MORGAN & CLAYPOOL PUBLISHERS

Community Search over Big Graphs

Xin Huang
Laks V.S. Lakshmanan
Jianliang Xu

SYNTHESIS LECTURES ON DATA MANAGEMENT

H.V. Jagadish, Series Editor

CHAPTER 1

Introduction

Communities are ubiquitous in nature and society. Since communities serve as fundamental building blocks of networks, significant work has been done on their automatic detection. In contrast to community detection, the aim of community search is to find communities satisfying a given query. Over the last decade, significant strides have been made in the development of community search techniques, which overcome the expensive computation of global community detection by leveraging local network properties and incorporating the constraints in a given query. The main goal of this book is to introduce recent community search models, problems, algorithms, and techniques, and identify important further research that remains open in this important field. We start this chapter by motivating the study of community search using illustrative examples. In addition, we provide some basic concepts and provide an outline of this book.

1.1 GRAPHS AND COMMUNITIES

1.1.1 GRAPHS

Many complex systems in nature and society can be described in terms of graphs (networks) capturing the intricate web of connections among the units they are made of [139]. Graphs have emerged as a powerful model for representing different types of data. For instance, unstructured data (e.g., text documents), semi-structured data (e.g., XML databases), and structured data (e.g., relational databases) can all be modeled as graphs, where the vertices (nodes) are, respectively, documents, elements, and tuples, and the edges can, respectively, be hyperlinks, parent-child relationships, and primary-foreign-key relationships [93]. In addition, graphs naturally arise in application domains such as biological networks, social and information networks, and knowledge graphs, to name but a few.

1.1.2 COMMUNITIES

Community structures naturally exist in numerous real-world networks such as social, biological, collaboration, and communication networks being just a few examples, as shown in Figures 1.1a–1.1d. A question of interest is how to interpret the global organization of such networks in terms of the coexistence of their structural sub-units (communities) associated with more highly interconnected parts [139]. Identifying these *a priori* unknown building blocks (e.g., groups of people [150, 176], industrial sectors [137], and functionally related proteins [142, 158], etc.) is crucial to the understanding of the structural and functional properties of networks [139].

To illustrate, we introduce three notable examples of communities.

Online Social Networks. One type of network where communities are frequently observed is online social networks. Enabled by the Internet and sparked by the recent advent of online social networking sites such as Facebook, Google+, and Twitter [42], research on community discovery over online social networks has been booming. With the ready availability of large-scale social network data, the research has led to the development of many exciting applications, e.g., social circle discovery and influential community search. In addition, due to the prosperity of smartphone devices, online social networks have led to the rapid growth of geo-social networks (also known as location-based social networks), such as Foursquare, Yelp, Google+, and Facebook Places. In a geo-social network, users are associated with location information (e.g., hometowns and check-in places), and communities consist of users that are closely connected in the social layer as well as spatially proximate in the spatial layer.

Academic Collaboration Networks. A well-known academic collaboration network is the DBLP network, where a vertex represents an author and an edge between two authors indicates a collaborative relationship, i.e., they have co-authored publication(s). In addition, vertices can have attributes that represent the authors' areas of expertise. Communities in the DBLP network may represent a group of authors that frequently collaborate with each other and work on similar topics.

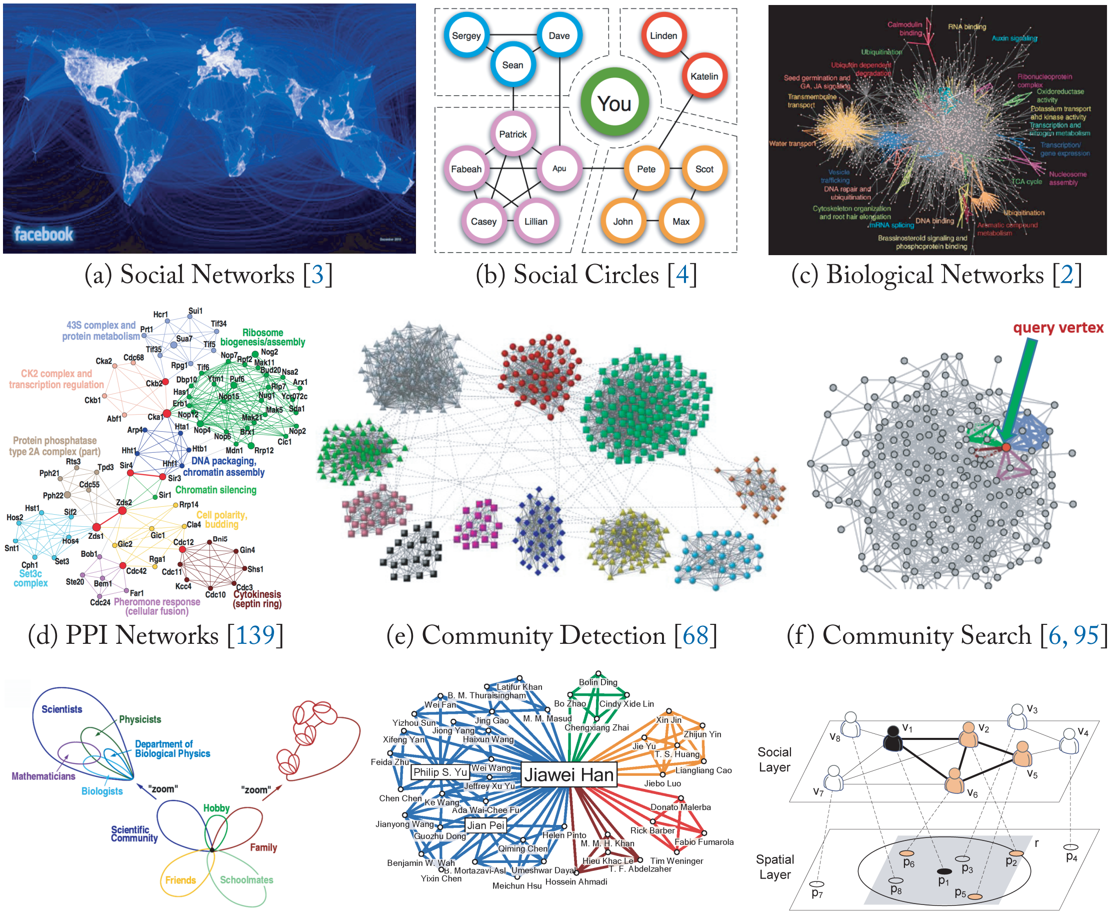
Heterogeneous Information Networks. Heterogeneous networks consist of vertices and edges both of different types [159]. For instance, in a healthcare network, vertices can be patients, doctors, medical tests, diseases, medicines, hospitals, treatments, and so on. On one hand, treating all the vertices as of the same type may miss important semantic information. On the other hand, treating every vertex as a distinct type may miss the big picture. This is a classic example of a heterogeneous network. Such multiple types of objects, interconnected, heterogeneous but often semi-structured information networks, make the communities over heterogeneous information networks complex and interesting.

1.2 COMMUNITY SEARCH

In this section, we provide some background on community search.

1.2.1 COMMUNITY SEARCH PROBLEM

The community search problem is, given one or more query vertices, to find densely connected communities containing the query vertices [96], as illustrated in Figure 1.1f. Since the communities defined by different vertices in a network may be quite different, community search with query vertices opens up the prospects of user-centered and personalized search, with the potential of producing meaningful answers to a user [54]. As just one example, in a social network,



(a) Social Networks [3]

(b) Social Circles [4]

(c) Biological Networks [2]

(d) PPI Networks [139]

(e) Community Detection [68]

(f) Community Search [6, 95]

(g) Overlapping Communities [139]

(h) Community Search on DBLP [89]

(i) Geo-social Group Search [199]

Figure 1.1: **(a)** Facebook social network where two friends have an edge connection. **(b)** Social circles naturally exist in online social networks. **(c)** Different communities in biological networks. **(d)** Different protein complexes in protein-protein-interaction (PPI) networks. **(e)** Community detection is finding all communities over the whole network. Each community is colored differently. **(f)** Community search is to find communities containing the given query nodes in the graph. It is a query-dependent community discovery. **(g)** Communities defined by different nodes in a network may be quite different. **(h)** Five communities containing the author of "Jiawei Han" on DBLP collaboration network. **(i)** Community search on geo-social networks aims at finding densely connected communities within the proximity of given queries.

the community formed by a person's high school classmates can be significantly different from the community formed by his/her family members, which in turn can be quite different from the one formed by his/her colleagues [117], as shown in Figure 1.1g.

1.2.2 A COMPARISON WITH COMMUNITY DETECTION

Community detection, which is to identify all communities in a network shown in Figure 1.1e, is another fundamental and well-studied problem in the literature [6, 46, 68, 133, 135, 139, 144, 154, 183]. Community detection techniques include graph clustering by means of optimizing specific functions (e.g., modularity [135], normalized cut [73, 155], low-conductance cuts [71, 111], personalized pagerank [11, 12, 161]), spectral methods [68, 125, 133, 134], generative models [187, 188, 191], and deep neural networks [193]. Community detection aims at revealing latent community structures from a given network, which provide insight into the underlying interactions and potential functions of networks. Applications include link prediction [34, 100, 115, 156, 164], eradicating social network worms [169], and terrorist group detection [173].

The two problems of community detection and community search have different goals: whereas community detection targets all communities in the entire network and usually applies a global criterion to find qualified communities [89], community search provides personalized community discovery for query vertices in an online manner. Specifically, the following differences between community detection and community search motivate the study of the online community search problem in its own right.

- Community detection finds all communities in a graph in a batched process. In contrast, community search only retrieves query-dependent communities containing the given vertices. To support the online retrieval of communities, the methods of community detection may detect all communities in an offline precomputation stage and construct an index of numerous communities, which is prohibitively costly for massive graphs with million or billions of nodes and edges. For example, Facebook networks had over 800 million nodes and 100 billion links [54] by 2013. As of the third quarter of 2018, Facebook had 2.27 billion monthly active users [1]. Thus, supporting online community detection is very expensive in terms of both space and time.

- Community detection usually uses a global criterion to find all communities in a graph [54], regardless of any notion of query vertices. However, the semantics of discovered communities may vary significantly for different query vertices. For example, in a research collaboration network, the communities of a famous scholar and a junior scholar can be dramatically different in terms of the community size and density. Furthermore, even for a particular researcher, her community w.r.t. her research colleagues can be quite different from her community w.r.t. family and friends, illustrating the difference that query vertices can make.

- It is a challenge for community detection to support dynamically evolving graphs, in which graph vertices and edges can be frequently inserted or deleted. Because it is expensive to re-run the community detection algorithm once the graph changes, the freshness of communities detected cannot be guaranteed [54]. In contrast, online community search provides real-time query services to users.

1.2.3 APPLICATIONS

Community search has many interesting applications in areas such as social network analysis, collaborative tagging systems, query-log analysis, biology, and others [157]. Several representative application scenarios of community search are briefly discussed in the following.

- *Event organization*: Suppose some scientists plan to organize an event like a research workshop. The chances of success of this event would be higher if they can invite a number of well-acquainted scientists that are specialized in this event's topic and with whom they perhaps have even collaborated, possibly together with other participants.

- *Tag suggestion*: In social media websites, many items (photos, videos, and music) are labeled with tags. For instance, in a photo-sharing portal, we can construct a tag-graph as follows: if two tags co-occur frequently in a number of photos, an edge is added between them. Now assume that a new photo is being uploaded, and some initial tags are provided by users. Then the system can suggest a number of additional tags for this new photo. A good suggestion would be the tags that are related to the initial ones and are densely connected to each other.

- *Protein discovery*: A biologist has identified a number of proteins that regulate a gene of interest, and would like to study further a candidate list of other proteins that are likely to participate in the regulation process. Such a candidate set can be obtained by finding a dense subgraph in the protein-protein-interaction network that contains the given proteins [93, 157].

- *Spatial task outsourcing*: Given a set of spatial tasks, each associated with a spatial location, one needs to distribute them to a set of workers, each having a service region. To successfully accomplish the tasks, the service regions of the selected workers should cover all spatial tasks' locations, and the workers are expected to have good collaborative relationships so that the tasks can be efficiently completed. A geo-social community search can address this worker selection problem in spatial task outsourcing, as shown in Figure 1.1i.

Other applications of community search include academic research community discovery [54, 65, 89, 96, 157, 180], influential group search [122], social circle discovery [89, 117], ambiguous name identification [87, 88], analysis of diverse meanings of words [88], and so on. A few examples of networks, communities, and community search are illustrated in Figures 1.1a–1.1i.

1.2.4 DATASETS AND TOOLS

This book introduces five types of real-world graph datasets with ground-truth communities, including simple graphs, attributed graphs, ego social networks, geo-social networks, and public-private collaboration networks. Besides datasets, we also introduce the evaluation metrics and details of query generation, which can be helpful for further study and for evaluating research ideas on community search. Moreover, we list publicly available software tools and demo systems for assisting further study of community search. All these datasets and tools can be found in Chapter 7.

1.3 PREREQUISITE AND TARGET READER

The prerequisite of this book includes basic knowledge of graph theory, network science, data structures and algorithms, and database indexing and query processing. The target reader of this book is anyone who is interested in modeling and searching communities over large graphs, from data mining and data management researchers to practitioners from the industry. Specifically, this book can serve as a textbook for graduate students and as a compact research monograph for a junior researcher to quickly get up to speed and find new research topics on community search. For those new to the area, the book will cover the necessary background material to help understand the topics and offer a comprehensive survey of the state-of-the-art techniques. Moreover, the book aims to provide new perspectives in regards to community search that will be interesting and valuable to the researchers with more experience in the field. For those having worked on classic community detection and graph clustering, we will demonstrate how the problem of community search interacts with commonly used models in terms of algorithmic efficiency and network dynamics, and poses new challenges compared to community detection. For those that have worked on community search, we hope to provide a comprehensive survey of latest work on community search and inspire new research directions by establishing connections with recent developments.

1.4 OUTLINE OF THE BOOK

Community search and its study is the central theme of this book, which consists of eight chapters. On a high level, the book first introduces the basic concepts of communities and networks and then gives an overview of the state-of-the-art research. Community search on different types of networks requires appropriate community models and search algorithms. Each chapter discusses one type of community search on a specific type of networks by presenting detailed community models, the intuition behind them, and the corresponding search algorithms. The presented techniques are illustrated with examples and comparisons are drawn between different community models. We summarize the content of each chapter in what follows.

- **Cohesive Subgraphs.** Chapter 2 presents classical concepts of cohesive subgraphs. In many real applications where information is modeled using graphs, communities are formed by a set of similar entities that are densely connected with certain relationships. Various kinds of cohesive subgraphs, including clique, quasi-clique, k-DBDSG, k-clan, k-club, k-plex, k-core, k-truss, k-vertex-connected, and k-edge-connected, which are widely used in the literature as building blocks for communities in graphs, are introduced.

- **Cohesive Community Search.** Chapter 3 introduces the problem of community search in simple graphs. In simple terms, a graph represents a structure of interactions within a group of vertices. Community models in this class can only leverage the structural characteristics of networks, essentially focusing on the density of the connection structure. Given a set of query vertices, community search is to find a densely connected subgraph containing all query vertices. Recently, several community models based on different dense subgraphs have been proposed, including quasi-clique [54], densest subgraph [180], k-core [18, 55, 122, 157], and k-truss [89, 96]. Our discussion in this chapter covers these various community models.

- **Attributed Community Search.** Chapter 4 discusses the problem of attributed community search in attributed networks, where nodes are associated with attributes or predicates. Many real-world networks contain attributes or predicates on vertices, e.g., a person may have information such as name, interests, and skills. In addition to the network structure, users may aim to search for attribute-related communities, or attributed communities. An attributed community is a group of vertices that are connected with a cohesive structure, and share homogeneous query attributes [65, 93]. The latter property bears some resemblance to keyword search over databases and graphs, but has important differences.

- **Social Circle Discovery.** Chapter 5 presents the problem of social circle discovery in social networks. Online social networks allow users to manually categorize their friends into social circles within their ego-networks (e.g., "circles" on Google+) [117, 170]. As one special kind of community, social circles are communities formed by friends only. The problem of social circles discovery is to automatically identify all social circles for a given user. Social circles can be used for content filtering, privacy protection, or sharing groups of users that others may wish to follow. The number of distinct social contexts also affects the process of information diffusion in social contagion [87, 163].

- **Geo-Social Group Search.** Chapter 6 describes the problem of finding geo-social groups in location-based social networks. In such networks, many users share their locations, which enables a new computing paradigm that explicitly combines the location and social factors to generate useful information for either business or social good. Geo-social group search looks for a group of users densely and closely connected in terms of both social and spatial proximity [123, 124, 199]. Relevant applications include recommending a group of friends for meet up and pushing mobile coupons to a group of close friends in location-based advertisements.

- **Datasets and Tools.** Chapter 7 lists a number of real-world datasets with ground-truth communities for further study in community search. Moreover, software systems implementing the representative community search algorithms presented in this book are also listed.

- **Conclusions.** Chapter 8 concludes this book. Whereas good progress has been made, research on community search is still in its infancy, and there are still many opportunities for further research. We briefly introduce the latest publications that are not covered in the early chapters of this book, discuss open problems, and highlight promising directions.

The content of this book is based on the tutorial entitled "Community Search over Big Graphs: Models, Algorithms, and Opportunities" [94] given by the authors at the 33rd IEEE International Conference on Data Engineering (ICDE) in April 2017. The tutorial slides are available online [95]. While the tutorial slides can serve as a companion to this book, the book includes more comprehensive and in-depth coverage of various community models, community search algorithms, as well as more recent developments in the area. Other supplementary materials of this book can be found at `http://db.comp.hkbu.edu.hk/csbook`.