A Global Local Modeling of Internet Usage in Large Mobile Societies

Abdullah Almutairi
Computer and
Information Science and
Engineering Department
University of Florida
Gainesville, Florida 32611
USA
aalmutai@cise.ufl.edu

Manas Somaiya
Computer and
Information Science and
Engineering Department
University of Florida
Gainesville, Florida 32611
USA
manas@computer.org

Saeed Moghaddam
Computer and
Information Science and
Engineering Department
University of Florida
Gainesville, Florida 32611
USA
saeed@cise.ufl.edu

Sanjay Ranka
Computer and
Information Science and
Engineering Department
University of Florida
Gainesville, Florida 32611
USA
ranka@cise.ufl.edu

Ahmed Helmy
Computer and
Information Science and
Engineering Department
University of Florida
Gainesville, Florida 32611
USA
helmy@cise.ufl.edu

ABSTRACT

Real-world wireless Internet usage data for any user is typically generated via an overlap of many correlations. These correlations could be based on hobbies (e.g. sports-fan), profession (e.g. work e-mail), day-to-day activities (e.g. news, Internet banking), communication (e.g. instant messaging, social networks), etc. The likelihood of appearance of these correlations in usage data may be influenced by the type of location the user is in. Hobbies and communication related web sites would be more likely to be accessed at home, Profession related web sites would usually be accessed at work. Understanding and capturing this generative process that is based on human interests, behavior and location is the key to the design of future mobile networks. We propose a novel Bayesian mixture model called the “Global Local” model based on the “POWER” model that can realistically describe Internet usage and correlations with various locations inside a large mobile society. The “POWER” model is a new class of mixture models where components compete to produce a single data point, this model allows for the discovery of complex overlapping patterns of user’s Internet behavior. The “Global Local” model learns a global template of user’s Internet behavior patterns using the “POWER” model first, then learns correlations between the templates and locations inside a large mobile society. We design a learning algorithm that can effectively learn the “Global Local” model from Internet usage data, and demonstrate its capabilities using synthetic data. Finally, we analyze a real-world Internet usage data for thousands of users collected via wireless LAN traces and discover many interesting correlations that can be explained very intuitively.

PM2HW2N'12, October 21–22, 2012, Paphos, Cyprus.
Copyright 2012 ACM 978-1-4503-1626-2/12/10 ...$15.00.

Categories and Subject Descriptors

C.2.5 [**COMPUTER-COMMUNICATION NETWORKS**]: Local and Wide-Area Networks—*Internet*; I.6.5 [**SIMULATION AND MODELING**]: Model Development—*Modeling methodologies*

Keywords

wireless networks, data-driven modeling, mixture models

1. INTRODUCTION

Wireless mobile networks are extensively used in many places throughout the world. They are ever growing to the point of being ubiquitous. The majority of people nowadays carry a device that can connect to these networks. This takes a huge toll on these networks and strains their ability to support the load and demand of their users. By using data-driven modeling and new design paradigms[19], better context-aware network protocols and services can be designed based on the users’ Internet usage behavior. In order to help build these network protocols and services, a “Global Local” model is proposed to model the location-based usage in a mobile society. The Internet usage behavior may be influenced by the location in the mobile society that the behavior was observed in. For example, profession-related web domains (e.g. work e-email) are most likely to be used in the work location. Hobby-related web domains are most likely to be used in the user’s residence. We wish first to learn the classes of usage behavior in accessing web domains from the overlap of usage patterns in the wireless data. Then, we wish to investigate whether, and to what extent, does the type of location a user is in can influence his Internet usage behavior and correlate with the type of web domain a user is visiting. Also, we wish to know the probability of each usage behavior class appearing in the locations. This gives the level of relationship between each location and all usage behavior classes.

To study the users' Internet usage behavior, an extensive *netflow*, DHCP and MAC trap traces for thousands of mobile users in a WLAN spanning over 79 buildings and including over 700 APs, were collected and processed[19]. This is the largest set of traces processed in any study of mobile networks to date. Moghaddam et al.[19] provided a systematic method to process the billions of records in this *netflow* to integrate and aggregate the multi-dimensional data.

The Global Local model mainly uses a generic Bayesian framework called the PrObabilistic Weighted Ensemble of Roles Model, or "POWER" model[24] for short. This framework is a new class of mixture models[7, 17] where multiple components can contribute to the generation of a single data point while simultaneously allowing each component to have a varying degree of influence on different data attributes. One of the challenges of using the classic mixture model with high dimensional data is that it allows only a single mixture component to generate each data point. There are many real-world high dimensional datasets where it makes more sense to model a data point as being generated using multiple overlapping components. An unintended consequence of the single-component data generation is that a component can not limit its influence to only a subset of the data attributes, making it difficult to capture patterns in data subspaces. Consider the scenario of building an informative model for the web usage patterns of users at a university campus given that we have usage logs for each user. Under the classic mixture model, we would assume that each user belongs to only one class. Membership in a given class should attempt to completely describe all of the web surfing patterns of each member user. Given the diversity in surfing patterns and the variety of websites, this is highly unrealistic.

In real life, a user could belong to many classes like news-junkie, social-network-fan, movies-fan, sports-fan, hacker, and gaming-enthusiast, and the usage patterns may be influenced by one or more of these classes. Hence, it makes more sense to model the behavior of each user as resulting from the influence of several classes. Now, consider a user that is a movies-fan, a sports-fan, a gaming-enthusiast, and a hacker. As this user is surfing the web, he finds a new gaming website based on a popular sport. It seems obvious that membership in both the sports-fan and the gaming-enthusiast classes should be relevant to the decision of visiting this website, but membership in the hacker or movies-fan classes should not be. Hence, we can conclude that it is probably not realistic for each class to influence each and every one of a userŠs website visits. Based on the concepts of multi-class membership and that each class should only influence a subset of data attributes, a generative process would allow each data point to be modeled with high precision, while still learning very general roles such as hacker and sports-fan that are important, and yet cannot describe any data point completely.

The "POWER" model suffers from a very long learning time. It has been shown that it takes 510 hours to learn the model on a data set collected from Reuters news stories (22,429 stories, 21 mixture components, 1,000 word attributes) using a 32-core machine[24]. Running the netflow dataset on the model will be impractical. Therefore, a faster version of the "POWER" model was developed to handle the large wireless data in a timely fashion[2]. The slow learning time is mainly due to the time complexity of the model which is $O(n \cdot k \cdot d^2)$, where n is the number of data points, k is the number of components and d is the number of data attributes. This operational cost is specifically inflicted while updating the weight parameter of the model, which dictates the varying influence of a mixture component on a data attribute, in the Gibbs sampler. This faster version is an approximated version that reduces the time complexity of learning the "POWER" model to $O(n \cdot k \cdot d \cdot p)$, where p is an approximation parameter that is much smaller than d. This allows for learning the model in a linearly scalable time w.r.t. the number of attributes, making the model practical for use on very large datasets such as the netflow dataset.

The contributions of this paper are:

- We present a novel Global Local model that learns the Internet usage behavior in a wireless mobile society as a whole and in locations inside the mobile society. The model gives the level of relationship between the classes of usage behavior learned globally in the mobile society and the locations inside the mobile society. It has one local model for all locations allowing different locations to be compared to each other based on the usage behavior.
- We realistically describe Internet usage in large mobile societies by an overlap of correlations by using a Bayesian mixture model called the "POWER" model. The model allows us to capture the mixture of hidden patterns in the user's Internet behavior. It can assign the users to multiple irrelevant classes of web users simultaneously based on the their usage behavior, a feat classical mixture models are incapable of. A fast version of the "POWER" model is used.

The rest of the paper is organized as follows: Section 2 reviews the related work in the area. Section 3 outlines our modeling approach of collecting and processing the wireless data and describes the Global Local model. Section 4 discusses experiments using synthetic and wireless data. Finally, Section 5 concludes.

2. RELATED WORK

Understanding the wireless mobile network users' behavior has been the scope of many papers. This is due to the wide spread of the wireless mobile networks. Among the topics covered is the evolution and usage of WLAN across time [25, 15, 8], user mobility [11, 5, 18], traffic flow statistics [27], user association patterns [20], application traffic characterization [21] and encounter patterns [12, 6]. Another work[4] analyzed a public WLAN user behavior based on user distribution, session duration, data rates, application popularity and mobility. From the analysis, wireless users were characterized in terms of a parametrized model for WLAN traffic studies and analysis. The user behavior only included the type of protocol used (TCP, UDP, etc.) and the type of application used (HTTP, SSH etc.), the type of web domain visited was not analyzed. Also, the public WLAN was in a 3-day ACM SIGCOMM'01 conference and not in a larger mobile society. Some previous works [11, 12] explore the space of understanding realistic users behavior empirically from data traces. The two main trace libraries for the networking communities can be found in the archives at [1] and [28].

Using the observed user behavior to design realistic and practical mobility models has been the focus of many work [9, 10, 16, 14]. It has been shown that most widely used existing mobility models fail to generate realistic mobility characteristics observed from the traces. Realistic mobility modeling is essential for protocol performance [3] . Correlating the user behavior with his location has rarely been covered in research. Ploumidis et al. [21] used a multi-level (network, AP and client) application-based traffic characterization, then grouped APs based on building category to examine variation in application use. A weak correlation has been found between the type of application used and some building categories. The traffic characterization was only based on AP traces and only on 7.5 days of traffic . Another application based study of a WLAN usage in a campus [8] evaluated the inbound and outbound traffic of web applications in a home and non-home location. Their work didn't find differences in user behavior based on the location. In a previous work [19] locations (i.e. buildings) in a mobile society were clustered together based on the similarity of Internet usage behavior and was found that locations of the same type actually cluster together, but no relation were drawn between the type of web domain clusters learned and and their appearance in similar type of locations. A smaller scale study of the influence of the region on the Internet usage behavior focused on the usage in a rural village[13], a difference was found between the dominant Internet traffic type (HTTP vs. Peer-to-peer) between urban areas and rural areas. Also, it has been found that in the village residence facebook web domains dominated the web traffic. This supports our idea that the location of the user influences his Internet behavior . Therefore, to the best of our knowledge this represents the first work in finding a correlation between the type of web domain cluster learned from Internet usage behavior from a mobile society as a whole and the type of location they most likely appear in.

A couple of studies were conducted on the Internet usage behavior on smartphones on wireless networks and 3G mobile networks [22, 26]. In one paper [22], influence of the location was found on the usage behavior without specifying a relationship between the type of location and the type of web domain visited. The other paper [26] found a correlation between the type of phone online application most used by the user at work and at home. Entertainment and social networks applications were found to be most used at home, while mail application most used at work. This kind of correlation were discovered also in our work.

The closest related work to the POWER model is the "Mixture Of Subsets" MOS model [23]. Like the POWER model, in the MOS model many components compete to influence each data attribute. However, the MOS model suffers from a complicated mechanism for allowing the many components to choose which attributes they will influence. Also, a main difference in the MOS model is the use of the EM algorithm to perform the learning instead of the bayesian approach used in the POWER model. This is the first work done in speeding up the POWER model and making it scalable as the number of attributes increase. Thus, allowing it to be practical for use on very large data such as the netflow dataset.

3. MODELING APPROACH

In this section we present The "Global Local" model which is used to learn the correlations between Internet usage patterns in a mobile society and the type of locations in the society. The wireless data's collection and processing step will be first presented. Then, the original POWER model and the fast POWER model are briefly described. We end this section with the formal description of the Global Local model.

3.1 The Wireless Data

Data-driven modeling of large mobile societies requires the collection and processing of multi-dimensional large datasets with fine granularity . In the first phase, extensive datasets are collected using the network infrastructure (or the mobile devices), plus augmenting information from 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 acquired information from different resources (e.g., access points, IP and MAC addresses), in which multiple datasets are manipulated, integrated and aggregated.

3.1.1 Data Collection

For the modeling of wireless users in the large campus, we collect different types of traces including netflows, DHCP and wireless AP session logs (MAC traps) through network switches. An IP flow is a sequence of packets with some common properties such as a source IP address and port number and a destination IP address and port number, that pass through a network router. Netflow records also include timestamps for the start and end time of a session (duration), protocol numbers and flow sizes. Websites accessed can be identified by destination IP addresses, while the application used can be identified by the port and protocol numbers. Wireless access points (APs) or switch ports which are the aggregate of APs in a building) collect the wireless session log. The trace also includes start and end events for device associations with APs or switches and its related details such as, the MAC addresses of the devices, the time and date of those events, and the IP and port numbers of the AP (or switch). From the above the location and time of user association for all MAC addresses can be derived. The DHCP log contains the dynamic IP assignments to MAC addresses. The IP in the DHCP log is given to the MAC address at the associated date and time.

3.1.2 Data Processing

Data processing consists of three steps, data manipulation, data integration and data aggregation to cross correlate the collected data.

Data Manipulation.

In the first step, due to the enormous size of the collected traces compression was used by substituting similar patterns with binary codes. Afterwards, mapping headers were created for use in future manipulation. The data was then exported into a MySQL database management system. Finally, customized store procedures were designed to manipulate the large data in a timely fashion.

The main challenges of this step is handling the variety and scale of the different collected traces. A naive approach for this step would have taken a time in the order of a month. Tens of manipulation operations would have been required. For example, The USC campus netflow dataset contains around 2 billions of flow records for each month in 2008, this equals to 2.5 terabytes of data per year.

Data Integration.

The second step of the processing is the integration of data. A semantic link must be created between the data from different sources due to the format difference between them. In our case, in wireless session logs users are represented by MAC addresses, while in netflow traces they are represented by IP addresses. However, the cost of this integration using SQL commands dramatically increases when the data scale for one of the traces (i.e. netflows) is very large. Thus, customized stored procedures were designed for this purpose.

Data Aggregation.

The output of the integration process cannot be directly fed to the analysis phase. This is due to the sheer size of the number of records in the output which are in the order of billions. Running any data mining method on this output will take years to finish. Therefore, an intermediate aggregation process is required for building design-specific views of the output. Records can be aggregated on a set of a fields or just one. The fields contain user, time, location, domain name and application. The appropriate aggregation scheme is chosen based on the modeling goal. An aggregation on domain

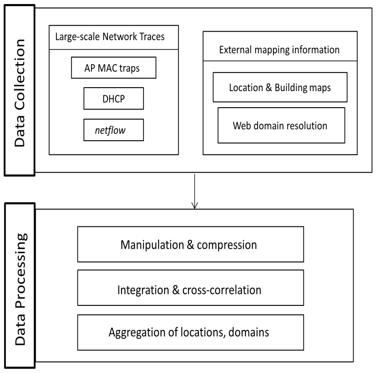
Data Collection
Large-scale Network Traces
AP MAC traps
DHCP
netflow
External mapping information
Location & Building maps
Web domain resolution
Data Processing
Manipulation & compression
Integration & cross-correlation
Aggregation of locations, domains

Figure 1: Data collection and processing phases.

name,location and time for the number of packets is chosen when we are interested in studying usage patterns for different domains at different locations. In our case, an aggregation on user id, domain name and months is chosen to study the users' spent time at different web domains for different months.

3.2 The POWER Model

The POWER model[24] is a new class of mixture models where unlike classic mixture models, a single data point is generated by the contribution of multiple classes, with each class having a varying influence on the generation. This model allows for discovering complex overlap of patterns in the data.

The Gibbs sampling algorithm, which is an iterative Monte Carlo Markov Chain method, is used in the POWER model to learn the parameters of the model. Expanded details of the POWER model is found in a paper by Somaiya et al.[24].

3.3 The Fast POWER Model

The POWER model suffers from a very long learning time. This is attributed to the weight parameter learning step in the Gibbs sampling algorithm. The learning time of the weight parameter is $O\left(n \cdot k \cdot d^2\right)$ where n is the number of data points, k is the number of components and d the number of attributes.

A novel approximation method is used in the computation of the weight parameter learning step to reduce the learning time. The new learning time became $O\left(n \cdot k \cdot d \cdot p\right)$ where p is a new parameter used in the approximation and is much smaller than d. Hence, the running time of the POWER model also drastically improved. Details of the fast POWER model can be found in the technical report[2].

3.4 The Global Local Model

The Global Local model was devised in order to model the influence of a location in a mobile society on a user's web activity and behavior. It is important to understand the web usage in a location and design efficient context-aware Internet protocols and services suitable for all locations. A previous work [19] showed how similar types of locations are clustered together based on the users' web behavior, our model shows how the type of a web domain cluster learned from users' behavior in the mobile society as a whole correlates with the type of a certain location in the society. The model also generates the likelihood of the web domain clusters appearing in the locations. This gives the level of relationship between each location and all web domain clusters.

The model first learns a global template of clusters from the global wireless data. The global template learned will be imposed on the model learning to each location in the mobile society. This relates the locations in a mobile society together and gives the ability to compare locations.

Table 1: The 4 patterns used to generate the synthetic data. the underbrace under a number is the string length of the number.

Id	Pattern
1	$\underbrace{1...1}_{10}\underbrace{0...0}_{30}$
2	$\underbrace{0...0}_{20}\underbrace{1...1}_{10}\underbrace{0...0}_{10}$
3	$\underbrace{0...0}_{10}\underbrace{1...1}_{10}\underbrace{0...0}_{20}$
4	$\underbrace{0...0}_{30}\underbrace{1...1}_{10}$

The Global Local model Consists of two phases. The first phase (Global phase) is basically the fast POWER model[2] run on global wireless data (wireless data from all locations of the mobile society). The global phase learning will reach a steady state where the model parameters will be learned. These parameters include the weights w which specifies the strength of influence of each learned pattern on various data attributes. Another parameter learned is the Θ, which parameterizes the probability density function of the random variable used to generate a data point in the generative process of the POWER model. The second phase (Local phase) continues the fast POWER model learning process of the global phase but while fixing the learning of the parameters of w and Θ. The local phase is run on each location in the mobile society separately with the w and Θ fixed with the values learned from the global phase. The input data to the local phase in a location is the subset of the global data corresponding to that location. The appearance probability of each pattern, denoted with α, is learned from this phase for each location in the mobile society. This gives the strength of the relationship between the type of usage behavior class and the type of the location and the correlation between them.

4. RESULTS

This section shows and discusses the results of the Global Local model on both synthetic and wireless data. The wireless data was collected from a campus-wide analysis from the University of Southern California (USC) in 2008. The Global Local model was written in C++ and run on a 1.73 GHz Intel Core i7 laptop with 6 GB of RAM.

4.1 Synthetic Data Results

The Global Local model was tested on synthetic data generated from a mixture of 4 simple patterns with varying appearance probabilities with 40 attributes across 10 locations. There were 1000 data points for each location making the total number of data points to 10,000. Table 1 shows the patterns used to generate the data.

The Global Local model correctly learned the patterns and both of their global and local appearance probabilities (α). Table 2 shows both the generating α and the learned α for the global part and Table 3 shows them for the local.

4.2 Wireless Data Results

In this subsection we test the Global Local model on wireless data gathered from the USC campus in 2008. The data consists of the top 30 web domains visited on campus by the users. The number of user records covered in the data is 6284 user records. Each user's visit to those web domains were converted to a row of

Table 2: The global generated appearance probability (α) and the learned appearance probability (α).

Global Appearance Probability (α)				
Generating α	0.30	0.39	0.21	0.28
Learned α	0.29	0.40	0.22	0.28

Table 3: The local generated appearance probability (α) and the learned appearance probability (α).

Location 1 Appearance Probability (α)				
Generating α	0.50	0.50	0.50	0.50
Learned α	0.48	0.48	0.53	0.50
Location 2 Appearance Probability (α)				
Generating α	0.75	0.75	0.25	0.25
Learned α	0.76	0.74	0.28	0.22
Location 3 Appearance Probability (α)				
Generating α	0.40	0.00	0.40	0.00
Learned α	0.41	0.00	0.44	0.00
Location 4 Appearance Probability (α)				
Generating α	0.00	0.40	0.00	0.40
Learned α	0.00	0.40	0.00	0.38
Location 5 Appearance Probability (α)				
Generating α	0.75	0.25	0.00	0.00
Learned α	0.72	0.25	0.00	0.00
Location 6 Appearance Probability (α)				
Generating α	0.00	0.50	0.75	0.00
Learned α	0.00	0.51	0.76	0.00
Location 7 Appearance Probability (α)				
Generating α	0.00	0.30	0.00	0.30
Learned α	0.00	0.32	0.00	0.31
Location 8 Appearance Probability (α)				
Generating α	0.00	0.50	0.00	0.75
Learned α	0.00	0.52	0.00	0.72
Location 9 Appearance Probability (α)				
Generating α	0.35	0.50	0.00	0.35
Learned α	0.37	0.49	0.00	0.36
Location 10 Appearance Probability (α)				
Generating α	0.25	0.25	0.25	0.25
Learned α	0.23	0.30	0.24	0.26

Table 4: The wireless data locations used in the USC campus.

Location Id	Location Name	Location code
1	Alpha Chi Omega Sorority	Sor1
2	Kappa Alpha Theta Sorority	Sor2
3	Alpha Tau Omega Fraternity	Frat1
4	Beta Omega Phi Fraternity	Frat2
5	Sigma Phi Epsilon Fraternity	Frat3
6	Zeta Beta Tau Fraternity	Frat4
7	Alpha Kappa Psi Fraternity / Business	Frat5/Business
8	Fluor Tower Housing	hous1
9	Annenberg House Apartment	hous2
10	Web Tower Housing	hous3
11	Annenberg School for Communication & Journalism	jour
12	George Lucas Building School of Cinematic Arts	lucas
13	Wilson Dental Library	dent
14	Norris Medical Library	med
15	University Computing Center	UCC

Table 5: The highest ranked web domains for some of the components learned from the USC campus wireless data set global phase.

Id	Web domains
1	theplanet, youtube, wikimedia, panthercdn, tfbnw
5	cnet, washingtonpost, apple, mac
6	live
8	facebook, tfbnw, mediaplex
9	co, mozilla

zeros and ones, zeros being the user not visiting the domain and one being visiting the domain. Thus, we obtain a 0/1 matrix of size 6284 by 30. This matrix can be divided into 15 disjoint parts. Each part represents the location that portion of the data was gathered from. Table 4 shows the locations used in gathering the data with locations of the same type colored with the same color. The number of components k was set to 11.

We also study the similarity in Internet usage between locations. The results from the local phase of the model, which are the appearance probability of the web domain components for all locations, are analyzed using a dissimilarity matrix. Then, a graph representation of the dissimilarity matrix is created and cliques are discovered between the locations. We want to observe if locations of a similar type have a similar Internet usage based on the discovered cliques.

4.2.1 Global Phase

The fast POWER Global phase is run on the global wireless data to learn the model. KL divergence is used on the results to rank the attributes according to importance for all components. Table 5 shows the highly ranked web domains for some of the components learned from the data set.

Discussion The global phase learning of the model captures a clear and intuitive cluster of web domains, based on user behavior, associated with the learned components. It was able to capture the known components usually clustered from this data set. It manages to cluster media related domains in component 1. Component 5 represents the "‘mac component"’ which is a group of domains that always cluster together and represent the behavior of mac users. Microsoft web domains are represented in component 6 which in our data is only represented by the live domain. Facebook and its supported domains is represented in component 8. Component 9 is the mozilla firefox users cluster, there is always a correlation between the mozilla domain and the co domain.

4.2.2 Local Phase

After running the local phase on each location we observe the appearance probability (α) of the components learned from the global

Table 6: The appearance probability (α) of locations in descending order, sorted per component.

Component	Location and corresponding α
C1 (media component)	Sor1 0.45, Frat3 0.45, hous1 0.45, Sor2 0.44, jour 0.43, hous2 0.40, Frat1 0.38, Frat2 0.37, Frat5/business 0.30, UCC 0.28, hous3 0.25, lucas 0.23, med 0.16, dent 0.12, Frat4 0.10
C5 (mac component)	jour 0.36, Lucas 0.34, Sor2 0.31, hous1 0.22, hous2 0.20, Frat3 0.14, Dent 0.11, Sor1 0.11, Frat5/business 0.09, Frat1 0.08, Frat2 0.08, hous3 0.07, med 0.07, UCC 0.05, Frat4 0.01
C8 (facebook component)	Sor2 0.43, hous1 0.29, jour 0.28, Frat3 0.25, hous2 0.25, UCC 0.18, Sor1 0.17, Frat4 0.15, Frat5/business 0.15, med 0.14, Frat1 0.13, hous3 0.12, dent 0.10, lucas 0.10, Frat2 0.06

phase in each location. The appearance probabilities of the component in the locations are displayed per component and are sorted in descending order by the probability's value in each location. The results are shown in Table 6.

Discussion The local phase of the model captures an intuitive correlation between the type of components to the types of locations in the wireless mobile society. As can be seen from Table 6, component 1 named the media component has a high probability of appearing in housing, fraternities and sororities locations. this is related to the youtube domain being in the component, the youtube domain is usually visited for entertainment purposes, which is an appropriate use for users being in their residence. Also, the media component has a high appearance probability in the school of communication and journalism. This is due to the youtube domain being also a video sharing domain, which is an important tool used by journalism students to research news items videos on the domain and to also share their class related videos. The media component is least likely to appear in library locations, this is related to the unsuitability of viewing videos in locations where a quiet environment is expected.

Component 5 (the mac component) is likely to appear in the school of communication and journalism and the school of cinematic arts. This is related to prominent use of mac machines in these schools for their great capabilities of image and video editing. Another reason for the high appearance of this component in the school of communication and journalism is having the washington post newspaper domain in the component. There is an obvious correlation between the newspaper domain and the school.

Component 8 (the facebook component), has a high appearance probability in Sororities and housing locations. There is an intuitive correlation between the social network website use and users being in their residence, users tend to use social network websites in the comfort of their homes. Another location where the domain is most likely to appear in is the annenberg school of communication and journalism, this is likely due to another use of the social network website as a news outlet for politicians and media figures. Other education locations get a lower appearance probability of this domain for the lack of use of the social network in their fields.

4.2.3 *Internet Usage Similarity Between Locations*

In this section we wish to study the similarity of Internet usage among locations in the mobile society. We achieve this by creating a dissimilarity matrix between the 15 locations available in our wireless data, the dissimilarity between locations will be computed by the cosine distance function between the locations' appearance probabilities (α) of the web domain components. Then, the dissimilarity matrix is mapped to an undirected graph as follows. Nodes in the graph will represent locations in the mobile society, an edge is drawn between two different nodes if their dissimilarity is less than a threshold. Finally, we find cliques within the graph to discover groups of locations with similar Internet usage.

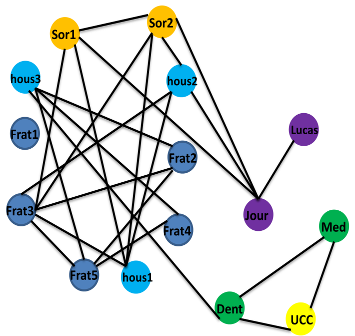

Figure 2: Graph representation of the dissimilarity matrix using the threshold of 0.1 for the locations in the mobile society.

Figure 2 shows the resulting graph with a threshold of 0.1. Locations of similar type have the same node color, location codes from Table 4 were used to denote the nodes.

Discussion The resulting graph includes around 11% of all possible edges using the mentioned threshold. As can be seen from the graph, a clear relationship exists between identified cliques and the types of locations based on the Internet usage. This shows that locations of the similar type actually have the same pattern of Internet users' behavior.

5. CONCLUSION

In this paper we have introduced a new model for users' web behavior in a wireless mobile society as a whole and in certain types of locations in the society. The model learns one global template for the wireless Internet usage behavior in the mobile society. This global template is used to learn the correlations in the different locations of the society. Having one global template allows for the different locations to have a common ground for comparison, instead of having separate models for each location. The model learned the usage behavior classes of the mobile society based on the users' web behavior. It has also produced the appearance probability of each usage behavior class (learned globally in the mobile society) in all locations of the mobile society. This showed the level of the relationship between the usage behavior classes and locations. We have shown that there is a correlation between the type of usage behavior class and the type of locations in the society where this class is more likely to appear in. This model helps in building better context-aware network protocols and services by using data-driven modeling and design paradigm.

6. REFERENCES

[1] Mobilib: Community-wide library of mobility and wireless networks measurements (investigating user behavior in wireless environments). `http://nile.cise.ufl.edu/MobiLib/`.

[2] A. Almutairi, S. Ranka, and M. Somaiya. A fast algorithm for learning weighted ensemble of roles. Technical report, University of Florida, 2012.

[3] F. Bai, N. Sadagopan, and A. Helmy. The important framework for analyzing the impact of mobility on performance of routing protocols for adhoc networks, 2003.

[4] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan. Characterizing user behavior and network performance in a public wireless lan. In *in: Proceedings of ACM SIGMETRICS, Marina Del Rey, 2002*, pages 195–205.

[5] M. Balazinska. Characterizing mobility and network usage in a corporate wireless local-area network. pages 303–316, 2003.

[6] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. Mob. Comp*, 6:606–620, 2007.

[7] B. S. Everitt and D. J. Hand. *Mixture Models: Inference and Applications to Clustering*. Chapman and Hall, London, 1981.

[8] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. In *In Proceedings of ACM MOBICOM*, pages 187–201. ACM Press, 2004.

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

[10] R. Jain, D. Lelescu, and M. Balakrishnan. Model t: a model for user registration patterns based on campus wlan data. *Wirel. Netw.*, 13(6):711–735, Dec. 2007.

[11] W. jen Hsu and A. Helmy. On modeling user associations in wireless lan traces on university campuses. In *In Proceedings of the Second Workshop on Wireless Network Measurements (WiNMee*, 2006.

[12] W. jen Hsu and A. Helmy. On nodal encounter patterns in wireless lan traces. In *IEEE Int.l Workshop on Wireless Network Measurement (WiNMee*, 2006.

[13] D. L. Johnson, E. M. Belding, K. Almeroth, and G. van Stam. Internet usage and performance analysis of a rural wireless network in macha, zambia. In *Proceedings of the 4th ACM Workshop on Networked Systems for Developing Regions*, NSDR '10, pages 7:1–7:6, New York, NY, USA, 2010. ACM.

[14] M. Kim and D. Kotz. Extracting a mobility model from real user traces. In *In Proceedings of IEEE INFOCOM*, 2006.

[15] D. Kotz and K. Essien. Analysis of a campus-wide wireless network. In *In Proceedings of ACM Mobicom*, pages 107–118. ACM Press, 2002.

[16] D. Lelescu, U. C. Kozat, R. Jain, and M. Balakrishnan. Model t++: an empirical joint space-time registration model. In *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '06, pages 61–72, New York, NY, USA, 2006. ACM.

[17] G. Mclachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, Oct. 2000.

[18] M. McNett and G. M. Voelker. Access and mobility of wireless pda users.

[19] S. Moghaddam, A. Helmy, S. Ranka, and M. Somaiya. Data-driven co-clustering model of internet usage in large mobile societies. In *Proceedings of the 13th ACM international conference on Modeling, analysis, and simulation of wireless and mobile systems*, MSWIM '10, pages 248–256, New York, NY, USA, 2010. ACM.

[20] M. Papadopouli, H. Shen, and M. Spanakis. Characterizing the mobility and association patterns of wireless users in a campus, 2004.

[21] M. Ploumidis, M. Papadopouli, and T. Karagiannis. Application-based characterization of the traffic of a campus-wide wireless network.

[22] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum. Livelab: measuring wireless networks and smartphone users in the field. *SIGMETRICS Perform. Eval. Rev.*, 38(3):15–20, Jan. 2011.

[23] M. Somaiya, C. Jermaine, and S. Ranka. Learning correlations using the mixture-of-subsets model. *ACM Trans. Knowl. Discov. Data*, 1(4):3:1–3:42, Feb. 2008.

[24] M. Somaiya, C. M. Jermaine, and S. Ranka. Mixture models for learning low-dimensional roles in high-dimensional data. In *KDD*, pages 909–918, 2010.

[25] D. Tang and M. Baker. Analysis of a local-area wireless network. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, MobiCom '00, pages 1–10, New York, NY, USA, 2000. ACM.

[26] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3g network. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 267–279, New York, NY, USA, 2009. ACM.

[27] S. Wong, Y. Yuan, and S. Lu. Characterizing flows in large wireless data networks. In *In Proceedings of ACM MOBICOM*, pages 174–186, 2004.

[28] J. Yeo, D. Kotz, and T. Henderson. Crawdad: a community resource for archiving wireless data at dartmouth. *SIGCOMM Comput. Commun. Rev.*, 36(2):21–22, Apr. 2006.