Community Search over Big Graphs: Models, Algorithms, and Opportunities

Xin Huang*†, Laks V.S. Lakshmanan*, Jianliang Xu†
*University of British Columbia, Vancouver, Canada
†Hong Kong Baptist University, Hong Kong, China
{xin0, laks}@cs.ubc.ca, {xinhuang, xujl}@comp.hkbu.edu.hk

***Abstract*—Communities serve as basic structures for understanding the organization of many real-world networks, such as social, biological, collaboration, and communication networks. Recently, community search over large graphs has attracted significantly increasing attention, from simple and static graphs to evolving, attributed, location-based graphs. Different from the well-studied problem of community detection that finds all communities in an entire network, community search is to find the cohesive communities w.r.t. the query nodes.**

In this tutorial, we survey the state-of-the-art of community search on various kinds of networks across different application areas such as densely-connected community search, attributed community search, social circle discovery, and querying geo-social groups. We first highlight the challenges posed by the community search problems. We continue the presentation of their principles, methodologies, algorithms, and applications, and give a comprehensive comparison of the state-of-the-art techniques. This tutorial finally concludes by offering future directions for research in this important and growing area.

I. INTRODUCTION

Community structures naturally exist in numerous real-world networks, social, biological, collaboration, and communication networks being just a few examples. The task of community detection is to identify all communities in a network, which is a fundamental and well-studied problem in the literature. Recently, several works have studied a related but different problem called community search, which is to find cohesive communities containing a given set of query nodes. Since the communities defined by different nodes in a network may be quite different, community search with query nodes opens up the prospects of user-centered and personalized search [16]. As just one example, in a social network, the community formed by a person's high school classmates can be significantly different from that formed by her family members which in turn can be quite different from the one formed by her colleagues [20].

In this tutorial, we will first introduce the basic background and concepts of communities and networks, then give an overview of the state-of-the-art research. For each proposed method, we will give a brief explanation of the community models, the intuition behind them, and the algorithms. We will show examples illustrating recent research using the techniques, and make a comprehensive comparison of different community models. A brief overview of the scope of the tutorial is as follows.

Cohesive Community Search. In the simplest way, 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 nodes, community search is to find a densely-connected subgraph containing all query nodes. Recently, several community models based on different dense subgraphs have been proposed, including quasi-clique [14], densest subgraph [46], k-core [39], [15], [5], [31] and k-truss [23], [26].

Attributed Community Search. Many real social networks contain attributes or predicates on the vertices, e.g., a person may have information including name, interests, and skills, etc. 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 cohesive structure, which share homogeneous query attributes [18], [25]. The latter property bears some resemblance to keyword search over databases and graphs, but has important differences.

Social Circle Discovery. Online social networks allow users to manually categorize their friends into social circles within their ego network (e.g., circles on Google+) [35], [44]. As one special kind of communities, social circles are communities formed by only friends. The task of social circles discovery is to automatically identify all social circles for a given user. Social circles can be used for content filtering, for privacy, and for sharing groups of users that others may wish to follow. The number of distinct social contexts also affects the process of information diffusion on social contagion [42], [21].

Querying Geo-Social Groups. In location-based social networks, many users share their locations, which enables a new computing paradigm that explicitly combines both location and social factors to generate useful information for either business or social good. Geo-social group queries look for a group of users densely and closely connected in terms of both social and spatial proximity [32], [33], [50]. Relevant applications include recommending a group of friends nearby for gathering, and a restaurant pushing mobile coupons to a group of close friends in location-based advertisements.

Last but not least, we will offer open problems and future directions.

II. TARGET AUDIENCE

This tutorial targets anyone interested in modeling and querying communities over large graphs, from data mining and

data management researchers to practitioners from academia and industry. For those new to the domain, this tutorial will cover the necessary background material to help understand the topics and will offer a comprehensive survey of the state-of-the-art. In addition, the tutorial is aimed at giving a new perspective that will be interesting and valuable even for researchers with more experience in the field. For those having worked in 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 having worked in querying communities, we hope to inspire new research directions through connecting with recent developments in a new public-private model of graphs and distributed graph processing systems.

III. OUTLINE

Our tutorial includes 3 sections where the 2nd section consists of 4 subsections, and will take 1.5 hours to cover in all. We summarize the content of each section as follows.

A. Introduction, Motivations, and Challenges

In this part, we provide the background of community search. It consists of an introduction to the research field and highlights the popularity, applications, and challenges in community search. Specifically, we first introduce various kinds of networks and some of the most notable examples of communities. For a good understanding of this problem, vivid examples are illustrated to distinguish community search with related problems such as, community detection [49], [10], [38], [24], [9], keyword search [1], [20], [6], [17], [30], [37], and team formation [29], [28], [19]. A brief comparison of different problems is shown in Table I. Next, we elaborate on the application scenarios and challenges. The challenges include, but are not limited to, the complexity of underlying community models, responsiveness requirements of query processing, dynamic structures and massive collection of networks. As an example to illustrate the difficulty in designing community models, in a research collaboration network, the communities of a famous scholar and of a junior scholar can be dramatically different in terms of the community size and density [23].

B. Existing Research

We review different community models on various types of networks and query processing techniques.

Densely-Connected Community Search. This part gives an overview of the community search in simple graphs, which only have structural characteristics of networks. Community search on a simple graph aims to find densely connected communities containing all query nodes. We survey the-state-of-art community search models in simple graphs. They are based on different densely-connected subgraph definitions, including quasi-clique [14], densest subgraph [46], k-core [39], [15], [5], [31] and k-truss [23], [26]. An important issue here is the "free rider effect" [26], [46], whereby nodes far away from query nodes and irrelevant to them are included in the detected community. Several different index structures are designed for the efficient search of k-core and k-truss based communities

Method	Topic	Participation Condition	Attribute Function	Cohesiveness Constraint	Communication Cost
[6]	KS	χ	✓	χ	✓
[17]	KS	χ	✓	χ	✓
[30]	KS	χ	✓	χ	✓
[29]	TF	χ	✓	χ	✓
[19]	TF	χ	✓	✓	✓
[28]	TF	χ	✓	χ	✓
[39]	DCS	✓	χ	✓	✓
[14]	DCS	✓	χ	✓	χ
[15]	DCS	✓	χ	✓	χ
[5]	DCS	✓	χ	✓	χ
[26]	DCS	✓	χ	✓	✓
[31]	DCS	χ	χ	✓	χ
[46]	DCS	✓	χ	✓	✓
[18]	ACS	✓	✓	✓	χ
[25]	ACS	✓	✓	✓	✓

TABLE I. A COMPARISON OF REPRESENTATIVE WORKS ON KEYWORD SEARCH (KS), TEAM FORMATION (TF), DENSELY-CONNECTED COMMUNITY SEARCH (DCS) AND ATTRIBUTED COMMUNITY SEARCH (ACS).

[5], [31], [23], [26]. The tutorial discusses these models in a comparative fashion, and points out their pros and cons in the context of desiderata of good query communities. More specifically, we make a comparison of these models w.r.t. three aspects: (i) consideration of query nodes, (ii) cohesive structure, and (iii) quality of approximation.

Attributed Community Search. In this part, we focus on the attributed community search in attributed networks, where nodes are associated with attributes or predicates. Many real networks contain attributes in vertices, e.g., in social networks, a person has attributes including name, interests, and skills, etc. Given a set of query nodes and attributes, the attributed community search is to find the communities containing query nodes with a cohesive structure and sharing homogeneous query attributes. Recently, [18] and [25] have proposed two different models for attributed community search over attributed graphs. The key distinction between two works is that the model [18] is based on k-cores in structure with a strict attribute function. In contrast, the model [25] is based on k-truss in structure with a relaxed attribute function. We will discuss the pros and cons of these design choices.

Social Circle Discovery. This part discusses one special kind of communities in social networks, called social circles. In social networks, for a query user, social circles are communities formed only by her friends. The induced subgraph of an entire network only by her friends and herself is called ego network. [36] proposed an unsupervised community model to automatically detect circles in ego networks. The discovered circles are disjoint, overlapping, and hierarchically nested. Social circles can affect the process of information diffusion on social contagion [42]. A social circle represents a distinct social context of a user, and the multiplicity of social contexts is termed structural diversity [42]. Taking one social contagion process in Facebook as an example, a user is much more likely to join Facebook and become engaged if he or she has a larger structural diversity. [21], [22] studied the problem of finding k users with the highest structural diversity in graphs, which can be beneficial to political campaigns, promotion of health practices, marketing, and so on.

Querying Geo-Social Groups. This part covers the group queries in location-based social networks. With the rapid

development of location-aware mobile devices, many users share their locations in social networks [32], [3]. The problem of querying geo-social groups looks for a group of users densely and closely connected in terms of both social and spatial proximity. [48] selected a group of nearby attendees with a tight social relationship, which aims to minimize the spatial distance among selected users. [50] proposed a new family of k-core based geo-social group queries with minimum acquaintance constraint. [32], [33] studied a minimum user group query, in which each user has k neighbors and the users' joint regions cover all query points. Variants of R-tree index structure integrating the social information are designed for different geo-social query processing. A general framework that offers flexible data management and algorithmic design for geo-social network queries has been proposed in [3], [2].

C. Future Directions

While good progress has been made, research on community search is still in its infancy, and there are many opportunities for further research. In the following, we highlight some of the promising directions.

Querying Communities on Heterogeneous Information Networks. Most of the current research on community search focuses on homogeneous networks. In heterogeneous real-world networks where nodes and relations are of different types [40], the study of community search has not been investigated as yet. For example, in a healthcare network, nodes can be patients, doctors, medical tests, diseases, medicines, hospitals, treatments, and so on. On one hand, treating all the nodes as of the same type may miss important semantic information. On the other hand, treating every node as of a distinct type may miss the big picture. Such multiple types of objects, interconnected, forming complex, heterogeneous but often semi-structured information networks, bring rich opportunities and challenges for community modeling and discovery [41].

Scalability. Scaling community search techniques to the massive and rapidly growing network datasets of the Big Data era is another important direction. Current graph processing techniques include I/O efficient algorithms for k-core decomposition [45], [43] and k-truss decomposition [11], distributed graph computing (including Pregel [34]and Blogel [47]), and sketching [7], [13]. Techniques for handling community indexes in highly evolving graphs [23] and streaming graphs [4] are also avenues for future work.

Public-Private Social Networks. Most existing works assume that the entire network structure is visible and assume unrestricted access to it. However, in real applications, due to privacy issues, social networks can be more complex than a simple fully visible graph. Social network providers allow users to control their privacy by controlling the information they are willing to share. E.g., as reported in a recent study, 52.6% of 1.4 million New York City Facebook users hid their friends list [16]. Such privacy protection leads to a novel graph model, called *public-private* graphs [12]. It contains a public graph, in which each node is also associated with a private graph. The public graph is visible to everyone, and each private graph is visible only to the corresponding user. Despite the long existence of such networks, they have started gaining attention from the research community very recently. Owing to the scale of the network and its associated idiosyncrasies, community search over such networks cannot be efficiently answered by traditional algorithmic tools and techniques, which remain largely unexplored.

Community Search on Probabilistic Graphs. A large number of real-world networks are associated with uncertainty, due to the data collection process, machine-learning methods employed at preprocessing, inherent uncertainty in link inference in biological networks, or privacy-preserving reasons. The discovery of communities in uncertain graphs can be beneficial for a wide range of application domains including functional module identification for helping critical clinical diagnosis of diseases such as cancer in biology. Given the recent surge of interest in dense subgraphs such as k-core [8] and k-truss [27] in probabilistic graphs, an exciting question is how to generalize various community models [5], [23], [18], [26] and search techniques to probabilistic graphs. The challenge is to develop extensions that are widely useful and tractable.

ACKNOWLEDGMENT

This work is supported by a Discovery grant and a Discovery Accelerator Supplements grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), and HK-RGC Grants 12200114, 12201615, 12244916.

REFERENCES

[1] S. Agrawal, S. Chaudhuri, and G. Das. Dbxplorer: A system for keyword-based search over relational databases. In *ICDE*, pages 5–16, 2002.

[2] N. Armenatzoglou, R. Ahuja, and D. Papadias. Geo-social ranking: functions and query processing. *The VLDB Journal*, 24(6):783–799, 2015.

[3] N. Armenatzoglou, S. Papadopoulos, and D. Papadias. A general framework for geo-social query processing. *PVLDB*, 6(10):913–924, 2013.

[4] B. Bahmani, R. Kumar, and S. Vassilvitskii. Densest subgraph in streaming and mapreduce. *PVLDB*, 5(5):454–465, 2012.

[5] N. Barbieri, F. Bonchi, E. Galimberti, and F. Gullo. Efficient and effective community search. *DMKD*, 29(5):1406–1433, 2015.

[6] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using banks. In *ICDE*, pages 431–440, 2002.

[7] P. Boldi, M. Rosa, and S. Vigna. Hyperanf: approximating the neighbourhood function of very large graphs on a budget. In *WWW*, pages 625–634, 2011.

[8] F. Bonchi, F. Gullo, A. Kaltenbrunner, and Y. Volkovich. Core decomposition of uncertain graphs. In *KDD*, pages 1316–1325, 2014.

[9] C. Bothorel, J. D. Cruz, M. Magnani, and B. Micenkova. Clustering attributed graphs: models, measures and methods. *Network Science*, 3(03):408–444, 2015.

[10] H. Cheng, Y. Zhou, X. Huang, and J. X. Yu. Clustering large attributed information networks: an efficient incremental computing approach. *DMKD*, 25(3):450–477, 2012.

[11] J. Cheng, Y. Ke, S. Chu, and M. T. Özsu. Efficient core decomposition in massive networks. In *ICDE*, pages 51–62, 2011.

[12] F. Chierichetti, A. Epasto, R. Kumar, S. Lattanzi, and V. Mirrokni. Efficient algorithms for public-private social networks. In *KDD*, pages 139–148, 2015.

[13] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *JCSS*, 55(3):441–453, 1997.

[14] W. Cui, Y. Xiao, H. Wang, Y. Lu, and W. Wang. Online search of overlapping communities. In *SIGMOD*, pages 277–288, 2013.

[15] W. Cui, Y. Xiao, H. Wang, and W. Wang. Local search of communities in large graphs. In *SIGMOD*, pages 991–1002, 2014.

[16] R. Dey, Z. Jelveh, and K. Ross. Facebook users have become much more private: A large-scale study. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 346–352. IEEE, 2012.

[17] B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin. Finding top-k min-cost connected trees in databases. In *ICDE*, pages 836–845, 2007.

[18] Y. Fang, R. Cheng, S. Luo, and J. Hu. Effective community search for large attributed graphs. *PVLDB*, 9(12):1233–1244, 2016.

[19] A. Gajewar and A. D. Sarma. Multi-skill collaborative teams based on densest subgraphs. In *SDM*, pages 165–176, 2012.

[20] V. Hristidis and Y. Papakonstantinou. Discover: Keyword search in relational databases. In *PVLDB*, pages 670–681, 2002.

[21] X. Huang, H. Cheng, R.-H. Li, L. Qin, and J. X. Yu. Top-k structural diversity search in large networks. *PVLDB*, 6(13):1618–1629, 2013.

[22] X. Huang, H. Cheng, R.-H. Li, L. Qin, and J. X. Yu. Top-k structural diversity search in large networks. *The VLDB Journal*, 24(3):319–343, 2015.

[23] X. Huang, H. Cheng, L. Qin, W. Tian, and J. X. Yu. Querying k-truss community in large and dynamic graphs. In *SIGMOD*, pages 1311–1322, 2014.

[24] X. Huang, H. Cheng, and J. X. Yu. Dense community detection in multi-valued attributed networks. *Information Sciences*, 314:77–99, 2015.

[25] X. Huang and L. V. Lakshmanan. Attribute truss community search. *arXiv preprint arXiv:1609.00090*, 2016.

[26] X. Huang, L. V. Lakshmanan, J. X. Yu, and H. Cheng. Approximate closest community search in networks. *PVLDB*, 9(4):276–287, 2015.

[27] X. Huang, W. Lu, and L. V. S. Lakshmanan. Truss decomposition of probabilistic graphs: Semantics and algorithms. In *SIGMOD*, pages 77–90, 2016.

[28] M. Kargar and A. An. Discovering top-k teams of experts with/without a leader in social networks. In *CIKM*, pages 985–994, 2011.

[29] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *KDD*, pages 467–476, 2009.

[30] G. Li, B. C. Ooi, J. Feng, J. Wang, and L. Zhou. Ease: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *SIGMOD*, pages 903–914, 2008.

[31] R.-H. Li, L. Qin, J. X. Yu, and R. Mao. Influential community search in large networks. *PVLDB*, 8(5), 2015.

[32] Y. Li, R. Chen, J. Xu, Q. Huang, H. Hu, and B. Choi. Geo-social k-cover group queries for collaborative spatial computing. *IEEE Transactions on Knowledge and Data Engineering*, 27(10):2729–2742, 2015.

[33] Y. Li, R. Chen, J. Xu, Q. Huang, H. Hu, and B. Choi. Geo-social k-cover group queries for collaborative spatial computing. In *ICDE*, pages 1510–1511, 2016.

[34] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: a system for large-scale graph processing. In *SIGMOD*, pages 135–146, 2010.

[35] J. J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *NIPS*, volume 272, pages 548–556, 2012.

[36] J. J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *NIPS*, pages 548–556, 2012.

[37] L. Qin, J. X. Yu, L. Chang, and Y. Tao. Querying communities in relational databases. In *ICDE*, pages 724–735, 2009.

[38] Y. Ruan, D. Fuhry, and S. Parthasarathy. Efficient community detection in large networks using content and links. In *WWW*, pages 1089–1098, 2013.

[39] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *KDD*, pages 939–948, 2010.

[40] Y. Sun and J. Han. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, 14(2):20–28, 2013.

[41] Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao. Community evolution detection in dynamic heterogeneous information networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 137–146, 2010.

[42] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. *PNAS*, (16):5962–5966.

[43] J. Wang and J. Cheng. Truss decomposition in massive networks. *PVLDB*, 5(9):812–823, 2012.

[44] Y. Wang and L. Gao. An edge-based clustering algorithm to detect social circles in ego networks. *Journal of computers*, 8(10):2575–2582, 2013.

[45] D. Wen, L. Qin, Y. Zhang, X. Lin, and J. X. Yu. I/O efficient core graph decomposition at web scale. In *ICDE*, pages 133–144, 2016.

[46] Y. Wu, R. Jin, J. Li, and X. Zhang. Robust local community detection: On free rider effect and its elimination. *PVLDB*, 8(7), 2015.

[47] D. Yan, J. Cheng, Y. Lu, and W. Ng. Blogel: A block-centric framework for distributed computation on real-world graphs. *PVLDB*, 7(14):1981–1992, 2014.

[48] D.-N. Yang, C.-Y. Shen, W.-C. Lee, and M.-S. Chen. On socio-spatial group query for location-based social networks. In *KDD*, pages 949–957, 2012.

[49] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. *PVLDB*, 2(1):718–729, 2009.

[50] Q. Zhu, H. Hu, J. Xu, and W.-C. Lee. Geo-social group queries with minimum acquaintance constraint. *arXiv preprint arXiv:1406.7367*, 2014.

Biographies

Xin Huang

Xin Huang is a Research Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. He has worked as a postdoctoral fellow at University of British Columbia in 2015-2016. He received the Ph.D. degree from the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. His research interests include data management, data mining, and social network analysis. He has recently been actively working on the topics of community discovery and graph query processing. He serves on program committees for conferences including VLDB, ICDE, WWW, SDM, and etc.

Laks V.S. Lakshmanan

Laks V.S. Lakshmanan is a Professor of Computer Science at University of British Columbia, Canada. His research covers a wide spectrum of topics including data management and mining, advanced data models for novel applications, OLAP and data warehousing, data integration, data cleaning, semi-structured data and XML, information and social networks and social media, recommender systems, and personalization. He is currently serving as an Associated Editor of the VLDB Journal (VLDBJ).

Jianliang Xu

Jianliang Xu is a Professor in the Department of Computer Science, Hong Kong Baptist University. He received his BEng degree in computer science and engineering from Zhejiang University, Hangzhou, China, in 1998 and his PhD degree in computer science from Hong Kong University of Science and Technology in 2002. He held visiting positions at Pennsylvania State University and Fudan University. His research interests include data management, mobile/pervasive computing, and networked and distributed systems. He has published more than 150 technical papers in these areas. He is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering (TKDE) and Proceedings of the VLDB Endowment (PVLDB 2018).