Community Detection for Emerging Networks

Jiawei Zhang* Philip S. Yu†

Abstract

Nowadays, many new social networks offering specific services spring up overnight. In this paper, we want to detect communities for emerging networks. Community detection for emerging networks is very challenging as information in emerging networks is usually too sparse for traditional methods to calculate effective closeness scores among users and achieve good community detection results. Meanwhile, users nowadays usually join multiple social networks simultaneously, some of which are developed and can share common information with the emerging networks. Based on both link and attribution information across multiple networks, a new general closeness measure, *intimacy*, is introduced in this paper. With both micro and macro controls, an effective and efficient method, CAD (Cold stArt community Detector), is proposed to propagate information from developed network to calculate effective intimacy scores among users in emerging networks. Extensive experiments conducted on real-world social networks demonstrate that CAD can perform very well in addressing the emerging network community detection problem.

1 Introduction

Clusters in networks are defined as groups of nodes which are strongly connected in the group but loosely connected to nodes in other groups. Depending on specific disciplines, networks studied in clustering problems can be very diverse, which include online social networks, e.g., Twitter and Facebook [18]; e-commerce networks, e.g., Amazon and Epinions [6]; and bibliographic networks, e.g., DBLP [15]. Meanwhile, discovering clusters of user in social networks is also formally defined as the *community detection* problem [10, 18, 17, 7, 15]. Community detection is very important for online social networks as it is a crucial prerequisite for many concrete social services, e.g., better organization of users' friends in online social networks by partitioning them into "schoolmates", "family", "celebrities", etc.

Nowadays, witnessing the incredible success of popular online social networks, e.g., Facebook and Twitter, a large number of new social networks offering specific services also spring up overnight. Generally, new emerging networks are networks containing very sparse information, which can be (1) networks starting to provide social services for only a very short period of time; or (2) even more mature ones that start to branch into new geographic areas or social groups [20]. The formal definitions of "*emerging networks*" and "*developed networks*" are available in Section 2. Considering its wide applications in various social services, community detection is important for emerging networks as high-quality community detection results enable emerging networks to provide better services, which will help attract more user registration effectively.

Problem: In this paper, we study the community detection problem for emerging networks, which is formally defined as the "*emerging network community detection*" problem. Furthermore, when the network is brand new (i.e., little information about the registered users exists in the network), the problem will be the "*cold start community detection*" problem. Few works have studied the *cold start problem* in community detection and we are the first to propose the concepts of "*emerging network community detection*" and "*cold start community detection*".

Community detection for emerging networks is a new problem and conventional community detection methods for developed networks cannot be applied. Compared with developed networks, information in emerging networks is too sparse to support traditional methods in calculating effective closeness scores and achieving good results. More information about related problems is available in Section 5.

Meanwhile, as proposed in [5, 19, 20, 21], users nowadays usually participate in multiple social networks simultaneously to enjoy more social services. Users who are involved in an emerging network may have been using other developed social networks for a long time. Furthermore, some of the developed networks can share common information with emerging networks either due to the network establishing purpose, e.g., Google Scholar (released in 2004) and Research Gate (launched in 2008) are both constructed for better academic communications, or because of specific network features, e.g., Twitter (created in 2006) and Foursquare

*University of Illinois at Chicago, IL, USA. jzhan9@uic.edu

†University of Illinois at Chicago, IL, USA. Institute for Data Science, Tsinghua University, Beijing, China. psyu@cs.uic.edu

(launched in 2009) can both offer geo-spatial services and allow users to follow other users. If the useful information in developed networks can be propagated to emerging networks, the information sparsity problem encountered in detecting communities for emerging networks can be solved promisingly.

Despite its importance and novelty, the "*emerging network community detection*" problem is very challenging to solve due to the following reasons:

- *effective closeness measure*: Effective definition and calculation of closeness measure which can capture the connections among users in various aspects is a prerequisite for effective community detection. The problem is more urgent in the *emerging network community detection* problem due to the information sparsity problem in emerging networks.

- *information weight control*: Users have both link and attribute information (i.e., multiple information types) in both emerging network and developed networks (i.e., multiple information sources). How to determine the weights of different information types and information sources in closeness score calculation is very challenging.

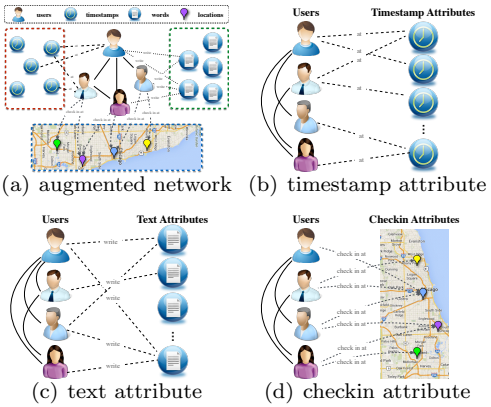
- *high time and space cost*: Community detection across multiple networks can involve too many nodes and connections, which will lead to high time and space cost.

To solve all the above challenges, a novel community detection method, CAD, is proposed in this paper: (1) CAD introduces a general closeness measure, *intimacy*, based on both link and attribute information in (and across) heterogeneous networks; (2) CAD can propagate useful information across developed and emerging networks to solve the shortage of information problem with both *micro* and *macro* information weight controls, whose parameters can be adjusted automatically; (3) effective and efficient techniques are proposed to help CAD overcome the high time and space cost problem.

This paper is organized as follows. We formulate the problem in Section 2. Detailed description of the methods is introduced in Section 3. In Section 4, we show the experiment results. Finally, we give the related works and conclusions in Sections 5 and 6.

2 Problem Formulation

In this paper, we will use the definitions of *anchor user*, *anchor link*, *aligned networks* proposed in [5, 19, 20, 21]. Definitions of other important terminologies and the formulation of the emerging network community detection problem will be introduced in this section.



(a) augmented network (b) timestamp attribute (c) text attribute (d) checkin attribute

Figure 1: An example of attribute augmented heterogeneous network. (a): attribute augmented heterogeneous network, (b-d): timestamp, text and location attributes.

2.1 Terminology Definition

Definition 2.1. *(Intimacy): Users in social networks can be correlated with each other closely and the correlation is quantified as the "intimacy" in this paper. Intimacy is a general closeness measure and can be applied to various networks, e.g., networks with link information only, networks with both link and attribute information as well as multiple aligned heterogeneous networks. The intimacy between user $u_i, u_j \in \mathcal{V}$ denotes the transition probability from u_i to u_j in the network.*

Definition 2.2. *(Intimacy Matrix): Matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is defined as the intimacy matrix among users in $\mathcal{V}$, where $H(i,j)$ is the intimacy between u_i and u_j.*

Definition 2.3. *(Attribute Augmented Heterogeneous Networks): Users can have both link and attribute information in social networks, which can be formulated as attribute augmented heterogeneous networks, $G = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V}$ and $\mathcal{E}$ are the user set and link set respectively. $\mathcal{A} = \{a_1, a_2, \cdots, a_m\}$ is the set of m different attributes that users have in the network and $a_i \in \mathcal{A}$ can have n_i different values.*

Including the attributes as nodes provides a conceptual framework to handle social links and node attributes in a unified framework. For example, in the social network shown in Figure 1(a), we can get not only user social link information, but also their active time, posting content and check-in locations and each of them can take on a set of values. By creating an augmented network, we can make the posting times, content keywords, and locations as augmented network nodes as shown in Figures 1(b)-1(d). The effect on increasing the dimensionality of the network will be handled in Lemma 3.1 in lower dimensional space.

DEFINITION 2.4. *(Aligned Attribute Augmented Heterogeneous Networks): Multiple aligned attribute augmented heterogeneous networks can be defined as* $\mathcal{G} = ((G^1, G^2, \cdots, G^n), (A^{1,2}, A^{1,3}, \cdots, A^{1,n}, A^{2,3}, \cdots, A^{(n-1),n}))$*, where* $G^i, i \in \{1, 2, \cdots, n\}$ *is an attributed augmented heterogeneous network and* $A^{i,j}$, $i \neq j, i, j \in \{1, 2, \cdots, n\}$ *is the set of undirected anchor links [21, 5] between* G^i *and* G^j.

DEFINITION 2.5. *(Average Degree) The average degree of a network denotes the average number of edges connected to each node in the network, i.e., connection density. The average degree of network* G *can be represented as* $AD(G) = \frac{|\mathcal{E}|}{|\mathcal{V}|}$.

DEFINITION 2.6. *(Emerging and Developed Networks): Concepts "emerging" and "developed" can depict the sparsity of information in networks. In this paper, emerging networks (or developed networks) are defined as networks whose average degree is lower than threshold* ϵ_{new} *(or larger than threshold* ϵ_{dev}*). In other words, network* G *is an emerging network iff* $AD(G) < \epsilon_{new}$ *and* G *is a developed network iff* $AD(G) > \epsilon_{dev}$.

2.2 Emerging Network Community Detection

In this paper, we will study the emerging network community detection problem based on two real-world *partially aligned networks*: Foursquare and Twitter, whose detailed information is available in Section 4. According to the given definitions, networks studied in this paper can be formulated as two partially aligned attribute augmented heterogeneous networks: $\mathcal{G} = ((G^t, G^s), (A^{t,s}))$, where G^t and G^s are the emerging and developed networks respectively and $A^{t,s}$ is the set of anchor links between G^t and G^s. Both G^t and G^s can be formulated as the attribute augmented heterogeneous network, e.g., $G^t = (\mathcal{V}^t, \mathcal{E}^t, \mathcal{A}^t)$. With information across $\mathcal{G}$, we can calculate the *intimacy matrix*, $\mathbf{H}$, among users in the emerging network G^t that we target on. *Emerging network community detection* problem aims at partitioning user set $\mathcal{V}^t$ of G^t into K disjoint clusters, $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$, based on the *intimacy matrix*, $\mathbf{H}$, such that users in each cluster are more similar to each other than those in different clusters, where $\bigcup_i^K C_i = \mathcal{V}^t$ and $C_i \cap C_j = \emptyset, \forall i, j \in \{1, 2, \cdots, K\}, i \neq j$. When the target network G^t is brand new, i.e., $\mathcal{E}^t = \emptyset$ and $\mathcal{A}^t = \emptyset$, the problem will be the *cold start community detection* problem.

3 Proposed Methods

We will introduce the emerging network community detection method, CAD, in this section. In Section 3.1, we first introduce the concept of intimacy based on the simple case with social links only and then extend the intimacy matrix to capture attribute similarity in Section 3.2. Generalization of the intimacy matrix to cover cross network information propagation is addressed in Section 3.3. The approximation to solve the high space and time cost is introduced in Section 3.4.

3.1 Intimacy Matrix of Homogeneous Network

For a given homogeneous network, e.g., $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$, $\mathcal{E}$ are the user set and social link set in the network respectively, we can define the adjacency matrix of G to be $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where $Z(i,j) = 1$, if $(u_i, u_j) \in \mathcal{E}$. Meanwhile, via the social links in $\mathcal{E}$, information can propagate within the network, whose propagation paths can reflect the closeness among users [12]. Formally, we define $p_{ij} = \frac{Z(i,j)}{\sqrt{\sum_m Z(j,m) \sum_n Z(i,n)}}$ to be the information *transition probability* from u_j to u_i. Let's assume that user $u_i \in \mathcal{V}$ injects a stimulation into network G initially and the information will be propagated to other users in G afterwards. During the propagation process, users receive stimulation from their neighbors and the amount is proportional to the amount difference of stimulation reaching the user and his neighbors. Let vector $\boldsymbol{f}^{(\tau)} \in \mathbb{R}^{|\mathcal{V}|}$ denote the states of all users in $\mathcal{V}$ at τ, i.e., the proportion of stimulation at users in $\mathcal{V}$ at time τ. The change of stimulation at u_i at time $\tau + \Delta t$ is defined as follows:

$$\frac{f^{(\tau+\Delta t)}(i) - f^{(\tau)}(i)}{\Delta t} = \alpha \sum_{u_j \in \mathcal{V}} p_{ji}(f^{(\tau)}(j) - f^{(\tau)}(i)),$$

where coefficient α can be set as 1 as proposed in [23].

The *transition probabilities* $p_{ij}, i, j \in \{1, 2, \cdots, |\mathcal{V}|\}$ can be represented with the *transition matrix* $\mathbf{X} = (\mathbf{D}^{-\frac{1}{2}} \mathbf{Z} \mathbf{D}^{-\frac{1}{2}})$ of network G, where $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, $X(i,j) = p_{ij}$ and diagonal matrix $\mathbf{D} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ has value $D(i,i) = \sum_{j=1}^{|\mathcal{V}|} Z(i,j)$ on its diagonal.

DEFINITION 3.1. *(Social Transition Probability Matrix): The social transition probability matrix of network* G *can be represented as* $\mathbf{Q} = \mathbf{X} - \mathbf{D}_{\mathbf{X}}$*, where* $\mathbf{X}$ *is the transition matrix defined above and diagonal matrix* $D_{\mathbf{X}}$ *has value* $D_{\mathbf{X}}(i,i) = \sum_{j=1}^{|\mathcal{V}|} \mathbf{X}(j,i)$ *on its diagonal.*

Furthermore, by setting $\Delta t = 1$, denoting that stimulation propagates step by step through network, we can rewrite the propagation updating equation as:

$$\begin{aligned}\boldsymbol{f}^{(\tau)} &= \boldsymbol{f}^{(\tau-1)} + \alpha(\mathbf{X} - \mathbf{D}_{\mathbf{X}})^T \boldsymbol{f}^{(\tau-1)} = (\mathbf{I} + \alpha \mathbf{Q}^T) f^{(\tau-1)} \\ &= (\mathbf{I} + \alpha \mathbf{Q}^T)^\tau \boldsymbol{f}^{(0)} = (\mathbf{I} + \alpha \mathbf{Q})^\tau \boldsymbol{f}^{(0)},\end{aligned}$$

where $\mathbf{Q}$ is symmetric and $\mathbf{Q}^T = \mathbf{Q}$. Other *transition probability matrices* in the following parts of this paper

are all *symmetric* and we will use $(\mathbf{I}+\alpha\mathbf{Q})$ to denote $(\mathbf{I}+\alpha\mathbf{Q}^T)$ for simplicity.

The propagation process will stop when $\boldsymbol{f}^{(\tau)} = \boldsymbol{f}^{(\tau-1)}$ and the stationary transition matrix can be represented as $(\mathbf{I}+\alpha\mathbf{Q})^{(\tau)}$, where that smallest τ that can stop the propagation (i.e., $(\mathbf{I}+\alpha\mathbf{Q})^{(\tau)} = (\mathbf{I}+\alpha\mathbf{Q})^{(\tau-1)}$) is defined as the *stop step*. To obtain the *stop step* τ, we need to keep checking the powers of $(\mathbf{I}+\alpha\mathbf{Q})$ until it doesn't change as τ increases, i.e., the *stop criteria*.

DEFINITION 3.2. *(Homogeneous Network Intimacy Matrix): Matrix $\mathbf{H} = (\mathbf{I}+\alpha\mathbf{Q})^{\tau} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ is defined as the intimacy matrix of users in homogeneous network G, where τ is the stop step and $H(i,j)$ denotes the intimacy score between u_i and u_j in the network.*

3.2 Intimacy Matrix of Attribute Augmented Heterogeneous Network and Micro Control

Real-world social networks can usually contain various kinds of information, e.g., links and attributes. Besides among users, information can also propagate among users via shared attributes in heterogeneous networks.

DEFINITION 3.3. *(Attribute Transition Probability Matrix): The connections between users and attributes, e.g., a_i, can be represented as the attribute adjacency matrix $\mathbf{A}_{a_i} \in \mathbb{R}^{|\mathcal{V}|\times n_i}$, where n_i is the number of values that attribute a_i can have. Similar to the social transition probability matrix, based on $\mathbf{A}_{a_i}$, we formally define the attribute transition probability matrix from users to attribute a_i to be $\mathbf{R}_i \in \mathbb{R}^{|\mathcal{V}|\times n_i}$ and that from attribute a_i to users in $\mathcal{V}$ to be $\mathbf{S}_i = \mathbf{R}_i^T$.*

The importance of different information types in calculating the closeness measure among users can be different. To handle such a problem, we introduce the *micro* control by giving different information types distinct weights to denote their differences: $\omega = \{\omega_0, \omega_1, \cdots, \omega_m\}$, where $\sum_{i=0}^{m}\omega_i = 1.0$, ω_0 is the weight of link and ω_i is the weight of attribute a_i, $i \in \{1, 2, \cdots, m\}$.

DEFINITION 3.4. *(Weighted Attribute Transition Probability Matrix): With weights ω, we define $\tilde{\mathbf{R}} = [\omega_1\mathbf{R}_1, \cdots, \omega_n\mathbf{R}_n]$, $\tilde{\mathbf{S}} = [\omega_1\mathbf{S}_1^T, \cdots, \omega_n\mathbf{S}_n^T]^T$ to be the weighted attribute transition probability matrices between users and all attributes, where $\tilde{\mathbf{R}} \in \mathbb{R}^{|\mathcal{V}|\times(n_{aug}-|\mathcal{V}|)}$, $\tilde{\mathbf{S}} \in \mathbb{R}^{(n_{aug}-|\mathcal{V}|)\times|\mathcal{V}|}$, $n_{aug} = (|\mathcal{V}| + \sum_{i=1}^{m} n_i)$ is the number of all nodes in the attribute augmented heterogeneous network.*

DEFINITION 3.5. *(Network Transition Probability Matrix): Furthermore, the transition probability matrix of the whole attribute augmented heterogeneous network G is defined as $\tilde{\mathbf{Q}}_{aug} = \begin{bmatrix}\tilde{\mathbf{Q}} & \tilde{\mathbf{R}}\\ \tilde{\mathbf{S}} & \mathbf{0}\end{bmatrix}$, where $\tilde{\mathbf{Q}}_{aug}^T = \tilde{\mathbf{Q}}_{aug} \in \mathbb{R}^{n_{aug}\times n_{aug}}$ and $\tilde{\mathbf{Q}} = \omega_0\mathbf{Q}$ is the weighted social transition probability matrix of the network.*

In the real world, heterogeneous social networks can contain large amounts of attributes, i.e., n_{aug} is extremely large. The *weighted transition probability matrix*, i.e., $\tilde{\mathbf{Q}}_{aug}$, will be of extremely high dimensions and can hardly fit in the memory. As a result, it will be impossible to update the matrix until the *stop criteria* meets to obtain the *stop step* and the *intimacy matrix*. To solve such problem, we propose to obtain the *stop step* and the *intimacy matrix* by applying partitioned block matrix operations with the following Lemma 3.1.

LEMMA 3.1. $(\tilde{\mathbf{Q}}_{aug})^k = \begin{bmatrix}\tilde{\mathbf{Q}}_k & \tilde{\mathbf{Q}}_{k-1}\tilde{\mathbf{R}}\\ \tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-1} & \tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-2}\tilde{\mathbf{R}}\end{bmatrix}$, $k \geq 2$, *where* $\tilde{\mathbf{Q}}_k = \begin{cases}\mathbf{I}, & \text{if } k=0,\\ \tilde{\mathbf{Q}}, & \text{if } k=1,\\ \tilde{\mathbf{Q}}\tilde{\mathbf{Q}}_{k-1} + \tilde{\mathbf{R}}\tilde{\mathbf{S}}\tilde{\mathbf{Q}}_{k-2}, & \text{if } k\geq 2\end{cases}$, $\tilde{\mathbf{Q}}_k \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ *and the heterogeneous network intimacy matrix is defined as*

$$\begin{aligned}\tilde{\mathbf{H}}_{aug} &= \left(\mathbf{I}+\alpha\tilde{\mathbf{Q}}_{aug}\right)^{\tau}(1:|\mathcal{V}|, 1:|\mathcal{V}|)\\ &= \left(\sum_{t=0}^{\tau}\binom{\tau}{t}\alpha^t(\tilde{\mathbf{Q}}_{aug})^t\right)(1:|\mathcal{V}|, 1:|\mathcal{V}|)\\ &= \left(\sum_{t=0}^{\tau}\binom{\tau}{t}\alpha^t\left((\tilde{\mathbf{Q}}_{aug})^t(1:|\mathcal{V}|, 1:|\mathcal{V}|)\right)\right)\\ &= \left(\sum_{t=0}^{\tau}\binom{\tau}{t}\alpha^t\tilde{\mathbf{Q}}_t\right),\end{aligned}$$

where $\mathbf{X}(1:|\mathcal{V}|, 1:|\mathcal{V}|)$ is a sub-matrix of $\mathbf{X}$ with indexes in $[1, |\mathcal{V}|]$, τ is the stop step, achieved when $\tilde{\mathbf{Q}}_\tau = \tilde{\mathbf{Q}}_{\tau-1}$, i.e., the stop criteria, $\tilde{\mathbf{Q}}_\tau$ is called the stationary matrix.

Proof. The lemma can be proved by induction on k [22]. Considering that $(\tilde{\mathbf{R}}\tilde{\mathbf{S}}) \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ can be precomputed in advance, the space cost will be $O(|\mathcal{V}|^2)$, $|\mathcal{V}| \ll n_{aug}$.

Since we are only interested in the *intimacy* and *transition matrices* among users, not those between the augmented attributes and users, we create a reduced dimensional representation only involving users for $\tilde{\mathbf{Q}}_k$ and $\tilde{\mathbf{H}}$ such that we can capture the effect of "user-attribute" and "attribute-user" transitions on "user-user" transitions. $\tilde{\mathbf{Q}}_k$ is a reduced dimensional representation of $\tilde{\mathbf{Q}}_{aug}^k$, while eliminating the augmented items, it still maintains the "user-user" transitions effectively.

3.3 Intimacy Matrix across Aligned Attribute Augmented Heterogeneous Networks and Macro Control

When G^t is new, the *intimacy matrix* $\tilde{\mathbf{H}}$

among users calculated based on the information in G^t will be very sparse. Meanwhile, useful information propagated from other aligned developed networks can help solve the shortage of information problem in the emerging network [19, 20]. However, as proposed in [11], information propagated from the developed networks can be different from that in emerging networks. To handle this problem, we propose to apply the *macro control* technique by using weights, $\rho^{s,t}, \rho^{t,s} \in [0,1]$, to control the proportion of information propagated between G^s and G^t. If information from G^s is helpful for improving the community detection results in G^t, we can set a higher $\rho^{s,t}$ to propagate more information from G^s. Otherwise, we can set a lower $\rho^{s,t}$ instead.

DEFINITION 3.6. *(Anchor Transition Matrix): To propagate information across networks, we define the anchor transition matrices between G^t and G^s to be $\mathbf{T}^{t,s} \in \mathbb{R}^{|\mathcal{V}^t|\times|\mathcal{V}^s|}$ and $\mathbf{T}^{s,t} \in \mathbb{R}^{|\mathcal{V}^s|\times|\mathcal{V}^t|}$, where $\mathbf{T}^{t,s}(i,j) = \mathbf{T}^{s,t}(j,i) = 1$, if $(u_i^t, u_j^s) \in A^{t,s}, u_i^t \in \mathcal{V}^t, u_j^s \in \mathcal{V}^s$.*

DEFINITION 3.7. *(Weighted Network Transition Matrix): Meanwhile, with weights $\rho^{s,t}$ and $\rho^{t,s}$, we define the weighted network transition probability matrix of G^t and G^s to be $\bar{\mathbf{Q}}_{aug}^t = (1-\rho^{t,s})\begin{bmatrix}\tilde{\mathbf{Q}}^t & \tilde{\mathbf{R}}^t \\ \tilde{\mathbf{S}}^t & \mathbf{0}\end{bmatrix}$ and $\bar{\mathbf{Q}}_{aug}^s = (1-\rho^{s,t})\begin{bmatrix}\tilde{\mathbf{Q}}^s & \tilde{\mathbf{R}}^s \\ \tilde{\mathbf{S}}^s & \mathbf{0}\end{bmatrix}$, where $\bar{\mathbf{Q}}_{aug}^t \in \mathbb{R}^{n_{aug}^t \times n_{aug}^t}$ and $\bar{\mathbf{Q}}_{aug}^s \in \mathbb{R}^{n_{aug}^s \times n_{aug}^s}$, n_{aug}^t and n_{aug}^s are the numbers of all nodes in G^t and G^s respectively.*

DEFINITION 3.8. *(Weighted Anchor Transition Matrix): Furthermore, to accommodate the dimensions, we define the weighted anchor transition matrices between G^s and G^t to be $\bar{\mathbf{T}}^{t,s} = (\rho^{t,s})\begin{bmatrix}\mathbf{T}^{t,s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}$, $\bar{\mathbf{T}}^{s,t} = (\rho^{s,t})\begin{bmatrix}\mathbf{T}^{s,t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}$, where $\bar{\mathbf{T}}^{t,s} \in \mathbb{R}^{n_{aug}^t \times n_{aug}^s}$ and $\bar{\mathbf{T}}^{s,t} \in \mathbb{R}^{n_{aug}^s \times n_{aug}^t}$. Nodes corresponding to entries in $\bar{\mathbf{T}}^{t,s}$ and $\bar{\mathbf{T}}^{s,t}$ are of the same order as those in $\bar{\mathbf{Q}}_{aug}^t$ and $\bar{\mathbf{Q}}_{aug}^s$ respectively.*

DEFINITION 3.9. *(Aligned Network Transition Matrix): The transition probability matrix across aligned networks is defined as $\bar{\mathbf{Q}}_{align} = \begin{bmatrix}\bar{\mathbf{Q}}_{aug}^t & \bar{\mathbf{T}}^{t,s} \\ \bar{\mathbf{T}}^{s,t} & \bar{\mathbf{Q}}_{aug}^s\end{bmatrix}$, where $\bar{\mathbf{Q}}_{align}^T = \bar{\mathbf{Q}}_{align} \in \mathbb{R}^{n_{align}\times n_{align}}$, $n_{align} = n_{aug}^t + n_{aug}^s$ is the number of all nodes across the aligned networks.*

DEFINITION 3.10. *(Aligned Network Intimacy Matrix): According to Definition 7, with $\bar{\mathbf{Q}}_{align}$, we can obtain the the intimacy matrix, $\bar{\mathbf{H}}_{align}$, of users in G^t to be*

$$\bar{\mathbf{H}}_{align} = (\mathbf{I} + \alpha\bar{\mathbf{Q}}_{align})^\tau (1:|\mathcal{V}^t|, 1:|\mathcal{V}^t|),$$

where $\bar{\mathbf{H}}_{align} \in \mathbb{R}^{|\mathcal{V}^t|\times|\mathcal{V}^t|}$, τ is the stop step.

Meanwhile, methods introduced in Lemma 3.1 doesn't work well with $\bar{\mathbf{Q}}_{align}$ as the non-zero square matrix at the upper left corner of $\bar{\mathbf{Q}}_{align}$ is still of high dimension. To obtain the *stop step*, we have no choice but to keep calculating powers of $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})$ until the *stop criteria* can meet, which can be very time consuming. In this part, we propose to solve the problem with the following Lemma 3.2.

LEMMA 3.2. *For the given matrix $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})$, its k_{th} power meets $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})^k\mathbf{P} = \mathbf{P}\mathbf{\Lambda}^k, k \geq 1$, matrices $\mathbf{P}$ and $\mathbf{\Lambda}$ contain the eigenvector and eigenvalues of $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})$. The i_{th} column of matrix $\mathbf{P}$ is the eigenvector of $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})$ corresponding to its i_{th} eigenvalue λ_i and diagonal matrix $\mathbf{\Lambda}$ has value $\Lambda(i,i) = \lambda_i$ on its diagonal.*

Proof. The Lemma can be proved by induction on k [13]. The time cost of calculating $\mathbf{\Lambda}^k$ is $\mathrm{O}(n_{align})$, which is far less than that required to calculate $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})^k$.

DEFINITION 3.11. *(Eigen-decomposition based Aligned Network Intimacy Matrix): In addition, if $\mathbf{P}$ is invertible, we can have $(\mathbf{I}+\alpha\bar{\mathbf{Q}}_{align})^k = \mathbf{P}\mathbf{\Lambda}^k\mathbf{P}^{-1}$,, where $\mathbf{\Lambda}^k$ has $\Lambda(i,i)^k$ on its diagonal. And the intimacy calculated based on eigenvalue decomposition will be*

$$\bar{\mathbf{H}}_{align} = \left(\mathbf{P}\mathbf{\Lambda}^\tau\mathbf{P}^{-1}\right)(1:|\mathcal{V}^t|, 1:|\mathcal{V}^t|).$$

where the stop step τ can be obtained when $\mathbf{P}\mathbf{\Lambda}^\tau\mathbf{P}^{-1} = \mathbf{P}\mathbf{\Lambda}^{\tau-1}\mathbf{P}^{-1}$, i.e., stop criteria.

3.4 Approximated Intimacy to Reduce Dimension Eigendecomposition based method proposed in Lemma 3.2 enables us to calculate the powers of $(\mathbf{I}+\alpha\mathbf{Q}_{align})$ very efficiently. However, when applying Lemma 3.2 to calculate the *intimacy matrix* of real-world partially aligned networks, it can still suffer from the space problem. The reason is that the dimension of $(\mathbf{I}+\alpha\mathbf{Q}_{align})$, i.e., $n_{align}\times n_{align}$, is so high that matrix $(\mathbf{I}+\alpha\mathbf{Q}_{align})$ can hardly fit in the memory. To solve that problem, in this part, we propose to calculate the approximated *intimacy matrix* $\bar{\mathbf{H}}_{align}^{approx}$ with less space and time costs instead.

Let's define the transition probability matrices of G^t and G^s to be $\tilde{\mathbf{Q}}_{aug}^t$ and $\tilde{\mathbf{Q}}_{aug}^s$ respectively. By applying Lemma 3.1, we can get their *stop step* and the *stationary matrices* to be τ^t, τ^s, $\tilde{\mathbf{Q}}_{\tau^t}^t$ and $\tilde{\mathbf{Q}}_{\tau^s}^s$ respectively.

DEFINITION 3.12. *(Reduced Aligned Network Transition Matrix): Stationary matrices $\tilde{\mathbf{Q}}_{\tau^t}^t$, $\tilde{\mathbf{Q}}_{\tau^s}^s$ together with the anchor transition matrices, $\mathbf{T^{t,s}}$ and $\mathbf{T}^{t,s}$, can be used to define a low-dimensional reduced aligned network transition matrix, which only involves users explicitly, while the effect of "attribute-user" or "user-attribute" transition is implicitly absorbed into $\tilde{\mathbf{Q}}_{\tau^t}^t$ and*

$\tilde{\mathbf{Q}}^s_{\tau^s}$:

$$\bar{\mathbf{Q}}^{user}_{align} = \begin{bmatrix} (1-\rho^{t,s})\tilde{\mathbf{Q}}^t_{\tau^t} & (\rho^{t,s})\mathbf{T}^{t,s} \\ (\rho^{s,t})\mathbf{T}^{s,t} & (1-\rho^{s,t})\tilde{\mathbf{Q}}^s_{\tau^s} \end{bmatrix},$$

where $\bar{\mathbf{Q}}^{user}_{align} \in \mathbb{R}^{(|\mathcal{V}|^t+|\mathcal{V}^s|)^2}$ *and* $(|\mathcal{V}|^t + |\mathcal{V}^s|) \ll n_{align}$.

Definition 3.13. *(Approximated Aligned Network Intimacy Matrix): With Lemma 3.2, we can get intimacy matrix of users in* G^t *based on* $\bar{\mathbf{Q}}^{user}_{align}$ *to be:*

$$\bar{\mathbf{H}}^{approx}_{align} = \left(\mathbf{P}^*(\mathbf{\Lambda}^*)^\tau(\mathbf{P}^*)^{-1}\right)(1:|\mathcal{V}^t|, 1:|\mathcal{V}^t|),$$

where $(\mathbf{I} + \alpha\bar{\mathbf{Q}}^{user}_{align}) = \mathbf{P}^*\mathbf{\Lambda}^*(\mathbf{P}^*)^{-1}$, τ *is the stop step.*

3.5 Clustering and Weight Self-Adjustment Intimacy matrix $\bar{\mathbf{H}}_{align}$ (or $\bar{\mathbf{H}}^{approx}_{align}$) stores the intimacy scores among users in $\mathcal{V}^t$ and can be used to detect communities in the emerging network. In this paper, we will use the low-rank matrix factorization method proposed in [16] to get the latent feature vectors, $\mathbf{U}$:

$$\min_{\mathbf{U},\mathbf{V}} \left\|\bar{\mathbf{H}}_{align} - \mathbf{U}\mathbf{V}\mathbf{U}^T\right\|^2_F + \theta\left\|\mathbf{U}\right\|^2_F + \beta\left\|\mathbf{V}\right\|^2_F,$$
$$s.t., \mathbf{U} \geq \mathbf{0}, \mathbf{V} \geq \mathbf{0},$$

where $\mathbf{U}$ is the latent feature vectors, $\mathbf{V}$ stores the correlation among rows of $\mathbf{U}$, θ and β are the weights of $\left\|\mathbf{U}\right\|^2_F$, $\left\|\mathbf{V}\right\|^2_F$ respectively.

Detailed derivatives and solution to the above objective function is available in [16]. The latent feature vectors in $\mathbf{U}$ can be used to detect communities in some traditional clustering methods, e.g., Kmeans [3].

Meanwhile, to handle Challenge 2, we use weights, ω^t, ω^s, $\rho^{t,s}$ and $\rho^{s,t}$, to denote the importance of different information types and information sources respectively. For simplicity, we set $\omega^t = \omega^s = \omega = [\omega_0, \omega_1, \cdots, \omega_m]$ and $\rho^{t,s} = \rho^{s,t} = \rho$ in this paper.

Let $\mathcal{C}$ be the community detection result achieved by CAD in G^t. The optimal result $\mathcal{C}$, evaluated by some metrics, e.g., *entropy* $E(\mathcal{C})$ [23], can be achieved with the following equation:

$$\omega, \rho = \min_{\omega,\rho} E(\mathcal{C}).$$

The optimization problem is very difficult to solve. Next, we will propose a method to adjust ω and ρ automatically to enable CAD to achieve better results.

The weight adjustment method used to deal with ω can work as follows: for example, in network G^t, we have relational information and attribute information $\mathcal{E}$ and $\mathcal{A} = \{A_1, A_2, \cdots, A_m\}$, whose weights are initialized to be $\omega = \{\omega_0, \omega_1, \cdots, \omega_m\}$. For $\omega_i \in \omega, i \in \{0, 1, \cdots, m\}$, we keep checking if increasing ω_i by a ratio of γ, i.e., $(1+\gamma)\omega_i$, can improve the performance or not. If so, $(1+\gamma)\omega_i$ after re-normalization is used as the new value of ω_i; otherwise, we restore the old ω_i before increase and study ω_{i+1}. In the experiment, γ is set as 0.05. Similarly, for the weight of different networks, i.e, ρ, we can adjust them with the same methods to find the optimal ρ.

Table 1: Properties of the Heterogeneous Networks

		network	
	property	**Twitter**	**Foursquare**
# node	user	5,223	5,392
	tweet/tip	9,490,707	48,756
	location	297,182	38,921
# link	friend/follow	164,920	76,972
	write	9,490,707	48,756
	locate	615,515	48,756

4 Experiments

To demonstrate the effectiveness of CAD, in this section, we will conduct extensive experiments on two real-world aligned online heterogeneous networks: Foursquare and Twitter.

4.1 Dataset Description The datasets used in this paper are those proposed in [5, 19, 20, 21], crawled during November, 2012, whose statistical information is available in Table 1. The number of anchor links crawled between Foursquare and Twitter is 3,388. Foursquare and Twitter share lots of common information as users in both Foursquare and Twitter can make friends with other users, write posts and check in at locations. For more detailed information about the datasets, please refer to [5, 19, 20, 21].

4.2 Experiment Settings In this part, we will introduce the experiment settings in details, which include the comparison methods, evaluation metrics and experiment setups.

4.2.1 Comparison Methods We have different implementations of CAD, which are compared with both state-of-art and traditional community detection methods. All the comparison methods can be divided into 3 categories:

Methods with Parameter Adjustment

• CADe-a (Exact intimacy matrix based CAD with parameter Adjustment): CADe-a can calculate the exact intimacy matrix across aligned attribute augmented networks based on eigenvalue decomposition as proposed in Subsection 3.3, detect communities with matrix factorization and adjust parameter ρ and ω automatically.

• CADa-a (Approximated intimacy matrix based CAD with parameter Adjustment): CADa-a is similar to CADe-a except that CADa-a calculate the *intimacy matrix* with the lower-dimensional *reduced aligned network transition probability matrices* method as proposed in Subsection 3.4.

Methods without Parameter Adjustment

• CADe (Exact intimacy matrix based CAD): CADe is identical to CADe-a except that in CADe, ω and ρ are fixed as $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$ and 0.8 respectively.

• CADa (Approximated intimacy matrix based CAD): CADa is identical to CADa-a except that in CADa, ω and ρ are fixed as $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$ and 0.8 respectively.

Single Network Clustering Methods

• Sinfl (Social Influence-based clustering): Sinfl proposed in 2013 [23] can detect the communities with the influence matrix calculated based on the emerging network only.

• Ncut (Normalized Cut): Ncut [14] aiming at minimizing the normalized cut between different clusters can be used to detect the communities based on the influence matrix obtained by Sinfl in the emerging network.

• Kmeans (Kmeans): Kmeans [3] is a traditional clustering methods, which can also detect social communities in online social networks based on the influence matrix obtained by Sinfl in the emerging network.

4.2.2 Evaluation Metrics Evaluation metrics used to evaluate the performance of all the comparison methods in the experiment include:

• *normalized Davies-Bouldin index*: $ndbi(\mathcal{C}) = \frac{1}{K}\sum_{i=1}^{K} \min_{j\neq i} \frac{d(c_i,c_j)+d(c_j,c_i)}{\sigma_i+\sigma_j+d(c_i,c_j)+d(c_j,c_i)}$, where c_i is the centroid of $U_i \in \mathcal{C}$, $d(c_i, c_j)$ is the distance between c_i and c_j, σ_i denotes the average distance between items in U_i and centroid c_i [23].

• *Silhouette*: Let $a(u) = \frac{1}{|U_i|-1}\sum_{v\in U_i, v\neq u} d(u,v)$ and $b(u) = \min_{j,j\neq i}\left(\frac{1}{|U_j|}\sum_{v\in U_j} d(u,v)\right)$, the *Silhouette index* is defined to be $silhouette(\mathcal{C}) = \frac{1}{K}\sum_{i=1}^{K}(\frac{1}{|U_i|}\sum_{u\in U_i}\frac{b(u)-a(u)}{\max\{a(u),b(u)\}})$ [9].

• *Entropy*: $E(\mathcal{C}) = -\sum_{i=1}^{K} P(i)\log P(i)$, where $P(i) = \frac{|U_i|}{|\mathcal{V}|}$ [23].

4.2.3 Experiment Setups In the experiment, Foursquare and Twitter are used as the emerging and developed networks respectively. As proposed in [19, 20], to obtain networks of different degrees of newness, we randomly sample a proportion of information from Foursquare, which include both link and attribute information controlled by $\sigma_F \in [0, 1]$. If $\sigma_F = 0.0$, then Foursquare is brand new; if $\sigma_F = 0.8$, then the Foursquare network is more developed and 80% of the information is preserved. Meanwhile, considering the abnormally large number of locations and words used in each network, only top 5000 locations (words) that users frequently visited (used) in each network are used in the experiment. Different methods applied to the new Foursquare network can obtain different clustering results. To check whether these clustering methods can discover the communities in the real world, we evaluate the clustering results based on the similarity matrix among users calculated with original complete social information and the similarity measure used is *Jaccard's Coefficient*. If methods can obtain enough reliable information from either the emerging or other developed networks, then their performance will be very good evaluated by different metrics.

4.3 Experiment Results The experiment results are shown in Table 2. Parameter K is fixed as 50 and the ratio of anchor links σ_A is fixed as 0.8 but change the information sampling rate (i.e., σ_F) with values in $\{0.0, 0.1, \cdots, 1.0\}$ to denote different degrees of newness. The results are evaluated by metrics: *ndbi*, *entropy* and *silhouette*.

As shown in Table 2, Sinfl, Ncut and Kmeans cannot work when $\sigma_F = 0.0$ due to the cold start problem. However, CADe-a, CADa-a, CADe and CADa, based on the intimacy matrix across aligned networks, can still work well when $\sigma_F = 0.0$. For example, when $\sigma_F = 0.0$, the *ndbi* score of CADe-a is 0.954; the *entropy* is 3.001; the *silhouette* is -0.396. In addition, for different σ_F, CADe-a, CADa-a, CADe and CADa can perform better than Sinfl, Ncut and Kmeans consistently. It shows that information propagated from aligned network can (1) overcome the cold start problem, and (2) solve the information sparsity problem in emerging network community detection.

Compared with CADe (or CADa), CADe-a (or CADa-a) can perform better in most cases when evaluated by *ndbi*, *silhouette* and *entropy*. For example, when $\sigma_F = 0.8$, the *ndbi* of CADe-a is 0.984, which is 2% higher than that of CADe; the *silhouette* of CADe-a is -0.150, which is 23.4% better than that of CADe. CADe-a (or CADa-a) can always perform better than CADe (or CADa) for $\sigma_F \in \{0.0, 0.1, \cdots, 1.0\}$ when evaluated by *entropy*, as *entropy* is used as the metric when adjusting the parameters. This shows that parameter adjustment method can work very well in determining better parameters.

By comparing CADe, CADa with CADe-a, CADa-a respectively, methods based on approximated intimacy matrix can achieve very similar results as those based on matrix eigendecomposition. Meanwhile, as shown in Table 3 the memory space and time needed

Table 2: Community Detection Result of Foursquare.

measure	methods	Information Sampling Rate σ_F 0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
ndbi	CADE-A	**0.954**	**0.959**	**0.966**	**0.969**	**0.968**	**0.972**	**0.974**	**0.979**	**0.984**	**0.989**	**0.991**
	CADA-A	0.917	0.922	0.923	0.925	0.938	0.946	0.946	0.946	0.947	0.949	0.950
	CADE	0.938	0.944	0.949	0.949	0.954	0.957	0.959	0.966	0.966	0.969	0.969
	CADA	0.914	0.914	0.918	0.923	0.932	0.936	0.939	0.940	0.942	0.942	0.946
	SINFL	-	0.881	0.889	0.901	0.907	0.913	0.913	0.916	0.916	0.917	0.917
	NCUT	-	0.864	0.870	0.889	0.889	0.893	0.894	0.894	0.894	0.897	0.897
	KMEANS	-	0.842	0.859	0.881	0.886	0.887	0.889	0.890	0.892	0.893	0.894
entropy	CADE-A	**3.001**	**2.859**	**2.753**	**2.482**	**2.361**	**2.342**	**2.167**	**2.25**	**2.140**	**1.994**	**1.932**
	CADA-A	4.150	4.137	4.133	4.108	4.084	4.025	4.013	3.856	3.506	3.70	3.68
	CADE	3.751	3.751	3.726	3.718	3.621	3.585	3.38	3.233	3.173	3.005	2.998
	CADA	4.360	4.237	4.213	4.211	4.102	4.061	4.021	4.015	3.97	3.851	3.823
	SINFL	-	5.147	5.105	5.063	4.981	4.968	4.934	4.892	4.856	4.768	4.668
	NCUT	-	5.823	5.691	5.618	5.517	5.512	5.494	5.485	5.473	5.467	5.459
	KMEANS	-	6.182	5.993	5.909	5.888	5.878	5.829	5.812	5.762	5.730	5.699
silhouette	CADE-A	**-0.396**	**-0.272**	**-0.28**	**-0.257**	**-0.251**	**-0.244**	**-0.224**	**-0.216**	**-0.150**	**-0.147**	**-0.132**
	CADA-A	-0.401	-0.384	-0.380	-0.377	-0.287	-0.279	-0.271	-0.270	-0.260	-0.237	-0.238
	CADE	-0.401	-0.302	**-0.275**	-0.270	-0.264	-0.262	-0.242	**-0.222**	**-0.196**	-0.186	**-0.129**
	CADA	-0.401	-0.381	-0.380	-0.372	-0.272	-0.260	-0.259	-0.251	-0.247	-0.246	-0.204
	SINFL	-	-0.482	-0.472	-0.469	-0.463	-0.462	-0.461	-0.459	-0.457	-0.428	-0.408
	NCUT	-	-0.415	-0.413	-0.413	-0.412	-0.410	-0.410	-0.408	-0.408	-0.345	-0.336
	KMEANS	-	-0.515	-0.515	-0.510	-0.508	-0.504	-0.498	-0.467	-0.464	-0.452	-0.434

Table 3: Space and time costs in calculating $\bar{\mathbf{H}}_{align}$.

emerging network	cost	method **exact**	**approx.**
Foursquare	space cost(MB)	19526	1627
	time cost(s)	65996.17	6499.97

by CADA and CADA-A to calculate $\bar{\mathbf{H}}^{approx}_{align}$ is much less than that used by CADE and CADE-A to calculate the exact *intimacy matrix*. So, calculating intimacy matrix with approximation would not harm the performance but can save lots of space and time.

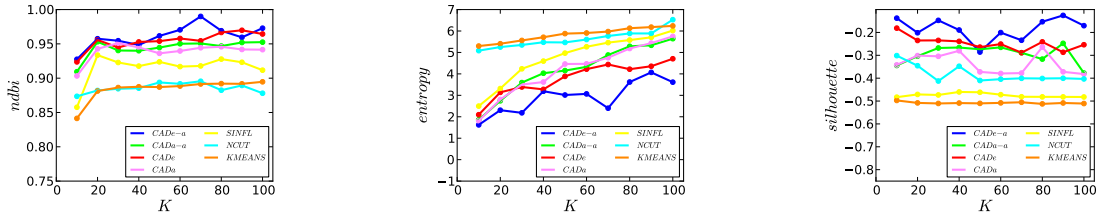
4.4 Parameter Analysis In this part, we will analyze the effect of parameter K (i.e., the number of clusters) on the clustering results. We fix σ_F and σ_A as 0.5 and 0.8 respectively, but change K with values in $\{10, 20, \cdots, 100\}$. The results are shown in Figure 2. As shown in Figures 2(a)-2(c), different methods can achieve the best performance at different Ks when evaluated by different metrics. For example, in Figure 2(a) when evaluated by *ndbi*, CADE-A can perform the best at $K = 70$, but CADE performs the best at $K = 90$. In Figure 2(b), CADE-A performs the best at $K = 70$ and $K = 10$ when evaluated by *entropy* and in Figure 2(c) under the evaluation of *silhouette*, CADE-A can achieve the best performance at $K = 90$.

5 Related Work

Clustering aims at grouping similar objects in the same cluster and many different clustering methods have also been proposed. One type is the hierarchical clustering methods [2], which include agglomerative hierarchical clustering methods [1] and divisive hierarchical clustering methods [1]. Another type of clustering methods is partition-based methods, which include K-means for instances with numerical attributes [3].

In recent years, many community detection works have been done on heterogeneous online social networks. Zhou et al. [22] propose to do graph clustering with relational and attribute information simultaneously. Zhou et al. [23] propose a social influence based clustering method for heterogeneous information networks. Some other works have also been done on clustering with incomplete data. Sun et al. [15] propose to study the clustering problem with complete link information but incomplete attribute information. Lin et al. [8] try to detect the communities in networks with incomplete relational information but complete attribute information.

Multiple aligned heterogeneous networks first studied by Kong et al. [5] have become a hot research topic in recent years. Kong et al. [5] are the first to propose the concept of "anchor link", "aligned heterogeneous networks" and study the anchor link prediction problem across aligned networks. Zhang et al. [19] are the first to study link prediction problem for new users with information transferred from other aligned source networks via anchor links. Zhang et al. [20] are the first to study collective link prediction across "partially aligned location-based social networks". Zhang et al. [21] propose the concepts of collective PU link prediction and extends the traditional intra-network meta paths to inter-network meta paths. Jin et al. study the community detection problem of multiple aligned large-scale networks simultaneously in [4].



(a) evaluated by *ndbi* (b) evaluated by *entropy* (c) evaluated by *silhouette*

Figure 2: Experiment results with different K.

6 Conclusion

In this paper, we have studied the community detection problems for emerging networks. A novel community detection method, CAD, has been proposed to solve the problem. CAD can calculate the intimacy matrix among users across aligned attribute augmented heterogeneous networks with efficient information propagation model. CAD can handle the network heterogeneity and difference problems very well with micro and macro controls, whose parameters can be adjusted automatically. Extensive experiments have been done on real-world partially aligned networks and the results demonstrate effectiveness of CAD in address the emerging network community detection problem.

7 Acknowledgement

This work is supported in part by NSF through grants CNS-1115234, and OISE-1129076, and the Pinnacle Lab at Singapore Management University.

References

[1] P. Cimiano, A. Hotho, and S. Staab. Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text. In *ECAI*, 2004.

[2] T. Hastie, R. Tibshirani, and J. Friedman. Hierarchical clustering. In *The Elements of Statistical Learning*. 2009.

[3] Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowledge Discovery*, 1998.

[4] S. Jin, J. Zhang, P. Yu, S. Yang, and A. Li. Synergistic partitioning in multiple large scale social networks. In *IEEE BigData*, 2014.

[5] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.

[6] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW*, 2008.

[7] C. Lin, Y. Cho, W. Hwang, P. Pei, and A. Zhang. *Clustering Methods in a Protein-Protein Interaction Network*. 2007.

[8] W. Lin, X. Kong, P. Yu, Q. Wu, Y. Jia, and C. Li. Community detection in incomplete information networks. In *WWW*, 2012.

[9] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *ICDM*, 2010.

[10] N. Mishra, R. Schreiber, I. Stanton, and R. Tarjan. Clustering social networks. In *WAW*. 2007.

[11] S. Pan and Q. Yang. A survey on transfer learning. *TKDE*, 2010.

[12] R. Panigrahy, M. Najork, and Y. Xie. How user behavior is related to social affinity. In *WSDM*, 2012.

[13] P. Petersen. *Linear Algebra*. 2012.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.

[15] Y. Sun, C. Aggarwal, and J. Han. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *VLDB*, 2012.

[16] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting homophily effect for trust prediction. In *WSDM*, 2013.

[17] J. Tobias, R. Planqué, D. Cram, and N. Seddon. Species interactions and the structure of complex communication networks. *PNAS*, 2014.

[18] L. Wang, T. Lou, J. Tang, and J. Hopcroft. Detecting community kernels in large social networks. In *ICDM*, 2011.

[19] J. Zhang, X. Kong, and P. Yu. Predicting social links for new users across aligned heterogeneous social networks. In *ICDM*, 2013.

[20] J. Zhang, X. Kong, and P. Yu. Transferring heterogeneous links across location-based social networks. In *WSDM*, 2014.

[21] J. Zhang, P. Yu, and Z. Zhou. Meta-path based multi-network collective link prediction. In *KDD*, 2014.

[22] Y. Zhou, H. Cheng, and J. Yu. Graph clustering based on structural/attribute similarities. *VLDB*, 2009.

[23] Y. Zhou and L. Liu. Social influence based clustering of heterogeneous information networks. In *KDD*, 2013.