Jonathan Chang, **Jordan Boyd-Graber**, and David M. Blei. **Connections between the Lines: Augmenting Social Networks with Text**. *Refereed Conference on Knowledge Discovery and Data Mining*, 2009.

```
    @inproceedings{Chang:Boyd-Graber:Blei-2009,
Title = {Connections between the Lines: Augmenting Social Networks with Text},
Booktitle = {Refereed Conference on Knowledge Discovery and Data Mining},
Author = {Jonathan Chang and Jordan Boyd-Graber and David M. Blei},
Year = {2009},
Location = {Paris, France},
}
```

Connections between the Lines: Augmenting Social Networks with Text

Jonathan Chang
Electrical Engineering
Engineering Quadrangle
Princeton, NJ 08544
jcone@princeton.edu

Jordan Boyd-Graber
Computer Science
35 Olden St.
Princeton, NJ 08544
jbg@cs.princeton.edu

David M. Blei
Computer Science
35 Olden St.
Princeton, NJ 08544
blei@cs.princeton.edu

ABSTRACT

Network data is ubiquitous, encoding collections of relationships between entities such as people, places, genes, or corporations. While many resources for networks of interesting entities are emerging, most of these can only annotate connections in a limited fashion. Although relationships between entities are rich, it is impractical to manually devise complete characterizations of these relationships for every pair of entities on large, real-world corpora.

In this paper we present a novel probabilistic topic model to analyze text corpora and infer descriptions of its entities and of relationships between those entities. We develop variational methods for performing approximate inference on our model and demonstrate that our model can be practically deployed on large corpora such as Wikipedia. We show qualitatively and quantitatively that our model can construct and annotate graphs of relationships and make useful predictions.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms

Keywords

statistical topic models, social network learning, graphical models

1. INTRODUCTION

Network data—data which express relationships between ensembles of entities—are becoming increasingly pervasive. People are connected to each other through a variety of kinship, social, and professional relationships; proteins bind to and interact with other proteins; corporations conduct business with other corporations. Understanding the nature of these relationships can provide useful mechanisms for suggesting new relationships between entities, characterizing new relationships, and quantifying global properties of naturally occurring network structures [2, 6, 31, 33, 34].

Many corpora of network data have emerged in recent years. Examples of such data include social networks, such as LinkedIn or Facebook, and citation networks, such as CiteSeer, Rexa, or JSTOR. Other networks can be constructed manually or automatically using texts with people such as the Bible, scientific abstracts with genes, or decisions in legal journals. Characterizing the networks of connections between these entities is of historical, scientific, and practical interest. However, describing every relationship for large, real-world corpora is infeasible. Thus most data sets label edges as merely on or off, or with a small set of fixed, predefined connection types. These labellings cannot capture the complexities underlying the relationships and limit the applicability of these data sets.

In this paper we develop a method for augmenting such data sets by analyzing document collections to uncover the relationships encoded in their texts. Text corpora are replete with information about relationships, but this information is out of reach for traditional network analysis techniques. We develop *Networks Uncovered By Bayesian Inference* (Nubbi), a probabilistic topic model of text [5, 12, 29] with hidden variables that represent the patterns of word use which describes the relationships in the text. Given a collection of documents, Nubbi reveals the hidden network of relationships that is encoded in the texts by associating rich descriptions with each entity and its connections. For example, Figure 1 illustrates a subset of the network uncovered from the texts of Wikipedia. Connections between people are depicted by edges, each of which is associated with words that describe the relationship.

First, we describe the intuitions and statistical assumptions behind Nubbi. Second, we derive efficient algorithms for using Nubbi to analyze large document collections. Finally, we apply Nubbi to the Bible, Wikipedia, and scientific abstracts. We demonstrate that Nubbi can discover sensible descriptions of the network and can make predictions competitive with those made by state of the art models.

2. MODEL

The goal of Nubbi is to analyze a corpus to describe the relationships between pairs of entities. Nubbi takes as input very lightly annotated data, requiring only that entities within the input text be identified. Nubbi also takes as input the network of entities desired to be annotated. For some corpora this network is already explicitly encoded as

KDD'09, June 28–July 1, 2009, Paris, France.

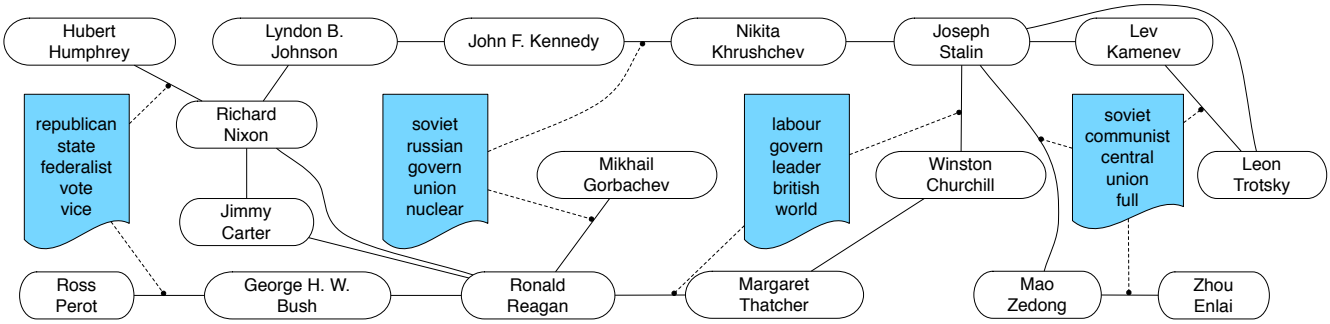

Figure 1: A small subgraph of the social network Nubbi learned taking only the raw text of Wikipedia with tagged entities as input. The full model uses 25 relationship and entity topics. An edge exists between two entities if their co-occurrence count is high. For some of the edges, we show the top words from the most probable relationship topic associated with that pair of entities. These are the words that best explain the contexts where these two entities appear together. A complete browser for this data is available at *http://topics.cs.princeton.edu/nubbi*.

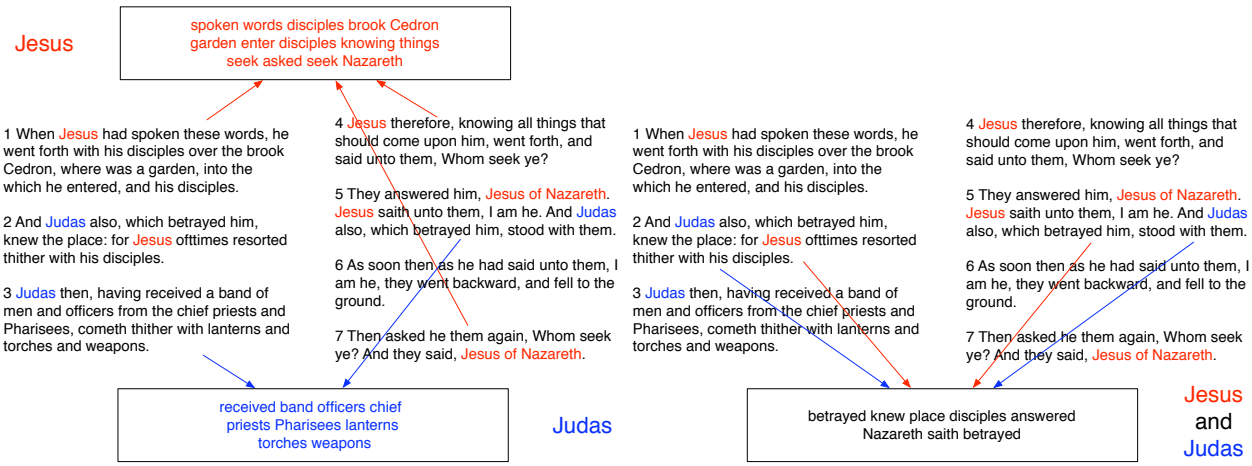

Figure 2: A high-level overview of Nubbi's view of text data. A corpus with identified entities is turned into a collection of bags-of-words (in rectangles), each associated with individual entities (left) or pairs of entities (right). The procedure in the left panel is repeated for every entity in the text while the procedure in the right panel is repeated for every pair of entities.

a graph. For other text corpora this graph must be constructed. One simple way of constructing this graph is to use a fully-connected network of entities. One can prune the edges in this graph using simple statistics such as entity-entity co-occurrence counts.

From the entities in this network, the text is divided into two different classes of bags of words. First, each entity is associated with an *entity context*, a bag of words co-located with the entity. Second, each pair of entities is associated with a *pair context*, a bag of words co-located with the pair. Figure 2 shows an example of the input to the algorithm turned into entity contexts and pair contexts.

Nubbi learns two descriptions of how entities appear in the corpus: entity topics and relationship topics. Following [5], a *topic* is defined to be a distribution over words. To aid intuitions, we will for the moment assume that these topics are given and have descriptive names. We will describe how the topics and contexts interplay to reveal the network of relationships hidden in the texts. We emphasize, however, that the goal of Nubbi is to analyze the texts to learn both the topics and relationships between entities.

An *entity topic* is a distribution over words, and each entity is associated with a distribution over entity topics. For example, suppose there are three entity topics: POLITICS, MOVIES, and SPORTS. Ronald Reagan would have a distribution that favors POLITICS and MOVIES, athlete actors like Johnny Weissmuller and Geena Davis would have distributions that favor MOVIES and SPORTS, and specialized athletes, like Pelé, would have distributions that favor SPORTS more than other entity topics. Nubbi uses entity topics to model entity contexts. Because the SPORTS entity topic would contain words like "cup," "win," and "goal," associating Pelé exclusively with the SPORTS entity topic would be consistent with the words observed in his context.

Relationship topics are distributions over words associated with pairs of entities, rather than individual entities, and each pair of entities is associated with a distribution over relationship topics. Just as the entity topics cluster similar people together (e.g., Ronald Reagan, George Bush, and Bill Clinton all express the POLITICS topic), the relation-

ship topics can cluster similar *pairs* of people. Thus, Romeo and Juliet, Abelard and Heloise, Ruslan and Ludmilla, and Izanami and Izanagi might all share a LOVERS relationship topic.

Relationship topics are used to explain pair contexts. Each word in a pair context is assumed to express something about either one of the participating entities or something particular to their relationship. For example, consider Jane Wyman and Ronald Reagan. (Jane Wyman, an actress, was actor/president Ronald Reagan's first wife.) Individually, Wyman is associated with the MOVIES entity topic and Reagan is associated with the MOVIES and POLITICS entity topics. In addition, this pair of entities is associated with relationship topics for DIVORCE and COSTARS.

Nubbi hypothesizes that each word describes either one of the entities or their relationship. Consider the pair context for Reagan and Wyman:

> 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.

We have marked the words that are not associated with the relationship topic. Functional words are gray; words that come from a POLITICS topic (associated with Ronald Reagan) are underlined; and words that come from a MOVIES topic (associated with Jane Wyman) are *italicized*.

The remaining words, "1938," "co-starred," "engaged," "Glendale," "filed," "divorce," "1948," "divorced," and "ex-wife," describe the relationship between Reagan and Wyman. Indeed, it is by deducing which case each word falls into that Nubbi is able to capture the relationships between entities. Examining the relationship topics associated with each pair of entities provides a description of that relationship.

The above discussion gives an intuitive picture of how Nubbi explains the observed entity and pair contexts using entity and relationship topics. In data analysis, however, we do not observe the entity topics, pair topics, or the assignments of words to topics. Our goal is to discover them.

To do this, we formalize these notions in a generative probabilistic model of the texts that uses hidden random variables to encode the hidden structure described above. In *posterior inference*, we "reverse" the process to discover the latent structure that best explains the documents. (Posterior inference is described in the next section.) More formally, Nubbi assumes the following statistical model.

1. For each entity topic j and relationship topic k,
   (a) Draw topic multinomials $\beta^{\theta}_{j} \sim \text{Dir}(\eta_\theta + 1)$, $\beta^{\psi}_{k} \sim \text{Dir}(\eta_\psi + 1)$
2. For each entity e,
   (a) Draw entity topic proportions $\theta_e \sim \text{Dir}(\alpha_\theta)$;
   (b) For each word associated with this entity's context,
      i. Draw topic assignment $z_{e,n} \sim \text{Mult}(\theta_e)$;
      ii. Draw word $w_{e,n} \sim \text{Mult}(\beta^{\theta}_{z_{e,n}})$.
3. For each pair of entities e, e',
   (a) Draw relationship topic proportions $\psi_{e,e'} \sim \text{Dir}(\alpha_\psi)$;
   (b) Draw selector proportions $\pi_{e,e'} \sim \text{Dir}(\alpha_\pi)$;
   (c) For each word associated with this entity pair's context,
      i. Draw selector $c_{e,e',n} \sim \text{Mult}(\pi_{e,e'})$;
      ii. If $c_{e,e',n} = 1$,
         A. Draw topic assignment $z_{e,e',n} \sim \text{Mult}(\theta_e)$;
         B. Draw word $w_{e,e',n} \sim \text{Mult}(\beta^{\theta}_{z_{e,e',n}})$.
      iii. If $c_{e,e',n} = 2$,
         A. Draw topic assignment $z_{e,e',n} \sim \text{Mult}(\theta_{e'})$;
         B. Draw word $w_{e,e',n} \sim \text{Mult}(\beta^{\theta}_{z_{e,e',n}})$.
      iv. If $c_{e,e',n} = 3$,
         A. Draw topic assignment $z_{e,e',n} \sim \text{Mult}(\psi_{e,e'})$;
         B. Draw word $w_{e,e',n} \sim \text{Mult}(\beta^{\psi}_{z_{e,e',n}})$.

This is depicted in a graphical model in Figure 3.

The hyperparameters of the Nubbi model are Dirichlet parameters α_θ, α_ψ, and α_π, which govern the entity topic distributions, the relationship distributions, and the entity/pair mixing proportions. The Dirichlet parameters η_θ and η_ψ are priors for each topic's multinomial distribution over terms. There are K_θ per-topic term distributions for entity topics, $\beta^{\theta}_{1:K_\theta}$, and K_ψ per-topic term distributions $\beta^{\psi}_{1:K_\psi}$ for relationship topics.

The words of each entity context are essentially drawn from an LDA model using the entity topics. The words of each pair context are drawn in a more sophisticated way. The topic assignments for the words in the pair context for entity e and entity e' are hypothesized to come from the entity topic proportions θ_e, entity topic proportions $\theta_{e'}$, or relationship topic proportions $\psi_{e,e'}$. The switching variable $c_{e,e',n}$ selects which of these three assignments is used for each word. This selector $c_{e,e',n}$ is drawn from $\pi_{e,e'}$, which describes the tendency of words associated with this pair of entities to be ascribed to either of the entities or the pair.

It is $\psi_{e,e'}$ that describes what the relationship between entities e and e' is. By allowing some of each pair's context words to come from a relationship topic distribution, the model is able to characterize each pair's interaction in terms of the latent relationship topics.

3. COMPUTATION WITH NUBBI

With the model formally defined in terms of hidden and observed random variables, we now turn to deriving the algorithms needed to analyze data. Data analysis involves inferring the hidden structure from observed data and making predictions on future data. In this section, we develop a variational inference procedure for approximating the posterior. We then use this procedure to develop a variational expectation-maximization (EM) algorithm for parameter estimation and for approximating the various predictive distributions of interest.

3.1 Inference

In posterior inference, we approximate the posterior distribution of the latent variables conditioned on the observations. As for LDA, exact posterior inference for Nubbi is intractable [5]. We appeal to variational methods.

Variational methods posit a family of distributions over the latent variables indexed by free variational parameters. Those parameters are then fit to be close to the true posterior, where closeness is measured by relative entropy. See [13]

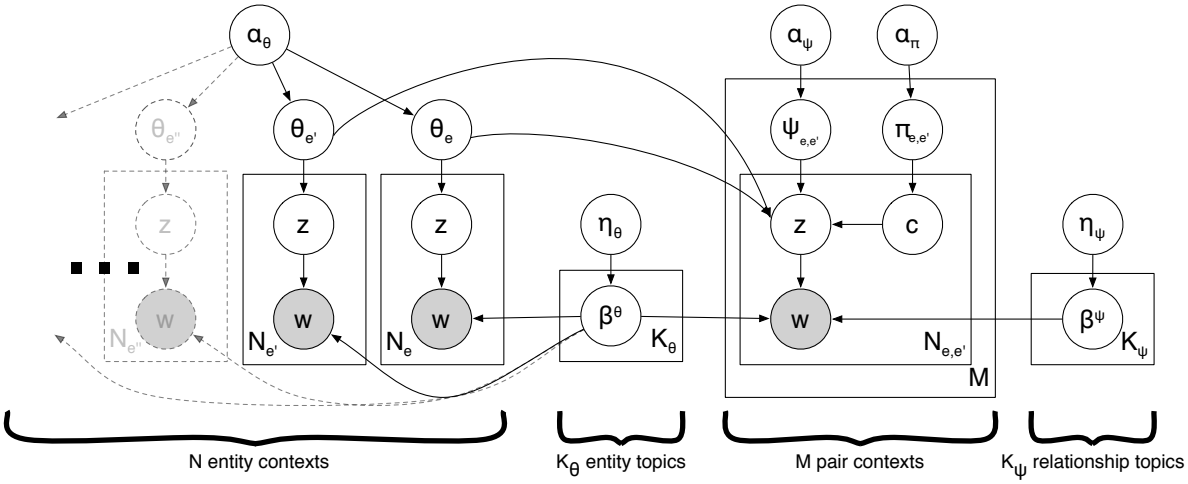

Figure 3: A depiction of the Nubbi model using the graphical model formalism. Nodes are random variables; edges denote dependence; plates (i.e., rectangles) denote replication; shaded nodes are observed and unshaded nodes are hidden. The left half of the figure are entity contexts, while the right half of the figure are pair contexts. In its entirety, the model generates both the entity contexts and the pair contexts shown in Figure 2.

for a review. We use the factorized family

$$q(\boldsymbol{\Theta}, \mathbf{Z}, \mathbf{C}, \boldsymbol{\Pi}, \boldsymbol{\Psi} | \boldsymbol{\gamma}^\theta, \boldsymbol{\gamma}^\psi, \boldsymbol{\Phi}^\theta, \boldsymbol{\Phi}^\psi, \boldsymbol{\gamma}^\pi, \boldsymbol{\Xi}) = \prod_e \left[q(\theta_e | \gamma_e^\theta) \prod_n q(z_{e,n} | \phi_{e,n}^\theta) \right] \cdot \prod_{e,e'} q(\psi_{e,e'} | \gamma_{e,e'}^\psi) q(\pi_{e,e'} | \gamma_{e,e'}^\pi) \cdot \prod_{e,e'} \left[\prod_n q(z_{e,e',n}, c_{e,e',n} | \phi_{e,e',n}^\psi, \xi_{e,e',n}) \right],$$

where $\boldsymbol{\gamma}^\theta$ is a set of Dirichlet parameters, one for each entity; $\boldsymbol{\gamma}^\pi$ and $\boldsymbol{\gamma}^\psi$ are sets of Dirichlet parameters, one for each pair of entities; $\boldsymbol{\Phi}^\theta$ is a set of multinomial parameters, one for each word in each entity; $\boldsymbol{\Xi}$ is a set of multinomial parameters, one for each pair of entities; and $\boldsymbol{\Phi}^\psi$ is a set of matrices, one for each word in each entity pair. Each $\phi_{e,e',n}^\psi$ contains three rows — one which defines a multinomial over topics given that the word comes from θ_e, one which defines a multinomial given that the word comes from $\theta_{e'}$, and one which defines a multinomial given that the word comes from $\psi_{e,e'}$. Note that the variational family we use is *not* the fully-factorized family; this family fully captures the joint distribution of $z_{e,e',n}$ and $c_{e,e',n}$. We parameterize this pair by $\phi_{e,e',n}^\psi$ and $\xi_{e,e',n}$ which define a multinomial distribution over all $3K$ possible values of this pair of variables.

Minimizing the relative entropy is equivalent to maximizing the Jensen's lower bound on the marginal probability of the observations, i.e., the evidence lower bound (ELBO),

$$\mathcal{L} = \sum_{e,e'} \mathcal{L}_{e,e'} + \sum_e \mathcal{L}_e + \mathrm{H}(q), \tag{1}$$

where sums over e, e' iterate over all pairs of entities and

$$\mathcal{L}_{e,e'} = \sum_n \mathbb{E}_q \left[\log p(w_{e,e',n} | \beta_{1:K}^\psi, \beta_{1:K}^\theta, z_{e,e',n}, c_{e,e',n}) \right] + \sum_n \mathbb{E}_q \left[\log p(z_{e,e',n} | c_{e,e',n}, \theta_e, \theta_{e'}, \psi_{e,e'}) \right] + \sum_n \mathbb{E}_q \left[\log p(c_{e,e',n} | \pi_{e,e'}) \right] + \mathbb{E}_q \left[\log p(\psi_{e,e'} | \alpha_\psi) \right] + \mathbb{E}_q \left[\log p(\pi_{e,e'} | \alpha_\pi) \right]$$

and

$$\mathcal{L}_e = \sum_n \mathbb{E}_q \left[\log p(w_{e,n} | \beta_{1:K}^\theta, z_{e,n}) \right] + \mathbb{E}_q \left[\log p(\theta_e | \alpha_\theta) \right] + \sum_n \mathbb{E}_q \left[\log p(z_{e,n} | \theta_e) \right].$$

The $\mathcal{L}_{e,e'}$ term of the ELBO differentiates this model from previous models [5]. The connections between entities affect the objective in posterior inference (and, below, in parameter estimation).

Our aim now is to compute each term of the objective function given in Equation 1. After expanding this expression in terms of the variational parameters, we can derive a set of coordinate ascent updates to optimize the ELBO with respect to the variational parameters, $\boldsymbol{\gamma}^\theta, \boldsymbol{\gamma}^\psi, \boldsymbol{\Phi}^\theta, \boldsymbol{\Phi}^\psi, \boldsymbol{\gamma}^\pi, \boldsymbol{\Xi}$. Because of space limitations, we must refer the reader to the longer version of this paper for a full derivation of the following updates.

The updates for $\phi_{e,n}^\theta$ assign topic proportions to each word associated with an individual entity,

$$\phi_{e,n}^\theta \propto \exp\left(\log \beta_{w_n}^\theta + \Psi\left(\gamma_e^\theta \right) \right),$$

where $\log \beta_{w_n}^\theta$ represents the logarithm of column w_n of β^θ and $\Psi(\cdot)$ is the digamma function. (A digamma of a vector is the vector of digammas.) The topic assignments for each word associated with a pair of entities are similar,

$$\phi_{e,e',n,1}^\psi = \exp\left(\log \beta_{w_n}^\theta + \Psi\left(\gamma_e^\theta \right) - \Psi\left(\mathbf{1}^\mathsf{T} \gamma_e^\theta \right) - \lambda_{e,e',n,1} \right)$$

$$\phi_{e,e',n,2}^\psi = \exp\left(\log \beta_{w_n}^\theta + \Psi\left(\gamma_{e'}^\theta \right) - \Psi\left(\mathbf{1}^\mathsf{T} \gamma_{e'}^\theta \right) - \lambda_{e,e',n,2} \right)$$

$$\phi_{e,e',n,3}^\psi = \exp\left(\log \beta_{w_n}^\psi + \Psi\left(\gamma_{e,e'}^\psi \right) - \Psi\left(\mathbf{1}^\mathsf{T} \gamma_{e,e'}^\psi \right) - \lambda_{e,e',n,3} \right),$$

where $\lambda_{e,e',n}$ is a vector of normalizing constants. These normalizing constants are then used to estimate the probability that each word associated with a pair of entities is assigned to either an individual or relationship,

$$\xi_{e,e',n} \propto \exp\left(\lambda_{e,e',n} + \Psi\left(\gamma_{e,e'}^\pi \right) \right).$$

The topic and entity assignments are then used to estimate the variational Dirichlet parameters which parameterize the

latent topic and entity proportions,

$$\gamma^{\pi}_{e,e'} = \alpha_\pi + \sum_n \xi_{e,e',n}$$
$$\gamma^{\psi}_{e,e'} = \alpha_\psi + \sum_n \xi_{e,e',n,3}\phi_{e,e',n,3}.$$

Finally, the topic and entity assignments for each pair of entities along with the topic assignments for each individual entity are used to update the variational Dirichlet parameters which govern the latent topic assignments for each individual entity. These updates allow us to combine evidence associated with individual entities and evidence associated with entity pairs.

$$\gamma^{\theta}_{e} = \sum_{e'}\sum_n \left(\xi_{e,e',n,1}\phi^{\psi}_{e,e',n,1} + \xi_{e',e,2}\phi^{\psi}_{e',e,n,2}\right) + \alpha_\theta + \sum_n \phi^{\theta}_{e,n}.$$

3.2 Parameter estimation

We fit the model by finding maximum likelihood estimates for each of the parameters: $\pi_{e,e'}$, $\beta^{\theta}_{1:K}$ and $\beta^{\psi}_{1:K}$. Once again, this is intractable so we turn to an approximation. We employ variational expectation-maximization, where we iterate between optimizing the ELBO of Equation 1 with respect to the variational distribution and with respect to the model parameters.

Optimizing with respect to the variational distribution is described in Section 3.1. Optimizing with respect to the model parameters is equivalent to maximum likelihood estimation with expected sufficient statistics, where the expectation is taken with respect to the variational distribution. The sufficient statistics for the topic vectors β^{θ} and β^{ψ} consist of all topic-word pairs in the corpus, along with their entity or relationship assignments. Collecting these statistics leads to the following updates,

$$\beta^{\theta}_{w} \propto \eta_\theta + \sum_e\sum_n \mathbb{1}(w_{e,n}=w)\phi^{\theta}_{e,n} + \sum_{e,e'}\sum_n \mathbb{1}(w_{e,e',n}=w)\xi_{e,e',n,1}\phi^{\psi}_{e,e',n,1} + \sum_{e,e'}\sum_n \mathbb{1}(w_{e',e,n}=w)\xi_{e',e,n,2}\phi^{\psi}_{e',e,n,2}$$
$$\beta^{\psi}_{w} \propto \eta_\psi + \sum_{e,e'}\sum_n \mathbb{1}(w_{e,e',n}=w)\xi_{e,e',n,3}\phi^{\psi}_{e,e',n,3}.$$

The sufficient statistics for $\pi_{e,e'}$ are the number of words ascribed to the first entity, the second entity, and the relationship topic. This results in the update

$$\pi_{e,e'} \propto \exp\left(\Psi\left(\alpha_\pi + \textstyle\sum_n \xi_{e,e',n}\right)\right).$$

3.3 Prediction

With a fitted model, we can make judgments about how well the model describes the joint distribution of words associated with previously unseen data. In this section we describe two prediction tasks that we use to compare Nubbi to other models: word prediction and entity prediction.

In word prediction, the model predicts an unseen word associated with an entity pair given the other words associated with that pair, $p(w_{e,e',i}|\boldsymbol{w_{e,e',-i}})$. This quantity cannot be computed tractably. We instead turn to a variational approximation of this posterior,

$$p(w_{e,e',i}|\boldsymbol{w_{e,e',-i}}) \approx \mathbb{E}_q\left[p(w_{e,e',i}|z_{e,e',i})\right].$$

Here we have replaced the expectation over the true posterior probability $p(z_{e,e',i}|\boldsymbol{w_{e,e',-i}})$ with the variational distribution $q(z_{e,e',i})$ whose parameters are trained by maximizing the evidence bound given $\boldsymbol{w_{e,e',-i}}$.

In entity prediction, the model must predict which entity pair a set of words is most likely to appear in. By Bayes' rule, the posterior probability of an entity pair given a set of words is proportional to the probability of the set of words belonging to that entity pair,

$$p((e,e')|\boldsymbol{w}) \propto p(\boldsymbol{w}|\boldsymbol{w_{e,e'}}),$$

where the proportionality constant is chosen such that the sum of this probability over all entity pairs is equal to one.

After a qualitative examination of the topics learned from corpora, we use these two prediction methods to compare Nubbi against other models that offer probabilistic frameworks for associating entities with text in Section 4.2.

4. EXPERIMENTS

In this section, we describe a qualitative and quantitative study of Nubbi on three data sets: the *bible* (characters in the bible), *biological* (genes, diseases, and proteins in scientific abstracts), and *wikipedia*. For these three corpora, the entities of interest are already annotated. Experts have marked all mentions of people in the Bible [23] and biological entities in corpora of scientific abstracts [26, 30], and Wikipedia's link structure offers disambiguated mentions. Note that it is also possible to use named entity recognizers to preprocess data for which entities are not previously identified.

The first step in our analysis is to determine the entity and pair contexts. For *bible*, verses offer an atomic context; any term in a verse with an entity (pair) is associated with that entity (pair). For *biological*, we use tokens within a fixed distance from mentions of an entity (pair) to build the data used by our model. For *wikipedia*, we used the same approach as *biological* for associating words with entity pairs. We associated with individual entities, however, all the terms in his/her Wikipedia entry. For all corpora we removed tokens based on a stop list and stemmed all tokens using the Porter stemmer. Infrequent tokens, entities, and pairs were pruned from the corpora.[1]

4.1 Learning Networks

We first demonstrate that the Nubbi model produces interpretable entity topics that describe entity contexts and relationship topics that describe pair contexts. We also show that by combining Nubbi's model of language with a network automatically estimated through co-occurrence counts, we can construct rich social networks with labeled relationships.

Table 1 shows some of the relationship topics learned from the Bible data. (This model has five entity topics and five

[1]After preprocessing, the *bible* dataset contains a lexicon of size 2411, 523 entities, and 475 entity pairs. The *biological* dataset contains a lexicon of size 2425, 1566 entities, and 577 entity pairs. The *wikipedia* dataset contains a lexicon of size 9144, 1918 entities, and 429 entity pairs.

	Topic 1	Topic 2
Entities	Jesus, Mary Terah, Abraham	Abraham, Chedorlaomer Ahaz, Rezin
Top Terms	father begat james daughter mother	king city smote lord thousand

Table 1: Examples of relationship topics learned by a six topic Nubbi model trained on the Bible. The upper part of the table shows some of the entity pairs highly associated with that topic. The lower part of the table shows the top terms in that topic's multinomial.

relationship topics.) Each column shows the words with the highest weight in that topic's multinomial parameter vector, and above each column are examples of entity pairs associated with that topic. In this example, Relationship Topic 1 corresponds to blood relations, and Relationship Topic 2 refers to antagonists. We emphasize that this structure is uncovered by analyzing the original texts. No prior knowledge of the relationships between characters is used in the analysis.

In a more diverse corpus, Nubbi learns broader topics. In a twenty-five topic model trained on the Wikipedia data, the entity topics broadly apply to entities across many time periods and cultures. Artists, monarchs, world politicians, people from American history, and scientists each have a representative topic (see Table 2).

The relationship topics further restrict entities that are specific to an individual country or period (Table 3). In some cases, relationship topics narrow the focus of broader entity topics. For instance, Relationship Topics 1, 5, 6, 9, and 10 in Table 3 help explain the specific historical context of pairs better than the very broad world leader entity Topic 7.

In some cases, these distinctions are very specific. For example, Relationship Topic 6 contains pairs of post-Hanoverian monarchs of Great Britain and Northern Ireland, while Relationship Topic 5 contains relationships with pre-Hanoverian monarchs of England even though both share words like "queen" and "throne." Note also that these topics favor words like "father" and "daughter," which describe the relationships present in these pairs.

The model sometimes groups together pairs of people from radically different contexts. For example, Relationship Topic 8 groups composers with religious scholars (both share terms like "mass" and "patron"), revealing a drawback of using a unigram-based method. As another example, Relationship Topic 3 civil war generals and early Muslim leaders.

4.2 Evaluating the predictive distribution

The qualitative results of the previous section illustrate that Nubbi is an effective model for exploring and understanding latent structure in data. In this section, we provide a quantitative evaluation of the predictive mechanisms that Nubbi provides.

As with any probabilistic model, Nubbi defines a probability distribution over unseen data. After fitting the latent variables of our model to data (as described in Section 3.1), we take unseen pair contexts and ask how well the model predicts those held-out words. Models that give higher probability to the held-out words better capture how the two entities participating in that context interact. In a complimentary problem, we can ask the fitted model to predict entities given the words in the pair context. (The details of these metrics are defined more precisely in Section 3.3.)

We compare Nubbi to three alternative approaches: a unigram model, LDA [5], and the Author-Topic model [27]. All of these approaches are models of language which treat individual entities and pairs of entities alike as bags of words. In the Author-Topic model [27], entities are associated with individual contexts and pair contexts, but there are no distinguished pair topics; all words are explained by the topics associated with individuals. In addition, we also compare the model against two baselines: a unigram model (equivalent to using no relationship topics and one entity topic) and a mutual information model (equivalent to using one relationship topic and one entity topic).

We use the bootstrap method to create held-out data sets and compute predictive probability [10]. Figure 4 shows the average predictive log likelihood for the three approaches. The results for Nubbi are plotted as a function of the total number of topics $K = K_\theta + K_\psi$. The results for LDA and author-topic were also computed with K topics. All models were trained with the same hyperparameters.

Nubbi outperforms both LDA and unigram on all corpora for all numbers of topics K. For word prediction Nubbi performs comparably to Author-Topic on *bible*, worse on *biological*, and better on *wikipedia*. We posit that because the *wikipedia* corpus contains more tokens per entity and pair of entities, the Nubbi model is able to leverage more data to make better word predictions. Conversely, for *biological*, individual entities explain pair contexts better than relationship topics, giving the advantage to Author-Topic. For *wikipedia*, this yields a 19% improvement in average word log likelihood over the unigram model at $K = 24$.

In contrast, the LDA model is unable to make improved predictions over the unigram model. There are two reasons for this. First, LDA cannot use information about the participating entities to make predictions about the pair, because it treats entity contexts and pair contexts as independent bags of words. Second, LDA does not allocate topics to describe relationships alone, whereas Nubbi does learn topics which express relationships. This allows Nubbi to make more accurate predictions about the words used to describe relationships. When relationship words do find their way into LDA topics, LDA's performance improves, such as on the *bible* dataset. Here, LDA is able to obtain a 6% improvement over unigram; Nubbi obtains a 10% improvement.

With the exception of Author-Topic on *biological*, Nubbi outperforms the other all the other approaches on the entity prediction task. For example, on *wikipedia*, the Nubbi model shows a 32% improvement over the unigram baseline, LDA shows a 7% improvement, and Author-Topic actually performs worse than the unigram baseline. While LDA, Author-Topic, and Nubbi improve monotonically with the number of topics on the word task, they can peak and decrease for the entity prediction task. Recall that an improved word likelihood need not imply an improved entity likelihood; if a model assigns a higher word likelihood to other entity pairs in addition to the correct entity pair, the predictive entity likelihood may still decrease. Thus, while each held-out context is associated with a particular pair of entities, it does not follow that that same context could not

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Entities	George Westinghouse George Stephenson Guglielmo Marconi James Watt Robert Fulton	Charles Peirce Francis Crick Edmund Husserl Ibn al-Haytham Linus Pauling	Lindsay Davenport Martina Hingis Michael Schumacher Andre Agassi Alain Prost	Lee Harvey Oswald Timothy McVeigh Yuri Gagarin Bobby Seale Patty Hearse	Pierre-Joseph Proudhon Benjamin Tucker Murray Rothbard Karl Marx Amartya Sen
Top Terms	electricity engine patent company invent	work universe theory science time	align bgcolor race win grand	state american year time president	social work politics society economics
	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Entities	Betty Davis Humphrey Bogart Kate Winslet Martin Scorsese Audrey Hepburn	Franklin D. Roosevelt Jimmy Carter Brian Mulroney Neville Chamberlain Margaret Thatcher	Jack Kirby Terry Pratchett Carl Barks Gregory Benford Steve Ditko	Babe Ruth Barry Bonds Satchel Page Pedro Martinez Roger Clemens	Xenophon Caligula Horus Nebuchadrezzar II Nero
Top Terms	film award star role play	state party election president government	story book work fiction publish	game baseball season league run	greek rome history senate death

Table 2: Ten topics from a model trained on Wikipedia carve out fairly broad categories like monarchs, athletes, entertainers, and figures from myth and religion. An exception is the more focused Topic 9, which is mostly about baseball. Note that not all of the information is linguistic; Topic 3 shows we were unsuccessful in filtering out all Wikipedia's markup, and the algorithm learned to associate score tables with a sports category.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Pairs	Reagan-Gorbachev Kennedy-Khrushchev Alexandra-Alexander III Najibullah-Kamal Nicholas I-Alexander III	Muhammad-Moses Rabin-Arafat E. Brontë-C. Brontë Solomon-Moses Arafat-Sharon	Grant-Lee Muhammad-Abu Bakr Sherman-Grant Jackson-Lee Sherman-Lee	Paul VI-John Paul II Piux XII-Paul II John XXIII-John Paul II Pius IX-John Paul II Leo XIII - John Paul II	Philip V-Louis XIV Louis XVI-Francis I Maria Theresa-Charlemagne Philip V-Louis XVI Philip V-Maria Theresa
Terms	soviet russian government union nuclear	israel god palestinian chile book	union corp gen campaign richmond	vatican cathol papal council time	french dauphin spanish death throne
	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Pairs	Henry VIII-C. of Aragon Mary I (Eng)-Elizabeth I Henry VIII-Anne Boleyn Mary I (Scot)-Elizabeth I Henry VIII-Elizabeth I	Jefferson-Burr Jefferson-Madison Perot-Bush Jefferson-Jay J.Q. Adams-Clay	Mozart-Salieri Malory-Arthur Mozart-Beethoven Bede-Augustine Leo X-Julius II	George VI-Edward VII George VI-Edward VIII Victoria-Edward VII George V-Edward VII Victoria-George VI	Trotsky-Stalin Kamenev-Stalin Khrushchev-Stalin Kamenev-Trotsky Zhou Enlai-Mao Zedong
Terms	queen english daughter death throne	republican state federalist vote vice	music play film piano work	royal queen british throne father	soviet communist central union full

Table 3: In contrast to Table 2, the relationship topics shown here are more specific to time and place. For example, English monarch pairs (Topic 6) are distinct from British monarch pairs (Topic 9). While there is some noise (the Brontë sisters being lumped in with mideast leaders or Abu Bakr and Muhammad with civil war generals), these relationship topics group similar pairs of entities well. A social network labeled with these relationships is shown in Figure 1.

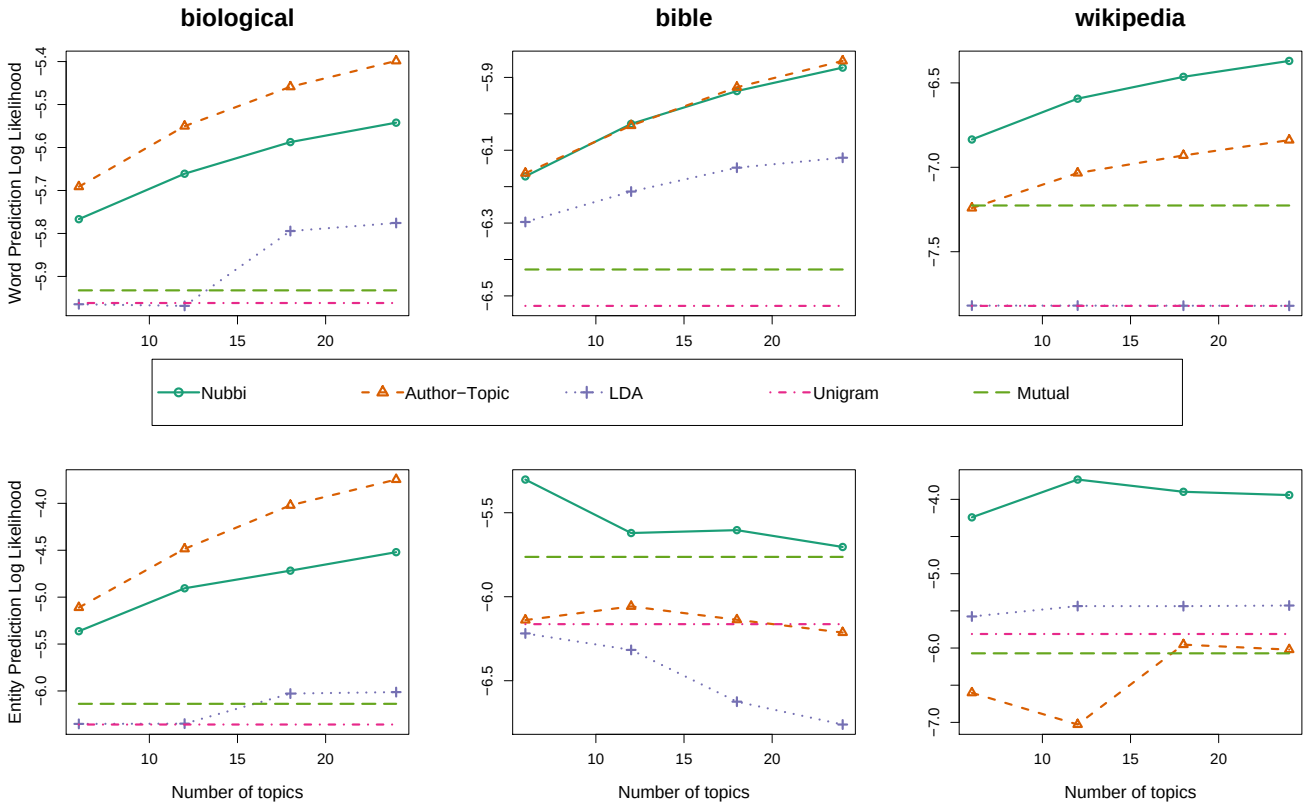

Figure 4: Predictive log likelihood as a function of the number of Nubbi topics on two tasks: entity prediction (given the context, predict what entities are being discussed) and relation prediction (given the entities, predict what words occur). Higher is better.

also be aptly associated with some other entity pair.

5. DISCUSSION AND RELATED WORK

We presented Nubbi, a novel machine learning approach for analyzing free text to extract descriptions of relationships between entities. We applied Nubbi to three corpora—the Bible, Wikipedia, and scientific abstracts. We showed that Nubbi provides a state-of-the-art predictive model of entities and relationships and, moreover, is a useful exploratory tool for discovering and understanding network data hidden in plain text.

Analyzing networks of entities has a substantial history [33]; recent work has focused in particular on clustering and community structure [2, 6, 11, 18, 25], deriving models for social networks [15, 16, 19, 31], and applying these analyses to predictive applications [34]. Latent variable approaches to modeling social networks with associated text have also been explored [17, 20, 22, 32]. While the space of potential applications for these models is rich, it is tempered by the need for observed network data as input. Nubbi allows these techniques to augment their network data by leveraging the large body of relationship information encoded in collections of free text.

Previous work in this vein has used either pattern-based approaches or co-occurrence methods. The pattern-based approaches [1, 9, 21, 28] and syntax based approaches [3, 14] require patterns or parsers which are meticulously hand-crafted, often fragile, and typically need several examples of desired relationships limiting the type of relationships that can be discovered. In contrast, Nubbi makes minimal assumptions about the input text, and is thus practical for languages and non-linguistic data where parsing is not available or applicable. Co-occurrence methods [7, 8] also make minimal assumptions. However, because Nubbi draws on topic modeling [5], it is able to uncover hidden and semantically meaningful *groupings* of relationships. Through the distinction between relationship topics and entity topics, it can better model the language used to describe relationships.

Finally, while other models have also leveraged the machinery of LDA to understand ensembles of entities and the words associated with them [4, 24, 27] these models only learn hidden topics for individual entities. Nubbi models individual entities and pairs of entities distinctly. By controlling for features of individual entities and explicitly relationships, Nubbi yields more powerful predictive models and can discover richer descriptions of relationships.

6. ACKNOWLEDGEMENTS

We would like to thank David Petrou, Bill Schillit, Casey Whitelaw, and Ryan MacDonald for their advice and support during the development of this work. We also thank the Office of Naval Research and Google for supporting this work.

7. REFERENCES

[1] E. Agichtein and L. Gravano. Querying text databases for efficient information extraction. *Data Engineering, International Conference on*, 0:113, 2003.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. *KDD 2008*, 2008.

[3] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI 2007*, 2007.

[4] I. Bhattacharya, S. Godbole, and S. Joshi. Structured entity identification and document categorization: Two tasks with one joint model. *KDD 2008*, 2008.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Mining hidden community in heterogeneous social networks. *LinkKDD 2005*, Aug 2005.

[7] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. *AAAI 2005*, 2005.

[8] D. Davidov, A. Rappoport, and M. Koppel. Fully unsupervised discovery of concept-specific relationships by web mining. In *ACL*, 2007.

[9] C. Diehl, G. M. Namata, and L. Getoor. Relationship identification for social network discovery. In *AAAI 2007*, July 2007.

[10] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382), 1983.

[11] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. *HYPERTEXT 1998*, May 1998.

[12] T. Hofmann. Probabilistic latent semantic indexing. *SIGIR 1999*, 1999.

[13] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Oct 1999.

[14] S. Katrenko and P. Adriaans. Learning relations from biomedical corpora using dependency trees. *Lecture Notes in Computer Science*, 2007.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. *KDD 2008*, 2008.

[16] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. *WWW 2008*, 2008.

[17] A. McCallum, A. Corrada-Emmanuel, and X. Wang. Topic and role discovery in social networks. *IJCAI 2005*, 2005.

[18] A. McGovern, L. Friedland, M. Hay, B. Gallagher, A. Fast, J. Neville, and D. Jensen. Exploiting relational structure to understand publication patterns in high-energy physics. *ACM SIGKDD Explorations Newsletter*, 5(2), Dec 2003.

[19] E. Meeds, Z. Ghahramani, R. Neal, and S. Roweis. Modeling dyadic data with binary latent factors. *NIPS 2007*, 2007.

[20] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. *WWW 2008*, Apr 2008.

[21] Q. Mei, D. Xin, H. Cheng, J. Han, and C. Zhai. Semantic annotation of frequent patterns. *KDD 2007*, 1(3), 2007.

[22] R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. *KDD 2008*, 2008.

[23] O. J. Nave. *Nave's Topical Bible*. Thomas Nelson, 2003.

[24] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD 2006*, pages 680–686, New York, NY, USA, 2006. ACM.

[25] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 2006.

[26] T. Ohta, Y. Tateisi, and J.-D. Kim. Genia corpus: an annotated research abstract corpus in molecular biology domain. In *HLT 2008*, San Diego, USA, 2002.

[27] M. Rosen-Zvi, T. Griffiths, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *AUAI 2004*, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[28] S. Sahay, S. Mukherjea, E. Agichtein, E. Garcia, S. Navathe, and A. Ram. Discovering semantic biomedical relations utilizing the web. *KDD 2008*, 2(1), Mar 2008.

[29] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 2007.

[30] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1, 2005.

[31] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. *NIPS 2003*, 2003.

[32] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and text. *Proceedings of the 3rd international workshop on Link discovery*, 2005.

[33] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 1996.

[34] D. Zhou, S. Zhu, K. Yu, X. Song, B. Tseng, H. Zha, and C. Giles. Learning multiple graphs for document recommendations. *WWW 2008*, Apr 2008.