Extract: Mining Social Features from WLAN Traces: A Gender-Based Case Study

Udayan Kumar
Computer and Information Science and Engineering
University of Florida
ukumar@cise.ufl.edu

Ahmed Helmy
Computer and Information Science and Engineering
University of Florida
helmy@cise.ufl.edu

ABSTRACT

The next frontier for sensor networks is sensing the human society. Several mobile societies are emerging, especially with wide deployment of wireless LANs (WLANs) on campuses. WLAN traces can provide much insight into mobile user behavior. Such insight is essential to develop realistic models and to design better networks, and analyze effects of social attributes on mobile network usage. The most extensive libraries of wireless traces are collected from university campuses, are anonymized and do not provide affiliation, gender or preference information explicitly. Hence, it becomes a challenge to analyze network usage characteristics for social groups using the existing traces. In this paper, we present two novel scientific techniques to classify WLAN users into social groups. The first technique uses mapping of the traces into buildings (e.g., dept. buildings, libraries, sororities and fraternities) to extract affiliation and gender information based on network usage statistics. The second technique utilizes directory information that can be linked to WLAN users to extract useful information. For example, usernames of the WLAN users (if available) can be used to find user's gender based on first name and databases. As a case study we perform classification and behavior analysis of users by gender. Extensive WLAN traces from two major universities are collected over three years and analyzed. Results from both the methods are cross-validated and show more than 90% correspondence.

Results of gender classification are then used to examine usage patterns and preferences across gender groups, including spatio-temporal distribution of wireless on-line activity, study majors and vendor preference. In some cases these metrics are equal across genders, however, there are several interesting cases that clearly indicate statistically significant and consistent effects of gender; e.g., males have longer online sessions in Engineering and Music, while females have longer sessions in Social Sciences and Sports areas. At one university female groups consistently preferred Apple computers. These findings can have a great impact on several mobile networking applications; they can be directly used for realistic modeling of wireless user on-line behavior, mobility and virus susceptibility, and for designing socially-aware protocols and class-based or gender-based services, to name a few.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Wireless Communication*

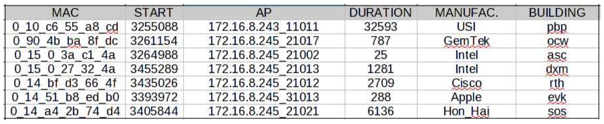
1. INTRODUCTION

In future mobile networks, with many hand held devices tightly coupled with a user, communication performance is bound to user mobility and behavior. This applies to various kinds of mobile networks, including cellular networks, but more particularly ad-hoc and delay tolerant networks (DTNs), because every node may act as a router and the network may be infrastructure-less. In such an environment, it is imperative to understand the various aspects of user behavior, including mobility, commonalities, differences in preference, and net activity between classes of users, in order to design efficient protocols and effective network models.

It is a challenge by itself to create a model/protocol that incorporates social behavior parameters (in terms of the numerousness of the parameters to choose from). This challenge is further aggravated by the unavailability of techniques that can classify users into social groups so as to extract the desired social parameters. Addressing the later challenge, in this work we propose a new approach to classification and feature analysis of user behavior based on social grouping. We provide a set of techniques that can be used to provide information about a user from social perspective. We use WLAN (Wireless LANs) traces (generally considered for studying network characteristics) to mine social behavior of the users based on gender, majors, and other interest groups. WLAN are the best source of information about real user mobility and network usage. These traces have been used in many studies whenever real user data is required. They have been previously used to validate mobility models [1, 2] and understand user associations [3] among other usages.

Our paper is the first, to our knowledge, to scientifically and automatically classify most of the WLAN users into social groups and extract parameters/feature-sets across user groups. This methodology provides a richer and more reliable data set (because data is based on user activities and not on what a user perceives as reported in surveys) that can be recomputed as and when required. We present the general methodology with an example case study of grouping by gender with investigation of gender gaps in WLAN usage. The lack of such empirical data poses an interesting challenge and raises several research (and privacy) questions: How can we meaningfully infer gender information from such anonymous traces? Does gender information influence user

MSWiM'10, October 17–21, 2010, Bodrum, Turkey.
Copyright 2010 ACM 978-1-4503-0274-6/10/10 ...$10.00.



MAC	START	AP	DURATION	MANUFAC.	BUILDING
0_10_c6_55_a8_cd	3255088	172.16.8.243_11011	32593	USI	pbp
0_90_4b_ba_8f_dc	3261154	172.16.8.245_21017	787	GemTek	ocw
0_15_0_3a_c1_4a	3264988	172.16.8.245_21002	25	Intel	asc
0_15_0_27_32_4a	3455289	172.16.8.245_21013	1281	Intel	dxm
0_14_bf_d3_66_4f	3435026	172.16.8.245_21012	2709	Cisco	rth
0_14_51_b8_ed_b0	3393972	172.16.8.245_31013	288	Apple	evk
0_14_a4_2b_74_d4	3405844	172.16.8.245_21021	6136	Hon_Hai	sos

Figure 1: A sample trace database snapshot

behavior and preference in a significant and consistent manner? Finally, what is the impact of these finding on network modeling, protocol and service design in the future?

Our study begins by introducing a location-based method for gender classification on campus. It provides robust filters, based on individual and group network behavior, in addition to clustering techniques, to identify males and females with high confidence. We analyze extensive Wireless LAN traces collected for over 3 years from 2 major universities covering more than 50,000 users. The findings are cross validated with ground truth using *Name based* method and yield over 90% success. Once the gender classification is performed, a thorough investigation of the spatio-temporal characteristics of the gender based network activity is conducted. Among the parameters we have considered for evaluating the gender gaps, we found enough statistical evidence to conclude that (for the traces used in our study) usage patterns of males and females are different, and that gender does affect user activity and vendor preference. We believe that such attributes will certainly enhance the understanding of the mobile society and is essential to provide efficient network protocols and services in the future.

Contributions: This paper provides following contributions: *i.* class and gender inference methods based on location, usage and name filtering from extensive WLAN traces, *ii.* providing the first gender-based trace-driven analysis in mobile societies, including study of majors and device preferences, *iii.* identifying unique features in the studied grouping that suggests consistent behavior and the design of potential future applications.

The rest of the paper is outlined as follows: Sec. 2 discusses multiple techniques for user classification, followed by Sec. 3, which provides several methods for validating the classification. Sec. 4 provides the gender-based feature analysis and results and Sec. 5 discusses potential applications. Conclusion and the future work is presented in Sec. 6.

2. APPROACH

In this work, we consider WLAN traces to understand usage characteristics/behavior pattern of social groups. WLAN traces are logs of user association with a Wireless Access Point (AP). Traces generally contain machine's MAC address, associating time, duration and associated AP. MAC address is always anonymized to protect privacy of the user. Having a meaningful classification (into social groups) with this partial information is the main challenge that we address in this work. Ideally, we would want to classify all users into groups. Taking a first step in this direction we present a general technique, which can be used to classify a smaller section of WLAN users into groups. Doing it for all the users still remains a challenge as we shall see. Instead, we focus on obtaining a sample significant enough for a statistical analysis.

Our technique works on raw WLAN SNMP and SYSLOG traces. The traces are accumulated for a time period and parsed into a standard format. The processed data is fed into a database on which SQL queries can be run easily (and generically) to extract information of interest. Fig. 1 illustrates the generic trace database layout, which is used in our experiment. The fields include the following: 1. anonymized MAC addresses of the wireless devices logged onto the WLAN, 2. the session start time (in seconds), 3. the AP with which the wireless device associated, 4. Duration of the association with the AP, 5. the manufacturer of the wireless card (which we inferred from partial MAC address), and 6. the building at which the AP is located (inferred based on a map), this field is external to the actual traces. Mobility of users can be tracked by looking at the approximate geographic locations of the APs. In some cases, if more information such as usernames are available, we can add more fields to the database. The advantage of having a standard schema for the database is that similar queries can be used on traces coming from multiple sources. We have used this same database framework to analyze traces from USC[5], Dartmouth [4], UF and UNC[6]; the method is general and applicable to many traces (campuses and urban) and several grouping criteria.

The trace collection process, environment, and anonymization used have a great impact on the utility of the traces. Hence, it is difficult to find one general method, which would classify users in all settings. Therefore we propose multiple methods. One challenge in this study is to validate the results obtained from trace analysis against the ground truth. We have used several statistical methods to give us confidence in the classification and cross-validated our results with the name-based approach; closest possible to the ground truth at a large scale.

We use traces from two universities, U1 and U2 (names withheld for privacy reasons) that provide information as shown in Fig. 1 except that university U2 trace also provides the usernames. Traces from U1 belong to Feb 2006, Oct 2006 and Feb 2007, and Traces from U2 belong to Nov 2007 and Apr 2008. The grouping parameter we use in this work for investigation is gender based. To do this categorization, we propose two novel techniques: *Location based Classification* (LBC) and *Name based Classification* (NBC), and subsequently, we examine and discuss their advantages. Both of these techniques are generic and have been presented in this work with an example case study of gender based classification.

2.1 Location Based Classification (LBC)

Most US universities have sororities (female organizations) and fraternities (male organizations) as social organizations. The buildings, which houses these organizations also serve as residences for most of the members. Given the physical location of APs on campus, APs located in sororities and fraternities are identified, and the users associated with them are classified as females or males respectively. This method can also be used to classify users by other grouping criteria such as study major. For example all users associating with Computer Science building AP can be classified as Computer Science major students. Since wireless networks may be used by anyone in the physical proximity to the AP, this kind of classification will also have un-related users or visitors accessing these APs, which can make the classification inaccurate. We next present techniques to filter out regular users from visitors at an AP.

Filtering: LBC requires filtering, as fraternities and sororities have male and female visitors. Without further refinements and filtering, this method would not be accurate. But even if we validate the presence of visitors, how can we filter them from our classification? First, visitors are infrequent users of the mobile network in the visited locations. Second, we expect a significant difference between residents and visitors in terms of network activity (in number and duration of on-line sessions). Third, a user who is visitor at one location can be a regular user at some other location. Hence, we can define a visitor as a user with less number of sessions and smaller duration of sessions than the average user in that location (group behavior) or as user who has more sessions and larger online duration at other locations (individual behavior). Our filtering techniques rate users based on two metrics: the number of sessions and session duration. Once we rate all the users on these two metrics, we apply cut-off thresholds to determine regular users. Filtering can be performed on these ratings considering individual and/or group behavior as described in rest of the section.

2.1.1 Individual Behavior Based filtering (IBF)

In Individual Behavior based Filtering (IBF), we find the probability of a user being male or female by counting the number of sessions and measuring the duration he/she spends in fraternities versus sororities. This can be done using the equations below.

The probability of a user being male, considering only session counts at fraternities and sororities is given by:

$$PCM(u) = \frac{C_f(u)}{C_f(u)+C_s(u)}$$

where function C_f gives session count for user u in fraternities and function C_s gives the session count for user u in sororities. Similarly, the probability of a user being male, considering only session durations at fraternities and sororities is given by:

$$PDM(u) = \frac{D_f(u)}{D_f(u)+D_s(u)}$$

where function D_f gives the total duration of sessions for user u in fraternities and function D_s gives the total duration of sessions for user u in sororities. Fig. 2 shows users who visited fraternity and/or sororities in decreasing order of $PCM(u)$ and $PDM(u)$ for Feb 2006 traces from university U1. Interesting observation is that both PCM and PDM follow a similar trend and there is a sudden transition from 1 to 0 (between 500th and 700th user), essentially separating males from females. Out of 1119 users, there is a large number ($\sim$ 425) of users whose probability of being male is 1. These users have never associated with sororities APs. We also have large number ($\sim$ 362) of users who have never associated with fraternities AP ($PCM = 0$ and $PDM = 0$), who we can classify as females. As fraternities and sororities have visitors, many males will have probability less than 1 (vice-versa for females), if we only consider users with probability 1 or 0, we would considerably remove legitimate users who have visited and used WLAN at other locations (sororities for males and fraternities for females).

We have instead classified all the users having $PCM >$ 0.80 and $PDM >$ 0.80 as males and $PCM <$ 0.20 and $PDM <$ 0.20 as females, using the 80-20 rule or the Pareto principle such that 80% of the regular users should fall in top 20% probability. Other users are discarded from the study. We also did the analysis for all the trace sets belonging to university U1 and U2, and saw a very similar distribution as seen in Fig. 2. This method, IBF, is generic and can also be used in other grouping criteria such as study major among others.

2.1.2 Group Behavior Based filtering (GBF)

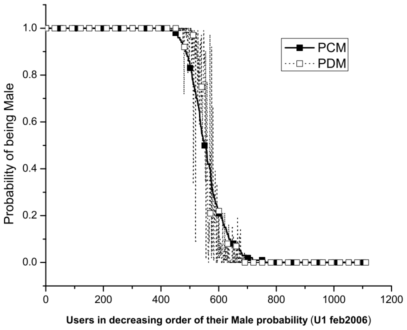


Figure 2: Users visiting Fraternity and/or Sorority in decreasing order of their Male probability (U1 feb2006)

In Group based Filtering (GBF), we filter a user based on where his usage pattern lies with respect to all the users at a particular location. GBF is also useful when traces are available only from limited number of buildings and we cannot use IBF due to lack of traces from all the buildings. For example lets consider that at a particular location, we discover that average session duration of regular users is 3000sec and their session count is 10 in a period of one month. So all users who at least meet these criteria can become regular users and are classified as male or female based on the location, everyone else is considered a visitor and therefore removed. Finding these thresholds is not a trivial task as these thresholds would vary from building to building and may also change with time. For this task we employ clustering techniques [7] (one of the key methods for unsupervised learning) to partition our data into regular users and visitors.

Clustering: Clustering can be used to divide a set of users into several subsets such that users in each subset are most similar based on WLAN usage metrics (duration, session count, distinct login days). From two general category of clustering algorithms; namely hierarchical and partition scheme, we choose a robust partitioning method called **Partitioning Around Mediods** (PAM) [8]. This method has distinct advantages (over standard *k-means* [7]) in that it uses dissimilarity score to minimize dissimilarity in the same cluster, making clusters robust to outliers. It also provides a novel method called Silhouette Widths and Plots for estimating cluster quality. The average Silhouette Widths are useful in estimating the number of clusters present in the data (often a challenging job in cluster analysis). One has to run PAM several times, each time for different number of clusters and then compare the resulting Silhouette Widths. The clustering size that produces maximum average width is the best clustering possible. The average width can also be used to estimate the quality of the clustering; above 0.70 for strong clustering, between 0.50 – 0.70 for a reasonable structure and below 0.50 for weak structure [8].

We use PAM to distinguish visitors from regular users (i.e residents). We use number of distinct days of login, session count, and sum of session durations as the metrics for user evaluation. This metrics can help identify and thus separate users who make several sessions only in few days (may be visitors) from users who make sessions everyday. We applied this clustering technique to Sororities and Fraternity user trace from both Universities U1 and U2. We found that the best cluster size in each case is 2. In each set we found that average silhouette width is above 0.65, 0.84 being the maximum in one of the cases (more results in Tab. 1). The cluster size of 2 clearly identifies our intuition of regular users and visitors and separates them using usage behavior in that particular building/location. Also, the high average silhouette width indicates the high quality of clustering. Detailed results of GBF are in middle column of Tab. 2.

Fig. 3 shows effect of total session duration, total number of sessions and unique days of login over clustering of users.

	U1			U2	
	Feb 2006	Oct 2006	Feb 2007	Nov 2007	Apr 2008
Fraternity	0.72	0.74	0.75	0.84	0.78
Sorority	0.65	0.72	0.69	0.78	0.76

Table 1: Average Silhouette Width for Sorority and Fraternities from University U1 and U2

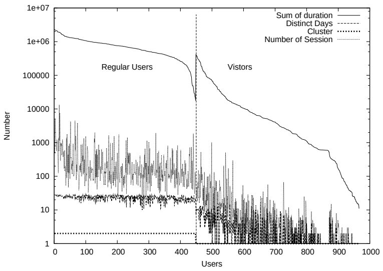


Figure 3: Clustering results for University U1 Sororities (feb2006)

We can see a clear drop in the number of sessions and unique days when the clustering changes from 2 to 1 (2nd cluster signifies the resident). We notice that at the beginning of cluster 1 there is a spike in the total duration but still these users are not included in the regular users as their number of sessions and unique days of login are comparatively less than users belonging to cluster 2. Clustering ensures that all three metrics are incorporated when making a decision. Similar results are obtained for other traces from university U1 and U2. GBF is generic and can be used to identify other social groupings such as study-major, which will be investigated in our future research.

2.1.3 Hybrid filtering (HF)

As we do not know the ground truth or have the real data about the users, it is difficult to validate the results of these classifications. In order to have a meaningful analysis after the classification, we need to validate the classification. We validate LBC via multiple techniques in Sec. 3. In one of the techniques, we compare the results from IBF and GBF. Results are tabulated in Tab.4. We find that both methods mainly select same set of users, which should be the case as both methods attempt to identify regular users (males in fraternities and females in sororities). Therefore, for higher confidence/correct classification and analysis in the later sections of the paper, we choose the users selected by both filtering methods. We call this method *Hybrid Filtering* (HF) as this uses results from both IBF and GBF. By doing so we successfully classify majority of the users (more than 90% of the users selected by GBF are common to users selected by IBF based method as shown in Tab. 4).

Our proposed scheme of LBC is generic and can classify users into social groups if these groups have inherent location preferences (Sororities are females residences, Computer Science major has strong ties with Computer Science buildings or Theater group meets often at the auditorium). One thing to note is that LBC and its filtering techniques do not need access to unanonymized MAC address. As long as the MAC addresses are consistently anonymized, LBC is applicable. This property makes LBC usable in most of the available WLAN traces.

For traces/environments where locations cannot be used to distinguish user groups (like non-existence of sororities or fraternities), we propose another method - Name Based Classification (NBC). Name Based Classification method can work in certain scenarios where LBC is constrained, but it requires additional information.

2.2 Name Based Classification (NBC)

In this technique, we augment the traces with user specific information from external sources. A few traces provide additional information like usernames (This field may be obtained on campuses and enterprises that require authorization mechanism such as passwords to access WLAN). This extra information can be used to link external data with the traces that can allow us to classify the users. External data can be extracted from directories, yellow pages, schedules and other public information sources. Traces coming from university U2 provide us with usernames. University U2 also host a directory that can be searched using these usernames (as usernames also serve as email addresses) and users have the option of not listing their names in the phone book directory. This allows us to search the directory and find the first names corresponding to the usernames for the users who have made their information available in the phone book directory. We then use the list of top 1000 males and females first names from the US Social Security administration website [9] and remove the names present in both lists (neutral names). Thus, we get the list of most popular male-only and female-only names. We run this list against the list of names we find from the phone book directory, thus finding the gender of the users [10, 11]. In this technique, we do not have problem of visitors thus we do not need any filtering. We observe that names from the US Social Security list may not be able to classify foreign national students and non-popular names into gender groups, this however is not a limitation of our method but of the name database. Using a more comprehensive database should provide better classification. In this paper, however, we are more concerned with a general methodology of classifying WLAN users, the details of how to acquire a better database are out of scope of the paper.

Using NBC classification, we could classify 11,000 as males or females out of 27,000 users in the trace period of Nov 2007, and 12,500 as males or females out of 30,000 users in the trace period of Apr 2008 at University U2. Details of the classification are listed in Tab. 2. This method is dependent on the kind of external data we can link the traces to. For the university U2, the directory also provides study-majors corresponding to a username, this information can thus be used to study the study-major classification of the users.

Compared to NBC, LBC requires less information (username not needed); however, we need to find a way validate LBC. One way to validate is to compare classification results of LBC with NBC as shown in Sec. 3.3. NBC method is much closer to the ground truth. The use of NBC is limited as the availability of usernames is limited to a very few currently available traces. Once we check the correctness of LBC, this can become the primary method for classification.

3. VALIDATION OF LBC

Validation of LBC is needed to raise confidence in the results from U1 i.e. users classified as visitors are indeed visitors and not the regular users of that Access Point (males in case of fraternities and females in case of sororities). Validation of the results with the ground truth/actual reality is difficult, especially when we have developed the methods for publicly available traces and information. Even if we get access to students' university records, we would not be able to match it with student's device (especially when MAC addresses are anonymized). Surveying 50,000 users in

	U1-IBF			U1-GBF			U2-NBC	
	Feb 2006	Oct 2006	Feb 2007	Feb 2006	Oct 2006	Feb 2007	Nov 2007	Apr 2008
Total Users	16416	22405	20302	16416	22405	20302	27068	29982
Males(only)	506	553	545	451	437	417	5245	5807
Females(only)	513	570	509	441	456	410	5955	6817
Common	0	0	0	22	37	29	0	0

Table 2: Results of classification of users from U1 (LBC) and U2 (NBC). 'Common' signifies the users which were common to both male and female population.

each campus may result incomplete and noisy (erroneous) data aside from the enormous efforts/resources needed if at all possible. Instead, we have devised three statistical methods to validate our filtering mechanisms. The first method finds out regular users in the trace-set belonging to adjacent months and compares this list to see how many are common (temporal consistency). The second method compares results from IBF and GBF to check the similarities in the results. The third method takes the classification achieved using NBC method and compares it with the results of LBC because NBC should be very close to the ground truth. The methods are discussed in detail below.

3.1 Temporal Consistency Validation

In this method of validation, we consider a pair of one month long trace-sets belonging to adjacent months in the same semester (such as February 2006 and March 2006 from Spring 2006 semester) and use IBF, GBF, and HF filtering techniques to find out how many users are common between the two adjacent months before and after filtering. Assumption being that the set of users living in fraternities and sororities do not change from one month to another in the same semester. If after filtering, the percentage of common users increases then it is likely that this method works correctly in identifying regular users. Tab. 3 shows the results we obtain for both fraternity and sorority users. We see that for fraternities, before filtering, the percentage of common MACs in two consecutive months is around 60% to 64% and after filtering it goes upto between 72% to 80% in all three filtering schemes. In case of sororities, before filtering, we see that common users are between 66% to 72% and after filtering the percentage of common users shoots up to 80% to 93%. This shows that filtering schemes are selecting regular users, as percentage of common users rises dramatically after filtering.

Before Filtering

Month(a)	Month(b)	% common (Fraternity)	% common (Sorority)
Feb2006	Mar-Apr2006	60.4	72.3
Oct2006	Nov2006	63.8	66.8
Feb2007	Mar-Apr2007	62.1	70.2

After Filtering- IBF

Month(a)	Month(b)	% common (Fraternity)	% common (Sorority)
Feb2006	Mar-Apr2006	76.2	87.7
Oct2006	Nov2006	72.5	80.9
Feb2007	Mar-Apr2007	76.5	81.9

After Filtering- GBF

Month(a)	Month(b)	% common (Fraternity)	% common (Sorority)
Feb2006	Mar-Apr2006	80.0	92.7
Oct2006	Nov2006	78.27	87.6
Feb2007	Mar-Apr2007	79.4	92.3

After Filtering- HF

Month(a)	Month(b)	% common (Fraternity)	% common (Sorority)
Feb2006	Mar-Apr2006	79.8	92.4
Oct2006	Nov2006	78.2	88.3
Feb2007	Mar-Apr2007	77.9	90.4

Table 3: Similarity in the user population selected after filtering fraternity users for U1

3.2 IBF vs GBF

The LBC technique in Sec.2.1 describes two main filtering techniques - IBF and GBF. Both use location information to identify the gender; however, cut-off thresholds for filtering regular users and visitors are set differently. Comparing the results of both methods provides us with another validation mechanism. Tab. 4 shows comparison of filtering results for 3 months long traces (Feb2006, Oct2007, Feb2007) from university U1. We can see that greater than 400 (75%) users are consistently common in both the methods. This points to the high degree of similarity, which validates the filtering that both methods remove visitors and result in similar regular users (increasing the confidence in our results). We note that GBF is more conservative (less number of regular users) than IBF, which could be attributed to the fact that GBF takes into consideration the usage attributes (session count, duration, distinct days of login) of an average user for comparison (by using clustering), which can be higher than a regular user selected by IBF. For the user behavior analysis, in the following section, we only consider the users selected by both filtering methods also referred to as *Hybrid Filtering (HF).*

Month	Gender	IBF	GBF	HF
Feb 2006	Male	506	451	416
	Female	513	441	435
Oct 2006	Male	553	437	418
	Female	570	456	454
Feb 2007	Male	545	417	399
	Female	509	410	406

Table 4: Validation - comparing users selected by IBF and GBF for U1

Month	FL	$FL \cap MN$	E_f	ML	$ML \cap FN$	E_m
Nov 2007	1280	74	0.058	334	25	0.074
Apr 2008	1690	123	0.072	349	29	0.083

Table 5: Cross validation of LBC by NBC for U2

3.3 Cross Validation

NBC does not classify all users as either male or female (Sec. 2.2), however, this classification has a low error rate because of using statistics from real data coming from the US Social Security Office. Using this property of NBC, we can find out the error bound for the LBC. Availability of the error percentage can help in realizing the error margins for LBC. To calculate the error bounds, the users (from sororities and fraternities) classified by LBC as females and males are put in sets FL and ML respectively.

Using NBC, we classify all users from Fraternities and Sororities and put them in different sets. Females in set FN and males in set MN, and remove the unclassified users. The unclassified set of users are those whose name existed in both male and female databases or whose name was not in the database. The error in female classification by LBC can be given by E_f, where $E_f = (FL \cap MN)/FL$ and the error in male classification by LBC can be given by E_m, where $E_m = (ML \cap FN)/ML$.

Tab. 5 provides results on the cross validation of LBC by NBC. We did the analysis for trace sets coming from university U2 as it provides usernames along with the information about AP located in the sororities and fraternities, which allows us to perform both NBC and LBC. For Apr 2008 traces from university U2, the set FL has 1690 users after doing LBC and E_f is equal to 7.2%. In case of set ML, which has 349 users, we find that E_m is 8.3%. Similarly, in Nov 2007 traces, E_m and E_f is less than 8.3%. The low value of error, E, further increases our confidence in the LBC and validates the classification method.

To sum, we find our location classification LBC (with three filtering techniques - IBF, GBF & HF) are supported by three validation techniques. Vali-

dation ensures the users selected by the filtering are indeed the regular users, which in sororities means selecting females and in fraternities selecting males. The filtering statistical errors were below 10%, and the confidence was found to be over 90%.

4. USER BEHAVIOR ANALYSIS

Classification of users into social groups is the first step in understanding the usage differences between the groups. The classification techniques discussed in Sec. 2 take all the WLAN users and divide them into various sets (depending on the grouping criterion). For the gender based grouping, we have three sets : Male, Female and Unclassified (grouping could not be determined). These groups can now be evaluated on multiple metrics depending on the application. In this work we have considered three generic metrics (not corresponding to any application). We investigate the spatio-temporal distribution for wireless usage across genders in addition to vendor preference. The main aim of these metrics is to examine the existence of differences between the groups. We attempt to identify differences that are statistically significant and consistent across the multiple traces we have studied. The three metrics are discussed below.

4.1 User Spatial Distribution

An example of a metric is the spatial distribution of the users. This metrics can identify where the classified users spend most of their time (regular users). For example, by searching the female users in the complete trace we can find out the locations visited by them. We refer these locations as "Area", since they also represent major/department housed at that location. Here we only look into major trends by the active user. A user is considered active (regular) at an area by using GBF. Difference in the number of users among the genders can tell us about the building preferences of the genders. Fig. 4 and Fig. 5 show percentage distribution for males and females at Universities U1 and U2 at various buildings. At both universities, we can see that there are more males than females in the areas of Economics (by 39% at U1 and 33% at U2), Engineering (5% at U1 and 89% at U2) and Law (by 83% at U1 and 6% at U2). Law area information for Feb2007 is a outlier as we do not have any male student during that period. Females are more in number than males in the area of Social Science (by 16% at U1 and 3% at U2) and Sports (by 41% at U1 and 2% at U2). We see that at U1 and U2 trends are opposite for the area of Music (U1 has 40% more females however U2 has 33% more males).

Existence of locations, which are consistently preferred by one of the two genders, highlights the existence of difference in WLAN usage by two genders. Many of the trends hold even across the two campuses. We believe this can be beneficial to several application as discussed in Sec. 5.

4.2 Average Duration or Temporal Analysis

Average duration of a session for males and females gives us an understanding of the extent of WLAN usage at different areas. From Fig. 6 and Fig. 7, we observe that males on average have longer sessions than females in most of the areas (on average by more than 9%, in extreme cases by as much as 200%). On average, male users tend to stay - as WLAN users - at certain places for longer times than females. At both universities, we see that females consistently have higher average duration than males in the area of Social Science (by 12.8% at U1 and 10% at U2) and Sports (by 17.2% U1 and 8% U2). Males consistently have higher duration session at both universities in the areas of Engineering (by 76% at U1 and 15.4% at U2) and Music (by 39.9% at U1 and 36.8% at U2). We see that females at university U1 consistently have higher average duration in the area of communication (by 12%) where as males have higher session duration at university U2 (by 10%). We also see clear trends at university U2 that males have higher session duration at area of Economics.

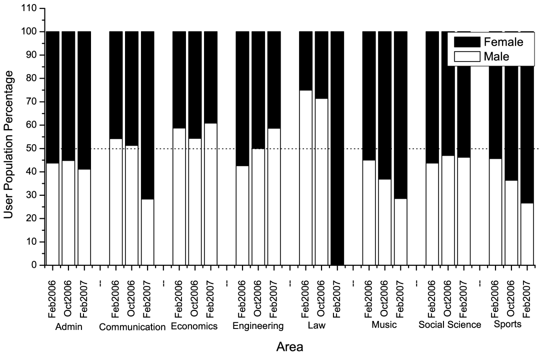


Figure 4: Comparison of user distribution across the university U1 campus (in Percentage)

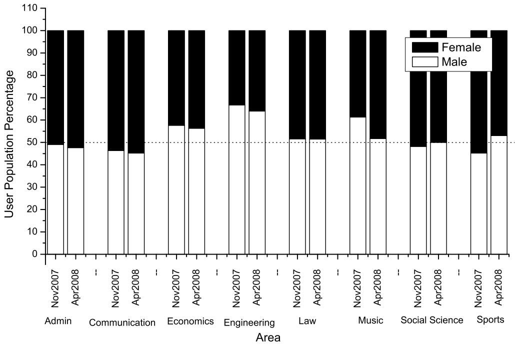


Figure 5: Comparison of user distribution across the university U2 campus (in Percentage)

Another observation of interest is that average duration per session decreases from Feb 2006 to Feb 2007 (from 2789 sec to 2454 sec) in almost all the cases for university U1 campus, we observe similar trend in university U2 (from 3800 sec in Nov 07 to 3609 sec in Apr 08). This points to the possibility that students are becoming more mobile, and thus have shorter sessions at the same location.

While in some cases the trends were equal across genders, ***in several scenarios we do find differences in WLAN usage*** among the genders. Some of these differences were found to be significant and spatio-temporally consistent even across campuses; females' wireless activity is stronger in Social Science and Sports areas, whereas males' activity is stronger in Engineering and Music. *In other scenarios each university campus had a different trend specific to it.* These findings are likely to have a significant impact on usage modeling in wireless networks

4.3 Device Preference

In many available traces, partial MAC anonymization is done, such that top three octets of the address (which identify the Manufacturer) are left unchanged. Traces from both U1 and U2 use partial anonymization. These top octets can be used to find preferred vendors for the groups (Male and Female). In this metric, we are only considering major vendors (by the number of users).

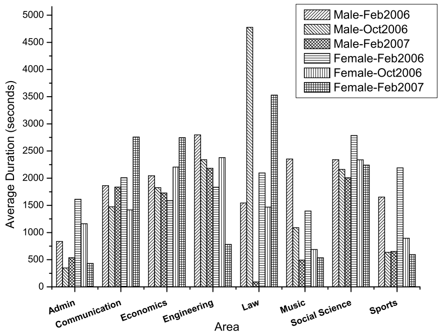


Figure 6: Average duration of male and females in different Areas of university U1 campus

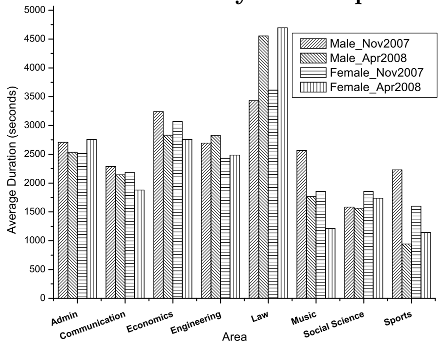


Figure 7: Average duration of male and females in different Areas of the university U2 campus

Fig. 8 and Fig. 9 show the number of users per vendor at University U1 and U2. At university U1, it is interesting to note that **Apple computers are more popular amongst females than males. Intel devices are more popular amongst males**. For example, using the Feb 2006 traces we find that 25% of the males use Apple and 32% use Intel, so that there are 28% more male users using Intel with respect to Apple users. In the case of Females, 30% use Apple and 27% use Intel, so 12% more female users use Apple than Intel. To test whether gender provides a bias towards specific vendors, we use the *Chi-Square* statistical significance test. The *Chi-Square* test shows with 90% confidence that *there is a bias between gender and vendor/brand*. This holds true for all the three trace sets from university U1. We also notice a consistent increase in percentage of Apple computer users of both genders over the three trace samples.

For comparison of the results from university U1 with university U2, for this case only, we considered users only from fraternities and sororities from university U2. The classification of users was performed using LBC (similar to university U1). At university U2, we do not find trends similar to university U1, we see that both the genders consistently prefer Intel devices more than the Apple devices. We tend to believe that preference of WLAN users can wary with geographic location and factors such as affluent society, presence of Apple store on campus among others.

We also observe that vendors like Enterasys, Linksys, D-link and Askey Corp. have a decreasing trend in terms of percentage of users. One of the reasons is that these manufacturers mostly make external Wi-Fi devices for old laptops (with no built-in Wi-Fi NICs). Currently almost all new laptops come with a built-in Wi-Fi, so the users of external devices are decreasing.

These results indicate once more that there are statistically significant differences in the usage pattern of the two gender. One possible implication of this device preference is that PC viruses or malware propagation in some female groups may be less effective, which will have a direct impact on security studies in future wireless societies as in DTN.

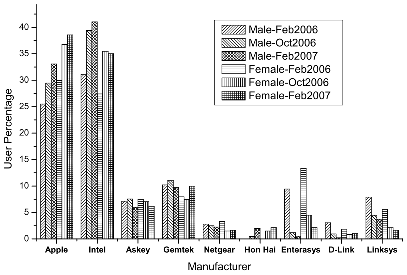


Figure 8: Device distribution by manufacturer at university U1

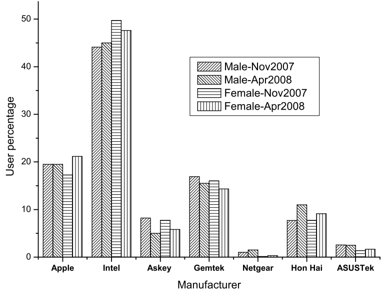


Figure 9: Device distribution by manufacturer at university U2

5. APPLICATIONS

Analysis of user behavior in the previous section highlights that statistically significant differences exist in the usage pattern of the two genders. There can be several metrics on which a group of users can be evaluated and their behavior quantified. The results from these metrics can then be applied to an existing or new application to make it context sensitive. In this section, we discuss few applications which will benefit from the quantified differences among the groups such as mobility modeling and protocol design. We also discuss impact of this analysis on user privacy, wireless network deployment, and resource management among others. For the lack of space, more details of the application are omitted.

5.1 Mobility Models

Mobility models are important tools to understand user movements and create models on which protocols can be tested. The knowledge of groups can be used to re-evaluate mobility models such as TVC [1], IMPORTANT [13], and several others [14]. This enhancement can allow us to model social groups on ‘behavioral’ aspects, load (sessions duration) and density among others. This kind of study can only be possible by using the methods mentioned in this work, other methods like taking a survey of 50,000 users would require tremendous effort and may still have similar error rates.

5.2 Protocol Design

Protocol and service design in Mobile Ad-Hoc networks can take features of various groups to evaluate its performance. It has been shown in Profile-Cast [15] that considering behavior of users (profiles), one can create efficient protocols for Mobile Ad-Hoc Networks. This work does not consider difference among groups of people. It has also been shown that users with similarities meet often and have closer ties [16]. Can similar people (belonging to same group) have higher chances of meeting more often? Can this knowledge increase the message delivery success? Our method helps in identifying the social groups, however, further investigation needs to be done such as combining this group information with services such as Profile-Cast

5.3 Privacy

A major impact of this work is bringing the privacy related issues with traces to forefront. Determining gender from the traces which were anonymized, shows weaknesses in current anonymization techniques. It may be argued that anonymization of location information may prevent this kind of classification, however, this not only decreases the utility of the traces, but also the authors in [17] show that location anonymization can be easily undone. The primary reason is the unique session patterns of the WLAN users. Anonymization of WLAN traces while maintaining utility of the traces is a challenging task.

We all have intuition where and how a certain group of users may use WLAN, our method allows to quantify this intuition. We believe that methods discussed in this work are the fundamental step for many interesting studies in the future.

6. CONCLUSION AND FUTURE WORK

In this study, we propose novel methods, which use WLAN traces to classify WLAN users in to social groups based on features such as gender and study-major among others. The work presents a general framework that can be applied to traces coming from multiple sources. As an example, traces from two university campuses have been used and gender based grouping classification is performed. Multiple techniques for grouping users are discussed since each one has slight advantages in certain scenarios. The study cross-validates the results by comparing results provided by each of the classification methods.

Results from this research are based on a sample of the user population, since gender may be identified based on sorority and fraternity wireless access point associations or based on name filter. We find that there is a distinct difference in WLAN usage patterns for different genders even with similar population sizes. Availability of results comparing groups of users can allow researchers to quantify the behavioral differences between the groups. We see that these trends and characteristics are consistent over periods of time and across different semesters and sometimes even across university campuses. We also see some trends that are not consistent across the two university campuses like the vendor preference. We think that some social characteristics are dependent on the location of the University campus and other facilities around the campus. Even though the results vary with time and location, it may be essential for a protocol designer of mobile networks to understand the characteristics of this network.

In the future, we plan to prepare mathematical models, which can represent a user in a particular group. This process would allow us to understand various features, which represent the user's WLAN usage characteristics. It would also allow us to classify users into groups by looking at the features only. User model would also be useful in tailoring the protocols for multicast and profile-cast to incorporate the group behavior.

We hope for this study to open the door for other mobile social networking studies and profile-based service designs based on sensing the human societies.

7. REFERENCES

[1] W. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling time-variant user mobility in wireless mobile networks," in *Proc. IEEE INFOCOM*, May 2007.

[2] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research," in *ACM REALMAN '06*.

[3] W. Hsu and A. Helmy, "On modeling user associations in wireless lan traces on university campuses," in *WiNMee '06*.

[4] "CRAWDAD: Community Resource for Archiving Wireless Data at Dartmouth," August 2008. [Online]. Available: http://crawdad.cs.dartmouth.edu/data.php

[5] W. Hsu and A. Helmy, "MOBILIB: Community-wide Library of Mobility and Wireless Networks Measurements," June 2008. [Online]. Available: http://nile.cise.ufl.edu/MobiLib/

[6] "UNC/FORTH: Repository of traces and models for wireless networks, Syslog Dataset #2," August 2007. [Online]. Available: http://netserver.ics.forth.gr/datatraces/

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[8] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, March 1990.

[9] "Popular baby names," September 2007. [Online]. Available: http://www.ssa.gov/OACT/babynames/

[10] A. Gallagher and T. Chen, "Estimating age, gender, and identity using first name priors," in *CVPR 2008*.

[11] M. O'Connell and G. E. Gooding, "The use of first names to evaluate reports of gender and its effect on the distribution of married and unmarried couple households," in *Population Association of America (PAA) 2006 Annual Meeting*.

[12] S. Tanachaiwiwat and A. Helmy, *Worm Propagation and Interaction in Mobile Networks in Handbook on Security and Networks*. World Scientific Publishing Co., 2010.

[13] F. Bai, N. Sadagopan, and A. Helmy, "The IMPORTANT framework for analyzing the impact of mobility on performance of routing protocols for adhoc networks," *AdHoc Networks Journal*, vol. 1, pp. 383–403, 2003.

[14] F. Bai and A. Helmy, *Chapter 1 A SURVEY OF MOBILITY MODELS in Wireless Adhoc Networks*. Springer, 2006.

[15] W. Hsu, D. Dutta, and A. Helmy, "Profile-Cast: Behavior-aware mobile networking," in *IEEE WCNC 2008*.

[16] M. Mcpherson, L. S. Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[17] U. Kumar and A. Helmy, "Human behavior and challenges of anonymizing WLAN traces," in *IEEE GLOBECOM 2009*.

[18] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, March 2002.