Exploration-Aided Downstream Graph Learning Tasks: A Survey on Exploratory Graph Learning

Yu Hou
School of Mathematics and Computing
(Computational Science and Engineering)
Yonsei University
Seoul, Republic of Korea
houyu@yonsei.ac.kr

Won-Yong Shin
School of Mathematics and Computing
(Computational Science and Engineering)
Yonsei University
Seoul, Republic of Korea
wy.shin@yonsei.ac.kr

***Abstract*—Graph learning is crucial for extracting meaningful information from graph-structured data, enabling effective solutions to various downstream tasks. However, existing methods for solving downstream graph learning tasks often rely on the availability of the graph structure, which may not always be accessible in real-world applications. To overcome this limitation, recent approaches have introduced *exploratory* learning techniques, which aim to tackle graph learning tasks on graphs with *unknown* topology. In this article, we provide a comprehensive overview of exploratory graph learning applied to two widely studied graph learning tasks: 1) influence maximization and 2) community detection. We delve into the problem formulation of both tasks concerning graphs with unknown topological information. Additionally, we explore the application of exploratory learning techniques to address these problems effectively.**

***Index Terms*—Community detection, exploratory learning, graph learning, influence maximization, unknown topology.**

I. Introduction

Graph learning is a fundamental technique for acquiring node or graph embeddings, while serving as a crucial tool for solving diverse downstream tasks on graphs, such as node classification [1], link prediction [2], influence maximization [3]–[5], and community detection [6]–[8]. The effectiveness of graph learning techniques relies heavily on the inherent structure of the underlying graph. For instance, connectivity information plays a vital role in extracting node embeddings, as interconnected nodes tend to share similar properties. By leveraging such embeddings, downstream tasks can be efficiently accomplished, leading to the improved overall performance.

Meanwhile, in real-world scenarios, complete access to the graph structure is often impractical, making existing graph learning methods ineffective in the absence of essential topology information. While one could consider investing additional efforts to uncover the entire graph structure before downstream applications, the process of collecting complete topological information proves to be prohibitively expensive and labor-intensive [9]. Consequently, this limitation has prompted the development of alternative approaches for addressing graph learning tasks in rather more feasible scenarios where the graph structure is incomplete or entirely unavailable.

The so-called *exploratory* learning has garnered significant attention to solve graph learning tasks in situations where the topological information of the underlying graph is unknown. This technique involves iteratively retrieving the neighbors of queried nodes within a predetermined query budget. Graph learning methods are then applied to accomplish the downstream tasks of interest with the help of the subgraph explored through these node queries, which serves as a surrogate for the underlying graph. By employing this node querying process, exploratory learning effectively overcomes the challenge of missing topology information, rendering it a valuable and effective strategy for graph learning in such situations.

Recent studies have shown that exploratory learning is a promising approach for tackling various graph learning tasks where the underlying topology information is unknown. Notably, exploratory learning has been applied to tasks such as influence maximization [10]–[13] and community detection [17], [18] when easy-to-collect node features are assumed to be available. First, influence maximization based on exploratory learning aims to identify a set of seed nodes from an explored subgraph that is expected to be as influential as the global optimal seed set, leveraging node features to enhance the identification of influential seed nodes. Second, community detection based on exploratory learning involves systematic exploration of multiple subgraphs using node queries to approximate the underlying graph structure. Leveraging node features enables to iteratively detect the community structure and aids to select more influential nodes to be queried. Consequently, this process continuously refines the outcome of community detection as the resulting subgraphs grow.

II. Basic Settings and Assumptions

Let us denote an underlying true graph, which is initially unavailable, as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of n nodes and $\mathcal{E}$ is the set of m edges. The graph $\mathcal{G}$ is assumed to be an undirected unweighted attributed graph without self-edges and repeated edges, having collectible node metadata (i.e., node features) $\mathcal{X} \in \mathbb{R}^{n \times d}$, where d is the dimension of each feature vector.

We assume a budget T of node queries. Upon querying a single node $v_{t \in [0, T-1]}$, we are able to discover its neighbors,

denoted as $\mathcal{N}_{\mathcal{G}}(v_t)$, and expand the observable subgraph accordingly. Let us denote the set of explored edges after t queries as $\mathcal{E}_t$. Specifically, during the $(t+1)$-th node query, we choose a node v_t from $\mathcal{V}$ to expand and update the set of explored edges $\mathcal{E}_{t+1} = \mathcal{E}_t \cup \mathcal{E}(\mathcal{N}_{\mathcal{G}}(v_t), v_t)$, where $\mathcal{E}(\mathcal{N}_{\mathcal{G}}(v_t), v_t)$ is a set of all edges to which each node in $\mathcal{N}_{\mathcal{G}}(v_t)$ and node v_t are incident. We also denote a subgraph and an inferred graph as $\mathcal{G}_{t+1} = (\mathcal{V}_{t+1}, \mathcal{E}_{t+1})$ and $\mathcal{G}^{(t+1)} = (\mathcal{V}, \mathcal{E}^{(t+1)})$, respectively, where $\mathcal{V}_{t+1} = \mathcal{V}_t \cup \mathcal{N}_{\mathcal{G}}(v_t) \cup v_t$, $\mathcal{E}^{(t+1)} = \mathcal{E}_{t+1} \cup \mathcal{E}'_{t+1}$, and $\mathcal{E}'_{t+1}$ is the inferred edges based on $\mathcal{E}_{t+1}$. Note that we can expand multiple subgraphs by selecting a queried node that is not connected to the currently explored subgraph.

III. Review on Graph Learning Tasks with Exploratory Learning

A. Problem Formulations

1) Exploratory learning-aided influence maximization: Let $f(S)$ denote the expected number of influenced nodes with the seed set $S \subseteq \mathcal{V}_t$. The objective function is formulated as:

$$(\mathcal{G}_T^*, S^*) = \underset{\mathcal{G}_T, S \subseteq \mathcal{V}_T, |S|=k}{\arg\max} f(S), \tag{1}$$

where k is the number of seed nodes.

2) Exploratory learning-aided community detection: Let $\mathbf{F}$ denote a non-negative weight affiliation matrix representing node-level community-affiliation embeddings. The objective function is formulated as:

$$(\mathbf{F}^*, \mathcal{Q}_T^*) = \underset{\mathbf{F} \geq 0, \mathcal{Q}_T \subset \mathcal{V}}{\arg\max} \mathbb{P}\left(\mathcal{G}^{(T)}, \mathcal{X} \,|\mathbf{F}\right), \tag{2}$$

where $\mathcal{Q}_T$ is the set of queried nodes and $\mathbb{P}\left(\mathcal{G}^{(T)}, \mathcal{X} \,|\mathbf{F}\right)$ is the likelihood to evaluate which affiliation embedding matrix $\mathbf{F}$ would make the given inferred graph $\mathcal{G}^{(T)}$ and node metadata $\mathcal{X}$ more probable.

B. Influence Maximization Using Exploratory Learning

Influence maximization using exploratory learning refers to the process of exploring subgraph $\mathcal{G}_T$ as a surrogate of the underlying graph by identifying a set of node queries. The objective is to maximize the spread of influence on $\mathcal{G}$ to identify the optimal seed nodes S^* from $\mathcal{G}_T$. The schematic overview of influence maximization using exploratory learning is illustrated in Fig. 1.

There have been several attempts to address the influence maximization problem in graphs with unknown topology using exploratory learning. Following the concept of active learning for classification [19], HEALER [20] was devised to address the dynamic influence maximization across a series of rounds, involving edge information collection after each round. The concept of exploratory influence maximization was introduced in [10] by providing a solution for querying individual nodes to retrieve their neighbors, leading to the construction of a subgraph $\mathcal{G}_t$. As follow-up studies, CHANGE [11] and Geometric-DQN [12] investigated the process of graph exploration node queries, utilizing the friendship paradox and

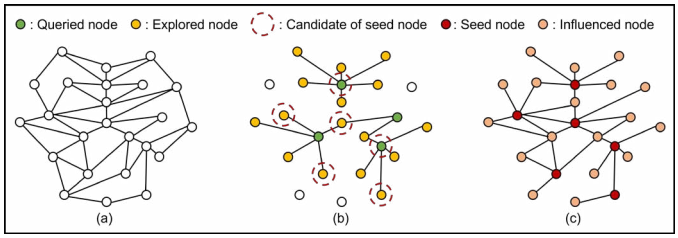

Fig. 1. The schematic overview of influence maximization using exploratory learning. (a) Underlying true graph $\mathcal{G}$. (b) Subgraph $\mathcal{G}_T$. (c) Seed nodes and influenced nodes.

patterns learned from a set of analogous graphs, respectively. A theoretical analysis on the performance of influence maximization was conducted in settings where a subgraph is retrieved via random node sampling [15], [16].

Now, let us focus on reviewing IM-META [13], pioneer work on leveraging the collected node metadata to aid the discovery of influential seed nodes, utilizing easy-to-collect node metadata. IM-META comprises two separate phases: graph exploration and seed set selection. During the graph exploration phase, we iteratively perform the following three steps. In Step 1, the relationship between node metadata in $\mathcal{X}$ and edges in the explored subgraph $\mathcal{G}_t$ is learned using a Siamese neural network model [14]. Specifically, the connectivity probabilities for unexplored edges are inferred by learning the similarity between nodes based on the currently explored edges retrieved from node queries and node metadata. The connectivity information can be learned by accurately capturing the homophily effect, which reveals the tendency of an individual node to associate with similar other nodes. In Step 2, a reinforced weighted graph is created by selecting a limited number of confidence edges whose edge probabilities exceeds a certain threshold. This edge selection can not only diminish noisy edges in the subsequent processes but also reduce the the computational complexity. In Step 3, a topology-aware ranking strategy is employed for query node selection. This is designed by measuring balances between the degree centrality of a target node and its geodesic distance to potential seeds. Then, the subgraph $\mathcal{G}_t$ is updated accordingly. This procedure is repeated until the T-th node query is reached. During the seed set selection phase, influential seed nodes are chosen using the greedy influence maximization algorithm [3].

C. Community Detection Using Exploratory Learning

Community detection using exploratory learning is a process involving iterative community detection and node query selection to explore (potentially multiple) subgraphs. The schematic overview of exploratory community detection is illustrated in Fig. 2.

As the first attempt, META-CODE [17] was first proposed to address community detection in graphs with unknown topology, utilizing easy-to-collect node metadata. META-CODE consists of three stages. In the first stage, an initial graph is inferred solely based on node metadata. In the second

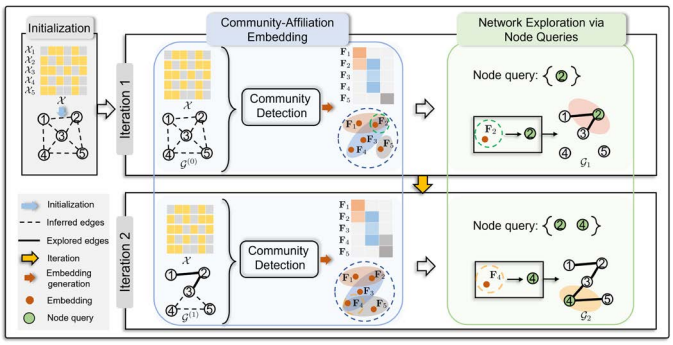

Fig. 2. The schematic overview of community detection using exploratory learning, where the first and second iterations are executed.

stage, graph representation learning based on a graph neural network (GNN) model [21] is performed to acquire community embeddings $\mathbf{F}$, by leveraging the node metadata and the inferred topological structure. In the third stage, a node to be queried aiming at faster graph exploration is selected based on two criteria: (i) the nodes lie within areas of overlapping communities and (ii) the selected nodes are distributed across diverse communities. Then, the inferred edges connected to the queried node are replaced with the explored edges $\mathcal{E}_t$, and new community-affiliation embeddings are generated by GNN-aid representation learning. This procedure is repeated until the T-th node query is reached. As a result, META-CODE provides increasingly improved community detection outcomes through exploratory learning.

Built upon the idea of META-CODE, a follow-up study [18] made full use of the explored edges for community detection in topologically unknown graphs. After exploring the neighbors of the queried node, connections between nodes in the unexplored portion of the underlying graph are further inferred by learning the connectivity information from the explored edges via a Siamese neural network model. The more accurate inferred graph $\mathcal{G}^{(t)}$, which incorporates the explored and inferred edges, is then used for community detection. This additional graph inference step enables to enhance the accuracy of community detection through exploratory learning.

IV. Conclusion

In this article, we have presented a comprehensive review of exploratory learning approaches for influence maximization and community detection in graphs with unknown topology. We summarized the key findings and contributions of exploratory learning in effectively solving influence maximization and community detection problems in such challenging yet realistic scenarios.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345, No. RS-2023-00220762) and by Institute of Information & communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00347, 6G Post-MAC (POsitioning- & Spectrum-aware intelligenT MAC for Computing & Communication Convergence)).

References

[1] Bhagat S, Cormode G, Muthukrishnan S. “Node classification in social networks,” Social network data analytics, 2011, 115-148.
[2] Lü L, Zhou T, “Link prediction in complex networks: A survey,” Physica A: statistical mechanics and its applications, vol. 390, no. 6, pp. 1150-1170, Mar. 2011.
[3] D. Kempe, J. Kleinberg, and ´E. Tardos, “Maximizing the spread of influence through a social network,” in Proc. KDD, 2003, pp. 137–146.
[4] Chen W, Wang Y, Yang S, “Efficient influence maximization in social networks,”, in Proc. KDD, 2009, pp. 199-208.
[5] Li Y, Fan J, Wang Y, et al, “Influence maximization on social graphs: A survey,” IEEE Trans. Knowl. Data Engineering, vol. 30, no. 10, pp. 1852-1872, Oct. 2018.
[6] X. Su, S. Xue, F. Liu, et al, “A comprehensive survey on community detection with deep learning,” IEEE Trans. Neural Networks Learn. Syst., pp. 1-21, Mar. 2022.
[7] Fortunato S, “Community detection in graphs,” Physics reports, vol. 486, no. 3-5, pp. 75-174, Feb. 2010.
[8] Xie J, Kelley S, Szymanski B K, “Overlapping community detection in networks: The state-of-the-art and comparative study,” ACM Computing Surveys, vol. 45, no. 4, pp. 1-35, Aug. 2013.
[9] T. W. Valente and P. Pumpuang, “Identifying opinion leaders to promote behavior change,” Health Educ. & Behav., vol. 34, no. 6, pp. 881–896, Dec. 2007.
[10] B. Wilder, N. Immorlica, E. Rice, and M. Tambe, “Maximizing influence in an unknown social network,” in Proc. AAAI, 2018, pp. 1–8.
[11] B. Wilder, L. Onasch-Vera, J. Hudson, J. Luna, N. Wilson, R. Petering, D. Woo, M. Tambe, and E. Rice, “End-to-end influence maximization in the field,” in Proc. AAMAS, 2018, pp. 1414–1422.
[12] H. Kamarthi, P. Vijayan, B. Wilder, B. Ravindran, and M. Tambe, “Influence maximization in unknown social networks: Learning policies for effective graph sampling,” in Proc. AAMAS, 2020, pp. 575–583.
[13] C. Tran, W.-Y. Shin, and A. Spitz, “IM-META: Influence maximization using node metadata in networks with unknown topology,” arXiv preprint arXiv:2106.02926, 2021.
[14] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a “Siamese” time delay neural network,” in Proc. NeurIPS, 1993, pp. 737–744.
[15] S. Eshghi, S. Maghsudi, V. Restocchi, S. Stein, and L. Tassiulas, “Efficient influence maximization under network uncertainty,” in Proc. Conf. Comput. Commun. Worksh., 2019, pp. 365–371.
[16] D. Eckles, H. Esfandiari, E. Mossel, and M. A. Rahimian, “Seeding with costly network information,” in Proc. Conf. Econ. Comput. (EC ’19), 2019, pp. 421–422.
[17] Y. Hou, C. Tran, and W.-Y. Shin, “META-CODE: Community detection via exploratory learning in topologically unknown networks,” in Proc. CIKM, 2022, pp. 4034–4038.
[18] Y. Hou, C. Tran, and W.-Y. Shin, “Graph neural network-aided exploratory learning for community detection with unknown topology,” arXiv preprint arXiv:2304.04497, 2023.
[19] B. Settles, “Active learning literature survey,” 2009.
[20] A. Yadav, H. Chan, A. X. Jiang, H. Xu, E. Rice, and M. Tambe, “Using social networks to aid homeless helters: Dynamic influence maximization under uncertainty.” in Proc. AAMAS, 2016, pp. 740–748.
[21] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. ICLR (Poster), 2017.