Data Transparency: Concerns and Prospects

By NIKOLAOS LAOUTARIS
Data Transaprency Lab, Barcelona 08019, Spain

I. INTRODUCTION

The question of "how far" technologies and business models of the web should go into collecting personal data of unassuming, or at best moderately informed citizens, appears to be one of the most timely questions of our times. Indeed, whenever we read a news article, "like" a page on a social network, or "check in" to a popular spot, our digital trace collected, processed, fused, and traded among myriads of tracking, analytics, advertising, and marketing companies becomes an ever more accurate descriptor of our lives, our beliefs, our desires, our likes and dislikes. The resulting revenue from marketing and advertising activities driven by the digital traces of millions of people is what funds the free online services we have come to rely upon.

In this opinion article, I will lay down my thoughts around data transparency and its role in ongoing data protection and privacy debates. The material draws upon my experience from conducting research in the area over the last six plus years, running the Data Transparency Lab's Grant Program[1] in 2015, 2016, and 2017, and attending several computer science, policy, and marketing and advertising events. The objective of the article is to discuss the possibility and the likelihood of data transparency acting as an important positive catalyser of data protection problems, as well as to point toward concerns and challenges to be addressed in order for this to materialize. Most of the discussion applies to the use of personal data by marketers on the fixed and mobile web, but some parts may also be relevant to other online and offline use cases and/or types of data (e.g., off-web health and financial data).

For years, the practice of collecting data on individuals at unprecedented scale was a nonissue for most people, for the simple reason that the public, and even governments, were just unaware of its magnitude, precision, and detail. The last few years, however, attitudes have started to change and the topic of privacy is increasingly appearing in the media and public discussions. This has stirred a huge public debate about who should be the rightful owner of personal data, and where to draw the red line of what is socially acceptable to track and monetize on, and what is not. Indeed, the same tracking technology of cookies and tracking pixels used to detect one's intention for buying new running sneakers, thus prompting an interesting discount at a nearby shop, can also be used to infer one's medical condition, political affiliation, or sexual preference, thus delivering offers at the wrong

Digital Object Identifier 10.1109/JPROC.2018.2872313
[1]http://www.datatransparencylab.org

time and place, or even worse, releasing the information to third parties that may use it in a discriminatory, excluding, or generally unfair manner. The former use of data has the potential to increase value for individuals, retailers, technology providers, and the society, whereas the latter can be detrimental to the trust put by individuals and institutions in technology.

II. TRAGEDY OF THE COMMONS AND THE WEB

What is particularly alarming is that the economics and incentives of exploiting personal data for advertising and marketing have all the characteristics of a "Tragedy of the Commons" (see [1]), in which consumer privacy and trust in the web and its business models are a shared commons that can be overharvested to the point of destruction. The essence of the problem is that even if most technology companies manage to agree on a set of principles, there will always be sufficient temptation for some to push the boundaries in pursuit of greater gains, while inflicting a heavy cost to society. Indeed, from the narrow perspective of some companies, all it takes to pursue a business that involves intrusive and unethical collection of very sensitive personal data, is a paying customer. The above seems to be verified by examples appearing in the press of trackers that compile lists of anything from suspected alcoholics and HIV-positive individuals, to active police officers.[2]

In essence, the narrow self-interest of a subset of data collection companies is eroding a valuable commons—the trust put by people in the web, or inversely their hope that nothing bad will happen to them by being carefree online. If, in the minds of citizens, ordering a drug online is associated with the risk of leaking medical information to health insurance companies, then they may very well abandon the web and just walk to a pharmacy. This means that the web, as big and successful as it is currently, is not invincible. It too can fall from grace like newspapers and broadcast TV have in the past, albeit for other reasons. Loss of public trust appears to be the Achilles' heel of the web and is being fueled by questionable practices from large and small companies alike.

III. THE ROLE OF TRANSPARENCY IN DATA PROTECTION DEBATES

Transparency is often heard in debates about governance, business, science, and matters of public life in general. According to Wikipedia, transparency is about: "operating in such a way that it is easy for others to see what actions are performed. Transparency implies openness, communication, and accountability." Transparency, in its different applications and contexts, largely embodies the famous quote of American Supreme Court justice Louis Brandeis that "Sunlight is said to be the best of disinfectants."

In the context of data protection and privacy of online services, transparency can be understood as the ability to credibly answer questions, such as the following.

- What information is being collected (stored, and processed) about individuals online?
- Who is collecting it?
- How is it being collected?
- How is it being used?
- Is it leaking to other unintended recipients?
- What are the consequences of such online leakage of private information?

Information leakage is a natural phenomenon in both the offline and online life. In the offline world, whenever we walk on a street or are seen at a public place, we are effectively giving up on our so-called "location privacy." Our clothes, hobbies, the car we may drive, or the house where we live convey information about our financial status, employment, and taste. Similarly, in the online world, networks need to know where we are in order to deliver our calls, e-mails, or chat requests. Social networks need to display our real names so that our offline friends can also befriend us online. The above realizations give rise to a simple alternative to trying to unknot the "utility versus privacy" tradeoff in data protection. Since we can neither stop all online leakage under the current technological paradigm, nor prescribe a generic, context-unaware solution to the tradeoff, we can instead try to reduce information leakage, while keeping an open eye for controversial practices driven by collecting personal data that go against public sentiment or the letter of the law. Such an objective can be achieved on top of existing web technologies and business models without requiring some radical redesign. Transparency is the guiding light pointing to problematic technologies and business practices that will require revision, if we are to keep a safe distance from a tragedy of the commons on the web.

Transparency has already proved its worth in what is probably the greatest technopolicy debate preceding data protection—the network neutrality debate. Network neutrality is the simple principle, now turned into regulation and telecommunications law, that a network operator cannot delay or drop one type of traffic from a certain application in order to protect or expedite the rest. Almost a decade ago, the network neutrality debate was ignited by reports that some telecommunications companies were using deep packet inspection (DPI) equipment to delay or block certain types of traffic, such as peer-to-peer (P2P) traffic from BitTorrent and other protocols. Unnoticed initially among scores of public statements and discussions, a group of computer scientists from Germany developed Glasnost [2], a set of tools for checking whether a broadband connection was being subjected to P2P blocking. All a user had to do to check whether their ISP was blocking BitTorrent was to visit a webpage and click on a button that launched a series of simple tests—basically streaming two flows of data toward the user, one

[2]https://money.cnn.com/2013/12/18/pf/data-broker-lists/.

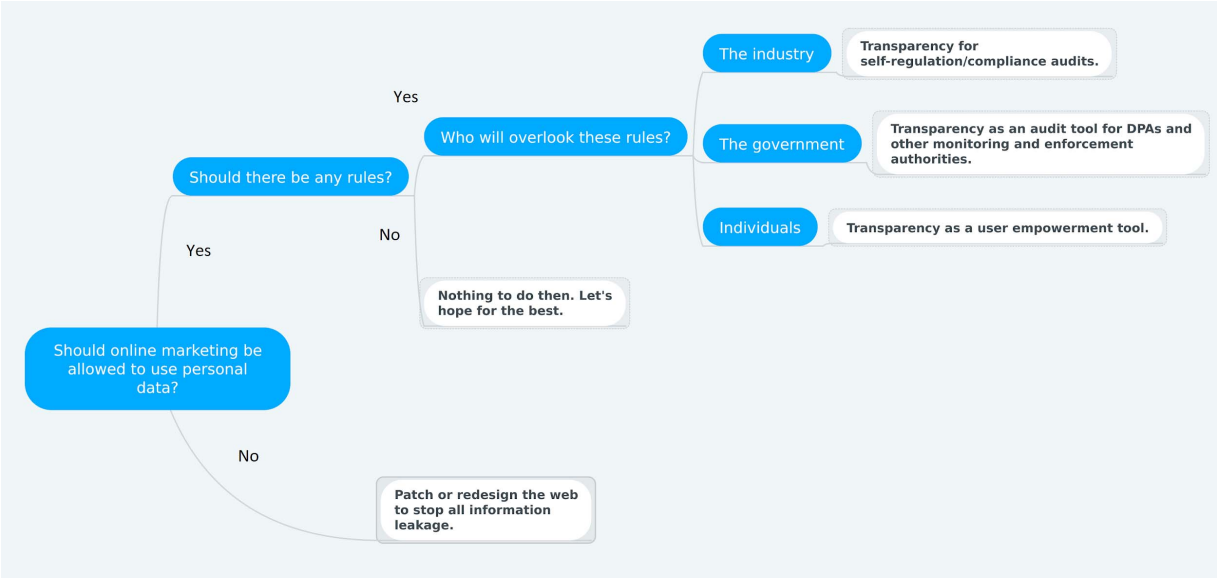

Fig. 1. ***The role of transparency in the context of the greater data protection debate.***

appearing to be P2P and one not. By comparing the corresponding reception data rates for the two flows the tool could say if P2P was throttled; anyone could check for themselves if their provider blocked P2P traffic or delivered the advertised speed of their plan. The existence of such easy-to-use tools created the right incentives that eventually obliged telecommunications companies to give up blocking or be open about it, and to indeed deliver the promised data rates.

The above network speed measurement tools are early examples of what I will henceforth call transparency software, to refer to any software purposefully built for providing transparency and for "shedding light" into public debates involving technology. An important idea, and frankly *raison d'être* for transparency software, is as follows: "complex technology can only be tamed by other, equally advanced, technology." Indeed, investigating complex data protection debates without specialized software for collecting evidence and testing hypotheses is like conducting preflight tests or periodic car inspections without specialized equipment that can test all the complex components of an airplane or a modern car. In all these domains, there exist entire fields dedicated to developing formal methods for testing and verification. In the same way that air transportation, the car industry, health, and other domains have benefited from purpose-built testing tools, online data protection needs to develop its transparency methods and software.

Although this article is not meant to be a complete survey, in the remainder I will discuss several examples of transparency software for things such as revealing personally identifiable information (PII) leakage in real-time detecting online price discrimination, detecting online targeted advertising, detecting advanced online tracking through fingerprinting, and others.

IV. TRANSPARENCY FOR WHOM?

Fig. 1 illustrates at the topmost level how and where transparency plugs into the general data protection debate. Assuming that some use of personal data for online marketing in exchange for free service is deemed acceptable, and that the Tragedy of the Commons is not something we should leave to luck, the figure suggests that transparency can be beneficial for all three main stakeholders of the data protection discussion. Transparency should not be seen as an alternative to existing efforts of these stakeholders, but as an extra tool at their disposal.

For the online tracking and advertising industry, transparency is essential to its efforts to convince government and citizens that it can effectively self-police and self-regulate the sector, making sure that individual companies do not perform actions that go against the public sentiment, or worse, data protection laws. The various Codes of Conduct and Best Practices documents issued by sector representative bodies and organizations make lots of use of the term transparency, but they often get criticized as being mere intentions without any real means for enforcement and actual demonstration of commitment and application. This is where transparency software can play a key role, by allowing anyone to independently check that companies make good on their promises. A smartphone app that commits to not communicate PII back to its servers or other third parties can be checked by software such as ReCon [3], Lumen [4], and AntMonitor [5]. A website and its

advertising partners that commit to not target minors can point users to Aditaur [6], a tool which anyone can use to verify the claim. In essence, the existence of such software allows the sector to make more credible and verifiable promises regarding its ability to self-regulate its data treatment practices.

For individual citizens, transparency is all about empowerment and freedom of choice. For every basic online service, there is typically a magnitude of alternative service providers offering it, each one with a potentially different approach and sensitivity toward data collection. Users are accustomed to rating online services in terms of performance, simplicity, feature richness, but gauging the quality of data management practices has for the most part been out of reach. Being able to evaluate the quality-over-privacy-risk ratio of different services empowers users to select the one providing the right balance. For example, PrivacyMeter [7] can display a risk score for every website visited by a user in real time. If a user deems the risk of visiting, say, a news portal to be too high, they can opt for an alternative one with similar content but better performance in terms of privacy. By doing so, users emit clear signals to the industry and contribute through market pressure to pushing it toward the right direction.

Last but not least, government agencies, especially Data Protection Authorities (DPA), need transparency for both their proactive monitoring activities, as well as their investigative activities following complaints from citizens, watchdog groups, or other companies. Transparency software can help DPAs scale up their investigations and even proactively monitor for offending practices, something that does not appear to be possible via *ad hoc* manual investigations. For example, with AppCensus [8], entire marketplaces of mobile apps can be checked for leakage of PII information, by automatically analyzing their binary executable distributions. Similarly, WebCensus [9] allows monitoring millions of domains every month to catalogue and rank their tracking practices. With Aditaur, DPAs can proactively check at scale thousands of popular domains for targeted advertising toward sensitive groups such as children, or driven by sensitive personal data about health, political, or sexual orientation that are protected under European Union's (EU's) GDPR law.

V. TRANSPARENCY OF WHAT?

As mentioned before, some perceive data protection as the challenge of understanding and limiting the amount of personal information that can be leaked and collected online. Although this has an importance of its own, for others the main motivation and driver for discussing data protection matters is understanding and limiting the consequences of personal data leakage. Transparency is important and can help at both levels of the debate. Data transparency and corresponding tools such as ReCon, Lumen, and AntMonitor are about revealing what data is leaking and where it is going. Algorithmic transparency, on the other hand, is looking at how personal data can end up fueling biased, discriminatory, and generally unfair automated decision making that impacts the life of real people. For example, the Price $heriff [10] reveals whether a consumer searching for a product online is being subjected to online price discrimination by algorithms that decide on the spot a dynamic price for each customer based on their perceived willingness to pay, extracted from information about them. FDVT [11] measures in real time the different economic valuation that Facebook advertisers have about different users, by tallying up their advertisement bids for product placements. Algorithmic transparency, of course, goes beyond the online services mentioned above, and can be applied to a range of offline contexts from health, to finance, to justice, but such areas and applications go beyond the scope of the current note.

VI. CHALLENGES

Next, I discuss some of the difficult challenges that need to be addressed if transparency is to make a positive dent upon privacy problems on the web.

A. Crowdsourcing Privacy Related Data

Several of the tools mentioned above are crowdsourced in nature, i.e., they rely on real users sharing their observations, in order to reverse engineer some aspect of an online service. eyeWnder [12], for example, relies on users reporting the advertisements seen on different pages in order to construct a crowdsourced database through which one can identify active advertising campaigns and the demographics of users targeted by them. Similarly, the Price $heriff relies on users reporting the price they see for the same product at the same site to detect instances of online price discrimination. Both tools use specialized encryption and anonymization techniques to protect the privacy of users that report back to the database the ads or the prices they have seen. Crowdsourcing is a powerful tool for detecting different types of discrimination and bias, but requires having in place a solution for protecting the privacy of users that contribute to the crowdsourced corpus of data. The above two tools use *ad hoc* techniques tailored to their specific function, but there is a clear need for developing more generic privacy-preserving crowdsourcing techniques and platforms that will make it easier to develop additional transparency tools for other problems.

B. Evaluation Criteria/ Reproducibility/Correctness

The appearance of several transparency tools in the last two to three years (of which I have touched only upon a small subset) is testament to the very significant amount of work that has been done in the area during a rather short amount of time. Still, the area is only in its infancy and thereby important requisites for

growth and eventual maturity are yet to be fulfilled. One of them is establishing common criteria and metrics upon which different transparency tools looking into the same task will be compared. Having the ability to directly compare different approaches is fundamental for the evolution of the area, the validity, and the correctness of the findings. In the same spirit, the findings of a tool need to be reproducible. This is difficult to achieve when the tool operates in the wild. eyeWnder, for example, can label an ad banner as targeted today but it might be impossible to reproduce the result after a week since the underlying advertising campaign may no longer exist, or may have changed its target audience. Reproducibility goes hand in hand with the ability to compile extensive crowdsourced data sets on privacy matters upon which the reproducibility of a tool can be checked or its performance compared with alternative approaches on the same problem.

C. Bootstrapping/UX Challenges/Outreach

Both privacy-preserving crowdsourcing as well as the establishment of common evaluation criteria are technical problems and, as such, something that a technical community of computer scientists and engineers knows how to handle. A different type of challenge for the area is finding enough users for these tools. For tools that are crowdsourced in nature, establishing an initial user base is a fundamental prerequisite for allowing them to derive credible and useful findings. Even for tools that can work by collecting data without user input (e.g., through crawling and scraping of information), having a user base outside research is largely a measure of true impact and success. To get there, we need to work on improving the usability aspects of such tools and adapting them to the needs and capacities of nonexpert users. We also need to work on disseminating them and putting them in front of end users.

VII. CONCLUSION

With the above, I hope I have convinced you that transparency can have an important role and contribution in contemporary data protection debates. In my mind, a very important first milestone is making sure that the online world is at least as transparent as the offline world. This may seem uninspiring on the surface, but it is actually a very difficult objective in practice. The scale of data collection online is at a totally different level from that offline. Rules and regulations are established for many offline activities, from credit rating, to equality of access to public services, whereas the online equivalents are left to chance. Finally, many transparency aspects that we take for granted in the offline world are hard to achieve online. Take price discrimination as an example. Two customers walking to a coffee shop see the same price for "cafe latte" written on the wall. If the clerk charges one of them a different price, questions will immediately follow. In the online world, the two customers can order the same coffee at the same time and pay a totally different price without any of them ever realizing it. This is because in the online realm, the "world" around us is dynamically generated, thereby, we do not even have the benefit of a common reference. Checking for being "followed" or discriminated against is more difficult online than offline. Of course, this is just under the current state of affairs. The same technology used for surveilling or discriminating against at scale can be flipped on its head and used instead for shedding light and providing transparency at scale. This means that an online world that is safer and fairer than the offline one is an open possibility that we should consider and pursue.

REFERENCES

[1] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, Dec. 1968.
[2] Glasnost. [Online]. Available: http://broadband.mpi-sws.org/transparency/
[3] Recon. [Online]. Available: https://recon.meddle.mobi/
[4] Lumen. [Online]. Available: https://www.haystack.mobi/
[5] *AntMonitor*. [Online]. Available: http://antmonitor.calit2.uci.edu/
[6] Aditaur. [Online]. Available: https://www.lstech.io/aditaur
[7] *PrivacyMeter*. [Online]. Available: https://chrome.google.com/webstore/detail/privacymeter/anejpkgakoflmgebgnombfjiokjdhmhg
[8] AppCensus. [Online]. Available: https://appcensus.mobi/
[9] WebCensus. [Online]. Available: https://webtransparency.cs.princeton. edu/webcensus/
[10] Price Heriff. [Online]. Available: http://www.sheriff-v2.dynu.net/views/home
[11] FDVT. [Online]. Available: https://fdvt.org/
[12] eyeWnder. [Online]. Available: http://www.eyewnder.com/