Spot the Hotspot: Wi-Fi Hotspot Classification from Internet Traffic

Andrey Finkelshtein(✉), Rami Puzis, Asaf Shabtai, and Bronislav Sidik

Ben Gurion University of the Negev, Beersheba, Israel
{andreyfi, sidik}@post.bgu.ac.il,
{puzis, shabtaia}@bgu.ac.il

Abstract. The meteoric progress of Internet technologies and PDA (personal digital assistant) devices has made public Wi-Fi hotspots very popular. Nowadays, hotspots can be found almost anywhere: organizations, home networks, public transport systems, restaurants, etc. The Internet usage patterns (e.g. browsing) differ with the hotspot venue. This insight introduces new traffic profiling opportunities. Using machine learning techniques we show that it is possible to infer types of venues that provide Wi-Fi access (e.g., organizations and hangout places) by analyzing the Internet traffic of connected mobile phones. We show that it is possible to infer the user's current venue type disclosing his/her current context. This information can be used for improving personalized and context aware services such as web search engines or online shops, without the presence on user's device. In this paper we evaluate venue type inference based on mobile phone traffic collected from 115 college students and analyze their Internet behavior across the different venues types.

Keywords: Smartphone · Machine learning · Classification · Wi-Fi · Hotspot

1 Introduction

The widespread use of ubiquitous devices with Internet access such as laptops, smartphones, and tablets is on the rise. According to Ofcom [2], in 2014, 62 % of adults in the UK used smartphones, and 52 % considered smartphones (22 %) and tablets (30 %) the most important devices for Internet access. Browsing the Internet is one of the most common activities (performed daily) of smartphone users. [11], usually via a Wi-Fi hotspot connection. According to "iPass Wi-Fi growth map" (http://www.ipass.com/wifi-growth-map), at the start of 2015, there were 50 million worldwide Wi-Fi hotspots, and their number is expected to hit 340 million by 2018.

Wi-Fi traffic analysis can provide interesting insights about users and their interests in various hotspot venues. Previous works have studied users' behavior and Wi-Fi hotspot properties in order to cluster users according to their engagement and length of stay [5, 10] or to improve network quality of service [3, 4].

We investigate the network usage patterns of mobile devices connected to hotspots located across different types of venues, such as home, work, or public transportation. These patterns can be used to infer user context and the type of venue a user is visiting

K.S. Xu et al. (Eds.): SBP-BRiMS 2016, LNCS 9708, pp. 239–249, 2016.
DOI: 10.1007/978-3-319-39931-7_23

without the need to know the user's exact location. Deriving user context can be used to provide context-aware personalized services such as recommendation systems or targeted advertisements [8]. For example, a recommendation in push notification can be appropriate while the user is waiting for a train but not while the user is at work. The type of venue a user is located in can also support context-aware access control policies [7]. For example, an organizational VPN can restrict access to specific resources to users located in public areas. In addition, venue type inference can be used to enrich maps and Wi-Fi hotspot databases such as WiGLE (https://wigle.net) in a non-intrusive manner.

In this paper we introduce the venue classification problem. Based on the properties extracted from the traffic of a single user the solver should infer the type of the venue the user is located in. Experimental results show that the type of venue can be identified with an accuracy rate of 75 %. In addition, in order to understand the users' behavior across venues, we distinguish between two types of properties: user agnostic *hotspot properties* such as quality of service (QoS) and *user behavior properties* such as categories of browsed websites. Analysis of *user behavior properties* distributions in the collected data and feature selection results show that the Internet behavior of smartphone users changes in different types of venues. On one hand, the most popular domain categories (e.g., social, search, business) are similar across different types of venues. On the other hand, the browsing patterns for less popular website categories (e.g., blogs, travel, and news), together with the domains' popularity and security ranks, differ significantly based on the venue type.

To the best of our knowledge, no work has been done on venue type inference based on the Internet traffic traces of mobile device users.

2 Related Work

Learning users' behavior and optimizing network performance are two important tasks associated with Internet communication, specifically when connected via Wi-Fi hotspot. A great deal of valuable information can be extracted from Internet traffic dumps. Previous works have studied users' behavior and hotspot properties in wireless networks. Afanasyev *et al*. [3] studied the Google Wi-Fi network deployed in Mountain View, California. In 2008, the authors collected information for a period of 28 days. They analayzed the diversity of coverage, temporal activity, traffic demands, and mobility of the network. The users were grouped by device: smartphones, laptops, and static devices. They found significant differences between the three groups in terms of session lengths, network usage, and the diversity of application layer protocols.

Balachandran *et al*. [4] also characterized user behavior in wireless networks using Internet traffic. Traffic traces were collected from a wireless network during a computer networking conference. These traces contained aggregated packet level statistics of the link and network, transport, and application layers, together with information about users in different hotspots such as MAC addresses and signal-to-noise-ratios (SNR). The authors investigated session duration, data rates, popular protocols, and user mobility. Moreover, they studied network performance at different hours of the day. Their main findings were as follows: (1) users primarily use the Internet for browsing and SSH communication, (2) they tend to connect for short time sessions, and (3) their data usage

is uneven. The authors observed specific data patterns from wireless hotspots at the conference, with patterns of use associated with specific concentrated locations and specific periods of time. This pattern is similar to that of classrooms, airport gates, etc.; however, it is not characteristic of all wireless hotspots in public areas.

Other studies employed machine learning in order to classify users' behavior based on Wi-Fi traffic. ToGo [5] is a system used to predict the length of a user's stay at a Wi-Fi hotspot. The authors presented SVM models for predicting this, starting with a basic model with no feedback, solely based on time and RSSI (signal strength)/bitrate, and progressing to models that use smartphones sensors such as the accelerometer. The paper also demonstrates the use of ToGo in an experiment involving 15 users. The Mo-Fi system [10] predicts the length of stay of users in a hotspot. It extracts features from three types of Wi-Fi packets: Wi-Fi probe requests, probe responses, and data messages. Using the k-means clustering algorithm, the system clusters Wi-Fi users into four different groups: outside, walkbys, bounced, or engaged. The writers evaluated their system in a real life office environment and achieved a human presence detection rate of 87.4 %.

Namiot [8] presented SpotEX – an Android application that displays ads to users based on the SSID and other publicly available information about nearby hotspots. Unfortunately, SpotEX requires an agent to be present on the user device and it relies on a-priory knowledge of the hotspot types in order to infer the user context. In contrast to SpotEX we infer the hotspot types from the Internet traffic only and do not require an agent on the user's phone.

3 Wi-Fi Traffic Dataset

In the course of the current study we collected the traffic dumps of smartphone users within a specific geographic area. In this section we describe the data collection process, cleanup, and feature extraction as presented in Fig. 1.

3.1 Data Collection

Internet traffic data was collected from the smartphone devices of 115 students from Ben-Gurion University of the Negev for a period of 30 to 60 consecutive days during

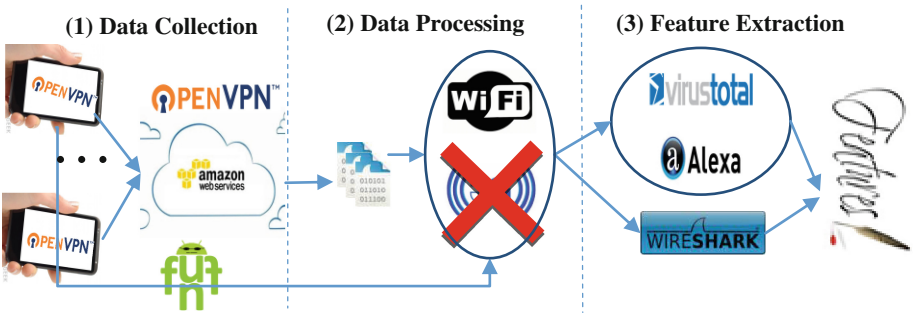

Fig. 1. Dataset processing scheme: (1) Client application redirects Internet traffic to a server and records WiFi connection/disconnection events. (2) Cellular traffic is filtered out. (3) Features are extracted for each session.

2014 and 2015. The traffic includes both Wi-Fi and cellular data that was collected using a dedicated VPN service. As such, traffic dumps contain IP packets but no data-link layer frame headers. In addition to the traffic dumps, we collected information about Wi-Fi hotspots using a dedicated Android application. All data was securely stored on a research server and analyzed using software tools for the purpose of Internet usage analysis in accordance with the permission of the university's ethics committee and subjects' consent.

3.2 Data Processing

We aggregated the collected IP packets into sessions (a session was either a TCP session or a UDP message and its response). First, we associated each session with a venue. Since much of the activity was performed on a university campus, an urban area with multiple venues such as offices and coffee shops located in the same buildings, we could not rely on GPS localization for reliable labeling of venues. Therefore, in the current study we used only Wi-Fi traffic. Every session was associated with a venue according to the SSID and BSSID of the hotspot which were collected by our application (a total of 978 hotspots). All sessions taking place between consecutive Wi-Fi connection and disconnection events were associated with the respective BSSID. The rest of the sessions were considered cellular traffic and were disregarded.

Next, we associated each hotspot with the venue it is located in. The hotspots were manually labeled based on SSID and BSSID. If the hotspot had very few sessions or it could not be labeled, its traffic was removed from the dataset. At the end of this process the data contained sessions for 738 different hotspots.

3.3 Feature Extraction

Following the hotspot labeling and aggregation of traffic into sessions, we extracted the following sets of features (summarized in Table 1).

Communication features focus on traffic volume, the ratio between sent and received traffic, duration of sessions, packet arrival times, and the amount of lost packets. Statistical values were extracted for each of those attributes: average, median, first quarter, third quarter, minimum, maximum, entropy, standard deviation, and variance.

Protocol-based features are extracted from the protocol headers. From the network layer (IP protocol) we extracted statistics about the TTL (time to live) values, and the GeoLite database was used to map IP addresses to their countries. Port usage and quality of service attributes, such as the number of lost segments and retransmitted packets, were evaluated from the transport layer protocols (TCP and UDP). The application layer was also used to extract features. We focused on HTTP, HTTPS, and DNS protocols as they were the most informative and prevalent in the data. HTTP cookies, bad DNS requests, and SSL certificate check are examples of the features we extracted. Communication features were also extracted for each of the three application layer protocols separately (e.g., traffic volumes of HTTP traffic).

The protocol-based features and the communication features were both extracted using T-Shark (www.wireshark.com).

Domain-based features are related to the popularity, security rank, and domain category (based on the browsing activity of users). These features were extracted using third party services. The domain names of sessions were taken from the DNS requests and HTTP/S host names. Domain categorization was based on Websense Threat Seeker and Bit Defender categories. A single category was derived by grouping similar categories and integrating the information provided by both services. The domains' security scores were based on the WoT (Web of Trust) rating tool, and the popularity rank was taken from the Alexa ranking service.

Table 1. Extracted features.

Feature Category	Sub-Category	Features
Communication	Sessions	total bytes, bytes out, bytes in, duration, in/out ratio
	Packets	Aggregation of the packets' arrival time, total bytes, bytes out, bytes in, and lost packets to sessions, computed by calculating the average, median, first quarter, third quarter, minimum, maximum, standard deviation, entropy, and variance
Protocol-based	Network Layer	TTL statistics, countries (nominal feature)
	Transport Layer	Port ratios including 80, 443, 5223 (used by Google Play), and other, lost segments, retransmitted/out of order/duplicated packets
	Application Layer	HTTP cookie count, HTTPS cipher key/certificate count, bad DNS requests. Communication features calculated separately for HTTP, HTTPS, and DNS traffic.
Domain-based	Category	Nominal features: Bit Defender + Websense category, WhatsApp/ Facebook
	Security	WoT score, Webutation score
	Popularity	Alexa Rank

Notice that the feature extraction process is automatic and does not require any manual process. The information is extracted from the traffic itself and from third party services (e.g., WoT) using their API. Therefore, the process can be deployed in real life applications.

4 Venue Type Inference

4.1 Types of Hotspots

Wi-Fi hotspots can be roughly divided into private and public hotspots. Private hotspots usually implement access control and are password protected. In contrast, public

hotspots are designed to serve a wide range of casual users and are usually open or their password is readily available to the targeted population. In this study, we divide the hotspots into four main classes: home, organization, waiting, and hangout hotspots. We consider the first two as private, while the latter two are considered public.

Home hotspots (H) are used to connect multiple devices to the Internet at private residences. Usually, they are characterized by a single access point, a small number of users, and high quality service. Most home hotspots are protected with passwords.

Organization hotspots (O) represent the second class of hotspots. This class of hotspot is maintained by IT professionals employed by commercial organizations, education facilities, etc. In most cases these are restricted password protected networks which are subject to strict access controls. Organizational hotspots are characterized by high QoS and tight security policies.

Hangout hotspots (HO) are located in public venues such as bars, restaurants, shopping malls, etc. Similar to waiting hotspots, the security and QoS of hangout hotspots are not usually high.

Waiting hotspots (W) are defined as public hotspots which are located in waiting areas such as public transportation, hair salons, etc. Hotspots of this type have high user turnover during short periods of time. The Internet behavior of users in waiting hotspots is expected to be characterized by brief browsing and "time killers" such as games and social networks. The security and QoS of waiting hotspots tends to be low.

4.2 Connection Windows

We infer the type of venues based on the traffic of smartphone users connected to Wi-Fi hotspots located in the respective venues. User's Wi-Fi traffic is generally available to ISP providers, VPN, and proxy servers. These parties can use user context for commercial and security applications as described in Sect. 1, even when other information sources (such as location services) are disabled.

The general data and statistics we collect for a single session are not sufficient to determine the type of the hotspot a user is connected to. Therefore, in the following analysis we aggregate sessions in small chunks denoted as connection windows (CW) during which the smartphone is connected to the same hotspot. Since devices are disconnected and reconnected to the same hotspot within minutes (for example, due to low signal strength), we allow an idle time (up to 30 min) between consecutive sessions within the same CW. In total we identified 37,714 CWs, most of which were associated with home hotspots, as shown in Table 2.

Table 2. Distribution of CWs based on hotspot type.

Hotspot type	Number of connection windows
Home (H)	27,367
Organization (O)	7,929
Hangout (HO)	2,708
Waiting (W)	720
Total	**37,714**

We aggregated the features of the sessions within each CW in order to define CW features; the following method was used: For every numerical session feature (see Sect. 3.3), we calculated the average, median, minimal, and maximal values across all of the sessions associated with the CW. For domain category features, we created numerical features that represent the categorical value's incidence in the CW. For example, if 50 sessions occurred in a CW, of which 30 are from the "search" category and 20 are categorized as "news," the value of the "search" and "news" features for the CW is 0.6 and 0.4, respectively, and the values of other domain category features are 0. The feature aggregation process resulted in $\sim$250 numerical features, after filtering out features that provided no information.

Three additional features were defined for CWs: the number of sessions in the CW, its day of the week (e.g., Sunday), and the time of day (8am-12am, 12am-16 pm, etc.). Both the "day of the week" and the "time of day" attributes were nominal, with seven possible values each. CWs that occurred in more than one "time of day" or "day of the week" were assigned based on the majority of session start times. For example, if a CW began at 7:55am and ended at 8:10am, with the majority of the sessions starting between 7:55 and 7:59, the label would be 5am-8am. Finally, each CW was labeled with the type of hotspot.

4.3 Hotspot Type Classification

Next we evaluate a model that classifies the type of hotspot a user was connected to. The model was based on the CW data defined above. Every hotspot was associated with numerous CWs creating a diverse dataset with sufficient representation of each hotspot type. However, there were significantly more CWs associated with home hotspots than those associated with other types of hotspots. Therefore, we generated three random balanced datasets (Dataset1, Dataset2, and Dataset3), each containing 3,600 CW instances. Waiting hotspots accounted for 20 % (720 instances) of the dataset, while home, organization, and hangout CWs accounted for 26.67 % each (960 instances).

For each of the datasets we selected the best features using the correlation feature selection (CFS) algorithm with GreadyStepwise search. Then, we built classification models using the rotation forest meta-classifier with random forest as the base classifier. The accuracy of the classifiers was around 57 % with the area under the ROC curve (AUC) ranging from 0.73 to 0.90 as presented in Table 3.

Table 3. Multiclass classification results.

	Dataset1	Dataset2	Dataset3
Accuracy	57.75 %	57.67 %	58.5 %
Weighted AUC	0.81	0.82	0.83
Home AUC	0.75	0.73	0.76
Hangout AUC	0.90	0.90	0.90
Org. AUC	0.80	0.81	0.83
Waiting AUC	0.80	0.81	0.82

These basic classifiers were able to distinguish well between waiting and hangout CWs; however they often confused home and organization CWs. To solve this issue, we used the "1-vs-all" (1vsA) approach. In this approach, four different classifiers are trained. Each classifier tries to classify whether the CW is within a single hotspot type or within the group consisting of the other three hotspot types, e.g., a classifier for "home" or "other." After training these classifiers, every CW has been classified by the four classifiers and the label of the CW is determined by the model with the highest confidence. This approach showed some improvement in the results. More importantly, we noticed that it was better at classifying home and organization CWs, while it often confused between hangout and waiting CWs.

In order to combine the pros of both approaches, we employed a meta-classifier that combines the multiclass and 1vsA classifiers. This model first classifies instances using the 1vsA classifier. In case the output label is "home" or "organization", we use this label. Otherwise, we classify it using the multiclass classifier. Experiments with the meta-classifier were performed on the same datasets as the previous models. Each dataset was randomly divided into train and test groups (test groups of 100, 200, 400, and 800 instances). Table 4 summarizes the performance of the meta-classifier.

Table 4. Results of combining multiclass and 1-vs-all classifiers.

	Dataset1				Dataset2				Dataset3			
	H.	O.	HO	W.	H.	O.	HO	W.	H.	O.	HO	W.
Avg. Precision	.73	.55	.55	.63	.70	.50	.46	.67	.66	.62	.51	.70
Avg. Recall	.81	.66	.48	.50	.84	.61	.34	.57	.81	.71	.46	.51
Avg. F-measure	.77	.60	.51	.56	.76	.55	.39	.62	.73	.66	.48	.59
Accuracy	**76.5 %**				**75.1 %**				**78.1 %**			

4.4 Public Vs Private Classification

The inference of hotspot type is important, but sometimes simpler classification is preferred in order to achieve higher accuracy. For example, specifying whether the hotspot is public or private may satisfy a context oriented access control application. Therefore, we decided to classify whether a user is connected to a private (home or organization) or public (waiting or hangout) hotspot. We balanced the CW dataset by randomly removing private instances until we obtained a ratio of 50:50 between the labels. This process resulted a dataset of 4,856 instances. The classifier we created used AdaBoost with resampling and random forest algorithms. In a 10-fold cross-validation, the model's accuracy rate was 78.95 %. The ROC graph and its AUC measure are presented in Fig. 2.

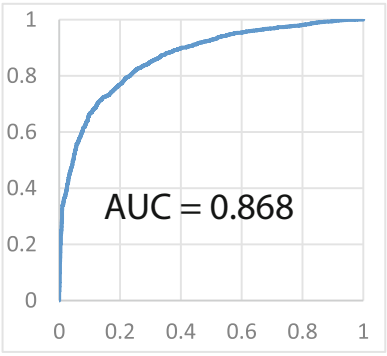

Fig. 2. ROC graph of public vs private CWs' classification.

5 Discussion

The experiments' results show that Internet behavior of users changes in different types of hotspots. Next we analyze the collected data in order to understand the nature of the differences in users' behavior in different types of venues. We define attributes of venue types similar to the CW features defined in Sect. 4.2. Specifically, we calculated the average values of all numeric session-features (e.g. WoT score) and derived the domain category incidence among all sessions associated with each type.

We depict the variability of the venue type attributes in Fig. 3. The bars in Fig. 3 present the normalized venue type attributes. For example, the incidence of search sessions is almost the same across different venue types while the incidence of entertainment sessions is larger in waiting venues than in all other venue types altogether. The relative incidences of popular domain categories (accounting for $\sim$66 % of sessions) are similar across venue types in contrast to the less popular categories. We attribute this phenomenon to the fact that popular information services became a part of the everyday life and are used always, regardless location and context. In addition a large amount of the traffic to domains in popular categories is generated by the smartphone regardless the user behavior. Therefore, the difference in user behavior across different hotspots is reflected in actions associated with less popular categories, such as playing games (entertainment category) and reading news.

In order to stress the importance of unpopular categories for hotspot type classification we present in Fig. 3 the results of the CFS algorithm. This algorithm selects a set of features that have high merit to the classification and low correlations between themselves. Unpopular categories were selected by the CFS algorithm because together they contribute to the classification. In contrast, the popular categories correlate to each-other and therefore, only one popular category was selected. In addition the differences in users' Internet behavior are reflected in the Alexa popularity rank[1],

[1] We reversed the Alexa rank such that popular domains receive high ranks.

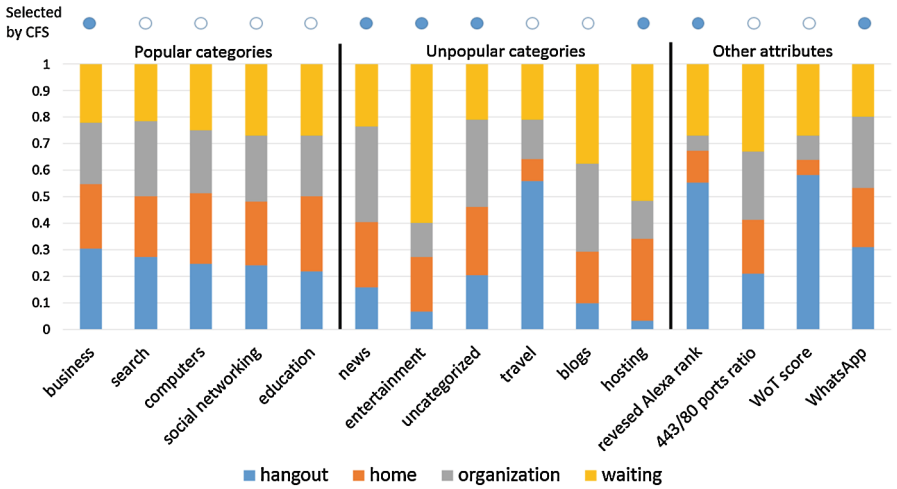

Fig. 3. Internet behavior feature selection results and comparison between hotspot types based on these features.

WoT security scores, and the use of the WhatsApp (instant messaging) application. We believe that the WoT feature was not selected by the CFS, because it correlates to the Alexa rank score, i.e., popular domains are often more secure than unpopular domains.

6 Conclusions

In this paper we study the user behavior in different types of venues and present the hotspot type classification problem. We show that venue type can be inferred from the Internet traffic of smartphone users. This type of inference can be used for advertisements, recommendations, access control and other context aware services. The prediction process can be automated as all the features.

The analysis of Internet behavior attributes shows that the majority of Internet traffic is similar in terms of domain categories and port usage. Nevertheless, users' behavior differs in access to less popular domain categories, instant messaging traffic, and in the domains' popularity and security ranks.

The subjects of the experiment were all students studying at a university in Israel; thus the data represents user behavior properties of only a segment of smartphone users. Despite the similar demographic properties of our subjects, the results show the feasibility of venue type classification based on user Internet traffic. In future we intend to expand and diversify the dataset in terms of demographics and geography.

Furthermore, we aim to classify the hotspot type using other information sources. For example, classify the hotspot type locally on the device using information that can obtained by applications (e.g., Wi-Fi connections, internet usage statistics and sensors).

References

1. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 217–253. Springer, US (2011)
2. Adults' Media Use and Attitudes Report. Ofcom (2014)
3. Afanasyev, M., Csirao, B.Q., Chen, T., Voelker, G., Snoeren, A.: Usage patterns in an urban WiFi network. IEEE/ACM Trans. Netw. **18**(5), 1359–1372 (2010)
4. Balachandran, A., Voelker, G.M., Bahl, P., Rangan, P.V.: Characterizing user behavior and network performance in a public wireless lan. In: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 195–205 (2002)
5. Manweiler, J., Santhapuri, N., Choudhury, R., Nelakuditi, S.: Predicting length of stay at WiFi hotspots. In: INFOCOM, 2013 Proceedings IEEE. Turin (2013)
6. Mark Hall, E.F.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 11 (2009)
7. Miettinen, M., Heuser, S., Kronz, W., Sadeghi, A.-R., Asokan, N.: ConXsense – context profiling and classification for context-aware access control. In: ASIACCS (2014)
8. Namiot, D.: Context-aware Browsing – a practical approach. In: 2012 6th International Conference on Next Generation Mobile Applications, Services and Technologies. Paris (2012)
9. Pentland, A.S., Aharony, N., Pan, W., Sumter, C., Gardner, A.: Funf: Open sensing framework (2013)
10. Qin, W., Zhang, J., Li, B., Zhu, H., Sun, Y.: Mo-Fi: discovering human presence activity with smartphones using Non-intrusive Wi-Fi sniffers. In: 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC) (2013)
11. The Infinite Dial 2013. Navigating Digital Platforms. Edison Reseach and Arbitron (2013)