LARGE-SCALE MINING OF MOBILE ONLINE BEHAVIOR: INTEREST-AWARE MODELING AND DESIGN

By

SAEED ABBASI MOGHADDAM

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012

To my Mom and Dad

ACKNOWLEDGMENTS

I would like to thank first and foremost my adviser, Dr. Ahmed Helmy, and my co-advisor, Dr. Sanjay Ranka, for their guidance and enlightenment through the course of my Ph.D. program. Dr. Helmy and Dr. Ranka are not only great academic advisers, but also great mentors and role models, who helped me significantly in pursuing my dream.

I would like to also express my deep gratitude to my supervisory committee member, Dr. Alin Dobra, who helped me significantly in the process of forming my dissertation. Also, I would like to extremely thank my supervisory committee members, Dr. Tamer Kahveci, and Dr. Michael Fang for their inputs in completion of this dissertation. I would like to also extend my thanks to many professors and colleagues I had the privilege to work with for improving my research work. I want to thank all the fellow A-groupers, Udayan Kumar, Gautam Takure, Sungwook Moon, Jeeyoung Kim, Yibin Wang and Guliz Tuncay for many helpful conversations and stimulating and pleasant environment we collectively created.

Furthermore, I would like to thank all my friends who helped and supported me during past few years including Masoumeh Rajabi, Farzad Fani Pakdel, Ashkan Behnam, Azam Feiz, Shahin Navardi, Niusha Nazar Kazemi, Mehdi Eshaghian, Hadi Hajimiri, Danial Sabri Dashti, Sahar Mirshamsi, Nima Afshar and Heshmat Saroui.

Finally, I would like to express my deepest gratitude to my family, especially my mother Effat Vahdati, my father Mohammad Ali Abbasi Moghaddam, my brother Masoud, and my sister Samaneh who have always given me tremendous support and encouragement. The completion of this dissertation was made possible by the many years of preparation for the great challenges of the world, and their endless support. Without them, I would not be where I am or have what I have achieved today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CDF	Cumulative Distribution Function
DHCP	Dynamic Host Configuration Protocol
EM	Expectation Maximization
GMM	Gaussian Mixture Model
ICAST	Interest-aware Casting
IP	Internet Protocol
KS	Kolmogorov-Smirnov
MAC	Media Access Control
NETFLOW	Network Flow
SOM	Self-organizing Map
WLAN	Wireless Local Area Network

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

LARGE-SCALE MINING OF MOBILE ONLINE BEHAVIOR: INTEREST-AWARE MODELING AND DESIGN

By

Saeed Abbasi Moghaddam

December 2012

Chair: Ahmed Helmy
Co-chair: Sanjay Ranka
Major: Computer Engineering

User online behavior and interests will play a central role in future mobile networks. People today are ever-increasingly utilizing online services on the move using their mobile devices for different purposes. This fast growing trend toward mobile Internet access creates a tight coupling between users and mobile networks, where various characteristics of user online activities, mobility, and trends can be captured and applied to provide new behavior-aware services and applications. In this dissertation, we introduce a novel systematic framework for large-scale mining of mobile users behavior and its application for interest-aware modeling and design in mobile networks. In this framework, we first provide a systematic method to process, integrate and aggregate billions of multi-dimensional mobile network traces including netflow logs, which is by far the largest dataset processed in any study of mobile networks to date. Then, we introduce three data-driven approaches for multi-dimensional mining and analysis of online activity, mobility, and trends in large mobile societies based on information-theoretic co-clustering, self-organizing maps and Gaussian mixture models. We show how these techniques can be effectively applied to discover, organize,

visualize, and preserve characteristics of mobile users' online behavior. Finally, we investigate three different problems including simulation of mobile users behavior, traffic analysis, and modeling and interest-aware casting and show how the proposed behavioral mining techniques can be effectively applied to provide interest-aware models and services for the mobile networks.

CHAPTER 1
INTRODUCTION

Emergence of the Mobile Internet

Wireless mobile networks are evolving and integrating with every aspect of our lives. Today, laptops, handhelds and smart phones are becoming ubiquitous providing almost continuous Internet access that had just been a distant dream till a few years ago. This significant shift toward mobile Internet access has been accelerating with the rise since 2007 of larger multi-touch smartphones, and of multi-touch tablet computers since 2010 by providing better and easier Internet access experiences than previous generations of mobile devices. In fact, the usage of mobile wireless Internet is progressing so fast that it is going to revolutionize 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 changed too. 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, mobility and trends can be captured and applied to provide new solutions to the networking problems. To achieve this goal we introduce a novel systematic framework which divides the process of problem solving for mobile environments into two major steps; a) behavioral mining and modeling and b) behavior-aware (interest-aware) problem solving. In the next section, we explain this new approach and each of its two main steps in detail.

Behavioral Modeling and Behavior-Aware Problem Solving Paradigms

In order to capture the actual characteristics of mobile users behavior and apply them to provide realistic solutions to the networking problems, e.g., simulation of users behaviors, traffic analysis or design of mobile services, we need to go through at least two major steps: a) Monitoring users behavior including all their interactions and moves b) Providing solutions based on the behavioral data. However, a major challenge here is the fact that considering all the interactions and moves for a large community, e.g., all students on a campus, needs dealing with tera-scale amount of data which is difficult, computationally costly and time-consuming. On the other hand, providing integrated realistic solutions requires us to consider different dimensions (e.g., activities on different websites) and various aspects (e.g., mobility, website visitation) of users behavioral patterns altogether, which is a challenging issue.

To overcome the above major challenges, we propose to use '*behavioral models*' as a bridge between the '*behavioral data*' and the '*behavior-aware solution*' (Figure 1-1). Generally speaking behavioral models provide compact abstract views of the whole data while preserving and revealing important characteristics of the dataset, e.g., existing correlations or behavioral groups. By developing such data-driven models from the real data we can reuse them to solve many different types of networking problems. For example, we can use them for simulation of the real dataset, behavior-aware traffic analysis or interest-based information delivery in mobile networks. To build such models, it is imperative to establish systematic and scalable methods for mining, analysis and modeling of massive multi-aspect/ multi-dimensional datasets of mobile users' online activities, mobility and traffic.

The data-driven paradigm is in sharp contrast to the general-purpose problem-solving paradigm conventionally used in the networking area in the past decades. The general-purpose paradigm focuses on the design of solution elements first, then the evaluation using generic statistical data (usually random, uniform). Often, these models deviate dramatically from reality, which leads to sub-optimal performance or outright failure of the designed solution during deployment. By contrast, the data-driven behavioral modeling approach starts by the analysis and realistic modeling of the target context and users, that then drives the design process of the solutions.

Research Framework and Organization

Figure 1-2 shows the research framework in this dissertation that includes three major phases. The first step is flow-level processing of very large datasets of mobile users activities on the Internet and also their movements during several months. This phase by itself includes three major steps of data collection from different types of sources, integration of all different types of data into a unified format, and finally an appropriate aggregation of data for the modeling problem. In this study, we collected information from the network infrastructure (including *netflow* logs, WLAN traces, and DHCP logs) and external sources (including whois lookup services, access points map, and buildings online catalogs). We developed a set of tools using Java programming language and MySQL database system for data processing, integration and aggregation of all the data (including billions of records) as per required by the modeling and design phases. In this dissertation, we mainly consider the interest level of mobile users in different websites or locations based on different aggregation criteria (e.g., the total online time or flow-based traffic).

The second phase is behavioral mining and modeling based on users interests. In this dissertation, we provide three types of techniques for this purpose including information theoretic co-clustering [1], self-organizing maps [2] and Gaussian mixture models [3]. Each of these models reveals and preserves some characteristics of the input data. Co-clustering model represents group characteristics of the data and provides a high level understanding of the interest groups within the society of mobile users. We apply this model considering users, activities and locations and analyze the similarity of locations based on their web access patterns as well. Self-organizing map model, which is based on neural networks, captures the existing trends inside the society. It also shows and keeps correlations, anti-correlations and anomalies considering individual features of users activities like accessing web domains or visitation of buildings. The third type of model, i.e., Gaussian mixture model, captures the main components of activity patterns that can be effectively applied for simulation of users visitation patterns.

As can be seen in Figure 1-2, the third part of this dissertation includes interest-aware modeling and design. In this phase, we try to solve a set of networking problems in an interest-aware approach in order to show how the acquired models can be applied in practice. As mentioned before, multivariate simulation of users' visitation patterns is one of these problems for which we provide a model-based solution utilizing the Gaussian mixture model and also an evaluation approach based on the Pearson's chi-square test [4]. Traffic analysis and modeling is another problem for which we provide domain specific, location based and behavior-aware solutions using KS test [5], a graph-based approach and the co-clustering model. Such traffic models and

simulation techniques can be effectively applied in the design of interest-aware solutions for the networking problems including caching of web objects, resource management, and also interest-aware casting, which we introduce in the last part of this dissertation. The interest-aware casting paradigm (iCast) is based on the concept of profile cast [6,7] that utilizes user profiles. In this casting approach, despite current blind way of casting, messages are delivered based on a pre-known interest model of the society. In other words, messages are delivered based on the existing interest groups in the society of mobile users. We achieve this goal by providing a set of recommendations to the users that are built based on the co-clustering model.

The framework in this dissertation provides a clear roadmap for data-driven modeling and design in mobile networks. This framework very well shows how to prepare the data, how to provide data-driven behavioral models and how to apply such models in practice to provide interest-aware solutions for networking problems. This dissertation is organized as follows. In Chapter 2, we review the related work. In Chapter 3, we describe the datasets and data preparation phase. Chapter 4 and 5 present the co-clustering model and the self-organizing map model respectively. Chapter 6 discusses the Gaussian mixture model and its application in simulation of mobile users activities. Chapter 7 presents domain, location and human-centered traffic modeling of mobile wireless network and Chapter 8 introduces the interest-aware casting (iCast) paradigm. Chapter 9 concludes the dissertation.

Contributions

This dissertation has the following major contributions:

1. We collect and process the largest set of mobile network usage traces (including billions of records) and provide practical techniques for integrating and aggregating the data.

2. We propose the co-clustering model as an effective approach for multi-dimensional analysis and modeling of very large datasets, and show how information theoretic co-clustering can be applied to create and correlate clusters of users, web domains and locations to capture group-specific characteristics.

3. We present the self-organizing map model that can be applied for modeling of minor and major trends in mobile society in terms of their online activity and mobility. We show how this method can be effectively applied to capture different types of correlations among different domains and locations. We also suggest the extended version of the model for multi-aspect modeling of users interests.

4. We suggest the Gaussian mixture model for capturing the main behavioral components in mobile users society and dynamic multivariate simulation of web access patterns and show how the accuracy of multivariate simulation can be measured using our modified version of Pearson's chi-square test.

5. We introduce domain, location and human-centered traffic analysis and modeling approaches and show how they can improve the accuracy of traffic models. We also show how our suggested graph-based approach can reveal similarities and differences between the traffic distributions of different domains or locations and how the co-clustering model can be effectively applied for behavior-aware traffic modeling of mobile networks.

6. We introduce a novel behavior-aware casting paradigm that can be effectively applied for interest-aware delivery of messages (*iCast*) in mobile networks. We show how this approach can be utilized in a semi-centralized architecture and how it can be applied to improve the performance of message delivery in a mobile society.

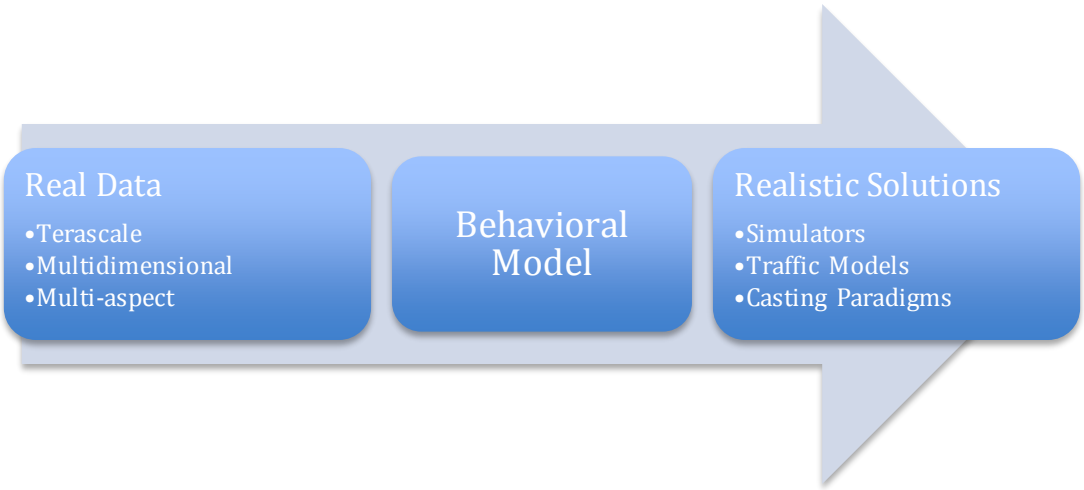


Figure 1-1. Behavioral model is the bridge between the real data and realistic solutions.

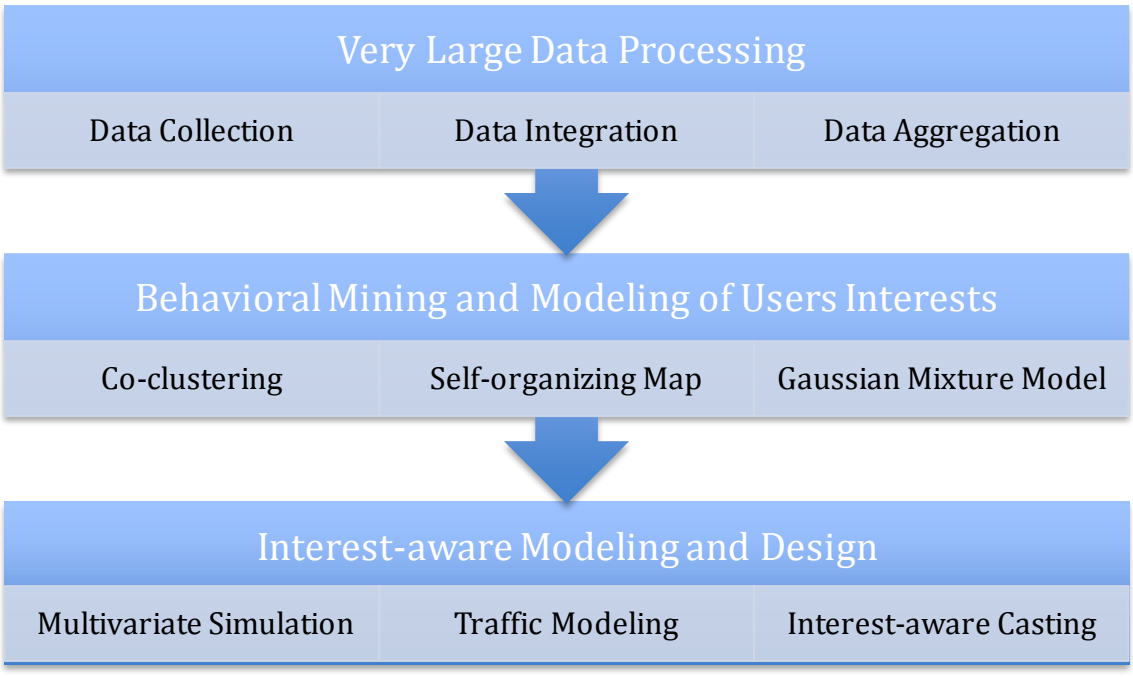


Figure 1-2. Research framework

CHAPTER 2
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. The scope of analysis includes WLAN usage and its evolution across time [8-10], user mobility [11-13], traffic flow statistics [14], user association patterns [15] and encounter patterns [16, 17]. Some previous works [11, 16] 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 [18] and [19]. 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 and locations 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 of wireless and mobile societies. Our work also includes novel data processing techniques to address the challenges provided by this large-scale multi-dimensional data.

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 [20-23]. They have shown that most widely used existing mobility models (mostly random mobility models, e.g., random walk, random waypoint; see [24] for a survey) fail to generate realistic mobility characteristics observed from the traces. Realistic mobility modeling is essential for protocol performance [25]. It has been shown that user mobility preference matrix representation leads to meaningful user clustering [26]. Several other works with focus on classifying users based on their mobility

periodicity [27], time-location information [28, 29], or a combination of mobility statistics [30]. The work on the TVC model [20] provides a data-driven mobility model for protocol and service performance analysis. Our work is complementary to TVC and can extend TVC dramatically to incorporate dimensions of load, interest and website visitation preferences. In addition, the netflow traces are over three orders of magnitude more than the WLAN traces and the techniques used for their analysis are quite different. In [14] it was shown that the performance of resource scheduling [31] and TCP vary widely between trace-driven analysis and non-trace-driven model analysis. Using multi-dimensional modeling, we can develop new behavioral models for online activities, mobility, trends and traffic and utilize the realistic profiles to enhance the performance of networking solutions.

On the scope of Internet traffic modeling, there has been many works as well and among all flow-level modeling has been one of the most popular approaches [32, 33, 34]. However, most of such studies use idealized models, e.g., Poisson process, to characterize flows. While such simplified models may be fine in the wired Internet, they are not appropriate for wireless networks. Among the works looking into heavy-tailed distributions, [35] propose to use several heavy-tailed distribution models to characterize the statistical process associated with TCP flows in a wide-area network. Ref [36] also suggests Weibull distribution as a better fit than other distributions for modeling of wired TCP flow arrivals. However, these works are mainly based on wired network and few studies have focused on wireless traffic modeling. While other works have characterized user and mobility patterns in wireless networks, most of them focused on host-level rather than flow-level. Researcher at [30, 8] studied users,

network activity and host mobility patterns in a metropolitan-area wireless network and also on a campus department. Other studies at [9, 10] investigated wireless user and access point (AP)/building activity and aggregate traffic for the Dartmouth campus wireless network. Ref [12] studied user population characteristics, network usage and load distribution in corporate networks and [37] characterized the aggregate network load and utilization and user patterns during a conference. For flow-level modeling of wireless networks, [14] propose a Weibull regression model to approximate the flow arrivals at individual APs. In another work [38], researchers found that accessed information by HTTP queries shows spatial locality in a wireless campus network. The key difference between the previous studies and ours is that we provide a very flexible flow-level traffic modeling approach based on domains, locations and behavioral subspaces instead of just individual entities or the whole network.

Our novel behavior-aware approach also provides user profiles that may be used in a myriad of networking applications. One network application for multi-dimensional modeling is profile-based services. Profile-cast [6, 7] 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 [6, 7] use location visitation preference and are not aware of Internet activity. Other previous works also rely on movement patterns. Our multi-dimensional modeling of mobile users, however, provides an enriched set of user attributes that relate to social behavior (e.g., interest, community as identified by web access, etc.) that has been largely ignored before. In our work, we apply such interest-based profiles to provide new type of behavior-aware applications and services.

CHAPTER 3
DATA PREPARATION

Data-driven modeling of large mobile societies requires three main phases to collect, process and analyze multi-dimensional large datasets with fine granularity (Figure 3-1). 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. The final phase is data analysis and modeling which includes the study of human behaviors based on their website and location visitations patterns.

Data Collection

In this dissertation, we collect different types of traces via network switches including netflows, DHCP and wireless access point (AP) session logs (MAC traps) from the University of Southern California (USC) campus. An IP flow is defined as a unidirectional sequence of packets with some common properties (e.g., IP address and port number of source and destination) that pass through a network device (e.g., router). This device can be used for flow collection. The collected data provides fine-grained metering for detailed usage analysis. 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) (Table 3-1). The destination IP address can be used to identify the websites accessed, while the port and protocol numbers can identify the application used. The wireless

session log is collected by each wireless access point (AP) or switch port (i.e., aggregate of APs in a building). The trace includes the 'start' and 'end' events for device associations (when they visited or left that specific AP), the device's MAC address, the date and time of those events, and the AP (or switch) IP and port numbers. From the above we can derive the association history (i.e., the location and time of user association) for all MAC addresses. The DHCP log contains the dynamic IP assignments to MAC addresses. The listed IP is given to the MAC address at the indicated date and time.

Large Data Processing

Data processing includes three steps of data manipulation, data integration and data aggregation. These steps are needed to cross correlate the collected data before data analysis. In the following, we briefly address the challenges involved in each step and explain our approach to resolve the problems.

Data Manipulation

The variety and scale of different collected traces introduce one of the main challenges with respect to data manipulation. The size of the underlying data is very large and therefore, with a naïve approach the required time for each manipulating action will be in the order of a month, with tens of manipulations needed. For example, the netflow dataset gathered from the USC campus includes around 2 billion of flow records for each month in 2008 which equals to 2.5 terabytes of data per year. Thus, appropriate methods for data manipulation are needed. Our approach to diminish the problem is to first compress the data via substituting similar patterns with binary codes and creating mapping headers to be used in future manipulation (using a set of Java

programs); then export the data into a database system (MySQL) and finally design customized store procedures for the manipulation of data in a reasonable time.

Data Integration

The second requirement of multi-dimensional modeling is data integration. Data from different sources are not gathered in the same format and therefore a semantic link is required to be created between them. For example, in our case study, users are represented by MAC addresses in wireless session logs and by IP addresses in netflow traces. However, when the data scale for one of the traces (in this case netflows) is very large, the cost of such integration using regular SQL commands increases dramatically. Thus, we designed customized stored procedures for this purpose.

Data Aggregation

Since the output of the integration process includes billions of records, we cannot directly feed the result to the analysis and modeling phase. Running rather any data mining method on such a large dataset will take years to accomplish. Therefore, we need an intermediate aggregation process for building design-specific views of the dataset. We can aggregate the records based on one or a set of fields e.g., time, user, location, domain name and application. The choice of appropriate aggregation scheme depends on the final design and modeling goals. If we are interested in studying usage patterns for different domains at different locations without considering single users or type of application, an aggregation on domain name, location and time for the number of bytes, packets or flows will be the best choice. If the goal is the study of users' spent time at different websites for different months, we need to aggregate based on user id, domain name and months.

Data Preparation Details

In this study, we processed netflow and DHCP traces from the USC campus for more than 22 thousand mobile users in a Wireless LAN spanning over 79 buildings and including over 700 APs 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, and then the February and April traces (including 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 and all the users interacting with those domains (e.g., Google, Facebook, etc.) were considered for the analysis and modeling.

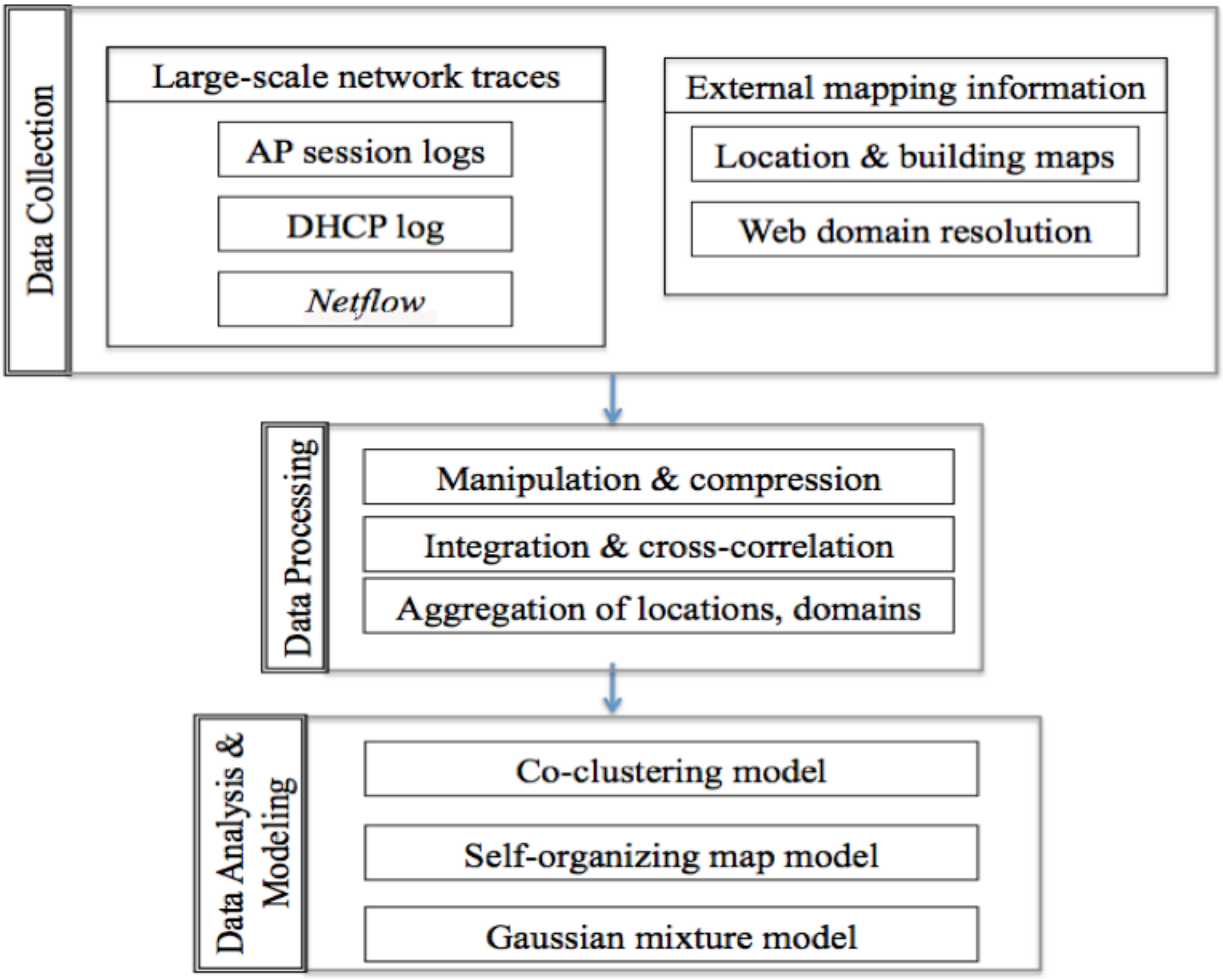


Figure 3-1. Phases of modeling: collection, processing, and analysis.

Table 3-1. Netflow sample

Start timestamp	Finish timestamp	Source IP	Source port	Destination IP	Dest port	Protocol number	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

. CHAPTER 4

DATA-DRIVEN CO-CLUSTERING MODEL OF MOBILE INTERNET USAGE

Behavioral mining and modeling is the basic building block for realistic design of networking solutions. Without behavioral models, we will not be able to use the characteristics of the real datasets to provide data-driven solutions. For this purpose, in this chapter, we present a novel approach based on co-clustering technique for modeling of Internet usage in large mobile networks [39]. The main goal of this study is to find behavioral groups inside the mobile society and analyze the similarity of different locations based on their web access patterns. We propose information-theoretic co-clustering technique [1] in a novel way to model and extract important relations between clusters of mobile users and clusters of accessed websites. We show that this method can provide accurate and efficient clustering with minimal information loss. A location-based clustering is then carried out based on website visitation pattern similarity. This method is systematic and can be generally applied to discover important spatio-temporal features of Internet behavior from any similar traces to our dataset.

We report two major findings in this chapter: 1- Mobile users cluster with respect to website visitation patterns into a small set of clusters with clearly distinct profiles. For example, Mac users consistently visit ‘washingtonpost’, ‘cnet’ and ‘apple’ websites and not ‘microsoft’. While PC users visit ‘yahoo’, ‘google’ and ‘microsoft’ websites, but not ‘apple’. 2- Locations in similar categories tend to cluster together, with a few exceptions, in terms of mobile website access. We establish the stability of these findings for three month-long samples. These findings provide the basis for mobile user behavioral models both qualitatively and quantitatively.

This study has the following key contributions:

1. We propose an effective approach for multi-dimensional analysis of very large datasets, and show how information theoretic co-clustering can be applied to create and correlate clusters of users and web domains to build group-specific profiles.
2. We conduct context-specific (location-based) analysis of mobile users' behavior, using two different methods (hierarchical clustering and graph clique detection) to effectively discover groups of locations with similar contexts.
3. We obtain consistent results for clustering of mobile users and locations behavioral similarity that provide the basis for a dependable model of mobile Internet usage.

Data Modeling

In this study, data modeling process is performed at three steps. The first step is to create a global model of dynamics within the network. This model reveals the behavioral groups and provides a big picture of dynamics inside the mobile society. The next step is to build and analyze location/context specific models, i.e., website access patterns in different buildings. The third step is to analyze the stability of learned models.

Global Analysis and Co-Clustering

The main goal of global analysis is to provide a big picture of dynamics within the network. For this purpose, a very well-known approach is to cluster entities (e.g., users, websites) with similar characteristics. However, a major challenge in modeling of multi-dimensional datasets is the fact that ordinary one-sided clustering algorithms like hierarchical clustering or k-means can only cluster data along different dimensions separately [40], i.e., either we get clusters of websites or clusters of users in our case. The proposed approach to resolve this problem is to apply co-clustering techniques, which cluster the input dataset along multiple dimensions simultaneously. In this way, we can correlate different dimensions in a unique model.

In our campus-wide case study, we first investigated applying bipartite graph co-clustering [41]. A graph formulation is used in this algorithm coupled with a spectral heuristic (using eigenvectors) to co-cluster the 2-dimensional input data. However, the restriction of this algorithm was that each row cluster was associated with one column cluster, a restriction which we found inappropriate to impose on our dataset due to the variety of users' trends. Therefore, we chose another approach; the information theoretic co-clustering [1] for simultaneous clustering of users and domains to obtain a global model of the mobile society. The input data for this algorithm is a matrix representing the time spent by users at different websites. The theoretical formulation of the co-clustering technique treats the (normalized) non-negative contingency table as a joint-probability distribution of two discrete random variables, whose values are given in the rows and columns, and poses the co-clustering problem as an optimization problem in information theory. In this technique, co-clustering is performed by defining mappings from rows to row-clusters and from columns to column-clusters. These mappings produce clustered random variables. The optimal co-clustering is one that leads to maximum mutual information between the clustered random variables, and minimizes the loss in mutual information between the original random variables and the mutual information between the clustered random variables. This algorithm monotonically increases the preserved mutual information and optimizes the loss function by intertwining both row and column clustering. Row clustering is performed by calculating closeness of each row distribution (in relative entropy) to row cluster prototypes. Column clustering is performed similarly. This iterative process converges to a local minimum. This algorithm differs from one-sided clustering in that the row cluster

prototypes incorporate column clustering information, and vice versa. The algorithm never increases the loss, and so, the quality of co-clustering improves gradually. It also ameliorates the problems of sparsity and high dimensionality. Iteratively, the method performs an adaptive dimensionality reduction and estimates fewer parameters than one-dimensional clustering approaches, resulting in a regularized clustering. In addition, the algorithm is efficient. The computational complexity of the algorithm is given by $O(N \cdot \tau \cdot (k + l))$ where k and l are the desired number of row and column clusters, N is the number of non-zeros in the input joint distribution and τ is the number of iterations; empirically 20 iterations are shown to suffice.

As the number of users and accessed domains is high and on the other hand the co-clustering algorithm has size limitations, we filter the dataset and limit ourselves to the most active websites and aggregate the data for only those domains. After performing the co-clustering algorithm and extracting clusters of users and domains, we can create an interest level matrix indicating the interest level of different groups of users in different clusters of domains. For each pair of user and domain clusters, the interest level is calculated by summing up the amount of all joint probabilities between them.

Location-Based Analysis

The main goal of location-based analysis is to discover different clusters of locations with similar context in terms of web access patterns. For this purpose, we first define a uniform way to describe the context of a location. Then, we formulate a comparison method between different locations to find their similarities. Finally, we devise an appropriate method to detect contextually similar locations.

As for the first step, the global acquired clusters of users and domains can be employed to provide a uniform way for context description. For each location, an interest level matrix for acquired clusters of users and domains can be formed using a uniform ordering of the clusters. We employ this location-specific interest level matrix as a context descriptor of the location.

In the second step, we provide a method for comparing the context descriptors of different locations. For this purpose, we treat the corresponding interest level matrices as vectors of all their values and employ cosine distance function. Using this method, we can create a dissimilarity matrix for different locations based on their context descriptors.

For the final step, we propose two different methods for finding groups of contextually similar locations. The first technique is to use hierarchical clustering to form clusters of locations. The second method is to map the dissimilarity matrix to an undirected graph as follows; considering a node for each location, we draw an edge between two different nodes if their dissimilarity is less than a threshold. Then, we find cliques within the graphs to discover groups of locations with similar contexts.

Stability Analysis

An important goal in data modeling is to discover stable models that can accurately describe not only the current state but also its time evolution and dynamics (i.e., its history and future). Such models are valuable in the sense that they can explain major trends during a long period of time (e.g., a semester) and thus can be effectively used for realistic and durable behavior-aware designs.

To assess the relative stability of patterns captured by global and location based models, we investigate whether the discovered clusters of users, domains and locations

are sufficient to describe the history and the future of the mobile Internet access. Our method for measuring this forward, backward stability of the discovered clusters over time is to: 1- take all the interactions for the same sets of users, domains and locations during the previous and the next periods of the analyzed period; then, 2- for each of the periods, recreate the global interest level matrix and the location dissimilarity matrix using the same acquired clusters; and finally, 3- calculate the distance in between corresponding matrices for the analyzed period and the previous/next one. In this calculation, matrices are again treated as vectors of all their values and their distance is determined by cosine distance function.

Experimental Results

In our case study, we conducted a campus-wide analysis on the data we collected from the University of Southern California (USC) in 2008. For this dataset, we performed global and location-based analysis based on the techniques explained in the previous section. In the following we present the results.

Global Analysis Results

For global analysis, a matrix was created associating the user IDs and domains (i.e., websites) using the corresponding total online time (per minute). For our analysis, we had 22,816 users, and 100 domains. The data is scaled using row-normalization of log the online time values. This was the input data for our modeling problem for which we applied the information theoretic co-clustering. In this case study, we discuss results for 10 clusters (i.e., with 10 as input to the algorithm for the number of output user and domain clusters[1]). Using this method of co-clustering, we produce two collections of

[1] Several values for the number of clusters (10, 20, 40) were investigated with 10 producing the best results.

domain clusters and user clusters, which are used to determine an interest level between each pair of user and domain clusters. Figure 4-1 shows the result of applying this method on the scaled data and Figure 4-2 depicts the interest level matrix between the resulting clusters. Each row in Figure 4-2 identifies a group of users in terms of their interest level to different groups of domains.

As shown, the co-clustering algorithm is able to group users with similar access patterns into clusters. In a way, users within the same cluster may be characterized by similar set of favorite wireless online activities. At a high level, we observe 4 general classes of user clusters with following characteristics:

1. Narrow access (cluster 1): users access two (or less) clusters of domains, in this case clusters *I* and *J* including 'usc' and 'infoave' (telecom and webhosting).
2. Narrow spread access (clusters 2,3): most user access time is spread over 3 or 4 domain clusters only.
3. Medium spread access (clusters 4-7): most user access time is spread over 5 to 8 domain clusters.
4. Wide spread access (clusters 8-10): with noticeable user access in all domains.

A deeper look into the clusters reveals some interesting facts. Clusters 2 and 3 include narrow spread users, but include clearly distinct sets of user interest. Cluster 2 shows users who mostly just utilize the Internet for search or email via 'yahoo' and 'google', and visit 'microsoft' for probably getting software updates, and thus are likely Microsoft/PC users. Cluster 3, by contrast, shows users frequently go to 'apple' and 'mac' sites but rarely go to 'microsoft', and thus are likely Mac users. Note that these users are commonly interested in 'washingtonpost' and 'cnet' but not interested in 'facebook' or 'yahoo' at all. Cluster 7 also depicts heavy Mac users. Again, as can be seen in the figure, these users rarely go to 'microsoft' but are interested in 'mac', 'apple',

‘washingtonpost’, ‘cnet’ and also visit ‘facebook’, ‘yahoo’ and some other websites frequently. Table 4-1 shows some other domains that are clustered together.

To assess the relative stability of trends, we process the records from Feb. 2008 and Apr. 2008 and recreate the global interest level matrices for them using the same clusters and ordering of users and domains in Figure 4-1. The results are depicted in Figure 4-3 and Figure 4-4 and indicate that the trends hold to a large extent; the interest level matrix for February is 92.25% and for April is 89.18% similar to that of March; plus, the interest level matrices of February and April are 98.51% similar. This indeed indicates the stability of the results.

Location-Based Analysis Results

In the second phase, we cluster locations based on their acquired context descriptors and analyze the results based on their actual context. For this purpose, for any interaction between a user and a domain, we first identify the switch-port that handle the connection using the WLAN traces. Then, we associate the interaction to its location among 84 buildings across the campus using a mapping table between the switch-ports and buildings. Next, all active buildings (handling at least one interaction) in March 2008 are selected (79 buildings) and their context descriptors are created as explained before. Finally, we create dissimilarity matrix for all the selected buildings based on their context descriptors. This matrix is used by two different techniques based on hierarchical clustering and graph clique detection to discover groups of contextually similar buildings. Figure 4-5 shows the result of applying hierarchical clustering for creating 10 clusters of locations. In the figure, all clusters can be identified using green line borders and all distances (dissimilarities) can be found using the z-axis. For each cluster, Figure 4-6 shows the average dissimilarity between each

corresponding building and all the others. To analyze the resulting clusters, we studied all the buildings and based on their actual context categorized them into 10 categories including: housing, auditorium, (outdoor) activity, sorority, fraternity, school, health, music, cinema and service (Table 4-2). In Figure 4-5, the category of each building is visualized by the assigned color in Table 4-2.

As can be seen in Figure 4-5, most of the buildings in the same category are clustered together into one or two clusters. For example sororities are all clustered together and fraternities form two major clusters and two uni-member clusters. The interesting point about the fraternities is the fact that those two uni-member clusters include professional fraternities and the other two contain social ones. We can also see that all auditoria are in the same cluster as well as cinema-related buildings. Regarding the “activity” category that includes buildings with different activity context including sports, religion, social and shopping, we notice that 6 out of 8 are in the same clusters while 3 of 6 are sports related. In addition, it can be observed that housing buildings form two major clusters and there is only one separated building in another cluster. The study of the building reveals the fact that it is the only housing complex that includes a plaza and a bookstore too. Health related buildings are also assigned into two main clusters. However, buildings in school and service categories are almost scattered among clusters because of the fact they include different types of schools for social work, journalism, humanities, letters and arts, law and leadership and different kind of centers for facilities management, financial, communication and computing services.

As before, to assess the relative stability of trends, we process the records from Feb. 2008 and Apr. 2008 and recreate the dissimilarity matrices for them using the

same acquired clusters for March. The results are depicted in Figure 4-7 and indicate that the trends hold to a large extent; the dissimilarity matrix for February is 92.72% and for April is 95.12% similar to that of March; plus, the dissimilarity matrices of February and April are 93.35% similar. This indeed indicates the stability of the results.

The graph-based method also detects cliques within the corresponding graph for dissimilarity matrix using the threshold of 0.06. Figure 4-8 shows the resulting graph layout for the data from March. As can be inferred from the histogram for the dissimilarity matrix (Figure 4-9), the resulting graph includes around 10% of all possible edges using the mentioned threshold. As can be seen in the graph, a clear relationship again exists between identified cliques and the actual categories of buildings.

Summary

In this study, we provided a systematic method to analyze the largest mobile trace to date, with billions of records of Internet usage from a campus network, including thousands of users and dozens of buildings. Novel analysis was conducted utilizing advanced data mining using efficient co-clustering, at the global and location-based levels. We have shown that mobile Internet usage can be modeled with a strikingly small number of clusters of distinct web access profiles. Similarly, building categories show very distinct Internet usage patterns and are often clustered together. These trends were found to be highly stable over time. The details of this study enable the parameterization of new and realistic models for mobile Internet with applications in several areas of networking, including simulation, traffic analysis and interest-aware design which we will discuss in the next chapters.

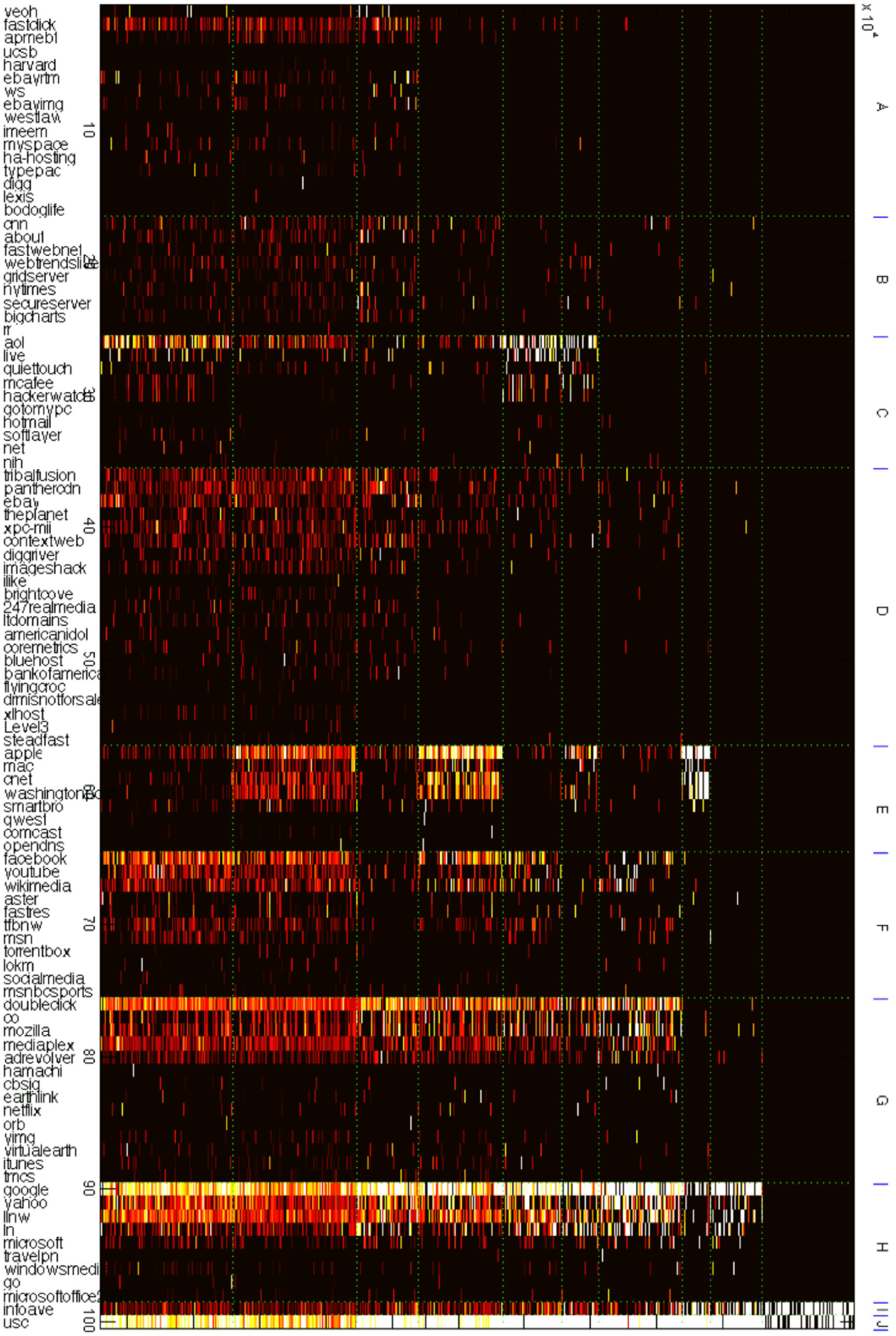


Figure 4-1. Information theoretic co-clustering on user-domain matrix, March 2008. The result is given for 10 clusters of users (1 through 10) and 10 clusters of domains (*A* through *J*). Colors in the range of black to white shows values in the range of minimum to maximum. X-axis shows the domain names and y-axis indicates the user ID.

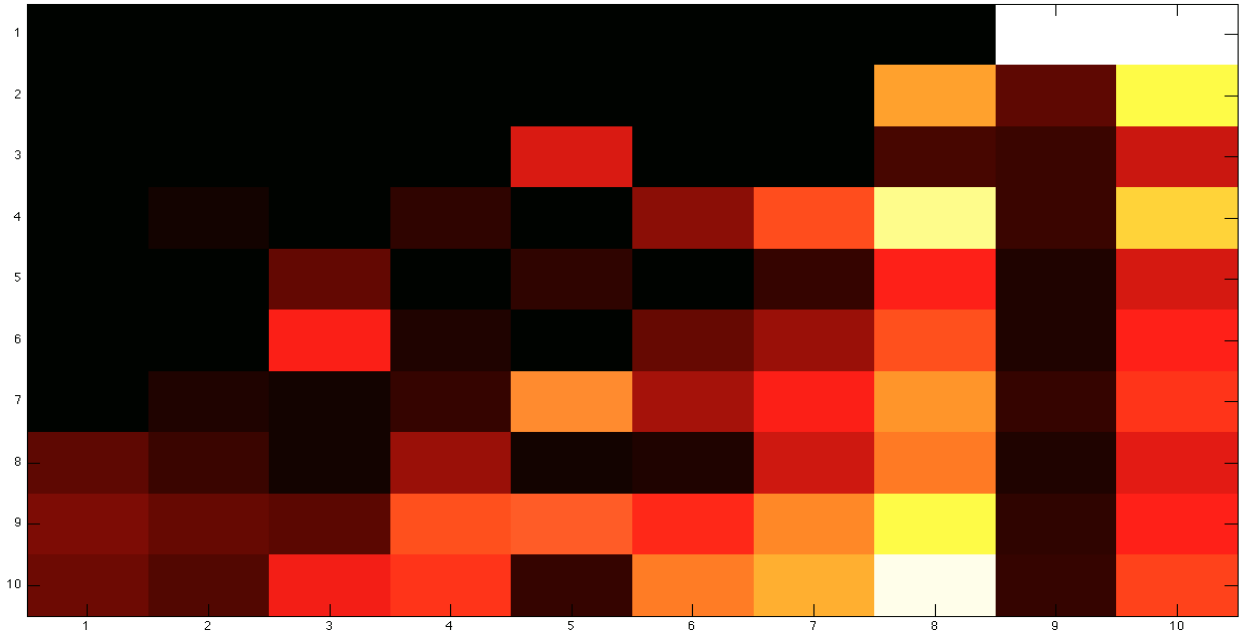
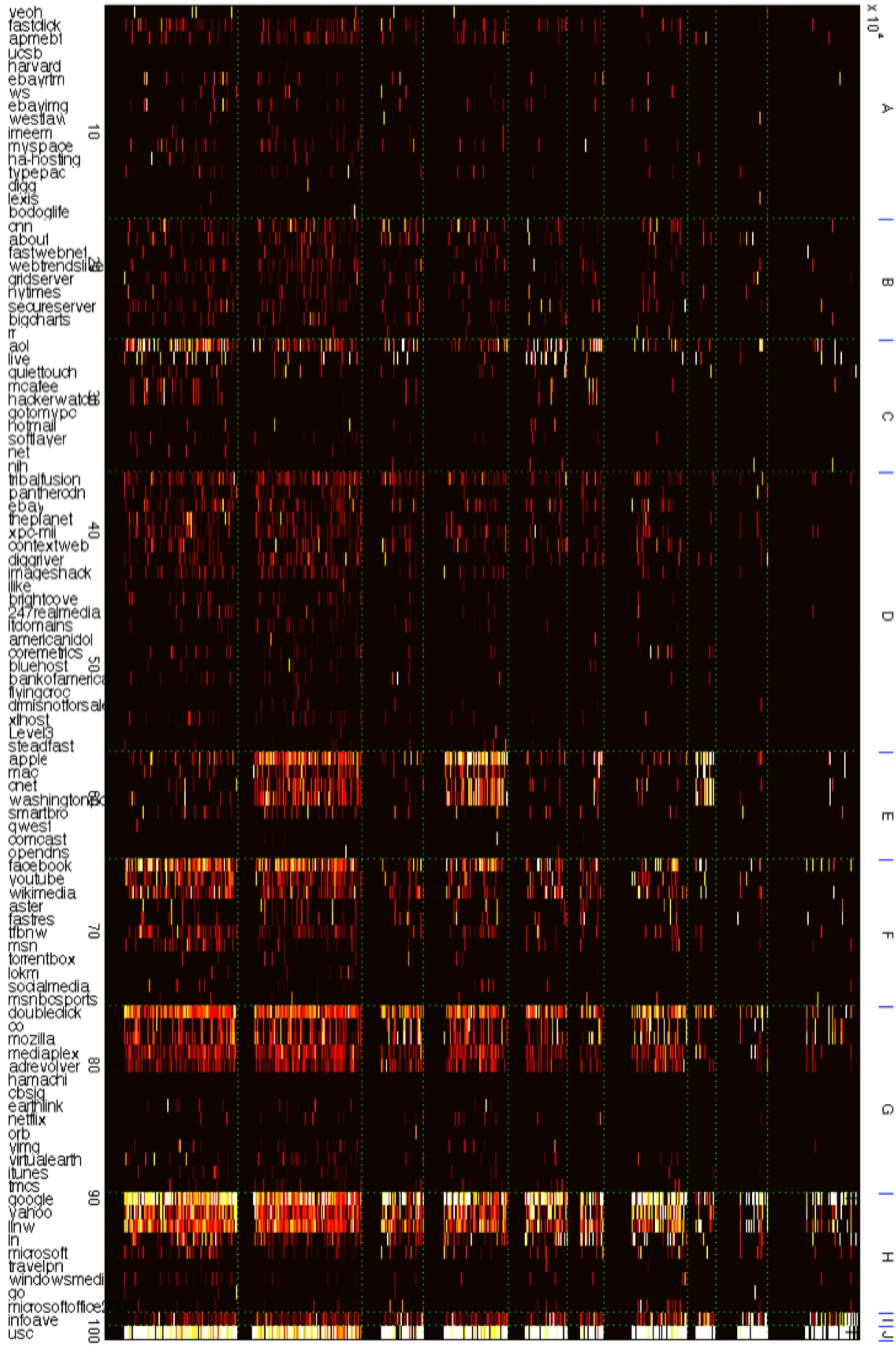


Figure 4-2. Interest level matrix for resulting user and domain clusters by applying information theoretic co-clustering (March 2008).

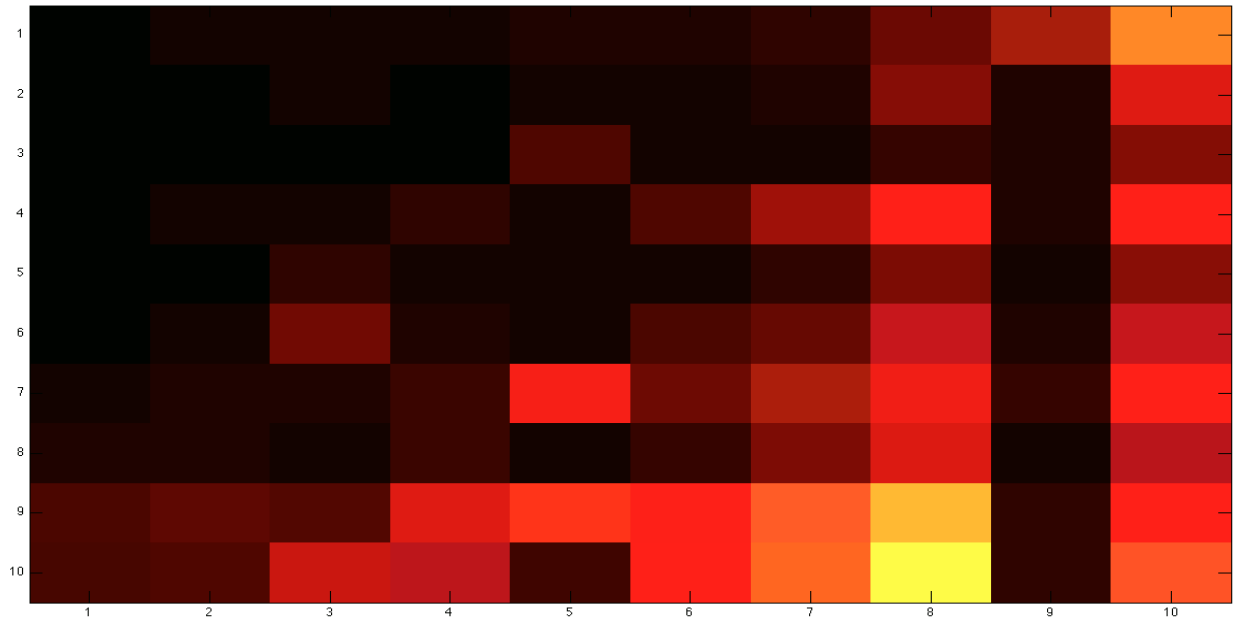
Table 4-1. Major related websites clustered together

Cluster	Domains
A	myspace - imeem (social media service) - digg (social news) typepad (blogging) - ebayrtm - ebayimg - wsj (business news) bodoglife (online gambling) - ucsb - harward - westlaw
B	cnn - nytimes (new york times)
C	mcafee – hackerwatch - live - hotmail
D	ebay - bankofamerica
E	apple - mac - washingtonpost - cnet
F	facebook - youtube - social media - msn - msnbcsports
G	netflix - itunes - orb (media cast) - tmcs (social city search) virtualearth (online map)
H	google - yahoo - microsoft - windowsmedia microsoftoffice2007



A

Figure 4-3. Stability analysis result for co-clustering model (Feb 2008). Using the same column and row ordering of users and domains in Figure 4-1, the graphs are constructed using Feb. 2008 measurements. The trends are relatively stable from one month to another especially for the narrow spread and wide spread clusters. A) Co-clustered matrix. B) Interest level matrix.



B

Figure 4-3. Continued

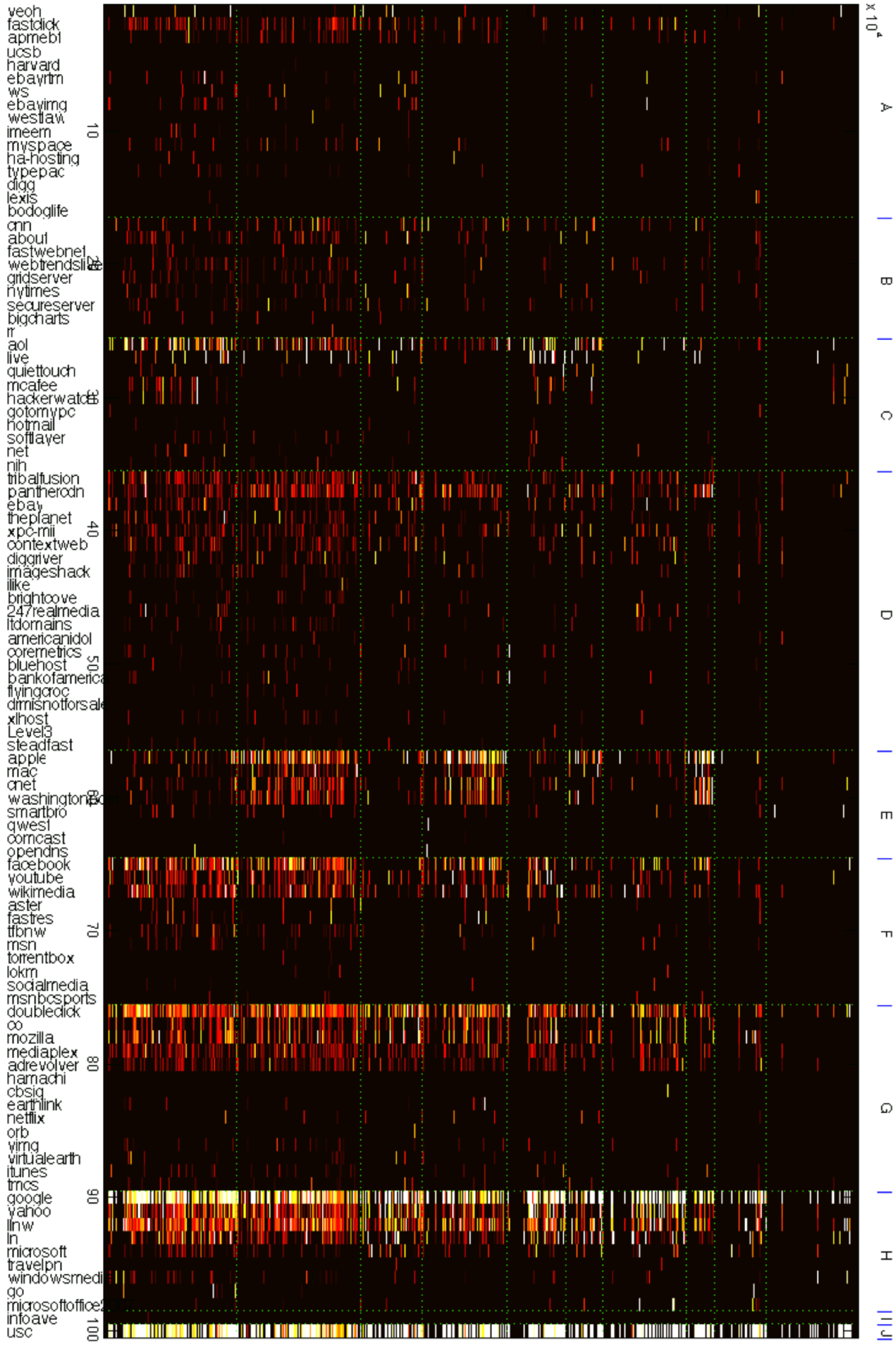
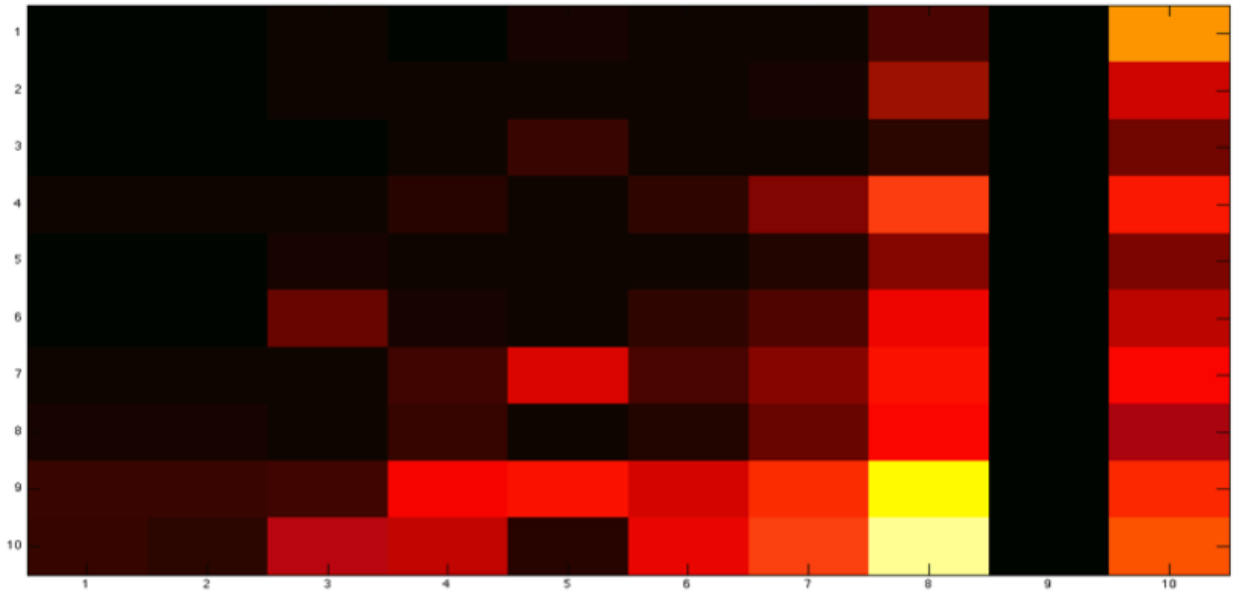


Figure 4-4. Stability analysis result for co-clustering model (Apr 2008). The trends are relatively stable from one month to another. A) Co-clustered matrix. B) Interest level matrix.



B

Figure 4-4. Continued

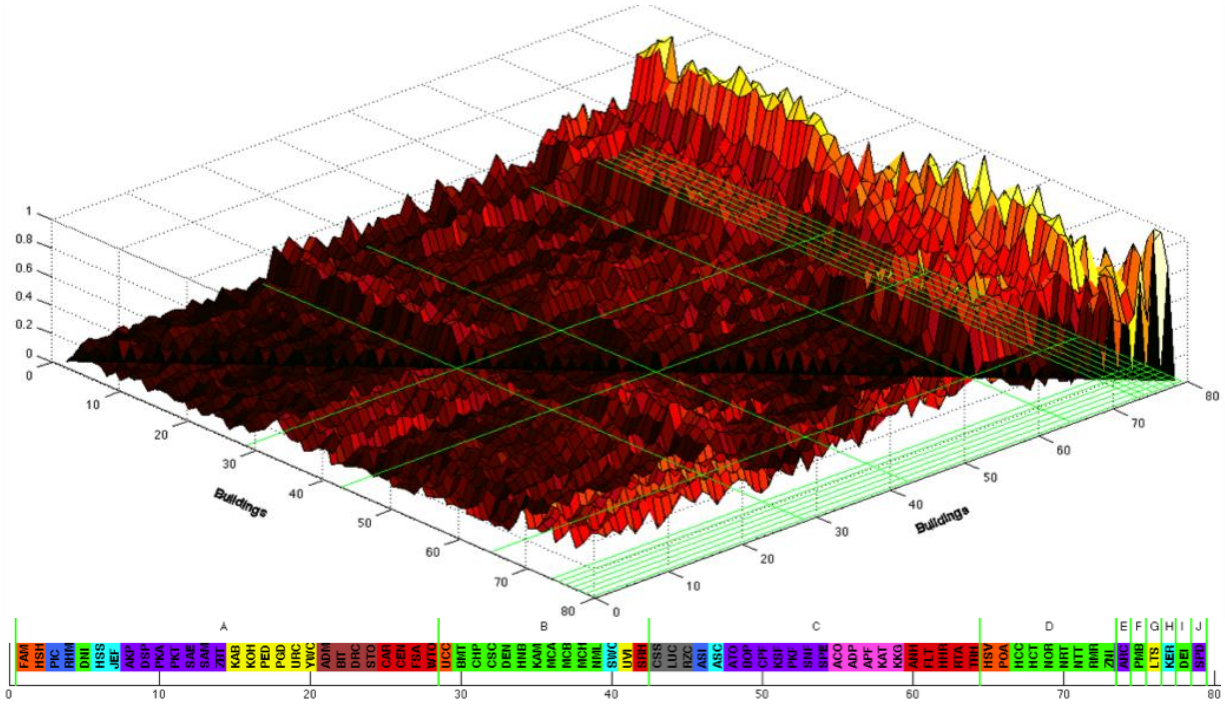


Figure 4-5. Clusters of buildings for March 2008 (distinguished by green lines); x and y axes show building IDs which are matched to the colored abbreviations below the chart and z-axis shows their dissimilarities (distances) in between 0 and 1. The background color of each abbreviation shows the category of the building based on the color mapping in Table 4-2.

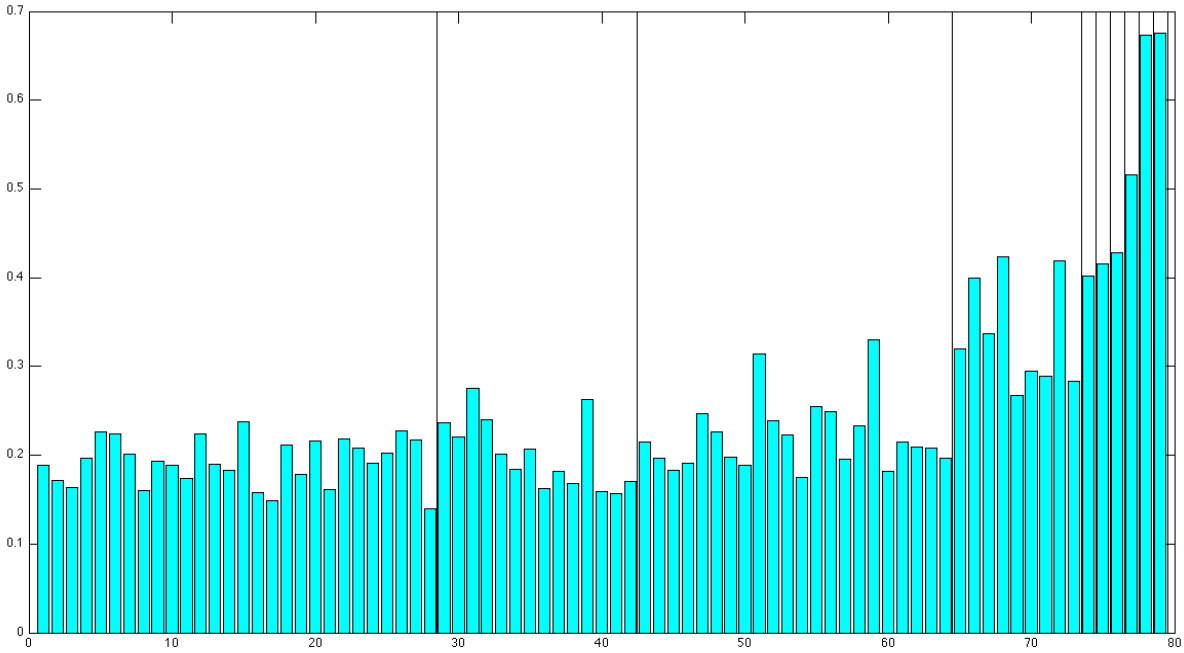
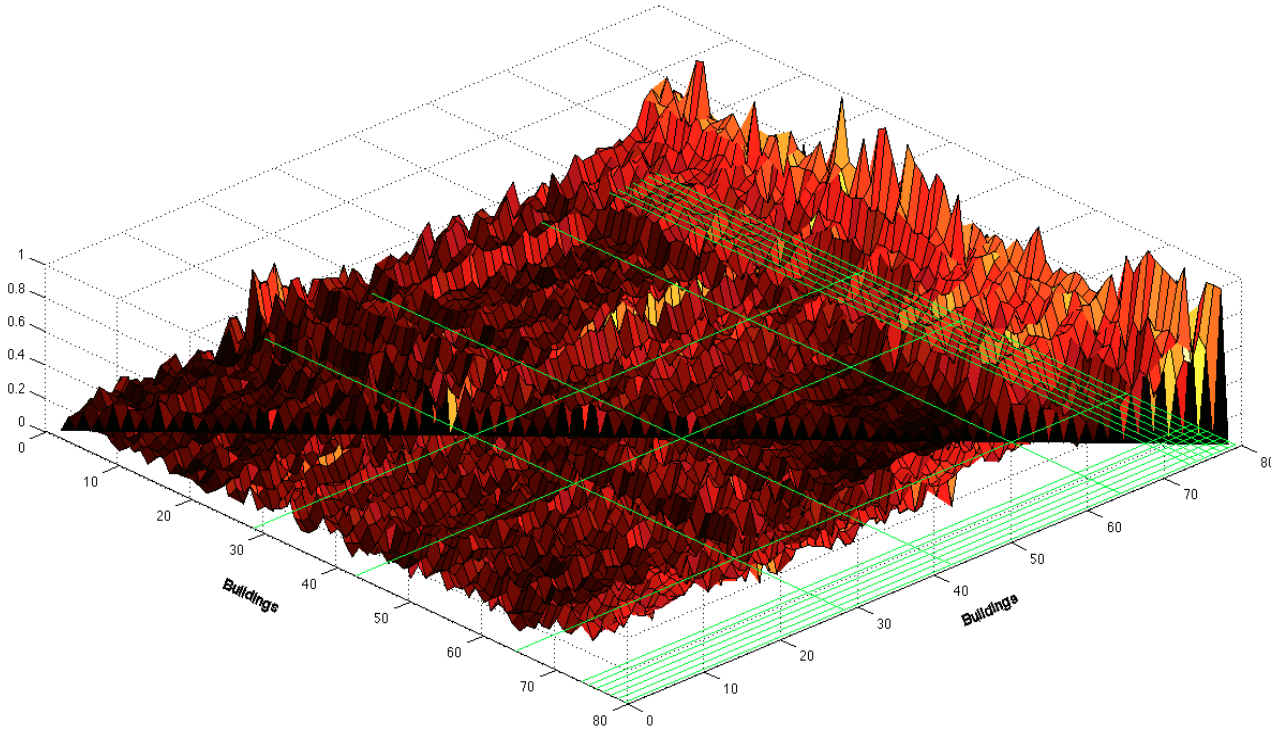
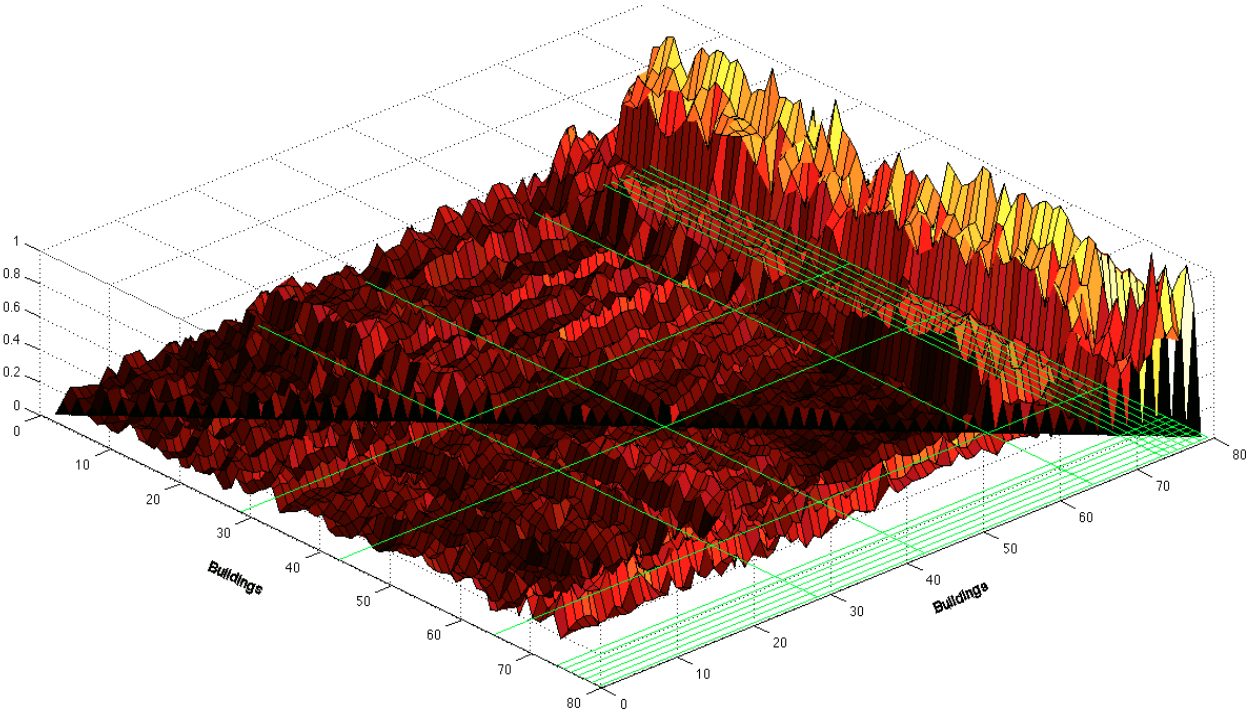


Figure 4-6. Average dissimilarity between each building in a cluster and all the other buildings (March 2008). Y-axis shows the dissimilarity between 0 and 0.7.



A



B

Figure 4-7. Clusters of buildings for a month before and after. A) February. B) April

Table 4-2. Building categories. The assigned colors to the categories are used in Figure 4-5 and Figure 4-8.

Category	Building abbreviation										
Activity	KAB	KOH	LTS	PED	PGD	URC	UVI	YWC			
Auditorium	ADM	BIT	DRC	STO							
Cinema	CSS	LUC	RZC								
Fraternity	AKP	ARC SAE	ATO SAM	BOP SNF	CPF SPD	DSP SPE	KSF ZBT	PKA	PKF	PKT	
Health	BMT	CHP MCA	CSC MCB	DEI MCH	DEN NML	DNI NOR	HCC NRT	HCT NTT	HNB PMB	KAM RMR	ZNI
Housing	ANH	CAR	CEN	FLT	FSA	HHR	RTA	SRH	TRH	WTO	
Music	ASI	PIC	RHM								
School	ASC	HSS	JEF	KER	SWC						
Service	FAM	HSH	HSV	POA	UCC						
Sorority	ACO	ADP	APF	KAT	KKG						

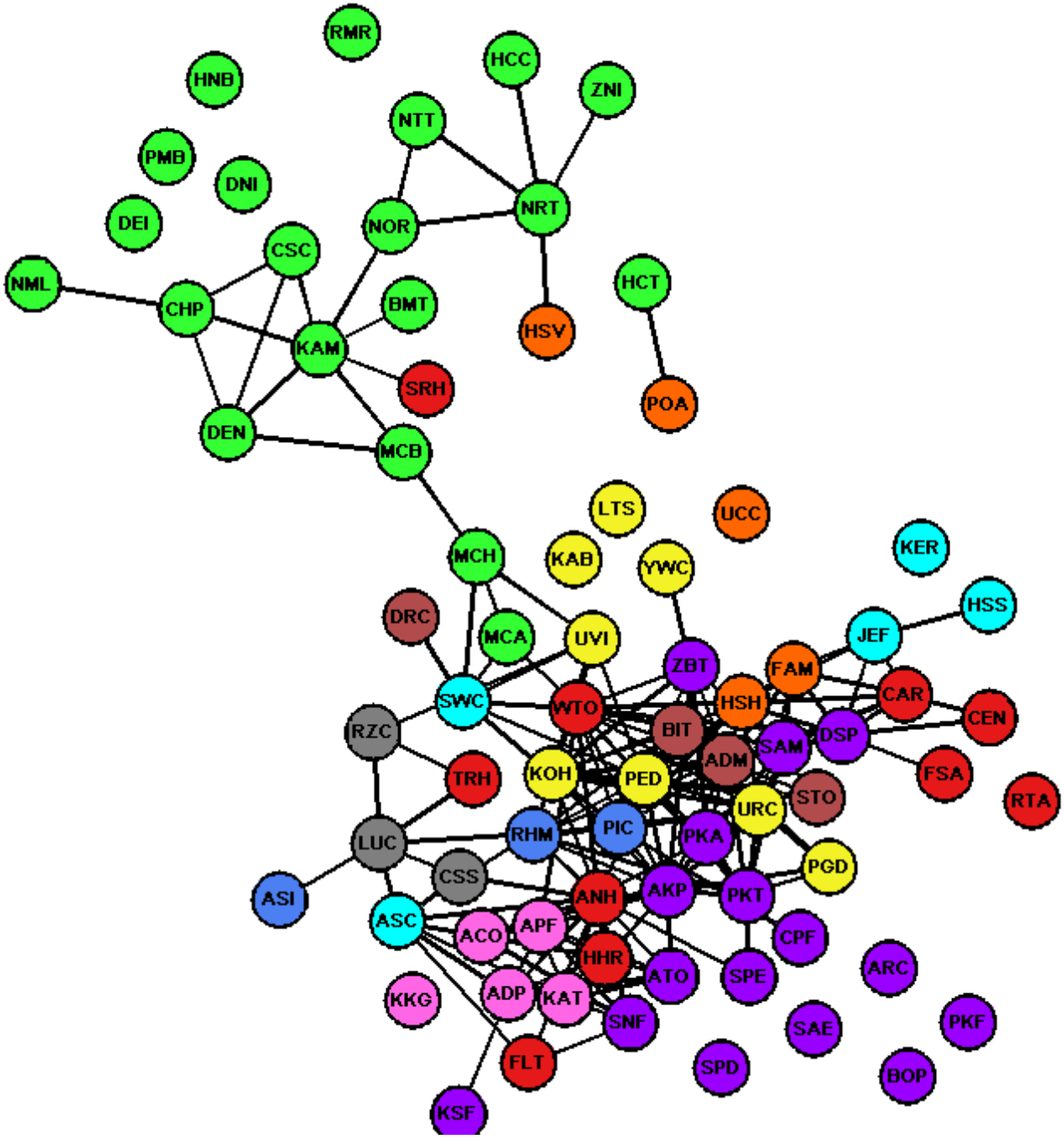


Figure 4-8. Graph representation of dissimilarity matrix using the threshold of 0.06 for March 2008. (Table 4-2 shows the mapping between colors and categories).

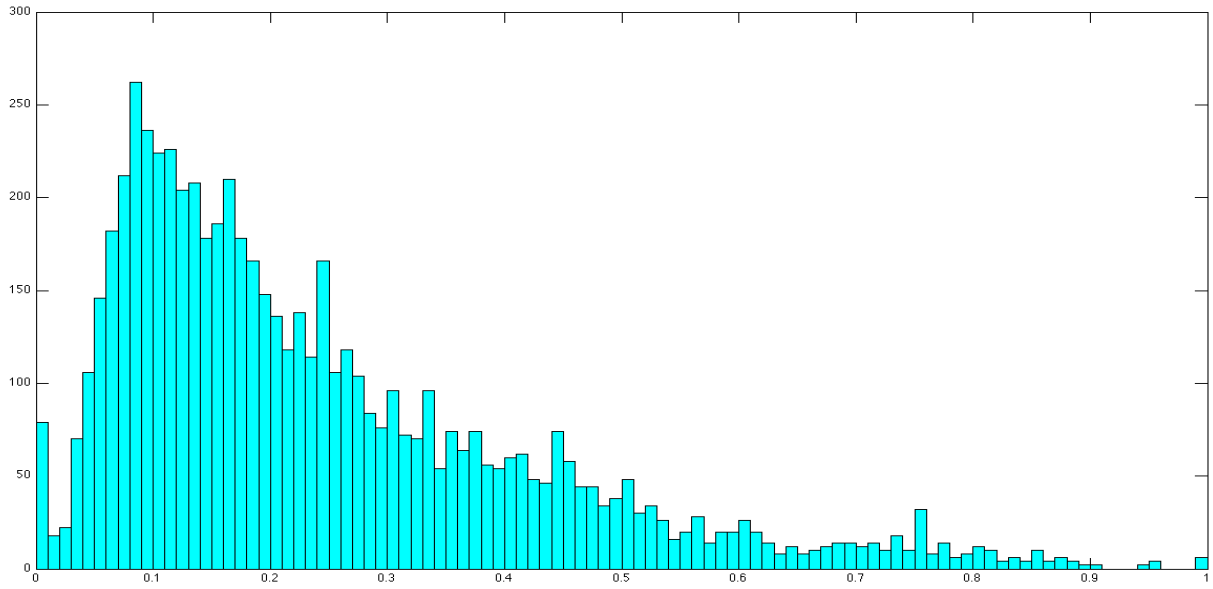


Figure 4-9. Histogram for the dissimilarity matrix. X-axis shows the dissimilarity in the range of 0 to 1.

CHAPTER 5
NEURAL NETWORK MAP MODEL OF MOBILE ONLINE BEHAVIOR

In today's mobile networks, many interactions and movements happen in every single second. In such a dynamic environment, discovering trends and correlations in users behavior can help us significantly to provide better solutions to the networking problems. For this purpose, in this chapter, we focus on the study of trends and behavioral correlations according to the website and location visitation patterns [42-45]. Although clustering techniques provide some insight on similarities between data items (and among features, if co-clustering [1] is used), they do not reveal intuitively how correlated they are. Therefore, they are not effectively useful for discovering trends and behavioral correlations. In this chapter, we introduce kind of neural networks called self-organizing maps (SOM) [2] to resolve this problem. While the co-clustering model provides a high level approach for the analysis of Internet usage (at cluster level), the SOM technique provides a tool for low level (feature level) analysis of online activities, mobility and trends. Using the SOM model, we can discover many correlations, anti-correlations and anomalies between domains and locations and infer '*trends*' inside the mobile society instead of focusing on individual behaviors.

A self-organizing map is an artificial neural network that is trained using unsupervised learning to generate a discretized low-dimensional representation (a map) of the input space of the data samples. Unlike other artificial neural networks, self-organizing maps use a neighborhood function to preserve the topological properties of the input space. The topology-preserving mapping keeps the more similar data groups closer together in the final map, which makes SOM useful for providing low-dimensional

views of high-dimensional data. These views can reveal the semantic behind major user trends and the correlation between different features.

Using the SOM technique, we can extract minor and major trends in mobile users' website/location visitation patterns and also important correlations between different web domains and locations. In this chapter, we show how to apply this technique methodically to the collected large-scale multi-dimensional dataset with minimal computational complexity to facilitate its meaningful analysis. This method is systematic and can be generally applied to discover important features of Internet behavior from other similar traces. It can also be applied for any other aspects of Internet usage, e.g., time and application in addition to the web domains and locations.

We report three major findings in this chapter. First, mobile users' behaviors based on web domain and location visitations can be accurately modeled by a small set of neurons (nodes) in the SOM, which can be further clustered into smaller number of major trends with clearly distinct characteristics. For example, a major trend represents Mac users who frequently visit 'mac' and 'apple' websites and have strong interest in 'washingtonpost' and 'cnet' too. Second, web domains / locations in similar categories tend to be modeled by an adjacent set of neurons. For example, most of the advertisement / marketing domains or fraternities are modeled together by neighboring neurons. Third, many nontrivial correlations exist between different kinds of web domains and locations. For example, we found that Music Practice Center is highly correlated to the Health Science Telephone Vault while they are located in two different campuses of USC.

This study has the following key contributions:

1. We propose an effective approach for multi-dimensional modeling of large mobile societies and show how self-organizing maps can be applied to model minor and major trends in wireless users online activity and mobility.
2. We conduct domain-specific and location-based analysis of mobile users' behaviors, using the feature maps extracted from the SOM and show how this method can be effectively applied to discover correlations among different domains and locations.
3. We suggest feature clustering technique on top of the SOM as a quantitative way for discovery of domains and locations correlations.
4. We propose an extension on top of the SOM for multi-aspect (concurrent) modeling of users behavior based on their online activity and mobility.

Modeling Approach

In this study, we perform the data modeling in two major steps. In the first step, we employ the self-organizing map to learn minor trends of users within the wireless society. The user trends may be learned based on website or location visitation preferences separately (using a common approach) or together (based on our proposed multi-aspect extension of the method). In any case, in the second step, we apply clustering technique on the map nodes to discover major trends inside the community.

Trend Modeling

The SOM technique [2] provides a powerful yet intuitively understandable tool for unsupervised learning and data visualization. The SOM is defined as a set of nodes which develop a mapping of high-dimensional input vectors (which may represent website or location visitation preferences) onto a discrete output space (the "map") such that each region on the map represents an area of the input space. This mapping preserves the topology of the input space in a way that local similarity of input patterns is reflected by proximity on the map. Therefore, it can be effectively applied in capturing the properties of the input space of users' behaviors and organizing their trends in an

ordered fashion. In a self-organizing map, a weight vector of the same dimension as the input data vectors and a position in the 2D map space are associated with each node (or neuron in neural networks). The usual arrangement of nodes is in the form of a hexagonal or rectangular grid. SOM training, i.e., the iterative adjustment of the weight vectors to acquire a desired mapping, is performed by successive presentation of all input data where each presentation leads to the adjustment of weights to the presented data. The training is based on two principles:

- Competitive learning: the weight vector most similar to a data vector is modified so that it is even more similar to it (the corresponding node is called Best Matching Unit or BMU). This way the map learns the position of the input data.
- Cooperative learning: not only the most similar weight vector, but also its neighbors are moved towards the data vector. This way the map self-organizes.

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

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

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

rate of the map decreases monotonically with the number of iterations to ensure convergence.

In this way, each neuron can learn a minor trend that represents a set of similar input data vectors. This is one of the major advantages of SOMs with respect to clustering techniques. While a clustering technique attempts to partition the input space (e.g. users' behaviors) by assigning each sample (e.g. a user) to a cluster, the SOM technique attempts to learn trends inside the input space form the samples. Note that each input data vector (e.g. a user) affects a set of neighboring neurons (trends) and therefore the input space is not distinctly partitioned by the neurons (unlike cluster assignment in conventional clustering techniques). This approach better adapts to the natural human behaviors with overlapping characteristics.

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

Trend Clustering

One way to visualize the resulting map after the training phase is to create U-matrix (unified distance matrix). The U-matrix shows the distance between the weight

of each node and the assigned weights of its neighbors after the learning process. Figure 5-1 A shows an instance of U-matrix with interpolated shading of colors. Small U-values (blue areas in the figure) indicate homogenous neighborhoods and large ones (red areas) depict heterogeneous neighborhoods. As large U-values mean large distances between the neighboring nodes, they can be interpreted as borders between clusters of neurons, i.e., trends. In order to find these borders (clusters), k-means [40] clustering algorithm can be applied. Because k-means result depends on the initial choice of cluster centroids, the algorithm is run multiple times for a given *k* and then the best result is selected based on the sum of the squared errors. Because the captured minor trends are already very well organized on the map, each resulting cluster maps into a contiguous area of neurons, representing a major trend (Figure 5-1 B). Clustering of trends instead of original data reduces the required computational time for any kind of clustering technique as the size of input is decreased. This is very important when dealing with massive amount of data. In addition, as the weight vectors are local averages of the data, the clustering result is less sensitive to random variations in the input data.

Multi-Aspect Modeling

The SOM technique in its original form is suitable for uni-aspect modeling of trends, i.e., based on either web domain or location aspect. However, in multi-aspect modeling the goal is concurrent modeling of trends based on all aspects together not separately. While uni-aspect modeling is good for intra-aspect analysis, multi-aspect modeling provides an opportunity for inter-aspect analysis. In multi-aspect modeling, instead of one general usage pattern vector per user, a set of localized vectors exists (i.e., a set of web access pattern vectors for different locations). Therefore, the regular

SOM learning method is not applicable. Our proposed approach to accommodate this situation is to consider a usage matrix for each user (representing website usage at different locations) and a weight matrix for each map unit and then get the map trained. This way each map unit can capture a multi-aspect trend. This is an extensible approach and can be applied for more than two aspects.

Data Analysis

We conduct uni-aspect and multi-aspect analysis considering web domain and location visitation aspects. For each of the analysis, we propose two qualitative and quantitative approaches. The qualitative approach relies on the visual inspection of extracted feature maps and is useful for discovering correlations, anti-correlations and anomalies among the features. The quantitative approach is based on our proposed feature clustering technique that applies a mathematical correlation function.

Feature Map Analysis

The feature maps are extracted from the SOM and show what kind of values the weight vectors of the map units have for each feature. In other words, a feature map shows the projection of the SOM for the corresponding feature (which can be a web domain, a location, or a web domain at a specific location in multi-aspect case). The value of each unit for the feature is presented with a color. Figure 5-2 shows a group of resulting feature maps in our study. By visual inspection of the feature maps, we can find many different interesting facts about the trends and features as follows:

1. Comparison of feature maps with the clustered SOM discovers the semantic behind each cluster of trends representing a major trend. For a cluster area, features whose maps looks red in the same area disclose the main captured trends by the cluster.
2. Similar feature maps reveals correlations between the corresponding features. The correlation can be partial or complete. If the maps seem highly similar, there exist rather complete correlation, but if they are partially similar, the correlation among

features will also be partial. In our case, correlation between a set of features means that they have the same visitation pattern.

3. Anomalies in a set of feature maps uncover anomalies regarding the corresponding features. In our case, for example, if for a category of web domains (e.g. marketing domains) all but one feature maps looks similar; the different one brings out an anomaly.

4. Feature maps that look inverted (i.e. red areas in one are blue in the other) disclose anti-correlations. Again, anti-correlations can also be rather complete or partial.

Feature Clustering

Taking the projection of all weight vectors (or weight matrix in multi-aspect case) on each feature, we propose to construct a description vector for the corresponding feature referred to as feature vector. By applying hierarchical clustering on the feature vectors, we can cluster features based on their correlation using the following correlation distance function:

$$D(i,j) = 1 - \frac{(v_i - \overline{v_i})(v_j - \overline{v_j})'}{\sqrt{(v_i - \overline{v_i})(v_i - \overline{v_i})'}\sqrt{(v_j - \overline{v_j})(v_j - \overline{v_j})'}}$$

where v_i and v_j are feature vectors. This procedure can also be interpreted as a quantitative way for comparing the feature maps.

Experimental Results

In our experiment, we conducted a campus-wide case study on the data we collected from the University of Southern California (USC) in 2008. For this dataset, we performed domain-specific, location-based and multi-aspect analysis based on the techniques explained in the previous section. In the following we present the results.

Modeling Results

For domain specific and location-based modeling, two separate matrices were created associating the user IDs with web domains and user IDs with locations using

the corresponding total online time (per minute). For our analysis, we had 22,816 users, and 100 domains and 79 buildings. The data for each matrix is scaled using row-normalization of log the online time values. The two input matrices trained two SOMs of 32 by 24 nodes separately. Figure 5-1 A shows the U-matrix for domains and Figure 5-1 B represents the corresponding SOM clustered into 20 clusters. For multi-aspect modeling, we chose the highest active 40 domains and 20 locations and created a 3D matrix associating the user IDs, web domains and locations using the corresponding online time and trained a SOM of 32 by 24 nodes (see [46] for location and multi-aspect maps).

Domain Specific Analysis

We created the feature maps for all the domains. Figure 5-2 and Figure 5-3 show several examples of resulting maps for different types of web domains. Inspection of the feature maps reveals many interesting facts. The following are some examples based on the presented feature maps here.

Figure 5-2 shows feature maps for advertisement and marketing domains. All these maps (except the right one) show a red area almost at the same neighborhood (right-bottom corner). This shows the major trend captured by the cluster depicted by orange at the same area in Figure 5-1 B is toward this kind of web domains.

High similarity between feature maps in Figure 5-2 shows that the corresponding domains for advertisement and marketing are highly correlated. We can also observe high correlations between the following groups of domains from Figure 5-3: i) security related domains, i.e., ‘mcafee’ and ‘hackerwatch’; ii) ‘itunes’ and ‘netflix’ (online media); iii)’mac’, ‘apple’, ‘washingtonpost’ and ‘cnet’ (showing a strong trend of Mac users toward ‘cnet’ and ‘washingtonpost’); iv) Windows related domains, i.e, ‘microsoft’,

‘windowsmedia’ and ‘microsoftoffice2007’. In the figure, we can see that ‘itunes’ is in one hand partly correlated to ‘netflix’ and on the other hand is partly correlated to ‘mac’ and ‘apple’. This may show the facts that i) Mac users dominantly use iTunes for online media and ii) Netflix costumers shop in iTunes store too. Different patterns of maps for ‘doubleclick’ among advertisement and marketing domains show an example of anomalies within a category of web domains. These anomalies might disclose different advertisement and marketing approaches taken by ‘doubleclick’.

Figure 5-3 reveals anti-correlation between Mac and Windows related domains as well. As can be noticed, the bight (red) area for ‘apple’ and ‘mac’ is almost dark (blue) for ‘windowsmedia’, ‘microsoftoffice2007’ and ‘microsoft’. We can also find anti-correlation between security related domains (i.e, ‘mcafee’ and ‘hackerwatch’) and ‘mac’ and ‘apple’, but partial correlation between them and Windows related domains.

We also applied our proposed feature clustering technique on top of the SOM for web domains and created 20 clusters. Table 5-1 shows some of the resulting clusters. As can be seen in the table, the two discussed categories of Apple and Microsoft correlated domains are clustered into two distinct clusters (Clusters A and B).

Location-Based Analysis

Similar to the domain-specific analysis, we can simply find the semantic behind each major trend for location visitations (see [46]). Inspection of the feature maps for the locations reveals many interesting correlations too. Figure 5-4 shows high correlations between social and professional fraternities. As can been seen, fraternities in the first row are highly correlated. We can also observe high correlation between ATO and ARC buildings. The feature map for PGD (playground) shows that both groups are partially correlated with the playground duplex too. Similar to domain-specific analysis,

discovered correlations among locations are not just between buildings of the same types. Figure 5-5 shows four pairs of highly correlated buildings that are not in the same category. As can be seen in the figure, the Music Practice Center (PIC) is highly correlated to Health Science Telephone Vault (HSV). The interesting point about these two buildings is the fact that they are located in two different campuses of USC and so relatively far from each other. However this is not the case for the Woman's Association (YMC) which is next to the Hall Building (HSH) and probably use the hall for their gatherings very frequently. We can also see the residents of housing complex TRH frequently go to the Healthcare Consultation Center (HCT). Also, fraternity PKT and sorority KAT are highly correlated which may reveal the fact that many of their members are in a relationship.

We also employed feature clustering on the location SOM and created 20 clusters. Figure 5-6 shows clustered heatmap of pair-wise correlation matrix for all the buildings. Darker blocks along the main diagonal in the figure show the fact that buildings within each cluster are highly correlated together but not much to the rest. To analyze the clusters, we studied all the buildings and based on their actual context categorized them into 10 categories including: housing, auditorium, (outdoor) activity, sorority, fraternity, school, health, music, cinema and service. These categories are available to the left of each abbreviation in the figure. As can be seen, many of the buildings in the same category are clustered together. For example many of fraternities and all sororities are placed in cluster 1 (cluster IDs are available at the right side of the heatmap). We can also observe that 4 building in ''activity'' category and 7 ones in ''health'' category are clustered into clusters 5 and 8 respectively ("activity" category includes buildings with

different activity context including sports, religion, social and shopping). We can also see that all of the discussed correlated buildings in Figure 5-4 and Figure 5-5 are also clustered into the same clusters.

Multi-Aspect Analysis

Figure 5-7 shows some examples of highly correlated domains at specific locations. The maps at the first row reveal a non-trivial correlation between visitation of 'yahoo' at ANH housing complex and 'live' at KAT sorority. We can also observe partial correlation between this pattern and visitation of 'mozilla' and 'google' at ATO fraternity. The second row shows two other examples of multi-aspect correlations: i) visitation of 'youtube' at LUC (cinema) and 'live' at ASC (Communication & Journalism school); ii) visitation of 'usc' at ACO (sorority) and 'yahoo' at PKF (fraternity).

One point in multi-aspect analysis is the fact that inspecting many feature maps for all the combination of aspects (in our case 800 combinations of 40 domains and 20 locations) is rather difficult. This was actually one of our main motivations for designing the feature clustering technique. By employing this technique, we can easily cluster all the maps and then use the visual inspection of feature maps for detailed analysis. Figure 5-8 shows the top-left quarter of acquired clustered heatmap for multi-aspect analysis (80 clusters in total) (see [46] for the complete map).

Summary

In this study, novel modeling and analysis were conducted utilizing self-organizing maps and our proposed extensions to the technique for multi-aspect trend modeling based on web domains and locations. We showed that online activity and mobility are modeled with an organized map of trends, which can be effectively used to find correlations, anti-correlations and anomalies inside the mobile societies. This study

provides another type of realistic models for mobile Internet which can be effectively applied in design of different networking solutions.

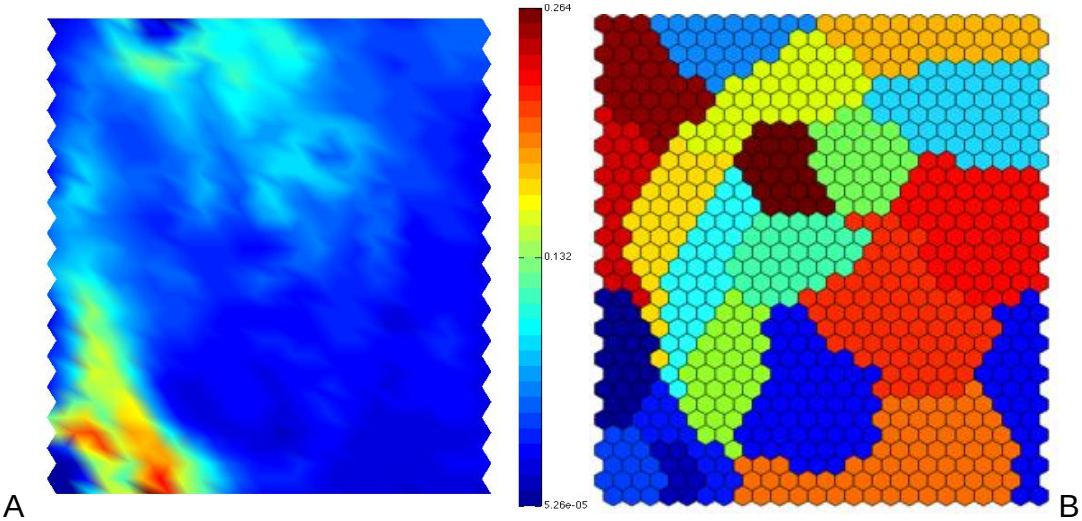


Figure 5-1. U-matrix and clustered SOM for WLAN Internet usage (for domains). A) U-matrix. B) Clustered SOM.

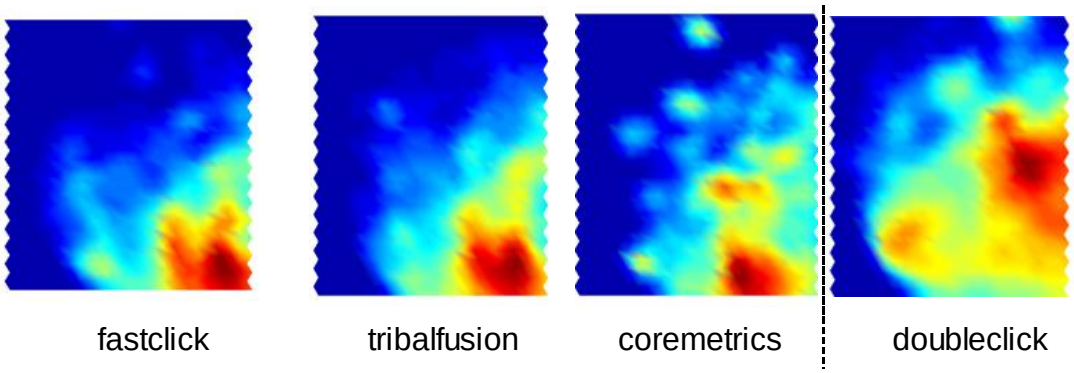


Figure 5-2. Feature maps for advertisement and marketing domains.

Table 5-1. Feature clustering result on web domains.

Cluster	Domain	Cluster	Domain
A	apple mac cnet washingtonpost itunes earthlink	B	microsoft windowsmedia microsoftoffice2007 mcafee hackerwatch quiettouch
C	google mozilla nih	D	live hotmail net hamachi
E	veoh secureserver	F	comcast fastwebnet
G	torrentbox rr	H	smartbro aster fastres opendns

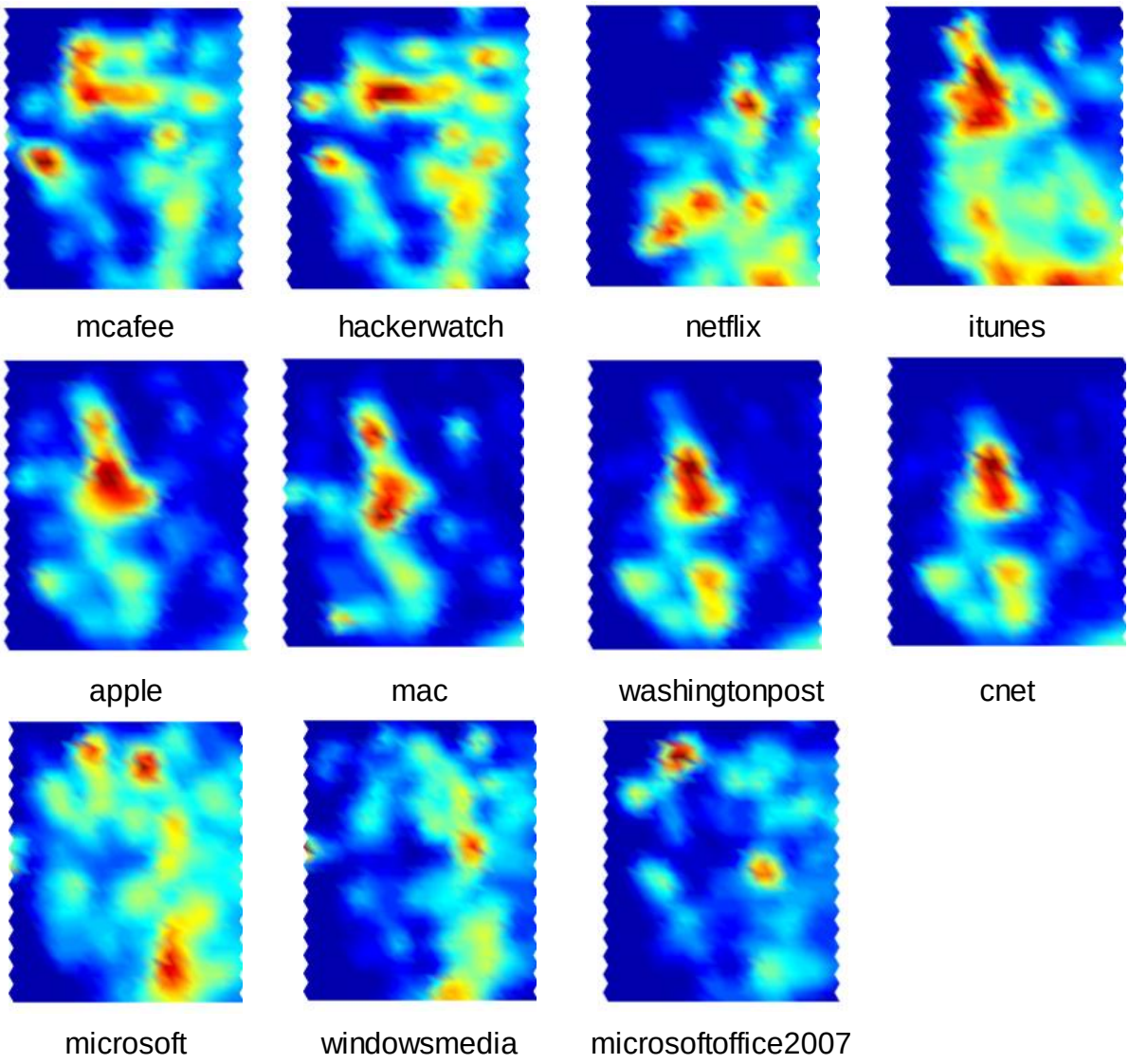


Figure 5-3. Feature maps for various types of domains.

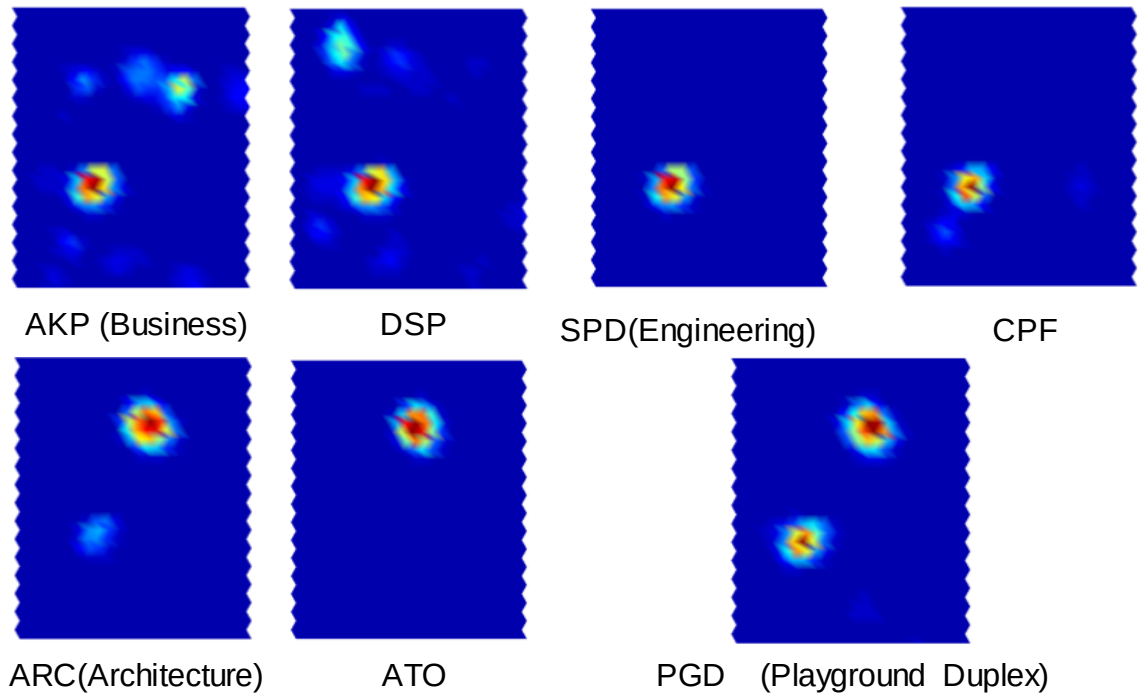


Figure 5-4. Feature maps for social & professional fraternities.

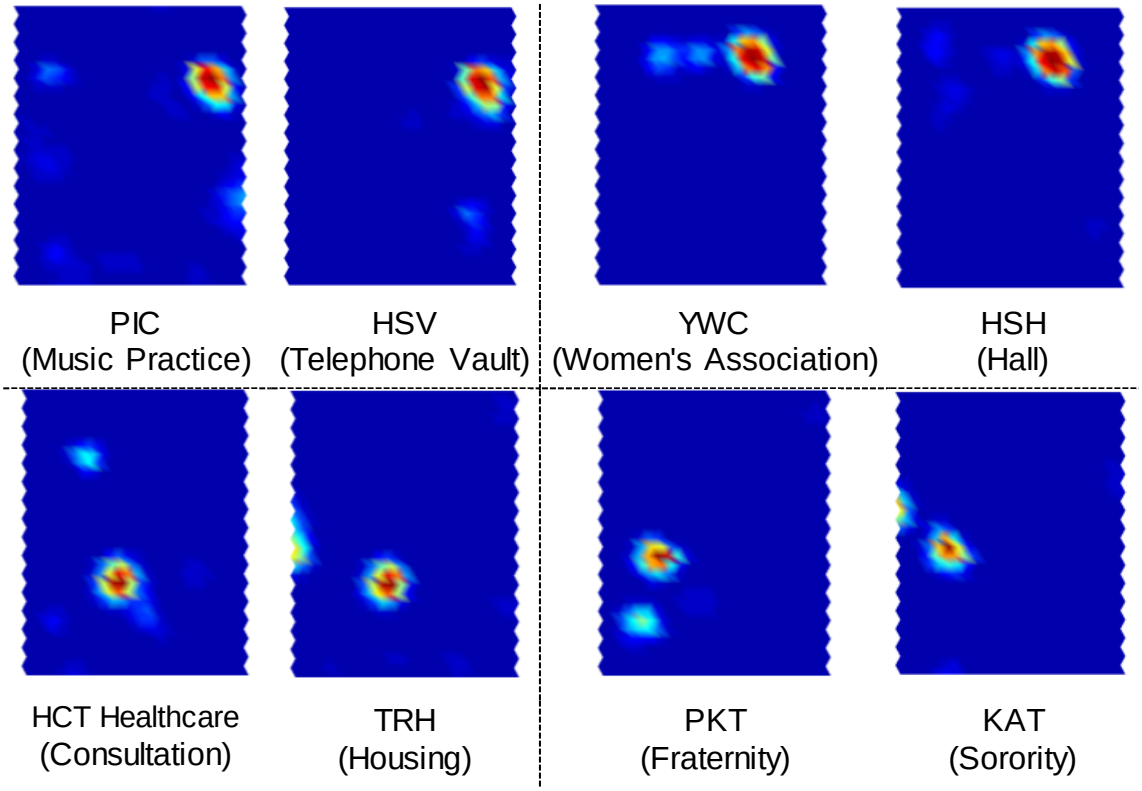


Figure 5-5. Feature maps for various types of locations.

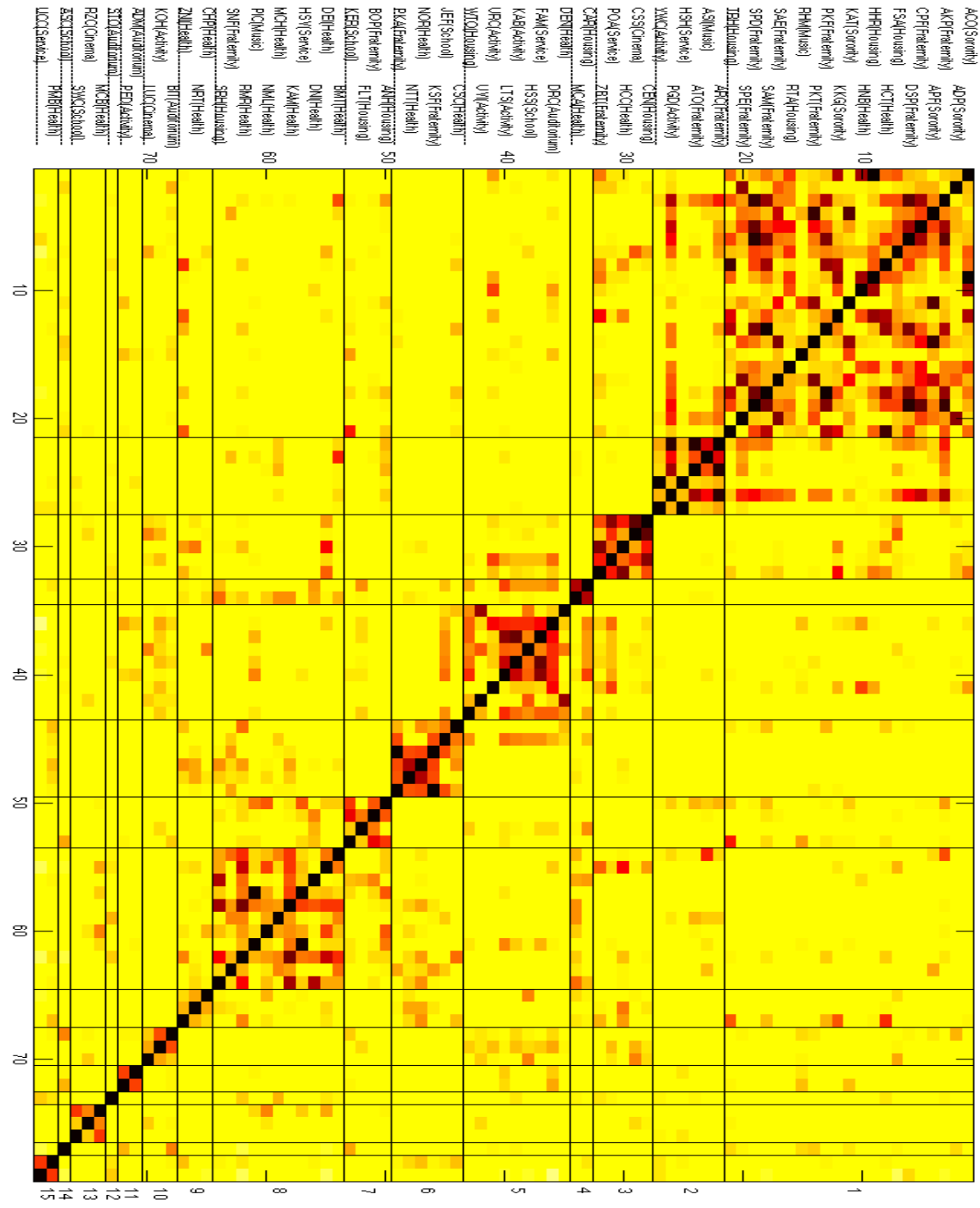


Figure 5-6. Clustered heatmap for all the buildings as features. Rows and columns represent building IDs and lines indicate cluster borders. Numbers at the right show cluster IDs and descriptions at the left include building abbreviation and category for each row. Darker colors show more correlation.

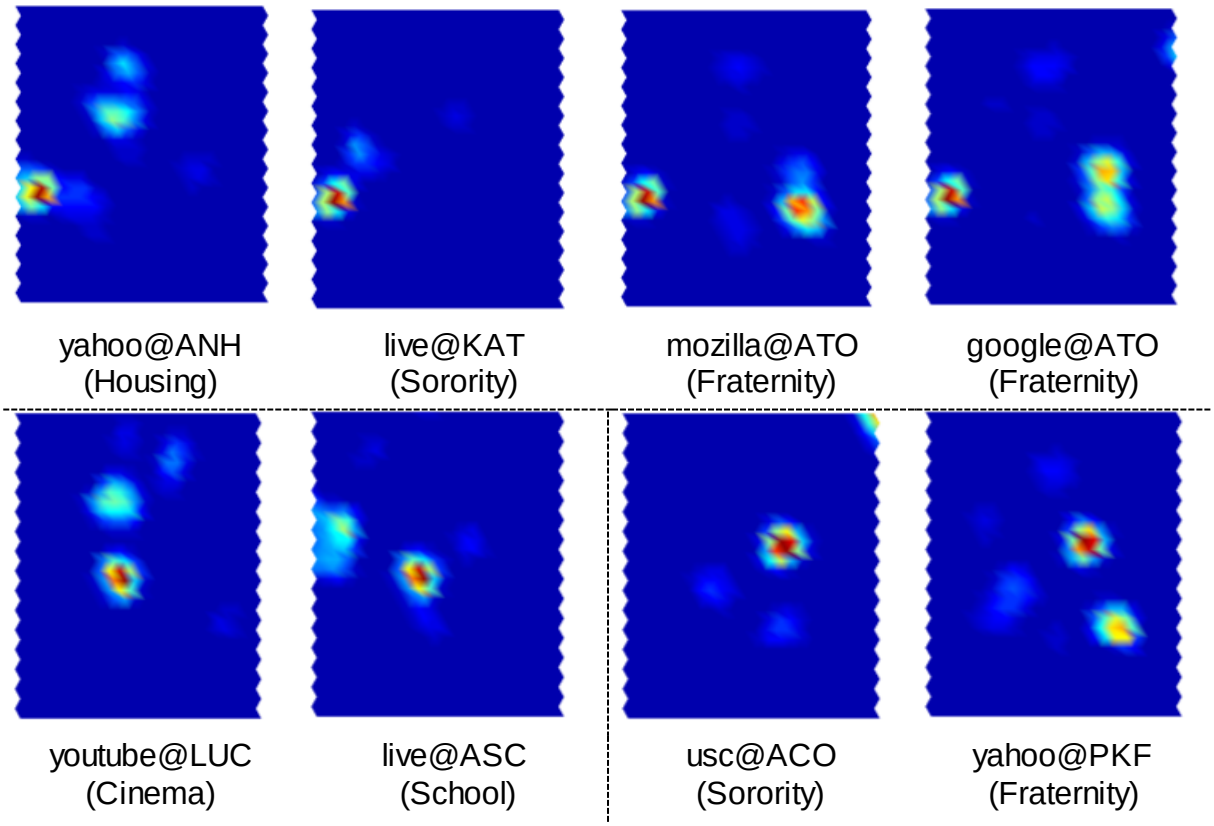


Figure 5-7. Multi-aspect feature maps for domain-location.

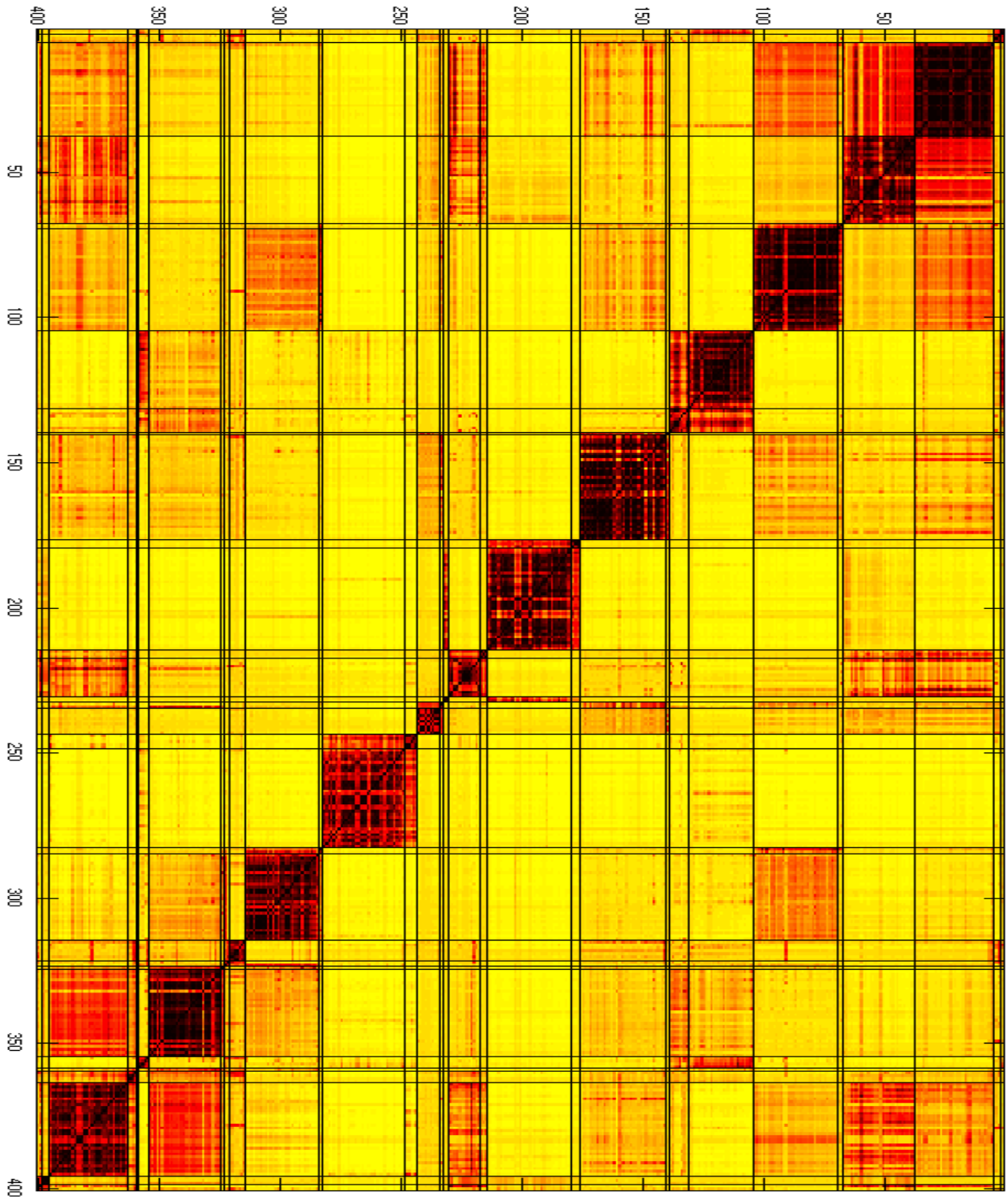


Figure 5-8. Clustered heatmap for multi-aspect feature analysis (zoomed on top-left quarter of the map.

CHAPTER 6
MODEL-BASED MULTIVARIATE SIMULATION OF MOBILE USERS 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 for simulation of multi-dimensional behaviors. Such modeling and simulation is imperative for design of novel behavior-aware networking services (e.g., for interest-aware casting). 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 features, 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 mobile networks.

In this chapter, we propose a new model-based simulation paradigm based on Gaussian Mixture Models (GMM) [3] for multivariate simulation of mobile users' behavior and interests [47]. 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.

This work has the following key contributions:

1. We propose an effective approach for multi-dimensional model-based simulation of mobile users activities extracted from one of the largest set of mobile network usage traces and show how Gausain mixture models can be applied to capture and preserve 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.

Model-Based Simulation Approach

Model-based simulation of mobile societies requires two main phases. The first phase is modeling of users' interests based on their web domain visitation patterns. The second phase includes generating simulated data based on the acquired parameters for the model and evaluating the quality of simulation output.

Data Modeling

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

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

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 that 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 that 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)).

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 that is essentially based on Pearson's chi-square test [4] 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 numbers of bins. However, for high dimensional 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-ranges 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 that 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 $E_i=1$ and $O_i=5$, 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:

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

Experimental Simulation Result

In our case study, we used our dataset for the University of Southern California (USC) in 2008 and applied our proposed method for simulating the real data samples. Figure 6-1 shows the clustering result on the real dataset. As can be seen, we can identify different behavioral groups based on users' interests. Figure 6-2 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.

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 bins for 2 to 15 sub-ranges along each dimension. Figure 6-3, Figure 6-4 A and Figure 6-4 B show the simulation accuracy for different numbers of domains and components while the number of sub-ranges along each dimension is set to 10, 15 and 5 respectively. As can be seen in Figure 6-3, generally speaking, as the number of components increases, 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% to around 60% 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 well shows the deficiency of uni-component simulation techniques for very high dimensional data. As can be seen in Figure 6-4, 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 6-5 and 6-6 show how the number of bins affects the accuracy metric. As can be seen in Figure 6-5, 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 6-6. 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 higher numbers of components lead to higher amounts.

Summary

In this study, we provided a systematic method for model-based simulation of web activities 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. This study is the first step toward realistic simulation of multi-dimensional users behavior and interest. 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.

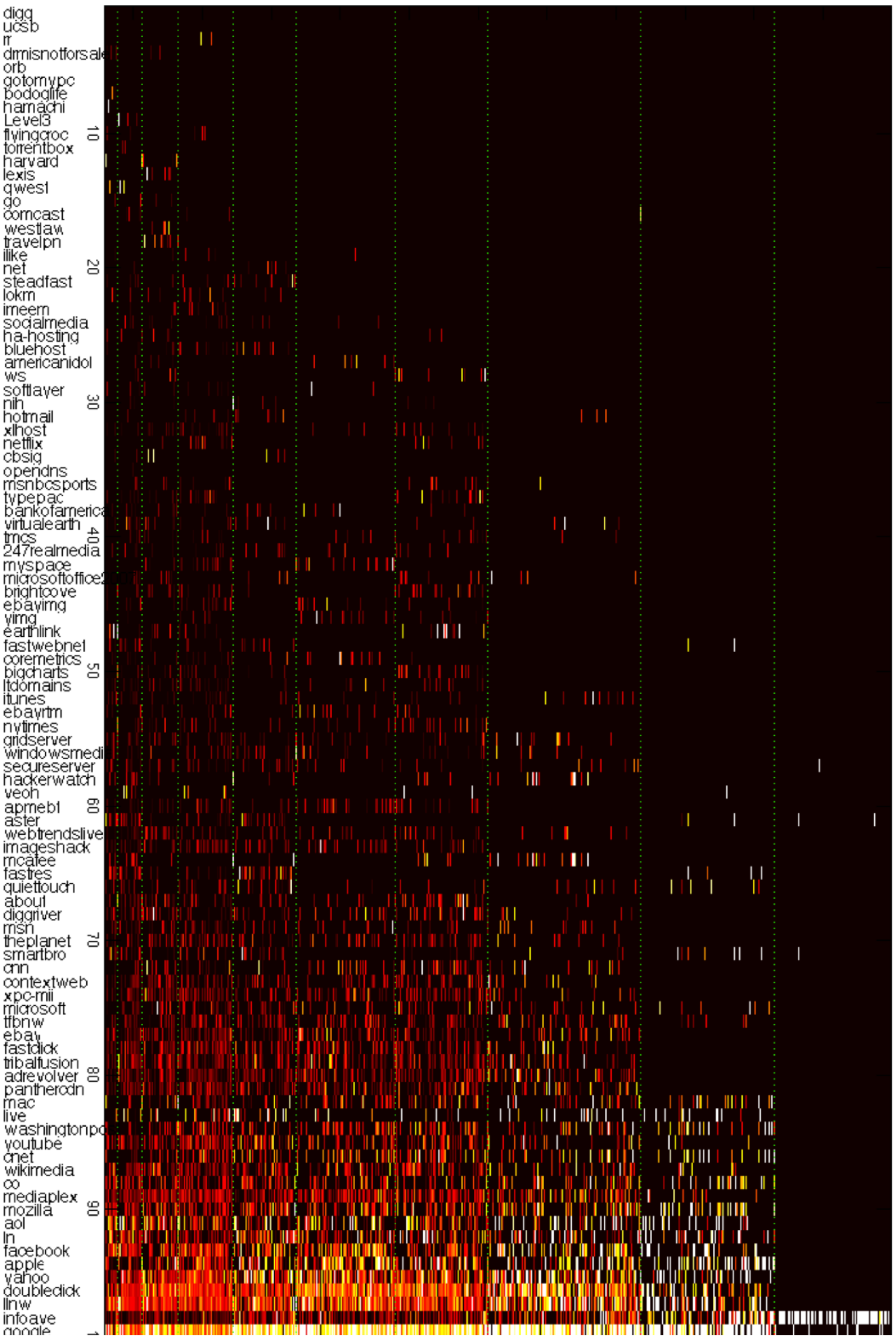
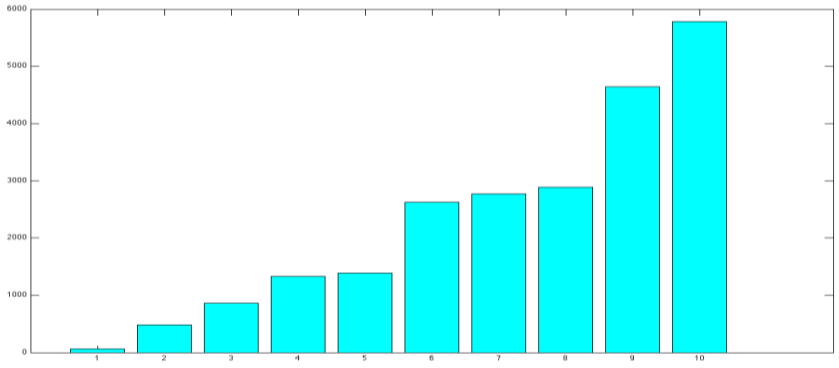
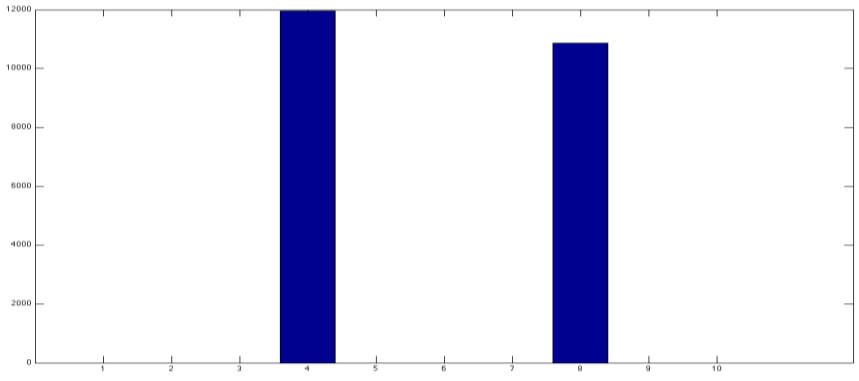


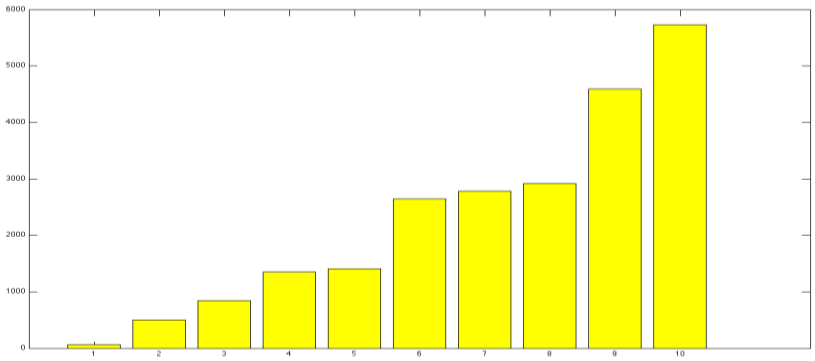
Figure 6-1. Behavioral clusters for the real samples. X-axis shows domain names and y-axis shows users and cluster IDs.



A



B



C

Figure 6-2. Distribution of real and simulated samples over behavioral clusters (sorted by the number of users). A) Real dataset. B) Simulated dataset using normal distributaion. C) Simulated dataset using our proposed techniqe (GMM)

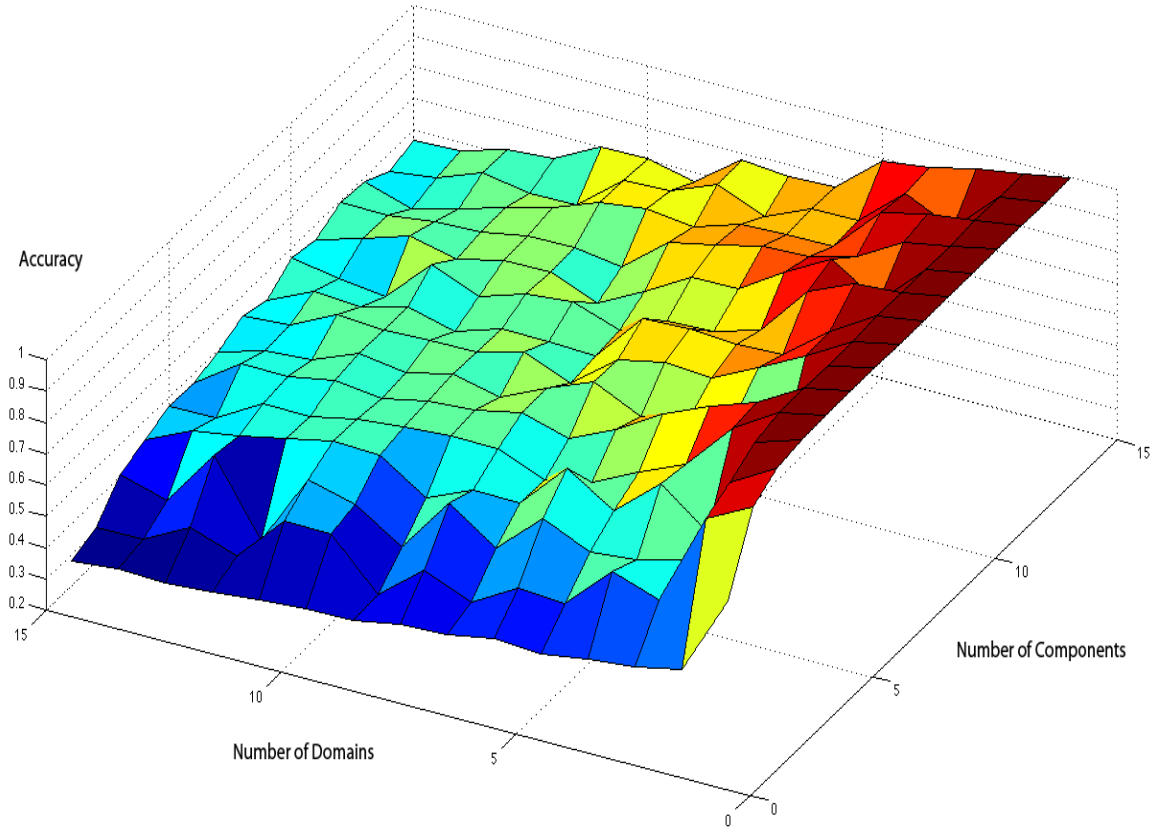
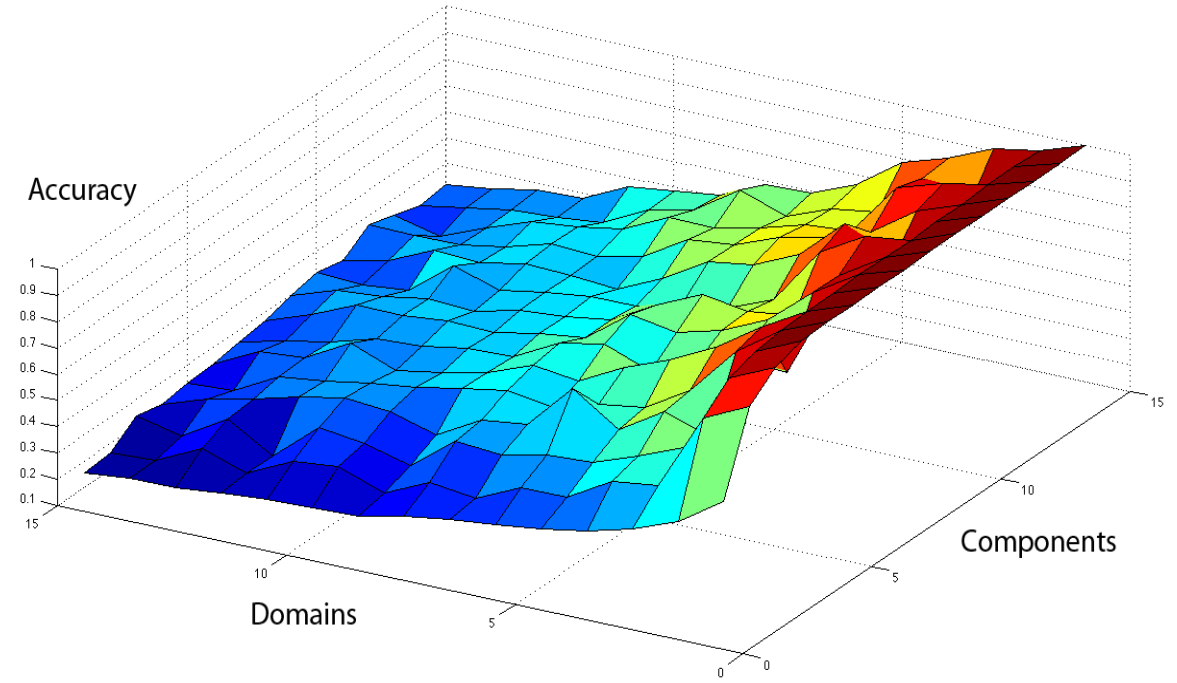
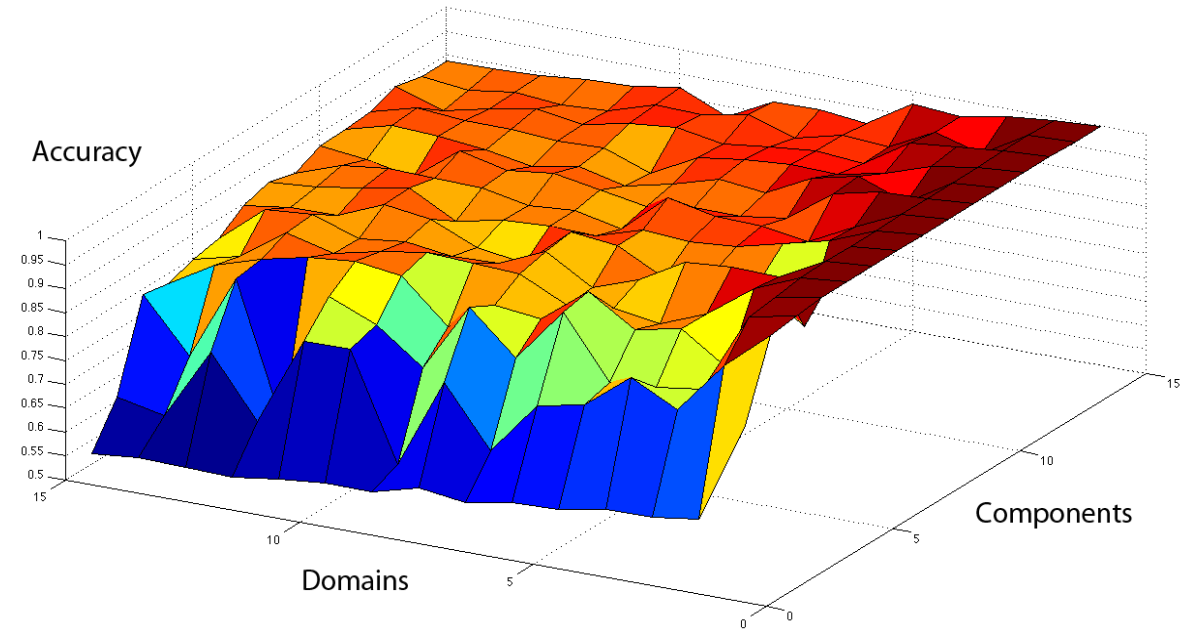


Figure 6-3. Simulation accuracy for different number of domains and components considering 10 sub-ranges along each dimension for creating the bins.



A



B

Figure 6-4. Simulation accuracy for different number of domains and components considering different sub-ranges along each dimension for creating the bins. A) 15 sub-ranges for bins. B) 5 sub-ranges for bins.

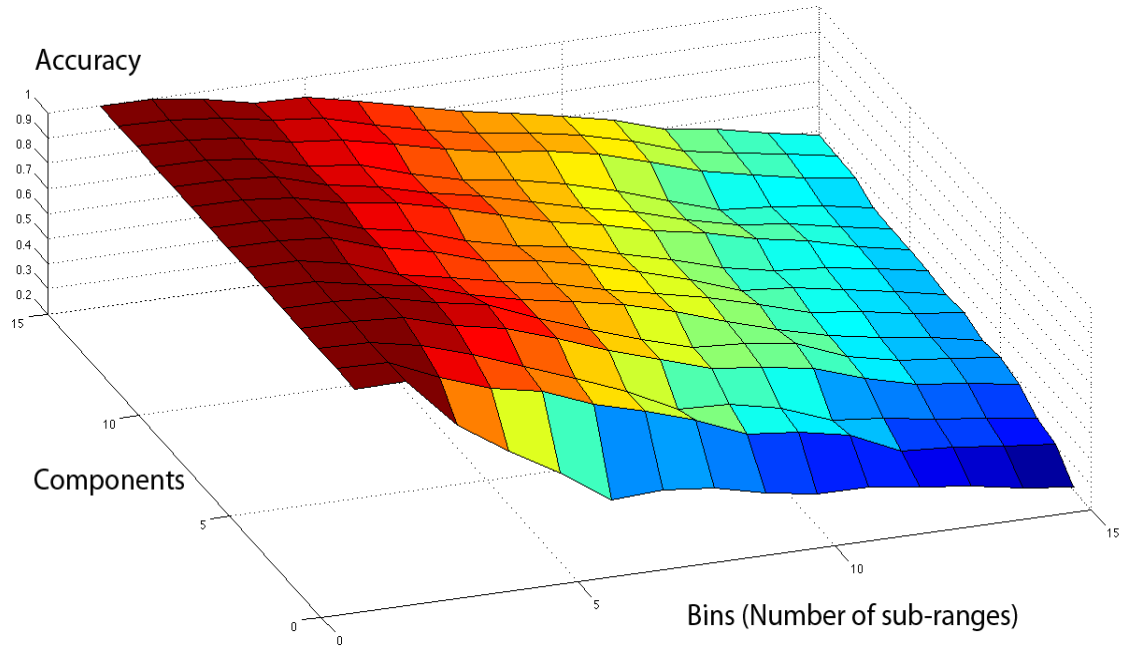


Figure 6-5. Simulation accuracy for different number of components and bins considering 10 domains.

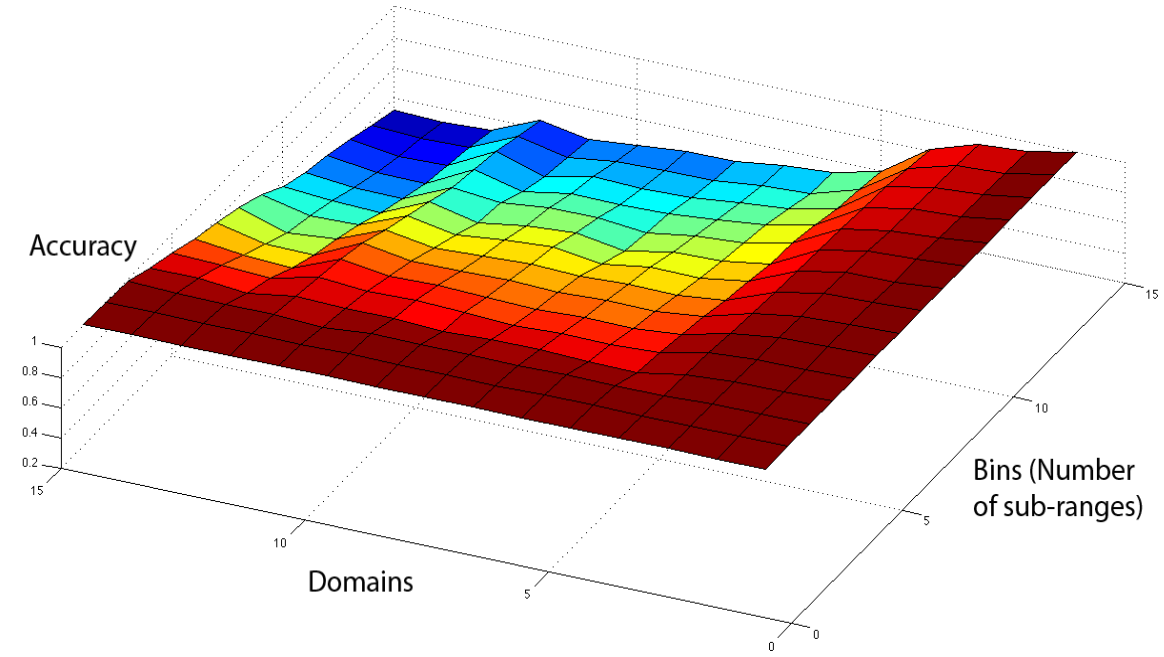


Figure 6-6. Simulation accuracy for different number of domains and bins considering 10 components.

CHAPTER 7
DOMAIN, LOCATION AND HUMAN-CENTERED TRAFFIC ANALYSIS AND MODELING OF MOBILE WIRELESS INTERNET

Mobile Internet traffic is going to experience a significant growth in the coming years. Different types of Internet-enabled mobile devices are getting more and more popularity and wireless Internet access infrastructures are growing faster than ever. The emergence of this radical change in availability of Internet raises a new need for modeling and analysis of mobile Internet characteristics. A traffic model in general is a model that can be used to regenerate the behavior of a real traffic stream. A major application of traffic models is in predicting the behavior of traffic as it passes through a network. The common approach toward traffic modeling is to find a generic model for the whole network. Although, such models provide good approximations for the wired Internet, but several studies have shown that they do not fit the dynamics of wireless networks. For example, [14] characterize the wireless traffic in different locations and shows that the dynamics of network follow a similar model but with different parameters. However, such models were generally based on small datasets of WLAN activities (e.g. 25000 flows a day), which are far from the full scale of dynamics in current networks (e.g., our dataset includes over 100 million flows per day). Moreover, most previous works have focused on individual access points, or short time periods but have not studied the characteristics of web domains, locations and human behavioral groups (subspaces) that can be beneficial to the realistic design of mobile network applications and services, e.g., caching of web objects at proxy servers.

In this chapter, we first present our domain and location based modeling approach including curve fitting and graph-based analysis parts. Our campus-wide case study shows that domain and buildings have distinct traffic characteristic that can also form

groups with distinct characteristics. Then, we introduce a two-tier modeling technique for human-centered modeling of mobile wireless Internet that includes behavioral modeling to form the behavioral subspaces, and behavior-aware traffic modeling to build the traffic models for the identified subspaces. A behavioral subspace represents all the interactions between a specific group of users and a specific group of domains; or; all the Internet interactions by a specific group of users at a specific group of locations. Our study shows that distinct behavioral subspaces with distinct characteristics exist within the mobile society. In our study, we investigate these characteristics by performing Internet traffic modeling for each of the domains, buildings or behavioral subspaces separately. For this purpose, we examine 9 different statistical distributions. In order to find the best fit in each case, we apply KS (Kolmogorov-Smirnov) test [5] as a measure for goodness of fit. To compare the traffic characteristics of domains or building together, we also apply Two-Sample KS test to either accept or reject the hypothesis of having the same traffic distribution. The experimental results show that our approach provides much more insight on the characteristics of the mobile wireless Internet and also improve the accuracy of traffic modeling.

This study has the following key contributions:

1. We provide domain and location based traffic modeling results for the top 100 active web domains and 68 different buildings (all active buildings) across the campus including 1800 users (all active users for the studied period). We present the results for 3 different days and explain how the results are interpreted. The studied dataset is one of the largest wireless mobile network traffic traces (including around 100 million records per day).
2. We provide a systematic method to discover similarities or differences between the traffic distributions of different domains or locations. We show how our graph-based method can be applied to identify groups of domains or locations with distinct traffic characteristics and reveal interesting facts about them.

3. We present a human-centered traffic modeling approach based on behavioral modeling using co-clustering technique. We present and analyze the modeling results for behavioral subspaces in terms of both accessed web domains and visited locations for different days.

4. We evaluate the accuracy of proposed approach using KS-test and weighted traffic intensity of domains, locations or behavioral subspaces. The evaluation results show that our approach can improve the accuracy of traffic modeling up to 62%.

Domain and Location Based Modeling

Traffic Modeling

In this section, we study the traffic behavior of mobile wireless Internet considering specific web domains and locations. The goal of this study is to find similarities or differences between the behavior of wireless Internet traffic for individual domains or locations, and the overall traffic of the wireless network. For this purpose, we find the best flow-level traffic model for each of the buildings or web domains. To achieve this goal, we first partition all the flows in terms of domains or buildings. Then, for each second during a day, we calculate the total number of flows for each domain or building. The output of this step for each domain or building represents its actual traffic distribution.

After obtaining the real distributions, we examine the dataset against 9 different statistical distributions to find the best curve fitted to the real distributions. The set of statistical distributions includes Weibull, Rayleigh, Poisson, Negative Binomial, Lognormal, Generalized Pareto, Generalized Extreme Value, Exponential and Gamma. We pick the best fit based on the well-known KS (Kolmogorov-Smirnov) test [5]. The KS test is a nonparametric test for the quality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution. The KS statistic quantifies a distance between the empirical distribution

function of the sample and the cumulative distribution function of the reference distribution. The KS-test has the advantage of making no assumption about the distribution of data. We repeat the curve fitting process for 3 different days in a week. Figure 7-1 shows the results for different domains and buildings. In addition to domain and location specific modeling, we also find the best fit for overall traffic during each day.

Domain-based modeling result

Figure 7-1 A shows the result of domain specific modeling for the top 100 active web domains during 3 different days. In this figure, the best fit for the traffic of each domain is presented using a letter representing one of the statistical distributions as follows: (L: Log-normal, W: Weibull, V: Generalized Extreme Value, P: Poisson, G: Gamma, R: Rayleigh, X: Exponential, V: Generalized Extreme Value, T: Generalized Pareto). Empty cells represent no traffic. In this figure the results for each day can be found in one of the rows. As can be observed in the figure, the traffic behavior of different domains follows different types of distributions. Table 7-1 shows the percentage of domains following a specific type of statistical distributions for different days. As can be seen in the figures and the table, we can find 4 categories of domains in terms of their best traffic distribution fit. On average, 25% of domains mostly follow Weibull, 23% mostly follow Lognormal, 21% follow Generalized Extreme Value and the rest follow other types of distribution. However, not all domains follow the same best fit during different days, which is of course not very surprising as the best generic traffic fits for the three days are also different. Our study shows that the best fit for traffic distribution of first and third days is Generalized Extreme Value while for the second day it is Generalized Pareto. However, the interesting point here is that for 43% of domains

we get the same best fit for different days which shows much more stability than the generic traffic modeling result. On the other hand, we also find a small number of domains (12%) showing completely different best fits for different days, which can be interpreted as kind of anomaly. These show that the best generic fit might not be the best model when considering specific domains.

Location-based modeling result

Figure 7-1 B and Table 7-1 show the traffic modeling results for different buildings during 3 different days. As shown, similar to domain specific modeling, location-based best traffic fits are also not the same as the overall best fits. We can again find 4 major categories of buildings. On average, 35% of buildings follow Weibull, 25% follow Lognormal, 18% follow Generalized Extreme Value and the rest follow other types of distributions. We can also observe that while the best traffic fit for the first day is Generalized Extreme Value, for the most of the buildings (88%) the best fit is one of other distributions at the same day. However, not very similar to the domain-based analysis, we can observe that only 15% of buildings follow the same distributions during different days. This reveals the fact that location-based traffic modeling is much more time-variant than the domain-based case. The reason for this might be the fact that for a specific building on a campus the schedule of classes might be in a way that very specific groups of users with specific traffic behavior go to that building at specific days.

Graph-Based Analysis

In the second part of domain and location based study, we investigate the similarities and differences between the traffic distributions of different domains or locations. While some domains or locations might follow the same type of statistical distribution, their models might follow different parameters. On the other hand, finding

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

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

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

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

In our study, we run the test at significant level of 5% for each pair of domains or buildings. The results form two matrices including a 100*100 matrix for domains and another 68*68 for buildings showing if two domains or buildings follow the same distribution or not. In order to analyze the results, each of the matrices can be interpreted and visualized as traffic similarity graphs. In such graphs nodes represents domains or buildings with an edge between nodes if the corresponding domains or

buildings have similar traffic distributions. Figure 7-2 and 7-3 show the resulting graphs for domains and locations for the middle day after running the algorithm presented in [48] for finding the modularity classes within the graphs and applying Fruchterman-Reingold algorithm [49] to form the graph layouts. In these graphs, the corresponding domain or building for each of the nodes can be identified from the node ID (IDs are the same as Figure 7-1). Modules are shown by different colors in the figure. The size of each node represents its degree in the graph that shows how unique the traffic distribution of the node is when compared to the other nodes (low degree is interpreted as uniqueness).

Domain-based analysis

As can be seen in Figure 7-2 for domain-based analysis, there are 21 domains with unique traffic characteristics. As can be observed, all top 5 active domains including 'usc', 'google', 'facebook' and 'aol' have unique characteristic. Among the rest, we can also find some other popular domains, e.g., 'apple', 'mozilla', 'veoh' and 'netflix'. As can be seen, in general, high traffic domains show more uniqueness in terms of their traffic characteristics.

The rest of domains form 12 groups with distinct traffic distributions. Half of the groups have a size of less than 5 and the size of rest is up to 16. Studying different groups reveals many interesting facts. For example, while video sharing domains like 'netflix' and 'veoh' show unique characteristics, we can see that 'youtube' traffic distribution is similar to 5 other domains including 'cnn', 'msnbcsport' and 'microsoft' (the group at top-left). This shows that even within a class of applications, e.g., online video sharing, we can find different classes of traffic characteristics, some with unique characteristics and some with more popular ones. Another interesting observation in

this case is that ‘msnbcsport’ is partially similar to the rest of 3 domains in terms of its provided services; both ‘msnbcsport’ and ‘cnn’ provide news, both ‘msnbcsport’ and ‘youtube’ provide online video, and both ‘msnbcsport’ and ‘microsoft’ are related to Microsoft. This might show that, in some cases, when different types of applications are combined together (as in ‘msnbcsport’) the traffic characteristics of the result is a combination of different application characteristics.

Another interesting finding is that many of domains related to high-speed Internet and phone providers like ‘comcast’, ‘charter’ and ‘qwest’ have similar traffic distributions (the group in middle-right). Interestingly, we can also find ‘shoutcast’ in this group that is not in that category but provides a similar type of service, i.e., Internet radio stations. Comparing Figure 7-1 and Figure 7-2 shows that the traffic of all domains in this group follows Rayleigh distribution during all days, which is also interesting.

Location-based analysis

Figure 7-3 presents the resulting graph for the location-based analysis showing three major groups of buildings with distinct traffic characteristics. In the center of this graph we see the lowest traffic buildings showing the most similarity to all the other buildings (building #36 is shown empty in Figure 7-1 as it is active for less than a threshold, i.e., 5 seconds).

By looking at the building categories (which can be find from the building IDs and category labels in Figure 7-1) we can again discover different interesting findings. For example, we can observe that more that 70% of buildings in Music, Cinema and Auditorium categories are in the same group (the group of nodes in right-bottom). We can also see most of fraternities (9 ones) are in the same group with similar traffic characteristics (the big one at the left). However, we can see that the buildings in Health

category are almost equally divided between the three groups. The interesting results of graph-based analysis suggest the idea of group-based modeling instead of individual analysis. The next part of study is inspired by this idea plus the idea of taking humans in to account too.

Human-Centered Modeling

While it is not impossible to conduct user specific traffic modeling for a small group of users, it is almost not practical and helpful to follow such an approach for a large community. First of all it takes a lot of time and resources to do the modeling for individual users and second, such models are not very useful to be applied by different applications and services (suppose we find the best traffic fit for 30,000 users, then the question is how to deal with so many models and their parameters). Of course, this problem is also the case if we have many locations or web domains. A natural solution to this problem is to consider classes of users instead of individual users. We can also consider classes of domains or locations as well. In order to form such classes of users we can think of different types of criteria. In this work, we create clusters of users in terms of the amount of exchanged traffic between them and different web domains. We also perform the clustering based on the exchanged traffic by users at different buildings. While we can do the modeling just based on the user classes, a good idea to reduce the required resources for modeling task and also the number of modeling parameters is to cluster domains or buildings at the same time too. In this way, we can identify the behavioral subspaces for each of the user classes and perform the traffic modeling in terms of the acquired subspaces instead of individual users, domains or locations. In this way, no matter how many users, locations or domains we deal with, we can always find a reasonable high granular traffic model for the network based on our

needs and resources. In this section, we introduce a novel two tier modeling approach which includes two steps of behavioral modeling, in order to discover behavioral subspaces inside mobile community, and behavior-aware traffic modeling, in order to provide the best possible traffic model considering the acquired subspaces.

Human Behavioral Modeling

The first step in human-centered traffic modeling is to provide a behavioral model for the mobile community. The main goal of behavioral modeling is to form the behavioral subspaces that represent distinct characteristics of human behavior in terms of their interactions with different domains or their Internet access pattern at different locations. For this purpose, we apply the information-theoretic co-clustering technique, explained in Chapter 4, on the traffic data. Using this modeling technique, we can identify distinct clusters of users-domains as well as users-locations and then partition the behavioral space into subspaces using the acquired clusters. Each of the behavioral subspaces formed by this approach represents a distinct class of behavior in a distinct class of users.

The input data for the co-clustering algorithm is a users traffic matrix, which represents the number of exchanged flows by users. In this study, we create two different matrices representing: i) the total number of flows exchanged between different users and domains and ii) the total number of flows for each user at different buildings. The experiment is performed for 1800 users, 100 domains and 68 buildings. The number of clusters in the study is set to 10 as it shows more distinct subspaces.

Behavior-Aware Traffic Modeling

After modeling and partitioning the behavioral space into subspaces, we partition the traffic in terms of the acquired subspaces which means that for each one we take all

the flows for the corresponding users and domains or the corresponding users and locations (depending on how we partitioned the behavioral space). For each subspace, we extract the real traffic distribution for 3 separate days. Then, we find the best fit for the traffic distribution of each subspace. For this purpose, as for domain and location based modeling, we examine the actual data against 9 different distributions and choose the best fit based on the KS-test.

Figure 7-4 and Figure 7-5 shows the modeling result for all the behavioral subspaces based on users-domains and users-buildings. In the figures, each row shows the best fits for different behavioral subspaces of a user class (10 user class). Each column in the figures shows the best fits for a group of domains or locations across different classes of users. As shown, in any of the cases there is no generic best fit for all the subspaces. For users-domains case, on average, Log-normal, Weibull and General Extreme Value are the best fits for 27%, 25% and 24% of the subspaces respectively. For users-locations case, on average, Log-normal, Weibull and General Extreme Value are the best fits for 26%, 26% and 20% of the none-empty sub-spaces respectively. In this case, 30% of the subspaces have no traffic (Table 7-2). Therefore, we can see again that the behavioral subspaces in terms of domains or locations form 4 major groups, three of which are best modeled using Log-normal, Weibull and Generalized extreme value. By looking at Table 7-2, we can also find that this trend is almost stable during 3 different days, although those days follow different generic traffic models as mentioned before.

By comparing Figure 7-4 B with Figures 7-4 A and Figure 7-4 C, we can also see that 45% and 54% of the subspaces in terms of users' domain visitations shows the

same best fit for the day before and day after the middle day respectively (note that the subspaces acquired based on the traffic for the middle day to analyze the stability of the subspace characteristics as well). As for location-based subspaces, we can see 40% and 32% of the subspaces show the same best fit for the day before and day after the middle day respectively. This shows that considering humans as well as domains or locations leads to more stable traffic models.

Another interesting observation in Figure 7-4 B is the fact that the best traffic model is majorly driven by the domain groups not the user classes (several columns have the same best fit for most of user classes). From this point of view, we can classify domain groups to three major categories. First category includes domain groups #4, #6, and #9 for which 70% or more of user classes follow the same type of best fits. Second category includes #3, #5, #7, #10 and the third one includes #1, #2 and #8 domain groups. The above amount is 50% or more for the second category and less than 50% for the third one. However, as for location-based behavioral groups, we cannot find such a trend for the building groups. The only noticeable observation is that around half of the none-empty user classes for building groups #5 and #6 follow the same best fits. Aside from that, we do not see a major dominance by the building groups over user classes to drive the traffic characteristics of the behavioral subspaces similar to what we observe in the domain-based case.

Accuracy Analysis

In this section, we evaluate the accuracy of proposed approach using the KS test. For this purpose, we compare the proposed approach with general modeling and two other approaches we refer to as ‘Adjusted’ and ‘Simplified’. These two approaches are somewhere between the to extreme of generic models and highly customized models in

our proposed approach. While in generic modeling, we apply the same acquired model for any domain, location or subspaces; in the adjusted approach we customize the parameters of the generic best fit for different domains, locations or subspaces. So, this approach is more similar to the generic approach. On the other hand, the simplified approach is more similar to the original approach with a little difference. In the simplified approach instead of choosing different distributions for different domains, locations or subspaces, we take one distribution but with customized parameters for them. In this approach we pick the best distribution with customized parameters. In order to compare different approaches, for each case, we calculate a weighted average of KS values for the estimated and the actual distributions for different domains, locations, or subspaces based on their traffic density. Table 7-3 shows the result of the evaluation.

As can be seen, for the domain and location based modeling approaches the average amounts of KS values are significantly reduced. The amounts of reduction are around 45% and 60% respectively. For the human-centered cases, we can see significant improvements as well. The amounts of reduction for user-domain and user-location cases are 62% and 54% respectively. As can be seen in the table, using the simplified approach, we can also get significant improvement close to that of the original approach. The advantage of this approach is the fact that the required information for applying the traffic model (which can be interpreted as the model complexity) is less, as the type of distribution is not required to be kept for each of the domains, locations or subspaces and only keeping the parameters for each one suffices. However, we lose some accuracy in this approach comparing to the original one. Thus, the best approach would depend on our requirements: if we are looking for more accuracy the original

approach gives us the best result, but if we are also concern about the complexity, the simplified version is preferred. The results for the adjusted approach shows that even if we adjust the parameters for the generic best fit, we would not get the best result comparing to the simplified and original approach.

Comparing different modeling approaches, we can see that the user-location based modeling shows the best accuracy (with the smallest average KS value of 0.0811). In addition, this approach provides another advantage in terms of complexity (as defined before). As 43% of the subspaces in this case are empty, we just need to keep the distribution parameters for 57 subspaces, which is less than the total number of domains (100 ones) and the total number of buildings (68 ones).

Applications

Domain, location and human-centered traffic models based on behavioral subspaces may be applied in different types of applications, e.g., caching of web objects at proxy servers. They also provide the opportunity of user-centered design for networking services and protocols. In a large network, it is not practical and of course affordable to consider characteristics of individual users and on the other hand generic models simply ignores such properties. By capturing and modeling the behavioral subspaces, it becomes possible to effectively consider the characteristics of classes of users in different behavioral subspaces.

Traffic modeling based on behavioral subspaces provides flexibility in modeling and analysis based on our requirements too. Subspace modeling approach provides an adjustable tool for zooming in and out on the dynamics of mobile community. Instead of a generic or individual traffic models, by taking this approach, we can choose an appropriate number of subspaces based on dynamics of mobile network, required

accuracy, desired number of parameters and available processing resources for building the model. More subspaces capture more details but require more parameters and processing resources to build and utilize the model.

Summary

In this study, we provided a systematic method for domain, location and human-centered modeling of mobile wireless Internet. For this purpose, we applied different techniques including different flavors of KS test as a measure of similarity, information-theoretic co-clustering as a tool for inferring the human behavioral sub-spaces and different graph-based techniques to analyze the traffic characteristics. We have shown that generic traffic models are not the best choice for mobile wireless Internet and taking domain, location and human behavioral subspaces into account (in terms of web domain accesses or location visitations) provides more realistic and accurate models. The details of our study enable the parameterization of new and realistic traffic models for future mobile Internet with potential applications in several areas of networking including web caching, interest-based casting, simulation and evaluation of protocols.

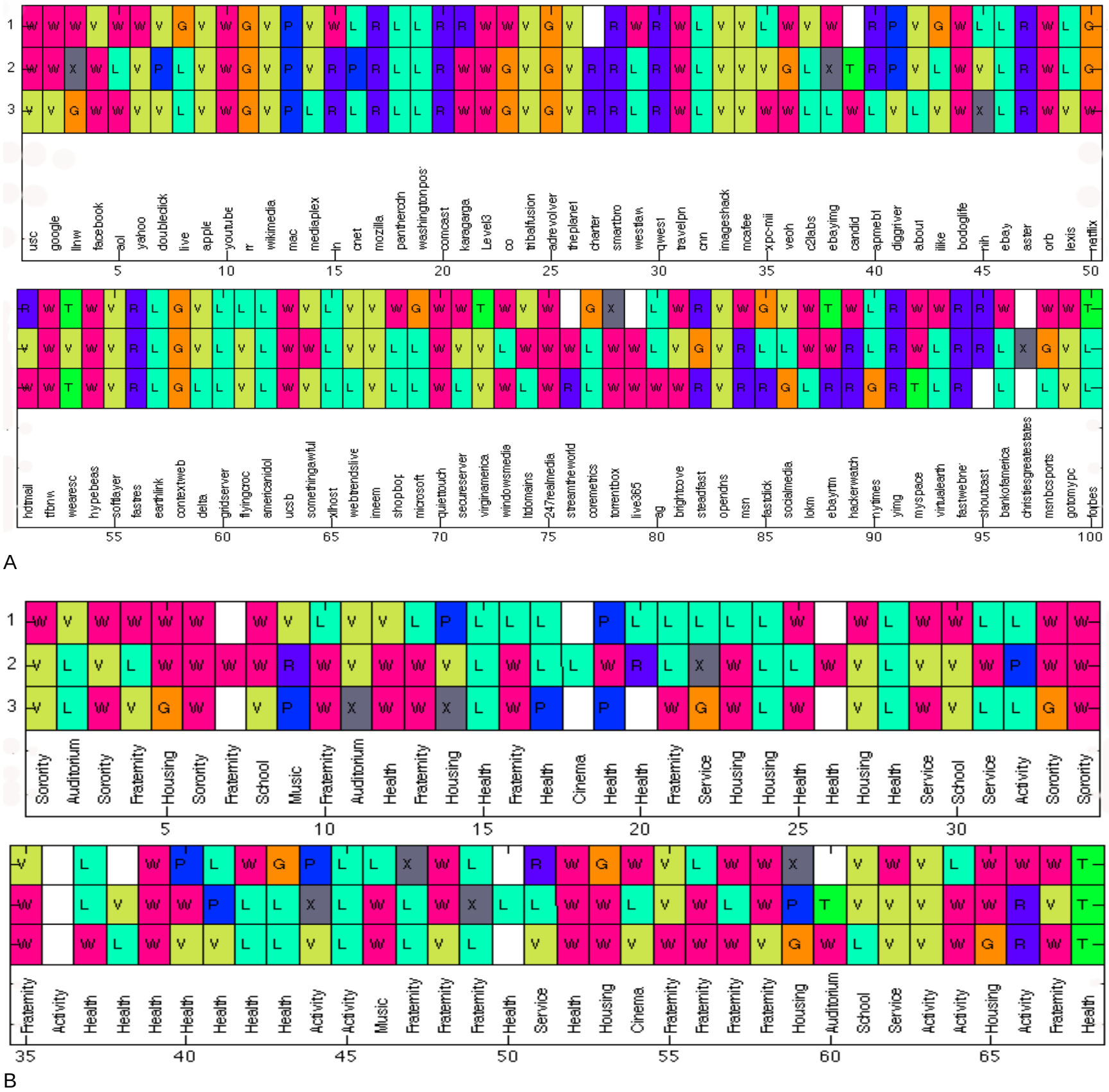


Figure 7-1. Best fits for traffic distributions of mobile wireless network in different campus buildings and for different visited domains. Each row shows the best fits for different domains or buildings during 3 days. (L: Log-normal, W: Weibull, V: Generalized Extreme Value, P: Poisson, G: Gamma, R: Rayleigh, X: Exponential, V: Generalized Extreme Value, T: Generalized Pareto, Empty cells represent no traffic, Each color represents a unique distribution). A) Best fits for top 100 active web domains. Column labels show domain names and IDs. Domains with higher IDs have lower traffic. B) Best fits for different buildings. Column labels show building types and IDs.

Table 7-1. Best fits for mobile Internet traffic for domain and location based modeling.

Rank	Best fit	1st day	2nd day	3rd day
		Percentage of domains		
1	Weibull	32	22	23
2	Lognormal	15	25	28
3	GEV	19	23	20
4	Rayleigh	13	14	16
5	Gamma	9	8	7
		Percentage of buildings		
1	Weibull	34	37	34
2	Lognormal	29	26	19
3	GEV	12	19	20
4	Empty	10	1	9
5	Poisson	6	4	4

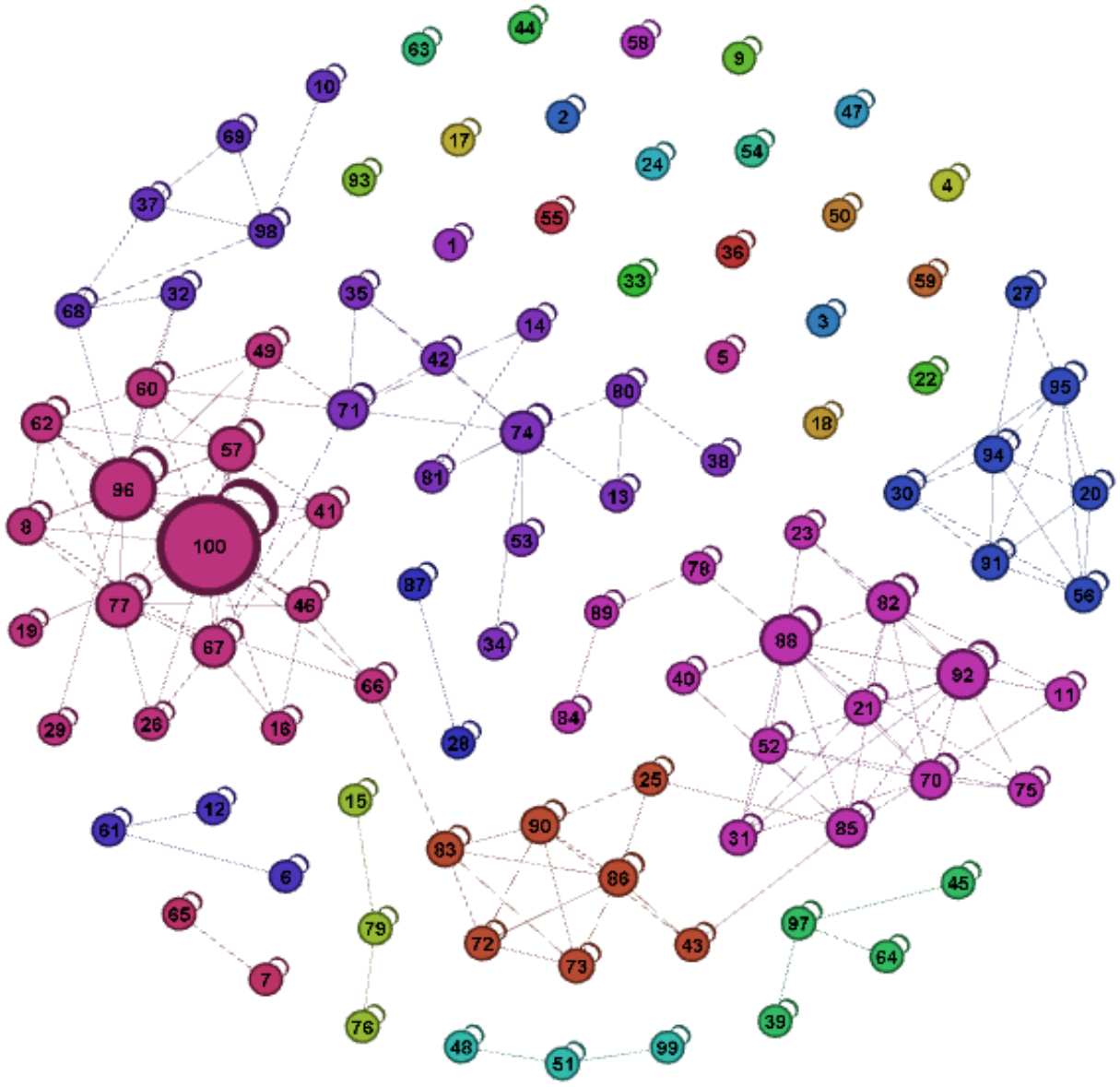


Figure 7-2. Traffic similarity graph for domains (for the middle day). Nodes represent domains and show their IDs (domain names can be find from Figure 7-1). Colors show different detected modules within the graph. Size of each node shows its degree in the graph.

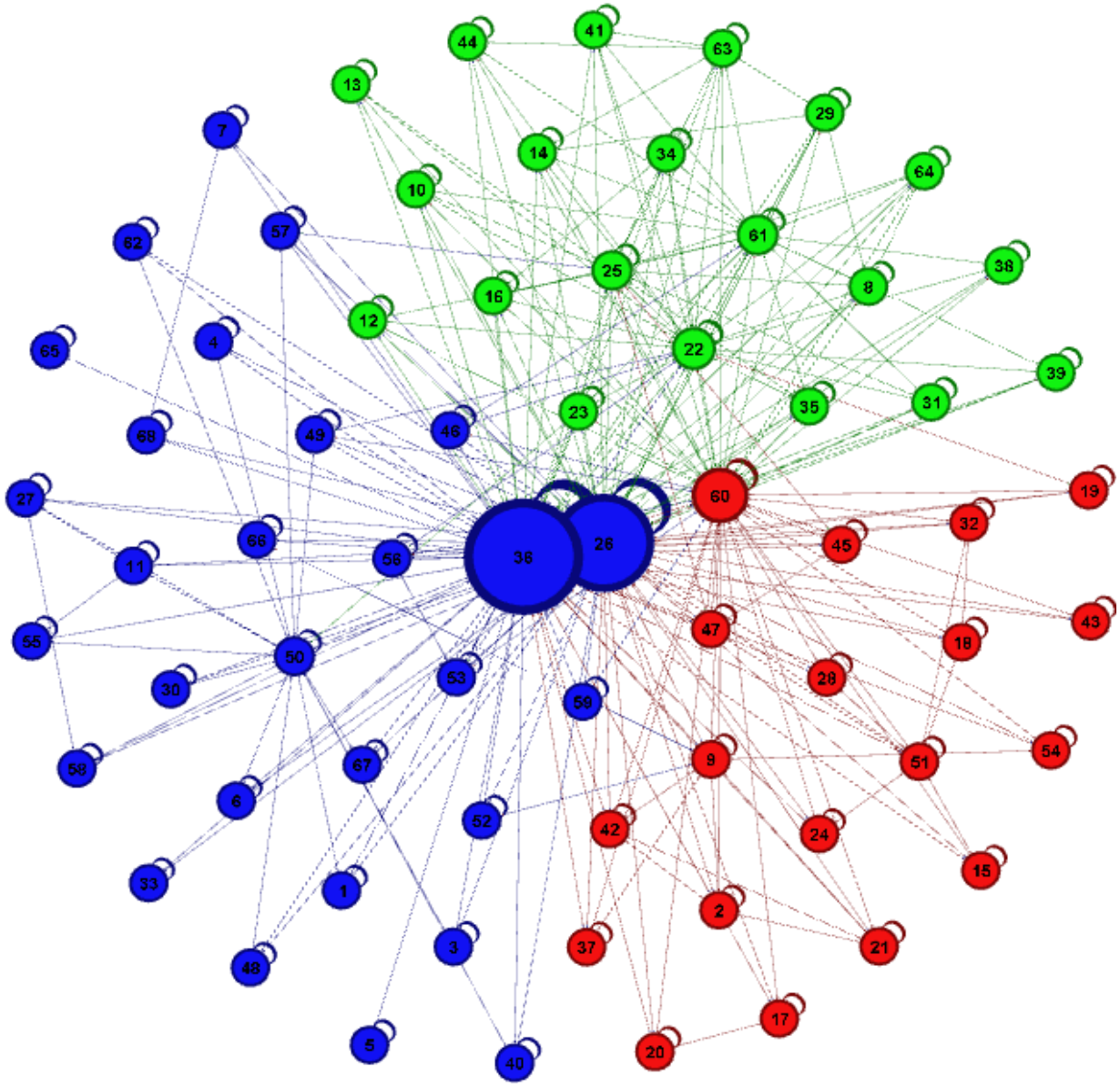
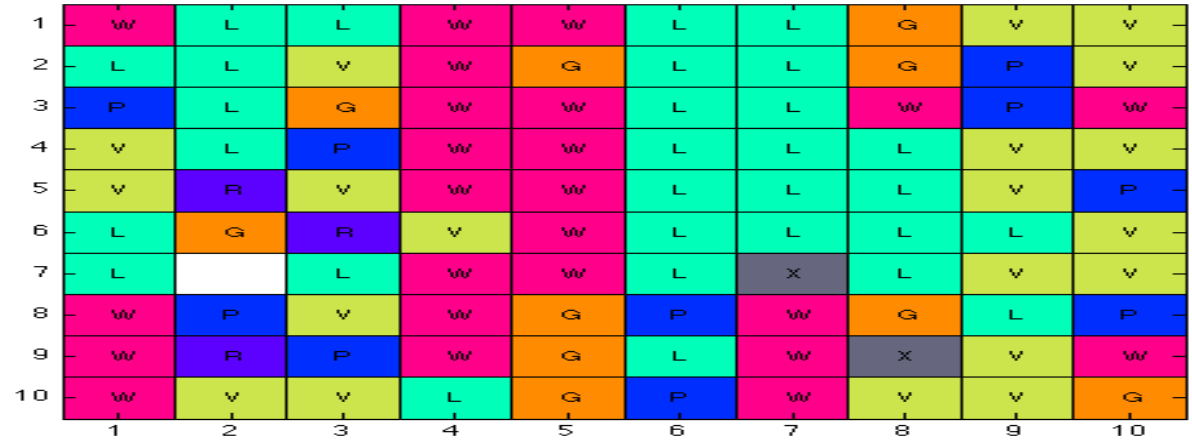
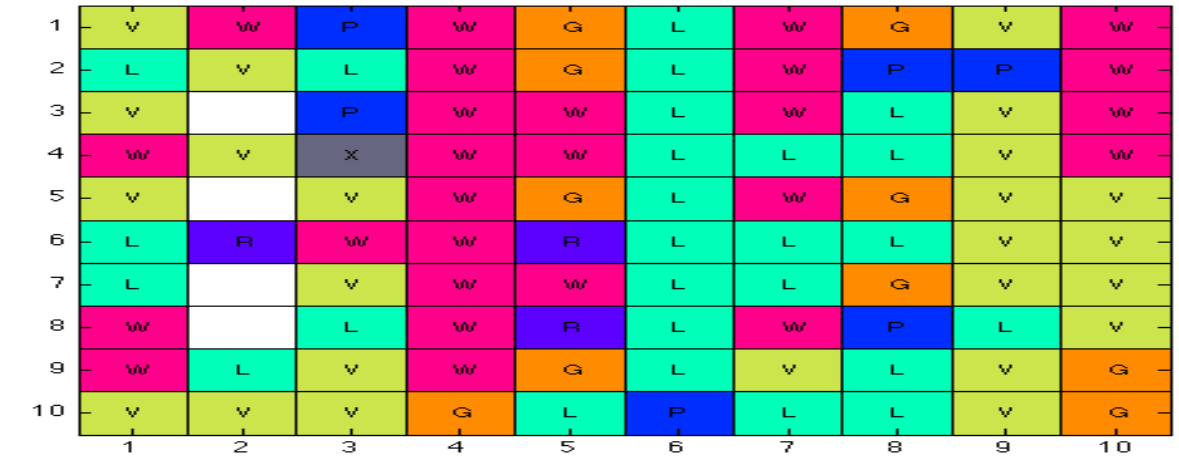


Figure 7-3. Traffic similarity graph for locations (for the middle day). Nodes represent buildings and show their IDs (Building types can be find from Figure 7-1). Colors show different detected modules within the graph. Size of each node shows its degree in the graph.



A

B



C

Figure 7-4. Best traffic fits for user-domain subspaces for three days. X and y axes show domain and user cluster IDs respectively. (L: Log-normal, G: Gamma, P: Poisson, V: Generalized Extreme Value, T: Generalized Pareto, W: Weibull, Empty cells represent no traffic). A) First day. B) Second day. C) Third day.

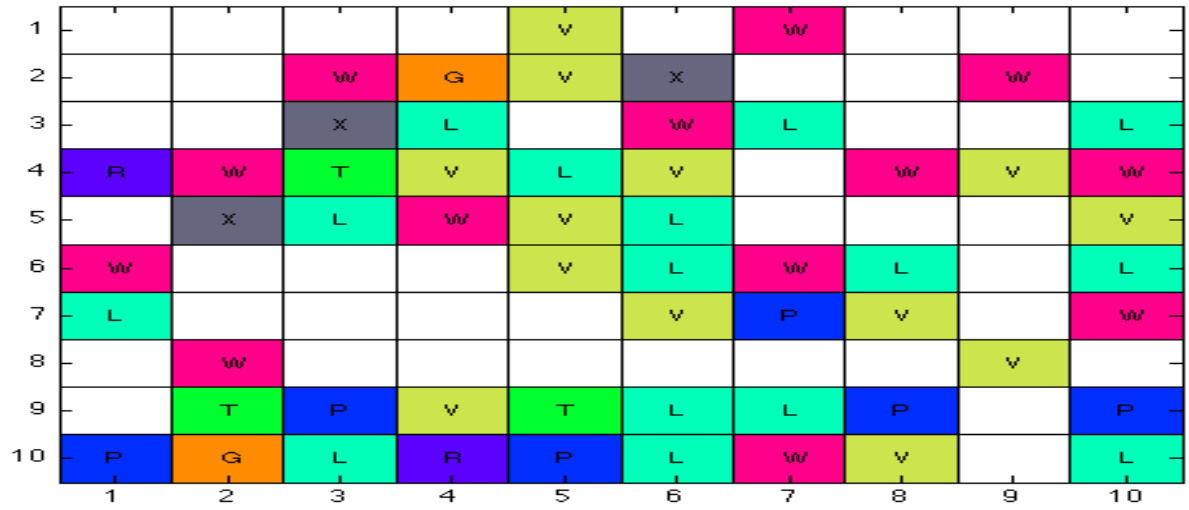


Figure 7-5. Best traffic fits for user-location subspaces for the second day. X and y axes show building and user cluster IDs respectively. (L: Log-normal, G: Gamma, P: Poisson, V: Generalized Extreme Value, T: Generalized Pareto, W: Weibull, Empty cells represent no traffic).

Table 7-2. Best fits for mobile Internet traffic for human-centered modeling.

Rank	Best fit	1st day	2nd day	3rd day
		Percentage of subspaces (domain-based)		
1	Lognormal	30	24	26
2	Weibull	24	24	26
3	GEV	20	28	24
4	Gamma	10	8	10
5	Poisson	10	10	6
		Percentage of subspaces (location-based)		
1	Empty	25	43	22
2	Lognormal	18	15	22
3	Weibull	21	13	19
4	GEV	14	13	15
5	Poisson	9	6	7

Table 7-3. Comparison of different traffic modeling approaches based on the KS test.

Approach	Domain based	Location based	User-domain based	User-location based
Generic	0.5643	0.7427	0.7765	0.6187
Adjusted	0.1517	0.1868	0.2256	0.1067
Simplified	0.1192 (Weibull)	0.1438 (Weibull)	0.1652 (Weibull)	0.0838 (Weibull)
Original	0.1159	0.1402	0.1565	0.0811

CHAPTER 8
ICAST: INTEREST-AWARE CASTING IN MOBILE NETWORKS

The fast growing trend toward mobile Internet access creates a tight coupling between users and mobile networks where various characteristics of user online activities, mobility and interests can be captured and applied to provide novel solutions to the networking problems. Information casting is one of the major services that can be re-designed based on the acquired user characteristics via the mobile network. In a mobile network where human behavior and interests can be inferred via their online activities and mobility, a new paradigm can be developed for behavior-aware casting of information. In such a paradigm messages are delivered based on users behavior and interests not their networking device addresses. However, current Internet communication abstractions (i.e., unicast, multicast) do not lend themselves readily to behavior-aware mobile networking. The nature of such communication paradigms is based on network characteristics not humans. On the other hand, today's interest-based services rely on elaborate information and settings explicitly provided by users, hindering the effective participation of classes of users including seniors, children, and handicapped. Therefore, we need a paradigm shift in the design of mobile networks to be able to provide effective and practical interest-aware services in a mobile network. In such a new approach, implicit inference of user interests shall be the first step toward the design of behavior-aware (interest-based) communications.

Mobile networks now have enough potential to accurately capture and infer multi-dimensional aspects of users' behavior, interests and preferences. Mobile devices can now act as distributed behavioral sensors of users to capture their interests and behavior and enable implicit behavioral profiling. Such enriched profiles can be then

analyzed for various mobile networking scenarios at various spatio-temporal granularities to facilitate the creation of accurate behavioral models. Based on the acquired interest profiles and behavioral models, we can develop a new type of communication paradigm that we refer to as *iCast* (or Interest-aware Casting). ICast extends the concept of profile-cast [6,7] and tries to deliver messages based on the user interests in a mobile network.

In this chapter, we first discuss different types of architectures to develop the iCast paradigm including centralized, semi-centralized, distributed and hybrid architectures. Then, we show how such a paradigm can be developed in practice via simulation of semi-centralized architecture. In this architecture, we suggest and evaluate three different approaches of message delivery including random, semi interest-aware and full interest-aware casting. We analyze the proposed approaches considering different parameters including target-receiver similarity score, computational and communication overhead and scalability, usability and privacy and show how they affect the performance of networking communications.

A Spectrum of Architectures for Interest-Aware Casting

Today's Internet uses machine-oriented abstractions (e.g., IP addresses) for routing unicast and multicast traffic. Moreover, users are required to explicitly input their interests and social information. Such architecture for behavior-aware networking creates inefficiencies due to sub-optimal indirections, and disadvantages for classes of technically challenged users. Therefore, new sets of mechanistic building blocks are needed to provide a native behavior-aware network, without the need for indirections or explicit collection of users interests. For this purpose, we propose a new network abstraction that embodies the notion of behavior-aware addressing. In this approach,

instead of sending a message to a destination address (unicast or multicast IP) the message is destined to a target behavioral profile. This notion is general and a profile can be defined as needed. However, we shall provide concrete behavioral profiles based on mobility preferences and online interests. Figure 8-1, shows a sample representation for the behavioral profile. In this case, the user profile shows the percentage of his online time at different buildings for different web domains over a period of time (which can be the last day, week, month, etc.). A simpler profile may just show the percentage of online time at different web domains without considering different locations or vice versa.

In order to develop the idea of interest-aware casting, different architecture can be considered; centralized, semi-centralized, and fully distributed architectures. A common set of essential building blocks is specified for all these architectures, including: i- behavioral-profile builder that consists of data collectors and profile processors, ii- behavioral-based dissemination that consists of target identification, behavioral matching, and behavior-based routing. In the following, we discuss these architectures and their corresponding building blocks.

Centralized Behavior-Aware Networking

This architecture is used to illustrate the role of the behavioral-profile abstraction (Figure 8-2). The behavioral data may either be collected by access points in the infrastructure, or reported by the users' mobile devices. The data is then processed by a centralized server (or a group of servers) and behavioral-profiles are generated. The process of sending a message starts by target identification (to be used as the destination profile) using behavioral metrics (e.g., target users with main interests in "netflix" and "mac" websites who visit the "library" and "business fraternity" frequently). A

sender generates the target profile representation, and then sends the message and the target to the server for matching and dissemination. Realization of this scheme may borrow from existing centralized services. In this architecture, issues of reporting overhead and privacy may arise.

Semi-Centralized Behavior-Aware Networking

In this architecture, data processing is done partially at the users' mobile devices, and only a general profile representation (showing major activities, but not details) is sent to the server to reduce privacy concerns (Figure 8-3). This profile representation not only preserves user privacy to good extend by just providing an overview of user's activities and keeping all other details at the user side, but also provides enough information for target matching of messages at the server side.

Distributed and Hybrid Behavior-Aware Networking

In the distributed architecture, mobile devices not only generate the behavioral profiles, but also disseminate and route the messages in a peer-to-peer fashion by locally performing the target profile matching (Figure 8-4). As no data is stored and processed on a server, this scheme preserves privacy but has lower performance than centralized and semi-centralized architectures. Therefore, a hybrid approach may also be considered depending on the required service and performance, and availability of infrastructure. For example, semi-centralized architecture may be used during normal operation while distributed architecture may be used during emergencies when servers fail.

A Spectrum of Problem Parameters

In any of the discussed architectures in the previous section, different input and output parameters can be considered to be assigned or measured. The input

parameters are generally determined by the context of the problem or our requirements, interests or constraints. The output parameters are determined by the solution we use based on the provided input parameters. Parameterization of the problem can enable us to compare different solutions and to choose the best one that shows better result for our studied context. In the following, we discuss such parameters.

Input Parameters

Three different input parameters can be considered for the interest-aware casting including: a) number of users; b) number of profile features, e.g. domains or buildings; and c) coverage rate showing the percentage of population we want to reach.

Number of users: This parameter is generally determined by the context. For example, in the campus-wide case study, the number of students and employees determines the number of users.

Number of profile features: This parameter is usually determined based on our interests or constraints. For example, in our case study, we are interested in top active web domains and among them we pick the top 100 ones because of the constraints imposed by the available computational recourses.

Coverage rate: This parameter shows the percentage of population we want to reach. The coverage rate is generally determined based on our requirements.

Output Parameters

Five different output parameters can be considered for the interest-aware casting problem including: a) similarity of receiver profiles and the target profile;
b) computational overhead and scalability c) communication overhead and scalability; d) usability; and e) privacy.

Target-receiver similarity score: The first parameter, similarity score, shows how well the target users are selected and how interesting the delivered messages are for them. Low similarity shows lows efficiency and high similarity score shows high efficiency in this regard.

Computational and communication overhead and scalability: The second and third parameters show the amount of computational and communication overhead for any provided solution. Based on these parameters we can also measure the scalability of the solutions as well.

Usability: The forth parameter shows how usable and convenient the provided solution is to the end users. The number of parameters a user needs to provide to be able to send a message, e.g., the amount of target user interest in different domains, is one of usability parameters that we can consider.

Privacy: The last parameter shows how well the privacy of users is preserved when they share their interest profiles. This parameter is important because if a solution does not preserve the privacy of users, they will not share their profiles and therefore the solution will not work in practice.

Case Study: Semi-Centralized Interest-Aware Casting

In our case study, we simulate the semi-centralized interest-aware casting. For this purpose, we use our integrated campus-wide dataset to create user profiles. In this study, we first build the profiles based on the percentage of online time in different web domains during a month. We use these offline-generated profiles instead of locally generated profiles by mobile devices in the actual architecture. Then, we employ the semi-centralize interest-aware casting using different message delivery approaches as explained in the following.

Interest-Aware Message Delivery

According to the semi-centralized architecture, when user profiles are built on the mobile devices, these profiles are sent to a central server. At the same time, when a user wants to send a message, he needs to send the message as well as its target profile to the server. When server receives the message, it needs to determine target users for the message. For this purpose, we suggest two interest-aware approaches that we refer to as *semi interest-aware* and *full interest-aware* casting. These approaches utilize user interest profiles to find the best matched users. In the following, we explain these two approaches as well as a none interest-aware approach, which we refer to as *random cast* and use it for the evaluation purpose.

Random cast

If we do not collect and form user interest profiles, the only way to transmit the messages to the interested users is to send every single message to all users and let them decide if they want to read that message or not. It is obvious that using this approach put a huge burden on the mobile network, mobile devices and of course users to transmit, receive and select the desirable/useful messages. To remedy this issue, an immediate approach is to send the messages to only a random portion of users not all depending on the available resources and the number of messages. We refer to this approach as random cast.

Semi interest-aware casting

In the first interest-aware approach, semi interest-aware casting, target users are selected based on user interest profiles maintained by the server. For each message, the server examines all the available profiles and sends the message to only top

matched users (considering a threshold). This approach, as we will discuss later, can improve the performance of message delivery to some extend, but does not scale well.

Full interest-aware casting

In this approach, we not only utilize user profiles on the server, but also try to build a big picture of the mobile society based on the available profiles and provide this big picture in the form of a set of recommendations to the sender. The rational behind this idea is the fact that when senders have no idea about the current interests in the community, they might frequently try to target users with specific interests who actually do not exist in the society. Providing a big picture of the community interests will help users to better target the receivers. For this purpose, senders are provided by a set of recommendations describing different existing interest groups inside the mobile society. In this way, senders can get an insight on people interests inside the mobile society and choose the appropriate target for their messages.

However, to take this approach we first need to find existing interest groups inside the mobile society. For this purpose, we need to create realistic models of user interests using their profiles. In Chapter 4, we showed how information theoretic co-clustering model can be effectively applied to solve this problem in large mobile societies. As discussed, this model can provide a multi-dimensional characterization of dynamics within the mobile society considering web or location visitations. In this model, users as well as their profile features (i.e., their interests in websites or locations) are clustered simultaneously. In this way, we can identify different behavioral groups representing similar points of interest (considering websites or location visitations). This model as we discussed is stable around 90% during several months and therefore we do not need to rerun the method very frequently to get the updated characteristics of the interest

groups. Considering these properties, we apply this method to detect the interest groups and then build the recommendations based on the characteristics of each group. Each recommendation shows the average interests of users in one of the groups.

Evaluation and Analysis

In this section, we evaluate and compare different message delivery approaches discussed in the previous section. For this purpose, we first briefly explain the input parameters. Then, each of the output parameters and the related measurements and results are discussed in detail.

Input parameters

We run the simulation for 22,816 users with interest profiles including 100 features representing the percentage of their online time at different web domains (for the top 100 active ones). The required coverage rate, i.e., the percentage of population we want to reach, is 10% in this simulation.

Target-receiver similarity score

In this case study, we use similarity score as the main metric for the evaluation purpose. The reason is that if we reach people who are not interested much in the received messages, the concept of interest-aware delivery will not be effectively useful. For this purpose, we measure similarity of receivers' interests and the targeted user profile by the sender. This metric shows how well the process of message delivery is performed. If the amount of similarity is high it shows that receivers will be actually interested in the received messages but if not it shows the opposite way. To measure this metric we simulated the three mentioned approaches using our dataset. In the experiments, we assumed that a sender wants to send a message to a target profile. In the first two approaches, i.e., random and semi interest-aware, target might be any

arbitrary profile, but in the third approach, i.e., full interest-aware, it can be one of the recommended ones. For each case, we repeated the experiment 10 times to remove any noises and measured the average performance of message delivery based on the similarity metric explained above.

Figure 8-5 shows the empirical CDF (cumulative distribution function) for interest similarity between receivers profile and an arbitrary target profile in random cast. X-axis shows similarity and y-axis shows the CDF value (representing the percentage of users with a similarity score less than or equal to the value on the X-axis at each point). In the experiment, a message with an arbitrary target was sent to a group of randomly chosen users (10% of population in this case). The experiment was repeated 10 times for different arbitrary targets and the plot shows the average result. As can be seen in the figure, in this approach more than 85% of receivers have a similarity score less than 40%. If we look at the maximum and average similarity score, we also find that the maximum similarity score that we can achieve in this approach is around 66% while the average is less than 27%. This shows that the random cast approach shows a very poor performance in terms of similarity metric.

Figure 8-6 shows the result for the second approach, i.e., semi interest-aware casting. The plot in the figure shows the empirical CDF of interest similarity between receivers profile and an arbitrary target profile. X-axis shows similarity and y-axis shows the CDF value. Again, this value is the average for 10 experiments. In each experiment, we matched the target against all the available profiles and send the message to only top 10% of users showing the highest similarity score. As can be observed form the plot as well, in this approach the maximum similarity score is the same as the random cast

(66%), which is trivial, however the minimum similarity is increased to 38%. Therefore, the average similarity score for the receivers in this approach is improved from 27% in random cast to more than 45%. While this approach shows relatively good improvement, it still suffers from two issues. First, the amount of similarity score is not very high. Second, it does not improve the upper bound for the similarity score. This is very important because it shows that even if we send multiple messages we will not (or rarely) get a similarity score better than 66%.

Figure 8-7 shows the result for the full interest-aware casting. In the experiment, we had 10 recommended interest groups that were built using co-clustering method. The plot in the figure shows the average empirical CDF for interest similarity between receivers' profile and a recommended target profile. X-axis shows similarity and y-axis shows the average CDF value. As can be observed form the figure, this approach does not improve the minimum similarity, however only less than 2% of receivers have a similarity score less than 60%. In this approach the maximum similarity score is close to 1 and the average is more than 82%. This shows that the full interest-aware approach can significantly improve the performance comparing to the other two approaches.

Computational overhead and scalability

In order to compare different message delivery approaches form the computational aspect, we need to investigate how much extra computational resources are needed to deploy each of the approaches. The shared computational overhead between different approaches that is tied to the nature of semi-centralized architecture is on the user device side, which needs to track user activities and build the interest profile. However, as the process of creating and updating the interest profile can be performed incrementally using simple arithmetic operations, it does not put a heavy

burden on the mobile devices. It is obvious that as this task is performed on the mobile devices not the server, it scales well even with huge number of users.

The second source of overhead is the computational operations needed on the server side. It is obvious that in the random cast there is no overhead but running a random function to choose the targets. However, in the semi interest-aware approach, we need to first choose the targets based on their interest profiles. Therefore, for every new message every single interest profile needs to be examined to find the best matched users. Thus, for *n* user and *m* messages the computational complexity of matching process will be *O(mn)* which can put a relatively high computational burden on the server if we need to deal with millions of users and messages. This problem is solved in the full interest-aware approach. In this approach, we do not need to run the matching process every time we get a new message. Instead, we just need to form the interest groups and recommendations for senders once a while (with a specific frequency e.g. a day or a couple of hours) and then reuse them for a period of time. Of course, this approach needs some computational resources to detect and build the interest groups. However, this process does not take so much resource (in our case study, the running time for the co-clustering method is in the order of seconds) and moreover it just needs to get run once a while. Therefore, the full interest-aware approach shows a significant improvement in terms of required computational resources and scalability as well.

Communication overhead and scalability

The common communication overhead for semi-centralized interest-aware casting is to transfer user interest profiles formed on the mobile devices to the server and keep them updated. The size of user profiles depends on the number of interest features,

e.g., the number of web domains (100 ones) in our case study. Therefore, for *n* number of users and profiles with *m* features we need to send *n* profiles of size *m* to the server. To keep the profiles updated on the server, mobile devices need to send update messages frequently. However, as the value of the profile feature shows the average of users' interests, which does not change dramatically, update messages are not required to be sent very frequently. Therefore, in general the process of transmitting and updating the user profiles on the server does not take much communication resource.

The main difference between different casting approaches discussed before is in their methods of choosing targets and message delivery. To compare different approaches form this perspective, for each one, we calculate the number of messages required to be sent to reach a certain percentage of population with a certain minimum amount of interest in the received messages. For this purpose, in our case study, we investigate the number of messages we need to send to reach 10% of population with a minimum interest level of 60% and then try to see how the number of required messages are changed when increasing the minimum interest level.

As can be observed form Figure 8-5, in random cast for each message only around 0.6% of recipients will have a similarity score more that 60%. Therefore, if we want to reach 10% of the whole population with a similarity score of at least 60%, we will need to send more than 170 messages to achieve this goal. By looking at Figure 8-6, we can observe that this number is significantly reduced by factor of 10 to 17 messages in the semi interest-aware approach. The reason is that in this case around 6% of recipients will have an interest profile with more than 60% similarity to the target profile. However, the number of required messages in this approach increases

dramatically as well if we want to achieve similarity score of more than 60%. For 61% we need around 35 messages, for 62% around 100 and for 63% it goes beyond 500 messages. Similarity score of more than 66% is almost impossible in this approach even with more increase in the number of messages. Therefore, although the semi interest-aware approach can significantly reduce the communication overhead, but it does not scale well for the higher required similarity scores. This problem has been solved in the full interest-aware approach. As can be seen in Figure 8-7, more than 97% of recipients have a similarity score of more than 60% in this approach. This means that sending one or at most two messages will suffice to achieve the goal of reaching 10% of population with minimum interest level of 60%. If we increase this to 70% or even 80%, two messages will still suffice as more than 65% of recipients have a similarity score of more than 80%. Therefore, the full interest-aware approach put a lot less overhead on the communication network and scales a lot better as well.

Usability

Usability metric shows how convenient the usage of a service is for the end users in practice. This is measured in our case study based on the number of parameters a user needs to set before sending a message. It is obvious that in the random cast a user does not need to provide any extra setting, however, as shown in previous sections, this approach is not efficient and practical in the real world considering other discussed metrics. So, we compare the other two interest-aware approaches from this perspective.

In semi interest-aware approach, senders need to determine a target profile for each message. This means that they need to set multiple values for all the profile features which in our case study is equal to 100. It is obvious that providing such details

for every single message (i.e. providing the percentage of online time for 100 different websites in this case) is not very easy and pleasant to the end user specially if he uses a mobile device. This problem is solved in the full interest-aware casting by providing recommendations to the end user. This approach makes it a lot easer for senders to choose the target groups for their messages based on the provided recommendations. In this approach, senders can easily pick one of the recommended options even from their mobile devices without spending too much time. Therefore, the full interest-aware approach shows better usability comparing to the semi interest-aware approach.

Privacy

When users share their interest profile, an important concern they usually have is the issue of privacy. In the semi-centralized architecture, the privacy of users is preserved in two ways. First, users do not share all the details of their interactions and they just share a summarized interest profile with the server. Second, the available interest profiles on the server are never shared with the senders. In semi interest-aware approach senders basically do not receive anything about the users form the server and in full interest-aware approach they just receive recommendations that are formed based on interest groups not the individual users. Therefore, in both suggested approaches the privacy of users is fully preserved.

Summary

In this study, we introduced a novel interest-aware approach for information casting called iCast. In the new paradigm, messages are sent to targets based on users' interests not their device or Internet addresses. To develop such a paradigm, we proposed three different architectures including centralized, semi-centralized, distributed and hybrid behavior-aware casting from which we simulated the second one. For the

simulation of semi-centralized architecture we proposed three different approaches for targeting the users including random, semi interest-aware and full interest-aware approaches. We evaluated the performance of these approaches based on different metrics. We showed that the full interest-aware approach can significantly improve the performance of user targeting from 27% for random cast and 45% for semi interest-aware approach to 82%. We also discussed how the later approach can significantly improve the scalability and usability in a real application and preserve the privacy as well.

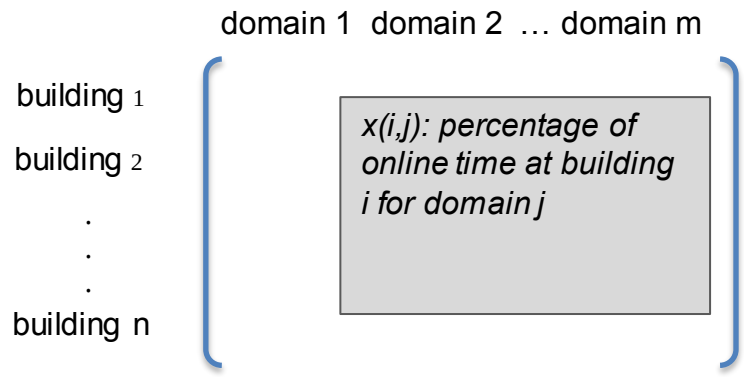


Figure 8-1. A sample representation for behavioral profile.

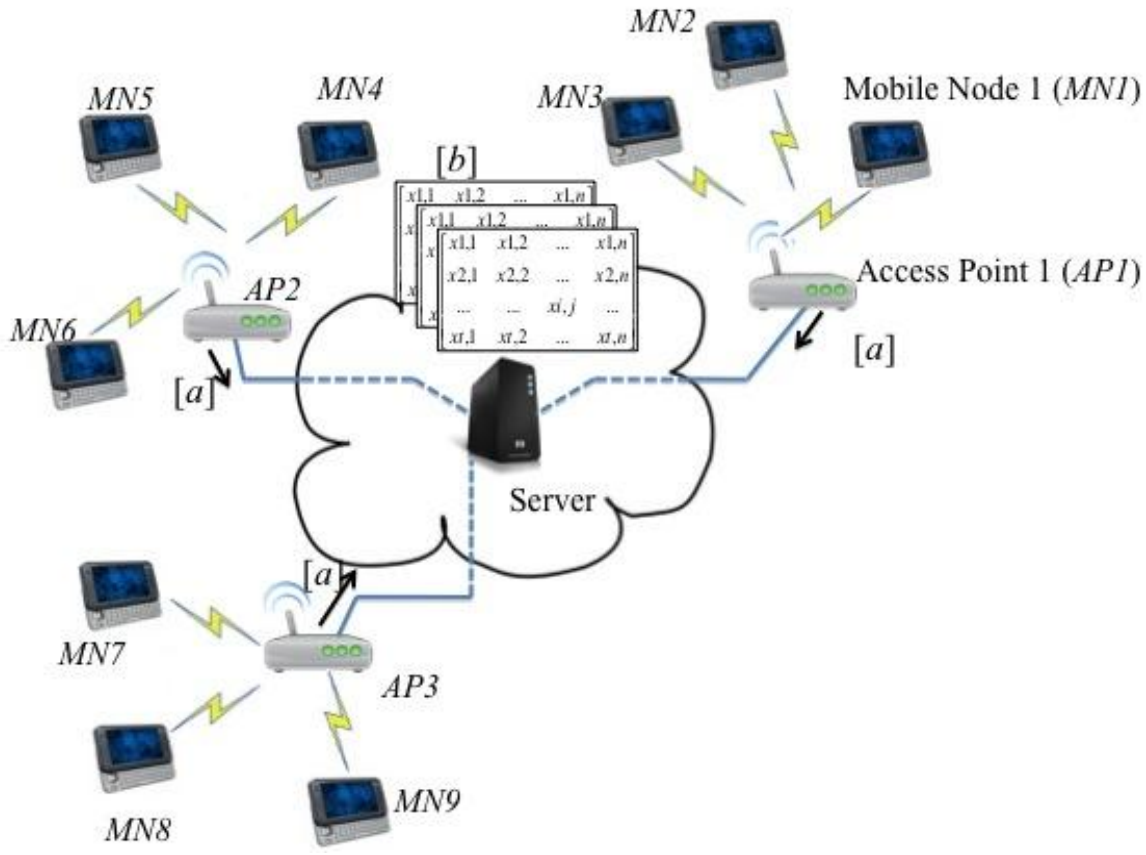


Figure 8-2. Centralized architecture: *a*. Wireless access points collect and relay information about activity and mobility of users. *b*. The information is relayed to the server that then constructs activity/mobility interest matrices for all users and processes them. Issues of reporting overhead and privacy arise.

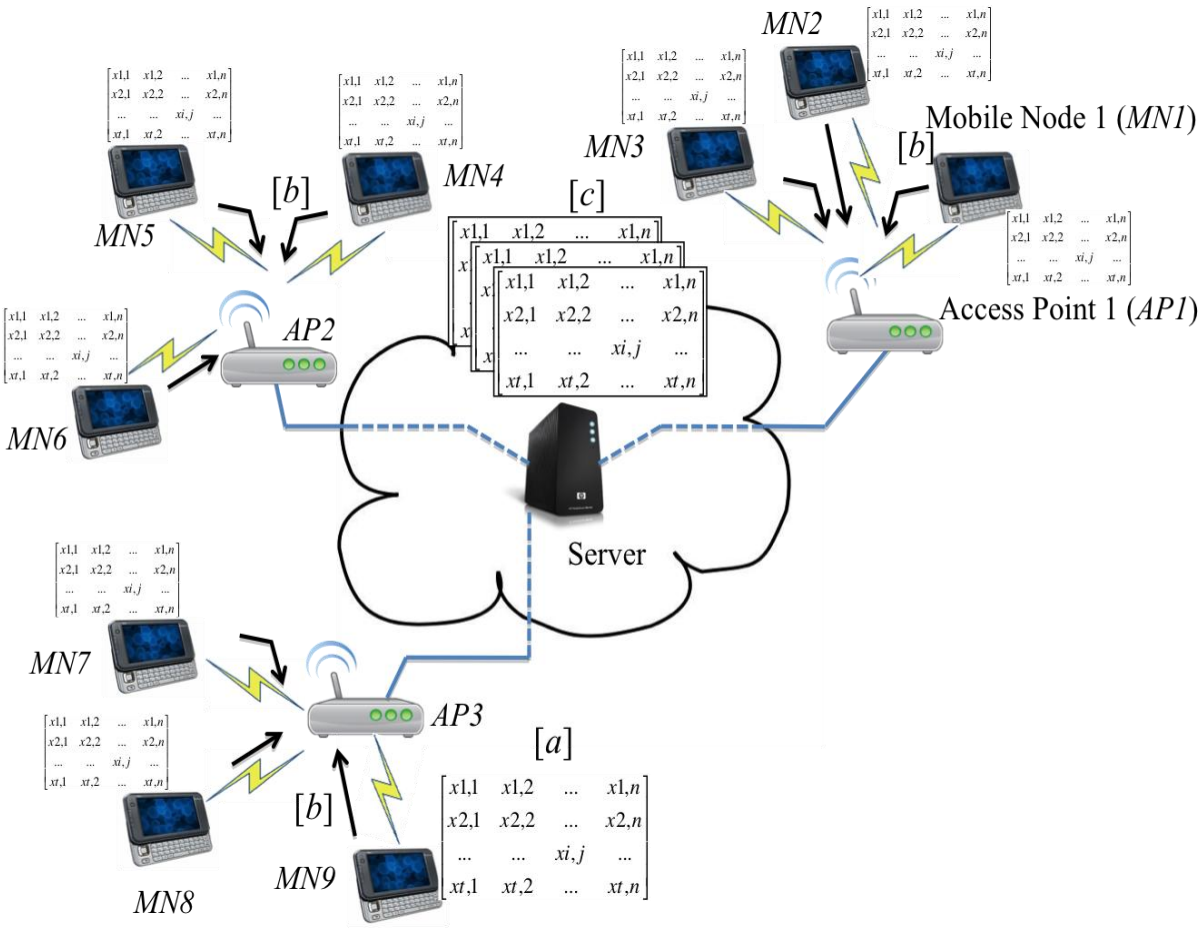


Figure 8-3. Semi-centralized architecture: *a*. Wireless devices collect and store activity and mobility interest matrices for individual users. *b*. The information is relayed to the server occasionally. *c*. Server processes information for all users. Reduced reporting overhead by the wireless devices. Users have control over profile reporting.

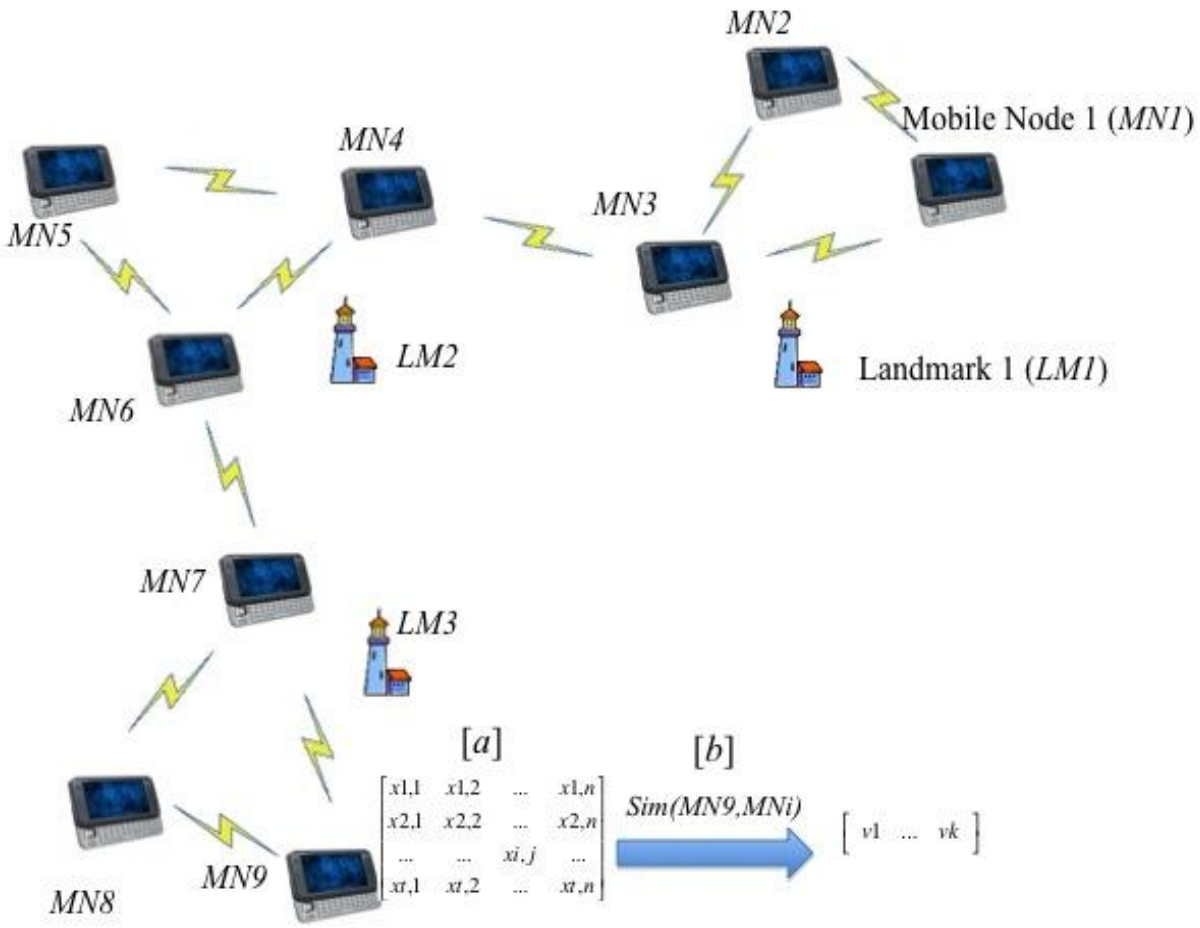


Figure 8-4. Distributed architecture: *a*. Wireless devices collect and store activity and mobility interest matrices for individual users. *b*. Similarity processing performed locally at each device during message forwarding. No reporting of profiles required, and no centralized server.

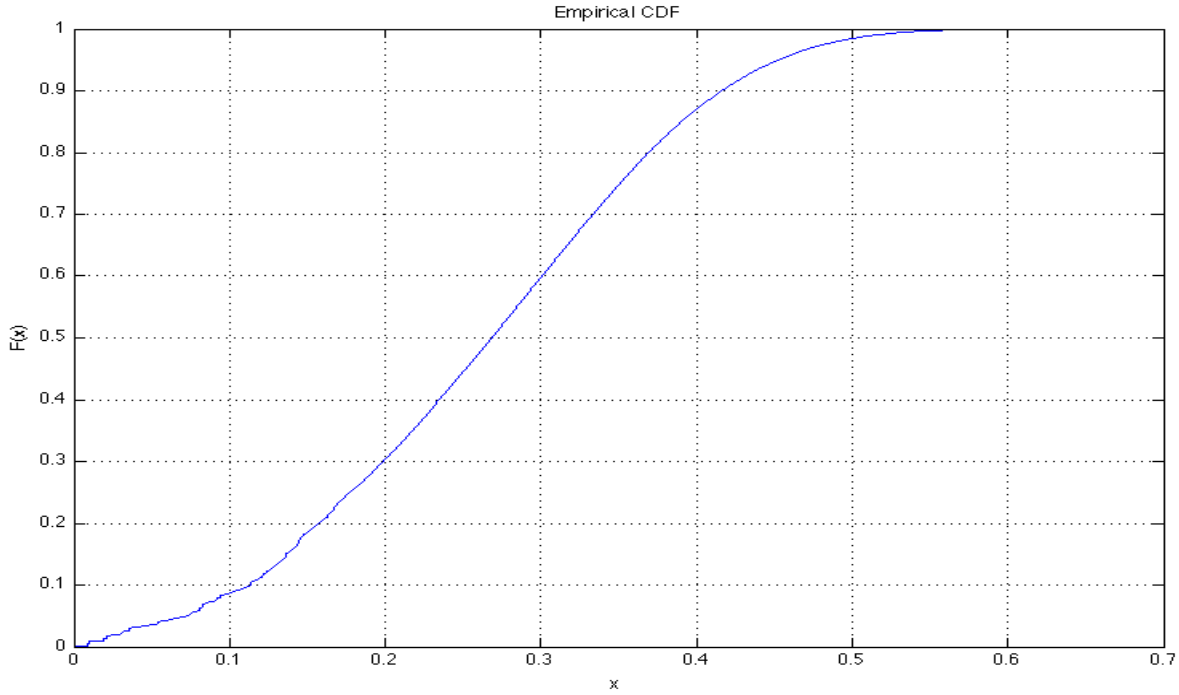


Figure 8-5. Empirical CDF for interest similarity between receivers' profiles and an arbitrary target profile in random cast approach. X-axis shows similarity and y-axis shows the CDF value.

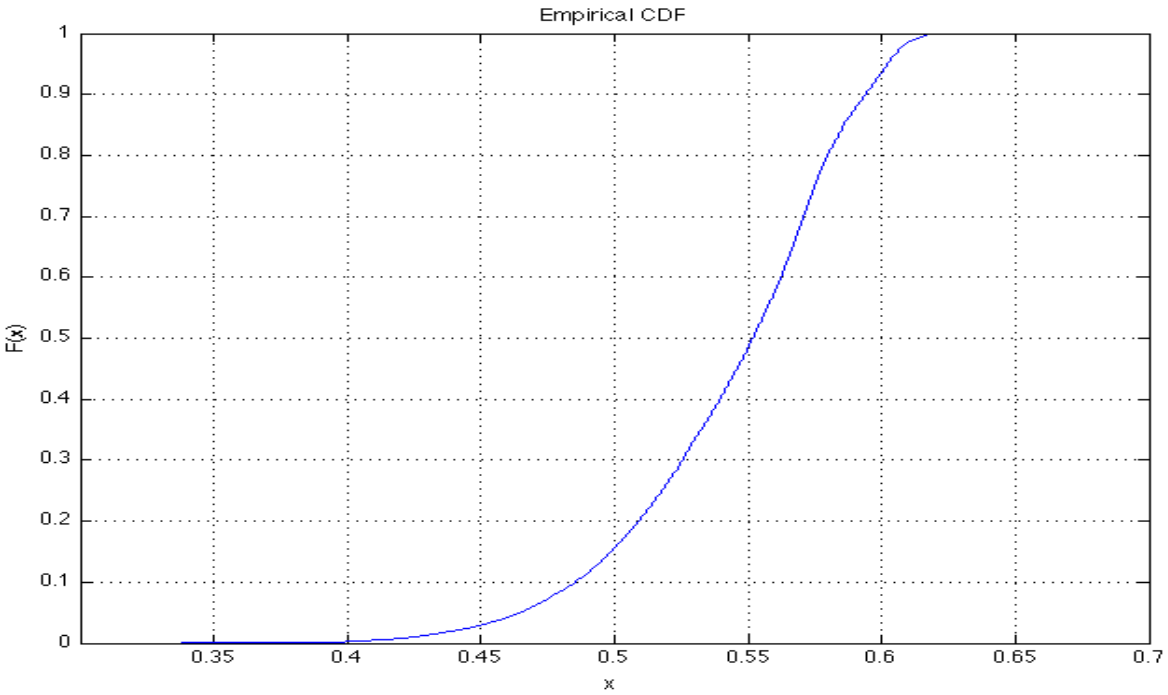


Figure 8-6. Empirical CDF for interest similarity between receivers' profiles and an arbitrary target profile in semi interest-aware casting approach. X-axis shows similarity and y-axis shows the CDF value.

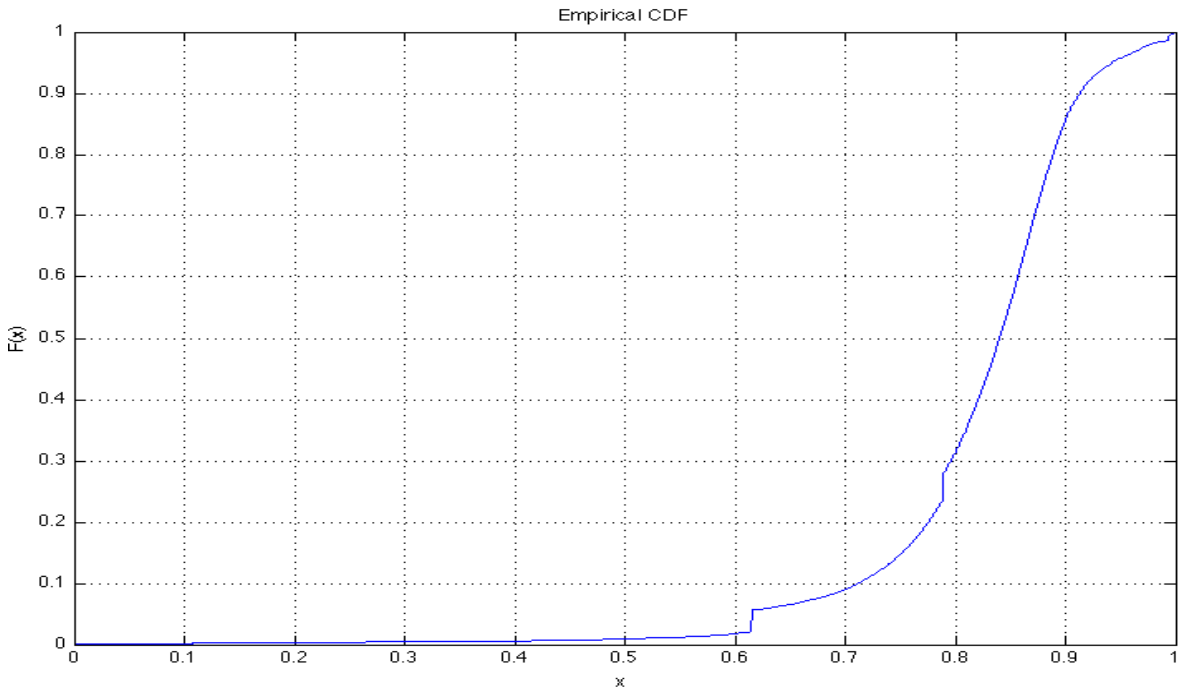


Figure 8-7. Empirical CDF for interest similarity between receivers' profiles and a recommended target profile in full interest-aware casting approach. X-axis shows similarity and y-axis shows the CDF value.

CHAPTER 9
CONCLUSION AND FUTURE WORK

This dissertation was motivated by the need for a paradigm shift that is data-driven to develop realistic behavioral models and to design behavior-aware solutions for the future mobile networks. In this dissertation, we provided systematic methods to process, analyze and utilize the characteristics of the largest mobile trace to date, with billions of records of Internet activities from a campus network, including thousands of users and dozens of web domains and buildings. For this purpose, in the first step, new processing techniques were applied for integration and aggregation of large datasets. Then, novel analysis and behavioral modeling was conducted utilizing advanced data mining techniques including information-theoretic co-clustering, at the global and location-based levels, multi-aspect trend modeling and correlation analysis using self-organizing maps, and modeling of behavioral components utilizing Gaussian mixture models. We showed that the co-clustering model can create and correlate multi-dimensional clusters of users, web domains and locations to capture group-specific characteristics. Our study revealed the fact that mobile online behavior can be modeled with a strikingly small number of clusters of distinct web access profiles. Similarly, buildings of the same actual categories (e.g. health) show very distinct Internet usage patterns and are often clustered together. We also showed that multi-dimensional behavioral trends and correlations in terms of online activity and mobility can be discovered, visualized and modeled using self-organizing maps. Moreover, we suggested an extended version of the model for multi-aspect modeling of users' interests in terms of both domain and locations visitations. We also proposed a

method based on Gaussian mixture model to capture the behavioral components inside the mobile society.

In this dissertation, we applied the proposed behavioral models to provide behavior-aware solutions to three different problems. First, we utilized the Gaussian mixture model for multivariate simulation of mobile users online activities and showed how it can be effectively applied for accurate simulation of users behavior in large mobile societies. Then, we investigated the problem of traffic analysis and modeling in mobile networks and showed how domain, location and human-centered modeling reveals new interesting facts and also improve the accuracy of traffic models. We explained how our suggested graph-based approach can reveal similarities and differences between the traffic distributions of different domains or locations and how the co-clustering model can be effectively applied for behavior-aware traffic modeling of mobile networks. We showed that mobile Internet traffic of different domains, locations and behavioral subspaces are best modeled by considering multiple statistical distributions not a generic one as commonly used. In the last part of this dissertation, we introduced a novel behavior-aware casting paradigm for interest-aware delivery of messages (iCast) in mobile networks. We suggested a spectrum of architectures including centralized, semi-centralized, distributed and hybrid to develop this paradigm and discussed the performance of message delivery in a semi-centralized architecture. We showed that our suggested full interest-aware approach can significantly improve the performance of interest-based casting in large mobile networks.

This study was the first step toward realistic behavioral modeling of online behaviors and behavior-aware design of networking solutions. Studying predictive

modeling approaches and also considering more behavioral dimensions e.g. application type are among potential future works on the modeling part. On the application part, implementation of iCast is one of the major directions that can be followed in future work. The provided behavioral models can also be used in a myriad of other networking applications in future. We hope for this study to provide an effective and practical approach and example for this purpose and lead to a paradigm shift in providing realistic solutions for future mobile networking problems.

LIST OF REFERENCES

[1] Dhillon, I. S. and Guan, Y. Information theoretic clustering of sparse co-occurrence data. *In Proceedings of the Third IEEE International Conference on Data Mining (2003).* IEEE Computer Society.

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

[3] McLachlan, G.J. and Peel, D. *Finite mixture models*, Wiley, 2000.

[4] Plackett, R.L. Karl Pearson and the Chi-squared test. *International Statistical Review,* 51, 1, (1983), 59–72.

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

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

[7] Hsu, W., Dutta, D. and Helmy, A. CSI: A paradigm for behavior-oriented profile-cast services in mobile networks*. Ad Hoc Networks, 10, 8,* (Nov 2012).

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

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

[10] 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.

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

[12] 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.

[13] McNett, M. and Voelker, G. M. Access and mobility of wireless PDA users. *SIGMOBILE Mobile Computing and Communications Rev*., 9, 2 (Apr 2005), 40-55.

[14] 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, 2004). ACM.

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

[16] Hsu, W. and Helmy, A. On nodal encounter patterns in wireless LAN traces. *In Proceedings of the IEEE WiNMee 2006* (Apr, 2006).

[17] 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.

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

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

[20] Hsu, W., 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.

[21] 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.

[22] 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.

[23] 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).

[24] Bai, F. and Helmy, A. A survey of mobility modeling and analysis in wireless adhoc networks, *Wireless Ad Hoc and Sensor Networks*, Springer, 2006.

[25] 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.

[26] 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.

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

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

[29] 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.

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

[31] Borst, S. User-level performance of channel-aware scheduling algorithms in wireless data networks. *IEEE ACM Trans. Network*, 13, 3 (Jun 2005), 636-647.

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

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

[34] Sarvotham, S., Riedi, R., Baraniuk, R. Connection-level analysis and modeling of network traffic. *ACM IMW* (2001).

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

[36] Feldmann, A. Characteristics of TCP connections. *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, eds., John Wiley, (2000), 367–399.

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

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

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

[40] Jain, A. K. and Dubes, R. C. *Algorithms for clustering data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[41] Dhillon, I. S. Co-clustering documents and words using bipartite spectral graph partitioning. *In Proceedings of the ACM SIGKDD 2001* (San Francisco, CA, 2001). ACM.

[42] Moghaddam, S., and Helmy, A., Multidimensional modeling and analysis of wireless users online activity and mobility: a neural-networks map approach, *In Proceedings of the ACM MSWiM* (USA, 2011). ACM.

[43] Moghaddam, S., and Helmy, A., Internet usage modeling of large wireless networks using self-organizing maps, *IEEE MASS SCENES* (USA, 2010).

[44] Moghaddam, S., and Helmy, A., Spatio-temporal modeling of wireless users Internet access patterns using self-organizing maps, *IEEE INFOCOM,* (China, 2011).

[45] Moghaddam, S., and Helmy, A., Modeling of Internet usage in large mobile societies using self-organizing maps, *ACM SIGMOBILE MC2R* (Oct. 2010).

[46] Moghaddam, S., and Helmy, A. Spatio-temporal modeling of wireless users Internet access patterns using self-organizing maps, *CoRR abs/1008.4904*: (2010).

[47] Moghaddam, S., and Helmy, A., SPIRIT: A simulation paradigm for realistic design of mature mobile societies, *IEEE IWCMC* (Turkey, 2011).

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

[49] Fruchterman, T. M. J. and Reingold, E. M. Graph drawing by force-directed placement, *Software – Practice & Experience,* Wiley, 21, 11 (1991), 1129–1164.

BIOGRAPHICAL SKETCH

Saeed Abbasi Moghaddam was born in Iran, in March 1980. He received his Bachelor of Science in computer engineering (software engineering) from Amirkabir University of Technology/Tehran Polytechnic (AUT) in 2003 and his Master of Science in information technology (software development & design) from Iran University of Science & Technology (IUST) in 2006. Saeed started his Ph.D. in 2008 at University of Florida focusing his research interests on mobile networks, data mining and very large data processing and received his Ph.D. in December 2012. Saeed's professional achievements include his role as the director of the Advanced Information Technology Lab at AUT and the principle founder of Rayneel Inc. His academic achievements include the Elite Student Award at the 6th National Olympiad for Computer and Informatics, membership of the National Elites Foundation and University of Florida's Outstanding International Student Award. He also serves as a program committee member and reviewer for several conferences and scientific journals and is a member of the Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM). In addition to his diverse professional interests, Saeed enjoys photography, calligraphy, piano, poetry and painting.