

Groups and Frequent Visitors Shaping the Space Dynamics

Karolina Baras^{1,*} and Adriano Moreira²

¹ Exact Sciences and Engineering Competence Center,
University of Madeira, Campus da Penteada, 9000-390 Funchal, Portugal
kbaras@uma.pt

² Department of Information Systems, University of Minho, Guimarães, Portugal

Abstract. Our research is about a dynamic symbolic space model that is fed with data from the environment by a set of processing modules that receive raw data from sensor networks. For the conducted experiments we have been using data from a WiFi network as it is a widely available infrastructure in our campus. Here we propose two processing modules which will provide more information about the spaces described in the model. The first one tries to implement our human perception of the usual visitors of a place using two measures, the long term and the short term tenant level. The second one detects where groups of users emerge, how many there are and what are their dimensions. Based on this new perspective of the campus we intend to realize how the presence of people shapes the dynamics of a space.

Keywords: groups of users, space dynamics, symbolic space model, WiFi.

1 Introduction

Public places tend to be extremely dynamic. This means that their contexts change instantaneously as a result of different factors. These include, in the first place, the presence of people and their activities, but also changes of physical characteristics, schedules and so forth. In this paper we will focus on detecting groups of people and their relation with the space dynamics.

Our research is based on a dynamic symbolic world model which is integrated in a larger system that comprises sensor networks, data acquisition and processing modules and applications. The world model representing a particular area is fed by the results obtained from the processing of the data coming from the locally deployed sensor networks. It acts as a repository of rich context information for applications aimed at the usual users of a place or the newcomers. Our main source of context information about spaces is the wireless network which is widely deployed in the University area. Our previous results [2] have shown that the WiFi network data is quite rich in context and many interesting details about spaces can be extracted from there.

* The first author would like to thank FCT for supporting this work through grant SFRH/BD/36046/2007.

2 Related Work

During the last decade wireless LANs have increasingly become commonplace on many university campuses, in enterprises and at other public venues providing access to the Internet on the move. Every laptop, and more recently, every smartphone have also become WiFi enabled. Extended research has been made on the wireless networks usage, e.g., [4] and user mobility and traffic patterns, e.g., [3], aiming to provide useful elements that might help to improve network performance [1] and management and planning [6].

As WiFi networks spread over many public and private places, the potential for this technology to capture the dynamics of the space is enormous. We are interested in metrics that have been used for the characterization of the wireless network usage and user mobility, but our prime objective is to characterize space dynamics and to understand the usage patterns and the pulse of each observed space. For instance, the number of users per access point can provide us with some information about the popularity of an access point and, consequently, about the popularity of a space where it is deployed. We infer the way a space is used and whether it is, normally, one of the most popular areas in the campus, through the analysis of the access point tenant levels and group detection. Daily, weekly and seasonal trends give us a picture of the space pulse, contributing to its characterization. Although our approach is different and we have not found any similar work in the literature, we describe below some projects that are somehow related to ours. In [7], the authors observed a change of habits of the students concerning the choice of a place to study or socialize as a consequence of the implementation and expansion of a WiFi network. The work described in [8] links together data about the human presence in spaces based on a WiFi network data and a system that monitors HVAC levels in order to improve energy efficiency in buildings. An example of a dynamic system that is constantly and automatically updated as new people join in or new events are about to happen is the Just-for-us project [5] developed in Melbourne. This system aims to encourage new forms of social interaction in public spaces.

Although the data source is the same, our objectives are significantly different as we are using this data to populate a world model with dynamic data coming from the environment. In this paper we give two examples, the tenant levels which describe to what extent a space is visited always by the same users or by many different users, and the groups detection which allows to add more characteristics to a space description according to the number of groups, their dimension and the time they spend at each place.

3 A Dynamic Space Model

The world model we propose was inspired in the human mental models of the world. It is a symbolic model and it consists of *objects*, *object attributes*, *relations* and *relation attributes* as illustrated in Fig. 1. These elements are created, either manually by the user, or automatically by the processing modules. Each object

has a name, a URI (Uniform Resource Identifier) that uniquely identifies it (the object id, a host name or IP address of the server where the object is stored and a port from which it is accessible), a type, an author, a creation date, availability (*public* or *private*) and a status (*active*, *idle*, *inactive*). An object can have an arbitrary number of attributes and an arbitrary number of relations with other objects.

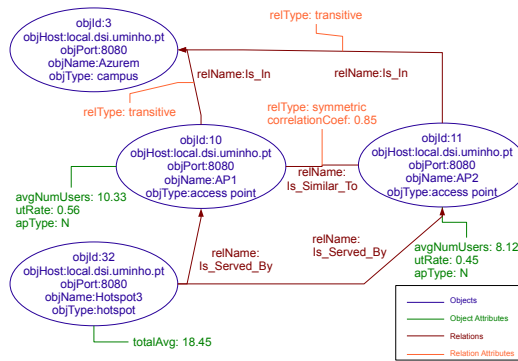


Fig. 1. A snapshot of the symbolic space model

$$\mathbf{object} = \{objid, \text{URI}, \text{name}, \text{type}, \text{author}, \text{cdate}, \text{availability}, \text{status}\}$$

An attribute is defined by the attribute id, the id of the described object, a timestamp and a name-value pair.

$$\mathbf{objAttribute} = \{attid, objid, \text{timestamp}, \text{name}, \text{value}\}$$

The semantics of relations is based on the way humans mentally relate objects or places. If we have a well known landmark, we will relate other objects to it, as in the following expression: "the University is near the castle". In our model this is achieved through semantic relations that may be established between objects. A relation is defined by a relation id, a name, a type, a creation date, the URI of the object in the domain and the URI of the object in the co-domain. It is characterized by an arbitrary set of relation attributes. Relations support inference processes that extract implicit knowledge from the model. The way a relation is handled by the inference algorithms is based on its name and type.

$$\mathbf{relation} = \{relid, \text{name}, \text{type}, \text{cdate}, \text{URIdom}, \text{URIdom}\}$$

Examples of the current relation names supported by our inference algorithms are: *Is_In*, which expresses spatial or administrative containment, *Is_Accessible_From*, which expresses accessibility or connectedness and *Is_Next_To* which expresses adjacency between two objects. Presently, the implemented inference algorithms support the *transitive* and the *symmetric* relation types. Other relation names and types are allowed but without any inference being performed.

A relation attribute is defined by its own id, the id of the relation to which it belongs, a timestamp and a name-value pair.

relAttribute = {*relattid*, *releid*, *timestamp*, *name*, *value*}

Currently, the model is implemented on top of a relational database. There is a service, the Symbolic Contextualizer, that provides the interface to the model, and allows querying and editing the model data. The model is integrated in a bigger system, as shown in Fig. 2. The model is updated in real time by a set of processing modules that use sensor data and the model data to infer characteristics of the objects stored in the model.

In order to take advantage of the already existing infrastructures which are deployed in our University campus, we are using data from the WiFi network as our main source of context information about spaces within the campus. This option was also motivated by the results of other research projects which studied WiFi networks and found that they were a very useful source of information about space usage [7]. The Location Server in Fig. 2 stores data containing the number of users per access point and the corresponding lists of MAC addresses every five minutes. All the acquired data is accessible through a query interface provided by the Sensor Data Service.

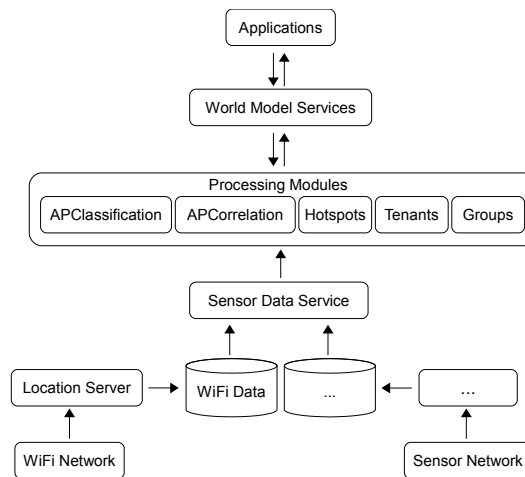


Fig. 2. Concise system architecture

In our previous work [2], we have analyzed data about the number of users per access point and identified useful information for creation and updating of the elements of our symbolic world model. This resulted in three processing modules, one for classification of access points according to their usage patterns, one for calculating correlation factor between access points and one for detecting hotspots that uses data produced by the two previous modules.

Next we explore the detection of frequent users and groups as two examples of processing modules that may feed the world model with data that contributes to space characterization. First, we explain the motivation for creating each of

these processing modules. Then, we identify the output that they are going to produce and, finally, the rules that should be followed for updating the model with these results. We assume that one MAC address corresponds to one person. Our objective is to observe if the users in a particular area of the campus are the usual visitors of that place or not by calculating the *tenant level* of a place as well as find and characterize the places in the campus where the groups emerge and spend their time.

4 Tenants, Visitors and Strangers

We propose a simple way of modeling our human perception of the usual users of a place. For instance, when we enter a familiar space, we easily notice whether the usual people are there or if there is someone new or someone missing. We propose a way of measuring this through the number of occurrences of each MAC address during a given time period. The process of calculating the *tenant level* of an access point, is the responsibility of one of the processing modules dedicated to this particular operation. Besides the calculations, it is also responsible for inserting the results in the model.

Motivation. Detecting a set of MAC addresses that repeatedly appear at some nodes of a network may allow for characterization of a place in terms of its users. Keeping track of the type of visitors of a particular place (place visitor profile) may let us discover whether a place is normally visited by its tenants, frequent or occasional visitors or complete strangers. This kind of knowledge may be used by context-aware applications running in the infrastructure.

Output for the model. Information about the frequent visitors of an access point may allow for the update of two object attributes. One describing the *long term tenant level* of a place and the other the *short term tenant level* as it will be explained in Sects. 4.1 and 4.2.

The rules. Update two attributes, one called *long term tenant level* and the other called *short term tenant level* with the values between 0 and 1.

For each access point, the list of connected devices' MAC addresses is recorded every five minutes. Frequencies of each detected MAC address are stored and updated. Fig. 3 shows an example in which data was stored during 24h resulting in 288 samples. In each sample the number of connected MACs per access point is recorded and the counter of occurrences for each present MAC is incremented. After the acquisition of samples during a time period, the available data consists of the frequencies of each detected MAC address. In our experiments the considered time periods were 20 days (corresponds to 5760 samples) for the long term and 5 minutes (a single sample) for the short term.

4.1 Long Term Tenant Level

The *long term tenant level* shows if a place is mostly visited by people who are always there, frequently dropping in or just occasionally passing by. It is

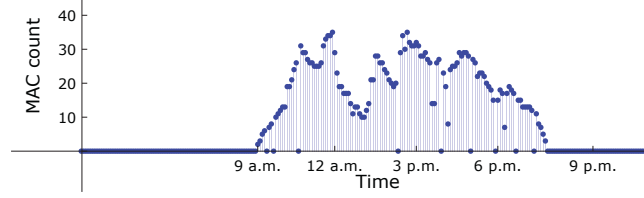


Fig. 3. 24h hours of MAC counts for an AP

calculated for the data collected during a fairly long time period, like several weeks, in the following way:

1. Calculate absolute frequencies for each observed MAC address i ($f_{abs,i}$);
2. Calculate normalized frequencies of observed MACs: $f_{norm,i} = f_{abs,i}/S$, where S is the number of samples in which at least one MAC was detected (valid samples). This value is chosen instead of the total number of samples that are taken during the considered time interval in order to discard periods in which nobody connects, e.g. during the night or during the weekend. In places that never close and where there are always people, it may be more suitable to consider the total number of samples. In Fig. 3, S is given by the number of time intervals between 9 a.m. and 7.30 p.m.;
3. Classify MACs based on their normalized frequencies and according to the rules described below in Table 1. The result is depicted in Fig. 4 for a set of 19 MACs randomly chosen out of the total of 720 MACs detected during the observation period (20 days).

Table 1. Place visitor profile based on detected MACs

Normalized frequency (f_{norm})	MAC Class
$f_{norm} \geq 0.2$	Tenants
$0.1 \leq f_{norm} < 0.2$	Frequent Visitors
$0.01 \leq f_{norm} < 0.1$	Visitors
$f_{norm} < 0.01$	Strangers

4. Count the total number of detected MACs (n_{MACs}).
5. Calculate the final value for the long term tenant level:

$$tenantLevel_{LT} = \frac{f_{normT} + f_{normFV} + f_{normV}}{n_{MACs}} \quad (1)$$

which is the sum of all the normalized frequencies higher or equal to 0.01 (all tenants, frequent visitors and visitors) divided by the total number of detected MACs. This results in a value between 0 and 1 for the *long term tenant level* of a place.

For the example illustrated in Fig. 4, the calculated normalized frequencies reveal the detection of 6 tenants, 1 frequent visitor, 7 visitors, and 5 strangers. The resulting tenant level follows from (1): $tenantLevel_{LT} = 0.13$.

4.2 Short Term Tenant Level

The *short term tenant level* shows, in turn, if the considered area is being visited at a particular moment by its tenants, frequent visitors, visitors or strangers. The *short term tenant level* is calculated for a single sample as follows:

1. Count the number of tenants, frequent visitors, visitors and strangers in the current sample by comparison to their class, if any, in the long term place visitor profile. MACs appearing for the first time are considered strangers;
2. Calculate the value for the *short term tenant level* as:

$$tenantLevel_{ST} = \frac{f_{normT} + f_{normFV} + f_{normV}}{m_{MACs}} \quad (2)$$

which is the sum of the normalized frequencies higher or equal to 0.01 divided by the total number of detected MACs (tenants, frequent visitors, visitors and strangers) in the current sample. This results in a value between 0 and 1 for the *short term tenant level* of a place.

For the same case from the Fig. 4, where we calculated the *long term tenant level* for an access point with 19 detected devices, we will now calculate the *short term tenant level*. To do this, we will consider that, the following devices were detected in a single sample: A, D, J, K, L and N. According to the figure, we have 2 tenants (D, K), 1 frequent visitor (N), 1 visitor (A) and 1 stranger (L). If we use the same values for their normalized frequencies as those shown in the figure, (2) yields $tenantLevel_{ST} = 0.21$. So it results in a slightly different value than the one obtained for the long term. Although a place may be mostly visited by occasional visitors, there may exist some time periods when it is visited by its tenants and frequent visitors.

For the following examples the *long term tenant level* was calculated for a 20 days long data set and the *short term tenant level* was calculated for each single sample during a 48h time period (from Sunday, 5 p.m. to Tuesday 5 p.m.). The first example (Fig. 5(a)) is from one of the library access points which is usually visited by a very large number of users during the opening hours. The second example (Fig. 5(b)) is from a rarely used access point, localized near the design rooms in the department of architecture. The third example (Fig. 5(c)) shows the results for the set of four access points in the students' dormitories which

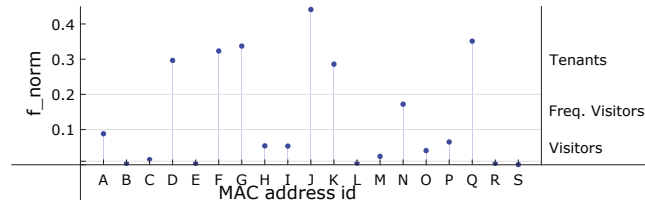
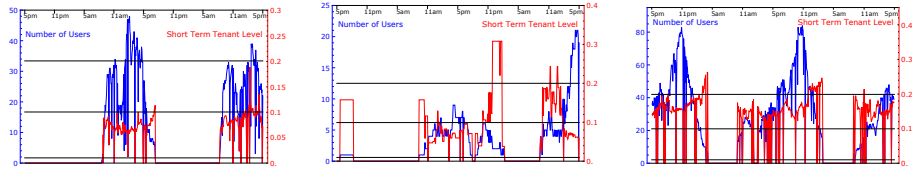


Fig. 4. An example of a place visitor profile based on a set of MACs detected at an access point during an observation period of 20 days



(a) The library is characterized by many occasional visitors. (b) Design rooms are characterized by a few tenants when no one else is there. (c) Dormitories are characterized by frequent visitors and tenants.

Fig. 5. Short term tenant level variation during two days time period. Left Y-axis represents the number of users and the right Y-axis, the short term tenant level.

are usually heavily used during all week. It shows strong frequent visitor, and sometimes, tenant profile.

It makes sense that in the library the visitor profile is characterized by many occasional visitors as it is a place where a large portion of the academic population spends some time during a week. However, the characterization of the design rooms profile is not as clear, because there are many fluctuations during the day and during the week. At some hours of the day only tenants are found, but as soon as the number of users increases, the tenant level value lowers. In the dormitories the profile is quite clear. Although there is a large number of users like in the library, this place is mostly used by its residents. Although it is not always clear whether the users connected at a particular access point are its tenants or occasional visitors, the results show a new perspective on the space usage. New knowledge about a space may be extracted from graphics like those shown in Fig. 5.

5 Groups of Users

In this Section we propose a processing module that is responsible for the detection of groups of users in a WiFi network coverage area. This is achieved based on the MAC address listings retrieved from the access points. This module also inserts the resulting data about groups into the space model.

Definition 1. *An iteration consists of six consecutive samples, taken every five minutes.*

This means that a single iteration is 30 minutes long. Due to overlapping (see Fig. 6) the duration of two consecutive iterations is 35 minutes.

Definition 2. *A group is defined as a set of two or more MAC addresses that are detected together at two different access points, during at least four iterations ($N_{\text{iter}} = 4$) at one of them.*

As explained above and shown in Fig. 6 the duration of this time interval may vary, depending on whether the MACs X and Y are found together in consecutive or non-consecutive iterations.

The minimum number of iterations for the creation of a group determines the number of detected groups. Four iterations for a week time is a reasonable value because it means that the members of the group are found together during at least 45 minutes, which is the minimum duration of a lecture. Also, the constraint of at least two different access points is important for it helps to eliminate laboratory equipment that is connected to the WiFi network unattended. An enhancement could be to establish a minimum distance between the two different access points in order to eliminate the "ping-pong" effect which may occur between nearby access points.

Motivation. The purpose of detecting groups of users is to extract more characteristics of the spaces where these groups are created, according to their number and their dimension. Identification of groups of devices, and to some extent, groups of users, allows for the identification of the areas in the campus where the groups of users work or meet. It also may add more dynamics to our model as it reflects their presence and movement within the area represented in the model.

Output for the model. When a group is identified, an object representing it may be created in the model. A relation with an access point may also be created every time a group connects to the network. There may be a list of all visited access points as an attribute of a group that may be used for inference of additional information about the group itself and the places it visits.

The rules. Two or more MAC addresses should be found connected to at least two access points during at least N_{iter} iterations for an object of type *group* to be created. Once created, this object will have a relation *Is_In_Range_Of* established with an access point.

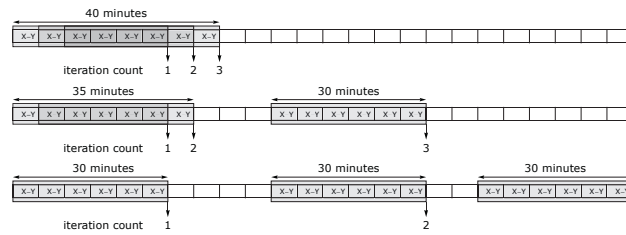
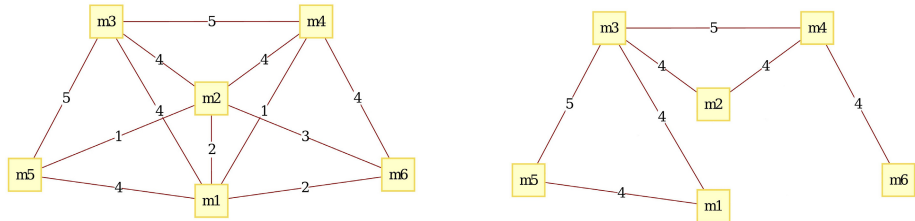


Fig. 6. Duration of the total number of counted iterations is variable. Three iterations may mean that the members of a group were found together during 40 min, 65 min or 90 min. The minimum of minutes, given by $30 + (N_{iter} - 1) \times 5$, is obtained if all the iterations are consecutive (partially overlapping) and the maximum, equal to $N_{iter} \times 30$ is obtained if there is no overlapping between iterations.

Fig. 7(a) shows a graph representation of a group in which vertices correspond to MAC addresses that were found connected to an access point during a considered time interval. The weight associated with each edge represents the number of iterations during which the pair of MACs was observed. If we set a threshold



(a) Edges between MAC addresses correspond to the number of iterations.

(b) When a threshold of 4 is set for the minimum number of iterations, edges with lower weights are removed.

Fig. 7. Graph representation of a group of users

to 4 iterations ($N_{iter} = 4$), as we did in our experiments, all the edges with weights lower than 4 are removed as shown in Fig. 7(b).

In our experiments, we detected 50 groups on average during one day in the campus. During one week we detected 830 groups on average. We also identified places in the campus where the largest number of groups emerge and also where the largest groups appear. Each of the considered places is covered by at least two and at most four access points. As shown in Fig. 8, the most frequent group dimension is 2. The largest group had 42 members and it was detected in the dormitories.

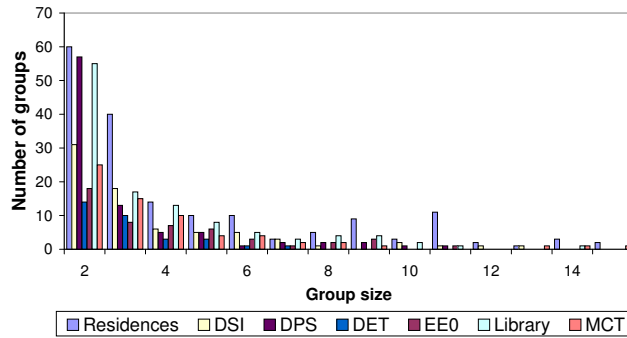


Fig. 8. Dimensions and number of groups for a week long data set. Larger groups were detected, 42 being the biggest. They are not shown here as they are not very frequent.

In comparison to the total number of unique MACs detected (3222), those participating in groups (945) represent nearly 30%. As expected, the most crowded places are the library and the dormitories and there appear the largest number of groups. We identified 7 places in the campus where 70% of the groups emerge (see Fig. 9). If we do not consider the dormitories, than we have 50% of all groups emerging in 6 places.

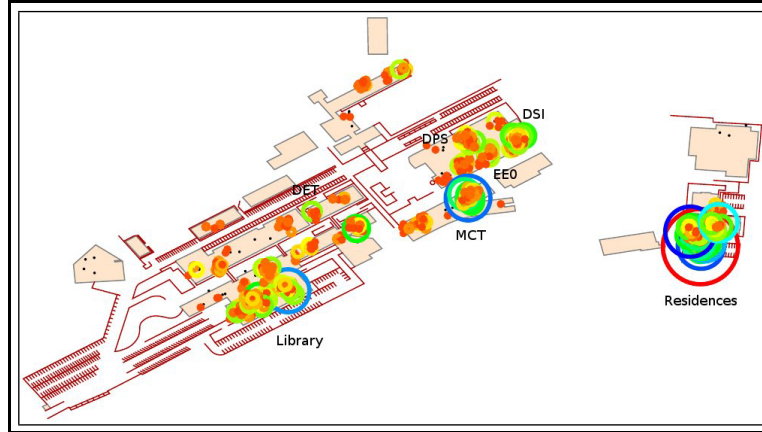


Fig. 9. Groups at the campus after analyzing a week long data set. The smallest groups have 2 members and the largest 42 (red circle). The radius of the circle is proportional to the dimension of the group.

6 Conclusions and Future Work

Starting with a structure of a dynamic world model and the data about the utilization of an University WiFi network, we have identified two additional sources of information for the update and the expansion of the developed model. We have defined a new parameter for characterization of spaces, the *tenant level*. This parameter translates the human ability to distinguish whether a familiar place is occupied by the usual users or there are new people around. The second source of data for our model comes from the detection of groups. We were able to detect where the groups emerge and what are their dimensions. These two new aspects add a new perspective on spaces derived from data gathered from the WiFi network.

Further developments may be done based on the results from this paper. For example, we may study the movement of groups around the campus and identify the most frequent trajectories. Besides the individual group detection, we can also observe the interconnections between all the existent groups in the campus. This happens due to the fact that people participate in several social groups and so they act as links between these groups. So, if we observe a week long data in all access points at once, we obtain a large group containing the great majority of detected MACs and several smaller groups that remain separated, probably representing occasional visitors. The social hierarchy of the users of a WiFi network may be studied through the observation of the groups that are formed at different levels of detail, according to the value of N_{iter} parameter. The largest group we found contains nearly all the users of the considered WiFi network. Inside the largest group of users, smaller groups consist of people belonging to

different courses and classes, inside the classes we can find groups of friends, and so on.

We have developed a prototype application that shows the current number of users for a set of APs through a color code in an attempt to illustrate the current state of the campus. For the future work, we intend to develop a new and more complete version of this application in order to allow for the visualization of the state of the campus in real time. A screen shot of that application will be something similar to Fig. 9. This application may be targeted at people coming to the campus for the first time to help them find the places of interest inside the campus. Further information about our ongoing research may be found at <http://ubicomp.algoritmi.uminho.pt/symbolic/>.

References

1. Balachandran, A., Voelker, G.M., Bahl, P., Rangan, P.V.: Characterizing user behavior and network performance in a public wireless LAN. In: SIGMETRICS 2002: Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 195–205. ACM, New York (2002)
2. Baras, K., Moreira, A.: Symbolic space modeling based on WiFi network data analysis. In: 2010 Seventh International Conference on Networked Sensing Systems, INSS (2010)
3. Henderson, T., Kotz, D., Abyzov, I.: The changing usage of a mature campus-wide wireless network. *Computer Networks* 52(14), 2690–2712 (2008)
4. Kim, M., Kotz, D.: Periodic properties of user mobility and access-point popularity. *Personal Ubiquitous Computing* 11(6), 465–479 (2007)
5. Kjeldskov, J., Paay, J.: Public pervasive computing: Making the invisible visible. *IEEE Computer* 39(9), 60–65 (2006)
6. Schwab, D., Bunt, R.: Characterising the use of a campus wireless network. In: Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2004, vol. 2, pp. 862–870 (March 2004)
7. Sevtsuk, A., Huang, S., Calabrese, F., Ratti, C.: Mapping the MIT campus in real time using WiFi. In: Foth, M. (ed.) *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*, pp. 326–337. IGI Global, Hershey PA (2008)
8. Vaccari, A., Samouhos, S., Glicksman, L., Ratti, C.: Mit enernet: Correlating WiFi activity to human occupancy. In: *Proceedings of Healthy Buildings 2009, the 9th International Conference and Exhibition* (2009)