Microsoft

Differential Privacy for Everyone

Differential Privacy for Everyone

Over the past few years, much has been written about the privacy risks inherent in data de-identification.[1] *Differential Privacy is a technology that enables researchers and analysts to extract useful answers from databases containing personal information and, at the same time, offers strong individual privacy protections. This seemingly contradictory outcome is achieved by introducing relatively small inaccuracies in the answers provided by the system. These inaccuracies are large enough that they protect privacy, but small enough that the answers provided to analysts and researchers are still useful. This whitepaper provides a non-technical description of how Differential Privacy works.*

The vast amounts of information that are collected today, coupled with innovative technologies used to analyze these "big data"[2] stores, offer tremendous promise to researchers, businesses, and society. Big data resources could help solve difficult problems in health care, economics, and other areas. They could also be used to provide new and innovative technology-based services. Unfortunately, because much of this data contains personal information, legitimate concerns about privacy can prevent many researchers and commercial organizations from accessing potentially valuable stores of data. Often, techniques used to protect privacy in these contexts, have proven to be insufficient.

A good example of a failed attempt to protect privacy occurred in the mid-90s, when the Commonwealth of Massachusetts Group Insurance Commission (GIC) released what at the time were believed to be anonymous[3] health records. They wanted to encourage research to benefit society while also protecting the identity of the people whose information was being released. The GIC took specific steps to protect privacy, such as suppressing street addresses and replacing people's names with random numbers. Latanya Sweeney (then a PhD student at MIT, later a professor at Carnegie Mellon University and Harvard) wanted to show how people's privacy could be compromised despite such precautions. For $20, she purchased a CD with the publicly available voter registration database for the city of Cambridge, Massachusetts. Next, by simply comparing and correlating the voter registration data with the GIC data, she was able to re-identify, among others, those health records in the GIC publication that belonged to the then governor of the Commonwealth of Massachusetts, William Weld. A few years later, she published a paper[4] in which she concluded that up to 87% of individuals living in the U.S. can be uniquely identified by using the same 3 data features she used to identify the governor's records in the GIC data: birth date, zip code, and gender.

---

[1] For an example see: Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," 57 UCLA Law Review 1701 (2010). http://paulohm.com/

[2] "Big Data" is a loosely defined term used in the Information Technology industry to designate a combination of extremely large databases, and the software and hardware necessary to manage, organize and process them. Big data tools and techniques are necessary because traditional database technologies that have for decades been deployed by governments, companies and other organizations become inefficient or useless when dealing with these huge bodies of data.

[3] The use of the word "anonymous" in this context is incorrect. If the published information had been really anonymous, it would not have been possible to single out the governor's records in the way that is described later in this paragraph. The term that should be used in cases like this is "de-identified," which means that the publisher has made an effort to remove information that connects the published data to the real identity of the individuals that data refers to.

[4] Latanya Sweeney, "Foundations of Privacy Protection from a Computer Science Perspective," Proceedings, Joint Statistical Meeting, AAAS, Indianapolis, IN. 2000 http://dataprivacylab.org/projects/disclosurecontrol/index.html

This problem is not new. The statistical, medical, and computer science research communities, as well as institutions like the U.S. Census Bureau, have long known that the alleged anonymity of individuals in data sets could be compromised using other publicly available information. Simply removing data that may uniquely identify individuals – like name, address, telephone number, and social security number – is often not enough to definitively conceal the identity of those whose information is in a database.

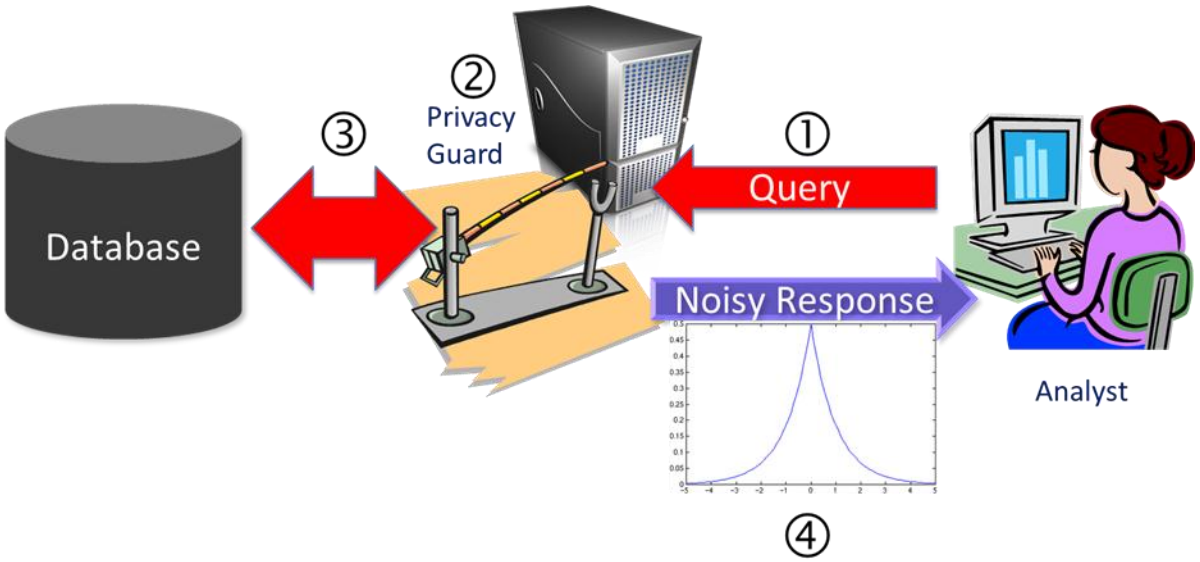
As the GIC example shows, attributes such as ZIP code, gender, and birth date may not by themselves immediately reveal an individual's identity, say, name or address, but when combined and correlated with other, possibly publicly available data like voter registration lists or property tax records, can make it possible to identify a significant number of individuals within a large population of allegedly anonymized data. For years, scientists have worked on developing techniques that will make it harder for people or organizations to do this re-identification of individuals in databases and have concluded that in many cases it is necessary to modify the data in order to make such goals difficult to achieve.[5] Oftentimes such modifications diminish the value that can be extracted from the database.

How Technology May Address the Problem

"Differential Privacy" (DP) was conceived to deal with privacy threats in this context. That is, to prevent unwanted re-identification and other privacy threats to individuals whose personal information is present in large datasets, while providing useful access to data. Under the DP model, personal information in a large database is not modified and released for analysts to use. Thus, in our example above for instance, Professor Sweeney would not have been given access to the GIC health records, even in their de-identified form. Instead, DP permits analysts to pose questions to the database by going through an intermediary piece of software, and to obtain answers with a minimum amount of distortion while acquiring essentially no information about any particular individual in the database, or determining if an individual is even in the database. This holds even if the data is correlated with information coming from other databases, as was the case with the GIC example.

Roughly speaking, DP works by inserting an intermediary piece of software between the analyst and the database. The analyst never gets to access or actually see the contents of the database; instead the intermediary acts as a privacy-protecting screen or filter, effectively serving as a privacy guard. The guard takes the questions (queries) that the analyst wishes to ask of the database and evaluates the combined privacy implications of that question and those that have preceded it. This evaluation depends only on the sequence of the queries, not on the actual data in the database. Once the guard establishes the privacy risk of the question, it then gets the answer from the database, and changes it to be slightly imprecise (we say that it injects a certain amount of "distortion" into the response, and the amount of distortion is calibrated to the privacy risk), before sending it back to the analyst. When the privacy risk is low, we can think of this distortion as inaccuracies that are small enough that they do not affect the quality of the answers significantly, but large enough that they protect the identities of individuals in the database. If, however, answering the question with relative accuracy opens up the possibility that somebody's privacy will be breached, then the guard will increase the amount of distortion to a level that may make the answer not useful. The analyst may then ask a more general question, or simply abandon it.

[5] Ibid.



Differential Privacy (DP) in action: ① Analyst sends a query to an intermediate piece of software, the DP guard. ② The guard assesses the privacy impact of the query using a special algorithm. ③ The guard sends the query to the database, and gets back a clean answer based on data that has not been distorted in any way. ④ The guard then adds the appropriate amount of "noise," scaled to the privacy impact, thus making the answer (hopefully slightly) imprecise in order to protect the confidentiality of the individuals whose information is in the database, and sends the modified response back to the analyst.

An Example

Assume that a hospital has a database of patients with a potentially life-threatening disease. The database contains a detailed record of the treatments each patient has undergone, dates and times for appointments, and records of prescriptions. It also contains general information about the patients, including places where they have lived in the past for five or more years. The hospital has deployed a DP guard for this database that keeps an eye out for patient privacy.

A researcher believes that the disease is more likely to manifest itself in people who have lived for long periods of time in certain regions, and wants to find out if the information in the database confirms this hypothesis. The researcher connects to the Differential Privacy guard and requests, for each town located in the suspected regions, the number of patients with the disease. The guard poses the question to the database and, when the answer comes back, it discovers that there are a significant number of individuals that lived in eight towns in the regions the researcher asked about, that there is one additional town, "Smallville," for which the number of patients is one, and a number of other towns for which the number is zero. We will call the single patient from Smallville, Bob, and use Bob's case to illustrate how DP works.

If, for instance, the researcher does other work at the hospital and because of this work has access to other, less detailed, patient records that show Bob recently moved in from Smallville, and she knows the town is located thousands of miles away from the hospital, she may reasonably conclude that Bob has the illness because it is unlikely that two people from such a small town would move such a long distance and end up in the same city and hospital at more or less the same time. Whether the researcher finds this out by accident or through a deliberate analysis is not relevant; the important thing is that Bob's privacy has been breached because he has been identified.

To avoid this situation, the DP guard will introduce a random but small level of inaccuracy, or distortion, into the results it serves to the researcher. Thus, instead of reporting one case for Smallville, the guard may report any number close to one. It could be zero, or ½ (yes, this would be a valid noisy response when using DP), or even -1. The researcher will see this and, knowing the privacy guard is doing its job, she will interpret the result as "Smallville has a very small number of cases, possibly even zero." In fact, and in order to maintain privacy, the guard may also report non-zero (but equally small) numbers for some of the towns that really have zero cases.

If we analyze these results we can conclude that:

- Bob's privacy is preserved because the researcher knows some level of noise is present in the results, thus she is not entirely sure if there is one patient who has spent time in Smallville or no patients at all. Therefore, she doesn't know if Bob has the illness, she only knows Bob has been to the hospital, which she would have known anyway from looking at the less detailed hospital records she has access to.

- From the research point of view, the results provided by the DP guard are still useful. The researcher has learned that:

    - There are a number of towns in the selected regions, including Smallville, that have very, very few and quite possibly no cases of the illness at all.

    - Eight other towns in the selected regions show a significant number of cases. It is important to point out that for these eight towns, the DP guard will have reported numbers that are slightly larger or smaller than the actual number of patients in each town.

Thus, the answers reported by the DP guard are accurate enough that they provide valuable information to the researcher, but inaccurate enough that the researcher cannot know if Bob's name is or is not in the database.

What is so special about Differential Privacy?

There are several aspects to DP that are revolutionary:

- The underlying data need not be modified or distorted in any way. In fact, while not absolutely necessary, it is recommended that it be left in pristine condition so that the answers the system provides are of the highest possible quality, without threatening privacy.[6]

- Distortion is introduced into the answers a posteriori. That is, the DP guard gets answers based on pristine data, and then mathematically decides the right amount of distortion that needs to be introduced, based on the type of question that was asked, on the size of the database itself, how much its data changes on a regular basis, etc.

[6] A final decision on whether data should be left in pristine condition or not will have to be made by the entity that controls the data. Such decision would, hopefully, balance the risk of a data breach and the strength of the security protections that are afforded to the data, against the specific needs of the application and the benefits this application provides to data subjects, to society, and other considerations, including technical ones, as appropriate.

- The DP guard keeps track of the cumulative privacy "cost" of all the questions that have been asked in the past, and also has a privacy "budget" that is assigned to the database as a whole. Should the cumulative cost of the questions asked by all analysts reach the budget, the privacy guard can raise the alarm and a policy decision can be made by the entity that controls the data, on whether the amount of distortion introduced to answers needs to be increased, or whether the risk is worth the reward.[7] The privacy equivalents of cost and budget are formalized into the privacy guard, so that DP can avoid scenarios in which the answers to different questions can be combined by multiple analysts in such a way that any individual's privacy is breached eventually.

- The DP guard works as a helpdesk. While analysts need to understand that the responses they will get from the DP guard are noisy in order to correctly interpret them, from the point of view of programming and submitting queries to the database, analysts do not need to understand DP, or privacy for that matter. They can simply keep asking questions and getting responses. The DP guard keeps track of the cumulative privacy cost of the questions that have been asked. If answering a particular question would mean exhausting the privacy budget, then the guard can inform the analyst that that particular question cannot be answered, or take other appropriate action as per the ruling policy.

- Research shows that DP's privacy protections are strong enough that, so long as the DP guard is built, protected and used properly, the analyst will not learn anything that can be combined with any other database to breach privacy.

Having said this, it is important to underscore the fact that DP is a still a research-level technology, not a commercial product, and that its potential implementation in real-life research and commercial scenarios, such as the one we described, will present mathematical, computational, and policy challenges that will need to be addressed before it can go into production.

Conclusion

Big data technologies offer promise and bring potential concerns. Society can only reap the full benefits offered by the data age if the privacy of individuals is protected at the same time. Microsoft believes that in order for society to reap the full benefits offered by the data age and the creative efforts of researchers and developers, without significantly eroding individual privacy, we will have address a variety of different needs and requirements. For some use cases, leveraging new and innovative privacy-protecting technologies like Differential Privacy will help meet those requirements.

For more information about Differential Privacy go to:

- Database Privacy: http://research.microsoft.com/en-us/projects/databaseprivacy/

- Privacy Integrated Queries (PINQ): http://research.microsoft.com/en-us/projects/PINQ/

[7] Such a decision would of course have to take into account user consent and the applicable regulatory framework.