Domain and Location Specific Modeling of Mobile Users Online Interests

Saeed Moghaddam[1,3], Marco Carvalho[2,3], Ahmed Helmy[1]
[1]Computer and Information Science and Engineering Department, University of Florida
[2]Department of Computer Sciences, Florida Institute for Technology
[3]Florida Institute for Human and Machine Cognition
saeed@cise.ufl.edu, mcarvalho@fit.edu, helmy@cise.ufl.edu

***Abstract*—User interests and behavior will play a key role in the design of future mobile networks. In this paper, we introduce a novel technique for the modeling of mobile users interests based on their online activity and mobility. In this study, we conduct domain-specific and location-based analysis and modeling of mobile user interests based on the amount of their online time. Using KS test, we show that domain-specific and location-based models provide more accuracy than a generic model for mobile users interests. The provided models can be applied for design and evaluation of a myriad of interest-based applications and services for future mobile global Internet.**

Keywords—data-driven; interest; modeling; wireless

I. INTRODUCTION

Wireless mobile networks are evolving and becoming increasingly more integrated with every aspect of our lives. Today, laptops, handhelds and smart phones are becoming ubiquitous, providing almost continuous Internet access. This significant shift toward a more persistent mobile Internet access has been accelerating with the rise of larger multi-touch smartphones and tablet computers, which provide a better and easier Internet access experience than previous generations of mobile devices. In fact, the usage of mobile Internet is progressing so fast that it is revolutionizing the entire framework of communication technology. In the last few years, not only has the use of cell phones increased in quite a dramatic way, but the way that people prefer to utilize them, communicate and stay in touch with the world has also changed significantly. People today are ever-increasingly utilizing online services on the move using their mobile devices for different purposes, e.g., listening to music, watching videos, sending and receiving emails, web browsing, and social networking.

This fast growing trend toward mobile Internet access creates a tight coupling between users and mobile networks where various characteristics of user online activities can be captured and applied to provide new human-centered solutions to the problems. However, an important step to achieve this goal is to model the behaviors of mobile users in a systematic way. Such behavioral modeling will provide a foundation for design and evaluation of behavior-aware applications for future mobile networks. In this regard, in our previous works, we proposed different modeling and simulation techniques for mobile Internet usage and users activities based on co-clustering [18], self-organizing maps [19] and Gaussian mixture models [20]. However, all these methods provide tools for high-level modeling and simulation of *usage patterns*, i.e., the aggregated amount of online time at different websites or locations *during a month*. None of the previous works provides mathematical models for the *distribution* of mobile users online time at *specific* websites or locations *during a day*. In this paper, we introduce an approach for domain and location specific modeling of mobile users online activities as the first step toward the design of future mobile communities. In our study, we collected netflow, DHCP and WLAN session logs from different building across campus and then extracted the actual distributions from the real data. Using KS test [21], we mathematically show that domain and location specific modeling provide more accuracy than the generic models for mobile Internet. For example taking Weibull for Google and Facebook and Generalized Extreme Value for YouYube and CNN provides the best result (with minimum KS value) for modeling of the amount of online time. Such realistic models can be applied in design and evaluation of many different applications of future global Internet including interest-aware services, web caching, protocol design and network planning to name a few.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we briefly address the requirements of realistic modeling and in Section 4 we present the proposed approach. Section 5 provides the modeling results based on our case study using campus traces and Section 6 concludes the paper.

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 [1, 2], traffic flow statistics [3], user mobility [4, 5], user association patterns [6] and encounter patterns [7]. Some previous works [4, 7] 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 [8] and [9]. However, none of the available traces provides *netflow* information coupled with DHCP and WLAN sessions to be able to map IP addresses to MAC addresses and detect locations.

978-1-4673-2480-9/13/$31.00 ©2013 IEEE

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 [10-12]. They have shown that most widely used existing mobility models (mostly random mobility models, e.g., random waypoint, random walk; see [13] for a survey) fail to generate realistic mobility characteristics observed from the traces. Realistic mobility modeling is essential for protocol performance. It has been shown that user mobility preference matrix representation leads to meaningful user clustering [14]. Several other works focus on classifying users based on their mobility periodicity [15], time-location information [16], or a combination of mobility statistics. The work on the *TVC* model [11] provides a realistic mobility model for protocol and service performance analysis. In [3] it is shown that the performance of resource scheduling and TCP vary widely between data-driven and non-data-driven model analysis. Using our methods for user modeling we can develop new applications and utilize the realistic models to enhance the performance of networking protocols. Our new user modeling technique may be used in a myriad of networking applications.

One network application for user modeling is profile-based services. *Profile-cast* [17] provides a one-to-many communication technique to send profile-aware messages to those who match a *behavioral profile.* Behavioral profiles in [17] use location visitation preference and are not aware of online Interests. Other previous works also rely on movement patterns. Our domain and location specific 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. Requirements of Realistic Modeling

Developing realistic user models for mobile societies requires three major phases including data processing, data modeling, and evaluation. In the following, we briefly explain each of these different phases.

A. Data Processing

Data processing is the first requirement of the process of users modeling. In order to provide any type of realistic model we first need to collect different types of datasets representing the real environment. In our study, we need to use network infrastructure for collecting different traces of users activities and utilize different online services (e.g., whois lookup service) to cross-correlate all the obtained information from different resources.

As the size of data is very large, this phase needs a considerable amount of time, processing power, memory and storage capacity. After the integration, the data is then aggregated in order to fit the size and format of data to the requirements of user modeling phase.

B. Data Modeling

The second phase is modeling of users activities based on their Internet usage as a measure of interest. The eventual goal of this step is to provide a mathematical model describing the behavior of mobile users at different websites and locations. Such a formulation needs to be in a way that keeps the characteristics of mobile community.

C. Evaluation

The third requirement of realistic modeling is the evaluation of acquired models. For this purpose, we need to compare important properties of the obtained model and actual samples. Without an appropriate evaluation method we will not be able to choose the best user model for realistic design of services for future mobile Internet.

IV. Realistic Users Modeling

A. Data Preparation

In our case study, we collect netflow, DHCP and WLAN session logs from USC campus. These traces are required to provide user, accessed IP address and the Access Point (AP) for each of the interactions. An IP 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 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). The source and destination IP addresses can be used to identify user device MAC addresses using DHCP log and the websites accessed respectively. The DHCP log contains the dynamic IP assignments to MAC addresses and includes date and time of each event. This information is needed to get a consistent 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 collected dataset includes around 70 million flow records per day in average.

B. User modeling

As mentioned before, the main goal of user modeling is to provide a mathematical model describing the behavior of mobile users at different websites and locations. For this purpose, we first need to choose the criteria on which we formulate the users behaviors. In this study, we choose the amount of online time as a measure of users interest to describe their behavior.

After choosing the measure, we need to process and aggregate the data based on the chosen measure to be able to build the models. For this purpose, we aggregated the data on users, domains and locations in terms of their amount of online time per minute during a day. The output of this step shows the amount of online time for each user at different websites and at specific locations (See Table 1).

In our case study, we did the aggregation for the wireless Internet traffic of all active users on 4 popular domains and 4 buildings during a complete day (266 users in total). We performed the aggregation for the total online time per minute (Table 1).

TABLE 1 – PROCESSED DATA SAMPLE

Used ID	Domain ID	Building ID	Online Time for a Day (Min)
11324	142	47	40
11324	386	47	4
11335	142	77	32
11349	386	77	1

Based on the aggregated dataset, we can now build users models based on their interests, i.e., their amount of online time. For this purpose, we first extract the actual distribution of users interest (online time) using the real data and then try to find a mathematical model describing the real interest distribution.

Basically, we can consider two types of modeling approach: a) generic modeling and b) context-specific modeling. A generic model provides a general model for users behaviors on the Internet regardless of their context. This is actually the common approach toward modeling problems in the scope of mobile networks. However, the second type of models, i.e., context-specific models provides customized models tailored for specific context, e.g., in the context of a specific website or a specific location.

In this study, we conduct three types of context specific users modeling; a) domains-specific user modeling, b) location-based user modeling, c) 3D user modeling based on both domains and locations. The main goal in any of above approaches is to study the behavior of mobile users in different contexts in terms of web domains, locations or both.

1) Domain-Specific Modeling

The main goal of domain-specific analysis is to find the best mathematical model describing the distribution of users interest at different web domains. For this purpose, we first extract the actual interest distribution of users at different web domains from the aggregated dataset showed in Table 1. Then, we look for a mathematical formulation describing the dynamics of each of the domains.

2) Location-Based Modeling

Similar to domain-specific modeling, we can find a mathematical model describing the distribution of users interest at different locations. Again, for this purpose, we first extract the actual interest distribution of users at different buildings and then, look for a mathematical formulation describing the dynamics of each of the buildings.

3) 3D User Modeling

Domain-specific or location-based models are based on either users and domain, or, users and locations. In 3D modeling we consider all three aspects of users, domains and locations altogether and provide models for dynamics of specific web domains at specific location. The process of finding the best model is similar to domain-specific and location-based modeling, however, we partition the data based on both domains and locations concurrently.

C. Modeling Evaluation

As explained in the previous section, in order to be able to choose the best user model out of different alternatives, we need an appropriate evaluation method. In our study, we require a mathematical approach to choose the best model describing the actual interest distribution. For this purpose, we apply the KS (Kolmogorov-Smirnov) test [21] to choose the best curve fitted to the real data. 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 shows the distance between the empirical distribution function of the real data and the cumulative distribution function (CDF) of the reference distribution. The best distribution is the one with the minimum distance between empirical distribution function and the cumulative distribution function. This test has the advantage of making no assumption about the distribution of data. In our approach, in order to find the best user interest model, we calculate the KS values considering different types of distributions. The set of distributions includes Weibull, Lognormal, Generalized Pareto, Generalized Extreme Value, Exponential and Gamma.

V. REALISTIC MODELING RESULTS

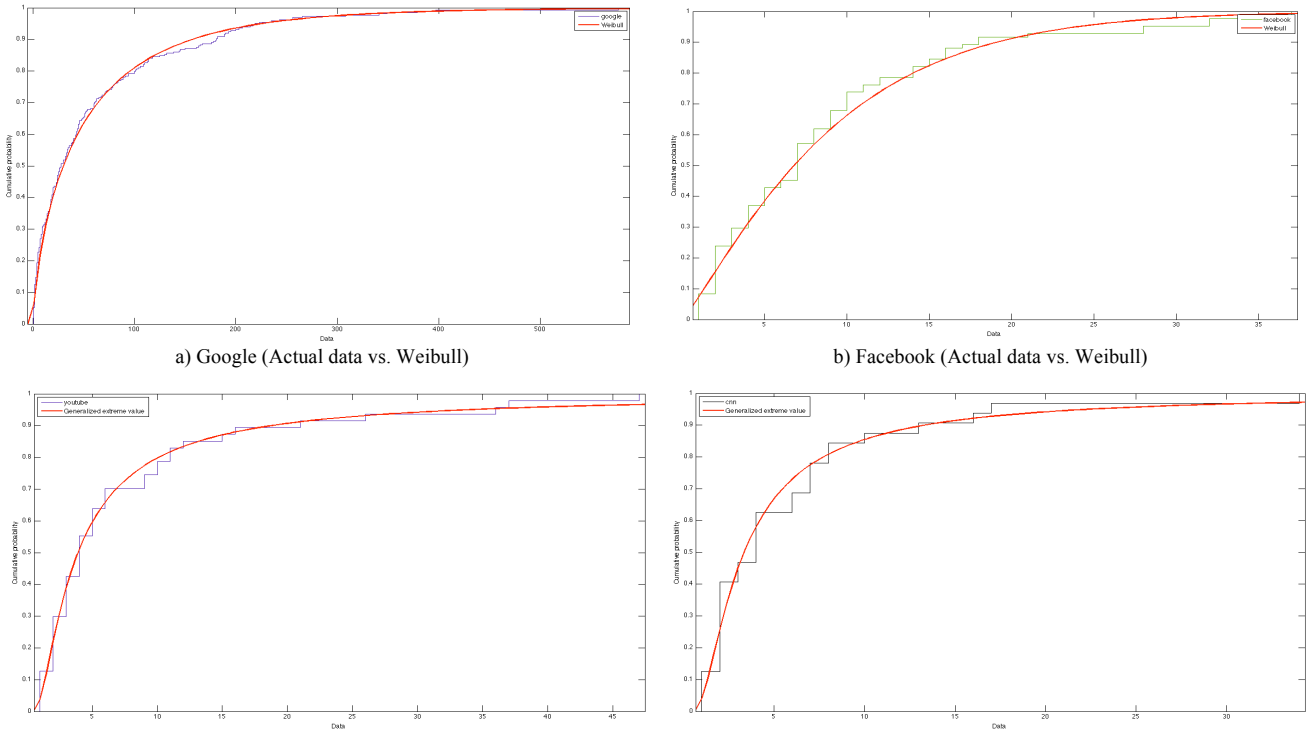
A. Domain-Specific Results

We conduct domains-specific modeling based on the explained approach in the previous sections. In our study, we used the actual data for 4 popular web domains including Google, Facebook, YouTube and CNN. Table 2 shows the resulting KS values for each of the domains for different distributions. As can be seen in the table, Weibull is the best model (with minimum KS value) for Google and Facebook with KS values of 0.0571 and 0.0841 respectively. However, the best fit for YouYube and CNN is Generalized Extreme Value with KS values of 0.0920 and 0.1512 respectively. This shows that we cannot find a generic best fit for all the domains and thus domain-specific user modeling provides more accurate result.

TABLE 2- KS VALUES FOR DIFFERENT WEBSITES AND DISTRIBUTIONS

	Google	Facebook	YouTube	CNN
Gamma	0.0661	0.0995	0.1766	0.1756
Exponential	0.1709	0.1198	0.1707	0.1648
Generalized Extreme Value	0.0929	0.1281	**0.0920**	**0.1512**
Generalized Pareto	0.0898	0.1077	0.1494	0.1833
Lognormal	0.0634	0.1214	0.1049	0.1706
Weibull	**0.0571**	**0.0841**	0.1569	0.1528

Figure 1 shows a comparison between the cumulative density function (CDF) of the actual interest data and the best



a) Google (Actual data vs. Weibull)

b) Facebook (Actual data vs. Weibull)

c) YouTube (Actual data vs. Generalized Extreme Value)

d) CNN (Actual data vs. Generalized Extreme Value)

Figure 1. Comparison of Cumulative Density Function (CDF) for the actual user interest data and the best fitted model for different web domains. Y-axis shows the cumulative probability.

fitted model for different web domains. As can be seen in the picture, as well as inferred from Table 2, these models can approximate the real data with a good accuracy.

B. Location-based Results

We conduct the location-based modeling based on the actual data for 4 different buildings including a health center, housing hall, school and computing center on the campus. Table 3 shows the resulting KS values for each of the buildings for different distributions. As can be seen in the table, Weibull is the best model (with minimum KS value) for the health center and computing center with KS values of 0.0627 and 0.0638 respectively. However, the best fit for the housing hall is Generalized Extreme Value with KS value of 0.0822 and for the school is Lognormal with KS value of 0.1161. This shows again that we cannot find a generic best fit for all the buildings too and thus location-based user modeling provides more accurate result than the generic model.

Figure 2 shows a comparison between the cumulative density function (CDF) of the actual interest data and the best fitted model for different buildings. As can be seen in the picture, as well as inferred from Table 3, these models can approximate the real data with a good accuracy.

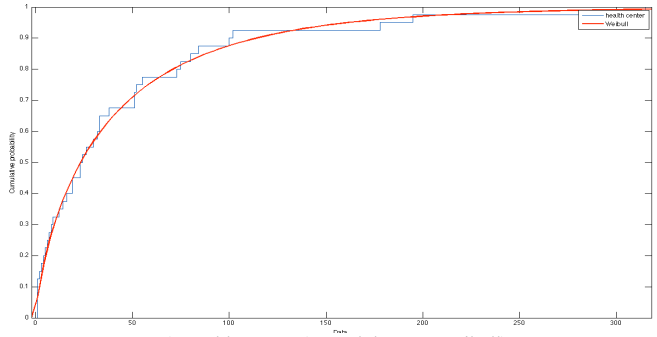
C. 3D Modeling Results

We conduct the 3D modeling based on the actual data for Google at 4 different buildings including the health center, housing hall, school and computing center on the campus. Table 4 shows the resulting KS values for each of the buildings for different distributions. As can be seen in the table, Weibull is the best model (with minimum KS value) for the health center and computing center with KS values of 0.0790 and 0.0742 respectively. However, the best fit for the housing hall is Generalized Extreme Value with KS value of 0.0733 and for the school is Lognormal with KS value of 0.1183. This shows again that we cannot find generic best fit for a specific domain at different buildings, but the best fits for different locations are similar to the location-based modeling result.

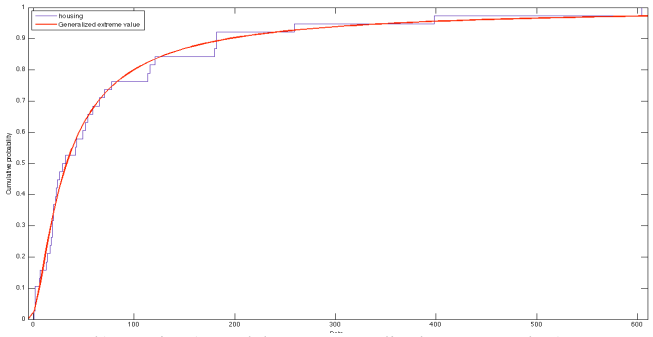
TABLE 3- KS VALUES FOR DIFFERENT LOCATIONS AND DISTRIBUTIONS

	Health Center	Housing	School	Computing Center
Gamma	0.0780	0.1137	0.1384	0.0667
Exponential	0.1439	0.1920	0.2398	0.1559
Generalized Extreme Value	0.1162	**0.0822**	0.1380	0.1104
Generalized Pareto	0.0898	0.0874	0.1469	0.1037
Lognormal	0.1145	0.1203	**0.1161**	0.0954
Weibull	**0.0627**	0.0955	0.1293	**0.0638**

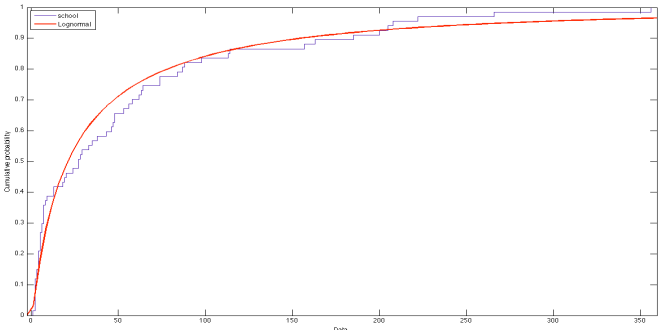
Figure 3 shows a comparison between the cumulative density function (CDF) of the actual interest data and the best fitted model for Google at different buildings. As can be seen



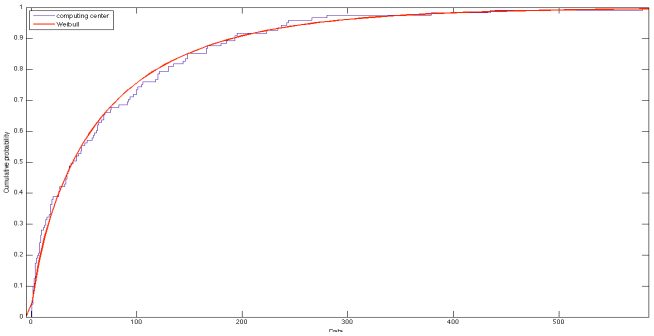
a) Health Center (Actual data vs. Weibull)



b) Housing (Actual data vs. Generalized Extreme Value)



c) School (Actual data vs. Lognormal)



d) Computing center (Actual data vs. Weibull)

Figure 2. Comparison of Cumulative Density Function (CDF) for the actual user interest data and the best fitted model for different buildings. Y-axis shows the cumulative probability.

in the picture, as well as inferred from Table 4, these models again can approximate the real data with a good accuracy.

TABLE 4- KS VALUES OF DIFFERENT DISTRIBUTIONS FOR GOOGLE AT DIFFERENT LOCATIONS

	Health Center	Housing	School	Computing Center
Gamma	0.0865	0.1363	0.1490	0.0778
Exponential	0.1597	0.2037	0.2668	0.1685
Generalized Extreme Value	0.1182	**0.0733**	0.1262	0.1203
Generalized Pareto	0.1097	0.0805	0.1374	0.1125
Lognormal	0.1127	0.1046	**0.1183**	0.0982
Weibull	**0.0790**	0.1037	0.1345	**0.0742**

VI. CONCLUSION

This study is motivated by the need for developing realistic models and efficient services for the future mobile Internet. We provided a systematic method for modeling of mobile users online Interests based on their spent time and different web domains and locations. We have shown that generic models are not the best choice for mobile Internet usage and considering domain and location specific characteristics provide more accurate models. The details of our study enable the parameterization of new and realistic models for future mobile Internet with applications in several areas of networking including interest-aware services, web caching, simulation and evaluation of protocols.

ACKNOWLEDGMENT

This material is partially based upon work supported by the Department of Energy, National Energy Technology Laboratory under Award Number DE-OE0000511.

REFERENCES

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

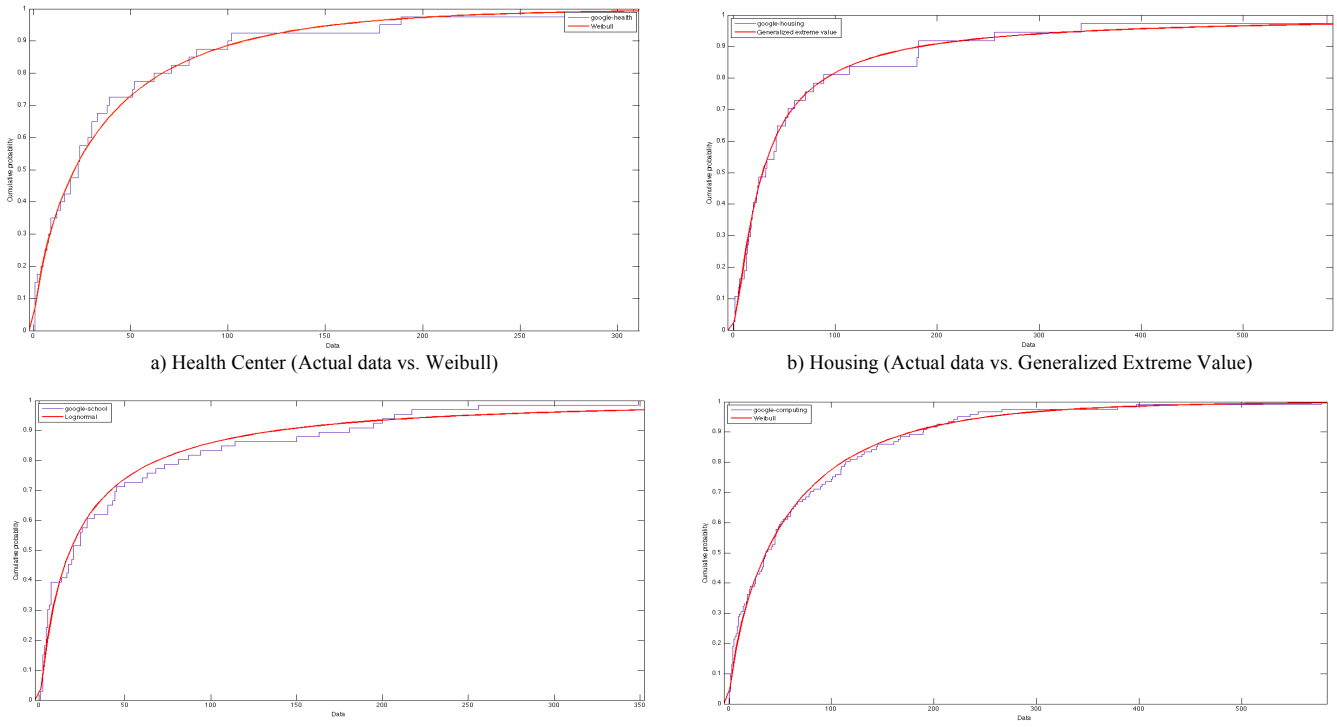
[2] Henderson, T., Kotz, D. and Abyzov, I. The changing usage of a mature campus-wide wireless network. Computer Networks, 52, 14 (Oct 2008), 2690-2712.

[3] 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.

[4] Hsu, W. and Helmy, A. On modeling user associations in wireless LAN traces on university campuses. In Proceedings of the IEEE Int'l Workshop on Wireless Network Measurements(WiNMee) (Apr, 2006).

[5] 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.

[6] Papadopouli, M., Shen, H. and Spanakis, M. Characterizing the duration and association patterns of



a) Health Center (Actual data vs. Weibull)

b) Housing (Actual data vs. Generalized Extreme Value)

c) School (Actual data vs. Lognormal)

d) Computing center (Actual data vs. Weibull)

Figure 3. Comparison of Cumulative Density Function (CDF) for the actual user interest data and the best fitted model for Google at different buildings. Y-axis shows the cumulative probability.

wireless access in a campus. In Proceedings of the 11th European Wireless Conference (Nicosia, Cyprus, Apr, 2005).

[7] Hsu, W. and Helmy, A. On nodal encounter patterns in wireless LAN traces. In Proceedings of the IEEE Int'l Workshop on Wireless Network Measurements(WiNMee) (Apr, 2006).

[8] MobiLib: Community-wide library of mobility and wireless networks measurements (Investigating user behavior in wireless environments). Available: http://nile.cise.ufl.edu/MobiLib/.

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

[10] 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.

[11] 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.

[12] 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).

[13] Bai, F. and Helmy, A. A Survey of Mobility Modeling and Analysis in Wireless Adhoc Networks, Wireless Ad Hoc and Sensor Networks, Springer, 2006.

[14] Hsu, W., Dutta, D. and Helmy, A. Mining behavioral groups in large wireless LANs. In Proceedings of the ACM MobiCom 2007 (Montral, Qubec, Canada, 2007). ACM.

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

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

[17] 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.

[18] Moghaddam, S., Helmy, A., Ranka, S. and Somaiya, M. Data-driven co-clustering model of Internet usage in large mobile societies. In Proceedings of the ACM MSWiM 2010 (Bodrum, Turkey, Oct, 2010). ACM.

[19] Moghaddam, S. and Helmy, A. Multidimensional modeling and analysis of wireless users online activity and mobility: A neural-networks map approach. ACM MSWiM 2011 (Oct, 2011). ACM.

[20] Saeed Moghaddam, Ahmed Helmy: SPIRIT: A simulation paradigm for realistic design of mature mobile societies. IWCMC 2011: 232-237.

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