Semantic Consistent Topic Discovery with Differential Privacy

Yexuan Shi[1], Zhiyang Su[2], Di Jiang[3], Zimu Zhou[4], Wenbin Zhang[1], Yongxin Tong[1]

[1] Beihang University, Beijing, China

[2] Hong Kong University of Science and Technology, Hong Kong SAR, China

[3] WeBank, Shenzhen, China

[4] ETH Zurich, Zurich, Switzerland

Outline

- Background
- Solution
- Experiments
- Conclusion

Background: Topic Modeling

- Understanding the topics of documents is important for many tasks:
  - Tag recommendation
  - Text categorization
  - Opinion mining
  - Statistical language modeling
  - …

Steyvers, M., Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*.

Background: Topic Modeling

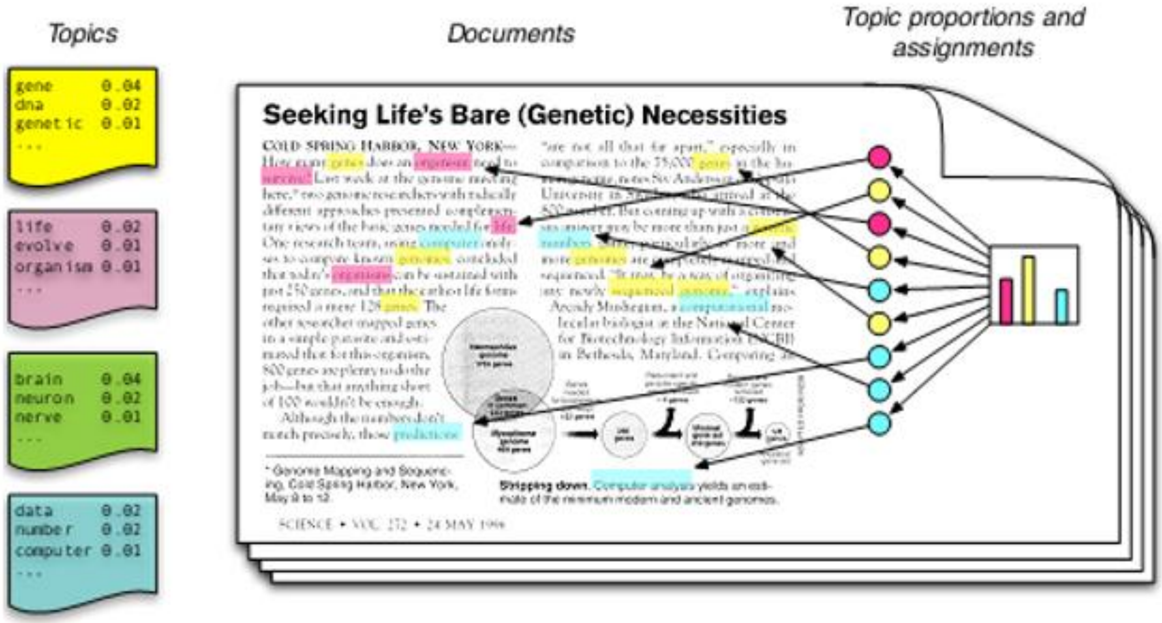
- Understanding the topics of documents is important for many tasks:
  - Tag recommendation
  - Text categorization
  - Opinion mining
  - Statistical language modeling
  - …

- Topic modeling is a powerful technique for the above applications.

Steyvers, M., Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*.

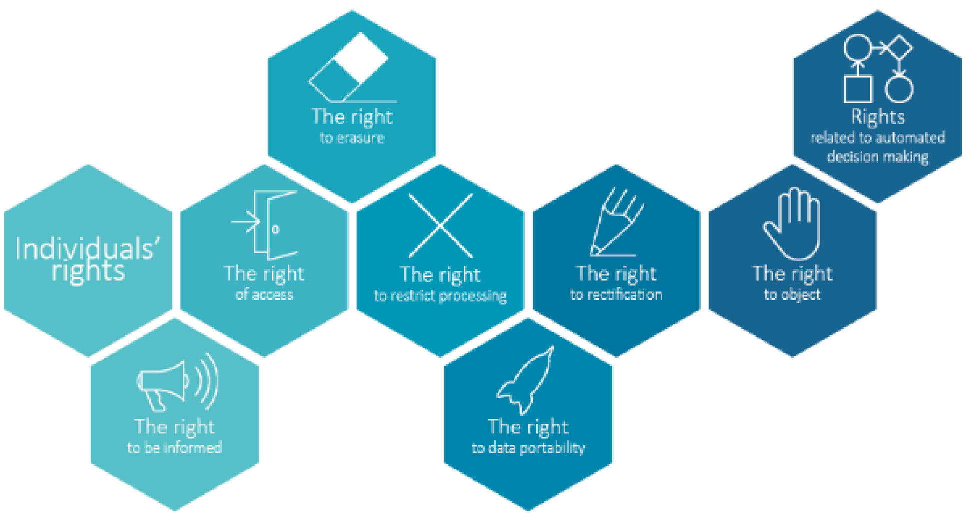
Background: Topic Modeling

- Topic modeling: model the generation of documents with latent topic structure
  - A topic ~ a distribution over words
  - A document ~ a mixture of topics
  - A word ~ a sample drawn from one topic



Background: Privacy Issue

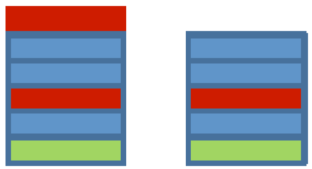
- The General Data Protection Regulation (GDPR), enacted by the European Union, aims primarily to give control to residents over their private data.
- Typical topic modeling does not consider the privacy issue, which may violate the GDPR.



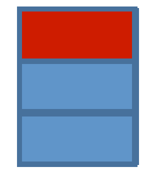
Background: Differential Privacy

For every pair of inputs that differ in one row

For every output



D_1 D_2



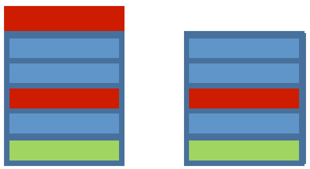
O

- Basic idea: adversary should not be able to distinguish between any D_1 and D_2 on any O

Dwork, C. (2006). Differential privacy. *International Colloquium on Automata Languages and Programming*.

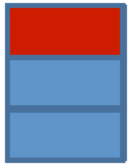
Background: Differential Privacy

For every pair of inputs that differ in one row



D_1 D_2

For every output



O

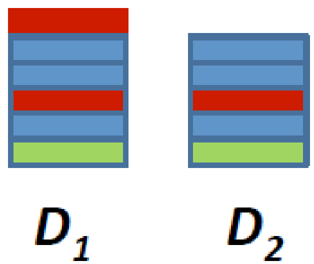
- Basic idea: adversary should not be able to distinguish between any D_1 and D_2 on any O
- Definition: mechanism M is (ϵ, δ)-differential private if

$$\Pr[M(D_1) = O] \leq e^{\epsilon} \Pr[M(D_2) = O] + \delta$$

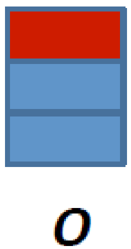
Dwork, C. (2006). Differential privacy. *International Colloquium on Automata Languages and Programming*.

Background: Differential Privacy

For every pair of inputs that differ in one row



For every output



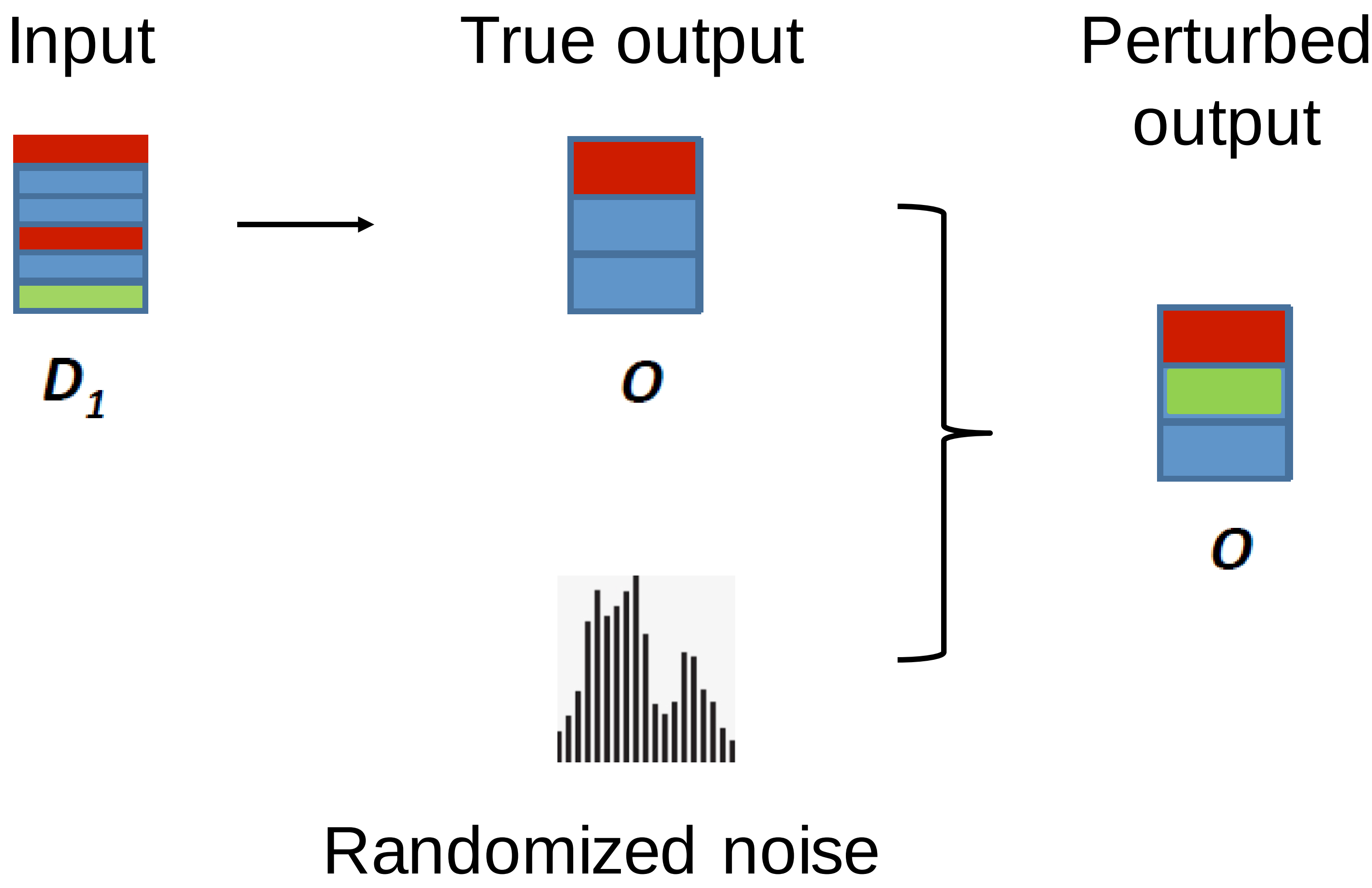
- Basic idea: adversary should not be able to distinguish between any D_1 and D_2 on any O
- Definition: mechanism M is (ϵ, δ)-differential private if

Smaller the ϵ, more the privacy

$$\Pr[M(D_1) = O] \leq e^{\epsilon} \Pr[M(D_2) = O] + \delta$$

Dwork, C. (2006). Differential privacy. *International Colloquium on Automata Languages and Programming*.

Background: Differential Privacy



Dwork, C. (2006). Differential privacy. *International Colloquium on Automata Languages and Programming*.

Background: Challenges

- Differential privacy (DP) introduces noise into topic models. It may degrade the accuracy of topic models.

Background: Challenges

- Differential privacy (DP) introduces noise into topic models. It may degrade the accuracy of topic models.

- The challenge is ***how to provide a reasonable degree of privacy while retaining the accuracy of topic modeling***.

Background: Challenges

- Differential privacy (DP) introduces noise into topic models. It may degrade the accuracy of topic models.

- The challenge is ***how to provide a reasonable degree of privacy while retaining the accuracy of topic modeling***.

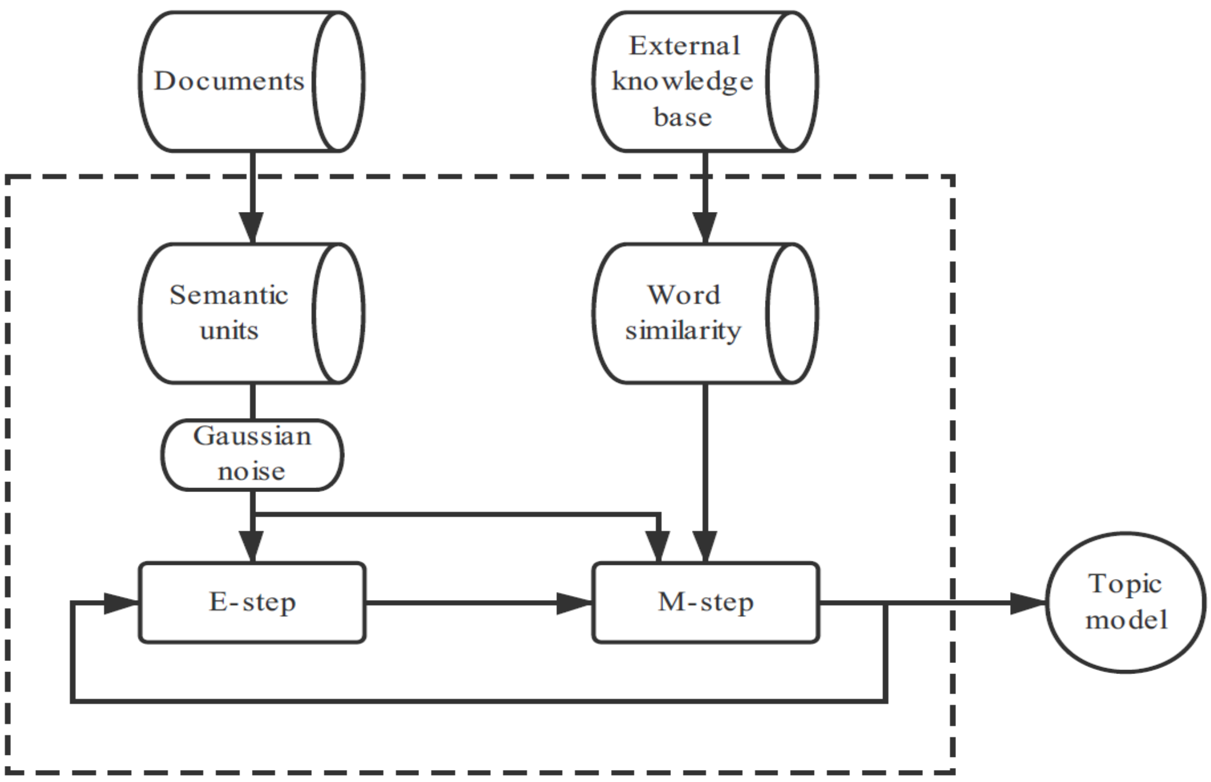
- Solution: Private and Consistent Topic Discovery (PC-TD)

Outline

- Background
- Solution
- Experiments
- Conclusion

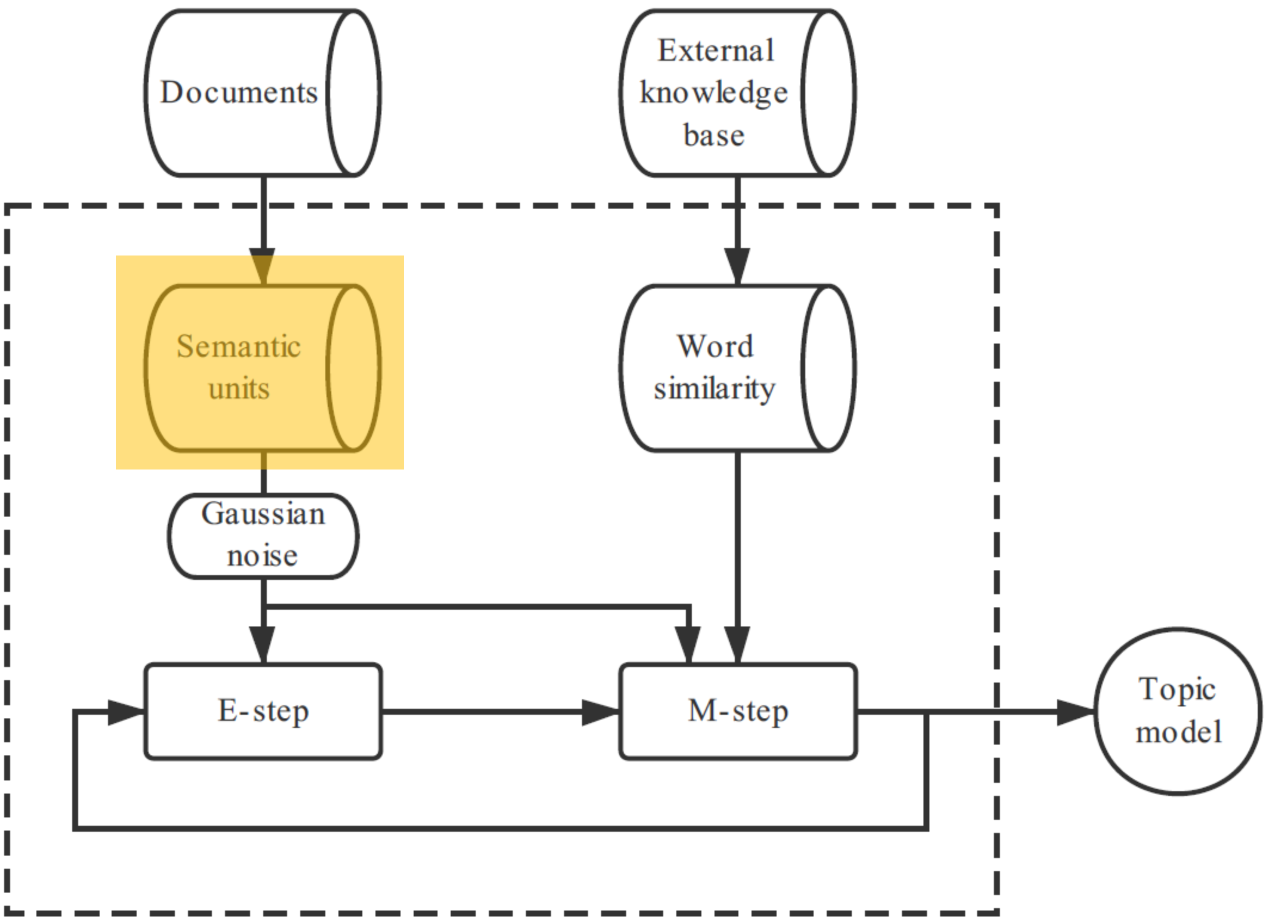
Overview

- A generative model to discover the topics of documents.
  - Local semantic consistency: semantic units
  - Global semantic consistency: word similarity



Local Semantic Consistency

- Basic idea: Several consecutive words should belong to the same topic



Local Semantic Consistency

- Semantic unit:
  - A semantic unit is a series of words generated by a single topic
  - It can be flexibly interpreted as:
    - N-gram
    - Sentence
    - Paragraph
    - …
  - A bi-gram example: (each color represent a topic)

Model Assumption

- Generation process:
  - 1. Select a document d_i with probability $p(d_i)$;
  - 2. For each semantic unit s_{ij} in d_i, pick a latent topic z_k with probability $p(z_k|d_i)$;
  - 3. For each position in s_{ij}, generate a word ω with probability $p(\omega|z_k)$.

Model Assumption

- Generation process:
  - 1. Select a document d_i with probability $p(d_i)$;
  - 2. For each semantic unit s_{ij} in d_i, pick a latent topic z_k with probability $p(z_k|d_i)$;
  - 3. For each position in s_{ij}, generate a word ω with probability $p(\omega|z_k)$.
- Complete data logarithm likelihood:

$$L(\mathrm{d},\mathrm{s},\mathrm{z}) = \sum_{i=1}^{D}\sum_{j=1}^{S_i}\sum_{k=1}^{Z} \log p(d_i, s_{ij}, z_k)$$

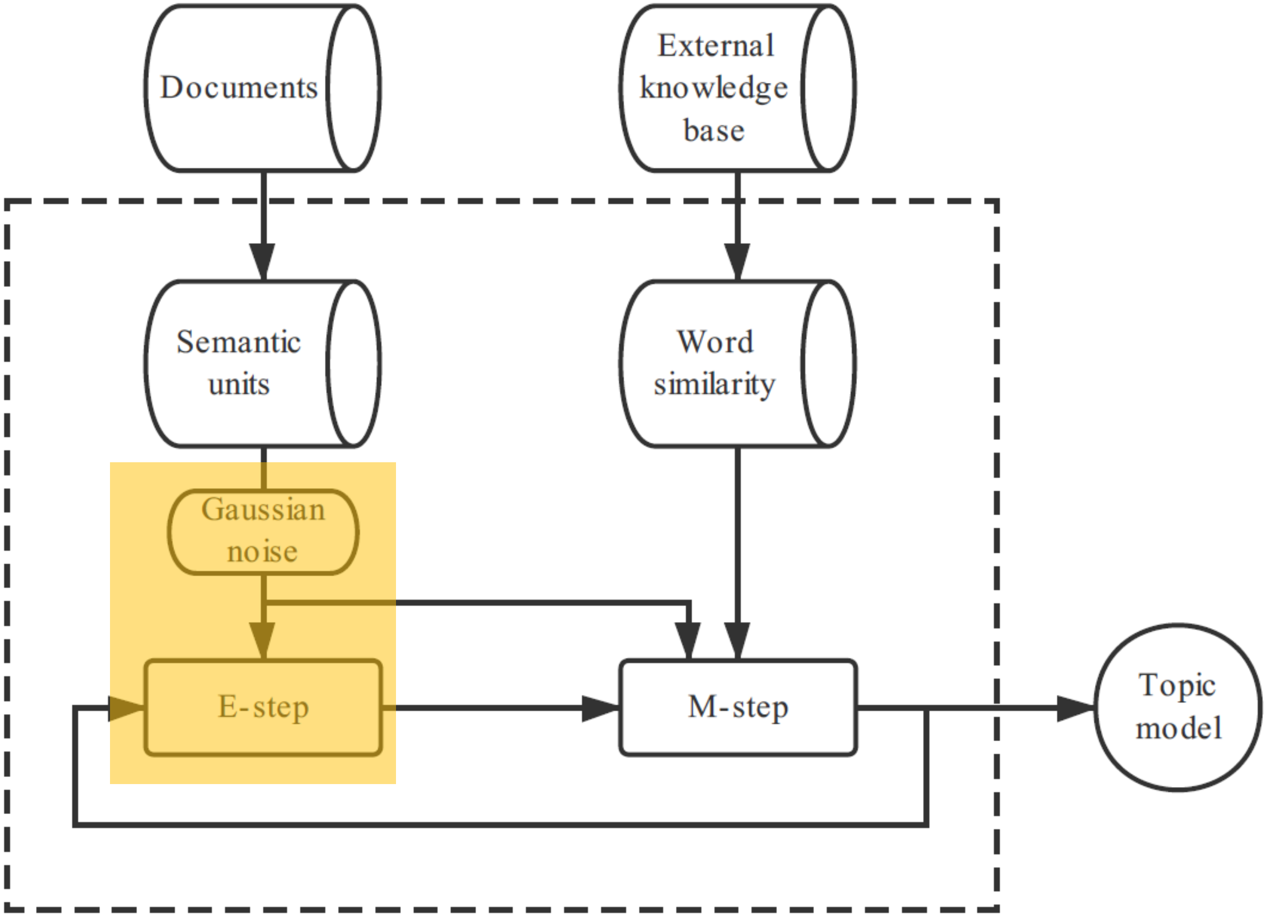
$$= \sum_{i=1}^{D}\sum_{j=1}^{S_i}\sum_{k=1}^{Z} \log \Big(p(d_i)p(z_k|d_i)p(s_{ij}|z_k) \Big)$$

D: number of documents

S_i: number of semantic units in the i-th document

Z: number of topics

Private Inference: E-step



Private Inference: E-step

- The posterior estimation of the latent topic z_k of semantic unit s_{ij} in document d_i:
  - $$p(z_k|d_i, s_{ij}) = \frac{p(z_k|d_i)p(s_{ij}|z_k)}{\sum_{k'=1}^{Z} p(z_{k'}|d_i)p(s_{ij}|z_{k'})}$$
    - $p(s_{ij}|z_k) = \prod_{\omega=1}^{W} p(\omega|z_k)^{N_{ij\omega}}$
    - $N_{ij\omega}$ is the number of ω in s_{ij}

Private Inference: E-step

- The posterior estimation of the latent topic z_k of semantic unit s_{ij} in document d_i:
  - $$p(z_k|d_i, s_{ij}) = \frac{p(z_k|d_i)p(s_{ij}|z_k)}{\sum_{k'=1}^{Z} p(z_{k'}|d_i)p(s_{ij}|z_{k'})}$$
    - $p(s_{ij}|z_k) = \prod_{\omega=1}^{W} p(\omega|z_k)^{N_{ij\omega}}$
    - $N_{ij\omega}$ is the number of ω in s_{ij}
- Which term will access the training data?

Private Inference: E-step

- The posterior estimation of the latent topic z_k of semantic unit s_{ij} in document d_i:
  - $$p(z_k|d_i, s_{ij}) = \frac{p(z_k|d_i)p(s_{ij}|z_k)}{\sum_{k'=1}^{Z} p(z_{k'}|d_i)p(s_{ij}|z_{k'})}$$
    - $p(s_{ij}|z_k) = \prod_{\omega=1}^{W} p(\omega|z_k)^{N_{ij\omega}}$
    - $N_{ij\omega}$ is the number of ω in s_{ij}
- Which term will access the training data?
  - $N_{ij\omega}$

Private Inference: E-step

- Apply DP mechanism to counting $N_{ij\omega}$
  - Add Gaussian noise to $N_{ij\omega}$: ($\Delta N = 1$)

$$\hat{N}_{ijw} = N_{ijw} + \Omega \qquad \Omega \sim \mathcal{N}(0, (\Delta N)^2 \sigma^2)$$

Private Inference: E-step

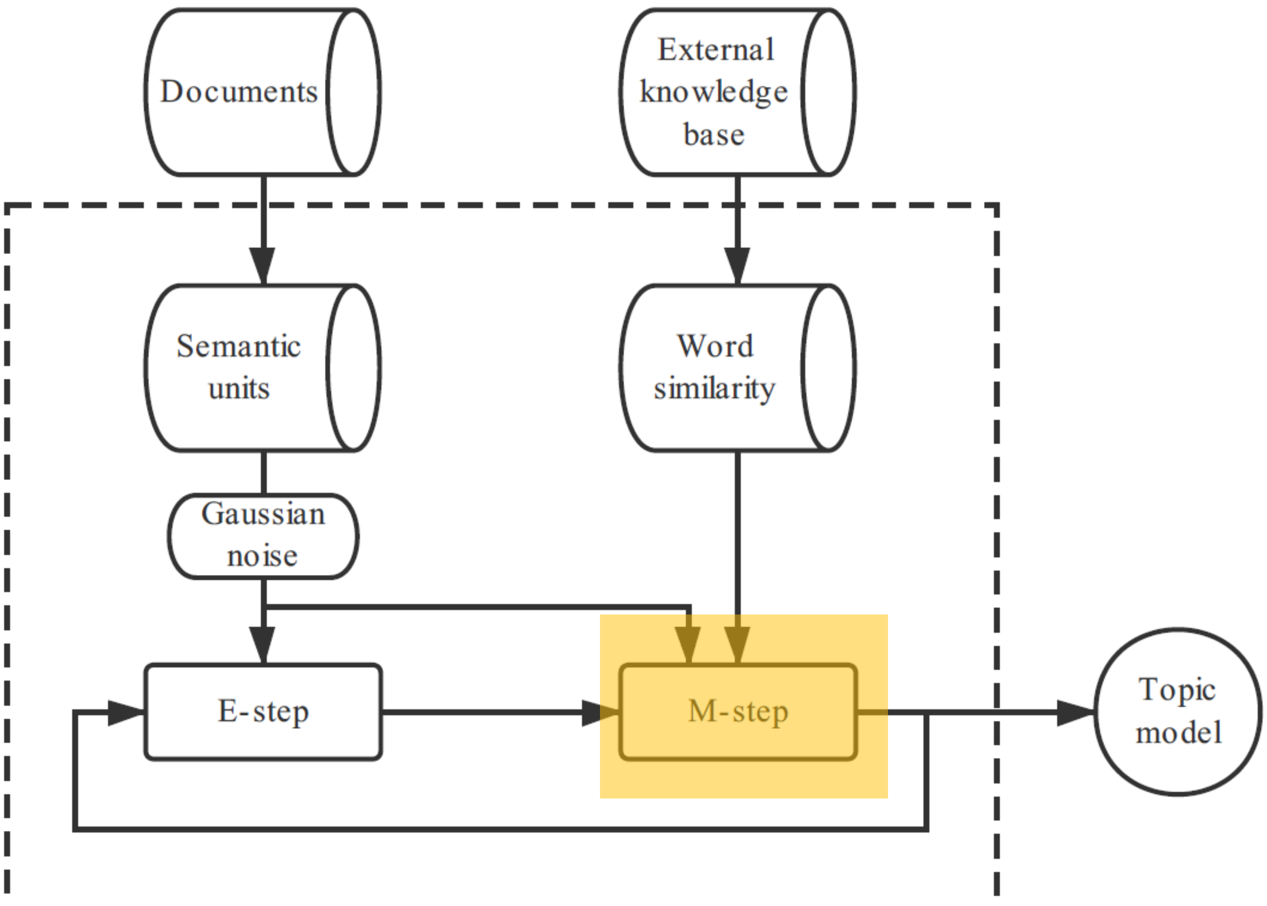
- Apply DP mechanism to counting $N_{ij\omega}$
  - Add Gaussian noise to $N_{ij\omega}$: ($\Delta N = 1$)

$$\hat{N}_{ijw} = N_{ijw} + \Omega \qquad \Omega \sim \mathcal{N}(0, (\Delta N)^2 \sigma^2)$$

- The perturbed posterior estimation:

$$\begin{aligned}\hat{r}_{ijk} &= \hat{p}(z_k | d_i, s_{ij}) \\ &= \frac{p(z_k|d_i) \prod_{w=1}^{W} p(w|z_k)^{\hat{N}_{ijw}}}{\sum_{k'=1}^{Z} p(z_{k'}|d_i) \prod_{w'=1}^{W} p(w'|z'_k)^{\hat{N}_{ijw}}}\end{aligned}$$

Private Inference: M-step



Private Inference: M-step

- The maximization of expected logarithm likelihood leads to:

$$p(z_k|d_i) = \frac{\sum_{j=1}^{S_i} \hat{r}_{ijk}}{\sum_{j=1}^{S_i} \sum_{k'=1}^{Z} \hat{r}_{ijk'}}$$

$$p(w|z_k) = \frac{\sum_{i=1}^{D} \sum_{j=1}^{S_i} N_{ijw} \hat{r}_{ijk}}{\sum_{i=1}^{D} \sum_{j=1}^{S_i} N_{ij} \hat{r}_{ijk}}$$

Private Inference: M-step

- The maximization of expected logarithm likelihood leads to:

$$p(z_k|d_i) = \frac{\sum_{j=1}^{S_i} \hat{r}_{ijk}}{\sum_{j=1}^{S_i} \sum_{k'=1}^{Z} \hat{r}_{ijk'}}$$

$$p(w|z_k) = \frac{\sum_{i=1}^{D} \sum_{j=1}^{S_i} N_{ijw} \hat{r}_{ijk}}{\sum_{i=1}^{D} \sum_{j=1}^{S_i} N_{ij} \hat{r}_{ijk}}$$

- The same as E-step, we add noise to $N_{ij\omega}$ and get the perturbed $\hat{P}(\omega|z_k)$:

$$\hat{p}(w|z_k) = \frac{\sum_{i=1}^{D} \sum_{j=1}^{S_i} \hat{N}_{ijw} \hat{r}_{ijk}}{\sum_{i=1}^{D} \sum_{j=1}^{S_i} \hat{r}_{ijk} \sum_{w=1}^{W} \hat{N}_{ijw}}$$

Analysis of Privacy

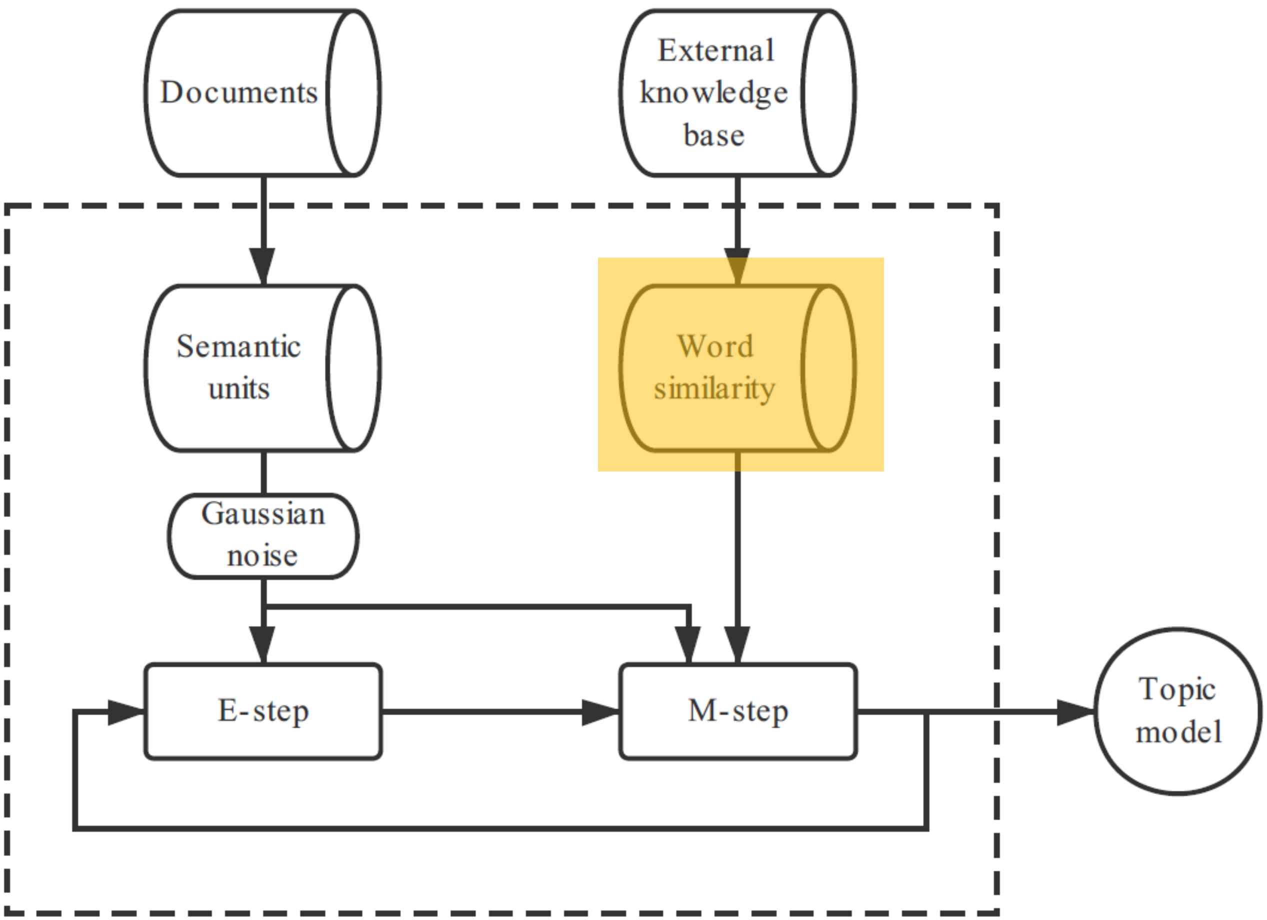
- We use the ***Moments Accountant*** (MA) composition method to account the privacy loss incurred by iterations of our EM algorithm.

Theorem 1. *For any $\epsilon < \Theta(T)$, PC-TD is (ϵ, δ)-differentially private for any $\delta > 0$ if we choose*

$$\sigma \geq \Theta\Big(\frac{\sqrt{T \log(1/\delta)}}{\epsilon}\Big)$$

Global Semantic Consistency

- Basic idea: The words with similar meanings should belong to the same topic.



Global Semantic Consistency

- Basic idea: The words with similar meanings should belong to the same topic.
  - Use word embedding to represent the words to vectors.
  - The similarity of two words a and b can be calculated by cosine similarity.

$$R_{ab} = \frac{v_a \cdot v_b}{||v_a||_2 ||v_b||_2}$$

Global Semantic Consistency

- For a given similarity matrix R, we adjust the topic-word distribution $P(\omega|z_k)$ as:

$$p'(w|z_k) \leftarrow p(w|z_k) + \tau \frac{p(w|z_k) \sum_{i=1}^{W} R_{iw} p(i|z_k)}{P(\cdot|z_k)^T R P(\cdot|z_k)}$$

- τ is a hyperparameter.

Outline

- Background
- Solution
- Experiments
- Conclusion

Experiments: Setup

- Dataset:
  - New York Times, June 26th-30th, 2016
  - 500 documents
  - 18,286 unique words
- Compared Algorithms:
  - PC-TD ($\tau = 0.3$)
  - LDA($\alpha = Z/50, \beta = 0.01$)

Experiments: Metric

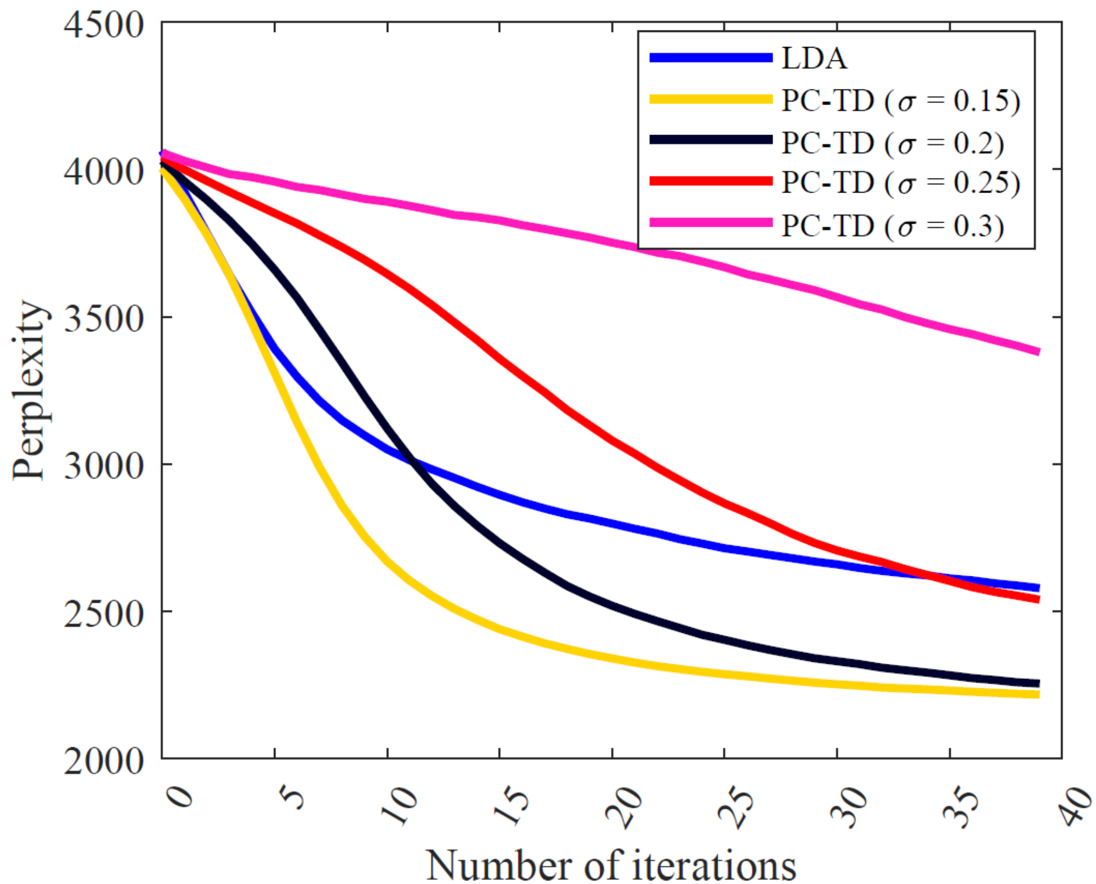
- We use the perplexity of documents to evaluate the performance of topic models.

- Perplexity =

$$\exp\left(-\frac{1}{\sum_{i=1}^{D}|d_i|}\sum_{i=1}^{D}\sum_{w\in d_i}\ln\left(\sum_{k=1}^{Z}p(w|z_k)p(z_k|d_i)\right)\right)$$

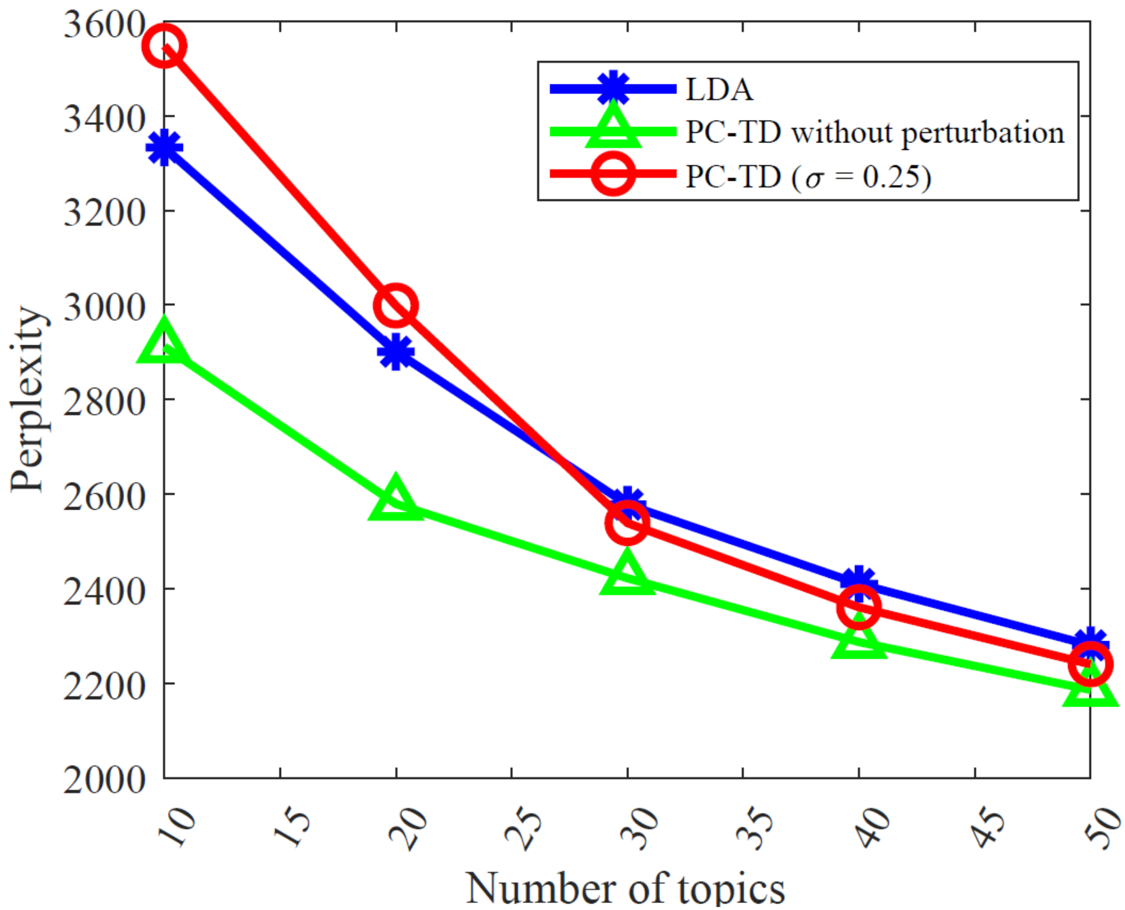
Experiments: Result

- Convergence curves of varying σ
  - the smaller the privacy budget is, the more slowly the algorithm converges.



Experiments: Result

- Perplexity of varying number of topics
  - As the amount of topics increases, the perturbed PC-TD performs gradually better than LDA.



Outline

- Background
- Solution
- Experiments
- Conclusion

Conclusion

- We propose PC-TD to discover latent topics with semantic consistency and privacy guarantee.
- We propose a differential private parameter inference algorithm for PC-TD to ensure the privacy of sensitive documents.
- Experiments on a corpus collected from New York Times show that the proposed method outperforms the conventional LDA.

Q & A

Thank You