

Statistical Analysis of Network Data

A Brief Overview

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu



Focus of this Talk

In this talk I will present a brief overview of the foundations common to the statistical analysis of network data across the disciplines, from a statistical perspective.

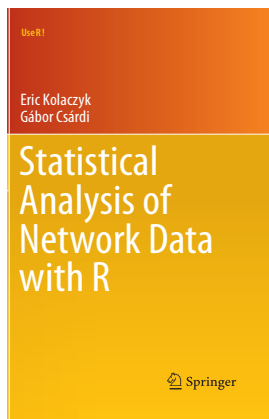
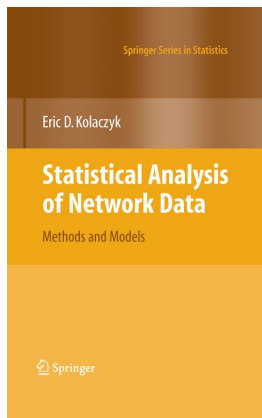
Approach will be that of a high-level, whirlwind overview of the topics of

- network summary and visualization
- network sampling
- network modeling and inference, and
- network processes.

Concepts will be illustrated drawing on examples from bioinformatics, computer network traffic analysis, neuroscience, and social networks.

Resources

Organization and presentation of material in this quick talk will largely parallel that in



Our Focus . . .

The statistical analysis of *network data*

i.e., analysis of measurements either of or from a system conceptualized as a network.

Challenges:

- relational aspect to the data;
- complex statistical dependencies (often the focus!);
- high-dimensional and often massive in quantity.

Outline

- 1 Introduction
- 2 Network Mapping**
- 3 Network Characterization
- 4 Network Sampling
- 5 Network Modeling
- 6 Network Inference
- 7 Wrap-Up

Descriptive Statistics for Networks

First two topics go together naturally, i.e.,

- network mapping
- characterization of network graphs

May seem 'soft' ... but it's important!

- This is basically descriptive statistics for networks.
- Probably constitutes at least 2/3 of the work done in this area.

Note: It's sufficiently different from standard descriptive statistics that it's something unto itself.

Network Mapping

What is 'network mapping'?

Production of a network-based visualization of a complex system.

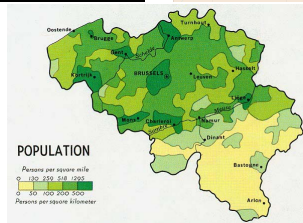
What is 'the' network?

- Network as a 'system' of interest;
- Network as a graph representing the system;
- Network as a visual object.

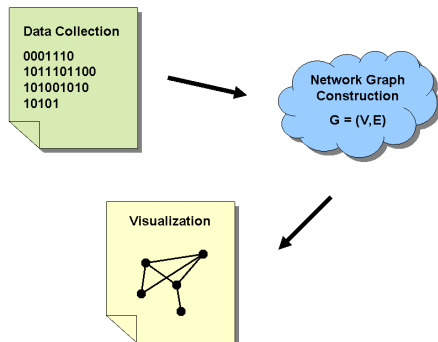
Analogue: Geography and the production of cartographic maps.

Example: Mapping Belgium

Which of these is 'the' Belgium?



Three Stages of Network Mapping



Question: What is the impact on network visualization of differential privacy applied at these various stages?

Outline

- 1 Introduction
- 2 Network Mapping
- 3 Network Characterization**
- 4 Network Sampling
- 5 Network Modeling
- 6 Network Inference
- 7 Wrap-Up

Characterization of Network Graphs: Intro

Given a network graph representation of a system (i.e., perhaps a result of network mapping), often *questions of interest* can be phrased in terms of *structural properties* of the graph.

- *social dynamics* can be connected to *patterns of edges among vertex triples*;
- routes for *movement of information* can be approximated by *shortest paths between vertices*;
- *'importance' of vertices* can be captured through so-called *centrality measures*;
- natural *groups/communities* of vertices can be approached through *graph partitioning*.

Characterization Intro (cont.)

Structural analysis of network graphs \approx descriptive analysis; this is a standard first (and sometimes only!) step in statistical analysis of networks.

Main contributors of tools are

- social network analysis,
- mathematics & computer science,
- statistical physics

Many tools out there ... two rough classes include

- characterization of vertices/edges, and
- characterization of network cohesion.

Characterization of Vertices/Edges

Examples include

- Degree distribution
- Vertex/edge centrality
- Role/positional analysis

We'll look at the *vertex centrality* as an example.

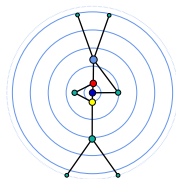
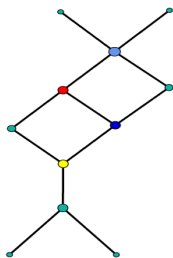
Centrality: Motivation

Many questions related to 'importance' of vertices.

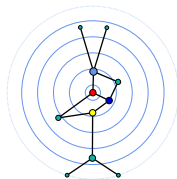
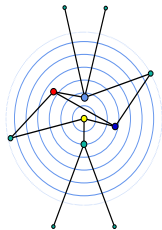
- Which actors hold the 'reins of power'?
- How authoritative is a WWW page considered by peers?
- The deletions of which genes is more likely to be lethal?
- How critical to traffic flow is a given Internet router?

Researchers have sought to capture the notion of vertex importance through so-called centrality measures.

Centrality: An Illustration



Clockwise from top left:
 (i) toy graph, with (ii) closeness, (iii) betweenness, and (iv) eigenvector centralities.



Example and figures
 courtesy of Ulrik Brandes.

Network Cohesion: Motivation

Many questions involve scales coarser than just individual vertices/edges.
More properly considered questions regarding 'cohesion' of network.

- Do friends of actors tend to be friends themselves?
- Which proteins are most similar to each other?
- Does the WWW tend to separate according to page content?
- What proportion of the Internet is constituted by the 'backbone'?

These questions go beyond individual vertices/edges.

Network Cohesion: Various Notions!

Various notions of 'cohesion'.

- density
- clustering
- connectivity
- flow
- partitioning
- ... and more ...

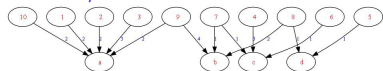
Illustration: Detecting Malicious Internet Sources

Ding *et al.*^a use the idea of **cut-vertices** to detect Internet IP addresses associated with malicious behavior.

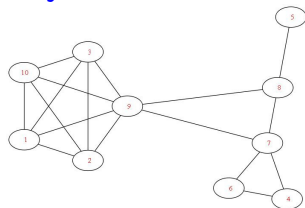
Corresponds to a type of (anti)social behavior.

^aDing, Q., Katenka, N., Barford, P., Kolaczyk, E.D., and Crovella, M. (2012). Intrusion as (Anti)social Communication: Characterization and Detection. *Proceedings of the 2012 ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Source/Destination Network



Projected Source Network



Outline

- 1 Introduction
- 2 Network Mapping
- 3 Network Characterization
- 4 Network Sampling**
- 5 Network Modeling
- 6 Network Inference
- 7 Wrap-Up

Network Sampling: Point of Departure ...

Common *modus operandi* in network analysis:

- System of elements and their interactions is of interest.
- Collect elements and relations among elements.
- Represent the collected data via a network.
- Characterize properties of the network.

Sounds good ... right?

Interpretation: Two Scenarios

With respect to what frame of reference are the network characteristics interpreted?

- 1 The collected network data are themselves the primary object of interest.
- 2 The collected network data are interesting primarily as representative of an underlying 'true' network.

The distinction is important!

Under Scenario 2, statistical sampling theory becomes relevant . . . but is not trivial.

Common Network Sampling Designs

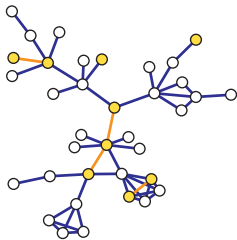
Viewed from the perspective of classical statistical sampling theory, the network sampling design is important.

Examples include

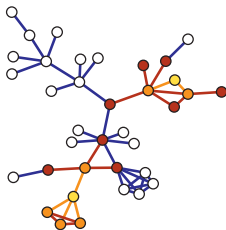
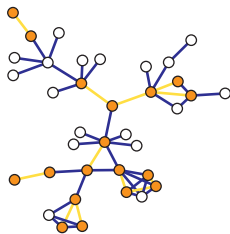
- Induced Subgraph Sampling
- Incident Subgraph Sampling
- Snowball Sampling
- Link Tracing

Common Network Sampling Designs (cont.)

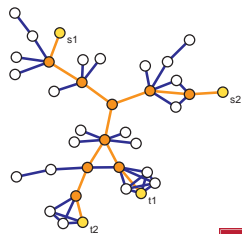
Induced Subgraph Sampling



Incident Subgraph Sampling



Snowball Sampling



Traceroute Sampling

Caveat emptor . . .

Completely ignoring sampling issues is equivalent to using 'plug-in' estimators.

The resulting bias(es) can be both substantial and unpredictable!

	BA	PPI	AS	arXiv
Degree Exponent	↑ ↑ ↓	↑ ↑ =	= = ↓	↑ ↑ ↓
Average Path Length	↑ ↑ =	↑ ↑ ↓	↑ ↑ ↓	↑ ↑ ↓
Betweenness	↑ ↑ ↓	↑ ↑ ↓	↑ ↑ ↓	= = =
Assortativity	= = ↓	= = ↓	= = ↓	= = ↓
Clustering Coefficient	= = ↑	↑ ↓ ↑	↓ ↓ ↑	↓ ↓ ↓

Lee *et al* (2006): Entries indicate direction of bias for induced subgraph (red), incident subgraph (green), and snowball (blue) sampling.

Accounting for Sampling Design

Accounting for sampling design can be non-trivial.

Classical work goes back to the 1970's (at least), with contributions of Frank and colleagues, based mainly on Horvitz-Thompson theory.

More recent resurgence of interest, across communities, has led to additional studies using both classical and modern tools.

See Kolaczyk (2009), Chapter 5.

Illustration: Estimation of Degree Distribution

Under a variety of sampling designs, the following holds:

$$E[\mathbf{N}^*] = P\mathbf{N} \quad , \quad (1)$$

where

- $\mathbf{N} = (N_0, N_1, \dots, N_M)$: the true degree vector, for N_i : the number of vertices with degree i in the original graph
- $\mathbf{N}^* = (N_0^*, N_1^*, \dots, N_M^*)$: the observed degree vector, for N_i^* : the number of vertices with degree i in the sampled graph
- P is an $M + 1$ by $M + 1$ matrix operator, where $M = \text{maximum degree in the original graph}$

Estimating Degree Distribution: An Inverse Problem

Ove Frank (1978) proposed solving for the degree distribution by an unbiased estimator of N , defined as

$$\hat{\mathbf{N}}_{\text{naive}} = P^{-1}\mathbf{N}^* . \quad (2)$$

There are two problems with this simple solution:

- 1 The matrix P is typically not invertible in practice.
- 2 The non-negativity of the solution is not guaranteed.

Does It Really Matter? Yes!

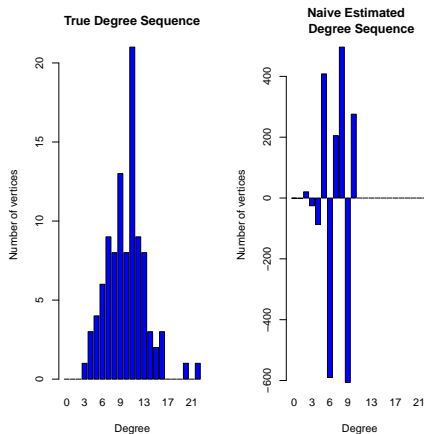


Figure : Left: ER graph with 100 vertices and 500 edges. Right: Naive estimate of degree distribution, according to equation (2). Data drawn according to induced subgraph sampling with sampling rate $p = 60\%$.

A Modern Variant: Constrained, Penalized WLS

We have recently proposed¹ a penalized weighted least squares with additional constraints.

$$\begin{aligned}
 & \underset{\mathbf{N}}{\text{minimize}} && (\mathbf{P}\mathbf{N} - \mathbf{N}^*)^T \mathbf{C}^{-1} (\mathbf{P}\mathbf{N} - \mathbf{N}^*) + \lambda \cdot \text{pen}(\mathbf{N}) \\
 & \text{subject to} && N_i \geq 0, \quad i = 0, 1, \dots, M \\
 & && \sum_{i=0}^M N_i = n_V,
 \end{aligned} \tag{3}$$

where

- $\mathbf{C} = \text{Cov}(\mathbf{N}^*)$,
- $\text{pen}(\mathbf{N})$ is a penalty on the complexity of \mathbf{N} ,
- λ is a smoothing parameter, and
- n_V is the total number of vertices of the true graph.

¹Zhang, Y., Kolaczyk, E.D., and Spencer, B.D. (2013). Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks. Under review by *Annals of Applied Statistics*. arxiv-1305.4977

Application to Online Social Networks

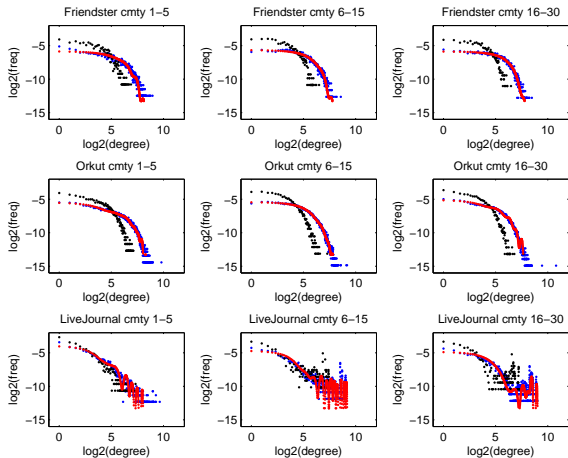


Figure : Estimating degree distributions of communities from Friendster, Orkut and Livejournal. Blue dots represent the true degree distributions, black dots represent the sample degree distributions, red dots represent the estimated degree distributions. Sampling rate=30%. Dots which correspond to a density $< 10^{-4}$ are eliminated from the plot.

Estimating Approximate Epidemic Thresholds: Friendster

- Moments of degree distributions can be used to obtain bounds of the network's epidemic threshold τ_c .
- An approximate threshold is given by the inverse of the largest eigenvalue λ_1 of the adjacency matrix (Mieghem, Omic, & Kooij '09).
- Simple bounds for λ_1 are

$$M_1 \leq \sqrt{M_2} \leq \lambda_1 \leq (2N_e)^{1/2} \quad (4)$$

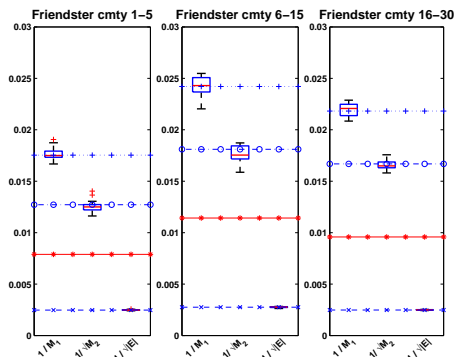


Figure : Estimated bounds for epidemic threshold in Friendster, based on 20 samples. Four horizontal lines are the true values for $\frac{1}{M_1}$, $\frac{1}{\sqrt{M_2}}$, λ_1 and $\frac{1}{\sqrt{2N_e}}$ from top to bottom.

Outline

- 1 Introduction
- 2 Network Mapping
- 3 Network Characterization
- 4 Network Sampling
- 5 Network Modeling**
- 6 Network Inference
- 7 Wrap-Up

Two Scenarios

We will look at two complementary scenarios²:

- 1 we observe a network G (and possibly attributes \mathbf{X}) and we wish to model G (and \mathbf{X});
- 2 we observe the network G , but lack some or all of the attributes \mathbf{X} , and we wish to infer \mathbf{X} .

These are, of course, caricatures. Reality can be more complex!

²A third option, that we observe attributes \mathbf{X} , but lack some or all of the network G , and we wish to infer G , is usually called network topology inference, which we'll talk about last.

High Standards

Statisticians demand a great deal of their modeling:

- 1 theoretically plausible
- 2 estimable from data
- 3 computationally feasible estimation strategies
- 4 quantification of uncertainty in estimates (e.g., confidence intervals)
- 5 assessment of goodness-of-fit
- 6 understanding of the statistical properties of the overall procedure

Still have a long way to go in the context of networks!

Classes of Statistical Network Models

Roughly speaking, there are network-based versions of three canonical classes of statistical models:

- 1 regression models (i.e., ERGMs)
- 2 latent variable models (i.e., latent network models)
- 3 mixture models³ (i.e., stochastic blocks models)

³These may be viewed as a special case of latent variable models.

Statistical Network Models: Progress and Challenges

This is one of the most active areas of research in statistics and networks.

Most work in ERGMs and SBMs.

A few high-level comments:

- ERGMs have the largest body of work associated with them ...
- ... but they also arguably have the greatest number of problems (i.e., degeneracy, instability, problems under sampling, as well as the least supporting formal theory).
- SBMs arguably have the most extensive theoretical development (i.e., fully general formulation, consistency and asymptotic normality of parameter estimates, etc.) ...
- ... but they can be still too simple for many modeling situations, and lack the link to regression possessed by ERGMs.

Processes on Network Graphs

So far we have focused on network graphs, as representations of *network systems of elements and their interactions*.

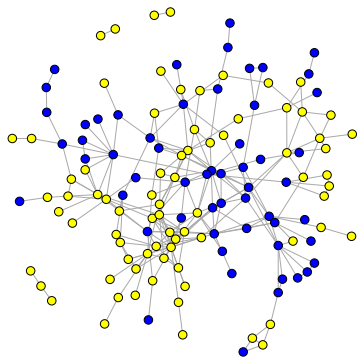
But often it is *some quantity associated with the elements* that is of most interest, rather than the network *per se*.

Nevertheless, such quantities may be influenced by the interactions among elements.

Examples:

- Behaviors and beliefs influenced by social interactions.
- Functional role of proteins influenced by their sequence similarity.
- Computer infections by viruses may be affected by 'proximity' to infected computers.

Illustration: Predicting Signaling in Yeast



- Baker's yeast (i.e., *S. cerevisiae*)
- All proteins known to participate in *cell communication* and their interactions
- *Question:* Is knowledge of the function of a protein's neighbors predictive of that protein's function?

In fact . . . yes!

A Simple Approach: Nearest-Neighbor Prediction

A simple predictive algorithm uses nearest neighbor principles.

Let

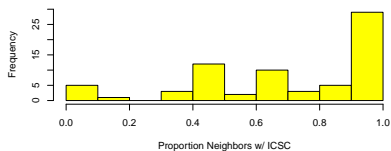
$$X_i = \begin{cases} 1, & \text{if corporate} \\ 0, & \text{if litigation} \end{cases}$$

Compare

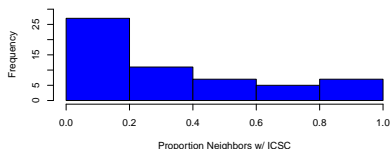
$$\frac{\sum_{j \in \mathcal{N}_i} X_j}{|\mathcal{N}_i|}$$

to a threshold.

Egos w/ ICSC



Egos w/out ICSC



Modeling Static Network-Indexed Processes

The nearest-neighbor algorithm (also sometimes called ‘guilt-by-association’), although seemingly informal, can be quite competitive with more formal, model-based methods.

Various models have been proposed for static network-indexed processes.

Two commonly used classes/paradigms:

- Markov random field (MRF) models
 - ⇒ Extends ideas from spatial/lattice modeling.
- Kernel-learning regression models
 - ⇒ Key innovation is construction of graph kernels

Outline

- 1 Introduction
- 2 Network Mapping
- 3 Network Characterization
- 4 Network Sampling
- 5 Network Modeling
- 6 Network Inference**
- 7 Wrap-Up

Network Topology Inference

Recall our characterization of *network mapping*, as a three-stage process involving

- 1 Collecting relational data
- 2 Constructing a network graph representation
- 3 Producing a visualization of that graph

Network topology inference is the formalization of Step 2 as a task in statistical inference.

Note: Casting the task this way also allows us to formalize the question of validation.

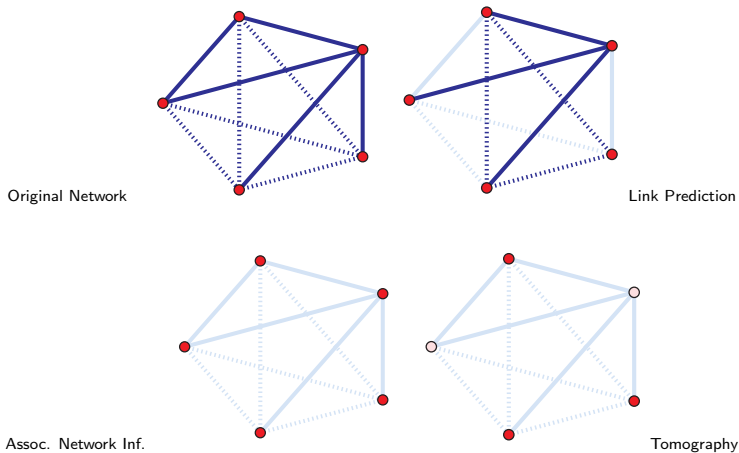
Network Topology Inference (cont.)

There are *many* variants of this problem!

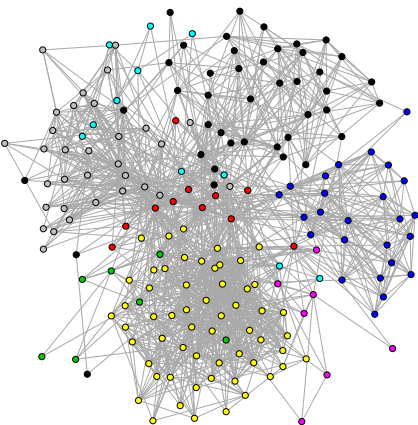
Three general, and fairly broadly applicable, versions are

- Link prediction
- Association network inference
- Tomographic network inference

Schematic Comparison of Inference Problems



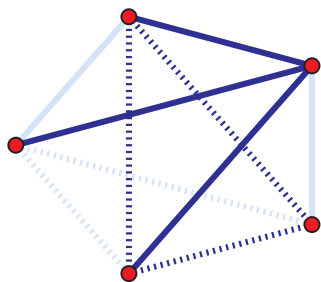
Link Prediction: Examples



Examples of link prediction include

- predicting new hyperlinks in the WWW
- assessing the reliability of declared protein interactions
- predicting international relations between countries

Link Prediction: Problem & Solutions



Goal is to predict the edge status' \mathbf{Y}^{miss} for all potential edges with missing (i.e., unknown) status, based on

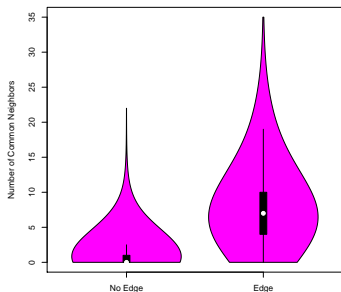
- observed status' \mathbf{Y}^{obs} , and
- any other auxiliary information.

Two main classes of methods proposed in the literature to date.

- Scoring methods
- Classification methods

See Kolaczyk (2009), Chapter 7.2.

Link Prediction: Illustration



Scoring methods come in all shapes and sizes.

Number of common neighbors a basic example.

Sufficient here (in a network of blogs) to obtain substantial discrimination (e.g., AUC of ~ 0.9 in leave-one-out prediction).

Outline

- 1 Introduction
- 2 Network Mapping
- 3 Network Characterization
- 4 Network Sampling
- 5 Network Modeling
- 6 Network Inference
- 7 Wrap-Up**

Wrapping Up

Lots of additional topics we have not touched upon:

- Dynamic networks
- Weighted networks
- Community detection
- Etc.

Wrapping Up (cont.)

Some of the things my group is working on currently include:

- Uncertainty in graph summary statistics under noisy conditions.
- Asymptotics for parameter estimation in network models.
- Estimation of degree distributions from sampled data.
- Bayesian latent-factor network perturbation models.
- Multi-attribute networks.