2015 IEEE International Conference on Data Mining

Community Detection Based on Structure and Content: A Content Propagation Perspective

Liyuan Liu*, Linli Xu*, Zhen Wang† and Enhong Chen*
*School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China
llychina@mail.ustc.edu.cn, linlixu@ustc.edu.cn, cheneh@ustc.edu.cn
†AVIC Helicopter Research and Development Institute, zwang25@mail.ustc.edu.cn

***Abstract*—With the recent advances in information networks, the problem of identifying group structure or *communities* has received a significant amount of attention. Most of the existing principles of community detection or clustering mainly focus on either the topological structure of a network or the node attributes separately, while both of the two aspects provide valuable information to characterize the nature of communities. In this paper we combine the topological structure of a network as well as the content information of nodes in the task of detecting communities in information networks. Specifically, we treat a network as a dynamic system and consider its community structure as a consequence of interactions among nodes. To model the interactions we introduce the principle of *content propagation* and integrate the aspects of structure and content in a network naturally. We further describe the interactions among nodes in two different ways, including a linear model to approximate influence propagation, and modeling the interactions directly with random walk. Based on interaction modeling, the nature of communities is described by analyzing the stable status of the dynamic system. Extensive experimental results on benchmark datasets demonstrate the superiority of the proposed framework over the state of the art.**

Keywords*—*information networks, community detection, content propagation

I. Introduction

With the rapid growth of digital and online storage and linkage of data entities, massive generation of relational or networked data has been witnessed during the last decade across various scientific disciplines, producing different types of information networks with certain structures. For an information network, one of the most important structural characteristics is *community* [1], which indicates the group structure of the network. The task of identifying latent groups in a network, also known as community detection or graph clustering, is one of the major topics in data mining and can help discover the structural characteristics such as functional modules of protein-protein interaction networks [2] or groups of people with similar interests in social networks [3].

An information network is usually represented by a graph with data entities corresponding to the nodes in the graph and edges indicating the relations among entities. In addition to the topological *structure*, nodes are usually associated with various types of attributes, which we refer to as *contents* of the nodes. The task of identifying communities involves discovering groups with common properties, such as similarity among group members or densely connected structure.

Most of the existing approaches tackle the task of community detection by considering a certain criterion of "groupness", such as sharing similar entity contents among group members or being densely connected inside the group. However, they mainly focus on analyzing either the topological structure of the graph, or the contents of nodes separately. Examples include methods based on structural criterion such as normalized cut [4] and modularity [5]; as well as methods based on attribute similarity [6].

On the other hand, both of the topological structure and contents of nodes provide valuable information to characterize the nature of communities, which should be compact in structure and similar in content simultaneously. Recently, some approaches have been developed to incorporate the structure and content information for community detection [7], [8], [9], [10]. Among them, probabilistic models have been applied to fuse content analysis and link analysis in a unified framework. Examples include generative models that combine a generative link model and a generative content model through some shared hidden variables [11], [12]. A discriminative model is proposed in [7] where a conditional model for link analysis and a discriminative model for content analysis are unified. In addition to the probabilistic models, some approaches integrate the two aspects of information from other directions. For instance, a similarity based method [13] adds virtual attribute nodes and edges in a network and computes the similarity based on the augmented network, which may suffer from significant expansion and complication of structure if the number of attributes is large.

Unlike the approaches discussed above, in this paper, we take a different perspective to combine the topological structure and content information of nodes in the task of community detection. Specifically, we treat a network as a dynamic system and focus on the adaptive formation of communities by considering the interactions among nodes and analyzing the nature of communities. Interactions in a network occur with information sent to or received from every node. Therefore we consider *information propagation*, which is a fundamental factor in the study of groups in sociology as well as a key element to identify community structure in networks. In general, the community structure can be viewed as a stable status in a dynamic system or network, which can be described as repeated interactions among nodes or information propagation.

In recent works, information propagation, especially influence propagation has been applied to tackle the community detection task in specific scenarios. For instance [14] considers the community detection problem when the network is not

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.105

available while a log of user activity is given, and [15] focuses on clustering heterogeneous networks where influence is calculated to measure the similarity between nodes. These methods share a common property that they are based on the strength of propagation, which in general is only related to the network structure.

In our framework, we detect communities by considering propagation as an essential element of groupness. On one hand, propagation is an abstraction of the interactions in a network; on the other hand, group characteristics such as similar node contents and densely connected structure can both be explained by intense information propagation among group members. Furthermore, to integrate the structure and content in a network naturally, we introduce *content propagation*. We describe the process of content propagation with principles of influence propagation as well as random walk, based on which we design two ways of calculating the content of propagation correspondingly. In the meantime, we consider a network as a dynamic system and a community as the stable status shared by the nodes in that community. Specifically, we assume nodes in the same community are likely to receive the same amount of content propagation, based on which communities can be identified.

To the best of our knowledge, the framework proposed in this paper is the first that describes the nature of communities with both dense structures and similar contents based on the interaction dynamics among nodes through information propagation. We conduct extensive experiments on benchmark datasets and show significant improvements over the state of the art.

The rest of the paper is organized as follows. In Section II we give a brief review of the related work. In Section III, we present the content propagation framework. The algorithm and optimization strategy are introduced in Section IV, followed by extensive experimental results in Section V to demonstrate the effectiveness of the proposed method. The paper is then concluded in Section VI.

II. Related Work

There exist two aspects of related work regarding the topic here, which are community detection with combination of structure and content, and information propagation. In Section I we have briefly discussed existing approaches that detect communities using structure and content of a network. In this section, we review related principles of information propagation, including influence propagation and random walk.

A. Influence Propagation

Influence defines the impact that an individual has on others which leads to the change of their attributes from their out-links. A large amount of studies have been conducted to analyze the patterns of influence propagation from nodes, which include the popular models of Independent Cascade (IC) [16] and Linear Threshold (LT) [17]. Specifically, in the independent cascade model, each activated node has a single chance to influence or activate its neighbors independently, and a node i activates a node j successfully with a probability. The propagation terminates at a step when there is no newly activated node.

In most scenarios, the models of independent cascade or linear threshold are used to calculate the influence spread and simulate the propagation of certain viral contagion. Based on the two models, various extensions have been proposed. In [18], it is proven that calculating the influence spread of a seed set in the IC model is #P-hard. Monte Carlo simulation is applied to approximate the influence spread in [19], which requires high computational costs. A quick approximation of influence spread is then proposed in [20] by solving a linear system. A similar linear model is proposed in [21] to interpret PageRank with influence propagation. In this paper, we also consider using a linear model to describe the content propagation.

B. Random Walk

Random walk has been widely used in graph-based learning. For example, the probability that an unlabeled node shares the same label with labeled nodes can be calculated based on this principle [22]. A variation of random walk is used to add ghost edges to a network to enable the flow of information from labeled to unlabeled nodes [23]. In addition, random walk is used to compute the similarity of an augmented network that adds virtual attribute nodes and edges to integrate structural and attribute information [13]. In this paper, we also employ the random walk principle to calculate the probability of propagating the content of a node to another node.

III. Framework of Content Propagation

A network with both structure and content can be represented by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{F}\}$ which includes a set of N nodes $\mathcal{V} = \{v_i | i \in [1, N]\}$ connected by a set of M edges $\mathcal{E}$. The network considered here is directed and $e_{i,j}$ indicates an edge from v_i to v_j. $\mathcal{F} = \{\mathbf{f}_i | i \in [1, N]\}, \mathbf{f}_i \in \mathbb{R}^d$ is the set of content features associated with nodes in V, and we record all the feature vectors in the feature matrix $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]$. Given $\mathcal{G}$, the task is to find a set of communities $\mathcal{C} = \{C_i | i \in [1, K]\}$, K is the number of communities.

To describe the process of content propagation, we take all possible relationship into account. Specifically, we not only consider "linking to" as a bridge of propagation, but also "being linked to", which implies that both directions of the edges are considered. To ensure that every node can receive content propagation from at least one node, we assume every node can propagate to itself. Therefore, the adjacency matrix L of the network $\mathcal{G}$ can be computed as

$$L_{ij} = \begin{cases} 1, & \text{if } e_{i,j} \in \mathcal{E}, \text{ or } e_{j,i} \in \mathcal{E}, \text{ or } i = j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Given L, D can be defined as a diagonal matrix with $D_{ii} = \sum_{j=1}^{N} L_{ij}, \forall i \in [1, N]$.

A. Content Propagation

Next we define the process of content propagation with two principles. First, we consider a linear approximate model of influence propagation to describe content propagation. Alternatively, we establish the process of content propagation with the random walk principle. Specifically, we calculate the probability that the content of a node propagates to another based on the influence propagation or random walk principle, and derive the content of propagation accordingly.

1) Content Propagation Based on Influence Propagation: As aforementioned in Section II, various models have been proposed to model the patterns of influence propagation. In general, these propagation processes end until no more inactive nodes become activated, therefore they usually diffuse rapidly and are less predictable in practice [24]. On the other hand, the formation of communities needs a long period of time and is more complicated. As a consequence, we consider a stochastic description about the probability of a node influencing another with a linear approximation of influence propagation, and we treat it as the probability that the content of a node propagates to another. Similar to [21], [20], we suppose the probability of a node receiving content from others depends on its neighbors, while this probability is always positive given that every node can propagate its content to itself according to our assumption. Thus, we propose the following propagation model:

Definition 1 (Content Propagation Model 1 (CPIP)): Denote the probability that the content of node i is propagated to node j as $w_{i,j}$, which satisfies $(\forall i, j \in [1, N], i \neq j)$:

$$w_{i,i} = \beta + \alpha \sum_{e_{l,j} \in \mathcal{E}} w_{i,l} t_{l,i} \quad w_{i,j} = \alpha \sum_{e_{l,j} \in \mathcal{E}} w_{i,l} t_{l,j}$$

where $t_{l,j}$ is the transition probability from node l to node j, $\beta > 0$ is a constant corresponding to the probability of a node propagating its content to itself, α is the damping coefficient of this propagation process.

The values of $w_{i,j}$ constitute a propagation matrix for CPIP which is denoted as R^{I}, $R^{\mathrm{I}}_{ij} = w_{i,j}$.

The transition probability $t_{l,j}$ can be computed by $\frac{L_{lj}}{\sum_{i=1}^{N} L_{ij}}$ similar to [21]. If we define a transition matrix T such that $T_{lj} = t_{l,j}$, T can be calculated by

$$T = LD^{-1} \tag{2}$$

Next if we let $\gamma = \frac{\alpha}{\beta}$, according to the constraints in Definition 1, R^{I} can be formulated as

$$R^{\mathrm{I}} = (\beta I - \alpha T)^{-1} = \beta(I - \gamma LD^{-1})^{-1} \tag{3}$$

Given $w_{i,j}$, the content of propagation received by node i can be defined as $\mathbf{g}^{\mathrm{I}}_i = \sum_{j=1}^{N} \mathbf{f}_j \cdot w_{j,i}$, and also written in the matrix form:

$$G^{\mathrm{I}} = FR^{\mathrm{I}} = \beta F(I - \gamma LD^{-1})^{-1} \tag{4}$$

As we will mention in the following section, the positive parameter β will not influence the structure of detected communities because it is the same for every node.

2) Content Propagation with Random Walk: Similar to [22], which uses random walk to calculate the probability of a pair of data points sharing the same label in semi-supervised learning, here we describe the process of content propagation with random walk.

The one step transition probability P_{ij} of random walk from node i to node j can be obtained from the network directly: $P_{ij} = \frac{L_{ij}}{\sum_{l=1}^{N} L_{il}}$, or

$$P = D^{-1}L \tag{5}$$

Notice the subtle difference between (5) and (2).

Furthermore, the t step transition probability $P_{t|0}(j \mid i)$, which denotes the probability of a node arriving at node j at time t given that it started from node i at time 0, can be computed as $P_{t|0}(j \mid i) = [P^t]_{ij}$. In the semi-supervised learning task in [22], t is treated as a time scale parameter and set to some given value, and the goal is to solve for $P_{0|t}(i \mid k)$ by assuming the start point of random walk is chosen uniformly at random, i.e., $P(i) = \frac{1}{N}$. On the other hand, here we use random walk to describe content propagation for the task of community detection, which is a long-lasting process involving a "mixture" of different scales of time, rather than a process limited to a single scale of time. Therefore we do not fix a specific value for t, instead we treat it as a random variable from $[0, \infty)$. We record the transition probability here as $P_{\mathbf{t}|0}(j \mid i)$, and it can be calculated by

$$P_{\mathbf{t}|0}(j \mid i) = \sum_{s=0}^{\infty} P_{s|0}(j \mid i) \cdot P(\mathbf{t} = s).$$

We assume t follows a geometric distribution, or $P(\mathbf{t} = s) = \lambda(1-\lambda)^s$, then

$$P_{\mathbf{t}|0}(j \mid i) = \sum_{s=0}^{\infty} [P^s]_{ij} \cdot \lambda(1-\lambda)^s$$

If we record $P_{\mathbf{t}|0}(j \mid i)$ as S_{ij}, we have

$$\begin{aligned} S &= \lambda P + \cdots + \lambda(1-\lambda)^s P^s + \ldots \\ &= \lambda(I - (1-\lambda)P)^{-1} \end{aligned} \tag{6}$$

Next, with the similar assumption that the start point of a random walk is chosen uniformly at random, we can calculate $P_{0|\mathbf{t}}(i \mid j)$ as

$$P_{0|\mathbf{t}}(i \mid j) = \frac{P_{\mathbf{t}|0}(j \mid i)}{\sum_{l=1}^{N} P_{\mathbf{t}|0}(j \mid l)} = \frac{S_{ij}}{\sum_{l=1}^{N} S_{lj}} \tag{7}$$

which denotes the probability that node j receives content propagation from node i.

The values of $P_{0|\mathbf{t}}(i \mid j)$ constitute a matrix R^{W}, which is similar to R^{I} in (3) of CPIP, while being slightly different in form due to the different random walk principle. R^{W} provides an alternative method to calculate the probability of content propagation with the random walk principle (CPRW).

Given R^{W}, the content propagation received by node i can be similarly calculated by $\mathbf{g}^{\mathrm{W}}_i = \sum_{j=1}^{N} \mathbf{f}_j \cdot R^{\mathrm{W}}_{ji}$, and written in the matrix form:

$$G^{\mathrm{W}} = FR^{\mathrm{W}}. \tag{8}$$

B. Stability of Content Propagation

We can now sum up the procedure of calculating the propagated content as follows:

- Start with calculating the probability of the content of a node propagating to another, i.e., the propagation probability, R^{I} according to (3) in CPIP or R^{W} according to (7) in CPRW.

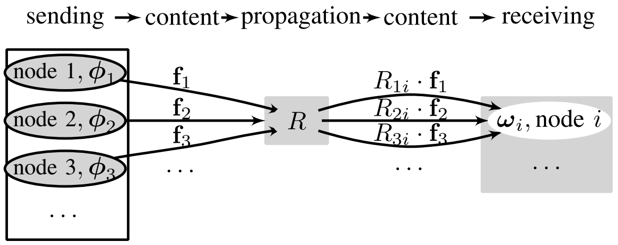

sending → content → propagation → content → receiving
node 1, ϕ_1
node 2, ϕ_2
node 3, ϕ_3
$\mathbf{f}_1$
$\mathbf{f}_2$
$\mathbf{f}_3$
R
$R_{1i} \cdot \mathbf{f}_1$
$R_{2i} \cdot \mathbf{f}_2$
$R_{3i} \cdot \mathbf{f}_3$
$\boldsymbol{\omega}_i$, node i
...

Fig. 1: The process of content propagation. R is the matrix of propagation probability.

- Next, the propagated content that node j receives is
$$\sum_{i=1}^{N} \mathbf{f}_i \cdot P(\text{content of node } i \text{ propagates to node } j),$$
which can be calculated according to (4) in CPIP or (8) in CPRW alternatively.

In this subsection, we analyze the stability of contents in the process of content propagation, assuming them to be random variables. Communities are detected by maximizing likelihood of the statistical model with this stability assumption. This analysis is not restricted to CPIP or CPRW. For simplicity, we denote R^{I} and R^{W} uniformly as R.

The underlying principle of content propagation is that: the content sent by a node reveals the characters of that node; and the content propagation a node receives would influence and change its characters, consequentially alter the distribution of the content propagation it sends. Based on that, we define the content in propagation below.

Definition 2 (Content in Propagation): For node i, it acts and propagates the content of itself $\mathbf{f}_i$ to the network, which is an observation of the random variable ϕ_i. In the meantime, it receives propagated contents from other nodes, marked as $\mathbf{g}_i$, which is an observation of the random variable $\boldsymbol{\omega}_i$.

The received contents for all the nodes can be calculated in a matrix G according to (4) or (8), which is regarded as an observation of the random variable $\boldsymbol{\omega}$. The process of content propagation is shown in Fig. 1.

The key element of community detection with content propagation is the assumption of community consistency. Specifically, when the propagation reaches stability, nodes in the same community are likely to achieve uniform stability, or in other words, receive the same amount of content propagation. Therefore, the task of detecting communities can be viewed as the problem of finding a set of communities C_k such that

$$\forall i, j \in C_k, \quad \mathbb{E}[\boldsymbol{\omega}_i] = \mathbb{E}[\boldsymbol{\omega}_j] = \boldsymbol{\mu}_k \tag{9}$$

where $\boldsymbol{\mu}_k$ denotes the expectation of content propagation received by the community C_k.

Based on the condition of stability (9) above, given an indicator Y we have

$$\mathbb{E}[\boldsymbol{\omega}_i] = \sum_{k=1}^{K} \boldsymbol{\mu}_k \cdot Y_{ik}, \qquad \text{where } Y_{ik} = \begin{cases} 0 & i \notin C_k \\ 1 & i \in C_k \end{cases} \tag{10}$$

At the same time, the content propagated to node i is

$$\boldsymbol{\omega}_i = \sum_{j=1}^{N} \boldsymbol{\phi}_j \cdot R_{ji}$$

which is a sum of random variables. Therefore the distribution of $\boldsymbol{\omega}_i$ can be approached by a Gaussian distribution $\mathcal{N}(\boldsymbol{\Theta}_i, \Sigma_i)$ due to the Central Limit Theorem (CLT). Combined with the expectation of the propagated content $\boldsymbol{\omega}_i$ in (10), the density function of the Gaussian distribution of $\boldsymbol{\omega}_i$ can be written as

$$f(\boldsymbol{\omega}_i = \mathbf{x} \mid Y, \boldsymbol{\mu}) = \frac{1}{(2\pi)^{\frac{q}{2}} \|\Sigma_i\|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{x} - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_k)^{\top} \cdot \Sigma_i^{-1}(\mathbf{x} - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_k))$$

Assuming all the variables in the distribution are independent of each other and the covariance matrix $\Sigma_i = I$, given the observation of $\boldsymbol{\omega}_i$, $\mathbf{g}_i$ is calculated, the log likelihood of the stable model of content propagation can be formulated as

$$\begin{aligned} \log \mathcal{L}(Y, \boldsymbol{\mu}) &= \sum_{i=1}^{N} \log(f(\mathbf{g}_i)) \propto -\sum_{i=1}^{N} \|\mathbf{g}_i - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_i\|_2^2 \\ &\propto -\sum_{i=1}^{N} \| \sum_{j=1}^{N} R_{ji} \cdot \mathbf{f}_j - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_k\|_2^2 \end{aligned} \tag{11}$$

Community detection based on content propagation can then be achieved by maximizing the likelihood in (11), or solving the following optimization problem

$$\begin{aligned} &\min_{Y, \boldsymbol{\mu}_k} \sum_{i=1}^{N} \| \sum_{j=1}^{N} R_{ji} \cdot \mathbf{f}_j - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_k\|_2^2 \\ &\text{subject to} \quad Y \in \{0,1\}^{N \times K}, \ \sum_{k=1}^{K} Y_{ik} = 1 \end{aligned} \tag{12}$$

IV. Optimization

The optimization problem (12) is not jointly convex of the variables Y and $\boldsymbol{\mu}_k$ in addition to the binary constraints on Y, which implies computational hardness of solving the problem globally. An alternative optimization algorithm is therefore applied here to get a local optimum solution. To achieve a high quality local solution, we design two strategies of initialization by first finding an approximate solution of (12).

A. Initialization

First, we normalize Y with $Z_{ik} = \frac{Y_{ik}}{\sum_{j=1}^{N} Y_{jk}}$ such that $\sum_{j=1}^{N} Z_{jk} = 1$. To solve the problem (12), we can first solve for $\boldsymbol{\mu}_k$ with Y fixed:

$$\boldsymbol{\mu}_{\text{opt},k} = \sum_{i=1}^{N} \frac{\mathbf{g}_i \cdot Y_{ik}}{\sum_{j=1}^{N} Y_{jk}} = \sum_{i=1}^{N} \mathbf{g}_i \cdot Z_{ik} = \sum_{i=1}^{N} \mathbf{g}_i \cdot [\sqrt{Z}]_{ik}^2 \tag{13}$$

where $\sqrt{Z}$ is a matrix defined as the element-wise square root of Z. In another word, we have:

$$\begin{aligned}\mathcal{J}_A &= \min_{\boldsymbol{\mu}_k} \sum_{i=1}^{N} \|\mathbf{g}_j - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_k\|_2^2 \\ &= \sum_{i=1}^{N} \|\mathbf{g}_j - \sum_{k=1}^{K} Y_{ik} \cdot \boldsymbol{\mu}_{\text{opt},k}\|_2^2\end{aligned}$$

By substituting (13) into (14), after some algebraic transformations, we obtain

$$\mathcal{J}_A = \sum_{1 \le i \le N} \|\mathbf{g}_i\|_2^2 - \mathcal{J}_B$$

where $\sum_{1 \le i \le N} \|\mathbf{g}_i\|_2^2$ is a constant and

$$\begin{aligned}\mathcal{J}_B &= \sum_{\substack{1 \le i,j \le N \\ 1 \le k \le K}} \{(\mathbf{g}_i \cdot \mathbf{g}_j)\sqrt{Z_{ik}Z_{jk}}\} \\ &= \sum_{\substack{1 \le i,j \le N \\ 1 \le k \le K}} \{\sqrt{Z_{ik}Z_{jk}} \sum_{l=1}^{N} (\mathbf{f}_l \cdot R_{li}) \sum_{m=1}^{N} (\mathbf{f}_m \cdot R_{mj})\} \\ &= \sum_{\substack{1 \le l,m \le N \\ 1 \le k \le K}} \{(\mathbf{f}_l \cdot \mathbf{f}_m) \sum_{i=1}^{N} (\sqrt{Z_{ik}} R_{li}) \sum_{j=1}^{N} (\sqrt{Z_{jk}} R_{mj})\} \\ &= \text{trace}(\sqrt{Z}^{\top} R^{\top} F^{\top} F R \sqrt{Z})\end{aligned}$$

Next, given $\boldsymbol{\mu}_{\text{opt},k}$, one can solve for Y by minimizing $\mathcal{J}_A$, or maximizing $\mathcal{J}_B$:

$$\max_{Z} \mathcal{J}_B \tag{14}$$

Since $\|\sqrt{\mathbf{z}_k}\| = 1$ due to $\sum_{j=1}^{N} Z_{jk} = 1$ where $\mathbf{z}_k$ is the k-th column of Z, we can relax the discrete constraints on Z and get an approximate solution to (14) by setting columns of $\sqrt{Z}$ to be the first K eigenvectors of the matrix $R^{\top} F^{\top} F R$, $\boldsymbol{\mu}_k$ can then be initialized according to (13). This strategy of initialization is called PI here [25], [26].

The PI strategy is based on a continuous relaxation of Y or Z. We further design another initialization strategy by first changing the elements of of $\mathcal{J}_B$,

$$\mathcal{T}_B = (\mathbf{f}_l \cdot \mathbf{f}_m) \sum_{i=1}^{N} (\sqrt{Z_{ik}} R_{li}) \sum_{j=1}^{N} (\sqrt{Z_{jk}} R_{mj})$$

to

$$(\mathbf{f}_l \cdot \mathbf{f}_m) \sum_{i=1}^{N} (\sqrt{Z_{ik}} R_{li} \cdot Y_{ik}^{\rho}) \sum_{j=1}^{N} (\sqrt{Z_{jk}} R_{mj} \cdot Y_{jk}^{\rho})$$

which does not change the value of the element when Y still takes discrete values. When $\rho \to \infty$, the local solution of Y by optimizing the objective will be encouraged to be binary even after we relax the discrete constraints on Y to be $\sum_{k=1}^{K} Y_{ik} = 1$. This is mainly because:

$$\lim_{\rho \to \infty} Y_{ik}^{\rho} = \begin{cases} 1 & Y_{ik} = 1 \\ 0 & 0 \le Y_{ik} < 1 \end{cases}$$

Although a larger ρ could result in better performance, it is harder to calculate. Meanwhile, it inspires us a new strategy of getting an approximate solution. Because $0 \le R_{ij} \le 1, \forall i, j$, we can derive the following

$$\begin{aligned}&\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}} \\ \ge &\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}} R_{li} R_{mj} (Y_{ik}Y_{jk})^{\rho} \\ \ge &\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}} \\ \cdot\ &\frac{\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}}}{\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}}}\end{aligned} \tag{15}$$

Notice that our task is to maximize $\mathcal{J}_B$, whose element $\mathcal{T}_B$ has a factor equal to $\sum_{i,j} \sqrt{Z_{ik}Z_{jk}} R_{li} R_{mj} (Y_{ik}Y_{jk})^{\rho}$ which satisfies

$$\sum_{i,j} \sqrt{Z_{ik}Z_{jk}} R_{li} R_{mj} (Y_{ik}Y_{jk})^{\rho} \le \sum_{i,j} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}}$$

according to the inequality (15). In the meantime, there exists a lower bound ξ of $\sum_{i,j} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}}$ during maximization; while based on AM-QM Inequalities, we have

$$\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}} \le n \cdot \sqrt{\sum_{1 \le i,j \le N} Z_{ik}Z_{jk}} = n$$

therefore

$$\frac{\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}}}{\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}}} \ge \frac{\xi}{n}$$

combined with (15), we get

$$\begin{aligned}&\sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}} R_{li} R_{mj} (Y_{ik}Y_{jk})^{\rho} \\ \ge\ &\frac{\xi}{n} \cdot \sum_{1 \le i,j \le N} \sqrt{Z_{ik}Z_{jk}R_{li}R_{mj}}(Y_{ik}Y_{jk})^{\frac{\rho}{2}}\end{aligned} \tag{16}$$

If we mark $\mathcal{T}_C = (\mathbf{f}_l \cdot \mathbf{f}_m) \sum_{i=1}^{N} (\sqrt{Z_{ik} R_{li}}) \cdot \sum_{j=1}^{N} (\sqrt{Z_{jk} R_{mj}})$, by combining (15) and (16) we obtain

$$\mathcal{T}_C (Y_{ik}Y_{jk})^{\frac{\rho}{2}} \ge \mathcal{T}_B (Y_{ik}Y_{jk})^{\rho} \ge \frac{\xi}{n} \mathcal{T}_C (Y_{ik}Y_{jk})^{\frac{\rho}{2}}$$

which establishes the equivalence between $\mathcal{T}_C (Y_{ik}Y_{jk})^{\frac{\rho}{2}}$ and $\mathcal{T}_B (Y_{ik}Y_{jk})^{\rho}$ to some extent. Since a larger ρ could result in better performance, we regard $\mathcal{T}_C$ as a reasonable alternative of $\mathcal{T}_B$. Correspondingly, we solve another optimization problem below to initialize using $\mathcal{T}_C$ as element of the objective:

$$\max_{\sum_{i=1}^{N} Y_{ik}^2 = 1} \mathcal{J}_C$$

where $\mathcal{J}_C = \sum_{\substack{1 \le l,m \le N \\ 1 \le k \le K}} \{(\mathbf{f}_l \cdot \mathbf{f}_m) \cdot \sum_{i,j} \{\sqrt{Z_{ik}Z_{jk}} \cdot \sqrt{R_{li}R_{mj}}\}\}$

$$= \text{trace}(\sqrt{Z}^{T} \sqrt{R}^{\top} F^{\top} F \sqrt{R} \sqrt{Z}) \tag{17}$$

An approximate solution to (17) can be obtained by setting $\sqrt{Z}$ to the first K eigenvectors of the matrix $\sqrt{R}^{\top} F^{\top} F \sqrt{R}$, followed by initializing $\boldsymbol{\mu}$ by (13). This strategy of initialization is called SI.

B. Alternative Optimization

Given the initialization of $\boldsymbol{\mu}$, alternative optimization can be applied to update the values of $\boldsymbol{\mu}$ and Y with:

$$\boldsymbol{\mu}_k^{t+1} = \sum_{i=1}^{N} \mathbf{g}_i \cdot \frac{Y_{ik}^t}{\sum_{j=1}^{N} Y_{jk}^t} \tag{18}$$

$$Y_{ik}^{t+1} = \begin{cases} 1 & \forall m \in [1, K], \|\boldsymbol{\mu}_k^t - \mathbf{g}_i\| \leq \|\boldsymbol{\mu}_m^t - \mathbf{g}_i\| \\ 0 & otherwise \end{cases} \tag{19}$$

C. Algorithm

The algorithm of community detection with content propagation based on influence propagation (CPIP) or random walk (CPRW) is summarized in Algorithm 1. Two strategies of initialization are applied, which are denoted as SI and PI. As a consequence, we have 4 algorithms in the framework of community detection with content propagation, which are CPIP-PI, CPRW-PI, CPIP-SI and CPRW-SI respectively. From the algorithm, it is easy to see that the β value in (3) and (4) does not affect the detected communities, so we can simply set its value to 1.

Algorithm 1 Algorithm of CPIP/CPRW

```
Input: adjacency matrix $L$, feature matrix $F$;
    number of clusters $K$, parameter $\lambda$;
Output: detected communities indicated by $Y$;
calculate $R$ with (3) for CPIP; or (7) for CPRW;
calculate $G$ with (4) for CPIP; or (8) for CPRW;
Initialization (SI): calculate the first $K$ eigenvectors of
    $\sqrt{R}^\top F^\top F \sqrt{R}$ and treat them as initialization of $\sqrt{Z}$;
    or
    Initialization (PI): calculate the first $K$ eigenvectors of
    $R^\top F^\top F R$ and treat them as initialization of $\sqrt{Z}$;
initialize $\boldsymbol{\mu}$ with (13);
while not converged do
    update $Y$ with $\boldsymbol{\mu}$ by (19);
    update $\boldsymbol{\mu}$ with $Y$ by (18);
end while
return $E_n$;
```

V. Experiments

In this section, we empirically validate our framework of content propagation by comparing to the state-of-the-art community detection methods on information or social networks. We evaluate the performance of these methods by examining the accuracy of the detected communities $\mathcal{C}$ with different metrics.

A. Datasets

In the experiments, we conduct investigations on 4 datasets that contain both network structure as well as node attributes:

Citeseer Dataset is a citation network consisting of 3312 scientific publications, each of which is labeled as one of 6 sub-fields[1]. This network contains 4732 links, and every publication is described by a 0/1-valued word vector, indicating whether a word in a dictionary of 3703 unique words appears in the corresponding publication.

Cora Dataset is a citation network with 2708 nodes and 5294 links [27]. Its nodes correspond to publications described by 0/1-valued word vectors and classified into 7 sub-categories. The dictionary here contains 1433 unique words.

Facebook Dataset contains 10 ego-networks from the online social network Facebook [8]. It consists of 4086 nodes and 170174 edges in total. Each node in that network represents an account on Facebook, which is described by a 0/1-valued vector, corresponding to the absence/presence of an attribute. The relations between accounts are represented by edges between the corresponding nodes. In the end, the ground truth communities of this network are defined by social circles.

PubMed Diabetes Dataset consists of 44338 links between 19717 scientific publications from the PubMed database pertaining to diabetes belonging to 3 classes[1]. Each publication in it is described by a tf/idf weighted word vector from a dictionary with 500 unique words.

B. Evaluation Metrics

We use 3 different metrics to evaluate the accuracy of the detected communities $\mathcal{C}$, which are calculated using the ground-truth communities $\mathcal{C}^*$.

F-score: Given the detected communities $\mathcal{C}$ and the ground-truth communities $\mathcal{C}^*$, the F-score between $\mathcal{C}$ and $\mathcal{C}^*$ is defined as

$$F(\mathcal{C}, \mathcal{C}^*) = \sum_{C_i \in \mathcal{C}} \frac{|C_i|}{\sum_{C_j \in \mathcal{C}} |C_j|} \max_{C_j^* \in \mathcal{C}^*} F(C_i, C_j^*)$$

where $F(C_i, C_j^*)$ represents the F-score between C_i and C_j^*.

Jaccard Similarity: The jaccard similarity could be defined for $\mathcal{C}$ and $\mathcal{C}^*$ as in [10]:

$$S(\mathcal{C}, \mathcal{C}^*) = \sum_{C_j^* \in \mathcal{C}^*} \frac{\max_{C_i \in \mathcal{C}} S(C_i, C_j^*)}{2|\mathcal{C}^*|} + \sum_{C_i \in \mathcal{C}} \frac{\max_{C_i^* \in \mathcal{C}^*} S(C_i, C_j^*)}{2|\mathcal{C}|}$$

where $S(C_i, C_j^*)$ is the jaccard similarity between C_i and C_j^*.

Normalized Mutual Information (NMI): The NMI metric can be calculated as:

$$\text{NMI}(\mathcal{C}, \mathcal{C}^*) = \frac{\widehat{\mathcal{MI}}(\mathcal{C}, \mathcal{C}^*)}{\max(H(\mathcal{C}), H(\mathcal{C}^*))}$$

where $H(\mathcal{C})$ is the entropy of the partition $\mathcal{C}$, and

$$\widehat{\mathcal{MI}}(\mathcal{C}, \mathcal{C}^*) = \sum_{C_i, C_j^*} p(C_i, C_j^*) \log \frac{p(C_i, C_j^*)}{p(C_i)p(C_j)}$$

evaluates the mutual information between $\mathcal{C}$ and $\mathcal{C}^*$.

The 3 metrics above all take values from $[0, 1]$, and larger values indicate better quality of communities.

[1] http://linqs.cs.umd.edu/projects/projects/lbc/index.html

TABLE I: Performance of various methods on Citeseer, Cora, Facebook and PubMed Diabetes datasets. Bold values indicate the best performance.

Method	Infomation	Citeseer			Cora			Facebook			PubMed Diabetes		
		F-score	NMI	Jaccard	F-score	NMI	Jaccard	F-score	NMI	Jaccard	F-score	NMI	Jaccard
CNM	Link	0.1735	0.2290	0.1094	0.4210	0.3681	0.2315	0.3964	0.1491	0.1807	0.3560	0.1519	0.1912
Big-CLAM	Link	0.5114	0.2197	0.0872	0.4826	0.2919	0.2340	0.2505	0.1634	0.1374	0.2167	0.1525	0.1111
GibbsLDA++	Content	0.5537	0.2699	0.3813	0.4390	0.1860	0.2833	0.3555	0.0405	0.1507	0.5837	0.1810	0.4160
Adamic Adar	Link+Content	0.6696	0.3671	0.4429	0.6041	0.3920	0.4247	0.6241	0.2490	0.3472	0.5364	0.0747	0.2759
PCL-DC	Link+Content	0.6228	0.3838	0.4568	0.6756	0.4694	0.5191	0.5824	0.2379	0.3044	0.6576	0.3036	0.4822
Circles	Link+Content	0.3405	0.0024	0.1867	0.3595	0.0064	0.1810	0.5449	0.1080	0.2684	0.4460	0.0004	0.2684
CODICIL	Link+Content	0.5953	0.3392	0.4041	0.5857	0.3947	0.4254	0.4479	0.1253	0.2016	0.6891	0.3023	**0.5325**
CESNA	Link+Content	0.5240	0.1158	0.1158	0.6059	0.4671	0.3254	0.4103	0.1836	0.1726	0.3842	0.2723	0.2293
CPRW-PI	Link+Content	**0.7001**	**0.4396**	**0.5031**	0.6247	0.4848	0.4515	0.5327	0.2403	0.2908	0.6821	0.3179	0.4949
CPIP-PI	Link+Content	0.6894	0.4252	0.4954	**0.7018**	0.5071	0.4920	0.5745	0.2801	0.3202	0.6853	0.3105	0.4968
CPRW-SI	Link+Content	0.6863	0.4353	0.5018	0.6893	0.5364	**0.5223**	**0.6277**	0.2717	0.3473	**0.7017**	**0.3378**	0.5216
CPIP-SI	Link+Content	0.6912	0.4253	0.4959	0.6921	**0.5390**	0.5192	0.6248	**0.3240**	**0.3605**	**0.7017**	0.3255	0.5212

C. *Comparison Methods*

We consider 3 classes of baseline community detection methods in our experiments :

Link Only: The first class of baselines consider the structure of networks only, and group densely connected nodes. Both *Big-CLAM* [28] and *Clauset-Newman-Moore (CNM)* [29] are the state-of-the-art overlapping community detection methods based on structure. We take implementations of the two methods from *SNAP*[2].

Content Only: The second class of baselines only consider the content of nodes. Here we use *GibbsLDA++*[3] as a representative, which is a C/C++ implementation of LDA using Gibbs sampling [30]. We treat the content vectors as documents, which belong to the groups identified by the latent topics.

Combining Link and Content: The third class of baselines consist of approaches that detect communities by considering both the content of nodes and the network structure. We choose 5 different state-of-the-art representatives. *Adamic Adar* [31] calculates pairwise node similarity by utilizing both kinds of information and then detects communities with spectral clustering [32]; *PCL-DC* [7] unifies a conditional model for link analysis and a discriminative model for content analysis to find non-overlapping communities; *Circles* [8] is developed to detect overlapping hard memberships in social networks; *CODICIL* [9] presents a biased edge sampling procedure and gets an edge set by leveraging both content and network structure; and *CESNA* [10] detects overlapping communities by assuming communities "generate" both the network structure and content. For the last 4 baseline methods, we use implementations provided by their authors.

D. *Experimental Setup and Evaluation*

In this subsection, we evaluate the quality of communities detected by our methods and compare with the approaches listed above. Our proposed framework consists of 4 methods, corresponding to content propagation based on influence propagation (CPIP) and random walk (CPRW) with 2 initialization strategies including PI and SI. Correspondingly, the 4 methods are denoted as CPIP-PI, CPRW-PI, CPIP-SI and CPRW-SI respectively.

For the proposed framework of content propagation, we set the parameter $\lambda = 1 - \gamma = 0.1$ and evaluate our methods with this single parameter value. In the meantime, the parameters of the comparison methods from the first two classes are set to their default values. For each approach in the third class, we choose the parameter which yields the best performance for each evaluation metric.

The number of communities is fixed to 7 on the Cora dataset, 6 on the Citeseer dataset and 3 on the PubMed Diabetes dataset according to the ground truth. For the methods that can automatically decide the number of communities including *Big-CLAM*, *CNM*, *CESNA* and *Circle*, we will compare the results given the predefined number with the automatically chosen number of communities, and report the best performance. For the Facebook dataset with irregular community properties, we choose from a set of numbers of communities for all the methods.

The networks included in the Cora, Citeseer and Pubmed Diabetes datasets are directed. In our framework we process the networks by considering both directions of the edges to cover different kinds of relations. Meanwhile, some baselines are evaluated on both the undirected and directed graphs for Cora and Citeseer datasets with the best performance chosen in comparison, without showing significant impact on the performance nevertheless.

Given the experimental setup described above, we summarize the results of various methods in Table I. Among the 3 metrics, F-score mainly depicts the "accuracy" of detected communities, while Jaccard Similarity also concerns about the "recovery" of communities given the ground truth, and NMI offers an entropy measure of the overall matching quality of the detected communities and the ground truth.

As demonstrated in Table I, the 4 methods in our framework of content propagation, shown in the bottom of Table I, outperform the baselines significantly on the 4 datasets across different evaluation metrics, except for performing closely to the best result regarding Jaccard Similarity on the PubMed Diabetes dataset, considering the fact that the proposed al-

[2] http://snap.stanford.edu/index.html

[3] http://gibbslda.sourceforge.net/

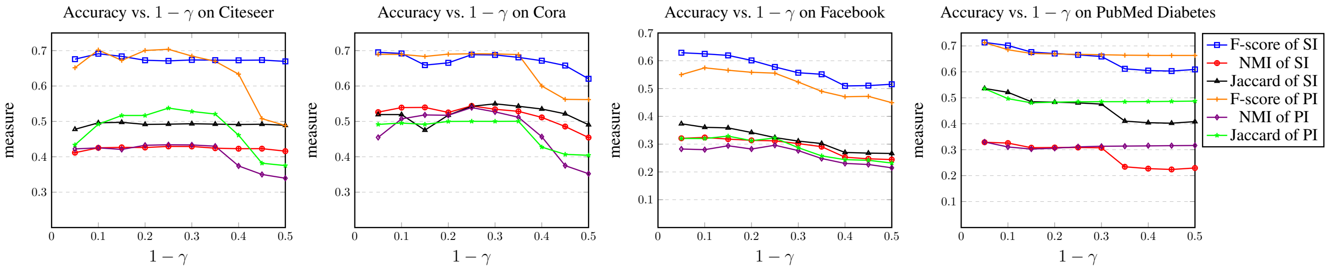


Fig. 2: Performance of CPIP with different values of $1-\gamma$

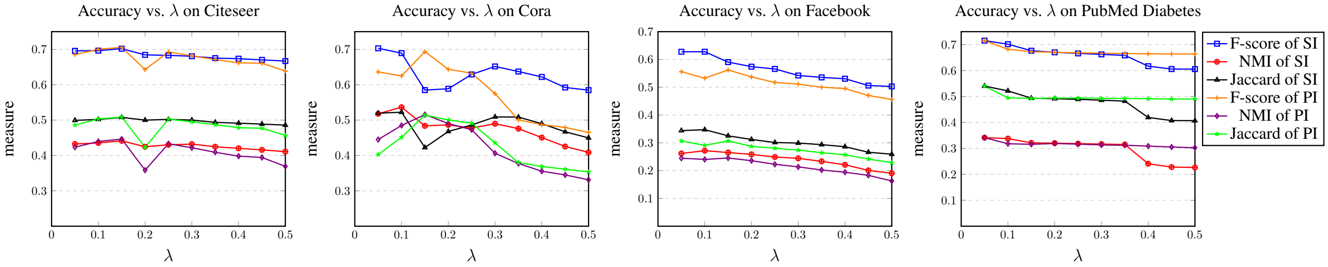
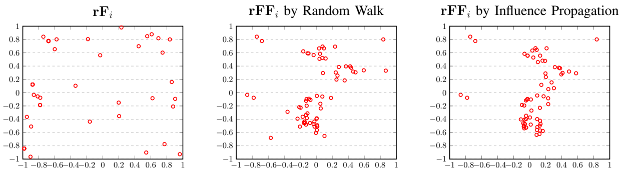


Fig. 3: Performance of CPRW with different values of λ



(a) Origin Feature (b) CPRW (c) CPIP

Fig. 4: Effect of Content Propagation: the original feature vectors $\mathbf{rF}_i$ and the propagated content $\mathbf{rFF}_i$

gorithms use a fixed parameter on all the datasets, while the best parameters are chosen for the comparison methods. This validates that our proposed framework describes the interactions between nodes better, and therefore captures the nature of communities more accurately.

From the comparison, it is shown that methods using both link and content yield better performance than approaches using a single source of information in general, which justifies the motivation of combining both structure and content in community detection. One could also observe that the overlapping community detection methods such as *Big-CLAM*, *CESNA* and *Circle* suffer from relatively low values of Jaccard Similarity and NMI while achieving higher F-Scores. This implies that the communities detected by the overlapping algorithms do not recover the ground truth very well, probably due to the reason that they may group some nodes into a community which is not aligned with the ground truth.

E. Effect of Content Propagation

In this subsection, we investigate the effect of content propagation. We conduct the experiment on the smallest network in the Facebook dataset, which consists of 66 nodes while each node is described by 48 features. We first project the feature vectors of all the nodes to 2d space with PCA, and the reduced feature vectors are denoted as $\mathbf{rF}_i, \forall i \in [1, N]$. Then we calculate the content propagation received by each node, denoted as $\mathbf{rFF}_i, \forall i \in [1, N]$. The 2d feature vectors $\mathbf{rF}_i$, as well as the received content propagation $\mathbf{rFF}_i$, calculated by random walk or influence propagation, are plotted in Fig. 4.

We can observe that content propagation indeed enhances the cohesiveness of nodes by interacting with each other, which verifies our motivation of using content propagation to facilitate nodes with attributes, which are initially diverse, to converge to communities.

F. Effect of Varying $1-\gamma$ or λ

In the experiments conducted above we fix the value of $1-\gamma$ or λ to 0.1. In this subsection, we investigate the influence of the parameter $1-\gamma$ or λ on the performance of the proposed content propagation based methods. The value of $1-\gamma$ is varied from 0.05 to 0.5 with step size equal to 0.05 and we plot the performance of CPIP and CPRW. Results are summarized in Fig. 2 and Fig. 3. We can observe that our methods achieve better performance at smaller λ or $1-\gamma$ values, and the performance decreases slowly in general when they get larger. This is probably due to the fact that a small value of λ or $1-\gamma$ may result in a longer-term propagation process, which helps improve the quality of communities discovered.

G. Effect of Varying K

When the number of communities is given in the ground truth or could be readily estimated, we can set the K value correspondingly. On the other hand, there are also scenarios where the number of communities is hard to determine. Here, we conduct a brief investigation on how the quality of discovered communities is impacted by different values of K.

In Fig. 5, we demonstrate the experimental results on Cora as we vary the value of K from 4 to 10 after setting the value of λ or $1-\gamma$ to 0.1. We observe relatively stable performance with different K values, and the best performance is achieved when $K = 6$ for CPRW-PI, CPRW-SI, CPIP-SI, and $K = 7$ for CPIP-PI; while according to the ground truth, $K = 7$. The discrepancy here is probably due to the fact that there exist

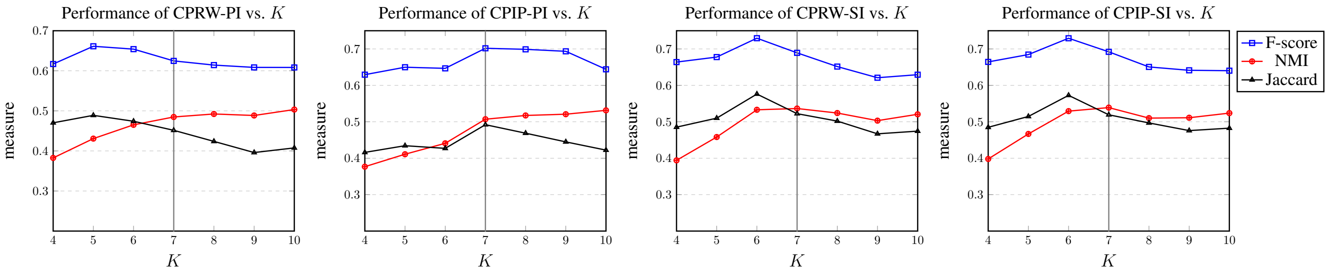
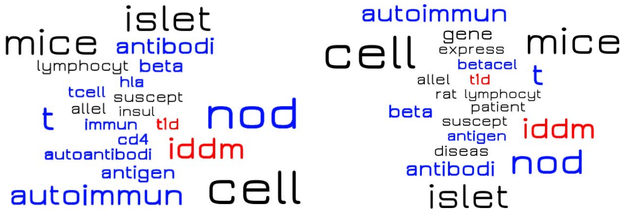


Fig. 5: Performance of our framework with different number of communities (K) on Cora

two communities in Cora with significant overlap. From Fig. 5 one can observe that the proposed methods achieve relatively better performance when K is close to the ground-truth value, and declines slowly as K moving away from it, however the influence is not very significant.

H. Analysis of the Content of Communities

Our proposed framework of content propagation calculates the content received by each community, which could be viewed as its description. In this subsection, we examine the semantic accordance between this description and the ground truth. We choose the PubMed Diabetes dataset in this investigation, because it is the only dataset that we can obtain the meaning of the attributes. There exist three ground truth communities in the PubMed Diabetes dataset: "Diabetes Mellitus, Experimental", "Diabetes Mellitus Type 1" and "Diabetes Mellitus Type 2".

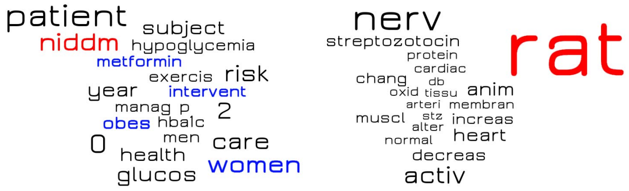


(a) Most Distinguishing (b) Most Related

Fig. 6: Most distinguishing and related attributes of the 2nd Community detected by CPRW-SI

Given the communities discovered with our framework, we identify the most related or positively distinguishing attributes of a community by its corresponding $\boldsymbol{\mu}$ vector calculated in (12). We try two different ways to identify them:

- For the i-th community, treat the first 20 attributes with largest values in $\boldsymbol{\mu}_i$ as the most related attributes.
- For the i-th community, calculate $\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}$ where $\overline{\boldsymbol{\mu}} = \frac{\sum_{j=1}^{K} \boldsymbol{\mu}_j}{K}$, and treat the first 20 attributes with largest values in $\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}}$ as the most positively distinguishing attributes.



(a) 1st Community (b) 3rd Community

Fig. 7: Most distinguishing attributes of the 1st and the 3rd communities detected by CPRW-SI

TABLE II: Pairwise F-score between ground-truth communities on PubMed Diabetes dataset and communities detected by CPRW-SI

F-Score	Experimental	Type One	Type Two
3rd Community	**0.7525**	0.0173	0.1003
2nd Community	0.3918	**0.5869**	0.0962
1st Community	0.0297	0.1620	**0.7517**

The most distinguishing and related attributes of the 2nd detected community are summarized in Fig. 6, demonstrating strong accordance with Type 1 Diabetes. The size of an attribute indicates the degree of being distinguishing or related. We notice that attributes t1d and iddm (marked in red) show up in both figures, and both of them indicate the same disease as Diabetes Mellitus Type 1. Besides, most of the other attributes are also related to Diabetes or Diabetes Mellitus Type 1, and we mark part of the highly distinguishing attributes of Type 1 Diabetes in blue. We also observe that the positively distinguishing attributes identified in the second way are indeed more distinguishing than the related ones.

Next, we summarize the most distinguishing attributes of the other communities in Fig. 7, from which we can observe the abbreviation of NonInsulin-Dependent Diabetes Mellitus (Diabetes Mellitus Type 2), niddm (marked in red), is among the distinguishing attributes in the 1st Community. Another interesting observation is that women appears to be a more distinguishing attribute than men, which agrees with the fact that women seem to be at a greater risk of Diabetes Mellitus Type 2[4]. As to the 3rd Community, the big attribute rat (marked in red) implies that this group may be related to experiments regarding Diabetes.

We also report the pairwise F-score between the ground-truth communities and the communities discovered by our framework in Table II. It demonstrates that the 1st community discovered with niddm (Diabetes Mellitus Type 2) among the most distinguishing attributes is really similar to the ground-truth community named "Diabetes Mellitus Type 2". Similar observation can be made on the 2nd and 3rd communities, from which we can conclude that our detected communities align with the ground truth well, both in structure and in semantic meaning.

[4] http://en.wikipedia.org/wiki/Diabetes_mellitus_type_2

VI. Conclusion

In this paper, we propose a framework of community detection which combines the network structure and node attributes from the perspective of *content propagation*. It treats a network as a dynamic system and considers its community structure as a consequence of interactions among nodes. We model the interactions with two principles—influence propagation and random walk. Experimental evaluation fully justifies the effectiveness of the proposed framework from multiple aspects, including accuracy, semantic meaningfulness, etc. An interesting direction to pursue in future work would be augmenting the description of the propagation process by incorporating heterogenous information on networks.

Acknowledgment

Research supported by the National Natural Science Foundation of China (No. 61375060) and the National Science Foundation for Distinguished Young Scholars of China (No. 61325010).

References

[1] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[2] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[3] D. J. Watts, P. S. Dodds, and M. E. Newman, "Identity and search in social networks," *science*, vol. 296, no. 5571, pp. 1302–1305, 2002.

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[5] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[6] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '08, 2008, pp. 567–580.

[7] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 927–936.

[8] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.

[9] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proceedings of the 22nd international conference on world wide web*, 2013, pp. 1089–1098.

[10] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proceedings of 2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 1151–1156.

[11] D. A. Cohn and T. Hofmann, "The missing link - a probabilistic model of document content and hypertext connectivity," in *Neural Information Processing Systems*, 2001, pp. 430–436.

[12] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 542–550.

[13] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 718–729, 2009.

[14] N. Barbieri, F. Bonchi, and G. Manco, "Influence-based network-oblivious community detection," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 955–960.

[15] Y. Zhou and L. Liu, "Social influence based clustering of heterogeneous information networks," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 338–346.

[16] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[17] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, pp. 1420–1443, 1978.

[18] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1029–1038.

[19] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[20] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. A. Shad, "On approximation of real-world influence spread," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 548–564.

[21] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, "Pagerank with priors: An influence propagation perspective," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 2740–2746.

[22] M. S. T. Jaakkola and M. Szummer, "Partially labeled classification with markov random walks," *Advances in neural information processing systems (NIPS)*, vol. 14, pp. 945–952, 2002.

[23] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, "Using ghost edges for classification in sparsely labeled networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 256–264.

[24] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Velocity and hierarchical spread of epidemic outbreaks in scale-free networks," *Physical Review Letters*, vol. 92, no. 17, p. 178701, 2004.

[25] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

[26] L. Xu, M. White, and D. Schuurmans, "Optimal reverse prediction: A unified perspective on supervised, unsupervised and semi-supervised learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1137–1144.

[27] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Magazine*, vol. 29, no. 3, pp. 93–106, 2008.

[28] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 587–596.

[29] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[30] X.-H. Phan and C.-T. Nguyen, "Gibbslda++: Ac/c++ implementation of latent dirichlet allocation (lda)," 2007.

[31] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[32] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.