Final manuscript of article to be published in The ACM Magazine XRDS: Crossroads. 21(3), Spring 2015 - DOI: 10.1145/2730912

Exploring the Risks to Identity Security and Privacy in Cyberspace

Jason R. C. Nurse

Cyber Security Centre, Department of Computer Science, University of Oxford

jason.nurse@cs.ox.ac.uk

Cyberspace, a world of great promise, but also, of great peril. Pirates, predators and hackers galore, are you and your online identity at risk in this wild frontier?

Setting the Stage

The Internet – that burlesque theater where the good, the bad, and the ugly of digital technology all plays out. As I write this article, Staples Inc. has announced the possibility of around 1.16 million payment cards being compromised by a data breach that took place between July and September 2014 [1]. This is but one of many recent high-profile leaks that have exposed the personal identity information of millions. While it is imperative that we reflect on how and why such breaches occur, it would also be prudent to consider areas where the security and privacy of our online identities might be at risk in the future. This article reflects on the most topical of these areas with the aim of elucidating the risks that we all face in our use of technology and by our very existence in cyberspace. We shall embark on a research journey that explores identity theft in the realm of social media, email and everyday computing devices. These are technologies familiar to us all and, for many people, embedded in the very fabric of their lives. But I often wonder if we are aware of the risks involved in their usage?

Read on, I guarantee you will discover something new and, perhaps, even something scary.

Not-So-Social Media

I would be willing to place a small wager (in Galactic credits) that everyone reading this article has a presence, however inconspicuous, on social media. This fact is testament to the scale of social media, which is nothing short of astounding: there are hundreds, if not thousands, of social-media services today catering to every interest and allowing a wide range of interactions from information publishing and networking, to gaming and purchasing. Unfortunately, as social media has grown, so too has its allure for hackers. The large number of teenage and elderly users, the expanding demographic, and the sad reality that many users are unaware of the security risks involved or how to protect from them, all combine to make social media an ideal hunting ground for online predators.

An especially risky practice of social media users is *oversharing*. Oversharing, as the name suggests, refers to the sharing of excessive amounts of information online about one's identity, movements, and

activities. We all have that one friend or colleague that tweets at least twenty times a day, documenting when they leave home, the Starbucks they are currently at, the newly-released film that they are about to watch, and how bad the traffic is on the way to Chiang Mai, their favorite local Thai restaurant. The problem here is not the number of posts but their content, and the fact that this information can be used for various malevolent purposes including stalking or home burglary. There have been a number of cases in recent times where criminals used Facebook and Twitter to find empty homes to rob [2]. The issue of oversharing goes beyond what people post online to what information they make available on their public profiles. Posting addresses, birth dates, and phone numbers all have associated risks. Of course, for many individuals, the question is, what are those risks?

In our efforts to provide an answer, my colleagues and I have designed and implemented a model that can elucidate the security and privacy risks of online information sharing (and oversharing) [3]. The model is realized in three key tasks.

The first task is the definition of the full range of identity attributes, hereafter referred to as *elements*, people tend to make use of online. Examples include the name, username, avatar, address, and interests of that particular individual. This allows us to gain a thorough understanding of the identity information present online. The second task is the discovery and documentation of the various methods and techniques published that enable new identity information to be inferred from known elements. A simple example of this is a heuristic that states how an individual's username and email address (jsmith and jsmith@gmail.com) can be inferred from their real name (John Smith). More complex techniques may dig deeper, allowing inference of everything from age and friends to ethnicity and birthplace. The final task is the combination of the elements and inferences into a comprehensive, chained model, where elements are linked by inferences, and risks can be seen based on actual information exposed.

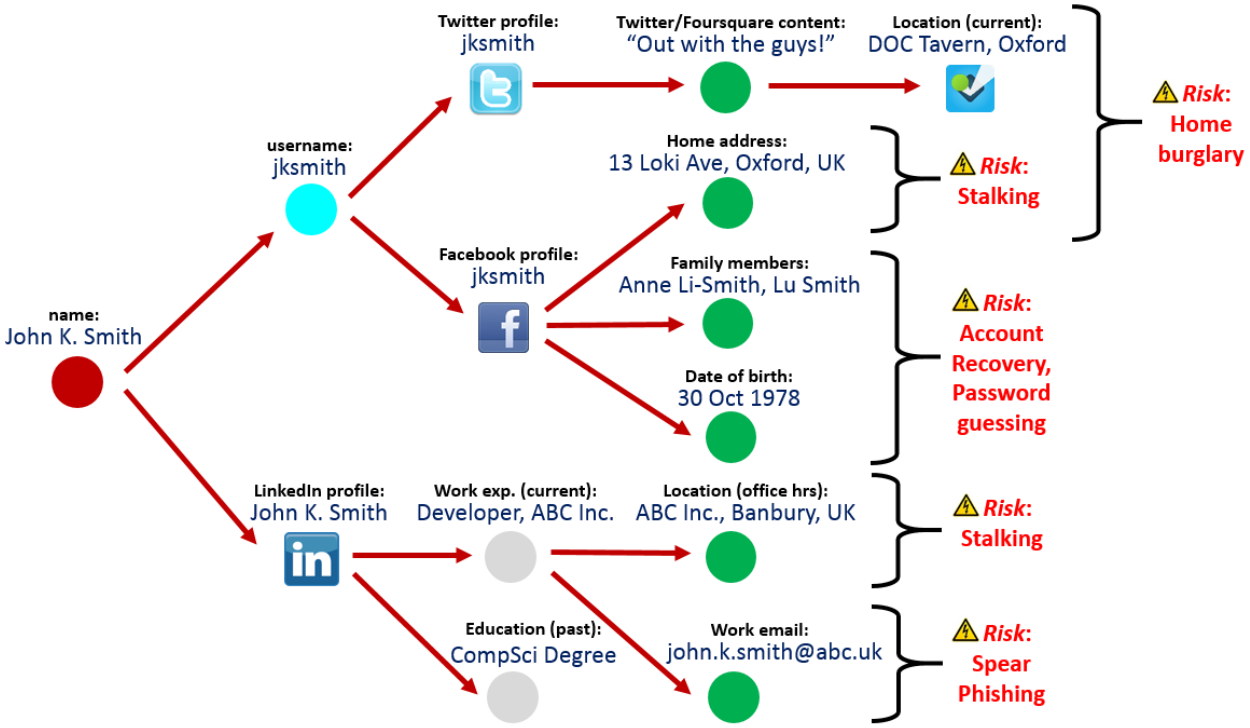

Figure 1. Applying the model to a fictitious scenario to assess the potential risk of sharing information online. Identity elements are represented as nodes, the inferences possible from those elements as arrows, and the risks to particular information being used by an attacker is presented in red text.

Figure 1 demonstrates some of the possible inferences and, thus, the new information that might be discovered about an individual given only their name. It can be seen that even with the most basic initial information an attacker could discover where John K. Smith works, where he lives and his exact current location.

The real benefit of our model is that it highlights the risks of sharing information on social media in an easy to understand way, making the data more palatable for lay users. In essence, it allows users to reflect in a way they never could before; to ask the question, 'If I share information X (or indeed, a set of information Y) online, what further information could others users infer about me?' And more crucially, 'What are the risks, both offline and online, that could result from such inference?' This can be an extremely useful tool in the fight against oversharing and in raising awareness about the risks involved.

Send Me an Email and I Will Tell Your Life Story

Since its inception in the 1960s, email has revolutionized the way we communicate allowing near-instantaneous interaction between individuals across the globe. More importantly, it's completely free for individuals and cheap for businesses. This has led to the monopolistic popularity of email as a communication medium within business, government and social spheres. How popular you ask? Well, current estimates suggest that 204 million emails are sent per minute [4]. That's right, per minute!

The security and privacy of email communications has come a long way as well. There have been proposals to protect the confidentiality of emails through encryption, the authenticity and integrity of emails using digital signatures, the availability of emails by increasing server resilience, and email recipients themselves by continuous scanning for spam, viruses and phishing attempts. A conclusion that might be drawn from the current emphasis of research and practice, therefore, is that the entities posing a threat are all external and must, at all costs, be prevented from getting in. The question I pose to you is this: is there any risk to the identity of an individual or their employer based on what is transmitted in an *outgoing* email?

The simple answer: yes, there is. And it gets worse; the content of the message is irrelevant and even blank mails can leak sensitive identity or organizational information. The problem lies in email headers. These are an essential component of every message and contain metadata about the sender, the receiver, and the route traversed between the two parties. The worry with these headers is that the email applications in use today are including an increasing amount of metadata about individuals with little regard for the privacy implications and associated risks. For instance, headers can expose the type of device used, the email client, its version number, and, in some instances, the internal username of the individual; see Figure 2 for an example. Moreover, as in the case of social media, many individuals – not you, of course – are unaware of the risks involved and the identity information that is very likely being exposed.

```
Received: from mailer.abc.uk ([XXX.XXX.XXX.XX2]) by
  mailtr.an.organisati.on with esmtp (Exim 4.80) id 1Y25L6-0002zV-oA
  for joe.monday@an.organisati.on; Sun, 14 Dec 2014 18:30:01 +0000
Received: from [XXX.XXX.XXX.XX1] (helo=[192.168.1.8]) by mailer.abc.uk
  (Postfix) with ESMTPSA (envelope-from <john.smith@abc.uk>) id
  2312312132; Sun, 14 Dec 2014 18:30:00 +0000
Content-Type: multipart/alternative;
  boundary="Apple-Mail-QWNHJDNF-KJDD"
MIME-Version: 1.0 (1.0)
Subject: Information on the purchase order
From: John Smith <john.smith@abc.uk>
X-Mailer: iPad Mail (12B440)
Date: Sun, 14 Dec 2014 18:30:05 +0000
CC: Pat Tuesday <pat.tuesday@an.organisati.on>
Content-Transfer-Encoding: 7bit
Message-ID: <KJDKUJSK-EOI-K2LS-DJFD-KJSD49SEI@abc.uk>
To: Joe Monday <joe.monday@an.organisati.on>
X-Username: jksmith3
Return-Path: john.smith@abc.uk
```

Figure 2. A simple example of the contents of an email header. Here, we can observe that the sender, John Smith, used an Apple Mail app on an iPad to send the email. Furthermore, we can see that John's internal username is jksmith3. Regarding the organizations, the headers show that the mail transfer agent in use is Exim 4.80 and the email protocols preferred are ESMTP and ESMTPSA.

To investigate whether the leakage of sensitive information in email headers is as real a problem as we imagine it to be, we recently conducted a study where 225 emails from various willing individuals and organizations worldwide were collected and analyzed [5]. Participants in the study were asked to send us empty emails from a variety of devices and applications including desktop email clients, web browsers, and mobile device apps. Our findings were, to say the least, very insightful. An assessment of the email headers collected allowed us to learn much more about the participants than was explicitly known and to determine how the networks of their employers were internally set up.

In terms of individual identity risk, we were able to discover information such as the internal usernames of individuals, that is, usernames used to login to internal organizational sites and systems, the specific email clients that were used to send email (Thunderbird, MacOutlook, iPhone Mail), details of the devices they use (iPhone, Android, Mac OS X), their IP addresses, and their Internet Service Providers (ISPs).

What are the risks? Well, usernames constitute half the information needed to access an employee's email and files, and it could certainly help launch a convincing spear-phishing campaign. With knowledge of the specific email client that an individual uses, a motivated hacker could research and craft attacks meant to exploit weaknesses in the client. This is especially easy in the case where individuals are using outdated clients with well-known vulnerabilities, such as Thunderbird 24.2.0. The details of a person's device can also be used to create customized system attacks. IPs can be useful for locating individuals via IP geo-locators, if only to discover when they are away from home. A simple IP lookup can also furnish details about the ISP of the email sender.

Armed with the information gathered above, a sophisticated social-engineering attack could be launched where an attacker pretends to be the ISP and either gathers more intelligence on the person by challenging for answers to security questions, or directly commits fraud by requesting credit card

details to cover a balance they claim is past due. As social-engineering becomes more common, such attacks will become more likely, posing even higher risks for unwitting individuals.

What Does Your Tech Leak About You? Yes, You!

We live in the digital age. Everything from our computers to our homes are connected to the Internet. The Internet-of-Things is no longer a thing of the distant future; it is already upon us. And does it promise wonders! Let's be honest, who can argue against a refrigerator that senses you are low on milk, and automatically places an online order for groceries to be delivered when your calendar has you scheduled to be at home? Or a car that can relay your location to a nearby automobile-assistance unit if it breaks down?

As with all great inventions however, there are risks. A software bug could result in 10,000 bottles of milk being ordered instead of one. More seriously, an employee at the automobile-assistance company could use their privileged access to GPS data for stalking or home burglary on a massive scale. Especially perturbing is the threat posed by technology that is closest to our lives such as our computers and mobile devices. We use these for everything from checking status updates on social media to gaming and writing emails. But, let me ask again, are you aware of the associated risks? These devices collect, store and even share a vast range of identity information about you. What do they collect? It depends on the device. Who do they share the data with? It depends on who is listening!

A few years ago, I read an enthralling research article that conducted a comprehensive study of the digital footprint that we create through our use of portable technology devices, and how revealing this footprint is of our identities [6]. The authors relied on a peculiar feature of many current portable devices called *active Wi-Fi probing*. This is where a device openly broadcasts a list of *all* wireless networks that it has connected to in the past (Linksys-MyHOME, ISPName-EJGD, StarbucksHS-Oxford) to see if any of them are currently available. You do know that your portable devices can do this once Wi-Fi is switched on, right?

For their study, the researchers set up a monitoring station in a busy city and for a period of one hundred days passively monitored the area for active Wi-Fi probes. Over that period they were able to collect data from over 8,000 devices and on more than 24,000 Wi-Fi access points. This data was then analyzed and ultimately used to infer relationships between the device owners using similarity metrics based on the wireless network lists. Think that's bad for privacy? Subsequent research has demonstrated that it is possible to infer various identity characteristics and even the socio-economic status of device owners based on these probes! This is a serious invasion of privacy. And remember, all this can be achieved without the knowledge of the device owner, that is, you and me.

We conducted a study to determine what other information our much-loved devices might be leaking about us. We focused our efforts at the system layer of devices, particularly the system logs. Why? Because these native logs are found on all technology devices, so any risk to an individual's identity here is of real potential significance. For our experiment, we collected log files from 23 Unix-based systems, including Linux and Mac OS, belonging to 17 willing individuals in our department. An analysis of the scanned log data revealed several facts worthy of the attention of all who use the aforementioned devices. We present only a few of our findings here. Details can be found in [7].

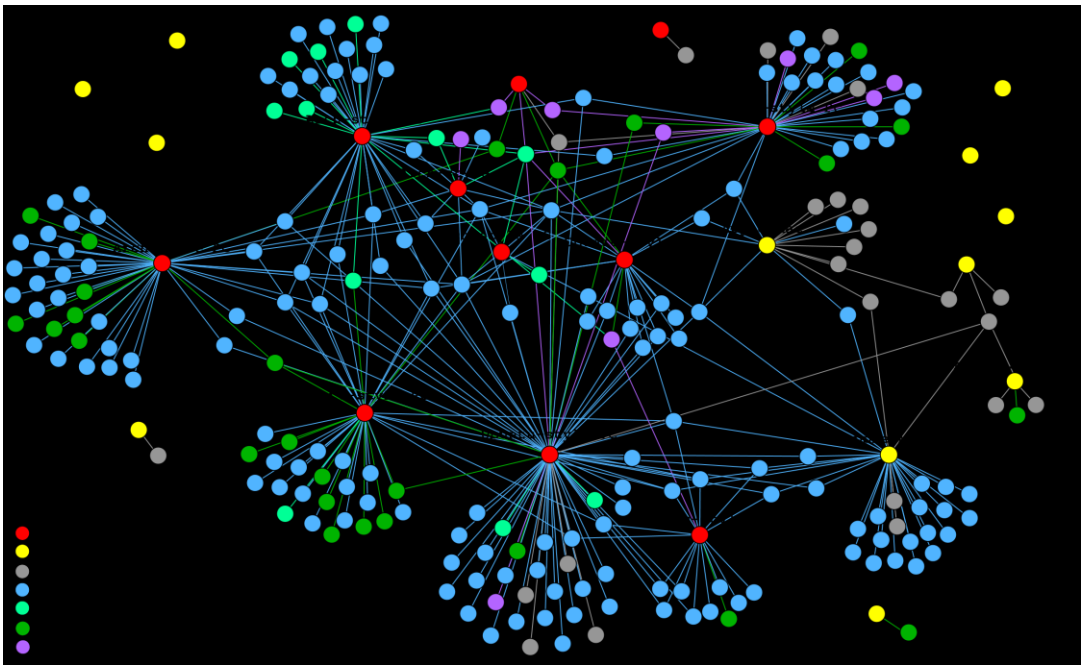

Figure 3. The social network graph of the 23 scanned systems and 249 additional devices discovered in their log files. These include USB and Bluetooth devices, Wi-Fi access points, and gateways. The nodes are color-coded with scanned systems labelled according to their computer type and shown in red and yellow, and discovered systems presented in other colors. Graph edges represent connections between the devices (for example, a USB plugged into a system, or a system connected to a Wi-Fi access point) [7].

We found that devices log much more information than the average person might expect, and while there are probably justifiable reasons from a system-management and debugging perspective, there is a notable risk to a device owner's privacy. Not only were we able to discover relationships between the device owners via shared Wi-Fi points, USB drives and gateways (as shown in Figure 3), but we could also make assertions about the strength of those relationships. Moreover, from an assessment of Wi-Fi names and the dates, times and length of connections, we could make intelligent guesses about which Wi-Fi points were used at home, which at work, and which for social activities. If one were to combine this data with information available on wireless network mapping sites such as Wigle.net, one could identify the location of an individual's home, where they work, and places like gyms or coffeehouses they frequent.

While the data leakage by the device in this case is not as disturbing as the identity information gathered via Wi-Fi probes, given that the log files of a system are not readily available, it does highlight a risk to privacy associated with one of the most fundamental parts of all technological devices. I shudder to think what other sensitive identity information may be exposed by the range of smart devices now available and what a motivated attacker – insider or outsider – might achieve by its acquisition.

Final Thoughts… For Now

Our world today is an amazing place to live in and one of the main reasons for its beauty is technology. It has put man on the moon, created an Internet infrastructure spanning the globe, and put a smartphone in the hands of billions of people. However, as we saw, the use of technology can be a double-edged

sword: in addition to the benefits, there can be some significant risks to the security and privacy of our identities online.

Whether we like it or not, we (and our devices on our behalf) are leaking oodles of sensitive identity information, and others are listening. Thus far, we only know of the burglars (and job recruiters) that monitor social network feeds. However, I cannot help but wonder who else may be snooping and what they plan to do with the data they have gathered. Targeted hacking and widespread identity theft are definitely possibilities. But, with the online advertising market expected to be worth 220.38 billion dollars by 2019, the identity data of individuals, their preferences, and movements are of immense value to advertising agencies for profiling and targeted marketing. Depending on who you are, both the aforementioned scenarios are equally unwelcome.

Finally, consider this. What situation would you prefer the least: a breach at a large corporation revealing your credit card details to hackers, or having a malicious entity capable of gathering all the data you expose and using it to gain access to the very fabric of your identity?

Until next time, I bid you adieu.

Biography

Dr. Jason R.C. Nurse is a Researcher and Lecturer at the Cyber Security Centre, in the Department of Computer Science at Oxford University. His research interests include assessing the risks to identity security and privacy in cyberspace, human factors in security (security usability and human-computer interaction), information trust, and insider threat.

References

[1] Bloomberg. (2014). Staples Says 1.16 Million Cards Affected in Breach. http://www.bloomberg.com/news/2014-12-19/staples-says-1-16-million-cards-affected-in-breach.html [Accessed 21 December 14]

[2] WIBW. (2014). Burglars Use Social Media to Target Wichita Homes. http://www.wibw.com/home/headlines/Burglars-Use-Social-Media-To-Target-Wichita-Homes-258705321.html [Accessed 21 December 14]

[3] Creese, S., Goldsmith, M., Nurse, J.R.C. and Phillips, E., (2012) “A Data-Reachability Model for Elucidating Privacy and Security Risks Related to the Use of Online Social Networks”, in proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2012). IEEE. pp. 1124–1131.

[4] Domo Inc. (2014). http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/ [Accessed 21 December 14]

[5] Nurse, J.R.C., Erola, A., Goldsmith, M. and Creese, S., (2014) “Investigating the leakage of sensitive personal and organizational information in email headers”, in the Journal of Internet Services and Information Security, volume 5, number 1.

[6] Cunche, M., Kaafar, M. A. and Boreli, R., (2012) “I know who you will meet this evening! Linking wireless devices using Wi-Fi probe requests,” in IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2012). IEEE, pp. 1–9.

[7] Nurse, J.R.C., Pumphrey, J., Gibson-Robinson, T., Goldsmith, M. and Creese, S., (2014) “Inferring Social Relationships from Technology-Level Device Connections”, in proceedings of the 12th International Conference on Privacy, Security and Trust (PST 2014). IEEE. pp. 40–47.