Tutorial: Statistical Analysis of Network Data

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu

BOSTON UNIVERSITY

Goals of this Tutorial

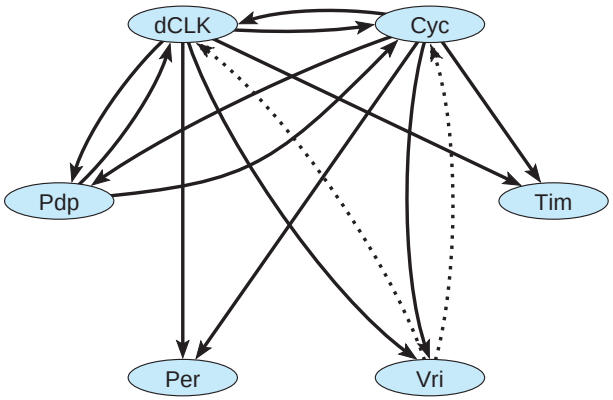
The SAMSI Program on Complex Networks seeks to bring together a critical mass of researchers in '*network science*', with emphasis on expertise in the statistical and applied mathematical sciences.

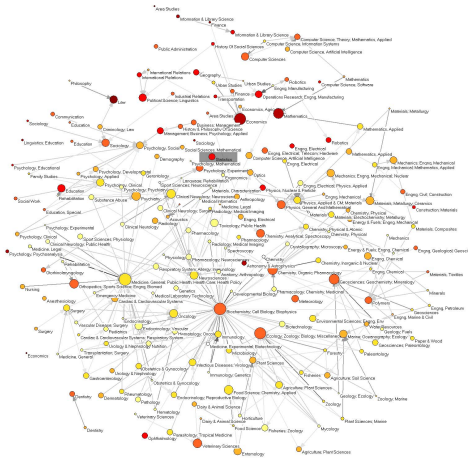
This tutorial will focus on statistical aspects of analyzing network data.

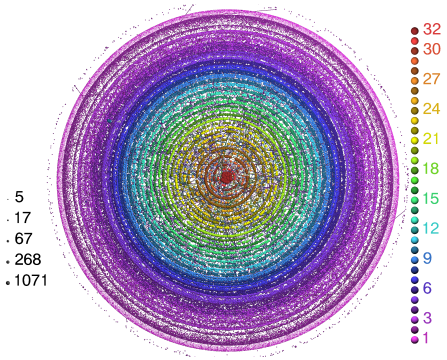
Goal is to present an overview, with an eye towards illustrating the types of statistical problems encountered in this area ... as well as some of the solutions!

BOSTON UNIVERSITY

Why Networks?







- Relatively small 'field' of study until past 10-15 years
- Epidemic-like spread of interest in networks since mid-90s
- 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.

BOSTON UNIVERSITY

What Do We Mean by 'Network'?

Definition (OED): *A collection of inter-connected things.*

Caveat emptor: The term 'network' is used in the literature to mean various things.

Two extremes are

1. a system of inter-connected things
2. a graph representing such a system[1]

Often is not even clear what is meant when an author refers to 'the' network!

[1] I'll use the slightly redundant term 'network graph'.

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.

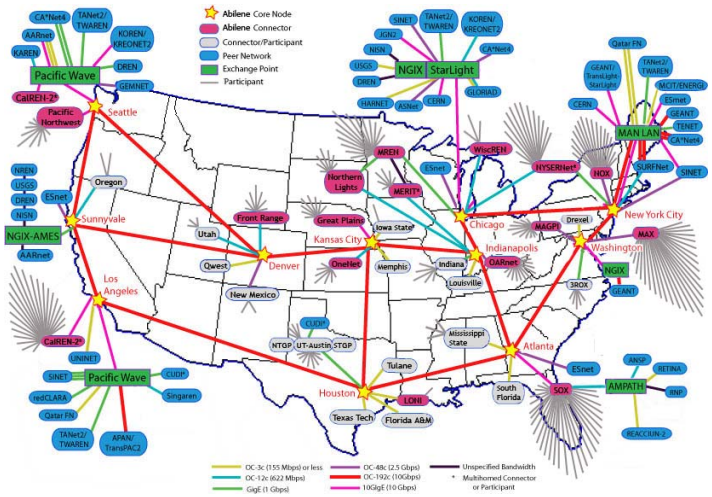
Examples of Networks

Network-based perspective has been brought to bear on problems from across the sciences, humanities, and arts.

To set some context, let's look quickly at examples from four general areas:

- Technological
- Biological
- Social
- Informational

Technological Nets



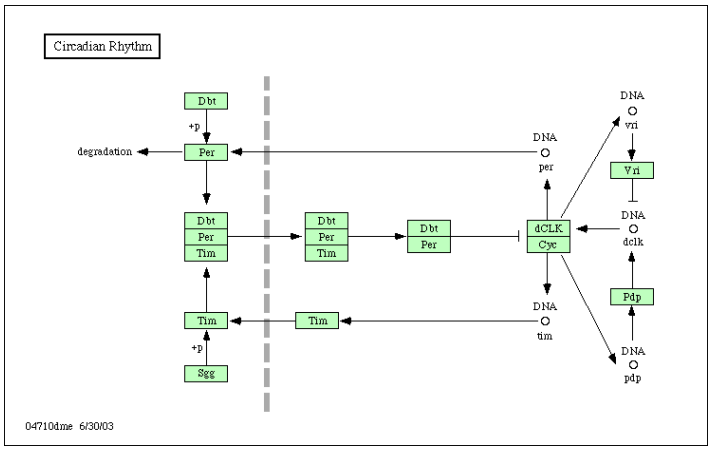
Includes communication, transportation, energy, and sensor networks.

The Internet: Questions

- What does the Internet look like today?
- What will Géant traffic through Belgium look like tomorrow?
- How can I detect anomalous traffic patterns?

BOSTON UNIVERSITY

Biological Nets



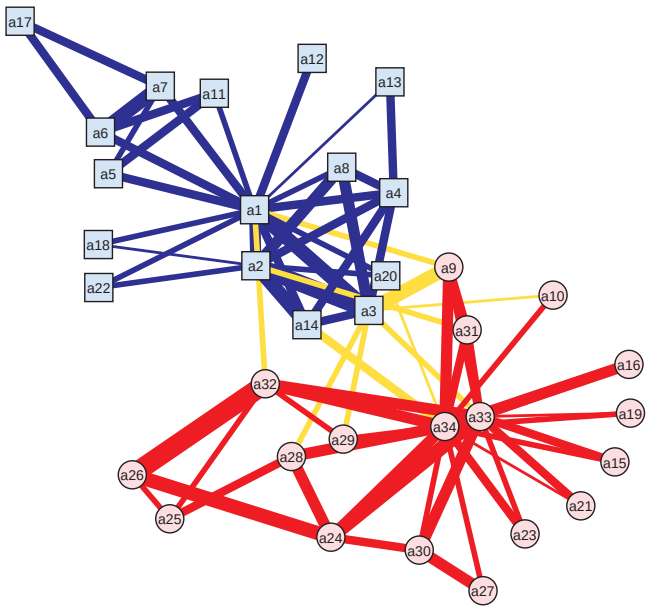
Includes networks of neurons, gene interactions, metabolic paths, predator/prey relationships, and protein interactions.

Questions

- Are certain patterns of interaction among genes more common than expected?
- Which regions of the brain 'communicate' during a given task?

BOSTON UNIVERSITY

Social Nets



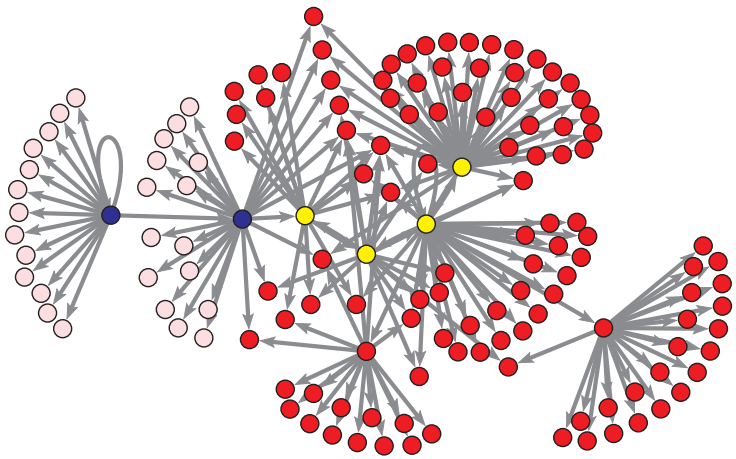
Examples include friendship networks, corporate networks, email networks, and networks of international relations.

Questions include

- Who is friends with whom?
- Which 'actors' are the 'power brokers'?
- What social groups are present?

BOSTON UNIVERSITY

Information Nets



Examples include the WWW, Twitter, and peer-to-peer networks (e.g., Gnutela, Bit Torrent).

Questions

- What does this network look like?
- How does information 'flow' on this network?

Statistics and Network Analysis

The (emerging?) field of 'network science' appears, at present, to be very *horizontal*.

Lots of 'players' ... uneven depth across the 'field' ... mixed levels of communication/cross-fertilization.

Note: Statisticians arguably a minority in this area!

But from a statistical perspective, there are certain *canonical* tasks and problems faced in the questions addressed across the different areas of specialty.

Better vertical depth can achieved in this area by viewing problems – and pursuing solutions – from this perspective.

BOSTON UNIVERSITY

Plan for the Remainder of this Tutorial

We'll look at some illustrative problems relating to

- network mapping
- network characterization
- network sampling
- network inference
- network processes

A more leisurely trip through this material will be taken during the first three weeks of the accompanying SAMSI course[2].

[2] See also the book,

Kolaczyk, E.D. (2009). Statistical Analysis of Network Data: Methods and Models. Springer, New York.

which the SAMSI course will follow for the first few weeks.

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.

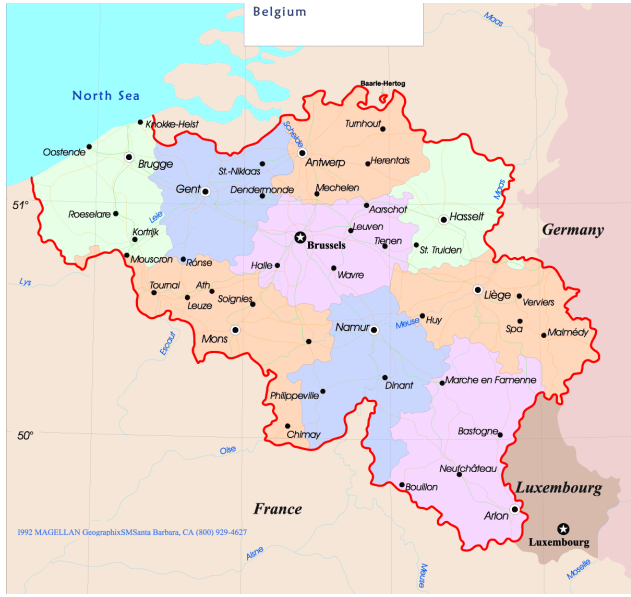
What is 'the' network?

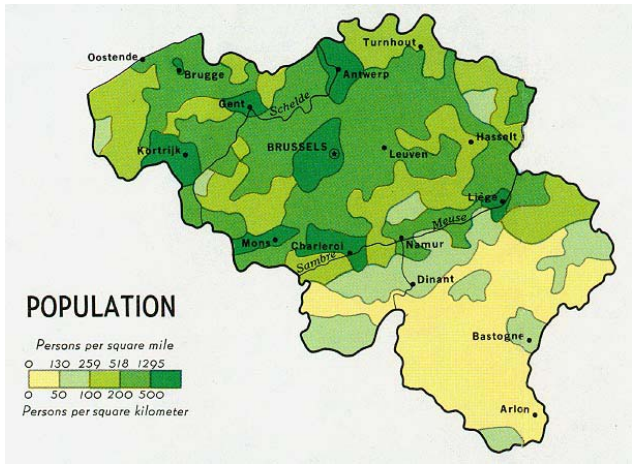
- Network as a 'system' of interest;
- Network as a graph representing the system;
- Network as a visual object.

Analogue: Geography and the production of cartographic maps.

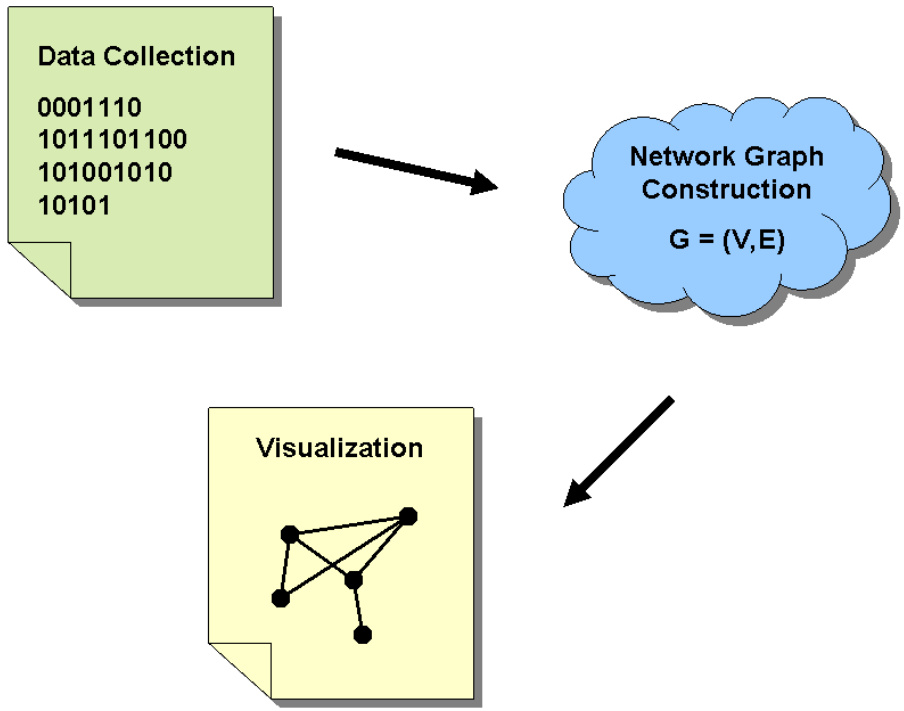
Example: Mapping Belgium

Which of these is 'the' Belgium?





Three Stages of Network Mapping

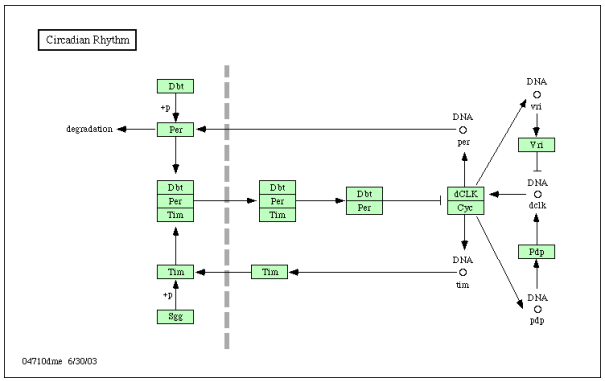


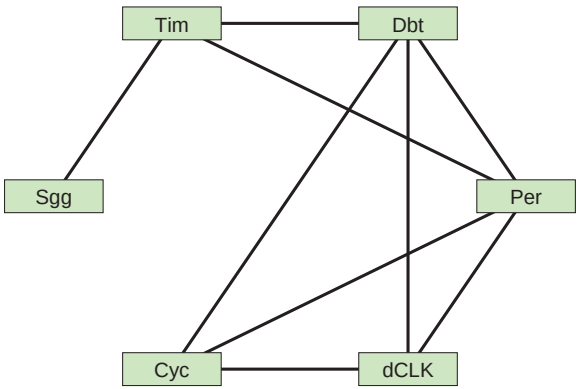
Continuing our geography analogue . . . a fourth stage might be 'validation'.

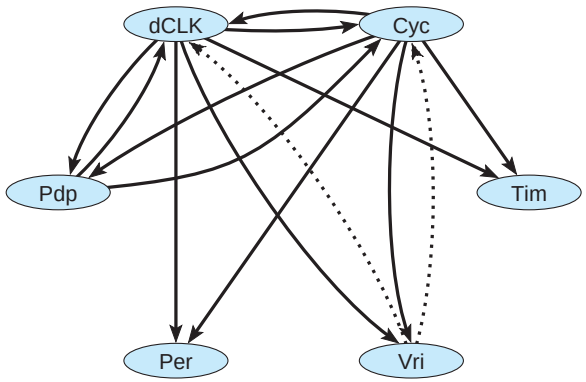
Stage 1: Collecting Relational Network Data

Begin with measurements on system 'elements' and 'relations'.

Note that choice of 'elements' and 'relations' can produce very different representations of same system.







Standard Statistical Issues Present Too!

- Type of measurements (e.g., cont., binary, etc.) can influence quality of information they contain on underlying 'relation'.
- Full or partial view of the system?
  (Analogues in spatial statistics ...)
- Sampling, missingness, etc.

Stage 2: Constructing Network Graphs

Sometimes measurements are direct declaration of edge/non-edge status.

More commonly, edges dictated after processing measurements

- comparison of 'similarity' metric to threshold
- voting among multiple views (e.g., router I-net)

Frequently *ad hoc* . . .
. . . sometimes formal (e.g., network inference).

Even with direct and error free observation of edges, decisions may be made to thin edges, adjust topology to match additional variables, etc.

BOSTON UNIVERSITY

Stage 3: Visualization

Goal is to embed a combinatorial object $G = (V, E)$ into two- or three-dimensional Euclidean space.

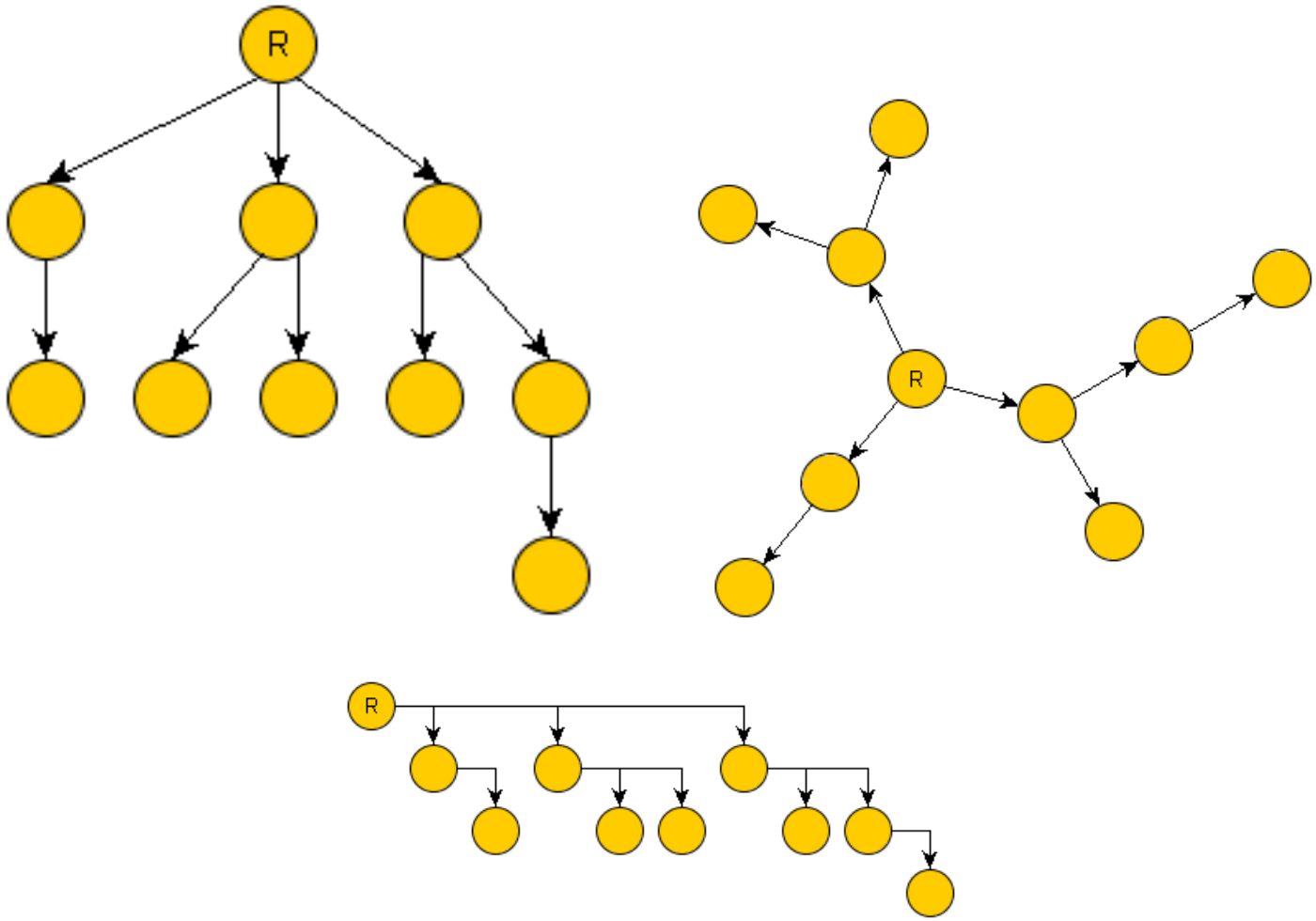
Non-unique ... not even well-defined!

Common to better define / constrain this problem by incorporating

- conventions (e.g., straight line segs)
- aesthetics (e.g., minimal edge crossing)
- constraints (e.g., on relative placement of vertices, subgraphs, etc.)

Layout ... Does it Matter?

Yes!



Layered, circular, and h-v layouts of the same tree.

Where are we at?

- network mapping
- network characterization
- network sampling
- network inference
- network processes

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: What Does It Mean?

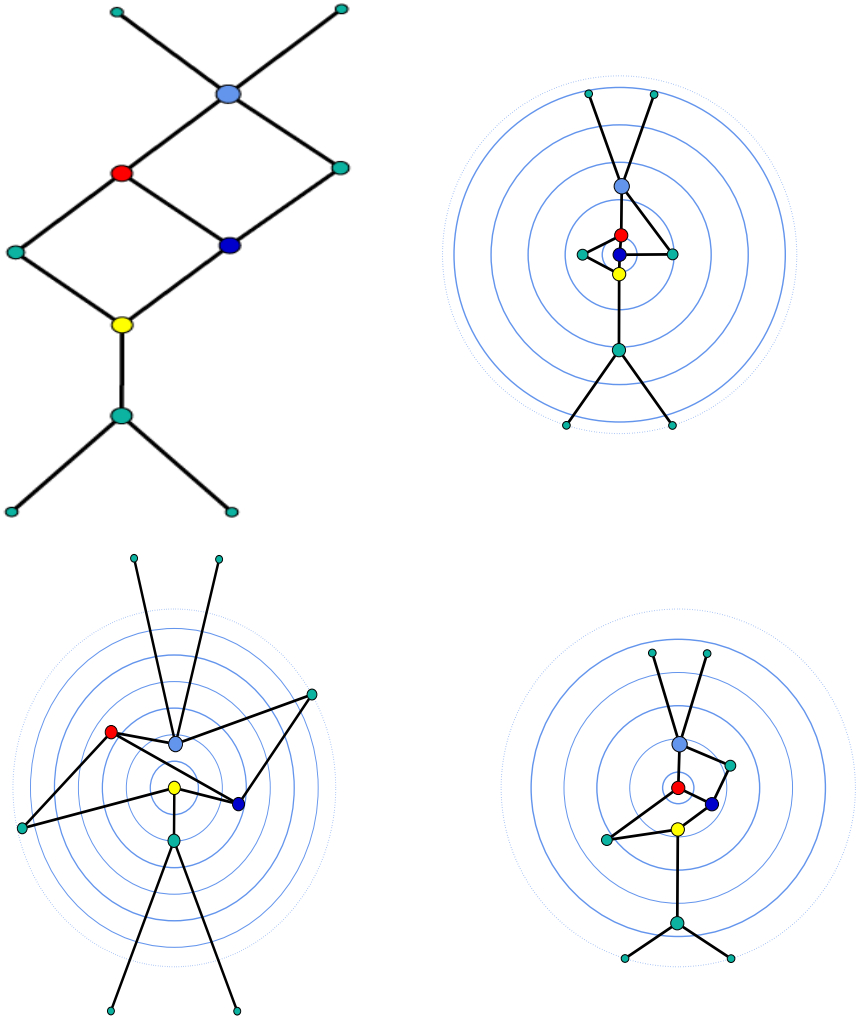
A *vast* number of measures have been introduced.

Useful, on the one hand, but indicative of something problematic, on the other hand!

> *There is certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is little agreement on the proper procedure for its measurement.*
>
> *L. Freeman, 1979*

Arguably still true today!

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 more 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 . . .

We'll look quickly at just one example: components.

Components

Not uncommon in practice that a graph be unconnected!

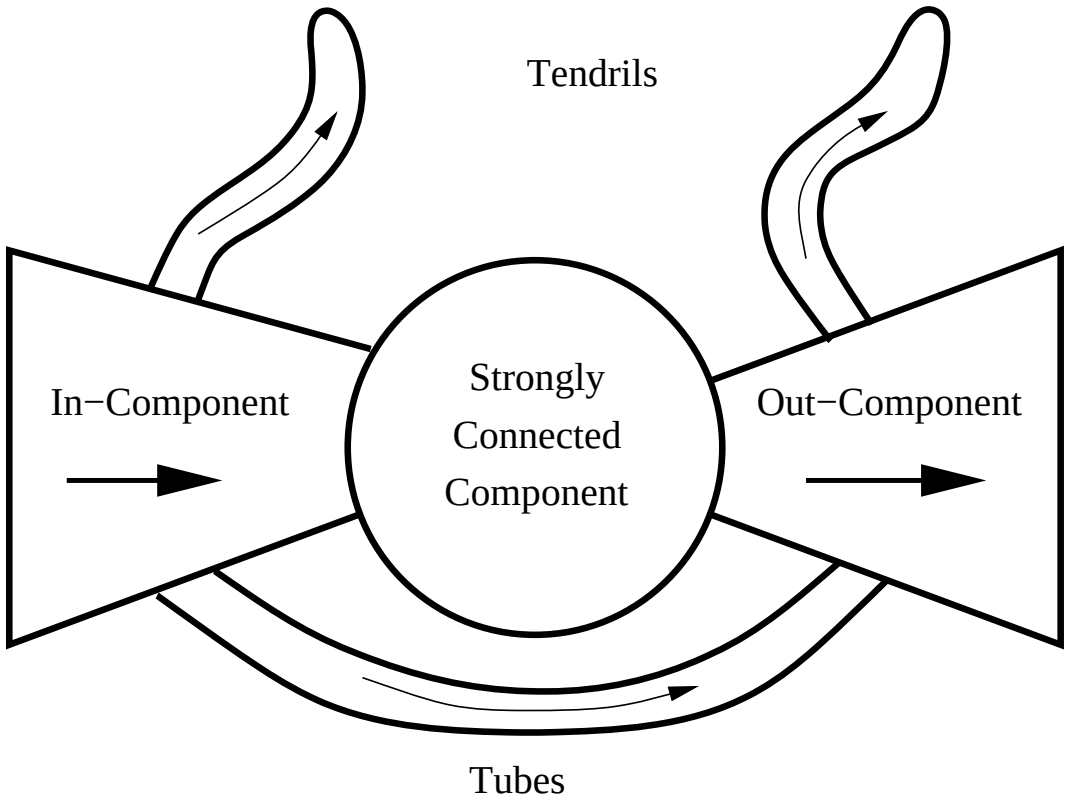
A *(connected) component* of a graph G is a maximally connected sub-graph.

Common to decompose graph into components. Often find this results in

- giant component
- smaller components
- isolates

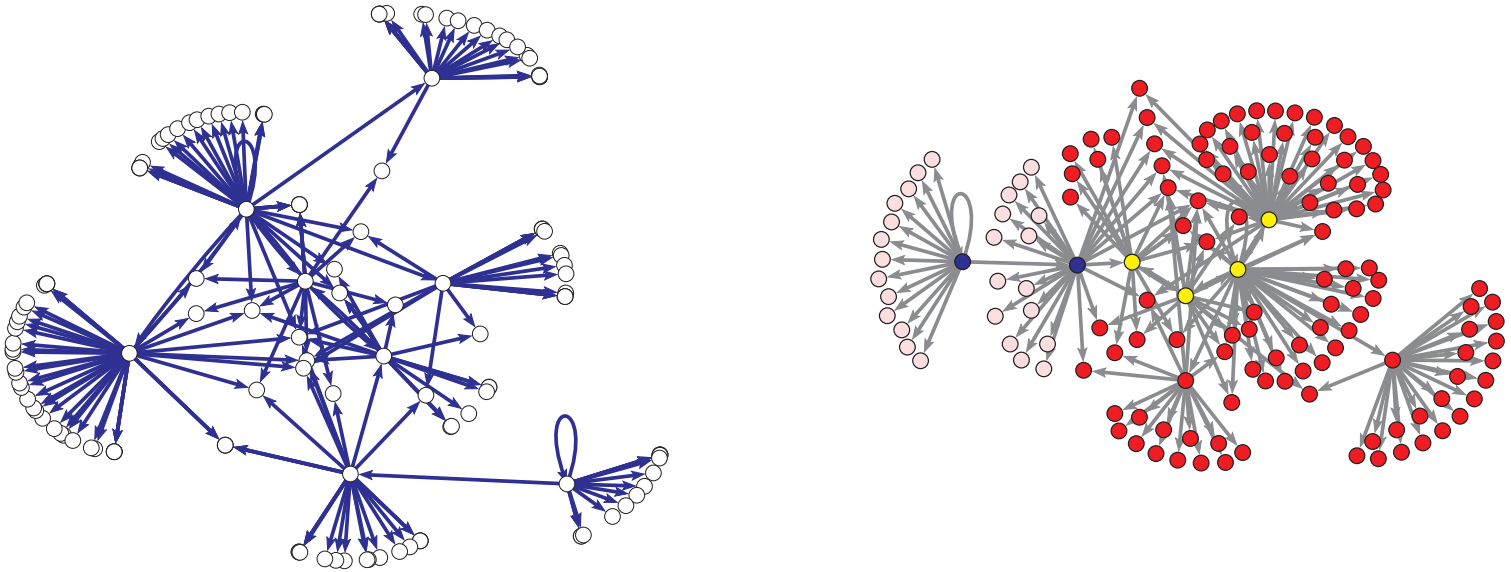
Frequently, reported analyses are for the *giant component*.

Components in Directed Graphs



Due to Broder *et al.* '00.

Example: AIDS Blog Network



Left: Original network. Right: Network with vertices annotated by component membership i.e.,

- strongly connected component (yellow)
- in-component (blue)
- out-component (red)
- tendrils (pink)

Where are we at?

- network mapping
- network characterization
- network sampling
- network inference
- network processes

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.

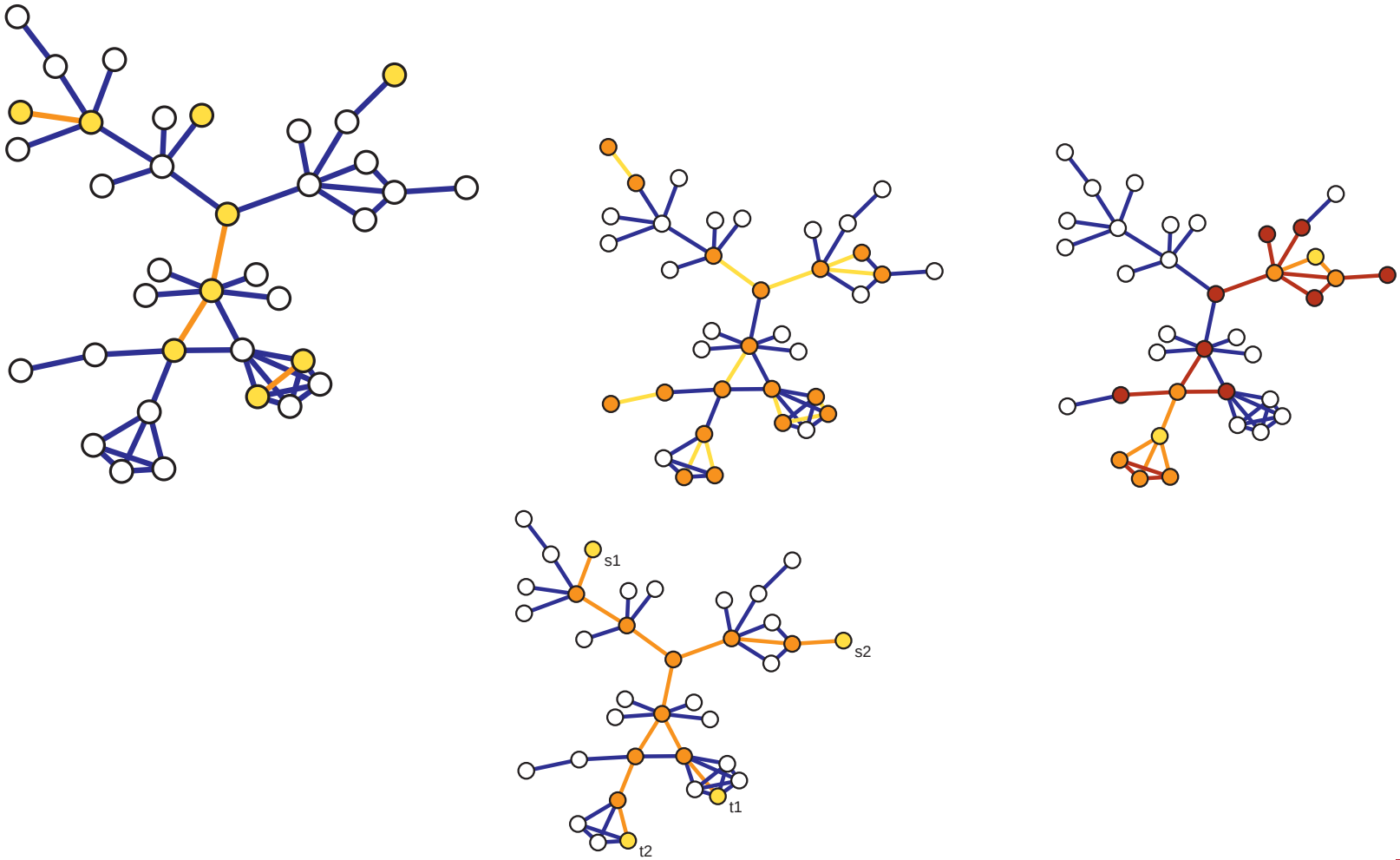
Common Network Sampling Designs

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.)



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 vertex (red), edge (green), and snowball (blue) sampling.

BOSTON UNIVERSITY

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 the talk by Matt Salganik Monday.

Where are we at?

- network mapping
- network characterization
- network sampling
- network inference
- network processes

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.

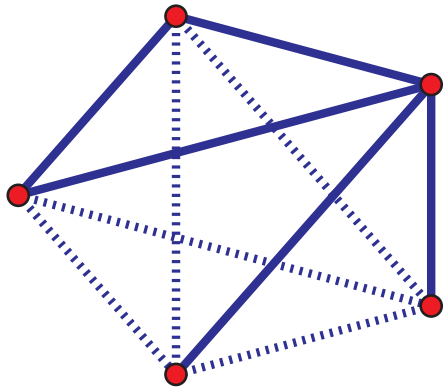
Network Topology Inference (cont.)

There are *many* variants of this problem!

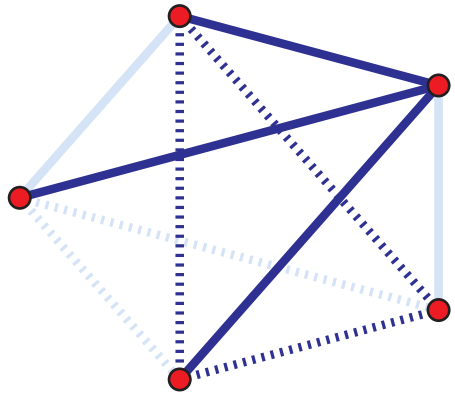
Three general, and fairly broadly applicable, versions are

- Link prediction
- Association network inference
- Tomographic network inference

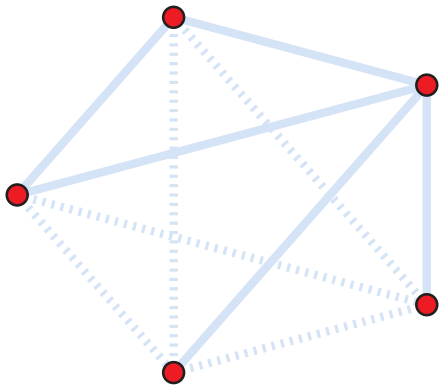
Schematic Comparison of Inference Problems



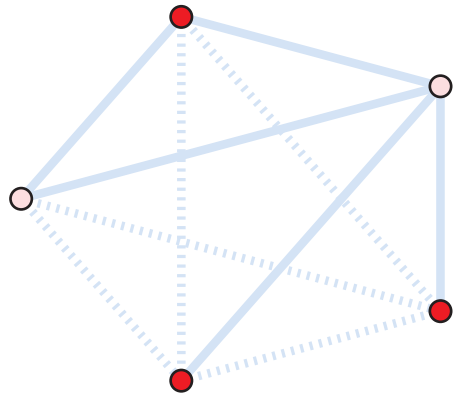
Original Network



Link Prediction



Assoc. Network Inf.



Tomography

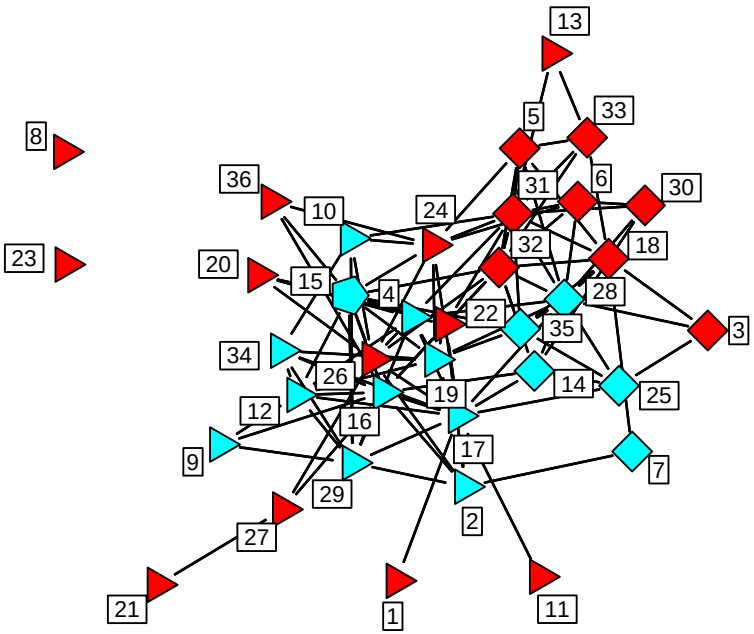
An Aside: Network Modeling

With the emphasis on a statistical perspective in this tutorial, in focusing our brief discussion of *network topology inference*, we are by-passing the important and intimately related topic of *network graph modeling*.

In the latter, the primary emphasis arguably is on the specification of models (typically motivated by 'real-world networks') and the study of their properties (often focused on 'explaining' observed fundamental topological characteristics).

More synergy is needed between the statistical and mathematical activities in this area!

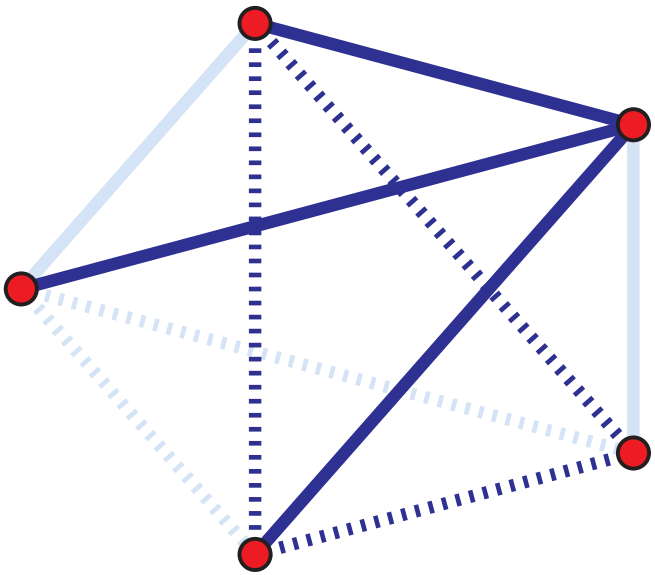
Link Prediction: Examples



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



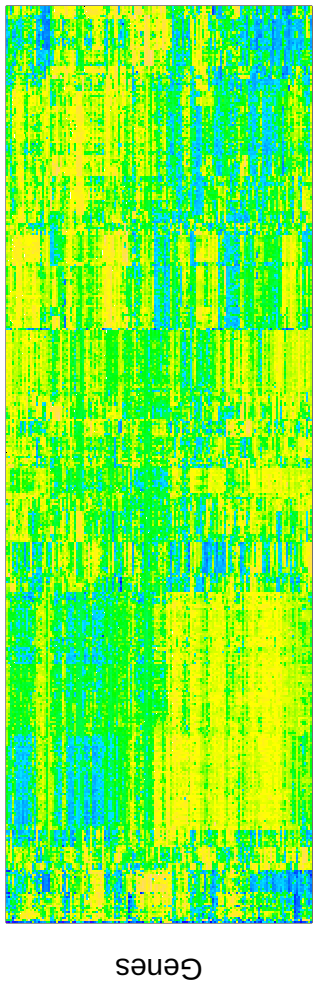
Goal is to predict the edge status' $\mathbf{Y}^{miss}$ for all potential edges with missing (i.e., unknown) status, based on

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

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

- Scoring methods
- Classification methods

Association Network Inference



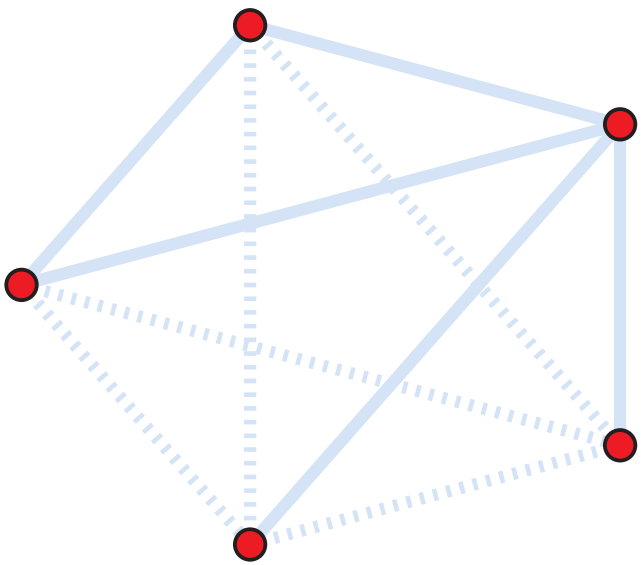
Networks where vertices are linked if there is a sufficient level of 'association' between attributes of vertex pairs.

Examples include

- citation networks
- movie networks
- gene regulatory networks
- neuro functional connectivity networks

BOSTON UNIVERSITY

Association Network Inference: Problem



Associate an attribute vector $\mathbf{x}$ corresponding to each vertex.

Observe $\{\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}\}$.

(Unknown) edges in $G = (V, E)$ defined based on some notion of $\text{sim}(i, j)$.

Goal: Infer nontrivial values of sim from attributes $\mathbf{x}_i$ and/or corresponding edge status'.

Association Network Inference: Solutions

This is arguably one the major and currently most active areas of contribution to 'network science' from statistics[3].

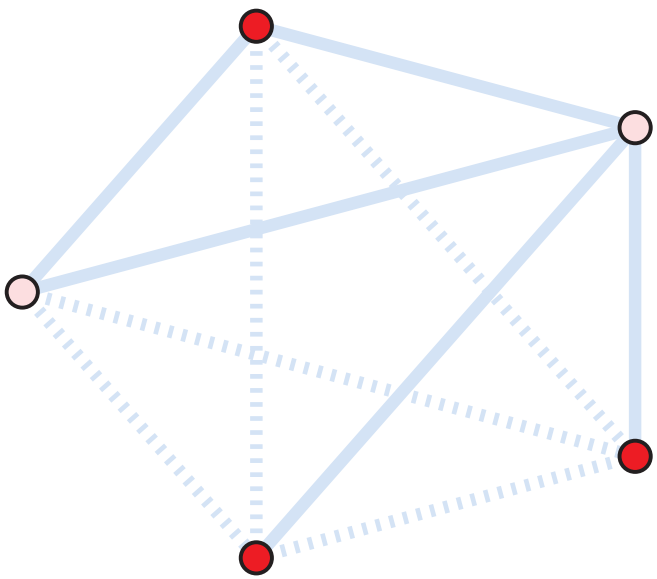
Similarity $\text{sim}(i, j)$ between vertices most commonly defined in terms of variations on correlation.

Inference pursued through two paradigms:

- testing, and
- penalized regression

[3] Discussion topic: Is this good or bad? ☺

Tomographic Topology Inference



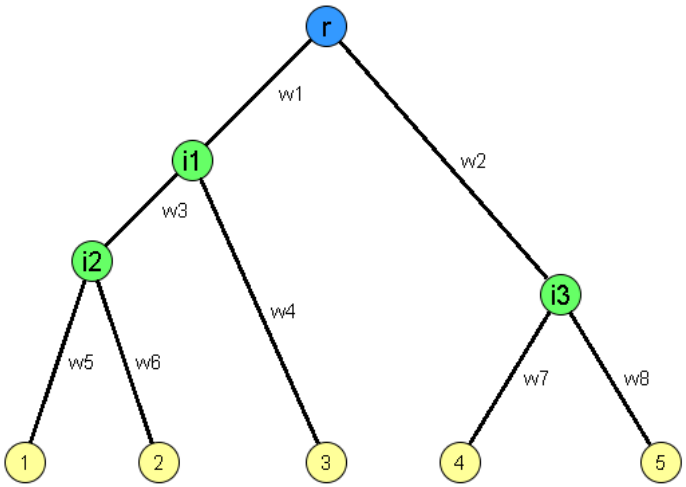
Arguably the most difficult of the network topology inference problems.

Nearly all work done to date assumes $G = (V, E)$ has a *tree* topology.

Two major (but, interestingly, independent!) literatures are

- phylogentic tree inference
- computer network topology identification

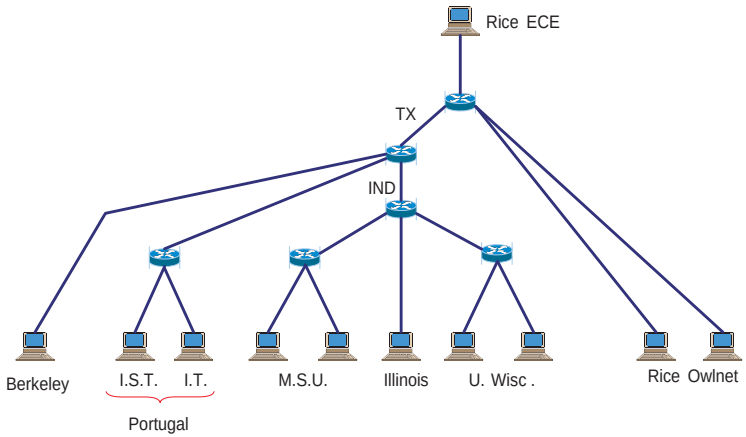
Tomographic Inference of Tree Topologies



Graph assumed to be a tree $T = (V_T, E_T)$.

Leaf vertices (and perhaps the root) known to us.

Goal: Infer the internal vertices and edges, based on measurements at the leaves.



Two major classes of methods are those based on

- hierarchical clustering
- likelihoods.

Where are we at?

- network mapping
- network characterization
- network sampling
- network inference
- network processes

Processes on Network Graphs

So far in this tutorial 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*.

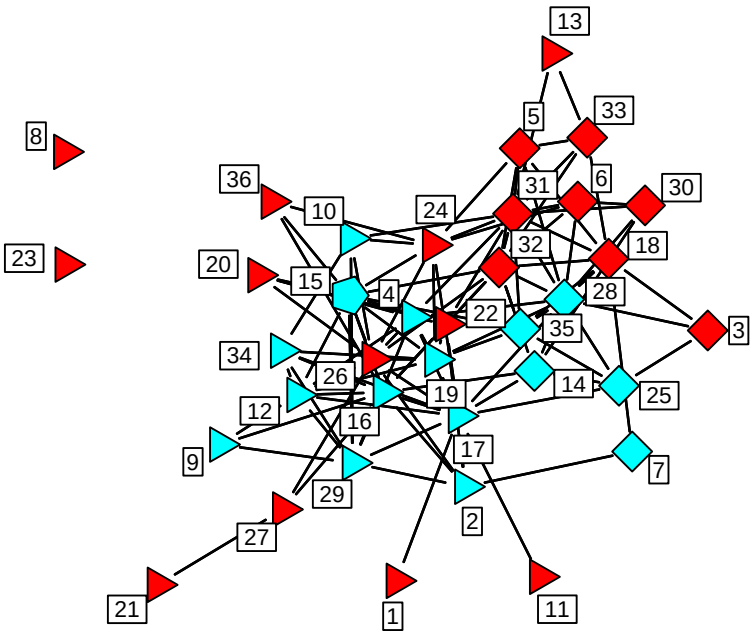
Nevertheless, such quantities may be *influenced by* the interactions among elements.

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.

BOSTON UNIVERSITY

Illustration: Predicting Lawyer Practice

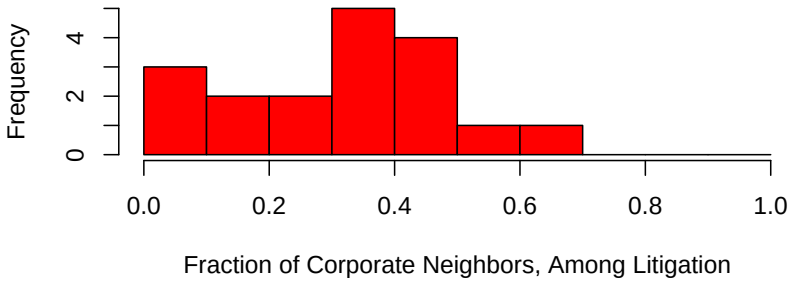


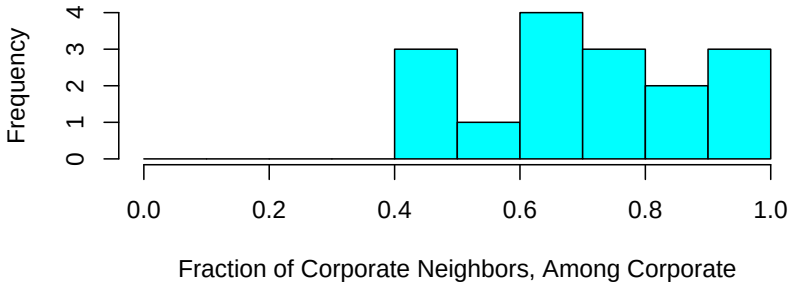
Suppose we observe type of practice – i.e., corporate or litigation – of all but *one* of the 36 lawyers in this company.

Question: Is knowledge of that lawyer's collaborators and their practice predictive of that lawyer's practice?

In fact . . . yes!

Predicting Lawyer Practice (cont.)





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.

Modeling Static Network-Indexed Processes

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

Two commonly used classes are

- Markov random field (MRF) models

  $\Longrightarrow$ Extends ideas from spatial/lattice modeling.

- Kernel regression models

  $\Longrightarrow$ Key innovation is construction of graph kernels

Dynamic Network-Indexed Processes

Many (most?) network-index processes are dynamic

Two classes of processes that have received a great deal of attention are

- epidemics/diffusions on networks, and

  See tutorials by Vespignani and Durrett for more details!

- flows

Network Flows

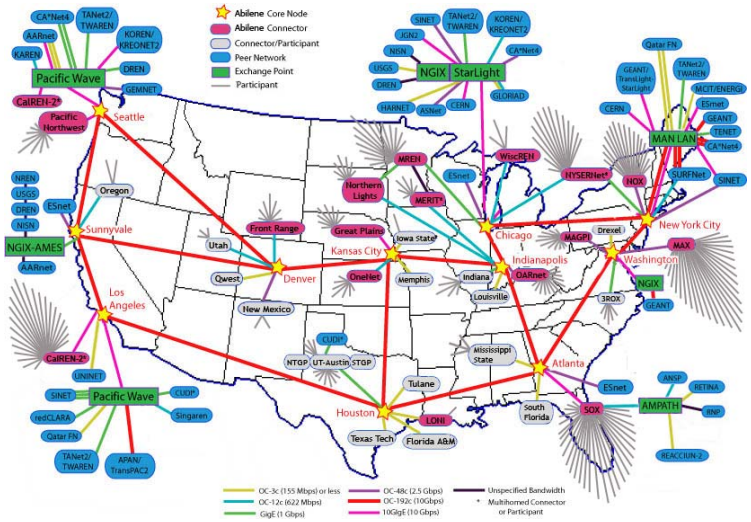
Often networks serve as conduits – either literally or figuratively – for *flows*.

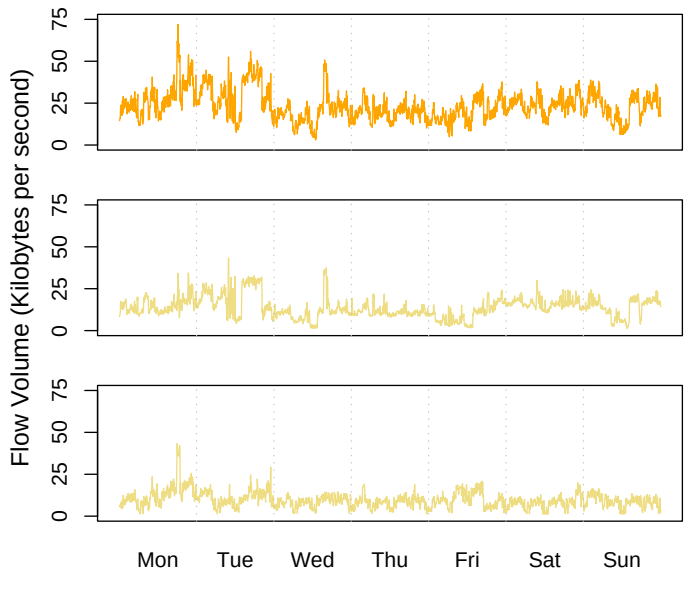
I.e., they facilitate the movement of something.

Examples include

- transportation networks
  i.e., flows of commodities and/or people
- communication networks
  i.e., flows of data
- networks of international trade relations
  i.e., flows of capital

Statistical Analysis of Network Flow Data





Many statistical challenges for modeling/analysis of network flow data.

Examples include prediction of

- origin-destination (OD) flow volume *(Gravity Modeling)*
- link volumes *(OD Traffic Matrix Estimation)*
- OD and/or link flow costs *(Quality of Service (QoS) Monitoring)*

Wrapping Up . . .

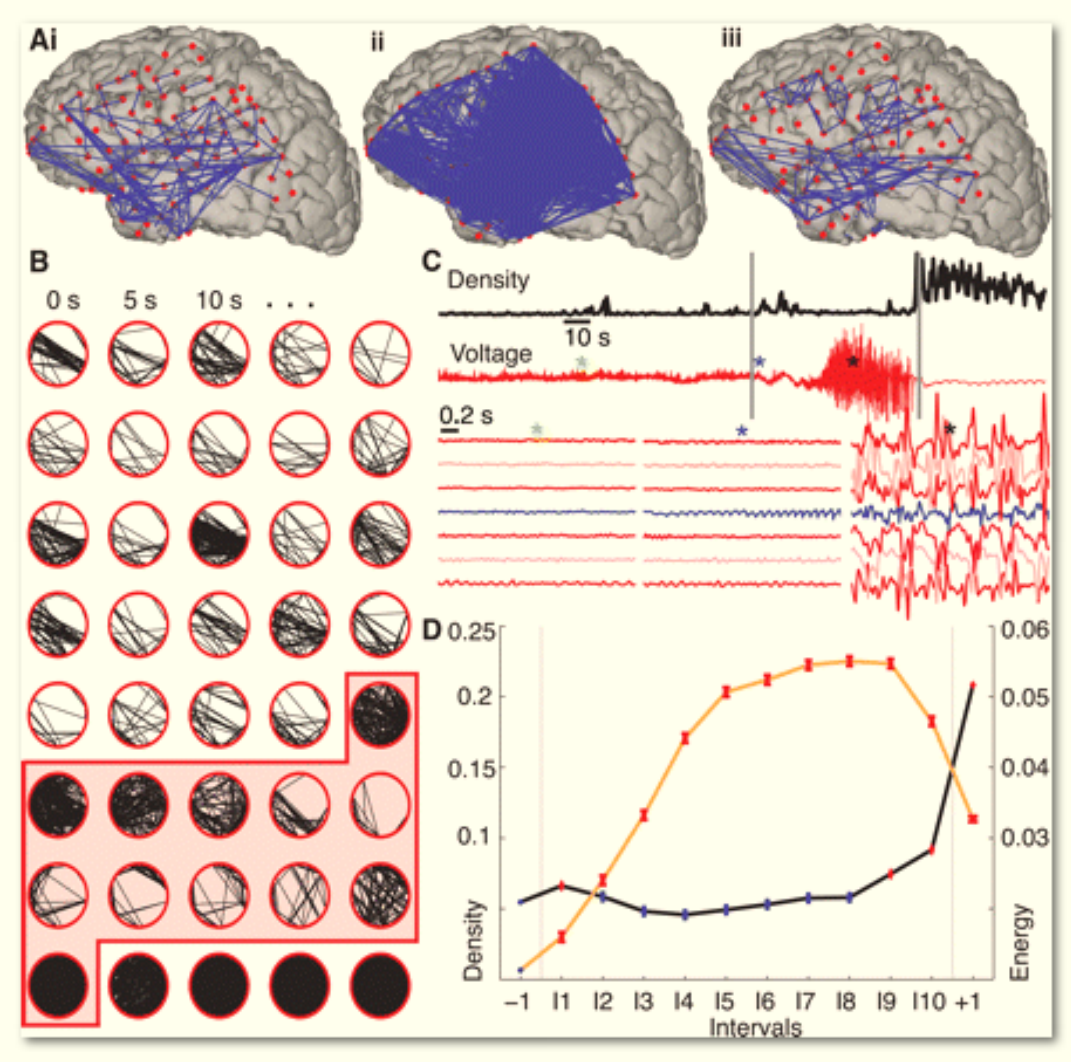
This SAMSI program has in mind (at least) 4 sub-themes for focus:

- Network sampling and inference
- Dynamic networks
- Epidemic processes on networks
- Spectral and geometric methods for networks

The success of these sub-programs depends on you . . .
. . . and will necessarily draw on efforts across the disciplines!

BOSTON UNIVERSITY

Lots of work to be done!



Kramer *et al.* (2010). Coalescence and fragmentation of cortical networks during focal seizures. *J. of Neuroscience*