7-2014

Modeling Citation Networks using Latent Random Offsets

Willie Neiswanger
Carnegie Mellon University

Chong Wang
Princeton University

Qirong Ho
Carnegie Mellon University

Eric P. Xing
Carnegie Mellon University, epxing@cs.cmu.edu

Published In
Proceedings of the 30th International Conference on Conference on Uncertainty in Artificial Intelligence (UAI 2014).

Modeling Citation Networks using Latent Random Offsets

Willie Neiswanger willie@cs.cmu.edu
Chong Wang chongw@cs.princeton.edu
Qirong Ho qho@cs.cmu.edu
Eric P. Xing epxing@cs.cmu.edu

Abstract

Out of the many potential factors that determine which links form in a document citation network, two in particular are of high importance: first, a document may be cited based on its subject matter—this can be modeled by analyzing document content; second, a document may be cited based on which other documents have previously cited it—this can be modeled by analyzing citation structure. Both factors are important for users to make informed decisions and choose appropriate citations as the network grows. In this paper, we present a novel model that integrates the merits of content and citation analyses into a single probabilistic framework. We demonstrate our model on three real-world citation networks. Compared with existing baselines, our model can be used to effectively explore a citation network and provide meaningful explanations for links while still maintaining competitive citation prediction performance.

1 Introduction

Many large citation networks—Wikipedia, arXiv, and PubMed[1], to name a few—continue to quickly grow in size, and the structure of these networks continues to increase in complexity. To effectively explore large-scale and complex data like these and extract useful information, users rely more and more on various types of guidance for help. An important type of guidance comes from the citations (or links) in the network. Citations serve as paths that users can easily follow, and do not require users to specify certain keywords in advance. In scientific research, for example, researchers often find potentially interesting articles by following citations made in other articles. In Wikipedia, users often find explanations of certain terms by following the links made by other Wikipedia users. Thus, generating relevant citations is important for many users who may frequently rely on these networks to explore data and find useful information.

We believe that, among many, two important factors largely determine how a document citation network is formed: the documents' contents and the existing citation structure. Take as an example a citation network of computer science articles. A research paper about "support vector machines (SVMs)", for instance, might be cited by several other articles that develop related methods, based on the subject matter alone. This type of information can be well captured by analyzing the content of the documents. However, the existing citation structure is also important. If this SVM paper included great results on a computer vision dataset, for example, it might be cited by many vision papers that are *not* particularly similar in content. Though different in content, this SVM paper could be very important to users in a different topic area, and should be considered by these users when choosing citations. This type of information cannot be easily captured by analyzing document content, but can be discovered by analyzing the existing citation structure among documents while studying the contents of the papers that generated these citations.

Given these observations, we present a probabilistic model to accurately model citation networks by integrating content and citation/link information into a single framework. We name our approach a *latent random offset* (LRO) model. The basic idea is as follows: we first represent the content of each document using a latent vector representation (i.e. "topics") that summarizes the document content. Then, each latent representation is augmented in an additive manner with a random offset vector; this vector models information from the citation structure that is not well captured by document content. The final augmented representa-

[1] http://www.wikipedia.org/, http://arxiv.org/, and http://www.ncbi.nlm.nih.gov/pubmed

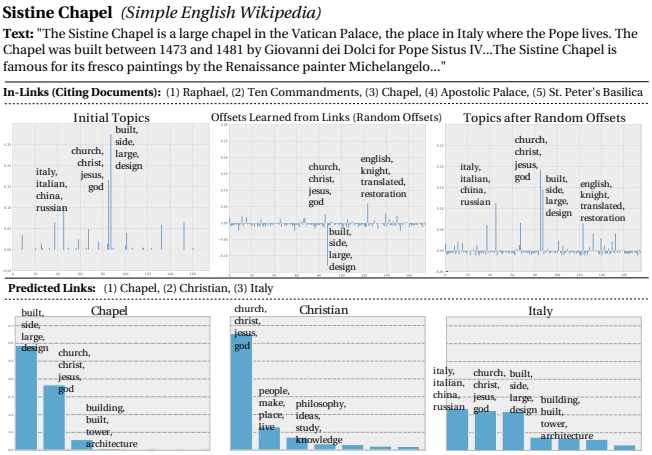

Figure 1: Analysis of content, latent offsets, and predicted links for the *Sistine Chapel* document in the Simple English Wikipedia dataset. The first row shows an example passage from the document. The next row shows the names of the documents that cite *Sistine Chapel*. The next row shows the initial latent topics (first column), the latent offsets learned from links (second column), and the latent topics after applying the offsets (third column). The final row shows interpretable link predictions; for each predicted link, we show the relative weight that each latent topic contributed to the prediction.

tion is then used to model how this document is cited by other documents. To motivate this representation, we present sample outputs from running LRO on the Simple English Wikipedia.

Examples from Simple English Wikipedia. The first graph in the top row of Figure 1 shows, for the *Sistine Chapel* article in the Simple English Wikipedia, the latent vector representation, which is concentrated around three topics: `countries` (*italy, italian, china, russian*), `Christianity` (*church*, *christ*, *jesus*, *god*), and `architecture` (*built, side, large, design*). Here we've listed the top four words in each topic (in parens). The incoming links to the *Sistine Chapel* article are also shown; these citing documents determine the random offsets for *Sistine Chapel*. The random offsets can be thought of as "corrections" to the latent vector representation, based on the content of citing documents—for example, the two largest positive offsets are `Christianity` (*church, christ, jesus, god*) and `Anglicanism` (*english, knight, translated, restoration*), meaning that the citing documents strongly exhibit these two topics (compared to the *Sistine Chapel* article). On the other hand, there is a large negative offset on `architecture` (*built, side, large, design*), indicating that the citing documents do not exhibit this topic as much as *Sistine Chapel*.

Notably, the topic `Anglicanism` (containing words related to Christianity in England) is found in the random offsets for *Sistine Chapel*, but is absent from its latent vector representation. This is because the Sistine Chapel is in the Vatican City, and thus its article does not emphasize content relating to England or Anglicanism (even though they are all related to Christianity). However, documents that link to *Sistine Chapel*, such as *Chapel*, talk about the Anglican Church in England. This is an example where pertinent information is found in the citation structure, but not in the document content. By capturing this citation information, the LRO model provides insights into the context surrounding a document.

Following this idea, we can add the latent vector and random offsets together to obtain the "augmented representation" of a document (i.e. the "topics after random offsets" graph in Figure 1), which takes into account not just its content, but the content of its citing documents as well. Link predictions in the LRO model are based upon the intuition that a document i cites document j only if both documents have similar representations. This intuition is captured in the bottom row of graphs in Figure 1, which explains three out-links predicted by the LRO model for the *Sistine Chapel* document. For each predicted link, we show the topics that contributed most to the prediction, and not surprisingly, the most important topics for each link also feature strongly in the augmented representation for the *Sistine Chapel*. Knowing which topics contributed to the prediction of links not only helps users interpret existing links within a document corpus, but also gives users an explanation for every new link predicted by the LRO model—for instance, a user might invoke LRO to recommend citations for an academic paper, and such "link explanations" give the user a quick overview of why each recommendation is relevant.

We note that of the three predicted out-links for *Sistine Chapel*, two of them (*Chapel*, *Italy*) are actual out-links in *Sistine Chapel*, while the third, *Christian*, is obviously relevant but not found in the document. This motivates another application of LRO: predicting relevant but missing links in document corpora; in this case, we are completing the references for a Wikipedia article. Another application context is academic paper writing: LRO can be used to recommend important (but otherwise overlooked) citations for a newly-written academic paper.

The rest of this paper is organized as follows: we begin by formalizing latent random offset modeling, and then show how we can use it to model citation networks. We then develop a fast learning algorithm with linear complexity in the size of the number of citations, and empirically evaluate our approach using

three real-world citation networks. Compared with several baselines, our model not only improves citation prediction performance, but also provides meaningful explanations for citations within the networks. By studying latent random offset representations, we show these explanations can be used to effectively interpret why our model predicts links for given documents and to explore citation networks.

2 Latent Random Offset Models

We introduce the general framework of latent random offsets for citation network modeling. Suppose our citation network consists of D documents (i.e. nodes), $\mathcal{D} = \{x_1, x_2, ..., x_D\}$. We use $y_{ij} = 1$ or 0 to indicate whether document i cites document j or not. Note that y_{ij} is *directed*, meaning y_{ij} is not necessarily the same as y_{ji}.

Each document x_j is usually a high-dimensional vector in $\mathbb{R}^V$, where V is the vocabulary size, so it is desirable to represent x_j using a low-dimensional vector θ_j. In other words, the mapping

$$\theta_j = \theta_j(x_j) \tag{1}$$

serves as a summarization of the original document content x_j, and these summarizations can be used to measure the content similarities of different documents.

However, in real citation networks, a document can be cited by others for reasons outside of its content information. For example, a target document might provide an influential idea that can be used in many different fields and thus be cited by a diverse set of documents. This information is encoded not in the document content but in the citation network structure. We choose to model this phenomenon by allowing a random offset vector ϵ_j to augment the low-dimensional vector θ_j, which gives the augmented representation

$$v_j = \theta_j + \epsilon_j. \tag{2}$$

The offset vector ϵ_j is used to capture the network structure information that is *not* contained in the document's content. One important property of this augmented representation is that the random offset ϵ_j is aligned in the same space as θ_j. If the dimension of θ_j has some semantic explanations, then ϵ_j can be understood as modifications of those explanations.

Finally we consider using a function f to model the citation from document i to document j, such that

$$f(\theta_i, \theta_j + \epsilon_j) \approx y_{ij} \ \ \text{(for all } i, j)$$

where y_{ij} is the citation indicator from document i to document j. Notice the asymmetric structure here for document i and j—we do not consider the offset vector ϵ_i for document i in our function f. In real citation networks, when a new document joins the citation network by citing some other documents, this new document is effectively "not in" the network. It will be most likely to cite other documents based only on their content and their citations, as no network information exists for this new document. One advantage of this formulation is that we can make citation predictions for a brand new document by only using its content information.

In the next two sections, we first describe how we create the low-dimensional document content representation θ_j and how we use the latent random offset model for citation network modeling.

2.1 Probabilistic topic models for document content representation

There are many potential ways to create the low-dimensional document content representation described in Eq. 1. Here we choose to use probabilistic topic models. Topic models [5] are used to discover a set of "topics" (or themes) from a large collection of documents. These topics are distributions over terms, which are biased to be associated under a single theme. One notable property of these models is that they often provide an interpretable low-dimensional representation of the documents [10]. They have been used for tasks like corpus exploration [8], information retrieval [23] and recommendation [22].

Here we describe the simplest topic model, latent Dirichlet allocation (LDA) [7] and use it to create the low-dimensional document content representations. Assume there are K topics, β_k, $k = 1, ..., K$ and each β_k is a distribution over a fixed vocabulary. For each document j, the generative process is as follows,

1. Draw topic proportions $\theta_j \sim \text{Dirichlet}(\alpha)$
2. For each word x_{jn} in document j,
   (a) Draw topic assignment $z_{jn} \sim \text{Mult}(\theta_j)$
   (b) Draw word $x_{jn} \sim \text{Mult}(\beta_{z_{jn}})$

This process describes how the words of a document are generated from a mixture of topics that are shared by the corpus. The topic proportions θ_j are document-specific and we use these topic proportions as our low-dimensional document content representation.

Given a document collection, the only observations are the words in the documents. The topics, topic proportions for each document, and topic assignments for each word, are all latent variables that have to be determined from the data. LDA has been extensively studied in the literature and many efficient algorithms

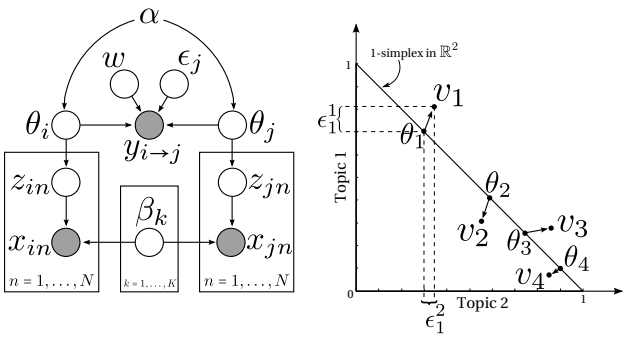

Figure 2: Left: The LRO graphical model. Only two documents (i and j) and one citation (from i to j) are shown. The augumented latent representation representation for document j is $v_j = \theta_j + \epsilon_j$. Right: An illustration of the random offsets. We show each document's content vector θ_j (which lies on the simplex), its offsets ϵ_j due to link structure (the superscript indicates the dimension for ϵ_j), and the resulting augmented latent representation v_j.

have been proposed to fit the LDA model variables [7, 12, 21]. For example, standard learning algorithms like variational EM or Gibbs sampling can be used to estimate these quantities [7]. These methods give us the estimated document content representations θ_j in terms of an approximate posterior distribution or point estimates.

2.2 Modeling citations via random offsets

Having described how we represent the documents in a low dimensional space, we now consider how to create the augmented representations introduced in Eq. 2. We model our latent random offset vector ϵ_j with a multivariate Gaussian distribution

$$\epsilon_j \sim \mathcal{N}(0, \lambda^{-1} I_K).$$

where λ is a scalar precision parameter for the latent random offsets.

Using the general idea of latent random offset modeling shown in Eq. 2 and probabilistic topic models described in Section 2.1, our *latent random offset model* (LRO) for citation network modeling has the following generative process (Figure 2 shows the graphical model). Assuming K topics, $\beta_{1:K}$,

1. For each document j,
   (a) Draw topic proportions $\theta_j \sim \text{Dirichlet}(\alpha)$
   (b) Draw latent random offset $\epsilon_j \sim \mathcal{N}(0, \lambda^{-1} I_K)$ and set the document augmented representation as $v_j = \theta_j + \epsilon_j$
   (c) For each word x_{jn},
      i. Draw topic assignment $z_{jn} \sim \text{Mult}(\theta)$
      ii. Draw word $x_{jn} \sim \text{Mult}(\beta_{z_{jn}})$
2. For each *directed* pair of documents (i, j), draw the citation indicator

$$y_{ij} \sim \mathcal{N}(y | w\theta_i^\top v_j, \tau_{ij}^{-1}).$$

where $w \in \mathbb{R}_+$ is a global scaling parameter to account for potential inefficiencies of the topic proportions θ_i, which are constrained to the simplex.[2] We chose a Gaussian response to model the citations, in similar fashion to [22]. Notation τ_{ij}^{-1} is the precision parameter for the Gaussian distribution such that

$$\tau_{ij} = \begin{cases} \tau_1 & \text{if } y_{ij} = 1 \\ \tau_0 & \text{if } y_{ij} = 0 \ . \end{cases}$$

Here, τ_1 specifies the precision if a link exists from document i to j, while τ_0 is for the case where the link does not exist. We set τ_0 to be much smaller (i.e. higher noise) than τ_1 — this is similar to the assumption made in [22], which models the fact that $y_{ij} = 0$ could either mean it is not appropriate for document i to cite document j, or simply that document i should cite document j but has inadvertently neglected to cite it. This also enables a fast learning algorithm with complexity linear in the number of citations (See Section 3 for details).

The expectation of the citation can be computed as

$$\mathbb{E}[y_{ij}] = w\theta_i^\top v_j = w(\theta_i^\top \theta_j) + w(\theta_i^\top \epsilon_j).$$

This reveals how likely it is for a citation from document i to document j to occur under our model. If the documents have similar content or document j has certain large positive offsets, it is more likely to be cited by document i.

For a document j, our latent representation θ_j is over a simplex. In Figure 2 (right), we show how the random offsets ϵ_j produce the augmented representation v_j.

2.3 Citation prediction

In a system for citation prediction, it is more realistic to suggest citations than to make hard decisions for the users. This is common in many recommender systems [13, 22]. For a particular document i, we rank the potential citations according to the score

$$\text{S}_{ij} = w\theta_i^\top v_j,$$

for all other documents j, and suggest citations based on this score (excluding document i and all pre-existing citations).

[2] Our experiments show that optimizing the global scaling parameter w is important for obtaining good results.

3 Learning Algorithm

We use maximum a posteriori (MAP) estimation to learn the latent parameters of the LRO, where we perform a coordinate ascent procedure to carry out the optimization. Maximization of the posterior is equivalent to maximizing the complete log likelihood of $v_{1:D}$, $\theta_{1:D}$ and $\beta_{1:K}$, which we can write as

$$\mathcal{L} = -\frac{\lambda}{2}\sum_j (v_j - \theta_j)^\top (v_j - \theta_j) - \sum_{i \neq j} \frac{\tau_{ij}}{2}(y_{ij} - w\theta_i^T v_j)^2 + \sum_j \sum_n \log\left(\sum_k \theta_{jk}\beta_{k,x_{jn}}\right).$$

where we have omitted a constant and set $\alpha = 1$.

First, given topics $\beta_{1:K}$ and augmented representations $v_{1:D}$, for all documents, we describe how to learn the topic proportions θ_j. We first define $\phi_{jnk} = q(z_{jn} = k)$. Then we separate the items that contain θ_j and apply Jensen's inequality,

$$\begin{aligned}\mathcal{L}(\theta_j) &\geq -\frac{\lambda}{2}\sum_j (v_j - \theta_j)^\top (v_j - \theta_j) \\ &\quad + \sum_n \sum_k \phi_{jnk}\left(\log \theta_{jk}\beta_{k,x_{jn}} - \log \phi_{jnk}\right) \\ &= \mathcal{L}(\theta_j, \boldsymbol{\phi}_j).\end{aligned}$$

where $\boldsymbol{\phi}_j = (\phi_{jnk})_{n=1,k=1}^{D\times K}$. The optimal ϕ_{jnk} then satisfies

$$\phi_{jnk} \propto \theta_{jk}\beta_{k,x_{jn}}.$$

The $\mathcal{L}(\theta_j, \boldsymbol{\phi}_j)$ gives the *tight* lower bound of $\mathcal{L}(\theta_j)$. We cannot optimize θ_j analytically, but we can use the projection gradient [3] method for optimization.[3]

Second, given this $\boldsymbol{\phi}$, we can optimize the topics $\beta_{1:K}$ with

$$\beta_{kx} \propto \sum_j \sum_n \phi_{jnk} 1[x_{jn} = x].$$

This is the same M-step update for topics as in LDA [7].

Next, we would like to optimize the augmented representations $v_{1:D}$. We can write the component of the log likelihood with terms containing v_j as

$$\mathcal{L}(v_j) = -\frac{\lambda}{2}(v_j - \theta_j)^\top (v_j - \theta_j) - \sum_{i,i\neq j} \frac{\tau_{ij}}{2}(y_{ij} - w\theta_i^\top v_j)^2.$$

To maximize this quantity, we take the gradient of $\mathcal{L}(v_j)$ with respect to v_j and set it to 0, which gives an update for v_j

$$v_j^* \leftarrow \left(\lambda I_K + w^2\left((\tau_1 - \tau_0)\sum_{i\in\{i:i\to j\}} \theta_i\theta_j^\top + \tau_0 \sum_{i,i\neq j} \theta_i\theta_j^\top\right)\right)^{-1} \times \left(\theta_j + w\tau_1 \sum_{i\in\{i:i\to j\}} \theta_i\right) \quad (3)$$

where $\{i : i \to j\}$ denotes the set of documents that cite document j. For the second line of Eq. 3, we can see that the augmented representation v_j is affected by two main parts: the first is the content from document j (topic proportions θ_j) and the second is the content from other documents who cite document j (topic proportions θ_i, where $i \in \{i : i \to j\}$).

Next, we want to optimize the global scaling variable w. Isolating the terms in the complete log likelihood that contain w gives

$$\mathcal{L}(w) = -\sum_{i\neq j} \frac{\tau_{ij}}{2}(y_{ij} - w\theta_i^\top v_j)^2.$$

In a similar manner as the previous step, to maximize this quantity we take the gradient of $\mathcal{L}(w)$ with respect to w and set it to 0, which gives its update[4]

$$w^* \leftarrow \left(\sum_j \left((\tau_1 - \tau_0)\sum_{i\in\{i:i\to j\}} (\theta_i^\top v_j)^2 + \tau_0 \sum_{i,i\neq j} (\theta_i^\top v_j)^2\right)\right)^{-1} \times \left(\tau_1 \sum_j \sum_{i\in\{i:i\to j\}} \theta_i^\top v_j\right). \quad (4)$$

Empirically, we found that an optimal trade-off between computation time and performance involves performing LDA [7] initially to learn the latent representations θ_j, and then performing coordinate ascent to learn the augmented representations v_j and global parameter w. We detail this procedure in Algorithm 1.

Computational efficiency. We now show that our learning algorithm (Algorithm 1) has runtime complexity linear in the number of documents and citations.

First, estimating the topic proportions θ_j, $j = 1, ..., D$ has the same complexity as the standard learning algorithm for LDA, which is linear in the number of documents.

Second, the augmented representations v_j, $j = 1, ..., D$ and global scaling parameter w can be estimated in linear time, via a caching strategy — this is similar to the method adopted by [13, 22]. We now describe this strategy.

[3] On our data, we found that simply fixing θ_j as the estimate from the LDA model gives comparable performance and saves computation.

[4] In theory, this update could lead to a negative value. However, in our experiments, we did not see this happen.

Algorithm 1 MAP Parameter Learning

Input: A citation network of documents $\{x_j\}_{j=1}^D$ with directed links y_{ij} for $i, j \in \{1, \ldots, D\}$, and stopping criteria δ

Output: Latent content representations θ_j, link-offset representations v_j, and global scale parameter w

Run LDA [7] on $\{x_j\}_{j=1}^D$ to learn $\theta_{1:D}$
Initialize $v_{1:D} = \theta_{1:D}$ and eps $= \infty$
while eps $> \delta$ **do**
  Update $w \leftarrow w^*$ ▷ Equation 4
  for $j = 1$ **to** D **do**
    Update $v_j \leftarrow v_j^*$ ▷ Equation 3
  end for
  Set eps $\leftarrow \|v_{1:D} - \tilde{v}_{1:D}\|$
end while

For the augmented representation v_j (Eq. 3), we cache $\theta_0 = \sum_i \theta_i$. This allows us to update v_j (Eq. 3) using the identity

$$\textstyle\sum_{i,i\neq j} \theta_i = \theta_0 - \theta_j.$$

Every time we update a θ_j, we also update the cache θ_0, and this takes constant time w.r.t. the number of documents and citations.

For the global scaling parameter w (Eq. 4), we can compute

$$\begin{aligned}\textstyle\sum_{i,i\neq j}(\theta_i^\top v_j)^2 &= \textstyle\sum_{i,i\neq j} v_j^\top \theta_i \theta_i^\top v_j \\ &= v_j^\top (\textstyle\sum_{i,i\neq j} \theta_i \theta_i^\top) v_j \\ &= v_j^\top (\textstyle\sum_i \theta_i \theta_i^\top) v_j - v_j^\top \theta_j \theta_j^\top v_j\end{aligned}$$

in $O(K^2)$ time (constant in the number of docs and citations) by simply caching $\Theta_0 = \sum_i \theta_i \theta_i^\top$. This cache variable also requires $O(K^2)$ time to update whenever we modify some θ_j.

The remaining sums in Eqs 3,4 touch every citation exactly once, therefore a single update sweep over all v_j and w only requires constant work per edge (treating K as constant). We have therefore shown that Algorithm 1 is linear in the number of documents and citations. Moreover, we have attained linear scalability without resorting to treating missing citations as hidden data. This gives our LRO a data advantage over methods that hide missing citations, such as the RTM [9].

4 Related Work

Our proposed work focuses on two aspects of citation network modeling: 1) network understanding/exploration and 2) citation prediction. We therefore divide the related work section into these two categories.

Network understanding/exploration. Network exploration is a broad empirical task concerned with, amongst other things, understanding the overall structure of the network [19], understanding the context of individual nodes [2], and discovering anomalous nodes or edges [20]. In addition to methods that operate on purely graph data, there are techniques that leverage both the graph as well as textual content, such as relational topic models (RTM) [9], Link-PLSA-LDA [17], and TopicFlow [18]. The idea behind such hybrid methods is that text and graph data are often orthogonal, providing complementary insights [11].

Our LRO model incorporates network information by modeling per-document random offsets that capture topical information from connected neighbors. These random offsets represent relevant topics that would otherwise not be found in the documents through content analysis. The Simple English Wikipedia analysis from the introduction provides a good example: the *Sistine Chapel* article's random offsets (the top row of Figure 1) contain the topic `Anglicanism` (which is also related to Christianity), even though the article text's latent topic representation makes no mention of it. In this manner, the LRO model helps us understand the context of network nodes (a.k.a. documents), and helps us to detect anomalous nodes (such as documents whose random offsets diverge greatly from their latent topic vectors).

Citation prediction. The citation prediction task can be approached by considering text features, network features, or a combination of both. In the text-only setting, approaches based on common text features (e.g., TF-IDF scores [4]) and latent space models (e.g., topic models [5]) can be used to the measure similarities between two documents, allowing for ranking and prediction. However, text-only approaches cannot account for citation behavior due to the network structure.

In the network-only setting without document content, there are a number of commonly-used measures of node similarity, such as the Jaccard Coefficient, the Katz measure [14] and the Adamic/Adar measure [1]. Latent space models such as matrix factorization (MF) methods [15] can be used here. However, when test documents are out-of-sample with respect to the network (when we consider newly-written papers with no preexisting citations), these measures are inapplicable.

Finally, there are methods that combine both document content and network structure to predict citations. One such method is the relational topic models (RTM) [9], in which link outcomes depend on a reweighted inner product between latent positions (under the LDA model). The weights are learned for each latent dimension (topic), but are not specific to any document,

and thus only capture network behavior due to topic-level interactions. In contrast, our random offsets are learned on a per-document basis, capturing interaction patterns specific to each document, which in turn yields better predictive performance as shown in our empirical study. In [16], in addition to the document content, author information is also considered to model the citation structure. In [17], citations were treated as a parallel document (of citations) as to the document content of words. Neither of these methods use per-document offsets to model citation structure.

5 Empirical Study

We will empirically demonstrate the use of our model for modeling citation networks. We will first show quantitative results for citation prediction then present qualitative results using our model to explore citation networks.

Datasets. We use three citation network datasets,

1. The ACL Anthology paper citation network (*ACL*) contains 16,589 documents and 94,973 citations over multiple decades.
2. The arXiv high energy physics citation network (*arXiv*) contains 34,546 arXiv/hep-th articles and 421,578 citations from January 1993 through April 2003.
3. The Simple English Wikipedia citation network (*Wikipedia*) contains 27,443 articles, and 238,957 citations corresponding to user-curated hyperlinks between articles.

5.1 Citation prediction

For citation prediction, we compare against the RTM [9], matrix factorization (MF) [15], LDA-based predictions [7], and three common baseline algorithms. A detailed description is given below.

The first task is predicting held-out citations. Here we used a five-fold cross validation: for each document that has cited more than 5 documents, we held out 20% of the documents into test set and the rest into the training set.

The second task is predicting citations for new documents. To simulate this scenario, we train our model using all the citations before a certain year and predict the citations of the new documents published in that year. This task is important for a real citation prediction system, where user may input some text without existing citations. For this experiment, we excluded MF from the comparisons, because it cannot perform this task.

Evaluation metric. Our goal is to make citation predictions, where it is more realistic to provide a rank list of citation predictions than to make hard decisions for the users. For a given set of M predicted citations, we use a performance metric, Recall@M,

$$\text{Recall@}M = \frac{\text{number of citations in the predicted set}}{\text{total number of citations}}$$

which can be viewed as the proportion of "true" citations successfully predicted by a given method, when the method is allowed to provide M guesses.

Comparison methods. We compare our model with a number of competing strategies, starting with the RTM [9]. In order to make predictions using the RTM, we learn a latent representation for each document and predict citations using a similarity function between these representations (detailed in [9]). The second comparison is an LDA-based prediction strategy, in which document predictions are determined by the similarity between the latent document representation vectors θ_j. The similarity is computed using inverse of the Hellinger distance [6]

$$\mathrm{S}_{ij} = \mathrm{H}(\theta_i, \theta_j)^{-1} = \sqrt{2}\left\|\sqrt{\theta_i} - \sqrt{\theta_j}\right\|^{-1}.$$

Third, we compare with matrix factorization (MF), but only on the first task. (MF cannot make the citation predictions for a brand new document.) Finally, we compare with three simple baseline methods on both tasks. The first is that of Adamic/Adar [1], described in Section 4. The second is based on term frequence-inverse document frequency (TF-IDF) scores, where citations are predicted based on similarities in the documents' scores [4]. The third baseline is called "in-degree", where each document is given a score proportional to the number of times it is cited; in this case, the same set of predictions are given for every test document. Hyperparameters are set via cross validation.

Task one: predicting held-out citations. Given the document contents and the remaining links, the task is to predict the held out citations for each document. We show results for our model and six comparison methods on the ACL dataset in Figure 3. Our model (LRO) achieves a significantly higher recall over all ranges of the number of predictions, and we observed similar results for the other two datasets.

We also wanted to determine how our method performs across different datasets. To make the results comparable, we normalized the number of predictions M by setting it to a fraction of the total number of documents in each respective dataset. The results are shown in Figure 4: LRO performs well on all three datasets, though we note that ACL has a much better score than

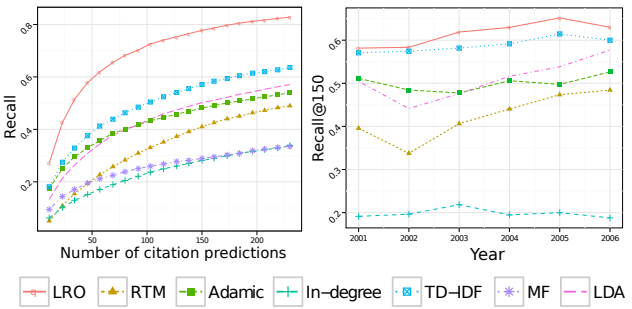

Figure 3: Left: Citation prediction performance on the ACL dataset for task one (predicting held-out citations).Right: Citation prediction performance on task two (predicting citations for new documents) on subsets of the ACL dataset for 7 years. In both cases, the LRO yields the highest recall over all ranges.

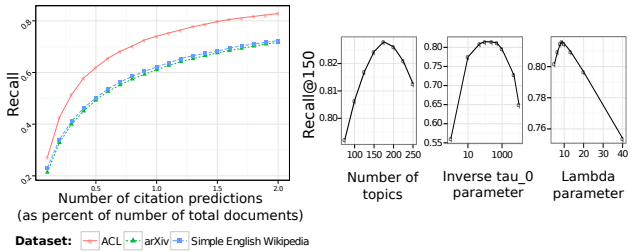

Figure 4: Left: citation prediction performance of our LRO model on three real-world datasets. The ACL dataset has a better score than the other two datasets. See main text for details. Right: citation prediction performance for a range of hyperparameter settings, including the number of topics K, the non-link variance parameter τ_0, and the latent random offset variance parameter λ.

the other two. We attribute this to the fact that ACL contains only refereed academic papers, and is therefore more structured than either arXiv (which is unrefereed) or Simple English Wikipedia (whose articles are not always subject to editorial attention).

Task two: predicting citations for new documents. The second task is to predict citations for documents with no prior citation information, corresponding to scenarios in which one needs to suggest citations for newly written documents. This task is often referred to as the "cold start problem" in recommender systems.

We simulate the process of introducing newly written papers into a citation network by dividing them according to publication year. Specifically, from the ACL citation network dataset, we select the citations and documents that existed before the year Y as training data, for Y ranging from 2001 to 2006. After training on this subset, the task is then to predict the citations occurring in year Y for the new documents written in year Y.

For this task, we compared our model against the same comparison methods used in the previous task, except for matrix factorization, which cannot make citation predictions for new documents. Figure 3 (right) shows the results. We fix the number of citation predictions $M = 150$ (other M values have similar trends). Again, our model achieves the best performance over a majority of the M values in all six years, and increases its lead over the comparison methods in later years, after a larger portion of the citation network has formed and can be used as training data.

Hyperparameter sensitivity. We also study how different hyperparameters affect performance, including the number of topics K, precision parameters τ_0 and τ_1, and latent random offset precision parameter λ (Figure 4, right). Again, we fix $M = 150$. First, we varied the number of topics from 75 to 250, and found an optimal value of approximately 175 topics. Next, in order to find the optimal balance between parameters τ_0 and τ_1, we fixed $\tau_1 = 1$ and varied τ_0 from 1/10000 to 1, finding an optimal value of approximately $\tau_0 = 1/100$. Finally, we varied the parameter λ from 5 to 40, and found an optimal value at approximately $\lambda = 9$.

5.2 Exploring citation networks

The latent random offsets can yield useful information that allows for analysis and exploration of documents in the citation network. Our model provides, for each document, a content representation vector θ_j, which captures the topics associated with the content of the document, and a latent offset vector ϵ_j, which captures topics not necesarily contained within the document but expressed by others who cited the document. Highly positive latent offsets may capture the topics where a given document has been influential within the context of the citation network; alternatively, negative offsets can represent topics that are expressed highly in a document, but that have not proven to be influential within the context of the network.

Given a document, we can therefore explore its contents by examining the learned set of topics, and we can explore its role in the citation network (and see the topics of documents that it has influenced) by examining the latent offsets. In Figures 1 and 5 we show the latent topic representations of document contents, the learned random offsets, and the final augmented representations (the sum of topic representations and random offsets), for a document in each of the Simple English Wikipedia and ACL datasets. The augmented representations provide information on both the content

and context of a document: they incorporate information contained in the document as well as in other documents that cite it.

For highly cited documents, we have a great deal of information from the citing documents (i.e. the in-links), and this information can be used to more strongly offset the latent topic representations. Intuitively, the content is like a prior belief about a document's latent representation, and as more sources start citing the document, this outside information further offsets the latent topic representations. Additionally, the offsets do not only "add" more information to the latent representation from the citing documents. In Figure 5 (top row), the offsets acted primarily to reduce the weights of many of the largest topics in the content representation, and only added weight to two topics. Here, the offsets served to dampen many of the content topics that did not appear to be relevant to the citing documents, and for this reason, the augmented representation is more sparse than the initial content representation.

Interpreting predictions. In addition to maintaining competitive prediction performance, our model allows for interpretable link prediction: for each predicted link we can use our latent representations to give users an understanding of why the link was returned. In particular, we can find the contribution that each topic provides to the final prediction score in order to determine the "reasons" (in terms of the latent topics) why a given document was predicted. We illustrate this in Figures 1 and 5 (bottom row of graphs). In Figure 1, for the *Sistine Chapel* document, *Chapel* is cited largely due to three topics (`architecture`, `Christianity`, and `buildings`), *Christian* is cited primarily due to a single topic (`Christianity`), and *Italy* is mainly cited due to six lower-weighted topics (`countries`, `Christianity`, `architecture`, `buildings`, `music`, and `populace`). Since *Italy* is a highly cited document and its augmented latent representation emphasizes a large number of topics (many of those expressed by its in-links), it was predicted due to a slight similarity in a number of topics as opposed to a strong similarity in just a few.

In Figure 5 we show three predictions for the document *Automatic Recognition of Chinese Unknown Words Based on Roles Tagging*. We can see that each of the predicted documents was due to a different aspect of this paper: the document *Automatic Rule Induction For Unknown-Word Guessing* was chosen primaily due to the `unknown-word` topic (related to the paper's goal of recognizing unknown words), the document *Word Identification for Mandarin Chinese Sentences* was chosen primarily due to the `China` topic (related to the paper's language domain area), and the document *A Knowledge-Free Method For Capitalized Word Disambiguation* was chosen primarily due to the `pronoun` topic (related to the paper's use of names, locations, and roles).

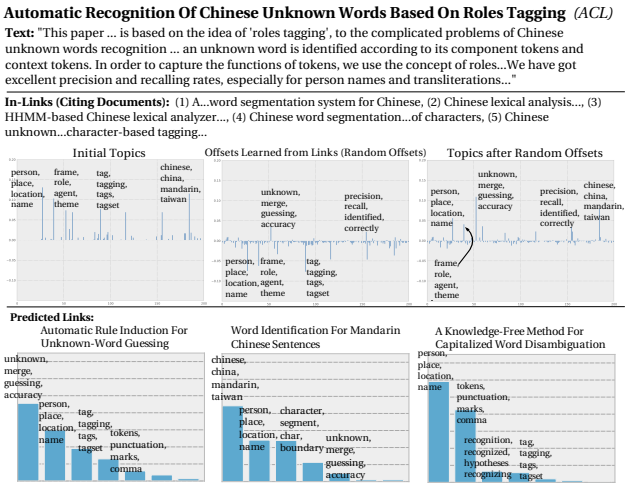

Figure 5: Interpreting citation predictions for the document *Automatic Recognition Of Chinese Unknown Words Based On Roles Tagging* in the ACL dataset. For each predicted link, we show the relative weight that each latent topic (denoted by the top four words) contributed to the prediction. These provide reasons why each predicted link was chosen, in terms of the topics.

6 Conclusion

In this paper, we proposed a probabilistic approach for citation network modeling that integrates the merits of both content and link analyses. Our empirical results showed improved performance compared with several popular approaches for citation prediction. Furthermore, our approach can suggest citations for brand new documents without prior citations—an essential ability for building a real citation recommendation system.

Qualitatively, our approach provides meaningful explanations for how predictions are made, through the latent random offsets. These explanations provide additional information that can be useful for making informed decisions. For example, in a citation recommendation system, we can inform users whether a citation is suggested more due to content similarities or due to the existing network structure, and we can show the relative amounts that individual topics contributed to the prediction. In future work, we would like to conduct user studies to quantify how this additional information helps users find more relevant citations in a more efficient way.

References

[1] Lada A Adamic and Eytan Adar, *Friends and neighbors on the web*, Social networks **25** (2003), no. 3, 211–230.

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

[3] D. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.

[4] Steven Bethard and Dan Jurafsky, *Who should I cite: learning literature search models from citation behavior*, International Conference on Information and Knowledge Management, 2010, pp. 609–618.

[5] D. Blei, *Probabilistic topic models*, Communications of the ACM **55** (2012), no. 4, 77–84.

[6] D. Blei and J. Lafferty, *Topic models*, Text Mining: Theory and Applications (A. Srivastava and M. Sahami, eds.), Taylor and Francis, 2009.

[7] D. Blei, A. Ng, and M. Jordan, *Latent Dirichlet allocation*, Journal of Machine Learning Research **3** (2003), 993–1022.

[8] Allison June-Barlow Chaney and David M. Blei, *Visualizing topic models*, The International AAAI Conference on Weblogs and Social Media, 2012.

[9] J. Chang and D. Blei, *Relational topic models for document networks*, Artificial Intelligence and Statistics, 2009.

[10] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. Blei, *Reading tea leaves: How humans interpret topic models*, Advances in Neural Information Processing Systems (NIPS), 2009.

[11] Qirong Ho, Jacob Eisenstein, and Eric P Xing, *Document hierarchies from text and links*, Proceedings of the 21st international conference on World Wide Web, ACM, 2012, pp. 739–748.

[12] M. Hoffman, D. Blei, C. Wang, and J. Paisley, *Stochastic Variational Inference*, ArXiv e-prints (2012).

[13] Yifan Hu, Yehuda Koren, and Chris Volinsky, *Collaborative filtering for implicit feedback datasets*, IEEE International Conference on Data Mining, 2008.

[14] Leo Katz, *A new status index derived from sociometric analysis*, Psychometrika **18** (1953), no. 1, 39–43.

[15] Yehuda Koren, Robert Bell, and Chris Volinsky, *Matrix factorization techniques for recommender systems*, Computer **42** (2009), no. 8, 30–37.

[16] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc, *Topic-link LDA: joint models of topic and author community*, International Conference on Machine Learning, ACM, 2009, pp. 665–672.

[17] Ramesh Nallapati and William Cohen, *Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs*, International Conference for Weblogs and Social Media, 2008.

[18] Ramesh Nallapati, Daniel A Mcfarland, and Christopher D Manning, *Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents*, International Conference on Artificial Intelligence and Statistics, 2011, pp. 543–551.

[19] Mark EJ Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences **103** (2006), no. 23, 8577–8582.

[20] Taeshik Shon and Jongsub Moon, *A hybrid machine learning approach to network anomaly detection*, Information Sciences **177** (2007), no. 18, 3799–3821.

[21] Alexander Smola and Shravan Narayanamurthy, *An architecture for parallel topic models*, Proc. VLDB Endow. **3** (2010), no. 1-2, 703–710.

[22] Chong Wang and David Blei, *Collaborative topic modeling for recommending scientific articles.*, ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2011.

[23] X. Wei and B. Croft, *LDA-based document models for ad-hoc retrieval*, SIGIR, 2006.