

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*

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 and social pressure on companies to employ “privacy by design.”
- Hence, privacy expertise can become a benefit in SE job applications.

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

01/18 Lecture

01/25 Reading & discussion

Topic 2: Differential privacy (DP)

02/01 Lecture

02/08 Reading & discussion

Topic 3: DP deployments and systems

02/15 Reading & discussion

02/22 Reading & discussion

Topic 4: Homomorphic encryption (HE)

02/29 Lecture

03/07 Reading & discussion

Topic 5: Secure multiparty computation (MPC)

03/28 Lecture

04/04 Reading & discussion

Topic 6: Private web advertising

04/11 Reading & discussion

Topic 7: Compositions and tensions of privacy technologies

04/18 Lecture, reading & discussion

Topic 8: Beyond technology: legal and policy perspectives of privacy

04/25 Invited lecture

Assignments

1. Basic Concept Homeworks (first half of the semester)
2. Team Project (second half of the semester)

Instructions for both will appear on Courseworks/Files after first 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 02/08 10:59:59am EST)
 - Homework 2: Differential privacy (due 02/22 10:59:59am EST)
 - Homework 3: Homomorphic encryption (due 03/21 10:59:59am 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 project updates are provided weekly in class and are graded
- A 20-minute final project presentation will be delivered on final exam date
- Will introduce the project in the second half of the semester, but you can consult courseworks project file for info/project list.

Deadlines (subject to adjustments, follow [class page](#))

first half of semester

Topic 1: Privacy risks and attacks

01/18 Lecture

01/25 Reading & discussion

01/25 HW 1 assigned (due 02/08)

Topic 2: Differential privacy (DP)

02/01 Lecture

02/08 Reading & discussion

02/08 HW1 due

02/08 HW2 assigned (due 02/22)

Topic 3: DP deployments and systems

02/15 Reading & discussion

02/22 Reading & discussion

02/22 HW2 due

Topic 4: Homomorphic encryption (HE)

02/29 Lecture

02/29 HW3 assigned (due 03/21)

03/07 Reading & discussion

second half of semester

03/21 Midterm quiz & team project launch

03/21 HW3 due

Topic 5: Secure multiparty computation (MPC)

03/28 Lecture

03/28 Project status update: team, selected project

04/04 Reading & discussion

04/04 Project status update: three-step execution plan

Topic 6: Private web advertising

04/11 Reading & discussion

04/11 Project status update: step 1 report

Topic 7: Compositions and tensions of privacy technologies

04/18 Lecture, reading & discussion

04/18 Project status update: step 2 report

Topic 8: Beyond technology: legal and policy perspectives of privacy

04/25 Invited lecture

04/25 Project status update: step 3 report

Grading

- 30%: Three individual homeworks (10% each)
- 20%: Midterm quiz
- 30%: Team project (10% intermediary updates + 20% final presentation)
- 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 Lead(s): 15 minutes (engage with other students)
 - May split this role into multiple ones, with specific “duties” to think abt. and lead discussions on.
 - Roles **randomly chosen**, at most one role per class per student
 - If you participate as a non-lead, you help the Discussion Lead(s), but you can also gain points if your contributions are good.

Next class papers

01/25 Reading & discussion

- Nasr, Carlini, Hayase, et al., ArXiv 2023. Scalable Extraction of Training Data from (Production) Language Models. See also the associated blog post.
- Cohen, USENIX Security 2022. Attacks on Deidentification's Defenses.

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 friends, etc.
- As Discussion Lead, you shouldn't do all the talking! Engage your friends in an honest, interest-driven discussion about the paper
- **10-15 minutes**

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

No Lead Role?

- **Please participate in the Discussion part!**
- Helps the Discussion Lead (who may help you in the future :))
- And you can get some points for it

Grading (RG still deciding whether/how it's extra credit):

- 0: all except below
- 2: insightful comments/questions/points of view

But you cannot bail on your subsequent role assignments!

Resources

- Class website: <https://systems.cs.columbia.edu/private-systems-class/>.
- EdStem (will set up and invite soon).
- Homework instructions and HW1 Colab are available on Courseworks/Files and will be reviewed next time by CA.

Staff

- **CA: Pierre Tholoniati**
 - Ph.D. student working on privacy-enhancing systems.
 - Very familiar with most of the technologies we will discuss.
 - OH: See class web page.

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.](#)”