The Social Technology and Research (STAR) Lab in the University of Hong Kong

Reynold Cheng, Chenhao Ma, Xiaodong Li, Yixiang Fang*, Ye Liu†, Victor Y.L. Wong, Esther Lee, Tai Hing Lam, Sai Yin Ho, Man Ping Wang, Weijie Gong, Wentao Ning, and Ben Kao

The University of Hong Kong

{ckcheng, chma2, xdli}@cs.hku.hk, fangyixiang@cuhk.edu.cn, yliu03@scut.edu.cn, {vylwong, estherst, hrmrlth, syho, mpwang, gweijie}@hku.hk, {wtning, kao}@cs.hku.hk

1. INTRODUCTION

The main goal of the Social Technology and Research Laboratory (STAR Lab) in the University of Hong Kong (`https://star.hku.hk`) is to develop novel IT technologies for serving the society. Our team has more than three years of experience in project development, web, app, and game design, photography, and video production. We are interested in "Data Science for Social Good", researching data-driven approaches that can benefit the public, NGOs, and the government.

As of Fall 2021, the STAR lab is comprised of four professors, five postdoc researchers, ten PhD students, and more than twenty software developers. The lab are working on different aspects towards "Data Science for Social Good":

1. **research** – interdisciplinary research on data and social science;
2. **technology** – developing technologies for social services;
3. **synergy** – connecting with the public, NGOs, and the government;
4. **teaching** – educating NGOs and the public on social technologies; and
5. **impact** – bringing benefit to the public, NGOs, and the government.

To achieve the above goals, the STAR lab has been working on two frontiers: fundamental research (Section 2) and social and legal applications (Section 3). For fundamental research, we tackle the challenge of the huge volume and complexity of graph data, and develop efficient, scalable, and efficient algorithms on different kinds of graphs. We also develop novel graph-based recommender systems. We collaborate with universities including the University of Illinois at Urbana-Champaign and the University of British Columbia on these efforts. As for social and legal applications, we have been collaborating with more than 20 organizations in universities, governments, NGOs, and commercial organizations. The lab has acquired more than HKD $ 35M funding. We have recently received a SIGMOD Research Highlights Award, two industry awards, and one university knowledge-exchange award. A PhD graduate of the lab has been selected by Baidu Scholar as one of the 2021 Global Top 100 Chinese Rising Stars in Artificial Intelligence.

2. FUNDAMENTAL RESEARCH

Graph data are prevalent in different social applications. For example, as we will discuss in Section 3.1, one of our projects, called HINCare, utilize big graph technologies to facilitate the matching of volunteers to elders. Hence, the STAR lab has been recently engaged in various of fundamental research activities in graph mining, with the goal of advancing "data science in social good". In Section 2.1, we will discuss the main work that we have done in the past few years, namely densest subgraph discovery, motif analysis, and community search. We then describe the research problems we address regarding the heterogeneous information networks (HINs), in Section 2.2. These work can be useful for social applications, as we will describe in Section 3.

2.1 Big graph technologies

1. Densest subgraph discovery is fundamental to a wide range of applications, such as fraud

*Yixiang Fang joined the Chinese University of Hong Kong-Shenzhen in 2021.

†Ye Liu joined the South China University of Technology in 2021.

detection, community mining, and graph compression. We have examined this problem on undirected and directed graphs, as described below.

• **Undirected densest subgraph (UDS):** Given an undirected graph G, UDS aims to find a subgraph D of G with the highest density (e.g., the number of edges over the number of vertices in D). Because UDS is difficult to solve, we propose a new solution paradigm [10]. Our main observation is that the densest subgraph can be accurately found through a k-core (a kind of dense subgraph of G), with theoretical guarantees. Based on this intuition, we develop efficient exact and approximation solutions for UDS. Moreover, our solutions support a wide range of graph density definitions, including clique-based and general pattern-based density. Extensive experimental evaluation shows that our algorithms are up to four orders of magnitude faster than existing approaches.

• **Directed densest subgraph (DDS):** Given a directed graph $\hat{\mathrm{G}}$, DDS refers to the finding of a subgraph from $\hat{\mathrm{G}}$, whose "directed density" is the highest among all the subgraphs of $\hat{\mathrm{G}}$. [1] Essentially, we aim to find two sets of vertices, S^* and T^*, from G, where (1) vertices in S^* have a large proportion of outgoing edges to those in T^*, and (2) vertices in T^* receive a large proportion of edges from those in S^*. Existing DDS solutions suffer from efficiency and scalability problems. Hence, we develop an efficient and scalable DDS solution [19]. We introduce the notion of $[x, y]$-core, which is a dense subgraph for $\hat{\mathrm{G}}$, and show that the densest subgraph can be accurately located through the $[x, y]$-core with theoretical guarantees. Extensive experiments show that our proposed solutions are up to six orders of magnitude faster than the state-of-the-art. This year, we have a follow-up DDS work based on convex programming [18].

Our work in DDS [20] received the SIGMOD Research Highlight Award 2021, and its journal extension has been recently accepted by TODS as one of the Best of SIGMOD 2020 papers [21]. Recently, we have been collaborating with the HK Applied Science and Technology Research Institute on using the densest subgraphs found on their user-website-click graphs to find fraudulent clicks. We also plan to extract densest subgraphs from the volunteer-elderly graph (details in Section 3.1), and examine how these subgraphs can be used to provide recommendations of volunteers to elders.

[1]The directed density of a subgraph induced by two subsets S and T is computed by the number of edge connecting from S to T divided by the square root of the product of the sizes of S and T.

2. Motif-based graph analysis. This kind of analysis has recently emerged as an important tool for discovering insight from graphs, e.g., biological and social networks. A motif, or a small graph with a few nodes, is a fundamental building block of large and complex networks [17, 14, 15, 16]. Motif-based graph analysis enables "higher-order semantics" analysis, and performs better than traditional "edge-based" solutions in a range of graph analytics tasks, such as link prediction [16], graph clustering [15], and node ranking [14]. These tasks are often important for extracting insights and patterns from the graphs collected in our lab, as well as predicting user behaviors. We studied two problems about motifs, and developed a system prototype, as detailed below.

• **Counting motifs on uncertain graphs.** Given a graph G and a graph pattern m, e.g., a "triangle", a fundamental task in motif analysis is to count the number of instances (or frequency) of m on G; if m occurs frequently, then m can be considered as a motif of G. Motif counting enables the understanding of the characteristics of the graph, and also the discovery of the right motifs for motif-based graph analysis and visualization. We recently examine how to count motifs on uncertain graphs, whose edges exist probabilistically [17]. Although researchers have developed several fast motif counting solutions, they assume that graphs are deterministic, i.e., the graph edges are certain to exist. However, this assumption may not always hold. For example, in a social network, a link between two nodes (representing two users), which represent the friendship between these users, may only exist probabilistically. Ignoring this issues can lead to a wrong counting result, and affect motif-based graph analysis. We propose a solution framework, called LINC, which can be used by existing deterministic motif counting algorithms. Extensive experiments on real datasets show that LINC is more effective and efficient than existing motif-counting solutions for uncertain graphs.

• **Motif-paths.** We propose a new notion of motifs, known as *motif-path*, which is conceptually a concatenation of one or more motif instances between two given nodes on G. We use motif-paths to develop algorithms for three graph mining tasks, namely link prediction, local graph clustering and node ranking [14]. These tasks are important to the analysis of the graphs collected in our lab, and enables the performance of various tasks (e.g., graph integration and cleaning). Our experiments show that motif-paths are more effective than traditional motif-based analysis and "path-based solutions", i.e.,

those that use shortest path distance as a dissimilarity metric. We also develop a motif-path-based drug analysis algorithm based on a COVID-19 knowledge graph [16]. The algorithm can be used to trace the origins of COVID-19 variant strains.

• **System prototype.** Existing graph database systems are not designed to support queries that involve motifs. We develop M-Cypher, which is a system prototype designed to enable expression and execution of motif-related queries through a user-friendly graph query language [15].

3. Community search. A fundamental component of big graphs is the network community. Essentially, a community is a group of vertices which are densely connected. Community retrieval can be used in many real applications, such as event organization, friend recommendation, and network analysis. How to effectively and efficiently find high-quality communities from big graphs is an important research topic in the era of big data. A large group of research works, called community search (CS), have been proposed, which aim to provide efficient solutions for searching high-quality communities from large networks in real time. Nevertheless, earlier CS solutions mainly focused on simple undirected graphs, so they could not be applied to perform CS on more complicated graphs such as attributed graphs, directed graphs, and Heterogeneous Information Networks (HINs).

To overcome the above limitations, we extensively study the problems of CS over these graphs. Specifically, for attributed graphs, we consider several kinds of vertex attributes such as keywords, spatial locations, and profile information, and for each of them, we formulate novel models of communities by considering both link relationship and vertex attributes, propose efficient CS solutions, and experimentally evaluate them on real large attributed graphs [5, 4, 2]. For directed graphs, we formulate a novel community model by carefully considering the vertices' in-degrees and out-degrees, develop efficient both online and index-based CS algorithms, and evaluate the proposed solutions on billion-scale directed graphs [8, 3]. For HINs, we propose three community models and develop efficient algorithms to perform CS [9]. To the best of our knowledge, our works of CS on keyword-based attributed graphs [5], directed graphs [8], and HINs [9] are the first works in these problems.

We also develop a system, called C-Explorer [6], to assist users in extracting, visualizing, and analyzing communities. C-Explorer implements several state-of-the-art CS and community detection solutions, and various functions for analyzing the effectiveness of the communities of these solutions. Recently, we conduct a survey of existing CS works, compare the quality of communities with different cohesive subgraph models, and point out promising research directions [7].

2.2 Heterogeneous information networks

1. Meta path discovery. A heterogeneous information network (HIN) is a graph whose nodes and edges are tagged with "type labels" to express their meanings [13, 28, 25, 22]. Given two HIN nodes s and t, and a set S of *example node pairs* (e.g., pairs of nodes representing celebrity star couples), in [22] we developed a machine-learning model for discovering *meta paths* [25], which is essentially a sequence of node types and edge types that characterize the important relationships between node pairs in S. The meta paths found can be used to support graph-based applications such as friend search, product recommendation, anomaly detection, and graph clustering. More recently, we developed efficient algorithms for discovering the k most important meta paths in real-time, based on the occurrence frequency and rarity of meta-paths [28].

2. Meta structure relevance. An important problem in HIN is the computation of closeness, or relevance, between two HIN objects. We propose to use meta structure, which is a *directed acyclic graph* of object types connected by edge types, to measure the proximity between objects in [13]. The strength of meta structure is that it can describe complex relationships between two HIN objects (e.g., two papers in DBLP share the same authors and topics). We develop three relevance measures based on meta structure, and an efficient algorithm proposed to support the relevance evaluation.

3. Web query recommendation. A web query is a string of keywords posted by users for finding information in the Internet. Typically, web search engines provide alternative query formulations, which can be more articulate and interesting to users. A *long-tail query* is an uncommon request that rarely occurs in query logs. Traditional approaches, which rely solely on query logs, could perform poorly on long-tail queries because they rarely occur in query logs. However, it is relatively easy to extract HIN entities from long-tail queries. We have studied how to utilize HIN entity relationship information effectively to provide a recommendation solution for long-tail queries. In [11, 12], we study the use of meta-paths, a form of HIN-entity relationships, for query recommendation. Next, we will examine the use of HIN-based algorithms in social applications.

3. SOCIAL APPLICATIONS

We now discuss three social-related projects being done in the STAR lab: (1) data-driven recommendation for elderly care; (2) big data analytics for social services; and (3) legal data analysis.

3.1 Data-driven elderly care

Many metropolitan cities are facing sharp increase in aging population. In Hong Kong, for instance, the number of elderly citizens is estimated to rise to one third of the population, or 2.37 million, in year 2037. About 13% and 24% of these people are living alone or with their spouses only respectively. As they age and become more frail, the demand for formal support services will increase exponentially in the coming years. However, there is a severe lack of manpower to meet these needs: in HK, on average, each NGO employee needs to manage 10 elderly people at the same time. Some elderly-care homes also reported a 70% shortage in staff. There is thus a strong need of helpers for taking care of elderly people on a full-time, part-time, or voluntary basis.

We have been working on HINCare [2], a HKD $4M project supported by HK Innovation and Technology Commission. The HINCare is a volunteer management system with timebanking facilities (i.e., each person, after providing a service, can earn a time credit. The time credits are stored in the person's time bank account. He/she can later use the earned time credits to purchase other services.) We have designed an elderly-user-friendly mobile app. The system backend, designed for NGO administrators, is cloud-based and *generic*. Essentially, any organization can use our system to support their voluntary work services easily. Instead of having sophisticated software installation, only a few customization steps are needed. The platform supports multiple organizations, which can enable more sharing of data and collaboration. Currently, HINCare has been serving 5000 elders in 6 NGOs. We won one local (HKICT) and one international information technology awards (Asia Smart Apps) and a HKU Faculty Knowledge Exchange Award. Recently, the HK government's Community Investment and Inclusion Fund (CIIF) has provided 4 contract research projects to our team to further support their associated NGOs.

HIN-based matching. The core of HINCare employs novel heterogeneous information network (HIN) and AI technologies to recommend helpers to elders. Here, the HIN stores the relationship information among elders, helpers, and NGOs. It originates from various Big Data sources, such as social networks and senior citizen's profiles. We use the HIN to find out the best helpers for assisting elders. For example, a living-alone elder may want someone to repair a light bulb; the HIN reveals that a certain helper living close to the elder has the expertise and availability to do so, and the system will recommend the helper to the elder. In detail, we use the HIN built to develop a recommender system [1, 27, 24]. These meta-path-based solutions leverage semantic relationship among graph nodes. They provide rich information of interaction among users and items, and help to comprehend a user's interest.

We remark that this is the first time that HIN is used to support elderly care. Experiment results tested on two collected NGO datasets show good performance compared with other systems.

3.2 Big data analytics for social services

Applying big data and artificial intelligence in behavioral and social science are promising but limited currently [23]. Family is the core in shaping individual behavior and influencing social capital, particularly in Chinese societies. Family services provide an important source to understand the influences of family on individual and society by using the large volume of data regularly collected in territory-wide family services, including the information on service users, groups and programmes, and counselling case recordings. These data are multi-dimensional, containing text, numeric, audio, or video forms.

The Hong Kong Jockey Club SMART Family-Link (JCSFL) Project[3], initiated and funded by the Hong Kong Jockey Club Charities Trust in 2018, is a large scale (HKD $80M), 4-year cross-sectoral collaboration among (1) the School of Public Health, and (2) Technology-Enriched Learning Initiative of HKU, and (3) STAR lab, with 26 Integrated Family Service Centers and Integrated Service Centers for advance the use of Information and Communications Technology and big data analytics for enhancing family services in Hong Kong. Being the first in Chinese population, we are collaborating with family services providers to analyze aggregated anonymous data of a large number of IFSCs/ISCs. The findings will be useful for informing family services and policy in the future.

i-Connect. A key element of JCSFL is i-Connect – a software service management platform developed by the STAR lab to support the service operations of the NGO-operated IFSCs. The system enables the users, who are social workers and clerical staff in IFSCs, to perform their operational pro-

[2] https://www.hincare.hku.hk/blog

[3] https://jcsmartfamilylink.hk/en/

cesses and workflow on a centralised platform. Different kinds of data about clients, staff, services, workflow, and operations are processed and stored through this system. The system has been deployed on a secure cloud platform. Industrial standards and various security measures are also adopted to ensure data security and data privacy. The system was reviewed by different professional parties including Office of the Government Chief Information Officer of Hong Kong SAR, the Information Security Team of Hong Kong Jockey Club, and a third-party security audit company.

3.3 Legal data analysis in collaboration with LawTech Centre

We collaborate with HKU Law Faculty in joint projects that study the problems of machine-assisted extraction and modeling of legal knowledge from legal texts, leveraging domain knowledge provided by law experts. Our knowledge models have led to the development of a number of essential legal applications that facilitate legal studies and research. For example, we developed a prediction model for illegal drug trafficking sentencing, which can be accessed online by the public (`http://wwwnew2.hklii.hk/predictor`). The sentencing predictor is used by some NGOs in youth education and crime prevention programs. Moreover, our sentencing prediction model helps address a number of interesting issues in legal information processing, which include judgment recommendation, fairness and explainability in machine predictions [26]. We also develop an AI-driven conversational system that helps people who have not received any legal training to effectively locate relevant legal information on our *Community Legal Information Centre* (CLIC) website. CLIC is an online information source covering 32 legal topics with contents such as FAQs, reading guides and explanatory notes on illustrative court cases, and short videos with hypothetical illustrative stories. The CLIC project is part of our continuing efforts in promoting free public legal education.

4. EDUCATION AND ENGAGEMENT

The STAR lab has dedicated a lot of effort in education. We have been producing video clips and conducting seminars, in order to train social workers to use our systems. We have organized press conferences and international symposiums for sharing our knowledge and experience with the public. Some materials related to the projects have also been used in courses taught by the lab leader, to educate students about how data science can be used in the social domain. We plan to recruit students to assist NGOs in our projects, in order to enrich their social awareness and practical experience.

5. BUILDING THE LAB

Finally, we share our experience of building the capabilities of our lab. We have built a core team of professors, who have worked with each other, and have expertise in data science, gerontology, and social science. We assembled a competent software development team, through fundings provided by government and charity organizations. It is very important for the NGOs to participate actively and provide their data for our projects. This requires a huge effort in understanding their needs and establishing trust. We have also organized public seminars, conducted interviews in newspapers and radio stations, and participated in exhibitions.

Acknowledgement

This research is supported by the University of Hong Kong (Projects 104005858, 104005994, KE award 2100048), HKU Innovation Wing Two Research Fund, the Innovation and Technology Commission of Hong Kong (ITF project MRP/029/18), the Hong Kong Jockey Club Charities Trust (HKJC project no. 2018-0025-001), and HKU-TCL Joint Research Center for Artificial Intelligence (Project no. 200009430). This research is supported by the WYNG Foundation; the Innovation and Technology Commission fund ITS/234/20; and The University of Hong Kong Knowledge Exchange Fund KE-ID-2018/19-21. Yixiang Fang was supported by NSFC under Grant 62102341 and CUHK-SZ grant UDF01002139. The Hong Kong Jockey Club Smart Family-Link Project was funded by the Hong Kong Jockey Club Charities Trust.

6. REFERENCES

[1] Y. Cao, X. Wang, X. He, Z. Hu, and T.-S. Chua. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *WWW*, pages 151–161, 2019.

[2] Y. Chen, Y. Fang, R. Cheng, Y. Li, X. Chen, and J. Zhang. Exploring communities in large profiled graphs. *TKDE*, 2018.

[3] Y. Chen, J. Zhang, Y. Fang, X. Cao, and I. King. Efficient community search over large directed graph: An augmented index-based approach. In *IJCAI*, pages 3544–3550, 2020.

[4] Y. Fang, R. Cheng, X. Li, S. Luo, and J. Hu. Effective community search over large spatial graphs. *PVLDB*, 10(6):709–720, 2017.

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

[6] Y. Fang, R. Cheng, S. Luo, J. Hu, and K. Huang. C-explorer: browsing communities in large graphs. *PVLDB*, 10(12):1885–1888, 2017.

[7] Y. Fang, X. Huang, L. Qin, Y. Zhang, W. Zhang, R. Cheng, and X. Lin. A survey of community search over big graphs. *VLDBJ*, 29(1):353–392, 2020.

[8] Y. Fang, Z. Wang, R. Cheng, H. Wang, and J. Hu. Effective and efficient community search over large directed graphs. *TKDE*, 31(11):2093–2107, 2019.

[9] Y. Fang, Y. Yang, W. Zhang, X. Lin, and X. Cao. Effective and efficient community search over large heterogeneous information networks. *PVLDB*, 13(6):854–857, 2020.

[10] Y. Fang, K. Yu, R. Cheng, L. V. S. Lakshmanan, and X. Lin. Efficient algorithms for densest subgraph discovery. *PVLDB*, 12(11):1719 – 1732, jul 2019.

[11] Z. Huang, B. Cautis, R. Cheng, and Y. Zheng. KB-enabled query recommendation for long-tail queries. In *CIKM*, 2016.

[12] Z. Huang, B. Cautis, R. Cheng, Y. Zheng, N. Mamoulis, and J. Yan. Entity-based query recommendation for long-tail queries. *TKDD*, 12(6):64:1–64:24, 2018.

[13] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta structure: Computing relevance in large heterogeneous information networks. In *SIGKDD*, pages 1595–1604, 2016.

[14] X. Li, R. Cheng, K. Chang, C. Shan, C. Ma, and H. Cao. On analyzing graphs with motif-paths. volume 14, pages 1111–1123. VLDB Endowment, 2021.

[15] X. Li, R. Cheng, M. Najafi, K. Chang, X. Han, and H. Cao. M-cypher: A gql framework supporting motifs. In *CIKM*, pages 3433–3436, 2020.

[16] X. Li, V. K. Yan, X. Ye, R. Cheng, and et al. Drug repurposing for the treatment of covid-19: A knowledge graph approach. *Advanced Therapeutics*, page 2100055, 2021.

[17] C. Ma, R. Cheng, L. V. Lakshmanan, T. Grubenmann, Y. Fang, and X. Li. Linc: a motif counting algorithm for uncertain graphs. *PVLDB*, 13(2):155–168, 2019.

[18] C. Ma, Y. Fang, R. Cheng, L. V. Lakshmanan, and X. Han. A convex-programming approach for efficient directed densest subgraph discovery. In *SIGMOD*, 2022.

[19] C. Ma, Y. Fang, R. Cheng, L. V. Lakshmanan, W. Zhang, and X. Lin. Efficient algorithms for densest subgraph discovery on large directed graphs. In *SIGMOD*, pages 1051–1066, 2020.

[20] C. Ma, Y. Fang, R. Cheng, L. V. Lakshmanan, W. Zhang, and X. Lin. Efficient directed densest subgraph discovery. *ACM SIGMOD Record*, 50(1):33–40, 2021.

[21] C. Ma, Y. Fang, R. Cheng, L. V. Lakshmanan, W. Zhang, and X. Lin. On directed densest subgraph discovery. *ACM Transactions on Database Systems (TODS)*, 46(4):1–45, 2021.

[22] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang. Discovering meta-paths in large heterogeneous information networks. In *WWW*, pages 754–764, 2015.

[23] M. Robila and S. A. Robila. Applications of artificial intelligence methodologies to behavioral and social sciences. *Journal of Child and Family Studies*, 29(10):2954–2966, 2020.

[24] C. Shi, B. Hu, W. X. Zhao, and P. S. Yu. Heterogeneous information network embedding for recommendation. *TKDE*, 2018.

[25] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

[26] T. Wu, B. Kao, A. S. Y. Cheung, M. M. K. Cheung, C. Wang, Y. Chen, G. Yuan, and R. Cheng. Integrating domain knowledge in ai-assisted criminal sentencing of drug trafficking cases. In *Legal Knowledge and Information Systems - JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334, pages 174–183. IOS Press, 2020.

[27] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. Recommendation in heterogeneous information networks with implicit user feedback. In *RecSys*, pages 347–350, 2013.

[28] Z. Zhu, T. N. Chan, R. Cheng, L. Do, Z. Huang, and H. Zhang. Effective and efficient discovery of top-k meta paths in heterogeneous information networks. *TKDE*, 2020.