

# Discovering Contexts and Contextual Outliers Using Random Walks in Graphs

Xiang Wang\*     Ian Davidson†

## Abstract

The identifying of contextual outliers allows the discovery of anomalous behavior that other forms of outlier detection cannot find. What may appear to be normal behavior with respect to the entire data set can be shown to be anomalous by subsetting the data according to specific spatial or temporal context. However, in many real-world applications, we may not have sufficient *a priori* contextual information to discover these contextual outliers. This paper addresses the problem by proposing a probabilistic approach based on random walks, which can **simultaneously** explore meaningful contexts and score contextual outliers therein. Our approach has several advantages including producing outlier scores which can be interpreted as stationary expectations and their calculation in closed form in polynomial time. In addition, we show that point outlier detection using the stationary distribution is a special case of our approach. This allows us to find both global and contextual outliers simultaneously and create a meaningful ranked lists consisting of **both** types of outliers. This is a major departure from existing work where an algorithm typically produces one type of outlier. The effectiveness of our method was justified by empirical results on real data sets, with comparison to previous work.

---

\*Department of Computer Science, University of California, Davis.  
xiang@ucdavis.edu

†Department of Computer Science, University of California, Davis.  
davidson@cs.ucdavis.edu

## 1 Introduction

### 1.1 Motivation

Outlier detection, also called anomaly detection, is an important but understudied branch of the data mining research: only recently did the first data mining survey on this topic become available [2]. Conventionally, the goal of outlier detection is to find data instances that do not conform to the *normal* behavior, which is typically defined by the **entire** data set. Many methods have been proposed in the literature and achieved success in numerous real-world applications, such as network intrusion detection, fraud detection, and medical informatics, to name just a few [2, 6].

Most of the existing approaches identify outliers from a global point of view, where each data instance is examined as deviating from normality as defined by the entire data set. This type of outlier detection is called *global outlier detection* [2]. However, sometimes an instance may not be an outlier when compared to the rest of the data set but maybe an outlier in the context of a subset of the data. This type of outlier detection is called *contextual outlier detection*, where the subset with respect to which the outlier is examined is called the *context*. For example, in a population demographic data set, a six-foot person may not be anomalous, but in the context of individuals aged under ten years of age would be an outlier.

As compared to global outlier detection, contextual outlier detection is even more understudied [2]. A major challenge of contextual outlier detection is identifying the contexts which then allow the identification of outliers. A data instance may appear anomalous in one context but not in others. Therefore, the meaningfulness of the context essentially decides the interestingness of the contextual outliers. In order to define the proper contexts, ex-

isting contextual outlier detection techniques require the user to *a priori* specify the contextual information, or the contextual attributes, in the data set. Typical contextual attributes used by previous work include partition labels [16], spatial and/or temporal information [13, 14, 9], adjacency in graphs [17], and profiles [5].

Unfortunately, the existing approaches on contextual outlier detection, though very useful, have two limitations. Firstly the *a priori* contextual information is not always available in practice. For example, a cellphone company wants to find anomalous user behavior by analyzing the user logs. But the spatial information of the users is part of the users' privacy and cannot be provided (explicitly) to a third-party analyst. Secondly, even if we have well-defined contexts within the data set, it is nontrivial to find contextual outliers therein. A naïve approach is to firstly partition the data set into individual contexts and then find outliers therein, separately and respectively, using traditional global outlier detection techniques. However, this **divide and conquer** style approach may not work since by partitioning the data set, some important structural information could be lost as we shall see in the example below.

Since defining contexts and detecting contextual outliers are mutually dependent, a logical extension of existing contextual outlier detection work is to **fold identifying the contexts into the outlier detection question itself** by asking under which natural contexts do outliers occur. Our work explores this more elaborate question.

**Example** Assume we want to find contextual outliers in the graph as shown in Fig. 1(a). There are two obviously contexts in the graph, namely  $\{1, 2, 3, 4\}$  and  $\{5, 6, 7, 8\}$ . Node 4 and 5 are contextual outliers, considering that their connectivity is different from the majority of nodes in their respective contexts. However, if we cut the graph first and remove the edge  $(4, 5)$ , the connectivity of node 4 and 5 will become the same as that of the other nodes in their respective contexts (Fig. 1(b)). Consequently, we can no longer find any type of outliers in either subgraph.

## 1.2 An Overview of Our Work

We propose a graph based random walk model that can formally define contexts and contextual outliers therein.

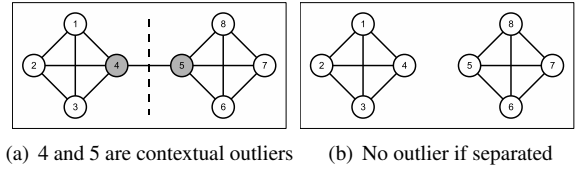


Figure 1: Contextual outlier detection  $\neq$  outlier detection in (separated) contexts.

Modeling the problem using graph model does not limit our work to graph based data such as social networks. Our work is applicable so long as a transition (probability) matrix can be generated from the data set. For example, we can build a random walk graph by representing each data instance as a node and by converting the similarity between data instances into a transition probability between the two nodes.

The main technical result of the work is as follows. Given a random walk graph we can easily determine a stationary probability distribution which describes a random walk's steady state behavior, regardless of where the walk starts. This steady state probabilities are easily calculated as the *principal eigenvector* of the transition matrix of the random walk graph and others have highlighted the least visited nodes are naturally global outliers [11]. Our technical contribution is to interpret the *non-principal eigenvectors* to help identify contextual outliers. Each and every non-principal eigenvector provides a cut of the graph as is well-known in the spectral clustering literature [18]. We illustrate how to interpret these eigenvectors as the stationary expectations of a random indicator variable. The indicator variable effectively calculates the difference in the chance of a node being visited if the random walk starts off in one subgraph (defined by respective non-principal eigenvectors) as opposed to starting in the other subgraph. Then we can use the stationary expectation as a contextual outlier score to identify contextual outliers as nodes that are (almost) equally likely to be visited by random walks starting from **either** subgraph. Conversely, we can identify contextual *inliers* if the chance of visiting a node is much larger when the random walk starts from a particular context (subgraph) as opposed to the other. It is important to note that our work identifies contexts and contextual outliers by performing random walks in the **entire graph**. We do not remove any edge or alter the

structure of the original graph in any way.

Based on our random walk graph model, we develop an algorithm using eigendecomposition, which can automatically and simultaneously find different contexts in the data set and rank all the data instances by a probabilistically principled outlier score with respect to their contexts.

To sum up, our contributions are:

1. To the best of our knowledge, this is the first work to find contextual outliers without *a priori* contextual information by automatically discovering the contexts.
2. We provide a flexible method of finding the contexts and the contextual outliers, which is applicable to both graphical and vector data.
3. We create an easily interpretable contextual outlier score for any node as being the chance that a random walk **on the entire graph** starting from both contexts, respectively, will visit the node. This allows us to meaningfully rank both **global** and **contextual** outliers.
4. We propose an **efficient** polynomial time algorithm based on eigendecomposition that can automatically and simultaneously find contexts as well as contextual outliers that are **interpretable from a probabilistic perspective**.

The rest of the paper is organized as follows: In Section 2 we outline the preliminaries and some well-known results so that in Section 3 we can establish a theoretical relationship between our contextual outlier score and the stationary expectation of a node being visited under a contextual random walk. The notion and analysis of a contextual random walk is novel. In Section 4 we outline our algorithm that iteratively explores contexts and contextual outliers in the graph, in a hierarchical fashion. In Section 5 we evaluate our algorithm on several real-world data sets. Empirical results validated its effectiveness and advantage against existing approach. Finally in Section 6 we outline related work and conclude our contributions in Section 7.

Table 1: Symbols and their meanings

$W$	The transition matrix of a random walk
$\mathbf{u}$	The principal eigenvector of $W$
$\mathbf{v}$	A non-principal eigenvector of $W$
$\pi$	The stationary distribution of a global random walk
$\mu$	The stationary expectation of a contextual random walk

## 2 Background and Preliminaries

In this section, we survey some previous results related to random walk graph for completeness and to introduce notation. We first introduce the notion of a random walk graph, which is essentially a homogeneous Markov chain as characterized by a transition matrix. We show the well-known result that the principal eigenvector of the transition matrix gives the stationary distribution of the nodes being visited in the graph under a global random walk. We then survey the previous work [11] that uses the principal eigenvector to score and rank global outliers. Readers who are familiar with these materials can skip to Definition 2. Symbols used throughout the rest of the paper are summarized in Table 1. It is important to note that the matrix  $W$  we analyze throughout the paper is *not* a Graph Laplacian [18], but rather a transition matrix.

### 2.1 Random Walk on Graphs

Given a data set  $\mathcal{D}$  with  $n$  data instances  $\{d_1, \dots, d_n\}$ ,  $f$  is a similarity function defined as:

$$f : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}^+. \quad (1)$$

We assume that  $f$  is symmetric yet does not necessarily satisfy the triangle inequality. Then we can model  $\mathcal{D}$  into a Markov random walk graph. Specifically, let  $A = (a_{ij})$  be the similarity matrix where  $a_{ij} = a_{ji} = f(d_i, d_j)$ ,  $\forall i, j, 1 \leq i, j \leq n$ . We can construct a random walk graph  $G = (V, E)$  from  $A$  as follows. Each node in  $G$  represents a data instance in  $\mathcal{D}$  and the directed edge from node  $i$  to node  $j$  means that the transition from node  $i$  to  $j$  happens with the probability as specified by a **transition**

**matrix**  $W$ . Let  $D$  be a diagonal matrix where

$$d_{ij} = \begin{cases} \sum_{i=1}^n a_{ij} & i = j \\ 0 & i \neq j \end{cases} \quad (2)$$

Then the transition matrix  $W$  is defined as

$$W = AD^{-1}. \quad (3)$$

The entries in  $W$  represents the probability of the transition from one node in  $G$  to another. Formally:

$$w_{ij} = p(X^{t+1} = i | X^t = j), \forall i, j, 1 \leq i, j \leq n, \quad (4)$$

where  $X^t \in \{1, \dots, n\}$  is the state random variable of the Markov chain. Note that we assume the Markov chain is time-homogeneous, i.e.  $w_{ij} = p(X^{t+1} = i | X^t = j)$  remains the same for any time  $t \geq 0$ .

## 2.2 Principal Eigenvector and Stationary Distribution

In this section we relate the principal eigenvector of a transition matrix to the stationary distribution of a random walk. Consider the eigendecomposition of the transition matrix  $W$ :

$$W\mathbf{u}_i = \lambda_i\mathbf{u}_i, \forall i, 1 \leq i \leq n, \quad (5)$$

where  $\mathbf{u}_i$  is the  $i$ -th eigenvector associated with eigenvalue  $\lambda_i$ . According to the Perron-Frobenius Theorem [7], the following two properties hold:

**Property 1.** *If we sort the eigenvalues in descending order, we have*

$$1 \geq \lambda_1 \geq \dots \geq \lambda_n \geq -1. \quad (6)$$

**Property 2.** *Given a transition matrix  $W$  and all its entries strictly positive, there exists an eigenvector  $\mathbf{u}$  associated with the largest eigenvalue 1, whose entries satisfy*

$$\mathbf{u}(i) > 0, \forall i, 1 \leq i \leq n, \quad (7)$$

where  $\mathbf{u}(i)$  is the  $i$ -th entry of  $\mathbf{u}$ ; and

$$\sum_{i=1}^n \mathbf{u}(i) = 1. \quad (8)$$

We call  $\mathbf{u}$  the (normalized) principal eigenvector of  $W$ .

The stationary distribution is a time-invariant measure that characterizes the behavior of a Markov random walk. Specifically, given a  $n$ -state time-homogeneous Markov random walk, as defined by the  $n \times n$  transition matrix  $W$ , we can define its stationary distribution as follows:

**Definition 1** (Stationary Distribution). *Let  $\pi = (\pi_1, \dots, \pi_n)^T$ , where  $\pi_i = p(X^t = i)$  is the probability of node  $i$  being visited by a random walk (as defined by  $W$ ) at time  $t$ . If at any time  $t \geq 0$ ,  $\pi$  satisfies:*

$$\pi_i = \sum_{j=1}^n \pi_j w_{ij}, \forall i, 1 \leq i \leq n, \quad (9)$$

then  $\pi$  is called the stationary distribution of the random walk.

It is well-known that the stationary distribution of a given random walk can be derived from the (normalized) principal eigenvector of the transition matrix. Formally, given a strictly positive transition matrix  $W$ , from Property 2 and Definition 1, we have

**Property 3.** *Given a strictly positive transition matrix  $W$ , the stationary distribution of the random walk is equal to the (normalized) principal eigenvector of the transition matrix  $W$ :*

$$\pi = \mathbf{u}. \quad (10)$$

## 2.3 Global Outlier and Stationary Distribution

Intuitively, given a (global) random walk in graph  $G$ , the less likely a node is visited by the random walk, the more likely it is a (global) outlier. Therefore previous work [11] used the stationary distribution as the global outlier score. Formally:

**Definition 2** (Global Outlier Score). *Given a random walk graph  $G$  and its transition matrix  $W$ ,  $\pi_i$  is the global outlier score for node  $i$ ,  $\forall i, 1 \leq i \leq n$ .*

The smaller the score is, the more likely node  $i$  is a (global) outlier. Informally, the global outlier score of a node is the chance that a random walk in the graph will visit that node.

The limitation of Definition 2 is that the stationary distribution only identifies outliers based on a global random

walk, where all the nodes are treated indifferently. It is nontrivial to extend the notion of a stationary distribution to contextual outlier detection, where different nodes in the graph may belong to different contexts. In the following section, we will show how to score contextual outliers using the **stationary expectation**.

### 3 Contextual Random Walks and Contextual Outliers

In the previous section we discussed finding global outliers using the global random walks, where an outlier is a node that is unlikely to be visited regardless of where the random walk starts. Though useful, this approach cannot identify contextual outliers since no contextual information is present or used. We now discuss our approach which can identify contextual outliers using the **non-principal eigenvectors** of a transition matrix and interpret them as the stationary expectation of contextual random walks.

In our model, each non-principal eigenvector of the transition matrix uniquely defines a 2-labeling/2-coloring of the graph. Intuitively, given a 2-coloring of the graph, each subgraph can be considered as a context. Let  $S^+$  be one subgraph and  $S^-$  the other, we can then determine the chance of a node being visited given the random walk starts from  $S^+$  and  $S^-$ , respectively. Without loss of generality, if a node in  $S^+$  is much more likely to be visited by the random walk starting from  $S^+$  than from  $S^-$ , then it can be considered as a contextual inlier w.r.t.  $S^+$ . On the other hand, there will be some unusual nodes whose chance of being visited by the random walk starting from either  $S^+$  or  $S^-$  is **about the same**, i.e., these nodes don't belong strongly to either  $S^+$  or  $S^-$ . We call these nodes **contextual outliers**. We assign contextual outlier scores to every node in the graph so that the contextual outliers can be discovered.

#### 3.1 Contextual Random Walk and Stationary Expectation

We first introduce the definitions and properties of the contextual random walk and the stationary expectation. Assume  $G$  is a random walk graph associated with a

strictly positive transition matrix  $W$ . First we define contexts in a graph:

**Definition 3** (Contexts and Contextual Random Walk). *Let  $(S^+, S^-)$  be a 2-coloring of  $G$ , where  $S^+$  is the index set of nodes labeled as + while  $S^-$  is the index set of nodes labeled as -.  $S^+$  and  $S^-$  satisfy*

$$S^+ \neq \emptyset, S^- \neq \emptyset, S^+ \cup S^- = \{1, \dots, n\}. \quad (11)$$

*We call  $(S^+, S^-)$  a pair of contexts of the graph  $G$ . A random walk in  $G$  with the existence of contexts is then called a contextual random walk.*

Now we consider the following indicator random variable:

$$Y_i^t = \begin{cases} 1 & X^t = i, X^0 \in S^+ \\ -1 & X^t = i, X^0 \in S^- \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

where  $1 \leq i \leq n$  and  $t = 0, 1, 2, \dots$ . To put it into words, if node  $i$  is visited by a random walk at time  $t$ , and if the random walk started from the context  $S^+$ , we set  $Y_i^t$  to 1; if the random walk started from  $S^-$ , we set  $Y_i^t$  to -1; if node  $i$  is not visited at time  $t$ , then we set  $Y_i^t$  to 0. We calculate the mathematical expectation of  $Y_i^t$  as follows:

$$\mathbf{E}(Y_i^t) = p(X^t = i, X^0 \in S^+) - p(X^t = i, X^0 \in S^-). \quad (13)$$

Consequently, if  $\mathbf{E}(Y_i^t)$  is (relatively) close to 1 (or -1), it indicates that node  $i$  is more likely to be visited by the random walk starting from  $S^+$  (or  $S^-$ ). However, if the expectation is close to 0, it means that node  $i$  is (almost) equally likely to be visited by a random walk starting from either context, which effectively makes the node anomalous as compared to the other nodes which are more strongly "aligned" with  $S^+$  or  $S^-$ .

Though  $\mathbf{E}(Y_i^t)$  is informative for identifying contextual outliers, it cannot be used as a contextual outlier score directly as it has the problem of being time-dependent, i.e., it is **not** a constant and always changes as  $t$  increases. Therefore we introduce a time-invariant measure which can better help characterizing the structure of the random walk graph and identifying contextual outliers. The time-invariant measure is, similar to the stationary distribution of a global random walk, the stationary expectation of a contextual random walk. Though related, these two

measures should be interpreted differently. The stationary probability distribution measures the chance of visiting a node regardless of where the random walk starts. Whereas the stationary expectation (as defined below) measures the **difference** in the chances of a node being visited by random walks starting from  $S^+$  and  $S^-$ , respectively. First we define the stationary expectation of a contextual random walk:

**Definition 4** (Stationary Expectation). *Given the random walk graph  $G$  and its transition matrix  $W$ , we say the expectation of  $Y_i^t$ , which is  $\mu_i$ , is stationary if for all  $t$  the following condition holds:*

$$\mu_i = c \sum_{j=1}^n \mu_j w_{ij}, \quad \forall i, 1 \leq i \leq n. \quad (14)$$

where  $c$  is a time-independent constant. We shall refer to  $\mu = (\mu_1, \dots, \mu_n)^T$  as the stationary expectation of the contextual random walk in  $G$  w.r.t.  $S^+$  and  $S^-$ .

Now the question becomes how we can find a stationary expectation  $\mu$  given the transition matrix  $W$ . We will show that if  $W$  is strictly positive, each of its non-principal eigenvectors uniquely determines a pair of contexts and the corresponding stationary expectation. Specifically, following Property 1, let  $\mathbf{v}$  be an eigenvector of  $W$  associated with the eigenvalue  $\lambda < 1$ . We call  $\mathbf{v}$  a non-principal eigenvector of  $W$  and the following lemma holds:

**Lemma 1.** *Given a non-principal eigenvector  $\mathbf{v}$  of a strictly positive transition matrix  $W$ , we have*

$$\sum_{i=1}^n \mathbf{v}(i) = 0, \quad (15)$$

where  $\mathbf{v}(i)$  is the  $i$ -th entry of  $\mathbf{v}$ .

*Proof.* The proof is trivial and omitted due to page limit. Please refer to textbooks on spectral analysis, say [3].  $\square$

With Lemma 1, we can use  $\mathbf{v}$  to define a 2-coloring of  $G$ , which gives us a pair of contexts:

$$S^+ = \{i : \mathbf{v} > 0\}, \quad S^- = \{i : \mathbf{v} < 0\}. \quad (16)$$

Now consider the contextual random walk in  $G$  w.r.t.  $(S^+, S^-)$ , we have the following theorem:

**Theorem 1** (The Stationary Expectation of a Contextual Random Walk). *If we set  $\mu = (\mu_1, \dots, \mu_n)^T$  to be*

$$\mu_i = \frac{\mathbf{v}(i)}{\sum_{j=1}^n |\mathbf{v}(j)|}, \quad \forall i, 1 \leq i \leq n, \quad (17)$$

where  $\mathbf{v}$  is a non-principal eigenvector of  $W$  associated with the eigenvalue  $\lambda$ , then Eq.(14) will hold. Hence  $\mu$  as defined in Eq.(17) is a stationary expectation of the contextual random walk.

*Proof.* First we need to show  $\mu_i$  is a valid expectation of  $Y_i^t$ , i.e. it should be achievable under proper initial conditions. Specifically, let

$$p(X^t = i, X^0 \in S^+) = \begin{cases} \frac{\mathbf{v}(i)}{\sum_{j=1}^n |\mathbf{v}(j)|} & i \in S^+ \\ 0 & i \in S^- \end{cases} \quad (18)$$

and

$$p(X^t = i, X^0 \in S^-) = \begin{cases} 0 & i \in S^+ \\ -\frac{\mathbf{v}(i)}{\sum_{j=1}^n |\mathbf{v}(j)|} & i \in S^- \end{cases} \quad (19)$$

where  $(S^+, S^-)$  are defined after Eq.(16). The initial probabilities in Eq.(18) and (19) are valid because:

$$\begin{aligned} & \sum_{i=1}^n (p(X^t = i, X^0 \in S^+) + p(X^t = i, X^0 \in S^-)) \\ &= \sum_{i \in S^+} (p(X^t = i, X^0 \in S^+) + p(X^t = i, X^0 \in S^-)) \\ & \quad + \sum_{i \in S^-} (p(X^t = i, X^0 \in S^+) + p(X^t = i, X^0 \in S^-)) \\ &= \sum_{i \in S^+} p(X^t = i, X^0 \in S^+) + \sum_{i \in S^-} p(X^t = i, X^0 \in S^-) \\ &= \frac{\sum_{i \in S^+} \mathbf{v}(i)}{\sum_{j=1}^n |\mathbf{v}(j)|} + \frac{-\sum_{i \in S^-} \mathbf{v}(i)}{\sum_{j=1}^n |\mathbf{v}(j)|} \\ &= \frac{\sum_{i=1}^n |\mathbf{v}(i)|}{\sum_{j=1}^n |\mathbf{v}(j)|} = 1 \end{aligned} \quad (20)$$

Following Eq.(18) and (19), the expectation of  $Y_i^t$  becomes:

$$\begin{aligned} E(Y_i^t) &= p(X^t = i, X^0 \in S^+) - p(X^t = i, X^0 \in S^-) \\ &= \frac{\mathbf{v}(i)}{\sum_{j=1}^n |\mathbf{v}(j)|} = \mu_i, \end{aligned}$$

$\forall i, 1 \leq i \leq n.$

Therefore  $\mu$  is a valid expectation of  $Y_i^t$ , by assuming the contextual random walk starts with probabilities shown in Eq.(18) and (19). Next we show that it satisfies Eq.(14). Since  $\mathbf{v}$  is an eigenvector associated with the eigenvalue  $\lambda$ ,  $\mu$  as defined in Eq.(17) is also an eigenvector associated with  $\lambda$ , thus at any time  $t \geq 0$  we have

$$\sum_{j=1}^n w_{ij} \mu_j = \lambda \mu_i,$$

which can be rewritten as

$$\mu_i = c \sum_{j=1}^n w_{ij} \mu_j, \quad (21)$$

where  $c = 1/\lambda$  is time-invariant. Hence Eq.(14) holds and  $\mu$  is a stationary expectation of the contextual random walk w.r.t.  $S^+$  and  $S^-$ .  $\square$

Theorem 1 shows that each non-principal eigenvector uniquely determines a 2-coloring of the graph,  $(S^+, S^-)$ , and its stationary expectation,  $\mu$ .

### 3.2 Contextual Outlier and Stationary Expectation

With Theorem 1, we can now define the contextual outlier score using the stationary expectation.

**Definition 5** (Contextual Outlier Score). *Given a random walk graph associated with the transition matrix  $W$ , the contextual outlier score of node  $i$  is  $|\mu_i|$ , where  $\mu_i$  is the stationary expectation as defined in Eq.(17), w.r.t. the contexts  $(S^+, S^-)$  as defined in Eq.(16).*

According to our definition, the contextual outlier score of any node is always between 0 and 1. A large score means that the node is highly expected to be visited by a random walk starting from one of the two contexts, and is thus a contextual *inlier*; a small score means that the node is equally likely to be visited by the random walk starting from either context, and is thus a contextual *outlier*.

Our contextual outlier score is time-invariant, and is solely determined by the structure of the random walk graph. Note that since the transition matrix  $W$  has  $n - 1$  non-principal eigenvectors, thus we can potentially have

$n - 1$  pairs of contexts, and we can compute for every node in the graph a contextual outlier score w.r.t. each pair of contexts.

An important advantage of our contextual outlier score is that it covers the global outlier score (based on the stationary distribution) as a special case. Formally, we have the following corollary:

**Corollary 1.** *The stationary distribution  $\pi$  is a special case of stationary expectation, where  $\lambda = 1$  and  $S^+ = \{1, \dots, n\}$ ,  $S^- = \emptyset$ .*

*Proof.* Let

$$p(X^t = i, X^0 \in S^+) = \frac{\mathbf{v}(i)}{\sum_{j=1}^n \mathbf{v}(j)}, \quad \forall i, 1 \leq i \leq n. \quad (22)$$

Then  $\pi$  is a valid expectation of  $Y_i^t$  and apparently we have

$$\pi_i = \sum_{j=1}^n w_{ij} \pi_j. \quad (23)$$

$\square$

Corollary 1 says that we can re-interpret the global outlier score within our framework and compare it directly to our contextual outlier score. Consequently, we can produce a unified ranked list containing both global outliers and contextual outliers, ordered by their anomalousness.

### 3.3 Remarks

**Choosing Contexts** As we mentioned above, any strictly positive transition matrix  $W$  has  $n - 1$  non-principal eigenvectors, each of which decides a pair of contexts in the random walk graph. Consequently a node would have different contextual outlier scores w.r.t. different contexts. However, in certain applications, we are more interested in eigenvectors associated with larger eigenvalues, especially the one associated with the second largest eigenvalue, which is sometimes called the *Fiedler eigenvector*. Previous work shows that the bi-partition of a random walk graph determined by the Fiedler eigenvector corresponds to the *normalized Min-Cut* of the graph [18]. Specifically, following Eq.(16), the Fiedler eigenvector partitions the graph into two most separated subgraphs, which are more interesting to study because the

two most separated subgraphs normally imply the two most obvious contexts in the random walk graph.

**Extension of Our Model** For the simplicity of formulation, we have always assumed that the transition matrix  $W$  is strictly positive, which makes  $G$  a complete graph. However, with some trivial modifications, our model can be extended to more general cases. For example, let  $A$  be an affinity graph, where a node is only connected to its  $k$  nearest neighbors. As a result the random walk graph will not be a complete graph, or not even connected. In this case, assuming it has  $p$  connected components, we can still construct the transition matrix  $W$ , and among its  $n$  eigenvalues, we will have  $\lambda_1 = \dots = \lambda_p = 1$ . The nonzero entries in  $\mathbf{v}_i$  are the nodes belong to the  $i$ -th connected component of the graph. Each of the  $(n - p)$  non-principal eigenvectors will define a 2-coloring of one of the  $p$  connected components. Therefore our model will simply treat each connected component as a graph and score outliers therein. Detailed formulation is omitted from this paper due to space limit.

## 4 Algorithm

In this section, we discuss the implementation of our contextual outlier score in practice. We propose a hierarchical algorithm which iteratively partitions the data set until the size of the subgraph is smaller than a user-specified threshold  $\alpha$ . Both global and contextual outliers are detected and ranked during each iteration. The outline of our algorithm is shown in Algorithm 1.

The input of our algorithm is a graph  $G$  and its associated transition matrix  $W$ . The transition matrix  $W$  is generated by normalizing a given similarity matrix  $A$ , where  $a_{ij}$  is the similarity between the  $i$ -th and  $j$ -th data instances. The choice of similarity function is application-dependent. In our experiments we show that promising results are obtained using the Euclidean distance as well as the inner product.

While we hierarchically partition the graph  $G$  into smaller subgraphs, we use a queue  $Q$  to store the subgraphs to be partitioned. A user-specified threshold  $\alpha$  is used to decide when we stop to further partition a subgraph, since as the subgraph becomes smaller, it's less likely to have meaningful contexts within itself.

---

### Algorithm 1: Hierarchical contextual outlier detection

---

**Input:** Random walk graph  $G$  with transition matrix  $W$ , queue  $Q$ , threshold  $\alpha$ ;  
**Output:** A sorted list  $L$ , consisting of tuples as defined in Eq.(24);

```

1  $Q \leftarrow \emptyset$ ;  $Q.enqueue(G, W)$ ;
2  $L \leftarrow \emptyset$ ;
3 repeat
4    $(G, W) \leftarrow Q.dequeue()$ ;
5   if  $|G| > \alpha$  then
6     Compute the (normalized) principal
7     eigenvector of  $W$ , which is  $\mathbf{u}$ ;
8     foreach  $i \in G$  do
9       Add  $\{i, G, \mathbf{u}(i)\}$  to  $L$ ; /* global
10      outliers */
11     end
12     Compute the Fiedler eigenvector of  $W$ ,
13     which is  $\mathbf{v}$ ;
14      $S^+ \leftarrow \{i : \mathbf{v}(i) > 0\}$ ,  $S^- \leftarrow \{i : \mathbf{v}(i) < 0\}$ ;
15     foreach  $i \in S^+$  do
16       Add  $\{i, S^+, |\mathbf{v}(i) / \sum_{j=1}^n \mathbf{v}(j)|\}$  to  $L$ ;
17       /* contextual outliers in
18        $S^+$  */
19     end
20     foreach  $i \in S^-$  do
21       Add  $\{i, S^-, |\mathbf{v}(i) / \sum_{j=1}^n \mathbf{v}(j)|\}$  to  $L$ ;
22       /* contextual outliers in
23        $S^-$  */
24     end
25     Generate the transition matrices for  $S^+$  and
26      $S^-$ , respectively;
27      $Q.enqueue(S^+, W^+)$ ;
28      $Q.enqueue(S^-, W^-)$ ;
29   end
30 until  $Q$  is empty;

```

---

The output of our algorithm is a ranked outlier list  $\mathbb{L}$ , whose entries are tuples in the form of

$$\{\text{instance}, \text{context}, \text{score}\}, \quad (24)$$

where *instance* is the index of the data instance; *context* is the context with respect to which that data instance is examined; *score* is the outlier score of that data instance. Note that one instance may appear more than once in  $\mathbb{L}$  because it has different outlier scores with respect to different contexts.

Our algorithm involves computing the first and second largest eigenvalues and eigenvectors of an  $n \times n$  matrix, where  $n$  is the number of data instances. Therefore its complexity is dominated by the complexity of eigendecomposition. Note that if the transition matrix is generated from the  $k$ -nearest-neighbor graph, then it will be very sparse when  $k$  is small, which leads to much faster eigendecomposition.

## 5 Empirical Study

### 5.1 Illustrative Example

We first illustrate that our method is able to find meaningful contexts and contextual outliers therein. Fig. 2(a) is a random headshot image. We converted this image into a graph such that each node corresponds to a pixel in the image; for each pixel we extracted its RGB color as well as its position; then we built the transition matrix based on the Euclidean distance between the color and spatial information of two nodes. Our method discovered two most stable/meaningful contexts in the image: the face and the background. As shown in Fig. 2(b), our method also identified contextual outliers in the face context, including the glasses (dark), the eyes (dark), the cheeks (highlight), the forehead (highlight), etc. This means that a random walk starting on the face is unlikely to visit those areas since they are locally different to their surroundings. As a contrast, there is no significant outlier in the background context, because the color of the background is uniform (Fig. 2(c)). We can see that our method captured natural contexts and scored contextual outliers therein in a way that conforms well to human observation.

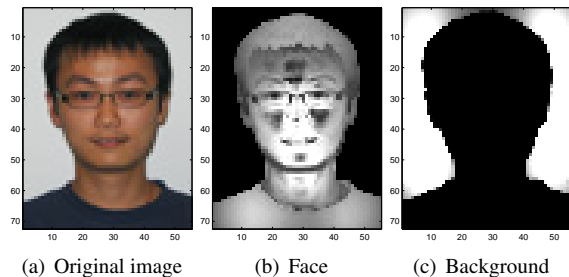


Figure 2: Contexts and contextual outliers in a headshot image. Dark pixels are outliers.

## 5.2 Quantitative Evaluation

### 5.2.1 Methodology

The quantitative evaluation of contextual outlier detection itself is an open problem because there is no commonly accepted ground truth for contextual outliers. An easy way to generate such ground truth is to use synthetic data, where we insert outliers based on human intuition or domain knowledge. However, synthetic data could often differ too much from real data. Therefore, instead of synthetic data, we generated ground truth using the class labels of real-world data sets. Specifically, given a data set with class labels, we first convert it into a random walk graph. Then we discover two contexts within it using the 2-coloring indicated by the Fiedler eigenvector. Note that this partition is equivalent to a normalized Min-Cut of the graph. Next we label each context by the label of the majority class in that context. The minority class in each context is then labeled as *true contextual outliers*. This can be interpreted as the contextual outliers being the instances most likely to contain class label noise. The ground truth contexts are given by the class labels and contextual outliers are those instances most likely not to be of this class. Note that due to the limited availability of ground truth (labels), we did not let the threshold  $\alpha$  to decide when the iterative partition should end: the number of partition performed was decided by the number of classes in the data set.

We apply our method to rank contextual outliers in each context, respectively, and compare our answer to the ground truth. Hence we essentially turn the outlier detection problem into a retrieval problem, where

we can compute the precision, recall and f-score of our method. Specifically, *precision* is the fraction of true contextual outliers among the contextual outliers found by our method; *recall* is the fraction of contextual outliers found by our method among all true contextual outliers; and *f-score* is the average of precision and recall, defined as  $f\text{-score} = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$ .

We also implemented a baseline method for comparison. It uses the same contexts defined by the Fiedler vector. But instead of computing the contextual outlier score, it ranks global outliers within each context, separately, using the stationary distribution based method as described in [11]. Note that we chose this method instead of popular outlier detection techniques say LOF[1] because this method adopted the same random walk graph model as the one used by us, and it ranks outlier in a probabilistic framework, thus is ready for direct comparison with our algorithm.

### 5.2.2 Results and Analysis

The first data set we used was *Iris* from the UCI Archive. It has 3 classes: *setosa*, *versicolor* and *virginica*. Each class has 50 instances and each instance has 4 attributes. For visualization purposes, we projected the data set onto a 2-dimensional space, using the standard PCA technique [4]. Then we converted the data set into a transition matrix using Euclidean distance.

When we applied our method to discover contexts, we noticed that the *setosa* class can be perfectly partitioned from the remaining 2 classes, which means that there’s no contextual outliers in these contexts (*setosa* vs. the remaining two). Thus we removed *setosa* and continued to partition the rest of data (Fig. 3(a)). As a result, the first context contained 54 instances, among which 43 were *versicolor* and 11 *virginica*. Thus the first context was labeled as *versicolor* and had 11 true contextual outliers. Similarly, the other context was labeled as *virginica* and had 7 true contextual outliers (Fig. 3(b)).

We scored contextual outliers using our method (Contextual Outlier Detection, COD) as well as the baseline method (Baseline), respectively. Both methods reported the top-10 contextual outliers from each context. We can clearly see in Fig. 3(c) that our method effectively identified most of the contextual outliers, while the baseline

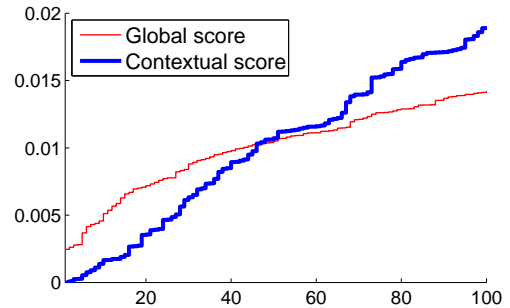


Figure 5: Global and contextual outlier scores for all data instances in the trimmed *iris* data set, sorted in ascending order.

method tended to report data points that are far away from the majority of the entire data set, but ignored the true contextual outliers (Fig. 3(d)). In fact, as shown in Fig. 4, our method consistently outperformed the baseline method in terms of precision, recall and f-score, in both contexts. More importantly, *our method had high precision when only reporting a small number of outliers, which is favorable in practice*. Recall that our contextual outlier score covers the global outlier score as a special case and thus makes it possible to measure the interestingness of global and contextual outliers in a unified framework. In Fig. 5, we show both the global and contextual outlier scores for all data instances in the *iris* data set, which are sorted in ascending order. We can see that even if we report global outliers together with contextual outliers, our algorithm would report 15 contextual outliers before reporting the most probable global outlier.

We also chose the *wine* data set from the UCI Archive, which has three classes. We converted the data set into a random walk graph, also using Euclidean distance. The normalized Min-Cut then partitioned the data set into two parts, where one subgraph contained 59 instances from Class 1, 34 from Class 2; the other subgraph contained 48 instances from Class 3, 37 from Class 2. Thus instances from Class 2 became natural contextual outliers in both subgraphs and were labeled as ground truth outliers. We applied our method as well as the baseline method to rank contextual outliers in both contexts. The result is shown in Fig. 6, which is consistent to the result on the *Iris* data set.

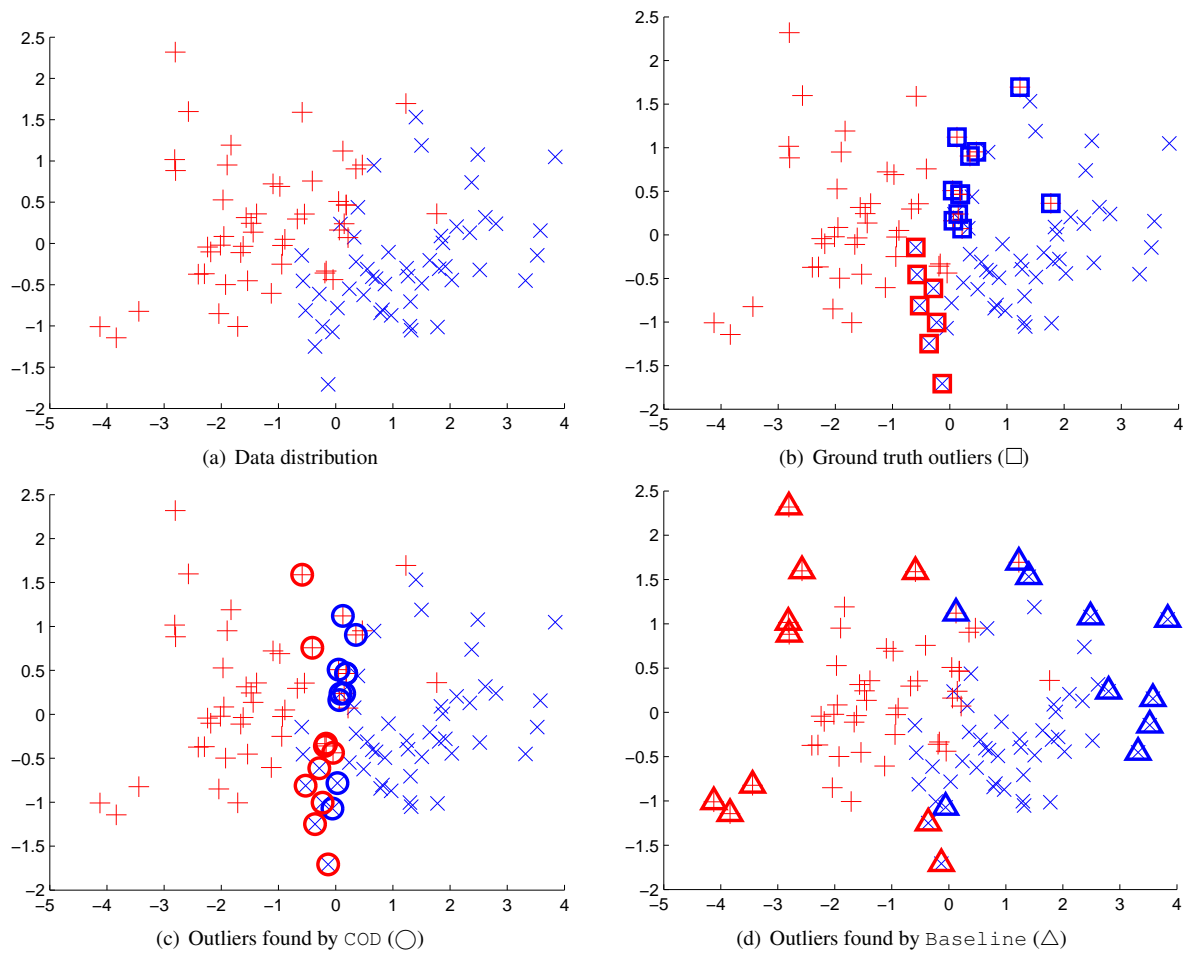


Figure 3: Trimmed Iris data set with two classes (versicolor and virginica).

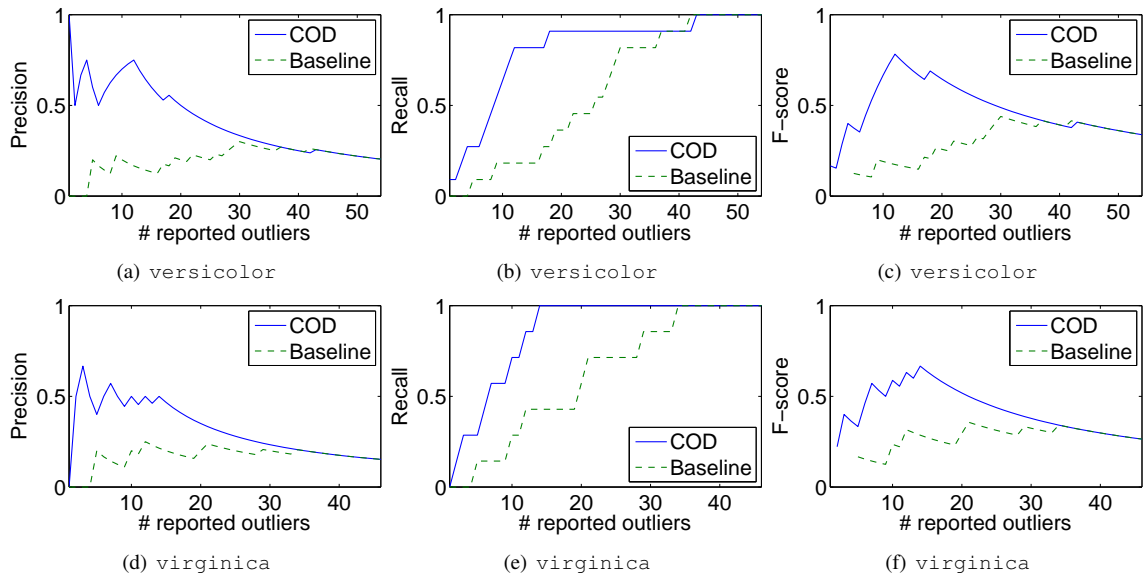


Figure 4: Precision, recall, and f-score on trimmed `iris` data, against # outliers reported.

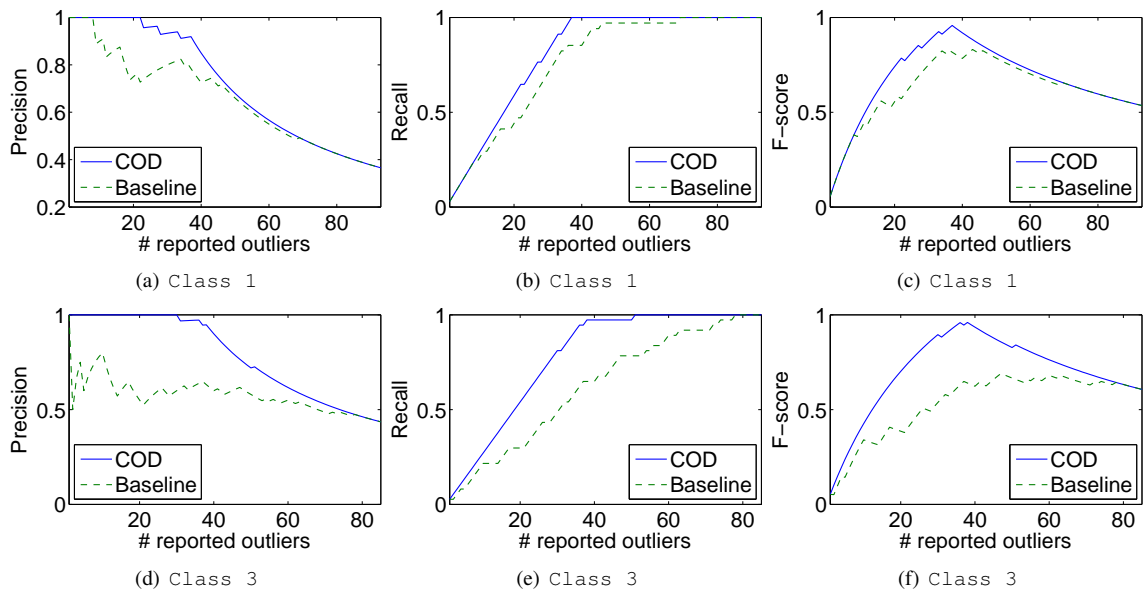


Figure 6: Precision, recall, and f-score on `wine` data, against # outliers reported.

To further justify the effectiveness and advantage of our method, we used the 20 Newsgroups data set, which consists of articles from 20 different topics, 1000 in each. Each time we extracted two classes from the data set and constructed the random walk graph using inner product distance. We tried random combinations of classes and computed the average precision, recall and f-score. Specifically, we chose four science related topics, namely `sci.crypt`, `sci.electronics`, `sci.med`, and `sci.space`. We enumerated all the 6 possible combinations between them. We report the average precision, recall, and f-score at the point where the number of reported outliers is equal to the number of true outliers, as shown in Table 2. Outlier ratio is the ratio of true contextual outliers in the selected data set. Again, the results showed that our method (left columns) performed twice as good as the baseline method (right columns) in terms of precision, recall, and f-score.

## 6 Related Work

Since the quality of identified contextual outliers heavily relies on the meaningfulness of the specified context [10], many existing approaches require *a priori* contextual information, which can be sequential [13], spatial [14, 9], profile [5], statistical [16], and so forth. However, these approaches won't work well in the case where we don't have sufficient *a priori* information of the potential contexts.

Our work uses spectral analysis on a random walk graph to simultaneously explore meaningful contexts and contextual outliers therein. Spectral analysis has been proved to be a powerful tool to study the structure of a graph [3]. Its main application in data mining is spectral clustering [12, 18], where the eigenvectors can be interpreted as a series of bi-partition of the graph. However, most spectral clustering techniques use the eigenvectors of graph Laplacian to partition the graph, whereas we use transition matrix and have different interpretations of the eigenvalues and eigenvectors.

Moonesinghe and Tan [11] introduced random walk model for *global* outlier detection and used the principal eigenvector of the transition matrix as *global* outlier score. The equivalence between the principal eigenvector and the stationary distribution has also been used to rank

web pages and achieved tremendous success [8]. Skillicorn [15] used spectral graph analysis to explore anomalous structure in a graph. Both of their methods were focused on outlier detection in a global point of view, and did not address contextual outliers.

## 7 Conclusion and Future Work

Contextual outlier detection typically requires a context to be specified *a priori*. In this work, we explore automatic and unsupervised identification of contexts and the contextual outliers therein. Our approach is applicable to graph data as well as vector data if this data can be converted to a graph where the edge weights correspond to the similarity between points.

We identify contexts as a 2-coloring of a random walk graph. We introduce the notion of stationary expectation, which is a generalization of the stationary distribution, as our contextual outlier score. For a given node its contextual outlier score characterizes difference in the chance of a random walk performed in the *entire* graph (not just the subgraph) visiting the node given the walk starts from either context. Our contextual outlier score is time-invariant and is solely determined by the structure of the random walk graph. Note that when we identify contexts, we do not modify the graph structure in any way such as by removing edges. Therefore, our approach is **not the same** as performing a multi-way cut on the graph and then applying global outlier detection separately in each individual subgraph.

Our algorithm produces a ranked list of tuples like {instance, context, score}. Note that an instance may appear multiple times in this list, but w.r.t. different contexts and different contextual outlier scores. Our method also covers the global outliers as a special case. We validate the effectiveness of our method by empirical results on real-world data. Our method consistently outperforms the baseline method on different data sets.

Our next extension will be to look at time-evolving graphs. This can essentially model data sets where the points are not fixed but have a trajectory, or that graph edges (e.g. social networks) evolve over time. To analyze this more complex data we shall also explore multi-way partitions of the graph that can discover multiple contexts in a graph simultaneously.

Table 2: Results on 20 Newsgroups data (COD vs. Baseline)

Selected classes	Outlier ratio	Precision		Recall		F-score	
crypt & electr.	.0440	.5463	.2720	.2445	.1325	.3091	.1694
crypt & med	.0155	.5202	.2759	.2814	.1624	.3412	.1939
crypt & space	.0110	.3412	.2239	.1600	.1168	.1991	.1466
electr. & med	.2745	.5085	.4395	.2547	.2112	.3131	.2624
electr. & space	.0910	.5319	.2601	.2659	.1143	.3271	.1430
med & space	.0215	.4035	.2440	.1945	.1015	.2414	.1282

## References

- [1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *SIGMOD Conference*, pages 93–104, 2000.
- [2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.
- [3] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.
- [4] C. H. Q. Ding and X. He. *K*-means clustering via principal component analysis. In *ICML*, 2004.
- [5] T. Fawcett and F. J. Provost. Activity monitoring: Noticing interesting changes in behavior. In *KDD*, pages 53–62, 1999.
- [6] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.
- [7] R. Horn and C. Johnson. *Matrix analysis*. Cambridge Univ. Press, 1990.
- [8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
- [9] Y. Kou, C.-T. Lu, and D. Chen. Spatial weighted outlier detection. In *SDM*, 2006.
- [10] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *ICDM*, pages 413–422, 2008.
- [11] H. D. K. Moonesinghe and P.-N. Tan. Outlier detection using random walks. In *ICTAI*, pages 532–539, 2006.
- [12] M. Saerens, F. Fouss, L. Yen, and P. Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *ECML*, pages 371–383, 2004.
- [13] S. Salvador, P. Chan, and J. Brodie. Learning states and rules for time series anomaly detection. In *Proc. 17th Intl. FLAIRS Conf*, pages 300–305, 2004.
- [14] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In *KDD*, pages 371–376, 2001.
- [15] D. B. Skillicorn. Detecting anomalies in graphs. In *ISI*, pages 209–216, 2007.
- [16] X. Song, M. Wu, C. M. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Trans. Knowl. Data Eng.*, 19(5):631–645, 2007.
- [17] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.
- [18] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.