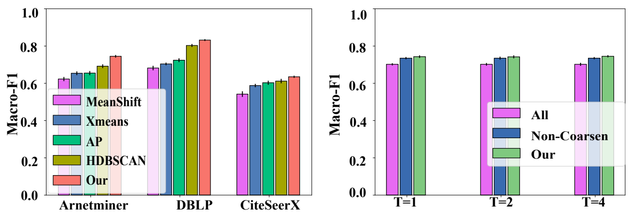

Figure 3: The performance of (Left) different clustering algorithms and (Right) alternative methods.

- Zhang et al. [23] use a Bayesian non-exhaustive classification framework for name disambiguation.
- Zhang et al. [24] use the PTE method to learn paper representations and use a hierarchical agglomerative clustering algorithm to cluster papers.

As it is shown in Table 2, our method **outperforms** all the other methods. For all the datasets, our method is about 7% to 50% better than others, and it is about 7% to 17% better than the second best results.

4.2 Network embedding methods

We evaluate the performance of different network embedding methods. For a fair comparison, the same network construction and clustering method proposed in this work is used. Further, we augment DeepWalk, LINE, and Node2Vec to incorporate heterogeneous networks by concatenating vectors learned from different networks as the final paper representations.

Table 2 summaries the results. Our method is 9% to 31% better than all the others. PTE and Hin2Vec obtain the second and third best performance due to their ability to encode multiple networks directly while the other methods (DeepWalk, LINE, Node2Vec, CANE) cannot. Our method performs better than PTE and Hin2Vec because we focus on the gap between positive and negatives edges, which is suitable for disambiguation task, and our method captures global graph properties via learning on coarsened networks.

4.3 Clustering methods

We evaluate the performance of different clustering algorithms for disambiguation with the proposed network embedding method. The clustering algorithms used for comparison, which does not require the input of the number of clusters, are HDBSCAN, AP, MeanShift [5], and Xmeans [6].

Figure 3 (Left) depicts the results. Our method is about 8% to 30% better than the other clustering methods. Our method outperforms HDBSCAN and AP, and can select better clustering results most of the time. Moreover, it has stable performance (the lowest standard deviation).

4.4 Alternatives and Sensitivity Results

When learning representations, for a network with $|E|$ edges, $T \times |E|$ triples are sampled. We evaluate the impact of T by varying it from 1 to 4. The result for the Artminer dataset is shown in Figure 3 (Right). It can be observed that T does not impact the performance significantly.

For each ambiguous name, we create networks based on papers whose author-list contains the name. An alternative way is to create big networks based on all the papers. The result is shown in Figure 3 (Right) with the legend *All*. Our method is about 6% better than the *All* method for the Arnetminer dataset. We use graph coarsening to learn the global network structure. As shown in Figure 3 (Right), with the coarsened network, the Macro-F1 score of our method is improved by about 2% comparing to Non-Coarsen.

5 CONCLUSION

In this work, we propose a network embedding method to learn the representations of papers for author disambiguation. The method can learn heterogeneous relationships between papers and can be easily adapted to many other scenarios. Furthermore, we propose a clustering algorithm which assigns papers to distinct authors. Through extensive experiments, we show that our method can significantly outperform state-of-the-art methods.

Acknowledgement: This work was partially supported by the National Basic Research Program of China (No. 2014CB340303), the National Science Foundation for Young Scholars of China (No. 61602500, 61502500, 61702529, 61702530, and 61402509).

REFERENCES

[1] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *PAKDD*.

[2] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. GraRep: Learning Graph Representations with Global Structural Information. In *CIKM*.

[3] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. 2018. HARP: Hierarchical Representation Learning for Networks.. In *AAAI*.

[4] Wei Sheng Chin, Yu Chin Juan, and Wei Cheng. 2013. Effective string processing and matching for author disambiguation. In *KDD Cup*.

[5] Dorin Comaniciu and Peter Meer. 2002. Mean shift: a robust approach toward feature space analysis. *PAMI* 24, 5 (2002), 603–619.

[6] Pelleg Dan and Andrew W. Moore. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *ICML*.

[7] Brendan J. Frey and Delbert Dueck. 2007. Clustering by Passing Messages Between Data Points. *Science* 315, 5814 (2007), 972–976.

[8] Tao-Yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *CIKM*.

[9] A Grover and J Leskovec. 2016. node2vec: Scalable Feature Learning. In *KDD*.

[10] Maria Halkidi, Michalis Vazirgiannis, and Yannis Batistakis. 2000. Quality Scheme Assessment in the Clustering Process. In *PKDD*.

[11] Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. 2004. Two supervised learning approaches for name disambiguation in author citations. In *Joint Conference on Digital Libraries (JCDL)*.

[12] Madian Khabsa, Pucktada Treeratpituk, and C. Lee Giles. 2015. Online Person Name Disambiguation with Constraints. In *JCDL*.

[13] Xueqin Lin, Fen Zhu, Bo Peng, and Weiling Li. 2017. A Novel Approach for Author Name Disambiguation Using Ranking Confidence. In *DSFAA*.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Computer Science* (2013).

[15] Bryan Perozzi, Rami Alrfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. *KDD* (2014).

[16] Yanan Qian, Junting Zheng, and Jun Liu. 2015. Author name disambiguation for growing digital libraries. *Information Retrieval Journal* 18, 5 (2015).

[17] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. 2017. An Attention-based Collaboration Framework for Multi-View Network Representation Learning. In *CIKM*. 1767–1776.

[18] Jie Tang, Alvis C. M. Fong, Bo Wang, and Jing Zhang. 2012. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. *TKDE* 24, 6 (2012).

[19] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks. In *WWW*.

[20] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *KDD*.

[21] Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. 2017. CANE: Context-Aware Network Embedding for Relation Modeling. In *ACL*. 1722–1731.

[22] Xuezhi Wang, Jie Tang, Hong Cheng, and Philip S. Yu. 2011. ADANA: Active Name Disambiguation. In *ICDM*.

[23] Baichuan Zhang, Murat Dundar, and Mohammad Al Hasan. 2016. Bayesian Non-Exhaustive Classification A Case Study:Online Name Disambiguation using Temporal Record Streams. In *CIKM*.

[24] Baichuan Zhang and Mohammad Al Hasan. 2017. Name Disambiguation in Anonymized Graphs using Network Embedding. In *CIKM*.