

# Human behavior and Challenges of anonymizing WLAN traces

Udayan Kumar, Ahmed Helmy

Department of Computer and Information Science and Engineering, University of Florida  
Email: {ukumar, helmy}@cise.ufl.edu

**Abstract**—With the wide spread deployment of wireless LANs (WLANs), it is becoming necessary to conduct analysis of libraries of measurements taken from such operational networks. The availability of these trace libraries can be quite useful to provide realistic models of network load and user mobility, among others. To maintain user privacy, techniques of trace anonymization may be used to hide information. In this paper, we propose to study the fundamental trade off between the utility of WLAN traces and its privacy. The study provides several realistic case studies in which privacy attacks may be conducted, and questions the efficacy of existing anonymization techniques. Our initial quantitative analysis to estimate mobile users' k-anonymity in WLAN traces shows surprisingly unique usage patterns, which may compromise anonymity. The main contribution of this paper is to articulate the compelling challenges facing anonymization of wireless networks traces and to shed some light on the answer to a most intriguing question: Just how private are wireless networks traces?

## I. INTRODUCTION

The advent of portable/mobile devices and availability of ubiquitous network coverage using heterogeneous wireless technologies like Wi-Fi (IEEE 802.11), GPRS, 3G and Wi-Max, has allowed humans to browse information on *the go*. From sharing a computing device at home, office, or a commercial establishment, we have come to an era where these devices have become very personal and customized to user's taste. A major impact of this change (apart from all the benefits of being mobile) is that these devices have become sensors of the human society. As these devices remain with their owners for many hours in a day, they can capture large amounts of user behavior patterns, which can be made available to researchers. On one hand, the study of such data can be used to develop better understanding of human behavior and provide improved services, on the other hand, availability of this kind of data can be considered an infringement on the privacy of the user.

Several researchers use WLAN traces for research and analysis purposes such as to examine usage behavior of users[1], [2], [3], discover characteristics for developing network protocols[4] or to study user mobility patterns[5], [6], [7]. Many of the WLAN traces are publicly available[8], [9]. It is, therefore, important to understand how the privacy of WLAN users gets affected. In this work, we investigate the extent of user's private information that can be extracted from the *anonymized* Wireless Local Area Network (WLAN) traces. Even though most of the trace libraries anonymize/sanitize the traces to protect user's privacy, we present several methods, which can be used to reverse the anonymization. We attempt to expose the weakness in the currently used anonymization

techniques and bring attention of the WLAN research community on this fundamental problem. We find that WLAN traces are unique in the sense that human movement pattern gets embedded in them, which can have unique signatures. These signatures can be later combined with publicly available information from such sources as directories or schedules to identify a user even after anonymization. Despite the importance of privacy issues in WLAN traces, there is a lack of significant research in this field. The purpose of this study, therefore, is to shed light on the need of better anonymization techniques and identify a rich set of plausible scenarios in which anonymity can be compromised.

The issues of privacy and anonymization have always been present in network traces. Researchers have also faced challenges in anonymizing the wired traces[10]. Recently, wireless traces have also been collected and archived at on-line public libraries like CRAWDDAD[8] and MobiLib[9] that collectively hold well over 50 traces. As these are pervasively captured user information, several questions have been raised about the process of collecting traces[11], [12]. Techniques are being researched such that users themselves can share their traces[13]. However, the pertinent question, which still remains unanswered is that once traces are collected, how can they be prepared for distribution such that they have a good utility, as well as, they do not compromise the privacy of the users. Our efforts are targeted at this question, which has become even more challenging with the WLAN traces, as we shall discuss in this paper. In this work, we present our analysis of the currently used anonymization methods and their shortcomings.

The next section presents the information available in the WLAN traces. Sec. III presents example scenarios where identifying a user and monitoring his usage pattern can be detrimental to his privacy. These cases justify the need for fail-proof anonymizing/sanitizing of WLAN traces. We discuss prevalent methods of anonymizing WLAN traces in Sec. IV, following which we discuss attack scenarios and methods, which can be used to break WLAN anonymization. Sec. V presents an analysis of how the anonymization could be broken. Sec. VI provides an analysis of the attacks and discusses different possible approaches that can be used to prevent evasion of privacy, though this remains an open question. In the last section, we summarize our findings and present directions for future research.

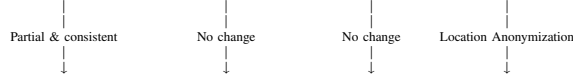
## II. INFORMATION IN WLAN TRACES

WLAN traces are logs of user association with wireless Access Points (AP). A generic information tuple, after some

processing of the raw trace, has MAC ID, Start time, Duration and Access Point/Location.

MAC	Start Time	Duration(sec)	AP/Location
00:11:22:33:44:55	01 Jun 2008 21:00:51 GMT	3000secs	CS_buildingAP1
11:22:33:44:55:66	01 Jun 2008 21:01:30 GMT	10secs	ECE_buildingAP2
01:02:03:04:05:06	01 Jun 2008 22:11:00 GMT	200secs	MSL_buildingAP1
10:20:30:40:50:60	01 Jun 2008 22:15:30 GMT	600secs	MACA_buildingAP1
11:22:33:44:55:66	01 Jun 2008 22:23:10 GMT	180secs	CS_buildingAP3

a. Sample un-anonymized trace



MAC	Start Time	Duration(sec)	AP/Location
00:11:22:0353	01 Jun 2008 21:00:51 GMT	3000secs	AcadBldg10AP1
11:22:33:0521	01 Jun 2008 21:01:30 GMT	10secs	AcadBldg2AP2
01:02:03:9877	01 Jun 2008 22:11:00 GMT	200secs	Library5AP1
10:20:30:3260	01 Jun 2008 22:15:30 GMT	600secs	AcadBldg22AP1
11:22:33:0521	01 Jun 2008 22:23:10 GMT	180secs	AcadBldg10AP3

b. Sample anonymized trace

TABLE I

WLAN TRACE SAMPLE: BEFORE AND AFTER ANONYMIZATION

A snapshot from an un-anonymized trace, is shown in Tab.Ia. Some traces may provide more information such as username. For the sake of simplicity, we have considered the basic tuple similar to shown in Tab.I. Using a tuple with less information makes the breaking of anonymity any easier as compromising anonymity with less information is more difficult.

### III. NEED FOR ANONYMITY

Although the implications of losing privacy in the real world are well known, in this section, we discuss the implications related to the loss of privacy in WLAN traces. As Tab.I shows, MAC address is one of the fields in the traces. This field is the link-layer address of the hardware/device used to access the WLAN network. Users generally do not change their MAC addresses between the sessions (perhaps due to lack of tools, which do it effortlessly or due to lack of awareness) and current protocols do not allow a user to change his MAC address during the session. This implies that MAC address becomes a permanent identifier of the machine. Since most of the machines using wireless are portable, they are less frequently shared by people. MAC address, thus, becomes associated to the person and hence his/her identifier. If we know MAC address of a device and its user, then we can search for that user in the WLAN traces and essentially know the places visited. MAC address of a device can be found by various methods such as sniffing the wireless channel.

Greenstein et al.[14], with the help of case studies, have shown how capturing and analyzing of 802.11 protocol packets can be used to evade user privacy. The cases, which we present, show similar threats as shown in this paper [14]; however, we are using only the WLAN traces and are not coupling it with actively captured data packets. In our case, threats become even more serious because the attacker need not be present in the same geographic location as the attacked/victim (traces are available on the Internet [9], [8]). Tracking the attacker can also be difficult due to the fact that some of the WLAN traces are publicly available with little or no security checks or log mechanism. Below are some cases that show possible attacks on user privacy:

Version
Header Lent
Type of Service
Identification
Flags
Fragment Offset
Time to Live
Protocol
Header Checksum
Source Address
Destination Address
Options
Data

TABLE II

SPECS OF EACH RECORD IN WIRED TRACES, BASICALLY A IP-HEADER

- 1) One can prove someone's presence at a location by showing the association of his machine with AP located in that vicinity.
- 2) If one knows MAC-to-name mapping of a user, he/she can trace the user by finding the location of AP with which the user associates. Therefore, he/she can get user's daily activity pattern/schedule (Imagine if a thief knows exactly when one is going to be away from house or in which time interval nobody is in the office).
- 3) By looking at the MAC addresses associated with a particular AP with which a user associates, one can make a guess about the people the user is meeting with. If MAC addresses to name mapping is available for all MACs, this would be a trivial task.
- 4) Information can be used as a forensic evidence against the user (or as an alibi).

These scenarios show us some of the possible privacy infringements, if the WLANs are available without anonymization. Trace providers are aware of these concerns and therefore anonymize the traces before making them public. In this study, however, we show that the anonymization techniques used can be compromised and users can be identified to some extent even after anonymization. The next section provides an in-depth discussion of the anonymization techniques used in WLAN, this would allow us to better appreciate the attack as well as the complexities involved in anonymizing the traces.

### IV. RELATED WORK

The wired network traces have existed for some time and many libraries have been created for sharing the traces[15], [16]. Researchers have developed several anonymization techniques[17], [10], [18] for wired network traces. Several tools have been developed such as Tcpcmpub[10] and Tcpcpriv[19]. We looked into these traces and techniques to investigate if they can be applied to WLAN traces. We, however, found that even though highly sophisticated techniques have been proposed to anonymize the wired traces they are not completely unbreakable[20]. In addition, there are fundamental differences between Wired and Wireless LAN (WLAN) traces, which makes it difficult to apply Wired trace anonymization on WLAN traces. In terms of anonymization goals, in wired traces, the goal is to prevent discovery of identities of network resources and leakage of security policies [10]; however, anonymization in WLAN traces, also requires protection of user's identity[14], [21] as the network resources are personal devices. Wired traces (also called netflow) have

fields as shown in Table II, which is essentially an IP header (IPv4). WLAN traces can have this information along with other information as in Table Ia, which is generated by association and disassociation of the device with the access points (AP). As this feature is unique to WLAN usage, we face newer challenges in anonymization. We can see that complete WLAN trace (along with netflow) is a super set of wired trace (only netflow). In WLANs, generally IP address are assigned using DHCP protocol and the subnet varies with WLAN access location. This reduces chances of same machine getting the same address on every session, which in wired traces can be considered 100% (assuming static assignments only). This makes anonymization of netflow information from WLAN traces much simpler than wired traces.

We find that in many studies[22], [2], [5], [3], [4] regarding WLAN traces, researchers have only used association traces such as shown in Table I. In fact, most of the WLAN trace libraries[9], [8], do not have comprehensive netflow traces as they have the association traces. One of the reason is the difference between association traces over netflow data in WLAN. Netflow information (like in wired traces) are usually used to understand the behavior of the applications[23], [24], to detect anomalies in the network[25], [26], network protocol designs, and network planning [27], [28]. Wireless traces have been used for network planning[3], [5], understanding user behavior[22], [2], [5], DTN protocol designs[4], and understanding societal interaction with technology [2].

Overall, we see that even though rich set of techniques are available for wired traces, their applicability to WLAN traces seems insufficient due to above reasons and because of similar reasons, the attacks on WLAN traces would be quite different than the attacks on wired traces.

Although anonymization is a very important step in releasing WLAN traces, we could not find any published work that deals with the techniques most suitable for WLAN. Most of the techniques used are not thoroughly investigated in the light of WLAN traces. This will be more clear in the next section where we talk about the possible attacks and drawbacks of the existing methods. Rest of this section examines the anonymization techniques currently used in WLAN trace anonymization.

**Current Techniques:**Anonymization in WLAN traces is done on field by field basis[29], [9]. Either a field is fully anonymized (mapped to a random number) or only a portion of the field is anonymized. In the traces having multiple sessions per MAC addresses, trace providers can either randomize the MAC address to a unique value for each session, or use the same anonymization mapping of the MAC address for all the sessions (consistent mapping). This step decides the information and utility of the traces. Consistent mapping for each MAC throughout the traces, provides ability to track a user through multiple sessions. Majority of the traces available at MobiLib[9] and Crawdad[8] provide the consistent mappings.

Some traces like Dartmouth traces[29] at Crawdad[8] anonymize the location field by giving a building level granularity of the AP's location or by anonymizing the building name with code names such as AcadBldg10AP3[29], which signifies an AP (numbered 3) located in a building used for academic purposes. In this case, all the buildings are grouped

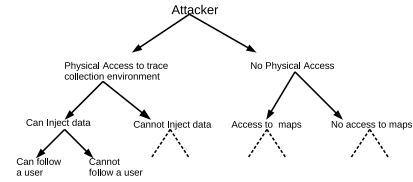


Fig. 1. Attacker capabilities

into building classes such as acadbldg, librarybldg etc. Tab.Ib shows how WLAN traces would look when anonymized for consistent and partial MAC anonymization with reduced location information. We will attempt to extract private information from traces which have been anonymized using this technique as this is used by many trace providers[29].

## V. ATTACK SCENARIOS

In this section, we present techniques where user privacy can be theoretically compromised. Fig. 1 shows attacker capabilities in terms of information related to the traces collection environment he can access. Attacker is assumed to have access to anonymized traces in all the scenarios. In this work we are, however, not dealing with all the possible scenarios as our aim here is to bring forth the shortcomings of the current anonymization, which can be achieved even if we can break the anonymization for one case. We are considering two possible attack scenarios: one where attacker can inject data into the traces by accessing the WLAN network (Sec. V-A, V-B and V-C) and second where attacker has physical access to the campus but cannot access the WLAN network (Sec. V-D). If we can identify anonymized MAC address in the traces for any user, we will consider that anonymity has been compromised. This can be justified since the main purpose of anonymization is to prevent user identification. Using this definition of compromise, we will show how an attacker can identify his own anonymized MAC address and then how can he identify any other user's MAC address.

### A. Identify Your Own MAC In Trace

Using the definition of anonymity compromise, even if an attacker can identify his own MAC address, it should be considered a failure of anonymization techniques. Although this is not a serious breach of privacy per se, yet an attacker can now use this information to find out building codes and identify MAC addresses for other users. Steps for obtaining one's own MAC address are as follows:

- 1) Go to a WLAN covered area in the campus, at a time when it is not frequently visited and the WLAN usage is minimum (find this pattern from the previous traces).
- 2) Associate with an AP belonging to campus network, and mark the start time and end time.
- 3) If there are some people around the area, move to a new location which is at least 100 ft away (beyond range of the previous wireless AP) and repeat Step 2.
- 4) Now go back to study the traces and find all the MAC addresses (anonymized though), which log-in at the same time and log-out at same time at the two locations visited.

- 5) If there are several MAC addresses, one needs to repeat this experiment from Step 1 to 4 and then take a intersection of the MACs. In the end, there should be only one MAC address left after the intersection.

This will provide ones MAC address's mapping in the traces. In Sec. VI, we mathematically show that even in a large environment (over 500 AP), at most 5 iterations of steps 1 to 4 would be enough to identify your own MAC address.

### B. Identifying Building Codes

Identifying the building codes is useful for finding users at a particular location. The attacker who knows his anonymized MAC address can visit all the buildings of interest in the campus and mark his login and logout time at each building. While looking back at the trace he can reverse map all the building codes to actual building codes/names by correlating the timings in the notes with the actual trace.

### C. Identifying A Person

Once we have the building codes, one can target a specific person, follow him and mark his device's start or end times (observing opening and closing of laptop lid). Filtering the traces with this approximate timing information and building information, one should not get many sessions. If one does then one can repeat this process and zero down to a single MAC address belonging to the target (publicly available schedules, status messages on social networking websites can also be used to find approximate login and logout timings).

To discover mapping of large number of MAC addresses to their real MAC address, one can sniff all the wireless traffic at a location (AP) whose trace mapping is already known, parsing this captured data for messages which clearly show that a machine is trying to associate with the AP [21]. In this case, we have the precise time of the user's log-in and also the MAC address with location. Identifying his anonymized MAC should be trivial. And once we know the mapping to real MAC address in the traces one can track that person anywhere on the campus.

Using the above methods, in theory, an attacker can track any person throughout the campus, causing a breach of privacy. This method presents a serious shortcoming to the prevalent methods. It shows a possibility of a privacy attack without much effort. If one does not have access to the campus Wi-Fi, one can ask a friend or one may use social networking skills to ask a complete stranger to do it. We also observe that even if the trace providers do not provide traces on daily basis, a careful planner can undertake several such experiments and then wait for the trace provider to release the trace and perform his attack.

### D. Multiple Filtering

In above described methods, the attacker has to have a capability to inject data into traces collection system (should have authorization to access the WLAN). In the current case, we consider an attacker with no ability to access (and inject data into) WLANs. He is limited to the physical access to the traces collection environment. Researchers have attempted to classify WLAN users based on their genders[2]. We extend this idea further by grouping users based on different categories like

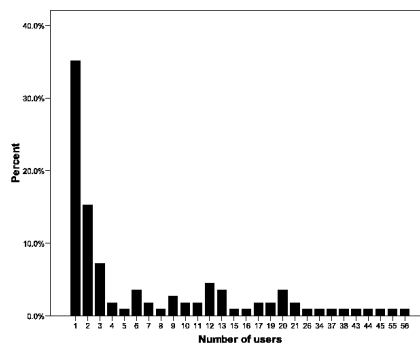


Fig. 2. Percentage of no. of users found, when 111 filters based on gender+major+manufacturer are applied

gender, login time, building, and manufacturer of the device. We, then attempt to identify users who appear under multiple categories (find intersection). In all these individual categories, the group size is large (~100). However, when we intersect the groups, this size drops rapidly. For example, female student going to Law building in the morning with an Apple computer resulted in a single user. This finding has privacy implications. Taking the above example, just by watching a female student going to a law school building with an Apple device in hand, should enable an attacker to go find the anonymized MAC ID of the student in the traces. Once it is accomplished, the attacker can trace the student's movement throughout the campus. This is a serious breach of privacy. We have conducted analysis to examine how many users can be identified using a filter using gender, study major and network card manufacturer (on a feb2006 trace downloaded from MobiLib[9]). We found that for 111 different filters (formed by different combinations of gender, study major and manufacturer), 35% resulted in a single user and 60% of the cases had less than 3 users (Fig. 2). We did the analysis for three different traces periods (feb2006, oct2006, feb2007) and found similar results. We also used different filters like gender-major-time, and again obtained a similar result. This method exposes a major flaw in the anonymization technique.

## VI. ANALYSIS AND MITIGATION

The attacks mentioned in the previous section were feasible because attacker could identify unique WLAN usage in the traces. The attacker could identify MAC address of his machine by creating usage patterns that were unique for that traces collection environment. Patterns are formed because MAC addresses are consistently anonymized. Therefore, considering all the sessions made by a device (identified by MAC address), one can identify individual usage sequences from fields in the trace like location, start time and duration. For example, a user who starts using WLAN everyday around 9 am is creating a pattern with respect to start time. This pattern may not be unique as there may be several users starting WLAN usage around 9 am. However, one can reduce the search space or may even make the pattern unique by combining location and duration patterns with start time. Consider employees working in same office space and having same office hours and

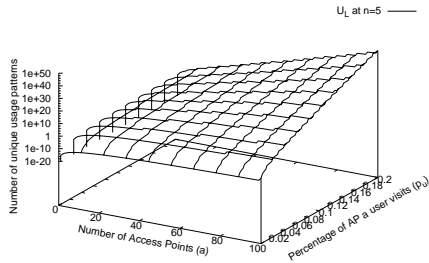


Fig. 3.  $U_L$  at  $n = 5$

work load. They would have similar start time, location and duration patterns. However, if the office and residences share a common WLAN service (say City-wide wifi or students living on-campus), the location, start time and duration of WLAN at residences would become different for all the users (unless each and every employee has the same residence and follows a similar lifestyle!). The argument here is that users can have sufficiently unique usage, which can be used to identify them even though traces are anonymized. In the next two subsections we present our reasoning in support of the above argument. We do a theoretically and a practical analysis on real WLAN traces.

#### A. Theoretical Analysis

Mathematically, it can be show that each field in the trace can create enormous amounts of patterns. For the sake of simplicity, we are only considering the patterns generated by location because similar equations can be used for other fields. Let  $U_L$  be the number of unique usage patterns possible using location field only.

$$U_L(a, p_u, n) = C_{(a.p_u)}^a \cdot (a.p_u)^n$$

where  $a$  is the total number of Access Points/locations,  $p_u$  is the percentage of total Access Points/locations a user visits,  $n$  is the number of sessions and  $C$  denotes the combination function. Fig. 3 shows the distribution of  $U_L$ .  $U_L$  is a product of the number of ways  $a.p_u$  Access points can be selected out of total  $a$  Access Points ( $C_{(a.p_u)}^a$ ) with number of ways in which  $a.p_u$  Access Points can be selected in  $n$  sessions ( $(a.p_u)^n$ ).

As an example, consider a university campus having hundreds of buildings, say University of Florida (UFL), which has over 500 hundred wireless access points, so we can have 500 different values in the location field. It has been shown, that users generally use less that 5% of the Access Points 90% of the time [30], [22]. Therefore, in our case ( $a = 500$  APs), we assume each person uses only 5% ( $= p_u$ ) of them ( $a.p_u = 25$ ). Because in a pattern not only visiting a location but also the order of visiting a location is important, we can see that total number of combinations of APs people can choose from is  $C_{25}^{500} \sim 10^{46}$ . Assuming that traces contain only 5 sessions per user ( $n = 5$ ), the total number of paths possible for a user, using 25 APs, is equal to  $25^5 = 9765625 \sim 10^6$ . Therefore, the total number of unique location pattern possible,  $U_L$  is  $\sim 10^{46} \times 10^6 = 10^{52}$ . Total number of students at

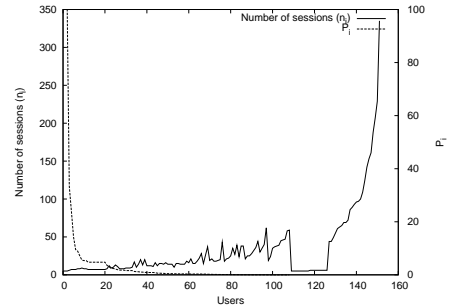


Fig. 4. Results of the combination generation and sequence matching for randomly chosen 230 users out of 27K users belonging to the month of Nov 2007. This graph shows  $P_i$  and  $n_i$ .

$UFL \sim 5 \times 10^4$ . So, theoretical number of unique location pattern per user  $= 20 \times 10^{46}$ . Even though this is a very loose upper bound and in reality this number can be smaller, what it shows is the enormous number of possible unique patterns that can be generated using just one field (location). This implies that theoretically every user can have a unique pattern in a short time, which can be used to identify him. This further implies that sanitization techniques cannot work well, if only the fields are anonymized; one should aim to anonymize the patterns. One of the ways is to use inconsistent MAC anonymization, which is extremely detrimental to the utility of the traces, the very reason traces are shared. A fundamental question about the relationship between the utility and the anonymization/privacy is evident here, which we plan to discuss in our future works.

#### B. Practical/Trace Analysis

To check the validity of the theoretically limits discussed above, we did an experiment on WLAN traces coming from UFL for a period of one year. Tab. III has the findings for users having same location visiting sequence. We calculate and distinguish users based on location field using Longest Common Subsequence algorithm [31]. We find the number of users having similar location visiting pattern with at least one other user, considering several time periods (1 day to 1 year), listing total WLAN users in that specific period. Tab. III also shows number of users who had number of sessions greater than 1 and 5. Results support our insight behind the theoretically limits. We notice, that for a period of one year, only 4880 users had a similar location visiting sequence with one or more users out of  $\sim 52K$  users (9%), if we consider 100% match. This means that almost 91% of users have distinct location visiting sequence and an attacker following a user can later identify him/her in the traces with probability greater than 0.9. Another result that further supports the above statement is that only 235 users (0.45%), who have same location visiting sequence with other users, have more than 5 sessions (in case of 100% match score). This further strengthens the theoretical limit we discussed earlier (to make it more interesting we found that most of these users had logged in to the same access point throughout their multiple sessions).

We also attempt to identify the source of these sequences, which become unique in a short time span. We note that not

	Period	(5 Nov 2007)	(5 to 11 Nov 2007)	(5 to 18 Nov 2007)	(Nov 2007)	(Aug to Dec 2007)	(Aug 2007 to Jul 2008)
	Total Users	9844	17602	22333	27068	47766	52217
100% match score	users	4288	4847	4969	4461	4288	4880
	> 1 session	1477	1872	2061	1928	1840	2186
	> 5 sessions	31	121	108	131	187	235
90% match score	users	4291	4494	5300	4879	4743	5486
	> 1 session	1480	2018	2391	2345	2294	2791
	> 5 sessions	34	268	439	548	642	839
80% match score	users	4473	6068	6924	6872	7484	8954
	> 1 session	1662	3092	4015	4339	5036	6260
	> 5 sessions	113	1085	1777	2272	3057	3930

TABLE III  
RESULT OF FINDING USERS WITH SIMILAR LOCATION VISITING SEQUENCES WITH VARYING DURATION OF THE TRACE

only each field can be used to form unique sequences but several fields may be combined to form unique sequences. We generate various sequences using several combinations of location field for a user, maintaining the temporal ordering in the combinations. This helps us to identify how much information an attacker may obtain about a user, even if the attacker follows him for only a few sessions. Because of this, attacker would find information holes in the observed sequence for the user. For example, he may be able to observe only 2<sup>nd</sup>, 3<sup>rd</sup>, 6<sup>th</sup>, 8<sup>th</sup> and 10<sup>th</sup> sessions of a user. We investigated 230 randomly selected users from a set of 27K users appearing in Nov 2007 WLAN traces from UFL. For each user, we created all the possible combinations of sequences of length 5 using Location field, maintaining temporal order (earlier we saw that users with number of sessions greater than 5 sessions have higher chances of being unique). Each combination represents a possible set of sequence an attacker may be able to capture by following a user, assuming attacker may not be able to capture all the user sessions. This simulates loss while capturing user information. Then we search for these sequences in traces belonging to all 27K users. Let  $P_i$  be the percentage of matches for user  $i$ , where  $P_i$  is defined as  $P_i = M_i/C_5^{n_i}$ . Here  $C_5^{n_i}$  represents the total number of combination of sequences possible of length 5 for user  $i$ ,  $n_i$  is the number of sessions for user  $i$  and  $M_i$  represents the number of matches found for  $C_5^{n_i}$  sequences in the trace belonging to 27K users. Fig. 4 shows the results for this experiment. We find that out of 230 users, 78 had less than 5 sessions in the whole month and were discarded. For the rest of the users we plot  $P_i$  in descending order along with  $n_i$ . One interesting result is that even when the total number of combinations generated is very high ( $n_i = 100$ ,  $C_5^{n_i} = 75287520 \sim 10^7$ ) and the number of matches is very low ( $M_i = 81$ ). This indicates that if the location information of 5 sessions is available in temporal order with many intermittently missing location information of a user, even then there is a very high chance of identifying the user in the trace.

As per the analysis we conducted, there can be two ways of mitigating the attacks discussed in the previous section. One is to manipulate the traces in such a manner that no one can identify unique patterns and the other is to prevent linking of usage patterns to users. Both these abstract ideas can be applied to the traces independent of each other. If one can identify usage pattern, but cannot assign it to a specific user, one can never be sure of identifying the correct user or the correct pattern of the user. On the other hand, if we can prevent linking of usage patterns to users, then no matter how many unique usage patterns one can identify, one would not be able

to link it back to a user. Both methods should individually provide sufficient privacy for the users. For the first method many techniques exist in literature such as  $k$ -anonymity [32] or  $l$ -diversity [33]. For the second method, we need to devise techniques, which can obscure linking information.

## VII. CONCLUSIONS AND FUTURE WORK

We have uncovered a serious problem in the way WLAN traces are anonymized. We believe that this kind of attack is possible as WLAN traces have human behavior pattern embedded in them, which can be easily observed by an attacker following the victim. The aim of any privacy protecting technique should be to ensure that even if attacker has access to all the publicly available information about a user or a group of users (but not the mapping between anonymized MAC and real MAC), he should not be able to reduce the sample size below a number, say  $K$ . This  $K$  should be a parameter configurable by the trace releasing authority.

In the future, we plan to work on the feasibility of anonymizing using techniques like perturbations and release of traces in multiple different formats like one with no location or time information. We would also like to investigate in further details how the fields like start time, duration and locations are responsible for generating unique patterns. It may be due to the atomic properties of these fields like periodicity and history. We would like to work on a system, which can generate anonymized traces according to the security clearance of the demanding user, this would allow us to serve traces with varying anonymization and privacy criterion and would make traces more useable. We also plan to investigate, if  $k$ -anonymity model [32] can be applied to WLAN trace.

Findings in this paper certainly call for a new research in the area of WLAN trace anonymization and privacy, details of which are to be pursued in our future work.

## REFERENCES

- [1] W. Hsu, D. Dutta, and A. Helmy, "Mining behavioral groups in large wireless lans," in *MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. New York, NY, USA: ACM, 2007, pp. 338–341.
- [2] U. Kumar, N. Yadav, and A. Helmy., "Gender-based grouping of mobile student societies," in *The International Workshop on Mobile Device and Urban Sensing (MODUS), IPSN 2008*, [http://www.motorola.com/innovators/ModusWorkshop/Gender\\_Based.pdf](http://www.motorola.com/innovators/ModusWorkshop/Gender_Based.pdf).
- [3] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," in *Proc. ACM MobiCom '04*, September 2004.
- [4] W. Hsu, D. Dutta, and A. Helmy, "Profile-cast: Behavior-aware mobile networking," in *Proc. IEEE WCNC*, Las Vegas, NV, March 2008.

- [5] G. Chen, H. Huang, and M. Kim, "Mining frequent and periodic association patterns," Dartmouth College, Computer Science TR2005-550, 2005.
- [6] M. Musolesi and C. Mascolo, "A community based mobility model for ad hoc network research," in *REALMAN '06: Proceedings of the 2nd international workshop on Multi-hop ad hoc networks: from theory to reality*. New York, NY, USA: ACM, 2006, pp. 31–38.
- [7] W. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling time-variant user mobility in wireless mobile networks," in *Proc. IEEE INFOCOM*, May 2007.
- [8] "CRAWDAD: Community Resource for Archiving Wireless Data at Dartmouth," August 2008. [Online]. Available: <http://crawdad.cs.dartmouth.edu/data.php>
- [9] 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/>
- [10] R. Pang, M. Allman, V. Paxson, and J. Lee, "The devil and packet trace anonymization," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 1, pp. 29–38, 2006.
- [11] D. C. Sicker, P. Ohm, and D. Grunwald, "Legal issues surrounding monitoring during network research," in *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2007, pp. 141–148.
- [12] M. Allman and V. Paxson, "Issues and etiquette concerning use of shared measurement data," in *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA: ACM, 2007, pp. 135–140.
- [13] K. Shilton, J. A. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Participatory privacy in urban sensing," in *MODUS: International Workshop on Mobile Device and Urban Sensing*, St. Louis, MO, USA, 2008.
- [14] B. Greenstein, R. Gummadi, J. Pang, M. Y. Chen, T. Kohno, S. Seshan, and D. Wetherall, "Can ferris bueller still have his day off? protecting privacy in the wireless era," in *HOTOS'07: Proceedings of the 11th USENIX workshop on Hot topics in operating systems*. Berkeley, CA, USA: USENIX Association, 2007, pp. 1–6.
- [15] "The Skitter Project," June 2008. [Online]. Available: <http://www.caida.org/tools/measurement/skitter/>
- [16] "The Passive Measurement and Analysis Project," June 2008. [Online]. Available: <http://pma.nlanr.net/>
- [17] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon, "Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme," in *Computer Networks*, 2002, pp. 280–289.
- [18] J. C. Mogul and M. Arlitt, "Sc2d: an alternative to trace anonymization," in *MineNet '06: Proceedings of the 2006 SIGCOMM workshop on Mining network data*. New York, NY, USA: ACM, 2006, pp. 323–328.
- [19] G. Minshall, "Tcprpriv, 1996."
- [20] S. E. Coull, C. V. Wright, F. Monrose, M. P. Collins, and M. K. Reiter, "Playing devils advocate: Inferring sensitive information from anonymized network traces," in *Proc. of the 14th Annual Network and Distributed System Security Symposium*, 2007, pp. 35–47.
- [21] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, "'802.11 user fingerprinting," in *MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking*. New York, NY, USA: ACM, 2007, pp. 99–110.
- [22] W. Hsu and A. Helmy, "On modeling user associations in wireless lan traces on university campuses," in *Proc. of International Workshop on Wireless Network Measurement (WiNMe)*, April 2006.
- [23] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos, "Is p2p dying or just hiding? [p2p traffic measurement]," *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, vol. 3, pp. 1532–1538 Vol.3, Nov.-3 Dec. 2004.
- [24] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the slammer worm," *Security & Privacy, IEEE*, vol. 1, no. 4, pp. 33–39, July-Aug. 2003.
- [25] "Predict: Protected Repository for the defense of the Infrastructure Against Cyber Attacks," June 2008. [Online]. Available: <http://www.predict.org>
- [26] B. Zdrnja, N. Brownlee, and D. Wessels, "Passive monitoring of dns anomalies," in *DIMVA*, 2007, pp. 129–139.
- [27] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 835–846, 1997.
- [28] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring isp topologies with rocketfuel," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 2–16, 2004.
- [29] D. Kotz, T. Henderson, and I. Abyzov, "CRAWDAD data set dartmouth/campus (v. 2007-02-08)," Downloaded from <http://crawdad.cs.dartmouth.edu/dartmouth/campus>, Feb. 2007.
- [30] M. Balazinska and P. Castro, "Characterizing mobility and network usage in a corporate wireless local-area network," in *MobiSys '03: Proceedings of the 1st international conference on Mobile systems, applications and services*. New York, NY, USA: ACM, 2003, pp. 303–316.
- [31] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (second edition ed.)*. MIT Press and McGraw-Hill, 2001, pp. 350–355.
- [32] 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.
- [33] A. Machanavajjhala, J. Gehrke, and D. Kifer, "l-diversity: Privacy beyond k-anonymity," April 2006, pp. 24–24.