COMP4801 Final Year Project Plan

Motif-G: A Motif-based Graph Analysis Platform

By

LI Boxuan

Supervised By

Dr. Reynold C.K. Cheng

Department of Computer Science
The University of Hong Kong

A project plan submitted to the Department of Computer Science in accordance with the requirements of the final year project in the Faculty of Engineering.

September 2018

Contents

1 Introduction

Motif and clique are two important concepts in graph theory. A recently proposed concept *motif-clique* combines motif and clique together, providing abundant semantic information based on heterogeneous information networks. This project will continue the work based on motif-clique concept and develop a fully functional motif based graph analysis platform together with a comprehensive algorithm.

2 Background

2.1 Heterogeneous Information Network

Heterogeneous information networks (HINs), such as bibliographical datasets, are widely used and discussed [1, 2]. Nodes of HINs are labeled, providing more abundant semantic meanings than unlabeled graphs [3]. Compared to homogeneous information networks, HINs distinguish different types of nodes and edges in the networks, consisting of rich semantic meaning of structural types of nodes [4].

2.2 Motif

A motif is essentially a small subgraph pattern, which is a foundational building block of complex HINs [5, 6]. Also known as higher-order structure, motif provides a tool to discover higher-order semantics of HINs [7]. It is widely used in graph analysis problems, such as graph clustering [8, 9], social network analysis [10].

2.3 Clique

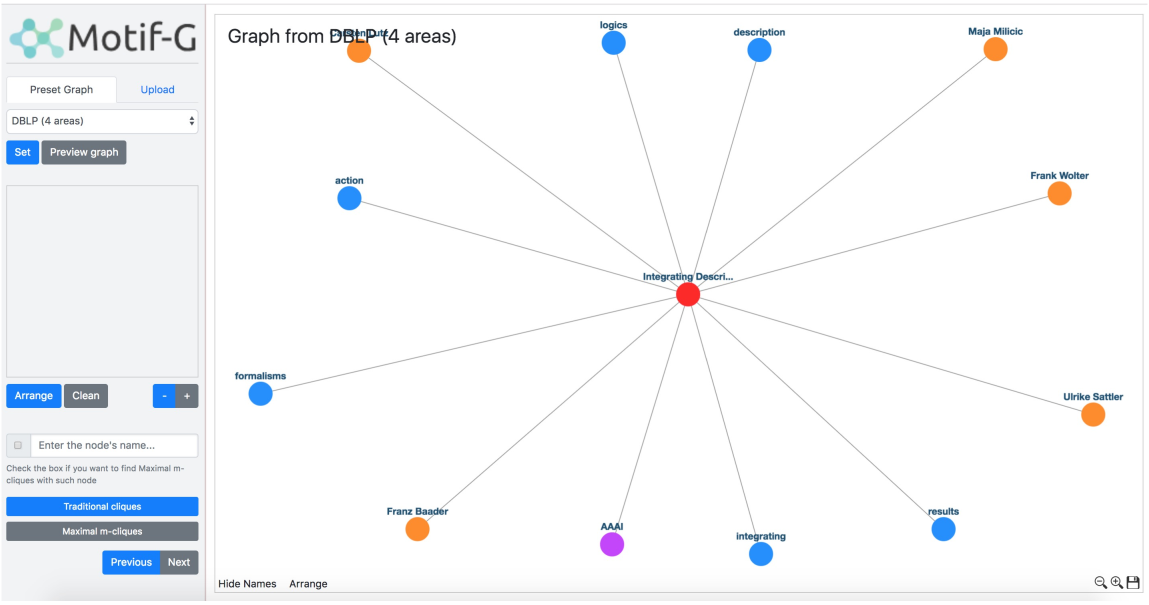
A clique is by definition a complete graph, i.e., every two nodes in the clique are adjacent. Thus, a clique represents a set of nodes that are closely relevant (e.g., a clique in a social network can represent a group of close friends). Cliques have been widely studied in both research and industry communities. Usages of cliques include social network detection [10], gene group detection [11], and transportation network analysis [12]. A maximal clique is a clique that is not a subgraph of any larger clique.

2.4 Motif-clique

Hu et al. [7] proposed a new concept, namely motif-clique or m-clique in short, which incorporates motifs to the clique definition. Recall a clique is a complete graph based purely on edges, i.e. it is complete since every two distinct vertices are connected by an edge. A motif-clique, as a generalization of a traditional clique, is a complete graph based on a user-defined pattern, i.e. motif, rather than edges. An m-clique is, therefore, a *higher-order* clique based on a user-given motif. Compared to traditional cliques, which treats nodes with different labels equally, an m-clique can capture the desired relationship among labeled nodes. A motif detection algorithm has been proposed and implemented.

2.5 Motif-clique Query System

As Hu et al. [7]'s subsequent work, a basic functional online motif-clique query system has been developed. Users can upload datasets following a specific format, or use a predefined dataset for demo purpose.



Check the box if you want to find Maximal m-cliques with such node

Figure 1: Motif-G - preview graph

As Figure 1 shows, after uploading or selecting a dataset, users can click *Preview graph* button to preview the graph. Before querying, users need to specify a motif, which is essentially a pattern that they want the clique to be based on.

As Figure 2 shows, a simple motif is defined on the left panel. There are four nodes in this motif, two of which are authors and the other two are papers. Each author is adjacent to both papers. This motif encapsulates the relationship among authors and papers, aiming to find co-authors who have at least two paper collaborations. Users can click *Maximal m-clique* button to see the result. The right panel on figure 2 shows an example of an m-clique based on the given motif. In this graph, two authors and several papers consist an m-clique. It is a motif-clique because, in this graph, any subgraph which consists of exactly two authors and two papers are connected in the same manner as the given motif. That is, pick two papers and two authors randomly in the motif-clique, each author must be adjacent to both papers and each paper must be adjacent to both authors. It is a maximal motif-clique because no other paper or author in the dataset can be added to the motif-clique such that it is still a valid motif-clique.

The algorithm, together with the demo system, is incomplete and still have much space to improve, which this project aims to tackle. Details will be covered in the project objective section.

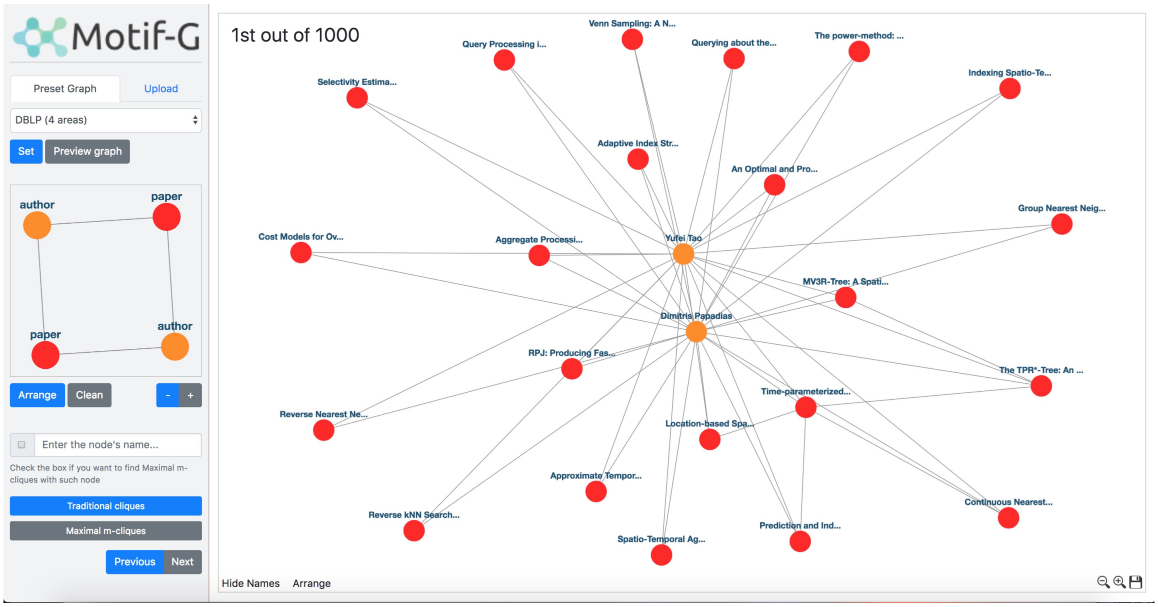


Figure 2: Motif-G - result of m-clique query

3 Objectives

This project aims to improve the algorithm and develop a fully functional motif-based analysis platform. The existing work by Hu et al. [7] will be utilized and extended. Currently, there are several limitations and problems with the algorithm and the system.

First, the m-clique search algorithm is not fast enough. Current experiments have shown that the search algorithm is much faster than a basic brute force search algorithm. However, it has perceivably long latency when being used on the platform. There are several steps which might be accelerated. For example, it uses a well-designed set-trie data structure to avoid duplication while searching. This pruning strategy is effective, however, can be improved. This project aims to improve the m-clique finding algorithm and use other ancillary data structure to speed up searching.

Second, the existing m-clique search algorithm only supports unweighted undirected graph. It does not support directed graph or weighted graph. DisGeNET, for example, is a dataset which focuses on gene-disease associations [13]. The dataset, however, has only directed and weighted edges. This provides abundant semantics but at the same time poses greater challenges to the existing algorithm. This project will make the m-clique search algorithm more generic so that it can be applied to different datasets.

Third, the platform does not support concurrency. Due to the limitation of the current implementation, the current online platform cannot support two or more queries at the same time. Considering the fact that a query might take rather long time, especially when the dataset is large or the motif is complicated, it could be the case that one person sends a query request before another person's query request is responded. Under such circumstance, the previous query would fail due to the interruption. This project aims to tackle this issue by adding concurrency support. Since the searching is resource intensive, the concurrency level would be limited, depending on the resources of the server. This project will also improve user experience by letting latter requests stay on hold rather than interrupt former requests when necessary.

Additionally, the query process itself is very basic and lacks some features. A dataset can be extremely large, containing hundreds of thousands of nodes and millions of edges. Under this circumstance, searching the whole graph would be rather slow. This project will mitigate the problem by allowing users to limit the search to only several parts of a large graph. The current searching process is also not configurable. Users cannot stop or continue searching or see the estimated time for the remaining process. Besides, the system returns all results as a whole instead of returning results in a stream, which might not be satisfactory because users would have to wait till the whole search process finishes. User experience would be better if users can start viewing results while the search process is going on. The algorithm itself supports this feature naturally, but the workflow implemented in the backend server does not yield results in a stream. This project will make users able to see a subset of results while waiting for rest results from the server at the same time.

Moreover, there are several frontend related problems with the current motif graph analysis system. It does not have export functionality, which means users can see results on the website, but they have no way to download the results. The export functionality will be added so that users can keep the results in archives or do further analysis based on motif-clique results. Another concern is that users do not have the flexibility to see more details on results. When a maximal motif-clique is shown on the panel, users might want to see more information such as surrounding nodes. For example, a motif-clique might only contain paper and author nodes, but users might also want to see venue and term nodes connected to paper nodes. This project will make the system more interactive and flexible to help users explore and utilize the discovered high-order semantics better. There is another frontend problem - some nodes might have the same names. Currently, the system allows users to specify a node which the discovered motif-clique must contain. However, this functionality does not work when two or more nodes share the same name. This issue will be tackled by this project.

Finally, this project will also seek chances to apply the algorithm to bioinformatics fields. Collaboration with bioinformatics researchers will be conducted to apply motif-clique detection techniques to bioinformatics datasets to discover the hidden relationship among gene and human diseases.

4 Methodology

The project consists of two parts, algorithm improvement, and web platform development.

4.1 Algorithm

Based on the analysis of the current algorithm, several potential improvements will be proposed. Pseudo-code for algorithms together with proof will be written, followed by actual implementation in C++. Extensive benchmarks and experiments would be conducted to compare the efficiency before and after the improvement.

Moreover, this project involves collaboration with bioinformatics research group at the Department of Computer Science, The University of Hong Kong. To fit datasets and generate more meaningful analysis and detections, the algorithm needs to be extended.

4.2 Web Platform

The existing platform is written in Java and JavaScript. This project aims to improve the platform comprehensively, which involves both backend and frontend work. The backend, written in Java, would be improved accordingly when the core algorithm is optimized. Concurrency support would also be added to make the platform more efficient and effective. The frontend, written in JavaScript, would be improved to enhance user experience. Several features will be implemented to make the platform more usable and comprehensive.

5 Schedule

Time Periods	Tasks & Milestones
September	• Detailed project plan submission • Setup final year project website • Related research paper review • Get familiar with current code base
October	• Discussion with bioinformatics researchers on the extension of the algorithm & platform • Enhance pruning strategies to speed up raw algorithm, implement enhanced algorithm in C++ and do extensive experiments
November - January	• Add concurrency support on the platform • Improve algorithm & platform to fit bioinformatics datasets and conduct data mining (write paper or journal article if applicable) • Interim report submission
January - March	• Make searching process configurable (ability to stop/resume, live view functionality, etc.) • Implement other frontend features (export functionality, interactive view, etc.) • Continuously collaborate with bioinformatics researchers to do data mining based on the platform
March - April	• System Evaluation • Final report submission & presentation

6 Conclusion

Motif-clique incorporates motif into clique definition, providing a new way to discover higher-order semantics of large heterogeneous information networks. The existing work is incomplete and has many limitations. This project aims to complete the motif-based graph analysis platform by improving the algorithm and enhancing the website. At the same time, this project involves collaboration with bioinformatics researchers, where the platform can be utilized to generate abundant higher-order semantics in the bioinformatics field. After the completion of this project, it would be flexible and convenient to conduct data mining on heterogeneous information networks.

References

[1] M. Ji et al. Graph regularized transductive classification on heterogeneous information networks. In ECML-PKDD, pages 570586, 2010.

[2] M. Ley. Dblp: some lessons learned. PVLDB, 2(2):14931500, 2009.

[3] Shi et al. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering 29(1):17-37, 2017.

[4] Sun et al. Mining heterogeneous information networks: a structural analysis approach. Acm Sigkdd Explorations Newsletter 14(2):20-28, 2013.

[5] R. Milo et al. Network motifs: simple building blocks of complex networks. Science, 298(5594):824827, 2002.

[6] N. Przulj and N. Malod-Dognin. Network analytics in the age of big data. Science, 353(6295):123124, 2016.

[7] J. Hu et al. Discovering Maximal Motif Cliques in Large Heterogeneous Information Networks. In ICDE, 2019.

[8] H. Yin et al. Local higher-order graph clustering. In KDD, pages 555 564, 2017.

[9] A. R. Benson et al. Higher-order organization of complex networks. Science, 353(6295):163166, 2016.

[10] R. A. Hanneman and M. Riddle. Introduction to social network methods, chapter 11: cliques., 2005.

[11] G. A. Pavlopoulos et al. Using graph theory to analyze biological networks. BioData mining, 4(1):10, 2011.

[12] X. Yang et al. Bus transport network model with ideal n-depth clique network topology. Physica A: Statistical Mechanics and its Applications 390(23-24):4660-4672, 2011.

[13] J. Piero et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database 2015, 2015.