How to Publish in a Top Tier Conference about Machine Learning on Heterogeneous Graphs?

Meng Jiang

I work on graph data mining and graph machine learning. But I am **NOT** an expert on mining or learning heterogeneous graph data. However, I am often asked by graduate students about "how to publish in a top tier conference about heterogeneous graphs." Multiple reasons that I receive such questions are:

- Heterogeneity is one of the most challenging aspects of modeling real-world graph data. Students understand the scientific and industrial value of studying heterogeneous graphs. The research experience is helpful for finding a competitive job in the market.
- It is truly a popular concept in data mining and machine learning. Several papers are receiving a great number of citations.
- I teach the courses of "Data Science" and "Computational Behavior Modeling" that cover the topics of representation learning from graph data. I introduce the challenges and a few existing methods on modeling heterogeneous graph data.

To answer the question, I searched "heterogeneous graph" and "heterogeneous information network" in Google Scholar, collected examples, and shared my observations in this blog. I list seven pairs of papers published in or after 2016 as follows. In the pairs, the first paper has "heterogeneous" in the title; the other paper does not emphasize the heterogeneity. The first paper can be considered as a nontrivial extension of the other work for heterogeneous graphs. I manually summarize the framework types, such as network embedding, GNN architectures, etc.

(1) Network embedding:

Dong et al. "metapath2vec: Scalable representation learning for **heterogeneous** networks." KDD 2017.
Grover et al. "node2vec: Scalable feature learning for networks." KDD 2016.

(2) GNN architecture 1: Sampling-based GNN:

Zhang et al. "**Heterogeneous** graph neural network." KDD 2019.
Hamilton et al. "Inductive representation learning on large graphs." NeurIPS 2017.

(3) GNN architecture 2: Graph attention network

Wang et al. "**Heterogeneous** graph attention network." WWW 2019.
Veličković et al. "Graph Attention Networks." ICLR 2017.

(4) GNN+GAN:

Hu et al. "Adversarial learning on **heterogeneous** information networks." AAAI 2019.
Wang et al. "Graphgan: Graph representation learning with generative adversarial nets." AAAI 2018. (Microsoft)

(5) GNN+Infomax:

Ren and Liu. "**Heterogeneous** deep graph infomax." Workshop of Deep Learning on Graphs: Methodologies and Applications co-located with AAAI 2020.
Velickovic, et al. "Deep graph infomax." ICLR, 2019.

(6) Graph architecture 3: Transformer

Hu et al. “**Heterogeneous** graph transformer.” WWW 2020.
Yun et al. “Graph transformer networks.” NeurIPS 2019.

(7) GNN+Pre-training:

Wang et al. “Self-supervised **heterogeneous** graph neural network with co-contrastive learning.” KDD 2021.
Qiu et al. “Gcc: Graph contrastive coding for graph neural network pre-training.” KDD 2020.

My observations (and suggestions) are as follows:

- Most of the “first papers” are published within one or two years after the second paper. That means, when a new framework is out, researchers work really fast to create an extension of it for heterogeneous graphs. The success rate of such extensions (in terms of experimental results) is very high: maybe because the authors are well experienced and/or very lucky. And the paper’s presentation quality is good for acceptance.
- If you are interested in doing such kind of research and writing such papers, you must be able to:
    - Follow the state-of-the-art graph learning frameworks/algorithms extensively;
    - Design a solution that addresses the heterogeneity issue;
    - Design and perform experiments on standard/popular graph datasets;
    - Write professionally.
- Keep in mind that it’d be very difficult to get to be the first accepted work, when a number of researchers are writing and submitting papers of the same or similar topics. Not all such papers have high citations. And quite a few have no opportunity to be exposed in any conferences or journals but just be accessible on arxiv.
- Personal opinion: When a graph machine learning algorithm is already extended for heterogeneous graphs and scaled up for web-scale data/systems, it would be hard to propose new fundamental algorithms, and there would be mainly two directions to go:
    - Uncovering the theory behind the uninterpretable algorithms;
    - Specializing techniques for specific applications that have more complex data: when heterogeneity meets sparsity, lack of labels, imbalance, dynamics, temporal patterns, fairness, etc.

Again, I do not publish on this topic. So my suggestions could be wrong. Good luck!!!

about 680 words
March 4, 2023 (created)
March 4, 2023 (completed)