Graph Sampling, Summarization, and Touch-Based Visual Analytics for Large Complex Systems

PROJECT ID | CS15-4 **TYPE |** [ ] New [X] Continuing **START DATE |** July 2015

PROJECT LEAD/PARTICIPANTS | Christoph Borst, Mehmet Engin Tozal, Nicholas Lipari, Henry Chu, Raju Gottumukkala, Ryan Benton

DESCRIPTION | We seek to enable interactive visual analytics of large-scale graphs using novel graph sampling methods and touch-based interfaces. Recently, there is a significant interest in modeling and studying real-world complex systems as large-scale graphs with numerous interconnected or interacting entities. Many real-world systems such as online social networks (OSN), world wide web (WWW) and Internet topology maps (ITMs) are very large, so capturing them in their entirety, analyzing them to extract useful information, and visualizing them for decision making are resource-consuming and challenging tasks. It is necessary to develop graph sampling approaches and integrate them with human-computer interfaces to study these large-scale graphs to understand the underlying real-world systems.

The PIs will investigate sources of information loss in a graph sampling process and identify fundamental factors that need to be carefully considered in a sampling design. We also plan to develop a software system as an extension to open source libraries (igraph/networkX/boost) that employs different sampling methodologies to estimate important graph characteristics. Developing an extension to igraph/networkX/boost, instead of a standalone application, allows more seamless integration of our work with other CVDI projects. Furthermore, the project will improve interactive visual analysis of large graphs by prototyping interface methods in combination with machine analytics. We will develop multitouch and gesture techniques to provide intuitive user control of navigation, filtering, clustering, and highlighting during visual analysis. The efficiency and clarity of interfaces is critical for the success of visual analytics systems and helps users understand results and analysis processes.

EXPERIMENTAL PLAN | During the project period, the team will work on the following tasks:

- Investigate the sources of information loss in a graph sampling process.
- Identify the fundamental factors that need to be carefully considered in a sampling design.
- Prototype a primitive interaction set for basic actions on graphs (e.g., like Semantic Action Taxonomy Gotz & Zhou, 2009).
- Design human-computer interaction techniques such as large graph search, controlled visualization, touch/gesture based graph component selection and manipulation.
- Develop a software system (as an extension to igraph/networkX/boost) that employs different sampling, visualization and interaction methodologies to display graphs and analyze their characteristics.

RELATED WORK | Sampling design considerations are not necessarily used in many studies involving analysis of large graphs. Existing studies often use sample graphs to make conclusions about the population graphs without noticing the discrepancy between the sample and the population. This contributes to incorrect inferences about the studied characteristics of the underlying systems.

On the interaction side, touch and gestural data exploration is a rapidly expanding field, but results for large graphs are minimal. Examples range from basic plots (MS Touchvis, Touchwave) to node-link diagrams (Max Planck Research Networks). Kinetica focuses on touch and point-based data but is not intended for large datasets. Apolo uses a relatively conventional interface but is interesting for integrating interaction with machine learning. Another tool, GRAPHITE, is notable for search based on user-drawn patterns.

HOW OURS IS DIFFERENT | The PIs plan a large-scale study to empirically demonstrate the relation between various sampling designs and the accuracy of resulting samples in estimating various population characteristics. The PIs further plan to integrate graph sampling into visualization to enable mining large graphs through a novel touch-based interface. The PIs believe the effectiveness of sampling, visualization, and interaction depend on: (a) the characteristic under study (e.g., degree, size, clustering coefficient, betweenness, density, etc.), (b) topology of the population graph (e.g., scale free, small world, random, or semi-hierarchical graph models), and (c) sampling scheme (e.g., node-based, link-based, walk-based, path-based, motif-based).

MILESTONES FOR YEAR 1

3 months: Obtain & prepare data sets for the empirical study and develop touch based browsing operations for features such as node aggregation and summarization.

6 months: Develop sampling design techniques for different networks and network characteristic of interest and design analytics commands via gesture or direct manipulation.

12 months: Develop an extension to igraph/networkX /boost for graph sampling, visualization and touch based interaction.

DELIVERABLES

- Practical knowledge about: (i) sources of information loss in a graph sampling process, (ii) fundamental factors that need to be carefully considered in a sampling design and (iii) graph visualization and multitouch exploration.
- Software and demonstration of visual analytics interface with multitouch exploration of large graphs and extension to igraph/networkX/boost with graph sampling methods.

BUDGET FOR YEAR 1

Students	$65,000
Supplies	$1000
Equipment	$16,000
Travel	$6,000
Overhead	$9,000
Total	**$97,000**

ECONOMICS | Navigating, mining, and presenting large datasets are essential to most industry sectors to maximize competitiveness. Real-time BI solutions with high end visual analytics can improve efficiency and quality of insight. Massive datasets are becoming common in government, enterprise, and science. Analysts need to rapidly gain insight to understand emerging business trends from customer & product graphs, identify disease spread across community networks, assess healthcare delivery trends across healthcare networks, etc. Advances in sampling can improve analysis and visualization of very large graphs. This study will enable us to gain practical knowledge about fundamental factors effecting different sampling schemes and develop accurate sampling designs to analyze and visualize graphs with minimum overhead.
POTENTIAL MEMBER COMPANY BENEFITS | CVDI members use graphs to model, analyze and visualize their underlying systems and business processes to make informed decisions and gain insights from complex heterogeneous data. Effective graph sampling, visualization, and interaction methods for various tasks will complement ongoing CVDI projects that employ graphs as a tool to model and analyze real-world systems and complex data. Novel approaches and methods may lead to advantages in monitoring operations, understanding data, retaining and acquiring customers, increasing revenue, etc. We aim to deliver potentially faster and more powerful visual analytic interfaces and more accessible and understandable graph browsing interfaces.
PROGRESS TO DATE | We previously prototyped related interfaces in earlier years, but for different device or data types (e.g., geospatial data, hand held devices). We will investigate and employ graph sampling and summarization to improve/extend our work.
KNOWLEDGE TRANSFER TARGET DATE | 12 months