

Connections between the Lines: Augmenting Social Networks with Text

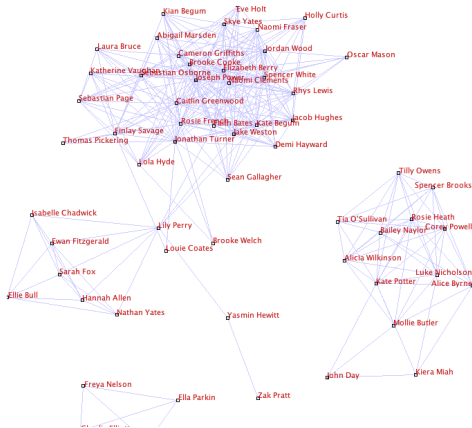
Jonathan Chang, Jordan Boyd-Graber, David M. Blei

KDD 09
June 29th, 2009



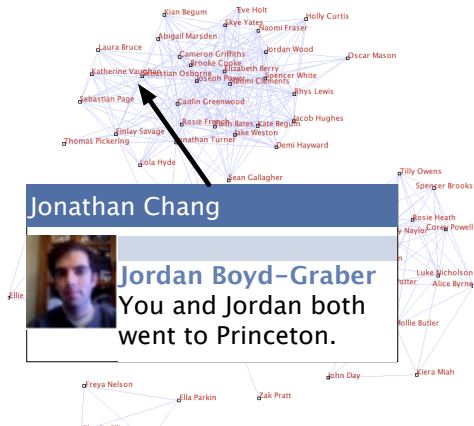
Motivation

Data that express relationships between ensembles of entities (people, places, companies, genes, etc.) is pervasive.



Motivation

Annotating the edges of these networks with rich information is often a desideratum.



Motivation

Unfortunately, much of the rich information about links is encoded *implicitly* as free text.

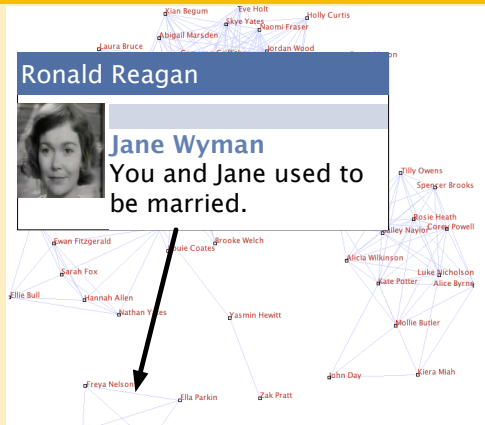
Example

In 1938, Wyman co-starred with Ronald Reagan. Reagan and actress Jane Wyman were engaged at the Chicago Theater and married in Glendale, California. Following arguments about Reagan's political ambitions, Wyman filed for divorce in 1948. Since Reagan is the only U.S. president to have been divorced, Wyman is the only ex-wife of an American President.

Motivation

Using the text as input, our goal is to annotate a network with this information.

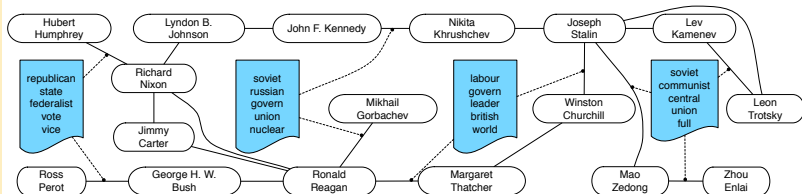
Example



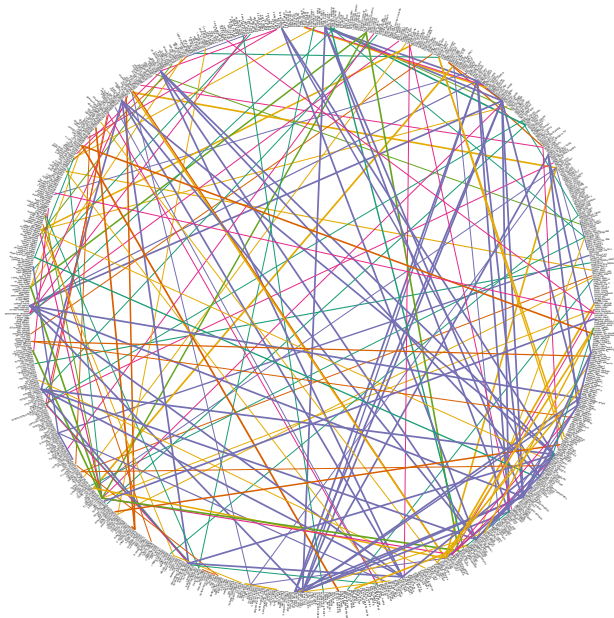
Main idea

- We define a probabilistic model **Networks Uncovered By Bayesian Inference (Nubbi)**.
- Using Nubbi, we can create a social network by discovering relations over the ensemble of people.
- Because Nubbi is a fully Bayesian approach, we can incorporate information from other data-mining approaches as priors (Agichtein and Gravano 2003; Diehl et al. 2007; Mei et al. 2007; Sahay et al. 2008; Banko et al. 2007; Katrenko and Adriaans 2007; Davidov et al. 2007).

Wikipedia



A larger example...



Topic Models

- Nubbi leverages the machinery of *topic modeling* (Blei et al. 2003; Hofmann 1999; Steyvers and Griffiths 2007), Bayesian mixture models of discrete data.
- Topic models have emerged as powerful tools for unsupervised analysis of large document collections.
- The multinomial parameters of the mixture components are known as “topics.”

Example (NYTimes using LDA)

LAW	ART	POLITICS	SPORTS
lawyer	music	republican	game
justice	film	democrat	coach
judge	artist	senate	player
investigate	art	campaign	play
prosecutor	ballet	mayor	match

Model input

We take as input text whose references to entities have been identified, and convert this into a collection of documents.

1 When **Jesus** had spoken these words, he went forth with his disciples over the brook Cedron, where was a garden, into the which he entered, and his disciples.

2 And **Judas** also, which betrayed him, knew the place: for **Jesus** oftentimes resorted thither with his disciples.

3 **Judas** then, having received a band of men and officers from the chief priests and Pharisees, cometh thither with lanterns and torches and weapons.

4 **Jesus** therefore, knowing all things that should come upon him, went forth, and said unto them, Whom seek ye?

5 They answered him, **Jesus of Nazareth**. **Jesus** saith unto them, I am he. And **Judas** also, which betrayed him, stood with them.

6 As soon then as he had said unto them, I am he, they went backward, and fell to the ground.

7 Then asked he them again, Whom seek ye? And they said, **Jesus of Nazareth**.

Model input

Jesus

spoken words disciples brook Cedron
garden enter disciples knowing things
seek asked seek Nazareth

1 When **Jesus** had spoken these words, he went forth with his disciples over the brook Cedron, where was a garden, into the which he entered, and his disciples.

2 And **Judas** also, which betrayed him, knew the place: for **Jesus** oftentimes resorted thither with his disciples.

3 **Judas** then, having received a band of men and officers from the chief priests and Pharisees, cometh thither with lanterns and torches and weapons.

4 **Jesus** therefore, knowing all things that should come upon him, went forth, and said unto them, Whom seek ye?

5 They answered him, **Jesus of Nazareth**. **Jesus** saith unto them, I am he. And **Judas** also, which betrayed him, stood with them.

6 As soon then as he had said unto them, I am he, they went backward, and fell to the ground.

7 Then asked he them again, Whom seek ye? And they said, **Jesus of Nazareth**.

received band officers chief
priests Pharisees lanterns
torches weapons

Judas

Model input

1 When **Jesus** had spoken these words, he went forth with his disciples over the brook Cedron, where was a garden, into the which he entered, and his disciples.

2 And **Judas** also, which betrayed him, knew the place: for **Jesus** oftentimes resorted thither with his disciples.

3 **Judas** then, having received a band of men and officers from the chief priests and Pharisees, cometh thither with lanterns and torches and weapons.

4 **Jesus** therefore, knowing all things that should come upon him, went forth, and said unto them, Whom seek ye?

5 They answered him, **Jesus of Nazareth**. **Jesus** saith unto them, I am he. And **Judas** also, which betrayed him, stood with them.

6 As soon then as he had said unto them, I am he, they went backward, and fell to the ground.

7 Then asked he them again, Whom seek ye? And they said, **Jesus of Nazareth**.

A diagram consisting of a white rectangular box with a black border. Four arrows point from the text above to the box: a blue arrow from 'disciples' in paragraph 1, a red arrow from 'betrayed' in paragraph 2, a red arrow from 'Jesus' in paragraph 4, and a blue arrow from 'Jesus' in paragraph 5. The box contains the following text: 'betrayed knew place disciples answered' on the top line and 'Nazareth saith betrayed' on the bottom line.

betrayed knew place disciples answered
Nazareth saith betrayed

Jesus
and
Judas

Intuition

- Both individuals and pairs of people can be described by topics.
- Nubbi hypothesizes that each word in a pair context describes either one of the entities or their relationship.

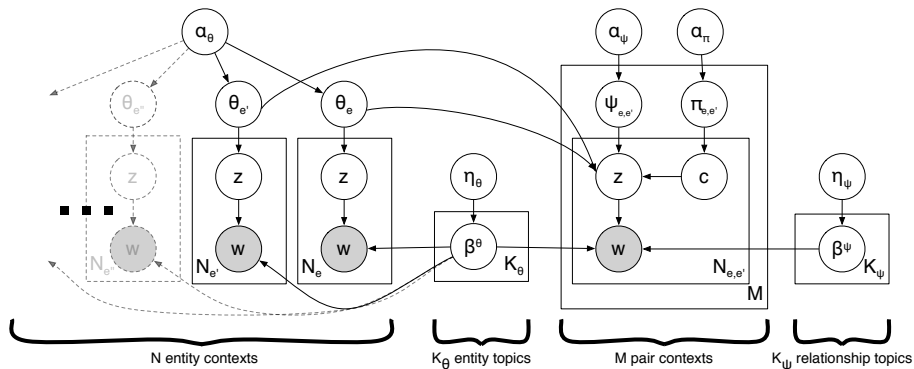
In 1938, Wyman co-starred with Ronald Reagan. Reagan and actress Jane Wyman were engaged at the Chicago Theater and married in Glendale, California. Following arguments about Reagan's political ambitions, Wyman filed for divorce in 1948. Since Reagan is the only U.S. president to have been divorced, Wyman is the only ex-wife of an American President.

Intuition

- Red words can be explained by a POLITICS topic.
- Blue words can be explained by an ACTING topic.
- POLITICS words can be attributed to Ronald Reagan.
- ACTING words can be attributed to either Ronald Reagan or Jane Wyman.
- The remaining words characterize their relationship.

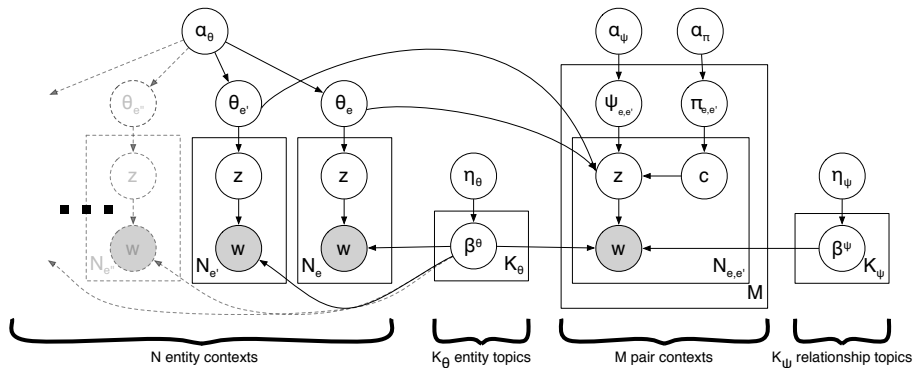
In 1938, Wyman co-starred with Ronald Reagan. Reagan and actress Jane Wyman were engaged at the Chicago Theater and married in Glendale, California. Following arguments about Reagan's political ambitions, Wyman filed for divorce in 1948. Since Reagan is the only U.S. president to have been divorced, Wyman is the only ex-wife of an American President.

Graphical Model



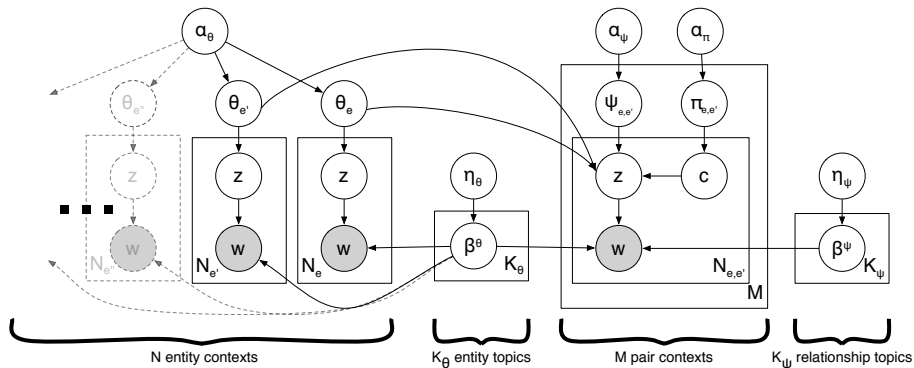
- These intuitions are encoded in this graphical model.

Graphical Model



- We must then infer the posterior probability over the hidden variables (e.g., the hidden topics and the assignments of topics to people and relationships) given the data.

Graphical Model



- The posterior lacks the structure for efficient computation, so exact posterior inference is intractable. We appeal instead to *mean-field variational inference* (Jordan et al. 1999).

Bible Topics

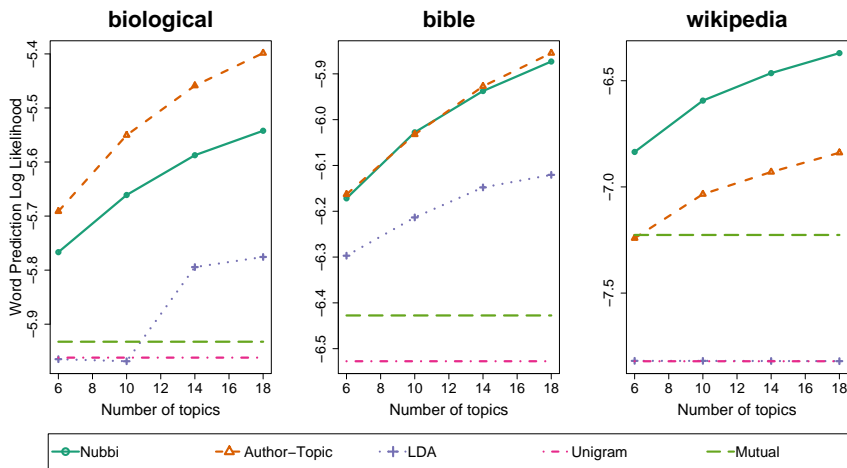
- Relationships are described by *relationship topics* (c.f. Bhattacharya et al. 2008; Newman et al. 2006; Rosen-Zvi et al. 2004; Culotta et al. 2005; Rabbat et al. 2006).
- Topics cluster words which are used to describe similar relationships and pairs who participate in similar relationships.

People in the Bible

	Topic 1	Topic 2
Top pairs	Jacob-Laban Miriam-Moses Ishmael-Sarah	Adonizedek-Piram Abraham-Birsha Birsha-Lot
Top words	wife abraham daughter child call	king cave lord smote great

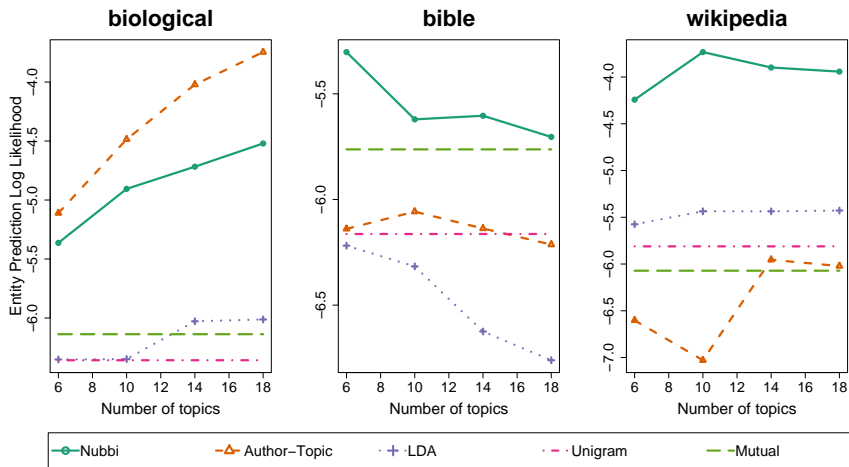
Describing relationships

Nubbi can predict words in each entity pair's text profile.
This describes the relationship between the entities.

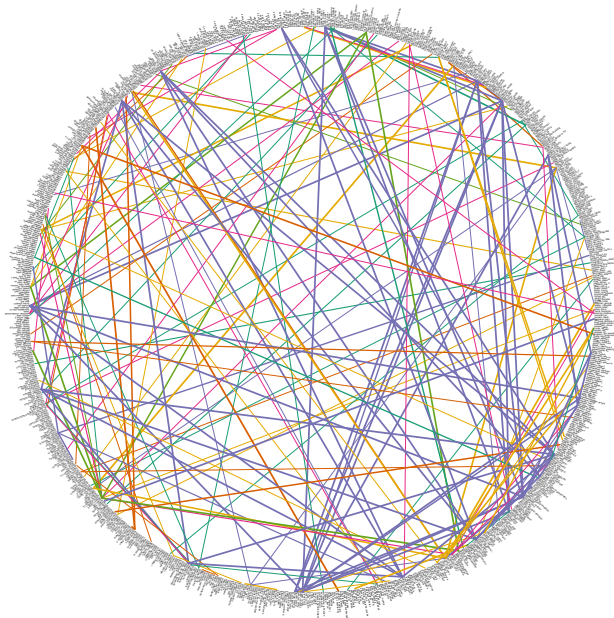


Predicting entities

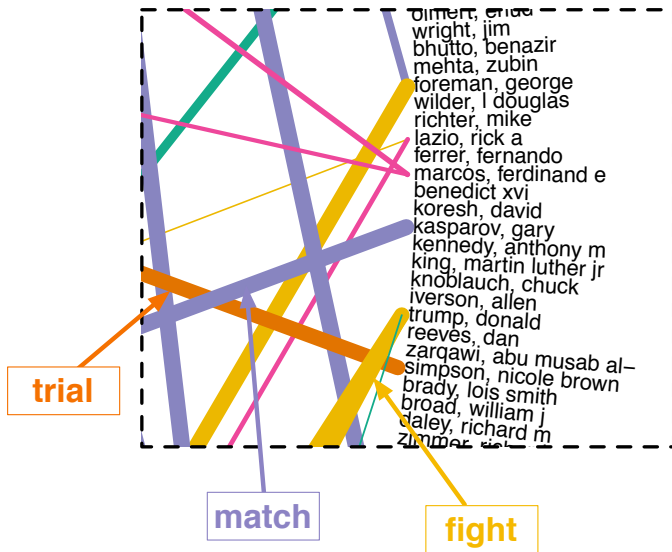
Nubbi can also predict which entities are best described by a text profile.



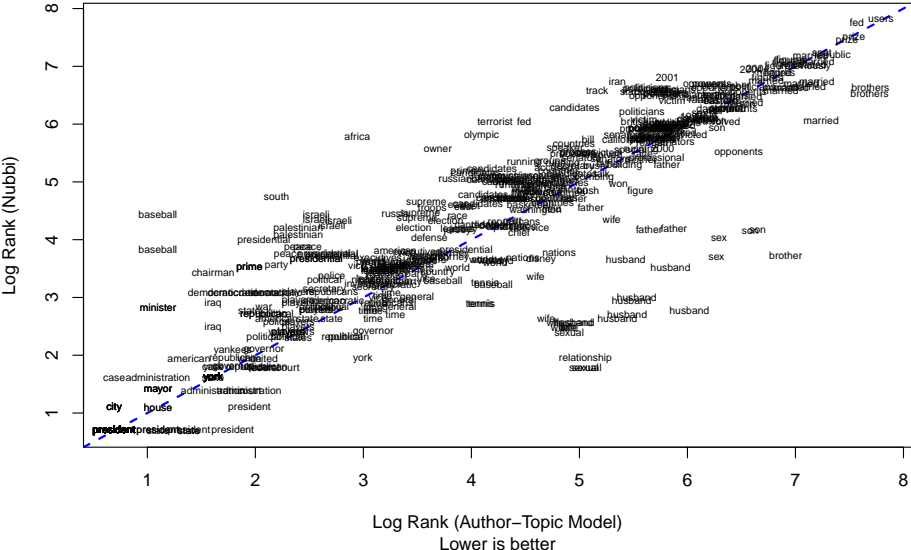
NY Times social network



NY Times social network



Examining predictions on solicited descriptions



Conclusion

- Networks of entities are a pervasive form of data.
- There are deep meanings associated with the links in these networks.
- These meanings are often not encoded in the network itself, but rather as free text elsewhere.
- The Nubbi model allows us to take this text and infer characterizations of the relationships between the ensemble of entities.
- This is an important step towards mining complete social networks from free text.

Thanks!

Wikipedia Topics

- Nubbi learns topics associated with individual entities.
- This clusters entities according to the words used to describe them.

People in Wikipedia

	Topic 1	Topic 2
Top Entities	Frederick Sanger Svante Arrhenius William Ramsay	Kate Winslet Ringo Starr Al Pacino
Top Terms	work universe year develop society	film music album award song

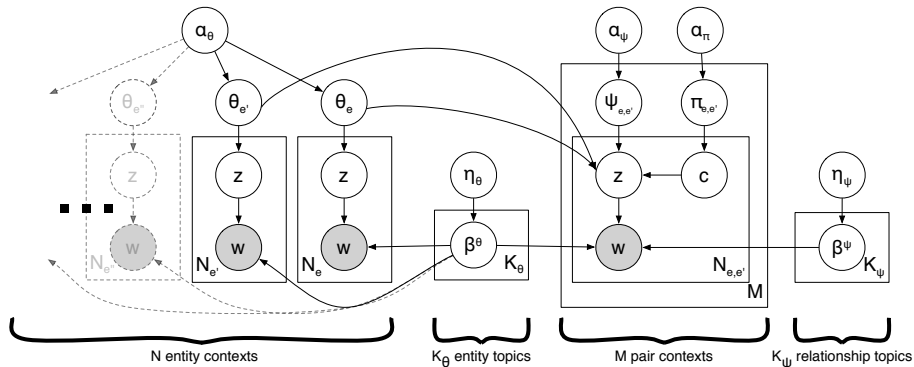
Wikipedia Topics

- Nubbi also learns topics associated with *pairs* of entities.
- These topics cluster together pairs with similar relationships.

People in Wikipedia

	Topic 1	Topic 2
Top Pairs	Tyler-Roosevelt McKinley-Roosevelt Taft-Roosevelt	Perot-Bush J.Q. Adams-Monroe J.Q. Adams-Clay
Top Terms	president vice roosevelt calvin johnson	republican march reagan vote state

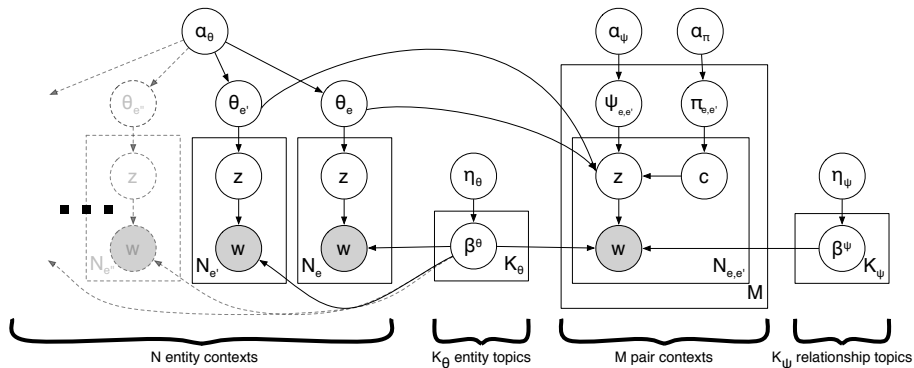
Graphical Model



1 For each entity e ,

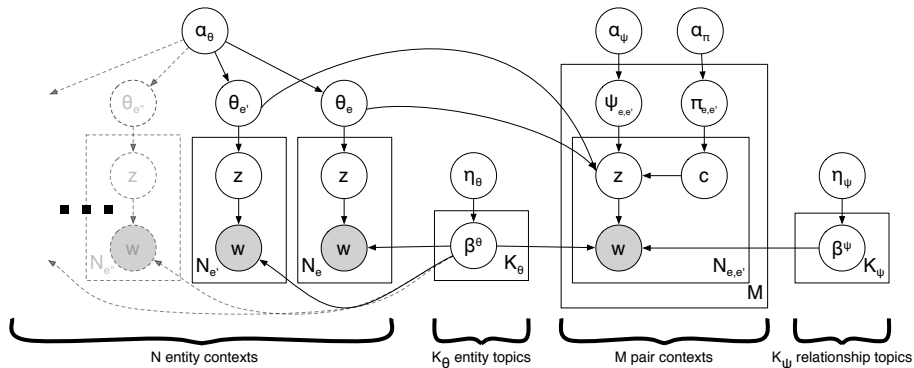
- 1** Draw entity topic proportions $\theta_e \sim \text{Dir}(\alpha_\theta)$;
- 2** For each word associated with this entity's context,
 - 1** Draw topic assignment $z_{e,n} \sim \text{Mult}(\theta_e)$;
 - 2** Draw word $w_{e,n} \sim \text{Mult}(\beta_{z_{e,n}}^\theta)$.

Graphical Model



- 1 For each pair of entities e, e' ,
 - 1 Draw relationship topic proportions $\psi_{e,e'} \sim \text{Dir}(\alpha_\psi)$;
 - 2 Draw selector proportions $\pi_{e,e'} \sim \text{Dir}(\alpha_\pi)$;
 - 3 For each word associated with this entity pair's context,
 - 1 Draw selector $c_{e,e',n} \sim \text{Mult}(\pi_{e,e'})$;

Graphical Model

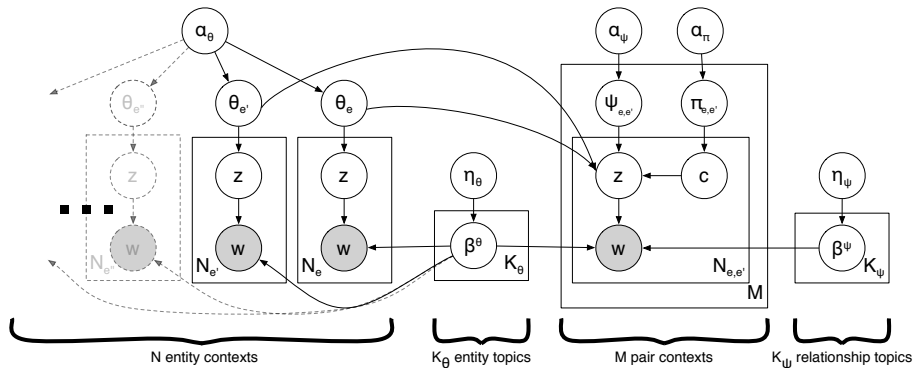


1 If $c_{e,e',n} = 1$,

1 Draw topic assignment $z_{e,e',n} \sim \text{Mult}(\theta_e)$;

2 Draw word $w_{e,e',n} \sim \text{Mult}(\beta_{z_{e,e',n}}^\theta)$.

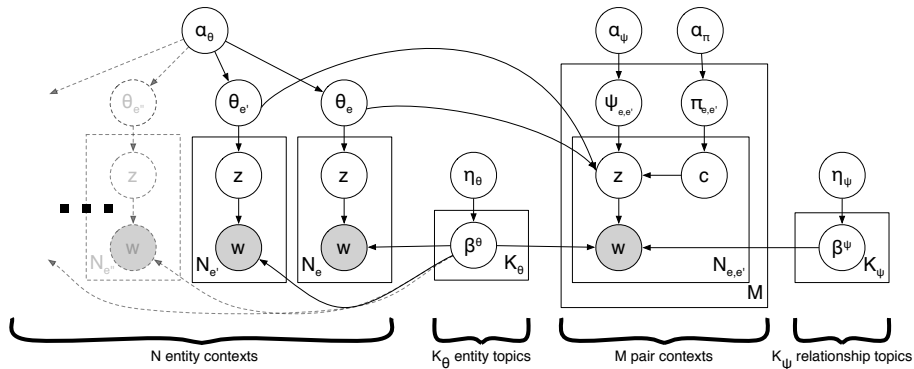
Graphical Model



1 If $c_{e,e',n} = 2$,

- 1 Draw topic assignment $z_{e,e',n} \sim \text{Mult}(\theta_{e'})$;
- 2 Draw word $w_{e,e',n} \sim \text{Mult}(\beta_{z_{e,e',n}}^\theta)$.

Graphical Model



1 If $c_{e,e',n} = 3$,

1 Draw topic assignment $z_{e,e',n} \sim \text{Mult}(\psi_{e,e'})$;

2 Draw word $w_{e,e',n} \sim \text{Mult}(\beta_{z_{e,e',n}}^\psi)$.

References

- Eugene Agichtein and Luis Gravano. Querying text databases for efficient information extraction. *Data Engineering, International Conference on*, 0:113, 2003. ISSN 1063-6382. doi: <http://doi.ieeecomputersociety.org/10.1109/ICDE.2003.1260786>.
- Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI 2007*, 2007. URL <http://www.ijcai.org/papers07/Papers/IJCAI07-429.pdf>.
- Indrajit Bhattacharya, Shantanu Godbole, and Sachindra Joshi. Structured entity identification and document categorization: Two tasks with one joint model. *KDD 2008*, 2008. URL http://portal.acm.org/ft_gateway.cfm?id=1401899&type=pdf&coll=ACM&dl=ACM&CFID=2420023&CFTOKEN=56734162.
- D Blei, A Ng, and M Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003. URL <http://www.mitpressjournals.org/doi/abs/10.1162/imlr.2003.3.4-5.993>.