Text Network Exploration via Heterogeneous Web of Topics

***Abstract*—A text network refers to a data type that each vertex is associated with a text document and the relationship between documents is represented by edges. The proliferation of text networks such as hyperlinked webpages and academic citation networks has led to an increasing demand for quickly developing a general sense of a new text network. To satisfy this requirement, the notion of *exploratory search* is proposed. While previous works in exploring text network mainly focus on projecting high dimensional words data into low dimensional space of word topics and mining relation patterns among these word topics, we further introduce document topics and thus allow people to investigate a text network on both word level and document level in a low dimensional space. To model word topic and document topic in a unified framework, we propose a probabilistic generative model named *MHT*(Model for Heterogeneous Topic web) for joint analysis of text and links. In MHT, three different relationships among these two types of topics are quantified, based on which we construct a heterogeneous web of topics for the exploration task. We develop a prototype demo system named *TopicAtlas* to exhibit such heterogeneous topic web, and demonstrate how this system can facilitate the task of text network exploration. Extensive qualitative analysis are included to verify the effectiveness of this heterogeneous topic web. Besides, we validate our model on real-life text networks, showing that it outperforms comparable baselines on objective evaluation metrics.**

I. Introduction

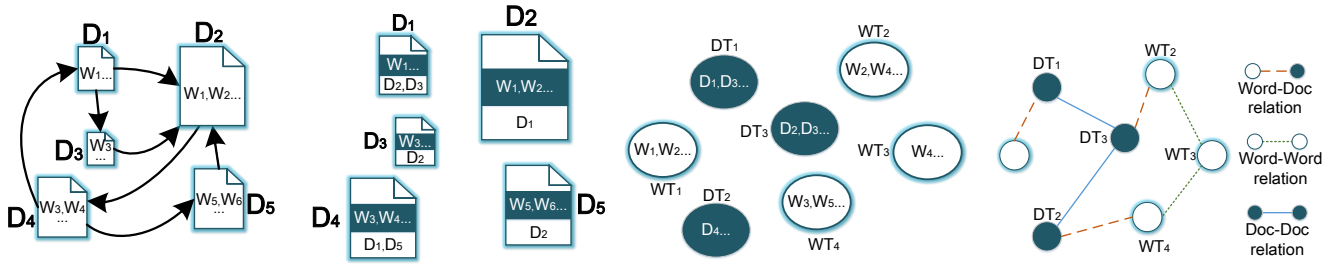
The information age has witnessed an increasing amount of unstructured data, most of which are in the form of text and possess high degrees of connectivity among themselves. We refer to this type of data as *text network* in which each vertex is associated with a text document and the relationship between documents is represented by the edges as shown in Fig.1a. Such text networks are ubiquitous in the real world. Typical representatives include hyperlinked webpages, online social network with user profiles, and academic citation network.

While the available text networks are continuously increasing, very often only little is known about the quantity, coverage and relations of their content [1]. Consequently, the general search engine, which requires users to have a specific set of keywords in mind before pursuing further investigation, fails to help [2]. When faced with a new or unfamiliar text network, people may first ask a more basic question: "What is there?". To answer this question, we resort to the notion of *exploratory search* [3] which is proposed to help people develop a general sense of the properties of new datasets before embarking on more specific inquiries [4].

Due to its importance, exploratory search has been investigated intensively. For example, Sinclair *et al.* use word frequency lists, frequency distribution plots and keyword-in-context models to enhance computer-assisted reading [5]. More recently, a computational technique named "topic modeling" is widely used to fulfill this task [4], [6], [7]. By exploiting the co-occurrence counts of words across the whole text network, such method can identity semantic clusters of words, defined as *topics* [8], which provides insight into a corpus' contents. Nevertheless, existing topic models are still far from adequate for text network exploration since the significance of topics represented only by words is limited for exploration task due to lack of an insight on document level.

To address these problems, we first dive deeper into the document level and exploit link structures in text network to obtain document clusters. We achieve this through viewing the text network as a set of "link documents" where each is a document represented by a "bag of links" [9], [10]. A document with k "neighbours" is viewed as a "document" with k "link tokens", each corresponding to the index of a document that links to it, thus we can also regard these "link tokens" as "document tokens". Then, we can model these "link documents" with a topic model framework where a new type of "topics" characterized by distributions over documents arises. Intuitively, this new type of topics captures the co-occurrence pattern of these "document tokens" and can be viewed naturally as a document cluster. By combining "word token" and "document token", each document is composed of two parts as shown in Fig.1b, and two different types of topics are included as illustrated in Fig.1c. To distinguish them, we call them *WordTopic* and *DocTopic*.

However, simple introduction of DocTopic is still insufficient for text network exploration since we cannot utilize the two types of topics in a united manner due to the lack of connection between them. Although different types of relationship between WordTopics have been investigated previously [10]–[17], a more important type of connection between WordTopic and DocTopic has never been studied before. This relationship is grounded in the intuition that there exists a co-occurrence pattern between WordTopic and DocTopic. For instance, WordTopic about "*robot navigation*" will frequently appear with DocTopic "*robot motion planning*" since planning is a core issue of navigation. Indeed, such Word-Doc relationship serves as a bridge that connects two types of topics and helps us to make the best sense of the underlying text network. Therefore, a heterogeneous web of topics as described in Fig.1d is required to identify two types of topics and uncover the multiple latent relationship, which includes the relation between WordTopic and WordTopic (*Word-Word relation*), DocTopic and DocTopic (*Doc-Doc relation*) and WordTopic and DocTopic (*Word-Doc relation*).



(a) Input Text Network (b) Two parts of a document (c) WordTopic and DocTopic (d) Heterogeneous Topic Web

Fig. 1. Illustration of some concepts. (a) Input text network. (b) Two parts of a document. W represents the "word token" part, and D below W represents "document token" part. (c) WordTopic (WT) and DocTopic (DT). (d) Heterogeneous topic web with two types of topics and three types of relationships.

To build such a heterogeneous topic web, we propose a probabilistic generative model called **MHT** (Model for Heterogeneous Topic web), to *jointly* cluster those semantic-alike words and documents into WordTopics and DocTopics, and the three types of relationships are included. Our experiments on two academic citation networks demonstrate that MHT not only produces reliable heterogeneous topic web with high-quality topics but also possesses strong generalizability and predictive power.

Furthermore, we construct **TopicAtlas**, a prototype demo system for convenient navigation in such heterogeneous topic web. TopicAtlas displays Word-Word relation, Doc-Doc relation, and Word-Doc relation in a unified framework. With TopicAtlas, users are able to freely wander around the text network via WordTopics and DocTopics. Also, users can pinpoint important documents (represented by DocTopics) with regard to a selected WordTopic, and identify important themes (represented by WordTopics) in a targeted document. We present the results of TopicAtlas in multiple views and show how it effectively facilitates the task of text network exploration. Finally, it is important to notice that TopicAtlas is not a general-purpose topic model browser such as [18]–[20]. The key difference is that these browsers are designed for understanding the results of topic models while TopicAtlas focuses on the specific exploration of text networks.

To summarize, our contributions are three folds:

- To the best of our knowledge, we are the first to present the idea of heterogeneous web of topics which consists of two types of topics: WordTopic and DocTopic, and three types of topic relations: Word-Word relation, Doc-Doc relation, and most importantly Word-Doc relation.
- We propose MHT, a probabilistic generative model that explicitly models Word-Doc relationship by capturing the joint co-occurrence pattern of semantic-alike word tokens and document tokens. Given a text network, MHT is able to extract two types of high-quality topics along with their heterogeneous relationships and produce a heterogeneous topic web.
- We construct TopicAtlas, a prototype system for text network exploration. TopicAtlas allows users to investigate the heterogeneous topic web with details and explore text network easily.

The rest of paper is organized as follows. In section II, we discuss some related works. We introduce MHT and its inference in section III and IV. In section V, we conduct the experiment and evaluate our model. Finally, we summarize this paper and discuss some future works in section VI.

II. Related Work

In terms of topic modeling. In this part we mainly focus on the exploring abilities of existing topic models.

Topic models are proposed to address the problem of topic identification in large document collections. In topic models, each document is associated with a topic distribution and each topic is associated with a word distribution. Two popular models in this field are Probabilistic Latent Semantic Analysis (PLSA) [21] and Latent Dirichlet Allocation (LDA) [22]. PLSA is a generative and unsupervised model, introducing latent topics into the generative process. Every word is assigned with a topic and every document has a document specific topic distribution. Compared to PLSA, LDA introduces a hyperparameter to address the overfitting issue. In LDA, the document specific topic distribution is generated according to Dirichlet distribution. These two models construct a generic framework for topic modeling.

Traditional topic models provide a convenient way to explore document collections and summarize high dimensional data with low dimensional topics. However, these models assume that topics are independent with each other and ignore their relationship. To obtain a deeper insight into the data, some works [10]–[17] investigate the connection between different topics. For example, CTM [13] models the topic correlation generatively, hoping to give a better fit on given data than LDA. Other works [11], [12] utilize topic relation to model the presence and absence of links in a pairwise manner. Citation-LDA [10] studies the topic dependency relationship and aims to obtain meaningful topic evolution pattern. CATHYHIN [15] hopes to build the topic hierarchy structure.

The *topic* mentioned so far is merely cluster of words, i.e. WordTopic. However, in text network those similar or linked documents can also be clustered to represent topics, i.e. DocTopic. In fact, Citation-LDA [10] models the links within topic model framework and clusters documents into different categories. By analogy with traditional topic models,

each cluster here is associated with a document distribution and the probability indicates the importance of one document in each category. Such DocTopic can be further exploited to classify documents [9].

Through the definition of WordTopic and DocTopic, there is a heterogeneous topic web hidden in a given text network. The connection exists between WordTopic and WordTopic, DocTopic and DocTopic, and WordTopic and DocTopic. However, all the above works fail to model these three relationships jointly nor construct a heterogeneous topic web comprehensively.

In terms of text network exploration. There are many recent efforts utilizing topic models to explore text networks. They focus on either new features or visualization techniques.

Chaney and Blei [20] make an early effort in this field. They present a method for visualizing traditional topic models, where a navigator of the documents is created, allowing users to explore the hidden structure discovered by a topic model. Later, Gretarsson *et al.* build a relatively mature system called TopicNets [6], which enables users to visualize individual document sections and their relations within the global topic document and topic nodes. Also, in TopicNets analysts can select relevant subsets of documents and perform real-time topic modeling on these subsets. Maiya *et al.* [23] develope the topic similarity network for exploration and recognize how topics form large themes. Some other works [14], [24] focus on topic dynamics and topic flow in text networks.

While the works mentioned above convey some information visually, they are orthogonal to our focus. The topic in these works is word clusters, and they do not study DocTopic and the hidden heterogeneous topic web.

III. Model for Heterogeneous Topic web

In this part we describe the framework and generative process of our model MHT (Model for Heterogeneous Topic web), whose graphical representation is illustrated in Fig.2.

A. Framework

We consider the input text network as a graph $G(V, E)$, where V is the set of document vertices and E is the set of directed edges or links. $v_i \in V$ represents the i^{th} document and $e_{ij} \in E$ connects two vertices v_i and v_j. Each document is associated with a bag of words and we denote w_{in} as the n^{th} word token in document v_i. To cluster links or documents into topics, we adopt the assumption of "bag of links" and y_{il} is denoted as the l^{th} link token (document token) in v_i.

Our work aims at detecting the relationship between WordTopic which is represented by distributions over the whole vocabulary and DocTopic which is represented by distributions over the entire corpora. Although WordTopic and DocTopic have been explored separately before, there is no mechanism to capture the relation between these two different but closely related topics.

In classical topic models each document is associated with a document specific topic distribution, which is used to draw a topic for each word in the generative process. Note that the "topic" here actually represents WordTopic in our notation framework. Similarly, adopt the assumption of "bag of links" and each document is able to be associated with a DocTopic distribution, which can generate documents. Since these two distributions are totally different, some transition procedure between them is required to jointly model text and links.

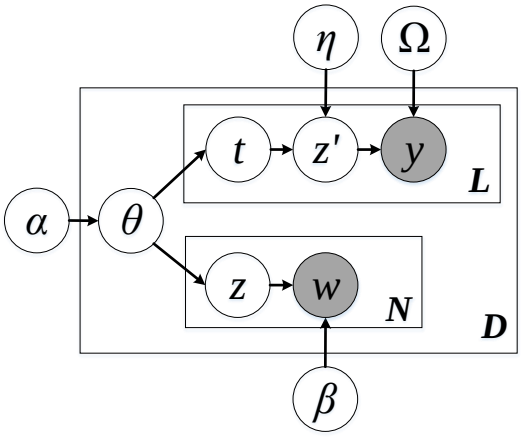


Fig. 2. Graphical Representation of MHT

Based on the discussion above, we employ a transition distribution over DocTopics η to depict the relation between the two types of topics.

B. Generative Process

Details for full generative process of our proposed model MHT is demonstrated below.

For each document v_i, where $i = 1, \cdots, D$:

1) Generate WordTopic distribution:
$$\theta_i \sim Dir(\cdot|\alpha)$$
2) For each word w_{in}, where $n = 1, \cdots, N_i$:
   a) Draw a WordTopic:
   $$z_{in} \sim Mult(\cdot|\theta_i)$$
   b) Draw a word:
   $$w_{in} \sim Mult(\cdot|\beta_{z_{in}})$$
3) For each link y_{il}, where $l = 1, \cdots, L_i$:
   a) Draw a transition topic:
   $$t_{il} \sim Mult(\cdot|\theta_i)$$
   b) Draw a DocTopic:
   $$z'_{il} \sim Mult(\cdot|\eta_{t_{il}})$$
   c) Draw a linked document:
   $$y_{il} \sim Mult(\cdot|\Omega_{z'_{il}})$$

Step 1 and Step 2 are the same as classical topic model to generate words. A major distinction of MHT from other models is Step 3, where we employ a latent variable t as an "intermediary" from WordTopic domain to DocTopic domain. In the transition stage, we introduce a transition parameter η to express the relation between WordTopic and DocTopic.

In other words, the generation for DocTopic is equivalent to drawing it from $\theta\eta$ and thus η serves as a transition matrix from θ to a "spurious" underlying mixed DocTopic distribution θ'. More explicitly, given WordTopic k the value of $\eta_{kk'}$ indicates the probability for generating DocTopic k', i.e. $p(z' = k'|z = k) = \eta_{kk'}$. With that in mind, we can see how η works on transforming WordTopic domain into DocTopic domain.

IV. Model Learning

To learn MHT, we resort to the variational EM inference framework. For each document v_i, we first decompose its corresponding log-likelihood as follows:

$$\ln p(\mathbf{w_i}, \mathbf{y_i}|\alpha, \eta, \beta, \Omega) = \mathcal{L}_i(q) + \mathrm{KL}_i(q\|p), \tag{1}$$

where q is the variational distribution used to approximate posterior distribution $p(\theta_i, z_i, t_i, z'_i|\mathbf{w_i}, \mathbf{y_i})$. Specifically,

$$\begin{aligned}\mathcal{L}_i(q) = \int_{\theta_i} \sum_{z_i} \sum_{t_i} \sum_{z'_i} q(\theta_i, z_i, t_i, z'_i) \\ \times \ln\left\{\frac{p(\mathbf{w_i}, \mathbf{y_i}, \theta_i, z_i, t_i, z'_i|\alpha, \eta, \beta, \Omega)}{q(\theta_i, z_i, t_i, z'_i)}\right\}.\end{aligned} \tag{2}$$

$$\begin{aligned}\mathrm{KL}_i(q\|p) = -\int_{\theta_i} \sum_{z_i} \sum_{t_i} \sum_{z'_i} q(\theta_i, z_i, t_i, z'_i) \\ \times \ln\left\{\frac{p(\theta_i, z_i, t_i, z'_i|\mathbf{w_i}, \mathbf{y_i}, \alpha, \eta, \beta, \Omega)}{q(\theta_i, z_i, t_i, z'_i)}\right\}.\end{aligned} \tag{3}$$

Here $\mathcal{L}$ is the lower bound of log-likelihood, and $\mathrm{KL}_(q\|p)$ is the KL-divergence between distribution q and p.

Furthermore, we construct q by introducing several free variational parameters as follows:

$$\begin{aligned}q(\theta_i, z_i, t_i, z'_i) &= q(\theta_i, z_i, t_i, z'_i|\gamma_i, \phi_i, \lambda_i, \sigma_i) \\ &= q(\theta_i|\gamma_i) \prod_{n=1}^{N_i} q(z_{in}|\phi_{in}) \\ &\times \prod_{l=1}^{L_i} q(t_{il}|\lambda_{il}) \prod_{l=1}^{L_i} q(z'_{il}|\sigma_{il}),\end{aligned} \tag{4}$$

where $q(\theta_i|\gamma_i)$ is Dirichlet distribution and $q(z_{in}|\phi_{in})$, $q(t_{il}|\lambda_{il})$ and $q(z'_{il}|\sigma_{il})$ are all multinomial distributions.

Finally, for the entire text network, we have the total log-likelihood as follows:

$$\ln p(\mathbf{w}, \mathbf{y}|\alpha, \eta, \beta, \Omega) = \sum_{i=1}^{D} \ln p(\mathbf{w_i}, \mathbf{y_i}|\alpha, \eta, \beta, \Omega) \tag{5}$$

In the E-step, we update γ, ϕ, λ and σ iteratively to approximate the posterior distribution. Then, in the M-step, α, β, η and Ω are renewed to maximize $\mathcal{L}$. Due to limitation of space, we only provide the updating equations here, and more details about the standard variational EM derivation can be referred to in [22].

$$\phi_{ink} \propto \beta_{kw_{in}} \exp(\Psi(\gamma_{ik})). \tag{6}$$

$$\gamma_{ik} = \alpha_k + \sum_{n=1}^{N_i} \phi_{ink} + \sum_{l=1}^{L_i} \lambda_{ilk}. \tag{7}$$

$$\lambda_{ilk} \propto \exp(\Psi(\gamma_{ik}) + \sum_{k'=1}^{K_y} \sigma_{ilk'} \log \eta_{kk'}). \tag{8}$$

$$\sigma_{ilk'} \propto \Omega_{k'y_{il}} \exp(\sum_{k=1}^{K_w} \lambda_{ilk} \log \eta_{kk'}). \tag{9}$$

$$\beta_{kx} \propto \sum_{i=1}^{D} \sum_{n=1}^{N_i} w_{in}^x \phi_{ink}. \tag{10}$$

$$\eta_{kk'} \propto \sum_{i=1}^{D} \sum_{l=1}^{L_i} \sigma_{ilk'} \lambda_{ilk}. \tag{11}$$

$$\Omega_{k'd} \propto \sum_{i=1}^{D} \sum_{l=1}^{L_i} y_{il}^d \sigma_{ilk'}. \tag{12}$$

Here, $\Psi(\cdot)$ is the digmma function, $w_{in}^x = 1$ if $w_{in} = x$, and 0 otherwise. Likewise, $y_{il}^d = 1$ if $y_{il} = d$, and 0 otherwise. α is updated by Newton-Raphson algorithm, the interested readers may refer to [22].

First, for each document, we execute step (6) through (9) iteratively until convergence. Next, we update α, β, η and Ω . The whole process is in an outer loop until the lower bound $\mathcal{L}$ converges.

V. Experiments

In this section, we demonstrate how our proposed system – TopicAtlas effectively explores text networks. We first describe the experiment setups such as dataset selection and parameter settings. Then, we show how to construct the heterogeneous topic web for TopicAtlas, and present some qualitative analysis of the constructed network. Finally, we validate the effectiveness of MHT, the backbone method for TopicAtlas, as a topic model for text network. Compared with some representative baseline methods, MHT achieves the best averaged performance in terms of topic interpretability and generalizability. For repeatability, the codes, datasets, results and the demo TopicAtlas are available to the public[1].

A. Datasets

We use the following two datasets in our experiments:

ACL Anthology Network (AAN). AAN [25] is a public scientific literature dataset with $20,989$ papers and $125,934$ citations. This is a quite "dense" dataset since each document has approximately 6 links on average. Furthermore, the abstract of each paper is available and we take them as the major textual contents. The papers in AAN dataset are mainly in the Natural Language Processing (NLP) field and thus the detected topics tend to be more specific.

CiteseerX. CiteseerX[2] is a well-known scientific literature digital library that primarily focuses on the literature in

[1] http://tinyurl.com/TopicAtlas

[2] http://citeseer.ist.psu.edu/oai.html

computer and information science. We collect a subset of CiteseerX dataset, which includes the abstracts of $716,800$ documents and $1,760,574$ links. Compared with AAN dataset, each document in CiteseerX only has roughly 2.5 links on average, which suggests that the citation network in CiteseerX is much sparser than that in AAN. Since CiteseerX contains a broader range of subjects, we find those identified topics are quite distinct from each other.

B. Parameter Setting

On the task of exploring heterogeneous topic web, we first need to select a reasonable topic number, which is a non-trivial task in topic models. To achieve this, we first preprocess the data using classical LDA model with varying topic numbers and evaluate the topic interpretability in terms of the topic coherence score [26]. Among the candidate topic numbers 50, 70, 90, 110, 130, and 150, topic number 70 leads to the highest topic coherence score for both AAN and CiteseerX. For simplicity, we set the topic number of WordTopic and DocTopic equal. Therefore, we implement MHT with 70 WordTopics and 70 DocTopics to explore the text networks in these two datasets. In addition, we follow the convention of [27] and initialize $\alpha = 0.01$. The parameters η, β and Ω are randomly initialized since we do not have any prior knowledge.

Furthermore, as discussed above, we use variational EM inference to learn the parameters in MHT. For standard variational EM inference framework, there are two iterative processes: the inner variational inference and the outer EM loop. Therefore, it is of great importance to determine the termination conditions for the two loops. In our experiments, for both datasets the inner variational inference loop terminates when the fractional increase of the lower bound of log-likelihood is less than 10^{-9} in two successive iterations, or when the number of iterations exceeds the maximum inner iteration number which we set to be 100. For the outer EM loop, we stop it when the relative increment ratio is less than 10^{-4}, or when the number of iterations exceeds the maximum outer iteration number which we set to be 50.

C. Heterogeneous Topic Web Construction

Here we show how to construct the heterogeneous topic web step-by-step. Recall that we aim to quantify the relations in heterogeneous topic web, which exist between WordTopic and WordTopic (Word-Word relation), DocTopic and DocTopic (Doc-Doc relation), and WordTopic and DocTopic (Word-Doc relation). We use co-occurrence probability to quantify the strength of the three types of relations, and thus our goal is to figure out $p(z = k_1, z = k_2|D)$, $p(z' = k'_1, z' = k'_2|D)$ and $p(z = k, z' = k'|D)$.

Word-Word Relation Strength. In MHT, we assume the generation of WordTopics are independent with each other when the document v is given. Therefore, the Word-Word relation strength can be calculated as follows:

$$\begin{aligned} p(z = k_1, z = k_2|D) = \sum_{z'} \sum_{i} p(z'|D) p(v_i|z'; D) \\ \times p(z = k_1|v_i; D) \\ \times p(z = k_2|v_i; D), \end{aligned} \tag{13}$$

where $p(z|v; D)$ and $p(v|z'; D)$ can be obtained from θ and Ω respectively. Posterior expectation of θ is given by:

$$\theta_{ik} = \frac{\#(v = i, z = k) + \alpha_k}{\sum_{k=1}^{K_w} (\#(v = i, z = k) + \alpha_k)}, \tag{14}$$

where $\#(v = i, z = k)$ represents the number of words assigned with WordTopic k in document v_i and the assignment can be obtained from ϕ. K_w is the number of WordTopics.

In addition, the empirical posterior distribution over DocTopics can be computed as:

$$p(z' = k'|D) = \frac{\#(z' = k')}{\sum_{k'} \#(z' = k')}, \tag{15}$$

where $\#(z' = k')$ represents the number of documents assigned with DocTopic k' in the dataset and the assignment can be obtained from σ.

Doc-Doc Relation Strength. Based on the assumption that DocTopics are generated independently given a WordTopic, we can compute Doc-Doc relation strength as:

$$\begin{aligned} p(z' = k'_1, z' = k'_2|D) = \sum_{z} p(z|D) p(z' = k'_1|z; D) \\ \times p(z' = k'_2|z; D). \end{aligned} \tag{16}$$

Similarly, η represents $p(z'|z)$ and the empirical posterior distribution over word topics is given by:

$$p(z = k|D) = \frac{\#(z = k)}{\sum_k \#(z = k)}. \tag{17}$$

Word-Doc Relation Strength. Word-Doc relation strength can be easily computed by Bayes' theorem:

$$p(z = k, z' = k'|D) = p(z' = k'|z = k; D) p(z = k). \tag{18}$$

Summarizing DocTopic. While top words are able to represent WordTopic explicitly, on the document side there are only distributions over documents to express DocTopics. However, generally it would be preferable to summarize topics with a few words. With that in mind, we leverage words in abstract to discriminate different DocTopics. Specifically, for a given DocTopic k', we compute the expectancy of word w as:

$$\mathbb{E}(w|z' = k') = \sum_{d=1}^{D} \Omega_{k'd} \cdot \#(w, d). \tag{19}$$

Then the words with high expectancy are selected as indicative words of this DocTopic.

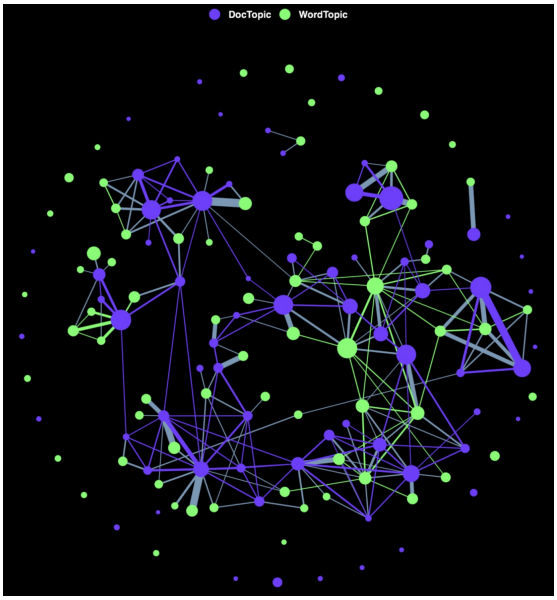


Fig. 3. An overview of TopicAtlas. Different colors indicate different types of topics, and the node size expresses the dominance of corresponding topic. Thickness of edges is proportionate to relation strength (best seen in color).

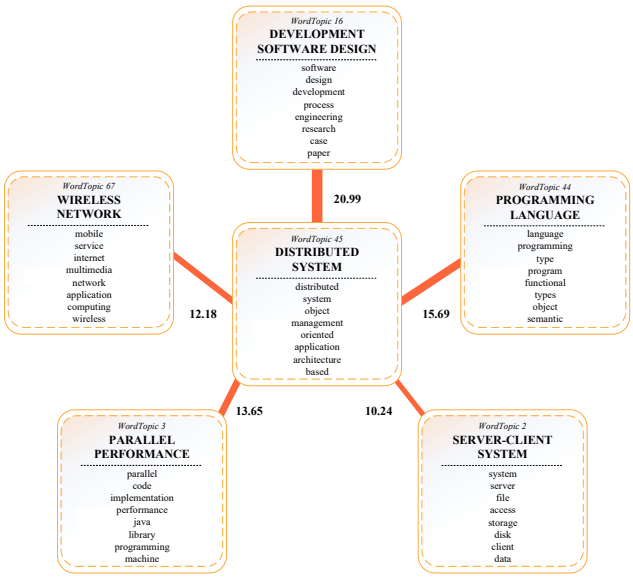


Fig. 4. "*Distributed system*" example of Word-Word subgraph. These topics are labeled manually. The weights of edges are the ratio of the Word-Word relation strength we calculate to the prior probability of a random edge (0.0002). The thickness of the edges is roughly proportionate to these values.

D. TopicAtlas

We design TopicAtlas based on the constructed heterogeneous topic web to exhibit WordTopic, DocTopic and the relationship among them. An overview of TopicAtlas is displayed in Fig.3. Aiming to help users navigate in an unfamiliar text network, TopicAtlas has the following primary features:

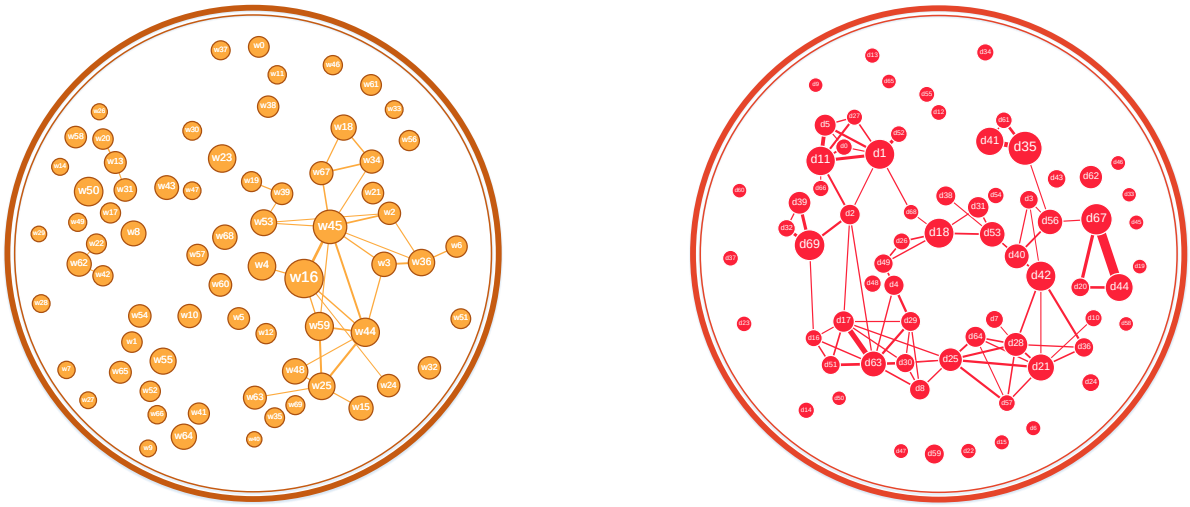
1) **Topic Landscape Exhibition.** We display top 10 keywords for each WordTopic and titles of top 5 representative documents for each DocTopic. In addition, different diameters of topic vertices express their corresponding *topic dominance* or *topic importance*, which is indicated by $p(z)$ for each WordTopic and $p(z')$ for each DocTopic.
2) **Accurate Relationship.** The three types of relations mentioned before correspond to three types of edges in the graph. The weights of these edges are the ratio of the co-occurrence probability we calculate to the prior probability of a random edge (0.0002). The thickness of the edges is proportionate to these values. Although the graph is fully connected, we removed some edges whose weights are negligible.
3) **User Friendliness.** TopicAtlas also allows users to zoom in to inspect details, or zoom out to see the big picture. For more details readers can refer to our public material.

E. Text Network Exploration via Heterogeneous Topic Web

As mentioned above, TopicAtlas is fairly comprehensive and informative. In this part, we engage in an in-depth exploration of the heterogeneous topic web in TopicAtlas and conduct qualitative analysis to illustrate how TopicAtlas assists in understanding large text network. To facilitate the analytic reasoning for three types of relations in this heterogeneous web, three auxiliary subgraphs of TopicAtlas are presented here: *Word-Word* subgraph, *Doc-Doc* subgraph and *Word-Doc* subgraph. As the name suggests, Word-Word subgraph only includes the edges between WordTopics, Doc-Doc subgraph contains merely the edges between DocTopics, and Word-Doc subgraph displays edges between WordTopics and DocTopics. Due to the limitation of space, we only give analysis for CiteseerX here. The TopicAtlas demos for both CiteseerX and AAN are available in our public website.

*1) **Word-Word Relation**:* The Word-Word subgraph is shown in Fig.5a. It is quite sparse and 62.87% of WordTopic nodes have no connection with others. It implies that the content of an individual paper is relatively "pure" and one paper mainly focuses on one WordTopic, leading to an unapparent co-occurrence pattern between WordTopics. This result agrees with our intuition: most of high quality scientific papers show clear themes.

Though the graph is sparse on the whole, there are still a few nodes of high degrees worth investigating. On the basis of previous recognition that the content of documents is generally "pure", we believe that those WordTopics which enjoy high co-occurrence probability with various other WordTopics are foundation of certain scientific fields and thus hold strong connection with many other WordTopics. In Fig.5a, WordTopic w45 (degree: 9), w44(degree: 6), w16 (degree: 5), and w25 (degree: 5) have the highest degrees. The corresponding WordTopics are "*distributed system*", "*programming language*" , "*software design*", and "*semantic reasoning*". We mark these topics as *FoundationTopic*. Obviously they are all general and basic. Take "*distributed system*" as an example, distributed system achieves efficiency improvement of solving computational



(a) Word-Word subgraph (b) Doc-Doc subgraph

Fig. 5. Subgraphs of heterogeneous topic web: (a) Word-Word Subgraph and (b) Doc-Doc Subgraph (best seen in color).

problems and therefore has broad applications in different fields such as telephone networks, routing algorithms, network file system etc. Interestingly, FoundationTopics we detect are also dominant WordTopics in the dataset. It is reasonable because general theory and basic tool gains more popularity than topics of a specific and narrow domain. As a case study, we show WordTopic w45 and its related WordTopics in Fig.4.

Based on the discussion above, the Word-Word graph is not only useful to explore the relationship between WordTopics, but also helps locate dominant WordTopics and Foundation-Topics for a novice.

*2) **Doc-Doc Relation**:* We show the Doc-Doc subgraph in Fig.5b. Compared with Word-Word subgraph, the Doc-Doc subgraph is densely connected. The close relation between DocTopics indicates that different from concentrating on one topic when writing word part of papers, authors tend to cover multiple DocTopics in the reference list. It is intuitive because a comprehensive reference section is desired for most authors, and this leads to a close connection between DocTopics. Furthermore, since ubiquitous techniques are likely to be cited in a variety of distinct domains, we expect nodes with high degrees in the Doc-Doc subgraph represent DocTopics about universal principle and method. In Fig.5b, the top four highest-degree nodes are DocTopic d63 (degree: 11), d28 (degree:7), d21 (degree:7), d17 (degree:7) and they represent "*linear system method*", "*logic programming*", "*model checking*" and "*conservation law*" respectively. Unsurprisingly, these Doc-Topics are basic techniques and laws. Hence we name them *BasicMethodTopic*.

In addition to examining DocTopics from a global perspective, inspecting details of specific DocTopic provides an insight into a text network on a document level. The DocTopic enables us to assess topic-aware impact of papers and recognize authoritative documents in one field. In our framework, documents are clustered with respect to their co-occurrence pattern and the document with more links is more likely to be assigned with a higher probability. Therefore, the top documents in one given DocTopic are generally the most popular and representative ones. In academic citation network, these top documents are supposed to be the most influential papers in a field with high citation numbers. In Fig.6 we list top 5 documents in the most dominant DocTopic d35 and its neighbours d41, d61, d56.

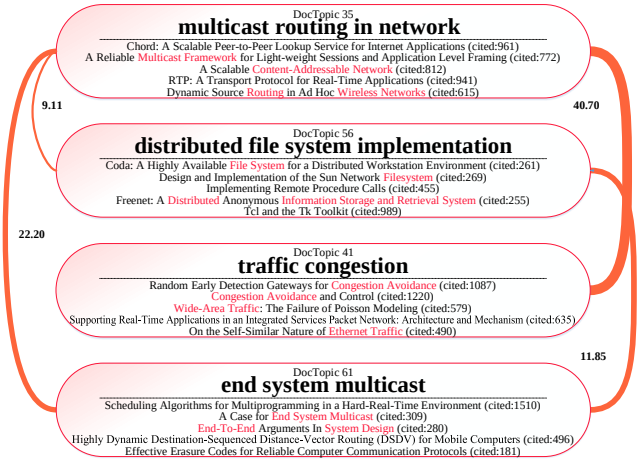


Fig. 6. "Multicast routing in network" example of Doc-Doc subgraph. These topics are labeled manually. The weights of edges are the ratio of the Doc-Doc relation strength we calculate to the prior probability of a random edge (0.0002). The thickness of edges is roughly proportionate to these values. For each document, we display its citation number in our dataset.

*3) **Word-Doc Relation**:* We believe that Word-Doc relation is the most important type of relationship since it builds a bridge between words and documents and delivers multi-aspect messages of the text network. We summarize the contributions of Word-Doc relation from three perspectives.

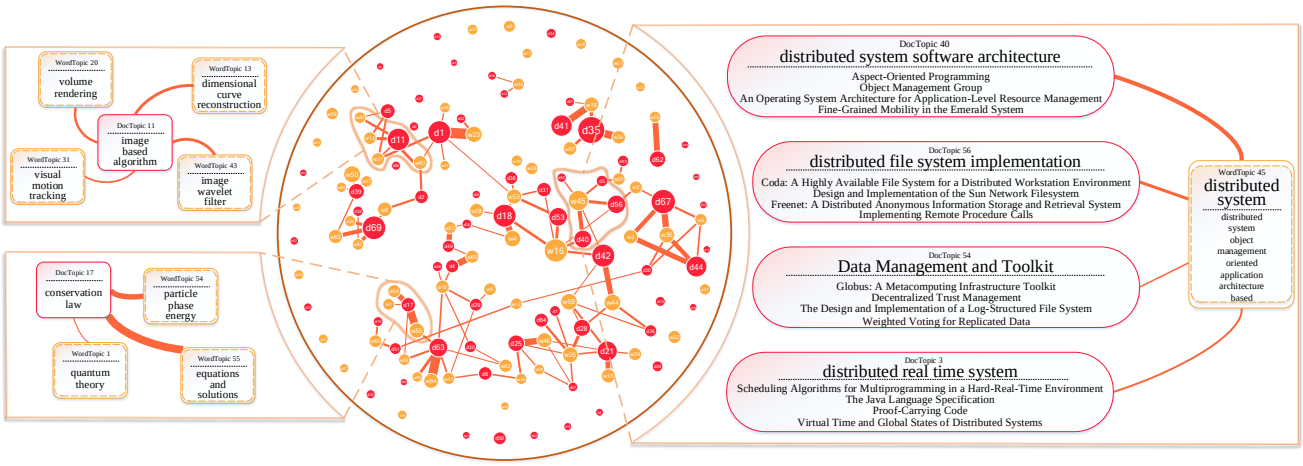


Fig. 7. Word-Doc subgraph and some instances. Red nodes (d) represent DocTopics and orange nodes (w) indicate WordTopics (best seen in color).

The Word-Doc subgraph and these examples are illustrated in Fig.7.

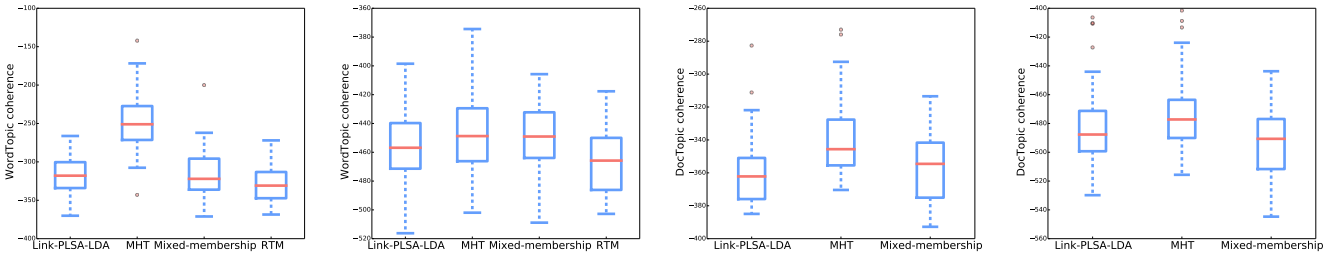
Connect WordTopic and DocTopic reasonably. As Fig.7 suggests, the BaseMethodTopic d17 is about "*conservation law*", and its neighbouring WordTopics are w54 "*particle phase energy*", w1 "*quantum theory*" and w55 "*equations and solutions*". These topics cover some basic components of quantum mechanics. In addition, WordTopic w36 is about "*shared memory processor*", and it has a strong link with DocTopic d44 "*shared memory system*" and d67 "*cache performance*". Also, it connects with DocTopic d20 "*power analysis of design*" through a edge weighting about 15 since energy reduction plays an important role in shared memory processor. Besides, WordTopic w57 "*mobile robot navigation*" is connected with DocTopic d49 "*mobile robot localization*" and d26 "*motion planning*". These connections expose the main structure of "*mobile navigation*". There are a lot of other examples in our heterogeneous topic web, readers can check them in our demo TopicAtlas.

Link WordTopics indirectly. As mentioned before, Word-Word subgraph is relatively sparse since the content of most papers only focuses on one or two topics. The missing co-occurrence phenomenon results in difficulty in spotting relevant WordTopics. However, DocTopics can serve as intermediaries between WordTopics and uncover the hidden relationship. More specifically, if two WordTopics co-occur frequently with the same DocTopic, then we can confidently say the two WordTopic are related. For example, WordTopic w13 "*dimensional curve reconstruction*", w20 "*volume rendering*" and w31 "*visual motion tracking*" are connected together in Word-Word subgraph Fig.5a. There is no edge between WordTopic w43 "*image wavelet filter*" and them, though many volume rendering and visual motion tracking models are wavelet-based. Nonetheless, DocTopic d11 "*image based algorithm*" completes the relation information as illustrated in Fig.7: all the four WordTopics enjoy strong relation with DocTopic d11, from which we can come to the conclusion that the four WordTopics are related to each other. There are other examples: WordTopic w1 "*quantum theory*", w54 "*particle phase energy*" and w55 "*equations and solutions*" are connected through DocTopic d17 "*conservation law*", WordTopic w41 "*random number set*", w64 "*numerical method*", w66 "*matrix factorization*", w52 "*dynamical model simulation*" and w55 "*equations and solutions*" are connected by the general and dominant DocTopic d63 "*linear system algorithm*".

Locate Relevant Documents. As we stressed before, a major distinction of TopicAtlas from other topic-based exploratory method is that the web we constructed is heterogeneous and provides an insight into text network on a document level. Through identifying the important documents in a DocTopic and establishing connection between DocTopic and WordTopic, users can investigate relevant documents for WordTopics. Note that instead of simply recognizing all related documents for WordTopics, TopicAtlas organizes the relevant documents according to DocTopic and allows for inspecting them in different aspects . We give an instance in Fig.7. If a researcher aims to find relevant documents for WordTopic w45 "*distributed system*", he can locate papers about implementation of distributed file or network system in DocTopic d56, examine distributed system architecture stuff in d40, get to know some data management or toolkit documents in distributed system from d54, or explore papers about distribution application in real-time system from d3. With the relevant documents sorted, the researcher is less prone to be swamped by the flood of information.

F. Topic Modeling

Since our main objective is to obtain highly effective heterogeneous topic web, it is important to evaluate the topic interpretability and ensure that the introduction of transition matrix has not come at the expense of the predictive power and generalizability of topic model. On the other hand, as



(a) WordTopic on AAN (b) WordTopic on CiteseerX (c) DocTopic on AAN (d) DocTopic on CiteseerX

Fig. 8. Topic coherence for WordTopic and DocTopic in two datasets (higher is better)

MHT models text and links simultaneously, the appropriate baselines are models which also jointly model both text and links.

1) Comparative Methods: We compare our method MHT with mixed-membership model [9], Link-PLSA-LDA [28] and RTM [11], all of which are joint models for both text and links. Mixed membership model is an early effort in this direction and can produce both cluster of words and cluster of documents like MHT. However, it neither distinguishes the two types of clusters, nor models their relationship. Nallapati *et al.* [28] propose two well-known joint topic models Pairwise-Link-LDA and Link-PLSA-LDA. Pairwise-Link-LDA models the presence and absence of links in a pairwise manner while Link-PLSA-LDA views links as "link tokens". Since Link-PLSA-LDA outperforms Pairwise-Link-LDA with respect to heldout likelihood and recall, we only include Link-PLSA-LDA in our baseline methods. The core idea of RTM is that topic relations directly account for the presence of links. To guarantee the justness, all these models are inferred through EM algorithm and parameters are initialized with the same way as MHT.

2) Topic Interpretability: Acceptable semantic quality of topics is a fundamental requirement for topic models, hence we try to quantify how interpretable the topics are.

Metric. There are some metrics for evaluating topic interpretability such as *PMI* [29], *word intrusion* [30], and *topic coherence* [26]. We adopt *topic coherence* in our experiment. For one thing, while word intrusion needs expert annotations, topic coherence is an automated evaluation metric and does not rely on human annotators. For another, topic coherence does not reference collections outside the training data as PMI dose. Also, topic coherence is proven more closely associated with the expert annotations than PMI [26].

Letting $D(w)$ be the *document frequency* of word type w and $D(w, w')$ be *co-document frequency* of word types w and w', *topic coherence* is defined as

$$C(k; W^{(k)}) = \sum_{m=2}^{M} \sum_{n=1}^{m-1} \log \frac{D(w_m^{(k)}, w_n^{(k)}) + 1}{D(w_n^{(k)})} \tag{20}$$

where $W^{(k)} = (w_1^{(k)}, \cdots, w_M^{(k)})$ is a list of the M most probable words in topic k. In our experiment, we choose $M = 10$.

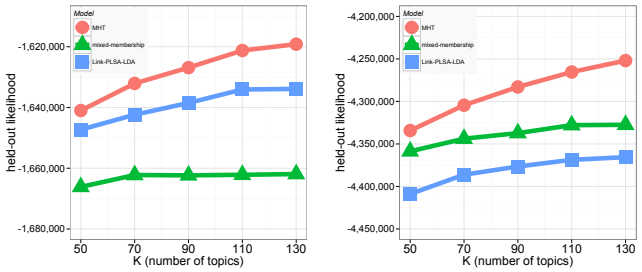
It is important to bear in mind that we have two types of topics, WordTopic and DocTopic, and topic coherence is originally designed for WordTopics. However, recall that we have obtained indicative words for each DocTopic, and these words can be regarded as a WordTopic. Intuitively, the indicative words for the documents with similar themes are more informative than that for a random set of documents, i.e. the interpretability of these words depicts the quality of corresponding DocTopic indirectly. Therefore, we compute topic coherence for these "spurious" WordTopics to evaluate DocTopics. To distinguish the two different topic coherence score, we denote them as *WordTopic coherence* and *DocTopic coherence*.

Since the heterogeneous topic web is constructed with a topic number 70, we compare the topic coherence score of different methods with 70 topics, and the result is illustrated in Fig.8. As RTM does not produce DocTopics, it is not included in DocTopic coherence comparison. The qualities of our WordTopic and DocTopic are better than baseline methods.

3) Held-Out Log Likelihood: Held-out Log Likelihood is a well-accepted metric to measure the generalizability and predictive power of topic models. When computing held-out log likelihood, we take both text and links into consideration. Since words generally dominates the data, likelihood of text is much larger. Therefore, to ease favor for text and obtain a more convincing and reasonable result, we filter out the documents with less than 3 links and 8 links for AAN and CiteseerX respectively, and we get a collection of AAN with $18,000$ documents and CiteseerX with approximately $60,000$ documents.

Our experimental set-up is as follows. We randomly split data into five folds and repeat the experiment for five times, for each time we use one fold for test, four folds for training, and we report the average values in Fig.9. The performance of MHT is better than mixed-membership model and Link-PLSA-LDA. Note that we exclude RTM in this part since held-out log likelihood favors RTM significantly due to its pairwise manner. While MHT, Link-PLSA-LDA and mixed-membership model generate links with respect to multinomial distribution, link in RTM is a Bernoulli random variable and is modeled with binomial distribution. More specifically, if links are generated without any training stages and prior knowledge (i.e. links are generated uniformly), the probability

for generating a link in RTM (0.5) is much larger than other models (0.00006 in AAN and 0.00002 in CiteseerX).



(a) AAN (b) CiteseerX

Fig. 9. Held-out log likelihood for both text and links on two datasets. (higher is better)

VI. Conclusion

In this paper, we present *MHT*, short for *Model for Heterogeneous Topic web*, a unified generative model tying together text, links as well as two types of topics, namely WordTopic and DocTopic. The relationships between WordTopic and WordTopic (Word-Word relation), DocTopic and DocTopic (Doc-Doc relation) and WordTopic and DocTopic (Word-Doc relation) are quantified through MHT, based on which we construct the heterogeneous web of topics to explore text network. In experiment, we construct the heterogeneous topic web of AAN and CiteseerX collection and build a prototype demo system, called *TopicAtlas* to exhibit the heterogeneous topic web and assist users' exploration. Qualitative analysis of the heterogeneous topic web is presented to demonstrate the effectiveness of TopicAtlas. Besides, we show that MHT outperforms existing methods as a topic model with respect to topic interpretability and held-out log likelihood. For future work, we plan to perfect TopicAtlas by displaying topic evolution pattern simultaneously and recommending documents automatically.

References

[1] P. Jahnichen, P. Oesterling, G. Heyer, T. Liebmann, G. Scheuermann, and C. Kuras, "Exploratory search through visual analysis of topic models," *Digital Humanities Quarterly (special issue)*, 2015.

[2] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant *et al.*, "Quantitative analysis of culture using millions of digitized books," *science*, vol. 331, no. 6014, pp. 176–182, 2011.

[3] G. Marchionini, "Exploratory search: From finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1121949.1121979

[4] L. F. Klein, J. Eisenstein, and I. Sun, "Exploratory thematic analysis for digitized archival collections," *Digital Scholarship in the Humanities*, p. fqv052, 2015.

[5] S. Sinclair, "Computer-assisted reading: Reconceiving text analysis," *Literary and Linguistic Computing*, vol. 18, no. 2, pp. 175–184, 2003.

[6] B. Gretarsson, J. Odonovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 23, 2012.

[7] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher, "Serendip: Topic model-driven visual exploration of text corpora," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 2014, pp. 173–182.

[8] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[9] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5220–5227, 2004.

[10] X. Wang, C. Zhai, and D. Roth, "Understanding evolution of research themes: a probabilistic generative model for citations," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1115–1123.

[11] J. Chang and D. M. Blei, "Relational topic models for document networks," in *International conference on artificial intelligence and statistics*, 2009, pp. 81–88.

[12] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 542–550.

[13] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[14] R. Nallapati, D. A. Mcfarland, and C. D. Manning, "Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents." in *AISTATS*, 2011, pp. 543–551.

[15] C. Wang, J. Liu, N. Desai, M. Danilevsky, and J. Han, "Constructing topical hierarchies in heterogeneous information networks," *Knowledge and Information Systems*, vol. 44, no. 3, pp. 529–558, 2015.

[16] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: how can citations help?" in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 957–966.

[17] L. Weng and T. M. Lento, "Topic-based clusters in egocentric networks on facebook." in *ICWSM*, 2014.

[18] M. J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi, "The topic browser: An interactive tool for browsing topic models," in *NIPS Workshop on Challenges of Data Visualization*, vol. 2, 2010.

[19] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 2012, pp. 74–77.

[20] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models." in *ICWSM*, 2012.

[21] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[23] A. S. Maiya and R. M. Rolfe, "Topic similarity networks: visual analytics for large document sets," in *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014, pp. 364–372.

[24] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 113–120.

[25] D. R. Radev, P. Muthukrishnan, and V. Qazvinian, "The acl anthology network corpus," in *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Association for Computational Linguistics, 2009, pp. 54–61.

[26] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[28] R. Nallapati and W. W. Cohen, "Link-plsa-lda: A new unsupervised model for topics and influence of blogs." in *ICWSM*, 2008.

[29] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 100–108.

[30] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.