Detecting Anomalies on Cryptocurrency Markets Using Graph Algorithms

Agata Skorupka
Collegium of Economic Analysis, Warsaw School of Economics, Warsaw, Poland

Keywords: Graph Embeddings, Anomaly, Anomaly Detection, Cryptocurrency.

Abstract: The low level regulation of cryptocurrency market as well as crucial role of trust and digital market specificity makes it a good environment for anonymous transactions without identity verification, therefore fraudulent activities. Examples of such anomalies may be failing to fulfil transaction, as well as different forms of market manipulation. As cryptocurrencies are incorporated in more and more investment portfolios, including big companies accepting payment by this means, anomalies on cryptocurrency may pose significant systemic risk. Therefore there is a need to detect fraudulent users in a computationally efficient way. This paper presents usage of graph algorithms for that purpose. While most of the literature is focused on using structural and classical embeddings, this research proposes utilizing nodes statistics to build an accurate model with less engineering overhead as well as computational time involved.

1 INTRODUCTION

In 2019 global economic cost associated with fraudulent activities was estimated around $5.12 trillion (Gee, 2019). The emergence of unregulated markets, such as cryptocurrency exchanges, often perceived as a "lawless territory", where one can perform activities which will be illegal anywhere else (Félez-Viñas, 2022), has elevated that number even higher (to the extent which is not always possible to determine due to the diffused nature of cryptocurrencies). Examples of such activities may include black-market trading (Foley, 2019), money laundering and terrorist financing (Fletcher, 2021), insider trading (Fratrič, 2022), market manipulations such as wash trading (Cong, 2020) and pump-and-dump schemes (Kamps, 2018) or just improper performance of obligations stemming from the sales contract (Kumar, 2018).

According to Félez-Viñas (2022), insider trading is estimated to occur in even 25% of listings at the largest cryptocurrency exchange in the US - Coinbase. Despite general lack of regulation of the cryptocurrency market, US Securities and Exchange Commission (SEC) already commenced the first prosecution regarding insider trading on that market, estimating the value of illegal profits for more than $1.1 million in that case (Fratrič, 2022). Apart from insider trading, one can also enumerate other types of fraudulent activities contributing to the global economic cost, such as market manipulation - especially wash sales (Victor & Weintraud, 2021) and pump and dump (Chen, 2019) are common practices on cryptocurrencies market. Both are coordinated actions to artificially increase the market price (or mislead investors in a different manner) in the short run, but in wash sales, both seller and buyer is the same market actor (Hamrick, 2019; Li, 2018). Besides monetary cost, there also exists significant systemic risk to the financial sector and to the entire economy associated with cryptocurrency volatility (Fratrič, 2022). Initially serving as a medium of exchange or a niche asset for a relatively small number of market actors, now cryptocurrencies are incorporated in more and more investment portfolios, including big companies accepting payment by this means (Robleh, 2014). Naturally, this kind of cost is more difficult to measure directly and even estimate. Risk modeling in portfolios has been approached in recent studies by methods such as multi-objective feature selection (Kou, 2021), clustering (Li, 2021) and network analysis (Anagnostou, 2018).

Taking the above mentioned into account, there exists a need to detect fraudulent activities in both computationally effective and relatively fast way. Traditionally, fraudulent activities on financial markets were examined by the regulatory organ

Skorupka, A.
Detecting Anomalies on Cryptocurrency Markets Using Graph Algorithms.
DOI: 10.5220/0012133900003541
In *Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023)*, pages 504-509
ISBN: 978-989-758-664-4; ISSN: 2184-285X

manually in an individual (case by case) manner. This approach requires gathering official documents, transaction reports, interviewing witnesses and therefore is time consuming (Dhanalakshmi, 2019). As an example here may serve SEC investigation guidelines.

With the advent of digitalisation and big data, as well as developments in software and computing power, machine learning techniques gained more popularity for the purpose of detecting fraudulent activities on financial markets (Kou, 2004; Nagi, 2011). In particular, neural networks and SVM algorithms for outlier detection were used (Ogut, 2009), hidden Markov chains (Song, 2012), analysis and Bayesian techniques regarding updating beliefs (Holton, 2009). What is worth bearing in mind, not all types of digital markets have text data just as information markets, hence such methods will not be always appropriate for analysis. Dhanalakshmi and Subramanian (2014) proposed usage of the clustering method, while Golmohammadi (2014) conducted an in-depth survey using methods such as decision trees, k-nearest neighbors analysis and Bayesian methods.

Nevertheless, usage of classical machine learning techniques has been recently criticized in the literature for not taking into account the complexity of the structure of the financial market and conducting analyses (Liu, 2019). For that purpose, graph methods were proposed (Tamersoy, 2016; Rayes and Mani, 2019). Although recent literature on anomaly detection using graphs is developing at a fast pace, it focuses mostly on anomalies in citation networks, product networks (fake reviews), not financial markets or social networks (Zhao, 2019; Liu, 2021; Zhang, 2022; Wang, 2021).

On the other hand, fraud detection is a relatively new topic in cryptocurrency research (Victor & Weintraud, 2021; Chen, 2019) and most of research focuses on simulating effects of fraudulent activities using agent-based modeling (Luther, 2013; Bornhold, 2014; Cocco; 2017 and 2019, Pyromallis, 2018; Zhou, 2017; Shibano, 2020; Bartolucci, 2020; Fratrič, 2022) or classical machine learning methods, as Random Forests (Baek, 2019) or Support Vector Machines (Sayadi, 2019).

2 RESEARCH MOTIVATION AND GOALS

The following paper aims to cover this gap focusing on graph anomaly detection methods, as well as mitigate another challenge often raised when detecting anomalies using graph data: lack of extensive dataset with included labels (ground truth). For that reason, mostly unsupervised techniques were developed (Zhao, 2019; Liu, 2021; Zhang, 2022; Wang, 2021). These have significant drawbacks, i.e. injecting synthetic fraudulent users according to the definition of developed algorithm, as in Liu (2021). In that way, authors ensure that their algorithm will outperform others, as it was designed specifically for that problem.

Classical anomaly detection algorithms were based on the network characteristics, however training models can be as good as data provided. Cryptocurrencies networks do not gather as much data about users as e.g. social networks. On the other hand, graph data is relatively easy to obtain even in case of sparsity of users' characteristics, as long as the network can be represented as a graph, which is the case with cryptocurrency market: users are represented by nodes, whereas transactions by edges. First category of graph features relatively easy to obtain is to compute nodes' statistics, such as number of neighbors (centrality) as well as a variety of importance measures (centralities). These features will be further referred to also as "graph features". On the other hand, state of the art in the literature is to use embedding algorithms, which, using deep learning techniques map nodes to the vector space. The idea behind it is to keep similar nodes close to each other in vector space. There are two ways of interpreting similarity: being neighbors of each other (node or classical embedding) and having an equivalent type of neighborhood (structural embedding). It is a common consensus in the literature that this type of sophisticated, deep learning based algorithms is a better predictor than simple node statistics. On the other hand, graph embeddings are not always computationally efficient, which is of a special importance with the constant increase of users in underlying graph networks representing markets and social networks. Furthermore, embeddings require careful choice of type of embedding as well as embedding dimension.

The aim of this paper is to contribute to the literature on anomaly detection on the cryptocurrency market in order to detect fraudulent transactions in an accurate and computationally efficient way. Especially the latter is of a particular importance given the ever-growing number of market actors and transactions performed. The following research aims to propose a computationally efficient graph algorithm for anomaly detection based on node statistics and to test hypotheses if the proposed

algorithm can outperform state-of-the-art approaches based on graph embeddings.

3 DATASETS

The following research examines anomaly detection algorithms on two datasets representing Bitcoin transaction markets: Bitcoin OTC [1] and Bitcoin Alpha [2]. These networks can be represented as directed graphs, with nodes denoting users and edges denoting transactions between them. Weights of edges are representing rating by a particular user, which can happen only after a transaction and can take values from -10 (full distrust) to 10 (full trust). The ratings were rescaled between -1 and 1. Benign users were determined in the following way: platform founders, as well as users rated positively (at least 0.5 after rescaling) by them. Fraudulent users, also referred to as anomalies, were considered as those who were rated negatively (at most -0.5 after rescaling) by a benign user group. Table 1 represents statistics of both datasets.

Table 1: Statistics of Bitcoin datasets.

	Bitcoin OTC	Bitcoin Alpha
Number of nodes	5881	3783
Number of edges	35592	2418
Average degree	12	13
Minimum degree	1	1
Maximum degree	1298	888
Number of components	4	5
Size of the largest component	5875	3775
Number of isolated nodes	0	0

One advantage of these datasets is that the graph represents the whole network, as sometimes it may be hard to obtain one and therefore graph sampling methods are used (Stella, 2019; Feng, 2021; Dehghan, 2022). This procedure can negatively influence accuracy of results. Another value Bitcoin datasets are providing is the existence of dataset labels based on objective criterion, rather than manual annotation by human using expert knowledge (Stella, 2019; Feng, 2021), which is also a significant factor able to influence model performance and possibility of generalizations.

[1] https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html

4 METHODS AND RESULTS

On the basis of two Bitcoin datasets the following models detecting fraudulent users were built: one group using embeddings and second one using node statistics such as degree centrality, harmonic centrality, pagerank, closeness and betweenness centrality, as well as local clustering coefficient. For the first group of models, following embeddings were used - each with its own model: two classical (node2vec and DeepWalk) and two structural (RolX and Struc2vec). Following dimensions of embeddings were used: 4, 8, 16, 32, 64, 128 in order to determine the best performing dimension. Furthermore, all embeddings having dimension above 64 were additionally compressed using PCA and UMAP algorithms in order to examine if noise reduction can help in model performance, or, in case of UMAP, taking non-linearity into account. In the case of UMAP, three versions were prepared with three different seeds to ensure that the algorithm is stable.

Models for two dataset features were chosen using AUC metrics among Random Forest, XGBoost and Generalized Linear Model as well as two ensemble models: one built on the top of all models and second built on the top of best models of their own class using AUC metrics. AutoML parameter tuning with 5-fold cross-validation on the test dataset was used. Then, the best model built on the top of embeddings was compared with the best model built using nodes' statistics using F-1 metrics.

All models were built using h2o and xgboost python libraries. Results were presented in Tables 2 and 3. For brevity purposes, only ten best models were shown.

There are following conclusions from the comparison of anomaly detection graph algorithms: first of all, nodes' statistics perform almost as good as the best model based on embeddings. In the case of the Bitcoin Alpha dataset, h2o model based on nodes' statistics achieved a 0.83 F-1 score compared to 0.86 in the case of h2o model based on Struc2vec with 32 dimensions. In the case of Bitcoin OTC, h2o model built on the top of nodes' statistics achieved 0.91 F-1 score outperforming h2o RolX of dimension 128, compressed to 16 using PCA. That means that similar results can be achieved by far less computational and engineering time. The latter refers to the choice of type of embedding, as well as its dimension. Each machine learning task requires a specific choice of

[2] https://snap.stanford.edu/data/soc-sign-bitcoin-alpha.html

Table 2: Comparison of anomaly detection model performance on Bitcoin Alpha network.

Rank	Embedding	Library	F1	Accuracy	MCC	Dimension	Compression	Original dimension
1	struc2vec (32)	h2o	0.857	0.833	0.679	32	no	NA
2	nodes' statistics	h2o	0.829	0.805	0.615	NA	no	NA
3	struc2vec (16)	h2o	0.827	0.805	0.611	16	no	NA
4	struc2vec (64)	h2o	0.825	0.805	0.611	64	no	NA
5	nodes' statistics	xgboost	0.825	0.805	0.611	0	no	NA
6	struc2vec (128)	h2o	0.825	0.805	0.611	128	no	NA
7	struc2vec (8)	h2o	0.820	0.805	0.609	8	no	NA
8	rolx (128)	h2o	0.820	0.805	0.610	128	no	NA
9	struc2vec (4)	h2o	0.818	0.777	0.580	4	no	NA
10	rolx (8)	h2o	0.814	0.792	0.585	8	no	NA

Table 3: Comparison of anomaly detection model performance on Bitcoin OTC network.

Rank	Embedding	Library	F1	Accuracy	MCC	Dimension	Compression	Original dimension
1	nodes' statistics	h2o	0.913	0.916	0.832	NA	no	NA
2	rolx (128 to 16), PCA	h2o	0.886	0.894	0.789	16	PCA	128
3	struc2vec (128)	h2o	0.878	0.873	0.759	128	no	NA
4	struc2vec (32)	h2o	0.872	0.873	0.750	32	no	NA
5	struc2vec (8)	h2o	0.869	0.873	0.747	8	no	NA
6	rolx (128 to 16), UMAP 1	h2o	0.860	0.863	0.728	16	UMAP	128
7	rolx (128 to 16), UMAP 2	h2o	0.860	0.863	0.728	16	UMAP	128
8	rolx (32 to 16), PCA	h2o	0.860	0.863	0.728	16	PCA	32
9	rolx (128)	h2o	0.857	0.852	0.717	128	no	NA
10	rolx (128 to 16), UMAP 0	h2o	0.857	0.863	0.726	16	UMAP	128

embeddings. Nevertheless, there is no specific principles or rule of thumb how to choose it, so either engineer choose embedding arbitrarily with a low chance of outperforming nodes' statistics algorithm, either will they build number of embedding types in different dimensional variants, which is very time consuming, especially when size of graph is significant.

Secondly, regarding types of embeddings, structural embeddings are performing significantly better than classical ones. There is not even one classical embedding in top ten models in the case of both Bitcoin Alpha and Bitcoin OTC. It is worth to note that with the first dataset the advantage of Struc2vec among others is prevalent, however it does not happen with Bitcoin OTC, as we can see both RolX and Struc2vec among top ten models. Another

Table 4: Comparison of the model performance of the best model of the given class embedding on the Bitcoin Alpha market.

Embedding	Dimension	Library	F1	Accuracy	MCC	Compression	Original dimension
struc2vec	32	h2o	0.857	0.833	0.679	no	NA
nodes statistics	NA	h2o	0.829	0.806	0.615	no	NA
rolx	128	h2o	0.821	0.806	0.610	no	NA
node2vec	8	h2o	0.740	0.653	0.347	no	NA
deepwalk	128	h2o	0.712	0.597	0.227	no	NA

Table 5: Comparison of the model performance of the best model of the given class embedding on the Bitcoin OTC market.

Embedding	Dimension	Library	F1	Accuracy	MCC	Compression	Original dimension
nodes statistics	NA	h2o	0.913	0.916	0.832	no	NA
rolx	16 from 128 (PCA)	h2o	0.886	0.895	0.789	PCA	128
struc2vec	128	h2o	0.878	0.874	0.760	no	NA
node2vec	16 from 64 (UMAP)	h2o	0.804	0.780	0.613	UMAP	64
deepwalk	16	h2o	0.727	0.726	0.453	no	NA

conclusion is that it is difficult to determine if dimensionality reduction help, as in the case of Bitcoin Alpha no compressed features found themselves in the top ten models, whereas in the case of Bitcoin OTC half of the best embedding models were characterized by compression (both UMAP and PCA). This means compression is task and dataset specific and adds engineering overhead to the model building. Tables 4 and 5 are presenting comparison of the model performance of the best model in the case of given class embedding, on the Bitcoin Alpha and Bitcoin OTC markets respectively.

5 CONCLUSIONS

In this work, graph anomaly detection methods were examined on the basis of cryptocurrency markets: two over-the-counter Bitcoin markets, namely Bitcoin OTC and Bitcoin Alpha. It was determined that although state-of-the-art embeddings have strong predictive power, they are often computationally inefficient, especially in the case of large graphs. Furthermore, they require case-by-case choice of type of embedding, as well as its dimension. Sometimes there is a need to determine if to use dimensionality reduction techniques, which adds engineering overhead and can be even more time consuming, as the choice of the embedding can either be random or informed after building a number of embeddings. On the other hand, anomaly detection models based on nodes' statistics turned out to be almost as good as the best model among embedding-based, while providing simplicity and computational efficiency. Presenting results on the two datasets show that results can be generalized, however, there is a need to extend the research on other datasets for further check of results stability. Another interesting research direction in the future is to build specific algorithms using nodes' statistics, e.g. involving dimensionality reduction or statistics for node neighbors.

REFERENCES

Gee, J., & Button, M. (2019). The financial cost of fraud. Retrieved from http://www.crowe.ie/wp-content/uploads/2019/08/The-Financial-Cost-of-Fraud-2019.pdf

Félez-Viñas, E., Johnson, L., & Putniņš, T. J. (2022). Insider Trading in Cryptocurrency Markets. *Available at SSRN 4184367*..

Fletcher E., Larkin C., Corbet S. (2021) Countering money laundering and terrorist financing: a case for Bitcoin regulation. Res Int Bus Finance 56(January):101387. https://doi.org/10.1016/j.ribaf.2021.101387

Kamps J., Kleinberg B. (2018) To the moon: defining and detecting cryptocurrency pump-and-dumps. Crime Sci 7(1):1–18. https://doi.org/10.1186/s40163-018-0093-5

Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., & Subrahmanian, V. S. (2018, February). Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 333-341).

Fratrič, P., Sileno, G., Klous, S., & van Engers, T. (2022). Manipulation of the Bitcoin market: an agent-based study. *Financial Innovation*, *8*(1), 1-29.

Hamrick J., Rouhi F., Mukherjee A., Feder A., Gandal N., Moore T., Vasek M. (2019) The economics of cryptocurrency pump and dump schemes. SSRN Electron J. https://doi.org/10.2139/ssrn.3303365

Chen W., Xu Y., Zheng Z., Zhou Y., Yang J. E., Bian J. (2019) Detecting 'Pump & dump schemes' on cryptocurrency market using an improved a priori algorithm. In: Proceedings—13th IEEE international conference on service-oriented system engineering, SOSE 2019, 10th international workshop on joint cloud computing, JCC 2019 and 2019 IEEE international workshop on cloud computing in robotic systems, CCRS 2019, pp 293–298. https://doi.org/10.1109/SOSE.2019.00050

Victor F., Weintraud A. M. (2021) Detecting and quantifying wash trading on decentralized cryptocurrency exchanges. In:The web conference 2021—proceedings of the world wide web conference, WWW 2021 2, pp 23–32. https://doi.org/10.1145/3442381.3449824. arXiv:2102.07001

Hamrick J., Rouhi F., Mukherjee A., Feder A., Gandal N., Moore T., Vasek M. (2019) The economics of cryptocurrency pump and dump schemes. SSRN Electron J. https://doi.org/10.2139/ssrn.3303365

Li T., Shin D., Wang B. (2018) Cryptocurrency pump-and-dump schemes. SSRN Electron J. https://doi.org/10.2139/ssrn.3267041

Robleh A., Barrdear, J., Clews, R., Southgate, J. (2014) The economics of digital currencies. Bank Engl Q Bull 2014 Q3(1):276–286.

Kou G., Xu Y., Peng Y., Shen F., Chen Y., Chang K., Kou S. (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Supp Syst 140:113429. https://doi.org/10.1016/j.dss.2020.113429

Anagnostou I., Sourabh S., Kandhai D. (2018) Incorporating contagion in portfolio credit risk models using network theory. Complexity 2018:6076173. https://doi.org/10.1155/2018/6076173

Li T., Kou G., Peng Y., Yu P. S. (2021) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3109066

Dhanalakshmi, S., & Subramanian, C. (2014). An analysis of data mining applications for fraud detection in securities market. International Journal of Data Mining Techniques and Applications, 3(1), 9–1. doi:10.20894/IJDMTA.102.003.001.003

Öğüt, H., Doganay, M., & Aktas, R. (2009). Detecting stock-price manipulation in an emerging market: The case of Turkey. Expert Systems with Applications, 36(9), 11944–11949. doi:10.1016/j.eswa.2009.03.065

Tamersoy, A. (2016). Graph-based algorithms and models for security, healthcare, and finance [Unpublished Doctoral dissertation]. Georgia Institute of Technology.

Rayes, J., & Mani, P. (2019). Exploring Insider Trading Within Hypernetworks. In P. Haber, T. Lampoltshammer, & M. Mayr (Eds.), Data Science – Analytics and Applications. Springer. doi:10.1007/978-3-658-27495-5_1

Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. Computers in human behavior, 24(5), 1816-1836.

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*.

Liu, Y., Li, Z., Pan, S., Gong, C., Zhou, C., & Karypis, G. (2021). Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, *33*(6), 2378-2392.

Dehghan, A., Siuta, K., Skorupka, A., Dubey, A., Betlen, A., Miller, D., Xu, W., Kaminski, B., and Pralat, P. Detecting Bots in Social-Networks Using Node and Structural Embeddings, Unpublished, 2022.

Zhang, F., Fan, H., Wang, R., Li, Z., & Liang, T. (2022). Deep Dual Support Vector Data description for anomaly detection on attributed networks. *International Journal of Intelligent Systems*, *37*(2), 1509-1528.

Wang, G., Xie, S., Liu, B., and Philip, S. Y.. Review graph based online store review spammer detection. In IEEE International Conference on Data Mining series, 2011.

Luther W. J. (2013) Crypto-currencies, network effects, and switching costs. SSRN Electron J. https://doi.org/10.2139/ssrn.2295134

Cocco L., Concas G., Marchesi M. (2017) Using an artificial financial market for studying a cryptocurrency market. J Econ Interact Coord 12(2):345–365. https://doi.org/10.1007/s11403-015-0168-2. arXiv:1406.6496

Pyromallis C., Szabo C. (2019) Modelling and analysis of adaptability and emergent behavior in a cryptocurrency market. In: 2019 IEEE Symposium series on computational intelligence, SSCI 2019, pp 284–292. https://doi.org/10.1109/SSCI44817.2019.9002829

Shibano K, Lin R, Mogi G (2020) Volatility reducing effect by introducing a price stabilization agent on cryptocurrencies trading. In: ACM International conference proceeding series, pp 85–89. https://doi.org/10.1145/3390566.3391679

Bartolucci S, Caccioli F, Vivo P (2020) A percolation model for the emergence of the Bitcoin Lightning Network. Sci Rep 10(1):1–14. https://doi.org/10.1038/s41598-020-61137-5

Kumar, S., Spezzano, F., Subrahmanian, V. S., & Faloutsos, C. (2016, December). Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (pp. 221-230). IEEE.

Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., & Subrahmanian, V. S. (2018, February). Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 333-341).