DANTE: Data and Information Management Research at Nanyang Technological University

Sourav S Bhowmick, Xiaokui Xiao*

School of Computer Science and Engineering
Nanyang Technological University, Singapore

1. INTRODUCTION

The data management group at the Nanyang Technological University (DANTE) was founded in 2009 by the first author when the School of Computer Science and Engineering hired two young faculty members in this area. The group currently consists of three faculty members and more than twenty graduate students, research assistants, and postdocs. Our alumni include faculty members at the Chinese University of Hong Kong, National University of Singapore, University of New South Wales, and several researchers and engineers at Facebook, Google, eBay, Huawei, and other technology companies. The group's major funding is from the Ministry of Education in Singapore, National Research Foundation, and companies such as Huawei.

In DANTE, we subscribe to the policy of conducting research mostly in small groups. Typically, one or two faculty members work together with their students, staffs, and collaborators (if any). Our members bring in diverse strengths, some have penchant for inventing efficient and scalable solutions to existing data management problems whereas others are more inclined toward inventing novel problems and efficient solutions to address them. Our research is often multi-disciplinary in flavour, bridging data management and analytics with social science and biology. In particular, our research have been nominated as one of the best papers in venues such as SIGMOD 2015, ICDE 2015, and ICDE 2010. We have also received best paper award in ACM BCB 2011. The common thread running through our research is a focus on going beyond papers to build usable novel prototypes. Specifically, we have successfully demonstrated more than 20 novel research prototypes in top-tier data and information management conferences, which is the highest among all data management research groups in Australasia. In this article, we present a brief overview of the key research themes in our group; more details are available on our website at `http://www3.ntu.edu.sg/scse/dmal/dante/`.

*The authors are the founding members of DANTE. The second author is a faculty member of National University of Singapore since January 2018.

2. GRAPH DATA MANAGEMENT

Graphs are used to model data in a variety of domains such as biology, chemistry, and social science. Even we can model the relational data as a graph. This prevalence of graph-structured data has led us to pursue several research directions in this arena as follows.

2.1 Graph Query Processing

Querying graphs has emerged as an important research problem since the last decade. Our group has invented a suite of efficient and scalable techniques to support a variety of graph queries such as distance queries [15], subgraph enumeration [31], supergraph search [16], personalized PageRank queries [24, 51, 52], SimRank queries [38, 48], and reachability queries [14, 70]. In particular, our solutions for personalized PageRank and SimRank queries provide superior practical efficiency while providing strong theoretical guarantees in terms of accuracy and time complexity. In a different project, we questioned the longstanding assumption that a subgraph search query must be a *connected* graph. Such assumption typically demands users to have precise knowledge of the topological structure of what they are seeking. Unfortunately, due to the topological complexity of the underlying graph data, it is often unrealistic to assume that an end user is aware of the precise relationships between nodes in a query graph. This led us to invent a novel subgraph query processing framework called PANDA [59] that can efficiently support formulation and processing of *partial topology queries*. Such queries are disconnected query graphs comprising of two or more disjoint *connected query components*. This framework can also be used to address the problem of keyword search for graphs as a keyword can be considered as a single-node query component.

2.2 Human Interaction with Graphs

In our HINT project[1], we explore pioneering techniques and paradigms for visually interacting with graphs using queries. It is well-known that visual query interfaces (*i.e.,* GUI) enable interactive construction of graph queries without demanding knowledge of a graph query language from end users. In a classical visual graph querying framework, the visual query interface is "loosely-coupled" with the underlying query engine. Typically, a visual query interface is designed and implemented by leveraging principles from the human-computer interaction (HCI) area to enhance its usability. On the other hand, the query engine is realized using data management principles to ensure efficient and scalable execution of graph queries. Seldom there is any meaningful interactions between these two components *concurrently*. Consequently, when an end user is visually formulating a graph query, the underlying query engine remains idle as human interactions at the GUI level are rarely communicated to the query engine. The query engine is only invoked when the complete query has been formulated and the `Run` icon is clicked to execute it. We refer to this loose coupling of these two key components as *shallow integration*.

The visual graph query formulation process demonstrates two key characteristics. First, a query is gradually exposed to the underlying query engine during its construction. Second, it gives rise to GUI *latency* (*i.e.,* the time spent by a human to complete certain query formulation task such as construction of an edge). In this research, we crystallize "tight coupling" between visual graph query interface and query engine components by exploiting these features. Instead of the query engine being oblivious to human interactions in the GUI during visual query formulation, we "track" these interactions and process them judiciously during query formulation by exploiting the GUI latency. We refer to this tight coupling of the visual query interface and the query engine as *deep integration*.

In our group, we have explored a suite of novel techniques to realize deep integration. Specifically, in [63] we report a technique that leverages on partially constructed query information during query formulation to present opportune suggestions to end users toward completion of the query. These efforts realize deep integration between the visual query interface and underlying query engine by generating data-driven suggestions while a graph query is being visually formulated. In [26, 28, 29, 44], we realize deep integration by blending visual graph query formulation with its processing to prune false results and prefetch partial query results by exploiting the GUI latency, leading to superior system response time. We investigate a variety of graph queries (subgraph matching, subgraph similarity, and homomorphic queries) in this paradigm. In particular, these frameworks allow a user to execute a query fragment any time during query formulation and not wait until the entire query is visually formulated. Consequently, this paradigm is exploited by PICASSO [25] to realize exploratory search on graphs. Lastly, query performance study in a deep integration-based graph querying framework demands exhaustive user study due to tight coupling between human interactions and the underlying query engine. However, such user study is expensive and time-consuming. In [3], we present a framework called VISUAL that draws upon the literature in HCI and graph data management to simulate visual subgraph query construction process. This paves the way for automated performance study without requiring users.

[1] `http://www.ntu.edu.sg/home/assourav/research/hint/index.html`

2.3 Graph Analytics

Lastly, we have contributed efficient and scalable algorithms for addressing a variety of graph analytics problems. For instance, we have invented scalable algorithms for finding maximal cliques [11,12], core and truss decomposition [13, 49], triangle listing [17], computation of maximum independent set [39], attributed graph clustering [60], and discovering frequent subgraph patterns using the MapReduce framework [36].

In particular, some of our research in graph analytics is multi-disciplinary in nature. For instance, we have explored the role of network analytics in biology. It is increasingly attractive to model biological systems from a broader, "systems" perspective instead of modeling its components in an isolated, reductionist manner [27]. The most well-known method to model biological systems in this manner is through *biological networks*. However, due to the complexity of such networks, it is challenging to uncover key system-wide properties and behaviors of a biological system from it. To this end, we have developed scalable techniques for discovering network motifs (*i.e.,* interaction patterns that recur throughout biological networks, much more often than in random networks) by leveraging modern hardware such as GPUs [37]. This work was nominated as one of the best in ICDE 2015. We have also developed novel techniques for *summarizing* static and dynamic biological networks [1] as well as predicting potential drug targets by analyzing dynamics of signaling networks [18]. In particular, our work on functional summarization of protein-protein interaction networks received the Best Paper Award in ACM BCB 2011 (flagship conference of the SIGBio group) [43].

3. SOCIAL DATA MANAGEMENT

In the social data management arena, DANTE members have primarily focused on two topics: social influence analysis in online social networks and social image exploration.

3.1 Social Influence Analysis

Our group has made significant contributions in social influence analysis especially in the context of the influence maximization problem. Given a social network G and a constant k, the influence maximization problem asks for k nodes in G that (directly and indirectly) influence the largest number of nodes under a pre-defined diffusion model. This problem originates from viral marketing, and it has been extensively studied in the literature since 2003. However, before 2014, there was a long-standing tension between the efficiency and accuracy of influence maximization algorithms. In particular, there exist a few methods that provide non-trivial approximation guarantees, but they require days to process even a small graph; meanwhile, methods with reasonable practical efficiency all rely on heuristics, due to which they fail to offer any worst-case accuracy assurance. We are the first to ease the tension by proposing Two-phase Influence Maximization (TIM) [46], an algorithm that runs in $O((k + l)(n + m) \log n/\epsilon^2)$ expected time, and returns a $(1-1/e-\epsilon)$-approximate solution to influence maximization, with at least $11/n^l$ probability. The time complexity of TIM is near-optimal, and it is empirically shown to be up to three orders of magnitude lower than any existing solution with non-trivial approximation guarantees. Subsequently, we develop an improvement of TIM [47] that retains its theoretical guarantees while improving its practical efficiency by up to another order of magnitude.

We also extended our research on influence maximization to competitive networks where several groups may simultaneously attempt to select seeds in the same network [33]. We proposed a framework called GetReal that finds the best solution for each group, who are maximizing their influences, based on game theory. This work was one of the nominees for the best paper award in SIGMOD 2015.

In a separate project, we took the first systematic step to discover k *influential event organizers* from online social networks (*e.g., Meetup* (`www.meetup.com`) who are essential to the overall success of social events [20]. These event organizers comprise a small group of people who not only have the relevant skills or expertise that are required for an event but they are also able to influence largest number of people to actively contribute to it.

The aforementioned efforts as well as numerous other social influence research in the data management and data mining communities have largely stripped off social psychology of users in their solution design. For example, these efforts ignore *conformity* of people, which refers to a person's inclination to be influenced by others. Consequently, despite the great progress made in terms of algorithmic efficiency and scalability, existing techniques may not necessarily produce high quality results in practice. In our PAELLA project[2], we investigate the interplay between psychology and social influence in online social networks and devise novel social influence solutions that are psychology-aware. Specifically, we are the first to explore techniques that incorporate conformity in computing social influence and influence maximization solutions [34, 35].

3.2 Social Image Search Results Exploration

Due to increasing popularity of social image sharing platforms (*e.g., Flickr*), techniques to support *Tag-based Social Image Retrieval* (TAGIR) [32] for finding relevant high-quality images using keyword queries have generated tremendous research and commercial interests. Many TAGIR studies attempt to improve its search accuracy or diversify its search results so as to maximize the probability of satisfying users' search intentions. In our SIERRA project, we go beyond retrieval and ranking of social images by facilitating deeper understanding through explanation and exploration of the result images.

Why-not questions on search results. Traditional TAGIR systems fail to provide a systematic framework for end users to ask why certain images are not in the result set of a given query and provide an explanation for such missing results. However, as humans, such *why-not* questions [7] are natural when expected images are missing in the query results returned by a TAGIR system. This may be due to the following reasons. First, the desired images may be ranked very low in the search results because the same keyword query (*e.g.,* "rock") may express very different search intentions for different users. Second, the set of tags associated with images may be noisy and incomplete. Consequently, not all keywords mentioned in the search query may appear as tags in relevant images. Third, the query formulated by the user maybe too restrictive due

[2] `http://www.ntu.edu.sg/home/assourav/research/paella/index.html`

to the user's limited understanding of the data collection. Indeed, it will be helpful to users if they could simply pose a follow-up *why-not* question to the retrieval engine to seek an explanation for desired missing images and suggestions on how to retrieve them. Our group developed a novel system called WINE [2] to address this challenge. Specifically, it leverages on three explanation models that exploit *Wikipedia* to automatically generates explanation to a *why-not* question posed by a user and recommends refined query, if necessary, whose result may not only includes images related to the search query but also to the why-not question.

Search results summarization. Social image search engines often diversify the search results to match all possible aspects of a query in order to minimize the risk of completely missing out a user's search intent. An immediate aftermath of such results diversification strategy is that often the search results are not semantically or visually coherent. For example, the results of a search query "fly" may contain a medley of visually and semantically distinct objects and scenes (*i.e.,* concepts) such as parachutes, aeroplanes, insects, birds, and even the act of jumping. Consequently, a thumbnail view of query results fails to provide a bird eye view of different concepts present in it. Our PRISM [42] system addresses this challenge by constructing high quality summary of top-k social image search results based on *concept-preserving* and *visually coherent clusters* which maximally cover the result set. Each cluster is represented by *minimal* tag(s) shared by *all* images in it. Due to the *concept-preserving* nature, the images in a cluster form an equivalence class with respect to the tags. Consequently, any image in each cluster can be selected as an exemplar without loss of accuracy to facilitate generation of high quality exemplar summary of the result set.

4. GEO-TEXTUAL DATA MANAGEMENT

The proliferation of GPS-equipped mobile devices has given rise to massive volumes of geo-textual or spatio-textual data (*e.g.,* points of interest, tweets, check-ins). Each geo-textual object is associated with a geo-location and a text value. In this section, we summarize the geo-textual data management challenges that we have addressed in our group.

4.1 Query Processing

A variety of classical spatial database queries and keyword queries have been revisited and rethought in the context of querying geo-textual data. Our research in this arena can be broadly classified into two streams: *spatial keyword queries* and *querying geo-textual streams*. In the former, we combine spatial functionality with keyword search (*e.g.,* find geo-tagged objects that best match the given location and keywords). Specifically, we have contributed to a variety of spatial keyword queries such as m-closest keywords [23], collective keyword [5], and keyword-aware route planning [4]. These queries typically find an aggregation of several geo-textual objects (ranked or otherwise) that are near each other. Some of our work have extended spatial queries on a spatial network that need to utilize spatial distance, which is more computationally expensive than Euclidean distance [6, 61, 62].

For the latter category, we focus on devising efficient solutions for querying streaming geo-tagged data (*e.g.,* microblog posts). Specifically, we have investigated techniques to support boolean subscription [8, 10] and similarity-based subscription [9] queries. These techniques aim to develop efficient spatial-keyword subscription strategies. More recently, we have looked into the problem of continuous queries on a stream of geo-tagged object (*e.g.,* detecting bursty region [22]).

4.2 Exploratory Search

We have also invented efficient techniques for exploring geo-tagged data. Our work can be broadly categorized into two streams: *region search* and *region exploration*. In the former, we aim to find a region for exploration that satisfies a user-defined condition (*e.g.,* size and shape of the region) and maximizes some aggregate score of the geo-tagged objects inside it [21]. In the latter category, we address the problem of exploring and discovering properties (*e.g.,* topics) of user-specified region [69].

5. INFORMATION PRIVACY

The era of big data has witnessed the collection, analysis, and sharing of individual data (*e.g.,* user behavioral records) at large scale. These data provide invaluable insights, but their usage often raises significant concerns about individual privacy. To address such concerns, a common practice is to anonymize the data by removing personal identifiers (such as names and IDs) and retaining all other information. This approach, however, has been shown to be vastly insufficient for privacy protection, since the information remained in the data may still be exploited to re-identify an individual. This motivated considerable research effort on systematic approaches for data privacy protection.

Our recent work on data privacy has focused on *differential privacy* [19], which is a strong and rigorous privacy model that has been adopted in Google Chrome and Apple iOS. In particular, we have developed techniques that improve the utility of differentially private algorithms for a number of important analytical tasks, including range count queries [56, 57, 64], model fitting [68], frequent itemset mining [50], histogram construction [55], and the synthesization of spatial, sequence, and high-dimensional data [30, 66, 67]. Most recently, we have investigated differentially private algorithms for collecting data from users who do not trust the data collector, and have devised solutions for collecting heavy hitters [41] and graph statistics [40]. In particular, our work in [56] was selected as one of the best papers in ICDE 2010.

6. FAIR PEER REVIEW MANAGEMENT

A fair peer-review process is a key ingredient for running a successful academic event. Fairness is affected by many factors, such as the expertise of reviewers, the quality of review comments, the design of the review form, etc. However, the most important factor is the relationships between authors and reviewers. In this research, we explore design and implementation of a novel reviewer suggestion system that focuses on declaration and detection of *conflicts of interest* (COIs) in the peer-review [53, 54], an issue that has received scant attention despite its significance in upholding quality and fairness of an academic event. This work is in collaboration with University of Macau and the Northeastern University, USA. Specifically, we extract relevant information related to authors by exploiting sources such as DBLP, *ResearchGate*, and *ArnetMiner*. Next, we mine relationships between the authors based on various strategies such as meta-path information [45]. Finally, we rank the COIs and display a recommended COI list of a given set of authors by utilizing a supervised ranking model that can be iteratively refined from the data collected from past COI declarations. A prototype of our system called PISTIS[3] will be demonstrated in SIGMOD 2018 [54].

Acknowledgments. First of all, we thank our graduate students, research assistants, and research fellows for their hard work and contributions. We thank our current and past faculty members (Gao Cong, Arijit Khan, and James Cheng) for their vision and scientific contributions. Last but not the least we thank our international collaborators - these include Byron Choi, Graham Cormode, Curtis Dyreson, Johannes Gehrke, Wook-Shin Han, Christian S. Jensen, Divesh Srivastava, Ryan U, Marianne Winslett, Jeffrey Xu Yu, and Shuigeng Zhou.

[3]In Greek mythology, PISTIS was the personified spirit (*daimona*) of trust, honesty, and good faith.

7. REFERENCES

[1] S. S. Bhowmick, B.-S. Seah. Summarizing Biological Networks. *Springer Verlag*, May 2017.

[2] S. S. Bhowmick, A. Sun, B. Q. Truong. Why Not, WINE?: Towards Answering Why-Not Questions in Social Image Search. *In ACM MM*, 2013.

[3] S.S. Bhowmick, H.-E. Chua, B. Choi, C. Dyreson. `ViSual`: Simulation of Visual Subgraph Query Formulation to Enable Automated Performance Benchmarking. *IEEE TKDE* 29(8), 2017.

[4] X. Cao, L. Chen, G. Cong, X. Xiao. Keyword-aware Optimal Route Search. *In PVLDB*, 5(11), 2012.

[5] X. Cao, G. Cong, T. Guo, C. S. Jensen, B. C. Ooi. Efficient Processing of Spatial Group Keyword Queries. *ACM Trans. Database Syst.,* 40(2):13, 2015.

[6] X. Cao, G. Cong, et al. Retrieving Regions of Interest for User Exploration. *In PVLDB*, 7(9), 2014.

[7] A. Chapman and H. V. Jagadish. Why not?, *In SIGMOD*, 2009.

[8] L. Chen, G. Cong, X. Cao. An Efficient Query Indexing Mechanism for Filtering Geo-textual Data. *In SIGMOD*, 2013.

[9] L. Chen, G. Cong, et al. Temporal Spatial-keyword Top-k Publish/Subscribe. *In ICDE*, 2015.

[10] Z. Chen, G. Cong, Z. Zhang, T. Z. J. Fu, L. Chen. Distributed Publish/Subscribe Query Processing on the Spatio-Textual Data Stream. *In ICDE*, 2017.

[11] J. Cheng, Y. Ke, A. Fu, J. Xu Yu, L. Zhu. Finding Maximal Cliques in Massive Networks by H*-graph. *In SIGMOD*, 2010.

[12] J. Cheng, et al. Fast Algorithms for Maximal Clique Enumeration with Limited Memory. *In KDD*, 2012.

[13] J. Cheng, Y. Ke, S. Chu, M. T. Özsu. Efficient Core Decomposition in Massive Networks. *In ICDE*, 2011.

[14] J. Cheng, S. Huang, H. Wu, A. Fu. TF-Label: A Topological-folding Labeling scheme for Reachability Querying in a Large Graph. *In SIGMOD*, 2013.

[15] J. Cheng, Y. Ke, S. Chu, C. Cheng. Efficient Processing of Distance Queries in Large Graphs: A Vertex Cover Approach. *In SIGMOD*, 2012.

[16] J. Cheng, Y. Ke, et al. Fast Graph Query Processing with a Low-cost Index. *In VLDB J.* 20(4), 2011.

[17] Shumo Chu, James Cheng. Triangle listing in Massive Networks. *In TKDD*, 6(4), 2012.

[18] H. E. Chua, S.S. Bhowmick, L. Tucker-Kellogg. Synergistic Target Combination Prediction From Curated Signaling Networks: Machine Learning Meets Systems Biology and Pharmacology. *Methods*, Vol. 129, 2017.

[19] C. Dwork. Differential Privacy. *In ICALP*, 2006.

[20] K. Feng, G. Cong, S. S. Bhowmick, S. Ma. In Search of Influential Event Organizers in Online Social Networks. *In SIGMOD*, 2014.

[21] K. Feng, G. Cong, S. S. Bhowmick, W.-C. Peng, C. Miao. Towards Best Region Search for Data Exploration. *In SIGMOD*, 2016.

[22] K. Feng, T. Guo, G. Cong, S. S. Bhowmick, S. Ma. SURGE: Continuous Detection of Bursty Regions Over a Stream of Spatial Objects. *In ICDE*, 2018.

[23] T. Guo, X. Cao, G. Cong. Efficient Algorithms for Answering the m-Closest Keywords Query. *In SIGMOD*, 2015.

[24] T. Guo, X. Cao, G. Cong, J. Lu, X. Lin. Distributed Algorithms on Exact Personalized PageRank. *In SIGMOD*, 2017

[25] K. Huang, S. S. Bhowmick, S. Zhou, B. Choi. PICASSO: Exploratory Search of Connected Subgraph Substructures in Graph Databases. *PVLDB*, 10(12), 2017.

[26] H. H. Hung, S. S. Bhowmick, B. Q. Truong, B. Choi,

S. Zhou. QUBLE: Towards Blending Interactive Visual Subgraph Search Queries on Large Networks. *VLDB J.* 23(3), 2014.

[27] T. Ideker, T. Galitski, L. Hood. A New Approach to Decoding Life: Systems Biology. *Annual review of genomics and human genetics*, vol. 2, January 2001.

[28] C. Jin, S. S. Bhowmick, X. Xiao, J. Cheng, B. Choi. GBLENDER: Towards Blending Visual Query Formulation and Query Processing in Graph Databases. *In SIGMOD*, 2010.

[29] C. Jin, S. S. Bhowmick, et al. PRAGUE: A Practical Framework for Blending Visual Subgraph Query Formulation and Query Processing. *In ICDE*, 2012.

[30] G. Kellaris, S. Papadopoulos, X. Xiao, D. Papadias. Differentially Private Event Sequences over Infinite Streams. *In PVLDB*, 7(12), 2014.

[31] H. Kim, J. Lee, S. S. Bhowmick, W. S. Han, J. H. Lee, S. Ko, M. Jarrah. DUALSIM: Parallel Subgraph Enumeration in a Massive Graph on a Single Machine. *In SIGMOD*, 2016.

[32] X. Li, C. G. M. Snoek, M. Worring. Learning Social Tag Relevance by Neighbor Voting. *IEEE Transactions on Multimedia*, 11(7), 2009.

[33] H. Li, S. S. Bhowmick, et al. GETREAL: Towards Realistic Selection of Influence Maximization Strategies in Competitive Networks. *In SIGMOD*, 2015.

[34] H. Li, S. S. Bhowmick, A. Sun. CASINO: Towards Conformity-aware Social Influence Analysis in Online Social Networks. *In ACM CIKM*, 2011.

[35] H. Li, S. S. Bhowmick, A. Sun, J. Cui. Conformity-aware Influence Maximization in Online Social Networks. *In VLDB J*, 24(1), 2015.

[36] W. Lin, X. Xiao, G. Ghinita. Large-scale Frequent Subgraph Mining in MapReduce. *In ICDE*, 2014.

[37] W. Lin, X. Xiao, X. Xie, X. Li. Network Motif Discovery: A GPU Approach. *In ICDE*, 2015.

[38] Y. Liu, B. Zheng, X. He, Z. Wei, X. Xiao, K. Zheng, J. Lu. ProbeSim: Scalable Single-Source and Top-k SimRank Computations on Dynamic Graphs. *In PVLDB*, 11(1), 2017

[39] Y. Liu, J. Lu, H. Yang, X. Xiao, Z. Wei. Towards Maximum Independent Sets on Massive Graphs. *In PVLDB*, 8(13), 2015.

[40] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, K. Ren. Generating Synthetic Decentralized Social Graphs with Local Differential Privacy. *In CCS*, 2017.

[41] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, K. Ren. Heavy Hitter Estimation over Set-Valued Data with Local Differential Privacy. *In CCS*, 2016.

[42] B. S. Seah, S. S. Bhowmick, and A. Sun. PRISM: Concept-preserving social image search results summarization. *In ACM SIGIR*, 2014.

[43] B.-S. Seah, S. S. Bhowmick, et al. FUSE: Towards Multi-Level Functional Summarization of Protein Interaction Networks.*In BCB*, 2011.

[44] Y. Song, H. E. Chua, S. S. Bhowmick, B. Choi, S. Zhou. BOOMER: Blending Visual Formulation and Processing of p-Homomorphic Queries on Large Networks. *In ACM SIGMOD*, 2018.

[45] Y. Sun, J. Han, et al. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. *In PVLDB*, 4(11), 2011.

[46] Y. Tang, X. Xiao, Y. Shi. Influence Maximization: Near-optimal Time Complexity meets Practical Efficiency. *In SIGMOD*, 2014.

[47] Y. Tang, Y. Shi, X. Xiao. Influence Maximization in Near-Linear Time: A Martingale Approach. *In SIGMOD*, 2015

[48] B. Tian, X. Xiao. SLING: A Near-Optimal Index Structure for SimRank. *In SIGMOD*, 2016.

[49] Jia Wang, James Cheng. Truss Decomposition in Massive Networks. *PVLDB* 5(9): 812-823, 2012.

[50] N. Wang, X. Xiao, et al. PrivSuper: A Superset-First Approach to Frequent Itemset Mining under Differential Privacy. *In ICDE*, 2017.

[51] S. Wang, R. Yang, X. Xiao, Z. Wei, Y. Yang. FORA: Simple and Effective Approximate Single-Source Personalized PageRank. *In SIGKDD*, 2017

[52] S. Wang, Y. Tang, X. Xiao, Y. Yang, Z. Li. HubPPR: Effective Indexing for Approximate Personalized PageRank. *In PVLDB*, 2016.

[53] S. Wu, Leong H. U, S. S. Bhowmick, W. Gatterbauer. Conflict of Interest Declaration and Detection System in Heterogeneous Networks. *In CIKM*, 2017

[54] S. Wu, Leong H. U, S. S. Bhowmick, W. Gatterbauer. PISTIS: A Conflict of Interest Declaration and Detection System for Peer Review Management. *In SIGMOD*, 2018.

[55] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu. Differentially Private Histogram Publication. *In ICDE*, 2012.

[56] X. Xiao, G. Wang, J. Gehrke. Differential Privacy via Wavelet Transforms. *In ICDE*, 2010.

[57] X. Xiao, G. Bender, M. Hay, J. Gehrke. iReduct: Differential Privacy with Reduced Relative Errors. *In SIGMOD*, 2011.

[58] X. Xiao, Y. Tao, M. Chen. Optimal Random Perturbation at Multiple Privacy Levels. *PVLDB*, 2(1), 2009.

[59] M. Xie, S. S. Bhowmick, G. Cong, Q. Wang. PANDA: Towards Partial Topology-based Search on Large Networks in a Single Machine. In *VLDB J*, 26(2), 2016.

[60] Z. Xu, Y. Ke, Y. Wang, H. Cheng, J. Cheng. A Model-based approach to Attributed Graph Clustering. *In SIGMOD*, 2012.

[61] B. Yao, F. Li, X. Xiao. Secure Nearest Neighbor Revisited. *In ICDE*, 2013

[62] B. Yao, X. Xiao, F. Li, Y Wu. Dynamic Monitoring of Optimal Locations in Road Network Databases. *In VLDB J.*, 23(5), 2014.

[63] P. Yi, B. Choi, S. S. Bhowmick, J. Xu. AutoG: A Visual Query Autocompletion Framework for Graph Databases. *In The VLDB Journal*, 26(3), 2017.

[64] G. Yuan, Z. Zhang, M. Winslett, X. Xiao, Y. Yang, Z. Hao. Optimizing Batch Linear Queries under Exact and Approximate Differential Privacy. *ACM Trans. Database Syst.*, 40(2), 2015.

[65] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, X. Xiao. Private Release of Graph Statistics using Ladder Functions. *In SIGMOD*, 2015.

[66] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, X. Xiao. PrivBayes: Private Data Release via Bayesian Networks. *In TODS*, 42(4), 2017.

[67] J. Zhang, X. Xiao, X. Xie. PrivTree: A Differentially Private Algorithm for Hierarchical Decompositions. *In SIGMOD*, 2016.

[68] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, M. Winslett. Functional Mechanism: Regression Analysis under Differential Privacy. *In PVLDB*, 5(11), 2012.

[69] K. Zhao, L. Chen, G. Cong. Topic Exploration in Spatio-temporal Document Collections. *In SIGMOD*, 2016.

[70] A. D. Zhu, W. Lin, S. Wang, X. Xiao. Reachability Queries on Large Dynamic Graphs: A Total Order Approach. *In SIGMOD*, 2014.

[71] A. D. Zhu, X. Xiao, S. Wang, W. Lin. Efficient Single-source Shortest Path and Distance Queries on Large Graphs. *In SIGKDD*, 2013.