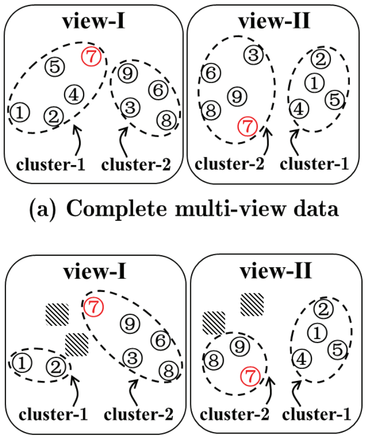
The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Partial Multi-View Outlier Detection Based on Collective Learning

Jun Guo,[1] **Wenwu Zhu**[1,2]
[1] Tsinghua-Berkeley Shenzhen Institue, Tsinghua University, Shenzhen 518055, China
[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
eeguojun@outlook.com, wwzhu@tsinghua.edu.cn

Abstract

In the past decade, various multi-view outlier detection methods have been designed to detect horizontal outliers that exhibit inconsistent across-view characteristics. The existing works assume that all objects are present in all views. However, in real-world applications, it is often the incomplete case that every view may suffer from some missing samples, resulting in partial objects difficult to detect outliers from. To address this problem, we propose a novel Collective Learning (CL) based framework to detect outliers from partial multi-view data in a self-guided way. More specifically, by well exploiting the inter-dependence among different views, we develop an algorithm to reconstruct missing samples based on learning. Furthermore, we propose similarity-based outlier detection to break through the dilemma that the number of clusters is unknown priori. Then, the calculated outlier scores act as the confidence levels in CL and in turn guide the reconstruction of missing data. Learning-based missing sample recovery and similarity-based outlier detection are iteratively performed in a self-guided manner. Experimental results on benchmark datasets show that our proposed approach consistently and significantly outperforms state-of-the-art baselines.



(a) Complete multi-view data

(b) Partial multi-view data

Figure 1: Illustrations of detecting horizontal outliers from (a) complete and (b) partial multi-view data. The shaded area means that the corresponding sample is actually missing.

1 Introduction

Nowadays, data are usually collected from various feature extractors or obtained from diverse domains, and each group of features is conceived as a particular view (Xu, Tao, and Xu 2013). For example, a video can be represented by visual and audio information; an image can be represented by color, shape and other features; and a webpage can be represented by images, words, and URLs on the page. Then, multi-view outlier detection, *i.e.*, identifying the abnormal objects in multi-view data, becomes more challenging due to the complicated distribution and organization of data.

To date, a number of multi-view outlier detection methods (Das et al. 2010; Janeja and Palanisamy 2013; Muller et al. 2012) have been developed. In the similar direction to conventional single-view outlier detection (Chandola, Banerjee, and Kumar 2009; Akoglu, Tong, and Koutra 2015), these multi-view approaches achieved good performance of detecting outliers that have abnormal behaviors in each view. Meanwhile, a new branch of multi-view outlier detection called horizontal anomaly detection is proposed by (Gao et al. 2011), which detects **horizontal outliers** that exhibit inconsistent characteristics (mainly referring to cluster memberships) across different views.

An example is illustrated in Figure 1(a), which can be described as follows. Objects {①,②,④,⑤} stand for a group of developers working on project A, while objects {③,⑥,⑧,⑨} are working on both projects A and B (*e.g.*, in a managerial role). After analyzing the file accesses of users (view-I), we find that object ⑦ has the same access right as the first group {①,②,④,⑤}. From analyzing view-II representing email interactions among users, we find two social network clusters: {①,②,④,⑤} and {③,⑥,⑦,⑧,⑨}. By examining both views, it is apparent that object ⑦ is anomalous due to its inconsistent across-view behaviors (*i.e.* cluster results).

Along this mainline, many effective methods (Liu and Lam 2012; Alvarez et al. 2013; Li, Shao, and Fu 2015; Zhao and Fu 2015; Iwata and Yamada 2016) have been proposed in the past decade. Detecting horizontal outliers from multi-view data has shown more and more significant potential in various applications, such as malicious insider detection (Liu

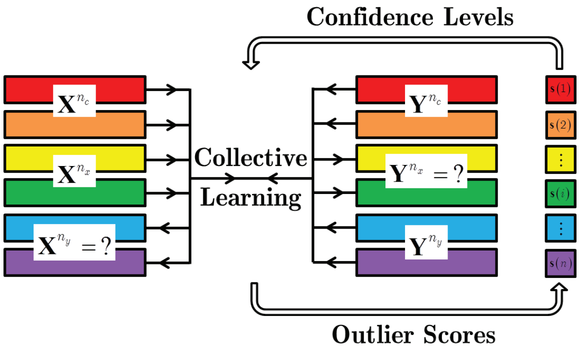


Figure 2: Framework of our proposed method to detect outliers from partial multi-view data in a self-guided manner.

and Lam 2012), purchase behavior analysis (Gao et al. 2013), information disparity management (Duh et al. 2013), and so on. These previous studies assume that all objects are present in all views.

However, in real-world applications, each view may suffer from some missing samples, which results in partial objects (Li, Jiang, and Zhou 2014). Continuing the aforementioned example, Figure 1(b) shows a partial multi-view scenario. Object ③ and ⑥ do not have the corresponding samples in view II, meanwhile, object ④ and ⑤ only have the corresponding samples in view II. This situation is more practical, *e.g.*, there may be new users without historical behaviors in a system. After clustering, the outlier (object ⑦) has been identified as an inlier owing to its consistent across-view cluster memberships. Hence, the incompleteness influences outlier detection. The more partial the multi-view data is, the much more difficult it is to identify outliers.

All previous methods may fail to detect horizontal outliers in partial multi-view scenarios. There are two distinguishing challenges summarized as follows. On the one hand, if an object is present in all views, there exists natural correspondence relationship among its corresponding samples. In (Wang et al. 2016), this relationship is dubbed "co-occurrence". The inter-view consistency of behaviors can be consequently guaranteed. However, the existence of horizontal outliers suppresses this inter-view consistency. It is challenging to determine which object appearing in all views can help to link up different distributions. On the other hand, the objects with missing samples act as troublemakers that disturb the intrinsic doublet/triplet based relations. Since they still provide some information, it is not suitable to remove these objects. The intention of horizontal outlier detection is also for all given objects. It is challenging to make full use of them as well.

In order to address the above challenges, we propose a Collective Learning (CL) based outlier detection approach, which can detect horizontal outliers in the scenario that every view suffers from some missing samples. We present a unified framework to conduct missing sample recovery and outlier detection in a self-guided manner. As illustrated in Figure 2, CL is utilized to reconstruct missing samples by enhancing the inter-dependence among all samples from different views. In order to identify outliers in the intractable case that the cluster number is unknown, we propose a similarity-based Hilbert-Schmidt Independence Criterion for outlier scores. Then, the calculated outlier scores are converted into confidence levels of all objects and in turn guide CL. Learning-based missing sample recovery and similarity-based outlier detection are iteratively integrated together. Comparison experiments on benchmark datasets convincingly demonstrate the superiority of our method.

To the best of our knowledge, this paper is the first attempt to detect outliers from partial multi-view data. In summary, our major contributions are as follows:

1) We propose a novel collective learning based partial multi-view outlier detection method, which iteratively integrates missing sample recovery and outlier detection together in a self-guided way.

2) An effective algorithm with closed-form solutions is developed for learning-based missing samples recovery.

3) Similarity-based outlier scores are designed for outlier detection, which can in turn guide the process of learning.

2 Preliminary

2.1 Notation Summary

Except in some specified cases, lowercase letters $(m, n, \cdots)$ are scalars. Bold lowercase letters $(\mathbf{x}, \mathbf{y}, \cdots)$ denote vectors, while bold uppercase letters $(\mathbf{X}, \mathbf{Y}, \cdots)$ are matrices. $\mathbf{x}(i)$ presents the i^{th} element of $\mathbf{x}$. $\mathbf{X}_{ij}$ is the j^{th} element in the i^{th} row of $\mathbf{X}$. $|\mathbf{X}_{ij}|$ is the absolute value of $\mathbf{X}_{ij}$. $\|\mathbf{X}\|_1$ and $\|\mathbf{X}\|_F$ denote the l_1 norm $(\sum_{i,j}|\mathbf{X}_{ij}|)$ and Frobenius norm $(\sqrt{\sum_{i,j}\mathbf{X}_{ij}^2})$, respectively. $Tr\left(\cdot\right)$, $\left(\cdot\right)^{-1}$, and $\left(\cdot\right)^T$ stand for the trace, inverse, and transpose, respectively. Moreover, $\mathbf{1}$ denotes an all-ones vector with a compatible length. $\mathbf{I}$ is an identity matrix with an appropriate size.

2.2 Hilbert-Schmidt Independence Criterion (HSIC) Revisit

HSIC (Gretton et al. 2005) is a kernel-based metric to measure the independence between two random variables $\mathbf{x}$ and $\mathbf{y}$ by computing the Hilbert-Schmidt-norm of the cross covariance operator over the domain $\mathcal{X} \times \mathcal{Y}$ in Reproducing Kernel Hilbert Spaces (RKHSs). Suppose that $\mathfrak{X}$ and $\mathfrak{Y}$ are two RKHSs in $\mathcal{X}$ and $\mathcal{Y}$, respectively. By the Riesz representation theorem, there are two feature mappings $\varphi\left(\mathbf{x}\right) : \mathcal{X} \rightarrow \mathbb{R}$ and $\phi\left(\mathbf{y}\right) : \mathcal{Y} \rightarrow \mathbb{R}$, such that the kernel function $L\left(\mathbf{x}, \mathbf{x}'\right)$ returns the inner product $\varphi(\mathbf{x})^T\varphi\left(\mathbf{x}'\right)$ in $\mathfrak{X}$ and $K\left(\mathbf{y}, \mathbf{y}'\right)$ returns the inner product $\phi(\mathbf{y})^T\phi\left(\mathbf{y}'\right)$ in $\mathfrak{Y}$.

HSIC can be empirically estimated in the RKHSs by a finite number of samples. Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ denote n observations that are independently and identically drawn from the joint distribution $\mathrm{Pr}_{\mathcal{X}\times\mathcal{Y}}$. Then,

$$\mathrm{HSIC} = \frac{1}{(n-1)^2} Tr\left(\mathbf{HLHK}\right), \tag{1}$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{n\times n}$ is called the centering matrix[1]. $\mathbf{L}, \mathbf{K} \in \mathbb{R}^{n\times n}$ are both kernel matrices with each element $\mathbf{L}_{ij} = L\left(\mathbf{x}_i, \mathbf{x}_j\right)$ and $\mathbf{K}_{ij} = K\left(\mathbf{y}_i, \mathbf{y}_j\right)$.

[1] In this paper, we absorb $\frac{1}{n-1}$ into $\mathbf{H}$ as $\mathbf{H} = \frac{\mathbf{H}}{n-1}$.

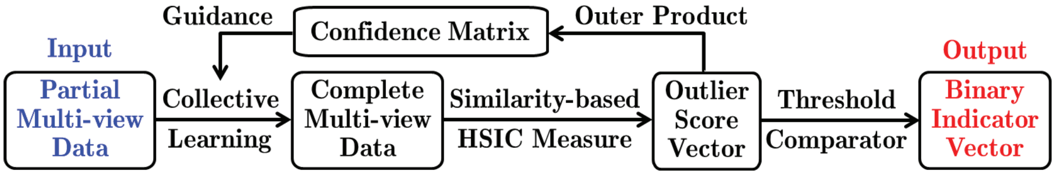


Figure 3: The flowchart of our proposed method to detect outliers for partial multi-view data.

3 The Proposed Method

In this section, we propose a Collective Learning (CL) based framework to detect outliers from partial multi-view data in a self-guided manner. Figure 2 illustrates our CL based framework. Here, we first introduce the mechanism of CL.

3.1 Collective Learning (CL)

Borrowed from the realm of pedagogy (Garavan and Carbery 2012), CL draws on a wide body of theories related to psychology and sociology. It is generally conceptualized as a dynamic process of knowledge production. Learning emerges due to cumulative interactions where individual knowledge is disseminated, diffused, and further shared. Therefore, CL can be conceived as an evolutionary process of perfecting collective knowledge.

Suppose there are n people with diverse professional backgrounds. Towards a novel concept, their opinions may be different and complementary. Then, through collective learning, each person can obtain new knowledge and correct his/her priori knowledge. There are three key components in collective learning:

- The learning process aims to let everyone have a complete knowledge at the end. Thus, CL should maximize the dependency among all people's updated knowledge.
- Each person should have confidence for what he/she has contributed. Thus, other people can determine the quality of knowledge they learned. Consequently, the learning process can prevent from misleading.
- The learning process should be dynamic as well as the individual confidence. The learning process and the update of confidence should be iteratively integrated together.

For partial multi-view data, the incompleteness accompanied by complementarity spontaneously motivates us to utilize CL to reconstruct missing samples (§3.2). Considering that different individuals may have different confidence in his/her knowledge, there is a demand to measure confidence levels. Fortunately, the outlier score is a quantitative estimation of across-view inconsistency, which is a naturally good choice. After missing sample recovery, similarity-based outlier scores are calculated (§3.3) and then converted into confidence levels to guide CL in turn. Learning-based missing sample recovery and similarity-based outlier detection are iteratively integrated together in a self-guided way (§3.4).

3.2 Learning-based Missing Sample Recovery

For the convenience of presentation, assume that we are given n objects with two views. Both of the two related views are incomplete, *i.e.*, $\mathcal{S}_c = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_c}$ is the set of samples present in both views, $\mathcal{S}_x = \{\mathbf{x}_{n_c+i}\}_{i=1}^{n_x}$ is the set of samples only present in the first view, and $\mathcal{S}_y = \{\mathbf{y}_{n_c+n_x+i}\}_{i=1}^{n_y}$ is the set of samples only present in the second view. Hence, we can formulate the above **partial multi-view data** as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{n_c} \\ \mathbf{X}^{n_x} \\ \mathbf{X}^{n_y} = ? \end{bmatrix} \text{ and } \mathbf{Y} = \begin{bmatrix} \mathbf{Y}^{n_c} \\ \mathbf{Y}^{n_x} = ? \\ \mathbf{Y}^{n_y} \end{bmatrix},$$

where $n = n_c + n_x + n_y$. The i^{th} row of $\mathbf{X}$ or $\mathbf{Y}$ stands for a sample of object i. The sub-matrices $\mathbf{X}^{n_y}$ and $\mathbf{Y}^{n_x}$ stand for the missing data in each view, respectively. To help identify outliers from the two incomplete views, we first introduce CL to reconstruct missing samples in this subsection.

The missing data $\mathbf{X}^{n_y}$ (or $\mathbf{Y}^{n_x}$) has inherent dependency upon the given data $\mathbf{Y}^{n_y}$ (or $\mathbf{X}^{n_x}$). According to Eq.(1), maximizing the empirical estimation of HSIC will lead to the maximization of the dependency between two random variables. We adopt linear kernel matrix, *i.e.*, $\mathbf{L} = \mathbf{X}\mathbf{X}^T$ and $\mathbf{K} = \mathbf{Y}\mathbf{Y}^T$. Therefore, we formulate Eq.(2) as the objective function of CL for learning-based missing sample recovery.

$$\max_{\mathbf{X}^{n_y}, \mathbf{Y}^{n_x}} Tr\left(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\right) \tag{2}$$

Without loss of generality, we can follow (Shao, He, and Yu 2015; Zhao, Liu, and Fu 2016) to fill in the missing data $\mathbf{X}^{n_y}$ with average features as initialization. Then, Eq.(2) can be alternately optimized as follows:

1) Solve $\mathbf{Y}^{n_x}$ with fixed $\mathbf{X}^{n_y}$. Denote $\mathbf{P} = \mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}$ as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}^{n_c n_c} & \mathbf{P}^{n_c n_x} & \mathbf{P}^{n_c n_y} \\ (\mathbf{P}^{n_c n_x})^T & \mathbf{P}^{n_x n_x} & \mathbf{P}^{n_x n_y} \\ (\mathbf{P}^{n_c n_y})^T & (\mathbf{P}^{n_x n_y})^T & \mathbf{P}^{n_y n_y} \end{bmatrix}.$$

Then, using the fact that $\mathbf{P}$ is symmetric, Eq.(2) reduces to

$$\max_{\mathbf{Y}^{n_x}} \left\{ \begin{array}{l} Tr\left[\mathbf{Y}^{n_x}(\mathbf{Y}^{n_x})^T\mathbf{P}^{n_x n_x}\right] \\ +2Tr\left[\mathbf{Y}^{n_x}(\mathbf{Y}^{n_c})^T\mathbf{P}^{n_c n_x}\right] \\ +2Tr\left[\mathbf{Y}^{n_y}(\mathbf{Y}^{n_x})^T\mathbf{P}^{n_x n_y}\right] \end{array} \right\}. \tag{3}$$

By setting the derivative w.r.t. $\mathbf{Y}^{n_x}$ to zero, we can acquire the solution[2]

$$\mathbf{Y}^{n_x} = -(\mathbf{P}^{n_x n_x})^{-1}\left[(\mathbf{P}^{n_c n_x})^T\mathbf{Y}^{n_c} + \mathbf{P}^{n_x n_y}\mathbf{Y}^{n_y}\right]. \tag{4}$$

2) Solve $\mathbf{X}^{n_y}$ with fixed $\mathbf{Y}^{n_x}$. Denote $\mathbf{Q} = \mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}$ as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{n_c n_c} & \mathbf{Q}^{n_c n_x} & \mathbf{Q}^{n_c n_y} \\ (\mathbf{Q}^{n_c n_x})^T & \mathbf{Q}^{n_x n_x} & \mathbf{Q}^{n_x n_y} \\ (\mathbf{Q}^{n_c n_y})^T & (\mathbf{Q}^{n_x n_y})^T & \mathbf{Q}^{n_y n_y} \end{bmatrix}.$$

[2] To guarantee the invertibility, we usually add a small perturbation $\epsilon = 10^{-6}$ to each main diagonal element of $\mathbf{P}^{n_x n_x}$ in practice.

Then, in a similar way, we obtain the solution w.r.t. $\mathbf{X}^{n_y}$:

$$\mathbf{X}^{n_y} = -(\mathbf{Q}^{n_y n_y})^{-1}\left[(\mathbf{Q}^{n_c n_y})^T\mathbf{X}^{n_c} + (\mathbf{Q}^{n_x n_y})^T\mathbf{X}^{n_x}\right]. \quad (5)$$

The above steps of CL are alternatively performed until the learning-based missing sample recovery converges.

3.3 Similarity-based Outlier Detection

After missing sample recovery, we construct similarity matrices $\mathbf{W}^{(X)}$and $\mathbf{W}^{(Y)} \in \mathbb{R}^{n\times n}$ for both views. Specially, the similarity matrix $\mathbf{W}^{(X)}$ for n samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is calculated as $\mathbf{W}_{ij}^{(X)} = \begin{cases} 1\ , & j \in \mathcal{N}_i \text{ or } i \in \mathcal{N}_j \\ 0\ , & \text{otherwise} \end{cases}$, where $\mathcal{N}_i$ is a set of indexes indicating the k nearest neighbors of $\mathbf{x}_i$. Similarly, we can obtain $\mathbf{W}^{(Y)}$. Note that both $\mathbf{W}^{(X)}$ and $\mathbf{W}^{(Y)}$ are symmetric matrices.

Remark 1: It is a frequent occurrence that distant samples most likely come from different clusters while samples close to each other are often within the same cluster. $\mathbf{W}_{ij}^{(X)} = 0$ means that $\mathbf{x}_i$ and $\mathbf{x}_j$ are completely dissimilar, and thus may have different labels, while $\mathbf{W}_{ij}^{(X)} = 1$ suggests that the two samples are likely to be grouped into the same cluster.

Remark 2: Despite the number of clusters is unknown a priori, we can denote it as c tentatively. Let $\mathbf{V} \in \mathbb{R}^{n\times c}$ be the cluster indicator matrix of $\mathbf{X}$. Each row of $\mathbf{V}$ is a one-hot label vector: $\mathbf{v}_i = [0, 0, \cdots, 1, \cdots, 0, 0]$, whose non-zero position indicates the cluster label of $\mathbf{x}_i$. Let $\mathbf{J} \in \mathbb{R}^{n\times n}$ denote the linear kernel matrix of $\mathbf{V}$, *i.e.*, $\mathbf{J} = \mathbf{V}\mathbf{V}^T$. Thus, if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, $\mathbf{J}_{ij} = \mathbf{v}_i\mathbf{v}_j^T = 1$; otherwise, $\mathbf{J}_{ij} = \mathbf{v}_i\mathbf{v}_j^T = 0$.

As analyzed in Remark 1 and 2, the similarity matrix is a naturally good surrogate for the linear kernel matrix of corresponding cluster indicators, especially when the total number of clusters is unknown priori. Consequently, we revive the kernel-based HSIC and design a similarity-based criterion for the quantitative estimation of inconsistency.

Denote $\mathbf{s} \in \mathbb{R}^n$ as the outlier score vector. Then, our criterion to estimate each object's outlier score is

$$\mathbf{s}\,(i) = \mathbf{\Delta}_{ii} \ \text{ with } \ \mathbf{\Delta} = \mathbf{H}\mathbf{W}^{(X)}\mathbf{H}\mathbf{W}^{(Y)}. \quad (6)$$

Note that the summation of all outlier scores is an approximation of HSIC, which measures the independence between the cluster indicators of two views. The smaller the score $\mathbf{s}\,(i)$ is, the more likely object i is a horizontal outlier.

3.4 Iterative Integration

After reconstructing samples and computing outlier scores, object i is marked as a horizontal outlier if $\mathbf{s}\,(i)$ is smaller than the threshold τ. One can regard this as a natural ending, but it is not done yet. Recall that our goal is to detect outliers from partial multi-view data. Do all objects contribute equally to missing sample recovery? Of course not. The reason comes from the fact that a horizontal outlier has inconsistent cluster memberships. The more likely an object is an outlier, the less effect it should have on the reconstruction of missing data.

Algorithm 1 Partial Multi-view Outlier Detection Based on Collective Learning

```
Input:
    Partial multi-view data $\mathbf{X}$ and $\mathbf{Y}$;
    the number of self-guided iterations $T$;
    the number of nearest neighbors $k$;
    the threshold $\tau$.
Output:
    Binary outlier indicator vector $\mathbf{o}$.
Initialize missing data with average features, $\mathbf{C} = \mathbf{1}\mathbf{1}^T$.
for $t = 1 : T$ do
    repeat
        Calculate $\mathbf{P} = diag\,(\mathbf{C})\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}$.
        Fix the others and update $\mathbf{Y}^{n_x}$ via Eq.(4).
        Calculate $\mathbf{Q} = \mathbf{H}\mathbf{Y}\mathbf{Y}^T diag\,(\mathbf{C})\mathbf{H}$.
        Fix the others and update $\mathbf{X}^{n_y}$ via Eq.(5).
    until convergence
    Construct similarity matrices $\mathbf{W}^{(X)}$ and $\mathbf{W}^{(Y)}$.
    Calculate the outlier score vector $\mathbf{s}$ via Eq.(6).
    Scale $\{\mathbf{s}(i)\}_{i=1}^n$ to $[0.1, 1]$ and update $\mathbf{C} = \mathbf{s}\mathbf{s}^T$.
end for
Generate the binary outlier indicator vector $\mathbf{o} \in \mathbb{R}^n$,
if $\mathbf{s}\,(i) < \tau$, $\mathbf{o}\,(i) = 1$; otherwise, $\mathbf{o}\,(i) = 0$.
```

To address this issue, we employ the outlier score vector $\mathbf{s}$ to integrate missing sample recovery and outlier detection in a self-guided manner. All scores $\{\mathbf{s}\,(i)\}_{i=1}^n$ are scaled to the range of $[0.1, 1]$ and a confidence matrix $\mathbf{C} = \mathbf{s}\mathbf{s}^T \in \mathbb{R}^{n\times n}$ is calculated. The optimization problem (2) for learning-based missing sample recovery is modified as

$$\max_{\mathbf{X}^{n_y}, \mathbf{Y}^{n_x}} Tr\left[\left(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\right) \odot \mathbf{C}\right], \quad (7)$$

where $\odot$ is the Hadamard product (element-wise multiplication). $\mathbf{C}_{ij} = \mathbf{C}_{ji}$, whose value reflects the pairwise confidence level in measuring the dependency of samples.

Considering the definition of $Tr(\cdot)$, Eq.(7) is equivalent to $\max_{\mathbf{X}^{n_y}, \mathbf{Y}^{n_x}} Tr\left[\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T diag\,(\mathbf{C})\right]$, where $diag\,(\mathbf{C})$ stands for a diagonal matrix with the same main diagonal elements as $\mathbf{C}$. The fact that the main diagonal elements of $\mathbf{C}$ are $\{\mathbf{s}(i)^2\}_{i=1}^n$ provides us a fresh perspective of self-guided mechanism. Just as in collective learning, if a person has less confidence in his/her knowledge, he/she will be shyer to share. Hence, if an object is more likely an outlier, it should provide less contribution to missing sample recovery. Regarding the optimization procedure, we only need to update $\mathbf{P} = diag\,(\mathbf{C})\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}$ and $\mathbf{Q} = \mathbf{H}\mathbf{Y}\mathbf{Y}^T diag\,(\mathbf{C})\mathbf{H}$. Algorithm 1 outlines the whole procedures of our method. As indicated in Algorithm 1, missing sample recovery and outlier detection are iteratively performed in a self-guided way.

3.5 Discussion

Algorithmic analysis: The time complexity for missing sample recovery approximates $\mathcal{O}(n_x^3 + n_y^3)$, which involves the most time-consuming matrix inversion. The time complexity for outlier detection approximates $\mathcal{O}(nd_x + nd_y)$ due to the most computational k nearest neighbor graph construction, where d_x and d_y are the feature dimensions of data.

In practice, we set a maximum iteration number T for the self-guided iteration. We find that the performance gradually levels out in less than 100 iterations for all datasets in our experiments. As for missing sample recovery, we judge the alternating optimization to be converged as long as the value of Eq.(7) changes not obviously ($\leq 10^{-7}$). The experimental results show that it also converges very fast, *i.e.*, at most 100 iterations. Hence, the overall time cost of our method tends to be small because Algorithm 1 converges rapidly.

Extension for multiple views: It is straight-forward to extend our method to the case with more than two views. Suppose there are m incomplete views ($m > 2$). For arbitrary two partial views in which some objects have co-occurring samples, Algorithm 1 can be implemented to generate a binary outlier indicator vector. Then, to decide whether an object is an outlier, we just need to take the majority vote of all its corresponding outlier indicators.

4 Experiment

In this section, we compare our proposed approach with four baseline multi-view outlier detection methods over one synthetic dataset and two real-world datasets.

4.1 Datasets

Synthetic dataset is composed of two views $\{\mathbf{X}, \mathbf{Y}\}$. For each view, we randomly select 200 data points from a two-component Gaussian mixture model as samples. There are two clusters (*i.e.*, cluster 1 and 2). Specifically, the cluster means are $\mu_1^{(X)} = [1, 1]$ and $\mu_2^{(X)} = [4, 2]$ in $\mathbf{X}$, $\mu_1^{(Y)} = [1, 3]$ and $\mu_2^{(Y)} = [3, 1]$ in $\mathbf{Y}$. The corresponding covariances are

$$\mathbf{\Sigma}_1^{(X)} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.4 \end{bmatrix}, \mathbf{\Sigma}_2^{(X)} = \begin{bmatrix} 0.2 & 0.15 \\ 0.15 & 0.35 \end{bmatrix};$$

$$\mathbf{\Sigma}_1^{(Y)} = \begin{bmatrix} 0.25 & -0.05 \\ -0.05 & 0.2 \end{bmatrix}, \mathbf{\Sigma}_2^{(Y)} = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}.$$

Real-world datasets are described as follows.

- **Oxford Flowers Dataset (Flowers)** (Nilsback and Zisserman 2006) is comprised of 17 flower classes, with 80 images per class. Each image is described by color and shape features. In this paper, we adopt the χ^2 distance matrices of different features as different views.
- **USPS-MNIST Dataset** combines two popular handwritten datasets, USPS (Hull 1994) and MNIST (LeCun et al. 1998). The USPS dataset contains 9298 digit images with the size of 16×16, and the MNIST dataset contains 70000 digit images with the size of 28×28. The same digits in two datasets can be regarded as two different views, since they were collected under different scenarios. In the experiments, we follow (Li, Shao, and Fu 2015) and randomly select 50 images per digit from each dataset. Thus, there are 500 samples in each view.

All the above datasets are naturally complete. To simplify the partial multi-view scenarios, we follow the settings in (Shao, He, and Yu 2015) and delete the same number of samples for all views. In specific, we set the Partial Object Ratio (POR) from 0% (all views are complete) to 75% (75% of the total objects have only one view) with 15% as interval. The missing samples are distributed evenly in all views. Note that for each object, it is available in at least one view. Then, we follow the strategy in (Gao et al. 2011) to generate outliers. We take two objects from different classes and swap their samples in one view but not in the others. We randomly perturb 10% of all data in that way.

4.2 Baseline Algorithms

For the compared methods, the following algorithms are considered as baselines:

- HOrizontal Anomaly Detection (HOAD) (Gao et al. 2011): Horizontal outliers are identified by comparing the calculated cosine distances among different spectral embeddings.
- Anomaly detection via Affinity Propagation (AP) (Alvarez et al. 2013): Horizontal outliers are detected via the distances among different clustering results from the affinity propagation algorithm.
- Multi-view Low-Rank Analysis (MLRA) (Li, Shao, and Fu 2015): Horizontal outliers are identified by $l_{2,1}$-norm regularized low-rank subspace learning.
- Dual-regularized Multi-view Outlier Detection (DMOD) (Zhao and Fu 2015): Horizontal outliers are detected via $l_{2,1}$-norm induced K-means clustering.

Detailed description of these methods is in §5. For AP, we utilize the l_2 distance with HSIC to yield better performance. For MLRA and DMOD, we focus on detecting across-view inconsistence by setting the trade-off parameters in outlier score estimation to zero. Since these works cannot directly deal with partial multi-view data, we pre-process the incomplete views by mean imputation (Shao, He, and Yu 2015) for these methods. That is to say, regarding each missing sample, we use the linear combination (weighed by similarities) of its 5 nearest neighbors that appear in both views.

4.3 Results and Analysis

Outlier detection for partial multi-view data is evaluated by AUC, *i.e.*, the area under Receiver Operating Characteristic (ROC) curve. We repeat all experiments 10 times and report the means and standard deviations of AUC for all algorithms in Table 1-3. The best results are in boldface.

It can be observed that our proposed method significantly outperforms other multi-view outlier detection methods. The main reason is that our unified framework integrates missing sample recovery and outlier detection in a self-guided manner. Without doubt, our proposed method can achieve better performance than using two-step strategies, *i.e.*, reconstructing missing samples followed by outlier detection. The other observation is that the performance of all methods drops with more missing samples (*i.e.*, POR increases). The possible reason is that when POR increases, the quality of information in the filling samples through mean imputation or collective learning tends to be bad or even misleading.

In addition, we conducted a statistical significance test for all the results to judge the significant improvements of

Table 1: AUC results and p-values on the synthetic dataset under different POR settings.

	0%	15%	30%	45%	60%	75%
HOAD	0.6512±0.0445	0.6036±0.0451	0.5815±0.0429	0.5598±0.0461	0.5302±0.0470	0.5067±0.0474
	(2.15×10^{-15})	(4.58×10^{-10})	(1.16×10^{-12})	(9.24×10^{-6})	(4.10×10^{-17})	(1.98×10^{-20})
AP	0.8946±0.0397	0.8739±0.0475	0.8519±0.0511	0.8340±0.0467	0.8017±0.0554	0.7811±0.0569
	(6.14×10^{-18})	(9.81×10^{-8})	(5.64×10^{-11})	(9.83×10^{-17})	(5.98×10^{-11})	(9.11×10^{-17})
MLRA	0.7038±0.0458	0.6798±0.0462	0.6512±0.0471	0.6264±0.0475	0.5984±0.0476	0.5603±0.0481
	(3.16×10^{-19})	(6.22×10^{-20})	(1.84×10^{-15})	(8.16×10^{-16})	(2.58×10^{-15})	(4.41×10^{-20})
DMOD	0.7612±0.0551	0.7408±0.0566	0.7054±0.0561	0.6819±0.0570	0.6590±0.0582	0.6244±0.0578
	(5.96×10^{-18})	(6.38×10^{-17})	(6.10×10^{-19})	(4.41×10^{-20})	(8.77×10^{-14})	(3.28×10^{-16})
Ours	**0.9385±0.0260**	**0.9177±0.0252**	**0.8826±0.0259**	**0.8543±0.0245**	**0.8291±0.0258**	**0.8055±0.0243**

Table 2: AUC results and p-values on the Flowers dataset under different POR settings.

	0%	15%	30%	45%	60%	75%
HOAD	0.6481±0.0725	0.6179±0.0755	0.5933±0.0711	0.5706±0.0697	0.5470±0.0684	0.5219±0.0715
	(1.68×10^{-16})	(3.09×10^{-14})	(5.29×10^{-15})	(8.16×10^{-10})	(9.47×10^{-18})	(7.13×10^{-19})
AP	0.8066±0.0457	0.7907±0.0481	0.7859±0.0412	0.7760±0.0501	0.7521±0.0496	0.7451±0.0483
	(1.08×10^{-14})	(4.09×10^{-15})	(2.96×10^{-16})	(8.35×10^{-9})	(1.58×10^{-12})	(5.01×10^{-15})
MLRA	0.7435±0.0502	0.7028±0.0513	0.6469±0.0469	0.6201±0.0487	0.5993±0.0495	0.5645±0.0504
	(3.56×10^{-18})	(2.19×10^{-20})	(3.00×10^{-16})	(7.01×10^{-18})	(6.48×10^{-18})	(5.04×10^{-20})
DMOD	0.7329±0.0531	0.6867±0.0542	0.6454±0.0563	0.6001±0.0551	0.5582±0.0536	0.5276±0.0545
	(6.00×10^{-19})	(3.18×10^{-12})	(4.85×10^{-14})	(6.91×10^{-12})	(9.07×10^{-15})	(3.12×10^{-18})
Ours	**0.8712±0.0215**	**0.8623±0.0268**	**0.8488±0.0247**	**0.8367±0.0251**	**0.8012±0.0272**	**0.7898±0.0274**

Table 3: AUC results and p-values on the USPS-MNIST dataset under different POR settings.

	0%	15%	30%	45%	60%	75%
HOAD	0.6611±0.0610	0.6407±0.0576	0.6196±0.0582	0.5876±0.0595	0.5523±0.0621	0.5259±0.0605
	(5.71×10^{-21})	(3.58×10^{-20})	(6.54×10^{-18})	(1.08×10^{-17})	(2.70×10^{-18})	(4.92×10^{-20})
AP	0.8738±0.0379	0.8569±0.0391	0.8301±0.0389	0.8127±0.0377	0.7814±0.0410	0.7470±0.0519
	(1.28×10^{-14})	(3.50×10^{-10})	(2.44×10^{-13})	(8.14×10^{-9})	(1.98×10^{-13})	(2.41×10^{-12})
MLRA	0.8361±0.0495	0.8102±0.0502	0.7839±0.0487	0.7564±0.0513	0.7301±0.0479	0.7037±0.0485
	(3.25×10^{-20})	(1.34×10^{-18})	(7.18×10^{-18})	(6.37×10^{-15})	(3.05×10^{-16})	(9.24×10^{-19})
DMOD	0.8272±0.0522	0.8003±0.0534	0.7792±0.0540	0.7386±0.0539	0.7015±0.0546	0.6593±0.0550
	(3.47×10^{-19})	(4.51×10^{-20})	(7.15×10^{-17})	(6.20×10^{-13})	(5.07×10^{-17})	(9.10×10^{-21})
Ours	**0.9147±0.0214**	**0.8938±0.0230**	**0.8784±0.0218**	**0.8501±0.0225**	**0.8308±0.0215**	**0.8123±0.0223**

the developed models in comparison with the state-of-the-art methods. The significance level, *i.e.*, p-value, is typically set to 0.05, which means that if the significance evaluation is lower than this level, the performance difference between the evaluated methods is statistically significant. The p-values between our proposed method and the compared methods are shown in the parentheses. We can see that the performance differences between our method and all the compared methods are statistically significant, which also improves the effectiveness of our method.

4.4 Parameter and Convergence Study

Since AUC is adopted as the evaluation metric, we do not need to specify the threshold τ in Algorithm 1. Then, there are two major parameters, *i.e.*, the number of self-guided iterations T and the number of nearest neighbors k.

As for parameter T, we have observed that the performance gradually levels out for all datasets in our experiments. Figure 4 shows the convergence trend against parameter T (iteration numbers) on the synthetic dataset. In general, there are two stages seen from each curve: in the first stage, the objective function value increases dramatically; in the second stage, the increment becomes gradually inconspicuous. Besides, the other observation is that our method converges faster under a larger POR. The reason is that the two data matrices for the n_c objects appearing in both views will inevitably affect the step of learning. A larger POR indicates that there are fewer differences between $\mathbf{X}^{n_c}$ and $\mathbf{Y}^{n_c}$, hence Algorithm 1 needs

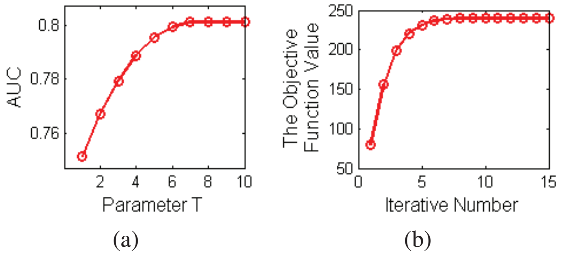


Figure 4: (a) The performance trend against parameter T on the synthetic dataset with POR $= 60\%$. (b) The convergence curve of missing sample recovery in the 4^{th} iteration of (a).

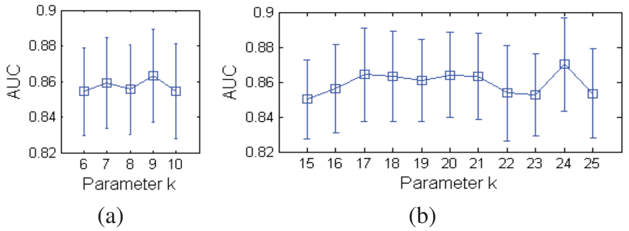


Figure 5: AUC curves with respect to parameter k on (a) synthetic and (b) USPS-MNIST datasets with POR $= 45\%$.

fewer iterations before converging to unified representations.

Regarding parameter k, we limit it to a certain percentage of the total number of objects. It is observed that our proposed method achieves a relatively good performance when the proportion is in the range of $[3\%, 5\%]$. Figure 5 plots the AUC curves w.r.t. parameter k on the synthetic dataset and USPS-MNIST dataset with POR $= 45\%$. The total numbers of objects in the above two datasets are 200 and 500, respectively. Figure 5 can illustrate our observation very well. Similar observations can be concluded for other datasets with different POR settings.

5 Related Work

Outlier detection is a fundamental data analysis technique which aims to identify the abnormal objects in a dataset. Over the past decades, a number of single-view outlier detection methods (Xiong, Chen, and Schneider 2011; Liu, Xu, and Yan 2012; Li, Shao, and Fu 2014; Hu et al. 2016) have been proposed and applied to many fields, such as web spam detection (Castillo et al. 2007), video surveillance (Krausz and Herpers 2010), network failure detection (Ding et al. 2012), and so on. Nowadays, heterogeneity has already been common in data mining applications. Multi-view data is usually collected from diverse domains (Xu, Tao, and Xu 2013; Bai et al. 2016; 2017). Along the above mainline, many single-view approaches have been extended to multi-view scenarios. These methods (Das et al. 2010; Gao et al. 2010; Janeja and Palanisamy 2013; Muller et al. 2012) have achieved impressive performance of detecting outliers that exhibit abnormal behaviors in each view.

Besides, there exist other plotlines (Shekhar, Lu, and Zhang 2002; Sun et al. 2005; Song et al. 2007; Christoudias, Urtasun, and Darrell 2008; Wang and Davidson 2009) to detect different types of outliers from multi-view data. Among them, Horizontal Anomaly Detection (Gao et al. 2011) may be an interesting branch, which detects horizontal outliers that have inconsistent across-view cluster memberships. Gao *et al.* firstly compute spectral embeddings with an ensemble similarity matrix, and then calculate the outlier score with the cosine distance between different embeddings. Subsequent works utilize sophisticated machine learning algorithms to detect inconsistent characteristics for each object, *e.g.*, consensus clustering (Liu and Lam 2012), affinity propagation (Alvarez et al. 2013), and probabilistic latent variable models (Iwata and Yamada 2016). To identify two types of outliers simultaneously, $l_{2,1}$-norm induced error terms are integrated into low-rank subspace learning (Li, Shao, and Fu 2015) and K-means clustering (Zhao and Fu 2015). In general, our paper is closely related to this practical branch. We focus on a real-world case that every view suffers from some missing samples. These previous methods assume that all objects are present in all views, consequently their performance may be dramatically harmed. It is necessary to develop a method as our proposed one in this paper that can work well in partial multi-view scenarios.

6 Conclusion and Future Work

In this paper, we have proposed a novel Collective Learning (CL) based framework for partial multi-view outlier detection. CL is first introduced for missing sample recovery by well exploiting the inter-dependence among different views. An optimization algorithm with closed-form solutions is developed accordingly. Moreover, we propose a similarity-based Hilbert-Schmidt Independence Criterion for outlier detection. Our device alleviates the dilemma that the number of clusters is unknown priori. Then, the above two key operations are iteratively integrated together through a confidence matrix calculated using outlier scores, *i.e.*, learning-based missing sample recovery and similarity-based outlier detection are alternatively performed in a self-guided way. Extensive experiments demonstrate the superior performance of our approach over state-of-the-art multi-view outlier detection methods.

As for future work, there will be more novel frameworks to deal with partial multi-view data. The two distinguishing challenges mentioned in §1 are not solved thoroughly. Regarding our proposed collective learning strategy, it is promising to integrate it into generative models. Besides, collective learning can contribute to the combination of data driven and knowledge driven methods.

Acknowledgments

This work is supported by National Program on Key Basic Research Project No. 2015CB352300 and National Natural Science Foundation of China Major Project No. U1611461. We thank Dr. Hang Su from Tsinghua University for reviewing an earlier version of this paper. Besides, we thank the anonymous reviewers for their helpful comments.

References

Akoglu, L.; Tong, H.; and Koutra, D. 2015. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery* 29(3):626–688.

Alvarez, A. M.; Yamada, M.; Kimura, A.; and Iwata, T. 2013. Clustering-based anomaly detection in multi-view data. In *CIKM*, 1545–1548.

Bai, S.; Sun, S.; Bai, X.; Zhang, Z.; and Tian, Q. 2016. Smooth neighborhood structure mining on multiple affinity graphs with applications to context-sensitive similarity. In *ECCV*, 592–608.

Bai, S.; Zhou, Z.; Wang, J.; Bai, X.; Latecki, L. J.; and Tian, Q. 2017. Ensemble diffusion for retrieval. In *ICCV*, 774–783.

Castillo, C.; Donato, D.; Gionis, A.; Murdock, V.; and Silvestri, F. 2007. Know your neighbors: web spam detection using the web topology. In *SIGIR*, 423–430.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly detection: A survey. *ACM CSUR* 41(3):15:1–15:58.

Christoudias, C. M.; Urtasun, R.; and Darrell, T. 2008. Multi-view learning in the presence of view disagreement. In *UAI*, 88–96.

Das, S.; Matthews, B. L.; Srivastava, A. N.; and Oza, N. C. 2010. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study. In *KDD*, 47–56.

Ding, Q.; Katenka, N.; Barford, P.; Kolaczyk, E.; and Crovella, M. 2012. Intrusion as (anti) social communication: characterization and detection. In *KDD*, 886–894.

Duh, K.; Yeung, C. M. A.; Iwata, T.; and Nagata, M. 2013. Managing information disparity in multilingual document collections. *ACM TSLP* 10(1):1:1–1:28.

Gao, J.; Liang, F.; Fan, W.; Wang, C.; Sun, Y.; and Han, J. 2010. On community outliers and their efficient detection in information networks. In *KDD*, 813–822.

Gao, J.; Fan, W.; Turaga, D.; Parthasarathy, S.; and Han, J. 2011. A spectral framework for detecting inconsistency across multi-source object relationships. In *ICDM*, 1050–1055.

Gao, J.; Fan, W.; Turaga, D.; Parthasarathy, S.; and Han, J. 2013. *A Multi-graph Spectral Framework for Mining Multi-source Anomalies*. Springer: New York.

Garavan, T. N., and Carbery, R. 2012. *Collective Learning*. Springer.

Gretton, A.; Bousquet, O.; Smola, A.; and Scholkopf, B. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, 63–77.

Hu, R.; Aggarwal, C. C.; Ma, S.; and Huai, J. 2016. An embedding approach to anomaly detection. In *ICDE*, 385–396.

Hull, J. J. 1994. A database for handwritten text recognition research. *TPAMI* 16(5):550–554.

Iwata, T., and Yamada, M. 2016. Multi-view anomaly detection via robust probabilistic latent variable models. In *NIPS*, 1136–1144.

Janeja, V. P., and Palanisamy, R. 2013. Multi-domain anomaly detection in spatial datasets. *Knowledge and Information Systems* 36(3):749–788.

Krausz, B., and Herpers, R. 2010. Metrosurv: detecting events in subway stations. *Multimedia Tools and Applications* 50(1):123–147.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haaffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–2324.

Li, S.; Jiang, Y.; and Zhou, Z. 2014. Partial multi-view clustering. In *AAAI*, 1968–1974.

Li, S.; Shao, M.; and Fu, Y. 2014. Locality linear fitting one-class SVM with low-rank constraints for outlier detection. In *IJCNN*, 676–683.

Li, S.; Shao, M.; and Fu, Y. 2015. Multi-view low-rank analysis for outlier detection. In *SDM*, 748–756.

Liu, A. Y., and Lam, D. N. 2012. Using consensus clustering for multi-view anomaly detection. In *IEEE Symp. SPW*, 117–124.

Liu, G.; Xu, H.; and Yan, S. 2012. Exact subspace segmentation and outlier detection by low-rank representation. In *AISTATS*, 703–711.

Muller, E.; Assent, I.; Sanchez, P. I.; Mulle, Y.; and Bohm, K. 2012. Outlier ranking via subspace analysis in multiple views of the data. In *ICDM*, 529–538.

Nilsback, M. E., and Zisserman, A. 2006. A visual vocabulary for flower classification. In *CVPR*, 1447–1454.

Shao, W.; He, L.; and Yu, P. S. 2015. Multiple incomplete views clustering via weighted NMF with $l_{2,1}$ regularization. In *ECML/PKDD*, 318–334.

Shekhar, S.; Lu, C. T.; and Zhang, P. 2002. Detecting graph-based spatial outliers. *Intelligent Data Anal.* 6(5):451–468.

Song, X.; Wu, M.; Jermaine, C.; and Ranka, S. 2007. Conditional anomaly detection. *TKDE* 19(5):631–645.

Sun, J.; Qu, H.; Chakrabarti, D.; and Faloutsos, C. 2005. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, 418–425.

Wang, X., and Davidson, I. 2009. Discovering contexts and contextual outliers using random walks in graphs. In *ICDM*, 1034–1039.

Wang, K.; Yin, Q.; Wang, W.; Wu, S.; and Wang, L. 2016. A comprehensive survey on cross-modal retrieval. *arXiv:1607.06215*.

Xiong, L.; Chen, X.; and Schneider, J. 2011. Direct robust matrix factorization for anomaly detection. In *ICDM*, 844–853.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv:1304.5634*.

Zhao, H., and Fu, Y. 2015. Dual-regularized multi-view outlier detection. In *IJCAI*, 4077–4083.

Zhao, H.; Liu, H.; and Fu, Y. 2016. Incomplete multi-modal visual data grouping. In *IJCAI*, 2392–2398.