CubeNet: Multi-Facet Hierarchical Heterogeneous Network Construction, Analysis, and Mining

Carl Yang*, Dai Teng*, Siyang Liu*, Sayantani Basu*, Jieyu Zhang*, Jiaming Shen*, Chao Zhang*, Jingbo Shang*, Lance Kaplan#, Timothy Harratty#, Jiawei Han*
*University of Illinois at Urbana Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA
#US Army Research Laboratory, 2800 Powder Mill Rd, Adelphi, MD 20783, USA
*{jiyang3, daiteng2, sliu134, basu9, jieyuz2, js2, czhang82, shang7, hanj}@illinois.edu,
#{lance.m.kaplan.civ, timpthy.p.hanratty.civ}@mail.mil

ABSTRACT

Due to the ever-increasing size of data, construction, analysis and mining of universal massive networks are becoming forbidden and meaningless. In this work, we outline a novel framework called CubeNet, which systematically constructs and organizes real-world networks into different but correlated semantic cells, to support various downstream network analysis and mining tasks with better flexibility, deeper insights and higher efficiency. Particular, we promote our recent research on text and network mining with novel concepts and techniques to (1) construct four real-world large-scale multi-facet hierarchical heterogeneous networks; (2) enable insightful OLAP-style network analysis; (3) facilitate localized and contextual network mining. Although some functions have been covered individually in our previous work, a systematic and efficient realization of an organic system has not been studied, while some functions are still our on-going research tasks. By integrating them, CubeNet may not only showcase the utility of our recent research, but also inspire and stimulate future research on effective, insightful and scalable knowledge discovery under this novel framework.

1 INTRODUCTION

Research Project, Goals and Partners. Real-world networks nowadays are becoming very large (*e.g.*, DBLP publication network with 4 millions of paper nodes, Facebook social network with 2 billions of user nodes, *etc*). It is hard and wasteful, if not impossible, for various algorithms to scale up to the sheer sizes of networks, and many network analytical measures and mining tasks become meaningless on the massive universal networks.

In this work, we demonstrate the ability and value of constructing and organizing massive networks *w.r.t.* an underlying data cube structure. Firstly, based on metadata and textual contexts, we propose to automatically construct multi-facet hierarchical data cube structures for semantic-aware organization of massive networks. Next, we enable various cube-based OLAP-style network analysis including contrastive network summarization, cell-based semantic backtracking and multi-granularity structure exploration. Finally, we develop novel concepts and techniques for flexible and insightful network mining including contextual pattern mining, query-specific network localization and conditional proximity search.

Partners of this project include current researchers and alumni of UIUC, and we thank the support from funding agencies including US Army Research Lab, DARPA, NSF and NIH.

Fit with the KDD ecosystem This project will be attractive to the large audience interested in network mining, network science, text mining, big data management and various downstream applications. While researchers and practitioners familiar with network data mining and data management may get mostly benefited, we believe it will also provide valuable insight towards the new era of network science and inspire general audience and even newcomers to KDD. Particularly, the series of state-of-the-art techniques we develop and showcase brings new light to how massive network data could be organized and utilized in the future. The project and leveraged techniques are all fully open-sourced, so the audience can build their own CubeNet following our approaches in an effortless manner.

2 MAIN INNOVATIONS

Figure 1 gives an overview (Ex. 1) of the proposed CubeNet system, where a large multi-facet heterogeneous network is organized *w.r.t.* a topic-objective-year data cube structure.

2.1 CubeNet Construction

Heterogeneous network enrichment. Without clear semantics, real-world networks are less informative. For more insightful data analysis and mining, we enrich the heterogeneity of networks by incorporating nodes from massive free texts. In this system, we leverage our recent research on text mining, *i.e.*, AutoPhrase [2] for phrasal node extraction and AutoNER [3] for typed link generation.

Multi-facet taxonomy generation. In this system, we leverage both existing metadata and our recent research on automatic taxonomy generation, *i.e.*, TaxoGen [9], to create multiple taxonomies for each network, essentially leading to a data cube structure [6].

Weakly-supervised network organization. To organize networks into the data cube structure, *i.e.*, allocate nodes to proper cells, we leverage our recent research on heterogeneous network classification based on AutoPath [8], which assigns similar labels in the taxonomy to nodes close in the network based on small sets of nodes weakly labeled via surface name matching.

2.2 CubeNet Analysis

Contrastive network summarization. Various traditional network statistical measures such as clustering coefficient, character path length and triangle count become hard to compute and meaningless in massive universal networks. However, in CubeNet where each cell holds a relatively small network, the structures can be efficiently summarized and contrasted across relevant cells by aggregating network statistics, which provides insight into network evolution along different semantic dimensions (Ex. 2 in Figure 1).

Cell-based semantic backtracking. While text cube supports the retrieval of most relevant cells *w.r.t.* unary queries, CubeNet

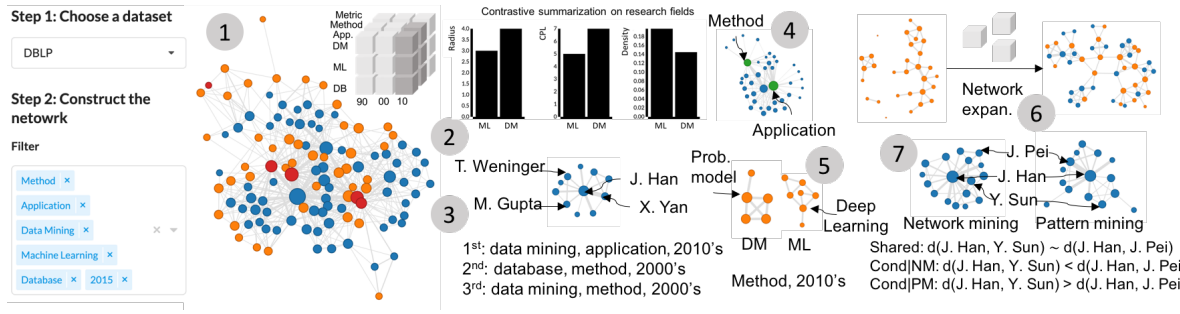

Figure 1: An overview of the proposed CubeNet system with examples illustrating the novel functions we developed.

further allows semantic backtracking *w.r.t.* network queries, such as pairs of nodes and small sub-networks. The idea is to combine the graph coverage [5] and top-k cell search [1] algorithms to find the k cells that mostly cover the network query from all cells with an optimized search order (Ex. 3 in Figure 1).

Multi-granularity structure exploration. By allocating nodes into hierarchically organized cells, CubeNet supports network roll-up and drill-down, which essentially merges nodes and edges into super-nodes and super-edges (or the other way around), to allow the exploration of network structures in preferred granularities. To make the process efficient, we implement the techniques developed in our previous research on graphcube [10] (Ex. 4 in Figure 1).

2.3 CubeNet Mining

Contextual network pattern mining. Traditional graph pattern mining does not consider the contexts of networks. To find more semantic-related patterns, we extend our previous work [7] to CubeNet by computing a mixture score of popularity, integrity and distinctiveness. Particularly, popularity is computed as the normalized frequency, integrity is the ratio of frequency between the pattern and its corresponding close pattern, and distinctiveness is the ratio between the frequency in a particular cell and the average in all cells. The three scores can be combined using customized weights to highlight user preference (Ex. 5 in Figure 1).

Query-specific network localization. Given data mining queries over particular sets of nodes, computation over the universal massive network is wasteful and hard to handle. Since CubeNet organizes networks by grouping semantically relevant nodes, it is possible to find a set of cells that mostly cover the queried nodes. To this end, we apply our on-going research on query-specific network construction, which leverages a light reinforcement learning algorithm to find the optimal combination of cells, from which a relevant and complete network can be constructed to support downstream data mining on queried set of nodes (Ex. 6 in Figure 1).

Conditional proximity search. While our recent research produces high-quality node embedding on universal networks [4, 8], they do not easily scale to millions of nodes and fail to consider node proximity under different semantic conditions. To deal with both challenges, we apply our on-going research on the co-embedding of network nodes and cube cells, which jointly learns node embedding in each sub-network and a set of embedding transformation functions that align relevant sub-networks. The embedding of sub-networks thus facilitates proximity search conditioned on corresponding cell semantics, while the alignment functions enable proximity transfer among similar cells (Ex. 7 in Figure 1).

3 DEMONSTRATION

The current CubeNet system indexes 4 real-world large-scale heterogeneous networks: (1) a DBLP network with 2M nodes of authors, topical phrases, venues, years and 0.4B links; (2) a Yelp network with 0.8M nodes of businesses, opinion phrases, locations, stars and 0.2B links; (3) a PubMed network with 0.2M nodes of genes, proteins, diseases, chemical compounds, species and 0.1B links; (4) a FreeBase network with 16M nodes of persons, locations, books, movies, *etc.* and 77M links. We further generate taxonomies of 11K topics for DBLP, 70K categories for Yelp, 197K diseases for PubMed, and 800K types for FreeBase, together with other trivial taxonomies like publication years and rating scores from metadata. A toy system is available at https://github.com/yangji9181/CubeNet, which currently supports main functionalities on DBLP and Yelp, while a full version will be gradually rolled out before demonstration.

ACKNOWLEDGEMENT

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

REFERENCES

[1] Bolin Ding, Bo Zhao, Cindy Xide Lin, Jiawei Han, Chengxiang Zhai, Asok Srivastava, and Nikunj C Oza. 2011. Efficient keyword-based search for top-k cells in text cube. *TKDE* 23, 12 (2011), 1795–1810.
[2] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *TKDE* 30, 10 (2018).
[3] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, and Jiawei Han. 2018. Learning Named Entity Tagger using Domain-Specific Dictionary. In *EMNLP*.
[4] Jingbo Shang, Meng Qu, Jialu Liu, Lance M Kaplan, Jiawei Han, and Jian Peng. 2016. Meta-Path Guided Embedding for Similarity Search in Large-Scale Heterogeneous Information Networks. *arXiv preprint arXiv:1610.09769* (2016).
[5] Jiaming Shen, Jinfeng Xiao, Xinwei He, Jingbo Shang, Saurabh Sinha, and Jiawei Han. 2018. Entity set search of scientific literature: An unsupervised ranking approach. In *SIGIR*. 565–574.
[6] Fangbo Tao, Honglei Zhuang, Chi Wang Yu, Qi Wang, Taylor Cassidy, Lance R Kaplan, Clare R Voss, and Jiawei Han. 2016. Multi-Dimensional, Phrase-Based Summarization in Text Cubes. *IDEB* 39, 3 (2016), 74–84.
[7] Xifeng Yan and Jiawei Han. 2003. CloseGraph: mining closed frequent graph patterns. In *KDD*. 286–295.
[8] Carl Yang, Mengxiong Liu, Frank He, Xikun Zhang, Jian Peng, and Jiawei Han. 2018. Similarity Modeling on Heterogeneous Networks via Automatic Path Discovery. In *ECML-PKDD*. 37–54.
[9] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *KDD*. 2701–2709.
[10] Peixiang Zhao, Xiaolei Li, Dong Xin, and Jiawei Han. 2011. Graph cube: on warehousing and OLAP multidimensional networks. In *SIGMOD*. 853–864.