Reverse Nearest Neighbors in Unsupervised Distance-Based Outlier Detection

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović

Abstract—Outlier detection in high-dimensional data presents various challenges resulting from the "curse of dimensionality." A prevailing view is that distance concentration, i.e., the tendency of distances in high-dimensional data to become indiscernible, hinders the detection of outliers by making distance-based methods label all points as almost equally good outliers. In this paper we provide evidence supporting the opinion that such a view is too simple, by demonstrating that distance-based methods can produce more contrasting outlier scores in high-dimensional settings. Furthermore, we show that high dimensionality can have a different impact, by reexamining the notion of reverse nearest neighbors in the unsupervised outlier-detection context. Namely, it was recently observed that the distribution of points' reverse-neighbor counts becomes skewed in high dimensions, resulting in the phenomenon known as *hubness*. We provide insight into how some points (antihubs) appear very infrequently in k-NN lists of other points, and explain the connection between antihubs, outliers, and existing unsupervised outlier-detection methods. By evaluating the classic k-NN method, the angle-based technique (ABOD) designed for high-dimensional data, the density-based local outlier factor (LOF) and influenced outlierness (INFLO) methods, and antihub-based methods on various synthetic and real-world data sets, we offer novel insight into the usefulness of reverse neighbor counts in unsupervised outlier detection.

Index Terms—Outlier detection, reverse nearest neighbors, high-dimensional data, distance concentration

1 Introduction

Outlier (anomaly) detection refers to the task of identifying patterns that do not conform to established regular behavior [1]. Despite the lack of a rigid mathematical definition of outliers, their detection is a widely applied practice [2]. The interest in outliers is strong since they may constitute critical and actionable information in various domains, such as intrusion and fraud detection, and medical diagnosis.

The task of detecting outliers can be categorized as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and/or regular instances. Among these categories, unsupervised methods are more widely applied [1], because the other categories require accurate and representative labels that are often prohibitively expensive to obtain. Unsupervised methods include distance-based methods [3], [4], [5] that mainly rely on a measure of distance or similarity in order to detect outliers.

A commonly accepted opinion is that, due to the "curse of dimensionality," distance becomes meaningless [6], since distance measures concentrate, i.e., pairwise distances become indiscernible as dimensionality increases [7], [8]. The effect of distance concentration on unsupervised outlier detection was implied to be that every point in high-dimensional space becomes an almost equally good outlier [9]. This somewhat simplified view was recently challenged [10].

Our **motivation** is based on the following factors:

(1) It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in [10] the actual challenges posed by the "curse of dimensionality" differ from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space [9]. We will present further evidence which challenges this view, motivating the (re)examination of methods.

(2) Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points [11], [12],[1] but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data [14] warrant their reexamination for the outlier-detection task. In this light, we will revisit the ODIN method [11].

Our **contributions** can be summarized as follows:

(1) In Section 3 we discuss the challenges that unsupervised outlier detection faces in high-dimensional space. Despite the general impression that all points in a high-dimensional data set seem to become outliers [9], we show that unsupervised methods can detect outliers which are *more pronounced* in high dimensions, under the assumption that all (or most) data attributes are meaningful, i.e. not noisy. Our findings complement the observations from [10] by

- *M. Radovanović and M. Ivanović are with the Faculty of Sciences, University of Novi Sad, Serbia. E-mail: {radacha, mira}@dmi.uns.ac.rs*
- *A. Nanopoulos is with the Ingolstadt School of Management, University of Eichstaett-Ingolstadt, Germany. E-mail: alexandros.nanopoulos@ku.de*

1. To prevent confusion, it needs to be noted that the paper [12] incorrectly cites earlier work [13] by the second author of the present article as the source of the reverse-neighbor method.

demonstrating such behavior on data originating from a single distribution without outliers generated by a different mechanism. Also, we explain how high dimensionality causes such pronounced outlierness in comparison with low-dimensional settings.

(2) Recently, the phenomenon of *hubness* was observed [14], which affects reverse nearest-neighbor counts, i.e. k-occurrences (the number of times point $\mathbf{x}$ appears among the k nearest neighbors of all other points in the data set). Hubness is manifested with the increase of the (intrinsic) dimensionality of data, causing the distribution of k-occurrences to become skewed, also having increased variance. As a consequence, some points (*hubs*) very frequently become members of k-NN lists and, at the same time, some other points (*antihubs*) become infrequent neighbors. In Section 4 we examine the emergence of antihubs and the way it relates to outlierness of points, also considering low-dimensional settings, extending our view to the full range of neighborhood sizes, and exploring the interaction of hubness and data sparsity.

(3) Based on the relation between antihubs and outliers in high- and low-dimensional settings, in Section 5 we explore two ways of using k-occurrence information for expressing the outlierness of points, starting with the method ODIN proposed in [11]. Our main goal is to provide insight into the behavior of k-occurrence counts in different realistic scenarios (high and low dimensionality, multimodality of data), that would assist researchers and practitioners in using reverse neighbor information in a less ad-hoc fashion.

(4) Finally, in Section 6 we describe experiments with synthetic and real data sets, the results of which illustrate the impact of factors such as dimensionality, cluster density and antihubs on outlier detection, demonstrating the benefits of the methods, and the conditions in which the benefits are expected.

2 Related Work

According to the categorization in [1], the scope of our investigation is to examine: (1) point anomalies, i.e., individual points that can be considered as outliers without taking into account contextual or collective information, (2) unsupervised methods, and (3) methods that assign an "outlier score" to each point, producing as output a list of outliers ranked by their scores. The described scope of our study is the focus of most outlier-detection research [1].

Among the most widely applied methods within the described scope are approaches based on nearest neighbors, which assume that outliers appear far from their closest neighbors. Such methods rely on a distance or similarity measure to find the neighbors, with Euclidean distance being the most popular option. Variants of neighbor-based methods include defining the outlier score of a point as the distance to its kth nearest neighbor [3] (henceforth referred to as the k-NN method), or as the sum of distances to the k nearest neighbors [4]. Related to these methods are approaches that determine the score of a point according to its relative density, since the distance to the kth nearest neighbor for a given data point can be viewed as an estimate of the inverse density around it [5]. A widely-used density-based method is the local outlier factor (LOF) [15], which influenced many variations, e.g., the local correlation integral (LOCI) [16], local distance-based outlier factor (LDOF) [17], and local outlier probabilities (LoOP) [18].

The angle-based outlier detection (ABOD) [19] technique detects outliers in high-dimensional data by considering the variances of a measure over angles between the difference vectors of data objects. ABOD uses the properties of the variances to actually take advantage of high dimensionality and appears to be less sensitive to the increasing dimensionality of a data set than classic distance-based methods.

The study in [20] distinguishes three problems brought by the "curse of dimensionality" in the general context of search, indexing, and data mining applications: poor discrimination of distances caused by concentration, presence of irrelevant attributes, and presence of redundant attributes, all of which hinder the usability of traditional distance and similarity measures. The authors conclude that despite such limitations, common distance/similarity measures still form a good foundation for secondary measures, such as shared-neighbor distances, which are less sensitive to the negative effects of the curse.

Zimek et al. [10] continue the discussion of problems relevant to unsupervised outlier-detection methods in high-dimensional data by identifying seven issues in addition to distance concentration: noisy attributes, definition of reference sets, bias (comparability) of scores, interpretation and contrast of scores, exponential search space, data-snooping bias, and hubness. In this article we will focus on the aspect of hubness, and assume that all attributes carry useful information, i.e., are not overly noisy.

Finally, the notion of reverse nearest neighbors, considered important in areas outside outlier detection [21], [22], was used to formulate outlier scores in various ways. In [11], the reverse k-nearest neighbor count is defined to be the outlier score of a point in the proposed method ODIN, where a user-provided threshold parameter determines whether a point is designated as an outlier or not. Experiments were performed on low-dimensional data, and offered little insight into the reason why reverse nearest neighbors should constitute meaningful outliers. In [12], a method for detecting outliers based on reverse neighbors was briefly considered, judging that a point is an outlier if it has a zero k-occurrence count. The proposed method also does not explain the mechanism which creates points with low k-occurrences, and can be considered a special case of ODIN with the threshold set to 0. In [23], the relation between

reverse nearest neighbors and outliers was explored, but again no investigation was performed on how reverse neighbors are connected with high-dimensional phenomena, focusing instead on application to stream mining and improving the execution time of reverse nearest-neighbor computation. Also, there exists the influenced outlierness measure (INFLO) [24], based on a symmetric relationship that considers both neighbors and reverse neighbors of a point when estimating its density distribution. INFLO is essentially a density-based technique (an extension of LOF that considers the density of reverse neighbors in addition to direct neighbors when estimating the outlierness of a point), designed to work in settings of low to moderate dimensionality. The main focus of [24] was on the efficiency of computing INFLO scores. In contrast to all approaches above, we focus on high-dimensional as well as low-dimensional data and use reverse nearest neighbors only through the distribution of k-occurrences, taking into account the inherent relationship between dimensionality, neighborhood size and reverse neighbors that was not observed in previous outlier-detection work. In doing so, we will revisit the outlier scoring method used in ODIN [11].

3 Outlier Detection in High Dimensions: Improving the Perspective

In this section we revisit the commonly accepted view that in high-dimensional space unsupervised methods detect every point as an almost equally good outlier, since distances become indiscernible as dimensionality increases [9]. In [10] this view was challenged by showing that the exact *opposite* may take place: as dimensionality increases, outliers generated by a different mechanism from the data tend to be detected as more prominent by unsupervised methods, assuming all dimensions carry useful information. We present an example revealing that this can happen even when no true outliers exist, in the sense of originating from a different distribution than other points.

Example 3.1: Let us observe $n = 10{,}000$ d-dimensional points, whose components are independently drawn from the uniform distribution in range $[0, 1]$. We employ the classic k-NN method [3] ($k = 50$; similar results are obtained with other values of k). We also examine ABOD [19] (for efficiency reasons we use the FastABOD variant with $k = 0.1n$), and use standard deviation to express the variability in the assigned outlier scores.

Fig. 1(a) illustrates the standard deviations of outlier scores against dimensionality d. Let us observe the k-NN method first. For small values of d, deviation of scores is close to 0, which means that all points tend to have almost identical outlier scores. This is expected, because for low d values, points that are uniformly distributed in $[0, 1]^d$ contain no prominent outliers. This assumption also holds as d increases, i.e., still there should be no prominent outliers in the data. Nevertheless, with increasing dimensionality, for k-NN there is a clear increase of the standard deviation. This increase indicates that some points tend to have significantly smaller or larger outlier scores than others. This can be observed in the histogram of the outlier scores in Fig. 1(b), for $d = 3$ and $d = 100$. In the former case, the vast majority of points have very similar scores. The latter case, however, clearly shows the existence of points in the right tails of the distributions which are prominent outliers, as well as points on the opposite end with much smaller scores.

The ABOD method, on the other hand, exhibits a completely different trend in Fig. 1(a), with the deviation of its scores quickly diminishing as dimensionality increases, which makes it appear that the method is severely "cursed" by the dimensionality. However, ABOD was specifically designed to take advantage of high dimensionality and shown to be very effective in such settings (cf. Section 6), meaning that the shrinking variability observed in Fig. 1(a) says little about the expected performance of ABOD, which ultimately depends on the quality of the produced outlier rankings [10]. However, when scores are regularized by logarithmic inversion and linearly normalized to the $[0, 1]$ range [25], a trend similar to k-NN can be observed, shown in Fig. 1(c). □

As discussed, high dimensionality causes the emergence of some points that tend to be clearly detected as outliers by common unsupervised methods. This happens despite the fact that the existence of prominent outliers is not expected. Apparently, it is only the increase of dimensionality that caused the generation of the prominently scored outliers. This observation raises several questions: Is such behavior an artefact of the selected data distribution? Is it a property of the distance function used? Can these prominent outliers somehow be characterized?

In the example above, we chose the setting involving uniformly distributed random points because of the intuitive expectation that it should not contain any really prominent outliers. Analogous observations can be made with other data distributions, numbers of drawn points, and distance measures. The demonstrated behavior is actually an inherent consequence of increasing dimensionality of data, with the tendency of the detected prominent outliers to come from the set of *antihubs* – points that appear in very few, if any, nearest neighbor lists of other points in the data.

4 Antihubs and Outliers

In this section, we observe antihubs as a special category of points in high-dimensional spaces. We explain the reasons behind the emergence of antihubs and examine their relation to outliers detected by unsupervised methods in the context of varying neighborhood size k. Finally, we explore the interplay of hubness and data sparsity.

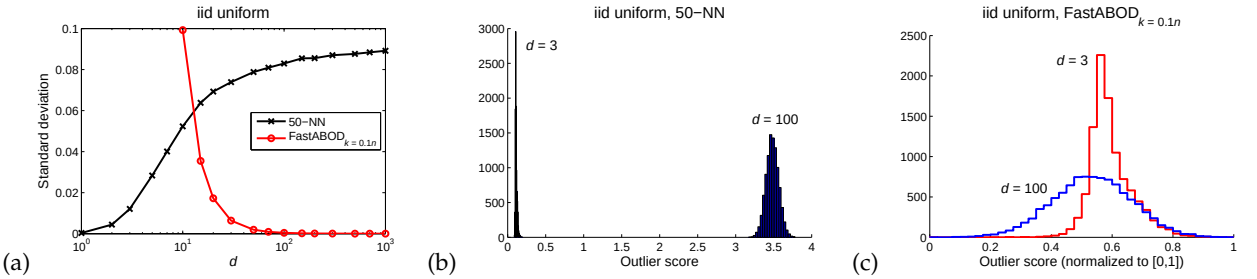

Fig. 1. Outlier scores vs. dimensionality d for uniformly distributed data in $[0,1]^d$: (a) Standard deviation; (b) Histogram of 50-NN scores; (c) Histogram of normalized ABOD scores.

4.1 Antihubs: Definition and Causes

The existence of antihubs is a direct consequence of high dimensionality when neighborhood size k is small compared to the size of the data. To understand this relationship more clearly, let us first briefly review the counterintuitive concentration behavior of distances as dimensionality increases [8].

Distance concentration refers to the tendency of distances in high-dimensional data to become almost indiscernible as dimensionality increases, and is usually expressed through a ratio of a notion of spread (e.g., standard deviation) and magnitude (e.g., the expected value) of the distribution of distances of all points in a data set to some reference point. If this ratio tends to 0 as dimensionality goes to infinity, it is said that distances concentrate. Considering random data with iid coordinates and Euclidean distance, concentration is reflected in the fact that, as dimensionality increases, the standard deviation of the distribution of distances remains constant, while the mean value continues to grow. More visually it can be said that, as dimensionality increases, all points tend to lie approximately on a hypersphere centered at the reference point, whose radius is the mean distance. It is important to note that in high-dimensional space *any point* can be used as the reference point, producing the concentration effect: the radius of the sphere (the expected distance to the reference point) increases with dimensionality, while the spread of points above and below the surface (e.g., the standard deviation of the distance distribution) becomes negligible compared to the radius.

Returning to antihubs, their emergence is an aspect of the "curse of dimensionality" related to distance concentration. This aspect will be generally referred to as *hubness* [14]. To describe hubness, let us define the notions of k-occurrences, hubs and antihubs.

Definition 1 (k-occurrences): Let $D \subset \mathbb{R}^d$ be a finite set of n points. For point $\mathbf{x} \in D$ and a given distance or similarity measure, the number of k-occurrences, denoted $N_k(\mathbf{x})$, is the number of times $\mathbf{x}$ occurs among the k nearest neighbors[2] of all other points in D. Equivalently, $N_k(\mathbf{x})$ is the reverse k-nearest neighbor count of $\mathbf{x}$ within D.

2. We use a fixed k, with ties broken randomly.

Definition 2 (hubs and antihubs): For $q \in (0,1)$, *hubs* are the $\lceil nq \rceil$ points $\mathbf{x} \in D$ with the highest values of $N_k(\mathbf{x})$. For $p \in (0,1), p < 1-q$, *antihubs* are the $\lceil np \rceil$ points $\mathbf{x} \in D$ with the lowest values of $N_k(\mathbf{x})$.[3]

Under widely applicable assumptions,[4] for $k \ll n$, as dimensionality increases the distribution of N_k becomes skewed to the right, with variance increasing, resulting in the emergence of hubs that appear in many more k-NN lists than other points, and conversely antihubs that appear in a much lower number of k-NN lists (possibly 0). These extreme cases of hubs and antihubs are the focal points of our interest.

Example 4.1: To illustrate the changes in the distribution of N_k with varying dimensionality, let us consider a random data set consisting of $n = 10{,}000$ d-dimensional points drawn uniformly from the unit hypercube $[0,1]^d$, and another random data set drawn from the standard multivariate normal distribution with iid components. Fig. 2(a–c) shows the empirically observed distributions of N_k ($k = 5$), with respect to Euclidean distance, for dimensionalities 3, 20, and 100.

For $d = 3$, the empirical distributions of N_5 (Fig. 2a) are consistent with the binomial distribution, which is expected if we view k-occurrences as node indegrees in the k-NN digraph. In low dimensions, the degree distributions of the digraphs closely follow the Erdős-Rényi (ER) random graph model, which is binomial and Poisson in the limit [26]. As dimensionality increases, the observed distributions of N_5 depart from the ER model and become more spread out and skewed to the right (Fig. 2b, c). □

We made similar observations as in Example 4.1 with various values of $k \ll n$, distance measures (l_p distances for $p \geq 1$ and $0 < p < 1$, cosine, Bray-Curtis, normalized Euclidean, Canberra), and data distributions (including Poisson, Chi-square, lognormal, exponential and Weibull). In virtually all cases, skewness exists and (anti)hubs emerge.

To provide an intuitive depiction of the mechanism through which hubs and antihubs emerge, let

3. Again, ties are broken randomly.

4. As explained in [14], the two assumptions required for the phenomenon to manifest for $k \ll n$ are centrality (i.e., the existence of meaningful centers in the data with respect to the distance measure used) and high intrinsic dimensionality.

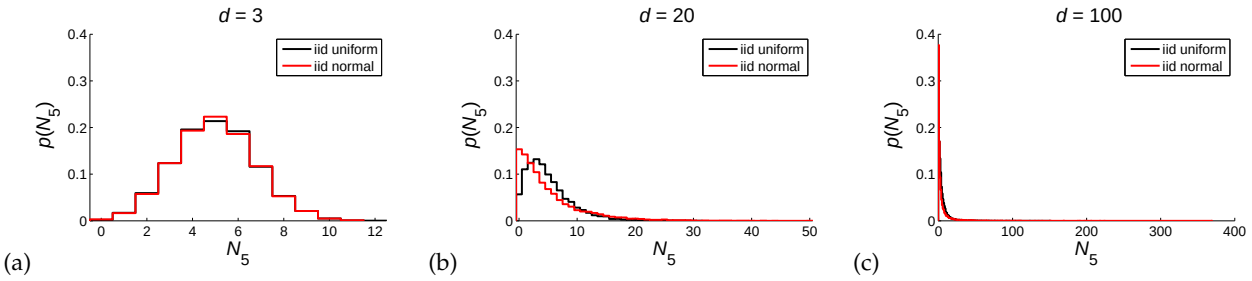

Fig. 2. Empirical distribution of N_5 for Euclidean distance on iid uniform, and iid normal random data sets with $n = 10{,}000$ points and dimensionality (a) $d = 3$, (b) $d = 20$, and (c) $d = 100$.

us return to the visual representation of distances in the data space via the hypersphere centered at some reference point. However, we deliberately select the reference point to be the data mean. The previous discussion on concentration indicates that some data points will lie very close to the hypersphere surface (e.g., less than one standard deviation away from the expected distance to the data mean), some will lie considerably below the surface (e.g., more than one standard deviation closer to the mean), and some will be situated considerably above (e.g., more than one standard deviation farther). Since the data mean is taken to be the center of the hypersphere, as dimensionality increases the points considerably below the surface tend to become closer, in relative terms, to other points in the data set, becoming hubs. Conversely, points considerably above the surface of the hypersphere become antihubs, while points near to the surface, i.e., the "regular" points, tend to have a close to expected value of N_k (which is k).

It follows that the principal mechanism which generates hubs and antihubs is *spatial centrality*: when a point is closer to the center, the distances to its neighbors become smaller (by virtue of the point being closer to all other points, assuming a unimodal data distribution, or to points from the same cluster, assuming a multimodal data distribution) [14]. Conversely, when a point is farther away from the center the distances to its nearest neighbors become larger. What is, however, remarkable, is that spatial centrality becomes amplified as dimensionality increases, producing more pronounced hubs and antihubs.

Example 4.2: To illustrate the effect of centrality on formation of hubs and antihubs, let us consider Spearman's rho and Kendall's tau-b correlations between Euclidean distance to the data center and N_5 scores, for the same iid uniform data used in the previous example. As dimensionality increases, stronger correlation emerges, increasing from -0.02 Spearman / -0.014 Kendall in the 3-dimensional case to -0.8 Spearman / -0.63 Kendall for 20 dimensions, and -0.867 Spearman / -0.715 Kendall for the 100-dimensional setting (weaker Kendall than Spearman correlations are predominantly due to ties in the integer N_5 scores). This implies that points closer to the center tend to become hubs. We made analogous observations with other values of $k \ll n$, and combinations of data distributions and distance measures for which hubness occurs. As noted in [14], proximity to one global data-set center correlates with hubness in high dimensions when the underlying data distribution is *unimodal*. For multimodal data distributions, e.g. those obtained through a mixture of unimodal distributions, hubs tend to appear close to the means of component distributions of the mixture [14]. □

The above discussion of (anti)hubness was given with the assumption that $k \ll n$. When k is large, i.e., comparable to data-set size n ($k \sim n$), the distribution of N_k is left with little "room to maneuver" for having large variance and becoming skewed: the expected value of N_k is still k (now large), but the maximal value of $n-1$ is now much easier to reach, resulting in variance remaining small relative to the mean value. More interestingly, for values of k comparable to n, the property of centrality amplification discussed before is observable in N_k scores regardless of dimensionality.

Example 4.3: Consider the same setting from Example 4.1 ($n = 10{,}000$ d-dimensional iid uniform and iid normal random points, Euclidean distance), with the only difference of using $k = 0.5 \cdot n = 5{,}000$. The observed distributions of N_k for all dimensionalities are much more bell-shaped, highly resembling binomial/Poisson distributions, and actually no points have zero N_k values. On the other hand, the Spearman correlations between distance to the center and $N_{5,000}$ scores are almost perfect for all dimensionalities, -0.999, with Kendall not far behind: -0.977 for $d = 3$ and -0.983 for $d \in \{20, 100\}$. We obtained the same trends with other data distributions. □

The strong correlations in Example 4.3 indicate that for k values comparable with n high dimensionality is not required for the creation of hubs and antihubs that correspond with point centrality. We will further discuss the differences between N_k distributions for small and large k values and their implications on unsupervised outlier detection in the next subsection.

4.2 The Relation Between Antihubs and Outliers

Outlier-detection methods can generally be categorized into global and local approaches, i.e., the de-

cision on the outlierness of some data object can be based on the complete (global) database or only on a (local) selection of data objects [27]. Naturally, there can exist a whole continuum of degrees between the two opposing extremes of "global" and "local," where the degree of locality may be tunable using parameters. For example, by raising the value of k when using the classic k-NN outlier detection method, one increases the set of data points used to determine the outlier score of the point of interest, moving from a local to a global notion of outlierness, and ending in the extreme case when $k = n - 1$. Likewise, raising k when determining reverse nearest neighbors, i.e., antihubs, raises the expected size of reverse-neighbor sets (while their size can still vary amongst points).[5]

Since antihubs have been defined as points with the lowest N_k values, we can explore the relation between N_k scores and outlierness by measuring the correlation between N_k values and outlier scores produced by unsupervised methods. For the data in Example 3.1, we measured the Kendall tau-b correlation coefficient between inverse N_k values (the lower the value, the stronger the antihub) and the outlier scores computed by the k-NN and ABOD methods (for efficiency reasons we use FastABOD [19] with $k = 0.1 \cdot n =$ 1,000). The measured correlations are plotted in Fig. 3(a) and (b), together with the correlation between inverse N_k values and the distance to the data set mean (Fig. 3c) for two values of dimensionality: low ($d = 2$) and high ($d = 100$). Furthermore, we consider two portions of points for computing correlations: all points ($p = 100\%$) and $p = 5\%$ of points with the highest distance from the data set mean as the strongest outliers. It can be seen that for the high-dimensional case correlations for $p = 100\%$ are very strong for a wide range of k values, with the exceptions being very low (close to 1) and very high values (close to $n =$ 10,000). For $p = 5\%$ agreement between N_k and k-NN/ABOD still exists, but is notably weaker. This means that N_k scores can be considered a feasible alternative to established k-NN and ABOD outlier scoring methods, since on one hand they produce very similar rankings overall, but on the other hand the rankings of the strongest outliers produced by N_k values do not completely agree with the established methods, suggesting that N_k is not redundant compared to them. The suitability of N_k for expressing outlierness is supported by the strong correlations with the "ground truth" shown in Fig. 3(c). For the low-dimensional case, the very strong correlations are also achievable, albeit for a more narrow range of k values when observing all points. Interestingly, this range widens when considering $p = 5\%$ of strongest outliers. The weak correlation between N_k and ABOD scores in the low-dimensional case can be attributed to ABOD's reliance on high dimensionality to produce meaningful scores, i.e. ABOD expectedly fails in this setting. Please note that this experiment has the intention to show the tendency of N_k to produce values that are similar to some common outlier-detection methods, and meaningful with respect to increase of dimensionality and neighborhood size k, with more general conclusions deferred for later sections.

In summary, the emergence of antihubs is closely connected with outliers both in high-dimensional and low-dimensional data. The examples above illustrate this connection, and suggest that antihubs can be used as an alternative to standard outlier-detection methods. However, from the discussion above one could deduce that antihubs simply provide a crude approximation of established outlier scoring methods for some ranges of values of parameter k. As we will see in the next section, this is not the case, since in more realistic settings involving multimodal data the correlations can behave quite differently.

4.3 Multimodality and Neighborhood Size

Real data differs from the synthetic examples from previous sections in many respects, including existence of multiple clusters in the data, and possibility that different regions where data resides have different densities. In this section we will explore the former aspect with respect to antihubs, and defer the discussion of the latter to Section 6.2.1.

In Fig. 4 we show the same correlations as in Fig. 3, for two-cluster random data with $n =$ 10,000 points. Half of the points are drawn with iid uniform components from range $[-0.5, 0.5]$, and the other half from $[0.5, 1.5]$, creating two well-separated uniformly distributed clusters. Fig. 4(a) and (b) depict correlations between N_k scores on the one hand, and k-NN and ABOD outlier scores on the other, while Fig. 4(c) shows the correlation of N_k with the "ground truth," in this case the distance of points to their respective cluster centers (not the global data-set mean).

It can be seen that for this setting N_k is not so strongly correlated with k-NN and ABOD scores as before. In Fig. 4(a) there is a visible "dent" in the correlation for $k =$ 5,000, which is the point where k reaches cluster size and for this setting produces a somewhat pathological case where all N_k scores are equal, due to the structure of the k-NN graph being reduced to two complete graphs that represent the two disjoint clusters. In Fig. 4(b), agreement between N_k and ABOD starts only for ks above 5,000 in the high-dimensional case. Fig. 4(c) suggests that for k values below 5,000 the N_k scores and k-NN distance correctly identify the outliers, while for higher k values both scoring methods "go global" and are no longer able to detect the cluster-specific local outliers.

While reiterating the "local vs. global" nature of outliers detected by different methods, the main aim

5. Within the framework from [27], the sets we discussed are referred to as *context sets*, while *reference sets* are global.

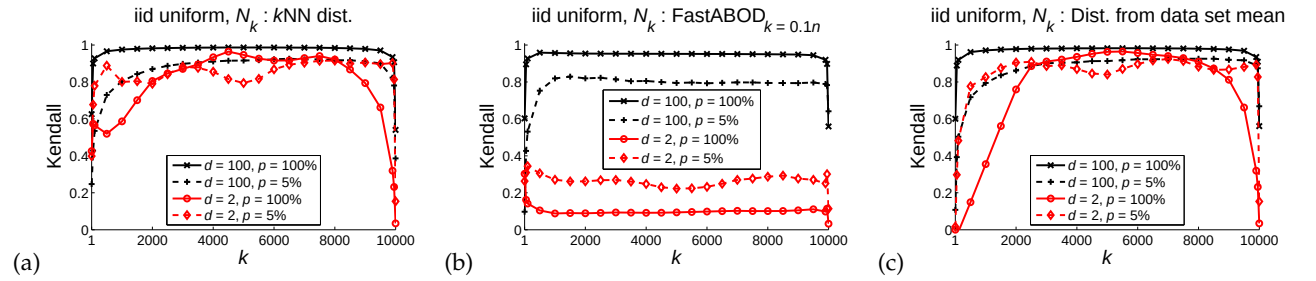

Fig. 3. Correlation between N_k values and outlier scores (a, b), and the distance from the data set mean as the "ground truth" (c), for iid uniform random data ($n = 10{,}000$ points).

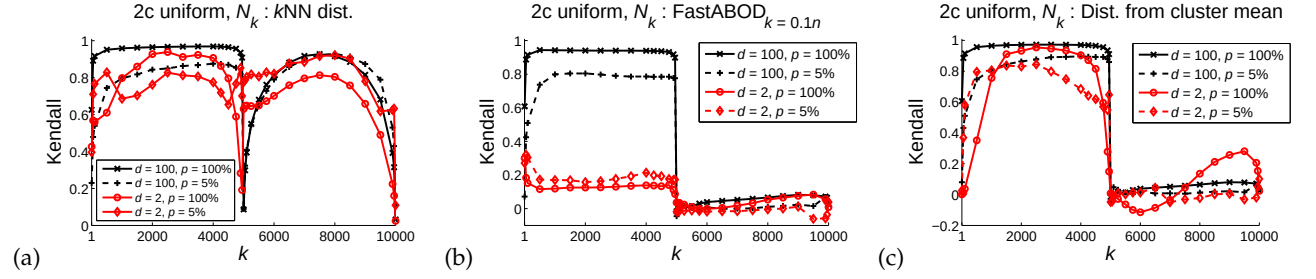

Fig. 4. Correlation between N_k values and outlier scores (a, b), and the distance from the data set mean as the "ground truth" (c), for two-cluster uniform random data ($n = 10{,}000$ points).

of this subsection is to demonstrate the existence of realistic settings where N_k values produce different outlier orderings than established outlier scoring methods, demonstrating that N_k scores are not redundant. Fig. 5 shows the correlations between N_k and k-NN/ABOD for several real data sets described in Section 6.1 (on kdd99-r2l we use FastABOD with $k = 0.1n$, on the other data sets exact ABOD), selected to illustrate the variety of behavior that can be observed, with k values given as fractions of data-set size n. It can be seen that correlation of N_k values with k-NN and ABOD scores can be weak, strong, and manifest in different combinations (e.g., for the thyroid-sick data set, correlation of N_k with k-NN distance is weak, and with ABOD is strong, while for the us-crime data set both correlations are strong).

4.4 Hubness and Sparsity

So far we have shown that there exists a relationship between N_k and outlierness of points which makes antihubs good candidates for outliers in the data. On the other hand, we also saw that for $k \ll n$ high dimensionality can induce strong hubness in the data, resulting in a large number of points with low or 0 N_k score, hindering discrimination. In the next section we will discuss discrimination in more detail, while here we will focus on the interplay between the number of data points (n) and dimensionality (d), and the question of how data sparsity affects hubness. We will first state a theoretical result by Newman et al. [28] which directly motivates our discussion.

Let $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$, $i = 0, \ldots, n$ be iid random vectors with a continuous distribution $\mathbf{F}(\mathbf{X})$, $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ of the form $\mathbf{F}(\mathbf{X}) = \prod_{i=1}^{d} G(X_i)$, i.e. the coordinates $X_1, \ldots, X_d$ are iid. Let $N_1^{n,d}$ be the number of points from $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ whose nearest neighbor with respect to Euclidean distance is $\mathbf{x}^{(0)}$.

Theorem 1: (Newman et al. [28], p. 730, Theorem 7) Suppose that G has a finite fourth moment, then

$$\lim_{n\to\infty} \lim_{d\to\infty} \mathrm{Var}(N_1^{n,d}) = \infty\,. \tag{1}$$

Theorem 2: (Newman et al. [28], p. 730, Theorem 8)

$$\lim_{d\to\infty} \lim_{n\to\infty} \mathrm{Var}(N_1^{n,d}) = 1\,. \tag{2}$$

In [14] we discussed a more general version of Theorem 1 presented in [29]. However, what is particularly interesting about the two theorems from [28] for the purposes of the present discussion is the completely different behavior of N_1 counts when limits in n and d are reversed: when d dominates n, $\mathrm{Var}(N_1^{n,d})$ diverges, while in the opposite case $\mathrm{Var}(N_1^{n,d})$ converges to 1. Although the limits can not offer definitive answers as to what happens in the finite case, they raise interesting questions of how data sparsity affects hubness: (1) Will dimensionality which dominates the number of points induce extremely strong hubness that will be hard to manage, and (2) Will increasing the number of points lead to a decrease and elimination of hubness from the data.

To explore the practical implications of Theorems 1 and 2, in the context of the two questions raised above, we ran simulations with iid uniformly distributed random data, in one case fixing the number of points n to 1,000 and raising d so it dominates n, and in the other case fixing $d = 100$ and varying n. Fig. 6

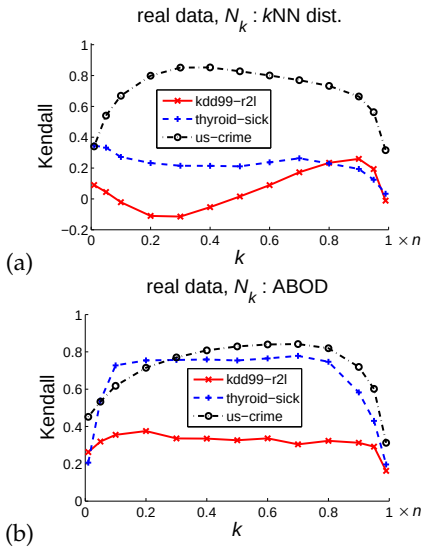

Fig. 5. Correlation between N_k values and outlier scores for a selection of real data sets.

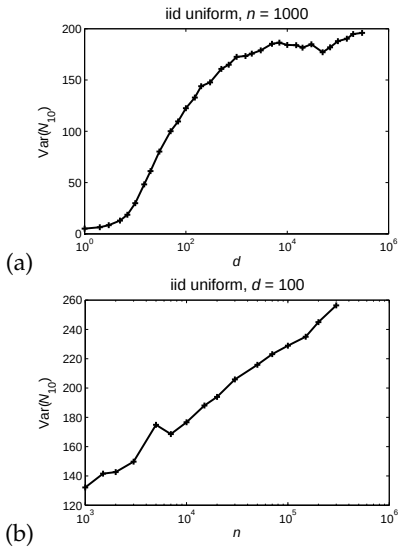

Fig. 6. Observed $\mathrm{Var}(N_{10})$ on iid uniform random data for (a) $n = 1{,}000$, varying d; (b) $d = 100$, varying n.

shows the resulting variance of N_{10}, with Fig. 6(a) reporting averages over 30 runs (we obtained analogous results with other values of $k \ll n$).

From Fig. 6(a) it can be seen that as d increases, $\mathrm{Var}(N_{10})$ initially grows slowly, then has a period of faster growth up to d of around 1,000, where the growth slows down. On the other hand, in Fig. 6(b), $\mathrm{Var}(N_{10})$ shows a steady logarithmic growth trend with respect to n.[6] These observations suggest that in the finite case the dominance of dimensionality (resulting in extreme sparsity) does not necessarily induce extreme levels of hubness, since the growth of the variance slows in Fig. 6(a). Conversely, adding more points as in Fig. 6(b) does not eliminate nor decrease hubness, but to the contrary increases it at a logarithmic rate. Of course, for some even higher values of d and n the trends may change, but overall we can say that sparsity (or lack of it) in itself is not expected to lead to extreme cases of strong hubness or absence of hubness in the data.

6. This observation is actually in slight contradiction with our earlier study that detected no correlation between n and hubness, which however examined this relationship only on a collection of real data sets [14].

Algorithm 1 AntiHub$_{dist}(D, k)$ (based on ODIN [11])

Input:
- Distance measure $dist$
- Ordered data set $D = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$, for $i \in \{1, 2, \ldots, n\}$
- No. of neighbors $k \in \{1, 2, \ldots\}$

Output:
- Vector $\mathbf{s} = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^n$, where s_i is the outlier score of $\mathbf{x}_i$, for $i \in \{1, 2, \ldots, n\}$

Temporary variables:
- $t \in \mathbb{R}$

Steps:
1) For each $i \in (1, 2, \ldots, n)$
2) $\quad t := N_k(\mathbf{x}_i)$ computed w.r.t. $dist$ and data set $D \backslash \mathbf{x}_i$
3) $\quad s_i := f(t)$, where $f : \mathbb{R} \to \mathbb{R}$ is a monotone function

5 Outlier Detection Methods Based on Antihubs

In this section we consider methods for outlier detection based on the properties of antihubs described in previous sections. Natural outlier scoring based on N_k counts was used in the ODIN method [11]. Since we do not consider threshold parameters, and apply normalization to the scores, we will reformulate this method as AntiHub, which defines the outlier score of point $\mathbf{x}$ from data set D as a function of $N_k(\mathbf{x})$, and is given in Algorithm 1.[7] We will discuss the properties of this method, and propose a refinement whose aim is to tackle some its weaknesses.

From the previous discussion and evaluation in Section 6, we summarize the properties of the AntiHub (ODIN) method by considering different factors:

(1) Hubness. High dimensionality induces very good *overall* correlation between N_k scores and "ground truth" about outlierness, even when $k \ll n$. However, presuming one is interested in detecting outliers that represent only a small proportion of the data set (the usual scenario), high dimensionality can cause problems in discriminating the scores, since the majority of candidate points will have similar low N_k values, as illustrated in Fig. 3(c) by the contrast between cases where $p = 100\%$ and $p = 5\%$ for $d = 100$.

On the other hand, in the low-dimensional case the overall correlation between N_k scores and "ground

7. The purpose of function f is to allow for changing the monotonicity of outlier scores, and confining the scores within a certain range. In our experiments, we use the function $1/(N_k(\mathbf{x}) + 1)$, which assumes that the higher the score, the more the point is considered an outlier, and maps the scores to the $(0, 1]$ range. The framework in [27] refers to this as the normalization step.

truth" is weak when $k \ll n$, but it increases significantly when considering a small portion of the data points, as can be seen in Fig. 3(c) by comparing the cases when $p = 100\%$ and $p = 5\%$ for $d = 2$. This means that in low dimensions the outlier scores produced by the AntiHub/ODIN method can still be meaningful when $k \ll n$.

(2) Locality vs. globality. AntiHub can be used both as a local and global scoring method, by adjusting the k parameter. However, as discussed above, the "local mode" when $k \ll n$ can have problems with discrimination of scores in high-dimensional settings.

(3) Discreteness of scores. The scores produced by AntiHub are inherently discrete regardless of dimensionality and hubness (since they are based on integer N_k counts), which can also hinder discrimination.

(4) Varying density. AntiHub is not sensitive to the scale of distances in the data, i.e., it can effectively detect (local) outliers in clusters of different densities without explicitly modeling density (see Section 6.2.1).

(5) Computational complexity. To use the AntiHub method in "global mode," i.e., with high values of k can be computationally expensive, since nearest-neighbor search and indexing methods typically assume that $k \ll n$ (see Section 6.3).

From the above summary it is evident that discrimination of scores represents a notable weakness of the AntiHub method, with two contributing factors: hubness (property #1) and inherent discreteness (property #3). In order to add more discrimination (primarily w.r.t property #1), one approach could be to raise k, possibly to some value comparable with n. In the next section we will explore this option, but the approach raises two concerns: (1) with increasing k the notion of outlierness moves from local to global, thus if local outliers are of interest they can be missed (property #2); (2) k values comparable with n raise issues with computational complexity (property #5).

For these reasons we propose a simple heuristic method AntiHub^2, which refines outlier scores produced by the AntiHub method by also considering the N_k scores of the neighbors of $\mathbf{x}$, in addition to $N_k(\mathbf{x})$ itself. For each point $\mathbf{x}$, AntiHub^2 proportionally adds $(1-\alpha) \cdot N_k(\mathbf{x})$ to α times the sum of N_k scores of the k nearest neighbors of $\mathbf{x}$, where $\alpha \in [0,1]$. We select summation as a simple and natural way to aggregate the neighbors' scores, while other heuristics such as averaging are also possible (used, e.g., in LOF to aggregate ratios of local reachability distances). The proportion α is automatically determined by maximizing discrimination between outlier scores of the strongest outliers, and controlled by two user-provided parameters: the ratio of strongest outliers for which to observe discrimination ($p \in (0,1]$) and step size when searching for the best α value ($step \in (0,1]$). Algorithm 2 describes the method in more detail.

To illustrate the improvement in discrimination of scores that AntiHub^2 introduces compared to AntiHub, Fig. 7 shows the values of discrimination scores computed by function discScore from Algorithm 2 (with parameter $p = 0.1$) on real data sets from Section 6.1. It is evident that AntiHub^2 offers considerable improvement, with values of $k < 100$ being sufficient for producing highly discriminated scores on all considered data sets.

Algorithm 2 AntiHub$^2_{dist}(\mathbf{x}, k, p, step)$

Input:
- Distance measure $dist$
- Ordered data set $D = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^d$, for $i \in \{1, 2, \ldots, n\}$
- No. of neighbors $k \in \{1, 2, \ldots\}$
- Ratio of outliers to maximize discrimination $p \in (0, 1]$
- Search parameter $step \in (0, 1]$

Output:
- Vector $\mathbf{s} = (s_1, s_2, \ldots, s_n) \in \mathbb{R}^n$, where s_i is the outlier score of $\mathbf{x}_i$, for $i \in \{1, 2, \ldots, n\}$

Temporary variables:
- AntiHub scores $\mathbf{a} \in \mathbb{R}^n$
- Sums of nearest neighbors' AntiHub scores $\mathbf{ann} \in \mathbb{R}^n$
- Proportion $\alpha \in [0, 1]$
- (Current) discrimination score $cdisc, disc \in \mathbb{R}$
- (Current) raw outlier scores $\mathbf{ct}, \mathbf{t} \in \mathbb{R}^n$

Local functions:
- discScore$(\mathbf{y}, p)$: for $\mathbf{y} \in \mathbb{R}^n$ and $p \in (0, 1]$ outputs the number of unique items among $\lceil np \rceil$ smallest members of $\mathbf{y}$, divided by $\lceil np \rceil$

Steps:
1) $\mathbf{a} := \text{AntiHub}_{dist}(D, k)$
2) For each $i \in (1, 2, \ldots, n)$
3) $\quad ann_i := \sum_{j \in \text{NN}_{dist}(k,i)} a_j$, where $\text{NN}_{dist}(k, i)$ is the set of indices of k nearest neighbors of $\mathbf{x}_i$
4) $disc := 0$
5) For each $\alpha \in (0, step, 2 \cdot step, \ldots, 1)$
5) $\quad$ For each $i \in (1, 2, \ldots, n)$
6) $\quad\quad ct_i := (1 - \alpha) \cdot a_i + \alpha \cdot ann_i$
7) $\quad cdisc := \text{discScore}(\mathbf{ct}, p)$
8) $\quad$ If $cdisc > disc$
9) $\quad\quad \mathbf{t} := \mathbf{ct}, disc := cdisc$
10) For each $i \in (1, 2, \ldots, n)$
11) $\quad s_i := f(t_i)$, where $f : \mathbb{R} \to \mathbb{R}$ is a monotone function

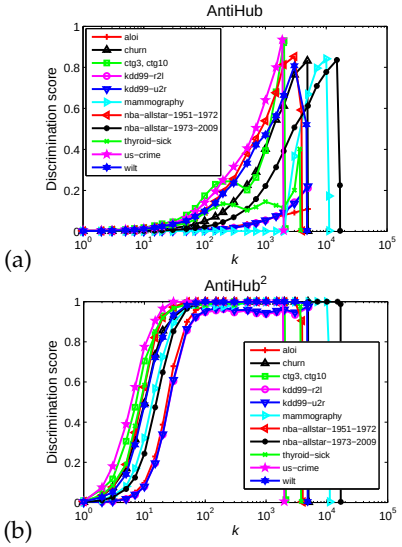

Fig. 7. Discrimination of AntiHub and AntiHub^2 scores.

6 Experimental Evaluation

The goal of our experimental evaluation is to examine the effectiveness of outlier mining methods described in the previous section. Its second objective is to examine the behavior of the methods with respect to the k parameter. The main overall aim of this section is to further support the observations that reverse-neighbor relations can be effectively applied to outlier detection in both high- and low-dimensional settings.

6.1 Experimental Procedure

6.1.1 Methods

Our experimental evaluation considers the two methods described in the previous section, denoted AntiHub_k and AntiHub_k^2, where k is the used number of nearest neighbors. We will always assume Euclidean distance. For convenience, k may be referred to as a fraction of data-set size n.

As the main baseline for comparison we examine the k-NN method [3]. For the second baseline we select ABOD [19] since it exploits the properties of high-dimensional data. We will use the original parameterless version of ABOD which observes angles between all pairs of points, and employ the version denoted FastABOD which restricts the point set to the k nearest neighbors on some larger data sets for feasibility of computation. We will also include LOF [30] as a classic representative of density-based methods. Finally, we will include the influenced outlierness method (INFLO) [24], a density-based method which also makes use of reverse nearest neighbors (through a symmetric neighborhood relationship). We employ the two-way search method [24], always using the default threshold value of $M = 1$. For LOF and INFLO we use the implementations provided by the Environment for Developing KDD-Applications Supported by Index-Structures (ELKI) [31], while for other methods we use our own implementations.

In all experiments we use areas under curve (AUC), which is a standard way to measure the effectiveness of outlier-detection methods [4], [18], [32].

6.1.2 Data Sets

We used both synthetic and real data sets. Synthetic data sets are employed because they allow for modifying crucial parameters, such as dimensionality and distribution of data. Comparison among methods is also performed with real data sets summarized in Table 1, which shows the number of points (n), dimensionality (d), skewness of the distribution of N_{10} ($S_{N_{10}}$), the percentage of points labeled as outliers, and data-set source. The associated class labels are used only for evaluation purposes. Unless otherwise stated in Table 1, we designate the minority class as outliers. All real data sets are z-score standardized.

The $S_{N_{10}}$ values are included following the approach from [14], in order to give an indication of data-set intrinsic dimensionality which affects the degree distribution of the k-NN graph (for $k \ll n$). Despite of the majority of data sets having high (embedding) dimensionality d, only us-crime appears to be intrinsically high-dimensional (with $S_{N_{10}} > 1$), with churn, ctg3, ctg10 and nba-allstar-1973-2009 being of moderate intrinsic dimensionality, and other data sets having low intrinsic dimensionality.

It should be noted that the KDD'99 network intrusion data, used to derive two data sets in Table 1, has received various criticism [41], [42]. Therefore, we use the cleaned-up version from [38] (full train set) which addresses some drawbacks such as redundant records and varying difficulty of groups. The data set is still not representative of real-world networks, however we believe its use is reasonable in our context.

Regarding NBA data [40], we used regular season player statistics which we split into two periods, 1951–1972 and 1973–2009, since two statistics (offensive and defensive rebounds) were introduced in 1973.

6.2 Experimental Results

6.2.1 Experiment with Synthetic Data

Our first experiment has the purpose to provide a demonstration of one plausible scenario where the methods based on antihubs are expected to perform well, which is in a setting involving clusters of different densities. For this reason, we use synthetic data in order to control data density and dimensionality.

We randomly generate $n = 10{,}000$ points divided into two clusters of equal size, by drawing 5,000 points from the multivariate normal distribution with mean $-\mathbf{1}$ and independent components with standard deviation 0.1, and the other 5,000 points from the normal distribution with mean $\mathbf{1}$ and component standard deviation 1. We consider dimensionalities 2 and 100. Since the result is a set of points that form two normally distributed clusters of different densities, we can characterize as outliers the points that lie in the exteriors of the two clusters, thus working with the notion of local outlierness. For each cluster, we take 5% of points with the largest distance from their cluster center, move them even farther from the center by 20% of the distance, and designate them as outliers.

The results of applying outlier detection methods are shown in Fig. 8. The chart in Fig. 8(a) shows the 2-dimensional setting, while Fig. 8(b) depicts 100-dimensional data. Both charts show the AUC of outlier detection methods as a function of parameter k, with the exception of ABOD which is used in its FastABOD variant with $k = 0.1n$. To facilitate a better view of both the local case (low k) and global case (high k), a logarithmic horizontal axis is used.

In all cases k-NN and ABOD do not perform particularly well, which is expected since they tend to assign larger outlier scores to points from the sparser cluster. AntiHub is able to achieve very good

TABLE 1
Real Data Sets Used in the Experiments

Name	n	d	$S_{N_{10}}$	Outlier%	Source and notes
aloi	50,000	64	0.260	3.016	[33], prepared by the developers of ELKI [34]
churn	5,000	17	0.849	14.140	[35], description of features available in [36]
ctg3	2,126	35	0.652	8.279	UCI [37]
ctg10	2,126	35	0.652	2.493	UCI [37]
kdd99-r2l	68,338	38	0.018	1.456	[38], R2L intrusions regarded as outliers
kdd99-u2r	67,395	38	0.031	0.077	[38], U2R intrusions regarded as outliers
mammography	11,183	6	0.103	2.325	[39]
nba-allstar-1951-1972	4,018	15	0.483	15.903	[40], regular seasons 1951–1972, all-star players regarded as outliers
nba-allstar-1973-2009	16,916	17	0.730	5.669	[40], regular seasons 1973–2009, all-star players regarded as outliers
thyroid-sick	3,772	52	0.371	6.124	[39]
us-crime	1,994	100	1.327	7.523	[39]
wilt	4,839	5	−0.075	5.394	UCI [37]

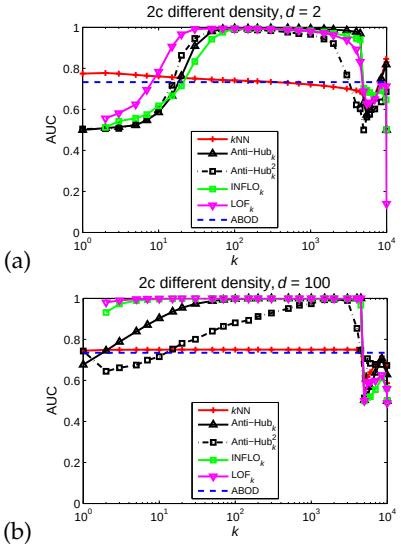

Fig. 8. Results for synthetic data with two clusters.

performance, with the only difference between the low- and high-dimensional settings being in the value of k where good performance is achieved: in low dimensions values between 100 and 1,500 are all sufficient, while in high dimensions it performs best for k ranging from 500 to just below 5,000. Density-based methods LOF and INFLO expectedly perform very well, but it is important to note that AntiHub exhibits robustness to different densities in a very simple and natural fashion, without explicitly modeling density.

6.2.2 Results on Real Data

Next, we examine the real data sets from Table 1. Our goal is to illustrate the situations where methods based on antihubs can achieve comparable or better performance than the reference methods. We do not claim overall superiority of the the proposed methods, but rather that they can contain valuable information concerning outlierness. The results are summarized in Figure 9, showing AUC for different k values given on a logarithmic scale. All methods are parameterized by k, with the exception of ABOD where we used the original exact version on the smaller data sets, and FastABOD on larger data sets with one fixed value of k for feasibility of computation ($k = \lfloor 0.1n \rfloor$ for mammography, $k = \lfloor 0.01n \rfloor$ for aloi, kdd99-r2l, kdd99-u2r, nba-allstar-1973-2009). For the same reason, AntiHub^2, LOF, INFLO are limited to k values up to several thousand on the largest data sets.

Regarding best-performing outlier-detection methods, two types of data sets can be discerned in Fig. 9: data sets with mostly local density-based outliers (aloi, thyroid-sick, wilt) where LOF, INFLO, AntiHub and AntiHub^2 perform well for small values of k, and other data sets where these methods fail with small k, but ABOD and k-NN perform well, indicating that outliers are more distance-based in nature. Generally, AntiHub and AntiHub^2 tend to follow the performance trends of LOF and INFLO, with AntiHub and/or AntiHub^2 having the edge on some data sets in terms of reaching better performance for some value of k (e.g., churn, us-crime, wilt), or performing better for more values of k (e.g., ctg3, ctg10, nba-allstar-1973-2009), and vice versa.

On data sets with predominantly distance-based outliers (where density-based methods fail for small k, e.g., KDD99, mammography, NBA data sets) k-NN and ABOD are usually the safest choice. However, it is interesting that LOF, INFLO, AntiHub and AntiHub^2 can reach and even surpass their performance for large k, suggesting there may exist a relationship between "global" density-based and distance-based outliers. It can be noted that AntiHub^2 offers improvement over AntiHub, as well as LOF and INFLO, on many such data sets (churn, ctg3, mammography, NBA data sets, thyroid-sick, us-crime). AntiHub^2 can also be worse than AntiHub, suggesting that discrimination of scores may not be the only factor to take into account for improving AntiHub.

Without relying on labels in the data for AUC evaluation, the choice of k can be based on prior domain knowledge (are local or global outliers expected), as well as agreement between various methods (e.g., on several charts in Fig. 9 similar high AUCs for k-NN and ABOD high agreement between the two methods and thus a dominance of global outliers in the data).

Finally, let us examine some interpretable output of outlier detection: the top 5 outliers from the nba-allstar-1973-2009 data set (Table 2). We used $k = 7000$,

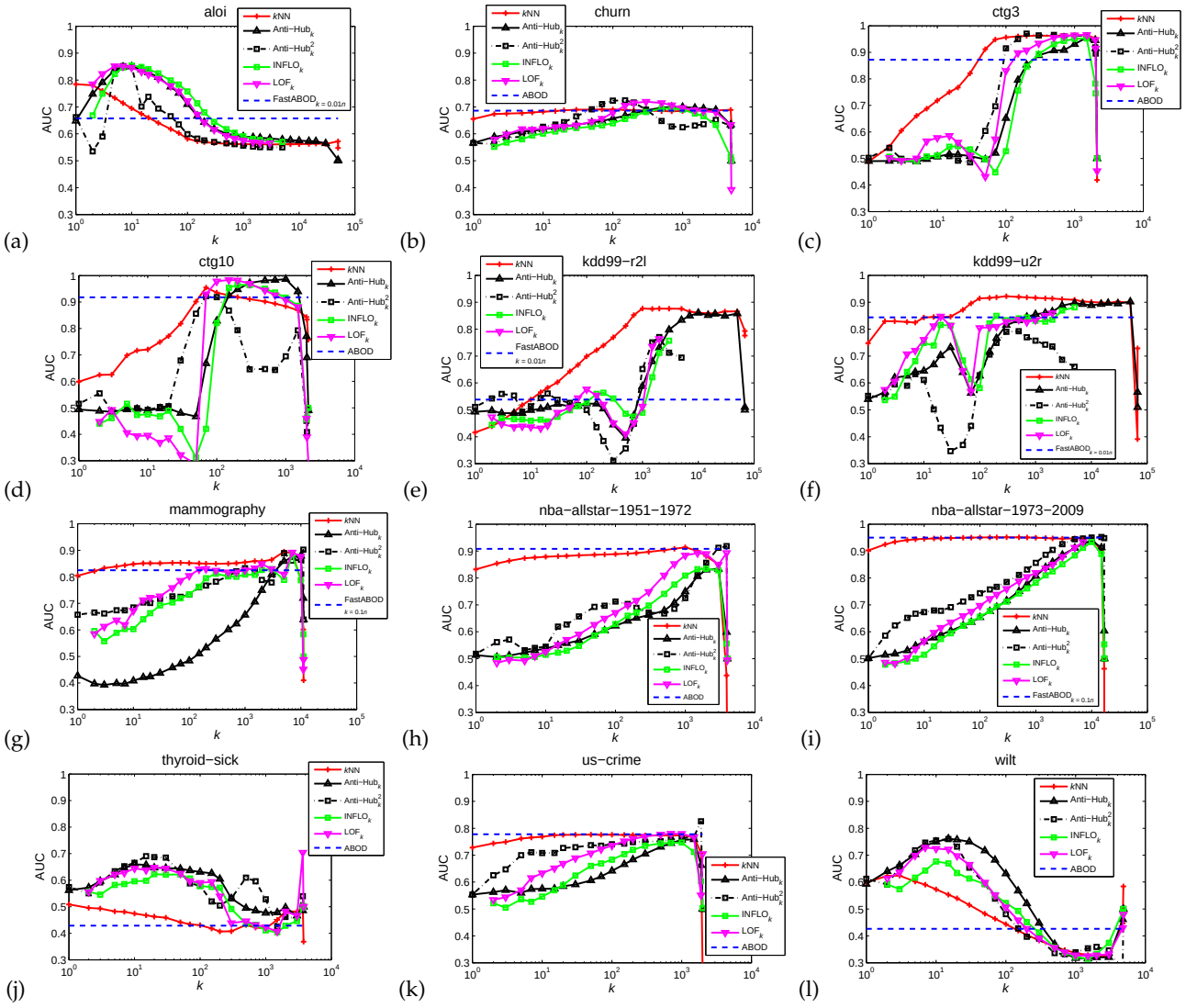

Fig. 9. Results on real data sets.

where according to AUC all methods perform comparably well. First, it needs to be noted that regardless of whether players' seasons in Table 2 (and further down the lists) were labeled all-star or not, the top-ranked seasons were all exceptional: Mark Eaton's and Manute Bol's seasons are top two in the number of blocks for the whole data set, while Antoine Walker and Dirk Nowitzki had exceptional years in 2000, comparable to other seasons when they were all-stars.

It can be seen that the top 5 of INFLO and LOF are in high agreement, which can be expected as they are both density-based methods. The list produced by Anti-Hub is very similar to the mentioned two, indicating that with this particular data set and k value Anti-Hub exhibits its density-sensitive side. The list produced by k-NN can perhaps be interpreted as the most "classic." ABOD shows a preference for older seasons, which could partly be attributed to missing data on offensive rebounds. Anti-Hub^2, could be viewed as preferring balanced seasons – being good in all stats without excelling in a particular one.

6.3 Discussion of Computational Complexity

Regarding execution time, the direct implementation of AntiHub has $O(n^2)$ time complexity, since it needs to compute all pairwise distances and calculate the N_k values by finding the k-nearest neighbors of each point (this reduces to a selection problem that can be solved in linear time by, e.g., the median of medians algorithm [43]). It directly follows that AntiHub^2 has $O(n^2 s)$ time complexity, since it requires $s = 1/step$ ($step \in (0, 1]$) additional scans of the data set to update the scores using already computed N_k counts.

Our experimental setting did not permit us to make exact empirical comparisons of running time. However, it can safely be said that k-NN is the fastest algorithm, and that AntiHub adds little overhead on top of it. AntiHub^2 runs expectedly slower, but is still at least as manageable as density-based methods LOF and INFLO. (Fast)ABOD was generally the slowest.

An approximate version of AntiHub could reduce its execution time by computing approximate neighbors, with a trade-off would be between reduction in

TABLE 2
Strongest Outliers in the nba-allstar-1973-2009 Data Set

k-NN ($k = 7000$)

Name	Year	All star?
Kareem Abdul-Jabbar	1979	yes
Artis Gilmore	1973	yes
Hakeem Olajuwon	1989	yes
Moses Malone	1978	yes
Michael Jordan	1986	yes

Anti-Hub$_{k=7000}$

Name	Year	All star?
Artis Gilmore	1973	yes
Manute Bol	1985	no
Mark Eaton	1984	no
Kareem Abdul-Jabbar	1975	yes
Artis Gilmore	1974	yes

Anti-Hub$^2_{k=7000}$

Name	Year	All star?
Shawn Marion	2002	yes
Shawn Marion	2004	yes
Shawn Marion	2005	yes
Antoine Walker	2000	no
Dirk Nowitzki	2000	no

INFLO$_{k=7000}$

Name	Year	All star?
Mark Eaton	1984	no
Manute Bol	1985	no
Kareem Abdul-Jabbar	1975	yes
Artis Gilmore	1973	yes
Hakeem Olajuwon	1989	yes

LOF$_{k=7000}$

Name	Year	All star?
Mark Eaton	1984	no
Manute Bol	1985	no
Artis Gilmore	1973	yes
Kareem Abdul-Jabbar	1975	yes
Hakeem Olajuwon	1989	yes

FastABOD$_{k=0.1n}$

Name	Year	All star?
Artis Gilmore	1973	yes
Artis Gilmore	1974	yes
Kareem Abdul-Jabbar	1975	yes
George McGinnis	1974	yes
Artis Gilmore	1974	yes

execution time and effectiveness. Since we demonstrated that AntiHub (and k-NN as well) can reach peak performance for values of $k \sim n$, the development of an approximate version of AntiHub (and k-NN) for such k values represents a significant challenge since, to the best of our knowledge, current approximate k-NN algorithms assume small constant k.

7 Conclusions

In this paper, we provided a unifying view of the role of reverse nearest neighbor counts in problems concerning unsupervised outlier detection, focusing on the effects of high dimensionality on unsupervised outlier-detection methods and the hubness phenomenon, extending the previous examinations of (anti)hubness to large values of k, and exploring the relationship between hubness and data sparsity.

Based on the analysis, we formulated the AntiHub method for unsupervised outlier detection, discussed its properties, and proposed a derived method which improves discrimination between scores. Our main hope that this article clarifies the picture of the interplay between the types of outliers and properties of data, filling a gap in understanding which may have so far hindered the widespread use of reverse-neighbor methods in unsupervised outlier detection.

The existence of hubs and antihubs in high-dimensional data is relevant to machine-learning techniques from various families: supervised, semi-supervised, as well as unsupervised [14], [44], [45]. In this paper we focused on unsupervised methods, but in future work it would be interesting to examine supervised and semi-supervised methods as well. Another relevant topic is the development of approximate versions of AntiHub methods that may sacrifice accuracy to improve execution speed. An interesting line of research could focus on relationships between different notions of intrinsic dimensionality, distance concentration, (anti)hubness, and their impact on subspace methods for outlier detection. Finally, secondary measures of distance/similarity, such as shared-neighbor distances [20] warrant further exploration in the outlier-detection context.

Acknowledgments. M. R. and M. I. thank the Ministry of Education, Science and Technological Development of the Republic of Serbia for support through project no. OI174023, "Intelligent techniques and their integration into wide-spectrum decision support."

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput Surv*, vol. 41, no. 3, p. 15, 2009.

[2] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., 1987.

[3] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *SIGMOD Rec*, vol. 29, no. 2, pp. 427–438, 2000.

[4] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," in *Proc Conf on Applications of Data Mining in Computer Security*, 2002, pp. 78–100.

[5] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *VLDB J*, vol. 8, no. 3–4, pp. 237–253, 2000.

[6] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *Proc 7th Int Conf on Database Theory (ICDT)*, 1999, pp. 217–235.

[7] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in *Proc 8th Int Conf on Database Theory (ICDT)*, 2001, pp. 420–434.

[8] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE T Knowl Data En*, vol. 19, no. 7, pp. 873–886, 2007.

[9] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *Proc 27th ACM SIGMOD Int Conf on Management of Data*, 2001, pp. 37–46.

[10] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

[11] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proc 17th Int Conf on Pattern Recognition (ICPR)*, vol. 3, 2004, pp. 430–433.

[12] J. Lin, D. Etter, and D. DeBarr, "Exact and approximate reverse nearest neighbor search for multimedia data," in *Proc 8th SIAM Int Conf on Data Mining (SDM)*, 2008, pp. 656–667.

[13] A. Nanopoulos, Y. Theodoridis, and Y. Manolopoulos, "C^{2}P: Clustering based on closest pairs," in *Proc 27th Int Conf on Very Large Data Bases (VLDB)*, 2001, pp. 331–340.

[14] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *J Mach Learn Res*, vol. 11, pp. 2487–2531, 2010.

[15] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *SIGMOD Rec*, vol. 29, no. 2, pp. 93–104, 2000.
[16] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *Proc 19th IEEE Int Conf on Data Engineering (ICDE)*, 2003, pp. 315–326.
[17] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in *Proc 13th Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD)*, 2009, pp. 813–822.
[18] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: Local outlier probabilities," in *Proc 18th ACM Conf on Information and Knowledge Management (CIKM)*, 2009, pp. 1649–1652.
[19] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proc 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD)*, 2008, pp. 444–452.
[20] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?" in *Proc 22nd Int Conf on Scientific and Statistical Database Management (SSDBM)*, 2010, pp. 482–500.
[21] A. Singh, H. Ferhatosmanoğlu, and A. Şaman Tosun, "High dimensional reverse nearest neighbor queries," in *Proc 12th ACM Conf on Information and Knowledge Management (CIKM)*, 2003, pp. 91–98.
[22] Y. Tao, M. L. Yiu, and N. Mamoulis, "Reverse nearest neighbor search in metric spaces," *IEEE T Knowl Data En*, vol. 18, no. 9, pp. 1239–1252, 2006.
[23] C. Lijun, L. Xiyin, Z. Tiejun, Z. Zhongping, and L. Aiyong, "A data stream outlier delection algorithm based on reverse k nearest neighbors," in *Proc 3rd Int Symposium on Computational Intelligence and Design (ISCID)*, 2010, pp. 236–239.
[24] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proc 10th Pacific-Asia Conf on Advances in Knowledge Discovery and Data Mining (PAKDD)*, 2006, pp. 577–593.
[25] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores," in *Proc 11th SIAM Int Conf on Data Mining (SDM)*, 2011, pp. 13–24.
[26] P. Erdős and A. Rényi, "On random graphs," *Publ Math-Debrecen*, vol. 6, pp. 290–297, 1959.
[27] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Min Knowl Disc*, vol. 28, no. 1, pp. 190–237, 2014.
[28] C. M. Newman, Y. Rinott, and A. Tversky, "Nearest neighbors and Voronoi regions in certain point processes," *Adv Appl Probab*, vol. 15, no. 4, pp. 726–751, 1983.
[29] C. M. Newman and Y. Rinott, "Nearest neighbors and Voronoi volumes in high-dimensional point processes with various distance functions," *Adv Appl Probab*, vol. 17, no. 4, pp. 794–809, 1985.
[30] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proc ACM Int Conf on Management of Data (SIGMOD)*, 2000, pp. 93–104.
[31] E. Achtert, S. Goldhofer, H.-P. Kriegel, E. Schubert, and A. Zimek, "Evaluation of clusterings – metrics and visual support," in *Proc 28th Int Conf on Data Engineering (ICDE)*, 2012, pp. 1285–1288.
[32] E. Müller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *Proc 27th IEEE Int Conf on Data Engineering (ICDE)*, 2011, pp. 434–445.
[33] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, "The Amsterdam library of object images," *Int J Comput Vision*, vol. 61, no. 1, pp. 103–112, 2005.
[34] "DataSets/MultiView – ELKI," 2014. [Online]. Available: http://elki.dbs.ifi.lmu.de/wiki/DataSets/MultiView
[35] "SGI – MLC++: Datasets from UCI," 2014. [Online]. Available: http://www.sgi.com/tech/mlc/db/
[36] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, 2005.
[37] K. Bache and M. Lichman, "UCI machine learning repository," 2014. [Online]. Available: http://archive.ics.uci.edu/ml
[38] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc 2nd IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 2009, pp. 1–6.
[39] Z. Ding, "Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics," Ph.D. dissertation, Computer Science Dissertations. Paper 60. http://scholarworks.gsu.edu/cs_diss/60, 2011.
[40] "databaseBasketball.com Stats," 2014. [Online]. Available: http://www.databasebasketball.com/stats_download.htm
[41] J. McHugh, "Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," *ACM T Inform Syst Se*, vol. 3, no. 4, pp. 262–294, 2000.
[42] T. Brugger, "KDD Cup '99 dataset (Network Intrusion) considered harmful," 2007. [Online]. Available: http://www.kdnuggets.com/news/2007/n18/4i.html
[43] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. MIT Press, 2009.
[44] N. Tomašev and D. Mladenić, "Nearest neighbor voting in high dimensional data: Learning from past occurrences," *Comput Sci Inf Syst*, vol. 9, no. 2, pp. 691–712, 2012.
[45] N. Tomašev, M. Radovanović, D. Mladenić, and M. Ivanović, "The role of hubness in clustering high-dimensional data," *IEEE T Knowl Data En*, vol. 26, no. 3, pp. 739–751, 2014.

Miloš Radovanović is an Assistant Professor at the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Serbia, where he received his BSc, MSc and PhD degrees. From 2009 he is Managing Editor of the *Computer Science and Information Systems* journal. He (co)authored one programming textbook, a research monograph, and over 40 papers in data mining and related fields.

Alexandros Nanopoulos is an Assistant Professor at the University of Eichstaett-Ingolstadt, Germany. His main research interests include Data Mining and Machine Learning with applications in Databases and Information Retrieval. Dr. Nanopoulos obtained his BSc and PhD from the Department of Informatics of Aristotle University of Thessaloniki, Greece, where he taught as Lecturer from 2004 to 2008. From 2008 to 2012 he was an Assistant Professor at University of Hildesheim, Germany. Dr. Nanopoulos is the co-author of more than 140 articles in international journals and conference proceedings. He has co-authored the monograph "R-trees: Theory and Applications" and is serving as program committee member of several international conferences on data mining and databases.

Mirjana Ivanović holds the position of Full Professor at Faculty of Sciences, University of Novi Sad. She is a member of the University Council for Informatics. She is author or co-author of 13 textbooks and more than 290 research papers on multi-agent systems, e-learning, intelligent techniques (CBR, data and web mining), most of which are published in international journals and conferences. She is/was a PC member of more than 150 international conferences, leader of numerous international research projects and is Editor-in-Chief of the *Computer Science and Information Systems* journal.