Gender-based Grouping of Mobile Student Societies

Udayan Kumar, Nikhil Yadav, Ahmed Helmy
Dept. of Computer and Information Science and Engineering, University Of Florida
Gainesville, Florida, USA
Email: {ukumar, nyadav, 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. With the rapid increase in WLAN deployment come various research challenges. WLAN traces pertaining to network usage contain very useful information and can provide us with a lot of insight about mobile user behavior and network usage. Such insight may be used to design better future networks, and analyze the effect of social attributes on usage of mobile technology. The most extensive libraries of wireless traces are collected from university campuses. The traces are anonymized and do not provide affiliation or preference information explicitly. Hence, it becomes a challenge to perform social or group-based analysis with existing traces. In this paper we present a novel technique to group WLAN users based on gender. By mapping the traces into buildings (including sororities and fraternities), we then extract affiliation (and hence gender) information based on statistical network usage. Once such grouping is attained, we examine the commonalities and differences of usage patterns and preferences between male and female groups. Parameters analyzed in our study include on-line session durations, vendor preference, and user distribution on campus and across study majors. Our results clearly indicate the effect of gender on on-line behavior and vendor preference. We find such effect to be statistically significant and consistent across various semesters. Our findings provide great promise in utilizing WLAN traces for future mobile applications, yet raise privacy concerns that we plan to investigate in future work. Also, our method provides a framework for analyzing group behavior in mobile networks in further studies.**

Keywords- Social grouping; group behavior; privacy; WLAN

I. INTRODUCTION

Most existing studies on sensor networks focus on sensing the physical world and phenomena. Although very useful, there have been very few (if any) studies on sensing the human society, which presents a new set of intricate and intriguing research questions. In this study we present one attempt to use WLAN traces to better understand certain behaviors, based on groupings, in wireless mobile societies.

There has been a rapid increase in WLAN deployment across university campuses. User activity on these networks is growing dramatically. As student societies become more mobile, with the ubiquity of wireless coverage and availability of new portable devices, there are great research opportunities to mine and study mobile student societies to understand their behavior, preferences, grouping, among other characteristics. Such understanding opens the door for the efficient design of future networks, and has been facilitated by the availability of recent libraries of WLAN traces [8][7]. The traces provide anonymized individual traces, but lack any information about the social context, attributes, affiliation or gender, and hence hide potentially very interesting characteristics of group behavior in mobile societies.

While previous works studied WLAN deployment issues [1], issues of mobility [2] and user association patterns [3][4], we aim to address issues of user classification based on social grouping. In this paper we set out to analyze WLAN usage patterns based on gender, majors and other interest groups. This allows us to examine trends among different social groups. Having an insight into user behavior from social standpoint can give us a lot of input in designing network protocols, such as delay tolerant networks (DTNs). Mobility analysis [2] and user predictions in mobile networks already gain a lot by incorporating user's social behavior. Context aware services [11] of mobile networks of the future would need to understand *the context* from user's perspective, so they may have to understand the social behavior of the user.

We propose to use WLAN traces, which are generally considered for studying network characteristics, to mine social behavior of the users. We present a general methodology with an example case study of grouping by gender, and investigate 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? Will such information influence user behavior and preference in a significant and consistent manner? In this paper, we introduce a novel technique to mine WLAN usage patterns based on gender, majors and other interest groups. Some of the central ideas in our paper include the user of building map information for the WLAN traces, knowledge of locations of departments, fraternities and sororities, and the use of sound statistical methods to classify users in majors, males and females and reason about the effect of such classification on network activity and preference. The method we provide can be used further to analyze behavior based on various other groupings.

Gender based studies have been conducted in the past to study issues such as the difference in technology adoption for the Internet [5]. This paper is the first, to our knowledge to analyze WLAN adoption patterns across these groups. Among the parameters we have considered for evaluating the gender gaps, we found enough statistical evidence to conclude that (for the traces in our study) usage patterns of males and females is different, and that gender does affect user activity and vendor preference. Our success also indicates that the problem of mobile user privacy should be re-visited.
The rest of the paper is outlined as follows. Section II discusses the main challenges in our study. Section III provides our approach for data analysis and our method. Section IV provides details about the traces used in our study, and Section V presents out data filtering technique for gender classification. Section VI provides the gender-based analysis and results. Section VII. Discusses potential applications and Section VIII concludes.

II. CHALLENGES

How can we begin to classify all the students into groups like gender and study major using only the publicly available information? The method we have used involves processing and analysis of the WLAN traces. Traces are logs of user association with wireless Access Points (AP). Traces generally contain only user's machine's MAC ID, associating time, duration and associated AP [7][8]. Often, because of user privacy issues, the MAC IDs are anonymized. Having a meaningful classification with this partial information is the main challenge we address. Ideally we would want to classify all students 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 the all students still remains a challenge as we shall see. Instead we focus on obtaining a sample significant enough for a statistical analysis.

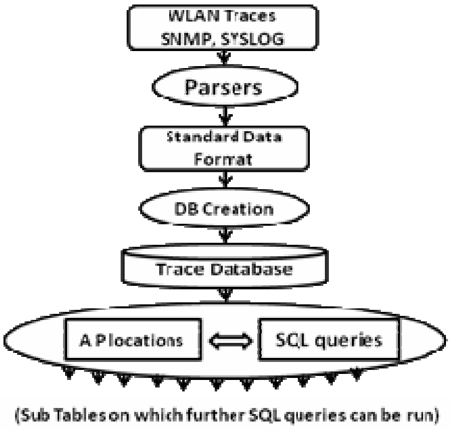
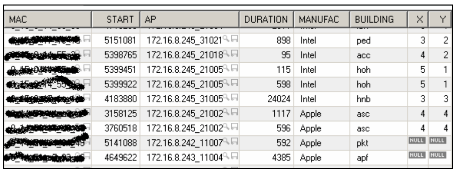


Fig 1: Query based User grouping Technique



MAC	START	AP	DURATION	MANUFAC	BUILDING	X	Y
[redacted]	5151081	172.16.8.245_31021	898	Intel	ped	3	2
[redacted]	5398765	172.16.8.245_21018	95	Intel	acc	4	2
[redacted]	5399451	172.16.8.245_21005	115	Intel	hoh	5	1
[redacted]	5399922	172.16.8.245_21005	598	Intel	hoh	5	1
[redacted]	4183880	172.16.8.245_31005	24024	Intel	hnb	3	3
[redacted]	3158125	172.16.8.245_21002	1117	Apple	asc	4	4
[redacted]	3760518	172.16.8.245_21002	596	Apple	asc	4	4
[redacted]	5141088	172.16.8.242_11007	592	Apple	pkt	NULL	NULL
[redacted]	4649622	172.16.8.243_11004	4385	Apple	apf	NULL	NULL

Fig 2: Trace Database

III. APPROACH and METHOD of DATA ANALYSIS

Our technique works on raw WLAN SNMP and SYSLOG traces. The traces are accumulated for the period we are interested in studying and parsed into a standard format as in Fig 1. We also use the location information of the APs, in the form of buildings in which they are located. This helps in knowing the geographic locations of a user at a later stage. Mobility of users can be tracked by looking at the approximate geographic locations of the APs.

The processed data is fed into a database on which SQL queries can be run easily (and generically) to extract information of interest to us. Fig. 2 illustrates the trace database layout which was used in our experimentation. The fields include the following: 1. MAC IDs of the wireless devices logged onto the WLAN, 2. the starting session time in seconds, 3. the AP which wireless device logs into, 4. Duration in seconds that the MAC stays logged into the AP, 5. the manufacturer (which can be inferred from MAC ID), and 6. the building the AP is located at (approximately), which can be checked based on access point location information which is external data to the actual traces. Two-dimensional co-ordinates can be inbuilt into the database based on a campus grid map to allow mobility based queries to be performed as well.

The trace database provides enough information on which to run SQL queries. For instance, a simple query returns the number of MACs logged into building *a* or *b* with durations within a certain range.

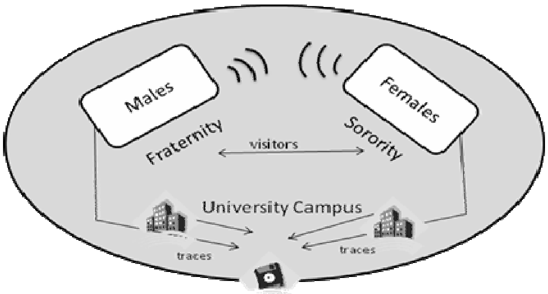


Fig 3: Gender grouping in Fraternities and Sororities

We have used this same database framework to analyze traces from USC[8], Dartmouth [7], UF and UNC[10], the method is general and applicable to many traces, campuses and societies. Completing these analyses is part of our future work.

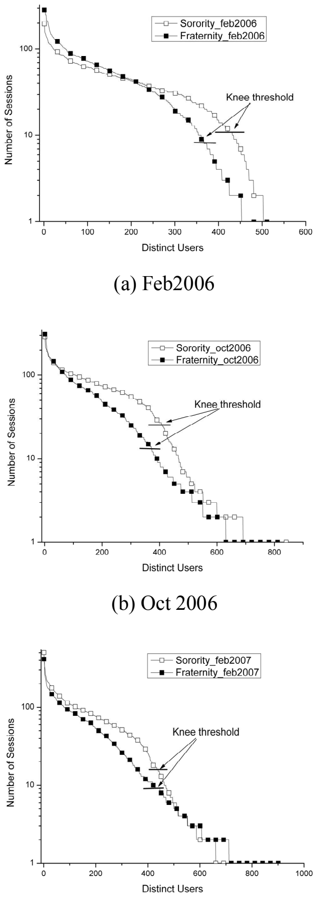
The grouping parameter we use for investigation in this document is gender. To do this categorization, we propose the following novel technique. Most universities have *sororities* and *fraternities* as social organizations. Sororities are female organizations while fraternities represent male organizations. 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 female or male. Fig. 3 shows how grouping is done in this setting. The fact that visitors may frequent these locations also needs to be taken into account. We deal with visitors in the filtering section.

IV. CHOICE OF TRACE

In order to carry out a meaningful analysis of the various groups of WLAN users, specifically targeting the differences (if any) between genders, the traces need to provide the following information: ***i.*** comprehensive Syslog (or SNMP) logs for various buildings, dorms and sororities/fraternities for the general population of students (without population bias) and for extended periods of time (30 days or more), ***ii.*** mapping between the APs (or point of collection) to specific building's designation. ***iii.*** differentiation between individual users and ability to track the same users along the whole duration of trace (without necessarily knowing their identity). We have investigated WLAN traces collected at USC[8], UNC and Dartmouth[7]. Dartmouth traces do not provide AP-to-building mapping, which makes it difficult to do this kind of study. UNC traces on the other hand have limited number of APs in sororities and fraternities. We chose the USC traces for our study as 12 fraternities and 7 sororities are included in WLAN traces and the AP-to-building mapping is also available. We have chosen 3 months for study - Feb 2006, Oct 2006 and Feb 2007. The reason for having traces from multiple periods is to look at consistency in the results and also at the trends. Traces have been taken from different semesters, so if there is something as a semester effect in the results it should also be visible.

V. FILTERING of VISITORS

As fraternities and sororities have male and female visitors, without further refinements and filtering our classification would not be very accurate. But even if we validate the presence of visitors, how can we remove 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



(a) Feb2006

(b) Oct 2006

(c) Feb 2007

Fig 4: Session count for Sorority and Fraternity users

network activity (in number and duration of on-line sessions). Hence, we define a visitor as a user with less number of sessions and smaller duration of sessions than the average user in that location. Therefore, our technique rates users based on two metrics, the number of sessions and session duration. Fig 4 represents session counts per MAC ID in decreasing order (for more detailed analysis see[6]). Fig 4 graphs are produced using the average session duration (in sororities and fraternities, respectively) as the threshold for session duration. We observe interesting, distinguished characteristic in Fig 4 that indicates the presence of a sharp bend (knee) as the number of sessions per MAC ID decreases. Intuitively, this means that MAC IDs below the knee have an order of magnitude less number of sessions (accounting for the difference between a regular user and a visitor). All users below the knee were classified as visitors and removed from the study. While users above the knee in sororities (fraternities) were classified as females (males). Changing the session duration had no effect on the shapes of the curves [6].

The general steps for visitor filtering are: ***i.*** Extract the number of sessions per MAC ID for each fraternity or sorority AP. ***ii.*** Vary the minimum session duration (as a threshold for regular users) and observe its effect on the number of sessions and distinct users. ***iii.*** Obtain a suitable threshold for the session duration and session count to classify users above these limits as being either males or females.

We also performed a time evolution study for number of sessions per MAC ID varying the minimum average session duration. In such study we perform the filtering for decreasing sample/trace sizes. An interesting observation to note from that is that the shape of the curves (with *knee*) becomes stable after the 4th day of the trace [6]. This indicates the suitability of our analysis to traces of shorter duration in similar environments.

VI. ANALYSIS OF MALE AND FEMALE WLAN USAGE

Once we have the MAC ID's of male and female users, we can analyze their behavior in the whole campus trace and identify and characterize their usage pattern over the whole campus. In this analysis we investigate the following usage and behavioral questions:

a) WLAN Usage and Gender Distribution: What are the trends in WLAN usage across different (buildings) areas on campus?

b) Average online time: Are there trends in the average online times of users and can differences be spotted based on gender and (building) areas within the campus?

c) Manufacturer preferences: Which device vendors do different genders prefer?

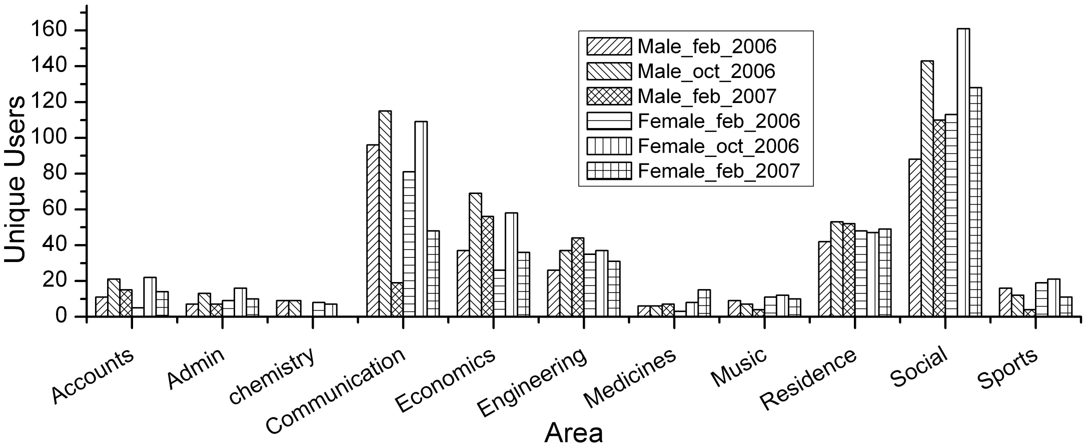


Fig 5: Unique user count across different areas

A. *WLAN usage by area*

We track MAC ID's of the previously identified/classified male and female students in different areas of the campus. If the session duration of the user is above a threshold (corresponding to the average user in that location), we consider that user a regular user of this area. Fig 5 shows the usage distribution per area type based on our definition of 'regular user'. *Economics* buildings show a higher

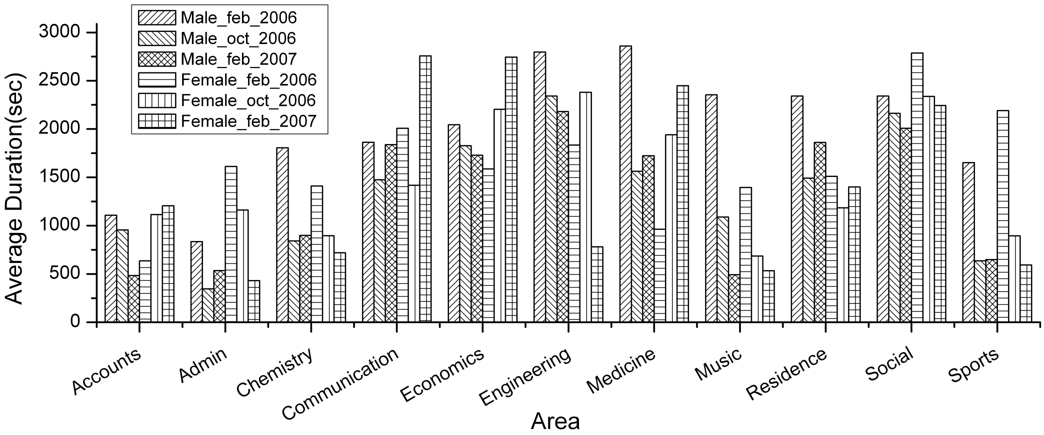


Fig 6: areas Average Session duration by Area

population of male users, social science buildings have a higher count of female users. It is interesting to see that there are more female WLAN users than males in *Engineering* buildings for the sample Feb 2006 however Males users take the lead by Feb 2007. We see that absolute number of students classified as male and females increase in Oct 2006 and then drop down in Feb 2007. This is perhaps due to the fact that more students join in the Fall semester than in any other period of the year and also that many students graduate after the Fall semester. It also indicates that several courses offered in the Fall require the use of laptops (and the wireless network) for the course work.

B. Average session duration

We now study the average session duration for male and female users across campus. From Fig. 6 we observe that males spend more time online than females in most of the areas. Females show dominant usage in the Social Science, Economics and Medicine areas across campus. We can deduce from this that on average, male users tend to stay - as WLAN users - at certain places for longer times than females. Another observation of interest is that average duration per session decreases from Feb 2006 to Feb 2007 in almost all the cases (Engineering, residence, social, sports, music). This points to the possibility that students are becoming more mobile, and thus having shorter sessions in the same location.

C. Manufacturer Preferences

The preference of manufacturer (based on the type of wireless card traced) is shown In Fig.7. It is interesting to note that *Apple* computers are more popular amongst *females* than males. Intel devices are more popular amongst males. For this study, only major vendors were considered. For example using the Feb 2006 we find that: In case of males there are ~25% using *Apple* and ~32% using *Intel*, so there are 28% more male users using *Intel* with respect to *Apple* users. In the case of Females: there are ~30 % using *Apple* and ~27% using *Intel*, so 12% more females users using *Apple* with respect to *Intel* users.

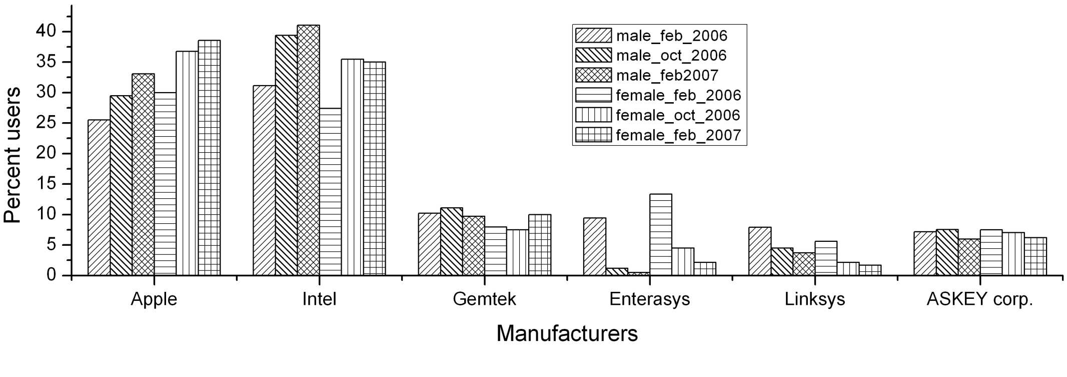


Fig 7: Device distribution by manufacturer

To test whether gender provides a bias towards specific vendors, we use the statistical significance test, *Chi-Square.* The *Chi-Square* test shows with 90% confidence that there is a bias between gender and vendor/brand.

Another interesting observation we make from Fig 7 is the consistent trend of increasing percentage of Apple computers usage in both the genders. We also see that vendors like Enterasys, Linksys 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.

VII. APPLICATIONS

In general our method for user classification and grouping can be used to group users. It may be used to profile users, grouping them based on their on-line activity. The current gender based analysis can be used to find out the extent of WLAN adoption amongst genders, which is of great interest to social scientists studying socio-economic and socio-cultural differences between genders. Further investigation of group behavior can be used to predict user movements. Trends in mobile social networking for these groups can be used to provide specific services. Protocols can be made context aware, if they have access to such user classification methods. Announcements and advertisements on campus can be directed based on the general psyche of males and females. The areas these users frequent more could serve as good places to advertise related interests, services or products. This is part of a paradigm our group is designing called *Profile-Cast* [9].

VIII. CONCLUSIONS AND FUTURE WORK

In this study we propose a novel technique to classify WLAN mobile users into groups by analyzing anonymized WLAN traces from a major university campus. We utilize mapping information of buildings and departments to obtain a meaningful, statistically sound classification. We focus our analysis on gender-based groups. 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. We find that there is a distinct difference in WLAN usage patterns for different genders even with similar population sizes. Females seem to dominate in WLAN usage in areas of Social Science and Economics and prefer Apple over Intel. Males have longer session durations than females in most cases. We see that these trends and characteristics are consistent over periods of time and across different semesters.

Our methodology of gender classification and the use of SQL queries on the WLAN traces are generic, and can be applied to classify users into groups like study major and various interests. In our future work, we plan to study similar characteristics across different university campuses, including a detailed time-based analysis of gender mobility based on different time periods of the day.

We interestingly note that we were able to classify users into male and female and were even successful in obtaining their preference of vendor, based on analysis of anonymized traces. Our study was based on group (not individual) behavior. Yet there are several privacy issues raised implicitly in our work. Can private information of individuals be identified by analyzing anonymized traces? What kind of anonymization algorithms should be used for mobile networks traces? And how can such algorithms provide a notion of *k-anonymity* for the mobile society while retaining useful information for researchers? These are questions that bear further research and we plan to address them in our future work.

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.

REFERENCES

[1] T. Henderson, D. Kotz and I. Abyzov, “The Changing Usage of a Mature Campus-wide Wireless Network,” in Proceedings of ACM MobiCom 2004, September 2004.

[2] W. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, “ Modeling Time-variant User Mobility in Wireless Mobile Networks ,” in Proceedings of IEEE INFOCOM, May 2007

[3] W. Hsu and A. Helmy, “On Modeling User Associations in Wireless LAN Traces on University Campuses,” The Second International Workshop on Wireless Network Measurement (WiNMee 2006), Boston MA, Apr. 2006.

[4] G. Chen, H. Huang, and M. Kim, “Mining Frequent and Periodic Association Patterns,” Dartmouth College Computer Science Technical Report TR2005-550, July 2005.

[5] Ruby Roy Dholakia et al. “Gender and Internet Usage, “*The Internet Encyclopedia”*, Wiley, 2003.

[6] http://nile.cise.ufl.edu/socnet

[7] CRAWDAD is the Community Resource for Archiving Wireless Data At Dartmouth http://crawdad.cs.dartmouth.edu/data.php

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

[9] W. Hsu, D. Dutta and A. Helmy, “Profile-Cast: Behavior-Aware Mobile Networking,” ACM MOBICOM poster and student research competition, Montreal, Canada, Sep. 2007

[10] UNC/FORTH repository of traces and models for wireless networks, Syslog dataset #2, http://netserver.ics.forth.gr/datatraces/

[11] Wei-jen Hsu, Debojyoti Dutta, and Ahmed Helmy, "Profile-Cast: Behavior-Aware Mobile Networking," to appear in IEEE WCNC, Las Vegas, NV, Mar. 2008.