Algorithmic approaches to privacy at WMF

Hal Triedman — Privacy Engineering Intern

Table of Contents

Privacy Policy vs. Open Access Policy

Privacy Policy? Or Open Access Policy?

Privacy policy and data retention guidelines:

- **Minimize harms of collecting user data**
- Clear guidelines around retaining data
- Anti-surveillance
- “Lean data diet”

Open access policy:

- As much transparency as possible
- Recognition that WMF controls resources that lots of people regularly access
- **Releasing more data could conceivably allow us to better understand the internet**

The stakes are high, because Wikipedia is inherently political — users and editors are pseudonymous for a very good reason

What does “privacy” even mean?

(... it’s complicated, and depends on who you ask)

What does “privacy” even mean?

Some potential definitions (from least to most technical):

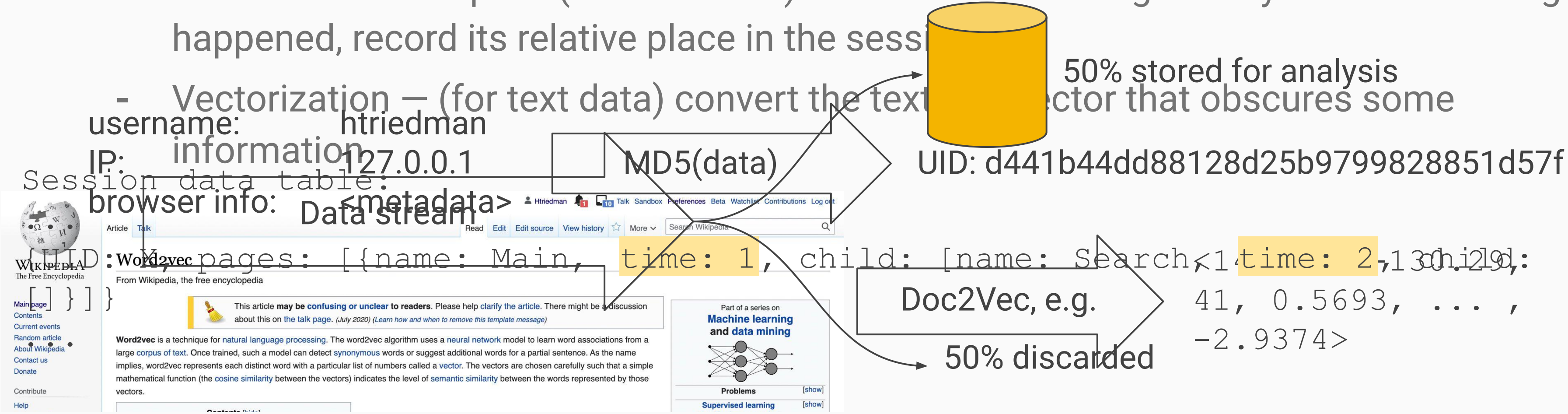
- Privacy is a vibe — you know it when you feel it
- US legal privacy — freedom from state search without a warrant/CCPA
- Bits of entropy — how much does a given piece of information identify you?
- Differential privacy — we’ll come back to this later

What are the tools in our privacy toolbox?

General Approaches

General Approaches

- Pseudonymization — hashing usernames/IPs
- Filtering — only save a percentage of the data for analysis purposes
- Relative timestamps — (for a session) rather than recording exactly when something happened, record its relative place in the sessi
- Vectorization — (for text data) convert the text ... ector that obscures some information



K-anonymity

K-anonymity: Definition

For some group size *K*, selectively aggregate fields in your database so that no subgroup is smaller than *K*

K-anonymity: Example

UID	Gender	Age	Location
1	F	29	Barcelona
2	M	64	Houston
3	M	54	DC
4	F	18	Berlin
5	F	44	Boston
6	M	66	London
7	F	38	SF
8	M	75	Rome

2-anonymity

UID	Gender	Age	Location
1	F	18-29	Europe
2	M	50-64	US
3	M	50-64	US
4	F	18-29	Europe
5	F	30-49	US
6	M	65+	Europe
7	F	30-49	US
8	M	65+	Europe

K-anonymity: Problems

- K-anonymity is an NP-hard problem — computationally intractable with large datasets
  - (or you can just sorta eyeball it)
- Same issue as with bits of entropy — what if a government is comfortable arresting *K* people?
- Vulnerable to re-identification attacks — can link this database with outside knowledge (a la Latanya Sweeney) to break privacy

Differential Privacy

Differential Privacy: Definition

- Imagine we had two databases, X and X’, owned by the **database owner**
- X and X’ are comprised of **entries**
- X and X’ are **adjacent** if they differ by **one and only one entry**
- An **analyst** can can send queries to the **database owner**

Differential Privacy: Definition

- Differential privacy: a **promise** between the database owner and participants who contribute entries:

"From the perspective of the analyst, your participation in this database will be completely hidden. Population-level information can be extracted, but no one will be able to infer your presence or absence (even if you're an outlier)."

Differential Privacy: Definition

Let ε be a positive real number, $\mathcal{M}$ be an algorithm that adds noise to a dataset, R is a result in the range of $\mathcal{M}$, and X and X' be two adjacent databases. $\mathcal{M}$ is ε-differentially private if:

$$\Pr[\mathcal{M}(X) = R] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(X') = R]$$

As $\varepsilon \to 0$, the bound e^{ε} gets tighter, and we need to add more noise.

Differential Privacy: Definition

Importantly, ε serves as a measurable, quantifiable metric called *privacy loss*:

$$\ln\left(\frac{\Pr[\mathcal{M}(X) = R]}{\Pr[\mathcal{M}(X') = R]} \right) = \varepsilon$$

Differential Privacy: Definition

Impossible for differentially-private algorithms to be subject to re-identification attacks

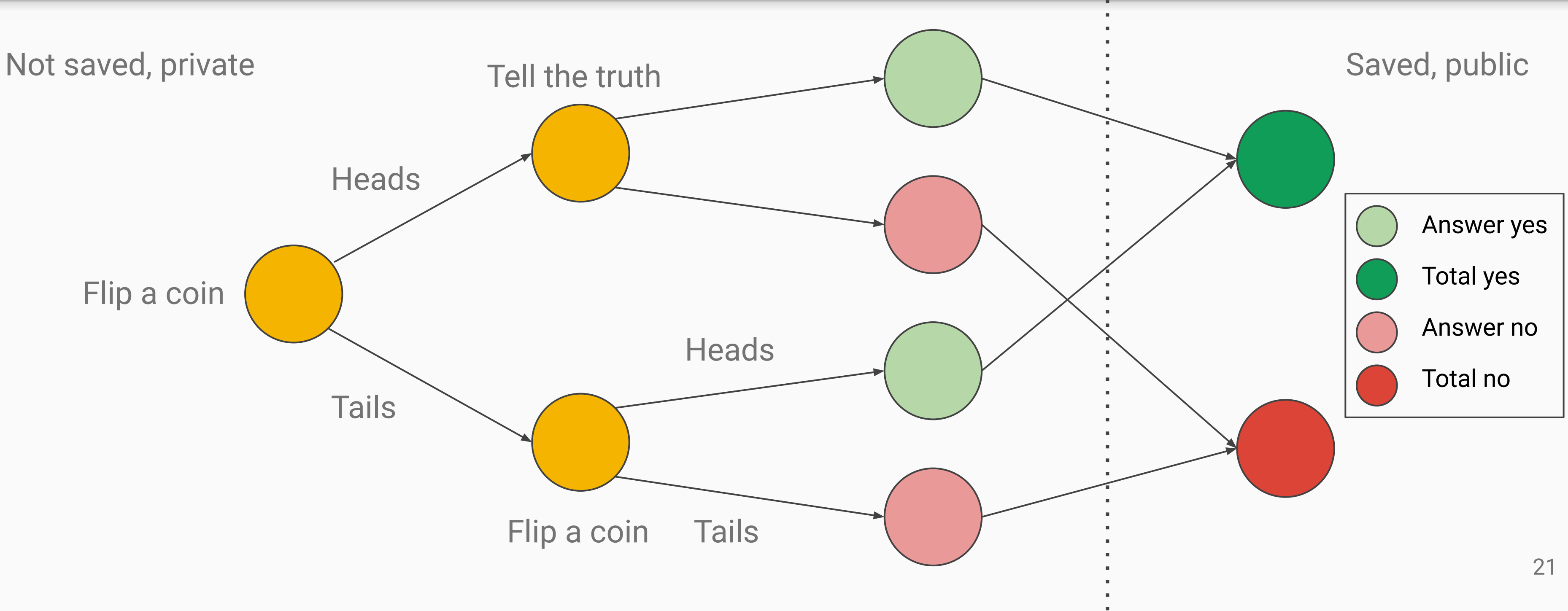
- $\mathcal{M}$ is independent from the world $\rightarrow$ data participants are protected from prior outside knowledge and protected from future outside knowledge

Differential Privacy: Randomized Response

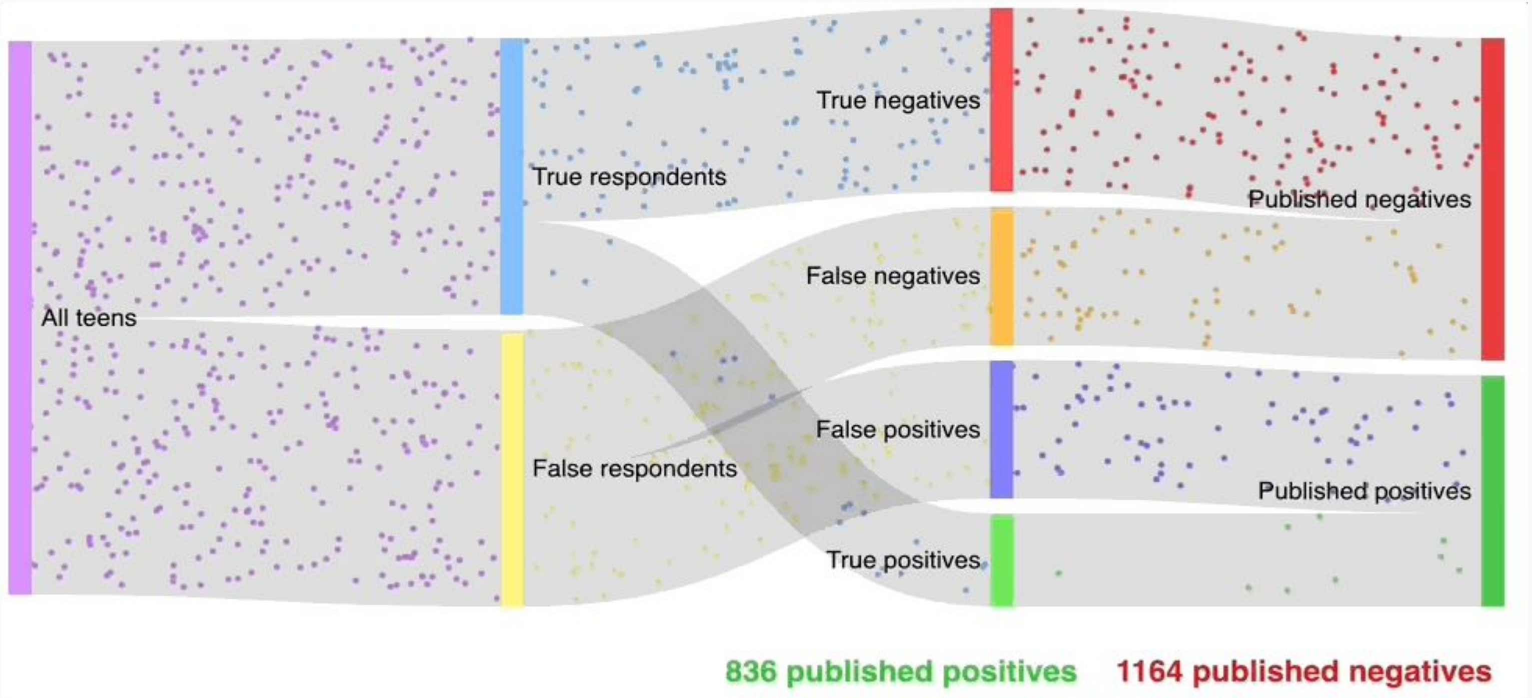
Let's imagine you were conducting a poll about teen drinking:

- NIH estimates 33.6% of teens have consumed alcohol in the past month
- The behavior is illegal/taboo → respondents might not answer truthfully
- Probabilistically add noise to the number
    - Not an accurate count, but you can track changes proportionally over time
    - If methodology is published, can work backwards to derive estimates

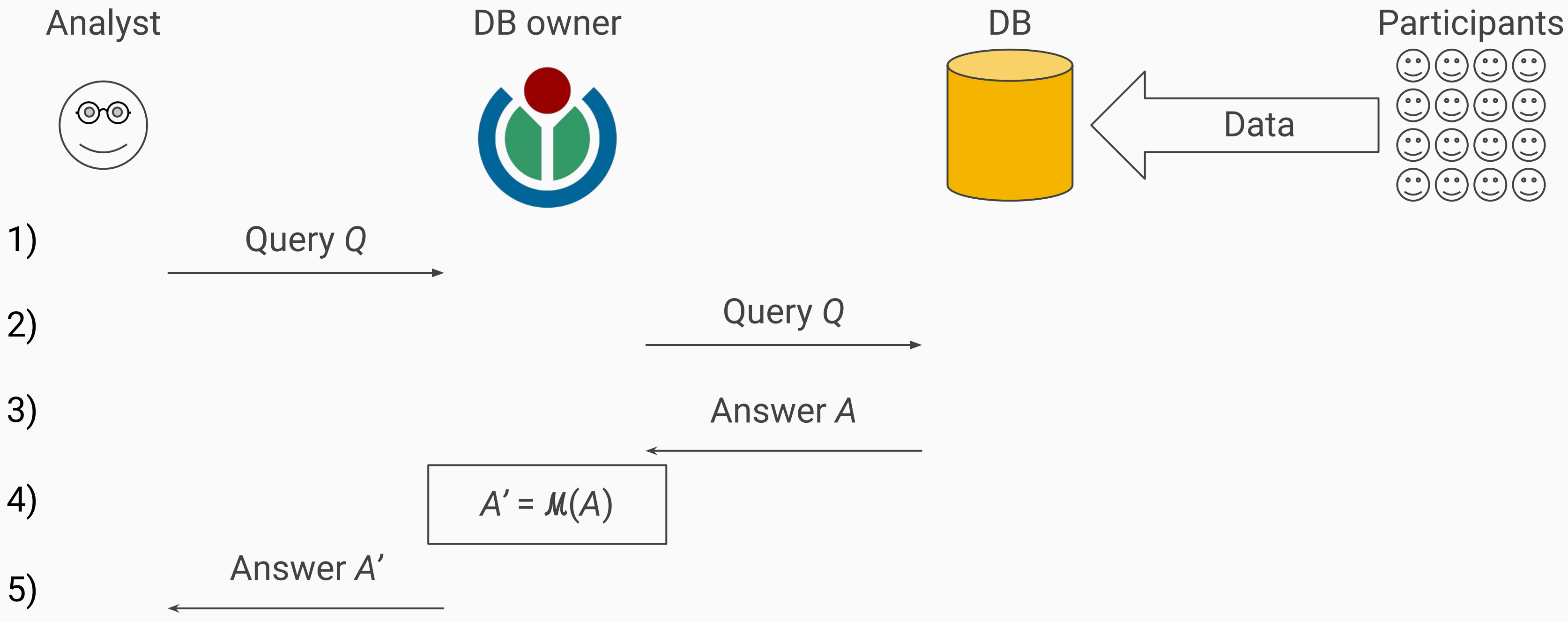
Differential Privacy: Randomized Response



Differential Privacy: Randomized Response



Differential Privacy: Real World



Differential Privacy: Limitations

- Counts are not exactly accurate
- Privacy loss is measurable (ε), but every time the DB owner answers a query, it increases
- Lots of these parameters are very abstract — how should we communicate with less- or non-technical community members about them?

Future Directions

Future Directions: Granular Pageview Data

Currently working to privately release (page, language, country, views) tuples:

- MediaWiki API currently has (page, language, views) and (page, country, views), but filtering by both could leak a lot of information
- Increase knowledge for editors in multilingual/less-connected places
  - e.g. editors in India, Anglophone/Francophone Africa, Vietnamese speakers in the US
- Disaggregate country-level trends
  - e.g. 2021 storming of the US Capitol was a top-10 enwiki article for two weeks in January, but probably not in South Africa

Beta version of this available at https://diff-privacy-beam.wmcloud.org

Future Directions: Session Data

Could combine differential privacy with existing Wikipedia ontologies to anonymize and release some form of session data

- Maybe a knowledge graph based on browsing history?

Questions?