NEIGHBORHOOD RECONSTRUCTION FOR GRAPH ANOMALY DETECTION

Amit Roy
Purdue ID: 34778209

1 INTRODUCTION

In this project, we focus on detecting anomalous nodes over attributed static graphs. An attributed graph $G = (V, E, X) \in \mathcal{G}$ consists of a vertex set $V = \{1, 2, \cdots, N\}$ and an edge set E. $X = [\cdots x_u^\top \cdots]^\top \in \mathbb{R}^{N \times k}$ collect all node attributes and $x_u \in \mathbb{R}^k$ is the attribute for node u. The degree of node u is denoted as d_u. The objective of this project is to perform unsupervised anomaly detection. Each node u has an anomaly label y_u where $y_u = 0$ or $y_u = 1$ implies node u is normal or anomalous respectively. The goal is to design a detection method $f(G) : \mathcal{G} \rightarrow \{0, 1\}^N$ that associates each node with a label. However, these node labels are assumed to be unknown when designing f.

Let $\mathcal{N}_u$ be the set of 1-hop neighbor nodes of node u. Let $\bar{\mathcal{N}}_u$ be an augmented set of 1-hop neighborhood of node u that includes the attribute of node u, the set of the attributes of its neighbors, i.e., $\bar{\mathcal{N}}_u\{x_u, \{x_v | v \in \mathcal{N}_u\}\}$. Our assumption to detect anomalous nodes is that given the label y_u, the distribution $p(\bar{\mathcal{N}}_u | y_u)$ are different across norms and anomalies. Here, we consider just one-hop neighborhood as a proof of concept, which is also often adequate for use cases in practice Akoglu et al. (2012). The neighborhoods considered can be extended to the multi-hop case, while extra computation costs need to be paid in that scenario.

2 DATASET

We utilize the following anomaly detection datasets which are real-world graph datasets with benchmarking anomalies.

Dataset	# Nodes	# Edges	# Feat.	Avg. Degree	Ratio
cora	2,708	11,060	1,433	4.1	5.1%
amazon	13,752	515,042	767	37.2	5.0%
flickr	89,250	933,804	500	10.5	4.9%
weibo	8,405	407,963	400	48.5	10.3%
reddit	10,984	168,016	64	15.3	3.3%
disney	124	335	28	2.7	4.8%
books	1,418	3,695	21	2.6	2.0%
enron	13,533	176,987	18	13.1	0.4%

Table 1: Statistics of graph anomaly detection datasets following the Benchmarking Outlier Node Detection (Liu et al. (2022a)) paper

3 DEEP LEARNING METHOD USED

To detect anomalies in unsupervised static attributed graphs, we assume that the neighborhood pattern between a normal node and an anomalous node is distinguishable. For instance, the regular nodes vs. spammers have different interaction activities in social media, or the normal account vs. the fake users have different transaction patterns in financial transaction networks. Motivated by these real-world examples, we wish to employ the reconstruction loss of neighborhood as the anomaly detection score. The objective is to learn the distribution of normal vs anomalous nodes based on the reconstruction of their neighborhood. The neighborhood of a target node u can be represented as its own features x_u, its number of neighbors or degree d_u, and the features of its

neighbors $x_{\mathcal{N}_u}$. To reconstruct these properties of a target node u, we will utilize Neighborhood Reconstruction-based Graph Auto-Encoder Tang et al. (2022). Traditional Graph Auto Encoder (GAE), reconstructs the links of a graph in the decoder part with the node representations obtained from the graph convolution. However, reconstructing the links of a graph might be useful for tasks like link prediction. For anomaly detection, the reconstruction of links might not be suitable. For this reason, the plan is to employ neighborhood reconstruction-based GAE for learning the distribution of normal and anomalous nodes.

4 Related Work

In real-world applications, most of the graphs have node attributes (features). Nodes with inconsistent attributes have a high chance to be an anomaly node. Moreover, considering the information on node attributes along with structure helps to locate anomalies more accurately. Detecting anomalies in attributed networks can be achieved by clustering methods Bojchevski & Günnemann (2018); Perozzi et al. (2014a), interaction with human experts Ding et al. (2019b), group merging technique Zhu & Zhu (2020). Network embedding methods Perozzi et al. (2014b); Grover & Leskovec (2016); Tang et al. (2015) can also be applied to GAD on attributed graphs Bandyopadhyay et al. (2019; 2020b).

Auto-Encoder framework that focuses on extracting principal components from the data via deep learning has been extensively applied in anomaly detection Bandyopadhyay et al. (2020a); Kipf & Welling (2016); Ding et al. (2019a); Fan et al. (2020); Luo et al. (2022). GAE built upon GNNs can combine node attributes and graph structure properly and can detect anomalies based on checking the reconstruction loss of node attributes or links Kipf & Welling (2016); Ding et al. (2019a); Fan et al. (2020). But these works do not reconstruct the entire neighborhood for GAD. Rather, they use reconstruction error, and estimating Gaussian mixture density is also applied for GAD Li et al. (2019). Some works view nodes with multiple views and a node may or may not be considered an anomaly in different views. These nodes hold attributes from multiple views of the identity. To capture such multi-view information, multiple GNNs are often applied Peng et al. (2020); Sheng et al. (2019); Wu et al. (2014; 2013); Liu et al. (2022b) for anomaly detection. GNNs have also been applied to detect anomalies in multiple scales Gutiérrez-Gómez et al. (2020), and to detect anomalies and solve recommendation tasks simultaneouly Wang et al. (2019); Zhang et al. (2020). More involved techniques such as contrastive learning, self-supervised learning Xu et al. (2022); Jin et al. (2021); Liu et al. (2021); Zhou et al. (2022); Corsini et al.; Huang et al. (2021) and reinforcement learning Morales et al. (2021); Ding et al. (2019b); Langford & Zhang (2007) have also been recently applied to GAD.

5 Expected Results

With the neighborhood reconstruction-based GAE, the performance of anomaly detection should outperform the performance of anomaly detection with the traditional GAE. For the evaluation purpose, we will compare the results with traditional GAE-based methods e.g. MLPAE, ANOMALOUS, DOMINANT, DONE, AnomalyDAE. Also, we wish to improve the scalability of the model with the reconstruction of the neighbor feature distribution compared to neighborhood alignment prediction of the NRGAE.

References

Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. Fast and reliable anomaly detection in categorical data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 415–424, 2012.

Sambaran Bandyopadhyay, N Lokesh, and M Narasimha Murty. Outlier aware network embedding for attributed networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 12–19, 2019.

Sambaran Bandyopadhyay, Saley Vishal Vivek, and MN Murty. Outlier resistant unsupervised deep architectures for attributed network embedding. In *Proceedings of the 13th international conference on web search and data mining*, pp. 25–33, 2020a.

Sambaran Bandyopadhyay, Saley Vishal Vivek, and MN Murty. Outlier resistant unsupervised deep architectures for attributed network embedding. In *Proceedings of the 13th international conference on web search and data mining*, pp. 25–33, 2020b.

Aleksandar Bojchevski and Stephan Günnemann. Bayesian robust attributed graph clustering: Joint learning of partial anomalies and group structure. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Benoıt Corsini, Pierre-André Noël, David Vázquez, and Perouz Taslakian. Self-supervised anomaly detection in static attributed graphs.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 594–602. SIAM, 2019a.

Kaize Ding, Jundong Li, and Huan Liu. Interactive anomaly detection on attributed networks. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 357–365, 2019b.

Haoyi Fan, Fengbin Zhang, and Zuoyong Li. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5685–5689. IEEE, 2020.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

Leonardo Gutiérrez-Gómez, Alexandre Bovet, and Jean-Charles Delvenne. Multi-scale anomaly detection on attributed networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 678–685, 2020.

Tianjin Huang, Yulong Pei, Vlado Menkovski, and Mykola Pechenizkiy. Hop-count based self-supervised anomaly detection on attributed networks. *arXiv preprint arXiv:2104.07917*, 2021.

Ming Jin, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. Anemone: graph anomaly detection with multi-scale contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3122–3126, 2021.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, 2016.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Yuening Li, Xiao Huang, Jundong Li, Mengnan Du, and Na Zou. Specae: Spectral autoencoder for anomaly detection in attributed networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 2233–2236, 2019.

Kay Liu, Yingtong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, Lichao Sun, Jundong Li, George H. Chen, Zhihao Jia, and Philip S. Yu. Benchmarking node outlier detection on graphs. *arXiv preprint arXiv:2206.10071*, 2022a.

Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6):2378–2392, 2021.

Zhiyuan Liu, Chunjie Cao, and Jingzhang Sun. Mul-gad: a semi-supervised graph anomaly detection framework via aggregating multi-view information. *arXiv preprint arXiv:2212.05478*, 2022b.

Xuexiong Luo, Jia Wu, Amin Beheshti, Jian Yang, Xiankun Zhang, Yuan Wang, and Shan Xue. Comga: Community-aware attributed graph anomaly detection. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 657–665, 2022.

Peter Morales, Rajmonda Sulo Caceres, and Tina Eliassi-Rad. Selective network discovery via deep reinforcement learning on embedded spaces. *Applied Network Science*, 6:1–20, 2021.

Zhen Peng, Minnan Luo, Jundong Li, Luguo Xue, and Qinghua Zheng. A deep multi-view framework for anomaly detection on attributed networks. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2539–2552, 2020.

Bryan Perozzi, Leman Akoglu, Patricia Iglesias Sánchez, and Emmanuel 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*, pp. 1346–1355, 2014a.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014b.

Xiang-Rong Sheng, De-Chuan Zhan, Su Lu, and Yuan Jiang. Multi-view anomaly detection: Neighborhood in locality matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4894–4901, 2019.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pp. 1067–1077, 2015.

Mingyue Tang, Carl Yang, and Pan Li. Graph auto-encoder via neighborhood wasserstein reconstruction. *arXiv preprint arXiv:2202.09025*, 2022.

Jianyu Wang, Rui Wen, Chunming Wu, Yu Huang, and Jian Xiong. Fdgars: Fraudster detection via graph convolutional networks in online app review system. In *Companion proceedings of the 2019 World Wide Web conference*, pp. 310–316, 2019.

Jia Wu, Xingquan Zhu, Chengqi Zhang, and Zhihua Cai. Multi-instance multi-graph dual embedding learning. In *2013 IEEE 13th International Conference on Data Mining*, pp. 827–836. IEEE, 2013.

Jia Wu, Shirui Pan, Xingquan Zhu, and Zhihua Cai. Boosting for multi-graph classification. *IEEE transactions on cybernetics*, 45(3):416–429, 2014.

Zhiming Xu, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 444–457. Springer, 2022.

Shijie Zhang, Hongzhi Yin, Tong Chen, Quoc Viet Nguyen Hung, Zi Huang, and Lizhen Cui. Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 689–698, 2020.

Shuang Zhou, Xiao Huang, Ninghao Liu, Fu-Lai Chung, and Long-Kai Huang. Improving generalizability of graph anomaly detection models via data augmentation. *arXiv preprint arXiv:2209.10168*, 2022.

Mengxiao Zhu and Haogang Zhu. Mixedad: A scalable algorithm for detecting mixed anomalies in attributed graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1274–1281, 2020.