ANOMALOUS: A Joint Modeling Approach for Anomaly Detection on Attributed Networks

Zhen Peng[1], **Minnan Luo**[1], **Jundong Li**[2], **Huan Liu**[1], **Qinghua Zheng**[1,3]
[1]MOEKLINNS, School of Electronic and Information Engineering, Xi'an Jiaotong University, China
[2]Computer Science and Engineering, Arizona State University, USA
[3]National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University, China
zhenpeng27@outlook.com, {minnluo,qhzheng}@xjtu.edu.cn, jundongl@asu.edu, huanliucs@gmail.com

Abstract

The key point of anomaly detection on attributed networks lies in the seamless integration of network structure information and attribute information. A vast majority of existing works are mainly based on the Homophily assumption that implies the nodal attribute similarity of connected nodes. Nonetheless, this assumption is untenable in practice as the existence of noisy and structurally irrelevant attributes may adversely affect the anomaly detection performance. Despite the fact that recent attempts perform subspace selection to address this issue, these algorithms treat subspace selection and anomaly detection as two separate steps which often leads to suboptimal solutions. In this paper, we investigate how to fuse attribute and network structure information more synergistically to avoid the adverse effects brought by noisy and structurally irrelevant attributes. Methodologically, we propose a novel joint framework to conduct attribute selection and anomaly detection as a whole based on CUR decomposition and residual analysis. By filtering out noisy and irrelevant node attributes, we perform anomaly detection with the remaining representative attributes. Experimental results on both synthetic and real-world datasets corroborate the effectiveness of the proposed framework.

1 Introduction

Anomaly detection [Chandola *et al.*, 2009; Aggarwal, 2015] aims to spot rare, unexpected and suspicious instances that significantly deviate from the patterns of majority in datasets. It has significant implications in various high-impact domains, such as financial fraud detection, intrusion detection and event detection [Rayana and Akoglu, 2015; Eskin *et al.*, 2002; Chen *et al.*, 2016]. Typically, when an instance is identified as anomalous, we can perform further analysis to investigate why it is abnormal and what makes it different from the other instances in order to gain more insights on the potential risks of the system. Recently, there is a surge of anomaly detection research focusing on attributed networks [Sánchez

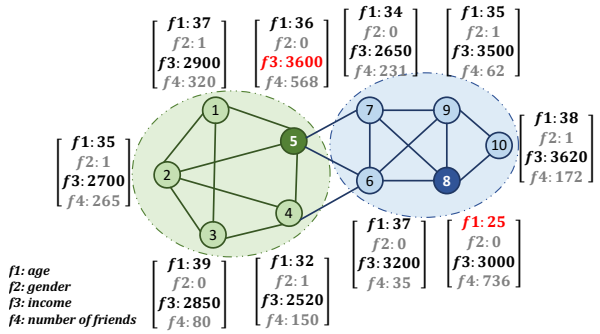

Figure 1: A toy example for anomaly detection on representative attributes via attribute selection.

et al., 2013; Liu *et al.*, 2017; Li *et al.*, 2017] as attributed networks are increasingly used to represent real-world information systems such as social networks, product co-purchase networks and protein-protein interaction networks. Networks of this particular type use edges to describe relationships between data instances, and also collect additional attributes to delineate the properties of nodes. Due to the unique characteristics of attributed networks, anomaly detection faces new challenges. First, how to leverage network structure information and nodal attributes seamlessly for anomaly detection is a crucial yet challenging problem. Secondly, the existence of noisy and structurally irrelevant attributes can impede us to accurately spot anomalies and may even lead to misjudges. Hence, identifying and filtering out these attributes is imperative for anomaly detection.

A vast majority of existing efforts [Davis *et al.*, 2011; Liu *et al.*, 2017; Li *et al.*, 2017] heavily rely on the assumption of Homophily [McPherson *et al.*, 2001] to detect anomalies on attributed networks. The principle of Homophily implies that there exists a strong correlation between network structure and node attributes, instances connected with each other in the network are also similar w.r.t. the nodal attributes. Despite its empirical success, when we conduct further analysis to examine the relations between each attribute and the network structure, we may observe that not all attributes are strongly hinged to the network structure. In other words, there may exist some outlying attributes that do not satisfy the Homophily hypothesis, and the existence of these struc-

turally irrelevant attributes could bring about adverse effects on anomaly detection. Figure 1 illustrates an example of the above-mentioned issues on a real-world social network. In the figure, each node denotes a user and she/he has four attributes (*age, gender, income, number of friends*), users are interconnected with their acquaintances. When performing anomaly detection with all four attributes, we are most likely to regard each node as an anomaly since each node does not conform to the patterns of the majority. Nonetheless, if we only focus on attributes f_1 and f_3, it is obvious that *node 5* is a community anomaly [Gao *et al.*, 2010] as its third attribute (f_3) value is relatively higher than the other nodes within the same community (*node 1, 2, 3* and *4*). Meanwhile, *node 8* is also an anomaly since its first attribute (f_1) value deviates significantly from its structurally connected nodes. This phenomenon indicates that not all attributes show dependencies with the network structure, examples include the binary attribute f_2, and the numerical attribute f_4 which is almost assigned with random values. These noisy and irrelevant attributes hinder us in capturing the correlation between network structure and node attributes, which may further make the true anomalies *node 5* and *node 8* undiscovered. Therefore, selecting representative attributes that are closely hinged with the network structure is essential for anomaly detection. Although there are a few attempts on performing subspace selection for spotting anomalies, they treat it as a pre-processing step before anomaly detection [Sánchez *et al.*, 2013; 2014] rather than optimizing them as a whole, it may lead to a suboptimal result.

In this paper, we propose a joint modeling approach for anomaly detection on attributed networks, called *ANOMALOUS*. Specifically, through optimizing attribute selection and anomaly detection as a whole, *ANOMALOUS* selects a subset of representative instances on the space of representative attributes that are closely hinged with the network topology based on CUR decomposition [Mahoney and Drineas, 2009], and then measures the normality of each instance via residual analysis [She and Owen, 2011]. The main contributions of our work are as follows:

- Examining the issues of existing attempts on performing anomaly detection on attributed networks.
- Introducing a new idea that optimizes attribute selection and anomaly detection as a whole instead of treating them as two separate steps.
- Proposing a novel joint anomaly detection framework *ANOMALOUS* for attributed networks by CUR decomposition and residual analysis.
- Evaluating the performance of the proposed framework on synthetic datasets and real-world datasets.

2 The Proposed Methodology

The notations used in this paper are introduced as follows. Following the standard notation, we use bold uppercase characters (*e.g.*, $\mathbf{A}$) to denote matrices and bold lowercase characters (*e.g.*, $\mathbf{b}$) to indicate vectors. Scalars are written as normal lowercase characters (*e.g.*, c) and uppercase italic letters (*e.g.*, F) for sets. Given a matrix $\mathbf{A}$, we use $\mathbf{A}(i,:)$, $\mathbf{A}(:,j)$ and $\mathbf{A}(i,j)$ to denote its i-th row, j-th column and (i,j)-th entry, respectively. As for the vector or matrix norms, the only used vector norm is the ℓ_2 norm, denoted by $\|\cdot\|_2$. The $\ell_{2,1}$-norm of matrix $\mathbf{A} \in \mathbb{R}^{d\times n}$ is written as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{n}\mathbf{A}(i,j)^2}$, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{n}\mathbf{A}(i,j)^2}$ represents its Frobenius norm and its $\ell_{2,0}$-norm means the number of nonzero rows, denoted by $\|\mathbf{A}\|_{2,0}$.

2.1 Problem Formulation

We first give the formal definition of anomaly detection on attributed networks: suppose $U = \{u_1, u_2, \ldots, u_n\}$ indicates a set of n instances, each instance is affiliated with a set of d-dimensional attributes $F = \{f_1, f_2, \ldots, f_d\}$. We adopt a matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ to represent the attribute information of all n instances, where $\mathbf{X}(:,i) \in \mathbb{R}^d$ denotes the attribute information of the i-th instance u_i. In addition, these instances are interconnected with each other to form a network, and we use the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ to describe the link relationships between them, where $\mathbf{A}(i,j) = 1$ indicates u_i and u_j is connected with each other. The goal of anomaly detection is to find a set of instances which are rare and differ significantly from the majority of the reference instances using the attribute information $\mathbf{X}$ and the network structure information $\mathbf{A}$.

2.2 Joint Anomaly Detection Framework

Residual analysis which aims to study the residuals between true data and estimated data can be applied to spot anomalies since anomalies usually have large residual errors caused by the huge deviations from majority reference instances in patterns [Tong and Lin, 2011]. A well-known method of building estimated data is by using some instances to reconstruct true data [Yu *et al.*, 2006]. In view of the existence of noisy and irrelevant attributes, we would like to simultaneously select structure-related attributes and representative instances to rebuild true data $\mathbf{X}$. Hence, we decide to expand our view from standard factorizations to CUR decomposition, which has better interpretability for instance and attribute selection. Mathematically, the proposed framework is formulated as:

$$\min_{\mathbf{C},\mathbf{U},\mathbf{R},\widetilde{\mathbf{R}}} \|\mathbf{X} - \mathbf{CUR} - \widetilde{\mathbf{R}}\|_F^2 + \Psi(\widetilde{\mathbf{R}}, \gamma, \varphi), \qquad (1)$$

where $\mathbf{C} \in \mathbb{R}^{d\times m}$ is a subset of m columns of $\mathbf{X}$, $\mathbf{R} \in \mathbb{R}^{r\times n}$ is a subset of r rows of $\mathbf{X}$, $\mathbf{U} \in \mathbb{R}^{m\times r}$ with m, r and $rank(\mathbf{U})$ as small as possible, $\widetilde{\mathbf{R}}$ is the residual matrix of node attributes and Ψ is a regularization term on $\widetilde{\mathbf{R}}$.

It can be observed from the above formulation that the process of finding matrix $\mathbf{C}$ and $\mathbf{R}$ indicates the selection of m instances and r attributes from the original data and attribute space. Accordingly, $\mathbf{CUR}$ is a combination of the selected instances and attributes which is supposed to approximate matrix $\mathbf{X}$ as closely as possible, thus the selected instances and attributes can be regarded as representative ones.

Here we give details about the regularization term Ψ as follows:

$$\Psi(\widetilde{\mathbf{R}}, \gamma, \varphi) = \gamma\|\widetilde{\mathbf{R}}^T\|_{2,0} + \varphi tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T), \qquad (2)$$

where $\mathbf{L}$ is a Laplacian matrix generated by the adjacency matrix $\mathbf{A}$ of networks. The first item $\|\widetilde{\mathbf{R}}^T\|_{2,0}$ is used to limit

the number of anomalies since a large norm of $\widetilde{\mathbf{R}}(:,i)$ indicates the i-th instance has a higher possibility to be anomalous, and the number of anomalies is much smaller than that of normal instances. The parameter γ controls the column sparsity of the residual matrix $\widetilde{\mathbf{R}}$. At this point, we have only dealt with attribute information with neglect of network structure. Nevertheless, network structure information provided by attributed networks is also particularly critical for anomaly detection. Theoretically, we expect attribute modeling can filter out noisy and irrelevant attributes so that the Homophily hypothesis holds for selected attributes. Therefore, Based on Homophily, we require that if two instances are connected in the network, their attribute patterns in the residual matrix $\widetilde{\mathbf{R}}$ ought to be similar after attribute reconstruction. Formally, we achieve network structure modeling by minimizing $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\widetilde{\mathbf{R}}^T(i,:)-\widetilde{\mathbf{R}}^T(j,:))^2\mathbf{A}(i,j) = tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T)$. In fact, $\varphi tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T)$ is modeling the correlation between network structure and node attributes, parameter φ controls the contribution of the modeling of this correlation.

2.3 Model Reformulation

Due to the under-determination of a general CUR model, we introduce two indicator vectors $\mathbf{v} = (\mathbf{v}_1,\ldots,\mathbf{v}_n)^T \in \{0,1\}^n$ and $\mathbf{p} = (\mathbf{p}_1,\ldots,\mathbf{p}_d)^T \in \{0,1\}^d$ to reformulate Eq. (1) as:

$$\min_{\mathbf{v},\mathbf{p},\widehat{\mathbf{U}},\widetilde{\mathbf{R}}} \|\mathbf{X}-\mathbf{X}diag(\mathbf{v})\widehat{\mathbf{U}}diag(\mathbf{p})\mathbf{X}-\widetilde{\mathbf{R}}\|_F^2 + \Psi(\widetilde{\mathbf{R}},\gamma,\varphi)$$
$$s.t.\ \ \mathbf{1}_n^T\mathbf{v} = m,\ \mathbf{1}_d^T\mathbf{p} = r, \tag{3}$$

where $diag(\mathbf{v})$ is a diagonal matrix with diagonal elements $\mathbf{v}$ and $\mathbf{1}_n$ is an n-dimensional vector with all elements being 1. $\mathbf{X}diag(\mathbf{v})$ keeps m columns of $\mathbf{X}$ unchanged as selected instances and sets remaining $n-m$ columns to zero vectors. $diag(\mathbf{p})\mathbf{X}$ holds r rows of $\mathbf{X}$ unaltered as selected attributes and sets the rest $d-r$ rows to zero vectors.

Then we reformulate Eq. (3) by adding the regularization term Φ:

$$\min_{\mathbf{W},\widetilde{\mathbf{R}}} \|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{X}-\widetilde{\mathbf{R}}\|_F^2 + \Phi(\mathbf{W},\alpha,\beta) + \Psi(\widetilde{\mathbf{R}},\gamma,\varphi), \tag{4}$$

where $\mathbf{W} = diag(\mathbf{v})\widehat{\mathbf{U}}diag(\mathbf{p}) \in \mathbb{R}^{n\times d}$, and it enables the task of selecting representative instances and attributes simultaneously through a regularization term Φ. The regularization term Φ is defined as:

$$\Phi(\mathbf{W},\alpha,\beta) = \alpha\|\mathbf{W}\|_{2,0} + \beta\|\mathbf{W}^T\|_{2,0}, \tag{5}$$

where α and β controls the row sparsity and the column sparsity of matrix $\mathbf{W}$, respectively. Here we make a key observation that the above definition achieves simultaneous instance and attribute selection. On one hand, $\mathbf{WX}$ in Eq. (4) can be regarded as a coefficient matrix, and when the i-th row of $\mathbf{WX}$ is not a zero vector, $\mathbf{X}(:,i)$ is chosen as a representative instance. Apparently, $\|\mathbf{W}\|_{2,0}$ ensures that only a few instances are chosen to be representative ones. On the other hand, $\mathbf{XW}$ can also be viewed as a coefficient matrix, and when the j-th column of $\mathbf{XW}$ is not a zero vector, $\mathbf{X}(j,:)$ is selected as a representative attribute. Obviously, $\|\mathbf{W}^T\|_{2,0}$ ensures only part of attributes become representative attributes which are leveraged to reconstruct the original data $\mathbf{X}$.

Considering the $\ell_{2,0}$-norm term in Eq. (2) and Eq. (5) will make Eq. (4) NP-hard because of its discrete nature. To address this issue, we relax the $\ell_{2,0}$-norm constraint as $\|\mathbf{W}\|_{2,1}$ which is the minimum convex hull of $\|\mathbf{W}\|_{2,0}$, the objective function of framework *ANOMALOUS* can be formulated as:

$$\begin{aligned}\min_{\mathbf{W},\widetilde{\mathbf{R}}} &\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{X}-\widetilde{\mathbf{R}}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{W}^T\|_{2,1} \\ &+ \gamma\|\widetilde{\mathbf{R}}^T\|_{2,1} + \varphi tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T).\end{aligned} \tag{6}$$

From a global perspective, the correlation modeling term $tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T)$ imposes constraint on instance and attribute selection. We can understand it that in order to minimize the value of $tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T)$, our framework tends to choose instances which are similar to great majority of instances as representative ones, and simultaneously select representative attributes which are closely hinged with the network structure and can represent the whole dataset most precisely. By analyzing the residual matrix $\widetilde{\mathbf{R}}$, we can rank anomalies according to residual errors.

3 Optimization Algorithm

The optimal solution of Eq. (6) is difficult to obtain as the objective function is not convex in terms of both $\mathbf{W}$ and $\widetilde{\mathbf{R}}$ simultaneously, and it is also not smooth due to $\ell_{2,1}$-norm term. Hence, we present an alternating way to solve this problem.

Update $\widetilde{\mathbf{R}}$: When $\mathbf{W}$ is fixed, Eq. (6) is convex $w.r.t.$ $\widetilde{\mathbf{R}}$. Therefore, we first fix $\mathbf{W}$ to update $\widetilde{\mathbf{R}}$, then the objective function can be reformulated as:

$$\min_{\widetilde{\mathbf{R}}} \|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{X}-\widetilde{\mathbf{R}}\|_F^2 + \gamma\|\widetilde{\mathbf{R}}^T\|_{2,1} + \varphi tr(\widetilde{\mathbf{R}}\mathbf{L}\widetilde{\mathbf{R}}^T). \tag{7}$$

By setting the derivative of Eq. (7) $w.r.t.$ $\widetilde{\mathbf{R}}$ to zero, we get

$$\mathbf{X}\mathbf{W}\mathbf{X} + \widetilde{\mathbf{R}} - \mathbf{X} + \gamma\widetilde{\mathbf{R}}\mathbf{D}_R + \varphi\widetilde{\mathbf{R}}\mathbf{L} = 0, \tag{8}$$

where $\mathbf{D}_R$ is a diagonal matrix with diagonal elements $\mathbf{D}_R(i,i) = \frac{1}{2\|\widetilde{\mathbf{R}}^T(i,:)\|_2}$[1]. It can be noticed that matrix $\varphi\mathbf{L}$, $\mathbf{I}$ and $\gamma\mathbf{D}_R$ are both positive semidefinite, so the summation of $\mathbf{I}+\gamma\mathbf{D}_R+\varphi\mathbf{L}$ is also a positive semidefinite matrix. Therefore, we can get a closed-form solution of $\widetilde{\mathbf{R}}$ as follows:

$$\widetilde{\mathbf{R}} = (\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{X})(\mathbf{I}+\gamma\mathbf{D}_R+\varphi\mathbf{L})^{-1}. \tag{9}$$

Update $\mathbf{W}$: When $\widetilde{\mathbf{R}}$ is fixed, Eq. (6) is convex $w.r.t.$ $\mathbf{W}$. Then the objective function can be reformulated as:

$$\min_{\mathbf{W}} \|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{X}-\widetilde{\mathbf{R}}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \beta\|\mathbf{W}^T\|_{2,1}. \tag{10}$$

Similarly, by setting the derivative of Eq. (10) $w.r.t.$ $\mathbf{W}$ to zero, we arrive at

$$\mathbf{X}^T\mathbf{X}\mathbf{W}\mathbf{X}\mathbf{X}^T + \alpha\mathbf{D}_W^{'}\mathbf{W} + \beta\mathbf{W}\mathbf{D}_W^{''} = \mathbf{H}, \tag{11}$$

[1]Considering $\|\widetilde{\mathbf{R}}^T(i,:)\|_2$, $\|\mathbf{W}(i,:)\|_2$ and $\|\mathbf{W}^T(i,:)\|_2$ could be zero theoretically, we redefine $\mathbf{D}_R(i,i) = \frac{1}{2\|\widetilde{\mathbf{R}}^T(i,:)\|_2+\varepsilon}$, $\mathbf{D}_W^{'}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2+\varepsilon}$ and $\mathbf{D}_W^{''}(i,i) = \frac{1}{2\|\mathbf{W}^T(i,:)\|_2+\varepsilon}$ in practical programming, where ε is a very small constant.

where $\mathbf{H} = \mathbf{X}^T\mathbf{X}\mathbf{X}^T - \mathbf{X}^T\widetilde{\mathbf{R}}\mathbf{X}^T$. $\mathbf{D}_W^{'}$ and $\mathbf{D}_W^{''}$ are diagonal matrices with the i-th diagonal element as $\mathbf{D}_W^{'}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2}$ and $\mathbf{D}_W^{''}(i,i) = \frac{1}{2\|\mathbf{W}^T(i,:)\|_2}$, respectively.

Then $\mathbf{W}$ in Eq. (11) can be solved easily according to the following lemmas and theorem:

Lemma 1. *For any matrix $\mathbf{B} \in \mathbb{R}^{m\times n}$ right multiplied by a diagonal matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$, the expression $\mathbf{BD}$ can be rewritten as $\mathbf{D}^* \odot \mathbf{B}$, where $\odot$ denotes the Hadamard product (component-wise multiplication) and $\mathbf{D}^* \in \mathbb{R}^{m\times n}$ is defined as $\mathbf{D}^*(i,j) = \mathbf{D}(j,j)$, $\forall i = 1,2,\ldots,m$, $\forall j = 1,2,\ldots,n$.*

Lemma 2. *For the following matrix equation on $\mathbf{U}$:*

$$\mathbf{C}_1\mathbf{U}\mathbf{C}_2 + a\mathbf{D}_1\mathbf{U} + b\mathbf{D}_2 \odot \mathbf{U} = \mathbf{V}, \tag{12}$$

where $\mathbf{C}_1$ and $\mathbf{C}_2$ are symmetric and positive semidefinite matrices, $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices, a and b are scalers, $\odot$ denotes the Hadamard product. We can reformulate Eq. (12) as follows:

$$\Theta_1\mathbf{K}\Theta_2 + a\mathbf{P}^T\mathbf{D}_1\mathbf{P}\mathbf{K} + b\mathbf{D}_2 \odot \mathbf{K} = \mathbf{P}^T\mathbf{V}\mathbf{Q}, \tag{13}$$

where we decompose $\mathbf{C}_1$ and $\mathbf{C}_2$ by non-negative eigenvalues such that $\mathbf{C}_1 = \mathbf{P}\Theta_1\mathbf{P}^T$, $\mathbf{C}_2 = \mathbf{Q}\Theta_2\mathbf{Q}^T$, $\mathbf{P}$ and $\mathbf{Q}$ are both orthogonal matrices, Θ_1 and Θ_2 are two diagonal matrices composed of the eigenvalues. In addition, $\mathbf{K} = \mathbf{P}^T\mathbf{U}\mathbf{Q}$.

Theorem 1. *The closed-form solution of $\mathbf{W}$ in optimization problem Eq. (10) is*

$$\mathbf{W} = \mathbf{P}\mathbf{Y}\mathbf{Q}^T. \tag{14}$$

Proof. Based on Lemma 1 and Lemma 2, Eq. (11) becomes

$$\Theta_1\mathbf{Y}\Theta_2 + \alpha\mathbf{Z}\mathbf{Y} + \beta\mathbf{D}_W^* \odot \mathbf{Y} = \mathbf{M}, \tag{15}$$

where $\mathbf{X}^T\mathbf{X} = \mathbf{P}\Theta_1\mathbf{P}^T$, $\mathbf{X}\mathbf{X}^T = \mathbf{Q}\Theta_2\mathbf{Q}^T$, $\mathbf{Y} = \mathbf{P}^T\mathbf{W}\mathbf{Q}$, $\mathbf{Z} = \mathbf{P}^T\mathbf{D}_W^{'}\mathbf{P}$ and $\mathbf{M} = \mathbf{P}^T\mathbf{H}\mathbf{Q}$. For a better representation, we denote $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_d] \in \mathbb{R}^{n\times d}$, $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_d] \in \mathbb{R}^{n\times d}$, $\mathbf{Z} = [\mathbf{z}^1; \mathbf{z}^2; \ldots; \mathbf{z}^n] \in \mathbb{R}^{n\times n}$, then the first three terms of Eq. (15) can be rewritten as

$$\Theta_1\mathbf{Y}\Theta_2 = [\Theta_2(1,1)\Theta_1\mathbf{y}_1, \Theta_2(2,2)\Theta_1\mathbf{y}_2, \ldots, \Theta_2(d,d)\Theta_1\mathbf{y}_d],$$

$$\alpha\mathbf{Z}\mathbf{Y} = \alpha\begin{bmatrix} \mathbf{z}^1\mathbf{y}_1 & \mathbf{z}^1\mathbf{y}_2 & \cdots & \mathbf{z}^1\mathbf{y}_d \\ \mathbf{z}^2\mathbf{y}_1 & \mathbf{z}^2\mathbf{y}_2 & \cdots & \mathbf{z}^2\mathbf{y}_d \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}^n\mathbf{y}_1 & \mathbf{z}^n\mathbf{y}_2 & \cdots & \mathbf{z}^n\mathbf{y}_d \end{bmatrix}, \text{and}$$

$$\beta\mathbf{D}_W^* \odot \mathbf{Y} = \beta[\mathbf{D}_W^*(1,1)\mathbf{I}\mathbf{y}_1, \mathbf{D}_W^*(1,2)\mathbf{I}\mathbf{y}_2, \ldots, \mathbf{D}_W^*(1,d)\mathbf{I}\mathbf{y}_d].$$

Plugging the equations above into Eq. (15), we find that Eq. (15) can be decomposed into d equations with respect to d column vectors of matrix $\mathbf{Y}$, *i.e.*,

$$\Theta_2(i,i)\Theta_1\mathbf{y}_i + \alpha\mathbf{Z}\mathbf{y}_i + \beta\mathbf{D}_W^*(1,i)\mathbf{I}\mathbf{y}_i = \mathbf{m}_i, \tag{16}$$

for $i = 1, 2, \cdots, d$. As a result, we have

$$\mathbf{y}_i = (\Theta_2(i,i)\Theta_1 + \alpha\mathbf{Z} + \beta\mathbf{D}_W^*(1,i)\mathbf{I})^{-1}\mathbf{m}_i. \tag{17}$$

Integrating each $\mathbf{y}_i$ obtained by Eq. (17), we achieve the closed-form solution of $\mathbf{Y}$. Recall $\mathbf{Y} = \mathbf{P}^T\mathbf{W}\mathbf{Q}$, it is accessible to $\mathbf{W}$ through basic matrix operations. □

Algorithm 1 *ANOMALOUS*: A joint modeling approach for anomaly detection on attributed networks

Input: Adjacency matrix $\mathbf{A}$, attribute matrix $\mathbf{X}$, parameters $\alpha, \beta, \gamma, \varphi$.
Output: Top t instances with the highest abnormal scores.

```
Build Laplacian matrix L from the adjacency matrix A;
Initialize D_R, D'_W and D''_W to be identity matrix.
Initialize R̃ = X(I + γD_R + φL)^{-1}.
Build orthogonal matrix P, Q and diagonal matrix Θ1, Θ2 based on Lemma 2.
while objective function in Eq. (6) not converge do
    Update W by Eq. (14);
    Update D'_W by setting D'_W(i,i) = 1/(2||W(i,:)||_2);
    Update D''_W by setting D''_W(i,i) = 1/(2||W^T(i,:)||_2);
    Update R̃ by Eq. (9);
    Update D_R by setting D_R(i,i) = 1/(2||R̃^T(i,:)||_2);
end while
Compute the abnormal score for the i-th instance as ||R̃(:,i)||_2;
Output top t instances with the highest abnormal scores.
```

The key steps of framework *ANOMALOUS* spotting anomalies on representative attributes via CUR decomposition and residual analysis are summarized in Algorithm1. At each iteration, the most cost operation is the matrix inverse operation requiring $O(n^3)$ caused by matrix update. Fortunately, the linear equation system can speed up the update of $\widetilde{\mathbf{R}}$, only needing $O(n^2d)$ (d is usually smaller than n). Similarly, the update of $\mathbf{W}$ requires $O(n^2)$ which can be accelerated through calculating each column of $\mathbf{Y}$ in parallel. Consequently, the total time complexity is $\#iterations*(O(n^2d)+O(n^2))$, *i.e.*, $\#iterations * O(n^2d)$.

4 Experiments

In this section, we empirically evaluate the effectiveness of the proposed framework *ANOMALOUS* on both synthetic and real-world datasets. Specifically, we compare our approach with six state-of-the-art methods which are popular in the field of anomaly detection. Here, the compared methods in experiments are listed as follows:

- **SCAN** [Xu *et al.*, 2007]: SCAN only considers the network structure information. It detects anomalies in a structural level.
- **LOF** [Breunig *et al.*, 2000]: LOF only considers attributes of nodes and makes use of all attributes. It detects anomalies in a contextual level.
- **ConSub+CODA** [Sánchez *et al.*, 2013]: ConSub performs the statistical selection of congruent subspaces as a pre-processing step and then uses CODA to detect subspace community anomalies.
- **ConOut** [Sánchez *et al.*, 2014]: For each node, ConOut determines its subgraph and its statistically relevant subset of attributes locally and then detects anomalies in selected local context.

	D10	D20	D40	D60	D80
#nodes	1,000	1,000	1,000	1,000	1,000
#edges	1,434	1,322	1,349	1,361	1,448
#attributes	10	20	40	60	80
ratio of anomaly	10%	10%	10%	10%	10%

Table 1: Details of synthetic datasets

- **AMEN** [Perozzi and Akoglu, 2016]: AMEN uses both attribute information and network structure information to detect anomalous neighborhoods. Since it is designed to detect an abnormal cluster rather than a single node, we regard all nodes in anomalous clusters as anomalies for comparison.
- **Radar** [Li *et al.*, 2017]: Radar detects anomalies by learning and analyzing the residuals of attribute information and its coherence with network information.

In these algorithms listed above, the latter four take both attribute information and structure information into consideration. Only ConSub+CODA, ConOut and AMEN carry out attribute selection, but they perform it as pre-processing step independently of the following anomaly detection which may easily lead to suboptimal results. We adopt the criteria of AUC which is the area under the ROC curve to evaluate the quality of various anomaly detection methods. The closer the AUC value is to 1, the better performance of the algorithm is. In addition, the parameter settings of these baseline methods excluding AMEN and Radar follow the settings of [Sánchez *et al.*, 2013]. For AMEN and Radar, we choose the parameter settings which can get the best experimental results. Considering the proposed framework has four different regularization parameters, we tune these parameters by a "grid-search" strategy from $\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$. The detailed experimental results will be demonstrated later.

4.1 Synthetic Datasets

Considering the detection performance may be affected by irrelevant attributes and increased attribute dimensions, we design an experiment to evaluate the performance of each baseline method and the proposed framework *ANOMALOUS* $w.r.t.$ different attribute dimensions. In order to make our experiment more convincing, we adopt a public synthetic dataset[2] in [Sánchez *et al.*, 2013] which contains attribute information with five different dimensions as well as the ground truth labels of anomalies. The generated network diagrams follow a power law distribution in order to reproduce the properties observed in real-world networks. Node attribute information is divided into network structure related and network structure irrelevant following uniform random distribution, each accounting for 50%. A brief description of this synthetic datasets is shown in Table 1.

Performance Evaluation

The experimental results of all seven algorithms $w.r.t.$ different attribute dimensions are depicted in Figure 2. We can observe from the figure that our framework *ANOMALOUS* achieves the best performance on five different dimensions of attributes. Hence, it is safe to obtain the conclusion that our proposed algorithm is most robust to the increase of attribute dimensions and irrelevant attributes. The main reason is that *ANOMALOUS* can effectively filter out irrelevant attributes and select representative attributes and instances simultaneously for anomaly detection. In addition, we can clearly see that the performance of LOF tends to decrease with increased attribute dimensions. Especially, since SCAN only considers the structure information for anomaly detection, the change in attribute dimensions has no effect on it.

[2]http://www.ipd.kit.edu/∼muellere/consub/

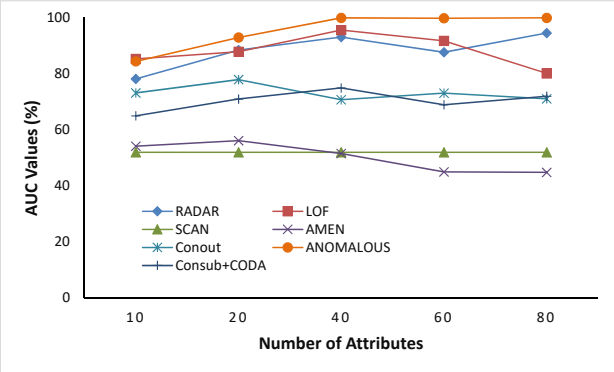

Figure 2: Anomaly detection performance as a function of the number of attributes.

	Disney	Amazon	Enron
#nodes	124	1,418	13,533
#edges	334	3,695	176,987
#attributes	28	28	20
ratio of anomaly	4.8%	2.0%	0.4%

Table 2: Details of real-world datasets

4.2 Real-World Datasets

We adopt three attributed networks that have been widely used in the previous research [Sánchez *et al.*, 2015; Li *et al.*, 2017] to evaluate *ANOMALOUS* on real-world datasets. Among them, Disney dataset is a network of movies including many attributes such as ratings, prices and the number of reviews. Amazon dataset is a network which contains various books as nodes and has similar attributes as Disney. And Enron is a communication network with email transmission as edges between email addresses, the attributes include average content length, average number of recipients, etc. Details about these three datasets are listed in Table 2.

Performance Comparison

The anomaly detection performance of different approaches on three real-world datasets is shown in Figure 3. Compared to other methods, *ANOMALOUS* achieves the best performance on all three datasets. We can attribute this improvement to the joint selection of representative attributes and instances for anomaly detection. In particular, the instances that deviate significantly from the representative instances are recorded as anomalies and the selection of representative attributes can help us target for these representative instances more accurately. As for ConSub+CODA, ConOut and AMEN, they treat subspace selection and anomaly detection as two independent steps, there is no guarantee that the sub-

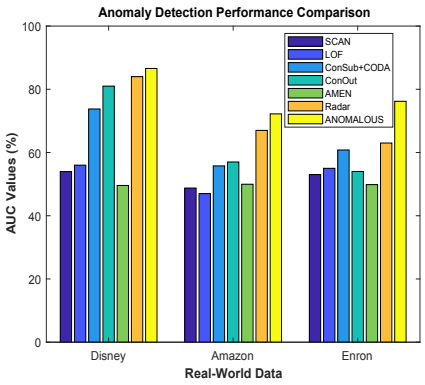

Figure 3: Anomaly detection performance of different approaches.

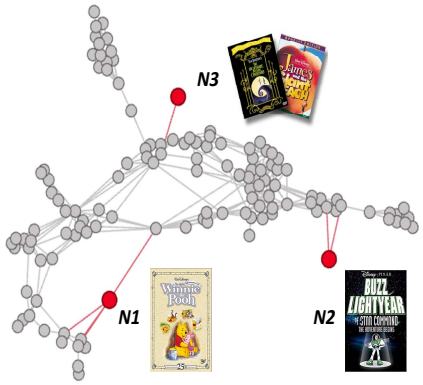

Figure 4: Visualization of three anomalies detected by *ANOMALOUS* on Disney dataset.

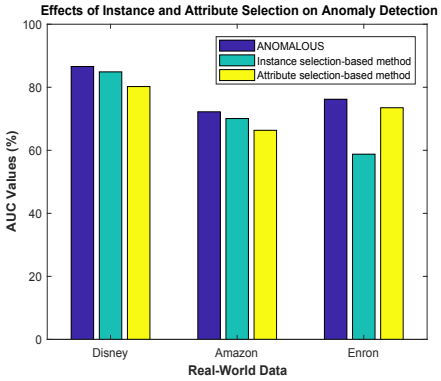

Figure 5: Effects of instance and attribute selection on anomaly detection.

space selection made by these methods can effectively improve the performance of anomaly detection. Moreover, Figure 4 shows the three detected anomalies by our method on Disney dataset. Specifically, node N_1 corresponds to the film *The Many Adventures of Winnie the Pooh* and node N_2 corresponds to the film *Buzz Lightyear of Star Command*, these two nodes deviate significantly in the rating attributes from the other co-purchased products. Node N_3 corresponds to the film *The Nightmare Before Christmas / James and the Giant Peach* is a typical structural anomaly since it is an isolated co-purchase product. In a nutshell, our framework *ANOMALOUS* can help us find anomalies of different formats.

Effects of Instance Selection and Attribute Selection

Among the four parameters in Eq. (6), α and β are relatively more important since they control instance selection and attribute selection respectively. In this subsection, we investigate the impact of instance selection and attribute selection on anomaly detection performance. We compare *ANOMALOUS* with the following methods by varying α and β:

- *Instance selection-based method*: We set the parameter β to be zero, therefore, only instance selection is taken into consideration in the reconstruction phase.
- *Attribute selection-based method*: We set the parameter α to be zero, therefore, only attribute selection is taken into consideration in the reconstruction phase.

We show anomaly detection performance of each method in Figure 5. As can be observed, the AUC values of instance selection-based method and attribute selection-based method are both lower than *ANOMALOUS* on all three datasets. This phenomenon indicates that compared to only selecting instances or attributes, considering both instance selection and attribute selection can indeed improve the performance since they choose representative instances and attributes simultaneously for anomaly detection which can alleviate the negative effects brought from the noisy and irrelevant attributes.

5 Related Work

Typically, existing efforts on anomaly detection for networked data are broadly divided into two classes: (1) anomaly detection on plain networks; and (2) anomaly detection on attributed networks [Akoglu *et al.*, 2015]. For a plain network, the only available information we can collect is the network topology. Therefore, detecting anomalies using structure information such as node degree and subgraph centrality is a major feature of this kind of algorithms. Existing methods such as [Xu *et al.*, 2007; Akoglu *et al.*, 2010; Tong and Lin, 2011] focus on spotting structural anomalies in networks which do not belong to any communities. As for an attributed network, it can be regarded as a richer graph representation having both structure information and attribute information. Hence, anomaly detection algorithms exploit the structure as well as the coherence with node attributes to detect anomalies. There are many existing methods [Long *et al.*, 2006; Gao *et al.*, 2010; Akoglu *et al.*, 2012] working on attributed networks attempt to partition the given graph into structurally dense and attribute-wise homogeneous clusters to find anomalies. But recently, researchers find that not all attributes are correlated to the network structure, thus they conduct unsupervised feature selection (a.k.a. subspace selection) [Gunnemann *et al.*, 2010; Tang and Liu, 2012] to find relevant attributes for anomaly detection. Among them, Consub+CODA carries out subspace selection first and then use CODA detects anomalies [Sánchez *et al.*, 2013]. ConOut determines a local context for each node and then calculates the anomaly score within the local context [Sánchez *et al.*, 2014]. However, these algorithms regard subspace selection and anomaly detection as two independent steps which may lead to suboptimal results. Different from existing works, our proposed algorithm treats anomaly detection and attribute selection in a joint optimization framework.

6 Conclusions and Future Work

In this work, we introduce a joint framework *ANOMALOUS* based on CUR decomposition and residual analysis for spotting anomalies on attributed networks, where attribute and network structure information are collectively used for simultaneous attribute selection and anomaly detection. Experiments on both synthetic and real-world datasets indicate the effectiveness of our approach. Future work will concentrate on including edge attributes or finding adaptive methods [Anava and Levy, 2016] to remodel the network structure.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2016YFB1000903), National Nature Science Foundation of China (61502377, 61532015, 61572399 and 61672418), Innovative Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Center for Engineering Science and Technology.

References

[Aggarwal, 2015] Charu C Aggarwal. Outlier analysis. In *Data Mining*, 2015.

[Akoglu *et al.*, 2010] Leman Akoglu, Mary McGlohon, and Christos Faloutsos. Oddball: Spotting anomalies in weighted graphs. In *PAKDD*, 2010.

[Akoglu *et al.*, 2012] Leman Akoglu, Hanghang Tong, Brendan Meeder, and Christos Faloutsos. Pics: Parameter-free identification of cohesive subgroups in large attributed graphs. In *SDM*, 2012.

[Akoglu *et al.*, 2015] Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015.

[Anava and Levy, 2016] Oren Anava and Kfir Levy. k*-nearest neighbors: From global to local. In *NIPS*, 2016.

[Breunig *et al.*, 2000] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM Sigmod Record*, 2000.

[Chandola *et al.*, 2009] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009.

[Chen *et al.*, 2016] Ting Chen, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, and Kai Zhang. Entity embedding-based anomaly detection for heterogeneous categorical events. In *IJCAI*, 2016.

[Davis *et al.*, 2011] Michael Davis, Weiru Liu, Paul Miller, and George Redpath. Detecting anomalies in graphs with numeric labels. In *CIKM*, 2011.

[Eskin *et al.*, 2002] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. *Applications of Data Mining in Computer Security*, 6:77–102, 2002.

[Gao *et al.*, 2010] Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, and Jiawei Han. On community outliers and their efficient detection in information networks. In *KDD*, 2010.

[Gunnemann *et al.*, 2010] Stephan Gunnemann, Ines Farber, Brigitte Boden, and Thomas Seidl. Subspace clustering meets dense subgraph mining: A synthesis of two paradigms. In *ICDM*, 2010.

[Li *et al.*, 2017] Jundong Li, Harsh Dani, Xia Hu, and Huan Liu. Radar: Residual analysis for anomaly detection in attributed networks. In *IJCAI*, 2017.

[Liu *et al.*, 2017] Ninghao Liu, Xiao Huang, and Xia Hu. Accelerated local anomaly detection via resolving attributed networks. In *IJCAI*, 2017.

[Long *et al.*, 2006] Bo Long, Zhongfei Mark Zhang, Xiaoyun Wu, and Philip S Yu. Spectral clustering for multi-type relational data. In *ICML*, 2006.

[Mahoney and Drineas, 2009] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[Perozzi and Akoglu, 2016] Bryan Perozzi and Leman Akoglu. Scalable anomaly ranking of attributed neighborhoods. In *SDM*, 2016.

[Rayana and Akoglu, 2015] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *KDD*, 2015.

[Sánchez *et al.*, 2013] Patricia Iglesias Sánchez, Emmanuel Muller, Fabian Laforet, Fabian Keller, and Klemens Bohm. Statistical selection of congruent subspaces for mining attributed graphs. In *ICDM*, 2013.

[Sánchez *et al.*, 2014] Patricia Iglesias Sánchez, Emmanuel Müller, Oretta Irmler, and Klemens Böhm. Local context selection for outlier ranking in graphs with multiple numeric node attributes. In *SSDBM*, 2014.

[Sánchez *et al.*, 2015] Patricia Iglesias Sánchez, Emmanuel Müller, Uwe Leo Korn, Klemens Böhm, Andrea Kappes, Tanja Hartmann, and Dorothea Wagner. Efficient algorithms for a robust modularity-driven clustering of attributed graphs. In *SDM*, 2015.

[She and Owen, 2011] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[Tang and Liu, 2012] Jiliang Tang and Huan Liu. Unsupervised feature selection for linked social media data. In *KDD*, 2012.

[Tong and Lin, 2011] Hanghang Tong and Ching-Yung Lin. Non-negative residual matrix factorization with application to graph anomaly detection. In *SDM*, 2011.

[Xu *et al.*, 2007] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas AJ Schweiger. Scan: a structural clustering algorithm for networks. In *KDD*, 2007.

[Yu *et al.*, 2006] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *ICML*, 2006.