2020 IEEE 36th International Conference on Data Engineering (ICDE)

MC-Explorer: Analyzing and Visualizing Motif-Cliques on Large Networks

Boxuan Li*, Reynold Cheng*, Jiafeng Hu†, Yixiang Fang‡, Min Ou*
Ruibang Luo*, Kevin Chen-Chuan Chang§, Xuemin Lin‡
*The University of Hong Kong, †Google, ‡The University of New South Wales
§University of Illinois at Urbana-Champaign
*{liboxuan@connect, ckcheng@cs, oumin717@connect, rbluo@cs}.hku.hk, †jfhu@google.com
‡{yixiang.fang@, lxue@cse.}unsw.edu.au, §kcchang@illinois.edu

***Abstract*— Large networks with labeled nodes are prevalent in various applications, such as biological graphs, social networks, and e-commerce graphs. To extract insight from this rich information source, we propose *MC-Explorer*, which is an advanced analysis and visualization system. A highlight of *MC-Explorer* is its ability to discover *motif-cliques* from a graph with labeled nodes. A motif, such as a 3-node triangle, is a fundamental building block of a graph. A motif-clique is a "complete" subgraph in a network with respect to a desired higher-order connection pattern. For example, on a large biological graph, we found out some motif-cliques, which disclose new side effects of a drug, and potential drugs for healing diseases. *MC-Explorer* includes online and interactive facilities for exploring a large labeled network through the use of motif-cliques. We will demonstrate how *MC-Explorer* can facilitate the analysis and visualization of a labeled biological network.**

An online demo video of *MC-Explorer* can be accessed from https://www.dropbox.com/s/vkalumc28wqp8yl/demo.mov

I. Introduction

Billion-node-scale graphs, such as biological networks, bibliographical databases, and co-purchasing graphs, are prevalent in different fields [8], [6], [10], [5], [4]. These graphs, which contain a gigantic number of inter-related nodes, as well as textual descriptions about edges and nodes, are "information gold mines". In recent years, research and industry communities are actively developing effective analysis tools to extract new insights from large graphs.

In this paper, we propose *MC-Explorer*, an analytics platform for facilitating online knowledge discovery from a large graph. The main goal of *MC-Explorer* is to extract and analyze *motif-cliques*, which are subgraphs of a given graph G. To understand a motif-clique, let us consider Fig. 1, which shows a graph with nodes representing diseases (in orange), drugs (in green), and genes (in blue). This graph contains a large number of instances following a pattern: a 'drug' node is connected to a 'disease' node through a 'gene' node (Fig. 2). In fact, this pattern, or *motif*, is a *fundamental* building block of Fig. 1. A motif, which is also regarded as a higher-order connection pattern, reveals important information about a graph, and has attracted a lot of attention in recent years [8]. A motif-clique is essentially a 'complete' subgraph of G in terms of a motif. For example, in Fig. 1, the subgraph S, enclosed by the red box, is a motif-clique for the motif M shown in Fig. 2. M is isomorphic to any subgraph of S containing a drug, a gene, and a disease node.

The motif-clique S shown in Fig. 1, which is "dense" in terms of the motif in Fig. 2, reveals interesting relationship between diseases and drugs. It was known that the drug *loxapine* has side effect *schizophrenia*. Because *loxapine* is connected to *obesity* nodes through the gene nodes that also link to *schizophrenia* nodes, our biological experts (co-authors Luo and Ou) hypothesized that *loxapine* could lead to *obesity*. Thus, the motif-clique here could be useful for *side effect analysis*. As we will discuss, the motif-cliques obtained from biological networks are also important for *disease subtyping* and *drug repurposing* applications. We will also show in the demo that motif-cliques are useful for bibliographical data analysis and online product recommendation.

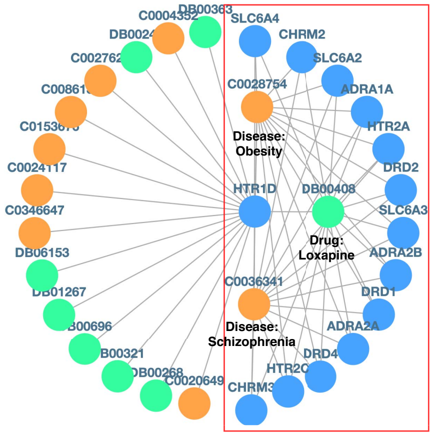


Fig. 1: Illustrating a motif-clique.

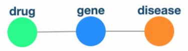


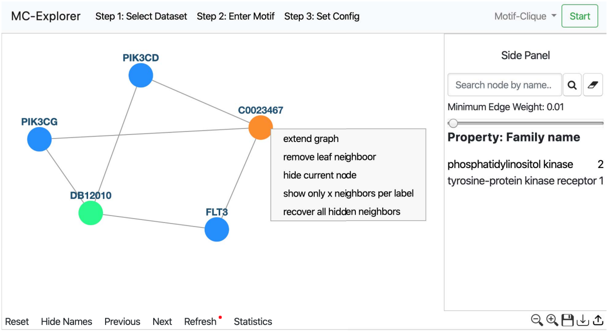
Fig. 2: A motif for Fig. 1.

Related works. A motif-clique is a generalization of a *clique*, where a motif-clique reduces to a clique when the motif given is an edge [8]. A clique of graph G is a subgraph of G where each node is connected to every other node in the subgraph. A maximal clique is a clique such that it is not a subgraph of another clique. The problem of detecting maximal cliques is well studied (e.g., Bron-Kerbosch algorithm [2], algorithm for large sparse graphs [15]), and clique-based analysis tools have been recently developed (e.g., [13], [1]). However, the maximal cliques found may not always be useful. For example, for the graph in Fig. 1, a maximal clique is just

2375-026X/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDE48307.2020.00154

an edge! On the other hand, we will adopt our previously-developed solutions for finding maximal motif cliques [8], which reveal important meanings (e.g., drug repurposing) that may not be found by maximal cliques.

Fig. 3 shows the user interface of *MC-Explorer* that runs on a biological dataset. A motif-clique found is displayed, which shows the properties of the node when the mouse hovers over it. On the right panel, a user can conduct a keyword search on the nodes and edges of the motif-clique found, or filter low-weighted edges for facilitating the viewing of the most important relationships. *MC-Explorer* also includes functions of discovering motifs from the user-given graph, based on state-of-the-art motif-discovery algorithms (e.g., [11], [12]).



Minimum Edge Weight: 0.01

Property: Family name

phosphatidylinositol kinase	2
tyrosine-protein kinase receptor	1

Fig. 3: Illustrating *MC-Explorer*.

The rest of the paper is organized as follows. We introduce the framework of *MC-Explorer* in Section II, discuss how it supports biological data analysis in Section III, and discuss the demonstration process in Section IV.

II. System Overview

We now discuss more about the main modules and workflow of *MC-Explorer*. As shown in Fig. 4, *MC-Explorer* adopts a *browser-server* model. The browser-side provides an interface for user input (e.g., on graphs and motifs), as well as facilities for displaying the information about the graph and motif-cliques found. It also supports visualization and statistical analysis of the motif-cliques.

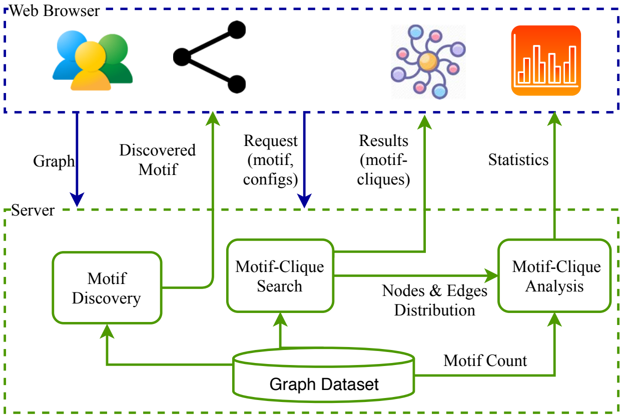


Fig. 4: The architecture of *MC-Explorer*.

The server-side contains three major modules:

- **Motif suggestion.** The function of this module is to allow a user to specify motif manually. We also provide facilities for finding motifs for a given graph G automatically.
- **Motif-clique discovery.** This module is responsible for finding maximal motif-cliques from G for a given motif M. Here we will adopt our recently developed algorithm called `META` [8], which is a highly efficient and scalable algorithm for enumerating maximal motif-cliques.
- **Motif-clique analysis.** This module provides statistical analysis related to motif-cliques found (e.g., the frequencies of motifs and motif-cliques, and node degree distribution of motif-cliques).

Workflow. The user first uploads a graph dataset to the system. Then, the user can visualize and explore the graph, and also use the motif suggestion module to obtain motif(s). The motif is then passed to the motif-clique discovery module for obtaining maximal motif-cliques. Afterward the motif-clique analysis module can then be used to perform analysis on the motif-cliques found.

Next, we discuss how *MC-Explorer* supports biological data analysis.

III. Biological Data Analysis

We have adapted *MC-Explorer* on a biological graph, and performed some case studies in the system. The features of *MC-Explorer* described here will be demonstrated to the conference participants.

Datasets. For the purpose of the demonstration, we have created a biological network, which consists of 46,186 nodes, 578,855 edges, and 3 node labels (i.e., drug, gene, disease). The graph is built by integrating two bipartite graphs, namely DisGeNet [14] and DrugBank [9]. The DisGeNet graph contains links between gene and disease nodes; a link between a gene and a disease means that the disease can be caused by some pathological changes to the gene involved. The DrugBank network contains relationship information between genes and drugs; a link between a gene and drug indicates that the drug could produce a therapeutic effect on some diseases associated with the gene. A subgraph of this network is shown in Fig. 1.

Next, we illustrate how the three modules of *MC-Explorer* described in the previous section are customized.

A. Motif Suggestion

To enable the discovery of motif-cliques, the user has first to specify the motif to be used. The *MC-Explorer* system provides two facilities for suggesting motifs. First, the user can find graph patterns that frequently occur on the graph. One such example is the motif in Fig. 2. As it turns out, this motif, which acts as a “bridge” between drugs and diseases, enable interesting knowledge discovery that will be later discussed. To implement this function, we use the algorithm [11], which efficiently finds the graph pattern that occurs the most frequently on a given graph.

Another motif suggestion function of *MC-Explorer* is based on the textual information given by the user. In particular, the user provides labels about the type of the nodes desired to appear in the motif (e.g., “disease”), and our system will return motifs containing labels given by the user. To achieve

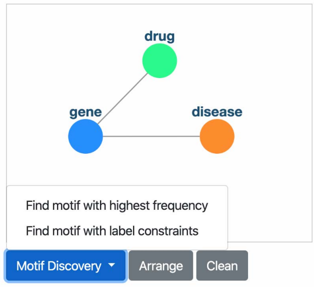

drug
gene
disease
Find motif with highest frequency
Find motif with label constraints
Motif Discovery
Arrange
Clean

Fig. 5: Motif suggestion.

this goal, we have adopted the label-driven motif discovery algorithm in [12]. Fig. 5 illustrates the user interface for motif suggestion in *MC-Explorer*.

B. *Motif-clique Discovery*

Based on the motifs produced by the motif suggestion module, the motif-clique discovery module generates maximal motif-cliques. Although maximal motif-clique enumeration problem is NP-hard, we use *META* algorithm [8], which employs advanced pruning strategies to effectively reduce the search space. It is discussed in [8] that the algorithm is highly efficient and effective. In the following, we discuss three medical applications based on the motif-cliques found. Fig. 6 shows the proposed functionalities of *MC-Explorer* extended to support these applications.

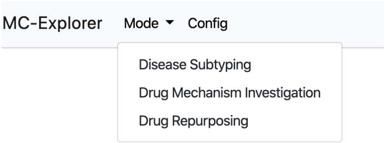


Fig. 6: The 3 functions of motif-cliques for medical analysis.

1) Disease subtyping: This is about discovering hidden intrinsic characteristics of drugs and finding subtypes of diseases. A better understanding of subtypes of a disease could help medical doctors to choose more precise medicine to treat patients. There is a lot of medical research focusing on discovering the subtypes of certain diseases, e.g., cancer. By subtyping a cancer disease, it is known that the complexity of analyzing cancer can be reduced; the cancer can be represented by a small number of underlying characteristics called *hallmarks* [7]. We have used the motif shown in Fig. 2 in *MC-Explorer* to find maximal motif-cliques related to breast cancer. This motif-clique, shown in Fig. 7, depicts that a known subtype of breast cancer called *sustaining proliferative signaling*. Moreover, this subtype can be targeted by a drug named *Marimastat*. We have found six more motif-cliques that correspond to other hallmarks of breast cancers. They were not shown here due to space limits.

2) Drug mechanism investigation: With the same biological graph and motifs used in Sec. III-B1, we now pick the motif-cliques with single drug nodes. An example is shown in Fig. 1, which shows that a drug called *Loxapine* is connected to a cluster of genes, which in turn link to Schizophrenia and Obesity nodes. It has been known that Loxapine is a common drug for Schizophrenia. Because Loxapine has a close relationship with the Obesity node, this indicates that obesity is likely a side effect of Loxapine. Hence, motif-cliques can allow the user to study medicine characteristics and find potential side effects of drugs.

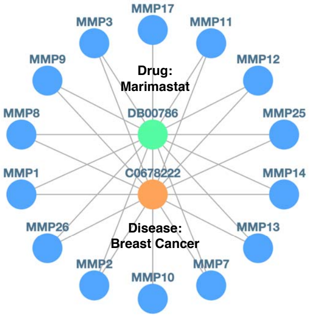


Fig. 7: A subtype of breast cancer.

3) Drug repurposing: This topic has gained much attention from pharmaceutical companies since it is important to use existing drugs for new therapeutic purposes. Based on the same biological graph and motifs used in Sec. III-B1, we select the motif-cliques with single disease nodes. This enables the finding of the new usage or purpose of an existing drug. For example, in Fig. 8, two drug nodes, i.e., Loxapine and Doxepin, and the Schizophrenia disease nodes, are all connected to the same set of genes. Since Loxapine is a common medicine for the treatment of Schizophrenia, the following conjecture can be produced: *Doxepin can heal Schizophrenia.*

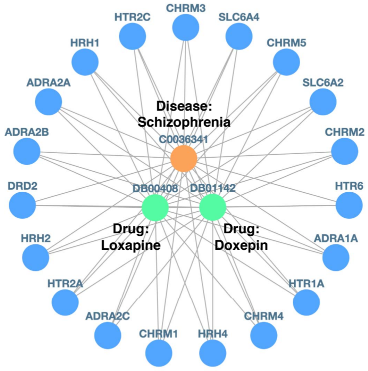
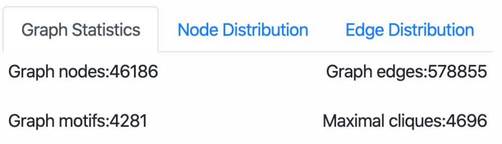


Fig. 8: Doxepin repurposing.

C. *Motif-clique Analysis*

The *MC-Explorer* has a motif-clique analysis module, which generates statistics for motifs and motif-cliques. Fig. 9 shows how *MC-Explorer* displays the number of nodes and edges of the input graph, as well as the number of motifs and maximal cliques. We used the state-of-the-art algorithm in [3] to count motifs, and the Bron-Kerbosch algorithm [2] to count the maximal cliques.

Another function of the module is to collect statistics about the motif-cliques generated for a given motif. In Fig. 10, the node and edge distributions of the motif-cliques used for disease subtyping are shown. The two distributions are quite similar. Also, the sizes of motif-cliques tend to be small, which have fewer than 100 nodes and 200 edges.



Graph nodes:46186

Graph edges:578855

Graph motifs:4281

Maximal cliques:4696

Fig. 9: Statistical analysis facilities of *MC-Explorer*.

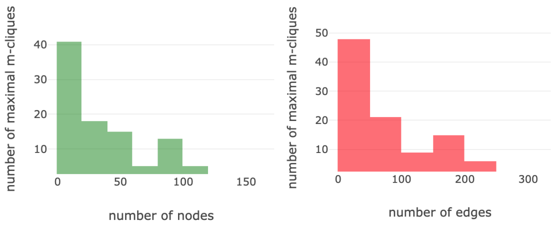
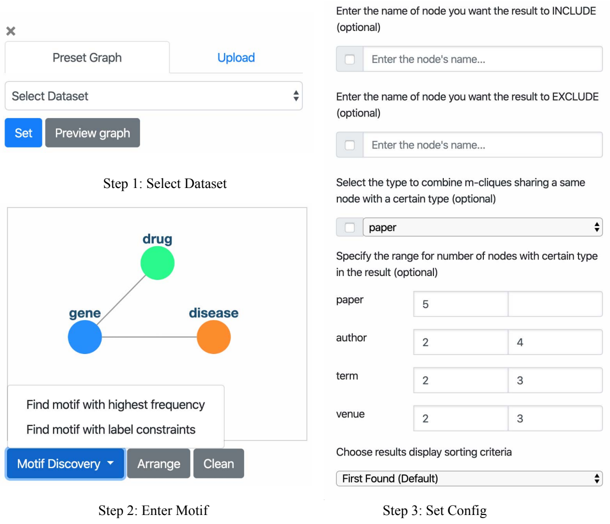


Fig. 10: Node and edge distributions of the first 100 discovered maximal motif-cliques for drug subtyping.

IV. Demonstration

We will engage the audience closely in the demo. The audience will be first asked about the the domain he/she is interested: biological analysis, e-commerce, and bibliographical network analysis. We will prepare the datasets corresponding to these domains. Based on the user's choice, the related data will be loaded, and the user can visualize and explore the related dataset in our web user interface. As shown in Fig. 3, the user can preview the graph, and inspect nodes and edges they are interested in. If the edges are weighted, the user can use the edge weight threshold sliding bar on the right panel to retain or remove the nodes and edges from the canvas. The user can highlight certain nodes by searching their names. If nodes are associated with properties in the dataset, properties will be collected, aggregated and displayed on the right panel. The user can click on one or more properties to highlight nodes with these properties.



Step 1: Select Dataset

Step 2: Enter Motif

Enter the name of node you want the result to INCLUDE (optional)

□ Enter the node's name...

Enter the name of node you want the result to EXCLUDE (optional)

□ Enter the node's name...

Select the type to combine m-cliques sharing a same node with a certain type (optional)

□ paper

Specify the range for number of nodes with certain type in the result (optional)

paper	5	
author	2	4
term	2	3
venue	2	3

Choose results display sorting criteria

First Found (Default)

Step 3: Set Config

Fig. 11: Steps to search motif-cliques.

By clicking on the "Enter Motif" button, a motif panel will be displayed, as shown in Fig. 11. The audience can draw a motif on canvas by adding nodes and edges. Alternatively, the user can select a motif suggested by the system. The audience can choose a motif which is statistically more significant than other subgraphs, or a motif that is both frequent and conforms to label constraints given by the user. The user can also set up custom configurations. For example, the user can require the results to include some specific nodes or to exclude some other nodes. Afterwards, the audience can click on the "Start" button on the top right to submit a motif-clique search request. The server will start searching and return discovery and analysis results. The user can then visualize motif-cliques and the related statistics. Finally, we will show the three applications of motif-cliques used in biological analysis.

Acknowledgement. Jiafeng Hu and Reynold Cheng were supported by the Research Grants Council of Hong Kong (RGC Projects HKU 17229116, 106150091, and 17205115), and the University of Hong Kong (Projects 104004572, 102009508, and 104004129), and the Innovation and Technology Commission of Hong Kong (ITF project MRP/029/18). Kevin Chen-Chuan Chang was supported by National Science Foundation IIS 16-19302 and IIS 16-33755, UIUC OVCR CCIL Planning Grant 434S34, and UIUC CSBS Small Grant 434C8U.

References

[1] J. Binchi, E. Merelli, M. Rucco, G. Petri, and F. Vaccarino. jholes: A tool for understanding biological complex networks via clique weight rank persistent homology. *Electronic Notes in Theoretical Computer Science*, 306:5–18, 2014.

[2] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

[3] V. Carletti, P. Foggia, A. Saggese, and M. Vento. Challenging the time complexity of exact subgraph isomorphism for huge and dense graphs with vf3. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):804–818, 2017.

[4] Y. Fang, R. Cheng, X. Li, S. Luo, and J. Hu. Effective community search over large spatial graphs. *PVLDB*, 10(6):709–720, 2017.

[5] Y. Fang, R. Cheng, S. Luo, and J. Hu. Effective community search for large attributed graphs. *PVLDB*, 9(12):1233–1244, 2016.

[6] Y. Fang, X. Huang, L. Qin, Y. Zhang, W. Zhang, R. Cheng, and X. Lin. A survey of community search over big graphs. *VLDBJ*, 29(1):353–392, 2020.

[7] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

[8] J. Hu, R. Cheng, K. C.-C. Chang, A. Sankar, Y. Fang, and B. Y. Lam. Discovering maximal motif cliques in large heterogeneous information networks. In *ICDE*, pages 746–757. IEEE, 2019.

[9] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2013.

[10] Z. Li, Y. Fang, L. Qin, J. Cheng, R. Cheng, and J. C. Lui. Walking in the cloud: parallel simrank at scale. *PVLDB*, 9(1):24–35, 2015.

[11] G. Micale, R. Giugno, A. Ferro, M. Mongiovì, D. Shasha, and A. Pulvirenti. Fast analytical methods for finding significant labeled graph motifs. *Data Mining and Knowledge Discovery*, 32(2):504–531, 2018.

[12] M. Mongioví, G. Micale, A. Ferro, R. Giugno, A. Pulvirenti, and D. Shasha. glabtrie: A data structure for motif discovery with constraints. In *Graph Data Management*, pages 71–95. Springer, 2018.

[13] G. Petri, M. Scolamiero, I. Donato, and F. Vaccarino. Topological strata of weighted complex networks. *PloS one*, 8(6):e66506, 2013.

[14] J. Piñero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, 2015.

[15] R. A. Rossi, D. F. Gleich, A. H. Gebremedhin, and M. M. A. Patwary. Fast maximum clique algorithms for large graphs. In *WWW*, pages 365–366. ACM, 2014.