AUT Journal of Mathematics and Computing

AUT J. Math. Comput., 3(2) (2022) 173-184
DOI: 10.22060/AJMC.2022.21715.1103

Review Article

A survey on hierarchical community detection in large-scale complex networks

Mojtaba Rezvani*[a], Fazeleh Sadat Kazemian[a]

[a]*College of Engineering and Computer Science, Australian National University, Canberra, Australia*

ABSTRACT: Vertices in a real-world social network can be grouped into densely connected communities that are sparsely connected to other groups, and these communities can be partitioned into successively more cohesive communities. Given the ever-growing pile of research on community detection, various researchers have surveyed the evolution of various community detection methods such as flat community detection, overlapping community detection, dynamic community detection and community search. Yet, the problem of hierarchical community detection, despite being well studied, has not been surveyed and the evolution of methods to identify hierarchies of communities in large-scale complex networks has not been documented. In this survey, we study the hierarchical community detection problem and formally define this problem. We then classify the existing works on hierarchical community detection and discuss some of the flat community detection approaches that are capable of producing hierarchies. We then introduce a set of empirical analysis tools, such as benchmark datasets and accuracy measures to evaluate the performance of a hierarchical community detection method.

Review History:
Received:22 August 2022
Accepted:28 August 2022
Available Online:01 September 2022

Keywords:
Hierarchical community detection
Large-scale networks
Complex networks

AMS Subject Classification (2010):
91D30; 90-02

(Dedicated to Professor S. Mehdi Tashakkori Hashemi)

1. Introduction

A *network*, which is also referred to as a graph in mathematics, consists of a set of entities, called *vertices*, and the relationships between those entities, called *edges*. Networks were studied as early as 1735 A.D., when Euler solved the Seven Bridges of Königsberg problem using a network model in which each land mass area was represented as a vertex, and each bridge was represented as an edge. Nowadays, networks are used as a modelling tool in a wider range of applications, such as transportation, communication infrastructures, power grids, information flow, social interactions and prediction of prospective friendships between people [4].

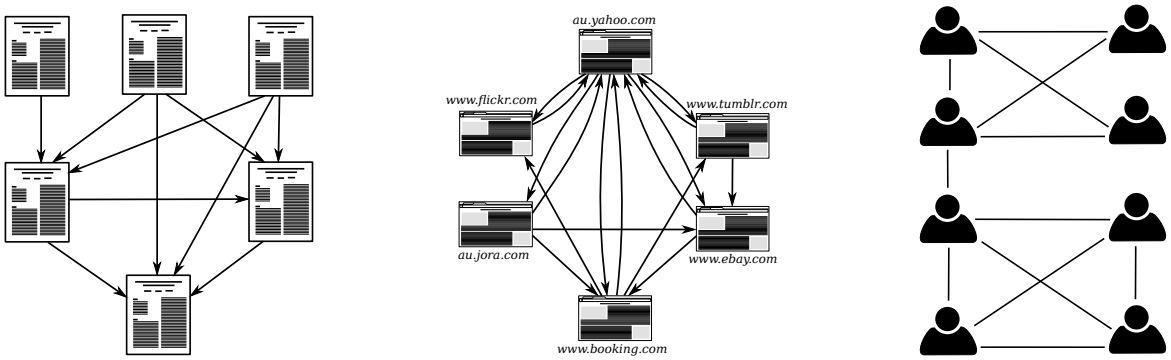
The term *complex network* refers to a network that has a non-trivial topological structure which does not appear in simple networks such as lattices and cliques but frequently occurs in real-world networks.

Complex networks are quite prevalent these days, as a networks are used as common representation for a variety of complex systems [10, 38] such as information networks [40], technological networks [3], and biological networks [9]. A social network of people is an instance of a complex network, where members of the social network represent the set of vertices, and different types of relationships between members such as friendship, follower/followee, messaging and endorsement represent the set of edges of this network. The World Wide Web forms a network where webpages are the vertices, and the hyperlinks among webpages form the edges. Fig. 1a illustrates a small part of the complex network of webpages in the World Wide Web. In a similar manner, citation among research articles forms a complex

*Corresponding author.
E-mail addresses: mojtaba.rezvani@alumni.anu.edu.au, fazelehsadat.kazemian@anu.edu.au

network, where each research article represents a vertex, and every citation from one article to another represents an edge between the corresponding vertices. Fig. 1b represents a small network of research articles and the citations between them. In biology, researchers have created networks of proteins based on the chemical interactions between proteins, where proteins are represented by vertices, and there is an edge between two proteins if there is a certain chemical interaction between those proteins. The availability of such networked data has provided us with an opportunity to understand the underlying structure of these complex systems.

Despite the differences in the way these complex networks are constructed, they all share several characteristics. One of the interesting characteristics is the small-world phenomenon, which states that the longest distance between any pair of vertices is usually a small constant [41]. Furthermore, navigability is one of the important characteristics in small-world networks [17]. Complex networks often share other characteristics such as clustering coefficient and power-law degree distribution, which can help us to make sense of the non-trivial nature of such networks.



(a) Citation network (b) World Wide Web network (c) Social Network

Figure 1: Examples of complex networks. (a) shows a citation network, which consists of research papers as vertices and citations as edges; (b) illustrates a world Wide Web network, in which webpages are represented by vertices, and hyperlinks are represented by edges; (c) depicts a social network, where users are represented by vertices and there is an edge between two vertices if there is a direct friendship between users.

It is well-known that communities in a network often exhibit a hierarchical structure [10, 28, 34, 36]. For instance, metabolic networks of organisms can be decomposed into highly connected communities, where communities form a hierarchy in which communities at lower levels of the hierarchy are more cohesive, and vertices within those communities are closer to each other [28]. Researchers in a collaboration network can be grouped into communities based on their research areas, from general areas such as computer science to more specific ones such as database and data mining, where information circulates more quickly among them. Therefore, small and cohesive communities are nested into larger and less cohesive communities in a hierarchical manner. Fig. 2 shows a hierarchical structure of communities detected in a real-world network Amazon, where vertices form a hierarchical structure in which the density of edges between vertices becomes higher as we navigate the hierarchy from top towards the leaves.

There have been a significant amount of work devoted to community detection in complex networks, and a noticeable proportion of the literature is either directly aimed at identifying hierarchical structure of communities or capable of identifying the hierarchical structure of communities. While the previous surveys on community detection [10, 36, 44] have described the hierarchical community detection problem in details due to a lack of literature at the time, they are not inclusive of the most recent developments on this problem. The purpose of this survey is to cover some of the most influential works on hierarchical community detection and document the evolution of academic works on this problem. In addition this document introduces a set of real-world benchmark datasets for scalability testing and performance measurement criteria to evaluate the accuracy of the hierarchical community detection approaches. The major contributions in this survey are as follows,

- The problem of hierarchical community detection is formally defined in this survey.
- The most influential, novel approaches in hierarchical community detection are categorised and discussed.
- The benchmark datasets and a set of accuracy measurement criteria are introduced in this survey.
- A discussion on the related works and the future directions is provided.

The rest of this survey is organized as follows. Section 2 introduces preliminaries and Section 3 covers the categorisation and discussion of the most recent and influential works on hierarchical community detection. Section 4

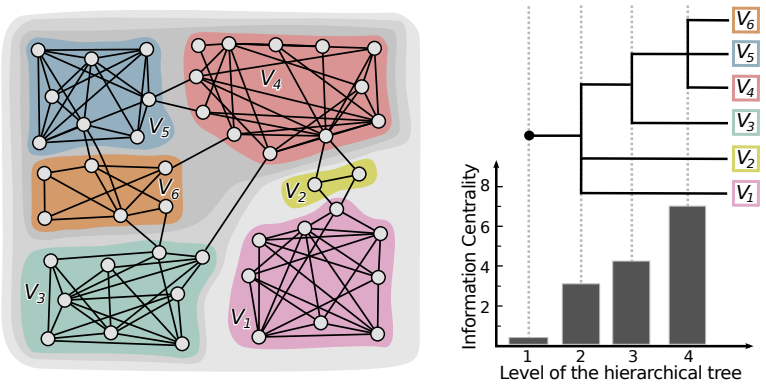


Figure 2: A hierarchical structure of communities in Amazon, where from the root to its leaves connections within communities become denser and the values of their information centrality increase.

introduces the benchmark datasets and measurement criteria. Section 6 and Section 5 discuss the related works and the future directions in the hierarchical community detection domain, respectively.

2. Preliminaries

A network can be modeled as an undirected connected graph $G = (V, E)$, where V is the set of vertices representing individuals and E is the set of edges representing relationships between individuals. Let $n = |V|$ and $m = |E|$. Denote by $\Gamma(v)$ the set of neighbours of vertex $v \in V$. The *degree* of a vertex v is the number of edges incident to it, denoted by $|\Gamma(v)|$. The *distance* between two vertices u and v in a graph G, denoted by $d(u, v)$, is the length of the shortest path between them. We have $d(v, v) = 0$ for any vertex $v \in V$. The *information centrality* of G, denoted by $\mathcal{D}(G)$, is the inverse of mean distance between every pair of vertices u and v [11, 30].

Let $E(S, T)$ be an edge cut between subsets S and T of vertices and $e(S, T)$ the cut size, i.e., $e(S, T) = |E(S, T)|$. For brevity, we simply write $E[S]$, whenever $S = T$. Two paths in G are called *edge-disjoint* if they do not share any edges. The number of edge-disjoint paths between two vertices u and v is the *edge-connectivity* between them, denoted by $\lambda(u, v)$.

Traditionally, communities are perceived as subsets of vertices of a graph G that the number of edges among them (density of connections) is large. We define the *flat community detection* as the problem of identifying a collection $\mathcal{C} = \{V_1, V_2, \ldots, V_{|\mathcal{C}|}\}$ of communities in G. The power set of a given set V, denoted by 2^V, refers to the set of all subsets of V, which is a super set of flat communities.

Following this perception, it is possible to find hierarchical communities recursively. Specifically, we here represent the hierarchy of communities as a rooted tree of subsets of vertices in G. Given two partitions $P = \{V_1, \ldots, V_{|P|}\}$ and $P' = \{V'_1, \ldots, V'_{|P'|}\}$ of V , we say that P has a higher *hierarchical order* than P', denoted by $P \succ P'$, if for every set $V'_i \in P'$ there is a strict superset $V_j \in P$ that includes V'_i, i.e. $V'_i \subset V_j$. Given a hierarchy $\mathcal{P} = \{P_1, \ldots, P_t\}$, we refer to P_t the lowest level of hierarchy and we refer to P_1 as the root of the hierarchy. For every partition $P_i \in \mathcal{P}$, we say that $P_j \in \mathcal{P}$ is at a lower level if $j > i$.

Example 1. *Let us consider the network illustrated in Fig. 2, where V is the set of vertices of the network and V_i is the set of vertices in subgraph G_i $(1 \leq i \leq 6)$. We show that $\mathcal{P} = \langle P_1, P_2, P_3, P_4 \rangle$ is a cohesive hierarchy, where $P_1 = \{V\}$, $P_2 = \{V_1, V_2, V_3 \cup V_4 \cup V_5 \cup V_6\}$, $P_3 = \{V_1, V_2, V_3, V_4 \cup V_5 \cup V_6\}$, and $P_4 = \{V_1, V_2, V_3, V_4, V_5, V_6\}$.*

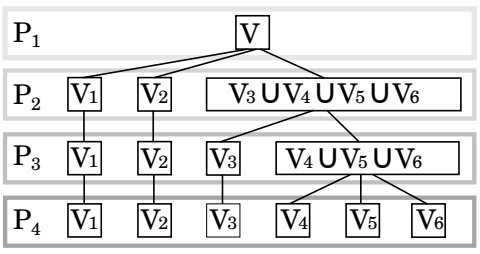


Figure 3: The hierarchy of communities in the network of Fig. 2

3. Hierarchical community detection approaches

In recent years, considerable efforts have been devoted to build efficient metrics and models that can accurately capture the properties of communities in real-world complex networks. In their comprehensive surveys, Xie *et al.* [44], Fortunato [10] and Shaeffer [36], surveyed state-of-the-art algorithms for community detection. Existing methods for detecting the hierarchical structure of communities in complex networks can be categorised into four major groups: (1) *fitness-metric based*, (2) *hierarchical graph clustering*, (3) *random walk models* and (4) *structured approaches*. In this section, we review some of the most influential works in each of these categories and further categories each group of works.

3.1. Fitness-metric based algorithms

A significant portion of the literature on community detection has focused on utilising a fitness metric that can quantify the quality a community in a network. An optimisation algorithm is then used to find the communities in network that optimises the fitness metric. Therefore, these works differ mainly in two aspects: (1) The definition of the fitness metric in terms of the network vertices and edges, and (2) The computational approach that have been used to optimise the fitness metric. In this section we discuss these two aspects in details.

3.1.1. Fitness metrics

A large number of fitness metrics have been developed to help measure the quality of a given community $C \subseteq V$ or a collection of communities $\mathcal{C} \subseteq 2^V$. Such fitness metrics are usually defined in terms of the density of connections inside a community and the sparsity of connections between two communities. The following list introduces some of the well-adopted fitness metrics in community detection.

- *Classic density* $\delta(C)$ of a community C [33] is referred to as the average degree of vertices within the community C, i.e., $\delta(C) = e(C)/|C|$, where $e(C)$ is the number of edges in the subgraph induced by vertices in C.
- *Relative density* $\rho(C)$ of a community C [21] is referred to as the ratio $e(C)$ of the number of edges in community C to the number of edges that have at least one vertex in C, i.e., $\rho(C) = e(C)/(e(C) + e(C, V \setminus C))$.
- *Subgraph modularity* $\psi(C)$ of a community C [20, 43] is referred to as the ratio of the number of edges in community C to the number of edges between vertices in C and the vertices in $V \setminus C$, i.e., $\psi(C) = e(C)/e(C, V \setminus C)$. Note that this subgraph modularity [20] is a variant of the traditional modularity [23].
- *Community Information Centrality* $\mathcal{D}(C)$ of a community C [31] is referred to as the average length of shortest paths in the induced subgraph of a community. The information centrality of a community is defined as $\mathcal{D}(C) = \sum_{u,v \in V} d^{G(C)}(u, v)/n(n-1)$, where $G(C)$ is the induced subgraph of G by C.
- *Global modularity* $Q(G)$ of a graph G [22] is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. The global modularity of a graph is defined as $Q(G) = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) I(C_i, C_j)$, where A_{ij} is corresponding element in the adjacency matrix, P_{ij} is the number of edges between i and j in the null model and $I(C_i, C_j)$ is 1 if and only if i and j are in the same community.
- *Global Information Centrality* $\mathcal{D}(G)$ of a graph G [11] is referred to as the average length of shortest paths in a graph. The information centrality of a graph is defined as $\mathcal{D}(G) = \sum_{u,v \in V} d(u, v)/n(n-1)$.

3.1.2. Optimisation Approaches

Various approaches have been used to optimise a fitness metric over a graph or a set of given communities. These optimisation approaches can be categorised into two groups: (1) top-down approaches, and (2) bottom-up approaches, where the former starts from a set of seed communities and expands the seed communities until the fitness metric can no longer be optimised, and the later starts with the set of vertices V as a single community and partitions the set of communities in each iteration until the fitness metric can no longer be optimised.

Top-down approaches. Given an undirected graph G, top-down approaches start with the whole network as a community and break the network into communities by removing edges or vertices, until a certain condition is met. For example, Fortunato *et al.* [11] exploited the information centrality of a network as a criteria for deciding which edges to be removed at each iteration. Fortunato *et al.* [11] suggested to iteratively remove the edges, whose removal will result in the maximum decrease in the information centrality of a network. While the measure used for the information centrality is inaccurate, the time-complexity of the proposed algorithm is $\Theta(nm^3)$, which is infeasible in networks that contain more than a few thousands of vertices.

Local expansion methods are based on growing a community, using a community fitness metric to measure the quality of the community. Whang *et al.* [42] used a personalized PageRank algorithm for finding cuts between communities, where a random walk in a network can start from seeds only. Since the vertices close by seed are more likely to be visited, thereby receive higher ranks and join the same communities. Among the methods, Algorithm LFM [18] chooses random seeds and then expands the seeds until the value of fitness function based on the number of edges in the community is locally maximal. While the fitness metric used in the local expansion methods can capture the community density, it suffers from free rider [43] or separation effect. The proposed method in this paper falls into this category but aims to minimize free-rider and separation effects on the found communities. Bandyopadhyay *et al.* devised an algorithm called FOCS [2], where initially communities are the neighbourhoods of all vertices in the network and these communities are then refined by adding and removing vertices from communities, using local modularity. However, it has been shown that both subgraph and local modularities suffer from free rider and separation effects [30, 43].

Bottom-up approaches. Unlike top-down approaches, bottom-up approaches start with seeds and expand those seeds gradually, until a certain threshold is met. However, one challenge is to choose appropriate seeds for the community expansion phase. For example, the clique expansion method [19] identifies distinct cliques as initial seeds, and then expands the seeds greedily using a local fitness metric. In clique percolations [25], a community is defined as the maximal union of maximal cliques that can reach each other through a series of adjacent maximal cliques. However, since some complex networks, such as collaboration networks, are fundamentally a union of cliques, this model may consider the whole network as a single community. To tackle this issue, Shen *et al.* proposed an algorithm called EAGLE [37], which merges two communities with the maximum similarity into one, where the similarity between two communities is proportional to the number of edges between them. Du *et al.* proposed the algorithm COCD [8], in which seeds are a set of maximal cliques and two maximal cliques are merged if their similarity is positive, where the similarity between two cliques is a proportional to the number of edges between non-overlapping vertices of those cliques. Even though cliques can guarantee a very strong connectivity among its members, it is considered as a very strict condition for real-world communities. Therefore, Lancichinetti *et al.* proposed the algorithm LFM [18], where random seeds are expanded until the value of a fitness function based on the number of edges in the community is locally maximal. Similarly, Whang *et al.* [42] used a personalized PageRank algorithm for finding cuts between communities, where a random walk in a network can start from vertex seeds only. Whang *et al.* [42] suggested the use of vertices with maximum degree, and dominating sets and random vertices as seeds for community expansion. Considering that networks contain many vertices that act as hubs and connect several communities, using vertices as seeds can affect the outcome of the algorithms. Therefore, finding an appropriate seed in bottom-up approaches ends up in an judgement call that is difficult to make, due to difference in topology of networks.

3.2. Hierarchical Graph Clustering

One of the key attributes of the community structure is the *cohesiveness*, which means that the members of the same community exhibit similar characteristics such as interests, location, beliefs, and activities. Therefore, various researchers have attempted to exploit the similarity measures between a pair of vertices to create a hierarchical structure of communities. While a similarity measure can indicate cohesiveness of the members of a community, it does not necessarily imply a direct edge between the associated vertices in the network. One of the core techniques in the clustering approaches that are introduced in this subsection is the definition of a similarity measure between vertices or groups of vertices.

Specifically, a clustering approach consists of two main steps: (1) choosing the similarity measure that needs to be used to cluster vertices, and (2) an iterative algorithm that creates the clusters of vertices using the given similarity measure. Therefore, the first step in the clustering approaches is defining a similarity measure. While there is no universally accepted similarity measure, some measures have shown to outperform others, such as Modularity that focuses on network structure and Markov random walks [32]. Here, a list of widely adopted similarity measures is provided and then some clustering algorithms are discussed.

- Local information centrality [12] - The local information centrality, also known as local closeness centrality, is a measure of closeness based on the distance between one vertex and all other vertices in the graph, i.e. $\mathcal{D}(v) = \sum_u d(u,v)/(n-1)$.
- Jaccard Similarity - The Jaccard similarity between two vertices is defined as the result of the number of common neighbours between those two vertices divided by the total number of neighbours of both those vertices, i.e. $\omega_{u,v} = (\Gamma(u) \cap \Gamma(v))/(\Gamma(u) \cup \Gamma(v))$. Similarly, the Jaccard similarity between two communities is defined as the number of vertices that have a neighbour in both communities, divided by the number of vertices that have a neighbour in either one of the communities, i.e. $\omega_{C,C'} = (\Gamma(C) \cap \Gamma(C'))/(\Gamma(C) \cup \Gamma(C'))$.

- Linkage in weighted graphs [7] - Various linkage measures have been used to show the similarity between two clusters of vertices in a graph. The single linkage is defined as the maximum weight of the link between two clusters, i.e. $\max(u,v) \in E(C,C')\{\omega(u,v)\}$, the complete linkage is defined as the minimum weight of the link between two clusters, i.e. $\min(u,v) \in E(C,C')\{\omega(u,v)\}$, while the average weighted linkage is defined in terms of the weight of the sub-clusters that formed a cluster at a higher level of the hierarchy.

- Cuts in weighted graphs [39] - The normalised cut is a global similarity measure of a partitioning of the graph that is defined in terms of a sum over the weight of the edges that run between each clusters and all other vertices in the network divided by the degree of that cluster, i.e. $ncut(C) = \sum_{k \leq i \leq k} (\omega(Ci, V \setminus C_i)/(\Gamma(C_i))$.

- *Conductance* $\sigma(C)$ of a community C [16] is referred to as the ratio of the size of the edge cut to the minimum of the number of edges that have at least one endpoint in C and number of edges that have at least one endpoint in $V \setminus C$, i.e., $\sigma(C) = e(C, V \setminus C)/\min\{vol(C), vol(V \setminus C)\}$. The smaller values of conductance are preferred for a community.

- *Betweenness Centrality* of an edge is defined in terms of the number of the shortest paths that an edge is bridging divided by the number of overall shortest paths between other pairs of vertices, i.e. $B(e) = \sum_{u,v \in E} \frac{g_e(u,v)}{g(u,v)}$, where, $g(u,v)$ is the total number of paths between nodes u and v, and $g_e(u,v)$ is the number of paths between u and v.

The clustering algorithms can be categories in two groups: (1) Agglomerative algorithms and (2) Divisive algorithms. In the following, the definition and the most influential works in each category are described.

Agglomerative algorithms. The agglomerative approaches are similar to the top-down approaches, where a set of seed communities are expanded, until the similarity has reached a threshold. The agglomerative approaches iteratively merge the set of communities/clusters based on a given graph similarity measure.

Dhulipala *et al.* [7] studied the hierarchical agglomerative clustering on edge-weighted graphs and proposed algorithms for various cluster similarity measures such as single linkage, complete linkage and weighted average linkage. In their paper, a heap data structure is used to provide theoretical guarantees of the performance of the algorithms into sub-quadratic time complexity.

Tabatabaei *et al.* [39] proposes an agglomerative clustering algorithm that strives to minimize the normalized cut (or equivalently, maximize the normalized association). The algorithm proposed is a greedy maximization of normalized association via an agglomerative hierarchical clustering. The algorithm iteratively identifies the hierarchies, where in iteration k, the k-th level of the hierarchy is identified by maximising the normalized association measure.

Divisive algorithms. The divisive algorithms utilise a similarity measure to iteratively split a cluster and create more fine-grained clusters by removing edges connecting vertices with low similarity. While the divisive approaches have a major resemblance to the top-down approaches that used a fitness metric, it must be noted that the two approaches are different due to a naturally different notion for deriving the communities, where the divisive algorithms use a similarity measure to increase the cohesiveness of a community, while the top-down approaches use a fitness metric to enhance the density of a community.

One of the distinguished divisive algorithms is the one proposed by Rattigan *et al.* [27] based on the k-means algorithm and the local information centrality of vertices as a measure of similarity. This algorithm requires the number of target communities k to be provided, as well as a distance measure that maps pairs of instances to a real value, i.e. $D : (u,v) \to R$. The algorithm consists of four phases: (1) randomly designate k instances to serve as "seeds" for the k clusters; (2) assign the remaining data points to the cluster of the nearest seed using D; (3) calculate the centroid of each cluster; and 4) repeat steps 2 and 3 using the centroids as seeds until the clusters stabilize.

Another significant work in divisive clustering is the Iterative Conductance Cutting (ICC) [16], which uses the minimum conductance cuts to iteratively divide clusters into smaller and denser ones. Since it is NP-hard to find cuts with minimum conductance, a poly-logarithmic approximation algorithm is adopted to find cuts in feasible time. Consider the vertex ordering implied by an eigenvector to the second largest eigenvalue of $\sigma(G)$. Among all cuts that split this ordering into two parts, one of minimum conductance is chosen. Splitting of a cluster ends when the approximation value of the conductance exceeds a given threshold.

Newman *et al.* [13, 24] proposed a divisive algorithm that aims at removing the edges with the highest betweenness centrality score, as such edges are known to play a central role in bridging different communities in a complex network. Betweenness centrality is the measure of the proportion of shortest paths between nodes that pass through a particular link. After every iteration of the algorithm, the Betweenness scores need to be calculated

on the residual graph, and the process is repeated. However, since the time-complexity of calculating betweenness centrality in a network is quite high $\Theta(nm)$, the overall time-complexity of this approach is $\Theta(nm^2)$, which is not practical in real networks.

3.3. Random walk models

Random walks have been used in a wide range of applications for network analysis [35], where depending on the application, a certain behaviour from a random walker is considered. In the context of hierarchical community detection, random walks have received a significant amount of attention, due to the flexibility of describing a random walker's behaviour, such as the probability that a random walker leaves a community, the probability that a random walker reaches one vertex from another vertex, and the expected number of edges to be visited while a random walker starts a journey from vertex u and aims at reaching a vertex v. In this subsection, the most influential works in the area of hierarchical community detection using random walks are discussed.

Zhou [48] introduced a distance measure between two vertices u and v, where $\overline{d}(u,v)$ is defined as the average number of edges on the path that is taken by a random walker moving from u to v. Due to the density of connections in a community, the expected distance between two vertices in the same community is expected to be small. Therefore, Zhou introduced two contraction methods to construct the communities in an agglomerative way, i.e. "global attractor" and "local attractor". The global attractor of a vertex u is the vertex v with minimum distance, and the "local attractor" of u its neighbour $v \in \Gamma(u)$ with minimum distance from u. Zhou [49] also extended Symmetric Adjacency Difference [5], i.e. $d_{uv} = \sqrt{\sum_{w \neq u,v} (A_{uw} - A_{vw})^2}$, and replaced the elements of adjacency matrix A_{vw} and A_{uw} with $\overline{d}(v,w)$ and $\overline{d}(u,w)$, respectively. A divisive algorithm is then utilised to identify minimal communities by merging vertices that are closest to each other. The time complexity of the algorithm using both distance measures is in the order of $O(n^3)$.

Zhou and Lipowsky [50] also introduced a biased random walker behaviour, where random walker has a smaller distance with vertices that share common neighbours. A proximity index is then defined that indicates the closeness between every pair of vertices. An agglomerative hierarchical clustering algorithm is then employed to construct the communities. The time complexity of the proposed algorithm is $O(n^3)$. Similarly, Latapy and Pons [26] studied the characteristic of a random walker given bounded number of hops to ensure that a random walker is unlikely to leave a community, and they used an agglomerative clustering algorithm to construct the communities, i.e. the Ward's method [15]. The worst-case time complexity of the algorithm is $O(n^3)$.

3.4. Structural approaches

A significant body of work in community detection has focused on the structural properties of a graph for identifying communities. Such structural properties include degrees, number of triangles, and edge-connectivity, in combination with other metrics such as information centrality. Some of these approaches are capable of creating a cohesive hierarchy of communities. In this subsection, some of the most influential works in this area are reviewed.

One of the most efficient ways for finding both hierarchical and flat communities is the use of vertex degrees in partitioning. The concept of k-core was propose by Zhang *et al.* [46] to distinguish the dense subgraphs in which the minimum degree of vertices is k. The k-cores decomposition of a graph can be identified by removing all the vertices with degree smaller than k, repeatedly. It has been shown that the k-core decomposition can be performed in $O(n+m)$. Since a k-core is a subset of a k'-core, where $k' < k$, one can start with $k = 1$ and iteratively identify k-cores and increment the k after each iteration to find a hierarchy of k-cores.

The key issue on community with degree-based models is that the members in each community have weak connectivity, i.e., they can be disconnected by removing a small number of edges. Cohen [6] suggested the notion of a k-truss, where every edge in a community forms at least k triangles with other edges in the community, and suggested an algorithm with the overall time complexity of $O(nm^{3/2})$. The k-truss can guarantee a strong edge connectivity in graph since they are $(k+1)$-edge-connected –won't be disconnected by removing less than $k+1$ edges. However, several types of networks (such as product-buyer networks) do not have any triangles.

Since edge connectivity is a major concern in the formation of close communities and k-truss components are not generalised to be used in all kinds of networks (such as bipartite networks), a great deal of effort is devoted to find tightly connected subgraphs that cannot be disconnected by removing only a few number of edges. In order to tackle this problem, Zhou *et al.* [51] suggested the notion of k-edge-connectivity in a network, and defined a community as a subgraph, in which every pair of vertices are k-edge-connected. While the k-edge-connectivity is general enough to be applicable to several types of networks, the distance in a k-edge-connected community is not bounded compared with the size of the detected community. Akiba *et al.* [1] proposed efficient algorithm for identifying the k-edge-connected components of a graph using random edge contraction.

Rezvani *et al.* [31] proposed the notion of a cohesive hierarchy as a hierarchy of communities $\mathcal{P}$, where the communities in level k are connected to communities in level $k+1$ by at most $k+1$ edges. In this paper, a

Table 1: Real datasets with their details.

Dataset	Number of vertices	Number of edges	Number of communities
Facebook	4,039	88,234	308
Twitter	81,306	2,420,766	4,065
Google Plus	107,614	30,494,866	468
Amazon	334,863	925,872	14,529
DBLP	317,080	1,049,866	7,556
LiveJournal	3,997,962	34,681,189	12,115
Orkut	3,072,441	117,185,083	9,120

systematic graph sparsification method is used to reduce the size of the graph and find communities at higher levels of the hierarchy, and then the sparse certificates are augmented by the spanning forests of the residual graph to construct denser communities in lower levels of the hierarchy. The approach proposed in [31] has shown to be scalable to networks with hundreds of millions of edges using a single-core personal computer.

4. Empirical analysis for hierarchical community structure

This section is devoted to introduce the large-scale real-world datasets that can be used to evaluate the performance of hierarchical community detection algorithms. In addition, measure of accuracy are introduced that are specifically developed to incorporate the hierarchical structure of communities in accuracy analysis. It must be noted that the datasets and performance measures that are introduced in this section are generalised versions of the datasets and accuracy measures that have been used for flat community detection analysis. Therefore, the accuracy measures and datasets that are introduced in this section can be used for other variants of community detection such as flat community detection and overlapping community detection as well [29].

4.1. Large-scale datasets

We introduce seven real datasets that are publicly available[1], and have been widely used in the literature [44]: (1) Facebook is a subgraph of the social network facebook, where communities are groups of members identified by surveyed users, (2) Twitter consists of 'lists' from Twitter. The social communities are the ground-truth communities in Twitter. (3) Google Plus is a social network in Google+. The groups that are defined by users represent ground-truth communities. (4) Amazon is a network in which vertices are products and there is an edge between two vertices i and j if product i is frequently co-purchased with product j. Products in each category are considered as ground-truth communities, (5) DBLP is a collaboration network of researchers, where communities are defined as journals and conferences, (6) LiveJournal is a friendship network of users in the LiveJournal website. Users can create groups, and these groups are considered as the ground-truth communities. (7) Orkut is the friendship network of Orkut members, where communities are groups created by users, where other users can join each group.

4.2. Accuracy measures

Many graph clustering algorithms perform successive divisions or aggregations of subgraphs leading to a hierarchical decomposition of the network. An important question in this domain is to know if this hierarchy reflects the structure of the network or if it is only an artifice due to the conduct of the procedure.

Measuring the quality of detected communities is challenging, as different metrics lead to different interpretations of communities. We employ F-measure that is widely-adopted in the literature [14, 42, 44, 45, 47] for quantifying the accuracy of detected communities. Let $\mathcal{C}^*$ be the set of ground-truth communities and let C be a detected community. The F-measure of C compared to $C^* \in \mathcal{C}^*$ is defined as follows,

$$F_k(C) = \max_{C^* \in \mathcal{C}^*} \{\frac{(k+1) \cdot p(C, C^*) \cdot r(C, C^*)}{k \cdot p(C, C^*) + r(C, C^*)}\}, \tag{1}$$

where $p(C, C^*) = |C \cap C^*|/|C|$ and $r(C, C^*) = |C \cap C^*|/|C^*|$ are the precision and recall, respectively. To calculate the accuracy of a flat community detection algorithm, one may calculate the average of F_1 and F_2-measures for all detected communities [14, 42, 44, 45, 47]. However, the situation is different for hierarchical community detection algorithms, as communities detected at each level of a hierarchy have different characteristics and can be interpreted

[1] http://snap.stanford.edu/data/index.html

differently. One general rule in hierarchical community detection is that communities at the lower levels are smaller, more connected and more cohesive than the ones at the higher levels. Therefore, we suggest a weighting method in calculating the F-measure of communities at different levels of a hierarchy, which provides us with the ability to put more weight on communities at lower levels. Specifically, we incorporate a weight α_i, called the weight of level i, into F-measure of communities at level i of a hierarchy. Given a detected hierarchy $\mathcal{P} = \{P_1, \ldots, P_{|\mathcal{P}|}\}$, we define the F-measure of $\mathcal{P}$ as follows,

$$F_k(\mathcal{P}) = \sum_{1 \le i \le |\mathcal{P}|} \frac{1}{|\mathcal{P}|} \sum_{C \in P_i} \alpha_i \frac{F_k(C)}{|P_i|}, \qquad \text{where } \alpha_i = \frac{i}{\sum_{1 \le j \le |\mathcal{P}|} j},$$

where the term α_i is called the weight of level i, which is used to emphasize on the lower levels of the hierarchy.

5. Future works

While significant amount of work have been done in developing large-scale benchmark datasets for flat community detection problem, the literature has failed to develop benchmark datasets for scalable analysis of hierarchical community detection. Researchers have mainly relied on small datasets for accuracy measurements and flat community detection benchmarks for scalability analysis of the algorithms. One of the key gaps in the area of hierarchical community detection is undoubtedly the absence of large-scale benchmark datasets.

The current survey only covered the high-level hierarchical community detection problem and the categorisation of the proposed methods in this domain. Due to a lack of literature, the analysis of other variants of hierarchical community detection such as overlapping hierarchical community detection, dynamic hierarchical community detection and hierarchical community search have not been covered in this survey. This paves the way for further development in this space.

6. Related works

Given the exponential growth of social networks and the adoption of graph databases and graph models in the industry, community detection and has received a significant amount of attention in the past few decades and its applications in graph analytics have been widely realised. Numerous approaches with delicate details and intricacies have been proposed to address the community detection in different kinds of networks and for various kinds of applications. These works have been surveyed in four different categories:

- Generic community detection – Fortunato [10] and Shaeffer [36] provided extensive studies of the community detection with its variants. These surveys cover a significant portion of the literature on community detection. While the last decade has witnessed a significant amount of development of research within this area, these surveys are still the main reference for flat community detection and the variants of community detection problem.

- Overlapping community detection – Xie *et al.* [44] provided a study of the overlapping community detection and the vast amount of work that have been developed for identifying communities with overlapping vertices. This survey has mainly focused on overlapping communities and different approaches, such as community expansion methods, that have been widely used to identify the overlap between communities in complex networks.

- Hierarchical community detection – To the best of my knowledge, this is the first survey that aims to cover the hierarchical community detection problem and aims at introducing the high-level approaches that have been adopted to identify the hierarchy of communities in a complex network. Furthermore, this survey introduces accuracy measures that can be utilised to measure the performance of hierarchical community detection methods and various datasets that can be used to measure the quality of algorithms.

References

[1] T. Akiba, Y. Iwata, and Y. Yoshida, *Linear-time enumeration of maximal k-edge-connected subgraphs in large networks by random contraction*, in Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013, pp. 909–918.

[2] S. Bandyopadhyay, G. Chowdhary, and D. Sengupta, *Focs: Fast overlapped community search*, TKDE'15, 27 (2015), pp. 2974–2985.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, *Complex networks: Structure and dynamics*, Physics reports, 424 (2006), pp. 175–308.

[4] J. A. Bondy, U. S. R. Murty, et al., *Graph theory with applications*, vol. 290, Citeseer, 1976.

[5] R. S. Burt, *Positions in networks*, Social forces, 55 (1976), pp. 93–122.

[6] J. Cohen, *Trusses: Cohesive subgraphs for social network analysis*, National Security Agency Technical Report, (2008), p. 16.

[7] L. Dhulipala, D. Eisenstat, J. Łacki, V. Mirrokni, and J. Shi, *Hierarchical agglomerative graph clustering in nearly-linear time*, in Proceedings of the 38th International Conference on Machine Learning, M. Meila and T. Zhang, eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, 18–24 Jul 2021, pp. 2676–2686.

[8] N. Du, B. Wang, and B. Wu, *Overlapping community structure detection in networks*, in CIKM'08, 2008, pp. 1371–1372.

[9] D. A. Fell and A. Wagner, *The small world of metabolism*, Nature biotechnology, 18 (2000), p. 1121.

[10] S. Fortunato, *Community detection in graphs*, Physics reports, 486 (2010), pp. 75–174.

[11] S. Fortunato, V. Latora, and M. Marchiori, *Method to find community structures based on information centrality*, Physical review E, 70 (2004), p. 056104.

[12] L. C. Freeman, *A set of measures of centrality based on betweenness*, Sociometry, (1977), pp. 35–41.

[13] M. Girvan and M. E. Newman, *Community structure in social and biological networks*, PNAS'02, 99 (2002), pp. 7821–7826.

[14] P. K. Gopalan and D. M. Blei, *Efficient discovery of overlapping communities in massive networks*, PNAS'13, 110 (2013), pp. 14534–14539.

[15] J. H. W. Jr., *Hierarchical grouping to optimize an objective function*, Journal of the American Statistical Association, 58 (1963), pp. 236–244.

[16] R. Kannan, S. Vempala, and A. Vetta, *On clusterings: Good, bad and spectral*, J. ACM, 51 (2004), pp. 497–515.

[17] J. M. Kleinberg, *Navigation in a small world*, Nature, 406 (2000), p. 845.

[18] A. Lancichinetti, S. Fortunato, and J. Kertész, *Detecting the overlapping and hierarchical community structure in complex networks*, New Journal of Physics, 11 (2009), p. 033015.

[19] C. Lee, F. Reid, A. McDaid, and N. Hurley, *Detecting highly overlapping community structure by greedy clique expansion*, SNA/KDD'10, (2010), pp. 33–42.

[20] F. Luo, J. Z. Wang, and E. Promislow, *Exploring local community structures in large networks*, Web Intelligence and Agent Systems, 6 (2008), pp. 387–400.

[21] M. Mihail, C. Gkantsidis, A. Saberi, and E. Zegura, *On the semantics of internet topologies*, Tech. Rep. GIT-CC-02-07, College of Computing, Georgia Institute of Technology, Atlanta, GA, 2002.

[22] M. E. Newman, *Coauthorship networks and patterns of scientific collaboration*, PNAS'04, 101 (2004), pp. 5200–5205.

[23] ——, *Modularity and community structure in networks*, PNAS'06, 103 (2006), pp. 8577–8582.

[24] M. E. Newman and M. Girvan, *Mixing patterns and community structure in networks*, in Statistical mechanics of complex networks, Springer, 2003, pp. 66–87.

[25] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, 435 (2005), pp. 814–818.

[26] P. Pons and M. Latapy, *Computing communities in large networks using random walks*, in Computer and Information Sciences - ISCIS 2005, p. Yolum, T. Güngör, F. Gürgen, and C. Özturan, eds., Berlin, Heidelberg, 2005, Springer Berlin Heidelberg, pp. 284–293.

[27] M. J. Rattigan, M. Maier, and D. Jensen, *Graph clustering with network structure indices*, in Proceedings of the 24th International Conference on Machine Learning, ICML '07, New York, NY, USA, 2007, Association for Computing Machinery, p. 783–790.

[28] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Hierarchical organization of modularity in metabolic networks*, Science, 297 (2002), pp. 1551–1555.

[29] M. Rezvani, W. Liang, C. Liu, and J. X. Yu, *Efficient detection of overlapping communities using asymmetric triangle cuts*, IEEE Transactions on Knowledge and Data Engineering, 30 (2018), pp. 2093–2105.

[30] M. Rezvani, W. Liang, W. Xu, and C. Liu, *Identifying top-k structural hole spanners in large-scale social networks*, in CIKM'15, 2015, pp. 263–272.

[31] M. Rezvani, Q. Wang, and W. Liang, *Fach: Fast algorithm for detecting cohesive hierarchies of communities in large networks*, in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 486–494.

[32] M. Rosvall and C. T. Bergstrom, *Maps of random walks on complex networks reveal community structure*, PNAS'08, 105 (2008), pp. 1118–1123.

[33] B. Saha, A. Hoch, S. Khuller, L. Raschid, and X.-N. Zhang, *Dense subgraphs with restrictions and applications to gene annotation graphs*, in Research in Computational Molecular Biology, Springer, 2010, pp. 456–472.

[34] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, *Extracting the hierarchical organization of complex systems*, PNAS'07, 104 (2007), pp. 15224–15229.

[35] P. Sarkar and A. W. Moore, *Random walks in social networks and their applications: A survey*, in Social Network Data Analytics, C. C. Aggarwal, ed., Springer US, Boston, MA, 2011, pp. 43–77.

[36] S. E. Schaeffer, *Graph clustering*, Computer science review, 1 (2007), pp. 27–64.

[37] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, *Detect overlapping and hierarchical community structure in networks*, Physica A: Statistical Mechanics and its Applications, 388 (2009), pp. 1706–1712.

[38] S. H. Strogatz, *Exploring complex networks*, nature, 410 (2001), p. 268.

[39] S. S. Tabatabaei, M. Coates, and M. Rabbat, *Ganc: Greedy agglomerative normalized cut for graph clustering*, Pattern Recognition, 45 (2012), pp. 831–843.

[40] H. Tong, S. Papadimitriou, C. Faloutsos, S. Y. Philip, and T. Eliassi-Rad, *Gateway finder in large graphs: problem definitions and fast solutions*, Information retrieval, 15 (2012), pp. 391–411.

[41] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, nature, 393 (1998), pp. 440–442.

[42] J. J. Whang, D. F. Gleich, and I. S. Dhillon, *Overlapping community detection using seed set expansion*, in CIKM'13, 2013, pp. 2099–2108.

[43] Y. Wu, R. Jin, J. Li, and X. Zhang, *Robust local community detection: on free rider effect and its elimination*, VLDB'15, 8 (2015), pp. 798–809.

[44] J. Xie, S. Kelley, and B. K. Szymanski, *Overlapping community detection in networks: The state-of-the-art and comparative study*, ACM Computing Surveys, 45 (2013), p. 43.

[45] J. Yang and J. Leskovec, *Overlapping community detection at scale: a nonnegative matrix factorization approach*, in WSDM'13, ACM, 2013, pp. 587–596.

[46] Y. Zhang and S. Parthasarathy, *Extracting analyzing and visualizing triangle k-core motifs within networks*, in 2012 IEEE 28th international conference on data engineering, IEEE, 2012, pp. 1049–1060.

[47] Y. Zhao, G. Karypis, and U. Fayyad, *Hierarchical clustering algorithms for document datasets*, Data mining and knowledge discovery, 10 (2005), pp. 141–168.

[48] H. Zhou, *Distance, dissimilarity index, and network community structure*, Phys. Rev. E, 67 (2003), p. 061901.

[49] ——, *Network landscape from a brownian particle's perspective*, Phys. Rev. E, 67 (2003), p. 041908.

[50] H. Zhou and R. Lipowsky, *Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities*, in International conference on computational science, Springer, 2004, pp. 1062–1069.

[51] R. Zhou, C. Liu, J. X. Yu, W. Liang, B. Chen, and J. Li, *Finding maximal k-edge-connected subgraphs from a large graph*, in EDBT'12, 2012, pp. 480–491.

Please cite this article using:

Mojtaba Rezvani, Fazeleh Sadat Kazemian, A survey on hierarchical community detection in large-scale complex networks, AUT J. Math. Comput., 3(2) (2022) 173-184
DOI: 10.22060/AJMC.2022.21715.1103