Editorial

Open Access

Big Data Privacy

Raymond Chi-Wing Wong*

Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China

Many companies store a lot of massive data and typically, the data itself contains a lot of non-trivial but useful information. Data mining techniques can be used to discover this information which can help the companies for decision-making. However, in real life applications, data is massive and is stored over distributed sites. One of my major research topics is to protect privacy over this kind of data.

Protecting privacy over massive data stored in distributed sites is very important in our daily lives because there can be serious consequences if the data owner releases data without considering the protection of sensitive information. Some real events that demonstrate the problem are the following. The first event is that in 2002, Sweeney [1] identified the insufficient protection of a medical dataset. In a real medical data, about 87% of individuals can be uniquely identified by matching certain attributes with a publicly available external table such as a voter registration list by a simple mapping operation. The second event is in 2006; AOL did not take sufficient precautions and encountered some undesirable consequences. A dataset including search logs was published in 2006. Later, AOL realized that a single 62 year old woman living in Lilburn (in Geogia) can be identified from the search logs by New York Times reporters using her several individual-specific queries (e.g., finding webpages with her last name and finding landscapers in Lilburn). The search logs were withdrawn and two employees responsible for releasing the search logs were fired [2]. The third event comes from Netflix. Netflix is a popular online movie rental service with a recommender system, called Cinematic that recommends movies to its customers based on their predicted movie preferences. Netfix released its data to the public on October 2, 2006 for a challenging competition called Netflix Prize with the aim of improving the prediction accuracy of the recommender system. However Narayanan [3], found that 96% of the subscribers could be uniquely identified by limited knowledge of at most 8 movie ratings with their corresponding rating dates. The fourth event is that the research project conducted by Beinat shows that 24% of the mobile clients using location-based services (LBS) have serious privacy concerns about disclosing their locations together with their personal information.

Privacy protection becomes more and more important when the data becomes massive and is stored in a lot of different sites. In many applications, data comes at a rapid rate and the data is now stored in gigabytes and terabytes. For example, in United States, the American supermarket chain Wal-Mart keeps about 20 million sales transactions per day. In Hong Kong, the Octopus system has over 7 million transactions per day in 2003. The growth of this massive data gives more chances for individual information disclosure because there are more records about an individual in the data. Furthermore, since the data is large and is stored in the cloud computing manner, the data is stored in different sites. For example, in United States, the famous supermarket chain Wal-Mart has over 2000 retail stores in different regions and the data over the distributed sites is collected in a centralized site running with at least 483 processors. In Hong Kong, the Octopus system involves over 100 service providers. In these scenarios, it is more likely that an adversary can breach individual privacy since the adversary can have more sites for breaching individual privacy. Thus, privacy issues become much more serious when the data is massive and is stored in different sites.

There are a lot of important privacy issues in the context of "big data".

Issue 1: How to protect individual privacy when the data is stored in a single site

Issue 2: How to protect individual privacy when the data is stored in multiple sites

Issue 3: How to protect individual privacy efficiently

Issue 4: How to protect individual privacy when data changes over time

References

1. Sweeny L (2002) k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge based Systems 10: 557-570.
2. Barbaro M, Zeller T(2006) A Face is Exposed for AOL Searcher No. 4417749, New York Times, USA.
3. Narayanan A, Shmatikov V (2006) How to Break Anonymity of the Netflix Prize Dataset. The University of Texas at Austin.

***Corresponding author:** Raymond Chi-Wing Wong, Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China, E-mail: raywong@cse.ust.hk

Received November 28, 2012; **Accepted** December 03, 2012; **Published** December 05, 2012

Citation: Wong RC (2012) Big Data Privacy. J Inform Tech Softw Eng 2:e114. doi:10.4172/2165-7866.1000e114