# Community Detection by Popularity Based Models for Authored Networked Data

Tianbao Yang[†] and Prakash Mandaym Comar[‡] and Linli Xu[♮]
[†]GE Global Research, San Ramon, CA 94583, USA
[‡]Michigan State University, East Lansing, MI 48823, USA
[♮]University of Science and Technology of China, Hefei, Anhui 246133, China
Email: tyang@ge.com, mc.prakash@gmail.com, linlixu@ustc.edu.cn

***Abstract*—Community detection has emerged as an attractive topic due to the increasing need to understand and manage the networked data of tremendous magnitude. Networked data usually consists of links between the entities and the attributes for describing the entities. Various approaches have been proposed for detecting communities by utilizing the link information and/or attribute information. In this work, we study the problem of community detection for networked data with additional authorship information. By authorship, each entity in the network is authored by another type of entities (e.g., wiki pages are edited by users, products are purchased by customers), to which we refer as authors. Communities of entities are affected by their authors, e.g., two entities that are associated with the same author tend to belong to the same community. Therefore leveraging the authorship information would help us better detect the communities in the networked data. However, it also brings new challenges to community detection. The foremost question is how to model the correlation between communities and authorships. In this work, we address this question by proposing probabilistic models based on the popularity link model [1], which is demonstrated to yield encouraging results for community detection. We employ two methods for modeling the authorships: (i) the first one generates the authorships independently from links by community memberships and popularities of authors by analogy of the popularity link model; (ii) the second one models the links between entities based on authorships together with community memberships and popularities of nodes, which is an analog of previous author-topic model. Upon the basic models, we explore several extensions including (i) we model the community memberships of authors by that of their authored entities to reduce the number of redundant parameters; and (ii) we model the communities memberships of entities and/or authors by their attributes using a discriminative approach. We demonstrate the effectiveness of the proposed models by empirical studies.**

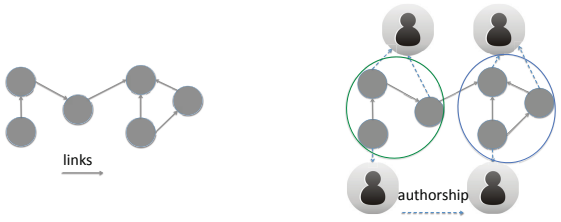
## I. Introduction

The last decade has witnessed massive generation of relational data or networked data from across different branches of scientific fields. From the well known social networks, to biological protein networks, gene networks, chemical networks, geographic weather networks and financial transaction networks, it has become ubiquitous for data mining practitioners to work with networked data of varied types from different domains and characteristics. An important mining task on the networked data is community detection, which involves identifying cohesive sub-group of nodes in the network. This task has a wide variety of applications involving storage, retrieval and inference on networked data.

Past research on community detection focused on networks from a single source. Lately, the networks have become more diverse and scattered across different sources (domains). It is very common to find cross-links between entities in different networks like Youtube, Flickr, Twitter and Facebook. It is also both interesting and insightful to extract groups of similar likings or taste in each network where the taste depends on the user information scattered across multiple networks. Thus, to improve the modeling of individuals, it is important to develop mining algorithms that combine information from the multiple networks.

In this paper we propose probabilistic models for community detection of target entities from two sources of links, namely within-domain links between target entities and cross-domain links between target entities and peripheral entities, as well as additional attributes for describing the entities that are tied to those links. In particular, we study a class of cross-domain links, where the target entities are authored by the second type of entities, to which we refer as authors. We also refer to such networked data as authored networked data. We note that authored networked data are pervasive in real world applications. A simple example is bibliographic data of authored articles that cite each other. Other examples include inter-connected webpages edited by different editors, co-purchased products brought by different customers, similar video clips commented by different users, connected people employed by different companies, interactive proteins reside in different cells and etc. An illustrative example in Figure 1 shows the authorship information can improve the clustering results of their owned entities. The community of the middle node is vague when we only utilize the links between the nodes. However, if we consider the authorship information, it is clearly identified to belong to the first community on the left.

The challenge in detecting communities for authored networked data is how to model the correlations between the communities and the links, the attributes and the authorships. Part of the problems have been addressed successfully in a prominent work [1], which proposed a popularity conditional link (PCL) model and a discriminative content (DC) model to leverage both the link information and the content information for community detection. The PCL model introduces random variables named popularities to model the links such that it can fit the scale-free behavior (or equivalently pow-law degree distribution) that occurs pervasively in real networks. The DC model alleviates the affect of those attributes that are irrelevant to community memberships. The PCL-DC model has




(a) a network (b) an authored network

Fig. 1. An Illustration Example: authorship information can help community detection.

observed significant improvements over traditional approaches of combining link information and attribute information for community detection, and therefore we build our models based on the PCL-DC model. However, the remaining question is how to model the authorship information. The contribution of the work is summarized as follows:

- We propose and study two different methods of modeling the authorship information for community detection.
- We complete the family of popularity models for community detection by leveraging links, attributes and authorships.
- We present EM algorithms for estimating the model parameters and conduct empirical studies on three data sets to demonstrate the effectiveness of the proposed models.

This paper is organized as follows. In Section II, we discuss the relevant literature and motivate the proposed approach. In Section III, we give a formal definition of the problem and present a generative model as the baseline. In Section III-B and Section III-C, we propose two models for detecting communities in authored networked data. In Section IV, we present extensions over the basic models. In Section V, we present the EM algorithms for maximizing the log-likelihood of the proposed models. Finally, in the Section VI, we present our experimental results.

## II. Related Work

Early research on community detection on networks has focused on graph partitioning techniques [2], [3], [4]. The goal is to partition the graph in such a way that the intra-cluster links (links within each partition) are maximized and inter-cluster links (links between partitions) are minimized. To this extent, several approaches have been successfully applied, including techniques based on multi-level graph partitioning [5], [6], and matrix factorization approaches [7], spectral clustering [8], [9], [10]. Recently, probabilistic models have evolved as a major approach to detect the communities in a network. Most cited work include stochastic block models [11], [12], [13], PHITS model [14] and Popularity Conditional Link model [1]. All of these algorithms focused on partitioning a homogeneous networks into clusters or communities.

Built upon Latent Dirichlet Allocation (LDA) [15], many generative models have been proposed and applied to networked data for mining topics and communities. We briefly describe some models by emphasizing what data are used and how they are modeled. LDA [15] is a Bayesian extension of PLSA model [16], which generates the words of documents by adding an intermediate layer of topics, namely to generate a word in a document, a topic is first generated from a dirichlet distribution and then the word is generated by a topic-dependent multinomial distribution. Author topic model [17] takes the authorship information into account when generating the words of a document. Topic is generated from an author-dependent dirichlet distribution, where an author is randomly generated from the document's author list. Link LDA model [18] extends LDA model not only to generate the words of a document but also to generate the links from the document, which share the same topic distribution of the document. Other works for modeling the words and links include Link-PLSA-LDA model [19], Author-Recipient Topic model [20], Relational Topic model [21], Latent Topic model for Hypertext [22]. It is notable that the author-recipient topic model is designed for modeling the email which is written by one author to many recipients. Liu et al. proposed a Topic-Link LDA model [23] by leveraging words, links and authorships, where the generation of links is conditioned on the topic similarity of documents and community similarity of corresponding authors. However, it treats the link as a Bernoulli random variable and suffers from poor scalability and robustness as we discussed shortly.

A major issue suffered by LDA based models for community detection is that generative topic models are vulnerable to words that are irrelevant to communities. Recently, Yang et al. [1] made a successful progress by proposing a popularity conditional link (PCL) model and a discriminative content (DC) model for modeling the links and the attributes of nodes in a network, respectively. Both PCL model and DC model have been demonstrated to be superior than PHITS model and topic models, respectively. PCL model is motivated by using the popularities to fit the power-law degree distribution of real scale-free networks. When both the in-degree and out-degree distribution follow a power-law, Yang et al. [24] extended PCL model to a big family of productivity and popularity link (PPL) models, where the productivity is introduced to model the power-law out-degree distribution. In this work, we complete the family of Popularity Link model and Discriminative Content model by incorporating cross-domain link information, namely authorship information.

Our work also falls in a broad class of methods for detecting communities in heterogeneous networks. Heterogeneous networks contain various types of objects and relations. A big challenge is how to model the heterogeneous information in a unified framework. Sun et al. [25], [26] proposed unified ranking and clustering algorithms for bi-typed and multi-typed information networks. Long et al. [27] proposed a generative model for performing relational clustering of heterogeneous entities. Tang et al. [28] explored the community detection in multi-dimensional heterogeneous networks, where a set of same typed nodes are linked with each other by multiple types of relationships.

Finally, we briefly discuss different modeling choices for links, attributes and authorships to motivate our approach. There are two different ways for modeling links. They can be either modeled by a Bernoulli distribution [11], [12], [13] to generate both the presence of link and the absence of link,

or modeled by a multinomial distribution to compute a conditional link probability [14], [1]. The multinomial distribution is more attractive than the Bernoulli distribution because (i) the conditional link probability only models the presence of links and therefore it is more scalable to networks of large size, and (ii) it is less vulnerable to the missing links. It therefore motivates us to choose the multinomial distribution to model the links.

For modeling attributes, one can use generative models by making certain assumption on the distribution of the attributes. For example, topic models [15] use multinomial distribution to generate the words in a document, and [27] uses an exponential family distributions for generating the attributes. The problems with generative models include (i) the assumed distribution may not be suitable for the data; (ii) the community memberships are vulnerable to irreverent attributes. In contrast, discriminative model [1] overcomes these problems by fitting the attributes to the community memberships and adding discriminative power to alleviate the affect of irrelevant attributes. It has also been demonstrated to yield significant improvements over generative models in the community detection task. That is why we adhere ourself to the discriminative model for modeling attributes.

Finally in terms of authorships, two competing methods have been considered before. Some works [17] use the authorship as an explanatory factor in modeling the links, while others treat it as another type of link and employ link models [27]. These two methods have been studied separately, and it is not clear which one is preferable than the other one for the purpose of community detection. Therefore, we consider both modeling choices and make a comparison by empirical studies.

## III. Preliminaries

Let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{A})$ denote an authored networked data on entities represented by the node set $\mathcal{V} = \{\mathrm{v}_i, i \in [n]\}$ ($[n]$ denotes the set $\{1, \ldots, n\}$). The entities could be articles, products, blogs, web sites and etc. The set $\mathcal{E} = \{s_{ij} \in \mathbb{N}^+, (i,j) \in [n] \times [n]\}$ contains the link information between node pairs $\mathrm{v}_i$ and $\mathrm{v}_j$. Typically, $s_{ij} > 0$ encodes the weight of a directed link from node $\mathrm{v}_i$ to node $\mathrm{v}_j$ and $s_{ij} = 0$ indicates the absence of link between the corresponding node pairs. The set $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^d, i \in [n]\}$ contains the attributes for describing each node in $\mathcal{V}$. The set $\mathcal{A}$ denote the authorship information, i.e., $\mathcal{A} = \{\mathbf{u}_i \in \{0,1\}^m, i \in [n]\}$, where $u_{ij} = 1/0$ indicates whether node $\mathrm{v}_i$ is authored by "author" $\mathrm{a}_j$ among all $m$ authors. A canonical example for this setup is the citation network of articles. The attribute $\mathbf{x}_i$ represents the words occurring in the article, and the authorship $\mathbf{u}_i$ codes the authors who write the article. Without incurring any ambiguity, we also use $\mathcal{A}$ to denote the set of all authors.

Let $\mathcal{S}_{ou}(i) = \{\mathrm{v}_j | s_{ij} > 0, j \in [n]\}$ be the set of nodes that are linked from node $\mathrm{v}_i$, and $\mathcal{S}_{in}(i) = \{\mathrm{v}_j | s_{ji} > 0, j \in [n]\}$ be the set of nodes that are linking to node $\mathrm{v}_i$. Let $\mathcal{A}(i) = \{\mathrm{a}_j | u_{ij} = 1, j \in [m]\}$ denote the set of authors associated with the node $\mathrm{v}_i$, and $\mathcal{O}(j) = \{\mathrm{v}_i | u_{ij} = 1, i \in [n]\}$ denote the set of nodes that are authored by author $\mathrm{a}_j$.

Our goal is to identify the communities of nodes in $\mathcal{V}$ by utilizing all the available links, attributes and authorships information, and to identify the communities of authors if necessary. Our model is based on a well known approach for combining links and attributes to detect communities that uses a popularity conditional link model and a discriminative content model [1]. This approach models the links via popularities and community memberships of nodes and models the community memberships via the attributes. Our contributions are incorporating the authorship information into the model.

A big challenge is how to model the correlation between the authorship information and the other observed information (the links and the attributes) and the hidden variables (community memberships and popularities). To address the challenge, we propose two alternative approaches to model the authorships. In the first approach, we model the authorship via community memberships and popularities by analogy of popularity conditional link model. In the second approach, we consider the authorship as an explanatory factor for generating the links, accompanied by the memberships and the popularities. The two models differentiate from each other in treating the authorship to be a response variable or an explanatory variable. In next sections, we present the two models in details and discuss several extensions. Before moving to the details of the proposed models, we first present a baseline approach that employs generative models in a unified framework.

### A. A baseline: Author Topic LDA-Link Model

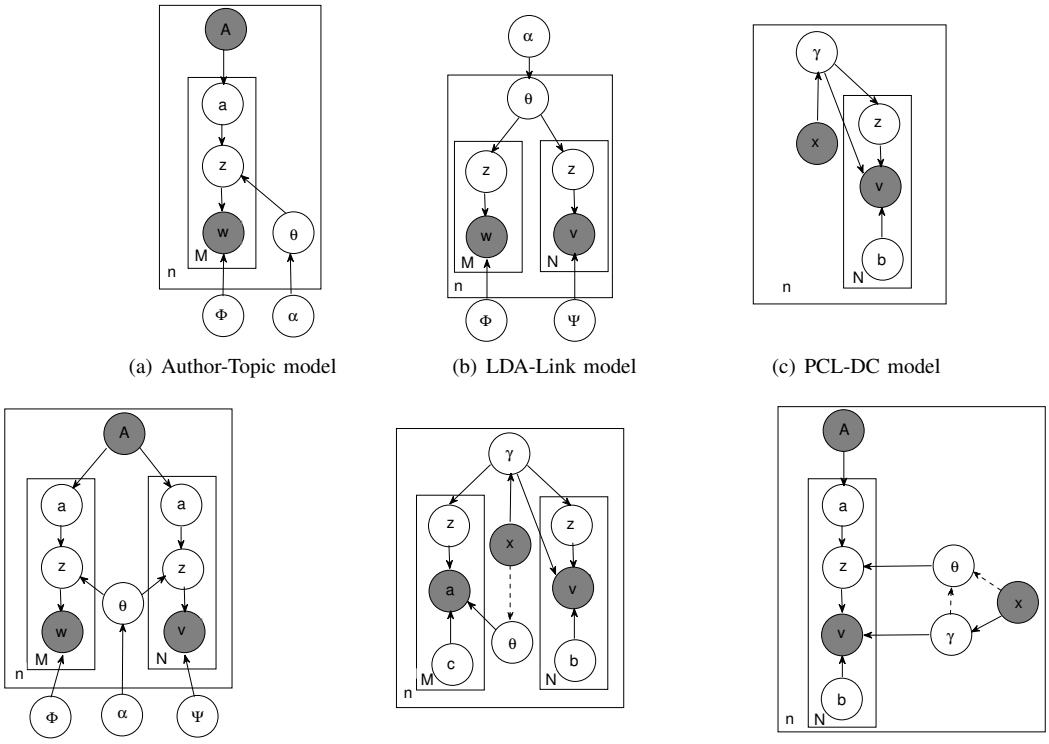
We present a straightforward extension of LDA-Link model that models the words and links via LDA model by adding the author-dependent topic distribution (i.e., the author-topic model) on the top. Figure 2 shows the graphical representation of the model. The generative process is similar to Author-Topic model on both words in documents and links from nodes. The parameters $\alpha, \Phi, \Psi$ and $\theta$ can be inferred by maximum posterior estimation using EM algorithm. Due to the limit of space, we omit the details. Finally, we can compute the community membership of each node $\mathrm{v}_i$ by $\gamma_{ik} = \sum_{\mathrm{a} \in \mathcal{A}(i)} \Pr(\mathrm{a}|\mathcal{A}(i))\theta_{\mathrm{a}k}$.

### B. Popularity Conditional Link & Author (PCLA) Model

The Popularity Conditional Link and Author (PCLA) Model consists of two parts: (1) the popularity conditional link (PCL) model and (2) the popularity conditional author (PCA) model. The popularity conditional link model is first proposed by Yang et. al [1]. It aims to model the conditional link probability $\Pr(\mathrm{v}_j|\mathrm{v}_i)$, i.e., given a node $\mathrm{v}_i$, how likely it will link to node $\mathrm{v}_j$ among all nodes. Let $\gamma_{ik}, i \in [n], k \in [K]$ denote the community memberships of node $\mathrm{v}_i$, i.e., how likely the node $\mathrm{v}_i$ belongs to community $C_k$, and $b_i$ denote the popularity of node $\mathrm{v}_i$. The conditional link probability in PCL is given by

$$\Pr(\mathrm{v}_j|\mathrm{v}_i) = \sum_{k=1}^{K} \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \gamma_{ik} \tag{1}$$

The underlying hypothesis is that if one node belongs to the same community of another node and has a higher popularity, it has a higher probability to be linked by the other node. In popularity conditional author (PCA) model, we make the same assumption for modeling the author of a node. In particular, we let $\theta_{jk}$ denote the community memberships of author $j$ and $c_j$




(a) Author-Topic model (b) LDA-Link model (c) PCL-DC model

(d) Author-Topic LDA-Link model (e) PCLA-DC model (f) ATPCL-DC model

Fig. 2. The graphical representation of (a) Author-Topic Model; (b) LDA-Link model; (c) PCL-DC model; (d) Author-Topic LDA-Link model; (e) PCLA-DC model and (f) ATPCL-DC model. The dashed line represents flexible modeling choices. Dark circles denote observed information: **A** denotes authorships, **w** denotes words, **v** denotes the linked nodes, **x** denotes attributes (of nodes and authors, which include bag-of-words representation **w** as a special case), $\gamma$ denotes community memberships of nodes and $\theta$ denotes community memberships of authors.

denote the popularity of author $\mathrm{a}_j$. The conditional probability that node $\mathrm{v}_i$ is authored by author $\mathrm{a}_j$ is given by

$$\Pr(\mathrm{a}_j|\mathrm{v}_i) = \sum_{k=1}^{K} \frac{\theta_{jk} c_j}{\sum_{j'} \theta_{j'k} c_{j'}} \gamma_{ik} \tag{2}$$

We combine the two separate models together to model the authored network by summing up the log-likelihood, i.e.,

$$\begin{aligned}\mathcal{L}(\gamma, \theta, b, c) = (1-\alpha) &\sum_{(\mathrm{v}_i, \mathrm{v}_j) \in \mathcal{E}} s_{ij} \log \Pr(\mathrm{v}_j|\mathrm{v}_i) \\ &+ \alpha \sum_{(\mathrm{v}_i, \mathrm{a}_j) \in \mathcal{A}} u_{ij} \log \Pr(\mathrm{a}_j|\mathrm{v}_i)\end{aligned} \tag{3}$$

The parameter $\alpha$ is added to the combined log-likelihood for more flexibility. When $\alpha = 1$, the model reduces to the PCL model and when $\alpha = 0$ it reduces to the PCA model. To infer the model parameters $\gamma$, $\theta$ and $b, c$, we can take the EM algorithm to maximize the log-likelihood, as discussed in Section V.

### C. Author-Topic Popularity Conditional Link (ATPCL) Model

In the popularity conditional author model, we treat the author to be a response variable. An alternative option is to consider the author as an explanatory variable. This modeling strategy has been adopted in the author-topic model. To motivate the author-topic popularity conditional link model, we first briefly describe the author topic model. In order to generate a word in an authored document, one author is sampled uniformly out of all authors of the document, then a topic is generated according to author-topic distribution , and finally a word is sampled by topic-word distribution. In mathematical form, the probability of word $\mathbf{w}_j$ in document $i$ with authors $\mathcal{A}(i)$ is given by

$$\Pr(\mathbf{w}_j|\mathcal{A}(i)) = \overbrace{\sum_k \beta_{jk}}^{\text{finally sample a word}} \underbrace{\sum_{\mathrm{a}_{j'} \in \mathcal{A}(i)} \theta_{j'k} \widehat{u}_{ij'}}_{\text{first sample an author by } \widehat{u}_{ij'}\text{, then sample a topic by } \theta_{j'k}}$$

where $\widehat{u}_{ij}$ is normalized such that $\sum_{\mathrm{a}_j \in \mathcal{A}(i)} \widehat{u}_{ij} = 1$. The above author-topic model has been extended to author-topic link model with document $i$ replaced with node $\mathrm{v}_i$ and word $\mathbf{w}_j$ replaced with node $\mathrm{v}_j$ in section III-A. In this section, we combine popularity conditional link (PCL) model and author-topic (AT) model to form an author-topic popularity conditional link model (ATPCL), which is given by

$$\Pr(\mathrm{v}_j|\mathrm{v}_i, \mathcal{A}(i)) = \sum_k \underbrace{\frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}}}_{\text{PCL}} \overbrace{\sum_{\mathrm{a}_{j'} \in \mathcal{A}(i)} \widehat{u}_{ij'} \theta_{j'k}}^{\text{AT}} \tag{4}$$

Compared with the author topic model, the free parameter $\beta_{jk}$ is molded by an explicit form $\frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}}$ using the popularity, which is important to model the real networks which usually reveal the scale-free property. In contrast to the PCL model,

the community membership $\gamma_{ik}$ of node $i$ is replaced by the author topic model $\sum_{j'} \theta_{j'k}\widehat{u}_{ij'}$ associated with node $i$. In the next section, we extend ATPCL from two aspects.

## IV. Extentions

### A. ATPCL$^{cn}$: computing author community memberships from nodes

If the problem is to detect the communities of nodes, we can reduce the number of free parameters by assuming that the community memberships of authors depend on the community memberships of their owned nodes. Simply, we let $\theta_{jk} = \sum_i \widetilde{u}_{ij}\gamma_{ik}$, where $\widetilde{u}_{ij}$ is the normalized such that $\sum_{\mathrm{v}_i \in \mathcal{O}(j)} \widetilde{u}_{ij} = 1$. Replacing such $\theta$ in equation (4), we have

$$\Pr(\mathrm{v}_j|\mathrm{v}_i, \mathcal{A}(i)) = \sum_k \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \sum_{j'} \sum_{i'} \widehat{u}_{ij'} \widetilde{u}_{i'j'} \gamma_{i'k} \quad (5)$$

Define $T = \widehat{A}\widetilde{A}^T$, which is the co-authorship matrix between nodes. We then have

$$\Pr(\mathrm{v}_j|\mathrm{v}_i, \mathcal{A}(i)) = \sum_k \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \sum_{i'} T_{ii'} \gamma_{i'k} \quad (6)$$

The log-likelihood is given by

$$\mathcal{L} = \sum_{(\mathrm{v}_i \to \mathrm{v}_j) \in \mathcal{E}} s_{ij} \log \sum_k \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \sum_{i'} T_{ii'} \gamma_{i'k} \quad (7)$$

### B. ATPCL-DC: ATPCL + Discriminative Content Model

When the attribute information of nodes is available, we can combine the discriminative content model with ATPCL or its variants as did in [1]. In particular, the community memberships $\gamma_{ik}$ of node $\mathrm{v}_i$ can be molded by

$$\gamma_{ik} = \frac{\exp(\mathbf{w}_k^\top \mathbf{x}_i)}{\sum_l \exp(\mathbf{w}_l^\top \mathbf{x}_i)} \quad (8)$$

The combined model is to simply replace $\gamma_{ik}$ in ATPCL with the above explicit form of community memberships.

Finally, we briefly mention the limitations and possibilities when applying these extensions to PCLA model. First, we can also compute the community memberships of authors from the memberships of nodes in PCLA model, however, it would make the maximum likelihood estimation much more involved. Second, the discriminative content model can be added to the PCLA model as straightforward as ATPCL-DC, to which we refer as PCLA-DC model.

## V. Maximum Likelihood Estimation

In this section, we present the maximum likelihood estimation for the models proposed in previous sections. The maximum likelihood estimation consists of two alternate steps: E-steps and M-steps. From the perspective of optimization, the E-step is essentially to lower bound the log-likelihood and the M-steps is to maximize the lower bound over the parameters. We first present the EM steps without considering the discriminative content model, and then discuss how to modify the basic algorithms when combined with discriminative content model. The computations of EM-steps for the proposed models are shown in Figure 3. The detailed derivations can be found in the appendix.

We make several remarks about the EM steps for each model. For PCLA model, it is easy to see when setting $\alpha = 0$, if we ignore $p_{ijk}, \eta_k$, $\theta_{jk}$ and $c_j$ associated with authorship, the EM-steps reduce to that of PCL model as presented in [1]. Similarly, when setting $\alpha = 1$, if we ignore $q_{ijk}, \tau_k, b_i$, we obtain the EM steps for PCA model. It is also interesting to note that the popularity of node $\mathrm{v}_i$ is proportional to $n_{in}(i) = \sum_j s_{ji}$, i.e., the indegree of node $\mathrm{v}_i$, and the popularity of author $\mathrm{a}_j$ is proportional to $o(j) = \sum_i u_{ij}$, i.e., the indegree of author $\mathrm{a}_j$ (how many papers he is involved). These results seem intuitive. For example, if a node has a higher popularity, it should receive more incoming links, and similarly an author who has a high popularity would publish more papers.

In contrast to PCL model in which the community memberships are affected by both the incoming links via $n_{in}(i,k)$ and the outgoing links via $n_{out}(i,k)$, the community memberships in ATPCL are only affected by the incoming links of associated nodes. However, ATPCL$^{cn}$ model by computing the community memberships of authors from their associated nodes has similar mechanism in inferring the community memberships of nodes, except that the impact from the outgoing links is propagated through all co-authored nodes (i.e., $n^c_{out}(i,k)$). Since ATPCL$^{cn}$ used more information to infer the community memberships than ATPCL model, it is expected to produce better results. This is further verified by our empirical studies as shown in section VI.

Finally, we briefly discuss how to compute $\gamma$ and $b$ or $\theta$ and $c$ in each M-step, which are connected by two equations in the general form of

$$\gamma_{ik} = \frac{A_{ik} + B_{ik}}{f_k b_i + \sum_k B_{ik}}, \quad b_i = \frac{\sum_k A_{ik}}{\sum_k f_k \gamma_{ik}}$$

where $A_{ik}, B_{ik}$ and $f_k$ are known. Noting that $\sum_k \gamma_{ik} = 1$, we can first compute $b_i$ by solving the following equation

$$\sum_k \frac{A_{ik} + B_{ik}}{f_k b_i + \sum_k B_{ik}} = 1$$

The root of above equation can be computed by Newton method [29]. Then the values of $\gamma_{ik}$ are computed by substituting the value of $b_i$ in the equation.

### A. Two-stage algorithms for PCLA-DC and ATPCL-DC

When the content information is available and is modeled by the discriminative model, the log-likelihood become more complex, which makes the EM steps more involved. Yang et al. [1] proposed a nice two-stage algorithm that separates inference of the link model and that of the content model, which can be applied to inferring models proposed in this work. In the two-stage algorithm, $\gamma_{ik}$ is first computed in M-steps from link and/or authorship information, and then projected back to the domain of explicit form by solving

$$\max_{\mathbf{w}} \sum_{ik} \gamma_{ik} \log \frac{\exp(\mathbf{w}_k^\top \mathbf{x}_i)}{\sum_l \exp(\mathbf{w}_l^\top \mathbf{x}_i)} - \frac{\lambda}{2} \sum_k \|\mathbf{w}_k\|_2^2 \quad (9)$$

where the last term is added intentionally to avoid overfitting, which usually yields good empirical performances. The E-step is then proceeded by using the projected solution $y_{ik} =$

PCL E-steps :

$$q_{ijk} \propto \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \gamma_{ik},\ s.t.\ \sum_k q_{ijk} = 1$$

$$\tau_k = \sum_{j'} \gamma_{j'k} b_{j'}$$

PCL M-steps :

$$\gamma_{ik} = \frac{n_{in}(i,k) + n_{out}(i,k)}{g_k b_i + n_{out}(i)}$$

$$b_i = \frac{n_{in}(i)}{\sum_k g_k \gamma_{ik}}$$

PCLA E-steps :

$$q_{ijk} \propto \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \gamma_{ik},\ s.t.\ \sum_k q_{ijk} = 1$$

$$p_{ijk} \propto \frac{\theta_{jk} c_j}{\sum_{j'} \theta_{j'k} c_{j'}} \gamma_{ik},\ s.t.\ \sum_k p_{ijk} = 1$$

$$\tau_k = \sum_{j'} \gamma_{j'k} b_{j'}, \quad \eta_k = \sum_{j'} \theta_{j'k} c_{j'}$$

PCLA M-steps :

$$\gamma_{ik} = \frac{(1-\alpha)[n_{in}(i,k) + n_{out}(i,k)] + \alpha a(i,k)}{(1-\alpha) g_k b_i + (1-\alpha) n_{out}(i) + \alpha a(i)}$$

$$b_i = \frac{n_{in}(i)}{\sum_k g_k \gamma_{ik}}, \quad \theta_{jk} = \frac{o(j,k)}{h_k c_j}, \quad c_j = \frac{o(j)}{\sum_k h_k \theta_{ik}}$$

ATPCL E-steps :

$$q_{ijk} \propto \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \sum_{j'} \widehat{u}_{ij'} \theta_{j'k},\ s.t.\ \sum_k q_{ijk} = 1$$

$$e_{ijk} \propto \widehat{u}_{ij} \theta_{jk}\ s.t.\ \sum_j e_{ijk} = 1$$

$$\tau_k = \sum_{j'} \gamma_{j'k} b_{j'}$$

ATPCL M-steps :

$$\gamma_{ik} = \frac{n_{in}(i,k)}{g_k b_i}, \quad b_i = \frac{n_{in}(i)}{\sum_k g_k \gamma_{ik}}, \quad \theta_{jk} \propto \sum_{ij'} s_{ij'} q_{ij'k} e_{ijk}$$

ATPCL$^{cn}$ E-steps :

$$q_{ijk} \propto \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \sum_{i'} T_{ii'} \gamma_{i'k},\ s.t.\ \sum_k q_{ijk} = 1$$

$$\zeta_{ii'k} = \frac{T_{ii'} \gamma_{i'k}}{\sum_j T_{ij} \gamma_{jk}}, \tau_k = \sum_{j'} \gamma_{j'k} b_{j'}$$

ATPCL$^{cn}$ M-steps :

$$\gamma_{ik} = \frac{n_{in}(i,k) + n^c_{out}(i,k)}{g_k b_i + \sum_k n^c_{out}(i,k)}, \quad b_i = \frac{n_{in}(i)}{\sum_k g_k \gamma_{ik}}$$

$$n_{in}(i,k) = \sum_j s_{ji} q_{jik},\ n_{out}(i,k) = \sum_j s_{ij} q_{ijk},\ a(i,k) = \sum_j u_{ij} p_{ijk},\ o(j,k) = \sum_i u_{ij} p_{ijk},\ \ n^c_{out}(i,k) = \sum_{i'j'} \zeta_{i'ik} s_{i'j'} q_{i'j'k}$$

$$n_{in}(i) = \sum_j s_{ji},\ n_{out}(i) = \sum_j s_{ij},\ a(i) = \sum_j \widehat{u}_{ij},\ o(j) = \sum_i \widehat{u}_{ij},\ g_k = \frac{\sum_{ij} s_{ij} q_{ijk}}{\tau_k},\ h_k = \frac{\sum_{ij} u_{ij} p_{ijk}}{\eta_k}$$

Fig. 3. EM steps for PCL (top left), PCLA (top right), ATPCL (middle left) and ATPCL-cn (middle right). The bottom box are some notations.

$\frac{\exp(\mathbf{w}_k^\top \mathbf{x}_i)}{\sum_l \exp(\mathbf{w}_l^\top \mathbf{x}_i)}$ in place of $\gamma_{ik}$. The problem in (9) can be viewed as a soft version of logistic regression where the binary class labels are replaced with the soft community memberships $\gamma_{ik}$. To solve the problem, many efficient algorithms can be employed (e.g., [30], [31], [32]). In the experiments, we use a LBFGS implementation which is available at http://www.cs.toronto.edu/~liam/software.shtml.

## VI. Experimental Evaluations

In this section, we present several experimental results. We first describe the data sets.

*a) Two Paper Citation Data sets:* We choose two paper citation data sets, Cora and Citeseer. The two data sets are preprocessed by the research group of Lise Getoor [33] and have been used by many works for detecting communities in networks [1], [24]. Each paper is described by a binary vector indicating presence and absence of a list of words, and each belong to one of several classes that denote the sub-category of the paper (e.g. artificial intelligence, machine learning, etc.). The two data sets only contain the citation information and the content information. We crawl the author information from the original Cora corpus [34] and the Citeseer bibtex library [1] by using the paper id from the data sets. By removing the records that lack of author information, the final statistics of the two data sets are reported in Table I. In the last two columns, we also report the average number of links per node and the average number of authors per node.

*b) A Wikipedia Data set:* We use the Wikipedia dump from Oct-09-2009 for our experiments. We collected roughly 20K articles from four categories–Biology, Natural Science, Computer Science and Social Science, with 5K articles in each category. Each of the four topics are further divided into three subtopics. After removing stubs and other smaller articles we were left with 10K articles and 53K editors (who have edited the articles). We removed articles/editors that do not have sufficient links (less than 3 links) with other articles/editors in our corpus. Our final data set contains 6403 articles and 5361 editors. After stemming and removing stopwords and words that appear in few documents, each article is described by a binary vector of dimension 4512. Our goal is to identify the 12 sub-categories of articles.

[1] http://citeseer.ist.psu.edu

TABLE I. STATISTICS OF DATA SETS

| name | #papers | #words | #authors | #classes | #links/node | #authors/node |
|---|---|---|---|---|---|---|
| Cora | 1915 | 1433 | 1625 | 7 | 3.64 | 1.94 |
| Citeseer | 1669 | 3703 | 3090 | 6 | 1.61 | 2.73 |
| Wikipedia | 6320 | 4512 | 5361 | 12 | 21.71 | 5.04 |

TABLE II. PERFORMANCE OF COMMUNITY DETECTION ON THE THREE DATASETS. IN THE FIRST COLUMN, L REPRESENTS LINK, A REPRESENTS AUTHORSHIP, AND C REPRESENTS CONTENT. THE VALUE OF $\alpha$ IN PCLA IS SET TO 0.5. THE VALUES OF $\lambda$ IN LINK-DC MODELS ARE SELECTED FROM $\{0, 1, 2, \ldots, 8\}$ FOR OBTAINING THE BEST PERFORMANCES. THE BEST VALUES OF $\lambda$ IN PCL-DC MODEL ON THE THREE DATA SETS ARE $1, 6, 4$, IN PCLA-DC MODEL ARE $1, 3, 4$, IN ATPCL MODEL ARE $4, 8, 2$ AND IN ATPCL$^{\text{CN}}$ MODEL ARE $1, 1, 1$.

| | | Cora | | | Citeseer | | | Wikipedia | | |
|---|---|---|---|---|---|---|---|---|---|---|
| info. | algo. | ACC | NMI | PWF | ACC | NMI | PWF | ACC | NMI | PWF |
| L | PCL | 0.2679 | 0.0668 | 0.1987 | 0.2331 | 0.0226 | 0.2098 | 0.3666 | 0.3523 | 0.2854 |
| A | PCA | 0.2037 | 0.0225 | 0.1654 | 0.2229 | 0.0118 | 0.1773 | 0.3864 | 0.3520 | 0.3077 |
| L + A | PCLA | 0.3143 | 0.1023 | 0.2113 | 0.2037 | 0.0137 | 0.1785 | 0.4400 | 0.4478 | 0.3551 |
| | ATPCL | 0.2256 | 0.0401 | 0.1759 | 0.2013 | 0.0085 | 0.1760 | 0.4796 | 0.5249 | 0.3937 |
| | ATPCL$^{\text{cn}}$ | **0.4470** | **0.2139** | **0.2990** | **0.2391** | **0.0315** | **0.2029** | **0.5068** | **0.5262** | **0.4375** |
| L + C + A | Author-Topic LDA-Link | 0.3029 | 0.0254 | 0.3010 | 0.2265 | 0.0075 | 0.2806 | | NA | |
| | PCLA-DC | 0.4198 | 0.2407 | 0.2977 | 0.3595 | 0.1015 | 0.2499 | 0.5434 | 0.5854 | 0.4542 |
| | ATPCL-DC | 0.6162 | 0.3979 | 0.4649 | 0.6010 | 0.3339 | 0.4552 | 0.5465 | 0.5629 | 0.4365 |
| | ATPCL$^{\text{cn}}$-DC | 0.6047 | 0.4251 | 0.4487 | 0.6327 | 0.3787 | 0.4942 | 0.5408 | 0.5667 | 0.4635 |

### A. Experimental Results

We report the experimental results in Table II. The performances are measured with three metrics, namely accuracy (ACC), normalized mutual information (NMI) and pairwise F-measure (PWF). The metrics are defined with respect to the ground truth labels. For more details, we refer to [1]. We first take an overview of the results. The performances vary on the three data sets. The performances of PCL model are influenced by how accurate the links are in telling the community relationship between nodes and how many links are available. Citeseer data set has the least number of links and the worst performance. In contrast, Wikipedia data set has the most number of links per node and the best performance. Similar phenomenon can be observed for PCA model. In addition, from the results we are able to answer the following questions.

- Does authorship help to improve the performance of community detection? The answer is yes. Comparing PCL and ATPCL$^{\text{cn}}$, we can conclude that considering the authorships in the link models can yield improvements on community detection.
- Authorship: Response vs Explanatory? Comparing ATPCL$^{\text{cn}}$ and PCLA, we find that ATPCL$^{\text{cn}}$ is much better than PCLA, which means the modeling the authorships as explanatory factors is more effective that modeling them as response variables.
- Does it help by computing the community memberships of authors from nodes? The answer is positive. Comparing ATPCL and ATPCL$^{\text{cn}}$, we can see that ATPCL$^{\text{cn}}$ gives better results. On Cora and Citeseer data sets, ATPCL is even worse than PCL, which is because the two data sets have sparse information and ATPCL has too many parameters. By computing the community memberships of authors from nodes, APTCL$^{\text{cn}}$ can have more accurate inference of the community memberships.
- How about incorporating links, contents, and attributes together? By incorporating the content information, the performances of the ATPCL or ATPCL$^{\text{cn}}$ are improved significantly by fitting the content information to the community memberships. We also include the performances [2] of a generative model described in Section III-A, namely Author-Topic LDA-Link model. In comparison to the Author-Topic LDA-Link model, the proposed models performs much better. In some cases, even without considering the content information, the ATPCL$^{\text{cn}}$ model gives us better results than that of Author-Topic LDA-Link model.

Finally, we briefly mention that the parameters in the various models (e.g. $\alpha$ in PCLA model and $\lambda$ in DC models) can be tuned based on performances on link prediction [1] without acquiring the ground true labels, because the performance of community detection has a positive relationship with the performance of link prediction.

## VII. CONCLUSIONS

In this paper, we have presented models for detecting communities in networks by incorporating the link, content and authorship information. We build our models on the top of popularity conditional link model and discriminative content model. We considered two modeling choices for modeling the relationship between community memberships and authorships and observed by empirical studies that treating the authorship as an explanatory variable yield better performances for community detection than that by treating it as a response variable. Our empirical studies also demonstrate that authorship information can improve the performances of popularity link models.

[2] We only report the results on Cora and Citeseer data set, because the model does not scale well to the relatively large Wikipedia data set.

## VIII. Appendix

### A. Maximum Likelihood Estimation for ATPCL$^{cn}$

In this section, we present the maximum likelihood estimation for ATPCL$^{cn}$. Following the same philosophy, one can easily derive the EM steps for PCLA and ATPCL model. Using the notations in Figure 3, in the E-step we derive the lower bound of the log-likelihood,

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{ij} s_{ij} \log \sum_k \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \sum_{i'} T_{ii'} \gamma_{i'k} \\
&\geq \sum_{ijk} s_{ij} q_{ijk}^t \Big( \log(\gamma_{jk} b_j) - \log \sum_{j'} \gamma_{j'k} b_{j'} + \log \sum_{i'} T_{ii'} \gamma_{i'k} \Big) \\
&- \sum_{ijk} s_{ijk} q_{ijk}^t \log q_{ijk}^t \\
&\geq \sum_{ij} s_{ij} \sum_k q_{ijk}^t \Big( \log(\gamma_{jk} b_j) + 1 - \sum_{j'} \frac{\gamma_{j'k} b_{j'}}{\tau_k^t} - \log \tau_k^t \\
&+ \sum_{i'} \zeta_{ii'k}^t \log \frac{T_{ii'} \gamma_{i'k}}{\zeta_{ii'k}} \Big) - \sum_{ijk} s_{ijk} q_{ijk}^t \log q_{ijk}^t
\end{aligned}$$

In the M-step, we maximize the lower bound over the parameters, i.e.,

$$\max_\theta \sum_{ijk} s_{ij} q_{ijk}^t \left( \log \gamma_{jk} b_j - \sum_{j'} \frac{\gamma_{j'k} b_{j'}}{\tau_k} + \sum_{i'} \zeta_{ii'k} \log \gamma_{i'k} \right)$$

Using the notations in Figure 3, it is not difficult to check that the optimal solution to $\gamma, b$ satisfy the equations given in Figure 3.

## References


[1] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *KDD*, 2009, pp. 927–936.

[2] L. Ford and D. Fulkerson, "Maximal flow through a network," vol. 8, pp. 399–404, 1956.

[3] B. Bollobs, *Modern Graph Theory. Graduate Text in Mathematics.* Springer-Verlag, 1998, vol. 184.

[4] Y.-C. Wei and C.-K. Cheng, "Towards efficient hierarchical designs by ratio cut partitioning," in *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*, 1989, pp. 298 –301.

[5] B. Hendrickson and R. W. Leland, "A multi-level algorithm for partitioning graphs," in *Proceedings of Supercomputing '95*, 1995.

[6] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 359–392, 1998.

[7] Z. Zhang, T. Li, C. Ding, and X. Zhang, "Binary matrix factorization with applications," in *Proceedings of the IEEE Int'l Conf on Data Mining*, 2007, pp. 391–400.

[8] M. Newman, "Modularity and community structure in networks," *Proceedings of National Academy of Science*, vol. 103, pp. 8577–8582, 2006.

[9] P. K. Chan, M. D. F. Schlag, and J. Y. Zien, "Spectral k-way ratio-cut partitioning and clustering," in *Proceedings of the 30th Int'l Design Automation Conf.* ACM, 1993, pp. 749–754.

[10] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proceedings of CVPR*, 1997.

[11] P. W. Holland and S. Leinhardt, "The statistical analysis of local structure in social networks," Tech. Rep., 1974.

[12] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic block models for relational data with application to protein-protein interactions," in *In Proceedings of the International Biometrics Society Annual Meeting*, 2006.

[13] J. M. Hofman and C. H. Wiggins, "A Bayesian approach to network modularity," *Physiccal Review L*, vol. 100, 2008.

[14] D. Cohn and H. Chang, "Learning to probabilistically identify authoritative documents," in *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[15] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.

[16] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of 15th Uncertainty in Artificial Intelligence*, 1999.

[17] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, 2004.

[18] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed membership models of scientific publications," in *Proceedings of the National Academy of Sciences*, 2004.

[19] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[20] A. Mccallum, A. Corrada-Emmanuel, and X. Wang, "The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email," Tech. Rep., 2004.

[21] J. Chang and D. M. Blei, "Relational topic models for document networks," *Journal of Machine Learning Research - Proceedings Track*, vol. 5, pp. 81–88, 2009.

[22] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Latent topic models for hypertext," in *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence*, 2008.

[23] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link lda: joint models of topic and author community," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 665–672.

[24] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, "Directed network community detection: A popularity and productivity link model," in *SDM*, 2010, pp. 742–753.

[25] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 797–806.

[26] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 2009, pp. 565–576.

[27] B. Long, P. S. Yu, and Z. Zhang, "A General Model for Multiple View Unsupervised Learning," in *Proceedings of the 2008 SIAM International Conference on Data Mining*, 2008.

[28] L. Tang, X. Wang, and H. Liu, "Community detection via heterogeneous interaction analysis," *Data Min. Knowl. Discov.*, pp. 1–33, 2012.

[29] T. J. Ypma, "Historical development of the newton-raphson method," *SIAM Rev.*, vol. 37, pp. 531–551, 1995.

[30] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Math. Program.*, pp. 503–528, 1989.

[31] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*, 2005.

[32] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets," INRIA, Tech. Rep. arXiv:1202.6258v1, 2012.

[33] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Magazine*, pp. 93–106, 2008.

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