Approaches and techniques for analysing WiFi location data

Clément Roux, James Little, John McAuley

Trinity College Dublin, ThinkSmarter, Dublin Institute of Technology
roux.cl@gmail.com, james@thinksmarter.io, john.mcauley@dit.ie

Abstract. Cities have dense WiFi networks that enable access to the web on a range of different devices. Although the data collected from these networks is anonymised, when aggregated, it provides a rich illustration of how cities, public places and urban spaces are used. However, making use of this data can prove challenging as the data is rarely labelled and the location accuracy can be questionable. Given these constraints, we have developed a set of methods that can be applied to anonymous WiFi location data. Through two use-cases, we show successful examples of each method, describe insights and intuitions along with an explanation of their application.

1 Introduction

The objective of our work is twofold. First, we investigate a range of techniques that can be applied to anonymous WiFi data and second, we seek to better understand how these techniques can support a range of different use cases. To do this, we are working closely with a team of WiFi engineers that develop WiFi networks for large indoor and outdoor installations. For example, the team work closely with a conference centre in Barcelona to establish how best to utilise WiFi networks during their events while at the same time, working alongside Dun Laoghaire Library in Dublin to see how WiFi can better support their patrons. In each situation, the datasets have specific constraints that can only be alleviated through better network configurations or through the deanonymisation of data.

Collected WiFi signals result in spatiotemporal data that combines geographic coordinates with a timestamp and mac address of a specific device. Often this data is gathered through a process of passive probing in which the access point connects momentarily to a device, identifies the device and then drops the connection. Our interest lies in maximising the potential of this data so that it can support a range of different applications such as urban planning, crowd management, public security, transportation optimisation and advertisement.

In this paper, we present two use cases in which we use WiFi data to address a set of common problems. The first investigates how WiFi data can be used for group identification from data gathered during the FA cup final at Wembley stadium. The second illustrates a "location accuracy" analysis performed on a dataset gathered during a conference at the Fira conference centre in Barcelona.

2 Related Work

Given the use cases addressed in this paper, we divide the related work between group identification and accuracy analysis.

2.1 Group Identification

Due to the constraints of working with unlabelled data, a lot of research on group identification use unsupervised techniques such as DensityJoin-Cluster [4] or graph techniques with spatial connectivity features [10]. The difficulty with this work is that although the algorithm will find groups, the groups may not necessarily be meaningful or useful. Rarely, are enough labels collected to conduct a meaningful analysis with supervised methods. Robol et al. use SVMs to detect crowds but this is a simple approach focused on finding whether a group is present or not - essentially binary classification on presence/absence [12]. Ruiz-Ruiz et al. employ local experts to label 5% of the data and then label the remaining 95% using a Bayesian Network [13]. The result is evaluated against national statistics, which provides some indication of success, however, the method requires a manual process and the need for a local expert (a category of individual that is context specific). Nie et al. extract the spatiotemporal transition features to predict place attributes [11]. Their assumption is based on flows - how people traverse a thoroughfare illustrates attributes of place. However, again, the approach cannot be applied to large crowds or events. Similarly, Zeng et al. classify a shoppers' state (walking vs standing, fast vs slowly, inside vs outside) based on flows obtained from indoor WiFi signals but the approach requires significant signal accuracy and relies on a small indoor experimental setup [21]. Meneses et al., on the other hand, study motion and flows using spatial connectivity features on different campuses [10]. The analysis is interesting as it does not rely on labels but looks at how different features suggest different spatial relationships. A similar approach, based on trajectory similarities, is described in [8]. Their evaluation techniques include a set of visual and intuitive heuristics for spatiotemporal data analysis.

2.2 Accuracy Analysis

Indoor and outdoor WiFi probe requests along with their accuracy indicators such as RSSI are often used for location-based analytics. Implementation of these techniques is cost effective and their accuracy is relatively reliable [9]. However, the majority of research focuses on datasets collected from indoor locations [20], as outdoor WiFi configurations are more challenging to implement effectively and can result in noisier data [18]. Similarly, there has been a lot of work in combining smart-phone sensor data, such as the GPS, Bluetooth, acoustic data [14], accelerometer and magnetic field data [2] in an attempt to generate better location accuracy. However, there is an additional cost when implementing denser networks and, generally, these points do have restricted access to the devices' information. Standard WiFi networks, on the other hand, remain

the least restrictive and cheapest solution, because of their ability to passively probe closely located devices. There are, of course, numerous applications for this data, such as queuing time prediction [15] or zone occupancy tracking [5, 17]. Location-based analytics, group identification and accuracy analysis often use data visualisation as a way to analyse behaviour, identify group structures and measure signal propagation [9, 16, 3, 6].

3 Use Case 1: Group identification at stadium events

The ability to identify and segment groups has a range of use cases such as targeted advertising, security management and urban planning. However, many existing approaches are based on data in which no clearly delineated or a priori groups exist, nurses versus patients for example [13]. In contrast, the 2017 FA cup Final event was organised so that supporters from each of the two teams were provided with their own fan zone outside of the stadium. Arsenal's fan zone was in the Arena square and Chelsea's fan zone was in the Event's pad. Also, the supporters arriving from the Olympic way (the main thoroughfare that brings supporters to the stadium) were separated so that each side of the avenue was dedicated to the supporters from one particular team (see figure 1). Out aim was to assess how we could identify these groups based on their behaviours at different points during the event and verify this identification using the fan-zones and the Olympic way.

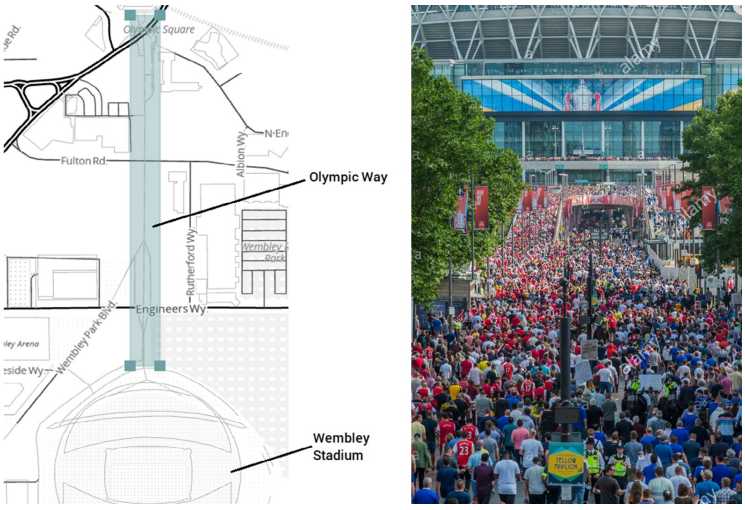

Fig. 1. Map showing Olympic way going to the main entrance and a picture of the Chelsea (blue) and Arsenal (red) supporters heading to the stadium before the FA cup final match, the 27th of May 2017. Credit: Guy Bell/Alamy Live News

3.1 Data Description

Each record in the dataset is the result of a triangulation process applied on multiple probe requests initiated by a single device. Each data point has a MAC address, a time stamp, a set of geographical coordinates (latitude and longitude), a geographical zone and a confidence factor (CF) related to the RSSI value. The CF is a distance representing half the side of a square of which the centre is the data point itself. This square represents the point's uncertainty surface, which shows where the device is located within a 95% confidence interval. The zones and their respective locations are shown in figure 2.a.

3.2 Methodology

We applied a set of unsupervised techniques for group identification and used a combination of analysis and visualisation for verification. Because the network covers a wide area, which can itself impact accuracy, we avoid using a data points' location estimations directly and instead focus on associated devices – two or more devices collocated (same place, same time) and their relationship with either fan-zone. Associated devices are defined by the number of shared locations of two or more devices - the larger the number of shared locations, the greater the similarity between devices. Similarly, the association between two zones can be defined as the number of shared devices - the greater the number of shared devices, the more similar the zones. In both definitions, we assume that the supporters of the same team spend more time in similar locations than supporters from the opposing team. A group of supporters can be considered as a group of devices that share the same locations at the same time throughout the event. By the same reasoning, a group of subzones sharing the same devices throughout the event will likely form a bigger zone dedicated to a team. This

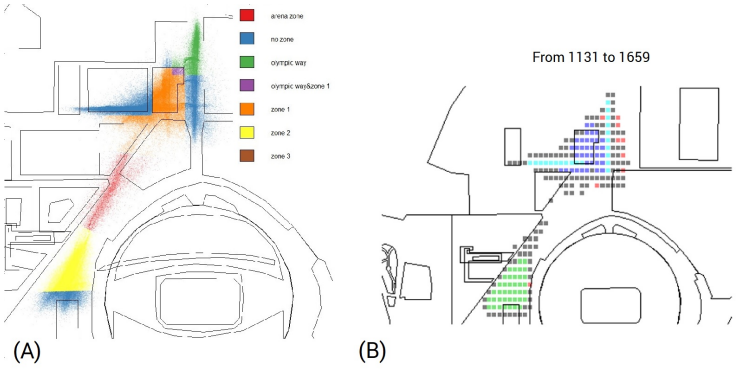

Fig. 2. (A) Raw data points split by their original zones. (B) Four groups of grid-cells resulting from the hierarchical clustering of the transformed grid-cells adjacency matrix between 11h31 and 16h59, taking into account the most 25% accurate devices' location and the 20% strongest links between pairs of grid-cells.

approach yields two adjacency matrices, the first, or grid-cell matrix, pools sub-locations together based on the number of shared devices, whereas the second, or devices-based matrix, pools devices together based on shared locations.

Clustering the grid-cell matrix: This method seeks to find groups of grid-cells based on the proportion of shared devices between each pair of grid-cells. In other words, we cluster grid-cells that have a large proportion of collocated devices. The binary matrix (grid-cells X devices) is multiplied against its transpose to generate a square (grid-cells X grid-cells) adjacency matrix, which represents the number of shared distinct devices between each pair of grid-cells. Each cell of the square matrix is then divided by the total number of distinct devices that were identified at least once in either one or both grid-cell locations. This converts the absolute-based matrix into a proportion-based matrix where each proportion represents the percentage of distinct shared devices against the total number of distinct devices between the corresponding pair of grid-cells. We assume the bigger the proportion, the closer the pair of grid-cells and vice-versa. We applied two techniques. First, we convert the proportion matrix into a distance matrix, apply Ward's hierarchical clustering to the distance matrix [19] and reduce the tree to the desired number of groups (3). Second, a fast-greedy modularity optimization algorithm was applied to the proportion adjacency matrix. The latter interprets dense sub-graphs as communities. The more relationships between a group of grid-cells, the more likely to be interpreted as a community by the algorithm [1].

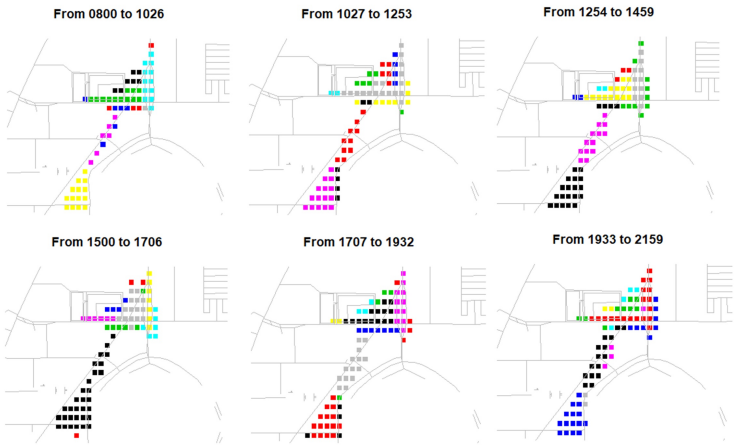

Fig. 3. Groups resulting from the fast-greedy modularity optimization algorithm applied on the adjacency graphs of shared devices between pairs of grid-cells (20 x 20 grid). The colours are not consistent in time. Only 75% of the most accurate data points (Confidence Factor first 0.75 quantile) from 8:00 to 22:00 with a minimum number of devices per grid-cell equal to its median.

Clustering the device matrix: This method consists in finding groups of devices based on the proportion of shared locations. As previously, the adjacency matrix is divided by the square matrix of weights containing the unique number of locations shared between devices. The resulting matrix represents the percentage of distinct shared locations against the total number of distinct locations where the corresponding pair of devices were seen. We assume the greater the number of shared locations between two devices, the greater the likelihood that both devices belong to the supporters of the same team. We apply Ward's hierarchical clustering on the distance-matrix. Unlike the groups of grid-cells, all the devices cannot be directly mapped on a 2D plane as it would suffer from occlusion.

3.3 Results

Each approach reveals clusters that reflects the topological configuration of the event. As illustrated in figures 2.b, 3 and 4 both fan zones, the Olympic way (vertically positioned) and Engineers way (horizontally positioned) are clearly visible. Four distinct zones are evident in figure 2.b, Chelsea's and Arsenal's fan zones appear in green and blue. In addition, three vertical zones (light blue, red and grey) appear on Olympic way illustrating different groups separated by Police. The light blue and red areas represent Arsenal supporters and Chelsea supporters respectively. The grey zone is most likely noise. Figure 3 shows the event as a set of small multiples with each image representing a time span between 8:00 am to 21:59 pm. Each time span has its own colour signature expressing dense and high levels of shared devices between grid-cells. The same colour between

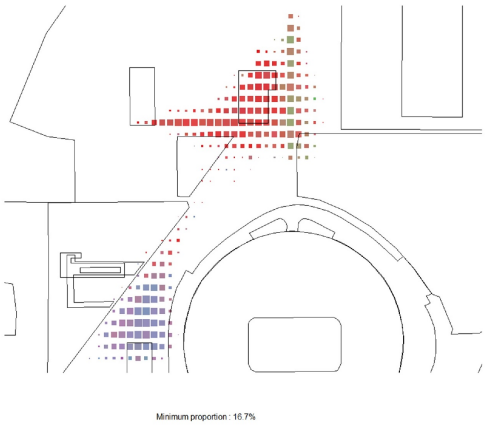

Fig. 4. Four groups of devices resulting from the hierarchical clustering of the transformed devices adjacency matrix between 11h00 and 16h59, taking into account the most 25% accurate devices' location and the 20% strongest links between pairs of devices. The groups' proportions are encoded with colours and transparency. The size of the grid-cells represents the number of distinct devices per location.

two small multiples does not represent the same group. As can be seen, both fan zones are always defined with different colours and the Olympic Way is separated vertically in two from 10:27 to 21:59. This supports our assumption that clustering the adjacency matrix can reveal subzones matching the environment's properties.

3.4 Discussion

As evident from the visualisation, we can clearly identify groups that reflect the order and composition of crowds at the event. The method is flexible and includes several parameters such as the number of groups and the size of the grid-cells that can be used to address a range of different scenarios. The difficulty, however, is that the approach may have limitations when applied to other contexts in which groups are not as strongly delineated.

4 Use Case 2: Accuracy analysis on large indoor spaces

The geographic location of a device is estimated using a triangulation of locations from several access points and is expressed as a measure of uncertainty in regard to that estimation [7]. There are multiple parameters that can impact this accuracy such as network configuration, number of access points, whether the connection area is outside and the types of access points (in this case it is Cisco Meraki). Comparative methods that address accuracy have focused on unipersonal data points with trajectories using a ground truth [9, 20]. However, there has been little work on developing and evaluating techniques based on aggregated data from multiples devices. Generally, zones, or locations that exist within a specific geographic area, are used to support location based analytics. Often the topographical configuration of the building, such as a room, a floor, a street portion, a fan zone, a road intersection, are used to define the zones. This approach, however, can hinder the accuracy of an estimation and is often not the most optimal way to configure the network. Zone analysis can evaluate a network and possibly suggest a better configuration.

4.1 Data Description

The dataset was collected during the ESC congress 2017 which took place in the Barcelona Fira conference centre. As can be seen in figure 5, the centre is a multi-room building of which we focus on the two biggest zones called hall 2 and hall 3. The dataset consists of 48 hours of data collected from the 26th to the 27th of August 2017.

4.2 Methodology

Two approaches are used for conducting accuracy analysis. First, data visualisation is used to plot the levels of accuracy on a map. Second, the confidence factor of each data point is used to measure the effectiveness of a zone configuration.

Data visualisation has been used as a tool for analysing subzones' accuracy [16]. In this work, we used level plots, with bilinear interpolation, to map confidence factors on a 2D plane. This enabled us to visualise the levels of uncertainty in the locations' estimation. Hall 2 and Hall 3 zones are randomly sampled down to approximately 10 thousand records each and used to compute the corresponding level plots. Figure 6 shows the Hall 2 and the Hall 3 confidence factors' level plots. The greener the sub-zone, the more accurate the location's estimation and the more likely the point's estimation to be located there. The redder the subzone, the less accurate the location's estimation. Both level plots show similar levels of accuracy of between 0 and 50 meters. However, both halls' border subzones have a higher level of inaccuracy of up to and between 100 to 150 meters.

The zone's accuracy indicator is a probabilistic measure of a zone's ability to match each point's inner and outer states between the points' location estimation and its unknown ground truth. To reflect reality, a zone's estimated inner-points

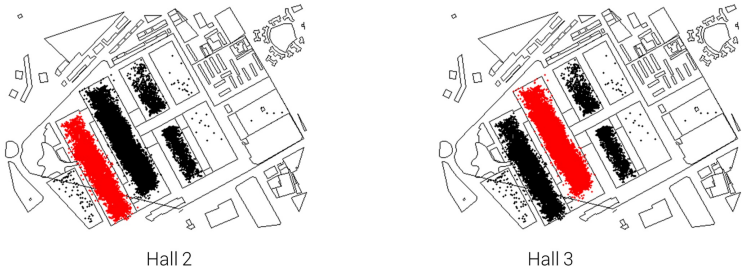

Fig. 5. The red data points are located inside Hall 2 and Hall 3 of the Barcelona Fira Gran Via conference centre. The black points are located outside the Hall 2 and the Hall 3.

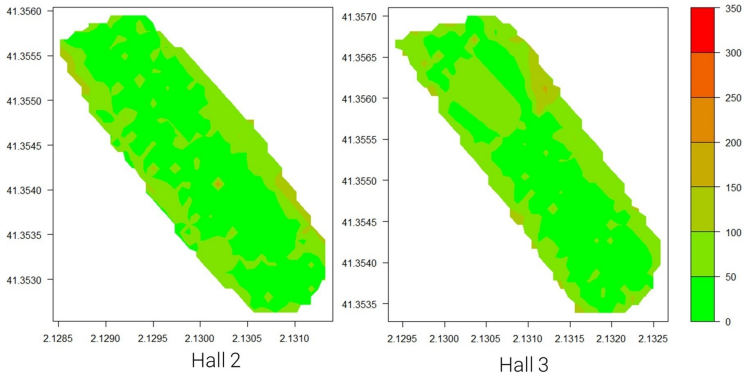

Fig. 6. Level plots of the Fira Barcelona Gran Via - Hall 2 and Hall 3 - over 48 hours of conference

should maximise their probability of being real inner-points, and the estimated zone's outer-points should maximise their probability of being real outer-points (1).

$$max(ZoneAccuracy) = \begin{cases} max(\overline{P(InsideZone|ZoneInnerPoint)} \\ max(\overline{P(InsideZone|ZoneOuterPoint)} \end{cases} \quad (1)$$

Based on this observation, we assume that the accuracy of a zone is closely related to its inner and outer points' average probability to be located inside the zone. More precisely, the bigger the zone's inner-points average probability and the smaller the zone's outer-points average probability, the better the accuracy and the more effective the zone's configuration.

We define uncertainty square as the surface where an estimated point is located within a 95% level of confidence. Two points' uncertainty squares are illustrated in figure 7. As illustrated in figure 7, 60% of the inner-point's uncertainty square resulting from its confidence factor is located inside the zone. This can be interpreted as a 60% likelihood of being a real inner-point and 40% likelihood of being a real outer-point. Similarly, 35% of the outer-point's uncertainty square resulting from its confidence factor is located inside the zone. This can be interpreted as a 35% likelihood of being a real inner-point (instead of an estimated outer-point) and 65% likelihood of being a real outer-point.

Finally, we assume a zone's aggregated accuracy indicator can be defined as the difference between both inner and outer points' average probabilities (2). The more accurate a zone configuration, the bigger the average probability of its inner-points to be located inside the zone and the smaller the average probability of its outer-points to be located inside the zone.

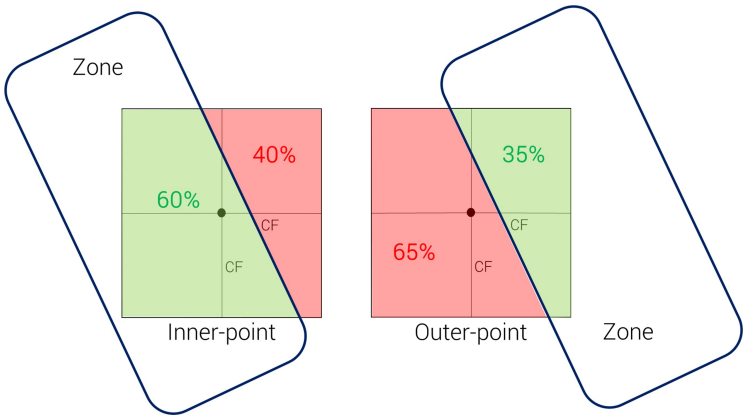

Fig. 7. The estimated inner-point has 60% chance to correctly be defined as an estimated inner-point and 40% to wrongly be defined as an estimated outer-point. The estimated outer-point has 65% chance to correctly be defined as an estimated outer-point and 35% chance to wrongly be defined as an estimated inner-point.

$$zoneAccuracy = \overline{P(InsideZone|ZoneInnerPoint)} - \overline{P(InsideZone|ZoneOuterPoint)} \quad (2)$$

Because inner and outer points are computed separately, data points from all zones are required. First, the data points identified in different floors are removed and then random sampling is performed on the dataset so that approximately 20 thousand data points were used in the analysis.

After every points' uncertainty squares are created, the proportion of each square's surface located inside the selected zone is computed.The bigger the inner point's surface being located inside the zone and the lower the outer point's surface being located inside the zone, the more likely the estimation to be true. The average uncertainty squares' proportions of inner points and outer points located inside the zone are computed separately. As can be seen in figure 8, a large proportion of inner-points' squares are located inside the zone and a large proportion of outer-points' squares are not located inside the zone. Finally, the aggregated zone's accuracy indicators are computed (2). The bigger the zone's accuracy indicator, the more likely the estimated points' locations to be truly inside the zone, and vice versa.

4.3 Results

Both halls' confidence factors' level maps (figure 6) provide an indication about the uncertainty of the location estimation. The aggregated accuracy indicator

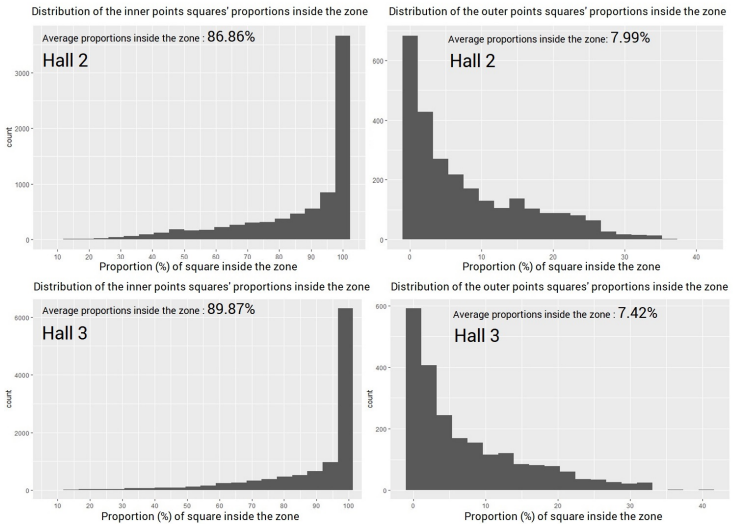

Fig. 8. Both distributions of inner and outer points squares' proportions located respectively inside the Hall 2 and the Hall 3.

was computed for each hall. Based on this metric, Hall 3, with an accuracy measure of 82.45, outperforms Hall 2 with an accuracy measure of 78.87. This measure provides a way to assess the effectiveness of zone configuration.

5 Summary and future work

We described a set of techniques that can support a range of use cases when working with location-based WiFi data. Although we have focused on unsupervised techniques and visualisation methods for verification, we are currently collecting labelled data to quantitatively evaluate these techniques. In future work, we aim to advance these methods using labelled data.

Acknowledgement. This research was supported by the CONNECT centre - Science Foundation Ireland for Future Networks and Communications.

References

1. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E 70(6) (Dec 2004), http://arxiv.org/abs/cond-mat/0408187, arXiv: cond-mat/0408187
2. Du, H., Yu, Z., Yi, F., Wang, Z., Han, Q., Guo, B.: Group mobility classification and structure recognition using mobile devices. In: Pervasive Computing and Communications (PerCom), 2016 IEEE International Conference on. pp. 1–9. IEEE (2016), http://ieeexplore.ieee.org/abstract/document/7456523/
3. Kim, J., Zheng, K., Ahn, S., Papamanolis, M., Chao, P.: Graph-based Analysis of City-wide Traffic Dynamics using Time-evolving Graphs of Trajectory Data. In: Australasian Transport Research Forum (ATRF), 38th, 2016, Melbourne, Victoria, Australia (2016), http://atrf.info/papers/2016/files/ATRF2016_paper_166.pdf
4. Kjærgaard, M.B., Wirz, M., Roggen, D., Tröster, G.: Mobile sensing of pedestrian flocks in indoor environments using WiFi signals. In: 2012 IEEE International Conference on Pervasive Computing and Communications. pp. 95–102 (Mar 2012)
5. Kontokosta, C.E., Johnson, N.: Urban phenology: Toward a real-time census of the city using Wi-Fi data. Computers, Environment and Urban Systems 64, 144–153 (Jul 2017), http://linkinghub.elsevier.com/retrieve/pii/S0198971516300928
6. Landesberger, T.v., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G., Kerren, A.: MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering. IEEE Transactions on Visualization and Computer Graphics 22(1), 11–20 (Jan 2016)
7. Liu, H., Darabi, H., Banerjee, P., Liu, J.: Survey of Wireless Indoor Positioning Techniques and Systems. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 37(6), 1067–1080 (Nov 2007), http://ieeexplore.ieee.org/document/4343996/
8. Liu, S., Wang, S.: Trajectory Community Discovery and Recommendation by Multi-Source Diffusion Modeling. IEEE Transactions on Knowledge and Data Engineering 29(4), 898–911 (Apr 2017), http://ieeexplore.ieee.org/document/7779106/

9. Mathisen, A., Sørensen, S.K., Stisen, A., Blunck, H., Grønbæk, K.: A comparative analysis of Indoor WiFi Positioning at a large building complex. In: 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN). pp. 1–8 (Oct 2016)
10. Meneses, F., Moreira, A.: Large scale movement analysis from WiFi based location data. In: International Conference on Indoor Positioning and Indoor Navigation (IPIN). pp. 1–9. IEEE (2012), http://ieeexplore.ieee.org/abstract/document/6418885/
11. Nie, S., Das, A., Gabrilovich, E., Lu, W.L., Mazniker, B., Schilling, C.: STEPS: Predicting place attributes via spatio-temporal analysis. arXiv preprint arXiv:1610.07090 (2016), https://arxiv.org/abs/1610.07090
12. Robol, F., Viani, F., Polo, A., Giarola, E., Garofalo, P., Zambiasi, C., Massa, A.: Opportunistic crowd sensing in WiFi-enabled indoor areas. In: 2015 IEEE International Symposium on Antennas and Propagation USNC/URSI National Radio Science Meeting. pp. 274–275 (Jul 2015)
13. Ruiz-Ruiz, A.J., Blunck, H., Prentow, T.S., Stisen, A., Kjærgaard, M.B.: Analysis methods for extracting knowledge from large-scale WiFi monitoring to inform building facility planning. In: 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom). pp. 130–138 (Mar 2014)
14. Sen, R., Lee, Y., Jayarajah, K., Misra, A., Balan, R.K.: GruMon: Fast and Accurate Group Monitoring for Heterogeneous Urban Spaces. In: Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems. pp. 46–60. SenSys '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2668332.2668340
15. Shu, H., Song, C., Pei, T., Xu, L., Ou, Y., Zhang, L., Li, T.: Queuing Time Prediction Using WiFi Positioning Data in an Indoor Scenario. Sensors 16(11), 1958 (Nov 2016), http://www.mdpi.com/1424-8220/16/11/1958
16. Tervonen, J., Hartikainen, M., Heikkila, M., Koskela, M.: Applying and Comparing Two Measurement Approaches for the Estimation of Indoor WiFi Coverage. In: New Technologies, Mobility and Security (NTMS), 2016 8th IFIP International Conference on. pp. 1–4. IEEE (2016), http://ieeexplore.ieee.org/abstract/document/7792436/
17. Vattapparamban, E., Çiftler, B.S., Güvenç, İ., Akkaya, K., Kadri, A.: Indoor occupancy tracking in smart buildings using passive sniffing of probe requests. In: Communications Workshops (ICC), 2016 IEEE International Conference on. pp. 38–44. IEEE (2016), http://ieeexplore.ieee.org/abstract/document/7503761/
18. Wang, J., Tan, N., Luo, J., Pan, S.J.: WOLoc: WiFi-only Outdoor Localization Using Crowdsensed Hotspot Labels. In: Proc. IEEE INFOCOM. Atlanta, GA, USA (2017), http://www3.ntu.edu.sg/home/sinnopan/publications/[INFOCOM17]WOLoc%20WiFi-only%20Outdoor%20Localization%20Using%20Crowdsensed%20Hotspot%20Labels.pdf
19. Ward, J.H.: Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58(301), 236–244 (1963), http://www.jstor.org/stable/2282967
20. Yassin, M., Rachid, E., Nasrallah, R.: Performance comparison of positioning techniques in Wi-Fi networks. pp. 75–79. IEEE, Al Ain, United Arab Emirates (Nov 2014), http://ieeexplore.ieee.org/document/6987565/
21. Zeng, Y., Pathak, P.H., Mohapatra, P.: Analyzing Shopper's Behavior through WiFi Signals. In: Proceedings of the 2nd workshop on Workshop on Physical Analytics. pp. 13–18. ACM Press (2015), http://dl.acm.org/citation.cfm?doid=2753497.2753508