Are We Really Measuring Progress? Transferring Insights from Evaluating Recommender Systems to Temporal Link Prediction

Filip Cornell*
filip.cornell@king.com
King/Microsoft
Stockholm, Sweden

Oleg Smirnov*†
oleg.smirnov@microsoft.com
King/Microsoft
Stockholm, Sweden

Gabriela Zarzar Gandler
gabriela.zarzar@king.com
King/Microsoft
Stockholm, Sweden

Lele Cao
lele.cao@king.com
King/Microsoft
Stockholm, Sweden

ABSTRACT

Recent work has questioned the reliability of graph learning benchmarks, citing concerns around task design, methodological rigor, and data suitability. In this extended abstract, we contribute to this discussion by focusing on evaluation strategies in Temporal Link Prediction (TLP). We observe that current evaluation protocols are often affected by one or more of the following issues: (1) inconsistent sampled metrics, (2) reliance on hard negative sampling often introduced as a means to improve robustness, and (3) metrics that implicitly assume equal base probabilities across source nodes by combining predictions. We support these claims through illustrative examples and connections to longstanding concerns in the recommender systems community. Our ongoing work aims to systematically characterize these problems and explore alternatives that can lead to more robust and interpretable evaluation. We conclude with a discussion of potential directions for improving the reliability of TLP benchmarks.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; *Evaluation of retrieval results*; • **Computing methodologies** → **Machine learning**; *Neural networks*.

KEYWORDS

temporal graph learning, temporal link prediction, ranking metrics, evaluation

ACM Reference Format:
Filip Cornell, Oleg Smirnov, Gabriela Zarzar Gandler, and Lele Cao. 2025. Are We Really Measuring Progress? Transferring Insights from Evaluating Recommender Systems to Temporal Link Prediction. In *Proceedings of Temporal Graph Learning Workshop, SIGKDD International Conference on Knowledge Discovery and Data Mining 2025 (TGL Workshop, KDD 2025).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.
†Corresponding author.

TGL Workshop, KDD 2025, August 03-07, 2025, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the increasing interest in learning over graphs, Temporal Graph Learning (TGL) has gained significant traction, leading to the development of numerous neural architectures. A variety of Temporal Graph Neural Networks (TGNNs) have been proposed [Cong et al. 2023; Ding et al. 2024; Gravina et al. 2024; Kumar et al. 2019; Rossi et al. 2020; Wang et al. 2021a,b; Xu et al. 2020; Yu et al. 2023], alongside several distinct evaluation protocols [Huang et al. 2024a; Poursafaei et al. 2022a; Poursafaei and Rabbany 2023]. Notably, under these evaluation schemes, prior studies have demonstrated that simple heuristic methods can outperform neural models in both temporal graph and Temporal Knowledge Graph tasks [Cornell et al. 2025; Dileo et al. 2024; Gastinger et al. 2024].

This observation raises concerns about the complexity of existing datasets. Yi et al. [2025] and Gastinger et al. [2024] suggest that repetitive structures in datasets may lead to predictable patterns. If such patterns dominate, learning-based methods should ideally capture and build upon them more effectively.

Another key limitation of many TGL models is their poor scalability to large datasets. These models typically maintain time-dependent node embeddings that evolve at each timestamp, making standard retrieval methods, such as kNN-based lookup, impractical. To manage this, prior work has adopted sampling-based evaluation strategies to enable experimentation on large-scale workloads.

A recent position paper [Bechler-Speicher et al. 2025] calls for a substantial revision of benchmarks in the broader graph learning field. Among the key concerns is the choice of evaluation strategy, which fundamentally influences how model performance is interpreted and compared.

This work contributes to that discussion by examining evaluation strategies used in TLP. We illustrate how specific metric choices may lead to inconsistent or misleading conclusions, which we illustrate through practical examples. Our analysis draws on prior work [Krichene and Rendle 2020; Poursafaei and Rabbany 2023], particularly from the recommender systems domain, where similar issues have long been recognized. We further present empirical results on TLP benchmarks that reveal critical flaws in current evaluation protocols. These findings further motivate the need for

Benchmark	Inconsistent Sampled Metrics	Hard Negative Sampling	Combined Nodes Predictions
DGB [Poursafaei et al. 2022a]	■	■	■
TGB [Huang et al. 2024a]	■	■	❑
BenchTemp [Huang et al. 2024b]	■	❑	■
TGB-Seq [Yi et al. 2025]	■	❑	❑
DyGLib [Yu et al. 2023]	■ / ❑	■	■

Table 1: Evaluation issues across TLP benchmarks and libraries. ■ indicates the issue is present; ❑ indicates absence. For DyGLib, the authors additionally report the uniformly sampled average rank, which avoids the inconsistency issue.

continued scrutiny in benchmark and metric design, aligning with concerns raised in prior work.

2 ISSUES IN TLP EVALUATION

This section reviews evaluation strategies in TLP, highlighting how they can produce misleading impressions of model performance. We illustrate these issues through prior work in recommender systems and with practical and illustrative examples.

We examine five established benchmarks and libraries: Dynamic Graph Benchmark (DGB) [Poursafaei et al. 2022a], Temporal Graph Benchmark (TGB) [Huang et al. 2024a], BenchTemp [Huang et al. 2024b], DyGLib [Yu et al. 2023], and TGB-Seq [Yi et al. 2025]. These, along with most TLP studies, adopt ROC-AUC and Average Precision (AP) for binary classification [Huang et al. 2024b; Li et al. 2023; Yu et al. 2023], or sampled Mean Reciprocal Rank (MRR) for ranking tasks [Huang et al. 2024a; Yi et al. 2025].

Rahman et al. [2025] frame these paradigms as questions: binary classification asks *"Does an edge exist between nodes u and v at time t?"*, while ranking asks *"Which nodes are likely to connect with node u at time t?"*

Both formulations exhibit limitations that merit closer examination. Table 1 summarizes the presence of three key concerns, across these benchmarks, discussed in details in following sections: (i) inconsistent sampled metrics, (ii) hard negative sampling, and (iii) combining predictions across nodes.

2.1 Inconsistent Sampled Metrics

Numerous works [Dallmann et al. 2021; Jin et al. 2021; Krichene and Rendle 2020; Li et al. 2020, 2023, 2024; Liu et al. 2023a,b; Zhao et al. 2022] in the recommender systems domain have examined the inconsistencies of top-k ranking metrics under sampling. Krichene and Rendle [2020] analytically demonstrate that such metrics are not consistent under uniform sampling, even in expectation. They identify AUC as the only truly consistent metric due to its linearity, and propose three rank estimators to mitigate inconsistency: an unbiased estimator, a minimal bias estimator, and a variance-reducing extension of the latter.

Concurrently, Cañamares and Castells [2020] analyzed target set effects and reached similar conclusions. Hidasi and Czapp [2023] argue that sampled ranking metrics represent a pervasive flaw in the field. Li et al. [2020] and subsequent works [Jin et al. 2021; Li et al. 2023, 2024] introduce a series of correction techniques aimed at approximating true top-k metrics from sampled data. More recently, Liu et al. [2023a] explored how the robustness, consistency, and discriminative power of ranking metrics depend on dataset properties. Liu et al. [2023b] later proposed a sampling strategy that adapts based on the number of items each user has interacted with.

Sampled ranking metrics in TLP. Two prominent benchmarks, TGB and TGB-Seq, employ ranking metrics in the form of sampled MRR. This is equivalent to the simplified sampled AP analyzed by Krichene and Rendle [2020], who showed that such metrics are inconsistent under sampling. Sampling is used because full-ranking evaluations are computationally expensive and do not scale well for many current TGNNs [Huang et al. 2024a; Yi et al. 2025]. To investigate alignment issues between full and sampled ranking metrics, we require evaluations in both settings. To this end, we employ scalable heuristics shown to be effective on TGB and BenchTemp [Cornell et al. 2025]. These heuristics score destination nodes based on either timestamp (recency) or occurrence count (popularity), operating at a local scale (per source node) or a global scale (entire dataset).

Using these heuristics, we observe a Simpson's paradox [Hernán et al. 2011] on certain datasets. Figure 1 illustrates this phenomenon using the Yelp dataset from TGB-Seq [Yi et al. 2025]. Heuristics with *local* granularity, which track destination scores per source node, show strong internal positive correlation. However, when combined with *global* heuristics, the overall correlation turns strongly negative. Conversely, excluding the local heuristics, the global ones remain positively correlated among themselves.

This suggests that small changes in inductive biases [Mitchell 1980] of a model yield consistent results within similar heuristic types, while more significant differences introduce greater inconsistency. While not rigorously tested on conventional TGNN models due to the prohibitive computational cost of running inference for each node at every timestamp in any real-world settings, we hypothesize that similar effects may also manifest in these models as well. In particular, the fact that algorithmic heuristics yield sampled metric scores closely aligned with those of TGNNs (ref. Table 2 in [Cornell et al. 2025]) suggests that correlation risks extending to other evaluation protocols.

As part of our ongoing work, we explore using covariance estimates between model predictions and negative samplers to assess bias in sampled ranking metrics. While challenging in practice due to non-uniform covariation and skewed score distributions, this approach may help identify when model comparisons are unfair and guide the development of principled correction methods.

2.2 Hard Negative Sampling

Sampling harder negatives for improved evaluation in TLP was first proposed by the authors of DGB [Poursafaei et al. 2022a], who observed that the binary classification task often becomes too easy. Under uniform sampling, evaluation metrics can yield near-perfect scores, a pattern also evident in the BenchTemp leaderboard.

Negative sampling strategies in TLP evaluation. While hard negative sampling is often introduced to improve evaluation robustness, it is well established in the recommender systems community [Dallmann et al. 2021; Hidasi and Czapp 2023; Liu et al. 2023a; Zhao et al. 2022] that it can lead to inconsistent conclusions. Its

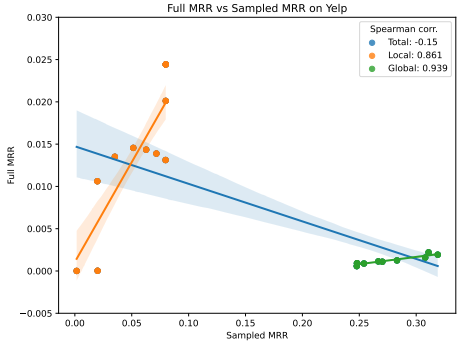

Figure 1: Sampled vs. full MRR for local and global heuristics, highlighting inconsistency across evaluation scales.

impact varies across models and workloads, and although it may improve robustness in specific cases, it is not considered a generally reliable evaluation strategy. For example, Dallmann et al. [2021] show that popularity-based sampling, similarly to uniform sampling, can produce results inconsistent with full-ranking metrics. Hidasi and Czapp [2023] also discuss candidate generation and note that although better candidate generation methods may mitigate the issue, no universally accepted solution currently exists.

While a general drop in model performance is expected under hard negative sampling as noted in prior work [Poursafaei et al. 2022a], the key question is: *given a specific sampler, does each model's metric converge to its true value at the same rate compared to uniform sampling?* If not, model rankings may shift, leading to potential underestimation or overestimation of performance metrics for some models. In certain settings, performance estimates are observed to be less reliable than those from uniform sampling, as illustrated by the `tgbl-review` dataset. We discuss this issue in further details in Appendix A.

Ranking discrepancies of this kind have already been highlighted in the TLP domain [Poursafaei et al. 2022a; Poursafaei and Rabbany 2023; Yu et al. 2023]. For example, based on results from Poursafaei et al. [2022b], we computed Spearman rank correlations of the average test AUCs of all models reported under different negative sampling strategies, as shown in Table 2. In at least one dataset for each of the three sampling strategies, the correlation is *negative*, indicating inconsistent or even reversed model rankings. Across all measurements, only 8 out of 39 correlations approach 0.9, underscoring the high variability in comparative model performance. Computing analogous statistics for nine models using the setup from Yu et al. [2023] yields the same pattern.

As several works have pointed out [Poursafaei et al. 2022a; Poursafaei and Rabbany 2023; Yu et al. 2023], different biased evaluations capture different aspects of model behavior. While this enables valuable analysis, it raises the fundamental question: *which sampled metric should be trusted when selecting a model?* Choosing the maximum result across samplers risks aligning with aspects not reflected in real deployment, leading to overestimation. Without alignment between the sampling strategy and the inference-time setup, real-world performance estimates may become difficult to interpret reliably. This concern has been raised in prior work, notably by Hidasi and Czapp [2023] and Castells and Moffat [2022], the latter arguing that the ideal candidate selector in offline evaluation should mirror the one used in production systems.

Our ongoing work investigates a debiasing strategy for hard negative sampling that aggregates multiple biased samplers. By combining diverse candidate sources, the approach aims to mitigate individual sampler biases and approximate a more balanced evaluation signal.

2.3 Combined Nodes Predictions

The final issue we identify in TLP evaluation arises when edge predictions from different source nodes are combined. This approach implicitly assumes equal base probabilities for all edges, regardless of the query (source) node. A similar concern was raised in the recommender systems domain by Tamm et al. [2021], who questioned the validity of aggregating recommendations across users. They proposed computing AUC on a per-user basis as a more appropriate alternative, though without fully addressing its limitations. This per-user AUC formulation is adopted in OpenRec [Yang et al. 2018] and used by Zhu et al. [2017].

To illustrate this issue, consider the example in Figure 2, showing two source nodes, U_1 and U_2, with different distances to their respective targets. The filtered ranking metrics yield perfect scores, as all queries $Q = \{(u_1, v_i) \in E\} \cup \{(u_2, v_i) \in E\}$ receive a filtered rank of 1. However, both ROC AUC and AP report much lower values. This occurs because the negative samples associated with U_1 are ranked above the true edges of U_2, resulting in poor rankings despite perfect per-query results.

The problem intensifies when a source node appears frequently in the dataset. Such a node, having a higher base probability of forming edges, may rank closer even to incorrect targets, causing the true edges of less frequent nodes to be ranked lower. In graphs with power law degree distributions, hub nodes can dominate the sampled evaluation, leading to misleading metric values.

How significant is this issue? From a graph-theoretic perspective, node degrees in real-world graphs often follow a power-law distribution [Barabási and Albert 1999], where a small number of nodes have very high degree while the majority have low degree. As a result, high-degree nodes inherently have a higher base probability of forming edges and are more likely to appear in the test set. We conjecture that this imbalance can substantially affect evaluation outcomes, particularly as a function of source node degree distribution in the evaluation set.

3 REMEDIES AND SOLUTIONS

Multiple solutions have been proposed in prior work, and here we outline a few promising directions in the context of TLP.

Avoiding sampling schemes completely. Several works [Hidasi and Czapp 2023; Krichene and Rendle 2020; Zhao et al. 2022] recommend avoiding sampled top-k ranking metrics entirely. One promising direction is to develop model architectures that support full-scale evaluation or kNN-based retrieval during inference, enabling accurate top-k metrics. Recent TLP benchmarks have made progress in this direction: both TGB and TGB-Seq report training

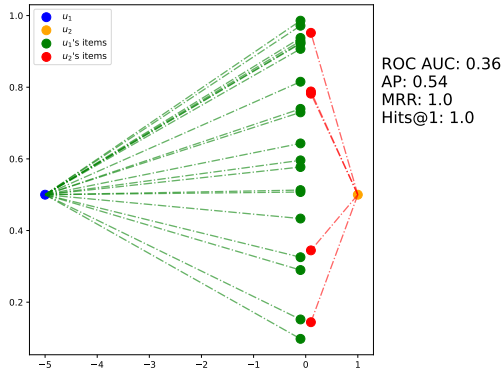

Figure 2: An example where ranking metrics computed per source node yield perfect scores, while aggregated metrics such as AUC and AP report substantially lower values.

Sampling strategy	Historical Inductive	Historical Random	Inductive Random
Can. Parl.	0.99	0.38	0.36
Contact	0.34	0.16	0.02
Enron	0.49	0.2	-0.27
Flights	-0.1	-0.22	0.52
LastFM	0.29	-0.11	-0.23
MOOC	0.85	0.94	0.93
Reddit	0.75	0.84	0.9
Social Evo.	0.93	0.89	0.89
UCI	0.71	0.63	0.7
UN Trade	0.11	0.52	0.16
UN Vote	0.64	0.41	0.77
US Legis.	0.83	0.54	0.21
Wikipedia	0.69	0.82	0.95

Table 2: Spearman rank correlations of seven models based on their AUC rankings, using Poursafaei et al. [2022b] results.

and inference runtimes, while BenchTemp incorporates runtime directly into its leaderboard metrics.

Pushing further remains valuable. A compelling example is the WikiKG90Mv2 benchmark from the OGB-LSC competition [Hu et al. 2021], where models must retrieve the top 10 entities from all 91 million entities, resulting in measuring *MRR*@10. This setup enforces scalability by design and highlights the importance of evaluation constraints aligned with deployment realities.

Using consistent metrics. If sampled metrics are still required, more reliable alternatives exist. The simplified AUC analyzed by Krichene and Rendle [2020] corresponds to a consistent, normalized inverse estimate of the mean rank, denoted $\widehat{MR}$. Closely related is the hit ratio, which Li et al. [2020] show is linearly related to the expected sampled Hits@$C \cdot K$ and Hits@K, where $C = \frac{|V|-1}{n_s}$.

These properties make $\widehat{MR}$ and Hits@$C \cdot K$ consistent even when sampling a large fixed set of nodes for all queries, assuming appropriate handling of true answer collisions. This approach reduces the sampling complexity from $O(|E_{\text{test}}|)$ to $O(1)$, while maintaining statistical reliability. However, it is important to note that AUC captures different performance aspects compared to commonly used top-k metrics in TLP [Zhao et al. 2022]. Some recent works [Ding et al. 2024; Yu et al. 2023] already report uniformly sampled mean rank, thereby adopting a more consistent evaluation.

It is important to emphasize that $\widehat{MR}$ assumes *uniform* negative sampling. When harder negatives are used, the estimate becomes biased. Furthermore, $\widehat{MR}$ is not without limitations. If other true targets are filtered out during evaluation, this can affect the metric's reliability. As a result, we are restricted to using the *raw* $\widehat{MR}$, where all unranked entities, including true ones, are treated as negatives. Under suitable conditions, e.g., when the number of filtered entities is small and the number of nodes is large, this bias can be considered negligible. However, if many true entities are filtered for a given query, the impact may become substantial.

Using statistical corrected estimates. Several works [Krichene and Rendle 2020; Li et al. 2020, 2023, 2024; Liu et al. 2023b] propose statistical correction techniques to improve the estimation of true ranking metrics. While often mathematically involved, these methods can reduce bias, though at the cost of increased variance. Their correction effectiveness has been questioned [Zhao et al. 2022], and further research is needed to adapt these approaches to TLP, where temporal dynamics must be taken into account.

Using alternative evaluation protocols. Some recent works have proposed alternative evaluation strategies to address limitations in existing metrics. For example, Rahman et al. [2025] suggest a counterfactual evaluation method that perturbs timestamps. However, this approach still relies on metrics such as AP and ROC-AUC, which remain sensitive to inconsistent sampling and combines predictions across source nodes.

Earlier, Poursafaei and Rabbany [2023] introduced EXHaustive, which avoids sampling by partitioning edges into time intervals, thereby reducing repeated evaluations. This strategy highlights performance drops and ranking inconsistencies also reported by Yu et al. [2023]. However, EXHaustive has practical limitations and may introduce its own biases. It requires domain knowledge, does not scale well with dense temporal datasets, and also introduces its own biases. In particular, it assumes that the model should prioritize edges selected by the evaluation strategy itself – an assumption that may not hold for many models.

While these methods represent useful steps forward, they may not fully address the underlying challenges and can shift the bias rather than remove it. As a result, their limitations need to be carefully considered before drawing conclusions.

4 CONCLUSION

This work highlights the need for a unified and reliable evaluation framework for TLP, building on insights extensively discussed in the recommender systems literature. Ideally, such a framework should be easy to use in practice and enable confident interpretation of model performance. However, the assumptions underlying current evaluation protocols often undermine this goal. These issues may not appear in every setting, as their impact depends on interactions between dataset characteristics, model behavior, and evaluation design. Still, even isolated counterexamples can demonstrate how current methods may yield unreliable or misleading results if their limitations are not well understood.

In the absence of a perfect solution, our suggestions are as follows. First, models should be designed with reliable evaluation in mind. Second, if sampling is used, it should yield statistically unbiased estimates or be paired with a suitable correction method. Third, biased negative sampling should be avoided when possible; if used, conclusions about model performance should be limited to scenarios where the sampling strategy aligns with the inference-time setting.

Looking ahead, our ongoing work aims to address these limitations through a covariance-based correction method for sampling bias and a multi-sampler strategy to mitigate distortions from hard negative sampling. We believe these directions can contribute to more robust and interpretable evaluation, and support the development of fair, reproducible benchmarks. More broadly, continued exchange between TGL and recommender systems domains may inspire more principled evaluation practices.

ACKNOWLEDGMENTS

This work was partially supported by Wallenberg AI, Autonomous Systems and Software Program (WASP).

REFERENCES

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M Bronstein, Mathias Niepert, Bryan Perozzi, et al. 2025. Position: Graph Learning Will Lose Relevance Due To Poor Benchmarks. In *Proceedings of the 42nd International Conference on Machine Learning*.

Rocío Cañamares and Pablo Castells. 2020. On target item sampling in offline recommender system evaluation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 259–268.

Pablo Castells and Alistair Moffat. 2022. Offline recommender system evaluation: Challenges and new directions. *AI magazine* 43, 2 (2022), 225–238.

Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. 2023. Do We Really Need Complicated Model Architectures For Temporal Networks?. In *The Eleventh International Conference on Learning Representations*.

Filip Cornell, Oleg Smirnov, Gabriela Zarzar Gandler, and Lele Cao. 2025. On the Power of Heuristics in Temporal Graphs. In *I Can't Believe It's Not Better: Challenges in Applied Deep Learning*. https://openreview.net/forum?id=cTckUeh0Sw

Alexander Dallmann, Daniel Zoller, and Andreas Hotho. 2021. A case study on sampling strategies for evaluating neural sequential item recommendation models. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 505–514.

Manuel Dileo, Matteo Zignani, et al. 2024. Link prediction heuristics for temporal graph benchmark. In *ESANN 2024: Proceedings*. i6doc. com, 381–386.

Zifeng Ding, Yifeng Li, Yuan He, Antonio Norelli, Jingcheng Wu, Volker Tresp, Yunpu Ma, and Michael Bronstein. 2024. DyGMamba: Efficiently Modeling Long-Term Temporal Dependency on Continuous-Time Dynamic Graphs with State Space Models. *arXiv preprint arXiv:2408.04713* (2024).

Julia Gastinger, Christian Meilicke, Federico Errica, Timo Sztyler, Anett Schülke, and Heiner Stuckenschmidt. 2024. History Repeats Itself: A Baseline for Temporal Knowledge Graph Forecasting. In *IJCAI*.

Alessio Gravina, Giulio Lovisotto, Claudio Gallicchio, Davide Bacciu, and Claas Grohnfeldt. 2024. Long range propagation on continuous-time dynamic graphs. *arXiv preprint arXiv:2406.02740* (2024).

Miguel A Hernán, David Clayton, and Niels Keiding. 2011. The Simpson's paradox unraveled. *International journal of epidemiology* 40, 3 (2011), 780–785.

Balázs Hidasi and Ádám Tibor Czapp. 2023. Widespread flaws in offline evaluation of recommender systems. In *Proceedings of the 17th acm conference on recommender systems*. 848–855.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. 2021. OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs. *arXiv preprint arXiv:2103.09430* (2021).

Qiang Huang, Xin Wang, Susie Xi Rao, Zhichao Han, Zitao Zhang, Yongjun He, Quanqing Xu, Yang Zhao, Zhigao Zheng, and Jiawei Jiang. 2024b. Benchtemp: A general benchmark for evaluating temporal graph neural networks. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 4044–4057.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. 2024a. Temporal graph benchmark for machine learning on temporal graphs. *Advances in Neural Information Processing Systems* 36 (2024).

Ruoming Jin, Dong Li, Benjamin Mudrak, Jing Gao, and Zhi Liu. 2021. On estimating recommendation evaluation metrics under sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4147–4154.

Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1748–1757.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 1269–1278.

Dong Li, Ruoming Jin, Jing Gao, and Zhi Liu. 2020. On sampling top-k recommendation evaluation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2114–2124.

Dong Li, Ruoming Jin, Zhenming Liu, Bin Ren, Jing Gao, and Zhi Liu. 2023. Towards reliable item sampling for recommendation evaluation. In *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 37. 4409–4416.

Dong Li, Ruoming Jin, Zhenming Liu, Bin Ren, Jing Gao, and Zhi Liu. 2024. On item-sampling evaluation for recommender system. *ACM Transactions on Recommender Systems* 2, 1 (2024), 1–36.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2023a. On the consistency, discriminative power and robustness of sampled metrics in offline top-n recommender system evaluation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1152–1157.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2023b. What we evaluate when we evaluate recommender systems: Understanding recommender systems' performance using item response theory. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 658–670.

Tom M Mitchell. 1980. The need for biases in learning generalizations. (1980).

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. 2022a. Towards better evaluation for dynamic link prediction. *Advances in Neural Information Processing Systems* 35 (2022), 32928–32941.

Farimah Poursafaei and Reihaneh Rabbany. 2023. Exhaustive Evaluation of Dynamic Link Prediction. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1121–1130.

Farimah Poursafaei, Zeljko Zilic, and Reihaneh Rabbany. 2022b. A Strong Node Classification Baseline for Temporal Graphs. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 648–656.

Aniq Ur Rahman, Alexander Modell, and Justin Coon. 2025. Rethinking Evaluation for Temporal Link Prediction through Counterfactual Analysis. In *I Can't Believe It's Not Better: Challenges in Applied Deep Learning*. https://openreview.net/forum?id=TKydQh6kOc

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637* (2020).

Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. 2021. Quality metrics in recommender systems: Do we calculate metrics consistently?. In *Proceedings of the 15th ACM conference on recommender systems*. 708–713.

Lu Wang, Xiaofu Chang, Shuang Li, Yunfei Chu, Hui Li, Wei Zhang, Xiaofeng He, Le Song, Jingren Zhou, and Hongxia Yang. 2021a. Tcl: Transformer-based dynamic graph modelling via contrastive learning. *arXiv preprint arXiv:2105.07944* (2021).

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. 2021b. Inductive Representation Learning in Temporal Networks via Causal Anonymous Walks. In *International Conference on Learning Representations (ICLR)*.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations*.

Longqi Yang, Eugene Bagdasaryan, Joshua Gruenstein, Cheng-Kang Hsieh, and Deborah Estrin. 2018. Openrec: A modular framework for extensible and adaptable recommendation algorithms. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 664–672.

Lu Yi, Jie Peng, Yanping Zheng, Fengran Mo, Zhewei Wei, Yuhang Ye, Yue Zixuan, and Zengfeng Huang. 2025. TGB-Seq Benchmark: Challenging Temporal GNNs with Complex Sequential Dynamics. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=8e2LirwiJT

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. 2023. Towards better dynamic graph learning: New architecture and unified library. *Advances in Neural Information Processing Systems* 36 (2023), 67686–67700.

Wayne Xin Zhao, Zihan Lin, Zhichao Feng, Pengfei Wang, and Ji-Rong Wen. 2022. A revisiting study of appropriate offline evaluation for top-N recommendation algorithms. *ACM Transactions on Information Systems* 41, 2 (2022), 1–41.

Han Zhu, Junqi Jin, Chang Tan, Fei Pan, Yifan Zeng, Han Li, and Kun Gai. 2017. Optimized cost per click in taobao display advertising. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2191–2200.

A EXEMPLIFYING THE ISSUE OF BIASED NEGATIVE SAMPLING

To illustrate the effects of biased negative sampling, we draw on an observation from Cornell et al. [2025]. While many TGB link prediction benchmarks contain a high degree of edge repetition, `tgbl-review`, which consists of Amazon product reviews, stands out with a surprise index of 98.7%. This means that only 1.3% of historical edges reoccur. Despite this, historical edges are still used as negative samples.

In this setting, simple heuristics can rule out a large portion of negative samples by assigning low scores to edges that have appeared in the past. Machine learning models, such as GNNs and other embedding-based approaches, may however still assign high scores to these edges due to similarities in their learned representations. As a result, models can be penalized for favoring edges that heuristics would easily dismiss. On the other hand, a model that learns to suppress historical edges may achieve artificially high scores, producing an overly optimistic estimate when compared to uniform sampling of a similar size.

This example highlights that alignment between a model's inductive bias and the sampling strategy may play a more influential role than the alignment between the sampler and the dataset itself.