Graph-theoretical Approach to Enhance Accuracy of Financial Fraud Detection Using Synthetic Tabular Data Generation

Dae-Young Park
School of Computing, Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea
AI Technology Team, Financial Security Institute (FSI)
Seoul, Republic of Korea
mainthread@kaist.ac.kr

ABSTRACT

Tabular data synthesis has become crucial for financial applications including fraud detection, especially where there are data privacy regulations such as General Data Protection Regulation (GDPR) restrict access to original data. Despite its importance, current generative models inadequately address key challenges in financial fraud detection (FFD) data, namely extreme class imbalance, high data sparsity, and non-normal attribute distributions. My research introduces novel graph-theoretical generative models, *SeparateGGM* and *SignedGGM*, designed to tackle these challenges. By integrating graph neural network-based feature engineering, graph topology and connectivity analysis, and novel graph centrality indicators, my models achieve optimal graph settings for enhanced fraud detection accuracy. This approach is pioneering in its application of diverse graph-theoretical methods to improve FFD performance. Preliminary results demonstrate my models' superiority over competing methods on multiple FFD benchmark datasets. The goal of this research is to significantly advance real-world financial fraud detection techniques and to show that several graph-theoretical methodologies can significantly contribute to the generation of high-quality tabular synthetic data for enhancing fraud detection accuracy to the data science community [1].

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Mathematics of computing → Graph theory**; • **Information systems → Data mining**.

KEYWORDS

graph-theoretical methodology, financial fraud detection, generative model, synthetic tabular data

ACM Reference Format:
Dae-Young Park. 2024. Graph-theoretical Approach to Enhance Accuracy of Financial Fraud Detection Using Synthetic Tabular Data Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3627673.3680267

1 INTRODUCTION AND PROBLEM

Generating synthetic tabular data is especially significant in financial fraud detection (FFD), where data sharing restrictions among companies and limited available original data necessitate the use of synthetic data for training models [2] [3]. The field has evolved from traditional statistical methods to advanced deep generative models, enhancing the utility of synthetic data for downstream tasks.

However, despite their success in other applications, current generative models struggle with the unique challenges of FFD data, which include extreme class imbalance, high data sparsity, and a large number of attributes with non-normal distributions. I confirmed that these characteristics significantly degrade the performance of existing models. As shown in Table 1, by demonstrating these challenges stem from FFD data's characteristics through a version of FFD data maintained by my institute [4], I confirmed that the performance of existing generative models deteriorates as the intensity of these characteristics increases. The existing works did not show a detailed evaluation of detection performance while considering the three characteristics of FFD data.

To address these challenges, I propose the novel *graph-theoretical generative models*, named *SeparateGGM* and *SignedGGM*, for tabular synthetic data by jointly leveraging several graph-theoretical methodologies, including a graph neural network-based feature engineering, graph topology and connectivity analysis, and new graph centrality indicators that I developed for this approach. The approach consists of several key steps. Specifically, a Graph Neural Network is first utilized to augment the features of the original data. Subsequently, based on similarities and class relationships between data instances, separate and signed directed K-NN graphs are constructed to create a candidate set of graphs. Next, graph topology and connectivity are analyzed to select effective K-NN graphs with the optimal number of positive and negative K in terms of detection accuracy. Through this graph analysis, it has been observed that certain positive and negative K ratios, which make several graph measurements approximately the highest points, consistently yield the highest detection accuracy of target FFD models, thereby facilitating the effective K-NN graphs selection with the optimal number of hyperparameters (negative and positive Ks). Thus, the graph analysis enables my approach to be essentially hyperparameter-free.

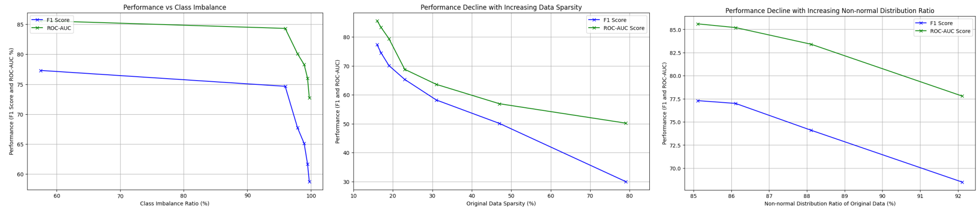
[1] This research is supervised by In-Young Ko (iko@kaist.ac.kr)

[2] https://www.statice.ai/case-study/provinzial-predictive-analytics-synthetic-insurance-data

[3] https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data/payments-data-for-fraud-detection

[4] https://www.fsec.or.kr

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680267



(a) Class imbalance **(b) Data sparsity** **(c) Non-normal distribution**

Figure 1: The three characteristics intensity in the original FFD data and the changes in detection accuracy of a target FFD model trained on the generated tabular synthetic data

Then, for the separate graphs, graph centrality indicators of data instances within the selected graphs are calculated by extending existing graph centrality indicators, called *C* and *CC* [8]. Next, for the signed graph, I develop new graph centrality indicators, named *Ambivalent Centeredness* and *Center-Closeness* (*AC* and *ACC*) to calculate graph centrality. The indicators can measure the degree to which a data instance is considered central and close to other centers, taking into account both positive and negative class relationships. Next, in the separate and signed graphs, the average values of centrality indicators are computed for a base generative model to pay attention to more important training data respectively.

The main contributions of my research is as follows. First, to handle the three challenges of FFD data, I propose a novel graph-theoretical method for tabular synthetic data generation, which jointly employs graph-based feature engineering, analysis of graph measurements, and new graph centrality indicators. Second, this study presents the first application of graph theory in the domain of tabular synthetic data generation. Experimental results demonstrate that several graph-theoretical methodologies can significantly contribute to the generation of high-quality tabular synthetic data, thereby enhancing fraud detection accuracy.

2 STATE OF THE ART

There are several works related to financial fraud detection. Research on FFD has been a longstanding topic of interest, aimed at protecting financial consumers and markets [6]. Especially, recent studies have focused on enhancing the effectiveness of fraud detection models by leveraging synthetic data where the use of synthetic data not only boosts the FFD models' accuracy but also lowers the risk of exposing personal information [1]. However, there is a notable scarcity of research on tabular synthetic data generation that directly handle the three characteristics of FFD data.

There are several works related to generative models for tabular synthetic data. Recently, there has been an attempt to propose tabular data synthesis using deep generative models because they show great performance of downstream tasks [2]. Generative Adversarial Network (GAN)-based tabular data synthesis models have mostly been proposed among the related works. For example, TableGAN is one of the most influential academic works to mark a milestone in the field of tabular data synthesis [10]. Next, CTGAN demonstrates notable efficacy, primarily attributed to its utilization of mode-specific preprocessing to handle the multi-modality of each features [12]. Furthermore, there are works based on other types of deep learning models. For example, TabDDPM is based on a diffusion probability model, which has recently shown successful image generation [4]. DPHFlow is based on a flow-based model that reduces data loss through inverse function transformation while adjusting the level of information protection [5]. Since there have been few generative models tailored for real-world FFD data, my research distinguishes itself from the existing works by conducting a detailed evaluation of fraud detection accuracy from the perspective of the three characteristics of FFD data.

3 PROPOSED APPROACH

My approach consists of several steps as shown in Figure 2.

3.1 Graph-based feature augmentation

I adopt GraphSAGE, a popular GNN, to augment the feature space because the process of feature augmentation can serve as a pivotal mechanism to enhance the intrinsic structure of tabular data in terms of consideration of pair-wise correlation. This is achieved by transforming the original data into a graph-based representation, wherein nodes encapsulate data instances and edges embody the relational dependencies among these instances.

3.2 Graph candidate set creation

Separate directed K-NN Graphs (G^+ and G^-) and signed directed K-NN graph G^s are created based on the similarities between each data instance (i.e., node) by adjusting hyperparameter values, positive K (indicating the number of positive directed edge for each node) and negative K (indicating the number of negative directed edge for each node). The values of positive K range from 2 to (the number of nodes with the same class - 1) and the values of negative K ranges from 2 to (the number of nodes with the different class - 1). The similarities between all possible pairs of the data instances are computed by the inverse of the Minkowski distance (IMD).

3.3 Graph topology and connectivity analysis

I validate how to determine the optimal number of positive and negative K (in terms of detection accuracy) for selecting effective graph settings via graph topology and connectivity analysis [7]. Specifically, graph topology and connectivity are analyzed by quantifying key graph measurements including (a) diameter, (b) average path length (APL), and (c) the number of reachable pair for directed graphs, (d) reciprocity, and (e) the number of strongly connected components (SCC) for directed graphs.

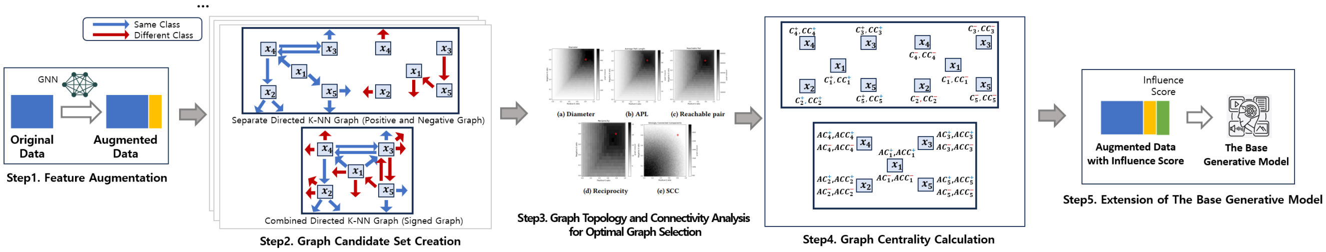

Same Class
Different Class
GNN
Original Data
Augmented Data
Step1. Feature Augmentation
Separate Directed K-NN Graph (Positive and Negative Graph)
Combined Directed K-NN Graph (Signed Graph)
Step2. Graph Candidate Set Creation
(a) Diameter
(b) APL
(c) Reachable pair
(d) Reciprocity
(e) SCC
Step3. Graph Topology and Connectivity Analysis for Optimal Graph Selection
Step4. Graph Centrality Calculation
Influence Score
Augmented Data with Influence Score
The Base Generative Model
Step5. Extension of The Base Generative Model

Figure 2: The overall process of my proposed approach

Specifically, to determine the optimal graph based on a objective (i.e., maximizing detection performance), I conduct graph-theoretical analysis (in terms of diameter, average path length, and reachable pair). Through the banking FFD data in our institute, I observe that certain positive and negative K ratios, which make the three graph measurements the highest points, consistently yield the highest detection accuracy, facilitating efficient graph selection with the optimal number of Ks. It is observed that when positive and negative K ratios are about 80% (between 79.18 and 80.54), graph measurements including (a) diameter, (b) APL, (c) reachable pair, and (d) reciprocity (except for (e) SCC) reach approximately the highest points (within the maximum range) where the highest detection accuracy of the target models is shown. This can also be interpreted that these four highest graph measurements, which correspond to the optimal numbers of positive and negative K (optimal hyperparameters for detection accuracy), thus enable the generation of high-quality synthetic data that shows the highest detection accuracy of the target detection model [5].

3.4 Graph centrality calculation

3.4.1 Separate directed K-NN graphs. Similar to a previous paper [8], the graph centrality indicators for G^+ and G^- can be represented as a linear algebra form: $\mathbf{cc}^{+(i+1)} = \kappa(\mathbf{W}^+\mathbf{W}^{+T})^{i+1}\mathbf{cc}^{+(0)}$, $\mathbf{c}^{+(i+1)} = \kappa(\mathbf{W}^{+T}\mathbf{W}^+)^i\mathbf{W}^{+T}\mathbf{cc}^{+(0)}$, $\mathbf{cc}^{-(i+1)} = \kappa(\mathbf{W}^-\mathbf{W}^{-T})^{i+1}\mathbf{cc}^{-(0)}$, and $\mathbf{c}^{-(i+1)} = \kappa(\mathbf{W}^{-T}\mathbf{W}^-)^i\mathbf{W}^{-T}\mathbf{cc}^{-(0)}$, where $\mathbf{c}^+$ and $\mathbf{cc}^+$ are the vector forms of centrality indicators in a positive graph G^+, $\mathbf{c}^-$ and $\mathbf{cc}^-$ are the vector forms of centrality indicators in a negative graph G^-, $\mathbf{W}^+$ and $\mathbf{W}^-$ is are edge weight matrices for positive and negative graphs respectively, $\mathbf{W}^+, \mathbf{W}^- \in \mathbb{R}^{n\times n}$, $\mathbf{c}^+, \mathbf{cc}^+, \mathbf{c}^-, \mathbf{cc}^- \in \mathbb{R}^n$. κ is a normalization factor such that $\|\mathbf{cc}^+\| = \|\mathbf{c}^+\| = \|\mathbf{cc}^-\| = \|\mathbf{c}^-\| = 1$. Next, in the separated directed graph G^+ and G^-, The final C and CC of data instances are calculated by taking the absolute difference between positive and negative C and CC.

3.4.2 Signed directed K-NN graph. I design AC and ACC to reflect the adaptation of the existing indicators for the signed directed graph G^s based on Heider's balance theory, aiming to stabilize cognitive balance within the signed graph G^s. Specifically, I can represent AC and ACC as a linear algebra form to generalize the indicators for G^s. First, I need to explain several symbols related to W to vectorize the four equations: Let D be the out-degree diagonal matrix of $|W|$ (i.e., $D_{ii} = \sum^j |W|_{ij}$). Then, $\tilde{W} = D^{-1}W$. Next, $\tilde{W}$ can be decomposed into positive matrix and negative matrix: $\tilde{W} = \tilde{W}_+ - \tilde{W}_-$, where $\tilde{W}_+$ includes values of positive elements in $\tilde{W}$ and $\tilde{W}_-$ includes absolute values of negative elements in $\tilde{W}$. Then, $\mathbf{ac}^+, \mathbf{acc}^+, \mathbf{ac}^-$, and $\mathbf{acc}^-$ can be generalized as follows: $acc^+ \leftarrow \tilde{W}_+ac^+ + \tilde{W}_-ac^-$, $ac^+ \leftarrow \tilde{W}_+^T acc^+ + \tilde{W}_-^T acc^-$, $acc^- \leftarrow \tilde{W}_+ac^- + \tilde{W}_-ac^+$, and $ac^- \leftarrow \tilde{W}_+^T acc^- + \tilde{W}_-^T acc^+$. the final AC and ACC of data instances are calculated by taking the absolute difference between positive and negative AC and ACC.

3.5 Extension of the base generative model

I define the average value of **C** and **CC** within separate directed graphs, and the average value of **AC** and **ACC** within a signed directed graph, as the *influence score* for each data instance respectively. In my base generative model, CTGAN [6], the integration of these scores into the critic loss significantly impacts the training process. This integration enhances the critic's capability to distinguish between essential and non-essential real and synthetic data. With respect to the generator loss, these scores guide the base model to concentrate on reproducing critical features, thereby facilitating the creation of synthetic data instances that closely align with representative real data instances.

4 RESEARCH METHODOLOGY AND RESULTS

I conduct comprehensive experiments across three FFD benchmark datasets to answer the following research questions (RQs):

1. **RQ1**: *How much performance does my method show compared to 7 baselines on popular benchmarks?*
2. **RQ2**: *How does my method perform when the three characteristics intensity of the FFD data increases?*
3. **RQ3**: *How much time does graph topology and connectivity analysis reduce hyperparameter search time in my method?*

4.1 Experimental settings

4.1.1 Datasets and Competing methods. I picked three commonly used FFD benchmark datasets from existing works: Ethereum transaction, online payment record, and credit card transaction datasets [2, 3, 9, 11, 13]. All three tabular benchmark datasets have a target class that indicates a fraud or normal label [7]. Next, my method is compared with seven baselines including popular and state-of-the-art models outlined in [9]. I adopt the models because each one is the leading method in different paradigm of generative model.

[5] For the graph analysis of signed graph G^s, same trend is consistently observed

[6] CTGAN is adopted as the base model in my approach since it shows the best accuracy

[7] I divide all of the datasets into training (80%) and test dataset (20%)

Table 1: Performance comparison of SeparateGGM , SignedGGM, and seven baselines in terms of fraud detection accuracy (%)

Methods	Ethereum				Online payments				Credit card			
	Macro-F1	Weighted-F1	ROC-AUC	PR-AUC	Macro-F1	Weighted-F1	ROC-AUC	PR-AUC	Macro-F1	Weighted-F1	ROC-AUC	PR-AUC
GC	61.129	72.378	72.611	74.714	57.112	64.345	68.678	69.122	62.412	67.501	70.278	71.789
CART	83.213	84.789	87.512	87.201	70.989	81.912	81.013	83.157	73.856	82.578	85.947	87.451
TVAE	84.579	88.902	87.134	89.015	71.456	79.997	83.547	85.989	76.036	84.311	89.998	89.567
TableGAN	85.312	88.512	86.497	87.844	70.978	79.956	83.279	85.867	76.123	85.678	89.112	89.365
CTGAN	85.987	88.312	88.594	89.101	72.689	80.314	84.578	86.712	76.582	86.412	90.567	91.002
DPHFlow	85.998	88.501	88.901	89.401	72.467	81.145	84.312	86.534	75.373	86.123	90.712	91.678
TabDDPM	86.771	86.903	89.123	90.312	73.412	82.567	85.729	87.978	77.118	87.011	90.789	92.123
SeparateGGM	88.101	**89.249**	90.878	91.312	**76.678**	84.969	87.001	88.412	**79.989**	88.567	91.901	93.112
SignedGGM	**88.112**	89.245	**90.944**	**91.678**	76.189	**84.991**	**87.516**	**88.795**	79.901	**88.683**	**92.001**	**93.615**

4.2 Preliminary results

*4.2.1 Performance comparison **(RQ1)**.* As shown in Table 1, it is clear that my method outperforms the best competing methods over all metrics across all FFD benchmark datasets [8].

*4.2.2 Performance change on three characteristics intensity of FFD data **(RQ2)**.* I compare the degree of performance degradation between my method (i.e., SignedGGM) and competing methods by adjusting the intensity of the three characteristics across three FFD benchmark datasets. I observe SignedGGM consistently outperforms all competing method over the degrees of adjusted class imbalance, sparsity, and non-normal distribution ratio [9] [10].

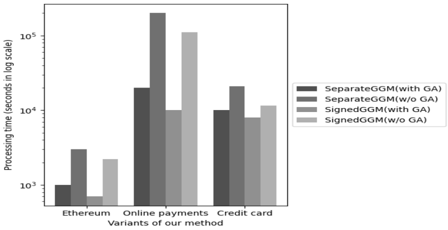


Figure 3: Processing time of the variants of my method

*4.2.3 The effect of graph topology and connectivity analysis **(RQ3)**.* As shown in Figure 3, I compare variants of my method, with and without graph topology and connectivity analysis (GA) to demonstrate that incorporating GA significantly reduces processing times in my method. Specifically, through GA, by identifying the optimal number of positive and negative Ks on separate and signed graphs in terms of detection accuracy, I eliminate the need to measure detection accuracy across all target detection models that are trained with a set of synthetic data generated from the entire augmented data's separate and signed graphs.

5 CONCLUSION AND FUTURE WORK

I propose novel graph-theoretical models for generating high quality of synthetic tabular data aimed at enhancing FFD model performance by addressing the three characteristics of FFD data: extreme class imbalance, high data sparsity, and a large number of attributes with non-normal distributions. The proposed models, SeparateGGM and SignedGGM, leverage graph-theoretical methodologies to significantly outperform existing methods across three benchmark datasets. In essence, my work demonstrates graph-theoretical methodologies can make generative models more effective in terms of detection accuracy. In future work, I will apply additional graph-theoretical concepts, such as assortativity, to observe whether other graph measurements also exhibit similar trends. In addition, I will employ additional base generative models to enhance the performance of SeparateGGM and SignedGGM.

ACKNOWLEDGEMENTS

I want to thank my supervisor, Professor In-Young Ko (iko@kaist.ac.kr), for his support and guidance. This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01795) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.

[2] Joao Fonseca and Fernando Bacao. 2023. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data* 10, 1 (2023), 115.

[3] Eunjin Jung, Marion Le Tilly, Ashish Gehani, and Yunjie Ge. 2019. Data mining-based ethereum fraud detection. In *2019 IEEE international conference on blockchain (Blockchain)*. IEEE, 266–273.

[4] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*. 17564–17579.

[5] Jaewoo Lee, Minjung Kim, Yonghyun Jeong, and Youngmin Ro. 2022. Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7345–7353.

[6] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, and Hao Fu. 2015. Financial fraud detection model: Based on random forest. *International journal of economics and finance* 7, 7 (2015).

[7] Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review* 45, 2 (2003), 167–256.

[8] Dae-Young Park and In-Young Ko. 2022. Urban Event Detection from Spatio-temporal IoT Sensor Data Using Graph-Based Machine Learning. In *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 234–241.

[9] Dae-Young Park and In-Young Ko. 2024. An Empirical Study of Utility and Disclosure Risk for Tabular Data Synthesis Models: In-Depth Analysis and Interesting Findings. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 67–74.

[10] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data Synthesis based on Generative Adversarial Networks. *Proceedings of the VLDB Endowment* 11, 10 (2018).

[11] Arjan Reurink. 2019. Financial fraud: A literature review. *Contemporary Topics in Finance: A Collection of Literature Surveys* (2019), 79–115.

[12] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems* 32 (2019).

[13] Chuang Zhang, Qizhou Wang, Tengfei Liu, Xun Lu, Jin Hong, Bo Han, and Chen Gong. 2021. Fraud detection under multi-sourced extremely noisy annotations. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 2497–2506.

[8] The values are the average accuracy of five popular fraud detection models (i.e., Random Forest, LightGBM, MLP, LSTM+CNN, and TabNet)

[9] To conduct experiments, I increase the three different characteristics intensity by 5%

[10] I omit the related figures due to space limitation