The Annals of Applied Statistics
2016, Vol. 10, No. 1, 1–31
DOI: 10.1214/15-AOAS887

TOPIC-ADJUSTED VISIBILITY METRIC FOR SCIENTIFIC ARTICLES

BY LINDA S. L. TAN*,†,[1], AIK HUI CHAN† AND TIAN ZHENG*,[2]

Columbia University and National University of Singapore†*

Measuring the impact of scientific articles is important for evaluating the research output of individual scientists, academic institutions and journals. While citations are raw data for constructing impact measures, there exist biases and potential issues if factors affecting citation patterns are not properly accounted for. In this work, we address the problem of field variation and introduce an article level metric useful for evaluating individual articles' visibility. This measure derives from joint probabilistic modeling of the content in the articles and the citations among them using latent Dirichlet allocation (LDA) and the mixed membership stochastic blockmodel (MMSB). Our proposed model provides a visibility metric for individual articles adjusted for field variation in citation rates, a structural understanding of citation behavior in different fields, and article recommendations which take into account article visibility and citation patterns. We develop an efficient algorithm for model fitting using variational methods. To scale up to large networks, we develop an online variant using stochastic gradient methods and case-control likelihood approximation. We apply our methods to the benchmark KDD Cup 2003 dataset with approximately 30,000 high energy physics papers.

1. Introduction. Measuring the impact and influence of scientific articles is important for evaluating the work of individual scientists [Abramo and D'Angelo (2011), Hirsch (2005)] and comparing journals [Garfield (2006), Moed (2010)]. For researchers, such information is one of the most considered factors for hiring, promotion, funding decisions, award consideration and professional recognition. For academic journals, it is an indicator of a journal's stature among its peers, which is valuable for various reasons, from being considered by prospective authors for paper submission to being sought after by readers who need authoritative opinions on a topic.

Due to the lack of a unified definition, the quality and importance of scientific articles are often judged based on the journal in which they are published [Simons (2008)]. In particular, a journal's *impact factor* [see Garfield (2006)], defined using a journal's average number of citations per article, is frequently used as an indicator of the "quality" of its articles and a means of evaluating the research output of

Received June 2015; revised October 2015.

[1]Supported by the National University of Singapore Overseas Postdoctoral Fellowship.

[2]Supported in part by NSF Grant SES-1023176.

Key words and phrases. Article level metric, citation network models, stochastic blockmodels, variational Bayes, stochastic variational inference.

individuals and institutions [Casadevall and Fang (2014)]. However, studies have shown that such usage can be misleading; the journal impact factor conceals differences in citation rates among articles, is research field dependent and does not measure the scientific quality of individual articles [Seglen (1997)]. To improve the assessment of scientific research by academic institutions, funding agencies and other parties, the *San Francisco declaration on research assessment*[3] recommends (among other proposals) placing greater emphasis on the scientific content of an article rather than the journal impact factor [see Alberts (2013)]. An increasing number of publishers and organizations are also providing article level metrics [Fenner (2014), Neylon and Wu (2009)] to enable users to gauge the impact of articles based on their own merits. These new indicators include data on usage activity, bookmarks (e.g., CiteULike and Mendeley) and discussions/recommendations on the Social Web (e.g., Twitter, Facebook, Blogs) in addition to citations.

Citations (and other reference counts alike) are raw data for constructing measures to evaluate the impact of scientific articles. The h-index [Hirsch (2005)], for instance, attempts to measure the impact of an author's published work using citations (a researcher who has published h papers each having at least h citations has index h). However, there exist biases and potential issues in using raw citations to compare the impact of scientific articles without accounting for other factors which may affect citation patterns. These factors include time from publication, journal profile, article type and social network of authors [Bornmann and Daniel (2008)]. A well-known and highly relevant factor is the variation in citation practices among different disciplines [Garfield (1979)]. Articles in certain disciplines (e.g., Social Science and Mathematics) are typically much less cited than others (e.g., Molecular Biology and Immunology) and comparing articles using raw citation counts would be inappropriate. To address this issue, different procedures of normalizing the citation counts with respect to some reference standard have been proposed [e.g., Radicchi, Fortunato and Castellano (2008), Schubert and Braun (1996), Vinkler (2003)]. More recently, Crespo, Li and Ruiz-Castillo (2013) and Crespo et al. (2013) consider a model where the number of citations received by an article depends on the subfield to which the article belongs and the scientific influence of the article in the subfield. Their model assumes that citation impact varies monotonically with scientific influence.

In this work, we introduce an article level measure (of citation likelihood) that accounts for the variation in citation practices in different fields and is potentially useful for evaluating the impact of scientific articles. This measure, named as *topic-adjusted visibility metric*, derives from joint probabilistic modeling of the content (text) in the articles and the citations (links) among them. We consider a framework whereby the connectivity of an article in a citation network depends on (1) the citation probability of the research fields (topics) that it belongs to and (2) its *visibility*

[3]Outcome of a gathering of scientists at the Annual meeting of the American Society for Cell Biology on December 2012.

to articles that are in a position to cite it. Our motivation is that while a citation is driven primarily by compatibility in research topics, the decision to cite an article over other equally relevant ones may be due to a complex mixture of attributes of the selected article which are unobserved/hard to quantify (e.g., research value and quality, profile of authors, journal readership) or difficult to model directly (e.g., time since publication, article type). Here we use the term *visibility* to capture collectively attributes of the article apart from research topics that accounts for its connectivity. In our model, the topics are discovered using only the text and connectivity information. It does not take into account the discipline classification of the article by the journal. We also use the term "field" to refer to a particular topic (area).

As citation networks are a type of relational data where content information is available on individual nodes, our proposed model combines two well-established models (for text and relational data, resp.): latent Dirichlet allocation (LDA [Blei, Ng and Jordan (2003)]) and the mixed membership stochastic blockmodel (MMSB [Airoldi et al. (2008)]). LDA is a generative probabilistic model which can uncover research topics from the text of scientific articles, while the MMSB can detect communities within the citation network and model inter- and intra-community citation probabilities. As the communities detected in citation networks often correlate well with research topics [Chen and Redner (2010)], these two models can be integrated by identifying the communities in MMSB with the topics in LDA (Pairwise-Link-LDA [Nallapati et al. (2008)]). We further introduce a latent variable at the article level into the MMSB, which scales the probability of a citation due to compatibility in research topics and acts as a measure of the visibility of individual articles. The proposed model provides a structural understanding of the field variation in citation behavior and a measure of visibility for individual articles adjusted for citation probabilities within/between topics.

Our model can also provide article recommendations which take into account individual articles' visibility and citation patterns across different topics. Consider a scenario where one is searching for papers on a computational technique applied in multiple topic areas by using keywords. A method which sorts relevant articles by citation counts may yield a list where papers in topics with higher citation rates are overrepresented at the top. We avoid this scenario, as articles are recommended based on the citation probability within/across topics as well as the visibility metric which has adjusted for field variation in citation behavior. Hence, high impact articles in topics with low citation rates will not be overlooked. As the MMSB is able to capture both inter- and intra-topic citation probabilities, relevant articles which integrate multiple topics can also be identified.

The proposed topic-adjusted visibility metric is novel and differs from approaches based on normalization of citation counts. While similar in motivation with Crespo, Li and Ruiz-Castillo (2013) and Crespo et al. (2013), our model is significantly different from theirs. First, they do not consider the text of articles and make use of an external system provided by Thomson Reuters for classification

(which may be limited in range), while we identify research topics in the articles using LDA and MMSB jointly. Moreover, our model does not assume that citation counts vary monotonically with visibility within each topic and is fully generative for text and citations. Besides citations, other reference data (e.g., usage activity) which are field dependent can also benefit from our proposed framework.

For model fitting, we adopt a Bayesian approach and develop efficient variational methods [Jordan et al. (1999)] for fast approximate posterior inference. As real-world citation networks are often massive and the computational cost of analyzing every pairwise interaction in the MMSB scales as the square of the number of nodes, we develop an online variant of our variational algorithm by subsampling the full network using case-control likelihood approximation techniques [Raftery et al. (2012)] and stochastic variational inference [Hoffman et al. (2013)]. Previously, stochastic variational inference has been employed successfully for LDA [Hoffman, Blei and Bach (2010)] and the hierarchical Dirichlet process [Wang, Paisley and Blei (2011)]. At each iteration, it subsamples the data and optimizes the variational objective using stochastic approximation methods [Robbins and Monro (1951)], thus reducing both computational and storage costs. Recently, Gopalan and Blei (2013) extended stochastic variational inference to massive networks for detecting overlapping communities by using a variant of the MMSB. They sampled node pairs using "informative set sampling," where the sets of pairs are defined using network topology information. A related idea is the stratified sampling scheme for MCMC estimation of latent space models [Raftery et al. (2012)], where strata are defined by shortest path lengths. Adapting case-control designs in epidemiology, they approximated the log-likelihood function by sampling all links for each node and only a small proportion of nonlinks from each stratum. This approach is feasible as large networks are often sparse. It assumes that "closer" nodes contain more information and are more relevant in estimating each other's latent position. Motivated by these methods, we propose a novel strategy for sampling node pairs that is suitably adapted to our model requirements.

We apply our methods to the Cora dataset with 2410 scientific publications in computer science research and the benchmark KDD Cup 2003 dataset with approximately 30,000 high energy physics papers. We also evaluate the performance of our model using a simulation study. A particularity of citation networks is that articles join the network over time, and published articles cannot cite articles appearing at a later date. Hence, the absence of such links cannot be construed as true "zeros" and should be omitted from the likelihood. We show that taking into account publication times (when available) can significantly improve performance of our model.

The rest of the paper is organized as follows. Section 2 lays out the details of our model and reviews closely associated models. Section 3 introduces a variational algorithm for obtaining approximate posterior inference and Section 4 describes how the algorithm can be scaled up to large networks using stochastic optimization methods. Section 5 discusses comparisons with alternative approaches and

Section 6 predictions and article recommendations based on our proposed model. Section 7 presents application results using simulations and real data. We conclude with discussion in Section 8.

2. Model description. Our proposed model combines LDA with the MMSB, and introduces a latent variable for each article which acts as a measure of its visibility adjusted for topic-level variation in activity level. Before describing our model, we review LDA, MMSB and other associated models.

2.1. *Review of LDA, MMSB and Pairwise-Link-LDA.* LDA is a generative probabilistic model for text copora which can be used for tasks such as detecting themes, summarization and classification. It assumes that word order can be disregarded ("bag-of-words" model) and that each document in the corpus exhibits K topics with varying proportions. Let the number of documents in the corpus be D and the size of the vocabulary be $\mathcal{V}$. Each topic β_k is a $\mathcal{V} \times 1$ vector with a Dirichlet(η) prior, representing a probability distribution over the vocabulary. For each document d, the topic proportion θ_d is a $K \times 1$ vector with a Dirichlet(α) prior, representing the probability of each topic occurring in the document. Let the number of words in document d be N_d. The nth word in document d, w_{dn}, is generated by drawing a topic assignment z_{dn} from Multinomial(θ_d) and the word from Multinomial($\beta_{z_{dn}}$). Both z_{dn} and w_{dn} are indicator vectors with a single one. If the kth element of z_{dn} is one, w_{dn} is drawn from the topic $\beta_{z_{dn}}$, which refers to β_k (slight abuse of notation).

On the other hand, MMSB is a mixed membership model for relational data that can detect communities within a network. Suppose the relational data is represented by a directed graph. For each node pair (d, d') where $d \neq d'$, we define the binary variable $y_{dd'}$ to be 1 if there is a directed edge from d to d' and 0 otherwise. The MMSB assumes that there are K latent communities (groups) and each node belongs to the K groups with varying degrees of affiliation. Specifically, each node d is associated with a $K \times 1$ membership vector, θ_d, drawn from a Dirichlet(α) prior, representing the probability of the node belonging to each of the K groups. Each node may assume different membership when interacting with different nodes. The blockmodel B is a $K \times K$ matrix where B_{ij} represents the probability of a directed link from a node in group i to a node in group j. For each (d, d'), membership indicator vectors for the *sender* ($s_{dd'}$) and *receiver* ($r_{dd'}$) are first drawn from Multinomial(θ_d) and Multinomial($\theta_{d'}$) respectively. If the ith and jth elements of $s_{dd'}$ and $r_{dd'}$ are ones respectively, the value of the interaction $y_{dd'}$ is sampled from Bernoulli($B_{s_{dd'}r_{dd'}}$), where $B_{s_{dd'}r_{dd'}}$ refers to B_{ij}. The generating process of LDA and MMSB is shown in Figure 1.

Citation networks are a type of relational data where the nodes are the documents/articles and the directed links are the citations between them (there is a directed link from d to d' if d cites d'). Pairwise-Link-LDA [Nallapati et al. (2008)]

LDA (For modeling text)

1. Draw topic $\beta_k \sim \text{Dirichlet}(\eta)$ for $k = 1, \ldots, K$.
2. For each document $d = 1, \ldots, D$:
   (a) Draw topic proportion $\theta_d \sim \text{Dirichlet}(\alpha)$.
   (b) For each position $n = 1, \ldots, N_d$, draw:
   - Topic assignment $z_{dn} \sim \text{Multinomial}(\theta_d)$.
   - Word $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$.

MMSB (For modeling links)

1. For each node $d = 1, \ldots, D$, draw mixed membership vector $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. For each node pair (d, d'), draw:
   - Membership indicator for sender $s_{dd'} \sim \text{Multinomial}(\theta_d)$.
   - Membership indicator for receiver $r_{dd'} \sim \text{Multinomial}(\theta_{d'})$.
   - Interaction $y_{dd'} \sim \text{Bernoulli}(B_{s_{dd'} r_{dd'}})$.

FIG. 1. *Generating process for LDA (left) and MMSB (right). These two models can be combined by identifying the communities in MMSB with the topics in LDA, that is, the topic proportions with the mixed membership vectors (both denoted by θ_d).*

combines MMSB with LDA to jointly model text in articles and the citations between them by identifying the topics in LDA with the communities in MMSB. As links between documents indicate a certain level of similarity in topics, it is believed that network information, when suitably incorporated, would improve topic modeling [Ho, Eisenstein and Xing (2012), Kleinberg (1999)].

2.2. *Proposed model: LMV.* In Pairwise-Link-LDA, any two documents with the same topic proportions have equal probability of being cited. This assumption is easily violated in real-world citation networks, as factors other than research topics affect the citation probability. For instance, a well-cited document's higher chances of being cited may be due to its quality, novelty and the authors' social networks. Our proposed model aims to capture, collectively via the *visibility* measure, attributes of the cited document that explain this variation in citation probability given compatibility in topics.

Our proposed model is presented in Figure 2. The text is generated as in LDA. For the generation of links, we introduce a latent variable $\tau_{d'}$ for each document d', drawn from a $\text{Beta}(g_0, h_0)$ prior, which modifies the citation probability by scaling the blockmodel. Given that the ith element of $s_{dd'}$ and the jth element of $r_{dd'}$ are ones, the probability of a document d from the ith topic citing a document d' from the jth topic becomes $\tau_{d'} B_{ij}$, which is dependent on the characteristic $\tau_{d'}$ of the document d' receiving the citation.

We define $\tau_{d'}$ as the *topic-adjusted visibility* of document d', which is a scaling factor that adjusts the universal probability of being cited based on citation probabilities within/between topics. It is a characteristic of d' that accounts for the variation in citation probability among documents with equal topic proportions. This variation might be due to attributes of d' which are unobserved, hard to quantify or not directly modeled. The topic-adjusted visibility is potentially useful as

Generative process of LMV

1. Draw topic $\beta_k \sim \text{Dirichlet}(\eta)$ for $k = 1, \ldots, K$.
2. For each document $d = 1, \ldots, D$:
   - Draw visibility $\tau_d \sim \text{Beta}(g_0, h_0)$.
   - Draw topic proportion $\theta_d \sim \text{Dirichlet}(\alpha)$.
   - For each position $n = 1, \ldots, N_d$, draw:
     - Topic assignment $z_{dn} \sim \text{Multinomial}(\theta_d)$.
     - Word $w_{dn} \sim \text{Multinomial}(\beta_{z_{dn}})$.
3. For $i, j \in \{1, \ldots, K\}$, draw $B_{ij} \sim \text{Beta}(a_0, b_0)$.
4. For each document pair (d, d') where $d \neq d'$:
   - Draw topic indicator for sender: $s_{dd'} \sim \text{Multinomial}(\theta_d)$. Receiver: $r_{dd'} \sim \text{Multinomial}(\theta_{d'})$.
   - Draw $y_{dd'} \sim \text{Bernoulli}(\tau_{d'} B_{s_{dd'} r_{dd'}})$.

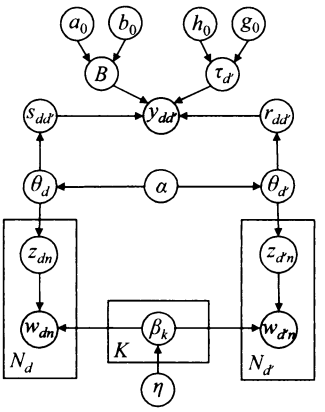


FIG. 2. *Left: Outline of LMV. Right: Graphical representation of a two-document segment. The complete model contains $y_{dd'}$ for every document pair. Circles denote variables and observed variables are shaded. The plates contain variables to be replicated and the number of times is indicated in the lower left corner.*

a descriptive article level measure of citation likelihood that adjusts for the differences in citation practices in different topic areas. We use the acronym LMV for our model, taking the first letter from LDA, MMSB and visibility.

Our model reduces to Pairwise Link-LDA [Nallapati et al. (2008)] when $\tau_{d'}$ is identically one. One computational issue associated with LDA and MMSB is the multi-modality of these models, whereby multiple model configurations give equivalent fits to the data [see, e.g., Ho, Parikh and Xing (2012)]. Combining content and connectivity information in a citation network may alleviate the multi-modality issue of MMSB. On the other hand, the communities detected in the citation network regularize the topic modeling on the content information.

We assume that the hyperparameters a_0, b_0, g_0, h_0, η and α are known. The latent variables $s_{dd'}$ and $r_{dd'}$ give rise to a more tractable joint distribution, which leads to significant simplification of the variational optimization procedure to be introduced in Section 3.

2.3. *Review of associated methods.* The relational topic model (RTM [Chang and Blei (2010)]) also uses LDA as a basis for modeling document networks. RTM does not consider every pairwise interaction, however, and models only observed links. It uses a symmetric probability function with a diagonal weight matrix, which allows only within topic interactions. To improve the RTM, extensions have been proposed. For instance, Chen et al. (2013) define generalized RTM with a full weight matrix and perform regularized Bayesian inference where a log-logistic

loss is minimized. A regularization parameter is used to control influence from link structures. Zhang, Zhu and Zhang (2013) propose sparse RTM where normal and Laplace priors are placed respectively on the topic and word representations, and word counts are Poisson distributed. Different regularization parameters are used for links and nonlinks in the minimization of a log-loss.

Some other extensions of LDA to document networks include the TopicBlock [Ho, Eisenstein and Xing (2012)], which uses text and links to induce a hierarchical taxonomy, and block LDA [Balasubramanyan and Cohen (2013)], which considers documents annotated with entities and models only realized links between entities using a stochastic blockmodel. Zhu et al. (2013) propose a Poisson mixed-topic link model that combines LDA with a variant of the MMSB, called the Ball–Karrer–Newman model, where the number of links between two documents is Poisson distributed instead of Bernoulli. Neiswanger et al. (2014) present a latent random offset model that augments the topic proportions of the cited document with a vector to capture contents of citing documents in link predictions. These models do not address the issue that documents with similar topics may have different connectivity due to unobserved factors of impact.

To improve the understanding of large text corpora, models that incorporate additional information about the corpus or document metadata into text analysis have also been introduced recently. These models investigate the relationship between text data and observed variables that may affect the text composition (e.g., authors, date, political affiliation and quality ratings). To overcome the difficulty of incorporating high-dimensional text data into statistical analyses for predicting sentiment variables, multinomial inverse regression [Taddy (2013)] uses the inverse conditional distribution of text given sentiment to obtain low-dimensional document representations that preserve sentiment information. The inverse regression topic model [Rabinovich and Blei (2014)] extends multinomial inverse regression to the mixed-membership (multiple topics) setting, while distributed multinomial regression [Taddy (2015)] tackles high-dimensional sentiment variables by modeling document-word counts as independent Poisson distributed variables. Roberts et al. (2013) propose the structural topic model, which uses generalized linear models as priors to incorporate document-level covariates. In this paper, we adopt a different approach to the exploration of large text copora by modeling text and links between documents jointly. Further incorporating document metadata into our proposed model will be an interesting direction for future research.

3. Posterior inference. As the true posterior of our model is not available in closed form, we develop an efficient variational algorithm for posterior approximation. Let Θ denote the set of unknown variables in the LMV. In variational methods, the true posterior is approximated by more tractable distributions which are optimized to be close to the true posterior in terms of Kullback–Leibler diver-

gence. Here we consider a fully-factorized family,

$$q(\Theta) = \prod_d \Big\{ q_D(\theta_d|\gamma_d) \prod_n q_M(z_{dn}|\phi_{dn}) \prod_{d' \neq d} [q_M(s_{dd'}|\kappa_{dd'}) q_M(r_{dd'}|\nu_{dd'})] \Big\}$$
$$\times \prod_k q_D(\beta_k|\lambda_k) \prod_{i,j} q_B(B_{ij}|a_{ij}, b_{ij}) \prod_{d'} q_B(\tau_{d'}|g_{d'}, h_{d'}),$$

where q_D, q_M, q_B denote the Dirichlet, multinomial and beta distributions respectively and $\{\lambda, \gamma, \phi, \kappa, \nu, a, b, g, h\}$ are variational parameters to be optimized. Discussion on assumptions made in the variational approximation and their implications can be found in the supplement [Tan, Chan and Zheng (2016)].

From Jensen's inequality, minimizing the Kullback–Leibler divergence between $q(\Theta)$ and the true posterior is equivalent to maximizing a lower bound $\mathcal{L}$ on the log marginal likelihood, where $\mathcal{L}$ is given by

$$\sum_{(d,d')} \mathrm{E}_q[\log p(y_{dd'}|\tau_{d'}, B, s_{dd'}, r_{dd'}) + \log p(s_{dd'}|\theta_d) + \log p(r_{dd'}|\theta_{d'})]$$

$$(3.1) \quad + \sum_{d,n} \mathrm{E}_q[\log p(z_{dn}|\theta_d) + \log p(w_{dn}|z_{dn}, \beta)] + \sum_{i,j} \mathrm{E}_q[\log p(B_{ij}|a_0, b_0)]$$
$$+ \sum_d \mathrm{E}_q[\log p(\tau_d|g_0, h_0) + \log p(\theta_d|\alpha)] + \sum_k \mathrm{E}_q[\log p(\beta_k|\eta)] + H(q).$$

In (3.1), E_q denotes expectation with respect to $q(\Theta)$ and $H(q)$ denotes the entropy of q. All terms in $\mathcal{L}$ can be evaluated analytically except $\mathrm{E}_q\{\log(1 - \tau_{d'} B_{ij})\}$. We expand this expectation using a first-order approximation about the mean [Braun and McAuliffe (2010)] so that

$$(3.2) \quad \begin{aligned} \mathrm{E}_q\{\log(1 - \tau_{d'} B_{ij})\} &\approx \log(1 - \mathrm{E}_q(\tau_{d'})\mathrm{E}_q(B_{ij})) \\ &= \log\left(1 - \frac{g_{d'}}{g_{d'} + h_{d'}} \frac{a_{ij}}{a_{ij} + b_{ij}}\right). \end{aligned}$$

The approximate lower bound obtained using (3.2) is denoted by $\mathcal{L}^*$. Discussion on the first-order approximation and the expression for $\mathcal{L}^*$ can be found in the supplement [Tan, Chan and Zheng (2016)].

We optimize $\mathcal{L}^*$ with respect to the variational parameters via coordinate ascent (see Algorithm 1). For $\{\lambda, \gamma, \phi, \kappa, \nu\}$, closed-form updates can be derived by differentiating $\mathcal{L}^*$ with respect to each parameter and setting the gradient to zero. For $\{a, b, g, h\}$, the likelihood is nonconjugate with respect to the prior and we use nonconjugate variational message passing [Knowles and Minka (2011)]. This is a fixed point iteration method for optimizing the natural parameters of variational posteriors in exponential families. The advantages of this approach are that it yields closed-form updates and extends to stochastic variational inference naturally. However, $\mathcal{L}^*$ is not guaranteed to increase at each step and updates for

Algorithm 1 Coordinate ascent procedure for the LMV

Initialize λ, γ, ϕ, κ, ν, a, b, g and h. Cycle the following updates until convergence is reached.

1. For each document pair (d, d'), cycle the following updates until $\kappa_{dd'}$ and $\nu_{dd'}$ converge.

$$\kappa_{dd'i} \propto \exp\left\{\psi(\gamma_{di}) - \psi\left(\sum_i \gamma_{di}\right) + \sum_j \nu_{dd'j}\varsigma_{dd'}(i,j)\right\} \qquad \text{for } i = 1,\dots,K,$$

$$\nu_{dd'j} \propto \exp\left\{\psi(\gamma_{d'j}) - \psi\left(\sum_j \gamma_{d'j}\right) + \sum_i \kappa_{dd'i}\varsigma_{dd'}(i,j)\right\} \qquad \text{for } j = 1,\dots,K,$$

where

$$\varsigma_{dd'}(i,j) = \begin{cases} \psi(a_{ij}) - \psi(a_{ij}+b_{ij}) + \psi(g_{d'}) - \psi(g_{d'}+h_{d'}), & \text{if } y_{dd'} = 1, \\ \log\left(1 - \dfrac{g_{d'}}{g_{d'}+h_{d'}}\dfrac{a_{ij}}{a_{ij}+b_{ij}}\right), & \text{if } y_{dd'} = 0. \end{cases}$$

2. For $d = 1,\dots,D$, $n = 1,\dots,N_d$, $k = 1,\dots,K$,

$$\phi_{dnk} \propto \exp\left\{\psi(\gamma_{dk}) - \psi\left(\sum_k \gamma_{dk}\right) + \sum_v w_{dnv}\left[\psi(\lambda_{kv}) - \psi\left(\sum_v \lambda_{kv}\right)\right]\right\}.$$

3. For $d = 1,\dots,D$, $\gamma_d = \alpha + \sum_n \phi_{dn} + \sum_{d' \neq d}(\kappa_{dd'} + \nu_{d'd})$.
4. For $k = 1,\dots,K$, $v = 1,\dots,V$, $\lambda_{kv} = \eta_v + \sum_d \sum_n w_{dnv}\phi_{dnk}$.
5. Cycle updates in (a) and (b) until convergence is reached.
   (a) For $i = 1,\dots,K$, $j = 1,\dots,K$, $\begin{bmatrix} a_{ij} \\ b_{ij} \end{bmatrix} \leftarrow (1 - s_t)\begin{bmatrix} a_{ij} \\ b_{ij} \end{bmatrix} + s_t\begin{bmatrix} \hat{a}_{ij} \\ \hat{b}_{ij} \end{bmatrix}$, where

$$\begin{bmatrix} \hat{a}_{ij} \\ \hat{b}_{ij} \end{bmatrix} = \begin{bmatrix} a_0 + \sum_{(d,d'):y_{dd'}=1} \kappa_{dd'i}\nu_{dd'j} \\ b_0 \end{bmatrix} + \frac{1}{|I_{a_{ij},b_{ij}}|(a_{ij}+b_{ij})^2} \times \begin{bmatrix} (a_{ij}+b_{ij})\psi'(a_{ij}+b_{ij}) - b_{ij}\psi'(b_{ij}) \\ a_{ij}\psi'(a_{ij}) - (a_{ij}+b_{ij})\psi'(a_{ij}+b_{ij}) \end{bmatrix} \sum_{(d,d'):y_{dd'}=0} \frac{\kappa_{dd'i}\nu_{dd'j}\frac{g_{d'}}{g_{d'}+h_{d'}}}{1 - \frac{g_{d'}}{g_{d'}+h_{d'}}\frac{a_{ij}}{a_{ij}+b_{ij}}}.$$

   Start with $s_t = 1$. If any $a_{ij} \leq 0$ or $b_{ij} \leq 0$, reduce s_t (say by half each time) until all $a_{ij} > 0$ and $b_{ij} > 0$. Accept update only if $\mathcal{L}^*$ increases.

   (b) For $d' = 1,\dots,D$, $\begin{bmatrix} g_{d'} \\ h_{d'} \end{bmatrix} \leftarrow (1 - s_t)\begin{bmatrix} g_{d'} \\ h_{d'} \end{bmatrix} + s_t\begin{bmatrix} \hat{g}_{d'} \\ \hat{h}_{d'} \end{bmatrix}$, where

$$\begin{bmatrix} \hat{g}_{d'} \\ \hat{h}_{d'} \end{bmatrix} = \begin{bmatrix} g_0 + \sum_d y_{dd'} \\ h_0 \end{bmatrix} + \frac{1}{|I_{g_{d'},h_{d'}}|(g_{d'}+h_{d'})^2} \times \begin{bmatrix} (g_{d'}+h_{d'})\psi'(g_{d'}+h_{d'}) - h_{d'}\psi'(h_{d'}) \\ g_{d'}\psi'(g_{d'}) - (g_{d'}+h_{d'})\psi'(g_{d'}+h_{d'}) \end{bmatrix} \sum_{i,j} \frac{\sum_{d:y_{dd'}=0}\kappa_{dd'i}\nu_{dd'j}\frac{a_{ij}}{a_{ij}+b_{ij}}}{1 - \frac{g_{d'}}{g_{d'}+h_{d'}}\frac{a_{ij}}{a_{ij}+b_{ij}}}.$$

   Start with $s_t = 1$. If any $g_{d'} \leq 0$ or $h_{d'} \leq 0$, reduce s_t (say by half each time) until all $g_{d'} > 0$ and $h_{d'} > 0$. Accept update only if $\mathcal{L}^*$ increases.

Note: $|I_{a,b}|$ denotes determinant of the Fisher information matrix of Beta(a, b). See supplement [Tan, Chan and Zheng (2016)].

$\{a, b, g, h\}$ may be negative at times. To resolve these issues, we use the fact that nonconjugate variational message passing is a natural gradient ascent method with step size 1 and smaller step sizes may also be taken. In Algorithm 1, we start with step size 1 and reduce the step size where necessary to ensure updates of $\{a, b, g, h\}$ are positive. If $\mathcal{L}^*$ increases, these updates are accepted. Otherwise, we revert to the former values. Updates for $\{a, b, g, h\}$ are derived in the supplement [Tan, Chan and Zheng (2016)]. As updates of $\{a, b\}$ and $\{g, h\}$ are highly interdependent, we introduce a nested loop for cycling these updates in step 5 of Algorithm 1.

4. Stochastic optimization of variational objective. We develop an online variant of Algorithm 1 that scales well to large networks using stochastic variational inference [Hoffman et al. (2013)]. In this approach, variational parameters are classified as *local* (specific to each node) or *global* (common across all nodes) parameters. At each iteration, a minibatch of nodes are randomly sampled from the whole dataset and local parameters corresponding to these nodes are optimized. Global parameters are then updated based on optimized local parameters using stochastic gradient ascent [Robbins and Monro (1951)]. The algorithm converges to a local maximum of the variational objective provided the step sizes and the objective function satisfy certain regularity conditions [see Spall (2003)].

Currently, Algorithm 1 has to update the variational parameters $\kappa_{dd'}$ and $\nu_{dd'}$ for each document pair (d, d') at every iteration. This computational cost scales as $\mathcal{O}(D^2)$ and makes our model infeasible for large networks. To apply stochastic variational inference, we regard $\kappa_{dd'}$ and $\nu_{dd'}$ as *local* parameters and perform these updates only for a random subset of all document pairs at each iteration. Remaining variational parameters are regarded as *global* parameters and are updated using stochastic gradient ascent.

4.1. *Proposed sampling strategy.* We devise a novel scheme for sampling document pairs. While simple random sampling is a possibility, it does not utilize information provided by the links. Raftery et al. (2012) propose a stratified sampling scheme where stratums are defined by shortest path lengths. Assuming "closer" nodes contain more information and that large networks are often sparse, they sample all links for each node and only a small proportion of nonlinks from each stratum. Gopalan and Blei (2013) consider "informative set sampling," where the "informative set" for a node consists of all links and nonlinks of path length 2. Remaining nonlinks are partitioned into "noninformative sets." At each iteration, either an "informative" set is chosen with high probability or one of the "noninformative sets" is chosen with low probability. These schemes are not directly applicable to our model. To update γ_d (Algorithm 1, step 2), unbiased estimates of $\sum_{d' \neq d} \kappa_{dd'}$ and $\sum_{d' \neq d} \nu_{d'd}$ are required. That is, samples have to be drawn from cases where d is the citing document as well as cases where d is the cited document. As Gopalan and Blei (2013) treat links as undirected while Raftery et al. (2012) do not subsample documents, they do not face these restrictions.

We propose the following sampling scheme. Consider the adjacency matrix y of ones and zeros, where the rows and columns denote the citing and cited documents respectively (diagonal is undefined). We associate each document pair (d, d') with an inclusion probability $\pi_{dd'}$, where $\pi_{dd'}$ is a decreasing function of the shortest path length from d to d'. While other definitions are plausible, we define, for simplicity,

$$\pi_{dd'} = \begin{cases} 1/l_{dd'}, & l_{dd'} > n_0, \\ 1/n_0, & \text{otherwise}, \end{cases} \tag{4.1}$$

where $l_{dd'}$ denotes the shortest path length from d to d' and n_0 is a positive integer. When $y_{dd'} = 1$, $\pi_{dd'} = 1$. Hence, all links are included and more "informative" nonlinks (in the sense of shorter path length) have a higher probability of being included. In the examples, we set $n_0 = 100$. This implies that document pairs with a shortest path length of 100 or more have a probability of 0.01 of being included. It is possible to experiment with other values depending on the application and computational constraints. A smaller n_0 implies a larger sample of document pairs. At each iteration, we:

1. Select a random sample $\mathcal{S}$ of $|\mathcal{S}|$ documents from the whole dataset.
2. Perform a Bernoulli trial with success probability $\pi_{dd'}$ for each (d, d') where $d \in \mathcal{S}$ or $d' \in \mathcal{S}$.
3. Select document pairs with successful trials (denote this set as $\mathcal{P}$).

The Bernoulli trial is performed only once for each document pair even if both d and d' are in $\mathcal{S}$. The sampling strategy is illustrated in the supplement [Tan, Chan and Zheng (2016)].

4.2. *Stochastic variational algorithm.* In the stochastic variational algorithm (Algorithm 2), updates for $\kappa_{dd'}$ and $\nu_{dd'}$ are cycled until convergence for each $(d, d') \in \mathcal{P}$ so that local parameters are optimized at the current global parameters. The global parameters are then updated using stochastic gradient ascent. For a parameter λ_i, we consider an update

$$\lambda_i \leftarrow \lambda_i + s_t \tilde{\nabla}_{\lambda_i} \mathcal{L}^*, \tag{4.2}$$

where s_t denotes a small step taken in the direction of $\tilde{\nabla}_{\lambda_i} \mathcal{L}^*$ (natural gradient of $\mathcal{L}^*$ with respect to λ_i). In variational Bayes and nonconjugate variational message passing, the natural gradient [Amari (1998)] is $\tilde{\nabla}_{\lambda_i} \mathcal{L}^* = \hat{\lambda}_i - \lambda_i$, where $\hat{\lambda}_i$ is the optimal update of λ_i. See supplement [Tan, Chan and Zheng (2016)] for details. Hence the update in (4.2) can be written as

$$\lambda_i \leftarrow (1 - s_t)\lambda_i + s_t \hat{\lambda}_i.$$

In stochastic gradient ascent, we replace the true natural gradients with unbiased estimates. For convergence, the step sizes should satisfy the conditions $s_t \to 0$, $\sum_{t=0}^{\infty} s_t = \infty$ and $\sum_{t=0}^{\infty} s_t^2 < \infty$.

Algorithm 2 Stochastic variational procedure for the LMV

Initialize $\gamma, \lambda, \kappa, \phi, \nu, a, b, g, h$. At each iteration:

1. Obtain a random sample $\mathcal{S}$ of $|\mathcal{S}|$ documents from the corpus.
2. For each document pair (d, d'), where $d \in \mathcal{S}$ or $d' \in \mathcal{S}$, perform a Bernoulli trial with success probability $\pi_{dd'}$. Let $\mathcal{P}$ denote the set of document pairs with successful trials. Let $\mathcal{P}_{d\cdot} = \{(l, l') \in \mathcal{P} | l = d\}$ and $\mathcal{P}_{\cdot d} = \{(l, l') \in \mathcal{P} | l' = d\}$.
3. Update $\kappa_{dd'}$ and $\nu_{dd'}$ iteratively as in Algorithm 1 for each $(d, d') \in \mathcal{P}$ until convergence.
4. Update ϕ_{dnk} for $d \in \mathcal{S}$, $n = 1, \ldots, N_d$ and $k = 1, \ldots, K$ as in Algorithm 1.
5. For $d \in \mathcal{S}$, $\gamma_d \leftarrow (1 - s_t)\gamma_d + s_t \hat{\gamma}_d$ where
$$\hat{\gamma}_d = \alpha + \sum_n \phi_{dn} + \sum_{(l,l') \in \mathcal{P}_{d\cdot}} \frac{\kappa_{ll'}}{\pi_{ll'}} + \sum_{(l,l') \in \mathcal{P}_{\cdot d}} \frac{\nu_{ll'}}{\pi_{ll'}}.$$
6. For $k = 1, \ldots, K$ and $v = 1, \ldots, V$, $\lambda_{kv} \leftarrow (1 - s_t)\lambda_{kv} + s_t \hat{\lambda}_{kv}$, where
$$\hat{\lambda}_{kv} = \eta_v + \frac{D}{|\mathcal{S}|} \sum_{d \in \mathcal{S}} \sum_n w_{dnv} \phi_{dnk}.$$
7. For $i = 1, \ldots, K$ and $j = 1, \ldots, K$, $\begin{bmatrix} a_{ij} \\ b_{ij} \end{bmatrix} \leftarrow (1 - s_t)\begin{bmatrix} a_{ij} \\ b_{ij} \end{bmatrix} + s_t \begin{bmatrix} \hat{a}_{ij} \\ \hat{b}_{ij} \end{bmatrix}$, where
$$\begin{bmatrix} \hat{a}_{ij} \\ \hat{b}_{ij} \end{bmatrix} = \begin{bmatrix} a_0 + \frac{D}{|\mathcal{S}|} \sum_{d' \in \mathcal{S}} \sum_{(l,l') \in \mathcal{P}_{\cdot d'} : y_{ll'} = 1} \kappa_{ll'i} \nu_{ll'j} \\ b_0 \end{bmatrix} + \frac{1}{|I_{a_{ij}, b_{ij}}|(a_{ij} + b_{ij})^2}$$
$$\times \begin{bmatrix} (a_{ij} + b_{ij})\psi'(a_{ij} + b_{ij}) - b_{ij}\psi'(b_{ij}) \\ a_{ij}\psi'(a_{ij}) - (a_{ij} + b_{ij})\psi'(a_{ij} + b_{ij}) \end{bmatrix}$$
$$\times \frac{D}{|\mathcal{S}|} \sum_{d' \in S} \frac{\sum_{(l,l') \in \mathcal{P}_{\cdot d'} : y_{ll'} = 0} \frac{\kappa_{ll'i} \nu_{ll'j}}{\pi_{ll'}} \frac{g_{d'}}{g_{d'} + h_{d'}}}{1 - \frac{g_{d'}}{g_{d'} + h_{d'}} \frac{a_{ij}}{a_{ij} + b_{ij}}}.$$
If any $a_{ij} \leq 0$ or $b_{ij} \leq 0$, reduce s_t for this update (say by half each time).
8. For $d' \in \mathcal{S}$, $\begin{bmatrix} g_{d'} \\ h_{d'} \end{bmatrix} \leftarrow (1 - s_t)\begin{bmatrix} g_{d'} \\ h_{d'} \end{bmatrix} + s_t \begin{bmatrix} \hat{g}_{d'} \\ \hat{h}_{d'} \end{bmatrix}$, where
$$\begin{bmatrix} \hat{g}_{d'} \\ \hat{h}_{d'} \end{bmatrix} = \begin{bmatrix} g_0 + \sum_d y_{dd'} \\ h_0 \end{bmatrix} + \frac{1}{|I_{g_{d'}, h_{d'}}|(g_{d'} + h_{d'})^2}$$
$$\times \begin{bmatrix} (g_{d'} + h_{d'})\psi'(g_{d'} + h_{d'}) - h_{d'}\psi'(h_{d'}) \\ g_{d'}\psi'(g_{d'}) - (g_{d'} + h_{d'})\psi'(g_{d'} + h_{d'}) \end{bmatrix}$$
$$\times \sum_{i,j} \frac{\sum_{(l,l') \in \mathcal{P}_{\cdot d'} : y_{ll'} = 0} \frac{\kappa_{ll'i} \nu_{ll'j}}{\pi_{ll'}} \frac{a_{ij}}{a_{ij} + b_{ij}}}{1 - \frac{g_{d'}}{g_{d'} + h_{d'}} \frac{a_{ij}}{a_{ij} + b_{ij}}}.$$
If any $g_{d'} \leq 0$ or $h_{d'} \leq 0$, reduce s_t for this update (say by half each time).

Using our proposed sampling scheme, unbiased estimates of the true natural gradients can be computed via the Horvitz–Thompson estimator [see, e.g., Kolaczyk (2009)]. Let $\mathcal{P}_{d\cdot} = \{(l, l') \in \mathcal{P} | l = d\}$ and $\mathcal{P}_{\cdot d} = \{(l, l') \in \mathcal{P} | l' = d\}$ denote samples in $\mathcal{P}$ lying in the dth row and column of the adjacency matrix respectively. We have, for example, the following unbiased estimates:

$$\sum_{(l,l')\in\mathcal{P}_{d\cdot}} \frac{\kappa_{ll'}}{\pi_{ll'}} \approx \sum_{d'\neq d} \kappa_{dd'} \quad \text{and} \quad \frac{D}{|\mathcal{S}|} \sum_{\substack{(l,l')\in\mathcal{P}_{\cdot d'} \\ d'\in\mathcal{S}}} \frac{\kappa_{ll'i}\nu_{ll'j}}{\pi_{ll'}} \approx \sum_{(d,d')} \kappa_{dd'i}\nu_{dd'j}. \tag{4.3}$$

The estimator on the right arises from a two-stage cluster sampling scheme which involves sampling the columns ($d' \in \mathcal{S}$) in the first stage and then sampling document pairs from these columns in the second stage. We show that this estimator is unbiased in the supplement [Tan, Chan and Zheng (2016)].

The stochastic variational algorithm for our model is outlined in Algorithm 2. Implementation details are given in the supplement [Tan, Chan and Zheng (2016)].

5. Comparison with alternative approaches. We compare our model with the most representative peer methods, RTM and Pairwise-Link-LDA. As a baseline for comparing methods which integrate the modeling of text and links, we also consider "LDA + Regression," which involves fitting an LDA model to the documents followed by a logistic regression model to the links. The covariates corresponding to the observation $y_{dd'}$ are $\bar{\gamma}_d \odot \bar{\gamma}_{d'}$, where $\odot$ denotes the Hadamard product and the kth element of $\bar{\gamma}_d$ is $\gamma_{dk}/\sum_l \gamma_{dl}$. This approach models text and links separately and information from topic modeling is not utilized in the link structure. The RTM accounts for both text and links structure. However, it assumes a symmetric probability function and considers a diagonal weight matrix that allows only within topic interactions. In addition, RTM only models observed links and does not deal explicitly with the imbalance between links and nonlinks. We consider the RTM with an exponential link probability function. Pairwise-Link-LDA combines LDA and MMSB. Comparing Pairwise-Link-LDA with LMV illustrates the importance of the visibility measure in link prediction.

All models are estimated using variational methods and the code for these algorithms are reproduced in `R` by ourselves. For LDA + Regression, RTM and Pairwise-Link-LDA, we have tried to follow the implementations suggested by the original authors as closely as possible. Variational parameters in LMV, RTM and Pairwise-Link-LDA are initialized using the fitted LDA and the same priors are used across all models. Details on priors and stopping criteria are given in the supplement [Tan, Chan and Zheng (2016)].

6. Prediction and article recommendations. We discuss how our model can be used to predict links for new documents assuming knowledge only of the text. An important application of the predictive probabilities is in recommending scientific articles, for instance, to researchers searching for information on certain

research topics or who are preparing manuscripts and looking for relevant articles to cite. Using a short paragraph of text or even just keywords (as text of the "new document"), predictive probabilities of links to documents in the training set can be computed and used as a means to rank documents and construct recommendation lists. As our model captures both inter- and intra-topic citation probabilities and estimates the visibilities of individual articles (which are adjusted for field variation in citation practices), it can identify relevant articles which are multi-topic or of high visibility but coming from topics with low citation rates. Wang and Blei (2011) and Gopalan, Charlin and Blei (2014) consider recommendation of scientific articles to readers of online archives based on article content and reader preferences. They do not consider citations among articles and make use of collaborative filtering.

The predictive probability of a link to a document in the training set can be computed as follows. First, we fit the LMV model to documents in the training set. Then we perform variational inference on the new document d to obtain its topic proportions [see, e.g., Nallapati et al. (2008)]. That is, we iterate till convergence the updates:

1. $\gamma_d \leftarrow \alpha + \sum_n \phi_{dn}$,
2. $\phi_{dnk} \propto \exp\{\psi(\gamma_{dk}) - \psi(\sum_k \gamma_{dk}) + \sum_v w_{dnv}[\psi(\lambda_{kv}) - \psi(\sum_v \lambda_{kv})]\}$ for $n = 1, \ldots, N_d$, $k = 1, \ldots, K$,

where λ is obtained from the fitted model. The first update is similar to step 3 of Algorithm 1. In this case, we do not assume knowledge of the links of the new document. Hence, parameters associated with the links are absent. Let w_d and w_T denote the words of the new document d and the training set respectively and let y_T denote links within the training set. Approximating the true posterior by the variational approximation $q(\Theta)$, the posterior predictive probability that d will cite any document d' in the training set is

$$p(y_{dd'} = 1 | w_d, w_T, y_T) \approx \mathrm{E}_q(\tau_{d'})\mathrm{E}_q(\theta_d)^T \mathrm{E}_q(B)\mathrm{E}_q(\theta_{d'}) \tag{6.1}$$
$$= \frac{g_{d'}}{g_{d'} + h_{d'}} \left(\frac{\gamma_d}{\sum_{k=1}^K \gamma_{dk}}\right)^T \frac{a}{a+b} \frac{\gamma_{d'}}{\sum_{k=1}^K \gamma_{d'k}}.$$

6.1. *Predictive rank.* In the examples, we compute the average predictive rank of held-out documents, following Chang and Blei (2010), as a way to evaluate the fit between considered models and the data. The predictive rank captures a model's ability to predict documents that a test document will cite given only its words. To compute the predictive rank of a test document for model $\mathcal{M}$, first fit model $\mathcal{M}$ to documents in the training set. Using the fitted topics, obtain the topic proportions of the test document using just its words as described above. Compute the posterior predictive probability that the test document will cite each document in the training set [use (6.1) for LMV] and rank the documents according to this

probability. The predictive rank is the average rank of the documents which the held-out document actually did cite. Lower predictive rank indicates a better fit, and it also implies that the articles that were actually cited are placed closer to the top of a recommendation list for a test article.

7. Applications. We apply our methods to two real datasets. To assist understanding of our proposed model and to evaluate Algorithms 1 and 2, a simulation study is also provided in the supplement [Tan, Chan and Zheng (2016)]. In this study, we generate datasets from the LMV and demonstrate that Algorithms 1 and 2 are able to recover the structure of the blockmodel B as well as the visibility of each document. We also show that our model performs significantly better than LDA + Regression, RTM and Pairwise-Link-LDA in link predictions. Predictive results from Algorithms 1 and 2 are very close and our subsampling strategy helps to reduce computation times.

In the following, the blockmodels, visibilities, topic proportions and topic assignments of documents are estimated using the posterior means of corresponding variational approximations. For each citation, a hard assignment of topics to the citing and cited documents is obtained by taking the positions of the maximum elements of $\kappa_{dd'}$ and $\nu_{dd'}$ respectively. For instance, if $\kappa_{dd'} = [0.8, 0.2, 0, \ldots, 0]$ and $\nu_{dd'} = [0.3, 0.7, 0, \ldots, 0]$, we interpret the citation as being from topic 1 to 2. For visualization of fitted topics, we order terms in the vocabulary using the term score [Blei and Lafferty (2009)],

$$\text{term-score}_{kv} = \bar{\lambda}_{kv} \log\left\{ \frac{\bar{\lambda}_{kv}}{(\prod_{k=1}^{K} \bar{\lambda}_{kv})^{1/K}} \right\},$$

where $\bar{\lambda}_{kv} = \lambda_{kv} / \sum_l \lambda_{kl}$ denotes the posterior mean probability of the vth term appearing in the kth topic. The second part of the expression downweights terms that have high probability of appearing in all topics. This term score is inspired by the TFIDF term score used in information retrieval [Baeza-Yates and Ribeiro-Neto (1999)]. Further discussion on ways to examine the quality of uncovered topics can be found in the supplement [Tan, Chan and Zheng (2016)].

We denote LDA + Regression and Pairwise-Link-LDA by LDA + Reg and PLLDA respectively. Methods with subsampling have a "S" added at the end, for example, LMVS denotes LMV with subsampling. If publication times are taken into account, we add a "(t)" at the end.

7.1. *Cora dataset.* The Cora dataset from the `R` package lda [Chang (2012)] has 2410 documents, 4356 links and a vocabulary of 2961 terms. This dataset is relatively small and it allows us to compare the predictive performance and CPU times of Algorithms 1 and 2. We randomly divide the dataset into five folds; each fold is used in turn as a test set and the remaining folds are used for training. During training, only documents in the training set and links within them are used. We investigate the predictive performance of different models for number of topics K ranging from 5 to 13. For Algorithm 2, we set the minibatch size as 200.

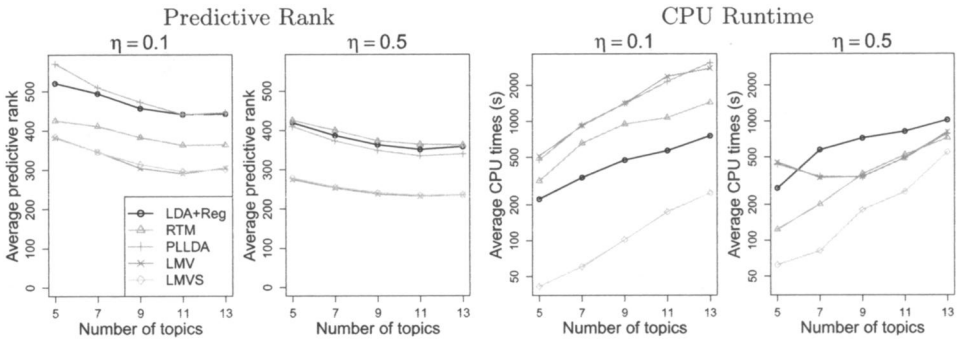


FIG. 3. *Cora: Average predictive ranks and CPU times in seconds for different methods.*

7.1.1. *Predictive performance and computation times.* The average predictive ranks and CPU times of different approaches are shown in Figure 3. We consider the hyperparameter η, which controls the concentration of the topic distribution, to be 0.1 or 0.5 (if η is small, the probability distribution on the vocabulary will be concentrated on a small number of terms). Except for the RTM, predictive performance of all other methods are better when η is 0.5 as compared to 0.1. LMV achieved significantly better predictive performance than the other models and attained 60–76% improvement in predictive rank over baseline.[4] Predictive performance of Algorithm 2 is close to that of Algorithm 1 even with subsampling and computation times reduced, particularly when $\eta = 0.1$. For large datasets, Algorithm 2 presents an avenue for overcoming computational and memory constraints while maintaining the same level of predictive performance. Predictive performance for the LMV stabilizes at around 9–11 topics when $\eta = 0.5$. In the following, we concentrate on the fitted models corresponding to $K = 9$ and $\eta = 0.5$ for one of the folds.

7.1.2. *Evaluating accuracy of Algorithm* 2. Repeating the LMVS (Algorithm 2) runs 50 times, we compute the average visibilities and blockmodel over these 50 runs and plot these quantities against corresponding values estimated by LMV (Algorithm 1) in Figure 4. There is very good agreement between the visibilities and blockdiagonal elements estimated by Algorithms 1 and 2. For the left plot, there is greater variation near zero, while the biggest two values appear to be slightly overestimated by LMVS in the right plot.

7.1.3. *Visibilities of individual articles.* Figure 5 plots the visibility of documents in the training set estimated by Algorithm 1 against their citation counts. There is a general trend of visibility increasing with citation counts. Hence, the

[4]Predictive rank computed by random guessing is $\frac{n+1}{2}$, where n is the number of documents in the training set.

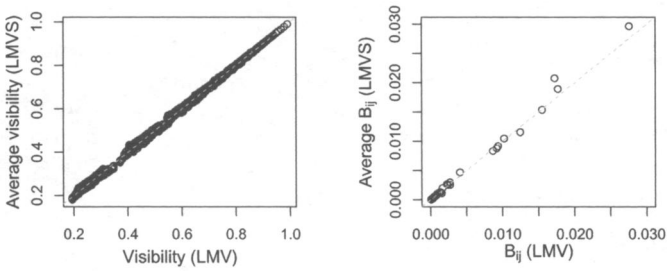


FIG. 4. *Cora*: *Visibilities* (*left*) *and blockmodel elements* (*right*) *estimated by LMVS and averaged over* 50 *runs against corresponding qualities estimated by LMV. Points lying close to the dotted line* ($y = x$) *indicate good agreement between LMV and LMVS.*

topic-adjusted visibility metric is capturing some degree of popularity. However, the relationship between visibility and citation counts is not monotone; a higher citation count does not necessarily imply a higher measure of visibility. The visibility metric $\tau_{d'}$ captures a complex mix of attributes of document d' that accounts for the variation in citation probability among documents with similar topic proportions. These include but are not limited to popularity.

To illustrate how the incorporation of visibilities improves predictive performance, let us consider as an example the test document: *"Some extensions of the k-means algorithm for image segmentation and pattern classification."* This article cited two documents from the training set: (1) *A Theory of Networks for Approximation and Learning* and (2) *Self-Organization and Associative Memory*. The citations, estimated visibility and predictive ranks of these documents are given in Table 1. The predictive ranks assigned by LMV and LMVS are significantly lower than the other methods. Interestingly, if visibility is not taken into account, the ranking by LMV will be similar to Pairwise-Link-LDA (120 and 357 for documents indexed 1 and 2 in order). However, factoring in the visibilities of these documents, which are much higher than the average of 0.42, improves predictive ranks tremendously. On average, LMV improves predictive performance by more than 30% over RTM, PLLDA and LDA + Reg when $\eta = 0.5$. This provides an indication of the favorable performance of LMV in article recommendation.

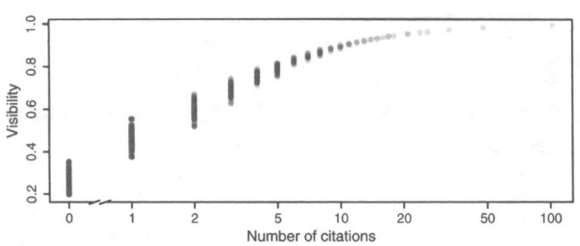
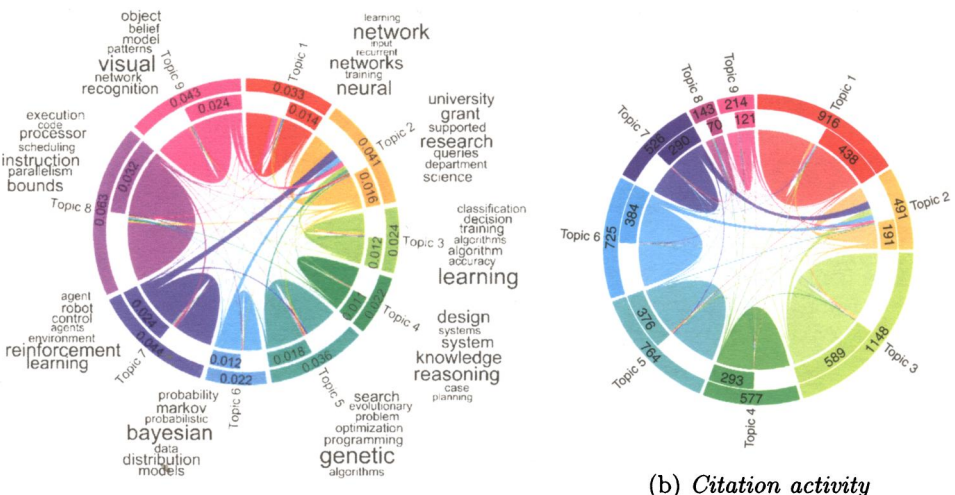


FIG. 5. *Cora*: *Visibility against number of citations for each document in training set.*

TABLE 1
Cora: Citations, visibility and predictive ranks of two cited documents

Index	Citations	$\hat{\tau}$	LDA + Reg	RTM	PLLDA	LMV	LMVS
1	26	0.96	98	108	120	10	9
2	48	0.98	385	1082	336	53	61

7.1.4. *Citation behavior among fitted topics.* Figure 6 provides a visualization of the LMV fitted using Algorithm 1, where (a) shows the estimated blockmodel B and top words of each topic and (b) shows the citation activity in the training set. Figure 6(b) is based on raw citation counts and tends to be dominated by topics with high citation frequency or a large number of publications. On the other hand, Figure 6(a) shows the citation probabilities within/between topics. It provides a more structured and unbiased understanding of the citation tendencies within/between topics. From Figure 6(a), topic 2 tends to be cited strongly by top-



(a) *Blockmodel B and top words of each topic*

(b) *Citation activity*

FIG. 6. (a) *Shows the citation patterns between different topics and the topic-to-topic citation tendencies. Each element B_{ij} is represented by a belt from topic i to j and the width of the belt is proportional to B_{ij}. The value on the inner arc of topic i represents the probability that topic i will cite any topic, while the value on the outer arc represents the total probability that topic i will cite or be cited by any topic. The top 7 words of each topic are displayed in word clouds and the font size is proportional to the term score.* (b) *Shows the actual citation activity among different topics. The width of each belt is proportional to the number of citations between the topics connected by the belt. The value on the inner arc is the total number of citations originating from a topic, while the value on the outer arc is the total number of citations coming from and going to that topic. The direction of the belts can be inferred from the presence (at the origin) and absence (at the destination) of the inner arcs.*

ics 3, 5–7 and 9. It also tends to cite nearly all other topics besides itself. Topic 9 tends to cite topics 1 and 2, and topic 8 has the highest tendency to cite documents within its own topic. Information such as this is helpful in understanding how citation behavior varies from one scientific area to another, and even among different research fields within a discipline.

7.2. *KDD high energy physics dataset.* The high energy physics (HEP) dataset for the KDD Cup 2003 competition [see Gehrke, Ginsparg and Kleinberg (2003)] contains 29,555 papers added to arXiv from 1992–2003 and is available at http://www.cs.cornell.edu/projects/kddcup/index.html. We consider the abstracts of 25,224 papers added between 1992–2001 and the 271,838 citations among them as the training set. We use 4064 papers which have cited at least one paper from the training set and were added between 2002–2003 as the test set. There are 60,914 citations from the test set to the training set. The vocabulary consists of 7211 terms after stemming and removal of stop words and infrequent terms. We have the dates when the papers were published online at the library of the Stanford Linear Acceleration Center (SLAC), as well as the year and month in which the papers were added on arXiv. As some papers were first published on SLAC and later added on arXiv or vice versa, we use the earlier of the two dates as the date when a paper is first available. Memory constraints render running the algorithms for Pairwise-Link-LDA and LMV in batch mode infeasible. Hence, we only use Algorithm 2 and Pairwise-Link-LDA is also implemented using our proposed subsampling strategy. A minibatch size of 2000 was used.

7.2.1. *Predictive performance and computation times.* Figure 7 shows the average predictive ranks and CPU times[5] of different approaches for number of topics K ranging from 10 to 30 and $\eta \in \{0.5, 0.1, 0.01\}$. LMV provides better predictive performance than Pairwise-Link-LDA and RTM for all η and K, and taking into account publication times yields significantly better predictions. This is likely due to more accurate estimation of the visibilities, as documents published at later dates are not penalized for not being cited by documents published before them. LMVS(t) achieved 73–83% improvement over baseline and optimal predictive performance at $K = 20$ and $\eta = 0.5$. In the following, we consider the model fitted by LMVS(t) when $K = 20$ and $\eta = 0.5$.

7.2.2. *Interpreting citation trends in HEP.* Figures 8 and 9 provide visualizations of the estimated blockmodel and the citation activity in the training set respectively. These plots may be read as in Figure 6. The blockmodel in Figure 8

[5]LDA was run in batch mode and CPU times are for model fitting only. We do not compute predictive ranks for LDA + Reg, as logistic regression for this dataset is prohibitively expensive. Note that LDA may also be implemented with stochastic variational inference.

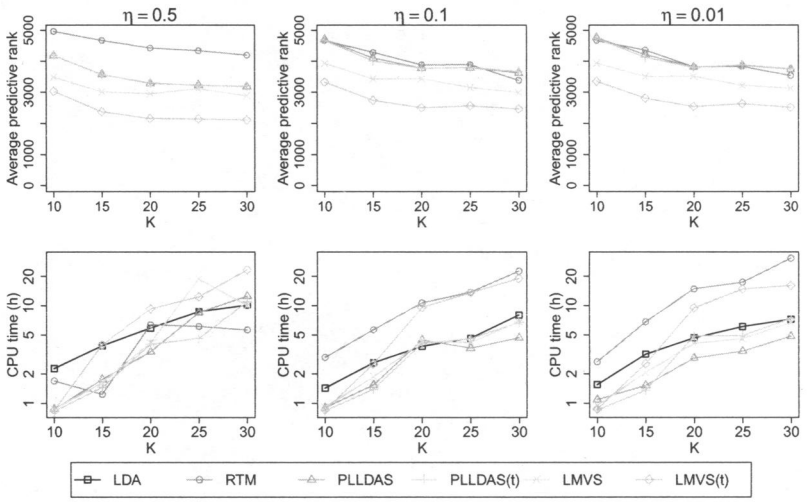


FIG. 7. *KDD: Average predictive ranks (first row) and CPU times in hours (second row) of different approaches. The columns correspond to* $\eta \in \{0.5, 0.1, 0.01\}$ *from left to right.*

indicates high probability of within-topic citation generally, while across-topic citation is much weaker in comparison. Figure 9 indicates that a large proportion of citations in the corpus occurred within topics 1–2.

Next, we focus on topics 1–4, 6 and 8–9 where interconnectivity is higher among them. Figure 10 provides a visualization of the citation patterns between these topics and reveals some interesting trends in the citation landscape of HEP. First, there is a strong tendency for within-topic citations, while the probability of across-topic citations is much weaker. However, the across-topic citation relations

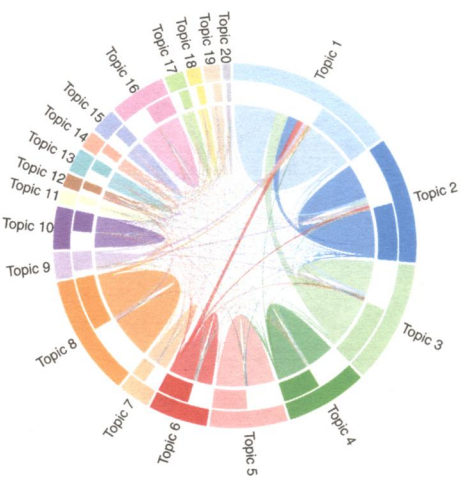


FIG. 8. *Blockmodel B.*

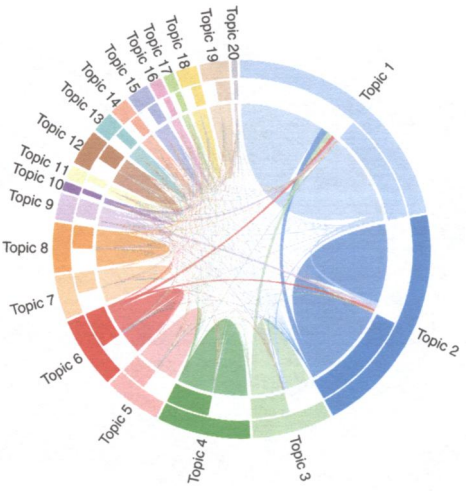


FIG. 9. *Citation activity.*

that do exist are unsurprising, as theoretical HEP deals with the fundamental aspects of Particle Physics and there are vast and deep links between these topics. It is also worth noting that certain topics do not have a high tendency to cite each other (e.g., topics 3 and 6) even though there are overlaps in the body of knowledge (e.g., supersymmetries and string theory). A possible reason is that HEP physicists may not always be aware of one another's work (especially among the 3 main

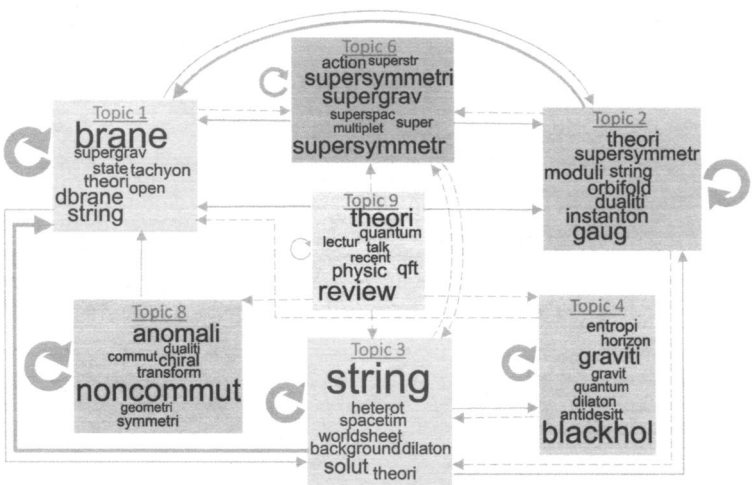


FIG. 10. *KDD: Visualization of citation patterns for topics* 1–4, 6 *and* 8–9. *Width of arrows are proportional to the citation strengths in the estimated blockmodel. Only elements greater than* 0.001 *are visualized. Dashed arrows represent elements less than* 0.005. *Top* 8 *words of each topic are shown. Font size within each topic is proportional to the term score.*

groups: experimental, phenomenological and theoretical). The articles from these groups have a different focus and may constitute different topics.

Topics 1 and 2 are both associated with string theory, which claims that the fundamental objects that make up all matter are strings (like rubber bands) instead of particles [e.g., Gubser (2010)]. The emphasis of topic 1 is on branes, which are multi-dimensional membranes that generalize the concept of particle (zero-dimensional brane) and string (one-dimensional brane) to higher dimensions. Topic 2 is associated with gauge theory and orbifold, which are both related to geometry. Duality, an important concept in string theory, relates branes (topic 1) to gauge theory and supersymmetry (topic 2). Thus, the citation relationship between topics 1 and 2 is expected. From Figure 10, these two topics also have a relatively higher probability of receiving citations from other topics. This could be socially driven by the authors' inclination to cite papers from which their respective topics originated. There may also be a tendency for researchers to appeal to popular topics and cite earlier successful theories such as the gauge theory, which is the fundamental edifice of high energy particle physics.

The prefix "super" is found in topics 1, 2 and 6. It is associated with supersymmetric theories which attempt to unify bosons and fermions under one generalized scheme by relating fractional spin to integral spins, and finally unifying force and matter particles. Links from topic 6 to topics 1 and 2 are therefore reasonable and expected. Topics 1 and 3 are clearly related, namely, brane, string and gravity. String theory holds the promise to unify gravity to the other fundamental forces in Physics at the expense of introducing more dimensions that need to be compactified mathematically. In the case of topic 4, research concerning entropy may be cited in relation to the event horizon of the black hole and this constitutes within-topic citations. Mini black holes may also be regarded as a collection of dbranes, resulting in citations from topics 4 to 1. For future investigation, it may be of interest to narrow the study to a certain HEP group (e.g., experimental, phenomenological or theoretical) and a period when a particular topic is in fashion.

7.2.3. *Visibilities of individual articles.* As in Figure 5, Figure 11 plots the estimated visibility of documents in the training set against their citation counts. There is a general trend of visibility increasing with citations as before. However,

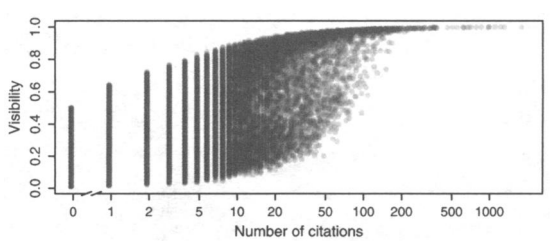


FIG. 11. *KDD: Visibility against number of citations for each document in the training set.*

the estimated visibility for each citation count now varies over a wide range. This range increases as the citation count increases from 0 but starts to gradually decrease as the citation count exceeds 20. It is more evident here than in Figure 5 that visibility correlates with but does not increase monotonically with citations. Hence, the visibility metric captures a complex mixture of attributes that includes but goes beyond popularity.

7.2.4. *Application: Article recommendation.* We use an example to illustrate the advantages that LMV can provide in article recommendation. Using the abstract of the article "Open Strings" from the test set as a query, we compute predictive probabilities of links from this article to each document in the training set using LMV, Pairwise-Link-LDA and RTM. In practice, key words or a short paragraph of text may be used as a search query if the abstract or manuscript is not available. Figure 12 lists the top fifteen recommended articles for each approach based on predictive probabilities (articles ranked closer to the top have higher probabilities of a link). Information such as topic proportions (in the form of barplots), title, citation counts, year of publication and visibility metric (for LMV) are also provided. This is a realistic representation of recommendation systems that can be constructed based on our proposed model.

In this example, five of the fifteen articles (marked by red asterisks) recommended by LMV were actually cited, while none recommended by RTM or Pairwise-Link-LDA was cited. Comparing the topic proportions of the test article with those of the recommended articles, we note that RTM tends to recommend articles with high proportions of topic 9. This is likely because RTM only allows within-topic interaction and the test article exhibits high proportions of topic 9. On the other hand, Pairwise-Link-LDA and LMV are able to model across-topic citation tendencies and the probability of topic 9 citing topics other than itself is quite high (see Figures 8 and 10). Hence, Pairwise-Link-LDA recommends articles mainly from topics 1 and 3, while articles recommended by LMV exhibit a higher degree of mixing with topics from 1–3 mainly and smaller proportions of some other topics. The articles recommended by LMV tend to have more citations than those recommended by Pairwise-Link-LDA and RTM. This is because the visibility metric captures a certain degree of popularity as shown in Figure 11. However, articles are not ranked based on citations alone; both topic compatibility and article visibility play a part in the ranking. For LMV, there is also a good mix of articles published from 1993–2000. While our model does not explicitly model time from publication, the visibility metric accounts for this to a certain degree in that old articles which have accumulated many citations will have high visibility, but so will recent articles which have managed to garner a proportionately smaller number of citations. This example highlights some of the advantages that LMV has to offer for article recommendation such as being able to identify relevant multi-topic articles and taking article visibility into account in rankings.

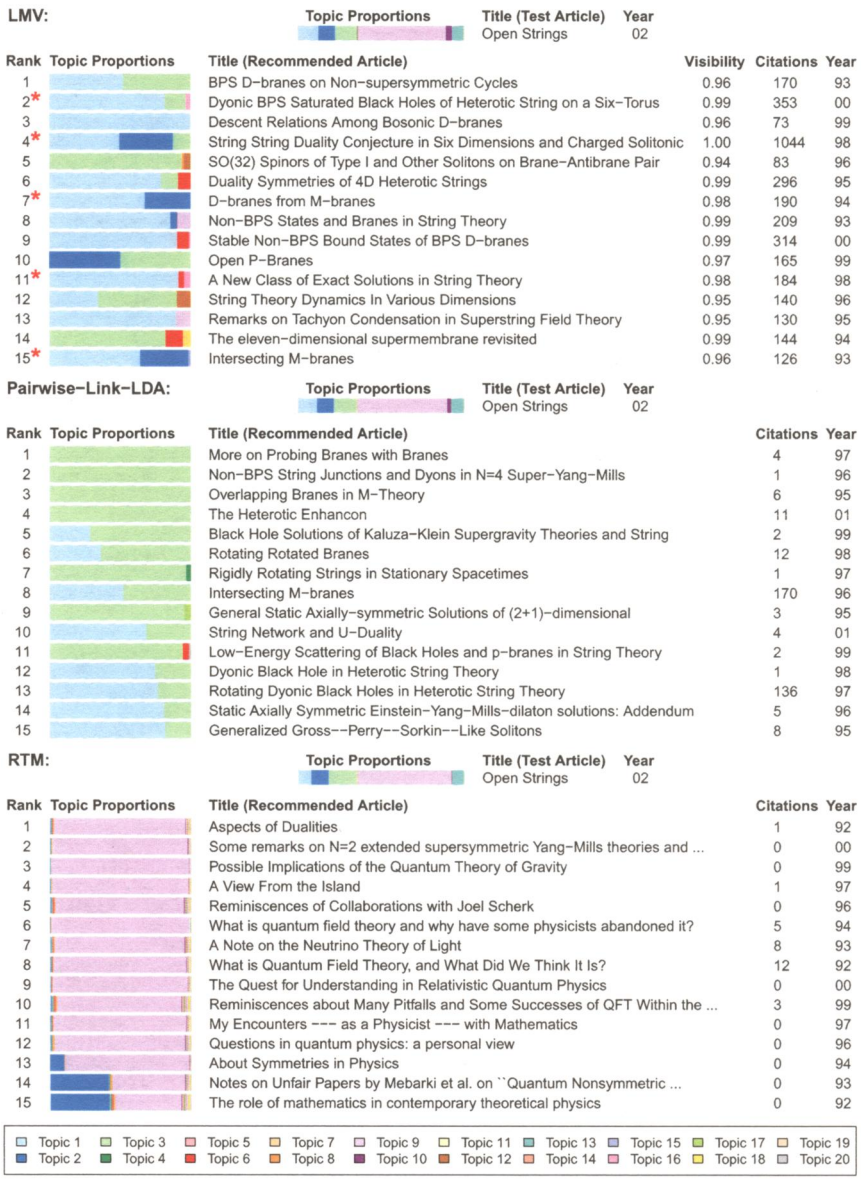


FIG. 12. *KDD: Top 15 articles recommended by LMV, Pairwise-Link-LDA and RTM. Asterisks indicate articles actually cited by test article. This figure is intended to be read in color (color version available in online publication).*

7.2.5. *Visibility as a topic-adjusted measure.* We examine the behavior of the visibility measure by plotting the estimated visibilities by topic and by year of publication in Figure 13. We note that the visibilities do not vary radically from one topic to another and are in fact much more comparable across different topics

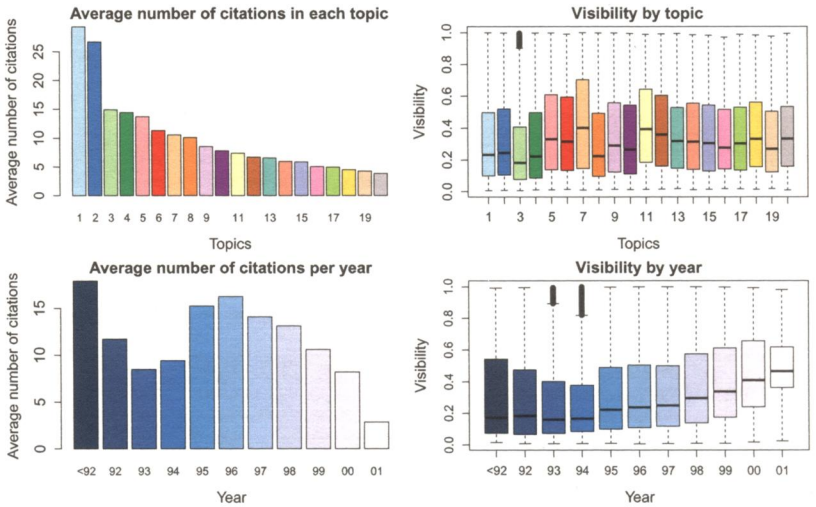


FIG. 13. *KDD: First row shows a barplot of the average number of citations in each topic (left) and boxplots of the visibilities of training documents classified by topic (right). Second row shows a barplot of the average number of citations in each year (left) and boxplots of the visibilities of training documents classified by year (right).*

than the average number of citations per article. The boxplots of visibilities by year does not display any systematic correlation with time of publication as well. Interestingly, even as the average number of citations per year is decreasing rapidly for articles published in the later years, the visibilities are not showing any decreasing trend. This could be due to the tendency to cite the newest and latest research.

Figure 14 is a barplot showing the topic proportions of documents with 40 citations. We have fixed the number of citations in order to study how visibility varies with topic proportions. As visibility increases, there is a transition in the color of the bars from being dominantly blue to red. That is, for the *same citation count* (40 in this case), the estimated visibility of an article from topics with low citation rates (red) tends to be greater than an article from topics with high citation rates (blue). This phenomenon demonstrates that the visibility metric adjusts for differences in citation frequency among different topics by assigning higher visibilities to articles from topics with lower citation rates for the same citation count. For example, consider an article from Mathematics and another from Molecular Biology having the same number of citations. As the citation rate in Mathematics is much lower than in Molecular Biology, it is inappropriate to compare the raw citation counts of these articles directly. Citations in Mathematics ought to be assigned higher weights and this can be achieved through normalization procedures, for example, by dividing citation counts by the average number of citations per article for a discipline [Radicchi, Fortunato and Castellano (2008)]. However, the choice of a suitable reference standard is a very intricate issue. The topic-adjusted

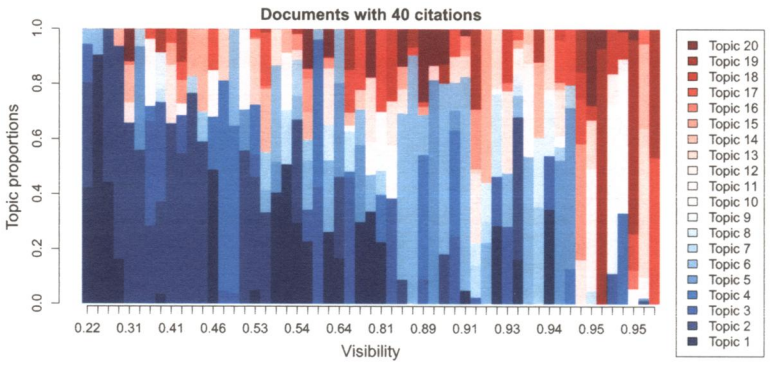


FIG. 14. *KDD barplot: Each bar represents the topic proportions of a document with* 40 *citations. The documents are arranged by visibility in increasing order. The legend indicates the color representing each topic. The topics are ordered by average number of citations received; topic* 1 *has the highest and topic* 20 *the lowest average number of citations. This figure is intended to be read in color (color version available in online publication).*

visibility metric that we propose offers an alternative to tackle the problem of field variation.

8. Conclusion. Citation activities of scientific applications and other article-centric activities (shares on social web, for instance) are of interest for the evaluation of scientific merit and impact of published research. In particular, the citation network among published articles is a special case of a relational network. Communities detected in a citation network have been shown to correlate well with scientific areas and research topics. A unique feature of a citation network is that content information is available on individual nodes, which is arguably influenced by the same mixed group structure underpinning the citation connectivity. Combining content and connectivity information in a citation network may alleviate the multi-modality issue of community detection in network analysis. On the other hand, communities detected in the citation network regularize the topic modeling on the content information.

The probability of being cited for a particular article is determined not only by its membership in one or more topic domains and the topic-level citation rates (within and between domains), but also factors that are unique to this publication such as timing, visibility of the authors, and novelty and importance of the results. In this paper we introduce a model for citation networks that infers the topic domain structure of the articles and their citation links, and estimates the citation activity rates at both the topic domain level and the article level. For each article, we introduce a latent variable that serves as a topic-adjusted visibility metric of this article. A higher value of this latent metric indicates that this article is more likely to be cited than other articles that are located close by in the topic domain. As we have shown in our application to real datasets, this metric correlates with

raw citation counts but is not merely a measure of popularity, as it accounts for variation in activity levels among different topics. The proposed model leads to significant improvement in link predictions, which can be helpful in article recommendation. It provides a better visibility metric for comparing individual scientific publications across different fields.

The inference of the proposed model is realized via a novel variational Bayes algorithm. For real-world large document networks, we propose a subsampling strategy that enables the use of stochastic variational inference, which is computationally efficient and achieves a similar level of predictive performance as the variational Bayes algorithm.

Acknowledgments. We thank the Editor, Associate Editor and the reviewers for their comments which have improved the manuscript greatly.

SUPPLEMENTARY MATERIAL

Supplement to "Topic-adjusted visibility metric for scientific articles" (DOI: 10.1214/15-AOAS887SUPP; .pdf). We provide additional material to support the results in this paper. This includes further discussions, detailed derivations, illustrations and a simulation study.

REFERENCES

ABRAMO, G. and D'ANGELO, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics* **87** 499–514.

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

ALBERTS, B. (2013). Impact factor distortions. *Science* **340** 787.

AMARI, S. (1998). Natural gradient works efficiently in learning. *Neural Comput.* **10** 251–276.

BAEZA-YATES, R. and RIBEIRO-NETO, B. (1999). *Modern Information Retrieval.* ACM Press, New York.

BALASUBRAMANYAN, R. and COHEN, W. W. (2013). Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In *Proceedings of the* 2011 *SIAM International Conference on Data Mining* (B. Liu, H. Liu, C. Clifton, T. Washio and C. Kamath, eds.) 450–461. SIAM Publications Online.

BLEI, D. M. and LAFFERTY, J. D. (2009). Topic models. In *Text Mining: Classification, Clustering, and Applications* (A. N. Srivastava and M. Sahami, eds.) 71–89. Chapman & Hall/CRC, Boca Raton, FL.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BORNMANN, L. and DANIEL, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation* **64** 45–80.

BRAUN, M. and MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **105** 324–335. MR2757203

CASADEVALL, A. and FANG, F. C. (2014). Causes for the persistence of impact factor mania. *The American Society for Microbiology* **5** e00064–14.

CHANG, J. (2012). Collapsed Gibbs sampling methods for topic models. R package: lda (version 1.3.2). Available at http://cran.r-project.org/web/packages/lda/index.html.

CHANG, J. and BLEI, D. M. (2010). Hierarchical relational models for document networks. *Ann. Appl. Stat.* **4** 124–150. MR2758167

CHEN, P. and REDNER, S. (2010). Community structure of the physical review citation network. *J. Informetr.* **4** 278–290.

CHEN, N., ZHU, L., XIA, F. and ZHANG, B. (2013). Generalized relational topic models with data augmentation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence* (F. Rossi, ed.) 1273–1279. AAAI Press, Menlo Park, CA.

CRESPO, J. A., LI, Y. and RUIZ-CASTILLO, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLOS ONE* **7** e33833.

CRESPO, J. A., HERRANZ, N., LI, Y. and RUIZ-CASTILLO, J. (2013). The effect on citation inequality of differences in citation practices at the web of science subject category level. *Journal of the Association for Information Science and Technology* **65** 1244–1256.

FENNER, M. (2014). Altmetrics and other novel measures for scientific impact. In *Opening Science* (S. Bartling and S. Friesike, eds.) 179–189. Springer, New York.

GARFIELD, E. (1979). *Citation Indexing. Its Theory and Applications in Science, Technology, and Humanities*. Wiley, New York.

GARFIELD, E. (2006). The history and meaning of the journal impact factor. *The Journal of the American Medical Association* **295** 90–93.

GEHRKE, J., GINSPARG, P. and KLEINBERG, J. M. (2003). Overview of the 2003 KDD cup. *SIGKDD Explorations* **5** 149–151.

GOPALAN, P. K. and BLEI, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proc. Natl. Acad. Sci. USA* **110** 14534–14539. MR3105375

GOPALAN, P., CHARLIN, L. and BLEI, D. M. (2014). Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems* **27** (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, eds.) 3176–3184. Curran Associates, Red Hook, NY.

GUBSER, S. S. (2010). *The Little Book of String Theory*. Princeton Univ. Press, Princeton, NJ. MR2655897

HIRSCH, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **102** 16569–16572.

HO, Q., EISENSTEIN, J. and XING, E. P. (2012). Document hierarchies from text and links. In *Proceedings of the 21st International Conference on World Wide Web* 739–748. ACM, New York.

HO, Q., PARIKH, A. P. and XING, E. P. (2012). A multiscale community blockmodel for network exploration. *J. Amer. Statist. Assoc.* **107** 916–934. MR3010880

HOFFMAN, M. D., BLEI, D. M. and BACH, F. (2010). Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* **23** (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel and A. Culotta, eds.) 856–864. Curran Associates, Red Hook, NY.

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KLEINBERG, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM* **46** 604–632. MR1747649

KNOWLES, D. A. and MINKA, T. P. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems* **24** 1701–1709. Curran Associates, Red Hook, NY.

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data. Methods and Models*. Springer, New York. MR2724362

MOED, H. F. (2010). Measuring contextual citation impact of scientific journals. *J. Informetr.* **4** 265–277.

NALLAPATI, R., AHMED, A., XING, E. P. and COHEN, W. W. (2008). Joint latent topic model for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discover and Data Mining* 542–550. ACM Press, New York.

NEISWANGER, W., WANG, C., HO, Q. and XING, E. P. (2014). Modeling citation networks using latent random offsets. In *Proceedings of 30th Conference on Uncertainty in Artificial Intelligence* (N. L. Zhang and J. Tian, eds.) 633–642. AUAI Press, Corvallis, OR.

NEYLON, C. and WU, S. (2009). Article-level metrics and the evolution of scientific impact. *PLOS Biology* **7** e1000242.

RABINOVICH, M. and BLEI, D. M. (2014). The inverse regression topic model. In *Proceedings of the 31st International Conference on Machine Learning, Beijing, China* (E. P. Xing and T. Jebara, eds.) *J. Mach. Learn. Res. Workshop and Conference Proceedings* **32** 199–207.

RADICCHI, F., FORTUNATO, S. and CASTELLANO, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. USA* **105** 17268–17272.

RAFTERY, A. E., NIU, X., HOFF, P. D. and YEUNG, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *J. Comput. Graph. Statist.* **21** 901–919. MR3005803

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668

ROBERTS, M. E., STEWART, B. M., TINGLEY, D. and AIROLDI, E. M. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, Nevada, US.*

SCHUBERT, A. and BRAUN, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics* **36** 311–324.

SEGLEN, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *Br. Med. J.* **314** 498–502.

SIMONS, K. (2008). The misused impact factor. *Science* **322** 165.

SPALL, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, Hoboken, NJ. MR1968388

TADDY, M. (2013). Multinomial inverse regression for text analysis. *J. Amer. Statist. Assoc.* **108** 755–770. MR3174658

TADDY, M. (2015). Distributed multinomial regression. *Ann. Appl. Stat.* **9** 1394–1414. MR3418728

TAN, L. S. L., CHAN, A. and ZHENG, T. (2016). Supplement to "Topic-adjusted visibility metric for scientific articles." DOI:10.1214/15-AOAS887SUPP.

VINKLER, P. (2003). Relations of relative scientometric indicators. *Scientometrics* **58** 687–694.

WANG, C. and BLEI, D. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 448–456. ACM Press, New York.

WANG, C., PAISLEY, J. and BLEI, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *Proc. of the 14th Int'l. Conf. on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA.* (G. Gordon, D. Dunson and M. Dudík, eds.) *J. Mach. Learn. Res. Workshop and Conference Proceedings* **15** 752–760.

ZHANG, A., ZHU, J. and ZHANG, B. (2013). Sparse relational topic models for document networks. In *Machine Learning and Knowledge Discovery in Databases* **8188** (H. Blockeel, K. Kersting S. Nijssen and F. Železný, eds.) 670–685. Springer, Heidelberg.

ZHU, Y., YAN, X., GETOOR, L. and MOORE, C. (2013). Scalable text and link analysis with mixed-topic link models. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 473–481. ACM, New York.

L. S. L. TAN
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
ROOM 1005 SSW, MC 4690
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
AND
DEPARTMENT OF STATISTICS & APPLIED PROBABILITY
BLOCK S16, LEVEL 7, 6 SCIENCE DRIVE 2
FACULTY OF SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
E-MAIL: statsll@nus.edu.sg

A. H. CHAN
PHYSICS DEPARTMENT
BLK S12 (MEZZANINE LEVEL)
FACULTY OF SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE
2 SCIENCE DRIVE 3
SINGAPORE 117551
E-MAIL: phycahp@nus.edu.sg

T. ZHENG
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
ROOM 1005 SSW, MC 4690
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
E-MAIL: tzheng@stat.columbia.edu