Bibliographic Analysis with the Citation Network Topic Model

Kar Wai Lim  
Australian National University, Canberra, Australia  
NICTA, Canberra, Australia  

Wray Buntine  
Monash University, Clayton, Australia

Keywords: author-citation network, topic model, Bayesian non-parametric

Abstract
Bibliographic analysis considers author’s research areas, the citation network and paper content among other things. In this paper, we combine these three in a topic model that produces a bibliographic model of authors, topics and documents using a non-parametric extension of a combination of the Poisson mixed-topic link model and the author-topic model. We propose a novel and efficient inference algorithm for the model to explore subsets of research publications from CiteSeerX. Our model demonstrates improved performance in both model fitting and a clustering task compared to several baselines.

1. Introduction
Models of bibliographic data need to consider many kinds of information. Articles are usually accompanied by metadata, for example, authors, publication data, categories and time. Cited papers can also be available. When authors’ topic preferences are modelled, we need to associate the document topic information somehow with the authors’. Jointly modelling text data with citation network information can be challenging for topic models, and the problem is confounded when also modelling author-topic relationships.

In this paper, we propose a topic model to jointly model authors’ topic preferences, text content and the citation network. The model is a non-parametric extension of previous models discussed in Section 2. We derive a novel algorithm that allows the probability vectors in the model to be integrated out, using simple assumptions and approximations, which give Markov chain Monte Carlo (MCMC) inference via discrete sampling. Section 3, 4 and 5 detail our model and its inference algorithm. Applying our model on research publication data, we demonstrate the model’s improved performance, on both model fitting and a clustering task, compared to baselines. We describe the datasets used in Section 6 and report on experiments in Section 7. Additionally, we qualitatively analyse the inference results produced by our model. We find that the topics returned have high comprehensibility.

2. Related Work

Latent Dirichlet Allocation (LDA) is the simplest Bayesian topic model used in modelling text, which also allows easy learning of the model. Teh and Jordan (2010) proposed the Hierarchical Dirichlet process (HDP) LDA, which utilises the Dirichlet process (DP) as a non-parametric prior which allows a non-symmetric, arbitrary dimensional topic prior to be used. Furthermore, one can replace the Dirichlet prior on the word vectors with the Pitman-Yor Process (PYP) (Teh, 2006b), which models the power-law of word frequency distributions in natural language (Sato and Nakagawa, 2010).

Variants of LDA allow incorporating more aspects of a particular task and here we consider authorship and citation information. The author-topic model (ATM) (Rosen-Zvi et al., 2004) uses the authorship information to restrict topic options based on author. Some recent work jointly models the document citation network and text content. This includes the relational topic model (Chang and Blei, 2010), the Poisson mixed-topic link model (PMTLM) (Zhu et al., 2013) and Link-PLSA-LDA (Nallapati et al., 2008). An extensive review of these models can be found in Zhu et al. (2013). The Citation Author Topic (CAT) model (Tu et al., 2010) models the author-author network on publications based on citations using an extension of the ATM. Note that our work is different to CAT in that we model the author-document-citation network instead of author-author network.

The Topic-Link LDA (Liu et al., 2009) jointly models author and text by using the distance between the document and author topic vectors. Similarly the Twitter-Network topic model (Lim et al., 2013) models the author (“follower”) network based on author topic vectors, but using a Gaussian process to model the network. Note that our work considers the author-document-citation of Liu et al. (2009) using the techniques developed in Lim et al. (2013), but using the PMTLM of Zhu et al. (2013) to model the network which lets one integrate PYP hierarchies with the PMTLM using efficient MCMC sampling.

There is also existing work on analysing the degree of authors’ influence. On publication data, Kataria et al. (2011) and Mimno and McCallum (2007) analyse influential authors with topic models. While Weng et al. (2010), Tang et al. (2009) and Liu et al. (2010) use topic models to analyse users’ influence on social media.

3. Citation Network Topic Model

In this section, we propose a topic model that jointly model the text, authors, and the citation network of research publications (documents). We name the topic model the Citation-Network Topic Model (CNTM). We first discuss the topic model part of CNTM where the citations are not considered, which will be used for comparison later in Section 7. The full graphical model for CNTM is displayed in Figure 1. To clarify the notations used in this paper, variables that are without subscript represent a collection of variables of the same notation. For example, \( w_d \) would represent all the words in document \( d \), that is, \( w_d = \{w_{d1}, \ldots, w_{dN_d}\} \) where \( N_d \) is the number of words in document \( d \); and \( w \) represents all words in a corpus, \( w = \{w_1, \ldots, w_D\} \), where \( D \) is the number of documents.
The CNTM uses the Griffiths-Engen-McCloskey (GEM) (Pitman, 1996) distribution to generate probability vectors and the Pitman-Yor process (PYP) (Teh, 2006b) to generate probability vectors given another probability vector (called mean or base distribution). Both GEM and PYP are parameterised by a discount parameter $\alpha$ and a concentration parameter $\beta$. PYP is additionally parameterised by a base distribution $H$, which is also the mean of the PYP. Note that the GEM distribution is equivalent to a PYP with a base distribution that generates an ordered integer label.

In modelling authors, CNTM modifies the approach of the author-topic model (Rosen-Zvi et al., 2004), which assumes that the words in a publication are equally attributed to the different authors. This is not reflected in practice since publications are often written more by the first author, excepting when the order is alphabetical. An approximation we make in this work is that the first author is dominant. We could model the influence of each author on a publication, say, using a Dirichlet distribution, but we found that considering only the first author gives a simpler learning algorithm and cleaner results.

IN CNTM, we first sample a root topic distribution $\mu$ with a GEM distribution, to act as a base distribution for the author-topic distributions $\nu_a$ for each author $a$:

$$
\mu \sim \text{GEM}(\alpha^\mu, \beta^\mu), \quad \nu_a|\mu \sim \text{PYP}(\alpha^{\nu_a}, \beta^{\nu_a}, \mu).
$$

Given the first author $a_d$ of each publication $d$, we sample the document-topic prior $\theta'_d$ and the document-topic distribution $\theta_d$:

$$
\theta'_d|a_d, \nu \sim \text{PYP}(\alpha^{\theta'_d}, \beta^{\theta'_d}, \nu_{a_d}), \quad \theta_d|\theta'_d \sim \text{PYP}(\alpha^{\theta_d}, \beta^{\theta_d}, \theta'_d).
$$

Note that instead of modelling a single document-topic distribution, we model a document-topic hierarchy with $\theta'$ and $\theta$. The primed $\theta'$ represents the topics of the document in the context of the citation network. The unprimed $\theta$ represents the topics of the text, naturally related to $\theta'$ but not the same. Such modelling gives citation information a higher impact to counter the relatively low amount of citations compared to the text. More details on the motivation of such modelling is presented in the supplementary materials.
For the vocabulary side, we generate a background word distribution $\gamma$, where $H^\gamma$ is a discrete uniform vector of length $|\mathcal{V}|$ and $\mathcal{V}$ is the set of distinct word tokens observed. Then, we sample a topic-word distribution $\phi_k$ for each topic $k$, with $\gamma$ as the base distribution:

$$
\gamma \sim \text{PYP}(\alpha^\gamma, \beta^\gamma, H^\gamma), \quad \phi_k|\gamma \sim \text{PYP}(\alpha^{\phi_k}, \beta^{\phi_k}, \gamma)
$$

Modelling word burstiness (Buntine and Mishra, 2014) is important since, as shown in Section 6, words in a document are likely to repeat in the document. This is addressed by making topics bursty, so each document only focuses on a subset of words in the topic. To generate $\phi'_{dk}$ for each topic $k$ in document $d$:

$$
\phi'_{dk}|\phi_k \sim \text{PYP}(\alpha^{\phi'_{dk}}, \beta^{\phi'_{dk}}, \phi_k)
$$

Finally, for each word $w_{dn}$ in document $d$, we sample the corresponding topic assignment $z_{dn}$ from the document-topic distribution $\theta_d$, while the word $w_{dn}$ is sampled from the topic-word distribution $\phi'_{d}$ given $z_{dn}$.

$$
z_{dn}|\theta_d \sim \text{Discrete}(\theta_d), \quad w_{dn}|z_{dn}, \phi'_d \sim \text{Discrete}(\phi'_{dz_{dn}})
$$

Note that $w$ includes words from title and abstract, but not the full article of a publication. This is because title and abstract provide a good summary of a publication’s topics, while the full article contains too much detail.

### 3.2. Citation Network Poisson Model

In CNTM, we assume that the citations are generated based on the topics relevant to the publications’ using the degree-corrected version of the PMTLM (Zhu et al., 2013). Denoting $x_{ij}$ as the number of times document $i$ citing document $j$, we model $x_{ij}$ with a Poisson distribution with mean parameter $\lambda_{ij}$:

$$
x_{ij}|\lambda_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad \lambda_{ij} = \lambda^+_i \lambda^-_j \sum_k \lambda^T_k \theta'_{ik} \theta'_{jk}
$$

Here, $\lambda^+_i$ is the propensity of document $i$ to cite and $\lambda^-_j$ represents the popularity of cited document $j$ and $\lambda^T_k$ scales the $k$-th topic. Hence, a citation from document $i$ to document $j$ is more likely when these documents are having relevant topics. The Poisson distribution is used instead of a Bernoulli because it leads to dramatically reduced complexity in analysis.

### 4. Model Representation and Posterior

Before presenting the posterior used to develop the MCMC sampler, we briefly review handling of hierarchical PYP models in Section 4.1. We cannot provide an adequately detailed review in this paper, thus we present the main ideas.

---

1. Note that Poisson distribution is similar to the Bernoulli distribution when the mean parameter is small.
4.1. Modelling with Hierarchical PYPs

The key to efficient Gibbs sampling with PYPs is to marginalise out the probability vectors (e.g. topic distributions) in the model and record various associated counts instead, thus yielding a collapsed sampler. While a common approach here is to use the hierarchical Chinese Restaurant Process (CRP) of Teh and Jordan (2010), we use another representation that requires no dynamic memory and has better inference efficiency (Chen et al., 2011).

We denote \( f(N) \) as the marginalised likelihood associated with the probability vector \( N \). Since the vector is marginalised out, the marginalised likelihood is in terms of — using the CRP terminology — the customer counts \( c_N = (\cdots, c^N_k, \cdots) \) and the table counts \( t^N = (\cdots, t^N_k, \cdots) \). The customer count \( c^N_k \) corresponds to the number of data points (e.g. words) assigned to group \( k \) (e.g. topic) for variable \( N \). Here, the table counts \( t^N \) represent the subset of \( c^N \) that gets passed up the hierarchy (as customers for the parent probability vector of \( N \)). We also denote \( C^N = \sum_k c^N_k \) as the total customer counts for node \( N \), and similarly, \( T^N = \sum_k t^N_k \) is the total table counts. The marginalised likelihood is:

\[
f(N) = \frac{(\beta^N N)^{T^N}}{(\beta^N N)^{C^N}} S_{x,y}^{c^N,\alpha^N} ; \quad \text{for } N \sim \text{PYP}(\alpha^N, \beta^N, \mathcal{P}) .
\]  

(2)

\( S_{x,y} \) is the generalised Stirling number; both \((x)_C\) and \((x|y)_C\) denote the Pochhammer symbol (rising factorial), see Buntine and Hutter (2012) for details. Note the GEM distribution behaves like a PYP in which the table count \( t^N_k \) is always 1 for non-zero \( c^N_k \).

The innovation of Chen et al. (2011) was to notice that sampling with Equation 2 directly led to poor performance due to inadequate mixing. They introduce a new Bernoulli indicator variable \( u^N_k \) for each customer who has contributed a “+1” to \( c^N_k \). A value \( u^N_k = 1 \) indicates that the customer has opened a new table, which also means the customer has also contributed a “+1” to \( t^N_k \) and thus has been passed up the hierarchy to the parent variable \( P \). The process repeats at the parent node because the “+1” to \( t^P_k \) is inherited as a “+1” to \( c^P_k \), and thus we now need to consider the value of \( u^P_k \). If \( u^N_k = 0 \) then a “+1” was not inherited and a corresponding \( u^P_k \) does not exist. The use of indicator variables has been empirically shown to lead to better mixing of the samplers.

Note that even though the probability vectors are integrated out and not explicitly stored, they can easily be estimated from the associated counts. The probability vector \( N \) is estimated from the counts and parent probability vector \( P \) using standard CRP estimation:

\[
N = \left( \cdots, \frac{(\alpha^N T^N + \beta^N)P_k + c^N_k - \alpha^N T^N_k}{\beta^N + C^N}, \cdots \right) .
\]  

(3)

4.2. Likelihood for the Hierarchical PYP Topic Model

We use bold face capital letters to denote the set of all relevant lower case variables, for example, \( Z = \{z_{11}, \cdots, z_{DN_D}\} \) denotes the set of all topic assignments. Variables \( W, T \) and \( C \) are similarly defined, that is, they denote the set of all words, table counts and customer counts respectively. Additionally, we denote \( \zeta \) as the set of all hyperparameters (such as the \( \alpha \)'s). With the probability vectors replaced by the counts, the likelihood of the topic
model can be written — in terms of $f(\cdot)$ — as $p(Z, W, T, C | \zeta) \propto f(\mu) \left( \prod_{a=1}^{A} f(\nu_a) \right) \left( \prod_{d=1}^{D} f(\theta_d') f(\theta_d) \right) \left( \prod_{k=1}^{K} f(\phi_k') f(\gamma) \right) \left( \prod_{v} \left( \frac{1}{|V|} \right)^{t_v} \right)$ \hspace{1cm} (4)

Note that the last term in Equation 4 corresponds to the parent probability vector of $\gamma$ (see Section 3.1), and $v$ indexes the unique word tokens in vocabulary set $V$.

4.3. Likelihood for the Citation Network Poisson Model

For the citation network, the Poisson likelihood for each $x_{ij}$ uses the definition of $\lambda_{ij}$ in Equation 1. Note that the term $x_{ij}!$ is dropped due to the limitation of the data that $x_{ij} \in \{0, 1\}$, thus $x_{ij}!$ is evaluated to 1. With conditional independence of $x_{ij}$, the joint likelihood for the whole citation network $X = \{x_{11}, \cdots, x_{DD}\}$ can be written as

$$p(X | \lambda, \theta') = \left( \prod_{i} (\lambda_i^+)^{g_i^+} (\lambda_i^-)^{g_i^-} \right) \prod_{ij} \left( \sum_{k} \lambda_k^T \theta_{ik} \theta_{jk}' \right)^{x_{ij}} \exp \left( - \sum_{ijk} \lambda_i^+ \lambda_j^- \lambda_k^T \theta_{ik} \theta_{jk}' \right),$$

where $g_i^+$ is the number of citations for publication $i$, $g_i^+ = \sum_j x_{ij}$, and $g_i^-$ is the number of times publication $i$ being cited, $g_i^- = \sum_j x_{ji}$. We also make a simplifying assumption\footnote{Technically, defining $x_{ii}$ allows us to rewrite the joint likelihood into another form for efficient caching.} that $x_{ii} = 1$ for all documents $i$, that is, all publications are treated as self-cited.

In the next section, we demonstrate that our model representation gives rise to an intuitive sampling algorithm for learning the model. We also show how the Poisson model integrates into the topic modelling framework.

5. Inference Techniques

Here, we derive the Markov chain Monte Carlo (MCMC) algorithms for learning the Citation Network Topic Model. We first detail the Gibbs sampler for the topic model and then discuss the Metropolis-Hastings (MH) algorithm for the citation network. The full inference procedure is performed by alternating between the Gibbs sampler and the MH algorithm.

5.1. Collapsed Gibbs Sampler for the Hierarchical PYP Topic Model

To jointly sample the words’ topic and the associated counts in the CNTM, we use a collapsed Gibbs sampler designed for the PYP (Chen et al., 2011). The concept of the sampler is analogous to LDA, which consists of decrementing the counts associated with a word, sampling the respective new topic assignment for the word, and incrementing the associated counts. Our collapsed Gibbs sampler is more complicated than LDA. In particular, we have to consider the indicators $u_{nk}$ described in Section 4.1 operating on the hierarchy of PYPs.

The sampler proceeds by considering the latent variables associated with a given word $w_{dn}$. First, we decrement out the effects of the latent variables, the topic $z_{dn} = k$ and the chain of indicator variables $u_{nk}^\theta, \nu_k^\theta, \mu_k^\nu, \mu_k^\mu$ (where they exist). After decrementing,
we jointly sample a new topic \( z_{dn} \) and the associated indicators (which contribute “+1” to counts) for word \( w_{dn} \) from their joint conditional posterior distribution:

\[
p(z_{dn}, T, C | Z^{-dn}, W, T^{-dn}, C^{-dn}, \zeta) = \frac{p(Z, W, T, C | \zeta)}{p(Z^{-dn}, W, T^{-dn}, C^{-dn} | \zeta)}.
\]

where the superscript \( \square^{-dn} \) indicates that the topic \( z_{dn} \), indicators and the associated counts for word \( w_{dn} \) are not observed in the respective sets, i.e. the state after decrement. The modularised likelihood of Equation 4 allows the conditional posterior (Equation 5) to be computed easily, since it simplifies to ratios of likelihood \( f(\cdot) \), which simplifies further as the counts differ by at most 1 during sampling. For instance, the ratio of the Pochhammer symbols, \( (x|y)_{C+1}/(x|y)_C \), simplifies to \( x + Cy \), while the ratio of Stirling numbers, such as \( S_{x+1,a}^{y+1}/S_{x,a}^{y} \), can be computed quickly via caching (Buntine and Hutter, 2012).

Sampling a new \( z_{dn} = k \) corresponds to incrementing the counts for variable \( \theta_d \), that is, “+1” to \( e_k \) and possibly also “+1” to \( t_{\theta_d}^k \). If \( t_{\theta_d}^k \) is incremented, then \( e_k \) will be incremented too but \( t_{\theta_d}^k \) may or may not be, as dictated by the sampled indicators \( u_k \). The process is repeated until the root \( \mu \), since \( \mu \) is GEM distributed, incrementing \( t_{\mu}^k \) is equivalent to sampling a new topic, i.e. the number of topics increase by 1. Procedure on the vocabulary side (\( \phi \) etc.) is similar.

### 5.2. Metropolis-Hastings Algorithm for Citation Network

We propose a novel MH algorithm that allows the probability vectors to remain integrated out, thus retaining the fast discrete sampling procedure for the PYP and GEM hierarchy, rather than, for instance, resorting to an expectation-maximisation (EM) algorithm or variational approach. We introduce an auxiliary variable \( y_{ij} \), named citing topic, to denote the topic that prompts publication \( i \) to cite publication \( j \). To illustrate, for a biology publication that cites a machine learning publication for the learning technique, the citing topic would be ‘machine learning’ instead of ‘biology’. From Equation 1, a citing topic \( y_{ij} \) is jointly Poisson with \( x_{ij} \):

\[
x_{ij}, y_{ij} = k| \lambda, \theta' \sim \text{Poisson} \left( \lambda_i^+ \lambda_j^- \lambda_k^T \theta_{ik} \theta_{jk}' \right).
\]

Incorporating \( Y \), the set of all \( y_{ij} \), we rewrite the citation network likelihood as

\[
p(X, Y | \lambda, \theta') \propto \prod_i (\lambda_i^+ \theta_i^+) \prod_k (\lambda_k^- \theta_k^-) \prod_k (\lambda_k^T \frac{1}{T} \sum_{i,j} x_{ij} \theta_{ik}^t \theta_{jk}^t \exp \left( - \sum_{ij} \lambda_{ij}^+ \lambda_{ij}^- \lambda_{ij}^T \theta_{ij}^t \theta_{ij}^t \right)
\]

where \( h_{ik} = \sum_j x_{ij} I(y_{ij} = k) + \sum_j x_{ji} I(y_{ji} = k) \) is the number of connections publication \( i \) made due to topic \( k \).

To integrate out \( \theta' \), we note the term \( \theta_{ik}^t \) appears like a multinomial likelihood, so we absorb them into the likelihood for \( p(Z, W, T, C | \zeta) \) where they correspond to additional counts for \( \theta_{ik}^t \), with \( h_{ik} \) added to \( c_{ik}^t \). To disambiguate the source of the counts, we will refer these customer counts contributed by \( x_{ij} \) as network counts, and denote the augmented counts (\( C \) plus network counts) as \( C^+ \). For the exponential term, we use Delta method approximation, \( \int f(\theta) \exp(-g(\theta)) \, d\theta \approx \exp(-g(\hat{\theta})) \int f(\theta) \, d\theta \), where \( \hat{\theta} \) is the expected value.
Bibliographic Analysis with the Citation Network Topic Model

according to a distribution proportional to \( f(\theta) \). This approximation is reasonable as long as the terms in the exponential are small (see supplementary material). The approximate full posterior of CNTM can then be written as

\[
p(\mathbf{Z}, \mathbf{W}, T, C, X, Y | \lambda, \zeta) \approx p(\mathbf{Z}, \mathbf{W}, T, C^+ | \zeta) \prod_i (\lambda_i^+) g_i^+ (\lambda_i^-)^g_i \prod_k (\lambda_k^T)^{\frac{1}{2} \sum_i h_{ik}} \exp \left( - \sum_{ij} \lambda_i^+ \lambda_j^- \lambda_k^T \hat{\theta}_{ijy} \hat{\theta}_{ijy} \right) \quad (7)
\]

The MH algorithm can be summarised in three steps: estimate the document topic prior \( \theta' \), propose a new citing topic \( y_{ij} \) from Equation 6, and accept or reject the proposed \( y_{ij} \) following an MH scheme with Equation 7. We present the details of the MH sampler in the supplementary material. Note that the MH algorithm is similar to the collapsed Gibbs sampler, where we decrement the counts, sample a new state and update the counts. Since all probability vectors are represented as counts, we do not need to deal with their vector form in the collapsed Gibbs sampler. Additionally, our MH algorithm is intuitive and simple to implement. Like the words in a document, each citation is assigned a topic, hence the words and citations can be thought as voting to determine a documents’ topic.

5.3. Hyperparameter Sampling

Hyperparameter sampling for the priors are important (Wallach et al., 2009). In our inference algorithm, we sample the concentration parameters \( \beta \) of all PYPs with an auxiliary variable sampler (Teh, 2006a)\(^3\), but leaving the discount parameters \( \alpha \) fixed. We do not sample the \( \alpha \) due to the coupling of the parameter with the Stirling numbers cache.

In addition to the PYP hyperparameters, we also sample \( \lambda^+ \), \( \lambda^- \) and \( \lambda^T \) with a Gibbs sampler. We let the hyperpriors for \( \lambda^+ \), \( \lambda^- \) and \( \lambda^T \) to be Gamma distributed with shape \( \epsilon_0 \) and rate \( \epsilon_1 \). With the conjugate Gamma prior, the posteriors for \( \lambda^+ \), \( \lambda^- \) and \( \lambda^T \) are also Gamma distributed, so they can be sampled directly.

\[
\begin{align*}
(\lambda^+_i | \mathbf{X}, \lambda^-, \lambda^T \theta') & \sim \text{Gamma} \left( \epsilon_0 + g^+_i, \epsilon_1 + \sum_k \lambda^T_{ik} \hat{\theta}^i_{ik} \sum_j \lambda^-_{ij} \hat{\theta}^j_{jk} \right), \\
(\lambda^-_i | \mathbf{X}, \lambda^+, \lambda^T \theta') & \sim \text{Gamma} \left( \epsilon_0 + g^-_i, \epsilon_1 + \sum_k \lambda^T_{ik} \hat{\theta}^i_{ik} \sum_j \lambda^+_{ij} \hat{\theta}^j_{jk} \right), \\
(\lambda^T_k | \mathbf{Y}, \lambda^+, \lambda^-, \theta') & \sim \text{Gamma} \left( \epsilon_0 + \frac{1}{2} \sum_i h_{ik}, \epsilon_1 + \lambda^T_{ik} (\sum_j \lambda^-_{ij} \hat{\theta}^j_{jk} + \sum_j \lambda^+_{ij} \hat{\theta}^j_{jk}) \right).
\end{align*}
\]

We summarise the full inference algorithm for the CNTM in Algorithm 1.

6. Data

We perform our experiments on subsets of CiteSeer\(^X\) data\(^4\) which consists of scientific publications. Each publication from CiteSeer\(^X\) is accompanied by title, abstract, keywords, authors, citations and other metadata. We prepare three publication datasets from CiteSeer\(^X\) for evaluations. The first dataset corresponds to Machine Learning (ML) publications.

\(3\). We outline the hyperparameter sampling for concentration parameters in the supplementary material.

\(4\). http://citeseerx.ist.psu.edu/
Algorithm 1 Inference Algorithm for the Citation Network Topic Model

1. Initialise the model by assigning a random topic assignment \( z_{dn} \) to each word \( w_{dn} \) and constructing the relevant customer counts \( c^N_k \) and table counts \( t^N_k \) for all variables \( N \).

2. For each word \( w_{dn} \) in each document \( d \):
   i. Decrement the counts associated with \( z_{dn} \) and \( w_{dn} \).
   ii. Blocked sample a new topic \( z_{dn} \) and associated \( T \) and \( C \) from Equation 5.

3. For each citation \( x_{ij} \):
   i. Decrement the network counts associated with \( x_{ij} \) and \( y_{ij} \).
   ii. Sample a new citing topic \( y_{ij} \) from the joint posterior of Equation 6.
   iii. Accept or reject the sampled \( y_{ij} \) with an MH scheme using Equation 7.

4. Update the hyperparameters \( \beta, \lambda^+, \lambda^- \) and \( \lambda^T \).

5. Repeat steps 2-4 until the model converges or a fix number of iterations reached.

which are queried from CiteSeer\(^X\) using the keywords from Microsoft Academic Search\(^5\). The ML dataset contains 139,227 publications.

Our second dataset corresponds to publications from 10 distinct research areas: agriculture, archaeology, biology, computer science, financial economics, industrial engineering, material science, petroleum chemistry, physics and social science. The query words for these 10 disciplines are chosen such that the publications form distinct clusters. We name this dataset M10 (Multidisciplinary 10 classes), which is made of 10,310 publications. For the third dataset, we query publications from both arts and science disciplines. Arts publications are made of history and religion publications, while the science publications contain physics, chemistry and biology researches. This dataset consists of 18,720 publications and is named AvS (Arts versus Science) in this paper.

The keywords used to create the datasets are obtained from Microsoft Academic Search, and are listed in the supplementary material. For the clustering evaluation in Section 7.1.2, we treat the query categories as the ground truth. However, publications that span multiple disciplines can be problematic for clustering evaluation, hence we simply remove the publications that satisfy the queries from more than one discipline. Nonetheless, the labels are inherently noisy. The metadata for the publications can also be noisy, for instance, the authors field may sometimes display publication’s keywords instead of the authors, publication title is sometimes an URL, and table of contents can be mistakenly parsed as the abstract. We discuss our treatments to these issues in Section 6.1. We also note that non-English publications are discarded using \texttt{langid.py} (Lui and Baldwin, 2012).

In addition to the manually queried datasets, we also make use of existing datasets from LINQS (Sen et al., 2008)\(^6\) to facilitate comparison with existing work. In particular, we use their CiteSeer, Cora and PubMed datasets. Their CiteSeer data consists of Artificial Intelligence publications and hence we name the dataset AI in this paper. Although these datasets are small, they are fully labelled and thus useful for clustering evaluation. However,

\(^5\) http://academic.research.microsoft.com/
\(^6\) http://linqs.cs.umd.edu/projects/projects/lbc/
they do not come with additional metadata such as the authors. Note that the AI and Cora
datasets are presented as Boolean matrices, i.e. the word counts information is lost and
all words in a document are assumed to occur only once. Although this representation is
less useful for topic modelling, we still use them for the sake of comparison. Also note that
the word counts were converted to TF-IDF in the PubMed dataset, so we recover the word
counts using a reasonable assumption, see supplementary material for the recovery process.
In Table 1, we present a summary of the datasets used in this paper.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Publications</th>
<th>Citations</th>
<th>Authors</th>
<th>Vocabulary</th>
<th>Words/Doc</th>
<th>%Repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ML</td>
<td>139227</td>
<td>1105462</td>
<td>43643</td>
<td>8322</td>
<td>59.4</td>
<td>23.3</td>
</tr>
<tr>
<td>2. M10</td>
<td>10310</td>
<td>77222</td>
<td>6423</td>
<td>2956</td>
<td>57.8</td>
<td>24.3</td>
</tr>
<tr>
<td>3. AvS</td>
<td>18720</td>
<td>54601</td>
<td>11898</td>
<td>4770</td>
<td>58.9</td>
<td>17.0</td>
</tr>
<tr>
<td>4. AI</td>
<td>3312</td>
<td>4608</td>
<td>–</td>
<td>3703</td>
<td>31.8</td>
<td>–</td>
</tr>
<tr>
<td>5. Cora</td>
<td>2708</td>
<td>5429</td>
<td>–</td>
<td>1433</td>
<td>18.2</td>
<td>–</td>
</tr>
<tr>
<td>6. PubMed</td>
<td>19717</td>
<td>44335</td>
<td>–</td>
<td>4209</td>
<td>67.6</td>
<td>40.1</td>
</tr>
</tbody>
</table>

Table 1: Summary of the datasets used in the paper, showing the number of publications,
citations, authors, unique word tokens, the average number of words in each doc-
ument, and the last column is the average percentage of unique words repeated in
a document. Note: author information is not available on the last three datasets.

6.1. Data Noise Removal

Here, we briefly discuss the steps taken in cleansing the noise from the CiteSeerX datasets
(ML, M10 and AvS). Note that the keywords field in the publications are often empty and
are sometimes noisy, that is, they contain irrelevant information such as section heading
and title, which makes the keywords unreliable source of information as categories. Instead,
we simply treat the keywords as part of the abstracts. We also remove the URLs from the
data since they do not provide any additional useful information.

Moreover, the author information is not consistently presented in CiteSeerX. Some of
the authors are shown with full name, some with first name initialised, while some others
are prefixed with title (Prof, Dr. etc.). We thus standardise the author information by
removing all title from the authors, initialising all first names and discarding the middle
names. Although standardisation allows us to match up the authors, it does not solve the
problem that different authors who have the same initial and last name are treated as a single
author. For example, both Bruce Lee and Brett Lee are standardised to B Lee. Note this
corresponds to a whole research problem (Han et al., 2004, 2005) and hence not addressed in
this paper. Occasionally, institutions are mistakenly treated as authors in CiteSeerX data,
example includes American Mathematical Society and Technische Universität München. In
this case, we simply remove the incorrect authors using a list of exclusion words7 for authors.

6.2. Text Preprocessing

Here, we discuss the preprocessing pipeline adopted for the queried datasets (LINQS data
were already processed). First, since publication text contains many technical terms that

7. The list of exclusion words is presented in the supplementary material.
are made of multiple words, we tokenise the text using phrases (or collocations) instead of unigram words. Thus, phrases like decision tree are treated as single token rather than two distinct words. The phrases are extracted from the respective datasets using LingPipe\(^8\).

In this paper, we use the word \textit{words} to mean both unigram words and phrases.

We then change all the words to lower case and filter out certain words. Words that are removed are \textit{stop words}, common words and rare words. More specifically, we use the stop words list from MALLET\(^9\), we define common words as words that appear in more than 18\% of the publications, and rare words are words that occur less than 50 times in each dataset. Note that the threshold are determined by inspecting the words removed. Finally, the tokenised words are stored as arrays of integers. We also split the datasets to 90\% training set for training the topic models, and 10\% test set for evaluations detailed in Section 7.

7. Experiments

In this section, we describe experiments that compare the CNTM against several baseline topic models. The baselines are HDP-LDA with burstiness (Buntine and Mishra, 2014), a non-parametric extension of the ATM, the Poisson mixed-topic link model (PMTLM) (Zhu et al., 2013) and the CNTM without the citation network. We evaluate these models quantitatively with goodness-of-fit and clustering measures. We qualitatively analyse the topics produced and perform topic analysis on the authors. Additionally, we experiment on merging authors who have low number of publications and grouping them based on categories. This gives us a semi-supervised topic modelling in which some labels are known for authors who do not publish much. Finally, we present a discussion on the algorithm running time and convergence analysis in the supplementary material.

In the following experiments, we initialise the concentration parameters \(\beta\) of all PYPs to 0.1, noting that the hyperparameters are updated automatically. We set the discount parameters \(\alpha\) to 0.7 for all PYPs corresponding to the \textit{“word”} side of the CNTM (\textit{i.e.} \(\gamma, \phi, \phi'\)). This is to induce power-law behaviour on the word distributions. We simply fix the \(\alpha\) to 0.01 for all other PYPs. Note that the number of topics grow with data in non-parametric topic modelling. To prevent the learned topics to be too fine-grained, we set a limit to the maximum number of topics that can be learned. In particular, we set the number of topics cap to 20 for the ML dataset, 50 for M10 and 30 for the AvS dataset. For all the topic models, our experiments find that the number of topics always converges to the cap. For AI, Cora and PubMed datasets, we fix the number of topics to 6, 7 and 3 respectively simply for comparison against PMTLM.

When training the topic models, we run the inference algorithm for 2,000 iterations. For the CNTM, the MH algorithm for the citation network is performed after 1,000 iterations, this is so the topics can be learned first. This gives a faster learning algorithm and also allows us to assess the \textit{“value-added”} by the citation network to topic modelling\(^{10}\). We repeat each experiment five times to reduce the estimation error of the evaluation measures.

---

8. \url{http://alias-i.com/lingpipe/}
9. \url{http://mallet.cs.umass.edu/}
10. This is elaborated further in the supplementary material with likelihood comparison.
7.1. Quantitative Results

7.1.1. Goodness-of-fit and Perplexity

Perplexity is a popular metric used to evaluate the goodness-of-fit of a topic model. Perplexity is negatively related to the likelihood of the observed words given the model, and lower is better. Perplexity, estimated using document completion, is given as:

$$\text{perplexity}(W) = \exp \left( - \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \log p(w_{dn} | \theta_d, \phi)}{\sum_{d=1}^{D} N_d} \right),$$

where $p(w_{dn} | \theta_d, \phi)$ is obtained by summing over all possible topics:

$$p(w_{dn} | \theta_d, \phi) = \sum_k p(w_{dn} | z_{dn} = k, \phi_k) p(z_{dn} = k | \theta_d) = \sum_k \phi_{kw_{dn}} \theta_{dk}.$$

The topic distribution $\theta$ is unknown for the test documents. Instead of using half of the text in the test set to estimate $\theta$, which is a standard practice, we used only the words from the title to estimate $\theta$. One of the reasons behind this is that although title is usually short, it is a good indicator of topic. Moreover, using only the title allows more words to be used to calculate the perplexity. The technical details on estimating $\theta$ is presented in the supplementary material. Note that the perplexity estimate is unbiased since the data used in estimating $\theta$ is not used for evaluation.

We present the perplexity result in Table 2, which clearly shows the significantly better performance of CNTM against the baselines. Inclusion of citation information also provides significant improvement for model fitting, as shown in the comparison of CNTM with and without network component.

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>Test</th>
<th>M10</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bursty HDP-LDA</td>
<td>4904.24 ± 71.34</td>
<td>4992.94 ± 65.57</td>
<td>1959.36 ± 32.77</td>
<td>2265.18 ± 68.19</td>
</tr>
<tr>
<td>Non-parametric ATM</td>
<td>2238.19 ± 12.22</td>
<td>2460.28 ± 11.34</td>
<td>1562.85 ± 18.11</td>
<td>1814.03 ± 23.18</td>
</tr>
<tr>
<td>CNTM w/o network</td>
<td>1918.21 ± 4.31</td>
<td>2057.61 ± 3.56</td>
<td>912.69 ± 10.94</td>
<td>1186.11 ± 8.32</td>
</tr>
<tr>
<td>CNTM w network</td>
<td>1851.82 ± 8.50</td>
<td>1990.78 ± 11.36</td>
<td>824.04 ± 11.96</td>
<td>1048.33 ± 21.39</td>
</tr>
</tbody>
</table>

Table 2: Perplexity for the train and test documents on ML and M10, lower is better.

7.1.2. Document Clustering

Next, we evaluate the clustering ability of the topic models. Recall that topic models assign a topic to each word in a document, essentially performing a soft clustering in which the membership is given by the document-topic distribution $\theta$. For the following evaluation, we convert the soft clustering to hard clustering by choosing a topic that best represents the documents, hereafter called the dominant topic. The dominant topic corresponds to the topic that has the highest probability in a topic distribution $\mathcal{N}$.

11 In this paper, significance is quantified at 5% significance level.
Table 3: Comparison of clustering performance on the M10 and AvS dataset.

<table>
<thead>
<tr>
<th></th>
<th>M10 Purity</th>
<th>M10 NMI</th>
<th>AvS Purity</th>
<th>AvS NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bursty HDP-LDA</td>
<td>0.66 ± 0.02</td>
<td>0.67 ± 0.01</td>
<td>0.75 ± 0.03</td>
<td>0.66 ± 0.01</td>
</tr>
<tr>
<td>Non-parametric ATM</td>
<td>0.58 ± 0.01</td>
<td>0.63 ± 0.00</td>
<td>0.69 ± 0.02</td>
<td>0.64 ± 0.01</td>
</tr>
<tr>
<td>CNTM w/o network</td>
<td>0.61 ± 0.04</td>
<td>0.67 ± 0.01</td>
<td>0.72 ± 0.03</td>
<td>0.66 ± 0.01</td>
</tr>
<tr>
<td>CNTM w network</td>
<td>0.67 ± 0.03</td>
<td>0.69 ± 0.02</td>
<td>0.72 ± 0.01</td>
<td>0.66 ± 0.00</td>
</tr>
</tbody>
</table>

As mentioned in Section 6, we assume the ground truth classes correspond to the query categories used in creating the datasets. We evaluate the clustering performance with purity and normalised mutual information (NMI)\(^\text{12}\) (Manning et al., 2008). Purity is a simple clustering measure which can be interpreted as the proportion of documents correctly clustered. For ground truth classes \(S = \{s_1, \ldots, s_J\}\) and obtained clusters \(R = \{r_1, \ldots, r_K\}\), the purity and NMI are computed as

\[
\text{purity}(S, R) = \frac{1}{D} \sum_k \max_j \left| r_k \cap s_j \right| ,
\]

\[
\text{NMI}(S, R) = \frac{2 I(S; R)}{H(S) + H(R)} ,
\]

where \(I(S; R)\) denotes the mutual information and \(H(\cdot)\) denotes the entropy:

\[
I(S; R) = \sum_{k,j} \frac{|r_k \cap s_j|}{D} \log_2 \frac{D |r_k \cap s_j|}{|r_k||s_j|} ,
\]

\[
H(R) = -\sum_k \frac{|r_k|}{D} \log_2 \frac{|r_k|}{D} .
\]

The clustering results are presented in Table 3 and Table 4. We can see that the CNTM greatly outperforms the PMTL in NMI evaluation. Note that for a fair comparison against PMTL, the experiments on the AI, Cora and PubMed datasets are evaluated with a 10-fold cross validation. Additionally, we would like to point out that since no author information is provided on these 3 datasets, the CNTM becomes a variant of HDP-LDA, but with PYP instead of DP. We find that the clustering performance of CNTM with or without network is similar in Table 4. This is likely because the publications in each dataset are highly related to one another\(^\text{13}\), and thus the citation information is not discriminating enough for clustering.

7.2. Author-merging for Semi-supervised Learning

Author modelling allows topic sharing of multiple documents written by the same author. However, there are many authors who have authored only a few publications, thus their treatment can be problematic. In this section, we experiment on merging these authors into groups to improve document clustering. We merge authors who have authored less than \(\eta\) publications, to clarify, \(\eta = 2\) means authors who have only a single publication are merged, while \(\eta = 1\) corresponds to no merging. Additionally, we use the category labels for

\(^{12}\) Note that the NMI in Zhu et al. (2013) is slightly different to ours, we use the definition in Manning et al. (2008). This penalises our NMI result when compared against the result in Zhu et al. (2013) since our normalising term will always be equal or greater than that of Zhu et al. (2013).

\(^{13}\) See the list of category labels of these datasets in supplementary material.
Bibliographic Analysis with the Citation Network Topic Model

<table>
<thead>
<tr>
<th></th>
<th>AI</th>
<th>Cora</th>
<th>PubMed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purity</td>
<td>N/A</td>
<td>0.51</td>
<td>N/A</td>
</tr>
<tr>
<td>NMI</td>
<td>0.41</td>
<td>0.63</td>
<td>0.27</td>
</tr>
<tr>
<td>PMTLM*</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>CNTM w/o network</td>
<td>0.51 ± 0.07</td>
<td>0.67 ± 0.02</td>
<td>0.37 ± 0.03</td>
</tr>
<tr>
<td>CNTM w network</td>
<td>0.51 ± 0.08</td>
<td>0.66 ± 0.02</td>
<td>0.39 ± 0.03</td>
</tr>
</tbody>
</table>

Table 4: Comparison of clustering performance of CNTM against PMTLM. The best PMTML results are chosen for comparison, from Table 2 in Zhu et al. (2013).

We present the clustering results for \( \eta = \{2, 3, 4, 5\} \) as a plot in Figure 2 (results in table format are shown in the supplementary material). We find that increasing \( \eta \) generally improves the clustering performance, although the effect is not too significant for successive \( \eta \). Note that if \( \eta \) is set to be too large, most of the author information will be replaced by the category labels, which defeats the purpose of author modelling.

7.3. Qualitative Analysis

We can obtain a summary of a text corpus from a trained CNTM, this is done by analysing the topic-word distribution \( \phi \). In Table 5, we display some major topics extracted from the ML dataset (M10 and AvS in supplementary material). The topics are represented by the top words, which are ordered based on \( \phi_{kw} \). The labels of the topics are manually assigned.

Additionally, we analyse the author-topic distributions \( \nu \) to find out about authors’ interests. We focus on the M10 dataset since it covers a wider area of research topics. For each author \( a \), we determine their dominant topic from their author-topic distribution \( \nu_a \).

We display the interests of some authors in Table 6. Again, the topic labels are manually picked given the dominant topics and the corresponding top words from the topics.


8. Conclusions

In this paper, we have proposed the Citation-Network Topic Model (CNTM) to jointly model research publications and their citation network. CNTM performs text modelling with a hierarchical PYP topic model and models the citations with the Poisson distribution. We also proposed a novel learning algorithm for the CNTM, which exploits the conjugacy of the Dirichlet and Multinomial distribution, allowing the sampling of the citation networks to be of similar form of the collapsed Gibbs sampler of a topic model. As discussed, our learning algorithm is intuitive and easy to implement.

The CNTM offers substantial performance improvement over previous work (Zhu et al., 2013). On three CiteSeerX datasets and three existing datasets, we demonstrate the improvement of joint topic and network modelling in terms of model fitting and clustering evaluation. Additionally, we experiment on merging authors who do not have many publications into groups of similar authors based on the query categories, giving us a semi-supervised learning. We find that clustering performance improves with the level of merging.

Future work includes learning the influences of the co-authors, utilising them for author merging and further speed up non-parametric modelling with techniques in Li et al. (2014).

Acknowledgments

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. The authors wish to thank CiteSeerX for providing the data.

Table 5: Topic Summary for ML Dataset

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinforcement Learning</td>
<td>reinforcement, agents, control, state, task</td>
</tr>
<tr>
<td>Object Recognition</td>
<td>face, video, object, motion, tracking</td>
</tr>
<tr>
<td>Data Mining</td>
<td>mining, data mining, research, patterns, knowledge</td>
</tr>
<tr>
<td>SVM</td>
<td>kernel, support vector, training, clustering, space</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>recognition, speech, speech recognition, audio, hidden markov</td>
</tr>
</tbody>
</table>

Table 6: Example of authors and their topic preference learned by the CNTM.

<table>
<thead>
<tr>
<th>Author</th>
<th>Topic</th>
<th>Top Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Aerts</td>
<td>Quantum Theory</td>
<td>quantum, theory, quantum mechanics, classical, quantum field</td>
</tr>
<tr>
<td>Y. Bengio</td>
<td>Neural Network</td>
<td>networks, learning, recurrent neural, neural networks, models</td>
</tr>
<tr>
<td>C. Boutilier</td>
<td>Decision Making</td>
<td>decision making, agents, decision, theory, agent</td>
</tr>
<tr>
<td>S. Thrun</td>
<td>Robot Learning</td>
<td>robot, robots, control, autonomous, learning</td>
</tr>
<tr>
<td>M. Baker</td>
<td>Financial Market</td>
<td>market, risk, firms, returns, financial</td>
</tr>
</tbody>
</table>

Furthermore, we can graphically visualise the author-topics network extracted by CNTM with Graphviz\(^\text{14}\)\. This is detailed in the supplementary material due to space.

### 8. Conclusions

In this paper, we have proposed the Citation-Network Topic Model (CNTM) to jointly model research publications and their citation network. CNTM performs text modelling with a hierarchical PYP topic model and models the citations with the Poisson distribution. We also proposed a novel learning algorithm for the CNTM, which exploits the conjugacy of the Dirichlet and Multinomial distribution, allowing the sampling of the citation networks to be of similar form of the collapsed Gibbs sampler of a topic model. As discussed, our learning algorithm is intuitive and easy to implement.

The CNTM offers substantial performance improvement over previous work (Zhu et al., 2013). On three CiteSeerX datasets and three existing datasets, we demonstrate the improvement of joint topic and network modelling in terms of model fitting and clustering evaluation. Additionally, we experiment on merging authors who do not have many publications into groups of similar authors based on the query categories, giving us a semi-supervised learning. We find that clustering performance improves with the level of merging.

Future work includes learning the influences of the co-authors, utilising them for author merging and further speed up non-parametric modelling with techniques in Li et al. (2014).

Acknowledgments

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. The authors wish to thank CiteSeerX for providing the data.

\(^\text{14}\) \url{http://www.graphviz.org/}
References


