Cross-domain Collaboration Recommendation

Jie Tang[†], Sen Wu[†], Jimeng Sun[‡], and Hang Su[†]
[†]Department of Computer Science and Technology, Tsinghua University
[‡]IBM TJ Watson Research Center, USA
jietang@tsinghua.edu.cn, ronaldosen@gmail.com, jimeng@us.ibm.com, suhang@sse.buaa.edu.cn

ABSTRACT

Interdisciplinary collaborations have generated huge impact to society. However, it is often hard for researchers to establish such cross-domain collaborations. What are the patterns of cross-domain collaborations? How do those collaborations form? Can we predict this type of collaborations?

Cross-domain collaborations exhibit very different patterns compared to traditional collaborations in the same domain: 1) **sparse connection:** cross-domain collaborations are rare; 2) **complementary expertise:** cross-domain collaborators often have different expertise and interest; 3) **topic skewness:** cross-domain collaboration topics are focused on a subset of topics. All these patterns violate fundamental assumptions of traditional recommendation systems.

In this paper, we analyze the cross-domain collaboration data from research publications and confirm the above patterns. We propose the Cross-domain Topic Learning (CTL) model to address these challenges. For handling sparse connections, CTL consolidates the existing cross-domain collaborations through topic layers instead of at author layers, which alleviates the sparseness issue. For handling complementary expertise, CTL models topic distributions from source and target domains separately, as well as the correlation across domains. For handling topic skewness, CTL only models relevant topics to the cross-domain collaboration.

We compare CTL with several baseline approaches on large publication datasets from different domains. CTL outperforms baselines significantly on multiple recommendation metrics. Beyond accurate recommendation performance, CTL is also insensitive to parameter tuning as confirmed in the sensitivity analysis.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining; J.4 [**Social Behavioral Sciences**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Collaboration recommendation, Social network, Social influence

KDD'12, August 12–16, 2012, Beijing, China.

1. INTRODUCTION

Social network analysis focuses on modeling interactions between people. Researchers have studied various issues in social networks, such as network properties [6, 11] and generation processes [18], link predictions [19, 20, 21, 32] and recommendations [2, 7, 17]. Despite all the existing research in social networks, little has been done on analyzing collaborations across two different domains.

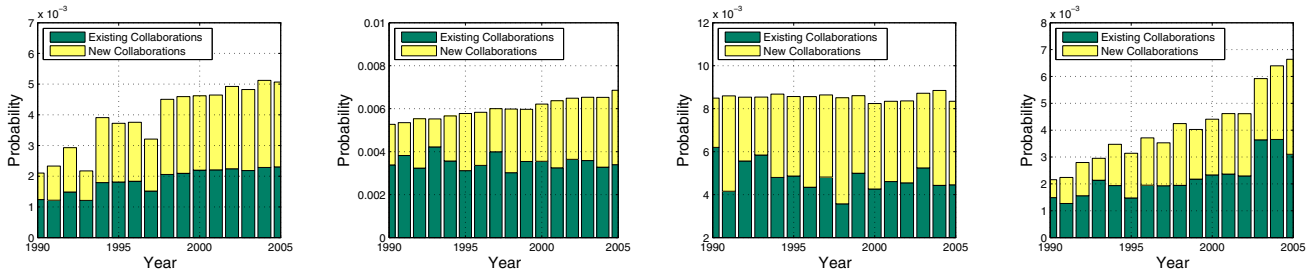
Interdisciplinary collaborations have generated huge impact to society. For example, collaborations between biology and computer science revolutionized the field of bioinformatics. Because of these cross-domain collaborations, originally extremely expensive tasks such DNA sequencing have become scalable and affordable to a much broader population. Now medicine and data mining are working together in the field of medical informatics, which is a big growth area that is expected to have huge impact on medicine [24]. Indeed, cross-domain collaboration has become increasingly important. Figure 1 shows the increasing trend of cross-domain collaborations over the past fifteen years across different domains in a publication database (Cf. § 4 for details). In most of the cases, there exists a clear increasing trend of the cross-domain collaborations.

However, it is often hard for researchers to establish such cross-domain collaborations. What are the patterns of cross-domain collaborations? How do those collaborations form? Can we predict this type of collaborations? Cross-domain collaborations often exhibit very different challenges compared to traditional collaborations in the same domain:

First, **sparse connection**, cross-domain collaborations are rare compared to traditional collaborations within a domain, partly because it is difficult for an outsider to find the right collaborator in the field that one does not know. This also makes it challenging to directly use a supervised learning approach due to the lack of training samples.

Second, **complementary expertise**, cross-domain collaborators often have different expertise and interest; For example, data mining researchers can easily identify who they want to work with in the data mining field, because the topics are known to them. However, for a cardiologist who wants to apply data mining techniques to predict heart failures, it will be difficult for her to find the right collaborators in data mining. Because these two fields (cardiology and data mining) are completely different with different terminology and problems. It is very difficult for one from cardiology to identify the right topics in data mining to look for collaborators.

Third, **topic skewness**, not all topics are relevant for cross-domain collaborations. In fact, in our study, only less than 9% of all possible topics pairs across domains have collaborations. Therefore, for the task of cross-domain collaboration recommendation,



(a) DM - TH (b) DM - MI (c) DM - VIS (d) MI - DB

Figure 1: The comparison of existing collaboration and new collaboration trends over years. DM - Data Mining domain; MI - Medical Informatics domain; TH - Theory domain; VIS - Visualization domain; DB - Database domain. The trends of cross-domain collaborations in all but one case are growing (The exception between DM and VIS remain roughly constant over time). Newly formed cross-domain collaborations are significantly in all cases.

we should focus on better modeling those topics with high probability of having cross-domain collaborations.

Despite of the above challenges, once such cross-domain collaboration is successfully formed, its impact is usually tremendous. In our study, cross-domain collaborations constitute a small portion of all possible collaborations as shown in Figure 1. The trends of cross-domain collaboration in many cases are growing. Newly formed cross-domain collaborations are significant in all cases, which confirmed the potential need for cross-domain collaborations.

Based on these observations, we propose the Cross-domain Topic Learning (CTL) method that addresses all three challenges including sparse connection, complementary expertise and topic skewness. CTL is a generative topic model that differentiates relevant topics to cross-domain collaboration from other topics.

We compare CTL with several baseline approaches on large publication data sets of different domains. CTL outperforms others significantly on recommendation metrics. Beyond accurate recommendation performance, CTL is also insensitive to parameter tuning as confirmed in the sensitivity analysis. Finally, we integrate CTL into a large-scale web application for recommending cross-domain research collaborators, which further demonstrates the scalability of CTL in handling real-time queries.

The rest of this paper is organized as follows: Section 2 formulates the cross-domain recommendation problem formally; Section 3 presents our proposed methods on cross-domain recommendation; Section 4 describes the experiments; Section 5 presents the related work; then we conclude in Section 6.

2. PROBLEM DEFINITION

We present required definitions and formulate the problem of cross-domain collaboration recommendation. Without loss of generality, we assume there are two domains, the source domain and the target domain. Our goal is to recommend potential collaborators in the target domain for a specific user from the source domain.

Definition 1. **Source/Target domain.** The source (or target) domain can be represented as a social network $G = (V, E, X)$, where V is a set of $|V| = N$ users and $E \subseteq V \times V$ is a set of undirected (collaborative) relationships between users, X is an $N \times d$ attribute matrix in which every row corresponds to a vector of attribute values of a user. We use x_j to denote the j^{th} attribute.

We use superscript ^{S} and ^{T} to differentiate the source domain and the target. If there is no ambiguity, we will omit ^{S} for the source domain and use superscript $'$ for the target, for brevity. Suppose each user v_i is associated with d attributes. For example, in the research collaboration network, each user is associated with a set of publication papers or a set of words appearing in those papers. Given this, we have the following definition:

Definition 2. **Domain-specific topic models.** A topic model θ_i of a user v_i is a multinomial distribution of attributes $\{P(x_j|\theta_i)\}_j$. Then a domain is considered as a mixture of multiple user-specific topic models. The assumption behind is that attributes associated with the user are sampled following a distribution corresponding to each topic, i.e., $P(x|\theta_i)$.

Such a definition is usually used in the LDA/PLSI style topic models [4, 15]. According to the above definition, attributes with the highest probability associated with each topic would suggest the semantics represented by the topic. For example, a "Data Mining" topic discovered from the publication data can be represented by keywords "clustering", "learning", "classification", etc.

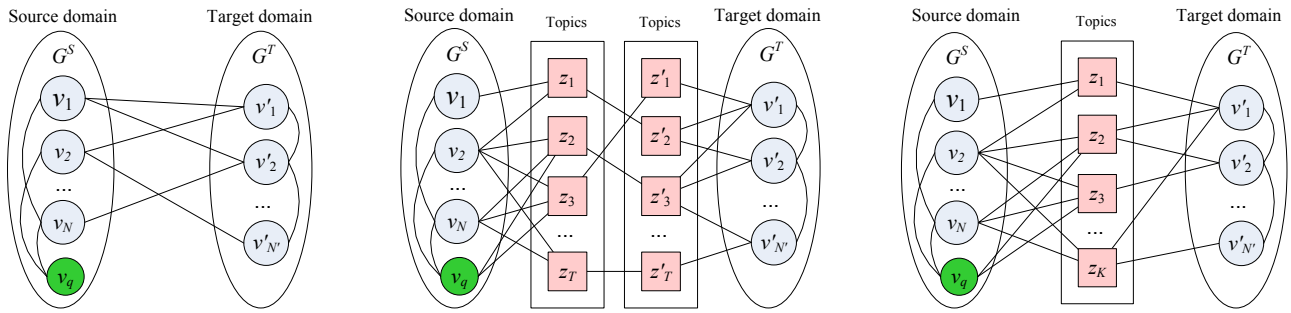
The input of our problem consists of a source domain G^S and a target domain G^T, each associated with topic models. Please note that the source domain and the target domain can be overlapping, i.e., $V^S \cap V^T \neq \varnothing$. Given this, we can precisely define the following problem:

Problem 1. **Cross-domain collaboration recommendation.** Given (1) a source domain G^S and a target domain G^T, (2) topic models θ and θ' associated with users in the two domains respectively, the goal is to rank and recommend potential collaborators in the target domain for a specific user v_q from the source domain.

The fundamental challenge of this problem is how to capture the collaboration patterns across different domains. Within the same domain, homophily is often considered as the driven force for the formation of collaborative relationships, which suggests that people with the similar interest (topic model θ) tend to collaborate with each other. However, in the cross-domain setting, the problem is very different. Technically, it is challenging to extract and discriminate topics in the two domains. In particular, given a specific user and her topic distribution from the source domain, on *which topics* and with *whom* should she collaborate in the target domains?

3. CROSS-DOMAIN TOPIC LEARNING

We begin by considering some baseline solutions and then propose our cross-domain topic learning approach. A simple approach to the problem is to construct a collaboration graph connecting users between source and target domains and then use a random walk with restart algorithm [28] to rank collaborators in the target



(a) Author matching (b) Topic matching (c) CTL

Figure 2: Graphical representation of the three recommendation models: author matching, topic matching, and CTL.

domain. We call this method *Author Matching*. The details of the algorithm are described in Section 3.1.

The problem with Author Matching is the sparse connections between authors across two domains. To alleviate this problem, the second model is to consolidate the correlation between the underlying topics. Suppose each domain has T different topics and each user has a distribution over the topics. We can augment the collaboration graph with two topic layers (as shown in Figure 2(b)). The links between the two topic layers indicate the alignment between topics, which implicitly represents the complementary expertise between users. Based on this representation, a random walk with restart algorithm can be again applied to the graph to rank (both topic and user) nodes in the target domain. We call this method *Topic Matching*, and details are described in Section 3.2.

Topic matching improves the cross domain connections through a subset of topic pairs from source domain to target domain. However, not all topic pairs are relevant for collaboration (topic skewness). Therefore, blindly computing all topics from source and target domains are not necessary for collaboration recommendation and often lead to sub optimal results. One challenge here is how to differentiate relevant "collaboration" topics from other topics. We further design a Cross-domain Topic Learning (CTL) algorithm to address this challenge in Section 3.3.

3.1 Author Matching

Based on the historic cross-domain collaborations, we create a collaboration graph, as shown in Figure 2(a). The problem is to rank relevant nodes in the target domain G^T for a given query node v_q in the source domain G^S. Measuring the relatedness of two nodes in the graph can be achieved using the Random Walks with Restarts (RWR) theory [22, 28]. Starting from node v_q, a RWR is performed by following a link to another node according to the weight of the link at each step.[1] Also, in every step, there is a probability τ to return the node v_q. The relatedness score of node v_i wrt node v_q is defined as the stationary probability r_i that the random walk will finally arrive node v_i, i.e.,

$$\mathbf{r}^{(t+1)} = (1-\tau)\mathbf{S}\cdot\mathbf{r}^{(t)} + \tau\mathbf{q} \tag{1}$$

where $\mathbf{r}^{(t)}$ is a vector with each element r_i^t denoting the probability that the random walk at step t arrives at node v_i; $\mathbf{q}$ is a vector of zero with the element corresponding to the starting node v_q set to 1, i.e., $q_{v_q} = 1$; $\mathbf{S}$ defines the transition probability of the random walk, with element S_{ij} denoting the random walking probability from node v_i to node v_j.

[1]In the author matching method, we use a uniform weight, i.e., weights of links of a node v to its neighbors are defined as $\frac{1}{NB(v)}$, where $NB(v)$ is the number of neighbors of node v. In §3.2, we will introduce how to define the weight based on topic model.

3.2 Topic Matching

The author matching method only considers the network structure information, but ignores the content (topic) information. How do people collaborate across different domains? And what are the hottest topics on which people from different domains tend to collaborate?

Recently, probabilistic topic models have been successfully applied to multiple text mining tasks to extract topics from text [4, 15, 27]. We employ an Author-Conference-Topic (ACT) model [31], which utilizes the topic distribution to represent the interdependencies among authors, papers, and publication venues.[2] The model simulates the process when people collaborate on a work, e.g., writing a scientific paper, using a series of probabilistic steps. In essence, for each object it estimates a mixture of topic distributions which represent the probability of the object associated with every topic. Such as for each author v, we have a set of probabilities $\{P(z_i|v)\}_i$ or $\{\theta_{vz_i}\}_i$, respectively denoting how likely author v is interested in topic z_i. Similarly, we have $\{P(x_j|z)\}_j$ or $\{\phi_{zx_j}\}_j$, the probability of attribute x_j (e.g., a keyword) given topic z. We use Gibbs sampling to learn the probabilities. The interested reader can refer to [31] for more details.

Combining topic model into random walk. We now discuss how to combine the topic model into the random walk framework. First, we apply the ACT model to the source and the target domains respectively and obtain two sets of topic distributions. Then we estimate the alignment between topics of these two domains. We calculate the alignment according to the historic cross-domain collaborations. Specifically, the strength of the alignment between topic z_i from the source domain and topic z'_j from the target domain is estimated by:

$$S_{z_i z'_j} = \frac{1}{\kappa}\sum_{(v,v')\in E^{ST}} [P(z_i|v) + P(z'_j|v')] \tag{2}$$

where κ is a normalization factor; $(v, v') \in E^{ST}$ indicates a cross-domain collaboration between v and v'.

We augment the graph generated in the author matching method with topic nodes $\{z\}$ and $\{z'\}$ extracted from the two domains. Figure 2(b) shows the graphical structure, which suggests that a random walk can be performed from a user v to a topic z and from

[2]The ACT model can be considered as an extension of LDA [4], but considers the collaborative relationships between users and the difference of various objects (e.g., author, paper, and conference/journal).

Input: a source domain G^S and a target domain G^T
Output: estimated parameters θ,θ',ϕ, ϑ, and λ

Initialize an ACT model in G^S by learning from documents written by authors only from G^S;
Similarly, initialize an ACT model for target domain G^T;
foreach *collaborated document* d **do**
  foreach *word* $x_{di} \in d$ **do**
    Toss a coin s_{di} according to $bernoulli(s_{di}) \sim beta(\gamma_t, \gamma)$, where $beta(.)$ is a Beta distribution, and γ_t and γ are two parameters;
    if $s_{di} = 0$ **then**
      Randomly select a pair (v, v') from d's authors, where v is an author from G^S and v' from G^T;
      Draw a topic $z_{di} \sim multi(\vartheta_{vv'})$ from the topic mixture $\vartheta_{vv'}$ specific to (v, v');
    end
    if $s_{di} = 1$ **then**
      Randomly select a user v;
      Draw a topic $z_{di} \sim multi(\theta_v)$ from the topic model of user v;
    end
  end
  Draw a word $x_{di} \sim multi(\phi_{z_{di}})$ from z_{di}-specific word distribution;
end

Algorithm 1: Probabilistic generative process in CTL.

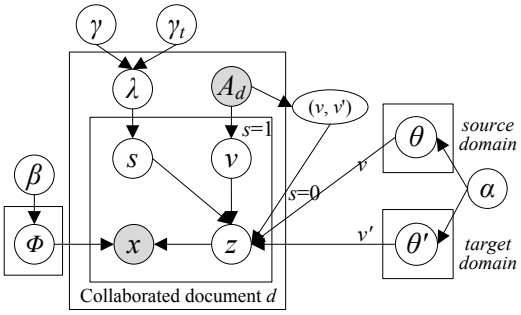


Figure 3: Graphical representation of CTL model.

a topic z of the source domain to a topic z' of the target domain (and vice versa). The link weight between user node v and topic node z is defined as the probability $P(z|v)$ obtained from the ACT model. Then the relatedness of the query node to a target topic z' is defined by a similar formula as that in Eq. 1 and analogously we can define the relatedness between the query node and user nodes in the target domain.

3.3 Cross-domain Topic Learning (CTL)

The topic matching method does not discriminate the "collaboration" topics from those topics existing in only one domain. As a result, the "irrelevant" topics (irrelevant to collaboration) may hurt the collaboration recommendation performance. We develop a new topic modeling approach called Cross-domain Topic Learning (CTL) to model topics of the source domain and the target domain simultaneously.

Model description. The basic idea here is to use two correlated generative processes to model the source and the target domains together. The first process is to model documents written by authors from single domain (either source or target). The second process is to model collaborated documents. For each word in a collaborated document, we use a Bernoulli distribution to determine whether it is generated from a "collaboration" topic or a topic-specific to one domain only. Figure 3 shows the graphical structure of the CTL model. (For simplicity, we omit the modeling part for single domain and focus on the modeling of the collaborated documents.) CTL models each cross-domain collaborated document using topic models of authors from the source domain and the target domain.

Table 1: Notations in the CTL model.

SYMBOL	DESCRIPTION
T	number of topics
d	a collaborated document
A_d	a set of authors of document d
x_{di}	the ith attribute (word) in document d
z_{di}	the topic assigned to attribute x_{di}
s_{di}	if x_{di} is a word from a single domain or a cross domain
θ_v	multinomial distribution over topics specific to author v
$\vartheta_{vv'}$	multinomial distribution over topics specific to author pair (v,v')
ϕ_z	multinomial distribution over words specific to topic z
α, β	Dirichlet priors to multinomial distributions θ, θ' and ϕ
λ	parameter for sampling the binary variable s
γ, γ_t	Beta parameters to generate λ

Let us briefly introduce notations. A_d is a set of authors of document d; v is an author and (v, v') is an author pair randomly sampled to be responsible for word x; s is a binary variable indicating whether the current word inherits the topic from a single domain ($s = 1$) or by a cross-domain collaboration $s = 0$; θ and θ' are topic models from the source domain and the target domain, respectively; $\vartheta_{vv'}$ is a collaboration topic model specific to author pair (v, v'); α is the Dirichlet hyperparameter; λ is a parameter for sampling the binary variable s; γ and γ_t are Beta parameters to generate λ. Table 1 summarizes the notations used in the CTL model.

Formally, the generative process is described in Algorithm 1: first, documents of the two domains G^S and G^T are partitioned into three clusters: documents written by authors only from the source domain, documents written by authors only from the target domain, and documents collaborated by authors from both domains. Then CTL respectively extracts topics of authors from the first two document clusters (without cross-domain collaborations) according to the distribution$p(\theta_v|\alpha)$ and $p(\theta'_{v'}|\alpha)$, where α is the Dirichlet prior. For simplicity, we use the same prior α for both source and target domains.

Second, CTL models the cross domain collaboration documents. For each word x_{di} in document d, a coin s is tossed according to $p(s|d) \sim beta(\gamma_t, \gamma)$, where $beta(.)$ is a Beta distribution. When $s = 1$, a single user v (or v') is chosen according to a uniform distribution, then the word x_{di} is sampled from a selected topic z_{di} specific to the user v, according to θ_v (therefore, this is not a cross-domain collaboration). When $s = 0$, a pair of cross-domain collaborators (v, v') are selected, and a new multinomial distribution $\vartheta_{vv'}$ is constructed by combining θ_v and $\theta_{v'}$ (therefore, cross-domain collaboration is formed). More specifically, we first expand the source and target topic spaces to be of the same dimension. For example, if source domain has 10 topics and target domain 5 topics, the expanded topic space will have 15 topics (10 from source domain and 5 from target domain). The expanded source topic distribution $\tilde{\theta}_v =< \theta_v, 0, \ldots, 0 >$, where we set 0 on the target topics. Similarly, we define the expanded target topic distribution to be $\tilde{\theta}'_{v'} =< 0, \ldots, 0, \theta'_v >$. The new distribution $\vartheta_{vv'}$ is then defined as $\tilde{\theta}_v + \tilde{\theta}_{v'}$, a simple mixture of the two expanded multinomials of θ_v and $\theta_{v'}$ [5]. Finally the word x_{di} is sampled from a collaboration topic z_{di} according to the new distribution $\vartheta_{vv'}$.

Figure 4 illustrates an example of CTL learning. Before CTL learning, each author only has topic distribution in either source or

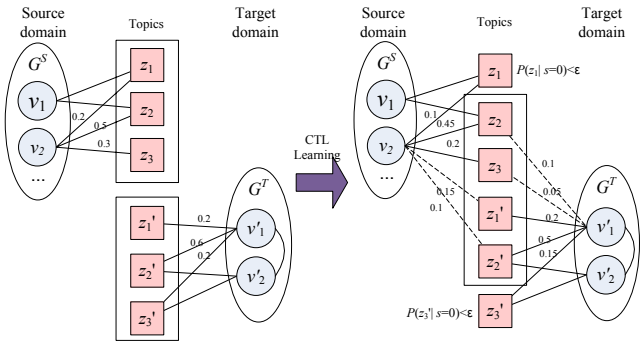


Figure 4: Intuitive explanation of the CTL learning. ϵ is a parameter to select collaboration topics.

target domain (zero probability on topics from the other domain). Then, CTL smoothes topics distributions across the two domains. Users from the source domain will also have a probability over topics extracted from the target domain, and vice versa. After training the CTL model, we also obtain a set of "collaboration topics" between the two domains, i.e., topics with the highest posterior probabilities $P(z|s=0,\cdot)$ (or $P(z|s=0,\cdot)>\epsilon$) in the collaborated documents. (Here, $\cdot$ indicates all the other parameters we should consider when calculating the probability.) For example in right hand side of Figure 4, the box indicates those collaboration topics.

Model inference. We use Gibbs sampling to estimate unknown parameters $\{\theta, \theta', \vartheta, \phi, \lambda\}$ in the CTL model. In particular, we evaluate (a) the posterior distribution on z' (or z) for each word in the document written by authors only from a single domain and then use the results to infer θ' (or θ); (b) the posterior distribution on s, and then use the sampling results of z and z' according to s to update ϑ, θ and θ'. Finally, λ and ϕ can be inferred from the obtained topic models. More specifically, we begin with the joint probability of all documents in the two domains, and then using the chain rule, we obtain the posterior probability of sampling the topic for each word. For (a) we use the same sampling algorithm as that for the LDA model (or the ACT model) (cf. [13] or [31]), i.e. with the posterior probability:

$$P(z_{di}|\mathbf{z}_{-di},\mathbf{x},\cdot) = \frac{n^{-di}_{vz_{di}}+\alpha}{\sum_z(n^{-di}_{vz}+\alpha)} \times \frac{m^{-di}_{z_{di}x_{di}}+\beta}{\sum_x(m^{-di}_{z_{di}x}+\beta)} \quad (3)$$

where n_{vz} is the number of times that topic z has been sampled from the multinomial distribution specific to a randomly selected author v; m_{zx} is the number of times that word x has been generated by topic z; the number n^{-di} with the superscript $-di$ denotes a quantity, excluding the current instance. We use a similar process for both domains.

For parameter estimation in (b), we consider a two-step Gibbs sampling. We first sample the coin s according to the posterior probability: (Detailed derivation is given in Appendix.)

$$P(s_{di}=0|\mathbf{s}_{-di},\mathbf{z},\cdot) = \frac{n^{-di}_{ds_0}+\gamma_t}{n^{-di}_{ds_0}+n^{-di}_{ds_1}+\gamma_t+\gamma} \times \frac{n^{-di}_{vv'z_{di}}+(n_{vz_{di}}+n_{v'z_{di}})+\alpha}{\sum_z(n^{-di}_{vv'z}+(n_{vz_{di}}+n_{v'z_{di}})+\alpha)} \quad (4)$$

where n_{ds_0} is the number of times that $s=0$ has been sampled in document d; (v,v') is the selected user pair to be responsible for x_{di}; $n_{vv'z}$ is the number of times that topic z has been sampled from user pair (v,v'). $P(s_{di}=1|\cdot)$ can be analogously defined as the above equation. The only difference is to replace the sum of the two terms $(n_{vz_{di}}+n_{v'z_{di}})$ with the number by a selected single user v (or v').

The posterior probability of topic z is defined as:

$$P(z_{di}|s_{di}=0,\mathbf{x},\mathbf{z}_{-di},\cdot) = \frac{m^{-di}_{z_{di}x_{di}}+m_{z_{di}x_{di}}+m'_{z_{di}x_{di}}+\beta}{\sum_x(m^{-di}_{z_{di}x}+m_{z_{di}x}+m'_{z_{di}x}+\beta)} \times \frac{n^{-di}_{vv'z_{di}}+(n_{vz_{di}}+n_{v'z_{di}})+\alpha}{\sum_z(n^{-di}_{vv'z}+(n_{vz}+n_{v'z})+\alpha)} \quad (5)$$

where m^{-di}_{zx} is the number of times that word x has been generated by topic z in the collaborated documents; m_{zx} and m'_{zx} respectively represents the number of times that word x has been generated by topic z in the source domain and that in the target domain.

During the parameter estimation, the algorithm keeps track of a $V\times T$ (user by topic) count matrix for both domains, a $D\times 2$ (collaborated document by coin), a $2\times T$ (coin by topic) count matrices, and a $AP\times T$ (user pair by topic) count matrix (AP is the number of user pairs). Given these matrices, we can estimate the probabilities of θ, θ', ϑ, ϕ, and λ.

Cross-domain recommendation via random walk. We combine the learned "collaboration" topics by CTL into the collaboration graph (Cf. Figure 2(c)). In principle, there could be a link between any user node and topic node (the difference is the link weight). To control the density of the constructed network, we define a parameter ϵ and add links between users and topics only when $P(z|s=0,\cdot)>\epsilon$. A smaller ϵ results in a more dense network. Random walk with restart is then performed on the topic augmented graph to calculate the relatedness between users from the target domain and the query user node in an analogous way as done in Eq. 1. Finally we rank users in the target domain according to the estimated relatedness scores and recommend users with the highest relatedness. One advantage of the CTL model is that it is able to recommend "related" collaboration topics based on the relatedness scores between the query node and the topic nodes. In topic matching, we could also consider recommending topics based on the relatedness scores; however, the recommended topics might be irrelevant to collaboration. In CTL, the recommended topics directly reflect existing collaborations across the two domains.

The CTL model can be also generalized to multiple domains. The basic idea is to use a multinomial distribution to replace the Bernoulli distribution. The multinomial represents collaboration topics among multiple domains, between two specific domains, or those in single domain. Based on the learned topics, we can construct a topic-centered network (similar to Figure 2(c)). Then the random walk with restart can be performed on the network to estimate the relatedness scores of users from different domains.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed methods on large publication datasets of different domains. All data sets and codes are publicly available[3].

4.1 Experimental Setup

Data sets. The data set is extracted from Arnetminer.org [31], an academic search system, which contains 1,436,990 authors and

[3]http://arnetminer.org/collaboration

1,932,442 publications. The data we used in our experiments spans from 1990 to 2005. We consider the following five sub-domains:

- **Data Mining:** We use papers of the following data mining conferences: KDD, SDM, ICDM, WSDM and PKDD as ground truth, which result in a network with 6,282 authors and 22,862 co-author relationships.
- **Medical Informatics:** We include the following journals: Journal of the American Medical Informatics Association, Journal of Biomedical Informatics, Artificial Intelligence in Medicine, IEEE Trans. Med. Imaging and IEEE Transactions on Information and Technology in Biomedicine, from which we obtain a network of 9,150 authors and 31,851 co-author relationships.
- **Theory:** We include the following conferences, i.e., STOC, FOCS and SODA, from which we get 5,449 authors and 27,712 co-author relationships.
- **Visualization:** We include the following conferences and journals, CVPR, ICCV, VAST, TVCG, IEEE Visualization and Information Visualization. The obtained coauthor network is comprised of 5,268 authors and 19,261 co-author relationships.
- **Database:** We include the following conferences, i.e., SIGMOD, VLDB and ICDE. From those conferences, we extract 7,590 authors and 37,592 co-author relationships.

Based on the above five sub domains, we create four cross-domain test cases: *Data Mining* to *Theory*, *Medical Informatics* to *Database*, *Medical Informatics* to *Data Mining*, and *Visualization* to *Data Mining*.

Comparison methods. We compare the following methods for collaboration recommendation:

Content Similarity (Content): It calculates similarity between authors based on papers published by them. Specifically, we construct feature vector $\mathbf{w}_q$ and $\mathbf{w}_{v'}$ of words used in papers published by query author q and target author v', respectively. Those feature vectors are normalized by TFIDF [1]. The similarity score is the Cosine similarity between $\mathbf{w}_q$ and $\mathbf{w}_{v'}$

$$Sim(v_q, v') = \frac{\mathbf{w}_q \cdot \mathbf{w}_{v'}}{\|\mathbf{w}_q\|\|\mathbf{w}_{v'}\|} \quad (6)$$

Collaborative Filtering (CF): It leverages the existing collaborations to make the recommendation. The basic idea is that if a query author q has the same or similar collaborators as a person x within the same domain, q is then likely to have the same cross-domain collaborators as x. We employ a memory-based collaborative filtering algorithm [8], in which recommendations are made for a query user q using the following formula:

$$CF\_score(q, v') = \sum_{x \in V^S} I(x, v') r(q, x) \quad (7)$$

where $r(q, x)$ is the similarity between authors in the source domain, e.g., Cosine similarity based on collaboration connections; the indicator variable $I(x, v')$ is 1 if the author x has a cross-domain collaboration with v' and 0 otherwise.

Hybrid: It considers a linear combination of the scores obtained by the Content and the CF methods, specifically,

$$Hybrid(v_q, v') = \mu CF\_score(v_q, v') + (1-\mu) Sim(v_q, v') \quad (8)$$

where μ is a balance parameter. We empirically set it as 0.5.

Katz: It is the best link predictor in [20]. It sums over all possible paths between the query user and a candidate user, and then use the summation score to rank all candidates.

Author Matching: (Cf. §3.1) It makes recommendation by performing the random walk with restart on the collaboration graph.

Topic Matching: (Cf. §3.2) It makes recommendation by combining the extracted topics into random walking algorithm.

CTL: (Cf. §3.3) It is the proposed method, which considers topic skewness and extracts relevant topics to cross-domain collaboration. The relevant topics are then integrated into the random walk framework for recommendation.[4]

Evaluation metrics. To quantitatively evaluate the proposed methods, in each test case, we use historic collaboration data (data before 2001) for training and the last four years (2001-2005) for validation. In evaluation, we consider those candidates who already have cross-domain collaborations and then our task is to predict if they will maintain the collaborations or expand new cross-domain collaborations. If the system recommends a cross-domain collaboration and later the collaboration has been built, then we say the system made a correct recommendation; otherwise we say the system made a wrong recommendation. Based on this, we evaluate the recommendation performance in terms of P@10 (Precision for the top 10 recommended results), P@20, R@100 (Recall for the top 100 results), MAP (Mean Average Precision), and Average Reciprocal Hit-Rank (ARHR) [9].

All codes are implemented in C++, and all the experiments are conducted on an x64 server with E7520 1.87GHz Intel Xeon CPU and 128G RAM. The operation system is Microsoft Windows Sever 2008 R2 Enterprise. For training the ACT and the CTL models, it takes about 12 hours and 15 hours respectively on the entire data set (1,436,990 authors and 1,932,442 publications). Recognizing the computation complexity of LDA style models, we are currently looking into developing more efficient computation mechanism to speed up the process.

4.2 Recommendation Performance Analysis

Table 2 lists the performance of cross-domain collaboration recommendation by the comparison methods on the four different test cases. The proposed CTL method clearly outperforms the baseline methods (+2.2-30% in terms of MAP). Content only considers the content information, which leads to a bad performance. The two methods (Hybrid and Topic Matching), combining the content and the network information, improve the recommendation performance compared to the simple baselines such as Content, CF and Author Matching. Moreover, Topic Matching considers the topic information extracted from the two domains, and thus performs better than the Hybrid method adopting a simple combination. CTL differentiates "collaboration topics" from those irrelevant topics and obtains significant improvement over both Hybrid and Topic Matching.

Cross-domain topics analysis. How many topics are enough for the cross-domain recommendation? We perform an analysis by varying the number of cross-domain topics in the proposed CTL method. Figure 5(a) shows its MAP performance with the num-

[4]As for the hyperparameters α, α_t, and β, following LDA [4], we empirically take fixed values (i.e., $\alpha = \alpha_q = 50/T$, and $\beta = 0.01$). γ and γ_t are defined to represent our preference for cross-domain collaborations (i.e., $\gamma_q = 3.0$ and $\gamma = 0.1$). We did try different settings and found that the estimated topic models are not very sensitive to the hyperparameters.

Table 2: Recommendation performance by different methods on the four cross-domain test cases (%). Content− Content Similarity; CF− Collaborative Filtering; Author− Author Matching; Topic− Topic Matching.

Cross domain	ALG	P@10	P@20	MAP	R@100	ARHR -10	ARHR -20
Data Mining (S) to Theory (T)	Content	10.3	10.2	10.9	31.4	4.9	2.1
	CF	15.6	13.3	23.1	26.2	4.9	2.8
	Hybrid	17.4	19.1	20.0	29.5	5.0	2.4
	Author	27.2	22.3	25.7	32.4	10.1	6.4
	Topic	28.0	26.0	32.4	33.5	13.4	7.1
	Katz	30.4	29.8	31.6	27.4	11.2	5.9
	CTL	**37.7**	**36.4**	**40.6**	**35.6**	**14.3**	**7.5**
Medical Info. (S) to Database (T)	Content	10.1	10.9	12.5	45.9	3.6	2.1
	CF	18.3	20.2	21.4	47.6	5.3	3.9
	Hybrid	25.0	26.5	28.4	59.1	6.4	4.2
	Author	26.2	29.6	32.2	54.8	10.5	**5.4**
	Topic	29.4	26.3	34.7	59.3	**11.5**	5.2
	Katz	27.5	28.3	30.7	57.2	10.5	5.0
	CTL	**32.5**	**30.0**	**36.9**	**59.8**	11.4	**5.4**
Medical Info. (S) to Data Mining (T)	Content	5.8	5.7	9.5	19.8	1.9	0.9
	CF	13.7	17.8	18.9	34.3	2.7	1.3
	Hybrid	18.0	19.0	19.8	36.7	3.4	1.3
	Author	20.1	23.8	29.3	**64.4**	5.3	2.1
	Topic	26.0	**25.0**	33.9	48.1	10.7	5.6
	Katz	21.2	23.8	32.4	48.1	10.2	4.8
	CTL	**30.0**	24.0	**35.6**	49.6	**12.2**	**6.0**
Visual. (S) to Data Mining (T)	Content	9.6	11.8	13.2	18.9	3.1	1.8
	CF	14.0	20.8	26.4	29.4	6.9	4.3
	Hybrid	16.0	20.0	27.6	30.1	6.3	4.4
	Author	22.0	25.2	27.7	31.1	11.9	6.7
	Topic	26.3	25.0	32.3	31.4	13.2	8.8
	Katz	23.0	25.1	29.3	30.2	10.4	5.4
	CTL	**28.3**	**26.0**	**32.8**	**36.3**	**14.0**	**9.1**

ber of cross-domain topics varied. We see, when the number is small (< 80), increasing the number often obtains a performance improvement. The trend becomes stable when the number is up to 150. This demonstrates the stability of the CTL method with respect to the number of topics.

Hyperparameter analysis. We use α as the example to analyze how hyperparameter influences the performance of the CTL method. Figure 5(b) shows the performance of CTL with the parameter α varied (all the other hyperparameters fixed and the number of topics is set as $T = 120$). We see although the performance changes when varying the value of α, the largest difference is less than 0.03 This confirms CTL method is not sensitive to the particular choice of α.

Restart parameter analysis. We study how the parameter τ influences the process of random walk with restart. Figure 5(c) plots the performance of the CTL method on the four test cases with the parameter τ varied. In general, the recommendation performance is not sensitive to the restart parameter τ. By a careful investigation, we find that a small τ makes the random walk diffuse too quickly thus can hurt the precision, while a large τ limits the diffusion process and thus can result in a lower recall.

Convergence analysis. We further investigate the convergence of the random walk with restart algorithm. Figure 5(d) shows the convergence analysis of different models on the test case of Visualization-Data Mining. We see all the three models converge within 10 iterations and CTL achieved even faster convergence

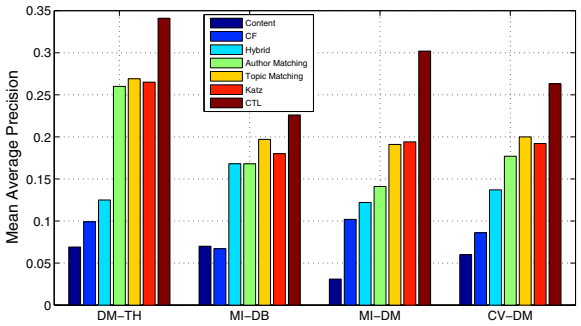


Figure 6: Performance on new collaboration prediction of all algorithms.

(within 5 iterations). This fast convergence on CTL model enable real time query support that is crucial in the deployed system we will discuss next.

New Collaboration Prediction The collaboration network is dynamic in nature, with collaborative relationships created over time. In general, there are two types of collaborative behaviors, maintaining existing collaborations and building new collaborations. Can we predict who will create a new collaboration in the future? This is a more difficult task. We conduct an experiment to evaluate the performance of the proposed method for new collaboration prediction. In particular, we still use the publication data before 2001 for training and the data between 2001-2005 for test, and in the evaluation, we only consider new collaborations in the test data. Figure 6 shows the performance of new collaboration prediction by the six comparison algorithms. On average, the performance of all algorithms drops a bit, but all algorithms have similar behaviors as that in Table 2. In particular, it is exciting to see that CTL can still maintain about 0.3 in terms of MAP which is significantly higher than the baseline methods.

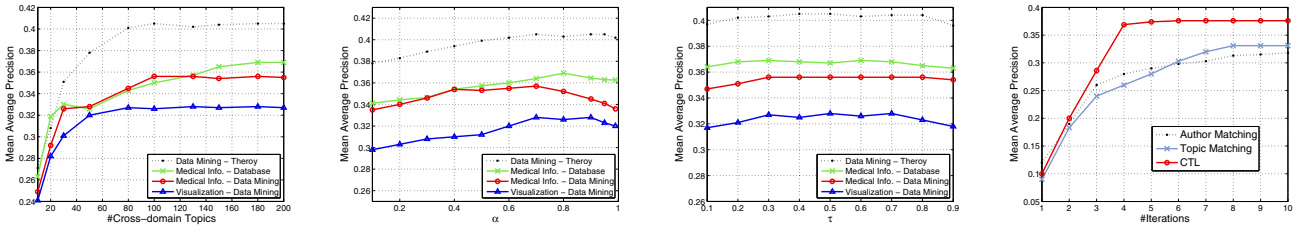
4.3 Prototype System

We have developed and deployed a web application for cross-domain recommendation based on the proposed CTL method[5]. The system trained a CTL model offline using all the publication data (about 1,932,442 publication papers) in Arnetminer.org. When a user wants to find cross-domain collaborators, he first inputs his profile (including organization and research interest) or use an existing author profile via the Arnetminer system, which includes more than 1 million researcher profiles. Then the user inputs the target domain (by keywords) in which he wants to find collaborations. The system performs the random walk with restart algorithm (Cf. §3.3) online against the CTL model to rank potential topics/collaborators in the target domain.

5. RELATED WORK

Collaboration recommendation plays an important role in many fields and has attracted a lot of research interest. Chen et al. [7] have developed a system called CollabSeer for discovering potential collaborators for a given author based on the structure of the coauthor network and the user's research interests. This is the most relevant paper to our work. However, it does not consider the cross-domain problem. Konstas et al. [17] investigated how social relationships can help recommendation. They developed a

[5]http://arnetminer.org/collaborator



(a) number of topics T (b) Hyperparameter α (c) RWR parameter τ (d) Convergence analysis

Figure 5: Parameter analysis. (a) Performance of cross-domain topic learning model by varying the number of topics T; (b) Performance of cross-domain topic learning (CTL) is stable when varying α parameter; (c) Performance of CTL is stable when varying the restart parameter τ in the random walk process on the four test cases; (d) Convergence analysis of different models on the test case of Visualization-Data Mining.

track recommendation system by considering both social annotation and friendship inherent in the social graph established among users, items and tags. Kautz et al. [16] introduced a system called ReferralWeb which attempts to combine social networks for collaborative filtering. There are a large body of research on collaborative filtering. For example [2] introduced a system called Fab by combining content-based filtering and collaborative filtering. Shi et al. [26] proposed a large scale machine learning system for recommending heterogeneous content in social networks and Sculley et al. [25] presented a method to rank which combines regression and ranking. Yuan et al. [35] aimed to fuse heterogeneous social relationships for recommendation using factorization and regularization technologies. Wang and Blei [34] developed an algorithm to recommend scientific articles to users of an online community by combining traditional collaborative filtering and probabilistic topic modeling. However, most existing works only consider the recommendation problem within one single domain, but do not consider the cross-domain recommendation problem. In addition, we propose a novel cross-domain topic learning method, which supports recommending collaboration topics as well.

Our work is also related to expert finding [3, 30, 36] and expertise matching [23, 33]. Mimno et al. [23] and Tang et al. [33] studied the problem of paper-reviewer recommendation, a subtask of expert finding. The proposed algorithms can be leveraged for collaboration recommendations. However, expert finding and expertise matching are in nature different from the problem of collaboration recommendation. The idea of differentiating irrelevant topics has been also studied in previous work such as the query-oriented topic model (qLDA) proposed in [29], which tries to identify relevant topics to queries in multi-document summarization.

6. CONCLUSION

In this paper, we study the problem of cross-domain collaboration recommendation. We precisely define the problem and present three models for ranking and recommending potential collaborators. A cross-domain topic modeling approach has been proposed to learn and differentiate collaboration topics from other topics. Experimental results in a coauthor network demonstrate the effectiveness and efficiency of the proposed approach.

As for the future work, it is intriguing to connect cross-domain collaborative relationships with social theories. For example, how cross-domain relationships correlate with strong/weak ties [12] and how such correlation can help spread knowledge from one domain to another domain. It would be also interesting to apply the proposed method to other networks, e.g., software development.

Acknowledgements. The work is supported by the Natural Science Foundation of China (No. 61073073, No. 61170061) and Chinese National Key Foundation Research (No. 60933013, No.61035004), 973 Program (No. 2011CB302302), a special fund for Fast Sharing of Science Paper in Net Era by CSTD, and Tsinghua-Tencent innovation funding.

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.
[2] M. Balabanović and Y. Shoham. Fab: content-based, collaborative recommendation. *Commun. ACM*, 40:66–72, March 1997.
[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR'06*, pages 43–55, 2006.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.
[5] W. Buntine and A. Jakulin. Applying discrete pca in data analysis. In *UAI'04*, pages 59–66, 2004.
[6] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.
[7] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Collabseer: a search engine for collaboration discovery. In *JCDL'11*, pages 231–240, 2011.
[8] A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: Scalable online collaborative filtering. In *WWW'07*, 2007.
[9] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 22(1):143–177, Jan. 2004.
[10] A. Doucet, N. de Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *UAI'00*, pages 176–183, 2000.
[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM'99*, pages 251–262, 1999.
[12] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.
[13] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS'04*, pages 5228–5235, 2004.
[14] G. Heinrich. Parameter estimation for text analysis. Technical report, University of Leipzig, Germany, 2004.
[15] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.
[16] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.
[17] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *SIGIR'09*, pages 195–202, 2009.
[18] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD'06*, pages 631–636, 2006.
[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW'10*, pages 641–650, 2010.
[20] D. Liben-Nowell and J. M. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.
[21] R. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD'10*, pages 243–252, 2010.

[22] L. Lovasz. Random walks on graphs: A survey. *Combinatorics*, 2(1):1?6, 1993.
[23] D. Mimno and A. McCallum. Expertise modeling for matching papers with reviewers. In *KDD'07*, pages 500–509, 2007.
[24] J. Quackenbush. Microarray analysis and tumor classification. *New England Journal of Medicine*, 354:2463–2472, June 2006.
[25] D. Sculley. Combined regression and ranking. In *KDD'10*, pages 979–988, 2010.
[26] Y. Shi, D. Ye, A. Goder, and S. Narayanan. A large scale machine learning system for recommending heterogeneous content in social networks. In *SIGIR'11*, pages 1337–1338, 2011.
[27] M. Steyvers, P. Smyth, and T. Griffiths. Probabilistic author-topic models for information discovery. In *KDD'04*, pages 306–315, 2004.
[28] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM'05*, pages 418–425, 2005.
[29] J. Tang, L. Yao, and D. Chen. Multi-topic based query-oriented summarization. In *SDM'09*, pages 1147–1158, 2009.
[30] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su. Topic level expertise search over heterogeneous networks. *Machine Learning Journal*, 82(2):211–237, 2011.
[31] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.
[32] L. Tang and H. Liu. Relational learning via latent social dimensions. In *KDD'09*, pages 817–826, 2009.
[33] W. Tang, J. Tang, T. Lei, C. Tan, B. Gao, and T. Li. On optimization of expertise matching with various constraints. *Neurocomputing*, 76(1):71–83, 2012.
[34] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD'11*, pages 448–456, 2011.
[35] Q. Yuan, L. Chen, and S. Zhao. Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. In *RecSys'11*, pages 245–252, 2011.
[36] J. Zhang, J. Tang, and J. Li. Expert finding in a social network. In *DASFAA'07*, pages 1066–1069, 2007.

8. APPENDIX

According to the generative process, we could integrate out the multinomial (Bernoulli) distributions $\theta, \theta', \vartheta, \lambda, \phi$, because the model only uses conjugate priors [10]. We use Eq. 4 as the example to explain its derivation. First we write the joint probability:

$$\begin{aligned}
&P(\mathbf{x}, \mathbf{x}', \mathbf{z}, \mathbf{z}', \mathbf{s}, \mathbf{v}, \mathbf{v}'|\alpha, \gamma, \gamma_t, \beta, \mathbf{A}) \\
\propto &\int P(\mathbf{s}|\lambda)P(\lambda|\gamma, \gamma_t)d\lambda \int P(\mathbf{v}|\mathbf{A})P(\mathbf{z}|\mathbf{v}, \mathbf{s}, \theta)P(\theta|\alpha)d\theta \\
&\int P(\mathbf{v}'|\mathbf{A}')P(\mathbf{z}'|\mathbf{v}', \mathbf{s}, \theta')P(\theta'|\alpha)d\theta' \int P(\mathbf{x}|\mathbf{z}, \phi)P(\phi|\beta)d\phi \\
&\int P((\mathbf{v}, \mathbf{v}')|\mathbf{A})P(\mathbf{z}|(\mathbf{v}, \mathbf{v}'), \mathbf{s}, \vartheta)P(\vartheta|\alpha)d\vartheta
\end{aligned} \tag{9}$$

The conditional of s_i is obtained by dividing the joint distribution of all variables by the joint with all variables but s_i (denoted by $\mathbf{s}_{-i}$) and canceling factors that do not depend on $\mathbf{s}_{-i}$.

$$\begin{aligned}
p(s_i = 0|\mathbf{s}_{-i}, \mathbf{z}, .) &= \frac{P(\mathbf{x}, \mathbf{x}', \mathbf{z}, \mathbf{z}', \mathbf{s}, \mathbf{v}, \mathbf{v}'|\alpha, \gamma, \gamma_t, \beta, \mathbf{A})}{P(\mathbf{x}, \mathbf{x}', \mathbf{z}, \mathbf{z}', \mathbf{s}_{-i}, \mathbf{v}, \mathbf{v}'|\alpha, \gamma, \gamma_t, \beta, \mathbf{A})} \\
&= \frac{\int P(\mathbf{s}|\lambda)P(\lambda|\gamma, \gamma_t)d\lambda}{\int P(\mathbf{s}_{-i}|\lambda)P(\lambda|\gamma, \gamma_t)d\lambda} \\
&\cdot \frac{\int P((\mathbf{v}, \mathbf{v}')|\mathbf{A})P(\mathbf{z}|(\mathbf{v}, \mathbf{v}'), \mathbf{s}, \vartheta)P(\vartheta|\alpha)d\vartheta}{\int P((\mathbf{v}, \mathbf{v}')|\mathbf{A})P(\mathbf{z}|(\mathbf{v}, \mathbf{v}'), \mathbf{s}_{-i}, \vartheta)P(\vartheta|\alpha)d\vartheta}
\end{aligned} \tag{10}$$

We now derive the first fraction of Eq. 10. As we assume that s_i is generated from a Bernoulli distribution λ whose Beta parameters are γ, γ_t, then we can get $p(\mathbf{s}|\lambda) = \prod_d \lambda_d^{n_{ds_0}} \cdot (1-\lambda_d)^{n_{ds_1}}$, where n_{ds_0} is the number of times that $s = 0$ has been sampled in document d and n_{ds_1} represents the number of times that $s = 1$ has been sampled in d. Because Beta is the conjugate prior of Bernoulli, we could solve the Bernoulli-Beta integral using Gibbs sampling. Specifically,

$$\begin{aligned}
&\int P(\mathbf{s}|\lambda)P(\lambda|\gamma, \gamma_t)d\lambda \\
&= \prod_d \frac{1}{B(\gamma_t, \gamma)} \int_0^1 \lambda_d^{n_{ds_0}+\gamma_t-1}(1-\lambda_d)^{n_{ds_1}+\gamma-1}d\lambda_d \\
&= \prod_d \frac{B(n_{ds_0}+\gamma_t, n_{ds_1}+\gamma)}{B(\gamma_t, \gamma)} \\
&= \prod_d \frac{\Gamma(n_{ds_0}+\gamma_t)\Gamma(n_{ds_1}+\gamma)\Gamma(\gamma_t+\gamma)}{\Gamma(n_{ds_0}+n_{ds_1}+\gamma_t+\gamma)}
\end{aligned} \tag{11}$$

To yield the first fraction of Eq. 10, we apply the above equation twice and obtain the following equation:

$$\begin{aligned}
\frac{\int P(\mathbf{s}|\lambda)P(\lambda|\gamma, \gamma_t)d\lambda}{\int P(\mathbf{s}_{-i}|\lambda)P(\lambda|\gamma, \gamma_t)d\lambda} &= \frac{\prod_d \frac{\Gamma(n_{ds_0}+\gamma_t)\Gamma(n_{ds_1}+\gamma)\Gamma(\gamma_t+\gamma)}{\Gamma(n_{ds_0}+n_{ds_1}+\gamma_t+\gamma)}}{\prod_d \frac{\Gamma(n_{ds_0}^{-di}+\gamma_t)\Gamma(n_{ds_1}^{-di}+\gamma)\Gamma(\gamma_t+\gamma)}{\Gamma(n_{ds_0}^{-di}+n_{ds_1}^{-di}+\gamma_t+\gamma)}} \\
&= \frac{n_{ds_0}^{-di}+\gamma_t}{n_{ds_0}^{-di}+n_{ds_1}^{-di}+\gamma_t+\gamma}
\end{aligned} \tag{12}$$

Here, we use the identity $\Gamma(x+1) = x\Gamma(x)$; the superscript $-di$ denotes a quantity, excluding the current instance. The second fraction of Eq. 10 can be derived analogously. Specifically, as $P((\mathbf{v}, \mathbf{v}')|\mathbf{A})$ is a uniform distribution, $P(\mathbf{z}|(\mathbf{v}, \mathbf{v}', \mathbf{s}, \vartheta)$ and $P(\vartheta|\alpha)$ are conjugate pair of Multinomial-Dirichlet, we can obtain [14]:

$$\begin{aligned}
&\int P((\mathbf{v}, \mathbf{v}')|\mathbf{A})P(\mathbf{z}|(\mathbf{v}, \mathbf{v}'), \mathbf{s}, \vartheta)P(\vartheta|\alpha)d\vartheta \\
&= \prod_d \frac{1}{\sigma(A_d)} \cdot \frac{1}{\Delta(\alpha)} \prod_z \vartheta_{vv'z}^{n_{vz}+n_{v'z}+n_{vv'z}+\alpha-1}d\vartheta_{vv'} \\
&= \prod_d \frac{1}{\sigma(A_d)} \frac{\Delta(\vec{n_d}+\alpha)}{\Delta(\alpha)}, \\
&\text{with } \vec{n_d} = \{n_{vz}+n_{v'z}+n_{vv'z}\}_{z=1}^T
\end{aligned} \tag{13}$$

where $\sigma(A_d)$ is the total number of cross-domain user pairs generated from authors of document d (for a specific document, the number will be a constant); $\Delta(\alpha) = \frac{\Gamma(\alpha)^T}{\Gamma(T\alpha)}$; $n_{vv'z}$ denotes the number of times that topic z has been sampled by user pair (v, v'); n_{vz} and $n_{v'z}$ are two numbers obtained when combining the two distributions θ_v and $\theta_{v'}$; please note that though we write it as the sum of the two numbers, in practice, when sampling a specific topic, we will only consider one of them. This is because, for example, if a topic z is from the source domain, the number $n_{v'z}$ will be 0. Accordingly, the second fraction of Eq. 10 can be written as:

$$\begin{aligned}
&\frac{\int P((\mathbf{v}, \mathbf{v}')|\mathbf{A})P(\mathbf{z}|(\mathbf{v}, \mathbf{v}'), \mathbf{s}, \vartheta)P(\vartheta|\alpha)d\vartheta}{\int P((\mathbf{v}, \mathbf{v}')|\mathbf{A})P(\mathbf{z}|(\mathbf{v}, \mathbf{v}'), \mathbf{s}_{-i}, \vartheta)P(\vartheta|\alpha)d\vartheta} \\
&= \frac{\prod_d \frac{1}{\sigma(A_d)} \frac{\Delta(\vec{n_d}+\alpha)}{\Delta(\alpha)}}{\prod_d \frac{1}{\sigma(A_d)} \frac{\Delta(\vec{n}_{d,\neg i}+\alpha)}{\Delta(\alpha)}} \\
&= \frac{\frac{\Gamma(n_{vv'z}+n_{vz}+n_{v'z}+\alpha)}{\Gamma(\sum_{z'}(n_{vv'z'}+n_{vz'}+n_{v'z'}+\alpha))}}{\frac{\Gamma(n_{vv'z}+n_{vz}+n_{v'z}+\alpha-1)}{\Gamma([\sum_{z'}(n_{vv'z'}^{-di}+n_{vz'}+n_{v'z'}+\alpha)]-1)}} \\
&= \frac{n_{vv'z_{di}}^{-di}+(n_{vz_{di}}+n_{v'z_{di}})+\alpha}{\sum_z(n_{vv'z}^{-di}+(n_{vz_{di}}+n_{v'z_{di}})+\alpha)}
\end{aligned} \tag{14}$$

Finally, by combining Eqs. 12 and 14, we obtain Eq. 4.