Supplementary material for: Evaluating explainability techniques on discrete-time graph neural networks

Anonymous authors

1 Additional information on datasets

We evaluate the explainability models on three real-world temporal graph datasets, covering three of the most important applicative domains of discrete-time networks, which are financial, social, and collaboration networks. Specifically, we consider:

- BitcoinOTC [Pareja *et al.*, 2020]: This is a who-trusts-whom network of people who trade using Bitcoin on a platform called Bitcoin OTC. Since Bitcoin users are anonymous, there is a need to maintain a record of users' reputations to prevent transactions with fraudulent and risky users. Members of Bitcoin OTC rate other members on a scale of -10 (total distrust) to +10 (total trust) in steps of 1.
- Reddit-title [You *et al.*, 2022]: It is a hyperlink network that represents the directed connections between two subreddits (a subreddit is a community on Reddit). The network is extracted from publicly available Reddit data covering 2.5 years, from Jan 2014 to April 2017. Specifically, it is extracted from the posts that create hyperlinks from one subreddit to another. A hyperlink originates from a post in the source community and links to a post in the target community. Each hyperlink is annotated with its timestamp. We consider the hyperlinks present in the title of the posts.
- Email-EU [Paranjape *et al.*, 2017]: The network was generated using email data from a large European research institution. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world. A directed edge (u, v, t) means that person u sent an e-mail to person v at time t. A separate edge is created for each recipient of the e-mail. We consider the subnetwork corresponding to the communication between members of a single department at the institution (Department 1, as defined by SNAP[1]).

A summary of the dataset properties is given in Table 1. For each dataset, in Table 2 we also report the global level of edge recurrence, as defined in [Poursafaei *et al.*, 2022; Gastinger *et al.*, 2024], edge reciprocity, analogously, and structural homophily, defined as the average structural homophily of the edges in the last train snapshot respect to the rest of the train set. Notice that all three networks are directed but Reddit-title and Email-EU exhibit no reciprocity mechanism in their evolution, and the level of structural homophily is generally quite low.

Dataset	#Edges	#Nodes	Frequency	#Snapshots
BitcoinOTC	35,588	6,005	Weekly	138
Reddit-title	571,927	54,075	Weekly	178
Email-EU	332,334	986	Daily	526

Table 1: Dataset statistics

Dataset	Recurrence	Reciprocity	Homophily
BitcoinOTC	0.45	0.31	0.24
Reddit-title	0.15	0.00	0.04
Email-EU	0.35	0.02	0.16

Table 2: Level of edge recurrence, reciprocity, and structural homophily in the three selected datasets. All metrics range between zero and one.

2 Additional information on base models

We adopt three state-of-the-art discrete-time graph neural networks as the base model:

- EvolveGCN [Pareja *et al.*, 2020]: it captures the dynamic of the graph sequence of snapshots by using an RNN to update the weights of each GNN layer. In this way, the RNN regulates the GCN model parameter directly and effectively performs model adaptation. Note that the GNN parameters are not trained and only computed from the RNN.
- GCRN-GRU [Seo *et al.*, 2018]: It is a generalization of the T-GCN model [Zhao *et al.*, 2020], which internalizes a GNN into the GRU cell by replacing linear transformations in GRU with graph convolution operators. GCRN uses ChebNet [Defferrard *et al.*, 2016] for spatial information and separated GNNs to compute different gates of RNNs.

[1] https://snap.stanford.edu/data/email-Eu-core-temporal.html, January 2025

Model	LR	WD	#Layers	d
EvolveGCN	0.010	5e-3	1	128
GCRN-GRU	0.010	5e-3	1	128
ROLAND-GRU	0.001	5e-3	2	128

Table 3: Best configuration of hyperparameters for Email-EU

- ROLAND-GRU [You *et al.*, 2022]: ROLAND is a framework that can help researchers re-purpose any static GNN to a dynamic graph learning task; consequently, adapting state-of-the-art designs from static GNNs and significantly lower the barrier to learning from dynamic graphs. Specifically, node embeddings at different GNN layers are viewed as hierarchical node states. To generalize a static GNN to a dynamic setting, you only need to define how to update these hierarchical node states based on newly observed nodes and edges. In this paper, we focus on the most effective node update solution, which is based on leveraging gated recurrent units (GRUs).

3 Link prediction experiments

Setting. We train and evaluate each base model on future link prediction tasks on the three considered datasets. We use our newly introduced training and evaluation setting, which is a slight variation of the recently proposed UTG framework [Huang *et al.*, 2024] in which the training phase is done using the live-update setting [You *et al.*, 2022]. As a standard practice [Pareja *et al.*, 2020], we perform random negative sampling for each test snapshot. We report the performance on the test set in terms of the Area Under Precision Recall Curve (AUPRC).

Hyperparameters. Base models are pre-trained using the best configuration of hyperparameters in the original papers, when available, or optimized using grid-search. In particular, the best configuration of hyperparameters can be found in the following work for BitcoinOTC and Reddit-title [Pareja *et al.*, 2020; You *et al.*, 2022]. The hyperparameter search spaces used for Email-EU are as follows: learning rate (LR) $\{0.1, 0.01, 0.001, 0.0001\}$, L2 weight-decay (WD) {5e-1, 5e-2, 5e-3, 5e-4}, number of hidden layers (#Layers) $\{1, 2\}$, representation dimension (d) $\{32, 64, 128, 256\}$. We report the best configuration of hyperparameters for the Email-EU dataset in Table 3.

Results. We report the test-set performance of the models in Table 4. Results are aggregated over the different test snapshots and averaged on 3 different random seeds. Overall, we reach a good level of link prediction performance with at least two models on all the considered datasets. However, GCRN-GRU can obtain performances close to a random edge predictor for Reddit-title and Email-EU datasets. Since the focus of our work is explaining the model's predictions, i.e. the TGNN logic, the level of performance does not impact the discussion of our results. Nevertheless, it could be interesting for future works to analyze if explaining wrong predictions only changes a lot the rank of the evaluated explainability techniques for discrete-time GNNs.

Model	BitcoinOTC	Reddit-title	Email-EU
EvolveGCN	84.48 ± 2.31	**88.93 ± 0.69**	66.91 ± 6.70
GCRN-GRU	96.31 ± 1.56	53.91 ± 1.43	57.79 ± 6.13
ROLAND-GRU	**96.89 ± 1.19**	77.93 ± 4.19	**70.56 ± 7.14**

Table 4: Link prediction performances of base models on the three datasets, in terms of AUPRC.

4 Fidelity trends.

We report all the fidelity trends for each base model and dataset in Figure 3. Thanks to the unlimited space of the Supplementary Material, we decided to report the trends using a subplot for each trend, for better clarity and readability. The figure reported in the paper is Figureh. The same questions discussed in the paper can be answered for each figure.

5 Additional information on the case study

In the paper, we present a case study where we focus on explaining the decisions of TGNNs in a specific real-world challenge. Specifically, we chose to explain the decisions behind the existence of total distrust edge in BitcoinOTC. In fact, the BitcoinOTC platform allows its members to rate other members on a scale of -10 (total distrust) to +10 (total trust) in steps of 1. Since their anonymity, this creates a record of users' reputations, which is needed to prevent transactions with fraudulent and risky users. Figure 1 shows the distribution of ratings in BitcoinOTC. Most users receive scores from 1 to 3, and only a few votes are negatives. Hence, predicting the existence of total distrust edges is important because it enables the identification of untrustworthy users, safeguards transactions, and protects the integrity of the platform. However, explaining the decisions made by a GNN in this context is fundamental to fostering trust in the model and ensuring its reliability. Without clear explanations, users and administrators may struggle to justify the model's decisions, particularly when false positives occur. Providing these explanations is crucial to avoid misclassifying good users as fraudulent, which could unfairly damage their reputation and undermine confidence in the system. To this end, we ask for an explainability model to obtain the important events related to the decisions of all the distrust edges in the first test snapshot. Overall, we obtain 70 target events. Specifically, we chose ROLAND-GRU as the base model and GNNExplainer, since they achieve the best performance on link prediction and fidelity on BitcoinOTC, respectively. In the paper, we report three of the most frequent kinds of explanation, observing highly human-readable explanations, and finding that most decisions are made based on edge recurrence, negative consensus on the target nodes, and authority of source nodes. We recall that recurrence is the mechanism by which two nodes that interacted in the past are likely to interact again in the future. Given a target event $e = (u, v, t)$, we define consensus as the average vote on the incoming edges of the destination node v before t, and authority as the in-degree centrality of the source node u before t, considering incoming edges with positive votes only. A negative consensus is an average consensus lower than zero. To evaluate quantitatively the presence of these patterns in the given explanations, we compare the distribution of authority and consensus in the explanatory subgraphs, computational graphs, and random vote networks. Random networks are generated using the Erdős–Rényi model [Barabasi and Posfai, 2016] with the probability of edge creation and number of nodes equal to the explanatory graph's density and number of nodes, respectively, and edge weights assigned uniformly at random in $[-10, 10]$. Comparing the two metrics with the ones obtained on a random network with an equal number of nodes and edges (on average) is a way to understand whether obtaining the described behavior for consensus and authority only happens by chance or not. We show the boxplot of the distributions of consensus and authority on the three graphs in Figure 2. Overall, we observe that only a few explanations leverage negative consensus, but they exhibit a value lower than the average consensus of both candidate and random graphs. Concerning authority, we notice that its value is generally far higher in the explanations than in the candidate or random graphs, confirming that it is a very leveraged pattern to decide whether a total distrust edge exists or not.

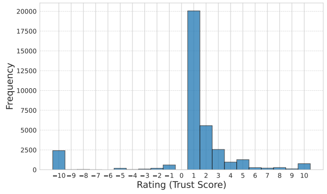


Figure 1: Distribution of ratings in BitcoinOTC.

References

[Barabasi and Posfai, 2016] Albert-Laszlo Barabasi and Marton Posfai. *Network science*. Cambridge University Press, Cambridge, 2016.

[Defferrard *et al.*, 2016] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc.

[Gastinger *et al.*, 2024] Julia Gastinger, Christian Meilicke, Federico Errica, Timo Sztyler, Anett Schülke, and Heiner Stuckenschmidt. History repeats itself: A baseline for temporal knowledge graph forecasting. In *IJCAI*, pages 4016–4024. ijcai.org, 2024.

[Huang *et al.*, 2024] Shenyang Huang, Farimah Poursafaei, Reihaneh Rabbany, Guillaume Rabusseau, and Emanuele Rossi. UTG: Towards a unified view of snapshot and event based models for temporal graphs. In *The Third Learning on Graphs Conference*, 2024.

[Paranjape *et al.*, 2017] Ashwin Paranjape, Austin R. Benson, and Jure Leskovec. Motifs in temporal networks. In *WSDM*, pages 601–610. ACM, 2017.

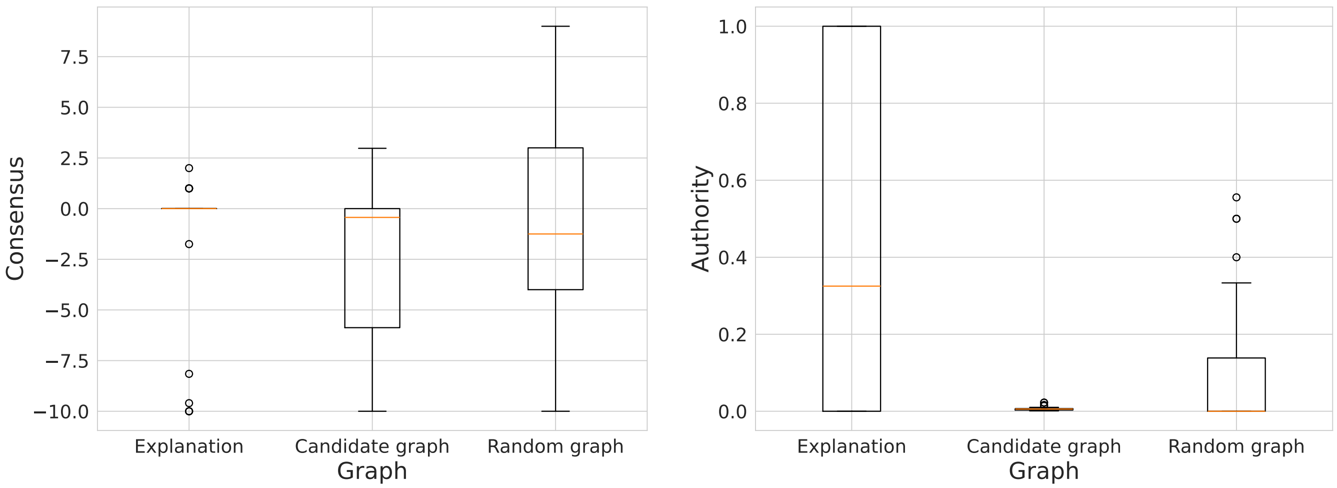
[Pareja *et al.*, 2020] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. Evolvegcn: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5363–5370, 2020.

[Poursafaei *et al.*, 2022] Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In *NeurIPS*, 2022.

[Seo *et al.*, 2018] Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I 25*, pages 362–373. Springer, 2018.

[You *et al.*, 2022] Jiaxuan You, Tianyu Du, and Jure Leskovec. Roland: Graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2358–2366, New York, NY, USA, 2022. Association for Computing Machinery.

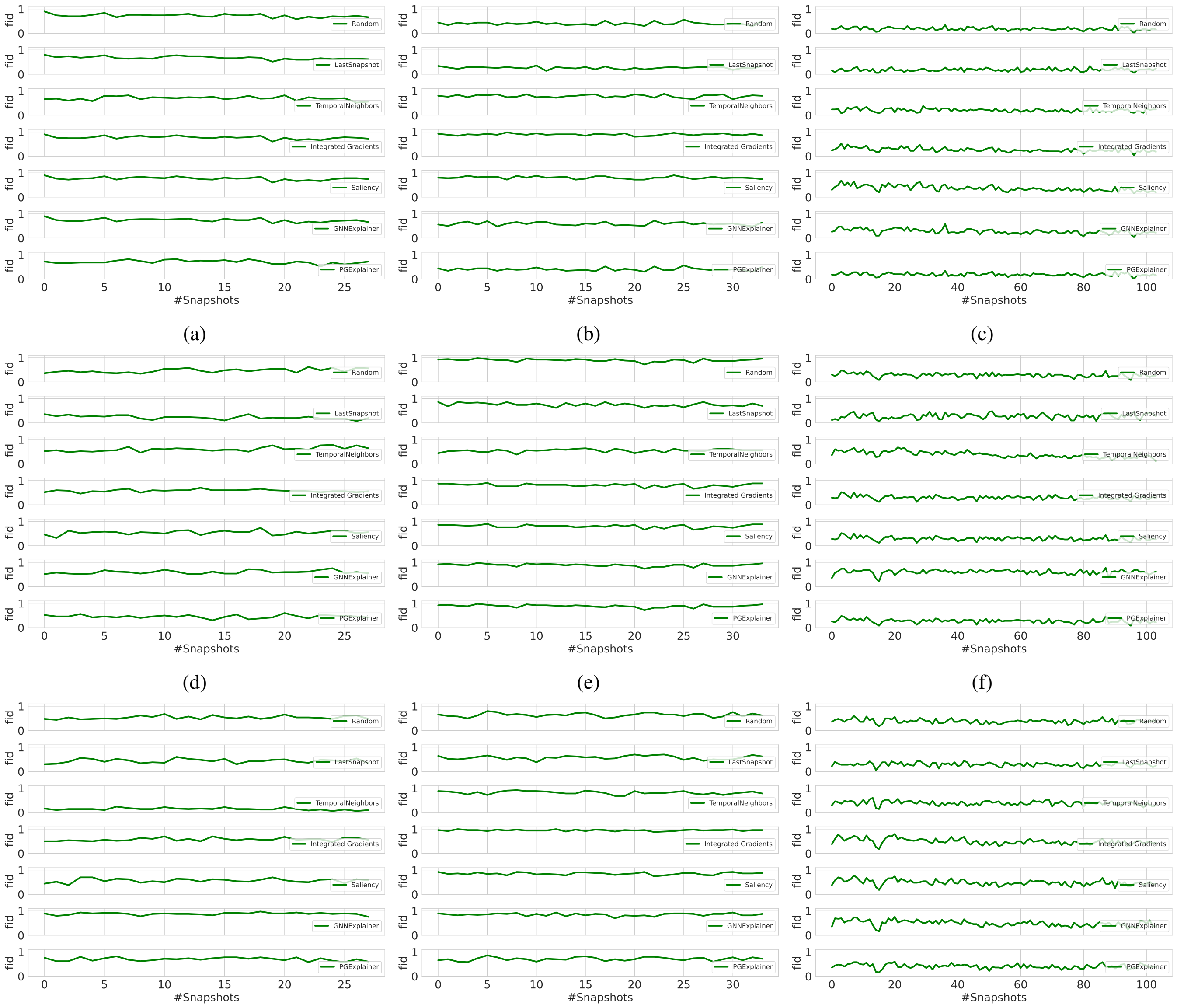
[Zhao *et al.*, 2020] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858, sep 2020.



(a)

(b)

Figure 2: Boxplot of the distributions of consensus (a) and authority (b) for existing target events corresponding to distrust edges in the first test snapshot. The three boxplots refer to the explanatory subgraphs (Explanation), the computational graph (Candidate graph), and the random graphs.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 3: Fidelity trends. The first column refers to the BitcoinOTC dataset, the second to Reddit-title, and the third to Email-EU. The first row refers to EvolveGCN, the second to GCRN-GRU, and the third to ROLAND-GRU.