Differentially Private Topic Modelling

Mapping Privacy Risks & Prototyping Secure Algorithms

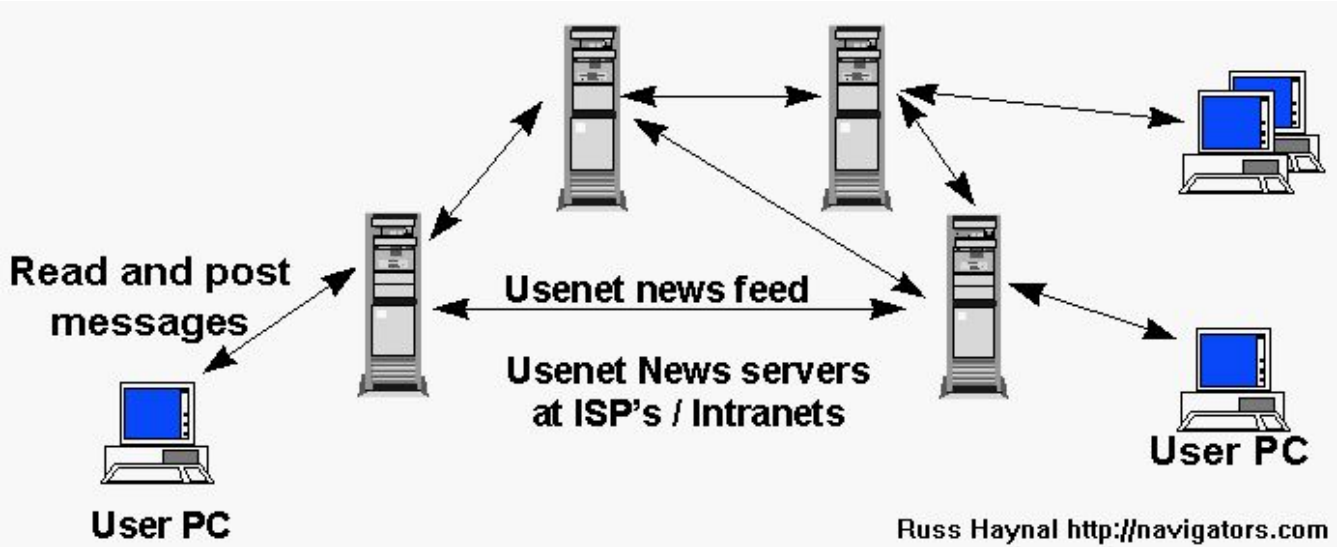
PPOL 6801 - **Muhammad Saad**

Project Overview

- **Latent Dirichlet Allocation** (LDA) used for text data applications including NLP
- **Uncover latent themes within a corpus** (*unsupervised*); documents as a mixture of topics inferred from their word distributions
- Are there **any privacy risks** associated with LDA?
- Can we implement **any privacy preserving techniques**?

Data Overview(I)

20 Newsgroup text dataset



Discussion groups on topic-specific forums called newsgroups

Decentralized microblogging platform before FB, Twitter, etc (early 1990s)

Data Overview(II)

Useful from a **privacy perspective**

```
Rows: 18,846
Columns: 5
$ text    <chr> "From: Mamatha Devineni Ratnam <mr47+@andrew.cmu.edu>\nSubject: Pens fans reactions…
$ group   <chr> "rec.sport.hockey", "comp.sys.ibm.pc.hardware", "talk.politics.mideast", "comp.sys.…
$ subject <chr> "Pens fans reactions", "Which high-performance VLB video card?", "Re: ARMENIA SAYS …
$ from    <chr> "Mamatha Devineni Ratnam <mr47+@andrew.cmu.edu>", "mblawson@midway.ecn.uoknor.edu (…
$ email   <chr> "mr47+@andrew.cmu.edu", "mblawson@midway.ecn.uoknor.edu", "hilmi-er@dsv.su.se", "gu…
```

Includes **personal identifiers**

Privacy risks in text-based machine learning models

How can we have a publicly available text dataset with identifiers??

Apparently, identifiers are '*mostly obsolete*' - ChatGPT

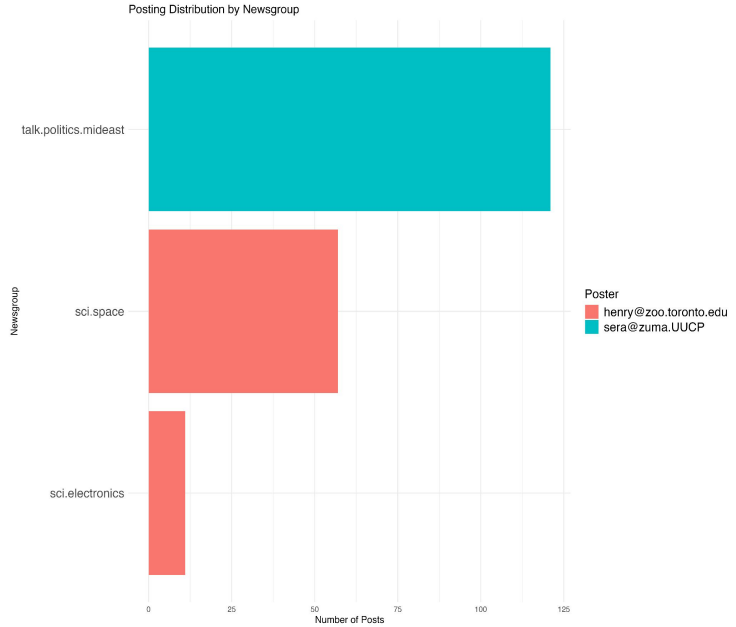
Data Overview(III)

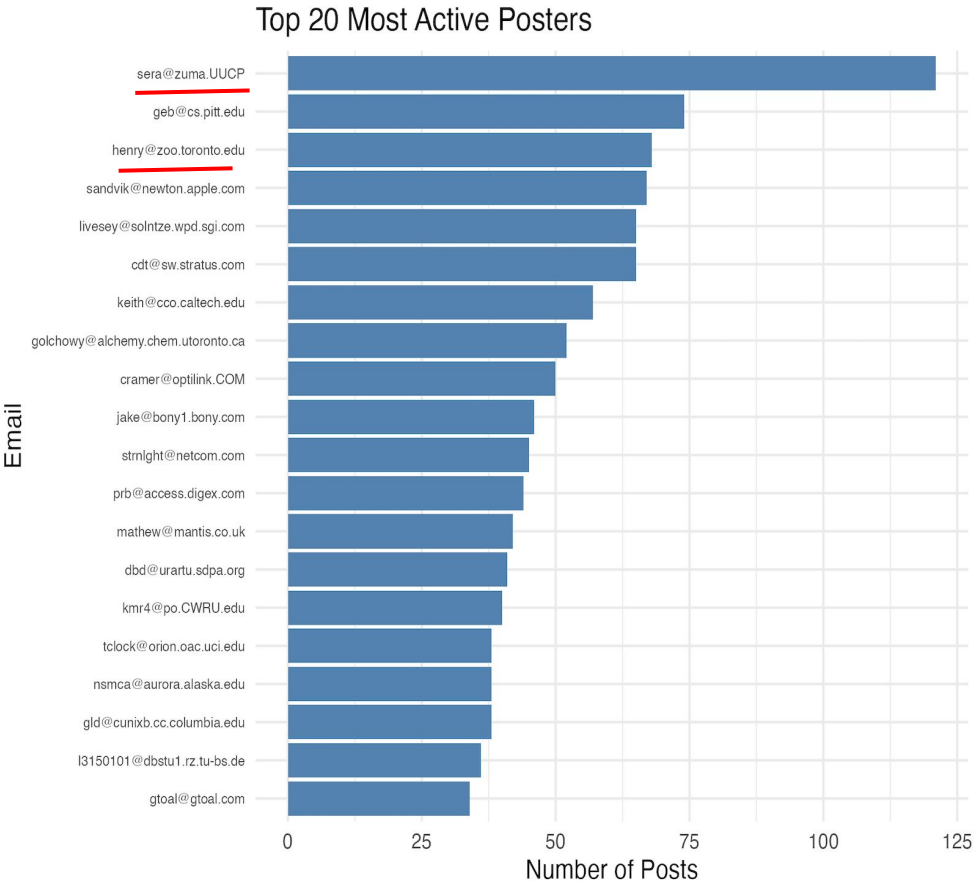
Total Documents: 18846

Users with Identifiable Emails: 8518

Max Posts by One User: 121

Median Posts per User: 1





Privacy Risks (I)

Google searched two of the top posters

sera@zuma.UUCP

ODTÜ Middle East Technical University
METU Research Information System

Prof. AHMET COŞAR

Faculty of Engineering, Department of Computer Engineering

WoS Research Areas: Computer Science, Engineering Computing & Technology (Eng)

Avesis Research Areas: Information Systems, Communication and Control Engineering, Communication Engineering, Computer Sciences, Engineering and Technology

Figure source.

ARMENIAPEDIA
A WIKI ENCYCLOPEDIA

'*Serdar Argic*' - spambot

henry@zoo.toronto.edu

Henry Spencer

Spencer is a Canadian programmer and space enthusiast who created three widely used, adapted, and influential regular expression libraries. In 1986, he was the first to release a regex library which could be freely included in other programs. Perl 2's regex package was based on and enhanced from Spencer's library, but Spencer's technological *tour de force* was creating the regex package used by Tcl. This implementation, Jeffrey Friedl writes, "is a hybrid [NFA/DFA engine] with the best of both worlds".

Websites: Wikipedia, O'Reilly bio, Lysator, Bio at NASA *(photo source)*

Figure source.

Privacy Risks (II)

Posts by top posters

sera@zuma.UUCP

'Muslim women and children were massacred by the Armenians.'

'Nazi Armenian Philosophy: Race above everything and before everything.'

henry@zoo.toronto.edu

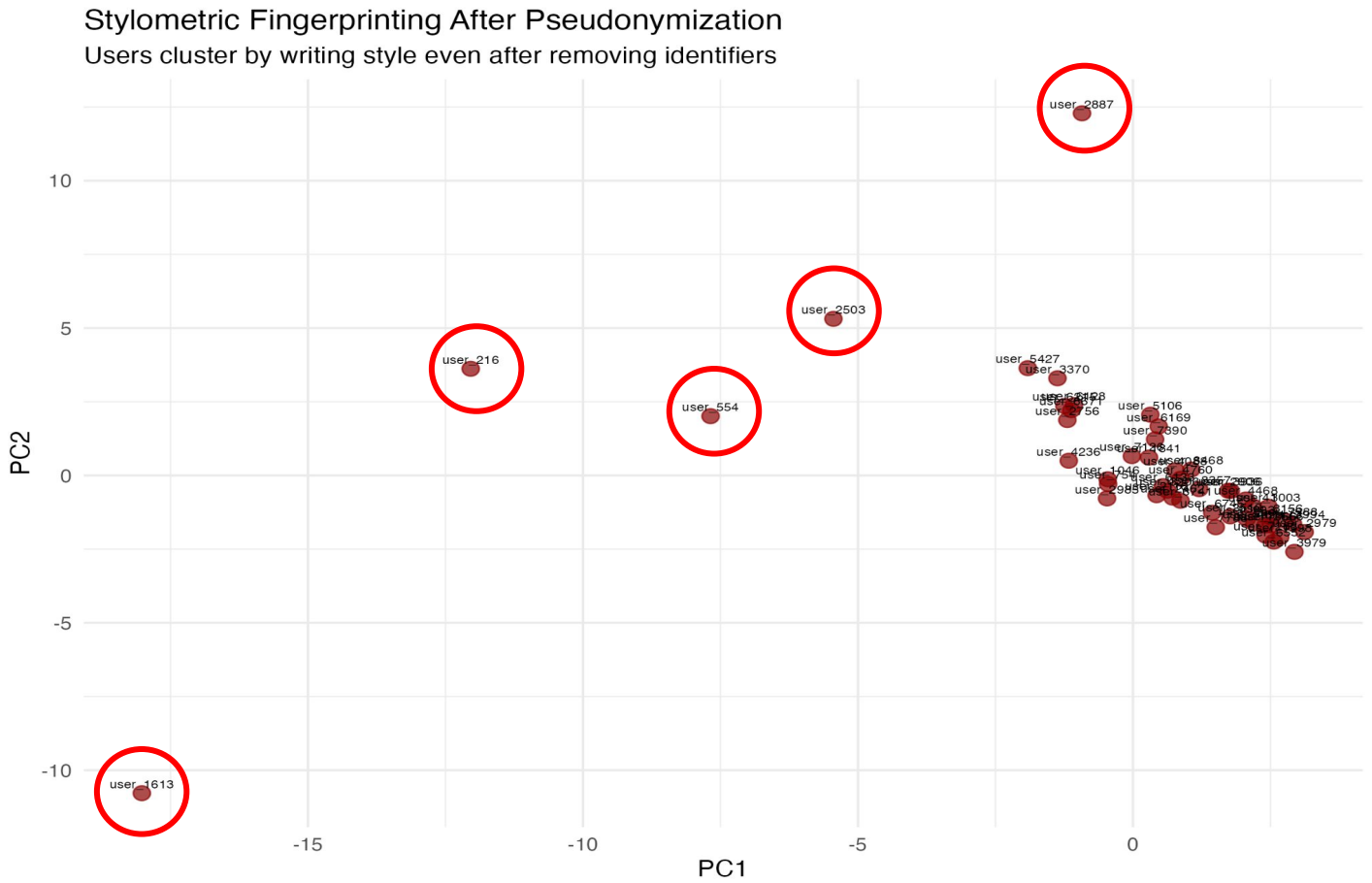
'We're repairing something with a 74ACT00 on it and the question arises, "well, do i really need the ACT part?"'

'That is new stuff for me. So it means that you just can not put a satellite around around the Moon for too long'

Each user leaves a unique, potentially identifiable digital fingerprint!

Pseudonymization Risks

Risk of re-identification based on unique stylistic contributions to text data (digital fingerprint)



What about Topic Modelling?

Latent Dirichlet Allocation (LDA) topic modelling technique

Identify **structure of documents**; assign each document a **distribution of labels**

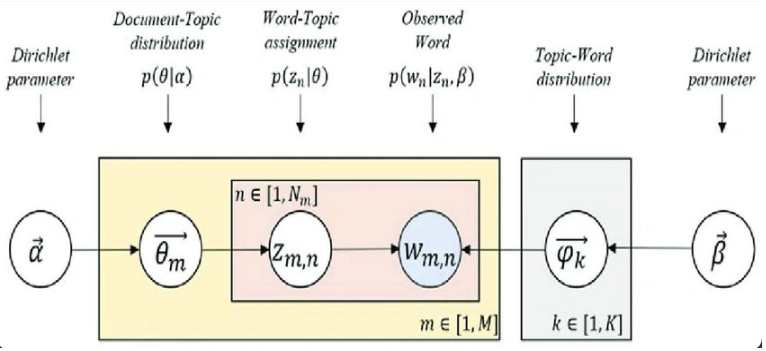


Figure source

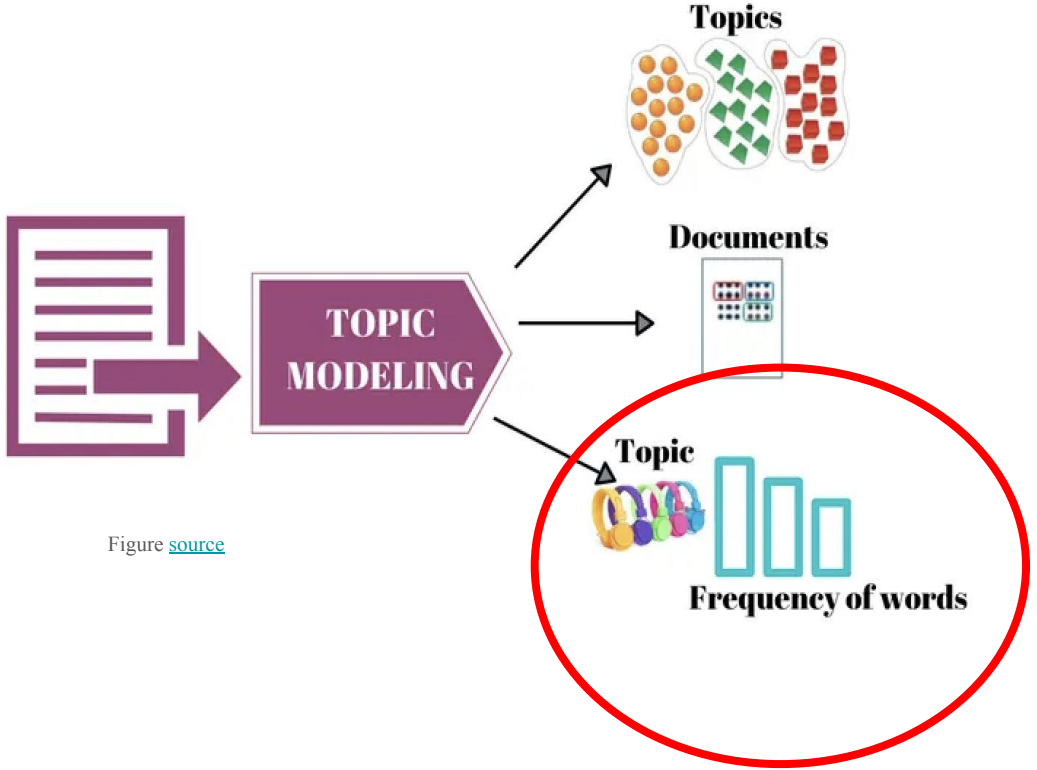
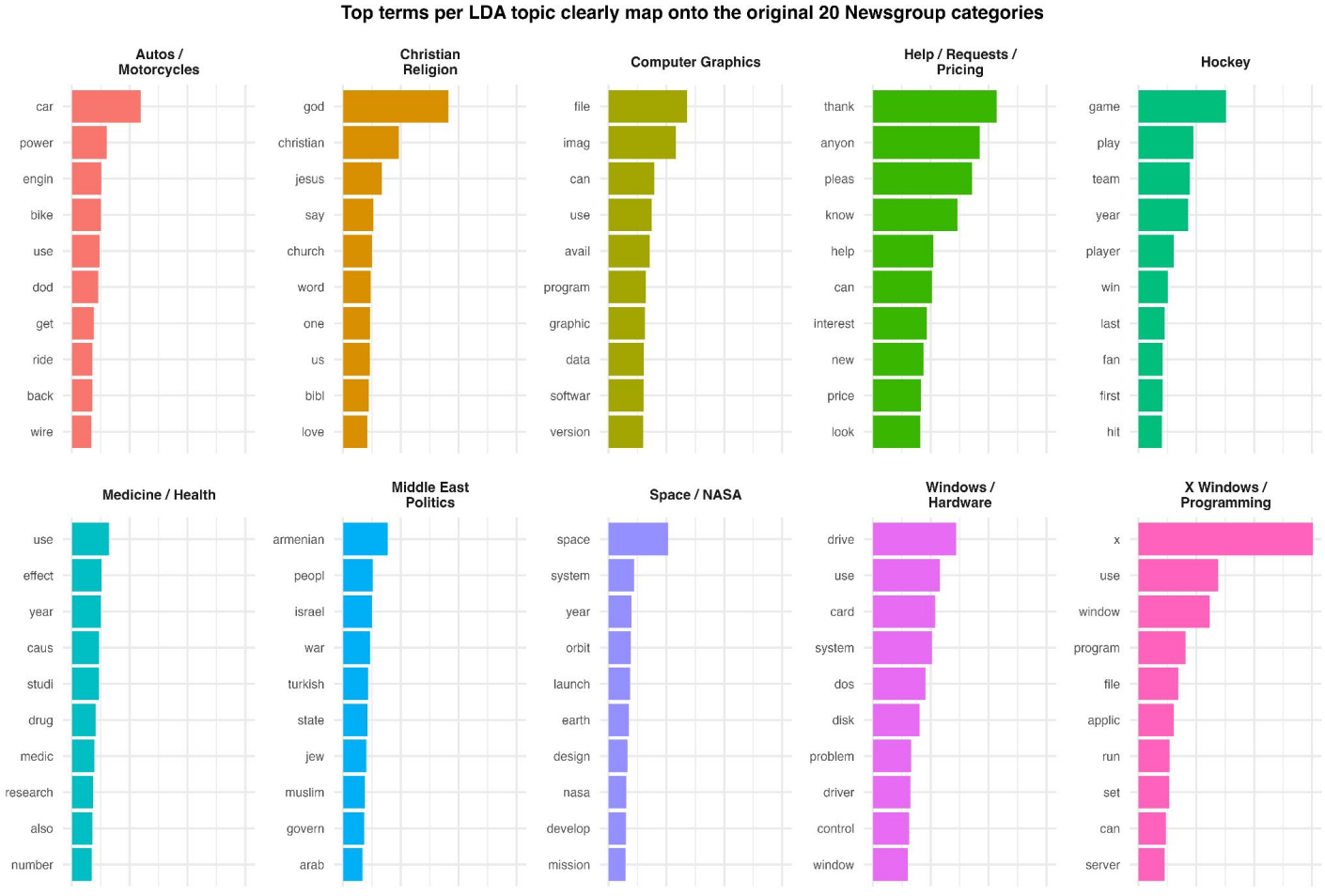


Figure source

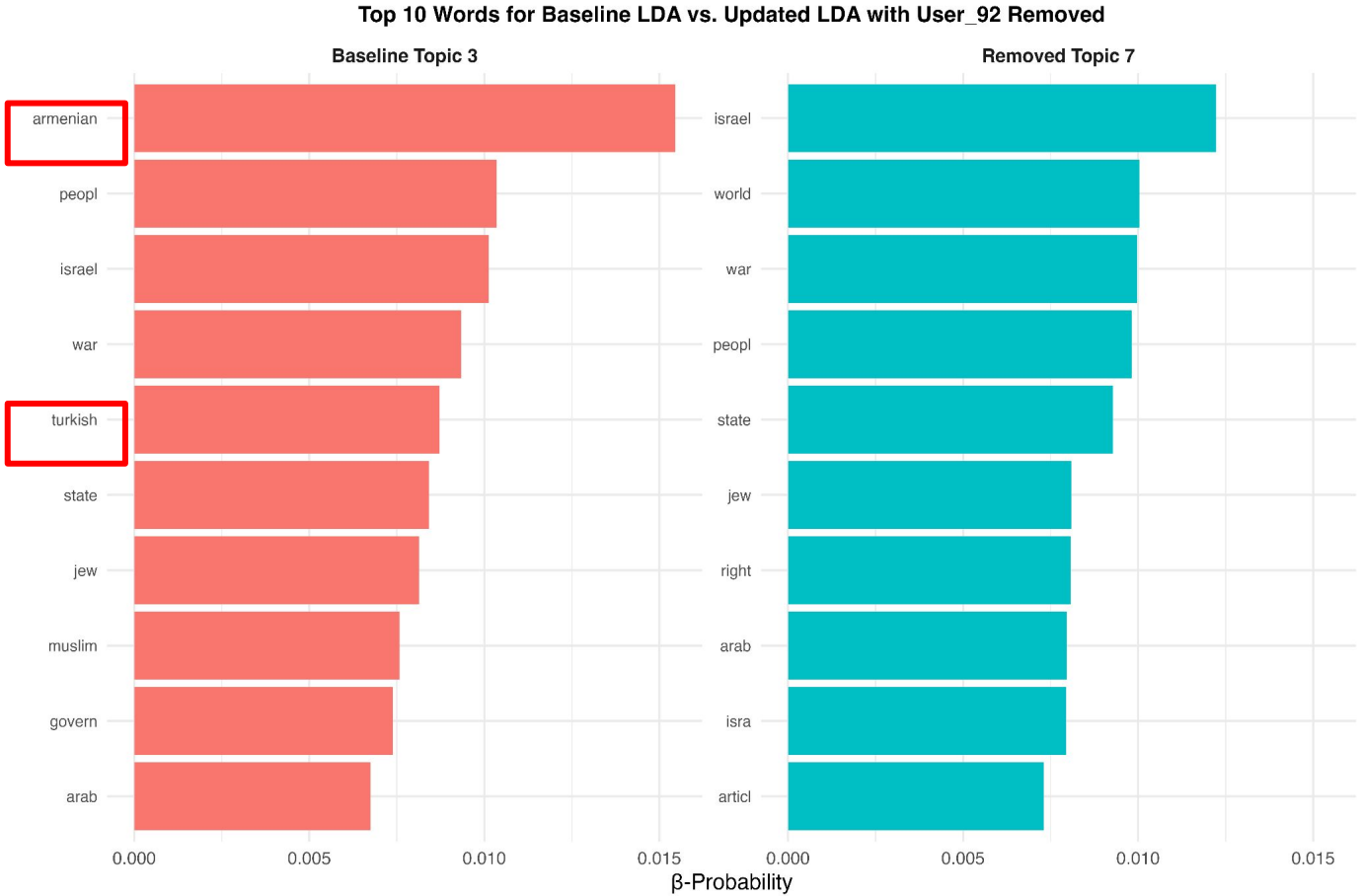
20 Newsgroup LDA

Ran LDA based on Gibbs sampling; k = 20 to mimic original groups



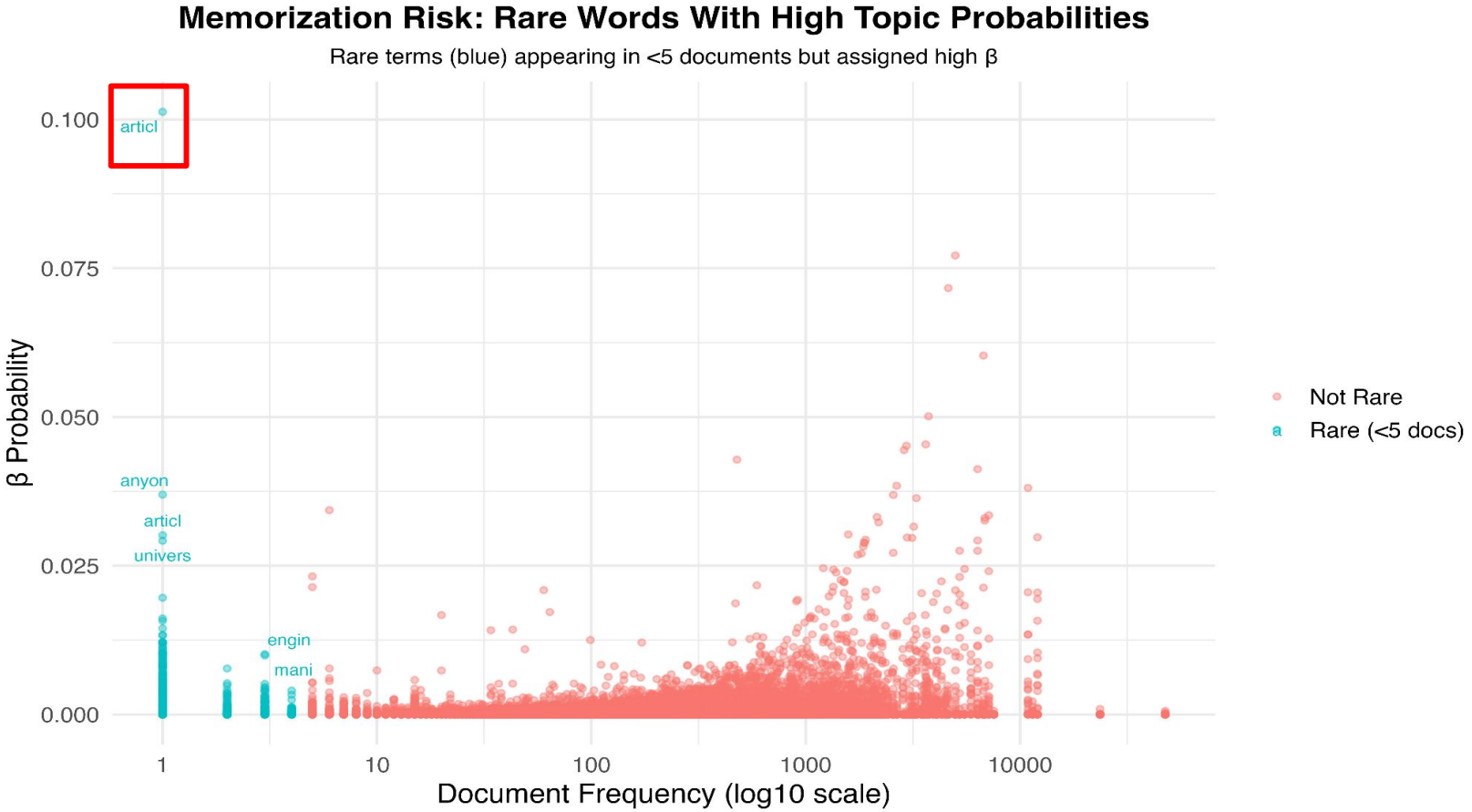
20 Newsgroup LDA - Memorization Risk (I)

If one user's data is removed, especially if that user is an outlier, it significantly affects LDA outputs, which increases privacy risks



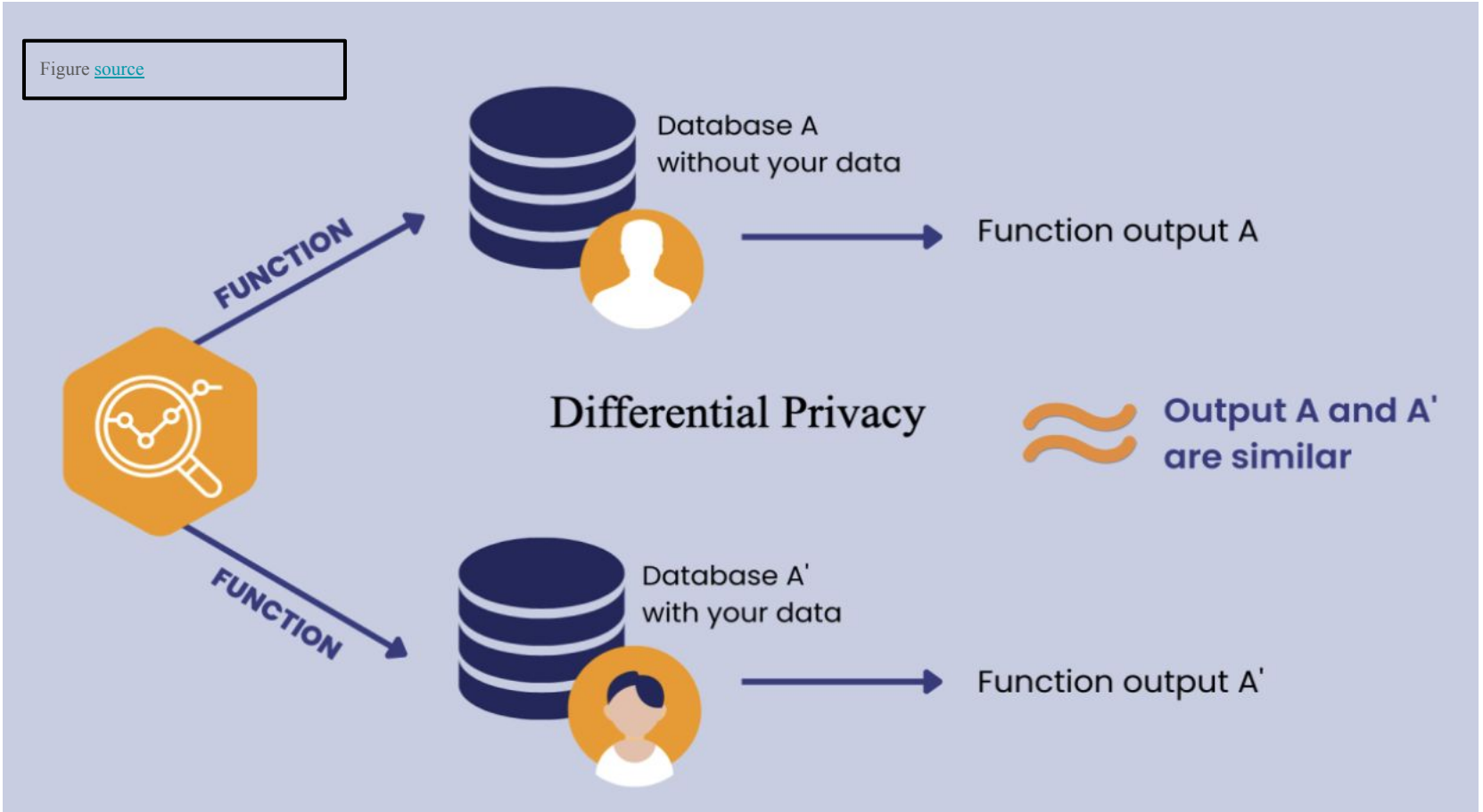
20 Newsgroup LDA - Memorization Risk (II)

An attacker who only sees the model outputs can infer rare term exists in the training corpus; strongly tied to a particular topic, possibly **one particular contributor** (**singling out!**)



Differential Privacy

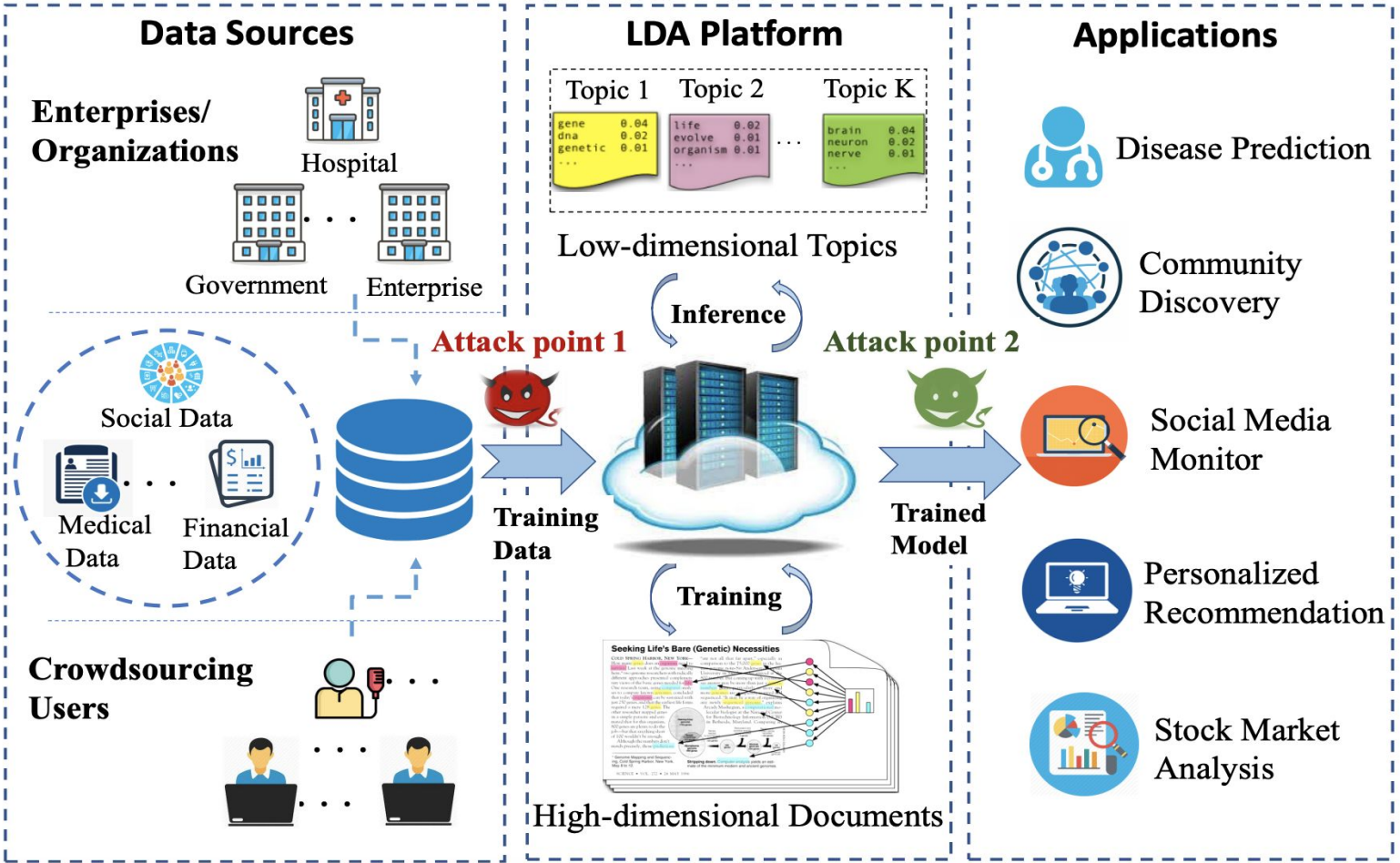
Mathematical definition of privacy; helps us induce calibrated noise in statistical outputs to mask individual contributions



Privacy Risks in LDA

Two attack points: *training*; *model output*. These are the two points where we can inject noise as well.

Figure source



Differentially Private LDA - Project Approach

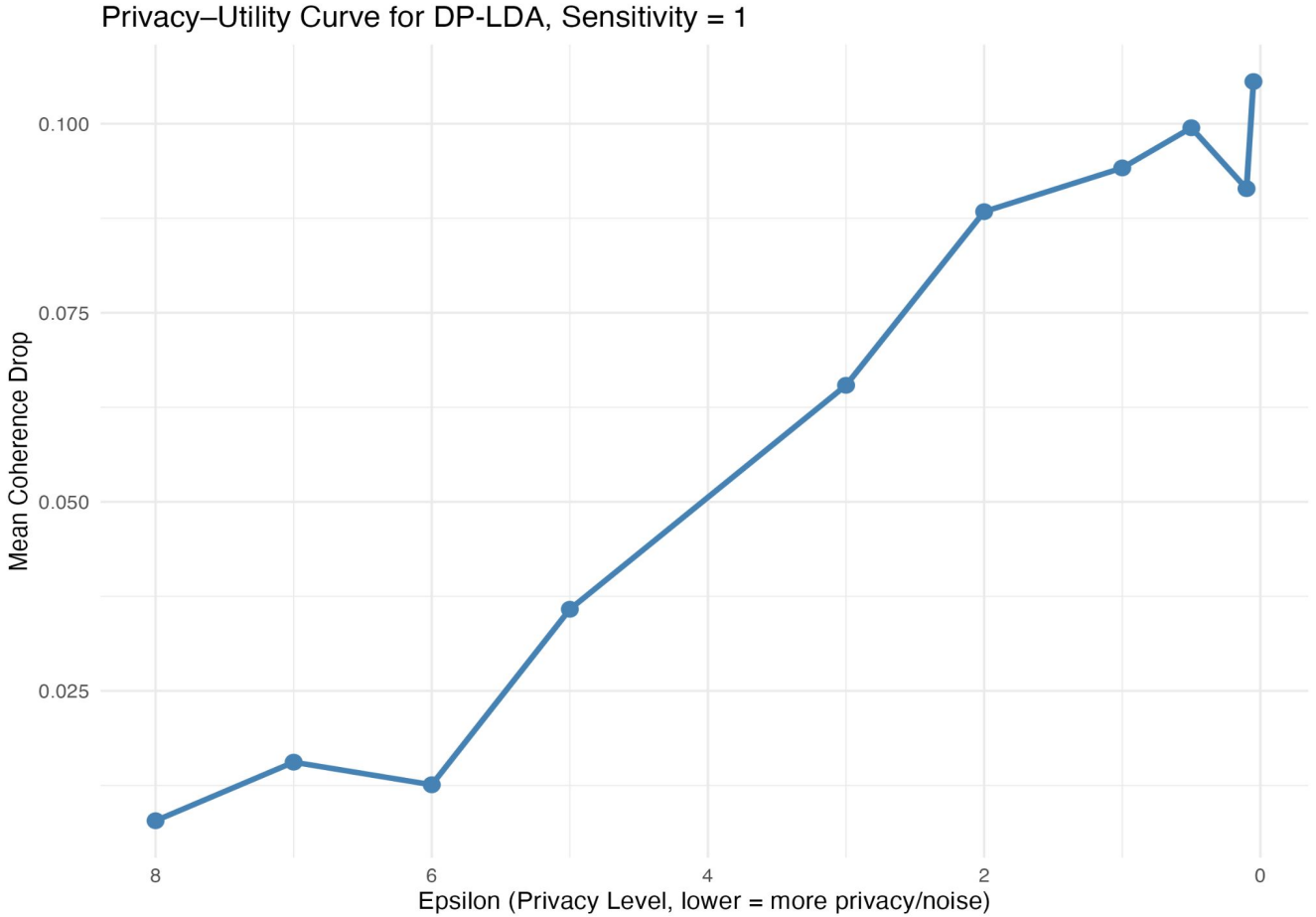
Multiple points where you can inject calibrated noise to perturb statistical outputs and mask individual contributions

Authors (Year)	Method Feature	Privacy Level / Mechanism
Zhao et al. (2020)	DP-LDA via Collapsed Gibbs Sampling	Adds Laplace noise to word–topic counts at each iteration; word-level, Laplace mechanism
Huang & Chen (2021)	Subsampled Laplace DP (SUB-LDA)	Poisson subsampling before noise addition; word-level, Laplace + subsampling
Huang et al. (2022)	Rényi-DP LDA	Replaces Laplace with Gaussian mechanism; word-level, Gaussian mechanism
Manzonelli et al. (2024)	Vocabulary-level DP (DP Set Union)	DP applied to vocabulary construction + training; vocabulary-level, DP Set Union
Saad (2025)	Statistics-Perturbation DP-LDA	Laplace noise added to topic–word (n_kw) and document–topic (n_dk) counts; count-level, Laplace mechanism

Injected noise directly into the ***document–topic*** + ***topic–word*** counts
Offered the most straightforward way to achieve differential privacy in Gibbs LDA
Simplistic application while still providing privacy guarantees

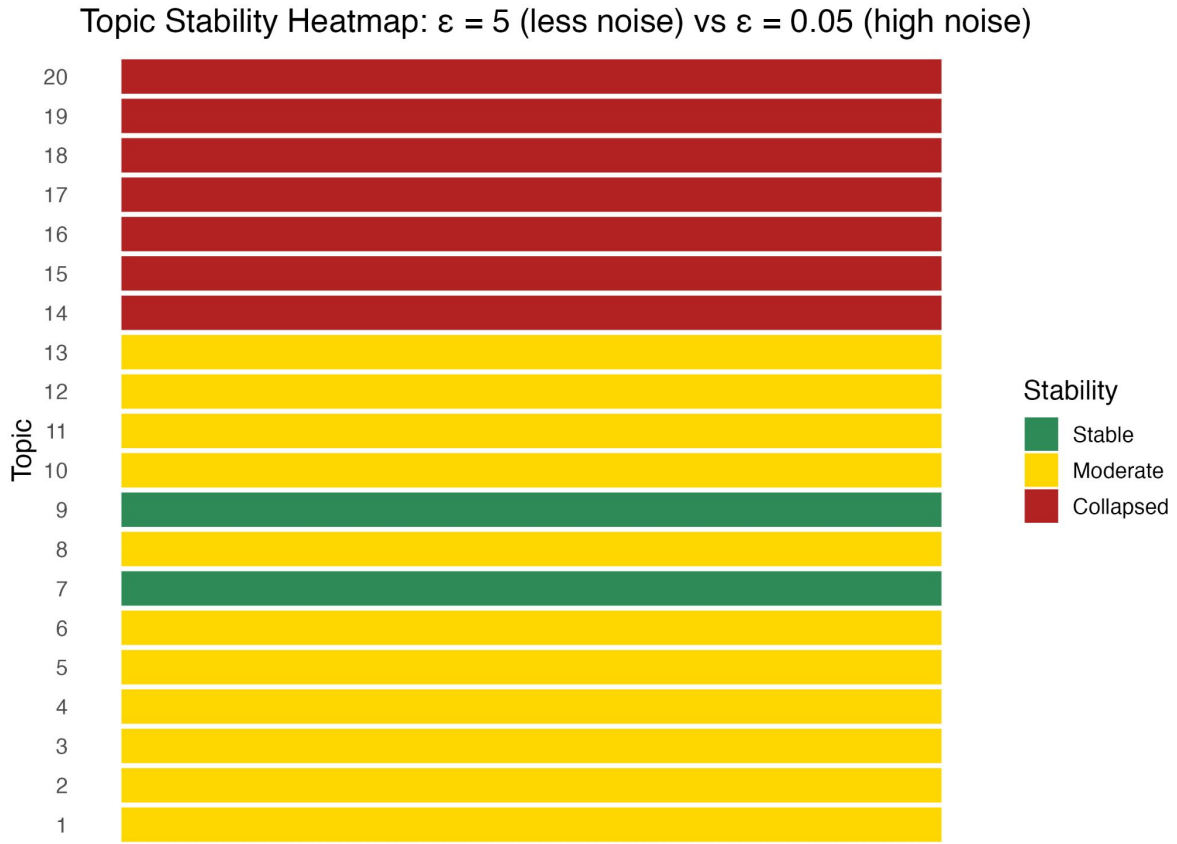
DP-LDA: Project Approach (Document-Level) (I)

Assume that addition/removal on one document in corpus can only change the topic-word count by one.



DP-LDA: Project Approach (Document-Level) (II)

Assume that addition/removal on one document in corpus can only change the topic-word count by one.

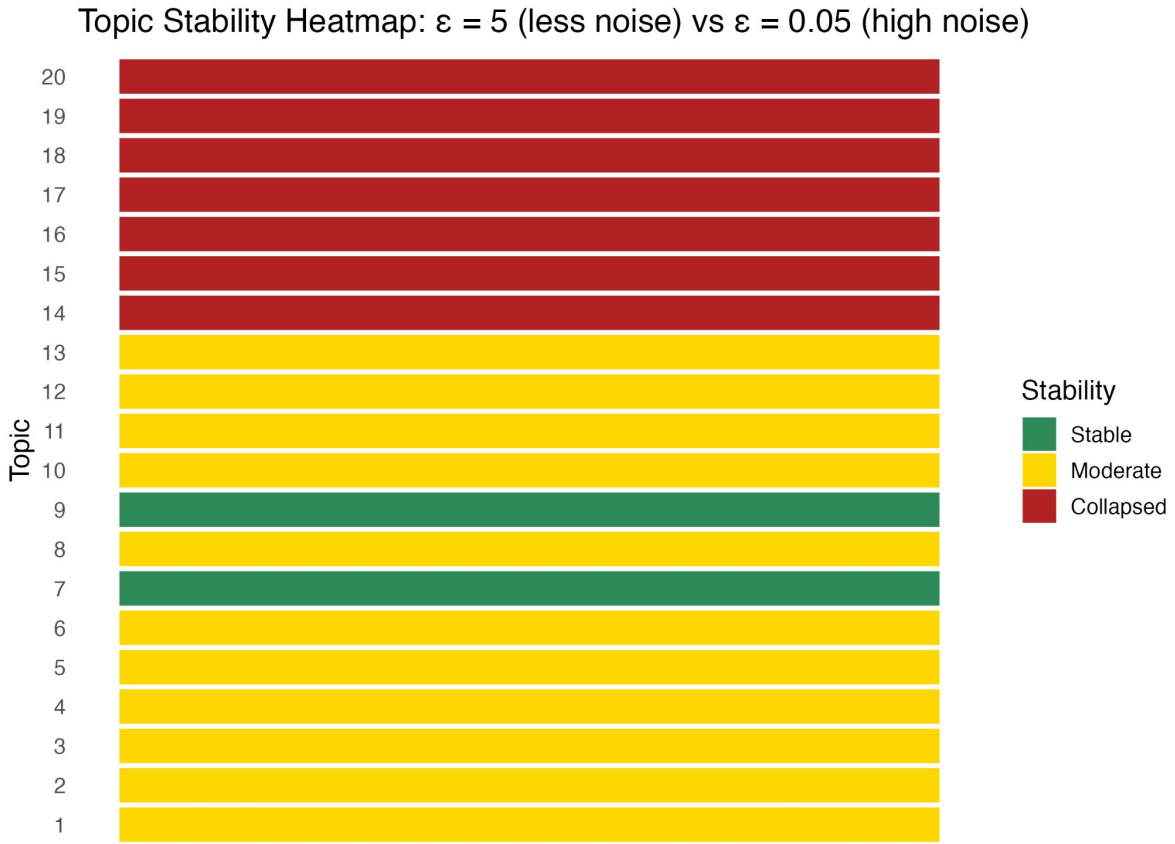


DP-LDA: Project Approach (Document-Level) (II)

Assume that addition/removal on one document in corpus can only change the topic-word count by one.

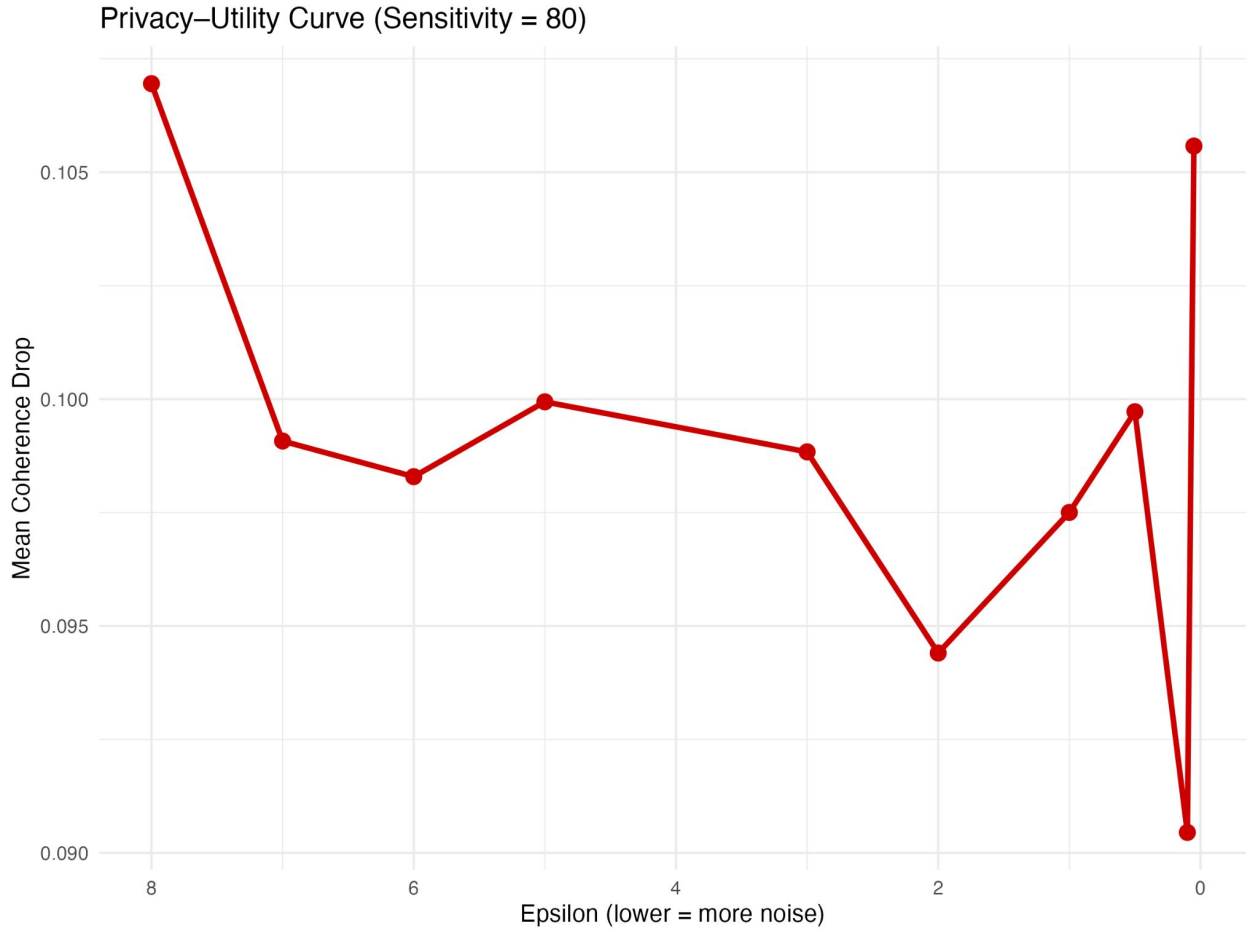
??

As much as 121 posts!



DP-LDA: Project Approach (User-Level) (III)

Set sensitivity = 80 to mask contribution by outlier user



DP-LDA: Project Approach (User-Level) (IV)

Next Steps: **User Level Privacy**

- Look at the distribution: number of posts by user
- Based on L1 norm histogram, look for a reasonable cut-off (ideally around 10 or as close as to 95th percentile)
- Clip document contributions at max of this cutoff

Questions!

‘Randomness is the protector of secrets!’