Researchpedia
Open Access Journals

Received: 5 July 2020; **Revised:** 13 Dec 2021; **Accepted:** 19 Dec 2021; **Published:** 26 Dec 2021
Journal Name [Online ISSN Coming Soon], Volume 2, Issue 1, Article 8, Pages 81-86, December 2021

Digital Object Identifier 10.1111/RpJC.2020.DOINumber

Anomaly Detection in Academic Social Networks using Deep Clustering

Khurram Shehzad[1], Safiullah Afzal[1], Muhammad Waqar[1], and Malik Khizar Hayat[2]

[1]Department of Software Engineering, Foundation University Islamabad, Rawalpindi Campus, Pakistan
[2]Department of Information Technology, University of Haripur, Pakistan

Corresponding author: Malik Khizar Hayat (e-mail: khizerhayat92@gmail.com)

ABSTRACT Detection of anomalies that are evolutionary by nature has emerged as a trending research topic in many areas, such as security, bioinformatics, education, economy and so on. Although, most of the research has focused on detecting anomalies using evolutionary behavior among objects in a network. However, in the real-world heterogenous networks, multiple types of objects co-evolve together with their attributes. To understand the deviant co-evolution of multi-typed objects in heterogeneous information networks (HINs), a special approach is required that can capture abnormal co-evolution of multi-typed objects. Detecting co-evolution-based anomaly in heterogeneous bibliographic information network can portray better the object-oriented semantics than just analyzing the co-author or citation network alone. In this paper, we propose a deep clustering-based model for anomaly detection in Microsoft Academic Graph (MAG). The star-network schema is used to process the MAG data. Feature learning and clustering tasks are combined using deep learning. Experimentation on the MAG data shows the efficiency of the proposed model.

Keywords Anomaly detection, Clustering, Deep learning, Heterogenous networks, Microsoft academic graph.

I. INTRODUCTION

With the upsurge in social web usage as a communication platform, these days, it has intensified the speed of data generation. For instance, the social networks such as Facebook, Twitter, Wordpress, as communication platforms, and DBLP, and ArnetMiner, ACM, and MAG as academic social networks. These networks are dynamic in terms of their structure (entities/objects and attributes) and constantly changing with respect to time which also makes them evolutionary in nature. Hence, these provide a comprehensive and cohesive structure among different types of objects and their attributes. MAG is also such a type of academic social network which comprise different types of objects and their inter-linked attributes. Several problems have been discussed in bibliographic networks including rising star prediction [1], dynamic research interests finding [2], topic-based heterogeneous ranking [3], anomaly detection, and so on. However, deep learning has been studies for social media analytics [4], however, regarding anomaly detection, it still needs exploration in bibliographic networks. It also has applications to other problem domains including spam detection, network intrusion detection, and fake news detection in social networks [5]. There are different platforms which provide bibliographic data such as DBLP, ArnetMiner, Scopus, Google Scholar, and Microsoft Academic Graph. In this research work, MAG is used for experimentation purpose.

The co-evolution analysis in networked data needs a distinct method from the typical methods used in evolutionary computing [6]–[9]. Importantly, in the task anomaly detection, the focus should be on the reason of being anomalous. Though, it is not practical to analyze the co-evolution for every possible node or a subnetwork in the input graph. Therefore, using a pre-defined structured schema like star-network is feasible [10]. Many earlier studies are available discussing the anomaly detection problem for evolutionary networks. For instance, community outlier detection [11], community-trend, evolutionary, and community distribution outlier detection [12]. Regarding anomalous changes in the graph structure, localizing relationships of nodes that are responsible for anomalousness is presented in [13]. To be brief, most of the existing techniques detect anomalies based on the evolutionary behavior of objects in the HINs and ignoring the co-evolutionary aspect. Particularly, we present a deep clustering-based anomaly detection approach which detects anomalies in a MAG.

In a bibliographic network, different objects and attributes are interlinked with each other. However, not every attribute is of equal significance, or influence. Therefore, it is necessary to extract meaningful features from the pool and use them for anomaly detection. Deep learning is extensively being used for feature learning from the networked data. In this work, we use deep autoencoders for feature learning, and then use k-means clustering to detect anomalies from MAG. The proposed method

can be applied to the evolutionary networks. The contributions made in this study are as follows:

1) Proposed a deep clustering-based method to detect anomalies in academic networks
2) Explored MAG for the anomaly detection problem
3) Experiments show the efficiency of the proposed method

The rest of the paper is organized as follows: related work is discussed in Section 2, proposed method is described in Section 3, experimentation is done in Section 4, and section 5 concludes the study.

II. RELATED WORK

Extensive work has been done on the problem of anomaly detection [14], [15]. However, this work is focused on the co-evolution among attributes of objects in the MAG. There are some related studies exist [5], [11], [12], [16], which studies the same problem. Though, most of them are dealing with homogeneous networks. Also, in the previous studies, the objects under consideration are of single-typed, whereof, in heterogeneous networks, both the objects and attributes are of multi-typed.

Graphs are of great interest when it comes to the identification of what is not regular – anomalous. A number of studies exist which explores graphical data to detect anomalies. It comprise the community detection-based methods [11], [16], where authors rely on interactions between objects and attributes of the same type whereas, in this work, the focus is on dealing multi-typed objects and attributes in MAG. Authors also discussed the density-based methods [17], [18], query-based [6], evolving graph-based [19], and subgraph-based [15], [20] methods to detect anomalies in the networked data. Nevertheless, all these approaches focus on a network of single-typed objects and attributes such as friends, authors, or events in their studies. The influence of attributes over multi-typed objects is not considered for anomaly detection. [21] tried to detect anomalies based on a single snapshot of a static network. [22] made an effort by considering anomaly detection problem as a binary classification task.

Authors [15], [23], [24] also explored the anomaly detection using pattern mining in graphs. It is to detect insignificant patterns and co-evolution patterns [25] from the graphs as subgraphs based on a user-specified threshold value. In view of author collaboration graph, for instance, there are nodes representing authors with their research areas and edges showing collaboration frequency between different research area authors namely DB, DM, AI, IR, and ML. The pattern that regularly exists is authors having related area of research as cooperation between DB/DB, or DM/DM researchers and so on. At this instance, detecting subgraphs of authors having collaborations between DB/IR, AI/DM, or IR/ML depict interdisciplinary collaborations [26] that are interesting to know if deliberated as detected anomalies.

III. ANOMALY DETECTION USING DEEP CLUSTERING

In this section, we formally define the problem and explain the proposed methodology in detail.

A. PROBLEM DEFINITION

Anomaly detection in MAG is to detect the deviant evolutionary behavior of objects and attributes. Several of the previous studies discussed single-typed anomaly detection with focus on the evolutionary behavior of objects, however, co-evolution among objects and attributes is overlooked. The MAG data has different objects such as *authors, affiliations, field of study, papers, coauthors, citations,* and so on. Each object has a number of attributes defined along with the time information in the form of *year*. The research problem can be divided into two following two parts:

1) What are the attributes which significantly deviate from the other in a specific timestamp?
2) How the extracted attributes influence the objects in their anomalous declaration?

B. FEATURE EXTRACTION

Feature extraction is the major task in anomaly detection. It is because from a pool of candidate features of objects, it is essential to extract meaningful features using deep learning. Later, clustering will be applied on these features to detect anomalies. We adopt Deep AutoEncoder (DAE) as used in [27] to extract features from MAG objects.

A DAE is a neural network of three layers – input layer, hidden layer, and output layer. Input layer is knowns as encoder and the output layer is decoder. The encoder at input layer is formulated as:

$$a_i = p(W_i x + b_i) \tag{1}$$

where a_i is the hidden features of the input data, W_i is the weight and b_i is the bias of encoder at layer i. In the same way, the decoder is defined as:

$$\bar{h} = q(W_j a_i + b_j) \quad (2)$$

where $\bar{h}$ is the reconstruction of the encoded data, W_j is the weight and b_j is the bias of encoder at layer j.

In these equations, i and j shows the number of timestamps under consideration from the MAG data. DAE may have several layers stacked over one another to learn the hidden features at lower-level which is given as input to a DAE of higher-level. Particularly, the first layer always takes the raw data as input to process it further.

C. CLUSTERING

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific. We use Euclidean distance as the similarity measure in k-means clustering algorithm.

Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. Unlike supervised learning, clustering is considered an unsupervised learning method since we do not have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups. Consequently, the objects which are far from the rest of the clusters are considered as anomalies in MAG. Noticeably, the clusters represent the research areas and the objects which deviate significantly from one research area to the other in consecutive timestamps are the actual anomalies in MAG. Consequently, it also shows the co-evolutionary behavior of attributes of objects in different timestamps. Clustering of the obtained features from the previous step is done using k-means clustering algorithm. Following is the pseudocode followed for anomaly detection in MAG:

1. Using MAG data, learn the features for objects using Eq. (2)
2. Select the objects from MAG with the feature information from step 1
3. Define the number of clusters k
4. Initialize the centroids at random
5. Iterate through the objects assignment to clusters until centroids do not change
6. Compute Euclidean distance between the objects and all centroids
7. Assign each object to the closest centroid

Finally, the objects which comprise their own cluster of single objects or the objects which belong to different clusters in different timestamps are the actual anomalies.

IV. EXPERIMENTS

In this section, we explain the dataset, the evaluation of clustering, and the results.

A. DATASET

The dataset MAG is used for experimentation purpose that is pre-processed for outliers, missing data, and the normalization. Based on the author keywords, we selected four different research areas to cluster – image processing, data mining, information retrieval, and computer networks. Moreover, following the star network schema, we select the authors as target object and following as attribute objects from MAG:

- Papers
- Affiliations
- FieldOfStudy
- ConferenceSeries

Each of these objects have different features/attributes which are learned using DAE as explained in Section III. The object *Paper* has an attribute named *year* which is used for distributing the dataset into different timestamps. We use 5 timestamps where each timestamp is of 3 year which overlapped as following:

- Timestamp 1: 2000 – 2003
- Timestamp 2: 2002 – 2005
- Timestamp 3: 2004 – 2007
- Timestamp 4: 2006 – 2009
- Timestamp 5: 2008 – 2011

The purpose of using overlapping timestamps is to make sure that there is no biasness in the data.

B. EVALUATION

In each timestamp, the results of clustering are separately used in order to analyze the co-evolution behavior of attributes over the objects while moving from one cluster to another in different timestamps. Clustering quality is subjective to measure. Among two well-known clustering quality approaches: *Internal Evaluation* is used when a result of clustering is assessed based on the data that was clustered itself, and *External Evaluation* is used when we have labels for classes and external benchmarks. These benchmarks comprise a set of pre-classified objects, however, we do not have benchmarks in this dataset. That is the reason we use internal evaluation measure - ***Davies-Bouldin Index*** [22] for measuring clustering quality. It evaluates intra-cluster similarity and inter-cluster differences as desired. Davies-Bouldin Index (DBI) is defined as the ratio of within and between cluster similarities. Mathematically, it can be represented as follows:

$$\boldsymbol{DB} = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \max_{\boldsymbol{j \neq i}} \{\boldsymbol{D_{i,j}}\} \tag{3}$$

where $\boldsymbol{D_{i,j}}$ is within to between clusters ratio for i^{th} and j^{th} cluster, ***max*** represents the worst case of the DBI. The smallest value is the optimal value considered for this index that refers to as low as the value of DBI, the higher will be the quality of clusters.

C. RESULTS

The clustering results are evaluated using the evaluation measure discussed in the previous section. We used the same value of DB index as used in the [19], because the research areas are same in number as used in that study. Figure 1 shows the DB index comparison with different number of clusters; however, we selected the value of DB index as 0.2 with 5 clusters which is possibly near to the total clusters we want to have at the end.

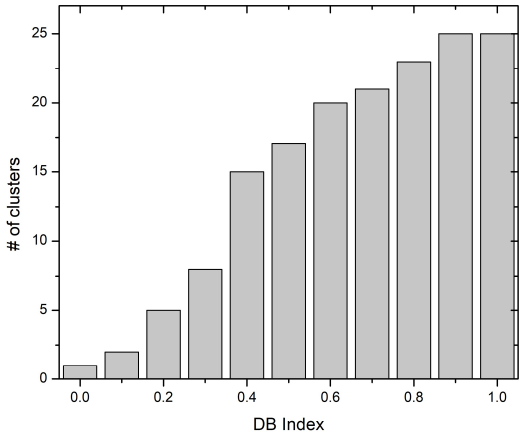

Figure 1: DB Index Comparison

In order to depict the clustering results, we computed similarity between clusters members in each timestamp. This similarity is based on the research area of the clustered objects. The greater the similarity, it is more likely that clustered members will keep the same cluster membership through each timestamp. In this way, it is easier to spot the anomalous timestamps with attributes

having deviated influence over objects that are declared as anomalies. Figure 2 shows the similarity value comparison of clustering results in different timestamps

.

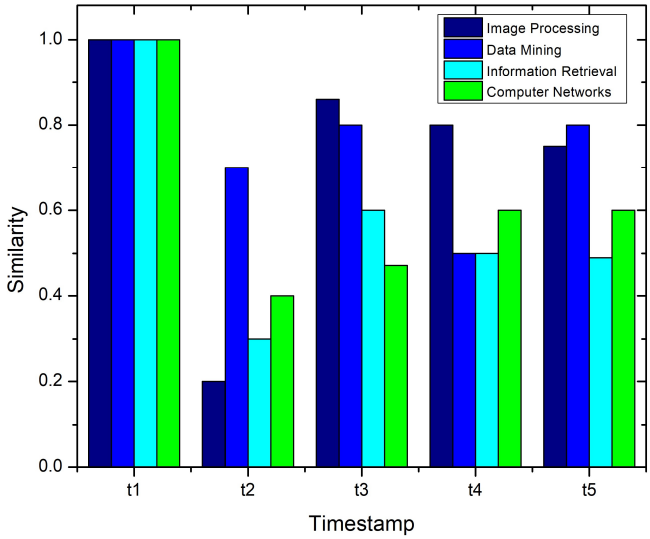

Figure 2: Cluster Membership Similarity Comparison for each Timestamp

For timestamp t_1, all of the research clusters are 100 percent similar because the similarity comparison starts in the second timestamp. In t_2, data mining field of study has the highest similarity which shows that attributes are not deviating in t_2 for this field of study. However, image processing field has the lowest similarity which shows that comparing t_1 and t_2, the attributes have deviated influence on the objects which caused less similarity between consecutive timestamps. During t_3, there is a rapid rise in similarity for image processing field of study. Comparing t_1 and t_3, image processing has the highest similarity and the computer networks field has the lowest similarity. It shows that the attributes in the computer networks has the most deviant influence on the objects during t_3. During t_4, every field of study has the similarity value higher or equal to 50 percent. During t_5, almost every field has again the similarity value higher or equal to 50 percent. However, information retrieval is somehow less than 50 percent which shows that influence has deviant influence on objects during t_5 for the information retrieval field of study.

V. CONCLUSION

In this research work, we introduced deep clustering-based anomaly detection for the MAG data using the star network schema. It is prevalent to analyze the deviated influence of attributes over objects in an academic/bibliographic network. Because anomalies occur due to the co-evolution of different attributes over time. It is concluded that the proposed method can better be utilized for analyzing exchangeability of research areas in MAG, hence spot out the anomalies in the network. Additionally, identifying the most influential attributes for an anomalous target object may help in anomaly prevention as well. From the future perspective, other bibliographic networks such as ACM, Scopus, can be explored for the anomaly detection by employing schemas other than the star network schema. It would reveal more hidden knowledge and rich semantics from the bibliographic networks. The proposed method can be applied to other heterogeneous networks such as Facebook, Twitter, or healthcare information networks.

REFERENCES

[1] A. Daud, M. Song, M. K. Hayat, T. Amjad, and R. A. Abbasi, “Finding rising stars in bibliometric networks,” *Scientometrics*, pp. 1–29, 2020.
[2] A. Daud, “Using Time Topic Modeling for Semantics-Based Dynamic Research Interest Finding,” *Journal of Knowledge-Based Systems*, vol. 26, pp. 154–163, 2012.
[3] T. Amjad, Y. Ding, A. Daud, J. Xu, and V. Malic, “Topic-based Heterogeneous Rank,” *Journal of Scientometrics*, vol. 104, no. 1, pp. 313–334, 2015.
[4] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood, “Towards deep learning prospects: Insights for social media analytics,” *IEEE Access*, vol. 7, pp. 36958–36979, 2019.
[5] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller, “Focused clustering and outlier detection in large attributed graphs,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1346–1355.
[6] A. Dalmia, M. Gupta, and V. Varma, “Query-based Evolutionary Graph Cuboid Outlier Detection,” Barcelona, Spain, 2016.
[7] M. Gupta, C. C. Aggarwal, J. Han, and Y. Sun, “Evolutionary Clustering and Analysis of Bibliographic Networks,” Kaohsiung, Taiwan, 2011.

[8] X. Sun, K. Ding, and Y. Lin, "Mapping the Evolution of Scientific Fields Based on Cross-Field Authors," *Journal of Informetrics*, vol. 10, no. 3, pp. 750–761, 2016.
[9] M. Gupta, J. Gao, Y. Sun, and J. Han, "Integrating community matching and outlier detection for mining evolutionary community outliers," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 859–867.
[10] Y. Sun, Y. Yu, and J. Han, "Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema," Paris, France, 2009.
[11] 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 SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 813–822.
[12] M. Gupta, J. Gao, and J. Han, "Community Distribution Outlier Detection in Heterogeneous Information Networks," Prague, Czech Republic, 2013.
[13] K. Sricharan and K. Das, "Localizing Anomalous Changes in Time-evolving Graphs," Snowbird, UT, USA, 2014.
[14] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *Journal of ACM Computing Surveys*, vol. 41, no. 3, pp. 1–72, 2009.
[15] 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.
[16] M. Gupta, J. Gao, Y. Sun, and J. Han, "Community trend outlier detection using soft temporal pattern mining," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 692–708.
[17] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," Dallas, TX, USA, 2000.
[18] F. Angiulli and F. Fassetti, "Toward Generalizing the Unification with Statistical Outliers: The Gradient Outlier Factor Measure," *ACM Transactions on Knowledge Discovery from Data*, vol. 10, no. 3, pp. 1–27, 2016.
[19] P. Leto and A. Clauset, "Detecting Change Points in the Large-Scale Structure of Evolving Networks," Austin, Texas, USA, 2015.
[20] M. Gupta, A. Mallya, S. Roy, J. H. D. Cho, and J. Han, "Local Learning for Mining Outlier Subgraphs from Network Datasets," Pennsylvania, USA, 2014.
[21] P. Bindu, P. S. Thilagam, and D. Ahuja, "Discovering Suspicious Behavior in Multilayer Social Networks," *Computer in Human Behavior*, vol. 73, pp. 568–582, 2017.
[22] Y. Yasami and F. Safaei, "A Statistical Infinite Feature Cascade-Based Approach to Anomaly Detection for Dynamic Social Networks," *Computer Communications*, vol. 100, pp. 52–64, 2017.
[23] P. Bindu and P. S. Thilagam, "Mining Social Networks for Anomalies: Methods and Challenges," *Journal of Network and Computer Applications*, vol. 68, pp. 213–229, 2016.
[24] Y. Sun and J. Han, "Mining Heterogeneous Information Networks: A Structural Analysis Approach," Beijing, China, 2013.
[25] M. K. Hayat and A. Daud, "Anomaly detection in heterogeneous bibliographic information networks using co-evolution pattern mining," *Scientometrics*, vol. 113, no. 1, pp. 149–175, 2017.
[26] W. Wei and K. M. Carley, "Measuring Temporal Patterns in Dynamic Social Networks," *Journal of Knowledge Discovery from Data*, vol. 10, no. 1, pp. 1–27, 2015.
[27] K. Tian, S. Zhou, and J. Guan, "Deepcluster: a general clustering framework based on deep learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017, pp. 809–825.