Using Graphical Features To Improve Demographic Prediction From Smart Phone Data

Syeda Akter
School of Electrical Engineering & Computer Science
Washington State University
P.O. Box 642752
Pullman, Washington 99164-2752
+1-509-332-9264
selina.akter@email.wsu.edu

Lawrence Holder
School of Electrical Engineering & Computer Science
Washington State University
P.O. Box 642752
Pullman, Washington 99164-2752
+1-509-335-6138
holder@wsu.edu

ABSTRACT

Demographic information such as gender, age, ethnicity, level of education, disabilities, employment, and socio-economic status are important in the area of social science, survey and marketing. But it is difficult to obtain the demographic information from users due to reluctance of users to participate and low response rate. Through automated demographics prediction from smart phone sensor data, researchers can obtain this valuable information in a nonintrusive and cost-effective manner. We approach the problem of demographic prediction, namely, classification of gender, age group and job type, through the use of a graphical feature based framework. The framework represents information collected from sensor networks as graphs, extracts useful and relevant graphical features, and predicts demographic information. We evaluated our approach on the Nokia Mobile Phone dataset for the three classification tasks: gender, age-group and job-type. Our approach produced comparable results with most of the state of the art methods while having the additional advantage of general applicability to sensor networks without using sophisticated and application-specific feature generation techniques, background knowledge and special techniques to address class imbalance.

CCS Concepts

• Computing methodologies → Machine learning → Supervised learning • Human-centered computing → Ubiquitous and mobile computing → Smartphones

Keywords

Graph mining, feature extraction, demographic prediction, graph representation.

NDA'17, May 19, 2017, Chicago, IL, USA
 ISBN 978-1-4503-4990-1/17/05…$15.00
DOI: http://dx.doi.org/10.1145/12345.67890

1. INTRODUCTION

We envision a framework based on graphical features that will collect data from wireless sensor network (WSN) applications, represent the data as a graph, extract graphical features, add these to a typical set of non-graphical features, and improve the prediction performance of corresponding applications. Most WSN applications focus on monitoring living beings, sensing object interaction and tracking locations visited. Previous work approached predicting activity recognition from smart home motion sensor data. [1] They represented the motion sensor data as a graph, extracted graphical features from it, and classified activities performed by residents. The approach achieved significant improvement in classification accuracy compared to the benchmarks in the area. As part of our goal of evaluating the use of graph representations and graph mining to improve performance on recognition and prediction tasks for WSNs, in this paper we represent smart phone sensor data as a graph for the task of demographic classification.

Inferring demographic characteristics has been of great interest both in academia and industry. In social sciences, demographic characteristics are a fundamental building block for analysis. In business, demographic information plays a key role for user profiling and customer-specific behavior targeting for marketing, advertising and personalization of online products and phone applications. The surveys are done by randomly selecting a sample population and by asking questions about the issue under research. Adding these socio-demographic questions to the survey lengthens the survey and increases the dropout rate. Only 18% of the population, who were approached, responded to survey in [2]. Many efforts have been made to design responsive surveys and online surveys [3]. Those approaches avoid the pitfalls of missing data, unacceptable responses, duplicate submissions and web abuse, but they still need active participation from users. The ability to perform the data collection and predictions without users' active participation and without interference with questions will substantially increase demographic data availability to social scientist researchers and marketing experts. [4]

Smart phones are increasingly becoming part of an individual's daily living and social life. They can collect sensor data related to social communication such as call and message logs, contacts and calendar; movement and location related data such as accelerometer and Global Positioning System (GPS); smart phone

app usage; physical proximity to other devices; and many other types of sensor data in an unobtrusive and cost-effective manner. As a result, it can be used as a tool for long-term large scale studies eliminating challenges of physical or virtual surveys such as researcher's intervention, outdated data, small and non-representative samples and low response rate. We recognize that GPS information is private information and users might not be willing to provide it. But for users who are willing to provide GPS information we can utilize it to avoid further inconveniencing the users by requesting participation in many potential surveys including demographic surveys.

In our experiment, we use Nokia Mobile Dataset that collected many different types of mobile phone sensor data over the duration of one year in Lausanne, Switzerland. [14] In our work, we choose only one type of sensor information, Global Positioning System (GPS), that collects users' location information over time, and we use it to infer demographic characteristics of individuals, namely, gender, age group and job type. We represent GPS information as a graph, extract graphical features from the graph representation, select useful features, and train a support vector machine to perform the classification tasks. Our approach improves performance over baseline methods. Most existing approaches focus on employing all types of sensor data and numerous measures, complicated multi-layer feature construction methods, feature selection, and adjustment of the classifications based on background information. Our approach outperforms most of the benchmarks in the field of demographic prediction from sensor data while using only one type of generic sensor data through leveraging graph structure and movement patterns.

2. RELATED WORK

Several approaches have been developed for demographics classification from mobile phone data. Mo et al. [5] approached the Nokia Mobile Phone Data challenge using contextual feature construction from the raw data and feature selection. While this work has a high classification accuracy, the number of initial features constructed from raw data is huge, more than tens of thousands. Several layers of complicated feature construction have been used on all types of data such as accelerometer, application usage, bluetooth; use of mean and variance of continuous variables such as the time length of calling, ratio of top-5 user-specific independent discrete variables such as the number of places visited by the user; frequency of common discrete variables such as frequent applications; and finally, specific features designed for specific tasks, such as the average number of accelerometer records for gender and the number of in-calls for job type. Temporal condition based probabilistic features have been computed for all these types of actions. General statistical and temporal conditions are considered together to generate numerous sophisticated feature sets. They also adjusted the prediction result based on the background information that there are dependencies among subtasks of demographic classes.

Nadeem et al. [6] used mean, variance, frequency and many other measures to produce thousands of features employing call log, visited GSM cells information, visited Bluetooth devices, visited Wireless LAN devices, accelerometer data, and so on. They have used SVM and Random Forest with bagging scheme over all these features.

Brdar et al. [7] used Relief for scoring features from the huge dataset of smart phone sensor information and then represented the data as a graph where each node is a user. Standard KNN and mutual KNN graph has been used with cosine similarity of feature vectors as a measure of similarity among users. They also tried RBFN and random forests. For different tasks, different methods gave better performance, e.g., random forest for gender and marital status, and KNN + Feature Selection for job type. All of the above works use leave one out as evaluation criteria.

Dong et al. [8] presented a basic correlation analysis between call log and sms based network characteristics and user demographics to see the relationship among populations of different gender and age group. Based on this analysis, the prediction tasks of gender and age have been formulated through the use of call and sms logs along with leveraging interrelations between gender and age. Ying et al [9] uses and computes 45 features related to users' behavior and environment, e.g, maximum movement in a location, kind of application usage per day, call and sms related features, kind of bluetooth device detected per day etc. Along with these features, Ying et al.'s work focused on solving the imbalanced class problem in the data by using a multi-level classification technique. In Koppel's work [10], bloggers' writing styles are used to predict their actual gender and age information. However, only 8% of Internet users write blogs [11]. The majority of Internet users browse news, products, or other webpages, which provides a large number of web-page click-through log data. [12] This method works for people who do a lot of online browsing, but for people who infrequently browse online, we can try to predict their demographics through smart phone sensor data. Using both results together can increase the accuracy even more. In [13], Weber shows that using demographic information has a potential to improve the state of the art web search results, especially for difficult queries, and that it leads to improvements in query suggestions.

In "Limits of Predictability in Human Mobility"[20], Song et al. seek the interplay between the regularity leading to predictability and randomness leading to un-foreseeability in human movements. As one of the important measures of predictability they used the probability that an appropriate predictive algorithm can predict correctly the user's future whereabouts. From statistical evidence of combination of Entropy and Fano's inequality the authors find that user mobility has 93% predictability on average. Furthermore, despite the variability of travel behavior that demographics, age, population density and variability in predictability is small. Authors found that no user's predictability was below 80%. This result indicates that our daily mobility is characterized by regularity and it might be possible to build accurate predictive models for processes driven by human mobility.

Montjoye et al. showed [21] that the uniqueness of human mobility traces is high and that mobility datasets can be re-identified using only a few outside locations. The authors claimed that four spatio-temporal points are enough to uniquely identify 95% of the individuals and therefore, individual's privacy concerns should be considered while designing and making policies for mobility data collection. In our work we use the GPS data that the users agreed to reveal. As described in our experiment and results section later, we experimented with absolute address visited by users and because this information is unique to each user, it did not help demographic classification, which are more generalized characteristics. In our next step, we chose location category as feature, which is more generalized

information and showed improved correlation with demographic characteristics.

3. GRAPHICAL FEATURE BASED FRAMEWORK

We propose a graphical feature based framework where we represent one type of sensor data in a graph, extract graphical features, apply feature selection techniques and then apply classification algorithm for prediction task.

3.1 Graph Representation

For the task of demographic prediction, we consider location data from GPS of mobile phones and categorize the locations visited. We represent each location category as a node in the graph and whenever the participant moves from one location to another location, we find the nodes in the graph for the corresponding location categories and add an undirected edge between the two nodes. In this way, we create a graph for each user. In the Experimental Setups section, we describe details of construction of such graphs and demonstrate some example graphs in Figure 6 to Figure 11.

3.2 Graphical Feature Generation

We construct one learning instance per user where we extract the existence of nodes and edges from the user's graph as features. We use the list of all nodes and edges triggered in all users' graphs as the feature set. For each task three different kinds of graphical features have been used: existence of nodes, existence of edges and existence of both nodes and edges. For the existence of nodes experiment, our feature set is all unique location categories that are represented as nodes in the graphs across all users. If a node is present in the graph for a user, we set value 'ON' for that feature in the corresponding instance otherwise we set the value to 'OFF'. For the existence of edge experiment, we compute all unique edges triggered by all users in their location-trajectory graph. We use this list of all edges as the feature set for the existence of edge experiment. To construct an instance for each user, we consider the user's location graph and check the existence of the edges. If the edge exists, we mark the value as 'ON'; otherwise 'OFF'. In the third experiment, we combined existence of both nodes and edges that we used as features in the previous two experiments as the combined feature set.

3.3 Feature Selection

Since we are using nodes and edges triggered across all users' graphs as our feature set, there might be a huge feature set compared to number of instances. To avoid overfitting in this condition, feature selection is employed for reducing the number of features. In the literature, there are two main approaches for choosing useful and relevant attributes, namely, filtering, and wrapper approach. [19] Filtering ranks each individual feature based on different measures such as information gain, correlation, gain ratio, and symmetrical uncertainty. In the wrapper approach, all possible subsets of the feature set are checked with a classification algorithm and the subset that performs the best is provided as the optimal and selected feature subset. When the size of the feature set is huge with thousands of features, the wrapper approach might be computationally expensive. For a large number of features, the hybrid approach is used where a feature set of reduced size is constructed based on the filtering approach and then the wrapper approach is applied to the reduced feature set to find an optimal feature subset. Then we apply the SVM classification technique on this optimally selected set of features.

Other than the above approach, we also tried a frequency-based approach where we select the top 20 features based on their frequency of occurrence in the dataset. We experimented with different types of features such as absolute addresses and location categories under the frequency-based approach while using those values as nodes in each user-graph and movement between location categories or absolute addresses as edges.

To avoid overfitting due to the large number of features, we also experimented by super-categorizing location categories to reduce the number of features. We created a mapping from super-category to sub-categories in order to group similar categories in the same group and thus reducing hundreds of different location category features to 9 different super-categories. Then we represent the super categories as nodes and transitions between them as edges.

We obtained better accuracy with the current Information-Gain based approach for selecting discriminating nodes and edges using the set of all location categories visited by participants, as compared to frequency-based graphical feature selection and super-categorization. These results show that less frequently visited location categories can be useful for discriminating among population characteristics along with the frequently visited locations.

4. EXPERIMENTAL SETUP

4.1 Dataset

We used the Nokia Mobile Data Challenge 2012 dataset, which was collected by Idiap and NRC-Lausanne through Lausanne Data Collection Campaign (LDCC) from 2009 to 2011. LDCC reached out to 185 participants to collect different types of mobile phone data related to location (GPS, WLAN), motion (accelerometer), proximity (Bluetooth), communication (phone call and SMS logs), multimedia (camera, media player), application usage (user-downloaded applications in addition to system ones) and audio environment (optional). [14] Among all these different types of data, we chose GPS location data to demonstrate the validity of our hypothesis. We chose the task of demographic attribute classification, namely, gender, age group and job type prediction from the proposed tasks in Nokia Mobile Data Challenge. Ground truth class distribution for these three tasks is shown in Figures 1, 2 and 3.

4.2 Extract Location Category

The Nokia dataset collected geolocation (latitude and longitude) information from users' smart phones along with a timestamp for all the participants. We extract users' geolocations sequenced in time from the whole dataset. We used OpenStreetMap (OSM), which is a map of the world built by a community of mappers that contribute and maintain data about roads, trails, cafes, railway stations and much more. [15] We have hundreds of thousands of geo-location data per user. It is time consuming to access existing world maps that are available only online to probe about categories of each geo-location. To address this issue, we used a tool called Nominatim [16] through which we can download OSM data, import to a local database, and do reverse geo-coding for large amounts of geo-location data locally in significantly less time. For our dataset, there are 171 unique location categories that

have been visited by all participants such as road, school, university, restaurant, post office, hotel, mall and others.

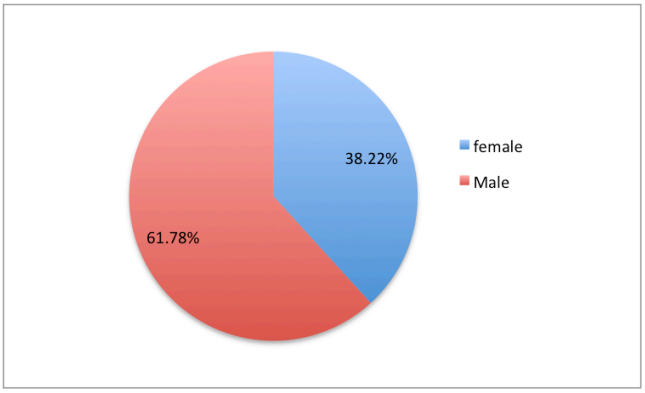

Figure 1. Class Distribution for Gender

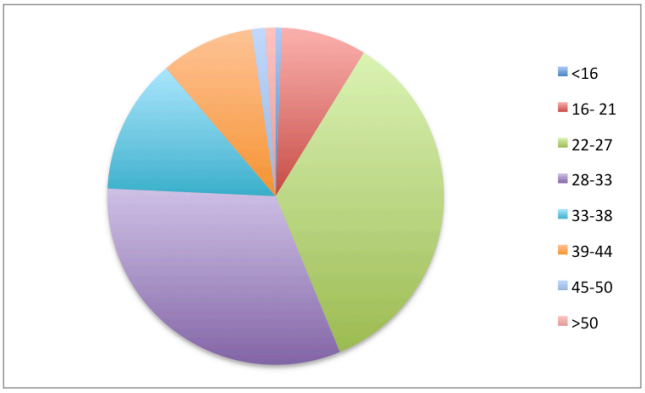

Figure 2. Class Distribution for Age Group

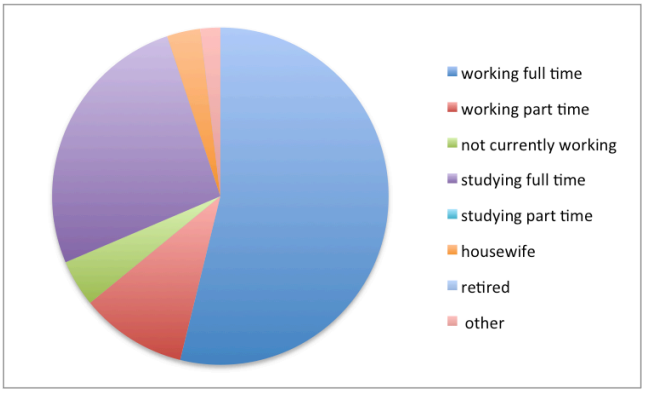

Figure 3. Class Distribution for Job Type

4.3 Construct User Wise Graphs

In Figure 5, we presented an example graph drawn from the data of a male participant where the spheres represent the nodes of the graph indicating all different unique location categories visited by the participant and the straight lines represent edges indicating movements between those location categories occurred in the dataset.

For our current dataset, there are a total of 171 unique location categories that are visited by participants, hence 171 features. In the dataset, we found 7430 unique edges visited by all the users in total. We also show the graph representations of the movements of participants with selected edges after filtering and wrapping in Figures 6 through 11. Figures 6 and 7 represent nodes and edges that are selected by the feature filtering and wrapping procedure as discriminating for the task of gender prediction and are triggered at least once in a female and in a male participant's dataset respectively. We also present figures 8 and 9 as example user graphs to show the useful nodes and edges that have been selected for classification of age-group and have been observed in participant's movement information for a participant of age more than 50 years and for a participant of age between 22 to 27 years respectively. Similarly, we show two example graphs with selected and triggered edges for the job-type classification task for a participant who is working full-time and for a participant who is not currently working in figures 10 and 11 respectively.

4.4 SVM To Graphical Feature Set

For existence of nodes experiment, we applied the wrapper approach directly as we have less than 200 nodes. For existence of edges experiment and for existence of both nodes and edges, we have thousands of nodes and edges together, and hence applied the hybrid approach. To filter, we use 'InfoGainAttributeEval' as the attribute evaluator with default parameters from Weka [17] and we use 'Ranker' as the search method with parameter N=200 to get top 200 highly ranked features based on the Information Gain measure. After filtering 200 attributes from thousands of features, we apply CfsSubsetEval as the wrapper approach with default parameters from Weka and with 'Greedy Stepwise' as the search method. We perform this procedure for all three tasks of classification resulting in different selected features and provide these selected features to the SVM classifier. We apply SMO for classification that implements John Platt's sequential optimization algorithm [18] for training a support vector classifier.

5. EXPERIMENTAL RESULTS

In this work, we report demographic prediction, namely, accuracy for classifying gender, age group and job type using our proposed graphical feature based framework. We measure accuracy and use 10-fold cross validation for training and testing. We compare our graphical feature based framework with the baseline of Majority Class prediction. We also compare our results with the benchmark works of Dong et al. [8], Ying et al. [9], Mo et al. [5], Brdar et al. [7], and Nadeem et al. [6].

We use majority voting as one of the baselines for evaluating the results of our experiment. Majority class for gender is male, for age group is 22-27 years old, and for job type is working-full-time. We see from Table 1 and Figure 4 that our existence of nodes approach results in significantly higher accuracy compared with the baseline for each task of demographics classification. Using only existence of edges as features improved the accuracy to a great extent for gender, age group and job type classification over the node only approach and baseline. Furthermore, using selected features from the combination of nodes and edges increased job-type accuracy significantly. In summary, our graphical feature based framework provides accuracy of 85.99% for gender with 39.19% improvement over baseline of majority class. Similarly, the graphical feature based framework demonstrated prediction accuracy of 66.45% for age group and 76.92% for job type, which outperformed majority class prediction by 87.5% for age and by 42.84% for job type.

The effectiveness of our method is three-fold. Firstly, the use of location type instead of absolute address led to better results. Absolute address is unique to each user and might not generalize well across all users for prediction tasks. On the other hand, location category holds enough information to discriminate among smart phone users' demographics but still represents

generalized information across all users. Secondly, we added transition between locations as features and from the results it is evident that users' movements between location categories are useful for predicting users' demographics. Thirdly, as oppose to using all possible features or selecting features based on their frequencies, we used feature selection techniques to find the optimum set of locations visited and movements between these locations that will differentiate users in this population.

Table 2 compares performance of our proposed graphical feature based framework method with the state of the art methods. All works mentioned in this Table use the Nokia Mobile dataset except Dong et al. [8]. Our method clearly outperforms the approach in Brdar et al. [7] where users are represented as nodes and cosine similarity between them is used as the feature vector.

Table 1. Accuracy (%) of Majority Class Vs Graphical Features

Approach	Gender	Age Group	Job Type
Majority Class	61.78	35.44	53.85
Existence of Nodes	68.79	50.00	58.97
Existence of Edges	85.99	66.45	72.43
Existence of Nodes and Edges	85.99	63.29	76.92

Our work outperformed Brdar et al.'s work by 3% for prediction of gender and 90% for prediction of job-type. Our approach also outperformed Nadeem et al.'s [6] approach that uses SVM and random forest on raw and computed feature sets by 3% for gender prediction and by 53.84% for job-type. The work of Mo et al. [5] provides a better accuracy than ours. But they used tens of thousands of features, a sophisticated feature generation process, and background knowledge to yield high performance of 89% for gender classification and 78% for job-type classification. In [5], Mo et al. used all sensor types provided in the dataset, while we obtained performance close to theirs using just one type of sensor information. Our simplistic graphical feature based framework provides result of 85.99% for gender and 76.92% for job-type, which is close to their result.

We obtained this comparable accuracy without the need for handcrafted application-specific features and by just leveraging the inherent graph structure of sensor networks. The approach in Mo et al. [5] used correlation between classification tasks to improve the result even more. In our graphical feature based framework we avoid use of task-specific feature engineering since our goal is to build a generic framework that can be used across different sensor network applications to improve predictions for corresponding tasks in general.

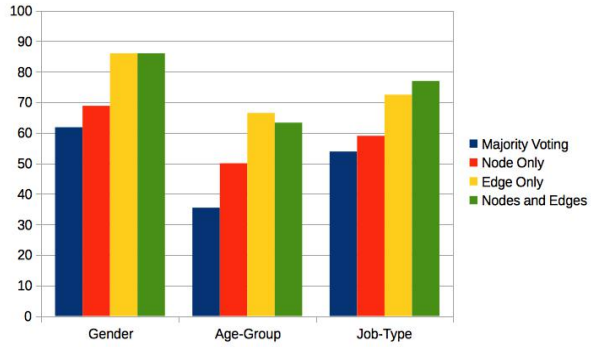

Figure 4. Accuracy (%) of Majority Class Vs Graphical Features

The hierarchical model used by Ying et al. [9] also used different types of sensor information related to behavioral features, movement, location, phone usage and communication whereas we computed graphical features just from one type of sensor. They approached the imbalance class problem in the Nokia mobile dataset through the use of multiple classifiers at multiple levels to reach the higher accuracy. For gender classification, the dataset does not exhibit the imbalanced class problem and in this case, our result is better than theirs with just location category information and with use of graphical features.

Dong et al. [8] use a different dataset, which contains 7 million users and 1 billion communication records. They used background information such as inference of social strategies from the population and used the insight from mining results for demographic predictions. Interdependency of age and gender has been employed for improving the predictions. Their approach performs better for gender with CALL information and for age with SMS information. Our simplistic approach of graphical features outperformed their prediction accuracy of gender by 7% and our age-group prediction achieved a comparable performance with social-strategy based approach with the use of just one type of sensor information and without any application based knowledge inference.

Table 2. Accuracy of Graphical Feature Framework Vs Other State of the Art Methods

Prediction Task	Dong et al. [8]	Ying et al. [9]	Mo et al. [5]	Brdar et al. [7]	Nadeem et al. [6]	Graphical Feature Framework
Gender	80.00	85.47	89.00	83.33	83	85.99
Age Group	73.00	77.77	NA (Regression)	NA (Regression)	NA (Regression)	66.45 (Classification)
Job Type	NA	83.33	78.00	40.28	50	76.92

6. CONCLUSION

In this paper, we study and assess the use of graph representation and graphical features in sensor networks to improve prediction tasks in general. We propose a graphical feature based framework to apply to sensor network data. The purpose of construction and use of this kind of framework is multi-fold. First, the framework leverages inherent graph structure of sensor networks to represent sensor network data. Second, the framework offers a generic approach of applying graphical features to improve performance accuracy of prediction tasks across different sensor networks. Third, the framework improves predictions without using application-specific and prediction-task specific feature crafting. Previous work [1] represented motion sensor data from smart home sensor network in a graph and extracted graphical features that improved activity recognition performance over all existing benchmarks. In this work of demographic attribute prediction, our graphical feature based framework outperformed most of the state of the art methods while using no background knowledge and using no application-specific and task-specific feature computation.

As next steps, we would like to explore other sensor network applications, extract larger subgraphs and path sequences along with edges as features, and evaluate the effect of these graphical features on prediction tasks of sensor network applications.

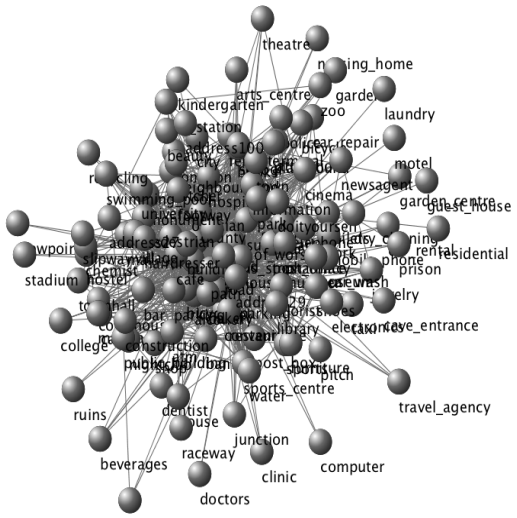

Figure 5. Undirected graph representation for a male participant showing all the edges triggered

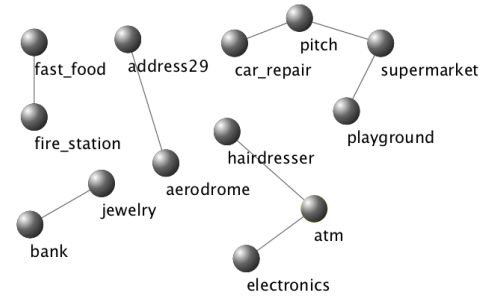

Figure 6. Edges triggered for a female participant

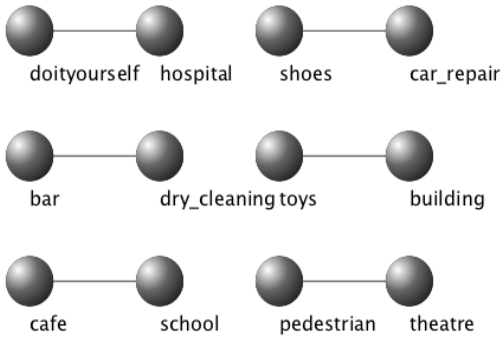

Figure 7. Edges triggered for a male participant

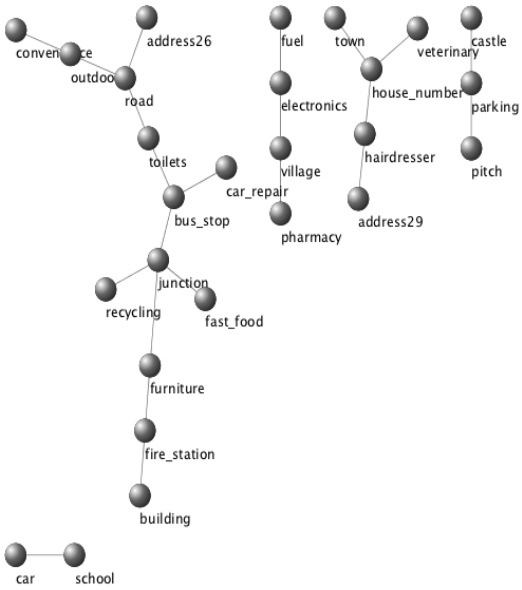

Figure 8. Edges triggered for a participant with age > 50

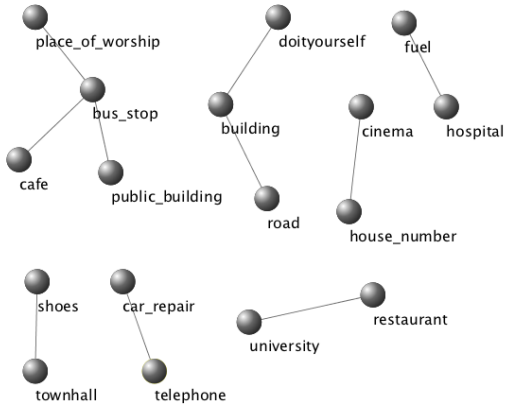

Figure 11. Edges triggered for a participant who does not currently have a job

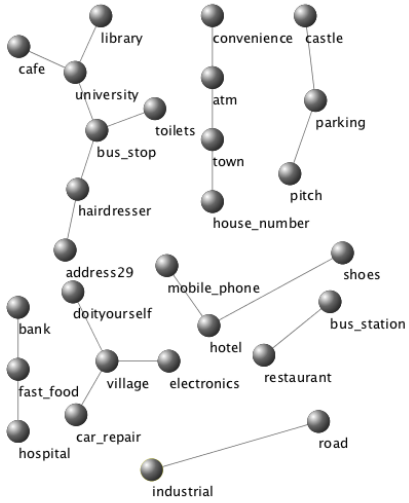

Figure 9. Edges triggered for a participant of age 22-27

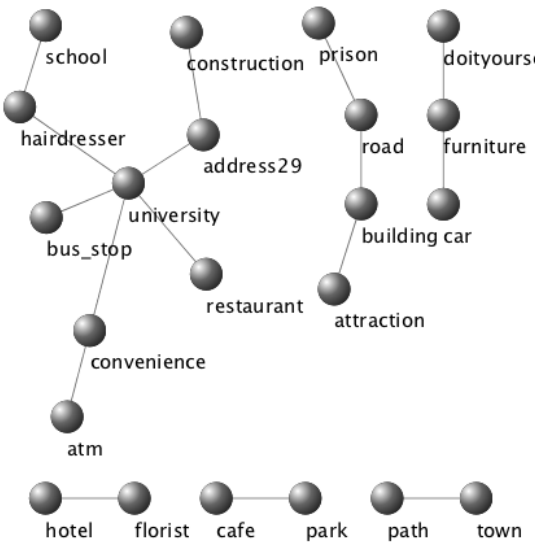

Figure 10. Edges triggered for a participant having full-time job

7. ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1646640.

8. REFERENCES

[1] Akter, S. S., & Holder, L. B. (2014, December). Activity Recognition Using Graphical Features. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on* (pp. 165-170). IEEE.

[2] Diamantopoulos, A., Schlegelmilch, B. B., Sinkovics, R. R., & Bohlen, G. M. (2003). Can socio-demographics still play a role in profiling green consumers? A review of the evidence and an empirical investigation. *Journal of Business research*, *56*(6), 465-480.

[3] Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 274-279.

[4] Axinn, W. G., Link, C. F., & Groves, R. M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*, *48*(3), 1127-1149.

[5] Mo, K., Tan, B., Zhong, E., & Yang, Q. (2012, June). Report of task 3: your phone understands you. In *Nokia mobile data challenge 2012 workshop, Newcastle, UK* (pp. 18-19).

[6] Nadeem, S. M. S. J. T., & Weigle, M. C. (2012). Demographic prediction of mobile user from phone usage. *Age*, *1*, 16-21.

[7] Brdar, S., Culibrk, D., & Crnojevic, V. (2012, June). Demographic attributes prediction on the real-world mobile data. In *Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK*.

[8] Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014, August). Inferring user demographics and social

strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 15-24). ACM.

[9] Ying, J. J. C., Chang, Y. J., Huang, C. M., & Tseng, V. S. (2012). Demographic prediction based on users mobile behaviors. *Mobile Data Challenge*.

[10] Koppel, M., Schler, J., Argamon, S., & Pennebaker, J. (2006, March). Effects of age and gender on blogging. In *AAAI 2006 spring symposium on computational approaches to analysing weblogs* (pp. 1-7).

[11] Lenhart, A., & Fox, S. (2010). Bloggers-a portrait of the internet's new storytellers (2006). *URL http://www. pewinternet. org/Reports/2006/Bloggers. aspx*.

[12] Hu, J., Zeng, H. J., Li, H., Niu, C., & Chen, Z. (2007, May). Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web* (pp. 151-160). ACM.

[13] Weber, I., & Castillo, C. (2010, July). The demographics of web search. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 523-530). ACM.

[14] Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T. M. T., Dousse, O., ... & Miettinen, M. (2012). The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing* (No. EPFL-CONF-192489).

[15] Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12-18.

[16] https://nominatim.openstreetmap.org/

[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

[18] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

[19] .Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE transactions on knowledge and data engineering*, *15*(6), 1437-1447

[20] Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018-1021.

[21] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.