Metadata & Community Detection

Daniel Larremore
Assistant Professor
Dept. of Computer Science
& BioFrontiers Institute

CSCI 5352

University of Colorado **Boulder**

daniel.larremore@colorado.edu
@danlarremore

community detection:
find groups of nodes in networks, based only on patterns of edges

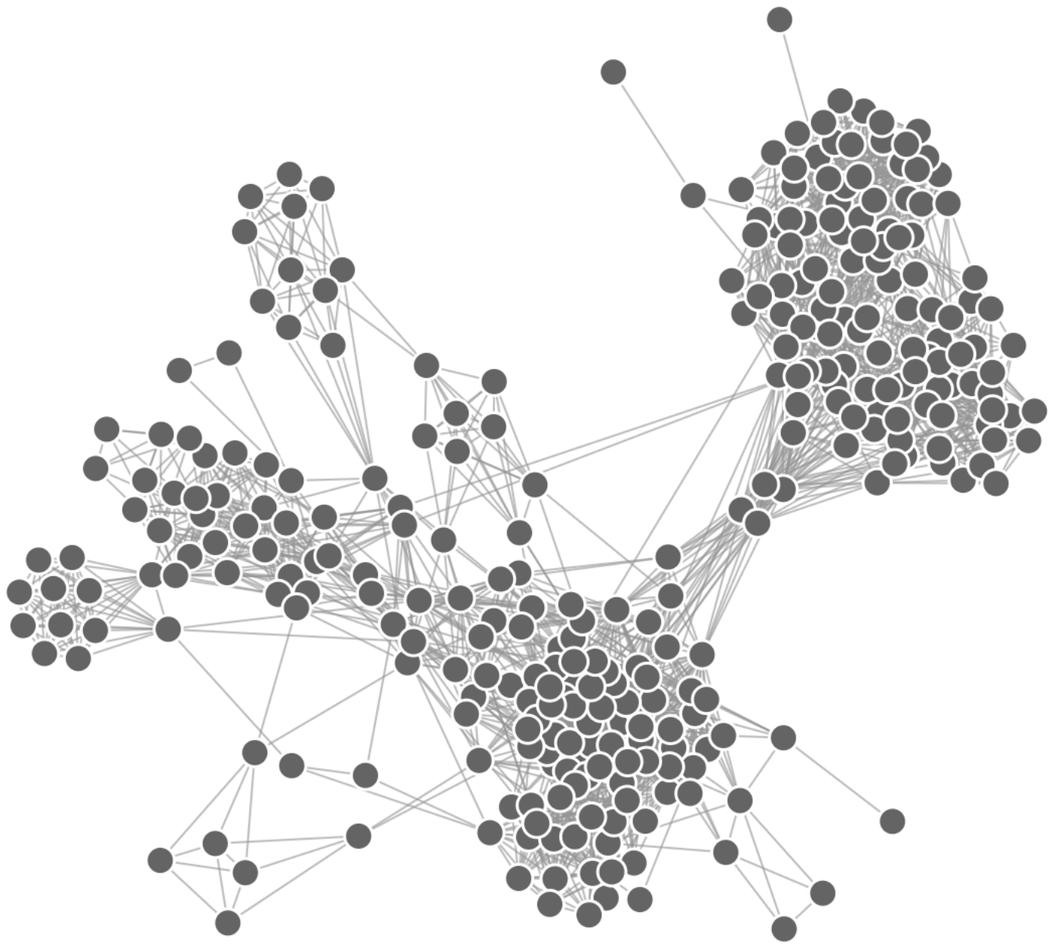
ground truth:
the right answer

metadata:
well, data.

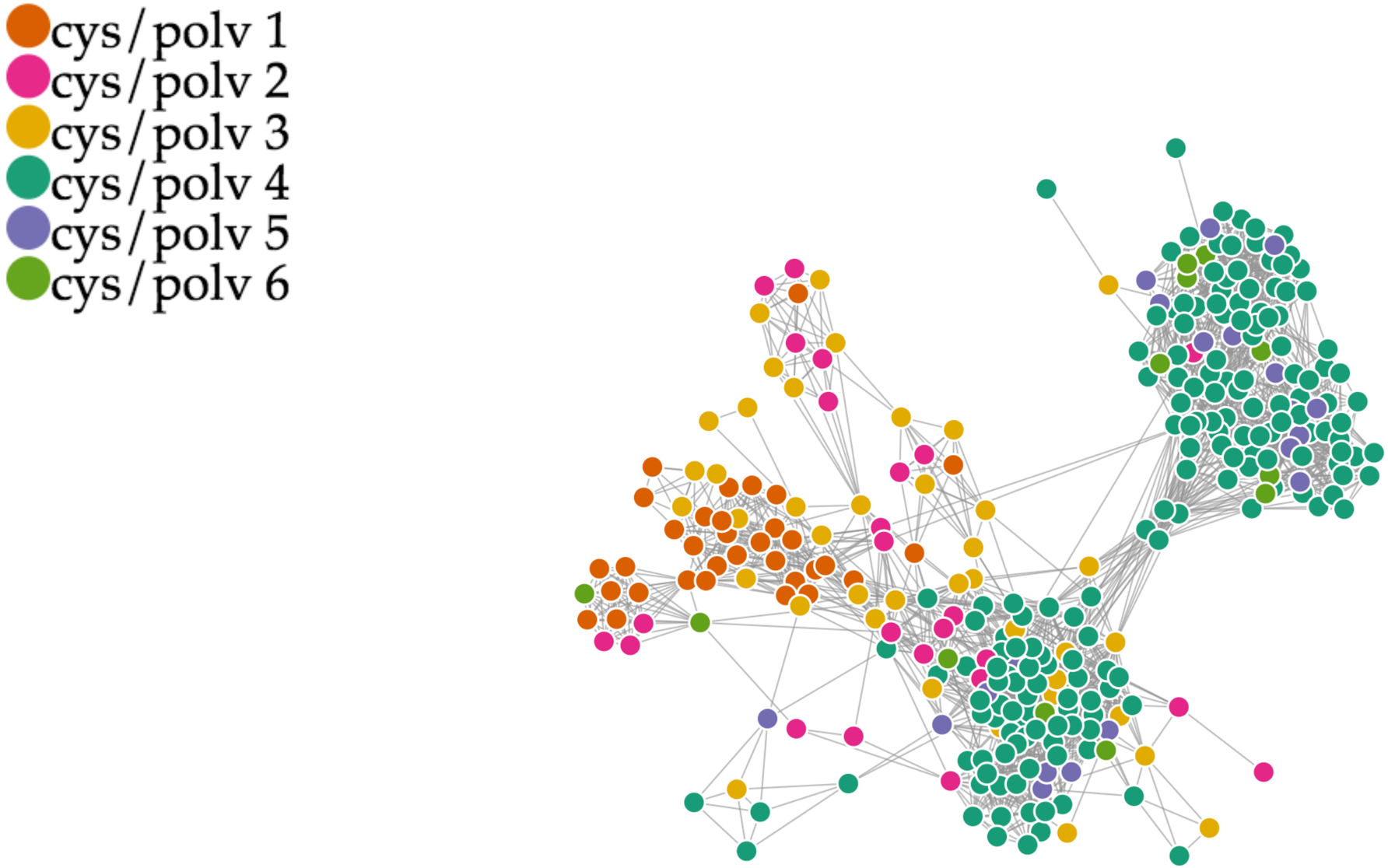
metadata:
categorical vertex attributes or labels.

Real world network vertices have metadata

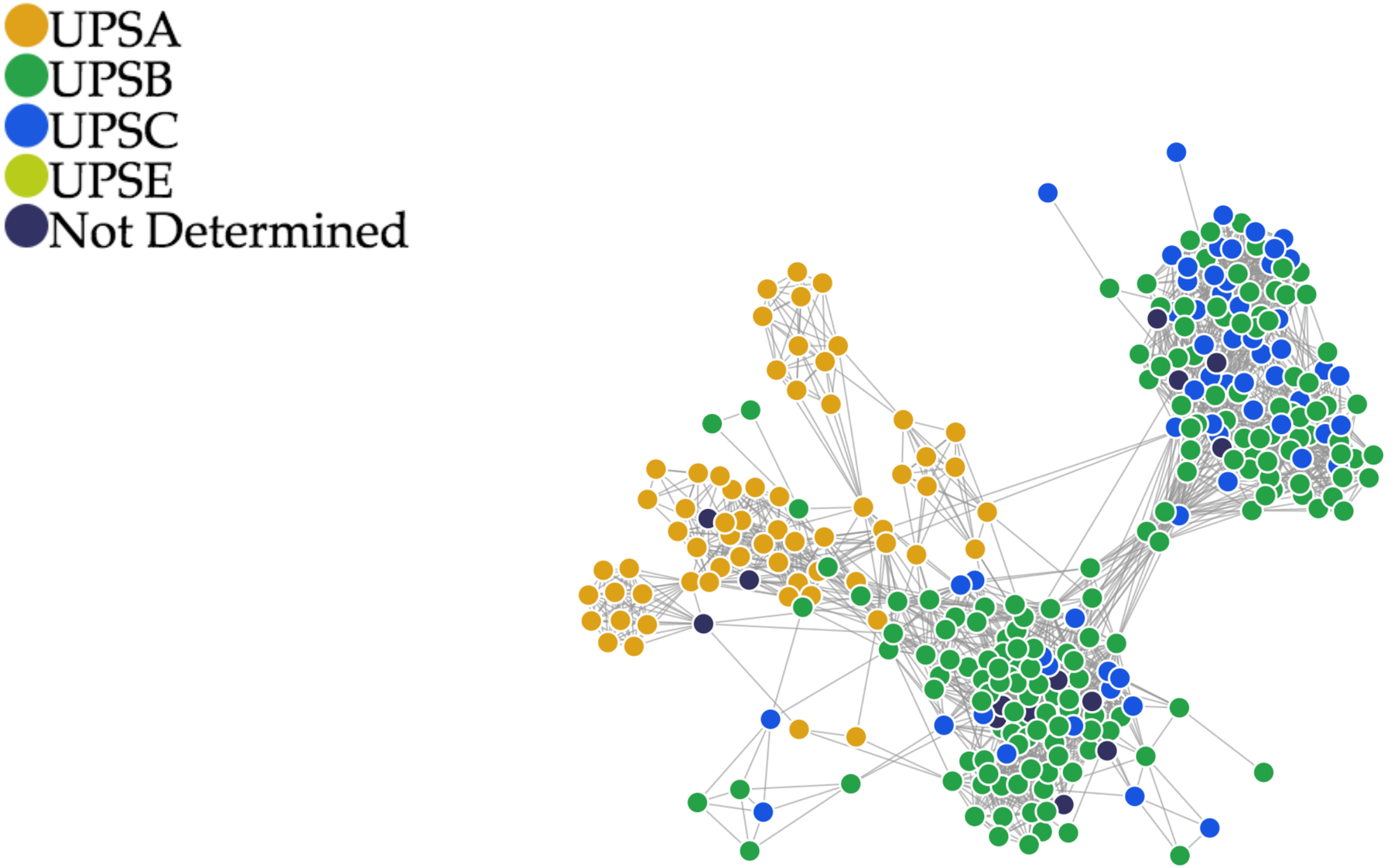
social network	age, gender, ethnicity, education
food web	trophic level, body mass
Internet	capacity, physical location
protein interaction	association with disease



DBL, Clauset, Buckee. PLoS Comp. Biol. **9**(10), e1003268 (2013)

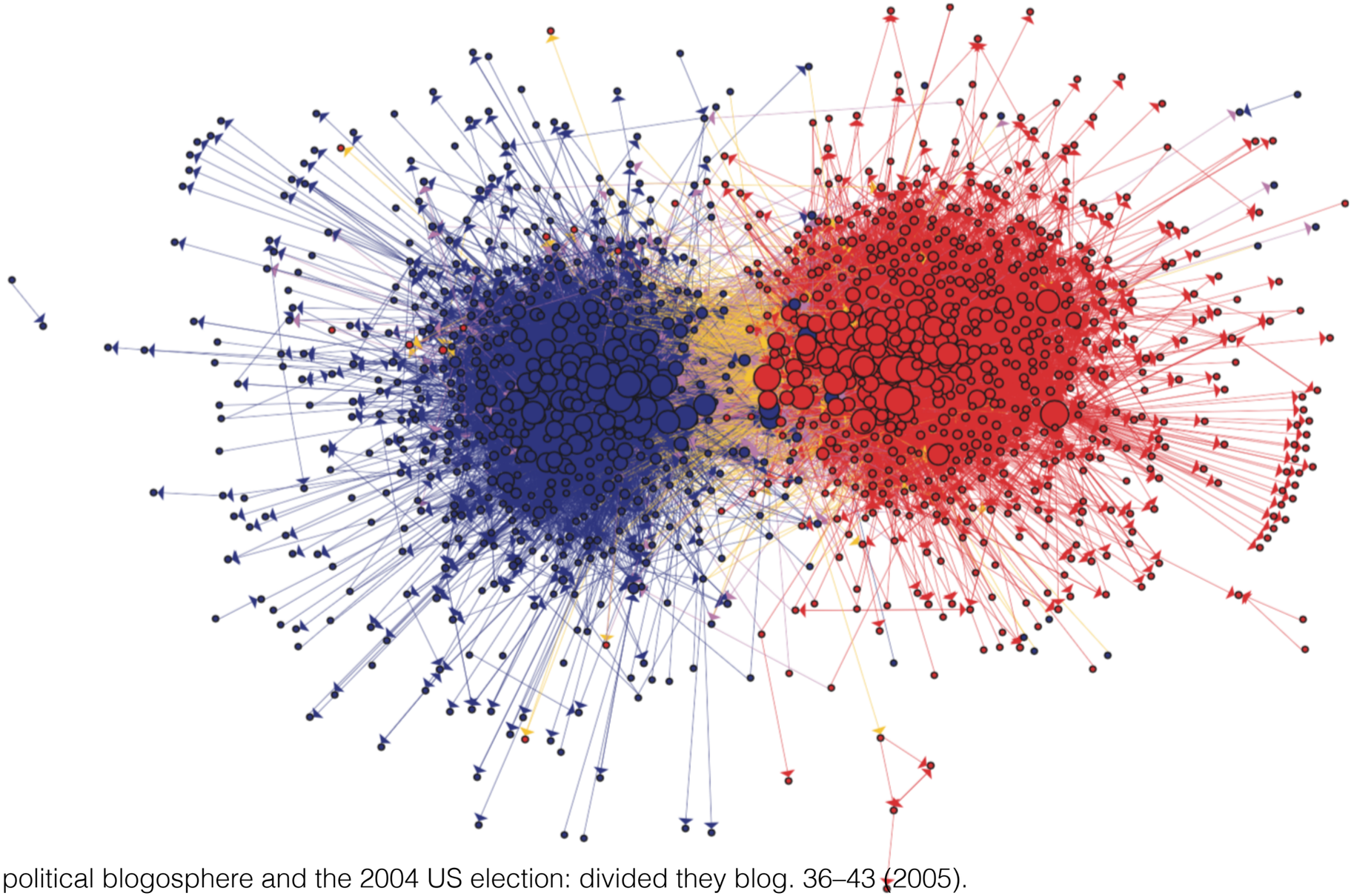


DBL, Clauset, Buckee. PLoS Comp. Biol. **9**(10), e1003268 (2013)



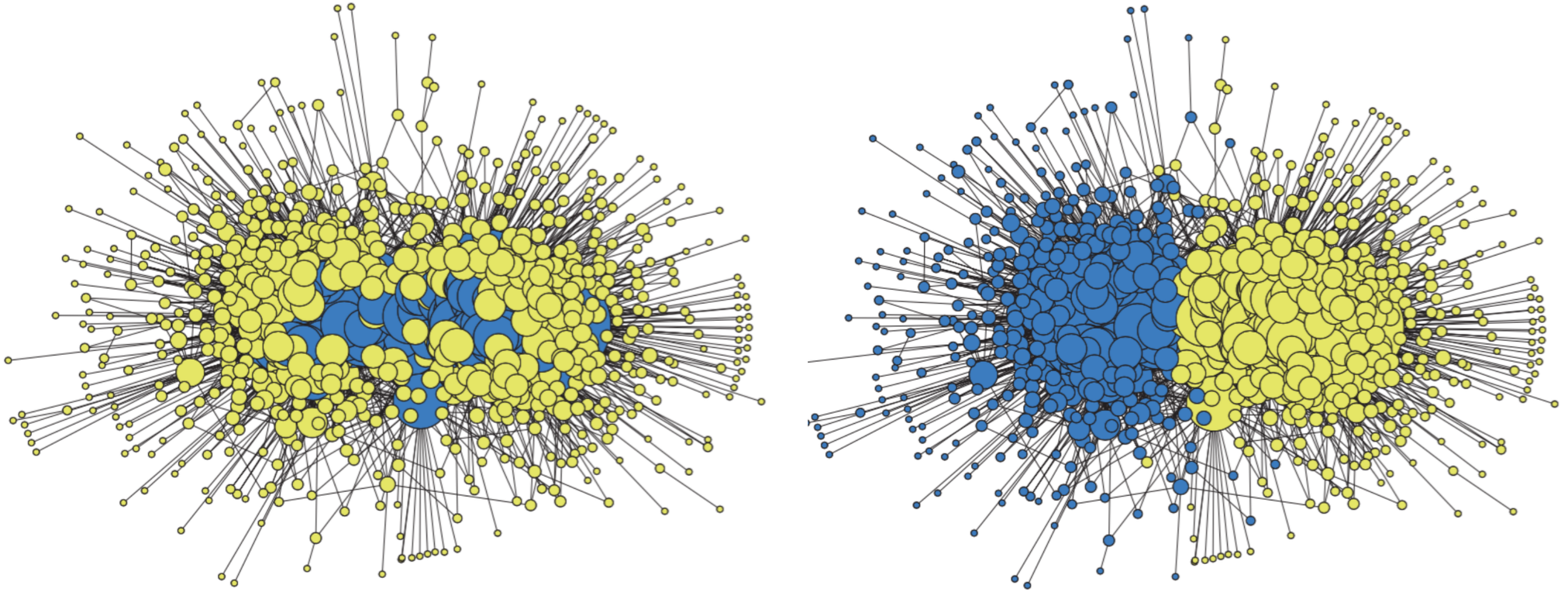
DBL, Clauset, Buckee. PLoS Comp. Biol. **9**(10), e1003268 (2013)

Metadata offer clues for network formation



Adamic, Glance. The political blogosphere and the 2004 US election: divided they blog. 36–43 (2005).

Recovering metadata implies sensible methods



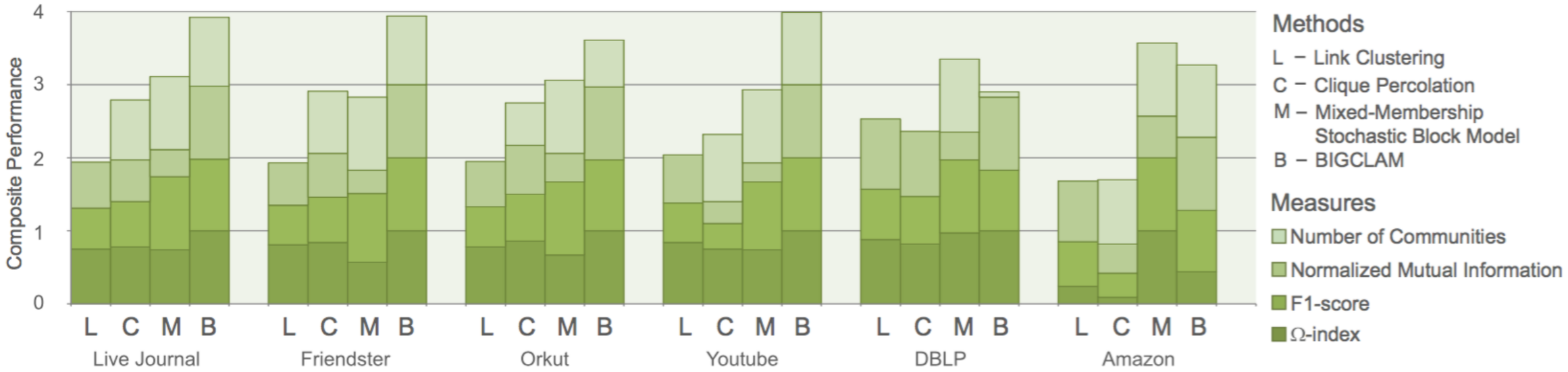
(a) Without degree-correction

(b) With degree-correction

The degree-corrected stochastic block model (right) finds communities which correspond to known metadata.

Karrer, Newman. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107 (2011).
Adamic, Glance. The political blogosphere and the 2004 US election: divided they blog. 36–43 (2005).

Metadata may become the target for comm. det.



Metadata M are often treated as *ground truth* T,
sometimes explicitly (and very often implicitly),
to evaluate the accuracy of community detection algorithms.

Yang & Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach (2013).

Outline

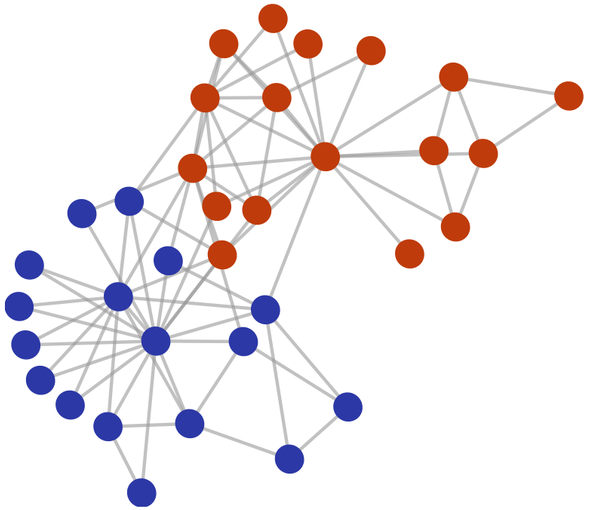
1. **Metadata are not ground truth** for community detection.

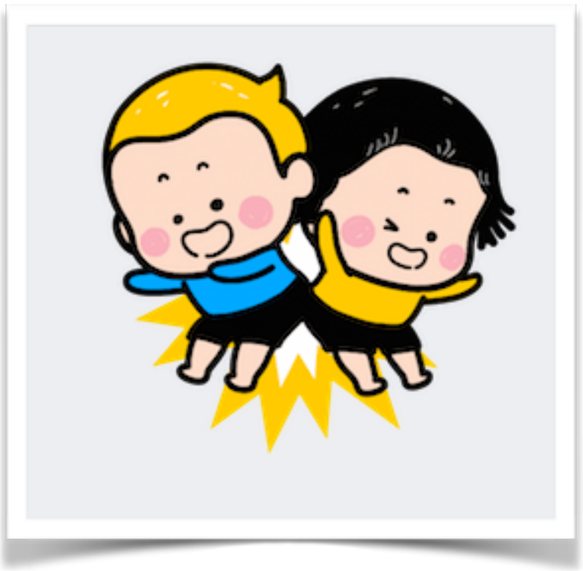
2. **There is no ground truth** for community detection in real world networks.

3. **Metadata are data**; we present two methods to treat them as such.

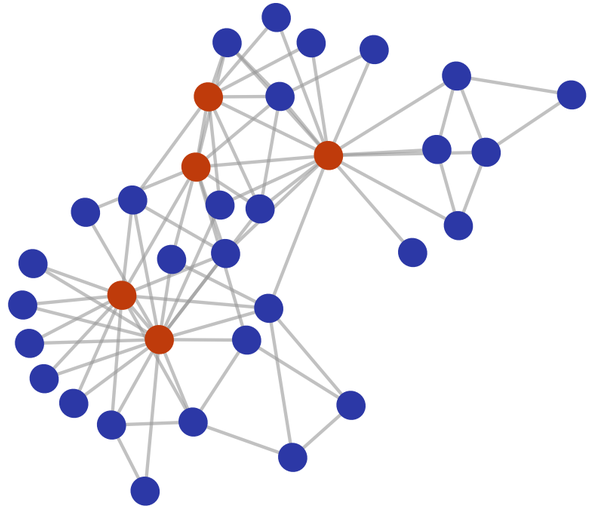
What happens when we fail?

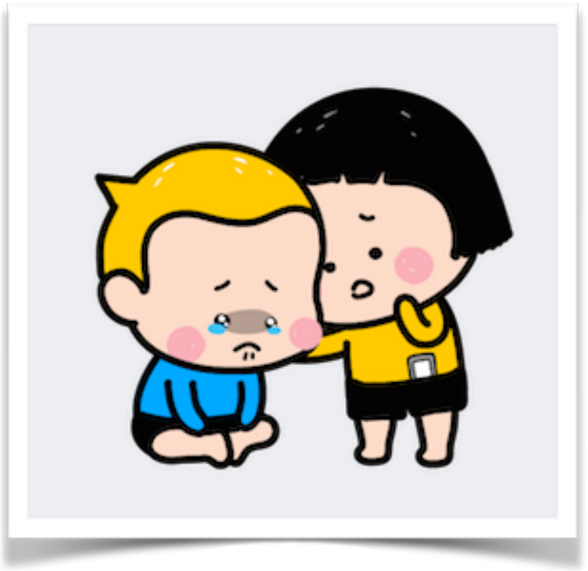
network G + method $f \rightarrow$ communities $C = f(G)$ vs M metadata





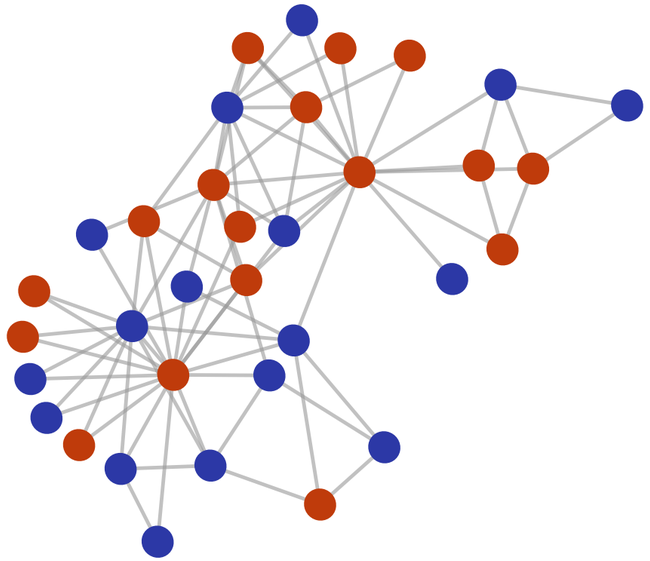
"i like this method"





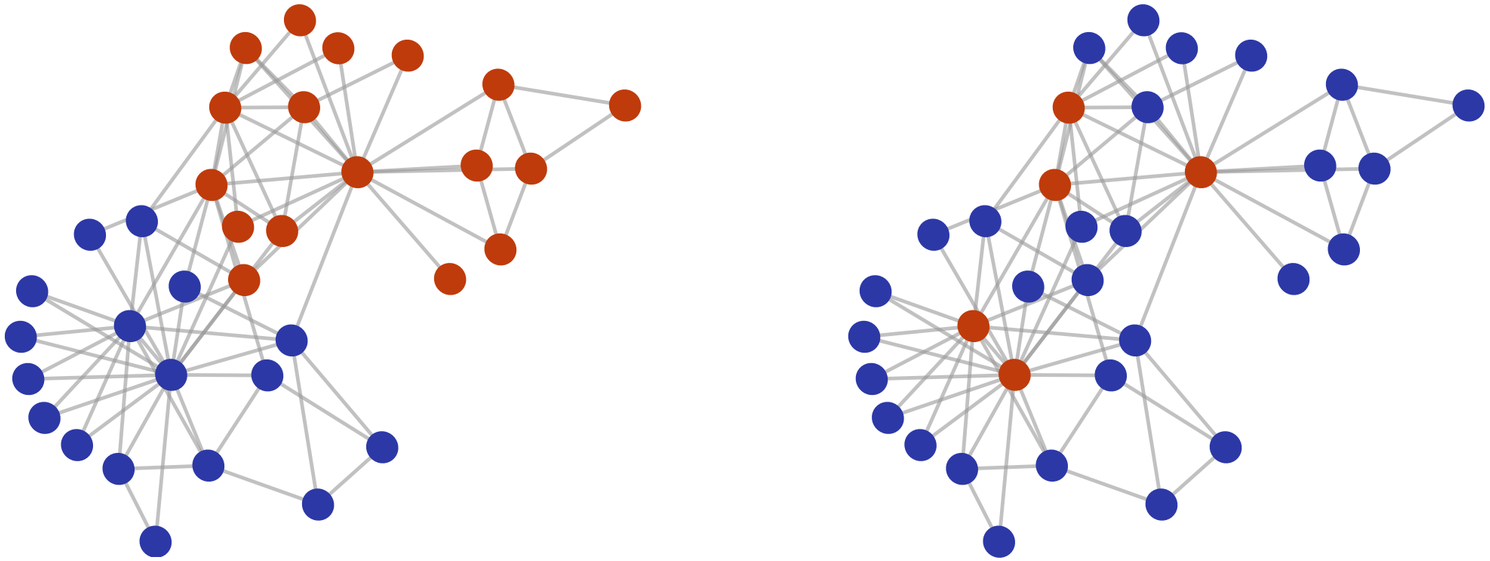
"i do not like this method"

When communities ≠ metadata



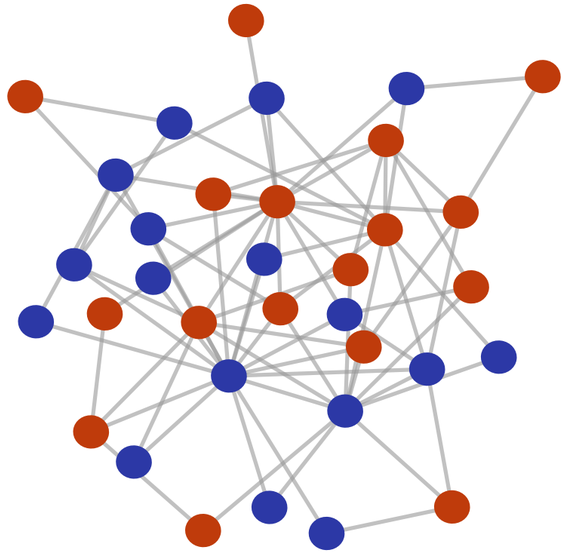
(i) metadata are unrelated to network structure

When communities ≠ metadata



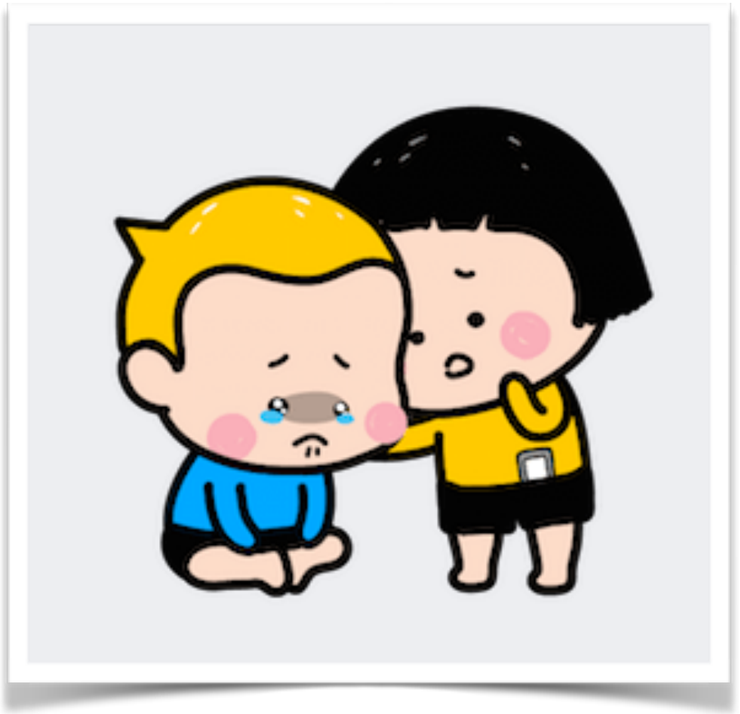
(ii) detected communities and metadata capture different aspects of network structure

When communities ≠ metadata



(iii) the network has no structure

When communities ≠ metadata

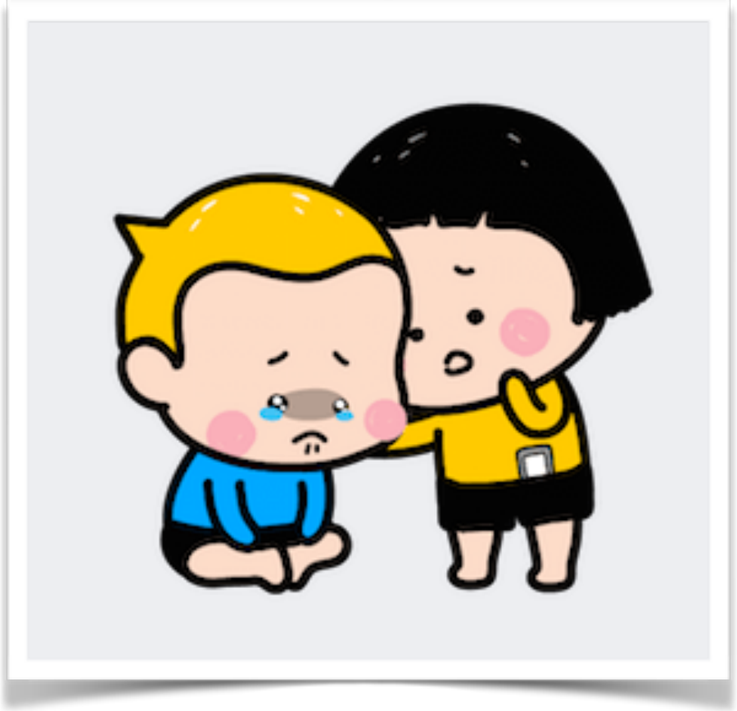


(iv) the algorithm does not perform well

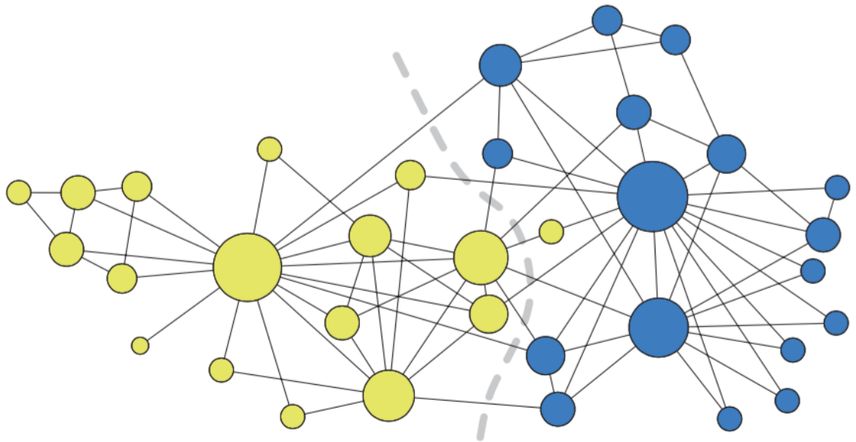
Typically, we assume this is the only possible cause.

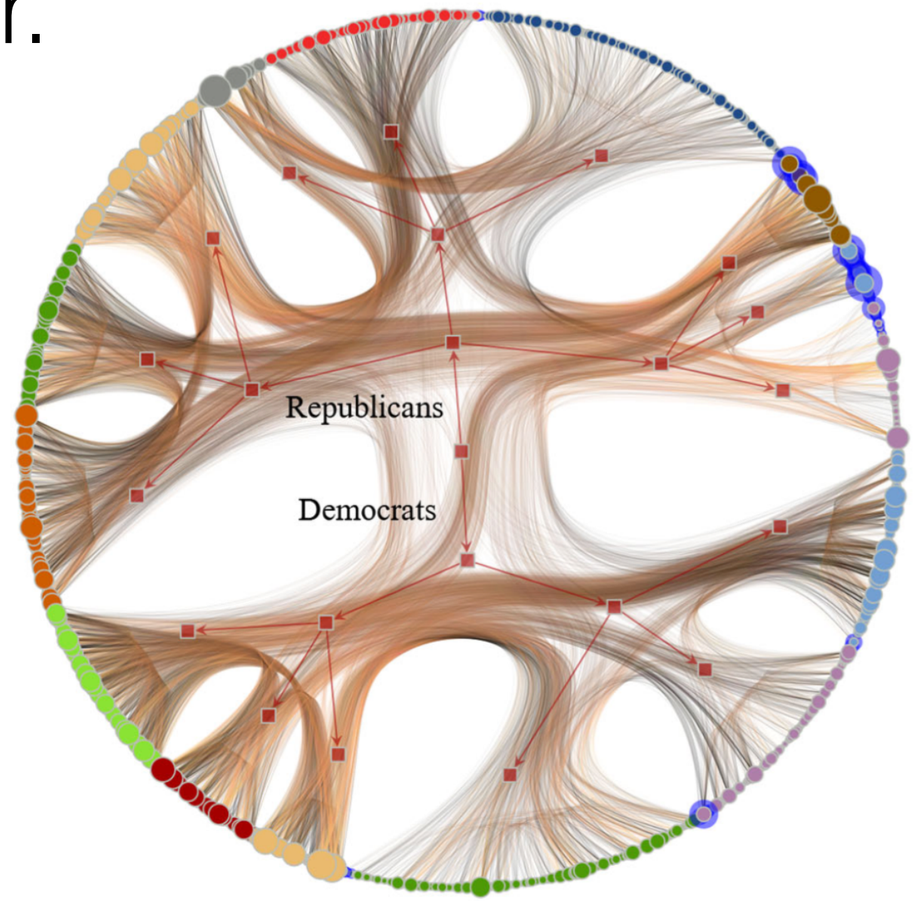
Disambiguation problems

(i) M unrelated to network structure
(ii) C and M capture different aspects of network structure
(iii) the network has no structure
(iv) the algorithm does not perform well



You see (only) what you look for.



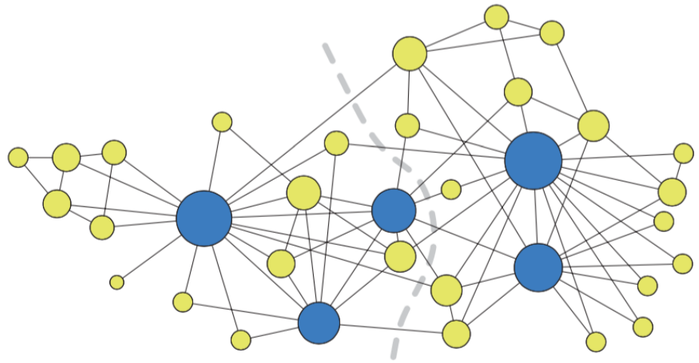


node 9 is *interesting*, not misclassified.

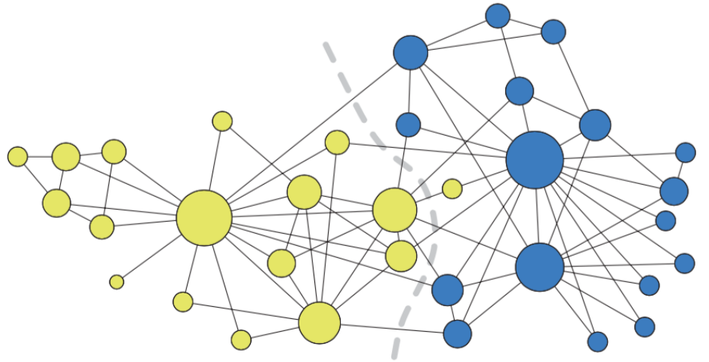
US politics is more than just red and blue.

Zachary, W. W. An information flow model for conflict and fission in small groups. Journal of anthropological research 452–473 (1977).
Karrer, Newman. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107 (2011).
Peixoto, T. P. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Phys. Rev. X 4, 011047 (2014).

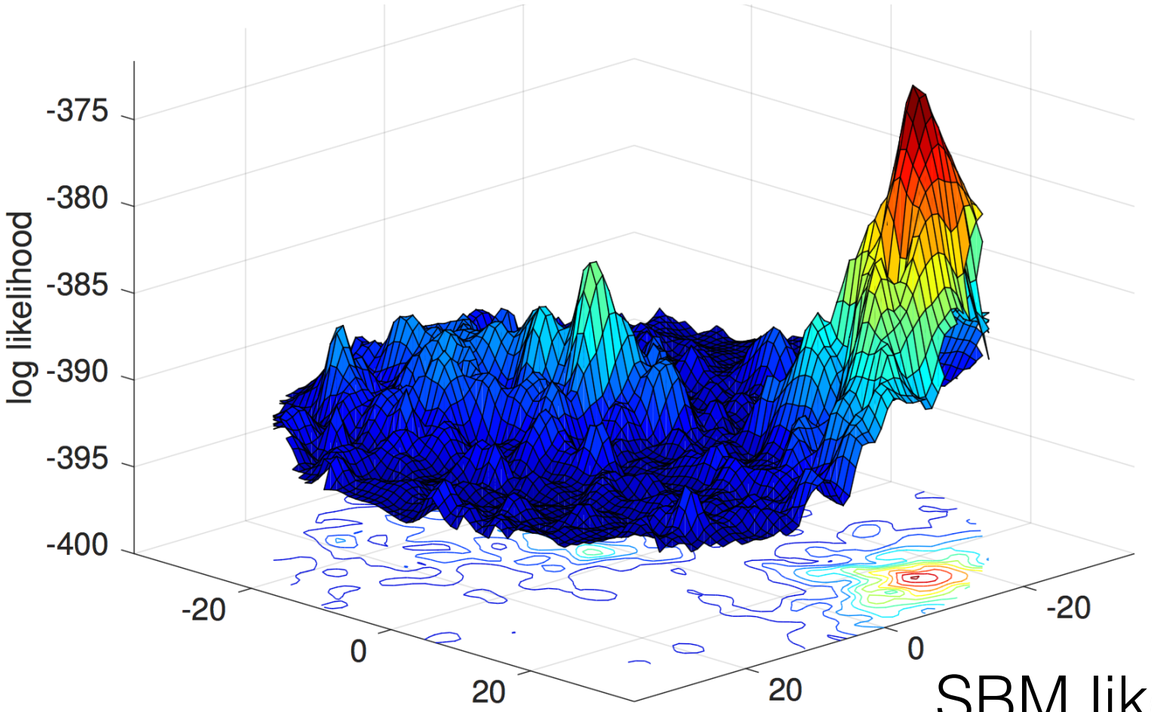
Different **models** see different **communities**

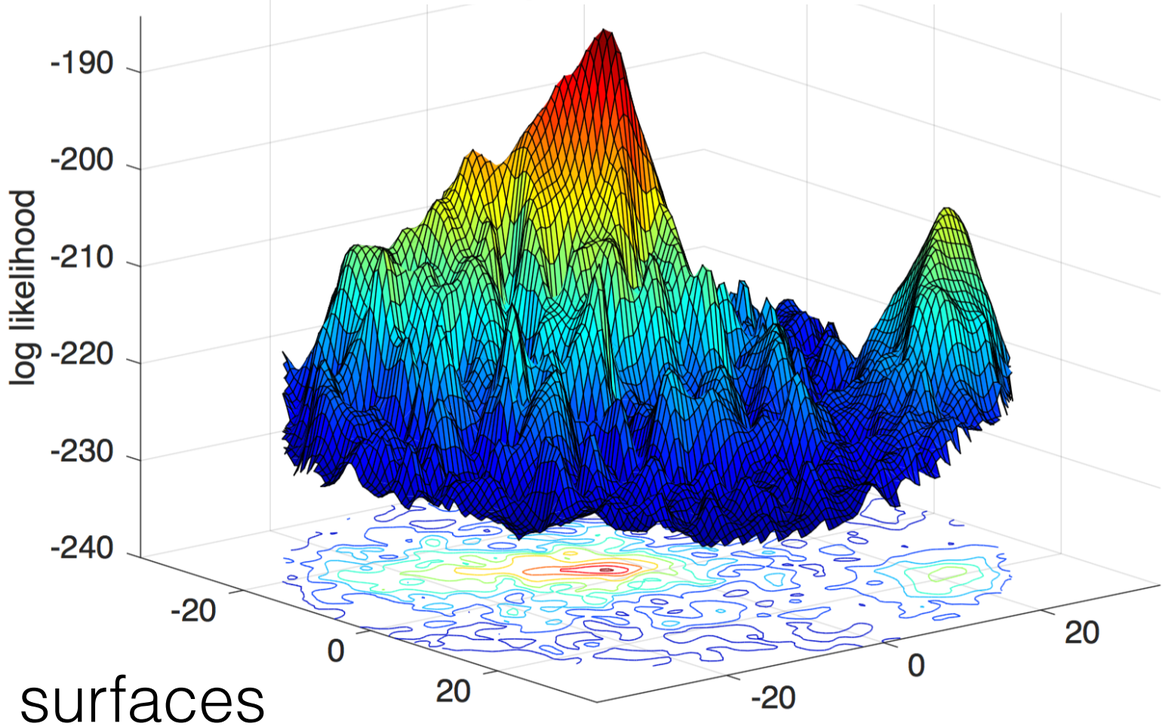


without degree correction



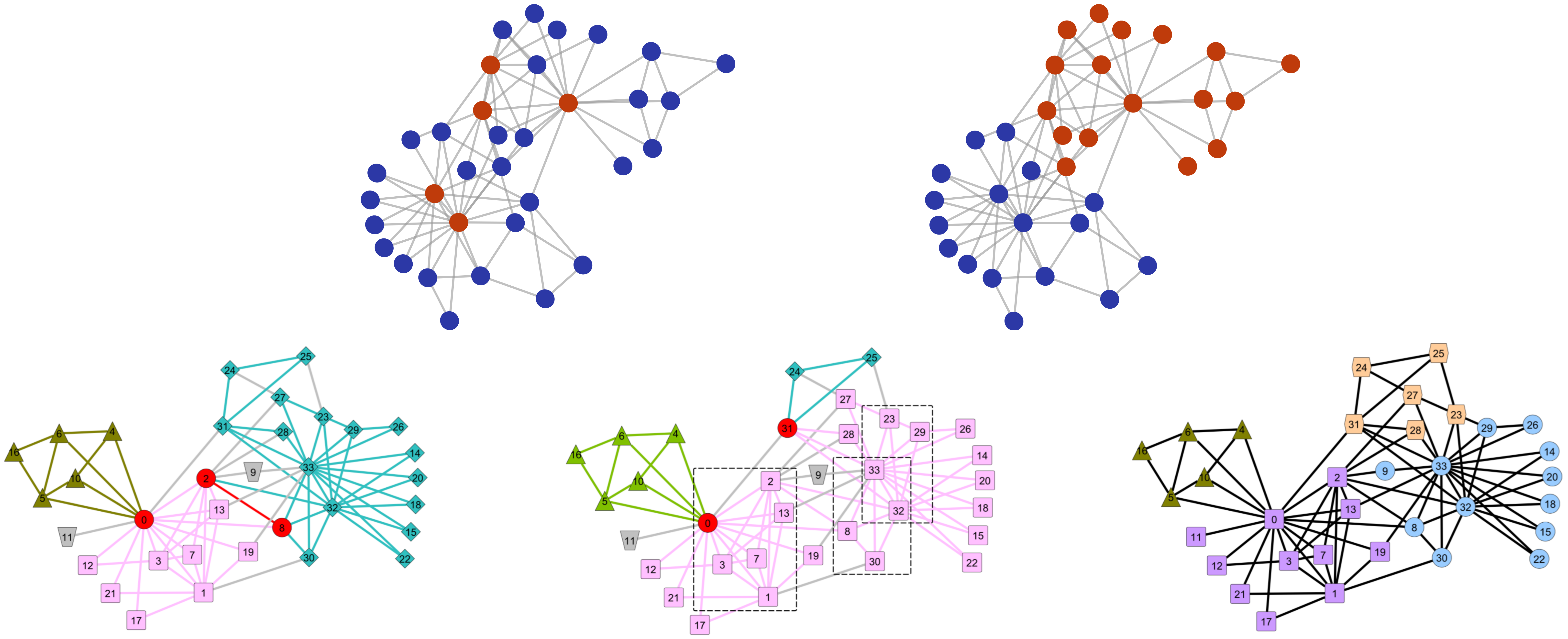
with degree correction





SBM likelihood surfaces

. . . many good partitions

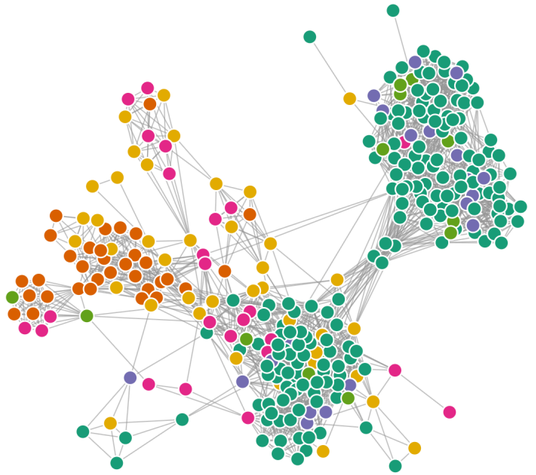


Evans, T. S. Clique graphs and overlapping communities. J. Stat. Mech. 2010, P12037–22 (2010).

Metadata are not ground truth for community detection.

No interpretability of negative results.

(i) M unrelated to network structure
(ii) C and M capture different aspects of network structure
(iii) the network has no structure
(iv) the algorithm does not perform well

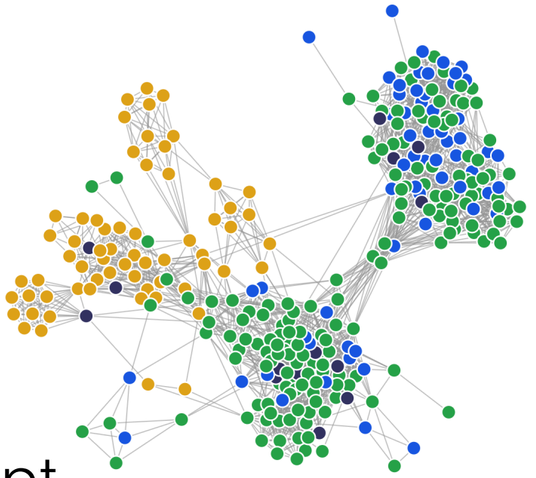


Multiple sets of metadata exist.

Which set is ground truth?

We see what we look for.

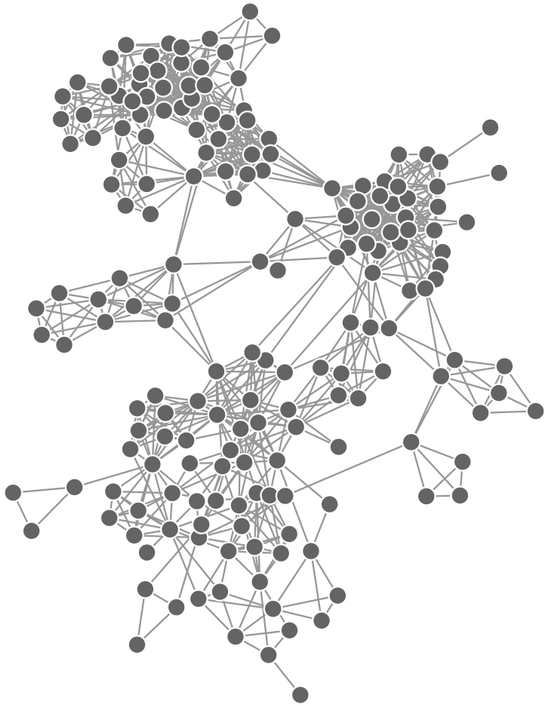
Confirmation bias. Publication bias.



“Community” is model dependent.

When faced with the same question, should all methods attempt to provide the same answer?

The ground truth detection problem

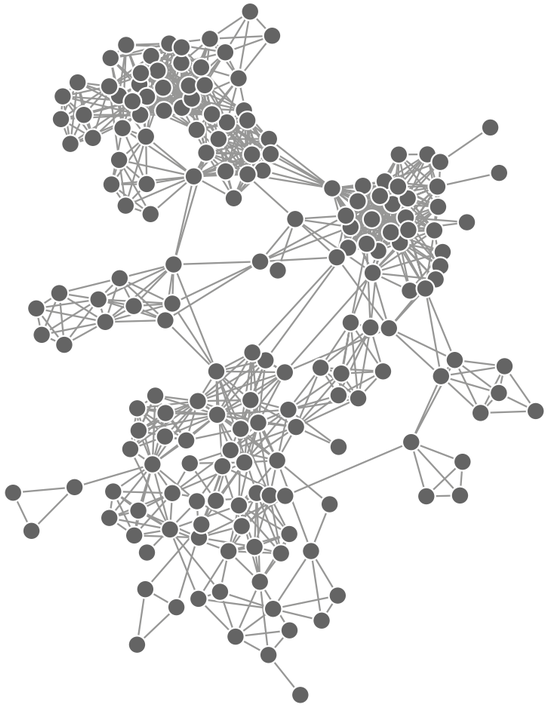


graph G

community detection →

partition T × data-generating process g

The ground truth implanting problem



graph G

graph creation

partition T $\times$ data-generating process g

An ill-posed inverse problem

graph G

graph creation

partition T_1		data-generating process g_1
partition T_2		data-generating process g_2
partition T_3		data-generating process g_3
partition T_4	$\times$	data-generating process g_4
...		...
partition T_n		data-generating process g_n

For any graph there exist a (Bell) number of possible “ground truth” partitions,
and an infinite number of capable generative models.

There *is* no ground truth for community detection.

No free lunch for community detection

NFL theorems (machine learning) state that there cannot exist a supervised classifier that is *a priori* better than any other, averaged over all possible problems.

Theorem (NFL for CD, paraphrased):

For the community detection problem, with accuracy measured by **adjusted mutual information**, the uniform average of the accuracy of any method *f* over all possible community detection problems is a constant which is independent of *f*.

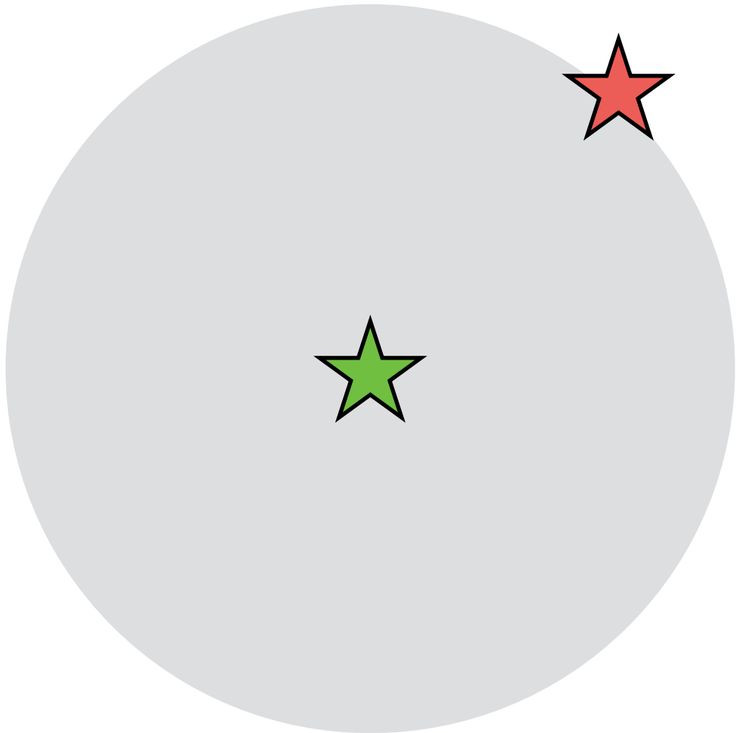
Suggested: Wolpert, D. H. The lack of a priori distinctions between learning algorithms. Neural Computation 8, 1341–1390 (1996).

Hard part of proof: show AMI is homogeneous

$$L(u) = \sum_{v \in \Omega} \mathbf{AMI}(u, v)$$

(all possible partitions)

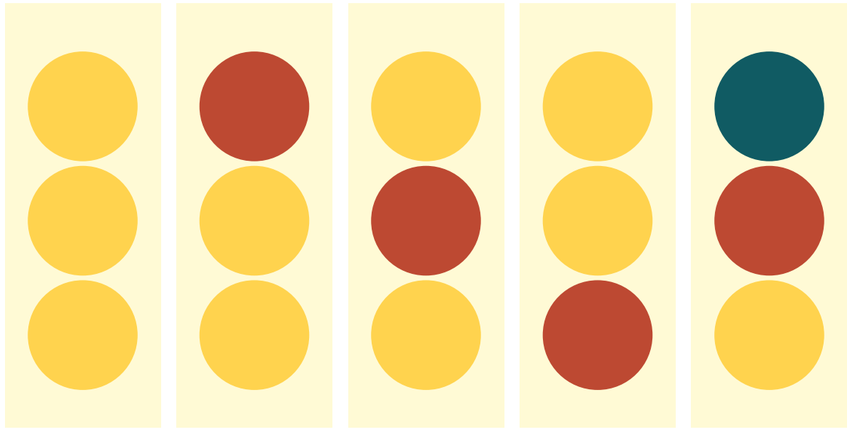
i.e., show that this quantity is independent of *u.*



Why do we have to show this? Geometry!

$$\mathbf{AMI}(u, v) = \frac{I(u, v) - E[I(u, v)]}{\sqrt{H(u)H(v)} - E[I(u, v)]}$$

Ex: all partitions of 3 objects



Partition 1	Partition 2				
	1	**2**	**3**	**4**	**5**
1	1	0	0	0	0
2	0	1	0.27	0.27	0.76
3	0	0.27	1	0.27	0.76
4	0	0.27	0.27	1	0.76
5	0	0.76	0.76	0.76	1
$\mathbb{E}$[NMI]	0.20	0.46	0.46	0.46	0.66

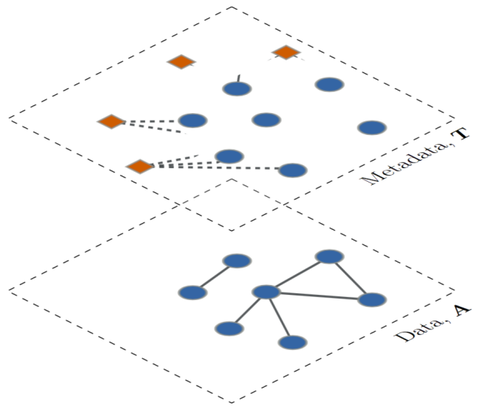
Partition 1	Partition 2				
	1	**2**	**3**	**4**	**5**
1	1	0	0	0	0
2	0	1	-0.5	-0.5	0
3	0	-0.5	1	-0.5	0
4	0	-0.5	-0.5	1	0
5	0	0	0	0	1
$\mathbb{E}$[AMI]	0.20	0	0	0	0.20

Metadata are still [just] data

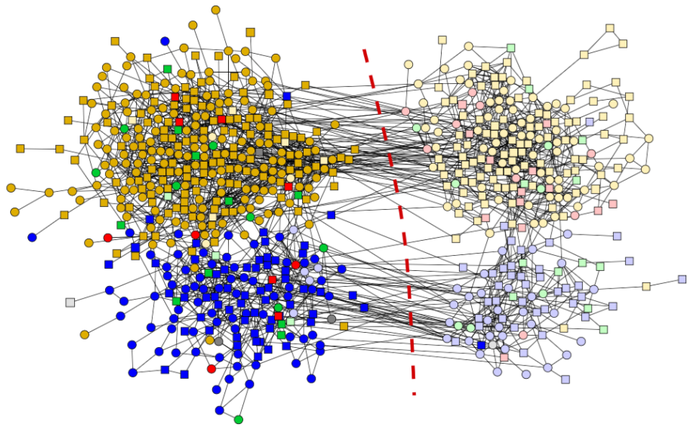
Metadata = types of nodes

Communities = large-scale patterns of how nodes interact

Metadata + Communities = how different types of nodes interact with each other.



Joint inference of
network+metadata model
Hric, Peixoto, Fortunato



Inference of network model
with model for metadata as “prior”
Newman, Clauset

Hric, D., Peixoto, T. P. & Fortunato, S. Network structure, metadata and the prediction of missing nodes. arXiv.org physics.soc-ph, (2016).

Newman, M. E. J. & Clauset, A. Structure and inference in annotated networks. Nature Communications 7, 1–11 (2016).

Two methods for exploring networks + metadata

Are the metadata related to network structure?
blockmodel entropy significance test

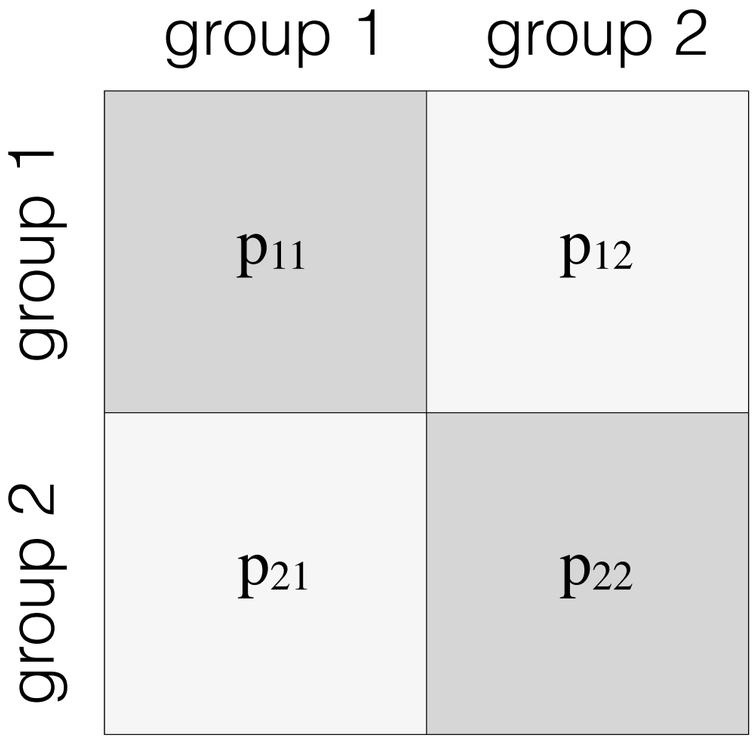
(i) M unrelated to network structure

Do metadata and detected communities capture different aspects of network structure?
neoSBM

(ii) C and M capture different aspects of network structure

Stochastic block model

- generative model for networks with community structure
- nodes are divided into groups
- the probability that an edge exists between any pair of nodes depends only on their group affiliations
- edges are conditionally independent



For community detection, SBM parameters can be fitted to data using maximum likelihood or fully Bayesian techniques.

Blockmodel entropy significance test

How well do the metadata explain the network?

randomly assigned metadata
→ model gives no explanation, high H

metadata correlated with communities
→ model gives good explanation, low H

1. Divide the network G into groups according to metadata labels M.

2. Fit the maximum likelihood parameters of an *a posteriori* SBM and compute the entropy $H(G,M)$ of the corresponding ensemble.

3. Compare the entropy of this SBM ensemble to distribution of entropies from SBMs partitioned using shuffled metadata $\underline{M}$.

p-value = $\Pr[H(G,\{\underline{M}\}) \leq H(G,M)]$

Multiple network layers; multiple metadata attributes

Network	Status	Gender	Office	Practice	Law School
Friendship	$< 10^{-6}$	0.034	$< 10^{-6}$	0.033	0.134
Cowork	$< 10^{-3}$	0.094	$< 10^{-6}$	$< 10^{-6}$	0.922
Advice	$< 10^{-6}$	0.010	$< 10^{-6}$	$< 10^{-6}$	0.205

model = SBM

Multiple sets of metadata **significantly explain** multiple networks.
[Should one particular set of metadata be ground truth?]

BESTest accommodates many models of group structure

Network	Model	
	SBM	DCSBM
Malaria 1	0.566	0.066
Malaria 2	0.064	0.126
Malaria 3	0.536	0.415
Malaria 4	0.588	0.570
Malaria 5	0.382	0.097
Malaria 6	0.275	0.817
Malaria 7	0.020	0.437
Malaria 8	0.464	0.143
Malaria 9	0.115	0.104

metadata = parasite origin

A negative result: parasite origin is irrelevant to genetic substring-sharing.

Malaria parasites *do not* have a strong strain structure, with implications for **diversifying selection** among parasites.

DBL, Clauset, A. & Buckee, C. O. A Network Approach to Analyzing Highly Recombinant Malaria Parasite Genes. PLoS Comp Bio 9, e1003268 (2013).

Two methods for exploring networks + metadata

Are the metadata related to network structure?
blockmodel entropy significance test

(i) M unrelated to network structure

Do metadata and detected communities capture different aspects of network structure?
neoSBM

(ii) C and M capture different aspects of network structure

neoSBM

Choose between the SBM partition and the metadata partition.

$$\underset{\substack{\text{neoSBM}\\ \text{log likelihood}}}{\mathcal{L}_{\text{neoSBM}}} = \underset{\substack{\text{SBM}\\ \text{log likelihood}}}{\mathcal{L}_{\text{SBM}}} + \underset{\text{cost}}{f(\theta)}$$

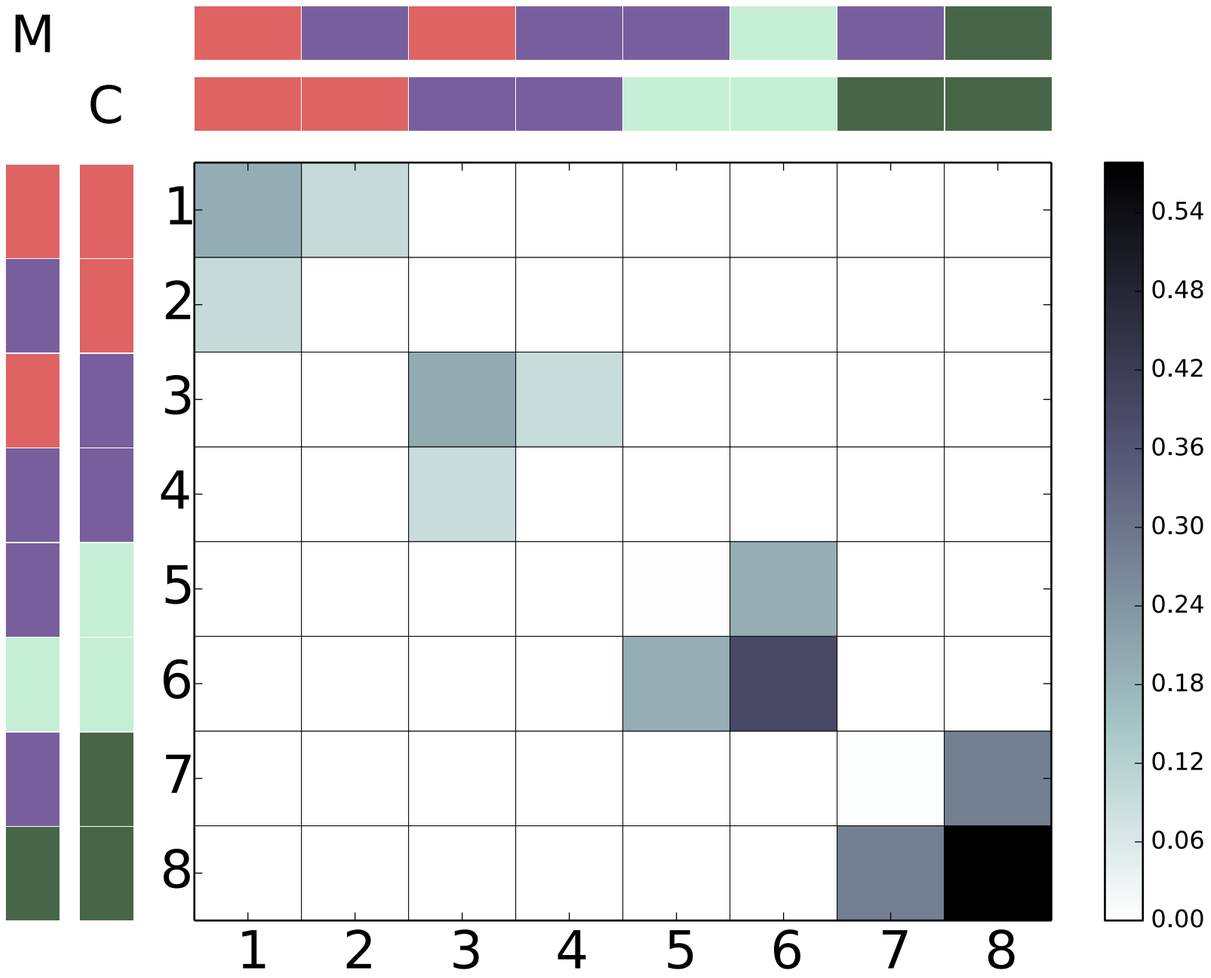
Log likelihood with parameterized prior:

θ is the parameter of a Bernoulli prior on whether the node is free to choose its own community or held fixed at its metadata label.

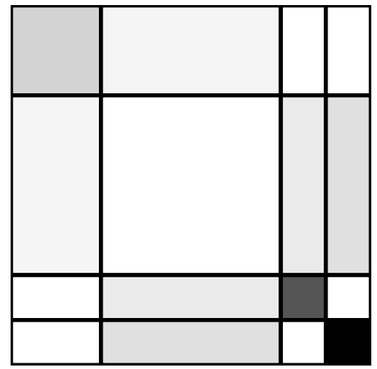
As θ increases, the cost of freeing a node decreases.

Varying θ in the unit interval **explores the space of partitions** between M and C.

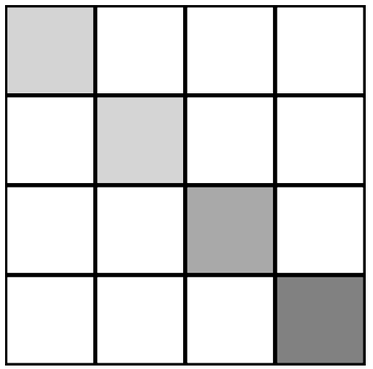
A clever test that Leto Peel devised



SBM with 8 groups and
two interesting 4-group partitions:

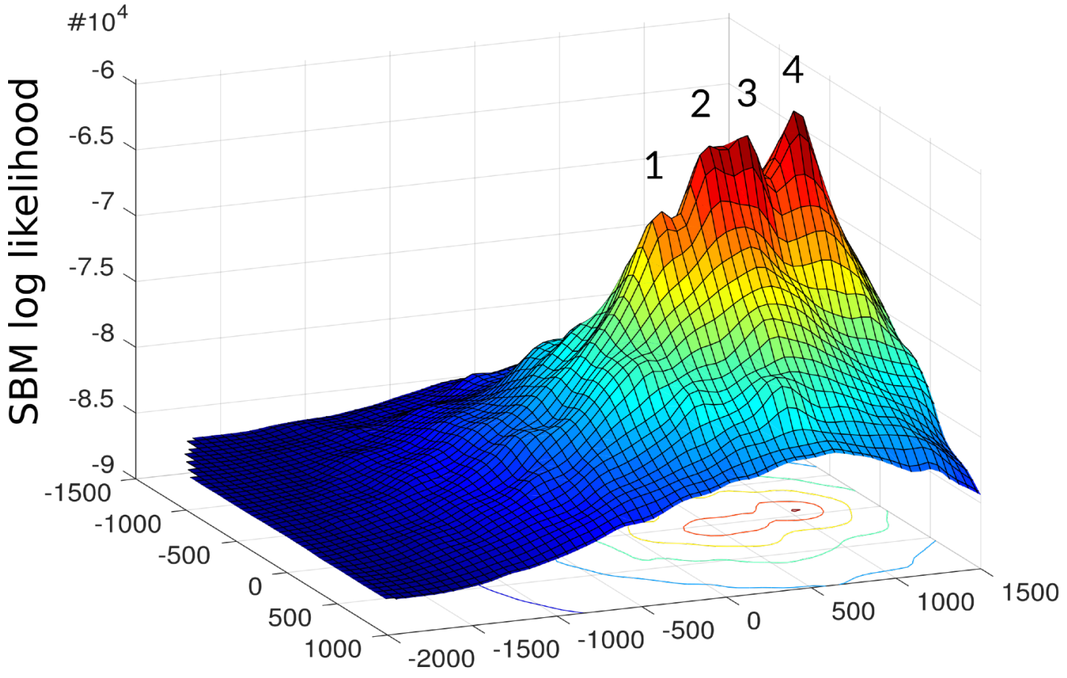


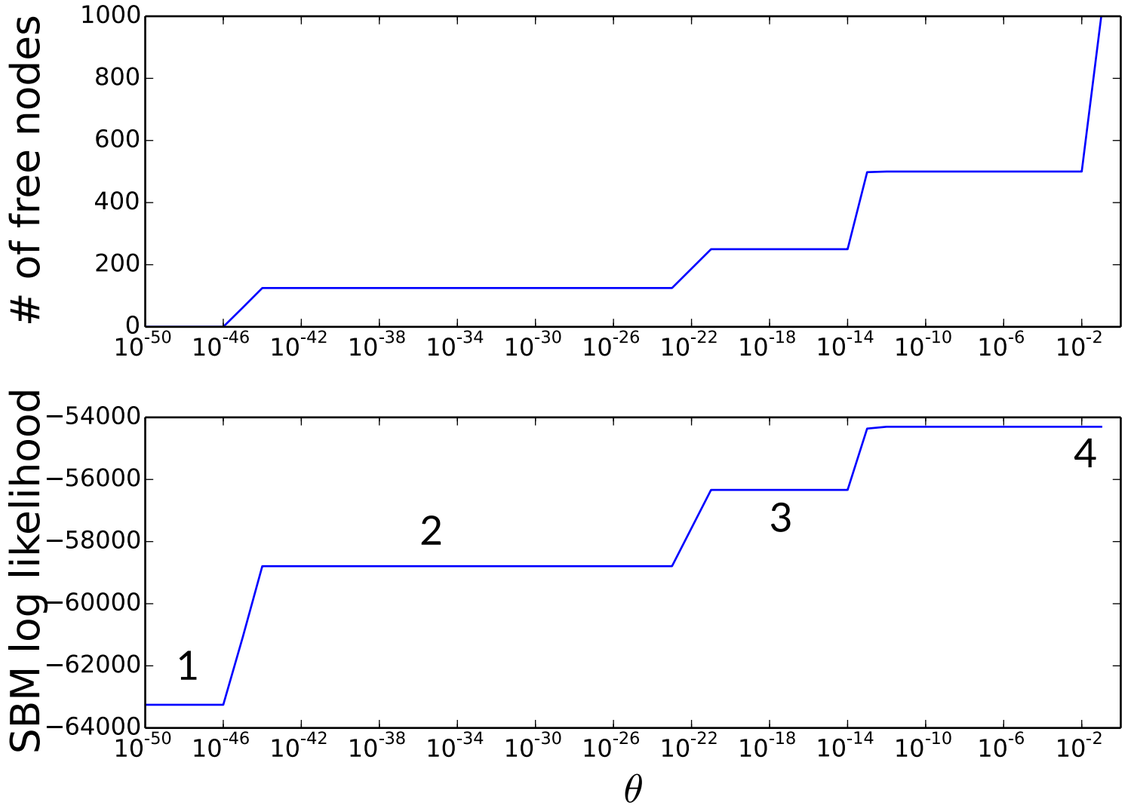
i. core-periphery

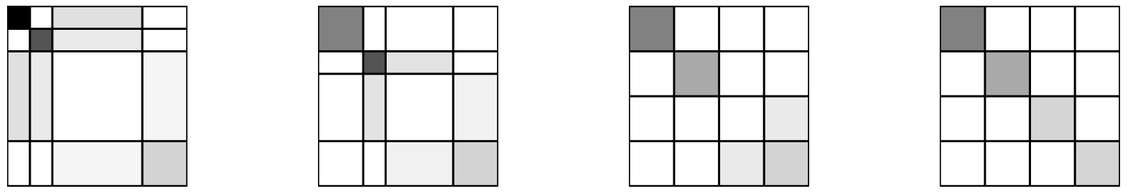


ii. assortative

The neoSBM identifies four interesting partitions



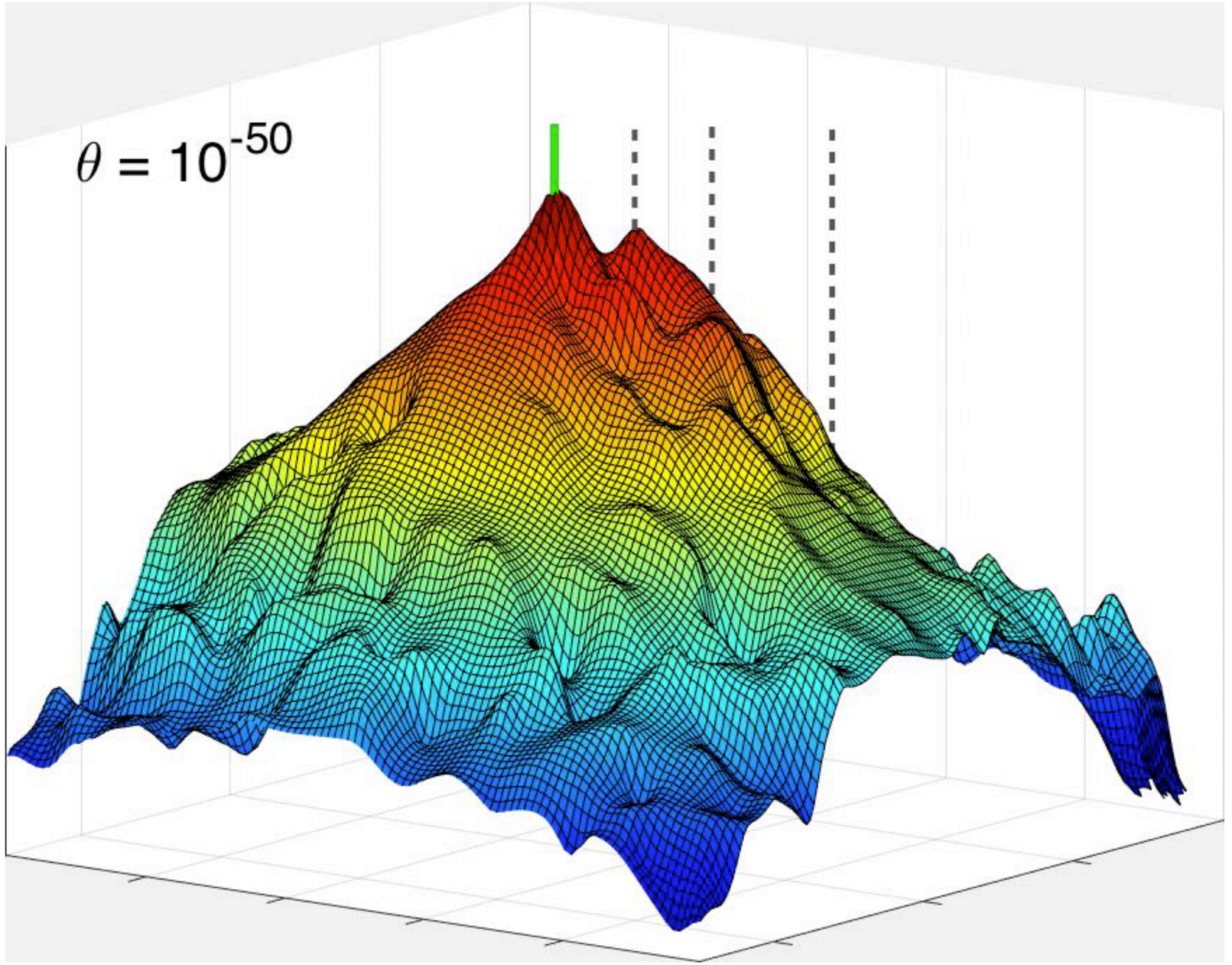




metadata

SBM

The prior parameter changes the posterior surface



Conclusions & Remarks

Metadata ≠ ground truth for community detection.

i. cannot directly interpret metadata-community mismatch. [+ result bias]
ii. many metadata attributes – which is ground truth? [confirmation bias]
iii. if we try to find metadata, we find what we are looking for [confirmation bias 2]

There are no ground truth communities in real networks.

i. exact recovery is ill-posed; there is no bijection between partitions and graphs.
ii. no free lunch for community detection

Metadata are data; Data are useful. Two exploratory and versatile tools:

i. Blockmodel Entropy Significance Test: are metadata related to network structure?
ii. neoSBM: do metadata and communities captures different structural patterns?

SCIENCE ADVANCES | RESEARCH ARTICLE

APPLIED MATHEMATICS

The ground truth about metadata and community detection in networks

Leto Peel,[1,2*†] Daniel B. Larremore,[3*†] Aaron Clauset[3,4,5†]

http://advances.sciencemag.org/content/advances/3/5/e1602548.full.pdf

https://github.com/piratepeel/**neoSBM**

https://github.com/dblarremore/**BESTest**