Statistical Network Analysis and Data Science: Things we do well, and things we’re still working on

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu
Networks and Data Science

Network-based perspectives are now ubiquitous in data science. Google Scholar reports \( \sim 436,000 \) articles with ‘network’ in the title published since 1999.

Contributions have been from across the data sciences:
- Computer Science, Mathematics, Statistics
- Signal processing, Statistical Physics, Information Sciences
- Bioinformatics, Economics, Neuroscience, Sociology
- Digital Humanities
Networks Analysis

Network-based analysis traditionally a relatively small ‘field’ of study
Epidemic-like spread of interest in networks since mid/late-1990s
Arguably due to various factors, such as
  • Increasingly systems-level perspective in science, away from reductionism;
  • Flood of high-throughput data;
  • Globalization, the Internet, etc.
Where is Statistics Amid All of this Work?

Everywhere!

Statistical aspects of network analysis include problems involving

- sampling and design
- description and visualization
- modeling and inference
- prediction

for data both of and on networks.

Much of this work occurs in domain-specific areas; a nontrivial amount of it is general.

This is an actively-evolving field . . . much of network analysis we can do well, while for a good deal of the rest we still have a ways to go!
Statistical Challenges Working with Network Data

Broadly speaking, the primary statistical challenge(s) in most network problems comes from nontrivial interplay between

- relational/dependent nature of the data;
- lack of (traditional) geometry; and
- ‘big data’.

**Question:** Despite a fairly well-developed ability to deal successfully with many classes of problems in many contexts and domain areas, how well do we truly understand the implications of network data (e.g., as opposed to IID, temporal, or spatial data) on statistics at a foundational level?

**My Answer:** Somewhat . . . but there’s still a long way to go!
Illustrations

Through the use of *three vignettes* of current/ongoing work from my research group, I will illustrate

- what I mean by ‘foundations’;
- the novelty lent to foundational topics when intersected with complex networks; and
- some of the resulting challenges and open problems.

**Vignette Topics**

1. Adjusting for Bias in Network Sampling
2. Propagation of Uncertainty To Network Summary Statistics
3. Inference for Large Samples of Network Data Objects
Figure: Estimating degree distributions of communities from Friendster, Orkut and Livejournal. Blue dots represent the true degree distributions, black dots represent the sample degree distributions, red dots represent the estimated degree distributions. Sampling rate=30%. Dots which correspond to a density $< 10^{-4}$ are eliminated from the plot.

BU Data Science Day, Jan 2016
Examples particularly prevalent in biology (e.g., gene regulatory networks, protein-protein interaction networks, and neural functional connectivity networks), but some noise likely present in most network applications.
‘Statistics 101’ for Collections of Networks

Current state of the art (aka ‘mass univariate’) uses edge-wise testing (presence/absence) and multiple correction.

(A) Mass-univariate analysis (Sex)  (B) Mass-univariate analysis (Age)

Comparison: Mass-Univariate vs Multivariate

- Both methods detect differences in mean networks across gender and age, when using the full 1000 connectomes; but . . .

- Only our multivariate method detects those differences at small sample sizes (i.e., relevant to single labs).
Collaborators and Support

Supported in part by AFOSR, NIH, NSF, and ONR.

BU Data Science Day, Jan 2016