Afternoon Tutorial

Network Sampling

Lise Getoor
University of Maryland, College Park

Ashwin Machanavajjhala
Duke University

Abstract

Network data appears in various domains, including social, communication, and information sciences. Analysis of such data is crucial for making inferences and predictions about these networks, and moreover, for understanding the different processes that drive their evolution. However, a major bottleneck to perform such an analysis is the massive size of real-life networks, which makes modeling and analyzing these networks simply infeasible. Further, many networks, specifically those that belong to social and communication domains, are not visible to the public due to privacy concerns, and other networks, such as the Web, are only accessible via crawling. Therefore, to overcome the above challenges, researchers use network sampling overwhelmingly as a key statistical approach to select a sub-population of interest that can be studied thoroughly.

In this tutorial, we aim to cover a diverse collection of methodologies and applications of network sampling. We will begin with a discussion of the problem setting in terms of objectives (such as, sampling a representative subgraph, sampling graphlets, etc.), population of interest (vertices, edges, motifs), and sampling methodologies (such as Metropolis-Hastings, random walk, and snowball sampling). We will then present a number of applications of these methods, and will outline both the resulting opportunities and possible biases of different methods in each application.

Instructors

Mohammad A. Hasan is an Assistant Professor of Computer Science at Indiana University–Purdue University, Indianapolis (IUPUI). Before that, he was a Senior Research Scientist at eBay Research Labs, San Jose, CA. He received a Ph.D. degree in Computer Science from Rensselaer Polytechnic Institute (RPI) in 2009, and an MS degree in Computer Science from the University of Minnesota, Twin Cities in 2002. His research interest focuses on developing novel algorithms in data mining, data management, information retrieval, machine learning, social network analysis, and bioinformatics. One of his particular interests is to develop algorithms for sampling small substructures from large networks. He developed methods for: (1) sampling frequent subgraphs from a graph database, (2) sampling triangles and graphlets from a large network, and (3) Sampling interesting subgraph patterns using interactive feedbacks, all using Markov Chain Monte Carlo (MCMC) sampling algorithm. His doctoral dissertation won the ACM SIGKDD doctoral dissertation award in 2010. He is also a recepient of NSF CAREER award in 2012.

Jennifer Neville is an assistant professor at Purdue University with a joint appointment in the Departments of Computer Science and Statistics. She received her PhD from the University of Massachusetts Amherst in 2006. In 2012, she was awarded an NSF Career Award, in 2008 she was chosen by IEEE as one of ”AI’s 10 to watch”, and in 2007 was selected as a member of the DARPA Computer Science Study Group. Her research focuses on developing data mining and machine learning techniques for relational domains, including citation analysis, fraud detection, and social network analysis.

Nesreen Ahmed is a 5th year Ph.D. student working with Jennifer Neville in the Computer Science Department at Purdue University. Her Ph.D research is focused on statistical network sampling and network stream sampling. She has worked on research developing machine learning algorithms for time series forecasting, statistical predictive analysis of social media, and digital marketing. She has worked as a research intern at Adobe ATL labs and Intel Corporation, a research assistant at the data mining and computer modeling center of excellence in Egypt, and a teaching assistant at Cairo University..

KDD’13, August 11–14, 2013, Chicago, Illinois, USA.
ACM 978-1-4503-2174-7/13/08.