Block-LDA: Jointly modeling entity-annotated text and entity-entity links

Ramnath Balasubramanyan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
rbalasub@cs.cmu.edu

William W. Cohen
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
wcohen@cs.cmu.edu

Abstract

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 block models 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 model is evaluated by inspecting induced topics to understand the nature of the data and by quantitative methods such as functional category prediction of proteins and perplexity which exhibit improvements when joint modeling is used over baselines that use only link or text information.

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 block models (MMSB) [1, 2] 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) [3] 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 [4, 5] 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. [6].

In this paper, we present a model, *Block-LDA*, that jointly generates text documents annotated with entities and external links between pairs of entities allowing it to use supplementary annotated text to influence and improve link modeling. The model merges the idea of latent topics in topic models with blocks in stochastic block models. 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 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. The rest of the paper is organized as follows. Section 2 introduces the model and presents a Gibbs sampling based method for performing approximate inference with the model. Section 3 discusses related work and Section 4 provides details of datasets used in the experiments. Sections 5 and 6 presents the results of our experiments on two datasets from different domains. Finally, our conclusions are in Section 7.

2 Block-LDA

The Block-LDA model (plate diagram in Figure 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,
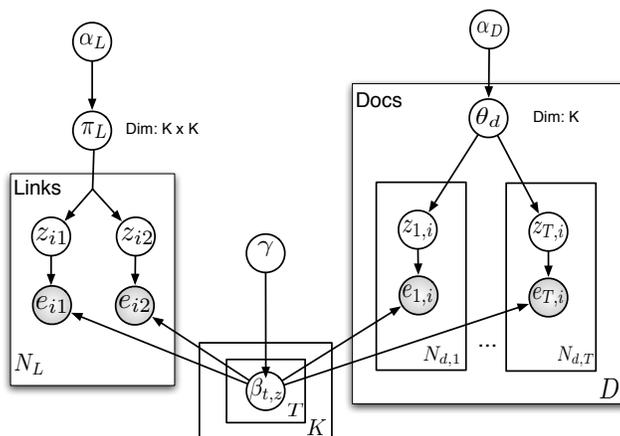
α_L - Dirichlet prior for the topic pair distribution for links
α_D - Dirichlet prior for document specific topic distributions
γ - Dirichlet prior for topic multinomials
π_L - multinomial distribution over topic pairs for links
θ_d - multinomial distribution over topics for document d
$\beta_{t,z}$ - multinomial over entities of type t for topic z
$z_{t,i}$ - 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
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

Figure 1: Block-LDA

through shared latent topics. More specifically, the distribution over the entities of the type that are linked is shared between the block model and the text model.

The component on the right, which is an extension of the LDA models documents as sets of "bags of entities", 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 in the figure is a generative model for graphs representing entity-entity links with an underlying block structure, derived from the sparse block model introduced by Parkkinen et al. [2]. 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 graph are between entities of type t_l, 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 $\beta_{t,z} \sim \text{Dirichlet}(\gamma)$, the topic specific entity distribution.

2. Generate documents. For every document $d \in \{1 \ldots D\}$:

- Sample $\theta_d \sim \text{Dirichlet}(\alpha_D)$ 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 $z_{t,i} \sim \text{Multinomial}(\theta_d)$
    - Sample an entity $e_{t,i} \sim \text{Multinomial}(\beta_{t,z_{t,i}})$

3. Generate the link matrix of entities of type t_l:

- Sample $\pi_L \sim \text{Dirichlet}(\alpha_L)$ where π_L describes a distribution over the Cartesian product of topics, for links in the dataset.

- For every link $e_{i1} \to e_{i2}$, $i \in \{1 \cdots N_L\}$:
  - Sample a topic pair $\langle z_{i1}, z_{i2} \rangle \sim$ Multinomial(π_L)
  - Sample $e_{i1} \sim \text{Multinomial}(\beta_{t_l, z_{i1}})$
  - Sample $e_{i2} \sim \text{Multinomial}(\beta_{t_l, z_{i2}})$

Note that unlike the MMSB model introduced by Airoldi et al. [1], this modle 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

$$(2.1) \quad p(\pi_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{e}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{e_1}, \mathbf{e_2} \rangle | \alpha_D, \alpha_L, \gamma) \propto$$

$$\prod_{z=1}^{K} \prod_{t=1}^{T} \text{Dir}(\beta_{t,z} | \gamma_t) \times$$

$$\prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha_D) \prod_{t=1}^{T} \prod_{i=1}^{N_{d,t}} \theta_d^{z_{t,i}^{(d)}} \beta_{t, z_{t,i}^{(d)}}^{e_{t,i}} \times$$

$$\text{Dir}(\pi_L | \alpha_L) \prod_{i=1}^{N_L} \pi_L^{\langle z_{i1}, z_{i2} \rangle} \beta_{t_l e, z1}^{e_{i1}} \beta_{t_l e, z2}^{e_{i2}}$$

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):

$$(2.2) \quad p(z_{t,i} = z | e_{t,i}, \mathbf{z}^{\neg i}, \mathbf{e}^{\neg i}, \alpha_D, \gamma)$$

$$\propto (n_{dz}^{\neg i} + \alpha_D) \frac{n_{zte_{t,i}}^{\neg i} + \gamma}{\sum_{e'} n_{zte'}^{\neg i} + |E_t|\gamma}$$

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:

$$(2.3) \; p(\mathbf{z_i} = \langle z_1, z_2 \rangle | \langle e_{i1}, e_{i2} \rangle, \mathbf{z}^{\neg i}, \langle \mathbf{e_1}, \mathbf{e_2} \rangle^{\neg i}, \alpha_L, \gamma)$$

$$\propto \left(n_{\langle z_1, z_2 \rangle}^{L \neg i} + \alpha_L \right) \times$$

$$\frac{(n_{z_1 t_l e_{i1}}^{\neg i} + \gamma)(n_{z_2 t_l e_{i2}}^{\neg i} + \gamma)}{\left(\sum_e n_{z_1 t_l e}^{\neg i} + |E_{t_l}|\gamma \right) \left(\sum_e n_{z_2 t_l e}^{\neg i} + |E_{t_l}|\gamma \right)}$$

E_t refers to the set of all entities of type t. The n's are counts of observations in the training set.

- 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 MLE estimates after inference using the counts of observations.

$$(2.4) \quad \beta_{t,z}^{(e)} = \frac{n_{zte} + \gamma}{\sum_{e'} n_{zte'} + |E_t|\gamma},$$

$$(2.5) \quad \theta_d^{(z)} = \frac{n_{dz} + \alpha_D}{\sum_{z'} n_{dz'} + K\alpha_D} \text{and}$$

$$(2.6) \quad \pi_L^{\langle z_1, z_2 \rangle} = \frac{n_{\langle z_1, z_2 \rangle} + \alpha_L}{\sum_{z'_1, z'_2} n_{\langle z'_1, z'_2 \rangle} + K^2\alpha_L}$$

A de-noised form of the entity-entity link matrix can also be recovered from the estimated parameters of the model. Let B be a matrix of dimensions $K \times |E_{t_l}|$ where row $k = \beta_{t_l, k}$, $k \in \{1, \cdots, K\}$. 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 $M = B^T Z B$

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[6] combines MMSB with LDA by modeling documents using LDA and generating links between them using MMSB. The Relational Topic Model [7] generates links between documents based on their topic distributions. The Copycat and Citation Influence models [8] also model links between citing and cited documents by extending LDA and eliminating independence between documents. The Latent Topic Hypertext Model (LTHM) [9] presents a generative process for documents that can be linked to each other from specific words in the citing document. The model proposed in this paper, Block-LDA, is different from this class of models in that they model links between entities in the documents rather than links between documents.

The Nubbi model [10] 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

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 block model (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 1: Related work

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 [11] 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) [12] 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)[13] 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 [14] describe an application of an approach that uses both content and network information to analyse 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 1.

4 Datasets

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 2: List of functional categories

The Munich Institute for Protein Sequencing (MIPS) database [15] 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 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 2). The 844 proteins participating in interactions are mapped to these 15 functional categories with an average of 2.5 annotations per protein. In addition to the MIPS PPI data, we use a text corpus that is derived from the repository of scientific publica-

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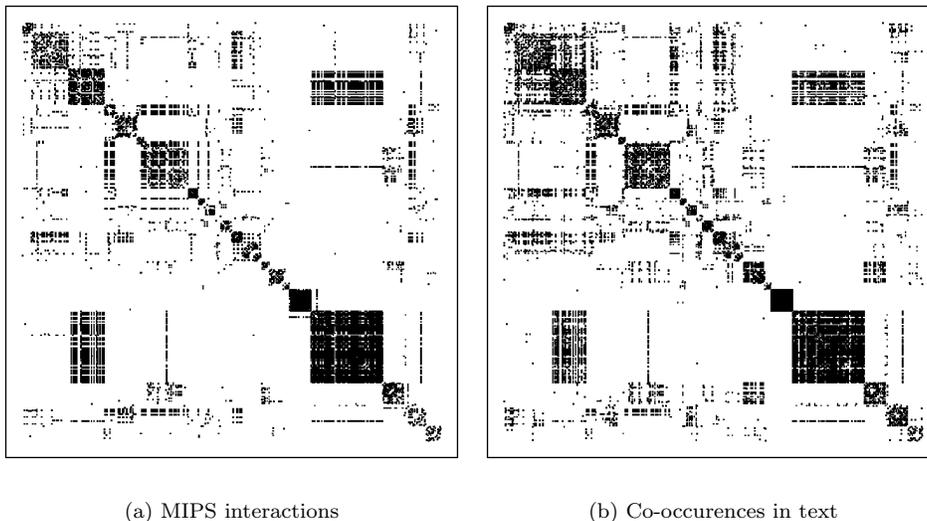
(b) Co-occurences in text

Figure 2: Observed protein-protein interactions compared to thresholded co-occurrence in text

tions at PubMed Central. PubMed is a free, open-access on-line archive of over 18 million biological abstracts and bibliographies, including citation lists, for papers published since 1948 (U.S. National Library of Medicine 2008). The subset we work with consists of approximately 40,000 publications about the yeast organism that have been curated in the Saccharomyces Genome Database (SGD) [16] 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.

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 2(b) shows that the resultant matrix which looks very similar to the MIPS PPI matrix in Figure 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 2(b) and 2(a) are ordered by their cluster ids, obtained by clustering them using k-means clustering, treating proteins as a 15-bit vectors of functional category annotations.

The Enron email corpus[17] 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 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"

---

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

Method	F_1	Precision	Recall
Block-LDA	**0.249**	0.247	0.250
Sparse Block model	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 4: Functional category prediction

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 5(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.

5 Results from the protein-protein interactions and abstracts dataset

5.1 Topic analysis An useful application of latent block modeling approaches is to understand the underlying nature of data. The topic wise multinomials for each type of entity induced by Block-LDA provide an overview of this nature. Table 3 shows the top words, proteins and authors for six topics induced by running Block-LDA over the full PPI+SGD dataset. 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 and a title and an analysis of the topic added, after the inference procedure. Details about proteins and yeast researchers were obtained on 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.

5.2 Matrix reconstruction Next, we investigate the ability of the model to recover the block structure inherent in the protein protein interactions. Figure 3 shows the reconstructed protein-protein interaction matrix using the sparse block model and Block-LDA. It can be seen that both matrices approximately resemble the observed PPI matrix in Figure 2(a) with Figure 3(b) being a crisper reconstruction.

5.3 Functional category prediction Proteins are identified as belonging to multiple functional categories in the MIPS dataset, as described in Section 4. We use

Block-LDA and baseline methods to predict proteins' functional categories and evaluate it by comparing it to the ground truth in the MIPS dataset using the method presented in prior work [1]. 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 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 [1]. 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 4. 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 1's 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 [1].

We see that the F_1 scores for the baseline sparse block model 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 as 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.

5.4 Perplexity and convergence Next, we investigate the convergence properties of the Gibbs sampler by observing link perplexity on heldout data at different epochs. Link perplexity of set of links L is defined as (5.7)

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

---

[2]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

Words	rna, mrna, nuclear, translation, pre, ribosomal, processing, complex, rrna, export, splicing, factor, required, prion, binding
Proteins	sup35, rpl3, rps2, rpl18a, rpl6a, rpl7a, rpl42b, rpl5, rpl18b, rps0b, rpl22a, rps11b, rpl27b, rpl32, rpl7b
Authors	tollervey_d, hurt_e, parker_r, wickner_rb, seraphin_b, corbett_ah, silver_pa, hinnebusch_c, baserga_sj, rosbash_m, beggs_jd, jacobson_a, liebman_sw, linder_p, petfalski_e, luhrmann_r, fromont-racine_m, ter-avanesyan_md, johnson_aw, raue_ha, keller_w, schwer_b, wente_sr, tuite_mf
Analysis	Translation is conversion of DNA to mRNA, a process that is followed by splicing (in which parts of the mRNA are removed). sup35 is a protein that terminates transcription; it also exists as a misfolded protein called a "prion". Tollervey, Hurt, and Parker study RNA processing and export.

(d) RNA maturation

| Words	dna, repair, replication, recombination, damage, cerevisiae, strand, saccharomyces, double, checkpoint, induced, telomere, role, homologous, complex
Proteins	rad52, rad51, rad54, rad57, rad55, msh2, mre11, rad50, xrs2, rad1, rad14, rfa1, rad10, rfa2, rfa3
Authors	haber_je, prakash_s, prakash_l, kolodner_rd, sung_p, burgers_pm, kunkel_ta, petes_w, jinks-robertson_s, resnick_ma, johnson_re, zakian_va, jackson_sp, enomoto_t, seki_m, heyer-wd, rothstein_r, alani_e, gasser_sm, campbell_jl, haracska_l, boiteux_s, symington_ls, foiani_m
Analysis	DNA repair is required because errors sometimes occur in when DNA is replicated. RAD52, RAD51, RAD54, RAD57, RAD55, MSH2, and MRE11 are involved in DNA repair. Haber and S. Prakash study DNA repair, and L. Prakash is a frequent co-author with S. Prakash.

(e) DNA repair

Table 3: Top words, proteins and authors: topics obtained using Block-LDA on the PPI+SGD dataset
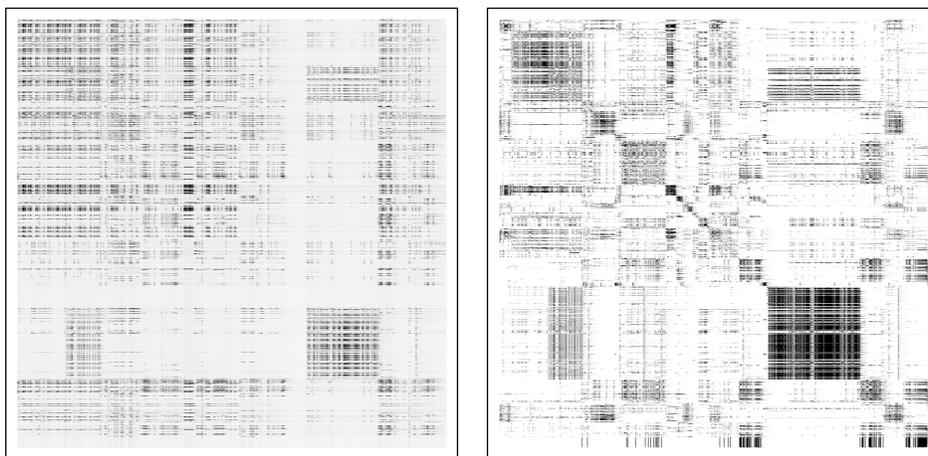
(a) Sparse block model

(b) Block-LDA

Figure 3: Inferred protein-protein interactions
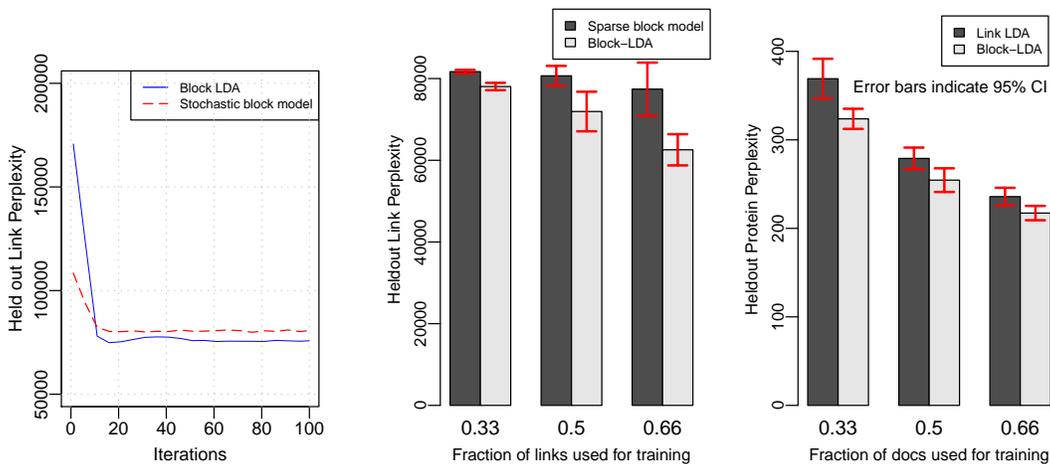
(a) Gibbs sampler convergence

(b) Gain in perplexity through joint modeling

Figure 4: Evaluating perplexity in the PPI+SGD dataset

Figure 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 heldout 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+SGD dataset. The SGD text data has 3 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 heldout protein-protein interactions using increasing amounts of the PPI data for training.

All the 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 1. Figures 3(a) and 3(b) shows the posterior likelihood of protein-protein interactions

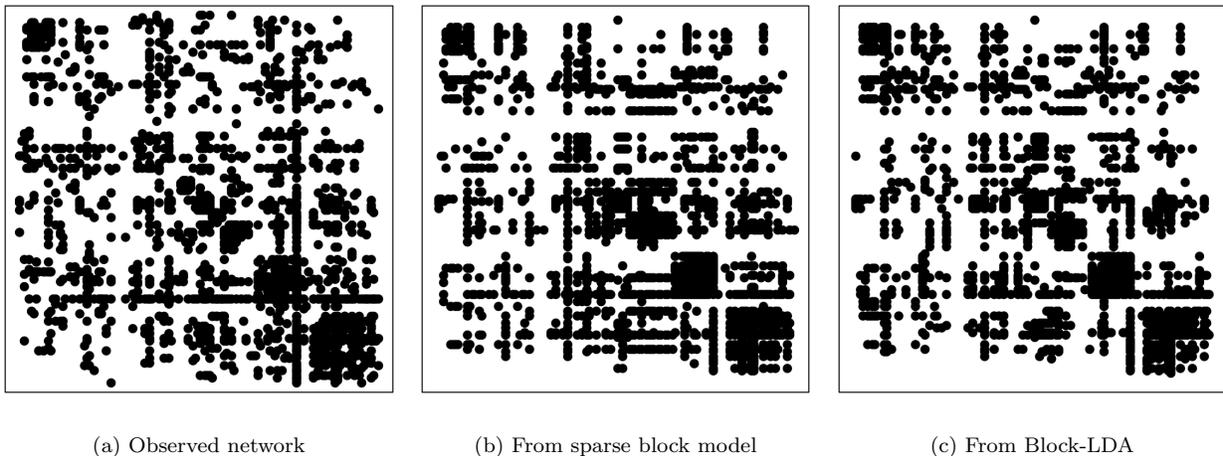(a) Observed network   (b) From sparse block model   (c) From Block-LDA

Figure 5: Enron network and its de-noised recovered versions

recovered using the sparse block model and using Block-LDA respectively. In the other set of experiments, we evaluate the model using protein perplexity in heldout 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 is used. When link data is 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 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. 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.

6 Enron email corpus

As described in Section 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 the 96,103 emails in the sent folders and the 200,404 links obtained from the full email corpus. Figures 5(a), 5(b)

and 5(c) show the true communication matrix, the matrix reconstructed using the sparse mixed membership stochastic block model 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.

Next, we study the top words and people in the topics induced by Block-LDA shown in Table 5. The table shows sample topics induced after running Block-LDA with K set to 15. We present only a subset of the fifteen topics due to space limitations. The topic labels and notes were hand created after looking at the top words and employees and by using the partial knowledge available about the roles of the employees in the Enron organization [17]. It can be seen that the people within the recovered topics are likely to need to communicate with each other. These instances of topics suggest that the topics capture both notions of semantic concepts obtained from the text of the emails and sets of people who need to interact regularly about the concepts.

Figure 6(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

Words	contract, party, capacity, gas, df, payment, service, tw, pipeline, issue, rate, section, project, time, system, transwestern, date, el, payment, due, paso
Employees	fossum, scott, harris, hayslett, campbell, geaccone, hyatt, corman, donoho, lokay
Notes	Geaconne was the executive assistant to Hayslett who was the Chief Financial Officer and Treasurer of the Transwestern division of Enron.

(a) Financial contracts

Words	power, california, energy, market, contracts, davis, customers, edison, bill, ferc, price, puc, utilities, electricity, plan, pge, prices, utility, million, jeff
Employees	dasovich, steffes, shapiro, kean, williams, sanders, smith, lewis, wolfe, bass
Notes	Dasovitch was a Government Relations executive, Steffes the VP of government affairs, Shapiro, the VP of regulatory affairs and Haedicke worked for the legal department.

(b) Energy distribution

Words	enron, business, management, risk, team, people, rick, process, time, information, issues, sally, mike, meeting, plan, review, employees, operations, project, trading
Employees	kitchen, beck, lavorato, delainey, buy, presto, shankman, mcconnell, whalley, haedicke
Notes	The people in this topic are top level executives: Kitchen was the President of Enron Online, Beck the Chief operating officer and Lavarato the CEO.

(c) Strategy

Words	deal, deals, dec, mid, book, pst, columbia, please, pl, kate, desk, west, changed, file, questions, mike, report, books, mw, thanks
Employees	love, semperger, symes, giron, keiser, williams, mclaughlin, white, forney, grigsby
Notes	This topic about trading has Semperger in the most likely list of people who was a senior analyst dealing with cash accounts and Forney who worked as a trader at the real time trading desk.

(d) Trading

Words	legal, trading, credit, master, energy, eol, isda, list, counterparty, company, financial, agreement, power, trade, inc, access, products, mark, approval, swap, request
Employees	dasovich, sanders, haedicke, kean, steffes, derrick, harris, williams, shapiro, davis
Notes	As noted before, Dasovich, Haedicke and Steffes performed roles that involved interacting with government agencies.

(e) Legal and regulatory affairs

Words	gas, storage, volumes, volume, demand, capacity, transport, ces, deal, price, day, month, daily, market, ena, contract, power, prices, cash, index
Employees	germany, farmer, grigsby, tholt, townsend, smith, parks, neal, causholli, hernandez
Notes	Farmer was a logistics manager and Tholt who was the VP of the division.

(f) Logistics

Table 5: Top words and people from latent topics in the Enron corpus

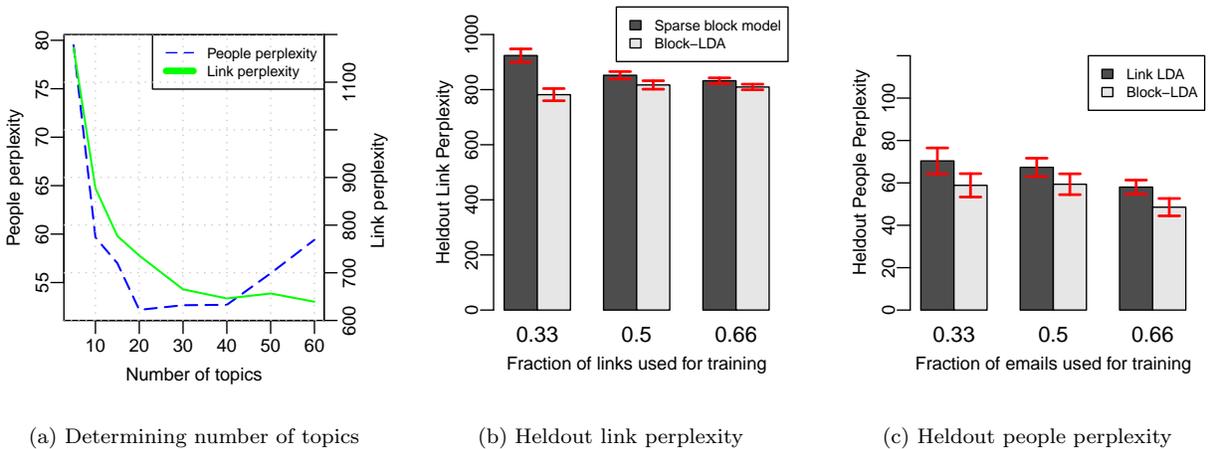| (a) Determining number of topics | (b) Heldout link perplexity | (c) Heldout people perplexity |

Figure 6: Experiments with the Enron corpus

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 (≈ 80 iterations). The perplexity values are averaged over 10 runs. Figure 6(c) shows the person perplexity in text in 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 is 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 block model which serves as a baseline does not use any text information. Figure 6(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 the 96,103 emails are used. The histogram shows that Block-LDA obtains lower perplexities than the sparse block model 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.

7 Conclusion

We proposed a model that jointly models links between entities and text annotated with entities that permits co-occurrence 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. Moreover, the topics induced by the model when examined subjectively appear to be useful in understanding the structure of the data both in terms of the topics discussed and in terms of the connectivity characteristics between entities.

Acknowledgements

References

[1] Edoardo M. Airoldi, David Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, September 2008.

[2] Juuso Parkkinen, Janne Sinkkonen, Adam Gyenge, and Samuel Kaski. A block model suitable for sparse

graphs. In *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009)*, Leuven, 2009. Poster.

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

[4] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220, 2004.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.

[6] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, Las Vegas, Nevada, USA, 2008. ACM.

[7] J. Chang and D. M Blei. Relational topic models for document networks. In *Proc. of Conf. on AI and Statistics (AISTATS 09)*, 2009.

[8] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007)*, page 233240, 2007.

[9] Amit Gruber, Michal Rosen-zvi, and Yair Weiss. Latent topic models for hypertext. *UAI*, 2008.

[10] Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 169178, 2009.

[11] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and their attributes. *In Advances in Neural Information Processing Systems 18*, pages 1449—1456, 2006.

[12] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 494, 487. AUAI Press, 2004.

[13] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. *IN IJCAI*, pages 786—791, 2005.

[14] Zhen Wen and Ching-Yung Lin. Towards finding valuable topics. In *SDM*, pages 720–731, 2010.

[15] H. W. Mewes, D. Frishman, K. F. X. Mayer, M. Mnsterktter, O. Noubibou, T. Rattei, M. Oesterheld, and V. Stmpflen. Mips: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32:41–44, 2004.

[16] S.S. Dwight, R. Balakrishnan, K.R. Christie, M.C. Costanzo, K. Dolinski, S.R. Engel, B. Feierbach, D.G. Fisk, J. Hirschman, E.L. Hong, et al. Saccharomyces genome database: Underlying principles and organisation. *Briefings in bioinformatics*, 5(1):9, 2004.

[17] Jitesh Shetty and Jafar Adibi. The enron email dataset database schema and brief statistical report, 2004.