Our work furthers this direction of investigation by designing the *metapath2vec* and *metapath2vec++* models to capture heterogeneous structural and semantic correlations exhibited from large-scale networks with multiple types of nodes, which can not be handled by previous models, and applying these models to a variety of network mining tasks.

6 CONCLUSION

In this work, we formally define the representation learning problem in heterogeneous networks in which there exist diverse types of nodes and links. To address the network heterogeneity challenge, we propose the *metapath2vec* and *metapath2vec++* methods. We develop the meta-path-guided random walk strategy in a heterogeneous network, which is capable of capturing both the structural and semantic correlations of differently typed nodes and relations. To leverage this method, we formalize the heterogeneous neighborhood function of a node, enabling the skip-gram-based maximization of the network probability in the context of multiple types of nodes. Finally, we achieve effective and efficient optimization by presenting a heterogeneous negative sampling technique. Extensive experiments demonstrate that the latent feature representations learned by *metapath2vec* and *metapath2vec++* are able to improve various heterogeneous network mining tasks, such as similarity search, node classification, and clustering. Our results can be naturally applied to real-world applications in heterogeneous academic networks, such as author, venue, and paper search in academic search services.

Future work includes various optimizations and improvements. For example, 1) the *metapath2vec* and *metapath2vec++* models, as is also the case with DeepWalk and node2vec, face the challenge of large intermediate output data when sampling a network into a huge pile of paths, and thus identifying and optimizing the sampling space is an important direction; 2) as is also the case with all meta-path-based heterogeneous network mining methods, *metapath2vec* and *metapath2vec++* can be further improved by the automatic learning of meaningful meta-paths; 3) extending the models to incorporate the dynamics of evolving heterogeneous networks; and 4) generalizing the models for different genres of heterogeneous networks.

Acknowledgments. We would like to thank Reid Johnson for discussions and suggestions. This work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and the National Science Foundation (NSF) grants CNS-1629914 and IIS-1447795.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and others. 2016. TensorFlow: A system for large-scale machine learning. In *OSDI '16*.
[2] Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, and Alexander J. Smola. 2013. Distributed Large-scale Natural Graph Factorization. In *WWW '13*. ACM, 37–48.
[3] Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *IEEE TPAMI* 35, 8 (2013), 1798–1828.
[4] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. 2015. Heterogeneous Network Embedding via Deep Architectures. In *KDD '15*. ACM, 119–128.
[5] Ting Chen and Yizhou Sun. 2017. Task-Guided and Path-Augmented Heterogeneous Network Embedding for Author Identification. In *WSDM '17*. ACM.
[6] Yuxiao Dong, Jing Zhang, Jie Tang, Nitesh V. Chawla, and Bai Wang. 2015. CoupledLP: Link Prediction in Coupled Networks. In *KDD '15*. ACM, 199–208.
[7] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR* abs/1402.3722 (2014).
[8] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In *KDD '16*. ACM, 855–864.
[9] Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. 2012. Rolx: structural role extraction & mining in large graphs. In *KDD '12*. ACM, 1231–1239.
[10] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical association* 97, 460 (2002), 1090–1098.
[11] Xiao Huang, Jundong Li, and Xia Hu. 2017. Label Informed Attributed Network Embedding. In *WSDM '17*. na.
[12] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. 2016. Meta structure: Computing relevance in large heterogeneous information networks. In *KDD '16*. ACM, 1595–1604.
[13] Ming Ji, Jiawei Han, and Marina Danilevsky. 2011. Ranking-based classification of heterogeneous information networks. In *KDD '11*. ACM, 1298–1306.
[14] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD '08*. ACM, 426–434.
[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[16] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *WSDM '11*. 287–296.
[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781
[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS '13*. 3111–3119.
[19] Jennifer Neville and David Jensen. 2005. Leveraging relational autocorrelation with latent group models. In *Proceedings of the 4th international workshop on Multi-relational mining*. ACM, 49–55.
[20] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. 2016. Asymmetric Transitivity Preserving Graph Embedding. In *KDD '16*. ACM, 1105–1114.
[21] Siddharth Pal, Yuxiao Dong, Bishal Thapa, Nitesh V Chawla, Ananthram Swami, and Ram Ramanathan. 2016. Deep learning for network analysis: Problems, approaches and challenges. In *Military Communications Conference, MILCOM 2016-2016*. IEEE, 588–593.
[22] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *KDD '14*. ACM, 701–710.
[23] Xiang Ren, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, and Jiawei Han. 2016. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *KDD '16*. ACM.
[24] Xin Rong. 2014. word2vec Parameter Learning Explained. *CoRR* abs/1411.2738 (2014). http://arxiv.org/abs/1411.2738
[25] Yizhou Sun and Jiawei Han. 2012. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Morgan & Claypool Publishers.
[26] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB '11*. 992–1003.
[27] Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S. Yu, and Xiao Yu. 2012. Integrating Meta-path Selection with User-guided Object Clustering in Heterogeneous Information Networks. In *KDD '12*. ACM, 1348–1356.
[28] Yizhou Sun, Yintao Yu, and Jiawei Han. 2009. Ranking-based Clustering of Heterogeneous Information Networks with Star Network Schema. In *KDD '09*. ACM, 797–806.
[29] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. PTE: Predictive Text Embedding Through Large-scale Heterogeneous Text Networks. In *KDD '15*. ACM, 1165–1174.
[30] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding.. In *WWW '15*. ACM.
[31] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD '08*. 990–998.
[32] Lei Tang and Huan Liu. 2009. Relational learning via latent social dimensions. In *KDD '09*. 817–826.
[33] Lei Tang and Huan Liu. 2011. Leveraging social media networks for classification. *DMKD* 23, 3 (2011), 447–478.
[34] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE TPAMI* 29, 1 (2007).
[35] Jing Zhang, Jie Tang, Cong Ma, Hanghang Tong, Yu Jing, and Juanzi Li. 2015. Panther: Fast top-k similarity search on large networks. In *KDD '15*. ACM, 1445–1454.