TopicFlow Model: Unsupervised Learning of Topic-specific Influences of Hyperlinked Documents

Ramesh Nallapati
nmramesh@cs.stanford.edu

Daniel McFarland
dmcfarla@stanford.edu

Christopher Manning
manning@stanford.edu

Stanford University, Stanford, CA 94305, USA

Abstract

Popular algorithms for modeling the influence of entities in networked data, such as PageRank, work by analyzing the hyperlink structure, but ignore the contents of documents. However, often times, influence is topic dependent, e.g., a web page of high influence in politics may be an unknown entity in sports.

We design a new model called TopicFlow, which combines ideas from network flow and topic modeling, to learn this notion of topic specific influences of hyperlinked documents in a completely unsupervised fashion. On the task of citation recommendation, which is an instance of capturing influence, the TopicFlow model, when combined with TF-IDF based cosine similarity, outperforms several competitive baselines by as much as 11.8%. Our empirical study of the model's output on ACL corpus demonstrates its ability to identify topically influential documents. The TopicFlow model is also competitive with the state-of-the-art Relational Topic Models in predicting the likelihood of unseen text on two different data sets. Due to its ability to learn topic-specific flows across each hyperlink, the TopicFlow model can be a powerful visualization tool to track the diffusion of topics across a citation network.

1 Introduction

Finding authoritative entities in hyperlinked data such as the world-wide-web, academic literature, blogs, social media, *etc.* is an important problem in data mining and information retrieval. Although popular algorithms such as PageRank [13] and HITS [9] have been very effective in addressing this problem, one of their main shortcomings is that they model only the hyperlink structure and completely ignore the contents of the documents. However, the influence of an entity is highly dependent on the topical context. For example, Andrew Sullivan is much more influential on the topic of politics than when writing about rap music.

This problem was addressed by the Topic Sensitive PageRank (TSP) algorithm [6], which essentially runs the PageRank algorithm [13] for each topic independently on the whole corpus such that for each topic, the 'teleportation' probability mass is distributed only among the seed documents on that topic. The resulting Topic Sensitive PageRank scores are therefore biased towards documents on the given topic. A key limitation of TSP is that it requires that the topics to be pre-specified and that a few labeled documents for each topic are available. However in many cases, all the documents in the corpus are unlabeled and the topics of the corpus are unknown. Hence a technique that simultaneously discovers topics in a corpus as well as the authoritative documents on each topic is desirable.

In this work, we will present a new model called *TopicFlow*, that simultaneously learns the topics, as well as topic-specific global influence of hyperlinked documents in a completely unsupervised fashion, thereby overcoming a key limitation of the TSP algorithm. In addition, the new model is able to quantify the flow of topics across the citation network, offering a powerful visualization tool to data mining practitioners.

2 TopicFlow model

The TopicFlow model consists of two distinct but mutually dependent components: (a) A flow network for various topics that describes how each topic spreads across the citation network, and (b) a generative model for text that assigns a topic to each word in a document. In this section we will present each of these components and then describe how they are tied together.

2.1 Topic attribution network

We consider a directed graph $G = (V, E)$ where $V = \{d_i\}_{i=1}^{M}$ is the set of all M documents in the corpus. We define K directed edges $e_{ij}^{(k)}$ from document d_i to d_j if d_i

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

cites d_j where K represents the number of topics in the corpus. Each directed edge $e_{ij}^{(k)}$ acts as an infinite capacity channel for "flow of attribution" from document d_i to document d_j on topic k. The actual flow for topic k, represented by $f_{i,j}^{(k)}$ is restricted to be non-negative and in the same direction as the edge, and quantifies the degree to which document d_i 'relies' on document d_j for its content on topic k. Conversely, this quantity also represents the amount of 'influence' document d_j has on document d_i on topic k.

We now define an augmented graph $G' = (V', E')$ called the *topic attribution network*, such that $V' = V \cup \{s, t\}$ where s and t represent a fictitious source and sink respectively, and

$$E' = E \cup \{e_{s,d}^{(k)}\}_{d \in V; k \in K} \cup \{e_{d,t}^{(k)}\}_{d \in V; k \in K}. \quad (1)$$

In other words, for each topic k, we add edges from the source to each document and from each document to the sink. We also assume that a unit flow arises out of the source s, flows into the network through the augmented edges from s and reaches the sink t through its augmented edges, as shown in Fig. 1.

The source and sink are introduced to account for the flow of information across the document network that is unexplained by the set of hyperlinks in the graph G. For example, in academic literature, this scenario of missing links may occur when the author forgets to cite relevant work, or is simply unaware of other relevant work that was developed recently or simultaneously by other researchers. In less formal domains such as blogs or web pages, this situation is more common because the authors are not obligated to attribute every idea explicitly using hyperlinks. The source-sink formalism also accounts for the missing links that result from the 'edge-effects' in a finite corpus[1].

Conceptually, the flow $f_{s,d}^{(k)}$ from the source into the document d on topic k represents the amount of topical information on topic k in d that is not 'credited' by any citation due to missing incoming links, and the flow $f_{d,t}^{(k)}$ from document d to sink t represents the amount of topical information in d that is not attributed to any document due to missing outgoing links.

We assume that the flows are balanced for each document-topic pair (d, k) as follows:

$$\sum_{i \in \mathtt{Pa}(d)} f_{i,d}^{(k)} = \sum_{j \in \mathtt{Ch}(d)} f_{d,j}^{(k)} \quad (2)$$

where $\mathtt{Pa}(d)$, read as parents of d, represents the set of vertices that have an outgoing edge into d in the augmented graph G', and $\mathtt{Ch}(d)$, read as children of d, is the set of vertices that have an incoming edge from d in G' [2]. At each

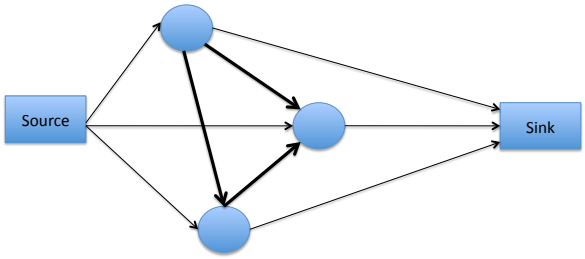

Figure 1: Topic-attribution network: the circular nodes are documents, the thick edges are the citations and the light edges are the augmented edges introduced from the source to each document and from each document to the sink. These edges represent the topic attribution network and should not be confused with the directed edges in a graphical model. We display edges corresponding to only one topic to prevent clutter.

document, we allow the incoming flow on each topic to be split arbitrarily across its children with the exception that a uniform fraction of the flow always flows into the sink as shown below.

$$f_{d,t}^{(k)} = f_{\cdot,d}^{(k)} / (|\mathtt{Ch}(d)|) \quad (3)$$

where $f_{\cdot,d}^{(k)} = \sum_{i \in \mathtt{Pa}(d)} f_{i,d}^{(k)}$ is the net incoming flow on topic k into d[3] and $|\mathtt{Ch}(d)|$ is the size of children of d in G'. This condition naturally satisfies the flow balance condition for documents that have no outgoing edges to other documents. For these documents, $|\mathtt{Ch}(d)| = 1$ (since the sink is their only child). Hence the entire incoming flow into such documents on a topic flows entirely into the sink.

At the source, there are two possibilities. We can either assume multiple sources, one for each topic, or a single source for all topics. Accordingly, we have the following source flow balance constraints:

$$\sum_{d=1}^{M} f_{s,d}^{(k)} = 1 \text{ for multiple sources, and}$$
$$\sum_{k=1}^{K} \sum_{d=1}^{M} f_{s,d}^{(k)} = 1 \text{ for a single source.} \quad (4)$$

In the single source paradigm, we allow the net flow originating at the source to be split arbitrarily not only among the documents but also across topics, thus modeling potential dominance of one topic over the others in the corpus.

2.2 Generative model for text

For generating the words in each document, we use a process similar to Latent Dirichlet Allocation [1], with one significant difference. Instead of using a Dirichlet prior

[1] In any finite corpus that has a notion of temporal ordering, documents at the beginning of time may have no outgoing links and documents at the end of time may have no incoming links.

[2] By definition, $\forall_{d \in V}\ s \in \mathtt{Pa}(d)$ and $t \in \mathtt{Ch}(d)$.

[3] We will also use the notation $f_{d,\cdot}^{(k)}$ to represent $\sum_{j \in \mathtt{Ch}(d)} f_{d,j}^{(k)}$, the net outgoing flow from a document d.

to generate the document's multinomial distribution over topics θ_d, we assume that the distribution is given by a deterministic process that we describe shortly. Accordingly, we generate document d with N_d words as follows:

For each position $i = 1, \ldots, N_d$:
Generate $z_i \in \{1, \ldots, K\} \sim \text{Mult}(\cdot|\theta_d)$
Generate $w_i \in \{1, \ldots, V\} \sim \text{Mult}(\cdot|\beta_{z_i})$

where V is the vocabulary size and β_k is the multinomial distribution over the vocabulary for topic K. The K topics in the generative model are the same as the topics we defined in the topic attribution network.

We define θ_d in terms of the incoming flows as follows:

$$\theta_d^{(k)} = (f_{\cdot,d}^{(k)})/(\sum_{k'=1}^{K} f_{\cdot,d}^{(k')}). \quad (5)$$

This is a key assumption that ties the network flow model with the topic model that generates words on specific topics in each document. The underlying hypothesis is that the more documents d' that assign 'votes' to document d on a given topic in terms of the attribution flow $f_{d',d}^{(k)}$, the greater is the probability that the words in d are on that topic. Clearly, this is not a model of the true process of document generation, since in the real world, the text of a document is generated before it receives citations from other documents. In this work, we exploit the "wisdom of the crowds" in hindsight, wherein we estimate the probability that a document discusses topic K based on the vote of confidence assigned on that topic by other documents through citations.[4]

Since the topical distribution θ_d for a document is defined based on the topical flows in the network, the model does not allow generation of a completely new document using a probabilistic generative process such as the one used in LDA. In other words, the TopicFlow model assumes a closed world of documents and is therefore not a fully generative model for new documents. In this regard, the TopicFlow model is very similar to PLSA [7]. Like PLSA, TopicFlow does offer a folding-in approach that estimates the topic assignments for a new document by estimating its new θ_d, as described in Sec. 4.4.

2.3 Discussion

In this framework, a dynamic interplay between citations and words determines the influence of a document on a given topic. If a document d discusses a topic k with high probability, it must have a high $\theta_d^{(k)}$ to explain its words, which induces higher topical flow $f_{\cdot,d}^{(k)}$ into the document relative to other topics. Conversely, if the document d is highly cited by other documents on topic k, it must have heavy incoming flow $f_{d,\cdot}^{(k)}$ on that topic, which will in turn induce high $\theta_d^{(k)}$ for that topic, resulting in the assignment of many words in d to that topic. Further, the topical flow balance conditions at every vertex ensure that the incoming flow at every document depends on the 'supply' of flow from vertices 'upstream' of d. Likewise, the outgoing topical flow at each document is influenced by the 'demand' for topical flow by vertices in its 'downstream'. Due to such network effects, the flow parameters learned by the model capture truly global influence. We define the topical influence of a document d on topic k as:

$$I(d,k) = (f_{\cdot,d}^{(k)} - f_{s,d}^{(k)})\theta_d^{(k)} \quad (6)$$

In other words, the influence is the product of topical attribution inflow from other documents not including the source, and the topical relevance in its text.[5]

3 Related Work

3.1 Comparison to Topic Sensitive Page Rank

Like TopicFlow model, TSP (as well as the original PageRank algorithm) has an implicit flow balance condition at each vertex, where the PageRank of each vertex is distributed typically equally among its children. Although the algorithm itself allows arbitrary splitting, it provides no guidance on how the mass should be split among its children. TopicFlow, on the other hand, learns to split a document's incoming flow on a topic among its children based on their topical relevance. In other words, the model can learn to distinguish between strong and less relevant citations on a given topic, which could be an especially attractive feature in the web context in culling out spam links.

TSP has a 'teleportation' feature where a small amount of the PageRank mass at each document is allocated uniformly to all documents on the topic of interest, to ensure irreducibility conditions for the Markov process associated with it. This is analogous to the source-sink formalism in the TopicFlow model, where the source supplies topical flow to all the documents and each document in turn drains a small amount of topical flow into the sink.

While TSP presents only a final set of PageRank values for each document on a given topic, TopicFlow can also quantify the amount of flow along each edge, providing a powerful visualization tool to track the diffusion of topics across the citation network. Finally, as we noted in the introductory section, TSP is a semi-supervised model that requires a set of seed labeled documents on each topic, while TopicFlow is a completely unsupervised model.

[4] We also examined a version where flows are in the opposite direction of citations, i.e. where flow can be interpreted as influence instead of attribution. We found the attribution model to be superior both qualitatively and quantitatively because, like PageRank, the attribution model treats links as votes and so better captures the notion of the "wisdom of the crowds".

[5] We also experimented with other definitions of influence, but the current definition gave us the best performance.

3.2 Related Work in Topic Modeling

Recently, many researchers have extended topic models to capture relational structure in document corpora. Topic models such as Joint Topic Models for Text and Citations [12], Latent Topic Models for Hypertext [5] and the more recent state-of-the-art Relational Topic Model [2] learn the topical correlations in documents connected by hyperlinks and thereby improve on the performance of basic LDA. Of these, the model in [5] also captures the notion of global influence of each document, but it is not topic specific. Another class of topic models such as Markov Topic Models [18] and Markov Topic Fields [8] capture topical correlations among documents related by venues, hyperlinks and time. Topic models such as the Citation Influence model [3] and HTM [15] both model the influence of a document's cited documents on itself as a multinomial distribution over the cited documents. However, the influence captured by these models is only local to each document. In another related work, Gerrish and Blei [4] estimate the influence of a document by modeling how the thematic content of that document is adopted by other documents over time. Their focus is in the context where hyperlink information is unavailable, which is different from the present work.

4 Learning and Inference

To learn the parameters of the model, we optimize the observed data log-likelihood for the whole corpus with respect to the flow parameters as well as the generative model parameters. The likelihood is given as:

$$\log P(\mathbf{w}|\beta,\mathbf{f}) = \sum_{d=1}^{M}\sum_{n=1}^{N_d}\log(\sum_{k=1}^{K}\beta_{kw_n}\theta_d^{(k)}), \quad (7)$$

The optimization problem for our model is the following:

$$\max_{\mathbf{f},\beta}\mathcal{F}(\mathbf{f},\beta) = \log P(\mathbf{w}|\beta,\mathbf{f})$$
$$-\frac{1}{2}\lambda\left(\sum_k\|\mathbf{f}_{s,\cdot}^{(k)}\|^2 + \sum_d\sum_k\|\mathbf{f}_{d,\cdot}^{(k)}\|^2\right)$$
$$\text{s.t. } \forall\, d\in V'-\{s\}\ :\ \sum_{i\in \texttt{Pa}(d)} f_{i,d}^{(k)} = \sum_{j\in\texttt{Ch}(d)} f_{d,j}^{(k)} \text{ and}$$
$$\sum_{d=1}^{M} f_{s,d}^{(k)} = 1 \text{ for multiple sources or}$$
$$\sum_{k=1}^{K}\sum_{d=1}^{M} f_{s,d}^{(k)} = 1 \text{ for one source} \quad (8)$$

where λ is the coefficient of regularization and $\mathbf{f}_{s,\cdot}^{(k)}$ is a vector consisting of all the flows from the source on topic k, while $\mathbf{f}_{d,\cdot}^{(k)}$ is the vector of all flows from document d to its children on topic k. L2 regularization is introduced into the objective function to ensure that all the flows remain small and as close to uniform as possible unless required by the data, and also to facilitate identifiability of the solution.

4.1 Elimination of equality constraints

We eliminate the equality constraints in Eq. 8 using the following equivalent flow balance condition.

$$\forall_{j\in\texttt{Ch}(d)}\ f_{d,j}^{(k)} = f_{\cdot,d}^{(k)}\psi_{d,j}^{(k)} \text{ s.t. } \sum_{j\in\texttt{Ch}(d)}\psi_{d,j}^{(k)} = 1 \text{ and } \forall_j\ \psi_{d,j}\geq 0 \quad (9)$$

where the new multinomial variable $\psi_d^{(k)}$ for each document-topic pair (d,k) determines how the net incoming flow into d on topic k is split among its children $\texttt{Ch}(d)$.

At the source, we define ψ as follows:

$$f_{s,d}^{(k)} = 1\cdot\psi_{s,d}^{(k)} \text{ s.t.} \forall_{d,k}\ \psi_{s,d}^{(k)}\geq 0, \text{and}$$
$$\sum_{k=1}^{K}\sum_{d=1}^{M}\psi_{s,d}^{(k)} = 1 \text{ for one src.;}$$
$$\sum_{d=1}^{M}\psi_{s,d}^{(k)} = 1 \text{ for multi-src.} \quad (10)$$

Equations 9 and 10 still have equality constraints, but these are much easier to handle, as we describe in Section 4.3.

4.2 Variational approximations

Using a variational posterior multinomial distribution ϕ_{dn} over topics for each position n in document d, we can define a lower bound on the log-likelihood of observed data as follows:

$$\log P(\mathbf{w}|\beta,\mathbf{f}) \geq$$
$$\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\phi_{dnk}(\log\beta_{kw_n} + \log\theta_d^{(k)} - \log\phi_{dnk}) \quad (11)$$

Maximizing the lower bound w.r.t ϕ_{dnk} and β_{kw_n} yields the following update rules:

$$\beta_{kw_n} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d}\phi_{dnk} \quad ; \quad \phi_{dnk}\propto\beta_{kw_n}\theta_d^{(k)} \quad (12)$$

However, estimating the parameter θ_d is non-trivial since it involves the flow parameters as given by Eq. 5. Substituting this equation into Eq. 11 yields the following:

$$\log P(\mathbf{w}|\beta,\mathbf{f}) \geq \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\phi_{dnk}(\log\beta_{kw_n}$$
$$+\log f_{\cdot,d}^{(k)} - \log(\sum_{k'} f_{\cdot,d}^{(k')}) - \log\phi_{dnk}) \quad (13)$$

The only parameters that remain to be estimated are the flow parameters $f_{ij}^{(k)}$ for each edge e_{ij} and topic k. However, instead of estimating the flow parameters directly, we estimate ψ's, the flow splitting proportions using the relation in Eq. 9. Notice that the multinomial parameter vector

$\psi_d^{(k)}$ enters into the lower bound in Eq. 13 only through the log likelihood terms for the children of d. Hence we only consider the observed data log-likelihood for $\texttt{Ch}(d)$ to optimize $\psi_d^{(k)}$:

$$\log P(\mathbf{w}_v|\beta,\mathbf{f})_{v\in\texttt{Ch}(d)} \geq \sum_{v=1}^{|\texttt{Ch}(d)|}\sum_{n=1}^{N_v}\sum_{k=1}^{K}\phi_{vnk}\,(\log\beta_{kw_n} + \log f_{\cdot,v}^{(k)} - \log(\sum_{k'} f_{\cdot,v}^{(k')}) - \log\phi_{vnk}) \quad (14)$$

We can now express $f_{\cdot,v}^{(k)}$ in terms of $\psi_{d,v}^{(k)}$ as follows:

$$f_{\cdot,v}^{(k)} = f_{-d,v}^{(k)} + f_{d,v}^{(k)} = f_{-d,v}^{(k)} + f_{\cdot,d}^{(k)}\psi_{d,v}^{(k)} \quad (15)$$

where $f_{-d,v} = \sum_{u\in\texttt{Pa}(j)-\{d\}} f_{u,v}^{(k)}$. The second step in the above equation arises from the relation in Eq. 9.

4.3 Logistic transformation of ψ's:

Although we have succeeded in eliminating equality constraints from the original problem in Eq. 8, we still have additional equality and inequality constraints expressed in Eq. 9 that guarantee that ψ_d^k remains a multinomial vector. We handle these constraints by further using a multinomial logistic transformation as shown below:

$$\psi_{d,v}^k = \begin{cases} (1-\frac{1}{|\texttt{Ch}(d)|})\frac{\exp(\eta_{d,v}^k)}{\sum_{v'\in\texttt{Ch}(d)-s}\exp(\eta_{d,v'}^k)} & \text{if } v\neq t \\ \frac{1}{|\texttt{Ch}(d)|} & \text{if } v=t, \end{cases} \quad (16)$$

where the variable $\eta_{d,v}^k$ is now unconstrained. Note that the probability of outflow into sink is obtained from Eq. 3. At the source, we define the logistic transformation for $\psi_{s,d}^k$ in line with its definitions in Eq. 10 as follows respectively:

$$\psi_{s,d}^k = \begin{cases} \frac{\exp(\eta_{s,d}^k)}{\sum_{d'\in\texttt{Ch}(s)}\exp(\eta_{s,d'}^k)} & \text{for multiple sources} \\ \frac{\exp(\eta_{s,d}^k)}{\sum_{k'}\sum_{d'\in\texttt{Ch}(s)}\exp(\eta_{s,d'}^{k'})} & \text{for single source} \end{cases} \quad (17)$$

We can now substitute Eqs. 15 and 16 into Eq. 14 to optimize η's of the documents directly in an unconstrained way. The final equations for derivatives of the objective function w.r.t. $\eta_{u,d}^k$ below, where u is a parent document of d, i.e., $u\in\texttt{Pa}(d)-s$ are given by:

$$\frac{\partial}{\partial\eta_{u,d}^k}\mathcal{F}(\mathbf{f},\beta) = \psi_{u,d}^{(k)}f_{\cdot,u}^{(k)}(1-\frac{1}{|\texttt{Ch}(u)|}) \left((\frac{\phi_{d\cdot k}}{f_{\cdot,d}^{(k)}} - \frac{N_d}{f_{\cdot,d}^{(\cdot)}}) - \sum_{d'\in\texttt{Ch}(u)}\psi_{u,d'}^{(k)}(\frac{\phi_{d'\cdot k}}{f_{\cdot,d'}^{(k)}} - \frac{N_{d'}}{f_{\cdot,d'}^{(\cdot)}})\right) - \lambda(f_{\cdot,d}^{(k)})^2\psi_{u,d}^{(k)}(\sum_{d'\in\texttt{Ch}(u)} -(\psi_{u,d'}^{(k)})^2 + \psi_{u,d}^{(k)}) \quad (18)$$

where $\phi_{d\cdot k} = \sum_{n=1}^{N_d}\phi_{dnk}$ and the second term above is computed only once for all $d'\in\texttt{Ch}(u)$.

Similarly, at the source, using the outflow relations in Eq. 10, and the logistic transformations in Eq. 17, we get the following equations for the derivative:

$$\frac{\partial}{\partial\eta_{s,d}^k}\mathcal{F}(\mathbf{f},\beta) = \frac{\partial}{\partial\eta_{s,d}^k}(\sum_{d'}\sum_{n=1}^{N_{d'}}\sum_{k'=1}^{K}\phi_{d'nk'}(\log\beta_{k'w_n} + \log(\frac{f_{\cdot,d'}^{(k')}}{\sum_{k''}f_{\cdot,d'}^{(k'')}}) - \log\phi_{d'nk'})$$

$$= \left((\frac{\phi_{d\cdot k}}{f_{\cdot,d}^{(k)}} - \frac{N_d}{f_{\cdot,d}^{(\cdot)}}) - \sum_{d'}\psi_{s,d'}^k(\frac{\phi_{d'\cdot k}}{f_{\cdot,d'}^{(k)}} - \frac{N_{d'}}{f_{\cdot,d'}^{(\cdot)}})\right)\psi_{s,d}^{(k)}f_{\cdot,s} - \lambda(f_{\cdot,s})^2\psi_{s,d}^{(k)}\left(\sum_{d'} -(\psi_{s,d'}^{(k)})^2 + \psi_{s,d}^{(k)}\right) \text{ for mult.src. ;} \quad (19)$$

$$= \left((\frac{\phi_{d\cdot k}}{f_{\cdot,d}^{(k)}} - \frac{N_d}{f_{\cdot,d}^{(\cdot)}}) - \sum_{d'}(\sum_{k'}\frac{\phi_{d'\cdot k'}\psi_{s,d'}^{(k')}}{f_{\cdot,d'}^{(k')}} - \frac{N_{d'}\psi_{s,d'}^{(\cdot)}}{f_{\cdot,d'}^{(\cdot)}})\right)(\psi_{s,d}^{(k)}f_{\cdot,s}) - \lambda(f_{\cdot,s})^2\psi_{s,d}^{(k)}\left(\sum_{k'}\sum_{d'} -(\psi_{s,d'}^{(k')})^2 + \psi_{s,d}^{(k)}\right) \text{ for one src.} \quad (20)$$

Notice that the derivative in Eq. 19 for the multiple-sources version is analogous to the document derivative in Eq. 18, owing to their analogous behavior as sources of flows.

We optimize $\eta_{u,d}^{(k)}$ and $\eta_{s,d}^{(k)}$ by performing gradient ascent using the derivatives in Eqs. 18 and 19. The model is initialized randomly before commencing gradient ascent.

4.4 Inference

At inference time, given a completely new document network, we optimize the objective function in Eq. 8 with respect to the flow parameters only, keeping β's, the topic distributions over the vocabulary, fixed at values learned at training time. Since the normalized flow parameters give us the values of each document's distribution over topics as shown in Eq. 5, it effectively means that we are learning the θ for each document in the test set. This is unlike the fully generative approach in LDA, but is analogous to the PLSA model, that learns the θ's for test documents using a folding-in approach as described in Section 4.2 of [7].

Although the TopicFlow model is intended to run on networked data, it can also learn from and infer on pure text alone. On a non-networked dataset, the only edges in the topic attribution network are the augmented edges from the source to each document, and from each document to the sink. The learning and inference is run normally on this network. In this scenario, the TopicFlow model would simply collapse to PLSA.

5 Experiments

5.1 Data sets

The first dataset we considered is the ACL anthology [14] dataset comprising full text and abstracts of all papers published in NLP conferences such as ACL, EMNLP, NAACL, *etc*., over a period of over 30 years. For our experiments, we used the full text of 9,824 papers published before or in 2005 as the training set. There are 33,604 hyperlinks in total in the training set and some of the documents contain no incoming or outgoing hyperlinks. We used 1,041 abstracts of papers published in 2006 with no hyperlink information within this data, as the test set. However, we do have the hyperlinks that arise from the test documents and point to the ones in training, but we use it for only evaluation purposes in the citation recommendation experiments, described in Section 5.2 below. After stopping and stemming, our vocabulary size is 46,160. The average training full text document is 1848.07 words long while the average test abstract is only 61.59 words long.

As an additional dataset, we also used the Cora dataset [10] that consists of abstracts of Computer Science research papers. After removing documents with no incoming or outgoing hyperlinks and randomly sampling the remaining documents, we are left with 11,442 documents with 24,582 hyperlinks. We removed stop-words and performed stemming, resulting in a vocabulary size of 7,185 words. Each document length is 72.19 words.

5.2 Citation Recommendation

This task consists of predicting the true citations (outgoing links) of a document based on its textual content and that of the other documents. The choice of this task is based on our hypothesis that documents tend to cite other documents that are not only topically relevant but also influential. We believe this is a reasonable assumption for academic datasets. Assuming this assumption holds good, this task should allow us to distinguish between models such as TopicFlow and TSP that can potentially model influence and other models such as LDA and RTM that do not.

In our experimental setup, for each abstract in the test set of the ACL corpus, we score documents in the training set based on a "citability score". We then rank the training documents in decreasing order of the citability score and evaluate the quality of the ranked list using Average Precision [16] measured with respect to the test document's true citations in the training set. We compute the mean of these average precisions over all test documents, called MAP, which we use as the evaluation metric.

As a first baseline, we used TF-IDF based cosine similarity between the test document's content and the training document's content. We used the basic LDA [1] that only models text, and RTM [2], the state-of-the-art joint topic model for text and hyperlinks, as additional baseline topic models. We used an open-source Gibbs sampling based implementation for RTM[6] while for LDA, we used David Blei's variational inference based implementation[7].

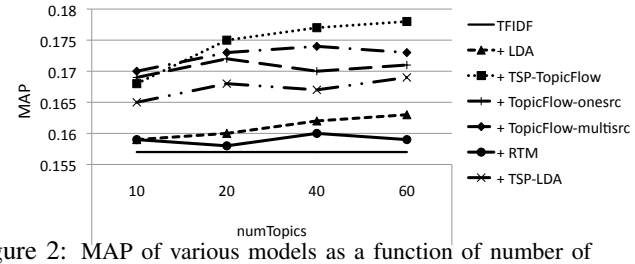

Figure 2: MAP of various models as a function of number of topics on the citation recommendation task. The symbol '+' in the legend indicates that the corresponding model is combined with TF-IDF score. Both versions of TopicFlow are significantly better than all models except TSP-TopicFlow, with the best TopicFlow model outperforming the strong TF-IDF baseline by as much as 10.82%. TopicFlow as well as TSP-TopicFlow are significantly better than RTM, LDA and TF-IDF as measured by Wilcoxon's signed rank test at 99% confidence, but the differences between TSP-TopicFlow and both TopicFlow models are not statistically significant at the same confidence level.

For all topic models we used in our experiments, we trained the models on the training set consisting of textual as well as network information, and inferred the distribution over topics for only the text of each test document. We then defined the citability score for each train and test document pair as ζ(Model-score) + $(1-\zeta)$(TF-IDF-score) where Model-score is the cosine similarity between the test document and the train document in the topic space and $0 \leq \zeta \leq 1$ is a tunable free parameter.

For LDA, cosine similarity is computed between the test document's inferred topic distribution and the training document's learned topic distribution. For the TopicFlow model, we computed cosine between the train document's topical influence vector $\{I(d,k)|k \in 1,\ldots,K\}$ given by Eq. 6 and the test document's inferred topic distribution θ. For RTM inference, we estimated the β's, the topic-specific distributions over vocabulary from the counts in the training set, and sampled the topics for documents in the test set, keeping the β's fixed. Then we estimated the θ's for each test document by averaging the topic counts across several Gibbs samples. Finally, we also used Topic Sensitive PageRank as an additional model for comparison. Since TSP requires a pre-defined set of topics and a seed set of labeled documents for each topic, we used topics from both the multi-source TopicFlow model as well as the basic LDA as input to TSP, where for each topic k, we used all documents

[6] http://cran.r-project.org/web/packages/lda/. We took assistance from the first author of RTM in performing all RTM experiments, for which we remain thankful.

[7] http://www.cs.princeton.edu/∼blei/lda-c/

d that satisfy $\arg\max_{k'} \theta_{dk'} = k$ as the seed examples.

We tuned all the free parameters of the models on a development set, which is derived from further splitting the training set into all documents preceding 2004 as development-training and all documents in 2004 as development-test. For all the models except the baseline TF-IDF model, we tuned their respective free parameters with the number of topics K fixed at 30. For all models, we found that the best performance is reached in the interval $\zeta \in [0.05, 0.2]$. For TSP, the optimal value of the teleportation probability is found to be 0.30. For TopicFlow models, the optimal value for regularization is 1.0 for multi-source version and 10.0 for single source version. For all models, we report the MAP scores for the test set for $K = 10, 20, 40$ and 60.

The results of our experiments, displayed in Fig. 2, show that both versions of TopicFlow significantly outperform TF-IDF and LDA that use only textual information, as well as RTM that uses both text and hyperlinks, with the best TopicFlow model achieving 10.82% improvement over the TF-IDF baseline. The differences between TopicFlow and the above mentioned models are also statistically significant as measured by Wilcoxon's signed rank test at 99% confidence. Although RTM is a good model for capturing topical correlations in a citation network, it has no mechanism to capture the notion of global influence. We believe this could be one of the distinguishing features of TopicFlow that allows it to outperform RTM as well as LDA. More interestingly, our experiments show that the TSP algorithm, that uses the topic distributions from TopicFlow (we call this run 'TSP-TopicFlow'), is able to achieve the best performance (a maximum of 13.37% improvement over baseline TF-IDF). However, both TopicFlow models and the TSP-TopicFlow are statistically indistinguishable by the Wilcoxon test at 99% confidence level. To understand whether the high performance of TSP-TopicFlow is due to TSP alone or by virtue of TopicFlow's topics, we also ran TSP on the output of LDA topics (which we call 'TSP-LDA'). As shown in the figure, this combination performs significantly worse than TopicFlow as measured by the same Wilcoxon's test, indicating that the topics learned by TopicFlow are superior to those learned by LDA.

5.3 Empirical Analysis

Fig. 3 presents a visualization of the TopicFlow model run on the ACL training corpus. Since the corpus consists of papers in the ACL conference, we see mostly Natural Language Processing topics such as 'Machine Translation', 'Parsing' and 'Discourse Analysis'. As indicated in the bottom row of the table, the model is also able to numerically quantify the influence of documents on each topic. In addition, we also display the flows in the neighborhood of the most influential document on the topic of 'Machine Translation', which helps us understand how influence has

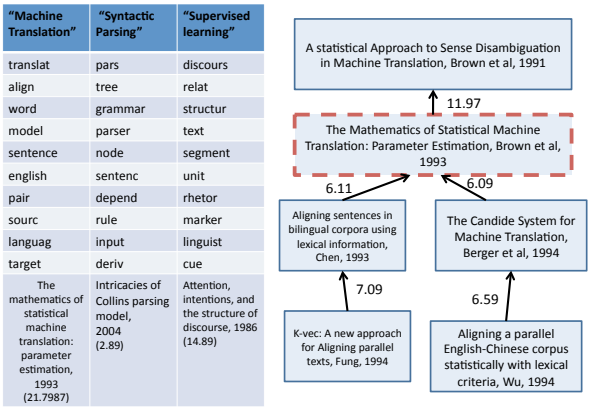

Figure 3: Visualization of a 60 topic Multi-source TopicFlow model on the ACL training corpus: Left: top 10 words and the most influential document (bottom row) for three representative topics. The numbers in the bottom row in braces indicate the topic-specific influence of the document as measured by $I(d, k) \times 100$. Right: a slice of the TopicFlow model in the neighborhood of the most influential document (bordered in broken lines in dark red) on the topic of "Machine Translation". The numbers next to the arrow are the topic-specific flows, times 100.

spread across the network on this topic. This feature, to the best of our understanding, is unique to TopicFlow.

We also compared the influence rankings of TopicFlow with the list of all-time most cited ACL papers[8]. We found that 6 out of these 10 most cited papers also occur in the TopicFlow's list of top 10 most influential papers on at least one topic, and that all 10 papers occur in the top 52 most influential papers on at least one topic. Further, the very same papers are ranked as low as 300-2500 on irrelevant topics, demonstrating the model's ability to capture only topic-specific influence. Upon further inspection of the most cited papers that are not ranked as high by TopicFlow, we realized that these are actually broad dataset or methodology papers that are widely cited for reasons other than their topical influence. For example, the most cited paper in the ACL corpus, titled "Building a large annotated corpus of English: The Penn TreeBank" is ranked at the highest rank of 14 by TopicFlow on the topic of 'Parsing'. We conjecture that most papers that present techniques for parsing use the Penn TreeBank corpus in their experiments, and therefore cite this paper for completeness, which does not necessarily reflect the influence of the TreeBank paper on them. Similarly, we found that the third most cited paper in the ACL corpus, titled "Bleu: A method for automatic evaluation of Machine Translation", is ranked at its highest rank of 52 on the topic of 'Machine Translation'. Again, we argue that although this paper is a very important contribu-

[8] Available at http://clair.si.umich.edu/clair/anthology/rankings.cgi.

tion, most papers cite this work primarily because they use the evaluation measure proposed in this paper. We found that although these papers have a huge number of citations, the topical flow from each citation is quite small, resulting in a small overall influence. The TopicFlow model is thus able to discount the citations to these papers by virtue of the differences in word and topic usage between the citing and cited papers.

5.4 Document Completion Log-likelihood

We also compare the performance of the TopicFlow model with LDA and RTM on the task of predicting unseen text. Comparing the likelihood of TopicFlow with that of fully generative models such as LDA on new documents would be unfair to the latter since TopicFlow learns the θ parameters for new documents, while LDA only marginalizes θ with respect to the learned Dirichlet prior. Hence we opted for the *Document completion* likelihood where we learn the models' parameters based on the first half of each text document as well as the network information, and estimate the model's likelihood on the second halves of the same documents as described in section 5.1 of [17]. Since all models use their learned estimates of θ to estimate the likelihood of the second half of each document, the comparison is the fairest possible. For LDA, we estimated θ as the expectation of the variational posterior parameters defined in Eq. 7 in [1], while for RTM, we used point estimates from Gibbs samples as shown in Eq. 27 in [17]. For TopicFlow, we used Eq. 5 to estimate θ, and set the regularization coefficient λ to zero, to facilitate fair comparison with all the other models that have no regularization terms in their objective functions.

The results presented in Fig. 4 on both the Cora and the ACL datasets show that both RTM and TopicFlow, which exploit citation information, are able to better predict the unseen half of the documents than LDA which uses only textual data for learning. Both versions of the TopicFlow model are quite competitive with RTM on the Cora dataset and consistently outperform RTM on the ACL data.

6 Conclusion

In this paper, we presented a new model that combines network flow with topic modeling approach and learns topic-specific influence of documents in a completely unsupervised fashion. Our experiments on citation retrieval as well as the empirical analysis on the ACL corpus demonstrated that the new model is not only competitive with state-of-the-art models in modeling topic-specific influences of documents, but is also successful in filtering out documents that are highly cited for reasons other than their topical influence. Besides, our experiments on log-likelihood show that TopicFlow is a good model for text as well.

The TopicFlow model can serve as a powerful visualization tool for information analysts to track the diffusion of topics across citation networks, and also to identify the topic-specific influential documents. In the future, we plan to build a user-friendly web based graphical browser based on the model's output on various corpora. We also plan to explore the applicability of the model to adversarial and non-temporal corpora such as the web.

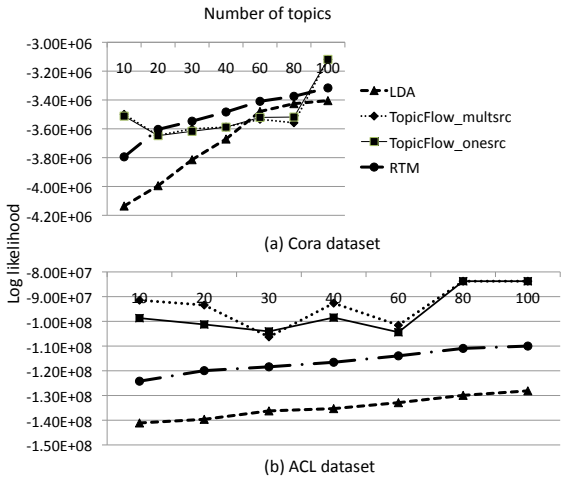

Figure 4: Log-likelihood as a function of number of topics on two datasets. Both versions of TopicFlow are competitive with RTM on Cora, while they consistently outperform RTM on ACL. Both RTM and TopicFlow are better than LDA because they are able to capture topical correlations across hyperlinked documents.

Although the model's computational complexity is linear in the number of topics and number of documents, the current implementation is not easily scalable to millions of documents[9]. We have developed a simple algorithm to parallelize the learning which we plan to implement shortly.

Acknolwedgments

This research was supported by NSF grant NSF-0835614 CDI-Type II: What drives the dynamic creation of science? We wish to thank our anonymous reviewers for their deeply insightful comments and feedback. In particular, we would like to thank Daniel Ramage for his Topic Sensitive PageRank implementation and helpful suggestions, Jonathan Chang for his assistance with RTM, and David Vickrey and Rajat Raina for their suggestions on optimization techniques.

[9]On the ACL corpus size of 9,824 training documents and 1,041 test documents, learning and inference together took about 47 min. for 10 topics and 12 hrs. 10 min. for 100 topics on a Dual Core 2.4MHz AMD processor with 16G RAM, running Linux.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 2003.

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

[3] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *International Conference on Machine Learning*, 2007.

[4] S. Gerrish and D. Blei. A language-based approach to measuring scholarly impact. In *International Conference on Machine Learning*, 2010.

[5] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Latent topic models for hypertext. In *Uncertainty in Artificial Intelligence*, 2008.

[6] T. H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. In *IEEE Transactions on Knowledge and Data Engineering*, 2003.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[8] H. Daume III. Markov random topic fields. In *ACL-IJCNLP Conference*, 2009.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999.

[10] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000. www.research.whizbang.com/data.

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

[12] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *KDD*, 2008.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. In *Stanford University technical Report*, 1998.

[14] D. Radev, M. T. Joseph, B. Gibson, and P. Muthukrishnan. A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Society for Information Science and Technology*, 2009.

[15] C. Sun, B. Gao, Z. Cao, and H. Li. HTM: a topic model for hypertexts. In *Empirical Methods in natural Language Processing*, 2009.

[16] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *ACM SIGIR Conference*, 2006.

[17] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *ICML*, 2009.

[18] C. Wong, B. Thiesson, C. Meek, and D. Blei. Markov topic models. In *AISTATS*, 2009.