Database Systems Project Report

(Implementation flavour)

Submitted By: Sai Sankar Bhuvanapalli (SB23W), Sai Tharun Reddy Darukumalli (SD23BB) ,

Siva Prasad Reddy Bandi (SB23BW)

Title of our Paper : On Analysing Graphs with Motif-Paths

Paper Authors: Xiaodong Li† , Reynold Cheng† , Kevin Chen-Chuan Chang‡ , Caihua Shan† , Chenhao Ma† , Hongtai Cao‡

1. Abstract:

Path-based techniques have proven useful for a number of graph analysis tasks, including graph clustering and link prediction, but they have issues with large and complicated graphs. Recently, there has been an increase in interest in motif-based analysis, which views motifs as basic building blocks of a small graph with a few nodes. This method performs better than conventional edge-based solutions and captures higher-order structures between nodes. The study investigates the use of motif-paths concatenations of one or more motif instances in three path-based mining tasks: node ranking, local graph clustering, and link prediction. The usefulness of motif-paths in enhancing graph analysis is demonstrated by experimental findings on actual graph datasets.

2. Introduction:

Analysing the complexity and interconnection within graphs, particularly in domains like social network analysis and biological network analysis, has posed challenges. Despite various attempted techniques, motif-paths have emerged as effective tools for extracting meaningful information from graphs. Motif-paths are collections of motifs or subgraphs that occur in a graph, providing a comprehensive understanding of information flow in social networks or genetic pathways in biological networks. With graphs becoming integral data formats in diverse fields such as social networking, biology, and transportation, they serve as essential tools for modelling and analysing complex processes. The careful analysis of these graphs is crucial for comprehending the significance of depicted facts. A promising avenue for this field involves exploring motif-paths within graphs, offering a fresh perspective on graph analysis.

2.1 Graph Analysis: Motif-Paths:

Graph sequences of motifs are called motif-paths, and they provide a more detailed view of the topology of the graph than do individual motifs. The ability to capture intricate relationships and uncover hidden patterns has drawn attention to the study of motif-paths. The analysis of motif-paths in various contexts has been investigated in several papers. In a graph, motif-paths are collections of motifs connected by a network of edges. Compared to traditional motifs, this method seeks to capture more complex structural information by taking into account the arrangement and connectivity of motifs within the network. A fundamental concept in graph theory is the idea of motif pathways, which associates motifs with paths in graphs. A subgraph's recurring patterns are called motifs, while a route is a group of nodes connected sequentially. Using motif-path analysis in big networks has immense potential for finding significant patterns and outliers.

2.2 Graph Mining of Motif Path:

This paper presents a novel motif-path framework for graph mining, focusing on link prediction, local clustering, and node ranking. The authors alter well-known baseline algorithms like Betweenness Centrality, Graph Distance, Katz Index, and Local Graph Clustering to use motif-paths instead. The authors compare the accuracy of the modified algorithms with the baselines to demonstrate enhanced performance. The suggested motif-path technique is superior to earlier methods taken separately, with certain baselines like PageRank indirectly compared based on published results. The algorithms are chosen to encompass methods that use paths or motifs independently, emphasizing the benefits of merging both within the motif-path structure.

2.3 Research Problem:

The paper highlights drawbacks in conventional edge-based approaches for graph analysis, underscoring their failure to account for the higher-order structure of graphs, resulting in potentially inaccurate outcomes. This higher-order structure encompasses intricate connectivity patterns among nodes that surpass simple pairwise relationships, such as the formation of triangles or cycles. Traditional methods focusing solely on pairwise interactions may overlook these crucial higher-order patterns, impeding a comprehensive understanding of the graph's overall structure and function. Additionally, the paper notes that the higher-order graph is often fragmented into multiple connected components, even when the original graph is connected. So we have implemented a motif-path based solution to overcome all these problems in our project and used them in various fields like graph clustering, link prediction, and node ranking.

2.4 Contributions:

Our main contribution involves the successful adaptation and implementation of the provided Java code base into the C++ programming language. This crucial step not only enhances the accessibility of the proposed methodology but also addresses compatibility issues for a broader range of users. Our meticulous translation ensures that the core functionality and structure of the algorithm remain intact, maintaining fidelity to the original design. Additionally, we conducted extensive experiments using the C++ implementation to validate its performance and verify its consistency with the results presented in the paper. By providing a C++ version, we aim to facilitate broader adoption and usage of the proposed approach within the programming community.

3. Related Work:

The study presented in this paper [1] has greatly contributed in the field of a network analysis by proposing tensor spectral clustering for breaking down higher-order network structures. Their study stresses the necessity of recognizing higher-order patterns in networks, which goes beyond standard methodologies that exclusively focus on motifs. While this strategy has proven useful in specific instances, it does have limitations. Tensor spectral clustering can be computationally expensive and may struggle with scalability when dealing with large networks, but motif-based analysis frequently lacks the ability to capture more complicated higher-order patterns and relationships inside the network. In this context, our research offers the novel notion of motif-path analysis. It intrinsically captures higher-order relationships inside the network, providing a more sophisticated view of network topologies. This method addresses the constraints of prior methods by combining the capabilities of tensor spectral clustering with motif-based analysis.

A general framework for random walk-based graphlet statistics estimation was presented by Xiaowei Chen et al. [2]. A wider understanding of network architecture is possible with this method since it is more flexible when dealing with larger and more complex motifs. It considers a wider range of subgraph structures, which helps to mitigate some of the drawbacks of combinatorial techniques. When handling specific themes or capturing incredibly fine-grained, precise patterns, it could still encounter difficulties. Additionally, employing random walks might add a stochastic component, which isn't always ideal, especially when deterministic outcomes are required. Our study tackles these issues by providing motif paths as a deterministic and extremely flexible method, guaranteeing both accuracy and adaptability in identifying intricate network architectures.

The research [3] focuses on motif-aware graph clustering, emphasizing the importance of scalable clustering techniques in vast networks for successful community detection. It investigates approaches for recognizing motifs in the context of community detection, with the goal of revealing hidden network structures. This paper's scalability when dealing with large networks is a problem. It may not completely exploit motifs' potential for capturing complicated community patterns. Our approach, on the other hand, overcomes these constraints by introducing motif paths and providing a thorough framework for graph analysis. We overcome scalability issues and successfully employ motifs for community discovery by going beyond earlier motif-aware clustering algorithms. Motif-paths enable us to capture both local and large-scale motifs, resulting in a dependable approach for scaling motif-aware community recognition.

This work [4] proposes a link prediction method based on higher-order motif features. It extends the common neighbours-based techniques by introducing the use of motifs of different sizes for link prediction. It approaches link prediction as a supervised classification issue, proposing a structured framework for extracting motif-based characteristics. The approach's reliance on certain motifs, however, may limit its capacity to capture the full spectrum of high-order structural dependencies inside the network. Here is where our model comes in. It broadens the concept of motifs by incorporating motif-paths, which capture not just motifs but also structural links and sequences of motifs in the graph. This novel method overcomes the drawbacks of motif-based link prediction, increasing its adaptability to a variety of graph mining applications and improving link prediction accuracy by considering both short- and long-range structural relationships.

The 2014 paper by Xin Huang and co-authors [5] focuses on community detection in vast and dynamic networks, notably presenting the concept of k-truss communities, which are created by nodes connected by triangles in the graph. While it gives useful insights on community identification, the paper's technique relies heavily on triangle-based motifs for structural analysis, which has limits. It may not efficiently address more complicated structural patterns in networks beyond triangles, restricting its application to contexts where triangles are dominant motifs. In contrast, our approach introduces the concept of motif paths, extending motif analysis to a broader spectrum of high-order structural patterns. This method improves network analysis's adaptability and scalability, making it appropriate for a range of real-world situations in which distinct motifs are essential to comprehending the structures and behaviours of networks. The shortcomings of the 2014 research are overcome by our model's adaptability and wider motif coverage, providing a better and more thorough answer for network analysis tasks.

In this study [6], the authors presented the idea of edge-centric local graphlets, or subgraphs defined by the relationships between edges in a graph. To count these graphlets in large-scale networks and disclose information about local network designs, they develop the efficient algorithm E-CLoG. The emphasis on a certain kind of graphlet, however, is a shortcoming of this method that can limit its use to different graph analysis tasks. Alternatively, our work introduces motif-paths, which provide a holistic picture of the graph's structure, a more flexible and comprehensive methodology, and the ability to adapt to various graph mining issues. This allows for a more comprehensive understanding of high-order structural connections. This adaptability makes the model useful in the field of graph analysis, as it provides a more comprehensive understanding of the underlying structure of the graph.

The reference [7] which it mainly focuses on describes a node's graphlet orbit degree, which offers detailed details about the nearby local structures. Understanding node roles and functions in complex networks is made easier with the help of this metric. This paper focuses on the computationally intensive task of efficiently estimating node orbit degrees for large-scale graphs, which requires counting graphlets. The authors present an innovative sampling technique called SNOD for accurate node orbit degree estimation. They also offer a means of measuring the error in these approximations. In order to better understand the roles and functions of nodes in social and biological networks, the paper provides a fine-grained metric for analysing complex networks. This approach has the limited scope that is unable to link graphs that involve the prediction, local graph clustering and the node ranking that is also unable to capture the higher order structures whereas the motif paths approach is advanced and easy to solve these problems within the provided valuable approach.

This paper [8] introduces a Tensor Spectral Clustering (TSC) algorithm, addressing limitations in traditional spectral methods when dealing with higher-order network structures. Spectral methods, rooted in first-order Markov chains, often struggle to capture complex network patterns like triangles and cycles. TSC offers a solution by enabling explicit modeling and preservation of higher-order structures during graph partitioning. Leveraging tensor representations, the algorithm provides flexibility in modeling, while its multilinear spectral approach ensures efficient partitioning. The paper demonstrates the effectiveness of TSC in applications like discovering layered flows and anomaly detection, outperforming standard spectral methods, especially in preserving directed 3-cycles.

4. Technical Section:

The baseline algorithm establishes a foundation for the shortest motif-path search, providing a generic approach to finding the shortest motif-path from a source node s to a target node t in a graph. It introduces key concepts such as motif-instances, motif-connectivity, higher-order graphs, and motif-paths.

Input: $G = (V, E), \tau, (s, t) \in V \times V$;
Output: $|\mathcal{P}_{s,t}|$
$D \leftarrow \infty$, construct the higher-order graph $\mathcal{G}$ from G;
$S \leftarrow \{m | s \in V_m, m \simeq \tau\}, T \leftarrow \{m | t \in V_m, m \simeq \tau\}$;
for $m_i \in S, m_j \in T$ **do**
    $\mathcal{P} \leftarrow$shortest sequence of $m \simeq \tau$ linking m_i and m_j in $\mathcal{G}$.
    Update $D \leftarrow |\mathcal{P}|$ if $D > |\mathcal{P}|$;
return D.

Figure 1 : baseline algorithm

SMP Algorithm:

SMP is an algorithm that finds the shortest path between two nodes in a graph adhering to a specified motif. It uses the MOD-Index to efficiently search for motif-instances and calculate the shortest motif-path distance. The algorithm provides a lower-bound to the shortest motif-path distance by using the diameter of the motif to cover the shortest path. Pre-calculating shortest-path distances in the offline phase optimizes motif-path distance calculation, reducing redundancy and speeding up the search.

Input: $G = (V, E)$, τ, $(s, t) \in V \times V$
Output: $|\mathcal{P}_{s,t}|$

$Q \leftarrow \emptyset$, $d \leftarrow \infty$;
$Q.enqueue(s)$;
while $Q \neq \emptyset$ & t is not discovered **do**
  Rank $p \in Q$ according to $|\mathcal{P}^{p}_{s,t}|$;
  $q \leftarrow Q.dequeue()$;
  $M \leftarrow$ MODQ(τ, q, Φ), and remove m from M where m contains "searched" nodes or all node in m are marked "discovered";
  Mark q as "searched";
  for $v \in V_m$, $m \in M$, v is "undiscovered" **do**
    Mark v as "discovered";
    if $v = t$ **then return** $|\mathcal{P}_{s,t}|$;
    $Q.enqueue(v)$;
return ∞.

Figure 2: SMP Algorithm

Link Prediction :

The algorithm employs motif-paths to predict missing links between nodes, extending popular path-based approaches like Katz Index and Graph Distance. Motif-path versions, MKI and MGD, use motif-paths instead of traditional paths, enhancing link prediction accuracy by capturing structural patterns. The algorithms generate scores denoting the probability of missing links, providing a nuanced approach to link prediction in graph structures.

Input: $G = (V, E)$, τ, $(s, t) \in V \times V$ & $(s, t) \notin E$
Output: $g(s, t)$

#Motif-path based Graph Distance (MGD)
$|\mathcal{P}_{s,t}|$ =SMP(G, τ, s, t);
return $g_{MGD}(s, t) = \frac{1}{|\mathcal{P}_{s,t}|}$.

#Motif-path based Katz Index (MKI)
$C_1 \leftarrow \{s\}$;
for $l \leftarrow 1 : L$ **do**
  $M \leftarrow$ MODQ(τ, p, Φ), $p \in C_l$;
  $|\mathbb{MP}^{l}_{s,t}| \leftarrow |\{t' | t' \in V_m, m \in M, t' = t.\}|$;
  $C_{l+1} \leftarrow \{x | x \notin C_l, x \in Vm, m \in M.\}$;
return $g_{MKI}(s, t) = \sum_{l=1}^{L} \epsilon^{l-1} \cdot |\mathbb{MP}^{l}_{s,t}|$.

Figure 3: Link prediction

Local Graph Clustering:

Motif-path is integrated into Local Graph Clustering (LGC), where the goal is to find the k-nearest neighbours of a query node based on shortest motif-path distance. MLGC extends LGC by considering motif-paths, improving the effectiveness of local clustering. The algorithm identifies the k-nearest neighbours with respect to motif-paths, enabling a more context-aware exploration of local clusters in the graph. This approach is particularly useful for applications where understanding higher-order structural motifs is critical for accurate clustering.

Input: $G = (V, E)$, τ, k, $s \in V$
Output: C

Run SMP(G, τ, s, ∞) until k nodes are marked as "discovered";
$C \leftarrow \{c | c$ is marked as "discovered"$\}$;
return C.

Figure 4 : graph clustering

Node Ranking:

Motif-path is leveraged for node ranking, specifically using the motif-path-based Betweenness Centrality (MBET) for measuring node influence. The algorithm calculates the centrality of a node by considering the number of motif-paths passing through it. MBET provides a more nuanced measure of node influence, incorporating structural motifs. This approach is especially valuable for ranking nodes in graphs where higher-order structures play a crucial role in determining node importance. The algorithm involves a two-phase process, ensuring efficient computation of motif-path-based centrality scores.

Input: $G = (V, E)$, τ, $v \in V$
Output: $C(v)$

$\delta_v \leftarrow 0$, $\delta \leftarrow 0$, generate W from G;
for $s, t \in V \times V$, $s \neq t \neq v$ **do**
    $l \leftarrow 1$;
    while t is not discovered in W **do**
        Calculate $\mathbb{P}^l_{s,t}$ on W;
        if $|\mathbb{P}^l_{s,t}| > 0$ **then**
            $\delta_v \leftarrow \delta_v + |\{P^l_{s,t} | v \in P^l_{s,t}, P^l_{s,t} \in \mathbb{P}^l_{s,t}\}|$;
            $\delta \leftarrow \delta + |\mathbb{P}^l_{s,t}|$, break;
        $l \leftarrow l + 1$;
return $C(v) = \delta_v / \delta$.

Figure 5: Node Ranking Algorithm

5.1 Methodology:

The methodology used in this paper involves proposing several algorithms for analysing graphs with motif-paths, including the Motif-Path-Based Betweenness Centrality (MBET) algorithm for node ranking and the Multi-Level Graph Clustering (MLGC) algorithm for local graph clustering and . The paper also explores the use of motif-paths in other graph analysis tasks, such as link prediction and local graph clustering. The MBET algorithm uses a two-phase approach to calculate the all-pairs shortest motif-path and then applies the Betweenness Centrality algorithm to obtain the centrality score for each node. The algorithm is designed to handle large datasets and can be used to rank nodes based on their importance or centrality in the graph.

The MLGC algorithm uses a recursive approach to cluster nodes based on their motif-paths. The algorithm is designed to handle higher-order graph fragmentation and can be used to reveal higher-order graph semantics. The algorithm works by recursively clustering nodes based on their motif-paths, starting with the most frequent motifs, and then expanding to less frequent motifs. The paper also proposes several algorithms for analysing graphs with motif-paths, including node ranking, and motif-based link prediction. Each of these algorithms uses a different approach to analyse the graph and can be used to reveal different aspects of the graph's structure.

5.2 Software Tools:

The language that we aimed to implement in the paper is C++ programming, and we took two types of real-world data sets to evaluate the algorithms, which are social networks and protein-protein interaction networks . The libraries that are used to implement this model are the boost graph libraries. The main data sets that are used are GAVI and KCORE for the PPI networks and AMAZ, YOUT for the social networks. The algorithms were tested on a computer with an 32 GB of RAM and Intel Core i7-6700k CPU.

5.3 Implementation:

The proposed algorithms were implemented using C++. We implemented the shortest motif path algorithm which with the help of motif instances gives the shortest path between two nodes in a graph.

```
Base Graph:
Node 0:
Node 1: 4 5 2 3
Node 2: 4 1 4 3 5
Node 3: 1 2
Node 4: 1 7 2 6 2
Node 5: 1 2
Node 6: 4 7 9 8
Node 7: 4 8 6
Node 8: 7 9 6
Node 9: 8 6

Motif Instances:
Motif Instance: 1 2 3
Motif Instance: 1 2 4
Motif Instance: 1 2 5
Motif Instance: 4 6 7
Motif Instance: 6 7 8
Motif Instance: 6 8 9

Shortest Motif-Path: 1 -> 4 -> 7 -> End
```

Figure 6: implementation of SMP algorithm

With the implementation of the link prediction algorithm, it gives us MKI (Katz index) and MDK (Graph distance) scores. MKI has demonstrated effectiveness in predicting missing links within graphs by leveraging motif-based k-nearest neighbors, offering a simple and efficient solution for large datasets. The choice between MKI and MGD depends on the specific requirements of the graph analysis task, with MKI being more suitable for efficiency and scalability, and MGD excelling in capturing intricate higher-order structures within the graph.

```
Motif-path based Katz Index (MKI) score: 3
Motif-path based Graph Distance (MGD) score: 1

--------------------------------
Process exited after 0.02029 seconds with return value 0
Press any key to continue . . .
```

Figure 7: Implementation of link prediction algorithm

Graph clustering algorithm gives us a motif count to each node. the Multi-Level Graph Clustering (MLGC) algorithm, leveraging motif-paths for local graph clustering. This approach identifies local clusters based on higher-order structural patterns. Graph clustering algorithms, like MLGC, are crucial for uncovering meaningful patterns in large-scale graphs, providing insights into functional, social, or topological properties. Overall, these algorithms reveal cohesive node groups, exposing inherent communities not immediately evident from raw node connectivity.

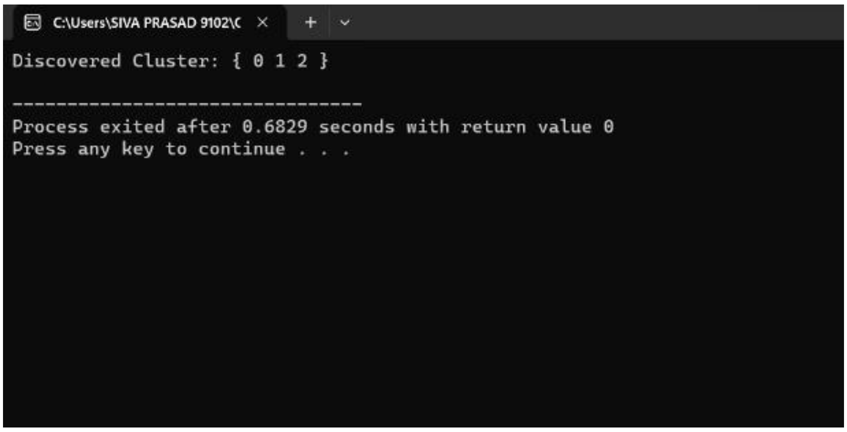

Figure 8 : Implementation of graph clustering

Node ranking serves as a valuable graph analysis task, offering insights into the structure and significance of nodes within a graph. The paper introduces the MBET algorithm, presenting an innovative method for node ranking that utilizes motif-paths to grasp higher-order graph semantics.

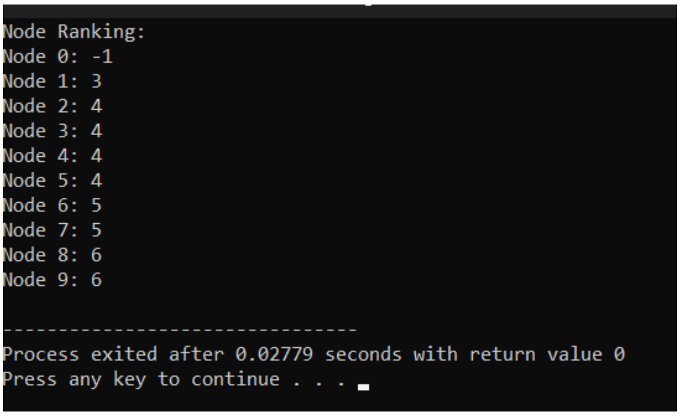

Figure 9: Node ranking algorithm Implementation.

6. Evaluation:

The evaluation methodology employed in the study encompasses the utilization of diverse real-world graph datasets, distinguishing between protein-protein interaction (PPI) networks (GAVI and KCORE) and social networks (AMAZ, and YOUT). In addition to metrics, the paper introduces several experimental details. The proposed motif-path methods are benchmarked against baseline approaches (e.g., Katz Index, Graph Distance, traditional clustering methods) for each task. Comparative analyses with state-of-the-art motif-based techniques are conducted, and runtimes of different methods are reported. The evaluation encompasses various motif patterns, such as triangles, cycles, and cliques, and assesses the impact of handling graph fragmentation through bridging edges.

7. GitHub Repo: https://github.com/saisankar20/cop5725projectdemo/tree/main

8. Conclusions:

The paper's key discoveries highlight the ability of motif-paths to unveil higher-order graph semantics, enhancing the efficacy of various graph analysis tasks. The research introduces algorithms tailored for motif-path analysis, including the Motif-Path-Based Betweenness Centrality (MBET) for node ranking, Multi-Level Graph Clustering (MLGC) for localized graph clustering and Motif-path based Link Prediction for link prediction. Notably, MBET demonstrates superior performance compared to traditional methods like PageRank and Weighted PageRank, while MLGC adeptly addresses higher-order graph fragmentation, exposing nuanced graph semantics. These findings signify the potential of motif-paths to enhance the precision of tasks such as node ranking and local graph clustering, revealing higher-order graph semantics overlooked by conventional analysis methods. The implications extend across diverse domains, including social network analysis, bioinformatics, and recommendation systems.

Future research avenues may explore the integration of motif-paths into other graph analysis tasks like community detection and link prediction. The study underscores motif-paths as a potent tool for uncovering higher-order graph semantics, prompting further investigation into its application across varied graph analysis domains. Optimization strategies for the proposed algorithms in the context of extensive datasets could also be a focus of future research.

To conclude, this paper introduces motif-paths as a robust mechanism for exposing higher-order graph semantics, offering multiple algorithms tailored for motif-path analysis. The outcomes suggest that motif-paths can significantly enhance the effectiveness of graph analysis tasks, presenting implications across a broad spectrum of applications.

9. Team Contributions:

9.1 Team Member 1:

Sai Tharun Reddy Darukumalli (SD23BB): Sai Tharun Reddy Darukumalli contributed to the project, providing valuable insights and ideas for the implementation of motif-path analysis techniques and also implemented C++ code for both motif-path algorithm SMP and also implemented C++ code for Graph clustering. He was responsible for conducting research on existing graph analysis methods and identifying areas where our work could make a unique contribution. Additionally, he played a major role in the development of the algorithms used to analyse motif paths in complex graph structures.

9.2 Team Member 2:

Siva Prasad Reddy Bandi (SB23BW): Siva Prasad Reddy Bandi was instrumental in the implementation of our motif-path analysis techniques. He was responsible for writing the code in C++ and using the boost graph libraries to create efficient algorithms for identifying and analysing motif paths in large datasets. Additionally, he played a key role in testing and refining our algorithms to ensure that they were accurate and effective.

9.3 Team Member 3:

Sai Sankar (SB23W): Sai Sankar contributed to the development of our motif-path analysis techniques and provided insights into the practical applications of our work. He was responsible for conducting experiments on real-world datasets to test the effectiveness of our algorithms and identify areas where further improvements could be made. Additionally, he played a key role in analysing the results of our experiments and presenting our findings in a clear and concise manner.

10. References:

[1] Austin R Benson, David F Gleich, and Jure Leskovec. 2015. Tensor spectral clustering for partitioning higher-order network structures. In Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM, 118–126.

[2] Xiaowei Chen, Yongkun Li, Pinghui Wang, and John Lui. 2016. A general framework for estimating graphlet statistics via random walk. Proceedings of the VLDB Endowment (2016).

[3] Charalampos E Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. 2017. Scalable motif-aware graph clustering. In Proceedings of the 26th International Conference on the World Wide Web. International World Wide Web Conferences Steering Committee, 1451–1460.

[4] Ghadeer AbuOda, Gianmarco De Francisci Morales, and Ashraf Aboulnaga. 2019. Link prediction via higher-order motif features. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 412–429.

[5] Xin Huang, Hong Cheng, Lu Qin, Wentao Tian, and Jeffrey Xu Yu. 2014. Querying the k-truss community in large and dynamic graphs. In Proceedings of the 2014 ACM SIGMOD. ACM, 1311–1322.

[6] Vachik S Dave, Nesreen K Ahmed, and Mohammad Al Hasan. 2017. E-CLoG: counting edge-centric local graphlets. In 2017 IEEE International Conference on Big Data (Big Data). IEEE, 586–595.

[7] Pinghui Wang, Junzhou Zhao, Xiangliang Zhang, Jing Tao, and Xiaohong Guan. 2018. SNOD: a fast sampling method of exploring node orbit degrees for large graphs. Knowledge and Information Systems (2018), 1–26.

[8] Austin R. Benson, David F. Gleich, and Jure Leskovec. 2015. Tensor spectral clustering for partitioning higher-order network structures. In Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM, 118–126.