Internet Usage Modeling of Large Wireless Networks Using Self-Organizing Maps

Saeed Moghaddam, Ahmed Helmy
Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL
{saeed, helmy}@cise.ufl.edu

***Abstract*—Human behavior and interests will play a central role in future mobile networks. We introduce a systematic method for large-scale multi-dimensional analysis of online activity for thousands of mobile users over 100 web domains. We propose a modeling approach based on self-organizing maps for discovering, organizing and visualizing different mobile users' trends from billions of WLAN and netflow records. We find surprisingly that users' trends can be modeled using a self-organizing map with clearly distinct characteristics while similar domains are modeled by neighboring nodes. This is the first study to acquire such detailed results for mobile Internet usage.**

Keywords- self-organizing map; data-driven; trend; wireless

I. INTRODUCTION

Wireless mobile networks have evolved into one of the most growing aspect of our lives. Laptops, handhelds and smart phones are becoming ubiquitous providing almost continuous Internet access. This provides a new window of opportunity for analyzing and modeling the behavior of mobile users and their online activities on the Internet. Such behavioral modeling will aid in the understanding of the load distribution on the network and users trends, and thus inform the design of important classes of applications, including modeling and scenario generation for network simulations, network capacity planning, web caching and behavior-aware networking protocols [1] to name a few. However, such behavioral analysis on extensive traces of Internet access is difficult, as it is large-scale, computationally costly and time-consuming. Such traces usually contain billions of records and therefore, the analysis may even be impossible when the dataset exceeds certain limits of size and complexity. As a result, it is imperative to establish systematic and scalable methods for the analysis and modeling of massive multi-dimensional datasets of mobile users' online activities.

Much of the previous mobility or web usage modeling have focused on individual behavior. While individual behavior is important, investigating group behavior (trend) is more challenging and involved. In this paper, we focus on group behavioral modeling, and study behavioral correlations based on groups of mobile users and their trends toward various websites.

One conventional approach to group data analysis is to cluster the data items with respect to some similarity or distance measure. However, many common clustering algorithms are either computationally intensive, e.g., hierarchical clustering [2], or provide no organization for the resulting clusters, e.g., k-means [2]. Moreover, although the resulting clusters provide some insight on similarities between data items (and among features, if co-clustering [3] is used), they do not reveal intuitively how similar they are. Therefore, they are not effectively useful for finding complete and partial correlations when there exist many features.

Our proposed approach to address the above limitations in data analysis is to apply self-organizing maps (SOM)[4]. A self-organizing map is an artificial neural network which is trained using unsupervised learning to generate a discretized low-dimensional representation (a map) of the input space of the data samples. Unlike other artificial neural networks, self-organizing maps use a neighborhood function to preserve the topological properties of the input space. The topology-preserving mapping keeps the more similar data groups closer together in the final map, which makes SOM useful for visualizing low-dimensional views of high-dimensional data. With further examination, we then reveal what features the members of a group have in common and how different features are correlated together.

In this study, we apply SOM in a novel way on a dataset provided by the processing of extensive *netflow*, DHCP and WLAN traces for more than 22 thousand mobile users in a Wireless LAN spanning over 79 buildings and including over 700 APs, that we have collected. This dataset, including billions of records, represents by far the largest set of traces analyzed in any study of mobile networks to date. Using this technique, we extract minor and major trends in mobile users' website access patterns and important correlations between different accessed web domains. We show how to apply this techniue methodically to the collected large-scale multi-dimensional dataset with minimal computational complexity to facilitate its meaningful analysis. Our method is systematic and can be generally applied to discover important features of

Internet behavior from other similar traces.

We report two major findings in this paper. First, mobile users' access patterns of web domains can be accurately modeled by a small set of neurons which can be further clustered into smaller number of major trends with clearly distinct characteristics. For example, Mac users who frequently visit 'mac' and 'apple' websites have strong trends toward 'washingtonpost' and 'cnet' too. Second, web domains in similar categories tend to be modeled by an adjacent set of neurons. For example, most of the hosting domains or advertisement and marketing domains are modeled together by close neurons. These findings provide the basis for mobile user behavioral models both qualitatively and quantitatively, as we discuss in our applications section.

Our work has the following key contributions:

1. We propose an effective approach for multi-dimensional analysis of one of the largest set of mobile network usage traces (including billions of records) and show how self-organizing map can be applied to model minor and major trends in mobile users' behaviors.
2. We conduct domain-specific analysis of mobile users' behaviors using the feature maps extracted from the SOM and show how this method can be effectively applied to discover correlations among different domains.

This work provides the necessary first step towards realistic modeling of mobile users' Internet behaviors, which is part of our future work.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we address challenges associated with the collection, processing and analysis of large-scale wireless traces. Section 4 provides our case study using campus traces and self-organizing maps to develop realistic models of wireless network. Section 5 discusses modeling and applications. Section 6 concludes.

II. Related Work

The rapid growth of wireless communication technologies has led to a widespread interest in analyzing the traces to understand user behavior. The scope of analysis includes WLAN usage and its evolution across time [5, 6], traffic flow statistics [7], user mobility [8, 9], user association patterns [10] and encounter patterns [11]. Some previous works [8, 11] attempt to understand user behaviors empirically from data traces. The two main trace libraries for the networking communities can be found in the archives at [12] and [13]. None of the available traces provides large-scale *netflow* information coupled with DHCP and WLAN sessions to be able to detect web domains and locations. Therefore, (to the best of our knowledge) our work is the first one to address large-scale multi-dimensional modeling of wireless networks. We analyze wireless data around three orders of magnitude above any existing study, providing richer semantics, finer granularity and potentially more accurate models. In addition, our work includes novel data analysis techniques to address the challenges provided by this large-scale multi-dimensional data.

There are several noticeable examples of utilizing the data sets for context specific study. Mobility modeling is a fundamentally important issue, and several works focus on using the observed user behavior characteristics to design realistic mobility models [14-16]. The work on the *TVC* model [15] provides a realistic mobility model for protocol and service performance analysis. Our work is complementary to *TVC* and can extend *TVC* dramatically to incorporate dimensions of load, interest and website visitation preferences. In [7] it is shown that the performance of resource scheduling and TCP vary widely between data-driven and non-data-driven model analysis. Using multi-dimensional modeling, our method has the potential to develop new mobility-aware Internet-usage models, and utilize the realistic profiles to enhance the performance of networking protocols. Our new application of self-organizing maps may be extended to incorporate online activity, location and mobility, and provides user profiles that may be used in a myriad of networking applications.

One network application for multi-dimensional modeling is profile-based services. *Profile-cast* [1] provides a new one-to-many communication technique to send profile-aware messages to those who match a *behavioral profile.* Behavioral profiles in [1] use location visitation preference and are not aware of online activity. Other previous works also rely on movement patterns. Our multi-dimensional modeling of mobile users, however, provides an enriched set of user attributes that relate to social behavior (e.g., interest, community as identified by web access, application, etc.) that has been largely ignored before.

III. Modeling Approach

Realistic modeling of large wireless networks requires three main phases to collect, process and analyze multi-dimensional large datasets with fine granularity. In the first phase, extensive datasets are collected using the network infrastructure which may be augmented using online directories (e.g., buildings directory, maps) and the web services (e.g., whois lookup service). Data processing is the second phase to cross-correlate obtained information from different resources (e.g., IP and MAC addresses), in which multiple datasets are manipulated, integrated and aggregated. The final phase is data analysis that includes trend and domain-specific study of human behaviors based on their website access preferences.

A. Data Collection

For the campus-wide modeling of wireless users, we collect different types of traces via network switches including netflows, DHCP and wireless AP session logs. An IP flow is defined as a unidirectional sequence of packets with some common properties (e.g., IP address and port number source and destination) that pass through a network device (e.g., router) which can be used for flow collection. Network flows are highly granular; flow records include the start and finish times (or duration), source and destination IP addresses, port numbers, protocol numbers, and flow sizes (in packets and bytes) (see Table I). The source and destination IP addresses can be used to identify user device MAC addresses and the websites accessed respectively. The DHCP log contains the dynamic IP assignments to MAC addresses and includes date and time of each event.

B. Data Processing

The variety and scale of different collected traces introduces one of the main challenges with respect to data processing. The size of the underlying data is very large and therefore, with a naïve approach the required time for this task will be in the order of months. For example, the netflow dataset gathered from USC campus includes around 2 billion of flow records for each month in 2008 which equals to 2.5 terabytes of data per year. To circumvent the problem, we first compress the data via substituting similar patterns with binary codes and creating mapping headers to be used in the analysis step; then get the data exported into a database management system (MySQL) and design customized stored procedures for data integration (mapping source IPs to MAC addresses (user ids) and destination IPs to domain names). In the last step, we aggregate the integrated data based on user id, domain name and month and calculate the total online time for each resulting record.

C. Data Analysis

The data analysis phase includes two major parts. In the first part, we employ the self-organizing map to learn trends of website visitation patterns within the mobile society. This part itself contains two steps to first learn minor trends and then discover major trends using clustering. In the second part, we conduct domain-specific analysis using the feature maps extracted from the SOM to discover correlations among different domains in the studied mobile society.

1) Trend Modeling: The SOM technique [4] provides a powerful yet intuitively understandable tool for unsupervised learning and data visualization. The SOM is defined as a set of nodes which develop a mapping of high-dimensional input vectors onto a discrete output space (the “map”) such that each region on the map represents an area of the input space. This mapping preserves the topology of the input space in a way that local similarity of input patterns is reflected by proximity on the map. Therefore, it can be effectively applied for capturing the properties of the input space of users’ behaviors and organizing their trends in an ordered fashion. In a self-organizing map, a weight vector of the same dimension as the input data vectors and a position in the 2-D map space are associated with each node (or neuron in neural networks). The usual arrangement of nodes is in the form of a hexagonal or rectangular grid. SOM training, i.e., the iterative adjustment of the weight vectors to acquire a desired mapping, is performed by successive presentation of all input data where each presentation leads to the adjustment of weights to the presented data. The training is based on two principles:

a) Competitive learning: the weight vector most similar to a data vector is modified so that it is even more similar to it (the corresponding node is called Best Matching Unit or BMU). This way the map learns the position of the input data. We use standard Euclidean metric as similarity function.

b) Cooperative learning: not only the most similar weight vector, but also its neighbors are moved towards the data vector. This way the map self-organizes.

The neighborhood function *h* regulates the weight changes based on the map distance between BMU and the neuron being adapted. In the case of a Gaussian shaped neighborhood function, the expression of *h* is given by:

$$h(i,j) = \exp\left(-\frac{dist_{map}(i,j)^2}{2r(n)} \right) \tag{1}$$

TABLE I – NETFLOW SAMPLE

Start Timestamp	Finish Timestamp	Source IP	Source Port	Dest IP	Dest Port	Protocol Num	ToS	Packet Count	Flow Size
0618.00:00:07.184	0618.00:00:07.184	128.125.253.143	53	207.151.245.121	64209	17	0	1	469
0618.00:00:07.184	0618.00:00:07.472	207.151.241.60	52759	74.125.19.17	80	6	0	4	1789
0618.00:00:07.188	0618.00:00:07.188	193.19.82.9	31676	207.151.238.90	43798	17	0	1	103

where $dist_{map}(i,j)$ measures the distance on the map between two neurons and $r(n)$ is a global parameter that controls the *width* of the neighborhood function. According to this expression, the amount of changes is maximum for the BMU and decreases for nodes that are far from it. The value of $r(n)$ decreases with the number of iteration; a relatively large radius during the initial iterations allows the map to quickly organize the nodes, while a smaller value toward the end determines localized changes in a way that different parts of the map become sensitive to different input features. The learning rate of the map decreases monotonically with the number of iterations to ensure convergence.

In this way, *each neuron can learn a minor trend that represents a set of similar input data vectors*. This is one of the major advantages of SOMs with respect to clustering techniques. While a clustering technique attempts to partition the input space (e.g. users' behaviors) by assigning each sample (e.g. a user) to a cluster, the SOM tries to learn trends inside the input space from the samples. Note that each input data vector (e.g. a user) affects a set of neighboring neurons (trends) and therefore the input space is not distinctly partitioned by the neurons like what clusters do in conventional clustering techniques. Therefore, this approach much better accommodates natural human behaviors with no clear-cut distinctions.

Map Creation - The side lengths of the map grid are determined based on the ratio of two biggest eigenvalues of the training data. For initializing the SOM, first, linear initialization along two greatest eigenvectors is attempted, but if the eigenvectors cannot be calculated, random initialization is used instead. After the initialization, the SOM is trained by normalized input data. The normalization of the input features is very important in determining what the map will be like. If the ranges of value of some features are much bigger than of the others, those features will probably dominate the map organization completely and the resulting map will not be useful. The computational complexity of SOM algorithm scales linearly with the number of data samples and quadratically with the number of map units

Map Clustering - One way to visualize the resulting map after the training phase is to create U-matrix (unified distance matrix). The U-matrix shows the distance between the weight of each node and the assigned weights of its neighbors after the learning process. Fig. 1(a) shows an instance of U-matrix with interpolated shading of colors. Small U-values (Blue areas in the figure) indicate homogenous neighborhoods and large ones (Red areas) depict heterogeneous neighborhoods. As large U-values mean large distances between the neighboring nodes, they can be interpreted as borders between clusters of neurons, i.e., trends. In order to find these borders (clusters), k-means clustering algorithm can be applied. Because k-means result depends on the initial choice of cluster centroids, the algorithm is run multiple times for a given k and then the best result is selected based on the sum of the squared errors. Because the captured minor trends are already very well organized on the map, each resulting cluster maps into a contiguous area of neurons, representing a major trend (Fig. 1(b)). Clustering of trends instead of original data reduces the required computational time for any kind of clustering technique as the size of input is decreased. This is very important when dealing with massive amount of data. In addition, as the weight vectors are local averages of the data, the clustering result is less sensitive to random variations (noises) in the input data.

2) Domain-Specific Analysis: We conduct domain-specific analysis using the feature maps extracted from the SOM. The feature maps show what kind of values the weight vectors of the map units have for each feature. In other words, a feature map shows the projection of the SOM for the corresponding feature which is a web domain in this case. The value of each unit for the feature is presented with a color. Fig. 2 to Fig. 4 show a group of resulting feature maps in our study. By visual inspection and comparison of the feature maps we can find many different interesting facts about the trends and features as below:

a) Comparison of feature maps with the clustered SOM discovers the semantic behind each cluster of trends representing a major trend. For a cluster area, features whose maps looks red in the same area disclose the main captured trends by the cluster.

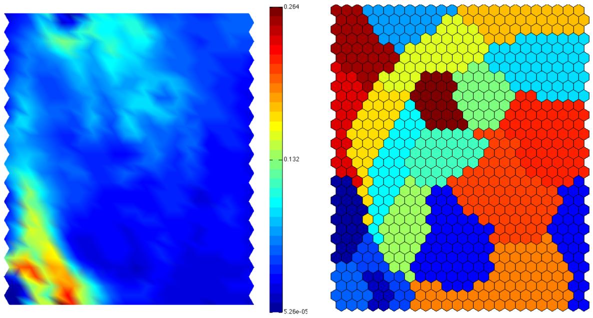
b) Similar feature maps reveals correlations between the corresponding features. The correlation can be partial or complete. If the maps seem highly similar, there exist rather complete correlation, but if they are partially similar the correlation among features will also be partial. In our case, correlation between a set of features, i.e., domains means that they have the same visitation pattern (if one is visited, the others are visited too with similar rates)

IV. CASE STUDY AND EXPERIMENTAL RESULTS

In our case study, we conduct a campus-wide analysis on the data collected from the USC in 2008 based on the approach and techniques explained in the previous section.

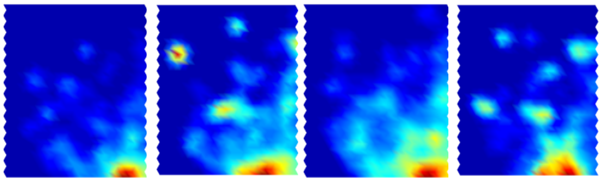
A. Data Processing Details

The *netflow* and DHCP traces from the USC campus (over 700 access points) were processed to identify mobile user IDs (MAC addresses), and destinations, or 'peers' (usually web servers) using IP address prefixes. Over a billion records (for Mar. 2008) were considered. Then, the IP prefixes (first 24 bits) were filtered using a



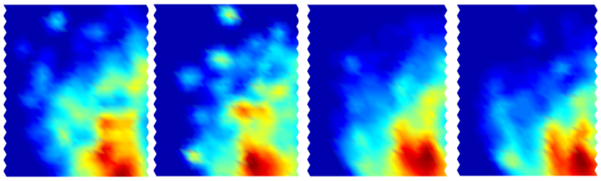
(a) U-matrix (b) Clustered SOM

Figure 1. U- matrix and clustered SOM for WLAN Internet usage in USC campus.



xlhost softlayer theplanet steadfast

Figure 2. Feature maps for hosting domains



webtrendslive coremetrics tribalfusion fastclick

Figure 3. Feature maps for advertisement and marketing domains

threshold of 100,000 flows (the reason for using 24 bits filter is that popular websites usually use an IP range instead of a single IP address). For the filtered IP prefixes, their domains were resolved. Among the resolvable domains, the top 100 active ones were identified and all the users interacting with those domains were considered for the analysis.

B. Trend Modeling Results

A matrix was created associating the user IDs and web domains using the corresponding total online time (per minute). For our analysis, we had 22,816 users, and 100 domains. The data is scaled using row-normalization of log the online time values. This is the input data for our modeling problem for which we applied the SOM technique. A map of 32 by 24 neurons was trained by this input dataset and the U-matrix was created (Fig. 1 (a)). Then, the SOM was clustered into 20 clusters (Fig. 1 (b)).

C. Domain Analysis Results

The feature maps were created for all the domains. Fig. 2 to Fig. 4 show several examples of resulting maps for different types of web domains. Inspection of the feature maps reveals many interesting facts. The following are some examples based on the presented feature maps here.

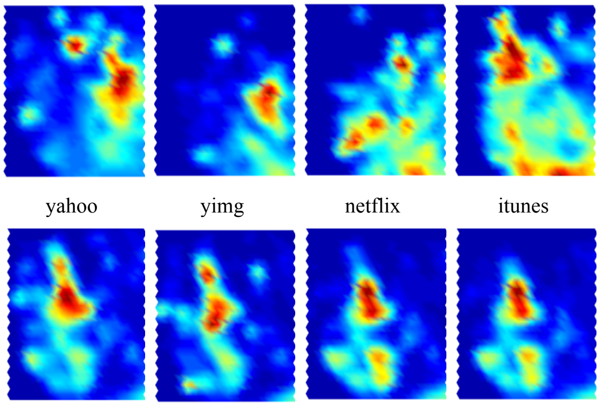
a) Fig. 2 and Fig. 3 show feature maps for two categories of web domains including: i) web hosting and ii) advertisement and marketing. All these maps show a red area almost at the same neighborhood (right-bottom corner). This shows the major trend captured by the cluster depicted by orange at the same area in Fig. 1(b) is toward these kinds of web domains.

b) High similarity between feature maps in Fig. 2 shows that the corresponding domains for web hosting are highly correlated. That is also the case for advertisement and marketing domains in Fig. 3. We can also observe high correlations between the following groups of domains from Fig. 4: i) 'yahoo' and 'yimg' (yahoo image); ii) 'itunes' and 'netflix' (online media); iii)'mac', 'apple', 'washingtonpost' and 'cnet' (showing a strong trend of Mac users toward 'cnet' and 'washingtonpost'). In the figure, we can see that 'itunes' is in one hand partly correlated to 'netflix' and on the other hand is partly correlated to 'mac' and 'apple'. This may show the facts that i) Mac users dominantly use iTunes for online media and ii) Netflix costumers shop in iTunes store too.

V. DISCUSSION: MODELING AND APPLICATIONS

The systematic realistic mining method proposed in this paper can be applied with any set of wireless data to reveal significant facts that may be used in several important applications in mobile networking research. Here, we briefly address three such major applications:

1- Modeling and simulating spatio-temporal web usage for mobile users: Network simulations are imperative for the design and evaluation of mobile networks (e.g., ns-2). To provide realistic input to the simulations, realistic models of users' behaviors are required, along with scenarios of events and dynamics of mobility, traffic and Internet access. While earlier work has focused on mobility and traffic modeling, we provide the first work on modeling of mobile Internet



yahoo yimg netflix itunes

apple mac washingtonpost cnet

Figure 4. Feature maps for various types of domains

usage. The parameters of on-line activity along with trend characteristics and correlations in the simulation can be easily derived from our analysis in this paper. None of the existing models captures such characteristics across website access. Recreating network usage more accurately will result in significantly different mobile node density, load, and similarity distributions from those created by today's models. Developing and releasing the code for the mobile Internet access model is part of our future work. Similarly, we plan to conduct an extensive study on the spatio-temporal parameters for mobile traffic modeling in the future.

2- Interest-based protocols and services: A new class of protocols and services center around user-interest and similarity, including profile-cast, participatory sensing [17], trust establishment [18], location-based services, crowd sourcing, alert notification and targeted announcements and ads. So far, mobility patterns have been used to infer interest. Website access patterns can remarkably enhance the accuracy of interest inference and provide much needed granularity for these protocols and services. The interest models developed based on our analysis can help both the informed design of such efficient protocols and the realistic evaluation thereof.

3- Network planning and web caching: Load distribution on the network is imperative for network capacity planning and on-going configuration and management issues, and is definitely related to web access patterns and its characteristics. Also, the caching of web objects for mobile users can only be efficient if informed by the history of access patterns. These applications for mobile networks are becoming more compelling especially with the significant growth of usage of smart phones, iphones, ipads, and the like.

VI. Conclusion

This study is motivated by the need for developing realistic models and efficient protocols for the future mobile Internet. We provided a systematic method to analyze the largest wireless trace to date, with billions of records of Internet usage from a campus network, including thousands of users. Novel analysis was conducted utilizing advanced data mining using self-organizing map for trend modeling and domain-specific analysis. We have shown that mobile Internet usage can be modeled with an organized map of trends which can be effectively used to find correlations. The details of our study enable the parameterization of new and realistic models for mobile Internet usage with applications in several areas of networking, including mobile web caching, simulation and evaluation of protocols, interest-aware services and network planning, to name a few. We hope for our analysis and method to provide an example for large-scale data-driven modeling of mobile networks in the future. With more measurements from mobile and sensor networks becoming available, we expect our method to facilitate analysis of many other large datasets in future studies.

Acknowledgment

This work was partially funded by NSF award number 0832043.

References

[1] Hsu, W., Dutta, D. and Helmy, A. Profile-cast: Behavior-aware mobile networking. *ACM SIGMOBILE Mobile Computing and Communications Review*, 12, 1 (Jan 2008), 52-54.

[2] Jain, A. K. and Dubes, R. C. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[3] Dhillon, I. S. and Guan, Y. Information Theoretic Clustering of Sparse Co-Occurrence Data. In *Proceedings of the Third IEEE ICDM* (2003). IEEE Computer Society.

[4] Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43, 1 (Jan 1982), 59-69.

[5] Kotz, D. and Essien, K. Analysis of a campus-wide wireless network. *Wirel. Netw.*, 11, 1-2 (Jan 2005), 115-133.

[6] Henderson, T., Kotz, D. and Abyzov, I. The changing usage of a mature campus-wide wireless network. *Computer Networks*, 52, 14 (Oct 2008), 2690-2712.

[7] Meng, X., Wong, S. H. Y., Yuan, Y. and Lu, S. Characterizing flows in large wireless data networks. In *Proceedings of the ACM MobiCom 2004* (Philadelphia, PA, USA, 2004). ACM.

[8] Hsu, W. and Helmy, A. On modeling user associations in wireless LAN traces on university campuses. In *Proceedings of the IEEE WiNMee 2006* (Apr, 2006).

[9] Balazinska, M. and Castro, P. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proceedings of the ACM MobiSys 2003* (San Francisco, CA, 2003). ACM.

[10] Papadopouli, M., Shen, H. and Spanakis, M. Characterizing the duration and association patterns of wireless access in a campus. In *Proceedings of the 11th European Wireless Conference* (Nicosia, Cyprus, Apr, 2005).

[11] Hsu, W. and Helmy, A. On Nodal Encounter Patterns in Wireless LAN Traces. In *Proceedings of the IEEE WiNMee 2006* (Apr, 2006).

[12] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements. http://nile.cise.ufl.edu/MobiLib/.

[13] Kotz, D. and Henderson, T. Crawdad: A community resource for archiving wireless data at dartmouth. *IEEE Pervasive Computing*(Dec 2005), 12-14.

[14] Lelescu, D., Kozat, U. C., Jain, R. and Balakrishnan, M. Model T++: an empirical joint space-time registration model. In *Proceedings of the 7th ACM MOBIHOC* (Florence, Italy, May, 2006). ACM.

[15] Hsu, W.-J., Spyropoulos, T., Psounis, K. and Helmy, A. TVC: Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE/ACM Trans. Netw.*, 17, 5 (Oct 2009), 1564-1577.

[16] Kim, M., Kotz, D. and Kim, S. Extracting a Mobility Model from Real User Traces. In *Proceedings of the IEEE INFOCOM 2006* (Barcelona, Spain Apr, 2006).

[17] Reddy, S., Estrin, D. and Srivastava, M. Recruitment Framework for Participatory Sensing Data Collections. *Pervasive Computing*(May 2010), 138-155.

[18] Kumar, U., Thakur, G. and Helmy, A. PROTECT: proximity-based trust-advisor using encounters for mobile societies. In *Proceedings of the IWCMC 2010* (Caen, France, Jun, 2010). ACM.