Big Privacy: Protecting Confidentiality in Big Data

A tremendous amount of data about individuals – e.g., demographic information, internet activity, energy usage, communication patterns and social interactions – are being collected and analyzed by many national statistical agencies, survey organizations, medical centers, and Web and social networking companies. Wide dissemination of microdata (data at the granularity of individuals) facilitates advances in science and public policy, helps citizens to learn about their societies, and enables students to develop skills at data analysis. Often, however, data producers cannot release microdata as collected, because doing so could reveal data subjects' identities or values of sensitive attributes. Failing to protect confidentiality (when promised) is unethical and can cause harm to data subjects and the data provider. It even may be illegal, especially in government and research settings. For example, if one reveals confidential data covered by the U. S. Confidential Information Protection and Statistical Efficiency Act, one is subject to a maximum of $250,000 in fines and a five year prison term.

At first glance, sharing safe microdata seems a straightforward task: simply strip unique identifiers like names, addresses, and tax identification numbers before releasing the data. However, anonymizing actions alone may not suffice when other readily available variables, such as aggregated geographic or demographic data, remain on the file. These quasi-identifiers can be used to match units in the released data to other databases. For example, computer scientist Latanya Sweeney showed as part of her PhD thesis at MIT that 97% of the records in publicly available voter registration lists for Cambridge, MA, could be uniquely identified using birth date and nine digit zip code. By matching on the information in these lists, she was able to identify Governor William Weld in an anonymized medical database. More recently, the company Netflix released supposedly de-identified data describing more than 480,000 customers' movie viewing habits; however, computer scientists Arvind Narayanan and Vitaly Shmatikov were able to identify several customers by linking to an on-line movie ratings website, thereby uncovering apparent political preferences and other potentially sensitive information. Probably the most sensational privacy breach occurred in 2006 when America Online (AOL) released 20 million search queries posed by users over a three month period in order to facilitate research on information retrieval. They knew the information in web-searches contained potentially identifying and sensitive information (including social security and credit card numbers!), and hence attempted to anonymize the data by replacing user identifiers with random numbers. However, within a couple of hours of releasing the anonymized data, two reporters from New York Times were able to uncover the identity of user No. 4417749 based on just her search history, e.g., "landscapers in Lilburn, Ga," several people with last name Arnold, and "numb fingers." This breach had far reaching consequences: several high ranking officials at AOL were fired, and search companies are now reluctant to release search logs and other personal information. In fact, even researchers are wary of using the now publicly available AOL data for research, since it reveals too much about individuals. Similar re-identification is possible from social network data, location traces, and even from power usage patterns.

Although these re-identification exercises were done to illustrate concerns over privacy, one easily can conceive of re-identifications attacks for nefarious purposes, especially for large databases on individuals. A nosy neighbor or family relative might search through a public database in an attempt to learn sensitive information about someone who they knew participated in a survey or administrative database. A journalist might try to identify politicians or celebrities. Marketers or creditors might mine large databases to identify good, or poor, potential customers. And, disgruntled hackers might try to discredit organizations by identifying individuals in public use data.

The threat of breaches, whether perceived or imminent, has serious implications for the practice and

scope of data sharing, especially with the availability of massive and richly detailed data. These threats have created a fascinating area of research for aspiring computer scientists, mathematical and statistical scientists, and social scientists. This area goes by names like privacy preserving methods (computer science) and statistical disclosure limitation (statistical science). It is an area where the research challenges are grand and interdisciplinary, the opportunities for high profile publications and external funding are strong, and the potential to impact the practice of data sharing is genuine and significant.

In this article, we describe some general themes in research in this area, with the aim of pointing out opportunities for students. Keeping with its interdisciplinary nature, we present perspectives from both computer science and statistical science, which are our two home departments. We begin by describing research into how to define and measure the risks of confidentiality breaches. We then describe some approaches to data protection. We end with a few general areas where students can engage in research.

We note that there are many more topics in big privacy that we do not cover for lack of space. These include, for example, systems for collecting data privately, access control in web and social networking applications, data security and cryptography, and protocols for secure computation. These are equally rich and complementary areas for research that are important for secure and confidential use of big data.

Defining and Measuring Confidentiality Risks

Both the computer science and statistical science communities have developed a variety of criteria and methods for quantifying confidentiality risks. Indeed, a major thrust of research funded by the US National Science Foundation (including grants to us) is to integrate these two perspectives, taking the best of what both have to offer. In reviewing some of the risk metrics, we do not attempt to cover all approaches. Rather, we cover a few important ones that we are most familiar with.

In statistical science, measures used in practice tend to be informal and heuristic in nature. For example, a common risk heuristic for publishing tabular magnitude data for business establishments (e.g., tables of total payroll within employee size groupings) is that no one establishment should contribute in excess of p% of the cell total, and no cell should comprise less than 3 establishments. Cells that do not meet these criteria are either suppressed or perturbed. The most general and mathematically formal method of disclosure risk assessment is based on Bayesian probabilities of re-identification, by which we means posterior probabilities that intruders could learn information about data subjects given the released data and a set of assumptions about the intruder's knowledge and behavior. Agencies can compute these measures across a variety of intruder knowledge scenarios as a way of identifying particularly risky records and making an informed decision about data release policy in the face of uncertainty (the goal of statistical science in general). Computing these probabilities in practice is computationally demanding and requires innovative methodology, especially for big data.

In computer science, some of the early efforts to quantify confidentiality risk were targeted to thwart re-identification attacks (which we described in the introduction) by ensuring that no individual's record is unique in the data. This motivated a popular notion of privacy called K-Anonymity, which required that microdata be released in a manner that no individual's record is distinguishable from at least K-1 other records. While this seemingly avoids the privacy breaches discussed in the introduction, it has two drawbacks. An adversary (especially one with prior knowledge) can learn sensitive information. For instance, suppose a hospital releases K-anonymous microdata about patients, and you know your neighbor Bob is in the data. If individuals in the anonymous group containing Bob all have either cancer or the flu, and you know for a fact that Bob does not have the flu, then you can deduce that Bob has

cancer. K-Anonymity has been extended in a number of ways to handle this shortcoming. One example is L-Diversity, which requires that each group of individuals who are indistinguishable via quasi-identifiers (like age, gender, zip code, etc.) not share the same value for the sensitive attribute (like disease), but rather has L distinct well represented (of roughly same proportion) values.

The current state of the art disclosure metric is called differential privacy. It eliminates (to a large extent) the confidentiality issues in K-anonymity, L-diversity and their extensions. Differential privacy can be best explained using the following opt-in/opt-out analogy. Suppose an agency (e.g., the Census Bureau or a search engine) wants to release microdata. Any individual has two options: opt-out of the microdata so that their privacy is protected, or opt-in and hope that an informed attacker can't infer sensitive information using the released microdata. A mechanism for microdata release is said to guarantee ε-differential privacy if for every pair of inputs D1 and D2 that differ in one individual's record (e.g., D1 contains record t and D2 does not contain t), and every microdata release M, the probability that the mechanism outputs M with input D1 should be close to (within an exp(ε) factor of) the probability that the mechanism outputs M with input D2. In this way, the release mechanism is insensitive to a single individual's presence (opt-in) or absence (opt-out) in the data. Thus, differential privacy represents a strong guarantee.

Moreover, differential privacy satisfies an important property called composability -- if M1 and M2 are two mechanisms that satisfy differential privacy with parameters ε1 and ε2, then releasing the outputs of M1 and M2 together also satisfies differential privacy with parameter ε1+ε2. Other known privacy conditions (e.g. k-anonymity and l-diversity) do not satisfy composability, and hence, two privacy preserving releases using these definitions can result in a privacy breach.

Methods for Protecting Public Release Data

Like risk measures, both computer scientists and statistical scientists have developed methods of altering or perturbing data before release. Indeed, sometimes very similar methods are developed independently in both communities! Apart from whether a privacy protection method results in low disclosure (according to one of the metrics described in the previous section), there are two important considerations when designing a privacy protection method. First, the method must result in outputs that retains useful information about the input. Note that every privacy protection must result in some loss in utility (after all we are trying to hide individual specific properties). Hence, it is usually a good idea to list the types of statistical analyses that must be run on the data, and then tuning/optimizing the output to best answer those analyses. Some techniques assume an interactive setting, where a data analyst queries the datasets, and perturbed results are returned for these queries. Second, a privacy protection method should be simulatable -- an attacker must be assumed to know the privacy protection method. For instance, a method that reports the age of an individual (x) as [x-10, x+10] is not simulatable, since an attacker who knows this algorithm can deduce the age of the individual to be x. We next present a few important types of privacy protection methods.

Aggregation. Aggregation reduces disclosure risks by turning atypical records---which generally are most at risk---into typical records. For example, there may be only one person with a particular combination of demographic characteristics in a city, but many people with those characteristics in a state. Releasing data for this person with geography at the city level might have a high disclosure risk, whereas releasing the data at the state level might not. Aggregation is very similar to K-anonymity. Unfortunately, aggregation makes analysis at finer levels

difficult and often impossible, and it creates problems of ecological inferences (relationships seen at aggregated levels do not apply at disaggregated levels). There is a significant literature on aggregation techniques that satisfy k-anonymity, l-diversity (and variants), as well as differential privacy.

Suppression. Agencies can delete sensitive values from the released data. They might suppress entire variables or just at-risk data values. Suppression of particular data values generally creates data that are missing because of their actual values, which are difficult to analyze properly. For example, if incomes are deleted because they are large, estimates of the income distribution based on the released data will be biased too low.

Data swapping. Agencies can swap data values for selected records---for example, switch values of age, race, and sex for at-risk records with those for other records---to discourage users from matching, since matches may be based on incorrect data. Swapping is used extensively by government agencies. It is generally presumed that swapping fractions are low -- agencies do not reveal the rates to the public (and hence, these algorithms are not simulatable) -- because swapping at high levels destroys relationships involving the swapped and unswapped variables.

Adding random noise. Agencies can protect numerical data by adding some randomly selected amount – e.g., a random draw from a normal distribution with mean equal to zero – either to the observed values or to answers to statistical queries. Adding noise to values can reduce the possibilities of accurate matching on the perturbed data, and distort the values of sensitive variables. The degree of confidentiality protection depends on the nature of the noise distribution; e.g., using a large variance provides greater protection. However, adding noise with large variance introduces measurement error that stretches marginal distributions and attenuates regression coefficients. Adding noise from a heavy tailed distribution (like a Laplace distribution) to query answers can provide strong privacy guarantees like differential privacy.

Synthetic data. The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions. These distributions are specified to reproduce as many of the relationships in the original data as possible. Synthetic data approaches come in two flavors: partial and full synthesis. Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values. For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables.

Fully synthetic data comprise an entirely simulated data set; the originally sampled units are not on the file. Typical algorithms either create a detailed model of the data, or transform the data into a different space (e.g., Fourier). The alternate representation is then perturbed, from which synthetic data is sampled. In fact, one of the US Census data products releases statistics about commute patterns of individuals using a synthetic data generation technique that provably guarantees a variant of differential privacy.

Research Challenges

While recent research has shed much light on formal disclosure metrics and new provably private methods that provide useful statistical information, there are still many intriguing research challenges in this area that students can be involved in. For instance, most work on privacy has considered data where each record corresponds to a unique individual, and the different records are typically considered independent. One important problem is ensuring privacy of linked data, e.g. social networks, where people are linked to other people, and relational data, where different types of entities maybe linked to one another. Reasoning about privacy in such data is tricky since information about an individual may be leaked through links to other individuals. Another interesting problem is that of releasing sequential releases of the same data over time. Attackers may link individuals across releases and infer additional sensitive information that they could not have from a single release. Finally, as the data we deal with become extremely high dimensional, we need to develop techniques that can protect privacy while guaranteeing utility. Understanding theoretical trade-offs between privacy and utility is an important open area for research.

To close then, we encourage students to learn more about opportunities in this research area. We would be delighted to point you to places to learn more; just drop one of us an email expressing your interests. And, of course, we pledge to keep your email confidential!