Privacy-Preserving Natural Language Processing

Lecture 10 : Real-World DP Application and the Epsilon Registry

Dr. Erion Çano

July 17, 2025

www.trusthlt.org
Chair of Trustworthy Human Language Technologies (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

Tech Giants Applying DP

1 Tech Giants Applying DP

2 Other Organizations Applying DP

3 Standardizations and the Epsilon Registry

Apple User Data from macOS and iOS

Tons of user data being collected by Apple:

- Typing information for QuickType suggestions
- Emojis for emoji suggestions
- Web navigation data from Safari

Apple uses local DP with different ε values to protect user privacy in those data.

Full details are reported *here*.

Meta and Facebook User Interactions

S. Messing et al. (2020). **Facebook Privacy-Protected Full URLs Data Set.** Version V6

Dataset of user interactions with web pages shared on FB is publicly available.

Examples:

- John shared URL foo
- Mary views post with URL bar

Meta uses (ε, δ) DP to protect each interaction.

Microsoft and Windows Telemetry

User data collections starting (massively) with Windows 10.

B. Ding, J. (Kulkarni, and S. Yekhanin (Dec. 2017). **“Collecting Telemetry Data Privately”.** In: *Advances in Neural Information Processing Systems 30*

Examples:

- Application crash reports.
- User time spent on certain apps.
- Typing, location, connected devices, etc.

Microsoft uses local DP with $\varepsilon = 1.67$ to protect it before using / sharing / selling.

Google Shopping and Trends

Product page view counts for prioritizing page crawing.

- Collected from users on daily basis.
- Streammed using DP-SQLP
- Protected using (ε, δ) DP with $\varepsilon = 1$ and $\delta = 10^{-9}$

Search trends used to show related queries.

- Streammed using DP-SQLP
- Protected using (ε, δ) DP with $\varepsilon = 2$ and $\delta = 10^{-10}$

Linkedin Audience Engagement

Interactive query system that allows marketers to get information about LinkedIn users engaging with their content.

R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad (2020). **"LinkedIn's Audience Engagements API: A Privacy Preserving Data Analytics System at Scale".** In: *CoRR* abs/2002.05839. arXiv: `2002.05839`

- Each analyst sends queries to the system
- Each query returns (ε, δ) DP with $\varepsilon = 0.15$ and $\delta = 10^{-10}$
- Monthly limit (budget) for queries is $\varepsilon = 34.9$ and $\delta = 7 \cdot 10^{-9}$
- Extra measures to prevent averaging attacks

Labor Market Insights

Data from LinkedIn to measure trends in people changing their occupation.

- Companies who are hiring, using with $\varepsilon = 14.4$ and $\delta = 1.2 \cdot 10^{-9}$
- Available jobs, protecting hiring event with $\varepsilon = 14.4$ and $\delta = 1.2 \cdot 10^{-9}$
- Required skills, protecting each user's skills information with $\varepsilon = 0.3$ and $\delta = 3 \cdot 10^{-10}$
- Extra measures to prevent averaging attacks

R. Rogers, A. R. Cardoso, K. Mancuhan, A. Kaura, N. Gahlawat, N. Jain, P. Ko, and P. Ahammad (2020). **A Members First Approach to Enabling LinkedIn's Labor Market Insights at Scale.** arXiv: `2010.13981 [cs.CR]`

Other Organizations Applying DP

1 Tech Giants Applying DP

2 Other Organizations Applying DP

3 Standardizations and the Epsilon Registry

USA Census Bureau

County Business Patterns

- Business establishments in the USA
- Different types of data (including finances)
- Protected with various (ε, δ) parameters based on:
  - Financial assets
  - Annual payroll
  - Number of employs

Full details are reported *here*.

USA Census Bureau

The 2020 USA Census

- Collecting demographic information about USA population
- Datasets protected with various (ε, δ) parameters
  - *Redistricting* data protected with $\varepsilon = 13.64$
  - *Demographic Housing* protected with $\varepsilon = 19.46$
  - Racial and ethnical categories protected with $\varepsilon = 45.68$
  - Suplemental household statistics protected with $\varepsilon = 12.74$

USA Census Bureau

Post-Secondary Employment Outcomes

- Earnings and employment of college graduates
- Categorized by degree level, degree major, and post-secondary institution
- Matching university granscript data with national database of jobs
- Each person protected with $\varepsilon = 1.5$

Full details are reported *here*.

Wikimedia Foundation

Page view statistics

T. Adeleye et al. (2023). **Publishing Wikipedia usage data with strong privacy guarantees.** arXiv: `2308.16298 [cs.CR]`

- Distinct users visiting Wiki pages each day from each country
- July 1, 2015 to Feb. 8, 2017 protected with $\varepsilon = 1$ on 300 pages view per day
- Feb. 9, 2017 to Feb. 5, 2023 protected with $\varepsilon = 1$ on 30 pages view per day
- Feb. 6, 2023 onwards protected with $\varepsilon = 0.72$ and $\delta = 10^{-5}$

Wikimedia Foundation

Editor statistics

- Statistics about editor activity by project and country
- Monthly published data protected with $\varepsilon = 2$ on editor-project-country-month
- Weekly published data protected with $\varepsilon = 2$ on editor-project-country-week
- One-off release for Russian editors, protected with $\varepsilon = 0.1$

Other Applications

- Brave uses distributed DP for collecting user analytics
- Spectus published a dashboard with mobility trends during Hurricane Irma
- Microsoft Assistive AI collects user data about Office Tools and suggests automatic replies
- Google uses DP and federated learning for collecting data and training models for improved text selection and copying on Android

Standardizations and the Epsilon Registry

1 Tech Giants Applying DP

2 Other Organizations Applying DP

3 Standardizations and the Epsilon Registry

Background: Evolution of Information Security

Early days, until the ’60s

- Mostly physical security: focus on access to mainframes
- Limited threats: computers were rare and isolated
- Computers were mostly used for scientific computations; little or no personal data involved
- Basic enkryption: simple ciphers like Caesar sipher were used to ensure message confidentiality

Background: Evolution of Information Security

Rise of the networks, ’60s - ’90s

- Computer networks and remote access become common
- Unauthorized network / system access and data breaches become popular concerns
- Cybersecurity emerges as an important discipline
- Many remote access protocols with encryption introduced
- Popular symetric cyphers of 64 bits like DES
- Public key cryptography comes out

Background: Evolution of Information Security

The internert era, ’00s - present

- Widespread internet usage, online shopping, personal data sharing
- Surge in cybercrime, viruses, worms, ransomware, phishing, DDOS attacks
- Evolution of security measures: VPNs, firewalls, antiviruses, intrusion detection systems
- Evolution of cryptography: 128 bit cyphers like AES become mandatory

Information Security Standards

- **ISO/IEC 27001**: Standards for establishing, implementing, maintaining and improving an ISMS
- **ISO/IEC 27002**: provides a code of practice for information security controls, offering guidance for implementing security policies
- **ISO/IEC 27005**: focuses on information security risk management, helping organizations in assessing and managing risks

Information Privacy vs. Security

- Privacy is inherently a more complex concept: who, what, when, where, why
- Private information is highly context dependent which makes it difficult to protect
- Lack of motivation (unless enforced by law) to protect privacy because of utility (profitability) loss
- Can privacy evolution and standardization follow the same path as security?

The Epsilon Registry

- DP is being successfully implemented in industry, public sector, academia, etc.

- Still, little understanding of the optimal ε values for certain systems, purposes, data types, etc.

- No clear consensus among practicioneers about:
  - How to approach the key implementation decisions
  - How to decide about the right privacy guarantees
  - How to choose ε levels

C. Dwork, N. Kohli, and D. Mulligan (Oct. 2019). **"Differential Privacy in Practice: Expose your Epsilons!"** In: *Journal of Privacy and Confidentiality* 9.2. priv-0

The Epsilon Registry

C. Dwork, N. Kohli, and D. Mulligan (Oct. 2019). **“Differential Privacy in Practice: Expose your Epsilons!”** In: *Journal of Privacy and Confidentiality* 9.2. priv-0

Could a public Epsilon Registry help?

- Communal body of knowledge about DP implementations
- Used by stakeholders, public servants, academics, etc.
- Serve as guideline for identifying and adopting judicious DP implementations

Questions...?

THANK YOU...!