Quão efetivas são Redes Neurais baseadas em Grafos na Detecção de Fraude para Dados em Rede?

Ronald D. R. Pereira[1], **Fabricio Murai**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{ronald.pereira, murai}@dcc.ufmg.br

Abstract. *Graph-based Neural Networks (GNNs) are recent models created for learning representations of nodes (and graphs), which have achieved promising results when detecting patterns that occur in large-scale data relating different entities. Among these patterns, financial fraud stands out for its socioeconomic relevance and for presenting particular challenges, such as the extreme imbalance between the positive (fraud) and negative (legitimate transactions) classes, and the concept drift (i.e., statistical properties of the data change over time). In this work, we conduct experiments to evaluate existing techniques for detecting network fraud, considering the two previous challenges. For this, we use real data sets, complemented by synthetic data created from a new methodology introduced here. Based on this analysis, we propose a series of improvement points that should be investigated in future research.*

Resumo. *Redes Neurais baseadas em Grafos (GNNs) são modelos recentes criados para o aprendizado de representações de nós (e de grafos), que alcançaram resultados promissores na detecção de padrões que ocorrem em dados de larga escala que relacionam diferentes entidades. Dentre esses padrões, fraudes financeiras se destacam por sua relevância socioeconômica e por apresentarem desafios particulares, tais como o desbalanceamento extremo entre as classes positivas (fraudes) e negativas (transações legítimas), e o desvio de conceito (i.e., propriedades estatísticas dos dados mudam ao longo do tempo). Neste trabalho, realizamos uma série de experimentos para avaliar técnicas existentes de detecção de fraudes em rede, considerando os dois desafios anteriores. Para isso, utilizamos conjuntos de dados reais, complementados por dados sintéticos criados a partir de uma nova metodologia introduzida aqui. Baseado nessa análise, propomos uma série de pontos de melhoria a serem investigados em pesquisas futuras.*

1. Introdução

Transações financeiras tornaram-se muito mais comuns nos últimos anos com o aumento de criptomoedas, bancos digitais e *gateways* de pagamento online. Essas tecnologias transferiram o controle dos bancos para os usuários, tornando o processo de compra e transferência mais distribuído e acessível para a população em geral, mas também mais suscetível a fraudes. Esse problema afeta as métricas e os resultados financeiros de muitas empresas, já que o valor total da fraude às vezes pode ultrapassar sua receita ou tornar a margem de lucro tão pequena (mesmo que o valor da fraude seja pequeno quando comparado a transações legítimas) e, a prazo, tornar a falência apenas uma questão de tempo.

O projeto de sistemas de antifraude tem como desafios adicionais, em relação a sistemas de previsão tradicionais, limitar o número de alarmes falsos (ou seja, evitar uma alta taxa de falsos-positivos) e lidar com a diminuição na precisão das previsões conforme seus dados de treinamento se tornam obsoletos, devido ao desvio de conceito. Desvio de conceito se refere ao fenômeno de ter a distribuição de dados subjacentes mudando ao longo do tempo, tornando difícil manter a qualidade das previsões. Outro desafio é o desequilíbrio de classes, pois há muito menos fraude do que instâncias legítimas, o que torna ainda mais difícil aprender a distribuição intrínseca dos dados.

Recentemente, uma importante relação entre entidades envolvidas em transações financeiras que vai além das relações de primeira ordem "origem-destino", foi incorporada a técnicas de aprendizagem de grafos para detectar fortes relações entre instâncias distintas de dados por meio de pesos em arestas usando grafos [Weber et al. 2018, Wagner 2019, Weber et al. 2019], mas nenhuma deles tenta mitigar o desbalanceamento do conjunto de dados ou os problemas de desvio de conceito. Esses problemas são abordados em outros trabalhos de pesquisa, mas sem que se tente detectar essas relações usando modelos estruturados de grafos [Wang et al. 2017, Wang et al. 2020, Yang and King 2009, Gao et al. 2020]. Passos iniciais foram dados recentemente para solucionar o problema de desvio de conceito usando redes convolucionais de grafos sem, no entanto, considerar o problema de classes desequilibradas em suas avaliações [Pareja et al. 2020]. Por outro lado, outra pesquisa recente tenta mitigar o problema do desbalanceamento dos dados também usando redes neurais baseadas em grafos, mas sem levar em conta a desvio de conceito [Liu et al. 2021].

O objetivo desse artigo é propor uma nova metodologia para avaliar a robustez de métodos de detecção de fraudes financeiras baseada em geração de dados sintéticos, quantificando os avanços alcançados por pesquisas recentes na área ao implementá-los e compará-los com modelos anteriormente desenvolvidos. A partir desses experimentos foi possível concluir que estamos realmente progredindo para melhores desempenhos dada a complexidade inerente do domínio de detecção de fraude em dados em rede. No entanto, observamos que ainda não há um método capaz de lidar com a combinação dois maiores desafios da área e alcançar métricas de desempenho consistentes. A versão completa deste artigo pode ser vista em [Pereira and Murai 2021].

2. Trabalhos Relacionados

[Weber et al. 2018] é um dos poucos trabalhos atuais que aborda a detecção de fraude usando técnicas de aprendizagem de grafos, como a construção de Redes Convolucionais de Grafos, um tipo recente de redes neurais profundas. Posteriormente, foram abordados os cenários de lavagem de dinheiro em Bitcoin usando técnicas de aprendizagem baseadas em grafos sofisticadas [Weber et al. 2019], como Skip-GCN, uma Rede Convolucional de Grafos contendo conexões de salto via compartilhamento de pesos de aresta entre níveis de profundidade distintos da rede neural, e EvolveGCN, por [Pareja et al. 2020], uma arquitetura de rede convolucional baseados em grafos que lida com as mudanças temporais comuns desses tipos de conjuntos de dados (desvio de conceito) usando técnicas de memória de longo prazo. Um estudo mais recente [Liu et al. 2021] realizou um avanço significativo ao conseguir adaptar e utilizar redes neurais baseadas em grafos para conjuntos de dados altamente desbalanceados aplicados ao contexto de detecção de fraudes em dados em rede. Nesse estudo é realizado uma etapa de amostragem nos vértices do

Dataset	Trans. Legit.	Trans. Ilic.	Usuários Legit.	Usuários Ilic.	TD	Desv. Conc.?
AMLSim 1	5000	5000	500	500	1	Não
AMLSim 2	9500	500	950	50	19	Não
AMLSim 3	9500	500	950	50	19	Sim

Tabela 1. Estatísticas do conjunto de dados sintético AMLSim.

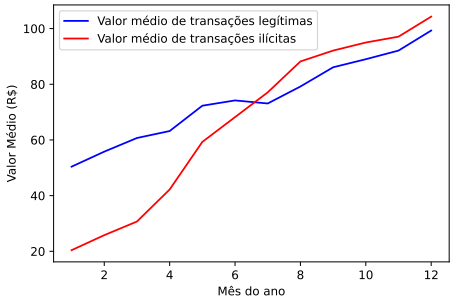

Figura 1. Variação do valor médio em dinheiro de transações legítimas e ilícitas ao longo de um ano para o AMLSim 3.

grafo de forma a equilibrar o número de vértices pertencentes a cada classe e facilitar a propagação do sinal da classe minoritária aos vértices mais próximos. Os autores concluíram que essa estratégia melhora efetivamente o desempenho dos modelos quando aplicados a esse problema.

3. Conjuntos de Dados

Dados sintéticos gerados com *Anti-money Laundering Simulator* (AMLSim): O AMLSim[1] [Pareja et al. 2020, Weber et al. 2018] é um simulador multiagente proposto para gerar conjunto de dados para o domínio de detecção de fraudes baseado em outro simulador de transações mais genérico (não possui nenhuma especificidade para gerar comportamentos de fraude), denominado PaySim [Lopez-Rojas et al. 2016]. A partir do AMLSim foram gerados três conjuntos de dados, conforme demonstrado na Tabela 1, com o objetivo de estudar a robustez dos métodos existentes em relação ao desbalanceamento de classe e ao desvio de conceito.

O primeiro conjunto de dados, AMLSim 1, é perfeitamente balanceado e servirá como um *baseline* para os demais. O segundo, AMLSim 2, possui um forte desbalanceamento de classes e servirá para demonstrar o impacto desse fenômeno no desempenho dos modelos estudados. O terceiro, AMLSim 3, também foi gerado com uma alta taxa de desbalanceamento entre as classes, porém com um efeito adicional de desvio de conceito. Mais precisamente, os valores médios das transações legítimas e ilícitas é variado mês-a-mês, durante 12 meses de simulação, conforme pode-se observar na Figura 1. Observe ainda que os valores médios chegam a se cruzar após o mês 7.

Elliptic Data Set: é um conjunto de dados coletado durante cerca de duas semanas diretamente do *blockchain* da criptomoeda Bitcoin e mapeia transações de entidades reais anonimizadas como legítimos (bolsas de valores, fornecedores de carteira, mineradores, serviços lícitos, etc) ou como ilícitos (golpes, *malware*, organizações terroristas, *ransomware*, etc) [Weber et al. 2019]. Ele contém 203769 vértices (transações) e 234355

[1] `https://github.com/IBM/AMLSim/`

Métrica	YelpChi F1-macro	YelpChi F1-fraud	YelpChi AUC	Elliptic Data Set F1-macro	Elliptic Data Set F1-fraud	Elliptic Data Set AUC
XGBoost	,546 ± ,054	,261 ± ,027	,655 ± ,045	,564 ± ,058	,259 ± ,026	,583 ± ,043
CatBoost	,518 ± ,055	,213 ± ,026	,654 ± ,046	,543 ± ,059	,201 ± ,025	,571 ± ,042
GCN	,576 ± ,059	,245 ± ,028	,620 ± ,042	,458 ± ,041	,044 ± ,024	,590 ± ,037
GAT	,551 ± ,059	,117 ± ,025	,643 ± ,045	,463 ± ,045	,026 ± ,024	,527 ± ,035
GraphSAGE	,471 ± ,054	,108 ± ,025	,579 ± ,040	,434 ± ,045	,055 ± ,022	,542 ± ,036
GraphSAINT	,605 ± ,067	,282 ± ,034	,727 ± ,052	,568 ± ,060	**,378 ± ,027**	,669 ± ,044
EvolveGCN	**,667 ± ,067**	**,459 ± ,033**	**,756 ± ,051**	**,623 ± ,064**	**,285 ± ,031**	**,744 ± ,049**
PC-GNN	**,664 ± ,063**	**,370 ± ,032**	**,757 ± ,049**	**,612 ± ,060**	,259 ± ,032	**,699 ± ,049**

Tabela 2. Desempenho dos modelos nos conjuntos de dados reais.

arestas (operações financeiras). Dentre os vértices, apenas 23% desses vértices são rotulados, sendo 2% ilícitos (4545 nós) e 21% (42019 nós) legítimos, enquanto os 77% restantes não possuem rotulação alguma.

YelpCHI: é um conjunto de dados coletado do site yelp.com que consiste em 67395 avaliações para um conjunto de 201 hotéis e restaurantes na área de Chicago e contém informações dos produtos e de seus 38063 usuários, carimbos de data e hora, classificações e a avaliação em texto livre [Mukherjee et al. 2013, Rayana and Akoglu 2015, Rayana and Akoglu 2016].

Implementações: Todas as implementações foram disponibilizadas pelos autores em um repositório público: `github.com/ronaldpereira/brasnam-experiments`.

4. Resultados

Nesta seção iremos descrever os resultados dos experimentos realizados nessa pesquisa.

Degradação em função do Desbalanceamento de Classes Para avaliar a robustez dos métodos em relação ao desbalanceamento de classes, realizamos experimentos utilizando os dois conjuntos de dados sintéticos, AMLSim 1 e AMLSim 2, cujas taxas de desbalanceamento TD são iguais a 1 (perfeitamente balanceado) e 19 (extremamente desbalanceado), respectivamente.

A Tabela 3 apresenta o desempenho de cada método para estes datasets. Analisando apenas o AMLSim 1, observa-se que os modelos do estado da arte para detecção de fraude podem ser superados pelos outros métodos quando os dados estão balanceados. Por outro lado, os resultados no AMLSim 2 mostram que quando há desbalanceamento, todos os métodos, com exceção do EvolveGCN e do PC-GNN, apresentam um baixíssimo desempenho, com AUC muito próximo de 0,5. Surpreendentemente, o EvolveGCN obteve desempenho significativamente superior (do ponto de vista estatístico) neste cenário do que no cenário balanceado, em relação às métricas F1-macro e F1-fraud.

Degradação ao longo do tempo devido ao Desvio de Conceito. Para avaliar a robustez dos métodos a um desvio de conceito gradual de classes, realizamos experimentos utilizando o conjunto de dados sintético AMLSim 3. Utilizamos os meses 1 a 4 para definir o conjunto de dados de treinamento, enquanto os dados restantes foram usados para definir dois conjuntos de teste: o primeiro contendo os meses 5 a 8, e o segundo contendo os meses 9 a 12. Esses dados sofrem de uma mudança na distribuição de suas características mês a mês, de modo que quanto maior a distância entre os meses de treinamento e os meses de teste, menor o desempenho esperado do modelo.

Métrica	AMLSim 1 (TD = 1) F1-macro	F1-fraud	AUC	AMLSim 2 (TD = 19) F1-macro	F1-fraud	AUC
XGBoost	,841 ± ,091	,906 ± ,041	,939 ± ,067	,440 ± ,045	,054 ± ,021	,535 ± ,037
CatBoost	**,865 ± ,082**	**,956 ± ,042**	**,964 ± ,057**	,473 ± ,047	,070 ± ,022	,567 ± ,035
GCN	,840 ± ,083	,714 ± ,041	,910 ± ,062	,497 ± ,048	,242 ± ,023	,554 ± ,035
GAT	,833 ± ,080	,712 ± ,041	,932 ± ,061	,461 ± ,044	,008 ± ,020	,542 ± ,039
GraphSAGE	**,874 ± ,089**	,825 ± ,043	**,994 ± ,066**	,512 ± ,052	,127 ± ,029	,620 ± ,044
GraphSAINT	,840 ± ,082	**,975 ± ,041**	,947 ± ,062	,566 ± ,052	,292 ± ,029	,616 ± ,045
EvolveGCN	,718 ± ,074	,642 ± ,037	,864 ± ,054	**,781 ± ,072**	**,826 ± ,037**	**,863 ± ,054**
PC-GNN	,748 ± ,076	,529 ± ,037	,854 ± ,056	**,747 ± ,077**	**,762 ± ,038**	**,831 ± ,057**

Tabela 3. Desempenho dos modelos no experimento sobre desbalanceamento.

Métrica	AMLSim 3 (TD = 19; Meses 5, 6, 7 e 8) F1-macro	F1-fraud	AUC	AMLSim 3 (TD = 19; Meses 9, 10, 11 e 12) F1-macro	F1-fraud	AUC
XGBoost	,419 ± ,056	,390 ± ,027	,698 ± ,045	,287 ± ,026	,158 ± ,012	,321 ± ,025
CatBoost	,434 ± ,052	,398 ± ,032	,699 ± ,057	,365 ± ,029	,048 ± ,011	,356 ± ,026
GCN	,432 ± ,063	,188 ± ,032	,542 ± ,047	,269 ± ,048	,000 ± ,000	,339 ± ,039
GAT	,379 ± ,052	,107 ± ,027	,508 ± ,043	,166 ± ,052	,000 ± ,000	,298 ± ,042
GraphSAGE	,491 ± ,065	,305 ± ,031	,697 ± ,048	,328 ± ,055	,032 ± ,023	,435 ± ,041
GraphSAINT	,534 ± ,062	,277 ± ,031	,707 ± ,049	,317 ± ,040	**,156 ± ,020**	,498 ± ,033
EvolveGCN	**,568 ± ,069**	**,425 ± ,032**	**,711 ± ,050**	**,453 ± ,048**	**,292 ± ,024**	**,579 ± ,035**
PC-GNN	**,684 ± ,065**	**,532 ± ,033**	**,777 ± ,052**	**,469 ± ,054**	,129 ± ,026	**,547 ± ,042**

Tabela 4. Desempenho dos modelos no experimento sobre desvio de conceito.

A Tabela 4 mostra os resultados obtidos para os dois conjuntos de teste. Analisando os resultados para o primeiro conjunto (meses 5 a 9), observa-se que os métodos tiveram um baixo desempenho geral em termos de F1-macro e F1-fraud, mas o AUC obtido por alguns modelos se manteve próximo ou superior a 0,7. A explicação para isto é que, embora o limiar de classificação definido com base no conjunto de treinamento cometa muitos erros nos casos de teste, o ranqueamento dos nós segundo a probabilidade de serem fraudulentos continua razoavelmente boa, fazendo com que o AUC não se degrade muito. Neste caso, um sistema de detecção de fraude poderia ser "corrigido" com um simples ajuste deste limiar.

Por outro lado, os resultados para a segundo conjunto (meses 10 a 12) mostram um resultado bem mais negativo. Além da queda ainda mais acentuada no F1-macro e F1-fraud, o AUC de todos os métodos, com exceção do EvolveGCN e do PC-GNN ficam abaixo de 0,4, indicando que tais métodos costumam classificar instâncias negativas como pais prováveis de serem fraudes do que as instâncias positivas. Isto é explicado pelo cruzamento das curvas na Figura 1. Desse modo, nenhum dos métodos do estado da arte foram capazes de manter um desempenho razoável quando expostos ao desvio de conceito, explicitando a necessidade de se realizar progressos nessa vertente.

5. Conclusões e Trabalhos Futuros

Nesse trabalho, propusemos uma nova metodologia para se avaliar a robustez de métodos de detecção de fraudes financeiras baseadas em geração de dados sintéticos utilizando o AMLSim. Desse modo, conseguimos visualizar os pontos fortes e pontos fracos de cada um dos modelos anteriormente apresentados. Também utilizando conjuntos de dados reais, mostramos que técnicas propostas recentemente, como o EvolveGCN e o PC-GNN alcançam um excelente desempenho na tarefa de classificação, a despeito da complexi-

dade inerente ao problema. No entanto, a combinação das duas características que distinguem a detecção de fraudes de tarefas comuns de classificação – o desbalanceamento de classe e o desvio de conceito – ainda impõe desafios para estas técnicas. Concluímos que ainda não existe um modelo que seja capaz de lidar com estas duas características simultaneamente. Como direções futuras, acreditamos que técnicas que utilizem amostragem para compensar o desbalanceamento e que incorporem, ao mesmo tempo, um modelo para a dinâmica da relação entre *features* e o rótulo possam ser promissoras para o avanço na área de detecção de fraudes para dados em rede.

Referências

Gao, L., Yang, L., Arefan, D., and Wu, S. (2020). One-class classification for highly imbalanced medical image data. In *Medical Imaging 2020*, pages 342 – 347.

Liu, Y., Ao, X., Qin, Z., Chi, J., Yang, H., and He, Q. (2021). Pick and choose: A gnn-based imbalanced learning approach for fraud detection. In *WWW*.

Lopez-Rojas, E., Elmir, A., and Axelsson, S. (2016). Paysim: A financial mobile money simulator for fraud detection. In *EMSS*, pages 249–255. Dime University of Genoa.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013). What yelp fake review filter might be doing? In *ICWSM*, volume 7.

Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T. B., and Leiserson, C. E. (2020). EvolveGCN: Evolving graph convolutional networks for dynamic graphs. In *AAAI*.

Pereira, R. D. R. and Murai, F. (2021). How effective are graph neural networks in fraud detection for network data? *arXiv preprint arXiv:2105.14568*.

Rayana, S. and Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In *KDD*, pages 985–994.

Rayana, S. and Akoglu, L. (2016). Collective opinion spam detection using active inference. In *SDM*, pages 630–638. SIAM.

Wagner, D. (2019). Latent representations of transaction network graphs in continuous vector spaces as features for money laundering detection. In *SKILL*, pages 143–154.

Wang, G., Yang, J., and Li, R. (2017). Imbalanced svm-based anomaly detection algorithm for imbalanced training datasets. *Etri Journal*, 39(5):621–631.

Wang, X., Du, Y., Cui, P., and Yang, Y. (2020). Ocgnn: One-class classification with graph neural networks. *arXiv preprint arXiv:2002.09594*.

Weber, M., Chen, J., Suzumura, T., Pareja, A., Ma, T., Kanezashi, H., Kaler, T., Leiserson, C. E., and Schardl, T. B. (2018). Scalable graph learning for anti-money laundering: A first look. *arXiv preprint arXiv:1812.00076*.

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*.

Yang, H. and King, I. (2009). Ensemble learning for imbalanced e-commerce transaction anomaly classification. In *ICONIP*, pages 866–874. Springer.