

Data Privacy: the Non-interactive Setting

Arvind Narayanan

May 19, 2009

1 Introduction

The ease of large-scale data collection. Before the growth of the Internet, collection of data from individuals on a national or global scale was feasible only for governments and very large corporations. The infrastructure required for collecting and aggregating data was either in the form of a door-to-door survey as in a census, or a pervasive physical presence, such as a large supermarket chain collecting data on people's shopping habits. Clearly, the Internet has changed the equation dramatically: one might grasp the magnitude of the change by considering that the 2010 U.S. Census, which does not take advantage of information technology, is estimated to cost 14 billion USD, whereas a purely electronic survey of that scale, although hardly as rigorous, would cost orders of magnitude less.

Data sharing. The other emerging trend is sharing of collected data. There are several ways in which data collected about people is shared:

- Internet-scale data warehousing for tasks such as Customer Relationship Management is a specialized capability and is often outsourced, occasionally to offshore providers. In addition, lower costs and higher reliability are cited as reasons for outsourcing.
- Highly targeted marketing and advertising are an important part of the business model of many Web 2.0 businesses. In fact, customer data allowing targeted advertising is considered the primary monetizable asset of companies such as Facebook. It is often necessary to share detailed customer data with advertisers for this purpose. Several companies such as Phorm, NebuAd and Front Porch are developing advertising systems based on behavioral targeting, which involve building detailed profiles of user behavior [26].
- Companies also make data available for exploratory research purposes. Such research is

usually done in-house in corporations with dedicated research departments, but with the increasing ease of data collection by smaller companies, academia is often recruited for this purpose. Recent data releases by AOL [13] and Netflix [12] come to mind.

- Allowing users to search for and access information about other individuals is frequently an essential part of the functionality for which the data was collected. Social-network services are the most visible example of such functionality.
- More complex sharing situations arise when there is more than one party who owns a part of a dataset and joint analyses or computations need to be performed on the data. A joint medical study between two hospitals might involve sensitive patient data that needs to be aggregated to enable analysis. A network-security response center might aggregate audit log data from dozens of institutions.

Finally, a truly distributed computation involves thousands or millions of users participating. Such systems are increasingly coming into fruition in the form of cloud computing, raising serious and largely unexplored privacy questions [14].

The factors described above have led to an explosion in the amount of personal data collected. This has given rise to complex privacy questions related to leakage of sensitive data on individuals and the mass harvesting of personal information by unscrupulous parties. Before we can discuss the privacy questions, however, it helps to have an understanding of the nature of the data being collected and the data-release processes.

2 Understanding data

2.1 Types of data

While all databases can be viewed simply in terms of rows and columns, it is useful to make further distinctions based on the semantics of the data.

Micro-data vs. aggregate data. Micro-data refers to data about individuals while aggregate data is information about a group of users or about the database as a whole. The terms originate from census terminology. Data when collected is always micro-data, whereas when sharing a database, one might choose not to make micro-data available. It is increasingly common to share micro-data because of the increasing complexity of the data being collected and the algorithmic complexity of the analyses that need to be performed. When micro-data is shared, it is often in an anonymized form.

High-dimensional data. The rows or records in a database typically represent individuals, and the columns represent attributes. By dimensionality we mean the number of columns or attributes in the database. The terminology comes from viewing a record as a point in a vector-space. Let us empirically define high-dimensional databases as those for which algorithms that are exponential in the dimension are infeasible. While these could theoretically include any type of data, databases in current practice that we consider high-dimensional are those containing *transaction profiles* of users, which are vectors consisting of their preferences, behavior, purchase or transactional history for a variety of items of the same type.

Graph-structured data. Graph-structured data is derived from real-world networks of individuals and consists of nodes and edges, possibly with attributes corresponding to each node and edge. Since edges and edge-attributes cannot be associated with any single individual, a different abstraction becomes necessary.

Sensitive attributes. The semantics of the attributes sometimes make it meaningful to distinguish between sensitive and non-sensitive attributes. For instance, name and gender may not be sensitive but attributes pertaining to medical history might be. A similar notion is that of access vs. data attributes, for instance name vs. phone number in the context of a telephone directory. The statistical-database literature uses the notion of quasi-identifiers [6], which are attributes that are not structurally unique but potentially empirically unique, either by themselves or in

combination with other quasi-identifiers, for instance ZIP code together with birth date in the context of a census database.

2.2 Release process

Interactive vs. non-interactive sharing. We can identify four categories of data sharing in the decreasing order of interactivity of the process. Traditionally, release of sensitive data involved manual auditing of each data access request to ensure compliance with privacy policies. For a recent example of HIPAA compliance guidelines involving manual auditing, see [20]. As privacy-preserving data mining techniques have matured—for example, the SuLQ framework [4] enables sophisticated query control and perturbation of query outputs—it has become possible to automate this process. This type of data sharing has not yet seen widespread adoption, as we will explain presently.

Moving further down on the scale of interactivity, the data collector might implement a lightweight query interface to the data, without additional safeguards such as output perturbation. Essentially, the data can be retrieved by crawling such a system. This allows some query control, but when such controls are not implemented, it can be viewed as equivalent to the final scenario, which is simply publishing the data, perhaps after “sanitization.”

The appeal of non-interactivity. Non-interactive data release has been pursued aggressively in recent years because it avoids the costs and delays of implementing manual privacy safeguards. Compared to automated but interactive data sharing, non-interactive data release avoids the need to set up a reliable, high-performance, low-latency infrastructure for performing computations on databases. The added overhead of the privacy requirement means that interactive data release is often infeasible given current technological constraints.

In addition, rationing resources such as queries, computation and memory when there are multiple third parties interested in data analysis might very well prove insoluble in an interactive setting. Data mining often has a competitive aspect, even if it may not always be explicit. Genetic association studies are a good example [11].

Further, most work on interactive data mining does not address the requirement of privacy for the *client*, which is also important given the competitiveness of data mining. Thus, choosing an interactive setting

might discourage interested third parties from participating.

2.3 Utility of data

Finally, we classify datasets based on the purpose for which they are shared. There are three categories that we will encounter: directory, data mining and statistical utility. This is not meant to be an exhaustive classification.

I use the term directory database or electronic directory to refer to databases where the utility comes from the information associated with individual users (*e.g.*, a college alumni directory).

In a statistical database, the utility comes from learning statistical relationships (*e.g.*, census data). Typical uses are computing marginals, joint distributions and cross-attribute correlations. By data mining we mean various computations like clustering, classifiers, and collaborative filtering. The distinction between the latter two categories is fuzzy. In both of these categories, however, the utility comes from aggregate information and not individual data.

3 Protecting privacy

The nature of the privacy requirement depends on the intended utility of the data. In databases with an aggregate utility, the privacy constraint is that information associated with individuals should not be revealed. Directory databases, on the other hand, exist to make individual data available. However, one would like mass harvesting to be infeasible.

I will now describe five broad strategies for protecting privacy. Note that these are not mutually exclusive.

Access control. Some form of access control is necessary for most datasets. The most common form of coarse-grained access control is passwords. Password-based authentication is often sufficient protection for data when the sharing requirements are not very complex. More fine-grained access control based on roles and privileges is an active research area and is available in commercial relational database systems [22, 21].

Query control can be a powerful tool for privacy protection in the interactive setting. It involves query filtering, such as allowing only SUM, COUNT and other aggregate queries; query logging and monitoring; and finally query auditing [2, 16]. Automated filtering of queries to ensure that responses do not

leak sensitive information is hard; for instance, even denials might violate privacy. Technical solutions include simulatable auditing [15], but in many situations, occasional review of audit logs combined with authentication and the threat of punitive measures such as the revocation of query privileges might be sufficient to enforce self-policing of queries.

Perturbation-based techniques have a long history and include tools such as generalization, suppression, cell swapping and addition of noise. (However, the term perturbation is used in the literature to refer only to the last of these.)

Secure multi-party computation is a set of general cryptographic techniques that allow a set of players to compute joint functions of their inputs while leaking no information other than their respective outputs [28, 10]. It has been adopted to the problem of privacy-preserving data mining with some success [19, 18].

Anonymity is increasingly being used as a tool for protecting privacy in databases with aggregate utility in the non-interactive setting [24]. The rationale is that users are “de-identified”, *i.e.*, identifying information about users is removed, then privacy is protected as long as the adversary cannot re-link the identity of an individual with their record in the database.

There are two reasons why anonymity has been especially popular as a privacy protection technique. First, de-identification can be implemented easily and cheaply, as opposed to alternatives such as secure multi-party computation. Secondly, anonymized data release is often motivated by privacy laws. Even though such laws might not explicitly require anonymity as a precondition of release of sensitive data, they are often interpreted as such.

The use of de-identification as a “catch-all” privacy-protection mechanism, especially in the non-interactive data release scenario, is very tempting because it avoids the need to reason about the types of computations that users should (or should not) be able to perform on the released data. However, this thinking is flawed: privacy can only be meaningfully defined as a property of specific computation — as differential privacy does [7] — and not as a property of the data itself.

I believe that the dual trends of increasing non-interactivity in data sharing and the increasing reliance on (syntactically defined) anonymity for privacy protection have serious implications for privacy in data sharing.

4 Applications

In this section we describe common applications where privacy is a concern in sharing data.

Electronic directory. Electronic directory databases have a precursor in the form of telephone directories, with a long history going back to 1878 [5]. Interestingly, *reverse telephone directories* have been compiled for decades, although their availability to the general public has sometimes raised legal questions because of privacy issues.

Electronic directories potentially allow finer-grained functionality and privacy control. The operator may choose to make the existence of individuals in the database hidden unless there is enough information to identify them; access might be permitted based on certain fields and not on others; and mass harvesting may or may not be allowed. In fact, it is usually preferable to make mass harvesting infeasible. In the context of email, this concern is well known as the *directory harvest attack*; however, it is an issue in any electronic-directory setting.

This flexibility leads to some interesting and complex privacy issues. For instance, Facebook recently faced a privacy gaffe where users who chose to make the values of some attributes (such as religion and sexual orientation) hidden found their privacy violated when their profiles showed up in searches based on specific values for those attributes [23].

Collaborative filtering. The term collaborative filtering refers to algorithms for predicting future user behavior by analyzing transactional profiles of a large number of users in conjunction. It is used primarily in online recommendation systems. Collaborative filtering algorithms operate, in an abstract sense, on a matrix (database) that contains a score for each (user, item) pair.

An alternate output of collaborative filtering algorithms is item-similarity information, *i.e.*, numerical similarity scores between each pair of items. This is useful, for instance, in deciding what items to put next to each other in a supermarket, or related items to show on a web page pertaining to the item currently being viewed

At an intuitive level, users do not like their purchase history being revealed, and yet would like to be able to feed this history into some system that is capable of making useful predictions. A formal framework for defining this notion is necessary. Purchase histories are usually protected by the privacy policy, and furthermore, there are often strong legal protections in place such as the Video Privacy Protection

Act [8].

Social networks. Unique privacy challenges arise with sharing social-network data because the data is non-relational. Therefore, it is perhaps unsurprising that there currently exists no comprehensive framework for analyzing privacy and anonymity in social networks.

Online social-network services, which are a new development, face even more challenges because the data has both directory and aggregate purposes: users and application developers need access to the individual information of other users. On the other hand, researchers and advertisers need access to aggregate information. For instance, Facebook alone has faced storms of privacy-related criticism of the way it shares data with users [25], applications [9], and advertisers [27].

Telephone-call graphs arguably contain much more sensitive information. The AT&T call graph, for instance, contains 1.9 trillion edges going back decades.¹ Law enforcement regularly mines information from such graphs. More worryingly, anonymized versions of call graphs are often published or shared for research purposes [1, 17].

The release of social-network data in anonymized form is hardly limited to the above situations, however. To give just one example, sexual relations between students in a number of high schools (totalling about 90,000 participants) have been collected as part of the Add Health dataset; the data from each school (often involving a thousand students) can be viewed as a social network. Such graphs are often published with identifying information removed [3].

References

- [1] J. Abello, P. Pardalos, and M. G. C. Resende. On maximum clique problems in very large graphs. In J. Abello and J. Vitter, editors, *External memory algorithms and visualization*, volume 50 of *DMACS Series on Discrete Mathematics and Theoretical Computer Science*, pages 119–130. American Mathematical Society, 2001.
- [2] N. Adam and J. Worthmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.
- [3] P. S. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110(1):44–91, 2004.

¹An edge in this context represents a single call.

- [4] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The SuLQ framework. In *Proc. 24th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 128–138. ACM, 2005.
- [5] The Connecticut District Telephone Company. The telephone directory. Reproduced at http://www.christies.com/LotFinder/lot_details.aspx?intObjectID=5084352, 1878.
- [6] T. Dalenius. Finding a needle in a haystack—or identifying anonymous census record. *Journal of Official Statistics*, 3:329–336, 1986.
- [7] C. Dwork. Differential privacy: A survey of results. *Theory and Applications of Models of Computation—TAMC*, 4978:1–19, 2008.
- [8] Electronic Privacy Information Center. The Video Privacy Protection Act (VPPA). <http://epic.org/privacy/vppa/>.
- [9] A. Felt and D. Evans. Privacy protection for social networking APIs. In *Proc. Web 2.0 Security & Privacy (W2SP)*, 2008.
- [10] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game. In *Proc. 19th ACM Symposium on Theory of Computing (STOC)*, pages 218–229. ACM, 1987.
- [11] The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies. <http://www.genome.gov/Pages/About/OD/OPG/NewModel-Gain.pdf>, 2007.
- [12] K. Hafner. And if you liked the movie, a Netflix contest may reward you handsomely. *New York Times*, 2006.
- [13] S. Hansell. AOL removes search data on vast group of web users. *New York Times*, 2006.
- [14] S. Hansell. Does cloud computing mean more risks to privacy? <http://bits.blogs.nytimes.com/2009/02/23/does-cloud-computing-mean-more-risks-to-privacy/>, 2009.
- [15] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing. In *Proc. 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 118–127. ACM, 2005.
- [16] J. Kleinberg, C. Papadimitriou, and P. Raghavan. Auditing boolean attributes. In *Proc. 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 86–91. ACM, 2000.
- [17] M. Kurucz, A. Benczur, K. Csalogany, and L. Lukacs. Spectral clustering in telephone call graphs. In *Proc. 9th WebKDD and First SNA-KDD Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD)*, pages 82–91. ACM, 2007.
- [18] Y. Lindell and B. Pinkas. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality (to appear)*.
- [19] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Proc. Advances in Cryptology - CRYPTO*, volume 1880 of *LNCS*, pages 36–54. Springer, 2000.
- [20] Office of Compliance and SUNY Downstate Medical Center Audit Services. HIPAA audit physical rounds checklist. <http://www.downstate.edu/hipaa/audit.html>, 2003.
- [21] Oracle Technical White Paper. The virtual private database in oracle9ir2. <http://www.oracle.com/technology/deploy/security/oracle9ir2/pdf/VPD9ir2twp.pdf>, 2002.
- [22] S. Rizvi, A. Mendelzon, S. Sudarshan, and P. Roy. Extending query rewriting techniques for fine-grained access control. In *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 551–562. ACM, 2004.
- [23] R. Singel. Private Facebook pages are not so private. <http://www.wired.com/software/webservices/news/2007/06/facebookprivacysearch>, 2007.
- [24] L. Sweeney. k-anonymity: A model for protecting privacy. *International J. of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.
- [25] Techcrunch. Facebook news feed reports on you behind your back. <http://www.techcrunch.com/2008/04/14/facebook-newsfeed-reports-on-you-behind-your-back/>, 2007.
- [26] Wikipedia. Phorm — Wikipedia, the free encyclopedia, 2008. [Online; accessed April-2008].
- [27] PC World. Facebook’s beacon more intrusive than previously thought. <http://www.pcworld.com/article/id,140182-c,onlineprivacy/article.html>, 2007.
- [28] A. Yao. Protocols for secure computations. In *Proc. 23rd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 160–164. IEEE, 1982.