Self-Portrayals of GI Junior Fellows

Dominik Herrmann*

Online privacy: Attacks and defenses

Abstract: I approach privacy issues on the Internet from two ends. On the one hand, I design and evaluate defensive measures, so-called privacy enhancing technologies (PETs), which can be used by individuals to protect themselves against surveillance on the Internet. On the other hand, I study the efficacy of offensive techniques. I am especially interested in passive surveillance techniques that cannot be detected. For instance, I have shown how machine learning techniques can be used to infer the contents of encrypted traffic and how to track users solely based on characteristic behavioral patterns.

Keywords: Privacy enhancing technologies, online tracking, linkability, machine learning, website fingerprinting, Domain Name System.

ACM CCS: Security and privacy → Security services → Pseudonymity, anonymity and untraceability, Networks → Network properties → Network privacy and anonymity

DOI 10.1515/itit-2015-0001
Received January 19, 2015; accepted January 28, 2015

1 Introduction

To what extent are we monitored on the Internet? How can we protect our privacy online? These are two of the questions that motivate my research. The importance of privacy is on the rise. We are trusting more and more online services with our personal data. Due to the increasing number of omnipresent smart devices and due to the overall complexity of technology, many citizens have lost track of their personal data, i. e., they do not know what information they have disclosed to whom.

More importantly, many of them are not even aware of the fact that there are so-called third parties on the Internet that are constantly monitoring their behavior without their consent. This kind of information asymmetry is characteristic for online privacy [20]. My research aims to improve this situation twofold: firstly, by laying the grounds for the development of more usable and thus more practical privacy tools that can be used as a means of self-defense, and secondly, by raising awareness, i. e., informing users, software developers and privacy regulators about methods and means used for online surveillance.

Before reviewing my work, I will state the general motivation for PETs in Section 2. Sections 3–6 introduce my research fields and present selected findings. I will conclude with an outlook in Section 7.

2 Motivation for PETs

On the Internet, communication is not encrypted by default. The traffic of a user can be monitored with little effort, either by the Internet service provider (ISP) or by operators of network routers that forward traffic towards its destination.

Messages can be encrypted before transmission, for instance with the TLS protocol used in HTTPS [5]. This protects *message contents* but not the *circumstances of the communication* such as timing, size, and the identity of the communication partners (i. e., their IP addresses). However, the mere circumstances may be sensitive as well: an ISP that observes one of its customers frequenting the homepage of a debt counseling service may conclude that this customer has financial problems.

Protecting the circumstances of communication *against the ISP* can be achieved by routing all traffic through an encrypted VPN tunnel, a service offered by providers such as Cyberghost or *anonymizer.com*. At first sight, a VPN seems to be a satisfactory solution. However, VPNs merely shift the privacy issue to the VPN provider, i. e., now users have to trust the VPN operators with their data.

This limitation is lifted by "real" *privacy enhancing technologies* [9] like Tor and AN.ON [2, 6], two anonymity services that provide *relationship anonymity on the network layer*. They obscure IP addresses of senders and receivers, which effectively hides who communicates with whom. Anonymity is achieved by encrypting messages multiple times and re-routing the traffic over two or more anonymizing nodes (so-called *onion routers* or *mixes*). The construc-

***Corresponding author: Dominik Herrmann,**
Universität Hamburg, Fachbereich Informatik,
e-mail: herrmann@informatik.uni-hamburg.de

tion of such anonymity services is based on Chaum's mix networks [3]. Properly implemented mix networks ensure that no single node can link senders and receivers.

3 Performance evaluation of anonymity services

While Chaum designed mix networks to protect asynchronous communication, AN.ON and Tor are specifically designed for low-latency applications, for instance, surfing on the World Wide Web. However, anonymity always comes at a cost: re-routing traffic over multiple nodes introduces additional delays (increasing latency), and the anonymity service may become a bottleneck when many concurrent users are connected (decreasing available bandwidth). According to [18], performance is one of the critical factors influencing usability of an anonymity service.

One pillar of my research is the evaluation of the performance of anonymity services under real-world conditions from a user's perspective. In 2007, I was among the first to perform systematic measurements using the live Tor and AN.ON networks. According to our results [22], Tor offered much more bandwidth at that time (good for downloading files) than AN.ON, while AN.ON introduced smaller delays than Tor (good for surfing on the web, which is characterized by many short-lived connections). A closer look revealed that the attainable performance depended considerably on the time of day (diurnal pattern). In the afternoon, when many users were connected, the systems performed much worse than in the morning. This finding could be explained for AN.ON, which was mainly used by users from European countries at that time. However, we also found strong diurnal patterns in the results for Tor, which is counterintuitive given Tor's global user group and its global routing.

Since then, there has been a growing interest in the evaluation of the performance of anonymity services. For instance, members of the Tor project have set up their own dedicated measurement infrastructure, which publishes data at http://metrics.torproject.org/.

4 PETs for DNS query privacy

Besides my aforementioned activities of evaluating existing designs, I have also studied how to design and construct novel PETs that balance privacy and performance.

Tor and AN.ON are *generic anonymity services* that provide an HTTP or SOCKS proxy interface on the client-side. In principle, these services are compatible with all applications that have clients connecting to servers via TCP connections. However, due to their distributed nature, the generic services are less suitable for applications that demand *very* low latencies. In my research, I have looked at one of those applications, namely the Domain Name System.

The Domain Name System (DNS) is the name resolution service on the Internet. It is used to translate domain names like *www.google.com* to IP addresses. Clients offload most of the work to so-called "DNS resolvers", dedicated servers that have mostly been provided by ISPs in the past. However, during the last years a "third-party ecosystem" for DNS services has evolved. Besides well-known offers like Google Public DNS and OpenDNS, there are many more public resolvers (cf. http://public-dns.tk). Unfortunately, switching to a public resolver inevitably discloses one's online activities to the DNS provider. This gives rise to privacy concerns: due to their central role, DNS resolvers are a preeminent entity for behavioral monitoring [10]. Nevertheless, there has been increasing demand for third-party DNS resolvers: Google's DNS resolvers have answered more than 70 billion queries per day in 2012 [4].

Relaying DNS queries via Tor is not practicable due to its prohibitively high latency: according to [7], it typically takes longer than 1 s until DNS queries are resolved via Tor. However, our "DNS mix" design demonstrates that mix networks can be *tailored* to the properties of DNS traffic to achieve sufficient performance [8]. Experiments with a cascade of three "DNS mixes" and the traffic of more than 2000 concurrent users showed promising results: 50% of the queries were answered within at most 171 ms. As the resulting latency mainly consists of the network latencies *between the mixes*, user-perceived latency can be further optimized by clever placement of the mixes.

An alternative to mix networks are *lightweight* PETs that give up some security features in order to achieve better performance. In this vein, we have proposed the EncDNS system [14], which repurposes the conventional DNS resolver (the one that is provided by the ISP) as a proxy server. In EncDNS, the original DNS queries are encrypted by the client and relayed via the conventional DNS resolver to an EncDNS server for name resolution. On the one hand, EncDNS provides query privacy by ensuring that none of the servers can learn the identity (IP address) of the user *as well as* the desired hostnames. On the other hand, the design minimizes additional latency (as low as 5 ms in our experiments), because it requires only one additional hop.

Anonymity, i. e., hiding one's identity, is only one way to achieve privacy. In my work, I have also studied approaches for DNS query privacy that achieve privacy by other means. One promising approach seems to be a broadcast service that pushes popular DNS records to all clients. This allows the clients to resolve the majority of their DNS queries locally, i. e., without querying a DNS resolver (ensuring unobservability). Such alternative approaches to privacy are a fruitful, yet mostly overlooked area of research.

5 Traffic analysis attacks

Apart from evaluating and designing PETs, I also study attacks on anonymity services in order to understand their limitations and their weak spots. In particular, I am interested in the field of *traffic analysis*, which is concerned with methods and techniques that can be used by an adversary to infer pieces of information based on the observed network traffic [19]. To this end, seemingly innocuous properties of the traffic, which are sometimes referred to as "metadata", are analyzed to uncover characteristic patterns.

In [16] I have evaluated the so-called *website fingerprinting* attack [17]. Website fingerprinting allows an observer on the network (for instance, the ISP or the provider of the first mix or Tor router) to infer which websites a user visits, even though the user sends all of her traffic through an encrypted tunnel. In theory, observers that see encrypted traffic only should not be able to gain any information about the contents. However, when a browser downloads a website, it issues a series of requests to retrieve all the embedded images. This pattern creates a characteristic traffic silhouette, which is not fully obscured by encryption. Website fingerprinting can thus be modeled as a pattern recognition problem or a classification task, i. e., a supervised learning problem. We have used a Multinomial Naive Bayes classifier to match the fingerprints of more than 770 popular websites. The results of our experiments indicate that website fingerprinting is quite effective against popular encrypted tunnels: more than 90% of the considered websites were detected correctly for OpenSSL, OpenVPN, and IPsec. Since then, others have improved upon our work and demonstrated that the Tor system may also be affected [21].

6 Tracking on the application layer

Tools like AN.ON and Tor provide privacy on the *network layer*, i. e., they hide the IP address of the users from the visited websites. However, many privacy issues on the Internet arise from handling data on the *application layer*.

In particular, I am interested in advanced tracking techniques that can be leveraged by third parties to monitor the behavior of users on the Internet without their consent. The data collectors apply statistical models and machine learning techniques to extract behavioral profiles from the browsing history of a user. From such profiles, sensitive pieces of personal information can be inferred, among them demographic properties such as gender and age as well as moral and political beliefs. Behavioral profiles are not only used to individualize the banner ads that are embedded within websites, but also for more questionable practices such as price discrimination.

Traditionally, data collectors have stored tracking cookies within the browser and relied on the fact that the IP address of most users remains unchanged for some time (many ISPs assign dynamic IP addresses with a lifespan of 24 hours). However, more and more users are now deleting cookies regularly, and some browsers have started to reject third-party tracking cookies altogether. Therefore, data collectors are on the lookout for new tracking opportunities, as demonstrated by the recent discovery of browser fingerprinting techniques in the wild [1].

In my research, I have analyzed the practicability of a novel tracking technique, which exploits behavioral characteristics of individuals to link multiple surfing sessions of the same user [13, 15]. Each session is modeled as a vector containing the usage intensity (number of requests) for all websites (hostnames) a user has visited within the respective session. In analogy to the website fingerprinting attack mentioned in Section 5, the task of linking sessions can be framed as a supervised learning problem. Our behavior-based tracking technique consists of a 1-Nearest Neighbor classifier that uses cosine similarity (the angle between two vectors) for comparison. Results obtained on a real-world dataset indicate that most users exhibit sufficiently regular and characteristic behavioral patterns: given the DNS queries of 12 000 users, more than 75% of the sessions were linked correctly over a course of two months [11].

What makes behavior-based tracking so worrisome is the fact that it can be carried out entirely passively, i. e., in contrast to cookies or browser fingerprinting, it cannot be detected on the client side at all. In order to prevent this form of tracking, users have to change their IP addresses (and delete cookies) very frequently – ideally every few

minutes, which is quite cumbersome at the moment, but will become easier with IPv6 [12].

7 Outlook

Traditional privacy paradigms are often perceived as obstacles that hinder progress – especially in light of technical advances like mobile networking, cloud computing, and big data analytics, all of which promise unprecedented opportunities. Many users choose to profit from novel features, rather than abstaining from them only for the sake of better privacy. While any impact on convenience is immediately perceptible, a potential loss of privacy is perceived as something very remote, uncertain, and intangible. On the other hand, those users that *do* care for privacy are caught up in a state of analysis paralysis, because they are faced with a multitude of recommendations regarding tools, techniques, and strategies.

Therefore, moving forward, a major challenge for online privacy research may not consist in coming up with more convenient tools that help users to protect privacy online, but in finding ways to take the burden of self-defense off their shoulders.

References

1. Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juárez, Arvind Narayanan, and Claudia Díaz. The web never forgets: Persistent tracking mechanisms in the wild. In *CCS 2014*, pages 674–689. ACM, 2014.
2. Oliver Berthold, Hannes Federrath, and Stefan Köpsell. Web MIXes: A System for Anonymous and Unobservable Internet Access. In *International Workshop on Design Issues in Anonymity and Unobservability*, volume 2009 of *LNCS*, pages 115–129. Springer, 2001.
3. David Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2), 1981.
4. Jeremy K. Chen. Google Public DNS: 70 billion requests a day and counting. Official Google Blog, http://googleblog.blogspot.de/2012/02/google-public-dns-70-billion-requests.html (visited on 6 Jan 2015), 2012.
5. Tim Dierks and Eric Rescorla. The Transport Layer Security (TLS) Protocol Version 1.2. RFC 5246, RFC Editor, August 2008.
6. Roger Dingledine, Nick Mathewson, and Paul F. Syverson. Tor: The Second-Generation Onion Router. In *13th USENIX Security Symposium*, pages 303–320. USENIX, 2004.
7. Benjamin Fabian, Florian Goertz, Steffen Kunz, Sebastian Müller, and Mathias Nitzsche. Privately Waiting — A Usability Analysis of the Tor Anonymity Network. In *AMCIS 2010, SIGeBIZ track*, volume 58 of *LNBIP*, pages 63–75. Springer, 2010.
8. Hannes Federrath, Karl-Peter Fuchs, Dominik Herrmann, and Christopher Piosecny. Privacy-Preserving DNS: Analysis of Broadcast, Range Queries and Mix-Based Protection Methods. In *ESORICS 2011*, volume 6879 of *LNCS*, pages 665–683. Springer, 2011.
9. Ian Goldberg, David Wagner, and Eric Brewer. Privacy-enhancing Technologies for the Internet. In *42nd IEEE Spring COMPCON*. IEEE, 1997.
10. Scott Goodson. If You're Not Paying For It, You Become The Product. Forbes.com, http://onforb.es/wVrU4G (visited on 6 Jan 2015), 2012.
11. Dominik Herrmann. *Beobachtungsmöglichkeiten im Domain Name System: Angriffe auf die Privatsphäre und Techniken zum Selbstdatenschutz*. PhD thesis, Universität Hamburg, 2014.
12. Dominik Herrmann, Christine Arndt, and Hannes Federrath. IPv6 Prefix Alteration: An Opportunity to Improve Online Privacy. In *1st Workshop on Privacy and Data Protection Technology, co-located with Amsterdam Privacy Conference*, 2012.
13. Dominik Herrmann, Christian Banse, and Hannes Federrath. Behavior-based Tracking: Exploiting Characteristic Patterns in DNS Traffic. *Computers & Security*, 39A:17–33, November 2013.
14. Dominik Herrmann, Karl-Peter Fuchs, Jens Lindemann, and Hannes Federrath. EncDNS: A Lightweight Privacy-Preserving Name Resolution Service. In *ESORICS 2014*, volume 8712 of *LNCS*, pages 37–55. Springer, 2014.
15. Dominik Herrmann, Christoph Gerber, Christian Banse, and Hannes Federrath. Analyzing Characteristic Host Access Patterns for Re-identification of Web User Sessions. In *NordSec 2010*, volume 7127 of *LNCS*, pages 136–154. Springer, 2012.
16. Dominik Herrmann, Rolf Wendolsky, and Hannes Federrath. Website Fingerprinting: Attacking Popular Privacy Enhancing Technologies with the Multinomial Naïve-Bayes Classifier. In *CCSW 2009*, pages 31–42. ACM, 2009.
17. Andrew Hintz. Fingerprinting Websites Using Traffic Analysis. In *PET Workshop 2002*, volume 2482 of *LNCS*, pages 171–178. Springer, 2003.
18. Stefan Köpsell. Low Latency Anonymous Communication - How Long Are Users Willing to Wait? In *ETRICS 2006*, volume 3995 of *LNCS*, pages 221–237. Springer, 2006.
19. Jean-François Raymond. Traffic analysis: Protocols, attacks, design issues, and open problems. In *International Workshop on Design Issues in Anonymity and Unobservability*, volume 2009 of *LNCS*, pages 10–29. Springer, 2001.
20. Tony Vila, Rachel Greenstadt, and David Molnar. Why we can't be bothered to read privacy policies: models of privacy economics as a lemons market. In *ICEC 2003*, pages 403–407. ACM, 2003.
21. Tao Wang and Ian Goldberg. Improved website fingerprinting on Tor. In *WPES 2013*, pages 201–212. ACM, 2013.
22. Rolf Wendolsky, Dominik Herrmann, and Hannes Federrath. Performance comparison of low-latency anonymisation services from a user perspective. In *PET Workshop 2007*, volume 4776 of *LNCS*, pages 233–253. Springer, 2007.

Bionotes

Dr. Dominik Herrmann
Universität Hamburg, Fachbereich Informatik, Vogt-Kölln-Str. 30, D-22527 Hamburg
herrmann@informatik.uni-hamburg.de

Dominik Herrmann is a post-doctoral researcher in the group "Sicherheit in verteilten Systemen" at Universiät Hamburg. Besides his research activities in the field of online privacy, he supports administrative tasks at various universities with tailored software tools. He also helps a number of schools with running their IT infrastructure. He was honored with a junior fellowship of Gesellschaft für Informatik in 2014.