# Incompleteness in Networks: Biases, Skewed Results, and Some Solutions

Tina Eliassi-Rad
tina@eliassi.org
Northeastern University
Boston, MA

Rajmonda Caceres
rajmonda.caceres@ll.mit.edu
MIT Lincoln Laboratory
Lexington, MA

Timothy LaRock
larock.t@husky.neu.edu
Northeastern University
Boston, MA

## ABSTRACT

Most network analysis is conducted on existing incomplete samples of much larger complete, fully observed graphs. For example, many researchers obtain graphs from online data repositories without knowing how these graphs were collected. Thus, these graphs can be poor representations of the fully observed networks. More complete data would lead to more accurate analyses, but data acquisition can be at best costly and at worst error-prone. For example, think of an adversary that deliberately poisons the answer to a query. Given a query budget for identifying additional nodes and edges, how can one improve the observed graph sample so that it is a more accurate representation of the complete, fully observed network? How does the approach change if one is interested in learning the best function (e.g. node classifier) on the network for a down-stream task? This is a novel problem that is related to, but distinct from, topics such as graph sampling and crawling. Given the prevailing use of graph samples in the research literature, this problem is of considerable importance, even though it has been ignored. In this tutorial, we discuss latent biases in incomplete networks and present methods for enriching such networks through active probing of nodes and edges. We focus on active learning and sequential decision-making formulations of this problem (a.k.a. the *network discovery* problem). We present distinctions between learning to grow the network (a.k.a. active exploration) vs. learning the "best" function on the network (a.k.a. active learning). In addition, we will discuss issues surrounding adversarial machine learning when querying for more data to reduce incompleteness.

## CCS CONCEPTS

• **Mathematics of computing** → **Graph algorithms**; • **Information systems** → **Incomplete data**; • **Theory of computation** → **Sequential decision making**; • **Computing methodologies** → **Online learning settings**; **Active learning settings**.

## KEYWORDS

Incomplete networks, sequential decision making, active exploration, active learning, online learning.


*KDD '19, August 4–8, 2019, Anchorage, AK, USA*

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332276

**ACM Reference Format:**
Tina Eliassi-Rad, Rajmonda Caceres, and Timothy LaRock. 2019. Incompleteness in Networks: Biases, Skewed Results, and Some Solutions. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332276

## TUTORIAL DESCRIPTION

### Relevance

In data mining, we are interested in discovering useful patterns in large amounts of data. But, no matter how large (or "big") the data is, it is often incomplete because the underlying process generating the data is partially observed. This tutorial is relevant to the KDD community because it covers the important problem of latent biases in incomplete data and discusses models and algorithms for alleviating it. We focus on graph/network data because such data are ubiquitous in many applications.

### Outine

The outline of our 3-hour tutorial is as follows:

- Introduction and motivation (10 minutes)
- Graph crawling and sampling (15 minutes)
  - Graph crawling
  - Graph sampling
- Parameter estimation (20 minutes)
  - Estimating network parameters
- Explore vs. exploit methods (45 minutes)
  - Multi-armed bandits
  - Reinforcement learning: Sequential decision-making formulations
- Solutions and applications (60 minutes)
  - Enriching nodes and edges
  - Active learning vs. active exploration
  - Applications and downstream tasks
- Limits and adversaries (20 minutes)
  - Limits of learning in incomplete networks
  - Adversarial machine learning
- Wrap-up, future work, and Q&A (10 minutes)

### Previous Editions and Related Tutorials

An earlier version of this tutorial was presented at SIAM SDM 2016 and 2018:

- SIAM SDM'16: Problems with Incomplete Networks: Biases, Skewed Results, and Solutions (with Sucheta Soundarajan, Ali Pinar, and Brian Gallagher), Miami, FL, May 2016. See http://eliassi.org/sdm16tut.html.

- SIAM SDM'18: Problems with Partially Observed (Incomplete) Networks: Biases, Skewed Results, and Solutions (with Sucheta Soundarajan and Sahely Bhadra), San Diego, CA, May 2018. See http://eliassi.org/sdm18tut.html.

In SIAM SDM'18, we introduced new solutions to enhancing partially observed networks from the areas of multi-armed bandits and online reinforcement learning. The only overlap with the SIAM SDM'16 version was in the introduction, motivation, and part of the related work. The rest was new material.

Our KDD'19 tutorial focuses on active learning and sequential decision-making formulations of this *network-discovery* problem. We will also present distinctions between learning to grow the network (a.k.a. active exploration) vs. learning the "best" function on the network (a.k.a. active learning) for a down-stream task. In addition, we will discuss issues around adversarial attacks in the network-discovery setting.

The problem of enriching incomplete networks is related to graph sampling. However, unlike much of the work in graph sampling, when enriching an incomplete network, one is not attempting to generate a sample from scratch. Instead, one is studying how to enhance or improve an existing sample, without control over how that sample was generated. There have been five recent tutorials on graph sampling in prominent data mining venues:

- KDD'13: Network Sampling by Mohammad A. Hasan, Jennifer Neville, and Nesreen K. Ahmed. See http://www.kdd.org/kdd2013/accepted-tutorials.
- IEEE ICDM'13: Methods and Applications of Network Sampling by Mohammad A. Hasan, Nesreen Ahmed, and Jennifer Neville. See http://icdm2013.rutgers.edu/tutorials.
- KDD'14: Sampling for Big Data by Graham Cormode and Nick Duffield. See http://www.kdd.org/kdd2014/tutorials/t10_part1.pptx and http://www.kdd.org/kdd2014/tutorials/t10_part2.pptx.
- SIAM SDM'15: Methods and Applications of Network Sampling by Mohammad A. Hasan, Nesreen K. Ahmed, and Jennifer Neville.
  See http://cs.iupui.edu/~alhasan/SDM15-tutorial.html.
- KDD'15: VC-Dimension and Rademacher Averages: From Statistical Learning Theory to Sampling Algorithms by Matteo Riondato and Eli Upfal. See http://bigdata.cs.brown.edu/vctutorial/.

## Target Audience and Prerequisites

Our target audience includes researchers and practitioners in data mining, machine learning and data science, with an interest in incomplete (a.k.a. partially observed) networks and graphs. We are targeting people who are concerned about the latent biases in the "real-world" data being used in research and industry. We expect the audience to come away with an overview of the state-of-art in enriching incomplete networks and have a better understanding of the challenges in this area.

No assumption is made about familiarity with complex networks, graph mining, graph sampling, and incomplete data. A brief overview of them will be included in the tutorial.

## Presenters' Brief Biography

Tina Eliassi-Rad is an Associate Professor of Computer Science at Northeastern University in Boston, MA. She is also a core faculty member at Northeastern University's Network Science Institute. Prior to joining Northeastern, Tina was an Associate Professor of Computer Science at Rutgers University; and before that she was a Member of Technical Staff and Principal Investigator at Lawrence Livermore National Laboratory. Tina earned her Ph.D. in Computer Sciences (with a minor in Statistics) at the University of Wisconsin-Madison. Her research is rooted in data mining and machine learning; and spans theory, algorithms, and applications of data from networked representations of physical and social phenomena. She has over 80 peer-reviewed publications and has given over 180 invited talks and 13 tutorials. Her algorithms have been incorporated into systems used by the government and industry (e.g., IBM System G Graph Analytics) as well as open-source software (e.g., Stanford Network Analysis Project).

Rajmonda Caceres is a senior technical staff at MIT Lincoln Laboratory, in the Informatics and Decision Support Group. Her primary research interests are in the areas of network science, computational biology and data mining. Rajmonda earned her PhD degree in mathematics and computer science from the University of Illinois at Chicago in 2012. Her current work focuses on methods for learning robust representations of complex networks, as well as learning in resource-constrained environments.

Timothy LaRock is a third year doctoral student at Northeastern University's Network Science Program. His research falls at the intersection of network science and data mining. In particular, he develops models and algorithms that identify and understand sequential patterns and dependencies in network data. Tim completed his bachelors of science in Computer Science and Applied Mathematics with a minor in Philosophy at the State University of New York at Albany.

## ACKNOWLEDGMENTS

The contributors to this tutorial are: Timothy Sakharov (Northeastern University), Benjamin A. Miller (Northeastern University), Sucheta Soundarajan (Syracuse University), and Sahely Bhadra (Indian Institute of Technology (IIT) Palakkad. We thank them for their work on the material presented in this tutorial.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

This research was also sponsored by the Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here on.