Research Statement

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

1. Background and Overview

Large graphs are pervasive as people and things are digitally connected in a complex way. The information embedded in graphs brings immense opportunities to discover valuable insights that continuously power the development of data-driven economy. According to the Gartner 2020 report, graph analytics is becoming a key technology and supports decision making in 30% of organizations globally.

At SMU, my primary research focus is on a fundamental area in graph analytics: ***information dissemination on graphs (IDG).*** One novel application of IDG is viral marking in online social networks where users are represented as nodes in the network and relationships between users are presented as edges. A piece of information could quickly become pervasive though the "word-of-mouth" propagation among social network friends. Viral marking takes advantage of IDG to conduct effective campaign by spreading information on commercial products, elections and events on social network cites. Furthermore, the analysis of IDG has been applied to numerous application domains such as epidemic analysis, fraud detection, recommendation and many others.

Under the broad area of IDG analysis, my research studies can be categories by three layers: *application-layer, theoretical-layer and system-layer*, which are elaborated as follows:

- **Application layer: provide effective IDG analysis in different application contexts.**
  In many scenarios, in additional to graph topology, there are complex features to consider for a comprehensive IDG analysis such as topical, temporal and spatial information. For example, a user is more likely to propagate a piece of information if its topic matches the user's interests. Thus, it is crucial to develop context-aware IDG analysis for supporting novel applications.
- **Theory layer: design theoretically efficient algorithms for IDG analysis.**
  IDG analysis requires solving theoretically challenging problem. In many cases, an exact solution often incurs prohibitive execution costs. Motivated by this challenge, I propose efficient approximate IDG solutions which provide theoretically proven quality guarantees to the exact solution while often reduce the run time by orders-of-magnitudes.
- **System layer: develop parallel systems to enable scalable IDG analysis.**
  To further scale IDG analysis to large graphs, we develop several parallel systems/tools for general graph processing tasks. We also identify a set of high-level primitives for graph processing and end-users can easily implement their customized IDG analysis on our develop system without worrying about the low-level design of performance optimizations.

In summary, I aim to provide an end-to-end framework for IDG analysis from application, theory to system prototyping. The goal of the framework is to enable actionable and fast intelligence for users by providing effective and real-time IDG solutions on large-scale graphs with billions of nodes and edges. My aim and work fall under SMU's research area of "**Data Science & Engineering**", and they align with SMU's focus on addressing societal challenges related to "**Advancing Innovation & Technology**".

2. Research Map

Figure 1 provides an overview of my research roadmap and my key research works. I make continuous efforts according to this roadmap. In particular, "**context-aware information dissemination**" corresponds to the application layer of my work, "**graph analysis & algorithm design**" correspond to the theory layer and "**parallel processing and system prototyping**" corresponds to the system layer. The followings are a more detailed enumeration of my research areas and activities.

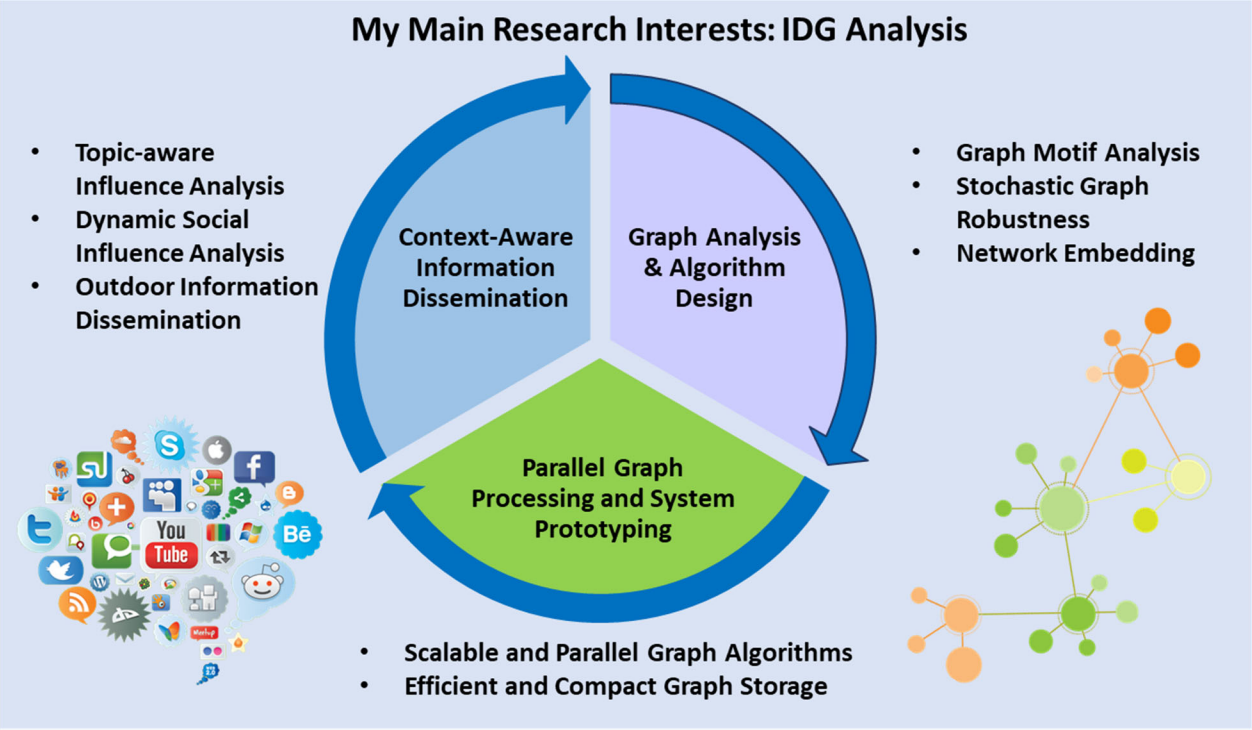

Figure 1. Research Map

2.1 Context-aware Information Dissemination

The existing works on IDG analysis mostly focus on homogeneous graphs where only the topological structure is considered. However, there are many different types of contextual information on graphs that governs information disseminations, which includes but not limited to: topical, temporal, and spatial information. My research works provide novel methods by incorporating the above information and enable contextual IDG analysis.

Topical-aware influence analysis: my early research (my PhD 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 topic 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: my research works as a postdoc focus on studying influence analysis under the dynamic network setting. Social network data is under frequent updates where new users and posts are continuously generated in 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 occurred and user these interactions to detect influential users [VLDB 2017 DIM] 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: my most recent works focus on applying information dissemination for outdoor advertising with the help of geospatial data. Outdoor advertising has been a market of multi-billion dollars and is expected to reach 33 billion dollars by 2021. Moreover, 74% of its growth comes for the outdoor billboard segment. The main audiences of billboards are people moving along their trips, by vehicles, motorcycles, bikes, etc. Enabled by the positioning devices, tremendous amounts of trajectories have been generated and records. We take advantage of the trajectory data and the spatial information of the billboards to model the influence impact from a billboard to a user traveling pass the billboard. Subsequently, we develop a novel method for billboard advertisers to select the best billboards that influence the largest number of trajectories for disseminating their ads [KDD 2018, TKDD 2020]. We further propose an advance method to optimize for user impression counts: users tend to take meaningful actions only when they see the same ads multiple times [KDD 2019, VLDB 2019 ITAA]. Our research works have received multiple awards including best paper candidate [KDD 2018], audience appreciate award nomination [KDD 2018] and best research paper award [KDD 2019].

2.2 Graph analysis & algorithm design

In addition to the application layer of IDG, my second research area explores the theoretical aspect on how graph structure affects the spread of information. We propose novel problems for the fundamental understanding of IDG. Furthermore, the analysis of IDG often requires expensive algorithms to execute if exact solutions are required. Hence, we develop approximate algorithms which have theoretical guarantees to the exact solution but run dramatically faster to enable real-time IDG analysis.

Graph 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 efficiently and effectively estimate the motif-adjacency matrix, where the approximated motif-adjacency matrix can still produce theoretically guaranteed motif-aware partitions [ICDE 2021]. Furthermore, we also study motifs on temporal network

where the edges are associated with time stamp 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 pose 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 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 network. We propose three novel robustness measures for networks hosting an 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 extends of node attacks. We introduce novel algorithms from building blocks found in previous works to evaluate the proposed measures. The measures shed light the differences among networks in terms of robustness and the surprise they furnish when attacked.

Network embedding: there is a growing interest in employing network embedding methods for IDG analysis. For example, network embeddings can be used to predict if a user can share a piece of information with another user on social networks. My research studies how to incorporate multi-modal data for network embeddings. We develop a heterogeneous graph embedding method to capture different types of node and edge information [ICDM 2018 HIN]. Furthermore, we propose an optimized embedding method for bipartite graphs [SIGIR 2020], which are used to model a range of real-world scenarios. For example, user activities on an e-commerce website can be formalized as a bipartite graph: nodes are categorized into users and items, while edges among nodes represent the interactive behaviors between users and items. Training embeddings is expensive, we thus propose a dynamic positive sampling approach to enable scalable learning on large graphs. We also develop a novel embedding method that incorporate propagation structures among nodes to improve the prediction on the influence outcome between two friends on large scale social networks [AAAI 2020].

2.3 Parallel graph processing and system prototyping

My third research area focuses on leveraging parallel computing platforms such as multi-core central processing units (CPUs) and graphics processing units (GPUs) to accelerate graph processing primitives, which can be used for more complex graph analysis such as IDG. We develop efficient parallel algorithms which take advantage of different hardware characteristics. To support large graph processing with different application requirements, we also propose novel graph representation formats. We implement prototype systems and make it easy for non-expert users.

Scalable and parallel graph algorithms:

We develop multiple parallel graph algorithms on emerging hardware such as CPUs and GPUs, which can achieve orders of magnitude speedup compared with the traditional single thread approach. We devise a dynamic personalized PageRank method that can be efficiently parallelized on CPUs and GPUs [VLDB 2017

PPR]. In [SIGMOD 2020], we propose a series of optimizations for subgraph enumeration on heterogeneous CPU-GPU platform. In order to support large graphs with billions of nodes and edges on a single machine, we divide the graph into partitions. Then, the workload of enumerating subgraphs within a partition is executed in GPUs whereas CPUs process the workloads which involve accessing data from different partitions. We further optimize the GPU algorithm by devising a novel reuse strategy that avoids processing repetitive tasks [TKDE 2020]. Most recently, we build a prototype system on GPUs to support constrained shortest path queries on large graphs. The system allows users to easily deploy their applications by providing application-specific constraints without exposing to the low-level details of GPU implementation and performance optimizations.

Efficient and compact graph storage: there have been a rich study of large graph storage format that optimizes both space overhead and efficiency for executing graph processing tasks. In my research, we focus on efficient storage format on GPUs. There are two major challenges: (1) the graph format must be general enough to host different graph algorithms; (2) the graph format should consider the characteristics of GPU platform. In [VLDB 2017 GPMA], we develop GPMA for dynamic graph format on GPUs. GPMA supports rapid dynamic graph updates on GPUs with millions of updates per second. Meanwhile, GPMA is general enough to support any graph processing tasks. Subsequently, we focus on optimizing space overhead of graph storage on GPUs. GPUs have significant less memory than its CPUs counterpart. In order to support large graph processing on GPUs, we propose a compressed graph storage (GCGT) [SIGMOD 2019]. GCGT combines existing techniques to effectively compress a large graph to be loaded on GPUs. To efficiently process graph processing tasks on GCGT, we propose novel on-the-fly decompression techniques which leverage the unique characteristic of GPU hardware. Extensive experiments reveal that GCGT shows prominent compression rate while support a wide range of graph applications. The efficiency of GCGT is also competitive with those running on non-compressed graphs.

3. Research Impact and Recognition

The impact of my research works is endorsed by international peers, as evidenced by his publications in top venues such as SIGMOD, VLDB, ICDE, KDD, WWW, TKDE, VLDBJ for data management and data mining, SIGIR, WSDM, CIKM, TOIS for information retrieval. Since 2015, I have published 30+ articles in referred journals and conferences. My recent works on outdoor information dissemination receives the best research paper award (KDD 2019), the best paper candidate (KDD 2018) and the audience appreciation award nomination (KDD 2018). I have also served as the contest chair of ICDM 2019 and PC members for over 10 top conferences. I was recognized as outstanding reviewer by PAKDD 2020.

In addition to publications, my research works are transformed to deployed and prototyped systems with significant practical impact. SHOAL [VLDB 2019 SHOAL] is a deployed large-scale taxonomy system to serve millions of searches per day on Taobao, one of the largest global e-commerce platforms. An intelligence outdoor advertising assistant called ITAA [VLDB 2019 ITAA], which finds the most influential advertisement placement based on large-scale trajectory data, at the intersection of non-equal costs of billboards, budget constraint and orthogonality of different influence modes. River [ICDM 2018 RIVER] is a real-time monitoring system on dynamic social streams that provides effective geospatial and temporal visualizations for influential topic exploration. Our prototype of compressed graph format on GPUs [SIGMOD 2019] was also recognized by the ACM SIGMOD reproducibility committee, where only 9 papers are recognized among 88 accepted papers.

4. Selected Publications:

- **[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 DIM]:** Wang, Y., Fan, Q., Li, Y., & Tan, K. L. (2017). Real-time influence maximization on dynamic social streams. In *PVLDB*, *10*(7), 805-816.
- **[VLDB 2017 PPR]:** Guo, W., Li, Y., Sha, M., & Tan, K. L. (2017). Parallel personalized pagerank on dynamic graphs. *PVLDB*, *11*(1), 93-106.
- **[VLDB 2017 GPMA]:** Sha, M., Li, Y., He, B., & Tan, K. L. (2017). Accelerating Dynamic Graph Analytics on GPUs. *PVLDB*, *11*(1), 107-120.
- **[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, July). Optimizing impression counts for outdoor advertising. In *SIGKDD* (pp. 1205-1215).
- **[VLDB 2019 ITAA]:** 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.
- **[VLDB 2019 SHOAL]:** Li, Z., Chen, X., Pan, X., Zou, P., Li, Y., & Yu, G. (2019). Shoal: Large-scale hierarchical taxonomy via graph-based query coalition in e-commerce. *PVLDB*, *12*(12), 1858-1861.
- **[ICDE 2019]:** Li, Y., Fan, J., Ovchinnikov, G., & Karras, P. (2019, April). Maximizing multifaceted network influence. In *ICDE* (pp. 446-457).
- **[SIGMOD 2019]:** Sha, M., Li, Y., & Tan, K. L. (2019, June). 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).
- **[SIGIR 2020]:** Huang, W., Li, Y., Fang, Y., Fan, J., & Yang, H. (2020). BiANE: Bipartite Attributed Network Embedding. In *SIGIR* (pp. 149-158).
- **[SIGMOD 2020]:** Guo, W., Li, Y., Sha, M., He, B., Xiao, X., & Tan, K. L. (2020). GPU-Accelerated Subgraph Enumeration on Partitioned Graphs. In *SIGMOD* (pp. 1067-1082).
- **[TKDE 2020]:** Guo, W., Li, Y., & Tan, K. L. (2020). Exploiting Reuse for GPU Subgraph Enumeration. *IEEE TKDE*.
- **[ICDE 2021]:** Huang, S., Li, Y., Bao, Z., & Li, Z. (2021). Towards Efficient Motif-based Graph Partitioning: An Adaptive Sampling Approach. In *ICDE* (to appear).
- **[VLDB 2021]:** Lu, S., He, B., Li, Y., & Fu, H. (2021). Accelerating Exact Constrained Shortest Paths on GPUs. In *PVLDB* (to appear)