Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

Restore the Fourth

Restore the Fourth Comment on Privacy-Enhancing Technologies (PETs)

The Office of Science and Technology Policy (OSTP)—on behalf of the Fast Track Action Committee on Advancing Privacy-Preserving Data Sharing and Analytics of the Subcommittee on Networking and Information Technology Research and Development (NITRD) of the National Science and Technology Council, the National Artificial Intelligence Initiative Office, and the NITRD National Coordination Office—recently requested public comment to help inform development of a national strategy on privacy-preserving data sharing and analytics, along with associated policy initiatives.

Restore the Fourth is a 501(c)(4) nonprofit organization committed to strengthening the United States' notion of privacy as enshrined in the Fourth Amendment and other applicable statutes and case law. While our focus is on curbing mass government surveillance, any and all correlation of data and data analytics has surveillance implications, whether directly by the government or nefarious actors. This particular comment will focus on privacy-enhancing technologies for medical, financial, and location data.

There are several broad principles that apply to collection and application of this data:

- Data, and thus surveillance applicability, begins at the point of collection
- Not all data is alike in terms of its risk profile
- Once data is in the hands of entities outside of its original scope, it can be bought, sold, and exploited completely free of privacy protections accorded to the original research

Data, and thus surveillance applicability, begins at the point of collection

In the modern Internet-connected information landscape, every US resident has multiple devices on their person and in their homes that are capable of collecting private information beyond the scope of its design and primary implementation. In the case of medical and financial data, the usual devices are Internet-connected smartphones and personal computers, as well as implantable and wearable medical devices that collect real-time biological information. Many of these datasets are geolocated to varying degrees of precision, and thus have embedded location data with its own privacy implications.

Financial data is particularly thorny from a data privacy and security perspective because of its relative centralization; while ATMs and credit card processing operate closest to the consumer, Americans already are governed by opaque algorithms derived from central processing of their transactions: the credit rating system. Modern advances in payment processing, from electronic payments (PayPal, Stripe, Plaid, Spreedly) to gamification/incentivization of preferred spending patterns (store rewards cards, digital healthcare platforms) have enabled financial transactions to originate from websites and regions more and more remote from this centralization.

Regardless of its scope and content, the data generated is stored onboard, *somewhere*, before transmission to other systems or over the Internet. Thus, common data security and privacy protection technologies, like "access control, data anonymization, data encryption, differential privacy protection, digital watermarking…identity authentication" [1] should be performed on-device wherever possible. In addition, the particular use case of machine learning algorithms being applied to large datasets allow for the possibility of federated learning, in which training algorithms are executed "across [decentralized] edge devices (e.g., individual mobile phones) or servers hosting different local samples (e.g., data owned by different samples). Data samples are not shared or [centralized] and only the trained models are communicated, which might improve data security and privacy of patient data" [2]. This of course does not limit the scope of surveillance that can be collected by law enforcement officers (LEOs) at the device level, many techniques of which carefully skirt US privacy law via overbroad and intentionally backdoored legal regimes, but it can reduce the ability of other nefarious actors to collect and act on this information.

Contacting research participants is also a day-to-day activity in research with this kind of data; unfortunately, the risks of email makes it trivial to de-anonymize research communications, even without a 1:1 match to exact data points. Implementing robust, scalable pseudonymous remailers allows for one more layer of data privacy; e-commerce sites like Craigslist have implemented this already with a minimal tech footprint.

Not all data is alike in terms of its risk profile

One of the most exciting prospects in current medical research is the ability to examine large swathes of disparate data algorithmically and extract useful patterns. While there is ultimately human oversight of the analysis of the data itself, the techniques and the access to the data vary depending on budget, centralization of resources, and the composition and behavior of researchers and supervisors. The federal government, despite its best efforts, is not a monolith; project-by-project myopia (and thus duplication of resources) virtually ensures that data privacy, and PETs by extension, vary widely in existence, implementation, and maturity.

Access control technology in particular has become a hotspot of current research, but by being targeted mostly at the operating system level, ignores the differential risk associated with different kinds of medical data. While role-based access control (RBAC) is a fundamental, if antiquated, tenet of identity access and management (IAM) practices across governmental organizations, more care must be taken to closely examine the risk of exploitation by the *kind* of data, and not just the access to data itself.

Some researchers have expanded RBAC through mathematical methods to something they call Risk-Adaptive Access Control; in this case, risk is quantified as the "deviation degree between users' access to medical information and their work tasks…[the] greater the deviation degree is, the greater the risk". By calculating the information entropy of users accessing medical information, taking that entropy as an input data set for advanced data processing (K-means

clustering) and seeing what patterns emerge with respect to a defined baseline risk, supervisors and administrators can "dynamically access control policies based on users' access conditions" [1].

While participant data itself can be exploited by any number of entities, the analytical techniques themselves are at risk as well. A recent paper by researchers at UC Berkeley, MIT, and the Institute for Advanced Study (IAS) demonstrates the ability to plant undetectable backdoors in machine learning classifiers [3]—incidentally, a class of techniques used broadly in analysis of Big Data of this nature. Backdoors of this nature can be used for any purpose, from altering research results to enabling de-anonymization of research protected via differential privacy techniques.

Once data is in the hands of entities outside of its original scope, it can be bought, sold, and exploited completely free of privacy protections accorded to the original research

The biggest data privacy risk of medical, financial, and location data is the ability for it to be correlated to other large datasets, and used to re-identify participants via this correlation. Datasets, while large in size, are ultimately portable, and their ability to exploit their participants via this type of correlation is ultimately limited by computing power and time.

In an era of data leaks, third-party data brokers, and overt intragovernmental information sharing, surveillance and exploitation is always on the table no matter who is doing the watching. The long tail of the 2020 Minneapolis protests and the armed insurrection on January 6 has also shown that there is always a law enforcement and thus profiling use case for re-identifying of this data. It is never fully possible to eliminate this possibility, given that this data is often freely exchanged on the black market, but one way to reduce legally-sanctioned scope creep of the data's applicability is to ensure that use of the data is governed by licenses that require disclosure to, and agreement by, the original researchers for activities that re-identify it. In addition, the US government should adopt standard data retention policies for types of data, particularly those that are re-identifiable (whether at present or by undisclosed surveillance and analytical capabilities).

In conclusion, the expansion of Big Data capabilities of the US government comes with numerous open questions and pitfalls given by its applicability to surveillance, but extending back to data privacy and their corresponding PETs. Restore the Fourth as an organization focuses on the Fourth Amendment as a cornerstone of individual privacy, but actual policy that protects and enshrines this idea depends not just on accompanying statutes and case law, but specific principles that are obeyed in implementing PETs at any level of data collection and data sharing.

References

[1] - Rong Jiang, Shanshan Han, Mingyue Shi, Tilei Gao, Xusheng Zhao, "Healthcare Big Data Privacy Protection Model Based on Risk-Adaptive Access Control", Security and Communication Networks, vol. 2022, Article ID 3086516, 12 pages, 2022. https://doi.org/10.1155/2022/3086516
[2] - Yadi Zhou, Fei Wang, Jian Tang, Ruth Nussinov, Feixiong Cheng, Artificial intelligence in COVID-19 drug repurposing, The Lancet Digital Health, Volume 2, Issue 12, 2020, Pages e667-e676, ISSN 2589-7500, https://doi.org/10.1016/S2589-7500(20)30192-8.
[3] - https://arxiv.org/abs/2204.06974