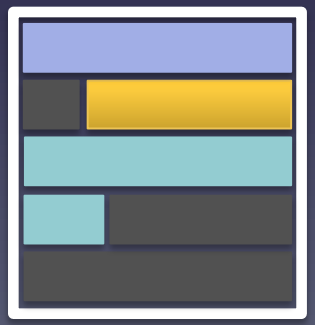
Thesis Contributions

Inferring hidden structures in government text → new insights into law and policy

Supreme Court authorship attribution (supervised learning)



Probabilistic Text Reuse for understanding large document collections (unsupervised learning)

Probabilistic Text Reuse (PTR)

Outline

- Motivation: Government text datasets
- Model
- Inference
- Results: 800,000 FCC net neutrality comments

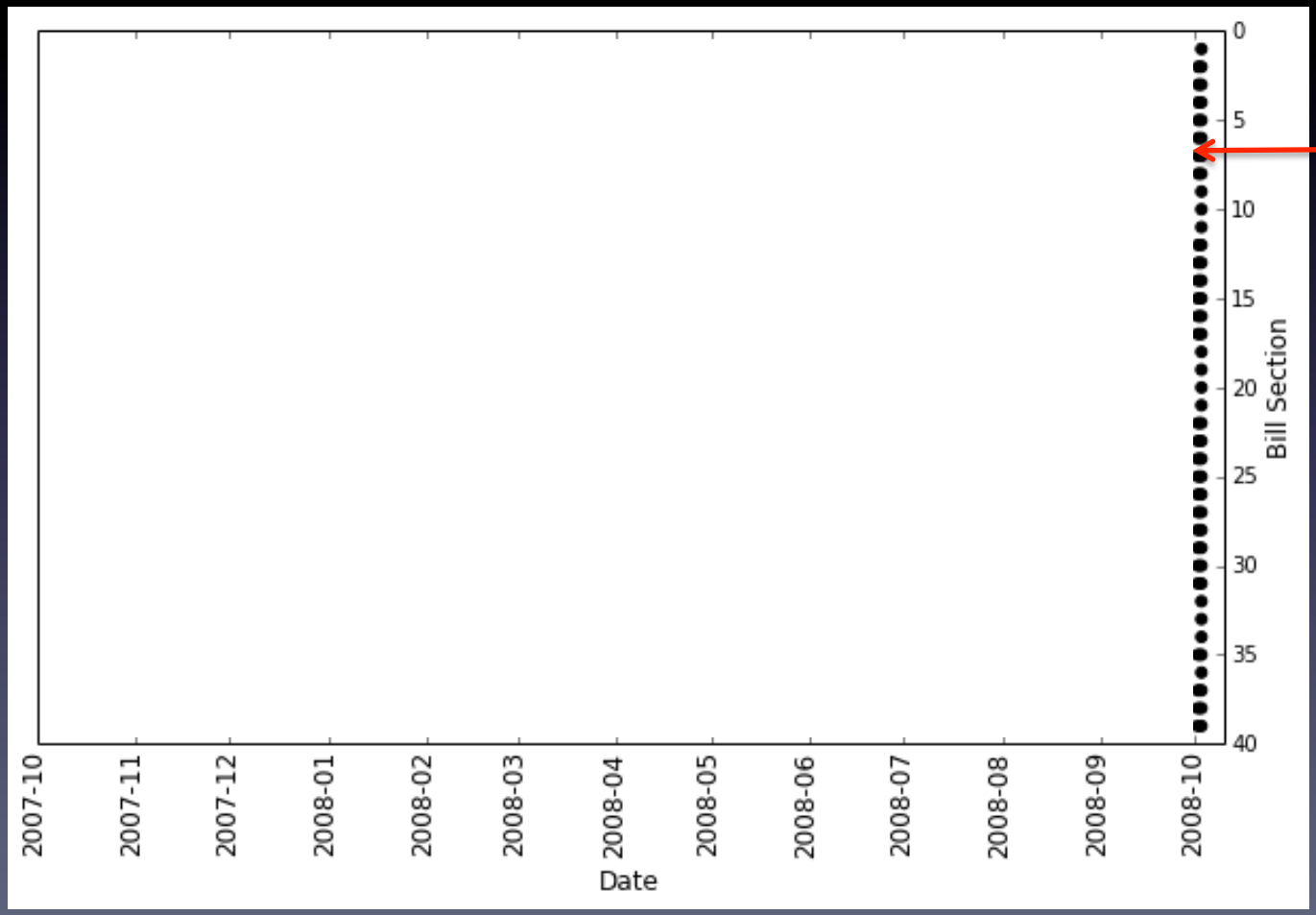
PTR: Motivation #1

Repeated text occurs widely in government documents

- Legislative bills (Wilkerson, Stramp, Smith 2015)
- 19th century American newspapers (Smith, Cordell, Dillon 2013)
- Public statements from Members of Congress (Lin, Margolin, Lazer 2015)

Text Reuse in Congress

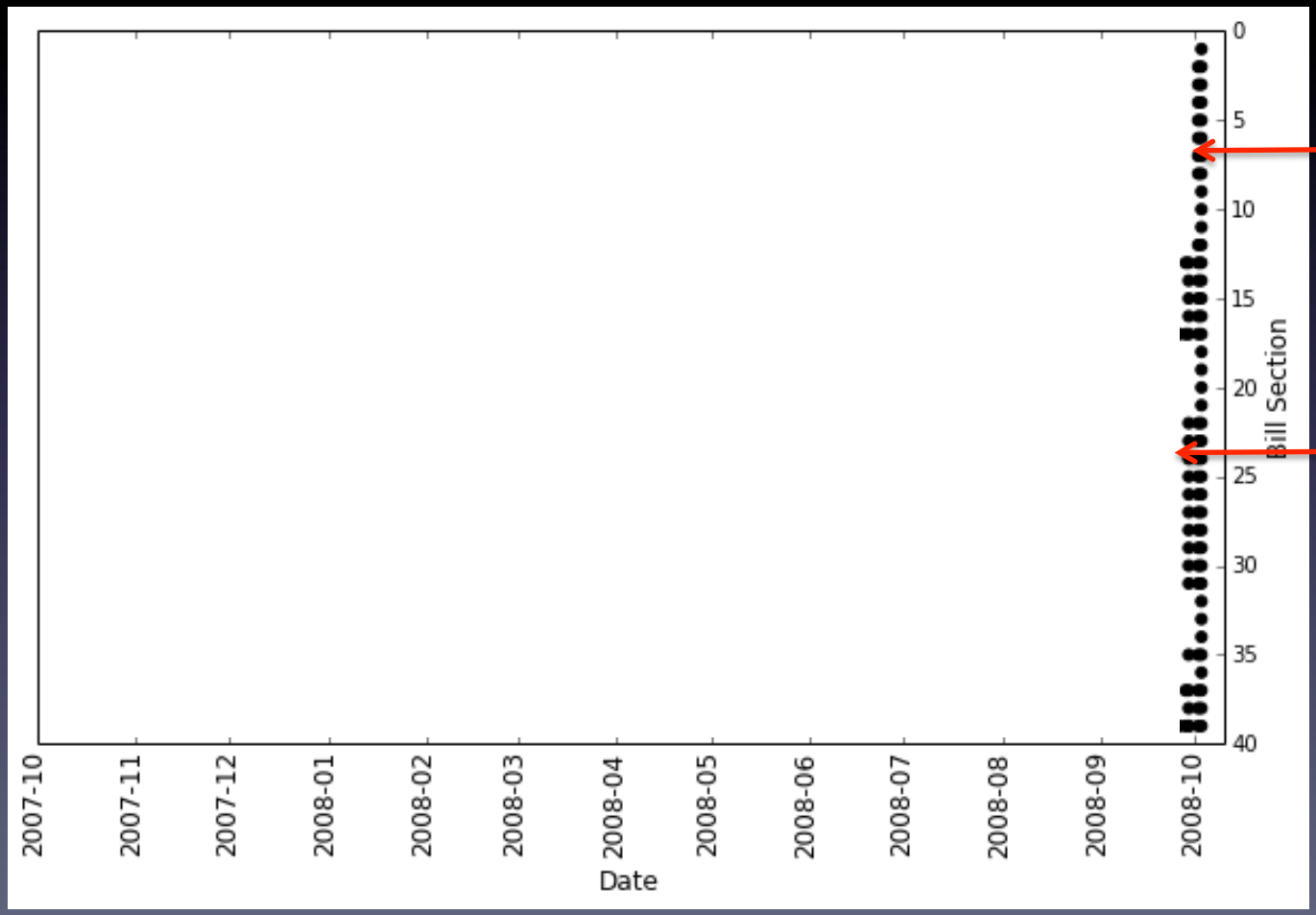
Troubled Asset Relief Program



Enacted
October 3, 2008

Text Reuse in Congress

Troubled Asset Relief Program

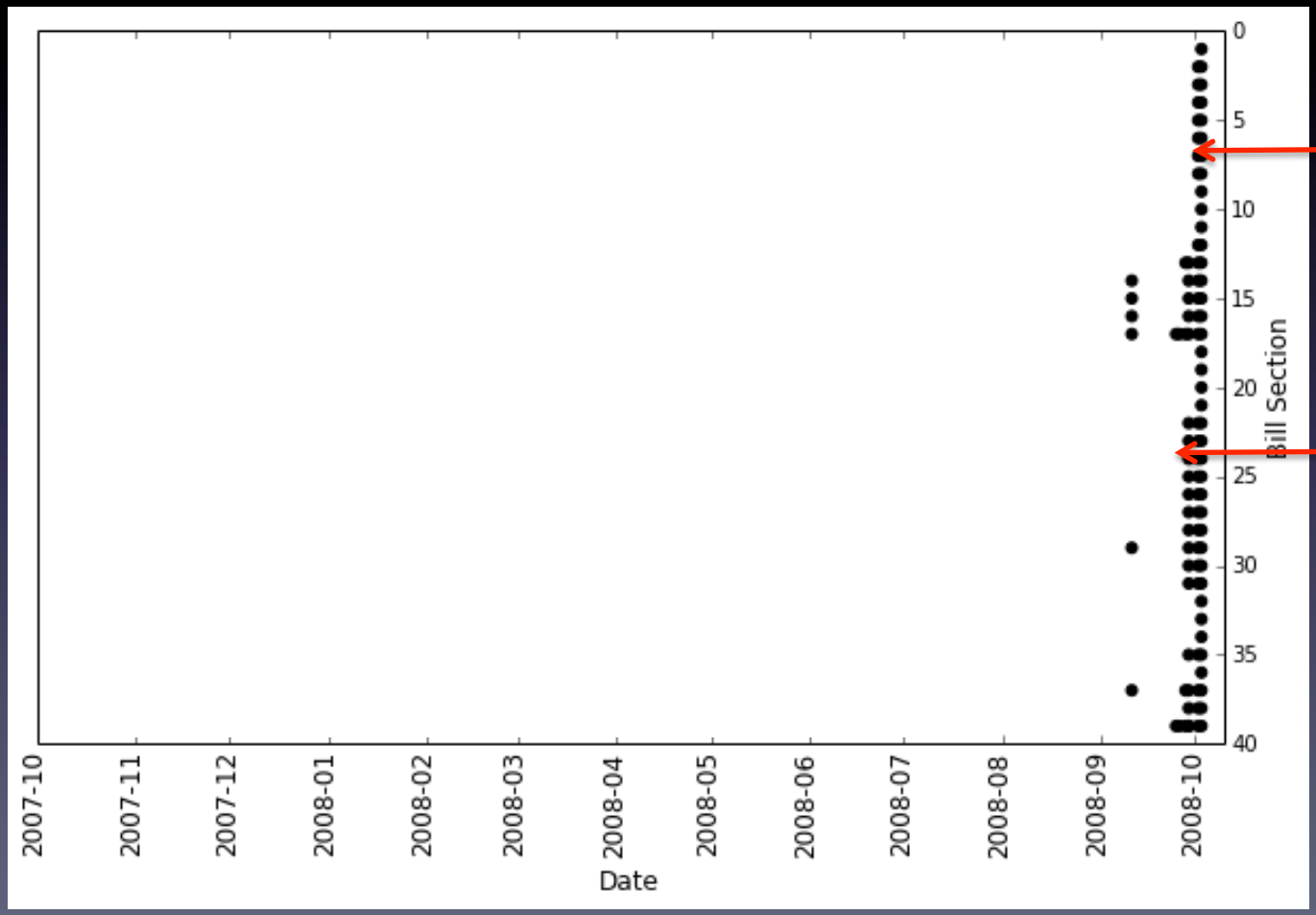


Enacted
October 3, 2008

Failed House Vote
(Dow Jones
drops 778 points)
September 29, 2008

Text Reuse in Congress

Troubled Asset Relief Program

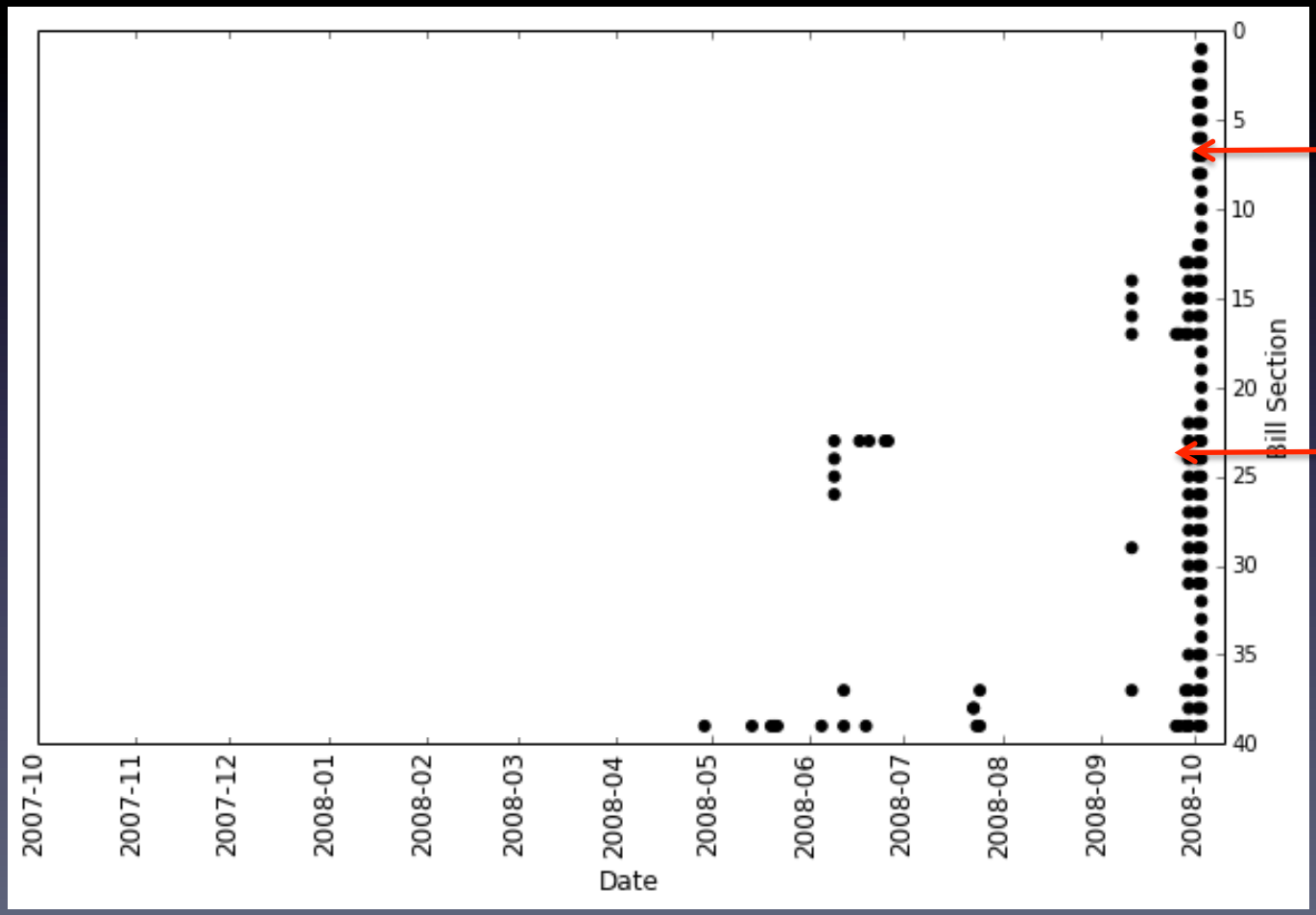


Enacted
October 3, 2008

Failed House Vote
(Dow Jones
drops 778 points)
September 29, 2008

Text Reuse in Congress

Troubled Asset Relief Program

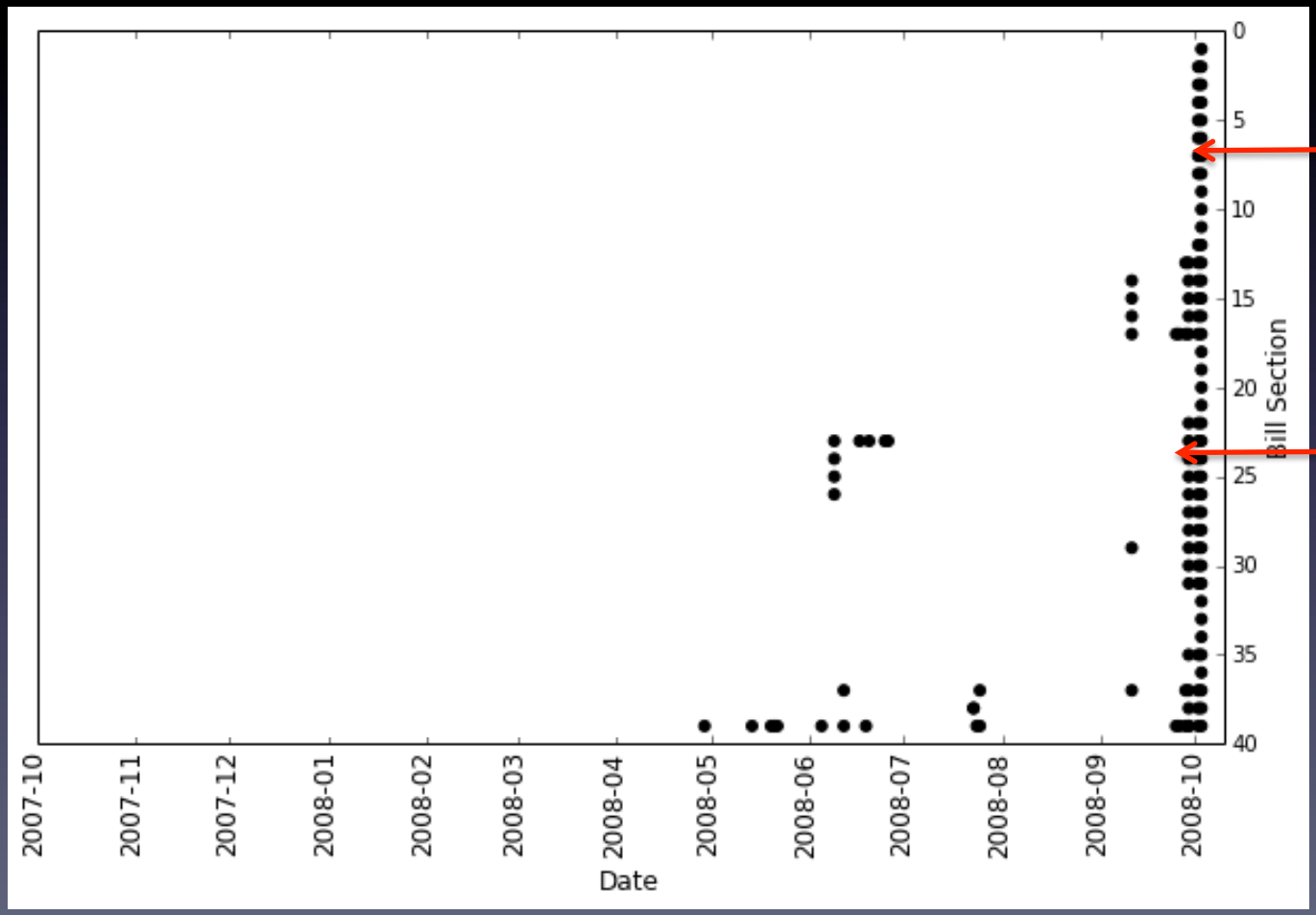


Enacted
October 3, 2008

Failed House Vote
(Dow Jones
drops 778 points)
September 29, 2008

Text Reuse in Congress

Troubled Asset Relief Program

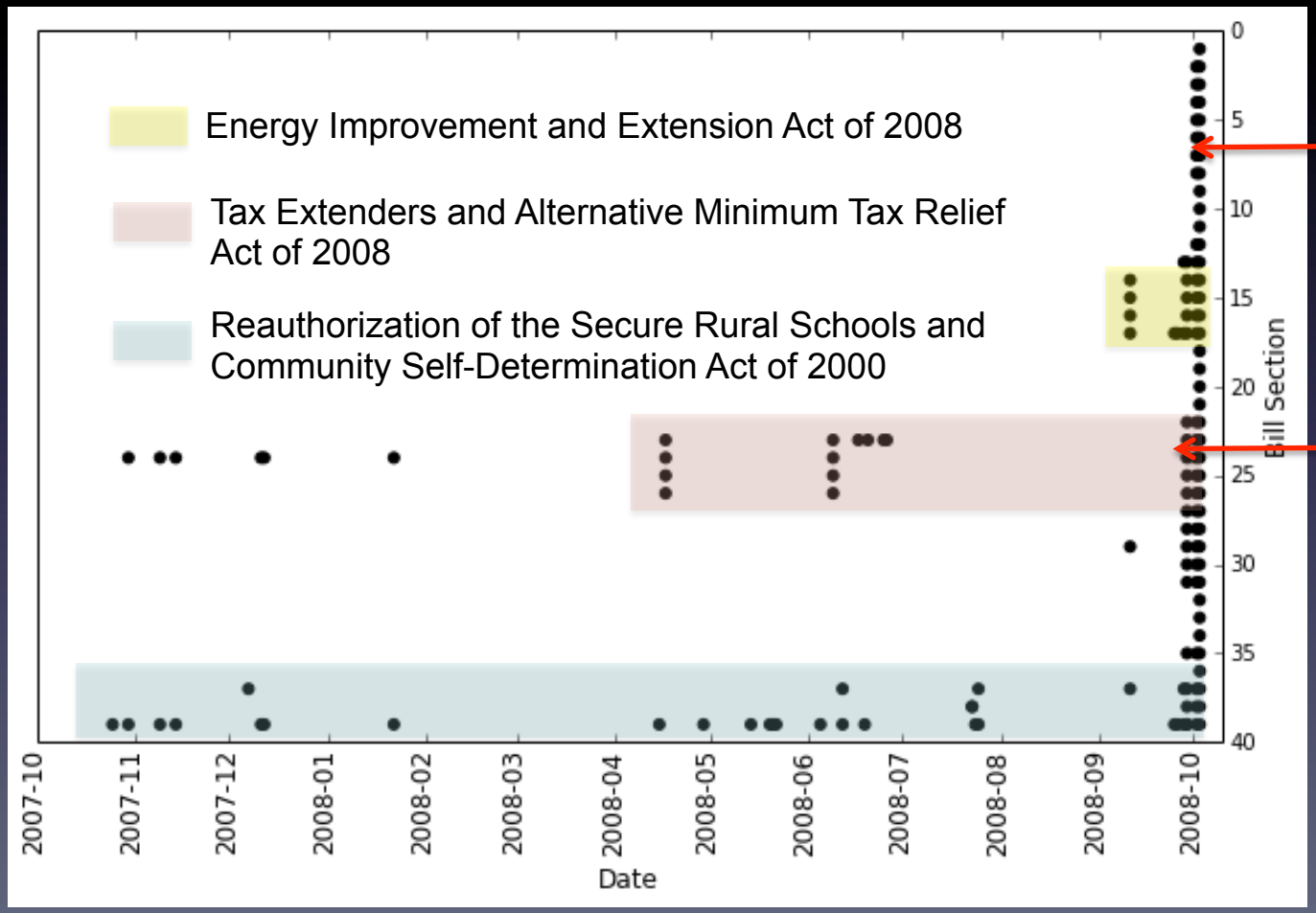


Enacted
October 3, 2008

Failed House Vote
(Dow Jones
drops 778 points)
September 29, 2008

Text Reuse in Congress

Troubled Asset Relief Program



Enacted
October 3, 2008

Failed House Vote
(Dow Jones
drops 778 points)
September 29, 2008

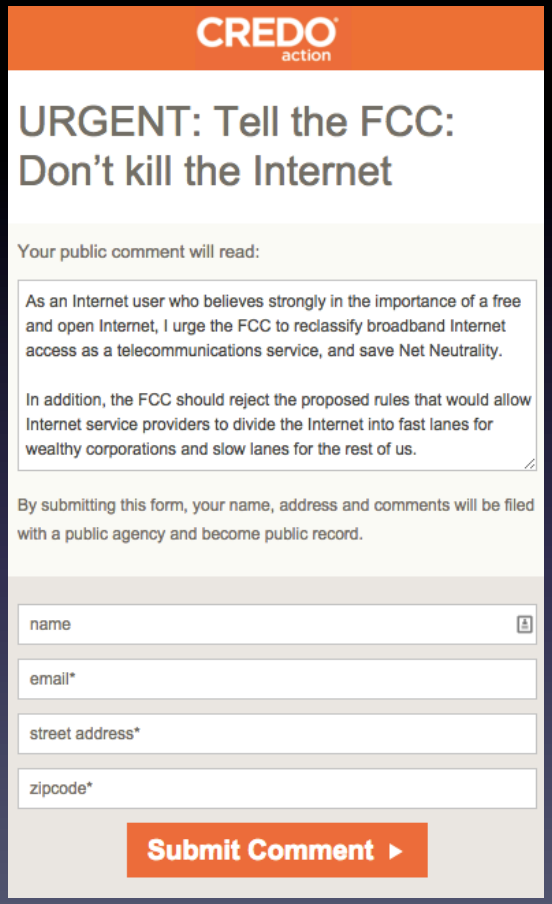
Text Reuse in Public Comments

- Proposed net neutrality regulations
- 800,000 public comments (May to July 2014)
  - 101 million words, 126 words/ comment

Template Ideas

"As an Internet user who believes strongly in the importance of a free and open Internet, I urge the FCC to reclassify broadband Internet access as a telecommunications service, and save Net Neutrality. In addition, the FCC should reject the proposed rules that would allow Internet service providers to divide the Internet into fast lanes for wealthy corporations and slow lanes for the rest of us."

(93,711 comments)



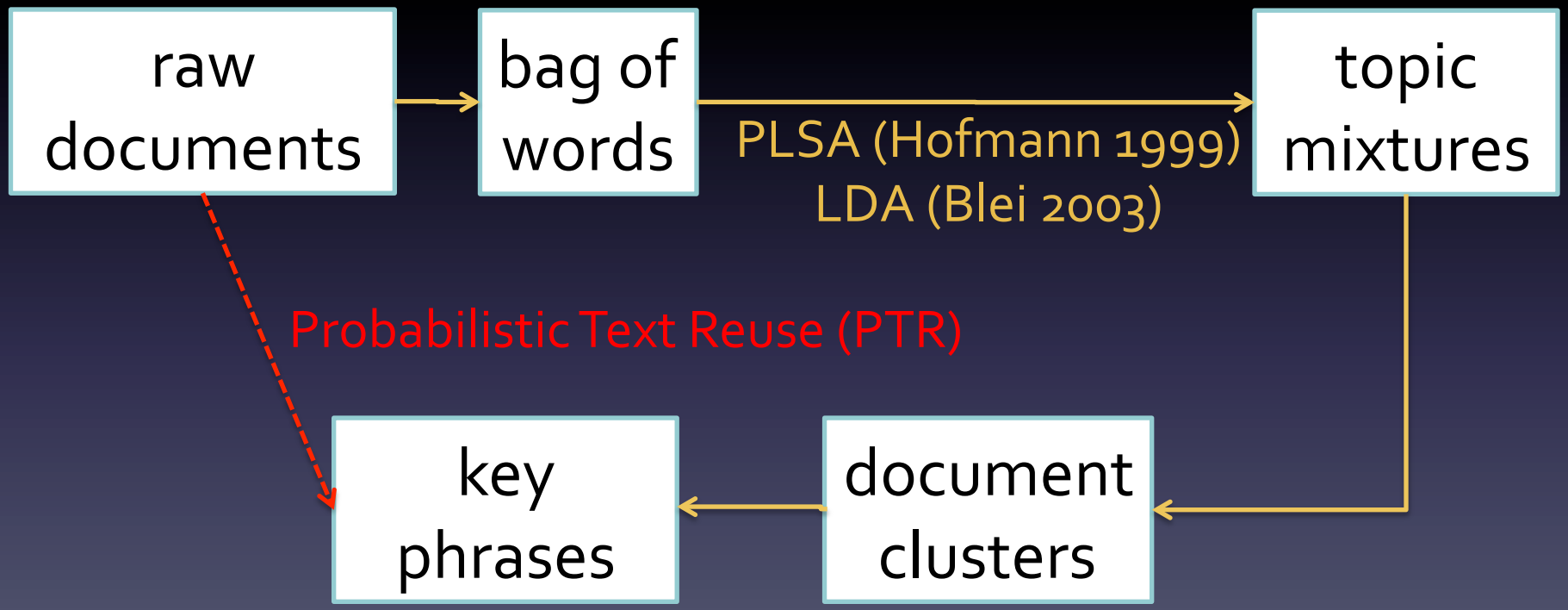
Text Reuse in Public Comments

- What are people saying?
  - Most common ideas and comments (templates)
  - Variations of ideas (different wording)
  - Less-common voices

PTR: Motivation #2

- **Current models of text corpora fail to capture text reuse**
  - Model mismatch
  - Limitations on interpretability and usefulness

Existing Work: Probabilistic Topic Modeling



- Text passages → bags of words → text passages

Probabilistic Text Reuse (PTR)

Represent documents with

reused text sequences

+

Probabilistic modeling

PTR: Intuition

Corpus:

- *D* documents (sequences of words)

Text generators:

- Background text model, π
- *K* ideas (text sequence generators)

Generative process:

1. Choose # of partitions, *Z*

PTR: Intuition

Corpus:

- *D* documents (sequences of words)

Text generators:

- Background text model, π
- *K* ideas (text sequence generators)

Generative process:

1. Choose # of partitions, *Z*
2. Assign each partition to one of K+1 text generators

PTR: Intuition

Corpus:

- *D* documents (sequences of words)

Text generators:

- Background text model, π
- *K* ideas (text sequence generators)

iii jjj kkk lll

π

fff zzz hhh

aaa ccc ddd eee yyy

π π π π π

π

aaa bbb ccc ddd eee

fff ggg hhh

iii jjj kkk lll

Generative process:

1. Choose # of partitions, *Z*
2. Assign each partition to one of K+1 text generators
3. Generate word sequence from text generator

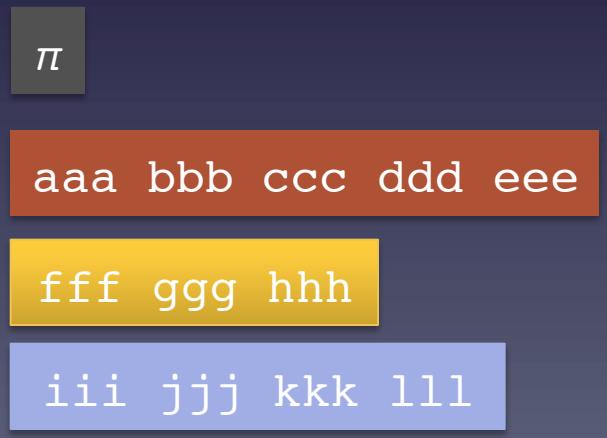
PTR: Intuition

Corpus:

- *D* documents (sequences of words)

Text generators:

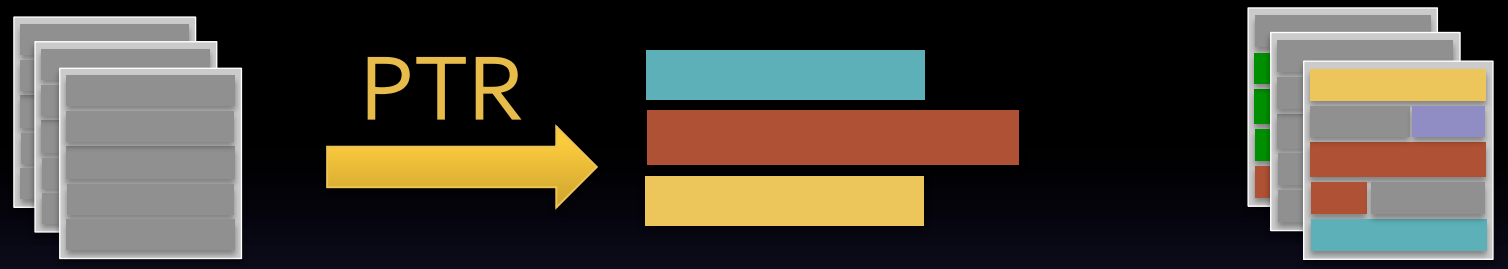
- Background text model, π
- *K* ideas (text sequence generators)



Generative process:

1. Choose # of partitions, *Z*
2. Assign each partition to one of K+1 text generators
3. Generate word sequence from text generator

PTR: Model



Documents, *D*

Ideas, *K*

Partitions, *Z*
Assignments, *A*

Objective Function:

$$P(K, Z, A \mid D) \propto P(K, Z, A) \cdot P(D \mid K, Z, A)$$

PTR: Model

Objective Function (document *n*, partition *m*)**:**

$$P(K, Z, A \mid D) \propto P(K, Z, A) \cdot P(D \mid K, Z, A)$$

$$= P(K) \cdot P(Z) \cdot P(A) \prod_{\text{partitions, } z_{nm}} P(d_{z_{nm}} \mid k_{a_{nm}})$$

idea model

partition model

assignment model

text passage model

Idea Model: P(K)

For each idea, *k*:

1. Choose length, *L*, from uniform distribution
2. Choose words, *w*, from background text model, *π*

$$
\begin{aligned}
L_i &\sim \mathrm{Unif}(N_{min}, N_{max}) \\
k_i(l) &\sim \pi(w) \\
P(K) &= \prod_{i=0}^{I} \frac{1}{N_{max} - N_{min}} \cdot \prod_{\text{words}, w_l \in k_i} \pi(w_l)
\end{aligned}
$$

Partition Model: P(Z)

For each document, *n*:

1. Choose number of partitions, *m*, from uniform distribution

$$\begin{aligned} \text{partitions per document} &\sim \text{Unif}(1, N_z) \\ P(Z) &= \prod_{\text{documents},n}^{N} \left(\frac{1}{N_z - 1}\right) \\ &= \left(\frac{1}{N_z - 1}\right)^N \end{aligned}$$

Assignment Model: P(A)

For each partition, *z*:

1. Choose an assignment, *a*, from multinomial distribution over ideas
   - Learn θ during inference

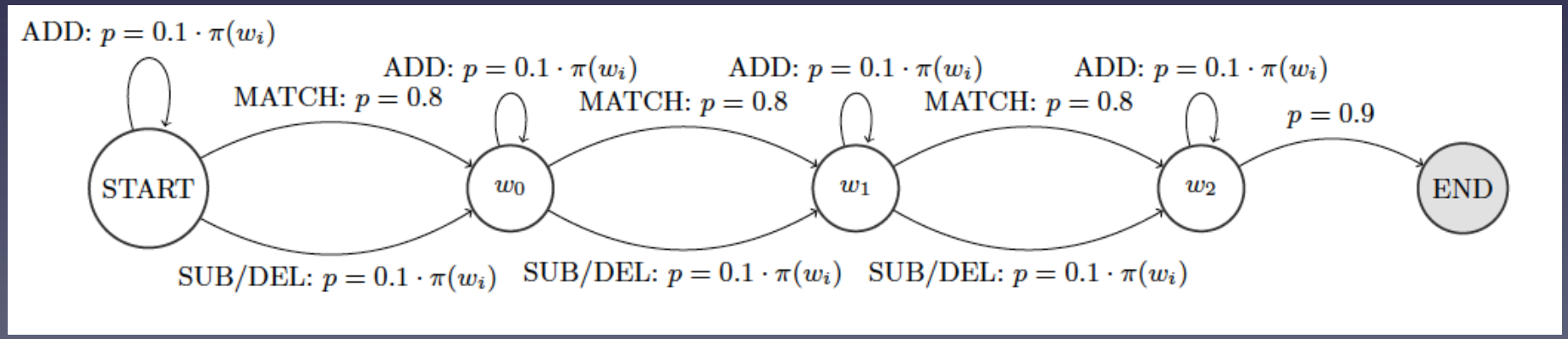
$$P(A) = \prod_{n,m} \prod_{i=0}^{I} \theta^{\mathbb{I}(a_{nm}=i)}$$

Text Passage Model: $P(d_{z_{nm}} \mid k_{a_{nm}})$

- Background text: unigram language model
- Ideas: Probabilistic finite state transducers (PFST)

$$P(d_z \mid k_a) = \text{stochastic edit distance between } d_z \text{ and } k_a$$

Example PFST for idea "$w_0\, w_1\, w_2$":



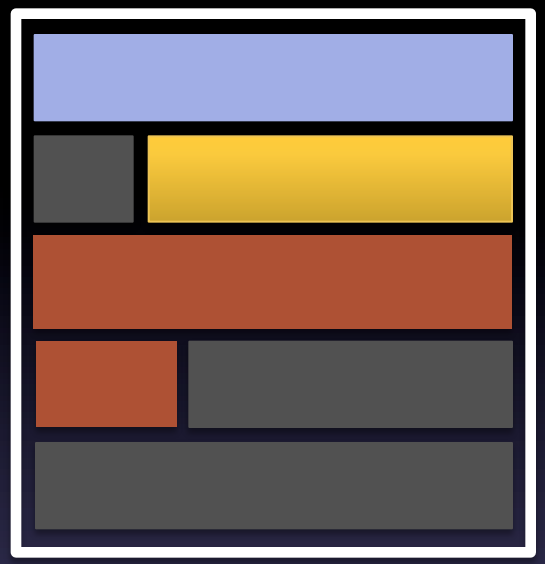
PTR: Model

Corpus:

- *D* documents (sequences of words)

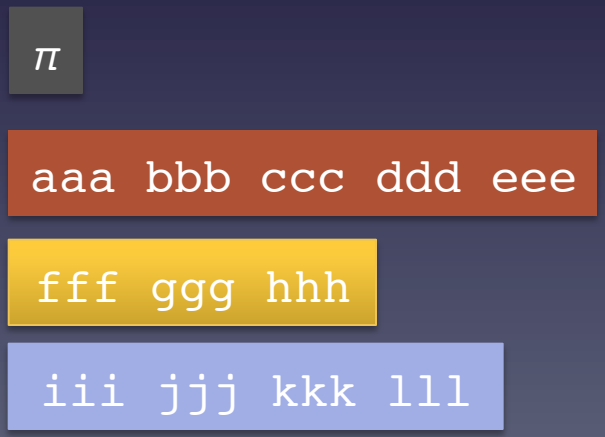
Text generators:

- Background text model, π
- *P(K)*



Generative process:

1. P(Z)
2. P(A)
3. $P(d_z|k_a)$



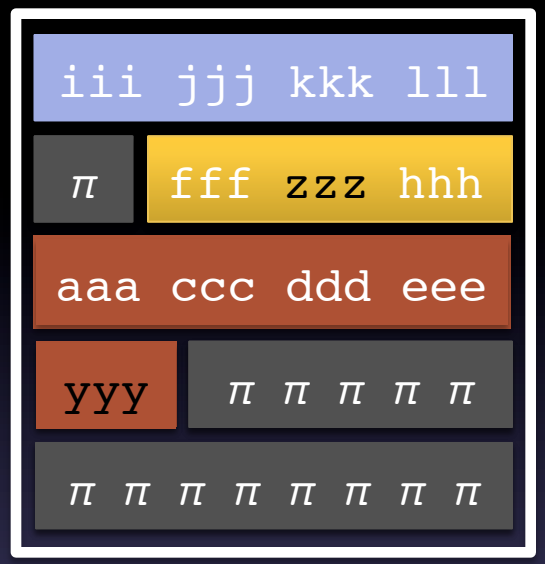
PTR: Model

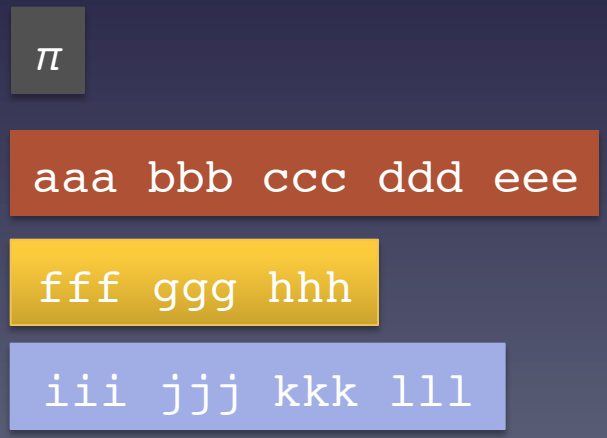
Corpus:

- *D* documents (sequences of words)

Text generators:

- Background text model, π
- *P(K)*





Generative process:

1. P(Z)
2. P(A)
3. $P(d_z|k_a)$

Parameter Estimation

- Goal: maximize $P(K,Z,A|D)$
- Coordinate descent:

  Iteratively optimize $P(d|k)$, $P(A)$, $P(K)$, $P(Z)$

  1. $P(d|k)$ and $P(A)$: assign pre-partitioned passages to best idea (or background)
  2. $P(K)$: re-estimate ideas given assigned partitions
  3. $P(Z)$: Propose concatenations and merges of ideas

Parameter Estimation: Assignments

For each passage:

1. Find highest-ranked candidate ideas
2. Compute PFST probabilities for top candidates (forward algorithm)
3. Choose assignment that maximizes probability

π

aaa bbb ccc ddd eee

fff ggg hhh

iii jjj kkk lll

fff zzz hhh

mmm nnn ooo

Parameter Estimation: Assignments

For each passage:

1. Find highest-ranked candidate ideas
2. Compute PFST probabilities for top candidates (forward algorithm)
3. Choose assignment that maximizes probability

π

aaa bbb ccc ddd eee

fff ggg hhh

iii jjj kkk lll

fff zzz hhh

mmm nnn ooo

Parameter Estimation: Ideas

- Find idea that maximizes sum of transduction probabilities (Steiner consensus string)
  - An NP-hard problem
- Approximation: Find best idea among current passages

Parameter Estimation: Partitions

- Proposals:
  - Concatenate ideas if they are frequently adjacent to each other
  - Merge most similar ideas
- Check if proposal increases log-likelihood

Evaluation

Dataset:

- FCC Net Neutrality comments
- 800,000 public comments (June-August 2014)
  - # unique comments: 650,300
  - # words/comment: mean = 131

Evaluation

- Use PTR to find:
  - Ideas with variations in wording
  - Template ideas
  - Document structure
- Infer partitions and assignments on held-out test set (160K documents)

Results: Word Variations

244 comments, "keep the internet a level playing field"

```
90 variations:
keep the internet an even playing field
keep the internet an open playing field
keep it a level playing field
keep net neutrality keep a level playing field
keep the net a level playing field
keep the internet a level playing field that it is
keep the internet a fair playing field
keep the net on a level playing field
keep the internet open as a fair playing field
keep the media landscape a level playing field
…
```

Results: Word Variations

168 comments, "the internet should be open"

```
21 variations:
the internet should be publicly owned
the internet should never be regulated
the internet should be divided
the internet should be open and neutral
the internet should be open to everyone equally
the internet should be an open platform
the internet should be equal opportunity
the internet should be taxed
the internet should be fair
the internet should absolutely be open
…
```

Results: Top Ideas

"Net neutrality is the First Amendment of the Internet, the principle that Internet service providers (ISPs) treat all data equally. As an Internet user, net neutrality is vitally important to me. The FCC should use its Title II authority to protect it…"

(89,989 comments)

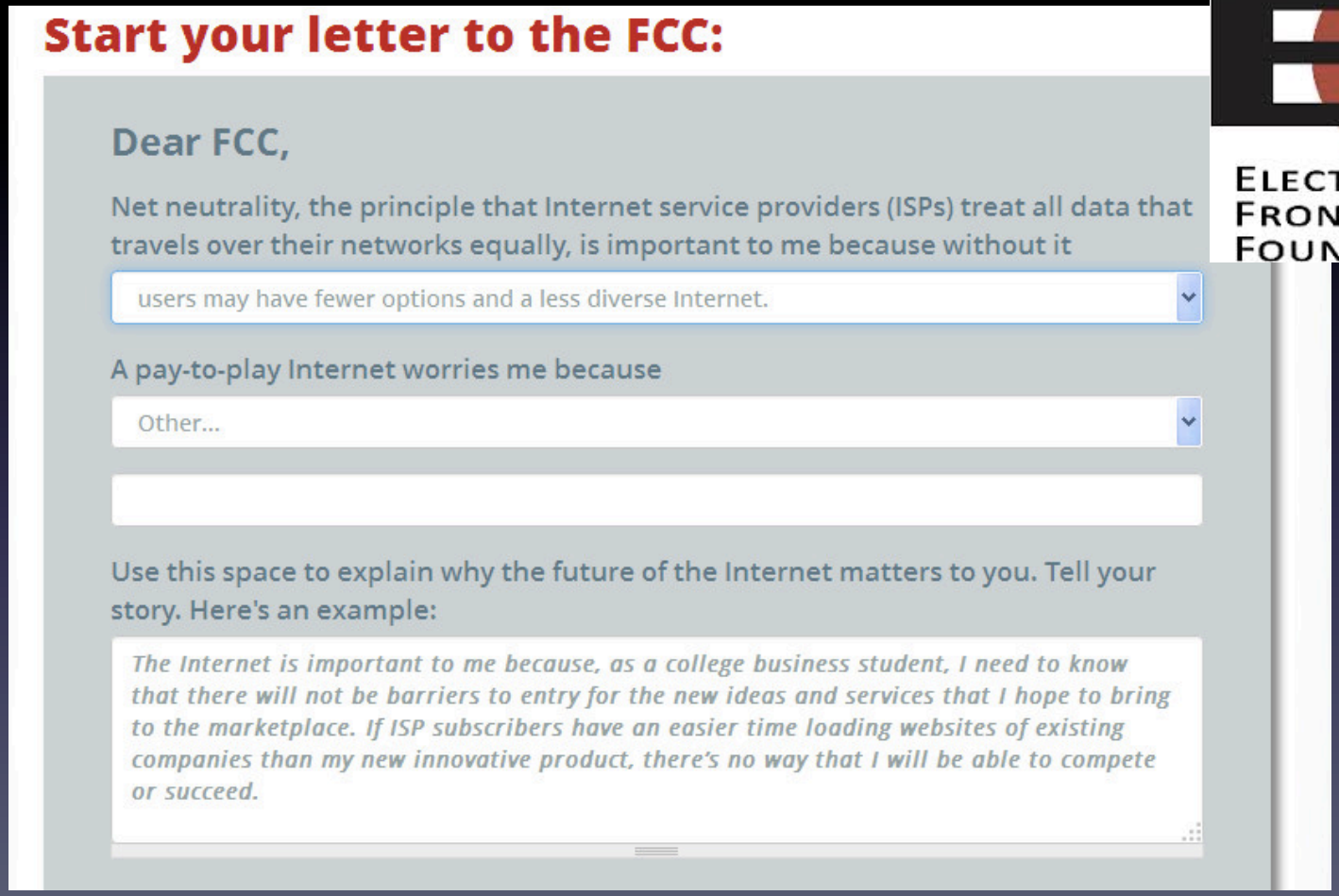
FIGHT FOR THE FUTURE

Results: Top Ideas

"As an American citizen, I wanted to voice my opposition to the FCC's crippling new regulations that would put federal bureaucrats in charge of internet freedom, and urge you to stop these regulations before they're enacted...Please stop the FCC's dangerous new regulations, and protect the future of internet freedom here in America."

(9331 comments)

Results: Document Structure



Results: Document Structure

"Dear FCC,

My name is Steve Roberts and I live in West Lafayette, IN. Net neutrality, the principle that Internet service providers (ISPs) treat all data that travels over their networks equally, is important to me because without it…

"…ISPs could have too much power to determine my Internet experience by providing better access to some services but not others." (41,524 comments)

"…users may have fewer options and a less diverse Internet." (28,173 comments)

Results: Document Structure

"A pay-to-play Internet worries me because...

"…new, innovative services that can't afford expensive fees for better service will be less likely to succeed." (36,509 comments)

"…ISPs could act as the gatekeepers to their subscribers." (32,252 comments)

Results: Idea Text Variations

The Internet is important to me because, as a IT professional, I need to know…

The Internet is important to me because, as a videographer & editor, I need to know…

The Internet is important to me because, as an artist and musician, I need to know…

The Internet is important to me because, as a business manager, I need to know…

Results: Document Structure

- Total EFF template comments: ~68,000

	"less diverse Internet"	"ISPs too much power"
"unaffordable expensive fees"	11,107	24,317
"pay-to-play Internet"	16,283	15,648

- Comments with >50 additional words: ~25,000

Results: Document Structure

Alex Jones, Manhattan, NY:
The internet's future is important to me because I am a Science student. By the time I enter the Science industry I hope to be able to use the internet to collaborate with out Scientists and teams around the world. If providers are allowed to choose which data is more important it may lead to the internet being used much less. This would remove any hopes of collaboration in the Science industry.

Frank Murphy, Colonia, NJ:
Look.. Long-story-short, if it ain't broke, don't fix it. If you want to improve anything why don't we start by upgrading infrastructure, giving lower-income and impoverished areas access to free Internet so that they may learn ways to improve their situation. Instead of catering to the powerful why don't we grant all citizens equal access to the pursuit of happiness as the founding American diplomats and their families intended.

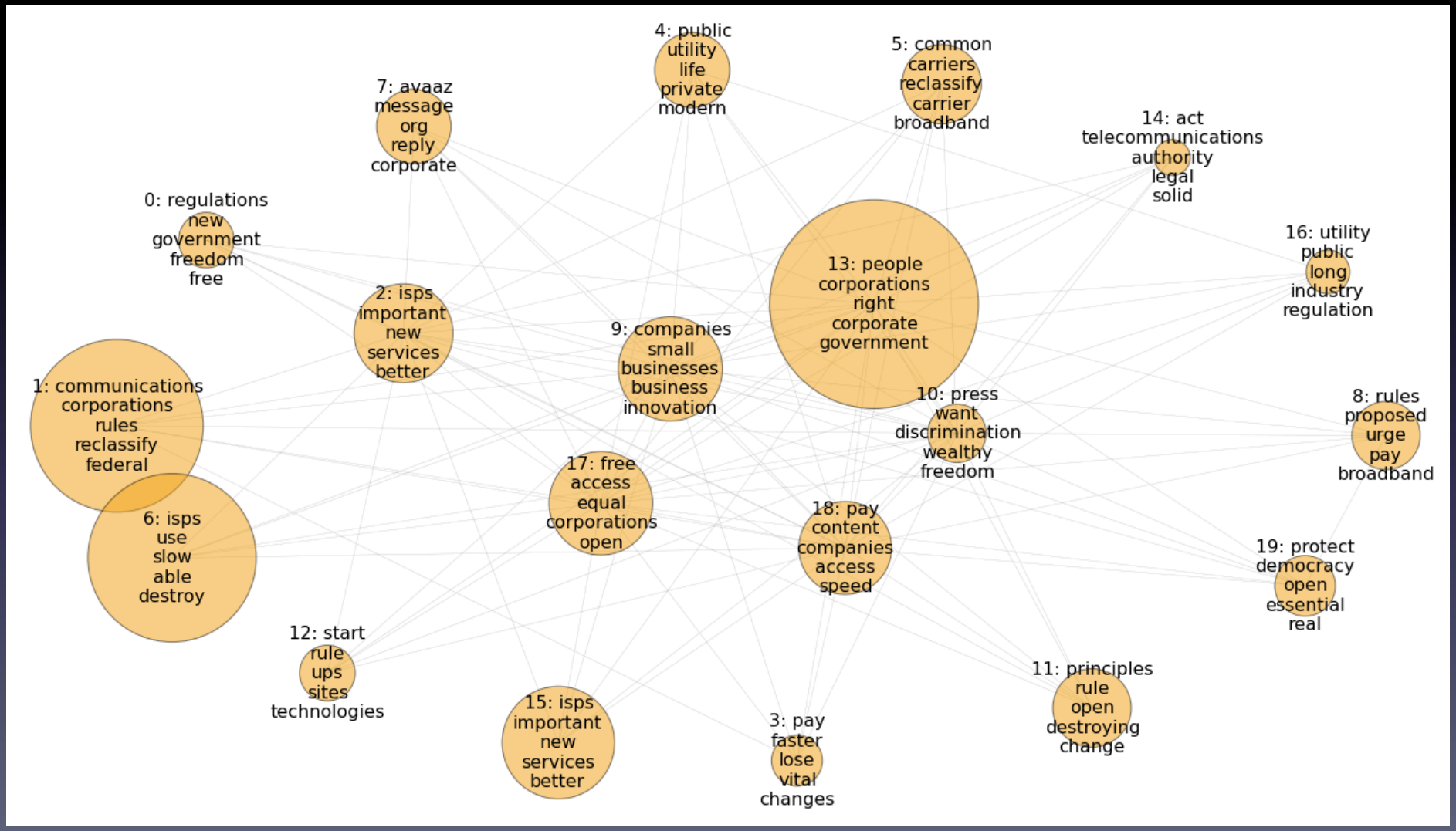
Kathleen Chisholm, Oakland, CA:
This country was founded on the idea of civil liberty, and the internet as it exists now is a great representation of how that ideal is fostered today. We have enough regulation in our everyday practices. The internet should remain exactly like it is; an open forum where everybody has an equal opportunity to have their voice heard.

Comparison: Topic Modeling

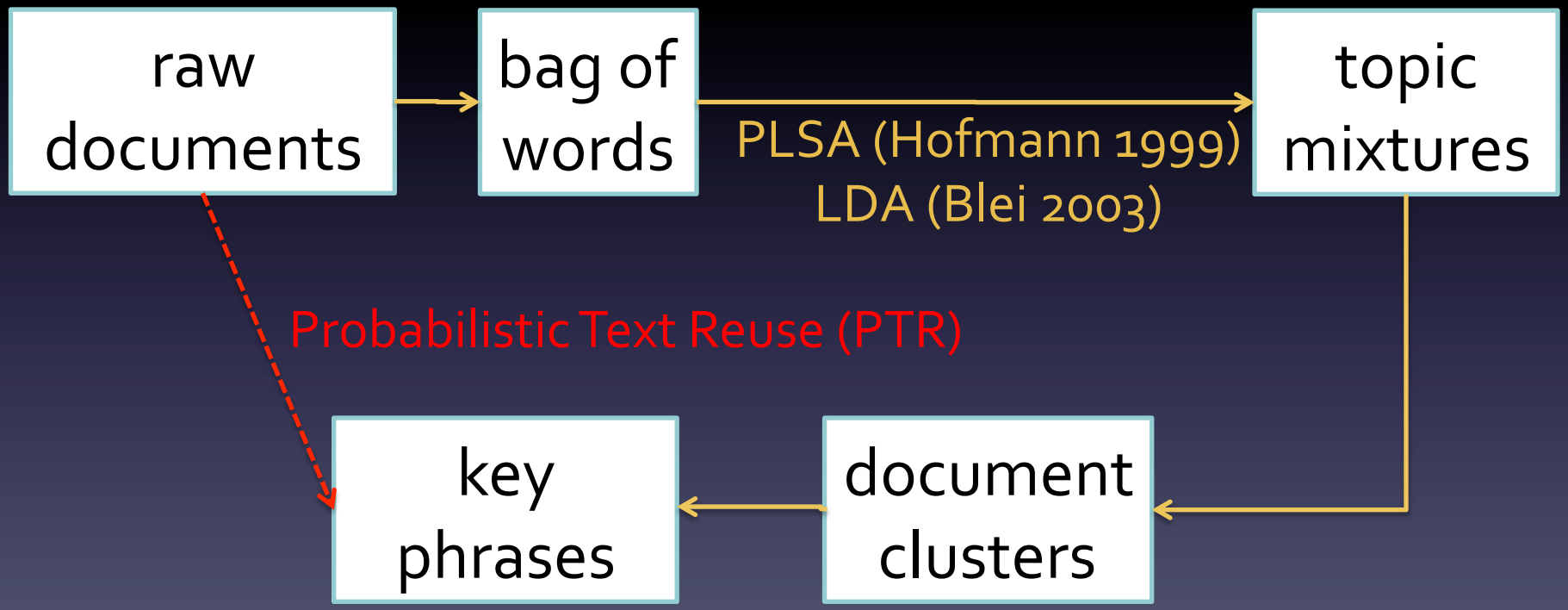
Top words from top 10 topics: coherent but difficult to interpret

- isps, use, slow, able, destroy
- isps, important, new, services, better
- communications, corporations, rules, reclassify, federal
- free, access, equal, corporations, open
- common, carriers, reclassify, carrier, broadband
- companies, small, businesses, business, innovation
- people, corporations, right, corporate, government
- pay, content, companies, access, speed
- wealthy, save, user, addition, believes
- comcast, like, verizon, cable, time

Comparison: Topic Modeling



Existing Work: Probabilistic Topic Modeling



- Text passages → bags of words → text passages

PTR: Quantitative Evaluation

	log-likelihood/token (higher is better)	
model	**training set**	**test set**
baseline unigram	-6.88	-14.01
100-topic LDA	-5.40	-12.44
PTR		

PTR: Quantitative Evaluation

	log-likelihood/token (higher is better)	
model	**training set**	**test set**
baseline unigram	-6.88	-14.01
100-topic LDA	-5.40	-12.44
PTR	-3.26	-10.24

- Percentage of comments with substantial non-idea text (>20 words): 37.4%
  - (Knight Foundation 2014 Report: 40%)

PTR: Summary

- A novel probabilistic model for text reuse
  - Latent partitions, assignments, and ideas (PFSTs)
- Useful for understanding large document collections (e.g. public speech)
  - Identifying common viewpoints, less-popular voices, and variations of ideas