Outlier Detection and Trend Detection: Two Sides of the Same Coin

Preprint — publisher version available at http://dx.doi.org/10.1109/ICDMW.2015.79

Erich Schubert, Michael Weiler, Arthur Zimek
Ludwig-Maximilians-Universität München
Oettingenstr. 67, 80538 München, Germany
http://www.dbs.ifi.lmu.de
{schube,weiler,zimek}@dbs.ifi.lmu.de

***Abstract*—Outlier detection is commonly defined as the process of finding unusual, rare observations in a large data set, without prior knowledge of which objects to look for. Trend detection is the task of finding some unexpected change in some quantity, such as the occurrence of certain topics in a textual data stream. Many established outlier detection methods are designed to search for low-density objects in a static data set of vectors in Euclidean space. For trend detection, high volume events are of interest and the data set is constantly changing. These two problems appear to be very different at first. However, they also have obvious similarities. For example, trends and outliers likewise are supposed to be rare occurrences. In this paper, we discuss the close relationship of these tasks. We call to action to investigate this further, to carry over insights, ideas, and algorithms from one domain to the other.**

***Keywords*-outlier detection; trend detection; stream outlier; event detection; textual outliers**

I. Introduction

Outliers are hard to define mathematically because we cannot expect them to follow a model or distribution known beforehand. Instead, most attempts at defining outliers focus on them being a rare observation, markedly different from the remainder of the data, such as, e.g., the well-known definition by Barnett and Lewis [15] of outliers as being "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". Yet, the notions of "inconsistence" and "remainder" remain vague. Various algorithms have been proposed that try to detect outliers in a way consistent with our intuition. Notable outlier detection algorithms include DB-Outlier [33], which reports the objects of lowest density as outliers and local outlier factor (LOF) [16] which uses a local neighborhood as "remainder" of the data set, instead of comparing to the complete data set every time.

Schubert et al. [66] propose a generalized framework for outlier detection that is applicable beyond the domain of vector spaces and show the applicability to graph and video data. In this article, we want to expand this generalized model to also cover what is known as *trend detection* and *emerging topic detection* in text mining. Here, the task is to detect changes in the distribution of a data stream that indicate the beginning of an event. The words used are slightly different, e.g. trend, emerging topic, bursty keyword. But essentially they refer to unusual, extreme topics: outliers in text streams.

The remainder of this paper is organized as follows. In Section II, we survey traditional outlier detection methods as well as adaptations to streaming data. In Section III, we point out why the research questions guiding the design of outlier detection methods might be misleading in some cases and why traditional outlier detection might also benefit from insights in trend detection. In Section IV, we discuss some methods for trend detection and their relationship to a generalized view of traditional outlier detection. Finally, in Section V, we summarize and identify challenges for future research.

II. Traditional Outlier Detection

A. Outlier Detection in Euclidean Space

Knorr and Ng [33] proposed a distance-based notion of outliers. This model is motivated by the intuition of statistical parametric approaches. It is aiming, however, not on a refinement of the statistical modelling of outliers but at designing efficient database-oriented approaches. This algorithm triggered the data mining community to develop many different approaches that have a less statistically oriented but a more spatially oriented notion to model outliers. The k-NN-outlier model [60] ranks the objects according to their distances to their k-th nearest neighbor. As a variant, the k-NN-weight model [12] uses the sum of distances to all objects within the set of k nearest neighbors (called the weight) as an outlier degree. While these models actually only use distances, the intuition is typically discussed with Euclidean data space in mind. In these distance-based approaches, for each object, a property (outlier model) [66] is learned based on a local neighborhood (radius ε, k nearest neighbors). However, the objects are eventually ranked according to this property ("outlier score") in a global way. For example, the object with the largest k-NN distance overall would be the most prominent outlier. Thus, these methods are best suited to identify *global* outliers. Recent global approaches base the decision not on Euclidean distances but on angle-variance (such as ABOD [40] and an efficient variant using random projections [57]), an intuition that is clearly also connected to a Euclidean data space.

Identifying *local* outliers (i.e., comparing local models with a local reference set [66]) started with the method LOF (local outlier factor) [16]. The basic idea is to assign a local density estimate (local reachability density, *lrd*) to each object of the database. Then, LOF considers ratios between the *lrd* of an object and the *lrd*s of its neighboring objects. Thus, the resulting outlier score is based on a local comparison rather than on a global comparison. Again, this notion of outlierness is a natural intuition for the Euclidean data space. Several extensions and refinements of the basic LOF model have been proposed, e.g. a connectivity-based outlier factor (COF) [74], or using the concept of micro-clusters to efficiently mine the top-n density-based local outliers in large databases (i.e., those n objects having the highest LOF value) [30]. A similar algorithm, named INFLO [29], for an extension of the LOF model is using also the reverse nearest neighbors additionally to the nearest neighbors and considering a symmetric relationship between both values as a measure of outlierness. The local distance-based outlier detection (LDOF) approach [81] merges the notion of local outlierness with the distance-based notion of outliers. LoOP [36] uses a density estimation based on the distance distribution of all nearest neighbors and formulates the local outlier score as a probability. COP [39] aims at detecting outliers in the presence of local correlations in the data set by measuring the deviation from the local model.

More or less explicitly, all these methods basically aim at providing rather simple approximations of statistical density estimates around data points in Euclidean space. Consequently, a recent evaluation study [17], discussing several of these methods, also focuses on numeric data.

B. Specialized Outlier Detection

Some approaches designed for high-dimensional data try to account for a local feature relevance and search outliers in subspaces of the data space [22], [31], [37], [50]–[54], [56], see the survey of Zimek et al. [85]. In the area of spatial data mining [61], the topic of spatial outliers has triggered several specialized methods [13], [20], [21], [34], [46], [47], [68], [71]. These approaches discern between spatial attributes (relevant for defining a neighborhood) and other attributes (usually only one additional attribute) where outliers deviate considerably from the corresponding attribute value of their spatial neighbors. How to derive spatial neighborhood and how to define "considerable deviation", however, differs from approach to approach. Other specialized approaches tackle for example outliers in time series [28], [73], outliers in graphs (e.g., in social networks or in DBLP) [5]–[7], [24], outlying trajectories [44], outliers in categorical or ordinal data [8], [80], or in uncertain data [4]. Aggarwal [3] provides more examples. Again, although they are not always operating in Euclidean space, all these methods aim eventually at some approximate descriptors of outlierness for the objects that ultimately should relate to statistical density estimates.

C. Outlier Detection in Data Streams

Recently, Sadik and Gruenwald [62] gave an overview on research issues for outlier detection in data streams. They follow the categorization of the well-known survey on outlier detection by Chandola et al. [19], where type II outliers, as opposed to type I outliers, are outliers with respect to some context, such as time or location. In database research, the concept of outliers w.r.t. some particular context is also known as a "local" outlier [16], [66] and is not restricted to special data types, although the concept of "locality" might be of paramount interest in such data types with special requirements for the outlier model [65], [66]. Consequently, Sadik and Gruenwald [62] are interested in streaming data and time series data without distinction, while we see time series as a special data type but consider here the scenario of streaming data as a more general scenario, that has been tackled in many studies with Euclidean data space in mind [11], [14], [26], [35], [41], [59], [70], [77]. There are two main categories of tackling dynamic data. First, the dynamic aspect of the data is tackled using an incremental approach, i.e., old data remain available while new data are coming in and the preliminary models are refined over time. The second possibility is to truly address the aspect of potential infinity of data, i.e., the fact that the complete data stream might not fit into the available memory or might actually never be completely available. In this case, the typical approach is a sliding "time window" that is oblivious of old data. The adapted approach therefore builds a model based only on the data within the time frame of the window.

D. Generalization of Outlier Detection

In a certain sense, the combination of different outlier detectors into an ensemble [25], [38], [43], [45], [55], [64], [82]–[84] can be seen as a generalization because, under certain conditions [38], [64], it becomes meaningful to combine even different methods that follow different intuitions about outlierness. But still, such combinations can only combine the available methods, that have been typically designed for Euclidean data space or for some particular use case. There is a recent line of reasoning, though, on truly generalizing the classic, abstract outlier methods to new use cases and data scenarios. Schubert et al. [66] modularized many existing outlier detection methods, demonstrating that there is a large conceptual overlap in these methods. Based on this modularized structure, they demonstrate how to modify existing methods to work on other data types such as geostatistical data, video streams, and graph data.

III. Limitations of Traditional Outlier Detection

Most outlier detection methods were designed with the intuition of low-density outliers in mind. In the following examples, we want to discuss some scenarios where data do not adhere to the intuition of low-density, and traditional outlier detection methods then do not work reliably.

A. Example: KDD Cup '99

On the popular KDD Cup '99 data set, one may argue that outliers are not at all rare instances. Depending on the exact version of this data set, $80\% - 94\%$ of the instances are attacks. As such, the legitimate connections may be considered the anomalies here. While this data set has been repeatedly used for evaluating outlier detection methods [1], [38], [42], [43], [45], [55], [64], [77], [78], the results of such analyses should be taken with a grain of salt.[1] The data set has many ($\approx 75 - 78\%$ [75]) duplicates and many established methods are not prepared for handling too many duplicates. Thus, unless the outlier methods are carefully implemented and parametrized, outlier scores may become undefined, and evaluation may be biased. For example, methods such as isolation forests [45] that work on random samples of the data set might then appear to perform better because they are less susceptible to the problem of duplicates. Because the data set contains categorical attributes, binary attributes, and integer valued attributes (including e.g., `num_compromised`), it is highly sensitive to preprocessing such as feature selection and data normalization. Furthermore, any intuition of density and distance, based on Euclidean space, is probably inappropriate for this data set.

B. Example: United States Census Data

In spatial outlier detection [2], observations consist of two kinds of data: a geographical location—which may be a point (a position) or an area (a polygon)—as well as a univariate or multivariate measurement. The US census data, for example, include statistics such as household size and population demographics at different spatial resolution such as census counting districts and county level. For some districts, sparsity of population causes artifacts: census counting districts include areas such as airports, graveyards, and ghost towns with a low population. Popular attributes such as relative ethnicity may be undefined for uninhabited districts, or show unusually extreme values for tiny populations. Therefore, popular outlier detection algorithms such as LOF cannot be meaningfully used on such data without modifications [66]. However, the methods can be easily generalized in a way to use a spatial context to determine the neighborhood and the non-spatial attributes for analysis and yield results competitive to those of existing geostatistics [66]. To make full use of this data set, the methods should be further customized to take uncertainty into account, in particularly those arising from a small population, which makes numbers such as ethnicity averages incomparable.

C. Example: Traffic Accidents

Schubert et al. [65] analyzed the density of traffic accidents in the UK, based on open government data. Again, results obtained by traditional outlier detection methods are not helpful: they will report accidents in low populated areas such as northern Scotland as low-density outliers. The data set contains 19% duplicated coordinates, probably due to measurement precision and reoccurring accident sites. Information available may include involvement of pedestrians, visibility conditions, severity, casualties, road numbers, authority IDs etc. that may also be missing or estimated. For their analysis, Schubert et al. only used the coordinates, and customized their approach for this data set by searching for areas with higher traffic accident density than expected, in order to find accident hotspots.

D. Observations

Above examples demonstrate how we might have been *asking the wrong question* in outlier detection research. By working primarily with data consisting of vectors in Euclidean space, and the intuition of low-density outliers, we designed our algorithms for this particular use case. This can be seen as a kind of "overfitting" at algorithm design time. (Similar observations have been made in other fields of data mining, e.g., clustering [23] and pattern mining [87].)

On the other hand, we have also been using our existing outlier detection tools the wrong way. Both on KDD Cup '99 and on the traffic accidents example our entities of interest are *not* the individual data samples. Instead, our potential outliers are some aggregation of the data: we need to aggregate the KDD Cup '99 data to *hosts* instead of processing individual connections if we want to detect attackers; Schubert et al. [65] implicitly used local maxima in density as aggregations of traffic accidents (which yields black spot crossroads, not "outlier car accidents"). On the U.S. Census data, the data were already aggregated by the Census Bureau. When analyzing trends in text, we are interested in *topics, not messages*. This also holds for first story detection (FSD), where the output is not the topic, but the first message of each topic. This is related to what has been termed "type III outliers" [19] with the distinction that we claim that the particular type of aggregation is typically not a priori *given* with the data but is a matter of an adequate *interpretation* of the given plain data. One may argue that this can be solved by improving feature extraction or preprocessing (e.g. converting textual data to numerical vectors, aggregated at the desired level) before analyzing with an outlier detection method and should be done in the preprocessing phase: convert your textual data to appropriate numerical vectors, aggregate and transform the numeric data to the desired level (i.e. topics, instead of messages), and then analyze the aggregated and transformed data with some (Euclidean space) outlier detection method. In practice, as we will see next, this does not work that easily: on data streams, preprocessing and transformation cannot be completely decoupled, and if we want to be able to explain the resulting outliers, we need to be able to return to the original data representation. Last but not least, because of efficiency considerations, it may be necessary to integrate outlier/trend detection much earlier in the analysis process, instead of first transforming all data objects.

IV. Outliers in Textual Streams

A. Some Methods for Trend Detection

Guzman and Poblete [27] discuss burst detection as anomaly detection in Twitter streams. Their objective is to detect keywords that occur in the stream at an unusually high rate, similar to TwitterMonitor [48]. Both approaches use a rather simple model of burstiness: TwitterMonitor [48] seems to track high frequency terms, whereas Guzman and Poblete [27] use

[1] See the discussion by McHugh [49] and by Tavallaee et al. [75] and kdnuggets n18 2007, "KDD Cup '99 dataset (Network Intrusion) considered harmful": http://www.kdnuggets.com/news/2007/n18/4i.html

the second derivative of the term frequency. enBlogue [10] detects trends that increase most over their moving average. The UMass system [9] generates a vector space representation (such as the term-frequency) for each incoming document. A nearest-neighbor search is then applied for each document-vector to identify outliers with a sufficiently large distance to the most similar one. Kleinberg [32] uses an infinite-state automaton to keep track of term frequencies of incoming documents (emails in particular). Frequency bursts are modeled as transitions within the automaton. The hierarchical structure of email-document topics is preserved by the state transitions. In Blogosphere [58], Kleinberg's approach was used to discover bursty terms with their correlations for extracted titles of blog posts. Bursty terms are outliers, that show a radical frequency increase within a short period of time compared to a long period. A distance metric based on Euclidean space was then used to discover potential correlations between terms. Another use of Kleinberg's algorithm was adapted by Takahashi et al. [72] to model the frequencies of topics determined by a dynamic topic model (DTM). Further work to determine trending terms are made using biology-inspired processes to model life cycles of terms [18] or wavelet-based analysis using signal processing measures [76]. A recent advancement, Signi-Trend [63], uses a heavy-hitters style algorithm to track both the average frequency and the moving standard deviation of any popular term combination. Closely related to trend detection is the problem of first story detection often referred to as 'detection of *events*'. Yang et al. [79] applied document clustering in their on-line event detection system. Events, which can be seen as outliers of the document stream, are identified by obtaining significant vocabulary shifts and rapid changes within the term frequency distribution. To distinguish events further, they also take the time gap between them into account, as discussed topics with large temporal differences are usually related to different events (e.g., different earthquakes).

By looking at all these studies, a common important component of trend detection is the definition of the properties (like frequency or density) with their expected *normal* range to determine outliers within the document space.

B. On the Relationship between Trends and Outliers

Where trend detection is interested in outliers in textual streams, these outliers are not individual instances—messages, news items, tweets—but rather *topics*. If we attempt to run traditional outlier detection methods on such a stream, we will get plenty of uninteresting outliers due to misspellings and rare words. By a classic notion of outlierness, such instances will most correctly appear as outliers. However, in most cases we probably do not want our algorithms to degenerate to counting the number of rare words per message and find the text with the most unusual vocabulary in the corpus.

Most trend detection methods—e.g. TwitterMonitor [48], Burst Detection [27], enBlogue [10], and SigniTrend [63]—perform some kind of aggregation. Often, a sliding window approach is used for aggregation of individual instances. With this technique, each such window corresponds to a point in time, and the aggregation yields one such time series for every word (or n-gram) in the data set. Outliers are those time series that show an unusual change in activity in the current time window. When interpreting these example algorithms for trend detection on the time series of a single term, we can roughly summarize their model as follows: TwitterMonitor uses the increase in term activity compared to the previous time window, Burst Detection uses the second derivative (the increase in the current time window, compared to the previous increase), enBlogue uses the relative increase in frequency over a moving average. SigniTrend uses the most complex model, consisting of exponentially weighted moving average and standard deviation, and it thus can also capture variance. All of these are fairly simple statistical models and in itself not spectacular. Much of the challenge in trend detection comes from scale: these values must be tracked and analyzed for every word (or n-gram) in the data set simultaneously, for millions of words. The main contributions of above articles are on scalability, not the statistical models used: enBlogue tracks only those word pairs where at least one word is considered a seed tag. SigniTrend uses a hashing-based approach, which is lossy on rare terms but accurate with high probability on frequent terms and uses a constant amount of memory.

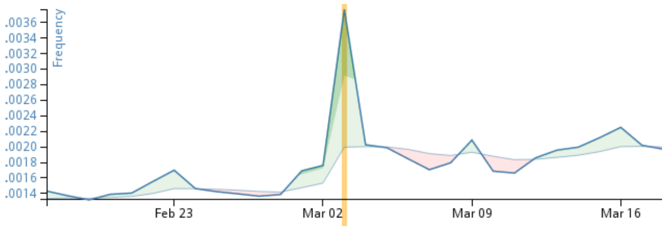

Fig. 1: Model used by SigniTrend: frequency (thick line), mean (thin line) and standard deviations (shaded areas) for the keyword "Selfie" on Twitter. The yellow bar highlights the outlier event. Screenshot from http://signi-trend.appspot.com/

In the general framework of Schubert et al. [66], all these methods can be seen as first-order outlier detection methods on time series. The earlier frequencies of a term are used to construct a model (frequency, moving average, moving average with standard deviation) and the new frequency is compared to the previous value. Figure 1 visualizes the model used by SigniTrend: from the observed frequency a moving average as well as a moving standard deviation are computed. The famous "Oscar Selfie" achieved 11 standard deviations over the previous average. We can also see that, after this selfie, the average volume substantially increased (note that the axis does not start at 0), but the model adapted quickly to this higher average volume. If we consider the standard deviation as used by SigniTrend to be a derived value from the mean, then this method can be seen as second-order. We have not yet, however, observed "locality" [66] in the sense of LOF [16]: each time series object is evaluated on its own, the significance is not compared to other time series. This calls for future work, as this will eventually allow to detect trends in smaller communities, that otherwise are masked by globally popular trends. On the other hand, several related terms (e.g. *boston*, *marathon*, *explosion*) may trend together, and the comparison of the series may help identifying the most explaining term combination. Judging the significance of a trend based on the scores of related terms can thus be expected to yield better results. This way, lessons learned for traditional outlier detection, reconsidering the notion of "locality" [66], could be transferred and boost research progress also in trend or event detection.

V. Conclusion

This article is intended as a call to action. Outlier detection research has been very much focused on Euclidean space, and the community has become detached from the actual data problems we want to solve. This manifests itself in a lack of good evaluation data, and in often incremental variations of the general theme [17]. *Data diversity* is a leading theme of data science: much of our data at hand cannot be squeezed into the rigid structure of a finite $\mathbb{R}^d$ vector space. If we want to obtain meaningful *descriptions* of outliers, we first need to work on meaningful *data*. Outlier detection, while meant to be an *unsupervised* task, is not—and cannot be—entirely free from assumptions on the characteristics of outliers. Every existing method embodies some implicit concept of outlierness. And while the methods do not use dedicated training data, they are still "trained" by the intuition of the method designer of what constitutes normal or abnormal. We suggest to formalize this notion, make it explicit, and design methods that allow customized notions of outlierness. Eventually, this will also lead to better explanation and description of outliers.

In this article, we focused on the peculiarities of textual data, which often comes streaming and in a high volume. We interpret existing trend detection methods as simple outlier detectors over time series. Many of the abstract problems become visible in this context, such as the difference between instances (e.g., messages) and outliers (i.e., trends or events) in this domain. Reflecting traditional outlier detection, we observe a similar pattern in density-based outlier detection: as much as we are seeking outlier instances, we are also seeking regions of low density. In summary, we suggest the following *research issues*:

- Improved statistical models need to be developed and used, to obtain meaningful conclusions and provide robustness against spam.
- Low latency is required for practical use for trend detection, which collides with the desire to use larger time windows to obtain more reliable statistics. This may be resolved by using an appropriate combination of models [82].
- Aggregation of results is important, e.g., merging overlapping trending topics, to avoid overloading the user with noisy and redundant results.[2]
- Scalability to a large number of instances, to a large number of aggregations, and to a fast data stream is required. This will usually require the use of approximation and indexing techniques, and will limit the complexity of models usable [67].
- Artifacts and domain-specific anomalies are omnipresent in real data, and it should be possible to customize and modify methods to handle these. On text data, stop words and spam constitute such artifacts. On census data areas with few inhabitants cause such anomalies. In traffic data, accidents attributed to the nearest milepost may mislead an algorithm.
- Bridge the gap! Outlier models should be developed that are applicable to Euclidean vector spaces, time series, text, and other data types in the same way. This will make it easier for one research domain to benefit from advances in the other.

References

[1] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA*, 2006, pp. 504–509.

[2] E. Achtert, A. Hettab, H.-P. Kriegel, E. Schubert, and A. Zimek, "Spatial outlier detection: Data, algorithms, visualizations," in *Proceedings of the 12th International Symposium on Spatial and Temporal Databases (SSTD), Minneapolis, MN*, 2011, pp. 512–516.

[3] C. C. Aggarwal, *Outlier Analysis*. Springer, 2013.

[4] C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in *Proceedings of the 8th SIAM International Conference on Data Mining (SDM), Atlanta, GA*, 2008, pp. 483–493.

[5] L. Akoglu and C. Faloutsos, "Anomaly, event, and fraud detection in large network datasets," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), Rome, Italy*, 2013, pp. 773–774.

[6] L. Akoglu, M. McGlohon, and C. Faloutsos, "OddBall: spotting anomalies in weighted graphs," in *Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Hyderabad, India*, 2010, pp. 410–421.

[7] L. Akoglu, H. Tong, and D. Koutra, "Graph-based anomaly detection and description: A survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[8] L. Akoglu, H. Tong, J. Vreeken, and C. Faloutsos, "Fast and reliable anomaly detection in categorical data," in *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM), Maui, HI*, 2012, pp. 415–424.

[9] J. Allan, V. Lavrenko, D. Malin, and R. Swan, "Detections, bounds, and timelines: UMass and TDT-3," in *Proceedings of Topic Detection and Tracking (TDT–3)*, 2000, pp. 167–174.

[10] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum, "See what's enBlogue: real-time emergent topic identification in social media," in *Proceedings of the 15th International Conference on Extending Database Technology (EDBT), Berlin, Germany*, 2012, pp. 336–347.

[11] F. Angiulli and F. Fassetti, "Distance-based outlier queries in data streams: the novel task and algorithms," *Data Mining and Knowledge Discovery*, vol. 20, no. 2, pp. 290–324, 2010.

[12] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Helsinki, Finland*, 2002, pp. 15–26.

[13] L. Anselin, "Local indicators of spatial association–LISA," *Geographical Analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[2] In many areas of data mining, redundancy of data mining results has been observed for early approaches and has been addressed later on in more mature approaches, well-known examples being subspace clustering [69] and frequent pattern mining [86].

[14] I. Assent, P. Kranen, C. Baldauf, and T. Seidl, "AnyOut: Anytime outlier detection on streaming data." in *Proceedings of the 17th International Conference on Database Systems for Advanced Applications (DASFAA), Busan, South Korea*, 2012, pp. 228–242.

[15] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. John Wiley&Sons, 1994.

[16] M. M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, 2000, pp. 93–104.

[17] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, 2015.

[18] M. Cataldi, L. D. Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proceedings of the 10th International Workshop on Multimedia Data Mining (MDM/KDD), Las Vegas, NV*, 2010, p. 4.

[19] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. Article 15, 1–58, 2009.

[20] S. Chawla and P. Sun, "SLOM: A new measure for local spatial outliers," *Knowledge and Information Systems (KAIS)*, vol. 9, no. 4, pp. 412–429, 2006.

[21] F. Chen, C.-T. Lu, and A. P. Boedihardjo, "GLS-SOD: A generalized local statistical approach for spatial outlier detection," in *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC*, 2010, pp. 1069–1078.

[22] X. H. Dang, I. Assent, R. T. Ng, A. Zimek, and E. Schubert, "Discriminative features for identifying and interpreting outliers," in *Proceedings of the 30th International Conference on Data Engineering (ICDE), Chicago, IL*, 2014, pp. 88–99.

[23] I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek, "On using class-labels in evaluation of clusterings," in *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC*, 2010.

[24] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han, "On community outliers and their efficient detection in information networks," in *Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC*, 2010, pp. 813–822.

[25] J. Gao and P.-N. Tan, "Converting output scores from outlier detection algorithms into probability estimates," in *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), Hong Kong, China*, 2006, pp. 212–221.

[26] D. Georgiadis, M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos, "Continuous outlier detection in data streams: an extensible framework and state-of-the-art algorithms," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), New York City, NY*, 2013, pp. 1061–1064.

[27] J. Guzman and B. Poblete, "On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model," in *Workshop on Outlier Detection and Description, held in conjunction with the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA*, 2013, pp. 31–39.

[28] H. V. Jagadish, N. Koudas, and S. Muthukrishnan, "Mining deviants in a time series database," in *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB), Edinburgh, Scotland*, 1999, pp. 102–113.

[29] W. Jin, A. K. H. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore*, 2006, pp. 577–593.

[30] W. Jin, A. Tung, and J. Han, "Mining top-n local outliers in large databases," in *Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA*, 2001, pp. 293–298.

[31] F. Keller, E. Müller, and K. Böhm, "HiCS: high contrast subspaces for density-based outlier ranking," in *Proceedings of the 28th International Conference on Data Engineering (ICDE), Washington, DC*, 2012, pp. 1037–1048.

[32] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, 2003.

[33] E. M. Knorr and R. T. Ng, "A unified notion of outliers: Properties and computation," in *Proceedings of the 3rd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach, CA*, 1997, pp. 219–222.

[34] Y. Kou, C.-T. Lu, and D. Chen, "Spatial weighted outlier detection," in *Proceedings of the 6th SIAM International Conference on Data Mining (SDM), Bethesda, MD*, 2006.

[35] P. Kranen, H. Kremer, T. Jansen, T. Seidl, A. Bifet, G. Holmes, B. Pfahringer, and J. Read, "Stream data mining using the MOA framework," in *Proceedings of the 17th International Conference on Database Systems for Advanced Applications (DASFAA), Busan, South Korea*, 2012, pp. 309–313.

[36] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: local outlier probabilities," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China*, 2009, pp. 1649–1652.

[37] ——, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, 2009, pp. 831–838.

[38] ——, "Interpreting and unifying outlier scores," in *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*, 2011, pp. 13–24.

[39] ——, "Outlier detection in arbitrarily oriented subspaces," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, 2012, pp. 379–388.

[40] H.-P. Kriegel, M. Schubert, and A. Zimek, "Angle-based outlier detection in high-dimensional data," in *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV*, 2008, pp. 444–452.

[41] L. I. Kuncheva, "Change detection in streaming multivariate data using likelihood detectors," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1175–1180, 2013.

[42] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM), San Francisco, CA*, 2003.

[43] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL*, 2005, pp. 157–166.

[44] J.-G. Lee, J. Han, and X. Li, "Trajectory outlier detection: A partition-and-detect framework," in *Proceedings of the 24th International Conference on Data Engineering (ICDE), Cancun, Mexico*, 2008, pp. 140–149.

[45] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 1, pp. 3:1–39, 2012.

[46] X. Liu, C.-T. Lu, and F. Chen, "Spatial outlier detection: Random walk based approaches," in *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), San Jose, CA*, 2010, pp. 370–379.

[47] C.-T. Lu, D. Chen, and Y. Kou, "Algorithms for spatial outlier detection," in *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM), Melbourne, FL*, 2003, pp. 597–600.

[48] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the Twitter stream," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Indianapolis, IN*, 2010, pp. 1155–1158.

[49] J. McHugh, "Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory," *ACM Transactions on Information and System Security*, vol. 3, no. 4, pp. 262–294, 2000.

[50] B. Micenková, R. T. Ng, X. H. Dang, and I. Assent, "Explaining outliers by subspace separability," in *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM), Dallas, TX*, 2013, pp. 518–527.

[51] E. Müller, I. Assent, P. Iglesias, Y. Mülle, and K. Böhm, "Outlier ranking via subspace analysis in multiple views of the data," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), Brussels, Belgium*, 2012, pp. 529–538.

[52] E. Müller, I. Assent, U. Steinhausen, and T. Seidl, "OutRank: ranking outliers in high dimensional data," in *Proceedings of the 24th International Conference on Data Engineering (ICDE) Workshop on Ranking in Databases (DBRank), Cancun, Mexico*, 2008, pp. 600–603.

[53] E. Müller, M. Schiffer, and T. Seidl, "Adaptive outlierness for subspace outlier ranking," in *Proceedings of the 19th ACM Conference on In-*

formation and Knowledge Management (CIKM), Toronto, ON, Canada, 2010, pp. 1629–1632.

[54] ——, "Statistical selection of relevant subspace projections for outlier ranking," in *Proceedings of the 27th International Conference on Data Engineering (ICDE), Hannover, Germany*, 2011, pp. 434–445.

[55] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan, "Mining outliers with ensemble of heterogeneous detectors on random subspaces," in *Proceedings of the 15th International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan*, 2010, pp. 368–383.

[56] H. V. Nguyen, V. Gopalkrishnan, and I. Assent, "An unbiased distance-based outlier detection approach for high-dimensional data," in *Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA), Hong Kong, China*, 2011, pp. 138–152.

[57] N. Pham and R. Pagh, "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data," in *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Beijing, China*, 2012, pp. 877–885.

[58] M. Platakis, D. Kotsakos, and D. Gunopulos, "Searching for events in the blogosphere," in *Proceedings of the 18th International Conference on World Wide Web (WWW), Madrid, Spain*, 2009, pp. 1225–1226.

[59] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Honolulu, HI*, 2007, pp. 504–515.

[60] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, TX*, 2000, pp. 427–438.

[61] J. F. Roddick and M. Spiliopoulou, "A bibliography of temporal, spatial and spatio-temporal data mining research," *ACM SIGKDD Explorations*, vol. 1, no. 1, pp. 34–38, 1999.

[62] M. S. Sadik and L. Gruenwald, "Research issues in outlier detection for data streams," *ACM SIGKDD Explorations*, vol. 15, no. 1, pp. 33–40, 2013.

[63] E. Schubert, M. Weiler, and H.-P. Kriegel, "SigniTrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds," in *Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), New York, NY*, 2014, pp. 871–880.

[64] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel, "On evaluation of outlier rankings and outlier scores," in *Proceedings of the 12th SIAM International Conference on Data Mining (SDM), Anaheim, CA*, 2012, pp. 1047–1058.

[65] E. Schubert, A. Zimek, and H.-P. Kriegel, "Generalized outlier detection with flexible kernel density estimates," in *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, 2014, pp. 542–550.

[66] ——, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 190–237, 2014.

[67] ——, "Fast and scalable outlier detection with approximate nearest neighbor ensembles," in *Proceedings of the 20th International Conference on Database Systems for Advanced Applications (DASFAA), Hanoi, Vietnam*, 2015, pp. 19–36.

[68] S. Shekhar, C.-T. Lu, and P. Zhang, "A unified approach to detecting spatial outliers," *GeoInformatica*, vol. 7, no. 2, pp. 139–166, 2003.

[69] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong, "A survey on enhanced subspace clustering," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 332–397, 2013.

[70] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB), Seoul, Korea*, 2006, pp. 187–198.

[71] P. Sun and S. Chawla, "On local spatial outliers," in *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), Brighton, UK*, 2004, pp. 209–216.

[72] Y. Takahashi, T. Utsuro, M. Yoshioka, N. Kando, T. Fukuhara, H. Nakagawa, and Y. Kiyota, "Applying a burst model to detect bursty topics in a topic model," in *Advances in Natural Language Processing – Proceedings of the 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October*, 2012, pp. 239–249.

[73] J. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 482–492, 2006.

[74] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan*, 2002, pp. 535–548.

[75] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*, 2009.

[76] J. Weng and B.-S. Lee, "Event detection in Twitter," in *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain*, 2011.

[77] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixture with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, pp. 275–300, 2004.

[78] J. Yang, N. Zhong, Y. Yao, and J. Wang, "Local peculiarity factor and its application in outlier detection," in *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV*, 2008, pp. 776–784.

[79] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 32nd International Conference on Research and Development in Information Retrieval (SIGIR), Boston, MA*, 1998, pp. 28–36.

[80] J. X. Yu, W. Qian, H. Lu, and A. Zhou, "Finding centric local outliers in categorical/numerical spaces," *Knowledge and Information Systems (KAIS)*, vol. 9, no. 3, pp. 309–338, 2006.

[81] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand*, 2009, pp. 813–822.

[82] A. Zimek, R. J. G. B. Campello, and J. Sander, "Ensembles for unsupervised outlier detection: Challenges and research questions," *ACM SIGKDD Explorations*, vol. 15, no. 1, pp. 11–22, 2013.

[83] ——, "Data perturbation for outlier detection ensembles," in *Proceedings of the 26th International Conference on Scientific and Statistical Database Management (SSDBM), Aalborg, Denmark*, 2014, pp. 13:1–12.

[84] A. Zimek, M. Gaudet, R. J. G. B. Campello, and J. Sander, "Subsampling for efficient and effective unsupervised outlier detection ensembles," in *Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Chicago, IL*, 2013, pp. 428–436.

[85] 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.

[86] A. Zimek and J. Vreeken, "The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives," *Machine Learning*, vol. 98, no. 1–2, pp. 121–155, 2015.

[87] A. Zimmermann, "The data problem in data mining," *ACM SIGKDD Explorations*, vol. 16, no. 2, pp. 38–45, 2014.