Interest-based Mining and Modeling of Big Mobile Networks

Saeed Moghaddam, Ahmed Helmy

University of Florida
saeed@cise.ufl.edu, helmy@cise.ufl.edu

***Abstract*—Usage of mobile wireless Internet has grown very fast in recent years. This radical change in availability of Internet has led to communication of big amount of data over mobile networks raising new challenges and opportunities for modeling of mobile Internet characteristics. While the traditional approach toward network modeling suggests finding a generic traffic model for the whole network, in this paper, we show that this approach does not provide enough accuracy for big mobile networks. We show that user interest acquired based on accessed domains and visited locations has a significant effect on traffic characteristics of big mobile networks. Our case study based on a big dataset including billions of netflow record collected from a campus-wide wireless mobile network reveals the fact that domains and locations showing similar point of interests (e.g. domains of news agencies or locations of fraternities) mostly follow similar types of traffic distributions. For this purpose, we utilize a novel graph-based approach based on KS-test. We also show that interest-based modeling of big mobile networks based on visited domains and locations can significantly improve the accuracy and reduce the KS distances by factor of 5 comparing to the generic approach.**

Keywords- interest; mobile; traffic; big data

I. INTRODUCTION

Mobile Internet traffic has experienced a significant growth in the recent years. Different types of Internet-enabled mobile devices are getting more and more popularity and wireless Internet access infrastructures are growing faster than ever. The emergence of this radical change in availability of Internet raises a new need for modeling of Internet characteristics in big mobile networks. A traffic model in general is a model that can be used to regenerate the behavior of a real traffic stream. A major application of traffic models is in predicting the behavior of traffic as it passes through a network. The common approach toward traffic modeling is to find a generic model for the whole network. Although, such models provide good approximations for the old wired Internet, but several studies have shown that they do not fit the dynamics of wireless networks. For example, [1] characterizes the wireless traffic in different locations and shows that the dynamics of network follow a similar model but with different parameters. However, such models were generally based on small datasets of WLAN activities (e.g. 25000 flows a day), which are far from the full scale of dynamics in current big mobile networks (e.g., our dataset includes over 100 million flows per day). Moreover, most previous works have not studied the characteristics of big mobile networks from user interest point of view which can be acquired based on accessed domains (e.g. 'cnn') or visited locations (e.g. 'cinema'). Interest-based modeling of big mobile network traffic can be beneficial to the realistic design of applications, protocols and services (e.g. for resource allocating or content caching).

In this paper, we present a novel interest-based modeling approach based on accessed domains and visited locations using a novel graph-based technique. Our campus-wide case study shows that domains and buildings have specific traffic characteristic that can also form groups with distinct characteristics. In our study, we investigate interest-based characteristics by first filtering the Internet traffic for individual domains and buildings and then comparing their characteristics using Two-Sample KS (Kolmogorov-Smirnov) test [2] to either accept or reject the hypothesis of following similar distributions.

This work has the following key contributions:

1. We provide a novel interest-based traffic modeling technique for big mobile networks based on accessed domains (top 100 active domains) and visited locations (68 different buildings) across a campus with more than 32000 users. The studied dataset is one of the largest wireless mobile network traffic traces (including around 100 million records per day).

2. We provide a systematic method to discover similarities and differences between the traffic distributions of different domains or locations. We show how a novel graph-based technique can be applied to identify groups of domains or locations with distinct traffic characteristics which reveals the fact that members of each group represent a similar point of interest.

3. We show that the proposed interest-based approach can significantly improve the accuracy of traffic modeling for big

mobile network and reduce the KS distances by factor of 5 using KS-test and weighted traffic intensity of domains and locations.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we briefly describe the datasets and big data processing phase. Section 4 presents our interest-based traffic modeling approach. Section 5 discusses the accuracy of the proposed approach and Section 6 concludes the paper.

II. Related Work

There has been many works on Internet traffic modeling and among all flow-level modeling has been one of the most popular approaches [3,4,5]. However, most of such studies use idealized models, e.g., Poisson process, to characterize flows. While such simplified models may be fine for the old wired Internet, they are not appropriate for big mobile wireless networks. Among the works looking into heavy-tailed distributions, [6] propose to use several heavy-tailed distribution models to characterize the statistical process associated with TCP flows in a wide-area network. In [7] Feldmann suggests Weibull distribution as a better fit than other distribution for modeling of wired TCP flow arrivals. However, these works are mainly based on small wired network and few studies have focused on traffic modeling for big mobile networks. While several works have characterized user and mobility patterns in wireless networks, most of them focused on host-level rather than flow-level. In [8, 9] Tang et al. studied users, network activity and host mobility patterns in a metropolitan-area wireless network and also on a campus department. Other studies at [10, 11] investigated wireless user and AP/building activity and aggregate traffic for the Dartmouth campus wireless network. In [12] Balazinska et al. studied user population characteristics, network usage and load distribution in corporate networks and in [13] Balachandran et al. characterized the aggregate network load and utilization and user patterns during a conference. For flow-level modeling of wireless networks, [1] propose a Weibull regression model to approximate the flow arrivals at individual APs. In another work, [14] found that accessed information by HTTP queries shows spatial locality in a wireless campus network.

On behavioral analysis of mobile networks, there has been a widespread interest for understanding the user behavior. The scope of analysis includes WLAN usage and its evolution across time [9-11] and user mobility [12,15]. Some works focus on using the observed user behavior characteristics to design realistic and practical mobility models [16,17]. In [1] it was shown that the performance of resource scheduling and TCP vary widely between trace-driven analysis and non-trace-driven model analysis. Several other works focus on classifying users based on their mobility periodicity [18], time-location information [19], or a combination of mobility statistics [8]. The work on the TVC model [16] provides a data-driven mobility model for protocol and service performance analysis. The key difference between the previous studies and ours is that we provide an interest-based flow-level traffic modeling approach based on accessed domains and visited locations for big mobile networks.

The two main trace libraries for the networking communities can be found in the archives at [20] and [21]. None of the available traces provides big netflow data coupled with DHCP and WLAN sessions to be able to map IP addresses to MAC addresses and to AP, building and eventually to a context (e.g., history department or a fraternity). Our dataset is significantly larger and richer in semantic than the other mobile wireless network traces and includes around 100 million records per day. Our novel data-driven approach can develop realistic interest-based traffic models to enhance the performance of networking services design for big mobile networks.

One network application for interest-based traffic modeling is profile-based services. Profile-cast [22] provides a new one-to-many communication paradigm targeted at a behavioral groups. In the profile-cast paradigm, profile-aware messages are sent to those who match a behavioral profile. Behavioral profiles in [22] use location visitation preference and are not aware of Internet activity and traffic. Other previous works also rely on movement patterns. Our approach, however, provides interest-based models that have been largely ignored before.

III. Big Data Processing

Realistic traffic modeling and analysis of big mobile networks requires processing of huge amount of network traces. In our study, we process extensive traces collected via all network switches around the campus including netflows, DHCP and wireless session logs. A flow is defined as a unidirectional sequence of packets with some common properties (e.g., source IP address) that pass through a network device (e.g., router) which can be used for flow collection. Network flow records include the start and finish timestamps, source and destination IP addresses, port numbers, protocol numbers, and flow sizes (in packets and bytes). The source and destination IP addresses combined with DHCP logs 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. The DHCP log can be applied as a mapping of dynamically assigned IP addresses to the device MAC addresses. The wireless session log collected by each wireless access point (AP) includes the ‘start’ and ‘end’ events for device associations (when they visited or left that specific AP) which can be used to derive the location of users at any time. The location information can then be applied to find the buildings information.

The variety and scale of different described traces is a major processing challenge. As the size of netflow data is very large (our dataset includes around 100 million flow records per day), a naïve processing approach requires a significant amount of time. To circumvent the problem, we leveraged DataPath [23], a big data processing engine

developed at University of Florida and Rice University that allows complex queries to be defined and executed over TB-sized data in minutes. For this purpose, we used a machine that has 16 Cores, 64GB ram and 2 SSD 1TB disks. DataPath uses novel techniques to make this possible such as on-the-fly code generation, aggressive I/O, push-based data processing, a hybrid column/row store and multi-threaded database operators. Furthermore, DataPath allows seamless integration of aggregation and mining tasks. The setup we used could deal with data rates in excess of 2.6GB/s.

In our study, we first filtered the popular IP prefixes (first 24 bits) using a threshold (the reason for using 24 bits filter is the fact that popular websites usually use an IP range instead of a single IP address). Then, 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 (e.g., 'google', 'facebook', etc.) were considered for the modeling phase. Then, the location of each Internet access (per flow) was identified using the WLAN session logs.

IV. INTEREST-BASED MODELING

A. Domain and Location Specific Analysis

In this section, we study the traffic behavior of big mobile wireless networks considering specific user interests in terms of accessed domains and visited locations. The goal of this study is to find similarities or differences between the behavior of mobile Internet traffic for individual domains or locations, and the overall traffic of the mobile wireless network. For this purpose, we first extract the flow-level traffic distribution for different domains and buildings (per second). Then, we examine the dataset against different statistical distributions to find the best curve fitted to the real distributions. The set of distributions includes Weibull, Rayleigh, Poisson, Negative Binomial, Lognormal, Generalized Pareto, Generalized Extreme Value, Exponential and Gamma. We pick the best fit based on the KS (Kolmogorov-Smirnov) test [2]. The KS test is a nonparametric test for the quality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution. The KS statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. The KS-test has the advantage of making no assumption about the distribution of data. In our experiment, we used a confidence level of 5 percent for the KS test. In addition to domain and location specific modeling, we also find the best fit for overall traffic.

Our study reveals the fact that traffic behavior of different domains follows different types of distributions which forms four categories. On average, 25 percent of domains follow Weibull, 23 percent follow Lognormal, 21 percent follow Generalized Extreme Value and the rest follow other types of distribution. Our study also shows that traffic characteristics of different location are not the same as well. We can again find four major categories of buildings. On average, 35 percent of buildings follow Weibull, 25 percent follow Lognormal, 18 percent follow Generalized Extreme Value and the rest follow other type of distributions. This clearly shows that the best generic fit which is Generalized Extreme Value is not always the best model when considering specific domains or locations.

B. Graph-based Analysis

In the section, we investigate the similarities and differences between the traffic distributions of different domains and locations. While some domains or locations might follow the same type of statistical distribution, their models might follow different parameters. On the other hand, finding the best fit for different domains or locations does not provide us with a quantitative measure to compare their traffic similarities. Therefore, in this part of our study, we provide a method to compare the actual traffic distributions of different domains or locations. For this purpose, we apply another flavor of KS-test that is called Two-sample KS test. The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. This test compares the distributions of the values in the two input data samples. The null hypothesis is that the two samples are from the same distribution. The alternative hypothesis is that they are from different distributions. The two-sample Kolmogorov–Smirnov statistic for samples of size n and n' is:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|$$

where sup_x is the supremum of the set of distances, $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively. The null hypothesis is rejected at significance level α if:

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha.$$

In our study, we run the test at significant level of 5 percent for each pair of domains or buildings. The results form two matrices including a 100*100 matrix for domains and another 68*68 for buildings showing if two domains or buildings follow the same distribution or not. In order to analyze the result, each of the matrices can be interpreted and visualized as traffic similarity graphs. In such graphs nodes represents domains or buildings with an edge between nodes if the corresponding domains or buildings have similar traffic distributions. Figure 1 and 2 shows the resulting graph for domains and locations after running the algorithm presented in [25] for finding the modularity classes within the graphs and applying Fruchterman-Reingold algorithm [25] to form

the graph layouts. In these graphs, the corresponding domain or building for each of the nodes can be found from the node identifier number (mappings between identifier numbers and domains and buildings are available in Figure 3). Modules (or groups) are shown using different colors in the figures. The size of each node represents its degree in the graph that shows uniqueness of traffic distribution of the node compared to the other nodes (low degree is interpreted as uniqueness).

1) Domain-based Analysis

As can be seen in Figure 1 for domain-based analysis, there are 21 domains with unique traffic characteristics. As can be observed, most of very popular domains including 'google', 'facebook' and 'apple' have unique characteristic. In other words, high traffic domains show more uniqueness in terms of their traffic characteristics.

The rest of domains form 12 groups with distinct traffic distributions. Half of the groups have a size of less than 5 and the size of rest is up to 16. Studying different groups reveals many interesting facts. For example, video sharing domains like 'netflix' and 'veoh' show unique characteristics. We can also observe traffic distributions of 'cnn', 'msnbcsport' and 'microsoft' (the group at top-left) are similar. The interesting fact here is that both 'cnn' and 'msnbc' provide news and on the other hand both 'microsoft' and 'msnbc' are provided by the same entity, i.e., Microsoft. This shows that the type of provided content by a domain and also its content provider may affect its traffic distribution. This might also show that, in some cases, when various types of domain attributes (e.g. content type and provider) are appeared together (as in 'msnbcsport') the traffic characteristics of the result is a combination of characteristics regarding those attributes.

Another example of interesting finding is that many of domains related to high-speed Internet and phone providers like 'comcast', 'charter' and 'qwest' have similar traffic distributions (the group in middle-right). Interestingly, we can also find 'shoutcast' in this group which is not in that category but provides a similar type of service, i.e., Internet radio stations (similar in the sense that both phone and radio services provide voice data). Our study shows that the traffic distribution of all domains in this group follows Rayleigh distribution which is different from the generic distribution.

2) Location-based Analysis

Figure 2 shows the resulting graph for the location-based analysis showing three major groups of buildings with distinct traffic characteristics. By looking at the building categories we can again discover different interesting facts. For example, we can observe that more that 70 percent of buildings in Music, Cinema and Auditorium categories are in the same group (the group of nodes in right-bottom). We can also see most fraternities (9 ones) are in the same group with similar traffic characteristics (the big one at the left). This shows that type of location and its context has also have an important effect on characteristics of its traffic distribution. In

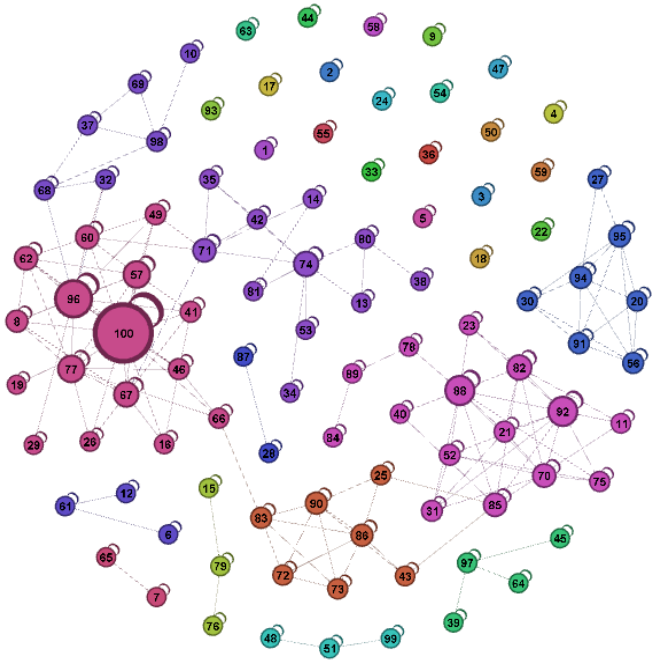

Figure 1. Traffic similarity graph for domains. Nodes represent domains and show their IDs (Domain names can be found in Figure 3). Colors show different detected modules (groups) within the graph. Size of each node shows its degree in the graph.

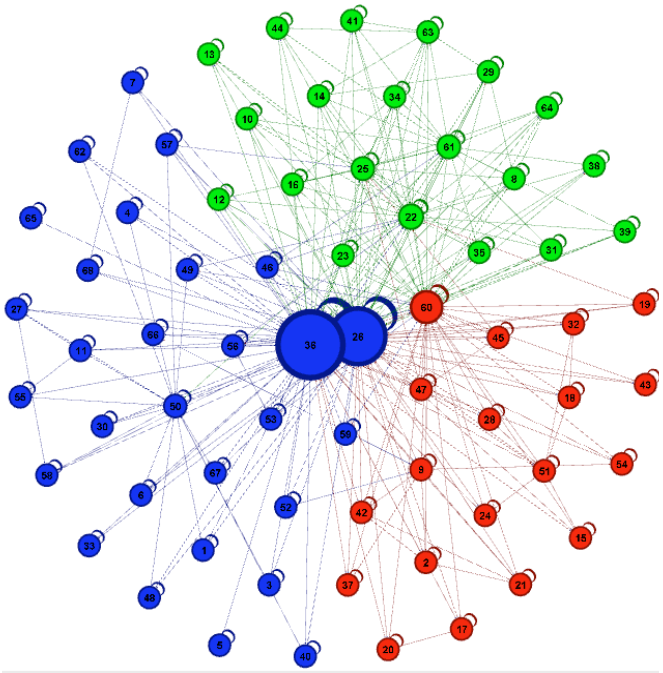

Figure 2. Traffic similarity graph for locations. Nodes represent buildings and show their IDs (Building types can be found in Figure 3). Colors show different detected modules (groups) within the graph. Size of each node shows its degree in the graph.

other words, locations with similar context mostly follow similar traffic characteristics.

importance of using interest-based mining and modeling for big mobile networks.

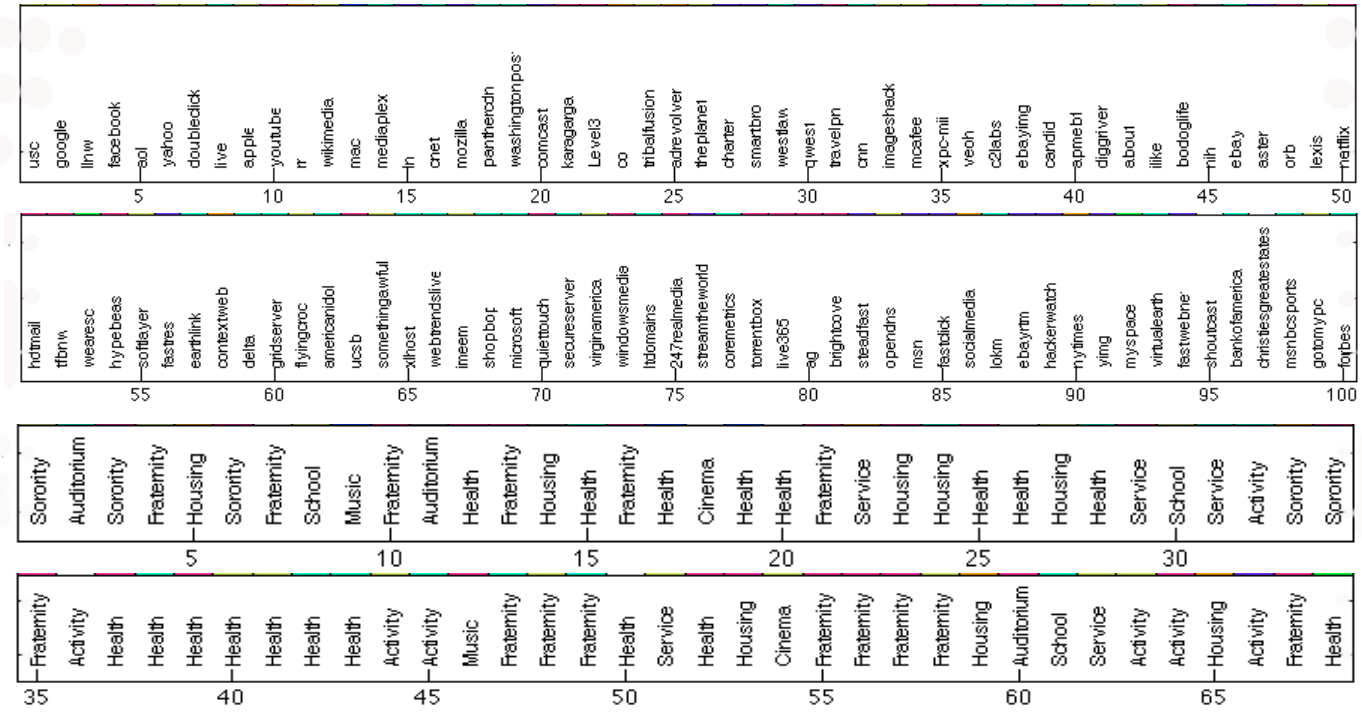

Figure 3. Identifier mappings for domains and buildings

V. ACCURACY ANALYSIS

In this section, we compare the accuracy of generic modeling vs interest-based modeling of big mobile networks. For this purpose, we calculate a weighted average of KS distances between the estimated and the actual distributions for different domains and locations based on their traffic density. Table 1 shows the result of the evaluation.

Table 1- Comparison of generic vs interest-based traffic modeling of big mobile networks based on KS test.

Approach	Domain Access	Location Visitation
Generic	0.5643	0.7427
Interest-based	0.1159	0.1402

As can be seen, KS distance for the interest-based approach based on both accessed domains and visited locations is significantly reduced. The analysis results shows that if we use generic model to reproduce traffic distribution of different domains or locations, the KS distance will be significantly large. The average KS distance for domains will be more than 56 percent and for locations more than 74 percent. However, if we use the proposed interest-based modeling technique the KS distance is reduced to around 11 percent for domains and 14 percent for locations. This means a significant improvement by factor of 5 which shows the

VI. CONCLUSION

This study is motivated by the need for developing realistic mining and modeling methods for big mobile networks. For this purpose, we proposed an interest-based approach based on accessed domains and visited location. Using a novel graph-based technique and two-sample KS test we showed that characteristic of big mobile network traffic largely depends on user interests captured based on domain accesses and location visitations. We also showed that the proposed interest-based approach can significantly improve the modeling accuracy of big mobile networks which is essential to the design of future mobile services and protocols. In future, we plan to provide an interest-based simulation tool for big mobile networks based on the proposed modeling approach.

REFERENCES

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

[2] Kolmogorov, A., Sulla determinazione empirica di una legge di distribuzione, G. Inst. Ital. Attuari, 4, 83, 1933.

[3] Barakat, C., Thiran, P., Iannaccone, G., C.Diot, and P.Owezarski. A flow-based model for Internet backbone traffic. In ACM IMC, 2002.

[4] Fredj, S.B., Bonald, T., Proutiere, A., Regnie, G., and Roberts, J., Statistical bandwidth sharing: A study of congestion at flow level. In ACM SIGCOMM, 2001.

[5] Sarvotham, S., Riedi, R., Baraniuk., R., Connection-level analysis and modeling of network traffic. IMW, 2001.

[6] Paxson, V., Empirically derived analytic models of wide-area TCP connections. IEEE/ACM Transactions on Networking, 2(4):316–336, 1994.

[7] Feldmann, A.. Characteristics of TCP connections. In K.Park and W.Willinger, editors, Self-similar Network Traffic and Performance Evaluation, pages 367–399. John Wiley and Sons, 2000.

[8] Tang, D. and Baker, M. Analysis of a metropolitan-area wireless network.Wirel.Netw.8,2/3 (Nov 2002), 107-120.

[9] Tang, D., and Baker, M.. Analysis of a local-area wireless network. In ACM MOBICOM, 2000.

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

[11] 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.

[12] Balazinska, M. and Castro, P. Characterizing mobility and network usage in a corporate wireless local-area network. ACM MobiSys 2003.

[13] Balachandran, A., Voelker, G.M, Bahl, P., and Rangan, V., Characterizing user behavior and network performance in a public wireless LAN. In ACM SIGMETRICS, 2002.

[14] Chinchilla, F., Lindsey, M.R., and Papadopouli, M.. Analysis of wireless information locality and association patterns in a campus. In IEEE INFOCOM, 2004.

[15] McNett, M. and Voelker, G. M. Access and mobility of wireless PDA users. SIGMOBILE Mob. Comput. Commun. Rev., 9, 2 (Apr 2005), 40-55.

[16] 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.

[17] Jain, R., Lelescu, D. and Balakrishnan, M. Model T: a model for user registration patterns based on campus WLAN data. Wirel. Netw., 13, 6 (Dec 2007), 711-735.

[18] Kim, M. and Kotz, D. Periodic properties of user mobility and access-point popularity. Personal Ubiquitous Comput., 11, 6 (Aug 2007), 465-479.

[19] Eagle, N. and Pentland, A. Reality mining: sensing complex social systems. Personal and Ubiquitous Computing, 10, 4 (May 2006), 268.

[20] MobiLib: http://nile.cise.ufl.edu/MobiLib/.

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

[22] Hsu, W., Dutta, D. and Helmy, A. CSI: A Paradigm for Behavior-oriented Profile-cast Services in Mobile Networks. IEEE/ACM Transactions on Networking.

[23] Moghaddam, S., Parthasarathy, Y., Dobra, A., Helmy, A., Efficient Large-Scale Data-driven Network Mining, The 31st Annual IEEE International Conference on Computer Communications (INFOCOM), March 2012.

[24] Blondel, V. D., Guillaume, J.L , Lambiotte, R., Lefebvre, E. , Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000.

[25] Fruchterman, T. M. J.; Reingold, E. M. Graph Drawing by Force-Directed Placement, Software – Practice & Experience (Wiley) 21 (11): 1129–1164 (1991)