Analyzing the quality of crowd sensed WiFi data

Carlos Pérez-Penichet, Adriano Moreira
Algoritmi Research Centre, University of Minho.
Guimarães, Portugal.

***Abstract*—The widely extended WLAN infrastructure has often been used as geographic landmark to support localization applications and human mobility studies. In these applications a WiFi network made of static nodes seems to be always assumed. Here, evidence is presented to show that this is hardly ever the case. Several independently collected datasets were analyzed to show that dynamic, moving Access Points are often present and could have a significant negative impact in this kind of applications. Additionally other irregularities in these traces are also exposed. The possible impact of these irregularities is evaluated in one specific application, Proximity Maps. The node degree distribution of Proximity Maps is studied and the influence of the proposed solution on the degree distribution is analyzed. Finally some possible simple solutions to mitigate the problem are presented.**

I. Introduction

In recent years, there has been an enormous increase in the popularity and number of WiFi networks deployed. The wide availability of such networks opens doors to a variety of applications allowing new ways of interaction among people and with the networks themselves. One of those applications is to use the WLAN Access Points (APs) as geographic clues to aid or even completely support localization services or other forms of context inference [1]–[4]. In some cases similar strategies are used as a proxy to conduct studies on fundamental aspects of human mobility [5]–[9]. In all of those cases, the underlying assumption that Access Points are immutable entities that never change and are permanently associated to a fixed physical location seems to be always made. Furthermore, it is generally assumed that the BSSID (MAC address) of these devices is universally unique and thus can be used to identify them. Here evidence is presented showing that these assumptions do not always hold.

Several independently collected datasets were analyzed in search for clues of the violation of these two common assumptions. Evidence of the presence of both of these irregularities was found in each one of the analyzed datasets, showing that the quality of this kind of data, in the sense of compliance with these two basic principles, is not necessarily guaranteed. Here, this evidence is presented along with a discussion on the possible causes for those irregularities and an analysis of their possible impact on the already mentioned applications as well as on a specific application aimed to automatically build Proximity Maps. Furthermore, some relatively simple solutions are proposed to mitigate the identified problems and an evaluation of its effectiveness is conducted for the specific case of Proximity Maps.

The rest of the paper is structured as follows: In section II, the different data sets used in this study are introduced and the structure of the data is described. Section III introduces the concept of Proximity Maps and presents examples. In section IV the evidence of data irregularities and assumption violations are presented. In section V possible causes of these problems are discussed while in section VI some solutions are proposed. In section VII the impact of the proposed solutions is evaluated for the case of Proximity Maps. Finally in section VIII conclusions and final remarks are drawn.

II. Data sources

In the past few years, our research group has been collecting data about the presence of WiFi networks in the wild, by using opportunistic collaborative sensing approaches or specific hardware/software tools. These data have been used to support research activities in the areas of indoor positioning based on WiFi fingerprinting, human mobility analysis, and place and trip learning and modeling. In this particular study, several independent data sources were analyzed. These data sources range from data collected using applications developed as part of an internal project to applications developed by colleagues to applications developed by completely independent research teams. At the same time data has been collected by volunteers as part of data gathering campaigns organized both by our team and the independent teams and colleagues. In total, data has been collected in 17 different countries in varied proportions (Figure 1).

The list of the different data sources that have been made available consists on:

1) The EPI system [10].
2) The MOVE Android application.
3) The MySteps Android application.
4) The Geo Anuncios Android application.
5) The LifeMap Android application [11], [12].

The EPI system [10] is built around an application for notebook computers running Windows that allows nearby users to exchange messages among them through WiFi networks. This system has been developed by our group to study the willingness of people to engage in collaborative sensing campaigns.

MOVE is a data logging Android application, developed by a partner research team, that also gathers observations

978-1-4799-2736-4/14/$31.00 ©2014 IEEE

about the WiFi environment on which its users are embedded. MySteps is an Android application developed in order to help analyze the mobility patterns of its users and also gathers observations in a similar fashion as MOVE does. The MySteps application has been developed by one of our partners within the context of the TICE.Mobility project aiming to model the transportation habits of people.

Geo Anuncios is an Android application to publish geo-referenced classified messages. It has also been developed by our research group in the context of a study about the impact of sensing applications on the energy consumption of smart phones. As part of its process it also collects data in the form of WiFi observations.

The LifeMap application [11]–[13] is a totally independently developed Android application to track the mobility of its users.

All five data sources continuously collect samples at more than one sample per minute. In the case of GeoAnuncios, data collection is only performed while the user is on the move in an attempt to save power.

Since each of these data sources was developed by independent parties and with different goals in mind, they each store their data in slightly different ways. To avoid problems associated with variable data representation and provide a consistent data source for the rest of the processes, a unified framework has been adopted [14].

Within this framework, the WiFi radio signature is mapped into an *observation.* Observations consist of the identification mID of the moving entity (a person, device) that made the observation, a timestamp t marking when the observation was made and a list of the Access Points visible to the client at that given moment, their SSIDs and their RSSIs among other things that are out of our interest for now. The input data is of the form:

$$O = \{(mID_1, t_1, BSs_1) \cdots (mID_{N_s}, t_{N_s}, BSs_{N_s})\} \quad (1)$$

where BSs_i is the list of APs of the i'th record. It is assumed that there are N_s records and that in the i'th record Na_i APs were detected. The lists BSs_i have the form:

$$BSs_i = \left\{ \begin{pmatrix} BSSID_{i,1} \\ RSSI_{i,1} \\ SSID_{i,1} \end{pmatrix} \cdots \begin{pmatrix} BSSID_{i,Na_i} \\ RSSI_{i,Na_i} \\ SSID_{i,Na_i} \end{pmatrix} \right\} \quad (2)$$

Each application collects other kinds of data, including GPS coordinates, optionally collected by all of the applications except for EPI. This will aid in analyzing the data irregularities.

Table I shows figures describing the various data sources. The *Edges* column shows the number of edges in the corresponding Proximity Map (Section III).

III. Proximity Maps

The aim of Proximity Maps is to infer the logical space structure by collaboratively sampling the radio landscape, specifically the WiFi networks [15]. This approach differs from the traditional two-phase fingerprinting localization methods in that there is no off-line or calibration phase. This advantage is reached at the cost of obtaining a map that only represents the space's logical structure as opposed to actual geographic relationships among places. This map is used to provide spatial context to the analysis of human mobility patterns. The provided context is intended to be used analyzing transportation efficiency.

A Proximity Graph $G_p = \{N_p, E_p\}$ is a map of the proximity of network access points, represented by the nodes N_p, based on their visibility to users through their portable and mobile devices. The main assumption used to build Proximity Maps is that if two access points are simultaneously visible to a user (device), then they must be close to each other. The nodes of the proximity graph N_p represent all APs as reported by one or more user observations. N_p is defined as:

$$N_p = \bigcup_{i,j=1,1}^{Na_i,N_s} BSSID_{i,j} \quad (3)$$

The edges of the graph E_p represent proximity between

Table I
Statistics on the various data sources.

	Users	Observations (N_s)	BSSIDs	Edges
MySteps	35	1494073	47101	769662
Move	77	7138239	102231	1192424
Geo Anuncios	118	575299	101164	106060
EPI	48	825969	5107	243377
LifeMap	12	119648	117276	1236701

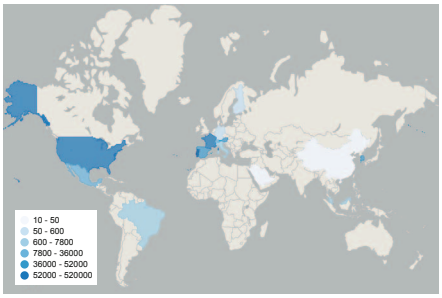

Figure 1. Geographic distribution of GPS observations.

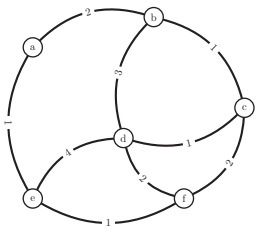

Figure 2. Conceptual example of a Proximity Map. The nodes represent the APs while the edges represent the vicinity of connected Access Points. The weight of the edges represents the number of times the connected nodes have been observed together.

pairs of nodes. An edge will be added between any pair of nodes if and only if those two nodes are ever detected simultaneously, meaning the nodes must not be too far apart. More concisely:

$$E_p = \left\{ \bigcup_i (BSSID_{i,j}, BSSID_{i,m}) : j \neq m \right\} \quad (4)$$

Figure 2 shows a simple example of a conceptual Proximity Map. Information about the number of times a given node or edge was observed by users is also stored and represented in the final graph as the data associated with the nodes and the edges respectively.

Figure 3 shows an example of a real Proximity Map. This map was generated from the LifeMap dataset, it contains the combined data of all the almost 120000 observations collaboratively gathered by 12 users. The map consists of 9325 connected components and Figure 3 shows only the largest one (so-called "giant component") which contains 80692 (69% of the total) nodes and 108267 (88% of the total) edges.

Table I shows the number of nodes (unique BSSIDs detected) and edges found in each of the Proximity Graphs generated with each of the data sets.

IV. Data irregularities

In this section, the assumption violations found are presented. The irregularities are the following: Very high node degree, mobile access points and non-unique MAC addresses.

A. Excessive node degree

When studying the node degree distribution[1] $P(k)$ of these Proximity Maps, it was found that, regardless of the data source, the degree distribution presents the long tail typical of scale-free networks [16] (Figure 4). What is more, the degree distribution of all the datasets gathered with mobile phones appear to match very closely to each other. In the case of EPI, it could be speculated that the difference may lie in the fact that the data collection with EPI is not performed in the same continued mobile fashion as the rest of the datasets. Instead EPI collects data while computers are in use and that tends to occur at intervals and from places where people remain for extended periods of time.

A least squares fit was performed to each distribution in order to estimate the power law exponent: $P(k) \sim k^{-\gamma}$. It was found that in most of the cases the exponent γ seems to be very similar regardless of the data source, except for the notable exception of the EPI dataset. Table II presents the values of γ for each case. The filters referred to in the table are discussed in Section VI, for now only the unfiltered results are relevant. Notice that, with the exception of the EPI dataset, the exponents are comparable to well known examples presented in [16].

The characteristics of these node degree distributions imply that the dynamic character of the Proximity Graphs and the irregular way in which Access Points are distributed in space could play the roles of openness and preferential attachment referred in [16] as the causes for the emergence of this kind of network.

Table II also shows the maximum node degree for each case. Notice that these figures appear to be disproportionately large.

B. Mobile APs

Every day the practice of users setting up a WiFi access point in their smart phones to be used for tethering becomes more popular. The same popularity surge is enjoyed by devices like 3G WiFi routers. Furthermore, the number of WiFi access points installed inside vehicles such as buses, trains or ferries is also raising. As a consequence, the number of APs that appear in many different locations at different times is increasing. This fact is in direct conflict with the assumption made in most of the applications that leverage WiFi networks as a geographic indicator because one implicit assumption in them is that APs do not move an thus can be used as a proxy to represent physical locations.

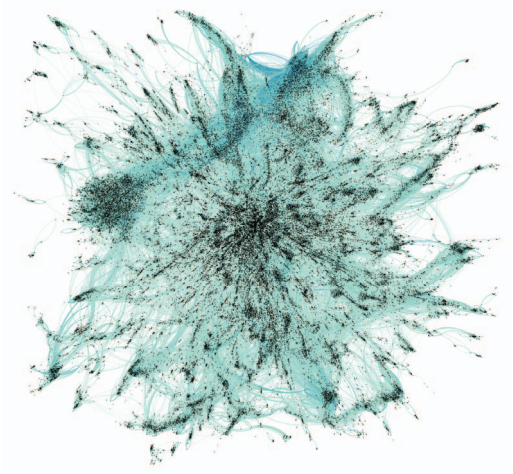

Figure 3. Example of Proximity Map from the LifeMap dataset. Only the main (giant) component is depicted here. There are other 9324 smaller components in this map. The edges are colored to reflect the degree of the connected nodes.

[1]The degree k of a node is the number of edges connected to the it. The degree distribution $P(k)$ is the probability distribution of those degrees in the network.

Table II
Characterization of the node degree distribution of the filtered Proximity Maps.

	Before filters		After filters	
	γ	max(k)	γ	max(k)
MySteps	2.4 ± 0.1	1124	2.5 ± 0.1	589
Move	2.4 ± 0.1	2592	2.5 ± 0.1	692
Geo Anuncios	2.5 ± 0.1	195	2.6 ± 0.1	180
EPI	1.2 ± 0.1	1180	1.2 ± 0.1	1137
LifeMap	2.5 ± 0.1	983	2.6 ± 0.1	885

In the case of EPI only the first filtering strategy was applied given that no GPS data is available.

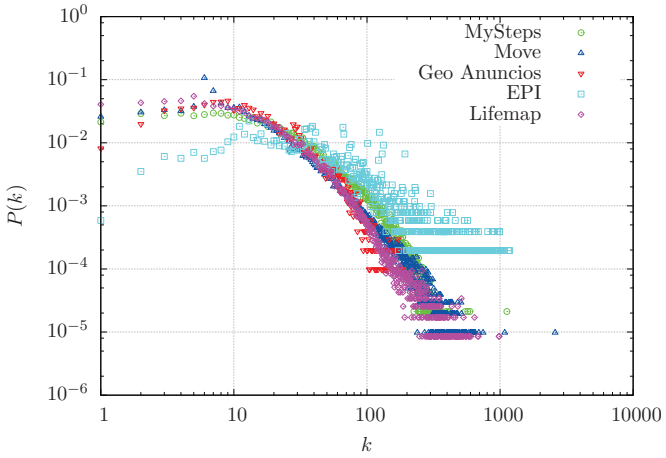

Figure 4. Node degree distribution for Proximity maps generated from different data sources.

The following analysis was performed after close inspection of the nodes with highest node degrees, the ones more likely to be mobile APs: the GPS positions collected within a short time (30 seconds wide window) of the detection of those APs were plotted on a MAP. It was noted that some of them have, in fact, been detected in different, very far places or over wide areas. This fact, paired with an analysis of those MACs' OUI[2] prefix and reported SSIDs, which lead to the identification of some of the affected MAC addresses as belonging to so-called "3G WiFi router" devices, has shown beyond any doubt that in some cases mobile APs are being detected in the present datasets. Similarly, there are many instances of SSIDs like "`Android AP`" and "`Jane's iPhone`" which are the default SSID patterns for Android and iPhone devices respectively.

Figure 5 shows one example of a mobile access point. This particular Access Point was, detected with what appears to be the default SSID ("`VodafoneMobileWiFi-2C2455`") of a 3G WiFi router.

C. Non-unique MAC address

Close inspection of the nodes with highest node degrees also lead to notice that some of them correspond to APs with particular MAC addresses. The most distinctive of which correspond to Access Points with all-zero MAC address (`00:00:00:00:00:00`). This BSSID was detected in many occasions and with dissimilar SSIDs, leading to the conclusion that somehow devices exist in different locations that carry the all-zero MAC address. There was another significant case where APs from the Linksys manufacturer (now Cisco) are reset to a default MAC address (`00:90:4c:91:00:01`) after a firmware upgrade.

By performing the same operation as described in section IV-B to correlate WiFi observations with GPS observations it was possible to verify that APs in this situation were detected in very different regions of the world. Figure 6 shows all the locations where the default MAC address set after the Linksys upgrade was detected by one or more of the users that contributed to the MOVE data set.

This kind of issue carries a problem because it directly challenges the assumption that access point MAC addresses are unique and they correspond roughly to a geographic region.

V. Discussion

The fact that the degree distribution of Proximity Maps match so well regardless of the data source constitutes a very interesting result from the point of view of complex networks, however, the maximum node degree (See Table II) reaches values that seem inordinately large. It seems intuitive to expect typical node degrees of around 20, perhaps even going up to 50 for the most densely populated areas with a lot of physical and virtual APs. Instead the degrees that are obtained go up to several hundred neighbors and beyond. This would mean that, at least in certain areas, there are many hundred access points in close proximity. This fact, compared to the scale-free characteristics previously described seems to be a contradictory result that, by itself cannot be considered a sign of poor data quality but is a sign that more inquiry is needed to make sure there is not any mistake producing these unexpected results.

It is not possible to assure that the cause of the exceptionally high node degrees discussed in Section IV-A is entirely, the combination of both the presence of Mobile

Figure 5. Locations where one of the mobile APs has been detected.

[2]Organizationally Unique Identifier

Figure 6. Locations where the "Linksys" MAC address was detected.

APs in the data, as discussed in section IV-B, and of APs with non-unique MAC addresses, however it is clear that many of the high-degree nodes do present these problems.

The proliferation of tethering the phone's 3G connection through WiFi and the popularity of 3G WiFi routers will continue to increase and thus it can be expected that mobile APs become increasingly more common.

In the case of non-unique MAC addresses, the potential problems are similar, however, it seems that this phenomenon occurs not as a result of legitimate usage of the devices but as a consequence of software errors or other problems. While this is an issue that should be taken into account when processing this kind of data it should not be expected to become a general trend.

In the particular case of Proximity Maps, the presence of these mobile APs in the data may have an important impact in the final results. From the way the edges are defined (equation 4) it can be seen that every single mobile AP would create edges to other APs in every location it would visit. This would create the possibly wrong impression that those places are all near to each other and to the mobile AP. At the same time, this problem has the potential to result in nodes with very high node degree and other distorted properties; limiting, in this way, the potential usability of the Proximity Map.

Proximity Maps are also affected by non-unique APs in a similar way that mobile APs affect them. The presence of non-unique APs in the raw data for Proximity Maps may lead to a map that erroneously represents two or more distant places as a single one.

It can be speculated that these problems are solved in solutions like the Android geolocation system, however the fingerprinting technique typically used there can be expected to be much more robust to these issues. What is more, those solutions use GPS as an out-of-band calibration source which is what being tried to avoid here.

VI. Solutions

Given the negative impact that the data irregularities so far illustrated can have in applications trying to leverage WiFi networks to do context inference it would most likely be preferable to discard any AP presenting any of these difficulties. Next, two possible strategies to automatically detect and discard affected APs are introduced.

The first strategy to identify a possible affected AP is to look for MACs that are detected with different SSIDs over short periods of time. The detection of such MACs may be an indication, although not with total certainty, of duplicated MACs. In the analyzed datasets there are indeed many cases of duplicate SSIDs that have been identified with this procedure (Table III), however the presence of multiple SSIDs alone is not a solid indication of duplicate MACs. For that reason, this strategy can be compound with the following.

The other identification criterion takes advantage of the GPS observations that are occasionally collected along with the WiFi observations (except on the EPI system). These two types of observation can be correlated to obtain a sampling of the locations where each MAC address has been detected. This information can, in turn, be analyzed and used as an exclusion criterion. For instance if the set of locations where a certain MAC address has been encountered presents an above-threshold geographic spread[3] (standard deviation of latitude or longitude larger than 5×10^{-3}), the BSSID in question could be excluded from processing. Figure 6 presents one example where this strategy would clearly succeed, but there are many other cases where this strategy would also work flawlessly. Table III shows the number of BSSIDs detected in this situation.

These two filtering strategies would, in an optimal situation, be applied each to the complete dataset under analysis. This would ensure that all nodes are checked against both criteria. However due to the computational demands that both filters impose, specially the second one, the presented results were obtained applying the filters in sequence. That is, the first strategy was applied first, resulting in the identification of all BSSIDs for which multiple SSIDs were detected. Of those, it is likely that a significant part of them constitute a legitimate SSID change; thus the second filter was then applied to all of the APs with multiple SSIDs. In the end, only those also excluded by the second filter were discarded from the final map. Table III shows the figures detailing the number of APs that were flagged by each filter for each of the datasets as well as the minimum node degree of the removed nodes, in all cases the maximum node degree of the removed nodes equals the maximum of the unfiltered graph as shown in Table II.

Another possible strategy consists on analyzing the time variability in the visibility of the neighbors of each node. If the nodes that are detected together with the given node are the same most of the time, it is to be expected that the node in question is a "normal" node whereas, if the node presents a considerable variability in neighbors, it is likely the case that the node in question is a mobile node. This strategy has not been investigated thoroughly in this opportunity, instead it will be left for future work.

It would be simple to filter out the mobile SSID patterns already mentioned using an appropriate regular expression. However this approach would lack generality as SSID

Table III
Statistics on the irregularities found in the data sets.

	Non-Unique SSID	Mobile	min removed k
MySteps	301	82	0
Move	1777	110	8
Geo Anuncios	94	25	0
EPI	101	N/A	5
LifeMap	2676	1624	30

[3]Equivalent to around $500m$ depending on the latitude.

are easily configurable by users. Furthermore, it would require manually maintaining a list of exclusion patterns.

VII. Impact

With the objective of evaluating the impact that the identification strategies have in the generated Proximity Maps beyond the results already presented in Table III, in this section the Proximity Maps of the various datasets are re-analyzed after the filtering.

After applying the filters, the power law exponent remains practically unchanged in all datasets. Table II shows the values of γ for the different datasets, before and after the filters where applied. It can be seen that there is no significant difference introduced with the filters. There are however small differences in the maximum node degree, showing that the filters eliminated some of the nodes with highest node degrees, even if the filtered maximum node degrees are still large. Table II also presents the minimum node degree of the removed nodes. These figures suggest that nodes were removed across the range of node degrees.

VIII. Conclusions

A set of data irregularities that can potentially affect applications trying to leverage the WiFi infrastructure to perform context inference and localization has been presented. Concrete examples of these problems are present using a variety of real datasets. What is more, specific examples of the occurrence of the presented issues have been shown. At the same time a particular case, Proximity Maps, was used to illustrate the effect that these irregularities could have in potential applications.

Simple solutions were proposed to avoid the exposed difficulties. An evaluation of these solutions was also presented by applying the proposed filtering schemes to the example datasets. The method showed that it can identify at some of the offending nodes with the potential to improve the final results. However, those results show (Tables II and III) that the solution have room for improvement. Those issues will be addressed elsewhere.

An interesting observation is also reported, namely the heavy-tailed node degree distribution in Proximity Maps was studied both before and after applying the filters. In both cases those distributions present remarkably similar properties among each other and they were shown not to change significantly after applying the filters. It can be speculated that these properties arise from fundamental characteristics of the underlying phenomenon, namely the openness of the network and the fact that the access points of the network are distributed in space in an irregular manner favoring more densely populated areas or places of higher interest for specific reasons.

It can be concluded that the reported problems are, in fact, present in all of the discussed dataset and they are probably widespread. Each application should consider these issues and deal with them in the appropriate manner. The proposed solutions discussed here can be seen as a modest first step towards that direction.

Acknowledgements

Research supported by FEDER funds through COMPETE and National funds through FCT – Fundação para a Ciência e a Tecnologia under project numbers 13843 and PEst-OE/EEI/UI0319/2014.

References

[1] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, pp. 775–784, 2000.

[2] J. Krumm and E. Horvitz, "Locadio: Inferring motion and location from wi-fi signal strengths," in *First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous 2004)*, pp. 4–13, 2004.

[3] A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, *et al.*, "Place lab: Device positioning using radio beacons in the wild," *Pervasive Computing*, pp. 301–306, 2005.

[4] T. Sohn, W. G. Griswold, J. Scott, A. LaMarca, Y. Chawathe, and I. Smith, "Place Lab—An open architecture for location-based computing," in *ESEC/FSE*, 2005.

[5] R. Jain, A. Shivaprasad, D. Lelescu, and X. He, "Towards a model of user mobility and registration patterns," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 8, p. 59–62, Oct. 2004.

[6] J. Hightower, S. Consolvo, A. LaMarca, I. Smith, and J. Hughes, "Learning and recognizing the places we go," *UbiComp 2005: Ubiquitous Computing*, p. 159–176, 2005.

[7] M. Kim and D. Kotz, "Extracting a mobility model from real user traces," *In Proceedings of IEEE INFOCOM*, 2006.

[8] M. Kim and D. Kotz, "Periodic properties of user mobility and access-point popularity," *Personal and Ubiquitous Computing*, vol. 11, pp. 465–479, Oct. 2006.

[9] J.-K. Lee and J. C. Hou, "Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application," in *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '06, (New York, NY, USA), p. 85–96, ACM, 2006.

[10] H. Rodrigues, M. J. Nicolau, R. J. José, and A. Moreira, "Engaging participants for collaborative sensing of human mobility," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, (New York, NY, USA), p. 729–732, ACM, 2012.

[11] J. Chon and H. Cha, "LifeMap: a smartphone-based context provider for location-based services," *IEEE Pervasive Computing*, vol. 10, pp. 58–67, Feb. 2011.

[12] Y. Chon, H. Shin, E. Talipov, and H. Cha, "Evaluating mobility models for temporal prediction with high-granularity mobility data," in *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*, p. 206–212, 2012.

[13] Y. Chon, E. Talipov, H. Shin, and H. Cha, "Mobility prediction-based smartphone energy optimization for everyday location monitoring," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, SenSys '11, (New York, NY, USA), p. 82–95, ACM, 2011.

[14] J. Peixoto and A. Moreira, "Dealing with multiple source spatio-temporal data in urban dynamics analysis," in *Computational Science and Its Applications – ICCSA 2012* (B. Murgante, O. Gervasi, S. Misra, N. Nedjah, A. M. A. C. Rocha, D. Taniar, and B. O. Apduhan, eds.), no. 7334 in Lecture Notes in Computer Science, pp. 450–465, Springer Berlin Heidelberg, Jan. 2012.

[15] Carlos Pérez-Penichet, Ângelo Conde, and Adriano Moreira, "Human mobility analysis by collaborative radio landscape observation," in *Proceedings of the AGILE'2012 International Conference on Geographic Information Science*, (Avignon, France), pp. 153–158, Digital Editions, Apr. 2012.

[16] A.-L. Barabási, R. Albert, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A: Statistical Mechanics and its Applications*, vol. 272, no. 1, p. 173–187, 1999.