13

Block-LDA: Jointly Modeling Entity-Annotated Text and Entity-Entity Links

Ramnath Balasubramanyan

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

William W. Cohen

Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

CONTENTS

Identifying latent groups of entities from observed interactions between pairs of entities is a frequently encountered problem in areas like analysis of protein interactions and social networks. We present a model that combines aspects of mixed membership stochastic blockmodels and topic models to improve entity-entity link modeling by jointly modeling links and text about the entities that are linked. We apply the model to two datasets: a protein-protein interaction (PPI) dataset supplemented with a corpus of abstracts of scientific publications annotated with the proteins in the PPI dataset and an Enron email corpus. The induced topics' ability to help understand the nature of the data provides a qualitative evaluation of the model. Quantitative evaluation shows improvements in functional category prediction of proteins and in perplexity, using the joint model over baselines that use only link or text information. For the PPI dataset, the topic coherence of the emergent topics and the ability of the model to retrieve relevant scientific articles and proteins related to the topic are compared to that of a text-only approach that does not make use of the protein-protein interaction matrix. Evaluation of the results by biologists show that the joint modeling results in better topic coherence and improves retrieval performance in the task of identifying top related papers and proteins.

13.1 Introduction

The task of modeling latent groups of entities from observed interactions is a commonly encountered problem. In social networks, for instance, we might want to identify sub-communities. In the biological domain we might want to discover latent groups of proteins based on observed pairwise interactions. Mixed membership stochastic blockmodels (MMSB) (Airoldi et al., 2008; Parkkinen et al., 2009) approach this problem by assuming that nodes in a graph represent entities belonging to latent blocks with mixed membership, effectively capturing the notion that entities may arise from different sources and have different roles.

In another area of active research, models like latent Dirichlet allocation (LDA) (Blei et al., 2003) model text documents in a corpus as arising from mixtures of latent topics. In such models, words in a document are potentially generated from different topics using topic-specific word distributions. Extensions to LDA (Erosheva et al., 2004; Griffiths and Steyvers, 2004) additionally model other metadata in documents such as authors and entities by treating a latent topic as a set of distributions, one for each metadata type. For instance, when modeling scientific publications from the biological domain, a latent topic could have a word distribution, an author distribution, and a protein entity distribution. We refer to this model as *Link LDA* following the convention established by Nallapati et al. (2008). The different types of data that are contained in a document (e.g., words in the body, words in the title, authors, list of citations, etc.) are referred to as *entity types*.

In this chapter, we present a model, *Block-LDA*, that jointly generates text documents annotated with metadata about associated entities and external links between pairs of entities. This allows the model to use supplementary annotated text to influence and improve link modeling. The text documents are modeled as bags of entities of different types and the network is modeled as edges between entities of a source type to a destination type. Consider the example of a corpus of publications about the yeast organism and a network of protein-protein interactions in yeast. These publications are further annotated by experts with lists of proteins that are discussed in them. Therefore, each publication could be modeled as a collection of bags *vis a vis* bag of body-words, bag of authors, bag of proteins discussed in the paper, etc. Similarly, the network could be a collection of protein-protein interactions independently observed. The model merges the idea of latent topics in topic models with blocks in stochastic blockmodels. The joint modeling permits sharing of information about the latent topics between the network structure and text, resulting in more coherent topics. Co-occurrence patterns in entities and words related to them aid the modeling of links in the graph. Likewise, entity-entity links provide clues about topics in the text. We also propose a method to perform approximate inference in the model using a collapsed Gibbs sampler, since exact inference in the joint model is intractable.

We then use the model to organize a large collection of literature about yeast biology to enable topic-oriented browsing and retrieval from the literature. The analysis is performed using the mixed membership topic modeling to uncover latent structure in document corpora by identifying broad topics that are discussed in it. This approach complements traditional information retrieval tasks where the objective is to fulfill very specific information needs. By using joing modeling, we are able to use other sources of domain information related to the domain in addition to literature. In the case of yeast biology, an example of such a resource is a database of known protein-protein interactions (PPI) which have been identified using wetlab experiments. We perform data fusion by combining text information from articles and the database of yeast protein-protein interactions by using a latent variable model—Block-LDA (Balasubramanyan and Cohen, 2011), that jointly models the literature and PPI networks.

We evaluate the ability of the topic models to return meaningful topics by inspecting the top papers and proteins that pertain to them. We compare the performance of the joint model, i.e., Block-LDA, with a model that only considers the text corpora by asking a yeast biologist to

evaluate the coherence of topics and the relevance of the retrieved articles and proteins. This evaluation serves to test the utility of Block-LDA on a real task as opposed to an internal evaluation (such as by using perplexity metrics). Our evaluaton shows that the joint model outperforms the text-only approach both in topic coherence and in top paper and protein retrieval as measured by precision@10 values.

The chapter is organized as follows: Section 15.2 introduces the model and presents a Gibbs sampling-based method for performing approximate inference with the model. Section 13.3 discusses related work, and Section 13.4 provides details of datasets used in the experiments. Sections 13.5.1 and 13.5.2 present the results of our experiments on two datasets from different domains. Finally, our conclusions are in Section 13.6.

13.2 Block-LDA

Variables in the model

K	-	the number of topics (therefore resulting in K^2 blocks in the network)
α_L	-	Dirichlet prior for the topic pair distribution for links
α_D	-	Dirichlet prior for document specific topic distributions
γ	-	Dirichlet prior for topic multinomials
π_L	-	multinomial distribution over topic pairs for links
θ_d	-	multinomial distribution over topics for document d
T	-	the number of types of entities in the corpus
$\beta_{t,z}$	-	multinomial over entities of type t for topic z
D	-	number of documents in the corpus
$z_{t,i}$	-	topic chosen for the i-th entity of type t in a document
$e_{t,i}$	-	the i-th entity of type t occurring in a document
N_L	-	number of links in the network
z_{i1} and z_{i2}	-	topics chosen for the two nodes participating in the i-th link
e_{i1} and e_{i2}	-	the two nodes participating in the i-th link

The Block-LDA model (plate diagram in Figure 13.1) enables sharing of information between the component on the left that models links between pairs of entities represented as edges in a graph with a block structure, and the component on the right that models text documents through shared latent topics. More specifically, the distribution over the entities of the type that are linked is shared between the blockmodel and the text model.

The component on the right, which is an extension of the LDA models, documents as sets of "bags of entities," with each bag corresponding to a particular type of entity. Every entity type has a topic-wise multinomial distribution over the set of entities that can occur as an instance of the entity type.

The component on the left is a generative model for graphs representing entity-entity links with an underlying block structure, derived from the sparse blockmodel introduced by Parkkinen et al. (2009). Linked entities are generated from topic-specific entity distributions conditioned on the topic pairs sampled for the edges. Topic pairs for edges (links) are drawn from a multinomial defined over the Cartesian product of the topic set with itself. Vertices in the graph representing entities therefore have mixed memberships in topics. In contrast to MMSB, only observed links are sampled, making this model suitable for sparse graphs.

Let K be the number of latent topics (blocks) we wish to recover. Assuming documents consist of T different types of entities (i.e., each document contains T bags of entities), and that links in the

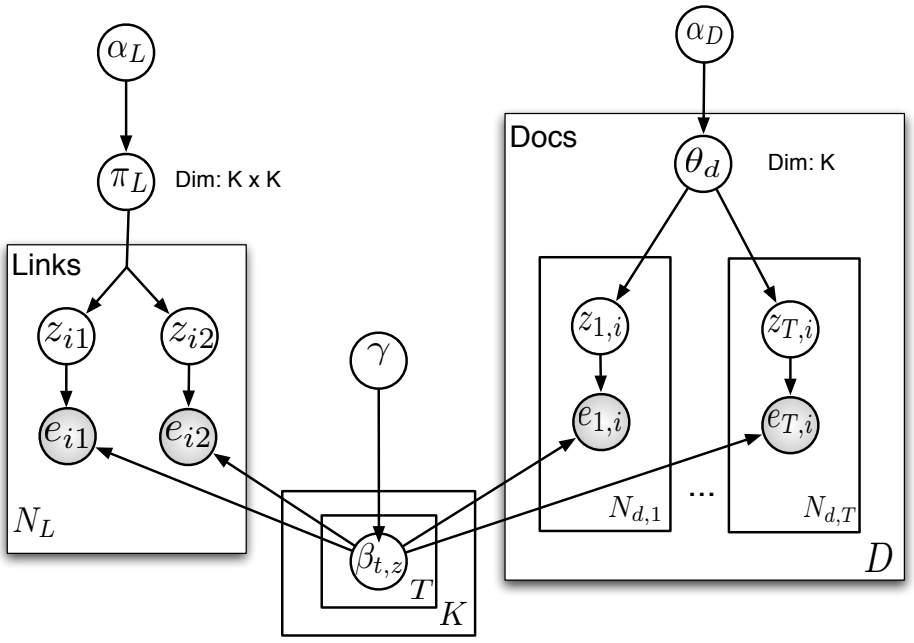


FIGURE 13.1
Block-LDA: plate diagram.

graph are between entities of type t_l and t_r, the generative process is as follows.

1. Generate topics:
   - For each type $t \in 1, \ldots, T$, and topic $z \in 1, \ldots, K$, sample $\beta_{t,z} \sim$ Dirichlet(γ), the topic specific entity distribution.
2. Generate documents. For every document $d \in \{1 \ldots D\}$:
   - Sample $\theta_d \sim$ Dirichlet(α_D) where θ_d is the topic mixing distribution for the document.
   - For each type t and its associated set of entity mentions $e_{t,i}$, $i \in \{1, \cdots, N_{d,t}\}$:
     - Sample a topic $z_{t,i} \sim$ Multinomial(θ_d).
     - Sample an entity $e_{t,i} \sim$ Multinomial$(\beta_{t,z_{t,i}})$.
3. Generate the link matrix of entities of type t_l:
   - Sample $\pi_L \sim$ Dirichlet(α_L) where π_L describes a distribution over the Cartesian product of topics for links in the dataset.
   - For every link $e_{i1} \rightarrow e_{i2}$, $i \in \{1 \cdots N_L\}$:
     - Sample a topic pair $\langle z_{i1}, z_{i2} \rangle \sim$ Multinomial(π_L).
     - Sample $e_{i1} \sim$ Multinomial$(\beta_{t_l,z_{i1}})$.
     - Sample $e_{i2} \sim$ Multinomial$(\beta_{t_r,z_{i2}})$.

Note that unlike the MMSB model introduced by Airoldi et al. (2008), this model generates only realized links between entities.

Given the hyperparameters α_D, α_L, and γ, the joint distribution over the documents, links, their topic distributions, and topic assignments is given by

$$p(\pi_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{e}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{e_1}, \mathbf{e_2} \rangle | \alpha_D, \alpha_L, \gamma) \propto \tag{13.1}$$
$$\prod_{z=1}^{K} \prod_{t=1}^{T} \text{Dir}(\beta_{t,z} | \gamma_t) \times$$
$$\prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha_D) \prod_{t=1}^{T} \prod_{i=1}^{N_{d,t}} \theta_d^{z_{t,i}^{(d)}} \beta_{t, z_{t,i}^{(d)}}^{e_{t,i}} \times$$
$$\text{Dir}(\pi_L | \alpha_L) \prod_{i=1}^{N_L} \pi_L^{\langle z_{i1}, z_{i2} \rangle} \beta_{tl,z1}^{e_{i1}} \beta_{tr,z2}^{e_{i2}}.$$

A commonly required operation when using models like Block-LDA is to perform inference on the model to query the topic distributions and the topic assignments of documents and links. Due to the intractability of exact inference in the Block-LDA model, a collapsed Gibbs sampler is used to perform approximate inference. It samples a latent topic for an entity mention of type t in the text corpus conditioned on the assignments to all other entity mentions using the following expression (after collapsing θ_D):

$$p(z_{t,i} = z | e_{t,i}, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{e}^{\neg \mathbf{i}}, \alpha_D, \gamma) \tag{13.2}$$
$$\propto (n_{dz}^{\neg i} + \alpha_D) \frac{n_{zte_{t,i}}^{\neg i} + \gamma}{\sum_{e'} n_{zte'}^{\neg i} + |E_t| \gamma}.$$

Similarly, we sample a topic pair for every link conditional on topic pair assignments to all other links after collapsing π_L using the expression:

$$p(\mathbf{z_i} = \langle z_1, z_2 \rangle | \langle e_{i1}, e_{i2} \rangle, \mathbf{z}^{\neg \mathbf{i}}, \langle \mathbf{e_1}, \mathbf{e_2} \rangle^{\neg \mathbf{i}}, \alpha_L, \gamma) \tag{13.3}$$
$$\propto \left(n_{\langle z_1, z_2 \rangle}^{L \neg i} + \alpha_L \right) \times$$
$$\frac{(n_{z_1 t_l e_{i1}}^{\neg i} + \gamma)(n_{z_2 t_r e_{i2}}^{\neg i} + \gamma)}{(\sum_e n_{z_1 t_l e}^{\neg i} + |E_{t_l}| \gamma)(\sum_e n_{z_2 t_r e}^{\neg i} + |E_{t_r}| \gamma)}.$$

E_t refers to the set of all entities of type t. The ns refer to the number of topic assignments in the data.

- n_{zte}—the number of times an entity e of type t is observed under topic z.
- n_{zd}—the number of entities (of any type) with topic z in document d.
- $n_{\langle z_1, z_2 \rangle}^{L}$—count of links assigned to topic pair $\langle z_1, z_2 \rangle$.

The topic multinomial parameters and the topic distributions of links and documents are easily recovered using their MAP estimates after inference using the counts of observations:

$$\beta_{t,z}^{(e)} = \frac{n_{zte} + \gamma}{\sum_{e'} n_{zte'} + |E_t| \gamma}, \tag{13.4}$$

$$\theta_d^{(z)} = \frac{n_{dz} + \alpha_D}{\sum_{z'} n_{dz'} + K \alpha_D}, \tag{13.5}$$

$$\pi_L^{\langle z_1, z_2 \rangle} = \frac{n_{\langle z_1, z_2 \rangle} + \alpha_L}{\sum_{z_1', z_2'} n_{\langle z_1', z_2' \rangle} + K^2 \alpha_L}. \tag{13.6}$$

A de-noised form of the entity-entity link matrix can also be recovered from the estimated

parameters of the model. Let B_t be a matrix of dimensions $K \times |E_t|$ where row k = $\beta_{t,k}$, $k \in \{1, \cdots, K\}$. Let Z be a matrix of dimensions $K \times K$ s.t $Z_{p,q} = \sum_{i=1}^{N_L} \mathbf{I}(z_{i1} = p, z_{i2} = q)$. The de-noised matrix M of the strength of association between the entities in E_{t_l} is given by $M = B_{t_l}^T Z B_{t_r}$.

13.3 Related Work

Link LDA, and many other extensions to LDA, model documents that are annotated with metadata. In a parallel area of research, various different approaches to modeling links between documents have been explored. For instance, pairwise-link-LDA (Nallapati et al., 2008) combines MMSB with LDA by modeling documents using LDA and generating links between them using MMSB. The relational topic model (Chang and Blei, 2009) generates links between documents based on their topic distributions. The copycat and citation influence models (Dietz et al., 2007) also model links between citing and cited documents by extending LDA and eliminating independence between documents. The latent topic hypertext model (LTHM) (Gruber et al., 2008) presents a generative process for documents that can be linked to each other from specific words in the citing document. These classes of models are different from the model proposed in this paper, Block-LDA, in that they model links between entities in the documents rather than links between documents.

The Nubbi model (Chang et al., 2009) tackles a related problem where entity relations are discovered from text data by relying on words that appear in the context of entities and entity pairs in the text. Block-LDA differs from Nubbi in that it models a document as bags of entities without considering the location of entity mentions in the text. The entities need not even be mentioned in the text of the document. The group-topic model (Wang et al., 2006) addresses the task of modeling events pertaining to pairs of entities with textual attributes that annotate the event. The text in this model is associated with events, which differs from the standalone documents mentioning entities considered by Block-LDA.

The author-topic model (AT) (Rosen-Zvi et al., 2004) addresses the task of modeling corpora annotated with the IDs of people who authored the documents. Every author in the corpus has a topic distribution over the latent topics, and words in the documents are drawn from topics drawn from the specific distribution of the author who is deemed to have generated the word. The author-recipient-topic model (ART) (McCallum et al., 2005) extends the idea further by building a topic distribution for every author-recipient pair. As we show in the experiments below, Block-LDA can also be used to model the relationships between authors, recipients, and words in documents by constructing an appropriate link matrix from known information about the authors and recipients of documents; however, unlike the AT and ART models which are primarily designed to model documents, Block-LDA provides a generative model for the links between authors and recipients in addition to documents. This allows Block-LDA to be used for additional inferences not possible with the AT or ART models, for instance, predicting probable author-recipient interactions. Wen and Lin (2010) describes an application of an approach that uses both content and network information to analyze enterprise data. While a joint modeling of the network and content is not used, LDA is used to study the topics in communications between people.

A summary of related models from prior work is shown in Table 13.1.

The Munich Institute for Protein Sequencing (MIPS) database (Mewes et al., 2004) includes a hand-crafted collection of protein interactions covering 8000 protein complex associations in yeast. We use a subset of this collection containing 844 proteins, for which all interactions were hand-curated (Figure 13.2(a)). The MIPS institute also provides a set of functional annotations for each protein which are organized in a tree, with 15 nodes at the first level (shown in Table 13.2). The 844

Model	Links	Documents
LDA	-	words
link LDA	-	words + entities
relational topic model	document-document	words + document IDs
pairwise-link-LDA, link-PLSA-LDA	document-document	words + cited document IDs
copycat, citation influence models	document-document	words + cited document IDs
latent topic hypertext model	document-document	words + cited document IDs
author-recipient-topic model	-	docs + authors + recipients
author-topic model	-	docs + authors
topic link LDA	document-document	words + authors
MMSB	entity-entity	-
sparse blockmodel (Parkkinen et al.)	entity-entity	-
Nubbi	entity-entity	words near entities or entity-pairs
group topic model	entity-entity	words about the entity-entity event
Block-LDA	entity-entity	words + entities

TABLE 13.1
Related work.

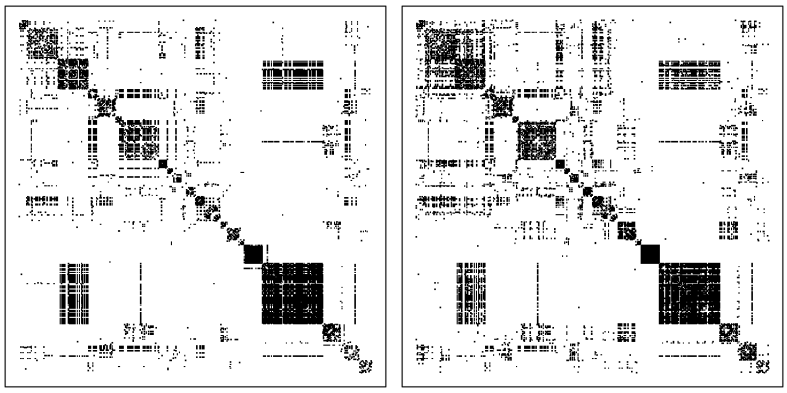
proteins participating in interactions are mapped to these 15 functional categories with an average of 2.5 annotations per protein.

We also use another dataset of protein-protein interactions in yeast that were observed as a result of wetlab experiments by collaborators. This dataset consists of 635 interactions that deal primarily with ribosomal proteins and assembly factors in yeast.

In addition to the MIPS PPI data, we use a text corpus that is derived from the repository of scientific publications at PubMed®. PubMed is a free, open-access, on-line archive of over 18 million biological abstracts, bibliographies, and citation lists for papers published since 1948 (U.S. National Library of Medicine, 2008). The subset we work with consists of approximately 40,000 publica-

Metabolism
Cellular communication/signal transduction mechanism
Cell rescue, defense and virulence
Regulation of / interaction with cellular environment
Cell fate
Energy
Control of cellular organization
Cell cycle and DNA processing
Subcellular localisation
Transcription
Protein synthesis
Protein activity regulation
Transport facilitation
Protein fate (folding, modification, destination)
Cellular transport and transport mechanisms

TABLE 13.2
List of functional categories.



(a) MIPS interactions (b) Co-occurences in text

FIGURE 13.2
Observed protein-protein interactions compared to thresholded co-occurrence in text.

tions about the yeast organism that have been curated in the Saccharomyces Genome Database (SGD) (Dwight et al., 2004) with annotations of proteins that are discussed in the publication. We further restrict the dataset to only those documents that are annotated with at least one protein from the MIPS database. This results in a MIPS-protein annotated document collection of 15,776 publications. The publications in this set were written by a total of 47,215 authors. We tokenize the titles and abstracts based on white space, lowercase all tokens, and eliminate stopwords. Low frequency ($<$5 occurrences) terms are also eliminated. The vocabulary contains 45,648 words.

13.4 Datasets

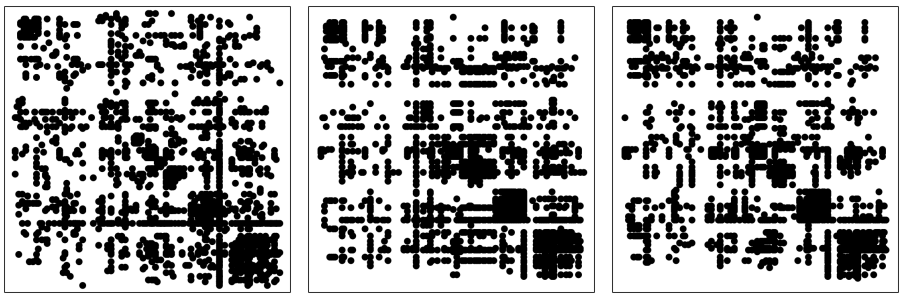
To investigate the co-occurrence patterns of proteins annotated in the abstracts, we construct a co-occurrence matrix. From every abstract, a link is constructed for every pair of annotated protein mentions. Additionally, protein mentions that occur fewer than 5 times in the corpus are discarded. Figure 13.2(b) shows that the resultant matrix looks very similar to the MIPS PPI matrix in Figure 13.2(a). This suggests that joint modeling of the protein-annotated text with the PPI information has the potential to be beneficial. The nodes representing proteins in Figures 13.2(a) and 13.2(b) are ordered by their cluster IDs, obtained by clustering them using k-means clustering, treating proteins as 15-bit vectors of functional category annotations.

The Enron email corpus (Shetty and Adibi, 2004) is a large publicly available collection of email messages subpoenaed as part of the investigation by the Federal Energy Regulatory Commission (FERC). The dataset contains 517,437 messages in total. Although the Enron Email Dataset contains the email folders of 150 people, two people appear twice with different usernames, and one user's emails consist solely of automated emails resulting in 147 unique people in the dataset. For the text component of the model, we use all the emails in the Sent[1] folders of the 147 users' mailboxes, resulting in a corpus of 96,103 emails. Messages are annotated with mentions of people

[1] "Sent", "sent_items," and "_sent_mail" folders in users' mailboxes were treated as "Sent" folders.

from the set of 147 Enron employees if they are senders or recipients of the email. Mentions of people outside of the 147 persons considered are dropped. While extracting text from the email messages, "quoted" messages are eliminated using a heuristic which looks for a "Forwarded message" or "Original message" delimiter. In addition, lines starting with a ">" are also eliminated. The emails are then tokenized after lowercasing the entire message, using whitespace and punctuation marks as word delimiters. Words occurring fewer than 5 times in the corpus are discarded. The vocabulary of the corpus consists of 32,880 words.

For the entity links component of the model, we build an email communication network by constructing a link between the sender and every recipient of an email message for every email in the corpus. Recipients of the emails include people directly addressed in the "To" field and people included in the "Cc" and "Bcc" fields. Similar to the text component, only links between the 147 Enron employees are considered. The link dataset generated in this manner has 200,404 links. Figure 13.3(a) shows the email network structure. The nodes in the matrix representing people are ordered by cluster IDs obtained by running k-means clustering on the 147 people. Each person s is represented by a vector of length 147, where the elements in the vector are normalized counts of the number of times an email is sent by s to the person indicated by the element.



(a) Observed network (b) From sparse blockmodel (c) From Block-LDA

FIGURE 13.3
Enron network and its de-noised recovered versions.

13.5 Experimental Results

We present results from experiments using Block-LDA to model the yeast and Enron datasets described in Section 13.4.

13.5.1 Results from the Yeast Dataset

Perplexity and Convergence

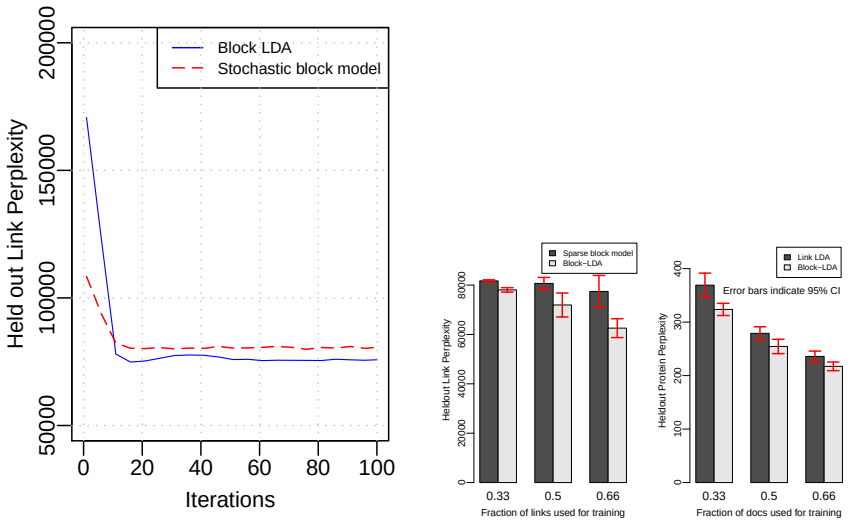
First, we investigate the convergence properties of the Gibbs sampler used for inference in Block-LDA by observing link perplexity on held-out data at different epochs. Link perplexity of a set of

links L is defined as

$$\exp\left(\frac{\sum_{e_1 \to e_2 \in L} \log\left(\sum_{\langle z_1, z_2\rangle} \pi^{\langle z_1, z_2\rangle} \beta^{(e_1)}_{t_l, z_1} \beta^{(e_1)}_{t_l, z_2}\right)}{|L|}\right). \tag{13.7}$$

Figure 13.4(a) shows the convergence of the link perplexity using Block LDA and a baseline model on the PPI+SGD dataset with 20% of the full dataset held-out for testing. The number of topics K is set at 15 since our aim is to recover topics that can be aligned with the 15 protein functional categories. α_D and α_L are sampled from Gamma$(0.1, 1)$. It can be observed that the Gibbs sampler burns-in after about 20 iterations.

Next, we perform two sets of experiments with the PPI+PubMed Central dataset. The text data has three types of entities in each document—words, authors, and protein annotations with the PPI data-linking proteins. In the first set of experiments, we evaluate the model using perplexity of held-out protein-protein interactions using increasing amounts of the PPI data for training.



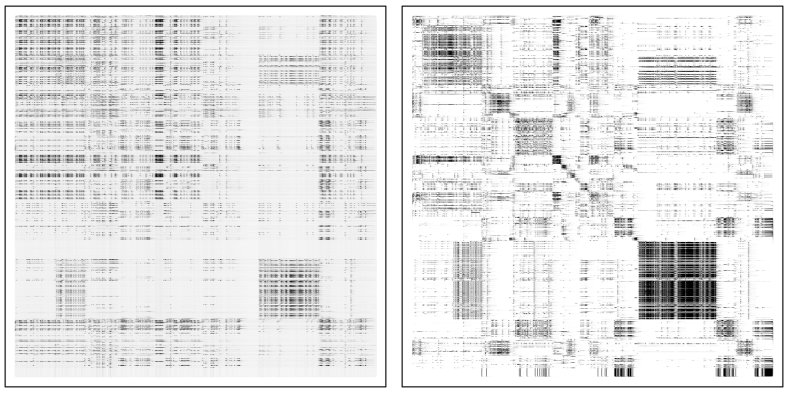
(a) Gibbs sampler convergence

(b) Gain in perplexity through joint modeling

FIGURE 13.4
Perplexity in the MIPS PPI+SGD dataset.

All 15,773 documents in the SGD dataset are used when textual information is used. When text is not used, the model is equivalent to using only the left half of Figure 13.1. Figures 13.5(a) and 13.5(b) show the posterior likelihood of protein-protein interactions recovered using the sparse blockmodel and Block-LDA, respectively. In the other set of experiments, we evaluate the model using protein perplexity in held-out text using progressively increasing amounts of text as training data. All the links in the PPI dataset are used in these experiments when link data are used. When link data are not used, the model reduces to Link LDA. In all experiments, the Gibbs sampler is run until the held-out perplexity stabilizes to a nearly constant value ($\approx$ 80 iterations)

Figure 13.4(b) shows the gains in perplexity in the two sets of experiments with different amounts of training data. The perplexity values are averaged over 10 runs. In both sets of experiments, it can be seen that Block-LDA results in lower perplexities than using links/text alone.



(a) Sparse blockmodel (b) Block-LDA

FIGURE 13.5
Inferred protein-protein interactions.

These results indicate that co-occurrence patterns of proteins in text contain information about protein interactions, which Block-LDA is able to utilize through joint modeling. Our conjecture is that the protein co-occurrence information in text is a noisy approximation of the PPI data.

Table 13.3 shows the top words, proteins, and authors for sample topics induced by running Block-LDA over the full PPI+SGD dataset. These topics provide a qualitative feel for the topics that emerge using the model. The Gibbs sampling procedure was run until convergence (around 80 iterations) and the number of topics was set to 15. The topic tables were then analyzed, with a title and an analysis of the topic added after the inference procedure was completed. Details about proteins and yeast researchers were obtained from the SGD [2] website to understand the function of the top proteins in each topic and to get an idea of the research profile of the top authors mentioned.

Topic Coherence

A useful application of latent blockmodeling approaches is understanding the underlying nature of data.

We conduct three different evaluations of the emergent topics. First, we obtain topics from only the text corpus using a model that comprises the right half of Figure 13.1, which is equivalent to using the Link-LDA model. For the second evaluation, we use the Block-LDA model that is trained on the text corpus and the MIPS protein-protein interaction database. Finally, for the third evaluation, we replace the MIPS PPI database with the interaction obtained from the wetlab experiments. In all cases, we set K, the number of topics, to be 15. In each variant, we represent documents as three sets of entities, i.e., the words in the abstracts of the article, the set of proteins associated with the article as indicated in the SGD database, and the authors who wrote the article. Each topic therefore consists of three different multinomial distributions over the sets of the three kinds of entities described.

Topics that emerge from the different variants can possibly be assigned different indices even when they discuss the same semantic concept. To compare topics across variants, we need a method to determine which topic indices from the different variants correspond to the same semantic concept. To obtain the mapping between topics from each variant, we utilize the Hungarian algorithm

[2] See http://www.yeastgenome.org.

Words	mutant, mutants, gene, cerevisiae, growth, type, mutations, saccharomyces, wild, mutation, strains, strain, phenotype, genes, deletion
Proteins	rpl20b, rpl5, rpl16a, rps5, rpl39, rpl18a, rpl27b, rps3, rpl23a, rpl1b, rpl32, rpl17b, rpl35a, rpl26b, rpl31a
Authors	klis_fm, bussey_h, miyakawa_t, toh-e_a, heitman_j, perfect_jr, ohya_y_ws, sherman_f, latge_jp, schaffrath_r, duran_a, sa-correia_i, liu_h, subik_j, kikuchi_a, chen_j, goffeau_a, tanaka_k, kuchler_k, calderone_r, nombela_c, popolo_l, jablonowski_d, kim_j
Analysis	A common experimental procedure is to induce random mutations in the "wild-type" strain of a model organism (e.g., saccharomyces cerevisiae) and then screen the mutants for interesting observable characteristics (i.e. phenotype). Often the phenotype shows slower growth rates under certain conditions (e.g. lack of some nutrient). The RPL* proteins are all part of the larger (60S) subunit of the ribosome. The first two biologists, Klis and Bussey's research use this method.

(a) Analysis of mutations

Words	binding, domain, terminal, structure, site, residues, domains, interaction, region, subunit, alpha, amino, structural, conserved, atp
Proteins	rps19b, rps24b, rps3, rps20, rps4a, rps11a, rps2, rps8a, rps10b, rps6a, rps10a, rps19a, rps12, rps9b, rps28a
Authors	naider_f, becker_jm, leulliot_n, van_tilbeurgh_h, melki_r, velours_j, graille_m_s, janin_j, zhou_cz, blondeau_k, ballesta_jp, yokoyama_s, bousset_l, vershon_ak, bowler_be, zhang_y, arshava_b, buchner_j, wickner_rb, steven_ac, wang_y, zhang_m, forgac_m, brethes_d
Analysis	Protein structure is an important area of study. Proteins are composed of amino-acid residues, functionally important protein regions are called domains, and functionally important sites are often "converved" (i.e., many related proteins have the same amino-acid at the site). The RPS* proteins all part of the smaller (40S) subunit of the ribosome. Naider, Becker, and Leulliot study protein structure.

(b) Protein structure

Words	transcription, ii, histone, chromatin, complex, polymerase, transcriptional, rna, promoter, binding, dna, silencing, h3, factor, genes
Proteins	rpl16b, rpl26b, rpl24a, rpl18b, rpl18a, rpl12b, rpl6b, rpp2b, rpl15b, rpl9b, rpl40b, rpp2a, rpl20b, rpl14a, rpp0
Authors	workman_jl, struhl_k, winston_f, buratowski_s, tempst_p, erdjument-bromage_h, kornberg_rd_a, svejstrup_jq, peterson_cl, berger_sl, grunstein_m, stillman_dj, cote_j, cairns_br, shilatifard_a, hampsey_m, allis_cd, young_ra, thuriaux_p, zhang_z, sternglanz_r, krogan_nj, weil_pa, pillus_l
Analysis	In transcription, DNA is unwound from histone complexes (where it is stored compactly) and converted to RNA. This process is controlled by transcription factors, which are proteins that bind to regions of DNA called promoters. The RPL* proteins are part of the larger subunit of the ribosome, and the RPP proteins are part of the ribosome stalk. Many of these proteins bind to RNA. Workman, Struhl, and Winston study transcription regulation and the interaction of transcription with the restructuring of chromatin (a combination of DNA, histones, and other proteins that comprises chomosomes).

(c) Chromosome remodeling and transcription

TABLE 13.3
Top words, proteins, and authors: Topics obtained using Block-LDA on the PPI+SGD dataset.

(Kuhn, 1955) to solve the assignment problem where the cost of aligning topics together is determined using the Jensen-Shannon divergence measure.

Once the topics are obtained, we first obtain the proteins associated with the topic by retrieving the top proteins from the multinomial distribution corresponding to proteins. Then, the top articles corresponding to each topic are obtained using a ranked list of documents with the highest mass of their topic proportion distributions (θ) residing in the topic considered.

Manual Evaluation

To evaluate the topics, a yeast biologist who is an expert in the field was asked to mark each topic with a binary flag indicating if the top words of the distribution represented a coherent sub-topic in yeast biology. The top words of the distribution representing a topic were presented as a ranked list of words. This process was repeated for the three different variants of the model. The variant used to obtain results is concealed from the evaluator to remove the possibility of bias.

In the next step of the evaluation, the top articles and proteins assigned to each topic were presented in a ranked list and a similar judgment was requested to indicate if the article/protein was relevant to the topic in question. Similar to the topic coherence judgments, the process was repeated for each variant of the model. Screenshots of the tool used for obtaining the judgments can be seen in Figure 13.6. It should be noted that since the nature of the topics in the literature considered was highly technical and specialized, it was impractical to get judgments from multiple annotators.

To evaluate the retrieval of the top articles and proteins, we measure the quality of the results by computing the precision@10 score.

First, we evaluate the coherence of the topics obtained from the three variants described above. Table 13.4 shows that out of the 15 topics that were obtained, 12 topics were deemed coherent from the text-only model and 13 and 15 topics were deemed coherent from the Block-LDA models using the MIPS and wetlab PPI datasets, respectively.

Variant	Num. Coherent Topics
Only Text	12 / 15
Text + MIPS	13 / 15
Text + Wetlab	15 / 15

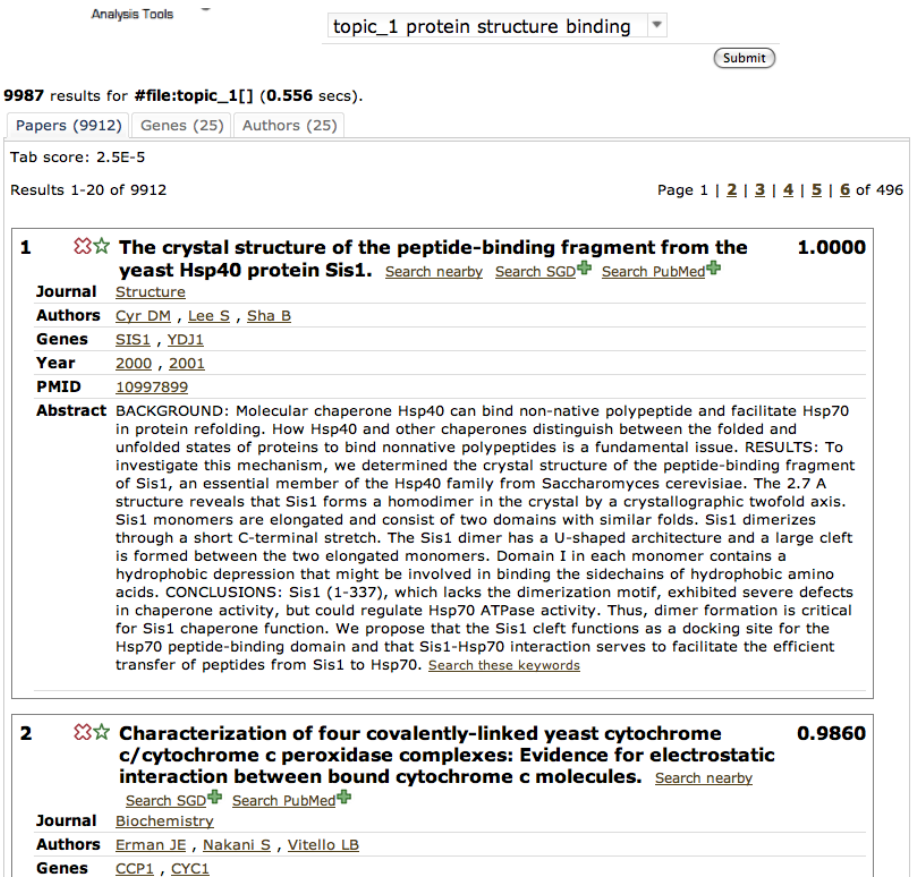
TABLE 13.4
Topic coherence evaluation.

Next, we study the precision@10 values for each topic and variant of the article retrieval and protein retrieval tasks (see Figures 13.7 or 13.8, respectively). The horizontal lines in the plots represent the mean of the precision@10 across all topics. It can be seen from the plots that for both the article and protein retrieval tasks, on average the joint models work better than the text-only model. For the article retrieval task, the model trained with the text + MIPS resulted in the higher mean precision@10 whereas for the protein retrieval task, the text + Wetlab PPI dataset returned a higher mean precision@10 value. For both the protein retrieval and paper retrieval tasks, the improvements shown by the joint models using either of the PPI datasets over the text-only model (i.e., the Link LDA model) were statistically significant at the 0.05 level using the paired Wilcoxon sign test. However, the difference in performance between the two joint models that used the two different PPI networks was insignificant, which indicates that there is no observable advantage in using one PPI dataset over the other in conjunction with the text corpus.

Functional Category Prediction

Proteins are identified as belonging to multiple functional categories in the MIPS PPI dataset, as described in Section 13.4. We use Block-LDA and baseline methods to predict proteins' functional

FIGURE 13.6
Screenshot: Article relevance annotation tool.



categories and evaluate them by comparing them to the ground truth in the MIPS dataset using the method presented in prior work (Airoldi et al., 2008). A model is first trained with K set to 15 topics to recover the 15 top-level functional categories of proteins. Every topic that is returned consists of a set of multinomials including $\beta_{\mathbf{t}_1}$, the topic-wise distribution over all proteins. The values of $\beta_{\mathbf{t}_1}$ are thresholded such that the top $\approx 16\%$ (the density of the protein-function matrix) of entries are considered as such a positive prediction that the protein falls in the functional category corresponding to the latent topic. To determine the mapping of latent topic to functional category, 10% of the proteins are used in a procedure that greedily finds the alignment resulting in the best accuracy, as described in Airoldi et al. (2008). It is important to note that the true functional categories of proteins are completely hidden from the model. The functional categories are used only during evaluation of the resultant topics from the model.

The precision, recall, and F_1 scores of the different models in predicting the right functional categories for proteins are shown in Table 13.5. Since there are 15 functional categories and a protein has approximately 2.5 functional category associations, we expect only $\sim$1/6 of protein-functional category associations to be positive. Precision and recall therefore depict a better picture of the predictions than accuracy. For the random baseline, every protein-functional category pair is randomly deemed to be 0 or 1 with the Bernoulli probability of an association being proportional to the ratio of 1s observed in the protein-functional category matrix in the MIPS dataset. In the

MMSB approach, induced latent blocks are aligned to functional categories as described in Airoldi et al. (2008).

We see that the F_1 scores for the baseline sparse blockmodel and MMSB are nearly the same, and that combining text and links provides a significant boost to the F_1 score. This suggests that protein co-occurrence patterns in the abstracts contain information about functional categories that is also evidenced by the better than random F_1 score obtained using Link LDA, which uses only documents. All the methods considered outperform the random baseline.

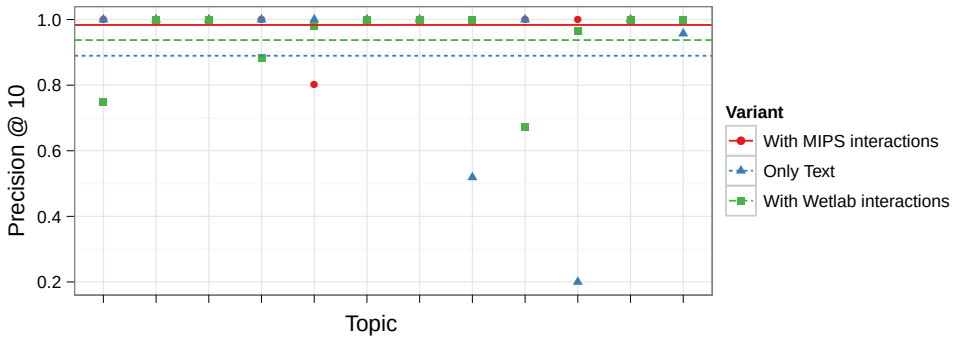


FIGURE 13.7
Retrieval performance - Article retrieval.

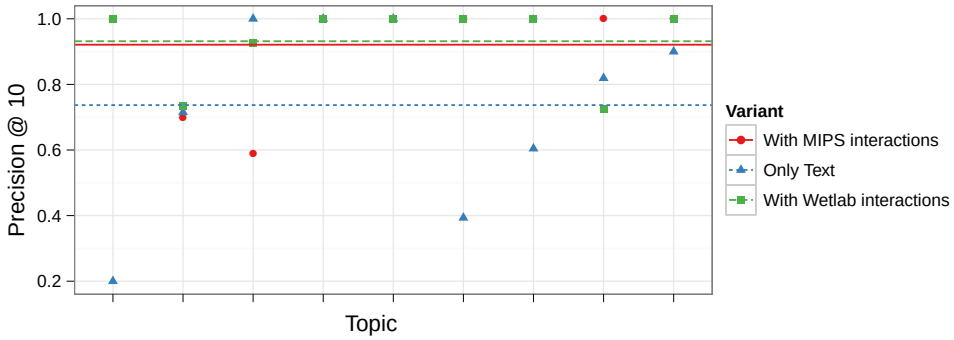


FIGURE 13.8
Retrieval performance - Protein retrieval.

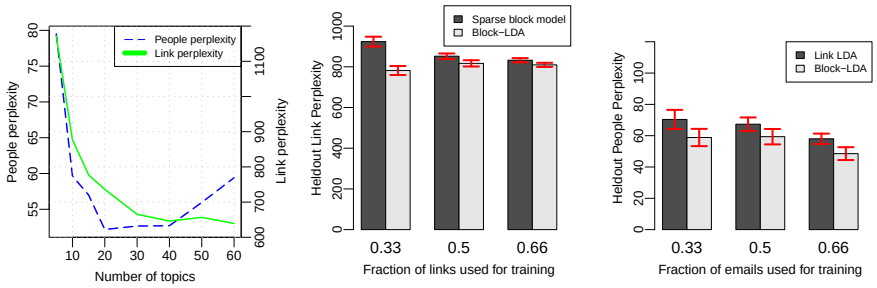
13.5.2 Results from the Enron Email Corpus Dataset

As described in Section 13.4, the Enron dataset consists of two components—text from the sent folders and the network of senders and recipients of emails within the Enron organization. Each email is treated as a document and is annotated with a set of people consisting of the senders and recipients of the email. We first study the network reconstruction capability of the Block-LDA model. Block-LDA is trained using all 96,103 emails in the sent folders and the 200,404 links obtained from the full email corpus. Figures 13.3(a), 13.3(b), and 13.3(c) show the true communication

Method	F_1	Precision	Recall
Block-LDA	**0.249**	0.247	0.250
Sparse Blockmodel	0.161	0.224	0.126
Link LDA	0.152	0.150	0.155
MMSB	0.165	0.166	0.164
Random	0.145	0.155	0.137

TABLE 13.5
Functional category prediction.

matrix, the matrix reconstructed using the sparse mixed membership stochastic blockmodel and the matrix reconstructed using the Block-LDA model, respectively. The figures show that both models are approximately able to recover the communication network in the Enron dataset.



(a) Determining number of topics

(b) Heldout link perplexity

(c) Heldout people perplexity

FIGURE 13.9
Enron corpus: Perplexity.

Figure 13.9(a) shows the link perplexity and person perplexity in text of held-out data, as the number of topics is varied. Person perplexity is indicative of the surprise inherent in observing a sender or a recipient and can be used as a prior in tasks like predicting recipients for emails that are being composed. Link perplexity is a score for the quality of link prediction and captures the notion of social connectivity in the graph. It indicates how well the model is able to capture links between people in the communication network. The person perplexity in the plot decreases initially and stabilizes when the number of topics reaches 20. It eventually starts to rise again when the number of topics is raised above 40. The link perplexity on the other hand stabilizes at 20 and then exhibits a slight downward trend. For the remaining experiments with the Enron data, we set $K = 40$.

In the next set of experiments, we evaluate Block-LDA and other models by evaluating the person perplexity in held-out emails by varying the training and test set size. Similar to the experiments with the PPI data, the Gibbs sampler is run until the held-out perplexity stabilizes to a nearly constant value ($\approx$ 80 iterations). The perplexity values are averaged over 10 runs. Figure 13.9(c) shows the person perplexity in text of held-out data as increasing amounts of the text data are used for training. The remainder of the dataset is used for testing. It is important to note that only Block-LDA uses the communication link matrix. A consistent improvement in person perplexity can be observed when email text data are supplemented with communication link data irrespective of the training set size. This indicates that the latent block structure in the links is beneficial while shaping latent topics from text.

Block-LDA is finally evaluated using link prediction. The sparse blockmodel, which serves as a baseline, does not use any text information. Figure 13.9(b) shows the perplexity in held-out data with varying amounts of the 200,404 edges in the network used for training. When textual information is used, all 96,103 emails are used. The histogram shows that Block-LDA obtains lower perplexities than the sparse blockmodel, which uses only links. As in the PPI experiments, using the text in the emails improves the modeling of the network of senders and recipients, although the effect is less marked when the number of links used for training is increased. The topical coherence in the latent topics induces better latent blocks in the matrix indicating a transfer of signal from the text to the network model.

13.6 Conclusion

We proposed a model that jointly models entity-entity links and entity-annotated text that permits co-occurence information in text to influence link modeling and vice-versa. Our experiments show that joint modeling outperforms approaches that use only a single source of information. Improvements are observed when the joint model is evaluated internally using perplexity in two different datasets and externally using protein functional category prediction in the yeast dataset. We also evaluated topics obtained from the joint modeling of yeast biology literature and protein-protein interactions in yeast and compared them to topics that were obtained from using only the literature. The topics were evaluated for coherence and by measuring the mean precision@10 score of the top articles and proteins that were retrieved for each topic. Evaluation by a domain expert showed that the joint modeling produced more coherent topics and showed better precision@10 scores in the article and protein retrieval tasks indicating that the model enabled information sharing between the literature and the PPI networks.

Acknowledgments

This work was funded by grant 1R101GM081293 from NIH, IIS-0811562 from NSF, and by a gift from Google. The opinions expressed in this paper are solely those of the authors.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9: 1981–2014.

Balasubramanyan, R. and Cohen, W. W. (2011). Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In *Proceedings of the 2011 SIAM Conference on Data Mining (SDM '11)*. SIAM/Omnipress, 450–461.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Chang, J. and Blei, D. M., (2009). Relational topic models for document networks. In *Proceedings*

of the of 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009). Journal of Machine Learning Research – Proceedings Track 5, 81–88.

Chang, J., Boyd-Graber, J., and Blei, D. M. (2009). Connections between the lines: Augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. New York, NY, USA: ACM, 169-178.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML '07)*. New York, NY, USA: ACM, 233-240.

Dwight, S. S., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J., Hong, E. L., Issel-Tarver, L., Nash, R. S., Sethuraman, A., Starr, B., Theesfeld, C. L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Weng, S., Botstein, D. and Cherry J., M. (2004). Saccharomyces genome database: Underlying principles and organisation. *Briefings in Bioinformatics* 5: 9.

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. D. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* 101: 5220.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* 101 Suppl 1: 5228–5235.

Gruber, A., Rosen-Zvi, M., and Weiss, Y. (2008). Latent topic models for hypertext. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*. Corvallis, OR, USA: AUAI Press, 230–239.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2: 83–97.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005). Topic and role discovery in social networks. *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI '05)*. IJCAI, 786—791.

Mewes, H. -W., Amid, C., Arnold, R., Frishman, D., Gldener, U., Mannhaupt, G., Mnsterktter, M., Pagel, P., Strack, N., Stmpflen, V., Warfsmann, J., and Ruepp, A. (2004). MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Research* 32: 41–44.

Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD '08)*. New York, NY, USA: ACM, 542–550.

Parkkinen, J., Sinkkonen, J., Gyenge, A., and Kaski, S. (2009). A block model suitable for sparse graphs. In *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009)*. Leuven, Belgium: poster presented.

Rosen-Zvi, M., Griffiths, T. L., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI 2004)*. Arlington, VA, USA: AUAI Press, 487–494.

Shetty, J. and Adibi, J. (2004). The Enron Email Dataset Database Schema and Brief Statistical Report. Tech. report, Information Sciences Institute.

Wang, X., Mohanty, N., and McCallum, A. (2006). Group and topic discovery from relations and their attributes. In Weiss, Y., Schölkopf, B., and Platt, J. (eds), *Advances in Neural Information Processing Systems 18*. Cambridge, MA: The MIT Press, 1449—1456.

Wen, Z. and Lin, C. -Y. (2010). Towards finding valuable topics. In *Proceedings of the 2010 SIAM Conference on Data Mining (SDM '10)*. Philadelphia, PA, USA: SIAM, 720–731.