Data Transparency Lab Kick Off Workshop (DTL 2014) Report

Rafael Gross-Brown
rafael.gross-brownabreu@telefonica.com

Nikolaos Laoutaris
nikolaos.laoutaris@telefonica.com

Michal Ficek
m.ficek@telefonica.com

Patrick Dressler
patrick.dressler@telefonica.com

Jose Luis Agundez
joseluis.agundez@telefonica.com

[This article is an editorial note submitted to CCR. It has NOT been peer reviewed]

ABSTRACT

On November 20 and 21 2014, Telefonica I+D hosted the Data Transparency Lab (”DTL”) Kickoff Workshop on Personal Data Transparency and Online Privacy at its headquarters in Barcelona, Spain. This workshop provided a forum for technologists, researchers, policymakers and industry representatives to share and discuss current and emerging issues around privacy and transparency on the Internet. The objective of this workshop was to kick-start the creation of a community of research, industry, and public interest parties that will work together towards the following objectives:

- The development of methodologies and user-friendly tools to promote transparency and empower users to understand online privacy issues and consequences;
- The sharing of datasets and research results, and;
- The support of research through grants and the provision of infrastructure to deploy tools.

With the above activities, the DTL community aims to improve our understanding of technical, ethical, economic and regulatory issues related to the use of personal data by online services. It is hoped that successful execution of such activities will help sustain a fair and transparent exchange of personal data online. This report summarizes the presentations, discussions and questions that resulted from the workshop.

Categories and Subject Descriptors

H.3.5 [**INFORMATION STORAGE AND RETRIEVAL**]: Web-based services

General Terms

Design

Keywords

Privacy, Web tracking, Measurement tools, Economics of the web, Internet advertising

1. INTRODUCTION

In November 2014, Telefonica I+D hosted the Data Transparency Lab Kickoff Workshop in Barcelona, Spain. Guests included DTL partners, technologists, researchers, policymakers and industry representatives. There were over 70 participants, including 27 speakers, from more than 10 countries. Among the speakers were 24 scientists from 14 academic institutions, including Columbia University, Northeastern University, the MIT Human Dynamics Lab, AT&T Research, and the Max Planck Institute of Software Systems. Other guests and speakers included representatives from more than 18 different companies, including start-ups such as YouTechnology, Disconnect.me, and TheGoodData Co-op, as well as key industry players like Microsoft, Mozilla, News UK Corp. and advertisement industry associations. The workshop marked the creation of a vibrant community that supports research by providing scalable technology infrastructure and funding via annual open call for research proposals, with the purpose of promoting the development of tools that expose and mitigate practices that do not support transparency around online personal data. DTL2014 provided a forum for DTL members to share initial ideas and vision for the community; discuss key current issues related to transparency and personal online data; present and discuss research work done in these areas and ways to contribute to DTL, and; discuss ideas on how to organize the community, including governance, allocation of resources, sponsorship, etc.

Within that context, the workshop held specific thematic sessions on:

- Setting the Stage for DTL2014
- Transparency Tools 1
- Breakout Session 1
- Value of Data
- Transparency Tools 2
- Handling Data
- Data Start-ups
- Breakout Session 2

The workshop focused on presentations and discussions around these topics. Frequently discussed concepts included: the challenges of developing and strengthening the DTL community; the need to put users in control of their data; the poor current state of the user experience of personal data transparency; the alarming lack of widespread user engagement with the issue of data transparency; questions about the feasibility of true anonymization of datasets without sacrificing their utility; the need to bridge conversations between technologists, regulators, users, and industry; the need for more agile regulation, etc. In all cases there were lively discussions on the need to shed light on the use of personal data online and on the kind of scientific research that would be needed to do so. Participants generally agreed that there is a need for this kind of work, that DTL could help to connect the different groups working on these issues, and that users should be at the center of the DTL vision.

2. SETTING THE STAGE

This session was focused on laying the groundwork for the Data Transparency Lab, the presentation of an overview of the current regulatory context, and a legal perspective on methodologies and practices for incorporating transparency into web services. Nikolaos Laoutaris (Telefonica I+D) kick-started the workshop by introducing the problems that made the participants come together to form the DTL community: the slowly eroding trust users have for online websites and platforms, and the possibility that this eventually could lead to the absolute depletion of trust online, which would mark the end of the web as the medium of choice for people around the world. He drew parallels between the privacy problems of the web and events that have plagued TV and newspapers in the past, and showed how all of the above are instances of the basic ”Tragedy of the Commons”. He summarized the mission of DTL as ”using transparency to avoid a Tragedy of the Commons in the web due to privacy and loss of trust”. The slides from this presentation can be downloaded at `http://www.datatransparencylab.org/docs/02_Nikos.pdf`.

Jose Luis Agundez (Telefonica I+D) laid out the initial blueprints of the DTL, explaining that Telefonica is part of a group of founding members who invite organizations and individuals to join the community. He detailed that the main stakeholders of DTL are consumers, universities, companies, advertisement industry representatives, NGOs, and any brand that could benefit from being associated with the high transparency standards that DTL aims for. The structure of DTL is designed to optimize the process of grant calls and awards for researchers. Agundez also highlighted that DTL has three fundamental tracks that drive will drive its evolution: a user-centric focus (dissemination and amplification through user experience), scientific research, and a strong support by a managing Executive Committee. The slides from this presentation can be downloaded at: `http://www.datatransparencylab.org/docs/JL.pdf`.

Balachander Krishnamurthy (AT&T Research Labs) presented his view on the goals, plans and metrics of success for DTL. He highlighted the benefits for researchers, such as access to infrastructure and resources (e.g.: platform, tools, datasets). He stated that the goals of DTL could be met by offering members the possibility to have their contributions be pluggable components, and not just stand alone entities. Platforms resulting from DTL need to be able to amplify the reach of participants, offering the possibility of reaching a much larger audience. He then dove into detail on performance metrics that could help evaluate the success of DTL, including those applicable to technologists, companies (including start-ups), regulators, and advocates.

Andrea Martens (European Commision) exposed current regulatory issues related to transparency for consumers of the European Union, focused on the idea of a digital single market. Martens stated that there is a strong need for better tools for regulators to aid in policymaking and regulation and for consumers to know more about behavioral pricing. She stated that the EU Directive on Consumer Rights is highly relevant to the discrimination of consumers online (e-commerce websites will have to satisfy certain transparency requirements). The EU is aiming for free cross-border participation without barriers and the elimination of discrimination based on nationality.

Urmika Devi (Mozilla) presented the work done as Product and Data Counsel at Mozilla, working with developers and designers to build tools. She pointed out that much of her work consists on clearly communicating with developers on the legal implications of their code, specifically in relation to data transparency. She suggested that companies should be focusing on good internal communication of transparency and privacy principles, getting lawyers to think about developers and users, as well as working on transparent interfacing with users. She also mentioned that there needs to be an improvement on tools for users to express their choices, and for more engaging interfacing in relation to transparency and privacy. The focus should be on getting users to care and engage.

3. TRANSPARENCY TOOLS I

This session was focused on tools that have been designed to shine a light on data transparency and privacy issues.

Max Tucker (Columbia University) presented XRay, a tool developed to detect the targeting of vulnerable groups, data sharing among web services, and to serve as a foundation for other privacy tools in the future. XRay infers which specific input (data from a web account) yields a specific output (e.g.: advertisements, recommended products, different prices). He explained that differential correlation requires accuracy, scalability and must be service agnostic. The most striking revelation from their research was targeting related to medical diagnosis, including depression and pregnancy.

Krishna P. Gummadi (Max Planck Institute) presented an ongoing study aimed at testing the idea that machine learning classifiers can discriminate certain groups of users. He proposed a question: does the automated program introduce a form of discrimination? Gummadi explained that, in cases of discrimination, several jurisdictions (such as the US) shift the burden of proof to the defendant once a plaintiff has proven that she has been discriminated. In the case of machine learning classifiers that potentially discriminate, the defendant might be deemed to be discriminating due to the output of these classifiers.

Ruben Cuevas (University Carlos III de Madrid) presented his work on reverse engineering of targeted advertisements, providing an overview of the different kinds of targeting (contextual, geographic targeting, demographic targeting, re-targeting, behavioral targeting and interest-based targeting). His team's research goal is to help users understand the type of advertisements they encounter. He presented a few challenges they are encountering, including challenges in establishing fine-grained correlation in this area, and making their research tool scalable.

Ronny Bjones (Microsoft) and his team at Microsoft are studying how tracking takes place under identity federation, i.e., how tracking of users occurs across different sites.

4. BREAKOUT SESSION I

Participants were divided into groups for the breakout sessions. The first session involved the discussions of the following themes, each group addressing a different theme.

The first group discussed and presented on the Questions that DTL should be answering. What do users care about? The group suggested that most users don't care or engage with the issue, probably due to technical illiteracy in this area, and that future questions should address this issue. What are some basic privacy red lines? The group proposed that health questions, insurance-related information, and events that create widespread shock (like Snowden and the Facebook beacons controversy) are most alarming. What are the priorities for regulators and the advertisement sector in terms of self-regulation? The upshot of this conversation was that DTL could be a great bridge between engineers and policy makers.

A second group presented on Tools that the Lab should be building, highlighting that tools related to user profile trading are desperately needed. They set forth a key question: how could DTL convince users that privacy matters, and giving them more control over their information. Future discussions should also consider the possibility of building a general platform that could be used to collect information for users to better understand how their information is used. They suggested that regulators could use these tools for auditing, and firms could use them for internal auditing of their transparency practices.

The third group discussed what type of Data should be collected and released in the context of the DTL. Their discussion focused on whether data should be released or not, considering that anonymization is a key challenge. DTL should quantify the risk of de-anonymization, and understand that there's no full anonymization for rich data: there is an inherent trade-off to be made when considering utility versus privacy risks. They raised concerns about potential biases that might arise from working with anonymized data. They also suggested that DTL should not incentivize users to share their data, but rather let this happen out of their own initiative. They also asked who would be the custodian of data collected in the context of DTL.

The fourth group presented on the Expertise needed for DTL success. This group focused on the crucial issues that should be addressed: education and conversation with users (user experience design and communications); better protocols and data manipulation (engineering); improve the state of current policymaking situation (bad implementations of bad laws, such as France and cookies opt-out law). They stated that the rhetoric around privacy issues should change. Ideas included crowdsourcing privacy policy tolerance taxonomies (i.e., users subscribe to certain policies that their friends are adopt for all services) and peer based reviews of privacy policies. The open source community should be engaged: a first step in that direction would be to map out key resource groups and figure out how to engage them.

5. VALUE OF DATA

Carla Bonina (Surrey Business School) presented an overview of the present value of open data, pointing out that it is a 3 trillion dollar economy, and that the use and reuse of open data sets is valued at around 40 billion euros. She proposed that there's a big tension on the role open data plays in the privacy discussion, and stated that trust has already been lost. She proposed that mechanisms must be set up to understand who is using open datasets, and that there's a real need to make regulation more agile.

Marcos Menendez (TheGoodData) opened his presentation by stating that transparency is not a goal, but a path. The first stop on this path is security, citing as an example that the eBay data breech lowered the company's value (by 7 percent), only to recover its pre-breech value two months later. Transparency involves letting users know what we do with their data and what decisions are taken based on it. Transparency stacks on security. Beyond transparency lies control of the data, which is governed by terms of services.

Dirk Stelzer (TU Ilumenau) aims to develop a transparency reference model. This requires an economic analysis of data transparency, which would entail identifying relevant measures and metrics, such as those that companies apply to enhance/undermine data transparency. He also proposed to perform a comparative analysis of the data transparency legal framework. His research would also comprise means to define levels of data transparency.

Jacopo Staiano (FKB) presented his team's work on the economics of personal mobile data, aimed at identifying the value people assign to different kinds of personal data. During a 6-week long user study in a living lab deployment with 60 participants, they collected their daily valuations of 4 categories of mobile Personally Identifiable Information ("PII") – communication, e.g. phone calls made/received, applications, e.g. time spent on different apps, location and media, photos taken – at three levels of complexity (individual data points, aggregated statistics and processed, i.e. meaningful interpretations of the data). Their findings show that the most sensitive and valued category of personal information is location. He highlighted that there are statistically significant associations between actual mobile usage, personal dispositions, and valuations.

6. TRANSPARENCY TOOLS 2

Pablo Rodriguez (Telefonica I+D) kicked off the session by providing examples of benefits and power that data could offer users in the future. He stated that Internet capacity would soon afford the capability of recording everything we capture with our 5 senses. This would have significant implications in the development of new technologies, forecasting that individuals would become a new platform. A personal data bank would give individuals control over their "data souls," giving them important benefits deriving from the use of their data, including health benefits (preventing them from becoming ill, avoiding hospital costs), better terms for loans and insurance agreements (you pay as well as you drive), to the ability to donate their data after death.

David Choffnes (Northeastern University) presented his research on measuring PII leaks from network flows. His team created Meddle, a framework that combines virtual private networks (VPNs) with middleboxes to provide an experimental platform that aligns the interests of users and researchers. This tool offers new opportunities for measuring and understanding mobile traffic, and designing new network features to improve the mobile experience. Meddle box is highly customizable in how it mediates a device's online connection, and its corresponding information flow. Challenges include successful training of classifiers and identifying appropriate measures of success. A future step for this project is accessing user feedback on their experience.

Alan Mislove (Northeastern University) presented his research on the personalization of online services. Personalization, now ubiquitous, is often not transparent or even announced. There is an ongoing practice of comprehensive user tracking, which involves the online and offline collection of user data. It is currently unclear how this tracking is used to users' advantage and/or disadvantage. The research scope was narrowed down to examine personalization due to client-side state associated with request. This research should be a part of a larger project aimed at understanding how web services collect data and how it affects the information presented to users.

Christo Wilson (Northeastern University) gave a brief overview of their personalisation research, specifically price discrimination. He pointed out that, in the United States, price discrimination is legal in some cases. He also gave examples of online price steering, most notably the Orbitz 2012 controversy (Mac users were shown more expensive hotels). Their research also shows that different e-commerce sites present results differently based on device geolocation. A telling example of "steered" content are Google Maps results that show different international borders depending on the origin of the query. Results that display the Sino-Indian border vary depending on whether the query originated in China or in Germany: the former partial to Chinese interests, the latter reflects the undefined and contentious nature of the border.

7. HANDLING DATA

Jonathan Ukena (Telefonica) presented his team's work on building data anonymization platforms. His team developed a platform designed to comply with German data transparency and privacy regulation, which could offer new business opportunities, and could be used in different cases because of its flexibility and generic quality.

Tristan Henderson (CRAWDAD) presented on his work for CRAWDAD, the world's largest community archive of wireless data. CRAWDAD is comprised of 116 datasets and tools that have been used in over 1,600 papers. However, for all its success, CRAWDAD faces significant challenges, which are relevant to the DTL community. Some of them include tracking usage of data (very few notify CRAWDAD of the use of their data), bad discoverability, privacy issues, anonymization/sanitization of data sets, and sustainability of the initiative. He concluded by asking important questions about the use of datasets: should DTL be the go-to place for all personal data? How can responsible usage and contribution be encouraged? Who should fund and protect this resource?

Arkadiusz Stopczynski (MIT Media Lab, Human Dynamics Group) focused his presentation on the impossibility of anonymizing data while preserving its utility. He proposed that researchers should be looking to gather real-time data that is granular, deep, with a feedback loop channel, and that gives the user control. A DTL goal should be to build tools with APIs that enable this. He also talked about the distinction between technical and operational privacy, and that modulating the technology to protect privacy inevitably entails a reduction of the value of the data.

Patrick Loiseau (Eurecom) proposed a framework to protect privacy when connecting with an untrusted provider based on adding noise before revealing data. Value depends on the amount/kind of data revealed by others, and information should be treated as a public good. He asserted that there's a need to accurately quantify the value of information.

8. DATA START-UPS

Kimon Zorbas (DBCG) went over a series of issues and challenges that the business of online advertising faces, including inadequate policymaking, messiness (frameworks, formats), the challenge of delivering a single user experience, questions about data ownership, inadequate pricing mechanisms, and lack of transparency in the market. He also highlighted that everybody is talking about user data, but nobody is talking to users.

Gus Warren (Disconnect) presented Disconnect's private and secure advertisement platform, where users have the option to opt-in to private and secure advertisement. If users opt-in they will see appropriately targeted advertisements based on their Disconnect "super profile." The role of Disconnect is to create a targeting signal based on a super profile, which would allow for targeting of users while preserving anonymity. This would benefit publishers (higher CRMs), advertisers (more precise campaigns), and agencies (increase value of industry due to more precise targeting).

Chris Reichert (YouTechnology) opened by stating that consumer generated data is growing at an unprecedented pace, and that a 2,000 percent increase in global data is expected by 2020. YouTechnology aims to put the control of data in user's hands. They propose privacy preservation through a distributed federated model, creating a robust ecosystem of databank contributors, seeded with core operator data: a personal data bank.

Alessandro de Zachne (News UK) stated that general credibility issues plague the market of third-party data trading, mainly because of problems such as data being collected in the incorrect context, aggregate level data used to infer traits of individual users, among others. Consequently, the market is wasteful, the user experience is subpar, and there is a widespread perception of lack of privacy.

9. BREAKOUT SESSION 2

A second breakout session was geared towards fostering a creative community and finding new ways to address the problem of transparency and privacy. Four different groups were prompted to generate ideas for platforms, products, games, services, etc. that could potentially widen user engagement, and discuss the implications of new approaches. Ideas included putting users in the role of an advertiser for a day and contests among users to generate cookie data that would reflect a real person (example: build the profile of a wealthy baby boomer). Some proposed that the focus should be on awareness and education. Others thought that games designed within the context of DTL could be an engaging way for users to contribute their data to DTL research.

10. FUTURE DIRECTIONS

The participants closed DTL2014 by discussing future paths for the community. Short-term milestones include consolidating the founding structures and inviting more organizations to join by promoting the community. Mid-term and long-term milestones include awarding the first research grants, hosting user-centric events, and setting up other mechanisms to effectively amplify the research produced by members of the DTL community.

11. WORKSHOPS PRESENTERS

Jose Luis Agundez Dominguez	Telefonica Digital
Pere Barlet-Ros	UPC Barcelona Tech
Ronny Bjones	Microsoft
Carla Bonina	Surrey Business School
David Choffnes	Northeastern University
Ruben Cuevas	University Carlos III de Madrid
Alessandro De Zanche	News UK
Urmika Devi	Mozilla
Krishna Gummadi	Max Planck Institute
Tristan Henderson	University of St. Andrews
Balachander Krishnamurthy	AT&T Research
Nikolaos Laoutaris	Telefonica Digital
Patrick Loiseau	Eurecom
Andrea Martens	European Commission
Marcos Menendez	TheGoodData
Alan Mislove	Northeastern University
Nuria Oliver	Telefonica Research
Chris Reichert	You Technology
Greg Riley	You Technology
Pablo Rodriguez	Telefonica Research
Jacopo Staiano	FKB
Dirk Stelzer	TU Ilmenau
Arkadiusz Stopczynski	MIT Media Lab, Human Dynamics
Max Tucker	Columbia University
Jonathan Ukena	Telefonica Digital Germany
Gus Warren	Disconnect
Christo Wilson	Northeastern University