SPIRIT: A Simulation Paradigm for Realistic Design of Mature Mobile Societies

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

Abstract—**In very close future, human interests and behavior will play a central role in design of mobile networks. Thus, simulation of user behaviors is imperative for the design and evaluation of future mobile networks. However, multi-dimensionality of human interests makes it difficult for us to provide realistic simulations. In this paper, we address challenges in this regard and propose a multivariate technique based on Gaussian mixture models as the first step for simulation of mobile users' interests considering website visitations. We also introduce an evaluation technique for measuring the quality of high-dimensional simulation output. Using our collected dataset including billions of WLAN netflow records for 100 web domains at 79 buildings, we show that our technique not only keeps the characteristics of hidden behavioral groups, but also is flexible enough to approximate the real dataset with given required accuracy.**

Keywords- Mixture models; Wireless networks; Internet usage

I. INTRODUCTION

Mobile networking is going to play a central role in supporting many important human activities. The explosion in the availability of information through tens of thousands of Internet websites means, on one hand, that the amount of data we can collect on mobile users behavior will continue to increase, and on the other hand, that we need to develop realistic paradigms to model and simulate multi-dimensional behaviors. Such modeling and simulation is imperative for design of novel behavior-aware network protocols (e.g., for targeted announcements/ads). However, for multi-dimensional behaviors and interests (e.g, websites visitations) classic univariate paradigms for network modeling and simulation are not appropriate. Applying these models means producing the outputs for different measures of interest separately. This method will lose significant information on potential correlations between different variables, e.g., web domains, and will lead the simulation to be far from the reality of mobile society. On the other hand, our study on billions of online activities reveals the fact that there exist different kinds of behavioral groups in a mobile society. For example, there exist groups with narrow website access and other groups with wide spread access. Keeping the characteristics of the behavioral groups in addition to the general characteristics of the behavioral space is another important requirement for realistic simulation of mature mobile networks.

Moreover, human behaviors can be viewed from several aspects, e.g., web accesses, location visitations and application usage. It is not far from realistic, if we expect to find inter-aspect behavioral patterns in addition to the intra-aspect ones. For example, considering web accesses and location visitations, we might find inter-aspect correlations (e.g., people who visit domain X at location A visit domain Y at location B too) or inter-aspect characteristics (e.g., locations X, Y, Z are similar on web access patterns for domains A, B, C, D). This possibility becomes exciting when sparking the idea that there might be even distinct multi-aspect multi-dimensional elements in human behavioral space, capable of explaining the whole space much like what chemical elements in periodic table do. Identifying such elements will be indeed a huge advancement toward the understanding of mobile societies. We might be able to explain very complex behavioral patterns using a mixture of these elements much like what chemists do for explaining the chemical reactions and the nature of substances. Maybe it is time for us to give up explaining the whole nature by just the (observable) classical elements of earth, water, air and fire. At this point, we are indeed in need of a significant paradigm shift from simple modeling techniques toward much more mature ones to be able to capture and simulate the spirit of dynamics within the mobile societies.

In this paper, we propose our simulation paradigm (SPIRIT) based on Gaussian Mixture Models (GMM) [27] for multivariate simulation of mobile users' behavior and interests. Multivariate distributions are generalization of one-dimensional (univariate) distributions to higher dimensions. The multivariate Gaussian distribution is often used to describe, at least approximately, any set of (possibly) correlated real-valued random variables. However, this classic model represents the whole feature (interest) distributions by just one position (mean vector) and an elliptic shape (covariance matrix). Therefore, it may simply ignore some underlying set of hidden behavioral classes and thus do not provide an accurate approximation of the distribution. To remedy this problem, mixture models comprise a set of component functions for modeling multiple classes of sample distributions. A GMM uses a discrete set of Gaussian functions as the components to provide not only a smooth overall distribution fit, but also details for multi-modal nature of the density.

For evaluating our simulation paradigm (SPIRIT), we apply a dataset provided by processing of extensive netflow, DHCP and WLAN session logs for more than 22 thousand mobile users in a Wireless LAN spanning over 79 buildings (including over 700 APs), that we have collected. This original dataset includes billions of records, represents by far the largest set of traces analyzed in any study of

This work was partially funded by NSF award number 0832043.

mobile networks to date. In our case study, we apply GMM for multivariate simulation of users interests on the top 100 active web domains.

Our work has the following key contributions:

1. We propose an effective approach for multi-dimensional modeling and simulation of mobile users interests extracted from one of the largest set of mobile network usage traces (including billions of records) and show how Gausain mixture models can be applied to keep behavioral characteristics.
2. We suggest an evaluation technique based on *Pearson's chi-square test* for measuring the accuracy of multivariate simulation output and show how the involved complexity problem for high-dimensional datasets can be resolved in practice.
3. We analyze how the choice of number of components affects the accuracy of simulation output and show how the chosen number of bins affects the evaluation scheme.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we briefly address challenges associated with collection and processing of large-scale wireless traces and then explain our modeling, simulation and evaluation technique in detail. Section 4 provides our case study using campus traces and the experimental results. Section 5 discusses applications and Section 6 concludes

II. RELATED WORK

The rapid adoption of wireless communication technologies and devices has led to a widespread interest in analyzing the traces to understand user behavior and to simulate their behaviors. The scope of analysis includes WLAN usage and its evolution across time [1-3], user mobility [4, 5], traffic flow statistics [6], and encounter patterns [7, 8]. Some previous works [4, 7] explore the space of understanding realistic user behaviors empirically from data traces. The two main trace libraries for the networking communities can be found in the archives at [9] and [10]. None of the available traces provides large-scale *netflow* information coupled with DHCP and WLAN sessions to be able to map IP addresses to MAC addresses to AP to location and eventually to a context (e.g., history department). Therefore, (to the best of our knowledge) our work represents the first one to address large-scale multi-dimensional modeling and simulation of wireless and mobile societies while providing finer granularity, richer semantics and more accuracy.

There are several prominent 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 and practical mobility models [11-14]. They have shown that most widely used existing mobility models (mostly random mobility models, e.g., random walk, random waypoint; see [15] for a survey) fail to generate realistic mobility characteristics observed from the traces. Realistic mobility modeling and simulation is essential for protocol performance [16]. It has been shown that user mobility preference matrix representation leads to meaningful user clustering [17]. Several other works with focus on classifying users based on their mobility periodicity [18], time-location information [19, 20], or a combination of mobility statistics [21]. The work on the *TVC* model [11] provides a data-driven mobility model for protocol and service performance analysis. In [6] it was shown that the performance of resource scheduling [22] and TCP vary widely between trace-driven analysis and non-trace-driven model analysis. Using multi-dimensional modeling, our simulation technique can help to develop new mobility-aware Internet-usage models, and utilize the realistic profiles to enhance the performance of networking protocols. Our propose simulation paradigm has the potential to incorporates web activity, location and mobility, and provides user profiles that may be used in a myriad of networking applications.

One network application for multi-dimensional simulation is profile-based services. *Profile-cast* [23, 24] 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 [23, 24] use location visitation preference and are not aware of Internet activity. Other previous works also rely on movement patterns. Our multi-dimensional simulation 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. SIMULATION APPROACH

Developing a realistic simulation paradigm for mobile societies requires four main phases. 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 third phase is modeling of users' interests based on their web domain visitation patterns. The fourth phase includes generating simulated data based on the acquired parameters for the model and evaluating the quality of simulation output.

A. Data Collection

We collect different types of extensive traces via network switches (in USC campus) including netflows, DHCP and wireless session logs. 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) (see Table 1). 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.

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 would be in the order of months. For example, the netflow dataset gathered from USC campus includes around 2 billions 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, location and month and calculate the total online time for each resulting record.

C. *Data Modeling*

Gaussian Mixture Model (GMM) is a type of density models which comprise a number of Gaussian component functions. A mixture of *K* Gaussian is defined as follows:

$$p(x) = \sum_{k=1}^{K} \alpha_k G\left(x, \mu_k, \Sigma_k\right)$$

where α_k is the mixing parameter satisfying $\Sigma\alpha_k = 1$ and $G(x,\mu_k,\Sigma_k)$ is the probability density function (pdf) for the k^{th} Gaussian component. The Gaussian mixture model contains the following adjustable parameters: α_k, μ_k and Σ_k. For estimating the parameters of the GMM which in some sense best matches the distribution of the training input patterns, we use maximum likelihood (ML) estimation method. The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. We apply Expectation Maximization (EM) algorithm for finding the maximum likelihood. EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the latent variables, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. (Latent variables are variables that are not directly observed but are rather inferred from other variables that are observed (directly measured)).

D. *Data Simulation and Evaluation*

After estimating the GMM parameter, we can simply produce synthetic data based on the model. A simulated data point can be generated by first choosing one of the multivariate Gaussians (with the probability of α_k) and then sampling based on the parameters for the chosen distribution (μ_k and Σ_k).

For the evaluation of simulation output, we apply two methods. In the first method, we first cluster the real dataset into a number of behavioral groups and acquire the distribution of real samples in different groups. Then, using the same clusters, we partition the synthetic samples and obtain their distribution over different clusters. Finally, we compare the two acquired distribution to see if the simulation output keeps the behavioral groups or not.

Although the first method provides a general insight on the quality of simulation output, it can be controversial in the sense that the quality metric depends on the clustering technique we use. Hence, we propose a second evaluation technique which is essentially based on *Pearson's chi-square test* [28] with some modifications. Pearson's chi-square is used to assess goodness of fit for a dataset. The test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution. In our test, we partition N observations (real samples) as well as N simulated samples into k sub-space (bins). Then we verify the hypothesis that, in the general population, real and synthetic samples would occur in each bin with equal frequency. The amount of discrepancy from this hypothesis is generally measured using the following formula:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where X^2 is the Pearson's cumulative test statistic, O_i is the observed frequency from the real dataset, E_i is the theoretical frequency from the simulation and k is the total number of bins.

This metric is easy to calculate when dealing with one-dimensional data as we can simply partition the data space using a few number of bins. However, for high dimensional

Table 1. 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

data the number of required bins grows exponentially fast. *For an n-dimensional data space we require b to the power n (b^n) bins to equally partition the space considering b sub-range along each dimension (e.g., 10^15 bins for 15 dimensions and 10 sub-ranges).* This means that the computational complexity of calculating this metric and the required memory is significantly high if we want to deal with huge number of bins. To resolve this problem, we suggest to just keep track of non-empty bins as in practice a huge number of bins will remain empty. For each new sample, we look into the set of non-empty bins which we keep track of. If we do not find the corresponding bins, we create a new one and keep track of that bin afterward. This way we can simply resolve the complexity and memory problem.

However, on the other hand, this metric is based on the assumption that sufficient samples will exist in each bin, which found is not always true in our case. This metric is essentially designed to show the fraction of samples that deviate from the analytical model; the smaller the value, the better the model. Therefore, for each bin it tries to calculate the fraction of deviated samples. The calculated fraction should not be more than the total number of samples in the bin, which is approximately true when having sufficient samples in the bin. However, when this assumption is not true for a bin, we will get wrong result. For example, if Ei=1 Oi=5, we get 16 from the formula while the result should not be more than 1 (the total number of verified samples). To resolve this problem, we modify the metric as follows:

$$X^2 = \sum_{i=1}^{k} \min\left(\frac{(O_i - E_i)^2}{E_i}, E_i\right)$$

Using the above discrepancy measure, we define simulation accuracy as follows:

$$Accuracy = 1 - \frac{X^2}{N}$$

IV. CASE STUDY

In our case study, we collected data from the University of Southern California (USC) in 2008 and conduct the simulation 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 using MAC addresses, and destinations, or 'peers' (usually web servers) using IP address prefixes. Over a billion records (for the month of March 2008) were considered initially, then the February and April traces (over two billion records) were considered for the stability analysis. The IP prefixes (first 24 bits) were filtered using a threshold of 100,000 flows (the reason for using 24 bits filter is the fact 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. Then, a dataset was created describing the total online time of all users (22,816) at different web domains (per minute). The data is finally scaled using row-normalization of log the online time values. This dataset forms our real data samples.

B. . Simulation Result

We applied our proposed method for simulating the real data samples. Figure 1 shows the clustering result on the real dataset. As can be seen, we can identify different behavioral groups based on users' interests. Figure 2 shows the distribution of users in different groups for the real dataset and two simulated datasets; one based on the classic multivariate normal distribution and the other based on our proposed technique. As can be seen in the figure, our technique is able to generate almost the same distribution for the behavioral groups while the other one fails to do so.

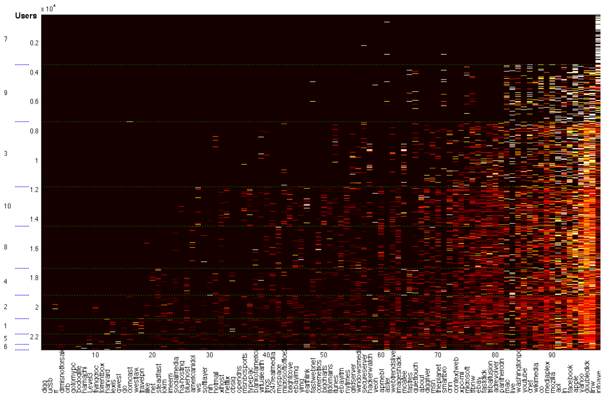
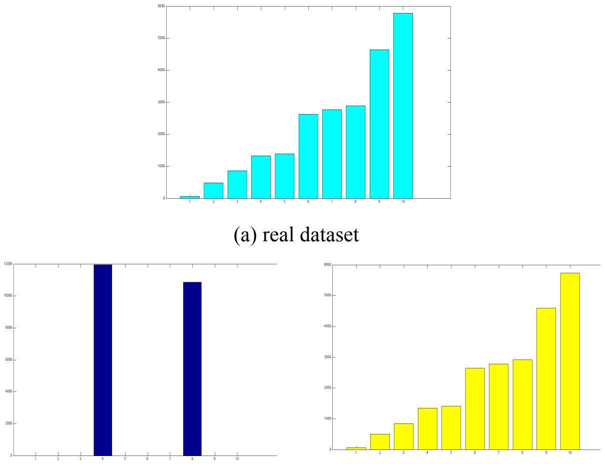


Figure 1. Behavioral clusters for the real samples. X-axis shows domain names and y-axis shows users and cluster IDs.



(a) real dataset

(b) simulated dataset using normal distributaion

(c) simulated dataset using our proposed techniqe (GMM)

Figure 2. Distribution of real and simulated samples over behavioral clusters (sorted by the number of users).

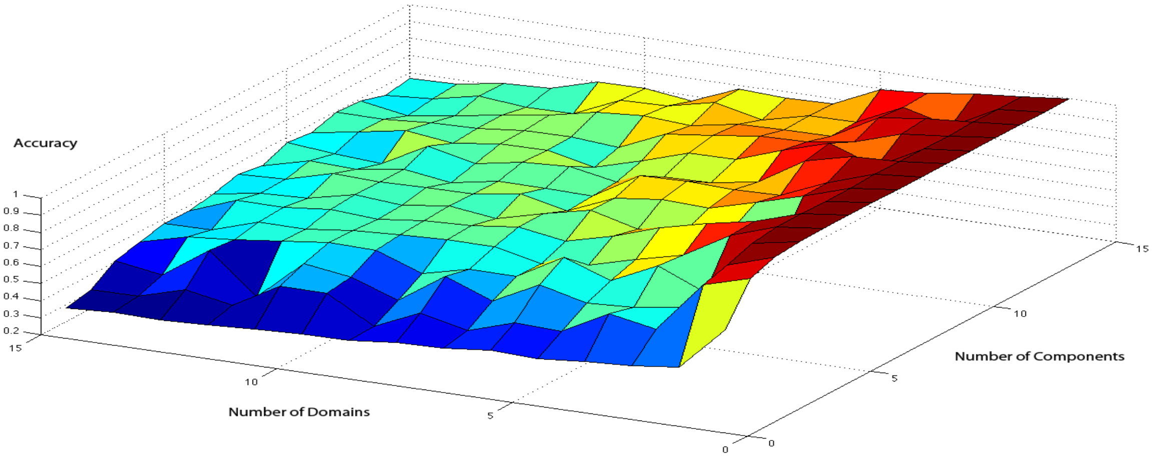
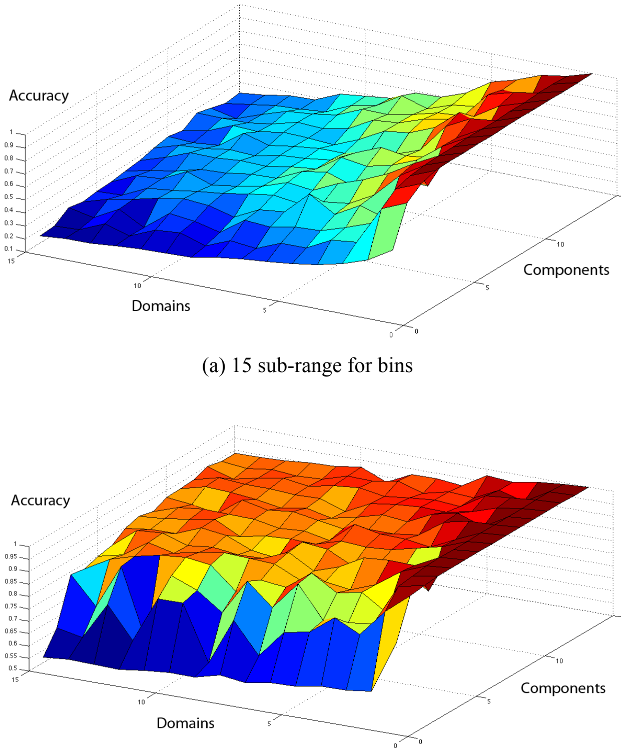


Figure 3. Simulation accuracy for different number of domains and components considering 10 sub-ranges along each dimension for creating the bins

For measuring and analyzing the simulation accuracy, we repeated the simulation process 225 times for different numbers of domains and components from 1 to 15. For each case, we measured the accuracy, using different numbers of bin for 2 to 15 sub-ranges along each dimension. Figure 3, Figure 4(a) and Figure 4(b) show the simulation accuracy for different numbers of domains and components while the sub-range along each dimension is set to 10, 15 and 5 respectively. As can be seen in Figure 3, generally speaking, as the number of components increase, the accuracy of simulation increases too. Therefore, depending upon the degree of required accuracy, we can choose an appropriate number of components for the GMM. For example, for 15 domains the simulation accuracy varies from less than 30 percent to around 60 percent for 1 to 15 components. However, as can be noticed in the figure, the level of accuracy is not always increasing and there exist several local maximum. For example, for 15 domains, we get the maximum accuracy at 12 in the range of 1 to 15 components. Another important point, which can be inferred from the figure, is the fact that as the dimensionality of data increases; the simulation accuracy decreases for the same number of components. This fact pretty good shows the deficiency of uni-component simulation techniques for very high dimensional data. As can be seen in Figure 4, these findings hold true for different number of sub-ranges along the dimensions for creating the bins. However, taking more bins into consideration leads to a smoother and more realistic measurement.

Figures 5 and 6 show how the number of bins affects the accuracy metric. As can be seen in Figure 5, for 10 domains and the same number of components (which in fact leads to the same simulation output), higher number of bins lead to lower accuracy measure, but they can better differentiate the quality of different simulation schemes. For example, the left-side columns in the graph (for lower numbers of bins) suggest not much accuracy difference for different numbers of components, while the right side of the graph (for higher numbers of bins) reveals a significant difference. We can infer a similar relationship for the number of domains and bins from Figure 6. The noticeable difference in the two graphs is basically because of the fact that higher numbers of domains lead to lower accuracy in general but the higher numbers of components lead to higher amounts.



(a) 15 sub-range for bins

(b) 5 sub-range for bins

Figure 4. Simulation accuracy for different number domains and components considering 15 and 5 sub-ranges along each dimension for creating the bins

V. Discussion: Applications

The systematic realistic simulation method proposed in this paper can be applied with any set of wireless data and can be used in several important applications in mobile networking research. Here, we briefly address two such major applications:

1- Interest-based protocols and services: A new class of protocols and services center around user-interest and similarity, including profile-cast, participatory sensing [25], trust establishment [26], location-based services, crowd sourcing, alert notification and targeted announcements and ads. So far, mobility patterns (e.g., in profile-cast) 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 developed simulation method can help both the informed design of such efficient protocols and the realistic evaluation thereof.

2- 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 users web access patterns which can be provide through our simulation technique. These applications for mobile networks are becoming more compelling with the significant growth of usage of smart phones, iphones, ipads, and the like.

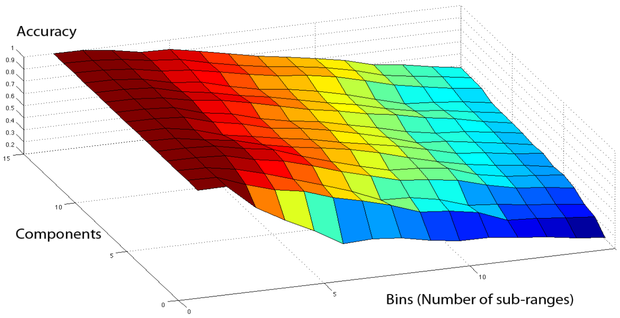


Figure 5. Simulation accuracy for different number of components and bins considering 10 domains

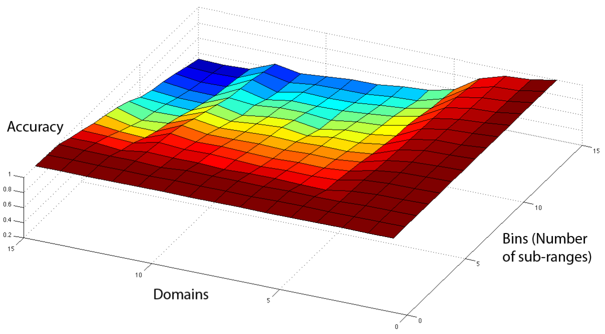


Figure 6. Simulation accuracy for different number of domains and bins considering 10 components

VI. Conclusion

This study is motivated by the need for developing realistic simulation paradigms required for design of efficient protocols and services for the future mobile Internet. We provided a systematic method to process the largest wireless trace to date, with billions of records of Internet usage from a campus network, and simulate web interests for thousands of users. We have shown that mobile Internet usage can be simulated using Gaussian mixture models with enough flexibility to acquire the required accuracy. Our study is the first step toward realistic simulation of multi-dimensional users behaviors and interest with many applications in several areas of networking, including mobile web caching, evaluation of protocols, interest-aware services and network planning, to name a few. We hope for our method to provide an example for realistic simulation of mobile societies and lead to a paradigm shift in simulation techniques in the future. With more measurements from mobile and sensor networks becoming available, we expect our method to get extended and matured in order to facilitate simulation of many other large datasets in future studies.

REFERENCES

[1] Tang, D. and Baker, M. Analysis of a local-area wireless network. In *Proceedings of the ACM MobiCom 2000* (Boston, Massachusetts, United States, Aug, 2000). ACM.

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

[3] 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.

[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] 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.

[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] Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R. and Scott, J. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*(Jun 2007), 606-620.

[9] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements (Investigating User Behavior in Wireless Environments). Available: http://nile.cise.ufl.edu/MobiLib/.

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

[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] 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.

[13] 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.

[14] 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).

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

[16] Bai, F., Sadagopan, N. and Helmy, A. The IMPORTANT framework for analyzing the Impact of Mobility on Performance Of RouTing protocols for Adhoc NeTworks. *Ad Hoc Networks*, 1, 4 (Nov 2003), 383-403.

[17] 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.

[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] Ghosh, J., Beal, M. J., Ngo, H. Q. and Qiao, C. On profiling mobility and predicting locations of wireless users. In *Proceedings of the 2nd international workshop on Multi-hop ad hoc networks: from theory to reality* (Florence, Italy, 2006). ACM.

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

[22] Borst, S. User-level performance of channel-aware scheduling algorithms in wireless data networks. *Ieee-Acm Transactions on Networking*, 13, 3 (Jun 2005), 636-647.

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

[24] 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.

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

[26] 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.

[27] McLachlan, G.J. and Peel, D, Finite Mixture Models, Wiley (2000).

[28] Plackett, R.L., Karl Pearson and the Chi-Squared Test. International Statistical Review (International Statistical Institute (ISI)) 51 (1): 59–72 (1983).