Distributed Graph Algorithmics: Theory and Practice

Silvio Lattanzi
Google Research, New York
silviol@google.com

Vahab Mirrokni
Google Research, New York
mirrokni@google.com

Categories and Subject Descriptors

F.0 [**Theory of Computation**]: General

Keywords

Parallel computing, MapReduce algorithms, Large scale data-mining

ABSTRACT

As a fundamental tool in modeling and analyzing social, and information networks, large-scale graph mining is an important component of any tool set for big data analysis. Processing graphs with hundreds of billions of edges is only possible via developing distributed algorithms under distributed graph mining frameworks such as MapReduce, Pregel, Gigraph, and alike. For these distributed algorithms to work well in practice, we need to take into account several metrics such as the number of rounds of computation and the communication complexity of each round. For example, given the popularity and ease-of-use of MapReduce framework, developing practical algorithms with good theoretical guarantees for basic graph algorithms is a problem of great importance.

In this tutorial, we first discuss how to design and implement algorithms based on traditional MapReduce architecture. In this regard, we discuss various basic graph theoretic problems such as computing connected components, maximum matching, MST, counting triangle and overlapping or balanced clustering. We discuss a computation model for MapReduce and describe the sampling, filtering, local random walk, and core-set techniques to develop efficient algorithms in this framework. At the end, we explore the possibility of employing other distributed graph processing frameworks. In particular, we study the effect of augmenting MapReduce with a distributed hash table (DHT) service and also discuss the use of a new graph processing framework called ASYMP based on asynchronous message-passing method. In particular, we will show that using ASyMP, one can improve the CPU usage, and achieve significantly improved running time.

1. SHORT BIOGRAPHIES OF THE TUTORS

Silvio Lattanzi

Silvio Lattanzi is a Senior Research Scientist at Google Research New York. He received his Phd from University La Sapienza of Rome. During his PhD he interned twice at Google Research Mountain View and once at Yahoo! Research Santa Clara, he also spent a semester visiting University of Texas Austin. His main research interests are in large-scale graph mining, information retrieval and probabilistic algorithms. Silvio has published several papers in top tier conferences in information retrieval, social network analysis and algorithms. He also served as program committee or senior program committee for several conferences including WWW, WSDM and KDD. Recently, he serves as a tech lead for the large-scale graph mining project at Google Research New York.

Vahab Mirrokni

Vahab Mirrokni is leading the algorithms research group at Google Research New York where he is a Senior Staff Research Scientist. He joined Google after research appointments at Microsoft Research, MIT, and Amazon.com. He received a B.Sc. from Sharif University in 2001 and a Ph.D from MIT in 2005. He has published more than 100 papers in journals and conferences and has filed more than 30 patents in these areas. He has been the co-winner of the SODA 2005 best student paper award, and the ACM EC 2008 outstanding paper award. Vahab's main research interests are algorithms design, large-scale graph mining, and the Internet Economics. He has served as a PC, and Senior PC member of several top conferences in these areas including EC, WWW, KDD, WSDM, and SODA. Previously, he has presented tutorials at ICML, WWW, ACM EC, and SIGMETRICS conferences. At Google, he leads several efforts in optimization for online advertising and large-scale graph mining. Vahab is now the tech lead manager for the large-scale graph mining project at Google Research New York.