PROJECT SUMMARY: CAREER: Signed Graph Analysis for Real Data:

Overview

The underlying relationship of data exists in the context of application, time, setting, and purpose. Graph representation is an intuitive way to capture contextual relations, and it is straightforward for the structured data. To date, life sciences and industry have benefited from structured data graph representation analysis models that efficiently solve complex recommendation, matching, and optimization problems. Unstructured data are a collection of various data types and formats, e.g. records, surveys, experimental readings, tweets, and purchase reviews data. The underlying relationships of these data items might not be clearly defined, so the chosen graph representation (which relations to model and how) has a direct impact on the interpretation of the graph analysis outcomes. The proposed work focuses on formalizing graph representation of the unstructured data and expanding graph analysis for the data collected in real-life processes. Such graphs are often large (tens of millions of nodes), sparse and structure-free, and existing algorithmic assumptions fail. The project will develop consensus attributes based on balanced graph theory, graph community discovery and recommendation methods that mitigate assumptions and use the data context for the the graph representation. Graph consensus analysis will produce an unbiased interpretation of the graph and identify weak relationships, groups of vertices, and disparately impacted groups of vertices. Depending on the domain, this can transfer to unreliable recommendations, influential cliques, or experimental anomalies. This will help drive the work around measuring the impact of graph construction from unstructured data sources on algorithmic performance and algorithmic evaluation in terms of effectiveness.

Intellectual Merit

The proposed work will propose the solution for graph construction that meets the algorithmic assumptions on small diameter, density, size, and topology. It will produce training- and assumption-free approaches for large and sparse graph analysis and datasets and metrics to measure community discovery and bias in the graph. The project will formalize the mapping of unstructured data to network graphs to inform subsequent assumptions and limitations for graph analysis.

Broader Impacts

The project will drive the development of course materials on the impact of unstructured data representation and analysis, as well as outreach activities and training for data scientists in the public and health sectors, specifically the use and impact of unstructured data representation and analysis methods for public health resilience, physiochemical data, and engineering data. The project will seed the growth of the network science branch at the newly established Data Science Center at Texas State University. The project will provide graph representation guide for the scientific community to use, and the domain data scientists will benefit from documenting the impact of data representation on the analysis outcome and from new analysis tools. We will release all project products as open source and provide documentation. Many dozens of students from underrepresented groups will be involved in the network science project within the proposed work.