Scalable Subgraph Counting: The Methods Behind The Madness

C. Seshadhri
University of California, Santa Cruz, CA
sesh@ucsc.edu

Srikanta Tirthapura
Iowa State University, Ames, IA
snt@iastate.edu

ABSTRACT

Subgraph counting is a fundamental problem in graph analysis that finds use in a wide array of applications. The basic problem is to count or approximate the occurrences of a small subgraph (the pattern) in a large graph (the dataset). Subgraph counting is a computationally challenging problem, and the last few years have seen a rich literature develop around scalable solutions for it. However, these results have thus far appeared as a disconnected set of ideas that are applied separately by different research groups. We observe that there are a few common algorithmic building blocks that most subgraph counting results build on. In this tutorial, we attempt to summarize current methods through distilling these basic algorithmic building blocks. The tutorial will also cover methods for subgraph analysis on "big data" computational models such as the streaming model and models of parallel and distributed computation.

CCS CONCEPTS

• **Information systems → Graph-based database models**; • **Theory of computation → Graph algorithms analysis**.

KEYWORDS

subgraph counting; motif counting; graphlet counting; sampling; edge orientation

ACM Reference Format:
C. Seshadhri and Srikanta Tirthapura. 2019. **Scalable Subgraph Counting: The Methods Behind The Madness. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308560.3320092**

1 INTRODUCTION AND GOAL

This tutorial focuses on *subgraph counting* (also called motif counting or graphlet analysis), which is an umbrella term that refers to problems where we wish to count or approximate the number of occurrences of a (small) subgraph H in a large graph G. An example problem is triangle counting, where we seek to approximate the number of instances of a *triangle*, a complete subgraph on three vertices, in a graph G. The term "subgraph counting" also encompasses versions where vertices have attributes, edges have timestamps, and cases where we wish to get (local) counts for each vertex. Depending on the application, there may be a need for algorithms in different computational models (streaming, distributed, sublinear, etc.).

Subgraph counting has a wide variety of uses, including:

- Social network analysis: Classifying social behavior, social behavior in gaming networks, explaining roles of nodes, predicting social tie strength, analyzing collaboration patterns. Subgraph counts form the basis of the widely used definition of graph clustering coefficients.
- Graph processing: Detecting dense subgraphs and communities, spam detection, modeling real-world graphs, semantic user search.
- Bioinformatics: Characterizing networks and roles of nodes within them, through the concepts of network motifs and graphlets.

Recent years have seen a surge of interest in subgraph counting in large graphs, in areas such as web search, social and biological network analysis. There is a dizzying array of papers on subgraph counting/approximation that have appeared in traditional data mining venues (The Web Conference, KDD, SDM, WSDM, etc.), database venues (VLDB, SIGMOD, PODS, etc.), and traditional theoretical computer science venues (FOCS, STOC, SODA, etc.) Such a large body of literature is overwhelming to track even for an active researcher in the area. Moreover, it is challenging for a practitioner to understand which result is best suited for their specific subgraph mining task.

Goal: The authors, while following the subgraph counting literature, have observed a set of algorithmic building blocks that are common to most of the results. Some examples of these building blocks are: edge sparsification, color coding, using graph orientations, and path sampling. Significant ingenuity is required to adapt these ideas to the problem and computational model at hand. Nonetheless, these common algorithmic building blocks form the foundation of much of this literature. Our aim is to highlight these techniques, and classify results according to the techniques they build on. The specific goals of the tutorial are to:

- Present algorithmic building blocks of scalable subgraph counting.
- Explain research results in the context of the above building blocks, and survey the research landscape.

Tutorial Structure: The tutorial will begin with an introduction to the problem of subgraph counting, its importance and applications. The tutorial will then introduce the various algorithmic building blocks. through existing results, and discuss how these building blocks can be implemented in different computational models. Given this foundation, we will classify many important results in the literature by the tools they build upon. We will not attempt an exhaustive survey, but we hope to give a bird's-eye view

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3320092

of the research landscape. We will provide examples to illustrate algorithmic ideas, and provide proof sketches as appropriate.

2 ORGANIZERS

C. Seshadhri is an Associate Professor of Computer Science at the University of California, Santa Cruz. His primary interest is in mathematical foundations of big data, especially modeling and algorithms. His work spans many areas: sublinear algorithms, graph algorithms, graph modeling, scalable computation, and data mining. Seshadhri has worked extensively on the topic of graph algorithms and data mining. Related to the topic of the tutorial, he has contributed several significant results. These include the technique of "wedge sampling" for subgraph counting, fast exact algorithms for counting small subgraphs, and state-of-the-art theoretical and practical algorithms for approximating counts of k-cliques.

Srikanta Tirthapura is the Kingland Professor of Data Analytics at Iowa State University. His research is centered on the foundations of large-scale data analysis, especially streaming and parallel algorithms applied to large-scale data. He also has interests in applications of data analytics to areas such as security, and has worked extensively on algorithms in the streaming and incremental models. On the topic of the tutorial, he has contributed ideas such as "neighborhood sampling" for sampling and counting subgraphs from a data stream, which has been implemented in state-of-the-art systems for subgraph analysis, as well as works on parallel streaming algorithms for subgraph counting.

3 DURATION, AUDIENCE, OUTCOMES

This is a half-day tutorial, whose expected audience is:

- Algorithm designers and practitioners interested in large graph analysis
- Researchers in the domains of social network analysis and network measurement
- Researchers interested in the use of randomized methods for big data

The only prerequisites required is basic familiarity with graphs and probability, and some knowledge of undergraduate graph algorithms (such as adjacency list representations, BFS, DFS, etc).

Expected Outcomes: The audience will get an overview of the main algorithmic ideas used in subgraph counting results. Practitioners will get guidance on which ideas or tools to use for their specific problems (and models of computation), and will gain an understanding of where to look within the literature. Researchers in the areas of subgraph counting, graph algorithms, and randomized methods will gain from a survey of the cutting-edge results in this area.