A NEW APPROACH FOR DATA SHARING - DATA TRANSPARENCY LAB WORKS TOGETHER WITH AIRCLOAK AND THE OFFICE FOR CREATIVE RESEARCH EXPLORING NEW WAYS TO SHARE DATASETS WHILE PRESERVING USERS PRIVACY

The primary goal of Data Transparency Lab is to shed light on online personal data practices. In cooperation with researchers and key industry players, DTL supports the development of tools and research that help end-users better understand how their online personal data is used.

Many of those tools must collect data from end-users to shed light on how certain industries are using personal data (e.g. price discrimination, targeted advertisement). DTL aims to make this data available to the broader research community to enable more research and gain a better understanding of online privacy.

Sharing personal data has always been risky and difficult. Traditional strong anonymization techniques like K-anonymity distort data too much, so typically data is shared after simple de-identification (removing personally identifying data like names and account numbers). However, the risk of re-identification remains.

For instance, in 2006 Netflix launched a public competition [1] to provide a better movie-recommendation algorithm by releasing a dataset of 100 million movie ratings made by 480,000 customers. Although dataset was de-identified (e.g. real names were replaced with random unique identifiers) researchers showed how many users could be re-identified by comparing them against non-anonymised movie ratings posted at the Internet Movie Database [2].

To avoid this danger, personal data is normally shared only with selected trusted partners under controlled conditions. This limits the amount of sharing, and in particular discourages researchers who wish to just “play around” with the data to see what interesting insights can be gathered.

To overcome these limitations, DTL is exploring new techniques that allow researchers to share datasets quickly, easily, and broadly with negligible risk of re-identification. This exploratory work is being done in cooperation with two companies: The Office for Creative Research (OCR) [3] and Aircloak, a spin-off of the Max Planck Institute for Software Systems [4].

The Office for Creative Research developed Floodwatch, a browser application [5] that records the ads that users see on their browsers, and provides a visual collage of the ads back to the user. OCR is eager to make this data available to researchers so as to better understand the online advertising ecosystem, but only if they are confident that personal data is protected.

Aircloak [4] has implemented a “cloaked database” that combines new anonymization techniques with “Trusted Computing” secure hardware and ``zero-access’’ system hardening. Once data is uploaded to the cloaked database, it can be only accessed via the anonymization interface. Queries run over raw data, but the answers are filtered and noised so that it is extremely difficult to obtain personal data even by a determined attacker.

DTL, OCR, and Aircloak completed an initial study to gain experience with the cloaked database, and determine if useful and accurate answers could be obtained from the database while still protecting personal data. We loaded an initial version of the Floodwatch Dataset into both a cloak and an unprotected database. Among other things, the data included: a non-identifying user ID, an identifier for the ad’s image, the page on which the ad was seen, and the time when the ad was viewed. A set of queries were performed against both the cloak and the unprotected database, and compared for accuracy, ease-of-use, and privacy protection. A report of this study is available online [6].

The key conclusions we achieved were:

1. Similar insights to the ones achieved by interacting directly with the end-user could be found. For instance, we executed queries to calculate statistics such as number of ads received by each user (total and distinct ads), websites that deliver ads, ads seen by the most distinct users getting similar statistical results that the ones calculated directly using the raw data.
2. The cloaked approach was demonstrated to hide data that could be used to re-identify individual users, and that was readily available through the unprotected database. In particular, images of private individuals seen by only a single user could be found in the unprotected database, but where hidden by the cloaked database.
3. Usability of the cloaked database is somewhat more difficult than the unprotected database. In particular, analysts must be aware of the kinds of data that can be hidden by the cloak, and must take care not to let this distort the statistical properties of the answers.

We would like to give researchers the opportunity to understand and use this new technology. To obtain access to data via a cloak, please email us at [xxxxx@xxxxx.xxx](mailto:xxxxx@xxxxx.xxx).

This proof of concept is just an initial step towards new mechanisms for sharing Datasets. The Data Transparency Lab is planning to launch a Data Sharing program in order to maximise the number of researchers working in Data Transparency. Stay tuned!

[1] <http://www.netflixprize.com/>

[2] <http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf>

[3] <http://ocr.nyc/>

[4] <https://aircloak.com/>

[5] <https://floodwatch.o-c-r.org/>

[6] The url of the report