Title: Combining Link and Content for Community Detection

Name: Tianbao Yang[1], Rong Jin[2], Yun Chi[3], Shenghuo Zhu[3]

Affil./Addr. 1: Machine Learning Lab, GE Global Research

San Ramon, CA 94583

E-mail: tyang@ge.com

Affil./Addr. 2: Department of Computer Science and Engineering

Michigan State University, East Lansing, MI 48824

E-mail: rongjin@cse.msu.edu

Affil./Addr. 3: NEC Laboratories America, Inc.

Cupertino, CA 95014

E-mail: {ychi, zsh}@sv.nec-labs.com

Combining Link and Content for Community Detection

Synonyms

Clustering, Graph Partitioning, Information Fusion

Glossary

Network: a set of nodes that are connected by relationships.

Community: a subset of nodes in the network that are densly connected and have similar attributes.

Community Detection: finding the communities in a network.

Link Analysis: using the link information to detect the communities.

Content Analysis: using the attribute information to detect the communities.

Generative Model: a model for randomly generating observable data given some hidden parameters.

EM Algorithm: an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical model.

Definition

In the contexture of networks, community structure refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network. When it comes to networked data (namely a network of nodes with each described by a number of attributes), the task of *community detection* is to find the cohesive groups of nodes which are densely connected within the group and sparsely connected with others, and share similar attributes as well. The attributes are usually referred to as "contents" in the context. The goal is to improve the performance of community detection by combining both the link and the content information of nodes.

Introduction

As online repositories such as digital libraries and user-generated media (e.g., blogs) become more popular, analyzing such networked data has become an increasingly important research issue. One major topic in analyzing such networked data is to detect salient communities among individuals. Community detection has many applications such as understanding the social structure of organizations and modeling large-scale networks in Internet services [Wang et al 2005].

A networked data set is usually represented as a graph where individuals in the network are represented by the nodes in the graph. The nodes are tied with each other by either directed links or undirected links, which represent the relations among the

individuals. In addition to the links that they are incident to, nodes are often described by certain attributes, to which we refer as *contents* of the nodes. For example, when it comes to the web pages, online blogs, or scientific papers, the contents are usually represented by histograms of keywords; in the network of co-authorship, the contents of nodes can be the demographic or affiliation information of researchers.

Besides community structure, real networks usually reveal many other interesting properties, among which two important properties are scale-free, small-word. Scale-free refers to the link structure in which a few nodes have a tremendous number of connections to the other nodes while most nodes in the network only have a handful of connections. A small-world network is a type of network in which although most nodes are not directed connected with each other, most nodes can be reached from each other by a small number of hops or steps. A small-world network is usually a scale-free network. It is very important to consider these properties of real networks when modeling the links for community detection [Yang et al 2010].

Many existing studies on community detection focus on either link analysis or content analysis. However, neither information alone is satisfactory in determining accurately the community memberships: the link information is usually sparse and noisy and often results in a poor partition of networks; the irrelevant content attributes could significantly mislead the process of community detection. It is therefore important to combine the link analysis and content analysis for community detection in networks. Recently, several approaches have been proposed to combine link and content information for community detection. Most of these approaches adopted a *generative* framework where a generative link model and a generative content model are combined through a set of shared hidden variables of community memberships. We argue that such a generative framework suffers from two shortcomings. First, community membership by itself is insufficient to model links—link patterns are usually affected by factors other

than communities such as the popularity of a node (i.e., how likely the node is cited by other nodes). Second, the content information often include irrelevant attributes and as a result, a generative model without feature selection usually leads to poor performance.

Community Detection

Link Analysis

Link based approaches for community detection only utilize the link information. They can be classified into two categories, namely measure-based algorithms and model-based algorithms. In measure-based algorithms, a measure is first defined to quantify the quality of communities, and then communities are identified by optimizing the measure. Such measures include graph cuts [Kolmogorov and Zabih 2004], modularity [Newman and Girvan 2003], centrality [Wasserman and Faust 1994], and density [Baumes et al 2005a]. Model based algorithms for community detection often define a generative process for the links observed in a network. Hidden variables are introduced for each node to represent its community memberships. By making certain statistical assumptions regarding the hidden variables of community memberships and the generative process for the observed links, we can write down the likelihood function for the observed links, and the optimal community assignment is decided by maximizing the likelihood of the observed links. There are various models that have been proposed for community detection, as briefly reviewed below.

Stochastic Block Model

Stochastic block model is a popular class of probabilistic models for relational data analysis pioneered by Holland and Leinhardt [1974], and later on was extended in various contexts [Airoldi et al 2006, Hofman and Wiggins 2008, Kemp et al 2004].

In stochastic block models, a community variable $z_i \in \{1, \cdots, K\}$ is introduced for each node i, which is a random variable indicating which community the node i belongs to and is drawn from a multinomial distribution $Mult(\gamma_{i1}, \cdots, \gamma_{iK})$, where γ_{ik} stands for the probability of assigning node i to community k. The probability of creating a link between two nodes i and j is assumed to depend only on the community memberships of the nodes, and to be independent from the entities of the two nodes. In the simplest scenario, assuming each link is a binary variable, i.e., $w_{ij} \in \{0, 1\}$, given the community variables of z_i, z_j, the probability of creating a link from node i to node j is given by

$$\Pr((i, j)|z_i, z_j) = \eta_{z_i,z_j}^{w_{ij}}(1 - \eta_{z_i,z_j})^{1-w_{ij}} \tag{1}$$

where η_{z_i,z_j} specifies the probability of creating a link from a node in community z_i to a node in community z_j. For simplicity, we introduce the matrix $\eta = [\eta_{k,l}]_{K\times K}$ to include all the probabilities of creating links between communities. The probability matrix η could be symmetric or asymmetric, dependent on whether the network is undirected or directed. Note that the probability defined in (1) models not only the presence of a link but also the absence of a link, which cause the stochastic block model to suffer from a high computational cost. The parameters η and the community membership γ of nodes are obtained by maximum likelihood. Differeent variants of stochastic block model differ in the process of generating community variables and the algorithm used for inference.

PHITS model and LDA-Link model

PHITS [Cohn and Chang 2000] is a probabilistic model that extends the Probabilistic Latent Semantic Analysis (PLSA) [Hofmann 1999] to network analysis. Similarly, LDA-Link [Erosheva et al 2004] extends Latent Dirichlet Allocation (LDA) [Blei et al 2003] to network analysis. Both PHITS and LDA-Link are generative models that are designed to model the directed links. These two models address the problem of high

computational cost suffered by stochastic block models. In PHITS and LDA-link, in order to generate a link (i, j) from node i to node j, they first sample a community variable z_{ij} for node i by following a node-dependent multinomial distribution, i.e, $z_{ij} \sim Mult(\gamma_{i1}, \cdots, \gamma_{iK})$. The conditional probability of creating a link from node i to j given z_{ij} is given by

$$\Pr(j|i, z_{ij}, \beta) = \beta_{jz_{ij}} \tag{2}$$

where β_{jk} is the probability for node j to be linked by any node in community k. By integrating out z_{ij},, we have

$$\Pr(j|i, \gamma, \beta) = \sum_k \gamma_{ik}\beta_{jk} \tag{3}$$

Then the parameters γ and β are obtained by maximizing the log-likelihood or computing the posterior distribution. PHITS and LDA-Link differ in the procedures for inference. In PHITS, the optimal values for β and γ are obtained by maximizing the log-likelihood of $\Pr(\mathcal{E}|\gamma, \beta)$. In LDA-Link, instead of computing the most likely values for γ, it infers the posterior distributions for γ by assuming that $\gamma_i, i = 1, \cdots, n$ are sampled from a dirichlet distribution of $Dir(\alpha_1, \cdots, \alpha_K)$. β can be viewed as parameters and obtained by maximum likelihood or treated as random variables sampled from a Dirichlet distribution and the posterior of β is computed.

Graph factorization model

Graph factorization models are probabilistic models that are only designed for analyzing undirected graphs. Similar to PHITS and LDA-Link, in GFM, additional variables of communities are introduced to capture the relationships among nodes. Let C_k denote the community k. The key quantity modeled by GFM is the link probability between node i and node j, denoted by $\Pr(i, j)$. It is the modeling of this joint probability that allows us to decide appropriate community assignments of individual nodes in a network. In [Yu et al 2005], K. Yu et al. factorized the joint probability $\Pr(i, j)$ as

$$\Pr(i,j) = \Pr(i)\Pr(j|i) = \Pr(i)\sum_k \Pr(j|C_k)\Pr(C_k|i)$$
$$= \sum_k \frac{\Pr(i,C_k)\Pr(j,C_k)}{\Pr(C_k)} = \sum_k \frac{b_{ik}b_{jk}}{\lambda_k} \quad (4)$$

where $\lambda_k = \Pr(C_k)$ and $b_{ik} = \Pr(i, C_k)$. Both parameters b_{ik} and λ_k are solved by maximum likelihood estimation. Finally the membership of node i is given by $\Pr(C_k|i) = \frac{\Pr(i,C_k)}{\sum_l \Pr(i,C_l)} = \frac{b_{ik}}{\sum_l b_{il}}$. In [Ren et al 2007], W. Ren et al. factorized the joint probability as

$$\Pr(i,j) = \sum_k \Pr(j|C_k)\Pr(i|C_k)\Pr(C_k) = \sum_k \beta_{ik}\beta_{jk}\pi_k \quad (5)$$

Similarly, the unknown parameters β, π are solved by maximum likelihood estimation. The membership for node i is given by $\Pr(C_k|i) = \frac{\Pr(i|C_k)\Pr(C_k)}{\sum_l \Pr(i|C_l)\Pr(C_l)} = \frac{\beta_{ik}\pi_k}{\sum_l \beta_{il}\pi_l}$. Note that the above two models are closely related to PHITS. Note that using the above derivation, we have $\Pr(v_j|v_i)$ derived as

$$\Pr(j|i) = \sum_k \Pr(j|C_k)\Pr(C_k|i), \quad (6)$$

which is equivalent to the PHITS model in equation (3) with $\gamma_{jk} = \Pr(j|C_k)$ and $\beta_{ik} = \Pr(C_k|i)$. Hence, PHITS and GFM are the essentially the same probabilistic model with PHITS for directed graphs and GFM for undirected graphs.

Popularity Conditional Link (PCL) model

In the models described above, the link probability between two nodes only depends on the community memberships of the two nodes. However, there are many other factors could affect the link generation between nodes. For example, an university website may link to Facebook other than LinkedIn, though Facebook is in the same community (the community of social networking websites) as LinkedIn. To address this issue, the present authors proposed a Popularity Conditional Link (PCL) model [Yang et al 2009b] that introduces a new variable for each node, named "popularity", to model the

difference of nodes in receiving links. The nodes with high popularity would have high probabilities to receive a link. Given the popularities and community memberships, the link probability $\Pr(j|i)$ conditioned on the community variable z_{ij} of node i associated with this link is given as follows

$$\Pr(j|i; z_{ij}, b) = \frac{\gamma_{jz_i} b_j}{\sum_{j'} \gamma_{j'z_i} b_{j'}} \tag{7}$$

which gives the conditional link probability $\Pr(j|i)$ by integrating out z_{ij} as

$$\Pr(j|i; b) = \sum_k \frac{\gamma_{jz_i} b_j}{\sum_{j'} \gamma_{j'z_i} b_{j'}} \gamma_{ik} \tag{8}$$

where γ_{ik} denotes the community membership of node i in community k, and b_j denotes the popularity of node j. As indicated by the above expression, the conditional link probability $\Pr(j|i)$ is proportional to b_j, the popularity of the ending node of the link. It was show that PCL model is an extension of PHITS model in (6) by restricting $\Pr(j|C_k)$ to an explicit form, i.e.,

$$\Pr(j|C_k) = \frac{\Pr(C_k|j)\Pr(j)}{\sum_{j'} \Pr(C_k|j')\Pr(j')} = \frac{\gamma_{jk} b_j}{\sum_{j'} \gamma_{j'k} b_{j'}} \tag{9}$$

It was demonstrated by the authors that the PCL model outperforms PHITS model in both link prediction and community detection [Yang et al 2009b].

Popularity and Productivity Link (PPL) model

PCL model is later on extended by the authors to a general popularity and productivity link (PPL) model [Yang et al 2010]. The motivation is that the probabilistic models proposed before are either *symmetric* (e.g., graph factorization model) in which incoming links and outgoing links are treated equally or *conditional* (e.g., PHITS or PCL) in which only one type (i.e., either incoming or outgoing) of links is modeled, and therefore these models are not suitable for real networks which usually reveals the heavy-tailed degree distribution. PPL model is a non-symmetric link model for directed

network community detection that aims to model both incoming links and outgoing links *simultaneously* and *differentially*. In particular, we introduce latent variables *node productivity* and *node popularity* to explicitly capture the differences of nodes in producing links (outdegree) and receiving links (indegree), respectively. PPL models the joint link probability $\Pr(i, j)$, i.e., how likely there is a directed link from node i to node j. In order to emphasize the different roles played by i and j, we write $\Pr(i, j)$ as $\Pr(i_{\rightarrow}, j_{\leftarrow})$, denoting that node i plays the role of producing the link, and node j plays the role of receiving the link. We model $\Pr(i_{\rightarrow}, j_{\leftarrow})$ as follows

$$\begin{aligned}\Pr(i_{\rightarrow}, j_{\leftarrow}) &= \sum_k \Pr(i_{\rightarrow}|C_k)\Pr(j_{\leftarrow}|C_k)\Pr(C_k) \\ &= \sum_k \left(\frac{\gamma_{ik} a_i}{\sum_{i'} \gamma_{i'k} a_{i'}} \frac{\gamma_{jk} b_j}{\sum_{i'} \gamma_{i'k} b_{i'}} \sum_{i'} \gamma_{i'k} c_{i'} \right) \end{aligned} \tag{10}$$

where a_i denotes the productivity of node i, b_j denotes the popularity of node j, and c_i denotes the weight for computing the prior probability of each community. These variables are normalized such that $\sum_i a_i = \sum_i b_i = \sum_i c_i = 1$.

The authors presented and analyzed three variants of PPL model by imposing different constraints on the parameters c. When setting $c_i = a_i$, the PPL model is reduced to Popularity Link model, from which one can derive the conditional link model of PCL. When setting $c_i = b_i$, we get a Productivity Link model, which only modes the difference of nodes in producing links. If a_i, b_i, c_i are set to be equal, the PPL model reduces to previous symmetric link models (e.g., graph factorization model). By imposing a dirichlet prior on c_i, we obtain another variant of PPL, namely PPL-D model. An important property of PPL model is that it can fit the power-law degree distribution (both indegree and outdegree) of real networks exactly, which was proved and empirically verified by the authors.

Content Analysis

In many networks, such as World Wide Web, online blogs, and citation networks, the contents of each node are usually available and can be represented by a vector of attributes. For instance, each node in the World Wide Web is a web page, and its content can be represented by a vector of word histograms. In addition to the link information between nodes, the contents of individual nodes also provide valuable information for deciding the community structure. For example, in a paper citation network, the contents of papers in machine learning are significantly different from the contents of the papers in natural language processing despite the potential citation links between papers in the two areas.

Traditional content analysis for clustering include k-means [Jain and Dubes 1988], single-linkage [Sibson 1973], complete-linkage [Defays 1977], and etc. These methods are usually applied to unstructured data. In the context of networks, in order to combine with model based link analysis methods, model based methods for content analysis are usually used. For example, Gaussian mixture model is a traditional model for clustering that assumes the data points in each cluster are generated from a Gaussian distribution. To generate a data point, it first samples a community from its community memberships by a multinomial distribution and then samples the data point from a Gaussian distribution parameterized by some unknown parameters. However, GMM makes a strong assumption about the data distribution and therefore limits its application to a narrow field. Another popular model for content analysis, in particular for document analysis, is the topic model [Ho et al 2002, Blei et al 2003].

Topic models originate from the document analysis. It aims to identify document topics from a collection of documents by assuming each document is essentially a mixture of multiple topics. Each topic, in the sense of statistical modeling, is represented by a probability distribution over words. Most topic models are generative models that

describe the generation of a textual document by a stochastic process. More specifically, to generate a document **d**, one first sample a topic from a prior distribution, and then sample words for the given topic. Many well known topic models have been proposed, including probabilistic latent semantic analysis (PLSA) [Ho et al 2002], latent dirichlet allocation (LDA) [Blei et al 2003], hidden topic markov model (HTMM) [Gruber et al 2007], correlated topic model (CTM) [Blei and Lafferty 2006] and author topic model (ATM) [Rosen-Zvi et al 2004]. Note that all these model can be directly applied to community detection based on document content if we map a document in topic models to a node in network, and a document topic to a network community.

Combined Link and Content Analysis

In many applications of network analysis, both link and content information are available. Most existing work on community detection focus on either link analysis or content analysis. However, neither information alone is sufficient for accurately determining the community memberships: the link information can be sparse and noisy, and often results in a poor partition of networks; the irrelevant content attributes could significantly mislead the process of community detection. It is therefore important to combine the link analysis and content analysis for community detection in networks. In the literature, most research model the content information and the link information by two separate generative processes, and combine them via the shared community memberships of nodes. The hidden community memberships are either obtained by maximum likelihood estimation or Bayesian inference via approximate inference. In this section, we first review the PHITS-PLSA model, a well known model for combining link and content information. In this model, PHITS is used to model the link information, and PLSA is used to model the content information; both probabilistic models are combined through the topic mixtures. Another example is the LDA-Link-Word model, which can

be viewed as a Bayesian extension of PHITS-PLSA. In addition to the LDA-Link-Word model, a number of algorithms [Nallapati et al 2008, Gruber et al 2008] were developed to combine various link models with the LDA model. Besides probabilistic models, several non-probabilistic approaches [Zhu et al 2007], such as matrix factorization, are developed to combine the link and content information for community finding. Finally and most importantly, we present a state-of-the-art approach for combing link and content by the present authors, namely a discriminative approach for combining link and content.

PHITS-PLSA

In PHITS-PLSA [Cohn and Hofmann 2001], PHITS is used to model link information, and PLSA is used to model content information. It is the community memberships that allow us to combine these models. More specifically, the log-likelihood of data for PHITS-PLSA is simply a sum of both models, computed as

$$\log \mathcal{L} = \sum_i \left[\alpha \sum_j s_{ij}^w \log \sum_k \beta_{jk}^w \gamma_{ik} + (1-\alpha) \sum_j s_{ij}^l \log \sum_k \beta_{jk}^l \gamma_{ik} \right] \tag{11}$$

where β_{*k}^w specifies the word distribution for community k; β_{*k}^l specifies the link distribution for community k; s_{i*}^w is the word histogram for node i; s_{ij}^l encodes the weight for the link between node i and node j; α is the combination coefficient that balances the effect between PHITS and PLSA.

LDA-Link-Word

LDA-Link-Word model [Erosheva et al 2004] employs LDA to model both the content information and the link information. For each community k, its distributions on words and nodes are denoted by β_{*k}^w and β_{*k}^l, respectively. To generate words and links for a node i, the community memberships are first sampled from a Dirichlet prior, i.e., $\gamma_i \sim Dir(\alpha_1, \cdots, \alpha_K)$. For each word j, a community variable z_{ij}^w is sampled from the

community membership by a Multinomial distribution, i.e., $z_{ij}^w \sim Mul(\gamma_{i1}, \cdots, \gamma_{iK})$. Similarly, for each link to node j, a community variable z_{ij}^l is sampled from the same community membership $z_{ij}^l \sim Mult(\gamma_{i1}, \cdots, \gamma_{iK})$. Words and links are sampled from distributions $Mult(\beta^w_{*z_{ij}^w})$ and $Mult(\beta^l_{*z_{ij}^l})$, respectively. Given this, we can write the joint probability of words and links and the community variables z_{ij}^w and z_{ij}^l for node i by

$$\begin{aligned} &\Pr(w_1^i, \cdots w^i_{N_i^w}, l_1^i, \cdots, l^i_{N_i^l}, z_{i1}^w, \cdots, z^w_{iN_i^w}, z_{i1}^l, \cdots, z^l_{iN_i^l} | \alpha, \beta^w, \beta^l) \\ &= \int d\gamma_i Dir(\gamma_i|\alpha_1, \cdots, \alpha_K) \prod_{j=1}^{N_i^w} \prod_k (\beta^w_{jk}\gamma_{ik})^{z_{ij}^w} \prod_{j=1}^{N_i^l} \prod_k (\beta^l_{jk}\gamma_{ik})^{z_{ij}^l} \end{aligned} \tag{12}$$

Then the log-likelihood is

$$\log \Pr(d^i, l^i) = \sum_i \log \sum_{z_i^w} \sum_{z_i^l} \int d\gamma_i Dir(\gamma_i|\boldsymbol{\alpha}) \prod_{j=1}^{N_i^w} \prod_k (\beta^w_{jk}\gamma_{ik})^{z_{ij}^w} \prod_{j=1}^{N_i^l} \prod_k (\beta^l_{jk}\gamma_{ik})^{z_{ij}^l} \tag{13}$$

where d^i denotes the set of words in document i, l^i denotes the set of links from node i. A variational inference method is used to efficiently derive the posterior distribution of γ_i, which in return determines the community memberships of node i.

Link models and LDA

In the literature, several other probabilistic models [Nallapati et al 2008, Gruber et al 2008] were proposed to combine link information and content information in the framework of LDA. In these methods, words are assumed to be generated following the process of LDA, while the generative process of links often differ from one method to another. For example, in Pairwise Link LDA model [Nallapati et al 2008], the mixed membership stochastic block model [Airoldi et al 2006] is combined with LDA via the shared community memberships γ. Link-PLSA-LDA model proposed by Nallapati et al [2008] makes a simplifying assumption that the link structure is a bipartite graph with all links emerging from the set of citing documents and pointing to the set of cited documents and uses different processes to model the citing documents and the cited

documents. In Latent Topic Model for Hypertext [Gruber et al 2008], the authors assumed that the links originate from a word, and each word can have at most one link associated with it. The generation of links is carried out by iterating over all the words in the document and for each word determining whether to create a link and if so, what is the target document.

Matrix Factorization

Since LSI [Deerwester et al 1990], matrix factorization has been widely used in document analysis. Essentially, these approaches tried to map the documents into a latent space, which gives reduced dimension and high quality representation. LSI [Deerwester et al 1990] used SVD to decompose the document term matrix. After that many matrix factorization methods have been used for document clustering. Xu et al [2003] used non-negative matrix factorization to cluster documents. The term-document matrix $X \in \mathbb{R}^{d\times n}$ is factorized into two non-negative matrices $U \in \mathbb{R}^{d\times K}$ and $V \in \mathbb{R}^{n\times K}$ by minimizing the squared error, i.e., $\min_{U\geq 0, V\geq 0} \|X - UV^T\|_F$, where each column of matrix U can be viewed as the latent representation of cluster centers, the elements of row i of matrix V gives the combination weight on each cluster. For cluster analysis, each document is assigned to the one that has the largest weight in the row of matrix V corresponding to document i. Zhu et al [2007] extended the matrix factorization method for document clustering by combing content and link. Besides factorizing the document-term matrix, they also tried to factorize the link matrix denoted by $W \in \mathbb{R}^{n\times n}$. These two factorizations are combined by the same latent representation of each document. For link factorization, they tried to factorize the link matrix W into ZUZ^T, where $Z \in \mathbb{R}^{n\times K}$ is the latent representation matrix of document. They obtained Z by minimizing the objective of $\min_{Z,U,V} \|W - Z^TUZ\|_F + \alpha\|X - Z^TV\|_F + \beta\|U\|_F + \gamma\|V\|_F$.

A discriminative approach for combined link and content analysis

As we survey above, most approaches that combine link and content for community detection adopt a generative framework where a generative link model and a generative content model are combined through a set of shared hidden variables of community memberships. We argue that such a generative framework suffers from two shortcomings. First, community membership by itself is insufficient to model links; link patterns are usually affected by factors other than communities such as the popularity of a node (i.e., how likely the node is cited by other nodes). Second, the content information often include irrelevant attributes and as a result, a generative model without feature selection usually leads to poor performance.

To address these issues explicitly, the present authors proposed a discriminative approach for combining link and content [Yang et al 2009b]. The approach consists of two parts:

- A popularity (and productivity) link model. In contrast to previous generative link models that only depend on the community memberships; instead, in our model we introduce hidden variables to capture the popularity (and productivity) of nodes in terms of how likely each node is cited by other nodes (and how likely each node is citing other nodes) .
- A discriminative content model. To alleviate the impact of irrelevant content attributes, we adopt a discriminative approach to make use of the node contents. As a consequence, the attributes are automatically weighed by their discriminative power in terms of telling apart salient communities.

We combine the above two models into a unified framework and propose a novel two-stage optimization algorithm for the maximum likelihood inference.

Popularity (and Productivity) Link model

For the link model, we can use popularity conditional link (PCL) model [Yang et al 2009b] or popularity and productivity link (PPL) model [Yang et al 2010]. Here we take PCL as an example. The conditional link probability is given by

$$\Pr(j|i;b) = \sum_k \gamma_{ik} \frac{\gamma_{jk} b_{jk}}{\sum_{j'} \gamma_{j'k} b_{j'k}} \tag{14}$$

Discriminative Content model

Let $\mathbf{x}_i \in \mathbb{R}^d$ denote the content vector of node i. The content information is used to model the memberships of nodes by a discriminative model, given by

$$\Pr(z_i = k) = y_{ik} = \frac{\exp(\mathbf{w}_k^\top \phi(\mathbf{x}_i))}{\sum_l \exp(\mathbf{w}_l^\top \phi(\mathbf{x}_i))} \tag{15}$$

where $\mathbf{w}_k \in \mathbb{R}^m$ is the weighting vector on the features for community k, and $\phi(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^m$ is a feature mapping. We can see that by incorporating the content model, the community membership is no longer specified by parameters γ_{ik}, but rather conditioned on the content through y_{ik} by a softmax transformation. Then, the conditional link probability $\Pr(j|i)$ expressed in Eq. (14) is modified as follows

$$\Pr(j|i;b,\mathbf{w}) = \sum_k y_{ik} \frac{y_{jk} b_j}{\sum_{j'} y_{j'k} b_{j'}}$$

where y_{ik} depends on $\mathbf{w}$ as given in Eq. (15). As revealed in the above expression, we do not generate the content attributes as most topic models do. Instead, by using the discriminative model, with an appropriately chosen weight vector $\mathbf{w}_k$ that assign large weights to important attributes and small weights or zero weights to irrelevant attributes, we avoid the shortcoming of the generative models, i.e., being misled by irrelevant attributes. Finally, the log-likelihood of the combined model is written as

$$\log \mathcal{L} = \sum_{(i \to j) \in \mathcal{E}} s_{ij} \log \sum_k y_{ik} \frac{y_{jk} b_j}{\sum_{j'} y_{j'k} b_{j'}} \tag{16}$$

where s_{ij} denote the weight for the link from node i to node j. To infer the parameters $\mathbf{w}$ and b, a two stage optimization algorithm is proposed in [Yang et al 2009b].

Future Directions

In the following we list several directions that are important for future research in this area.

- Community Detection for Dynamic Networks: extending the community detection algorithms to handle the dynamics in the networks and to detect the evolutions of communities along with the time. Several studies have been devoted to such extensions. Various clustering or community detection algorithms have been extended to their dynamic versions, e.g., evolutionary k-means [Chakrabarti et al 2006], evolutionary spectral clustering [Chi et al 2009], dynamic graph factorization model [Lin et al 2008], dynamic stochastic block model [Yang et al 2009a]. It still remains an unsolved problem how to extend the recently proposed improved link models, e.g., PCL [Yang et al 2009b], PPL [Yang et al 2010] into their dynamic versions. We believe such extensions can not only capture the evolutions of communities of individual nodes but also track the changes of popularities (or productivities) of nodes.
- Community Detection in Heterogeneous networks: incorporating multiple related networks in different domains to improve the performance of community detection in each domain or in one target domain. For example, in order to detect the communities of wikipedia pages, besides the central links between pages we can consider the peripheral connections between pages and their editors and also the networks between editors. We believe by exploring such peripheral connections can yield improved performance in detecting the communities.

- Community Detection in other applications: applying the model based community detection algorithms to other applications, e.g., link prediction, online recommendation. Model based community detection algorithms can be cast into a larger category of algorithms, namely factor (or prototype) based algorithms, where communities can explain the intrinsic factors that trigger the connections between entities (e.g., different interests cause customers to choose different products). It still needs efforts to compare the performance of model based community detection algorithms for other applications to that of existing works.

Cross-references

00006: Communities Discovery and Analysis in Online and Offline Social Networks

00010: Communities in Social Networks, Evolution of

00027: Community Detection, Current and Future Research Trends

00215: Community Discovery and Analysis in Large-Scale Online/Offline Social Networks

00223: Community Evolution

References

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2006) Mixed membership stochastic block models for relational data with application to protein-protein interactions. In: Proceedings of the International Biometrics Society Annual Meeting

Baumes J, Goldberg M, Krishnamoorty M, Magdon-ismail M (2005a) Finding communities by clustering a graph into overlapping subgraphs. In: Proceedings of the 2nd IADIS Appliedd Computing

Baumes J, Goldberg M, Magdon-ismail M (2005b) Efficient identification of overlapping communities. In: Proceedings of the 3rd IEEE International Conference on Intelligence and Security Informatics

Blei DM, Lafferty JD (2006) Correlated topic models. In: Proceedings of the 23rd International Conference on Machine Learning

Blei DM, Ng AY, Jordan MI, Lafferty J (2003) Latent dirichlet allocation. Journal of Machine Learning Research 3

Chakrabarti D, Kumar R, Tomkins A (2006) Evolutionary clustering. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, pp 554–560

Chi Y, Song X, Zhou D, Hino K, Tseng BL (2009) On evolutionary spectral clustering. ACM Trans Knowl Discov Data 3:17:1–17:30

Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Physical Review E 70

Cohn D, Chang H (2000) Learning to probabilistically identify authoritative documents. In: Proceedings of the 17th International Conference on Machine Learning

Cohn D, Hofmann T (2001) The missing link - a probabilistic model of document content and hypertext connectivity. In: Proceedings of the 13th Advanced in Neural Information Processing Systems

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. Journal of the American Society for Information Science 41

Defays D (1977) An efficient algorithm for a complete link method. The Computer Journal 20:364–366

Erosheva E, Fienberg S, Lafferty J (2004) Mixed membership models of scientific publications. In: Proceedings of the National Academy of Sciences

Gregory S (2007) An algorithm to find overlapping community structure in networks. In: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases

Gruber A, Rosen-Zvi M, Weiss Y (2007) Hidden topic markov models. In: Proceedings of the 11st Artificial Intelligence and Statistics

Gruber A, Rosen-Zvi M, Weiss Y (2008) Latent topic models for hypertext. In: Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence

Ho PD, Raftery AE, H MS (2002) Statistical analysis of multiple sociometric relations. Latent space approaches to social network analysis 97

Hofman JM, Wiggins CH (2008) A Bayesian approach to network modularity. Physiccal Review L 100

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of 15th Uncertainty in Artificial Intelligence

Holland PW, Leinhardt S (1974) The statistical analysis of local structure in social networks. Tech. rep.

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Inc.

Kemp C, Griffiths TL, Tenenbaum JB (2004) Discovering latent classes in relational data. Tech. rep.

Kolmogorov V, Zabih R (2004) What energy functions can be minimized via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 26

Lin YR, Chi Y, Zhu S, Sundaram H, Tseng BL (2008) Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In: Proceedings of the 17th international conference on World Wide Web, WWW '08, pp 685–694

Nallapati RM, Ahmed A, Xing EP, Cohen WW (2008) Joint latent topic models for text and citations. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining

Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74

Newman MEJ, Girvan M (2003) Finding and evaluating community structure in networks. Phyiccal Review E 69

Pinney J&WD (2006) Betweenness–based decomposition methods for social and biological networks. In: Proceedings of the 25th Interdisciplinary Statistics and Bioinformatics

Ren W, Yan G, Liao X, Cheng Y (2007) A simple probabilistic algorithm for detecting community structure in social networks. Physical Review E 79

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in Artificial Intelligence

Sibson R (1973) SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 16:30–34

Wang X, Mohanty N, McCallum A (2005) Group and topic discovery from relations and their attributes. In: Proceedings of the 18th Advances in Neural Information Processing Systems

Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press

Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval

Yang T, Chi Y, Zhu S, Gong Y, Jin R (2009a) A bayesian approach toward finding communities and their evolutions in dynamic social networks. In: Proceedings of the 9th SIAM International Conference on Data Mining

Yang T, Jin R, Chi Y, Zhu S (2009b) Combining link and content for community detection: a discriminative approach. In: Proceedings of the 15th ACM SIGKDD Conference on Knowlege Discovery and Data Mining, pp 927–936

Yang T, Chi Y, Zhu S, Gong Y, Jin R (2010) Directed network community detection: A popularity and productivity link model. In: Proceedings of the 10th SIAM International Conference on Data Mining, pp 742–753

Yu K, Yu S, Tresp V (2005) Soft clustering on graphs. In: Proceedings of 18th Advances in Neural Information Processing Systems

Zhu S, Yu K, Chi Y, Gong Y (2007) Combining content and link for classification using matrix factorization. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval

Recommended Reading

`http://en.wikipedia.org/wiki/Community_structure`

Newman MEJ (2003) Fast algorithm for detecting community structure in networks. Physical Review E 69

Newman MEJ (2006b) Modularity and community structure in networks. In: Proceedings of the National Academy of Sciences

Yang T, Chi Y, Zhu S, Gong Y, Jin R (2011) Detecting communities and their evolutions in dynamic social networks–a bayesian approach. Mach Learn 82:157–189

Fu W, Song L, Xing EP (2009) Dynamic mixed membership blockmodel for evolving networks. In: Proceedings of the 26th Annual International Conference on Machine

Learning, pp 329–336

Hagen L, Kahng AB (1992) New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 11

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Transaction on Pattern Analysis and Machine Intelligence 22

Ding CHQ, He X, Zha H, Gu M, Simon HD (2001) A min-max cut algorithm for graph partitioning and data clustering. In: Proceedings of 1st IEEE International Conference on Data Mining