

Research Statement

Hanghang Tong

1 Overview

Graphs appear in a wide range of settings and account for a large portion of real world data sets. For example, in sociology, the nodes are individuals and the edges represent the interaction between two persons (e.g., collaboration, trust, contact, etc); in computer networks, the nodes are routers or autonomous systems and edges represent the connection between two routers/autonomous systems; in user psychology, the nodes are people and items, and the edges represent some actions between the user and the items (e.g., the user *clicks* the web page, the user *recommends* some product, etc.); in ecology, the nodes are species, and edges represent prey-predator relationship; in biology, the nodes are proteins and the edges represent the interaction between two proteins (e.g., both are critical for some biological process to happen). Given such a social network, how to measure the closeness, and how to track it over time? How to identify abnormal behaviors of computer networks? In the case of virus attacks, which nodes are the best to immunize given the limited resources?

My research theme is to help the user to better *understand* and *utilize* large real graph data sets. More specifically, there are three closely related dimensions of this research goal:

- G1. (**Querying**) Given a graph (say, a social network), how to help the user to find things according to his/her particular interest?
- G2. (**Mining**) Given a graph, how to succinctly describe it, and report anomalies?
- G3. (**Scalability**) How to scale our querying and mining algorithms to large graphs, spanning multiple machines?

2 Motivation: Why Study Graphs?

These real-world graphs have posed a wealth of fascinating research questions and high-impact applications. Among others, our motivating applications are:

- (Social Networks) Effective querying and recommendation tools are playing an important role in on-line social network sites, - with hundreds of millions of users.
- (Security) Graph querying algorithms can help to find suspicious subgraphs (e.g., mastermind criminal in law enforcement, money-laundering ring in financial fraud, suspicious communication patterns, etc).

- (Epidemiology) A good immunization strategy might help to prevent an epidemic from out-breaking with the lowest cost.
- (E-commerce/Viral Marketing) A good immunization strategy can also help to spot the ‘best’ customers for advertisement (‘k-advertisement’) in viral marketing, which can largely improve the revenue.
- (Communication networks) Graph mining algorithms can help to detect abnormal behaviors in both computer networks and phone networks (e.g., port scanning, router mis-configuration, telemarketing, etc)

The relationship between these applications and our long term research goal is summarized in table 1, where rows are the research goals and columns are the driving applications.

Table 1: Applications of Long Term Research Goals

	Social Networks	Security	Epidemiology	E-commerce	Communication Networks
G1: Querying	✓	✓			
G2: Mining	✓		✓	✓	✓
G3: Scalability	✓	✓	✓	✓	✓

3 Current Achievements

In my thesis, we address the above challenges in multiple dimensions, by focusing on two types of tasks according to the interaction with users: querying and mining. For the task of querying, we want to answer the complex user-specific patterns, such as Center-Piece Subgraphs (*Given three criminals, who is the master-mind?*). For the task of mining, the goal is to summarize/compress a graph, and report anomalies. The main contributions of our current work can be summarized as follows:

3.1 Querying Graphs

- **Complex User-Specific Patterns.** We found that many complex user-specific patterns on large graphs can be answered by means of proximity measurement. In other words, *proximity allows us to query large graphs on the atomic levels.* We support our claim by conducting four case studies (center-piece subgraphs (KDD’06), “best-effort pattern match” (KDD’07 b), interactive querying (ICDM’08, CIKM’09), and recommendation (KDD’09)), all of which (despite the difference in applications) rely on proximity measurement as their building block.

Impact/Results. The proposed algorithms are operational, with careful design and numerous optimizations. They led to 3 patents pending. Our techniques for “best-effort pattern

match” can spot potential money-laundry ring, and Lawrence Livermore National Laboratory (LLNL) has obtained a copy of my code. The proposed algorithms for both center-piece subgraphs and interactive querying are to be deployed into a real product (*Cyano*) in IBM (Qu+ SCC08).

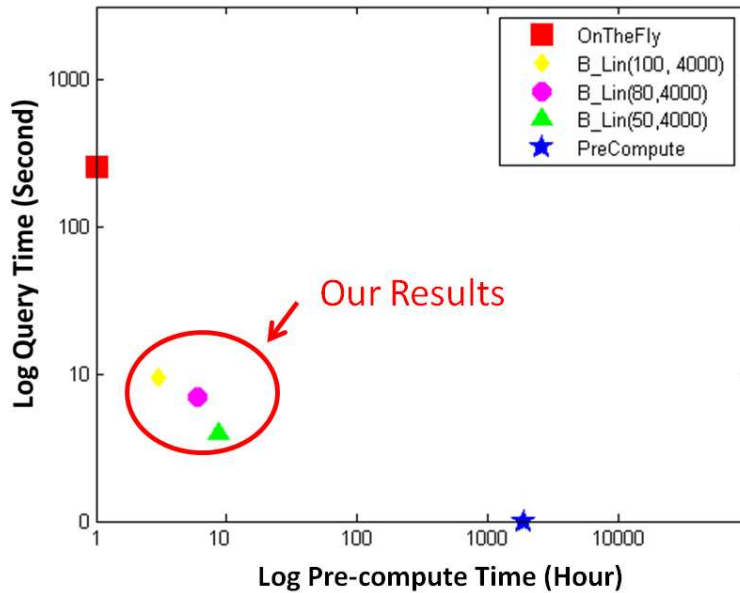
- **Proximity Tracking.** We proposed an efficient algorithm *pTrack* (SDM’08) to track proximity on time-evolving graphs.

Impact/Results. It enables us to do trend analysis on the graph level. The proposed algorithm (*pTrack*) is up to $176x$ faster than competitors and has *no* quality loss. This work won the *Best Paper* award in SIAM-DM 2008.

- **Fast Proximity Computations.** We developed a family of fast solutions (*FastRWR*) (ICDM’06, KDD’07 a, SDM’08, KAIS’08, SAM’08) to compute proximity in several different scenarios. The idea was to leverage important properties shared by many real graphs (e.g., the block-wise structure, the linear correlation, the skewness of real bipartite graphs, etc).

Impact/Results. We can often achieve orders of magnitude (*up to 6,000,000x*) speedup with little, etc) or *no* quality loss. Figure 1 presents some experimental results by one of these works (“B_Lin”) on a DBLP co-authorship graph. This work (“B_Lin”) won the *Best Research Paper* award in ICDM 2006.

Figure 1: Evaluation of the proposed “B_Lin” on a DBLP co-authorship graph. It achieves a good balance between the pre-compute time and the on-line query time, with more than 90% quality preserving. Note that both axes are in the logarithm scale.



3.2 Mining Graphs

- **Vulnerability Analysis.** We proposed an algorithm “NetShield” for immunization under the SIS (susceptible-infectious-susceptible) model.

Impact/Results. While straight-forward methods are computationally intractable ($O(\binom{n}{k}m)$), the proposed algorithm is *near-optimal*, *fast* (over to 7 orders of magnitude speedup), and *scalable* ($O(nk^2 + m)$), that is, linear on the number of nodes n and linear on the number of edges. Figure 2 presents a case study of the proposed “NetShield” on the Zachary’s karate graph.

- **Anomaly Detection.** We proposed a family of example-based low-rank matrix approximation methods “Colibri” (KDD’08) for anomaly detection.

Impact/Results. The proposed algorithms are provably equal to or better than the best known methods in both space and time, with the same accuracy. On real data sets, it is up to $112x$ faster than the best competitors.

- **Mining Complex Time-Stamped Events.** We show that graphs also provide a very powerful tool to solve some complex problems. As a case study (CIKM’08), we proposed a general framework (“ $T3$ ”) to mine complex time stamped events, by casting the problem as a graph analysis problem. We further proposed “MT3” to handle multiple-scale analysis.

Impact/Results. The proposed “ $T3$ ” is able to find similar time stamps, find abnormal time stamps and provide interpretations for our findings. The proposed “MT3” achieves up to 2 orders of magnitude speedup, with the same quality.

Figure 2: Evaluation of the proposed “NetShield” on the Zachary’s karate graph. The best-5 nodes selected by “NetShield” are in black. The results agree with the intuition: deleting these black nodes will make the remaining graph the most robust (i.e., the least vulnerable) to the virus attack.



3.3 Other Achievements

I have also worked on numerous topics on data mining, including image retrieval, blur detection, image quality assessment, within network classification, graph visualization, link prediction, community detection, etc. For all these problems, there are important real applications behind them. For example, while I was an intern in Microsoft Research Asia (MSRA), my group was among the first to study blur detection for digital images back in 2004. My understanding is that MSRA was trying to file a patent based on my work and several companies (e.g., Nikon and Canon) were interested in licensing the code. Currently, this functionality is common in many digital cameras (e.g., Nikon and Canon). To our best knowledge, we were the first to study no-reference holistic image quality assessment. One of my papers for image retrieval received more than *100* times citations (according to Google Scholar, Dec. 2, 2009).

4 Vision for the Future

Graphs provide a very powerful and unified tool to handle data heterogeneity, with an intuitive user interface. On themselves, graphs pose a wealth of fascinating research questions and high-impact applications. It is my belief that graphs will continue to play an even more important role in our lives: more and more real applications will rely on graphs; much richer types of graphs will show up; and the scales of real-world graphs will continue to grow.

My long-term research theme is to help the user to better *understand* and *utilize* large real graph data sets. Thus, my research will focus on the following five aspects, which are separated in medium term goals (M1-M3) and long term goals (L1-L2).

M1. Design new algorithms for recommendations on large graphs.

M2. Design new algorithms for immunization.

M3. Improve the usability of graph querying and mining results, by giving interpretation and summarization of querying and mining results.

L1. Address the scalability issue.

L2. Address rich types of data, specifically weighted graphs, attributed graphs, time-evolving graphs, and geo-coded graphs.

Next, we will present our medium term plan and long term plan, respectively. These steps, and their relationship with my thesis work, are summarized in table 2.

4.1 Medium Term Plan

In the near future, we will focus on the following three tasks, all of which are built on the thesis work:

M1: Broad Spectrum Recommendation Systems

A large portion of the thesis work focuses on querying large graphs. In other words, if the user

Table 2: Vision for the Future

Plans Goals	Step 1 (thesis work)	Step 2 (medium term)	Step 3 (long term)
G1 (Querying)	Chapter 2-6	Recommendation Interpretable querying	Querying rich types of data
G2 (Mining)	Chapter 7-9	Immunization Interpretable mining	Mining rich types of data
G3 (Scalability)	Chapter 2-9	O(E) or better (a single machine)	Scalable on Map-Reduce Scalable on rich types of data

knows what s/he exactly wants, we are now in a better position in helping them to find such things (e.g., center-piece subgraphs, gateway, etc). In the next step, we would like to help the user to find things that s/he might not (or partially) know, where recommendation plays a crucial role.

While most of the existing work focuses on relevance (i.e., find things that are most relevant to the user’s interest), there are other important aspects in recommendation, e.g., novelty, diversity etc. For example, our preliminary work (KDD’09) shows that by taking into account the novelty in recommendation, we can broaden user’s horizon.

Here, our ultimate goal is to provide the user a subset of items which covers the broad spectrum of his/her interest (e.g., relevance, diversity and novelty). In order to achieve this goal, we need to work on ‘*broad spectrum recommendation*’, where we aim to *collectively* find the whole recommended subset, instead of a list of *individual* items.

M2: Immunization

In the thesis, we have designed a very promising immunization algorithm for SIS (susceptible-infectious-susceptible) model. We will generalize our work to (1) immunize under other types of virus propagation models (e.g., SIR (susceptible-infectious-recovered), or the mixture of SIS and SIR, etc); (2) immunize in the case the graph structure is changing over time).

M3: Interpretation of Querying and Mining Results

Most real data sets do not have labels. It usually takes a lot of time for the analyst to check/understand the mining results. Therefore, it is important to generate a concise and intuitive explanation for the user to better understand the mining results. In the thesis, we show that a few representative examples are usually very helpful to interpret the querying and mining results (e.g., communities, anomalies, etc).

We will continue on this line of research to further improve the usability of mining results. Here, the two main research questions we will address are (1) how to select a few examples/nodes as ‘basis’; (2) how to use the selected examples to interpret the remaining nodes (e.g., by a sparse nonnegative linear combination).

4.2 Long Term Plan

In the long run, we will focus on the following two directions, all of which are common to both G1 (querying) and G2 (mining):

L1: Scalability

As the scale of the real data continues to grow, scalability is a ‘never-ending’ question in large graph mining. Here, we will deal with this issue through the following two orthogonal efforts: (1) continue to design scalable (linear or better) algorithms on a single machine; (2) explore map-reduce type abstractions for large scale computation on graphs, where the challenge is how to de-couple the computation among different machines.

L2: Rich Types of Graph Data

Most existing algorithms work on plain undirected graphs. We plan to extend our work to graphs with attributes (both on nodes and edges), time-evolving graphs, directed weighted graphs. As the main tool for analyzing single plain graphs is matrix algebra, in order to extend our algorithms to such types of graphs, we need to simultaneously analyze multiple inter-correlated matrices or to analyze tensor (the generalization of matrices). On the other hand, although graphs account for a large portion of real data sets, there are other types of data sets (e.g., spatial, temporal, etc). In my thesis, we show that we can handle complex time-stamped events by casting the problem as a graph analysis problem. We will continue on this line of research. Ideally, we would like to develop a unified model to handle such complex data (the mixture of relational, temporal and spatial data).