GPSP: Graph Partition and Space Projection based Approach for Heterogeneous Network Embedding

Wenyu Du*, Shuai Yu*, Min Yang*, Qiang Qu*, Jia Zhu**

*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

*{wy.du, shuai.yu, min.yang, qiang}@siat.ac.cn

**School of Computer Science, South China Normal University

**jzhu@m.scnu.edu.cn

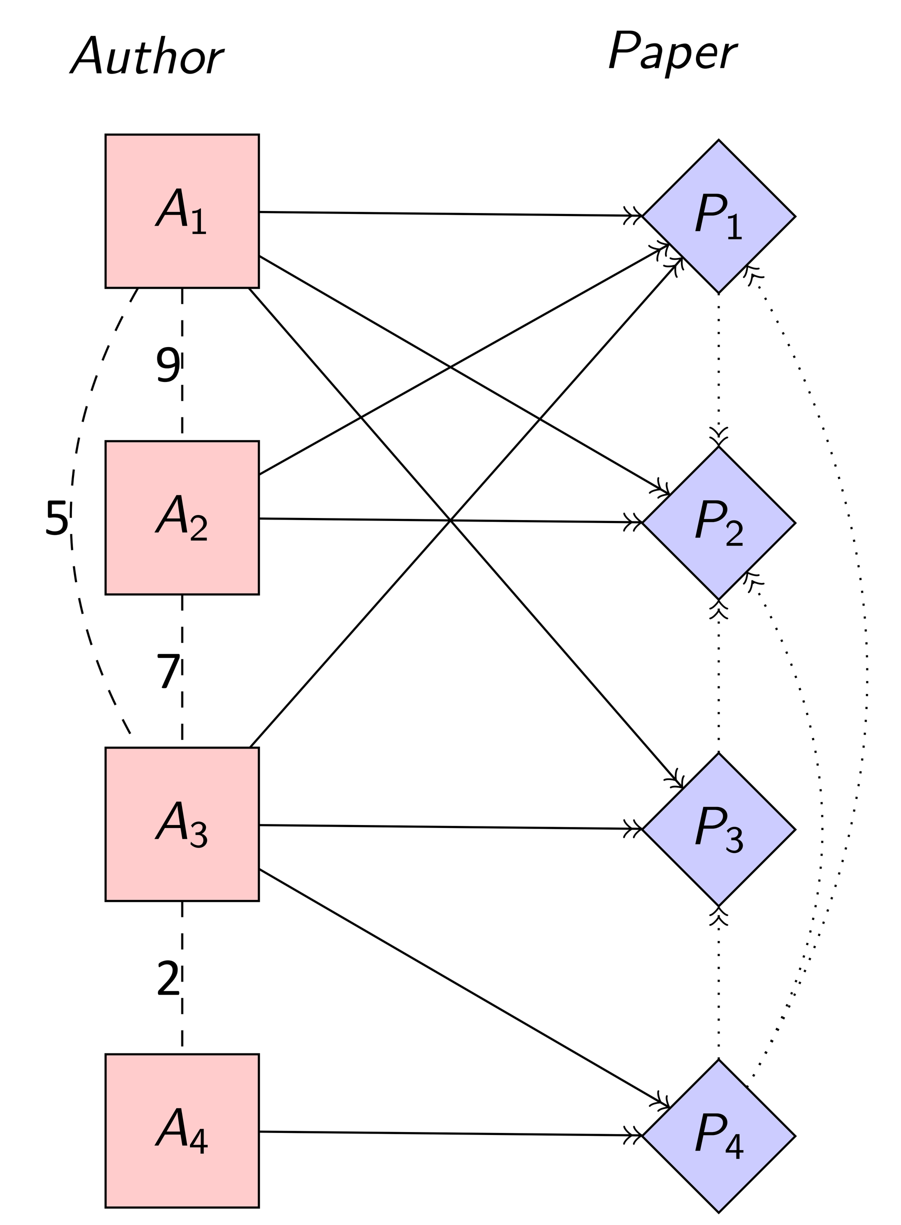
Introduction

In this paper, we propose *GPSP*, a novel **G**raph **P**artition and **S**pace **P**rojection based approach, to learn the representation of a heterogeneous network that consists of multiple types of nodes and links. Concretely, we first partition the heterogeneous network into homogeneous and bipartite subnetworks. Then, the projective relations hidden in bipartite subnetworks are extracted by learning the projective embedding vectors. Finally, we concatenate the projective vectors from bipartite subnetworks with the ones learned from homogeneous subnetworks to form the final representation of the heterogeneous network. Extensive experiments are conducted on a real-life dataset. The results demonstrate that *GPSP* outperforms the state-of-the-art baselines in two key network mining tasks: node classification and clustering.

Model

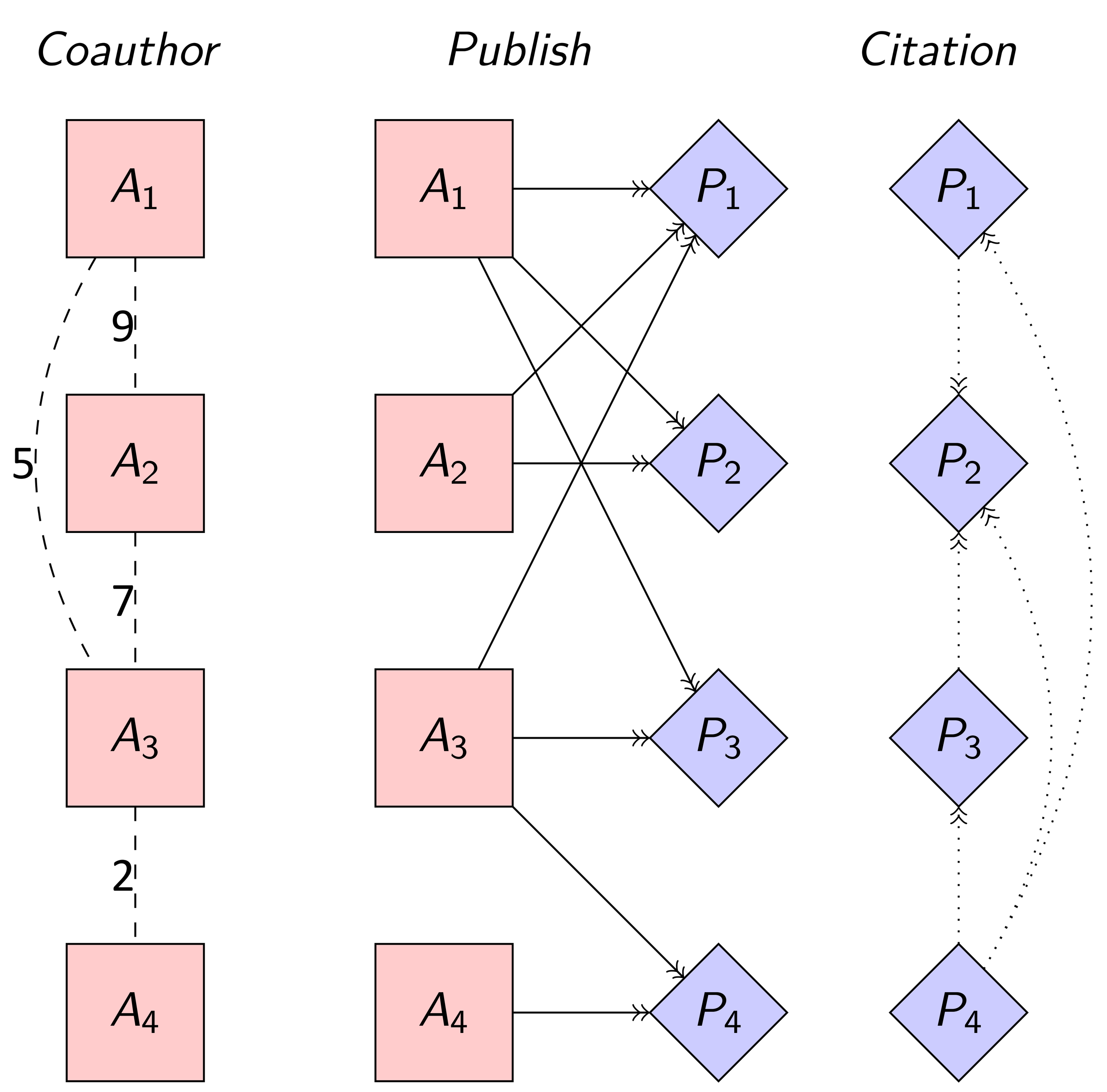
Heterogeneous Network Embedding

- Network embedding, or network representation learning, is the task of learning latent representation that captures the internal relations of rich and complex network-structured data.
- In practice, many networks are often heterogeneous, i.e., involving multiple types of nodes and relationships.



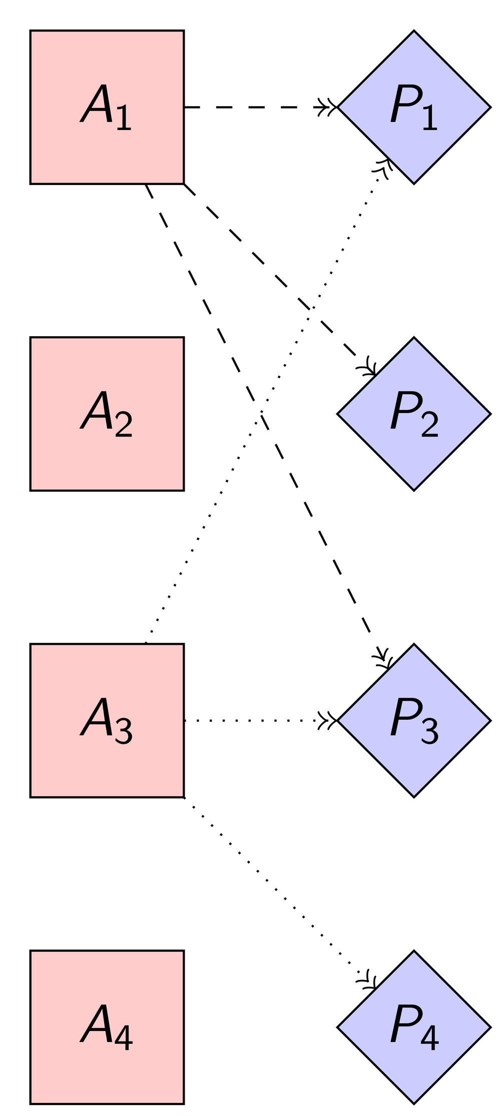
Graph Partition

- Firstly, an edge-based graph partition method is used to partition the heterogeneous network into two types of atomic subnetworks:
  1. Homogeneous networks that contain singular type of nodes and relationships.
  2. Bipartite networks that contain two types of vertices and one type of relationship.



Space Projection

- Second, we apply classic network embedding models [1,2] to learn the representations of homogeneous subnetworks.
- Third, for each bipartite subnetwork, the hidden projective relations are extracted by learning the projective embedding vectors for the related types of nodes.
- Finally, GPSP concatenates the projective node vectors from bipartite subnetworks with the homogeneous node vectors.



Experiments

Multi-label node classification

Metric	Model	10%	30%	50%	70%	90%
Micro-F1	LINE	0.7062	0.7067	0.7074	0.7062	0.7075
	DeepWalk	0.6992	0.7010	0.6992	0.6986	0.6988
	metapath2vec	0.6546	0.6549	0.6547	0.6552	0.6529
	metapath2vec++	0.6692	0.6681	0.6676	0.6677	0.6651
	GPSP-LINE	**0.7512**	**0.7557**	**0.7564**	**0.7554**	**0.7552**
	GPSP-DeepWalk	**0.7275**	**0.7318**	**0.7324**	**0.7320**	**0.7318**
Macro-F1	LINE	0.7032	0.7036	0.7043	0.7035	0.7036
	DeepWalk	0.6964	0.6982	0.6965	0.6963	0.6961
	metapath2vec	0.6307	0.6313	0.6322	0.6328	0.6301
	metapath2vec++	0.6478	0.6473	0.6478	0.6473	0.6445
	GPSP-LINE	**0.7482**	**0.7527**	**0.7534**	**0.7526**	**0.7522**
	GPSP-DeepWalk	**0.7253**	**0.7290**	**0.7298**	**0.7295**	**0.7289**

Node Clustering

LINE	DeepWalk	metap2v	metap2v++	GPSP-LINE	GPSP-DeepWalk
0.2516	0.2873	0.2403	0.2470	**0.3118**	**0.3555**

Dataset

We construct an academic heterogeneous network, based on the dataset from AMiner Computer Science [3]. The constructed network consists of two types of nodes and three types of relations.

- Computer Scientist (Author)
- Paper
- Authors coauthor with each other
- Authors write papers
- Papers cite other papers

References

1. Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. *Deepwalk: Online learning of social representations.* In SIGKDD. 701–710.
2. Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. *Line: Large-scale information network embedding.* In WWW. 1067–1077.
3. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. *Arnetminer: extraction and mining of academic social networks.* In SIGKDD. 990–998.