0	39620508	5.122 (8.11)	0.635 (0.28)	0.596 (0.24)	0.557 (0.22)	0.526 (0.21)	0.495 (0.20)
33	33	33	43290708	4.967 (7.82)	0.620 (0.28)	0.579 (0.25)	0.538 (0.23)	0.507 (0.21)	0.476 (0.20)
50	0	50	25271814	4.871 (8.82)	0.593 (0.29)	0.559 (0.26)	0.522 (0.24)	0.495 (0.23)	0.466 (0.21)
0	50	50	25861694	3.917 (4.64)	0.621 (0.28)	0.591 (0.25)	0.556 (0.23)	0.528 (0.22)	0.498 (0.21)

Table 7: Per day of the week metrics for setup with the highest improvement (wCD=50,wCWD=50)

	rfm	p10	p25	p50	p75	p100
Monday	3.15 (4.62)	0.4 (0.38)	0.38 (0.35)	0.35 (0.32)	0.34 (0.31)	0.32 (0.29)
Tuesday	4.13 (7.57)	0.45 (0.38)	0.42 (0.35)	0.39 (0.33)	0.37 (0.31)	0.35 (0.3)
Wednesday	3.97 (8.26)	0.41 (0.38)	0.39 (0.34)	0.36 (0.32)	0.34 (0.3)	0.32 (0.28)
Thursday	3.67 (6.65)	0.41 (0.38)	0.38 (0.35)	0.36 (0.32)	0.34 (0.31)	0.32 (0.29)
Friday	3.04 (5.65)	0.34 (0.36)	0.32 (0.33)	0.29 (0.3)	0.27 (0.28)	0.26 (0.26)

Table 8: Taxis in Rome trace results

w_{CD}	w_{CWD}	Rank Totals	rfm (σ)	p10 (σ)	p25 (σ)	p50 (σ)	p75 (σ)	p100 (σ)
50	50	3487	1.884 (1.39)	0.426 (0.49)	0.409 (0.43)	0.380 (0.33)	0.340 (0.28)	0.312 (0.25)

Table 9: Per day of the week metrics for Taxis in Rome, setup with the highest improvement (wCD=50,wCWD=50)

	rfm	p10	p25	p50	p75	p100
Monday	1.27 (0.61)	0.19 (0.39)	0.2 (0.33)	0.18 (0.21)	0.16 (0.17)	0.15 (0.16)
Tuesday	1.87 (1.18)	0.48 (0.5)	0.42 (0.42)	0.41 (0.3)	0.39 (0.28)	0.35 (0.21)
Wednesday	1.95 (1.41)	0.44 (0.5)	0.41 (0.45)	0.4 (0.33)	0.34 (0.28)	0.31 (0.25)
Thursday	2.03 (1.5)	0.45 (0.5)	0.43 (0.45)	0.4 (0.35)	0.35 (0.3)	0.32 (0.25)
Friday	2.16 (1.58)	0.57 (0.5)	0.54 (0.41)	0.49 (0.3)	0.46 (0.27)	0.42 (0.24)

algorithm which did not adapt to both scenarios.

6.3 Discussion

In contrast with our expectations, differences in results between CD and CWD temporal patterns tend to be orthogonal to the environment. As an example, one could consider that the CD temporal pattern better represents faculty (with a daily schedule), and CWD would better represent students that meet in classrooms following a weekly schedule. And, that this would be tightly connected to the environment where the data was gathered. However, the results obtained using the data from MobIPLity present similar outcomes to the ones obtained in Sec. 5.2, using the same data but on different years. Furthermore, and considering that the Taxis in Rome trace also present the same results, to an extent, we can consider that this supports the usage of our algorithm and probability modelling on multiple environments.

Results suggest that performance could be improved by considering weekdays in the definition of the ranking algorithm. However, this claim must be supported by additional experiments in other traces, and therefore, is left as future work.

7. CONCLUSIONS

Developers of mobile applications can be faced with the need of anticipating the number or the affiliation of the groups of devices to be found in the future. This paper leverages on a large dataset of accesses to a number of eduroam network sites at an academic institution to extract the complete set of temporal communities observed between 2005 and 2013. The paper derives a statistical model that characterizes the different temporal community sizes assuming four distinct recurrence patterns that mirror likely schedules of the users.

The fitting of the observations showed that the recurrence of three temporal patterns can be modelled by generalized pareto distributions, confirming the results already observed for the Inter-Contact Times.

The paper presents an algorithm to predict contacts with peer devices on a nearby future using the knowledge of the previous observed contacts and temporal patterns. The algorithm was evaluated against two datasets: one prepared by extracting data from a different year of the same dataset that inspired this research and another prepared from a pub-

licly available trace of taxis in Rome. Evaluation results showed that the algorithm presents good prediction capability and that its performance is comparable in both scenarios, what raises our expectations of its applicability as a generic prediction tool.

Analysis of the data continues. As future work, authors plan to apply the lessons learnt in the modulation of recurrence of temporal communities on real world applications. In addition, work will continue in the analysis and expansion of the dataset and on the verification if the results presented in the paper are reproducible in other datasets.

8. REFERENCES

[1] N. Aschenbruck, R. Ernst, E. Gerhards-Padilla, and M. Schwamborn. Bonnmotion: A mobility scenario generation and analysis tool. In *Procs. of the 3rd Int'l ICST Conf. on Simulation Tools and Techniques*, 2010.

[2] C. Boldrini and A. Passarella. Hcmm: Modelling spatial and temporal properties of human mobility driven by users' social relationships. *Computer Communications*, 2010.

[3] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, and A. Rabuffi. CRAWDAD data set roma/taxi (v. 2014-07-17). Downloaded from http://crawdad.org/roma/taxi/, July 2014.

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

[5] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[6] M. Conti, S. Giordano, M. May, and A. Passarella. From opportunistic networks to opportunistic computing. *Communications Magazine, IEEE*, 48(9):126–139, Sept 2010.

[7] P. Costa, C. Mascolo, M. Musolesi, and G. Picco. Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks. *Selected Areas in Communications, IEEE Journal on*, 26(5):748–760, June 2008.

[8] N. Cruz and H. Miranda. MobIPLity: A trace-based mobility scenario generator for mobile applications. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (Mobiquitous'14)*, London, UK, Dec. 2–5 2014. EAI.

[9] N. Cruz, H. Miranda, and P. Ribeiro. The evolution of user mobility on the eduroam network. In *Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 249–253, Mar. 24 2014.

[10] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, Jun 2008.

[11] D. Huang, S. Zhang, P. Hui, and Z. Chen. Link pattern prediction in opportunistic networks with kernel regression. In *Proceedings of The 7th International Conference on COMmunication Systems and NETworkS*, COMSNETS 2015, Jan. 2015.

[12] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: Social-based forwarding in delay-tolerant networks. *Mobile Computing, IEEE Transactions on*, 10(11):1576–1589, Nov 2011.

[13] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic. Power law and exponential decay of intercontact times between mobile devices. *Mobile Computing, IEEE Transactions on*, 9(10):1377–1390, Oct 2010.

[14] T. Karagiannis, J. Y. Le Boudec, and M. Vojnovic. Power law and exponential decay of intercontact times between mobile devices. *Mobile Computing, IEEE Transactions on*, 2010.

[15] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, 2009.

[16] H. Miranda, S. Leggio, L. Rodrigues, and K. Raatikainen. An algorithm for dissemination and retrieval of information in wireless ad hoc networks. In A.-M. Kermarrec, L. Bougé, and T. Priol, editors, *Proceedings of the 13th International Euro-Par Conference, Euro-Par 2007*, volume 4641 of *Lecture Notes in Computer Science*, pages 891–900, Rennes, France, Aug. 28–31 2007. Springer.

[17] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70:056131, Nov 2004.

[18] M. Orlinski and N. Filer. The rise and fall of spatio-temporal clusters in mobile ad hoc networks. *Ad Hoc Networks*, 11(5):1641 – 1654, 2013.

[19] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, Jun 2005.

[20] L. Pelusi, A. Passarella, and M. Conti. Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. *Communications Magazine, IEEE*, 44(11):134–141, November 2006.

[21] A.-K. Pietilänen and C. Diot. Dissemination in opportunistic social networks: The role of temporal communities. In *Proceedings of the Thirteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '12, pages 165–174, New York, NY, USA, 2012. ACM.

[22] C. Rigney. RADIUS Accounting. RFC 2866 (Informational), June 2000.

[23] J. Su, J. Scott, P. Hui, J. Crowcroft, E. de Lara, C. Diot, A. Goel, M. Lim, and E. Upton. Haggle: Seamless networking for mobile applications. In J. Krumm, G. Abowd, A. Seneviratne, and T. Strang, editors, *UbiComp 2007: Ubiquitous Computing*, volume 4717 of *Lecture Notes in Computer Science*, pages 391–408. Springer Berlin Heidelberg, 2007.