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

University of Southern California Information Sciences Institute

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government. We bear no responsibility for the accuracy, legality, or content of the responses and external links included in this document.

RFI Response: Request for Information on Advancing Privacy-Enhancing Technologies

July 8, 2022

These comments are submitted on behalf of the University of Southern California's Information Sciences Institute.

1. Response to Topic # 1: Specific research opportunities to advance PETs

1.1. Understanding and Regulating How Companies Collect, Store and Use Private Data

Startup landscape today is mostly unregulated when it comes to private data collection. What data is collected and how, how it is stored and how it is used is left fully to companies to decide. At best companies must inform users of data collection and get them to accept the terms of services. Companies misuse this by asking too many privileges from users, and justifying this by saying that users have freedom to not use a service. It is well known that users are not rational when it comes to deciding which service to use and how to reason about various privacy/utility trade-offs. It is also very likely that companies collect more data than they need, and store it in an ad-hoc manner. When that data is leaked, customers' privacy is put at risk.

We need thorough analysis of business practices around private data collection, storage and use. We need to understand how much of the data is really needed for each business model, and how that data can best be protected. We then need common APIs for data collection and common, secure services for storing and querying this data. Finally, we need to ensure that customers can view and delete their data at will and that they can revoke access to certain data while still being able to access services (potentially at a lower utility level). For example, a customer may revoke access of an online service to their address, which should lead to deletion of this address from the company's systems. The user should still be able to browse and order items, but would then be prompted to input their address each time they check out.

Further, we need studies on how users reason about privacy choices, and how well they can meet their privacy goals with current user interfaces. We also need to shift to a private-by-default model for user data. Instead of asking users to give access to their private data to the company offering a free service, the company should ask for data access only when absolutely necessary. No data should be assumed public and no data should be kept longer than absolutely necessary.

2. Response to Topic # 2: Specific technical aspects or limitations of PETs

2.1. Privacy-infringement of PETs

Privacy preserving technologies enable sharing of sensitive information while preserving privacy. However, there are open questions about the privacy implications of some of the used technologies. In this response, we specifically discuss two of these technologies and the questions around them.

The first technology is using AI generated face images to replace real face images in the shared data. The problem with this technology is that there is no guarantee that the generated face images do not correspond to real people. Therefore, if AI generated images are used to replace real images, the privacy of an individual might be compromised in order to preserve the privacy of another individual. In a pilot study we conducted at USC ISI, we estimated that around 90% of face images in the CelebA-HQ face dataset can be faithfully generated by a face image generation model (StyleGAN2). That indicates the high recall of such a model with respect to the distribution of real face images. Estimating the precision of these models is a more interesting and challenging task that requires further research.

The second technology is using adversarial perturbation to obfuscate the real identity of a face image. In this process, the face image is made to match an identity that is different from the original one by adding an imperceptible perturbation. However, the privacy of the new identity can be infringed by the practice. Therefore, methods for adversarial face image obfuscation should be designed such that there is some level of guarantee that the target identity of the obfuscation process does not match a real identity.

2.2. Privacy-preserving Entity Resolution

Entity resolution is the task of identifying records in different datasets or databases that refer to the same entity in the real world. For example, consider patient databases in two hospitals. The entity resolution task in this case is to identify the records of the same patients in the two databases. To accomplish this task, an algorithm will compare the information present in the records to determine if the records refer to the same person. The algorithm would compare the names, home or work address, phone numbers, etc. to make the determination. The principal challenge is that the two hospitals may record the information in different formats. For example, one hospital may use separate fields for first name and last name, and the other hospital may story the full name in a single field, including a middle name. The privacy preserving entity resolution task is to solve the entity resolution task without revealing sensitive information. Specifically, in the privacy preserving setting, each party will know which of its records are present in the databases of the other parties, but will learn nothing about any other records that the other parties may have in their databases. For example, if one hospital has records about Mary, Joe and Bob, and the other hospital has records about Joe, Bob and Susan (and assuming that Joe and Bob are the same person), then the first hospital will know that the second hospital has records about Joe and Bob, but will not learn any information about the two people unless explicitly disclosed by the second hospital. Furthermore, the first hospital will not know that the second hospital has a record about a person named “Susan”. In fact, the first hospital will not learn anything about the distribution of names, locations or any other information present in the database of the second hospital. Likewise, the second hospital will not learn any information about the first hospital’s database beyond knowing that it has records for Joe and Bob.

Privacy preservation creates a significant challenge for entity resolution as the data must be encrypted to preserve privacy, but once encrypted, similarity comparisons become difficult. This is especially the case for fuzzy or approximate matching, which is required to account for differences in formats and data structures. While entity resolution is a well-studied area of computer science (a query for “entity resolution ” in Google Scholar returns over 2 million results), privacy preserving entity resolution has received little attention in the scientific community.

The University of Southern California’s Information Sciences Institute (ISI) carries out basic and applied research in artificial intelligence, networks and cybersecurity, high- performance computing, microelectronics, and quantum information systems. Its $100M annual external funding comes from the NSF, DoD, IC, NIH, DoE, industry, foundations, and other sponsors. ISI is home to the first quantum computer in academia. Part of the USC Viterbi School of Engineering, ISI has more than 400 personnel that includes 28 faculty that advise 65 PhD students.

USC Viterbi
Information Sciences Institute