Research Proposal

Title	**Mining Graphs for Network Motifs and Frequent Patterns**
Applicant	Lincheng Jiang
Email	linchjiang@gmail.com
Domestic Supervisor	Prof. Weidong Xiao
Domestic Institution	College of Information System and Management, National University of Defense Technology
Host Supervisor	Prof. Dennis Shasha
Host Institution	Department of Computer Science, Courant Institute of Mathematical Sciences, New York University
Date of Study	September,2015 — August,2016

March 2015

1 Introduction

Recent scientific and technological advances have resulted in an abundance of data modeled as graphs. It is one of the most important data structure in computer science, which not only focuses on the attributes of the data object itself, but also pays attention to the interaction between data objects. As a general data structure representing relations among entities, graph has been used extensively in modeling complicated structures and schemaless data, such as proteins [1], XML documents [2], images [3], program flows [4], the Web [5] and so on.

A graph data structure consists of a finite (and possibly mutable) set of nodes or vertices, together with a set of ordered pairs of these nodes (or, in some cases, a set of unordered pairs). These pairs are known as edges or arcs. As in mathematics, an edge (x,y) is said to point or go from x to y. The nodes may be part of the graph structure, or may be external entities represented by integer indices or references. A graph data structure may also associate to each edge some edge value, such as a symbolic label or a numeric attribute (cost, capacity, length, etc.). A labeled graph is always denoted as six-tuple, i.e.

$$G = \langle V, E, L_v, L_e, F_v, F_e \rangle \tag{1}$$

where V is the set of vertices, E is the set of edges, L_v is the set of vertex labels, L_e is the set of edge labels, F_v is a function: $V \rightarrow L_v$ that assigns labels to vertices and F_e is a function: $E \rightarrow L_e$ that that assigns labels to edges.

The dominance of graphs in real-world applications asks for effective graph data management so that users can organize, access, and analyze graph data in a way one might have not yet imagined. Among myriad graph-related problems of interest, a common and critical one shared in many applications in science and engineering is the finding significantly overrepresented subgraphs in a (large) network. The overrepresented subgraphs can help us to better get information of graphs, which plays a great role in graph application systems. In my research, two main sub-areas are distinguished:

(1) Network motif discovery [6] usually refers to the discovery of subgraphs that are overrepresented with respect to network randomizations, with p-value higher than a certain threshold.

(2) Frequent subgraph mining [7] refers to the discovery of subgraphs that occur more than a specified threshold.

All the two sub-areas have been widespread concerned and deeply studied in the past decades, which has made a large number of research outcomes. But there are still many problems and to be solved and difficulties to be overcome. I will tackle some interesting problems and try to solve them.

2 Research contents

The proposed research covers the aforementioned two sub-areas for mining overrepresented subgraphs. In particular, I plan to investigate and advance the solutions to network motif discovery and frequent subgraph mining. In addition, as a key component involved in any algorithmic solutions to subgraph mining, I will look into subgraph isomorphism test too. I detail the three problems and review related work below, followed by some preliminary thoughts.

2.1 Network motif discovery

A motif in a network G is a connected graph H that occurs significantly more frequently as an induced subgraph than would be expected in a similar random network. The term network motif was coined by Alon et al [6,8,9], who discovered that they occur in several biological and artificial network, and thought that motifs might play a more important role than arbitrary subgraphs. Recently, network motifs have been found in a vast range of networks, and, in some cases, have been identified as functionally important.

Many algorithms have been proposed that enable motif discovery. Mainstream algorithms, those that can perform a k-node subgraph census, include Mfinder [6,9] (introduce network motifs, brute force and edge sampling), ESU(FANMOD) [10,11] (avoid duplication without simmetry breaking, node sampling), MODA [12] (extract larger motifs efficiently), and NeMoFinder [13] (maximal, not necessarily induced motifs) and gTrie [14] (a data structure whose authors claim impressive speedups vs. FanMod).

Motif discovery is typically performed by enumerating subgraphs in an input network and in an ensemble of comparison networks, which poses a significant computational problem. The problem is more prominent when finding rather large motifs. However, with the coming of the big data era the motifs are larger and larger, making the problem more difficult. In my research, I try to adopt better designed sampling methods and parallel algorithms to solve this problem.

2.2 Frequent subgraph mining

Frequent pattern mining has been a focused theme in data mining for more than a decade, making remarkable progress. Graph patterns, or frequent subgraphs, are of particular interest lately, which are subgraphs found from a collection of small graphs or single large graph with support no less than a user-specified threshold. Frequent subgraphs are useful at characterizing graph datasets, classifying and clustering graphs, and building structural indices [15].

The straightforward idea behind frequent subgraph mining (FSM) is to grow candidate subgraphs, in either a breadth first or depth first manner (candidate generation), and then determine if the identified candidate subgraphs occur frequently enough in the graph data set for them to be considered interesting (support counting) [7]. The two main research issues in FSM are generating the candidate frequent subgraphs and determining the frequency of occurrence of the generated subgraphs efficiently and effectively. Some of the most known algorithms are gSpan, FSG, FFSM, Gaston and SUBDUE. For more and various types of graph pattern mining problems and algorithms, a recent survey paper [15] is a good choice.

My research will focus on FSM mining on uncertain graphs. Lately, research effort has been dedicated to FSM on a collection of small uncertain graphs. Being equally important, however, the problem on single large uncertain graphs remains open, given that real-life large networks are increasingly involved with uncertainty in nature. For example, the relation of one person influences another in social network is probabilistic; the protein-protein interaction found in biological network is not error-proof due to measurement limit, and so forth. My research comes in response trying to fill the gap. I will propose to an approximation algorithm with

accuracy guarantee. Optimization techniques will also be developed to share computation among samples and prune non-promising subgraphs to further enhance efficiency.

2.3 Subgraph isomorphism test

Given a query graph q and a data graph g, subgraph isomorphism test refers to find all occurrences of q in g, which is considered one of the most fundamental query types for many real applications. Adopting better and more suitable subgraph isomorphism algorithm can effectively speed up network motif discovery and frequent subgraph mining. Subgraph isomorphism test is considered as a NP-hard problem, but many algorithms have been proposed to solve it in a reasonable time for real datasets.

Most mainstream subgraph isomorphism algorithms are based on backtracking [16], include Ullmann algorithm [17] (first algorithm to tackle this problem), VF2 [18] (exploits constraints to prune out candidate vertices), QuickSI [19] (access vertices having infrequent vertex labels and infrequent), SPath [20] (minimize the depth of the recursion tree by matching a path per call), etc.

However, a recent study has shown, through an extensive benchmark with various real datasets, that all existing algorithms have serious problems in their matching order selection. Furthermore, all algorithms blindly permutate all possible mappings for query vertices, often leading to useless computations. In the proposed research, I try to design a better vertices matching order to solve this problem.

3 Research Background and Preparation

I have solid background of computing and programming skills, and have published a few excellent papers, which will benefit my research at New York University. In recent two years, my main work is related to graph query and graph mining, and I am familiar with the basic theory in this field, the development trends and research priorities.

I am very interested in network motif discovery and frequent subgraph mining, and have done some related research in finding significantly overrepresented subgraphs in a (large) network. In my master's thesis, I proposed a new over-represented subgraphs mining demand, named minimal unique induced subgraph (MUIS). MUIS mining refers to find out a unique induced subgraph with a minimum number of vertices which contain the given query vertex. I have given the formal definition of MUIS, explored its property and given its coding methods. A filter-validation framework for solving the MUIS mining problems is also proposed. MUIS mining provides a new data access and using method for graph data management. MUIS mining can not only tap the special structure of the vertex neighborhood, but also can be used for visualization and to explore the property of vertices.

Moreover, my domestic laboratory has undertaken some national research projects, relating to graph data management. I can get great help from my domestic team. With the advising of Prof. Shasha, I am sure I can finish this project on time and make advances in my research.

4 Expected goals

I have several goals to fulfill during my visiting study at New York University.

(1) Through communication and cooperation, I will learn some advanced concepts and methodology concerning graph data mining, and attempt to develop relationships with the world-class academic teams and international peers.

(2) I will systematically study the latest progress in this research area and try my best to do some innovative and frontier research work in this research area under the guidance of Prof. Shasha.

(3) I will publish at least 3 international journal papers indexed by SCI and try to submit some papers to top international conference, such as SIGMOD and VLDB. What is more, I will integrate the academic achievements with my PhD dissertation.

References

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. Nucleic Acids Research, 28(1):235-242, 2000.

[2] J. McHugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom. Lore: A database management system for semistructured data. SIGMOD Record, 26(3):54-66, 1997.

[3] S. Berretti, A. D. Bimbo, and E. Vicario. Efficient matching and indexing of graph models in content-based retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 23(10):1089-1105, 2001.

[4] C. Liu, X. Yan, L. Fei, J. Han, and S. P. Midkiff. Sober: statistical model-based bug localization. In Proceedings of the 10th European software engineering conference (ESEC/FSE'05), pages 286-295, 2005..

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In Proceedings of SIGCOMM'99, pages 251-262, 1999.

[6] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. Science Signalling, 298(5594):824, 2002.

[7] C. Jiang, F. Coenen, and M. Zito. A survey of frequent subgraph mining algorithms. Knowledge Engineering Review, 28(1):75C105, 2013.

[8] S. Shen-Orr, R. Milo, S. Mangan, U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31: 64-68, 2002.

[9] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 20(11), 2004.

[10] S. Wernicke. Efficient detection of network motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 3(4):347C359, 2006.

[11] S. Wernicke, F. Rasche. FANMOD: a tool for fast network motif detection. Bioinformatics 22: 1152-1153, 2006.

[12] S. Omidi, F. Schreiber, and A. Masoudi-Nejad. MODA: an efficient algorithm for network motif discovery in biological networks. Genes and genetic systems, 84(5):385C395, 2009.

[13] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng. Nemofinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 106C115. ACM, 2006.

[14] P. Ribeiro and F. Silva. G-Tries: a data structure for storing and finding subgraphs. Data Mining and Knowledge Discovery, 28(2):337C377, 2014.

[15] H. Cheng, X. Yan, and J. Han. Mining graph patterns. In Frequent Pattern Mining, pages 307C338. 2014.

[16] J. Lee, W.-S. Han, R. Kasperovics, and J.-H. Lee. An in-depth comparison of subgraph isomorphism algorithms in graph databases. PVLDB, 6(2), 2013,

[17] J. R. Ullmann. An algorithm for subgraph isomorphism. J. ACM, 23:31C42, January 1976.

[18] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. IEEE PAMI, 26(10):1367C1372, 2004.

[19] H. Shang, Y. Zhang, X. Lin, and J. X. Yu. Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. PVLDB, 1(1):364C375, 2008.

[20] P. Zhao and J. Han. On graph query optimization in large networks. PVLDB, 3(1):340C351, 2010.

Supervisor: (Signature)