ECCS WARM-UP

School on Complex Networks, Sept 13-15

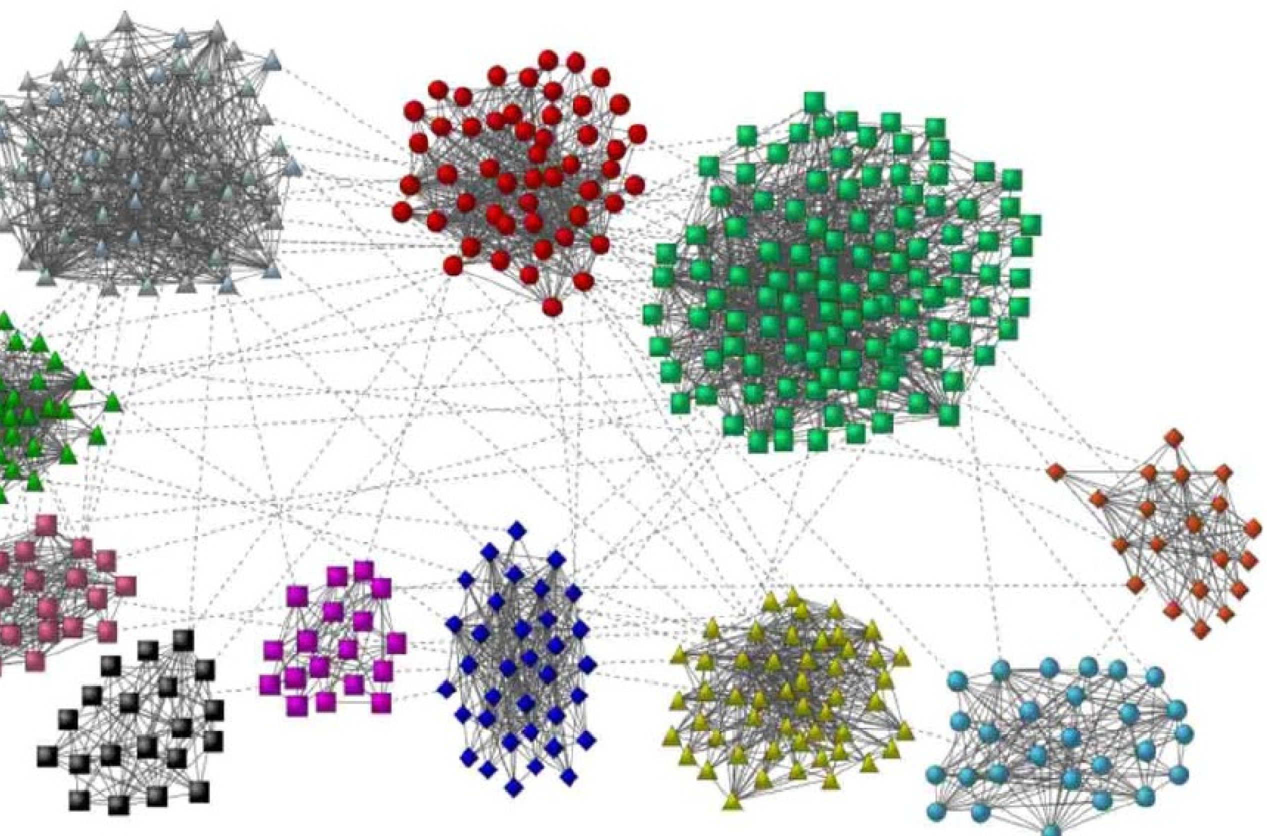
Marta Sales-Pardo

Department of Chemical Engineering,
Universitat Rovira i Virgili

From network modules to network inference

ECCS 2013
Barcelona

September 12, 2013

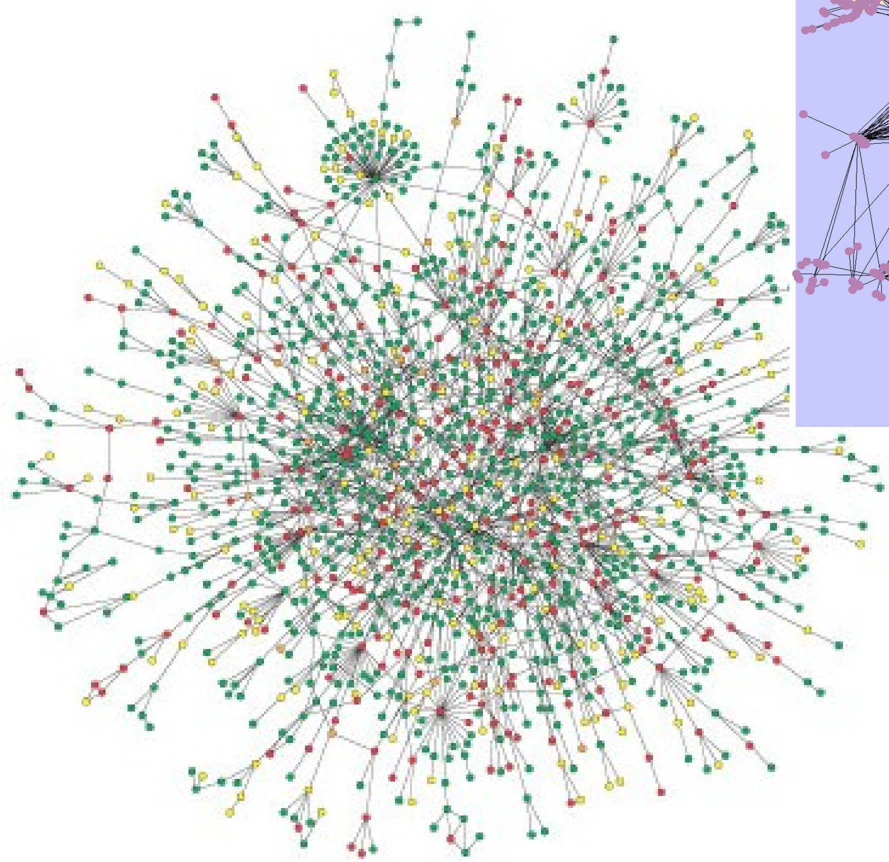


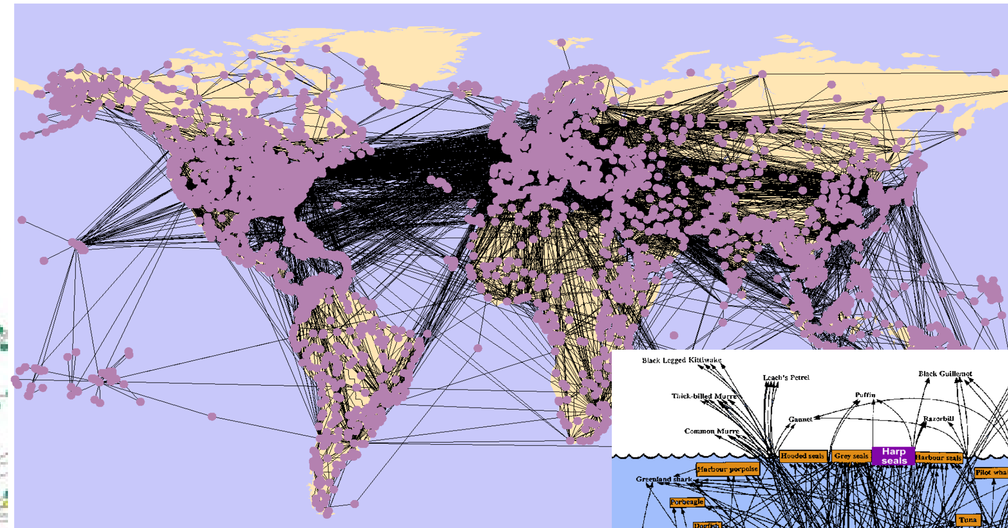
UNIVERSITAT ROVIRA I VIRGILI

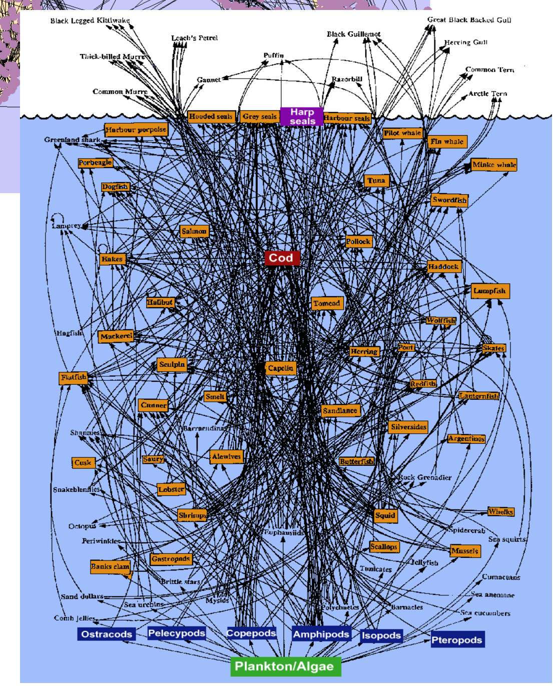
ECCS WARM-UP

School on Complex Networks, Sept 13-15

The promise of networks research





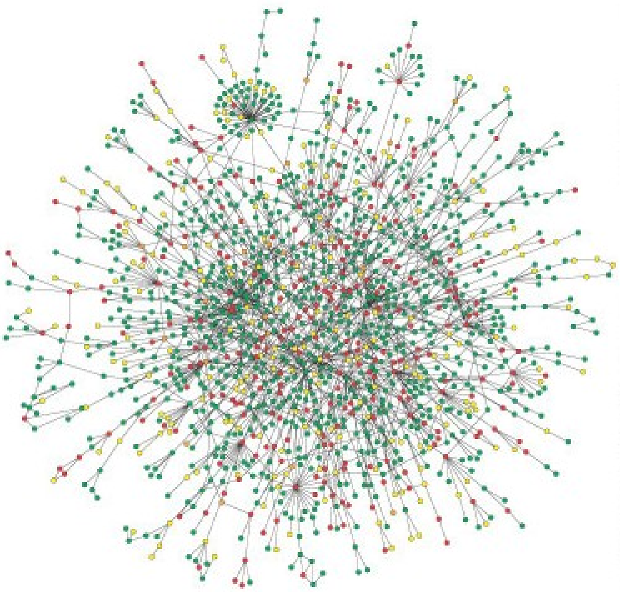


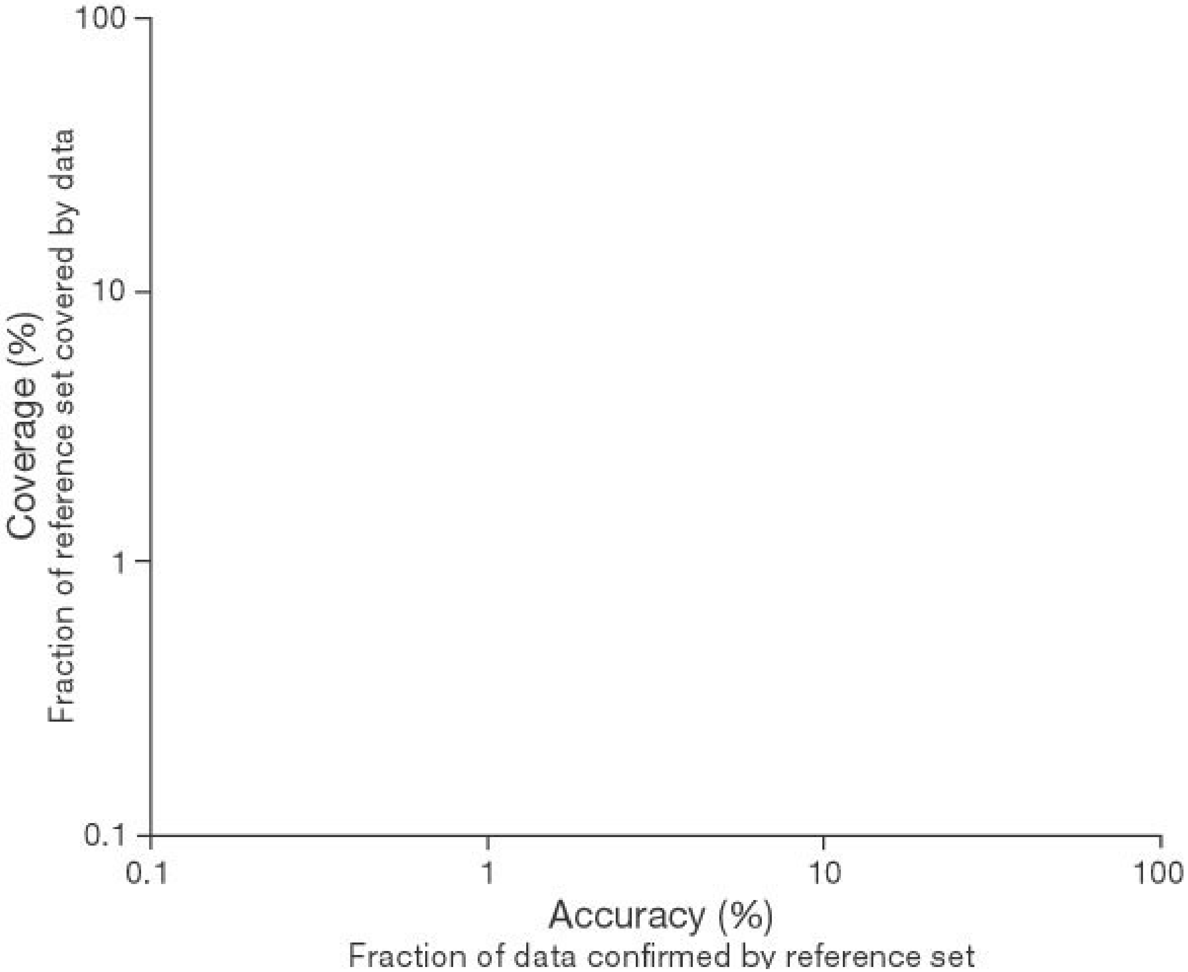
➔ **What can we *learn* about a system by studying the topology of the corresponding interaction network?**

ECCS WARM-UP

School on Complex Networks, Sept 13-15

Challenge #1: There is much about the interactions in the networks we study that we don't know



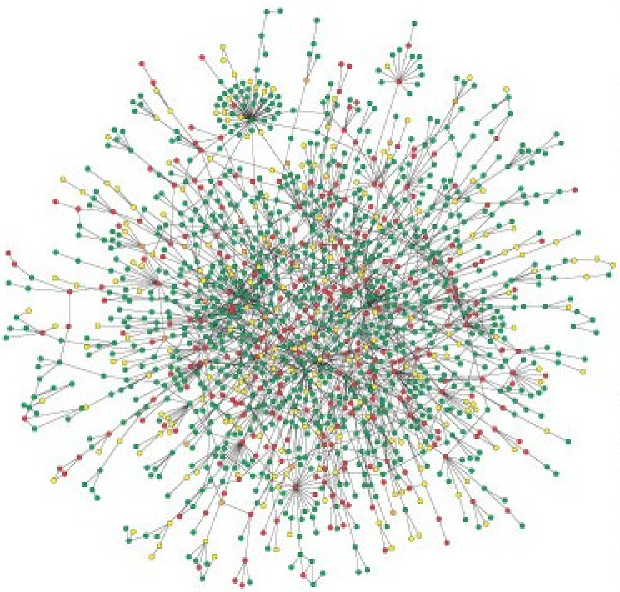


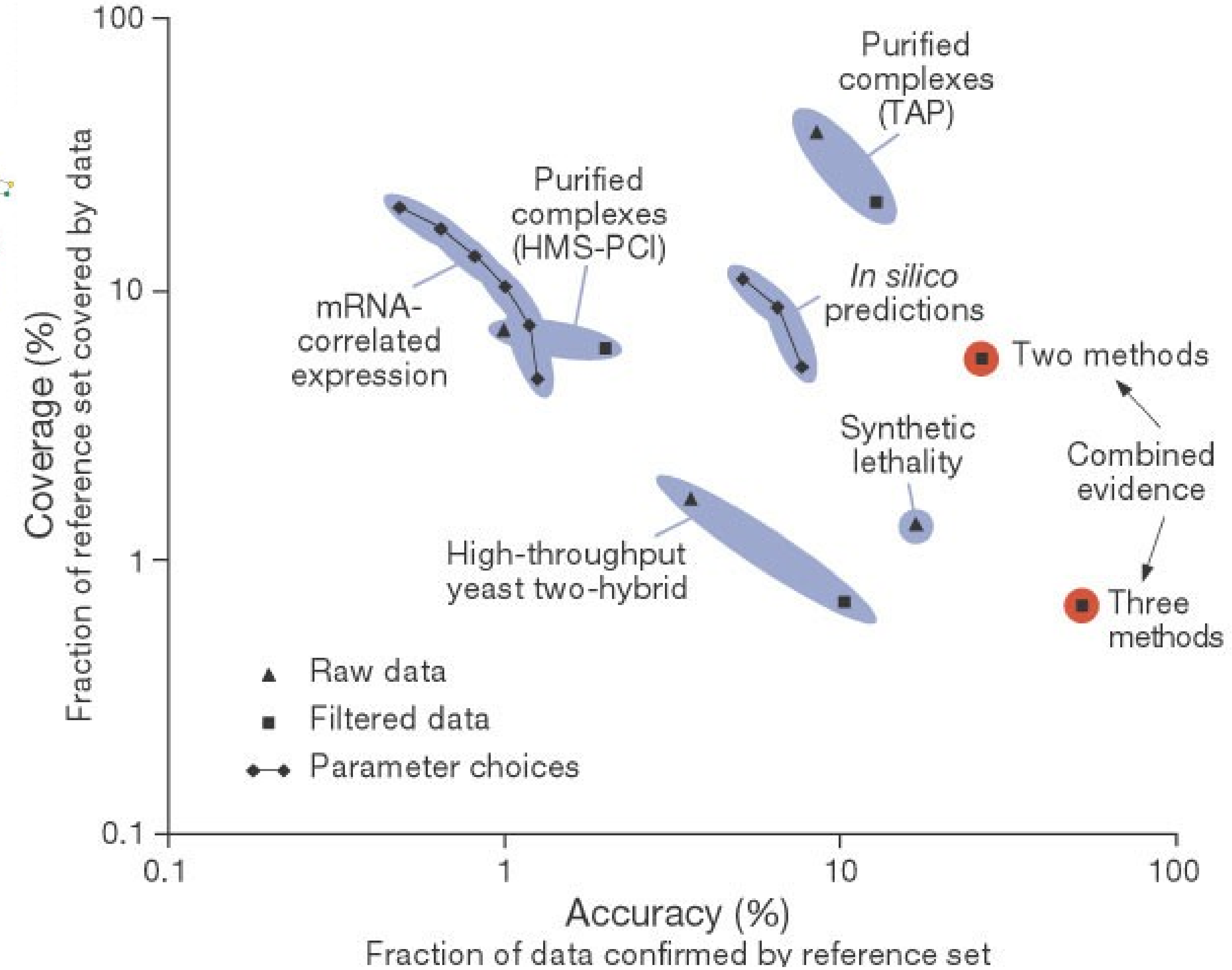
von Mering et al., *Nature* (2002)

ECCS WARM-UP

School on Complex Networks, Sept 13-15

Challenge #1: There is much about the interactions in the networks we study that we don't know



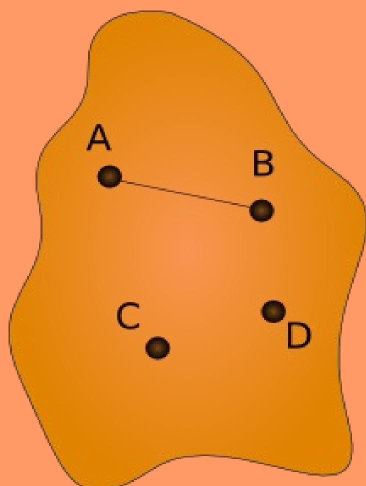
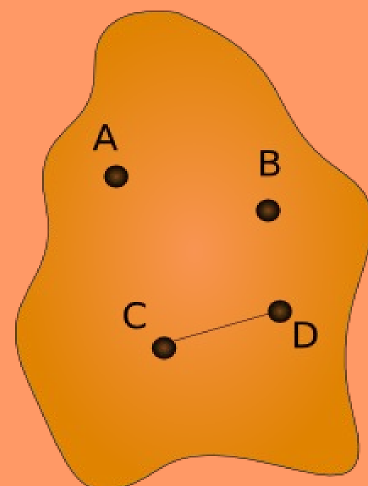


von Mering et al., *Nature* (2002)

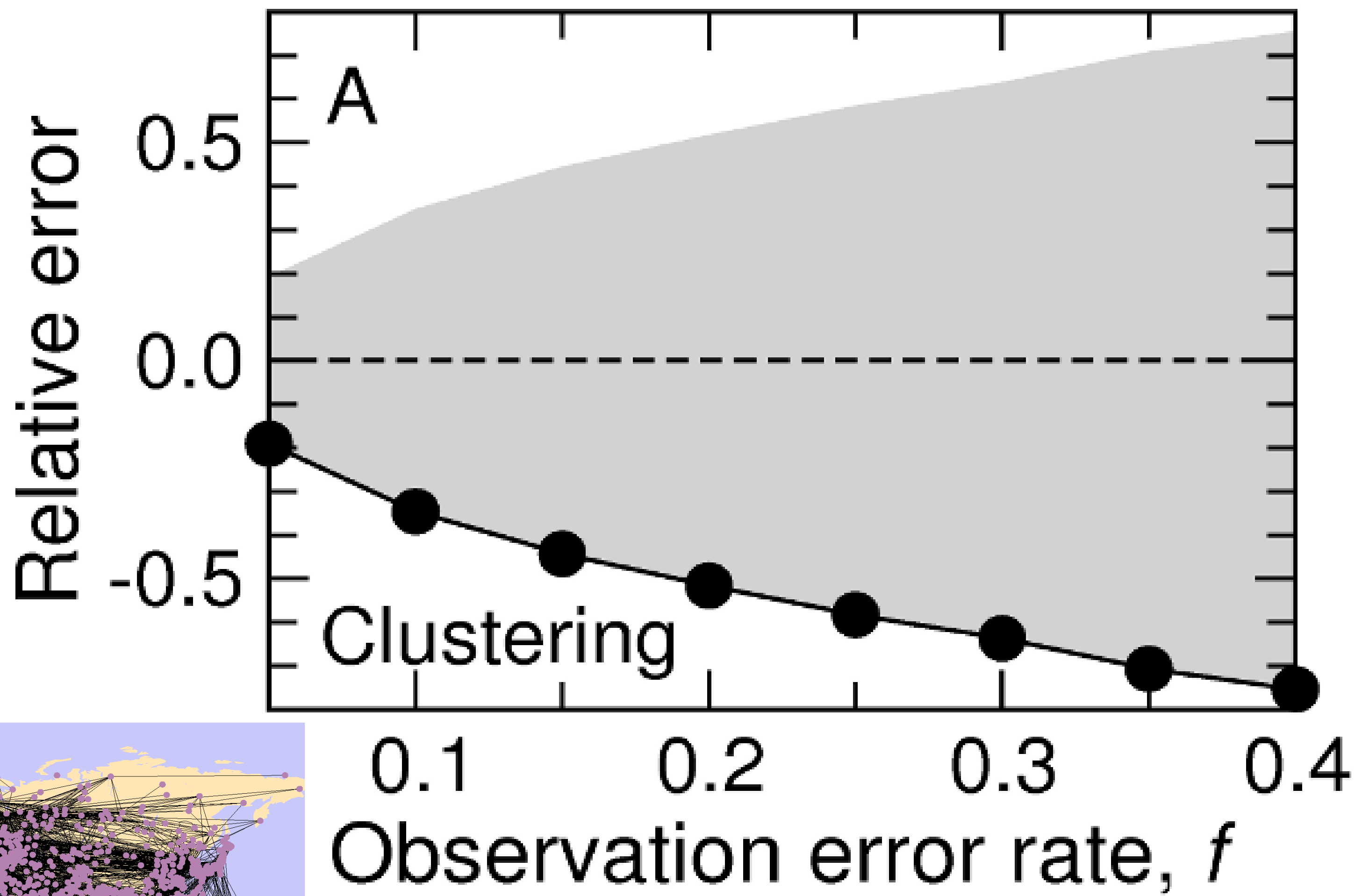
ECCS WARM-UP

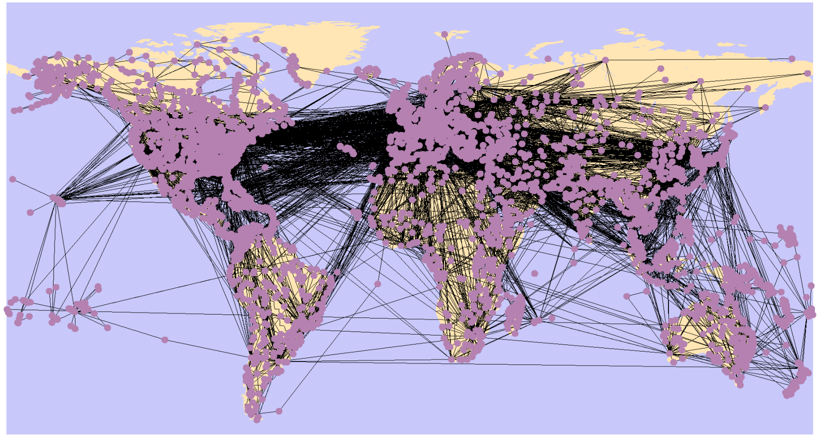
School on Complex Networks, Sept 13-15

We can test what is the effect of random errors in our network observations

	True network	Observed network	Test
Random errors			How do network properties change?

Network properties are often sensitive to even low error rates

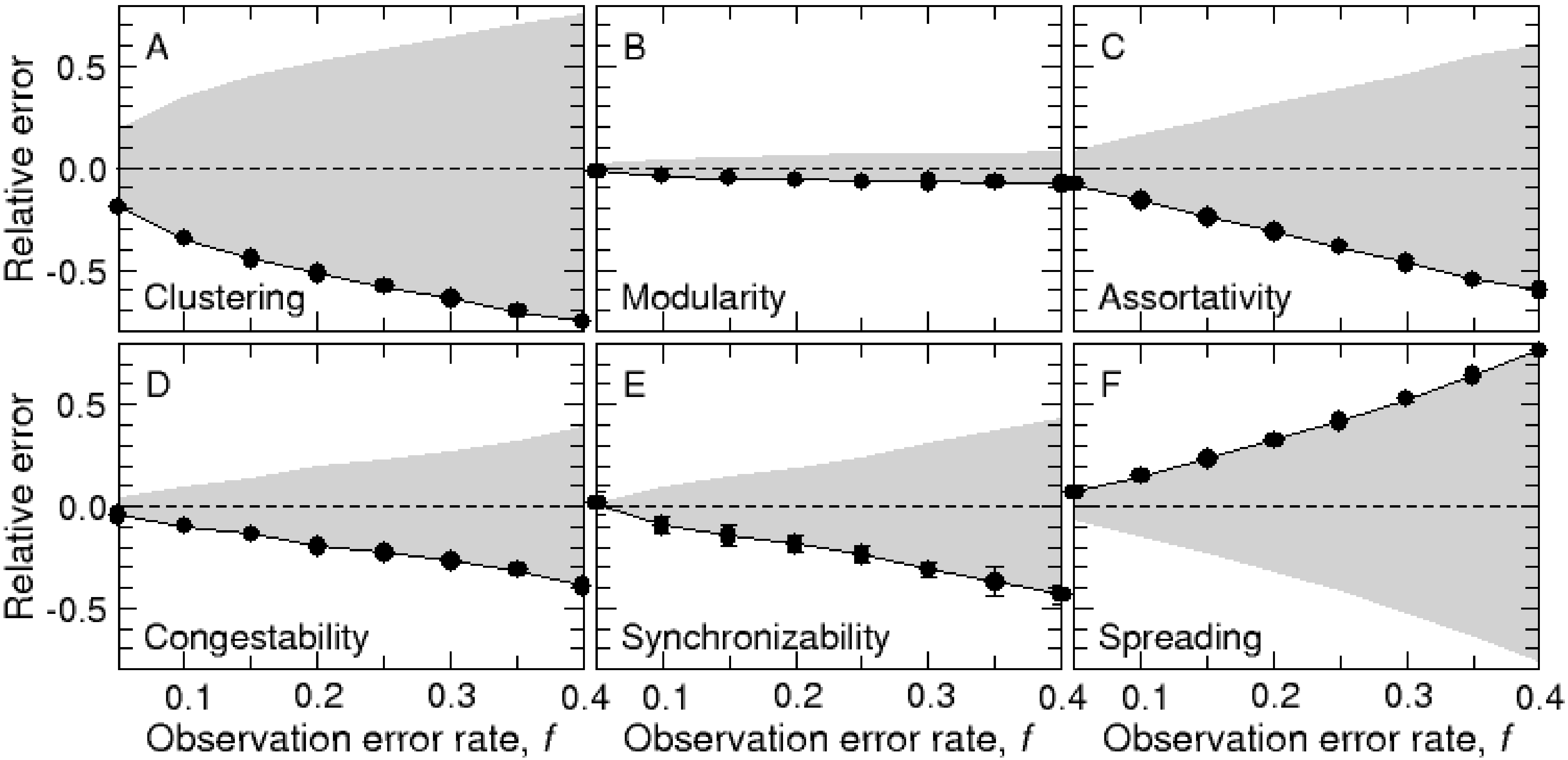




ECCS WARM-UP

School on Complex Networks, Sept 13-15

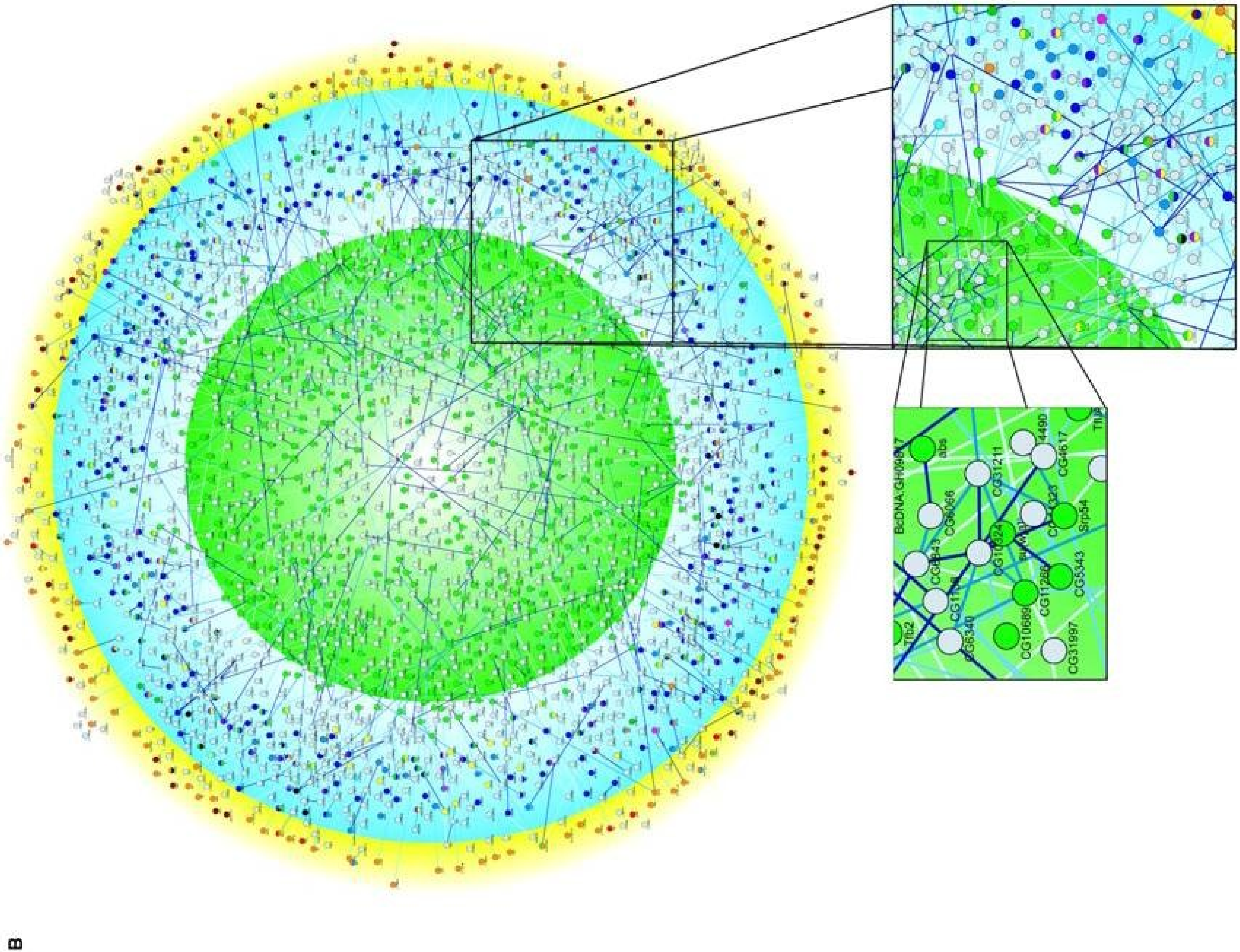
Network properties are often sensitive to even low error rates



ECCS WARM-UP

School on Complex Networks, Sept 13-15

Challenge #2: From data to knowledge (learning)



Giot *et al.*, *Science* (2003)

ECCS WARM-UP

School on Complex Networks, Sept 13-15

With the same tools we can predict if combining two drugs poses a risk to health...

ECCS WARM-UP

...or whether you are going to like "The Dark Knight rises"!

Outline

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message

ECCS WARM-UP

School on Complex Networks, Sept 13-15

We need a “cartography” of complex networks

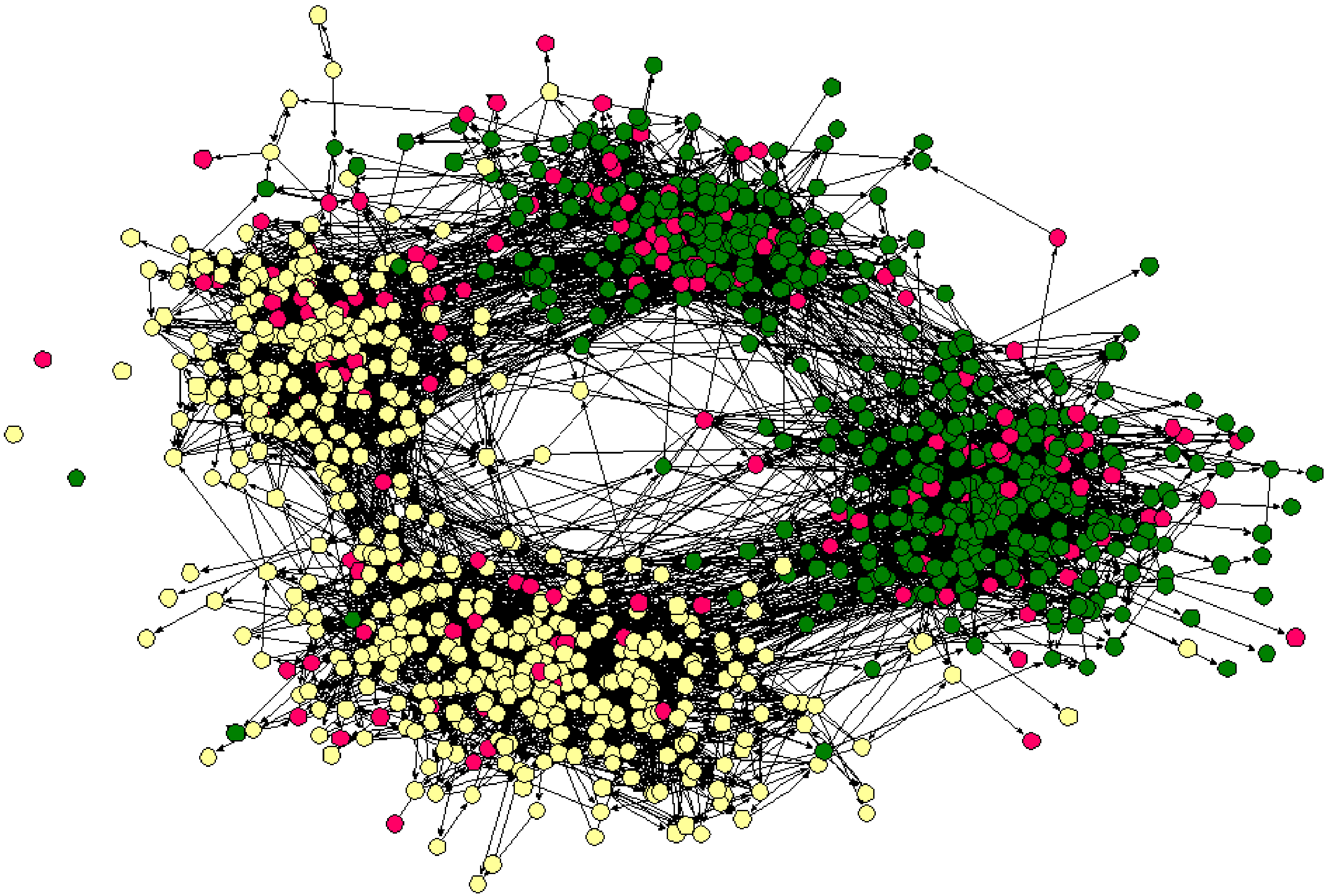
➔ Modules

- ➔ We divide the system into “regions”

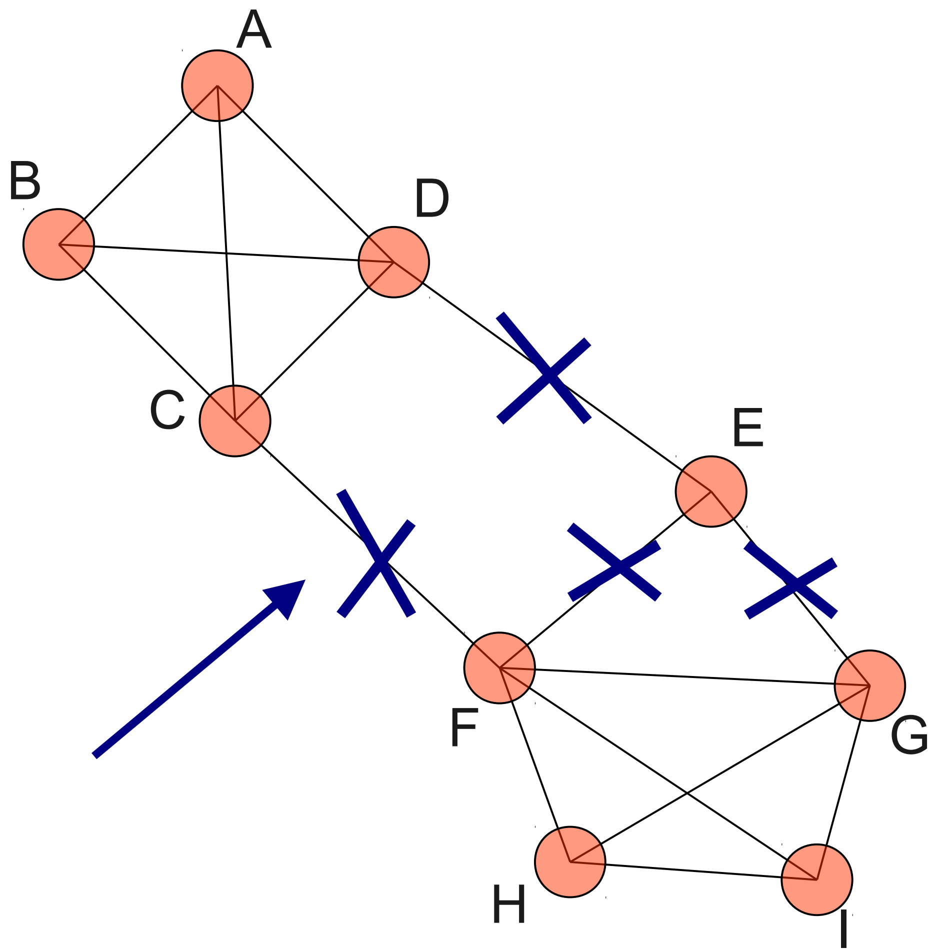
ECCS WARM-UP

School on Complex Networks, Sept 13-15

Densely connected groups of nodes (modules or communities) are good candidates for our "regions"



Source: http://www-personal.umich.edu/~mejn/networks/school.gif



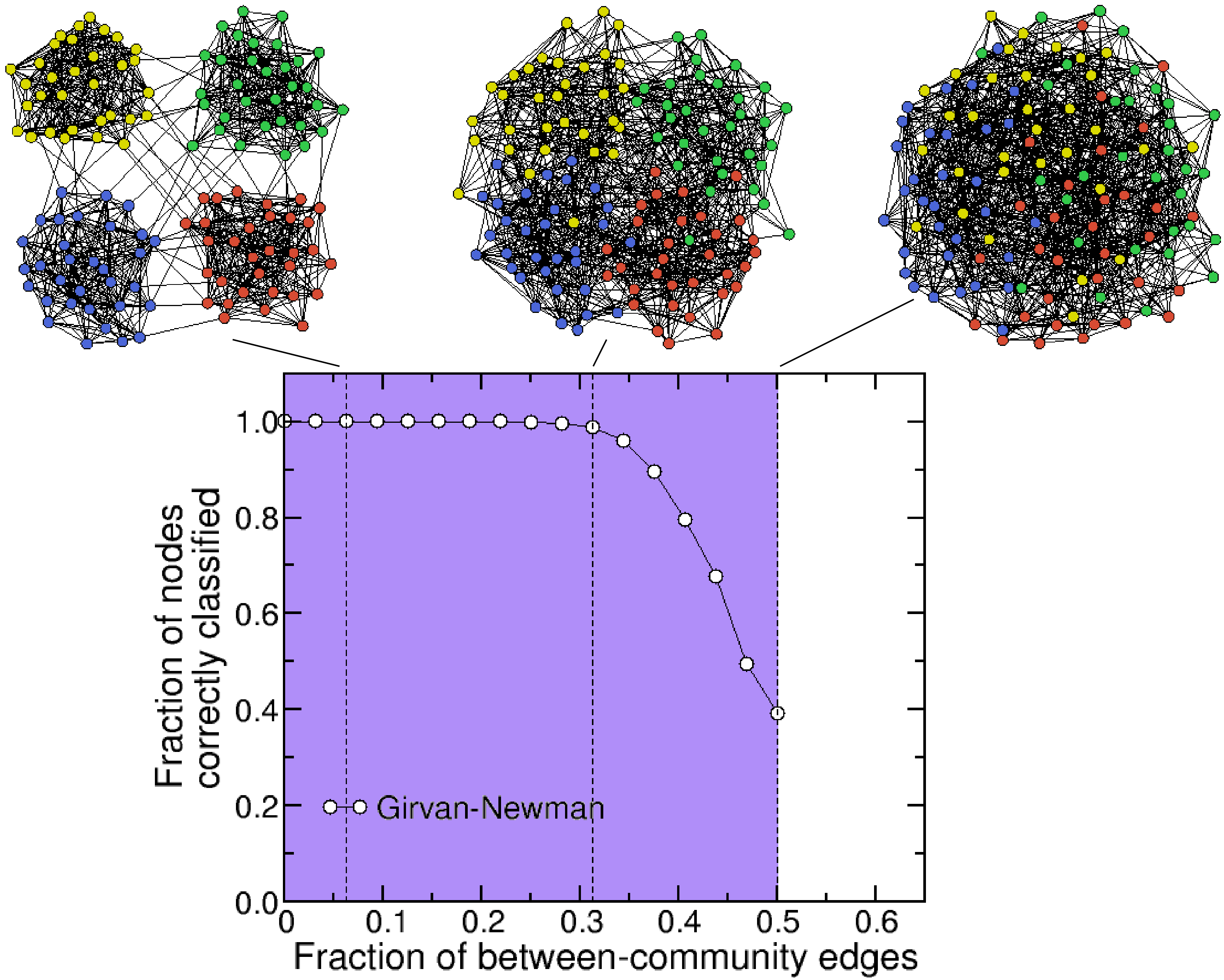
- Identify the most *central* edge in the network
- Remove the most central edge in the network
- Iterate the process

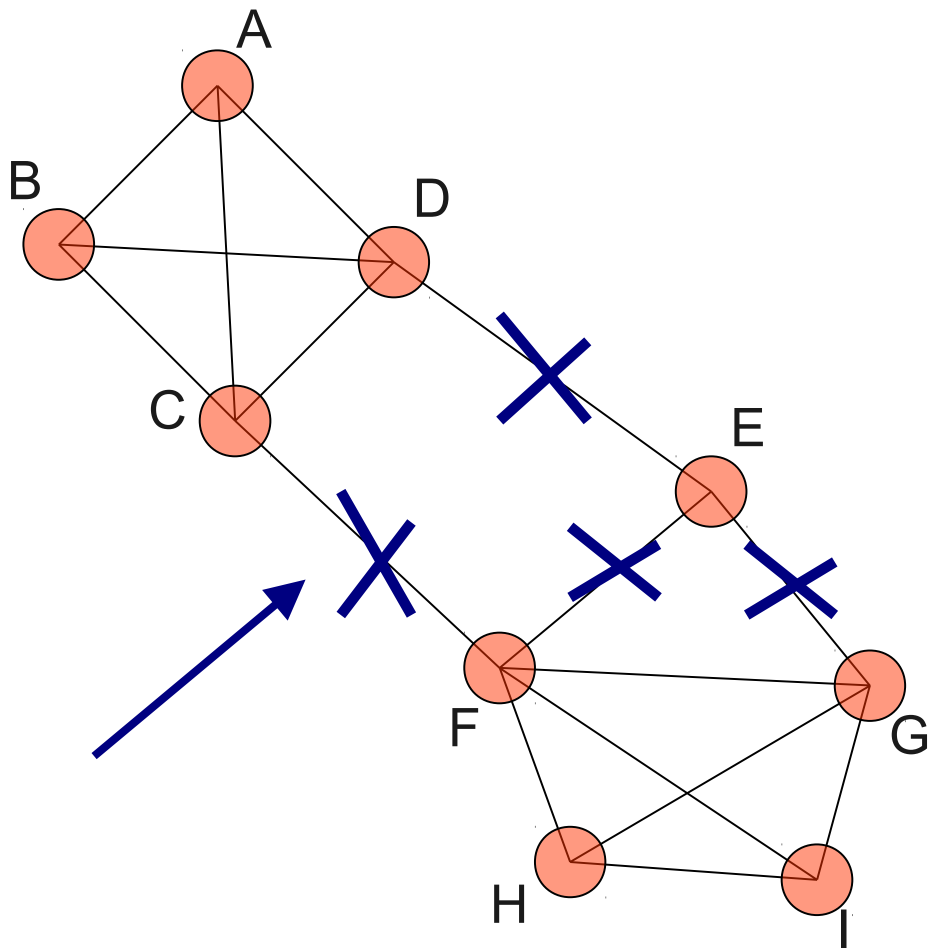
Girvan & Newman, *PNAS* (2002)

ECCS WARM-UP

School on Complex Networks, Sept 13-15

We can evaluate the performance of the Girvan-Newman algorithm using model network with known communities





- Identify the most *central* edge in the network
- Remove the most central edge in the network
- Iterate the process

PROBLEM

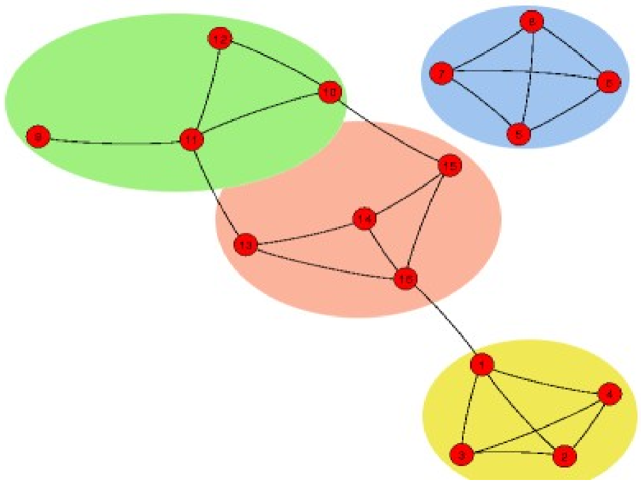
When do we stop?

Girvan & Newman, *PNAS* (2002)

ECCS WARM-UP

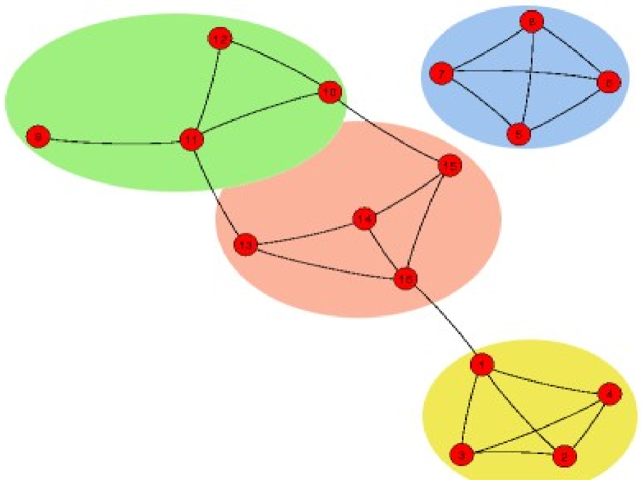
A quantitative measure of *modularity*

School on Complex Networks, Sept 13-15

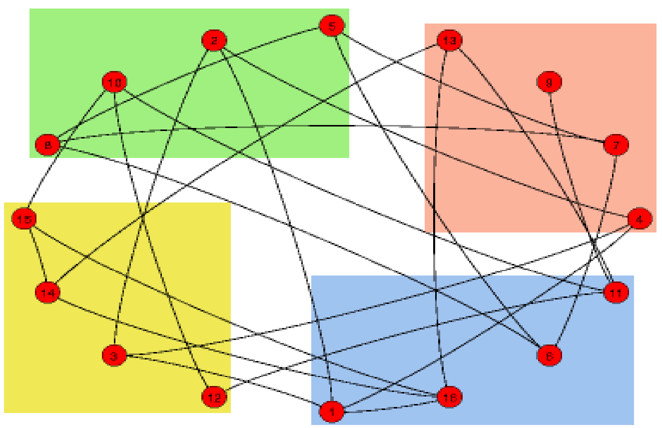


f_s: fraction of links within module s

A quantitative measure of *modularity*



f_s: fraction of links within module s



F_s: expected fraction of links within module s, for a random partition of the nodes

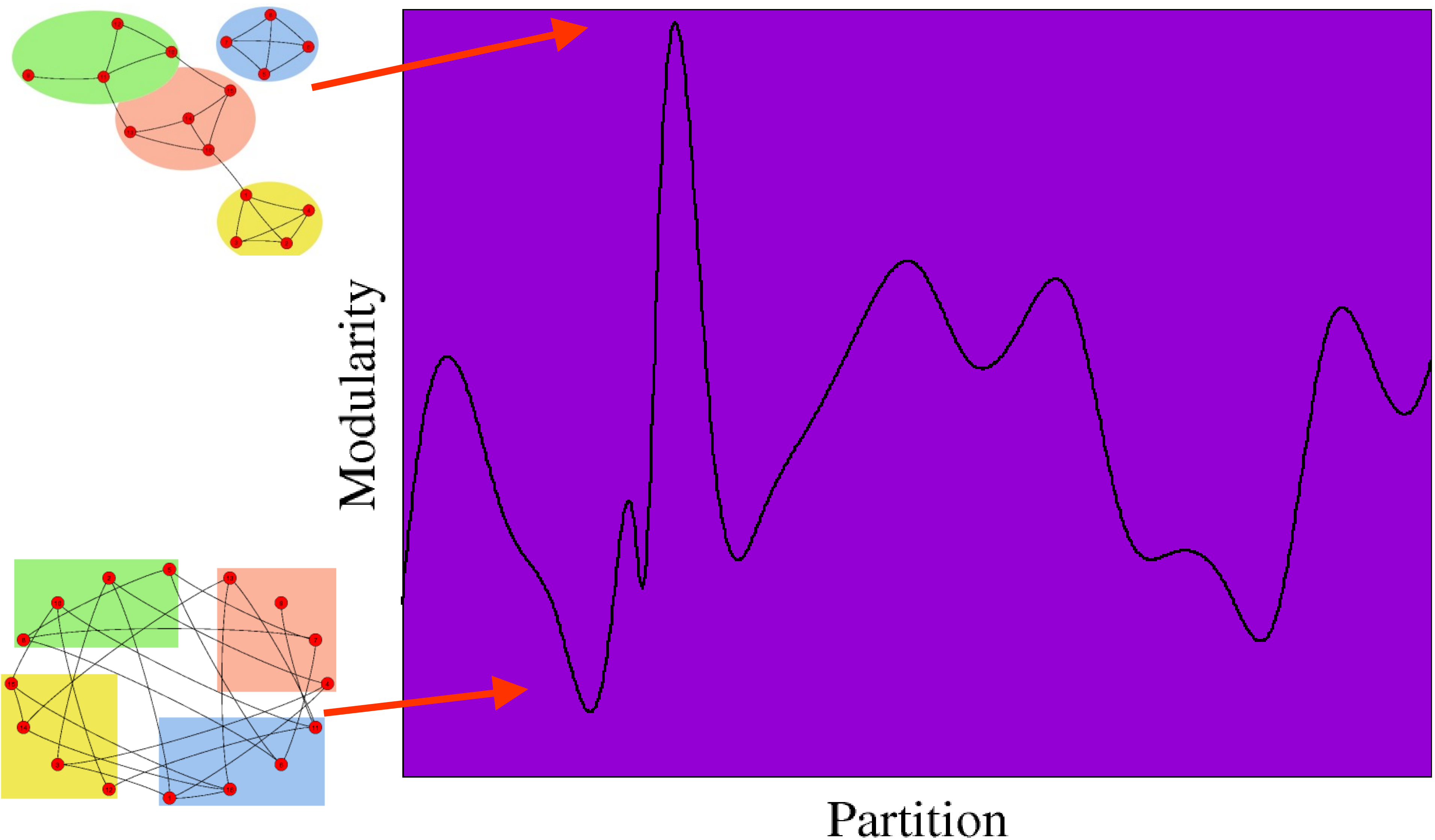
Modularity of a partition: $M = \sum_s (f_s - F_s)$

Newman & Girvan, *PRE* (2003)

ECCS WARM-UP

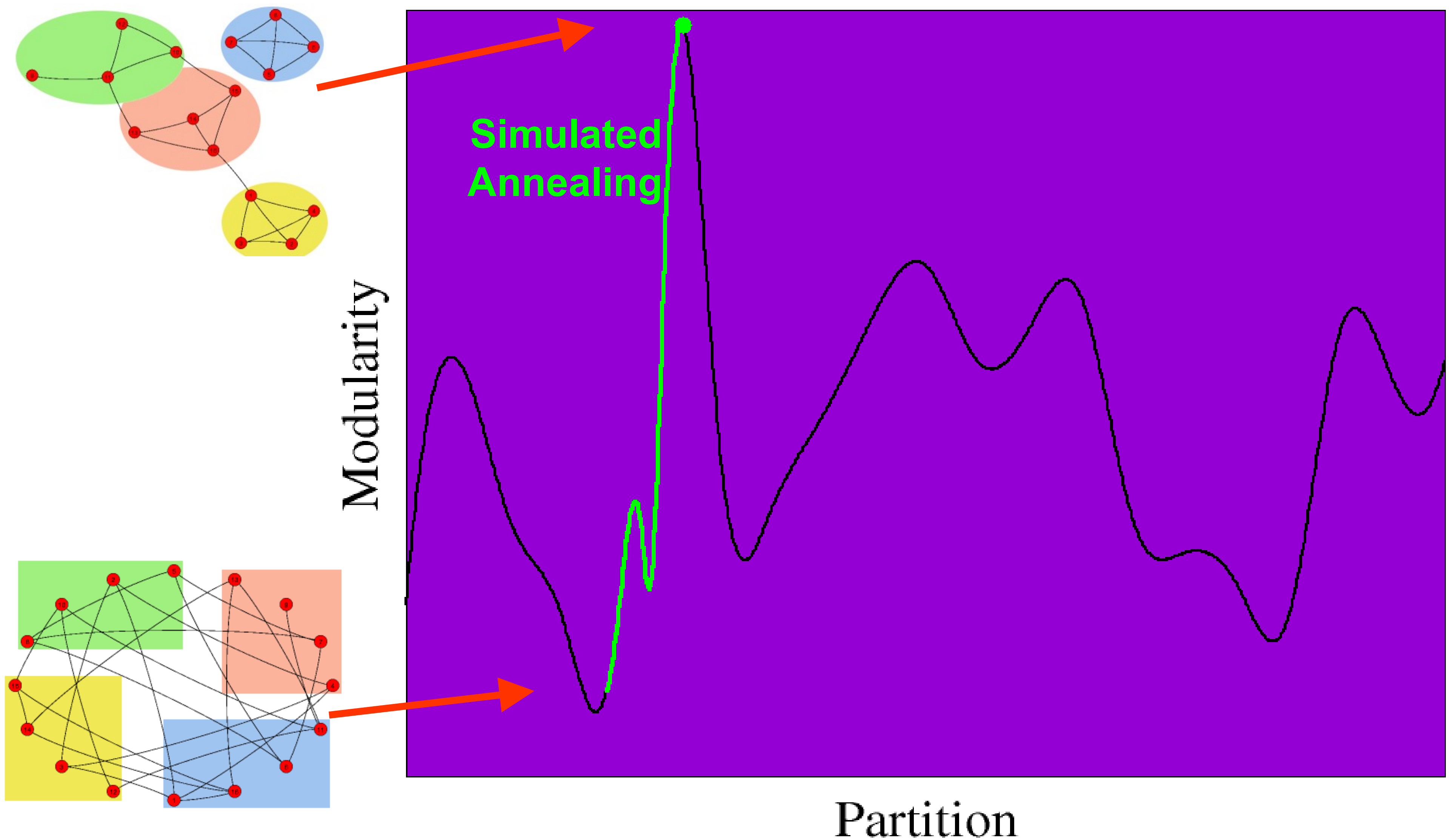
School on Complex Networks, Sept 13-15

Finding the maximum modularity is a difficult (NP-complete) combinatorial optimization problem



Guimera, Sales-Pardo & Amaral, *PRE* (2004); Guimera & Amaral, *Nature* (2005)

We use *simulated annealing* to obtain the partition with largest modularity

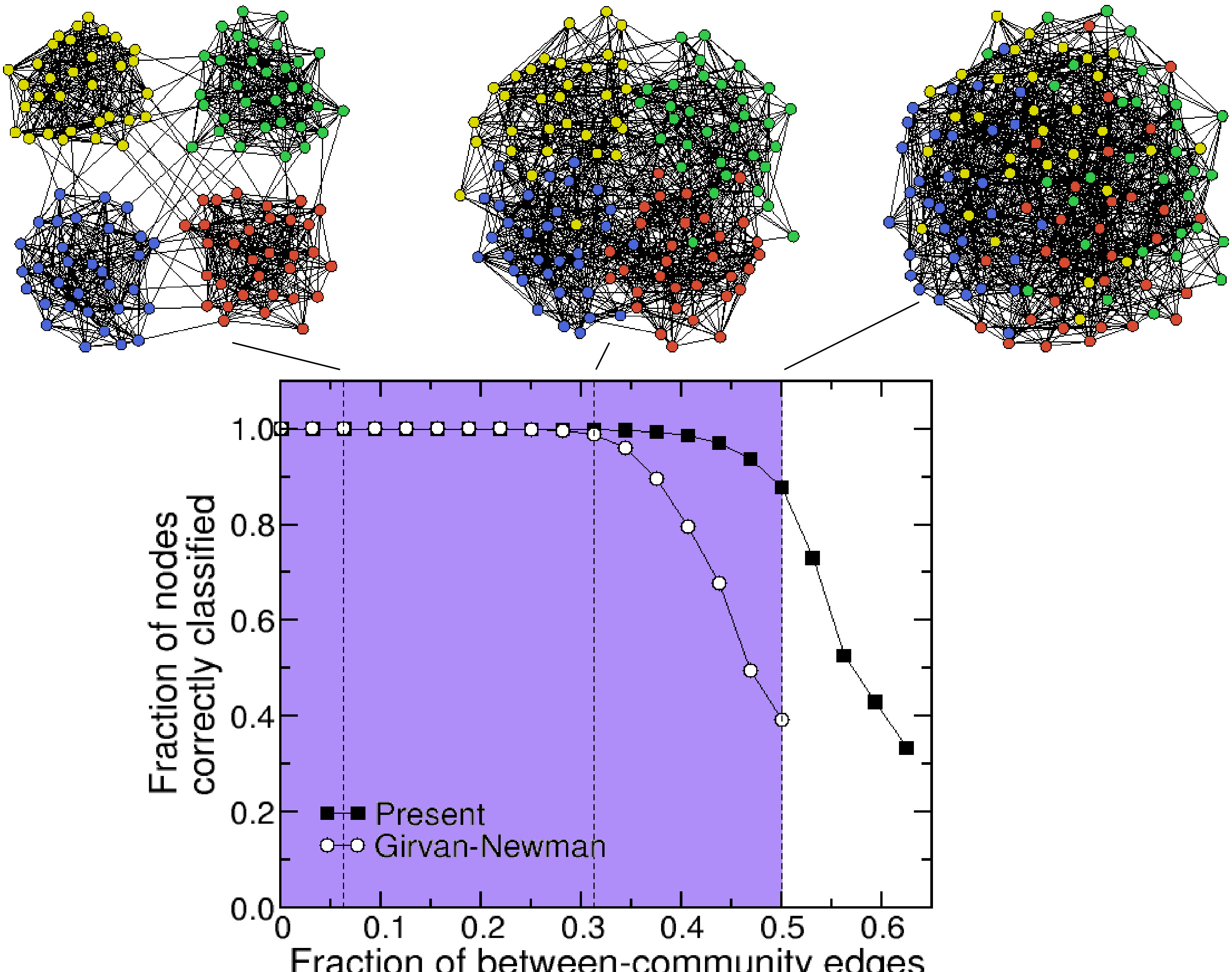


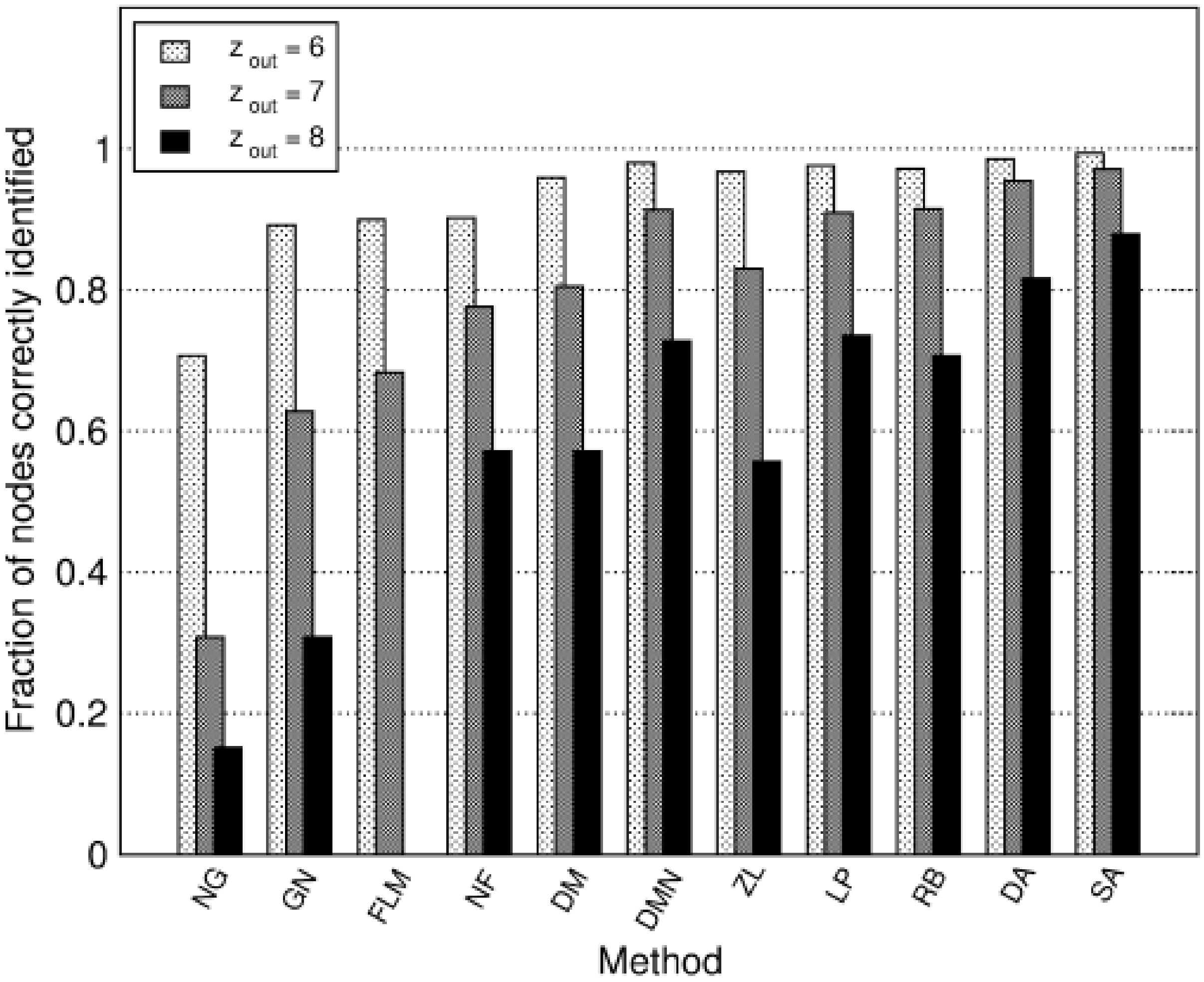
Guimera, Sales-Pardo & Amaral, *PRE* (2004); Guimera & Amaral, *Nature* (2005); Sales-Pardo et al. PNAS (2007).

ECCS WARM-UP

School on Complex Networks, Sept 13-15

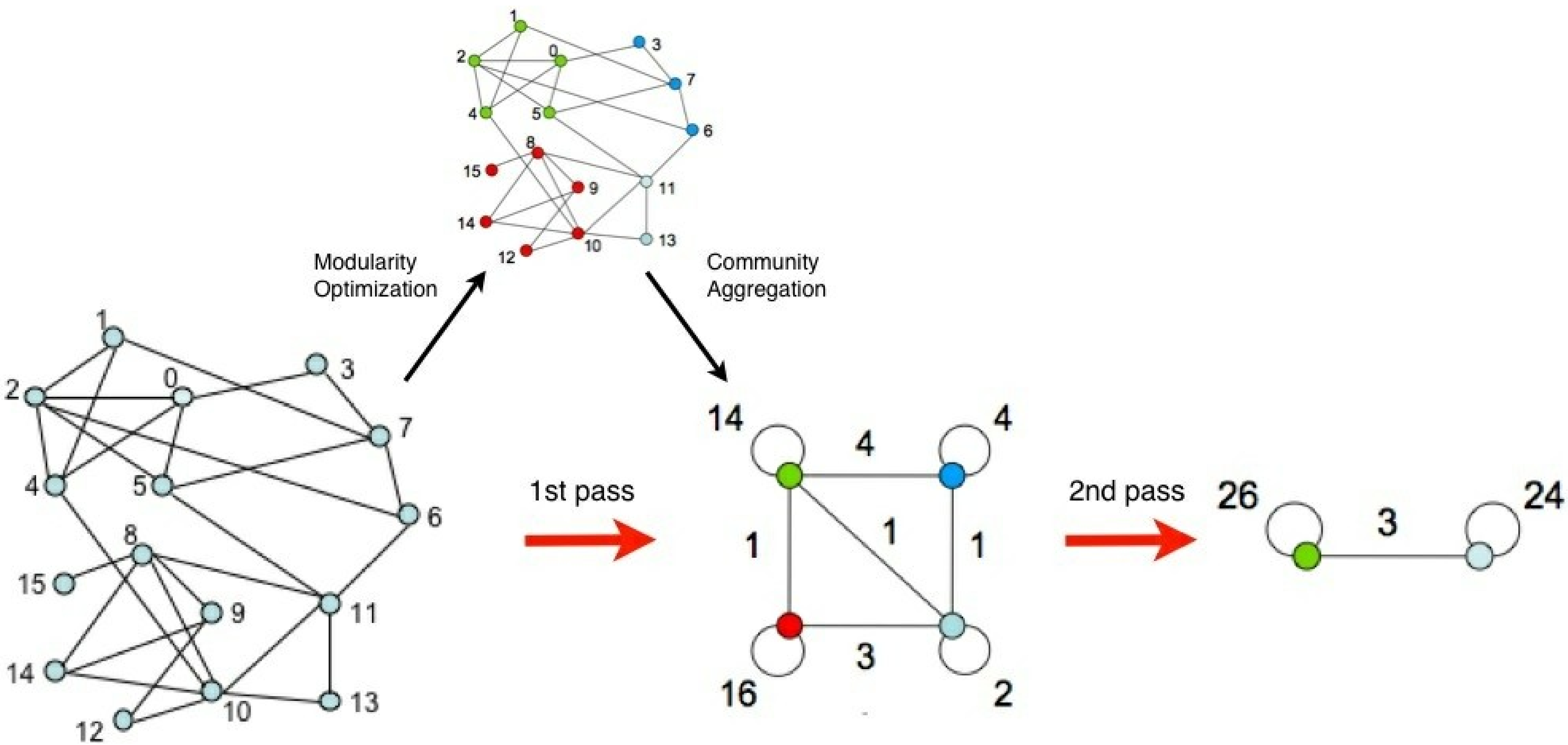
We can evaluate the performance of the Girvan-Newman algorithm using model network with known communities





Danon, Díaz-Guilera, Duch, Arenas, *JSTAT* (2005)

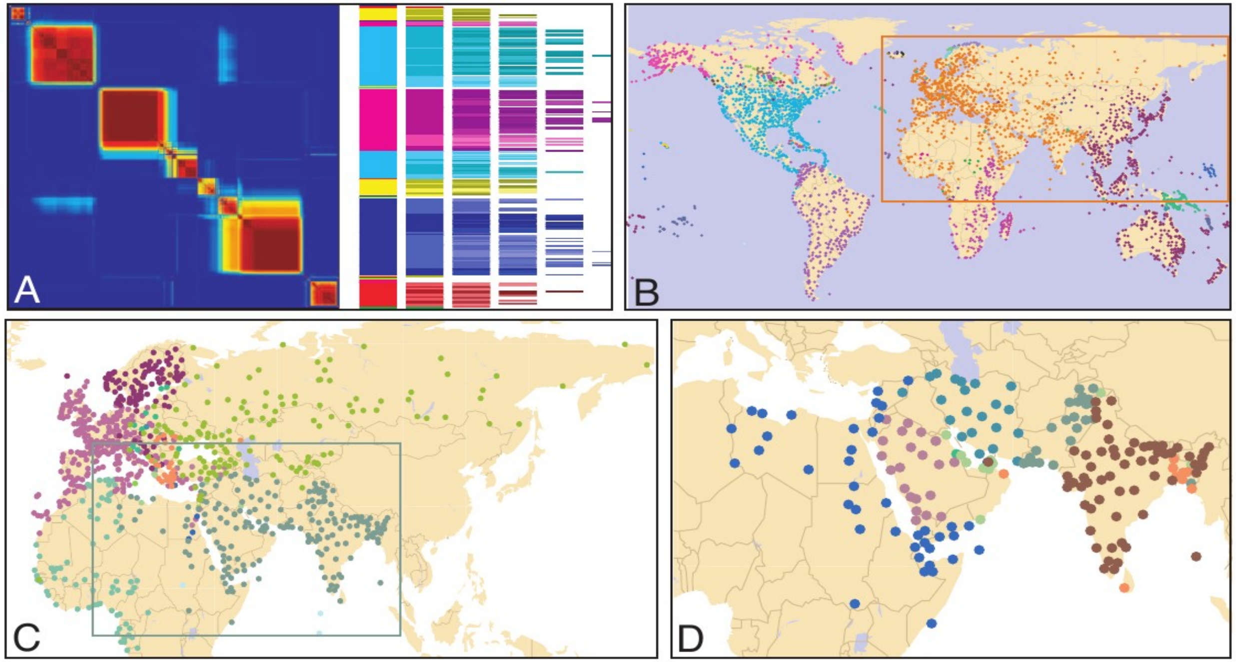
The "Louvain method" is a fast and quite accurate modularity-maximization method that works with multi-million node networks



Blondel, Guillaume, Lambiotte, Lefebvre, *JSTAT* (2008)

There are problems with modularity maximization

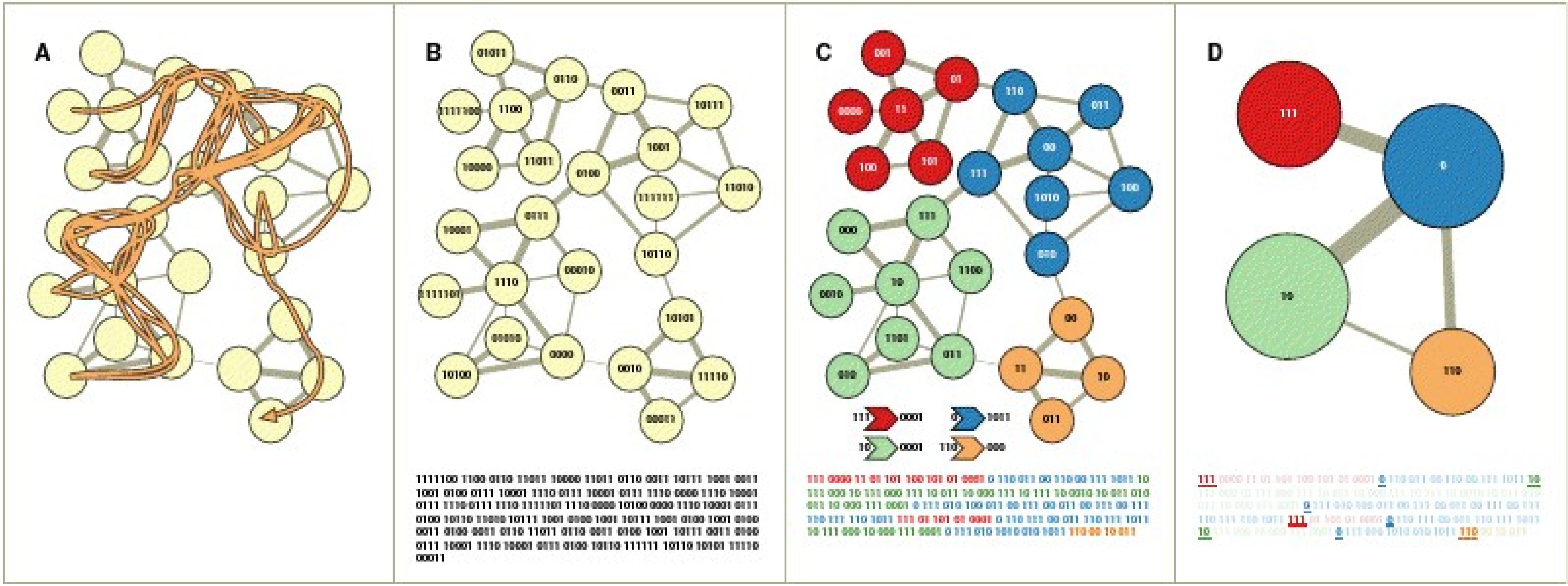
- Resolution limit: modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined (Fortunato, Barthelemy, *PNAS* 2006).
- Modular structure may be hierarchical (modules within modules) and modularity maximization only captures one scale (or, worse, a mixture of scales) (Sales-Pardo, Guimera, Moreira, Amaral, *PNAS* 2007).



ECCS WARM-UP

School on Complex Networks, Sept 13-15

Infomap is a very accurate algorithm not based on modularity maximization



Rosvall & Bergstrom, *PNAS* (2008)

There are some problems with the benchmark networks I have discussed so far

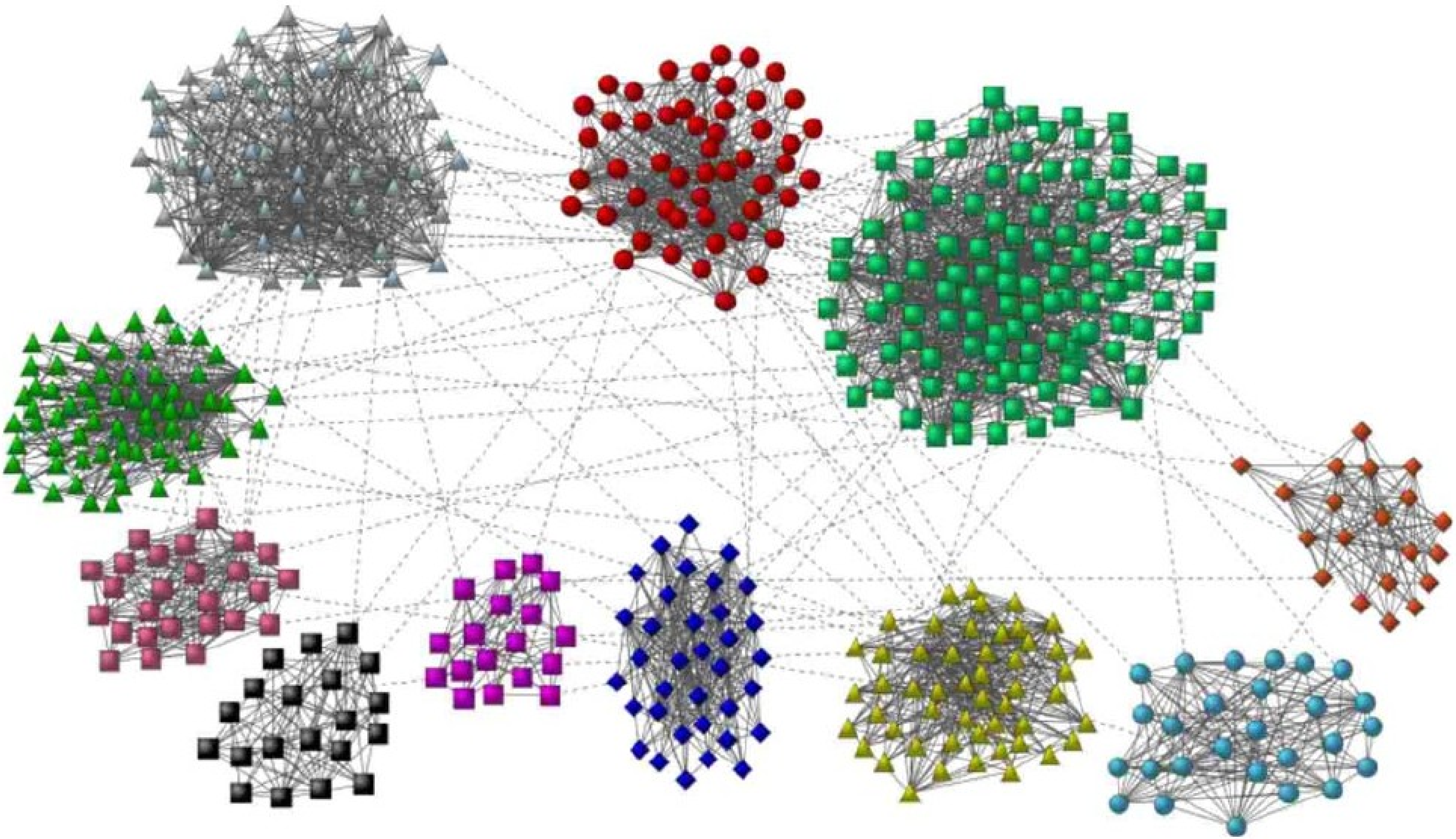
- Problems with the benchmark networks I have discussed so far:
  - All modules have the same size
  - All nodes in a module have more or less the same connections (Poisson degree distribution)

Lancichinetti, Fortunato & Radicchi, *PRE* (2008)

ECCS WARM-UP

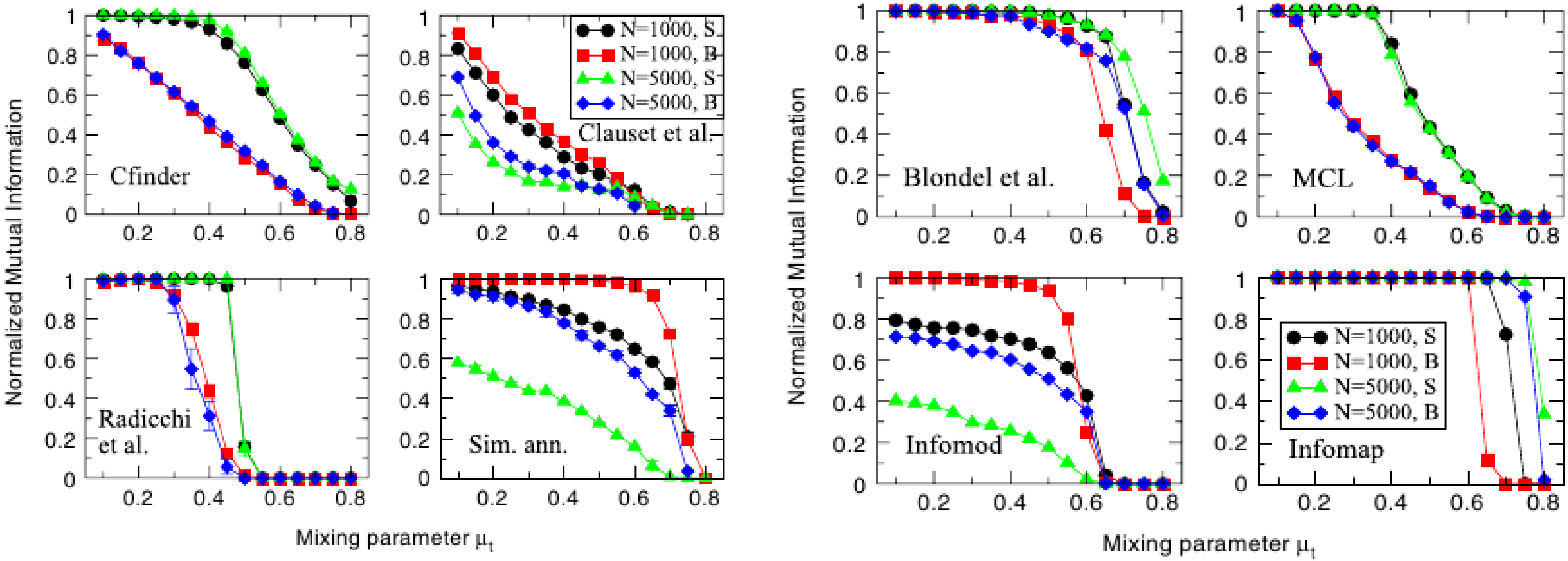
School on Complex Networks, Sept 13-15

LFR benchmark networks have broad community size and degree distributions



Lancichinetti, Fortunato & Radicchi, *PRE* (2008)

Algorithm comparison on LFR benchmark networks



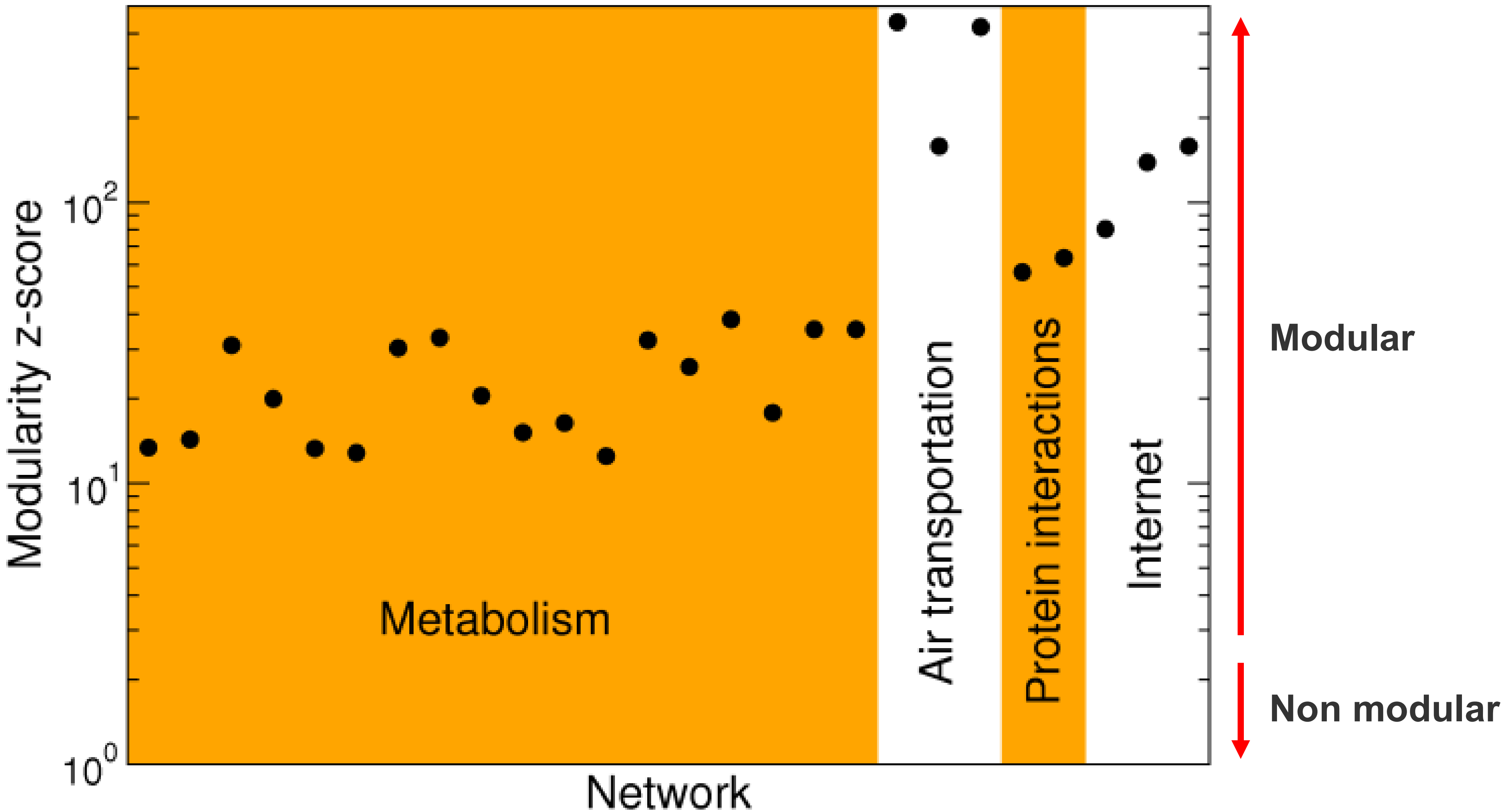
Fortunato, *Phys. Rep.* (2010)

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message

- Problem: If you look for modules, you find them (even in purely random graphs!!)
- Solution:
  - Obtain the modularity *M* for the real network
  - Compare *M* to the distribution of modularities in an ensemble of random networks with the same degree sequence as the real network

Guimerà, Sales-Pardo, Amaral, *PRE* (2004)

All networks we have studied are significantly modular



Guimerà, Sales-Pardo, Amaral, *Nature Physics* (2007)

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message

ECCS WARM-UP

School on Complex Networks, Sept 13-15

Connectors that span several modules are often key for system-wide behavior

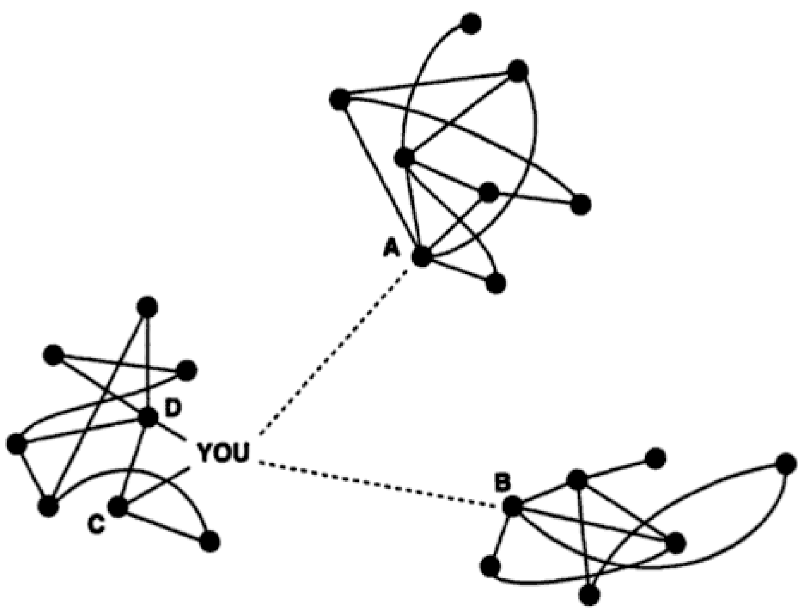
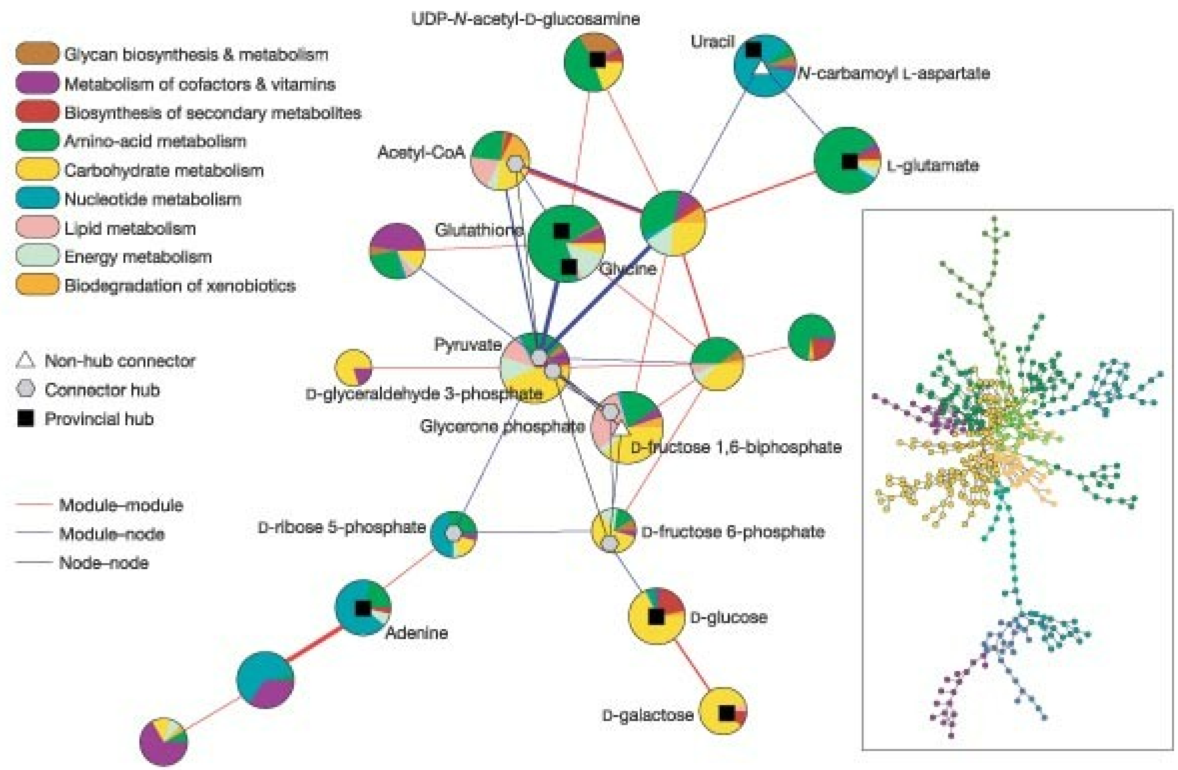


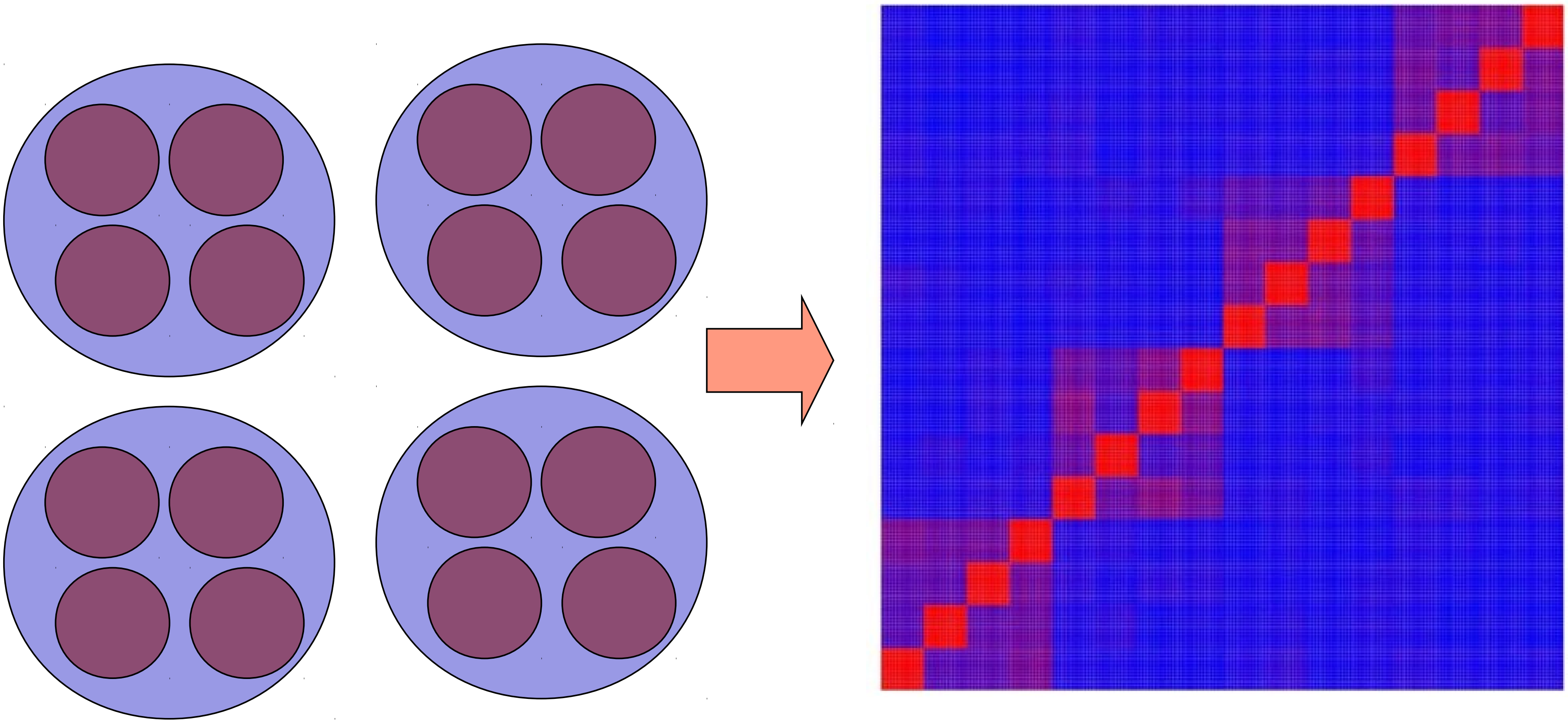
Figure 1.6 **Structural holes and weak ties**



Guimera, Amaral, *Nature* (2005)

Burt, *Structural Holes: The Social Structure of Competition* (1995)

The modular structure of a network determines its dynamic behavior

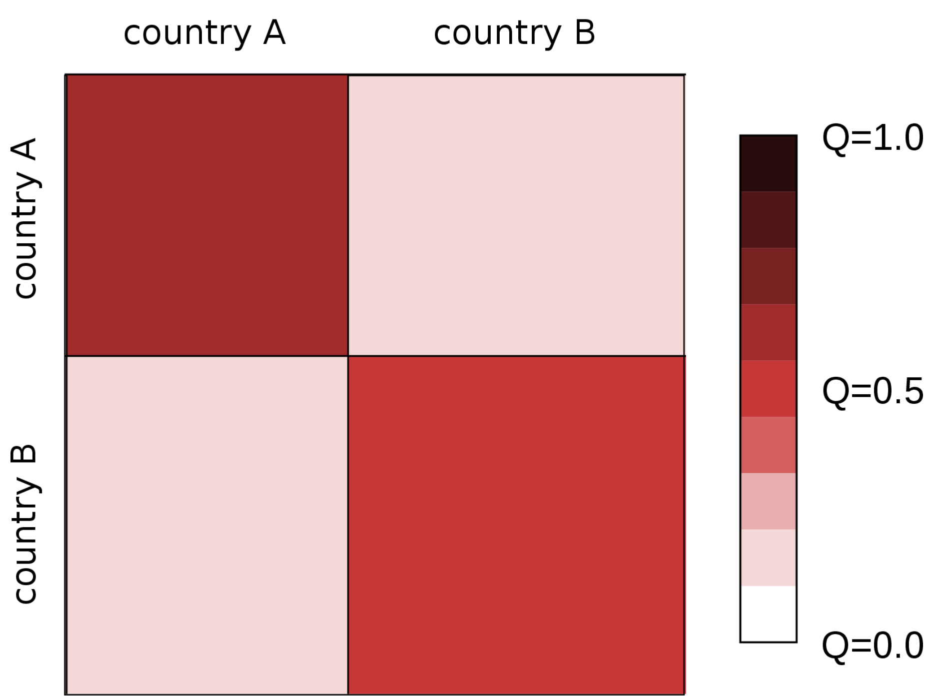


Arenas, Díaz-Guilera, Pérez-Vicente, *PRL* (2006)

Outline

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message

Modularity



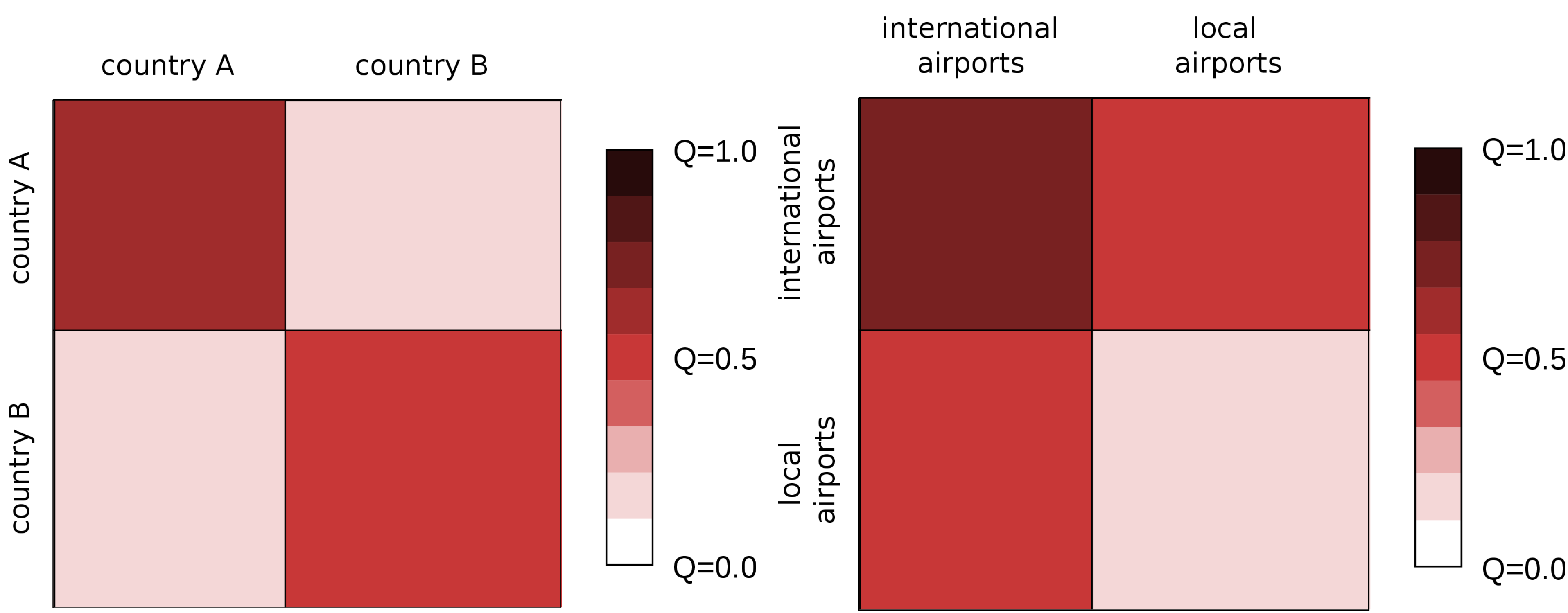
ECCS WARM-UP

School on Complex Networks, Sept 13-15

Stochastic block models are network models that account for modularity and other group-based features

Modularity

Role-to-role correlations



Outline

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
  - Hands-on: module-identification algorithms
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message and more hands-on

- Suppose that A and B are two “events”:
  - p(A,B) is the probability of both events
  - p(A|B) is the probability of A given B
  - p(A) is the probability of event A “regardless of B”

- We have that

$$p(A,B) = p(A|B)p(B)$$

$$p(B,A) = p(B|A)p(A)$$

- But since p(A,B)=P(B,A) we arrive at

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Bayes formula / rule / theorem

- Suppose that A and B are two "events":
  - p(A,B) is the probability of both events
  - p(A|B) is the probability of A given B
  - p(A) is the probability of event A "regardless of B"

- We have that

$$p(A, B) = p(A|B)p(B)$$

$$p(B, A) = p(B|A)p(A)$$

- But since p(A,B)=P(B,A) we arrive at

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- Suppose we have some data D and we want to be able to say something about a model M (estimate the parameters of the model, compare to other models, et c.)
- Using Bayes formula

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

- Since we (usually) only care about terms that depend on the model

$$p(M|D) \propto p(D|M)p(M)$$

- Suppose we have some data D and we want to be able to say something about a model M (estimate the parameters of the model, compare to other models, et c.)
- Using Bayes formula

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

- Since we (usually) only care about terms that depend on the model

$$p(M|D) \propto p(D|M)p(M)$$

Posterior
Plausibility of the Model given the Data

- Suppose we have some data D and we want to be able to say something about a model M (estimate the parameters of the model, compare to other models, et c.)
- Using Bayes formula

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

- Since we (usually) only care about terms that depend on the model

$$p(M|D) \propto p(D|M)p(M)$$

Posterior
Plausibility of the Model given the Data

Likelihood
Plausibility of the Data given the Model

- Suppose we have some data D and we want to be able to say something about a model M (estimate the parameters of the model, compare to other models, et c.)
- Using Bayes formula

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

- Since we (usually) only care about terms that depend on the model

$$p(M|D) \propto p(D|M)p(M)$$

Posterior
Plausibility of the Model given the Data

Likelihood
Plausibility of the Data given the Model

Prior
Plausibility of the model given previous information

- Imagine that we toss a coin 5 times and get {H,H,T,H,T}
- How do we estimate the bias *h* of our coin towards H?

- High school (naïve frequentist) approach: $h = 3/5$.

- Imagine that we toss a coin 5 times and get {H,H,T,H,T}
- How do we estimate the bias *h* of our coin towards H?

- **Bernoulli process** At each toss, independently of the previous ones, the probability of getting H is *h*. The model is fully specified by *h* (therefore, M := h)

- Imagine that we toss a coin 5 times and get {H,H,T,H,T}
- How do we estimate the bias *h* of our coin towards H?

- **Bernoulli process** At each toss, independently of the previous ones, the probability of getting H is *h*. The model is fully specified by *h* (therefore, M := h)
- Then, the probability of getting {H,H,T,H,T} is

$$p(\{H,H,T,H,T\}|h) = h \times h \times (1-h) \times h \times (1-h) = h^3(1-h)^2$$

Let's estimate the bias of a coin towards heads using Bayesian inference

- Imagine that we toss a coin 5 times and get {H,H,T,H,T}
- How do we estimate the bias *h* of our coin towards H?

- **Bernoulli process** At each toss, independently of the previous ones, the probability of getting H is *h*. The model is fully specified by *h* (therefore, M := h)
- Then, the probability of getting {H,H,T,H,T} is

$$p(\{H,H,T,H,T\}|h) = h \times h \times (1-h) \times h \times (1-h) = h^3(1-h)^2$$

- If, a priori, we don't know anything about the right value of *h*, we can assume that the prior is uniform

$$p(h) = 1, h \in [0,1]$$

- Imagine that we toss a coin 5 times and get {H,H,T,H,T}
- How do we estimate the bias *h* of our coin towards H?

- **Bernoulli process** At each toss, independently of the previous ones, the probability of getting H is *h*. The model is fully specified by *h* (therefore, M := h)
- Then, the probability of getting {H,H,T,H,T} is

$$p(\{H,H,T,H,T\}|h) = h \times h \times (1-h) \times h \times (1-h) = h^3(1-h)^2$$

- If, a priori, we don't know anything about the right value of *h*, we can assume that the prior is uniform

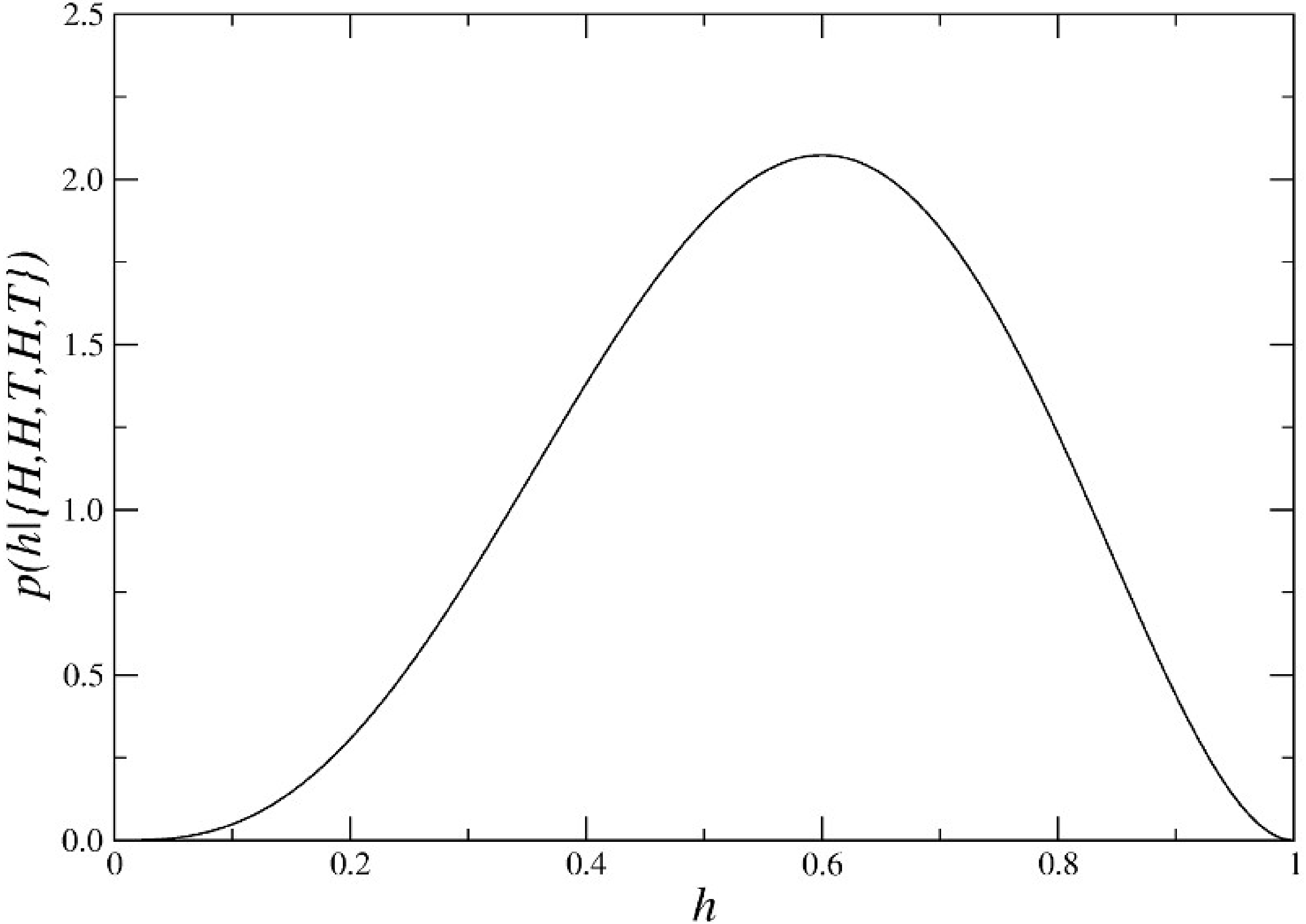
$$p(h) = 1, h \in [0,1]$$

- Then, we finally have that

$$p(h|\{H,H,T,H,T\}) \propto p(\{H,H,T,H,T\}|h)\, p(h) = h^3(1-h)^2$$

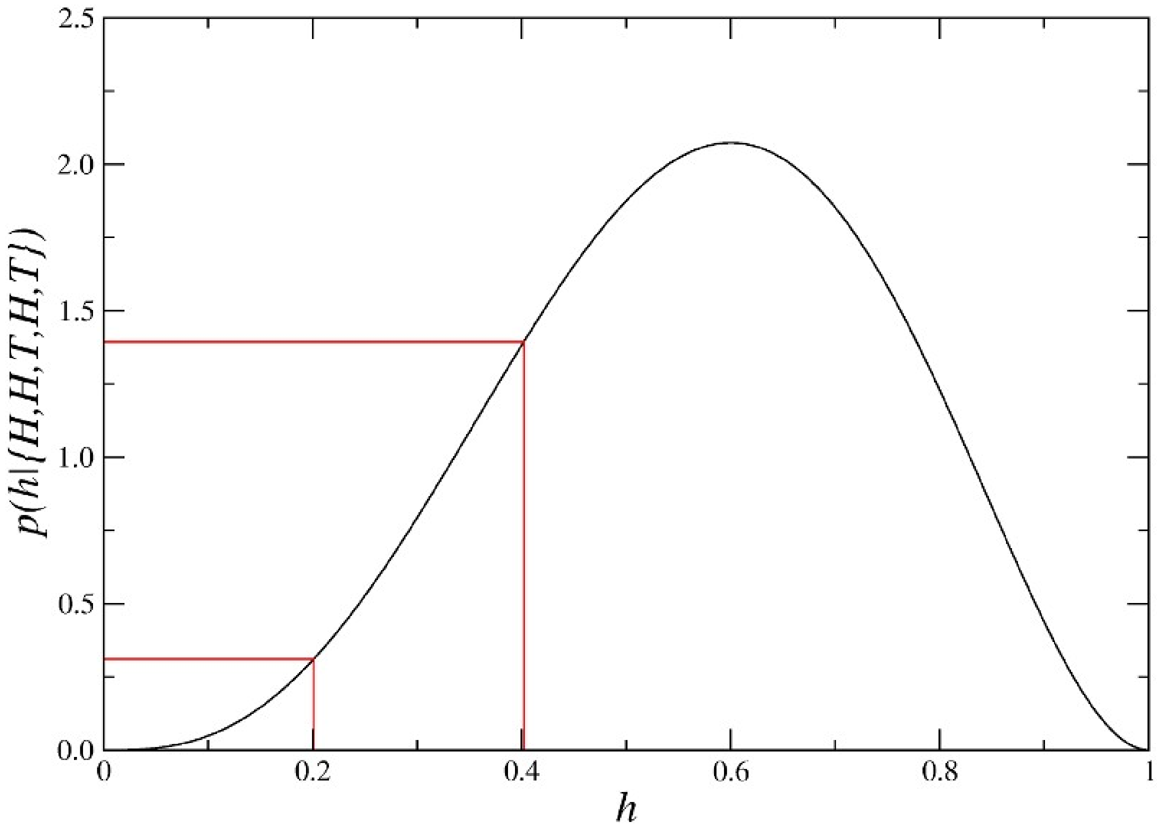
Let's estimate the bias of a coin towards heads using Bayesian inference

$$p(h|\{H, H, T, H, T\}) \propto h^3(1-h)^2$$

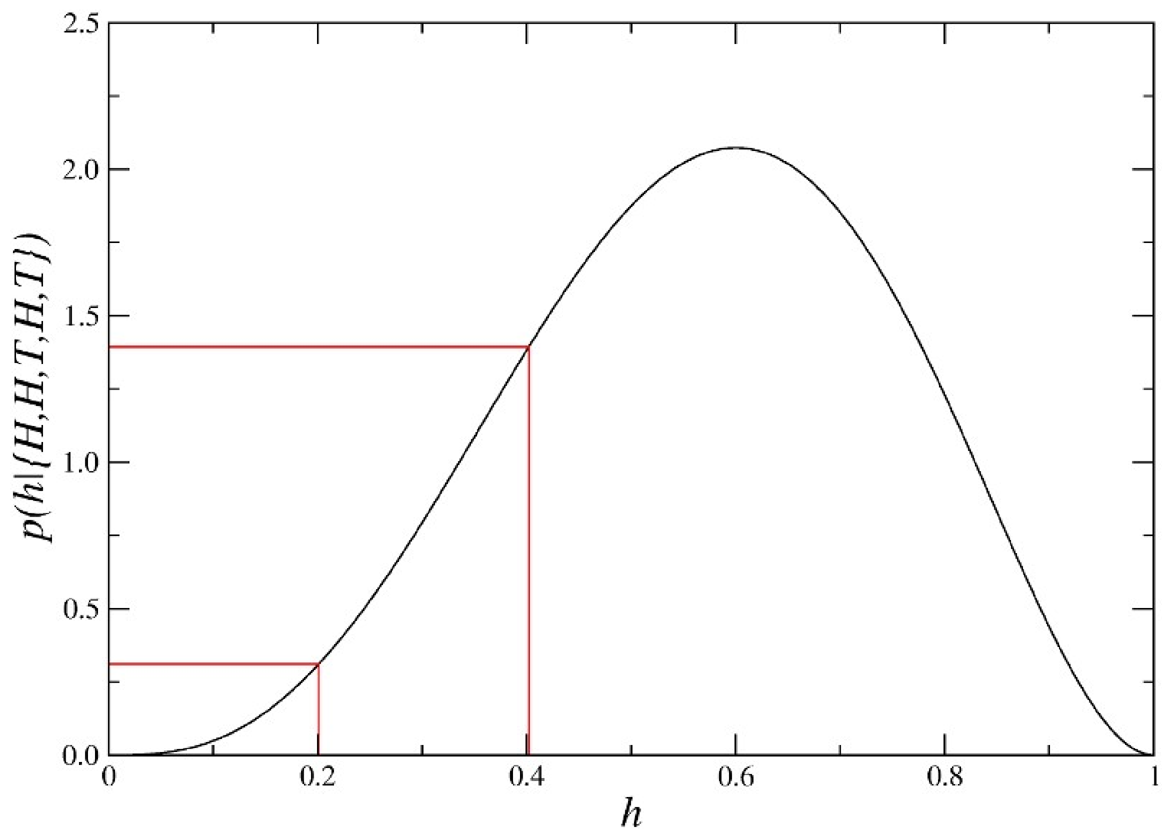


Let's now estimate the probability that the next toss gives H

- From the naïve frequentist approach: h=3/5, so that's the probability of getting H in the next toss
- Within the Bayesian approach, we can/should consider all evidence we have:



- From the naïve frequentist approach: h=3/5, so that's the probability of getting H in the next toss
- Within the Bayesian approach, we can/should consider all evidence we have:



$$p(\text{next toss} = H|\{H,H,T,H,T\}) = \int_0^1 h \times p(h|\{H,H,T,H,T\})dh$$

- Like in the previous example, we are often interested in evaluating integrals of the form

$$p(\text{next toss} = H|\{H,H,T,H,T\}) = \int_0^1 h \times p(h|\{H,H,T,H,T\})dh = \frac{4}{7}$$

$$\langle f(M) \rangle = \int f(M) \times p(M|D)dM$$

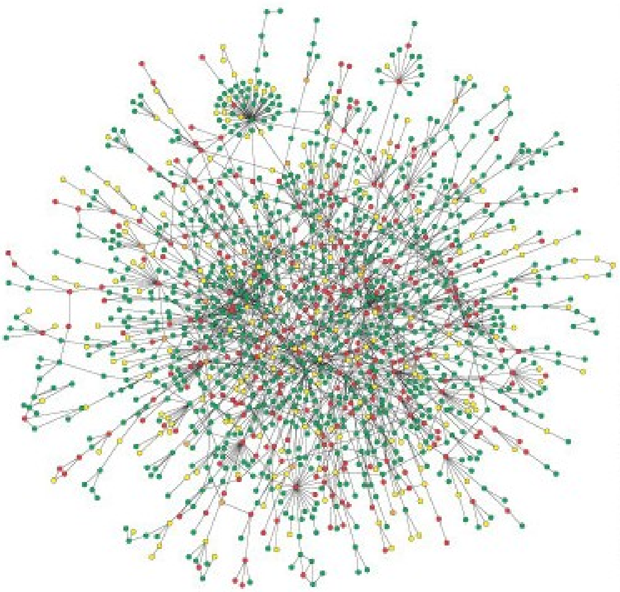
- Unlike the previous example, more often than not these integrals cannot be calculated exactly
- In such cases, we can use Markov Chain Monte Carlo (MCMC)

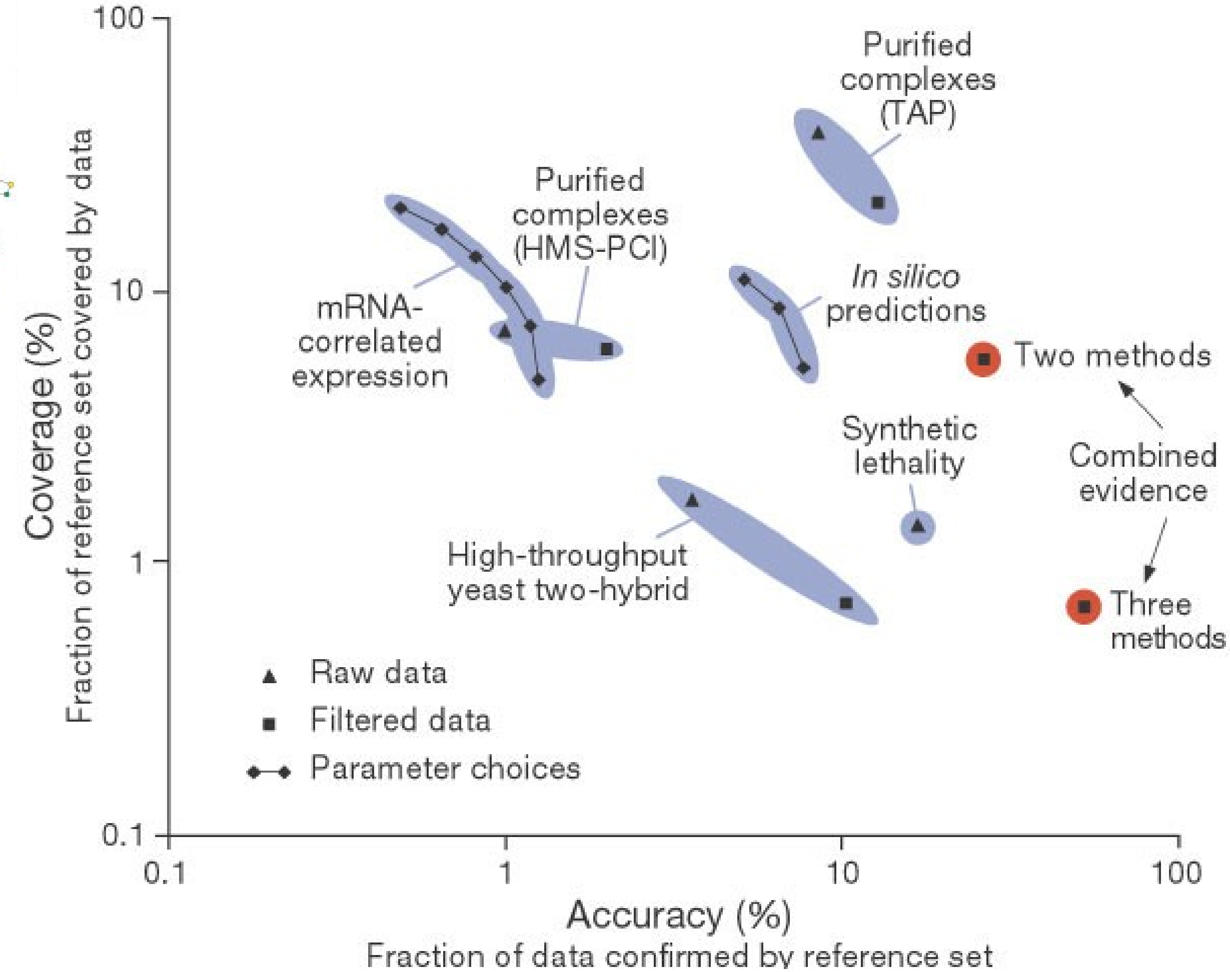
$$\langle f(M) \rangle = \frac{1}{N} \sum_i f(M_i)$$

where the sum is over *N* models sampled (using the Gibbs sampler or the Metropolis-Hastings algorithm) from the distribution *p(M|D)*

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message

Challenge #1: There is much about the interactions in the networks we study that we don't know





von Mering et al., *Nature* (2002)

- Given a single noisy observation of a network, determine:
  - ***Missing interactions*** Interactions that exist but are not captured in our observation of the system
  - ***Spurious interactions*** Interactions that do *not* exist but, for some reason, are included in our observation
- ***Reconstruct the network***, so that our reconstruction has properties that are closer to the properties of the true network

- Given a single noisy observation of a network, determine:
  - ***Missing interactions*** Interactions that exist but are not captured in our observation of the system
  - ***Spurious interactions*** Interactions that do *not* exist but, for some reason, are included in our observation
- ***Reconstruct the network***, so that our reconstruction has properties that are closer to the properties of the true network

- But:
  - We want to be able to do this for arbitrary real networks about which we don't know anything
  - There seems to be a paradox in trying to identify what is wrong in a network observation---from *the network observation itself* !

There are two possible scenarios when in comes to solving the paradox

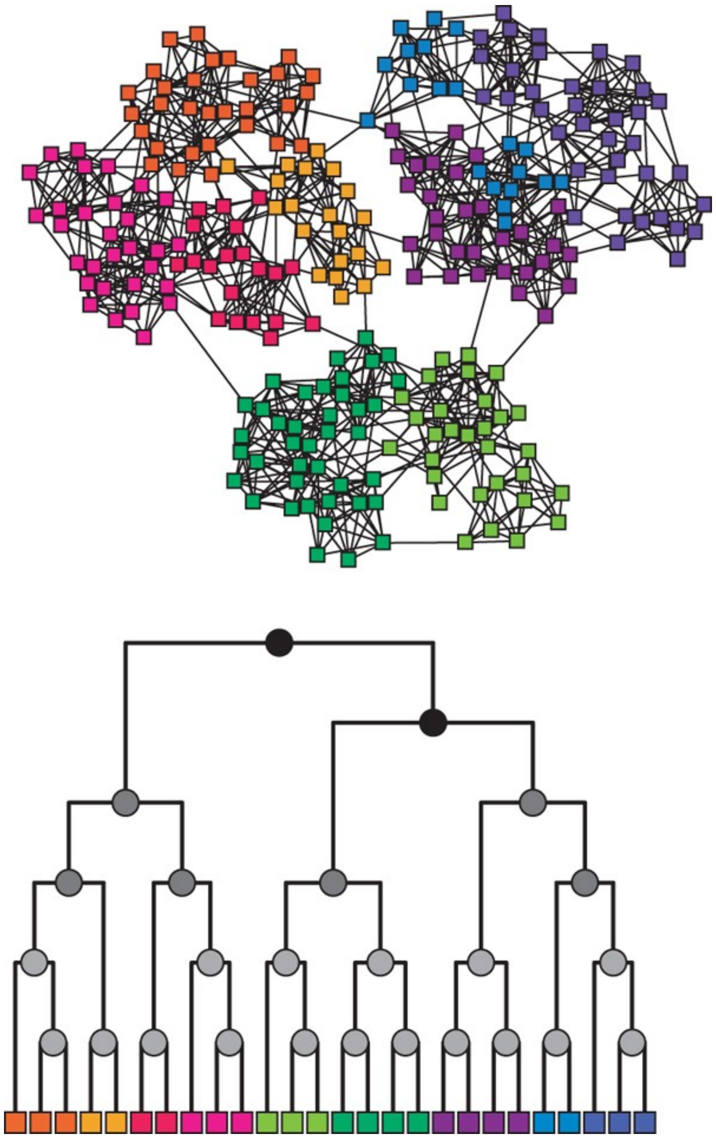
- Scenario 1: We *don't* have a clue about what the network should look like, or where does it come from (mechanistically or statistically):
  - We cannot do anything
- Scenario 2: We *do* have some ideas about the structure of the network:
  - We can formalize these ideas into a set of models
  - We can use the models to assess what is likely to be missing/wrong

- We assume our network is the outcome of an undetermined model *M*, from a (potentially infinite) collection of models $\mathcal{M}$
- We observe a network A^O
- Given my observation A^O, what is the probability that a property *X* takes the value *X=x* if we generate a new network (with the same model)?

$$p(X = x|A^0) = \int_{\mathcal{M}} p(X = x|M)\, p(M|A^0)\, dM$$

- We call $p(X=x|A^O)$ the ***reliability of the X=x measurement***

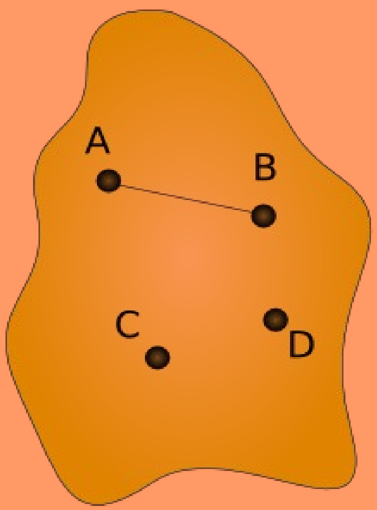
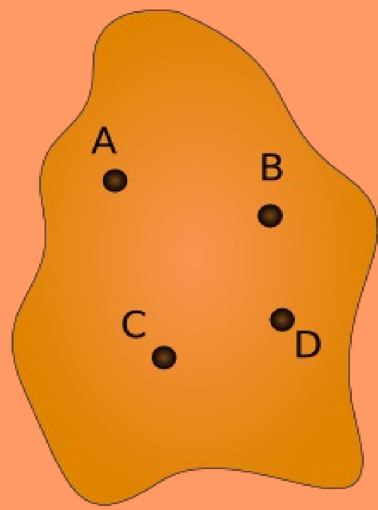
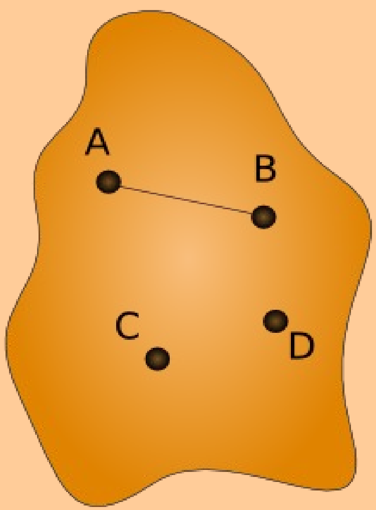
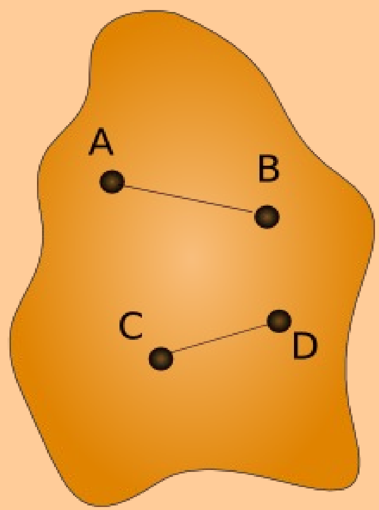
- In particular, we can calculate the probability $p(A_{ij} = 1|A^O)$ that a link exists



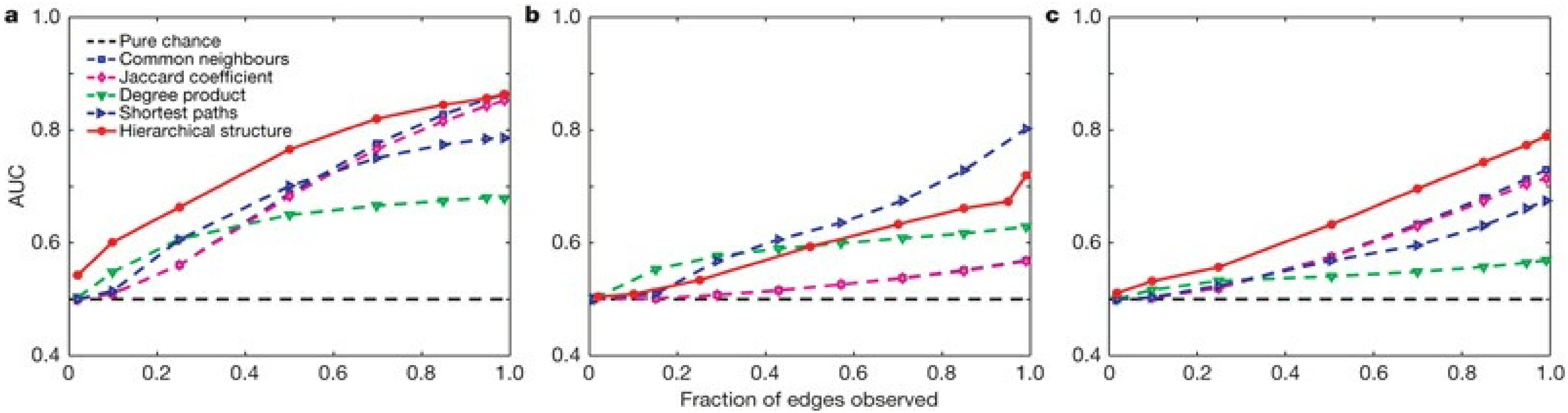
- ➔ A hierarchical network with structure on many scales, and the corresponding hierarchical random graph.
- ➔ Each internal node of the dendrogram is associated with a probability that a pair of vertices in the left and right subtrees of that node are connected. (The shades of the internal nodes in the figure represent the probabilities.)

Clauset, Moore, Newman, *Nature* (2008)

One can test if inference methods can identify missing and spurious interactions in real networks

<table>
<tr><th></th><th>True network</th><th>Observed network</th><th>Test</th></tr>
<tr><td>Missing interactions</td><td></td><td></td><td>How often is AB more reliable than CD?</td></tr>
<tr><td>Spurious interactions</td><td></td><td></td><td>How often is CD less reliable than AB?</td></tr>
</table>

Inference with the hierarchical random graph is often more accurate than "local" metrics

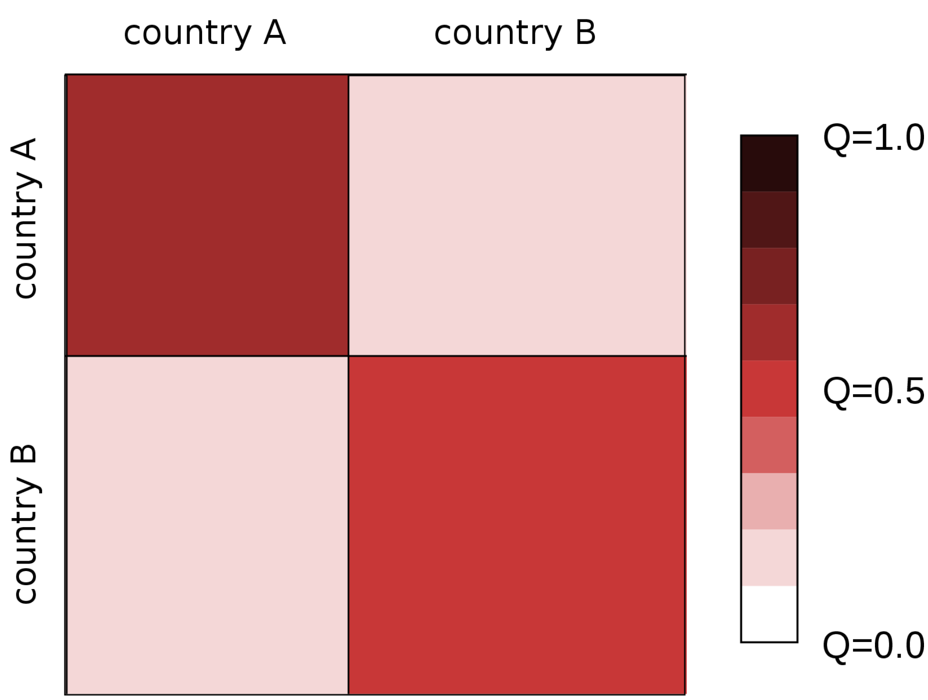


Clauset, Moore, Newman, *Nature* (2008)

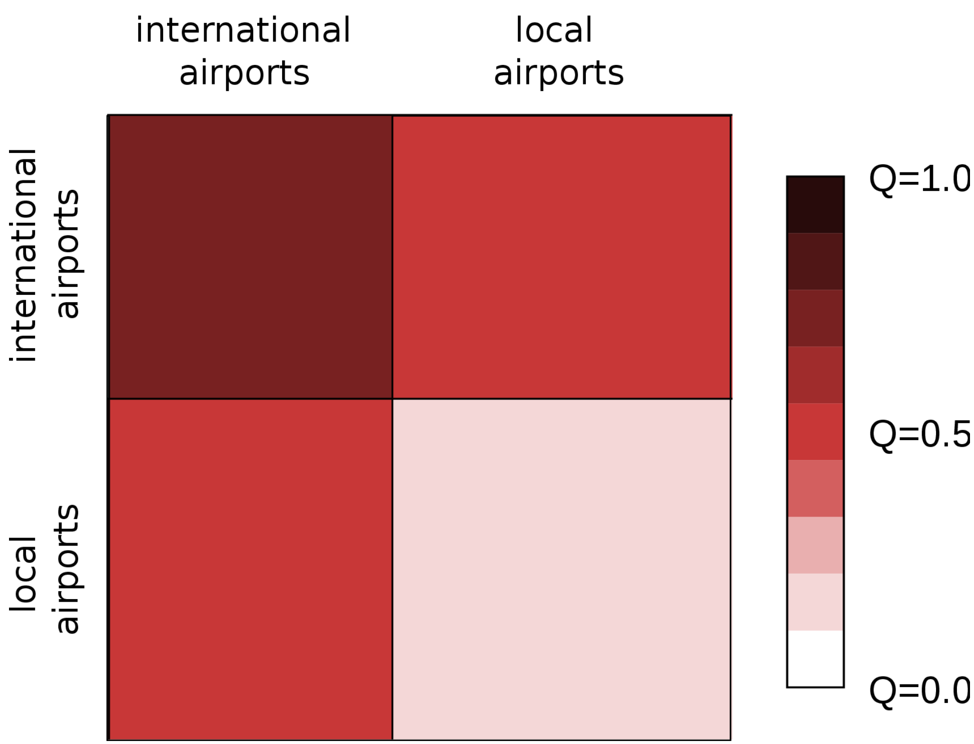
Outline

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message

Modularity

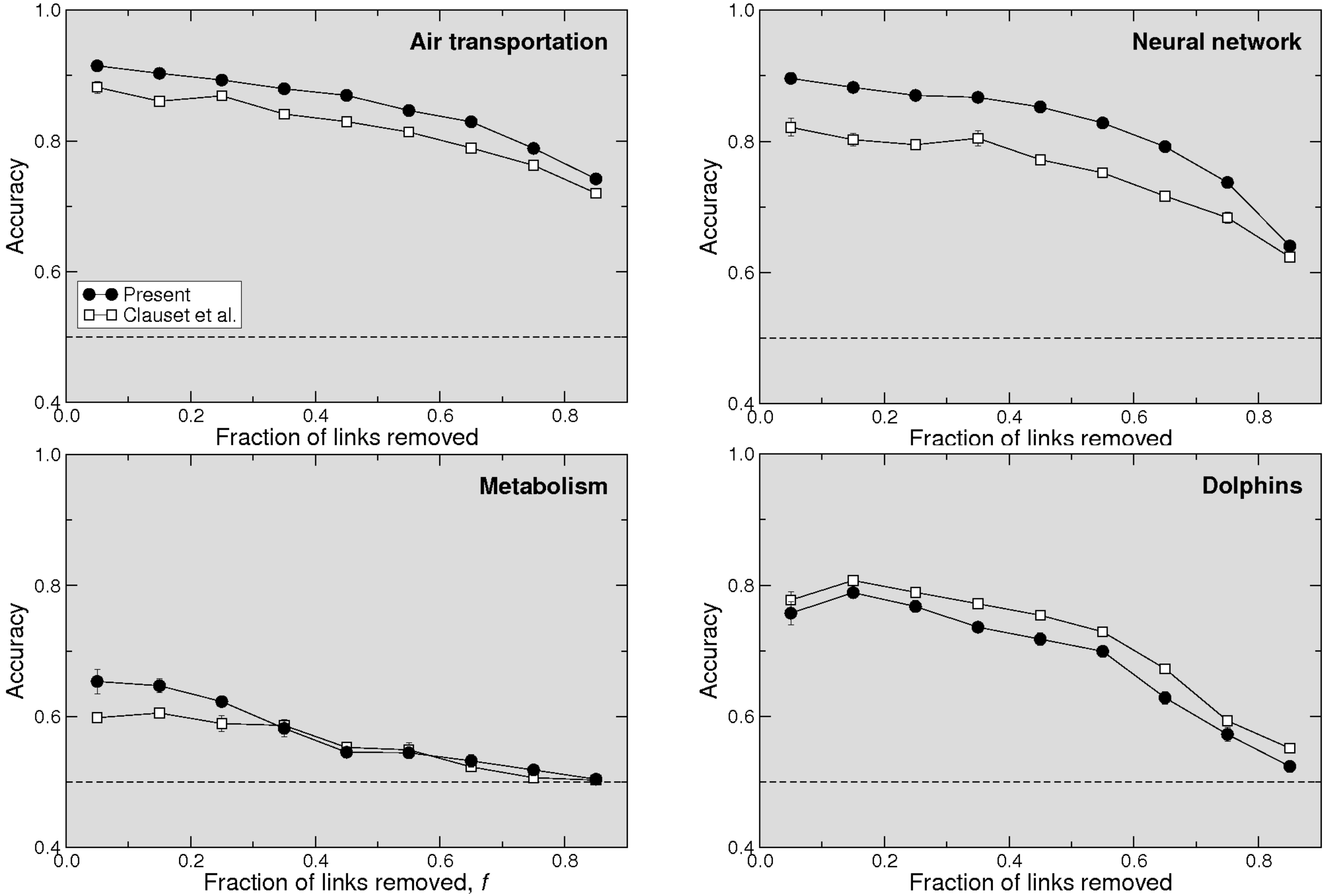


Role-to-role correlations



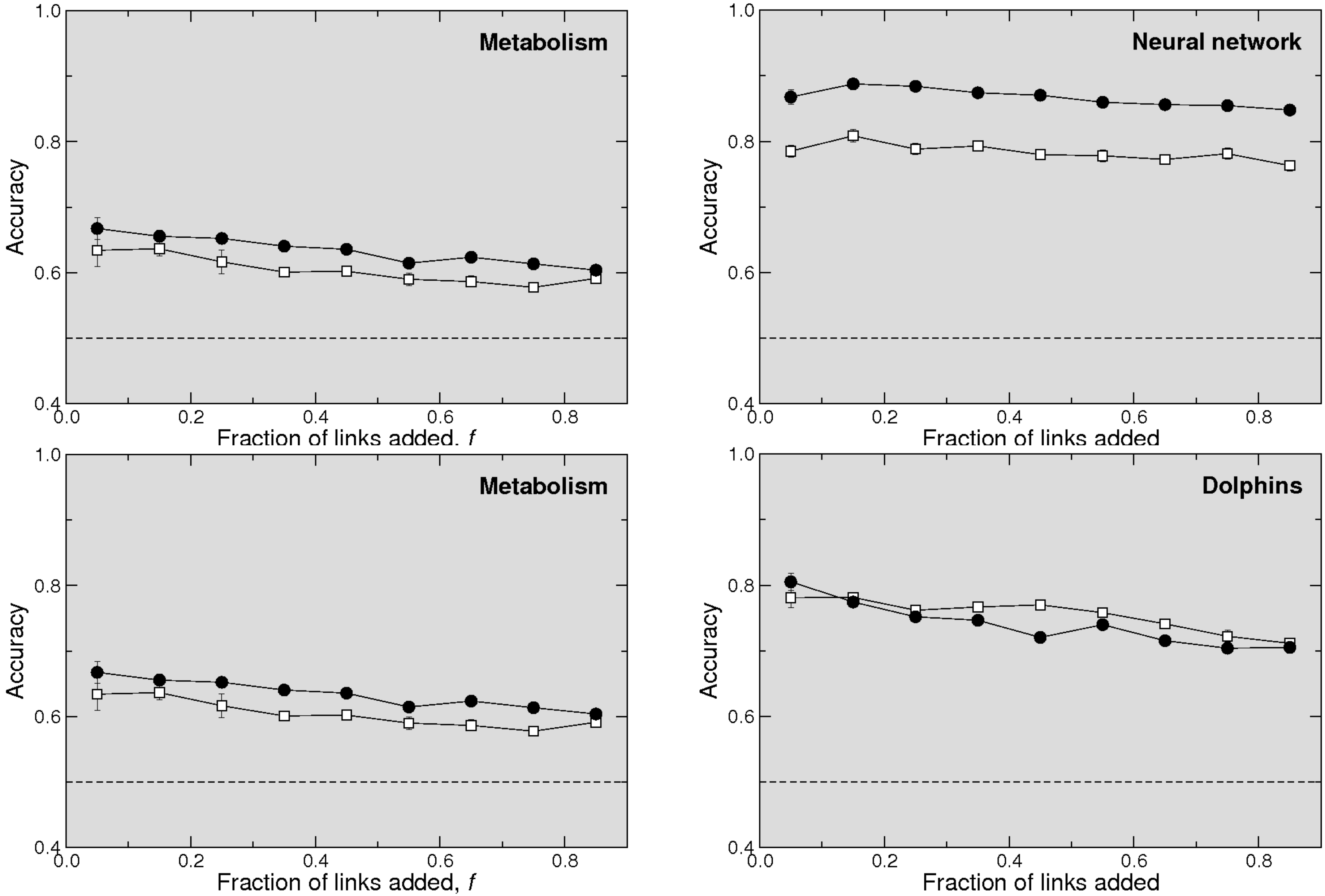
Guimera, Sales-Pardo, *PNAS* (2009)

Block models often outperform hierarchical random graphs at identifying missing interactions



Guimera, Sales-Pardo, *PNAS* (2009)

Block models often outperform hierarchical random graphs at identifying spurious interactions



Guimera, Sales-Pardo, *PNAS* (2009)

Reconstructing a network is more complicated than just adding missing interactions and removing spurious interactions

- Challenges:
  - We don't know how many links need to be added and removed
  - Links cannot be added and removed independently of each other

We define a network reliability: the *network reconstruction* is the most reliable network

- The reliability of a network is $R^N_A = p(A|A^O)$

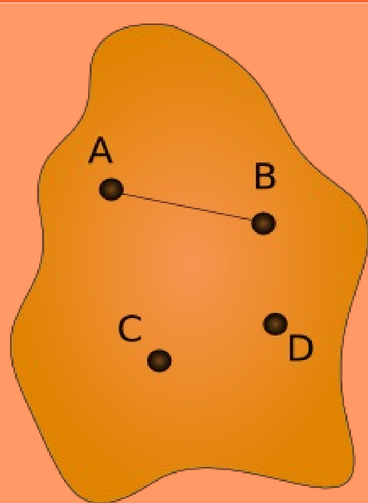
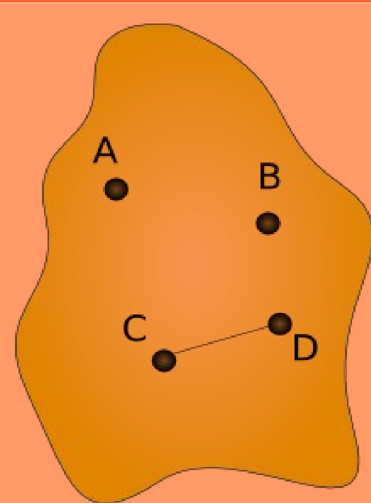
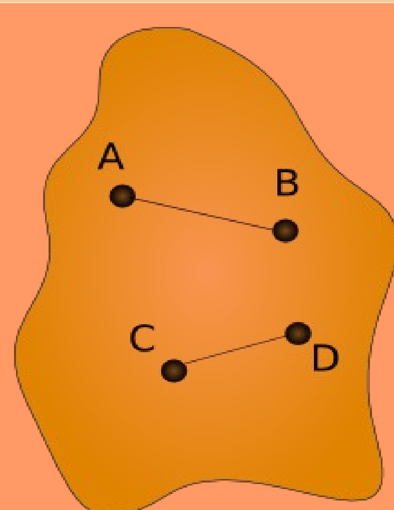
$$p(A|A^0) = \int_{\mathcal{M}} p(A|M)\, p(M|A^0)\, dM$$

- The reconstruction A^R is the network that maximizes R^N_A
- We obtain A^R using uphill search

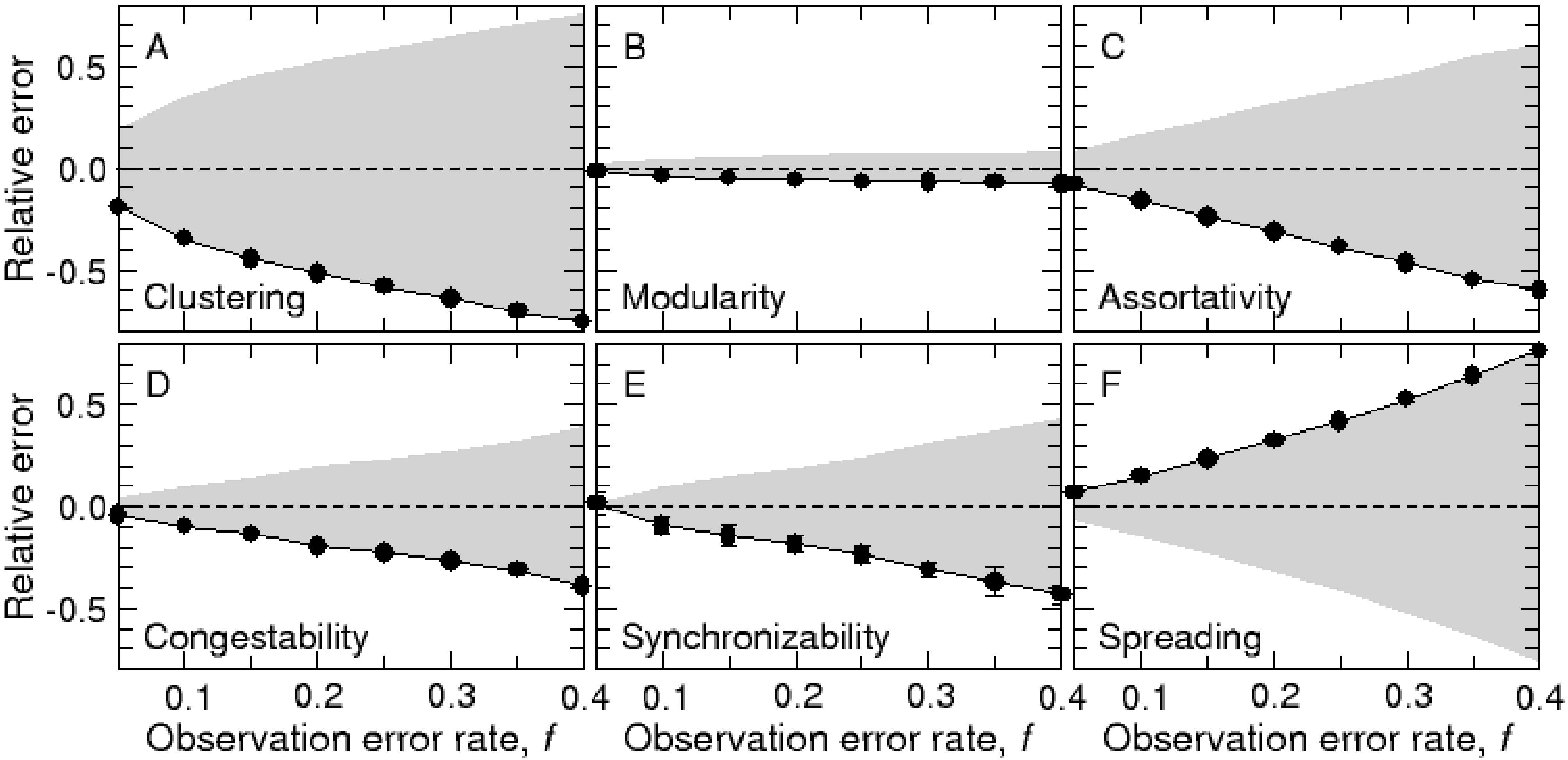
ECCS WARM-UP

Do reconstructions improve estimates of network properties?

School on Complex Networks, Sept 13-15

<table>
<tr><th></th><th>True network</th><th>Observed network</th><th>Test</th></tr>
<tr><td>Random errors</td><td>
</td><td>
</td><td>How do network properties change?</td></tr>
<tr><td></td><td></td><td>Reconstructed network</td><td></td></tr>
<tr><td></td><td></td><td>
</td><td>How do network properties change?</td></tr>
</table>

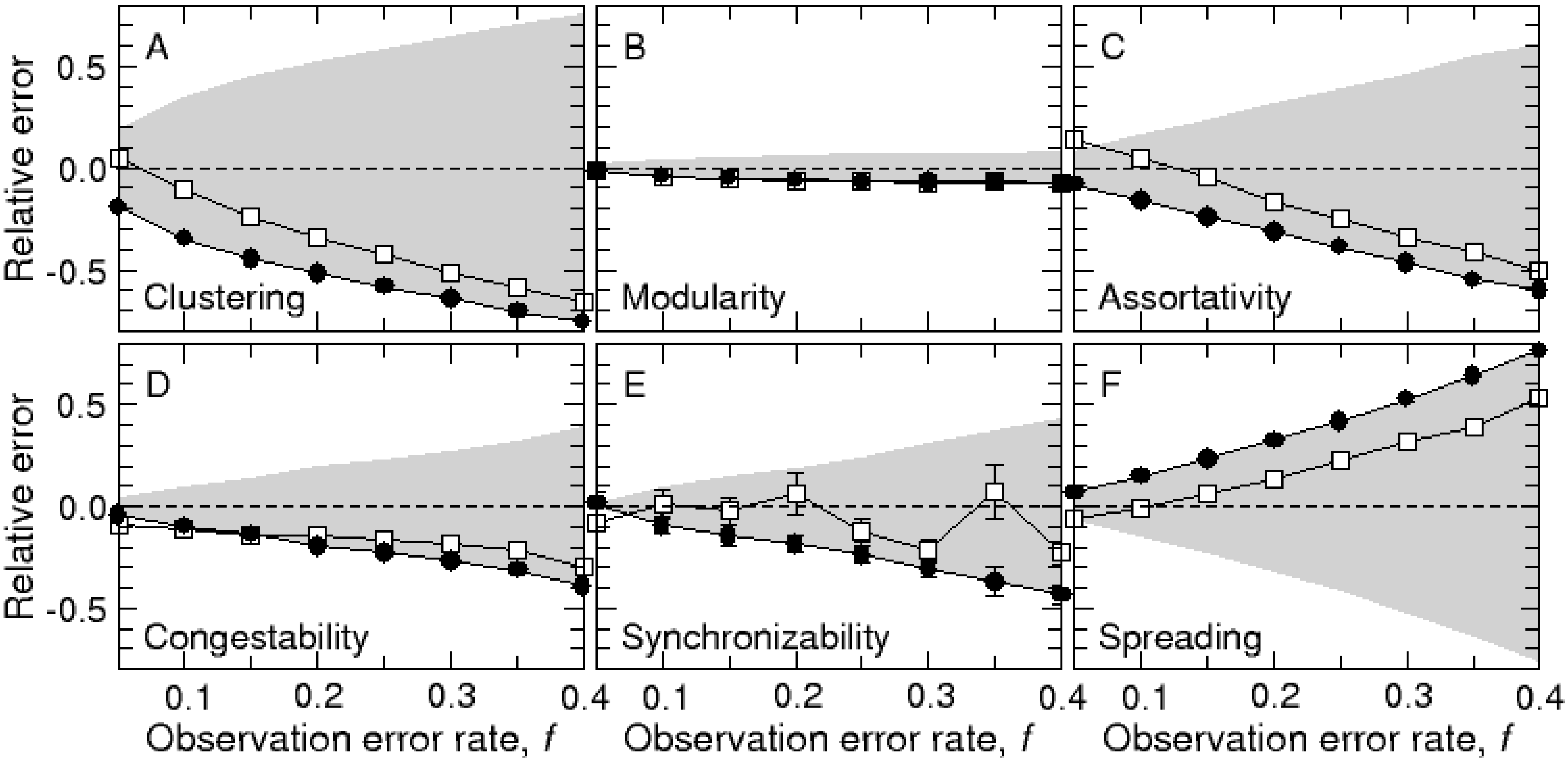
Network reconstructions provide better estimates of global network properties than the observations themselves



ECCS WARM-UP

School on Complex Networks, Sept 13-15

Network reconstructions provide better estimates of global network properties than the observations themselves



Guimera, Sales-Pardo, *PNAS* (2009)

- Network modularity
  - The problem
  - Algorithms and their evaluation
  - Are networks really modular?
  - So what, if real networks are modular?
  - Beyond modules: positions and block models
  - Hands-on: module-identification algorithms
- BREAK
- Network inference
  - Shortest tutorial ever on Markov chain Monte Carlo for Bayesian inference
  - Network inference using hierarchical random graphs
  - Network inference using stochastic block models
- Back to drugs and movies, take-home message and more hands-on

ECCS WARM-UP

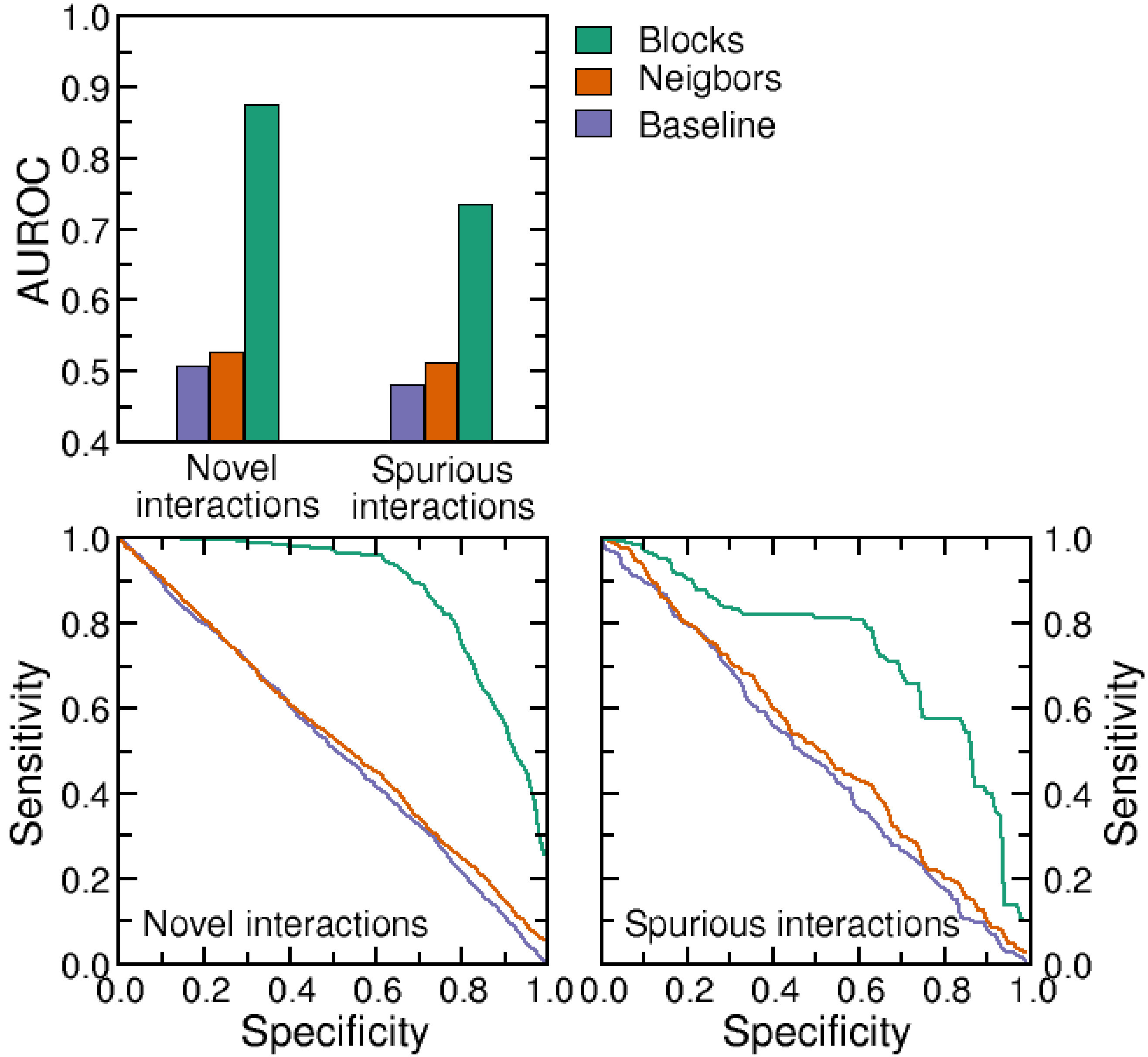
School on Complex Networks, Sept 13-15

With the same tools we can predict if combining two drugs poses a risk to health...

ECCS WARM-UP

School on Complex Networks, Sept 13-15

We can predict which severe drug interactions will be removed from and added to a database



Guimera, Sales-Pardo *PloS Comp Biol* (2013)

ECCS WARM-UP

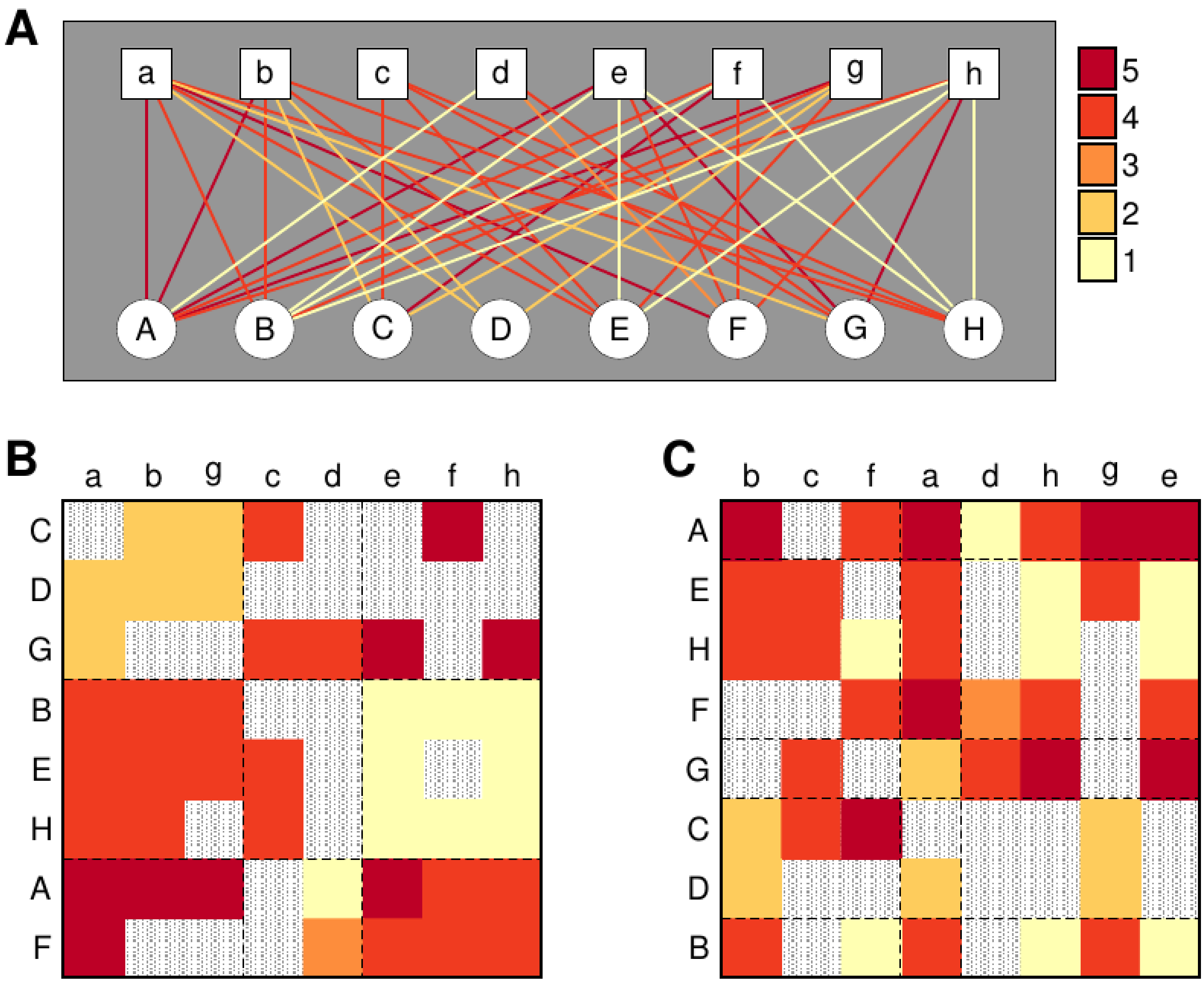
School on Complex Networks, Sept 13-15

...or whether you are going to like "The Dark Knight rises"!

ECCS WARM-UP

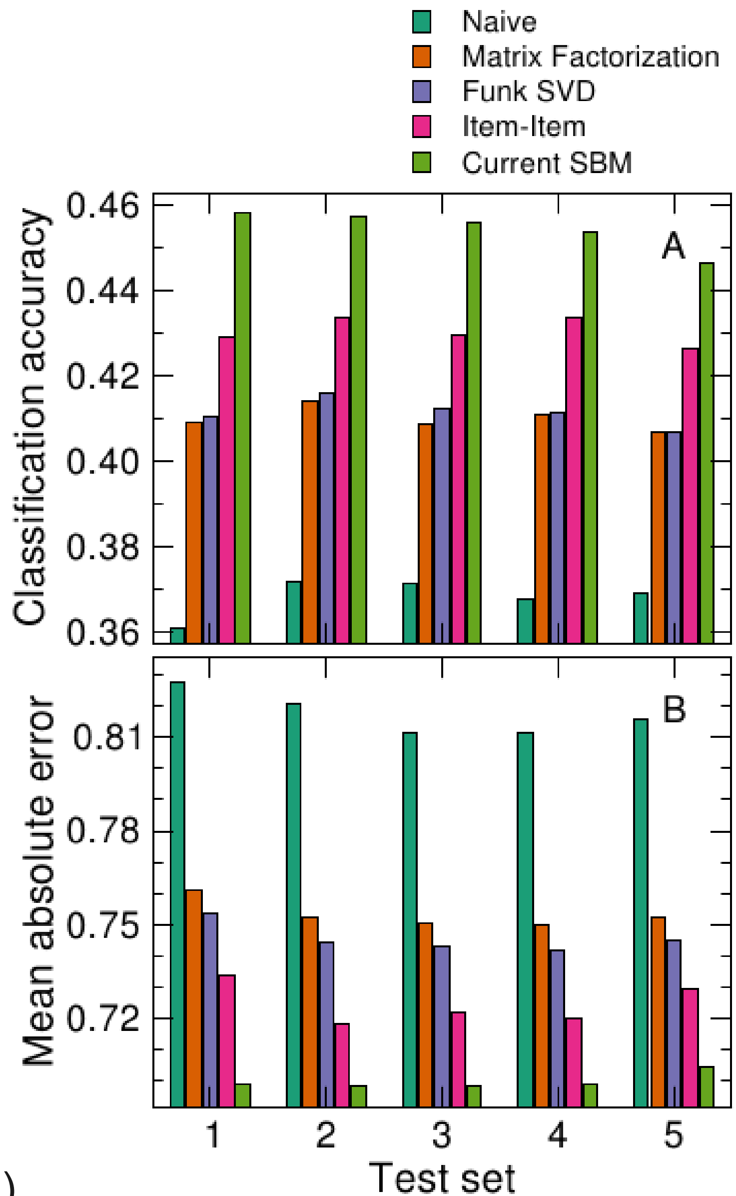
School on Complex Networks, Sept 13-15

Predicting human preferences can be reformulated as a problem of network inference and tackled, in particular, using SBM



Our approach predicts human preferences considerably better than some of the best collaborative filtering algorithms

- MovieLens set: 100,000 real 1-5 movie ratings by ~1,000 users
- 5 independent splits of the data into 80,000 observed ratings and 20,000 validation ratings



Guimera, Llorente, Moro, Sales-Pardo *PloS ONE* (2012)

- Funding

- More about our research:
  - http://seeslab.info
  - @sees_lab