

# Enhancing Recommendation Systems' Performance with Network-Oriented Negative Sampling

Elif Ece Erdem<sup>1,\*</sup>, Günce Keziban Orman<sup>1</sup>

<sup>1</sup>Galatasaray University, Istanbul, Türkiye

## Abstract

Recommendation systems often struggle with data sparsity and class imbalance, affecting their ability to suggest relevant items. An efficient solution to these problems is negative sampling. This study introduces three novel network-oriented negative sampling strategies—Path-Length (PL-NS), Diffusion Distance (DD-NS), and Path-Density (PD-NS)—to enhance the GCN-based recommendation systems' performance. Integrated into the LightGCN framework and tested on the Last.fm dataset, our methods showed up to a 5% improvement in Precision@10, 6.7% in NDCG@10, and reduced training time by 30-35%, highlighting both performance and efficiency gains.

## Keywords

Recommendation Systems, Negative Sampling, Network Topology, Data Sparsity.

## 1. Introduction

Recommendation systems (RS) [1] are essential for aiding user decisions in various applications, from e-commerce to entertainment [2]. The effectiveness of these systems largely depends on the quality and quantity of user interaction data. In recent years, collaborative filtering (CF) algorithms, which utilize user-item interactions, have become popular for their superior RS performance [3], relying on explicit interactions like ratings or implicit ones like clicks to infer user preferences [2].

However, two key challenges hinder RS performance: data sparsity and class imbalance. Data sparsity occurs when users interact with only a small subset of available items, leading to incomplete data that limits the system's ability to provide personalized recommendations [4]. Class imbalance, where interacted items are vastly outnumbered by non-interacted ones, further complicates accurate learning of user preferences [5].

Negative sampling, the task of selecting items a user is unlikely to be interested in, has become a practical solution, especially in Graph Convolutional Network (GCN)-based CF methods [6]. The most common negative sampling approach, uniform random sampling (RNS), treats all non-interacted items equally, often leading to the selection of irrelevant items that do not contribute effectively to the learning [6]. However, RNS may fall short due to the abundance of irrelevant items.

Our study introduces novel negative sampling techniques that leverage the structure of user-item bipartite networks, aiming to identify more informative negative examples and improve the GCN-based recommendation systems' accuracy. The key contributions include: (1) proposing ten new methods that address RNS inefficiencies in sparse, imbalanced data; (2) integrating these methods into LightGCN, one of the top-performing GCN-based CF models [7]; (3) evaluating their impact on a well-known RS dataset, focusing on accuracy, convergence time, and comparisons with two baselines; and (4) analyzing the effects of combining our techniques with RNS at different levels of uniform selection.

The paper is organized as follows: Section 2 introduces our method, introducing Path-Length Negative Sampling (PL-NS), Diffusion Distance Negative Sampling (DD-NS), and Path-Density Negative Sampling (PD-NS). Section 3 presents our experiments and the results. Finally, Section 4 concludes with a summary and future research directions.

*Discovery Science - Late Breaking Contributions 2024*

\*Corresponding author.

✉ eceerdem@gsu.edu.tr (E. E. Erdem); korman@gsu.edu.tr (G. K. Orman)

ORCID 0009-0002-8219-2132 (E. E. Erdem); 0000-0003-0402-8417 (G. K. Orman)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

## 2. Method

We introduce three novel strategies for finding more informative negative examples than traditional methods based on different network topology aspects.

**Path-Length Negative Sampling (PL-NS):** This strategy prioritizes negative samples based on the shortest path distance within the user-item bipartite network. Items farther away from the user are assumed to be less relevant and more informative. We use two methods for selecting these negative samples. In **Long-Distance Weighted Selection**, items are weighted according to their distance from the user, with longer paths being prioritized. In **Quartile-Based Selection**, items are selected based on specific quartiles of the distance distribution, targeting items that are neither too close nor too far from the user.

**Diffusion Distance Negative Sampling (DD-NS):** This approach measures node similarity using diffusion distance [8], which reflects how information propagates through the network via random walks. The diffusion distance provides a nuanced measure of similarity that considers both direct and indirect connections. We propose several strategies for selecting negative samples based on diffusion distance.

In **Farthest Node Selection** and **Nearest Node Selection**, we select items with the highest and smallest diffusion distance from the user, respectively. In **Long-Distance Weighted Selection** and **Quartile-Based Selection**, similar to PL-NS, but applied to diffusion distance metrics.

**Path-Density Negative Sampling (PD-NS):** This strategy focuses on the density of paths between user and item nodes within the bipartite network. Path density is estimated using random walks, where the number of successful walks reaching an item from a user node indicates the path density. Two selection strategies are employed.

In **Max Path Density Selection** and **Min Path Density Selection**, we prioritize items with the highest and lowest path density, respectively, to identify less relevant and more informative negative samples.

## 3. Experiments and Results

We integrate our proposed negative sampling strategies into the LightGCN framework which is known for its simplicity and efficiency in recommendation tasks <sup>1</sup>. We tested these methods on the Last.fm dataset, a standard benchmark in recommendation systems, using Precision@10, Recall@10, and NDCG@10, comparing our techniques with traditional RNS, which selects negative samples randomly, and PNS (popularity-based negative sampling) [9], which favors frequently interacted items as negatives. The LightGCN model was configured as in its original implementation, with a maximum of 10 positive training examples per user, randomly selected for users with more interactions. Table 1 summarizes our experimental results, showing that our negative sampling strategies outperformed the baselines across all metrics, effectively improving recommendation performance.

The Long-Distance approach achieved the highest performance among the PL-NS methods, with a 4% improvement in Precision@10 and Recall@10 over the PNS baseline. The Quartile-Based Selection performed slightly lower, likely due to excluding informative negatives at other path lengths. In the DD-NS, the Long-Distance approach also led, with up to a 2.3% improvement in Precision@10 and a 2.5% boost in NDCG@10 compared to RNS. However, farthest-node and nearest-node selections underperformed, suggesting extreme diffusion distances may be less effective. The PD-NS generally underperformed against the baselines. Although the PD-Max approach showed some potential, it was not as effective as the other network-oriented strategies, indicating that path density alone may not strongly indicate negative sample relevance in this dataset. Combining proposed strategies with RNS, hybrid approaches showed mixed results: random negatives sometimes improved performance slightly in some cases, but often diluted the benefits of network-oriented methods. This suggests that while some randomness can be helpful, most negative samples should be selected based on network topology

---

<sup>1</sup>Our code is available at [https://github.com/tripledoubleE/network\\_oriented\\_NS](https://github.com/tripledoubleE/network_oriented_NS)

**Table 1**

Proposed NS Approach Results with Baselines in Last.fm Dataset

Strategy	Sampling Approach	Precision@10	Recall@10	NDCG@10
Baseline	RNS	0.0811	0.1483	0.1347
	PNS	0.0790	0.1435	0.1295
Path-Length	<b>PL-Long-Distance</b>	<b>0.0825</b>	<b>0.1496</b>	0.1351
	PL-Q1	0.0797	0.1450	0.1324
	PL-Q2	0.0771	0.1379	0.1250
Diffusion Distance	DD-NS-Farthest	0.0109	0.0234	0.0175
	DD-NS-Nearest	0.0161	0.0318	0.0255
	DD-NS-Q1	0.0379	0.0673	0.0552
	DD-NS-Q2	0.0375	0.0674	0.0560
	<b>DD-NS-Long-Distance</b>	<b>0.0830</b>	<b>0.1515</b>	<b>0.1382</b>
Path-Density	PD-Max	0.0667	0.1237	0.0186
	PD-Min	0.0804	0.1437	0.1314

for optimal performance. Besides our methods consistently achieved faster convergence than the RNS baseline, reducing training time by 30-35%, demonstrating that more informative negative sampling improves both recommendation accuracy and training efficiency.

## 4. Conclusion

In this study, we introduced ten novel methods leveraging the user-item bipartite network to prioritize more informative negatives. Integrating these into the LightGCN framework significantly improved recommendation accuracy and model convergence. Specifically, our best-performing method, Diffusion Distance-based Long-Distance Negative Sampling, achieved up to a 5% improvement in Precision@10 and 6.7% in NDCG@10 compared to traditional static negative sampling baseline, PNS.

## Acknowledgments

This work is supported by the Galatasaray University Research Fund (BAP) within the scope of project number FBA-2024-1240, titled *Bağlantı Tahmini Yöntemleri ve Sınıf Dengesizliği Sorununa Çözüm Yaklaşımları*, and the bilateral project of the Scientific and Technological Research Council of Türkiye (TÜBİTAK), under grant number 122N701, with CampusFrance, within the scope of Hubert Curien Partnerships (PHC) project number 49032VB.

## References

- [1] D. Kosiur, Understanding Policy-Based Networking, 2nd. ed., Wiley, New York, NY, 2001.
- [2] F. Ricci, L. Rokach, B. Shapira, Recommender Systems Handbook, 3rd ed., Springer New York, NY, 2022.
- [3] S. Wu, F. Sun, W. Zhang, X. Xie, B. Cui, Graph neural networks in recommender systems: A survey, ACM Comput. Surv. 55 (2022). doi:10.1145/3535101.
- [4] S. Natarajan, S. Vairavasundaram, S. Natarajan, A. H. Gandomi, Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data, Expert Systems with Applications 149 (2020) 113248. doi:10.1016/j.eswa.2020.113248.
- [5] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, Journal of Big Data (2019). doi:10.1186/s40537-019-0192-5.

- [6] Z. Yang, M. Ding, C. Zhou, H. Yang, J. Zhou, J. Tang, Understanding negative sampling in graph representation learning, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1666–1676. doi:10.1145/3394486.3403218.
- [7] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 639–648. doi:10.1145/3397271.3401063.
- [8] G. Bertagnolli, M. De Domenico, Diffusion geometry of multiplex and interdependent systems, *Phys. Rev. E* 103 (2021) 042301. doi:10.1103/PhysRevE.103.042301.
- [9] Z. Yang, M. Ding, X. Zou, J. Tang, B. Xu, C. Zhou, H. Yang, Region or global? a principle for negative sampling in graph-based recommendation, *IEEE Transactions on Knowledge and Data Engineering* 35 (2023) 6264–6277. doi:10.1109/TKDE.2022.3155155.