Differential Privacy: What Is It?

1 March 2019 1,675 views One Comment

Joshua Snoke and Claire McKay Bowen on Behalf of the ASA Privacy and Confidentiality Committee

As statisticians, many of us use public-use data files and internal institutional or product data under nondisclosure agreements. We also analyze restricted data obtained via data use agreements, remote access, or research data centers such as those supported by the US Census Bureau. Statisticians in industry, though likely not using data use agreements or restricted data centers, still face questions about data sharing, either internally or for products developed using sensitive data. A much smaller fraction of our community is involved in the development and implementation of statistical disclosure control and data privacy techniques permitting access to confidential data and publication of the results based on such data.

Statistical disclosure control (SDC), or limitation, has existed as a subfield of statistics since the mid-20th century. Using these techniques, data maintainers have for years provided both the public and research communities high-quality data products that preserve the confidentiality of sensitive data.

The data landscape has changed dramatically over the past two decades. Advances in modern information infrastructure and computation have made it easier to collect, store, and analyze vast amounts and varieties of data. These advances have also enabled us to reconstruct databases and identify individuals from supposedly anonymized data.

Even for those who advocate only giving data to trusted researchers, recent misuses of data access such as those by Cambridge Analytica further call into question who can truly act as a trusted third party. The difficult question of how to achieve the proper trade-off of data quality and disclosure risk, while enabling data sharing and supporting scientific endeavors and sound policy decisions, has become even more difficult.

There exists a need for increased efforts to tackle statistical data privacy and disclosure control that match the shifting data culture. We, the statisticians, can play a more significant role. Addressing this problem starts with a healthy interest and discussion among statisticians about privacy demands in our current data culture and how we should best address those concerns. A comparable situation is emerging in economics with several debates about the meaning of privacy. Demography and health policy, among others, are also having similar discussions, with each field contributing research based on their unique contexts and problems.

A proper approach to data confidentiality and privacy has been and continues to be an area of hot debate that differs widely across contexts. While traditional methods of SDC and secure data centers are still used extensively, varying opinions about procedures have been developed across academia, government, and industry and in different countries. A definition known as Differential Privacy (DP) proposed by Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith has garnered much attention, and many researchers and data maintainers are moving to develop and implement differentially private methods.

Notably, research using this definition has been dominated by researchers in computer science, with few statisticians pursuing the topic. This situation has led to an imbalance in the goals and scenarios for which differentially private applications have been produced. Traditional inferential analysis, for example, has been under-researched compared to machine learning topics and prediction tasks. As statisticians become more involved, they should seek further collaborations with computer scientists, increasing the statistical perspective in the field and building on the work already performed.

The DP 'Hype'

So, why all the "hype" for DP? DP's primary selling point, in contrast to its SDC methodological ancestors, is the offer of privacy according to a provable and quantifiable amount, sometimes referred to as the privacy-loss budget. A lack of formal quantification has often been a critique of older methods, which instead implement ad-hoc procedures with subject-matter experts determining how to quantify the privacy protection under assumed hypothetical scenarios. The formal quantification offered by DP has inspired new research efforts, generally falling under the term "formal privacy," that seek to develop approaches with quantifiable privacy loss. These methods do not have to adhere to DP structures, but draw on many of its philosophical underpinnings.

How does DP exactly add quantifiable noise to the true data? It uses the concept of a privacy-loss budget, often denoted mathematically as e. This concept is useful to explain the definition in nontechnical terms and it allows the data curator, or steward, to know how much information is being leaked to researchers accessing the data. Specifically, if the data curator "spends" more of the privacy-loss budget (a larger value of e), researchers should gain more accurate information about the data. However, providing more information and greater accuracy typically means less privacy is guaranteed because information is being "leaked." Inversely, a data curator spending a smaller privacy-loss budget will result in less accurate information from the data but ensure a higher privacy protection. Additionally, with a set privacy-loss budget, DP adjusts the amount of noise being added to the data based on how sensitive the information is the researchers want. This sensitivity is not in terms of personal or private information, but how robust the information is to outliers.

Specifically, DP works by tying privacy to how much the answer to a question or statistic is changed given the absence or presence of the most extreme possible person in the population. For example, suppose the data we want to protect is income data and the statistic we want answered is, "What is the median income?" The most extreme person who could possibly be in any given income data could be Jeff Bezos. If he is absent or present in the data set, the median income shouldn't change drastically. This means we can provide a more accurate answer for us and the researchers about the median income that satisfies DP without using much privacy-loss budget.

What if the question is, "What is the maximum income?" Unlike the previous statistic, the answer would significantly change if Bezos is absent or present in the data set. A DP algorithm would provide a less accurate answer, or require more privacy-loss budget, to answer this query and protect the extreme case, Bezos.

Another strength of DP is it proposes a fully transparent approach, contrasting many older techniques that use the lack of transparency as additional protection. Although security through obscurity might seem to provide protection, it detracts from the credibility and robustness of the procedures for many researchers.

Additionally, from a statistical perspective, transparency is as important for preserving analyses on the perturbed data as it is for guaranteeing privacy. The typical argument in favor of older approaches is they recommend less perturbation than needed to satisfy DP, but they do not typically provide users a way to account for the impact of the additional noise, which leads to the analyses being either biased or an under-estimation of the standard errors. Knowing the distribution and magnitude of the noise added should allow researchers to account for this bias in the modeling procedure.

'Relaxed' DP

In some cases, protecting the most extreme outlier can be difficult, if not impossible. To remedy this, researchers have developed alternative definitions, or "relaxed" versions, of DP. One common relaxation is approximate DP, which has an additional probability (albeit a tiny one) that

someone within the data will be fully exposed while everyone else is protected under the full privacy-loss budget. This small probability tends to be on the order of 10-3 to 10-5.

Another popular alternative definition is local DP, which has the same spirit as regular DP but is conceptually different. Local DP focuses on how individual data are collected and aggregated for analysis. Each individual protects his or her data before sending it to the data curator to be used, and the total privacy-loss budget is divided evenly among all individuals. This means local DP methods trust no one, including the data curator, so the information the data curator receives is already noisy. These approaches are commonly combined with multi-party computation techniques.

Practical Issues

Overall, DP has inspired a new era of data privacy research and been incorporated into applications such as shrinkage regression, principal component analysis, genetic association tests, Bayesian learning, location privacy, recommender systems, and deep learning. Despite the rapid growth in papers published on the topic, issues remain. As the field has grown, there have been cases in which published attempts at formal privacy methods do not actually satisfy DP. The challenge of guaranteeing research quality (echoing the reproducibility issues in other fields) comes as no surprise, considering how young the field is.

In addition to the demand for transparency in data privacy research, most of the work is theoretical and there are few implemented applications on real research data. Because many proofs require theoretical conditions not typically met, problems exist for implementation on real-life data. Many approaches only apply to particular data, such as univariate data, or a particular type of data, such as categorical and not continuous. Another common problem is some DP methods make unrealistic assumptions about publicly available knowledge of the data to improve the proposed method's results. Furthermore, several theoretical approaches are computationally demanding or unfeasible, hindering the applicability for an average data curator with limited computational resources. Computational complexity is a particular issue for techniques that seek to release entire data sets.

Many of these practical issues represent an area in which increased input from statisticians and other collaborations should lead to improved solutions. In fact, the underlying concept is not new for statistics. Randomized response, originally proposed by Stanley Warner, is equivalent to a form of local DP. Approaches that enable feasible data analysis under DP will require statistical expertise and techniques that understand the effect of perturbations on the data and analyses. For instance, noise added to a statistic from different distributions may offer equal protection but significantly different amounts of bias or variance to an estimate of interest.

Alternatively, adding noise in one fashion may preserve predictive power of a model, but it may remove any meaningful inference from the estimated parameters.

Most papers on privacy by computer scientists focus on prediction tasks or models, which dictate the design of privacy algorithm. Few of these methods are suitable for statistical inference. The problem becomes even more complicated when considering complex survey designs, weighting schemes, missing data, or many of the common issues faced with real data. With the under-representation of statisticians in data privacy, these research questions have been barely touched.

Moving Forward

A few applications exist that try to move DP from the theoretical to the practical and begin to address the problems that inevitably arise. These can be used as examples for efforts moving forward, both for statisticians and computer scientists. Within industry, Google, Apple, and Uber have explored versions of DP.

Google created a local DP method called RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response), an end-user client software on Chrome browser data for crowd-sourcing statistics. Upon initial release, RAPPOR only worked on a few cases and the examples in the original paper were idealized.

Apple successfully implemented a local DP approach on their iPhone to gain information about suggested emojis based on keyboard strokes (in iOS 10 and 11). However, multiple sources suggest the privacy-loss budget employed by Apple to collect users' data on mobile devices is too high to be acceptable for privacy protection.

Uber took a different route and developed an approximate DP method instead of using local DP, a substantial and unique attempt on practical DP SQL queries. Like Google and Apple, Uber's approach also suffered conceptual and technical flaws.

All three companies are continually updating their implementations and working toward improving on their current shortcomings.

Consequently, much work remains in the early stages of progress with respect to DP and other formal privacy techniques. The US Census Bureau is undertaking a major overhaul of the data privacy approaches for the 2020 Decennial Census and considering carrying out similar changes to other data products.

The bureau has already spearheaded applications of DP and other formal privacy methods. One of the more popular examples is the US commuter data called OnTheMap, which is protected by approximate DP.

Less well-known examples are the Post-Secondary Employment Outcomes longitudinal data and Opportunity Atlas on childhood social mobility. While these examples are introducing formal privacy approaches, many standard disclosure limitation approaches are still used by the US Census Bureau and remote access data-sharing centers remain essential for providing original data to responsible researchers.

A new opportunity geared toward practical DP is the National Institute of Standards and Technology's (NIST) Differentially Private Synthetic Data Challenge. Synthetic data sets are a leading method for releasing publicly available data that can be used for numerous analyses. Combining this with DP can maximize both rigor and flexibility. The challenge seeks submissions of practically viable DP data sets that will be tested for accuracy on summary statistics, classification tasks, and regression models. Detailed proofs and code must be made available for the submissions, and the highest-scoring submissions will receive cash prizes. This competition encourages in-demand work, moving DP from the theoretical to the practical as NIST seeks to establish a measurement-based approach to fostering data-driven research and development in this area.

In addition to these practical implementation hurdles, there is no consensus on the appropriate amount of privacy-loss budget for practical use. As mentioned earlier, some may argue Apple's local DP approach use of the privacy-loss budget is too high for practical use, but what is the threshold of "too high"? Any attempt to answer this question should recognize it will always be contextual. It is both a policy and social question, which will likely be answered by stakeholders who are able to bear responsibility for the decision. Furthermore, data privacy researchers, including statisticians, should be involved, since we will be responsible for communicating the best trade-off curve and informing decision-makers about appropriate interpretations of the privacy-loss budget. Finally, the participants in the data may have their own views on this value. Incorporating all this to determine the right budget is an open question.

The decision about how to implement DP at the US Census Bureau and elsewhere is far from made; a robust conversation with the research community is just beginning. Although statisticians have representation in the field, the statistics community should seek a more active role. Without the involvement from the statistics community, we risk more biased research data. Advancing privacy methodology is necessary to maintain peoples' trust in institutions such as tech companies and the US Census Bureau. Public opinion of many institutions, particularly those responsible for collecting and maintaining data, has degraded over the past few years and will continue to do so without significant work overhauling our privacy methods.