Differential Privacy

Salil Vadhan
School of Engineering and Applied Sciences, Harvard University
salil_vadhan@harvard.edu

2019

WHAT IS DIFFERENTIAL PRIVACY?

(This section, except for the fifth paragraph, is a verbatim extract from Wood et al. 2019)

Differential privacy is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. It is used to enable the collection, analysis, and sharing of a broad range of statistical estimates based on personal data, such as averages, contingency tables, and synthetic data, while protecting the privacy of the individuals in the data.

Differential privacy is not a single tool, but rather a criterion, which many tools for analyzing sensitive personal information have been devised to satisfy. It provides a mathematically provable guarantee of privacy protection against a wide range of privacy attacks, defined as attempts to learn private information specific to individuals from a data release. Privacy attacks include re-identification, record linkage, and differencing attacks, but may also include other attacks currently unknown or unforeseen.[1] These concerns are separate from security attacks, which are characterized by attempts to exploit vulnerabilities in order to gain unauthorized access to a system.

Computer scientists have developed a robust theory for differential privacy over the last fifteen years, and major commercial and government implementations are starting to emerge. Differential privacy mathematically guarantees that anyone viewing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis.

What can be learned about an individual as a result of her private information being included in a differentially private analysis is limited and quantified by a privacy loss parameter, usually denoted epsilon (ε). Privacy loss can grow as an individual's information is used in multiple analyses, but the increase is bounded as a known function of ε and the number of analyses performed.

Differentially private algorithms are constructed by carefully introducing "random noise" into statistical analyses so as to obscure the effect of each individual data subject. Thus, differential privacy

[1] Even seemingly innocuous aggregate statistics can be exploited to reveal sensitive individual-level information. See Dwork et al. (2017).

reduces the accuracy of statistical analyses, but does so in a quantifiable manner that introduces an explicit privacy-utility tradeoff. As the number n of observations in a dataset grows sufficiently large, the loss in accuracy due to differential privacy generally becomes much smaller than that due to statistical sampling error. However, it can be challenging to maintain high accuracy for studies on modest-sized datasets (or modest-sized subsets of large datasets).

The differential privacy guarantee can be understood in reference to other privacy concepts:

- Differential privacy protects an individual's information essentially as if her information were not used in the analysis at all, in the sense that the outcome of a differentially private algorithm is approximately the same whether the individual's information was used or not.
- Differential privacy ensures that using an individual's data will not reveal essentially any personally identifiable information that is specific to her, or even whether the individual's information was used at all. Here, specific refers to information that cannot be inferred unless the individual's information is used in the analysis.

As these statements suggest, differential privacy is a new way of protecting privacy that is more quantifiable and comprehensive than the concepts of privacy underlying many existing laws, policies, and practices around privacy and data protection. The differential privacy guarantee can be interpreted in reference to these other concepts, and can even accommodate variations in how they are defined across different laws. In many cases, data holders may use differential privacy to demonstrate that they have complied with legal and policy requirements for privacy protection.

Differential privacy is currently in initial stages of implementation and use in various academic, industry, and government settings, and the number of practical tools providing this guarantee is continually growing. Multiple implementations of differential privacy have been deployed by corporations such as Google, Apple, and Uber, as well as federal agencies like the US Census Bureau. Additional differentially private tools are currently under development across industry and academia.

Some differentially private tools utilize an interactive mechanism, enabling users to submit queries about a dataset and receive corresponding differentially private results, such as custom-generated linear regressions. Other tools are non-interactive, enabling static data or data summaries, such as synthetic data or contingency tables, to be released and used.

In addition, some tools rely on a curator model, in which a database administrator has access to and uses private data to generate differentially private data summaries. Others rely on a local model, which does not require individuals to share their private data with a trusted third party, but rather requires individuals to answer questions about their own data in a differentially private manner. In a local model, each of these differentially private answers is not useful on its own, but many of them can be aggregated to perform useful statistical analysis.

Differential privacy is supported by a rich and rapidly advancing theory that enables one to reason with mathematical rigor about privacy risk. Adopting this formal approach to privacy yields a number of practical benefits for users:

- Systems that adhere to strong formal definitions like differential privacy provide protection that is robust to a wide range of potential privacy attacks, including attacks that are unknown at the time of deployment. An analyst using differentially private tools need not anticipate particular types of privacy attacks, as the guarantees of differential privacy hold regardless of the attack method that may be used.

- Differential privacy provides provable privacy guarantees with respect to the cumulative risk from successive data releases and is the only existing approach to privacy that provides such a guarantee.
- Differentially private tools also have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. This feature distinguishes differentially private tools from traditional de-identification techniques, which often conceal the extent to which the data have been transformed, thereby leaving data users with uncertainty regarding the accuracy of analyses on the data.
- Differentially private tools can be used to provide broad, public access to data or data summaries while preserving privacy. They can even enable wide access to data that cannot otherwise be shared due to privacy concerns. An important example is the use of differentially private synthetic data generation to produce public-use microdata.

Differentially private tools can, therefore, help enable researchers, policymakers, and businesses to analyze and share sensitive data, while providing strong guarantees of privacy to the individuals in the data.

WHAT ROLE CAN DP PLAY IN ISSOD?

There are several ways in which differential privacy could be used in ISSOD.

Public-Use Data Summaries

Differential privacy can be used to produce rich statistical summaries of sensitive datasets that can be shared widely without worry that the combination of released statistics will reveal individual-level information (in contrast to data de-identified using traditional means, which have repeatedly been shown to be vulnerable to re-identification). This is similar to how the US Census Bureau plans to use differential privacy to produce public-use microdata samples for the 2020 Decennial Census. In addition to tables of statistics that would be of common interest, in principle it is possible to generate differentially private "synthetic data" that reflects many statistical properties of the original dataset and thus can be treated as a safe-to-release proxy for the original dataset. (For example, this can be done by estimating the parameters of a statistical model in a differentially private way, and then generating new data points using the model with estimated parameters.)

Interactive Queries for Approved Researchers

ISSOD could also provide approved researchers with a query interface to run differentially private analyses of interest to them on the data. The reason to limit such access to approved researchers is that every query made increases the privacy loss (ε) measured by differential privacy, and thus there is a finite "privacy budget" of queries that can be made while maintaining a desired level of privacy protection. The privacy budget could be quickly exhausted with a public query interface.

On their own, differential privacy tools may not provide sufficient accuracy or support for the statistical methods that a researcher needs to use to obtain publishable results. In such a case, a differentially private query interface can be used for *exploratory data analysis*, for the purpose of selecting datasets among the many in ISSOD and for formulating hypotheses. The final analysis could then be carried out in a more controlled manner, for example in an enclave similar to Census Research Data

Centers or by having an ISSOD statistician vet and run the analysis for the researcher. A side benefit to using differential privacy for hypothesis generation is that it provides automatic protections against overfitting and false discovery (Dwork et al. 2017).

In general, in the near-term, we advocate thinking of differential privacy as part of a *tiered access model*, offering a way to provide wider access to sensitive data, but not necessarily being the only way in which researchers will be able to access and analyze data.

CHALLENGES TO OVERCOME

Inherent limitations of differential privacy

It is a consequence of the definition of differential privacy that one cannot carry out analyses that focus on the data of specific individuals or small groups of individuals. Thus, to allow such research (e.g, qualitative research on specific social media posts), it is necessary to have a different model for privacy (e.g, one based on informed consent). Indeed, differential privacy is designed to allow (only) for "statistical" research.

Relatedly, the accuracy of differentially private computations (relative to the magnitude of statistical sampling errors) inherently degrades as the number of data subjects decreases and/or the number of queries increases. While further engineering of differentially private algorithms will improve this situation, it will likely remain the case that statistical analyses of "small" subsets of datasets (e.g, hundreds of records) will incur a substantial loss in accuracy, and thus may require augmenting differential privacy with other modes of access (e.g, data enclaves as suggested above).

The setting of the overall privacy budget (ε) for a dataset, taking into account both its privacy and utility implications, and deciding how it should be managed among many potential data analysts can be difficult policy questions. One can have a standardized set of choices for the privacy budget, depending on the qualitative risks of privacy harms associated with the dataset (e.g, minor embarrassment vs. life-or-death consequences), but this omits consideration of the utility implications. As far as the management of the privacy budget, one option is to give approved researchers each separate privacy budgets for their queries, but this requires trusting that those researchers will follow an agreement to not collude.

Finally, it is important to note that differential privacy does not protect everything that one might consider "sensitive" or "confidential" in a dataset. It is targeted specifically at preventing the leakage of individual-level information, while supporting statistical analyses where the unit of observation is an individual data subject. There may be other concerns associated with the sharing of a dataset, such as a company's intellectual property or collective rights of a large subgroup (cf. Harding et al. 2012), and differential privacy does not address such concerns. Moreover, even with differential privacy, individuals may be harmed by the knowledge gained and actions taken as the result of a statistical analysis on a dataset; differential privacy only promises that any such harms are not due to the contribution of one's own data. Thus, differential privacy does not eliminate the need for a consideration of how the value of the research being performed relates to potential adverse consequences.

The Need for Software Tools

As mentioned in the first section, differential privacy has started to have large-scale deployments in industry and government (Google, Apple, US Census Bureau) and there is also an extensive body of implementation and experimental work in academia.

As far as we know, all of this work falls short of the needs of ISSOD in that it is:

1. highly tailored to a specific application, with particular data sets and types of analyses to be supported,
2. requires more expertise in computer science or differential privacy than a practicing social scientist would have, and/or
3. has not been vetted by the differential privacy community at large.

In the Privacy Tools Project at Harvard, we have been developing a differential privacy tool, PSI (Private data-Sharing Interface), that aims to address Items 1 & 2 above (Gaboardi et al. 2016). PSI is being designed to integrate into data repositories like Dataverse, to allow for releasing public statistical summaries and providing interactive queries for exploratory data analysis. PSI is planned for deployment when Dataverse starts to accept sensitive data. (See Crosas whitepaper "Dataverse, DataTags, and a decade building a widely-used data repository platform.")

However, PSI's functionality (in terms of the statistical analyses it supports) is still fairly limited, and the underlying library of implemented differentially private algorithms has not been vetted by the research community at large. Thus, we are also discussing the possibility of expanding our efforts to launch a larger open-source project for the differential privacy community to develop general-purpose and usable tools for differential privacy in which users can have confidence.

Relatedly, there is still a significant amount of research and engineering to be done to close the gap between many of the theoretical algorithms in the differential privacy research literature and implementations of optimized methods that meet the needs of practicing data analysts, such as those who would study the datasets in ISSOD. In particular, much more work needs to be done on designing practical differential privacy tools for statistical inference (as opposed to descriptive statistics) and for generating synthetic data. Additionally, the standard and most widely studied form of differential privacy is meant for tabular data where the unit of observation is an individual person. There are forms of differential privacy that are applicable to network data (e.g, where a node corresponds to an individual person), but these have been less well-studied than differential privacy for tabular data and are likely a bit further away from yielding practical tools.

References

Dwork, Cynthia, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. "Exposed! A Survey of Attacks on Private Data." *Annual Review of Statistics and Its Application* 4 (1): 61–84. Accessed May 2, 2019. doi:10.1146/annurev-statistics-060116-054123. http://www.annualreviews.org/doi/10.1146/annurev-statistics-060116-054123.

Gaboardi, Marco, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. 2016. "PSI (Ψ): A Private Data Sharing Interface." Accessed May 2, 2019. arXiv: 1609.04340 [cs.SR]. http://arxiv.org/abs/1609.04340.

Harding, Anna, Barbara Harper, Dave Stone, Catherine O'Neill, Patricia Berger, Stuart Harris, and Jamie Donatuto. 2012. "Conducting Research with Tribal Communities: Sovereignty, Ethics, and Data-Sharing Issues." *Environmental Health Perspectives* 120 (1): 6–10. doi:10.1289/ehp.1103904.

Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. 2019. "Differential Privacy: A Primer for a Non-Technical Audience." *Vanderbilt Journal of Entertainment and Technology Law.*