Statistical Analysis of Network Data

Lecture 1 – Network Mapping & Characterization

Eric D. Kolaczyk

Dept of Mathematics and Statistics, Boston University

kolaczyk@bu.edu

BOSTON UNIVERSITY

Setting Context

Over the past decade, the study of so-called 'complex networks' – that is, network-based representations of complex systems – has taken the sciences by storm.

Researchers from biology to physics, from economics to mathematics, and from computer science to sociology, are more and more involved with the collection, modeling and analysis of network-indexed data.

With this enthusiastic embrace of networks across the disciplines comes a multitude of statistical challenges of all sorts – many of them decidely non-trivial.

Focus of these Talks

In this series of three lectures 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 topics like

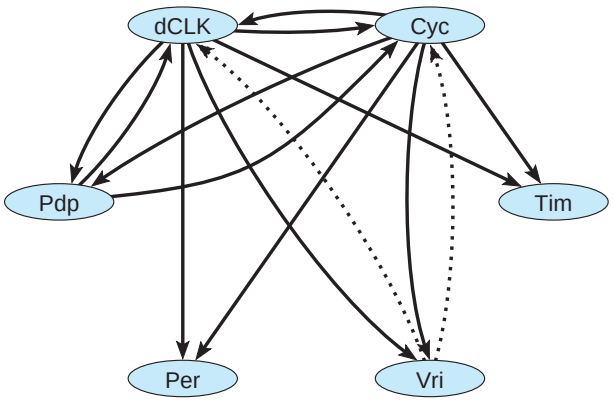
- network summary and visualization
- network sampling
- network modeling and inference, and
- network processes.

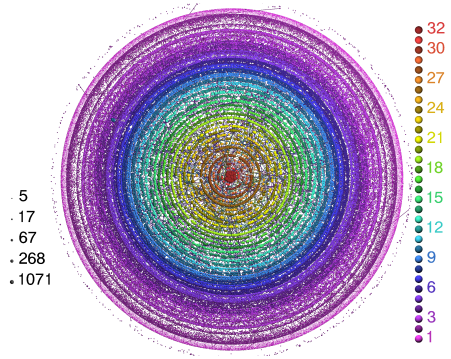
Concepts will be illustrated drawing on examples from bioinformatics, computer network traffic analysis, neuroscience, and social networks.

For a more complete coverage, see

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

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'.

BOSTON UNIVERSITY

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.

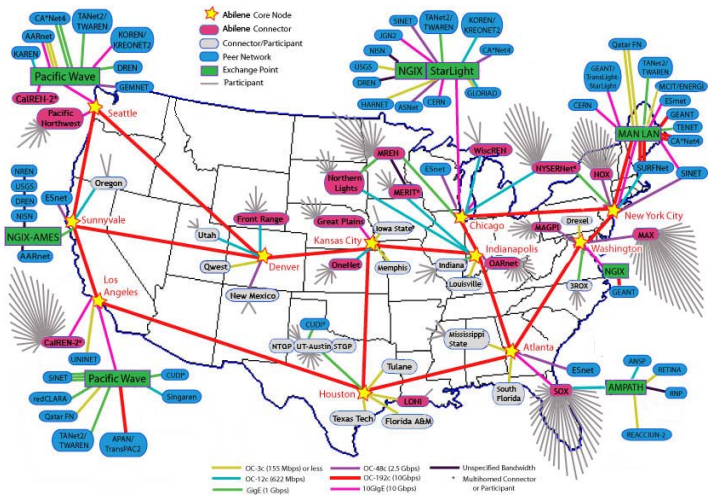
Examples of Networks

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

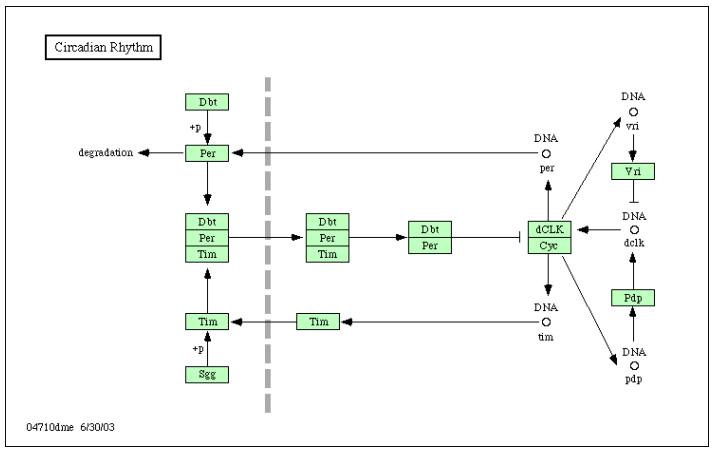


Includes communication, transportation, energy, and sensor networks.

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?

Biological Nets



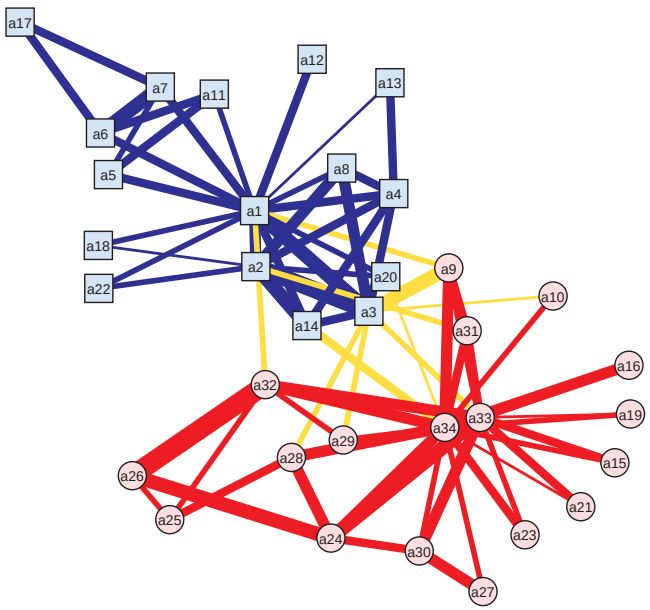
Includes networks of neurons, gene interactions, metabolic paths, predator/prey relationships, and protein interactions.

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

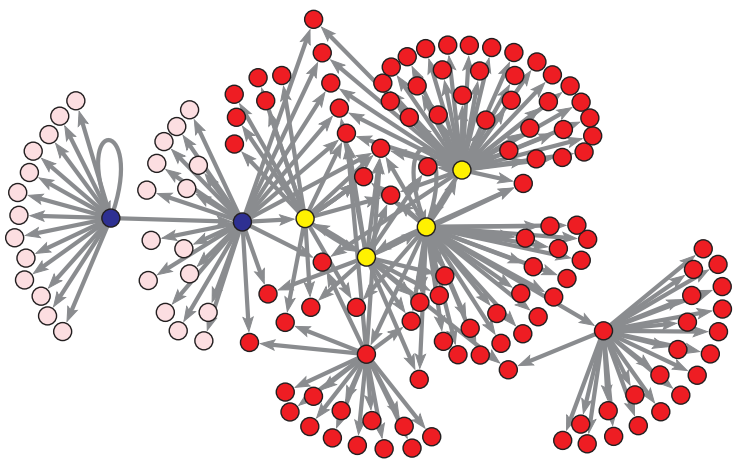


Examples include friendship networks, corporate networks, email networks, and networks of international relations.

Questions include

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

Information Nets



Examples include the WWW, Twitter, and peer-to-peer networks (e.g., 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.

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 be achieved in this area by viewing problems – and pursuing solutions – from this perspective.

BOSTON UNIVERSITY

This Lecture: 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.

BOSTON UNIVERSITY

Outline

1 Introduction

2 Network Mapping

3 Network Characterization

BOSTON UNIVERSITY

Network Mapping

What is 'network mapping'?

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

What is 'the' network?

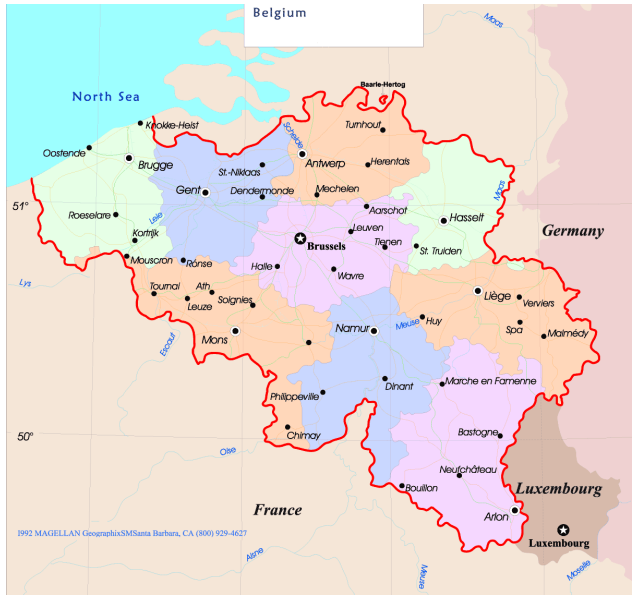
- Network as a 'system' of interest;
- Network as a graph representing the system;
- Network as a visual object.

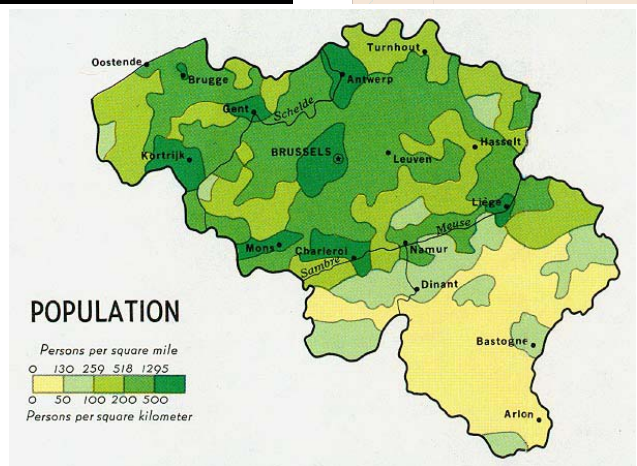
Analogue: Geography and the production of cartographic maps.

BOSTON UNIVERSITY

Example: Mapping Belgium

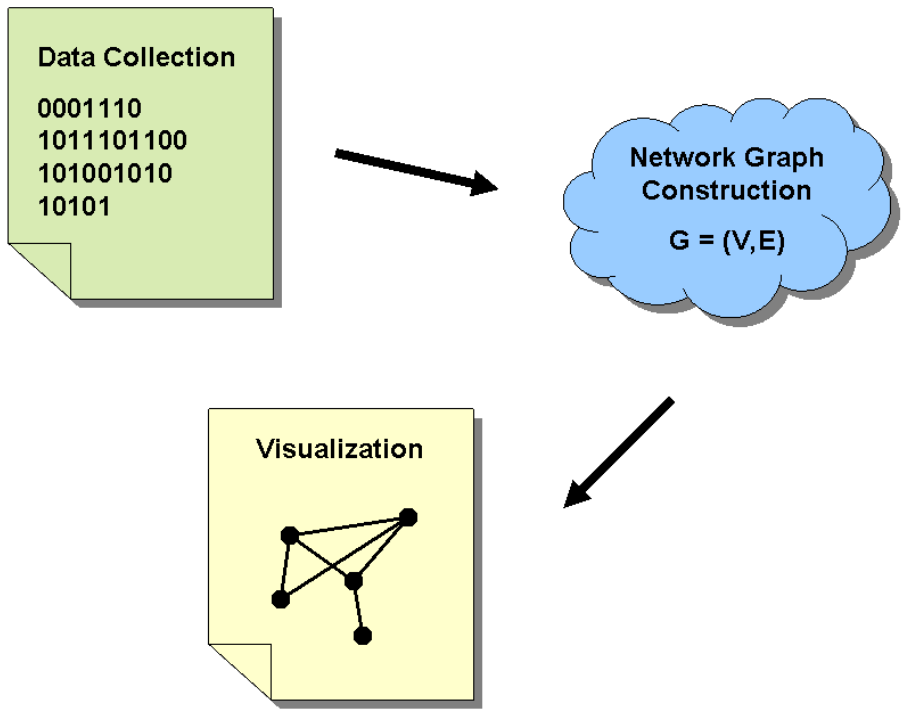
Which of these is 'the' Belgium?





BOSTON UNIVERSITY

Three Stages of Network Mapping

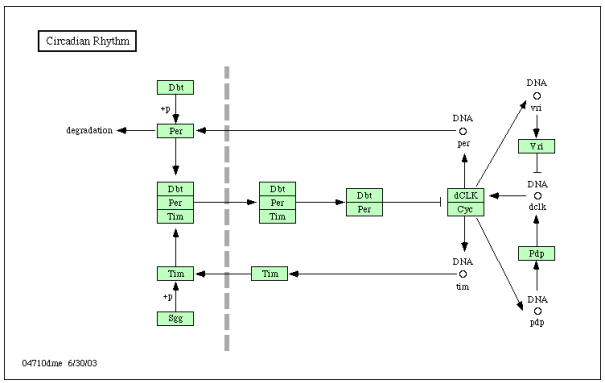


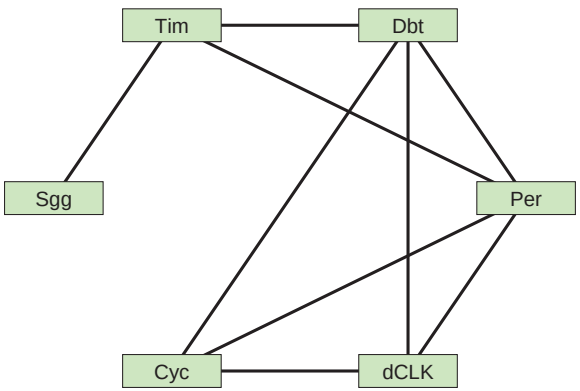
Continuing our geography analogue . . . a fourth stage might be 'validation'.

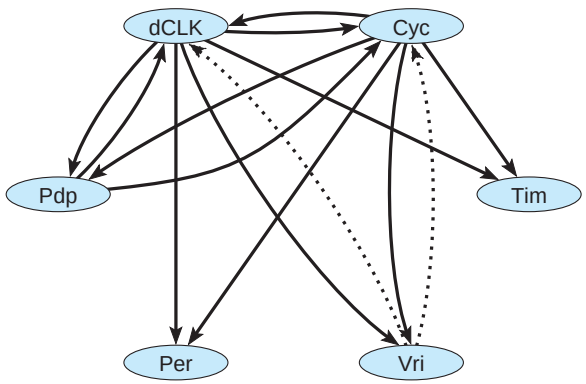
Stage 1: Collecting Relational Network Data

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 l-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!

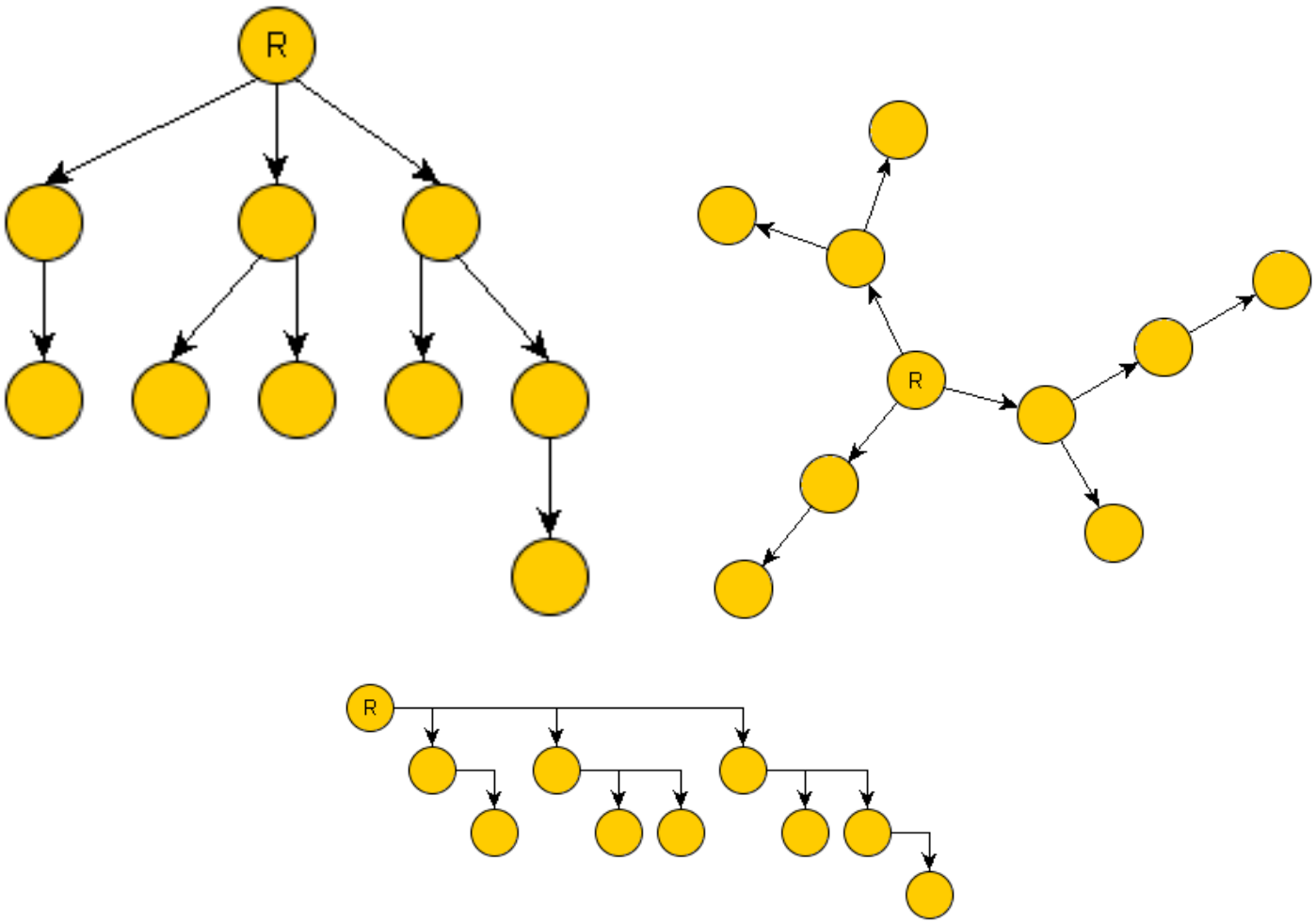
Common to better define / constrain this problem by incorporating

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

BOSTON UNIVERSITY

Layout ... Does it Matter?

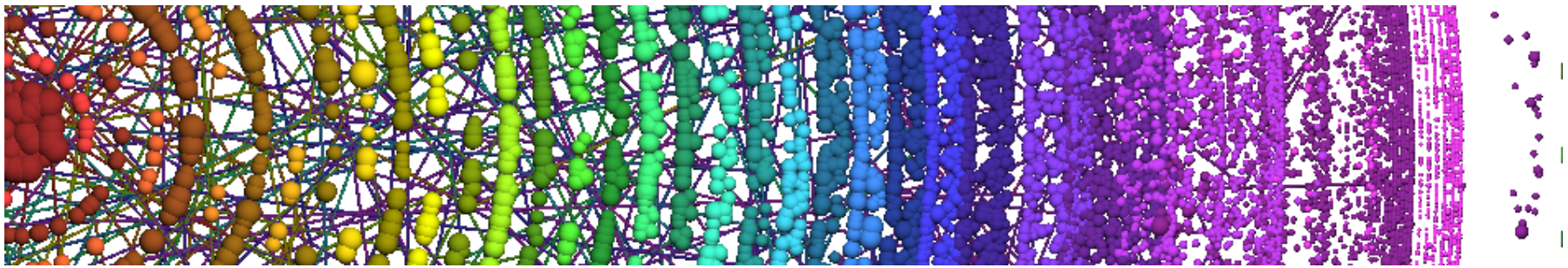
Yes!



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

BOSTON UNIVERSITY

Illustration: Visualizing Large-Scale Networks – A Cross-sectional View of the Internet

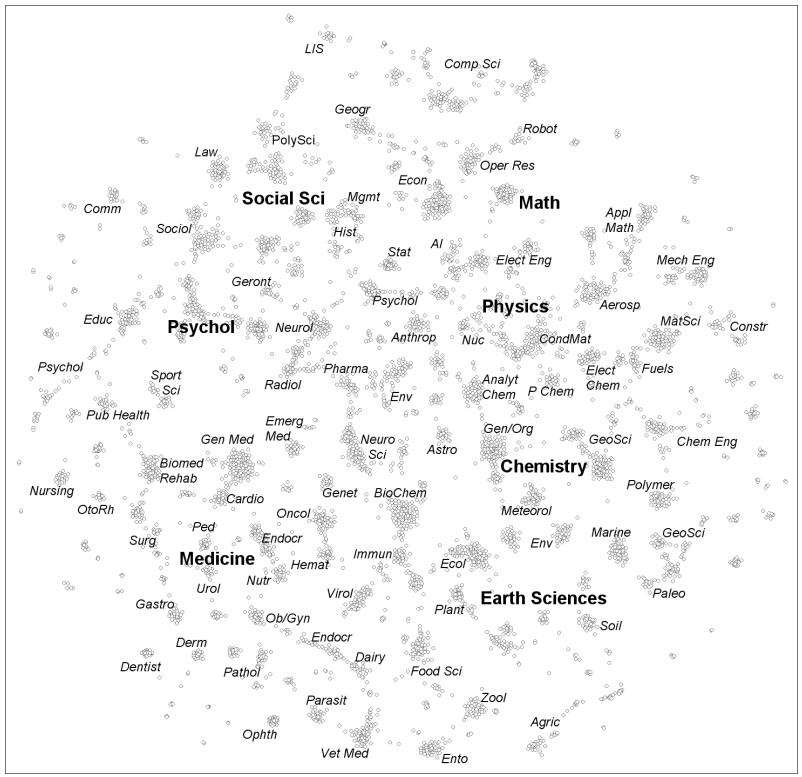


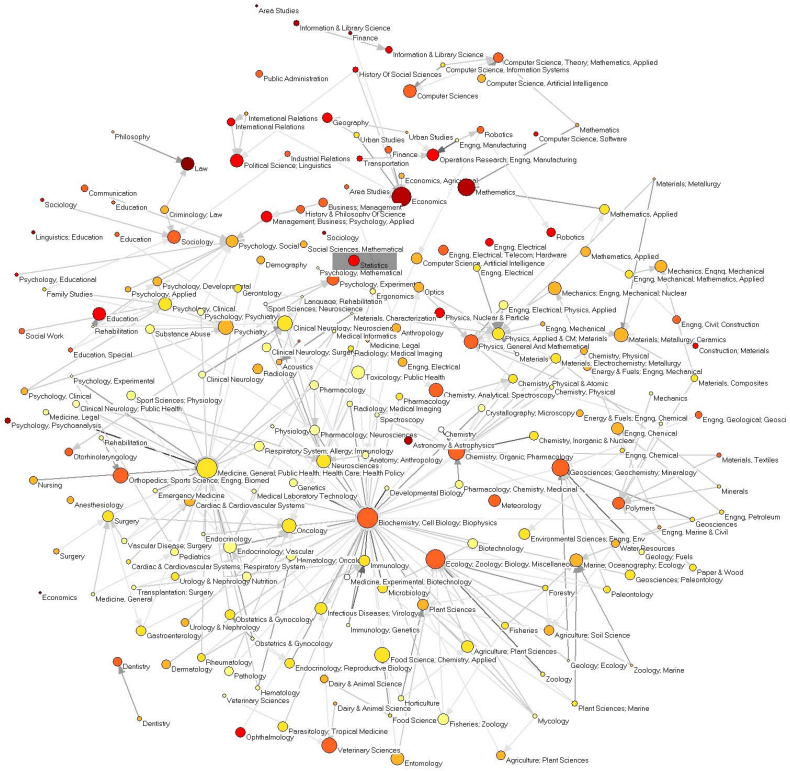
- CAIDA data on 192,244 nodes and 609,066 (logical) edges in the router-level Internet, obtained using `traceroute`.
- Visualization using LaNet-vi tools[2]
- Vertices nested according to k-core 'shells' and laid out using principles from radial and spring-embedder methods.

[2] http://lanet-vi.soic.indiana.edu/

BOSTON UNIVERSITY

Illustration: Mapping 'Science'





- Boyack, Klavens, & Borner (2005)
- 16.24M citations among 7121 scientific journals
- Substantial preprocesing + human interpretation

BOSTON UNIVERSITY

Outline

1 Introduction

2 Network Mapping

3 Network Characterization

BOSTON UNIVERSITY

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*.

BOSTON UNIVERSITY

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.

BOSTON UNIVERSITY

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.

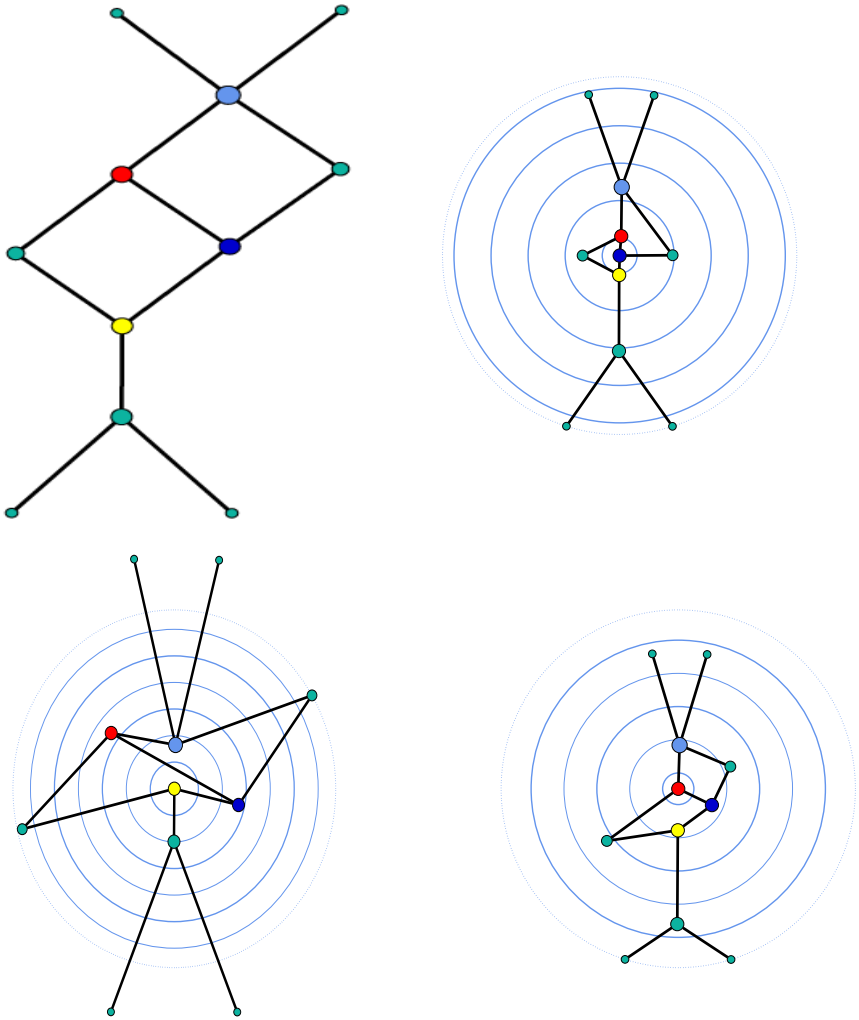
Useful, on the one hand, but indicative of something problematic, on the other hand!

> *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!

BOSTON UNIVERSITY

Centrality: An Illustration

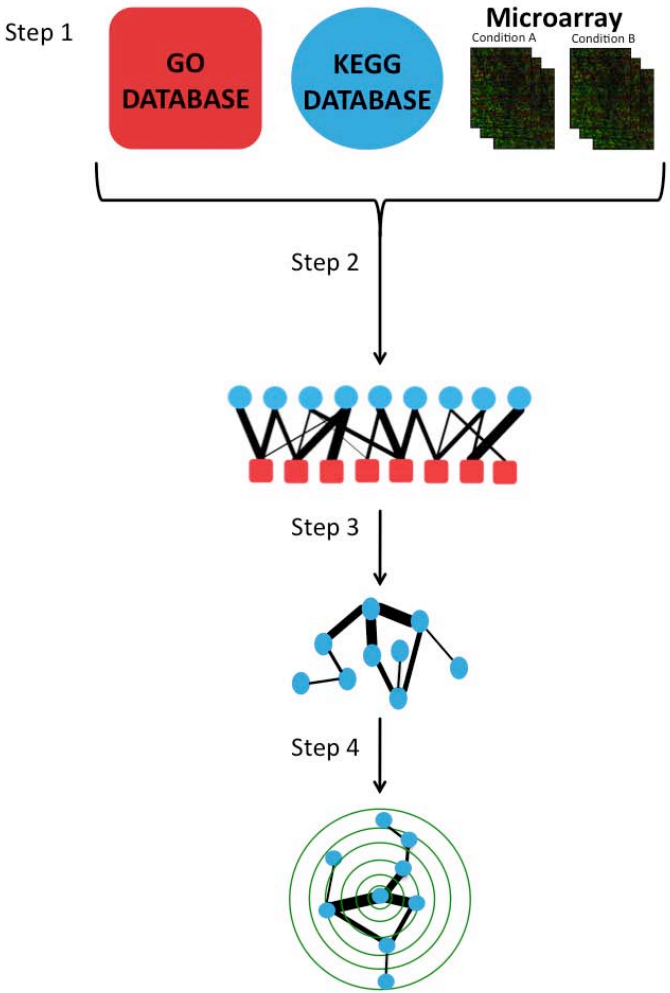


Clockwise from top left: (i) toy graph, with (ii) closeness, (iii) betweenness, and (iv) eigenvector centralities.

Example and figures courtesy of Ulrik Brandes.

BOSTON UNIVERSITY

Centrality in Practice: Cellular Mechanism of Action



LPIA[a] combines biological

1. pathways, and
2. function,

weighted by differential expression, to create a network-based representation of the extent to which pathways are disrupted by a 'perturbation'.

Pathways are ranked using test statistics based on eigenvector centrality.

[a] * Pham, L., Christadore, L., Schaus, S., and Kolaczyk, E.D. (2011). Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proceedings of the National Academy of Sciences*, 108(32), 13347 - 13352.

BOSTON UNIVERSITY

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.

BOSTON UNIVERSITY

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!

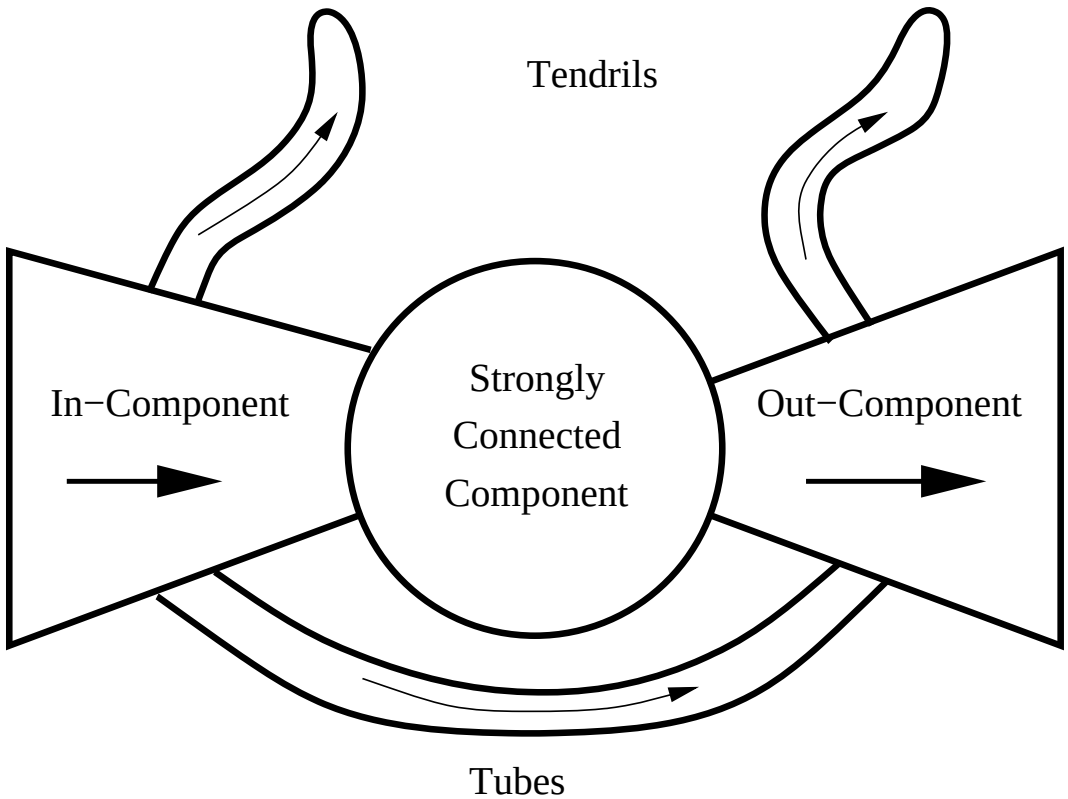
A *(connected) component* of a graph G is a maximally connected sub-graph.

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

- giant component
- smaller components
- isolates

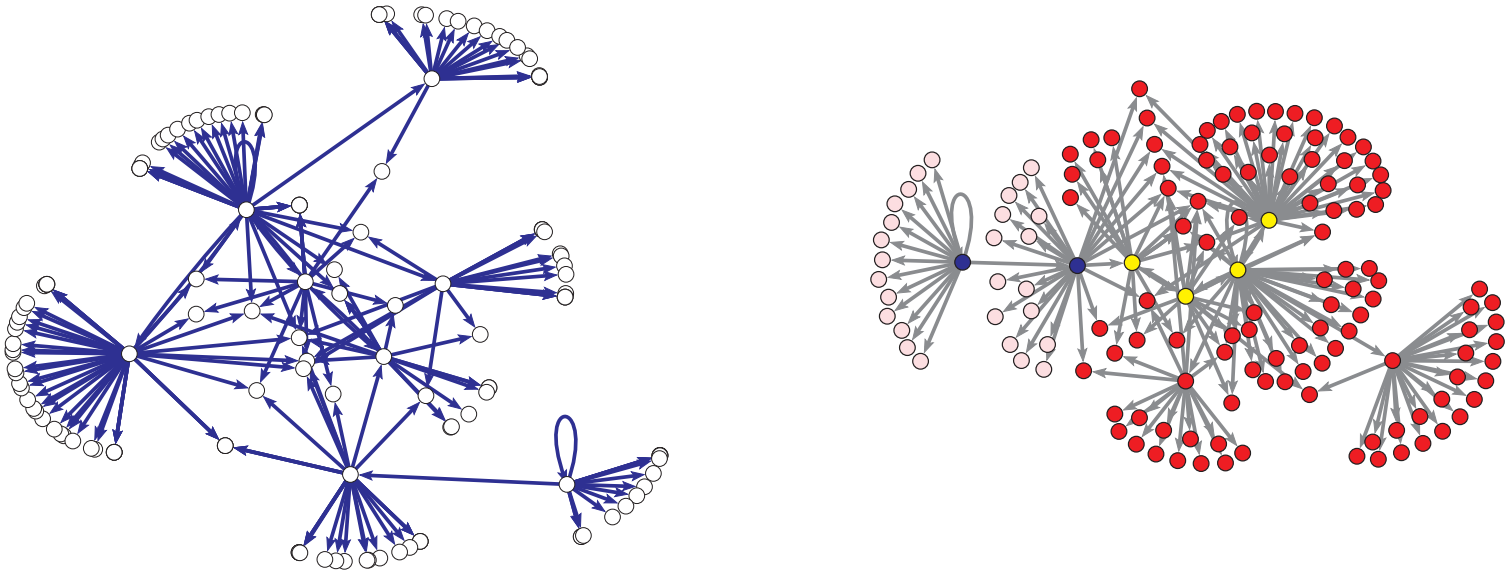
Frequently, reported analyses are for the *giant component*.

Components in Directed Graphs Become Interesting!



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)

BOSTON UNIVERSITY

Putting it All Together: Hypersynchronization in Epilepsy?

Some of our current work is focused on characterization of time-indexed sequences of networks during epileptic seizures[3]

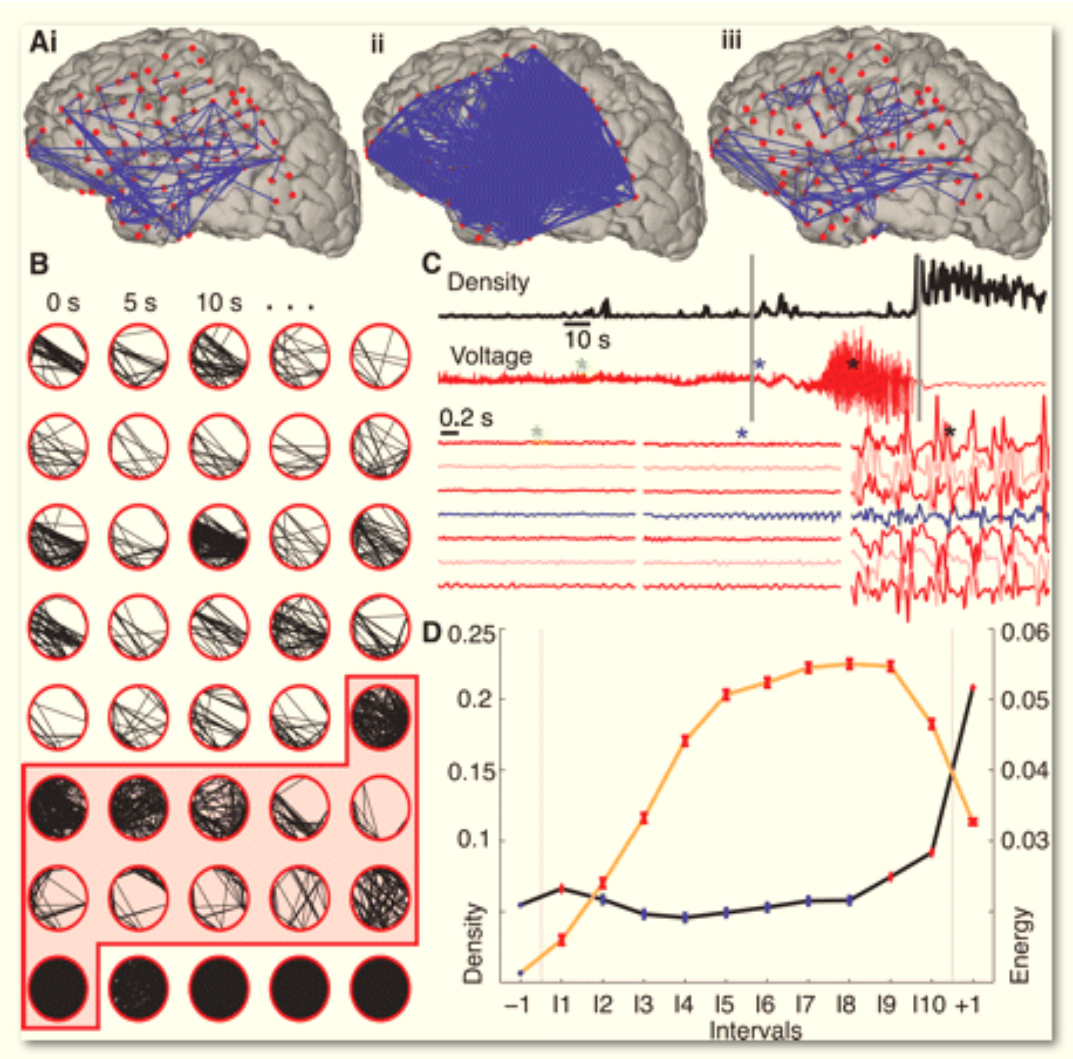
- (Focal) epileptic seizures understood to begin with a local event
- Changes in synchronization among brain regions is presumed to underlie mechanisms of seizure propagation
- Classically, seizures through to represent a hypersynchronous state
- Recent work has begun to challenge this assertion at larger spatial scales

Our results suggests there is actually a phenomenon of coalescence and fragmentation at work.

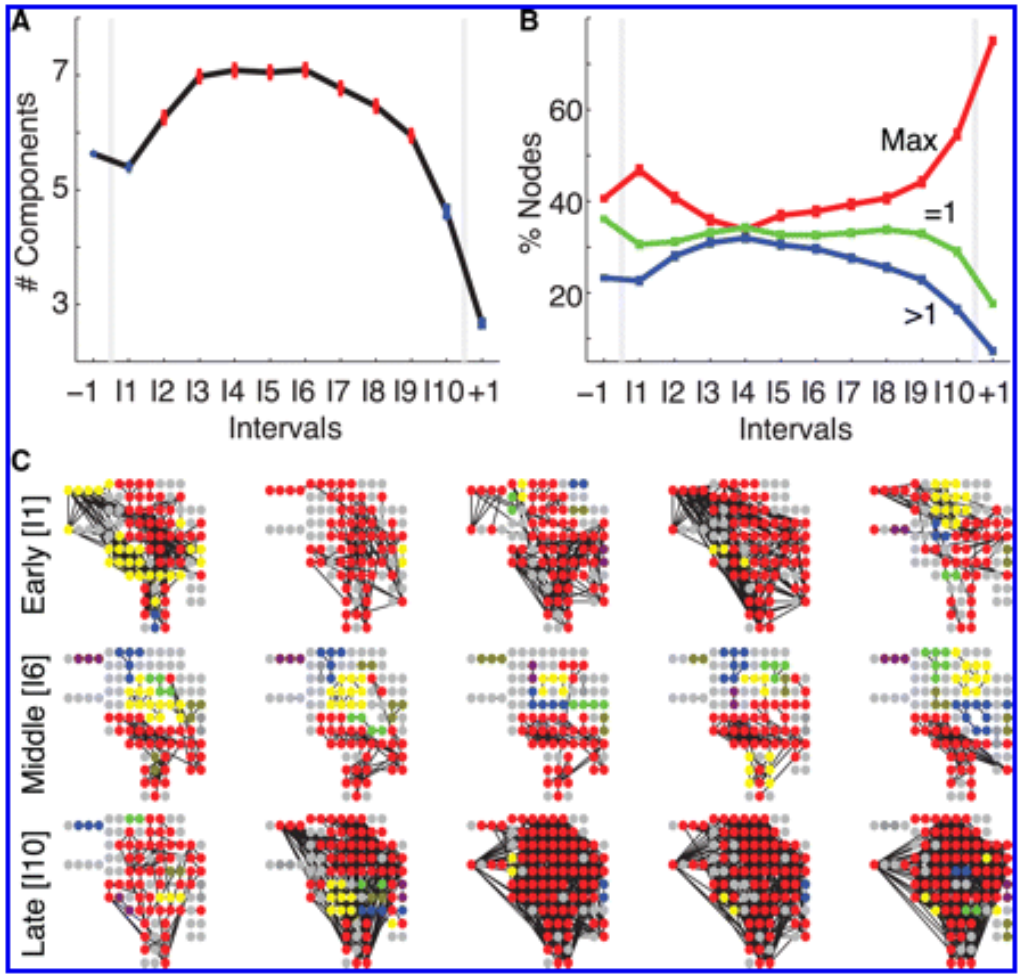
[3] Kramer, M.A., Eden, U.T., Kolaczyk, E.D., Zepeda, R., Es kandar, E.N., Cash, S.S. (2010). Coalescence and fragmentation of cortical netwo rks during focal seizures. *Journal of Neuroscience,* 30(30), 10076-10085.

BOSTON UNIVERSITY

Absence of Hypersynchrony During Seizure



Fragmentation & Coalescence



Looking Ahead . . .

Following two lectures will look at

1. Network sampling and modeling (today)
2. Network processes (tomorrow).