Anomaly Alignment Across Multiple Attributed Networks

Jie Zhang
State Key Laboratory of Communication Content Cognition College of Intelligence and Computing, Tianjin University, China
jie_zhang@tju.edu.cn

Nannan Wu*
State Key Laboratory of Communication Content Cognition College of Intelligence and Computing, Tianjin University, China
wunannan@act.buaa.edu.cn

Wenjun Wang*
State Key Laboratory of Communication Content Cognition, College of Intelligence and Computing, Tianjin University, China, Shihezi University,China,
wjwang@tju.edu.cn

Ying Sun
College of Intelligence and Computing, Tianjin University, China
yingsun@tju.edu.cn

Siddharth Bhatia
National University of Singapore,Singapore
siddharth@comp.nus.edu.sg

Abstract

Anomaly subgraph detection has been widely used in various domains, ranging from epidemics, transportation to computer networks and social networks. Despite an increasing need to detect anomaly on multiple networks due to complexity of real-world data, only a limited number of works are available for this task. Most existing methods focus on the anomaly detection on a single attributed network or single domain's multilayer networks, and rarely analyze the correlation among anomalies from different layers. Anomaly detection on attributeless networks is also difficult. Here we propose a novel method *anomaly alignment across multiple attributed networks* (A3MAN), which introduces a representation-based network alignment work to detect the correlated anomaly subgraphs on multiple attributed networks and obtain their related connections. Besides, A3MAN obtains the anomaly subgraph on an attributeless network by detecting and aligning the anomalies of multiple related attributed networks to the attributeless network. We constructed two scenarios to validate A3MAN on two tasks — related anomaly detection on multiple attributed networks and anomaly detection on an attributeless network — using five real-world datasets. In the real computer network scenario, we show that A3MAN outperforms competitive methods by at least 11% accuracy at 10% noise level and the number of related connections among anomaly subgraphs are 13.6 times that of the competitive method (63 times at 30% noise level). In the multi-type traffic scenario of Tianjin, we reveal that the "Yingfengdao" subway station located in the center of the "Nankai University Town" is the peak location of the bicycle-sharing and car-hailing region.

*The corresponding authors are Nannan Wu and Wenjun Wang.

ODD '2021, Aug 15, 2021, Virtual

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

CCS Concepts

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Keywords

anomaly alignment, anomaly subgraph detection, network alignment

ACM Reference Format:
Jie Zhang, Nannan Wu, Wenjun Wang*, Ying Sun, and Siddharth Bhatia. 2018. Anomaly Alignment Across Multiple Attributed Networks. In *6TH OUTLIER DETECTION AND DESCRIPTION WORKSHOP, Aug 15, 2021,Virtual.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

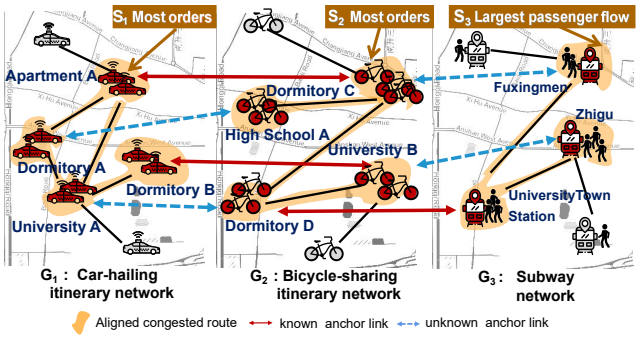

Figure 1: Related abnormal passenger flow itineraries with similar structures detected by A3MAN across three traffic datasets (Car-hailing & Bicycle-sharing & Subway). Within a certain period, different types of traffic usually show similar abnormal patterns, and are highly correlated in geographic attributes and specific locations. A3MAN detected their correlated abnormal structures (yellow shades) and related links (blue lines) by detecting and aligning the abnormalities of their itinerary networks.

1 Introduction

The problem of anomaly detection has recently attracted much more attention. Many methods have been proposed to spot anomalies in different scenarios, such as disease outbreak detection in

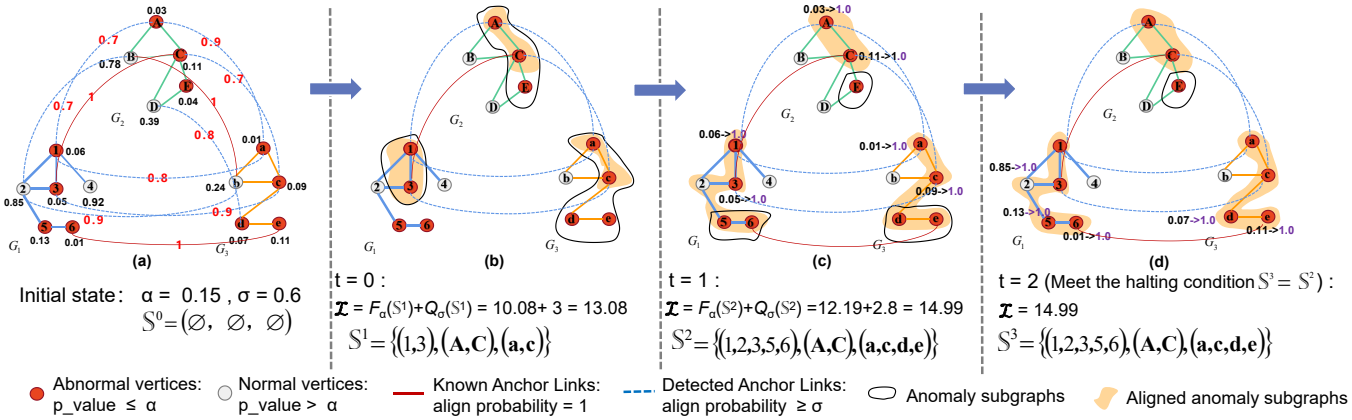

Figure 2: Illustration of A3MAN. The subgraphs within solid black line freeform shapes are the largest anomaly subgraphs (i.e., $\max F_\alpha$), and the yellow shaded subgraphs represent the aligned subgraphs (i.e., $\max Q_\sigma$). The red solid and blue dotted lines connect anomaly subgraphs that meet the alignment conditions. The vertex values and red edge weights are empirical p-values and alignment probabilities.

health alert networks, traffic jam detection in road networks, and event detection in social networks [12, 18, 19, 22, 24]. Most of these works are applied in a single network and cannot handle graph[1] anomalies across multiple networks or multiple datasets. Recently, several papers have mentioned anomaly detection on multiple networks[4, 21]. They focused on anomalies with the same anomaly characteristics on multiple attributed networks in a specific field, and improved the accuracy of the overall anomaly detection. However, no further exploration has been carried out on the correlation among anomalies from different networks. For example, ADOMS[4] focuses on multi-layer social networks, and uses hierarchical information to assist the overall abnormal ranking of nodes, which lacks analyses of the correlation among abnormalities from different social networks. A3MAN pays attention to the anomalies of multiple networks and their correlations in various fields. As Figure 1, in the multi-type traffic scenarios, not only the correlated abnormal itineraries from different networks can be detected but also the correspondence among their related locations (Dormitory B in S_1, University B in S_2 and Zhigu in S_3).

Anomaly detection on the attributeless network is also difficult. The latest work ASD-FT[20] introduces a related attributed network and alignment method to transfer anomalous features to obtain the abnormal information of the attributeless network. ASD-FT focuses on two networks, has not introduced multiple related networks, and relies on spectral method for alignment, which will be difficult to apply to multiple networks. A3MAN introduces a representation-based alignment method, which can use sufficient complementary information from multiple networks to obtain a more comprehensive anomalies and their correlation (As Table 4's *Achor_count*, A3MAN detected more related abnormal IPs).

Anomaly alignment across multiple attributed networks (A3MAN), by alternately performing two steps of anomaly detection and anomaly alignment, to obtain correlated anomaly subgraphs and their related links among multiple networks. Taking Figure1 as an example, we first detect the optimal anomaly subgraphs on the three networks and then align them to get the most correlated parts (yellow shades) with related links. The anomaly subgraphs (S_1,S_2,S_3) present anomaly human activity subgraphs in one city. We can observe that the "*the largest passenger flow*" subway station network aligning with "*the most car-hailing and bicycle-sharing orders*", which means that most people arrive or leave at a subway stations with transfer to bicycle or car at the certain relevant location. We define the problem of anomaly alignment across multiple attributed networks as:

$$\max F_\alpha(\mathbb{S}) + Q_\sigma(\mathbb{S}) \qquad s.t. \quad \mathbb{S} \subseteq \mathbb{G} \tag{1}$$

Where $\mathbb{G} = \{G_i\}$ is an attributed graph set, the graph index $i \in \{1, \cdots, N\}$. $\mathbb{S} = \{S_i\}$ is a set of subgraphs. Each subgraph S_i is connected, whose vertices and edges belong to G_i [24] . F is the abnormal score of $\mathbb{S}$ (e.g., the work TSPSD [24] as F). Q is the alignment score of $\mathbb{S}$ (e.g., the work CrossMNA[6] as Q). The parameters α and σ are significant level and alignment threshold respectively.

Our proposed method A3MAN consists of the main steps: 1) for each attributed G_i, detect its largest anomaly subgraph, and combine with the previous results to get a new subgraph S_i^*; 2) align all S_i^* to obtain the aligned anomaly subgraph S_{i+1} of each network; 3) consider the nodes of S_{i+1} as normal nodes in next step to avoid repeated anomaly detection; and 4) when no update on S_{i+1}, and return the optimal anomaly subgraphs. We have conducted our method on five real datasets and demonstrate the effectiveness of our method. We summarize our main contributions as follows:

1) Innovative work. A3MAN is a pioneer work in detecting associated anomaly subgraphs on multiple attributed networks and provides a method to detect anomalies on the attributeless network.

2) Effectiveness and robustness. Extensive experiments on real datasets have verified that A3MAN can be effectively applied to different multi-network scenarios. On the real computer network dataset, our method achieves 97% accuracy at the ten percent noise level, which is 11% higher than the best competitive baseline.

[1]In this paper, we use 'graph' and 'network' interchangeably, and 'vertex' and 'node' interchangeably, and 'attribute' and 'feature' interchangeably.

Table 1: Representative symbols

Symbol	Description
G_i	$G_i = \{V_i, E_i, P_i\}$, The i-th network with anomaly feature
V_i	The vertex set of G_i
E_i	The edge set of G_i
$\mathbf{p}$	Mapping function to get anomaly features
P_i	The anomaly feature set of G_i
S_i	The anomaly subgraph of G_i
V_{S_i}	The vertex set of S_i
E_{S_i}	The edge set of S_i
$\mathbf{1}_{S_i}$	$\mathbf{1} \in \{1\}^{\lvert V_i \rvert}$, $(\mathbf{1}_{S_i})_v \leftarrow 1$ if $v \in V_{S_i}$, $(\mathbf{1}_{S_i})_v \leftarrow 0$ otherwise
A_{ij}	the set of anchor links between S_i and S_j, $i \neq j$

Besides, our algorithm performs better than all baselines under different noise levels.

3) Comprehensiveness. Our method can detect anomaly subgraphs on the attributeless network because of the introduction of multiple related attributed networks. The information among multiple networks is complementary, so the detection results are more comprehensive. On the real computer network dataset, the anchor links in anomalies are 13.6 times that of the competitive method at ten percent noise level (63 times at thirty percent noise level).

Reproducibility: Our code is publicly available (anonymously) at https://github.com/joyce-hehe/A3MAN.

2 Related work

Our work is related to anomaly detection and network alignment. Here we briefly review these two aspects of work.

2.1 Anomaly detection

Anomaly detection has always been the focus of attention. Point anomalies only assign outliers to nodes, which can be regarded as a binary 0/1 classification problem [1], and there is no connection between the detected abnormal nodes. However, with the expansion of anomaly detection in the field of graphics and the needs of actual scenes, abnormal nodes usually need to be displayed as connected subgraphs. On the other hand, the discovery of anomalies is often inseparable from statistical data. Compared with traditional parameterized scanning statistics (Kulldorff statistical data [12]), nonparametric graph scan statistics (NPGS) can be applied to heterogeneous graph data because it is free of distribution assumption. Therefore, many NPGS-based abnormal connected subgraph detection algorithms were born in combination with actual scenarios and performance requirements. They can be divided into exact algorithms [18, 22] and approximate algorithms [5, 19, 24]. Among them, Wu[24] proved that anomaly subgraph detection is an NP-hard problem, and proposed TSPSD based on dynamic programming. The algorithm approximates the graph to a tree topology and can be used for the large-scale dataset. Most of these algorithms are used in single network, and only a few are used in multi-layer network scenarios [4, 7], and they are all used to identify abnormal nodes, not subgraphs. To the best of our knowledge, the latest work used to detect multiple networks' anomaly subgraphs is ASD-FT [20]. It detects anomaly subgraphs of the graph based on anomaly features of another graph. It introduces network alignment to capture anomaly features' transmission by inferring the basic edges between multiple entity networks. However, it is suitable for two network scenarios, which is different from our work.

2.2 Network alignment

Network alignment is the basic problem of cross-network mining, and many papers have proposed solutions. Most of them are based on attributes and structure. Traditional methods [17, 23] mostly use entity tag information to achieve alignment, such as user nicknames in social networks and entity names in knowledge graphs. Manually defining features is another method [2]. This method needs to carefully design features manually for specific problems, and it is not easy to migrate to other scenarios. Most of the above two types of methods only consider attribute information, while some methods consider both network structure and attribute information (COSNET [28], HYDRA [16], REGAL [10]). They want to complement the network structure and attribute information to achieve a better alignment effect. In addition, because the attribute information may be falsely fabricated or lost or hidden due to privacy, there are many alignment algorithms based only on structural information (BigAlign [11], UMA [27], IONE [15], CrossMNA [6]). Among them, UMA, REGAL, and CrossMNA can be applied to multiple network scenarios. Nevertheless, UMA and REGAL follow the assumption of topological consistency and cannot handle networks with different structures. However, CrossMNA does not follow topological consistency and can learn a common structure across network diversity. By integrating information from different networks, enhances the effect of embedding and effectively reduces space overhead, so it is suitable for large-scale multi-network scenarios.

Our work is based on the TSPSD and CrossMNA algorithms. Compared with the existing work, we are innovative and superior to the baselines in terms of efficiency and comprehensiveness.

3 Problem definition

Definition1 (Multiple Attributed Networks). Given an attributed graph set $\mathbb{G} = \{G_i\}$, where the graph index $i \in \{1, \cdots, N\}$. $G_i = (V_i, E_i, P_i)$ denotes the i-th graph, $V_i = \{v_1, \cdots, v_n\}$, $E_i \subseteq V_i \times V_i$ are sets of vertices and edges in G_i, and P_i is the anomaly feature set of G_i. N is the number of graphs. P_i obtained by the mapping function $\mathbf{p} : V_i \rightarrow [0, 1]$ defines a single empirical p-value corresponding to each node $v \in V_i$ [24], the smaller the p-value, the more abnormal the node. For the attributeless graph, we set the p-values of all its nodes to 1. We use $\mathbb{P}, \mathbb{V}, \mathbb{E}$ to represent the $\{P_i\}$, $\{V_i\}$, $\{E_i\}$ of all graphs. We define an anchor link between G_i and G_j as (v_k^i, v_k^j), where $v_k^i \in V_i$, $v_k^j \in V_j$. Anchor links have transitivity property in networks [27].

Definition 2 (Subgraph). We denote $S_i \subseteq G_i$ as subgraph whose vertices and edges are subset of G_i. We write V_{S_i}, E_{S_i} as the vertex set and edge set of S_i, respectively. Similar to $\mathbb{G}$, we define $\mathbb{S} = \{S_i\}$. Besides, we define $\mathbf{1}$ and $\mathbf{1}_{S_i}$. Among them, $\mathbf{1} \in \{1\}^{\lvert V_i \rvert}$, where $\lvert V_i \rvert$ is the number of vertices in G_i. $\mathbf{1}_{S_i}$ means that the

Algorithm 1 A3MAN

Input: $\mathbb{G} = (G_i), i \in \{1, \cdots, N\}$, $\mathbb{G}$'s Anomalous feature set $\mathbb{P}$, significant level α and alignment threshold σ

Output: The set of aligned anomaly subgraphs $\mathbb{S}$ and the set of $\mathbb{S}$'s anchor links $\mathbb{A}$

Initiate variables $\mathbb{S}^0 = \varnothing, \mathbb{P}^0 = \mathbb{P}, t = 0$
while $\mathbb{S}^t \neq \mathbb{S}^{t+1}$ **do**
  $\mathbb{S}^* \leftarrow \arg\max(F_\alpha(\mathbb{S} \cup \mathbb{S}^t))$, for $\mathbb{S} \subseteq \mathbb{G}$
  $\mathbb{S}^{t+1} \leftarrow \arg\max(Q_\sigma(\mathbb{S}^{'}))$, for $\mathbb{S}^{'} \subseteq \mathbb{S}^*$
  $\mathbb{P}^{t+1} \leftarrow \mathbb{P}^t \cup \mathbb{1}_{\mathbb{S}^{t+1}}$
  $t = t + 1$
end while
return $\hat{\mathbb{S}} = \mathbb{S}^t$

corresponding value of $v \in V_{S_i}$ in $\mathbf{1}$ is set to 1, otherwise it is set to 0, that is, $(\mathbf{1}_{S_i})_v \leftarrow 1$ if $v \in V_{S_i}$, $(\mathbf{1}_{S_i})_v \leftarrow 0$, otherwise. We use $\mathbb{1}_{\mathbb{S}}$ to summarize all $\mathbf{1}_{S_i}$. We define A_{ij} as the set of anchor links between S_i and S_j, $i \neq j$, and use $\mathbb{A}$ to represent all A_{ij}.

This paper aims to detect the aligned anomaly subgraphs among multiple attributed networks to discover the correlation among abnormal patterns. For the obtained anomaly subgraph S_i, we need to evaluate its anomaly score and its alignment score. Therefore, we call this problem "*Anomaly Alignment Across Multiple Attributed Networks*" (A3MAN), and its objective function is defined as follows:

$$\hat{\mathbb{S}} = \arg\max_{\mathbb{S} \subseteq \mathbb{G}} (F_\alpha(\mathbb{S}) + Q_\sigma(\mathbb{S})) \tag{2}$$

where it is the general form of problem (1). The anomaly subgraph $\hat{\mathbb{S}}$ is the optimal solution, and $\mathcal{L} = F_\alpha(\hat{\mathbb{S}}) + Q_\sigma(\hat{\mathbb{S}})$ is the corresponding score. The problem (2) has the following three intuitive properties:

- (P1) F is monotonically ***increased*** with the number of abnormal nodes in S_i.
- (P2) F is monotonically ***decreased*** with the number of normal nodes in S_i.
- (P3) Q is monotonically ***increased*** with the number of node pairs aligned between S_i and S_j.

These properties follow naturally because P1-P2 are widely used in connected anomaly subgraph detection [5, 24], and the possibility of same anomaly increases with aligned links (P3).

4 Methodology

To optimize the problem of *Anomaly Alignment Across Multiple Attributed Networks*, we propose the algorithm A3MAN, which combines two aspects of anomaly subgraph detection and network alignment. The A3MAN method is illustrated in Figure 2, and shows in Algorithm 1. The input of A3MAN is the edge set and anomaly feature set of multiple attributed networks, and the output is the aligned anomaly subgraphs. Besides, A3MAN needs to pre-set significant level α (e.g., 0.15) and alignment threshold σ (e.g., 0.6).

4.1 Detection of anomaly subgraphs

In order to obtain the anomaly score of each anomaly subgraph, we employ the non-parametric graph scanning statistic F as the scoring function, and its form is defined as follows:

$$F_\alpha(S) = \varphi(\alpha, N_\alpha(S), N(S)). \tag{3}$$

where S is a set of connected vertices, that is, a subgraph, and α is the significant levelt (the smaller the α, the higher the abnormal threshold), $N_\alpha(S)$ is the number of anomaly vertices in S whose p-value is less than or equal to α, and $N(S)$ is the total number of vertices in S.

In this paper, A3MAN employs two non-parametric graph scanning statistics as the score function (3): Berk-Jones (BJ) statistic [3] and Higher Criticism (HC) statistic [9]. They are defined as follows:

$$\varphi_{BJ}(\alpha, N_\alpha(S), N(S)) = N(S) \times KL(\frac{N_\alpha(S)}{N(S)}, \alpha), \tag{4}$$

$$\varphi_{HC}(\alpha, N_\alpha(S), N(S)) = \frac{N_\alpha(S) - N(S)\alpha}{\sqrt{N(S)\alpha(1-\alpha)}}. \tag{5}$$

where KL is Kullback-Liebler divergence between the observed and expected proportions of p-values less than α, its formulation is

$$KL(a, b) = \begin{cases} a\log(\frac{a}{b}) + (1-a)log(\frac{1-a}{1-b}), & if\, a \geq b \\ 0, & if\, a < b \end{cases} \tag{6}$$

We use non-parametric graph scanning statistics to specify the abnormality of the subgraph as a numerical value and obtains an overall abnormality score $F_\alpha(\mathbb{S}) = \sum F_\alpha(S_i)$ by accumulating the subgraph scores of each network.

4.2 Alignment of multiple anomaly subgraphs

In order to obtain the alignment score of each subgraph, we define Q as the following form:

$$Q_\sigma(S) = \frac{H_\sigma(S)}{H(S)} \tag{7}$$

where S is an anomaly subgraph, and σ is the predefined alignment threshold (the larger the σ, the higher the alignment threshold). $H_\sigma(S)$ is the number of aligning nodes in S that have alignment with other subgraph nodes, and the alignment probability is greater than or equal to σ. $H(S)$ is the number of all nodes in S.

The node's alignment probability is obtained through the network alignment work CrossMNA [6]. Based on graph embedding, it studies the multi-network alignment problem and can integrate information from different networks to improve alignment performance. By introducing this algorithm, we pre-align $\mathbb{G}$ and obtain the alignment probabilities of all node pairs among networks.

We first use network alignment to map the similarity of anomaly subgraphs, and compute the overall alignment score $Q_\sigma(\mathbb{S}) = \sum Q_\sigma(S_i)$ by accumulating the alignment scores of each subgraph.

4.3 Update of anomaly feature set

In Algorithm 1, we improve $\mathbb{S}^t$ at each iteration, in terms of detecting new anomaly subgraphs. We need to exclude the previous anomaly subgraph $\mathbb{S}^t$, and detect the new anomaly subgraph by updating the anomaly feature set $\mathbb{P}$ of $\mathbb{G}$. The specific operation is: $\mathbb{P}^{t+1} \leftarrow \mathbb{P}^t \cup \mathbb{1}_{\mathbb{S}^{t+1}}$, the p-values of the nodes of $\mathbb{S}^{t+1}$ are set to 1.

We set the anomaly subgraph $\mathbb{S}^{t+1}$ as a normal subgraph, and F_α will be maximized at the new anomaly subgraph in the next iteration. The update operations of the anomaly feature set avoid repeating detection of the same abnormal nodes.

Table 2: Summary of Datasets

Dataset	Division rule	Time	Graph	Node		Edge	
				property	#	property	#
Computer Network	Time	20140531-20140731	$G1$	IP/website	2, 639	IP visited the website	4, 203
		20140801-20140930	$G2$		1, 963		2, 887
		20141001-20141131	$G3$		2, 012		3, 528
		20141201-20150131	$G4$		2, 192		3, 860
		20150201-20150331	$G5$		1, 330		2, 069
		20150401-20150513	$G6$		124, 089		165, 224
*C*ar-*H*ailing	Traffic type	20191220 0:00-24:00	CH	itinerary	54, 962	the same	158, 674
*B*icycle-*S*haring		20191220 0:00-24:00	BS	start/end point	20, 858	itinerary	229, 816
Sub*W*ay (Metro)		20191220 0:00-24:00	SW	station	143	subway line	153

4.4 Theoretical Analysis

Time Complexity. A3MAN's time complexity is $O(k(N|V|^2 + |V|N^2)$, where k is the number of iterations, N is the number of networks, $|V|$ is the number of nodes in the network.

Derivation process: The time complexity of A3MAN is mainly composed of two parts, namely the third step $F_\alpha()$ and the fourth step $Q_\sigma()$ of the algorithm1.

$F_\alpha()$ is an anomaly subgraph detection step based on NPGS, and A3MAN introduces tree-priors-based NPGS method TSPSD to achieve it. For a single attribute network, the algorithm can get the best approximate solution, and the time complexity is $O(|P||V|^2/\epsilon)$ [24]. $|P|$ is the number of feature types in the network anomaly feature set (features with the same numerical value are regarded as the same type), $|V|$ is the number of nodes, and $1+\epsilon$ is the approximate factor. Since we pre-set the significant level α in A3MAN, $|P|$ here is regarded as a constant in A3MAN. Then the time complexity of $F_\alpha()$ is $O(|V|^2/\epsilon)$

$Q_\sigma()$ is a subgraph alignment step based on network alignment. In order to achieve this function, A3MAN introduces CrossMNA[6], a multi-network alignment algorithm based on graph embedding, to obtain the alignment probabilities of all node pairs among networks. CrossMNA uses cross-network information to express the network vector as a combination of two types of node embedding vectors, i.e., inter-vector for network alignment and intra-vector for other downstream network analysis tasks. Their respective dimensions are d, d1, d2, and $d2 \ll d1 \approx d$. The time complexity of CrossMNA is approximately $O(tN(d1d2|V| + d2|E|))$, where t is the number of iterations of CrossMNA, N denotes the number of networks, and $|V|$, $|E|$ denote the number of nodes and edges in each network respectively. In addition, we use a dictionary to store the alignment probability between node pairs in practical applications, so the search time complexity of each iteration of $Q_\sigma()$ is $O(|V|N^2)$.

In summary, the time complexity of A3MAN is $(k(N|V|^2/\epsilon + |V|N^2) + tN(d1d2|V| + d2|E|)$, $N \ll |V|$, expressed as $O(k(N|V|^2 + |V|N^2)$, where k is the number of iterations, N is the number of networks, and $|V|$ is the number of nodes of a network.

Theorem 4.1. ***Convergence and Optimality of A3MAN****. Within the pre-aligned domain (integrates the alignment probability information of all node pairs between the networks obtained by network alignment work), Algorithm 1 converges to the optimal solution of the problem (2).*

Proof. We use the contradiction method to prove Theorem 4.1. Suppose the output of Algorithm 1: $\hat{\mathbb{S}} = (\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_i, \cdots, \hat{S}_j, \cdots, \hat{S}_N)$ is not the global optimal solution. So there is at least one pair nodes (v_k^i, v_k^j) between G_i and G_j, which satisfies the condition that $v_k^i \notin \hat{S}_i$ but connect with $\hat{S}_i$, $v_k^j \notin \hat{S}_j$ but connect with $\hat{S}_j$, $p(v_k^i) \leq \alpha$, $p(v_k^j) \leq \alpha$, and the alignment probability of $(v_k^i, v_k^j) > \sigma$. Let $\hat{S}_i' = \hat{S}_i \cup \{v_k^i\}, \hat{S}_j' = \hat{S}_j \cup \{v_k^j\}, \hat{\mathbb{S}}' = (\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_i', \cdots, \hat{S}_j', \cdots, \hat{S}_N)$, then $N_\alpha(\hat{\mathbb{S}}') > N_\alpha(\hat{\mathbb{S}})$, $P_\sigma(\hat{\mathbb{S}}') > P_\sigma(\hat{\mathbb{S}})$ are held. According by properties of $\mathcal{L}$, the value of $\mathcal{L}$ corresponding to $\hat{\mathbb{S}}'$ is greater than $\hat{\mathbb{S}}$. Therefore, $\mathcal{L}$ of $\hat{\mathbb{S}}$ is not the maximum value that is inconsistent with equation (2). So this inference is contradicted by the assumption. □

By the theorem, our algorithm guarantees on detecting the most anomaly subgraphs on multiple attributed networks.

5 EXPERIMENTS

In this section, we performed a series of experiments to verify A3MAN. We applied the algorithm to two actual scenarios and verified the effectiveness of the algorithm through ground truth. We compared it with three competitive baselines.

5.1 Datasets

We constructed two multi-network scenarios (Table 2) based on the following four real datasets: Computer network, Car-hailing & Bicycle-sharing & Subway (Metro). Moreover, we introduce POI (Point of Interest) dataset as the ground truth of the second scene.

(1) **Computer network:** An Internet company provided browsing logs from the ****edu.cn*** websites, which involved a total of 996 websites, 131, 205 IPs, and the time range was from May 31, 2014 to May 13, 2015, with a total of 3, 978, 073 logs. In this time range, for a certain website/IP on the t-th day, we take the number of logs related to that website/IP on the t-th day as the observed value c_t. By comparing c_t with the daily $c_i, (i < t)$ of the website/IP before the t-th day, the empirical p-value of the website/IP on the t-th day is obtained. Therefore, for all websites/IPs involved in the daily log, we have their corresponding p-value snapshots. We divided these logs into six parts, each containing two months of data, and constructed six computer networks based on this.

(2) **Car-hailing:** This dataset includes 58, 674 online car-hailing orders in Tianjin on December 20, 2019. Each record contains the time of the order and the start and end of the itinerary. We processed

Table 3: Comparison of detecting abnormalities on multiple attributed computer networks.

Algorithms	noise	Recall	Precision	F1	Acc	TPR	FNR
A3MAN (BJ)	0	**1.00**	**0.99**	**1.00**	**0.99**	**1.00**	**0**
A3MAN (BJ)	10	**0.97**	**0.99**	**0.98**	**0.97**	**0.97**	**0.03**
A3MAN (BJ)	30	**0.94**	**0.97**	**0.95**	**0.91**	**0.94**	**0.06**
ASD-FT (BJ)[2020]	0	0.94	0.94	0.94	0.89	0.94	0.06
ASD-FT (BJ)[2020]	10	0.93	0.93	0.92	0.86	0.92	0.08
ASD-FT (BJ)[2020]	30	0.93	0.93	0.93	0.85	0.93	0.07
TSPSD (BJ)[2018]	0	0.96	0.95	0.97	0.95	0.96	0.04
TSPSD (BJ)[2018]	10	0.86	0.97	0.92	0.86	0.86	0.14
TSPSD (BJ)[2018]	30	0.70	0.95	0.81	0.69	0.70	0.30
NPHGS (BJ)[2014]	0	0.91	0.97	0.93	0.89	0.90	0.10
NPHGS (BJ)[2014]	10	0.82	0.95	0.88	0.80	0.82	0.18
NPHGS (BJ)[2014]	30	0.69	0.93	0.79	0.64	0.69	0.31

Table 4: Comparison of detecting abnormalities on attributeless computer network

Algorithms	noise	Anchor_Count	TPR (prediction)	FNR (prediction)
A3MAN (BJ)	0	**1,012**	**0.98**	**0.02**
A3MAN (BJ)	10	**1,011**	**0.95**	**0.05**
A3MAN (BJ)	30	**1,010**	**0.93**	**0.07**
ASD-FT (BJ)[2020]	0	115	0.94	0.06
ASD-FT (BJ)[2020]	10	74	0.88	0.12
ASD-FT (BJ)[2020]	30	16	0.86	0.14

[a] Higher Criticism (HC) statistic results are the same as BJ.

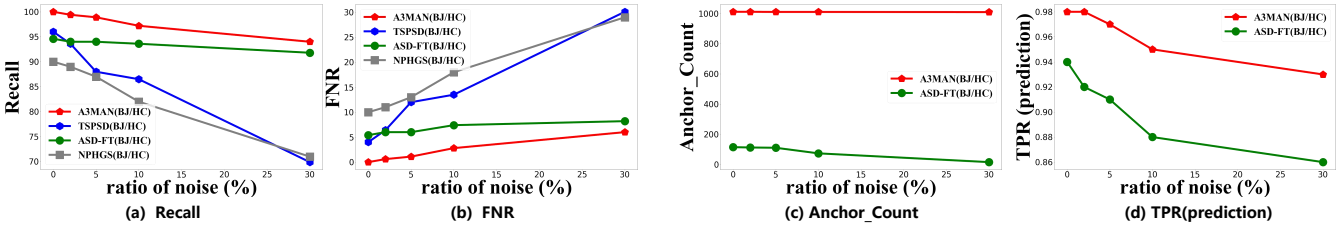

Figure 3: Effect of noise (noise level = 0%, 2%, 5%, 10%, 30%) on Recall, FNR, Anchor_Count and TPR (prediction) of algorithms on the computer network dataset, A3MAN achieves the best performance.

the dataset as follows: First, we divide the dataset into 24 parts according to the time (hours). For hourly data, we used the start point and endpoint of the itinerary as a node in the network and regarded the start point and endpoint of an itinerary is connected so that we can obtain the corresponding itinerary network of this hour. For the node v in the itinerary network of the hour t, we regard the number of times the node v is used as the start point or the endpoint in the hour t as the observed value c_t, and regard observed values c_i, $(i \neq t)$ of v in other hours as compared values, the empirical p-value of v at hour t can be obtained. As a result, we obtained 24 car-hailing itinerary networks.

(3) **Bicycle-sharing:** This dataset includes 229,814 bicycle-sharing orders in Tianjin on December 20, 2019. Each order contains time, start location, and end location. We perform the same processing on this dataset as the car-hailing dataset.

(4) **Subway:** This dataset includes subway traffic data in Tianjin on December 20, 2019. The data recorded the hourly passenger flow of 143 subway stations in Tianjin that day, totaling 3,432. According to the Tianjin subway route, connections are established between interconnected subway stations to form a subway network with 143 nodes and 153 edges. For the subway station s at hour t, we regard the passenger flow of s in that hour as the observed value c_t. Next, perform the same processing as the car-hailing dataset. As a result, we obtained 24 subway networks.

(5) **POI:** This dataset comes from Baidu Maps, recording the information of 242,189 interest points in Tianjin. Each message contains the name, location, telephone number, longitude, latitude, id, and keyword. This dataset can map the longitude and latitude information of the Car-hailing, Bicycle-sharing, and Subway datasets to specific locations, thereby obtaining more realistic results.

5.2 Methods.

To show the effectiveness of A3MAN in multi-network anomaly mining, we compared it with the following baselines in experiments on the computer network dataset. Below we will briefly introduce these methods and their corresponding experimental settings.

Our method. We express Algorithm1 as anomaly alignment across multiple attributed networks (A3MAN) and use BJ and HC statistics to conduct experiments. The parameters to be set for A3MAN include significant level α and alignment threshold σ. In the comparative experiment, we set $\alpha = 0.15$ and $\sigma = 0.8$.

Baselines. Anomaly subgraph detection algorithms:

1) NPHGS [5] is a method that considers the entire heterogeneous network for event detection: it first model the network as a "sensor" network, in which each node senses its "neighborhood environment" and reports an empirical p-value measuring its current level of anomalousness for each time interval (e.g., hour or day). It efficiently maximizes the nonparametric scan statistic over connected subgraphs to identify the most anomalous network clusters. This method's input is the edge set, node-set, and p-value set of a single network, and the output is the largest anomaly subgraph of the network. In the comparative experiment, we set the value of the parameter α_{max} and the number of seed entities K to 0.15 and 5 respectively. Execute the algorithm on each network, compare all the anomaly subgraphs detected with the ground truth, and obtain the corresponding evaluation metrics. **2) TSPSD [24]** is also an anomaly detection algorithm based on non-parametric scanning statistics. It implements efficient anomaly sub-graph detection by reformulating the problem as a series of Budget Price-Collecting Steiner Tree (B-PCST) subproblems. The input of TSPSD is the maximum connected subgraph of a single network and the set of p-values of the network nodes, and the output is the maximum

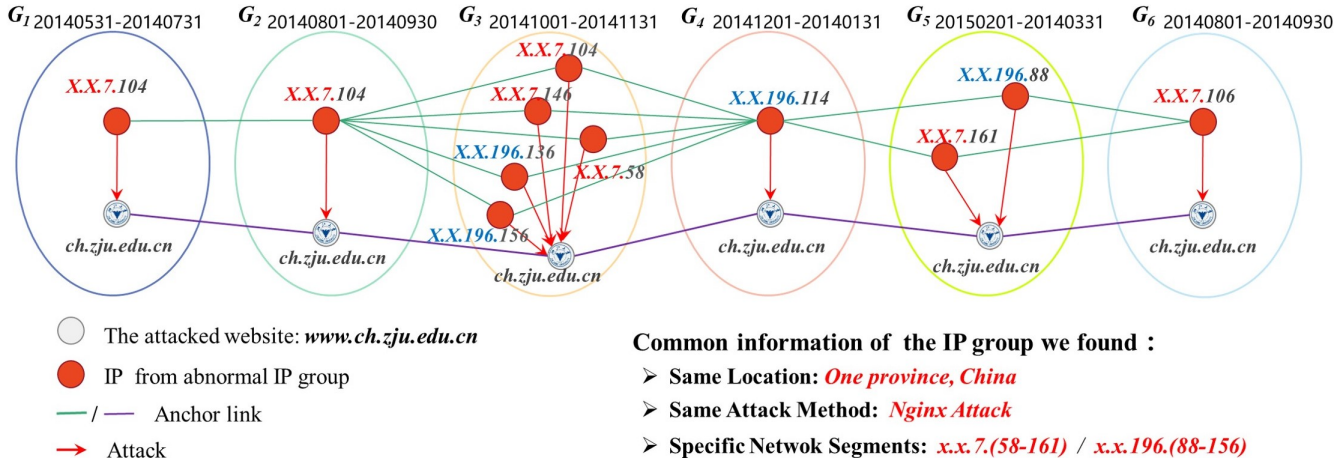

Figure 4: A set of related abnormal IPs detected by A3MAN across multiple networks. It was detected that the abnormal IPs in the set have a certain correlation, they all attacked *www.ch.zju.edu.cn*, and mainly came from two network segments *x.x.7.(58-161)* and *x.x.196.(88-156)*. In addition, through records, it was found that the addresses of these IPs were all *on the same province, China*, and the attack methods were all *Nginx Attack*.

anomaly subgraph of the network. In the experiment, we set the anomaly parameter α_{max} of TSPSD to 0.15, and by separately executing the TSPSD algorithm for each of the multiple networks, we can get the result of the anomaly subgraph of each network. Finally, we compare these results with the ground truth and get the corresponding evaluation metrics of TSPSD.

Baselines. Anomaly Alignment algorithm:

ASD-FT [20] detects anomaly subgraphs of a graph lacking anomaly features through anomaly features of another graph. It captures the transmission of anomaly features by introducing network alignment and inferring the underlying edges between entity graphs. For the network G_1 with anomaly features and the network G_2 without anomaly features, ASD-FT input includes their largest connected subgraph, the set of anomaly features of G_1, the known anchor chain matrix between G_1 and G_2 and the significant level α. To obtain anomalous results across multiple networks and compare them with A3MAN, we set $\alpha = 0.15$ in the experiment. We use ASD-FT to align the each anomaly subgraph S_i of G_i, $i \neq 6$ to G_6, and aggregate them to obtain the anomaly subgraph S_6 of G_6, so that we get the anomaly subgraphs of all networks. We get its evaluation metrics by comparing the result with ground truth.

5.3 Metrics

We use **Recall**, **Precision**, **F1**, **Acc (Accuracy rate)**, **TPR (True Positive Rate)**, **FNR (False Negative Rate)** to evaluate algorithms' ability to detect the overall anomaly on multiple attributed networks, and use **Anchor_Count**,**TPR (prediction)**, **FNR (prediction)** to evaluate the algorithms' ability to detect anomalies on attributeless networks and discover related links among anomalies across the network.

5.4 Experiment Results

We conduct comparative experiments on the computer network dataset with ground truth data, set $\alpha = 0.15$, $\sigma = 0.8$. The experimental results are shown in Table 3 and Table 4.

1) Ability to detect anomalies on multiple attributed networks: Taking all the attributed computer networks as the input of A3MAN and getting the metrics in Table 3, it can be seen that A3MAN outperforms all baselines. For Recall, F1 and TPR, our algorithm reaches 1.00, which is better than all baselines. For Acc, we can see that at ten percent noise level, our algorithm's evaluation index's value is 0.97, which improves at least 11% than the competitive methods. For FNR, we observed that A3MAN's minimum value is 0.00, which is much lower than all baselines, which fully proves the effectiveness of A3MAN in detecting anomalies on multiple attributed networks.

2) Ability to detect anomalies on the attributeless network: By setting the p-values of all nodes in G_6 to 1, it is regarded as a network with missing features. Then run A3MAN to get the metrics in Table4. For comparable TPR (prediction) and FNR (prediction), the A3MAN algorithm reached 0.98 and 0.02, which is significantly better than the two baselines. Moreover, compared with the ASD-FT anomaly alignment algorithm, the total number of abnormal anchor links obtained by our algorithm is $1,012$, which is 8.8 times the 115 of ASD-FT. These results mean that A3MAN has an excellent performance in predicting anomalies.

3) Robustness: To evaluate the robustness of A3MAN, we randomly flip the set of abnormal attribute values $\mathbb{P}$ of $\mathbb{G}$, and make its noise level reach 2%, 5%, 10%, and 30%. The experimental results are shown in Figure 3. For Recall (Figure 3(a)), before the noise level reaches 10%, our algorithm maintains at a level of about 98%. Even if the noise level reaches 30%, it can be maintained at 94%, which is better than all baselines. For FNR (Figure 3(b)), our algorithm has the smallest index value, and the highest value is only 0.06. For Anchor_Count (Figure 3(c)), the number of abnormal anchor links detected by our algorithm is much higher than that of ASD-FT and has been maintained at about 1010, showing stable performance. Besides, when the noise level reaches 30%, this indicator even reaches 63 times the baseline. For the indicator TPR (prediction) (Figure

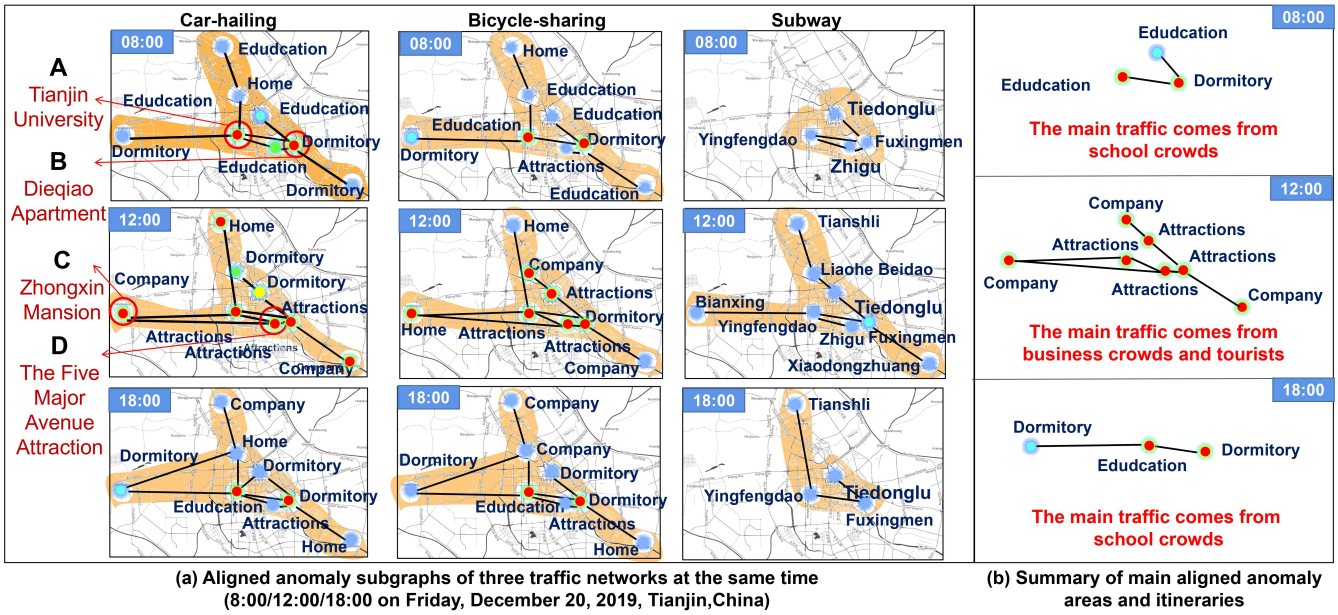

Figure 5: Associated abnormal distribution of Car-hailing & Bicycle-sharing & Subway detected by A3MAN. At 8:00, 12:00 and 18:00 on December 15, 2019 (Friday), by using the A3MAN on the three traffic networks, we can get their aligned anomaly subgraphs at different times. Where (a) is a heat map of the aligned abnormal distribution of the traffic networks. Each hot spot represents an abnormal area, and the redder the color, the higher the abnormality of the area. The black lines indicate that there are itineraries between the two areas, the dark blue text indicates the main position type of the start and end points of the itineraries included in the area. (b) is a summary of the aligned abnormalities at three periods in (a). From the extracted main aligned anomaly areas and itineraries, it can be seen that morning and evening traffic flow mainly comes from school crowds, while at noon, it is mainly business crowds and tourists.

3(d)), A3MAN still performs better than the baseline, and when the noise level reaches 30%, the performance exceeds baseline by 7%.

5.5 Case study in Computer network dataset

Run A3MAN on computer network dataset, input all networks, and set $\alpha = 0.15$, $\sigma = 0.8$.

1) Discovery of related abnormal IP group: Our algorithm can obtain the abnormal IP group and mine the hidden attacking IP information (Figure 4). A3MAN can mine abnormal anchor links across multi-layer networks through the network structure. By summarizing the anchor nodes corresponding to these anchor links, we can obtain an abnormal IP group. Although these IPs appear in different periods, their attack behaviors are similar. Through their log information, we found these IPs come from several fixed network segments, and their attack methods and locations are also the same, which means that these IPs may come from the same attack source. Based on the information obtained, we can prevent their attacks by intercepting IPs from these fixed segments.

2) Prediction of network attacks: We treat G_6 as an attributeless network by setting the p-value of all nodes in G_6 to 1 and use it with other networks as the input of A3MAN to obtain its anomaly subgraph S_6. Regard S_6 as the prediction result, which summarizes the IPs that may attack the website during the period of G_6. We compare it with the real attacks that occurred during this period, and get the TPR (prediction) and FNR (prediction) in Table 4. It can be seen from the metrics that A3MAN can make reasonably accurate predictions of future attacks. Our algorithm can detect the abnormal situation of the target network through networks with sufficient abnormal characteristics, even if the target network does not have any abnormal information.

5.6 Case study in Traffic datasets

Run A3MAN on multiple networks composed of the car-hailing itinerary network, the bicycle-sharing itinerary network, and the subway (metro) network. The experiment aims to discover the correlated anomalies of these three types of traffic networks in the same period. Therefore, we selected three networks with the same period (8:00/12:00/18:00) and set $\alpha = 0.05$, $\sigma = 0.8$.

1) Discovery of associated anomaly distribution and evolution mode: From Figure5(a), it can be found that in the same period, Car-hailing, Bicycle-Sharing, and Subway exhibited similar anomaly distributions. Especially the first two, not only the geographic location of the abnormal nodes but also the anomaly magnitude and type of location all show consistency. For example, at 8:00 and 18:00, Both of their main types of abnormal locations were dormitories and educational institutions, while at noon were attractions and companies. Meanwhile, if we only focus on one traffic type, we find that its anomaly distribution results mined by

A3MAN were similar in the morning and evening (Few abnormal locations, Low abnormality) but different at noon (More abnormal locations, Higher abnormal amplitude). The above results can show that A3MAN can get the associated anomaly distribution and their common anomaly evolution mode.

2) Real abnormal itinerary mining: From Figure 5(a), we know that most of the orders for bicycle-sharing and car-hailing were from A "Tianjin University"and B "Dieqiao Apartment" at 8:00. The passenger flow of the subway stations nearby was also abnormal. Through the ground truth provided by the POI dataset, we learned that "Tianjin University" is in "Nankai University Town", and the nearest subway station is "Yingfengdao". People usually transfer here through car-hailing and bicycle-sharing. "Dieqiao Apartment" is located in the dormitory area. Most people here go to different schools through "Fuxingmen", get off at the subway stations near the corresponding schools, and then transfer by car-hailing and bicycle-sharing. This is why in the educational institutions around "Dieqiao Apartment", the orders were also abnormal. At noon, the peak orders appeared in C "Zhongxin Mansion" and D "The Five Major Avenue Attraction". "Zhongxin Mansion" is located in "Xiqing Industrial Park", where many people go home through "Bianxing" station. Therefore, not only the passenger flow of this station has peaked, but also the orders of car-hailing and bicycle-sharing nearby. "The Five Major Avenue Attraction" is a famous attraction. The nearest station is "Zhigu", so the orders and passenger flow here were abnormal. Besides, from the Figure 5(b) we can intuitively see that at 8:00 and 18:00, the primary sources of orders were educational institutions and dormitories, and at noon they became companies and attractions. It can be inferred that the traffic peaks in the morning and evening were mainly due to the school crowds, while the business crowds and tourists at noon.

6 Conclusion

In this paper, we study the problem of anomaly alignment across multiple attributed networks and propose a solution, A3MAN. A3MAN introduces the network alignment method to the anomaly subgraph detection, mines the correlation of anomalies among multiple attributed networks, and provides a way to detect anomalies on attributeless networks. Extensive experiments show that the algorithm is indeed useful and has better performance than other current work. Future work will further improve the algorithm and make it suitable for more practical scenarios in the economy field, the agricultural field, and the social network field.

7 Acknowledgments

This work was supported by National Key R&D Program of China (No.2018YFC0832103), partly supported by NSFC (No.61902279), and China Postdoctoral Science Foundation (No.2019M650048).

References

[1] Leman Akoglu, Hanghang Tong, and Danai Koutra. 2015. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery* 29, 3 (2015), 626–688.

[2] Mishari Almishari and Gene Tsudik. 2012. Exploring linkability of user reviews. In *European Symposium on Research in Computer Security*. Springer, 307–324.

[3] Robert H Berk and Douglas H Jones. 1979. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 47, 1 (1979), 47–59.

[4] PV Bindu, P Santhi Thilagam, and Deepesh Ahuja. 2017. Discovering suspicious behavior in multilayer social networks. *Computers in Human Behavior* 73 (2017), 568–582.

[5] Feng Chen and Daniel B Neill. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1166–1175.

[6] Xiaokai Chu, Xinxin Fan, Di Yao, Zhihua Zhu, Jianhui Huang, and Jingping Bi. 2019. Cross-network embedding for multi-network alignment. In *The World Wide Web Conference*. 273–284.

[7] Manlio De Domenico, Albert Solé-Ribalta, Elisa Omodei, Sergio Gómez, and Alex Arenas. 2015. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature communications* 6 (2015), 6868.

[8] Youcef Djenouri, Asma Belhadi, Jerry Chun-Wei Lin, Djamel Djenouri, and Alberto Cano. 2019. A survey on urban traffic anomalies detection algorithms. *IEEE Access* 7 (2019), 12192–12205.

[9] David Donoho, Jiashun Jin, et al. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* 32, 3 (2004), 962–994.

[10] Mark Heimann, Haoming Shen, Tara Safavi, and Danai Koutra. 2018. Regal: Representation learning-based graph alignment. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 117–126.

[11] Danai Koutra, Hanghang Tong, and David Lubensky. 2013. Big-align: Fast bipartite graph alignment. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 389–398.

[12] Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26, 6 (1997), 1481–1496.

[13] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1269–1278.

[14] Jure Leskovec and Rok Sosič. 2016. Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 1 (2016), 1–20.

[15] Li Liu, William K Cheung, Xin Li, and Lejian Liao. 2016. Aligning Users across Social Networks Using Network Embedding.. In *Ijcai*. 1774–1780.

[16] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. 2014. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 51–62.

[17] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. 2011. How unique and traceable are usernames?. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 1–17.

[18] Skyler Speakman, Edward McFowland III, and Daniel B Neill. 2015. Scalable detection of anomalous patterns with connectivity constraints. *Journal of Computational and Graphical Statistics* 24, 4 (2015), 1014–1033.

[19] Skyler Speakman, Yating Zhang, and Daniel B Neill. 2013. Dynamic pattern detection with temporal consistency and connectivity constraints. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 697–706.

[20] Ying Sun, Wenjun Wang, Nannan Wu, Wei Yu, and Xue Chen. 2020. Anomaly Subgraph Detection with Feature Transfer. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'20), October 19–23, 2020, Virtual Event, Ireland.*

[21] Maciej Szmit, Anna Szmit, Sławomir Adamus, and Sebastian Bugała. 2012. Usage of Holt-Winters model and multilayer perceptron in network traffic modelling and anomaly detection. *Informatica* 36, 4 (2012).

[22] Kunihiko Takahashi, Martin Kulldorff, Toshiro Tango, and Katherine Yih. 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics* 7, 1 (2008), 14.

[23] Jan Vosecky, Dan Hong, and Vincent Y Shen. 2009. User identification across multiple social networks. In *2009 first international conference on networked digital technologies*. IEEE, 360–365.

[24] Nannan Wu, Feng Chen, Jianxin Li, Jinpeng Huai, Baojian Zhou, Naren Ramakrishnan, et al. 2018. A Nonparametric Approach to Uncovering Connected Anomalies by Tree Shaped Priors. *IEEE Transactions on Knowledge and Data Engineering* 31, 10 (2018), 1849–1862.

[25] M. Yoon, B. Hooi, K. Shin, and C. Faloutsos. 2019. Fast and Accurate Anomaly Detection in Dynamic Graphs with a Two-Pronged Approach. In *the 25th ACM SIGKDD International Conference*.

[26] S. Yoon, J. G. Lee, and B. S. Lee. 2020. Ultrafast Local Outlier Detection from a Data Stream with Stationary Region Skipping. In *The 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[27] Jiawei Zhang and S Yu Philip. 2015. Multiple anonymized social networks alignment. In *2015 IEEE International Conference on Data Mining*. IEEE, 599–608.

[28] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. 2015. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1485–1494.