Relational subgraph analysis in large scale complex networks

Project Title: Diffusion on Complex Networks: Algorithmic Foundations

Anil Kumar S. Vullikanti
Dept. of Computer Science and
Virginia Bioinformatics Institute
Virginia Tech, Blacksburg, VA 24061, USA
E-mail: vsakumar@vt.edu

Abstract

A canonical problem in several applications, such as social network analysis, data mining, fraud detection, chemical informatics, web information management and bioinformatics, involves relational queries in complex networks and semi-structured datasets– is there a subgraph that is "similar" to a specific template? In general, the nodes and edges of the networks and templates have different kinds of attributes (e.g., demographic and temporal properties and relationships), referred to generically as "labels", and such queries can capture complex relationships between different entities. Subgraph queries have numerous uses, such as: finding hidden relationships between different entities in the network (e.g., identifying specific interaction patterns in social, financial or terrorist networks to detect fraud or suspicious activities), characterizing graphs by means of measures that go beyond the degree distribution (e.g., using "graphlets" or "network motifs" for characterizing and comparing graph families), and understanding the dynamics of diffusion processes on graphs (e.g., characterizing vulnerable nodes or super-spreaders by means of subgraph properties).

Many versions of such relational subgraph analysis problems arise in different applications, including identifying one or all possible matches for a query, as well as computing specific functions (e.g., averages of the properties of the nodes matching the queries) over all possible matchings. In their most general form, queries can represent complex relationships which may be more involved than direct subgraph isomorphism; formal languages, such as graph grammars can be used to capture such relationships in a more systematic manner.

Subgraph analysis is computationally very challenging, especially for large templates and networks (which cannot be stored in memory), which is our focus. The labeled structure of these networks makes the problems even more involved. There is a large body of research for exact subgraph analysis, e.g., using the Apriori method based on breadth-first or depth-first ordering, and approximate subgraph analysis, e.g., using color coding. However, it is generally difficult to parallelize these approaches, and be able to handle very general kinds of queries, especially when the data is stored across different computing platforms, and computing resources are diverse.

We discuss a general and flexible framework for approximate relational subgraph analysis that scales to networks with millions of nodes, and can handle diverse kinds of queries. Our techniques are based on a parallel and streaming implementation of color-coding, and the main technical contribution involves an efficient implementation of the underlying dynamic programming step. We discuss SAHad, a Hadoop based implementation that can handle large templates on networks with a few million nodes and edges. Further, the algorithm and its implementation are very flexible and do not require complex development and machine dependent optimization. We demonstrate this by running our implementation on the Amazon EC2 cluster without any significant changes. We illustrate the power of such relational subgraph analysis by using it to analyze and characterize synthetic social contact networks and diffusion processes on these networks.