Combatting the Challenges of Local Privacy for Distributional Semantics with Compression

Alexandra Schofield
Harvey Mudd College
xanda@cs.hmc.edu

Gregory Yauney
Cornell University
gyauney@cs.cornell.edu

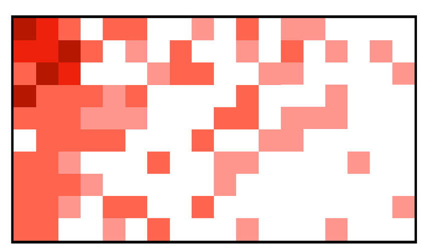
David Mimno
Cornell University
mimno@cornell.edu

Why limited-precision privacy?

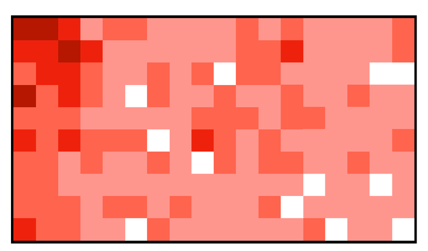
Privatizing text data is hard under local privacy.

- Bag-of-words data is: 1. high-dimensional
  2. sparse
  3. bursty
- No bound on the frequencies in each observation.
- Local setting: can't compute ℓ_1-sensitivity.

Standard mechanisms for local privacy:



Original data



Geometric noise

Isn't text just histograms? Can't we locally privatize those?

- We don't need to worry about whole documents.
- We don't want to care whether a single word shows up.

Privacy definitions

local privacy INDISTINGUISHABILITY

Consider a database D with rows in $\mathbb{R}^m$. A randomized mechanism R is ϵ**-locally private** if, for all pairs of possible rows $y, y' \in \mathbb{R}^m$, and a set of possible outputs $S \subset \mathbb{R}^m$:

$$\Pr\left[R(y) \in S\right] \leq e^{\epsilon} \cdot \Pr\left[R(y') \in S\right]$$

limited-precision local privacy (LPLP) PLAUSIBLE DENIABILITY

Consider a database D with rows in $\mathbb{R}^m$. A randomized mechanism R is **(N,ϵ)-limited-precision locally private** if, for all pairs of possible rows $y, y' \in \mathbb{R}^m$ with ℓ_1 difference $\|y - y'\|_1 \leq N$, and a set of possible outputs $S \subset \mathbb{R}^m$:

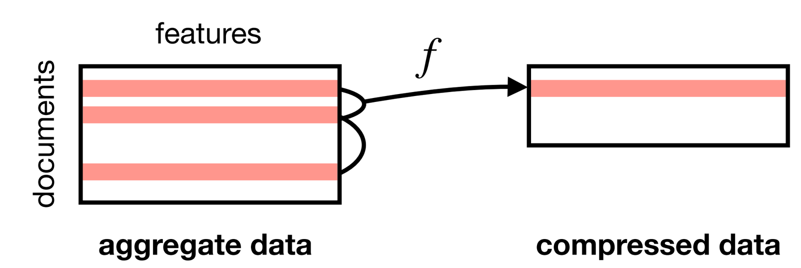
$$\Pr\left[R(y) \in S\right] \leq e^{\epsilon} \cdot \Pr\left[R(y') \in S\right]$$

Informally: LPLP only guarantees documents are hard to distinguish from *similar* documents.

New mechanisms for limited-precision local privacy

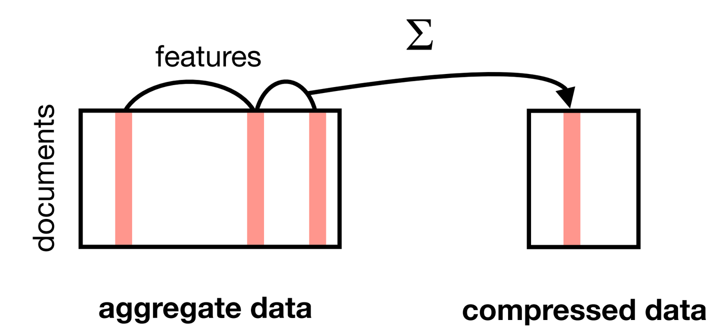
Compression: First compress data and then add random noise, retaining large-scale correlations.

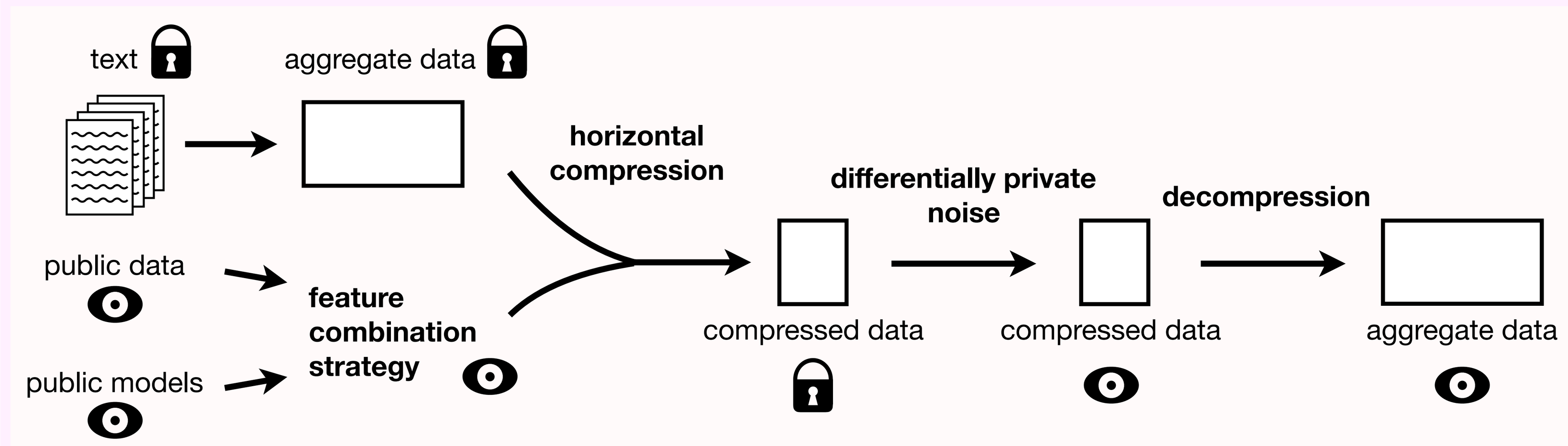
Vertical compression: combines documents



Horizontal compression: combines features

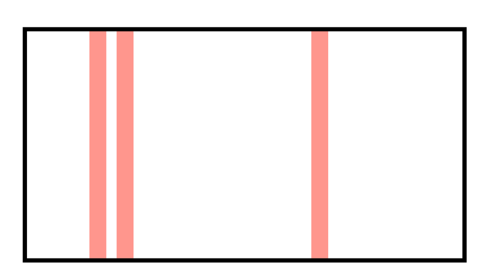
OUR FOCUS



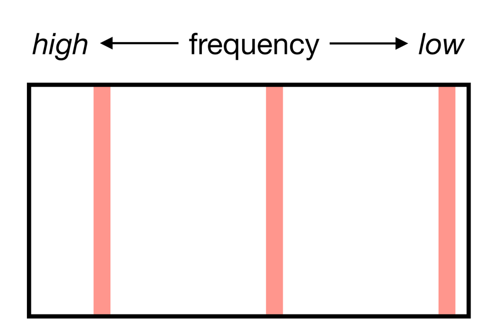


How should features be combined?

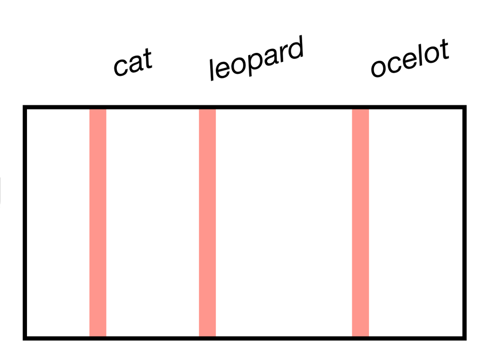
1. Random



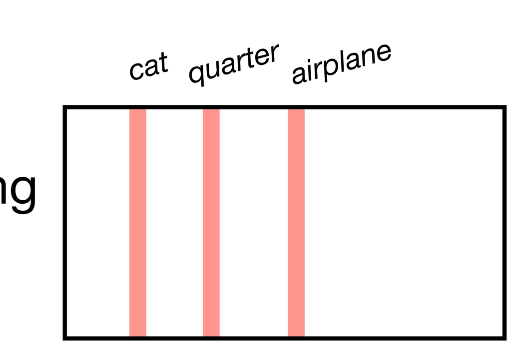
2. Frequency



3. Embedding cluster



4. Embedding disperse



Experiments

Goal: Evaluate whether data with horizontal compression and private noise can produce useful semantic models

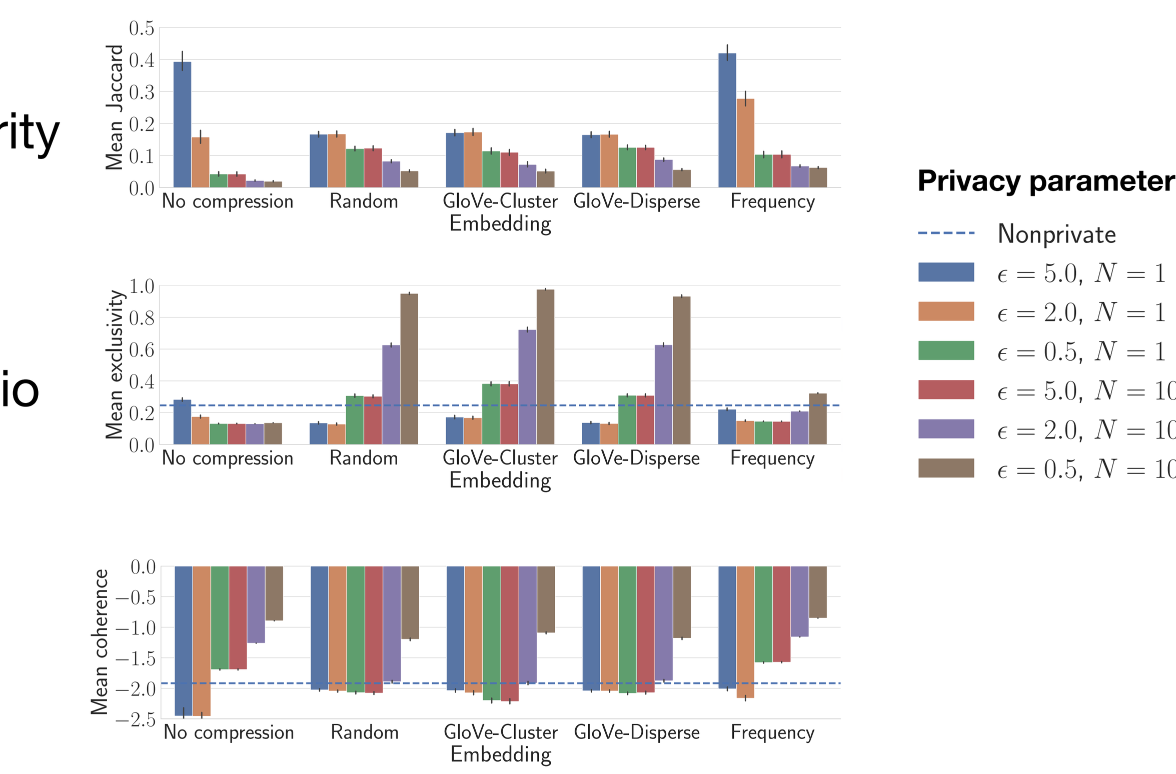
Dataset:
- 9,528 consumer complaints about financial products and services across 7 categories released by the U.S. Consumer Finance Protection Bureau
- 100-500 words in each document

1. LDA: Similarity between private and non-private topics

Jaccard similarity

Exclusivity ratio

Per-word coherence



FREQUENCY COMPRESSION WORKS WELL

2. LSA: Predict category of private and non-private documents

LSA + random forest classification

Takeaways

- High-dimensional bags-of-words are a challenge for local privacy.
- Compression helps with stronger privacy guarantees within LPLP.
- Promising feature combination approaches:
  - Distributing high-frequency features
  - Random feature combination

https://priml-workshop.github.io/priml2019/papers/PriML2019_paper_29.pdf