Report from Dagstuhl Seminar 17162

Online Privacy and Web Transparency

Edited by

Nataliia Bielova[1], Nikolaos Laoutaris[2], Arvind Narayanan[3], and Nick Nikiforakis[4]

1 **Université Côte d'Azur, Inria, FR,** `nataliia.bielova@inria.fr`
2 **Telefónica Research – Barcelona, ES,** `laoutaris@gmail.com`
3 **Princeton University, US,** `arvindn@cs.princeton.edu`
4 **Stony Brook University, US,** `nick@cs.stonybrook.edu`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 17162 "Online Privacy and Web Transparency". The seminar brought 29 participants in computer science, law and policy together, coming from companies and research institutions across Europe and the US.

The 2.5-days seminar had a well-filled program, with 25 research talks, followed by 7 short panel discussions, and 6 5-minute talks. Online privacy and Web transparency is a broad research field, that includes detection of privacy leaks on the Web and mobiles, measurement of tracking technologies on the Web, transparency tools to detect bias and discrimination, as well as how laws and regulations address these problems from a law research perspective, and how technical solutions can influence standards and laws.

Seminar April 17–20, 2017 – http://www.dagstuhl.de/17162

1998 ACM Subject Classification D.4.6 Security and Protection (K.6.5), H.3.4 Systems and Software – User profiles and alert services, H.3.5 Online Information Services – Web-based services, K.4.1 Public Policy Issues – Ethics, Privacy, Regulation

Keywords and phrases Big data, Privacy Enhancing Technologies, Transparency, User Discrimination, Web Privacy

Digital Object Identifier 10.4230/DagRep.7.4.83

1 Executive Summary

Nataliia Bielova
Nikolaos Laoutaris
Arvind Narayanan
Nick Nikiforakis

License 🅭 Creative Commons BY 3.0 Unported license
© Nataliia Bielova, Nikolaos Laoutaris, Arvind Narayanan, and Nick Nikiforakis

The Dagstuhl Seminar on *Online Privacy and Web Transparency* was the first seminar at Dagstuhl that gathered together researchers working in web applications, online privacy, transparency on the web, privacy enhancing technologies, privacy measurement, and network economics, as well as several representatives of law and policy discipline.

Except where otherwise noted, content of this report is licensed
under a Creative Commons BY 3.0 Unported license

Online Privacy and Web Transparency, *Dagstuhl Reports*, Vol. 7, Issue 04, pp. 83–106
Editors: Nataliia Bielova, Nikolaos Laoutaris, Arvind Narayanan, and Nick Nikiforakis
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Research context

The web has become an essential part of our society and is currently the main medium of information delivery. Billions of users browse the web on a daily basis, and there are single websites that have reached over one billion user accounts. In this environment, the ability to track users and their online habits can be very lucrative for advertising companies, yet very intrusive for the privacy of users.

Recent research has shown that third-party advertising networks and data brokers use a wide range of techniques in order to track users across the web – these techniques are used to reconstruct browsing sessions and to create profiles of users, inferring, among others, their hobbies, health status, political inclinations, and level of wealth. This information can be used to, not only deliver better targeted advertisements to users, but also to discriminate users, for example by providing customized prices for products based on a user's willingness and ability to pay. To protect users, several solutions have been proposed, ranging from the laws, policies and the W3C Do-Not-Track candidate recommendation, to browser tools developed by companies and volunteers, and other client-side mechanisms proposed by researchers. At the same time, a number of tools have been developed to increase transparency on the web and allow end users to know when they are being tracked and when discrimination happens.

The seminar aimed to address the open questions of how to protect user privacy and how to increase transparency on the web. The key objectives of the seminar are (i) review the state of the art in the field; (ii) identify key technical challenges and brainstorm potential solutions; (iii) understand how computer science research results can influence law and policy; (iv) discuss ethical and legal issues in privacy research.

The seminar brought together scientists from the privacy and transparency communities, as well as policy makers interested in understanding how existing privacy laws and policies can be implemented, and representatives of Internet users organisations. The discussions at this Dagstuhl Seminar were strongly inspired by the following questions and challenges:

Technology

- How can we detect tracking and algorithmic discrimination most effectively? What are the scientific and engineering challenges to overcome? What are the relative merits of automated, semi-automated, and crowdsourced approaches?
- How can we ensure that methodologies, techniques, and tools are shared across different communities working on this topic?
- How can we design the next generation of privacy tools, get users to actively use the tools, and generate data for privacy researchers to scrutinize?
- What are the tracking techniques and data collection practices on mobile devices and how do they compare to those on the web?
- What are the privacy and transparency issues raised by the Internet of Things, and how do we address them?

Law

- Do the current laws and policies cover existing tracking technologies? What is the process for reporting newly discovered tracking techniques to the appropriate Data Protection Authorities?
- Even with the appropriate legislation in place, how can we ensure that companies comply with the law? What can researchers do to help enforce compliance?

Ethical issues

- What is the best way to conduct web privacy research taking ethical issues into account?
- When does a study necessitate ethical review?

Users

- What is the most efficient way to raise user awareness about web tracking and transparency tools?

2 Table of Contents

3 Seminar program

The seminar attracted 29 participants, in privacy and transparency, and law and policy together, coming from companies and research institutions across Europe and the US. The group presented a nice mix of participants from computer science community, industry (including researchers from AT&T, Telefónica Research and Cliqz), as well we researchers in law and policy (KU Leuven and Leiden University), and also representatives of data protection authorities (French CNIL and Federal Trade Commission).

The seminar had a well-filled program, with 25 research talks, followed by 7 short panel discussions, 3 break-out sessions and 6 5-minute talks. The organisers tried to keep enough time during breaks and in the evening for informal discussion, however because the seminar was only 2.5 days, we did not have the social event. To compensate for the possible lack of informal discussions, the last morning of the seminar was kept for the conclusion and free discussions.

Research talks

The organizers invited all the participants to take the floor during the seminar, and encouraged them to have rather short 15 minutes talks (differently from standard 25 minutes talks at conferences). This allowed to organise small panel discussions after each session where participants could engage in discussions.

Online privacy and web transparency is a new research field, where researchers from different backgrounds try to propose solutions to the problems. The organizers decided that each session will have a separate topic or open question, where researchers from different domains, as well as companies, can be mixed. As a result, different research results and open problems were presented in 5 working topics:

1. Advertisement and Real-Time Bidding
2. How technical solutions can influence standards/regulations?
3. Transparency and Web tracking
4. Bias and discrimination
5. Privacy by construction

For each of the sessions, the list of talks is presented in this section. For more detailed information about each talk, we refer to the talk abstracts in section 4.

Session 1: Advertisement and Real-Time Bidding

- Krishna Gummadi gave us insights on how to explain online advertisement in Facebook, trying to answer the questions such as what's the best way to explain online ads, what properties that tell us the explanations are good enough and how to automatise explanations?
- Nicolas Kourtellis illustrated how to bring more transparency into online ads by analysing how much advertisers pay to show ads.
- Robbert van Eijk discussed how to classify trackers and ad exchange companies and presented a demo showing how ad networks are connected to each other.
- Panel discussed the problem of utility from explanations (Netflix shows explanations to manipulate users to watch certain movies), what's the cost and benefit for the society from RTB, and how privacy leakage can be measured.

Session 2: How technical solutions can influence standards/regulations?

- Joe Calandrino presented the activities of FTC in investigation, education and research and in particular around topics related to web tracking.
- Lukasz Oleynik gave us insights on new browser APIs that allows to acquire user's behavior from sensors: he gave us history of battery status API leakage that was first revealed in 2011 and finally removed from Firefox in 2017.
- Damian Clifford explained us that the existing Web tracking technologies are covered by the current EU laws (GDPR), and require user consent.
- Timothy Libert presented his webXRay platform for online privacy measurements based on PhantomJS and motivated that chilling effects for OBA industry would be desirable.
- Panel was skeptic about enforceability of laws, due to many technical problems: how to distinguish functional cookies from tracking ones? How to detect other tracking methods that don't have any cookie at all? Participants recalled alternative methods, such as class actions and made analogy with environmental regulations.

Session 3: Transparency and Web tracking

- Günes Acar gave us insights on online crawlers and pointed that one can have different results when using a headless browser. He suspects that companies try to detect crawlers and probably try to evade transparency tools.
- Angel Cuevas presented his FDVT Facebook tool that approximately estimates the revenue users generate for Facebook based on the ads they receive and ads they click on.
- Costas Iordanou demonstrated two tools: $sheriff that detects price discrimination and eyeWnder that provides information to the user about web advertisements.
- Steven Englehardt presented OpenWPM, a platform for online privacy measurement and gave us insights on statefull and stateless techonogies detected on 1 million websites.
- Konrad Rieck discussed the new study on ultrasonic device tracking that is found in three commercial solutions.
- David Choffnes discussed longitudinal measurements of leaks in mobile apps and underlined that we need relations between the third parties for better measurements.
- Oleksii Starov gave us insights that browser extensions have more tracking powers than scripts and showed that some extensions leak browser and search history.
- Josep M. Pujol proposed to distinguish "parties that rely on tracking" from "trackers", because in some cases tracking is unintentional, an may be a by-product of a design choice.

Session 4: Bias and discrimination

- Michael Carl Tschantz discussed the problem of accountability for showing certain ads to certain groups of population. He questioned whether companies are accountable for discrimination of showing ads to a certain group or whether an advertiser should be accountable?
- Vincent Tobiana presented the first analysis on Facebook ethnic affinity and tried to evaluate whether FB users receive different ads based on their ethnic affinity.
- Aniko Hannak gave us insights on bias in job search (LinkedIn, Indeed, etc.) and professional communities (GitHub) and discussed whether online bias is different from offline, and started a discussion whether large-scale crawling may break terms of service of companies that are analysed.

- The panel engaged in discussion on offline vs. online behaviour, whether it should be copied, on discrimination/bias versus targeting. The panel tried to find the fundamental difference between "bias" and "discrimination".

Session 5: Privacy by construction

- Diana Vlad-Calcic discussed European policy overview and presented existing EU programs to support research on online privacy and web transparency.
- Benoit Baudry gave us insights about the problem of code diversity that leads to efficient fingerprinting and discussed soultions for browser randomisation.
- Francis Some presented a tool that automatically prevents third-party tracking on the server side and protects users from unintentional tracking by third-party content.
- Steven Englehardt explained that sandboxing tracking scripts breaks many sites, and therefore we need other methods to detect and protect from web browser fingerprinting.

Break-out sessions

After the first day of the seminar, the organisers proposed several topics for break-out sessions, that were voted by the participants. As a result of the vote, three break-out sessions were chosen – they enabled participants to discuss selected topics in privacy research in smaller teams (around 10 people per team). The three topics were:

- Security and privacy trade off
- Research feasibility and validity
- Bluesky proposals

The purpose of the break-out sessions was to informally discuss the most important problems in privacy and transparency, state research challenges, and legal problems. As part of the break-out sessions, the teams identified the most relevant problems in the field and main challenges for the specific Web privacy and transparency area. The break-out sessions lasted 90 minutes and were held in parallel, on Wednesday afternoon. Each participant joined a break-out sessions of her choice. The last 30 minutes session on Thursday was used to report back the results of the three break-out sessions to the full group by means of an informal discussion. The reports of the tree break-out sessions are summarized in section 5.

5-minute talks

To encourage participants to share their new ideas and results, we had one 5-minute session, where the following speakers presented their work:

- Update on the Data Transparency Lab by Nikolaos Laoutaris
- Browser Extension and Login-Leak Experiment by Nataliia Bielova
- Harvest documentary by Dave Choffnes
- Stealing browsing history using light sensors by Lukasz Olejnik
- Use of browser fingerprinting for security purposes by Gunes Acar
- How news media use Twitter to attract traffic? by Arnaud Legout

Conclusion

The seminar brought together 29 participants in computer science, law and policy, coming from companies and research institutions across Europe and the US. The seminar had a well-filled program, with 25 research talks, followed by 7 short panel discussions, and 6 5-minute talks. Online privacy and Web transparency is a broad research field, and hence a diverse set of recent research results were presented. They covered Web tracking technologies and transparency tools, behavioural advertisement, privacy protection mechanisms and technologies, bias and discrimination. The representatives from FTC and CNIL gave us insights on how to influence standards and regulations, while law and policy researchers explained how the current Web technologies are covered by the EU laws.

The seminar also featured three break-out sessions on Security and Privacy trade offs, Research feasibility and validity (in Web crawling and bias analysis), and Bluesy proposals. The goal of the break-out session was to discuss the most important open problems in the Web privacy and transparency research, and the special brainstorming session on bluesky proposals tried to propose completely new solutions and approaches to improve user's privacy on the Web – the summaries are documented in this report.

Finally, several new collaborations have been created as a result of this seminar, and at least one person has received a job offer due to the discussions that took place in Dagstuhl. A group of participants have organised a *Slack*[1] community to exchange news and ideas in the area.

4 Overview of Talks

4.1 How can we reconcile diversity and privacy?

Benoit Baudry (INRIA – Rennes, FR)

License 🅭 Creative Commons BY 3.0 Unported license
© Benoit Baudry

Main reference P. Laperdrix, W. Rudametkin, B. Baudry, "Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints", in Proc. of the IEEE Symposium on Security and Privacy (SP 2016), pp. 878–894, IEEE, 2016.
URL https://doi.org/10.1109/SP.2016.57

With this talk, I wish to trigger a discussion about possible ways to reconcile two values, which are fundamentally good, and yet, seem hardly compatible: diversity and privacy.

Diversity is good. Diversity is an essential property of natural systems, it is a characteristic sought in most human organizations and it is a moral value sherished by mankind. Diversity is key for the robustness of complex systems, it is essential to prevent monocultures and the risk of single points of failures. Diversity is also extremely beneficial in software systems for security and safety.

However, diversity threatens online privacy. Meanwhile, individualization is the counterpart of diversity. Consequently, multiple forms of diversity have become threats for the privacy of web users. For example, browser fingerprinting has emerged from the massive diversity of software and hardware components that users can assemble to set their environment. The diversity of online behaviors can be tracked, analyzed and learned to create filter bubbles.

[1] https://onlineprivacy.slack.com

Here, I will present our observations about the massive diversity of browser fingerprints. Then, I will discuss some solutions we have investigated to reconcole this with privacy, as well as some of the limitations we encountered.

4.2 Browser Extension and Login-Leak Experiment

Nataliia Bielova (INRIA Sophia Antipolis, FR), Claude Castelluccia, and Gábor Gy. Gulyás

License 🅭 Creative Commons BY 3.0 Unported license
© Nataliia Bielova, Claude Castelluccia, and Gábor Gy. Gulyás
URL http://extensions.inria.fr

When a user browses the web, various trackers are spying on her online activities. Even though such trackers are invisible, they collect information about her, such as which pages she visits, which buttons clicks, and what text she types. This information is often used to show her targeted advertisements and may require her to pay a higher price during online shopping depending on the collected information.

Recent studies [6, 12, 5, 7, 3, 4, 1, 13, 10, 2] show that users can be identified based on their device characteristics: this tracking method is called *device fingerprinting*. Such unique collection of device's properties, or a fingerprint, can often uniquely identify the user who visited the website. Usually, the fingerprint includes technical parameters like what browser and operating system a visitor is using, what timezone she is from, what fonts she has in her system, or what audio card her device supports. Beyond pure technical characteristics, which are not explicitly chosen by the user, a visitor can be also identified by more *behavioral* characteristics, such as the browser extensions she has installed and the websites where she has logged in. Detecting extensions and website logins can clearly make a significant contribution to fingerprinting.

In our new experiment at http://extensions.inria.fr, we demonstrate how websites can use *behavioral fingerprinting* and detect two aspects of user's online behavior: web browser extensions and websites a user has logged in[2]. Using a detection method based on Web Accessible Resources, [11], we are able to detect more than 13,000 Chrome browser extensions, including AdBlock, Pinterest, and Ghostery. Our experiment demonstrates an important privacy concern: the more privacy extensions you install, the more identifiable you are!

References

1 E. Abgrall, Y. L. Traon, M. Monperrus, S. Gombault, M. Heiderich, and A. Ribault. XSS-FP: browser fingerprinting using HTML parser quirks. *CoRR*, abs/1211.4812, 2012.

2 G. Acar, C. Eubank, S. Englehardt, M. Juárez, A. Narayanan, and C. Díaz. The web never forgets: Persistent tracking mechanisms in the wild. In G. Ahn, M. Yung, and N. Li, editors, *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 674–689. ACM, 2014.

3 G. Acar, M. Juárez, N. Nikiforakis, C. Díaz, S. F. Gürses, F. Piessens, and B. Preneel. Fpdetective: dusting the web for fingerprinters. In A. Sadeghi, V. D. Gligor, and M. Yung, editors, *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*, pages 1129–1140. ACM, 2013.

[2] In the experiment, we collect user's browser fingerprint, together with the browser extensions installed and a list of websites the user has logged in. We only collect anonymous data during the experiment (see our Privacy Policy at https://extensions.inrialpes.fr/privacy.php). We securely store the data on an Inria server, use it only for research purpose and not share it with anyone outside of Inria.

4 K. Boda, Á. M. Földes, G. G. Gulyás, and S. Imre. User tracking on the web via cross-browser fingerprinting. In P. Laud, editor, *Information Security Technology for Applications – 16th Nordic Conference on Secure IT Systems, NordSec 2011, Tallinn, Estonia, October 26-28, 2011, Revised Selected Papers*, volume 7161 of *Lecture Notes in Computer Science*, pages 31–46. Springer, 2011.

5 Y. Cao, S. Li, and E. Wijmans. (cross-)browser fingerprinting via os and hardware level features. In *24th Annual Network and Distributed System Security Symposium, NDSS 2017, San Diego, California, USA, 26 February – 1 March, 2017*, 2017. To Appear.

6 P. Eckersley. How unique is your web browser? In M. J. Atallah and N. J. Hopper, editors, *Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings*, volume 6205 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2010.

7 S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. In E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1388–1401. ACM, 2016.

8 E. Homakov. Using content-security-policy for evil. http://homakov.blogspot.fr/2014/01/using-content-security-policy-for-evil.html, 2014.

9 R. Linus. Your social media fingerprint. https://robinlinus.github.io/socialmedia-leak/, 2016.

10 N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*, pages 541–555. IEEE Computer Society, 2013.

11 A. Sjösten, S. Van Acker, and A. Sabelfeld. Discovering browser extensions via web accessible resources. In *Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy*, CODASPY '17, pages 329–336, New York, NY, USA, 2017. ACM.

12 O. Starov and N. Nikiforakis. Extended tracking powers: Measuring the privacy diffusion enabled by browser extensions. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3 – 7, 2017*, 2017. To Appear.

13 N. Takei, T. Saito, K. Takasu, and T. Yamada. Web browser fingerprinting using only cascading style sheets. In L. Barolli, F. Xhafa, M. R. Ogiela, and L. Ogiela, editors, *10th International Conference on Broadband and Wireless Computing, Communication and Applications, BWCCA 2015, Krakow, Poland, November 4-6, 2015*, pages 57–63. IEEE Computer Society, 2015.

4.3 A Brief History of Mobile Privacy Leaks

David Choffnes (Northeastern University – Boston, US)

License 🅭 Creative Commons BY 3.0 Unported license
© David Choffnes

Joint work of Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, Christophe Leung, Christo Wilson
Main reference J. Ren, A. Rao, M. Lindorfer, A. Legout, D. R. Choffnes, "ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic", in Proc. of the 14th Annual Int'l Conf. on Mobile Systems, Applications, and Services (MobiSys 2016), pp. 361–374, ACM, 2016.
URL http://doi.acm.org/10.1145/2906388.2906392

Mobile devices have unparalleled access to our daily lives, but give us little access to how they gather and share such information over time. In this talk, I summarized some research [1, 2] my group has been doing to address this problem, using a project we call ReCon. I will cover how we identify personally identifiable information exposed by mobile devices to other parties over the Internet, some of the interesting findings from analyzing hundreds of apps and users, how the nature of data collection is changing over time, and what are the implications for users, policymakers, and regulators. As part of ongoing work, we have been analyzing how data collection from apps changes over time, and what are the corresponding privacy and security implications for users.

References

1 Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, David Choffnes. *ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic.* In Proceedings of MobiSys (MobiSys '16), Singapore, June 2016.

2 Christophe Leung, Jingjing Ren, David Choffnes, and Christo Wilson. *Should You Use the App for That? Comparing the Privacy Implications of Web- and App-based Online Services.* In Proceedings of the 16th ACM Internet Measurement Conference (IMC'16), Santa Monica, CA, November 2016.

4.4 FDVT: Facebook Data Valuation Tool

Ángel Cuevas Rumin (Univ. Carlos III – Madrid, ES), Rubén Cuevas Rumin, and José González Cabañas

License 🅭 Creative Commons BY 3.0 Unported license
© Ángel Cuevas Rumin, Rubén Cuevas Rumin, and Jos'e González Cabañas

Main reference J. González Cabañas, Á. Cuevas, R. Cuevas, "FDVT: Data Valuation Tool for Facebook Users", in Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems (CHI'17), pp. 3799–3809, ACM, 2017.
URL https://doi.org/10.1145/3025453.3025903

The OECD, the European Union and other public and private initiatives are claiming for the necessity of tools that create awareness among Internet users about the monetary value associated to the commercial exploitation of their online personal information. In this talk we present a recent developed tool addressing this challenge, the Data Valuation Tool for Facebook users (FDVT). The FDVT provides Facebook users with a personalised and real-time estimation of the revenue they generate for Facebook based on the ads they receive and the ads they click on while browsing in this social network. The FDVT has been implemented as a web browser extension available for Google Chrome and Firefox through fdvt.org. Currently, more than 5000 users have installed the FDVT.

4.5 Analyzing the Impact of Large Scale Online Tracking Measurement

Steven Englehardt (Princeton University, US) and Arvind Narayanan (Princeton University, US)

License Creative Commons BY 3.0 Unported license
© Steven Englehardt and Arvind Narayanan
Main reference S. Englehardt, A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis", in Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS 2016), pp. 1388–1401, ACM, 2016.
URL https://doi.org/10.1145/2976749.2978313

In this talk I'll summarize the findings of a 1-million-site measurement of online tracking. I'll share results of stateful (cookie-based) and stateless (fingerprinting-based) tracking measurements. I'll examine the impact our measurements have had on the adoption of new techniques by trackers and implementation of defenses by browsers. By extracting lessons from this research and the ensuing impact, I'll propose several new directions for the tracking measurement field.

4.6 Sandboxing Trackers with Resource Blocking Lists

Steven Englehardt (Princeton University, US)

License Creative Commons BY 3.0 Unported license
© Steven Englehardt
Joint work of Steven Englehardt, Tanvi Vyas, Eric Rescorla
URL https://bugzilla.mozilla.org/show_bug.cgi?id=1298207

The most effective consumer privacy tools available today work by blocking advertising and tracking resources from loading in a user's browser. Resource blocking frequently leads to lost revenue and site breakage. We aim to design a tracking protection feature that minimizes breakage and preserves the revenue stream provided by advertisements while limiting the tracking capabilities of third-party content. We explore several possible client-side solutions and evaluate their effectiveness at preventing tracking and minimizing site breakage, and in their engineering feasibility.

We find that sandboxing tracking resources is either ineffective or infeasible depending on the configuration. Sandbox configurations which don't provide Javascript sandboxing either continue to block a large percentage of resources or fail to significantly impact the level of tracking. Sandboxing javascript has the potential to both reduce the level of tracking and amount of breakage, but requires a design that's heavily coupled to current implementation of the tracking scripts. Our work highlights several hurdles to user privacy caused by the "mash-up" nature of the modern web and raises questions to the extent that purely client-side solutions can protect users.

4.7 The Diffix Framework: Noise Revisited, Again

Paul Francis (MPI-SWS – Kaiserslautern, DE)

License Creative Commons BY 3.0 Unported license
© Paul Francis

A longstanding problem is that of obtaining high-quality statistical data from a dataset about individuals while protecting the privacy of those individuals. We believe that we have broken new ground on this problem. Diffix is an approach to database anonymization that: has minimal distortion (noise with standard deviation of only 2 for counting queries); places no limit on the number of queries; has rich query semantics (most of SQL, many statistical operations); is easy to configure; can comfortably be called anonymous according to European privacy law. In this talk, I will give a short demo of a commercial-quality implementation of Diffix, and give a brief overview of the main concepts.

4.8 Explaining Online Ads

Krishna P. Gummadi (MPI-SWS – Saarbrücken, DE)

License Creative Commons BY 3.0 Unported license
© Krishna P. Gummadi

Abstract: Online service providers like Netflix and Facebook are beginning to offer automated explanations to their consumers, e.g., Facebook's "why am I seeing this ad?" feature. While these explanations have the potential to make the inner working of the services transparent to consumers, they raise several questions. Specifically, (i) What are the types of explanations that are being provided by these services today? (ii) What properties or standards should explanations offered to consumers satisfy (in order to be useful and / or meaningful)? (iii) How can we (automatically) construct explanations that meet specific standards? In an ongoing effort, we are attempting to answer these questions in the context of Facebook targeted advertisements. In this presentation, I will discuss our preliminary findings.

4.9 New Faces Of Bias in Online Labor Markets

Aniko Hannak (Central European University – Budapest, HU)

License Creative Commons BY 3.0 Unported license
© Aniko Hannak

The internet is fundamentally changing the labor economy. Millions of people use sites like LinkedIn, Upwork or Dribbble to find employment. These services are often driven by algorithms that rate, sort, recommend, and match workers and employers. In theory, many of the mechanisms that cause discrimination in traditional labor markets – cognitive bias, network homophily, statistical discrimination – should be absent from online markets. However, recent studies indicate that these mechanisms do transfer to online platforms, where they may be exacerbated by seemingly harmless design choices.

In this talk I will investigate three techniques that online platforms use to match users with content: social network algorithms, search algorithms and public review systems. Specifically, I present case studies of 6 different employment platforms, using large scale user data from the employers perspective. I show that biases known from traditional labor markets are indeed present in online platforms, although they manifest in new ways. First, I present results that focus on the visibility of users, which directly impacts the chances of being selected for a job or selling a product. I find that women often receive lower visibility either due to their ranking in the sites' search interface, or their positions in the underlying social network. Furthermore, I investigate social feedback and other success measures found on user profiles, another important factor in hiring decisions. Overall, my investigations show that demographic features are often correlated with the attention and the social feedback workers and employees receive. Exploring these new forms of inequalities, understanding where social biases enter systems and which mechanisms reinforce them, can be crucial for developing mitigation strategies.

4.10 Web transparency tools demo

Costas Iordanou (Telefónica Research – Barcelona, ES)

License 🅭 Creative Commons BY 3.0 Unported license
© Costas Iordanou

The first tool is related to online price discrimination called $heriff. The tools allows internet users to check the prices of different products and services available on the web and look for evidence of price discrimination. The tool is available for two web browsers. Google chrome version is available at https://chrome.google.com/webstore/detail/heriffv2/emobhicogmenmngifjhjbfliohhjijjl and Mozilla Firefox version is available at https://addons.mozilla.org/en-US/firefox/addon/sheriff_v2/.

The second tool is related to web advertisements called eyeWnder. The tool is following the crowdsourced approach and provides information to the user regarding web advertisements in real time. The tool is also able to visualise the web browsing history of the user annotated with the interest categories assigned by advertisers. The tool is available at http://www.eyewnder.com/.

4.11 Transparency in the era of programmatic, real-time bidding advertising and Cookie Synchronization

Nicolas Kourtellis (Telefónica Research – Barcelona, ES)

License 🅭 Creative Commons BY 3.0 Unported license
© Nicolas Kourtellis

Joint work of Panagiotis Papadopoulos, Pablo Rodriguez, Nikolaos Laoutaris, Evangelos Markatos
Main reference P. Papadopoulos, N. Kourtellis, P. Rodriguez, N. Laoutaris, "If you are not paying for it, you are the product: How much do advertisers pay to reach you?", arXiv:1701.07058 [cs.GT], 2017.
URL https://arxiv.org/pdf/1701.07058.pdf

This presentation was divided into two parts. The first part proposed a methodology on evaluating users' data from the advertising ecosystem side using RTB ads. The second part presented a longitudinal study on the cookie synchronization process used by advertisers and trackers to share data of users.

Part 1: Online advertising is progressively moving towards a programmatic model in which ads are matched to actual interests of individuals collected as they browse the web. Letting the huge debate around privacy aside, a very important question in this area, for which little is known, is: How much do advertisers pay for an individual's personal data?

In this study[1], we develop a first of its kind methodology for computing exactly that – the price paid for one's privacy – and we do that in real time. Our approach is based on tapping on the Real Time Bidding (RTB) protocol to collect cleartext and encrypted prices for winning bids paid by advertisers in order to place targeted ads. Our main technical contribution is a method for tallying winning bids even when they are encrypted. We achieve this by training a model using as ground truth prices obtained by running our own "probe" ad-campaigns. We implement our methodology through a browser addon and a back-end server that provides it with fresh models for encrypted bids. We validate our methodology using a one year long trace of 1600 mobile users and demonstrate that it can estimate a user's advertising worth with more than 82% accuracy.

Part 2: Cookies are still the dominant user targeting mechanism on the web. Third-party online companies maintain large user data stores for all unique users encountered and identified using anonymous cookies. In order to identify and track users across different publishers and through time, third parties connect with each other and synchronise their cookies. Consequently, Cookie Synchronization (CSync) is one of the de facto tracking mechanisms of modern web. CSync facilitates an information sharing channel between third parties that may or may not have direct access to the website the user visits. With CSync they can not only reconstruct the browsing history of a user by bypassing the same origin policy, but also merge the user data they own, in the background. In this paper [2], we perform a first to our knowledge longitudinal study of CSync in the wild, using a year-long dataset that includes browsing activity from 1600 real users. Through our study, we aim to understand the protocol's characteristics, growth and the flow graph of user personal information while being leaked to third parties. Our results show that 97% of the regular web users are exposed to Cookie Synchronization: most of them within the first week of their browsing. Our experiments also suggest that the average user is exposed to 63 distinct cookie synchronization events. This implies that despite the fact that all 63 of the trackers thought that they were tracking 63 different users (identities), after cookie synchronization, all 63 identities can point back to the same single user.

References

1 Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, Nikolaos Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay for your personal data? https://arxiv.org/pdf/1701.07058.pdf

2 Panagiotis Papadopoulos, Nicolas Kourtellis, Evangelos Markatos. Syncing the cookie monster: A longitudinal study of cookie syncing and its impact on user privacy.

4.12 How news media use Twitter to attract traffic?

Arnaud Legout (INRIA Sophia Antipolis, FR)

License 🅭 Creative Commons BY 3.0 Unported license
© Arnaud Legout

Joint work of Maksym Gabielkov, Arthi Ramachandran, Arnaud Legout, Augustin Chaintreau
Main reference M. Gabielkov, A. Ramachandran, A. Legout, A. Chaintreau, "Social Clicks: What and Who Gets Read on Twitter?", in Proc. of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science, pp. 179–192, ACM, 2016.
URL https://doi.org/10.1145/2964791.2901462

Online news domains increasingly rely on social media to drive traffic to their website. Yet we know surprisingly little about how social media conversation mentioning an online article actually generates a click to it. Posting behaviors, in contrast, have been fully or partially available and scrutinized over the years. While this has led to to multiple assumptions on the diffusion of information, each were designed or validated while ignoring this important step.

We present a large scale, validated and reproducible study of social clicks – that is also the first data of its kind – gathering a month of web visits to online resources that are located in 5 leading news domains and that are mentioned in the third largest social media by web referral (Twitter). Our dataset amounts to 2.8 million posts, together responsible for 75 billion potential views on this social media, and 9.6 million actual clicks to 59,088 unique resources. We design a reproducible methodology, carefully corrected its biases, enabling data sharing, future collection and validation. As we prove, properties of clicks and social media Click-Through-Rates (CTR) impact multiple aspects of information diffusion, all previously unknown. Secondary resources, that are not promoted through headlines and are responsible for the long tail of content popularity, generate more clicks both in absolute and relative terms. Social media attention is actually long-lived, in contrast with temporal evolution estimated from posts or impressions. The actual influence of an intermediary or a resource is poorly predicted by their posting behavior, but we show how that prediction can be made more precise.

4.13 Surveillance as a Regulatory Model

Timothy Libert (University of Pennsylvania – Philadelphia, US)

License Ⓒ Creative Commons BY 3.0 Unported license
© Timothy Libert

Bruce Schneier has noted that "surveillance is the business model of the internet". One must therefore ask, what is the regulatory model of the Internet? Currently, the dominant regulatory stances in regards to online tracking rely on industry self-regulatory guidelines which are largely focused on the paradigm of "notice and choice". This paradigm is a vastly weakened version of the normatively-grounded Fair Information Practice Principles, is not practiced by industry in any meaningful way, and fails to protect user privacy in sensitive areas such as personal health. To the degree to which industry is regulated, it is often done through a mixture of fines, bad publicity, and user attempts at blocking tracking mechanisms. These scatter-shot approaches have failed to place significant limits on the spread of online tracking.

This talk proposes a new possibility: surveillance as a regulatory model. With many researchers now conducting large-scale censuses of web tracking practices, it is possible to constantly monitor the activities of companies tracking users on the web and provide regulators to both daily and historical reports on the state of tracking. If and when companies engage in deceptive practices it should be possible to spot such practices quickly, apply fines based on the nature of the deception, and multiply fines based on the number of sites affected over time.

While marquee names such as Facebook and Google receive the most media and regulatory attention, such a large-scale approach could be applied to a much larger variety of companies and result in a significantly expanded pool of companies who would be under constant scrutiny. This would facilitate moving from the current model of large fines levied infrequently, to smaller-scale fines levied on a regular basis.

In short, if the business model of the internet is surveillance, the regulatory model for online privacy should follow suit. The purpose of this talk is to be fairly short, and provoke discussion towards the technical requirements of such an approach, the receptiveness of regulators, and overall feasibility.

4.14 Modern Web Privacy: standards, implementations, deployments

Lukasz Olejnik (University College London, GB)

License 🅭 Creative Commons BY 3.0 Unported license
© Lukasz Olejnik

Joint work of Lukasz Olejnik, Steven Englehardt, Arvind Narayanan
Main reference Battery Status Not Included: Assessing Privacy in Web Standards, International Privacy Engineering Workshop, 2017
URL https://blog.lukaszolejnik.com/battery-status-not-included-assessing-privacy-in-w3c-web-standards/

For majority of users, web browser is the most important computer application. Increasingly complex, exciting and rich, features are standardized by W3C and implemented in web browsers on a normal basis. New browser features introduce interesting privacy challenges for standardization, research and development.

The importance of privacy engineering has become increasingly apparent to standardizers (W3C), implementers (browser vendors) and web developers. Standardized guidelines for privacy assessments exist and are continuously improved in response to research and experience. The ever increasing complexity and richness of the web ecosystem necessitates a continual reevaluation of privacy assessments.

I provide a case study analysis of the evolution of the W3C Battery Status API, and discuss the specification and related implementations through the previous disclosure of several privacy vulnerabilities. We examine how implementations change and adoption shifts in response to these vulnerabilities. To provide context we present new measurement results and usage statistics for the Battery Status API, showing that there is a heavy fingerprinting use on the modern web. I will mention a list of methodologies, design patterns and recommendations to improve the privacy engineering process during the drafting of specifications and preparing of implementations.

I point out the strong features such as new communication channels or access to low-level sensors that the Modern Web is starting to offer and discuss a number of possible consequences to privacy. Are we ready for the web as a part of the Internet of Things and the ensuing challenges? I discuss how to measure privacy in the new web paradigms.

4.15 Ultrasonic Device Tracking for Fun and Profit

Konrad Rieck (TU Braunschweig, DE)

License 🅭 Creative Commons BY 3.0 Unported license
© Konrad Rieck

Joint work of Christian Wressnegger, Erwin Quiring, Daniel Arp, Konrad Rieck
Main reference D. Arp, E. Quiring, C. Wressnegger, K. Rieck, "Privacy Threats through Ultrasonic Side Channels on Mobile Devices", in Proc. of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P), pp. 35–47, IEEE, 2017.
URL https://doi.org/10.1109/EuroSP.2017.33

Device tracking is a serious threat to the privacy of users. Recently, several companies have started to use ultrasound for tracking mobile devices. To this end, ultrasonic markers are embedded in an audio signal and unnoticeably tracked using the microphone of mobile devices. This side channel allows an adversary to identify a user's current location, spy on her TV viewing habits or link together her different mobile devices. In this talk, we explore the capabilities and the current prevalence of this new tracking technique based on three commercial solutions. We discuss detection and mitigation approaches, and present case studies on Web and TV media.

4.16 Control What You Include ! Server-Side Protection against Third Party Web Tracking

Dolière Francis Some (Université Côte d'Azur, Inria, FR), Nataliia Bielova (Université Côte d'Azur, Inria, FR), and Tamara Rezk

License Ⓒ Creative Commons BY 3.0 Unported license
© Dolière Francis Some, Nataliia Bielova, and Tamara Rezk

Third party tracking is the practice by which third parties recognize users accross different websites as they browse the web. Recent studies show that more than 90% of Alexa top 500 websites[1] contain third party content that is tracking its users across the web. Website developers often need to include third party content in order to provide basic functionality. However, when a developer includes a third party content, she cannot know whether the third party contains tracking mechanisms. If a website developer wants to protect her users from being tracked, the only solution is to exclude any third-party content, thus trading functionality for privacy. We describe and implement a privacy-preserving web architecture that gives website developers a control over third party tracking: developers are able to include functionally useful third party content, the same time ensuring that the end users are not tracked by the third parties.

References

1 Franziska Roesner and Tadayoshi Kohno and David Wetherall. Detecting and Defending Against Third-Party Tracking on the Web. Proc. of the 9th NSDI. pages 155-168, 2012.

4.17 Understanding the privacy risks of browser extensions

Oleksii Starov (Stony Brook University, US) and Nick Nikiforakis (Stony Brook University, US)

License Ⓒ Creative Commons BY 3.0 Unported license
© Oleksii Starov and Nick Nikiforakis

Main reference O. Starov, N. Nikiforakis, "Extended Tracking Powers: Measuring the Privacy Diffusion Enabled by Browser Extensions," in Proceedings of the 26th International Conference on World Wide Web (WWW 2017), pp. 1481–1490, ACM, 2017.
URL https://doi.org/10.1145/3038912.3052596
Main reference O. Starov, N. Nikiforakis, "XHOUND: Quantifying the Fingerprintability of Browser Extensions", in Proc. of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P), pp. 941–956, IEEE, 2017.
URL https://doi.org/10.1109/SP.2017.18

The increased popularity of browser extensions can have serious negative consequences to a user's online privacy. In this talk, we describe our findings regarding extension privacy from two recent studies. First, we will describe the threat of browser extensions leaking a user's browser history, on purpose or accidentally, to third parties. In our recent work (published at WWW 2017), we built a dynamic analysis system and discovered that above 6% of browser extensions leak sensitive information to one or more third parties. Second, we discuss the issue of extension fingerprinting. In our recent paper at IEEE S&P 2017, we showed that it is possible to fingerprint browser extensions automatically and at scale. Specifically, we present the details of our dynamic analysis system (XHound) which stimulates browser extensions in order to identify their on-page DOM side-effects which can be abused by a web page to infer an extension's presence or absence. Our tool was able to automatically extract fingerprintable vectors from 9.2% of the top 10K Chrome extensions that run on any page, and 16.6% of extensions that run on popular domains like Google, YouTube, and Facebook. We explain why the threat of extension fingerprinting is more serious than traditional fingerprinting and discuss possible solutions, both for data-leaking extensions as well as for protecting users against extension fingerprinting.

4.18 Accountability for Privacy and Discrimination

Michael Carl Tschantz (ICSI – Berkeley, US)

License Creative Commons BY 3.0 Unported license
© Michael Carl Tschantz

My prior work, Datta et al. [1], presented AdFisher, an automated tool that explores how user behaviors, Google's ads, and its Ad Settings interact. AdFisher found that setting a simulated user's gender to female resulted it in getting fewer instances of an ad related to high paying jobs than setting the gender to male. It also found that visiting webpages associated with substance abuse changed the ads shown but not the settings page. However, we cannot determine who or what caused these findings due to our limited visibility into the ad ecosystem, which includes interactions between Google, advertisers, websites, and users.

I believe Google owes us an account of why such discrimination occurred and that such accounts form the basis of "accountability", as opposed to responsibility or punishment. However, providing such accountability has its difficulties.

One difficulty is interpreting vague policies that refer to what information is "about" or the "subject" of data. As big data analytics find increasingly unexpected associations between unexpected features, it is becoming increasingly difficult to delimit data along these lines. Motivated by this difficulty, we are exploring ways of identifying data for protection based upon where the data comes from instead of what it is about.

Another difficulty is the need to account for responsibility when multiple actors interact to produce a result. Datta et al. [2] developed a theory of causal responsibility, Quantitative Input Influence, that assigns responsibility to the participants in a blackbox system based upon cooperative game theory. I will discuss the difficulties, both practical and conceptional, with applying this theory to large systems.

References

1 Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. Proceedings on Privacy Enhancing Technologies (PoPETs), 2015.

2 Anupam Datta, Shayak Sen, Yair Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. Proceedings of 37th IEEE Symposium on Security and Privacy, 2016.

4.19 Tracking Classification, a cluster analysis approach to identify missing values

Robbert van Eijk (Leiden University Dual PhD Centre, NL)

License 🅭 Creative Commons BY 3.0 Unported license
© Robbert van Eijk

Main reference R. van Eijk, "Tracking Classification: A Cluster Analysis Approach to Identify Missing Values" (Presentations Slides).
URL http://dx.doi.org/10.2139/ssrn.2954946
Main reference R. van Eijk, "A Brief Introduction to Real Time Bidding (RTB)"
URL https://ssrn.com/abstract=2856941
Main reference R. van Eijk, "Web Tracking Detection System (TDS): An Effective Strategy to Reduce Systematic Monitoring and Profiling of User Habits Across Websites"
URL https://ssrn.com/abstract=2696690

Presentation part of Dagstuhl Seminar 'Online Privacy and Web Transparency'. This talk is motivated by the changing legal framework in the European Union, i.e., the General Data Protection Regulation and the (draft) ePrivayc Regulation. Transparency and fairness are key elements when it comes to informed consent. The problem with real-time bidding however is that many third-parties are unknown at the time of asking consent from a user. The talk will explore a methodology for classification of third-parties by looking at information flows between real-time bidding networks using cluster analysis. Interactive HTML-widgets will be used to guide the exploration. The difference in effectiveness of ad-blockers in the EU versus the US will also be discussed.

The presentation leads to two conclusions against the background of regulatory changes in the EU (GDPR and ePrivacy Regulation). Conclusion on technology: with cluster analysis on a standard referer graph it is possible to differentiate between actors and their role in interconnected RTB systems. The following cluster analysis approaches have proven to be useful: (a) node betweeness (b) cluster-edge betweenness (c) eigenvector centrality. Conclusion on policy: cookie enforcement can be effective as the example in Slovenia shows.

Keywords: web tracking, big data, privacy, data protection

5 Break-out sessions

5.1 Security and privacy trade off

This break-out session focused on a stateless web tracking technique called browser fingerprinting. The main question on the trade off between security and privacy was: fingerprinting is used to track users, but in the same time is it very useful for security – for example, an attacker who stole user's credentials would not be able to access the system since his fingerprint is different than those of the legitimate user.

Law and policy researchers pointed out that the **usage of fingerprinting can also be bounded by law**: it may be used only under the conditions of necessity and proportionality. Moreover, with upcoming EU laws, such as GDPR and ePrivacy, any non-functional tracking will require user consent, while functional tracking (like the one for security) will not require consent. The group has also concluded that personalised ads are not considered necessary for the provision of the service.

From a technical perspective, the group discussed and compared using **fingerprinting as a security measure** versus using security questions. By using security questions, an attacker who knows enough information about the user could break in. Some researchers concluded that negative authentication may be a good option: instead of whitelisting "good" (known) fingerprints, the system may blacklist "bad" fingerprints thus blocking bot fingerprinting.

Researchers have agreed that it's **difficult to detect fingerprinting** since there are no "bad" APIs. The state-of-the-art literature doesn't propose 100% accurate detection of fingerprinting. Fingerprinting is also used in malware operations when an attacker finds a vulnerable target because of the browser's fingerprint. Therefore proposing solutions to detect and/or prevent fingerprinting is an important goal not only for privacy reasons, but also for security of web applications.

It was pointed out that the usage of **fingerprinting for authentication** may be profitable, but in this case two fingerprints must be collected (one to use, another to store) to beat replay attacks. It can be considered as "functional tracking" if it is ensured that fingerprinting is always used together with authentication. However, if fingerprinting is used on other landing pages, then it's not functional.

Finally, researchers have raised a question: **Is fingerprinting a lost cause?** Recent communications with Mozilla underline that the company will not expand the APIs of the FireFox browser without purpose, however they do not believe that they can have an unfingerprintable browser. Some preliminary research shows that fingerprinting may bot work very well in the regular population – there are no studies that show how fingerprintable are regular users, who are not computer science experts. Another reason for these results is that regular population uses more mobile devices, which are less fingerprintable and less unique than the desktop devices.

We would like to thank Nick Nikiforakis for taking notes during this break-out session, and presenting the break-out results to the full group.

5.2 Research feasibility and validity

This break-out session discussed the ethical problems of doing research in the field of privacy and transparency. Researchers have agreed that repeatability has become a standard way of performing research in this field, **most of the results in online privacy and transparency are reproducible**, and researchers share voluntarily their finding and data with the community.

Many of the researchers in this session have experience in interacting with companies, while trying to obtain some data from them. However, the companies are precocious and not willing to share data, and the main reason is that researchers have shown that any **data anonymisation algorithm is not providing 100% guarantee**, and therefore companies don't take the risk of revealing user's data.

Researcher discussed that our research may **impact users**, and therefore requires IRB or another ethical committee to validate the study that involves users. Research that involves automatic crawling of web services may also **impact companies**, for example fake profiles may generate revenues for advertisers, and therefore companies may lose some amount of money due to the automatic crawls. We, as researchers, should at least estimate how much our experiments may cost to the companies.

The participants also discussed the differences between **legal and ethical issues** related to transparency research. Creation of fake profiles may violate terms of service, and causes a legal issue. We may also ask a company, whose service we are evaluating, for a permission to perform our study, and thus behave ethically. However, this may influence the practices of the company, and all "unethical", "biased" or "discriminative" behaviour may be removed before we even start our analysis. For automatic crawling, there is already a convention “robots.txt” that describes how a service may be crawled. Violating this convention is considered an ethical violation, and not a legal one.

An interesting ethical question raised during the discussion was: **Is it ok for us, researchers, to analyse data that is collected unethically/illegally by someone else?** Is there a difference on using data collected unethically from leaked data (when we don’t know how it was collected)? Even the law researcher in the group could not answer this question immediately. More concrete question raised was: Should we delete tweets from our tweeter dataset if they are deleted in twitter?

We would like to thank Nataliia Bielova for taking notes during this break-out session, and presenting the break-out results to the full group.

5.3 Bluesky proposals

The break-out session started from a provocative question in order to initiate lively discussions: **How would you design a clean slate web, both technology and business model, to do all its doing now and solve all privacy problems?**

The participants discussed whether **we should try to mimic the offline world**, despite its imperfections, or should we go for utopia, for an online world that is even more private than the offline we have now. To answer this question, the group mentioned “Evolution vs Revolution” by T. Khun, and raised another question: How close are we to a revolution? If the conditions for a revolution are near, do people have the tools to realise it? Revolutions are usually local not global. Can you revolt against the tool that has been used for revolutions?

Researchers brainstormed on **a clean state web**: Shall anti-monopoly by design be a property of a clean slate web? If you we to break a current monopoly through regulator intervention, how would we do it? They discussed micro payments, and whether they should be a basic ingredient of the clean slate design. Music industry did it. We would never have iTunes and Spotify without Napster and BitTorrent. AdBlockers seem to play the role of Napster forcing the AdTech to sit on the table and discuss alternatives. If not micro-payments, what about an all you can eat subscription on the browser. Can we feed all the publishers with 5 euro per month a la Spotify?

Finally, should we have a data levy imposed on data processors? Similar to what is being imposed on storage devices or tobacco to compensate for harm done? An ideal solution would be clean slate design without middlemen and with the user in the loop.

We would like to thank Nikolas Laoutaris for taking notes during this break-out session, and presenting the break-out results to the full group.

Participants

- Günes Acar
KU Leuven, BE
- Benoit Baudry
INRIA – Rennes, FR
- Nataliia Bielova
INRIA Sophia Antipolis, FR
- Joe Calandrino
Federal Trade Commission – Washington, US
- David Choffnes
Northeastern University – Boston, US
- Damian Clifford
KU Leuven, BE
- Angel Cuevas Rumin
Univ. Carlos III – Madrid, ES
- Steven Englehardt
Princeton University, US
- Paul Francis
MPI-SWS – Kaiserslautern, DE
- Krishna P. Gummadi
MPI-SWS – Saarbrücken, DE
- Aniko Hannak
Central European University – Budapest, HU
- Costas Iordanou
Telefónica Research – Barcelona, ES
- Nicolas Kourtellis
Telefónica Research – Barcelona, ES
- Balachander Krishnamurthy
AT&T Labs Research – New York, US
- Nikolaos Laoutaris
Telefónica Research – Barcelona, ES
- Arnaud Legout
INRIA Sophia Antipolis, FR
- Timothy Libert
University of Pennsylvania – Philadelphia, US
- Arvind Narayanan
Princeton University, US
- Nick Nikiforakis
Stony Brook University, US
- Lukasz Olejnik
University College London, GB
- Josep M. Pujol
Cliqz – München, DE
- Konrad Rieck
TU Braunschweig, DE
- Dolière Francis Somé
INRIA Sophia Antipolis, FR
- Oleksii Starov
Stony Brook University, US
- Vincent Toubiana
CNIL – Paris, FR
- Michael Carl Tschantz
ICSI – Berkeley, US
- Peggy Valcke
KU Leuven, BE
- Robbert van Eijk
Leiden University Dual PhD Centre, NL
- Diana Vlad-Calcic
European Commission – Brussels, BE