Physics-Based Anomaly Detection Defined on Manifold Space

HAO HUANG, Computer Science Department, Stony Brook University
SHINJAE YOO, Computational Science Center, Brookhaven National Laboratory
HONG QIN, Computer Science Department, Stony Brook University
DANTONG YU, Computational Science Center, Brookhaven National Laboratory

Current popular anomaly detection algorithms are capable of detecting global anomalies but often fail to distinguish local anomalies from normal instances. Inspired by contemporary physics theory (i.e., heat diffusion and quantum mechanics), we propose two unsupervised anomaly detection algorithms. Building on the embedding manifold derived from heat diffusion, we devise Local Anomaly Descriptor (LAD), which faithfully reveals the intrinsic neighborhood density. It uses a scale-dependent umbrella operator to bridge global and local properties, which makes LAD more informative within an adaptive scope of neighborhood. To offer more stability of local density measurement on scaling parameter tuning, we formulate Fermi Density Descriptor (FDD), which measures the probability of a fermion particle being at a specific location. By choosing the stable energy distribution function, FDD steadily distinguishes anomalies from normal instances with any scaling parameter setting. To further enhance the efficacy of our proposed algorithms, we explore the utility of anisotropic Gaussian kernel (AGK), which offers better manifold-aware affinity information. We also quantify and examine the effect of different Laplacian normalizations for anomaly detection. Comprehensive experiments on both synthetic and benchmark datasets verify that our proposed algorithms outperform the existing anomaly detection algorithms.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.1 [**Pattern Recognition**]: Models—*Unsupervised anomaly detection*

General Terms: Design, Algorithms, Performance

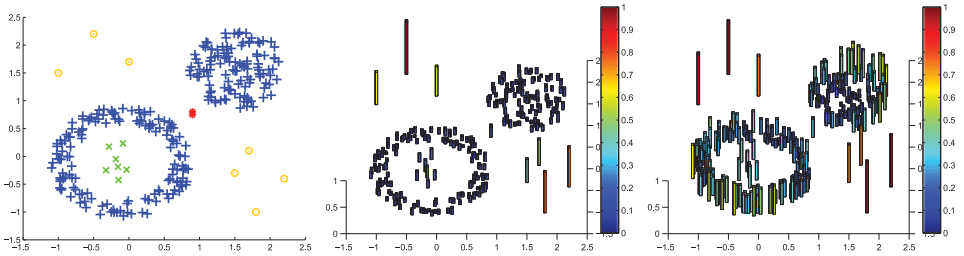
Additional Key Words and Phrases: Anomaly detection, Laplace operator, heat diffusion, quantum mechanics

ACM Reference Format:
Hao Huang, Hong Qin, Shinjae Yoo, and Dantong Yu. 2014. Physics-based anomaly detection defined on manifold space. ACM Trans. Knowl. Discov. Data 9, 2, Article 14 (September 2014), 39 pages.
DOI: http://dx.doi.org/10.1145/2641574

1. INTRODUCTION

Anomaly detection, or outlier detection, is of great significance to many real-world applications [Zhu et al. 2009; Pogorelc and Gams 2010], such as cancer diagnostics and virus detection. Its primary goal is to distinguish normal instances from a small

This paper is an extension of the work published in CIKM 2012 [Huang et al. 2012a] and ICDM 2012 [Huang et al. 2012b]. This research is supported in part by the National Science Foundation of the United States (No. IIS-0949467, IIS-1047715, and IIS-1049448) and the National Natural Science Foundation of China (No. 61190120, 61190121, and 61190125). It is also supported by the U.S. Department of Energy, Grant No. DE-SC0003361, funded through the American Recovery and Reinvestment Act of 2009.
Authors' addresses: H. Huang, Computer Science Department, Stony Brook University, Stony Brook, NY, U.S.A., 11794-4400; email: hao.huang.1@stonybrook.edu; H. Qin, Computer Science Department, Stony Brook University, Stony Brook, NY 11794-4400; S. Yoo and D. Yu, Computational Science Center, Brookhaven National Laboratory, Building 463, Upton, NY 11973.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2014 ACM 1556-4681/2014/09-ART14 $15.00
DOI: http://dx.doi.org/10.1145/2641574



(a) Synthetic Dataset Example (b) LOF Score ($k = 10$) (c) IForest Score ($nt = 100$, $\rho = 4000$)

Fig. 1. (a) Synthetic dataset with normal instances (blue), global anomalies (yellow), and local anomalies (red and green). (b) LOF score with $k = 10$. (c) IForest score. The anomalous score are visualized as a height bar over all instances. For each algorithm output, the anomalous score are normalized in the range of [0, 1] to have an easy comparison. We can see that both LOF and IForest fail to totally distinguish local anomalies from normal instances.

portion of new or abnormal instances (anomalies) [Chandola et al. 2009; Liu et al. 2008; Liu et al. 2011]. In many applications, anomalies are sparse and quite diverse, and learning with the known anomalies [Gao et al. 2006; Wu and Ye 2009; Blanchard et al. 2010] may not be necessarily useful in detecting the unknown ones in previously unseen data [Syed and Rubinfeld 2010]. On the other hand, manually labeling known datasets can be extremely time consuming for real-life applications and sometimes even unpractical to detect new types of rare events. Therefore, the key challenge of anomaly detection still lies in its ability to quantitatively characterize the intrinsic and informative density distribution around every instance in a unsupervised fashion.

In this article, we propose two different unsupervised anomaly detection algorithms: Local Anomaly Descriptor (LAD) and Fermi Density Descriptor (FDD). They measure instance anomalous score based on different physics theory—that is, heat diffusion and quantum mechanic theory, respectively. Compared with the existing algorithms [Breunig et al. 2000; Papadimitriou et al. 2003; Liu et al. 2008; Ting et al. 2010; Agovic et al. 2007], our methods are capable of measuring local density more effectively for the following reasons:

—Our methods have solid physics theory background.
—Our methods are based on manifold space, where the distance between anomalies and normal instances would be magnified. It makes anomalies more salient than in the input space.
—Our methods provide a more adaptive scope of neighborhood, which is of great importance to distinguish not only global but also local anomalies from normal instances.
—Our methods are highly desirable to combat scaling parameter tuning sensitivity.

These properties make our algorithms more informative and intrinsic to detect anomaly.

1.1. Related Work

According to the most classical definition by Hawkins [1980], an anomaly is "an observation which deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism." However, it is far from trivial to define the quantitative sense of "deviates so much from the other observations." As Figure 1(a) illustrates, global anomalies (in yellow) are those data points with low density in the entire data space. We can also say that these points are with globally low neighborhood density. On the other hand, local anomalies (in red and green) are data points with low local density in a constrained region. We call that these points are with locally

low neighborhood density. Profoundly speaking, local anomalies can be thought of as a generalization of global anomalies, as global anomalies will typically also be local anomalies, but not vice versa [De Vires et al. 2010].

In implementation, k-th nearest neighbor (kNN)-based algorithms such as LOF [Breunig et al. 2000], LDOF [Zhang et al. 2009], and LOCI [Papadimitriou et al. 2003] are defined on Euclidean distance. LOF [Breunig et al. 2000], one of the earliest works using kNN distance for anomaly detection, defines anomaly if its distance to its kNN is greatly larger than the distances of its neighbors to their own kNNs. Recent research [De Vires et al. 2010] extended LOF to a high-dimensional dataset by using random projection to reduce dimensions. Two major drawbacks of these approaches are as follows. First, they tend to miss local anomalies (Figure 1(b)) since it is not peculiar that kNN distances of local anomalies are similar to those of their normal instance neighbors. Second, it is of extreme importance to determine the value of k to faithfully reveal the instance anomalous score. On the one hand, k cannot be too small to avoid statistical error. Specifically, we need to ensure that for each instance, especially those forming a microcluster of anomalies, it covers a large enough neighborhood that includes more normal instances than anomalies. On the other hand, too large of a k will lead to overlooking some genuine anomalies. In Section 9.2, we will show that LOF is unpractical to detect anomalies in benchmark datasets by analyzing its sensitivity of k.

Instead of detecting anomalies based on average neighborhood distance, recent approaches such as IForest [Liu et al. 2008, 2011] and Mass [Ting et al. 2010] are to separate the anomalies from normal instances with their noteworthy attribute distribution. A representative anomaly definition [Liu et al. 2008] in these papers states that anomalies should have "attribute-values that are very different from those of normal instances" and at the same time should be "minority consisting of fewer instances." Therefore, these approaches have the capacity to handle anomalies with different attribute distribution compared with normal instances. Nonetheless, they may fail to detect local anomalies when their attributes have not-so-different distribution with some normal instances. From Figure 1(c), we can see that even though IForest does a good job on global anomaly detection, it fails to distinguish local anomalies (green and red instances in Figure 1(a)) from the "boundary" instances in the cluster of normal instances (blue instances in Figure 1(a)). This is because these anomaly detectors partition instances mainly based on observable attributes or, more precisely, the attribute distribution in input data space. Therefore, it will fail miserably when the anomaly distribution becomes far less discriminative if they share similar attribute range/distribution pattern with parts of the normal instances. In Figure 2, we can see that some anomalies have overlapping distribution with normal instances on the first four eigenvectors in ionosphere dataset (a popularly used dataset for anomaly detection [Liu et al. 2008; Hempstalk et al. 2008; Noto et al. 2010]). Such overlapping also appears at nonclassical multidimensional scaling (MDS) as well. This case, to a certain degree, shows that the aforementioned problem indeed exists in some real-world applications.

Since the neighborhood density is not as straightforward as pairwise distance or attribute distribution in the input space, many researches turned to manifold space. In an ideal manifold projection with enlarged distance between anomalous and normal instances, anomaly detection is no longer as hard as that in the input space. A few techniques [Agovic et al. 2007] tried to find an approximation of the data using a combination of attributes that capture the bulk of the variability in the data and then detect anomalies on the projected space. This kind of approach is to approximate the manifold subspaces in which the anomalous instances can be easily identified [Chandola et al. 2009]. However, the existing algorithms are based on suboptimal techniques such as isometric feature mapping (ISM) and locally linear embeddings (LLEs) [Agovic et al. 2007], which are highly sensitive to density-varying and complex data

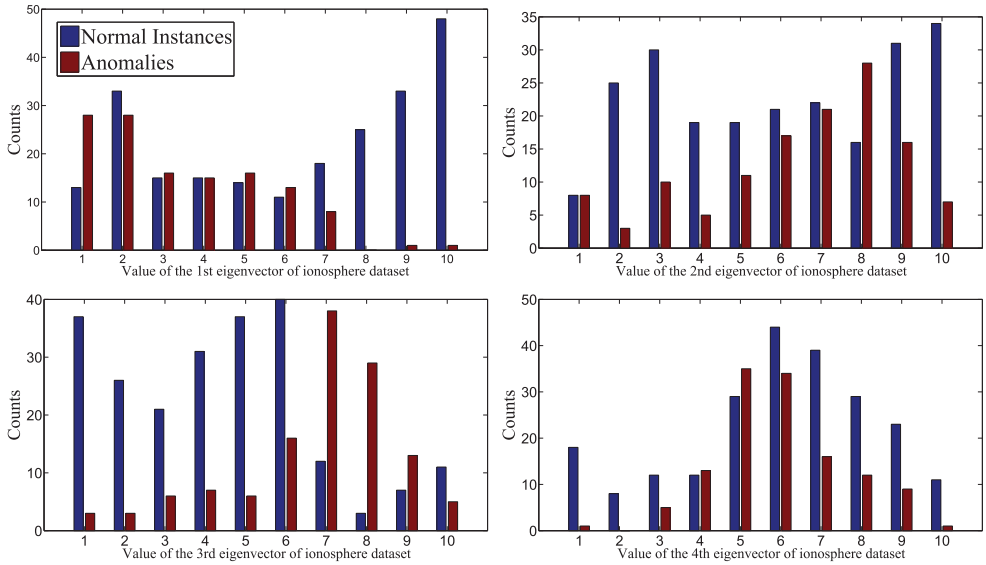


Fig. 2. Histogram of anomalies (red) and normal instances (blue) on the first four eigenvectors (*) of the ionosphere dataset (a popular benchmark dataset for anomaly detection [Liu et al. 2008; Hempstalk et al. 2008; Noto et al. 2010]). Some anomalies have overlapped distribution with parts of normal instances, and therefore it is nontrivial to separate them simply by difference between attribute distributions. The asterisk (*) designates that since the dataset is high dimensional, dimension reduction is imperative to provide a concise illustration. Although eigenvectors do not necessarily show full distribution of the input data, they tend to show certain patterns of original dimensions in the input space.

distribution [Lafon et al. 2006; Van der Maaten et al. 2009]. Therefore, anomaly detection algorithms based on such manifold reconstruction mechanism may fail miserably.

1.2. Motivation

Motivated by the aforementioned problems, we refine the definition of anomaly as follows.

Definition 1.1. Anomalies are those instances with (1) locally low neighborhood density and (2) small quantity of similar instances compared with normal instances.

To capture anomalies under such definition, we consider the Laplace operator in physics theory, which has solid foundation and intrinsic relationship with manifold reconstruction. The reason we resort to manifold space is that normal instances usually lie on low-dimensional embedding structures with high density. But the anomalies projected in manifold space tend to deviate from the normal instances, which makes them more discriminative. On the other aspect, measurement of anomalous score is highly related to similarity function in that the anomalous score of an instance is high if it has few similar neighbors. The Laplace operator is a differential operator given by the gradient divergence of a function on Euclidean space. Therefore, the Laplace operator, if it is performed on a similarity matrix, is capable of representing the flux density of the gradient flow of the neighborhood similarity. Consequently, it offers a natural mechanism to express intrinsic neighborhood density information. Furthermore, the Laplace operator occurs in differential equations that describe many physical phenomena, such as the diffusion equation for heat and quantum mechanics. These properties deliver inspiration and a solid theoretical foundation to our research in this work.

1.3. Contributions

This article articulates two physics-based unsupervised anomaly detection algorithms with the following contributions:

(1) We are the first to quantitatively characterize local density information based on heat diffusion theory (Section 3) and develop LAD (Section 5.1). This method has a locally adaptive scope of manifold-aware neighborhood and therefore can very well satisfy the first property of our proposed anomaly definition in Section 1.2.
(2) In favor of taking the amount of similar instances into account (the second property of the definition in Section 1.2), we integrate a scale-dependent umbrella operator (Section 5.1) into LAD that can bridge the gap between local and global information.
(3) We are the first to explore the use of quantum mechanics theory (Section 6) in anomaly detection and propose FDD (Section 7), which supplies rigorous probabilistic explanation for detecting anomalies and supreme stability to scaling parameter tuning.
(4) We first analyze different Laplacian normalization effects (Section 7.2) with the goal of anomaly detection. Both theoretical proof and quantitative experiments demonstrate that unnormalized Laplacian L_{nn} is the most responsive to local neighborhood density.
(5) We explore the use of anisotropic Gaussian kernel (AGK, Section 4), which more faithfully approximates the similarity between instances in the ideal manifold space and therefore can best help in manifold reconstruction with the goal of anomaly detection.
(6) We systematically evaluate the proposed algorithms with several closely related baseline algorithms on a number of benchmark datasets (Section 9). Our algorithms show not only better average performance but also more stable results than the other popular algorithms. Moreover, experiments confirm that FDD affords robustness for scaling parameter selection.

2. BACKGROUND OF GRAPH LAPLACIAN AND SPECTRAL ANALYSIS

The Laplace operator, when applied on spectral analysis methodology, is called *graph Laplacian*. In this section, we want to introduce the unnormalized and normalized graph Laplacians on finite weighted graphs.

In this work, we denote X, which is an $n \times m$ matrix, as a dataset with n instances, and each instance has m features. Its global similarity matrix W, an $n \times n$ matrix, represents the pairwise likeness of instances considering the whole feature space. Gaussian kernel (GAU) is one of the most generally used options for constructing W:

$$W^{(GAU)}(i, j) = exp\left(\frac{- \parallel x(i) - x(j) \parallel^2}{2\sigma^2}\right), \qquad i, j = 1, \ldots, n, \tag{1}$$

where σ controls the width of the neighborhood [Luxburg 2007]. The degree matrix D is defined by $D(i, j) = \sum_{p=1}^{n} W(i, p)$ if $i = j$, and 0 otherwise. Then, the unnormalized Laplacian matrix L_{nn} can be defined as

$$L_{nn} = D - W, \tag{2}$$

which is the difference between the degree matrix D and the similarity matrix W of the graph. The nice properties of L_{nn} have been discussed in Luxburg [2007]. One of the most important ones is that L_{nn} has as many eigenvalues 0 as there are connected components, and the corresponding eigenvectors are the indicator vectors of the connected components.

There are two common ways of normalizing L_{nn} to correct its bias of different density [Luxburg 2007; Coifman and Lafon 2006]: one is the symmetric normalized Laplacian

matrix L_{sym}, and the other is random walk normalized Laplacian matrix L_{rw}:

$$L_{sym} = D^{-1/2} L_{nn} D^{-1/2}. \tag{3}$$

$$L_{rw} = D^{-1} L_{nn}. \tag{4}$$

The matrix L_{sym} has the advantage of being symmetric; therefore, it has a more balanced view in the instance neighborhood, whereas L_{rw} is a stochastic matrix that can be viewed as the transition matrix of a Markov chain on each instance.

To better depict the global distribution, Coifman and Lafon [2006] analyzed these two normalization and proposed a new normalization family. It is shown by these authors that if we assume uniform sampling of data points from a submanifold $\mathcal{M}$, the eigenvectors of L_{rw} with $\sigma \to 0$ and $n \to \infty$, tend to approximate the Laplace-Beltrami operator on $\mathcal{M}$, which guarantees manifold structure reconstruction. However, in reality, the sampled data points tend to be nonuniform and show skewed density distributions, resulting in poor manifold structure reconstruction. To improve the global distributional sensitivity of traditional normalization, the following two additional normalizations are considered in Coifman and Lafon [2006]:

$$L_{fp} = I - D^{-1} W^{'}, \tag{5}$$

where $W^{'} = D^{-1/2} W D^{-1/2}$, and

$$L_{lbn} = I - D^{-1} W^{''}, \tag{6}$$

where $W^{''} = D^{-1} W D^{-1}$. L_{fp} is called *Fokker-Planck normalization*, and L_{lbn} is called *Laplace-Beltrami normalization*. Especially, Laplace-Beltrami normalization can remove the influence of the dataset density and recovers manifold structures on $\mathcal{M}$ with the condition of both $\sigma \to 0$ and $n \to \infty$ [Coifman and Lafon 2006]. In other words, the additional renormalization of affinity matrix W enables the reconstruction of manifold structures better under nonuniform density distribution for the purpose of clustering.

From any of the aforementioned L_{**}, we can obtain the corresponding eigenvectors. In spectral analysis theory, the first c ($c \ll m$) nontrivial eigenvectors with the smallest eigenvalues (except 0) are the most important signal components, which in theory form the manifold structure of X [Luxburg 2007]. Denote these c eigenvectors as Y, which is an $n \times c$ matrix. Each row of Y is the corresponding coordinates of each original instance in the manifold space, whereas each column of Y (eigenvector) represents an axis (dimension) in the manifold space. These eigenvectors are orthogonal to each other and together provide the compressed and embedding representation of a dataset's distribution.

As far as we know, there is no other research focus on the effect of different Laplacians on anomaly detection. In our article, we will analyze this problem with the two manifold-based techniques that we proposed.

3. HEAT KERNEL SIGNATURE BASED ON HEAT DIFFUSION

3.1. Introduction of Heat Diffusion

Our first proposed algorithm is strongly inspired by heat diffusion theory [Hsu 2002] in that it can provide information intimately related to local density. Heat theory can be interpreted as the transition density function of Brownian motion [Sun et al. 2009], which is the most fundamental continuous time Markov process. The Laplace operator is closely associated to heat diffusion, connecting geometry of a manifold with the properties of the heat flow. Using the discrete Laplace operator, the heat equation can be simplified and generalized to matrix operation over spaces with an arbitrary number of dimensions. Due to its intrinsic connection to the Markov process, in practice the heat equation is often coupled with random walk graph Laplacian [Coifman and Lafon

2006], L_{rw} (Equation (4)), which describes a stochastic process that randomly jumps from vertex to adjacent vertex. Heat equation therefore can be defined by

$$\frac{\partial H_t}{\partial t} = -L_{rw} H_t, \tag{7}$$

where $H_t = e^{-tL_{rw}}$ is the heat kernel on Riemannian manifold $\mathcal{M}$ and t is the time scaling parameter [Grigoryan 1999]. For $L_{rw} = \psi' \lambda \psi$ (ψ and λ are the eigenvectors and eigenvalues of L_{rw}), the heat kernel can be reformulated as follows:

$$H_t(i, j) = \sum_{p=1}^{N} [e^{-\lambda_p t} \psi_p(i) \psi_p(j)], \tag{8}$$

where λ_p is the p-th eigenvalue and $\psi_p(i)$ is the i-th element in the p-th eigenvector. $H_t(i, j)$ represents the amount of heat being transferred from i to j in time t given a unit heat source at i in the very beginning. The scaling parameter t in heat kernel is used to control the transitive connectivity: small t makes the loosely connected graph into slightly stronger connection, whereas large t makes the graph tend to be more strongly connected.

3.2. Heat Kernel Signature

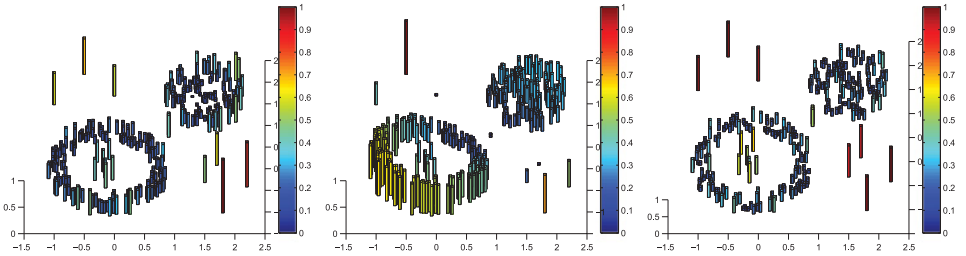
In 2009, Sun et al. [2009] proposed a concise form given by the heat kernel from one instance to itself:

$$\mathcal{H}_t(i) = H_t(i, i) = \sum_{p=1}^{N} [e^{-\lambda_p t} (\psi_p(i))^2], \tag{9}$$

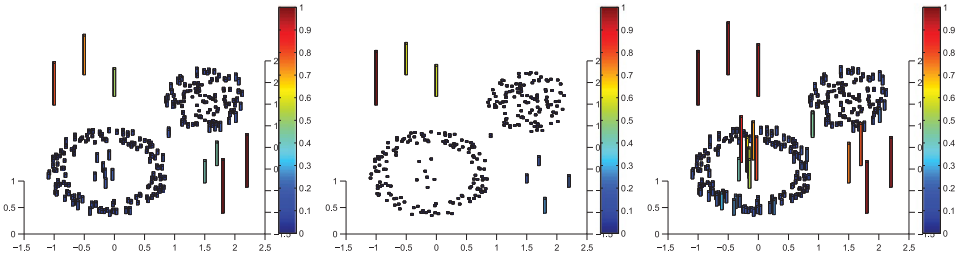
which is referred to as heat kernel signature (HKS). The physical meaning of HKS is the amount of heat each instance keeps within itself in time t. The property of the heat diffusion process states that heat tends to diffuse slower at instances with a more sparse neighborhood and faster at instances with a denser neighborhood. Therefore, HKS can intuitively depict the local density of each instance (the first property in our anomaly definition in Section 1.2). Besides, HKS also has the following properties that make it a very lucrative candidate for local density measurement:

—HKS is intrinsic to the local manifold structure.
—HKS is informative since it contains density information of the whole neighborhood in t scale.
—The stableness of HKS against small perturbation in the neighborhood can be well supported by the probabilistic interpretation of heat diffusion.

However, heat equation is assumed to build on the underlying manifold. But in most applications, the underlying manifold is unknown. In geometric modeling application, HKS is usually built on eigenvectors from GAU (Equation (1)) on observed space. Although graph Laplacian normalizations [Coifman and Lafon 2006] based on GAU on observed space can recover manifold structure to certain extent, nonuniformly sampled instances tend to show unpreserved density distribution on the reconstructed manifold. HKS on GAU will fail to reveal local density faithfully in such reconstructions. Figure 3(a) and 3(b) show the performance of HKS on anomaly detection with $t = 1$ and $t = 10$ based on GAU and random walk graph Laplacian normalization L_{rw}. When $t = 10$ (Figure 3(b)), the heat is extremely easy to dissipate, which blends both local and a few global anomalies into normal instances. Meanwhile, many marginal instances of the two normal instance clusters stand out due to the fact that HKS on GAU fails to show manifold-aware properties. When $t = 1$ (Figure 3(a)), although the short period of

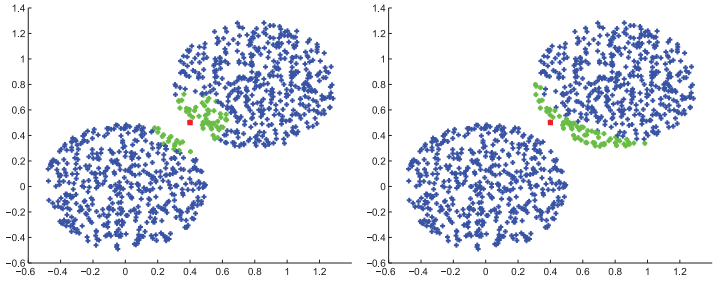


(a) HKS with GAU($t = 1$) (b) HKS with GAU($t = 10$) (c) LAD with GAU($t = 1$ and $k = 10$)



(d) HKS with AGK($t = 1$) (e) HKS with AGK($t = 10$) (f) LAD with AGK($t = 1$ and $k = 10$)

Fig. 3. HKS and LAD (see Equation (16)) score with GAU (Equation (1)) and AGK (see Equation (13)) of the synthetic dataset in Figure 1(a). For each algorithm output, the anomalous score are normalized in the range of [0, 1] to have an easy comparison. We can see that LAD with AGK is the most aware of both global and local anomalies.



(a) (b)

Fig. 4. The 70 nearest neighbors (in green) of red instance with GAU (a) and AGK (b), which shows that AGK has better manifold-aware property than GAU.

heat dissipation has salient effect on global anomalies, HKS on GAU still fails to distinguish local anomalies from normal instances on the boundary area of normal clusters. Therefore, an alternative way is indispensable to build a better manifold-aware affinity matrix. One of the most preferable candidates is AGK [Singer and Coifman 2008, 2011].

4. ANISOTROPIC GAUSSIAN KERNEL

In this section, we use AGK [Singer and Coifman 2008] to construct HKS in the interest of better manifold reconstruction. In Figure 4, we can see the 70 nearest neighbors of red instance when using GAU (Figure 4(a)) and AGK (Figure 4(b)), which shows that the intramanifold distances are much shorter than the intermanifold by using AGK.

Figure 3(d) and (e) show that anomaly detection can directly benefit from the use of AGK. In Figure 3(e) with $t = 10$, all of the global anomalies are highlighted even though the local anomalies are latent (compared with Figure 3(b)). This is because if the manifold is well reconstructed, global anomalies should be separated far away from normal instances even with large t scale. Furthermore, with small scope of $t = 1$ (Figure 3(d)), the difference of anomalous score between local anomalies and boundary normal instances are slightly more obvious than Figure 3(a), which illustrates that with the support from AGK, HKS is more capable of revealing the density information of the intrinsic manifold structure.

In the rest of this section, we briefly introduce AGK on the observed space X ($n \times m$ matrix) that approximates the GAU on the underlying manifold Y ($n \times d$ matrix, with $d \ll m$). The idea is to approximate the Euclidean distance between instances $y(j)$ in the manifold space Y using covariance matrix $C = JJ^T$, where J is the Jacobian matrix [Singer and Coifman 2008] and the instances $x(j) = f(y(j))$ in the observable space X. Let y, ϵ be two instances in the manifold space Y and $x = f(y)$, $\eta = f(\epsilon)$ be their mapping to the observable space X. Let $g : X \to Y$ be the inverse mapping of $f : Y \to X$—that is, $g(f(y)) = y$ and $f(g(x)) = x$, $\forall y \in Y$, $\forall x \in X$. Expanding the functions $y = g(x)$ in a Taylor series at the instance x gives

$$\begin{aligned}\epsilon(i) &= y(i) + \sum_j g^i_j(x)(\eta(j) - x(j)) \\ &\quad + \frac{1}{2}\sum_{kl} g^i_{kl}(x)(\eta(k) - x(k))(\eta(l) - x(l)) + O(\| \eta - x \|^3),\end{aligned} \tag{10}$$

where $g^i_j = \frac{\partial g(i)}{\partial y(j)}$. Therefore, the squared Euclidean distance in manifold space can be approximated by

$$\begin{aligned}\| \epsilon - y \|^2 = &\sum_{ijk} g^i_j(x) g^i_k(x)(\eta(j) - x(j))(\eta(k) - x(k)) \\ &+ \frac{1}{2}\sum_{ijkl} g^i_j(x) g^i_{kl}(x)(\eta(j) - x(j))(\eta(k) - x(k))(\eta(l) - x(l)) \\ &+ O(\| \eta - x \|^4).\end{aligned} \tag{11}$$

A similar expansion can be built at instance η and the average of these two equations can be produced as

$$\begin{aligned}\| \epsilon - y \|^2 = &\frac{1}{2}(\eta - x)^T \left[(JJ^T)^{-1}(x) + (JJ^T)^{-1}(\eta)\right](\eta - x) \\ &+ O(\| \eta - x \|^4),\end{aligned} \tag{12}$$

given that the Jacobian of the inverse g is the inverse of the Jacobian J (a detailed description of calculation can be referred to Singer and Coifman [2008]). So we can construct the AGK:

$$W^{(AGK)}(i, j) = exp\left(-\frac{\| J^{-1}x(i)(x(i) - x(j)) \|^2 + \| J^{-1}x(j)(x(j) - x(i)) \|^2}{2\sigma^2}\right), \tag{13}$$

where $i, j = 1, \ldots, n$.

AGK has the desired attributes that it is separable, and its first (nontrivial) eigenfunctions are monotonic functions of the independent parameters [Singer and Coifman 2011]. It also has been proved that the eigenvectors of AGK reveal the independent components [Singer and Coifman 2008]. HKS, built on such approximation of manifold

space, can better capture the embedding structure of data as shown in Figure 3(d) and 3(e), which is difficult or even impossible to achieve by using GAU or other similar techniques.

5. LOCAL ANOMALY DESCRIPTOR AND THE ALGORITHM FRAMEWORK

5.1. Local Anomaly Descriptor

Although HKS on AGK has the capability to offer desirable local density information, it is of importance to select the right time scaling parameter t, which provides a trade-off between the effects of local and global information. However, it is hard to get the "best of both worlds" with a single setting for this parameter. Even with better manifold reconstruction, if t is large the heat is still easy to dissipate regardless of normal instances or local anomalies (although not necessarily for global anomalies), which is shown in Figure 3(e). This is because with large t scale, the distance between local anomalies and the normal instances around them would still be close. As a result, local anomalies cannot retain their heat. On the other hand, if t is small, the heat diffusion runs for only a short period of time, and the resulting anomalous score captures very local information but almost carries the same value for instances with similar density inside a very restrained neighborhood, which is the major reason it sometimes confuses some normal instances with local anomalies. In Figure 3(d), we can see that HKS on AGK assigns similar scores to the local anomalies and some of the boundary normal instances. Intuitively speaking, HKS on AGK still fails to take the amount of similar instances into account with off-the-sweet-spot t setting.

As a means to handle the preceding problems, we propose the use of an umbrella operator [Taubin 1995; Desbrun et al. 1999]. An umbrella operator is an approximation of the Laplace operator measuring the vector from the vertex in question to the barycenter of its neighbors. In practice, umbrella operator U is usually implemented to compute the average difference between a point $x(i)$ and its kNNs $nb(x(i), k)$:

$$U(i) = \frac{1}{k} \sum_{x(j) \in nb(x(i),k)} (x(j) - x(i)). \tag{14}$$

In our research, we need to deliberate on the quantity of similar instances in neighborhood by bridging the gap between global and local properties. If an instance has a lot of close neighbors, the average value of the neighborhood should be very similar to the value of this instance. Therefore, we use the scale-dependent (weighted) umbrella operator $\mathcal{U}$:

$$\mathcal{U}(i) = \frac{1}{k} \sum_{x(j) \in nb(x(i),k)} W(i, j)(x(j) - x(i)), \tag{15}$$

where $W(i, j)$ is the weight between $x(i)$ and $x(j)$. If we replace $W(i, j)$ with $W^{(AGK)}(i, j)$, then we may use the scale-dependent umbrella operator on top of HKS ($\mathcal{H}$). LAD is defined for a point $x(i)$ as follows:

$$\mathcal{L}_t(i) = \mathcal{H}_t(i) - \frac{1}{k} \sum_{x(j) \in nb(x(i),k)} \mathcal{H}_t(j) \times W^{(AGK)}(i, j). \tag{16}$$

The geometric meaning of LAD is illustrated in Figure 5, where we measure the difference between a single $\mathcal{H}_t(i)$ and its neighborhood's weighted average $\mathcal{H}_t(j) \times W^{(AGK)}(i, j)$ value.

If an instance is globally anomalous, its HKS would be already high enough to discriminate itself to the other instances. Although it is locally anomalous, its HKS is

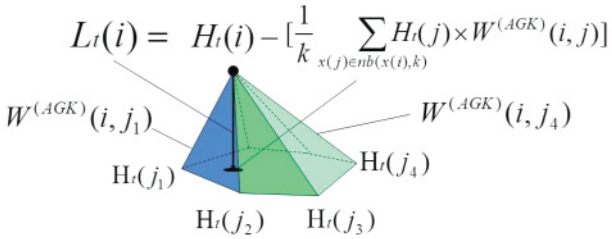


Fig. 5. Illustration of LAD (Equation (16)), which calculates the weighted average of neighbor differences. It is one of the ways to take the neighborhood distribution into consideration [Taubin 1995].

likely to be similar to some normal instances with a similarly sparse neighborhood. However, applying the scale-dependent umbrella operator, LAD can serve to recognize the local anomalies from normal instances with expanded horizon of neighborhood and reflection of the amount of similar instances inside. Local anomaly only has a small amount of neighbors with close HKS, but normal instances, on the other hand, have more such neighbors.

LAD has a very lucrative property in considering the amount of similar instances (the second property in our anomaly definition): since it not only measures a very constrained local area with small t, but also considers the heat diffusion area of the adjacent neighbors. It gives a measurement of an expanded horizon to capture how many similar instances there are inside a large enough neighborhood. If there are lots of similar neighbors (with similar local density), LAD will be quite small since the neighborhood difference of HKS is not large. On the contrary, if the neighbors' HKS are different, the LAD score tends to be assigned a greater value. So even though k is not large enough to include the whole appropriate neighborhood, LAD can still capture the information related to the amount of similar instances.

The benefits of LAD in comparison with HKS can be seen in Figure 3, which shows that our proposed LAD has a penetrating awareness on both global and local anomalies primarily because of the power of the scale-dependent umbrella operator on HKS.

Mathematically, LAD could also use GAU as the connection weighted function W (or $W^{(GAU)}$), which not only affects the term of subtrahend in Equation (16) but also the construction of HKS ($\mathcal{H}_t$). To have a concrete understanding of the effect of AGK, we also compare the different performance between GAU and AGK on LAD. LAD with GAU in Figure 3(c), although making some anomalies more salient, still fails to distinguish some local anomalies and normal instances. But LAD with AGK in Figure 3(f) clearly separates all global and local anomalies from the normal instances. This confirms that as a connection weighting function, AGK is more effective than GAU in that AGK is more aware of the differences between instances in the manifold space.

5.2. Algorithmic Framework of Local Anomaly Descriptor

In this subsection, we explain LAD framework step by step. Let X be a matrix of size $n \times m$, where n is the number of instances and m is the number of dimensions; our framework is detailed in Algorithm 1. This algorithm undergoes a kind of data warping process by using AGK (Step 1, Section 4) and Laplacian random walk normalization (Step 2, Section 2). Then we perform the eigen-decomposition (Step 3) and construct HKS for each instance (Step 4, Section 3). Equation (16) is used as the last step (Step 6, Section 5.1) to compute LAD as the final measurement of anomalous score.

Regarding computational complexity, affinity construction using GAU takes $O(n^2m)$. If using AGK, it takes $O(n^2m^2)$. Eigendecomposition (Step 3) is another time-consuming step. There are many iterative methods to conduct eigenvalue decomposition, but in

ALGORITHM 1: LocalAnomalyDescriptor(X, σ, t, k)

Input: Input data $X \in R^{n \times m}$, σ the Gaussian scaling parameter, t the time scaling parameter, k the neighborhood size.
Output: LAD score for each instance.
1 Construct anisotropic Gaussian kernel $W^{(AGK)}$ using Equation (13) and σ;
2 Construct Laplacian random walk normalization L_{rw} on $W^{(AGK)}$;
3 Compute generalized eigenvectors ψ_p and corresponding eigenvalues λ_p, $p = 1, 2, \ldots, n$;
4 Construct Heat Kernel Signature with time scale t using Equation (9);
5 Compute Local Anomaly Descriptor using Equation (16) with Heat Kernel Signature and anisotropic Gaussian kernel in the k nearest neighborhood for each instance.

general, finding the eigenvalues reduces to matrix multiplication by computing a symbolic determinant, which gives a running time of $O(n^3 + n^2 log^2 n)$ [Pan and Chen 1999]. An alternative way of estimating the heat kernel $H_t = e^{-tL_{rw}}$ is to use a partial sum of infinite series with

$$e^{-tL_{rw}} = \sum_{p}^{\infty} \frac{(-tL_{rw})^p}{p!}. \tag{17}$$

This method would be especially attractive for small values of t, since only a few terms would be needed to obtain an accurate estimation of $e^{-tL_{rw}}$ [Badeau et al. 2005], which is desired for our LAD calculation, since a small amount of t is good enough to reveal the anomalous score.

On the other hand, if we only use a small portion of eigensystem (say the first d eigenvalues and eigenvectors) to compute LAD, the eigendecomposition only takes $O(n^2 d)$. We will analyze the performance of this fast version in the experimental Section 9.6.

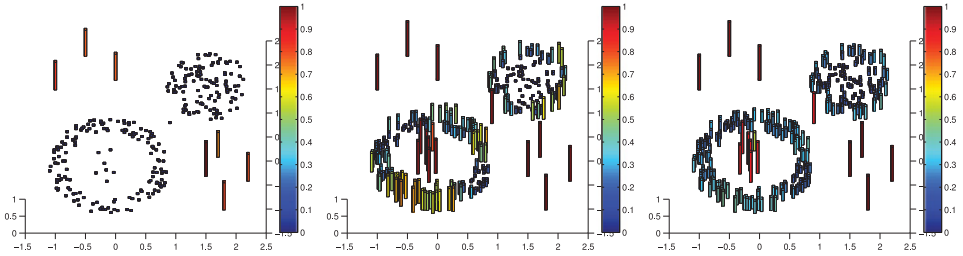
5.3. Discussion of Local Anomaly Descriptor

As we introduced in previous sections, LAD can capture the two properties of our anomaly definition (Section 1.2) effectively:

—The local diffusion process calculated by HKS with small t can intuitively depict the local density of each instance (Section 3.2).
—The umbrella operator provides a broader view even with too small t so that it has a lucrative property in considering the amount of similar instances (Section 5.1).

Although LAD gets over the instability of HKS to some degree by integrating a scale-dependent umbrella operator, which provides a more broader view of neighborhood distribution, LAD still suffers if the time scaling parameter t goes too large (see the example in Figure 6(a)). More completely, Figure 7 shows that the stability of LAD is below expectation when t is large. This is because as the diffusion time gets longer, HKS across all instances will all become the same. Subsequently, there is no difference between HKS of anomalies and their neighbors. This problem of LAD (and of course HKS as well) comes from the essential properties of heat diffusion naturally: once the dissipation time is large, heat will easily become overdiffused.

In the next two sections, we resort to quantum mechanics, whose research objects are in a discrete space, which has the potential to detect locally low neighborhood density more stably (Figure 6(b) and 6(c)). In quantum mechanics, particles jump from one quantum state to another, and the waves space is not continuous. The probability of a particle showing up at a certain place is highly related to the local density of this place. To a certain degree, quantum mechanics intuitively focuses on the intrinsic local density distribution while largely ignoring the extrinsic properties (pairwise distance, attribute distribution, etc.) of the ambient area of input space.



(a) LAD ($t = 100$ and $k = 10$) (b) FDD ($T = 1$) (c) FDD ($T = 0.01$)

Fig. 6. LAD with large t fails to reveal the local anomalous score (Figure 6(a)) due to the overdiffusion. Comparably, FDD acts robustly in measuring anomalous score regardless of small or large scaling parameter (Figure 6(b) and 6(c)). For each algorithm output, the anomalous score are normalized in the range of [0, 1] to have an easy comparison.

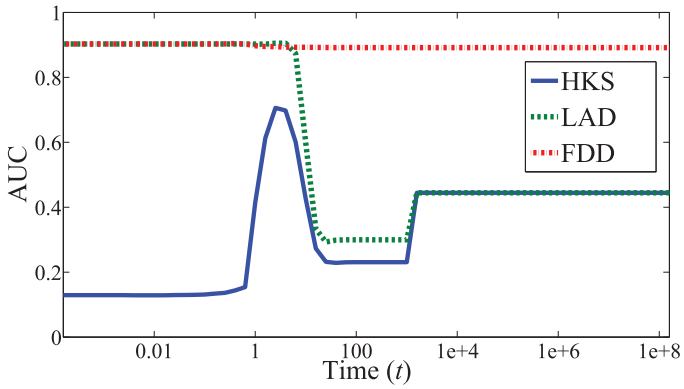


Fig. 7. Illustration of stability test on the ecoli dataset against time scaling parameter (t) tuning. We can see that although LAD (green curve) has better performance and stability than HKS (blue curve) when t is small, it still does not make accurate detection when t becomes larger ($t \geq 100$). Our ideal goal is to design an anomaly detection algorithm (red curve) that maintains a desirable result regardless of scaling parameter tuning.

6. SCHRÖDINGER EQUATION AND WAVE FUNCTION IN QUANTUM MECHANICS

Besides heat diffusion, another physics concept that is closely related to density measurement is quantum mechanics [Greenstein and Zajonc 2006], which also has strong connections to the Laplace operator. Quantum mechanics is a mathematical machine for predicting the behavior of microscopic particles. Anomalous instances can be treated as regions of low density that correspond to the aggregation area of maximal free energy, and such an area is easier to trap particles. On the other hand, normal instances indicate high-density regions with minima of the free energy in the system, so the probability for particles appearing in such an area is low.

The Schrödinger equation is the key equation in quantum mechanics, which describes how the quantum state of a physical system changes with time. One of the most famous examples is the nonrelativistic Schrödinger equation for a single particle moving in an electric field. If we ignore the potential energy, it is directly associated with the Laplace operator L as follows:

$$\iota \frac{\partial \phi}{\partial t}(x, t) = L\phi(x, t), \tag{18}$$

where ϕ is the space-time wave function of the quantum system, ι is the imaginary unit, x is the position, and t is time. The mod square $|\phi(x,t)|^2$ depicts the probability density of a particle at position x at time t, which satisfies

$$\int |\phi(x,t)|^2 dx = 1. \tag{19}$$

Assume that the Laplace spectrum has no repeated eigenvalues, and $L = \psi' \lambda \psi$ (ψ and λ are the eigenvectors and eigenvalues of L); the space-time wave function $\phi(x,t)$ can be expressed in the spectral domain as

$$\phi(x,t) = \sum_{p=1}^{\infty} e^{\iota \lambda_p t} \psi_p(x) f(\lambda_p), \tag{20}$$

where $f(\lambda)$ is the energy distribution. This is because in spectral domain, eigenvalue λ is approximately equivalent to energy level E [Greenstein and Zajonc 2006], so $f(\lambda)$ can also be rewritten as $f(E)$.

Integrating the mod square of wave function $|\phi(x,t)|^2$ over all time scales, we can get

$$\mathcal{P}(x) = lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} |\phi(x,t)|^2 dt = \sum_{p=1}^{\infty} f(\lambda_p)^2 \psi_p(x)^2. \tag{21}$$

The physical meaning of $\mathcal{P}(x)$ is the average possibility for a particle with energy distribution $f(\lambda)$ found at position x. The property of quantum mechanics states that due to the fast decaying nature of the evanescent wave, a particle tends to be trapped within the vicinity of region where the strong field enhancement occurs. In a high-dimensional dataset, the "tip" regions are those data points with a sparse neighborhood. In other words, the particle tends to stay at instances with a more sparse neighborhood and rarely shows up at instances with a denser neighborhood. Therefore, in theory, $\mathcal{P}(x)$ can intuitively represent the local density of each instance. In practice, however, there are two key challenges:

(1) What is the best energy distribution $f(\lambda)$?
(2) What is the best graph Laplacian L (which directly associates with λ and ψ)?

Section 7.1 will solve the first challenge, and the second challenge will be discussed and conquered in Section 7.2.

7. FERMI DENSITY DESCRIPTOR AND THE ALGORITHM FRAMEWORK

7.1. Energy Distribution Function and Definition of Fermi Density Descriptor

In this subsection, we will explore the best energy distribution function f for Equation (21). $f(E)$ ($f(\lambda)$ determines the probability that a particle is in energy state E. It can be viewed as a realization of the ideas of discrete probability in such a case that energy can be treated as a discrete variable. In quantum mechanics, there are three distinctly different distribution functions [Greenstein and Zajonc 2006], namely Maxwell-Boltzmann (MB) distribution, Fermi-Dirac (FD) distribution, and Bose-Einstein (BE) distribution. Besides quantum mechanics, existing research also explored distributions based on other theoretical assumptions. Section 3 already introduced heat diffusion (HD), which was used in Sun et al. [2009] to describe the heat diffusion given time t. In Aubry et al. [2011] chose Gaussian distribution (GD) in the logarithmic energy as $f(E)$ to define Wave Kernel Signature. Here we briefly introduce these five distribution functions and analyze their respective performance on anomaly detection.

—Maxwell-Boltzmann Distribution

$$f_{MB}(E) = \frac{1}{e^{E/\kappa T}}. \tag{22}$$

MB distribution depends on the energy level E of the single particle state, the absolute temperature T, and the Boltzmann constant κ. In quantum mechanics, the MB distribution usually applies to the particles at a high enough temperature and low enough density where quantum effects can be ignored [Greenstein and Zajonc 2006].

—Fermi-Dirac Distribution

$$f_{FD}(E) = \frac{1}{e^{(E-\mu)/\kappa T} + 1}, \tag{23}$$

where μ can be obtained from

$$\sum_{E} \frac{1}{e^{(E-\mu)/\kappa T} + 1} = n/2. \tag{24}$$

Beside the same parameters used in Equation (22), FD distribution is also conditional on a chemical potential μ and n the number of electrons in the whole systems. Equation (24) represents the number of orbitals since only two electrons (with opposite "spin") can occupy each orbital. In quantum mechanics, FD distribution applies to identical particles (fermions) with half-odd-integer spin in a system in thermal equilibrium [Greenstein and Zajonc 2006].

—Bose-Einstein Distribution

$$f_{BE}(E) = \frac{1}{e^{(E-\mu)/\kappa T} - 1}, \tag{25}$$

where μ can be obtained from

$$\sum_{E} \frac{1}{e^{(E-\mu)/\kappa T} - 1} = n/2. \tag{26}$$

The parameters used in the BE distribution function have the same physical meaning as those used in Equations (22) and (23). The BE distribution describes the statistical behavior of integer spin particles (bosons). At low temperatures, bosons can behave very differently than fermions because an unlimited number of them can be collected into the same energy state [Greenstein and Zajonc 2006].

—Heat Diffusion

$$f_{HD}(E) = e^{-Et}, \tag{27}$$

where t is the time for heat dissipation. Heat diffusion describes how the amount of heat dissipates from a heat source to its neighborhood at time t. Different from the three distributions in quantum mechanics that depict the discrete pattern of particle movement in terms of probability, heat diffusion has a continuous conception in both time and space domains.

—Gaussian Distribution

$$f_{GD}(E) = e^{-\frac{(e-log(E))^2}{2\sigma^2}}. \tag{28}$$

It is derived in Aubry et al. [2011] from a perturbation-theoretical analysis. Under the assumption that the eigenvalues (eigenenergies) of an articulated dataset are log-normally distributed random variables, the authors claimed that it is robust to small data perturbations while being as informative as possible.

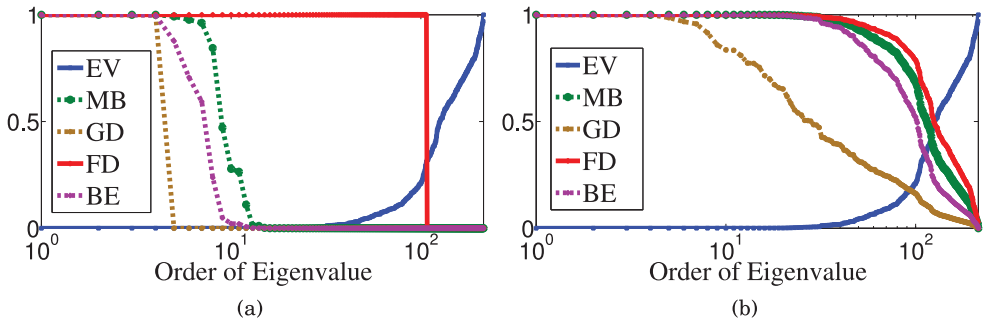


Fig. 8. Stability comparison between different energy distribution functions on a glass dataset. The blue curve is the eigenvalue (EV) ordered by increasing value (decreasing importance since EVs are derived from graph Laplacian). The green, red, purple, and brown curves are MB, FD, and BE distributions and GD, respectively. (a) The performance of four functions when $T = 0.001$. (b) The performance of four functions when $T = 50$. We can see that FD has the most stable performance as T changes.

Before comparison between the aforementioned different energy functions, we need to clarify a few points. First, since κ is a constant, from now on we will remove κ from the relative formulas (Equations (22) through (26)) in the interest of convenience. Second, although HD and MB distribution have different physical meaning, they indeed have similar mathematical performance if we simply replace t in Equation (27) with $\frac{1}{T}$ in Equation (22). In other words, small diffusion time t in heat diffusion has a similar effect as large environmental temperature T in MB distribution. Therefore, in the following analysis, we simply ignore heat diffusion (HD) and only compare the other four distribution functions. Third, although with different physical meanings, for the sake of mathematical convenience, we assign $2\sigma^2$ in GD (Equation (28)) with the same value as T in the quantum mechanics functions to compare the stability of different functions as scaling parameter changes.

Although GD and MB/HD, FD, and BE distributions have solid theoretical background, the differences of mathematical performance give rise to very different statistics, especially the stability of outcomes:

—Among these functions, FD is the most practical one for anomaly detection in terms of performance stability under different parameter (T) settings. In Equation (23), two special terms can stabilize the distribution function performance: the constant smoothing term "plus one" and the balancing term μ in the denominator part. The role of the smoothing term is to damp the contribution of the exponential part from being too small, which results from either extremely small λ (E) or large T. The balancing term μ (computed by Equation (24)) is a parameter controlling the trade-off between small and large λ (E). Besides, it helps to tune a sweet range for λ (E) according to T, since it has a positive side effect that it can accelerate the attenuation of contribution from those trivial eigenvalues in Equation (23).
—Comparably, MB/HD and GD without any smoothing term or balancing term are very sensitive to either extremely small λ (E) or large scaling parameter T. Although BE has balancing effect from μ, it actually suffers more from the "minus one" in the denominator part, since it lessens the stability by making the denominator part even smaller.

Figure 8 shows the value of different distribution functions across different eigenvalues (energy) of a glass dataset (statistic details of glass are in Section 9.1). In general, FD distribution tends to assign stable weights regardless of how temperature T changes compared with the other energy distribution functions. To have a broader

and fair comparison between the effect of different energy distribution functions, we test all of the distribution functions on seven datasets against changing T. The detailed results, which again confirm our findings, are recorded in Section 9.5.

Now we integrate the FD distribution function (Equation (23)) into Equation (21) and define FDD at a point $x(i)$ as

$$\mathcal{F}(i) = \frac{1}{C} \sum_{p=1}^{\infty} \left(\frac{1}{e^{(\lambda_p - \mu)/T} + 1} \right)^2 \psi_p(i)^2, \tag{29}$$

where $C = \sum_{p=1}^{\infty} (\frac{1}{e^{(\lambda_p - \mu)/T}+1})^2$, and μ can be derived from Equation (24) where n is set as the number of data instances in practice.

7.2. Laplace Operator for Fermi Density Descriptor

We discuss the best choice of graph Laplacian for FDD in this section. In Section 6, we showed that our proposed FDD is derived from the Schrödinger equation (Equation 18), which is strongly associated with the Laplace operator. The Laplace operator is intimately related to the "shape" of data, or more precisely, the density distribution of data. More precisely, the Laplace operator in Equation (18) is aiming to account for the kinetic energy of the particles constituting the system, which depends on the spatial configuration to conserve energy [Greenstein and Zajonc 2006]. Using the discrete Laplace operator, or graph Laplacian, the Schrödinger equation can be simplified and generalized to be a matrix operation over the space of an arbitrary number of dimensions.

Different graph Laplacian normalizations were introduced in Section 2. Although their effect on clustering has been thoroughly analyzed in Huang et al. [2011] and Luxburg [2007], it is still unclear what the best choice for FDD is with the purpose of anomaly detection.

In general, when the data points are sampled from the equilibrium distribution of a stochastic dynamical system, clustering algorithms tend to correct different density bias to obtain stable and balanced instance clusters. This is quite different from the need of anomaly detection applications when the density of the points is a quantity of interest and therefore cannot be ignored [Coifman and Lafon 2006]. For clustering purposes, we focus on normal instances and want to recover manifold insensitive to the existing anomalies (usually being treated as noise in such applications). In other words, the different density distribution prevents algorithms from the desire clustering result and therefore needs to be removed in clustering applications [Huang et al. 2011]. However, from anomaly detection's point of view, the focus is on the anomalies, and the recovered manifold should be aware of local density variation; therefore, in the manifold space, the density differences between anomalies and normal instances should be preserved or even magnified with respect to the input space distribution. In a nutshell, we need to find the graph Laplacian that is most reactive to local density distribution with the purpose of anomaly detection.

THEOREM 1. *The density impact power for L_{nn}, L_{rw}, L_{sym}, $L_{fp,}$ and L_{lbn} normalization are* 2, 1, 1, 0.5, *and* 0, *respectively.*

PROOF. Define $q(x)$ as the true density function of x, and a kernel function $k_\sigma(x, y)$ between x and y with σ as the neighborhood scaling parameter. Let

$$q_\sigma(x) = \int k_\sigma(x, y) q(y) dy, \tag{30}$$

which is an approximation of the true density $q(x)$; we can form the new kernel [Coifman and Lafon 2006]:

$$k_\sigma^\alpha(x, y) = \frac{k_\sigma(x, y)}{q_\sigma^\alpha(x) q_\sigma^\alpha(y)}, \tag{31}$$

where $\alpha \in R$. Apply the Laplacian operator to this kernel as follows:

$$d_\sigma^\alpha(x) = \int k_\sigma^\alpha(x, y) q(y) dy; \tag{32}$$

the new anisotropic kernel can be defined as

$$p_\sigma^\alpha(x, y) = \frac{k_\sigma^\alpha(x, y)}{d_\sigma^\alpha(x)}. \tag{33}$$

Therefore, based on the Laplacian operator, the infinitesimal generator of the Markov chain with $\sigma \to 0$ [Coifman and Lafon 2006] can be defined as

$$L_{\sigma,\alpha} = \frac{I - P_{\sigma,\alpha}}{\sigma}, \tag{34}$$

where $P_{\sigma,\alpha} f(x) = \int p_\sigma^\alpha(x, y) f(y) q(y) dy$ with any function f. If $\sigma \to 0$, we have

$$lim_{\sigma \to 0} L_{\sigma,\alpha} f = \frac{\Delta(f q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}} f. \tag{35}$$

Hence, the infinitesimal operator can be given by

$$\Delta\varphi - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}} \varphi, \tag{36}$$

where $\varphi = f q^{1-\alpha}$.

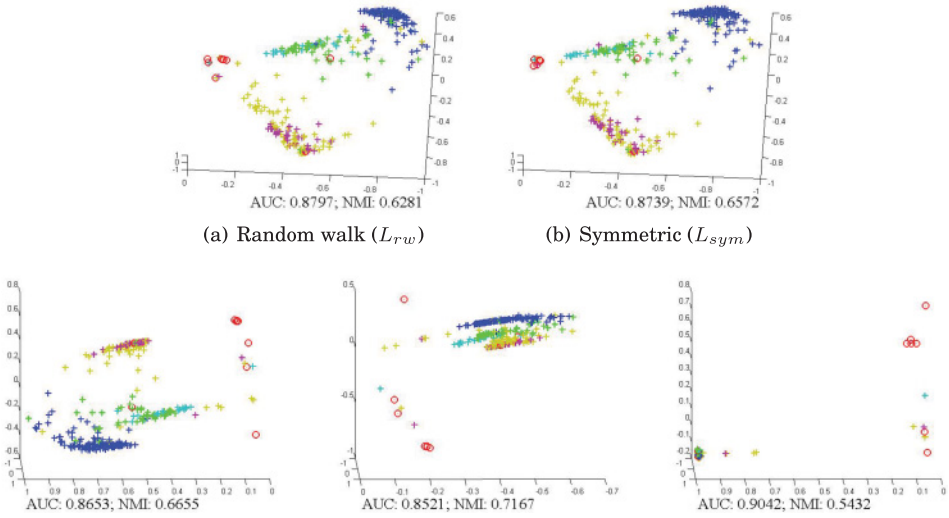
For L_{rw} normalization, $\alpha^{L_{rw}} = 0$ [Coifman and Lafon 2006], so the density impact power is $1 - \alpha^{L_{rw}} = 1$. For L_{fp} normalization, $\alpha^{L_{fp}} = 0.5$ [Coifman and Lafon 2006]; hence, the density impact power $1 - \alpha^{L_{fp}} = 0.5$. For L_{lbn} normalization, $\alpha^{L_{lbn}} = 1$ [Coifman and Lafon 2006]; thus, its density impact power is $1 - \alpha^{L_{lbn}} = 0$.

L_{sym} normalization can be transformed from L_{rw} normalization by $L_{sym} = D^{1/2} L_{rw} D^{-1/2}$. From Equation (32), we know that $D(d)$ is proportional to the density impact power q; therefore, $lim_{\sigma \to 0} L_{sym,\sigma} f$ depends on density function $q^{-1/2} q^{1-\alpha^{L_{rw}}} q^{1/2} = q^1$, where $\alpha^{L_{rw}} = 0$. On this account, its density impact power is also 1.

For L_{nn}, since $L_{nn} = D L_{rw}$, and $lim_{\sigma \to 0} L_{nn,\sigma} f$ depends on density function $q \times q^{1-\alpha^{L_{rw}}} = q^2$, where $\alpha^{L_{rw}} = 0$. Accordingly, L_{nn} has the greatest density impact power 2. □

Proof of Theorem 1 demonstrates that L_{nn} is the best option for FDD. As an illustration, Figure 9 shows the effects of different normalizations on the ecoli dataset (Section 9.1). We only plot the first three nontrivial eigenvectors derived from the graph Laplacian. The red circles indicate anomalous instances, and crosses with other colors represent different clusters of normal instances, respectively. We also show the AUC score (Section 9.1) of anomaly detection result, and the NMI score (the detailed definition of NMI can be referred to Hartigan and Wong [1978]) of clustering result from different graph Laplacians.

This experiment shows that the L_{lbn} normalization (Figure 9(d)) reorganizes points with larger intracluster similarity and smaller intercluster similarity. Therefore, L_{lbn} normalization has the highest NMI (0.7167). Nevertheless, the overdiffusion and the consequent unresponsiveness of density distribution generate a tail of normal instances connected to anomalous instances, which leads to the lowest AUC (0.8521). Compared



(a) Random walk (L_{rw}) (b) Symmetric (L_{sym})

(c) Fokker-Planck (L_{fp}) (d) Laplace-Beltrami (L_{lbn}) (e) No-normalization (L_{nn})

Fig. 9. The comparison from different graph Laplacians' effect on the ecoli dataset for the purpose of anomaly detection (measured by AUC) and clustering (measured by NMI). Red circles indicate anomalous instances, and crosses in other color represent different clusters of normal instances. We can see that L_{nn} is the best choice for anomaly detection since it magnifies the distance and density differences between anomalies and normal instances. On the contrary, L_{lbn} is the worst choice for anomaly detection purposes but the best option for clustering.

with L_{lbn} normalization, L_{fp} (Figure 9(c)) normalization spreads the instances with a slightly more dispersive distribution (e.g., cluster in dark yellow), which makes a lower NMI 0.6655 but a slightly higher AUC 0.8653.

L_{rw} (Figure 9(a)) and L_{sym} (Figure 9(b)) normalizations reconstruct circle-like shape in their manifold space. But they also show more mixture of different clusters since they preserve the same density as in the input space with impact power equal to 1. Consequently, it gives higher AUC (0.8739 for L_{sym} and 0.8797 for L_{rw}) but lower NMI (0.6572 for L_{sym} and 0.6281 for L_{rw}) compared with L_{lbn} and L_{fp}.

L_{nn} has the most polarized manifold reconstruction. The reason is that the density difference is amplified by L_{nn} compared with the four normalizations. It results in a situation where the normal instances with higher density shrink to a condensed area, whereas anomalous instances are far away from the collapsed center of normal instances. Consequently, L_{nn} has the strongest ability (with AUC 0.9042) to separate anomaly from normal instances even though the clustering based on it will miserably fail (with NMI 0.5432). Section 9.4 will show more convincing experiment results confirming that L_{nn} is the best Laplacian for FDD.

7.3. Algorithmic Framework of Fermi Density Descriptor

Let X be a matrix of size $n \times m$, where n is the number of instances and m is the number of dimensions; our algorithm is detailed in Algorithm 2.

Step 1 (details in Section 4) constructs an AGK similarity matrix and L_{nn} is operated on top of it in Step 2 (details in Section 2) to generate density polarized manifold projection. Then we perform the eigendecomposition (Step 3) and compute FDD for each instance (Step 4, details in Section 7). The FDD value is used as the final measurement of anomalous score.

ALGORITHM 2: FermiDensityDescriptorGlobal(X, σ, T)

Input: Input data $X \in R^{n \times m}$, σ the Gaussian scaling parameter, T is the environmental temperature.
Output: FDD score for each instance
1 Construct Anisotropic Gaussian Kernel (AGK) $W^{(AGK)}$ using Equation (13) and σ;
2 Construct L_{nn} (Equation (2));
3 Compute generalized eigenvectors $\psi(i)$ and corresponding eigenvalues λ_i of L, $i = 1, 2, \ldots, n$;
4 Construct Fermi Density Descriptor (FDD) with temperature T using Equation (29).

Similar to LAD in Section 5.2, the computation of FDD is dominated by affinity construction ($O(n^2m^2)$ if AGK and $O(n^2m)$ if GAU) and eigendecomposition in Step 3 ($O(n^3)$ [Pan and Chen 1999]). In addition, if we only use a small portion of the eigensystem (say the first d eigenvalues and eigenvectors) to compute FDD, the computational complexity of eigendecomposition would drop to $O(n^2d)$. We will analyze the performance of this fast version in Section 9.6.

7.4. Discussion of Fermi Density Descriptor

FDD satisfies the two properties of our anomaly definition (Section 1.2) in a more concise and effective way, in that FDD relies on the polarized manifold reconstruction that magnifies the distances between anomalies and normal instances. Consequently, anomalous instances will be more singular and distinctive. The dense neighborhood will become even denser, with analogous instances aggregated together. The sparse neighborhood, on the contrary, will be more sparse. In this fashion, FDD considers the locally low neighborhood density and the amount of similar instances simultaneously and effectively.

Besides, FDD has more robust performance against different physics parameter settings (especially the extreme cases). Part of the reasons lie in the stable energy function FD, which was already scrutinized in Section 7.1. The other reason is because the polarized manifold reconstruction "breaks" the connections between anomalies and normal instances. Figures 6 and 7 illustrate the stability comparison of FDD, LAD, and HKS, which once again confirm that FDD maintains the desired result more stably with different parameter tuning.

8. DISCUSSION OF THEORETIC PERSPECTIVES

We now justify the utility of our proposed two algorithms, LAD and FDD, by briefly documenting their theoretical connections with a few existing methods, which also lays a solid foundation for their attractive properties for practical use.

8.1. Comparison between Local Anomaly Descriptor and Fermi Density Descriptor

LAD and FDD are based on the Laplace operator on the affinity matrix, as well as the subsequent manifold reconstruction. They all try to describe the density information in a retained but informative neighborhood. However, their different theoretical background leads to quite different interpretation and performance:

(1) *Theoretical backgrounds*: LAD is inspired by heat diffusion, which is highly related to the Markov chain. It describes the amount of heat being transferred in a certain time scale; therefore, its conception is continuous in both time and space. On the contrary, FDD measures the probability that a particle (fermion) shows up at a certain position. It is built upon quantum mechanics, whose key idea is that the motion of a particle is discontinuous and random.

(2) *Manifold reconstruction*: Due to the close theoretical connections, LAD uses random walk normalization L_{rw} by natural and projects origin instances onto a diffusion space. However, the diffusion process is hard to control and usually becomes overdiffused, leading to a blending of local anomalies into normal instances. But FDD applies L_{nn} to construct a polarized manifold projection, which concentrates on magnifying the difference between anomalies and normal instances. Roughly speaking, in the polarized manifold, the similar points with higher density tend to collapse to the center of mass; therefore, clusters of normal instance are topologically isomorphic to extremely condensed convex sets. Conversely, anomalies will be more singular and distinctive from the normal instances. Although this type of mapping is nonisometric and the original distribution is changed, it is of central interest in anomaly detection, as it becomes more sensitive to locally low neighborhood density and the preservation of intracluster distance or distribution is not a concern at all. In Section 9.4, we will further confirm our choice of Laplacians for LAD and FDD with more experiment results.

(3) *Strategies against parameter sensitivity*: To overcome the narrow scope of small t, LAD integrates a scale-dependent umbrella operator on the projected diffusion space, which bridges the gap between global and local properties. Its advantage compared with HKS is that, although with the same small t, LAD covers a sufficiently large neighborhood for each instance $x(i)$ since LAD also considers the t scale neighborhood of $x(i)$'s neighbors. On the other hand, it takes the quantity of similar instances into consideration. But FDD approaches stability against parameter tuning in a different way: Equation (29) has two special terms that stabilize FDD performance: the constant smoothing term "plus one" and the balancing term μ in the denominator part. Both of these terms can damp the contribution of the denominator from being too small, which results from the extreme setting of scaling parameter.

(4) *Stability*: Although LAD provides more robust performance under very small t compared with HKS, it still suffers when t becomes too large due to overdiffusion of heat dissipation. But FDD has stronger stability than LAD in that it can conquer the negative side effect of extreme scaling parameter, regardless of whether it is too small or too large.

8.2. Connections between Local Anomaly Descriptor/Fermi Density Descriptor and Other Anomaly Detection Algorithms

kNN-based approaches. kNN-based methods [Breunig et al. 2000; De Vires et al. 2010; Zhang et al. 2009] approach local density for each instance using its neighborhood information. Like LAD and FDD, they require (scaling) parameters to capture a reasonably large neighborhood, and the density information is based on this prescribed local region. However, kNN-based methods have strictly local context in that they simply fix the neighborhood size with k. In contrast, LAD employs locally adaptive neighborhood size that directly benefits from the physics-inspired properties of heat diffusion, whereas FDD makes use of stabilization terms to smooth out the performance fluctuation from off-the-sweet-spot parameters. Moreover, kNN-based methods rely on Euclidean distance in the input space, which is a pairwise local quantity, whereas our methods consider the relationship between instances in manifold space, which is more comprehensive. For example, heat kernel used in LAD considers all possible paths between two instances within time t. Therefore, our proposed methods are more intrinsic and informative than kNN-based methods.

Attribute-based approaches. Attribute-based methods [Liu et al. 2008, 2011; Ting et al. 2010] try to compute local density by adding up a sequence of values from an

attribute-based function [Ting et al. 2010], which to some extent is equivalent to a kernel density function such as heat kernel. Their measurement of global instance distribution is based on each attribute and how deviated each instance is from the other instances in that specific attribute, which indeed is more informative than kNN-based approaches. However, the strong emphasis on input attribute distribution is also a double-edged sword: on the one hand, it is much faster without any distance calculation; on the other hand, such distribution simply hinged on attributes still fails to consider local anomalies. Although our methods undergo a step of dimension reduction or manifold projection at first, they map all correlated attributes onto a few lower dimensions. Therefore, both LAD and FDD are more capable of stably finding local anomalies.

8.3. Connections between Local Anomaly Descriptor and Other Related Techniques

Biharmonic operator. HKS is directly derived from the Laplace operator and its eigen-decomposition; therefore, HKS is intrinsically a second-order property relevant to the Laplace equation. The derivation of LAD, or the scale-dependent umbrella operator, can be intuitively related to the biharmonic process, because the Laplace operator is essentially applied twice (to compute both HKS and the subsequent scale-dependent umbrella operator). It provides a good balance in the sense that it decays slowly in small cluster around the source instance and fast enough to be structurally inherent in dense areas. This specific "balancing" is intimately derived from the biharmonic equation with properties such as local support and global informativeness [Lipman et al. 2010].

Signal processing. LAD also has strong connection to signal processing. In lowpass filtering, the divergence of a sample from its average neighborhood is the easiest way to pinpoint those inconsistent instances if the desired signal has significant high frequency content. As in traditional signal processing [Taubin 1995], it is possible for LAD to quantify the frequency response by computing an adjoining sum of the Laplace operator in its immediate vicinity. As a result, this enables LAD to distinguish between normal instances and inconsistent instances (anomalies) with greater precision.

Diffusion-based clustering. Some recent research [Richards et al. 2009; Qiu and Hancock 2007; Huang et al. 2011] proposed the probabilistic clustering approaches based on diffusion space. By integrating all time scales of kernel function into one single term, this kind of techniques completely removes the diffusion time scaling parameter and therefore has the built-in robustness to data perturbation and scaling parameter tuning [Huang et al. 2011]. However, as a side effect, this process of "integration" easily assimilates local anomalous instances into normal instance clusters since the excessive diffusion tends to connect everything together. LAD, in sharp contrast, is built upon kernel function with a small time scale and scale-dependent umbrella operator instead of integrating all time scales together. Therefore, it avoids the excessive connection problem.

8.4. Connections between Fermi Density Descriptor and Quantum-Based Clustering/Classification

Much data mining research [Horn and Gottlieb 2001, 2002; Nasios and Bors 2007; Weinstein 2010] has used the Schrödinger equation from quantum mechanics to allow the clusters, or overdense regions in the data, to reveal themselves.

As an example, the intuition behind quantum clustering [Horn and Gottlieb 2001] is based upon the fact that in the quantum system, local maxima in the ground state wave function correspond to the local minima of potential [Weinstein 2010]. And such minima are likely to be good candidates for the cluster centroid locations [Weinstein

Table I. Statistics of our Evaluation Datasets

	Dataset	Number of Instance	Number of Attribute	Percentage of Anomalies Classes)	References
1	breastcancer	683	9	35.0% (malignant)	Noto et al. [2010]
2	wdbc	569	29	37.3% (malignant)	Noto et al. [2010]
3	pima	768	8	34.9% (positives)	Liu et al. [2009]
4	arrhythmia	452	279	45.0% (abnormal)	de Vires et al. [2010]
5	arcene	200	10,000	44.0% (positives)	Guyon et al. [2007]
6	prostatetumor	102	10,509	50.9% (abnormal)	Statnikov et al. [2005]
7	gse24417	417	6,864	31.2% (abnormal)	Popovici et al. [2010]
8	hayesroth	132	5	22.7% (class 3)	Noto et al. [2010]
9	ecoli	336	7	2.7% (omL, imL, and imS)	Noto et al. [2010]
10	yeast	1,484	8	3.7% (vac, pox, and erl)	Noto et al. [2010]
11	abalone	4,177	7	8.0% (*age* <5 or >15)	Noto et al. [2010]
12	glass	214	9	4.2% (tableware)	Noto et al. [2010]
13	ionosphere	351	34	35.9% (bad)	Liu et al. [2009]
14	pageblocks	5,473	10	4.2% (graphic, vertline, and picture)	Noto et al. [2010]
15	magic04	19,020	10	35.2% (hadron)	Noto et al. [2010]

2010]. Instances lying in the basin of attraction of particular minima were identified as a single cluster. Advanced methods have been proposed [Horn and Gottlieb 2002; Weinstein 2010] that differ in how to handle the problem of identifying data points with local minima of the function in high dimensions. Specifically, the Schrödinger equation in Nasios and Bors [2007] is used to calculate the probability of locating a particle given its potential energy.

Although our proposed FDD also applies the Schrödinger equation, it ignores the potential energy in Equation (18), as we are not trying to cluster instances to certain centroids but rather are focusing on the local density measurement to distinguish anomalies. In our study, the Schrödinger equation acts as a cost function separating instances with different density instead of clustering/classifying instances according to the local potential.

9. EXPERIMENTAL ANALYSIS

9.1. Experimental Setup

9.1.1. Datasets. To demonstrate the performance of our proposed FDD and LAD, we evaluate our algorithms on 15 benchmark datasets, including 7 medical datasets (breastcancer, wdbc, pima, arrhythmia, arcene, prostatetumor, and gse24417), 4 biological datasets (hayeRoth, ecoli, yeast, and abalone), and 4 physics datasets (glass, ionosphere, pageblocks, and magic04), whose statistics are summarized in Table I. All of these datasets have been popularly used in anomaly detection research (related references for each dataset are listed in Table I). Such diverse combination of data is intended for our comprehensive studies. In the data preprocessing step, all nominal (including binary) attributes or attributes with missing values are removed.

Anomalies in some of the datasets (wdbc, arrhythmia, prostatetumor, etc.), although carrying a large number of instances, have scattered and sparse distribution as shown

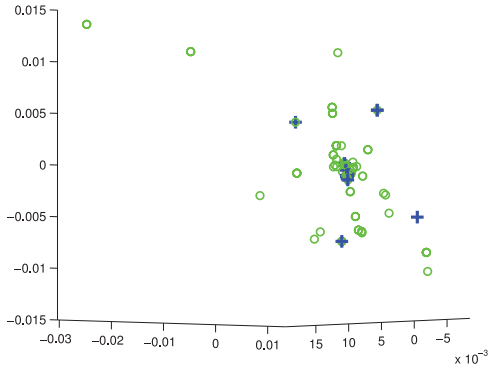


Fig. 10. Dataset "wdbc" shown on the first three nontrivial eigenvectors. Anomalous instances in green (37.3% of instances) are more scattered and sparse than normal instances in blue (62.7% of instances). Therefore, these anomalies, although having a large amount of instances, should be treated as many small abnormal clusters instead of a single cluster.

in Figure 10. Therefore, the anomalies in these datasets should be treated as a combination of many small anomalous clusters instead of one or a few normal clusters with high density [De Vires et al. 2010; Noto et al. 2010], which is consistent with our anomaly definition in Section 1.2.

9.1.2. Baselines. We choose seven state-of-the-art competitors in three categories to show the outstanding performance of our proposed FDD and LAD. For kNN-based algorithms, we choose Local Outlier Detection (LOF) [Breunig et al. 2000] and Local Correlation Integral (LOCI) [Papadimitriou et al. 2003]. LOCI especially provides an automatic, data-dictated cutoff to determine whether an instance is an anomaly based on probabilistic reasoning. For attribute-based methods, we include IForest [Liu et al. 2008] and Mass [Ting et al. 2010]. For manifold-based methods, we choose two different manifold-based techniques used in Agovic et al. [2007], including LLEs and ISM, followed by LOF to obtain anomalous score measurement. We also include the Strangeness-based Outlier Detection (StrOUD) algorithm presented in Barbara et al. [2006]. StrOUD is based on Transductive Confidence Machines, which have been proposed previously as a mechanism to provide individual confidence measures on classification decisions [Barbara et al. 2006].

9.1.3. Evaluation Metrics. Since we have the ground truth of labels for each dataset, we compare our anomaly detection results with labels. For the purpose of theoretical analysis and practical use, we use three evaluation metrics: AUC, F1-score, and macro paired t-tests.

AUC. AUC measures the area under the receiver operating characteristics curve, which can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. AUC is commonly used to evaluate anomaly detectors and it is cutoff independent. Detailed definition of AUC can be referred to Marzban [2004].

F1-score. In practical use, the anomalous score of all of the instances are usually sorted, and those instances with higher value are assigned as anomalies. We assume that the number of anomalies h is already known (calculated with the ground truth), then the first h instances with the highest anomalous score are selected. We evaluate the estimated results with F1-score of the anomaly class. For more details about F1-score, we refer readers to Powers [2011].

Macro paired t-tests. During the experiment, we also show that our FDD provides more stable anomaly detection accuracy for all of the datasets by using macro paired t-tests [Zimmerman 1997] against each competitor, respectively. Note that a score of macro paired t-tests (p-value) should be no more than 0.05 to be considered statistically significant.

9.1.4. Parameters. First of all, we introduce the parameter settings in our FDD and LAD. Our algorithm FDD has two scaling parameters: the Gaussian scales σ and the environmental temperature T. These two parameters are also used in our LAD (heat diffusion time t has been replaced by $\frac{1}{T}$; check Section 7.1 for details). Besides these two parameters, LAD also has another tuning parameter k, the size of neighborhood scope, which is used in the scale-dependent umbrella operator. In our experiments, they are set as follows:

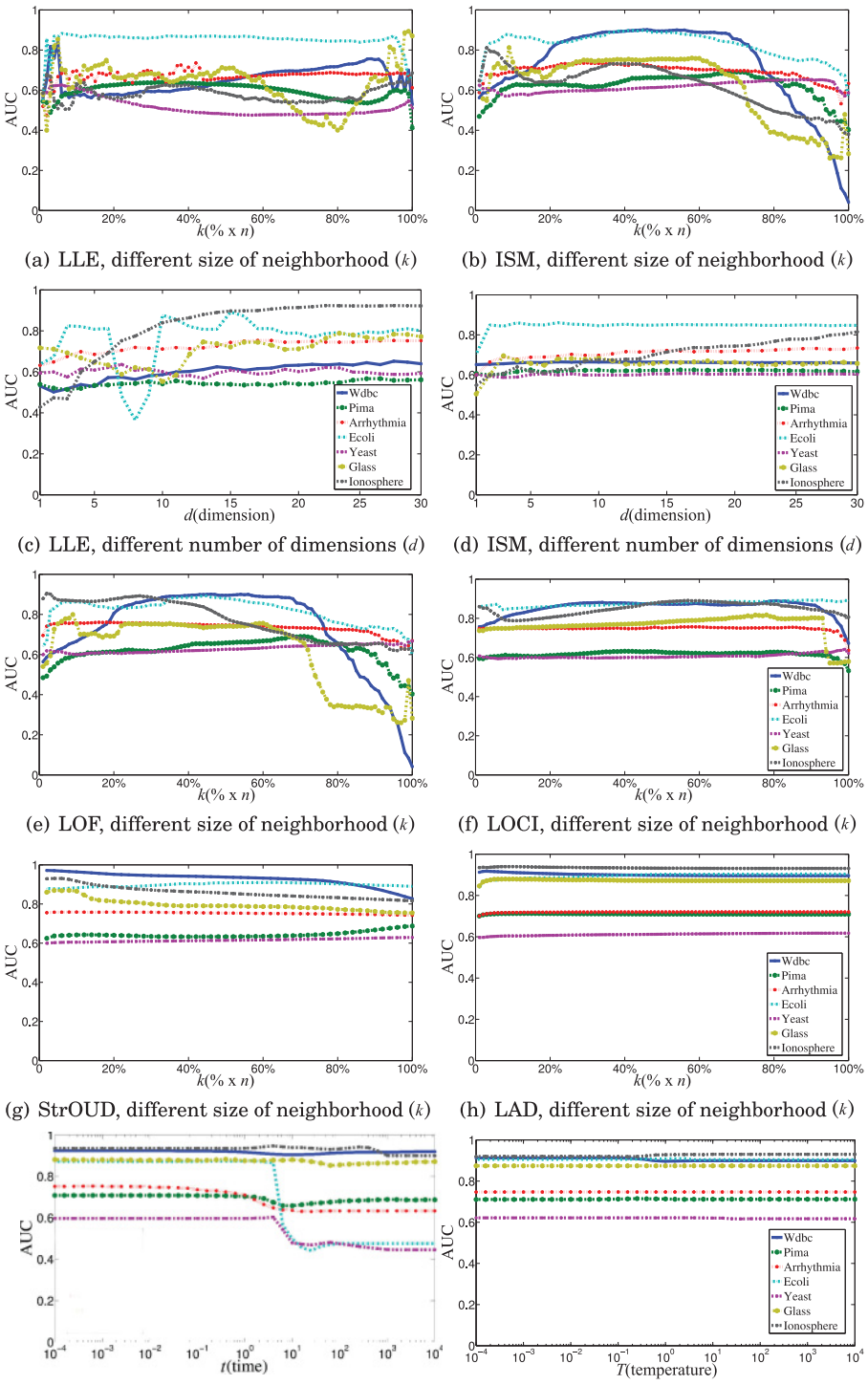
—$\boldsymbol{\sigma}$: For local sensitivity, σ in both FDD and LAD are always fixed to be the average distance of each point to its 2-NN (second nearest neighbor).
—$\boldsymbol{T(\frac{1}{t})}$: Specifically we fix $t = 1$ in LAD in all of the experiments (except Figures 11(h) and 12(h)) to avoid the heat dissipation from overdiffusion. For all FDD experiments and LAD in Figures 11(h) and 12(h), the range of $T(\frac{1}{t})$ is in $10\hat{}\{-4, -3.8, -3.6, \ldots, 3.8, 4\}$.
—$\boldsymbol{k}$: k is fixed to be $k = \lceil 1\% \times n \rceil$ (n is the number of instances) in Figures 11(i) and 12(i). But in the other LAD experiments, the stability of LAD with different k is tested with $k \in \lceil \{1\%, 2\%, 3\% \ldots, 100\%\} \times n \rceil$.

The size of neighborhood scope, k, is a commonly used parameter that also appears in LLE, ISM, LOF, LOCI, and StrOUD. For these algorithms, k is also tested in $k \in \lceil \{1\%, 2\%, 3\% \ldots, 100\%\} \times n \rceil$.

The parameter settings of the other algorithms in our experiments are briefly introduced as follows. For LLE and ISM, we fixed $d = 5$ to measure across different k in Tables II and III, as well as in Figures 11(a) and 11(b) and 12(a) and 12(b). But in Figures 11(c) and 11(d) and 12(c) and 12(d), we show the stability of LLE and ISM across different $d \in [1, 30]$ by choosing the best k from the previous test for each dataset. In LOCI, the radius coefficient is set as $\alpha = 0.5$, which is the same as in the work of Papadimitriou et al. [2003]. As for IForest, to conduct safe and fair comparison, we set ρ and the number of trees nt as the following six combinations: $\rho = 8$ and $nt = 100$ (the number of trees); $\rho = 8$ and $nt = 1{,}000$; $\rho = 256$ and $nt = 10$; $\rho = 256$ and $nt = 100$; $\rho = 256$ and $nt = 500$; $\rho = 256$ and $nt = 1{,}000$. For the same reason, in Mass we set the subsampling size ρ and the number of mass estimation ne as the following six combinations: $\rho = 8$ and $ne = 100$; $\rho = 8$ and $ne = 1{,}000$; $\rho = 256$ and $ne = 10$; $\rho = 256$ and $ne = 100$; $\rho = 256$ and $ne = 500$; $\rho = 256$ and $ne = 1{,}000$. On the other hand, IForest and Mass are based on random subsampling, which makes their performance unstable. In an attempt to get more stable statistics, for each dataset and parameter setting we run (30 times) and compute the average AUC and F1-score. In Tables II and III, we document the average AUC/F1-score of the best four (out of all six) parameter settings for each dataset.

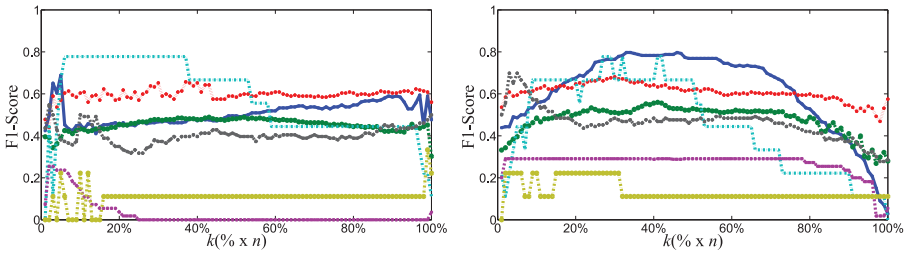
9.2. Comparison of Average Performance

In this section, we evaluate our proposed FDD and LAD, as well as the other seven anomaly detection algorithms. Table II documents the average AUC of each method across the corresponding tuning parameters and the relative p-value with regard to FDD, whereas Table III records the average F1-score of each method across the corresponding tuning parameters and the relative p-value with regard to FDD.



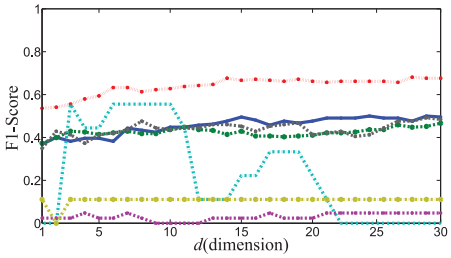
(a) LLE, different size of neighborhood (k)
(b) ISM, different size of neighborhood (k)
(c) LLE, different number of dimensions (d)
(d) ISM, different number of dimensions (d)
(e) LOF, different size of neighborhood (k)
(f) LOCI, different size of neighborhood (k)
(g) StrOUD, different size of neighborhood (k)
(h) LAD, different size of neighborhood (k)
(i) LAD, different diffusion time (t)
(j) FDD, different temperature (T)

Fig. 11. AUC stability with different parameters.

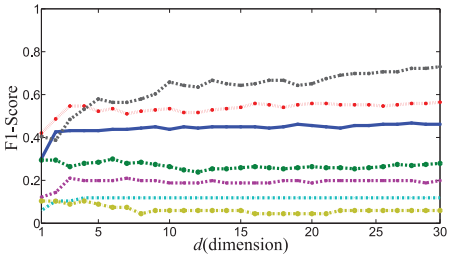


(a) LLE, different size of neighborhood (k)

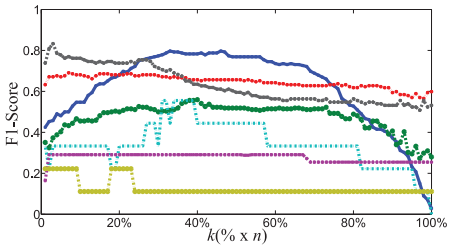
(b) ISM, different size of neighborhood (k)



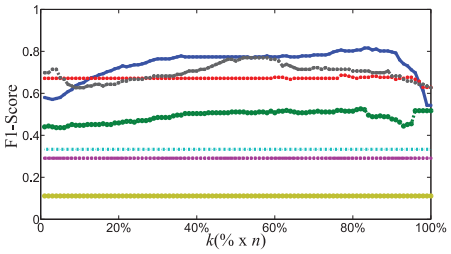
(c) LLE, different number of dimensions (d)



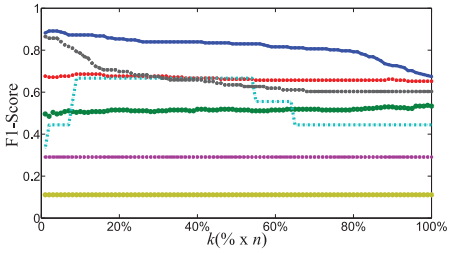
(d) ISM, different number of dimensions (d)



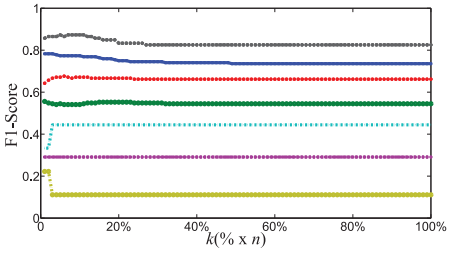
(e) LOF, different size of neighborhood (k)



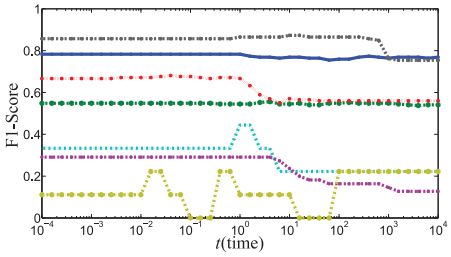
(f) LOCI, different size of neighborhood (k)



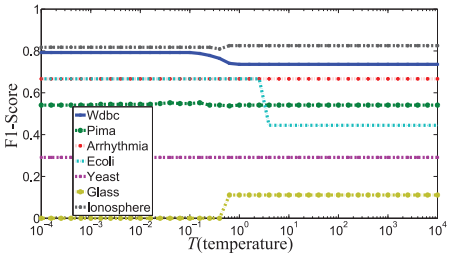
(g) StrOUD, different size of neighborhood (k)



(h) LAD, different size of neighborhood (k)



(i) LAD, different diffusion time (t)



(j) FDD, different temperature (T)

Fig. 12. F1-score stability with different parameters.

Table II. Comparison of Average AUC of Our FDD and LAD, and the Other Seven Popular Methods across Their Corresponding Parameters (indicated in the parentheses after each method in the first row)

Dataset	LLE(k)	ISM(k)	LOF(k)	LOCI(k)	Mass	IForest	StrOUD(k)	LAD(k)	FDD(T)
breastcancer	0.8448	0.6411	0.6423	0.8637	0.3572	0.9914	**0.9926**	0.9820 (4)	0.9870 (3)
wdbc	0.6502	0.7280	0.7289	0.8548	0.7823	0.8218	**0.9285**	0.9005 (3)	0.9049 (2)
pima	0.5964	0.6224	0.6188	0.6165	0.6094	0.6848	0.6434	0.7101 (2)	**0.7119** (1)
arrhythmia	0.6720	0.6942	0.7356	0.7475	0.6924	0.7425	**0.7530**	0.7192 (6)	0.7472 (3)
arcene	0.5118	0.5300	0.4417	0.3964	**0.5710**	0.3112	0.3242	0.5551 (2)	0.5551 (2)
prostatetumor	0.4637	0.4732	0.4757	0.4667	0.4203	0.4090	0.4279	0.5314 (2)	**0.5359** (1)
gse24417	0.5055	0.5176	0.5743	0.5625	0.5693	0.5688	0.5860	0.5860 (2)	**0.5895** (1)
hayesroth	0.8413	0.5713	0.5828	0.6438	0.6843	0.9846	0.7020	0.9903 (2)	**0.9905** (1)
ecoli	0.8502	0.8279	0.8178	0.8754	0.7690	0.8748	0.8972	0.8960 (3)	**0.9052** (1)
yeast	0.5146	0.6160	**0.6285**	0.6063	0.5665	0.6127	0.6149	0.6115 (6)	0.6212 (2)
abalone	0.6612	0.6793	0.6720	0.6767	0.6345	0.6933	0.6989	0.7299 (2)	**0.7332** (1)
glass	0.6346	0.6206	0.6205	0.7647	**0.8813**	0.6940	0.7962	0.8732 (3)	0.8737 (2)
ionosphere	0.5814	0.6206	0.7719	0.8482	0.8167	0.8517	0.8605	**0.9335** (1)	0.9253 (2)
pageblocks	0.7619	0.7740	0.7043	0.7759	0.8813	0.8561	0.5511	0.8893 (2)	**0.8939** (1)
magic04	0.5883	0.6466	0.6599	0.6747	0.6798	0.7214	0.6775	0.7286 (2)	**0.7520** (1)
Average	0.6452**	0.6396**	0.6390**	0.6916**	0.6611**	0.7214**	0.6861**	0.7758* (2)	**0.7818** (1)

Note: For each dataset, the bold-faced number indicates the best method, and the numbers in the parentheses indicate the ranks of our FDD and LAD. The Average row presents the average AUC of each method across all datasets, respectively. *, p-value of 5% or lower in the statistical significance test with regard to FDD. **, p-value of 1% or lower in the statistical significance test with regard to FDD.

Table III. Comparison of Average F1-Score of our FDD and LAD, and the Other Seven Methods across Their Corresponding Parameters (indicated in the parentheses after each method in the first row)

Dataset	LLE(k)	ISM(k)	LOF(k)	LOCI(k)	Mass	IForest	StrOUD(k)	LAD(k)	FDD(T)
breastcancer	0.7282	0.5277	0.5235	0.6971	0.2468	0.9411	**0.9439**	0.9144 (4)	0.9351 (3)
wdbc	0.5064	0.6149	0.6159	0.7406	0.6544	0.6745	**0.8161**	0.7445 (3)	0.7610 (2)
pima	0.4524	0.4772	0.4734	0.4892	0.4361	0.5289	0.5154	**0.5458** (1)	0.5426 (2)
arrhythmia	0.5968	0.6086	0.6469	**0.6711**	0.6067	0.6535	0.6654	0.6630 (4)	0.6667 (2)
arcene	0.4659	0.4639	0.3802	0.3786	0.4838	0.2730	0.3025	0.4882 (2)	**0.4941** (1)
prostatetumor	0.4808	0.4894	0.5046	0.5082	0.4849	0.4794	0.4569	0.5412 (2)	**0.5503** (1)
gse24417	0.3000	0.3017	0.3874	0.4005	0.3772	0.3766	0.4038	0.3774 (5)	**0.4200** (1)
hayesroth	0.5980	0.3233	0.3387	0.3760	0.4113	0.8888	0.3811	0.8997 (2)	**0.9042** (1)
ecoli	**0.5778**	0.4544	0.3444	0.3333	0.2570	0.5681	0.5578	0.4422 (6)	0.5691 (2)
yeast	0.0278	**0.2962**	0.2778	0.2909	0.0627	0.2505	0.2909	0.2909 (2)	0.2909 (2)
abalone	0.2061	0.2072	0.2845	0.2909	0.2891	0.3057	0.2402	0.3005 (3)	**0.3089** (1)
glass	0.1067	0.1367	0.1278	0.1111	**0.1875**	0.1083	0.1111	0.1133 (4)	0.1111 (5)
ionosphere	0.4047	0.4633	0.6312	0.6975	0.5955	0.6631	0.6587	**0.8331** (1)	0.8216 (2)
pageblocks	0.1991	0.2251	0.2009	0.2165	**0.3675**	0.2449	0.1905	0.3448 (3)	0.3588 (2)
magic04	0.4002	0.4079	0.4951	0.5166	0.5070	**0.5836**	0.4981	0.5081 (4)	0.5775 (2)
Average	0.4035**	0.3980**	0.4177**	0.4479**	0.3978**	0.5026**	0.4500**	0.5338* (2)	**0.5542** (1)

Note: For each dataset, the bold-faced number indicates the best method, and the numbers in the parentheses indicate the ranks of our FDD and LAD. The Average row presents the average F1-score of each method across all the datasets, respectively. *, p-value of 5% or lower in the statistical significance test with regard to FDD. **, p-value of 1% or lower in the statistical significance test with regard to FDD.

In Table II, our proposed FDD and LAD show the first and the second best average AUC score (0.7818 and 0.7758). They boost up the AUC close to or more than 8% compared with the best performance (0.7214 from IForest) among the other methods. For most of the datasets, FDD and LAD have the best or very close to the best performance. Specifically, FDD is the ranked in the top three for all datasets; meanwhile, our LAD, although not all the time, outperforms the other algorithms in most cases. In fact, in the only two cases when LAD is not among the best four ranks (arrhythmia and yeast), LAD still reaches more than 95% of the best AUC results.

Although some algorithms, such as LOF, IForest, and StrOUD, are more efficient in measuring the anomalous score, their methodologies are based on Euclidean space and therefore are under the curse of dimensionality. As the number of features increases, their performance drops significantly on the datasets, such as arcene and prostatetumor. The manifold-based algorithms, such as LLE and ISM, are to reduce the vulnerability of simple kernel under the high dimensions. Despite the fact that the LOF measurement on the projection of LLE and ISM shows better quality compared with LOF on the input space, it suffers from the inferior manifold reconstruction. Comparably, our FDD and LAD, built upon optimal embedding structure derived from solid physics theory, provide stronger capability of detecting anomalies in terms of AUC.

As for the macro paired t-tests across all datasets in Table II, compared with all the other algorithms, our quantum theory–based FDD has an extremely small p-value (less than 1%). Even compared with the other proposed method LAD, FDD still has a very small p-value of less than 5% with statistical significance. This proves that our FDD has the most stable average performance in terms of AUC.

To have a comprehensive test with a more practical view, we also measure the F1-score and document this in Table III. Here we only record the F1-score of the anomalous subset (class) because it is the focus of anomaly detection. Although the heat diffusion–based LAD shows slightly fluctuating performance compared with its AUC production in Table II, it still surpasses (0.5338 vs. 0.5026) IForest, the best among the other algorithms (except FDD) for more than 6%. But the quantum theory–based FDD acquires the best F1-score (0.5542), which improves more than 10% based on IForest. Furthermore, the same as demonstrated in AUC, the F1-score by FDD is almost persistently (except for glass) ranking in the top three among all algorithms. In the actual application of anomaly detection, the users tend to focus on the detected anomalous subset instead of the whole label distribution; therefore, the F1-score tells more of a story than AUC. In this case, FDD shows more convincing quality in terms of the F1-score. Thus, our proposed FDD has the capability of providing the most desirable label results of anomalies in practice.

Compared with the basic LOF algorithm, IForest shows a passable AUC and F1-score on average, which supports the argument that it is able to take both global and local contexts into consideration. This is different from kNN-based methods (LOF and LOCI), which only are concerned with instance-wise local context. Compared with LOF, LOCI has more than an 8% better AUC and more than a 7% better F1-score. This moderately stable and stronger performance comes from the built-in concept of a multigranularity deviation factor [Papadimitriou et al. 2003]. Although Mass cannot always maintain competitive quality of anomalous score measurement, it has the fastest computation speed compared with all other competitors.

9.3. Comparison of Stability

To systematically manifest the stability against parameter tuning of each method, we run experiments for LLE, ISM, LOF, LOCI, StrOUD, and our proposed FDD and LAD across their corresponding parameter tuning, respectively, and record the AUC

in Figure 11 and F1-score in Figure 12. We select seven small datasets—wdbc, pima, arrhythmia, ecoli, yeast, glass, and ionosphere—for the stability test. In theory, smaller datasets should be more sensitive to the change of scaling parameters. Therefore, these seven datasets are the more effective choices to show whether the algorithms perform robustly during the adjustment of their parameters.

For the size of neighborhood k and the number of embedding dimension d, LLE undergoes fluctuation especially on wdbc, ecoli, and glass. It is mainly because LLE has strong assumption that the data is densely sampled and the embedding structure is locally approximately linear, yet it is not true for many real-world datasets. Similarly, ISM's results vary dramatically as k changes, especially for ecoli, glass, and ionosphere, although later ISM is comparably stable while tuning d. This is because ISM is highly vulnerable to the local data perturbation, as the embedding given by the ISM tends to recover the geodesic distances between points on the manifold, which is very locally sensitive compared with random walk [Lafon et al. 2006; Van der Maaten et al. 2009].

Compared with LOF, LOCI performs robustly with different k, because its proposed multigranularity deviation factor can more intuitively cope with local density variations in the feature space and detect both isolated anomalies as well as outlying clusters [Papadimitriou et al. 2003]. LOF, although it occasionally beats LOCI with certain k, shows seriously unstable performance as k changes, which can be simply explained as follows: LOF is based on a direct normalization of anomaly scores for an inadaptive neighborhood.

StrOUD demonstrates not only its effectiveness and efficiency (since it is totally based on the input space without any projection) but also its AUC stability during the change of k. However, in terms of AUC result shown in Figure 11(g), the curves have different patterns: StrOUD reaches higher AUC with smaller k on the wdbc, glass, and ionosphere datasets, but it has a better AUC result with larger k on pima and ecoli. In the test of F1-score in Figure 12(g), StrOUD shows serious instability on ecoli as k changes, partly because StrOUD is principally built upon Euclidean distance on the input space, which cannot faithfully reveal the intrinsic dissimilarity and density on the nonlinear distributed data. Furthermore, it becomes even worse on the more complex datasets with large numbers of features, as already confirmed in Tables II and III.

Compared with the preceding algorithms, our proposed LAD shows the best stability against the change of k, as demonstrated in both AUC (Figure 11(h)) and F1-score (Figure 12(h)). This is because LAD has an inherent relationship with heat diffusion and random walk. More specifically, LAD has a strong probabilistic interpretation, which provides a power against noise appearance or neighborhood size perturbation, as long as they are not severe enough to perturb the general neighborhood statistics.

Importantly, we test the performance of our FDD and LAD with different physical parameters: heat diffusion time t in LAD and environmental temperature T. As we have already described in Figure 7 and Section 8.1, LAD may lose the power of local density description especially for local anomalies when t goes large, which means overdiffusion. Therefore, the AUC curves by LAD of ecoli and yeast significantly drop in Figure 11(i). Likewise, the F1-score by LAD in Figure 12(i) shows comparably unstable trends as t increases. On the contrary, Figure 11(j) establishes the robustness of FDD. Compared with LAD, FDD has more potential in combating off-the-sweet-spot physical parameter since it is constructed on polarized manifold space and has additional stabilizing factors that help to balance the riskiness from extreme cases. Another thing worth mentioning is that in Figure 12(j), FDD does not always maintain strong stability across all of the datasets. But comparatively, our proposed FDD still retains a certain level of anomaly detection quality as parameter changes. And most importantly, FDD outperforms the existing baselines in terms of average performance and

Table IV. Comparison of Average AUC by LAD with Different Laplacians

Dataset	LAD with L_{nn}	LAD with L_{sym}	LAD with L_{fp}	LAD with L_{lbn}	LAD with L_{rw}
breastcancer	0.9821(4)	0.9861(2)	0.9865(1)	0.9822(3)	0.9820(5)
wdbc	0.7241(5)	0.9044(2)	0.9047(1)	0.9015(3)	0.9005(4)
pima	0.5806(5)	0.7130(1)	0.7121(2)	0.7103(3)	0.7101(4)
arrhythmia	0.7473(1)	0.7403(2)	0.7261(3)	0.7183(5)	0.7192(4)
arcene	0.4316(5)	0.5369(3)	0.5101(4)	0.5523(2)	0.5551(1)
prostatetumor	0.5136(5)	0.5396(3)	0.5608(2)	0.5808(1)	0.5314(4)
gse24417	0.5587(4)	0.5888(1)	0.5791(3)	0.5426(5)	0.5860(2)
hayesroth	0.9916(3)	0.9926(2)	0.9929(1)	0.9903(4)	0.9903(4)
ecoli	0.4084(5)	0.8949(4)	0.8951(3)	0.8955(2)	0.8960(1)
yeast	0.4135(5)	0.6114(4)	0.6115(1)	0.6115(1)	0.6115(1)
abalone	0.6335(5)	0.7668(1)	0.7667(2)	0.7666(3)	0.7299(4)
glass	0.7132(5)	0.8765(1)	0.8744(2)	0.8732(3)	0.8732(3)
ionosphere	0.9466(1)	0.9336(2)	0.9324(5)	0.9335(3)	0.9335(3)
pageblocks	0.7091(5)	0.7958(2)	0.7932(4)	0.7947(3)	0.8893(1)
magic04	0.6513(5)	0.7300(1)	0.7289(3)	0.7296(2)	0.7286(4)
Average	0.6681(5)	0.7740(2)	0.7716(4)	0.7722(3)	0.7758(1)

Note: For each dataset, the numbers in the parentheses indicate the ranks of each Laplacian. The Average row presents the average AUC of each Laplacian across all datasets, respectively.

steadiness with the purpose of detecting anomalies. The robustness property is equally significant for domain experts who do not have a strong machine learning background. Since there are not many clues for tuning the traditional yet unstable algorithms such as LLE, ISM and LOF, it is much more comfortable for domain experts to utilize robust anomaly detection algorithms for the domain data analysis. Therefore, our proposed FDD is very hands-on and effective on many real-world applications.

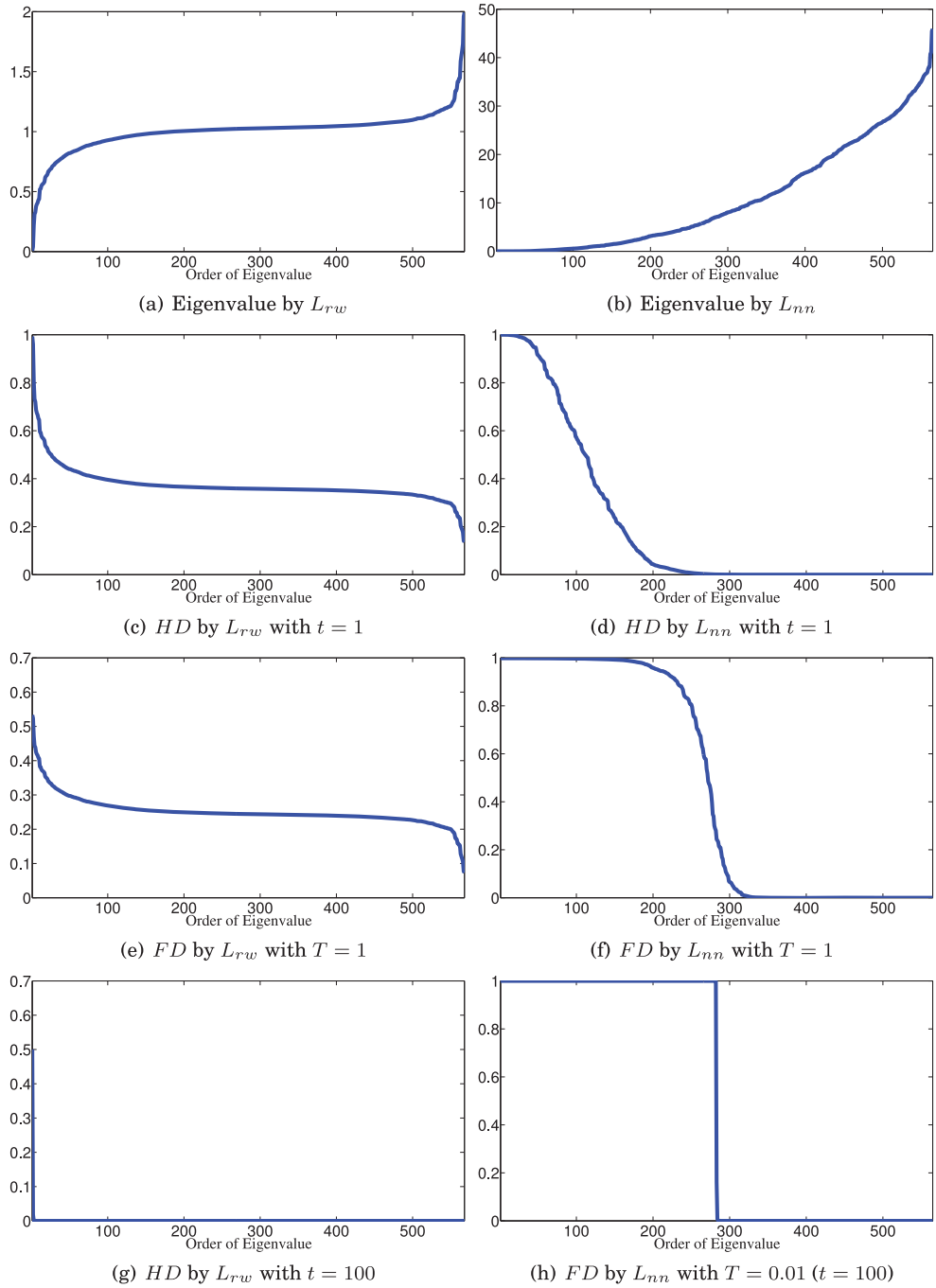
9.4. Comparison of Different Laplacians

In Sections 3.1 and 7.2, we introduced our selection of Laplacians in LAD and FDD. Here we analyze the reasons through experiments of LAD and FDD with the five Laplacians and 15 datasets, respectively. To save space, we only list the average AUC in Tables IV and V.

Table IV shows the effect of different Laplacians on LAD, and L_{rw} has the best average performance. We also note that except for L_{nn}, there is not too much difference among the four (normalized) Laplacians. Here is a detailed analysis:

—The similar performance of L_{rw}, L_{sym}, L_{fp}, and L_{lbn} can be explained by the use of an umbrella operator in LAD, which gives attention to the weighted distance between each point and its neighborhood. Therefore, as long as the eigenvalues are normalized, and there is deviation between the normalized eigencomponents, especially the corresponding value of any anomaly and its surrounding normal instances in the eigenvectors, LAD can capture such deviation regardless of the choice of normalized Laplacians. The reason we emphasize L_{rw} on LAD is that LAD is based on heat diffusion, and heat diffusion in classical physics has better interpretation with particles' random walk.

—The reason why LAD fails on top of L_{nn} relates to the unnormalized eigenvalue distribution. Figure 13(a) and 13(b) shows the eigenvalues (sorted in ascending order) derived from L_{rw} and L_{nn} on dataset wdbc correspondingly. Without normalization, the eigenvalues in Figure 13(b) increase exponentially, and only a small portion of eigencomponents in Heat Diffusion (Figure 13(d)) are given stable and large enough weights, whereas the other eigencomponents, even those informative,



(a) Eigenvalue by L_{rw} (b) Eigenvalue by L_{nn}

(c) HD by L_{rw} with $t = 1$ (d) HD by L_{nn} with $t = 1$

(e) FD by L_{rw} with $T = 1$ (f) FD by L_{nn} with $T = 1$

(g) HD by L_{rw} with $t = 100$ (h) FD by L_{nn} with $T = 0.01$ ($t = 100$)

Fig. 13. Eigenvalue by L_{rw} and L_{nn}, and the corresponding weighted function of HD (Equation (27)) and FD distribution (Equation (23)). The dataset is wdbc.

Table V. Comparison of Average AUC by FDD with Different Laplacians

Dataset	FDD with L_{nn}	FDD with L_{sym}	FDD with L_{fp}	FDD with L_{lbn}	FDD with L_{rw}
breastcancer	0.9870(1)	0.9187(2)	0.9088(3)	0.3171(5)	0.5612(4)
wdbc	0.9049(1)	0.8236(2)	0.5337(3)	0.2464(4)	0.2139(5)
pima	0.7119(1)	0.6886(3)	0.7027(2)	0.4344(5)	0.6032(4)
arrhythmia	0.7472(1)	0.7384(2)	0.7050(3)	0.6131(5)	0.6639(4)
arcene	0.5551(1)	0.5538(2)	0.5384(4)	0.5433(3)	0.4648(5)
prostatetumor	0.5359(5)	0.5415(4)	0.5792(2)	0.5427(3)	0.5904(1)
gse24417	0.5895(1)	0.5815(2)	0.5529(3)	0.4964(5)	0.5504(4)
hayesroth	0.9905(1)	0.6990(4)	0.9418(2)	0.1000(5)	0.9183(3)
ecoli	0.9052(1)	0.8913(3)	0.8977(2)	0.0832(5)	0.8906(4)
yeast	0.6212(2)	0.6205(3)	0.6614(1)	0.3654(5)	0.5927(4)
abalone	0.7332(2)	0.7526(1)	0.5190(3)	0.4291(4)	0.2504(5)
glass	0.8737(2)	0.7702(3)	0.8883(1)	0.3149(5)	0.6141(4)
ionosphere	0.9253(1)	0.6320(2)	0.3558(4)	0.1736(5)	0.5781(3)
pageblocks	0.8939(1)	0.6888(2)	0.2993(4)	0.4201(3)	0.2247(5)
magic04	0.7520(1)	0.7502(2)	0.3411(4)	0.6722(3)	0.3357(5)
Average	0.7818(1)	0.7100(2)	0.6283(3)	0.3835(5)	0.5268(4)

Note: For each dataset, the numbers in the parentheses indicate the ranks of each Laplacian. The Average row presents the average AUC of each Laplacian across all datasets, respectively.

go away quickly. Comparatively, the eigenvalues derived from L_{rw} show an inverse-hyperbolic-tangent-like distribution. So the consequent Heat Diffusion (Figure 13(c) with $t = 1$) gives very high weights on the first few eigencomponents and fewer but nonnegligible weights on most of the following ones. Therefore, the normalized Laplacians, such as $L_{rw,}$ weight the eigencomponents more safely, even conservatively, compared with L_{nn}.

—Note that if diffusion time goes too large, Heat Diffusion will only emphasize the first few eigencomponents and ignore all of the following, as shown in Figure 13(g). Therefore, HKS and even LAD fail with too large of a t.

In Section 7.2, we proved that L_{nn} is the best choice for FDD. Table V confirms that L_{nn} has the best average performance on FDD. Here we give a brief analysis:

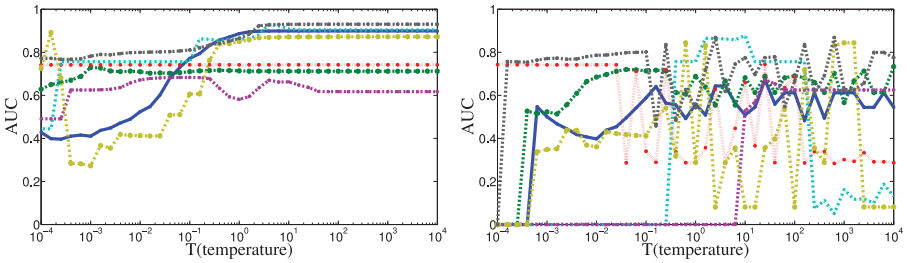
—FDD, different from LAD, does not use an umbrella operator but instead relies on the energy distribution functions and the eigendecomposition on Laplacians. Hence, the Laplace operator becomes more essential on the construction of FDD.

—In Figure 13(f) and 13(h), FD distribution function with L_{nn} robustly assigns similar and very stable weight to the first 200+ eigencomponents of the wdbc dataset, regardless of the value of T. Comparatively, FD with normalized Laplacians such as L_{rw} (Figure 13(e)) embraces too many eigencomponents, even including noisy ones. Without the help of an umbrella operator, these noisy components will bring unstable anomaly detection results, as shown in Table V.

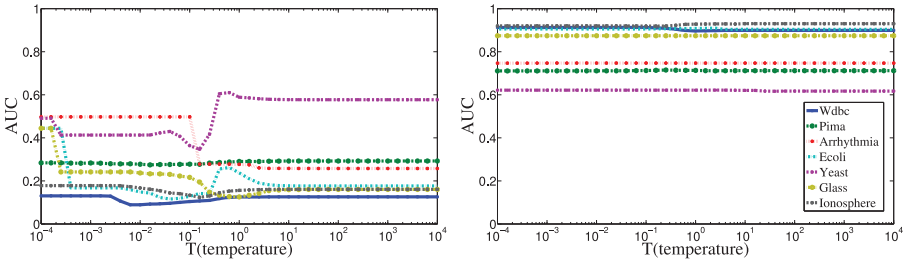
9.5. Comparison of Energy Distribution Functions

To have a better understanding of different distribution function effects (introduced in Section 7.1) on anomaly detection, we test their stability. Here we integrate all four functions, namely MB distribution, BE distribution, GD, and our chosen FD distribution, into Equation (21) with L_{nn} operator and calculate the anomaly detection scores in AUC and F1-score.

The results are illustrated in Figures 14 and 15. The stability of GD is reasonable, but the scores are apparently lower than the other three. BE shows the most fluctuant

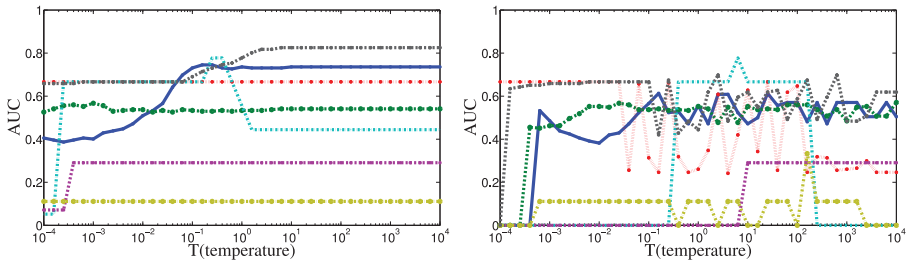


(a) Equation (21) with MB, different T (b) Equation (21) with BE, different T

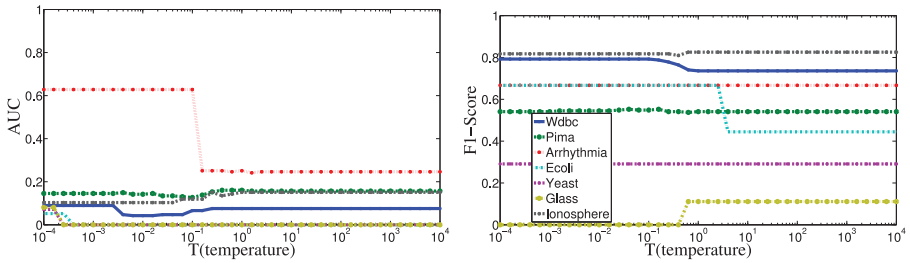


(c) Equation (21) with GD, different T (d) FDD (Equation (21) with FD), different T

Fig. 14. AUC stability with different energy distribution functions.



(a) Equation (21) with MB, different T (b) Equation (21) with BE, different T



(c) Equation (21) with GD, different T (d) FDD (Equation (21) with FD), different T

Fig. 15. F1-score stability with different energy distribution functions.

results in both AUC and F1-score because it does not have the smoothing term "plus one." MB suffers from extremely small temperature T, which is similar to the fact that HKS suffers from large diffusion time t; therefore, generally it has a dropping trend when T becomes smaller. Our FDD, although it not always maintains the best performance, has the best average result and the most stability in both AUC and F1-score.

Table VI. Comparison of AUC between Full and Fast Versions of LAD and FDD

Dataset	LAD(AGK)	LAD$_f(GAU)$	FDD(AGK)	FDD$_f(GAU)$
breastcancer	0.9820	0.9874	0.9870	0.4856
wdbc	0.9005	0.9681	0.9049	0.4883
pima	0.7101	0.4804	0.7119	0.5105
arrhythmia	0.7192	0.5666	0.7472	0.4973
arcene	0.5551	0.5510	0.5551	0.5055
prostatetumor	0.5314	0.6185	0.5359	0.5041
gse24417	0.5860	0.4449	0.5895	0.5054
hayesroth	0.9903	0.9311	0.9905	0.4908
ecoli	0.8960	0.9377	0.9052	0.8763
yeast	0.6115	0.6013	0.6212	0.5548
abalone	0.7299	0.7387	0.7332	0.4250
glass	0.8732	0.8232	0.8737	0.7545
ionosphere	0.9335	0.8241	0.9253	0.3000
pageblocks	0.8893	0.7188	0.8939	0.4992
magic04	0.7286	0.7307	0.7520	0.5064
Average	0.7758	0.7282	0.7818	0.5269

Table VII. Comparison of Running Time (in seconds)

Dataset	LAD(AGK)	LAD$_f(GAU)$	LOCI	StrOUD	IForest
breastcancer	0.9125	0.2732	45.5966	6.2400	4.8025
wdbc	0.8137	0.2822	129.7696	3.9162	3.8211
pima	1.3304	0.4589	231.3238	6.4149	5.3805
arrhythmia	0.5204	0.2265	78.8057	3.1139	3.8083
arcene	0.1772	0.0842	12.6367	8.3564	1.4613
prostatetumor	0.1442	0.0871	3.1904	7.0358	0.5022
gse24417	0.4080	0.1791	65.8869	8.6048	3.2912
hayesroth	0.1058	0.0735	4.2058	0.2284	0.6599
ecoli	0.3294	0.1223	32.6413	1.4403	2.7602
yeast	5.9344	0.5448	851.1367	25.5289	4.8630
abalone	72.5052	8.8877	17112.6534	192.0656	5.9136
glass	0.1414	0.0867	14.7054	0.6044	1.7353
ionosphere	0.3251	0.1003	41.3065	1.4513	2.9742
pageblocks	92.3223	35.4092	33725.4086	320.3760	6.3389
magic04	1297.0366	471.1242	252425.9877	864.9672	6.6197
Average	98.2004	34.5293	20318.3503	96.6896	3.6621

9.6. Comparison of Efficiency and Effectiveness

In this section, we analyze the efficiency and effectiveness of LAD and FDD with a small portion of eigencomponents. Additionally, to obtain a short amount of running time, we only use GAU ($O(n^2m)$) instead of AGK ($O(n^2m^2)$) here. Suppose that the size of dataset (number of instances) is n, and the first $max(\lfloor n/50 \rfloor, 10)$ eigencomponents are used to compute LAD and FDD, which are noted as LAD$_f(GAU)$ and FDD$_f(GAU)$. The AUC results are shown in Table VI. LAD$_f(GAU)$ obtains about 94% of AUC by full-version LAD(AGK), whereas FDD$_f(GAU)$ only gets 67% of the full FDD(AGK). Apparently, LAD does not suffer a lot from a small amount of eigenvectors compared with FDD, which can be explained by the effect of the umbrella operator and the illustration results in Figure 13. Table VII shows the running time comparison of a few algorithms. The parameter settings are documented in Section 9.1.4. Specifically, IForest's running time is measured by the average of the best four parameter settings listed in

Section 9.1.4. $\text{LAD}_f(GAU)$ is two times faster than LAD on average and also is more efficient than StrOUD and LOCI. The most efficient one among these five algorithms is IForest, which is extremely fast on large datasets, such as magic04 and pageblocks. However, it is worth noting that $\text{LAD}_f(GAU)$ is more efficient on the small datasets.

10. CONCLUSION

This article documents the physics-based methodology of unsupervised anomaly detection. The first algorithm that we propose is LAD, which is based on heat diffusion and a scale-dependent umbrella operator. Its capability of representing local density relies on a short time of heat dissipation and an informative neighborhood that is guaranteed by the scale-dependent umbrella operator. FDD is another anomaly detection method that we proposed. It is built upon a polarized manifold projection and a quantum motion probability measured by Fermi-Dirac energy distribution. We also analyze the utilization of AGK and the best choice of graph Laplacian with the purpose of anomaly detection. Compared with the existing algorithms, our proposed LAD and FDD exhibit better average performance and stability in our extensive experiments. Moreover, FDD demonstrates its robustness across different physics scaling parameters compared with LAD. We expect that the proposed physics-based methodology is useful for most types of data distribution. Nonetheless, much more extensive experiments are still required to validate this conjecture, which is part of our near-future research. Another direction is to investigate its possible connection with global structure and pattern mining, such as clustering and feature selection.

REFERENCES

A. Agovic, A. Banerjee, A. R. Ganguly, and V. Protopopescu. 2007. Anomaly detection in transportation corridors using manifold embedding. In *Proceedings of the 1st International Workshop on Knowledge Discovery from Sensor Data.*

M. Aubry, U. Schlickewei, and D. Cremers. 2011. The wave kernel signature—a quantum mechanical approach to shape analysis. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1626–1633.

R. Badeau, B. David, and G. Richard. 2005. Fast approximated power iteration subspace tracking. *IEEE Transactions on Signal Processing* 53, 8, 2931–2941.

D. Barbara, C. Domeniconi, and J. P. Rogers. 2006. Detecting outliers using transduction and statistical testing. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. 55–64.

G. Blanchard, G. Lee, and C. Scott. 2010. Semi-supervised novelty detection. *Journal of Machine Learning Research* 11, 2973–3009.

M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. 2000. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00)*. 93–104.

V. Chandola, A. Banerjee, and V. Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys* 41, 3, 1–72.

R. R. Coifman and S. Lafon. 2006. Diffusion maps. *Applied and Computational Harmonic Analysis* 21, 1, 5–30.

T. De Vires, S. Chawla, and M. Houle. 2010. Finding local anomalies in very high dimensional space. In *Proceedings of the 2010 IEEE 10th International Conference on Data Mining (ICDM'10)*. 128–137.

M. Desbrun, M. Meyer, P. Schroder, and A. Barr. 1999. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'99)*. 317–324.

J. Gao, H. Cheng, and P. N. Tan. 2006. Semi-supervised outlier detection. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC'06)*. 635–636.

G. Greenstein and A. Zajonc. 2006. *The Quantum Challenge: Modern Research on the Foundations of Quantum Mechanics*. Jones and Bartlett Publishers.

A. Grigoryan. 1999. Estimates of heat kernels on Riemannian manifolds. In *Proceedings of the ICMS Instructional Conference on Spectral Theory and Geometry*. London Mathematical Society Lecture Notes, Vol. 273. Cambridge University Press, 140–225.

I. Guyon, J. Li, T. Mader, P. A. Pletscher, G. Schneider, and M. Uhr. 2007. Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters* 28, 12, 1438–1444.

J. A. Hartigan and M. A. Wong. 1978. Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics* 28, 1, 100–108.

D. Hawkins. 1980. *Identification of Outliers*. Chapman and Hall, London.

K. Hempstalk, E. Frank, and I. H. Witten. 2008. One-class classification by combining density and class probability estimation. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*.

D. Horn and A. Gottlieb. 2001. The method of quantum clustering. In *Proceedings of NIPS*. 769–776.

D. Horn and A. Gottlieb. 2002. Algorithm for data clustering in pattern recognition problems based on quantum mechanics. *Physical Review Letters* 88, 1, 018702.

E. Hsu. 2002. *Stochastic Analysis on Manifolds*. AMS Graduate Series in Mathematics, Vol. 38. American Mathematical Society, Providence, RI.

H. Huang, H. Qin, S. Yoo, and D. Yu. 2012a. Local anomaly descriptor: A robust unsupervised algorithm for anomaly detection based on diffusion space. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 405–414.

H. Huang, H. Qin, S. Yoo, and D. Yu. 2012b. A new anomaly detection algorithm based on quantum mechanics. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM'12)*. 900–905.

H. Huang, S. Yoo, H. Qin, and D. Yu. 2011. A robust clustering algorithm based on aggregated heat kernel mapping. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM'11)*. 270–279.

S. Lafon, Y. Keller, and R. R. Coifman. 2006. Data fusion and multi-cue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 11, 1784–1797.

Y. Lipman, R. Rustamov, and T. Funkhouser. 2010. Biharmonic distance. *ACM Transactions on Graphics* 29, 3, Article No. 27.

F. T. Liu, K. M. Ting, and Z. H. Zhou. 2008. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*. 413–422.

F. T. Liu, K. M. Ting, and Z. H. Zhou. 2011. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data* 6, 1, Article No. 3.

U. V. Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4, 395–416.

C. Marzban. 2004. A comment on the ROC curve and the area under it as performance measures. *Weather and Forecasting* 19, 6, 1106–1114.

N. Nasios and A. G. Bors. 2007. Kernel-based classification using quantum mechanics. *Pattern Recognition* 40, 3, 875–889.

K. Noto, C. E. Brodley, and D. Slonim. 2010. Anomaly detection using an ensemble of feature models. In *Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM'10)*. 953–958.

V. Y. Pan and Z. Q. Chen. 1999. The complexity of the matrix eigenproblem. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC'99)*. 507–516.

S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. 2003. LOCI: Fast outlier detection using the local correlation integral. In *Proceedings of the 19th International Conference on Data Engineering, 2003*. 315–326.

B. Pogorelc and M. Gams. 2010. Discovery of gait anomalies from motion sensor data. In *Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI'10)*. 331–336.

V. Popovici, W. Chen, B. G. Gallas, and C. Hatzis. 2010. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Research* 12, 1, R5.

D. M. W. Powers. 2011. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2, 1, 37–63.

H. Qiu and E. R. Hancock. 2007. Clustering and embedding using commute times. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 11, 1873–1890.

J. W. Richards, P. E. Freeman, A. B. Lee, and C. M. Schafer. 2009. Accurate parameter estimation for star formation history in galaxies using SDSS spectra. *Monthly Notices of the Royal Astronomical Society* 399, 2, 1044–1057.

A. Singer and R. R. Coifman. 2008. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis* 25, 2, 226–239.

A. Singer and R. R. Coifman. 2012. Anisotropic diffusion on sub-manifolds with application to earth structure classification. *Applied and Computational Harmonic Analysis* 32, 2, 280–294.

A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. 2005. Gene Expression Model Selector. Retrieved August 21, 2014, from http://www.gems-system.org/.

J. Sun, M. Ovsjanikov, and L. Guibas. 2009. A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing (SGP'09)*. 1383–1392.

Z. Syed and I. Rubinfeld. 2010. Unsupervised risk stratification in clinical datasets: Identifying patients at risk of rare outcomes. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. 1023–1030.

G. Taubin. 1995. A signal processing approach to fair surface design. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'95)*. 351–358.

K. M. Ting, G. T. Zhou, F. T. Liu, and J. S. Tan. 2010. Mass estimation and its applications. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. 989–998.

L. Van der Maaten, E. Postma, and J. van der Herik. 2009. *Dimensionality Reduction: A Comparative Review*. Technical Report TiCC-TR 2009-005. Tilburg University, Netherlands.

M. Weinstein. 2010. Strange bedfellows: Quantum mechanics and data mining. *Nuclear Physics B-Proceedings Supplements* 199, 1, 74–84.

M. Wu and J. Ye. 2009. A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE Transactions on Pattern Analysis and Machine Learning* 31, 11, 2088–2092.

K. Zhang, M. Hutter, and H. Jin. 2009. A new local distance-based outlier detection approach for scattered real-world data. In *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, Vol. 5476, 813–822.

X. Zhu, X. Wu, and C. Zhang. 2009. Vague one-class learning for data streams. In *Proceedings of the 9th International Conference on Data Mining (ICDM'09)*. 657–666.

D. W. Zimmerman. 1997. A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics* 22, 3, 349–360.

Received December 2013; revised April 2014; accepted June 2014