Efficient Community Detection

Xufei Sun

A subthesis submitted in partial fulfillment of the degree of
Bachelor of Advanced Computing (Honours) at
The Department of Computer Science
Australian National University

August 2015

© Xufei Sun

Typeset in Palatino by TEX and LATEX 2ε.

Except where otherwise indicated, this thesis is my own original work.

Xufei Sun
7 August 2015

This work is dedicated to my parents and family, who make the strongest support behind me mentally and economically, during not only the tough individual research period, but the whole student life. Thank you.

It's also for my friends, without whom the extreme difficulties, loneliness and my weakness could eat me alive half way, make this dream come true. Yes I finally made it, and I'm very, very appreciated for everything on this journey. The submission of this thesis eventually ends this fantastic life stage of mine though, the friendship and memory with you guys, the most precious wealth I gained, accompanies me forever. Special thanks to Xindi Li, Yanning Liu and Renjing Xu.

Finally, I would like to literally kiss and hug myself for one thousand times. This work as well as all the challenges overcome and experience gained along the way, is dedicated to myself. And this individual research, especially the last few months of it, is the hardest thing I've ever attempted or seen. But you make it! You make it! You make it!

Xufei Sun

7 August, 2015

Acknowledgements

I would like to give my most sincere thank to my supervior, Weifa. It's his kindness, strictness and experience that makes this work possible.

Abstract

Given a large network, local community detection aims at finding the community that contains a set of query nodes and also maximises (minimises) a goodness metric. Furthermore, due to the inconvenience or impossibility of obtaining the complete network information in many situations, the detection becomes more challenging.

This problem has recently drawn intense research interest. Various goodness metrics have been proposed. And most of them base on the statistical features of community structures, such as the internal density or external spareness. However, the metrics often result in unsatisfactory results by either including irrelevant subgraphs of high density, or pulling in outliers which accidentally match the metric for the time being. Further more, when in a highly overlapping environment such as social networks, the unconventional community structures make these metrics usually end up with a quite trivial detection result.

In our work, we go for a alternative point of view on the formation of the communities, namely the assembly of nodes with different roles in the structure. With the new view point, we present two metrics which are proved to perform superiorly in traditional and complex environment respectively. Moreover, on realising a single metric is whatsoever limited in effectiveness as well as scope of application, we raise up a complete framework for the collaboration of metrics in the field, which also lands a base-stone for future innovations.

The experiment results collected from Amazon, DBLP, Youtube and LivingJournal well certifies the effectiveness of the metrics.

Contents

Chapter 1

An Introduction to My Thesis

1.1 Networks, Graphs and Community Structures

1.1.1 Networks in Real Life

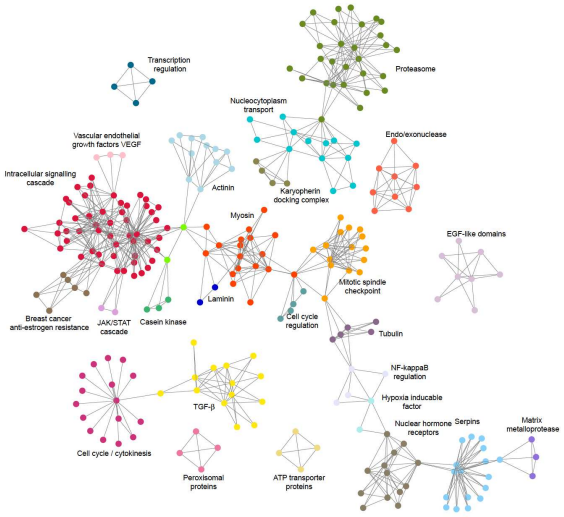

Figure 1.1: Protein-Protein Network, reprinted from [Jonsson PF 2006]

When we talk about network, we refer to the concept of a system with a set of components interacting or interdependent with each other, forming an integrated whole.

A basic knowledge is, where there is any connection or interaction between multiple entities, there is a network. And networks are all around us in life.

For example, the figure 1.1, 1.2, 1.3 respectively notates a network of different kind. Figure 1.1 describes a protein-protein interaction networks. The graph pictures the interactions between proteins in cancerous cells of a rat. Figure 1.2 visualise the data of internet at the AS level. Figure 1.3 depicts the relationship network between two families.

It's clear that these connections are the bonds that integrate the whole networks. And the study into these connections make great sense if we would like more infor-

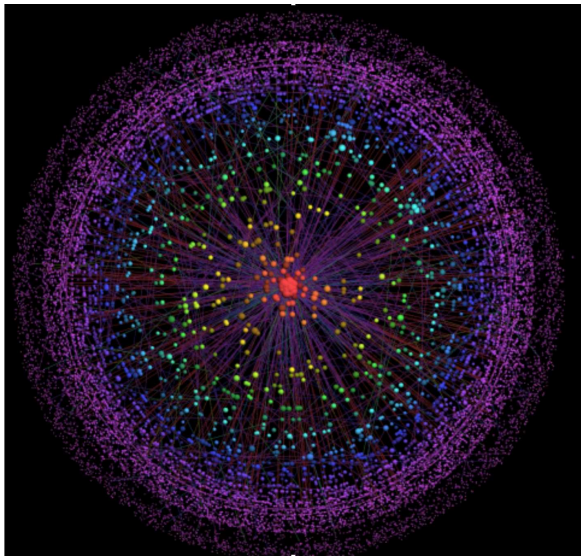

Figure 1.2: Internet Network, reprinted from [Shai Carmi 2007]

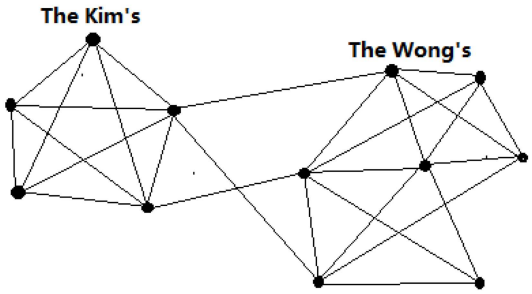

Figure 1.3: People Relationship Network

mation about the networks all around us.

1.1.2 Networks and Graphs

Specially, the networks usually can be studied as graphs.

In order to study the networks in depth, an adequate representative form of it is indispensable. And this is where the graph theory comes in. The origin of graph theory dates back to 1736. As the amount of understanding and knowledge on graphs obtained increases, especially their mathematical properties [Bollobás 1998], they gradually become integral in network researches, as representation of a wide variety of networks in different areas.

Especially, as shown with the example in last section, biological, social, technological, and information networks can be studied as graphs; and the graph analysis skills have become crucial to understand the features of these networks. For instance, social network analysis started in the 1930's and has become one of the most important

topics in sociology [Scott 2000]

In recent times, the age of information explosion, the technology dedicated in collecting and storing data of unbelievable size evolves at an amazon speed. The human kind is now capable of accessing seas of data which was like a dream. However, this need to deal with such a large number of units, in the order of millions or billions, has produced a deep change in the way graphs are approached [Albert and Barabasi 2002]

Before move on, we are going to formalise definitions for networks and graphs, for the convenience of demonstration in the rest part.

Definition 1. *Network, is a system in which the entities connect, interact and influence each other in one way or another. Network would be only referred to as the existed networks in our context.*

Definition 2. *Graph is a representation of a set of objects where some pairs of objects are connected by links.*

1.1.3 Graphs and Basic Concepts

Graphs are essentially data sets stored in the graph database. And now we are going to give a more official definition to them and talk about the details they concerns.

Definition 3. *Vertex (also called a 'node') is a fundamental part of a graph. In many applications the vertex is behind the entity in the real world.*

The vertexes can have a name, which we will call the ?key?.

In many applications the vertex is behind the entity in the real world. For example, in social network like Facebook, one vertex is usually representing a single user in real life; and for amazon commodity network, a node is often associated with a specific good. Besides that, a vertex may also have additional information. We will call this additional information the ?payload.? The quantity of the vertexes belong to a particular graph stands for the amount of concerned objects in it. For example the node number of a regional Facebook graph represents the Facebook user account amount in the area.

Definition 4. *Edge (also called an 'arc') is another fundamental part of a graph. An edge connects two vertices to show that there is a relationship between them in the graph.*

Edges are radically reflections of connections in networks to graphs. They may be one-way(undirected) or two-way(directed). If the edges in a graph are all two-way, we say that the graph is a directed graph, or a digraph. The class prerequisites graph shown above is clearly a digraph since you must take some classes before others. For instance, an undirected edge would be established between two users in Facebook as long as they friend each other on the website; and two goods may have a directed edge in between if one is relied upon another according to some rule. While any edge has to indicate explicitly the two nodes it sit between, sometimes it contains more information.

The quantity of edges shows the amount of existing connections between the nodes(objects) of the same graph. For example, the edge number of a company co-operation graph expresses the level of collaborating positiveness between the companies.

Definition 5. *The fundamental parts of the graph, namely nodes and edges, may be attached with a certain value demonstrate the difference between two in the graphs, which is called weight.*

Particularly, edges can be weighted to demonstrate a key attribute of the connections. For example, the attribute can be the cost to go from one vertex to another; as in a graph of roads that connect one city to another, the weight on the edge might represent the distance between the two cities. Or the length or strength(based on some particular rule) of a friendship on Facebook.

Some of the graphs also contain nodes with different weights. This kind of weighting usually showcases the nodes? fit against a given standard. For example, in the PageRank graph, the higher probability(weight) a website(node) with has, the bigger impact it has on the rest of the graph.

Definition 6. *The degree of a node is the number of edges connected to the node in the graph.*

As far as the graph is concerned, the degree of a node is the number of edges attached to it. And this also mirrors the relational density level of entities in the networks. Take social networks as examples, the vertex with very high node degree is, without doubt, playing a core role in the group under normal circumstances. At the same time, the node with few connections can be generally viewed as social-inactive.

1.1.4 The Existence of Community Structure

Definition 7. *Random graph: The paradigm of disordered graph is the random graph, introduced by P. Erdos and A. Renyi in [Erdos and I 1959]. In a random graph, the distribution of edges among the vertices is highly homogeneous. For instance, the distribution of the number of neighbours of a vertex, or degree , is binomial, so most vertices have equal or similar degree. In it, the probability of having an edge between a pair of vertices is equal for all possible pairs*

. There are some disciplines in the formation of real life networks, which means that the graphs representing real networks are not usually as regular as frames. Graphs are objects where order coexists with disorder. And they are not random graphs, for they do need to express some level of order and organisation to represent the inhomogeneities in real-life networks.

The degree distribution of these graphs is broad, with a tail that often follows a power law: therefore, many vertices with low degree coexist with some vertices with large degree. Furthermore, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups.

Definition 8. *Community(networks): Communities, also called clusters or modules , are groups of vertices which probably share common properties and/or play similar roles within the graph; usually marked with high concentrations of edges within these groups of vertices, and low concentrations in between*

.

If we have a look at the human society, we can find many organisations of clear order, such as military, governments, nations, schools, towns and friend circles. In recent years the online communities have made its appearance in organised groups as well.

Social communities have been studied for a long time [R. Edward Freeman 2004]. And communities also occur in many networked systems from biology, computer science, engineering, economics, politics, etc. If we have a look back at the network example given at the first section, we may notice the existence of communities as well, in the form of protein sub-structures, internet communities and families.

1.1.5 Other Special Concepts in Graphs Their Realistic Significance

For the rest of the work, we will give a formal definition and some notations to describe the graph. A graph can be represented by G where G=(V,E). For the graph G, V is a set of vertices and E is a set of edges. Each edge is a tuple (v,w) where w,v?V. We can add a third component to the edge tuple to represent a weight. G=(V,E,W)

Definition 9. *Subgraph : A subgraph, H, of a graph, G, is a graph whose vertices are a subset of the vertex set of G, and whose edges are a subset of the edge set of G. In reverse, a supergraph of a graph G is a graph of which G is a subgraph. A graph, G, contains a graph, H, if H is a subgraph of, or is isomorphic to G.*

Apart from all above, the example graph in the example helps illustrate some other key terms of concern:

Definition 10. *Path in a graph is a sequence of vertices that are connected by edges.*

Formally we would define a path as W1,W2,...,Wn such that (Wi,Wi+1)?E for all 1?i?n?1. The unweighted path length is the number of edges in the path, specifically n?1. The weighted path length is the sum of the weights of all the edges in the path. For example in Figure 1.4 the path from node i to node b is the sequence of vertices (i,j,f,b). The edges are (i,j),(j-,f),(f-b).

Definition 11. *Cycle in a graph is a path that starts and ends at the same vertex*

For example, in Figure 1.4 the path (i,j,f,h) is a cycle. A graph with no cycles is called an acyclic graph.

Definition 12. *Structural similarity structural equivalence: Structural equivalence describes the extent two nodes being similar to each other in the sense of inter-node structure. In another word, structural similarity shows how much two nodes are alike based on their interactions with each other and the rest of graph.*

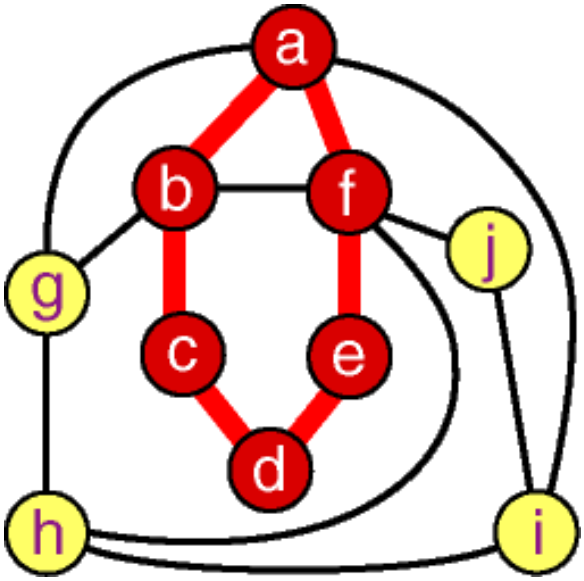

Figure 1.4: node a,b,c,d,e,f and the edges in between make up of a subgraph of the whole graph presented; On the other the entire graph is the super-graph of the graph made up of node a,b,c,d,e,f and the edges in between

Take the graph displayed below an example. We might try to assess which nodes are most similar to which other nodes intuitively by looking at a graph. We would notice some important things. It would seem that actors 2,5, and 7 might be structurally similar in that they seem to have reciprocal ties with each other and almost everyone else. Actors 6, 8, and 10 are "regularly" similar in that they are rather isolated; but they are not structurally similar because they are connected to quite different sets of actors. But, beyond this, it is really rather difficult to assess equivalence rigorously by just looking at a diagram.

Definition 13. *Density:the density of a graph(subgraph) is the measurement demonstrating how close the number of edges inside it is to the maximal number of edges for the same set of nodes.*

Intuitively, a dense graph is a graph with high density, while on the contrast, a graph with only a few edges, is a sparse graph. The distinction between sparse and dense graphs is rather vague, and depends on the context.

For undirected simple graphs, the graph density is defined as:

$$D = \frac{2 * |E|}{|V| * (|V| - 1)}$$

For directed simple graphs, the graph density is defined as:

$$D = \frac{|E|}{|V| * (|V| - 1)}$$

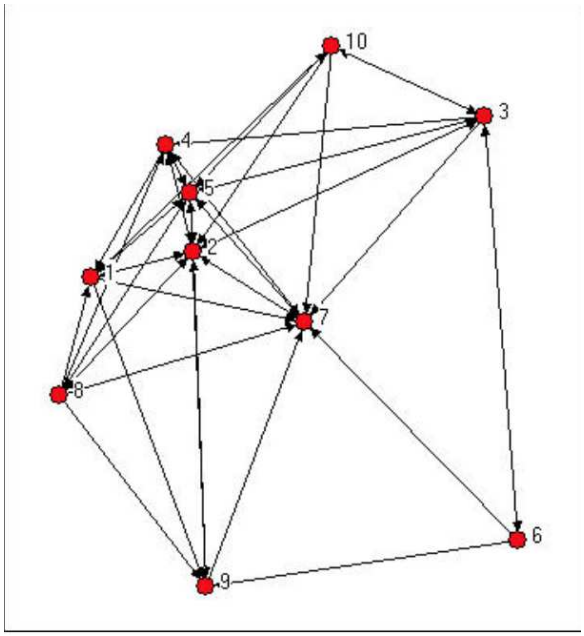

Figure 1.5: Knoke directed information network [Kno]

where E is the number of edges and V is the number of vertices in the graph. The maximum number of edges is —V— * (—V— - 1), so the maximal density is 1 (for complete graphs) and the minimal density is 0 [Coleman and Moré 1983].

Definition 14. *Isolation: Isolation expresses to which extent does a set of nodes(subgraph) in the graph is separated from the rest part, by means of edge amount between nodes within and out*

1.2 Community Structures

1.2.1 key features of the community

Community is a widely accepted existence in real-world networks; however in the context of graph, there is still no formal definition of it that is universally accepted; As a matter of fact, the definition often depends on the specific system at hand and/or application one has in mind.

On the other hand, we do have some concrete requirements that any community should satisfy. Firstly, there must be more edges inside the community than edges linking vertices of the community with the rest of the graph. This is the reference guideline at the basis of most community definitions. Besides that, the community has to be connected. It?s obvious that the combination of two disjoint subgraph would hardly make a high-edge-concentration component. And even if they do, the separated structures would probably make more sense.

The communities in the graphs are believed to exist as reflection of entities' grouping and dense interaction in real world. Hence it makes much sense that the cor-

responding nodes to entities belong to the same group in real life, sit in the same communities, share common properties and/or play similar roles in the graphs.

And the picture above showcases some common community structures in networks.

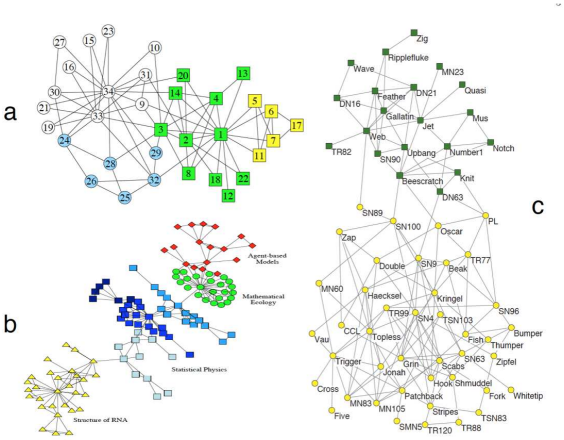

Figure 1.6: Community structure in social networks [Fortunato 2010]

High (Edge) Density Majority of the statistics definition of the community concerns fully or mostly about the edge density. As described in section 1.4, the nodes inside a same community are expected to have more connections between each other. Hence the edge density of the community is one of the most outstanding features of the community. Especially, a subgraph of high edge density is usually found with the existence of considerable cliques(subgraphs in which every node is connected to every other node in the clique). The example is the amazingly huge amount of connections between the students from the same college of the same university on social networks.

High Isolation Similar to the high density feature of the communities, according to formation of them in the graphs, the edge amount is expected to be relatively low in between the communities. The idea is intuitive that if a node has a lot of edges associated with the community, the chance is big that it belong to the community as well. Admitted that the node may have even more connections with the outside nodes, the choice of leaving it out of the community still help make a relatively higher isolation(less edges across the boarder). The example is that a collection of Chinese speakers(community) should have much lower connections with the speaker of other languages than with the other native Chinese speaking people.

Sparse Inner Shortest Path The paths that connect vertices of distinct communities must pass through at least one inter-cluster edge. Bearing in mind the fact that the communities are loosely connected, one can expect that the inter-cluster edges usually included in a rather big amount of shortest paths between node pairs. On the other hand, the vertices within a community are tightly connected, so the intra-cluster edges are associated with smaller number of shortest path between node pairs. [Network

community-detection enhancement by proper weighting]. For example, if one needs to contact someone of the same community, he might find multiple ways to get the information about the target from very different people/channels. However, if the target belongs to another community no matter who exactly he is, the chance is that there is a couple of people you have to contact to retrieve the information of the target.

Cycle Existence The basic intuition of cycle is similar to that of the shortest path: the nodes in a same community are expected to be connected in multiple ways. And the representation from the graph perspective of the intuition is the existence of the cycle (multiple cycles even) between the nodes in the cluster. The example is, when one needs to contact another belong to the same college, even if he is not able to get in touch with someone who can help him for sure, he will probably succeed in the task through another trail.

Structural Similarity The nodes in a same community should be similar to each other structurally, which means they share a great number of mutual friends. The example is that two students in the same college are expected to have relatively high rate of shared friends(other students in the college). Here the students of the college form the whole community, and the property should be reflected clearly in the graph.

Multi-hierarchy Not all communities are equal or of a same hierarchy: a community can contain sub-communities, or be contained by super-communities. The hierarchy is an organisation of actors in some latent space learned from the observed network. And an entity may belong to a series of these communities at the same time. This is a natural phenomenon for clustering and hence, for communities hidden in the graph databases [Qirong Ho 2012]. Many networks in real life exhibits hierarchy. For example, cold-blooded animals and mammals are large super-communities that can be sub-divided into smaller sub-communities, such as sharks and squid, or toothed whales and pinnipeds. These sub-communities can in turn be divided into even smaller communities (not shown).

Overlapping Since in most networks a single node is allowed (and very usually seen) to be a part of multiple communities at the same time and the communities may not always contain each other, it often appears that very different community pair shared a fraction of the common nodes, also referred to as the overlapping communities. Example are everywhere: I am a member of the college computer science association and am a part of the racing club as well. Overlapping is one of the most outstanding disturbing but ubiquitous features of the communities in the graph.

Definition 15. *Traditional community structural model, a traditional community is one that obeys the traditional view on the formation of a community, namely a subgraph within the graph with high density and high isolation value.*

Definition 16. *Traditional graph model: a traditional community is one that obeys the traditional view on the formation of a graph, namely a graph in which high concentration of edge signals the existence of community structures while low in it stands for the part is in between community structures.*

According to the definition of community, we may visualise a typical community structure as figure 1.7. For convenience but without losing generality, more connected

the node is to the outside, the the closer to the edge it would be located in the picture. And we place the node with all adjacencies within the structure in the centre. Further more, the nodes most connected externally (speaking of its edge distribution percentage, same with the following) form the boundary; while the nodes most internally connected make up of the core of the community structure. The definition is qualitative, and for the purpose of demonstration only.

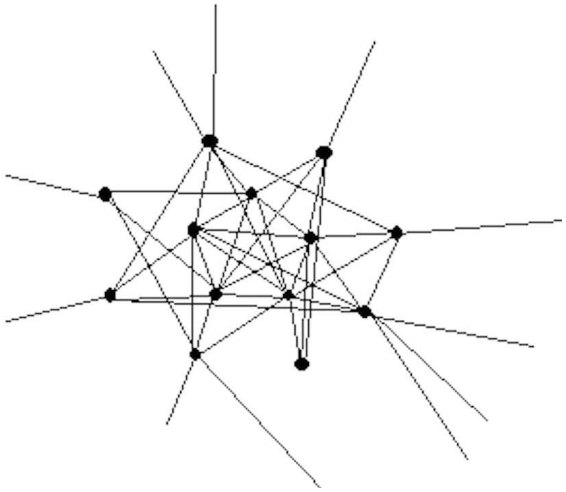

Figure 1.7: traditional community structure model

When a community is barely interacting with and very slightly influenpaced by the outside structures or nodes like this, we say the community is a traditional community structure. And with the features of a community, namely the high edge density within and low conductance with the rest graph given above, the nodes on the edges or the furthermost boundaries of the typical communities, is very likely to have a higher internal edge amount than external. Otherwise, due to the internal thin connection and unusual high conductance , the node might actually belong to some other communities. For example, under traditional graph model, the node A in figure 1.8 is very likely to be actually a part of the community made up by blue nodes.

Besides the model of subgraph analysed above(conductance of a subgraph can measure how well it is separated from the remaining graph), [Yubao Wu 2015] mentioned another common-used model, where local community is separated from the remaining graph by a set of low degree nodes. That is to say, the boundary nodes of the communities tend to have low degrees.

The graphs of Amazon and DBLP well represent the model.

1.2.2 why is community important and its application

As described above, the information of connection distribution in graphs have good reason to be looked into in detail. The graphs, especially unweighted indirected ones of them, are often suffering from the lack of expressive capability in picturing the real network. However, the network itself, especially the connections within, have

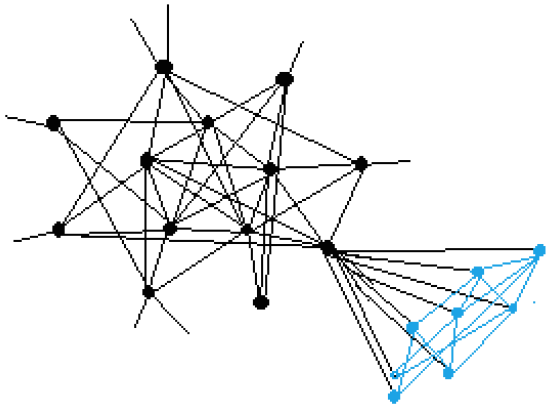

Figure 1.8: Outside Node included

always important information whenever we try to manipulate with it. Communities are among the hidden information of the graphs, and contain the potential of help graph express itself better.

Hence communities can have concrete applications. Clustering Web clients who have similar interests and are geographically near to each other may improve the performance of services provided on the World Wide Web, in that each cluster of clients could be served by a dedicated mirror server [Wang 2000].

In the network of purchase relationships between customers and products of online retailers (like, e. g., www.amazon.com), getting the idea of the customer groups with similar interests to a large degree help improve the efficient recommendation systems[Agrawal 2011]. And the application is capable of guiding them through the list of items of the retailer and enhancing the business opportunities.

Other than that, the clusters of large graphs can also be used to create data structures in order to generate compact touting tables while the choice of the communication paths is still efficient [clu 2001]. Identifying modules and their boundaries at the same time allows for a classification of vertices, according to their structural position in the modules(this idea is fundamental idea of the node-centric point of view on community structure to be mentioned). So, vertices with a central position in their clusters, i. e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group; vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities.

Another important aspect related to community structure is the hierarchical organisation displayed by most networked systems in the real world. Real networks are usually composed by communities including smaller communities, which in turn include smaller communities, etc. The human body offers a paradigmatic example of hierarchical organisation: it is composed by organs, organs are composed by tissues, tissues by cells, etc. One example is represented by business firms, who are characterised by a pyramidal organisation, going from the workers to the president, with

intermediate levels corresponding to work groups, departments and management. Herbert A. Simon has emphasised the crucial role played by hierarchy in the structure and evolution of complex systems [Simon 1991]. The generation and evolution of a system organised in interrelated stable subsystems are much quicker than if the system were unstructured, because it is much easier to assemble the smallest subparts first and use them as building blocks to get larger structures, until the whole system is assembled. In this way it is also far more difficult that errors (mutations) occur along the process.

1.3 what iscommunity detection and its application

As shown in the previous section, the community structure is of great use in many ways however usually implicit. The knowledge about the community structure in a graph requires the process of community detection, which is about exploration, extraction and analysis on the graph.

Community detection in graphs is the process of identifying the communities and, possibly, their hierarchical organisation, by only using the information encoded in the graph topology [Fortunato 2010].

When the graph is constrained to be of a small size, containing as many as tens of nodes, the detection process is trivial and fast. Merely with visualised presentation(with software like Graphviz) of the data set and human eyes can one identify the community structures at this scale in no time. However, as the size of data sets in graph database starts to rocket, the detection method evolved accordingly.

At this time and age, the community detection normally indicates the process of inputing the edge information (sometimes with node information) to the computing device, and receiving the formation of community(s) after a series of analysis and computation steps.

Especially, current community detection methods can be roughly classified into two categories, namely global ones and local ones.

Definition 17. *global community detection, is a school of community detection methods dedicated in identifying community structures underlying in the graphs,with always a reference to the whole graph. That means, the global methods needs consideration on the impact on the other parts of graph when determine wether if a subgraph is a community. A representative one belonging to this type is the algorithm of graph partitioning*

Definition 18. *local community detection is a school of community detection methods aiming at identifying the community structures in the graphs by identifying the formative features of subgraphs(how much this subgraph is like a community). Local methods make the decision with only concern on the properties of current subgraph. The method in this category is normally greedy node addition or greedy node deletion*

.

1.4 Motivations

1.4.1 Current Challenges

The area of community detection has been studied for several decades, being a particularly intensive interest in recent years. After decades of exhaustive study and experiments, a couple of detection methods with stable and robust performance has already been raised up in the field. However, the new and big challenges keep coming, causing problems to existed solutions. The major challenges are as below:

1.4.1.1 Retrieving and storing the information of the entire graph

When the node or edge amount in graph data set are frequently on the order of millions or even billions, you will notice the data retrieving and storage becomes a major concern in the process of community detection. In this time of big data the networks are often too large to comprehend and even a simple visualisation of the network is often impossible.[Fast community detection using local neighbourhood search]. This constraint is problematic for networks like the World Wide Web, which for all practical purposes is too large and too dynamic to ever be known fully, or networks which are larger than can be accommodated by the fastest algorithms [Clauset 2005].

Moreover, due to the computational complexity and network bandwidth, this action of retrievemcent from time to time take much more time than the community detection itself, causing an enormous challenge for problems with strict time constraint.

1.4.1.2 The rapid change on the graphs

As described in section 1, the graphs are essentially data sets saved in graph databases. Along with the fast growth in size, the rapid alterations on the existed data sets are causing harsh problems at the same time. It is just too hard to get whole knowledge of the networks evolving quickly or being too big ,such as the Internet [Tiantian Zhang 2012].

1.4.1.3 The problems with seeds

For most of the networks, a large proportion of it is out of our consideration ,we only care about the community structure or other statistical characteristics of some specific nodes. For example, in the book selling network, the purchaser only need the knowledge of books related to some subject, namely the book community which a certain book is in. Hence in the circumstance alike, it would be more appropriate for community detection method to behave as community relegation algorithm of particular nodes of interest [Tiantian Zhang 2012].

1.4.1.4 Dealing with the massive information

Given that the information of the entire graph is made clear and saved locally, to identify the communities within remains a challenge due to the computational complexity

on the massive data. That is also because of most of the detection methods described above (or ever existed) are products from the awareness and manipulations on the complete graph.

For example, finding the optimal partition for a given cost function is in general a difficult problem. Especially, maximising the modularity has been proven to be NP-hard [Brandes 2008]. Hence, different algorithms have been developed to approximate the optimal partition of a network. All the existing heuristics designed to extract community structures have to balance the quality of the partition with respect to the time complexity of the algorithm.

When it comes to the community detection tasks with pre-set query node sets, the challenge can get even harder. Due to the attempt to include all the seeds into one community and the intuition of executing the whole algorithm multiple times until the goal is achieved, the detection process would be rather computationally intensive for its pre-requirement of detecting all communities, especially for large scale networks [Kwan Hui Lim 2013].

1.4.1.5 The limitation on ability of graphs to demonstrate relationships

Among all the challenges given above, the the limitation of graph database itself is the most serious and almost unsolvable. The act of the graphs to express relationship as edges, while the properties of the relationships as the features given to the edges, is effective to some degree and the only way. However, it can also be quite problematic, so as that under many occasions, the graphdoesn?t represent the relationships(such as friendships /collaboration / chemical reactions) very well.

Furthermore, most current models make the assumption that networks are essentially some variation of a random graph, while we know that real networks are far from random on every level, e.g., certain motifs are much more likely than others [Santo Fortunato 2012].

1.4.2 Our Contribution

Our work dedicates in seed-centric local community detection. Firstly, on the basis of traditional graph model and community model, we raise up an alternative view on the community structures which well explains the state-of-art metric in local detection methods; Secondly, with the new view of point on the community formation in mind, we design a new metric(core-seeker metric), trying to help get rid of the outstanding issues in the process, namely outliers and free rider effect; thirdly, targeted at the severely invalid detecting result on multi-layer graphs, we put forward a new community structure theory as well as graph theory based on empirical observations; we then design another metric dedicated in the community identification in this overlapping environment(boundary-seeker); last, we propose our SLUD framework, which well describes the collaboration of metrics and formalises the majority attempts in the local community detection field.

1.5 The formation of the rest parts

In chapter two, we start with going through some state-of-art algorithms in the community detection field, with the focus on local community detection methods. After those, we put the common metrics in local methods under a metric formation framework for discussing and comparison; and we talk about the often ignorant way of making better use of the metrics : better designing the stop time for the local algorithms. At the end of this chapter, we mentioned a bunch of major problems with he state-of-art metrics.

We then describes our contribution to the field in detail in chapter three. At the beginning of chapter, we raise up a new viewpoint on the community structures, and discusses the cause to current metrics? disability to function in complex environment like social network graphs. Base on this alternative viewpoint, we then raise up two new metrics, with the aims in identifying community structures in traditional graph model and overlapping graph model environment respectively. At the end of this chapter we bring up the SLUD framework, which is able to describe nearly all the current attempts(metric+algorithm) in the field.

In chapter four, which is the experiment and discussion part, we gave proof in program result for all the statement above. Especially we prove the effectiveness of the new metrics as well as the application of SLUD metric.

Chapter 2

Literature Review

2.1 A brief introduction to detection methods

Given the increasing popularity of graph database and the wide-range application, the community detection has been a magnet attracting intensive research interest since its origin as early as 1927, when Stuart Rice looked for clusters of people in small political bodies, based on the similarity of their voting patterns. According to the definition5.2 in introduction part, many of the detection actions are based on the recognisable features of the community structures in the graph, among which the high edge density, high isolation are probably of the most common interest.

And other detection methods are usually associated with a statistical definition of community structure, such as modularity (Q) which was originally introduced by Girvan and Newman as a stopping criterion for their algorithm. Modularity has rapidly become an essential element of many clustering methods, which is by far the most used and best known quality function. And it represented one of the first attempts to achieve a principle understanding of the clustering problem, and it embeds in its compact form all essential ingredients and questions, from the definition of community, to the choice of a null model, to the expression of the strength of communities and partitions.

In summary, there have been quite a number of algorithms, based on various theories, attempting the task.

Up to today, the initial community detection area has been exhaustively studied and various different but effective-in-some-aspect methods brought about. It?s worth noting that, as the graph size explodes dramatically in the recent decades, nowadays the research focus largely shifts to the time and space consumption of the algorithm.

The focus of our work is on the improvement of seed-centric community detection methods. In the next sections we give out some common methods, of both global and local community detection methods, in the field to begin with.

2.1.1 global community detection methods

2.1.1.1 graph partitioning

The problem of graph partitioning consists in dividing the vertices in g groups of predefined size, such that the number of edges lying between the groups is minimal. The number of edges running between clusters is called cut size. Graph partitioning is a method mainly dedicated in optimising the isolation property. And the cut size is to be minimised for the target of a high isolation, which is also called the minimum cut problem.

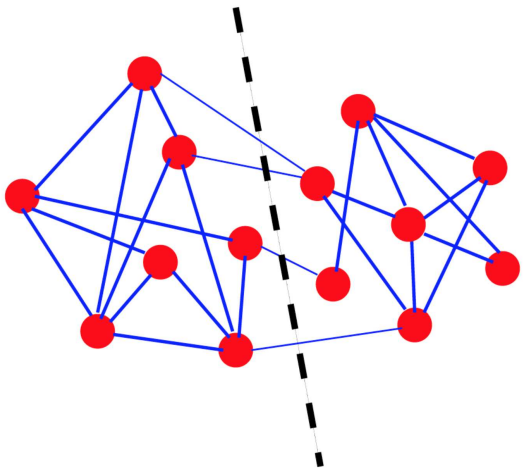

Figure 2.1: Graph partitioning. The dashed line shows the solution of the minimum bisection problem for the graph illustrated, i. e. the partition in two groups of equal size with minimal number of edges running between the groups. Reprinted figure from [Claudio Castellano and Loreto 2009]

One of the best-known solutions in this category is the max-flow min-cut theorem by Ford and Fulkerson [LR Ford 1956] . The theory states that the minimum cut between any two vertices s and t of a graph, i. e. any minimal subset of edges whose deletion would topologically separate s from t , carries the maximum flow that can be transported from s to t across the graph. Hence a equally reasonable partition can be given as the result of a max-flow algorithm on the graph.

2.1.1.2 Hierarchical clustering

Taking advantage of the similarity between nodes in the graph is another effective way to identify community structure in the graph. As raised by [J Friedman 2001], one may also use Hierarchical clustering algorithms, such as clustering techniques, to reveal the multilevel structure of the graph. The starting point of any Hierarchical clustering method is the definition of a similarity measure between vertices. After a measure is chosen, one computes the similarity for each pair of vertices, no matter if

they are connected or not. At the end of this process, one is left with a new n*n matrix X , the similarity matrix.

Hierarchical clustering techniques aim at identifying groups of vertices with high similarity, and can be classified in two categories: 1. Agglomerative algorithms , in which clusters are iteratively merged if their similarity is sufficiently high; 2. Divisive algorithms , in which clusters are iteratively split by removing edges connecting vertices with low similarity. In the context of normal community detection, we often regard the first type as the Hierarchical clustering detection method.

Hierarchical clustering takes advantage of the member similarity feature of the communities, and has very common applications in social network analysis, biology, engineering, marketing, etc.

2.1.1.3 Partitional clustering

Partitional clustering indicates another popular class of methods to find clusters in a set of data points. Here, the number of clusters is preassigned, say k. The points are embedded in a metric space, so that each vertex is a point and a distance measure is defined between pairs of points in the space. The distance is a measure of dissimilarity between vertices. The goal is to separate the points in k clusters such to maximise/minimise a given cost function based on distances between points and/or from points to centroids , i. e. suitably defined positions in space.

For example, for minimum k-clustering, the cost function here is the diameter of a cluster, which is the largest distance between two points of a cluster. The points are classified such that the largest of the k cluster diameters is the smallest possible. The idea is that the clusters of compact pattern are more likely to form communities.

With different definition of distance, partitional clustering is capable of making use of different features of the communities. For instance, the minimum k-cluster, where the distance stands for essentially the hop number from one node to another, also is based on the existence of massive edges inside the community structures.

2.1.1.4 Divisive Algorithms

The intuition of divisive algorithms is similar to that of graph partitioning, namely to divide the original graph up and get out of it a couple of communities. However, the divisive algorithms selects the edges based on the their chance of being the inter-community edges instead of the effort to minimise cut size. This is the philosophy of divisive algorithms. Hence the crucial point is to find a property of inter-community edges that could allow for their identification.

One of the most famous example of attempts in this field is brought by [Finding and evaluating community structure in networks]. Here edges are selected according to the values of measures of edge centrality , estimating the importance of edges according to some property or process running on the graph. The steps of the algorithm are:

1. Computation of the centrality for all edges; 2. Removal of edge with largest

centrality: in case of ties with other edges, one of them is picked at random; 3. Recalculation of centralities on the running graph; 4. Iteration of the cycle from step 2.

One of the metric they use as the standard of inter-community edge identification is the edge betweenness, by which they are making efforts on the distribution of shortest path between nodes on the graph(another feature of the communities).

2.1.2 local detection methods

2.1.2.1 Introduction

As defined in introduction part, Local detection methods, don?t require the information about the whole graph of concern. This school of methods compare a series of subgraphs containing the query nodes, and returns the one that looks most similar to a community structure.

In this time and age of big data, the local community detection method of many variants has become the trend for its superiority in time and space requirements. As described above, a fraction of the detection methods makes use of the accurate statistical numbers, such as the subgraph internal edge number, across-subgraph-boarder edge amount and the quantity of shortest path going through a particular edge, instead of general but vague features as the guide for community recognition. And the greedy algorithm basically aims at greedily forming a subgraph in order to best suit a given standard (some selected statistics) to achieve the community discovery goal.

For example, a greedy approach has been introduced by Blondel et al. [Fast unfolding of communities in large networks], for the general case of weighted graphs. Initially, all vertices of the graph are put in different communities. The first step consists of a sequential sweep over all vertices. Given a vertex i , one computes the gain in weighted modularity (Eq. 35) coming from putting i in the community of its neighbour j and picks the community of the neighbour that yields the largest increase of Q , as long as it is positive. At the end of the sweep, one obtains the first level partition. In the second step communities are replaced by super-vertices, and two super-vertices are connected if there is at least an edge between vertices of the corresponding communities.

2.1.2.2 Starting information of local community detection

Based on wether the entire knowledge on the graph is a necessity when one starts the greedy algorithms, they can be divided into two kinds: the global greedy algorithms omniscient about the graph, and the local greedy algorithms with merely the knowledge(precisely edge and node information) of particular nodes. Based on their exact action during the process, global greedy algorithm is also referred to as the greedy deletion algorithm and local greedy algorithm greedy addition algorithm. The choice of starting node set in greedy addition algorithms is also a research interest in the field.

Up-down local community detection When one has access to the information of the complete graph, one way of detecting the communities containing the seeds is to simply scan all results of a normal community detection method (excluding the local greedy algorithms) and retrieve the one including the seeds.

However, when it occurs that the seed amount exceeds one and they are organised into different communities by the community detection algorithm(graph partitioning for example), the result can be vague, overly general or inaccurate. Some methods in this type choose to repeat the detection process based on the previous result until the whole seed set being part of a single community, such as [Robust Local Community Detection: On Free Rider Effect and Its Elimination].

Bottom-up local community detection The community detection of this kind, due to the limitation of information, is hardly capable of making use of the features of the community structures. Hence the detection methods falling into this category normally make use of the statistical characteristics of the hidden communities, such as a set of nodes with high internal edge density with relatively low externe density. For the same reason, the greedy algorithm is often associated with the crawler-like seed-centric community detection process.

This is the main research focus in this work.

2.1.2.3 The process of bottom-up local community detection

We begin the description of the process with some basic definitions.

Definition 19. *Detected subgraph: detected subgraph is the latest result of the community detection algorithm; especially, in the bottom-up detection methods, it starts as the subgraph made of the query node set; and grows in size every iteration when another node gets pulled in*

Definition 20. *Ground-truth community: ground-truth communities are the given community structures in particular graphs. These are often for testing purpose*

Definition 21. *Candidates: in bottom-up local community detection methods, candidates are the node pool from which the algorithm needs to pick one up and merge it into the detected subgraph every iteration till end*

Definition 22. *Internal connection amount: internal connection amount is the number of edges between a node of concern and a ground-truth community it belongs to; if the node also belongs to the community, the internal connection amount is usually big in traditional graph models*

Definition 23. *External connection amount: external connection amount indicates the degree of a node of concern minus its number of connections with current ground-truth community; this is often used in the local detection methods, the internal connection amount is usually small in traditional graph models*

Definition 24. *Internal edge amount: external edge amount indicates the number of edges between a node of concern and the current detected subgraph; this is often used in the local detection methods as a indicator of internal connection amount*

Definition 25. *External edge amount: external edge amount indicates the degree of a node of concern minus its number of connections with current detected subgraph; this is often used in the local detection methods as a indicator of external connection amount*

The process of bottom-up local community detection can be described like this: Starting phase: The detection process starts with the query nodes being the detected subgraph. And all the whole neighbour node set makes up the initial candidate node set. Iteration phase: if the inclusion of any node in the current candidate node set wouldn?t make the detected subgraph more similar to a community structure, the detection algorithms stops; otherwise, the algorithm pull in the node making the most benefit for the structure, and update the candidate and detected subgraph information accordingly.

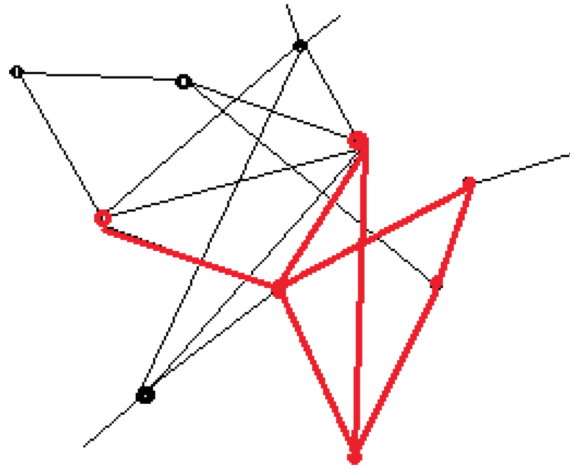

Figure 2.2: Detected subgraph before pulling in in this iteration

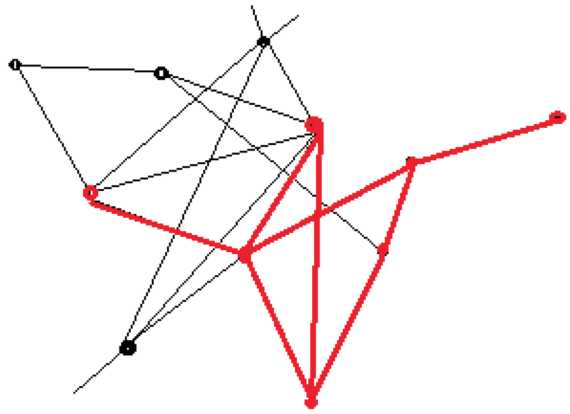

Figure 2.3: Detected subgraph after pulling in in this iteration

2.2 The introduction of metric

With the knowledge given above, now the main problem is, what defines a subgraph more or less like a community structure? Is there a line of similarity, above which we may say this subgraph makes a community? And this is where metric goes in. Metric is a standard to evaluating the similarity between subgraph structural information and community structure in practical use.

This similarity information may come from the comparison between the detected subgraph and the ground-truth communities. However, this can only be used for testing purpose, because the community information in the graphs is usually unknown(and that is above all why we need to detect the communities!). Especially, for better showing experiment result we give out some indicator stating the effectiveness of a metric(and its associated algorithm).

Suppose that the detected subgraph is notated by S, and the ground-truth community that contains the query nodes is represented by C.

Definition 26. *Precise: the accuracy of the detection result; Precise* $= \frac{|C \bigcap S|}{|S|}$

Definition 27. *Recall : the coverage of the detection result; Recall* $= \frac{|C \bigcap S|}{|C|}$

Definition 28. *Fscore: Fscore* $= \frac{Precise * Recall}{Precise + Recall} * 2$

In the context of local community detection, the criteria of similarity often comes from the outstanding features of community structure, including internal edge density and external isolation and others.

Definition 29. *Metric in local community detection process: in our work, we refer to the metric as a full-map function that defines the quantised quality for all subgraphs.*

Furthermore, in our context, metric is often the only reference for measuring how good a community the part of the graph is. And if the metric value associated with a particular subgraph meets the definition of community or makes the best-match against the community feature, we say this subgraph is the detected result of our metric on the graph.

2.3 Main Elements Metrics Concern

As described above, the criteria of similarity often comes from the outstanding features of community structure. In this section we are going to talk about some common elements in concern of metric formulation.

It?s worth noticing that the usage of random graph here. P. Erdos and A. Renyi [on random graph I] introduced the concept of random graph, in which the probability of having an edge between a pair of vertices is equal for all possible pairs. In a random graph, the distribution of edges among the vertices is highly homogeneous. For instance, the distribution of the number of neighbours of a vertex, or degree , is binomial, so most vertices have equal or similar degree. Real networks are not random

graphs, as they display big inhomogeneities, revealing the outline of the community structures[community detection in graphs]. A random graph, for instance, is not expected to have community structure, as any two vertices have the same probability to be adjacent, so there should be no preferential linking involving special groups of vertices. Therefore, one can dene a null model , i. e. a graph which matches the original in some of its structural features, but which is otherwise a random graph. The null model is used as a term of comparison, to verify whether the graph at study displays community structure or not. The most popular null model is that proposed by Newman and Girvan and consists of a randomised version of the original graph, where edges are rewired at random, under the constraint that the expected degree of each vertex matches the degree of the vertex in the original graph [Newman and Girvan , 2004].

Hence many of the elements in the metric are from statistical difference addressed between the graph formation with clusters and without.

As far as the author is concerned, the common elements metrics contain include but not least:

Internal Edge Quantity The number of edges existed between the node pairs situated inside the subgraph node sets. The internal edge quantity is highly associated with one of the key features of communities: high (edge) density. Hence an recognisable relatively high internal edge quantity, normally superior to that of a random graph, is expected with the presence of the community structure.

External Edge Quantity Similar with the internal edge quantity, the number of connected node pair with one end inside the subgraph and another outside is expected to differ from the same subgraph belong located in a random graph. The idea is from both another key feature of the community: high isolation, and a comparison to the internal edge quantity. And that is because of the outstanding but homogenised internal and external edge quantity hardly make clear of the existence of community structure.

Node Quantity The number of vertexes belonging to the community is another major concern when we design metrics. Though a strict limitation on node number is not common in any networks, but empirically we have the idea that there has to be a range of implicit size setting for any kind of communities, whether it?s social communities or good categories; Other than that, the node quantity makes good sense when it is used in association with internal external edge quantity. The reason is that the internal edge quantity and external edge quantity themselves are tremendously misleading. The collaboration with the edge quantities with the node amount actually produces the edge density, some particular distribution of which is the real indicator of the clustering in the graphs.

There are many other elements can be used to evaluate community structures for sure, for instance the betweenness centrality, the number of cliques included , the structural similarity between nodes inside and the cycle amount between them.

However many of them are not applicable for our purpose, a seed-centric local greedy search algorithm on large graphs. For example, the shortage of topology information about the entire graph disables the adoption of betweenness centrality in

the metrics, the calculation of which needs information about every possible shortest path between every pair of nodes. Another example is the existence of cliques. Although the clique is a relatively good sign of the community structure even in a complex overlapping context, as proved by [Uncovering the overlapping community structure of complex networks in nature and society.] with Clique Percolation Method (CPM), but the slim chance of cliques in considerable sizes containing the seeds makes it not practical at all.

2.4 State-of-art Metrics

And according to [robust Yubao 2015], a goodness metric is usually used to measure whether a subgraph forms a community in local community detection,. The existing goodness metrics for local community detection can be categorised into three classes. The first class optimises the internal denseness of a subgraph, i.e., the set of nodes in a community should be densely connected with each other. Such metrics include the classic density definition [Dense Subgraphs with Restrictions and Applications to Gene Annotation Graphs], edge-surplus [Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees], and minimum degree [The community-search problem and how to plan a successful cocktail party]. The second class optimises both the internal denseness and the external sparseness. That is, the set of nodes in the community are not only densely connected with each other, but also sparsely connected with the nodes that are not in the community. Such metrics include subgraph modularity [Exploring local community structures in large networks], density-isolation [Finding dense and isolated submarkets in a sponsored search spending graph]. The local modularity measures the sharpness of the community boundary and belongs to the third class [Finding local community structure in networks.]. Using this metric, the set of nodes in the boundary of the community are highly connected to the nodes in the community but sparsely connected to the nodes outside the community.

We now will go through a fraction of the metrics mentioned above.

2.4.1 Internal Denseness

The metrics fall into this category are ones taking merely the edge quantity and node quantity elements into account. The underlying idea is that the set of nodes in a community should be densely connected with each other.

2.4.1.1 Edge Density [$Metric_S = \frac{e(S)}{|S|}$]

Density is one quantitative measure of the connectedness of a subgraph and is defined as the ratio of the number of induced edges to the number of vertices in the subgraph.[Dense Subgraphs with Restrictions and Applications to Gene Annotation Graphs]. It's north noting that even though there are an exponential number of subgraphs, the problem of identifying optimal subgraph under the edge density metric

can be solved exactly in polynomial time[Dense Subgraphs with Restrictions and Applications to Gene Annotation Graphs].

2.4.1.2 Edge Surplus [

$$Metric_S = e(S) - \alpha * \binom{|S|}{2}$$

]

However, proved by [C.E. Tsourakakis, F. Bonchi etc] , densest subgraphs(with highest density) are typically large graphs, with small edge density and large diameter which in many occasion doesn't make an expected result. So people require another way of making use of the edge density.

A clique is a subset of vertices all connected to each other. And it has been proved that even in a most complex network the emergence of cliques probably lead to the structure of community, since it is very unlikely that inter-community edges form cliques: this idea was already used in the divisive method of [Defining and identifying communities in networks]. However even if in a dense graph, the cliques are not everywhere to be spotted; besides that, the problem of finding whether there exists a clique of a given size in a graph is NP-complete.

Hence [C.E. Tsourakakis, F. Bonchi etc] introduced the concept of quasi-cliques. A set of vertices S is an $\alpha s - quasi - clique if e[S](\binom{|S|}{2}), i.e., if the edge density of the induced subgraph G[S] exceeds a thres$ $(0,1). And the amount of internal edges a set of nodes share beyond the expected amount of this weakened clique definition$ $surplus. And the subgraphs that maximize f_\alpha(S) as are referred to as the optimal quasi - cliques.$

2.4.1.3 Minimum Degree [$\mathbf{min}_{degree}$]

The internal denseness related metrics discussed above are both upon the average internal degree of the nodes in the extracted community. However, the use of average degree type of metrics in local community detection methods has the drawback of being sensitive to free-riders, namely, irrelevant but dense subgraphs that may be attached to the query nodes and yield unintuitive solutions. For this reason, another a measure, the minimum degree, is attracting a part of research interest.

The density measure $f_m(H)$ based on minimum degree is defined to be the minimum degree of any node of V_H in the ein (V_H, E_H). As any measure that seeks to maximise a minimum, f_m has the drawback that it is sensitive to outliers. However, i $related community, as the measure f_a does.$

In the seed-centric detection context specially, with the collaboration of excluding nodes that are far from the query nodes, as usually these nodes are less related to the query nodes than those that are nearby, a somewhat satisfactory metric free from the free rider effect can be designed under this scheme[how to organise a cock-tail party].

2.4.2 Internal Denseness and External Sparseness

The second class optimises both the internal denseness and the external sparseness. That is, the set of nodes in the community are not only densely connected with each

other, but also sparsely connected with the nodes that are not in the community. And we are going through subgraph modularity and density-isolation here. According to [Finding Dense and Isolated Submarkets in a Sponsored Search Spending Graph], the tasks of identifying dense subgraphs and isolated subgraphs are different; the subgraphs that are most isolated, having the smallest ratio cut score or conductance, tend not to be dense, and the densest subgraphs tend to have large amounts of money crossing their boundaries. More generally, there is some unknown tradeoff between how dense a set can be and how isolated

2.4.2.1 Subgraph Modularity [

$$Metric_S = \frac{jn_d(S)}{out_d(S)}$$

]

[Exploring Local Community Structures in Large Networks] proposed an interesting metric based on both internal denseness and external sparseness: the subgraph modularity. The modularity M of a sub-graph S in a given graph G is defined as the ratio of its internal degree amount, ind(S), and external degree amount. And obviously the quantity of modularity will increase when sub-graph S has more internal edges and fewer external edges(internal density and external spareness).

On top of this definition,they further give the definition of a proper community structure: Given a graph G, a sub-graph S $\subset Gisamodule/communityifM > 1.Thissimplecommunityd$

2.4.2.2 Density Isolation [

$$Metric_S = M(H) - \beta B(H) - \alpha|H|$$

]

[Finding Dense and Isolated Submarkets in a Sponsored Search Spending Graph] put forward a metric for subgraphs that are simultaneously dense and isolated. For any subgraph H, M(H) indicated the total edge amount(or weight, as expressed by the original work) within the subgraph; B(H) means the total weight on edges crossing the boundary between the subgraph and the rest of the graph; —H— represents the total number of nodes inside the subgraph. And according to their work, for any fixed value of $\alpha 0 and \beta 0, thesubgraphthatoptimisestheobjectivefunctioncanbefound$.

2.4.3 Sharp Boundary

Given the features of communities such as high edge density and high isolation, it?s intuitive to make sense of the statement that the nodes on the boundary of communities, vertices situated in the communities that have at least one external neighbour, should make their appearance quite outstandingly for their edge distribution.

The emergence of this is also upon one of the main issues with the metrics that takes both internal denseness and external spareness into consideration: at the time communities are big enough, the big portion of total internal edges(neither end locates in the boundary) would make almost all subgraph look good under those metrics.

2.4.3.1 Local Modularity

If we restrict our consideration to those vertices in the subset of C that have at least one neighbour in U, i.e., the vertices which make up the boundary of C, we obtain a direct measure of the sharpness of that boundary. Additionally, this measure is independent of the size of the enclosed community. Intuitively, we expect that a community with a sharp boundary will have few connections from its boundary to the unknown portion of the graph, while having a greater proportion of connections from the boundary back into the local community

2.5 A General Framework of Metric Formation

2.5.1 The Metric Framework, and the relation to the state-of-art metrics

The framework is adapted and improved from the framework raised by [Denser than the Densest Subgraph: Extracting Optimal Quasi-Cliques with Quality Guarantees]. The idea is, the equation containing the very basic elements in metric formation is capable of expressing the constitution of most metrics in local search methods.

Let G = (V,E) be a graph, with —V — = n and —E— = m. For a set of vertices $S \subseteq V$, $let[S]be the number of edges in the subgraph induced by S. We define the following function.$

2.5.2 A Special Use of the Framework

Our framework is more than capable of describing most of the state-of-metric used in local community detection methods: its essence as a function can be made good use of.

For example, under the framework, to design a algorithm that gets out of a community containing the seeds in a given size from the graph is trivial; furthermore, with proper design such as the application of piecewise function, the metric under the framework can lead to communities of expected size.

$$Metric_S = \begin{cases} Function A(S) & \text{if } x \in (a,b) \\ Function B(S) & \text{otherwise} \end{cases}$$

2.6 The Stopping signal of a local community detection process

During the process of a local community detection, what matters not only include the choice of metric, but also the time for the community detection process to stop.

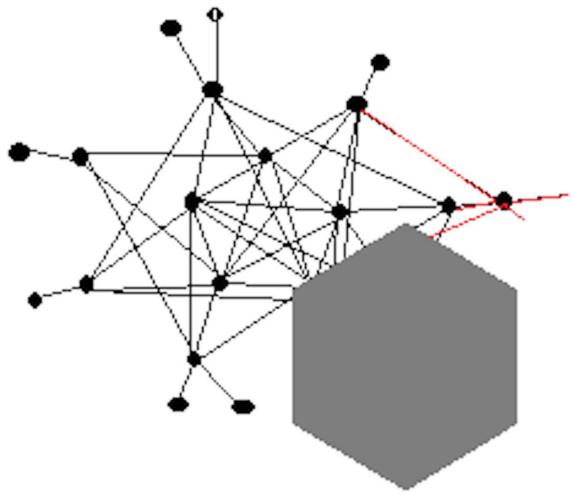

Figure 2.4: When the information about the whole graph topology is incomplete

This signal is particularly essential if we don?t have the detailed information about the whole graph(like figure 2.4), or the time or space constraints doesn't permit the algorithms running over the whole graph before the result comes out. Given that the seed-centric local bottom-up community detection method without the access to the complete graph information is the main focus of our work, this aspect is therefore of superior significance.

The ability of signalling a proper halt is not well made use of in algorithms, for instance the algorithm described in [local modularity]. The algorithm given in the work would not halt until the detected subgraph has reached a certain size. Figure 2.5 is the example.

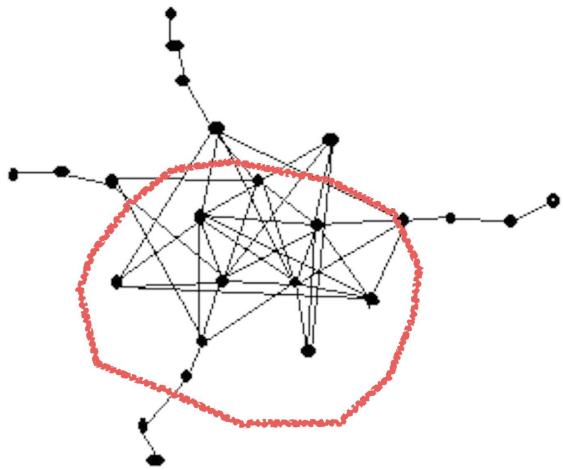

Figure 2.5: The red circle specify current detected subgraph; the metric value has dropped significantly and it?s not likely to recover though , the algorithm continues running

Some of the metrics, such as the first algorithm in 6.1.2, use the possibility of metric value increase as the measure of process: as soon as the metric stops increasing, the detecting process terminates. We will refer to this kind of measure of process as greedy addition in the later part of the work.

Meanwhile, last but not least, a third party of the metric usage exists which utilises some of the elements in its formation instead of the entire equation as the measure of the process. An outstanding example for this is [find local community structure in networks], where the processing terminating signal is a user-setting result node number.

There are often two types of signal indicting the end of the detecting, namely threshold signal and optimum signal.

2.6.1 Threshold Signal

With the threshold signal, a community is ascertained and algorithm stopped as long as the statistical information of the nodes and edges within met a fixed threshold. And the detected subgraph is to be recognised as the community structure. However, it?s evident that in this mode one may get back himself a great number of identified communities unless he stops the detection process manually at some point.

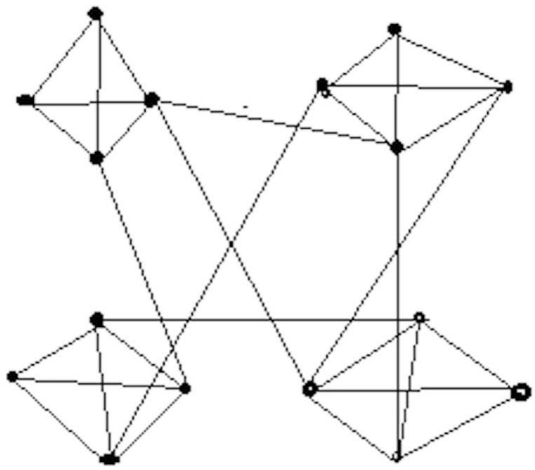

Figure 2.6: In this graph, every K4 subgraph makes a ground-truth community; however, with quasi-clique metric, it?s very likely to include unrelated parts

2.6.2 Optimum Signal

Another commonly accepted mode of signalling community existence is the optimum mode, in which only the optimal result out of all subgraphs encountered is to be returned as the community detected.

Take the local greedy approximation algorithm described in [Denser than the Densest Subgraph: Extracting Optimal Quasi-Cliques with Quality Guarantees] as an ex-

ample. This bottom-up algorithm starts initially with the prescribed seed set and then it keeps adding vertices to the current set S while the objective improves. When no vertices can be added, the algorithm tries to find a vertex in S whose removal may improve the objective. As soon as such a vertex is encountered, it is removed from S and the algorithm re-starts from the adding phase. The process continues until a local optimum is reached.

The work gives additionally another instance in optimum signal mode: The algorithm iteratively removes the vertex with the smallest degree from the graph. The output is the subgraph produced over all iterations that maximises the metric(measure of goodness).

2.7 problems with start-of-art metrics

2.7.1 Free Rider Effect with Raised Solutions

As defined and systematically proved by [Yubao 2015], most existing metrics tend to include irrelevant subgraphs in the detected local community, also referred to as the free riders. Specifically, if a goodness metric will include the global optimal subgraph,the subgraph with the largest possible goodness value, in the identified local community, we say this metric causes the global free rider effect; at the same time, if the metric will pull in the local optimal subgraphs, subgraphs whose goodness value is greater than that of their any subgraph, in the identified local community, it is said to be causing the local free rider effect. It?s obvious that the free rider emerges universally with almost every possible local detection metric.

Goodness metrics	Ref.	Formulas $f(S)$		Glo.	Loc.
Classic density	[29]	$e(S)/\|S\|$		✓	✓
Edge-surplus	[36]	$e(S)-\alpha h(\|S\|)$	concave $h(x)$	✓	✓
			$h(x)=\binom{x}{2}$	×	✓
Minimum degree	[31]	$\min_{u\in S} w_S(u)$		✓	✓
Subgraph modularity	[23]	$e(S)/e(S,\overline{S})$		✓	✓
Density-isolation	[22]	$e(S)-\alpha e(S,\overline{S})-\beta\|S\|$		✓	✓
External conductance	[2]	$e(S,\overline{S})/\min\{\phi(S),\phi(\overline{S})\}$		×	✓
Local modularity	[7]	$e(\delta S,S)/e(\delta S,V)$		×	✓

Figure 2.7: universality of free ride effect, table reprinted from [Yubao]

Especially, by definition, the global optimal subgraph also belongs as well to local optimal subgraph.

Definition 30 (Global Optimal Subgraph)**.** *The global optimal subgraph is the subgraph* $G[S_m]$ *whose goodness value* $f(S_m)$ $f(S)$, *for any* $S \subset V$.

Definition 31 (Local Optimal Subgraph)**.** *A local optimal subgraph is the subgraph* $G[S_lm]$ *whose goodness* S_lm. *That is, deleting any node* (s) *from a local optimal subgraph will decrease its goodness value. Note that by de*

Specially, given that our study focus is on identifying cluster structures with seed-centric bottom-up local detection techniques, the specific subgraph with high density however well separated from the seeds is out of concern due to the impossibility to

be encountered. Hence, the free rider effect issue in our context equals the local free rider effect as defined in [Yubao]'s work.

[The Architecture of Complexity] points out that hierarchic system structures, systems composed of interrelated subsystems, have crucial roles to play in the networks(complex systems). And each of the latter being, in turn, hierarchic in structure until we reach some lowest level of elementary subsystem. With the self-evident idea in mind that the subsystems presenting in the graphs tend to form internally dense subgraphs, we may further classify the free riders, with the knowledge, into two categories, as the figure 2.8 shows. More specifically, the first type is the dense subgraph that is a part of the ground-truth community at a particular scale; and the second is the dense subgraph which is not related to the seeds whatsoever. In other word, if the seeds belong to two communities at the same time, one containing another, the free rider of first kind might be included in the community of the larger scale while next to the smaller small one; and the second kind won?t interact, from a ground truth point of view, at any scale.

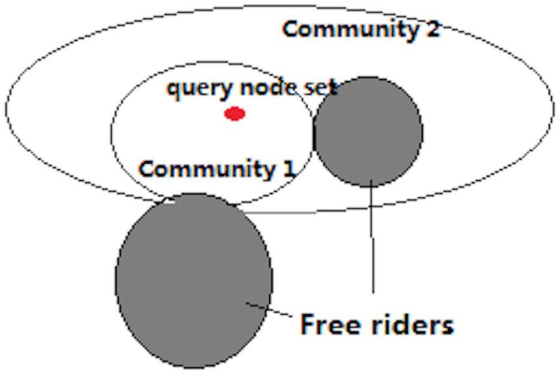

Figure 2.8: Some of the free riders may be members to community containing query node set but at a larger scale; while some don't interact with them whatsoever

Intuitively, the actual challenge brought by free riders, to during the detection how do we distinguish between the two kinds of free riders, allow the first kind at a proper time and keep rejecting the member nodes of the second type.

Luckily there are some previous studies on the problem already, among them is the idea adapting Minimum Degree in the detection process instead of the Average Degree, proposed by [The Community-search Problem and How to Plan a Successful Cocktail Party]. The work states that In order to find densely connected communities that contain the query nodes, one needs to define an appropriate measures of density. Such measures can be the average or the minimum degree of the nodes in the extracted community. While the average degree maximum often suffers from the free riders as described, their focus is on the latter measure, the minimum degree. Their exact goal is to find compact communities, containing the query nodes, and whose minimum degree is maximised, by excluding nodes that are far from the query nodes, as usually these nodes are less related to the query nodes than those that are nearby(nodes with

less connections with the seeds).

However as what they admitted, the attempt with minimum degree has the deadly drawback: it is sensitive to outliers. And this 'Rob Peter to pay Paul' kind of act obviously doesn't suit our goal.

Another solution is raised by [Yubao], with a random-walk-based weighting scheme beforehand to make nodes near the seeds more attractive. The solution seems to work fine however need much time and information particularly in the context of a complex graph. Hence it might not work as well under challenging constraints.

2.7.2 Outliers

This is another situation that leads to potential suboptimal solutions, which possibly doesn't make much sense. The situation can be explained as, some of the metrics, ones with less concern on the node number increase in particular, allow the introduction of the node that are in fact weakly connected to the community, and therefore result in an unsatisfactory accuracy ¡Detecting Communities in Social Networks using Local Information¿. For example, the subgraph modularity and the local modularity metrics would include the outliers, such as two-edge node with one connecting the community and another the outside, in the sparse graphs.

figure 2.9 shows a spare graph, where node 1,2,3,4 (a K4 subgraph) form a ground truth community. However, under the metric of subgraph modularity(equation), the metric value of this particular subgraph is less than 1; and this gives rise to greedy introduction of the outlier node from 11 to 14.

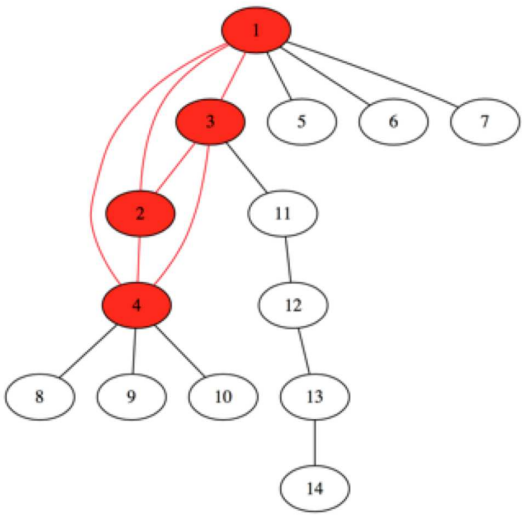

Figure 2.9: K4; Ground-truth community

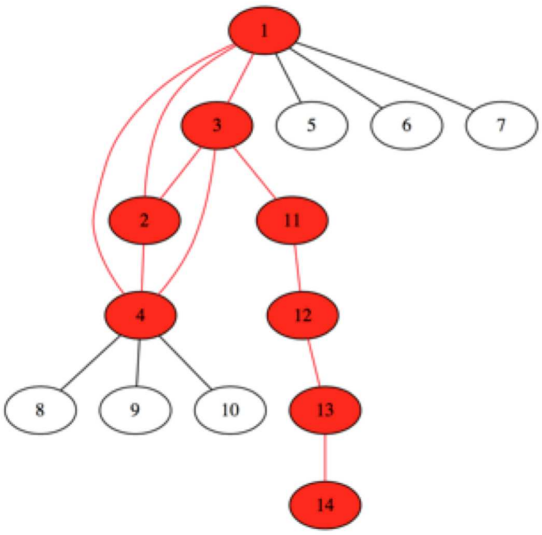

Figure 2.10: Detected Subgraph

2.7.3 Local Optimum Traps

Because the metric concerns only about info on current nodes and their adjacencies, and don't care about the graph as a whole, or simply the relationship and structure several steps away, the metric (often used in greedy addition algorithms) usually result in a solution that is better than all other solutions that are slightly different, but worse than theglobal optimum ¡Paul E. Black, "local optimum", inDictionary of Algorithms and Data Structures[online], Vreda Pieterse and Paul E. Black, eds. 17 December 2004. (accessed TODAY) Available from: http://www.nist.gov/dads/HTML/localoptimum.html ¿

In another word, when we use the metric in a node set expansion method looking for the best suited community from seeds, the process may risk turning for a non-member because of its inadvertently action in enhancing the goodness measurement.

For example, suppose the nodes in figure 3.1 are extracted from a larger graph and form a community. However if we take the node 3 as the seed and utilise the classical density metric, by the attempt to maximise the ratio of edge number and node number gradually, we probably get out nodes 1,2,3,4,5,6,7 and omit the others. It is obvious that 8,9,10,11,12,13 are forming a complete graph, and that is actually a major contribution to the metric goodness. And the classical density is not the only one that suffers the local optimum trap; as far as I know, the metrics of edge-surplus, density-isolation and more fail to well settle the issue as well.

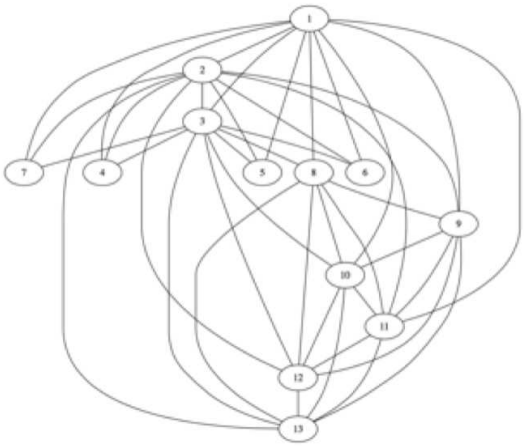

Figure 2.11: The graph of concern

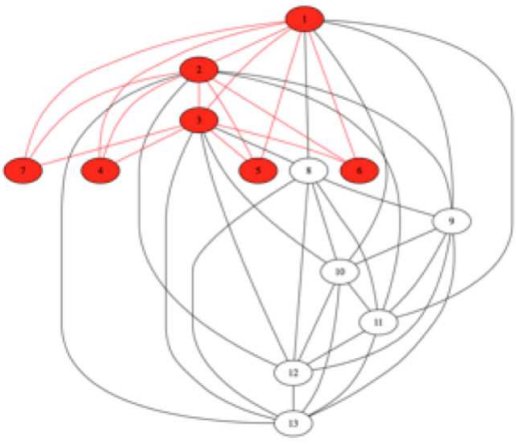

Figure 2.12: Detected Subgraph; result

Chapter 3

Our methods

3.1 The alternative view of point on community formation

3.1.1 Intuition: phenomenon in real-world networks

From empirical observation, we got this impression that the majority of communities have more constructional features in common, than ?high concentration of edges inside and lower in between the communities?. Especially we notice that the community member entities in real-world networks can be divided into three categories, namely the core community entities, boundary entities and other entities. For example, in a particular research field, there is supposed to be a small number of groundbreaking research works, which a considerable proportion of further papers base on and cite. And these pioneering works are refereed to as the core entities in this research community. At the same time, there should also be a proportion of efforts in this field absorbing the knowledge from another discipline(for instance, engineers learn from biological structures to build machines like airplanes); and these attempts would presumably associate above-average number of researches in other fields, however the base-stone conception and theories still belong to the field. And the works in this type are believed to be the boundary entities in the community, in other words, they will leave the community for one additional step. The examples are easily spotted in social networks as well. If a social circle bases on a shared interest, the people mastering the skills about the interest, offering to help others and making a great number friends in the community are seen as the core of it; meanwhile, those do enjoy this interest the most however have many other interests tent to form the boundaries of the communities.

It's worth noticing that the community formation we analysis here is based on traditional graph model. And these are the ones most fit the traditional model of community.

Correspondingly, the graphs dedicated to map these networks are expected to have a similar structure: specifically, the nodes in graph communities are expected to fall into three categories: the nodes on the boundary, core nodes, and other nodes in the middle.

The nodes in the core are hence marked with the high node degree and the majority(many connection within the communities) of its neighbours locate in the same

community; while the nodes on the boundary are believed to have a slightly big number of connections with outside nodes, however this number is lower than that of internal connections still. Therefore, we will refer to them as the core nodes, boundary nodes and other nodes respectively in the following parts.

3.1.2 the alternative community formation theory

Based on the intuition of node classification in community formation, we then propose the alternative community formation theory, base on which we will raise up some new metrics later. This alternative community formation theory is fairly plain but of good use. As far as we know, this formation theory is not previously adapted in the local community detection field.

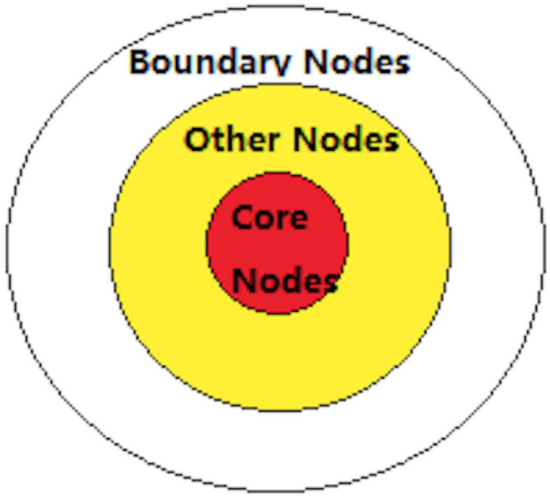

Figure 3.1: community structure under Node-centric view point

Definition 32. *Node-based community formation theory: The traditional communities' structural properties are the collective effect result from three types of nodes inside the community, namely core nodes, boundary nodes and other nodes. All the three types have big interactive connection amount. Especially, core nodes locate in the centre of the community, having good number of neighbours most of whom in the same community; boundary node are on the edge, related to outside nodes relatively more intensively but most adjacencies still in the community; other nodes have properties in between.*

Node-based community formation theory:

We are now going to review the state-of-art metric, with this alternative node-based community formation theory in mind. Furthermore , we would bring forward our innovative metric, which is based on the theory, in later sections.

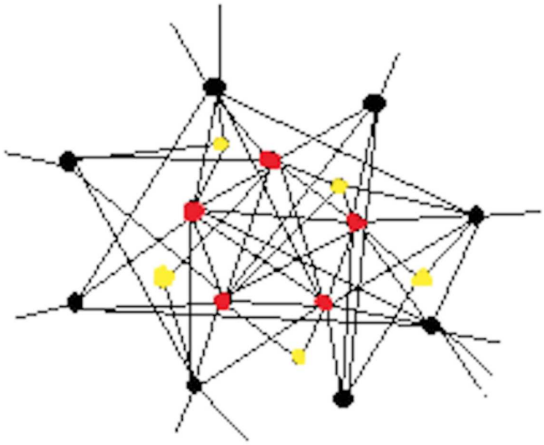

Figure 3.2: abstract community structure under Node-centric view point

3.2 A Review on State-of-art Metrics with the Node-centric Community Formation Theory

Edge Density: We discussed that the idea behind the edge density metric, is to pull in the node with the most connections with the current subgraph of interest; and stop at a given time (usually when the average edge density starts to drop or the subgraph reaches a certain size). And we know that in terms of connections with the query node set, the nodes within a same community, whether from boundary or core, make no outstanding difference; Hence, the goal of utilisation of edge density metric, from the node-centric community formation point of view, is to continually include the community member regardless of its role in it. Additionally, the algorithm(given that it is set to stop when edge density has to drop after inclusion) stops with the boundary nodes and prevents the inclusion of outside nodes.

Edge Surplus: The node goodness evaluation works identically for edge surplus: tries to make community member outstanding and selects it for addition.

Subgraph Modularity: The idea behind the subgraph modularity is different from above: the metric tries to underline the nodes maximising the ratio of internal connection amount against external connection amount. In other word, the node with relatively high level of association with query node and low with the rest parts would be regarded as the best fit under subgraph modularity. And according to the node-centric community formation theory, the core node, which is modelled with high-degree property would hardly make it to be included(when we try to minimise the external edge amount). And the nodes playing other parts of the community enjoy a better chance of being added, due to their low conductance with the outside part and presumably equal amount of connections to query node set.

Metric	Community Structural Feature	Intuition	Metric	Node Goodness Measurement	Algorithm Stops When
Edge Density	Community member nodes(core / boundary / other) should all be interactively connected severely	Adding the node, member to the community	**e(S) / lSl**	Δe(S)	Approaching the boundary
Edge Surplus	Community member nodes(core / boundary / other) should all be connected at a certain level	Adding the node, member to the community	**e(S) - α · (lSl 2)**	Δe(S)	Approaching the boundary
Subgraph Modularity	Ratio (Internal edge against conductance) is high	Adding the node, member to the community, especially the node in other parts than core or boundary	**e(S) / ρ(S)**	Δ(e(S) / ρ(S))	Approaching the boundary

Figure 3.3: State-of-art Metrics Under Node-centric View Point

3.3 Core-seeker metric

In this section we are going to raise up our first contribution to the field, the core-seeker metric; we will begin with the our original intuition about the design.

3.3.1 Intuition behind the metric

Based on the node-centric community formation model and especially the definition of the core nodes, we know that they are of high degree, and most of their neighbours belong to the community as well. The characteristics of the core make them the ideal targets to be included in the early stage of the detection: intuitively, from these node we may access quite a proportion of the nodes in the community; and the chance of ending up with more connections with outside nodes after the addition operation is small.

And we know that in the detection process if you

Especially, none of the metrics we talked about earlier have attempted for the identification of a certain type of community node to the most of our knowledge.

3.3.2 Core-seeker metric

We raise up the metric, which aims at identifying the nodes with most internal edges at the beginning in each iteration(maximum $\triangle e(S)$); $and from the nodes with the same biggest amount of internal edg$

Additionally, when it occurs that any node with the highest internal connection can not guarantee the increase of there conductance, the algorithm is to be stopped. The requirement is a little bit strict, however only through this can we maximise the possibility of the newly added nodes being core nodes.

Because of the metric is aiming at the most influential entities in the community and very strict, it is more likely to be used as a first iteration method under the SLUD framework(to be introduced later).

Essentially there isn't much difference between core-seeker metric and edge density in deciding which node to get pulled in for the iteration, it's the timing of stop

that mostly makes the metric different. The significance of stopping at a proper time has been discussed in 'Literature Review'.

Metric	Community Structural Feature	Intuition	Metric	Algorithm Stops When
Core-seeker	The core nodes have intensive connections internally, and high node degree as a whole	If the node is well connected with the query node set and has an even bigger number of external edges, it is most likely a core node	e(S) (+ρ(S))	The addition of no candidates with highest internal connection amount can increase the external connection amount

Figure 3.4: Core-seeker metric Under Node-centric View Point

3.3.3 Effectiveness Analysis

The core-seeker metric makes sense not only in the context of node-centric community formation theory, but also when dealing with different kinds of issues. And we are now going through some of them to illustrate the idea.

3.3.3.1 The Challenges Under Traditional Community Model

Let's have another look at the challenges in the local detection process, namely the free rider effect, outliers and local optimum effect. Intuitively, if we design an algorithm that is capable of getting rid of the problems and makes sense for community identification, we get ourselves a well qualified method.

Free Rider Effect Free rider effect is a most common phenomenon. Moreover, the detection methods are even more susceptible to them during the starting phase.

Theorem 3.3.1. *The earlier inclusion of a member of free rider would do more harm*

Intuitively, if the algorithm gets free rider members in the detected subgraph at the very starting phase, due to the big proportion of them at this stage(reference nodes incorrect), it is very likely for the algorithm to malfunction. However, if we start with a safe and sound metric, doing the best we can to include accurate nodes in the starting phase, the bigger amount of correct nodes(belong to the same ground-truth community) would reward us with smaller chance of making mistakes.

Especially, the more free rider nodes or outliers we get, the bigger chance we may include more of them. And the metric wouldn't be affected much(or sometimes getting even better!) by these mistakes. So as long as they are included, even with the SLUD framework we are proposing later, it's unlikely we can get rid of them. Hence, the early inclusion of a member of free rider would do great harm to the detection.

Theorem 3.3.2. *nodes in free riders to a community usually take up the boundary nodes of a community.*

The free riders would form a densely interacted subgraph. Hence the core nodes are not likely to be part of free riders.

This theorem demands us for an highly accurate result in the early stage of detection, to limit the free rider effect. Here is where our core-seeker metric goes in: this strict but accurate metric aims only at the core entities in the cluster. Given the theorem X and X+1, the idea is clear that the effort on digging out the core nodes of a community to begin at least to some degree lighten the free rider effect from the local detection process.

And after that we are left with many choices. If the goal of detection is to determine a bunch of nodes that are very sure to be in the same community with the query nodes, or to identify the most influential entities in the community, we may simply stop here; however, if the goal is, on the other hand, to retrieve a subgraph of relatively larger size, we may apply all kinds of local community detection methods after this prestage application, which better the result. This enhancement is proved by intuition as well as experiment results as shown in 'experimentdiscussion' section.

Outlier Outlier problem, as defined beforehand, is the issue of including unrelated vertexes which accidentally met the requirement of metric at the time. Take the outliers under edge density metric as examples: these nodes might not be that connected to the subgraph, however they are pulled in for sometimes the query node set situated right next to them. The coincidence makes the outliers look sometimes as good as the entities in the community against the metric.

Thus, to deal with the issue, the best and general solution is to take in only the nodes with the highest internal edge amount. Our metric is capable of doing it. Furthermore, the greatest hidden danger of putting emphasise on internal edge amount, the free rider effect, has been solved with our metric, as described in the section above.

Parameterized Metric Unmatched with Graph of concern On top of the common issues above, another issue concerning the design of metric is the parameter setting. For example, the subgraph modularity metric in many occasions take $\alpha as 1/3, and edge density metric usually take$ $2/5 * \varphi(S)) rather than (e(S)), the metric performs far better in the graph of Amazon.$

However, these parameter sets are usually empirical and not widely applicable. The good example is the (e(S) - 2/5* $\varphi(S))$ *metric, which performs superior lyon Amazon but not even as good as trivia*

Hence, another strong proof for the effectiveness of our core-seeker metric is its broad applicability on different kinds of graphs(under traditional community structure model). The newly proposed metric doesn't have any concerns on parameter settings, hence the effectiveness is to be irrelevant to the amount of knowledge or experience with the graph to be operated on.

3.4 A second community structure model

3.4.1 Intuition

As mentioned in introduction, at the current stage almost all of the local community detection methods suffer from intolerably low correct rate in dealing with graphs of social networks.

Figure 3.4 shows us clearly the situation on Youtube, a social network. This experiment is done by [Yubao], during which the access to the whole graph information is guaranteed.

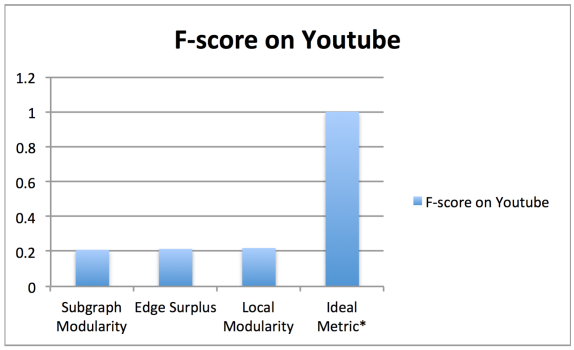

Figure 3.5: F-score of state-of-art metric on Youtube

Furthermore, the situation exists not as an extreme isolated case with the local bottom-up community detection methods, but a common one even the information of the whole graph is made good use of, as proved by [Yubao].

And [Jaewon Yang] and some fellow pioneers, with their observation and analysis, brought forward a new and organised explanation on the weird situation on social networks. Their work starts with a novel, and in retrospective very intuitive, observation that overlaps of communities tend to be more densely connected than the non-overlapping parts [33, 35]. In particular, they empirically observe that the more communities a pair of nodes shares the more likely they are connected in the network.

For example, people sharing multiple hobbies (i.e., interest based communities) have higher chance of becoming friends [23], researchers with many common interests (i.e., many common scientific communities) are more likely to work and publish together [26].

The result of experiments claims that the algorithm do identify the community structure under the traditional model(represented by the metric). However, the correctness of the detection under the traditional model doesn't stand for the success on identification of real communities in these graphs.

The situation in these complex graphs has its foundation in real-world networks. As discussed in the introduction part, the traditional community model often represents well the gathering of entities when the relationships of concern between them is of a single type(single-layer graph). For example, the relationship in Amazon network/graph is mostly about the commodity classification. However, it occurs that the network on human issues on the contrast is a multi-layer one. For example, within the network of Youtube, people tend to form communities based on their interest/taste as well as their background. And the different communities are not likely to influence the structure of each other but overlapping irregularly.

Even though the intuition and observations are self-evident and make sense in the board context, these empirical knowledge do not align with the traditional commu-

nity structure, where the intensive inter-node relationship signals the existence of a community.

3.4.2 A second community structural model : in the context of intensive overlapping

In this section we put forward another community structural model dedicated to make sense of the clustering conditions in the graphs of this kind. Additionally, we would also analysis the model under the node-centric point of view later.

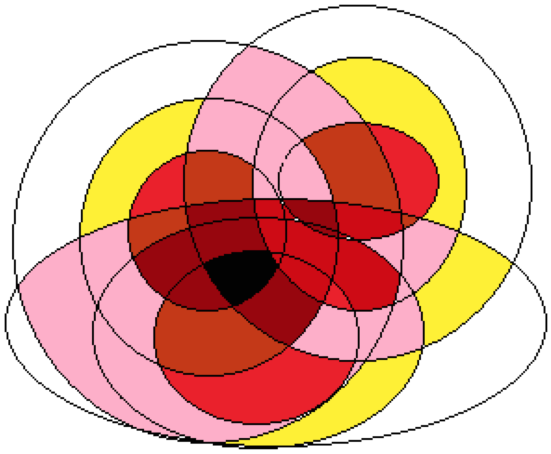

Figure 3.6: Overlapping graph under node-centric point of view

Definition 33. *Overlapping community model aims at describing the communities situated in intensive structural overlapping environment.*

Definition 34. *Overlapping graph model: this is a model to showcase the environment in which a majority of the community structures have severe mutual penetration. In this model, the high and low concentration of edges no longer indicates the community structures and inter-cluster edges respectively.*

This graph model well explains the failure of application of the state-of-art metrics, whose bases are still the graph model I and community formation model I, on social networks these graphs just don't follow the same rules.

Additionally, when we view the community model under our node-centric community formation theory, we may notice the difference layer overlapping brings to the nodes of different roles, as shown in the table below.

Node Role in the original community	Original node property(in graph model 1; without overlapping)	Connections within the community of concern	Connections with the nodes outside of community of concern	Overall Node Properties in graph model 2 context
Core Node	considerable connections to query nodes; relatively high level of conductance	not very much influenced; some increase slightly	Uncertain; Some increase dramatically	considerable connections to query nodes; relatively high(or very high) level of conductance
Boundary Node	considerable connections to query nodes; Low conductance	not very much influenced; some increase slightly	Uncertain; Some increase dramatically	considerable connections to query nodes; some nodes with low conductance, some influenced ones with high conductance
Other Node	considerable connections to query nodes; Limited conductance	not very much influenced; some increase slightly	Uncertain; Some increase dramatically	considerable connections to query nodes; conductance uncertain
Node outside of the community of concern	very limited connections to query nodes; conductance uncertain	Some may increase slightly	Uncertain	some connections to query nodes; conductance uncertain

Figure 3.7: Node property analysis under node-centric point of view

3.5 Boundary-seeker metric

3.5.1 intuition

As the topology rules being confused and complexed by the overlapping of multi-layer graphs/communities, the metric designed for single-layer graphs and traditional community structures suffer serious problem when dealing with them. Even the effective core-seeker metric we proposed earlier wouldn?t help. And the reason for that is trivial: the connection information of core node has now modified greatly, hence the result of core-seeker metric here might be simply a combination of nodes of different kinds. This idea is proved by experiment later.

Node Property	Possible Roles
Big internal edge amount; Big external edge amount	Core Nodes; Influenced Boundary Nodes; Influenced other nodes; outside nodes;
Big internal edge amount; Small external edge amount	Boundary nodes; outside nodes;
Small internal edge amount; Small external edge amount	Outside nodes;
Small internal edge amount; Big external edge amount	outside nodes;

Figure 3.8: Node property with their possible roles under node-centric point of view

It's worth noting that we are not stating the nodes belong to other parts cannot have high internal connection and low external; the information above are based on qualitative statistical analysis. And that means, the chance of a node with small internalexternal edge amount being of outside nodes is larger than that being of other types.

Metric	Community Structural Feature	Intuition	Metric
Boundary-seeker	The boundary nodes should have intensive connections internally, and relatively low conductance; Some of the boundary node would have significant increase in conductance while some don't	the nodes with Big internal edge amount as well as small external edge amount are very likely to be boundary nodes	e(S) (-p(S))

Figure 3.9: Boudary-seeker metric under node-centric point of view

However, when we try to have a look of above from another perspective, we can

get information above. Especially it is very interesting to notice that, when a node is associated with a big internal edge amount as well as small external amount, the chance is big that it falls into the category of boundary nodes or outside nodes. This information seems useless at a first glance, however, given that the whole traditional community features exist no more under the overlapping environment, it would do us great favour if we can get ourself back this group of nodes.

3.5.2 The Design of boundary-seeker metric

With the analysis we did above, we then raise up our metric dedicated to identify clusters in this complex overlapping environment(multi-layer graphs). This metric, called boundary-seeker, focuses on the identification of the boundary nodes of the community containing query nodes. The reason is that the boundary nodes are the only target we can possibly shoot at in the overlapping environment as shown above. Especially, the distinguish between the real boundary nodes and outside nodes can be left for further processing, under the SLUB framework to be introduced in the next section.

The whole idea is, in every iteration, we first compare the internal edge amount between the candidates; and then select the ones with the least external connection among the most internally connected nodes.

3.5.3 Effectiveness Evaluation

We did experiment with the metric on social network graphs, such as Youtube and LJ. And the result shows that they do perform superiorly than state-of-art metrics. The experiment result are given in 'experiment discussion' section.

3.6 Collaboration of Multiple Metrics

There are a number of main elements in the formation of metrics, and the difference in elements adoption and weight of them in the metric formation makes disparity between the results of them. And this is when the collaboration of metrics makes sense.

For example, the metrics focusing on the optimisation of internal edge density, such as edge density or edge surplus, are quite likely to include local free riders, as suggested by [Yubao]; meanwhile, a few other metrics (maybe metrics in global methods), such as distance threshold [The Community-search Problem and How to Plan a Successful Cocktail Party] are sensitive to a lot of mistakes but the free rider. A proper combination of them has the potential of making a better compound metric for local methods.

And in this section we will introduce several common metrics, other than the ones discussed above, which are repeatedly made use of in the compound metrics for their special features; and we will present two main schemes of multiple metrics collaborating.

3.6.1 Assisted Metrics

They could not be used directly for our purpose, still they can make good assisted schemes with local detection metrics.

3.6.1.1 Distance to the seeds

Intuitively enough, as the nodes in a same community as the seeds, they are not supposed to be well separated from the seeds themselves. Though the entities quite close to the seeds still have a good chance not staying in the same communities with them, the possibility for those far away from them is close to no. And the idea is well supported particularly in social networks, where the small-world theory illustrates it that the community should be quite limited in diameter (longest shortest path).

Based on the viewpoints above, [The Community-search Problem and How to Plan a Successful Cocktail Party] raised a distance constraint, which can be defined as follows. First let dG(v, q) denote the length of the shortest path between nodes v and q in the graph G. If v and q are in different connected components, then we define dG(v, q) to be infinity. Now, given a node v in the graph G, we define the distance of v from the query nodes Q to be DQ(G, v) = q?Q dG(v, q)2, (1) and we also define DQ(G) = max v?V (G) DQ(G, v), (2) the distance of the furthest node from the query nodes. For defining DQ(G, v) other alternatives are possible, for instance, not using squares, or using max instead of .

While [Yubao] adopts a different scheme in achieving this. They first compute the proximity value of each node with regard to the query nodes. The reciprocal of the proximity value is used as the node weight, thus the nodes closer to the query nodes will have smaller weights. Particularly, they chose a variant of degree normalised penalised hitting probability in realising it.

Although these solution may not accurate,mainly due to its failure in well taking the denseness of edges between the nodes inside the community (other than seeds themselves) into account, it can make a good weighting scheme for later processing in local community detection methods. And this has been proved by the good performance of [Yubao] algorithm.

3.6.1.2 Edge Betweenness

Another metric of this kind is the edge betweenness, served as a map of one of community key features: sparse inner shortest path distribution. [Network community-detection enhancement by proper weighting] discusses the action in detail. The basic idea of that is, the paths that connect vertices of distinct communities must pass through at least one inter-cluster edge. Bearing in mind the fact that the communities are loosely connected, one can expect that the inter-cluster edges have usually rather high EBC scores. On the other hand, the vertices within a community are tightly connected, so the betweenness centrality of intra-cluster edges is usually smaller[Network community-detection enhancement by proper weighting]. Hence a

reference to the edge betweenness distribution would make the difference between internal and external edges more clear.

3.6.1.3 Metric used in deletion

This is a little bit different from above, in which the metrics to be used would be still the state-of-art metrics ; however, the aim for the comparison is to pick up from the detected subgraph a node that could mostly improve it's performance in terms of metric value. The act is introduced by [subgraph modularity].

3.6.2 Collaboration Schemes

There are normally two ways to set up the chemical reaction on local detection metrics and the assisted metrics, respectively the weighting scheme and iteration scheme. And we will have a closer look at them, especially the latter which is another focus of our newly raised method.

3.6.2.1 Weighting Scheme

The basic function of the weighting scheme is to give nodes or edges of certain property a proper weight so as it is more or less likely to be recognised as a part of a community during the application of the next metric. The weighting scheme usually goes like this: the weighting metric is initially applied on the graph globally, following the algorithm with a second metric.

One example of this kind is the Query Biased Density(referred to as QBD in the rest part) metric proposed in [Robust Local Community Detection: On Free Rider Effect and Its Elimination]. QBD is essentially a metric focused on the internal edge density , with pioneering consideration to the diameter of the community structure. On observing that the local community is not very likely to include the nodes quite far away from it, QBD involves a node weighting scheme based on random walk.

Intuitively, if all the nodes weight 1, the query biased density is essentially the same as the classical density definition.

Another example is one of the weighting schemes adapted in the [Network community-detection enhancement by proper weighting], with regard to the edge betweenness centrality. The metrics is one based on modularity. And with the intuition that the nodes very frequently involved in the shortest paths between nodes are probably the intra-cluster nodes, the links to the nodes are somehow weighted for later computation. This method would not be discussed in detail since it doesn't have a lot to do with our work.

The weighting scheme is proved to be effective in improving community detection accuracy considerably, at least with the two examples given above. And that from a side delivers the message that the collaboration of metrics dedicated in different aspects of community structural characteristics is capable of bettering the result.

3.6.2.2 Iteration scheme

Besides the weighting scheme, another pattern of metric collaborating is also commonly accepted, which we name the iteration scheme. With local community detection algorithm of iteration scheme, the process would be repeated a couple of times with various metrics. It?s worth-noting that in iteration scheme, the latter iterations of the detection would be directly relied upon the results from the previous. The iteration scheme usually goes like this: the detecting algorithm with the first metric is applied on the graph, with a subgraph as its outcome; subsequent algorithms with other metrics are then used with the information on the initial graph as well as the outcome subgraphs of prior iterations.

The current number of research adapting the iteration scheme in the field is not great, however, as what we would prove with our own method later in the next section, the iteration scheme can be a good way to make metrics of different kind work together cohesively.

3.7 The SLUD framework

At last, we are going to put forward the SLUD framework in this section. SLUD framework is one capable of precisely describing almost every existing method in the local community detection field. It supports the weighting scheme and iteration scheme, with every weighting algorithm/iteration algorithm specified with its particular goodness metric and stop sign.

And this framework not only gives us a overall idea of the local community detection field, but also the possible ways of improvement.

We are now giving out the framework to begin with, followed by explanations and example at later sections.

The soul of the SLUD framework is to enhance collaboration between metrics. This collaboration can be expressed by not only their instinct complementarity, but also the extra and organised information multiple rounds of algorithms could bring.

3.7.1 SLUD framework

As shown in above, the SLUD framework is one utilising multiple metrics, and getting out of them a better-off combined method by selectively adapting weighting or iteration schemes.

3.7.2 State-of-art algorithms with metrics, under the viewpoint of SLUD framework

This is a good example of the application of SLUD metric, in which we further upgrade the core-seeker metric. By regarding it as a reasonable choice for the iteration scheme and end up with a bigger accuracy-guaranteed node set, the overall performance of the detection better off, as shown in the next part.

Chapter 4

experiment and discussion

We perform evaluation experiments to determine the effectiveness and efficiency of the proposed metrics,namely core-seeker and boundary-seeker, using a variety of real graphs with ground-truth communities know. All the programs are written in C++. All experiments are performed on local computer with 16G memory, 2.5 GHz Intel Core i7 CPU.

The statistics of the real networks used in the experiments are shown in right hand side Table . These datasets are provided with ground-truth community memberships and are publicly available at http://snap.stanford.edu.

Datasets	Abbr.	#Nodes	#Edges	#Communities
Amazon	AZ	334,863	925,872	151,037
DBLP	DP	317,080	1,049,866	13,477
Youtube	YT	1,134,890	2,987,624	8,385
Orkut	OR	3,072,441	117,185,083	6,288,363
LiveJournal	LJ	3,997,962	34,681,189	287,512
Friendster	FS	65,608,366	1,806,067,135	957,154

Figure 4.1: Test data sets information

We compare our core-seeker metric with several state-of-art local community detection metrics, which are summarised in Figure 4.2

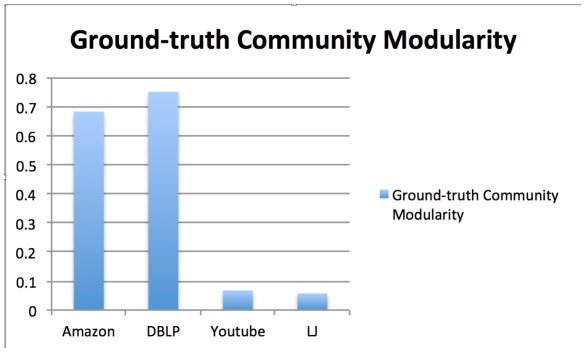

Figure 4.2: Ground-truth Community subgraph modularity

Before we move to the experiments, we need to bring about another experiment. The result of test on the subgraph modularity value on the ground-truth communities

in Amazon, DBLP, Youtube and LJ shows that, the former two graph data sets follow the definition of a typical graph model, where the community structures are clear and interacting less; while the latter graphs, Youtube and LJ, suffer a loss in an order of magnitude in subgraph modularity. Hence in the experiments we use Amazon and DBLP for the purpose of testing metric performance in traditional community structure environment; while Youtube and LJ would be used when we need evaluation on overlapping environments.

4.1 Evaluating Criteria

To evaluate the performance, we use three criteria to evaluate the selected methods: precise, recall and F-score, the definition of the which has been given in the introduction part. And it's their realistic significance that is noteworthy: precise measures the correctness of the detected community; recall weights how big part of the ground-truth community has been identified by the detection algorithm; while F-score is an overview of the precise and recall, often used as the core evaluation standard.

4.2 The Evaluation and Comparison on the New Metrics

The utilisation of edge surplus metric will be universally associated a $\alpha being 1/3, assuggestedintheexperimentof - artmetricofourknowledgetocomparewiththenewlyraisedmetric.$

4.2.1 The Evaluation and Comparison on the Typical Graphs with Core-seeker Metric

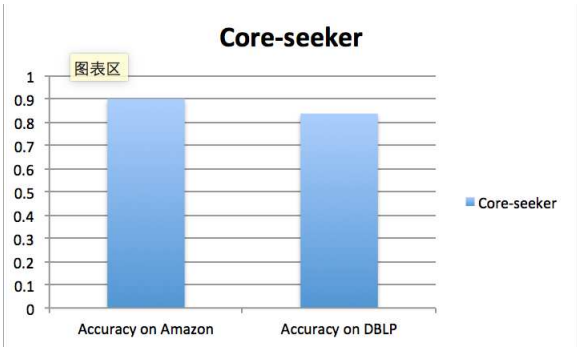

Figure 4.3: Accuracy of core-seeker test

The result shows that the core-seeker metric is a stable and broadly applicable metric, which doesn't rely on a lucky or experienced choice of any parameter. But its results are ideal either as the result of a core-node-seeking algorithm, or as the first iteration metric(under SLUD framework) targeting at a subgraph of small size but high accuracy.

Particularly some may notice the edge surplus metric perform equally as the core-seeker metric on Amazon graph, the reason for that is Amazon is a very typical single-layer graph. Hence, normally as long as a metric chooses the edge density as its major component and limit the expansion of detection process, the result quality is almost guaranteed. Edge surplus with $\alpha being 1/3 is surely one of those$.

4.2.2 The Evaluation and Comparison on the Overlapping Graphs with Boundary-seeker Metric

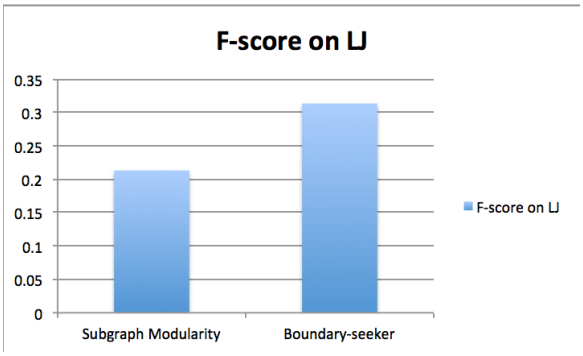

Figure 4.4: boundary-seeker F-score test 1

4.2.3 The Evaluation and Comparison on the Overlapping Graphs with Boundary-seeker Metric

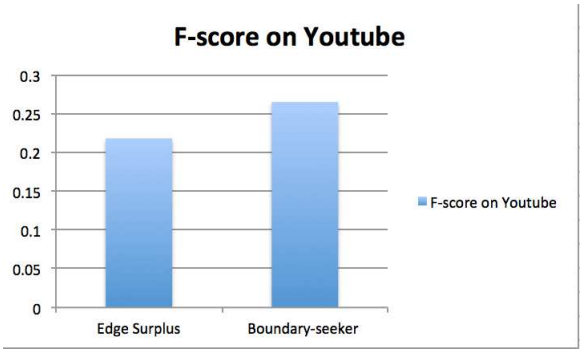

Figure 4.5: boundary-seeker F-score test 2

The result shows that the boundary-seeker metric is a stable and broadly applicable metric, which doesn't rely on a lucky or experienced choice of any parameter. But its results are ideal either as the result of a core-node-seeking algorithm, or as the first iteration metric(under SLUD framework) targeting at a subgraph of small size but high accuracy.

4.3 The Evaluation of the SLUD framework

The evaluation of weighting scheme under SLUD with the collaboration of hitting probability metric and edge density metric has proved its superiority in [Yubao]. And we are going to evaluate the application of SLUD framework on the enhancement of detection on typical graphs here.

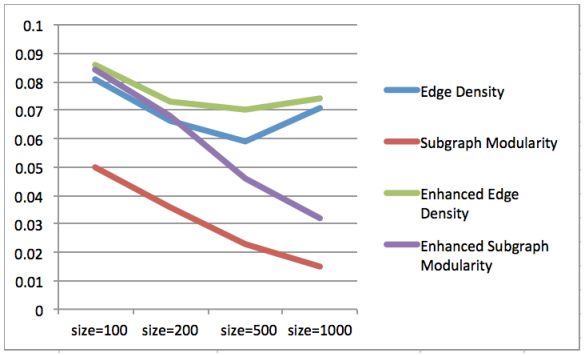

Figure 4.6: SLUD framework testing: core-seeker as pre-stage dealer

These results demonstrate that the metric collaboration has its significance in optimising the local community detection.

This result proved that the result of the edge surplus are more likely to be core node in the communities; Hence even if the accuracy of core-seeker is slightly second to edge surplus, when we need to expand the detected community from the first iteration result, core-seeker shows good superiority as first round metric .

Chapter 5

Conclusion

Local community detection is a fundamental problem in network analysis and has attracted intensive research interests. However, in many occasions the lack of information and the complexity and limitation on expressibility of graphs themselves make the problem challenging.

In this work, we try to analysis the community structure with an alternative viewpoint(node-centric viewpoint), and proposed two new metrics dedicated to better identify cluster structures in single-layer and overlapping environments respectively base on the new point of view. And the experiment results have shown the effectiveness of the new metrics.

Most importantly, we then propose a general framework of the local community detection. This framework well and precisely describes most of the state-of-art attempts and innovations in the field as discussed. With the help of the SLUD framework, the further study in the area may acquire some convenience, better organisation of ideas and make better use of current metrics by collaborating them in some new way.

Bibliography

(p. 7)
2001. *Cluster-Based Networks*. (Addison Wesley, Reading, USA). (p. 11)

AGRAWAL, R. 2011. Bi-objective community detection (bocd) in networks using genetic algorithm. *Contemporary Computing Communications in Computer and Information Science Volume 168, 2011, pp 5-15*. (p. 11)

ALBERT, R. AND BARABASI, A.-L. 2002. Statistical mechanics of complex networks. *REVIEWS OF MODERN PHYSICS, VOLUME 74, JANUARY 2002*. (p. 3)

BOLLOBÁS, B. 1998. *Modern Graph Theory*. Springer Verlag, New York, USA. (p. 2)

BRANDES, D. . G. M. . G. R., U; DELLING. 2008. On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on (Volume:20 , Issue: 2)*. (p. 14)

CLAUDIO CASTELLANO, S. F. AND LORETO, V. 2009. Statistical physics of social dynamics. *Rev. Mod. Phys. 81, 591 – Published 11 May 2009*. (p. 18)

CLAUSET, A. 2005. Finding local community structure in networks. *Phys. Rev. E 72, 026132 – Published 29 August 2005*. (p. 13)

COLEMAN, T. F. AND MORÉ, J. J. 1983. Estimation of sparse jacobian matrices and graph coloring blems. *SIAM J. Numer. Anal., 20(1), 187–209.* (p. 7)

ERDOS AND I, R. 1959. On random graph. (p. 4)

FORTUNATO, S. 2010. Community detection in graphs. *Physics Reports Volume 486, Issues 3–5, February 2010, Pages 75–174*. (pp. 8, 12)

J FRIEDMAN, R. T., T HASTIE. 2001. *The elements of statistical learning*. (p. 18)

JONSSON PF, Z. D. B. P., CAVANNA T. 2006. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics. 2006 Jan 6;7:2*. (p. 1)

KWAN HUI LIM, A. D. 2013. A seed-centric community detection algorithm based on an expanding ring search. *AWC '13 Proceedings of the First Australasian Web Conference - Volume 144 Pages 21-25*. (p. 14)

LR FORD, D. F. 1956. Maximal flow through a network. *Canadian Journal of Mathematics 1956*. (p. 18)

QIRONG HO, A. P. P. . E. P. X. 2012. Multiscale community blockmodel for network exploration. *ournal of the American Statistical Association Volume 107, Issue 499, 2012*. (p. 9)

R. EDWARD FREEMAN, B. P., ANDREW C. WICKS. 2004. Stakeholder theory and "the corporate objective revisited". *Organization Science Vol. 15, No. 3, May–June 2004,Organization Science, pp. 370–371*. (p. 5)

SANTO FORTUNATO, C. C. 2012. Community structure in graphs. *Computational Complexity 2012, pp 490-512*. (p. 14)

SCOTT, J. 2000. *Social Network Analysis: A Handbook*. (SAGE Publications, London, UK). (p. 3)

SHAI CARMI, S. K. Y. S. E. S., SHLOMO HAVLIN. 2007. A model of internet topology using k-shell decomposition. *PNAS July 3, 2007 vol. 104 no. 27 11150-11154*. (p. 2)

SIMON, H. A. 1991. The architecture of complexity. *Facets of Systems Science International Federation for Systems Research International Series on Systems Science and Engineering Volume 7, 1991, pp 457-476*. (p. 12)

TIANTIAN ZHANG, B. W. 2012. A method for local community detection by finding core nodes. *ASONAM '12 Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) Pages 1171-1176*. (p. 13)

WANG, J. 2000. On network-aware clustering of web clients. *SIGCOMM '00 Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication Pages 97-110*. (p. 11)

YUBAO WU, J. L. X. Z., RUOMING JIN. 2015. Robust local community detection: on free rider effect and its elimination. *Proceedings of the VLDB Endowment*. (p. 10)