How Private are WLAN traces?

Udayan Kumar(Student) and Ahmed Helmy(Faculty)

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

I. Introduction

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. Why do we need to talk about privacy of users in WLAN trace? One of the answers is that many researchers use WLAN trace for analysis and research purposes, such as, to find out usage behavior of users[3], [6] or to study user mobility patterns[1] and characteristics for developing network protocols. Therefore, it is important to understand how the privacy of WLAN users gets affected. Even though, most of the trace libraries anonymize/sanitize the traces to protect user's privacy, we present few methodologies, which can be used to reverse the anonymization. We hope that our study sheds light on the question of "How Private are WLAN traces" and how effective are exisiting anonymization techniques.

The issues of privacy and anonymization have been present timelessly for the network traces. Researchers have faced similar challenges in anonymizing the wired traces[7]. Recently, wireless traces have also been collected and archived at on-line public libraries like CRAWDAD[2] and MobiLib[4] that collectively have well over 25 traces. As these are pervasively captured user information, several questions have been raised about legality[9] of the process of collecting traces. Techniques are being researched such that users himself shares his traces[8].

The pertinent question, however, which still remains unanswered is that once traces are collected, how can they be prepared for distribution such that they have a good usability and do not compromise on privacy of the user. Our effort is targeted at this question, which has become more challenging, as we shall see, with the WLAN traces. In this work, we present our analysis of the currently used anonymization methods and their shortcomings.

II. Information In WLAN Traces

WLAN traces are the logs of users associating with wireless Access Points(AP). A generic information tuple that they provide has MAC ID, Start time, Duration and Access Point/Location. A snapshot from an un-anonymized trace,

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	ECE_buildingAP3

TABLE I
Sample Un-Anonymized trace

after some processing, is shown in Tab.I. Some traces may provide more information like username, etc. For the sake of simplicity, we have considered the basic tuple similar to shown in Tab.I. Having a tuple with less information does not make the breaking of anonymity any easier, as compromising anonymity with lesser information, is more difficult.

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	AcadBldg2AP3

TABLE II
Sample Anonymized trace

III. Prevalent Methods of Anonymization

Anonymization in WLAN traces is done on field by field basis[5], [4]. 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, 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 also 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[4] and Crawdad[2] provide the consistent mappings.

Some traces like Dartmouth traces[5] at Crawdad[2] also 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[5], which signifies an AP(numbered 3) located in a building used for academic purposes. In this case, all the buildings are grouped into building classes such as acadbldg, librarybldg etc. Tab.II shows how Tab.I 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[5].

IV. Analysis of Prevalent Methods

In this section, we present some techniques where user privacy can be theoretically compromised. We are considering two possible attack scenarios: one where attacker can inject data into the traces by accessing the WLAN network (Sec. IV-A, IV-B and IV-C) and second where attacker has physical access to the campus but cannot access the WLAN network(Sec. IV-D). So, how do we decide if the anonymization is compromised? If we can identify someones anonymized MAC address in the traces, we can then be sure that anonymity has been compromised. Using this definition of compromise, we will show how to identify ones own anonymized MAC address and then how to identify any other user's MAC address.

A. Identify Your Own MAC In Trace

Breaking the whole anonymization scheme can start by finding out mapping of ones own machine's MAC address. To obtain mapping of ones own MAC address, one can use the following scheme:

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.

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 in the campus and mark his login and logout time at each building. While looking back at the trace one can reverse map all the building codes to actual building codes/names by coorelating 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 victim.

Using the above methods, in theory, a 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 doesn't 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

We have found another technique which can break anonymity of a user. Researchers have attempted to classify WLAN users based on their genders[6]. We extend this idea further by grouping users based on different categories like 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 categories, the group size was large ($\sim$100). However, when we intersected the groups, this size dropped 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 a attacker to go back to the traces and find the anonymized MAC ID of the student. Once it is accomplished, the attacker can trace the student's movement throughout the campus (if the building codes anonymized, attacker can use technique in Sec. IV-B to find the code). This is a serious breach of privacy. We did an analysis to see how many users we can find out using one such filter which is gender with major and manufacturar(on USC feb2006 trace downloaded from MobiLib[4]). We found that for 111 resulting criterion Fig. 1, 35% resulted in a single user and 60% of the cases had less than 3 users. We did the analsis for three different traces periods(feb2006, oct2006, feb2007) and found very similar results. We also used different filter like gender-major-time, and again obtained a similar result. These example shows how easily one can infringe upon someones private life.

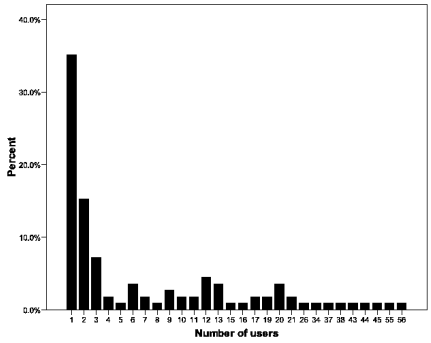

Fig. 1. Percentage of no. of users found, when gender+major+manufacturer filter is applied

V. Conclusions and Future Work

We have discovered a serious problem in the way WLAN traces are anonymized. We believe that this kind of attack is possible as WLAN traces have human behaviour pattern embedded in them, which can be easily observed by an attacker by following the victim. The aim of any privacy protecting technique should be to insure 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 future, we would want to work on the feasibility of anonymizing using other 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 a unique patterns. It may be due to the atomic properties of these fields like periodicity and history. We would also 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[10] can be applied to WLAN trace.

References

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

[2] CRAWDAD: http://crawdad.cs.dartmouth.edu/data.php.

[3] W. Hsu, D. Dutta, and A. Helmy, "Mining Behavioral Groups in Large Wireless LANs" ACM MobiCom, Sep. 2007.

[4] W. Hsu and A. Helmy, MobiLib USC WLAN trace data set. Downloaded from http://nile.cise.ufl.edu/MobiLib/USC_trace/

[5] D. Kotz, T. Henderson and I. Abyzov. CRAWDAD data set dartmouth/campus (v. 2007-02-08), Feb. 2007.

[6] U. Kumar, N. Yadav, A. Helmy."Gender-based Grouping of Mobile Student Societies," MODUS, IPSN workshop, Apr. 2008.

[7] R. Pang, M. Allman, V. Paxson and J. Lee, "The devil and packet trace anonymization," *ACM SIGCOMM Computer Communication Review*,vol. 36, issue 1, pp.29–38, Jan 2006.

[8] K. Shilton, J. Burke, D. Estrin, M. Hansen, and M. B. Srivastava, "Participatory Privacy in Urban Sensing," *MODUS*, IPSN workshop, Apr. 2008.

[9] D. C. Sicker, P. Ohm and D. Grunwald, "Legal issues surrounding monitoring during network research," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement,*, pp.141–148, 2007.

[10] 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, 2002.