# Mosaic: Quantifying Privacy Leakage in Mobile Networks

Ning Xia[§], Han Hee Song[†], Yong Liao[†], Marios Iliofotou[†],
Antonio Nucci[†], Zhi-Li Zhang[‡], Aleksandar Kuzmanovic[§]
[§] Northwestern University, [†] Narus Inc., [‡] University of Minnesota

**Abstract —** With the proliferation of online social networking (OSN) and mobile devices, preserving user privacy has become a great challenge. While prior studies have directly focused on OSN services, we call attention to the privacy leakage in *mobile network data*. This concern is motivated by two factors. First, the prevalence of OSN usage leaves identifiable digital footprints that can be traced back to users in the real-world. Second, the association between users and their mobile devices makes it easier to associate traffic to its owners. These pose a serious threat to user privacy as they enable an adversary to attribute significant portions of data traffic including the ones with *NO* identity leaks to network users' true identities. To demonstrate its feasibility, we develop the *Tessellation* methodology. By applying Tessellation on traffic from a cellular service provider (CSP), we show that up to 50% of the traffic can be attributed to the names of users. In addition to revealing the user identity, the reconstructed profile, dubbed as "mosaic," associates personal information such as political views, browsing habits, and favorite apps to the users. We conclude by discussing approaches for preventing and mitigating the alarming leakage of sensitive user information.

## Categories and Subject Descriptors

C.2.0 [**Computer-communication networks**]: General–*security and protection*

## Keywords

privacy;security;mobile network;user profile;online social network

## 1. INTRODUCTION

For a growing number of users, online social networking (OSN) sites such as Facebook and Twitter have become an integral part of their online activities. These OSN sites often function as launching points for users to receive news updates and venture over to other sites. In addition, many websites now have tie-ins with various OSN sites, so that users can recommend news items or Web posts via a simple click of Facebook's "Like" or Twitter's "Follow" buttons. With wide adoption of modern GPS-equipped mobile devices such as smart phones or tablets, as well as the emergence of various mobile applications and services, information access is nearly ubiquitous and literally at our fingertips.

With all the value and convenience it brings to our personal, social, and professional lives, this new era of mobile devices and online social networking also presents a quandary to users: *how to – or is it even possible to – preserve privacy in this new era*? Differing from their earlier incarnations, today's OSNs require users to register using their real names (at least in principle). In addition to personal data, such as age, gender, photos, and friends, these sites also track and record a variety of user online activities, such as messages exchanged and content shared with others, articles read and commented on, pictures browsed or video watched on the sites and other affiliated sites. At the same time, when accessing OSNs and mobile services on smart phones, users' current physical location may also be recorded and tracked due to the common use of automatic updates of location-specific contents.

Privacy issues directly related to OSNs are well-known and have been investigated by a number of recent studies [1,14,17,20,21,26] — this is *not* the primary focus of our paper. This paper calls attention to another important aspect of the privacy leakage problem: namely, *the potential danger to user privacy posed by a third party*, not simply by crawling data directly from OSN sites, but by gathering digital footprints left by users in cyberspace. As we explain next in more detail, the footprints can be collected by directly tapping into the wire, as well as by extracting information from the Web. Such a third party can be a hacker or a cyber criminal, a rogue employee in a cellular service provider (CSP) or an Internet service provider (ISP), a state agent of an authoritarian government, or any other "big brother" entities. Government agencies may target for surveillance or espionage. Other attackers may target to monetize user information by launching personalized spear-phishing attacks or spamming campaigns.

While extracting information from wireless or wireline packet traces is well explored, the prevalence of mobile devices and OSNs brings new possibilities that did not exist before: (i) Because such a device is typically tied to a specific user or a small, closely related user set, it is easier to associate traffic to specific users, e.g., via the spatio-temporal locality of user activities. Moreover, the prevalence of user OSN activities means that it is now plausible to further attribute traffic to an identifier used in the real-world, such as a user's first and last name extracted from her OSN profile, instead of simply using IP addresses or pseudonyms (e.g., email addresses) as before. (ii) With the real identifiers of users, one can paint more complete portraits of them by gleaning their network activities from the traffic, and then combining them with the data available on the Web, such as the users' OSN profile pages. (iii) Furthermore, the availability of GPS and other location information in mobile cellular data makes it possible to tie users' cyber

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*SIGCOMM'13,* August 12–16, 2013, Hong Kong, China.
Copyright 2013 ACM 978-1-4503-2056-6/13/08 ...$15.00.

activities to their presence in the physical world. Given the factors above, we see that the confluence of smart phones and OSNs renders the ability to glean personal information from mobile data a far more potent threat to user privacy than attacks on each individual service. An important question motivates our work: Is this concern merely hypothetical or real?

It is with this question in mind that we set out to study the *privacy leakage* problem from a network. Our goal is to first quantify the amount of privacy leakage in the online digital footprints left by users, and then characterize to what extent a third party may gather useful personal information from these disparate digital pieces. We refer to this problem as constructing a **mosaic** of a user from her online digital footprints, and correspondingly refer to the gathered footprint pieces as **tesserae**.

To demonstrate that this is indeed feasible, we develop a novel methodology referred to as **Tessellation**. Through Tessellation, we show how user identity information such as OSN IDs and device tracking cookies can be extracted from the traffic. Furthermore, we describe how the remaining pieces of traffic with no identity leakages can be attributed to the known user identities. Finally, with the additional information gleaned from the Web (e.g., data disclosed by the user in his/her OSN sites or by other sources), we further corroborate and augment the "mosaic" of online user portraits.

While Tessellation can be ran on any network, given the growth and ubiquity of mobile devices, in this work, we choose to focus on the privacy issues in mobile networks. From our evaluation on data trace collected from a CSP, Tessellation can attribute 50% of traffic to the owners with only 5% error. Optionally, the coverage can be increased to 80%, with just a 2% increase in the error rate. Using our methodology, we were able to create mosaics for more than 16,000 users and classify their personal information into 59 categories including user demographics, locations, affiliations, social activities, interests, etc. Our work makes the following specific contributions:

- We show that detailed personal information, even from traffic with *NO* obvious identity leaks, can be extracted and intelligently gleaned by any third party with access to the wire. More importantly, the third party can do this without direct mappings between network flows and their owners, and without information from full packet payloads.
- We design the Tessellation methodology that automatically brings together isolated pieces of mobile users' personal information. Using Tessellation, one can collect this information, even when the user activities are spread over different dynamic IPs, different devices, and over different time periods.
- We show that the combination of information reveals far greater knowledge of users than what can be obtained individually. Further, we quantify the amount of leaked information as a function of the duration of the vulnerability and number of compromised IPs, compare the information disclosure from network traffic versus public OSN profiles, and present case studies on how one can learn aspects of a specific user or user group.

**Vision of mosaic.** Our objective is to call attention to the potential risk to user privacy due to the *personal information leakage in network data*. Our work highlights the potential danger to user privacy posed by the prevalent usage of OSNs and mobile smart devices, both of which make it easier for a powerful and sophisticated adversary to attribute data traffic to specific users in the real-world and glean personal information about them. As illustrated by our *Tessellation* methodology, such capability is facilitated, in part, by some shortcomings of certain OSN design, as well as by the fundamental limitations of the current Web and Internet *from a user privacy perspective*, such as cookie mechanism used by the stateless HTTP protocol. Based on our analysis, in §5.4, we suggest possible countermeasures to safeguard against the alarming leakage of private information.

# 2. PRIVACY LEAKAGE IN MOBILE DATA

## 2.1 Motivation

The popularity of OSNs has increased the amount of sensitive information leaked into the network. On any day, a user (*i.e.*, Alice) may get onto the Internet using different devices over time: smart phones, tablets, or laptops. Even on a single device, her IP address will be allocated dynamically and randomly, depending on her mobility, traffic pattern, and the policies of her ISP/CSP. However, even though the IP address may change, Alice will be accessing the same sites. As a result, every time Alice logs on to the OSN sites and performs a variety of online activities, she leaves "islands of digital footprints" in the networks. Such digital footprints, once collected, can potentially be pieced together by someone to paint a digital "mosaic" of Alice and learn a lot about her.

**Attack model.** Clearly, the aforementioned danger to user privacy depends on one's ability to collect network data containing user "digital footprints". A powerful and sophisticated adversary may be capable of tapping into the wire, listening "in the air" [23, 24], or gaining access to stored network traffic (e.g., archived pcap files). The adversary can be a rogue employee in a CSP, a state agent, or a hacker. Even though an employee inside the origin CSP of a user can directly acquire user information through billing databases, such records are not available for employees present in transit Autonomous Systems (ASes). Given that network traffic, especially towards popular OSNs and e-commerce sites, is unlikely to remain local, any transit AS with access to traffic has the potential to launch such attacks. Government agencies have a similar observation point when they cannot directly collaborate with an origin ISP (e.g., CIA vs. a Middle-Eastern ISP). Agencies can acquire raw traffic data from a transit AS or even re-route traffic by launching a BGP/IP hijacking attack. Although such activities have not been publicized, they are theoretically possible. Finally, a hacker can perform Tessellation by gaining access to stored data collected from any network, such as [22], without direct access to ISP-wide traffic.

**Goals of the attack.** Government agencies may aim for surveillance or espionage. To this end, they can de-anonymize and track users using our proposed methodology. Other attackers may aim to monetize user information. By leveraging the profiles and interests of users, one can spear-phish a focused group of targets. For example, an attacker may gather information about a user, so as to craft a personalized email to trick the user into clicking a malicious URL. As we show next in this paper, one can collect such personal information about a user even with partial information of layer-7 headers (as our data do not have full packet payloads). Details of our datasets are provided in §2.3.

Assuming that an adversary has the ability to collect the CSP data, we address the following two intertwined questions in this paper: First, is it feasible to utilize users' OSN activities (and *user identifiable information* that may be leaked through such activities) to extract and attribute users' digital footprints to individual OSN users? Second, if the answer to the first question is affirmative, then how much and what type of information can be gleaned from the data, assisted and corroborated by whatever public information about the users available on the Web?

The first question essentially asks if we can associate network traffic to individual OSN users. A naive approach to address this question is to rely on the fact that many OSN sites incorporate the

| | Naive approach | Our approach | |
|---|---|---|---|
| | OSN ID Extraction | Traffic attribution | Activity analysis |
| Traffic coverage(%) | 2.4% | 49.8% | 78.6% |
| Error(%) | 0.0% | 5.5% | 7.5% |

**Table 1: Increase of traffic coverage in our approach**

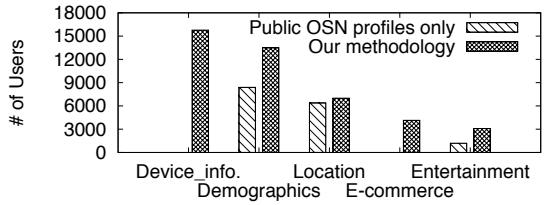



**Figure 1: Personal information gain in our approach**

user identifiers in HTTP headers (e.g., cookies) as plain text. Therefore, one can extract such OSN identifiers and directly assign those flows to an OSN user. We quantify this in Table 1, where we show that the amount of traffic attributed to OSN users by this naive approach compared to our proposed Tessellation approach. As we see, only 2.4% of the traffic is covered by simply identifying traffic with OSN IDs. Intuitively, we observe low coverage because OSN IDs are only leaked during some authentication phase, which comprises a relatively small fraction of the traffic.

In our Tessellation methodology, we devise a simple yet novel analytic to automatically associate traffic with no identity leakage to specific OSN users. As shown in the two columns titled "our approach" in Table 1, Tessellation can attribute almost half of the traffic to its owners with a small error of 5%. We can even raise the coverage rate to 78.6%, at the cost of an additional 2% error.

Regarding our second question, Figure 1 shows the information gained by crawling the public OSN profile of users (using the extracted OSN IDs) compared to the outcome of running our complete Tessellation methodology. As we see, various user attributes can be obtained from OSN profile pages, including demographics (*e.g.*, name, gender, birthday), location (*e.g.*, city, state), interest in entertainment (*e.g.*, favorite music, TV), etc. Unfortunately, this source of information carries inherent limitations of the OSN profiles – they are static (*e.g.*, interests a user declared at the time of joining the OSN) and coarse-grained (*e.g.*, location information only up to city level). However, with Tessellation, the information gained from combining raw traffic analysis corroborates and compliments the OSN profile information and reveals a wider variety of user activities (*i.e.*, device information, social associations, e-commerce activities, etc.). Moreover, it brings finer-grained and dynamic information, such as GPS coordinates of users' with timestamps, news, or shopping sites frequently visited, as well as videos users just watched. The full breadth of information extracted by Tessellation is the topic of a later section (§5.2).

## 2.2 Overview of Tessellation

**Traffic attribution.** Figure 2 overviews the workflow of "Traffic Attribution" of Tessellation (§3). Given a set of Layer-7 flows on a client IP address (marked as vertical lines in the figure), the first step of Tessellation, *OSN ID extraction*, extracts traffic that leaks OSN user IDs (marked with triangle flags). While dynamic IP assignment scatters a user's mobile traffic to multiple IP addresses in the long-run, the same IP address stays with the same mobile device for a short period until the device becomes idle for at least a few seconds. Leveraging this feature of mobile data networks, the second step, *Session block generation*, segments traffic on each IP address into blocks of generally short durations (shown as blocks in the second row of Figure 2). The challenge now becomes how to associate and attribute appropriate traffic blocks to individual

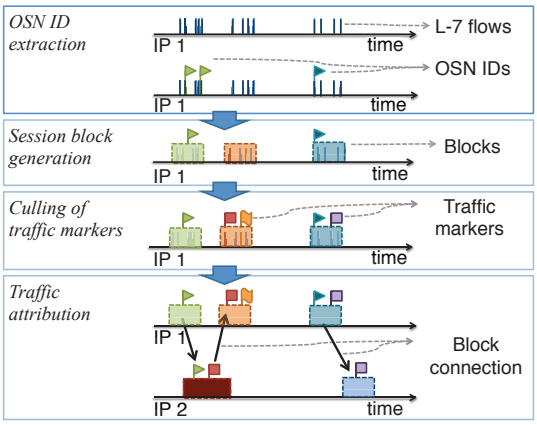



**Figure 2: Example of traffic attribution in Tessellation**

OSN users. In step three, *Culling of traffic marker*, we take advantage of ubiquitous cookies and related HTTP header fields (collectively referred to as *traffic markers*) that are used by Web services to keep track of users and devices (marked as square flags in the third row). Finally, in the fourth step, *Traffic association*, we associate the blocks that do not have OSN IDs (but have traffic markers) to the OSN users by connecting them through a block that has both OSN IDs and traffic markers (darkened block shown in the bottom of the last row). In the example, flows in the three blocks are now attributed to a single user (OSN ID).

**Mosaic construction.** Once we attribute traffic to individual OSN users, in the second stage, "Mosaic Construction", of Tessellation (§4), we attempt to mine and collect various kinds of user information that might be of interest to an adversary. For this, we conduct user activity analysis based on the DNS names associated with various services/sites they visit, classify them, and analyze users' distinct **Activity Fingerprints**. Furthermore, aided by the service classification, site-/service-specific information mining can be performed to gather specific types of information or interests (*e.g.*, GPS locations, device information, etc.). Such information gathering is further augmented by crawling the Web (*e.g.*, public profiles of OSN users). When combining information from all of these sources, we show that it is indeed possible to construct a "well-connected" content-rich user mosaic.

Figure 3 details the workflow of the proposed Tessellation. Next, in §3 and §4 we cover the Traffic Attribution and Mosaic Construction parts, respectively.

## 2.3 Dataset Description

Throughout this paper, we use network packet traces collected within the cellular data networks of two major CSPs (CSP-A, CSP-B). The main dataset used in this paper was collected from a backbone router of CSP-A for three hours, from 15:30 to 18:30 UTC in spring 2011 (referred as *3h-Dataset*). The dataset contains all traffic from a subset of areas in the North America the CSP serves. The data contain both layer-3/4 packet headers, as well as layer-7 headers, and span over 65,000 client IPs. We also have a second dataset collected from another backbone router of CSP-A in winter 2011, which lasts for 9 hours (the *9h-Dataset*) with 340,000 client IPs but contains primarily layer-7 HTTP headers.

The third hour-long dataset (Ground Truth Dataset) is collected from the content billing system of CSP-B in the summer 2012. Differing from the first two, this dataset provides details of beginning and ending time of users' Remote Authentication Dial In User Service (RADIUS) protocol [27] which associates each user with an ID for pay-per-use billing purpose. Because of its short duration, we only use the Ground Truth Dataset for accuracy evaluation.

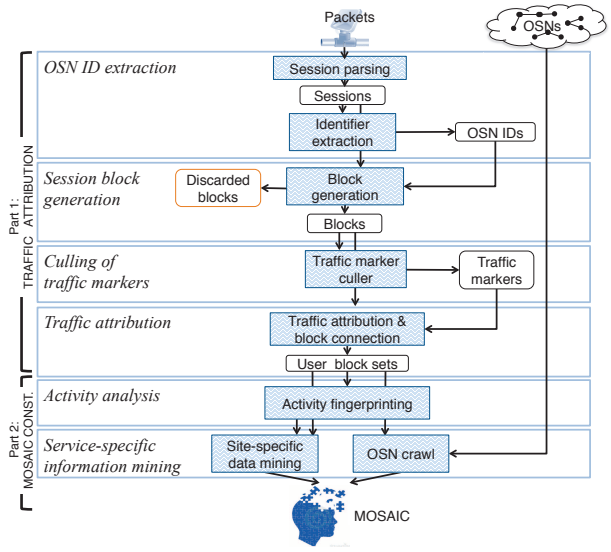



**Figure 3: User mosaic construction via Tessellation**

**Data preprocessing.** We analyze raw network traffic traces by first grouping packets into 5-tuple TCP/UDP flows using the standard TCP-level flow re-assembly. According to the specific protocol, the 5-tuple flows are then further parsed into (application) *sessions*, which are defined as a layer-7 interaction with one or more layer-4 connections. For example, all HTTP requests and replies of the same persistent TCP connections are grouped into a single HTTP session. Similarly, TCP flows belonging to an SMTP transaction become a single SMTP session. Likewise, UDP flows to the same DNS access turn into a DNS session. All traces were analyzed using the Narus Semantic Traffic Analyzer (STA) tool.

**Challenges with mobile traces.** Given the growth and ubiquity of mobile devices, in this paper we focus on mobility traffic of CSPs rather than static ISP traffic.Mobile traffic contains considerable amount of sensitive information that is important to our analysis, but at the same time, its mobile nature brings significant challenges. For instance, the dynamic nature of mobile IP assignment makes it hard to rely on IP addresses for identifying traffic owners. On mobile CSP networks, a client IP address of a device can change as frequently as once every 30 minutes, rendering current approaches of IP-based traffic associations ineffective [3]. Also, CSP traffic has a larger portion of encrypted flows (44% of them being HTTPS [7] in general) than that of some ISPs (3% HTTPS [10]). This fact highlights a significant advantage of our methodology: even with 38% of our CSP traffic encrypted, we can still attribute ∼80% of the unencrypted traffic to the users who generated it.

# 3. TESSELLATION PART I: TRAFFIC ATTRIBUTION VIA LEAKED OSN IDS

Tessellation is the process of annotating sessions in the captured network data to particular OSN user identifiers. In this section, we discuss the steps of Tessellation, which are depicted in Part 1 in Figure 3. The coverage analysis and accuracy evaluations of Tessellation are covered next in this section.

## 3.1 OSN ID Extraction

As mentioned in §2, the fact that many OSN sites "leak" their *user identifiers* allows Tessellation to attribute traffic to real users. In our paper, we focus on two popular OSN providers, labeled as OSN1 and OSN2, which account for over 95% of all OSN accesses in our datasets. Because each OSN has its own specific design, discovering which bit-/character-strings in HTTP headers are used by an OSN for uniquely identifying each user (namely, user identifiers) so as to extract them automatically, is not entirely trivial. It requires OSN-specific parsing and analysis.

| OSN IDs | Where to find | Keywords | Session coverage |
|---|---|---|---|
| *OSN1* ID | HTTP URL: *.osn1domain.com | session_key=#####-<**OSN1_ID**> | 166441/1.3% |
| | HTTP cookies | c_user=<**OSN1_ID**>; | |
| | HTTP cookies | m_user=email%3a<**OSN1_ID**> | |
| *OSN2* ID | HTTP URL: *.osn2domain.com | oauth_token=<**OSN2_ID**>-##### | 119849/1.0% |
| Email address | HTTP cookies | m_user=**email**%3aOSN1_ID | 24147/0.2% |
| | IMAP:payload | USER=**email@domain.com** | |
| | POP3:payload | LOGIN=**email@domain.com** | |
| | MSN:payload | MSNMSGR=**email@domain.com** | |

**Table 2: OSN User Identifiers**

Taking *OSN1* as an example, a numeric identifier can be found in either the URI or the authentication cookie, which is used as part of a session key. In addition to the user ID, the mobile pages of *OSN1* (*m.OSN1.com*) use a cookie that also leaks the email addresses of users (*m_user*). For the purpose of Tessellation, both the *OSN1* ID and email address are considered as the OSN user identifiers, as they both uniquely identify an *OSN1* user. In the case of *OSN2*, the user identifiers can be found in the authentication token, *oAuth*, included in the HTTP header field during user authentication. The numeric user ID is uniquely mapped to a user-generated *OSN2* screen name (*user*). See the snippet below as an example where the boldfaced line contains a user's *OSN2* ID:

```
1 authorization: OAuth realm="http://api.osn2.com/1/direct_messages.json",
2 oauth_nonce="1964799###", oauth_signature_method="HMAC-SHA1",
3 oauth_consumer_key="w1Gybt9LP9zG46mS1***",
4 oauth_token="<OSN2_ID>-OQyCfMaEcpYKQV7x***",
5 oauth_timestamp="########", oauth_signature="1Qt***"
```

In Table 2, we summarize the formats of common *OSN1* and *OSN2* user identifiers and report their locations inside the HTTP headers. Using the 3h-Dataset, which contains 12, 495, 482 (HTTP) sessions, we find 12,420 unique *OSN1* identifiers (users), which show up in a total of 166,441 sessions (about 1.3% of all sessions). Similarly, there are a total of 1,952 unique *OSN2* identifiers (users), that appear in 119,849 sessions (about 1.0% of all sessions). User identifiers of other OSN sites can be extracted in a similar fashion. However, since the identifiers from the two main sites comprise 95% of all observed IDs, we focus on *OSN1* and *OSN2* for our analysis. Apart from the user identifiers used by OSN sites, other user identifiers such as email addresses are often leaked by various services and protocols (*e.g.*, unencrypted Webmail, POP, or IMAP) and can therefore be used as user identifiers for the purpose of traffic attribution. We list some examples in the bottom part of Table 2.

## 3.2 Session Block Generation

As we show in Table 2, OSN sites tend to leak the identifiers of their users. Even though, the sessions containing such identifiers cover only a small fraction (2.5% in the 3h-Dataset) of all sessions. In this section we illustrate how we can use these few sessions as "anchors" to further expand our coverage of traffic attribution. To achieve this, we use the observation that CSPs commonly assign a single IP address to a device as long as the device is actively sending traffic. Therefore, traffic activities occurring on the same IP address within a short period of time are likely to belong to the same mobile device. Next, we describe how we utilize this observation to segment traffic into session blocks that likely belong to a single user.

At first, we begin with sessions from the same source IP address, and then group the sessions into distinct blocks of contiguous sessions using the following simple heuristic: two consecutive sessions belong to the same block if and only if the "idle" period (i.e. the ending time of the previous session and the starting time of the

next session) between them is less than $\delta$ seconds, where $\delta$ is a parameter depending on the dynamic IP assignment scheme used by the CSP. In other words, any two blocks (on the same source IP address) are separated from each other by an idle period longer than $\delta$ seconds. In our study, we use $\delta$=60 seconds, based on the analysis of the idle period distribution and confirmed by conversations with a network operator of the service provider. After applying this heuristic to the 3h-Dataset, the initial $12,495,482$ sessions are segmented into a total of $99,234$ session blocks.

Some factors complicate the problem and may cause the above heuristic to generate blocks that do not belong to a single user. One factor is the presence of network address translation (NAT) devices in the data. For example, phone tethering allows additional devices (e.g., a laptop or a tablet equipped only with WiFi) to access the Internet via a tethered mobile device. Another factor is that more than one user may share a phone within a short period of time.

To address these challenges, we apply the following steps. First, we filter out the blocks behind the NATed devices by testing for the existence of heterogenous IP TTL (time-to-live) values in a block as described in [4]. We also filter out the blocks shared by multiple users by determining the existence of two or more distinct user identifiers of the same OSN (*e.g.*, *OSN1*) from a single block. Applying these methods to the 3h-Dataset, we find $563$ blocks with conflicting TTL or OSN IDs. They account for a total of $993,171$ sessions, with an average of $1,764$ sessions per block.

## 3.3 Culling of Traffic Markers

While only 22,366 out of 97,117 blocks in the 3h-Dataset contain OSN user identifiers, intuitively we would expect that some of the blocks that are close in time, even though they do not contain any OSN user identifiers, are likely to belong to the same OSN user. In order to identify and attribute other session blocks (with no OSN identifiers) that are likely to be generated by the same OSN users, we leverage HTTP cookies and other <key-value> strings in HTTP headers, henceforth referred to as **Traffic Markers**. The traffic markers are generated and used by various Web services to bring together stateless HTTP request/reply messages and to keep track of the webpages users visited, user devices, or the users themselves. The challenges are that there are a huge variety of site-specific traffic markers, many of which are dynamically generated. For instance, a cookie's value may change as a user is tracked across pages within a website. Ideally, we would like to cull only those that are longer-lasting, *e.g.*, those that are used in tracking users or their devices.

One naive way to cull traffic markers is to perform site-specific analyses, which requires unscalable and error-prone manual inspection. To overcome this problem, we automate the process by relying on the co-occurrences of OSN and other Web traffic within those session blocks containing OSN user identifiers. First, we introduce two important concepts: *persistency* and *uniqueness*.

Let $U = \{u_i\}$ be a set of (OSN) users discovered in the data, where each user $u_i$ is defined by a set of OSN identifiers (e.g., *OSN1* ID and *OSN1* email address) she possesses. For simplicity, we treat an OSN user and her identifiers equivalently. Let $M = \{m_l\}$ be a set of *potential* candidate traffic markers, where each marker $m_l$ is typically expressed in the form of <key-value> pairs, i.e., $m_l \triangleq (k_l, v_l)$. We say that two (potential) markers $m_h \triangleq (k_h, v_h)$ and $m_l \triangleq (k_l, v_l)$ are of the same *type* if $k_h = k_l$ but $v_h \neq v_l$. Given a pair of $(u_i, m_l)$, $\mathbf{P}(u_i, m_l)$ denotes the probability that user $u_i$ and marker $m_l$ co-occur within a session block. $\mathbf{P}(u_i, m_l)$ is empirically computed as the total duration of the blocks that contain both $u_i$ and $m_l$ divided by the total duration of all blocks containing any $u \in U$. Let $\mathbf{P}(u_i) := \sum_{m_l \in M} \mathbf{P}(u_i, m_l)$.

| Traffic marker domain | Cate -gory | Where to find | Keywords |
|---|---|---|---|
| admob.com | Ad | HTTP: X-Admob-ISU | X-Admob-ISU |
| atdmt.com, msn.com, bing.com | Ad | HTTP:cookie | muid |
| doubleclick.net | Ad | HTTP:cookie | id |
| mydas.mobi | Ad | HTTP:cookie | mac-id |
| google.com | Sid | HTTP:cookie | sid |
| craigslist.org | Uid | HTTP:cookie | cl_b |
| yahoo.com | Uid | HTTP:cookie | c |
| scorecardresearch.com | Tid | HTTP:cookie | uid |
| quantserve.com | Tid | HTTP:cookie | mc |
| google-analytics.com | Tid | HTTP:cookie | utmcc |

**Table 3: The top-10 most commonly found traffic markers**

**Uniqueness.** Given a pair $(u_i, m_l)$, where $\mathbf{P}(u_i, m_l) > 0$, the *uniqueness* of $(u_i, m_l)$ (or simply, $m_l$), denoted by $\Psi(u_i, m_l)$, is defined as $\Psi(u_i, m_l) := 1 - \sum_{j \neq i : u_j \in U} \mathbf{P}(u_j, m_l)$.

From the above definition, if $\Psi(u_i, m_l) = 1$, *candidate* traffic marker $m_l$ is uniquely associated with user $u_i$. Otherwise, the same marker has been associated with another user, signifying that it is not a good traffic marker. Among all *candidate* traffic markers uniquely associated with each user, many of them may be "ephemeral" (*e.g.*, change from one webpage to another or from one user session to another). This leads us to define:

**Persistency.** Given a pair $(u_i, m_l)$ (where $\mathbf{P}(u_i, m_l) > 0$ and $\Psi(u_i, m_l) = 1$), the *persistency* of $(u_i, m_l)$ (or simply, $m_l$), denoted by $\Pi(u_i, m_l)$, is defined as:
$\Pi(u_i, m_l) := 1 - \sum_{h \neq l : m_h \in M} \mathbf{P}(u_i, m_h) / \mathbf{P}(u_i)$.

From the above definition, if $\Pi(u_i, m_l) = 1$ or $\Pi(u_i, m_l) \approx 1$ (say, $\geq 0.9$), the candidate marker co-occurs with $u_i$ almost all the time throughout the observation period. Hence $m_l$ serves as a good candidate traffic marker, and thus can be used to attribute other session blocks that contain $m_l$ but not $u_i$ to user $u_i$. In addition, using this persistency property, we can automatically filter out most session- or page-specific cookies whose values change from one webpage to another or from one user session to another as their persistency values are generally very low.

Applying our automated traffic marker culler to the 3h-Dataset, we cull 625 *types* of traffic markers. Table 3 lists 10 types that are most commonly seen in the data. Most of these markers are located inside the cookie field of the HTTP header listed in the "Keywords" column of the table. An exception, *admob.com* identifier, is found in a specific string ('X-Admob-ISU') in the HTTP GET requests. We use the total of 625 such traffic markers grouped into four categories based on the purpose they serve, *i.e.*, advertisement (Ad), personalized logins (Uid), tracking users (Tid), and tracking service sessions (Sid). We see that most of them are used for tracking the activity of users or for personalized advertising. The uniqueness and persistency values of the top-20 traffic marker types are shown in Figure 4. As expected, in the figure we see that the three OSN identifiers satisfy $\Pi = 1$, $\Psi = 1$ by definition.

## 3.4 Traffic Attribution

After culling the traffic markers, we next piece together the session blocks that are likely to only contain traffic generated by individual OSN users. Combined with the blocks annotated by the OSN IDs, they form the building blocks based on which the user mosaic will be pieced together (as we show later in §4.2).

Having the set of traffic markers $M(u_i)$ of user $u_i$, traffic attribution is straightforward: a block is attributed to user $u_i$, if and only if it contains either an OSN identifier of user $u_i$ or a traffic marker $m_l \in M(u_i)$.

| Tessellation steps | Coverage | | Coverage per-user | | Accuracy on covered set | | $\Psi$ |
|---|---|---|---|---|---|---|---|
| | Session | User | Session (avg / 90%) | Time (avg / 90%) | Session | User | |
| OSN ID extraction | 2.4% (297,358) | 15.7% | 14 / 7 | 11.8 / 8.3 [min] | 100% | 100% | 1 |
| Traffic attribution | 49.8% (6,217,036) | 43.2% | 326 / 176 | 65.4 / 62.3 [min] | 94.5% | 99.3% | 1 |
| Activity analysis | 78.6% (9,831,924) | 69.0% | 586 / 530 | 82.3 / 81.0 [min] | 92.5% | 96.4% | 0.98 |

**Table 4: Coverage and accuracy at each stage of the Tessellation**

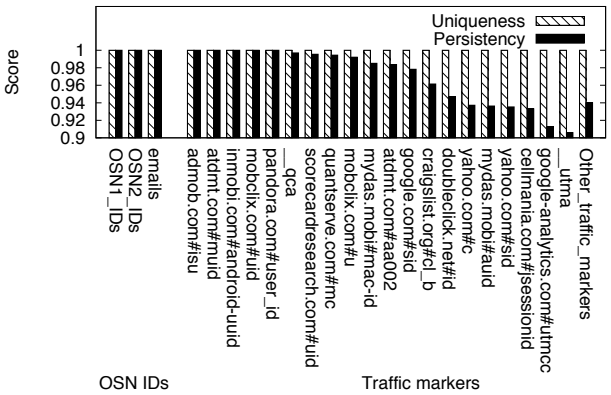



**Figure 4: Uniqueness $\Psi$ and persistency $\Pi$ for OSN identifiers (grouped on the left) and top-20 markers (right)**

## 3.5 Evaluation of Traffic Attribution

We evaluate the coverage and accuracy of traffic attribution using the Ground Truth Dataset. As explained in §2.3, in this dataset we have information to associate each session to a RADIUS user [28]. Figure 5 illustrates different example scenarios that may occur during our inference. The examples show the sessions of an inferred user (dubbed as Tessellation user) and how they compare with the ground truth (RADIUS user). In our examples, $R_i = (R_1, ..., R_6)$ and $T_j = (T_1, ..., T_5)$ denote the set of ground truth users and inferred users, respectively. Next, we explain the examples of Figure 5 in more detail.

(a) Tessellation does not identify any sessions from $R_1$. All these sessions are considered as false negatives.

(b) Tessellation associates some sessions of $R_2$ to $T_1$ but some are missed. This occurs when the sessions marked with "?" lack enough evidence to be attributed to $T_1$.

(c) Tessellation associates all of the sessions belonging to $R_3$ to $T_2$ (*i.e.*, $1:1$ match between Tessellation user and RADIUS user). This is the ideal case.

(d) Tessellation maps the traffic from two or more users ($R_4$ and $R_5$) to a single inferred user $T_3$ (*i.e.*, $1:$ *many* match between Tessellation user and RADIUS users). This is a misclassification because Tessellation wrongfully associates sessions from multiple users to a single real user.

(e) Tessellation infers some of $R_6$'s sessions to belong to $T_4$ and some to $T_5$ (*i.e.*, *many* $:1$ match between Tessellation users and RADIUS user).

For brevity, we do not list the cases of (d) and (e) with partial matches as they can easily be composed using the other examples. With the above cases in mind, we evaluate the coverage and accuracy using the following two metrics.

**Coverage (a.k.a. completeness).** Session-level coverage is the number of sessions that are given a prediction (*i.e.*, sum of sessions in all $T$s), divided by the total number of sessions. User-level coverage is the number of ground truth users for whom Tessellation identified all or a subset of their sessions divided by the total number of ground truth users. In the example in Figure 5, the only user that has no coverage is $R_1$.

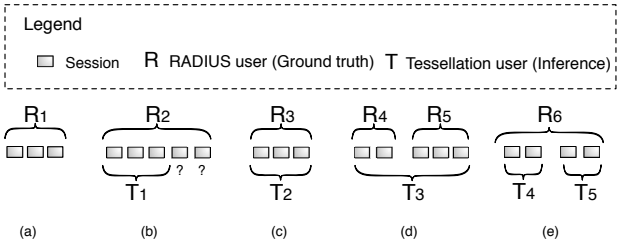



**Figure 5: Five cases of scenarios occurring in the accuracy and coverage measurement of Tessellation**

The results of our coverage evaluation are summarized in Table 4. The "coverage" column of the first two rows of Table 4 summarize the session and user coverage of Tessellation. A discussion on the third row (Activity analysis) will follow in §4.1. At the beginning of Tessellation, we only identify the sessions with OSN IDs and the coverage is low at 2.4% out of 12,495,482 sessions. User coverage is higher at 15.7% out of 22,862 users. This shows that a large number of users falls into case (b) rather than (c) (see Figure 5). By using the Traffic Attribution step we get a twenty-fold improvement in session coverage. We see this in the second row of Table 4, where the session coverage increases to 49.8%, and user coverage increases to 43.2%. This is a significant improvement because now we can associate half of the traffic (49.8%) with the users who generated it.

To measure how much of users' activities we capture, we include per-user coverage statistics in Table 4. Specifically, we report the average over all users and the bottom 90th percentile user ordered by their number of sessions. At first, we see that just 14 sessions are associated with each user on average. These sessions last for a total duration of 11.8 minutes. With the traffic attribution, an average of 326 sessions is attributed to a user, lasting a total of 65.4 minutes. For the bottom 90% of users, we also see a similar improvement; average session coverage increases from 7 to 176 and time coverage increases from 8.3 to 62.3 minutes. The results indicate that we can now view hour-long activities of users without interruptions.

**Accuracy on Covered set (AoC).** Session-level AoC is the number of correctly identified sessions (*i.e.*, sum of sessions in $T_1$, $T_2$, $T_4$, $T_5$), divided by the total number of predicted sessions (*i.e.*, sum of sessions in all $T_i$s where $i = 1 \cdots 5$). User-level AoC is the number of correctly identified users (*i.e.*, $R_2$, $R_3$, and $R_6$), divided by the total number of predicted users (*i.e.*, $T_1$, ..., $T_5$).

The "accuracy" columns of Table 4 show the AoC of each stage of Tessellation. At the OSN ID extraction stage, both the session AoC and user AoC are 100% because the sessions being extracted are only the ones with user identifiers. At the traffic attribution stage, the session AoC and the user AoC slightly decrease to 94.5% and 99.3%, respectively. The drop in the accuracy is due to the instability of traffic markers with persistency $\Psi < 1$. Breaking down the accurately inferred sessions into cases in Figure 5, 88.7% of them fall into (b) or (c). The remaining 10.6% of them fall into (e) indicating that not all of the users' blocks could be entirely culled into a single identity. Even if the mosaic of some users is incomplete, as we show next, the collected information allows the creation of detailed profiles with a number of practical applications.

| Service class | Keywords | Service provider |
|---|---|---|
| Banking | bank | wellsfargo, morganstanley |
| Blog | blog, buzz | huffingtonpost, boingboing |
| Book | book | barnesandnoble, half.com |
| Chat | talk, chat, messenger | skype, mtalk.google, aim |
| Dating | personals, harmony, match | plentyoffish, date |
| E-commerce | warehouse, market, buy | amazon, ebay, blockbuster |
| Education | .edu, college, education | medexch.med.unc.edu |
| Email | smtp, imap, pop, exchange | google, hotmail, yahoo |
| File hosting | upload, download, ftp | megaupload, dropbox |
| Gaming | game, casino | zynga, farmville, xbox |
| Map | maps, virtualearth | maps.google, wikimapia |
| Music | music, radio, playlist | pandora, itunes, zune |
| News | news | msnbc, ew, cnn |
| P2P | tracker, torrent, mininova | No specific domain |
| Picture | picture, photo | flickr, picasa.google |
| Search | search | google, bing, yahoo |
| Social | social | *OSN1*, *OSN2*, ning |
| Sports | sports | espn, bleacherreport |
| Travel | travel, hotel, flight | expedia, kayak, southwest |
| Video | video | netflix, youtube |
| Weather | weather, forecast | No specific domain |

**Table 5: Samples of service classes and providers**

# 4. TESSELLATION PART II: CONSTRUCTION OF USER MOSAIC

In this part of Tessellation, we create profiles of users by extracting key information from their sessions. The steps in this section correspond to the last two blocks of Figure 3.

## 4.1 User Activity Analysis & Fingerprinting

We start by analyzing and classifying online activities that users are engaged in, *e.g.*, websites they frequently visit. For this, we utilize the DNS names associated with various services, which often provide a good indication of the category of activities that a user is engaged in. For instance, `mail.yahoo.com` indicates that a user is checking her email; `www.youtube.com` indicates that she is likely to be browsing and watching videos online. By correlating the DNS query traffic, we are able to map the destination IP addresses in the dataset to their corresponding DNS names. After obtaining the DNS names, we further associate each DNS name to a *service class* and a *service provider*. We adopt a similar keyword matching scheme used in [12] to classify DNS names into 21 different service classes.

In the 3h-Dataset, we extract $54,426$ distinct domains and classify them into 21 distinct classes of services. In Table 5, we report an illustrative summary of the service classes, keyword samples used to classify DNS names into each service class, and service provider samples of each class. Classifying the DNS names into [service class, service provider] pairs reveals not only the types of activities a user is engaged in, but also the preferred service providers. Furthermore, it also enables us to study whether some users have distinct fingerprints in their activities that can be used for further traffic attribution, as we discuss below.

We are interested in obtaining a subset of services (*e.g.*, the most frequently accessed services) used by a user that can "fingerprint" the user with high confidence. In other words, these services represent a *distinct* activity pattern that distinguishes the user from all other users. For each user $u_i \in U$, let $s_i^j$ be the combination of a service class and a service provider identified from $u_i$'s traffic by our DNS name classification (each $s_i^j$ is a [service class, service provider] pair). Let $S(u_i)$ be the list of distinct $s_i^j$s associated with the traffic generated by the user, such that $S(u_i) := \{s_i^j\}$. Because $S(u_i)$ contains all the $s_i^j$ including the ones user $u_i$ visited only once during our observation, considering the entire $S(u_i)$ may introduce inconsistency in determining user $u_i$'s activity pattern.

To consider the most representative activities of $u_i$, we use the top $k$ most frequently accessed services, $F_i \subseteq S(u_i)$, where its length $|F_i| = k$. We refer to $F_i$ as the *activity fingerprint* of user $u_i$. Figure 6 shows the cumulative distribution of the user activity fingerprint lengths for all users in the 3h-Dataset. As expected, the number of users with an activity fingerprint length greater or equal to $k$ diminishes as $k$ increases. For example, about $60\%$ of the users have an activity fingerprint of length longer than 3; whereas only about $20\%$ of the users have an activity fingerprint of length longer than 8. Clearly, the longer an activity fingerprint, the more likely it will be uniquely associated with a user.

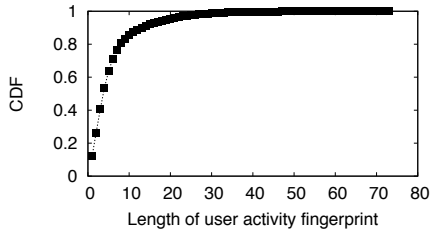



**Figure 6: CDF of the length of user activity fingerprints**

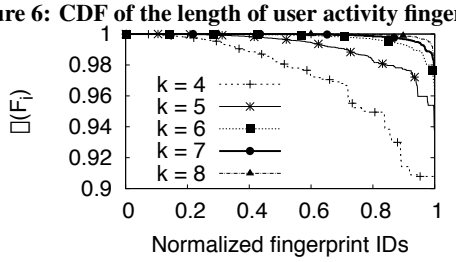



**Figure 7: CDF of the uniqueness score on user activity fingerprints of different size**

**Uniqueness.** Let $\mathcal{S}^k := \{u_j \in U || S(u_j)| \geq k\}$ and $\mathcal{F}_i^k := \{u_j (\neq u_i) \in U | S(u_j) \supseteq F_i\}$. In other words, $\mathcal{S}^k$ is the set of users whose activity fingerprint length is at least $k$, and $\mathcal{F}_i^k$ is the set of users (other than $u_i$) whose activity fingerprint contains $F_i$ as a subset. Analogous to $\Psi$ in §3.3, we define the *uniqueness* of $F_i$ as follows: $\Psi(F_i) := 1 - |\mathcal{F}_i^k|/|\mathcal{S}^k|$. The closer $\Psi(F_i)$ is to 1, the more distinct $F_i$ is as an *activity fingerprint* of user $u_i$.

Using the 3h-Dataset, Figure 7 plots $\Psi(F_i)$ with $F_i := S(u_i)$ for users whose activity fingerprint length $k$ ranges from 4 to 8. For each $k$, the x-axis represents the $F_i$s, ordered in decreasing value of $\Psi(F_i)$ (the y-axis). To allow comparison across different $k$s, the scale of x-axis is normalized by the total number of $F_i$s so that its range is [0,1]. We see that as $k$ increases, overall $\Psi(F_i)$ gets to closer to 1, which is expected. If we read Figure 7 together with Figure 6, we can see a clear tradeoff: with a larger $k$, the distinctness of an activity fingerprint $\Psi(F_i)$ increases. On the other hand, the utility of a longer activity fingerprint decreases with $k$, as fewer users have an activity fingerprint with length $\geq k$. (Furthermore, it requires a longer observation period to "fingerprint" a user's activity pattern and/or to attribute an unknown traffic block.) From the figures, we see that a good choice for $k$ is 5, which guarantees a reasonable $40\%$ user coverage while ensuring that $85\%$ of such fingerprints have $\Psi \geq 0.98$. Hence, using such an activity fingerprint, the probability that we erroneously attribute the activities of one user to another is at most $2\%$. Additional criteria may be used to reduce such fingerprinting or attribution errors, for example, by imposing certain closeness in time constraints.

Revisiting Table 4, the last row shows the coverage and accuracy after considering additional session blocks (with neither OSN identifiers nor traffic markers) attributed to users using their activity fingerprints. By setting $k = 5$ which yields $\Psi = 0.98$, we see that the session coverage increases by $28.8\%$, from $49.8\%$ in "Traffic attribution" to $78.6\%$ in "Activity analysis." The user coverage in-

| Tessera | Sub-classes | | | | | | |
|---|---|---|---|---|---|---|---|
| Demo-graphics | Name (12420/78.7%) | Email addresses (5336/33.8%) | Screen name (1696/10.8%) | Birthday (665/4.2%) | Phone number (138/0.9%) | Gender (9159/58.1%) | Web page (389/2.5%) | Profile picture (11758/74.5%) |
| Location | Current residence (4882/30.9%) | Prior residence (4472/28.3%) | Coordinates (618/3.9%) | Zip code (618/3.9%) | Time Zone (1090/6.9%) | City, state (1083/6.9%) | | |
| Affiliation | Employer (308/2.0%) | Current employer (2920/18.5%) | Prior employer (682/4.3%) | | | | | |
| Education | Current school (1395/8.8%) | Prior school (2257/14.3%) | High school (4387/27.8%) | College (2291/14.5%) | Visit to .edu (153/1.0%) | | | |
| Social Association | *OSN1* friends (6476/41.1%) | In relation with (339/2.1%) | Email exchanges (38/0.2%) | Chat (7163/45.4%) | *OSN2* followers (1696/10.8%) | | | |
| Social Activity | OSN (10492/66.5%) | Dating (372/2.4%) | Blog (946/6.0%) | | | | | |
| News Information | Search (4456/28.2%) | Search Queries (1256/8.0%) | News (2045/13.0%) | Map (1954/12.4%) | Weather (1351/8.6%) | | | |
| Content Exchange | File hosting (2835/18.0%) | P2P applications (295/1.9%) | | | | | | |
| Enter-tainment | Game (1661/10.5%) | Travel (237/1.5%) | Favorite team (1376/8.7%) | Sports (310/2.0%) | | | | |
| Art & Culture | Music (5298/33.6%) | Music genre (533/3.4%) | Video (5276/33.4%) | Book (2722/17.3%) | Picture (770/4.0%) | Religious views (351/2.2%) | Political views (280/1.8%) | |
| E-commerce | Shopping (3745/23.7%) | Shopping categories (451/2.8%) | Shopping queries (44/0.2%) | Banking (1121/7.1%) | | | | |
| Device & Traffic | Device name (15775/100%) | OS types (15775/100%) | App info (15775/100%) | Is hotspot (1037/6.5%) | Traffic info (15775/100%) | Session info (15775/100%) | Timing info (15775/100%) | |

Info. obtained from public OSN profiles.
Info. obtained by activity analysis on data trace.
Info. obtained from both data trace and public OSN profiles.

**Table 6: Tesserae and sub-classes of Mosaic with number of users annotated**

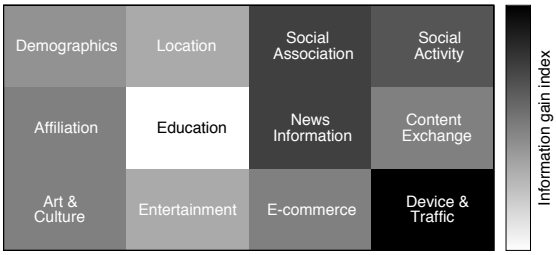



**Figure 8: Mosaic of an exemplary user Alice**

creases by 25.8% to 69.0%. An average of 586 sessions get labeled with a user, and on average those sessions last for 82.3 minutes. Session accuracy decreases by 2% as mentioned earlier and user accuracy drops by 2.9% to 96.4%.

## 4.2 Tessellating the User Mosaic by Gleaning Information from Various Sources

Here we describe the last step of the Tessellation methodology (see Figure 3). In summary, to build the user mosaic, we glean information from the following two sources: (i) user activity analysis as described earlier, which reveals not only the types of activities a user engages in, but also how much time she typically spends on each activity; and (ii) publicly available pieces of information about the user that can be crawled from the Web (e.g., those voluntarily disclosed in her public OSN profiles). For the remaining of this section, we describe this step using an example user from our datasets, which we refer to as "Alice."

Figure 8 shows a visual representation of the mosaic for Alice. We broadly classify the personal information gleaned from the data into twelve classes (tesserae), each of which contains a corresponding set of sub-classes described in Table 6. The hue of the tesserae indicates the amount of sub-class information gleaned. The darker the color, the more information we learn about the user. In the following, we present examples of the personal information that is gleaned from the network traffic, as well as the information crawled from the Web. For privacy concerns, we redact details that we deem sensitive (e.g., *OSN1*/*OSN2* IDs, physical addresses). For Alice, both her *OSN1* and *OSN2* IDs are leaked, which allows us to compare information across different OSNs and manually validate the created profile. In fact, we observed that her *OSN1* and *OSN2* ID's co-occur in multiple session blocks. In addition, by comparing her *OSN1* and *OSN2* public profiles, we find that the names, as well as key demographic information, match.

In general, the "publicly available" information extracted from crawling the OSN sites or searching the Web is at a coarser granularity and more static, as compared to the information collected from network traffic. For example, a user may disclose in her public OSN profile her city and state of residence, affiliation, education history, and her interests. But, typically, she will not disclose her precise home or work address, where she is right now, whom she has just messaged, what songs she listened in the past hour, and other pieces of information that are dynamic in nature. In the case of Alice, by crawling her *OSN1* page `http://www.osn1domain.com/profile.php?id=<OSN1_ID>`, we find her first and last name, the city where she lives in (City X, State Y) and where she comes from (City Z, State Y), her favorite TV shows (Sex and the city, etc.), and music artists (Bob Marley, etc.). By querying the *OSN2* API with her `OSN2_ID`, we obtain her *OSN2* screen name and time zone (GMT $-5:00$, Eastern time). In her profile, she does not disclose anything about her education background and searching the Web does not provide additional information either. This is why the "Education" tile in Figure 8 has the lightest color (white).

On the other hand, extracting information from the digital footprints left by Alice in the network reveals a lot more about her. From her activity analysis, we find that Alice spent 72% of her time (1.93 hrs out of 2.66 hrs) in shopping goods in three different e-commerce sites (`craigslist`, `amazon`, `ebay`). In the majority of her remaining time (0.6 hrs), she moved back and forth between `osn1domain.com` and an *OSN1* game app. In the mean time, her computer updated its OS from `windowsupdate.com` and virus signature from `symantecliveupdate.com` in the background. Aided by the user activity analysis and classification, we have also developed tools to mine and extract specific types of information. In the following, we provide some examples of such information.

**Location information.** We identify various location-based services (*e.g.*, map search, weather) that periodically transmit the devices' location information to servers. We then extract the users' coordinates in the from of longitude/latitude, and zip codes, along with precise timing information. For example, using Alice's cookies from `weather.com`, we match keywords such as `lat&lng` and extract her coordinate information over time. The extracted GPS locations are within the 10 mile perimeter of City X, State Y, confirming her residence as listed in her *OSN1* profile. Moreover,

as many automatically updating weather apps do, her GPS information is logged every 30 minutes, allowing us to build a trajectory of her whereabouts. For the first 1.5 hours (8:30-10:00am) of our trace, she stayed near a highway interchange. The next coordinate logs a shopping mall one exit south on the highway. Then after 30 minutes, her GPS indicates she returned to the first place she was located (which we believe is her residence). As shown in this case, a time-lapse analysis on the coordinates enables us to infer the residence and work place of a user.

**Social associations within and outside OSNs.** The digital footprints left by an OSN user can also reveal her social associations and interactions not only within an OSN (despite such information may not be disclosed in the public profile), but also *outside* of the OSN. An instance of social association *within* an OSN is a cyber-gift exchange: an HTML link to a game app's micro-credit transaction in *OSN1* sent by a friend of Alice is shown as below:

```
<div onclick = "...InterstitialOverlay('mystery
gift', <friend ID>, <friend name>)">
```

Social associations *outside* OSNs may be discovered from things like users' email exchanges or instant messages (IMs). For example, most instant messaging services transmit users' email or nicknames, as well as conversations in plain text as seen in the following IRC snippet:

```
NICK:BOBxxx, USER:CHUCKxxx <IP 1> <IP 2> <Message>.
```

**Device information.** By parsing HTTP header fields, such as the `User Agent`, we obtain information about the device a person uses to access the Internet. The device information includes device names (*e.g.*, Thunderc, iPad, VM670/8), OS types (*e.g.*, Android 2.2, iPhone OS 4.2.1, Windows 7 x64), Web browsers (*e.g.*, Mobile Safari), and some of the apps installed on the devices (*e.g.*, Pandora, eHarmony, Twitter, iTunes). In the case of Alice, she owns a computer with Windows XP and a smart phone with Android 2.2.1.

**Mining other static and dynamic information.** To obtain certain information of interests, we can further perform in-depth analysis of specific services or sites by parsing their HTTP headers. For example, from the music service provider `pandora.com`, we can extract musical genres that Alice likes. From her activities at e-commerce sites such as `ebay.com` and `craigslist.org`, Alice's shopping interests can be extracted based on the categories of goods she peruses and keywords she uses to search such goods. During the nearly 2 hours she spent on the 3 e-commerce sites, Alice performed four queries, visited 88 product pages, and browsed 229 item images.

# 5. QUANTIFYING PRIVACY LEAKAGE

In this section, we apply Tessellation to the cellular network datasets and quantify user privacy leakage. First, we measure the amount of privacy leakage as we vary the duration of observation and the number of IP addresses being analyzed (§5.1). Then, we study the leaked information by comparing the amounts and types of information disclosed on OSN profiles vs. those from network traffic (§5.2). We also demonstrate unique discoveries that can be made by associating OSN profiles with data extracted from the traffic (§5.3). We conclude by discussing ways to prevent the leakage of private information (§5.4).

## 5.1 Quantitative Analysis of Privacy Leakage

Running Tessellation on the 3h-Dataset, we extract significant information about each user. In Table 6, for specific types of information, such as the name and the political views, we list the number and percentage of users that we manage to get information about. Each row represents an information class and individual cells represent the sub-classes. The information for each class/subclass is

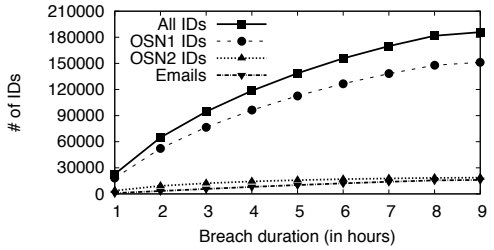



**Figure 9: User IDs captured in various time durations**

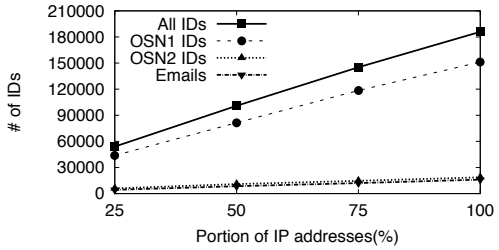


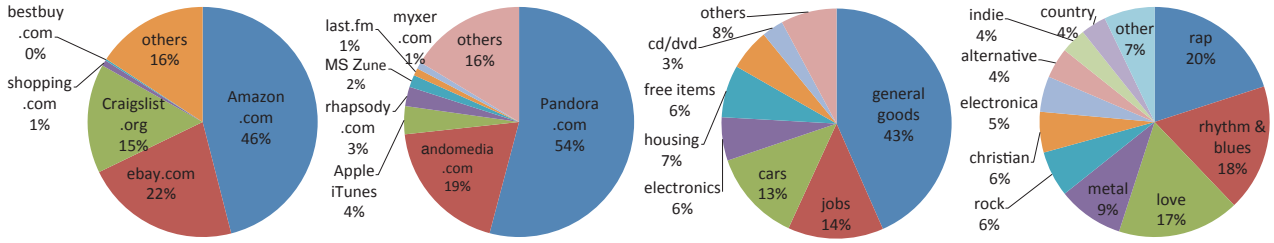

**Figure 10: Users captured in various portions of IP addrs.**

obtained via a combination of activity analysis, site-specific mining and crawling the OSN public profiles of the users as discussed in §4. From the table, we see that the Tessellation tool provides a wealth of information about the users.

**Privacy leakage as a function of breach duration and compromised IPs.** We quantify the amount of privacy leakage as we vary the duration of observation periods and the number of IP addresses being captured. Because the longevity of data is being tested, we use 9h-Dataset for these tests. The 3h-Dataset exhibited similar results as the 9h one; due to space limitation, they are omitted. Figure 9 shows the cumulative number of user identifiers leaked as the breach duration increases from 1 to 9 hours. Due to the sub-linear increase of leaked IDs over time, we see that around half of all user identifiers in the entire 9h-Dataset are leaked in the first three hours, and more than 80% of the identifiers are leaked before half (4.5 hours) of the total breach duration. The trend is most prominent with OSN1 IDs; similar observations apply to OSN2 IDs and email addresses as well. Hence, if only 1/3 or 1/2 of the total duration of 9h-Dataset were collected, an adversary may still be able to glean information about 50% or 80% of the users. Figure 10 shows how varying the number of IP addresses captured in the data affects the number of user identifiers leaked in the data. Out of 340,000 client IP addresses from 9h-Dataset, we randomly select 25%, 50%, 75% of the addresses and calculate the number of leaked user identifiers from them. Again, with somewhat limited data, an adversary may still be able to glean information for a significant number of users.

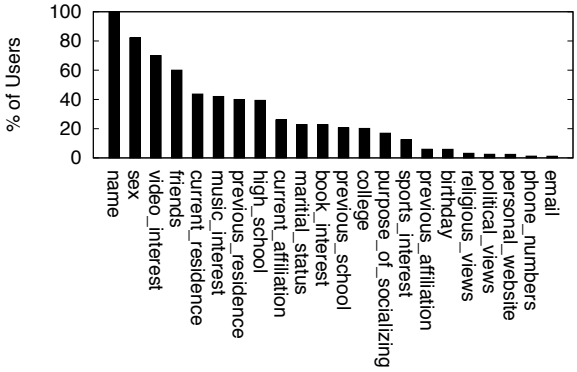
## 5.2 Comparison of Information Disclosed on OSNs vs. Leaked in the Network Data

**Information disclosed by users in their OSN profiles.** Most OSN sites provide privacy control "knobs" that allow a user to control what information is publicly disclosed (*i.e.*, in her "public" profile), what is only disclosed to "friends," and so forth. By crawling the *public* (*OSN1* and *OSN2*) profiles of the OSN users in the 3h-Dataset, we extract 25 personal attributes listed by these two OSNs. In Figure 11(a), we plot the percentage of users who have disclosed some information for each of these attributes. We see that all users have disclosed their names, and more than 80% have disclosed their gender. A plurality (more than 40%) of them have also disclosed coarse-grained information about their locations ("current_residence"), their online friends, personal interests ("video", "music", or "sports_interest"), and schools attended ("high_school"

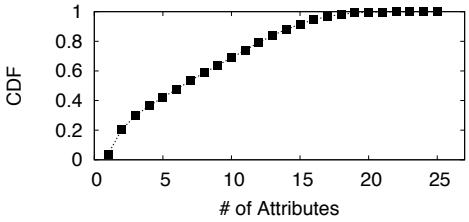



(a) Shopping by provider (b) Music by provider (c) Shopping by category (d) Music by genre

**Figure 12: Examples of detailed information disclosed by activity attributes and domain-specific mining**




(a) Amount of users with attributes open to public




(b) Cumulative distribution of attributes user put on public

**Figure 11: OSN attributes users disclose**

or "college"). On the other hand, very few have disclosed any information regarding their phone numbers, email addresses, personal websites, or their workplaces ("current_affiliation"). As explained in [13], the differential treatment by users to various types of personal information (personal attributes asked by the OSNs) can be attributed to users' concerns with sensitivity and identifiability. Less identifiable personal attributes, such as gender or interests in music, are generally deemed less sensitive by OSN users; further, they can help attract potential friends with shared interests or connect with old acquaintances (*e.g.*, high school classmates). On the other hand, highly sensitive and/or identifiable information such as emails, phone numbers, and work places are usually kept hidden to avoid being misused.

Figure 11(b) shows the number of attributes each person discloses. The average number of personal attributes disclosed is 5.3. About 20% of users disclose some information on 12 or more personal attributes, and a very small percentage have disclosed all 25 attributes. Further in-depth analysis shows that none of the users who disclose 5 or fewer attributes reveal their phone numbers or email addresses. In contrast, the 138 users (0.87%) who put their phone numbers in their public profiles disclose an average of 14.4 attributes, a three-fold increase compared to the average among all users. The high correlation between the disclosure of phone numbers and other attributes leads us to suspect that a portion of the 138

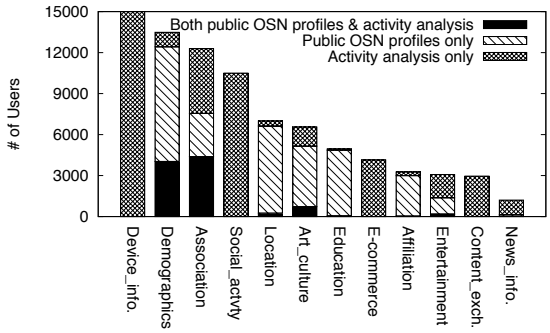



**Figure 13: Personal info gathered from public OSN profiles, activity analysis on network trace, and both**

users may have made a mistake with the privacy "knobs", showcasing the need for a better user-controlled privacy management [2,9].

**Information gleaned from the network data.** Having analyzed the publicly disclosed profiles, we now analyze and quantify what additional information can be gleaned from the network trace, as well as the granularity and category of the leaked information. For each major tessera of information (the row class) listed in Table 6, we count the number of users who either (i) disclose information in public profiles (cells shaded in light gray), (ii) have additional information leaked and mined in network data (cells in gray), or (iii) have both public profiles and leaked network data available (cells in dark gray). Figure 13 illustrates the proportion of personal information by each of the three sources. As previewed in Figure 1, while the information disclosed in the public profiles covers many classes (*e.g.*, demographics, location, education, and affiliation), it is generally coarse-grained (and thus less identifiable and less sensitive) and *static* (and thus less timely). In contrast, the information leaked on to the network is finer-grained and/or *dynamic* which closely reflects users' cyber and real-world activities and interests. The availability of overlapping information in both public profile and leaked network data helps to confirm our findings when comparing one side with the other.

As two illustrative case-studies, we provide an in-depth analysis of online shopping and music preferences as gleaned from running Tessellation. Using the 3h-Dataset, we first compute the amount of time the users spent on various e-commerce sites (resp., online music sites); for each service provider, we then tally the total amount of time that users spent on each site. Figure 12(a) and (b) show the major shopping service providers and music service providers, respectively. In both cases, a few service providers dominate each market (the top three services take up more than 75%), despite having a large number of services existing and competing with each (16% of the services account for less than 1% of the total time the user spent). Although here we use the duration of stay as the metric to compare various service providers, similar results are obtained when using frequency of visits, traffic volume, and so forth.

| Device type | Traffic (in average session count / device) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Web | Multimedia | Control | P2P | VoIP | Email | Chat | Total |
| Mobile | 37.94 | 1.30 | 15.00 | 0.01 | 0.02 | 0.54 | 0.21 | 55.02 |
| Stationary | 334.65 | 216.72 | 271.57 | 10.48 | 0.28 | 0.16 | 0.10 | 833.96 |

**Table 7: Traffic patterns on mobile vs. stationary devices**

In Figure 12(c) and (d), we provide a detailed breakdown of shopping categories and music genres. The shopping categories are extracted from HTTP GET messages of `craigslist.org`, where we decode its three-lettered category code (*e.g.*, AOS: automotive, BKS: books, ELE: electronics). In the case of music genres, we extract HTTP GET messages from `pandora.com` and extract nominal values from key-value pair `genre=<value>`. Here again we use the shopping categories/music genres simply as examples to illustrate the more specific information that may be gleaned from the network data. Similar analyses can be conducted on the queries sent to search engines, categories of videos and books being viewed, and so forth.

Through the experiments in this section, we show that the information disclosed by users on their public profiles and the information that is leaked and gleaned from the network data are often complementary and corroborative. When combined, they produce a richer mosaic of users, thereby posing a more severe threat to user privacy.

## 5.3 Other Usages of Tessellation: Examples

To demonstrate utilities of Tessellation other than quantifying privacy leakage, we introduce a few experiments that reveal deeper knowledge on users by associating different user attributes (Tesserae) together.

**Traffic breakdown among devices and apps.** One straightforward application of Tessellation is cellular data traffic analysis. For instance, using Tessellation we can sample and separate traffic generated by various mobile devices such as smart phones/tablets running Google Android, Apple iOS, and traditional "stationary" devices such as 3G/4G-equipped laptops/netbooks running Microsoft Windows OSes. In addition, Tessellation can provide application traffic breakdown and statistics for each type of device. Hence, this analysis is done by associating "OS types" and "Traffic info" sub-classes from the Device tessera together.

Not surprisingly, the cellular network is dominated by mobile devices. However, we also found a significant number of 3G-equipped stationary devices, as they are used by many businesses. Table 7 shows the traffic volume breakdown between mobile vs. stationary devices, in terms of the average number of sessions generated by each device among popular applications. We observe that, while there are relatively few of them, the stationary devices generate far more traffic per device, which is in agreement with the finding in [6]. The dominant applications on both categories of devices are Web based. Perhaps more interestingly, we see significantly more P2P (mostly Bittorrent) and Game/Video (mostly UDP-based) traffic from "stationary" devices, whereas there is considerably more Email (IMAP, POP3, SMTP) and Chat (XMPP, SIP, MSN, YahooIM) traffic from mobile devices. One possible reason for the dominance of mobile devices in email and chat traffic can be the periodic background updates and push services persistently running in smart phones and tablets.

**Age demographics of app usage.** Looking at users who disclose their age on the OSNs as samples, Tessellation can provide insight into how users of different age groups use a variety of apps. Figure 14 shows the usage statistics for five popular categories of apps running on all devices among age groups of 19-29, 30-39, 40-49, and 50+. For each app category, we count how many sessions in that category are generated by each user within an age group,

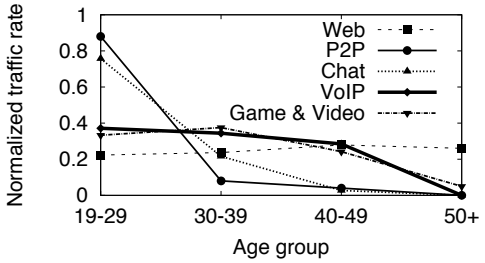



**Figure 14: Age demographics on the usage of apps**

and we compute the average number of sessions per user for each group. To compare the usage statistics across different app categories, the average session statistics per age group is normalized by the sum of all four groups and shown on the y-axis. As expected, we see that P2P and Chat applications are more common among young users. Moreover, VoIP and Game/Video apps are favored by the two younger groups, 19-29 and 30-39, and rarely used by the oldest age group, 50+. On the other hand, Web apps exhibit a relatively even distribution among all four age groups, suggesting that the Web is equally popular among all age groups.

As shown in the above examples, Mosaic can give deeper knowledge on the target users when different sources of information are put together. This is what makes the network-wide attacks to be far more threatening than security breaches on each individual service.

## 5.4 Preventing User Privacy Leakage

Here, we provide a brief discussion on the challenges in protecting user privacy, followed by approaches for preventing and mitigating the leakage of sensitive information.

From our study, we see that a key enabling element of privacy leakage is the identifiable information (*e.g.*, OSN IDs) leaked by OSNs and other services. Better Web design can alleviate the problem. However, due to the distributed and stateless nature of the Web, problems like cross-site scripting and the use of cookies may continue to cause problems. While growing number of services use encryption (*i.e.*, HTTPS/TSL), our study shows that an adversary can still attribute a significant portion of user traffic. For example, even for a service that fully operates over HTTPS, user IDs still leak from its third party mobile app.Provided that our method of traffic association relies heavily on leaked OSN IDs, use of authentication schemes that grant temporary access tokens to apps and browsers [5, 25], can prevent user credentials from being used as seeds for identity discovery. However, even without any OSN IDs, we can reconstruct a significant amount of information by using traffic markers (third-party tracking cookies), as shown in Table 4.

Although global adoption of end-to-end network traffic encryption could prevent the leakage, we envision that it is not likely to happen in the near future. Incremental adoption of encryption can alleviate the leakage problem and reduce the attack surface if it is implemented in the right way. We list some specific advice on preventing the leakage as following: (1) The usage of unique user/device identifiers should be carefully limited, and those identifiers should be strongly encrypted whenever it is necessary to transfer them in network traffic. (2) Tracking cookies and HTTP session identifiers, which are commonly used in today's Web services, should be encrypted or frequently updated. (3) The public profiles of OSN users should have certain attributes to be carefully obfuscated so it is hard for someone to link them together with the information in network traffic. (4) A service provider, such as an OSN, should have mechanisms to enforce third parties involved in the service, such as individual app developers, to obey its privacy guidelines.

# 6. RELATED WORK

**Leakage of personal information.** The leakage of personal information through online activities has attracted significant attention over the last years. Krishnamurthy et al. presented a range of studies [13, 15, 16, 18] that highlight how "personally identifiable information" (PII), such as email address, age, zipcode, and gender, is leaked via HTTP headers, URIs, and cookies. Working on privacy control, Persona [2] and NOYB [9] are online social networks that put users in control of their own social and privacy information. Fang et al. [8] presented a template for the design of a privacy wizard, which removes the burden of specifying security settings from the end users. Aggregating personal information across different services of same kind, HostTracker [30] employed a number of application level IDs in tracking hosts in email networks. Irani et al. [11] quantified leakage of personal information by combining different aspects of user footprints from multiple OSNs. All the above studies highlight the possibility of an entity collecting personal information from a single Web service or a group of similar Web services such as OSNs. Our work differs from them in that it operates on a network trace to attribute user sessions and construct a well-connected content-rich user mosaic by systematically combining isolated pieces of information across vastly different services, and without requiring explicit collaboration among them.

**Data de-anonymization.** Narayanan et al. [21] formulated identity discovery (or identity de-anonymization) into a sub-graph isomorphism in social graphs. Further on de-anonymization, Mudhakar et al. [19] provided a way to discover identities of mobile users by associating the users encounters in the physical world with their social graphs in the cyber space. In other words, all the above papers try to de-anonymize traces that were intentionally anonymized previously. This is a very different problem from the one we address here, because we focus on attributing network sessions to user identities extracted from traffic data.

**User activity profiling.** A group of studies attempted to profile users based on browsing habits and the types of applications used. Using data from a CSP, Keralapura et al. [12] showed that there exist distinct behavior patterns among mobile users. Trestian et al. [29] characterized the relationship between user application interests and their mobility properties. On a large scale, there have been studies characterizing mobile traffic and user interactive behaviors on embedded applications with smart phones [31]. While these studies focus on the potential of using the distinctive behavior of users as a way of identifying them, they do not propose actual methodology. In our work, we step forward by analyzing the distinction among different users' activity patterns and leveraging it to associate network traffic with known users' identities.

# 7. CONCLUSION

In this paper, we study the privacy leakage problem in mobile network data. We bring forth two key insights. First, the prevalence in the use of OSNs leaves identifiable digital footprints in the network. Second, the indiscriminate use of tracking techniques by mobile apps and services makes traffic attribution easier. The combination of these factors allows an adversary to attribute significant portions of traffic with *NO* explicit leaks of the users' true identities. To demonstrate the feasibility of the threat, we developed *Tessellation*. Using the network data from a CSP, we showed that up to 50% of the traffic can be attributed to users with high confidence. Further, we illustrated how various types of information can be gleaned about the user by painting a content-rich digital *mosaic*, and we demonstrated utility of this information to extract collective trends.

**Acknowledgments:** We would like to thank Nina Taft and the anonymous reviewers for their helpful comments and suggestions.

# 8. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *World Wide Web (WWW)*, May 2007.

[2] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: an online social network with user-defined privacy. In *SIGCOMM*, Aug 2009.

[3] M. Balakrishnan, I. Mohomed, and V. Ramasubramanian. Where's that phone?: geolocating ip addresses on 3g networks. In *IMC*, Nov 2009.

[4] S. M. Bellovin. A technique for counting natted hosts. In *ACM SIGCOMM Workshop on Internet measurment*, Nov 2002.

[5] E. D. Hardt. The oauth 2.0 authorization framework, ietf rfc 6749, 2012. http://tools.ietf.org/html/rfc6749.

[6] Ericsson. Traffic and market data report, Nov 2011. http://www.ericsson.com/res/investors/docs/2011/cmd/traffic_and_market_data_report_111107.pdf.

[7] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. A first look at traffic on smartphones. In *IMC*, Nov 2010.

[8] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *World Wide Web (WWW)*, Apr 2010.

[9] S. Guha, K. Tang, and P. Francis. NOYB: Privacy in Online Social Networks. In *WOSN*, Jun 2008.

[10] K. M. Hendrik Schulze. Internet study 2008/2009, ipoque. http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf.

[11] D. Irani, S. Webb, K. Li, and C. Pu. Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing*, pages 13–19, 2011.

[12] R. Keralapura, A. Nucci, Z. Zhang, and L. Gao. Profiling users in a 3g network using hourglass co-clustering. In *MOBICOM*, Sep 2010.

[13] B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. Protection measures: the growing disconnect. In *W2SP*, May 2011.

[14] B. Krishnamurthy and C. Wills. Characterizing privacy in online social networks. In *WOSN*, Jun 2008.

[15] B. Krishnamurthy and C. Wills. On the leakage of personally identifiable information via online social networks. In *WOSN*, Aug 2009.

[16] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In *World Wide Web (WWW)*, Apr 2009.

[17] F. Lardinois. PleaseRobMe and the Dangers of Location-Based Social Networks. *ReadWriteWeb*, Feb 2011.

[18] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *IMC*, Nov 2011.

[19] S. Mudhakar and M. Hicks. Deanonymizing mobility traces: Using social networks as a side-channel. In *CCS*, Oct 2012.

[20] C. Mulliner. Privacy leaks in mobile phone internet access. In *Intelligence in Next Generation Networks (ICIN)*, Oct 2010.

[21] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. In *IEEE Security and Privacy (S&P)*, 2009.

[22] Netresec. Publicly available PCAP files. http://www.netresec.com/?page=PcapFiles.

[23] K. Nohl. Wideband GSM sniffing. In *The 27th Chaos Communication Congress*, Dec 2010.

[24] K. Nohl. Defending mobile phones. In *The 28th Chaos Communication Congress*, Dec 2011.

[25] OpenID Foundation. Openid authentication 2.0, Dec 2007. http://openid.net/specs/openid-authentication-2_0.html.

[26] C. Riederer, V. Erramilli, A. Chaintreau, and P. Rodriguez. For sale: Your Data By: You. In *ACM HotNets*, Nov 2011.

[27] C. Rigney. Remote authentication dial in user service (radius), ietf rfc 2866, 2000.

[28] C. Rigney, S. Willens, A. Rubens, and W. Simpson. Radius accounting, ietf rfc 2865, 2000.

[29] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Googling the internet: Profiling internet endpoints via the world wide web. *IEEE/ACM Transactions on Networking (TON)*, 18(2):666–679, 2010.

[30] Y. Xie, F. Yu, and M. Abadi. De-anonymizing the Internet Using Unreliable IDs. In *SIGCOMM*, Aug 2009.

[31] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman. Identifying Diverse Usage Behaviors of Smartphone Apps. In *IMC*, Nov 2011.