==========

Sigcomm2013 Review #64A
Updated Wednesday 27 Mar 2013 6:00:42pm CDT

Paper #64: Mosaic: Quantifying Privacy Leakage in Mobile Networks

---

Overall merit: 4. Accept (top 15-5%)

===== Paper summary =====

This paper presents a study that quantifies the amount of information that can be compiled on individual users from data traces that can be collected in cellular networks. The authors use two data sets and demonstrate that first, many user identities and critical identifying information can be learned from the urls and other data contained in the traces. Then, the authors reconstruct a user profile, called a mosaic, which includes attributes of personal information such as political views, browsing habits and favorite apps. Based on this ability, the authors use the trace to quantify the amount of privacy leakage as a function of the duration of the vulnerability as well as the number of compromised IPs.

===== Paper strengths =====

This is an extremely well-written paper. It is easy to read, yet intellectually deep, and yields novel insights into the privacy (or lack there of) of users online.

===== Paper weaknesses =====

There are a few cases where something is introduced in advance of it's explanation, slightly decreasing the readability of the paper. However, this is a minor an easily fixable characteristic.

===== Comments for author =====

This paper was enjoyable to read and was very informative in the quantification of privacy leakage. A few minor points to improve the paper:

There are a number of instances where information could be explained earlier. For instance:
- define CSP in the abstract
- section 3.1 - the last paragraph is unclear the first time it is read. If there are 12,420 unique identifiers in 166,441 sessions, what happens in the other 98.7% of the sessions? It isn't until later in the paragraph that it is clear that these other sessions do not contain unique identifiers.
- similarly, in section 3.2, state earlier what a "short period of time" is.
- etc.

Otherwise, an excellent paper. I recommend acceptance.

==========

Sigcomm2013 Review #64B
Updated Wednesday 6 Mar 2013 3:55:19am CST

Paper #64: Mosaic: Quantifying Privacy Leakage in Mobile Networks

---

Overall merit: 2. Weak reject (top 60-30%)

===== Paper summary =====

The paper presents a methodology to correlate a user's offworld identity (via an identifier used by OSNs) with online behavior called Tessellation. The methodology runs on mobile network traffic that an ISP/CSP or a cyber criminal can get access to. The methodology is used to highlight privacy leakage by constructing a 'mosaic' of the user; associating behavioral information like shopping habits, political views etc to real names.
The methodology is tested out on real traces and the authors claim that 50-80% of the traffic can be attributed to real users with low error rate.

===== Paper strengths =====

Mobile privacy is an important topic. Correlating multiple sources of information -- via OSNs and browsing information and associating real identity is a useful exercise to raise awareness about privacy concerns.

===== Paper weaknesses =====

Not clear why the claimed privacy "leakage" is indeed a leakage. For Tessellation to work one needs access to the entire traffic of a mobile operator. The reported leakage is not something that an isolated hacker can pull off.

Not clear what is novel here compared to existing work in the field (eg work of Krisnamurthy et al.)

The writing can be improved -- the paper reads like a list of percentages and fractions with little insights to be gained.

===== Comments for author =====

Summary

=======

I enjoyed reading this paper, the first part primarily, the mosaic part does not come through as being coherent, looks more like an effort to add pages to the paper. I would be more positive about this paper if I felt that Tessellation can be achieved by partial access to mobile traffic as opposed to having access to an entire 3 hour long traffic dataset obtained from the operator. Having access to the raw traffic hardly qualifies as leakage. What is the attacker model here? A rogue employ of an ISP? Well, there is law to guard against that. I don't think that is anywhere implied that the current mobile Internet technology is designed to offer privacy protection from rogue employees. A rogue employ can do simpler things to link with the offline identity of a user, eg, he can go directly to the Radius server and get the associations between assigned IP address and customer Id. This is much simpler than running Tessellation.

In that sense I feel that the claimed leakage is exaggerated. I don't see how a small group of hackers that eavesdrop on a subset of base-stations can perform Tessellation with the same success that you can achieve by having the raw traffic.

Details

========

- The use of OSN ids as an 'anchor' point is understandable and useful.
However, not sure how easy it is to get this information -- many OSNs

are going for full HTTPS, including the header. Twitter for instance is full HTTPS.
Under this scenario, not sure how easy it is for an adversary to get access to
the headers that contains such information. It may be useful to do a spot
analysis on popular services today (FB, Twitter, gmail etc) and show which
of these services do not have their headers encrypted.

- The metric for accuracy is suspect. You always mention % of traffic 'attributed'
but this isn't the same as accuracy.
It is good that the information gain via Tessellation is more than that of
just looking up public profiles of OSN users (Fig. 2). However, most of the
other information was gathered via HTTP headers -- device info etc. and tying it to
an OSN id via Tessellation. This is proposed as the main contribution and hence
it is crucial that the method has a meaningful metric of accuracy.
This brings us to:

- It isn't clear how many users you were able to actually 'de-anonymize'
and you alway bring up success in terms of fraction of traffic attributed.
From a privacy perspective, the former is more important than the latter.
Case in point, sec 2.1, you say 2.4% of the traffic was identified. How many users
does this translate into? What percentage of users?
In sec 3.1, you claim OSN1 ID was 1.3% of all sessions and 1.0% for OSN2 ID. Again
what number of users are we talking about here? This is somewhat frustrating.
At a later stage (sec 3.5), you present the number of 15.7% of users in this
ground truth dataset. So this would be the upper bound on the number of users
you can de-anonymize. Would this be accurate? How did you increase this number to 43.2%?
The traffic markers are associated with an OSN id themselves, from Sec 3.3, 3.4.

- I liked the fingerprinting idea to deanonymize. Wonder how this connects to
the deanonmyization of sparse datasets by Narayanan et al.

- In general it would be better motivated if you gave real examples of
third parties getting access to this information. Clearly, the adversaries
are not service providers like Google etc nor are then ISPs/CSPs. You claim they
are rogue employees at ISPs/CSPs or a state agent in authoritarian govt. but such
cases can get access to more sensitive information in an easier way (billing information etc.)
You do make the assumption that CSP data can be collected but citing a few cases on how easy this is
can make your case stronger.

- Not sure how Table 7 is a contribution -- if we assume user's browsing behavior is very regular
and periodic, the result doesn't seem to add much at all.

- This may be hard but can you please cut down on the number of places you mention
fractions, percentages etc? It is overwhelming as there is hardly a sentence where
there is no number to be mentioned. I find too much emphasis on reporting statistics
has come at the cost of describing insights and takeaways.

Typos:

- Intro: can be ran on any network -> can be run on any network
59 categories on user demographics -> replace 'on'

===== Comments to address in the response =====

Can an outsider perform Tessellation without having direct access to ISP wide traffic data?

==========

Sigcomm2013 Review #64C
Updated Saturday 16 Mar 2013 2:50:10am CDT

Paper #64: Mosaic: Quantifying Privacy Leakage in Mobile Networks

---

Overall merit: 4. Accept (top 15-5%)

===== Paper summary =====

The paper shwos that the increasing prevalence of online social networks from mobile devices---and the leaks of identifiers from these applications---results in privacy leaks that can allow a user to be mapped to their mobile device and tracked. The attacks in the paper allow up to half of the network traffic traces to be attributed to the names of users. The paper shows how the attack can reconstruct a user profile and map that to both traffic flows and mobility. The reconstruction proceeds in two parts: the first attributes traffic to unique users via leaked social identifiers; the second reconstructs a user profile by correlating information from various sources.

===== Paper strengths =====

This paper represents an important wake-up call for privacy in mobile online social networks. The paper demonstrates a new class of attacks that correlates related pieces of information to reconstruct a user profile.

===== Paper weaknesses =====

Perhaps slightly out of scope for SIGCOMM, since the paper is really more about privacy and mobility than communications. Nevertheless, there's still a significant component of the paper that involves communications and traffic analysis, so I think it's fine. In short, no real major problems.

===== Comments for author =====

This is a deep, interesting paper that reveals a new class of attacks against users of online social networks on mobile devices. The paper shows how data gleaned from various sources can be used to reconstruct a user profile, and how that user profile can then be tracked, both in terms of traffic patterns and mobility. Of course, this type of correlation attack is known to be possible in theory, but the paper demonstrates the attack in practice.

Something that I must have missed in the attack, but seems like a rather obviosu fix: Wouldn't sending the traffic over an encrypted channel (e.g., SSL) prevent the data leaks that the paper describes? In that case, only the online social networks would have access to various the various "leaked" information. Perhaps the adversary has a capability or vantage point that I somehow missed, but it seems relatively

straightforward to prevent leaks of the OSN data that is outlined in the paper. The paper talks about gleaning this information from OSN profiles, but wouldn't most users also restrict their profiles so that these details are only visible to their friends (orm couldn't they do so to stymie this attack?). I must have missed something here. The paper briefly addresses this in Section 5.4, but I sitll can't quite figure out why encryption wouldn't just make these attacks infeasible.

In general, the paper's exposition could be made more clear with more precise statements about the goals and capabilities of the attacker. What traffic is the attacker capable of seeing, and from what sources? Is collusion necessary? To what extent does encryption or "noising" of various information frustrate the attack?

Similarly, the paper could be more clear with respect to the attacker's goals. Is the goal to deanonymize a user? Track a user? Discover a user's profile and interests? It was not entirely clear what aspects of privacy the attacker is attempting to compromise. The paper would benefit a lot from a section that clearly outlines the adversary's capabilities and goals (pretty standard for a security paper).

==========

Sigcomm2013 Review #64D
Updated Thursday 25 Apr 2013 5:30:36pm CDT

Paper #64: Mosaic: Quantifying Privacy Leakage in Mobile Networks

---

Overall merit: 3. Weak accept (top 30-15%)

===== Paper summary =====

The authors describe privacy threats that arise when it is possible to combine network traffic data from mobile phones together with some leaked OSN identifiers. The paper illustrates that it is fairly straightforward to associate traffic data without an ID to a specific user via heuristics such as “close in time” or “top-k web services used”. They are able to identify the users for about 50% of the network sessions (batched mobile traffic) in their ground truth dataset. After having identified the user (OSN ID) for many blocks of traffic, the authors illustrate how they can build a rich profile (called a “mosaic”) by supplementing their datasets with information obtained by crawling OSNs.

===== Paper strengths =====

Illustrating how data can be combined and the resulting privacy threat is an important topic, and provides important warnings for how OSNs are designed, Web services and mobile users. They demonstrate the richness of the profile that can be built for a person.

They have great data; the authors use network packet traces from 2 cellular service providers (CSP), spanning 65,000 clients; plus another dataset used as ground truth because it has user IDs in it derived from the CSP’s authentication protocol.

===== Paper weaknesses =====

While this is an important threat, the underlying methodology seems fairly straightforward. More importantly, the paper somehow doesn’t come together with a set of clear messages. Their message is garbled by lots of miscellaneous stuff in the paper that isn’t so interesting (Figure 10 on leakage of

location data – everyone knows this is possible, what's the point here?, why is Table 8 interesting?) Somehow everything isn't tied together, just a big statement about "wow, look what can be assembled about you".

In the end, I'm not sure they do what the title implies – privacy is not quantified on a per person basis but rather across the entire set of users in a mobile network. Overall, they do a lot more illustration than quantification as there are no new privacy metrics proposed.

===== Comments for author =====

They develop a method for associating network traffic to an OSN ID in 3 ways. First, if data from one session has an associated OSN ID (leaked explicitly or implicitly) and another session has no OSN ID, and these two sessions occur closely spaced in time, the second is assigned the ID of the first session. Second, if two blocks of network traffic data (one with an ID and one without ID) have the same "traffic markers", then they are associated to the same OSN ID. A traffic marker is stuff inside a cookie (strings such as "utmcc"/"udi" plus URL domains). Third, they count the "most frequently accessed web services" to create a fingerprint for each traffic block and associate a traffic block without an OSN ID to one with an OSN ID if the fingerprints are close.

While the authors illustrate the potential threat, they don't carefully discuss how easy or widespread such privacy attacks could be. Specifically, the analysis presented requires the use of network traces from cell service providers. So the CSPs themselves could do this mosaic building, but for fear of bad reputation and in order not to lose customers, they might restrain themselves. Hence the main threat is by an adversary who is able to collect mobile cellular network data illegally. I don't think this is so easy; even if an attacker could do it on one link, they aren't going to be able to penetrate a huge numbers of links, and thus the number of users affected will be limited (fewer users to match fingerprints against). I do agree that a rogue employee in a CSP is the best example.

Table 6 is very nicely presented. I liked that you can see what information is available only in OSN data, which information comes just from trace analysis, and which data comes from both.

Overall there is a lot of language employed that seems more grandiose that the work really is. Lots of comments about "combining static and dynamic data", but there aren't any really interesting specific examples of this. Sure your gender is static and your URL visits are dynamic, but there is nothing profound about observing that URL visits are dynamic that really influences your system. For example, does dynamic data need to be processed differently – to assess accuracy? Consistency? How many visits to a domain do you need to see before you decide it isn't ephemeral? (You do something else with persistency, looking for persistent markers rather than persistency of user activity.) You also talk about "user activity analysis" generally, but in fact the only activity captured is top-k services visited. It doesn't seem you are using "time-spent" (even though you use that language). Or did I miss something?

You say that can roughly attribute 50% of the sessions to the right user ID. That's pretty good. But what happens to the other 50% - do these result in associations with the wrong user ID? What are the implications – if this is being done by an adversary, then he won't care; but if it is being done by the CSP for the purposes of profiling users to provide better ads, then getting it wrong so often is problematic. (In all fairness, I guess this latter example is not the case you are focusing on.)

Section 5 entitled "Quantifying privacy leakage" is a title that sets the reader up for disappointment. Other than counting the number of users whose OSN ID leaks within an hour or two, you don't really *quantify* privacy leakage – you just give examples. With a title like that I had hoped to see something numerical about an individual. For example, "out of X pieces of PII in my profile, y% of them leaked before combining all the data, and z% leaked after combining all the data sources". Figure 12 shows information learned about an ensemble of all your users. It does NOT show

information leaked per individual. The amount of information leaked per individual will differ and that is hidden in the way you have plotted this (summing over all users). Fig 12d could be learned by Pandora.com and they wouldn't need cell data to do that. If you can identify 50% of users, does that mean that you can learn the shopping profile of 50% of users? Not exactly, because some of those users might not be shoppers. I wonder if there is another way to show the data in Fig 12 that focuses on a privacy leakage metric for individuals?

The authors should refer to the CCS 2012 paper that deanonymizes mobility traces using OSN data: "Deanonymizing Mobility Traces: Using a Social Network as a Side-Channel", by Mudhakar Srivatsa and Michael Hicks. Please clarify the difference between your technique and theirs.

You don't explicitly state the <key,value> pairs used – which is which in your tables?
What does the index 'j' refer to in the definition of s_i^j ?

You say that only the top 5 domains visited are needed to uniquely identify an individual. Wow! That isn't much (perhaps it is because we are talking about mobile users). Why not use 8, that looks even better. How much longer is the observation period between 5 and 8 – is it significant?

The most compelling example of a real threat enabled by tessallation is the one in section 5.3 learning the age of app users. That is a nice and concrete example!

==========

Sigcomm2013 Review #64E
Updated Saturday 13 Apr 2013 12:58:19am CDT

Paper #64: Mosaic: Quantifying Privacy Leakage in Mobile Networks

---

Overall merit: 2. Weak reject (top 60-30%)

===== Paper summary =====

The main purpose of this paper is to investigate the privacy leakage problem in the mobile network data. The authors design a system to illustrate the attribution of users information.

===== Paper strengths =====

THe strength is in design a system to demonstrate how easy it is to reveal users' privacy information for OSN.

===== Paper weaknesses =====

The fact that OSN and mobile devices have major "privacy" concerns is not new. I believe the authors design an interesting system (and experiments) to illustrate the attribution problem.

===== Comments for author =====

The main thesis of this paper is in demonstrating the privacy leakage problem in the mobile network data.

The author claim that their first key "insight" is the prevalence of OSN usages leaves identifiable digital footprints. This to me is not really a major contribution for it is well-known in the public.

The second insight provided by the authors is in the association between users and mobile devices enables the attribution of traffic to the users. This allows one to Tattribute significant portions of traffic with NO leaks to the users' true identities.

For me, the major bulk of the work is in the design a system and demonstrating the feasibility of the threat. The authors demonstrated that 50% of the traffic can be attributed to the leaked OSN identifiers with high confidence. The authors also showed that different types of information can be gleaned about any user by providing a digital mosaic. The information collection is interesting but I am not sure about overall novelty.

Frankly speaking, i am not sure this paper will have a chance of being accepted in Oakland conference or ACM CCS.

==========

Author's Response by ningxia2015@u.northwestern.edu
Paper #64: Mosaic: Quantifying Privacy Leakage in Mobile Networks

---

1. We agree with the reviewers that it is useful to include a description of the attack model in the paper. Next we explain the feasibility of the attack and the goals of the attacker.

FEASIBILITY OF ATTACKS
As correctly pointed out by reviewer 2, a rogue employee inside the origin CSP (ISP) of a user can directly acquire Rada IDs. However, the rogue employee may be present in any of the transit ASes and still have access to traffic data, but not to Rada IDs since they do not pass the borders of the origin CSP. In our paper we show that raw network traffic contains enough information to attribute traffic to real-users.

Government agencies have a similar observation point when they cannot directly collaborate with an origin ISPs (e.g., CIA vs an Iranian ISP). Agencies can acquire raw traffic data from a transit country or even re-route traffic by launching a BGP/IP hijacking attack.

GOALS OF ATTACKS
Government agencies may aim for surveillance or espionage. To this end, they can deanonymize and track users using our methodology. The rest of the attackers may aim for monetizing user information. By leveraging the profiles and interests of users, one can ``spear phish'' a focused group of targets.

In addition, any outsider can perform Tessellation without direct access to ISP wide traffic. One can run Tessellation on stored data collected from any network (e.g., public PCAPs on http://www.netresec.com/?page=PcapFiles).

2. While growing number of services use encryption, unless third party developers and ad partners of a service consistently secure their communications, the threat is real.

Even for a service operating over HTTPS, leakages from their third parties occur. Twitter's popular third party app, TwitPic, sends out user IDs unencrypted. A number of websites with Facebook 'like' buttons leak IDs as well. Both reconfirmed with data on March 2013.

As we show in Table 6, even without any OSN IDs, we can reconstruct significant information by using Traffic Markers (third-party tracking cookies). In order to prevent leakages of the cookies, it requires global adaptation of encryption for all ad traffic which may be challenging.

3. Regarding our accuracy evaluation, we only have Rada IDs for the 1-hour long dataset. Therefore we show accuracy numbers on that trace (Section 3.5). In the rest of the paper, we use the longer and richer, 3h-dataset, for which we can only show session coverage numbers.