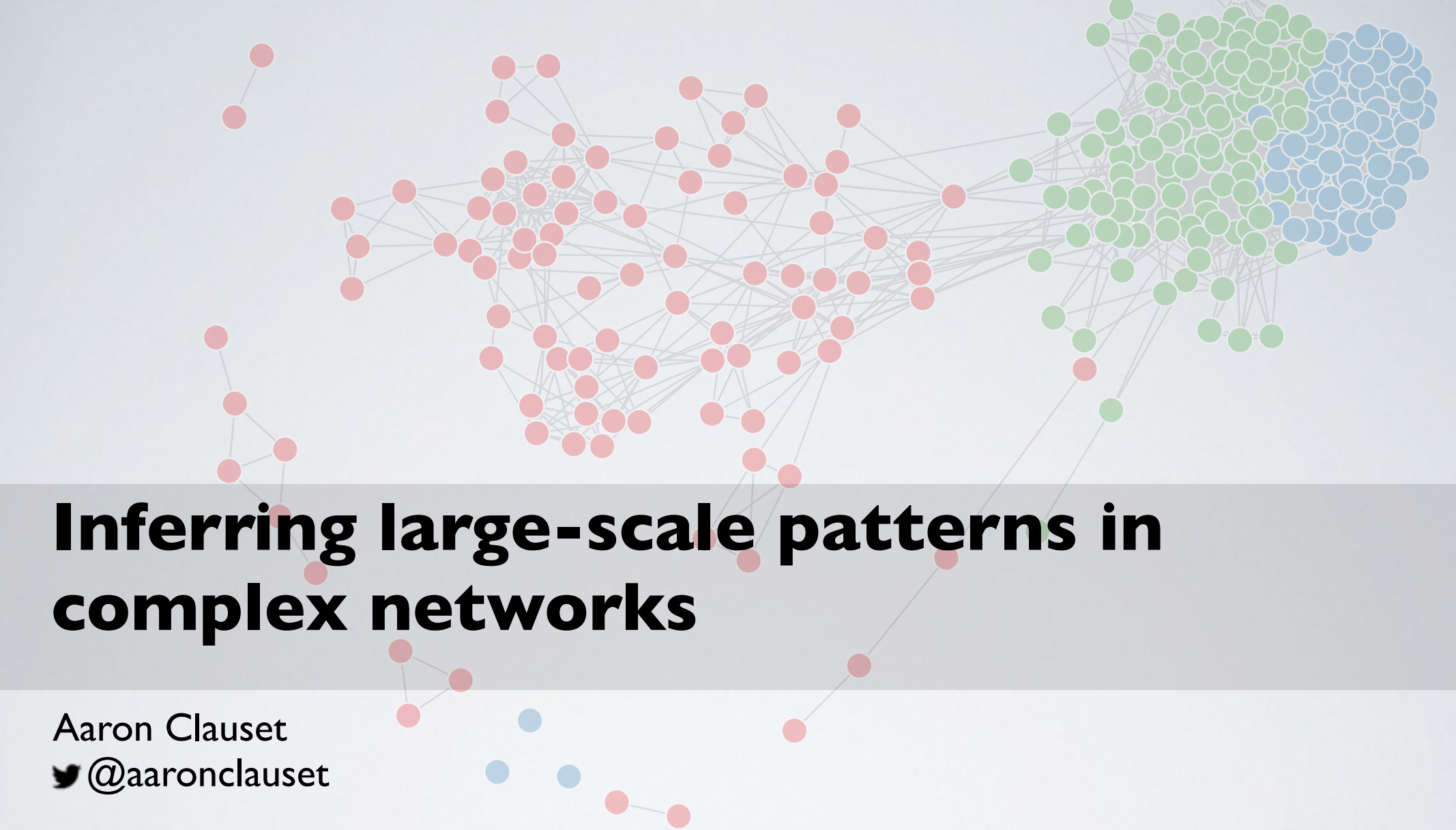


Inferring large-scale patterns in complex networks

Aaron Clauset
@aaronclauset

Computer Science Dept. & BioFrontiers Institute
University of Colorado, Boulder
External Faculty, Santa Fe Institute

joint work

Chris Aicher
(Colorado)

Abigail Z. Jacobs
(Colorado)

Dr. Dan Larremore
(Harvard)

Dr. Leto Peel
(Colorado)

Prof. Cris Moore
(Santa Fe)

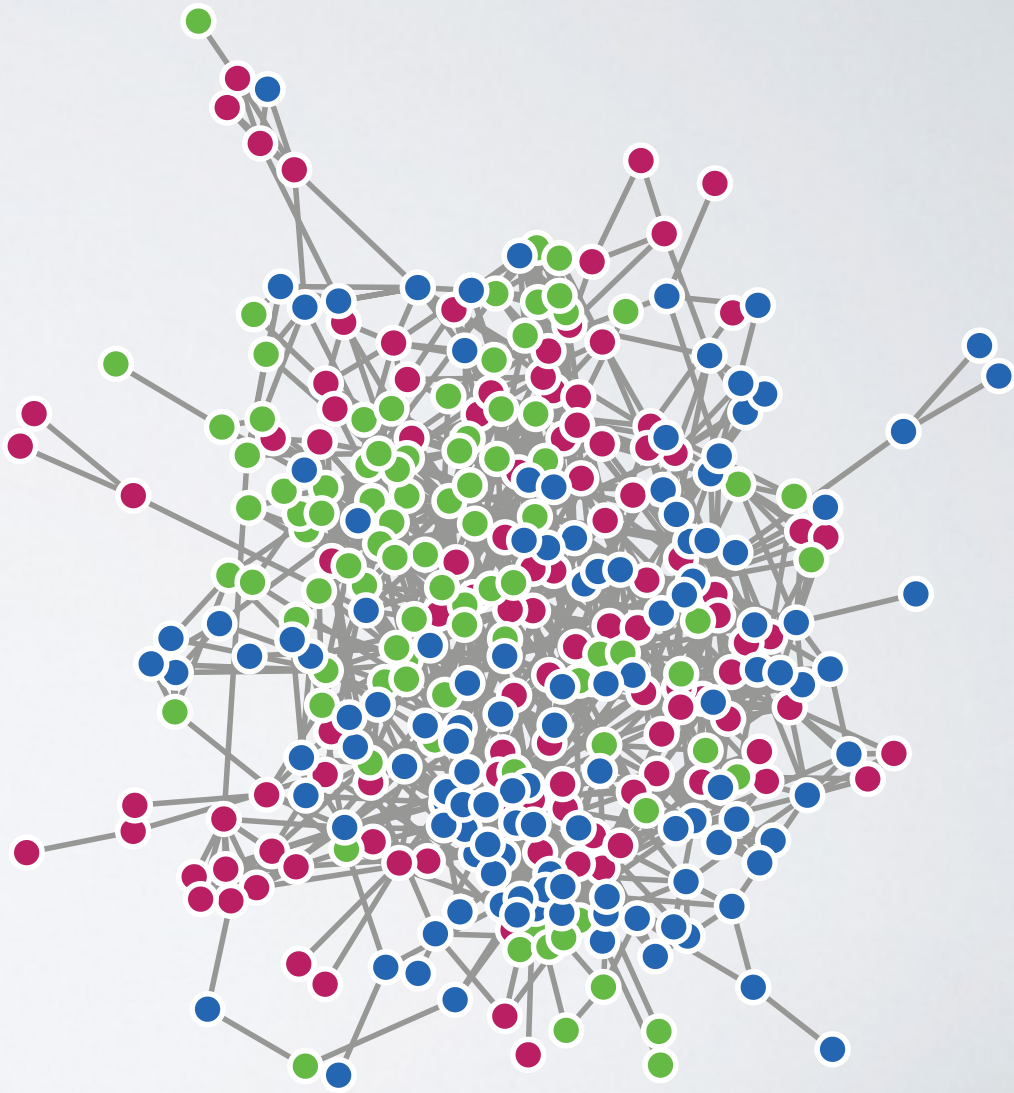
Prof. Mark Newman
(Michigan)

Prof. Caroline Buckee
(Harvard)

NSF
DARPA
AFOSR
James S. McDonnell Foundation
NIH
Center for Communicable Disease Dynamics

what is large-scale structure?

what networks look like



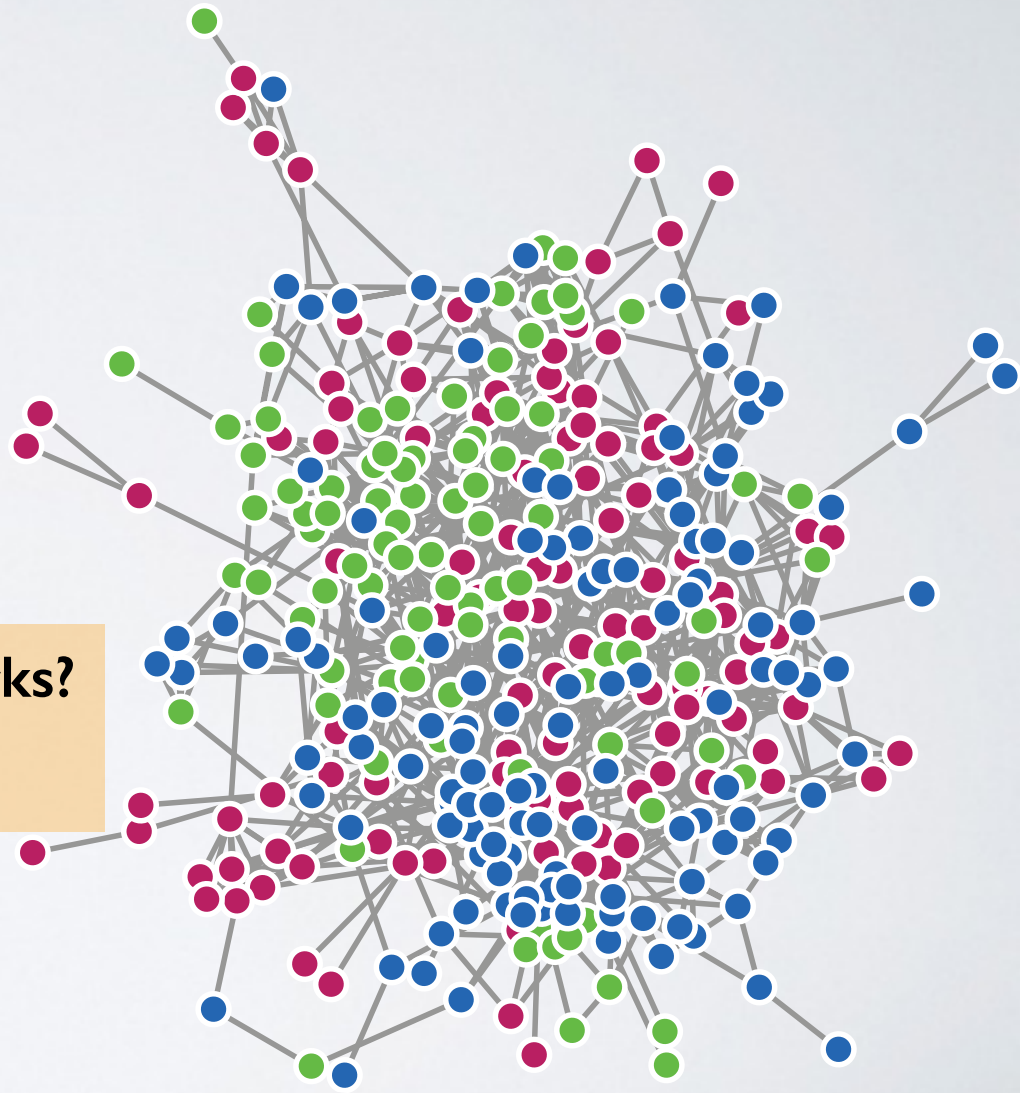
what is large-scale structure?

what networks look like

- **how are the edges organized?**
- **how do vertices differ?**
- **does network location matter?**
- **are there underlying patterns?**

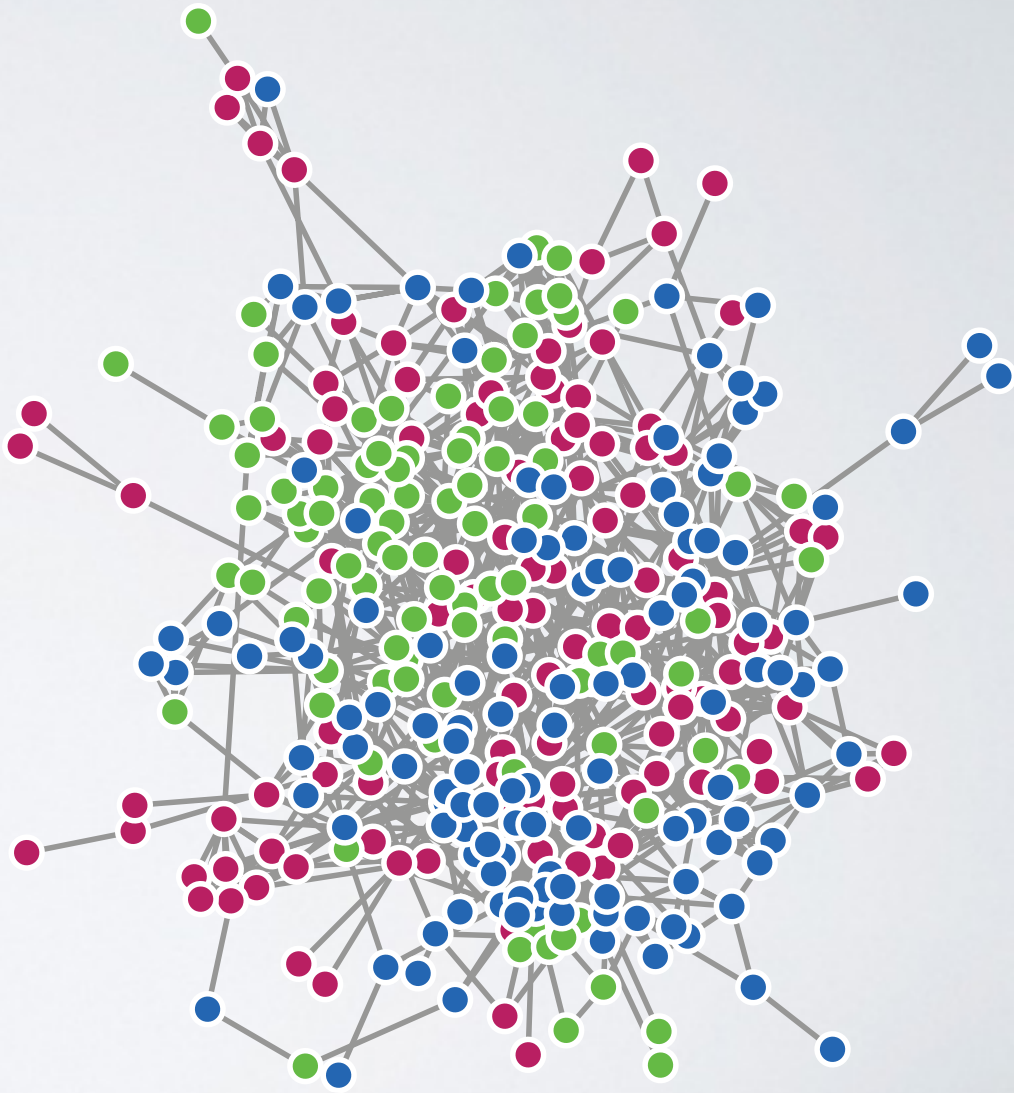
what we want to know

- **what processes shape these networks?**
- **how can we tell?**



what is large-scale structure?

what we usually do : **describe its features**

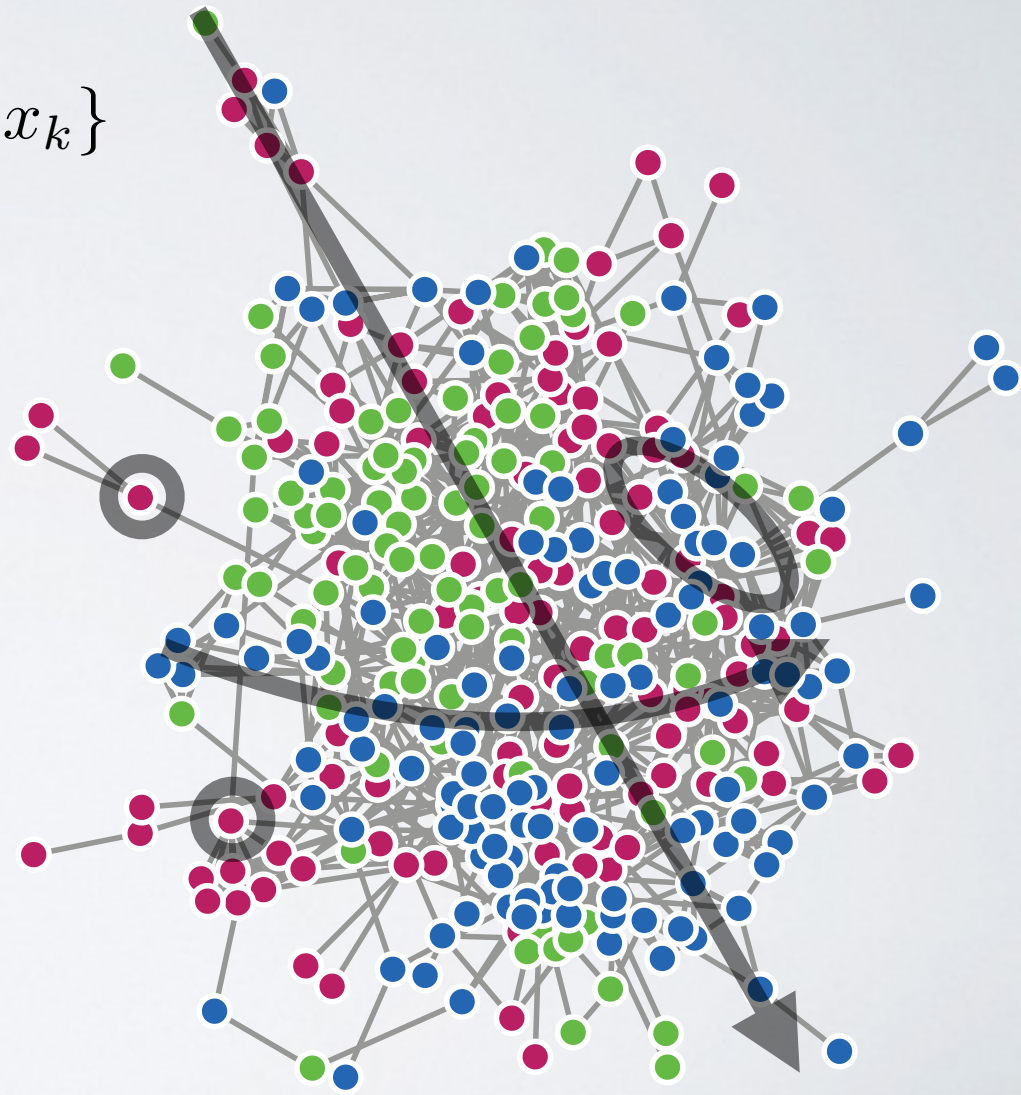


what is large-scale structure?

what we usually do : **describe its features**

$$f : G \to \{x_1, \ldots, x_k\}$$

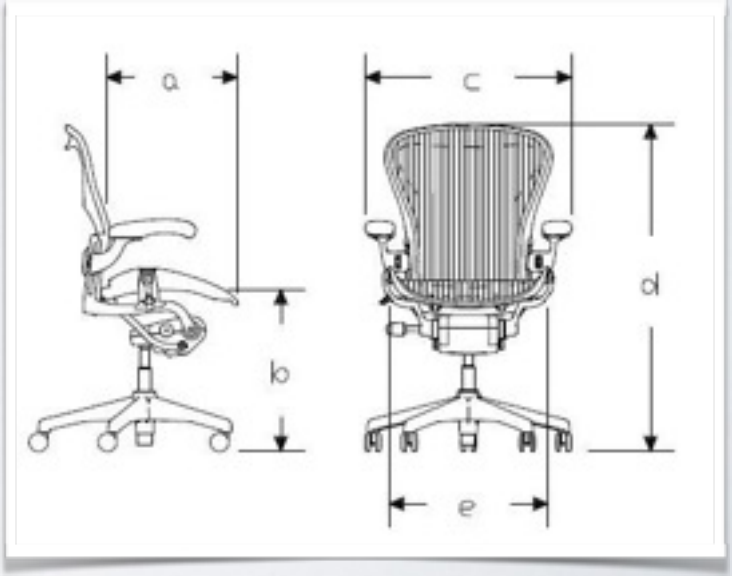
- **degree distributions**
- **short-loop density (triangles, etc.)**
- **shortest paths (diameter, etc.)**
- **centrality scores**
- **correlations between these**



what is large-scale structure?

what we usually do : **describe its features**

$$f : \text{object} \rightarrow \{x_1, \ldots, x_k\}$$



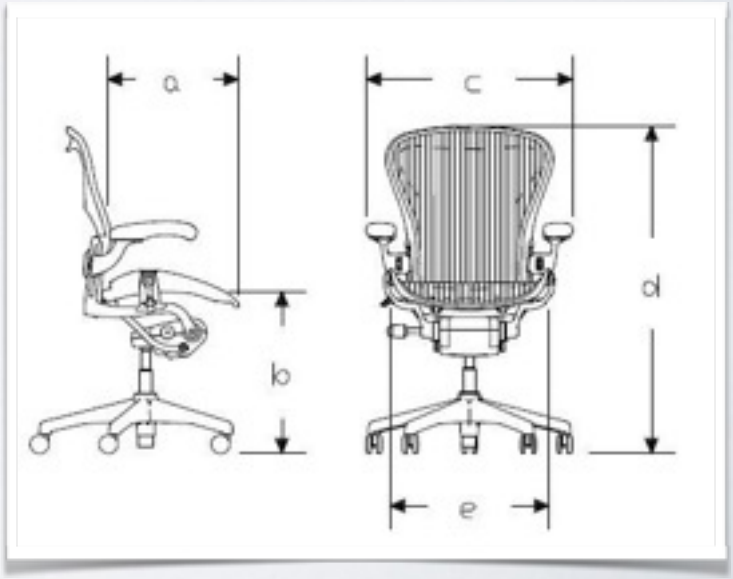
what is large-scale structure?

what we usually do : **describe its features**

$$f : \text{object} \rightarrow \{x_1, \ldots, x_k\}$$

- **physical dimensions**
- **material density, composition**
- **radius of gyration**
- **correlations between these**

helpful for intuition, but not what we want…



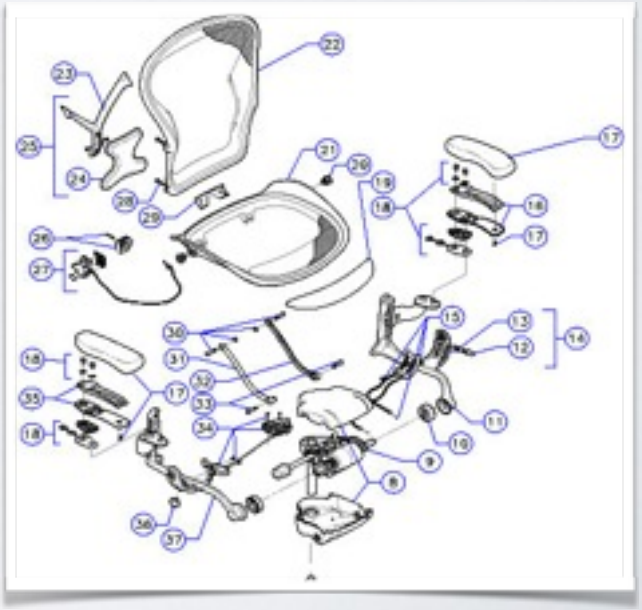
what is large-scale structure?

what we want : **understand its structure**

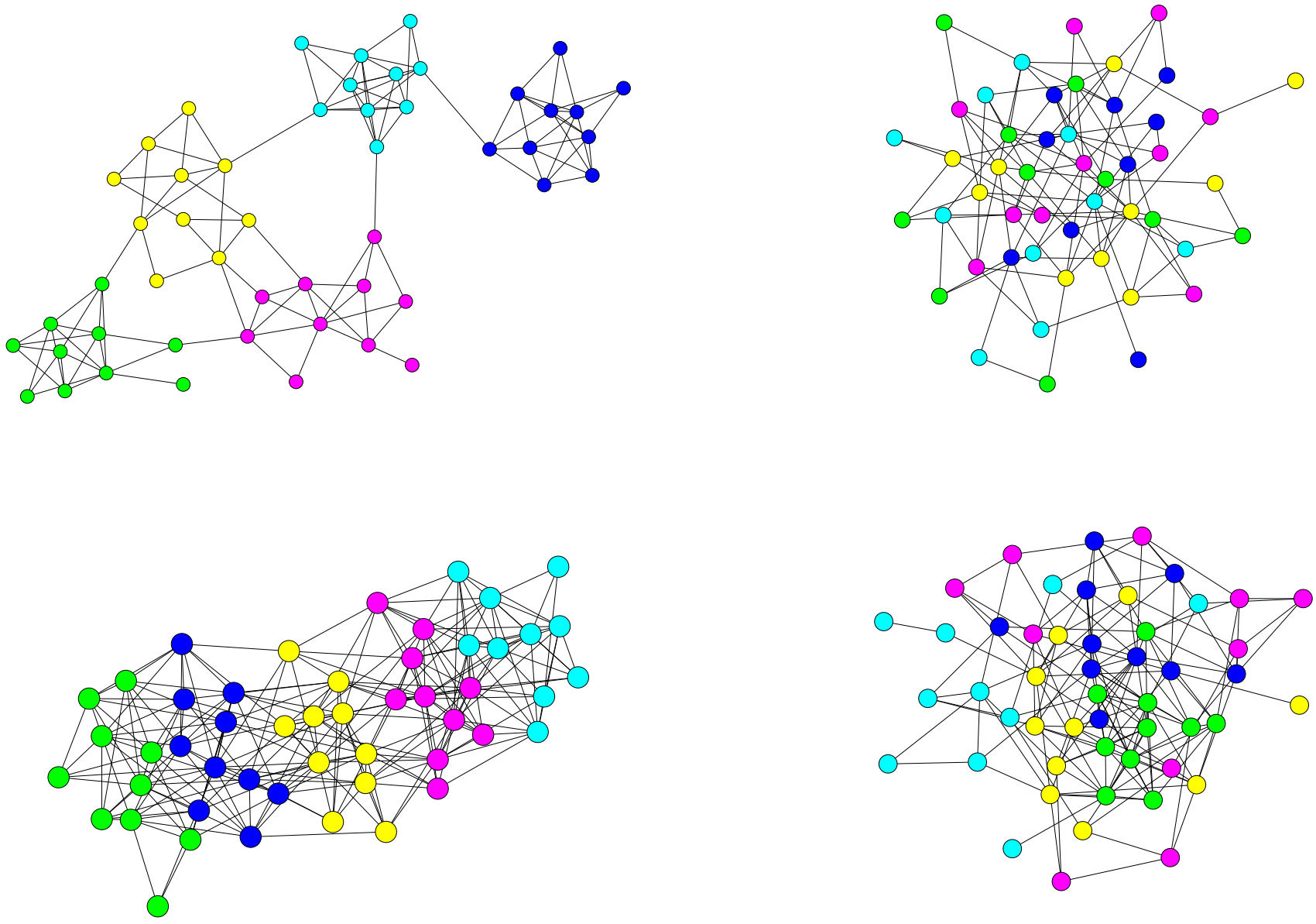
$$f : \text{object} \rightarrow \{\theta_1, \ldots, \theta_k\}$$

- **what are the fundamental parts?**
- **how are these parts organized?**
- **where are the degrees of freedom $\vec{\theta}$?**
- **how can we define an abstract class?**
- **structure — dynamics — function?**

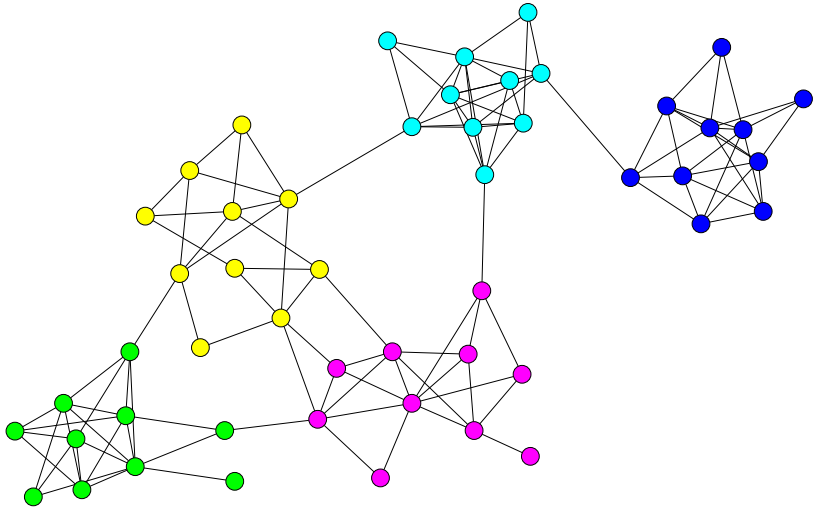
what does ***large-scale network structure*** look like?



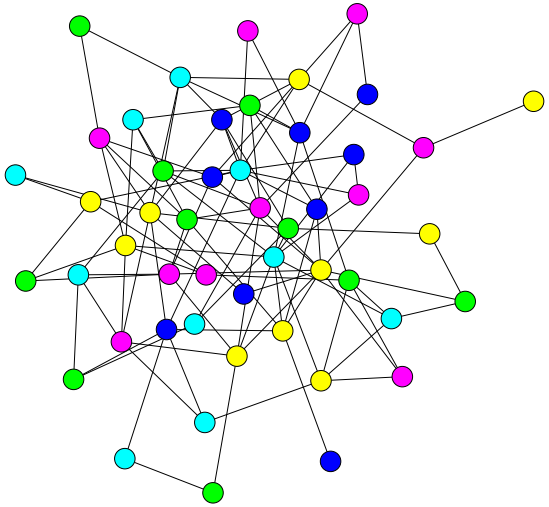
large-scale structure of networks



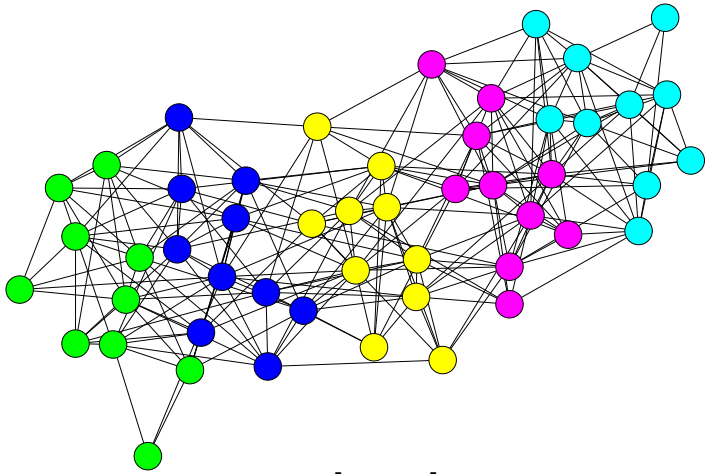
large-scale structure of networks



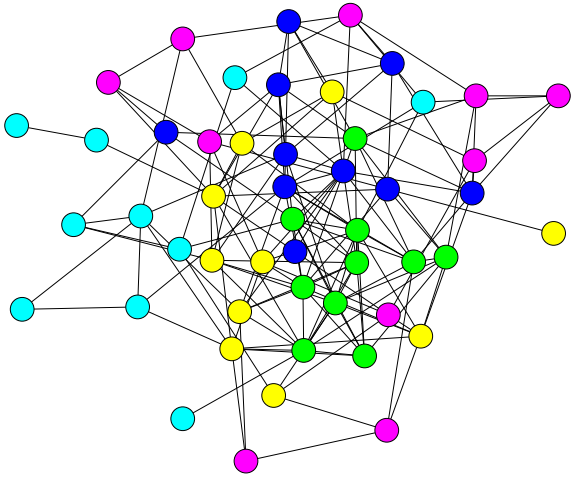
assortative
(edges within groups)



disassortative
(edges between groups)



ordered
(linear hierarchy of groups)

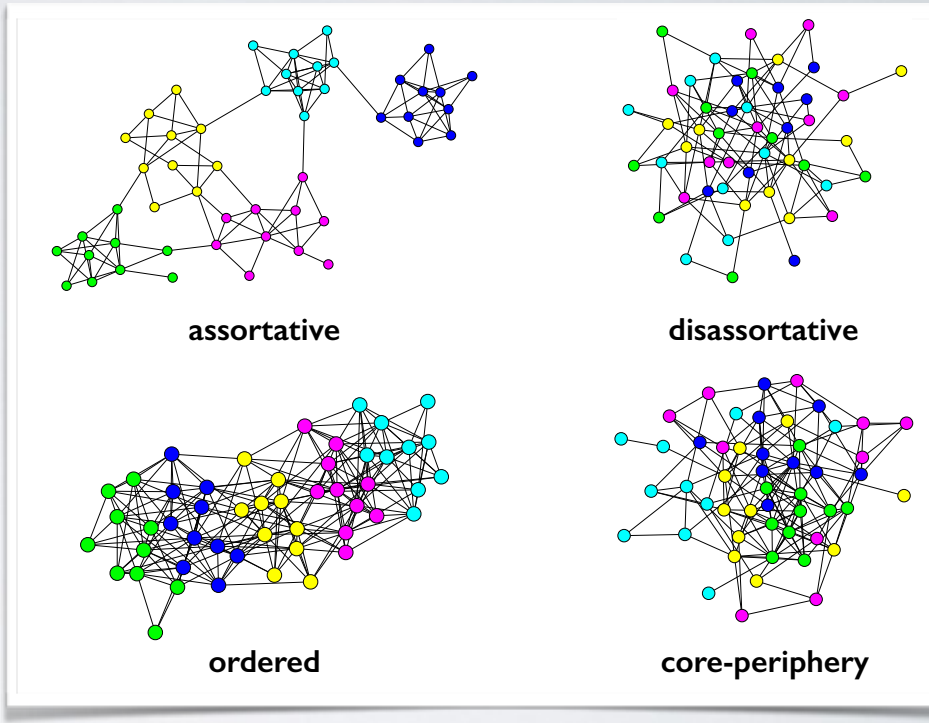


core-periphery
(dense core, sparse periphery)

large-scale structure of networks

large-scale structural analysis

- enormous interest, especially since 2000
- dozens of algorithms for extracting various large-scale patterns
- hundreds of papers published
- spanning Physics, Computer Science, Statistics, Biology, Sociology, and more
- this was one of the first:



Community structure in social and biological networks

M. Girvan*†‡ and M. E. J. Newman*§ *PNAS* 2002

5500+ citations on Google Scholar

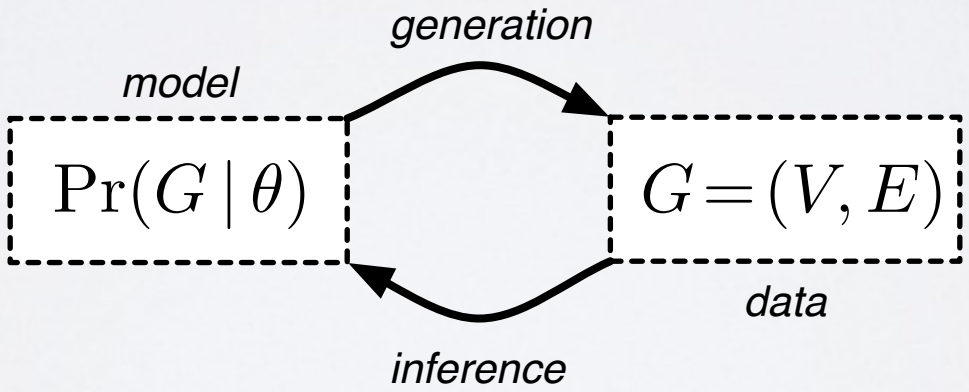
statistical inference and networks

a principled approach : generative models

statistical inference and networks

a principled approach : generative models

- define a parametric probability distribution over networks $\Pr(G \mid \theta)$
- *generation* : given θ, draw G from this distribution
- *inference* : given G, choose θ that makes G likely

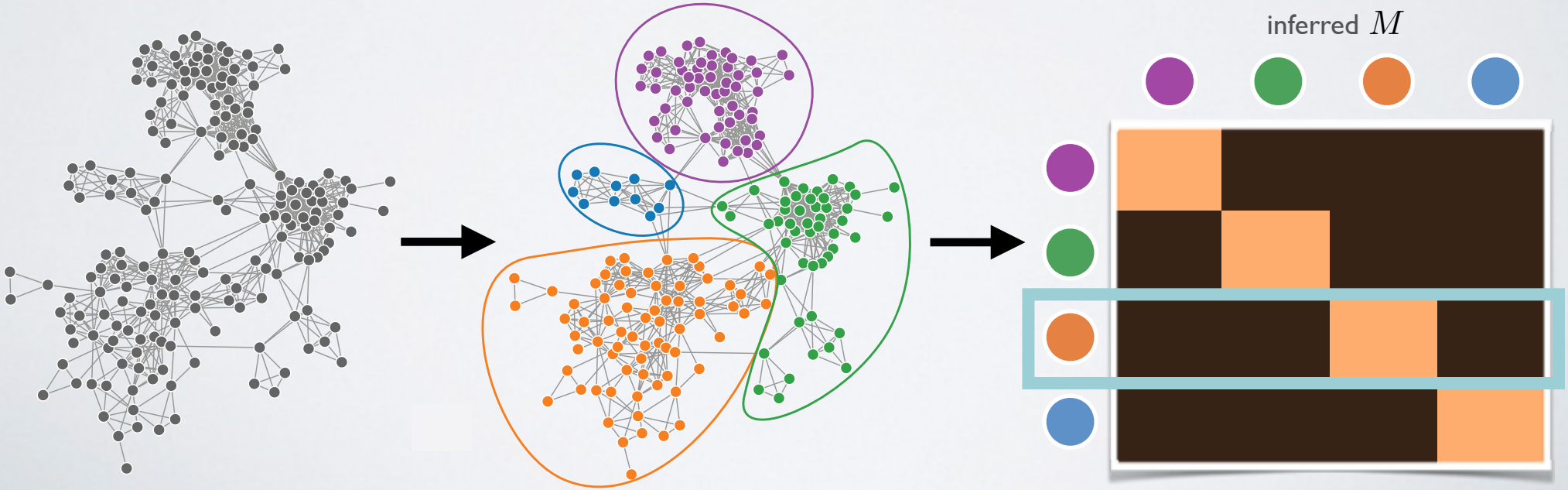


statistical inference and networks

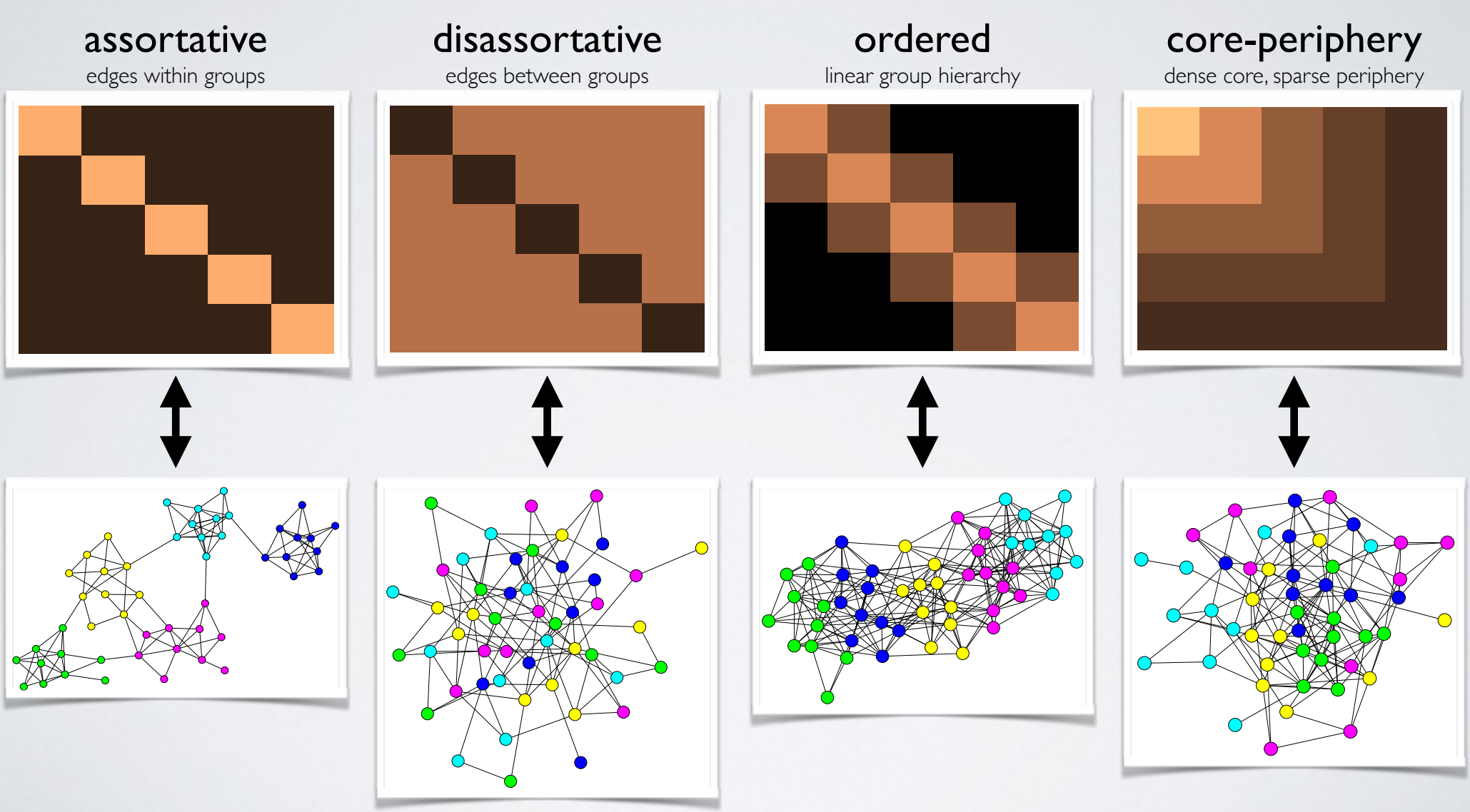
the stochastic block model

- each vertex i has type $z_i \in \{1, \ldots, k\}$ (k vertex types or groups)
- stochastic block matrix M of group-level connection probabilities
- probability that i, j are connected = M_{z_i, z_j}

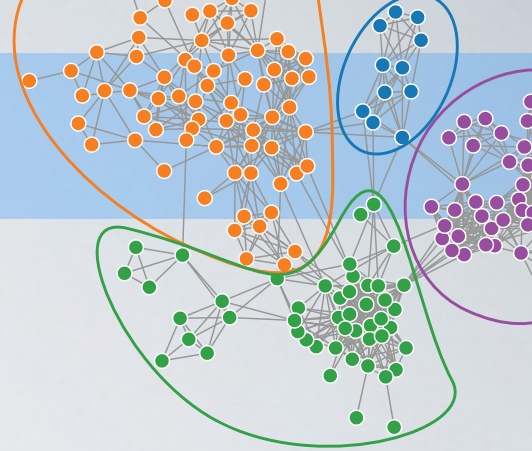
community = vertices with same pattern of inter-community connections



the stochastic block model



the stochastic block model

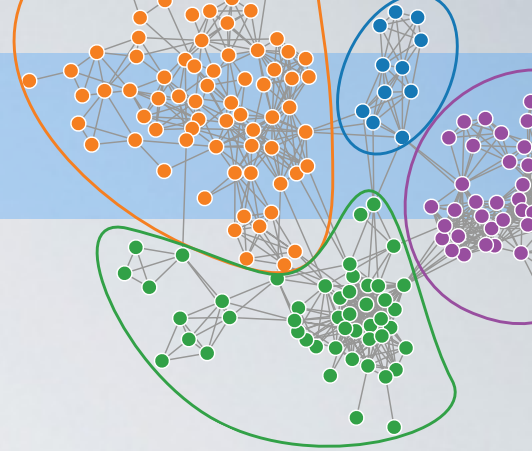


likelihood function

the probability of G given labeling z and block matrix M

$$\Pr(G \mid z, M) = \underbrace{\prod_{(i,j)\in E} M_{z_i,z_j}}_{\text{edge}} \quad / \quad \underbrace{\prod_{(i,j)\notin E} (1 - M_{z_i,z_j})}_{\text{non-edge probability}}$$

the stochastic block model



likelihood function

the probability of G given labeling z and block matrix M

$$\Pr(G \mid z, M) = \underbrace{\prod_{(i,j)\in E} M_{z_i,z_j}}_{\text{edge}} \underbrace{\prod_{(i,j)\notin E} (1 - M_{z_i,z_j})}_{\text{non-edge probability}}$$

or more generally

$$\Pr(A \mid z, \theta) = \prod_{i,j} f\left(A_{ij} \mid \theta_{\mathcal{R}(z_i,z_j)}\right)$$

A_{ij} : value of adjacency

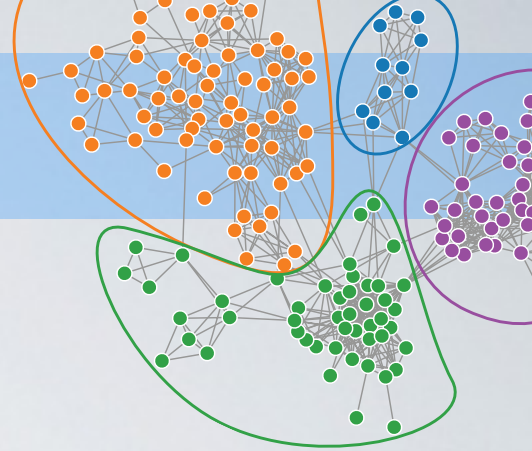
$\mathcal{R}$: partition of adjacencies

f : probability function

$\theta_{a,*}$: pattern for a-type adjacencies

Binomial = simple graphs
Poisson = multi-graphs
Normal = weighted graphs
etc.

the stochastic block model



asymptotically consistent model [see Airoldi et al. *NIPS* 2013, Bickel et al. 2012]

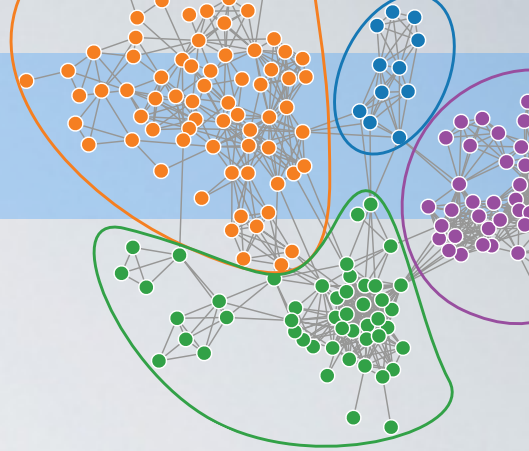
naturally models many large-scale patterns

highly effective in practice [see Karrer & Newman *PRE* 2011]

many nice mathematical features

- general definition of "community" or group
- learns from noisy or missing data [see Clauset et al. 2008]
- predicts missing or spurious or future data [see Clauset et al. 2008, Guimera et al. 2009]
- model comparison tools [this pattern or that pattern?]
- easily augmentable with auxiliary data
- inferred block matrix is interpretable for science
- naturally quantifies uncertainty

the stochastic block model



many flavors, depending on task

- binomial SBM [Holland, Laskey, Leinhardt 1983, Wang & Wong 1987]
- mixed-membership SBM [Airoldi, Blei, Feinberg, Xing 2008]
- hierarchical SBM [Clauset, Moore, Newman 2006,2008, Peixoto 2014]
- fractal SBM [Leskovec et al. 2005]
- infinite relational model [Kemp et al. 2006]
- simple assortative SBM [Hofman & Wiggins 2008]
- degree-corrected SBM [Karrer & Newman 2011]
- SBM + topic models [Ball, Karrer & Newman 2011]
- SBM + vertex covariates [Mariadassou, Robin & Vacher 2010]
- SBM + edge weights [Aicher, Jacobs & Clauset 2013,2014]
- bipartite SBM [Larremore, Clauset & Jacobs 2014]
- and many others

the stochastic block model

3 examples

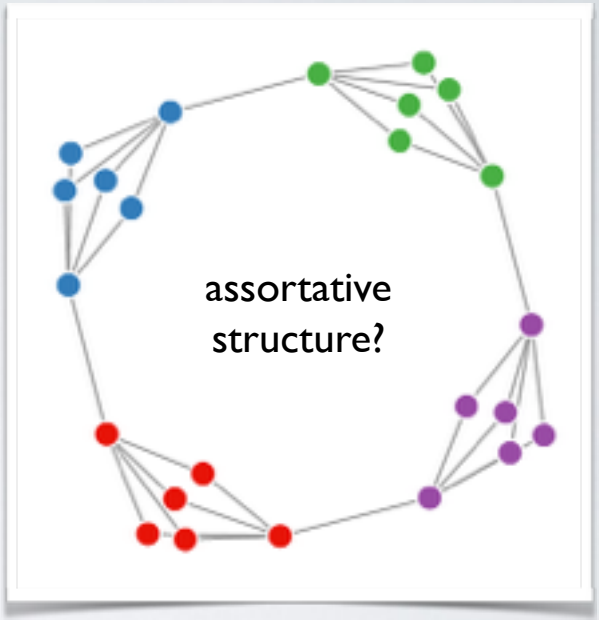
- bipartite community structure (biSBM)
- weighted community structure (WSBM)
- change-point detection in evolving networks (GHRG) *[see Leto Peel's talk Thursday at NetSci]*

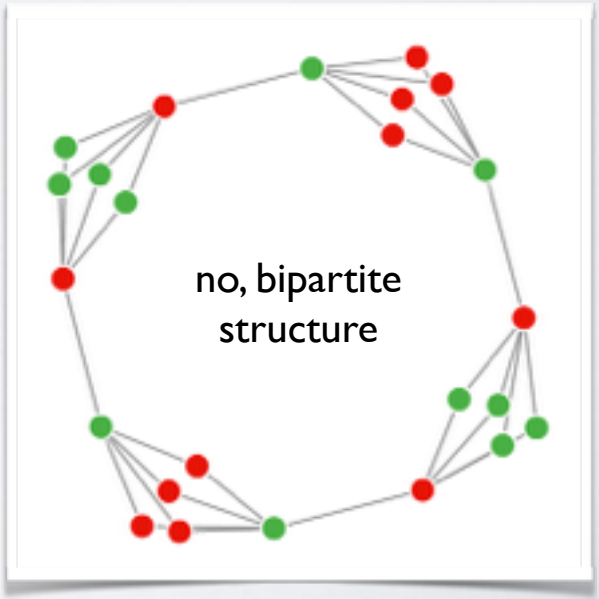
example 1: bipartite networks

many networks are bipartite

- scientists and papers (co-authorship networks)
- actors and movies (co-appearance networks)
- words and documents (topic modeling)
- plants and pollinators
- genes and genomes
- etc.

most analyses focus on one-mode projections which discard information



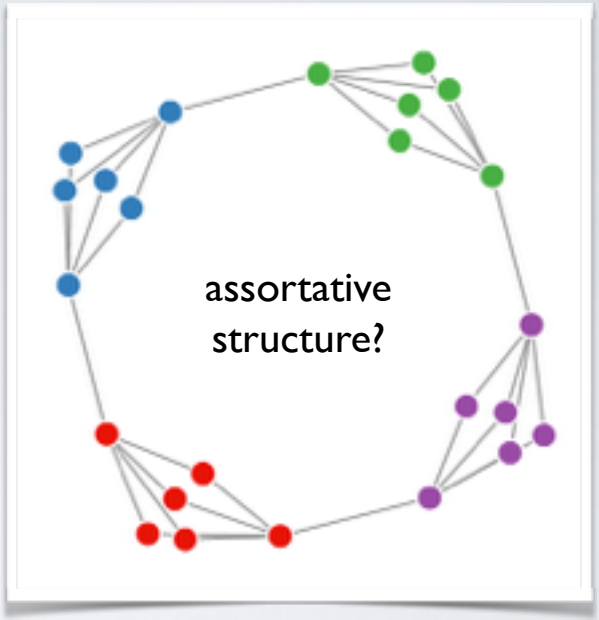


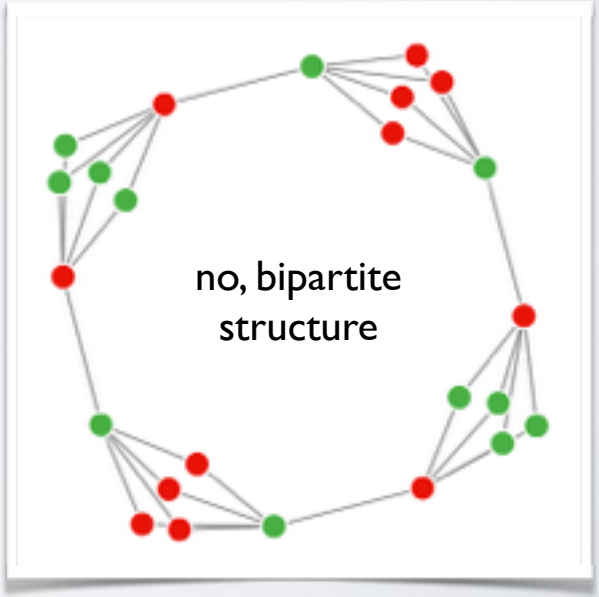
example 1: bipartite networks

bipartite stochastic block model (biSBM)

- exactly the SBM, but model knows network is bipartite
- if $\text{type}(z_i) = \text{type}(z_j)$

  then require $M_{z_i,z_j} = 0$
- inference proceeds as before

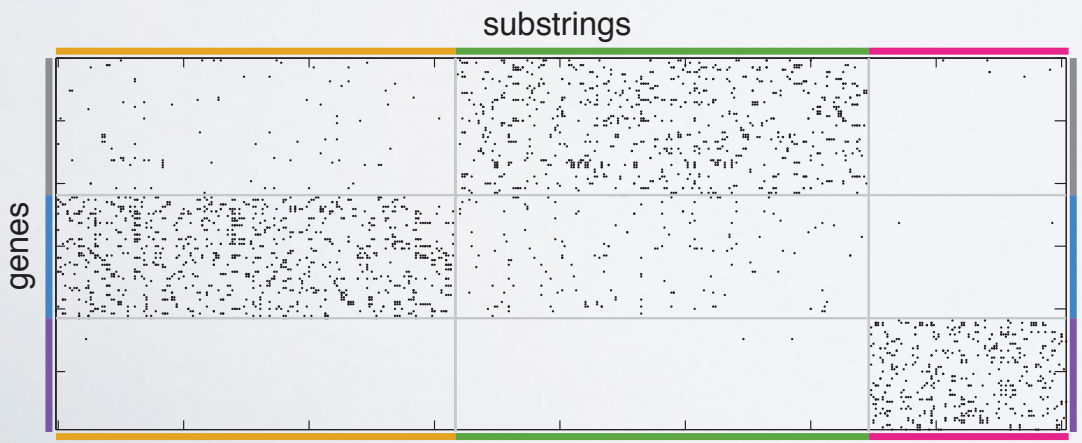


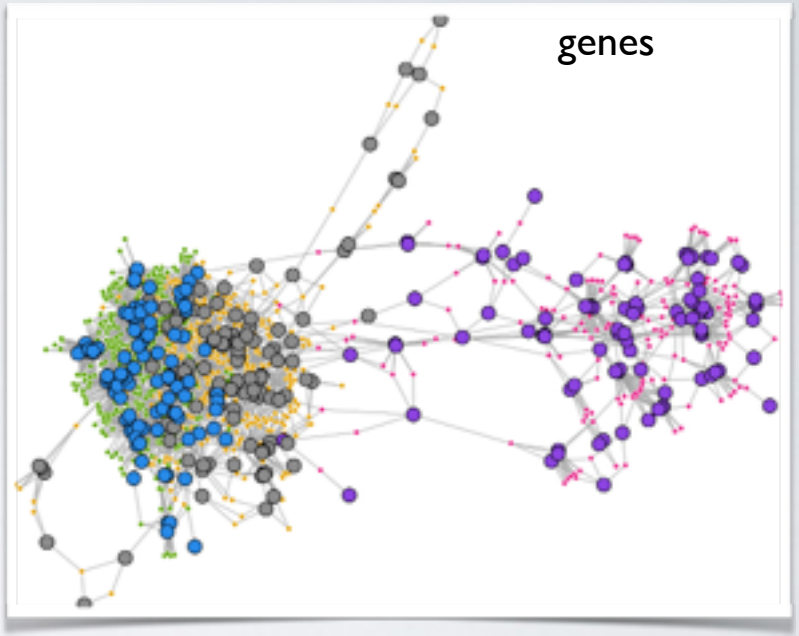


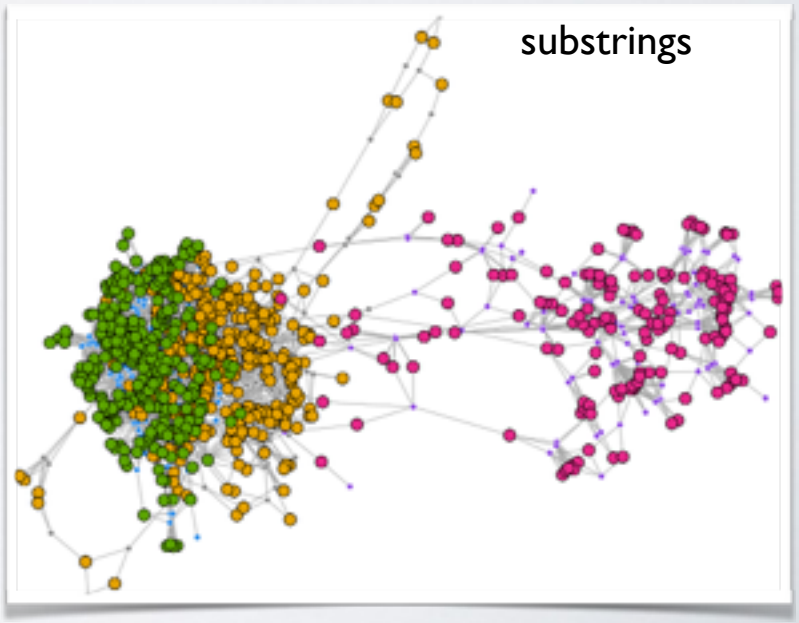
example 1: bipartite networks

bipartite stochastic block model (biSBM)

- SBM can *learn* bipartite structure, but biSBM much more efficient, accurate
- biSBM always find pure-type communities
- more accurate than modeling one-mode projections (even weighted projections)
- finds communities in both modes







example 2: weighted networks

most interactions are weighted

- interaction frequency, strength, character, outcome, etc.

- thresholding discards information, can obscure underlying structure

example 2: weighted networks

most interactions are weighted

- **interaction frequency, strength, character, outcome, etc.**
- **thresholding discards information, can obscure underlying structure**

weighted SBM:

$$\ln \Pr(G \mid M, z, \theta, f) = \alpha \ln \Pr(G \mid M, z) + (1-\alpha) \Pr(G \mid \theta, z, f)$$

edge-existence
[binomial distribution]
M_{z_i, z_j}

edge-weights
[exponential-family distribution]
θ_{z_i, z_j}

Poisson, Normal, Gamma,
Exponential, Pareto, etc.

infer z, M, θ

example 2: weighted networks

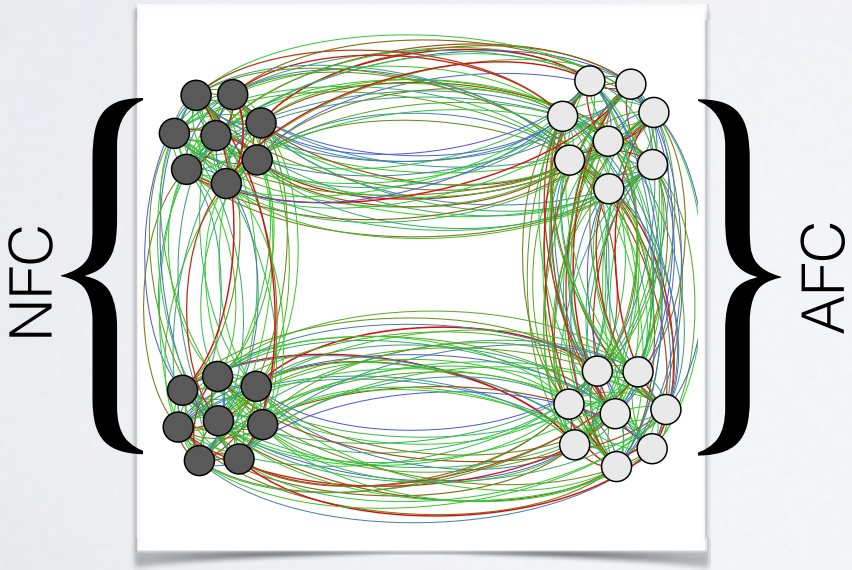
NFL 2009 season

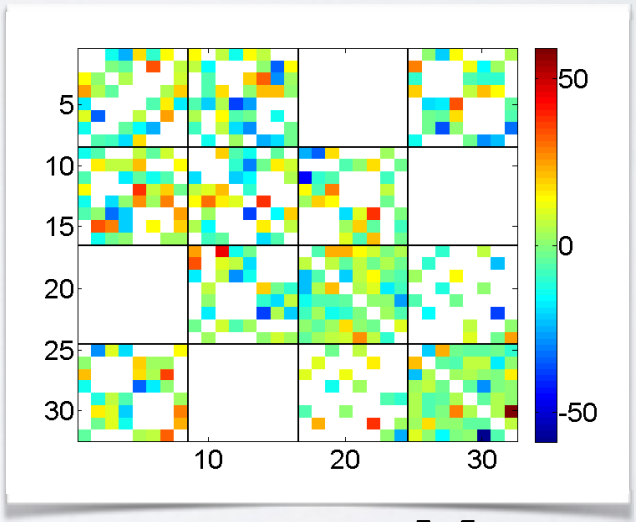
- 32 teams, 2 “divisions”, 4 “subdivisions”
- *edge existence*: who plays whom
- *edge weight*: mean score difference

example 2: weighted networks

NFL 2009 season

- 32 teams, 2 “divisions”, 4 “subdivisions”
- SBM ($\alpha = 1$) recovers subdivisions perfectly



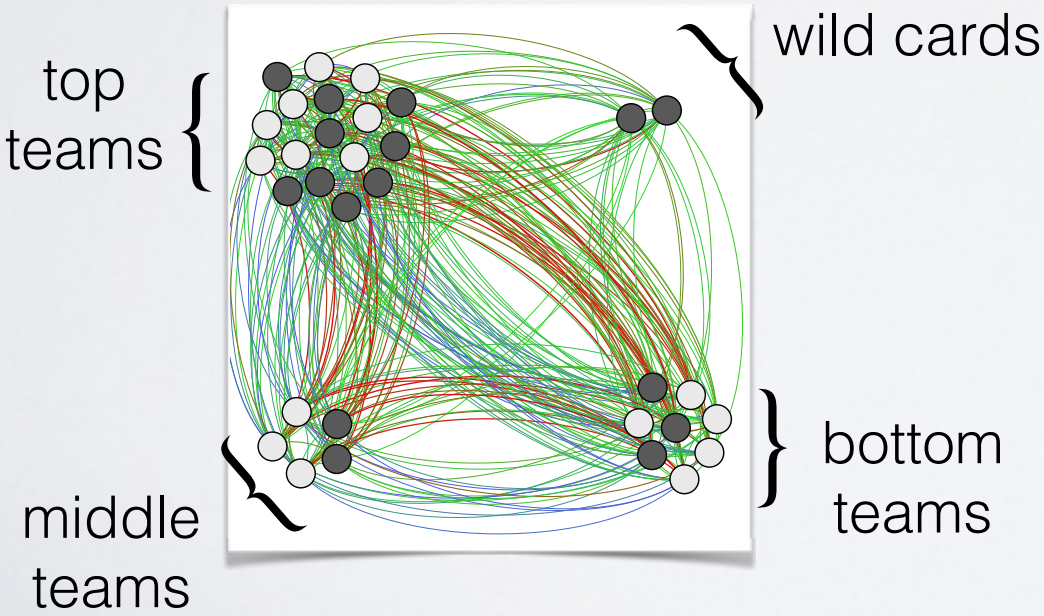


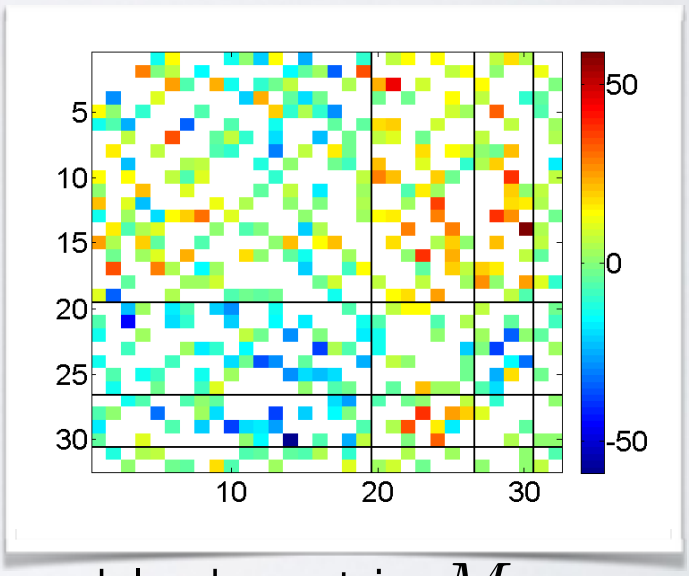
block matrix M, z

example 2: weighted networks

NFL 2009 season

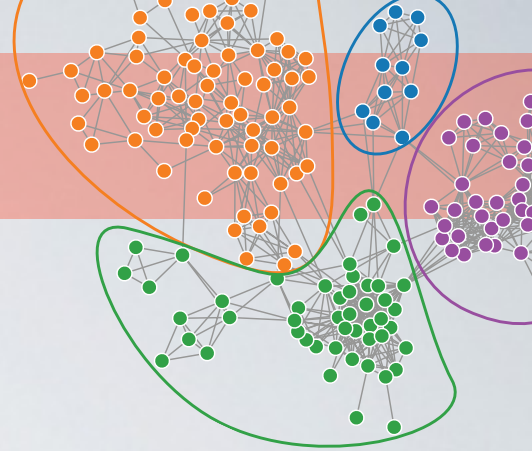
- 32 teams, 2 “divisions”, 4 “subdivisions”
- WSBM ($\alpha = 0$) recovers team skill hierarchy



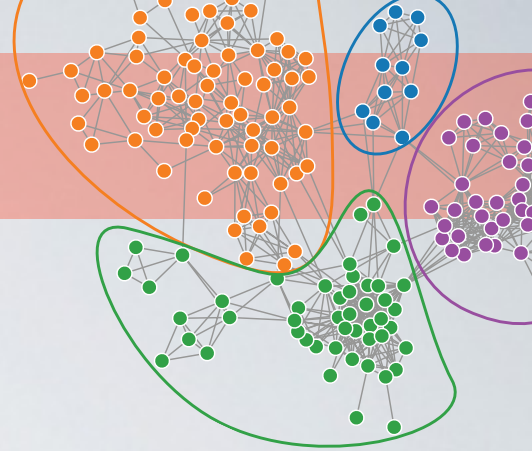


block matrix M, z

to summarize



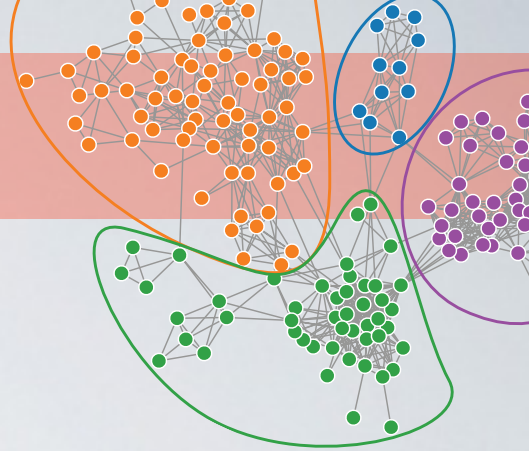
to summarize



generative models for networks

statistically principled approach for finding structure in networks

to summarize



generative models for networks

statistically principled approach for finding structure in networks

the stochastic block model

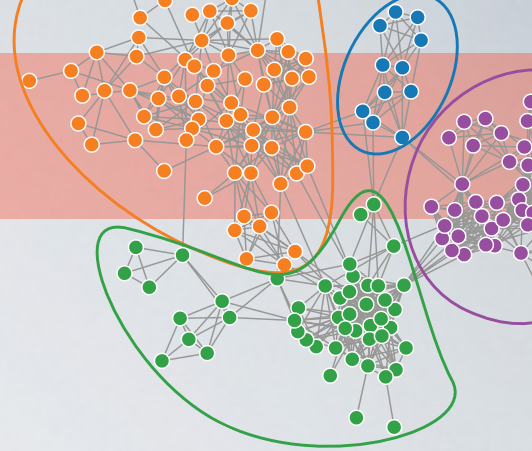
communities = vertices with similar community-connectivity patterns

general approach to infer such large-scale patterns

inference is fast, scalable

can incorporate auxiliary information [bipartite, weighted, directed, time, etc.]

to summarize



generative models for networks

statistically principled approach for finding structure in networks

the stochastic block model

communities = vertices with similar community-connectivity patterns

general approach to infer such large-scale patterns

inference is fast, scalable

can incorporate auxiliary information [bipartite, weighted, directed, time, etc.]

many opportunities

applications abound:

gene recombination, gene regulation, social interactions, etc. etc.

methodological tasks:

formalize specific structural hypotheses, model assessment, model comparison, etc.

fin

code + data available at

hierarchical SBM santafe.edu/~aaronc/hierarchy/

weighted SBM santafe.edu/~aaronc/wsbm/

bipartite SBM danlarremore.com/bipartiteSBM/

change-point detection SBM gdriv.es/letopeel/code.html

further reading

- Larremore, Clauset and Jacobs, "Efficiently inferring community structure in bipartite networks." Preprint (2014) [arxiv:1403.2933]
- Peel and Clauset, "Detecting change points in the large-scale structure of evolving networks." Preprint (2014) [arxiv:1403.0989]
- Aicher, Jacobs and Clauset, "Learning latent block structure in weighted networks." To appear, *Journal of Complex Networks* (2014) [arxiv:1404.0431]
- Larremore, Clauset and Buckee, "A network approach to analyzing highly recombinant malaria parasite genes." *PLOS Computational Biology* **9**, e1003268 (2013) [arxiv:1308.5254]
- Aicher, Jacobs and Clauset, "Adapting the stochastic block model to edge-weighted networks." *ICML Ws* (2013) [arxiv:1305.5782]
- Clauset, Moore, and Newman, "Hierarchical structure and the prediction of missing links in networks" *Nature* **453**, 98-101 (2008) [arxiv:0811.0484]