Tracking Personal Data Use: Provenance And Trust

Lucja Kot
Cornell University
Ithaca, NY 14853, USA
lucja@cs.cornell.edu

1. INTRODUCTION

In the era of Big Data, every individual is the target of intensive data collection by parties from the government to grocery store chains. Anecdotal evidence suggests that opting out of the data collection process is effectively impossible [3]. A recent report commissioned by the White House revealed a broad public concern about the collection and use of personal data by untrusted agencies and businesses [1].

As a result, we have seen an effort to improve the transparency of data collection and use. Due to legislative and public pressure, many data collectors now publish privacy policies that explain what personal data is stored and how it is processed. For example, Google's policy [2] states that "We may combine personal information from one service with information, including personal information, from other Google services [...]. We will not combine DoubleClick cookie information with personally identifiable information unless we have your opt-in consent." Such policies are useful but have shortcomings; as English-language documents, they are both too confusing for novice users and too vague for experts, and they require human effort to create and maintain.

A better solution is to create technological tools that empower individuals to track what happens to their data. The same problem has been addressed in scientific data processing through abstractions and algorithms for workflow provenance [5]. It is time to apply these techniques to the problem of personal data use; just like scientists can trace what happens to individual data points from a dataset, individuals should have access to a "Personal Data Use Workbench", where they can browse how a company or government agency is using their data.

2. RESEARCH CHALLENGES

Making Personal Data Use Workbenches a reality is not a simple application of scientific workflow provenance techniques; it requires addressing several unique challenges.

First, personal data pipelines sit in a special setting as far as trust is concerned. In scientific workflows, the user asking provenance queries owns the entire data and the pipeline. In our case, the individual who wants to trace his/her personal data does not own the rest of the data in the system, and the processing operators in the pipeline are highly proprietary and sensitive. We need a way to provide meaningful answers to provenance queries without disclosing operator code, based only on high-level specifications of the operators that the pipeline owner is willing to disclose. In addition, we need to preserve the privacy of other individuals' data that is fed into the same operators. In summary, we need to understand and resolve the tradeoff between allowing users to track their own data in detail, and violating the interests of the other parties in the system.

Second, the operators used in personal data processing are complex and heterogeneous. Tracing provenance in a pipeline with arbitrary operators is difficult, although there are solutions for frameworks like Map Reduce [6] and Pig [4]. When operators are user-defined functions, the most realistic approach is to provide a provenance API and ask a human developer to specify how the operator maps input data to output data. This approach was pioneered in the SubZero system [7] but needs to be extended beyond the setting of scientific data processing.

Third, users may want to ask provenance queries that do not occur in scientific workflows. In addition to normal queries that traverse the pipeline going forward ("how is my address used"?) and backward ("how do they know I am engaged"?), users may pose graph pattern-matching queries such as "is my photo ever combined with my location?". We need to decide what query languages and abstractions to use and support them with fast processing algorithms.

3. REFERENCES

[1] The 90-day review for big data. http://www.whitehouse.gov/issues/technology/big-data-review.

[2] Google privacy policy. http://www.google.com/policies/privacy/.

[3] My experiment opting out of big data made me look like a criminal. *Time Magazine*, May 1st, 2014, http://time.com/83200/privacy-internet-big-data-opt-out/.

[4] Y. Amsterdamer, S. B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen. Putting lipstick on Pig: Enabling database-style workflow provenance. *Proc. VLDB Endow.*, 5(4):346–357, Dec. 2011.

[5] S. B. Davidson and J. Freire. Provenance and scientific workflows: Challenges and opportunities. In *SIGMOD*, 2008.

[6] R. Ikeda, H. Park, and J. Widom. Provenance for generalized map and reduce workflows. In *CIDR*, 2011.

[7] E. Wu, S. Madden, and M. Stonebraker. SubZero: A fine-grained lineage system for scientific databases. In *ICDE 2013*.

7th Biennial Conference on Innovative Data Systems Research (CIDR '15)
January 4-7, 2015, Asilomar, California, USA.