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)
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

\[ f : G \rightarrow \{x_1, \ldots, x_k\} \]

- 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)

ordered
(linear hierarchy of groups)

disassortative
(edges between 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
a principled approach: generative models

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

```
NGDYKEKVSNNLRAIFNKIYENLNDPKLKKHYQKDAPNY
NGDYKKKVSNNLKTIFKKIYDALKDTVKETYKDDPNY
NGDYKEKVSNNLRAIFKKIYDALEDTVKETYKDDPNY
```

alignment position \( t \)

- calculate alignment scores
- convert to alignment indicators
- remove short aligned regions
- extract highly variable regions
the stochastic block model

assortative
edges within groups

disassortative
edges between groups

ordered
linear group hierarchy

core-periphery
dense core, sparse periphery
the stochastic block model

likelihood function

the probability of $G$ given labeling $z$ and block matrix $M$

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

edge / non-edge probability
the stochastic block model

likelihood function

the probability of $G$ given labeling $z$ and block matrix $M$

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

or more generally

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

$A_{ij}$: 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]
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]
inferred block matrix is interpretable for science
naturally quantifies uncertainty
model comparison tools  [this pattern or that pattern?]
the stochastic block model

many flavors, depending on task

- mixed-membership SBM [Airoldi, Blei, Feinberg, Xing 2008]
- hierarchical SBM [Clauset, Moore, Newman 2006, 2008]
- 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
malaria gene recombination networks

ultra-condensed malaria 101

- malaria kills ~1 million people / year
  [mostly children]

- caused by *Plasmodium* parasite

- *Plasmodium’s* var genes critical to infection

- frequent recombination "shuffles" var gene sequences = mosaic sequence pattern
  [similar to HIV and pneumococcus]

- can we find constraints on var recombination?

images: D. Ferguson (Oxford)
malaria gene recombination networks

a (bipartite) network hypothesis

- vertex $A = \text{var gene domain}$ [e.g. DBL$\alpha$]
- vertex $B = k$-mer [shared substring of length $k$]
- genes connect to all their $k$-mers

```
NGDYKEKVSNNLRAIFN
K
I
Y
E
N
L
N
D
P
L
K
K
H
Y
Q
K
D
A
P
N
Y

NGDYKKKVSNNLKTIFKK
I
Y
D
A
L
K
D
T
V
K
E
T
Y
K
D
D
P
N
Y

NGDYKEKVSNNLRAIFKK
I
Y
D
A
L
E
D
T
V
K
E
T
Y
K
D
D
P
N
Y
```
malaria gene recombination networks

a (bipartite) network hypothesis

- vertex $A = var$ gene domain [e.g. DBL$\alpha$]
- vertex $B = k$-mer [shared substring of length $k$]
- genes connect to all their $k$-mers

- what would different recombination patterns mean?

random recombination
no group structure

vs

constrained recombination
strong group structure
malaria gene recombination networks

data from Rask et al.,

- 7 whole-genome sequences
  [4 field isolates, 3 lab strains]
- 307 total DBLα domain genes
- apply bipartite SBM to gene/k-mer network
malaria gene recombination networks


- 7 whole-genome sequences
  [4 field isolates, 3 lab strains]
- 307 total DBLα domain genes
- apply bipartite SBM to gene/k-mer network
gene-gene networks

- each gene has multiple 'highly variable regions'
- different HVRs have different block structures
- some more / less similar

malaria gene recombination networks
malaria gene recombination networks

what do we learn?

• HVRs with similar recombination patterns may indicate shared functional constraints

• HVRs with uncorrelated patterns may indicate diversifying selection (immune evasion)

• this system for generating and managing recombinant sequences appears highly conserved across *Plasmodium* clade (evolved >2 million years ago)¹

• networks are key to understanding structure of recombined sequences

• stochastic block model is a *good* model of recombination

¹: see Larremore et al. "Ancient modularity maintains antigen mosaics across Laveranian malaria parasites" In prep (2014).
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.
code + data available at

- hierarchical SBM santafe.edu/~aaronc/hierarchy/
- weighted SBM santafe.edu/~aaronc/wsbm/
- bipartite SBM danlarremore.com/bipartiteSBM/

further reading

- 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]