

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**Local Information and Structures in Analysis and
Modelling of Complex Networks**

by

Mingshan Jia

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

July 2022

Certificate of Authorship/Originality

I, Mingshan Jia, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Mingshan Jia

Signature: Production Note:
Signature removed
prior to publication.

Date: July 2022

ABSTRACT

Local Information and Structures in Analysis and Modelling of Complex Networks

by

Mingshan Jia

Abstracting entities and their interactions as nodes and links, networks are a general representation for modelling and studying complex systems. Modelling relational structures of the underlying data, rather than only a set of isolated entities, allows us to build more accurate models for various types of domain data, such as social relationships, molecular interactions, program executions, and many more. Despite being powerful and ubiquitous, networks are also difficult to process, mainly due to their complex topological structures. Therefore, the study of network structure, especially local structure, has been the core theme of studying complex networks. This dissertation aims to provide new understandings of how local structure information is extracted and utilised in studying different types of complex networks.

The dissertation includes three original works in the direction of local structure and information on top of a comprehensive survey. In the review, we propose new taxonomies for graph structures that bring together the notions of centrality measures, motifs, and other local-level metrics. For theoretical understanding, we propose new metrics to quantify the formation of 3-node and 4-node subgraphs and develop new motif patterns that are distinctive features in both network- and node-level analysis. For methodological approaches, we propose the framework to effectively encode edge attributes into the typed-edge graphlet degree vector, for both sociocentric and egocentric networks. Moreover, for practical applications, the proposed metrics and approaches are applied in many different types of complex net-

works and case studies. They are not only proven to be effective in multiple learning and analytical tasks but also lead to new insights and interesting discoveries.

Dissertation directed by Professor Katarzyna Musial-Gabrys and Professor Bogdan Gabrys

School of Computer Science, Data Science Institute, Complex Adaptive Systems Lab, UTS

Acknowledgements

Although pursuing a PhD is not an easy path to take, looking back on the past three years and nine months, it might be one of the best decisions I have ever made. I am grateful that I did not give up in some very tough situations. It eventually led me to meet my current supervisors and other wonderful people, who have advised me, supported me, worked with me and motivated me.

First and foremost, I am eternally grateful to my supervisors, — Professor Katarzyna Musial-Gabrys and Professor Bogdan Gabrys, for all the support and guidance they have provided throughout my PhD study. They gave me the opportunity to continue my PhD at my most difficult time, and they have supported me, encouraged me and trusted me ever since. They introduced me to the world of network science and provided me with all kinds of advice, from idea selection and experiment design to paper writing and rebuttal. They not only helped me improve in academics, but also recommended me to participate in collaborative research projects with other universities, and gave me multiple opportunities to participate in teaching activities. Without them, I could not have achieved such an all-round growth. Thank you, Professor Katarzyna Musial-Gabrys and Professor Bogdan Gabrys.

Moreover, I would like to thank all the people who have advised me, worked with me and helped me along my PhD journey. I want to thank my co-authors, Maité Van Alboom, Liesbet Goubert and Piet Bracke. Thanks to you and my supervisors, we have conducted an exciting cross-disciplinary study about chronic pain. I want to thank Professor Wei Liu and Dr Yi Zhang for being on my candidature assessment panel. I also thank Professor Pasquale De Meo for the interesting and insightful discussions on multiple research topics.

Next, I want to express my gratitude to my colleagues and friends, who have

always been so kind and helpful whenever I seek advice or help. Their knowledge, diligence and perseverance are also what motivate me to move forward. Thank you, Joakim Skarding, Mohamad Barbar, Guanping Xiao, Bin Wang, Xiaolin Zhang, Yanbin Liu, Xiaohan Zhang, and Yu-Xuan Qiu.

Finally and most importantly, I would like to thank my family. It is your everlasting love and support that allow me to start this wonderful journey and make me fearless in the face of all difficulties. Thank you, my parents Lingxiang Feng, Zilai Jia, my wife Yuanyuan Liu, and my son Yiqian Jia. This is dedicated to you.

Mingshan Jia
Sydney, Australia, 2022

List of Publications

Journal Papers

- J-1. **M. Jia**, B. Gabrys and K. Musial, "Directed closure coefficient and its patterns," in Plos one 16.6 (2021): e0253822.
- J-2. **M. Jia**, B. Gabrys and K. Musial, "Measuring Quadrangle Formation in Complex Networks," in IEEE Transactions on Network Science and Engineering, vol. 9, no. 2, pp. 538-551, 1 March-April 2022.

Conference Papers

- C-1. **M. Jia**, B. Gabrys and K. Musial, "Closure Coefficient in Complex Directed Networks." International Conference on Complex Networks and Their Applications. Springer, Cham, 2020.
- C-2. **M. Jia**, M. Van Alboom, L. Goubert, P. Bracke, B. Gabrys, K. Musial. "Analysing Ego-Networks via Typed-Edge Graphlets: A Case Study of Chronic Pain Patients." International Conference on Complex Networks and Their Applications. Springer, Cham, 2021.
- C-3. **M. Jia**, M. Van Alboom, L. Goubert, P. Bracke, B. Gabrys, K. Musial. "Analysing Egocentric Networks via Local Structure and Centrality Measures: A Study on Chronic Pain Patients." 2022 International Conference on Information Networking (ICOIN). IEEE, 2022.

Contents

Certificate	ii
Abstract	iii
Acknowledgments	v
List of Publications	vii
List of Figures	xii
1 Introduction	1
1.1 Aim, Objectives and Significance	3
1.2 Methodology	6
1.3 Thesis Organisation	7
2 Literature Review	9
2.1 Motivation	9
2.2 Preliminaries and Background	11
2.2.1 Local vs. Global	11
2.2.2 Motifs vs. Graphlets	13
2.3 Graph structural measures	15
2.3.1 Subgraph Count Based Approaches	17
2.3.2 Subgraph Formation Based Approaches	25
2.3.3 Global Path Based Approaches	33
2.3.4 Message Passing Based Approaches	39

2.3.5	Hybrid Approaches	42
2.4	Discussion and Outlook	47
2.5	Conclusion	50
3	Directed Closure Coefficient	52
3.1	Introduction	52
3.2	Preliminaries	56
3.2.1	Clustering coefficient	56
3.2.2	Directed clustering coefficient	57
3.2.3	Closure coefficient	59
3.3	Closure Coefficient in Directed Networks	61
3.3.1	Closure coefficient in binary directed networks	61
3.3.2	Closure coefficients of particular patterns	63
3.3.3	Closure coefficient in weighted networks	66
3.3.4	Computational efficiency	68
3.4	Experiments and Analysis	68
3.4.1	Directed closure coefficient in real-world networks	69
3.4.2	Link prediction in directed networks	72
3.4.3	Case study in a weighted signed network	76
3.5	Additional Related Work Discussion	78
3.6	Conclusion	78
4	Measuring The Formation of Quadrangles	80
4.1	Introduction	80
4.2	Background and Motivating Example	84
4.2.1	Measuring Triangle Formation	85

4.2.2	A motivating example	86
4.3	Two Quadrangle Coefficients	87
4.3.1	I-quad coefficient	87
4.3.2	O-quad coefficient	90
4.3.3	Quadrangle coefficients in weighted networks	91
4.3.4	Computational cost	94
4.4	Experiments and Analysis	94
4.4.1	Quadrangle coefficients in real-world networks	94
4.4.2	Correlation with node degree	97
4.4.3	Network classification	101
4.4.4	Link prediction	104
4.4.5	Limitations and Future Directions	107
4.5	Related Work	109
4.6	Conclusion	111
5	Typed-Edge Graphlets	113
5.1	Introduction	113
5.2	Background and Preliminaries	115
5.2.1	Graphlets and orbits	116
5.2.2	Egocentric graphlets	117
5.3	Typed-Edge Graphlet Degree Vector	117
5.4	Typed-Edge Degree, Colored Graphlets and Heterogeneous Graphlets .	121
5.5	Experiments and Analysis	124
5.5.1	Dataset	125
5.5.2	Analysing pain grades via GDV and TyE-GDV	127

5.5.3 Predicting pain grades	130
5.6 Conclusion	132
6 Conclusion and future works	134
Bibliography	137

List of Figures

1.1	Methodology	5
1.2	Three verification steps in methodology	6
2.1	Structural measures on graphs.	9
2.2	Graphlets and their orbits [160]	13
2.3	Motifs vs. Graphlets	13
2.4	Subgraph count based measures.	15
2.5	Subgraph formation based measures.	23
2.6	Global path based measures.	31
2.7	Message passing based approaches.	37
2.8	Hybrid Approaches.	41
3.1	Classification diagram of local clustering measures.	52
3.2	Taxonomy of directed triangles.	53
3.3	Dealing with bidirectional edges.	55
3.4	Directed open triads.	62
3.5	Scatter plots of the local directed closure coefficient and the local directed clustering coefficient, with the Pearson correlation coefficient.	68

3.6	Average normalized closure coefficients of four patterns: head-of-path (HoP), mid-of-path (MoP), end-of-path (EoP) and cyclic (CYC). The dominant pattern in each network is labelled with its value.	69
3.7	Two scatter plots of the network BTC-ALPHA.	73
3.8	Two local enlarged scatter plots between weighted signed directed closure coefficient and node strength in the network BTC-ALPHA. . .	74
4.1	The i-quad coefficient and the o-quad coefficient in comparison with the clustering coefficient and the closure coefficient.	78
4.2	An example of the i-quad coefficient and the o-quad coefficient in a movie recommender network.	79
4.3	A motivating example.	83
4.4	Two types of open quadriads in a quadrangle.	86
4.5	Correlation of quadrangle coefficients and weighted quadrangle coefficients in three different networks.	90
4.6	Cumulative distribution curve of the i-quad coefficient $I(i)$ (in green colour) and the o-quad coefficient $O(i)$ (in purple colour) in six real-world networks of different types.	94
4.7	Correlation of two quadrangle coefficients with node degree in six real-world networks.	95
4.8	Two types of quadrangle formation via stub matching.	97
4.9	Two-dimensional visualisation of K-means clustering on PCA-reduced data, without and with quadrangle coefficients (left figure and right figure respectively).	97
4.10	Critical difference diagram of four classifiers with different feature sets.	104

4.11	An example of the coefficients proposed in related works, compared with our proposed quadrangle coefficients.	107
5.1	Graphlets of size 2–4 nodes with enumeration of orbits.	111
5.2	7 egocentric graphlets of 2 to 4 nodes. Ego node is painted in black. .	113
5.3	Degree distribution and edge type distribution of all patients.	122
5.4	Parallel coordinates plot of average GDV of different GCPS grades. Each coordinate represents the average number of graphlets belonging to that type.	123
5.5	Parallel coordinates plot of average TyE-GDV of different GCPS grades for two graphlets. Each coordinate represents the average number of edges belonging to that type.	125
5.6	Prototypes of GCPS grade-1 and GCPS grade-4.	126

Chapter 1

Introduction

Complex systems across various domains, such as biology, ecology, physics and social science, can be modelled as networks that abstract the interactions between system's components [16, 171]. Different from a simple grid graph or a line graph for image or text modelling respectively, the complexity of networks comes from their intricate topological structures. Therefore, understanding and exploiting graph structure, especially local structure, has always been a core theme in analysing complex networks and underlies various analytical and representative applications such as node-type classification [20, 115], link prediction [73, 120], anomaly detection [176, 6], and graph representation learning [86, 80].

Among manifold research approaches, local structural measures are one of the most important and influential ways of studying network topology. The study of structural measures is so ubiquitous that they often appear in different terms, such as the big family of centrality measures [148, 199, 43], the popular notion of motifs [163] and graphlets [161] and the set of subgraph formation measures such as the clustering coefficient [232], the closure coefficient [241], the square clustering coefficient [145], etc. These approaches, however, are still mostly limited to the oversimplified description of complex system, i.e., static, undirected and unlabelled networks and this thesis aims at addressing this limitation.

A preliminary and essential question to ask, when one ventures into the local structures' world, is what is local? Conventionally, it is assessed by the distance from a focal node, such as within 4- or 5-hop from the focal node. However, a subgraph

Table 1.1 : Number of different undirected and directed subgraphs, and their respective orbits (orbits are all the unique positions of a subgraph), depending on the size of the subgraphs[195].

	Undirected		Directed	
k	#Subgraphs	#Orbits	#Subgraphs	#Orbits
2	1	1	2	3
3	2	3	13	27
4	6	11	199	667
5	21	58	9,364	44,210
6	112	407	1,530,843	9,031,113

build from all nodes within a certain distance from a focal node can still be very complicated, simply because the number of nodes involved in it is theoretically unlimited. For example, even in an egocentric network where all alters are only 1-hop way from the focal node, the number of alters is still unlimited, which could result in complicated structures. Based on the current literature, the study of local structure seldom surpasses 5 nodes. The inherent complexity in structures beyond 5 nodes renders them infeasible for enumerating all possible subgraphs and their orbits (see Table 1). Therefore, our definition of local structure in complex networks is any structure containing no more than 5 nodes. Among them, we focus particularly on 3-node and 4-node structures, not only because they are efficient to compute, but also previous research has revealed that certain 3-node and 4-node structures are the critical building blocks or motifs (recurrent and statistically significant subgraphs or patterns of a larger graph) in different types of directed networks[163].

In parallel with the unremitting effort in studying graph structure, this thesis aims to make contributions to the area in three aspects. First, when it comes to the theoretical understanding of local structures, we first extend a recently proposed approach that assesses the edge clustering phenomenon (which is also a 3-node subgraph formation problem) to more complicated network models, including directed networks, weighted networks and signed networks. We then close a gap in measur-

ing the formation of 4-cycle structure by proposing the quadrangle coefficients, and further reveal their properties in various types of networks, and their correlations with node degree.

When it comes to the contributions to methodological approaches, we develop new methods for predicting missing links and classifying different types of networks using the developed metrics; we further propose a novel framework to encode edge type information in graphlets and generate a typed-edge graphlets degree vector, which contains both rich structural information and edge attributes. Last but not least, in the application area, through employing the proposed metrics and algorithms in different types of complex networks and case studies, we uncover interesting properties of those networks (such as the association of the perception of pain and the type of social ties), and demonstrate the usages and performances of our approaches in various analytical and machine learning tasks.

1.1 Aim, Objectives and Significance

The aim of this thesis is to provide new understandings, both theoretically and methodologically, on how local structure information is extracted and utilised in studying different types of complex networks.

On one hand, local structures and motifs, are building blocks of complex networks and they provide enlightening insight into network properties, functioning and analysis [172, 16]. There are fruitful results of their applications in biology, ecology, physics, social science, and many other disciplines. Various kinds of motifs have been found in different scenarios such as gene regulation, neurons, food webs, electronic circuits and World Wide Web [163, 162]. Therefore, studying the formation of local structures and motifs should be vital in understanding the property and functioning of complex networks. On the other hand, real networks are often accompanied by rich information about nodes and edges. And there are situations where we care

more about edge attributes than node attributes, for example, the cost of traffic in transportation networks [64], the type of interaction in biological networks [100] and the specific relationship in social networks [27]. However, fewer works have focused on leveraging edge attribute information in graph analysis, especially in the context of graphlets. Therefore, we aim to propose a new framework that effectively and meaningfully encodes edge attributes in complex networks.

Concretely speaking, first, we aim to deepen the understanding on subgraph formation and propose new metrics to measure them. We then aim to propose new structural approaches that also take link attributes into account. Finally, we aim to apply the proposed metrics and algorithms in different types of analytical and learning tasks in complex networks, such as node-type classification and link prediction, and network classification.

To fulfill the research aim, the following objectives are expected to be achieved:

- **Objective 1:** To conduct comprehensive literature review, and propose new taxonomies from the perspective of graph local structure.
- **Objective 2:** To advance the knowledge of assessing edge clustering in directed networks. The recently proposed closure coefficient [241] provides a new perspective on measuring local edge clustering. However, it cannot be applied to complex directed networks. We will close this gap by proposing the directed closure coefficient;
- **Objective 3:** To deepen the understanding of 4-node subgraph formation in complex networks. In many types of networks, quadrangles or 4-cycles appear at a much higher frequency than triangles, and become the most dominant motifs. However, there lacks the angle of measuring the formation of quadrangles based on the outer-node based open-quadrads, and we aim to close this gap;

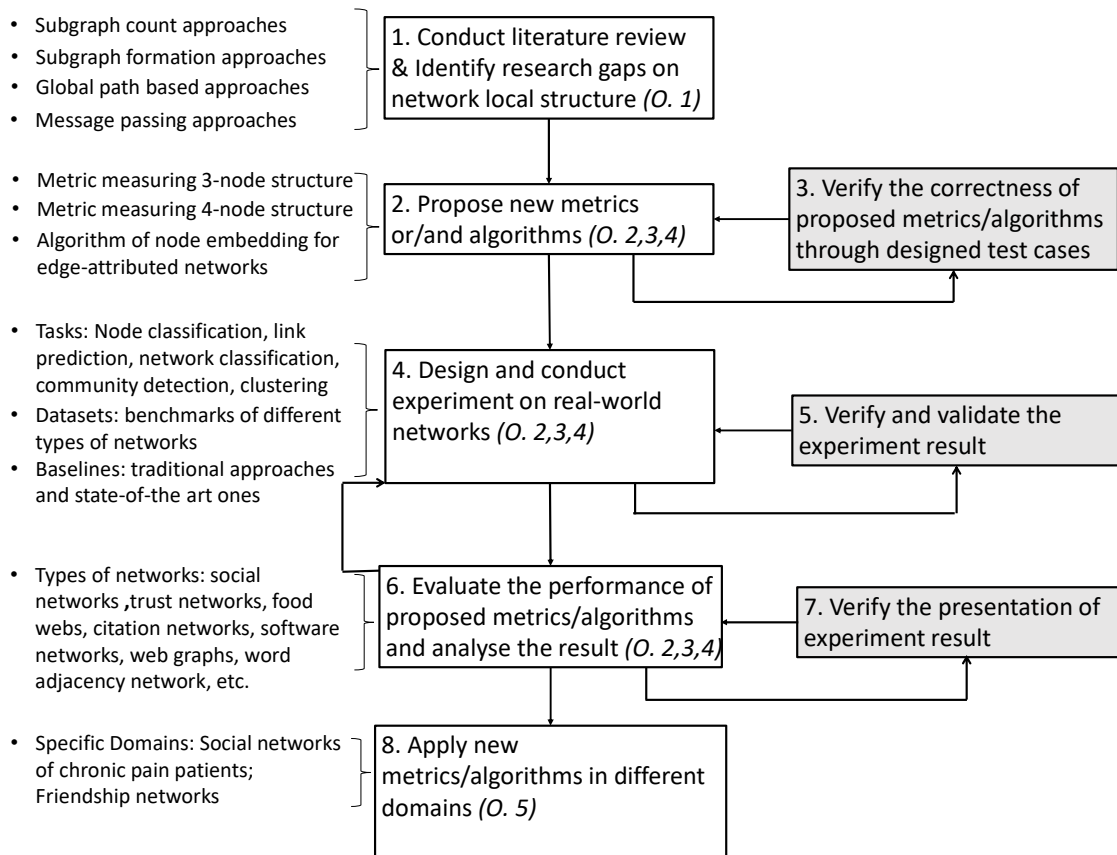


Figure 1.1 : Methodology

- Objective 4:** To bring new knowledge into the study of edge-attributed networks – Motivated by real-world network dataset and heterogeneous graphlets, we aim to propose a framework that can effectively and meaningfully embed edge attribute information in graphlets.
- Objective 5:** To verify the proposed approaches through their applications to various types of real-world networks and specific case studies, such as the social network of chronic pain patients, the friendship network of college students, etc.

1.2 Methodology

In order to attain the abovementioned objectives and close the research gaps, we have devised our research methodology which consists of eight major modules, as presented in Figure 1.1.

We first conduct a comprehensive literature review about network local structures, focusing on how the local information is extracted and applied in studying complex networks. We propose a taxonomy of five categories, i.e., subgraph count based measures, subgraph formation based measures, global path based measures, message passing based measures, and hybrid measures. This fulfills our first objective.

Then, based on the identified research gaps, we will propose two novel metrics for 3-node structures and 4-node structures respectively, and an algorithm that encodes edge attributes into graphlets (mapping back to our research objectives 2, 3 and 4 respectively). This stage also involves creating test cases to make sure that our proposed methods are validly implemented. We then perform multiple tasks on different types of real world networks, including node classification, link prediction, and network classification. Traditional approaches and state-of-the art methods are included as baseline approaches in these tasks.

To achieve the last objective 5, we apply our proposed metrics and algorithms in the study of particular types of networks, such as the social network of chronic pain patients and the friendship network of college students. An extra verification step is introduced to ensure that 1) our experiment setups are valid and the comparison is fair; 2) the dataset is valid (properly dealing without incorrect entries, repetitive entries and missing values) and when conducting experiments on multiple types of networks the dataset selection is balanced; and 3) our experiment result is valid. Finally, we evaluate the experiment results and discuss the main findings, after ver-

ifying that we present our results in an unbiased manner (choosing the appropriate metrics and visualisation techniques). The three-step verification flow is shown in Figure 1.2.

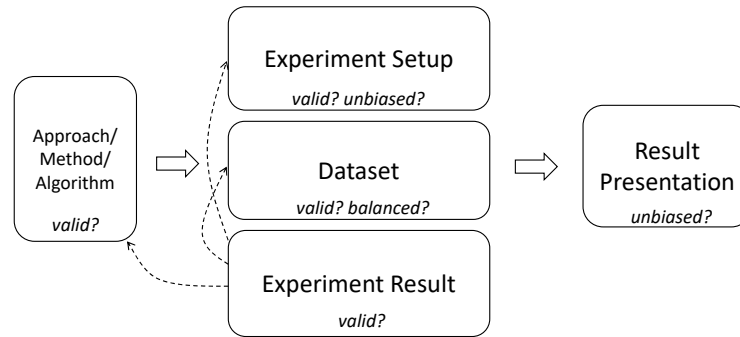


Figure 1.2 : Three verification steps in methodology

1.3 Thesis Organisation

This thesis is organised as follows:

- *Chapter 2*: This chapter presents the literature review on graph structural measures, including the discussion about the definitions of local measures and global measures, the differences and similarities between motifs and graphlets, and most importantly, our proposed taxonomies.
- *Chapter 3*: This chapter proposes novel approaches of assessing edge clustering in directed networks, including the novel metric of directed closure coefficient, the new link prediction method based on source and target closure coefficient and the utility of the four closure patterns. The content of this chapter is published in CNA 2020 and Plos One (C-1 and J-1 in the list of publications section).
- *Chapter 4*: This chapter formulates the formation of quadrangles in complex networks from a new perspective. The i-quad coefficient and the o-quad coefficient are proposed with the focal node at different positions. The content

of this chapter is published in IEEE Transactions on Network Science and Engineering (J-2 in the list of publications section).

- *Chapter 5*: This chapter proposes the framework to encode edge type information in graphlets, and generate a typed-edge graphlet degree vector. The vector enables us not only to better understand edge-labelled networks but also can be applied in node-level learning tasks. The content of this chapter is published in CNA 2021 (C-2 in the list of publication section) and we are invited to submit an extended work of it to Plos One.
- *Chapter 6*: The final chapter summarises the content and contributions of this thesis. Potential directions for future work are also given.

Chapter 2

Literature Review

This section covers our literature review on local network structures. First, we introduce the motivation for conducting this review. Then we discuss the basic notion of local and global in the context of complex networks as well as the important concepts of motifs and graphlets. Finally, we propose our taxonomy of graph structural measures.

2.1 Motivation

Understanding and exploiting graph structure has always been a core theme in analysing complex networks. The study of graph structures is so ubiquitous that they often appear in different terms, such as the big family of centrality measures [148, 199, 43], the popular notion of motifs [163] and graphlets [161] and the set of subgraph formation measurements such as the clustering coefficient [232], the closure coefficient [241], the square clustering coefficient [145], etc. How are these measurements different from each other and what are their usages in analysing complex networks? In this work, we aim to propose a new taxonomy and bring all these concepts together with a focus on graph structure. Specifically, as shown in Figure 2.1, we find most existing graph structural measures can be put in five categories: (i) subgraph count based measures, (ii) subgraph formation based measures, (iii) global path based measures, (iv) message-passing based measures, and (v) hybrid measures.

We now explain the logic behind our taxonomy. The first two categories both

covers a local area of the whole network (within a certain distance from the focal node, or with a limited number of nodes). The first category — subgraph count based approaches — is built from counting the number of particular local structures. For example, number of neighbours, local paths or subgraphs. The second category — subgraph formation based approaches — is uniquely defined based on the ratio of two subgraphs and thus bears the meaning of measuring the formation of certain local structures. To have both of them in the taxonomy instead of combining them in to one category is also to stress their differences.

Then, the third category expands its scope to the entire network. We name it as global path based approaches instead of global approaches. This is because all global approaches involves either the calculation of shortest paths or all paths originating from a node to any node in the entire graph. Notice here that path is also a particular type of graph. However, local path, such as 2-path or 3-path is in the category of subgraph count based measures, whereas global path or unbounded path is in another category. We choose to differentiate the third category with the previous two categories from the perspective the covered scope.

Next, the fourth category — message passing based approaches — is based on the idea of transmitting information along the edges. It is different from the above-mentioned three categories because it does not calculate any types of subgraphs of global paths. Instead, the structure is utilised in an implicit way. Every node is initialised with an importance score. Then iteratively, each node updates its score through aggregating the scores of its neighbours. Although these four categories are largely different from each other, there are many approaches that combine them together, which are naturally put into the fifth category — mixed approaches.

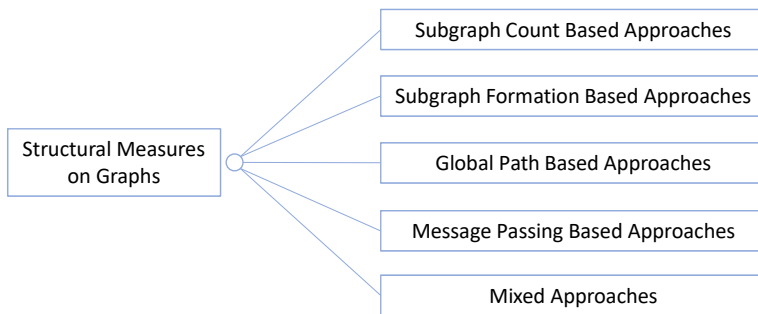


Figure 2.1 : Structural measures on graphs.

2.2 Preliminaries and Background

2.2.1 Local vs. Global

Before introducing the taxonomy of graph structural measures, it is helpful to first define what is local and what is global. Previous works [53, 99, 158, 155] either only focus on where the measures are defined on by dividing them into two or three categories: one is at the "local", "micro" or "individual" level, the other is at the "global", "macro" or "aggregate" level, and sometimes a third level of "mesoscopic", "quasi-local" or "subnetwork". Or they solely consider the range of information that is involved in their computation. This, however, leads to some problems. For example, betweenness centrality is defined at node-level, but it requires global information to compute. Should it be termed as local measure or global measure? Similarly, the average clustering coefficient is defined at network-level, but we only need local information, dividing the number of triangles over the number of wedges, at each node (then averaging over all nodes).

Therefore, we propose the following terms to distinguish both where the measures are defined on and the scope of information that is needed to calculate them:

- *Local-level measure* is a measurement defined on node-level or link-level (the

link here also includes non-existing link or potential link which is often used in link prediction task). Thus, It can be further divided into *node-level measure* and *link-level measure*.

- *Network-level measure* is a measurement defined for the whole network.
- *Local structural measure* is a measurement whose computation only involves the nearby neighbourhood of a node, i.e., within a range of k-hop away from a node. In most cases, k is less than or equal to 4. Many traditional measures only cares about the immediate neighbourhood around a node, and we name them as *Strict-local structural measure*.
- *Global structural measure* is a measurement that involves the global information in computation. Specifically, this type of measurement almost always involves the computation of paths between nodes in the network.

Now, when we revisit the previously mentioned betweenness centrality, it is both a local-level measure and a global structural measure. The average clustering coefficient, on the other hand, is both a network-level measure and a local structural measure. Some may argue that the average clustering coefficient involves the extra step of averaging over all nodes. Indeed, it is n times the complexity of computing the local clustering coefficient at a single node. However, when analysing networks, local-level measures are often calculated at the entire network, looping over all nodes or all edges. Moreover, any local-level measure can easily have an extended definition at network-level through aggregating over all nodes or edges. Therefore, when defining a measure as local structural or global structural, we choose to exclude this aggregation step and instead use the term network-level measure to distinguish it from its unextended counterpart.

2.2.2 Motifs vs. Graphlets

Next, we distinguish three similar concepts that are later used in our taxonomies, i.e., subgraphs, motifs and graphlets. Subgraph, as the name implies, is a smaller graph whose node set and edge set are subsets of those of the original graph. We then recap the notions of motifs [163] and graphlets [161] according to the papers that proposed them.

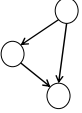
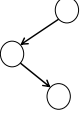
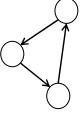
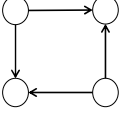
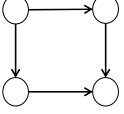
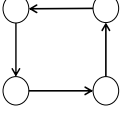
Network motifs [163] are subgraphs that recur much more frequently in the real network than in an ensemble of randomised networks. They are defined at network-level, in order to uncover the basic building blocks of directed networks across domains. Subgraphs having a p -value less than 0.01 are deemed as motifs, where p is the probability of the subgraph appearing more times in randomised networks than in the real network. The statistical significance of a motif can also be captured by Z-score, which is calculated as :

$$Z_i = (N_i^{\text{real}} - \bar{N}_i^{\text{rand}}) / \text{std}(N_i^{\text{rand}}),$$

where N_i^{real} is the number of subgraphs of type i in the real network, and N_i^{rand} is the number of subgraphs of type i in a randomised network. A natural downside of this approach, however, is that it needs to generate a large number of random networks (e.g. 100 or 1000) using certain configuration model. The original work only focuses on 3-node and 4-node directed subgraphs, finding that particular subgraphs such as 3 node feed-forward loop, 3-node feedback loop, bi-fan, bi-parallel, and 4-node feedback loop are significant building blocks in several different types of directed networks (Table 2.1).

Graphlets [161], are nonisomorphic induced subgraphs around a focal node. In the original work, it is defined for undirected networks. A key difference between

Table 2.1 : Some 3-node and 4-node motifs in directed networks[163]. Motifs containing bidirectional edges are not included.

Motif	Designation	Type of network
	3-node feed-forward loop	Gene regulation network Neural network Electronic circuits (forward logic chips)
	3-chain	Food webs
	3-node feedback loop	Gene regulation network Neural network Electronic circuits (forward logic chips)
	Bi-fan	Gene regulation network Neural network Electronic circuits (forward logic chips) Electronic circuits II
	Bi-parallel	Neural network Food webs Electronic circuits (forward logic chips)
	4-node feedback loop	Electronic circuits II

motifs and graphlets is that graphlets are defined at node-level. The term automorphism orbits, or orbits for short, are used to distinguish different positions of the focal node in a subgraph. Therefore, when subgraph size is limited to a range of 2 to 5 nodes, there are 73 different orbits on 30 different subgraphs. We recap graphlets with their orbits in Figure 2.2 (in order to save some space, the majority of 5-node graphlets are omitted). Different node colors within each subgraph are used to distinguish different node orbits.

To summarise, motifs and graphlets are both small induced subgraphs, but they are different in the following aspects (Figure 2.3): motifs are defined at network-level

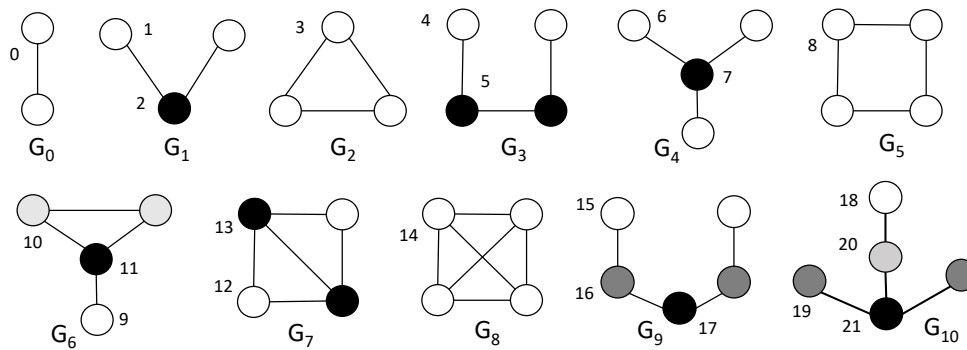


Figure 2.2 : Graphlets and their orbits. Different node color indicates nonisomorphic node position within a given graphlet. [161]

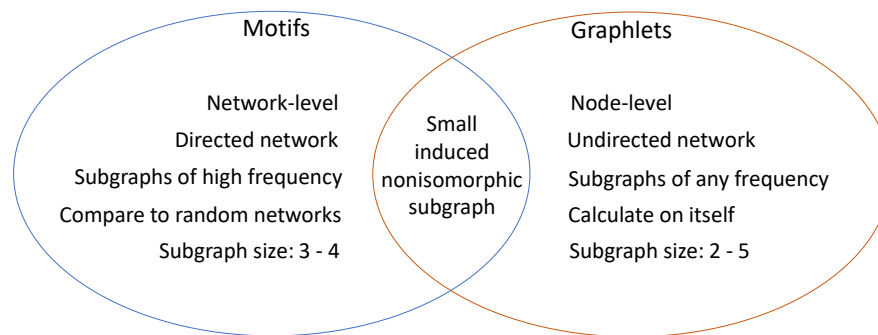


Figure 2.3 : Motifs vs. Graphlets

while graphlets are defined at node-level; motifs are proposed for directed networks while graphlets are for undirected networks; motifs are discovered from comparing real networks to randomised networks with the same degree sequence while graphlets are calculated on the network itself; lastly, motifs contain 3 - 4 nodes while graphlets have 2 - 5 nodes.

2.3 Graph structural measures

After introducing terms and notions, we propose the taxonomy of graph structural measures. We divide existing structural measures into five categories (Figure 2.1):

- *Subgraph Count Based Approaches.* These measures are defined based on the

number of particular subgraph or subgraphs.

- *Subgraph Formation Based Approaches.* In this category, the measures are defined by the ratio of the numbers of two subgraphs: one contains less edges (or nodes) and is viewed as the formation base of another.
- *Global Path Based Approaches.* As the name implies, these measures are based on unbounded paths. It involves the calculation of shortest paths or all paths originating from a node to any node in the entire graph.
- *Message Passing Based Approaches.* Unlike previous categories, message passing-based approaches utilise graph structural information in an implicit manner: every node is initialised with an importance score. Then iteratively, each node updates its score through aggregating the scores of its neighbours. Graph Neural Network approaches can be viewed as transforming this traditional message passing approach into a learnable process.
- *Hybrid Approaches.* These measures are simply some combinations of the previous four categories.

Evidently, the first two categories, i.e., subgraph count based approaches and subgraph formation based approaches are local structural measures. Global path based approaches, on the other hand, are global structural measures. Message passing based approaches are very different because they operate iteratively on all nodes of the graph. However, at each iteration and at each node, it only gathers information such as an influence or importance score, from its immediate neighbours. In this sense, they can be viewed as local structural measures. Below are detailed discussions about each category.

2.3.1 Subgraph Count Based Approaches

Subgraph count based measures are based on the number of particular subgraph or subgraphs. We further divide them into three subclasses, i.e., measures defined on 1-hop neighbours, measures defined on k-hop neighbours/local paths, and measures defined on multi-subgraphs. Figure 2.4 gives the detailed categorisation. Color of the block differentiates where the approach is defined on: grey is on node-level, blue is on link-level, and orange is on network-level.

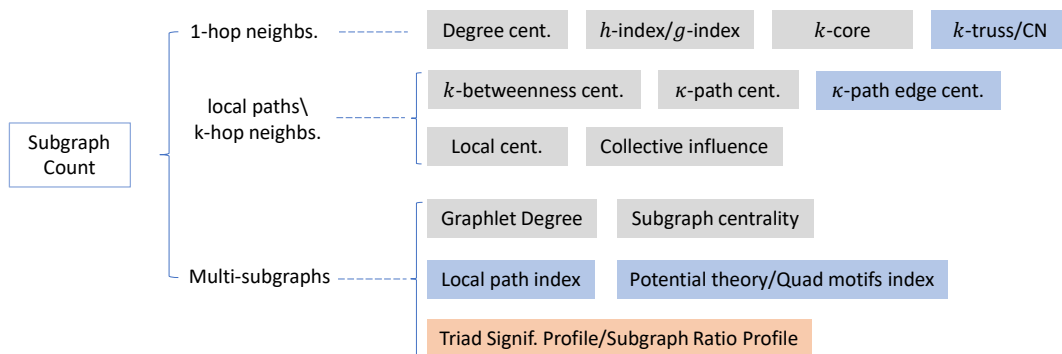


Figure 2.4 : Subgraph count based measures.

2.3.1.1 1-hop neighbours

As the name implies, the calculation of this category only requires the immediate neighbourhood around a node or a link.

- **Degree centrality.** Through calculating the number of nodes directly connected to a node, degree centrality is an easy and straightforward way to assess the importance or influence of the node[70]. In order to render it within the range of (0,1], it is often normalised by the size of the network minus one. Mathematically, the normalised degree centrality of node i is defined as:

$$\Theta_D(i) = \frac{d_i}{n-1}. \quad (2.1)$$

Despite being so simple, degree centrality has been widely applied in various domains. For example, in customer networks, degree centrality is used to find opinion leaders [198], and in biomedical semantic networks, it is effective in selecting crucial information for summarizing disease treatment [246]. Some interesting extensions of the degree centrality include the in-degree/out-degree centrality in directed networks, the strength centrality and weighted strength centrality in weighted networks [34] and the cross-layer degree centrality in multi-layered networks [31].

- ***h*-index/*g*-index.** *h*-index is proposed to evaluate the impact of an individual’s research output: A researcher has index *h* if *h* of his or her papers have at least *h* citations [91]. It is then used as a centrality measure in networks, and named as lobby index or *l*-index [119]. The *l*-index of a node is the largest integer *k* such that the node has at least *k* neighbours with a degree of at least *k*. Egghe argued that the influence of highly cited papers are underplayed in *h*-index, and proposed *g*-index to overcome this disadvantage [57]. After ranking a researcher’s papers according to their citations, *g*-index is defined as the highest rank *g* such that the top *g* papers together have at least *g*² citations. From its definition, *g*-index is always greater than or equal to *h*-index. To address the same issue, *e*-index is proposed to complement the *h*-index for excess citations[245]. Recently, local *h*-index centrality is proposed to identify influential spreaders by simultaneously considers the *h*-index values of the node and its neighbours [146]: $\Theta_{LH}(i) = h(i) + \sum_{j \in N_i} h(j)$.

- ***k*-core** [116]. Instead of only calculating the number of 1-hop neighbours at one node (as in degree centrality) or at both the node and its neighbours (as in *h*-

index), k -core or coreness takes into account the number of neighbours at every node. Specifically, a k -core is defined as a subgraph in which all nodes of degree smaller than k have been removed and the remaining nodes have a degree of at least k . A node located in a higher k -core is deemed as more important than a node in a lower k -core. k -core is calculated through the k -shell decomposition [35] which incrementally (from 1 to k) removes nodes with degree less than k (which in turn results in lowering the degree of remaining nodes) until no more nodes need to be removed. Given that the degree centrality, h -index and coreness are all based on the number of 1-hop neighbours, Lü et al. further revealed their relationships through proposing the high-order h -indices [151]. Bae et al. further propose a neighbourhood coreness that considers both the degree of a node and the coreness of its neighbours [14]:

$$\Theta_{NC}(i) = \sum_{j \in N(i)} ks(j). \quad (2.2)$$

The assumption is that a node having more connections to the neighbours located in the core of the network is more influential.

- **k -truss/Common neighbours.** k -truss is a subgraph where every edge is contained in at least $k - 2$ triangles[40, 227]. It is found through counting the number of common neighbours of a pair of nodes that forms an edge, i.e., the number of triangles that edge participates. A k -truss is also a $(k + 1)$ -core. Counting common neighbours around a pair of nodes that has not formed an edge (a non-edge) is also a basic approach in link prediction [144]. There is a big family of similar approaches that is based on the number of neighbours around non-edges, such as the Adamic-Adar index, the resource allocation index, the preferential attachment index, among the others [158]. Notice that k -truss and Common Neighbours-like

approaches are all on link-level. The block color is therefore blue in Figure 2.4.

2.3.1.2 local paths/ k -hop neighbours

The group of methods in this category requires the calculation of local paths or k -hop neighbours.

- **k -betweenness centrality** [26]. k -betweenness centrality or bounded-distance betweenness centrality is a variation of the well-known betweenness centrality that limits the length of shortest paths to a predefined value k . Specifically, the k -betweenness centrality of any node i is calculated by:

$$\Theta_{B_k}(i) = \sum_{s,t \in V} \frac{\sigma_k(s, t | i)}{\sigma_k(s, t)}, \quad (2.3)$$

where $\sigma_k(s, t)$ is the number of shortest paths of length at most k between node pair s and t , and $\sigma_k(s, t | i)$ is the number of those paths that pass through node i . The reason of proposing a bounded-distance betweenness centrality is that in some networks, long paths are rarely used for propagation of influence.

- **κ -path centrality** [8]. Instead of limiting the length of shortest paths between node pairs, κ -path centrality assumes that message traversals are along random simple paths of length at most k , and proposes to calculate the sum of the probability that a message originating from any possible node goes through the focal node. The κ -path centrality of node i is defined as:

$$\Theta_{P_k}(i) = \sum_{s \in V} \frac{\sigma_k(s | i)}{\sigma_k(s)}, \quad (2.4)$$

where s are all the possible source nodes, $\sigma_k(s | i)$ is the number of k -paths originating from s and passing through i , and $\sigma_k(s)$ is the overall number of k -

paths originating from s . In order to calculate it more efficiently in large networks, a randomised approximation algorithm called RA- κ path is also proposed. [8]

- **κ -path edge centrality** [47]. Moving κ -path centrality definition to link-level, we then have the κ -path edge centrality. The k -path edge centrality of any given edge e is defined as the sum of the frequency with which a message originated from any possible node traverses e , assuming that the message traversals are along random simple paths of length at most k :

$$\Theta_{P_k}(e) = \sum_{s \in V} \frac{\sigma_k(s | e)}{\sigma_k(s)}. \quad (2.5)$$

Quite similar to Equation 2.5, only here $\sigma_k(s | e)$ is the number of κ -paths originating from s that go over the edge e . The original κ -path edge centrality is very expensive to compute in large networks with a big k , therefore two randomised approximations have been further proposed, i.e., ERW- κ path and WERW- κ path [47].

- **Local centrality** [38]. Local centrality, sometimes summarised as LocalRank [148] utilises the information within a node’s 4-hop neighbourhood. Concretely, the local centrality of node i is defined as:

$$\Theta_{LR}(i) = \sum_{j \in N(i)} Q(j), \quad Q(j) = \sum_{k \in N(j)} R(k), \quad (2.6)$$

where $N(i)$ and $N(j)$ are the set of 1-hop neighbours of node i and j , and $R(k)$ is the number of both 1-hop and 2-hop neighbours of node k . It is said to perform better than betweenness centrality and almost as good as closeness centrality

to identify influential nodes under the setting of SIR model, with only a time complexity of $O(n\langle k \rangle^2)$.

- **Collective influence** [167]. Collective influence (CI) is another interesting method that takes higher-order neighbourhood into consideration. The idea is to find those nodes that will cause biggest drop in the “energy function” if being removed. Specifically, level k collective influence of node i is defined as:

$$\Theta_{CI_k}(i) = (d_i - 1) \sum_{j \in N_k(i)} (d_j - 1), \quad (2.7)$$

where $N_k(i)$ is k -hop neighbours of node i . After applying collective influence score, the paper finds that a large number of previously neglected weakly connected nodes (nodes of lower degree) emerges among the optimal influencers [167].

2.3.1.3 *Multi-subgraphs*

Methods of this category involves the count of multiple different subgraphs. They can be at node-level, link-level or network-level.

- **Graphlet degree** [161]. As discussed in Section 2.2.2, graphlets are nonisomorphic induced subgraphs around a node. Graphlet degree is a 73-dimensional vector formed by all different orbits in the subgraphs of size 2-5 nodes. The paper discovers that in protein-protein interaction(PPI) networks, nodes grouped together under this measure belong to the same protein complexes, perform the same biological functions and have the same tissue expressions. Some interesting extensions of graphlets include the dynamic graphlets for temporal networks[98], the directed graphlets for directed networks[12], the colored graphlets for heterogeneous networks[81], and the typed-edge graphlets for edge-labelled networks [102].

- **Subgraph centrality** [60]. Subgraph centrality focuses on subgraphs captured by closed walks of different length around a given node. For example, when the walk length is 4, three types of subgraphs are covered, which are 2-cliques, 2-paths, and 4-cycles. The number of closed walks of length k around node i can be calculated from the i^{th} diagonal entry of the k^{th} power of the adjacency matrix. When the walk becomes unbounded, the subgraph centrality of node i is calculated by:

$$\Theta_S(i) = \sum_{k=0}^{\infty} \frac{\mu_k(i)}{k!}, \quad (2.8)$$

where $\mu_k(i) = (\mathbf{A}^k)_{ii}$. It is shown to be more discriminative than many popular centrality measures such as the degree centrality, the betweenness centrality and the eigenvector centrality.

- **Local path index** [149]. Extended from common neighbours, local path index counts both the number of 2-paths and 3-paths between a pair of nodes. The approach is proposed for link prediction, and therefore focuses on non-connected node pairs. Concretely, the local path index of a node pair i and j is defined as:

$$\Theta_{LP}(i, j) = A_{ij}^2 + \epsilon A_{ij}^3, \quad (2.9)$$

where ϵ is a weigh parameter for 3-paths. The paper finds out that local path index remarkably outperforms common neighbours and can reach a competitive accuracy as Katz index where all paths are considered. Some other works compare 3-paths approaches against 2-paths approaches in link prediction and find out that 3-path approaches perform better in PPI networks and food webs [168, 120, 253].

- **Potential theory/Quad motifs index.** Potential theory aims to predict links in directed networks. By counting the numbers of 4 different directed 2-paths and 8 different directed 3-paths around a pair of nodes, the paper finds out that a link has higher probability to appear if it could generate more bi-fan subgraphs [249]. Very similar to the idea of potential theory, quad motifs index is proposed to count particularly three types of directed open-quadriad (3-paths) subgraphs: two of them are the bases for bi-parallel subgraphs and the other one is for bi-fan [97]. Specifically, the quad motifs index of a pair of nodes i and j is defined as:

$$\Theta_{QM}(i, j) = \alpha \times s_F(i, j) + \frac{(1 - \alpha)}{2} (s_{P1}(i, j) + s_{P2}(i, j)), \quad (2.10)$$

where $s_F(i, j)$ is the contribution from the bi-fan base while $s_{P1}(i, j)$ and $s_{P2}(i, j)$ are the contributions from two bi-parallel bases. Together with the local path index, it is interesting to see that 3-path subgraphs are of particular importance in link prediction.

- **Triad significance profile/Subgraph ratio profile** [162]. Extended from networks motifs [163], triad significance profile (TSP) is constructed from normalised Z scores of 13 different directed 3-node subgraphs.

$$TSP = \{SP_1, SP_2, \dots, SP_{13}\}, \quad SP_i = Z_i / (\Sigma Z_i^2)^{1/2}. \quad (2.11)$$

Z_i is in turn calculated from comparing with an ensemble of randomised networks with the same degree sequence, i.e., $Z_i = (N_i^{\text{real}} - \bar{N}_i^{\text{rand}}) / \text{std}(N_i^{\text{rand}})$. Subgraph ratio profile (SRP), on the other hand, is built from 6 undirected 4-node subgraphs

(G_3 to G_8 in Figure 2.2) :

$$SRP = \{SRP_1, SRP_2, \dots, SRP_6\}, \quad SRP_i = \Delta_i / (\Sigma \Delta_i^2)^{1/2}. \quad (2.12)$$

Unlike TSP, SRP uses abundance of each subgraph relative to random networks, i.e., $\Delta_i = \frac{N_{\text{real}_i} - \langle N_{\text{rand}_i} \rangle}{N_{\text{real}_i} + \langle N_{\text{rand}_i} \rangle + \varepsilon}$. Previously seemingly unrelated networks are found to belong to several superfamilies with very similar significance profile. Notice also that these two approaches are defined on network-level, not on node or link-level as we have seen often.

2.3.2 Subgraph Formation Based Approaches

Subgraph formation based measures view a subgraph being built from another less complex subgraph, i.e., with one link, multiple links, or one node less. We further divide them into three categories according to size of the subgraph, 3-node, 4-node and 4-node plus (Figure 2.5). Most of these approaches are defined at node-level, except that the edge clustering coefficient is at link-level and the interest clustering coefficient is at network-level.

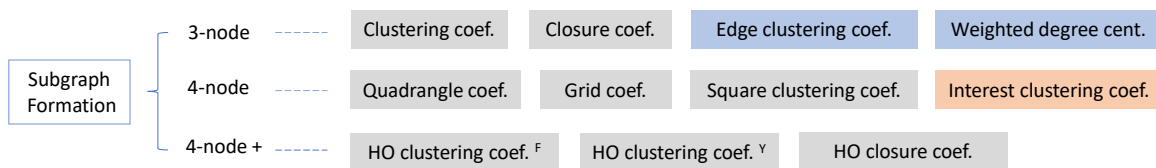


Figure 2.5 : Subgraph formation based measures.

2.3.2.1 3-node subgraph

3-node subgraph formation is the simplest situation since there is only one connected 2-node subgraph, with a link connecting 2 nodes.

- **Clustering coefficient** [232]. Clustering coefficient is the first and most influential measure in this category. It measures the extent to which the neighbours of a

node connect to each other. From a structural formation perspective, it measures the formation of triangles upon open-triads (also called wedges). Specifically, the clustering coefficient of node i is defined as the ratio between the number of triangles containing node i (denoted $T(i)$) and the number of open triads (denoted $OT(i)$):

$$\mathcal{C}_C(i) = \frac{T(i)}{OT(i)} = \frac{\frac{1}{2} \sum_{j \in N(i)} |N(i) \cap N(j)|}{\frac{1}{2} d_i (d_i - 1)}. \quad (2.13)$$

Due to its significance and simplicity in definition, the clustering coefficient has been widely applied in studying complex networks [186, 32, 204] and extended to directed networks [63, 5], weighted networks [17, 177, 244] and signed networks [129, 42].

- **Closure coefficient** [241]. Closure coefficient measures the formation of triangles from a new perspective, i.e., with the focal node located at the end of an open-triad. Different from the ordinary centre node perspective in clustering coefficient (orbit 2 of G_1 in Figure 2.2, denoted as $G_1^{(2)}$), the focal node in closure coefficient serves as the end node of an open triad (orbit type $G_1^{(1)}$). The closure coefficient of node i is calculated as the fraction of open triads ($OT_E(i)$), where i serves as the end node, that actually form triangles:

$$\mathcal{C}_E(i) = \frac{2 \cdot T(i)}{OT_E(i)} = \frac{\sum_{j \in N(i)} |N(i) \cap N(j)|}{\sum_{j \in N(i)} (d_j - 1)}. \quad (2.14)$$

Despite this subtle difference in definition, the closure coefficient has very different properties compared to the clustering coefficient. It has been extended to directed networks [243, 103] and weighted networks [104].

- **Edge clustering coefficient** [228]. Defined on link-level, edge clustering coeffi-

cient (ECC) evaluates to what extent nodes cluster around the focal edge. From a structure formation view, it measures the formation of triangle upon this link. The edge clustering coefficient of an edge e_{ij} is defined as:

$$\mathcal{C}_e(i, j) = \frac{T(i, j)}{\min(d_i - 1, d_j - 1)}, \quad (2.15)$$

where $T(i, j)$ is the number of triangles that e_{ij} participates, and $\min(d_i - 1, d_j - 1)$ is the number of triangles that edge could possibly form. Based on ECC, a node centrality measure is then defined as the sum of the edge clustering coefficients of all edges connected to it, i.e., $\mathcal{C}_N(i) = \sum_{j \in N_i} \mathcal{C}_e(i, j)$. This measure has been proven to be more efficient for identifying essential proteins than many other centrality measures.

- **Weighted degree centrality** [217]. Weighted degree centrality (WDC) is also proposed to identify essential proteins. Although this name seems related with degree centrality, it is in fact an extension of edge clustering coefficient. This approach is different in that it takes into account not only the PPI graph data but also the gene expression data. Specifically, a weight of an interaction is calculated as:

$$\mathcal{C}_w(i, j) = \mathcal{C}_e(i, j) + r(i', j'), \quad (2.16)$$

where $\mathcal{C}_e(i, j)$ is the edge clustering coefficient from the graph data, and $r(i', j')$ is the Pearson correlation coefficient calculated from the gene expression data. Similarly, the weighted degree centrality of a node is then defined as: $\Theta_W(i) = \sum_{j \in N_i} \mathcal{C}_w(i, j)$. This approach essentially integrates node features when analysing networks.

2.3.2.2 4-node subgraph

4-node subgraphs are much more complicated than 3-node subgraphs. There are in total 6 different subgraphs and 11 different orbits in 4-node subgraphs (Figure 2.2).

- **Quadrangle coefficients** [106]. Many real networks (such as PPI networks, neural networks and food webs) are naturally rich in quadrangles. Quadrangle coefficients, or i-quad coefficient and o-quad coefficient, are thus proposed to measure the formation of quadrangles upon open-quadrads (3-paths). As there are two orbits in an open-quadradiad ($G_3^{(5)}$ and $G_3^{(4)}$), i-quad coefficient has the focal node at $G_3^{(5)}$ while o-quad coefficient has the focal node at $G_3^{(4)}$. Specifically, the quadrangle coefficients of node i are defined as:

$$\mathcal{C}_I(i) = \frac{2Q(i)}{OQI(i)}, \quad \mathcal{C}_O(i) = \frac{2Q(i)}{OOO(i)}, \quad (2.17)$$

where $Q(i)$ is the number of quadrangles; $OQI(i)$ and $OOO(i)$ are number of open-quadrads with i as the inner node and outer node respectively. They are found to be more efficient than 3-node measures in classifying networks and predicting links.

- **Grid coefficients** [33]. Grid coefficients, including the primary grid coefficient and the secondary grid coefficient, also aim to measure the formation of 4-cycles. The primary grid coefficient measures the formation of “primary quadrilaterals” upon a node and three of its 1-hop neighbours, which is essentially the formation of chordal cycles (G_7) from tailed-triangles (orbit $G_6^{(11)}$). Concretely, the primary grid coefficient of node i is defined as:

$$\mathcal{C}_{G_p}(i) = \frac{Q_p(i)}{d_i(d_i - 1)(d_i - 2)/2}, \quad (2.18)$$

where $G_p(i)$ is the number of chordal-cycles containing i and the denominator being the number of possible chordal-cycles built from a node and its three neighbours. The secondary coefficient measures the formation of “secondary quadrilaterals” from a node, two of its 1-hop neighbours and one of its 2-hop neighbours:

$$\mathcal{C}_{G_s}(i) = \frac{Q_s(i)}{d_{i,2nd}d_i(d_i - 1)/2}. \quad (2.19)$$

Notice, however, in this definition the 2-hop neighbour could be at orbit $G_3^{(4)}$ or at orbit $G_{10}^{(20)}$. The latter essentially involves 5 nodes in total.

- **Square clustering coef.** As triangles (3-cycles) are absent in bipartite networks, square clustering coefficient is proposed to measure the formation of 4-cycles in the context of bipartite networks [145]. What is unusual about this approach is that it views 4-cycles being built from node overlapping instead of node connection. Specifically, the square coefficient of node i , with a pair of its neighbours m and n , is calculated as:

$$\mathcal{C}_S(i|m, n) = \frac{Q_{imn}}{(d_m - \eta_{imn})(d_n - \eta_{imn}) + Q_{imn}}, \quad (2.20)$$

where Q_{imn} is the number of 4-cycles containing nodes i, m, n ; and $\eta_{imn} = 1 + Q_{imn}$ if m and n are not connected (or $\eta_{imn} = 2 + Q_{imn}$ if m and n are connected). Zhang et al. [248] later proposed a modified version of square clustering coefficient: $\mathcal{C}_{S_z}(i|m, n) = \frac{Q_{imn}}{(d_m - \eta_{imn}) + (d_n - \eta_{imn}) + Q_{imn}}$. With this minor change at the denominator, 4-cycles are now built from connecting nodes. It is mainly applied in community detection.

- **Interest clustering coefficient** [222]. Interest clustering coefficient is intro-

duced to measure the “clustering of interest links” in directed social networks. It argues that the best way of defining a relationship between two individuals is through common interests, i.e., two individuals having links towards a common neighbour will have higher chance to follow other common neighbours. From a structural view, interest clustering coefficient essentially measures the formation of bi-fan subgraphs (Table 2.1) upon open bi-fans:

$$\mathcal{C}_I = \frac{4 \cdot \# \text{ bifan}}{\# \text{ open-bifan}}. \quad (2.21)$$

Note that this metric is defined at network-level. The paper finds out that the interest clustering coefficient of Twitter is higher than the traditional directed clustering coefficient, and further proves its usage in link recommendation.

2.3.2.3 *Beyond 4-node subgraph*

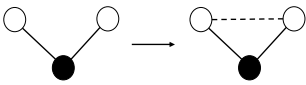
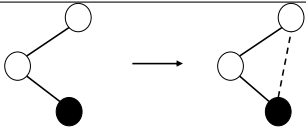
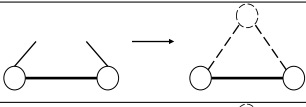
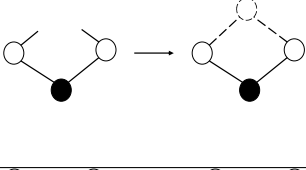
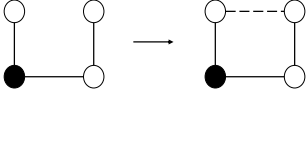
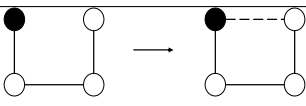
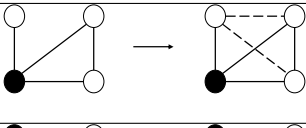
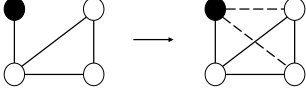
Some approaches are introduced with a variable subgraph size. In actual usage, however, due to high complexity, they seldom surpass the size of 6 nodes.

- **Higher-order clustering coefficients^F** [72]. Fronczak et al. propose the higher clustering coefficients to evaluate the probabilities that the shortest paths between any two neighbours of node i equal k , when all paths passing through node i are neglected. Particularly, a clustering coefficient of order k for node i is calculated as:

$$\mathcal{C}_{H_F}(i | k) = \frac{2E(i | k)}{d_i(d_i - 1)}, \quad (2.22)$$

where $E(i | k)$ denotes the number of shortest paths of length k between i 's neighbours. When k equals 1, it degrades to the standard clustering coefficient, and when k equals 2, it measures the formation of 4-cycles. Note that each pair of neighbours could have multiple shortest paths of the same length, and only one

Table 2.2 : Metrics for 3-node and 4-node subgraph formation.

3-node/4-node subgraph formation	Undirected networks	Directed networks	Weighted networks
	clustering coef.[232]	directed clustering coef.[63, 5]	wgtd. clustering coef. [17, 177, 244] wgtd. signed clustering coef. [129, 42] wgtd. directed clustering coef. [63]
	closure coef.[241]	directed closure coef. [243, 103]	weighted closure coef. [104]
	edge clustering coef.[228]		
	higher-order clustering coef. (Fronczak)[72] higher-order clustering coef. (Abdo)[1]	No	No
	square clustering coef. (Lind [145], Zhang [248]) i-quad coef. [106] primary grid coef. [33]	No	No
	o-quad coef. [106]	—	weighted o-quad coef. [106]
	higher-order clustering coef. (Yin)[240]	No	No
	higher-order closure coef. (Yin)[241]	No	No

of them should be counted so that the value of higher-order clustering coefficients is bounded by 1.

- **Higher-order clustering coefficient^Y** [240]. The higher-order clustering coefficient proposed by Yin et al. is another generalisation of the traditional clustering coefficient. It aims to measure the formation of higher-order cliques. Specifically, a k^{th} -order clustering coefficient of node i is defined as the probability that a k -clique plus an edge incident to i (termed as k -wedge) forms a $(k + 1)$ -clique:

$$\mathcal{C}_{HY}(i | k) = \frac{k \cdot |C_{k+1}(i)|}{|W_k(i)|} = \frac{k \cdot |C_{k+1}(i)|}{(d_i - k + 1)|C_k(i)|}, \quad (2.23)$$

where $C_{k+1}(i)$ is the set of $(k + 1)$ -cliques containing node i , and $W_k(i)$ is the set of k -wedges with i as the centre node. The properties of higher-order clustering coefficient in random graph and small-world model have also been thoroughly investigated [240].

- **Higher-order closure coefficient** [241]. Higher-order closure coefficient measures the formation of higher-order cliques from a different perspective, i.e., the focal node being the end-node of a k -wedge (instead of the centre-node). The k^{th} -order closure coefficient of node i is thus defined as the fraction of end-node based k -wedges that are closed (a closed k -wedge is a $(k + 1)$ -clique):

$$\mathcal{C}_{HE}(i | k) = \frac{k \cdot |C_{k+1}(i)|}{|W'_k(i)|} = \frac{k \cdot |C_{k+1}(i)|}{\sum_{j \in N(i)} [C_k(j) - (k - 1)C_k(i)]}, \quad (2.24)$$

where $C_{k+1}(i)$ is the set of $(k + 1)$ -cliques containing node i , and $W'_k(i)$ is the set of k -wedges with i as the end-node. Higher-order closure coefficient is proven to be useful in finding seeds for personalised PageRank community detection.

An illustrative summary for most abovementioned approaches is given in Table 2.2.

2.3.3 Global Path Based Approaches

Global path based approaches require structural information across the whole network, in the form of unbounded paths between nodes. One set of methods are based on the paths from one node to all other nodes, such as the well known closeness centrality and Katz index; another set of methods are based on paths between all node pairs, represented by the betweenness centrality (Figure 2.6).

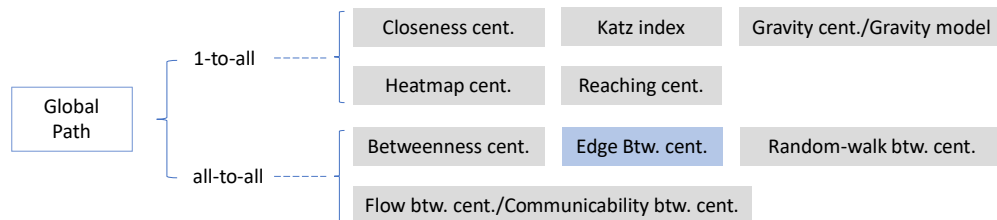


Figure 2.6 : Global path based measures.

2.3.3.1 One-to-all

The approaches of this type involve the paths from one node to all other nodes. They are also referred as radial measures.

- **Closeness centrality** [70]. Being one of the most classic centrality measures, closeness centrality is defined as the reciprocal of the average shortest path distance from a focal node i to all other nodes:

$$\Theta_C(i) = \frac{|V| - 1}{\sum_{j \in V, j \neq i} d(i, j)}. \quad (2.25)$$

Obviously, the original definition is not suitable for graphs with more than one connected components. To address this problem, a modified version of closeness

centrality is defined as [231]:

$$\Theta_{C'}(i) = \frac{n-1}{|V|-1} \frac{n-1}{\sum_{j=1}^{n-1} d(i,j)}, \quad (2.26)$$

where n is the number of nodes in one connected component. Due to its intuitiveness in definition, closeness centrality keeps being applied and extended in various fields. Some recent works include the neighbourhood closeness centrality in predicting essential proteins [139], and the backward/forward closeness in studying global value chain [82].

- **Katz index** [112]. Unlike closeness centrality that focuses on shortest paths, Katz centrality of a node considers all paths reaching other nodes, having longer paths contributing less. Concretely, the Katz centrality of node i is calculated as:

$$\Theta_K(i) = \sum_j \sum_{k=1}^{\infty} \beta^k \mathbf{A}_{ij}^k, \quad (2.27)$$

where k is path length and β is an attenuation parameter in range $(0, \frac{1}{\lambda})$, λ being the largest eigenvalue of \mathbf{A} . Further, the overall matrix $\mathbf{M} = \sum_{k=1}^{\infty} (\beta \cdot \mathbf{A})^k$ is a weighted ensemble of all paths. Thus, \mathbf{M}_{ij} represents the weighted sum of paths from i to j in all possible hops. Note that this definition is naturally suitable in directed networks and a recent work proposes to generate node embedding of directed graph by performing singular value decomposition on Katz index matrix [181].

- **Gravity model** [142] / **Gravity centrality** [152]. Inspired by Newton's gravity law formula, a gravity model is proposed by viewing the degree of a node as its

mass and the shortest path length between two nodes as their distance:

$$\Theta_G(i) = \sum_{j \in V, j \neq i} \frac{d_i \cdot d_j}{d(i, j)^2}. \quad (2.28)$$

In order to make it easier to compute in large networks, a modified version limits the radius from the entire network to a certain length. Adopting a similar idea, the gravity centrality is introduced by regarding the coreness of a node as its mass, and the shortest path length between nodes as their distance:

$$\Theta'_G(i) = \sum_{j \in N_k(i)} \frac{ks(i) \cdot ks(j)}{d(i, j)^2}, \quad (2.29)$$

where $N_k(i)$ is the neighbourhood of node i within k -hops, and $ks(i)$ is the coreness of node i . The two approaches are shown to be effective in identifying influential spreaders through analyses of the SIR model on real networks.

- **Heatmap centrality** [55]. Heatmap centrality measures the influence of a node by comparing the farness of the node with the average farness of its neighbours. Farness, the reciprocal of closeness, is defined as the sum of the lengths of shortest paths from a node to all other nodes, i.e., $f(i) = \sum_{j \in V, j \neq i} d(i, j)$. Therefore, the heatmap centrality of node i is quantified as:

$$\Theta_{HM}(i) = f(i) - \frac{\sum_{j \in N(i)} f(j)}{|N(i)|}. \quad (2.30)$$

The intuition of this metric is that if a node has smaller farness than its neighbours, the probability of information passing through it is higher. Note that according to heatmap centrality, a top-ranked node of influence should have the most negative value. Although the definition of heatmap centrality is more related to closeness

centrality, it is revealed that it is highly correlated with betweenness centrality.

- **Reaching centrality** [165]. Reaching centrality aims to rank the influence of a node in directed networks. Intuitively, the reaching centrality of node i is quantified as the proportion of nodes that can be reached by the node via outgoing edges, i.e., the number of nodes with a directed distance from i , divided by $|V| - 1$. Further, a global reaching centrality is then defined as:

$$GRC = \frac{\sum_{i \in V} [\Theta_R^{max} - \Theta_R(i)]}{|V| - 1}, \quad (2.31)$$

where Θ_R^{max} is the largest reaching centrality of all nodes. The meaning of GRC is the difference between the maximum reaching centrality and the average reaching centrality. Global reaching centrality is used as a hierarchy measure for directed networks, and is shown to be capable of capturing the degree of hierarchy in both synthetic and real networks.

2.3.3.2 All-to-all

The approaches here involve the count of paths between all node pairs, and among them the ones that passing through a focal node or edge. They are also referred to as medial measures.

- **Betweenness centrality** [69]. Betweenness centrality, or more precisely, shortest-path betweenness centrality is one of the best-known centrality measures. The betweenness centrality of node i is quantified as the sum of the fraction of all-pairs shortest paths going through i :

$$\Theta_B(i) = \sum_{s, t \in V} \frac{\sigma(s, t | i)}{\sigma(s, t)}, \quad (2.32)$$

where $\sigma(s, t | i)$ is the number of shortest paths between node pair s and t that pass through node i , and $\sigma(s, t)$ is the number of all shortest paths between s and t . It is often normalised by $\frac{(|V|-1)(|V|-2)}{2}$, in order to be compared in different networks. Betweenness centrality has also been generalised to directed networks [233] and weighted networks [178].

- **Edge betweenness centrality** [75]. With a small modification on the original betweenness centrality, Girvan and Newman propose edge betweenness centrality, in order to detect community structure in complex networks. The edge betweenness centrality of an edge e is quantified as the sum of the fraction of all-pairs shortest paths passing through e :

$$\Theta_B(e) = \sum_{s,t \in V} \frac{\sigma(s, t | e)}{\sigma(s, t)}, \quad (2.33)$$

According to the definition, edges between communities will have large edge betweenness. Therefore, the underlying communities of the network would be uncovered by removing edges of high edge betweenness centrality. It has been widely applied in community detection, and some recent applications include the study of anti-vaccination sentiment on Facebook [93] and the analysis of microbial diversity in marine sediment [96].

- **Flow betweenness centrality** [71]/ **Communicability betweenness centrality** [59]. A major limitation of betweenness centrality is that it exclusively focuses on shortest paths. In real situations, however, information often takes a more circuitous path randomly or intentionally [214]. Flow betweenness addresses this issue by considering all paths between nodes. Specifically, the flow

betweenness centrality of node i is defined as:

$$\Theta_F(i) = \sum_{s,t \in V} \frac{\phi(s, t | i)}{\phi(s, t)}, \quad (2.34)$$

where $\phi(s, t | i)$ is the maximum flow between s and t that passes through i , and $\phi(s, t)$ is the total flow between s and t . The maximum flow is in turn calculated by the minimum cut capacity [66]. Having the notion of “capacity ” on links, flow betweenness centrality is naturally suitable for weighted networks. Instead of treating each path equally, communicability betweenness centrality proposes to reduce weight for longer paths:

$$\frac{2}{(n-1)(n-2)} \sum_{s,t \in V} \frac{\sum_{k=0}^{\infty} \frac{1}{k!} \mu^k(s, t | i)}{\sum_{k=0}^{\infty} \frac{1}{k!} \mu^k(s, t)}, \quad (2.35)$$

where $\mu^k(s, t | i)$ is the number of paths between s and t passing i with length k , and $\mu^k(s, t)$ is the number of paths between s and t with length k .

- **Random-walk betweenness centrality** [173]. Random-walk betweenness centrality, also known as current-flow betweenness centrality, is another popular variant of betweenness centrality. It models information spreading in a network analogous to electrical current flow in a circuit. Concretely, the current-flow betweenness centrality of node i is defined as the amount of current flowing through i , averaged over all node pairs:

$$\Theta_{CF}(i) = \frac{\sum_{s,t \in V} I(s, t | i)}{(1/2)n(n-1)}, \quad (2.36)$$

where $I(s, t | i)$ is the current flowing from s to t that passes i . The paper then proves that a message spreading along random walks is equivalent to above

definition.

2.3.4 Message Passing Based Approaches

Abovementioned approaches depend solely on the topological information of a network, such as the number of particular subgraphs, the ratio between two subgraphs, the length of shortest paths, or the number of paths. Message passing based approaches further consider the information contained in each node. It is worth to notice that the popular graph convolutional network is also based on this idea, i.e, iteratively gathering information from nearby nodes.

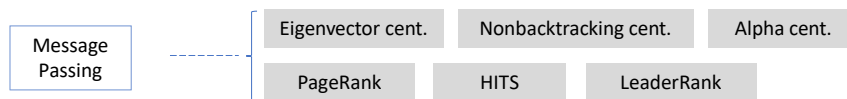


Figure 2.7 : Message passing based approaches.

- **Eigenvector centrality** [24]. Eigenvector centrality is another classic centrality measure. The idea is that a node’s centrality depends on the centralities of its neighbours:

$$x(i) = c \sum_{j \in N(i)} x(j), \quad (2.37)$$

where c is a normalisation constant. The equation is recursive and computed by starting with a set of initial influence scores and iterating the computation until it converges. In vectorisation form, i.e., $\vec{x} = c\mathbf{A}\vec{x}$, \vec{x} is found to converge to the dominant eigenvector of \mathbf{A} and c converges to the reciprocal of the dominant eigenvalue of \mathbf{A} . Eigenvector centrality has some problems in very sparse networks, i.e., the leading eigenvector is localised around nodes of highest degree and diminishes the effectiveness to quantify nodes’ importance [123].

– **Nonbacktracking centrality** [156]. Nonbacktracking centrality is proposed to address the abovementioned localisation issue. Same as in eigenvector centrality, a node's centrality is the sum of its neighbours' centralities, but now the neighbours' centralities are calculated without the influence of this node. This is achieved by using the nonbacktracking matrix [88]. Nonbacktracking matrix \mathbf{B} , is a $2m \times 2m$ matrix, defined on the directed edges of the graph (undirected edges are converted to bidirectional edges), and elements $\mathbf{B}_{i \rightarrow j, k \rightarrow l} = \delta_{il}(1 - \delta_{jk})$, where δ is Kronecker delta. Then, $e_{j \rightarrow i}$ of the leading eigenvector of \mathbf{B} gives the centrality of node j ignoring the contribution of i . Finally, the nonbacktracking centrality of node i is $x(i) = \sum_j \mathbf{A}_{ji} e_{j \rightarrow i}$. Eigenvalues of nonbacktracking matrix is also found to be useful in community detection [124].

– **Alpha centrality** [25]. When eigenvector centrality is applied in directed networks, a node's centrality is determined by those who pointed at it. Thus, the vector form becomes: $\vec{x} = \frac{1}{\lambda} \mathbf{A}^T \vec{x}$. The problem is that nodes with no incoming edges would have zero centrality value. Alpha centrality proposes to solve this problem by taking into account the "external status characteristics". The equation then becomes:

$$\vec{x} = \alpha \mathbf{A}^T \vec{x} + \vec{e}, \quad (2.38)$$

where \vec{e} is a vector of the exogenous sources of characteristics and α is a parameter which reflects the relative importance of topological structure versus exogenous factors.

– **PageRank** [30]. PageRank, a popular variation of eigenvector centrality, is proposed to rank the importance of web pages. Web pages and the links among them

are modelled as a directed network, and a page should have high rank if the sum of the ranks of pages that point to it is high. Specifically, the rank of page i is calculated as:

$$r(i) = c \sum_{j \in N_i^{in}} \frac{r(j)}{d_j^{out}}, \quad (2.39)$$

where N_i^{in} is the set of pages points to i (i 's in-neighbours), and d_j^{out} is out-degree of page j . In order to deal with the “rank sink” problem, where several pages form a loop without other outgoing links, a source of rank is introduced over all pages (also viewed as a random jumping factor), denoted as vector \vec{e} . Therefore, the rank of page i becomes: $r(i) = c(\sum_{j \in N_i^{in}} \frac{r(j)}{d_j^{out}} + e(i))$, and the corresponding vector form is $\vec{r} = c(\mathbf{A}^T + \vec{e} \times \mathbf{1})\vec{r}$. PageRank has also been extended in weighted networks [236], on nonbacktracking matrix [9], and applied to many different areas [76].

- **HITS** [118]. Unlike PageRank that focuses on pages having many incoming links, HITS, abbreviated from hyperlink induced topic search, proposes to distinguish two roles in the hyperlink structure, i.e., authorities and hubs. Authorities are reliable information sources, and hubs are the websites pointing to them. Based on the intuition that an authority should be pointed by hubs and a hub should point to authorities, an authority weight and a hub weight of page i are thus defined in a mutually dependent manner:

$$a(i) = \sum_{j \in N_i^{in}} h(j) \quad h(i) = \sum_{j \in N_i^{out}} a(j). \quad (2.40)$$

The corresponding vector forms are: $\vec{a} = \mathbf{A}^T \vec{h}$, and $\vec{h} = \mathbf{A} \vec{a}$. \vec{a} and \vec{h} are updated iteratively, and it is proven that \vec{a} converges to the leading eigenvector of $\mathbf{A}^T \mathbf{A}$, and \vec{h} converges to the leading eigenvector of $\mathbf{A} \mathbf{A}^T$. Based on HITS,

ARC (Automatic Resource Compilation) later proposes to incorporate textual information around the link by assigning each link a weight [36], and Co-HITS proposes to extend the idea to bipartite networks [48].

- **LeaderRank** [150]. In order to solve the abovementioned rank sink problem, LeaderRank proposes to add a ground node that connects to other nodes via bidirectional links. In the beginning, each node other than the ground node is initialised by one unit of score, and the ground node is initialised by zero. Then, same as PageRank, at each iteration, the score of node i is calculated as: $s(i)^{(t)} = c \sum_{j \in N_i^{in}} \frac{s(j)^{(t-1)}}{d_j^{out}}$. After the scores of all nodes reach steady state, the score of the ground node will be distributed evenly to other nodes, and thus the final score of node i is:

$$s(i) = s(i)^c + \frac{s(g)^c}{|V|}, \quad (2.41)$$

where $s(i)^c$ is the steady score of node i , and $s(g)^c$ is the steady score of the ground node. A major advantage of LeaderRank is that it has no additional parameter that needs to be optimised. Some interesting extensions of LeaderRank include the weighted LeaderRank that assigns degree-dependent weights onto links associated with the ground node [141] and the adaptive LeaderRank that introduces H-index into the weighted mechanism [238].

2.3.5 Hybrid Approaches

The methods in the fifth and final category are combinations of previously introduced approaches.

- **ClusterRank** [37]. Previous studies have shown that large clustering coefficient may slow the spreading process of disease in the entire network [58, 255]. Clus-

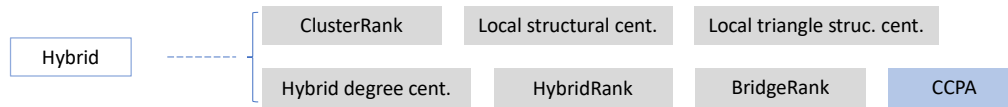


Figure 2.8 : Hybrid Approaches.

terRank thus proposes to consider not only the number of a node’s neighbours, but also the negative effect of local clustering when identifying influential nodes. The ClusterRank score of node i is defined as:

$$\Theta_{CR}(i) = f(c_i) \sum_{j \in N_i^{out}} (d_j^{out} + 1), \quad (2.42)$$

where $c_i = \frac{\sum_{j \in N_i^{out}} |N^{out}(i) \cap N(j)|}{d_i^{out}(d_i^{out}-1)}$ is a modified version of clustering coefficient in directed networks. $f(c_i)$ is a function that negatively correlates with c_i , for example an exponential function $f(c_i) = 10^{-c_i}$. Although ClusterRank is proposed for directed networks, it can be easily extended to undirected networks [37] and weighted networks [230]. Experiments on several real networks demonstrate that ClusterRank score outperforms PageRank and LeaderRank at identifying influential nodes while being more efficient in computation.

- **Local structural Centrality** [74]. Aiming to evaluate the spreading ability of nodes, local structural centrality essentially extends the local centrality (section 2.3.1.2) by further considering the connections between higher-order neighbours. The idea is that a node has better spreading ability when its neighbours are better connected because a neighbour node can be affected directly by the source node or indirectly by another neighbour node. The local structural centrality of

node i is defined as:

$$\Theta_{LS}(i) = \sum_{j \in N_i} (\alpha |N_j^{1,2}| + (1 - \alpha) \sum_{k \in N_j^{1,2}} c(k)), \quad (2.43)$$

where $N_j^{1,2}$ is the node set of 1-hop and 2-hop neighbours of node j , and $c(k)$ is the clustering coefficient of node k . α is a tunable parameter between 0 and 1, balancing direct and indirect spreading contribution. Notice that clustering coefficient is worked as a negative part when evaluating spreading speed as in ClusterRank, but a complementary part when measuring spreading ability here.

- **Local triangle structure centrality** [154]. Local triangle structure centrality (LTSC) proposes to include the triangle proportion of a node, instead of its clustering coefficient, when evaluating a node’s spreading ability. Triangle proportion is able to indicate the location of a node, whether it is located in a denser or sparser part of a network. LTSC partitions the spreading ability into two parts, i.e., inner spreading ability and outer spreading ability. Specifically, the local triangle structural centrality of node i is defined as:

$$\Theta_{TS}(i) = \sum_{j \in N_i} (d_j(1 + TP(j)) + (\sum_{k \in N_j} d_k - d_j)), \quad (2.44)$$

where $TP(j)$ is the triangle proportion of node j , calculated by the number of triangles containing j divided by the total number of triangles in the network. For each neighbour j of a given node i , the part of $d_j(1 + TP(j))$ is to measure its inner spreading ability, and the part of $\sum_{k \in N_j} d_k - d_j$ is to measure its outer spreading ability. Finally, the local triangle structure centrality of node i is the sum of the spreading abilities of its neighbours.

- **Hybrid degree centrality** [153]. The spreading probabilities of networks describing diseases, opinions, and rumours should obviously differ. Most existing centrality measures, however, fail to take that into consideration. The performance of centrality measures is sensitive to the spreading probability. Degree centrality, for example, works best with modest spreading probabilities, while local centrality (section 2.3.1.2) works better with higher ones [74]. In order to alleviate the sensitivity to different spreading probabilities, hybrid degree centrality is introduced by integrating the degree centrality and a modified local centrality. The hybrid degree centrality of node i is defined as:

$$\Theta_{HD}(i) = (\beta - p) \cdot \alpha \cdot \Theta_D(i) + p \cdot \Theta'_{LR}(i), \quad (2.45)$$

where $\Theta'_{LR}(i) = \Theta_{LR}(i) - 2 \sum_{j \in N_i} |N_j|$ is the modified local centrality, p is the spreading probability, α and β are two tunable parameters. The part contributed by degree centrality is viewed as near-source influence, and the part of modified local centrality as distant influence.

- **HybridRank** [3]. HybridRank proposes to identify influential spreaders by combining the neighbourhood coreness centrality (section 2.3.1.1) and eigenvector centrality. The reason of integrating these two measures is that they both regard a node as influential if the node is connected to other influential nodes. The hybrid centrality of node i is defined as:

$$\Theta_{HR}(i) = \Theta_{NC}(i) \times \Theta_E(i), \quad (2.46)$$

where $\Theta_{NC}(i) = \sum_{j \in N_i} ks(j)$ is the neighbourhood coreness of i , and $\Theta_E(i)$ is the

eigenvector centrality of node i . HybridRank algorithm further suggests that when selecting influential spreaders, the neighbours of selected ones should be neglected in order to maximise the spreading range. The effectiveness of HybridRank has also been tested in real networks using SIR model.

- **BridgeRank** [205]. In order to lower the time complexity of closeness centrality while keeping comparable performance, BridgeRank proposes to compute shortest paths to just a few core nodes in the network. In BridgeRank algorithm, at first, communities are identified by Louvain algorithm [22]. Then core nodes are discovered through calculating betweenness centralities within each community (one core node per community). Finally, the BridgeRank centrality of each node is defined as a filtered closeness centrality to these core nodes:

$$\Theta_{BR}(i) = \frac{1}{\sum_{j \in \mathcal{C}} d(i, j)}, \quad (2.47)$$

where \mathcal{C} is the set of identified core nodes in each community. The time complexity is therefore reduced from $O(|V|^3)$ to $O(|V| \log |V|)$. A modified version that allows multiple core nodes being selected in a community is also introduced [205]. Other community structure based methods include k -medoid that uses information transfer probabilities between any node pairs [250], and the influence maximization algorithm based on label propagation [252].

- **CCPA** [4]. Common neighbour and centrality based parameterised algorithm, or CCPA, is an approach for link prediction. It aims to bring together two essential properties of nodes, i.e., common neighbours and closeness centrality. The

similarity score between a pair of nodes i and j is defined as:

$$s(i, j) = \alpha \cdot (|N_i \cap N_j|) + (1 - \alpha) \cdot \frac{|V|}{d(i, j)}. \quad (2.48)$$

$|N_i \cap N_j|$ is obviously the part of common neighbours. $\frac{|V|}{d(i, j)}$, reciprocal of the normalised distance between two nodes, is deemed as the closeness centrality of them, since it has a similar form as the classic node closeness centrality. $\alpha \in [0, 1]$ is a user defined parameter controlling the weight of the two parts. Experiments on real-world datasets suggest that the change in performance (measured in average AUC) caused by the change of α is not significant.

2.4 Discussion and Outlook

In this section, we highlight some critical challenges and research avenues for future studies, and further discuss graph structural measures in different types of networks.

Network data, besides the pure topological presence, are often accompanied by rich information of node attributes and/or edge attributes, and they are also referred to as labelled networks or attributed networks. Most structural measures, as the name suggests, focus solely on capturing the part of topological properties. Theoretically, message passing approaches are able to include numeric node attributes, such as the initial rank and source of rank in PageRank [30], or the endogenous and exogenous status in alpha centrality [25]. In practice though, these features are usually set to identical values for all nodes, for example, all ones for initial rank and 0.15 for source of rank in PageRank. Multidimensional features are not supported in message passing approaches either. There are also attempts to integrating node/edge attributes with other graph structural measures. For instance, degree and betweenness centralities are combined with node attributes in studying crimi-

nal networks [29]; nodes attributes are used as threshold in LRIC index [10]; and node/edge attributes are fused into graphlets [201, 102]. It is also worth mentioning that one reason of the popularity of graph neural network is that it naturally enables integrating node attributes. Some recent works also propose to take edge attributes into account in GNNs [79, 108, 39]. We believe there is still great potential for developing novel structural approaches that integrate rich information on nodes and/or edges. Specifically, we find a research gap of leveraging edge attribute information in graphlets, which will be presented in Chapter 5.

Next open problem is benchmarking. An approach is usually proposed for some general tasks (such as ranking influential nodes, link prediction, network classification, etc), and tested on limited datasets. Researchers may test their approaches for different purposes on different datasets, but only report the most promising results. Some benchmarking methods have been recently proposed, but they only focus on some particular approaches with limited number of datasets [28, 11]. Therefore, we are still in need of an extensive graph structural benchmark that encompasses important graph analysis tasks and covers a diverse range of datasets. Further, when it comes to identifying and ranking influential spreaders in complex networks, susceptible-infected-recovered model (SIR) is dominantly chosen in most experiment settings [14, 147, 153, 229]. The simplicity of SIR model makes it popular to use, but also oversimplifies the complex spreading processes [221]. A recent study has shown that SIR model is unable to forecast the actual spreading pattern of epidemic in the long term [164]. Thus, it is desirable to test the performances of those identifying and ranking approaches with more complicated spreading models, where upon recovery there is no immunity (SIS model [185]), where immunity lasts only for a short period of time (SIRS model [138]), and where the disease has a latent period during which the person is not infectious (SEIS/SEIR model [140]).

Most approaches covered in the literature review assume that networks are static

or time-independent. Many real-world networks, however, are in fact dynamic, nodes and edges appearing and disappearing over time [95, 143]. In telecommunication networks, the connection between agents is often bursty and fluctuates across time; in social networks, relationships among people are typically intermittent and recurrent; in transportation networks, the frequency of public transport service is usually higher in rush hours. This extra dimension of time adds richness and complexity to the graph representation of a system, necessitating the development of more advanced approaches that can leverage temporal information. Many studies have generalised the classic graph structural measures to dynamic networks, including temporal degree centrality [114], temporal clustering coefficient [175], temporal closeness and betweenness centrality [113], temporal eigenvector centrality [218], temporal Katz centrality [175], temporal motifs [121, 184] and temporal graphlets [98]. Despite the large number of structural measures proposed for dynamic networks, there are still many open questions to be tackled. For example, what is the impact of the temporal network's structure on the dynamics of processes that occur on it; how to apply temporal measures in inferring spreading chains in incomplete temporal networks, etc. Furthermore, the previously introduced category of subgraph formation based approaches is exactly based on the dynamic nature of networks — a particular local structure is build from a less complex structure plus edges or nodes that would appear in the future. We find two research gaps regarding the subgraph formation approaches, which will be presented in Chapter 3 and Chapter 4.

Sometimes, systems are so complicated that multiple-layered networks are needed to better represent and study them [49, 44, 117, 23, 21]. For example, a multilayer social network incorporates both friendship and financial relationships among individuals; a multilayer brain network contains both anatomical brain layer and functional brain layer; and a multilayer transportation network integrates all sorts of transportation. Since interlayer connections cause new structural and dynamic cor-

relations between components, neglecting them or simply aggregating over layers will alter the original topological properties. Therefore, it is desirable to develop structural measures taking interlayer relationships into consideration. Not surprisingly, fundamental single-layer approaches have been largely generalised to multilayer networks, such as multilayer degree, clustering coefficient, closeness and betweenness centralities, [52, 31, 44, 23], multilayer motifs and graphlets [19, 50], multilayer eigenvector, PageRank and HITS centralities [46, 85, 45]. Some tailor-made approaches for multilayer networks are also recently introduced, for example, the minimal-layers power community index [18] and the singular vector of tensor centrality [226]. The study of multilayer structures, however, is still in an early stage. There is still much room for developing new cross-layer structural approaches that better model inter-layer spreading processes [206] and capture multiplex dynamics and controllability [107].

2.5 Conclusion

Since the emergence of network science, graph structural measures have been practical and powerful tools for analysing and understanding graph data in various domains. In this survey, we extensively reviewed the state-of-the-art progress in graph structural measures and proposed to divide them into five categories, i.e., subgraph count measures, subgraph formation measures, global path measures, message passing measures and hybrid measures. The first two categories are efficient to compute as they only require local information; the third category, based on unbounded paths over the entire network, is defined at global-level; the fourth category, utilises graph structure in an implicit way, i.e., gathering information from neighbours in an iterative manner; and the fifth category is a mix of previous categories. Finally, we discussed four open problems indicating the major challenges and future research directions of graph structural measures, which are limited work combining structure

with attributes, limited benchmark data and processes, and limited work concerning temporal and multilayer networks. We hope this comprehensive review and new taxonomies of graph structural measures would bring new perspectives in understanding existing approaches and serve as a starting point for future approaches. This work fulfils our first research objective. Based on the literature review, the following three chapters aim to close gaps in graph local structural measures, especially on subgraph formation measures and approaches that deal with attributed networks.

Chapter 3

Directed Closure Coefficient

3.1 Introduction

Networks, abstracting the interactions between components, are fundamental in studying complex systems in a variety of domains ranging from cellular and neural networks to social, communication and trade networks [172, 16]. Small subgraph patterns (also known as motifs [163] or graphlets [190]) that recur at a higher frequency than those in random networks are crucial in understanding and analysing networks. Motifs underlie many descriptive and predictive applications such as community detection [75, 182, 211, 235], anomaly detection [176, 130], role analysis [90, 169], and link prediction [249, 208].

Among them, 3-node connected subgraphs, which are the building blocks for higher-order motifs, are explored most often. Further, the 3-clique, or the triadic closure [56] from a temporal perspective, has been revealed to be a natural phenomenon of networks across different areas [163, 110]. Nodes sharing a common neighbour are more likely to connect with each other. For example, in an undirected friendship network, there is an increased likelihood for two people having a common friend to become friends [193]; in a directed citation network, a paper cites two papers where one tends to cite the other [234]; and in a signed directed trust network, when Alice distrusts Bob, Alice discounts anything recommended by Bob [109].

The classic measure of a 3-clique formation is the *local clustering coefficient* [232], which is defined by the percentage of the number of triangles formed with a

node (referred to as node i) to the number of triangles that i could possibly form with its neighbours. Note that in this definition, the focal node i serves as the centre-node in an open triad [219]. To emphasize, an open triad is an unordered pair of edges sharing one node. With a focus on node i , it describes the extent to which edges congregate around it. The extensions of local clustering coefficient have been thoroughly discussed for weighted networks [17, 177, 244], directed networks [63] and signed networks [129, 42]. Another metric for 3-clique formation, with a focus on an edge (referred to as e_{ij} connecting node i and j), is the *edge clustering coefficient* [228] which evaluates to what extent nodes cluster around this edge. It is calculated as the number of triangles containing e_{ij} , divided by the number of all possible triangles e_{ij} could form with other edges incident to nodes i and j .

A recent study has proposed another interesting local edge clustering measure, i.e., the *local closure coefficient* [241]. With the focal node i as the end-node of an open triad, it is quantified as the percentage of two times the number of triangles containing i to the number of open triads with i as the end-node. Conceptually, the local clustering coefficient measures the phenomenon that two friends of mine are also friends themselves, while the local closure coefficient is focusing on a friend of my friend is also a friend of mine. This new metric has been proven to be a useful tool in several network analysis tasks such as community detection and link prediction [241]. Together with the two measures mentioned above, we propose a classification diagram of all three local clustering measures (Figure 3.1).

The local closure coefficient is originally defined for undirected binary networks. However, in real-world complex networks, the relationships between components can be nonreciprocal (a follower is often not followed back by the followee), heterogeneous (trade volumes between countries vary significantly), and negative (an individual can be disliked or distrusted).

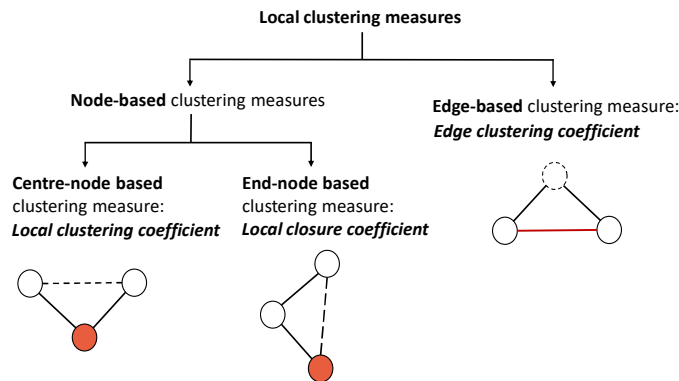


Figure 3.1 : Classification diagram of local clustering measures. In each of the two node-based clustering measures, the focal node is painted in red, and the dotted edge represents the potential closing edge in an open triad. In the edge-based clustering measure, the focal edge is in red, and the dotted outline circle represents the potential node that forms a triangle.

In this chapter, with an end-node focus, we propose the *local directed closure coefficient* to measure local edge clustering in binary directed networks, and we extend it for weighted directed networks and weighted signed directed networks. Since in a directed 3-clique, each of the three edges can take either direction, there are eight different triangles in total. According to the direction of the closing edge, i.e., the edge that closes an open triad and forms a triangle, we classify them into two groups (emanating from or pointing to the focal node, as shown in Figure 3.2(a)). Based on that, we propose the *source closure coefficient* and the *target closure coefficient* respectively.

Further, from a transitive perspective, we categorize all directed triangles into four patterns: (i) a head-of-path pattern, where the focal node is at the beginning of the length-2 path; (ii) a mid-of-path pattern, where the focal node serves as an intermediate node in the length-2 path; (iii) an end-of-path pattern, where the focal node is the endpoint of the length-2 path; (iv) a cyclic pattern, where the triads are not transitive with the focal node in a cyclical path (Figure 3.2(b)). The definition of the directed closure coefficient for each pattern is also given explicitly.

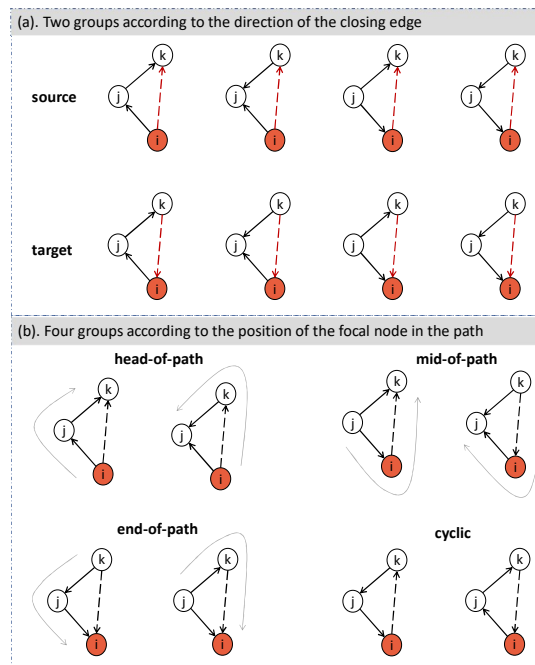


Figure 3.2 : Taxonomy of directed triangles. Two solid edges connecting nodes i , j and k form an open triad, which is closed by a dotted edge connecting nodes i and k . Focal node i , painted in red, is the end-node of an open triad. (a) Eight triangles are classified into two groups according to the direction of the closing edge. First row shows a group where the focal node serves as the source node of the closing edge; second row is another group where the focal node serves as the target. (b) Eight Triangles are classified into four groups from a transitive perspective. In six transitive triads, three different patterns are distinguished by the position of node i in a length-2 path (emphasized by grey curved arrows): head-of-path, mid-of-path, and end-of-path patterns. The remaining two non-transitive triads are classified as a cyclic pattern.

Our evaluations have revealed some interesting properties of the proposed metric. Through a correlation analysis on various networks, it is shown that the directed closure coefficient provides complementary information to the classical metric, i.e., the directed clustering coefficient. We also demonstrate how the four patterns can be used in analysing different types of directed networks.

In a link prediction task, we propose two indices that include the source closure coefficient and the target closure coefficient. We show that in most networks, adding closure coefficients leads to better performance. Finally, the usage of the *weighted signed directed closure coefficient* is illustrated through a case study. We see that it

is a more accurate measure compared to its original counterpart.

In summary, the main contributions of this chapter are as follows:

- We propose the directed closure coefficient as a novel measure of edge clustering in directed networks;
- We extend the directed closure coefficient to weighted and signed networks;
- We propose the four closure patterns for end-node-based directed triangles.
- We formulate the source and target closure coefficients and propose an algorithm of link prediction for directed networks;
- Theoretical and empirical studies demonstrate the intrinsic properties of the proposed metrics and their utilities in multiple network analysis tasks.

This work attains research objectives 2 and 5.

3.2 Preliminaries

This section introduces the preliminary knowledge of our work, including the classic clustering coefficient, its extension in directed networks, and the recently proposed closure coefficient.

3.2.1 Clustering coefficient

The notion of local clustering coefficient was originally proposed bearing the name clustering coefficient, in order to measure the cliquishness of a neighbourhood in an undirected graph [232].

Let $G = (V, E)$ be an undirected graph on a node set V (the number of nodes is $|V|$) and an edge set E , without multiple edges and self-loops. The adjacency

matrix of G is denoted as $\mathbf{A} = \{a_{ij}\}$. $a_{ij} = 1$ if there is an edge between node i and node j , otherwise $a_{ij} = 0$. We denote the degree of node i as $d_i = \sum_j a_{ij}$.

For any node $i \in V$, the *local clustering coefficient* is calculated as the number of triangles formed with node i and its neighbours (labelled as $T(i)$), divided by the number of open triads with i as the centre-node (labelled as $OT_c(i)$):

$$C_c(i) = \frac{T(i)}{OT_c(i)} = \frac{\frac{1}{2} \sum_j \sum_k a_{ij} a_{ik} a_{jk}}{\frac{1}{2} d_i (d_i - 1)}. \quad (3.1)$$

The subscript c here emphasizes that the focal node i serves as the centre-node of an open triad. We assume that $C_c(i)$ is well defined. Clearly, $C_c(i) \in [0, 1]$.

In order to measure clustering at the network-level, the *average clustering coefficient* is introduced by averaging the local clustering coefficient over all nodes (an undefined local clustering coefficient is treated as zero): $\overline{C_c} = \frac{1}{|V|} \sum_{i \in V} C_c(i)$.

Another frequently used measure of clustering at the network-level is the *global clustering coefficient* [174], which is defined as the fraction of open triads that form triangles in the entire network:

$$C_c = \frac{\sum_i \sum_j \sum_k a_{ij} a_{ik} a_{jk}}{\sum_{i \in V} d_i (d_i - 1)}. \quad (3.2)$$

Note that the global clustering coefficient is not equivalent to the average clustering coefficient. In fact, they can be very distinct from each other.

3.2.2 Directed clustering coefficient

Fagiolo[63] proposed an extension of the local clustering coefficient to directed networks, which takes into account all possible directed triangles formed around a focal node. In total, there are eight different triangles (each of the three edges can have two directions). When a directed open triad (or a directed triangle) contains

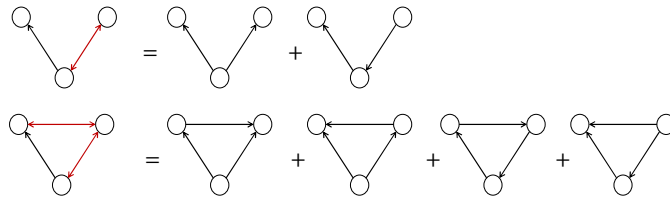


Figure 3.3 : Dealing with bidirectional edges. The first row shows that an open triad with one bidirectional edge is counted as two unidirectional open triads; the second row shows that a triangle with two bidirectional edges is counted as four unidirectional triangles.

bidirectional edges, they are treated as a combination of open triads (or triangles) with only unidirectional edges (Figure 3.3).

Let us denote $\mathbf{A} = \{a_{ij}\}$ as the adjacency matrix of a directed graph $G^{\mathcal{D}} = (V, E)$. $a_{ij} = 1$ if there is an edge from node i to node j , otherwise $a_{ij} = 0$. The degree of node i is denoted as d_i , including both outgoing edges and incoming edges: $d_i = d_i^{\text{out}} + d_i^{\text{in}} = \sum_j a_{ij} + \sum_j a_{ji}$. d_i^{\leftrightarrow} denotes the degree of bidirectional edges of i : $d_i^{\leftrightarrow} = \sum_j a_{ij}a_{ji}$.

The *local directed clustering coefficient* is thus defined as the number of directed triangles formed with node i and its neighbours (counted as unidirectional ones, labelled as $T^{\mathcal{D}}(i)$), divided by twice the number of directed open triads with i as the centre-node (labelled as $OT_c^{\mathcal{D}}(i)$):

$$\begin{aligned}
 C_c^{\mathcal{D}}(i) &= \frac{T^{\mathcal{D}}(i)}{2OT_c^{\mathcal{D}}(i)} \\
 &= \frac{(1/2) \sum_j \sum_k (a_{ij} + a_{ji}) (a_{ik} + a_{ki}) (a_{jk} + a_{kj})}{d_i (d_i - 1) - 2d_i^{\leftrightarrow}}.
 \end{aligned} \tag{3.3}$$

Note that $OT_c^{\mathcal{D}}(i)$ equals to $(1/2) [d_i (d_i - 1) - 2d_i^{\leftrightarrow}]$. $OT_c^{\mathcal{D}}(i)$ is multiplied by two because the edge closes a directed open triad can take two directions.

Similarly, the *average directed clustering coefficient* of the entire network is defined as: $\overline{C_c^{\mathcal{D}}} = |V|^{-1} \sum_{i \in V} C_c^{\mathcal{D}}(i)$.

One might have expected the existence of a directed version of the global clustering coefficient. However, somewhat surprisingly, such a measure has not appeared in the literature (Seshadhri et al.[209] introduced a global directed clustering measure, but only for each type of directed triangles, including triangles with bidirectional edges). We therefore give a definition here, which is a natural extension of the local directed clustering coefficient.

Definition 1. *The **global directed clustering coefficient** of a directed network, denoted $C_c^{\mathcal{D}}$, is defined as the fraction of directed open triads that form triangles in the entire network:*

$$C_c^{\mathcal{D}} = \frac{\frac{1}{2} \sum_i \sum_j \sum_k (a_{ij} + a_{ji})(a_{ik} + a_{ki})(a_{jk} + a_{kj})}{\sum_{i \in V} (d_i(d_i - 1) - 2d_i^{\leftrightarrow})}. \quad (3.4)$$

The numerator here equals three times the number of directed triangles in the entire network (each node of a triangle contributes an open triad with it as the centre-node).

3.2.3 Closure coefficient

Recently Yin et al.[241] proposed the *local closure coefficient* and thus closed a gap in the clustering measure on undirected networks. Different from the ordinary centre-node focus in the local clustering coefficient, this definition is based on the end-node of an open triad. Recall that an open triad is an unordered pair of edges sharing one node. For example, in an open triad ijk with two edges ij and jk , there is no difference between (ij, jk) and (jk, ij) .

Using the notations for undirected graph, the local closure coefficient of node i is defined as two times the number of triangles formed with i (labelled as $T(i)$), divided by the number of open triads with i as the end-node (labelled as $OT_e(i)$):

$$C_e(i) = \frac{2T(i)}{OT_e(i)} = \frac{\sum_j \sum_k a_{ij} a_{ik} a_{jk}}{\sum_{j \in N(i)} (d_j - 1)}, \quad (3.5)$$

where $N(i)$ denotes the set of neighbours of node i . $T(i)$ is multiplied by two for the reason that each triangle contains two open triads with i as the end-node. When a triangle is actually formed (e.g., with nodes i , j and k), the focal node i can be viewed as the centre-node in one open triad ($j\dot{i}k$) or as the end-node in two open triads ($\dot{i}jk$ and $\dot{i}kj$). Obviously, $C_e(i) \in [0, 1]$.

At the network-level, the *average closure coefficient* is then defined as the mean of the local closure coefficient over all nodes (undefined local closure coefficient is treated as zero): $\overline{C_e} = \frac{1}{|V|} \sum_{i \in V} C_e(i)$. When we consider a random network where each pair of nodes is connected with a probability p , its expected value is also p , i.e., $\mathbb{E}[\overline{C_e}] = p$.

Analogous to the global clustering coefficient (see Equation 3.2), we give a global version of the closure coefficient.

Definition 2. *The **global closure coefficient** of an undirected network, denoted C_e , is defined as :*

$$C_e = \frac{2 \sum_{i \in V} T(i)}{\sum_{i \in V} \sum_{j \in N(i)} (d_j - 1)}. \quad (3.6)$$

The numerator is equal to six times the number of triangles in the entire network (each node of a triangle contributes two open triads with it as the end-node), then divided by twice the number of open triads constructed from the end-node in the entire network.

This definition is actually equivalent to the global clustering coefficient (Equation 3.2) as globally the difference of the position of the focal node will not surface.

Proposition 1. *In any undirected network, $C_e = C_c$.*

Proof. Since globally the neighbourhood relationship is reciprocal, $\sum_{i \in V} \sum_{j \in N(i)} (d_j - 1)$ can be written as $\sum_{j \in V} \sum_{i \in N(j)} (d_j - 1)$ which equals $\sum_{j \in V} d_j (d_j - 1)$. Then we have $\sum_{i \in V} \sum_{j \in N(i)} (d_j - 1) = \sum_{i \in V} d_i (d_i - 1)$. Thus, $C_e = C_c$. \square

3.3 Closure Coefficient in Directed Networks

The local closure coefficient has been proven to be a useful metric in undirected networks [241]. In this section, we provide a general extension of it to directed networks, i.e., the local directed closure coefficient. We further propose the closure coefficients of particular patterns. Finally, we extend it into weighted directed networks and signed weighted directed networks.

3.3.1 Closure coefficient in binary directed networks

Motivated by the closure coefficient and the directed clustering coefficient, we aim to measure the directed 3-clique formation from the end-node of an open triad. There are eight different directed triangles, and similarly a triangle (or an open triad) with bidirectional edges is treated as a combination of triangles (or open triads) with only unidirectional edges (Figure 3.3).

Using the notations from Section 3.2, we now give the definition of the closure coefficient in directed networks.

Definition 3. *The **local directed closure coefficient** of node i in a directed network, denoted $C_e^{\mathcal{D}}(i)$, is defined as twice the number of directed triangles formed with node i (labelled as $T^{\mathcal{D}}(i)$), divided by twice the number of directed open triads with i as the end-node (labelled as $OT_e^{\mathcal{D}}(i)$):*

$$\begin{aligned} C_e^{\mathcal{D}}(i) &= \frac{2T^{\mathcal{D}}(i)}{2OT_e^{\mathcal{D}}(i)} \\ &= \frac{\sum_j \sum_k (a_{ij} + a_{ji})(a_{ik} + a_{ki})(a_{jk} + a_{kj})}{2 \sum_{j \in N(i)} (a_{ij} + a_{ji})(d_j - (a_{ij} + a_{ji}))}. \end{aligned} \quad (3.7)$$

When the neighbours of i are solely connected to i , the local directed closure coefficient is undefined. In real-world networks, however, nodes with undefined closure coefficient are very rare.

$T^{\mathcal{D}}(i)$ is multiplied by two since each triangle contains two open triads with i as the end-node. $OT_e^{\mathcal{D}}(i)$ is multiplied by two because the closing edge of a directed open triad can take two directions. Obviously, $C_e^{\mathcal{D}}(i) \in [0, 1]$. When the adjacency matrix \mathbf{A} is symmetric (the network becomes undirected), Equation 3.7 reduces to Equation 3.5, i.e., $C_e^{\mathcal{D}}(i) = C_e(i)$.

Similarly, in order to measure at the network-level, we propose the definition of an average directed closure coefficient and a global directed closure coefficient.

Definition 4. *The **average directed closure coefficient** of a directed network, denoted $\overline{C_e^{\mathcal{D}}}$, is defined as the average of the local directed closure coefficient over all nodes:*

$$\overline{C_e^{\mathcal{D}}} = \frac{1}{|V|} \sum_{i \in V} C_e^{\mathcal{D}}(i), \quad (3.8)$$

in which an undefined local directed closure coefficient is treated as zero. In a random network, where each directed edge occurs with a probability p , we also have $\mathbb{E}[C_e^{\mathcal{D}}(i)] = p$.

Definition 5. *The **global directed closure coefficient** of a directed network, denoted $C_e^{\mathcal{D}}$, is defined as:*

$$C_e^{\mathcal{D}} = \frac{2 \sum_{i \in V} T^{\mathcal{D}}(i)}{2 \sum_{i \in V} \sum_{j \in N(i)} (a_{ij} + a_{ji}) (d_j - (a_{ij} + a_{ji}))}, \quad (3.9)$$

where the numerator equals six times the number of directed triangles in the entire network (each node of a triangle contributes two open triads with it as the end-node), divided by twice the number of directed open triads across the network.

Similar to Proposition 1 and its proof, the global directed closure coefficient is equivalent to the global directed clustering coefficient (Equation 3.4).

Proposition 2. *In any directed network, $C_e^{\mathcal{D}} = C_c^{\mathcal{D}}$.*

3.3.2 Closure coefficients of particular patterns

In addition to a general measure, we propose to have a closer look at the directed closure coefficients of particular patterns in order to gain a deeper understanding and fully realise the potential of this metric.

First and foremost, we classify directed triangles into two groups according to the direction of the closing edge: one group where the focal node serves as the source node of the closing edge, another group where the focal node serves as the target (Figure 3.2(a)). Two definitions are given accordingly.

Definition 6. For a given node i in a directed network, the **source closure coefficient**, denoted $C_e^{src}(i)$, and the **target closure coefficient**, denoted $C_e^{tgt}(i)$ are defined as:

$$C_e^{src}(i) = \frac{\sum_j \sum_k (a_{ij} + a_{ji}) (a_{jk} + a_{kj}) a_{ik}}{2 \sum_{j \in N(i)} (a_{ij} + a_{ji}) (d_j - (a_{ij} + a_{ji}))},$$

$$C_e^{tgt}(i) = \frac{\sum_j \sum_k (a_{ij} + a_{ji}) (a_{jk} + a_{kj}) a_{ki}}{2 \sum_{j \in N(i)} (a_{ij} + a_{ji}) (d_j - (a_{ij} + a_{ji}))}.$$

Please note that $C_e^{src}(i) + C_e^{tgt}(i) = C_e^{\mathcal{D}}(i)$. These two metrics evaluate the extent to which the focal node is acting as the source node or the target node of the closing edges in a triangle formation. Note that there are no analogous definitions for the clustering coefficient because the closing edge is not incident to the focal node that serves as the centre-node of the open triad. In the next section, we show how the source/target closure coefficients can be used to improve the performance in a link prediction task.

Secondly, several studies have shown that the three-node transitive closure (also called the feedforward loop) prevails in many real-world networks [163, 63, 110]. Thus, we propose to categorize the eight directed triangles into four patterns from a transitive perspective: three transitive patterns distinguished by the position of

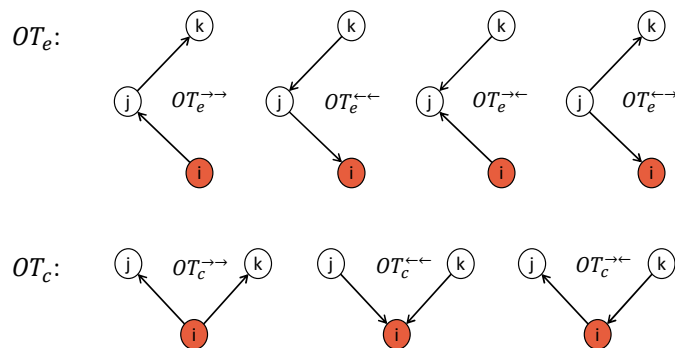


Figure 3.4 : Directed open triads. Upper row illustrates four different open triads with the focal node i as the end-node. Two arrows of the superscript describe the directions of two edges: First arrow depicts an edge from i to j (\rightarrow) or from j to i (\leftarrow); second arrow depicts the direction of an edge between j and k with regard to i (\rightarrow denotes an edge from j to k ; \leftarrow denotes an edge from k to j). As a comparison, lower row illustrates three different open triads with i as the centre-node. First arrow depicts the edge direction between i and j while second arrow depicts the edge direction between i and k . There are three instead of four since $OT_c^{\rightarrow\leftarrow}$ is equivalent to $OT_c^{\leftarrow\rightarrow}$.

the focal node in a length-2 path, plus one non-transitive pattern (Figure 3.2(b)).

Before introducing the definitions of directed closure coefficients of these four patterns, we first characterize four types of directed open triads with the focal node as the end-node, a comparison with centre-node focused triads is also provided (Figure 3.4). Then we give the following definitions.

Definition 7. *The local directed closure coefficients of four patterns, i.e., the **head-of-path closure coefficient**, denoted $C_e^{\text{head}}(i)$; the **end-of-path closure coefficient**, denoted $C_e^{\text{end}}(i)$; the **mid-of-path closure coefficient**, denoted $C_e^{\text{mid}}(i)$ and the **cyclic closure coefficient**, denoted $C_e^{\text{cyc}}(i)$ are defined as:*

$$\begin{aligned}
 C_e^{\text{head}}(i) &= \frac{2T^{\text{head}}(i)}{OT_e^{\rightarrow\rightarrow}(i) + OT_e^{\rightarrow\leftarrow}(i)} \\
 &= \frac{\sum_j \sum_k a_{ij} a_{ik} (a_{jk} + a_{kj})}{\sum_{j \in N(i)} a_{ij} (d_j - (a_{ij} + a_{ji}))},
 \end{aligned}$$

$$\begin{aligned}
C_e^{end}(i) &= \frac{2T^{end}(i)}{OT_e^{\leftarrow\leftarrow}(i) + OT_e^{\leftarrow\rightarrow}(i)} \\
&= \frac{\sum_j \sum_k a_{ji} a_{ki} (a_{jk} + a_{kj})}{\sum_{j \in N(i)} a_{ji} (d_j - (a_{ij} + a_{ji}))}, \\
C_e^{mid}(i) &= \frac{2T^{mid}(i)}{OT_e^{\rightarrow\leftarrow}(i) + OT_e^{\leftarrow\rightarrow}(i)} \\
&= \frac{\sum_j \sum_k (a_{ji} a_{ik} a_{jk} + a_{ki} a_{ij} a_{kj})}{\sum_{j \in N(i)} (a_{ij} (d_j^{in} - a_{ij}) + a_{ji} (d_j^{out} - a_{ji}))}, \\
C_e^{cyc}(i) &= \frac{2T^{cyc}(i)}{OT_e^{\rightarrow\rightarrow}(i) + OT_e^{\leftarrow\leftarrow}(i)} \\
&= \frac{\sum_j \sum_k (a_{ji} a_{ik} a_{kj} + a_{ki} a_{ij} a_{jk})}{\sum_{j \in N(i)} (a_{ji} (d_j^{in} - a_{ij}) + a_{ij} (d_j^{out} - a_{ji}))}.
\end{aligned}$$

As shown above, the numerator of each coefficient equals twice the number of particular triangles; the denominator can be calculated with the neighbourhood information of node i and the degree information of i 's neighbours.

The significance of defining closure coefficients of these four patterns are twofold. First, at the node-level analysis, they can be applied directly to measure whether a node of interest is more of an initiator (higher head-of-path closure coefficient), an intermediary (higher mid-of-path closure coefficient) or a target (higher end-of-path coefficient).

Secondly, at the network-level, they can also serve as interesting features. Adopting a similar approach which uses clustering signatures to classify networks [5], we introduce a normalised closure coefficient of each pattern for a given node i :

$$\tilde{C}_e^*(i) = \frac{C_e^*(i)}{C_e^{head}(i) + C_e^{mid}(i) + C_e^{end}(i) + C_e^{cyc}(i)},$$

where $*$ = {*head, mid, end, cyc*}. We then average it over the entire network:

$$\tilde{C}_e^* = \frac{1}{|V|} \sum_{i \in V} \tilde{C}_e^*(i). \quad (3.10)$$

This *average normalised closure coefficient* of each pattern can be used to describe or analyse networks.

3.3.3 Closure coefficient in weighted networks

So far, the study is focusing on binary networks, where the value of every edge is either 1 or 0. In many networks, however, we need a more accurate representation of the relationships between nodes, such as the frequency of contact in a social network, the traffic flow in a road network, etc. This is why we are also interested in giving a definition of a closure coefficient for weighted networks.

We begin with weighted undirected networks. Several versions of weighted clustering coefficients have been summarised in [207]. Among them, a definition given by Onnela et al. [177] and another given by Zhang and Horvath [244] are often employed. After normalisation (maximum weight normalised to 1), the former takes a geometric average of weights of actually formed triangles, divided by the number of potential triangles, which implies all edges taking the maximum weight in the denominator. The latter chooses a simple product of weights of formed triangles, divided by the product of two weights of an open triad, implying the potential triadic closing edge taking the maximum weight.

In our definition of weighted closure coefficient, similar to the method proposed by Zhang and Horvath[244], we choose to only assign a maximum weight to the closing edge. In a weighted graph G^w described by its weight matrix $\mathbf{W} = \{w_{ij}\}$, we suppose $w_{ij} \in [0, 1]$ (normalised by the maximum weight), and the strength of node i is $s_i = \sum_j w_{ij}$.

Definition 8. The *weighted closure coefficient* of node i in a weighted network, denoted $C_e^{\mathcal{W}}(i)$, is defined as:

$$C_e^{\mathcal{W}}(i) = \frac{\sum_j \sum_k w_{ij} w_{ik} w_{jk}}{\sum_{j \in N(i)} w_{ij} (s_j - w_{ij})}. \quad (3.11)$$

Obviously, $C_e^{\mathcal{W}}(i) \in [0, 1]$. When the weight matrix becomes binary, Equation 3.11 degrades to Equation 3.5, i.e., $C_e^{\mathcal{W}}(i) = C_e(i)$.

In a similar approach, the definition of closure coefficient in weighted directed networks can be extended from Equation 3.7. Let us denote $\mathbf{W} = \{w_{ij}\}$ as the weight matrix of a weighted directed graph $G^{\mathcal{W}, \mathcal{D}}$, $w_{ij} \in [0, 1]$. The strength of node i is denoted by s_i ($s_i = \sum_j w_{ij} + \sum_j w_{ji}$).

Definition 9. The *weighted directed closure coefficient* of node i , denoted $C_e^{\mathcal{W}, \mathcal{D}}(i)$, is defined as:

$$C_e^{\mathcal{W}, \mathcal{D}}(i) = \frac{\sum_j \sum_k (w_{ij} + w_{ji}) (w_{ik} + w_{ki}) (w_{jk} + w_{kj})}{2 \sum_{j \in N(i)} (w_{ij} + w_{ji}) (s_j - (w_{ij} + w_{ji}))}. \quad (3.12)$$

Last but not least, we discuss the closure coefficient in weighted signed networks. In many settings, the weights of relationships can be both positive and negative, as a person may trust or distrust others with different levels of intensity.

Let $G^{\mathcal{W}^{\pm}, \mathcal{D}}$ be a weighted signed directed graph. Its signed weight matrix is denoted by $\mathbf{W} = \{w_{ij}\}$, $w_{ij} \in [-1, 1]$. The absolute weight matrix is denoted by $\mathbf{P} = \{p_{ij}\}$, where $p_{ij} = |w_{ij}|$. And the strength of node i is indicated by \bar{s}_i ($\bar{s}_i = \sum_j p_{ij} + \sum_j p_{ji}$).

Definition 10. The *weighted signed directed closure coefficient* of node i ,

denoted $C_e^{\mathcal{W}^\pm, \mathcal{D}}(i)$, is defined as:

$$C_e^{\mathcal{W}^\pm, \mathcal{D}}(i) = \frac{\sum_j \sum_k (w_{ij} + w_{ji})(w_{ik} + w_{ki})(w_{jk} + w_{kj})}{2 \sum_{j \in N(i)} (p_{ij} + p_{ji})(\bar{s}_i - (p_{ij} + p_{ji}))}. \quad (3.13)$$

Obviously, $C_e^{\mathcal{W}^\pm, \mathcal{D}}(i)$ varies in $[-1, 1]$. It is positive when positive triangles formed around the focal node outweigh negative ones. It equals zero when no triangles formed with the focal node or positive triangles and negative triangles are balanced.

3.3.4 Computational efficiency

To end this section, we give a brief discussion about the computational efficiency of the aforementioned metrics. For the sake of explanation and understanding, we use the adjacency matrix of the network to present equations, which leads up to $O(|V|^3)$ in computation.

In actual development, however, after conveniently obtaining the neighbourhood information (both successors and predecessors in directed networks) of each node, the computational cost is $O(|V| \cdot \bar{k}^2)$, where \bar{k} is the average degree of the network. As in most real networks $\bar{k} \ll |V|$, the computation of these proposed metrics is therefore fast in large networks.

3.4 Experiments and Analysis

In this section, we evaluate the proposed directed closure coefficient in real-world networks. First, we compare it with the classic directed clustering coefficient. Then, we show how it can be applied in link prediction to improve the performance. We finish with a case study in a weighted signed directed network.

3.4.1 Directed closure coefficient in real-world networks

Datasets. We run experiments on 12 directed networks from different domains:

1. Six social networks.
 - (a) Two friendship networks. ADO-HEALTH[166]: a positively weighted friendship network created from a survey; DIGG-FRIENDS [94]: an online friendship network of news aggregator Digg.
 - (b) Three trust networks. BTC-ALPHA [126]: a weighted and signed trust network of users on Bitcoin Alpha; EPINIONS [159]: a weighted and signed trust network of online product rating site Epinions; WIKI-VOTE[133]: a network describing Wikipedia elections.
 - (c) One communication network. COLLEGEMSG[183]: a network comprised of messages between students.
2. Two citation networks. ARXIV-HEPPH[134]: a citation network from arXiv; US-PATENT[84]: a citation network of patents registered in the US.
3. Two online Q&A networks. ASKUBUNTU and STACKOVERFLOW[184]: two networks from Stack Exchange.
4. Two other networks. AMAZON[132]: a network describing co-purchased products on Amazon; GOOGLE[135]: a hyperlink network.

Table 3.1 lists some key statistics of these datasets. To better compare with Definition 9 and Definition 10, we calculate the weighted directed clustering coefficient and the weighted signed directed clustering coefficient following [244] and [42]:

$$C_c^{\mathcal{W}, \mathcal{D}}(i) = \frac{\sum_j \sum_k (w_{ij} + w_{ji}) (w_{ik} + w_{ki}) (w_{jk} + w_{kj})}{\sum_j \sum_k (w_{ij} + w_{ji}) (w_{ik} + w_{ki})},$$

Table 3.1 : Statistics of datasets, showing the number of nodes ($|V|$), the number of edges ($|E|$), the average degree (\bar{k}), the proportion of reciprocal edges (r), the average directed clustering coefficient ($\overline{C_c^D}$), and the average directed closure coefficient ($\overline{C_e^D}$). Datasets having timestamps on edge creation are superscripted by (τ). Positively weighted networks are superscripted by (+), and networks having both positive and negative weights are superscripted by (\pm).

Network	$ V $	$ E $	\bar{k}	r	$\overline{C_c^D}$	$\overline{C_e^D}$
COLLEGE MSG^τ	1,899	20,296	10.69	0.636	0.087	0.017
ADO-HEALTH $^+$	2539	12,969	5.11	0.388	0.090	0.071
BTC-ALPHA $^{\pm,\tau}$	3783	24,186	6.39	83.2	0.046	0.006
WIKI-VOTE	7,115	104K	14.57	0.056	0.082	0.017
EPINIONS $^{\pm,\tau}$	132K	841K	6.38	0.308	0.085	0.010
DIGG-FRIENDS $^\tau$	280K	1,732K	6.19	0.212	0.075	0.008
ARXIV-HEP PH	34,546	422K	12.2	0.003	0.143	0.053
US-PATENT	3,775K	16,519K	4.38	0.000	0.038	0.019
ASKUBUNTU $^\tau$	79,155	199K	2.51	0.002	0.028	2e-4
STACKOVERFLOW $^\tau$	2,465K	16,266K	6.60	0.002	0.008	2e-4
AMAZON	403K	3,387K	8.40	0.557	0.364	0.234
GOOGLE	876K	5,105K	5.83	0.307	0.370	0.097

$$C_c^{W^{\pm,D}}(i) = \frac{\sum_j \sum_k (w_{ij} + w_{ji})(w_{ik} + w_{ki})(w_{jk} + w_{kj})}{\sum_j \sum_k (p_{ij} + p_{ji})(p_{ik} + p_{ki})}.$$

We see from Table 3.1 that in all 12 networks, the average directed closure coefficient is less than the average directed clustering coefficient. In these types of networks, we may say that compared to a triangle formation from centre-node based open triads, fewer triangles are formed from the end-node based open triads. In some networks (ADO-HEALTH and AMAZON), the difference between them is not very big; while in Q&A networks, the difference is more than 40 times.

From the scatter plots of the local directed closure coefficient and the local directed clustering coefficient (Figure 3.5), we can see their relationship more clearly. First, the Pearson correlation is positive but weak (ranging from 0.134 to 0.759). Secondly, similar networks exhibit similar relationships between the two variables, as

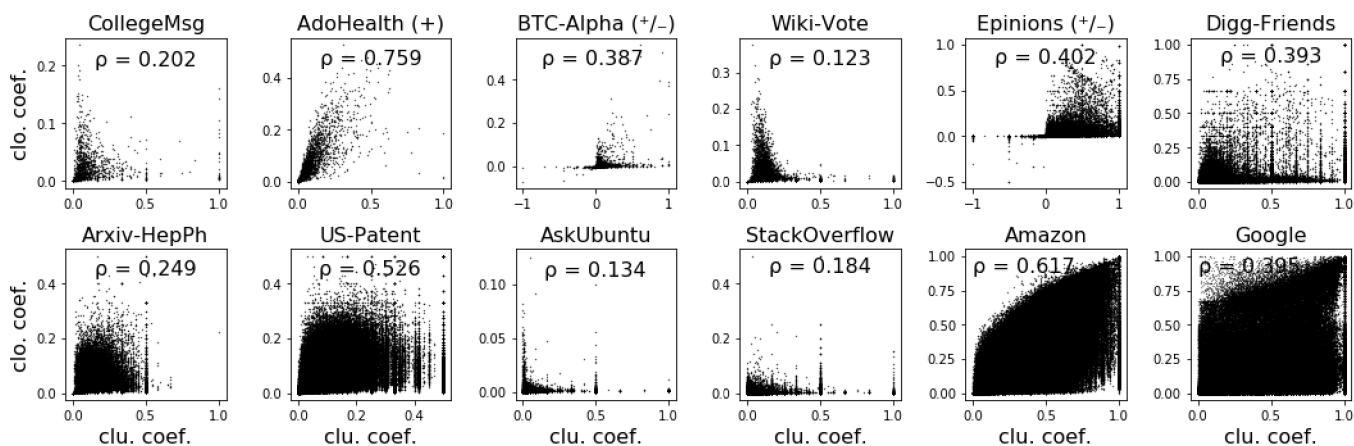


Figure 3.5 : Scatter plots of the local directed closure coefficient and the local directed clustering coefficient, with the Pearson correlation coefficient.

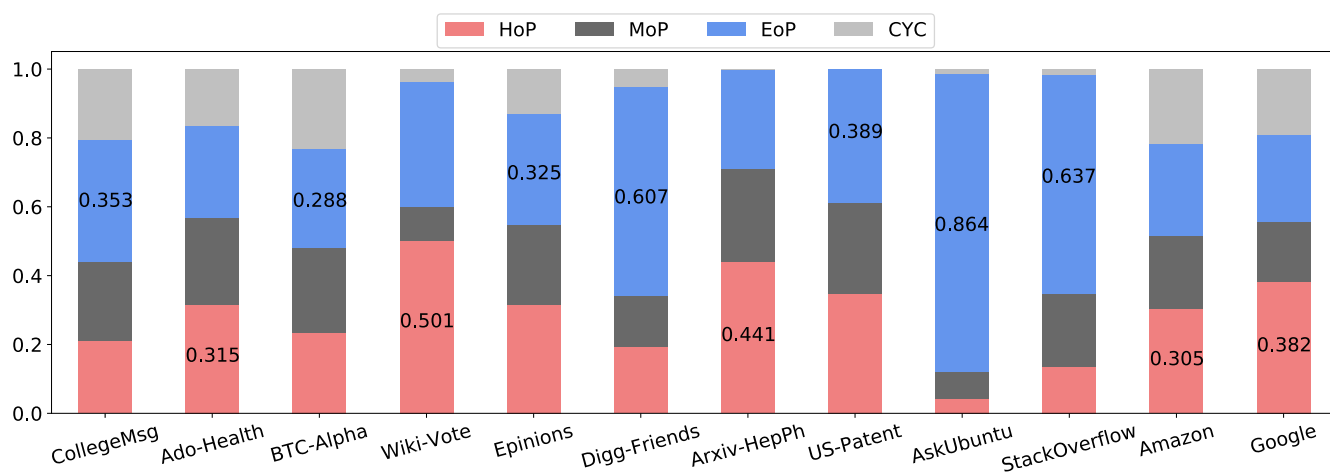


Figure 3.6 : Average normalized closure coefficients of four patterns: head-of-path (HoP), mid-of-path (MoP), end-of-path (EoP) and cyclic (CYC). The dominant pattern in each network is labelled with its value.

in two trust networks BTC-ALPHA and EPINIONS, in two citation networks ARXIV-HEPPH and US-PATENT or in two Q&A networks ASKUBUNTU and STACKOVERFLOW.

Applying Equation 3.10 in these networks, we get their average normalized closure coefficients of four patterns (Figure 3.6). It can be seen that the dominant pattern is either the head-of-path pattern or the end-of-path pattern, and that the cyclic pattern is substantially suppressed in many networks. Note that the neutral value of each pattern, corresponding to an undirected network, is 0.25.

To understand the meaning of the dominant head-of-path pattern, we take the trust network WIKI-VOTE as an example. Consider three people A , B and C , if A votes for B and B votes for C (or if A votes for B and C votes for B), A would probably vote for C . Similarly, a dominant end-of-path pattern in two Q&A networks implies that if C answers B 's question and B answers A 's question (or if B answers A 's question and B answers C 's question), it is likely that C would also answer A 's question.

3.4.2 Link prediction in directed networks

Many studies [144, 2, 254, 188, 112, 160] have shown that future interactions among nodes can be extracted from the network topology information. The key idea is to compare the proximity or similarity between pairs of nodes, either from the neighbourhoods [2, 254], the local structures [188] or the whole network [112, 160].

Most existing methods, however, focus solely on undirected networks. In this experiment, we show whether the information provided by the local directed closure coefficient can be used to enhance the performance of link prediction approaches for directed networks. As shown in [144], the neighbourhood based methods are simple yet powerful. We choose three classic similarity indices extended for directed networks as the baseline methods[251].

Let $N_{out}(i)$ be the out-neighbour set of node i (consisting of i 's successors); $N_{in}(i)$ be the in-neighbour set (consisting of i 's predecessors). The set of all neighbours $N(i)$ is the union of the two: $N(i) = N_{out}(i) \cup N_{in}(i)$. For an ordered pair of nodes (s, t) , the three baseline indices are defined below:

- Directed Common Neighbours index (DiCN)

$$DiCN(s, t) = |N_{out}(s) \cap N_{in}(t)|,$$

- Directed Adamic-Adar index (DiAA)

$$DiAA(s, t) = \sum_{u \in N_{out}(s) \cap N_{in}(t)} \frac{1}{\log |N(u)|},$$

- Directed Resource Allocation index (DiRA)

$$DiRA(s, t) = \sum_{u \in N_{out}(s) \cap N_{in}(t)} \frac{1}{|N(u)|}.$$

Proposed indices. Combining the idea of the Common Neighbours index and the source/target closure coefficients (Definition 6), we propose two indices to measure the *directed closeness* in directed networks.

Definition 11. For an ordered pair of nodes (s, t) , the **closure closeness index**, denoted $CCI(s, t)$; and the **extra closure closeness index**, denoted $ECCI(s, t)$ are defined as:

$$CCI(s, t) = |N_{out}(s) \cap N_{in}(t)| \cdot (C_e^{src}(s) + C_e^{tgt}(t)),$$

$$ECCI(s, t) = |N(s) \cap N(t)| \cdot (C_e^{src}(s) + C_e^{tgt}(t)).$$

Different from the closure closeness index, the extra closure closeness index uses the set of all neighbours, because the source closure coefficient of node s and the target closure coefficient of node t can also bring in the direction inclination.

Setup. We model a directed network as a graph $G^D = (V, E)$. For networks having timestamps on edges, we order the edges according to their appearing times and select the first 50% edges and related nodes to form an “old graph”, denoted $G_{old} = (V^*, E_{old})$. For networks not having timestamps, we randomly choose 50% edges and related nodes as G_{old} and repeat 10 times in the experiment ($r_1 = 10$).

Let E_{new} be the set of future edges among the nodes in V^* , which is also what

Table 3.2 : Performance comparison of six methods on link prediction in directed networks (Precision %). RP (second column) gives the probability that a random prediction is correct. The best performance in each network is in bold type. The number at the foot of certain datasets indicates the total repeated times. Weights are ignored in this task.

Network	RP	DiCN	DiAA	DiRA	CCI	ECCI
COLLEGEMSG ^τ	0.30	2.546	2.763	3.533	3.395	3.730
ADO-HEALTH ₍₁₀₎	0.10	8.404	8.406	8.304	10.23	11.07
BTC-ALPHA ^τ	0.05	8.588	9.269	7.313	8.418	9.226
WIKI-VOTE ₍₁₀₎	0.15	21.96	22.51	20.32	22.55	19.08
EPINIONS ₍₂₀₎ ^τ	0.37	3.613	3.662	3.531	3.490	5.106
DIGG-FRIENDS ₍₂₀₎ ^τ	0.33	6.649	6.709	6.685	7.135	5.569
ARXIV-HEPPH ₍₅₀₎	0.16	20.35	21.51	20.72	20.07	21.49
US-PATENT _(1,000)	0.04	9.787	10.14	9.987	11.67	11.31
ASKUBUNTU ₍₁₀₎ ^τ	0.03	4.100	4.912	4.163	5.412	4.697
STACKOVERFLOW ₍₁₀₀₎ ^τ	0.16	7.433	8.129	7.472	8.792	6.388
AMAZON ₍₅₀₀₎	0.06	23.71	27.94	27.43	26.76	29.46
GOOGLE ₍₅₀₀₎	1.19	44.48	52.32	50.29	49.39	46.24
<i>Average over all networks</i>	0.245	13.468	14.856	14.146	14.776	14.447

we aim to predict. Apparently, the total number of potential links on node set V^* is: $|V^*|^2 - E_{old}$. We apply each prediction method to output a list containing the similarity scores for all potential links in descending order, denoted L_p . An intersection of $L_p[0 : |E_{new}|]$ and E_{new} gives us the set of correctly predicted links, denoted E_{true} . The precision is then calculated by $|E_{true}|/|E_{new}|$.

For large networks ($|V| > 10K$), we randomly sample $5K$ connected nodes on G^D and repeat the above procedures r_2 times according to the size of the dataset. Therefore, for large networks without timestamps we run $r_1 * r_2 = 10 * r_2$ times in the experiment. For instance, we sample 50 times in the dataset AMAZON which does not have timestamps. Thus the total repeated times equals $10 * 50 = 500$.

Results and discussion. We compare three baseline methods with two pro-

posed methods (Definition 11) in Table 3.2. We see that the closure closeness index (CCI) has recorded the highest precision in 5 networks, and the extra closure closeness index (ECCI) has recorded the highest precision in 4 networks. It suggests that in most networks, including the local structure information of closure coefficient leads to improvement in link prediction. Sometimes the improvement is significant: In ADO-HEALTH and EPINIONS, ECCI is over 30% better than the baseline methods. In the other six networks (COLLEGEMSG, DIGG-FRIENDS, US-PATENT, ASKUBUNTU, STACKOVERFLOW and AMAZON), the precision of CCI or ECCI is over 5% higher than that of the baselines.

In order to offer a different perspective, we also calculated the average precision value over the 12 networks for each method. The result shows that the best performance is from the traditional DiAA approach, which is slightly higher than the CCI (around 0.5%) and ECCI approaches (less than 2%). However, we argue that the average precision score could be biased by some rare but extra-large values. For example, in the GOOGLE dataset, the traditional approach DiAA achieves a very high precision of 52.32 compared to other methods, which is the main cause of the higher average precision in DiAA. Therefore, to evaluate the performance of different methods in different types of networks, the win count is a more meaningful metric to use.

We also notice that in three networks (WIKI-VOTE, DIGG-FRIENDS and STACKOVERFLOW), where CCI records the highest precision, ECCI is, however, worse than the baseline methods. This suggests that sometimes the information provided by the extra neighbours without considering direction inclination conflicts with that provided by the source/target closure coefficients. Finding a method that better combines the information of common neighbours and closure coefficients is an interesting avenue for future study.

3.4.3 Case study in a weighted signed network

In this experiment, through a case study on the dataset of Bitcoin Alpha trust network (BTC-ALPHA) [127], we illustrate how the proposed weighted signed directed closure coefficient and the four patterns can serve as features in network analysis.

BTC-ALPHA is a trust network on a blockchain asset trading platform, where users rate other traders in a range of $[-10, 10]$ in steps of 1, from total distrust to total trust. This is a weighted signed directed network. A rating is a weighted edge from the rater (the source node) to the ratee (the target node).

First, without considering weights on edges, we conduct a correlation analysis of the directed closure coefficient with the node degree (left figure in Figure 3.7). We find that the directed closure coefficient is positively related to node degree ($\rho = 0.714$), implying big traders (who trade with a large number of people) tend to form more trustful cliques. However, when we put back the weights, the Pearson correlation score reduces to 0.265 (right figure in Figure 3.7). Big traders are not significantly better in forming trustful cliques. At the same time, we detect some nodes with negative closure coefficients, meaning the negative triangles outweigh the positive ones around them. From the balance theory [89], we know that negative triangles are rare in a trust relationship. Indeed, the percentage of such nodes is about 3.6%.

With a closer look at the nodes whose closure coefficients are negative (left figure in Figure 3.8), we find that these nodes have relatively small strength, and the absolute value of closure coefficient is not large. It implies that distrusted cliques are only formed around small traders, and the rated degree of distrust is not high. This phenomenon aligns with the intuition that an untrustworthy trader cannot build a large trading network. Interestingly, a detailed inspection of the nodes with

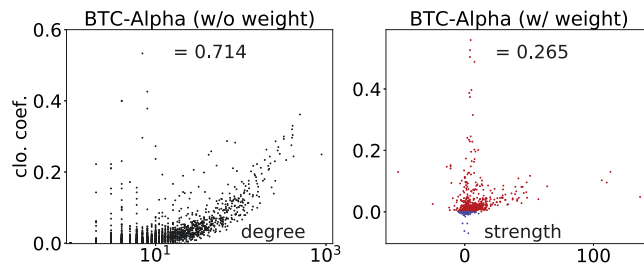


Figure 3.7 : Two scatter plots of the network BTC-ALPHA. Left one shows the correlation between directed closure coefficient and node degree (weights ignored); right one shows the correlation between weighted signed directed closure coefficient and node strength (weights taken into account).

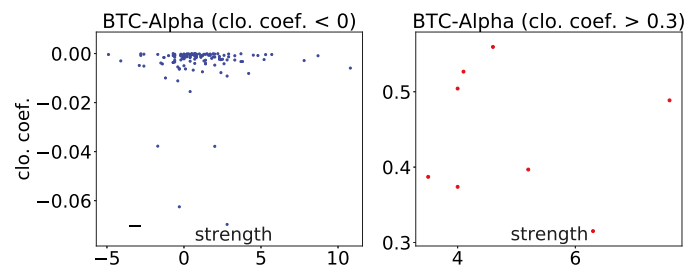


Figure 3.8 : Two local enlarged scatter plots between weighted signed directed closure coefficient and node strength in the network BTC-ALPHA. Left one is for the nodes with a negative closure coefficient; right one is for the nodes with a closure coefficient greater than 0.3.

a large closure coefficient (greater than 0.3) shows that only nodes with a small strength form highly trusted cliques (right figure in Figure 3.8), i.e., a high-trust group is often a relatively small group. This is because a trader with large strength could be involved in too many trading relationships to build a high-trust network around him/her.

Last, after calculating the four average normalized closure coefficients according to Equation 3.10, we find that the dominant pattern of nodes with a negative closure coefficient is the end-of-path pattern (Table 3.3). This implies that these traders act more as the ones being rated in the formed cliques. In contrast, the dominant pattern of nodes with a large closure coefficient (greater than 0.3) is the head-of-path pattern, implying these traders are more active in the assessment.

Table 3.3 : Comparison of head-of-path closure coefficient (HoP), mid-of-path closure coefficient (MoP), end-of-path closure coefficient (EoP) and cyclic closure coefficient (CYC) on three node sets of the dataset BTC-ALPHA. The value of the dominant pattern (> 0.30) in each group is put in bold type.

Network (selected nodes)	HoP	MoP	EoP	CYC
BTC-ALPHA (all nodes)	0.233	0.248	0.288	0.230
BTC-ALPHA ($C_e^D < 0$)	0.202	0.246	0.320	0.233
BTC-ALPHA ($C_e^D > 0.3$)	0.345	0.251	0.187	0.208

3.5 Additional Related Work Discussion

Yin et al.[242] proposed a family of eight metrics as the extension of the local closure coefficient in directed networks. The key differences of our work are that 1) we give a general definition of the local directed closure coefficient so that it can be easily used as a metric to describe networks; 2) we propose the source closure coefficient and the target closure coefficient, based on which two indices of directed closeness are introduced to improve the performance on link prediction; and 3) we extend it into weighted networks and weighted signed networks as well.

Fagiolo[63] also introduced four patterns when he proposed the directed clustering coefficient. Similarly, Ahnert and Fink[5] proposed clustering signatures to classify directed networks. The four patterns of the directed closure coefficient we propose here are different in that our definitions are end-node based and therefore asymmetric in nature.

3.6 Conclusion

In this chapter, we introduce the directed closure coefficient and its extension as another measure of edge clustering in complex directed networks. To better use it, we further propose the source/target closure coefficients and the closure coefficients of four patterns.

Through experiments on 12 real-world networks, we show that the proposed metric not only provides complementary information to the classic directed clustering coefficient but also helps to make some interesting discoveries in network analysis. Furthermore, we demonstrate that including closure coefficients in link prediction leads to significant improvement in most directed networks. We anticipate that the directed closure coefficient can be used as a descriptive feature as well as in other network analysis tasks.

This work fulfills research objectives 2 and 5.

Chapter 4

Measuring The Formation of Quadrangles

4.1 Introduction

Complex systems across various domains, such as biology, ecology, physics and social science, can be modelled as networks that abstract the interactions between system's components [16, 171, 170]. Different from a simple grid graph or a line graph for image or text modelling respectively, the complexity of networks comes from their intricate topological structures. Therefore, the study of network structure, especially local structure, underlies a number of representative and analytical applications such as representation learning of graphs [86, 80], node-type classification [20, 115], link prediction [73, 120] and anomaly detection [176, 6].

One fundamental and classic statistical metric to assess the local structure of complex networks is the *local clustering coefficient* [232, 63]. It is defined as the percentage of the number of triangles formed with a focal node to the number of triangles that the focal node could form with all its neighbours. Note that the focal node here serves as the centre node in an open triad (the middle of a length-2 path). Since many of the real-world networks are triangle-rich, the clustering coefficient — a measure of triangle formation — has become a standard metric to describe networks. It has also been used in numerous applications such as malware detection [131], language learning [78] and structural role discovery [90].

A recent study has proposed another interesting measure of triangle formation, i.e., the *local closure coefficient* [241]. With the focal node as the end node of an open triad (the head of a length-2 path), it is quantified as the percentage of twice the

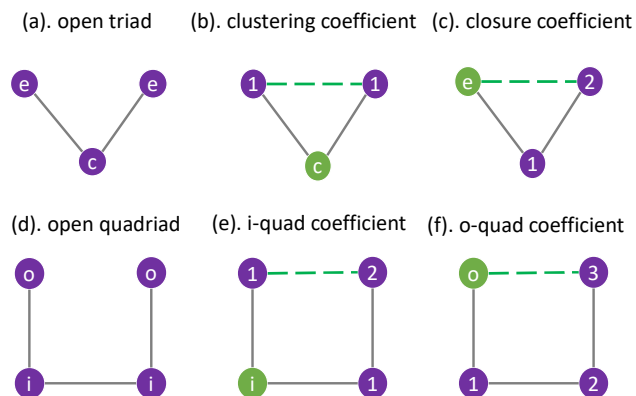


Figure 4.1 : The i-quad coefficient and the o-quad coefficient in comparison with the clustering coefficient and the closure coefficient. Letters c , e , i and o denote centre node, end node, inner node and outer node respectively. Node in green colour is the focal node in each subfigure. Number on node indicates the node’s distance from the focal node in the open triad or the open quadriad, which might be closed by an edge in dotted green line style.

number of triangles containing the focal node to the number of all length-2 paths starting from the focal node. Specifically, the classic local clustering coefficient measures the extent to which the 1-hop neighbours of a given node connect to each other, while the local closure coefficient measures the extent to which the 2-hop neighbours of a given node connect to the given node itself. This new metric has been proven to be a useful feature in network analysis tasks such as community detection and link prediction [241].

In many types of networks, however, quadrangles appear at a much higher frequency than triangles, and thus become the most dominant motifs [163]. For instance, in gene regulatory networks, logical circuits networks and neuron networks, the over-represented ”bi-fan” structure (a specific directed quadrangle) serves to carry information or signals from previous units to following ones; while in food webs, the highly recurring ”bi-parallel” structure (another type of directed quadrangle) describes how energy flows in an ecosystem.

In order to better describe and analyse the local structure of networks, we propose

two metrics quantifying the formation of quadrangles, i.e., the *i-quad coefficient* and the *o-quad coefficient*. There are two definitions in that two categories of nodes — the inner node or the outer node — can be distinguished from the node’s position in an open quadriad (also called intransitive quadriad in some works [192]). The i-quad coefficient, with the focal node functioning as the inner node of an open quadriad, measures the extent to which the focal node’s 2-hop neighbours connect to its 1-hop neighbours. The o-quad coefficient, having the focal node as the outer node of an open quadriad, measures the extent to which the focal node’s 3-hop neighbours connect to itself (Figure 4.1).

Although the focus in this chapter lies on the general unipartite networks, the proposed i-quad and o-quad coefficients provide interesting insights into bipartite networks as well. Suppose that in a recommender network where node type x denotes users and node type y denotes movies, an edge between x_i and y_i represents user x_i likes movie y_i . Take the i-quad coefficient for instance (Figure 4.2a), given x_1 , the focal user, likes movies y_1 and y_2 , while x_2 likes y_1 , it measures whether x_2 likes y_2 . In other words, the i-quad coefficient gives the extent to which other users have a similar preference as the focal user. Likewise, for the o-quad coefficient, given x_2 likes y_1 and y_2 , while x_1 , the focal node, likes y_1 , it measures whether x_1 likes y_2 (Figure 4.2b). That is to say, the o-quad coefficient gives the extent to which the focal user shares a similar opinion with other users. Interestingly, this explanation coincides with the idea of collaborative filtering [77, 215].

In addition to the basic network structure, a deeper understanding of complex systems sometimes requires taking into account the intensity or the strength of interactions between components. This is achieved by assigning weights to links. For instance, in unipartite networks, weighted links are used to represent the frequency of contact in a communication network, or the intensity of the traffic flow in a transportation network; in bipartite networks, especially recommender networks,

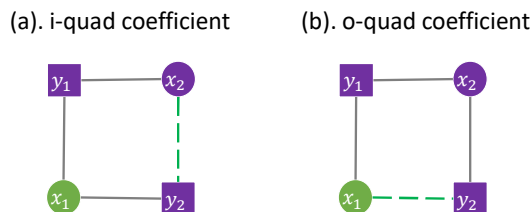


Figure 4.2 : An example of the i-quad coefficient and the o-quad coefficient in a movie recommender network. Circle nodes represent users, and square nodes represent movies. Node x_1 , marked in green, is the focal node. Four nodes and three solid links form an open quadriad, which if is closed by a dotted link will form a quadrangle.

weights are added to indicate how much a person likes a product or how often he or she purchases it. Accordingly, we introduce the *weighted i-quad coefficient* and the *weighted o-quad coefficient* in order to unveil the quadrangle formation in real weighted networks.

Our empirical study on 16 real-world networks from six domains has revealed several basic and interesting properties of the two proposed coefficients. First, we find that in most types of networks, the average o-quad coefficient is smaller than the average i-quad coefficient, which is also demonstrated through their cumulative density distributions. Secondly, we show that the o-quad coefficient has a strong positive correlation with node degree, whereas the correlation between the i-quad coefficient and node degree is very weak. We then provide a theoretical justification of this phenomenon under the configuration model.

Last but not least, we illustrate how the proposed quadrangle coefficients can be powerful features for network analysis and inference tasks. In a network classification task, we show that different types of real-world networks are significantly better clustered by adding the two quadrangle coefficients. Furthermore, in a link prediction task, we also show that the i-quad and o-quad coefficients can be used as effective predictors to improve the performance, especially in food webs, protein-protein interaction networks and infrastructure networks.

To sum up, the main contributions of this chapter are as follows:

- In order to measure the formation of 4-cycles in complex networks, we propose the i-quad coefficient and the o-quad coefficient, based on the inner node and the outer node of an open quadriad, respectively;
- We further extend two quadrangle coefficients to weighted networks;
- Empirical studies reveal that the average o-quad coefficient is smaller than the average i-quad coefficient in most types of networks;
- Theoretically, we prove that the o-quad coefficient tends to increase with node degree while the i-quad coefficient does not change too much as the node degree increases.
- Extensive experiments demonstrate that including the two coefficients leads to significant improvement in both network-level and node-level analysis tasks.

This work attains research objectives 3 and 5.

The remainder of this chapter is organised as follows. Section 4.2 introduces notations and background knowledge of clustering coefficient and closure coefficient. Section 4.3 presents and exemplifies the proposed quadrangle coefficients, whereas Section 4.4 provides details of the evaluation, including the datasets, experiment setups, performance measures, experiment results and our findings. Section 4.5 briefly contemplates the related works, and finally we conclude this chapter in Section 4.6.

4.2 Background and Motivating Example

This section first introduces the approaches that measure the formation of triangles. Then, we illustrate how these coefficients are calculated in the case of a small-scale network that serves as an example.

4.2.1 Measuring Triangle Formation

As introduced in Section 3.2.1, the local clustering coefficient is proposed to measure the cliquishness of a node's neighbours, or more specifically, the degree to which the neighbours of a node connects to each other. Barabási[16] gives the following equation to calculate the local clustering coefficient of a given node i : $C(i) = \frac{L(i)}{\frac{1}{2}d_i(d_i-1)}$, where $L(i)$ represents the number of links between i 's neighbours.

When we examine this definition from the perspective of subgraph formation, the denominator is in fact the number of open-triads, the numerator is the number of triangles, and the ratio between them is the percentage of triads that form triangles. Therefore, we give the following equation for defining the local clustering coefficient:

$$C(i) = \frac{T(i)}{OTC(i)} = \frac{\frac{1}{2} \sum_{j \in N(i)} |N(i) \cap N(j)|}{\frac{1}{2}d_i(d_i-1)}, \quad (4.1)$$

which gives the fraction of open triads, that actually form triangles. Also notice that here the focal node serves as the centre node of an open triad.

At the network-level, the *average clustering coefficient* is then defined as the mean of the local clustering coefficient over all nodes:

$$\bar{C} = \frac{1}{|V|} \sum_{i \in V} C(i). \quad (4.2)$$

Then recently the local closure coefficient is proposed to measure the edge clustering phenomenon from a novel perspective, by having the focal node at the end of an open-triad. As introduced in Section 3.2.3, the local closure coefficient of node i is defined as twice the number of triangles formed with i , divided by the number of open triads with i as the end node. (denoted $OTE(i)$):

$$E(i) = \frac{2T(i)}{OTE(i)} = \frac{\sum_{j \in N(i)} |N(i) \cap N(j)|}{\sum_{j \in N(i)} (d_j - 1)}. \quad (4.3)$$

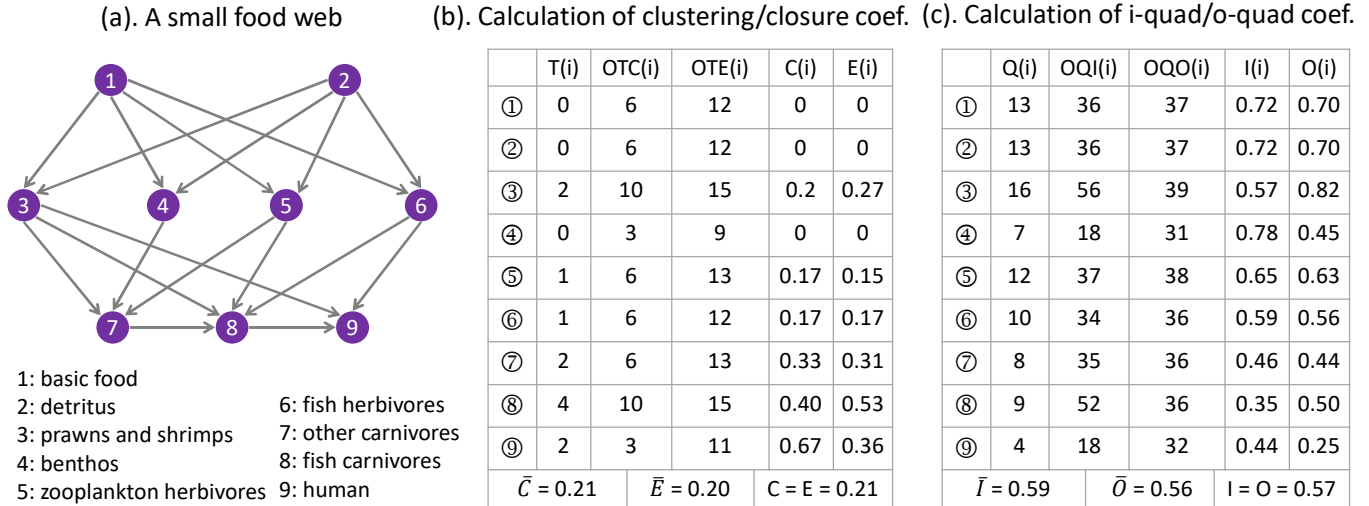


Figure 4.3 : A motivating example.

In other words, it is the fraction of open triads, where the focal node serves as the end node, that actually form triangles. $T(i)$ is multiplied by two because each triangle contains two open triads with i as the end node.

And at the network-level, the *average closure coefficient* is then defined as the mean of the local closure coefficient over all nodes (an undefined local closure coefficient is treated as zero):

$$\bar{E} = \frac{1}{|V|} \sum_{i \in V} E(i). \quad (4.4)$$

4.2.2 A motivating example

We illustrate how the two coefficients of triangle formation are calculated via a small yet real network. Figure 4.3a shows a simplified food web of the backwaters of Kerala, India [191]. It is composed of 9 nodes and 18 edges. Each node represents a species and each edge represents the flow of food energy from one species to another.

Figure 4.3b gives a detailed table of the number of triangles $T(i)$, the number of centre-node-based open triads $OTC(i)$, the number of end-node-based open triads $OTE(i)$, the local clustering coefficient $C(i)$ and the local closure coefficient $E(i)$

for each node. Also, the last row gives the average clustering coefficient, the average closure coefficient and the global clustering/closure coefficient, all of which are around 0.20.

Different from some triangle-rich networks, we find many more quadrangles than triangles in this food web (23 versus 4), which motivates us to propose measuring quadrangle formation instead. In the next section, new measures to quantify information about quadrangles in complex networks are proposed, and we show how we can leverage the fact that some networks are quadrangle and not triangle rich.

4.3 Two Quadrangle Coefficients

The clustering coefficient and the closure coefficient provide us two ways of measuring triangle formation. In some networks however, we care more about the formation of quadrangles. Also, triangles do not exist in bipartite networks and the most basic enclosed structure in this representation of networks is quadrangle. In this section, we first propose two coefficients measuring quadrangle formation, based on two different positions of the focal node in an open quadriad. Then, we further extend them to weighted networks.

4.3.1 I-quad coefficient

Recall that an open quadriad is a directionless length-3 path (Figure 4.1d). In an open quadriad $ijkl$, for instance, where three edges exist between node pairs (i, j) , (j, k) and (k, l) , we name nodes j and k as inner nodes. In contrast, nodes i and l are outer nodes. Obviously, an inner node has a degree of two, and an outer node has a degree of one. Further, an open quadriad with the focal node acting as the inner node is called inner-node-based open quadriad of that node; an open quadriad with the focal node acting as the outer node is named outer-node-based open quadriad of that node.

Conforming with the definition of the classic clustering coefficient which measures whether the two endpoints of a centre-node-based open triad are connected by a closing edge, we propose the *i*-quad coefficient that measures whether the two endpoints of an inner-node-based open quadriad are connected by a closing edge. It is quantified as the fraction of inner-node-based open quadriads that actually form quadrangles. Concretely, the ***i*-quad coefficient** of node i , denoted $I(i)$, is defined as twice the number of quadrangles formed with i (denoted as $Q(i)$), divided by the number of open quadriads with i as the inner node (denoted as $OQI(i)$):

$$\begin{aligned} I(i) &= \frac{2Q(i)}{OQI(i)} \\ &= \frac{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} |N(k) \cap N(i) - j|}{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} |N(i) - j - k|}. \end{aligned} \quad (4.5)$$

In the above equation, j is in i 's neighbour set, and k is in j 's neighbour set excluding i . $Q(i)$ is multiplied by two because each quadrangle can be viewed as constructed from two open quadriads with i as the inner node. By definition, it is obvious that $I(i) \in [0, 1]$.

Then, we define the **average *i*-quad coefficient** at the network-level, as the mean of the *i*-quad coefficient over all nodes (undefined ones are treated as zeros):

$$\bar{I} = \frac{1}{|V|} \sum_{i \in V} I(i). \quad (4.6)$$

In the case of a random network where each pair of nodes is connected with a probability p , the expected value of the average *i*-quad coefficient is also p , i.e., $\mathbb{E}[\bar{I}] = p$.

An alternative way of measuring quadrangle formation at the network-level is the **global *i*-quad coefficient**, which is defined as the fraction of inner-node-based

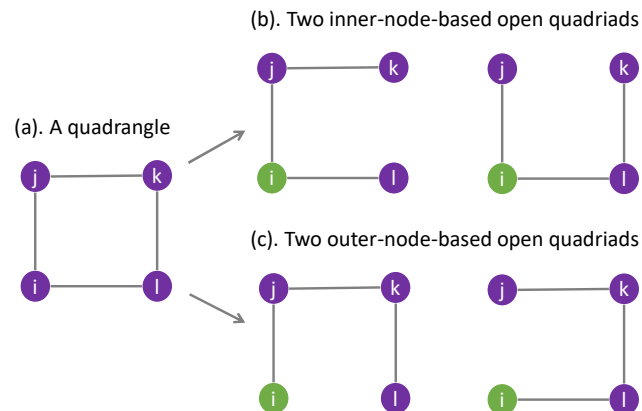


Figure 4.4 : Two types of open quadriads in a quadrangle. Node i , depicted in green, is the focal node, among four nodes i , j , k and l .

open quadriads that form quadrangles in the entire network:

$$I = \frac{\sum_{i \in V} \sum_{j \in N(i)} \sum_{k \in (N(j) - i)} |N(k) \cap N(i) - j|}{\sum_{i \in V} \sum_{j \in N(i)} \sum_{k \in (N(j) - i)} |N(i) - j - k|}. \quad (4.7)$$

The numerator of the above equation can be viewed as eight times the number of quadrangles in the entire network (each node of a quadrangle contributes two counts), then divided by twice the number of open quadriads with each node acting as the inner node.

Although both the average i-quad coefficient and the global i-quad coefficient can be used as metrics to describe quadrangle formation in the entire network, they are calculated differently. The average i-quad coefficient adds up the i-quad coefficient of every node then divides it by the number of nodes, giving each node equal weight. In contrast, the global i-quad coefficient gives nodes that form numerous quadrangles more weight, by first totalling the numerator of the i-quad coefficient then dividing it by the sum of the denominator of the i-quad coefficient.

4.3.2 O-quad coefficient

Inspired by the closure coefficient in measuring triangle formation, we move the focal node from the inner node to the outer node of an open quadriad, thus proposing the o-quad coefficient in order to measure the formation of quadrangle from a different perspective.

The significance of introducing the o-quad coefficient is twofold. First, the o-quad coefficient takes into account length-3 paths emanating from the focal node, and therefore has a larger scope of the network structure. Second, when a quadrangle is formed, the closing edge (the edge that closes the outer-node-based open quadriad) is incident to the focal node. This leads to some special properties, comparing to the i-quad coefficient where the closing edge is not incident to the focal node. We show in Section 4.4 that the cumulative distribution curve of the o-quad coefficient is above that of the i-quad coefficient, and that the o-quad coefficient tends to increase with node degree.

In a similar way, the *o-quad coefficient* of node i , denoted as $O(i)$, is defined as the fraction of open quadriads with i as the outer node that are closed:

$$\begin{aligned} O(i) &= \frac{2Q(i)}{OQO(i)} \\ &= \frac{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} |N(k) \cap N(i) - j|}{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} |N(k) - j - i|}, \end{aligned} \quad (4.8)$$

where $OQO(i)$ is the number of outer-node-based open quadriads of node i , and $Q(i)$ is the number of quadrangles containing i . $Q(i)$ is multiplied by two because each quadrangle contains two open quadriads with i as the outer node. In a quadrangle, the focal node can serve as the inner node in two open quadriads or as the outer node in another two open quadriads (Figure 4.4). Obviously, $O(i) \in [0, 1]$.

In order to measure at the network level, the *average o-quad coefficient*

is defined by averaging the o-quad coefficient over all nodes (an undefined o-quad coefficient is treated as zero):

$$\bar{O} = \frac{1}{|V|} \sum_{i \in V} O(i). \quad (4.9)$$

Analogous to the global i-quad coefficient, the *global o-quad coefficient* can be defined as the fraction of outer-node-based open quadriads that form quadrangles in the entire network:

$$O = \frac{\sum_{i \in V} \sum_{j \in N(i)} \sum_{k \in (N(j)-i)} |N(k) \cap N(i) - j|}{\sum_{i \in V} \sum_{j \in N(i)} \sum_{k \in (N(j)-i)} |N(k) - j - i|}. \quad (4.10)$$

As the equivalence between the global clustering coefficient and the global closure coefficient, this definition of global o-quad coefficient is actually not different from the global i-quad coefficient (Equation 4.7) since globally the difference of the position of the focal node will not arise.

Revisiting the motivating example, Figure 4.3c gives a detailed table of the number of quadrangles $Q(i)$, the number of inner-node-based open quadriads $OQI(i)$ and the number of outer-node-based open quadriads $OQO(i)$ of each node, based on which the i-quad coefficient $I(i)$ and the o-quad coefficient $O(i)$ are calculated. Also, the last row of this table gives the three network-level measures, i.e., the average i-quad coefficient, the average o-quad coefficient and the global i-quad/o-quad coefficient, which are more than 2.5 times larger than those metrics measuring triangles formation.

4.3.3 Quadrangle coefficients in weighted networks

Until now, the discussion has been focused on binary networks, where the value of each link is either one or zero. In many networks, however, we need a more accurate representation of the relationships between nodes, such as the frequency of

contact in a communication network, or the rating of a product given by a consumer in a recommender network, etc. This kind of information is usually expressed as a strength of the relationship and we use weighted networks to represent it. Therefore, we are interested in extending the two quadrangle coefficients to networks that allow for weights of the relationships.

Several versions of weighted clustering coefficient have been proposed in order to measure triangle formation in weighted networks [17, 177, 244, 207]. For example, Onnela et al. [177] proposed to sum over the geometric averages of the three weights in formed triangles, divided by the number of potential triangles. Alternatively, Zhang and Horvath. [244] chose to sum simply over the products of the three weights in formed triangles, divided by the total of products of the two weights of all open triads, implying the triadic closing edges taking the maximum weight.

Adopting a strategy similar to the one proposed by Zhang and Horvath [244], we introduce the weighted i-quad coefficient and the weighted o-quad coefficient to measure quadrangles formation in weighted networks. Let $G^{\mathcal{W}} = (V, E)$ be a weighted graph without self-loops and multiple edges. The weight of a link between any node i and j is denoted w_{ij} ($w_{ij} \in [0, 1]$ after normalisation by the maximum weight). For any node $i \in V$, the **weighted i-quad coefficient**, denoted as $I^{\mathcal{W}}(i)$, and the **weighted o-quad coefficient**, denoted as $O^{\mathcal{W}}(i)$, are defined as:

$$I^{\mathcal{W}}(i) = \frac{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} \sum_{l \in (N(i) \cap N(k)-j)} w_{ij} w_{jk} w_{il} w_{lk}}{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} \sum_{l \in (N(i)-j-k)} w_{ij} w_{jk} w_{il}}, \quad (4.11)$$

$$O^{\mathcal{W}}(i) = \frac{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} \sum_{l \in (N(i) \cap N(k)-j)} w_{ij} w_{jk} w_{il} w_{lk}}{\sum_{j \in N(i)} \sum_{k \in (N(j)-i)} \sum_{l \in (N(k)-j-i)} w_{ij} w_{jk} w_{kl}}. \quad (4.12)$$

When the graph becomes binary (unweighted), i.e., $w_{ij} = 1$, the above two

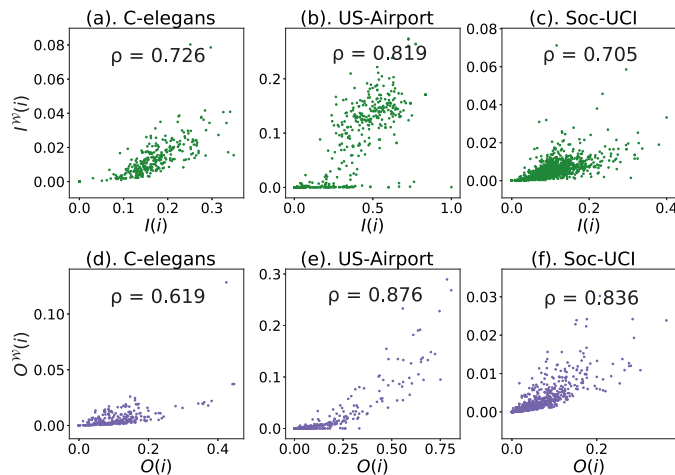


Figure 4.5 : Correlation of quadrangle coefficients and weighted quadrangle coefficients in three different networks. First row is the correlation of i-quad coefficient $I(i)$ and weighted i-quad coefficient $I^{\mathcal{W}}(i)$, second row is the correlation of o-quad coefficient $O(i)$ and weighted o-quad coefficient $O^{\mathcal{W}}(i)$. The weighted networks are: (1) C-elegans, the neural network of the Caenorhabditis elegans worm [232]; (2) US-Airport, the network of the 500 busiest commercial airports in the United States [41]; (3) Soc-UCI, the social network of online community for students at University of California, Irvine [179].

weighted quadrangle coefficients degrade to their unweighted versions (Equation 4.5 and Equation 4.8). The average weighted i-quad coefficient and the average weighted o-quad coefficient are then defined respectively as: $\overline{I^{\mathcal{W}}} = \frac{1}{|V|} \sum_{i \in V} I^{\mathcal{W}}(i)$, $\overline{O^{\mathcal{W}}} = \frac{1}{|V|} \sum_{i \in V} O^{\mathcal{W}}(i)$.

We can see from Figure 4.5 that in different weighted networks, the correlation of i-quad coefficient and weighted i-quad coefficient (and the correlation of o-quad coefficient and weighted o-quad coefficient) is also different. In other words, when weights are considered in calculating quadrangle coefficients, the weighted i-quad coefficient and the weighted o-quad coefficient capture different information compared to their unweighted counterparts.

4.3.4 Computational cost

At the end of this section, we give a brief discussion about the computational efficiency of the above mentioned metrics. From Equation 4.5 and Equation 4.8, we can see that to compute the i-quad coefficient or the o-quad coefficient for a single node, the worst-case cost is $O((k_{max})^3)$, where k_{max} is the maximum degree of the network. Therefore, the worst-case cost for computing the two coefficients for every node in a network is $O(|V| \cdot (k_{max})^3)$, which is not cheap. Fortunately, however, since most real-world networks are scale-free and exhibit heavy-tailed degree distribution, the actual cost is far less expensive than this. For example, it takes about 22.5 seconds to compute the average i-quad coefficient on the CORA citation network which contains 23,166 nodes and 89,157 edges (test on Intel Xeon Gold 6238R @ 2.2GHz with 180GB of RAM).

4.4 Experiments and Analysis

In this section, we analyse the proposed quadrangle coefficients on different types of real-world networks and demonstrate their usage in some common applications*.

4.4.1 Quadrangle coefficients in real-world networks

Datasets. We run experiments on 16 networks of six categories (collected from Konect[128] and Snap[136]):

1. Food webs. FW-FLORIDADRY[223] and FW-LITTLE ROCK [157]: energy transfer relationships collected from the cypress wetlands of South Florida and the Little Rock Lake of Wisconsin. Nodes represent species and an edge denotes that one species feeds on another (edge direction and weight are ignored).

*Our code is available at <https://github.com/MingshanJia/explore-local-structure>.

Table 4.1 : Statistics of datasets, showing the number of nodes ($|V|$), the number of edges ($|E|$), the average degree ($\langle k \rangle$), the average clustering coefficient (\bar{C}), the average closure coefficient (\bar{E}), the average i-quad coefficient (\bar{I}) and the average o-quad coefficient (\bar{O}). In order to facilitate comparison, the last four columns give the quotient of \bar{C} and \bar{E} , the quotient of \bar{I} and \bar{O} , the quotient of \bar{I} and \bar{C} , and the quotient of \bar{O} and \bar{E} respectively. Datasets having timestamps on edge creation are superscripted by (τ).

Network	$ V $	$ E $	$\langle k \rangle$	\bar{C}	\bar{E}	\bar{I}	\bar{O}	\bar{C}/\bar{E}	\bar{I}/\bar{O}	\bar{I}/\bar{C}	\bar{O}/\bar{E}
FW-FLORIDADRY	128	2,106	32.91	0.335	0.261	0.428	0.353	1.280	1.213	1.280	1.351
FW-LITTLEROCK	183	2,452	26.80	0.323	0.208	0.550	0.339	1.553	1.622	1.704	1.631
SOC-EMAIL τ	986	16,064	32.58	0.407	0.153	0.231	0.102	2.659	2.267	0.568	0.667
SOC-CLGMSG τ	1,899	13,838	14.57	0.109	0.022	0.081	0.029	5.082	2.806	0.744	1.347
SOC-BTCALPHA τ	3,783	14,124	7.47	0.177	0.020	0.058	0.013	8.937	4.448	0.326	0.655
SOC-TWITCHFR	6,549	113K	34.41	0.222	0.029	0.109	0.034	7.557	3.202	0.493	1.163
PPI-STELZL	1,706	3,191	3.74	0.006	0.002	0.038	0.021	3.827	1.806	6.332	13.416
PPI-FIGEYS	2,239	6,432	5.75	0.040	0.005	0.082	0.043	7.321	1.908	2.064	7.918
PPI-VIDAL	3,133	6,726	4.29	0.064	0.025	0.040	0.018	2.531	2.291	0.632	0.698
PPI-INTACT	8,077	26,085	6.46	0.083	0.016	0.063	0.021	5.101	2.993	0.750	1.278
CIT-DBLP τ	12,590	49,651	7.89	0.117	0.026	0.060	0.014	4.529	4.175	0.510	0.553
CIT-CORA	23,166	89,157	7.70	0.266	0.100	0.107	0.047	2.667	2.285	0.402	0.469
RD-NEWYORK	264K	365K	2.76	0.021	0.021	0.068	0.069	1.012	0.990	3.291	3.365
RD-BAYAREA	321K	397K	2.47	0.017	0.016	0.038	0.038	1.020	0.992	2.284	2.350
QA-MATHOVFL. τ	21,688	88,956	8.20	0.094	0.005	0.031	0.004	17.956	7.305	0.333	0.817
QA-ASKUBUNTU τ	138K	262K	3.81	0.015	5e-4	0.004	5e-4	31.708	7.867	0.243	0.981

2. Social networks. EMAIL τ [184]: a temporal email network from a European research institution (a temporal edge denotes that an email is exchanged between two persons at a certain time); CLGMSG[183]: temporal online message interactions between UC Irvine college students (a temporal edge means that a message is exchanged between two students at a certain time); BTCALPHA [126]: a temporal who-trusts-whom network of users on a Bitcoin trading platform Bitcoin Alpha (edge direction and weight are ignored); TWITCHFR [202]: a network of gamers who stream in French, where nodes are the users and edges are mutual friendships between them.

3. Protein-protein interaction (PPI) networks. STELZL[213], FIGEYS[62], VIDAL[203] and INTACT[180]: four networks of interactions between proteins in Homo sapiens. Nodes represent proteins and an edge denotes the physical contact between two proteins in the cell.
4. Citation networks. DBLP[137] and CORA[216]: two academic publication citation networks. DBLP contains temporal information on edges. Nodes represent papers, and an edge means that one paper cites another paper (direction is ignored).
5. Infrastructure networks. RD-NEWYORK and RD-BAYAREA[128]: two road networks for New York City and San Francisco Bay Area. Nodes represent intersections and endpoints, and the roads connecting them are represented by edges.
6. Q&A networks. MATHOVFL. and ASKUBUNTU[184]: two temporal Q&A networks derived from Stack Exchange. Nodes represent users, and a temporal edge means that one user answers another user's question at a certain time (edge direction is ignored).

Observations. Table 4.1 lists some key statistics including the proposed coefficients of these networks. We observe that in most types of networks (except road networks), the average o-quad coefficient is smaller than the average i-quad coefficient. That is to say, for the majority of nodes in these types of networks, fewer quadrangles are built from the outer-node-based open quadriads, compared to the number of quadrangles constructed from the inner-node-based open quadriads. This phenomenon is better revealed through the cumulative distribution function (CDF) in Figure 4.6: the CDF curve of the o-quad coefficient is above that of the i-quad coefficient when the coefficient value is small (except in RD-NEWYORK).

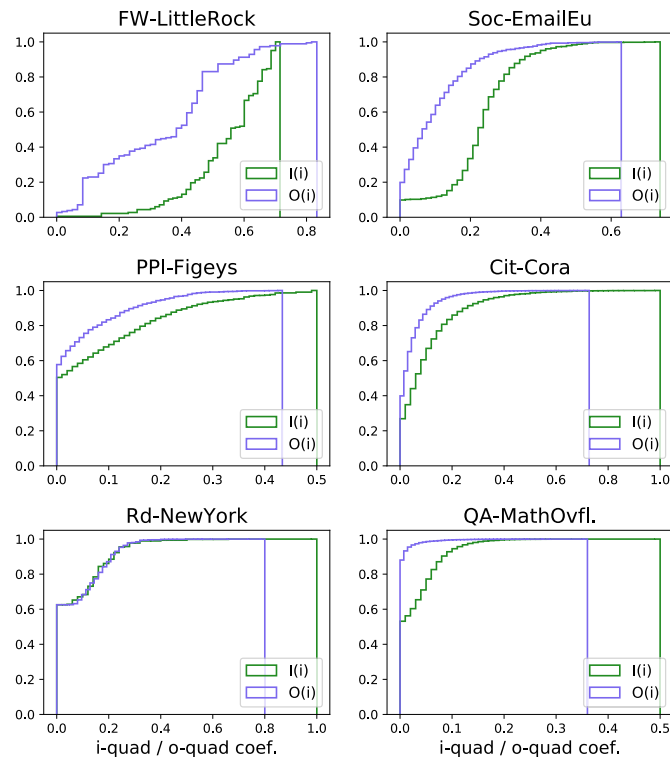


Figure 4.6 : Cumulative distribution curve of the i-quad coefficient $I(i)$ (in green colour) and the o-quad coefficient $O(i)$ (in purple colour) in six real-world networks of different types.

We can also observe that in all food webs, two PPI networks (PPI-STELZL and PPI-FIGEYS) and all road networks, the average i-quad coefficient is larger than the average clustering coefficient ($\bar{I} > \bar{C}$); and the average o-quad coefficient is larger than the average closure coefficient ($\bar{O} > \bar{E}$). In other words, these networks are more inclined to form quadrangles than to form triangles, which leads us to the following experiments.

4.4.2 Correlation with node degree

Since node degree is one of the most important and widely used concepts in network science, we study how the two quadrangle coefficients vary with it. We start by conducting an empirical analysis in real networks, followed by a theoretical justification under the degree-preserving random graph model.

We choose one network in each category and plot the correlation of quadrangle coefficients and degree (Figure 4.7). We observe a strong positive correlation between the o-quad coefficient and the node degree: the average o-quad coefficient is small among nodes with small degree and becomes larger as the average node degree increases. In contrast, the correlation between the i-quad coefficient and the degree is weak: the average i-quad coefficient is large (compared to the average o-quad coefficient) when the average node degree is small and does not change too much as the average degree increases. Since most real-world networks are scale-free and exhibit heavy-tailed degree distribution, it also explains why the average i-quad coefficient is bigger than the average o-quad coefficient in most networks studied in our work (Table 4.1).

To better understand the correlation between the quadrangle coefficients and the node degree, we give a theoretical explanation under the configuration model [68]. Constrained by a given degree sequence, the configuration model generates a network by placing edges between nodes uniformly at random. This can be achieved through a stub-matching process, in which the probability of forming an edge between node i and node j equals $d_i \cdot d_j / 2m$ (assuming $d_i^2 \leq 2m$ for all i). Now we give the following proposition.

Proposition 3. *Let V be a set of n nodes with specific degrees d_1, d_2, \dots, d_n , on which graph G is generated from the configuration model. Let $m = \frac{1}{2} \sum_{i=1}^n d_i$ denote the number of edges and $\bar{k} = (\sum_i d_i^2) / (\sum_i d_i)$ be the expected degree when a node is chosen with probability proportional to its degree. As $n \rightarrow \infty$, for any node $i \in V$, its local i -quad coefficient satisfies:*

$$\mathbb{E}[I(i)] = \frac{(\bar{k} - 1)^2}{2m},$$

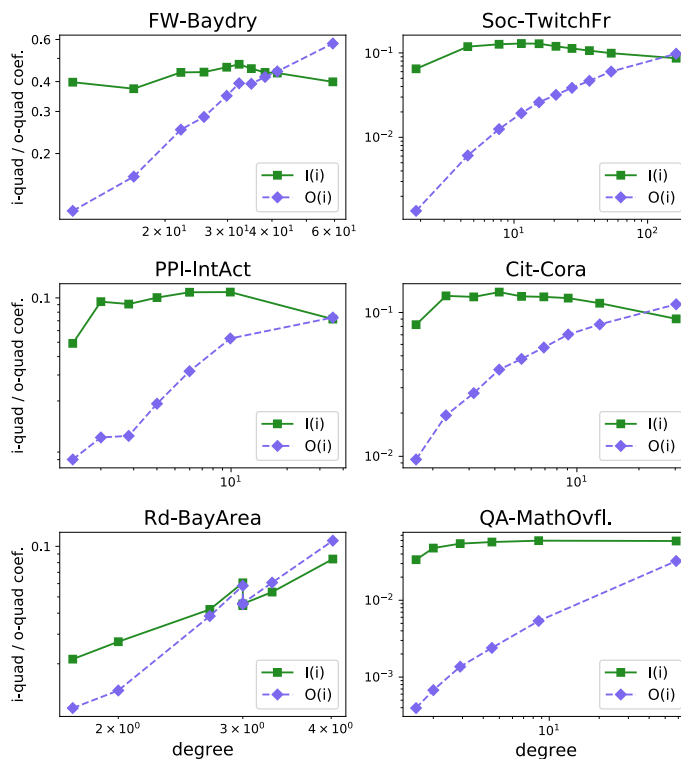


Figure 4.7 : Correlation of two quadrangle coefficients with node degree in six real-world networks. Nodes are grouped into logarithmic bins in ascending order by degree, then average i-quad and o-quad coefficients are calculated in each bin.

and its local o-quad coefficient satisfies:

$$\mathbb{E}[O(i)] = \frac{(d_i - 1) \cdot (\bar{k} - 1)}{2m}.$$

Proof. For any open quadriad with node i as an inner node, we denote one outer node by k and another outer node by l (Figure 4.8a). The probability that this open quadriad is closed equals the probability of having an edge between node k and l , which is $(d_k - 1)(d_l - 1)/2m$ in the configuration mode. The reason of subtracting 1 from d_k and d_l is that one stub of node k (and node l) has already been used in forming the open quadriad.

Now, we show that as $n \rightarrow \infty$, $\mathbb{E}[d_k] = \mathbb{E}[d_l] = \bar{k}$. Via stub matching, any node, other than node i and j , can form an edge with node j and thus become

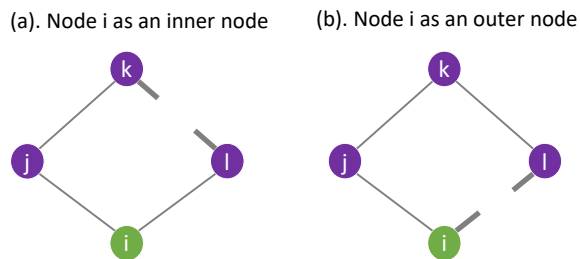


Figure 4.8 : Two types of quadrangle formation via stub matching. (a) Quadrangle is potentially formed with the focal node i acting as the inner node. The closing edge is between node k and l . (b) Quadrangle is potentially formed with the focal node i acting as the outer node. The closing edge is between node i and l .

one outer node of the open quadriad. The probability of node k being this node is proportional to its degree, which is $\frac{d_k}{\sum_{k \in V, k \neq i, j} d_k}$. Therefore, we have $\mathbb{E}[d_k] = \sum_{k \in V, k \neq i, j} d_k \cdot \frac{d_k}{\sum_{k \in V, k \neq i, j} d_k}$. When $n \rightarrow \infty$, $\mathbb{E}[d_k] = \sum_{k \in V} d_k \cdot \frac{d_k}{\sum_{k \in V} d_k} = \bar{k}$. Similarly, we have $\mathbb{E}[d_l] = \bar{k}$.

In short, we have:

$$\begin{aligned} \mathbb{E}[I(i)] &= \mathbb{E}[(d_k - 1)(d_l - 1)/(2m)] \\ &= \frac{(\mathbb{E}[d_k] - 1) \cdot (\mathbb{E}[d_l] - 1)}{2m} = \frac{(\bar{k} - 1)^2}{2m}. \end{aligned}$$

Likewise, for any open quadriad with node i as an outer node, we denote the other outer node by l (Figure 4.8b). And we have:

$$\begin{aligned} \mathbb{E}[O(i)] &= \mathbb{E}[(d_i - 1)(d_l - 1)/(2m)] \\ &= \frac{(d_i - 1) \cdot (\mathbb{E}[d_l] - 1)}{2m} = \frac{(d_i - 1) \cdot (\bar{k} - 1)}{2m}. \end{aligned}$$

□

Although Proposition 3 is given under the configuration model, we see from Figure 4.7 that this property is well preserved in most real-world networks. Only that in road networks, i.e., RD-NEWYORK and RD-BAYAREA, the average i-quad

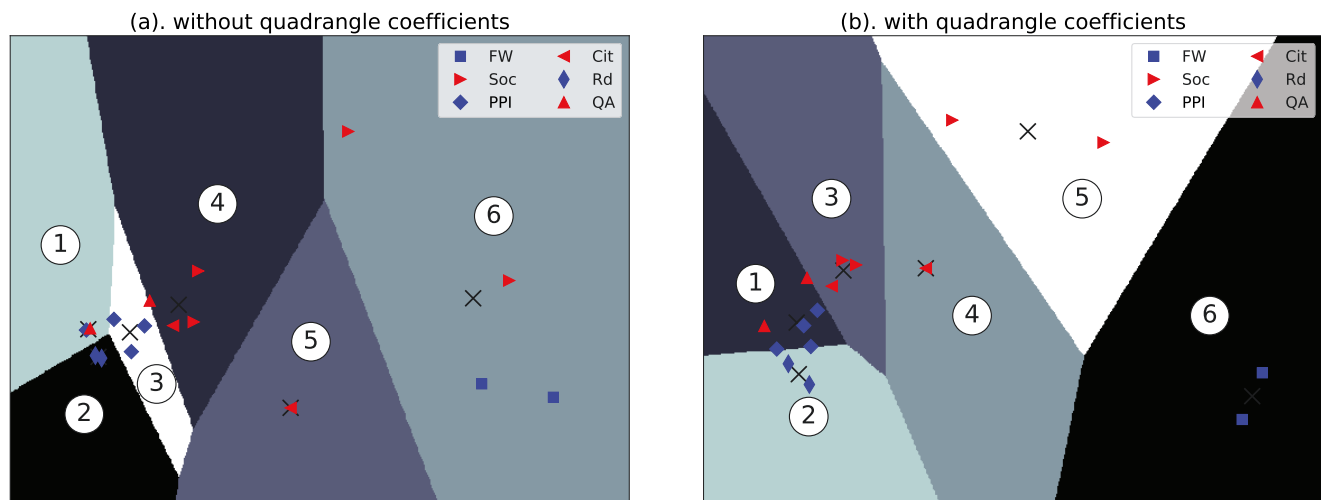


Figure 4.9 : Two-dimensional visualisation of K-means clustering on PCA-reduced data, without and with quadrangle coefficients (left figure and right figure respectively). Six clusters are labelled from 1 to 6, and painted in different colours. Centroids of clusters are marked as black crosses. Data points are plotted in different shapes and colours representing their ground truth categories, as shown in the legend.

coefficient and the average o-quad coefficient are very similar (Table 4.1), and they exhibit similar correlations with node degree. This is because the variance of node degree is extremely small (less than one) in this type of network, resulting in d_i close to \bar{k} , and thus $\mathbb{E}[O(i)]$ close to $\mathbb{E}[I(i)]$.

4.4.3 Network classification

In this section, we exhibit how useful the proposed quadrangle coefficients are in classifying different types of networks. Previous works have shown that normalized number of triads and triangles (triad significance profile[162] and clustering signatures[5]) are effective attributes in a network classification task. It motivated us to use the two quadrangle coefficients in the network classification, as they represent a normalized number of quadrangles.

We can see in Table 4.1 that the quotient of the average i-quad coefficient and the average clustering coefficient (\bar{I}/\bar{C}), and the quotient of the average o-quad

coefficient and the average closure coefficient ($\overline{O}/\overline{E}$) are contrasting in different types of networks. It is intuitive to expect the two quadrangle coefficients will be able to add useful discriminative information to a set of features, in addition to the average clustering coefficient and the average closure coefficient, for improving of the network classification accuracy.

Setup. We first prepare the data by using the three classic topological features of undirected networks, i.e., the average node degree $\langle k \rangle$, the average clustering coefficient \overline{C} and the average closure coefficient \overline{E} . We then employ a K-means clustering algorithm to partition the 16 networks into 6 clusters. The initial centroids are chosen randomly, and we repeat the algorithm with different sets of initial centroids for 1000 times, returning the best results in terms of homogeneity, completeness and V-measure score [200]. The maximum number of iterations for a single run is set to 300. To compare, we keep the experiment setting unchanged, but add the proposed quadrangle coefficients (i.e., the average i-quad coefficient \overline{I} and the average o-quad coefficient \overline{O}) to the baseline features.

Results and discussion. The classification results are given in Table 4.2. Homogeneity measures whether the samples from a single class belong to a single cluster; completeness measures whether all members of a class are assigned to the same cluster; V-measure score is the harmonic mean between homogeneity and completeness. After adding the two quadrangle coefficients, we observe significant improvement in all three measures (13% increase in homogeneity, 10% increase in completeness and 15% increase in V-measure score). It indicates that the information contained in the quadrangle coefficients are complementary to the information contained in the clustering and closure coefficients, making them discriminative features in classifying networks.

In order to further analyse the results, we adopt the Principal Component Anal-

Table 4.2 : Homogeneity (Homo.), completeness (Compl.) and V-measure score of the K-means clustering on 16 real-world networks, without and with the quadrangle coefficients (first row and second row respectively).

Features	Homo.	Compl.	V-measure
without quadrangle coefs.	0.700	0.764	0.707
with quadrangle coefs.	0.793	0.841	0.816

ysis (PCA) algorithm to compress the data to a two-dimensional space, and thus visualise the classification results (Figure 4.9). We can see from Figure 4.9(a) that the networks are poorly classified by just using three classic topological features (without the two quadrangle coefficients). Only two road networks are correctly allocated to cluster 2. Four PPI networks are separated into two clusters, resulting in a low completeness score; and two food webs are grouped together with two social networks, leading to a low homogeneity score. In contrast, when the quadrangle coefficients are included in the feature set, these networks are better clustered, especially the types of networks that are relatively rich in quadrangles (Figure 4.9(b)). Two food webs and two road networks are perfectly allocated to cluster 6 and cluster 2, respectively. In addition to that, four PPI networks are kept together within the same cluster, increasing, therefore, the completeness score. We observe, however, no obvious improvement in clustering social networks, citation networks and Q&A networks. This is because quadrangles are relatively underrepresented in these types of networks (for example, their average i -quad coefficients are less than their average clustering coefficients).

Since two more dimensions are added in the comparison, is the result statistically significant, i.e., would any added features lead to the same level of improvement? To answer this question, we conduct a significance test on V-measure score. First, we state the null hypothesis: adding two random features to the baseline feature set will achieve at least the performance of adding two quadrangle coefficients. Then we

generate two random features from a uniform distribution over 0 to 1, and append them to the baseline feature set. As previously, we employ the same algorithm and the same setup to group these networks and report the best V-measure score.

To get the distribution, we repeat the experiment 1,000 times with 1,000 different sets of randomly generated features. There are only 26 out of 1,000 sets that achieve a score higher than 0.816. Thus, we have the p-value of the null hypothesis equal to 0.026, meaning the probability of achieving such a result with random features is 0.026. As this p-value is lower than the default threshold of 0.05, the null hypothesis is confidently rejected and the statistical significance of the improvement brought by adding quadrangle coefficients is proved.

4.4.4 Link prediction

As two new metrics measuring quadrangle formation, the i-quad coefficient and the o-quad coefficient provide additional topological features for a node-level network analysis and inference. As an example, we show their utilities in missing link prediction, where significant improvement is brought by adding them.

Many studies have shown that common neighbours index and its variations such as Adamic-Adar index and resource allocation index perform well in the link prediction problem [144, 2, 254]. Besides, the clustering coefficient and the closure coefficient are proven to be useful features to improve the performance [7, 241]. Therefore, we use these five features as the baseline features in our prediction model, and then test the performance by adding the proposed i-quad and o-quad coefficients. XGBoost, the gradient boosted trees, is used as the prediction model due to its speed and performance.

Setup. We model a network as a graph $G = (V, E)$. For networks having timestamps on edges, we order the edges according to their appearing times and select the first 70% edges and related nodes to form an “old graph”, denoted $G_{old} = (V^*, E_{old})$.

The remaining 30% edges filtered by node set V^* will form a “new graph”, denoted $G_{new} = (V^*, E_{new})$. For networks not having timestamps, we randomly shuffle the edges then perform the partition, and we repeat 100 times in order to assess variance and reduce the impact of a single partition on the possible conclusions. The test set is built by node pairs, that appear in the old graph, but do not form a link. Each such pair of nodes indicates a positive or a negative example depending on whether a link between them appears in the new graph.

The training set is built on the old graph, on which we fit four XGBoost models with four sets of features: 1) baseline feature set which includes common neighbours, Adamic-Adar, resource allocation, clustering coefficient and closure coefficient; 2) baseline features plus i-quad coefficient; 3) baseline features plus o-quad coefficient; 4) baseline features plus both i-quad coefficient and o-quad coefficients. Then we evaluate their prediction performances on the test set. For large networks ($|V| > 10K$), we perform a randomised breadth first search sampling [51] of $3K$ nodes on the original graph and repeat 10 times.

Results and discussion. Since network link prediction is a highly unbalanced task, we choose the Area Under the ROC Curve (ROC-AUC) as the metric and report the prediction result on the test set, as shown in Table 4.3. First, we discover that adding the i-quad (3rd column) or the o-quad coefficient (4th column) leads to improvement in most networks. Furthermore, we find that adding the o-quad coefficient outperforms adding the i-quad coefficient in 14 out of 16 networks. One possible explanation of this phenomenon is that the o-quad coefficient looks 3-hop away from the focal node, which is in line with the recent discovery that 3-hop paths are more powerful predictors in link prediction [120, 253]. When both quadrangle coefficients are added to the baseline features (5th column), the performance is improved in all networks. The average ranking (last row) also shows that adding both i-quad and o-quad coefficients at the same time leads to the best overall performance,

Table 4.3 : Test set performance comparison measured in ROC-AUC score of four XGBoost classifiers with different features. Second column lists the scores with baseline features (BL), third column adds i-quad coefficient to baseline features, fourth column adds o-quad coefficient to baseline features, and fifth column adds both i-quad and o-quad coefficients to baseline features. An improvement of more than 2% is put in bold type, and an improvement of more than 5% is indicated by dagger. Last row gives the average (over the datasets) ranking of the four classifiers for comparison, where smaller is better. A classifier receives rank 1 if it has the highest ROC-AUC score, rank 2 if it has the second highest, and so on. If two classifiers share the best score, they both get rank 1.5, and so on. The best ranking is put in bold italic.

Network	w/ baseline features (BL)	add I(i) to BL	add O(i) to BL	add I(i) & O(i) to BL
FW-FLORIDADRY	0.6703	0.6779	0.6834	0.6886
FW-LITTLE ROCK	0.8077	0.8357	0.8421	0.8521 [†]
SOC-EMAIL EU ^τ	0.9076	0.9070	0.9090	0.9084
SOC-CLGMMSG ^τ	0.7831	0.7873	0.7879	0.7920
SOC-BTCALPHA ^τ	0.8588	0.8601	0.8679	0.8697
SOC-TWITCHFR	0.9160	0.9176	0.9192	0.9202
PPI-STELZL	0.6565	0.7778 [†]	0.7809 [†]	0.7764 [†]
PPI-FIGEYS	0.8171	0.8644 [†]	0.8668 [†]	0.8650 [†]
PPI-VIDAL	0.7566	0.7973 [†]	0.8009 [†]	0.7992 [†]
PPI-INTACT	0.8524	0.8808	0.8839	0.8842
CIT-DBLP ^τ	0.7294	0.7261	0.7336	0.7310
CIT-CORA	0.8700	0.8705	0.8726	0.8734
RD-NEWYORK	0.5268	0.5529	0.5538 [†]	0.5538 [†]
RD-BAYAERA	0.5218	0.5353	0.5353	0.5356
QA-MATHOVFL. ^τ	0.8546	0.8554	0.8541	0.8551
QA-ASKUBUNTU ^τ	0.8746	0.8791	0.8765	0.8777
Avg. ranking	3.8	2.8	1.9	<i>1.5</i>

closely followed by just adding the o-quad coefficient.

Second, we find that the improvement is particularly significant in food webs, protein-protein interaction networks and road networks (more than 2% in all eight

networks of these three types, and more than 5% in five networks when both quadrangle coefficients are added). The common characteristic of these types of networks is that they tend to have larger quadrangle coefficients compared to the clustering and closure coefficients. In other words, the extra information brought by the proposed coefficients is particularly useful in networks that are rich in quadrangles.

To give more statistical insight into these results, we adopt the non-parametric Wilcoxon Signed-Rank Test [210] to quantify the difference between classifiers with different feature sets, reporting the p-value where applicable. Note that this method is rank-based and essentially tests the null hypothesis that two paired samples come from the same distribution. In our setting, paired samples are paired columns from the result table, and rejected null hypothesis means that we would expect one approach to outperform another in a new dataset.

We find that adding the i-quad coefficient, adding the o-quad coefficient, and adding both of them to the baseline features all provide statistically significant gains over only using the baseline feature set (p-values are far less than 0.001 for all three). Moreover, the gains of adding the o-quad coefficient and adding both quadrangle coefficients to baseline features over adding the i-quad coefficient to baseline features are also critically different ($p = 0.005$, comparing adding the o-quad coefficient with adding the i-quad coefficient; $p = 0.003$ comparing adding both quadrangle coefficients with adding the i-quad coefficient). However, there is no significant difference between adding the o-quad coefficient and adding both quadrangle coefficients ($p = 0.35$). Accordingly, we create the critical difference diagram in Figure 4.10.

4.4.5 Limitations and Future Directions

Now, we describe several limitations of our work and outline how these might be overcome in future studies.

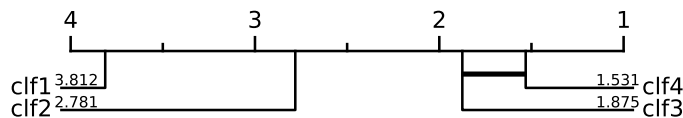


Figure 4.10 : Critical difference diagram of four classifiers with different feature sets. Classifier 1 (clf1) uses baseline features; classifier 2 (clf2) uses baseline features plus i-quad coefficient; classifier 3 (clf3) uses baseline features plus o-quad coefficient; classifier 4 (clf4) uses baseline features plus i-quad and o-quad coefficients.

Directed edges. Our work currently is limited to undirected networks (unweighted or weighted). A natural extension is to further propose the directed quadrangle coefficients in a similar approach as in extending the clustering coefficient and closure coefficient to directed networks [63, 105]. The complexity of this approach comes from the 16 different directed quadrangles. Another possible direction is to focus on one or two directed quadrangles that are proved to be more important in many types of networks, such as the bi-fan or the bi-parallel structures [249, 97].

Network dynamics. Both the i-quad coefficient and the o-quad coefficient are motivated by the view of network evolution — a closing edge appears between the two endpoints of an existing open quadriad and forms a quadrangle. Their definitions, however, do not take into consideration the dynamics of the network. An interesting future direction is to develop the notion of temporal open quadriad, meaning that an open quadriad is present at a certain timestamp while its two endpoints are not connected by a closing edge. Then we can define the temporal quadrangle coefficients as the fraction of temporal open quadriads that are closed at a later time point. With extra temporal information, these counterparts could therefore be more powerful in predicting future links.

Potential applications. Being new metrics of measuring quadrangle formation, the proposed coefficients could be promising in studying networks that are rich

in quadrangles — discovering similarities among protein-protein interaction networks [125], detecting compartments in food webs [122], and exploring how robust ecological systems are in the face of species loss [54]. More generally, the quadrangle coefficients also have the potential to be applied in community detection, as shown by the clustering and closure coefficients [241, 101]. Plus, although Graph Neural Networks have achieved state-of-the-art results in various applications, a recent study has exposed their shortcomings in capturing network structures [237]. Therefore, an interesting avenue is to incorporate the structural information brought by the proposed coefficients in the message passing scheme.

4.5 Related Work

We here recapitulate some related works that proposed other metrics to measure quadrangle formations in networks. Fronczak et al. [72] proposed a higher order clustering coefficient for random networks. It is defined as $C_i(x) = \frac{2E_i(x)}{k_i(k_i-1)}$, where i is the focal node and x is the length of path. $E_i(x)$ denotes the number of x -length paths between the neighbours of i . When x equals 2, this definition deals with the formation of quadrangles. The limitation of this definition is that the normalisation only takes the degree of the focal node i into account while neglects the degree of i 's neighbours. Since each pair of neighbours could have multiple length-2 paths between them, the clustering value can be larger than one.

Aiming to measure the formation of 4-cycles, Caldarelli et al. [33] proposed two grid coefficients, i.e., the primary grid coefficient and the secondary grid coefficient. The former is defined as: $G^p(i) = \frac{Q^p(i)}{Z^p(i)}$, where $Q^p(i)$ is the number of actual “primary quadrilaterals” containing node i , and $Z^p(i)$ is calculated by: $Z^p(i) = \frac{k_i(k_i-1)(k_i-2)(k_i-1)}{2}$. With this definition, however, it actually deals with the formation of 4-cycle with an extra diagonal edge. The secondary grid coefficient is defined as: $G^s(i) = Q^s(i)/Z^s(i)$, where $Q^s(i)$ is the number of actual “secondary

quadrilaterals” containing node i , and $Z^s(i)$ is calculated by: $Z^s(i) = \frac{k_{i,2nd}k_i(k_i-1)}{2}$. A potential problem within this definition is that it does not rule out the possibility that the 2-hop neighbour connects to two other 1-hop neighbours, making the formed structure containing five nodes.

Lind et al. [145] later proposed a square clustering coefficient in the context of bipartite networks by taking into consideration the degree of the neighbours, in other words, the length-2 paths starting from the focal node. It is defined as $C_{4,mn}(i) = \frac{q_{imn}}{(k_m - \eta_{imn})(k_n - \eta_{imn}) + q_{imn}}$, where m and n are a pair of neighbours of the focal node i , and q_{imn} denotes the number of squares containing the three nodes. What is uncommon about this definition is that it deems squares are formed via node overlapping, which is not a standard approach. Zhang et al. [248] then modified the equation and proposed another more standard square clustering coefficient for bipartite networks. Their definition is: $C_{4,mn}(i) = \frac{q_{imn}}{(k_m - \eta_{imn}) + (k_n - \eta_{imn}) + q_{imn}}$. However, in both of these definitions, there is no notion of open quadriad introduced, and the scope is limited within 2-hop distance from the focal node.

The proposed i-quad and o-quad coefficients are different from previous works in that 1) the scope of the o-quad coefficient is larger since it takes into account length-3 paths emanating from the focal node, whereas the square clustering coefficients or the grid coefficients only calculate length-2 paths in the normalisation; 2) the quadrangle coefficients proposed by us view a formed quadrangle as being built from open quadriads via connecting two endpoints with one edge, which conform with the classic clustering and closure coefficients (in their definitions a formed triangle is viewed as being built from open triads). In contrast, two edges are required to form a quadrangle in the grid coefficients; 3) the quadrangle coefficients are proposed for the general unipartite networks on which multiple experiments are conducted. In Figure 4.11, we provide a simple example to illustrate the five coefficients proposed by previous works and the two quadrangle coefficients proposed by us.

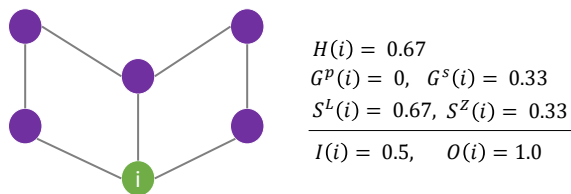


Figure 4.11 : An example of the coefficients proposed in related works, compared with our proposed quadrangle coefficients. $H(i)$ is the higher order clustering coefficient proposed by Fronczak et al.[72]; $G^p(i)$ and $G^s(i)$ are the primary grid coefficient and the secondary grid coefficient proposed by Caldarelli et al.[33]; $S^L(i)$ is the square clustering coefficient proposed by Lind et al.[145]; $S^Z(i)$ is another square clustering coefficient proposed by Zhang et al.[248]; $I(i)$ and $O(i)$ are the two quadrangle coefficients proposed by us.

4.6 Conclusion

In this chapter, we introduced the i-quad coefficient and the o-quad coefficient to measure quadrangle formation in networks, according to the different location of the focal node in an open quadriad. We also extended them to weighted networks. Through experiments on 16 real-world networks from six domains, we revealed that 1) in most types of networks, the average o-quad coefficient is smaller than the average i-quad coefficient; 2) in food webs, protein-protein interaction networks and road networks, the i-quad and o-quad coefficients are larger than the clustering and closure coefficients respectively; 3) the o-quad coefficient tends to increase with node degree while the i-quad coefficient does not change too much as the node degree increases.

We also demonstrated that including the two coefficients leads to improvement in both network-level and node-level analysis tasks, such as network classification and link prediction. The improvement is especially significant in food webs, protein-protein interaction networks and road networks in link prediction task. Additionally, we plan to further consider the dynamics of time-varying networks and link directions of directed networks when measuring quadrangle formation in the future. Due to the

simplicity and interpretability in the definitions, we anticipate that the i-quad and o-quad coefficients will become standard descriptive features and be incorporated in other network mining tasks.

This work fulfills research objectives 3 and 5.

Chapter 5

Typed-Edge Graphlets

5.1 Introduction

Underlying the formation of complex networks, topological structure has always been a primary focus in network science. Among numerous analytical approaches, graphlets [190] have gained considerable ground in a variety of domains. In biology, it is revealed that proteins performing similar biological functions have similar local structures depicted by the graphlet degree vector [161]. In social science, egocentric graphlets are used to represent the patterns of people's social interactions [220]. More broadly, the notion of graphlets is introduced in computer vision to capture the spatial structure of superpixels [247], or in neuroscience to identify structural and functional abnormalities [13].

However, the original graphlets concept is unable to capture the richer information in networks that contain different types and characteristics of nodes or edges. Specifically, there are situations in which we are more interested in edge-labelled networks. For example, in a routing network where edges represent communication links, the label of each edge indicates the cost of traffic over that edge and is used to calculate the routing strategy. Or in an egocentric social network, the different types of social relationships between the ego and the alters are essential in analysing ego's behaviour and characteristics. Some studies have extended graphlets to attributed networks (also called heterogeneous networks). Still, they either only deal with different types of nodes [201] or are not capable of encoding specific type information in graphlets [197, 81].

In this work, we introduce an approach to embedding edge type or edge attribute information in graphlets, named Typed-Edge Graphlets Degree Vector, or TyE-GDV for short. We employ both the classic graphlets degree vector [161] (GDV) and the proposed TyE-GDV to represent and analyse 303 egocentric social networks of chronic pain patients. The real-life data is collected from three chronic pain leagues in Belgium. Each patient selects up to ten connections and each edge is labelled with one social relationship type. After grouping the patients into four groups according to their self-perceived pain grades, we find that patients with higher grades of pain have more star-like structures (3-star graphlets) in their social networks, while patients in lower pain grades groups form more 3-cliques, tailed-triangles, 4-chordal-cycles and 4-cliques. With the additional edge type information provided by TyE-GDV, we further discover that the outnumbered 3-star graphlet in higher pain grade patients is mainly formed of friends or healthcare workers; and that in 3-cliques and 4-cliques, friends and colleagues appear more frequently among patients with lower pain grades.

We further apply TyE-GDV into a node classification task. The dataset contains demographic attributes, detailed information about chronic pain (duration, diagnosis, pain intensity, etc.), and other related data such as the physical functioning score, depression score, social isolation score, etc. We show that the edge-type encoded graphlet features depicted by TyE-GDV are more distinctive than the classic non-typed graphlet features given by GDV in telling apart patients of different pain grades.

To summarise, the main contributions of this chapter are as follows:

- In order to effectively encode edge type information, we propose a novel framework to generate a Typed-Edge Graphlet Degree Vector;
- We further modify the TyE-GDV framework so that it is applicable for ego-

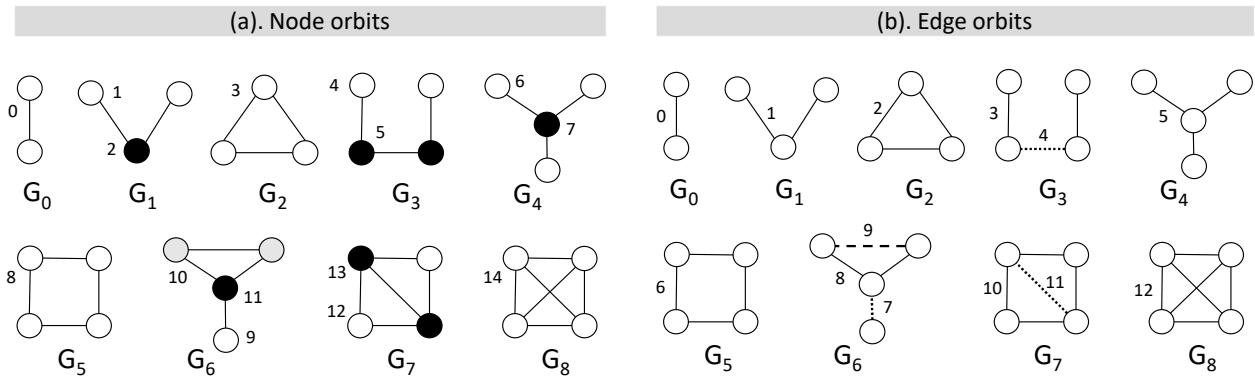


Figure 5.1 : Graphlets of size 2–4 nodes with enumeration of orbits.

centric networks;

- We extend colored graphlets and heterogeneous graphlets approaches for edge-typed networks.
- Case study on chronic pain patients shows that particular types of social relationships bear more importance in understanding the effect of chronic pain and could lead to more effective therapeutic interventions.

This work attains research objective 5.

The remainder of this chapter is organised as follows. Preliminary knowledge is provided in Section 5.2. Our proposed approach is introduced in Section 5.3. Experiments, results and analysis are presented in Section 5.5. And finally we conclude in Section 5.6 and discuss future directions.

5.2 Background and Preliminaries

In this section, we introduce the concepts of graphlets and graphlets in the context of egocentric networks.

5.2.1 Graphlets and orbits

Graphlets are small non-isomorphic induced subgraphs of a network [190]. Non-isomorphic means that two subgraphs need to be structurally different, and induced means that all edges between the nodes of a subgraph must be included. With a range of size from 2 to 5 nodes, there are 30 different graphlets in total. And, when the non-symmetry of node position is taken into consideration, there are 73 different local structures, which are also called automorphism orbits [161]. Simply put, orbits are graphlets that distinguish the position of a focal node (we use orbits and node-orbit graphlets interchangeably in this work). For any given node, a vector of the frequencies of all 73 orbits is then defined as the node's Graphlet Degree Vector (GDV). GDV or normalised GDV is often used as node feature to measure the similarities or differences among all nodes. We summarise node-orbits graphlets of size 2 to 4 nodes in Figure 5.1(a). Take G_6 for example, orbit-11 touches orbit-0 three times, orbit-2 twice, orbit-3 once and orbit-11 itself once. Thus, its GDV has 3 at the 0th coordinate, 2 at the 2nd coordinate, 1s at the 3rd and 11th coordinates, and 0s at the remaining coordinates.

The original notion of orbits is at node-level, distinguishing node position when counting graphlets. Hočevár and Demšar later propose to count graphlets at link-level and introduce the notion of edge orbits [92]. Figure 5.1(b) gives all edge orbits of size 2 to 4 nodes. Apparently, edge orbits are different from node orbits. For example, there is only one edge orbit in graphlet G_1 , but two node orbits in it. We also refer to edge orbits as edge-orbit graphlets in this work. The concept of heterogeneous graphlets is built upon edge orbits, and we will discuss more about it in Section 5.4.

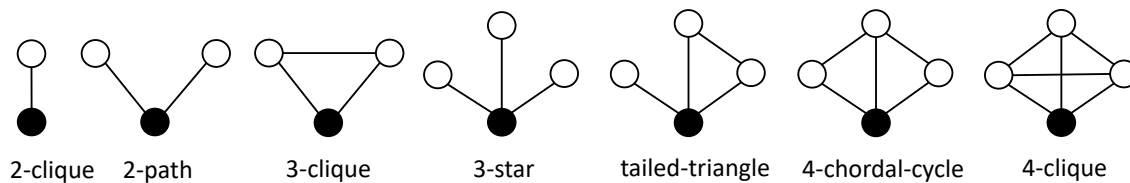


Figure 5.2 : 7 egocentric graphlets of 2 to 4 nodes. Ego node is painted in black.

5.2.2 Egocentric graphlets

In social network analysis, egocentric networks are sometimes of particular interest when we care more about the immediate environment around each individual than the entire world [189]. We may want to learn why some people behave the way they do, or why some people develop certain health problems. Since the notion of graphlets is defined at node-level, it is naturally suitable to be applied in egocentric networks, with two modifications. First, some graphlets that do not meet the requirement of being an egocentric network are excluded. For example, in graphlets of size up to 4 nodes (Figure 2.2), G_3 and G_5 are eliminated because any node in them serving as an ego cannot reach all other nodes with 1-hop. Second, there is no need to distinguish different orbits in egocentric graphlets because only one orbit can act as an ego. Therefore, there are in total 7 egocentric graphlets of size 2 to 4 nodes, which are 2-clique, 2-path, 3-clique, 3-star, tailed-triangle, 4-chordal-cycle and 4-clique (Figure 5.2).

5.3 Typed-Edge Graphlet Degree Vector

This section describes the framework for generating edge-type embedded graphlet degree vector.

The original concept of graphlets manages to capture rich connectivity patterns in homogeneous networks. However, many real-world networks are more complex by containing different types of nodes and edges, making them heterogeneous net-

Algorithm 1: Typed-Edge Graphlet Degree Vector (TyE-GDV).

input : preprocessed graph $G = \langle V, E, \mathcal{T}_e \rangle$, set of node-orbits \mathcal{O} , node set V' .
output: dictionary dic of vectors for all nodes $\in V'$.

- 1 initialise: $dic = \{\}$;
- 2 **foreach** $i \in V'$ **do**
- 3 initialise a 2d-vector vec of size $|\mathcal{O}| \times |\mathcal{T}_e|$ with zeros;
- 4 **foreach** $o \in \mathcal{O}$ **do**
- 5 $L_e = \text{GETEDGELIST}(o)$;
- 6 $\text{UPDATE}(vec, o, L_e)$
- 7 $dic[i] = vec$;

Algorithm 2: Update Vector.

- 1 **Function** UPDATE
- 2 **input** : 2d-vector vec , type of node orbit o , edge list L_e .
- 3 **foreach** $e \in L_e$ **do**
- 4 $\tau_e = \text{GETTYPE}(e)$;
- 5 /* o and τ_e are used as indices in vec . */
- 6 $vec[o][\tau_e]$ increase by 1;

works. Specifically, edge type information is crucial in that it indicates the specific relationship between the nodes. For example, in the dataset of this study, each chronic pain patient describes their egocentric social network, including up to ten actors, and each edge is labelled with 1 of 13 types of social relationships. In order to analyse edge-labelled networks at a finer granularity, we propose to embed edge-type information in graphlets. The original graphlet degree vector counts the occurrences of each type of graphlet, and as a result, a one-dimensional vector is created. Here, we propose to construct a two-dimensional vector by counting each type of edge touched by each type of graphlet.

To begin with, we give the formal definition of an edge-labelled network.

Definition 12. An edge-labelled network G is a triple $\langle V, E, \mathcal{T}_e \rangle$, where $V = \{v_1, v_2, \dots, v_n\}$ is the set of nodes, $E = \{e_{ij}\} \subset V \times V$ is the set of edges where e_{ij} indicates an edge between nodes v_i and v_j , and \mathcal{T}_e is the set of edge types, where $\tau_{e_{ij}}$ denotes the

type of edge e_{ij} .

The first step of the framework is graph preprocessing, in which the set of edge types is mapped to integers ranging from 0 to $|\mathcal{T}_e|$. For instance, the 13 types of social relationships in the targeted dataset are denoted from 0 to 12 ($\tau_e \in [0, 12]$). Also, the set of orbits \mathcal{O} is mapped to integers ranging from 0 to $|\mathcal{O}|$. In this work, we consider all possible orbits up to 4 nodes (Figure 5.1(a)). Therefore, there are 15 orbits coded from 0 to 14 ($o \in [0, 14]$).

Algorithm 3: Code Snippet for Orbit-6, 9 and 10.

```

1 foreach  $i \in V'$  do
2   initialise a 2d-vector  $vec$  of size  $|\mathcal{O}| \times |\mathcal{T}_e|$  with zeros;
3   foreach  $u \in N_i$  do
4     foreach  $v, w \in C(N_u, 2)$  do
5       if  $v \notin N_i \wedge w \notin N_i \wedge v \notin N_w$  then
6         | UPDATE( $vec, \tau_{o_6}, [e_{iu}, e_{uw}, e_{uw}]$ );           ▷ orbit-6
7       if  $v \notin N_i \wedge w \notin N_i \wedge v \in N_w$  then
8         | UPDATE( $vec, \tau_{o_9}, [e_{iu}, e_{uw}, e_{uw}, e_{vw}]$ );   ▷ orbit-9
9       if  $v \in N_i \wedge w \notin N_i \wedge v \notin N_w$  then
10        | UPDATE( $vec, \tau_{o_{10}}, [e_{iu}, e_{uv}, e_{uw}, e_{iv}]$ );
11        | UPDATE( $vec, \tau_{o_{10}}, [e_{iu}, e_{uv}, e_{uw}, e_{iv}]$ );   }
12        | UPDATE( $vec, \tau_{o_{10}}, [e_{iu}, e_{uv}, e_{uw}, e_{iw}]$ );   }           ▷ orbit-10
13
```

Algorithm 1 shows the approach of generating a two-dimensional vector of size $|\mathcal{O}| \times |\mathcal{T}_e|$, i.e., the Typed-Edge Graphlet Degree Vector (TyE-GDV) for any nodes of interest. Specifically, after initialisation, for each node in a given node set V' and for each type of the 15 node-orbit graphlets, the vector is updated through the UPDATE function (Algorithm 2). The calculation of each orbit in Algorithm 1 is omitted for a more concise expression. To demonstrate the detailed process, we give a program snippet for calculating orbit-6, 9 and 10 in Algorithm 3. $C(N_u, 2)$ denotes all possible 2-combinations of the set of neighbours of node u . The use of combinations is to avoid repetitive calculation. Due to the preprocessing step, o and τ_e are conveniently used as indices when updating the vector. At the end of

Algorithm 4: Typed-Edge Ego-Graphlet Degree Vector (TyE-EGDV).

```

input : preprocessed graph  $G = \langle V, E, \mathcal{T}_e \rangle$ , set of egocentric node-orbits  $\mathcal{O}$ ,
         node set  $V'$ .
output: dictionary  $dic$  of vectors for all nodes  $\in V'$ .
1 initialise:  $dic = \{\}$ ;
2 foreach  $i \in V'$  do
3   initialise a 2d-vector  $vec$  of size  $|\mathcal{O}| \times |\mathcal{T}_e|$  with zeros;
4   foreach  $u \in N_i$  do
5     UPDATE( $vec, o_0, e_{iu}$ ); ▷ 2-clique
6   foreach  $u, v \in C(N_i, 2)$  do
7     if  $v \notin N_u$  then ▷ 2-path
8       UPDATE( $vec, o_1, [e_{iu}, e_{iv}]$ );
9     else ▷ 3-clique
10      UPDATE( $vec, o_2, [e_{iu}, e_{iv}, e_{uv}]$ );
11   foreach  $u, v, w \in C(N_i, 3)$  do
12     if  $u \notin N_v \wedge u \notin N_w \wedge v \notin N_w$  then ▷ 3-star
13       UPDATE( $vec, o_3, [e_{iu}, e_{iv}, e_{iw}]$ );
14     else if  $v \in N_u \wedge w \notin N_u \wedge w \notin N_v$  then
15       UPDATE( $vec, o_4, [e_{iu}, e_{iv}, e_{iw}, e_{uw}]$ );
16     else if  $w \in N_u \wedge v \notin N_u \wedge v \notin N_w$  then
17       UPDATE( $vec, o_4, [e_{iu}, e_{iv}, e_{iw}, e_{uw}]$ ); ▷ tailed-tri
18     else if  $w \in N_v \wedge u \notin N_v \wedge u \notin N_w$  then
19       UPDATE( $vec, o_4, [e_{iu}, e_{iv}, e_{iw}, e_{vw}]$ );
20     else if  $u \in (N_v \cap N_w) \wedge w \notin N_v$  then
21       UPDATE( $vec, o_5, [e_{iu}, e_{iv}, e_{iw}, e_{uv}, e_{uw}]$ );
22     else if  $v \in (N_u \cap N_w) \wedge w \notin N_u$  then
23       UPDATE( $vec, o_5, [e_{iu}, e_{iv}, e_{iw}, e_{uv}, e_{vw}]$ ); ▷ 4-chord-cyc
24     else if  $w \in (N_u \cap N_v) \wedge v \notin N_u$  then
25       UPDATE( $vec, o_5, [e_{iu}, e_{iv}, e_{iw}, e_{uv}, e_{vw}]$ );
26     else ▷ 4-clique
27       UPDATE( $vec, o_6, [e_{iu}, e_{iv}, e_{iw}, e_{uv}, e_{vw}, e_{uw}]$ );
28    $dic[i] = vec$ ;

```

Algorithm 1, a dictionary of nodes as keys and their corresponding TyE-GDV as values is returned. For example, if an orbit-9 is detected and its four edges are of type ‘0’, ‘1’, ‘2’ and ‘2’, vector elements at coordinates (9, 0), (9, 1), (9, 2) and (9, 2) will increase by 1. Obviously, the time complexity of generating TyE-GDV is the same as counting graphlets.

As discussed earlier in Section 5.2.2, egocentric networks are sometimes of special interest, especially when edge type information is included (as in our case study dataset of chronic pain patients). With the restriction of being egocentric, there

are fewer orbits in graphlets that need to be considered. Therefore, we also propose a tailor-made version of the framework for egocentric networks, called TyE-EGDV (see Algorithm 4). $C(N_i, 2)$ and $C(N_i, 3)$ denotes all possible 2-combinations and 3-combinations of the set of neighbours of node i . Note that in TyE-EGDV, there are in total 7 orbits in \mathcal{O} , instead of 15. Therefore, the algorithm is more efficient in both time and space.

5.4 Typed-Edge Degree, Colored Graphlets and Heterogeneous Graphlets

Since node degree is the simplest network structural metric, a naive way of encoding edge type information in network structure is first to have the notion of typed-edge degree. Formally, the typed-edge degree of node i with edge type t , i.e., d_i^t , is defined as the number of edges of type t that are connected to i . Then, a typed-edge degree vector (TyE-DV), can be defined as a vector containing typed-edge degrees of all types.

Some other approaches that also aim to take node type and/or edge into consideration include the colored motifs [197], colored graphlets [81] and heterogeneous graphlets [201]. Colored motifs, as the name suggests, extended G-Tries algorithm that counts motifs [196] by including the information of node or edge type. This approach, however, is at network-level and is therefore not suitable for node-level analysis.

Colored graphlets approach [81] is at node-level, and proposes to distinguish different graphlets according to all combinations of node types. Although the paper claims that this approach also works with typed edges, they did not theoretically or experimentally demonstrate that. The paper alleges that the total number of combinations equals to $2^T - 1$, where T is the total number of possible node types.

This is incorrect as it fails to take graph size into consideration. We give the amended equation for calculating the number of combinations in a given graphlet g :

$$\mathcal{C}(g) = \sum_{n=1}^{\min(K(g), T)} \binom{T}{n}, \quad (5.1)$$

where $K(g)$ is the number of nodes of the graphlet when T refers to node type, and number of edges when T is edge type. We then develop a colored graphlets approach for edge-typed networks, named ColoredE-GDV, which is also applied to case studies in the next section.

The recently proposed heterogeneous graphlets approach [201] also considers node type in graphlets. It is different from the colored graphlets approach in two ways. First, heterogeneous graphlets are computed at link-level. It distinguishes the position of a given edge, instead of a given node (refer to the notion of edge-orbit graphlets in Section 5.2.1). The benefit of a link-based computation is that it is more time-efficient in sparse networks than node-based approaches. The downside, apparently, is that it is not suitable for node-level analysis. Second, heterogeneous graphlets propose to use combinations with repetitions of node types, rather than just combination, when distinguishing different graphlets. The total number of possible heterogeneous graphlets is calculated as:

$$\mathcal{H}(g) = \sum_{n=1}^T \binom{T}{n} \cdot \binom{K(g) - 1}{n - 1} = \binom{T + K(g) - 1}{K(g)}. \quad (5.2)$$

Similarly, $K(g)$ is the number of nodes of the graphlet when T refers to node type, and number of edges when T is edge type. Since repetition is allowed in heterogeneous graphlets, the number of possible heterogeneous graphlets is larger than that of colored graphlets.

In order to extend the idea of heterogeneous graphlets to node-level analysis

and to deal with typed edges, we propose a node-based typed-edge heterogeneous graphlets approach, named HeteroE-GDV^N (the original link-based typed-node approach is noted as HeteroN-GDV^L). The approach of HeteroE-GDV^N is demonstrated through Algorithm 5. We see clearly that its time complexity stays the same as counting untyped graphlets, but the space complexity grows fast with the number of edge types.

Algorithm 5: Node-based Heterogeneous Graphlets Degree Vector
(Hetero-GDV^N)

```

input : preprocessed graph  $G = \langle V, E, \mathcal{T}_e \rangle$ , set of node-orbits  $\mathcal{O}$ , node set  $V'$ .
output: dictionary dic of vectors for all nodes  $\in V'$ .
1 initialise:  $dic = \{\}$ ;
2  $L^{T_e} = [0, 1, \dots, |T_e| - 1]$ ;
   /* range of edge number of graphlets of size 2 - 4 nodes */
3 for  $k \leftarrow 1$  to 6 do
4    $L_k = [\text{GETCOMBWITHREP}(L^{T_e}, k)]$ ;
5 foreach  $i \in V'$  do
6   for  $o \leftarrow 0$  to  $|\mathcal{O}| - 1$  do
7     initialise  $vec_o$ ;
8     foreach  $o \in \mathcal{O}$  do
9        $k = \text{GETNUMOFEDGE}(o)$ ;
10       $L_e = \text{GETEDGELIST}(o)$ ;
11       $tup = (\text{SORT}(L_e))$ ;
12       $vec_o[\text{GETINDEX}(L_k, tup)]$  increase by 1;
13    $vec = [vec_0, vec_1, \dots, vec_{|\mathcal{O}|-1}]$ ;
14    $dic[i] = vec$ ;

```

Although the above approaches seem powerful to capture all possible combinations (or combinations of repetitions) of different types of nodes or edges, their numbers of possible graphlets, which are also their space complexities, grow near-exponentially with the number of node or edge types. For example, with 9 node types, in colored graphlets approach, there are 255 possible colored graphlets for a graphlet of 4 nodes; and in heterogeneous graphlets approach, there are 495 possible graphlets. In comparison, the space complexity grows linearly with the number of edge types in the proposed TyE-GDV approach. Moreover, out of this large number

of possible graphlets, only a tiny percentage of them actually exists in real networks. For example, in Cora citation network [216], only 19 heterogeneous graphlets exist out of 210 possible ones in a 4-clique graphlet.

In order to utilise the colored graphlets and heterogeneous graphlets approaches in egocentric networks, we further develop their egocentric versions, and apply them in the chronic pain case study. With fewer node orbits to consider, egocentric colored graphlets and egocentric heterogeneous graphlets are faster and more space-saving than the original ones. The implementation of these algorithms is available at <https://github.com/MingshanJia/explore-local-structure>.

To conclude this section, we summarise the time and space complexities of four main approaches in Table 5.1. Colored-GDV, HeteroE-GDV^N and TyE-GDV share the same time complexity because they are all node-based algorithms. Hetero-GDV^L as the only link-based algorithm, could be faster in sparse networks. When it comes to space complexity, the proposed TyE-GDV grows linearly with the number of edge types, while the other three methods grow near exponentially with it.

Approach	Time complexity	Space complexity
Colored-GDV [81]	$O(V \cdot k_{\max}^{S-1})$	$O(V \cdot \mathcal{O} \cdot 2^{ \mathcal{T}_e })$
Hetero-GDV ^L [201]	$O(E \cdot k_{\max}^{S-2})$	$O(E \cdot \mathcal{O}_e \cdot {}^K C_{ \mathcal{T}_e +K-1})$
HeteroE-GDV ^N	$O(V \cdot k_{\max}^{S-1})$	$O(V \cdot \mathcal{O} \cdot {}^K C_{ \mathcal{T}_e +K-1})$
TyE-GDV	$O(V \cdot k_{\max}^{S-1})$	$O(V \cdot \mathcal{O} \cdot \mathcal{T}_e)$

Table 5.1 : Time and space complexities of four approaches that deal with edge type information. S is the maximum number of nodes in graphlets, K is the maximum number of edges in graphlets, $|\mathcal{O}_e|$ is the number of edge-orbit graphlets.

5.5 Experiments and Analysis

In this section, we apply the proposed method to analyse egocentric social networks of chronic pain patients. Our code is available at <https://github.com/>

MingshanJia/explore-local-structure.

5.5.1 Dataset

The dataset is collected from chronic pain patients of the Flemish Pain League, the League for Rheumatoid Arthritis and the League for Fibromyalgia [224]. Each patient uses the graphical tool GENSI [212] to generate their egocentric social networks containing up to 10 alters. The types of relationship between the ego and the alters are explicitly given (all 13 types of social relationships are listed in Table 5.2). Participants were also asked to fill out a sociodemographic/pain questionnaire. After excluding inconsistent and incomplete entries, 303 patients' egocentric social networks and their sociodemographic/pain characteristics constitute the final dataset. The average age of all patients is 53.5 ± 12 years (248 females and 55 males).

Relationship	Type	Total number of occurs.
Partner	T-1	222
Father/Mother	T-2	209
Brother/Sister	T-3	293
Children/Grandchildren	T-4	493
Friend	T-5	506
Family-in-law	T-6	207
Other family	T-7	142
Neighbour	T-8	69
Colleague	T-9	57
Healthcare worker	T-10	233
Member of organisations	T-11	74
Acquaintance	T-12	15
Other	T-13	17

Table 5.2 : Edge type and total number of occurrences of each type in all networks.

Figure 5.3 gives some basic information about these egocentric networks, including the ego nodes' degree distribution and their edge-type distribution. The edge-type distribution is calculated by summing over all ego nodes on each type of the edges, which is also shown in the third column of Table 5.2. From the degree distribution (Figure 5.3a), we know that most patients (62%) have 10 social contacts in their networks. However, we don't expect degree being a discriminative

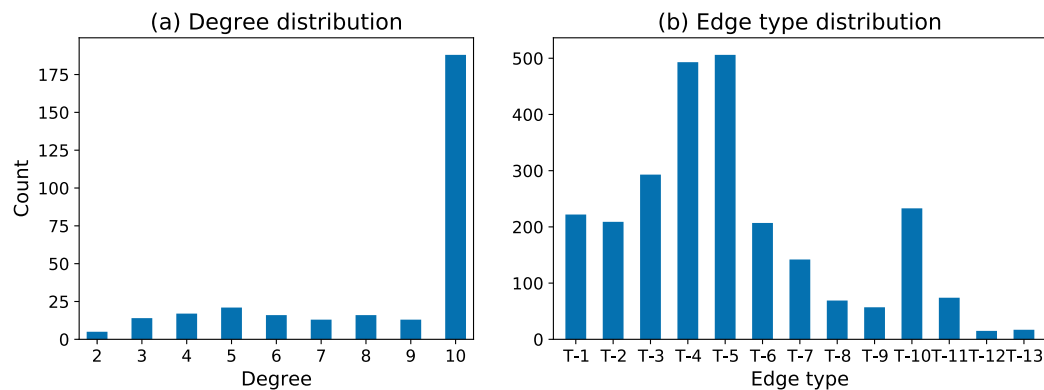


Figure 5.3 : Degree distribution and edge type distribution of all patients.

feature in the following analysis because 10 alters is the upper limit in the dataset. The edge-type distribution (Figure 5.3b) informs us that “friend” and “children” are the most frequent types appearing in these networks. In contrast, edges of types “neighbour”, “colleague” and “member of organisations” are underrepresented; “acquaintance” and “other” are almost negligible simply because if somebody is asked to name 10 contacts, they will name strongest contacts and there is no space for “acquaintance” or “other” relationships.

Furthermore, pain grades are calculated by means of the Graded Chronic Pain Scale (GCPS), which assesses both pain intensity and pain disability [225]. Patients are then classified into 5 grades based on their average intensity and disability scores: grade-0 no pain; grade-1 low intensity and low disability; grade-2 high intensity and low disability; grade-3 moderate disability regardless of pain intensity; and grade-4 high disability regardless of pain intensity. Because all participants are chronic pain patients, their GCPS grades range from grade-1 to grade-4. Specifically, we have 21 patients of grade-1, 33 patients of grade-2, 67 patients of grade-3 and 182 patients of grade-4. In this work, we aim to explore whether the graphlets and typed-edge graphlets are beneficial to recognising GCPS grades of chronic pain patients.

5.5.2 Analysing pain grades via GDV and TyE-GDV

Previous studies have revealed that social interactions play an important role in the perception of pain [111]. For example, a strong association was found between perceived social support and pain inference [65]; and improvements in social isolation lead to significant improvements in patients' emotional and physical functioning [15]. Usually, the social context of a patient is measured by means of the Patient Reported Outcome Measurement Information System (PROMIS[®])[83] or the Social Support Satisfaction Scale (ESSS) [194]. These measurements, however, are not based on patients' actual social networks and therefore cannot provide insights on the impact of network structures or specific types of interactions. To cope with this issue, we apply the classic graphlets and the proposed typed-edge graphlets to analyse patients' social networks.

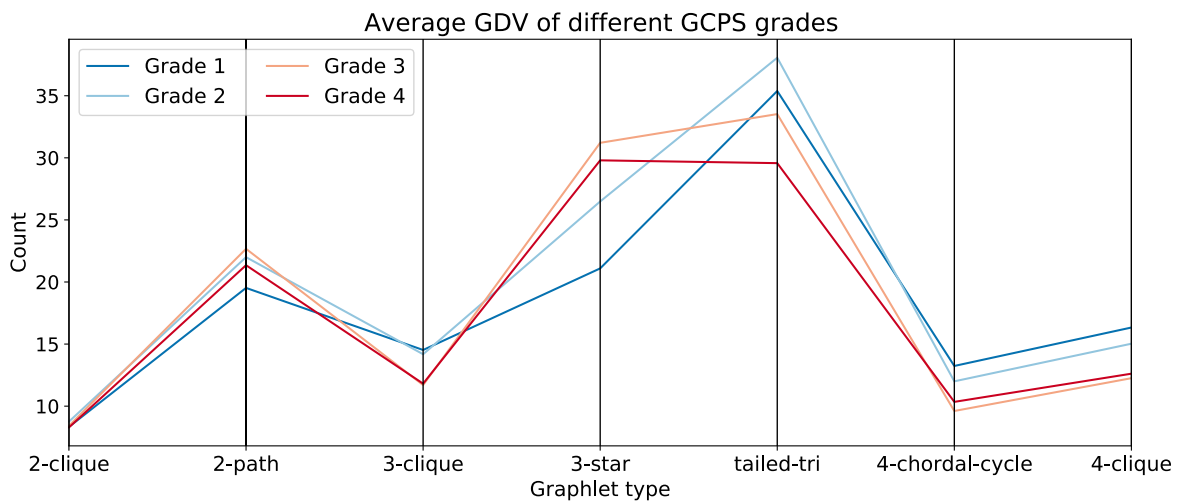


Figure 5.4 : Parallel coordinates plot of average GDV of different GCPS grades. Each coordinate represents the average number of graphlets belonging to that type.

First, we calculate the average Graphlet Degree Vectors of patients from each GCPS grade. A parallel coordinates plot shows the average degrees of all seven egocentric graphlets at each grade (Figure 5.4). We see that patients of higher-grade pains (grade 3 and grade 4) have more star-like structures (3-star graphlets) in their

social networks, and patients of lower pain grades (grade 1 and grade 2) form more 3-cliques, tailed-triangles, 4-chordal-cycles and 4-cliques. A worse connected star-like structure indicates a more isolated social environment, and a better connected structure such as the 3-clique or the 4-clique could be a sign of better social support. These findings are consistent with the previously mentioned studies [111, 65, 15] and provide further evidence that a patient's social network could inform the perceived pain grade. In addition, we find that the number of connections (2-cliques) does not help distinguish pain grades. This may result from the limited number of contacts in the dataset. Still nevertheless, another work also found that the size of a patient's egocentric social network is not significantly related to changes in pain [61]. This also explains why more complicated network structures should be considered in the analysis of patients' social networks.

Further, in order to investigate the relationship between the types of interactions and the pain grades, we employ the Typed-Edge Graphlet Degree Vector and focus on two particular graphlets, i.e. the poorly connected 3-star graphlet and the well connected 4-clique graphlet. These two graphlets are chosen because they present distinct differences between patients of lower pain grades and patients of higher pain grades. For each of the graphlets, we calculate the average TyE-GDV of patients from every pain grade and generate a parallel coordinates plot (Figure 5.5). We find that in the 3-star graphlet (Figure 5.5a), higher-grade pain patients have significantly more edges of type '5' (friend) and type '10' (healthcare worker) than lower-grade pain patients. In other words, friends and healthcare workers are not well connected in higher-grade pain patients. It thus provides the potential for interventions that increase the social involvements of a patient's friends and healthcare workers to improve the management of chronic pain.

Then from the average TyE-GDV of the 4-clique graphlet (Figure 5.5b), we observe that lower-grade pain patients have more edges of type '5' (friend) than

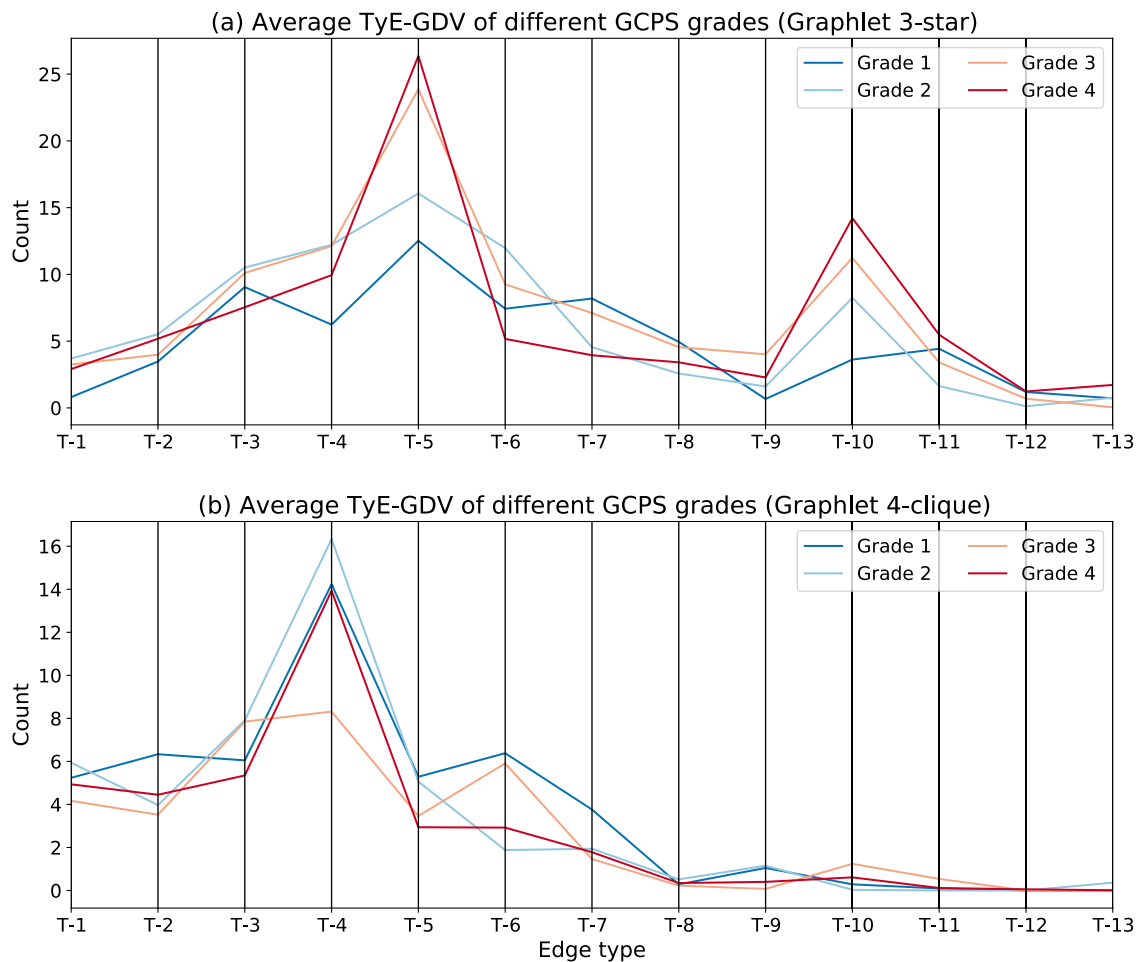


Figure 5.5 : Parallel coordinates plot of average TyE-GDV of different GPCS grades for two graphlets. Each coordinate represents the average number of edges belonging to that type.

higher-grade pain patients (5.2 compared to 3.2). That is to say, friends appear more often in these tightly connected groups among patients of lower-grade pain. The importance of the friend relationship is revealed in both 3-star and 4-clique graphlets. As pointed out by other studies [67, 239], patients with severe chronic pain may be at risk of deterioration in their friendships and are in need of supportive behaviours from friends. Another marked contrast between the higher-grade and lower-grade pain patients is in edge type '9' (colleague). Colleagues hardly appear (0.24 on average) in these closely connected structures among the former group, whereas more than one colleague (1.1 on average) emerges among the latter group.

It may reflect the adverse effects of severe chronic pain on patients’ professional activities [87]. To give an intuitive understanding of the structural differences, we give two actual examples from the dataset as the social network prototypes of pain grade-1 and pain grade-4, respectively (Figure 5.6).

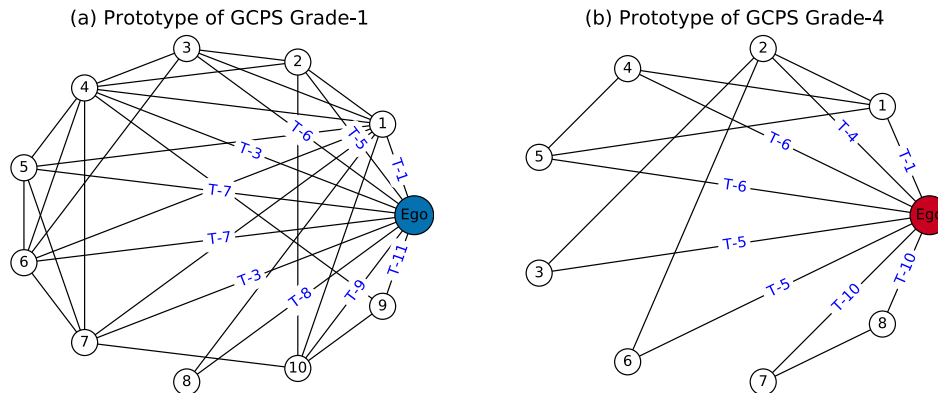


Figure 5.6 : Prototypes of GCPS grade-1 and GCPS grade-4.

This experiment shows that the extra information brought by TyE-GDV provides us with more insights into the relationship between patients’ social link types and their pain grades. Therefore, it has implications for how therapeutic interventions could be improved by increasing particular types of social connections.

5.5.3 Predicting pain grades

We now apply the proposed TyE-GDV, and the extended egocentric versions of colored graphlets (ColoredE-GDV) and heterogeneous graphlets (HeteroE-GDV^N) approaches in a typical machine learning task.

Node classification is one of the most popular and widely adopted tasks in network science [20], where each node is assigned a ground truth category. Here, we aim to predict the GCPS grade of chronic pain patients. In order to test the utility of the proposed approaches, we fit six sets of features into a random forest classifier. The first set includes patients’ demographic attributes, pain-related descriptions and

physical/psychological well-being indicators. Since it contains no structural information, we identify it as raw feature set. The second set includes the raw features plus the typed-edge degree vector (TyE-DV). The third set includes the raw features plus the classic GDV. The fourth set includes the raw features plus the proposed TyE-GDV. The fifth set includes the raw features plus the colored graphlets degree vector (ColoredE-GDV), and the sixth set includes the raw features plus the heterogeneous graphlets degree vector (HeteroE-GDV^N).

As the dataset is not large and the distribution of four grades is not balanced (see Section 5.5.1), we adopt a stratified 5-fold cross-validation [187] to evaluate the classification performance with different sets of features. Also, because decision tree-based models are inherently stochastic, we repeat the above step 500 times and report the mean metric score.

	Macro F1 (Mean \pm Std)	Gain over raw feat. (Mean)	Time in sec. (Sum)
Stratified	0.248 \pm 0.024	—	3
Raw feat.	0.578 \pm 0.005	—	116
Raw feat. + TyE-DV	0.600 \pm 0.005	3.8%	130
Raw feat. + GDV	0.597 \pm 0.008	3.3%	138
Raw feat. + ColoredE-GDV	0.608 \pm 0.006	5.2%	2091
Raw feat. + HeteroE-GDV ^N	0.638 \pm 0.006	10.4%	8230
Raw feat. + TyE-GDV	0.619 \pm 0.004	7.1%	252

Table 5.3 : Prediction results in average macro-F1 score (\pm standard deviation), average gain over raw features, and total running time of 500 repetitions.

We report the average macro-F1 scores of three models in Table 5.3. The macro-F1 score is chosen because this is a multi-class classification problem and the distribution of the four classes is unbalanced. A naive classifier (Stratified) is also added in the table, which generates predictions by respecting the class distribution in the training set. We see clearly that the bottom three approaches that encode type information in graphlets (raw features plus ColoredE-GDV, raw features plus

Approach	GDV	TyE-DV	TyE-GDV	ColoredE-GDV	HeteroE-GDV
Len. of vector	7	13	91	12367	38870

Table 5.4 : Comparison of vector length of different approaches.

HeteroE-GDV^N, and raw features plus TyE-GDV) perform better than the set of raw features plus TyE-DV and the set of raw features plus GDV. Recall that TyE-DV captures edge type information but with very limited structural information, and GDV, on the other hand, captures the rich structural information but without edge type information. This evidently shows that combining edge type information and rich structural information could lead to more distinctive features in network learning tasks.

We also observe large differences in the running time of those methods. The running time of the set of raw features plus ColoredE-GDV, and especially the set of raw features plus HeteroE-GDV^N are many times higher than other methods. This is because our dataset has 13 types of edges and the lengths of vectors generated from these two methods grow near exponentially with $|\mathcal{T}_e|$. Correspondingly, the speed of the machine learning algorithm will slow down as the feature vector becomes larger. Table 5.4 gives the vector lengths of all five approaches. Note that there is no edge type information between alter nodes in many egocentric networks, including this case study dataset, our implementations of ColoredE-GDV and HeteroE-GDV^N has excluded all the impossible combinations. Overall speaking, the proposed TyE-GDV achieves competitive performance while maintaining a small vector length.

5.6 Conclusion

In this chapter, we proposed to embed edge type information in graphlets and introduced the framework for calculating Typed-Edge Graphlets Degree Vector for both sociocentric and egocentric networks. Moreover, we extended the colored

graphlets approach and the heterogeneous graphlets approach to edge-typed networks and egocentric networks. After applying GDV and TyE-GDV to the chronic pain patients dataset, we found that 1) a patient's social network structure could inform their perceived pain grade; and 2) particular types of social relationships, such as friends, colleagues and healthcare workers, could bear more importance in understanding the effect of chronic pain and therefore lead to more effective therapeutic interventions. We also showed that the rich structural information combined with the edge type information results in significant improvement of a typical machine learning task that predicts patients' pain grades.

This work fulfills research objectives 4 and 5.

Chapter 6

Conclusion and future works

This thesis focuses on furthering the understanding of local structures and local information in complex networks, by proposing new approaches to measuring 3-node and 4-node structure formation, as well as algorithms encoding rich edge attributes in graphlets.

For measuring the formation of 3-node local structures, we introduced the directed closure coefficient and its patterns to measure directed triangle formation from an end-node perspective. Through extensive experiments, we demonstrated that, at network-level, including the four closure patterns leads to significant improvement in classifying different types of directed networks; while at link-level analysis, the source and target coefficients can be fused together with common neighbours as effective predictors, especially in food webs, software networks, web graphs and word adjacency networks. To deepen the understanding of the 4-cycle structure, we proposed i-quad and o-quad coefficients in order to better describe and analyse networks that contain fewer triangles and are rich in quadrangles. We then revealed empirically that the average o-quad coefficient is smaller than the average i-quad coefficient in most types of networks; and that the i-quad and o-quad coefficients are significantly larger than the traditional clustering coefficient. We also prove theoretically that under a configuration model, the o-quad coefficient increases with node degree while the i-quad coefficient does not change too much as the node degree varies. For encoding edge attributes in complex networks, we proposed a new framework to effectively generate a typed-edge graphlet degree vector for each node.

The approach is applied to a recently collected dataset of chronic pain patients. We uncovered that a patient’s social network structure could indeed inform his/her perceived pain grade, and that specific types of social relationships, such as friends, colleagues and healthcare workers, are more important in influencing the perception of pain.

To summarise, our contributions are:

- We proposed new taxonomies for various graph structural measures (attained objective 1);
- We proposed a new approach to assessing the edge clustering phenomenon in directed networks (attained objective 2);
- We proposed new metrics for measuring quadrangle formation in complex networks (attained objective 3);
- We proposed a new framework for effectively embedding edge type information in graphlets (attained objective 4);
- We applied our approaches to different types of real-world networks, and proved their performance in multiple learning and analysis tasks (attained objective 5).

For future studies, we plan to investigate further the local structures and information in a more complicated dynamic network model that adds the dimension of time, and a multilayered network model that integrates data from different sources. In addition, we plan to focus on applying the promising approaches of graph local structure to more real-world problems. For example, we will continue our joint study of chronic pain patients by exploring the possible relationships between local structures in social networks and the mental/physical well-beings (such as the degree

of depression/anxiety, and the ability to participate in social roles and activities). We will also commence another joint research project about disrupting organised criminal networks through studying the rich-labelled local structures.

Bibliography

- [1] A. H. Abdo and A. de Moura, “Clustering as a measure of the local topology of networks,” *arXiv preprint physics/0605235*, 2006.
- [2] L. A. Adamic and E. Adar, “Friends and neighbors on the web,” *Social networks*, 2003.
- [3] S. Ahajjam and H. Badir, “Identification of influential spreaders in complex networks using hybridrank algorithm,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.
- [4] I. Ahmad, M. U. Akhtar, S. Noor, and A. Shahnaz, “Missing link prediction using common neighbor and centrality based parameterized algorithm,” *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.
- [5] S. E. Ahnert and T. M. Fink, “Clustering signatures classify directed networks,” *Physical Review E*, 2008.
- [6] L. Akoglu, H. Tong, and D. Koutra, “Graph based anomaly detection and description: a survey,” *Data mining and knowledge discovery*, 2015.
- [7] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using supervised learning,” in *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.
- [8] T. Alahakoon, R. Tripathi, N. Kourtellis, R. Simha, and A. Iamnitchi, “K-path centrality: A new centrality measure in social networks,” in *Proceedings of the 4th workshop on social network systems*, 2011, pp. 1–6.

- [9] D. Aleja, R. Criado, A. J. G. del Amo, Á. Pérez, and M. Romance, “Non-backtracking pagerank: from the classic model to hashimoto matrices,” *Chaos, Solitons & Fractals*, vol. 126, pp. 283–291, 2019.
- [10] F. T. Aleskerov, N. Meshcheryakova, and S. Shvydun, “Centrality measures in networks based on nodes attributes, long-range interactions and group influence,” *Long-Range Interactions and Group Influence*, 2016.
- [11] Z. AlGhamdi, F. Jamour, S. Skiadopoulou, and P. Kalnis, “A benchmark for betweenness centrality approximation algorithms on large graphs,” in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017, pp. 1–12.
- [12] D. Aparicio, P. Ribeiro, and F. Silva, “Extending the applicability of graphlets to directed networks,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 6, pp. 1302–1315, 2016.
- [13] S. Ataei, N. Attar, S. Aliakbary, and F. Bakouie, “Graph theoretical approach for screening autism on brain complex networks,” *SN Applied Sciences*, 2019.
- [14] J. Bae and S. Kim, “Identifying and ranking influential spreaders in complex networks by neighborhood coreness,” *Physica A: Statistical Mechanics and its Applications*, vol. 395, pp. 549–559, 2014.
- [15] S. Bannon, J. Greenberg, R. A. Mace, J. J. Locascio, and A.-M. Vranceanu, “The role of social isolation in physical and emotional outcomes among patients with chronic pain,” *General Hospital Psychiatry*, 2021.
- [16] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.
- [17] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, “The architecture of complex weighted networks,” *PNAS*, 2004.

- [18] P. Basaras, G. Iosifidis, D. Katsaros, and L. Tassiulas, “Identifying influential spreaders in complex multilayer networks: A centrality perspective,” *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 1, pp. 31–45, 2017.
- [19] F. Battiston, V. Nicosia, M. Chavez, and V. Latora, “Multilayer motif analysis of brain networks,” *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 4, p. 047404, 2017.
- [20] S. Bhagat, G. Cormode, and S. Muthukrishnan, “Node classification in social networks,” in *Social network data analytics*. Springer, 2011.
- [21] G. Bianconi, *Multilayer networks: structure and function*. Oxford university press, 2018.
- [22] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.
- [23] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin, “The structure and dynamics of multilayer networks,” *Physics reports*, vol. 544, no. 1, pp. 1–122, 2014.
- [24] P. Bonacich, “Power and centrality: A family of measures,” *American journal of sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.
- [25] P. Bonacich and P. Lloyd, “Eigenvector-like measures of centrality for asymmetric relations,” *Social networks*, vol. 23, no. 3, pp. 191–201, 2001.
- [26] S. P. Borgatti and M. G. Everett, “A graph-theoretic perspective on centrality,” *Social networks*, vol. 28, no. 4, pp. 466–484, 2006.

- [27] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing social networks*. Sage, 2018.
- [28] A. Bramson and B. Vandermarliere, “Benchmarking measures of network influence,” *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.
- [29] D. A. Bright, C. Greenhill, M. Reynolds, A. Ritter, and C. Morselli, “The use of actor-level attributes and centrality measures to identify key actors: A case study of an australian drug trafficking network,” *Journal of contemporary criminal justice*, vol. 31, no. 3, pp. 262–278, 2015.
- [30] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.
- [31] P. Bródka, K. Skibicki, P. Kazienko, and K. Musiał, “A degree centrality in multi-layered social network,” in *2011 International Conference on Computational Aspects of Social Networks (CASoN)*. IEEE, 2011, pp. 237–242.
- [32] M. R. Brust, D. Turgut, C. H. Ribeiro, and M. Kaiser, “Is the clustering coefficient a measure for fault tolerance in wireless sensor networks?” in *2012 IEEE International Conference on Communications (ICC)*. IEEE, 2012, pp. 183–187.
- [33] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, “Structure of cycles and local ordering in complex networks,” *The European Physical Journal B*, vol. 38, no. 2, pp. 183–186, 2004.
- [34] L. Candeloro, L. Savini, and A. Conte, “A new weighted degree centrality measure: The application in an animal disease epidemic,” *PloS one*, vol. 11, no. 11, p. e0165781, 2016.

- [35] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, “A model of internet topology using k-shell decomposition,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 27, pp. 11 150–11 154, 2007.
- [36] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, “Automatic resource compilation by analyzing hyperlink structure and associated text,” *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 65–74, 1998.
- [37] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, “Identifying influential nodes in large-scale directed networks: the role of clustering,” *PloS one*, vol. 8, no. 10, p. e77455, 2013.
- [38] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, “Identifying influential nodes in complex networks,” *Physica a: Statistical mechanics and its applications*, vol. 391, no. 4, pp. 1777–1787, 2012.
- [39] J. Chen and H. Chen, “Edge-featured graph attention network,” *arXiv preprint arXiv:2101.07671*, 2021.
- [40] J. Cohen, “Trusses: Cohesive subgraphs for social network analysis,” *National security agency technical report*, vol. 16, no. 3.1, 2008.
- [41] V. Colizza, R. Pastor-Satorras, and A. Vespignani, “Reaction–diffusion processes and metapopulation models in heterogeneous networks,” *Nature Physics*, 2007.
- [42] G. Costantini and M. Perugini, “Generalization of clustering coefficients to signed correlation networks,” *PloS one*, 2014.
- [43] K. Das, S. Samanta, and M. Pal, “Study on centrality measures in social networks: a survey,” *Social network analysis and mining*, vol. 8, no. 1, pp. 1–11, 2018.

- [44] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, “Mathematical formulation of multilayer networks,” *Physical Review X*, vol. 3, no. 4, p. 041022, 2013.
- [45] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, “Centrality in interconnected multilayer networks,” *arXiv preprint arXiv:1311.2906*, 2013.
- [46] —, “Ranking in interconnected multilayer networks reveals versatile nodes,” *Nature communications*, vol. 6, no. 1, pp. 1–6, 2015.
- [47] P. De Meo, E. Ferrara, G. Fiumara, and A. Ricciardello, “A novel measure of edge centrality in social networks,” *Knowledge-based systems*, vol. 30, pp. 136–150, 2012.
- [48] H. Deng, M. R. Lyu, and I. King, “A generalized co-hits algorithm and its application to bipartite graphs,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 239–248.
- [49] M. E. Dickison, M. Magnani, and L. Rossi, *Multilayer social networks*. Cambridge University Press, 2016.
- [50] T. Dimitrova, K. Petrovski, and L. Kocarev, “Graphlets in multiplex networks,” *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.
- [51] C. Doerr and N. Blenn, “Metric convergence in social network sampling,” in *Proceedings of the 5th ACM workshop on HotPlanet*, 2013.
- [52] J. F. Donges, H. C. Schultz, N. Marwan, Y. Zou, and J. Kurths, “Investigating the topology of interacting networks,” *The European Physical Journal B*, vol. 84, no. 4, pp. 635–651, 2011.

- [53] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, “Complex networks in climate dynamics,” *The European Physical Journal Special Topics*, vol. 174, no. 1, pp. 157–179, 2009.
- [54] J. A. Dunne, R. J. Williams, and N. D. Martinez, “Network structure and biodiversity loss in food webs: robustness increases with connectance,” *Ecology letters*, vol. 5, no. 4, pp. 558–567, 2002.
- [55] C. Durón, “Heatmap centrality: A new measure to identify super-spreader nodes in scale-free networks,” *Plos one*, vol. 15, no. 7, p. e0235690, 2020.
- [56] D. Easley, J. Kleinberg *et al.*, *Networks, crowds, and markets*. Cambridge university press Cambridge, 2010, vol. 8.
- [57] L. Egghe, “Theory and practise of the g-index,” *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.
- [58] V. M. Eguiluz and K. Klemm, “Epidemic threshold in structured scale-free networks,” *Physical Review Letters*, vol. 89, no. 10, p. 108701, 2002.
- [59] E. Estrada, D. J. Higham, and N. Hatano, “Communicability betweenness in complex networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 5, pp. 764–774, 2009.
- [60] E. Estrada and J. A. Rodriguez-Velazquez, “Subgraph centrality in complex networks,” *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.
- [61] A. W. Evers, F. W. Kraaijaat, R. Geenen, J. W. Jacobs, and J. W. Bijlsma, “Pain coping and social support as predictors of long-term functional disability and pain in early rheumatoid arthritis,” *Behaviour research and therapy*, 2003.
- [62] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O’Connor, M. Li *et al.*, “Large-scale mapping

- of human protein–protein interactions by mass spectrometry,” *Molecular systems biology*, 2007.
- [63] G. Fagiolo, “Clustering in complex directed networks,” *Physical Review E*, 2007.
- [64] R. Z. Farahani, E. Miandoabchi, W. Y. Szeto, and H. Rashidi, “A review of urban transportation network design problems,” *European Journal of Operational Research*, vol. 229, no. 2, pp. 281–302, 2013.
- [65] M. A. Ferreira-Valente, J. L. Pais-Ribeiro, and M. P. Jensen, “Associations between psychosocial factors and pain intensity, physical functioning, and psychological functioning in patients with chronic pain: a cross-cultural comparison,” *The Clinical journal of pain*, 2014.
- [66] L. R. Ford and D. R. Fulkerson, *Flows in networks*. Princeton university press, 2015.
- [67] P. A. Forgeron, P. McGrath, B. Stevens, J. Evans, B. Dick, G. A. Finley, and T. Carlson, “Social information processing in adolescents with chronic pain: My friends don’t really understand me,” *Pain*, 2011.
- [68] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, “Configuring random graph models with fixed degree sequences,” *SIAM Review*, 2018.
- [69] L. C. Freeman, “A set of measures of centrality based on betweenness,” *Sociometry*, pp. 35–41, 1977.
- [70] —, “Centrality in social networks conceptual clarification,” *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.
- [71] L. C. Freeman, S. P. Borgatti, and D. R. White, “Centrality in valued graphs: A measure of betweenness based on network flow,” *Social networks*, vol. 13,

- no. 2, pp. 141–154, 1991.
- [72] A. Fronczak, J. A. Hołyst, M. Jedynek, and J. Sienkiewicz, “Higher order clustering coefficients in barabási–albert networks,” *Physica A: Statistical Mechanics and its Applications*, 2002.
- [73] F. Gao, K. Musial, C. Cooper, and S. Tsoka, “Link prediction methods and their accuracy for different social networks and network metrics,” *Scientific programming*, 2015.
- [74] S. Gao, J. Ma, Z. Chen, G. Wang, and C. Xing, “Ranking the spreading ability of nodes in complex networks based on local structure,” *Physica A: Statistical Mechanics and its Applications*, vol. 403, pp. 130–147, 2014.
- [75] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” *PNAS*, 2002.
- [76] D. F. Gleich, “Pagerank beyond the web,” *Siam Review*, vol. 57, no. 3, pp. 321–363, 2015.
- [77] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, “Using collaborative filtering to weave an information tapestry,” *Communications of the ACM*, 1992.
- [78] R. Goldstein and M. S. Vitevitch, “The influence of clustering coefficient on word-learning: how groups of similar sounding words facilitate acquisition,” *Frontiers in psychology*, 2014.
- [79] L. Gong and Q. Cheng, “Exploiting edge features for graph neural networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9211–9219.
- [80] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *KDD*, 2016.

- [81] S. Gu, J. Johnson, F. E. Faisal, and T. Milenković, “From homogeneous to heterogeneous network alignment via colored graphlets,” *Scientific reports*, vol. 8, no. 1, pp. 1–16, 2018.
- [82] J. Guan, Y. Li, L. Xing, Y. Li, and G. Liang, “Closeness centrality for similarity-weight network and its application to measuring industrial sectors’ position on the global value chain,” *Physica A: Statistical Mechanics and its Applications*, vol. 541, p. 123337, 2020.
- [83] E. A. Hahn, R. F. DeVellis, R. K. Bode, S. F. Garcia, L. D. Castel, S. V. Eisen, H. B. Bosworth, A. W. Heinemann, N. Rothrock, and D. Cella, “Measuring social health in the patient-reported outcomes measurement information system (promis): item bank development and testing,” *Quality of Life Research*, 2010.
- [84] B. H. Hall and B. Adam, “13 the nber patent-citations data file: Lessons, insights, and methodological tools,” *Patents, citations, and innovations: A window on the knowledge economy*, 2002.
- [85] A. Halu, R. J. Mondragón, P. Panzarasa, and G. Bianconi, “Multiplex pagerank,” *PloS one*, vol. 8, no. 10, p. e78293, 2013.
- [86] W. L. Hamilton, R. Ying, and J. Leskovec, “Representation learning on graphs: Methods and applications,” *arXiv preprint arXiv:1709.05584*, 2017.
- [87] S. Harris, S. Morley, and S. B. Barton, “Role loss and emotional adjustment in chronic pain,” *Pain*, 2003.
- [88] K.-i. Hashimoto, “Zeta functions of finite graphs and representations of p-adic groups,” in *Automorphic forms and geometry of arithmetic varieties*. Elsevier, 1989, pp. 211–280.

- [89] F. Heider, “Attitudes and cognitive organization,” *The Journal of psychology*, 1946.
- [90] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, “Rolx: structural role extraction & mining in large graphs,” in *KDD*, 2012.
- [91] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.
- [92] T. Hočevar and J. Demšar, “Computation of graphlet orbits for nodes and edges in sparse graphs,” *Journal of Statistical Software*, vol. 71, pp. 1–24, 2016.
- [93] B. L. Hoffman, E. M. Felter, K.-H. Chu, A. Shensa, C. Hermann, T. Wolyynn, D. Williams, and B. A. Primack, “It’s not all about autism: The emerging landscape of anti-vaccination sentiment on facebook,” *Vaccine*, vol. 37, no. 16, pp. 2216–2223, 2019.
- [94] T. Hogg and K. Lerman, “Social dynamics of digg,” *EPJ Data Science*, 2012.
- [95] P. Holme and J. Saramäki, “Temporal networks,” *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.
- [96] T. Hoshino, H. Doi, G.-I. Uramoto, L. Wörmer, R. R. Adhikari, N. Xiao, Y. Morono, S. D’Hondt, K.-U. Hinrichs, and F. Inagaki, “Global diversity of microbial communities in marine sediment,” *Proceedings of the national academy of sciences*, vol. 117, no. 44, pp. 27 587–27 597, 2020.
- [97] X. Hu, S. Liu, S. Chang, and H. Li, “A quad motifs index for directed link prediction,” *IEEE Access*, vol. 7, pp. 159 527–159 534, 2019.

- [98] Y. Hulovatyy, H. Chen, and T. Milenković, “Exploring the structure and function of temporal networks with dynamic graphlets,” *Bioinformatics*, vol. 31, no. 12, pp. i171–i180, 2015.
- [99] M. O. Jackson, B. W. Rogers, and Y. Zenou, “The economic consequences of social-network structure,” *Journal of Economic Literature*, vol. 55, no. 1, pp. 49–95, 2017.
- [100] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic *et al.*, “String 8—a global view on proteins and their functional interactions in 630 organisms,” *Nucleic acids research*, vol. 37, no. suppl-1, pp. D412–D416, 2009.
- [101] Q. Ji, D. Li, and Z. Jin, “Divisive algorithm based on node clustering coefficient for community detection,” *IEEE Access*, vol. 8, pp. 142 337–142 347, 2020.
- [102] M. Jia, M. V. Alboom, L. Goubert, P. Bracke, B. Gabrys, and K. Musial, “Analysing ego-networks via typed-edge graphlets: A case study of chronic pain patients,” in *International Conference on Complex Networks and Their Applications*. Springer, 2021, pp. 514–526.
- [103] M. Jia, B. Gabrys, and K. Musial, “Closure coefficient in complex directed networks,” in *International Conference on Complex Networks and Their Applications*. Springer, 2020, pp. 62–74.
- [104] —, “Directed closure coefficient and its patterns,” *Plos one*, vol. 16, no. 6, p. e0253822, 2021.
- [105] —, “Directed closure coefficient and its patterns,” *PLOS ONE*, vol. 16, pp. 1–23, 06 2021.
- [106] —, “Measuring quadrangle formation in complex networks,” *IEEE Transactions on Network Science and Engineering*, 2021.

- [107] L. Jiang, L. Tang, and J. Lü, “Controllability of multilayer networks,” *Asian Journal of Control*, 2021.
- [108] X. Jiang, R. Zhu, S. Li, and P. Ji, “Co-embedding of nodes and edges with graph neural networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [109] A. Josang, R. F. Hayward, and S. Pope, “Trust network analysis with subjective logic,” 2006.
- [110] K. Juszczyszyn, P. Kazienko, and K. Musiał, “Local topology of social network based on motif analysis,” in *KES*. Springer, 2008.
- [111] N. V. Karayannis, I. Baumann, J. A. Sturgeon, M. Melloh, and S. C. Mackey, “The impact of social isolation on pain interference: a longitudinal study,” *Annals of Behavioral Medicine*, 2019.
- [112] L. Katz, “A new status index derived from sociometric analysis,” *Psychometrika*, 1953.
- [113] H. Kim and R. Anderson, “Temporal node centrality in complex networks,” *Physical Review E*, vol. 85, no. 2, p. 026107, 2012.
- [114] H. Kim, J. Tang, R. Anderson, and C. Mascolo, “Centrality prediction in dynamic human contact networks,” *Computer Networks*, vol. 56, no. 3, pp. 983–996, 2012.
- [115] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [116] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, “Identification of influential spreaders in complex networks,” *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

- [117] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, “Multilayer networks,” *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.
- [118] J. M. Kleinberg *et al.*, “Authoritative sources in a hyperlinked environment.” in *SODA*, vol. 98. Citeseer, 1998, pp. 668–677.
- [119] A. Korn, A. Schubert, and A. Telcs, “Lobby index in networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 11, pp. 2221–2226, 2009.
- [120] I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao *et al.*, “Network-based prediction of protein interactions,” *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.
- [121] L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, “Temporal motifs in time-dependent networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 11, p. P11005, 2011.
- [122] A. E. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. W. Taylor, “Compartments revealed in food-web structure,” *Nature*, vol. 426, no. 6964, pp. 282–285, 2003.
- [123] M. Krivelevich and B. Sudakov, “The largest eigenvalue of sparse random graphs,” *Combinatorics, Probability and Computing*, vol. 12, no. 1, pp. 61–72, 2003.
- [124] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, “Spectral redemption in clustering sparse networks,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20 935–20 940, 2013.

- [125] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, “Topological network alignment uncovers biological function and phylogeny,” *Journal of the Royal Society Interface*, vol. 7, no. 50, pp. 1341–1354, 2010.
- [126] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, “Rev2: Fraudulent user prediction in rating platforms,” in *WSDM*. ACM, 2018.
- [127] S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos, “Edge weight prediction in weighted signed networks,” in *ICDM*. IEEE, 2016.
- [128] J. Kunegis, “Konekt: the koblenz network collection,” in *Proceedings of the 22nd International Conference on World Wide Web*, 2013.
- [129] J. Kunegis, A. Lommatzsch, and C. Bauckhage, “The slashdot zoo: mining a social network with negative edges,” in *WWW*, 2009.
- [130] T. LaFond, J. Neville, and B. Gallagher, “Anomaly detection in networks with changing trends,” in *ODD² Workshop*, 2014.
- [131] T. Lee, B. Choi, Y. Shin, and J. Kwak, “Automatic malware mutant detection and group classification based on the n-gram and clustering coefficient,” *The Journal of Supercomputing*, 2018.
- [132] J. Leskovec, L. A. Adamic, and B. A. Huberman, “The dynamics of viral marketing,” *ACM Transactions on the Web (TWEB)*, 2007.
- [133] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” in *WWW*, 2010.
- [134] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM transactions on Knowledge Discovery from Data (TKDD)*, 2007.

- [135] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, “Statistical properties of community structure in large social and information networks,” in *WWW*, 2008.
- [136] J. Leskovec and R. Sosič, “Snap: A general-purpose network analysis and graph-mining library,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2016.
- [137] M. Ley, “The dblp computer science bibliography: Evolution, research issues, perspectives,” in *International symposium on string processing and information retrieval*. Springer, 2002.
- [138] C.-H. Li, C.-C. Tsai, and S.-Y. Yang, “Analysis of epidemic spreading of an sirs model in complex heterogeneous networks,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 19, no. 4, pp. 1042–1054, 2014.
- [139] G. Li, M. Li, J. Wang, Y. Li, and Y. Pan, “United neighborhood closeness centrality and orthology for predicting essential proteins,” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 4, pp. 1451–1458, 2018.
- [140] M. Y. Li and J. S. Muldowney, “Global stability for the seir model in epidemiology,” *Mathematical biosciences*, vol. 125, no. 2, pp. 155–164, 1995.
- [141] Q. Li, T. Zhou, L. Lü, and D. Chen, “Identifying influential spreaders by weighted leaderrank,” *Physica A: Statistical Mechanics and its Applications*, vol. 404, pp. 47–55, 2014.
- [142] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, and T. Zhou, “Identifying influential spreaders by gravity model,” *Scientific reports*, vol. 9, no. 1, pp. 1–7, 2019.
- [143] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou, “Ranking in evolving complex networks,” *Physics Reports*, vol. 689, pp. 1–54, 2017.

- [144] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *Journal of the American society for information science and technology*, 2007.
- [145] P. G. Lind, M. C. Gonzalez, and H. J. Herrmann, “Cycles and clustering in bipartite networks,” *Physical review E*, 2005.
- [146] Q. Liu, Y.-X. Zhu, Y. Jia, L. Deng, B. Zhou, J.-X. Zhu, and P. Zou, “Leveraging local h-index to identify and rank influential spreaders in networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 512, pp. 379–391, 2018.
- [147] Y. Liu, M. Tang, T. Zhou, and Y. Do, “Identify influential spreaders in complex networks, the role of neighborhood,” *Physica A: Statistical Mechanics and its Applications*, vol. 452, pp. 289–298, 2016.
- [148] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, “Vital nodes identification in complex networks,” *Physics Reports*, vol. 650, pp. 1–63, 2016.
- [149] L. Lü, C.-H. Jin, and T. Zhou, “Similarity index based on local paths for link prediction of complex networks,” *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.
- [150] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, “Leaders in social networks, the delicious case,” *PloS one*, vol. 6, no. 6, p. e21202, 2011.
- [151] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, “The h-index of a network node and its relation to degree and coreness,” *Nature communications*, vol. 7, no. 1, pp. 1–7, 2016.

- [152] L.-l. Ma, C. Ma, H.-F. Zhang, and B.-H. Wang, “Identifying influential spreaders in complex networks based on gravity formula,” *Physica A: Statistical Mechanics and its Applications*, vol. 451, pp. 205–212, 2016.
- [153] Q. Ma and J. Ma, “Identifying and ranking influential spreaders in complex networks with consideration of spreading probability,” *Physica A: Statistical Mechanics and its Applications*, vol. 465, pp. 312–330, 2017.
- [154] X. Ma and Y. Ma, “The local triangle structure centrality method to rank nodes in networks,” *Complexity*, vol. 2019, 2019.
- [155] X. Ma and L. Gao, “Biological network analysis: insights into structure and functions,” *Briefings in functional genomics*, vol. 11, no. 6, pp. 434–442, 2012.
- [156] T. Martin, X. Zhang, and M. E. Newman, “Localization and centrality in networks,” *Physical review E*, vol. 90, no. 5, p. 052808, 2014.
- [157] N. D. Martinez, “Artifacts or attributes? effects of resolution on the little rock lake food web,” *Ecological monographs*, 1991.
- [158] V. Martínez, F. Berzal, and J.-C. Cubero, “A survey of link prediction in complex networks,” *ACM computing surveys (CSUR)*, vol. 49, no. 4, pp. 1–33, 2016.
- [159] P. Massa and P. Avesani, “Controversial users demand local trust metrics: An experimental study on epinions. com community,” in *AAAI*, 2005.
- [160] P. D. Meo, “Trust prediction via matrix factorisation,” *ACM Transactions on Internet Technology (TOIT)*, 2019.
- [161] T. Milenković and N. Pržulj, “Uncovering biological network function via graphlet degree signatures,” *Cancer informatics*, vol. 6, pp. CIN–S680, 2008.

- [162] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, “Superfamilies of evolved and designed networks,” *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.
- [163] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks,” *Science*, 2002.
- [164] S. Moein, N. Nickaeen, A. Roointan, N. Borhani, Z. Heidary, S. H. Javanmard, J. Ghaisari, and Y. Gheisari, “Inefficiency of sir models in forecasting covid-19 epidemic: a case study of isfahan,” *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.
- [165] E. Mones, L. Vicsek, and T. Vicsek, “Hierarchy measure for complex networks,” *PloS one*, vol. 7, no. 3, p. e33799, 2012.
- [166] J. Moody, “Peer influence groups: identifying dense clusters in large networks,” *Social Networks*, 2001.
- [167] F. Morone and H. A. Makse, “Influence maximization in complex networks through optimal percolation,” *Nature*, vol. 524, no. 7563, pp. 65–68, 2015.
- [168] A. Muscoloni, I. Abdelhamid, and C. V. Cannistraci, “Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more,” *bioRxiv*, p. 346916, 2018.
- [169] K. Musiał and K. Juszczyszyn, “Motif-based analysis of social position influence on interconnection patterns in complex social network,” in *ACIIDS*. IEEE, 2009.
- [170] K. Musiał and P. Kazienko, “Social networks on the internet,” *World Wide Web*, 2013.

- [171] M. Newman, *Networks*. Oxford university press, 2018.
- [172] M. E. Newman, “The structure and function of complex networks,” *SIAM Rev.*, 2003.
- [173] —, “A measure of betweenness centrality based on random walks,” *Social networks*, vol. 27, no. 1, pp. 39–54, 2005.
- [174] M. E. Newman, S. H. Strogatz, and D. J. Watts, “Random graphs with arbitrary degree distributions and their applications,” *Physical review E*, 2001.
- [175] V. Nicosia, J. Tang, C. Mascolo, M. Musolesi, G. Russo, and V. Latora, “Graph metrics for temporal networks,” in *Temporal networks*. Springer, 2013, pp. 15–40.
- [176] C. C. Noble and D. J. Cook, “Graph-based anomaly detection,” in *KDD*, 2003.
- [177] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, “Intensity and coherence of motifs in weighted complex networks,” *Physical Review E*, 2005.
- [178] T. Opsahl, F. Agneessens, and J. Skvoretz, “Node centrality in weighted networks: Generalizing degree and shortest paths,” *Social networks*, vol. 32, no. 3, pp. 245–251, 2010.
- [179] T. Opsahl and P. Panzarasa, “Clustering in weighted networks,” *Social networks*, 2009.
- [180] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. Del-Toro *et al.*, “The mintact project—intact as a common curation platform for 11 molecular interaction databases,” *Nucleic acids research*, 2014.

- [181] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, “Asymmetric transitivity preserving graph embedding,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1105–1114.
- [182] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *nature*, 2005.
- [183] P. Panzarasa, T. Opsahl, and K. M. Carley, “Patterns and dynamics of users’ behavior and interaction: Network analysis of an online community,” *Journal of the American Society for Information Science and Technology*, 2009.
- [184] A. Paranjape, A. R. Benson, and J. Leskovec, “Motifs in temporal networks,” in *Proceedings of the tenth ACM international conference on web search and data mining*, 2017, pp. 601–610.
- [185] R. Parshani, S. Carmi, and S. Havlin, “Epidemic threshold for the susceptible-infectious-susceptible model on random networks,” *Physical review letters*, vol. 104, no. 25, p. 258701, 2010.
- [186] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, “Using graph theory to analyze biological networks,” *BioData mining*, vol. 4, no. 1, pp. 1–27, 2011.
- [187] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, 2011.
- [188] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *KDD*, 2014.
- [189] B. L. Perry, B. A. Pescosolido, and S. P. Borgatti, *Egocentric network analysis: Foundations, methods, and models*. Cambridge university press, 2018.

- [190] N. Pržulj, D. G. Corneil, and I. Jurisica, “Modeling interactome: scale-free or geometric?” *Bioinformatics*, 2004.
- [191] S. Qasim, “Some problems related to the food chain in a tropical estuary,” *Marine food chains*, 1970.
- [192] E. Rainone, “The network nature of over-the-counter interest rates,” *Journal of Financial Markets*, 2020.
- [193] A. Rapoport, “Spread of information through a population with socio-structural bias: I. assumption of transitivity,” *The bulletin of mathematical biophysics*, 1953.
- [194] J. L. P. Ribeiro, “Escala de satisfação com o suporte social (esss),” 1999.
- [195] P. Ribeiro, P. Paredes, M. E. Silva, D. Aparicio, and F. Silva, “A survey on subgraph counting: concepts, algorithms and applications to network motifs and graphlets,” *arXiv preprint arXiv:1910.13011*, 2019.
- [196] P. Ribeiro and F. Silva, “G-tries: an efficient data structure for discovering network motifs,” in *Proceedings of the 2010 ACM symposium on applied computing*, 2010, pp. 1559–1566.
- [197] —, “Discovering colored network motifs,” in *Complex Networks V*. Springer, 2014.
- [198] H. Risselada, P. C. Verhoef, and T. H. Bijmolt, “Indicators of opinion leadership in customer networks: self-reports and degree centrality,” *Marketing Letters*, vol. 27, no. 3, pp. 449–460, 2016.
- [199] F. A. Rodrigues, “Network centrality: an introduction,” in *A mathematical modeling approach from nonlinear dynamics to complex systems*. Springer, 2019, pp. 177–196.

- [200] A. Rosenberg and J. Hirschberg, “V-measure: A conditional entropy-based external cluster evaluation measure,” in *EMNLP-CoNLL*, 2007.
- [201] R. A. Rossi, N. K. Ahmed, A. Carranza, D. Arbour, A. Rao, S. Kim, and E. Koh, “Heterogeneous graphlets,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 1, pp. 1–43, 2020.
- [202] B. Rozemberczki, C. Allen, and R. Sarkar, “Multi-scale attributed node embedding,” *arXiv preprint arXiv:1909.13021*, 2019.
- [203] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou *et al.*, “Towards a proteome-scale map of the human protein–protein interaction network,” *Nature*, 2005.
- [204] A. Said, R. A. Abbasi, O. Maqbool, A. Daud, and N. R. Aljohani, “Cc-ga: A clustering coefficient based genetic algorithm for detecting communities in social networks,” *Applied Soft Computing*, vol. 63, pp. 59–70, 2018.
- [205] C. Salavati, A. Abdollahpouri, and Z. Manbari, “Bridgerank: A novel fast centrality measure based on local structure of the network,” *Physica A: Statistical Mechanics and its Applications*, vol. 496, pp. 635–653, 2018.
- [206] M. Salehi, R. Sharma, M. Marzolla, M. Magnani, P. Siyari, and D. Montesi, “Spreading processes in multilayer networks,” *IEEE Transactions on Network Science and Engineering*, vol. 2, no. 2, pp. 65–83, 2015.
- [207] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertesz, “Generalizations of the clustering coefficient to weighted complex networks,” *Physical Review E*, 2007.
- [208] D. Schall, “Link prediction in directed social networks,” *Social Network Analysis and Mining*, vol. 4, no. 1, p. 157, 2014.

- [209] C. Seshadhri, A. Pinar, N. Durak, and T. G. Kolda, “Directed closure measures for networks with reciprocity,” *Journal of Complex Networks*, 2017.
- [210] S. Siegel, “Nonparametric statistics for the behavioral sciences.” 1956.
- [211] R. W. Solava, R. P. Michaels, and T. Milenković, “Graphlet-based edge clustering reveals pathogen-interacting proteins,” *Bioinformatics*, 2012.
- [212] T. H. Stark and J. A. Krosnick, “Gensi: A new graphical tool to collect ego-centered network data,” *Social Networks*, 2017.
- [213] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen *et al.*, “A human protein-protein interaction network: a resource for annotating the proteome,” *Cell*, 2005.
- [214] K. Stephenson and M. Zelen, “Rethinking centrality: Methods and examples,” *Social networks*, vol. 11, no. 1, pp. 1–37, 1989.
- [215] X. Su and T. M. Khoshgoftaar, “A survey of collaborative filtering techniques,” *Advances in artificial intelligence*, 2009.
- [216] L. Šubelj and M. Bajec, “Model of complex networks based on citation dynamics,” in *Proceedings of the 22nd international conference on World Wide Web*, 2013.
- [217] X. Tang, J. Wang, J. Zhong, and Y. Pan, “Predicting essential proteins based on weighted degree centrality,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pp. 407–418, 2013.
- [218] D. Taylor, S. A. Myers, A. Clauset, M. A. Porter, and P. J. Mucha, “Eigenvector-based centrality measures for temporal networks,” *Multiscale Modeling & Simulation*, vol. 15, no. 1, pp. 537–574, 2017.

- [219] L. ter Haar-Pomp, C. de Beer, R. van der Lem, M. Spreen, and S. Bogaerts, “Monitoring risk behaviors by managing social support in the network of a forensic psychiatric patient: A single-case analysis,” *Journal of Forensic Psychology Practice*, 2015.
- [220] S. Teso, J. Staiano, B. Lepri, A. Passerini, and F. Pianesi, “Ego-centric graphlets for personality and affective states recognition,” in *SocialCom*. IEEE, 2013.
- [221] J. Tolles and T. Luong, “Modeling epidemics with compartmental models,” *Jama*, vol. 323, no. 24, pp. 2515–2516, 2020.
- [222] T. Trolliet, N. Cohen, F. Giroire, L. Hogue, and S. Pérennes, “Interest clustering coefficient: a new metric for directed networks like twitter,” in *International Conference on Complex Networks and Their Applications*. Springer, 2020, pp. 597–609.
- [223] R. E. Ulanowicz and D. L. DeAngelis, “Network analysis of trophic dynamics in south florida ecosystems,” *US Geological Survey Program on the South Florida Ecosystem*, 1999.
- [224] M. Van Alboom, L. De Ruddere, S. Kindt, T. Loeys, D. Van Ryckeghem, P. Bracke, M. M. Mittinty, and L. Goubert, “Well-being and perceived stigma in individuals with rheumatoid arthritis and fibromyalgia: A daily diary study,” *The Clinical Journal of Pain*, 2021.
- [225] M. Von Korff, J. Ormel, F. J. Keefe, and S. F. Dworkin, “Grading the severity of chronic pain,” *Pain*, 1992.
- [226] D. Wang and X. Zou, “A new centrality measure of nodes in multilayer networks under the framework of tensor computation,” *Applied Mathematical Modelling*, vol. 54, pp. 46–63, 2018.

- [227] J. Wang and J. Cheng, “Truss decomposition in massive networks,” *arXiv preprint arXiv:1205.6693*, 2012.
- [228] J. Wang, M. Li, H. Wang, and Y. Pan, “Identification of essential proteins based on edge clustering coefficient,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2011.
- [229] J. Wang, X. Hou, K. Li, and Y. Ding, “A novel weight neighborhood centrality algorithm for identifying influential spreaders in complex networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 475, pp. 88–105, 2017.
- [230] Y. Wang, G. Yan, Q. Ma, Y. Wu, and D. Jin, “Identifying influential spreaders on weighted networks based on clusterrank,” in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2. IEEE, 2017, pp. 476–479.
- [231] S. Wasserman, K. Faust *et al.*, “Social network analysis: Methods and applications,” 1994.
- [232] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *nature*, 1998.
- [233] D. R. White and S. P. Borgatti, “Betweenness centrality measures for directed graphs,” *Social networks*, vol. 16, no. 4, pp. 335–346, 1994.
- [234] Z.-X. Wu and P. Holme, “Modeling scientific-citation patterns and other triangle-rich acyclic networks,” *Physical review E*, 2009.
- [235] J. Xie, S. Kelley, and B. K. Szymanski, “Overlapping community detection in networks: The state-of-the-art and comparative study,” *Acm computing surveys (csur)*, vol. 45, no. 4, pp. 1–35, 2013.

- [236] W. Xing and A. Ghorbani, “Weighted pagerank algorithm,” in *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004*. IEEE, 2004, pp. 305–314.
- [237] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” *arXiv preprint arXiv:1810.00826*, 2018.
- [238] S. Xu and P. Wang, “Identifying important nodes by adaptive leaderrank,” *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 654–664, 2017.
- [239] Y. Yang and H. Grol-Prokopczyk, “Chronic pain and friendship among middle-aged and older us adults,” *The Journals of Gerontology: Series B*, 2020.
- [240] H. Yin, A. R. Benson, and J. Leskovec, “Higher-order clustering in networks,” *Physical Review E*, vol. 97, no. 5, p. 052306, 2018.
- [241] —, “The local closure coefficient: A new perspective on network clustering,” in *WSDM*, 2019.
- [242] H. Yin, A. R. Benson, and J. Ugander, “Measuring directed triadic closure with closure coefficients,” *arXiv*, 2019.
- [243] —, “Measuring directed triadic closure with closure coefficients,” *Network Science*, vol. 8, no. 4, pp. 551–573, 2020.
- [244] B. Zhang and S. Horvath, “A general framework for weighted gene co-expression network analysis,” *Statistical applications in genetics and molecular biology*, 2005.
- [245] C.-T. Zhang, “The e-index, complementing the h-index for excess citations,” *PLoS One*, vol. 4, no. 5, p. e5429, 2009.

- [246] H. Zhang, M. Fiszman, D. Shin, C. M. Miller, G. Rosembat, and T. C. Rindfleisch, “Degree centrality for semantic abstraction summarization of therapeutic studies,” *Journal of biomedical informatics*, vol. 44, no. 5, pp. 830–838, 2011.
- [247] L. Zhang, M. Song, Z. Liu, X. Liu, J. Bu, and C. Chen, “Probabilistic graphlet cut: Exploiting spatial structure cue for weakly supervised image segmentation,” in *CVPR*, 2013.
- [248] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, “Clustering coefficient and community structure of bipartite networks,” *Physica A: Statistical Mechanics and its Applications*, 2008.
- [249] Q.-M. Zhang, L. Lü, W.-Q. Wang, T. Zhou *et al.*, “Potential theory for directed networks,” *PloS one*, vol. 8, no. 2, p. e55437, 2013.
- [250] X. Zhang, J. Zhu, Q. Wang, and H. Zhao, “Identifying influential nodes in complex networks with community structure,” *Knowledge-Based Systems*, vol. 42, pp. 74–84, 2013.
- [251] X. Zhang, C. Zhao, X. Wang, and D. Yi, “Identifying missing and spurious interactions in directed networks,” *International Journal of Distributed Sensor Networks*, 2015.
- [252] Y. Zhao, S. Li, and F. Jin, “Identification of influential nodes in social networks with community structure based on label propagation,” *Neurocomputing*, vol. 210, pp. 34–44, 2016.
- [253] T. Zhou, Y.-L. Lee, and G. Wang, “Experimental analyses on 2-hop-based and 3-hop-based link prediction algorithms,” *Physica A: Statistical Mechanics and its Applications*, vol. 564, p. 125532, 2021.

- [254] T. Zhou, L. Lü, and Y.-C. Zhang, “Predicting missing links via local information,” *The European Physical Journal B*, 2009.
- [255] T. Zhou, G. Yan, and B.-H. Wang, “Maximal planar networks with large clustering coefficient and power-law degree distribution,” *Physical Review E*, vol. 71, no. 4, p. 046141, 2005.