Privacy-Preserving Systems (a.k.a., Private Systems)

CU Graduate Seminar

Instructor: Roxana Geambasu

Course Introduction

Privacy Concerns

- Cybersecurity Insiders' "Biggest Privacy Issues Associated with Big Data:"
  - #1- Obstruction of privacy through data breaches
  - #2- It becomes near-impossible to achieve anonymity
  - #3- Data masking met with failure in big data
  - #4- Big data analysis isn't completely accurate
  - #6- Discrimination issues

Concern: Data Breaches

- Big data presents lucrative opportunities for adversaries.
- Adversaries can be outsiders or insiders of a company, hackers, governments, competition, etc.
- Breaches occur not only via vulnerability exploitation and social engineering, but also from public releases of innocuous-looking data.

See Security magazine's 2021 data breaches list

Example 2021 Breaches

“533 million Facebook users’ phone numbers and personal data have been leaked online.”

—*Business Insider*, Apr. 2021

- “A database of that size containing the private information such as phone numbers of a lot of Facebook’s users would certainly **lead to bad actors taking advantage of the data to perform social engineering attacks [or] hacking attempts.”**

—Alon Gal for *Business Insider*

“Massive data leak exposes 700 million LinkedIn users’ information”

—*Fortune*, Jun. 2021

- “We want to be clear that this is not a data breach and no private LinkedIn member data was exposed. […] **this data was scraped from LinkedIn and other various websites** […].”

—LinkedIn for *Fortune*

(From *Security* magazine’s 2021 list)

Concern: Data Repurposing/Reselling

- Data collected for a primary, user-visible purpose is then repurposed or shared with partners.
- This creates a divide between user expectations and reality, and shakes users' trust in the company.
- Sometimes the data "masking" is attempted, but that is insufficient for privacy protection.

"Facebook Rebuked for Failing to Disclose Data-Sharing Deals"

—The *New York Times*, Dec. 2018

- "[…] Facebook had granted business partners, including Microsoft, Amazon, and Spotify, more intrusive access to user data than it had divulged — allowing some partners' access without users' permission."

Concern: Undesired/Discriminatory Decisions

"How Marketers Use Big Data to Prey on the Poor"

— *Business Insider*, Dec. 2013

- "[…] marketers who purchase this data then use the information to sell risky financial products to the people who can least afford them."

- User data can tell surprisingly much about the user.
- These additional "meanings" of data can be used to make decisions about the users in undesirable or discriminatory ways.

What Causes These Incidents?

What Causes These Incidents?

Big data and machine learning technologies **push away** from privacy principles.

E.g., Foundational Fair Information Practices (FIPs):

- Collection limitation ↘
- Data quality ↘
- Purpose specification ↘
- Use limitation ↘
- Security ↘
- Openness/notice ↘
- Individual participation ↘
- Accountability ↘

Privacy-Preserving Systems

Overview

- Take a systems perspective to analyzing the privacy risks in data-driven ecosystems.
- Survey privacy-enhancing technologies (PETs) that can help in addressing these risks.
- Learn both basic concepts related to PETs and how they are/can be incorporated in real systems.

- Other privacy classes at Columbia are **theory-** or **law-oriented**. This class is **systems-oriented**.

Target Audience

- Future software engineers, who will build infrastructure systems and applications.
- Often, privacy expertise is relegated to a small number of people on the “privacy team.” This makes privacy an after-thought in products.
- This situation is likely to change with increasing legal, social, and usability pressure on companies to adopt “privacy by design.”
- Hence, privacy expertise can become a benefit in SE job applications.

Examples

- Ongoing W3C standardization of new web advertising APIs to replace the need for online tracking.
  - These combine differential privacy with secure multi-party computation or hardware enclaves to achieve privacy while supporting (ideally even improving!) advertising.
- Growing interest in “data clean rooms,” which are cloud offerings that let companies share data while retaining control over it.
  - These combine hardware enclaves with differential privacy.
- There is great need for systems and tools to simplify the use of these privacy technologies, hence opportunity for startups.

Recent note from Private Systems graduate

“Your Private Systems (6998) class [...] still stands out to me as one of the best classes I have ever taken, both in its content and in the learning environment you created. Being able to think more rigorously about privacy with a systems perspective really helped me when I was recently helping redesign the social welfare application system in Minnesota, making sure that the state was aware of possible deanonymization risks when sharing parts of their user data. The state's outlook on privacy has become more rigorous because of what I learned in your class. Thank you!”

Topics and Structure

(still subject to adjustments, follow class page)

Topic 1: Privacy risks and attacks

01/31 Lecture

02/07 Reading & discussion: 2 papers

Topic 2: Differential privacy (DP)

02/14 Lecture

02/21 Reading & discussion: 2 papers

Topic 3: DP deployments

02/28 Reading & discussion: 2 papers

Topic 4: Secure multi-party computation (MPC)

03/07 Lecture

03/14 Reading & discussion: 2 papers

Topic 5: Midterm project update

03/28 Midterm project presentations

Topic 6: Homomorphic encryption (HE)

04/04 Lecture

04/11 Reading & discussion: 1 papers

Topic 7: Hardware enclaves

04/18 Lecture; Reading & discussion: 1 paper

Topic 9: Compositions and tensions of privacy technologies

04/25 Lecture

Topic 10: Final project presentations

05/02 Final project presentations

Assignments

1. Basic Concept Homeworks
2. Team Project

Instructions for both will appear on Courseworks/Files before the second class

Basic Concept Homeworks

- Resolved **individually** to deepen understanding of the basic concepts
- Collaboration, copying, AI use are disallowed and will be severely punished for these homeworks
- Deadlines:
  - Homework 1: Privacy attacks (due 02/20 11:59:59pm EST)
  - Homework 2: Differential privacy (due 03/13 11:59:59pm EST)
  - Homework 3: Secure multi-party computation (due 04/10 11:59:59pm EST)
- Deadlines are firm, however there is a 48-hour grace period, accumulated over all homeworks, for which you will not be downgraded
- See instructions file on courseworks

Team Project

- Instructions on Courseworks/Files
- Performed in **teams of 2 people**, to exercise a privacy-related concept in a larger context than individual homeworks can
- In most cases, should follow the applied scientific process:
  - Articulate a hypothesis; develop a methodology for testing it, which must involve design/implementation/measurement of some software artifact with specific metrics
- 5-minute informal project updates are provided in class at specific dates and are graded
- 20-minute formal project presentations will be given at specific midterm and final dates and are also graded
- We launch the project in a couple of weeks, but you can start forming groups already and consult the project instructions document to start thinking of a cool project

Connections with Other S25 CU Courses

- New this year, we aim to create project-driven connections between Private Systems and two courses at Columbia:
  - IEOR ORCS 4201 **Policy for Privacy Technologies -- Rachel Cummings** (syllabus)
  - CS COMS 6424E **Hardware Security -- Simha Sethumadhavan** (syllabus)
- We are working to support projects that span across these classes.

Deadlines (subject to adjustments, follow class page)

homework deadline

project deadline

Topic 1: Privacy risks and attacks

01/31 Lecture

02/07 Reading & discussion: 2 papers; Project rollout

Topic 2: Differential privacy (DP)

02/14 Lecture

02/20 HW1 due

02/21 Reading & discussion: 2 papers

Topic 3: DP deployments

02/27 Project proposal (one page, on courseworks)

02/28 Reading & discussion: 2 papers; Project update: proposal (informal)

Topic 4: Secure multi-party computation (MPC)

03/07 Lecture

03/13 HW2 due

03/14 Reading & discussion: 2 papers

Topic 5: Midterm project update

03/28 Midterm project update: progress and three steps (formal pres.)

Topic 6: Homomorphic encryption (HE)

04/04 Lecture

04/10 HW3 due

04/11 Reading & discussion: 1 paper; Project update: step 1 executed (informal)

Topic 7: Hardware enclaves

04/18 Lecture; Reading & discussion: 1 paper; Project update: step 2 executed (informal)

Topic 9: Compositions and tensions of privacy technologies

04/25 Lecture; Project update: step 3 executed (informal)

Topic 10: Final project presentation

05/02 Final project presentation (formal pres.)

Grading

- 40%: Three individual homeworks (split equally)
- 40%: Team project (15% x 2 formal presentations + 10% for informal updates)
- 20%: Paper discussion participation

Grades will be curved. Expect a similar distribution of grades as in typical 6000-level courses. Weekly load will be similar too.

Paper Reading and Discussions

- Every student must read every paper!
- Discussions (20% of the grade!)
  - Four Leads per paper:
    - Motivation and Overview Lead: 5 minutes (solo)
    - Detailed Design Lead: 10 minutes (solo)
    - Evaluation Lead: 5 minutes (solo)
    - Discussion Leads (multiple): 10 minutes each (engage with other students)
      - May split this role into multiple ones, with specific “duties” to think abt. and lead discussions on.
    - Discussion participants (rest): contribute especially during discussion
  - Roles **randomly chosen**, at most one role per class per student

Motivation and Overview Lead

- Introduces the motivation, problem statement, threat model, and gives a brief high-level overview of the approach.
- No slides please, just talk to friends -- help them understand what this paper is about
- **5 minutes** (longer will result in downgrade)

Grading in 0-2:

- 0: not show up, entirely unintelligible or plain-wrong presentation
- 1: presentation ineffective, beats around the bush, and doesn't demonstrate a clear grasp of the aspect requested to cover; goes over time
- 2: solid, in time overview

Detailed Design Lead

- Describes the technical approach and design in greater detail.
- Goes into technical detail, shows architecture, some math if relevant, etc.
- Can use slides or whiteboard, but also just talk to friends -- help them understand in more detail the technical approach of this paper
- OK to say that you didn’t understand certain technical aspects in the paper (but you need to have tried to understand them)!
- **10 minutes** (longer will result in downgrade)

Grading in 0-2 (same as before)

Evaluation Lead

- Describes the evaluation questions, methodology, and main results and take-aways.
- Best is probably slides, so you can project graphs/tables from the paper.
- **5 minutes** (longer will result in downgrade)

Grading in 0-2 (same as before)

Discussion Lead

- Leads student discussion by preparing a set of questions/topics of discussion
- No slides please, just raise questions, propose topics of discussion to your classmates, etc.
- As Discussion Lead, you shouldn't do all the talking! Engage your friends in an honest, interest-driven discussion about the paper.
- **10 minutes per discussion lead**

Grading in 0-2:

- 0: not show up, very unprepared
- 1: questions/topics of discussion are not interesting, talks way too much instead of engaging others in the discussion
- 2: good questions/topics, engaging discussion

Discussion Participants

- **You need to participate in the Discussion part!**

Grading:

- 0: not show up, does not participate, or very unprepared
- 1: limited but legitimate participation, or takes over the discussion needlessly
- 2: good points and participation

Class Attendance

- Attendance is required for all discussion classes and classes that have a project update/presentation.
- Engagement is required and graded during paper discussion sessions.
- Participation is in person. There is a zoom link, but it is reserved for CVN or for exceptional circumstances (such as sickness or necessity to travel). Please use sparingly!
- If you cannot attend a class, but have an assigned role for a paper discussion, you must message TA (on Ed) to re-schedule your assignment.
- Role assignment will be available shortly after Columbia's course drop-off deadline.

Resources

- Class website: https://systems.cs.columbia.edu/private-systems-class/
- EdStem (will set up and invite soon)
- Homework/project instructions, homework notebooks, and demo notebooks we use in class will be available on Courseworks/Files
- HW1-Attacks will be reviewed next time by the CA.

Staff

- CA: **Peihan Liu**
  - Ph.D. student working on privacy-enhancing systems
  - Very familiar with most of the technologies we will discuss
  - OH: See class web page (it is still TBD while TA finalizes own schedule)

Quoted Articles

- *Cybersecurity Insiders*. "What Are the Biggest Privacy Issues Associated with Big Data?"
- *Security* magazine. "*The top data breaches of 2021*."
- *Business Insider*. "*533 million Facebook users' phone numbers and personal data have been leaked online*."
- *Fortune*. "*Massive data leak exposes 700 million LinkedIn users' information*."
- The *New York Times*. "*Facebook rebuked for failing to disclose data-sharing deals*."
- *Business Insider*. "*How marketers use big data to prey on the poor*."