Research Statement

Yuan FANG
School of Computing and Information Systems
Singapore Management University
Tel: (65) 6808-5150; Email: yfang@smu.edu.sg
27 December 2022

Background & Overview

Graphs are prevalent in real-world datasets, for they can model not only individual data entities, but also interactions between these entities. Example graphs include the Web, social networks, transportation and telecommunication systems, scholarly citation networks, as well as protein interaction networks, entailing vast social, scientific, engineering and business significance. In particular, real-world graphs are often heterogeneous in nature, where there are different types of entity nodes, and different types of relationships between nodes. There could also be additional auxiliary information on nodes and edges, including structured attributes, unstructured texts or even audios and videos. Such graphs are often known as heterogeneous graphs or information networks.

To gain insights into such data, my research (Figure 1) has undertaken *learning and mining on graphs*. In particular, I focus on three sub-areas: designing and learning graph representations, multi-modal graph-based learning, and data efficiency and scalability for learning on graphs.

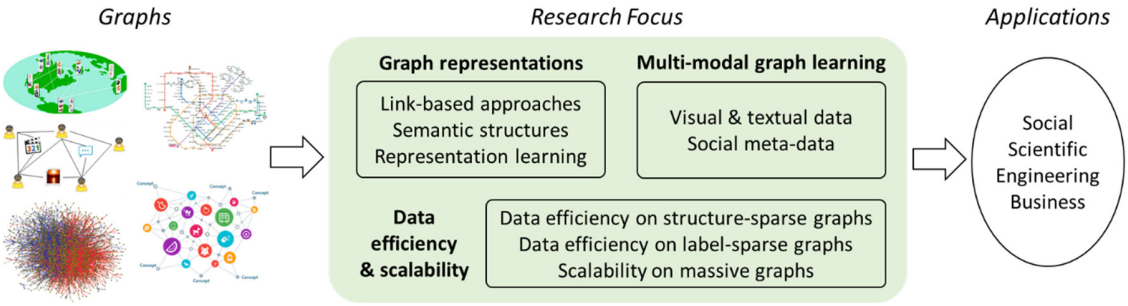

Figure 1: Overall research theme – Learning and mining on graphs.

Research Areas

A. Graph representations.

Link-based approaches. My earlier research towards my PhD dissertation mainly revolves around directly utilizing graph structures in learning, i.e., link-based approaches. Unlike traditional flat data, graph structures explain complex interactions between data entities, and thus are crucial towards data-driven tasks. Leveraging on link structures, we

investigated semi-supervised learning on graphs [ICML14]. The resulting graph-based probabilistic framework unifies the underlying principle in our previous random walk models [WSDM11, ICDE13]. We further considered heterogeneous graph structures [SIGIR12], as well as extended the learning objective on individual nodes to a set of nodes [ICDE16a]. In summary, link-based learning on graphs enable us to improve various tasks on graphs, including node classification and ranking, information extraction and data-driven crawling.

Semantic structure-based approaches. We also investigated higher-order semantic structures beyond simple link structures. In real-world scenarios, objects are often interlinked to form heterogeneous graphs, where different semantics exist between nodes. For instance, the below social network (Figure 2, left) contains users of different semantic relationships: some are classmates, some are family, and some are colleagues. The multitude of semantics arises from various types of nodes and their different interactions. We have proposed metagraph representations [ICDE16b] as a novel means to characterize these different semantic classes (Figure 2, right), which have shown very promising results in our studies on proximity ranking [ICDE16b] and node classification [Methods17]. Taking a step further, we have also explored metagraphs as a universal form of node and edge representations [TKDE19], demonstrating its superior performance in more downstream tasks including clustering and relationship prediction. Nevertheless, one limitation is the high computation complexity associated with the matching of metagraphs (i.e., counting of their instances) on large graphs. While using various heuristic pruning strategies can alleviate the problem, we plan to exploit machine learning-based method to predict the number of metagraph instances on graphs in future work.

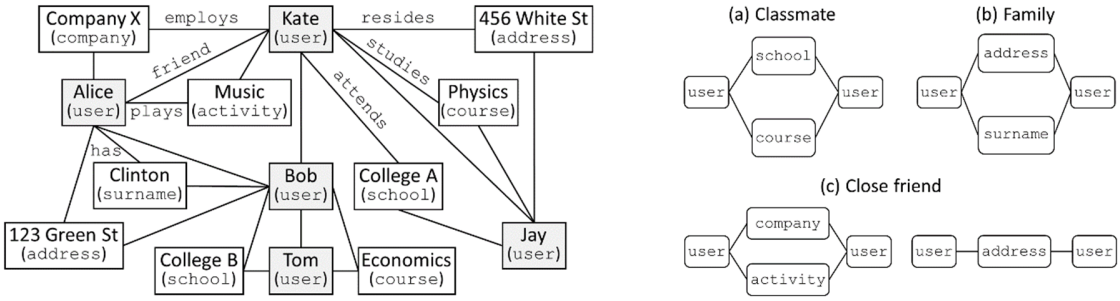

Figure 2. (left) Example heterogeneous social network. (right) Example metagraphs.

The community has also shown significant interest in leveraging metagraphs or similar semantic structures on heterogeneous graphs for representation learning through network embedding and graph neural networks. However, we observe that most prior approaches often under-utilize metagraphs, which are only applied in a pre-processing step and do not actively guide representation learning afterwards. In contrast, our recent work [TKDE20] proposes a novel framework called mg2vec, which learns the embeddings for metagraphs and nodes jointly. That is, metagraphs actively participate in the learning process by mapping themselves to the same embedding space as the nodes do, such that metagraphs are able to guide and constrain node embeddings during training.

General graph representation learning. We have also investigated the general problem of representation learning for different kinds of graphs, using different techniques. Specifically, we have studied neighborhood propagation [ICDM18] and adversarial learning [KDD19] for heterogenous graphs, neural attention mechanism [ECMLPKDD20a]

and Hawkes process [ECMLPKDD21] for dynamic graphs, the integration of structures and attributes on attributed graphs [IPM20, SIGIR20], complementary graph representations from multiple views on multi-view graphs [TKDD21, JBHI21], and the node-wise localization of graph neural networks [IJCAI21]. In summary, these works all leverage artificial neural networks for graph representation learning. Due to the ability to fit complex, nonlinear functions, neural networks-based graph representation learning often achieve state-of-the-art performance in various domains such as bioinformatics [BMC18, JBHI21] and recommendation systems [SDM20, ECMLPKDD20a].

B. Multi-modal graph-based learning.

Many problem statements often involve other kinds of data in addition to explicit graph structures, including visual and textual data and social meta-data. These data either enable us to construct new graphs, or to complement existing graphs to improve learning or to enable new tasks. We refer such research as multi-modal graph-based learning. Exploiting multi-modal data with graphs is a form of data enrichment to bring in knowledge that are not directly available from labeled data. There is a general consensus in the community that current machine learning approaches suffer from a significant knowledge gap. Additional knowledge from multi-modal data can potentially narrow the gap.

Vision & texts + Graph. In our work on object detection in images [IJCAI17], we exploit knowledge graphs to improve the visual detection task (Figure 3). In particular, knowledge graphs contain commonsense knowledge that relate different objects in images. An example piece of commonsense knowledge is that pets (e.g., cats) and furniture (e.g., table) often appear together. Such knowledge would improve detection recalls in home scenes: the detections of pet and furniture mutually reinforce each other, should one of them has low initial confidence. Alternatively, textual information, such as item descriptions in a recommendation scenario, can enrich the interactions between users and items to form a more dense graph structure, providing additional insights to boost recommendation performance [CIKM20a]. In another work, texts associated with graph nodes reveal multi-facet topical factors, which can guide finer-grained learning on graphs for both better model performance and interpretability [CIKM21b].

Figure 3. (left) An example task of object detection, to identify a dining table and a cat in the image. (right) Toy knowledge graph demonstrating the relationship between cats and tables.

Social meta-data + Graph. In our entity linking study [TACL14], we construct spatial and temporal graphs for entities appearing on Twitter, so that entities that are close to each other in either space or time are connected based on the meta-data of the tweets (i.e., timestamp and geotagging of the tweet). The connections reveal the relatedness between entities, which proves beneficial to the entity linking task on Twitter. In a more recent work,

we attempt to enrich collaborative filtering with a novel form of social meta-data known as "friend referral circle" [ECMLPKDD20b], where users are recommended with items liked or shared by their circle of friends. Leveraging the unique friend referral circle enables a more accurate modeling of social factors (e.g., user behaviors are more influenced by their friends who appears more authoritative), beyond just the homophily effect assumed in conventional social recommendation.

C. Data efficiency and scalability on graphs.

Learning with data efficiency has always been an important research topic, and has gained particular traction in recent years due to the rise of deep learning which often require large-scale, high-density data for optimal performance. To address the over-reliance on data, in particular on graphs, we have explored several directions of data efficiency. Besides, we have also studied the more conventional computational scalability problem on massive graphs.

Data efficiency on structure-sparse graphs. First, for very sparse graphs where the number of edges relative to the number of nodes is extremely small, we investigate a dual dropout strategy on both nodes and edges for graph neural networks [BIOINF20], to increase the overall robustness of learning. Second, even if a graph is considered dense on the whole, there still exist tail nodes with very few links. In other words, an individual tail node has very scarce structural connectivity, despite the abundance of links on other nodes. In general, the node degrees vary considerably across the network and are not uniformly distributed (Figure 4, left). The lack of structural connectivity on a tail node makes its representation more difficult to learn than nodes with abundant links (Figure 4, right). Representation learning for the tail nodes is thus a challenging and novel problem. Inspired by meta-learning, we formulate the problem as a few-shot regression task in our work meta-tail2vec [CIKM20b], a first attempt on this problem to our best knowledge. However, meta-tail2vec is a two-stage method that improves the tail node embedding through a post-processing step. Thus, we further propose an end-to-end tail node representation learning framework for graph neural networks [KDD21a]. Similarly, cold-start recommendation also suffers from the sparsity of links on new users and items. Thus, we formulate the cold-start problem as a few-shot link prediction task, and addressed it under the meta-learning paradigm as well [KDD20].

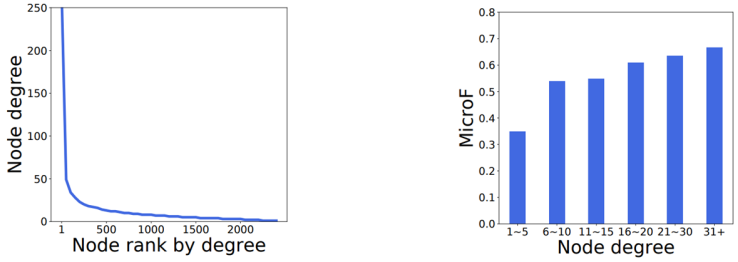

Figure 4. Relationship between node degree and the quality of learned representation on a typical graph. (left) Node degree distribution. (right) Classification performance w.r.t. node degrees.

Data efficiency on label-scarce graphs. Like any other supervised machine learning models, state-of-the-art graph neural networks rely on a large number of labels for good performance. However, in reality, many tasks often lack abundant labeled data. One common scenario is the few-shot node classification task on a graph, in which some novel

classes only have one or few examples. For instance, on a scholarly citation network, while Markov chains is a mature topic with many labeled examples, algorithmic explainability and fairness is relatively new with few labeled examples. To address few-shot learning on graphs, we resort to the framework of meta-learning, while simultaneously exploiting graph-specific characteristics including the long-range dependencies between nodes, and the global graph contexts [AAAI21a].

In another line, inductive graph learning trains a model on similar graphs, and the trained model can be generalized to a new graph in the same feature space. The inductive ability of the trained model effectively reduces the need of labels on the new graph, as opposed to transudative graph learning where the labels on the new graph are crucial as no information from other graphs can be leveraged. In our meta-inductive model [SIGIR21], we exploit both optimization-based and hypernetwork-based meta-learning to enhance the inductive ability on node classification. To take one step further, we exploit the pre-training of graph neural networks, which aims to learn a transferable prior from one or more graphs, to enable multiple downstream tasks with limited labeled data. In particular, the pre-trained model can generalize to different downstream tasks after a quick fine-tuning process on each task with a small number of labels. Some of our works [KDD21, CIKM21b] attempt to design better pre-training objectives to improve the capturing of the transferable prior on graphs. We further note that conventional pre-training is decoupled from fine-tuning, causing a divergence between their optimization objectives. Thus, we propose the concept of learning to pre-train [AAAI21b], by integrating and aligning the pre-training and fine-tuning stages under a meta-learning framework.

Scalability of learning on graphs. Finally, we explore scalable computational frameworks for massive graphs. Our earlier work [VLDB13, VLDBJ15] investigated fast approximation algorithms for computing Personalized PageRank on large graphs. The algorithm can speed up over existing methods by several folds with high accuracy and excellent scalability. More recently, as network embedding and graph neural networks emerge as the *de facto* standard on graphs, we have attempted to accelerate graph representation learning via importance sampling on large-scale heterogeneous graphs [TKDD20]. The key idea is to design an effective sampler that is aware of the multitude of node and edge types and their complex inter-dependence.

Select Publications and Outputs

A. Graph representations.

[WSDM11] **Y. Fang** and K. C.-C. Chang. “Searching Patterns for Relation Extraction over the Web: Rediscovering the Pattern-Relation Duality.” In *WSDM* 2011, pp. 825–834.

[SIGIR12] **Y. Fang**, P. Hsu and K. C.-C. Chang. “Confidence-Aware Graph Regularization with Heterogeneous Pairwise Features.” In *SIGIR* 2012, pp. 951–960.

[ICDE13] **Y. Fang**, K. C.-C. Chang and H. W. Lauw. “RoundTripRank: Graph-based Proximity with Importance and Specificity.” In *ICDE* 2013, pp. 613–624.

[ICML14] **Y. Fang,** K. C.-C. Chang and H. W. Lauw. “Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically.” In *ICML* 2014, Part 2, pp. 406–414.

[ICDE16a] **Y. Fang,** V. W. Zheng and K. C.-C. Chang. “Learning to Query: Focused Web Page Harvesting for Entity Aspects.” In *ICDE* 2016, pp. 1002–1013.

[ICDE16b] **Y. Fang**, W. Lin, V. W. Zheng, M. Wu, K. C.-C. Chang and X.-L. Li. “Semantic Proximity Search on Graphs with Metagraph-based Learning.” In *ICDE* 2016, pp. 277–288.

[Methods17] S. Kircali, **Y. Fang**, M. Wu, X. Xiao and X. Li. “Disease Gene Classification with Metagraph Representations.” *Methods* 131:83–92, 2017.

[ICDM18] V. W. Zheng, M. Sha, Y. Li, H. Yang, **Y. Fang**, Z. Zhang, K.-L. Tan and K. C.-C. Chang. “Heterogeneous Embedding Propagation for Large-scale E-Commerce User Alignment.” In *ICDM* 2018, pp. 1434–1439.

[BMC18] S. Kircali, L. Ou-Yang, **Y. Fang**, C.-K. Kwoh, M. Wu and X. Li. “Integrating Node Embeddings and Biological Annotations for Genes to Predict Disease-Gene Associations.” *BMC Systems Biology* 12(Supp 9): 31–44, 2018.

[KDD19] B. Hu, **Y. Fang** and C. Shi. “Adversarial Learning on Heterogeneous Information Networks.” In *KDD* 2019, pp. 120–129.

[TKDE19] **Y. Fang**, W. Lin, V. W. Zheng, M. Wu, J. Shi, K. C.-C. Chang and X. Li. “Metagraph-based Learning on Heterogeneous Graphs.” *IEEE TKDE* 33(1):154–168, 2019.

[TKDE20] W. Zhang, **Y. Fang**, Z. Liu, M. Wu and X. Zhang. “mg2vec: Learning Relationship-Preserving Heterogeneous Graph Representations via Metagraph Embedding.” *IEEE TKDE*, 2020.

[SDM20] X. Jiang, B. Hu, **Y. Fang**, C. Shi. “Multiplex Memory Network for Collaborative Filtering.” In *SDM* 2020, pp. 91–99.

[ECMLPKDD20a] Y. Ji, M. Yin, **Y. Fang**, H. Yang, X. Wang, T. Jia and C. Shi. “Temporal Heterogeneous Interaction Graph Embedding For Next-Item Recommendation.” In *ECML-PKDD* 2020, Part III, pp. 314--329.

[IPM20] Y. Ji, C. Shi, **Y. Fang**, X. Kong and M. Yin. Semi-supervised Co-Clustering on Attributed Heterogeneous Information Networks. *IPM* 57(6):102338, 2020.

[SIGIR20] W. Huang, Y. Li, **Y. Fang,** J. Fan and H. Yang. BiANE: Bipartite Attributed Network Embedding.” In *SIGIR* 2020, pp. 149–158.

[TKDD21] S. Ata, **Y. Fang**, M. Wu, J. Shi, C. Kwoh, X. Li. Multi-View Collaborative Network Embedding. In *ACM TKDD* 15(3), 2021.

[JBHI21] Z. Hao, D. Wu, Y. Fang, M. Wu, R. Cai and Xiao-Li Li. Prediction of Synthetic Lethal Interactions in Human Cancers using Multi-view Graph Auto-Encoder. *IEEE Journal of Biomedical & Health Informatics* 25(10), 2021, pp. 4041--4051.

[ECMLPKDD21] Y. Ji, T. Jia, **Y. Fang** and C. Shi. Dynamic Heterogeneous Graph Embedding via Heterogeneous Hawkes Process. In *ECML-PKDD* 2021, Part I, pp. 388–403.

[IJCAI21] Z. Liu, **Y. Fang**, C. Liu and S. C. H. Hoi. Node-wise Localization of Graph Neural Networks. In *Int’l Joint Conf. on Artificial Intelligence* (IJCAI) 2021, pp. 1520–1526.

B. Multi-modal graph-based learning.

[TACL14] **Y. Fang** and M.-W. Chang. “Entity Linking on Microblogs with Spatial and Temporal Signals.” *TACL* Vol. 2, October 2014, pp. 259-272. *Invited for oral presentation at EMNLP 2014.*

[IJCAI17] **Y. Fang**, K. Kuan, J. Lin, C. Tan and V. Chandrasekhar. “Object Detection Meets Knowledge Graphs.” In *IJCAI* 2017, pp. 1661–1667.

[CIKM20a] Y.-N. Chuang, C.-M. Chen, C.-J. Wang, M.-F. Tsai, **Y. Fang** and E. P. Lim. TPR: Text-aware Preference Ranking for Recommender Systems. In *CIKM* 2020, pp. 215–224.

[ECMLPKDD20b] Y. Lu, R. Xie, C. Shi, **Y. Fang**, W. Wang, X. Zhang and L. Lin. “Social Influence Attentive Neural Network for Friend-Enhanced Recommendation.” In *ECML-PKDD* 2020, Part IV (Applied Data Science), pp. 3–18.

[CIKM21a] S. Xu, C. Yang, C. Shi, **Y. Fang**, Y. Guo, T. Yang, L. Zhang and M. Hu. Topic-aware Heterogeneous Graph Neural Network for Link Prediction. In *CIKM* 2021, pp. 2261–2270.

C. Data efficiency and scalability on graphs.

[VLDB13] F. Zhu, **Y. Fang**, K. C.-C. Chang and J. Ying. “Incremental and Accuracy-Aware Personalized PageRank through Scheduled Approximation.” In *VLDB* 2013, pp. 481–492. *Extended version invited to the collection of best papers of VLDB.*

[VLDBJ15] F. Zhu, **Y. Fang**, K. C.-C. Chang and J. Ying. “Scheduled Approximation for Personalized PageRank with Utility-Driven Hub Selection.” *VLDBJ* 24(5):655–679, 2015.

[KDD20] Y. Lu, Y. Fang and C. Shi. “Meta-learning on Heterogeneous Information Networks for Cold-start Recommendation.” In *KDD* 2020, pp. 1563–1573.

[BIOINF20] R. Cai, X. Chen, **Y. Fang**, M. Wu and Y. Hao. Dual-Dropout Graph Convolutional Network for Predicting Synthetic Lethality in Human Cancers. *Bioinformatics* 36(16):4458–4465, 2020.

[CIKM20b] Z. Liu, W. Zhang, **Y. Fang**, X. Zhang and S. C. H. Hoi. Towards Locality-Aware Meta-Learning of Tail Node Embeddings on Networks. In *CIKM* 2020, pp. 975–984.

[TKDD20] Y. Ji, M. Yin, H. Yang, J. Zhou, V. W. Zheng, C. Shi and **Y. Fang**. Accelerating Large-Scale Heterogeneous Interaction Graph Embedding Learning via Importance Sampling. *ACM TKDD* 15(1), 2020.

[AAAI21a] Y. Lu, X. Jiang, **Y. Fang** and C. Shi. Learning to Pre-train Graph Neural Networks. In *AAAI* 2021, pp. 4276–4284.

[AAAI21b] Z. Liu, **Y. Fang**, C. Liu and S. C. H. Hoi. Relative and Absolute Location Embedding for Few-Shot Node Classification on Graph. In *AAAI* 2021, pp. 4267–4275.

[KDD21a] Z. Liu, K. Nguyen and **Y. Fang**. Tail-GNN: Tail-Node Graph Neural Networks. In *KDD* 2021, pp. 1109–1119.

[KDD21b] X. Jiang, T. Jia, C. Shi, Y. Fang, Z. Lin and H. Wang. Pre-training on Large-Scale Heterogeneous Graph. In *KDD* 2021, pp. 756–766.

[SIGIR21] Z. Wen, Y. Fang and Z. Liu. Meta-Inductive Node Classification across Graphs. In *SIGIR* 2021, pp. 1219–1228.

[CIKM21b] X. Jiang, Y. Lu, **Y. Fang** and C. Shi. Contrastive Pre-training of GNNs on Heterogeneous Graphs. In *CIKM* 2021, pp. 803–812.