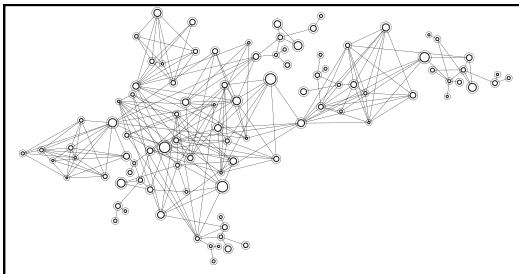
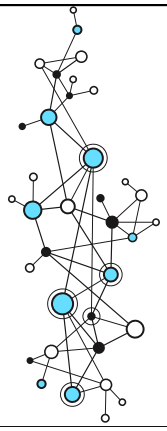


## Network science in biology

measuring, visualizing and modelling real world complex networks

Petra Vertes



## Overview of topics

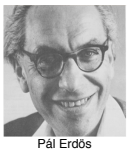
1. What is a network? – examples from social and biological sciences.
2. Constructing and representing complex networks.
3. Topological properties of networks – how to measure them and why they matter?
4. Network analysis in biological sciences – six examples
5. Generative modelling of networks – why and how?
6. Getting hold of data and code – tools and resources for network analysis

# 5.

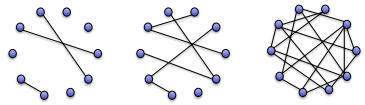
## Modelling Networks
## -
## What is a Network Model?

## The Erdös-Rényi model

N nodes, each pair connect with uniform probability p

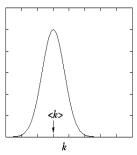


Pál Erdös



p=0.07 p=0.14 p=0.4

Alfréd Rényi

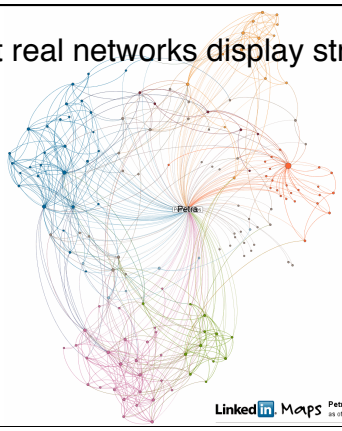



The resulting network has no visible structure, it has a binomial degree distribution (& some other interesting theoretical properties)

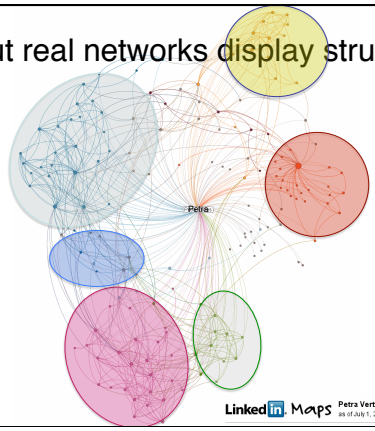
$$P(\deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Albert, Barabási *Reviews of Modern Physics* (2002) .

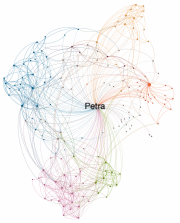
## But real networks display structure

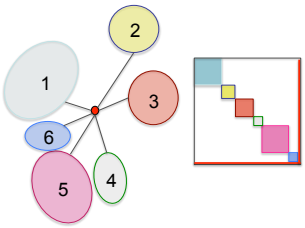



## But real networks display structure



## New models may explain this structure

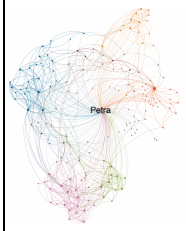


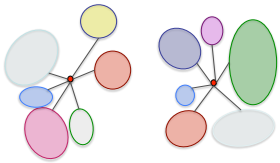




**New Model:** random connections between people in same country. No connections between countries.

**So:** we can compress the information in the network

## New models may explain this structure

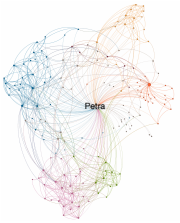


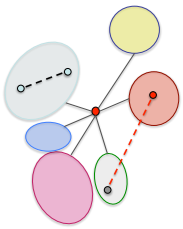



**New Model:** random connections between people in same country. No connections between countries

**So:** we can generalize from observed network to other similar networks

## New models may explain this structure

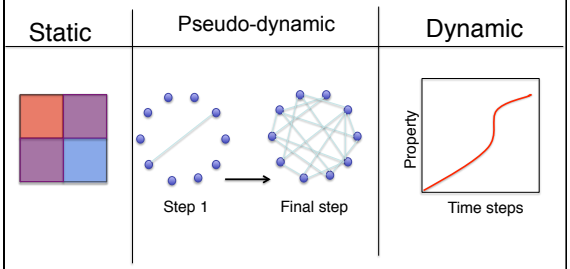


**New Model:** random connections between people in same country. No connections between countries

**So:** we can generalize from past network to future network. i.e. where will next link appear?

## Three classes of models



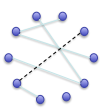

## Why Model Networks?

**Machine learning / Statistics:**

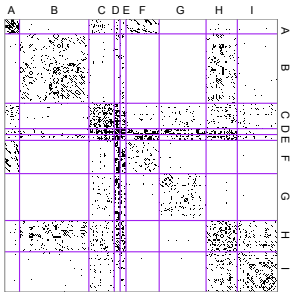
- Compressing information
- Generalizing from observed to unobserved network data:
  - Inferring latent properties that give rise to communities
  - Predicting missing links

**Physics:**

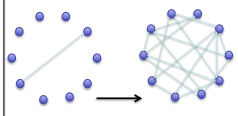
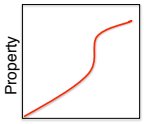
- Understanding the mechanisms for network formation:
  - Why are people friends?
  - How do neurons choose synaptic neighbours?




## How to Model Networks? (ML)



## Three classes of models

| Static | Pseudo-dynamic | Dynamic |
|---|---|---|
|  |  Step 1 Final step |  Property Time steps |

## How to Model Networks? (Phys)

1. **Make a list of stylized facts:**
   - What is special or interesting about the structure we observe?
   - Is this structure real and is it really interesting?

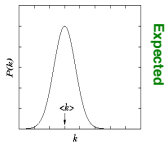
2. **Come up with simplest mechanism leading to this structure**
   - Use some domain knowledge
   - Ignore all the details (at least in the beginning)
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
   - Same data about different systems
   - Different data about the same system

5. **Think of what the model doesn't capture**
   - Improve the model

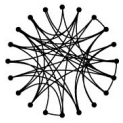
## Stylized fact 1:

Power-law degree distributions

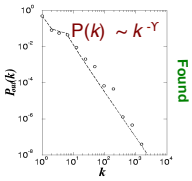
## The degree distribution



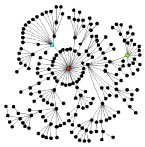

Expected



Random Network



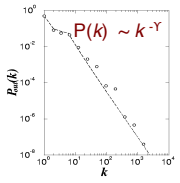

Found

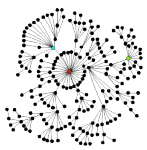


Scale-free Network

## Power-law degree distribution

- world-wide web
- electronic circuits
- coauthorship networks
- communication networks
- sexual web
- protein interaction networks
- linguistic networks
- international trade networks
- brain networks
- …



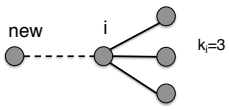
*Barabasi & Albert (1999) Science*

## The Barabási-Albert Model

aka:Preferential attachment

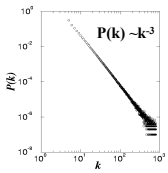
(1) Growth: Networks continuously expand by the addition of new nodes

**parameters: In practice, start with $m_0$ nodes give each new node $m \le m_0$ edges**




(2) Preferential attachment: New nodes prefer to link to highly connected nodes.

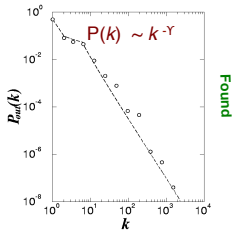
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$




**Exponent independent of $m_0$ and m**

*Barabási & Albert, Science 286, 509 (1999)*
*Albert & Barabási, Reviews of Modern Physics (2002)*

## Stylized fact – but is it real?




The distribution plotted on a log-log scale is a straight line.

But such a visual method is not reliable to correctly identifying power-laws and their exponents γ.

Instead, more rigorous methods have been proposed:

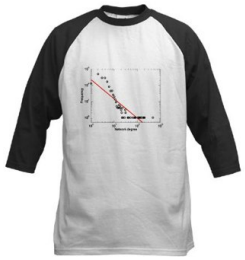
*Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51: 661-703.*

## Identifying power-laws (II)

1. The frequency of occurrence of unique words in the novel Moby Dick by Herman Melville.
2. The degrees of proteins in the partially known protein-interaction network of the yeast Saccharomyces Cerevisiae.
3. The degrees of metabolites in the metabolic network of the bacterium Escherichia Coli.
4. The degrees of nodes in the partially known network of the Internet at the level of autonomous systems for May 2006.
5. The number of calls received by customers of AT&Ts long distance telephone service in the US during a single day.
6. The intensity of wars (1816 to 1980), measured as the number of deaths per 10 000 in populations of the warring nations.
7. The severity of terrorist attacks from February 1968 to June 2006, measured as the number of deaths directly resulting.
8. The number of bytes of data received as the result of individual web (HTTP) requests from computer users at a large research laboratory during a 24-hour period in June 1996. Roughly speaking this distribution represents the size distribution of web files transmitted over the Internet.
9. The number of species per genus of mammals. This data set is composed primarily of species alive today but also includes some recently extinct species, where recent in this context means the last few tens of thousands of years.
10. The numbers of sightings of birds of different species in the North American Breeding Bird Survey for 2003.
11. The numbers of customers affected in electrical blackouts in the United States between 1984 and 2002.
12. The numbers of copies of bestselling books sold in the United States during the period 1895 to 1965.
13. The human populations of US cities in the 2000 US Census.
14. The sizes of email address books of computer users at a large university.
15. The sizes in acres of wildfires occurring on US federal land between 1986 and 1996.
16. Peak gamma-ray intensity of solar flares between 1980 and 1989.
17. The intensities of earthquakes in California between 1910 and 1992, measured as the maximum amplitude of motion during the quake.
18. The numbers of adherents of religious denominations, bodies, and sects, as published on the web site adherents.com.
19. The frequencies of occurrence of US family names in the 1990 US Census.
20. The aggregate net worth in US dollars of the richest individuals in the United States in October 2003.
21. The number of citations received (by June 1997) by papers published in 1981 and listed in the Science Citation Index.
22. The number of papers (co-)authored by mathematicians listed in the American Mathematical Societys MathSciNet database.
23. The number of hits received by web sites from customers of the America Online Internet service in a single day.
24. The number of links to web sites found in a 1997 web crawl of about 200 million web pages.

List of 24 datsets previously claimed to be power-law and re-analyzed in Clauset et al (2009). Shaded datasets were found to be inconsistent with power-laws.

## The Power-Law Shop



Men's Baseball Jersey $20.00

Infant Bodysuit $14.50

*http://www.cafepress.com/thepowerlawshop*

## How to Model Networks? (Phys)

1. **Make a list of stylized facts:**
   - What is special or interesting about the structure we observe?
   - Is this structure real and is it really interesting?
2. **Come up with simplest mechanism leading to this structure**
   - Use some domain knowledge
   - Ignore all the details (at least in the beginning)
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
   - Same data about different systems
   - Different data about the same system
5. **Think of what the model doesn't capture**
   - Improve the model

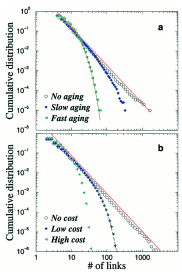
## Variants of preferential attachment (I)

Corrected stylized fact:

Many real world networks don't really have power-law degree distributions. E.g. C elegans connectome, electric power grid, actors network.

Variants of BA could explain such deviations:

- Nodes can be 'active 'or 'inactive' and only active ones receive new links.
- **Age**: Nodes start off 'active' but with each time step they transition to being 'inactive' with a certain probability P
- **Cost:** Nodes start off 'active' but they transition to being 'inactive' when they have reached a maximum degree $k_{max}$




*Amaral et al (2000) PNAS*

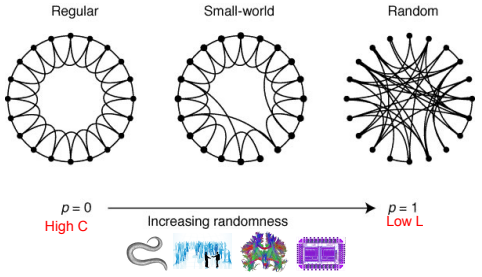
Internet Mathematics Vol. 1, No. 2: 226-251

## A Brief History of Generative Models for Power Law and Lognormal Distributions

Michael Mitzenmacher

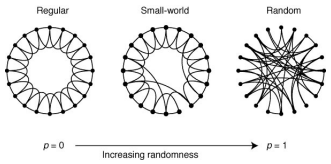
## Stylized fact 2:
Small Worlds: clustered and efficient

## The Watts-Strogatz Model




*Watts & Strogatz (1998) Nature*
*Albert & Barabási (2002) Reviews of Modern Physics*

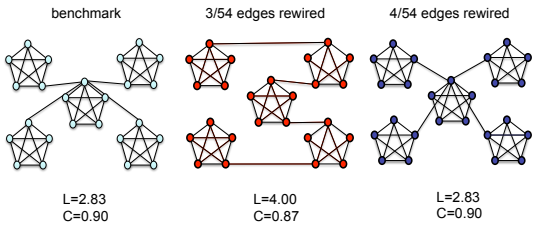
## Stylized fact – but is it interesting?




"Models of dynamical systems with small-world coupling display enhanced signal-propagation speed, computational power, and synchronizability.
In particular, infectious diseases spread more easily in small-world networks than in regular lattices."

*Watts & Strogatz (1998) Nature*

## Stylized fact – but is it interesting?

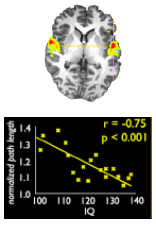
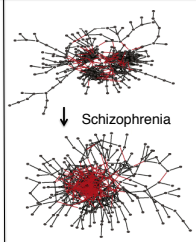
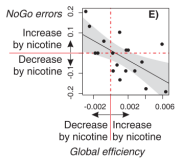



L=2.83
C=0.90

L=4.00
C=0.87

L=2.83
C=0.90

## Stylized fact – but is it interesting?

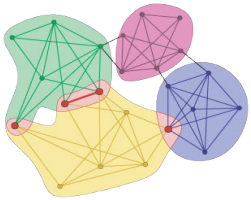
| Correlated with IQ | Affected by disease | Affected by drugs |
|---|---|---|
|  |  | Nicotine tends to increase Global Efficiency  |
| *Li et al (2009)*<br>*Van den Heuvel et al (2009)* | *Alexander-Bloch et al (2012)* | *Giessing et al (2013)* |



## Stylized fact 3:
Modularity

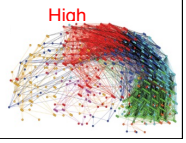
# Modularity

*M* = *Fraction of within-module connections compared to what is expected by chance (for the optimal partition).*



Low Q

High



# Capturing/Generating Modularity

This is more complicated than previous stylized facts because it requires us to differentiate between nodes by either:

(1) Assuming latent attributes

e.g. Blockmodels

(2) Using observed attributes in model

e.g. Spatial embedding

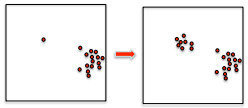
# Modularity through Space

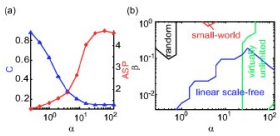
- Seed node at the centre of 2D square
- New nodes *i* are added and their probability of connecting to preexisting node j is:

$$P(i,j) = \beta e^{-\alpha d(i,j)}$$

where $d(i,j)$ is the distance from *i* to *j*

- Nodes that fail to connect to the rest of the network are removed
- Parameter β controls density and parameter α is distance penalty





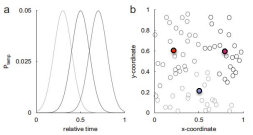

*Kaiser & Hilgetag (2004) Phys Rev E*

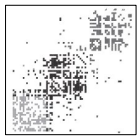
# Modularity through Space and Time

- Three seed nodes (far apart)
- Seed nodes *i* has time window: $P_{temp}(i)$
- New nodes are added and inherit $P_{temp}$ of nearest seed node
- Probability of connecting new node I to preexisting node j is:

$$P(i,j) = P_{temp}(i) \times P_{temp}(j) \times P_{dist}(i,j)$$

- $P_{dist}(i,j)$ decays exponentially with distance between *i* and *j*
- Nodes that fail to connect to the rest of the network are removed



*Kaiser & Hilgetag (2006) Neurocomputing*

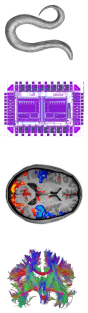
## How to Model Networks?

1. **Make a list of stylized facts**
2. **Come up with simplest mechanism leading to this structure**
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
5. **Think of what the model doesn't capture**

# CASE STUDY I.

Modelling human brain networks

## List of stylized facts (that matter) about brain networks



**Small-world**

- high clustering
- short minimum path length or high efficiency

**Cost-efficient**

- high efficiency for relatively low connection cost

**Hubby**

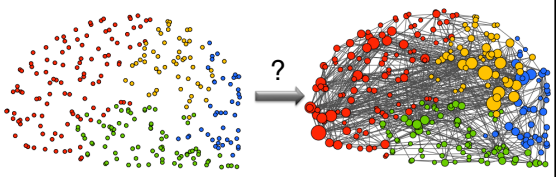
- fat-tailed degree distributions

**Modular**

- nodes are more densely connected to other nodes in the same module than to nodes in other modules

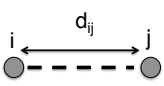
*Bullmore & Sporns (2009) Nat Rev Neurosci*
*Bassett et al (2010) PloS Comp Bio*

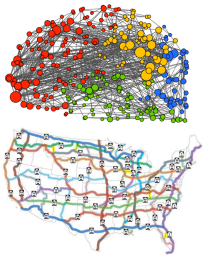
## Generating Brain-like Networks




## A Simple Economical Model

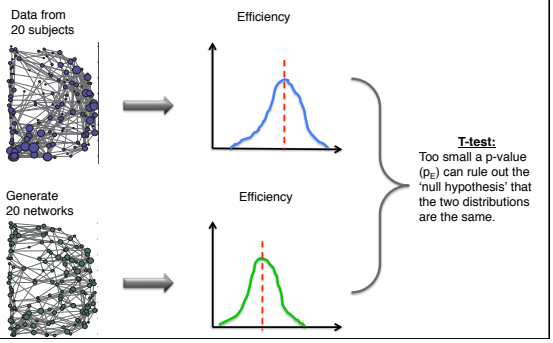
$P_{ij} \sim \exp(-\eta d_{ij})$



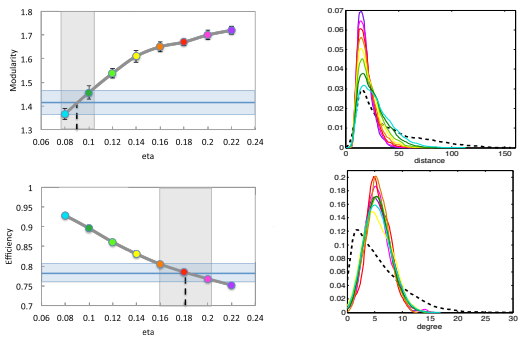



M. Kaiser and C. Hilgetag (2004) *Phys Rev E*
*"Spatial constraints are present during the development of many real networks. In biological systems, for instance, gradients of chemical concentrations, or molecule interactions, decay exponentially with distance."*

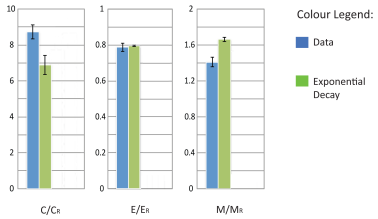
## Comparing Model Networks to Data



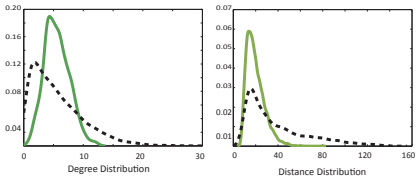

## No Parameter Setting is Satisfactory




## The Economical Model (I)

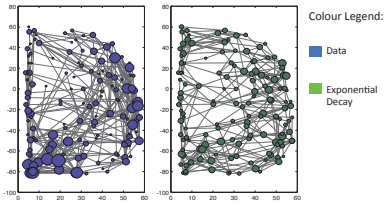


## The Economical Model (II)




Colour Legend: Data Exponential Decay

## The Economical Model (III)

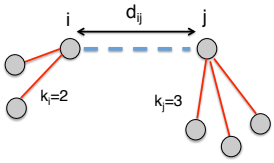


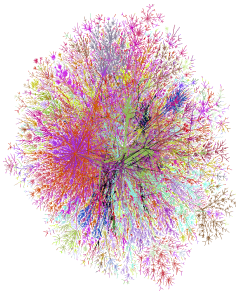

## How to Model Networks?

1. **Make a list of stylized facts**
2. **Come up with simplest mechanism leading to this structure**
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
5. **Think of what the model doesn't capture**
   - hubs

## The Economical Preferential Attachment Model

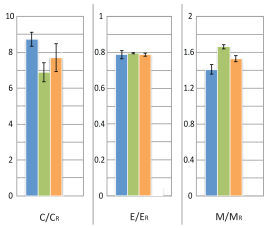
$P_{ij} = d_{ij}^{-\eta} (k_i k_j)^{\gamma}$



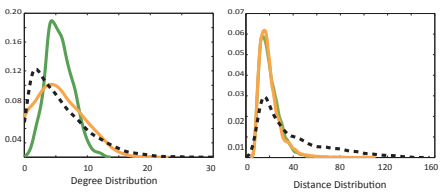
S. H. Yook, H. Jeong, A.-L. Barabási (2002) *PNAS*

## The Economical Preferential Attachment Model (I)



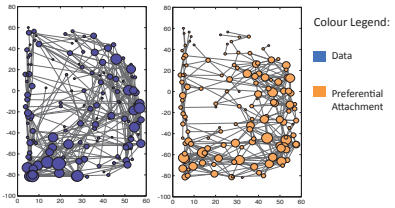

Colour Legend:
Data
Exponential Decay
Preferential Attachment

## The Economical Preferential Attachment Model (II)




Colour Legend: Data Exponential Decay Preferential Attachment

## The Economical Preferential Attachment Model (III)

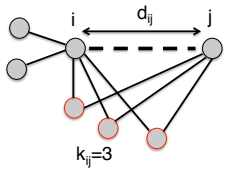


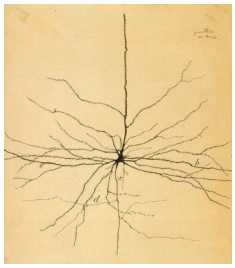

## How to Model Networks?

1. **Make a list of stylized facts**
2. **Come up with simplest mechanism leading to this structure**
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
5. **Think of what the model doesn't capture**
   - **Long distance connections**
   - **High C**

## The Economical Clustering Model

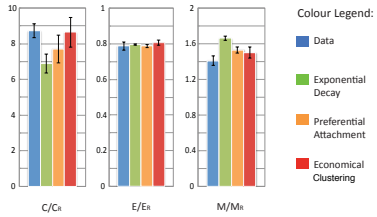
$$P_{ij} = d_{ij}^{-\eta} k_{ij}^{\gamma}$$



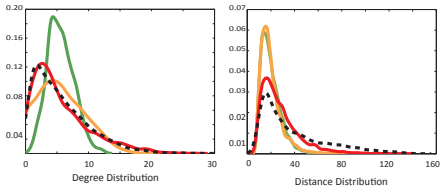



Vertes et al. (2012) *PNAS*

## The Economical Clustering Model (I)

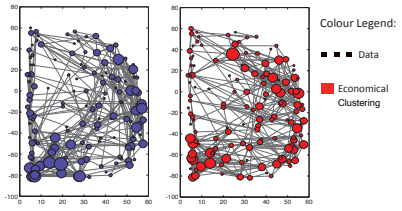



## The Economical Clustering Model (II)

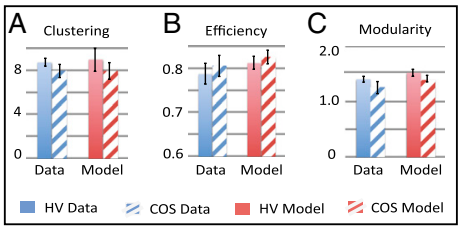




## The Economical Clustering Model (III)



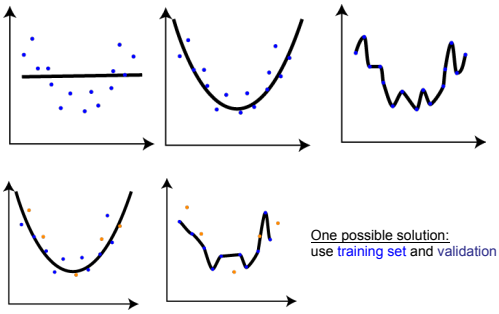
## Modeling Childhood Onset Schizophrenia (same data on different system)




## How to Model Networks?

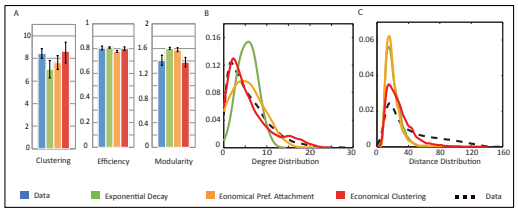
1. **Make a list of stylized facts:**
   - What is special or interesting about the structure we observe?
   - Is this structure real and is it really interesting? (Null models!)
2. **Come up with simplest mechanism leading to this structure**
   - Use some domain knowledge
   - Ignore all the details (at least in the beginning)
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
   - Same data about different systems
   - Different data about the same system
5. **Think of what the model doesn't capture**
   - Improve the model

## The problem of overfitting



One possible solution:
use training set and validation set

## Validation on Independent Dataset (same data on different system)




## How to Model Networks?

1. **Make a list of stylized fact:**
   - What is special or interesting about the structure we observe?
   - Is this structure real and is it really interesting? (Null models!)
2. **Come up with simplest mechanism leading to this structure**
   - Use some domain knowledge
   - Ignore all the details (at least in the beginning)
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
   - Different data about the same system
   - Same data about different systems
5. **Think of what the model doesn't capture**
   - Improve the model

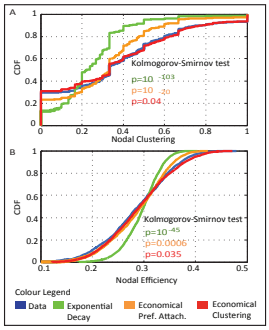
## SUMMARY 2.

1. **Make a list of stylized fact:**
   - What is special or interesting about the structure we observe?
   - Is this structure real and is it really interesting?
2. **Come up with simplest mechanism leading to this structure**
   - Use some domain knowledge
   - Ignore all the details (at least in the beginning)
3. **Fit the model to the data, compare to other models**
4. **Validate the model on independent data**
   - Same data about different systems
   - Different data about the same system
5. **Think of what the model doesn't capture**
   - Improve the model

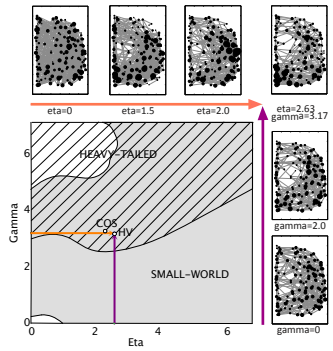
# Extra Slides

## Validate on Unconstrained Characteristics (different data on same system)

1. Distance distribution
2. Degree distribution can be left unconstrained
3. Nodal Measures

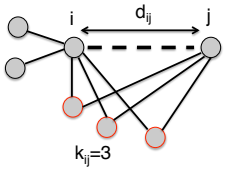



## Phase Diagram




## Interpreting the model

$$P_{ij} = d_{ij}^{-\eta} k_{ij}^{\gamma}$$




1. What does it mean?
   - Embodies trade-off between cost and topology
2. Are the parameters biologically plausible?
   - Cost of wiring
   - Hebbian analogy
   - Some experimental support at the neuronal level
3. Is the mechanism biologically plausible?
   - It is a local process