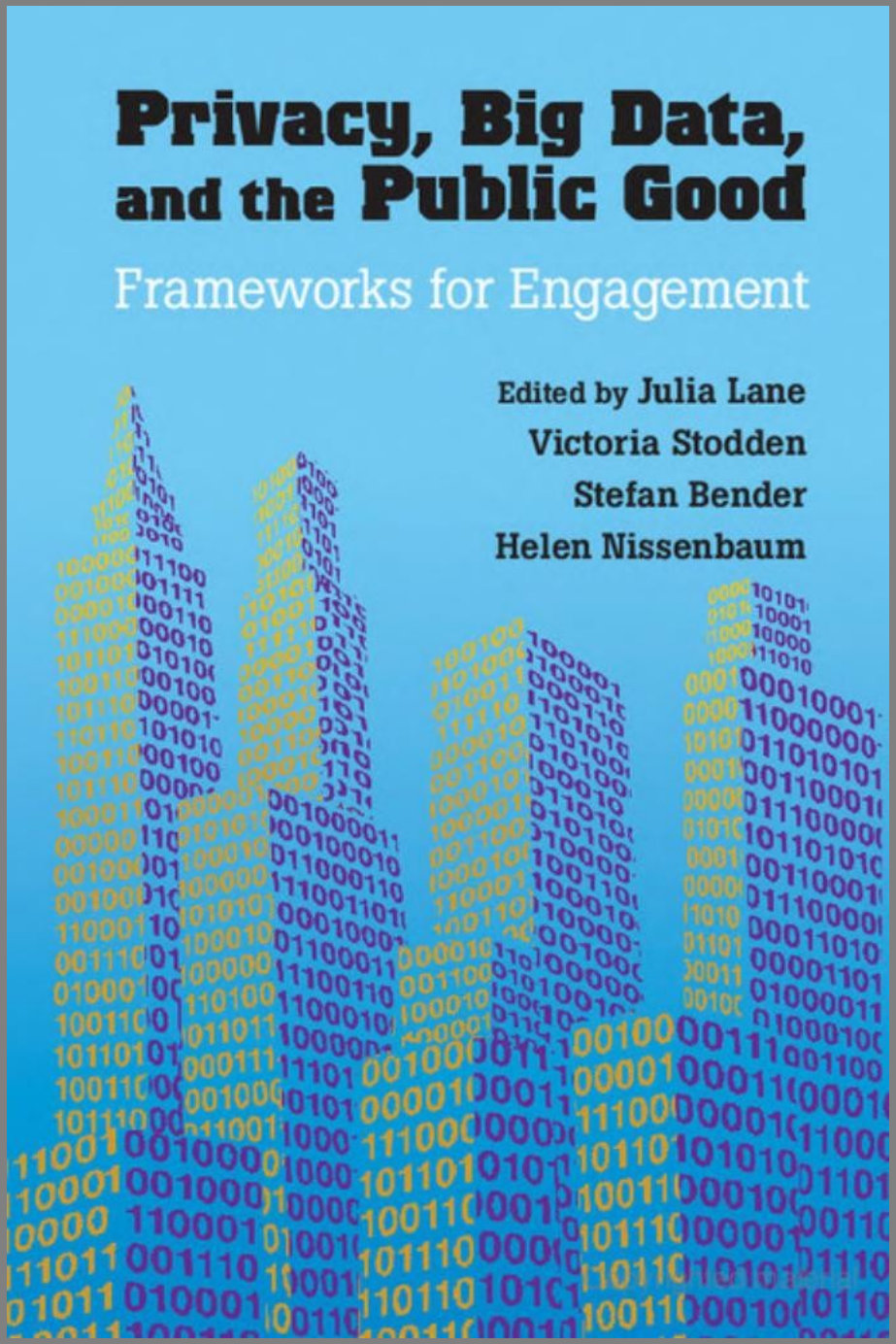
Privacy, Ethics, and Big Data

巨量資料的隱私與倫理議題

2015-07-08 @ TAHR

Tyng-Ruey Chuang (trc) 莊庭瑞



Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 75–98

Privacy and Data-Based Research

Ori Heffetz and Katrina Ligett

On August 9, 2006, the "Technology" section of the *New York Times* contained a news item titled "A Face Is Exposed for AOL Searcher No. 4417749," in which reporters Michael Barbaro and Tom Zeller (2006) tell a story about big data and privacy:

> Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield. No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything." And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia." It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga. . . . Ms. Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months' worth of them. "My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder." . . ."We all have a right to privacy," she said. "Nobody should have found this all out."

■ *Ori Heffetz is Assistant Professor of Economics, Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York. Katrina Ligett is Assistant Professor of Computer Science and Economics, California Institute of Technology, Pasadena, California. Their email addresses are oh33@cornell.edu and katrina@caltech.edu.*

http://dx.doi.org/10.1257/jep.28.2.75 doi=10.1257/jep.28.2.75

Privacy 隱私 （私隱）

- Space, Control, and Self-actualization
  - rights to be let alone
  - limited access to (and use of) information
  - “a fundamental aspect of personhood”
- (Not quite) secrecy and confidentiality
- Technologies and the concepts of privacy
  - meanings of space and communication
  - representations of the selves
  - “big data”

Anonymity and Consent

- Anonymity
  - disassociation of information about/to a self
- De-identification
  - Re-identification
- Profiling (about persons and groups)
- Modeling and predication

- Consent

Massachusetts (1997)

- William Weld, governor of Massachusetts, released de-identified medical (insurance) records of state employees to the public
  - without name, address, social security number
- Latanya Sweeney, a grad student at MIT, re-identified Weld's personal records in the dataset
  - with ZIP code, birth date, and sex

AOL (2006)

- AOL released “de-identified” search records of 650,000 users over a three-month period
  - 20 million search queries linked by unique identifiers (instead of usernames and IP addresses)
- NY Times found out user no. 4417749 is Thelma Arnold, a 62-year-old widow in Lilburn, Ga.
  - *There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”*

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

http://www.nytimes.com/2006/08/09/technology/09aol.html

Netflix (2006)

- Netflix Prize: To improve predictions about user ratings of films, using only past user ratings.
- Netflix released a “de-identified” database with100 million ratings of 17,770 films by about 500,000 subscribers covering a six-year period.
  - movie title, rating date, and a five-point rating
- *Used ratings from the Internet Movie Database (IMDB), which are publicly available and are linked to the raters' identities, a (Netflix) user could be identified knowing as little as that user's approximate viewing dates and ratings of two or three movies.*