Research Statement

Yinghui Wu (yinghui@cs.ucsb.edu)

My research interests are in the areas of *(graph) databases*, *data analytics* and *data quality*. My research is driven by a strong desire to make big, complex linked data easy to access and understand, for users at various knowledge levels.

No data is an island. The rising of big graph data, such as biological networks, social graphs, knowledge graphs, Web graphs, cyber networks and physical networks, are changing the way people produce, search, understand and exchange information. For example, the social graph of Facebook contains 1.15 billion active users, and 800 thousand new users join the network per day. Freebase, an open database of the world's knowledge, has more than 760 million entities and their relations. These real-life graphs, on the other hand, are typically "messy": *noisy*, *complex* and *large*. Freebase contains a huge amount of heterogeneous information that does not fit into a uniform, rigid schema, in contrast to relational databases. Moreover, such complexity grows as the graphs get bigger. Conventional data models and engines often fail to accommodate data management due to the complexity. These unique features create new research challenges which require a rethinking in the way we design data management solutions for graph data.

My research philosophy places a strong emphasis on the combination of fundamental theoretical analysis and system development. We design cost-effective solutions with solid theoretical results (*e.g.,* provable performance guarantees) and systems addressing the challenges. I will next describe my research contributions, its impact and future work, grouped by research themes.

Improving Graph Querying Usability

Among the vital issues to make big graphs useful is graph searching. Generally speaking, given a query and a large data graph, graph searching computes best answers for the query from the data graph. Two real-life examples are Google's Knowledge graph search over entire Web graph[1], and Facebook's Graph Search over social networks[2]. Nevertheless, searching *heterogeneous* graph data, such as knowledge graphs, is a challenging task for non-professional users. Due to the complex schemas and variational information, it becomes very hard for users to formulate queries that can be properly processed by existing systems. Traditional keyword queries are of high usability, but also come with high ambiguity. On the other hand, structured graph query languages (such as SPARQL) provide more specified semantics by supporting relationships, but are hard to write for end users. To release the users from daunting search tasks, we have been developing scalable, user-friendly searching frameworks over heterogeneous graphs.

Knowledge-based searching. The data diversity leads to low quality answers using fixed searching semantics; worse still, it is hard to determine good searching semantics using queries or data graphs alone. To address this challenge, we leverage external knowledge (*e.g.,* ontologies) to interpret both queries and data graphs via *transformations*, such that the answers that are close to queries can be identified. To support effective searching, I designed ontology-based indexes to obtain small summaries of transformed data graphs. These summaries can be efficiently traversed to find detailed matches in the original graph. Applying this framework for subgraph matching, we developed **OntQ** system[3], which finds more related matches in noisy knowledge graphs, and significantly outperforms state-of-the-art subgraph searching by both effectiveness and efficiency. The system was introduced in ICDE 2013, and was awarded the best poster award.

Query induced summarization. Understanding heterogeneous graphs is a daunting task for end users. Due to inherent ambiguity from both queries and heterogeneous data graphs, the searching process typically results in excessive number of (intermediate or final) answers that are too many to be inspected one by one. The second component of my work addresses the above challenge by leveraging data summarization techniques. We have proposed **gSum**, an effective summarization framework over the answers induced by the queries. In a nutshell, **gSum** constructs summary graphs that preserve relationships among the nodes related with queries, with requirements regarding conciseness and information coverage controlled by end users. I proposed three summary techniques, with emphasis on efficiently generating exact summary,

[1] http://www.google.com/insidesearch/features/search/knowledge.html
[2] https://www.facebook.com/about/graphsearch
[3] http://grafia.cs.ucsb.edu/ontq/

approximate summary and a set of diversified summaries, for tremendous number of answers. These algorithms effectively sketch heterogeneous answers as concise summary graphs, which in turn benefit users to better understand the semantics of the results without drilling down to details. Better still, graph queries can be induced from these summaries: a user can submit keyword queries and then pick up the desired graph queries, by adding suggested edges among the keywords. This allows users to take advantage of both keywords (of high usability) and graph query (with more expressive power). The keyword induced summarization framework **gSum** is accepted at VLDB 2013, and was introduced at ARL workshop and APG technique meeting.

Fast approximate query models. Conventional searching techniques, which resort to well-defined query languages and one-size-fits-all schema, easily become overkill in finding reasonable answers. They also lead to intractable querying tasks even for finding approximate answers, which hinder their application in big graph analytics. Although desirable, redesigning graph searching to strike balance between expressive power, usability and computation cost is challenging. We have designed a series of computationally efficient query models with provable performance guarantees on efficiency and result quality, addressing major challenges from computational complexity to system usability.

To address scalable query models in heterogeneous graphs, we have proposed a novel querying framework named **Nema**. It creates "sketches" of the neighborhoods for the nodes in a heterogeneous data graph. Given a query, *Nema* compares the neighborhood features of the query nodes with the encoded sketches of the neighborhoods of its possible matches in the data graph. The answers are induced by the data graph nodes with the best quality, which is dynamically maintained as a probabilistic inference process. Compared with conventional graph searching, *Nema* achieves great improvement in both match accuracy and efficiency. This line of work is presented at SIGMOD 2012 and VLDB 2013.

My work in this area also considers redefining "cornerstone" graph languages to cope with emerging heterogeneous graphs, including subgraph isomorphism, which is widely adopted in many state-of-the-art graph querying techniques. These techniques require strict constraints, *e.g.,* label equality and "edge to edge" matching (edge in a query graph can only be matched to another edge in data graph), which unnecessarily introduce inherent complexity to graph searching, and prevent more meaningful matches to be identified. (1) We have proposed **P-Hom**, a new graph similarity model equipped with entity similarity metrics and relaxed edge to path mapping. (2) We developed **simulation-based** graph searching framework, which consists of a range of novel query models that are highly effective at finding matches characterized by (1) many to many matching relations, and (2) flexible node and edge search conditions, all computable in polynomial time. These models are well-suited to detect "hidden" matches due to *e.g.,* heterogeneity from graph topology.

My work on novel query models has been published in premier conferences and journals. The **P-Hom** model is introduced at VLDB 2010 (Journal track). The **simulation-based searching** is presented at VLDB 2010. Its nontrivial extensions by incorporating regular expressions are introduced at ICDE 2011, and later invited to FCS Journal. My following up work on diversified social searching and view-based graph searching address further improving the usability and scalability of these query models. This line of work, introduced at VLDB 2013, supports user-defined relevance and diversified search, as well as fast searching using precached query results. We also developed an online expert searching system, **ExpFinder**, to find diversified matches for social searching of domain experts. This work is demonstrated at ICDE 2013. Our work for simulation-based, diversified and view-based searching have been adopted by *Audaque*, a startup that focuses on data quality, for complex Web object identification.

When multiple transformations (which specifies valid query answers) exist over schema-less heterogeneous graphs, it is hard to recommend good answers for users. Based on my previous work, I am working on a searching framework that can automatically determine proper ranking metrics. The goal is to seamlessly unifies offline ranking model learning and online searching. Given a library of transformations (*e.g.,* knowledge-based transformation, IR-based similarity, among others), the system should automatically learn the importance of their semantics *without* requiring manually labeled training examples. Top ranked answers should be identified accordingly via efficient online searching process. Moreover, the newly generated queries and answers can be reused for further offline training, and new transformations, indices and query logs can be readily integrated to the system. Our preliminary work of building such a system for both keyword and graph queries demonstrates its great potential for easy access of heterogeneous graphs.

Scalable Graph Processing: Beyond MapReduce

The growth of data heterogeneity is not the only force we need to contend with. The sheer volume of data makes emerging data intensive searching hard to scale, especially when it is theoretically hard, if not impossible, to reduce the complexity of the query models. Moreover, real-life graphs are often distributed. Mainstream big data processing rely on MapReduce and related techniques (*e.g.,* Google File System, Hadoop, Pregel), which is not a panacea for big graphs. To address these challenges, my work focuses on designing generic frameworks and principles applicable for a wide range of query models. In particular, we have developed generic techniques to compress and process dynamic and distributed graphs. These frameworks and principles significantly reduces *e.g.,* redundant computation, data access and communication cost, all with theoretical performance guarantees. In practice, they readily fit into MapReduce; on the other hand, they suggest effective solutions for big graphs beyond MapReduce.

Incremental searching. *How to handle dynamic real-life graphs which bear constant updates?* For dynamic graphs, incremental graph searching reduces unnecessary computation by guiding the algorithms to visit only the "affected" area in the graphs by modifications that may result changes to the query result. The outcome of this work is meaningful from both theoretical and practical perspective. From theory perspective, new complexity classes are proposed to measure the hardness of the incremental searching in terms of the size of the affected area. From practical perspective, I have proposed a range of incremental algorithms, indexes and optimization techniques for incremental searching problems, using subgraph searching, simulation-based model, and knowledge-based searching, respectively. Our incremental graph searching was introduced at SIGMOD 2011 and a top database journal TODS.

Distributed searching. *Real-life graphs are distributed.* Information sources are scattered at different sites, where the cost for inter-site communication and site access can be extremely expensive, if not impossible. For distributed graphs, our techniques deploys the search algorithms to the data that must be visited and shipped, by effective partition management, network traffic routine and site visit control, with maximized parallelism. Leveraging the idea of partial evaluation, we have developed effective distributed querying frameworks for a range of commonly used query models, including reachability queries, regular path queries, and simulation-based searching, all with provable bounds over communication cost and response time. Guaranteed by the bounds, these frameworks demonstrate performances that are not sensitive to the growth of data. In addition, we have also shown that these techniques can be readily integrated with MapReduce. Our distributed graph searching and its extensions was presented in VLDB 2012.

Query-able compression. *Can we make big graph "small" without affecting searching quality?* We proposed graph compression techniques that can be directly queried *without* decompression, which was introduced at SIGMOD 2012. The lossy compression schemes compute smaller graphs of the original data graphs that can be directly queried by *any* algorithm runnable on the original graphs, with guaranteed output quality. Better still, the searching time can be significantly reduced due to compressed input.

We are currently integrating our existing systems and the generic frameworks. The long-term goal is to develop a heterogeneous graph analytics engine that is capable to (1) efficiently extract knowledge over real-life graphs that can be accessed by a wide range of languages (including keywords, graph query and formal languages), (2) provides maximized usability in terms of good result presentation, summary and query suggestion mechanism, and (3) can be readily distributed for large-scale deployment, with effective support to handle dynamic, evolving world.

Information and Cyber Network Analytics: Application

Closely related with big graph processing, I am also working on several of its applications in the area of information network analytics and cyber network security. My research in this scope focus on how to extract useful knowledge to detect, explain and suggest solutions for network activities. This is a joint work with the students from UCSB (co-advised with Prof. Xifeng Yan and Prof. Ambuj Singh), and the scientists from Defense Advanced Research Projects Agency (DARPA), US Army Research Lab (ARL), and Aberdeen Proving Ground (APG). (1) While the need is evident, predicting the trace of information is challenging in emerging information networks especially in the context of incomplete information. Our work on the cascade inference proposes inference algorithms to predict the information flows using only small amount of

partially observed information (*e.g.,* the time and location the information is observed). Our work on cascade inference has been presented at ICDM 2012. (2) Within cyber security, I am working on techniques for (a) how to measure and improve the robustness of cyber networks in the presence of attacks and specific tasks, and (b) how to mine, summarize and query the causality of network activities to make suggestions for network performance monitoring and synopsis. This work is associated with Plan X, a long-term foundational cyberwarfare program of DARPA. We are investigating how our graph mining techniques can be integrated as toolbox for large-scale cyber network security evaluation and defense. Our preliminary work is introduced in the annual Science for Cybersecurity (S4C) 2013 workshop of ARL. We are keeping close collaboration with ARL and APG on network security analytics. Our network event summarization techniques are being pursued by LogicMonitor, a local cyber network analytics startup.

With the prevalence of large-scale cyber networks, the cyber threats has never been more apparent. Today's defense sectors require smart, scalable network monitoring, analysis and diagnosis systems. My research shall provide such techniques from the perspective of graph analytics over this important application area, including real-time network performance evaluation and diagnose, event pattern (*e.g.,* causality and correlation) detection and maintenance over cyber and information networks.

Collaboration

My research keeps close collaboration with the University of Edinburgh, UCSB, UIUC, IBM Research, BBN and research organizations from DARPA, ARL, APG, Salinas Police Department, as well as industry partners such as Aptima, LogicMonitor, and Audaque, among others. Our collaboration has made a range of contributions that address several fundamental challenges in big graph data processing, from new findings in complexity and algorithms to practical development and deployment of graph management systems. I highly appreciate all the support I have received from these ground organizations, researchers and partners.

Summary

Real-life graphs are large, heterogeneous and noisy. My current research aims to develop cost-effective analytics techniques for emerging big graph applications, with desirable guarantees on quality, efficiency and usability. My research emphasizes the combination of theoretical analysis and practical implementation with real-life impact. To date, my research has made contributions to the following areas: query models designing for big graphs, heterogeneous graph searching, distributed graph management, and information/cyber network analytics. My long-term goal is to make real-life graphs really easy to access and useful for a significant fraction of the world's population.