Research Statement

Yuchen Li
School of Computing and Information System (SCIS), Singapore Management University (SMU)
Email: yuchenli@smu.edu.sg; Tel: (65) 68289614

1. Background and Overview

Large graphs are increasingly important in the digital realm, representing complex connections between entities. Graph data science, recognized by Gartner as a critical emerging technology[1], plays a vital role in organizational decision-making. By 2025, it is expected that 80% of data and analytics innovations will utilize graph technology, compared to just 10% in 2021. This surge reflects graph technology's superior ability to connect multiple data points for deeper analysis, addressing the complexities of modern data and enabling more efficient and insightful decision-making across various sectors, including social marketing, supply chain management, and healthcare.

At SMU, my research primarily focuses on a critical aspect of graph analytics: ***information dissemination on graphs (IDG).*** A key application of IDG is evident in viral marketing within online social networks, where users are depicted as nodes and their connections as edges. In this context, information can rapidly spread through the network among connected users. This phenomenon is utilized in viral marketing to effectively disseminate information about products, political campaigns, and various events across social network platforms. Moreover, the principles and methods of IDG have been extended to a wide array of applications, including but not limited to epidemic analysis, fraud detection, and personalized recommendation systems. These diverse applications underscore the versatility and importance of IDG in understanding and leveraging the dynamics of information flow within complex networks.

Under the broad area of IDG analysis, I categorize my studies into three distinct layers: *application-layer, theoretical-layer, and system-layer*. These layers are further detailed as follows:

- **Application layer:** This layer focuses on delivering effective IDG analysis across various application contexts. In numerous scenarios, we must consider not just the graph topology but also intricate features like topical, temporal, and spatial information. For instance, a user's likelihood to share information is often influenced by how well the content aligns with their interests. Therefore, developing context-aware IDG analysis tools is essential for supporting innovative applications and ensuring relevance and effectiveness.
- **Theory layer:** Here, the goal is to design algorithms for IDG analysis that are not only effective but also theoretically efficient. IDG analysis often involves tackling complex problems where exact solutions can be computationally expensive. To address this, I work on developing efficient approximate IDG solutions. These solutions offer quality assurances close to exact solutions while significantly reducing runtime, often by several orders of magnitude. This balance between theoretical rigor and practical efficiency is key to advancing IDG analysis.
- **System layer:** To scale IDG analysis for larger graphs, this layer involves the development of parallel systems to enable scalable IDG analysis. A crucial part of this process is the identification and implementation of high-level primitives for graph processing. This approach allows end-users to execute their custom IDG analyses on our developed systems without delving into the complexities of low-level performance optimization. This layer is instrumental in making IDG analysis accessible and scalable for larger datasets and more complex graph structures.

[1] https://www.gartner.com/en/articles/what-s-new-in-the-2023-gartner-hype-cycle-for-emerging-technologies

In summary, my goal is to develop an end-to-end framework for IDG analysis, spanning application, theory, and system prototyping. The objective is to provide actionable and fast intelligence for users by delivering effective and real-time IDG solutions on large-scale graphs with billions of nodes and edges. My aims and work are in line with the SCIS core research area of "**Data Management & Mining**," and they align with SMU's focus on addressing societal challenges related to "**Advancing Innovation & Technology**."

2. Research Impact and Recognition

The international recognition of my research is evidenced by my publications in prestigious journals and conferences such as SIGMOD, VLDB, ICDE, KDD, WWW, SIGIR, TKDE, and VLDBJ, focusing on data management and data mining. Since 2015, my academic portfolio includes over 50 publications in peer-reviewed journals and conference proceedings. My research in outdoor information dissemination has been honored with several awards, including the Best Research Paper Award at KDD 2019, a nomination for Best Paper at KDD 2018, and an Audience Appreciation Award Nomination also at KDD 2018. In addition to my academic contributions, I have held significant roles in the academic community, serving as the Area Chair of ICDE 2024, Contest Chair of ICDM 2019, and a member of the program committees for over ten top-tier conferences. My commitment to high-quality scholarly review was recognized with Outstanding Reviewer awards at VLDB 2021 and PAKDD 2020.

Beyond publishing, my research has led to practical systems with substantial real-world impact. I have collaborated with Alibaba to develop an online multi-modal community query engine for their billion-scale heterogeneous network. This system aids business owners in understanding community characteristics and behaviors and was awarded runner-up for Best Demo System at CIKM 2021. Our partnership also produced SHOAL, a large-scale taxonomy system implemented on Taobao, supporting millions of daily searches on one of the world's largest e-commerce platforms. Another notable project, River [ICDM 2018 RIVER], is a real-time monitoring system for dynamic social streams, offering effective geospatial and temporal visualizations for influential topic exploration. Additionally, our prototype for a compressed graph format on GPUs [SIGMOD 2019] received recognition from the SIGMOD reproducibility committee, where only 9 papers are recognized among 88 accepted papers.

3. Research Roadmap

Figure 1 illustrates the main elements of my research contributions, which have developed over time. During my PhD at NUS, I concentrated on "context-aware information dissemination," focusing on applying the Information Dissemination Graph (IDG) framework in various scenarios. After moving to SMU in 2018, I shifted focus to "graph analysis & algorithm design," the theoretical foundation of my work, aiming to devise robust algorithms for the IDG framework. More recently, I've been working on making graph analysis more scalable and user-friendly through "parallel graph processing and system prototyping," which corresponds to the system layer of my research. Below is a detailed list of my research areas and activities.

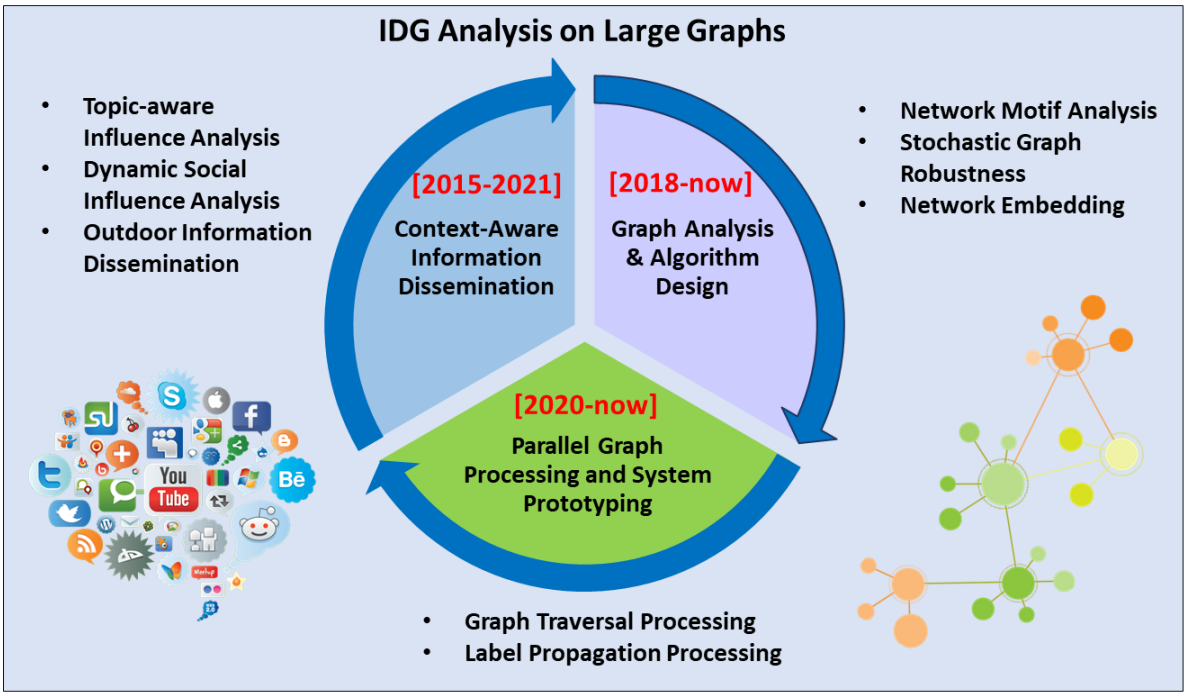

Figure 1. Research Roadmap

2.1 Context-aware Information Dissemination [my PhD study and beyond, 2015-2021]

Existing studies on IDG analysis predominantly target homogeneous graphs, where the focus is mainly on the topological structure. However, information dissemination is also governed by various types of contextual information, such as topical relevance, temporal patterns, and spatial relationships. My research extends beyond traditional models by integrating these contextual factors, offering novel approaches that enhance the scope and depth of IDG analysis.

Topical-aware influence analysis: my early research (my Ph.D. thesis) dealt with influence analysis in the context of topical information. In a topic-aware propagation network, each user in the network is associated with a profile of topics that he/she is interested in. Furthermore, the likelihood of a user influences another is also based on the topic information of the information under propagation. Under the topic-aware propagation network, I developed novel methods to find: (1) a set of seed users that can generate the maximum influence on a targeted group of users who are interested in a viral product matches to their interests [VLDB 2015, ICDE 2019]; (2) a set of influential topics/information that a given

user is shown to have the most influential power [SIGMOD 2017]; The developed methods provide effective and scalable tools for targeted advertising on large online social media.

Dynamic social influence analysis: during my tenure as a research fellow in NUS, my research focuses on studying influence analysis under the dynamic network setting. Social network data is under frequent updates where new users and posts are continuously generated at a rapid rate. It is crucial to understand how the trend of influence changes when the underlying network is subject to update. My research works employ the sliding window model to capture the up-to-date network for influence analysis. Under the sliding window model, we only keep track of the most recent interactions and use these interactions to detect influential users [VLDB 2017] and topics [ICDM 2018 RIVER]. We further integrate location information into dynamic influence analysis [TOIS 2018]. The methods that we developed not only provide theoretical guarantees on the quality of the solutions but also are efficient enough to process millions of updates per second occurred in online social media.

Outdoor information dissemination: Since joining SMU, I explore opportunities to leverage information dissemination in outdoor advertising, integrating geospatial data. This industry, already a multi-billion-dollar market, has seen substantial growth, reaching approximately $34.9 billion as of 2023 and is expected to climb to $59.24 billion by 2030. A significant driver of this growth, specifically in the outdoor billboard segment, is the strategic placement of ads along consumer travel routes—capturing the attention of drivers, cyclists, and pedestrians alike. Utilizing trajectory data and billboard spatial information, my work models the influence of billboards on passersby. We've developed methods for advertisers to select prime billboard locations for maximal ad dissemination impact, influencing the largest number of trajectories [KDD 2018, TKDD 2020, SIGMOD 2021 MROAM]. Further, we've innovated techniques to optimize for repeat impressions, which are crucial for converting user views into actions [VLDB 2019, KDD 2019].

2.2 Graph Analysis & Algorithm Design [After joining SMU, 2018-now]

My second line of research addresses the complex theoretical issues in understanding graph structures for IDG analysis. Given that precise analysis of these structures often involves computationally expensive algorithms, my research is geared towards the development of efficient approximate algorithms. These are meticulously crafted to closely mirror the results of exact solutions but with a significantly reduced computational load.

Network Motif Analysis: Network motifs are statistically frequent subgraph patterns appeared in a large graph. Recent works have discovered that network motifs can reveal fundamental understandings of IDG. For example, triangular motifs are crucial to finding ground-truth communities in social networks. Within a community, a piece of information spreads more rapidly compared with the spread across communities. Thus, existing works develop motif-aware graph partition to discover communities in a large graph. However, motif-aware graph partition requires an expensive two-step process: (1) construct a motif-adjacency matrix; (2) employ graph partitioning methods on the motif-adjacency matrix. Constructing the motif adjacency matrix is overwhelming expensive to compute as we need to enumerate all motif instances on a graph, which is exponential to the graph size. To enable real-time motif-aware graph partitioning, we propose sampling-based approaches to estimate the motif-adjacency matrix efficiently and effectively, where the approximated motif-adjacency matrix can still produce theoretically

guaranteed motif-aware partitions [ICDE 2021]. Furthermore, we also study motifs on temporal networks where the edges are associated with timestamp to indicate when the relationship occurs. Temporal networks can capture a more general class of applications for IDG, such as communication networks, computer networks, and financial transactions, etc. Enumerating temporal motifs poses even harder computing problems as there are more variations of temporal motifs compared with non-temporal motifs. We propose a hybrid approach where we combine a sampling method with exact enumeration to deliver a fast and accurate approximated solution [CIKM 2020].

Stochastic graph robustness: we pose a fundamental problem of IDG analysis: how can we access a network's ability to spread information under attack [WWW 2020]? Network robustness has been studied extensively in the case of deterministic networks. However, IDG applications raise a new question of robustness in probabilistic networks. We propose three novel robustness measures for networks hosting information diffusion, susceptible to node attacks. The outcome of such a process depends on the selection of its initiators, or seeds, by the seeder, as well as on two factors outside the seeder's discretion: the attack strategy and the probabilistic diffusion outcome. We consider three levels of seeder awareness regarding these two uncontrolled factors and evaluate the network's viability aggregated over all possible scenarios of node attacks. We introduce novel algorithms from building blocks found in previous works to evaluate the proposed measures. The measures shed light on the differences among networks in terms of robustness and the surprise they furnish when attacked [TKDE 2021].

Network embedding: My research taps into the burgeoning interest in using network embedding techniques for IDG analysis. A significant focus of my work is on integrating multi-modal data into network embeddings. To this end, we've developed a method for heterogeneous graph embedding, which effectively captures diverse node and edge types [ICDM 2018 HIN]. Additionally, we have introduced an innovative embedding method that incorporates propagation structures among nodes. This method enhances the accuracy of predicting influence outcomes between friends on large-scale social networks [referenced in [AAAI 2020]. For time-sensitive interactions and information dissemination, we propose novel network embedding architectures that leverage both structural and temporal information. These architectures are adept at predicting not only the sequence but also the precise timing of future interactions [KDD 2021 and TKDE 2022 GNPP]. Beyond practical applications, our research also includes a thorough investigation into the theoretical capabilities of different network embedding architectures, contributing to the foundational understanding of these approaches [TKDE 2022 WL].

2.3 Parallel graph processing and system prototyping [More recently, 2020-now]

In my third research area, I harness the power of parallel computing platforms, including multi-core CPUs and GPUs, to enhance graph processing techniques vital to IDG analysis. Our efforts are particularly geared towards improving graph traversal and label propagation processes. These specific aspects have gained increased attention in our recent research, recognizing their critical role in efficiently analyzing complex graph structures and disseminating information across networks.

Graph Traversal Processing: Graph traversal is a fundamental primitive in graph algorithms, known for its complexity due to frequent random memory accesses, which challenge cache efficiency. To address this, we developed ForkGraph, a cache-efficient graph processing system for multi-core architectures [SIGMOD 2021 FORK]. ForkGraph enhances cache efficiency by partitioning the graph into sections that match the

capacity of the LLC (last-level cache), and processes tasks in a buffered manner on a partition basis. We also implemented sophisticated intra- and inter-partition execution strategies for improved efficiency. Our evaluations on real-world graphs demonstrate that ForkGraph significantly surpasses existing graph processing systems like Ligra, Gemini, and GraphIt, achieving speedups by two orders of magnitude. For GPU graph traversal processing, we have introduced novel methods, including traversal on compressed graphs and self-adaptive graph traversal [SIGMOD 2021 SELF], aimed at optimizing cache performance. More recently, our research has extended to exploring constrained shortest path processing on both CPU and GPU architectures [VLDB 2020, ICDM 2023]. This processing differs from traditional graph traversal as it generates substantially larger intermediate results. To manage this, we developed efficient parallel pruning strategies and schedulers to balance memory overhead with processing efficiency. Our methods have achieved significant speedups over current state-of-the-art approaches, demonstrating their efficacy in handling complex graph traversal tasks.

Label Propagation Processing: Graph label propagation (LP) is a core routine in IDG analysis, widely used in applications such as fraud detection and recommendation systems. However, LP-based analysis typically incurs high processing overheads. For instance, in our analysis of the fraud detection process at TAOBAO, we found that the LP component alone accounted for 75% of the total processing overhead. To address this, we introduced a GPU-centric processing system, specifically designed to accelerate LP tasks [SIGMOD 2021 GLP]. Recognizing the diverse variants of LP used by data engineers to combat evolving fraud patterns, we acknowledged the challenges in implementing efficient and accurate GPU programs for graph-based applications like LP. To simplify this process, we developed a suite of expressive APIs that allow data engineers to easily customize and deploy LP algorithms on GPUs. Moreover, to further optimize performance, we tackled the issue of costly data transfers between CPUs and GPUs. By enabling LP processing on compressed graphs, we significantly reduced this expense [TKDE 2023]. Our solution not only supports real-world billion-scale graph workloads for fraud detection but also achieves a remarkable 13.2x speedup compared to existing in-house solutions on high-end multicore machines.

4. Future Directions

In my future research, I aim to advance graph analytics with an emphasis on enhancing system performance and user accessibility. A significant challenge in applying graph analytics at scale is the complexity of graph concepts and the difficulty of scaling analysis for large datasets. My goal is to create user-friendly systems that significantly improve efficiency. This effort will involve simplifying user interaction with the system, developing more effective system layers, and employing visualization techniques to make the results more understandable.

Furthermore, I plan to conduct in-depth research into various graph neural network (GNN) architectures. GNNs represent a versatile AI model that extends beyond traditional vision and language models. The current challenge lies in comprehending the behaviors of different GNN architectures and choosing the appropriate model for specific datasets. To address this, I will focus on simplifying the use of GNNs by creating an integrated system that covers data and model selection, as well as training and testing processes. This system will streamline the deployment of GNNs, making it easier for researchers and practitioners to utilize these powerful tools in their work.

5. Reference:

- **[VLDB 2015]:** Li, Y., Zhang, D., & Tan, K. L. (2015). Real-time Targeted Influence Maximization for Online Advertisements. PVLDB, 8(10), 1070-1081.
- **[SIGMOD 2017]:** Li, Y., Fan, J., Zhang, D., & Tan, K. L. (2017). Discovering your selling points: Personalized social influential tags exploration. In SIGMOD (pp. 619-634).
- **[VLDB 2017]:** Wang, Y., Fan, Q., Li, Y., & Tan, K. L. (2017). Real-time influence maximization on dynamic social streams. PVLDB, 10(7), 805-816.
- **[TOIS 2018]:** Wang, Y., Li, Y., Fan, J., & Tan, K. L. (2018). Location-aware influence maximization over dynamic social streams. TOIS, 36(4), 1-35.
- **[KDD 2018]:** Zhang, P., Bao, Z., Li, Y., Li, G., Zhang, Y., & Peng, Z. (2018). Trajectory-driven influential billboard placement. In SIGKDD (pp. 2748-2757).
- **[ICDM 2018 RIVER]:** Sha, M., Li, Y., Wang, Y., Guo, W., & Tan, K. L. (2018). River: A real-time influence monitoring system on social media streams. In ICDM (pp. 1429-1434).
- **[ICDM 2018 HIN]:** Zheng, V. W., Sha, M., Li, Y., Yang, H., Fang, Y., Zhang, Z., ... & Chang, K. C. C. (2018). Heterogeneous embedding propagation for large-scale e-commerce user alignment. In ICDM (pp. 1434-1439).
- **[KDD 2019]:** Zhang, Y., Li, Y., Bao, Z., Mo, S., & Zhang, P. (2019). Optimizing impression counts for outdoor advertising. In SIGKDD (pp. 1205-1215).
- **[VLDB 2019]:** Zhang, Y., Bao, Z., Mo, S., Li, Y., & Zhou, Y. (2019). ITAA: an intelligent trajectory-driven outdoor advertising deployment assistant. PVLDB, 12(12), 1790-1793.
- **[ICDE 2019]:** Li, Y., Fan, J., Ovchinnikov, G., & Karras, P. (2019). Maximizing multifaceted network influence. In ICDE (pp. 446-457).
- **[SIGMOD 2019]:** Sha, M., Li, Y., & Tan, K. L. (2019). Gpu-based graph traversal on compressed graphs. In SIGMOD (pp. 775-792).
- **[TKDD 2020]:** Zhang, P., Bao, Z., Li, Y., Li, G., Zhang, Y., & Peng, Z. (2020). Towards an Optimal Outdoor Advertising Placement: When a Budget Constraint Meets Moving Trajectories. TKDD, 14(5), 1-32.
- **[CIKM 2020]:** Wang, J., Wang, Y., Jiang, W., Li, Y., & Tan, K. L. (2020). Efficient Sampling Algorithms for Approximate Temporal Motif Counting. In CIKM (pp. 1505-1514).
- **[WWW 2020]:** Logins, A., Li, Y., & Karras, P. (2020, April). On the Robustness of Cascade Diffusion under Node Attacks. In WWW (pp. 2711-2717).
- **[AAAI 2020]:** Wang, H., Meng, Q., Fan, J., Li, Y., Cui, L., Zhao, X., ... & Du, X. (2020). Social Influence Does Matter: User Action Prediction for In-Feed Advertising. In AAAI (Vol. 34, No. 01, pp. 246-253).
- **[VLDB 2020]:** Lu, S., He, B., Li, Y., & Fu, H. (2020). Accelerating exact constrained shortest paths on GPUs. PVLDB, 14(4), 547-559.
- **[ICDE 2021]:** Huang, S., Li, Y., Bao, Z., & Li, Z. (2021). Towards efficient motif-based graph partitioning: An adaptive sampling approach. In ICDE (pp. 528-539).
- **[SIGMOD 2021 MROAM]:** Zhang, Y., Li, Y., Bao, Z., Zheng, B., & Jagadish, H. V. (2021). Minimizing the regret of an influence provider. In SIGMOD (pp. 2115-2127).
- **[SIGMOD 2021 FORK]:** Lu, S., Sun, S., Paul, J., Li, Y., & He, B. (2021). Cache-efficient fork-processing patterns on large graphs. In SIGMOD (pp. 1208-1221).
- **[SIGMOD 2021 GLP]:** Ye, C., Li, Y., He, B., Li, Z., & Sun, J. (2021). Gpu-accelerated graph label propagation for real-time fraud detection. In SIGMOD (pp. 2348-2356).
- **[SIGMOD 2021 SELF]:** Sha, M., Li, Y., & Tan, K. L. (2021, June). Self-adaptive graph traversal on gpus. In SIGMOD (pp. 1558-1570).
- **[TKDE 2021]:** Logins, A., Li, Y., & Karras, P. (2021). On the robustness of diffusion in a network under node attacks. TKDE, 34(12), 5884-5895.
- **[KDD 2021]:** Xia, W., Li, Y., Tian, J., & Li, S. (2021). Forecasting interaction order on temporal graphs. In SIGKDD (pp. 1884-1893).

- **[TKDE 2022 GNPP]:** Xia, W., Li, Y., & Li, S. (2022). Graph neural point process for temporal interaction prediction. TKDE, 35(5), 4867-4879.
- **[TKDE 2022 WL]:** Xia, W., Li, Y., & Li, S. (2022). On the Substructure Countability of Graph Neural Networks. TKDE.
- **[ICDM 2023]:** Xia, W., Li, Y., Guo, W., & Li, S. (2022). Efficient Navigation for Constrained Shortest Path with Adaptive Expansion Control. In ICDM (pp. 588-597).
- **[TKDE 2023]:** Ye, C., Li, Y., He, B., Li, Z., & Sun, J. (2023). Large-Scale Graph Label Propagation on GPUs. TKDE.