Ranking Outlier Nodes in Subspaces of Attributed Graphs

Emmanuel Müller [•○] Patricia Iglesias Sánchez [•] Yvonne Mülle [•] Klemens Böhm [•]

[•] *Karlsruhe Institute of Technology (KIT), Germany*
{emmanuel.mueller, patricia.iglesias, klemens.boehm}@kit.edu
yvonne.muelle@student.kit.edu

[○] *University of Antwerp, Belgium*
emmanuel.mueller@ua.ac.be

Abstract— **Outlier analysis is an important data mining task that aims to detect unexpected, rare, and suspicious objects. Outlier ranking enables enhanced outlier exploration, which assists the user-driven outlier analysis. It overcomes the binary detection of outliers vs. regular objects, which is not adequate for many applications. Traditional outlier ranking techniques focus on either vector data or on graph structures. However, many of today's databases store both, multi dimensional numeric information and relations between objects in attributed graphs. An open challenge is how outlier ranking should cope with these different data types in a unified fashion.**

In this work, we propose a first approach for outlier ranking in subspaces of attributed graphs. We rank graph nodes according to their degree of deviation in both graph and attribute properties. We describe novel challenges induced by this combination of data types and propose subspace analysis as important method for outlier ranking on attributed graphs. Subspace clustering provides a selected subset of nodes and its relevant attributes in which deviation of nodes can be observed. Our graph outlier ranking (GOutRank) introduces scoring functions based on these selected subgraphs and subspaces.

In addition to this technical contribution, we provide an attributed graph extracted from the Amazon marketplace. It includes a ground truth of real outliers labeled in a user experiment. In order to enable sustainable and comparable research results, we publish this database on our website[1] as benchmark for future publications. Our experiments on this graph demonstrate the potential and the capabilities of outlier ranking in subspaces of attributed graphs.

I. Introduction

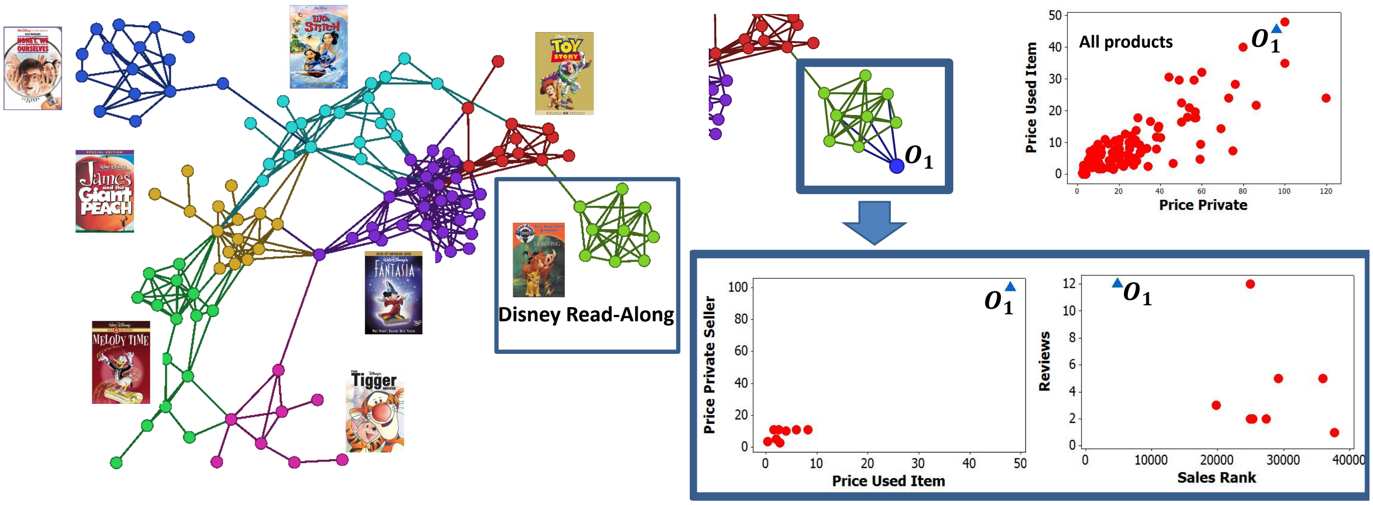
Outlier analysis is an important data mining task for fraud detection, network intrusion analysis, anomaly detection in e-commerce, and many more. In these applications one looks for highly deviating objects that show-up in contrast to the regular objects. Outlier ranking techniques score each object based on its degree of deviation. Hence, they overcome traditional outlier detection techniques [1], [2], which rely on a binary decision boundary and a difficult parametrization for this boundary. Outlier rankings enable a user-driven exploration of outliers by looking at the most promising objects first. They allow users to choose the decision boundary between outliers and regular objects in a flexible way.

In the past, outlier ranking techniques have focused on homogeneous vector data [3] or graph data [4]. However, in many of today's applications, information of both types is available. For instance, heterogeneous data can be found on e-commerce marketplaces such as Amazon. Their product databases store a large number of attributes for each product, e.g., prices, different rating ratios, product reviews. In addition, co-purchased products are stored as a graph structure. In this scenario, exceptional objects correspond to outstanding, fake, suspicious, or overpriced products. Not all of these outliers can be detected by a traditional outlier analysis restricted to attribute values or to graph structures only. For example, overpriced products might appear quite regular if one looks at the overall price distribution of the database. However, if one combines both price and co-purchases one might reveal its high deviation in price w.r.t. to this local subgroup of co-purchased products.

Our main hypothesis is that such complex outliers can only be detected by a combination of all available information. To this end, outlier mining techniques for heterogeneous databases have to be developed. They have to cope with information on relations between products, but also with a large number of attributes. Out of this large set of heterogeneous data, outlier ranking techniques have to automatically detect relevant data: (1) *subgraphs* as the relevant graph context of an outlier and (2) *subspaces* as the relevant attribute set in which an outlier is deviating. This is required as complex outliers deviate from their local context. For the attribute space, deviation might not be visible if one considers irrelevant attributes, e.g., randomly distributed attributes. Exceptional object deviation is also not recognized if one considers all given attributes simultaneously. This is due to the curse of dimensionality [5], as more and more attributes hinder the detection of outliers. Overall, outlier ranking has to measure the deviation of objects w.r.t. a subgroup of the data objects and a subset of the attributes. In this work, we consider outlier ranking based on this idea to tackle open challenges in

[1] http://www.ipd.kit.edu/~muellere/GOutRank/

978-1-4673-5304-5/13/$31.00 © 2013 IEEE



(a) Subgraph of Disney DVD products

(b) Different subspaces of the given attributes

Fig. 1. An outlier example in a subgraph of the Amazon co-purchased network

heterogeneous databases. We rank outlier nodes that are highly deviating from their local context in the attributed graph.

Let us illustrate this in a real world example. Figure 1(a) shows a part of the Amazon co-purchase network. In particular, we have selected *Disney DVD* products, which have been reviewed by a group of high school students in a user experiment at our university. Note that we also use this database (with the given ground truth of outliers) for our evaluation in Section IV. Product O_1 is one of the outliers, showing up due to its high price in attributes "price offered by private sellers" and "price for used products". This object shows high deviation w.r.t. these prices compared to its co-purchased products.

However, traditional outlier mining techniques can not detect this deviation. If we consider the graph structure only, the product is densely connected to other products. Based on the graph structure it seems to be regular. If we take only the attributes into account (cf. product prices in Fig. 1(b)), we observe many objects with high prices for new articles offered by private sellers and high prices for used articles. This seems to be quite regular over all products. Thus, graph structure or attributes alone can not reveal the deviation of object O_1. Nevertheless, O_1 is highly deviating in the densely connected group of *Disney Read-Along* products. All products of this subgraph have highly similar attribute values w.r.t. both prices, except for O_1. Note that this is only the case for this subspace. Other subsets of the attributes (e.g., Sales Rank and Reviews) form a very sparse subspace and do not indicate any high deviation of O_1. Overall, one can claim O_1 to be a true outlier w.r.t. to the *Disney Read-Along* products and the price attributes.

With this work we focus on the detection of such outliers that deviate w.r.t. a subgraph of highly connected nodes. The individual outlier shows high similarity to these nodes in the graph structure, but there exists a selection of attributes in which it deviates. We call this selection of attributes a relevant subspace. For the automatic selection of subgraphs and subspaces we rely on recent subspace analysis and graph clustering techniques. However, it is still unclear how to score individual objects based on these clusters.

In the following sections we will highlight the open issues with existing approaches (Section II), present the main challenges (Section III-B), define three scoring functions (Section III-C), present the Amazon co-purchase network as benchmark data, and demonstrate the potential and the capabilities of our approach on this real world database (Section IV). Finally we conclude and discuss several open questions for future work in this area (Section V and Section VI).

II. Related Work

We review existing approaches according to their data types. We discuss outlier mining in (1) vector data, (2) graph structures, and (3) combinations of both. We will highlight the emerging developments of outlier mining in combined vector data and graph structure, and derive the open challenges not yet addressed in literature.

Outlier Mining in Vector Data

Traditional outlier mining has focused on vector data [3]. Well known approaches have proposed outlier rankings based on density scores [6], [7]. These scores quantify the deviation of each object w.r.t. the local neighborhood of the object in the vector space. While traditional techniques are not able to detect objects that are outliers in a subset of the given attributes, recent development focuses on individual projections for each object [8], [9], [10], [11], [12], [13], [14]. They rank objects based on the selection of relevant subsets of the attributes, and tackle the curse of dimensionality for outlier ranking. However, they focus on vector data only and do not address relations between objects given in graph databases. We base our work on *OutRank* [10], [14] and extend its idea to both graph and vector data.

Anomalous Nodes in Graph Structures

In this work, we focus on outlier nodes and we do not consider anomalous edges, irregular subgraphs, and other suspicious structural anomalies [15], [16], [17]. A recent technique [18] uses the node neighborhood and its power-law characteristics to compute the outlierness score of each node. Furthermore, graph clustering algorithms [19] detect outliers as a byproduct of clustering. They detect sets of highly connected nodes as clusters and output the residual set of sparsely connected nodes as outliers. All these approaches succeed in the detection of outlier nodes based on the graph structure. However, they ignore additional information at each node such as numeric feature vectors. All of these graph mining techniques miss objects which deviate w.r.t. these node attributes.

Mining Graphs with Node Attributes

An emerging research field considers both graph and vector data. A first variant of graph clustering combines node attributes and graph structure to obtain better clustering results [20]. Its basic idea is to convert attribute values into graph nodes. However, it is limited to discrete attribute values, and it does not consider the selection of relevant attributes. More recent techniques have focused on graph clustering w.r.t. a selection of node attributes [21], [22], [23]. They address the selection of relevant attributes on the graph cluster level. In this work, we exploit the potential of these methods, which have not been designed for outlier analysis, for our graph outlier ranking. Regarding outlier mining, the most related algorithm [24] aims to detect outlier nodes that deviate from communities (e.g., in social networks). It combines information from the graph structure and the full dimensional space of the node attributes. Thus, this approach is hindered by the curse of dimensionality. In addition, it does not focus on the ranking of objects according to their degree of deviation.

III. GOutRank

Our graph outlier ranking method (*GOutRank*) aims to detect anomalous nodes in attributed graphs. It generalizes our previous outlier ranking method *OutRank* [10], [14], which has focused on high dimensional vector data without considering graph structures. Both techniques share the idea of computing a subspace clustering as pre-processing to the outlier ranking. As a general framework they can use any subspace cluster instantiation and improve with this emerging research area [25], [26], [27]. In addition, *GOutRank* exploits the hidden potential of graph clustering and its symbiosis with subspace analysis. *GOutRank* has been designed for complex outliers, which deviate only w.r.t. a local subgraph and a subset of relevant attributes. Thus, it tackles the challenges with outliers hidden in attributed graphs. *GOutRank*, detects outliers that can not be detected by traditional techniques. Our analysis on the Amazon co-purchase network will demonstrate this enhancement. Furthermore, we highlight future research potential for more enhanced scoring functions. Overall, *GOutRank* is the first solution to outlier ranking in subspaces of attributed graphs. However, it is only a first step with several open challenges for future research.

A. Basic Notions

The general aim of *GOutRank* is to provide a sorting of all the objects for the following database definition:

Definition 1: **Attributed Graph Database**
The database consists of a graph structure (V, E) and attribute information A. It is characterized as follows:
(1) Each object o is a graph vertex $o \in V$ and connected by edges $(o, p) \in E$ to other nodes $p \in V$ in the graph structure. The edges are undirected and unweighted.
(2) Each object o is described by a vector $(o_1, ..., o_d) \in \mathbb{R}^d$ in a d-dimensional continuous data space where the attributes are named $A = \{A_1, \ldots, A_d\}$.

An outlier ranking is a sorted list of all $o \in V$, in ascending order of a scoring function:

$$score(o) : V \to \mathbb{R}$$

The score represents a measure for the objects' regularity, and it considers both graph structure and attribute values. Outliers have low scores, and regular objects have high scores.

As a pre-processing step we build upon subspace clustering results obtained by different graph clustering algorithms. We abstract from their individual properties and assume that a clustering result is given as follows:

Definition 2: **Subspace Clustering Result**
A subspace clustering result in an attributed graph is a set of subspace clusters $Res = \{(C_1, S_1) \ldots (C_n, S_n)\}$, where $C_i \subset V$ is a densely connected subgraph with high attribute similarity in the subspace $S_i \subset A$.

Please note, that according to this definition, an object can be part of multiple clusters in several subspaces. This hinders the traditional detection of outliers (not assigned to any cluster), as typically each object occurs in at least one subspace cluster [25]. However, it provides us the means for a more enhanced outlier scoring, which evaluates cluster assignment of each object as an indication for its regularity.

B. Outlier Detection in Attributed Graphs

Outlier ranking in attributed graphs induces two main challenges: the selection of subgraphs with their individual subspaces and the scoring of objects in these multiple subspaces. In the following, we discuss the main challenges before presenting the *GOutRank* solution in detail.

Challenge 1:

Selection of subgraphs and subspaces

We deem the selection of subgraphs and subspaces the main challenge for outlier ranking in attributed graphs. In graph data, densely connected subgraphs stand for clusters with high intra-cluster similarity. Many relations between these clustered objects are a clear indicator for a homogeneous subgroup. Considering the attributes of each clustered node, we observe a correlation between the graph structure and some attribute values. Hence, a group of clustered nodes may only show high attribute similarity for a subset of relevant attributes. As illustrated in Figure 1(b), some subspaces show high correlation with the selected subgraph, while other attributes may tend to be irrelevant for this subgraph and show scattered attribute values.

As mentioned in Section II, recent techniques [21], [22], [23] set about solving Challenge 1. These approaches can detect subspace clusters in attributed graphs. We have consciously decided to take their results as input to our scoring functions in order to solve this first challenge. Our approach will even improve with future developments in this research area.

Challenge 2:

Scoring of objects in multiple subspace clusters

A naive outlier score would assign $score(o) = 1$ to all objects that occur in at least one cluster and $score(o) = 0$ to all objects that are not clustered. However, current graph clustering techniques in attributed graphs allow to obtain multiple views of an object w.r.t. the graph structure and the relevant subset of attributes. Such a function does not consider that an object might belong to several subspace clusters, and it misses thereby essential information about each object given by its cluster assignments. This information should be included for outlier ranking, and scoring should depend on the occurrence of objects in different subspace clusters.

OutRank is the first solution, which tackles Challenge 2 for high dimensional vector data [10], [14]. However, using the original *OutRank* score in case of attributed graphs is not enough. It only considers the attribute properties of subspace clusters. In Section IV, we will show that these properties are not sufficient for high quality outlier ranking on attributed graphs. Thus, the open challenge of outlier scoring on attributed graphs is very similar to the one of subspace selection. One has to consider both attribute and graph properties. Ranking functions require a unified score that incorporates all these properties in order to avoid loss of information. In particular, one has to incorporate the centrality of each node, its relevant attributes, and the objects it is clustered with.

C. Scoring based on Subgraphs and Subspaces

With *GOutRank*, we focus on Challenge 2 for both graph and attribute information in the following. We first review the solution presented in *OutRank* [10], [14]. Then, we present how we extend the basic idea of *OutRank* to graph information.

Definition 3: OutRank scoring

$$score_1(o) = \frac{1}{2} \cdot \sum_{\{(C,S) \in Res \;|\; o \in C\}} \frac{|C|}{c_{max}} + \frac{|S|}{s_{max}}$$

with $|C|$ being the number of objects in cluster C, and $|S|$ the number of attributes; c_{max} the maximal cluster size in Res and s_{max} the maximal dimensionality in Res.

This function defines outliers as objects that are found in abnormally few and low dimensional subspace clusters. Its core idea is that regular objects tend to cluster with many other similar objects. This is used as a first indication of the regularity of objects. The dimensionality of clusters is used as the second indication. Objects that are part of clusters with many attributes have strong dependencies in several properties. Hence, these regular objects get high scores. Please note, that in general [10], [14] one can weight the individual terms in the sum. For simplicity of presentation we skip this weighting parameter and use an equal weighting for cluster size and dimensionality.

This score generates high quality outlier rankings, as shown for several benchmark databases for vector data in a recent publication [12]. It has achieved higher quality results in comparison to several other outlier ranking techniques designed for high dimensional data [9], [11].

However, for attributed graphs $score_1(o)$ clearly misses some graph properties. To overcome this drawback, *GOutRank* defines two additional properties as indication for regular objects. They both utilize the centrality of a node in the graph structure.

First, we consider the local edge density to be a valuable criterion for our scoring. We search for isolated nodes in a strong connected graph structure. On the one hand, outliers are characterized by their low edge density. While on the other hand, highly connected subgraphs should be rated as indication for regular objects. In our example, co-purchases with many other products indicates the regularity of a product as a central hub from which other products are purchased. Outliers show only very few purchases and are clustered in sparsely connected subgraphs. Furthermore, this criterion can distinguish between nodes in multiple clusters with different edge densities. Overall, highly connected subgraphs are rated as better indication for regular objects than sparsely connected graphs.

Definition 4: GOutRank with node degree scoring

$$score_2(o) = \frac{1}{3} \cdot \sum_{\{(C,S) \in Res \;|\; o \in C\}} \frac{|C|}{c_{max}} + \frac{|S|}{s_{max}} + \frac{deg(o)}{deg_{max}}$$

with $deg(o) = |\{(o,p) \in E\}|$ and $\frac{deg(o)}{deg_{max}} \in [0,1]$ as the normalized edge degree of node o.

As second indication for regularity, we observe the centrality measure obtained by the Eigenvalues [28]. This measure has been used to immunize the most vulnerable node in a graph (e.g., to make it as robust as possible against a computer virus attack). It is based on a recent development in terms of graph centrality and provides an interesting indication for our regularity measure. The indicator is based on the observation that central nodes such as hubs form the core of the regular subgraph. Thus, high scores are assigned to these nodes.

Definition 5: GOutRank with eigenvalue scoring

$$score_3(o) = \frac{1}{3} \cdot \sum_{\{(C,S) \in Res \ | \ o \in C\}} \frac{|C|}{c_{max}} + \frac{|S|}{s_{max}} + \frac{|EV(o)|}{|EV|_{max}}$$

with $\frac{|EV(o)|}{|EV|_{max}} \in [0, 1]$ the normalized eigenvalue of node o.

Clearly, there are further centrality measures that could be used as instantiations of our model. We present only these two as a mixture of a basic degree scoring and a recent graph measure with eigenvalue scoring. Incorporating these basic graph properties shows significant quality improvement in our evaluation. But even more important, it highlights the potential for future regularity criteria in this scoring framework.

Finally, let us discuss the effects of the scoring functions and their intrinsic properties. They are designed as a conjunction of different indicators. Clear outliers are not part of any cluster, or they are part of clusters which only consist of nodes in very small, low dimensional, and sparsely connected subgraphs. All of these properties indicate a high deviation and lead to top ranking positions. Intermediate positions in the ranking are assigned to objects that show up in either large, high dimensional, or densely connected subgraphs. Finally, clear regular objects are clustered in large, high dimensional, and densely connected subgraphs, and thus, will be ranked at the bottom. For the graph-based components of $score_2$ and $score_3$ we expect centrality measures to provide an enhanced distinction between individual objects. In this respect, *GOutRank* can be considered as a general framework. It enhance its detection quality by novel developments in both centrality measures and subspace clustering.

IV. Experiments

In our empirical evaluation, we show the potential and the capabilities of our *GOutRank* method on a real world database. The dataset has been extracted from the Amazon co-purchase network [29] and restricted to Disney DVDs (124 nodes with 334 edges). In addition to this graph structure, we extracted further product information (e.g., product prices, different rating ratios, product reviews) from the Amazon website. Our attributed graph consists of 30 attributes per node. The existing graph clusters correspond to similar Disney films such as *Disney Pixar* Films or *Disney* classics. Product O_1 from Figure 1(b) is one of the real world outliers that corresponds to the overpriced film[2] *The Jungle Book (1994)* of *Rudyard Kipling's* hidden in the cluster of *Read-Along Disney* films.

For our quality assessment we use a ground truth of real outliers, which we obtained from a user experiment. Outliers have been labeled by a class of high school students as domain experts for the selected subgraph. We firstly obtain the graph clusters with a modularity based technique [30]. Thus, we have ensured that students do not simply label global outliers (e.g., product with the highest price of the database). Each co-purchased group was shown to the students as a product list, and they had to label one or two items that they considered deviating from the others in the group. For the ground truth we have deemed all products outliers that have been labeled as outlying by at least 50% of the students. Figure 1(a) shows the entire *Disney* network.

The dataset and a detailed description is publicly available on our website. It can be used as a benchmark database for outlier mining on attributed graphs. To the best of our knowledge it is the first attributed graph with a labeled outlier ground truth. Obviously, it is only a small data set. However, it is an interesting benchmark database due to its complex graph and attribute structure. In this line, we hope to facilitate comparability with future developments in this research area.

In our evaluation we compare *GOutRank* to the following outlier ranking techniques: LOF (only attributes, without subspace analysis) [6], SOF and RPLOF (only attributes with subspace analysis) [8], [9], SCAN (graph clustering that detects structural outliers) [19], and CODA (graph and attribute outlier mining, without subspace analysis) [24]. In addition, we evaluate our approach with different graph clustering approaches: CoPaM [21], GAMer [22] and an extension of Cocain [31]. All of these clustering techniques are publicly available [22].

A. Comparison to competing approaches

Figure 2 shows AUC (area under the ROC curve) measures and the runtimes for all approaches. The loss of information is clearly visible for both paradigms: (1) approaches using only attributes and (2) approaches using only the graph structure. For the first paradigm, we observe a higher quality of subspace outlier mining [8], [9] compared to the full space method [6]. This is due to the selection of relevant attributes for each individual outlier. However, they miss several outliers, hidden in combination of both data types, due to the loss of graph information. On the other hand, graph-based approaches [19], [24] show very low AUC. Although CODA has both graph and attribute information available, it fails due to the curse of dimensionality in the full attribute space. Overall, *GOutRank* is not as fast as the competing approaches. However, the runtimes depend heavily on the used subspace clustering technique (cf. Section IV-B) and *GOutRank* clearly outperforms all competitors with a significant quality enhancement. It is able to cope with both attribute and graph information and with large numbers of given attributes. It is a successful synthesis of both graph and attribute information with high quality due to its outlier detection in selected subspaces.

[2] http://www.amazon.com/dp/B00005T5YC

Used data	Paradigm	Algorithm	AUC[%]	Runtime[ms]
(1) attribute data only	full space outlier analysis	LOF [6]	56.85	41
	subspace outlier analysis	SOF [8]	65.88	825
		RPLOF [9]	62.50	7
(2) graph structure only	graph clustering	SCAN [19]	52.68	4
(3) both attributes and graph data	full space outlier analysis	CODA [24]	50.56	2596
	subspace outlier analysis	GOutRank	**86.86**	26648

Fig. 2. AUC results for all competitors on the Amazon database [Disney DVD selection]

Graph Subspace Clustering	Score Function	AUC[%]	Runtime Pre-processing [ms]	Runtime Score [ms]
GAMer [22]	$score_1$ (only attributes)	75.28	26648.60	0.20
	$score_2$ (node degree)	82.91		0.16
	$score_3$ (eigenvalue centrality)	**86.86**		0.25
extension of Cocain [31]	$score_1$ (only attributes)	75.85	123948.10	14.07
	$score_2$ (node degree)	76.97		16.91
	$score_3$ (eigenvalue centrality)	**77.96**		16.00
CoPaM [21]	$score_1$ (only attributes)	58.61	1615.20	1.34
	$score_2$ (node degree)	69.49		1.30
	$score_3$ (eigenvalue centrality)	**72.45**		1.31

Fig. 3. Quality w.r.t. different graph clustering techniques and scoring functions

B. Including different clusterings and scores

GOutRank allows to use any subspace graph clustering as pre-processing step. Thus, we compare *GOutRank* with the different scoring functions and three clustering inputs [22], [31], [21] in Figure 3. Regarding the different clustering schemes, we observe best results for GAMer, the most recent graph clustering approach based on subspace analysis. *GOutRank* finds most of the hidden outliers due to its high quality clustering. In comparison with $score_1$, we observe a clear benefit of the enhanced scoring functions $score_2$ and $score_3$, which take the centrality of the nodes into account. Both other clustering approaches (i.e. the extension of Cocain and CoPaM) have AUC values that are worse. In all cases $score_2$ and $score_3$ can improve over the traditional scoring of *OutRank*. Figure 3 shows also the runtimes of the pre-processing step and the calculation of each of the scores. In all cases, the overhead for scoring is negligible in comparison to the runtime of the subspace clustering algorithms.

Overall, our experiments show that *GOutRank* with $score_3$ performs best. The results also highlight the high outlier ranking quality of *GOutRank* for the most recent graph clustering technique. This indicates that improving the graph clustering techniques can lead to an increased outlier detection quality of *GOutRank*.

V. Conclusions

With *GOutRank* we have proposed a first solution for outlier ranking in subspaces of attributed graphs. Graph nodes are ranked according to their outlierness regarding both graph and attribute properties. We build upon graph clustering and subspace analysis as pre-processing steps to our outlier scoring. Both contribute to the high quality result in our evaluation. In all other cases we observe a significant decrease of the AUC values due to information loss w.r.t. graph data, or because there is no subspace analysis. We have made similar observations for our outlier scoring functions. They capture outlierness w.r.t. both subgraphs and subspaces. They are able to detect high quality outliers in attributed graphs. Our evaluation with the Amazon network proposes the first benchmark for outlier mining in attributed graphs and highlights that *GOutRank* has assigned high ranking positions to most of the user-labeled outliers.

VI. Open Challenges and Future Work

Our future work in this area will focus on several open challenges. In the following, we describe the most promising ones, which have been derived out of our case study on the Amazon network.

Open Challenge 1:

**Integration of outlier ranking
into graph clustering algorithms**

As first open challenge, we see high potential in the integration of outlier ranking into the actual graph clustering process. This would allow an interactive exploration of outliers during the cluster computation. Top-k results could be computed directly out of the clustering task without computing scores for all objects in the database as a post-processing.

Open Challenge 2:

**Scalable computation of outlier rankings
in large attributed graphs**

Our current two step processing has clear drawbacks w.r.t. scalability. It has to mine all subspace clusters first, before computing scores for each object in a second step. Integration of these two steps might lead to first efficiency improvements. However, further heuristics and approximations will be required to reduce the complexity in main bottlenecks such as subspace selection and complex scoring functions.

Open Challenge 3:

Scoring by comparison of graph clusters

As third open challenge, we observe the comparison of the obtained graph clusters. In the current solution, each cluster is considered individually in the scoring function. However, new indicators for outliers might be derived if one compares two clusters. In particular, one could exploit the redundancy of clusters (i.e. the coverage of objects in multiple clusters) to refine the indicators.

Open Challenge 4:

Incorporating more complex clustering models

Complex clustering models could be considered in order to exploit the hierarchical embedding of graph clusters for novel indicators. Hierarchies are of interest in many real world applications where clusters do not form a flat partitioning but a hierarchy of clusters that include each other. In particular, we observe the overlap of clusters to capture a high potential for future scoring functions. Overlap of clusters occurs in hierarchical clusterings, but also in recent multi-view clustering that searches for clusters in multiple perspectives of the database [32].

Open Challenge 5:

Enhanced graph measures and subspace selections for each individual node

Finally, we observe an open challenge in the extraction of further node properties as indicators for our scoring. There is a variety of centrality measures available that could be used for the structural outlierness of a node. However, we see even more potential in enhanced selection methods for individual subspaces. Current clustering techniques compute one subspace for the entire cluster. Individual sets of attributes for each node might provide even better outlier scores.

Overall, there are several directions that have been opened by the basic idea of *GOutRank* for future research. We are looking forward to exploit this potential for future improvements in graph outlier ranking.

Acknowledgments

This work is supported by the Young Investigator Group program of KIT as part of the German Excellence Initiative, by a post-doctoral fellowship of the research foundation Flanders (FWO), and by the German Research Foundation (DFG) within IME Graduate School at KIT.

References

[1] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Wiley, 1987.

[2] E. Knorr and R. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," in *VLDB*, 1998, pp. 392–403.

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

[4] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos, "Relevance search and anomaly detection in bipartite graphs," *SIGKDD Explorations*, vol. 7, no. 2, pp. 48–55, 2005.

[5] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbors meaningful," in *IDBT*, 1999, pp. 217–235.

[6] M. Breunig, H. Kriegel, R. Ng, J. Sander *et al.*, "LOF: identifying density-based local outliers," *Sigmod Record*, vol. 29, no. 2, pp. 93–104, 2000.

[7] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *ICDE*, 2003, pp. 315–326.

[8] C. Aggarwal and P. Yu, "Outlier detection for high dimensional data," *ACM Sigmod Record*, 2001.

[9] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *KDD*, 2005, pp. 157–166.

[10] E. Müller, I. Assent, U. Steinhausen, and T. Seidl, "Outrank: Ranking outliers in high dimensional data," in *DBRank Workshop*, 2008, pp. 600–603.

[11] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *PAKDD*, 2009, pp. 831–838.

[12] E. Müller, M. Schiffer, and T. Seidl, "Statistical selection of relevant subspace projections for outlier ranking," in *ICDE*, 2011, pp. 434–445.

[13] F. Keller, E. Müller, and K. Böhm, "HiCS: High contrast subspaces for density-based outlier ranking," in *ICDE*, 2012.

[14] E. Müller, I. Assent, P. Iglesias, Y. Mülle, and K. Böhm, "Outlier ranking via subspace analysis in multiple views of the data," in *ICDM*, 2012.

[15] D. Chakrabarti, "Autopart: Parameter-free graph partitioning and outlier detection," in *PKDD*, 2004, pp. 112–124.

[16] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *KDD*, 2003, pp. 631–636.

[17] W. Eberle and L. B. Holder, "Discovering structural anomalies in graph-based data," in *ICDM Workshops*, 2007, pp. 393–398.

[18] L. Akoglu, M. McGlohon, and C. Faloutsos, "oddball: Spotting anomalies in weighted graphs," in *PAKDD*, 2010, pp. 410–421.

[19] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "Scan: a structural clustering algorithm for networks," in *KDD*, 2007, pp. 824–833.

[20] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," *PVLDB*, vol. 2, no. 1, pp. 718–729, 2009.

[21] F. Moser, R. Colak, A. Rafiey, and M. Ester, "Mining cohesive patterns from graphs with feature vectors," in *SDM*, 2009, pp. 593–604.

[22] S. Günnemann, I. Färber, B. Boden, and T. Seidl, "Subspace clustering meets dense subgraph mining: A synthesis of two paradigms," in *ICDM*, 2010, pp. 845–850.

[23] L. Akoglu, H. Tong, B. Meeder, and C. Faloutsos, "Pics: Parameter-free identification of cohesive subgroups in large attributedgraphs," in *SDM*, 2012.

[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 *KDD*, 2010, pp. 813–822.

[25] E. Müller, S. Günnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data," *PVLDB*, vol. 2, no. 1, pp. 1270–1281, 2009.

[26] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," *TKDD*, vol. 3, no. 1, pp. 1–58, 2009.

[27] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong, "A Survey on Enhanced Subspace Clustering," *DMKD*, 2012.

[28] H. Tong, B. A. Prakash, C. E. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. Chau, "On the vulnerability of large graphs," in *ICDM*, 2010, pp. 1091–1096.

[29] J. Leskovec, L. Adamic, and B. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 5, 2007.

[30] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.

[31] Z. Zeng, J. Wang, L. Zhou, and G. Karypis, "Coherent closed quasi-clique discovery from large dense graph databases," in *KDD*, 2006, pp. 797–802.

[32] E. Müller, S. Günnemann, I. Färber, and T. Seidl, "Discovering multiple clustering solutions: Grouping objects in different views of the data," in *ICDM*, 2010, p. 1220.