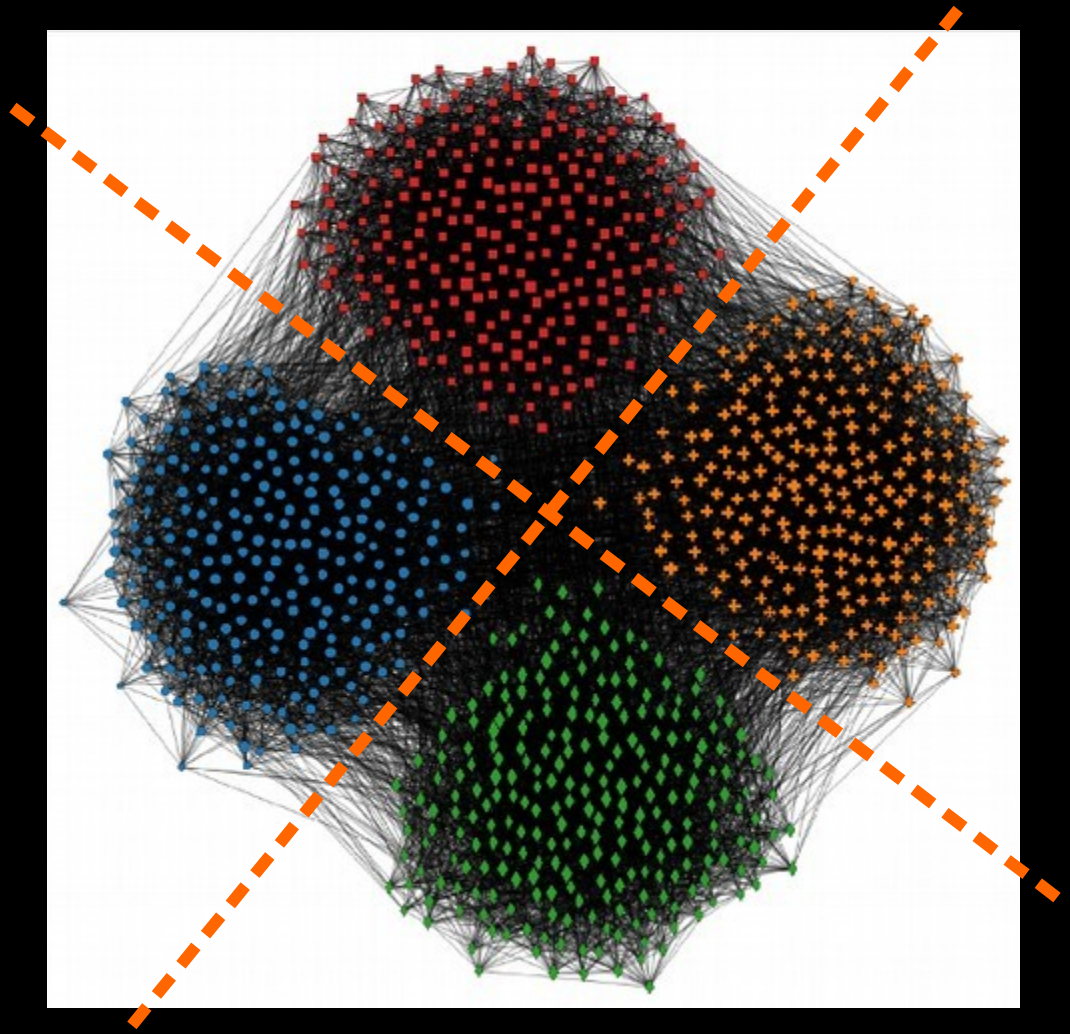
The ground truth about metadata and community detection in networks

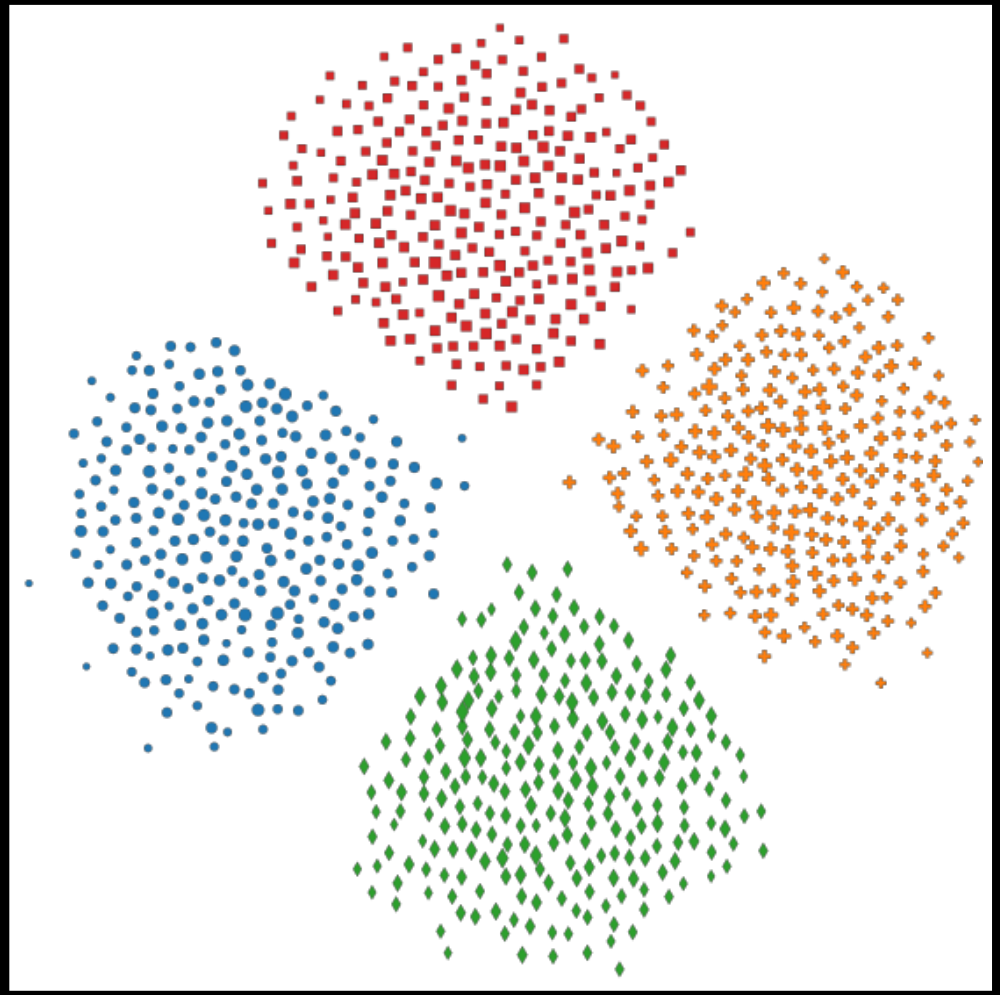
Leto Peel
Université catholique de Louvain

arXiv:1608.05878



Community detection:
Split nodes into groups based on their pattern of links

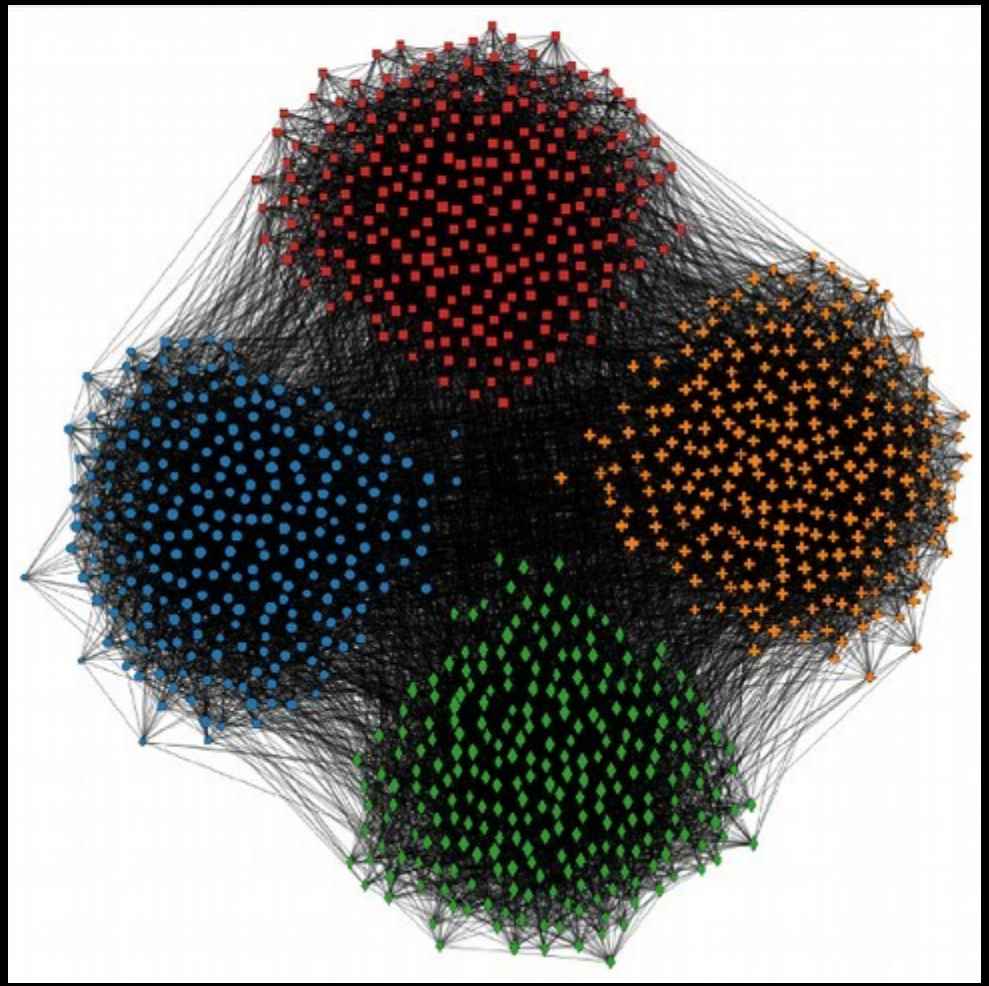
arXiv:1608.05878



Data generating process:

Generate nodes and assign to communities

arXiv:1608.05878



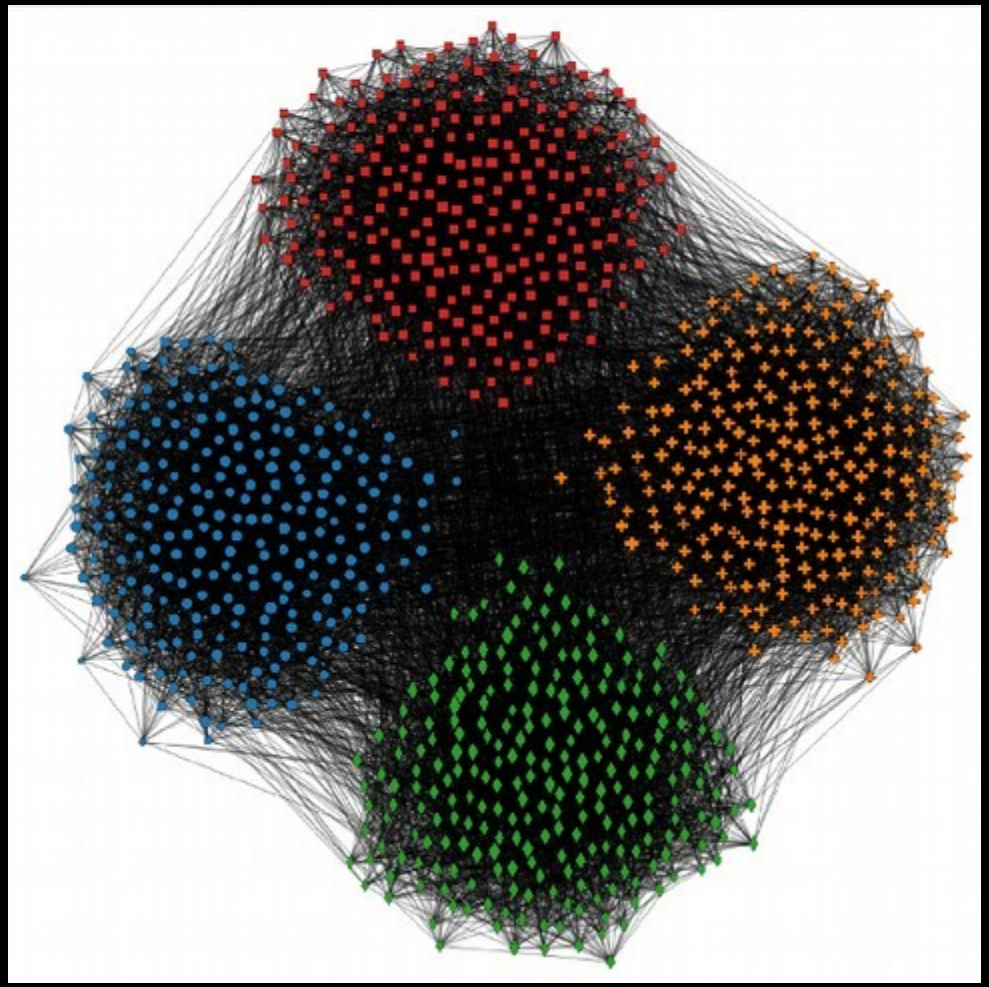
Data generating process:

Generate nodes and assign to communities, *T*

g(*T*)

Generate links in *G* dependent on community membership

arXiv:1608.05878



Community detection:

Infer *T*

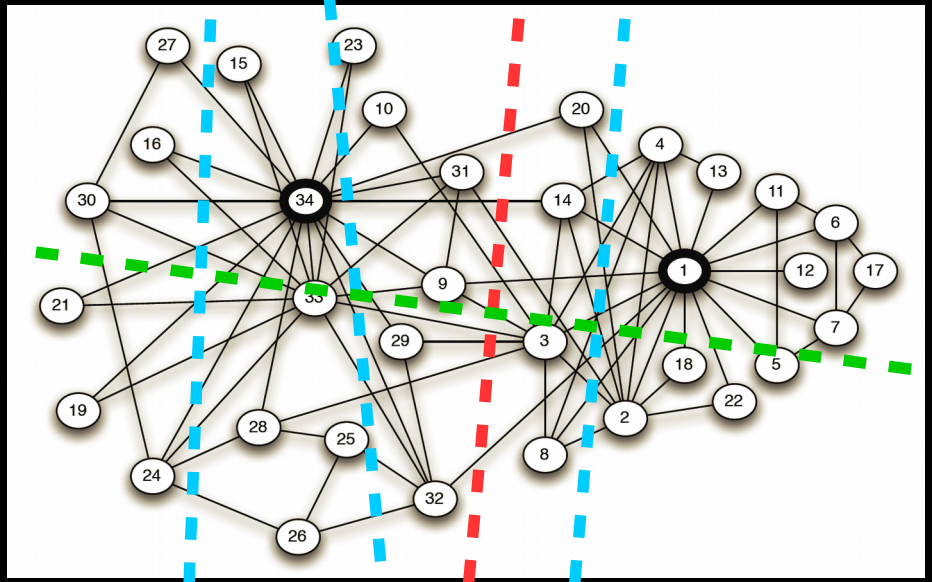
f(*G*)

Observe *G*

Assess performance on how well we recover *T*

Ground truth in real networks?

arXiv:1608.05878

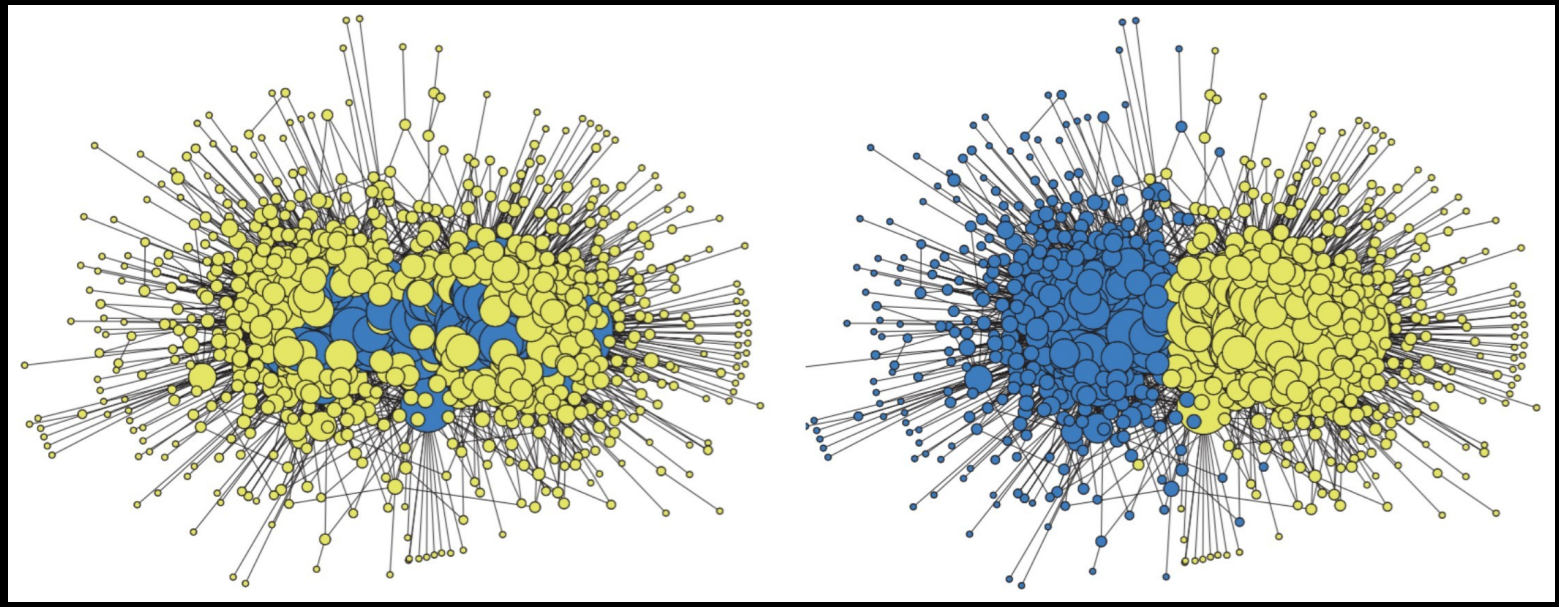


?

Networks can have *metadata* that describe the nodes

social networks	age, sex, ethnicity, race, etc.
food webs	feeding mode, species body mass, etc.
internet	data capacity, physical location, etc.
protein interactions	molecular weight, association with cancer, etc.

Recovering metadata implies sensible methods



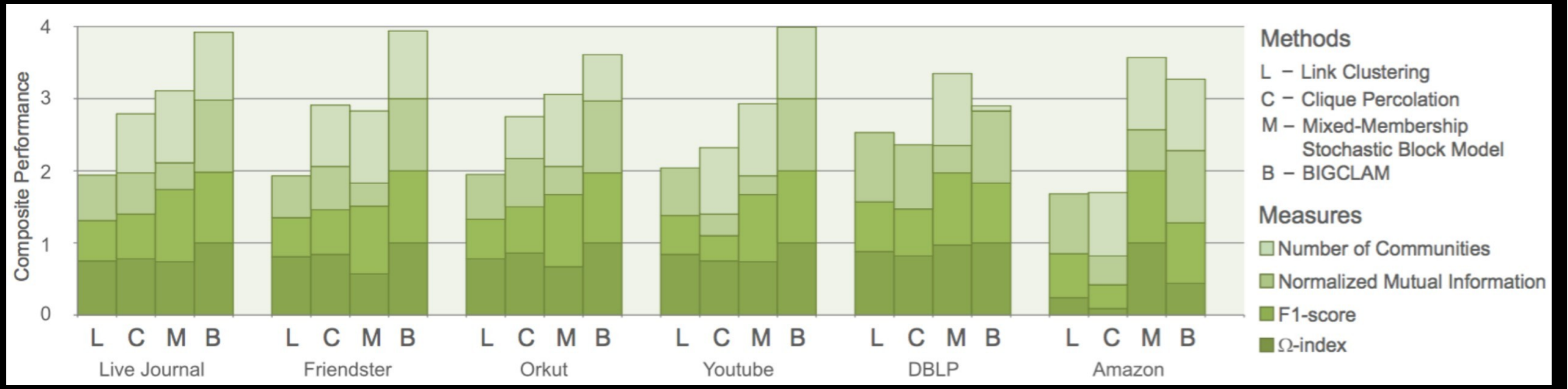
stochastic block model

stochastic block model
with degree correction

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 often treated as *ground truth*

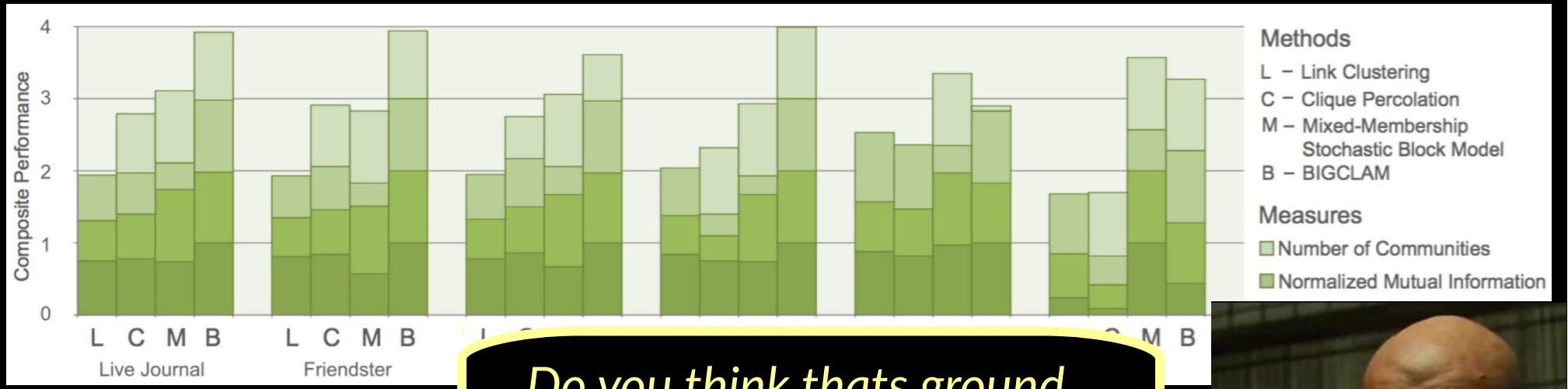
arXiv:1608.05878



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

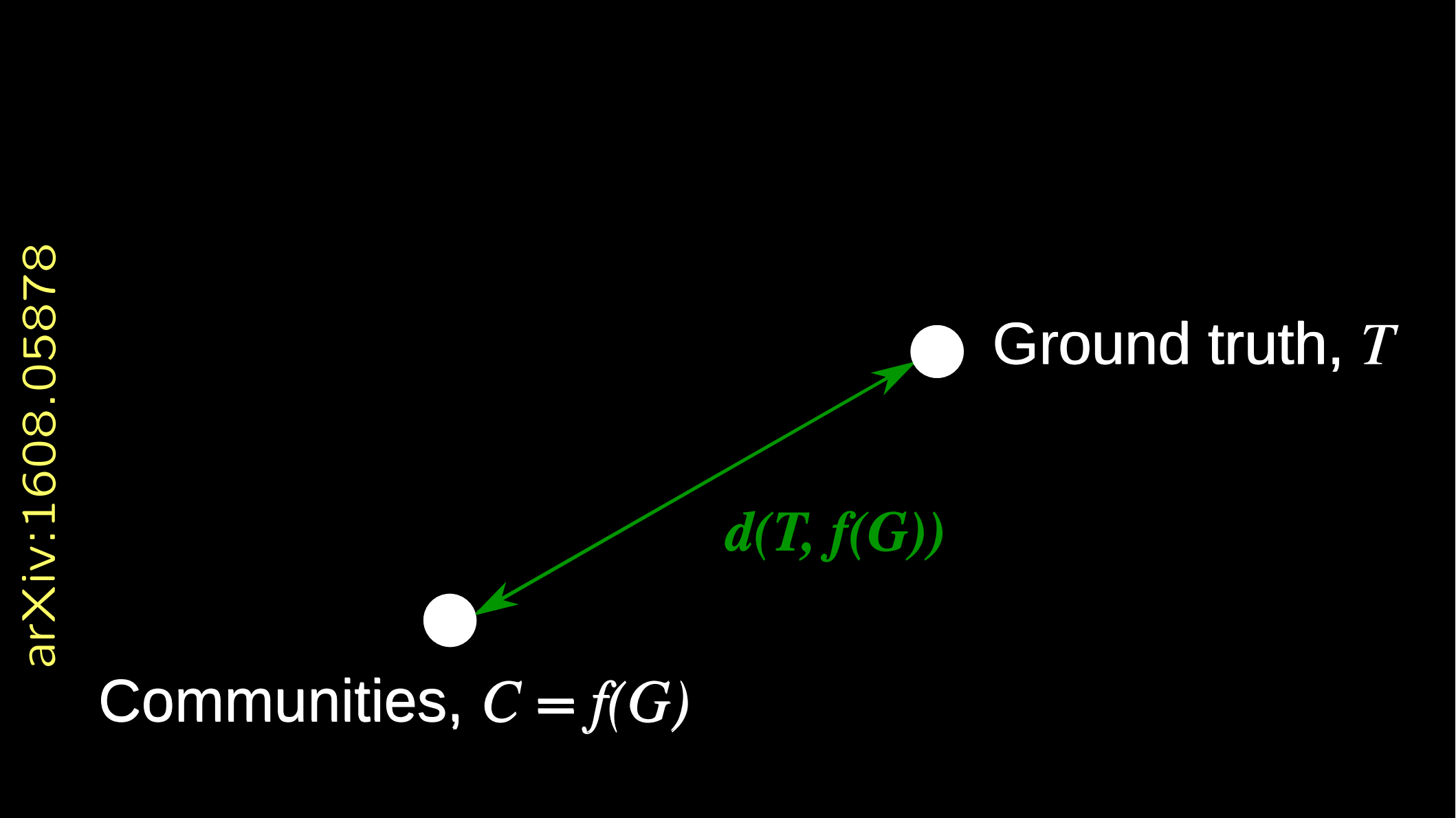
Metadata often treated as *ground truth*

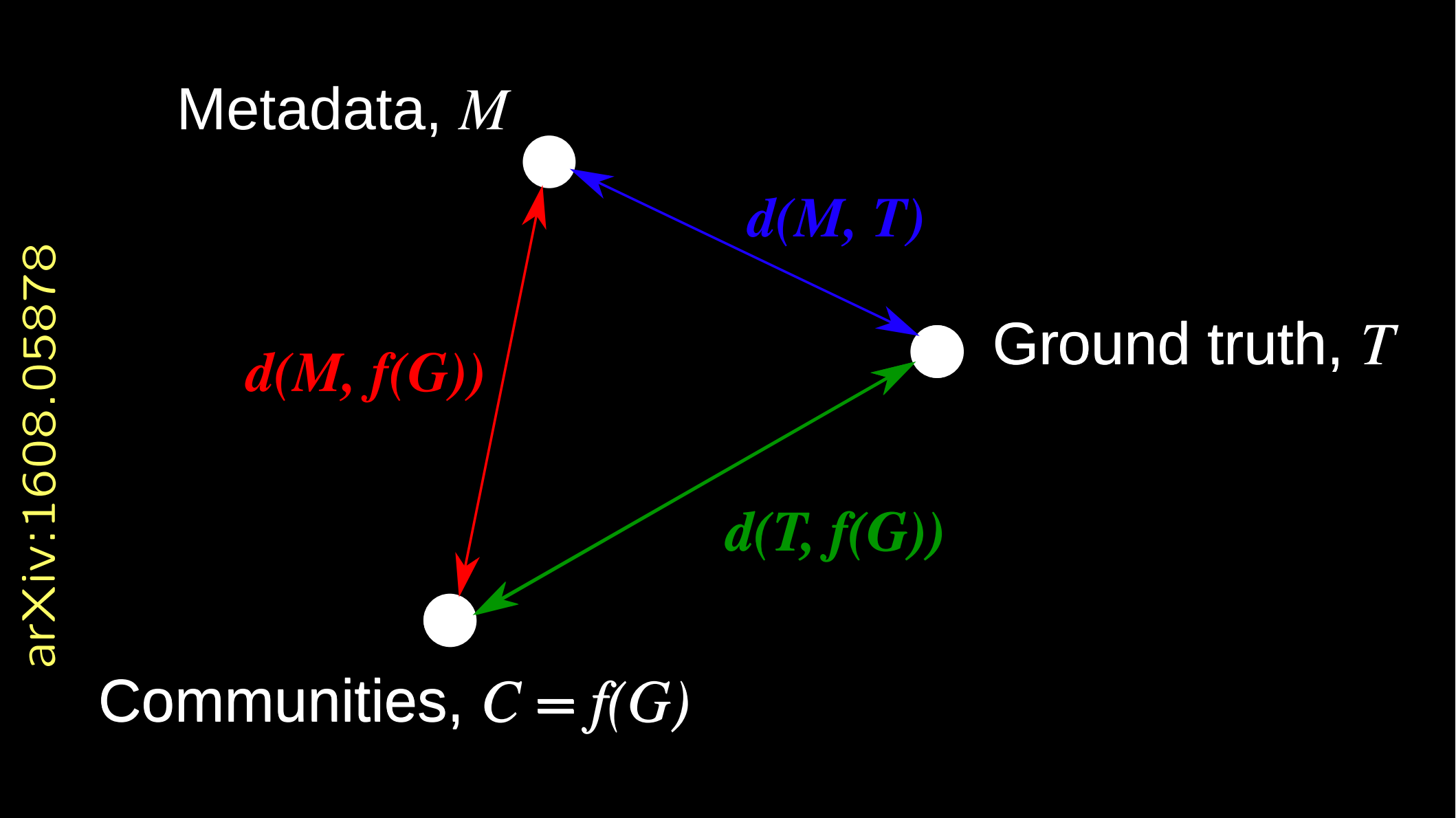
arXiv:1608.05878



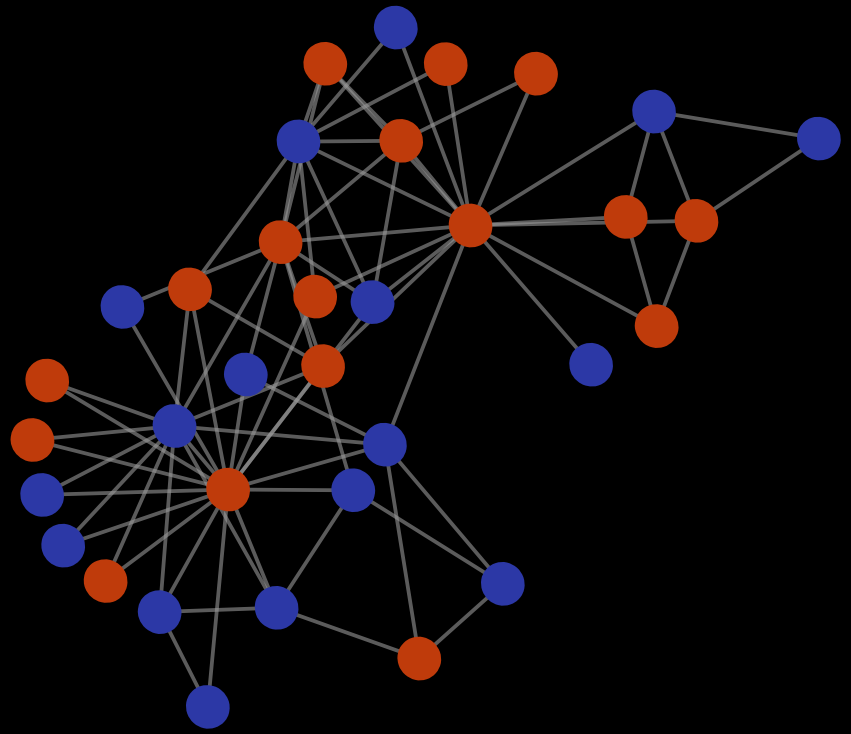
Do you think thats ground truth you're detecting?

Yang & Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization appr


arXiv:1608.05878
Ground truth, T
$d(T, f(G))$
Communities, $C = f(G)$

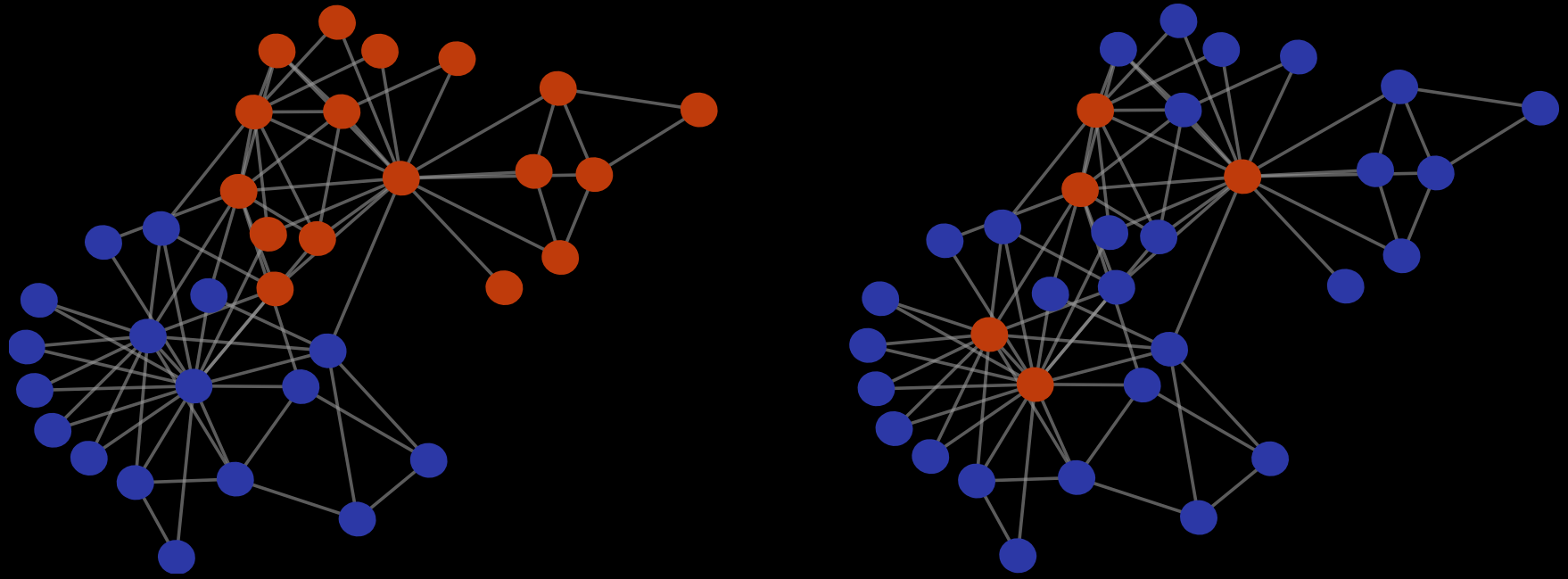


When communities ≠ metadata…



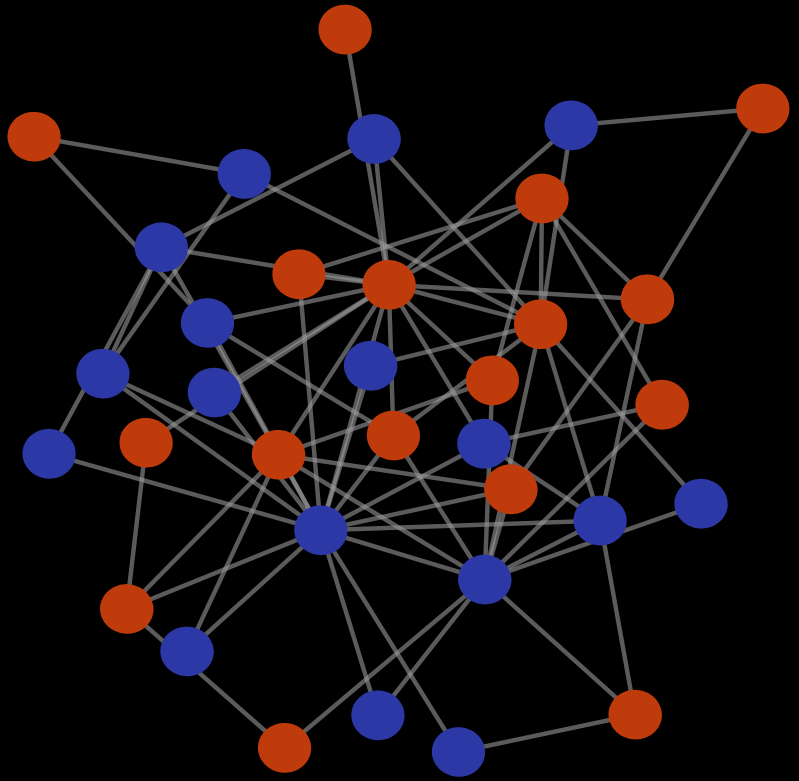
(i) the metadata do not relate to the network structure,

When communities ≠ metadata…



(ii) the detected communities and the metadata capture different aspects of the network's structure,

When communities ≠ metadata…



(iii) the network contains no structure (e.g., an E-R random graph)

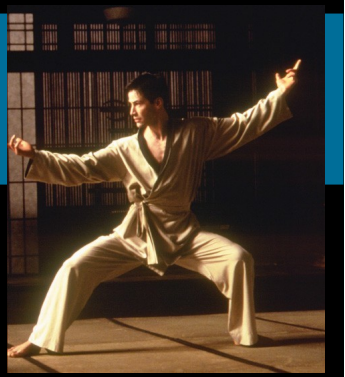
arXiv:1608.05878

When communities ≠ metadata…

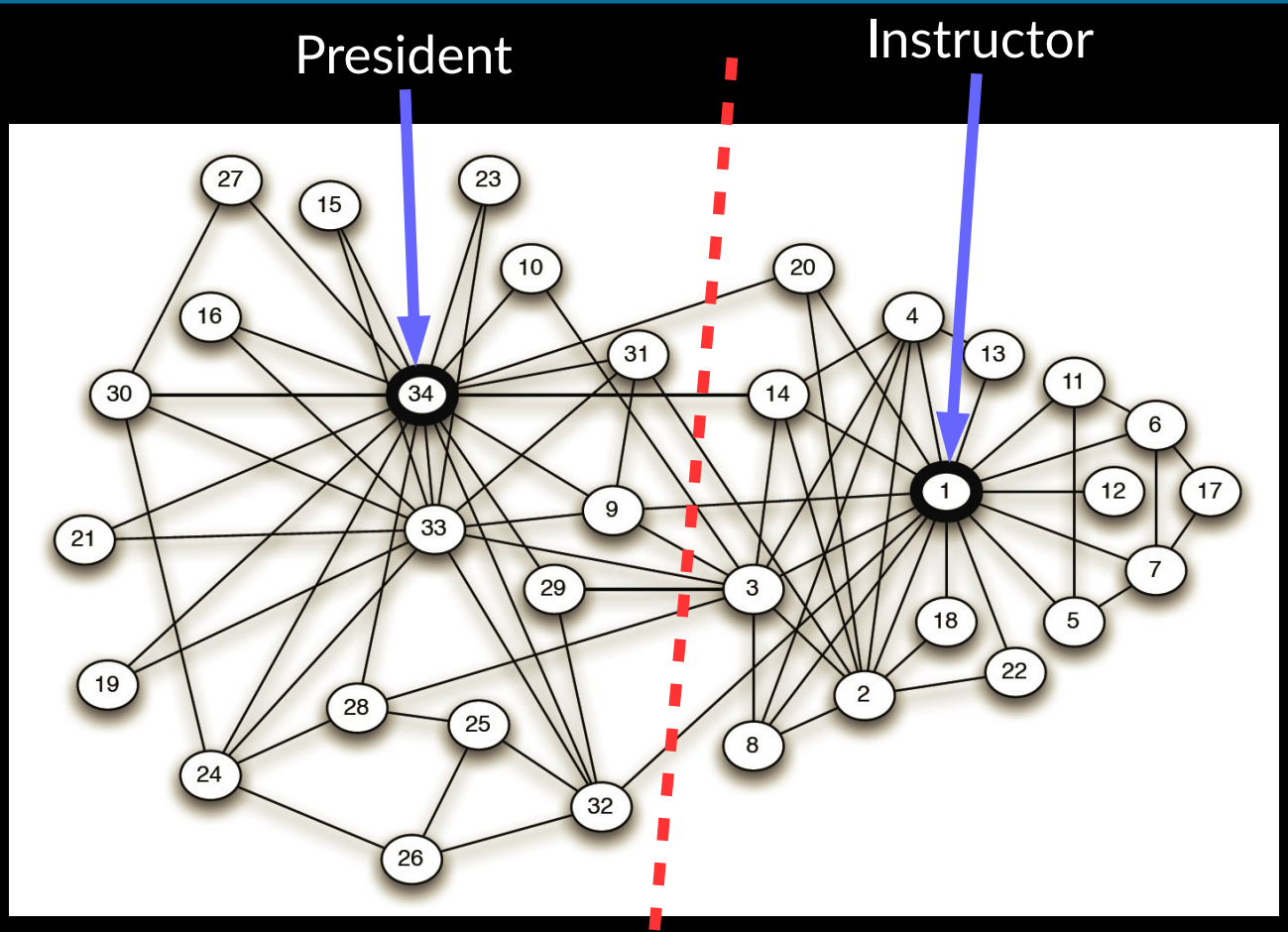
(iv) the community detection algorithm does not perform well.

Typically we assume this is the only possible cause

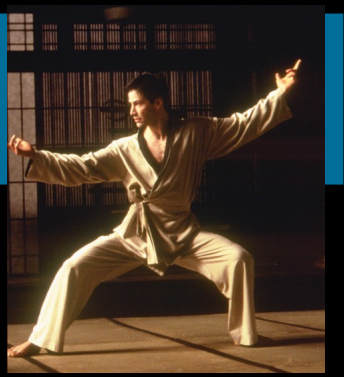
The Karate Club network



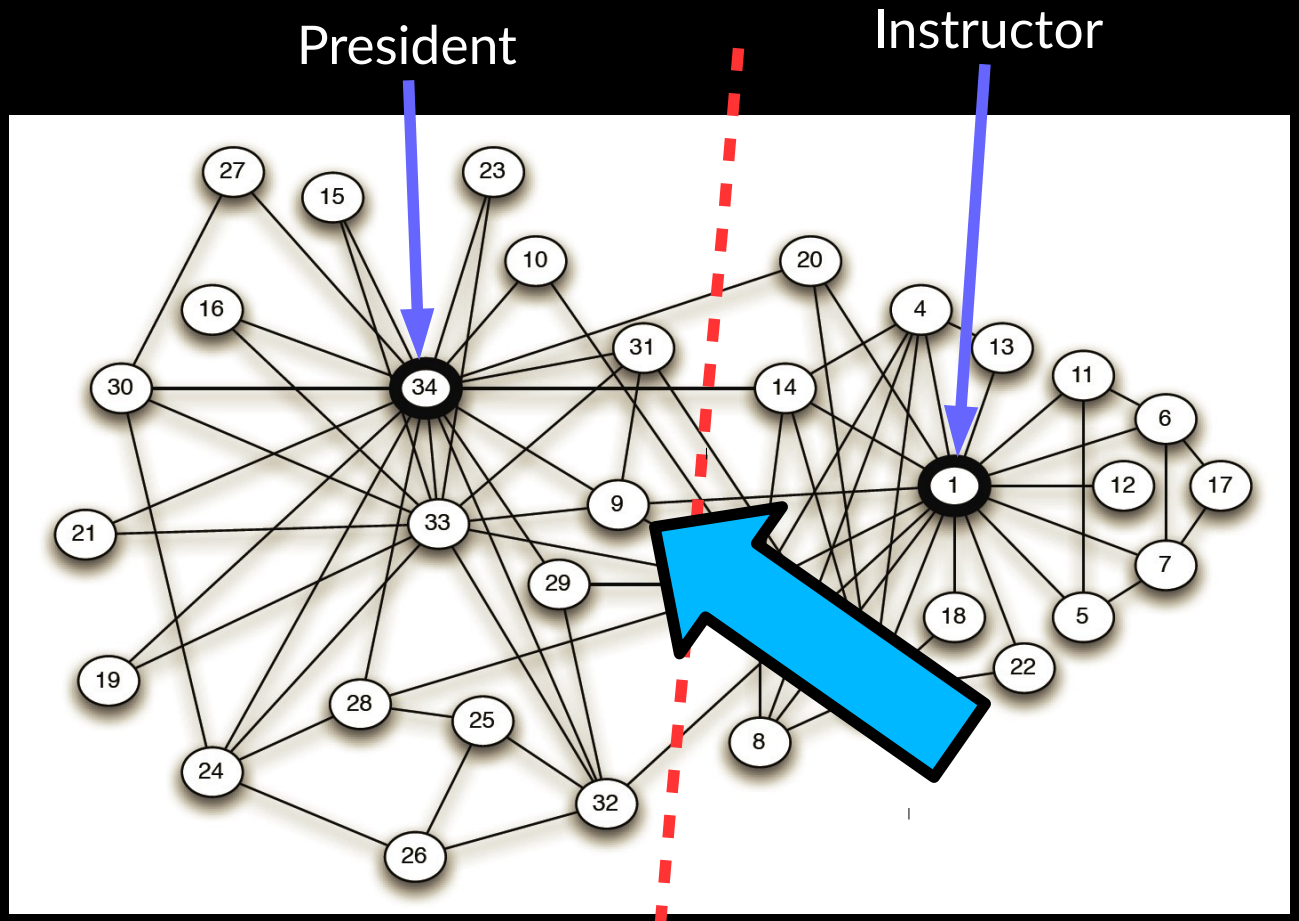
arXiv:1608.05878



The Karate Club network



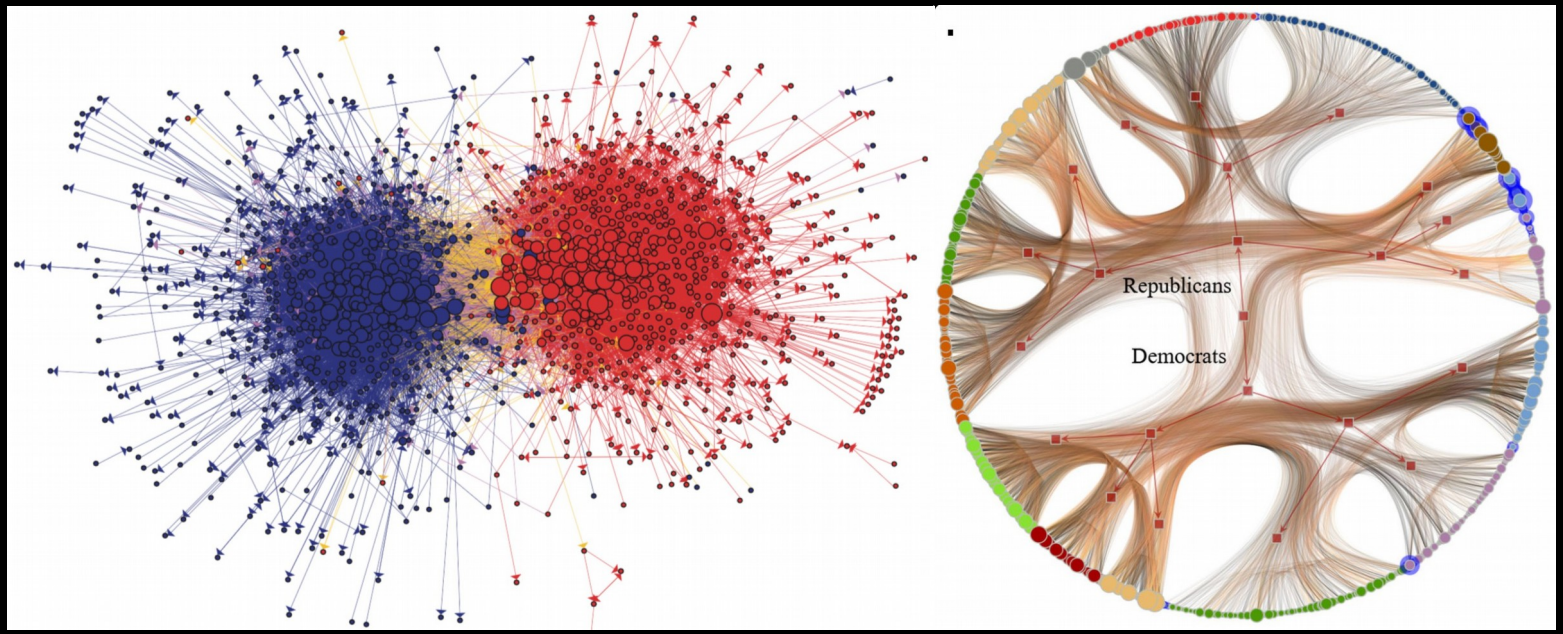
arXiv:1608.05878



‘This can be explained by noting that he was only three weeks away from a test for black belt (master status) when the split in the club occurred. Had he joined the officers’[President's] club he would have had to give up his rank and begin again in a new style of karate with a white (beginner’s) belt, since the officers had decided to change the style of karate practiced in their new club’

- Zachary 1977

You only see what you look for...

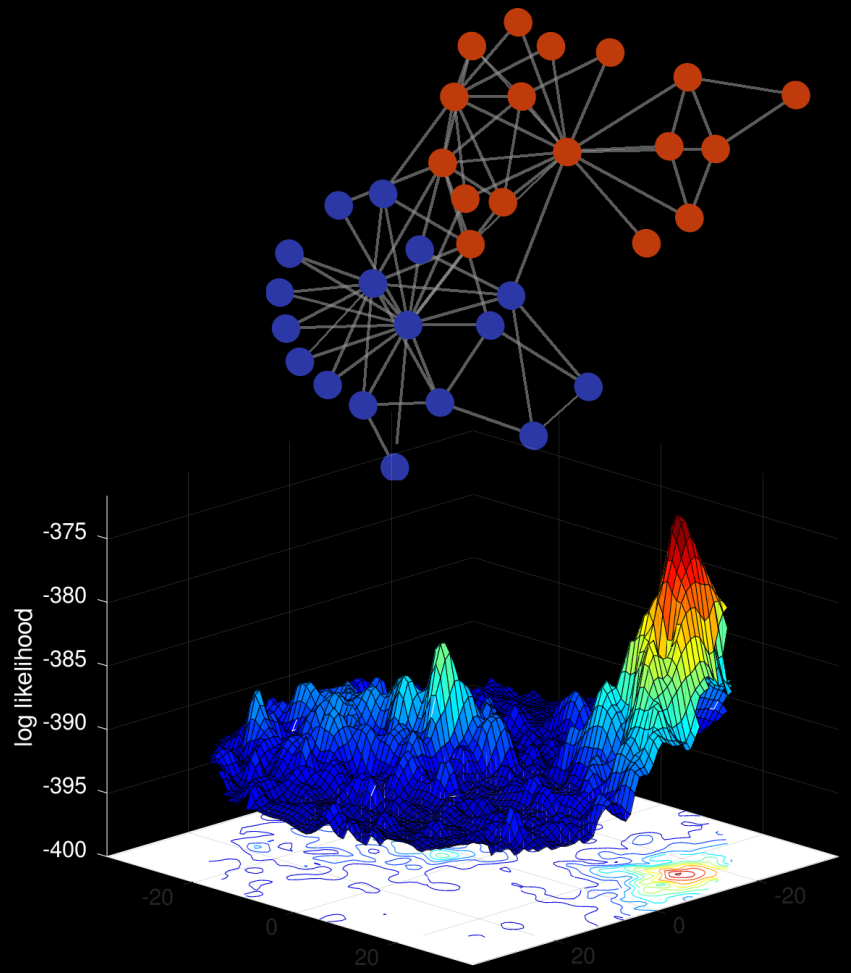


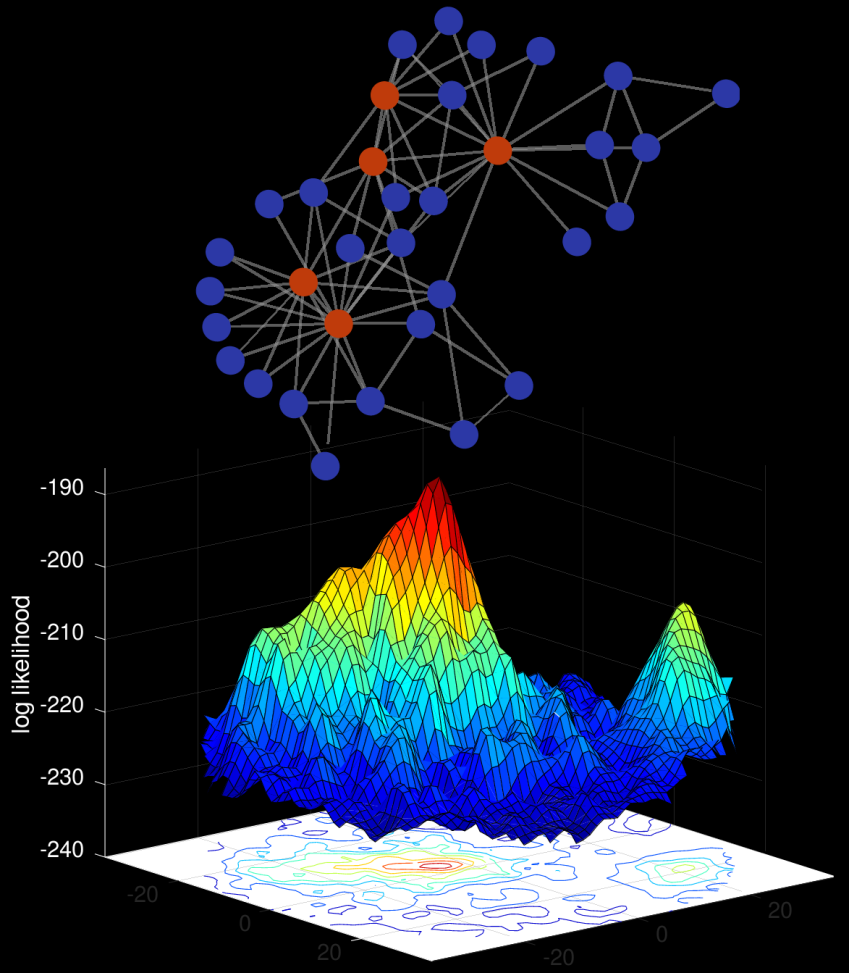
US politics is more than two opposing views

Adamic, Glance. The political blogosphere and the 2004 US election: divided they blog. 36–43 (2005).
Peixoto, T. P. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Phys. Rev. X 4, 011047 (2014).

Different generative processes = different community structures

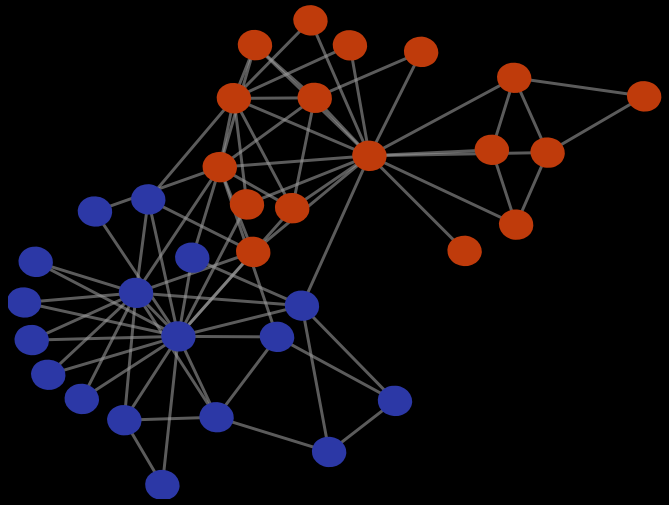
arXiv:1608.05878

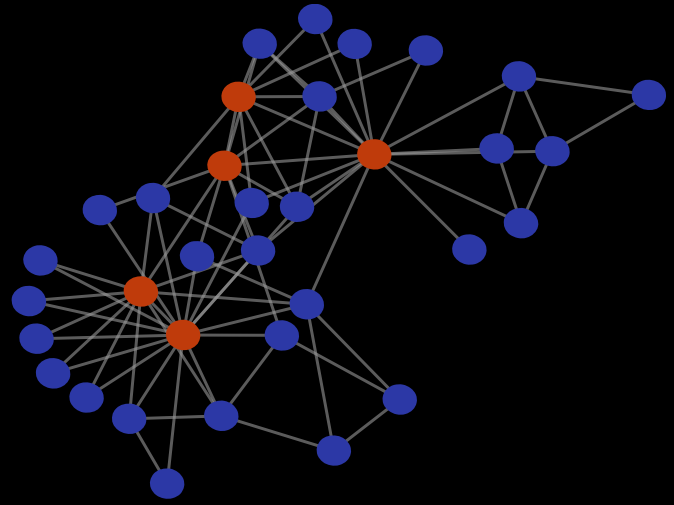


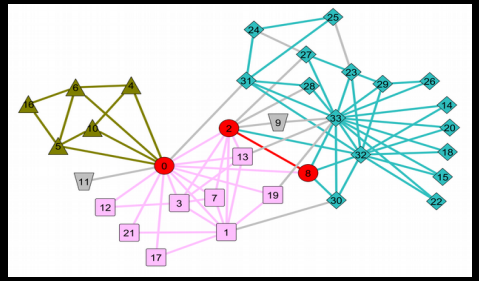


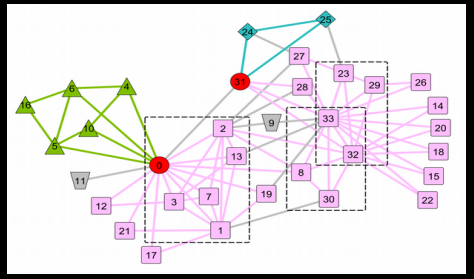
Many good partitions…

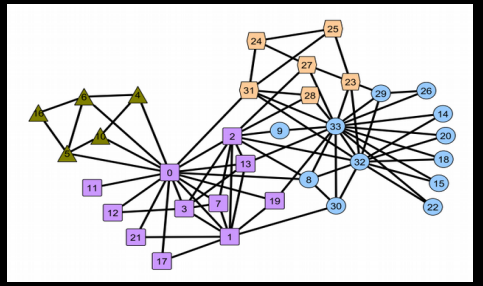
arXiv:1608.05878











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

Metadata are not ground truth for community detection

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

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?

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.

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?

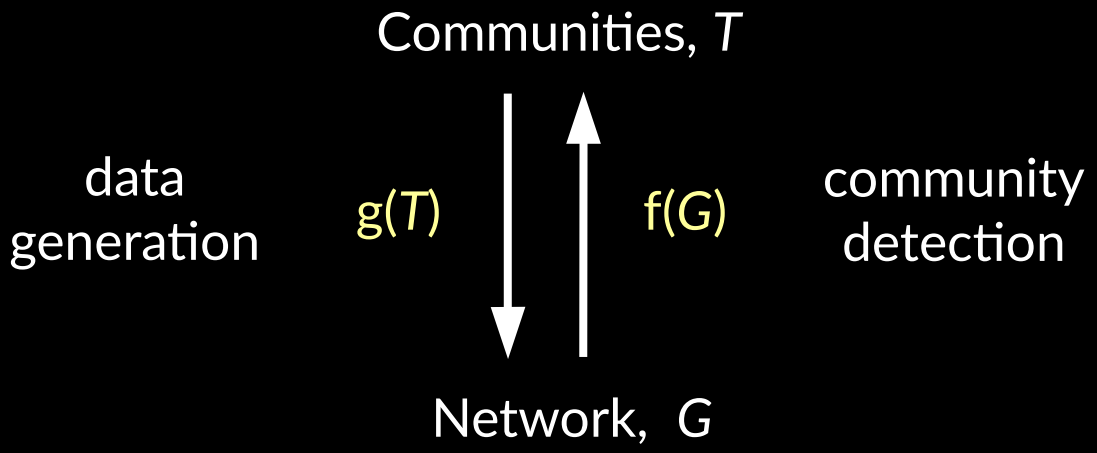
We see what we look for.

Confirmation bias. Publication bias.

“Community” is model dependent.

Do we expect all networks across all domains to have the same relationship with communities?

Community detection is an inverse problem



$$f^* = \arg\min_f d(\mathcal{T}, f(\mathcal{G}))$$

$$f^* = \arg\min_f d(\mathcal{T}, f(\mathcal{G}))$$

However, in real networks both *T* and *g* are unknown

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

{generative models, g} x {partitions, T} → {graph G}

many to one

see here for proof

The community detection problem is ill-posed (no unique solution)

A No Free Lunch Theorem for community detection?

arXiv:1608.05878

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

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

A No Free Lunch Theorem for community detection

arXiv:1608.05878

NFL Theorem for communtiy detection (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 .

On average, no community detection algorithm performs better than any other


DON'T TRY TO FIND THE GROUND TRUTH
INSTEAD... TRY TO REALIZE THERE IS NO GROUND TRUTH

So, what about metadata?

arXiv:1608.05878

Metadata = types of nodes

Communities = how nodes interact

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

we require new methods to understand the relationship between metadata and structure

Are the metadata related to the network structure?

Blockmodel Entropy Significance Test

Do metadata and detected communities capture different aspects network structure?

neoSBM

Are the metadata related to the network structure?

Blockmodel Entropy Significance Test

(i) the metadata do not relate to the network structure,

Do metadata and detected communities capture different aspects network structure?

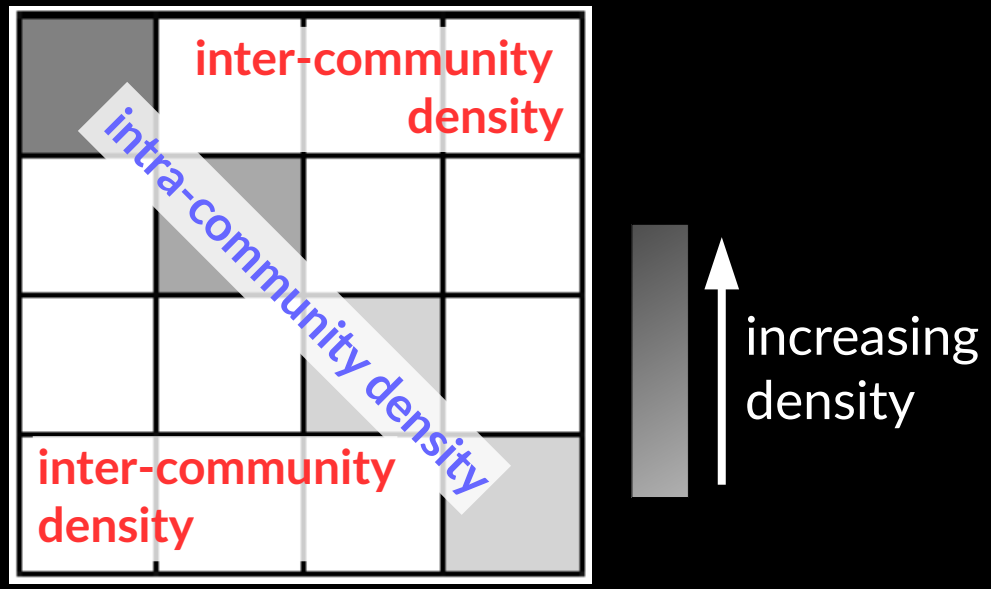
neoSBM

(ii) communities and metadata capture different aspects network structure,

The Stochastic Blockmodel

Edges are conditionally independent given community membership

$p_{ij} = p(e_{ij}|z_i,z_j,\omega) = \omega_{zi,zj}$



Blockmodel Entropy Significance Test

How well do the metadata explain the network?

arXiv:1608.05878

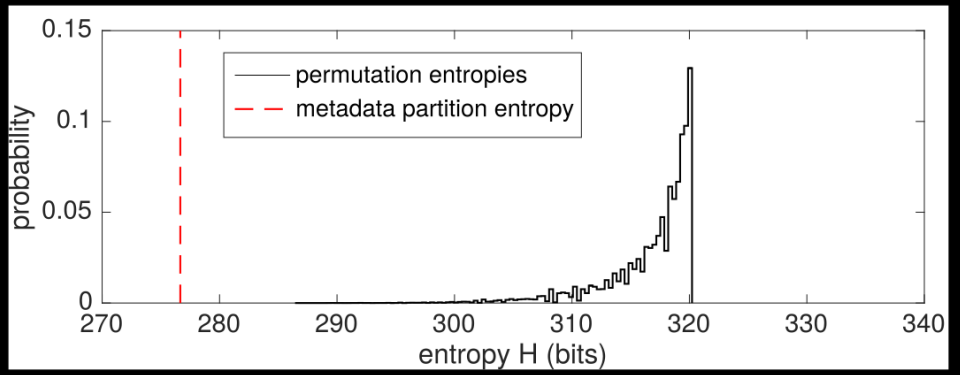
1. Divide the network *G* into groups according to metadata labels *M*.

2. Fit the parameters of an SBM and compute the entropy $H(G,M)$

3. Compare this entropy to a distribution of entropies of networks partitioned using permutations of the metadata labels.

metadata is randomly assigned
→ model gives no explanation, high H

metadata correlates with structure
→ model gives good explanation, low H



Multiple networks; multiple metadata attributes

arXiv:1608.05878

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

Multiple sets of metadata provide a significant explaination for multiple networks.

Are the metadata related to the network structure?

Blockmodel Entropy Significance Test

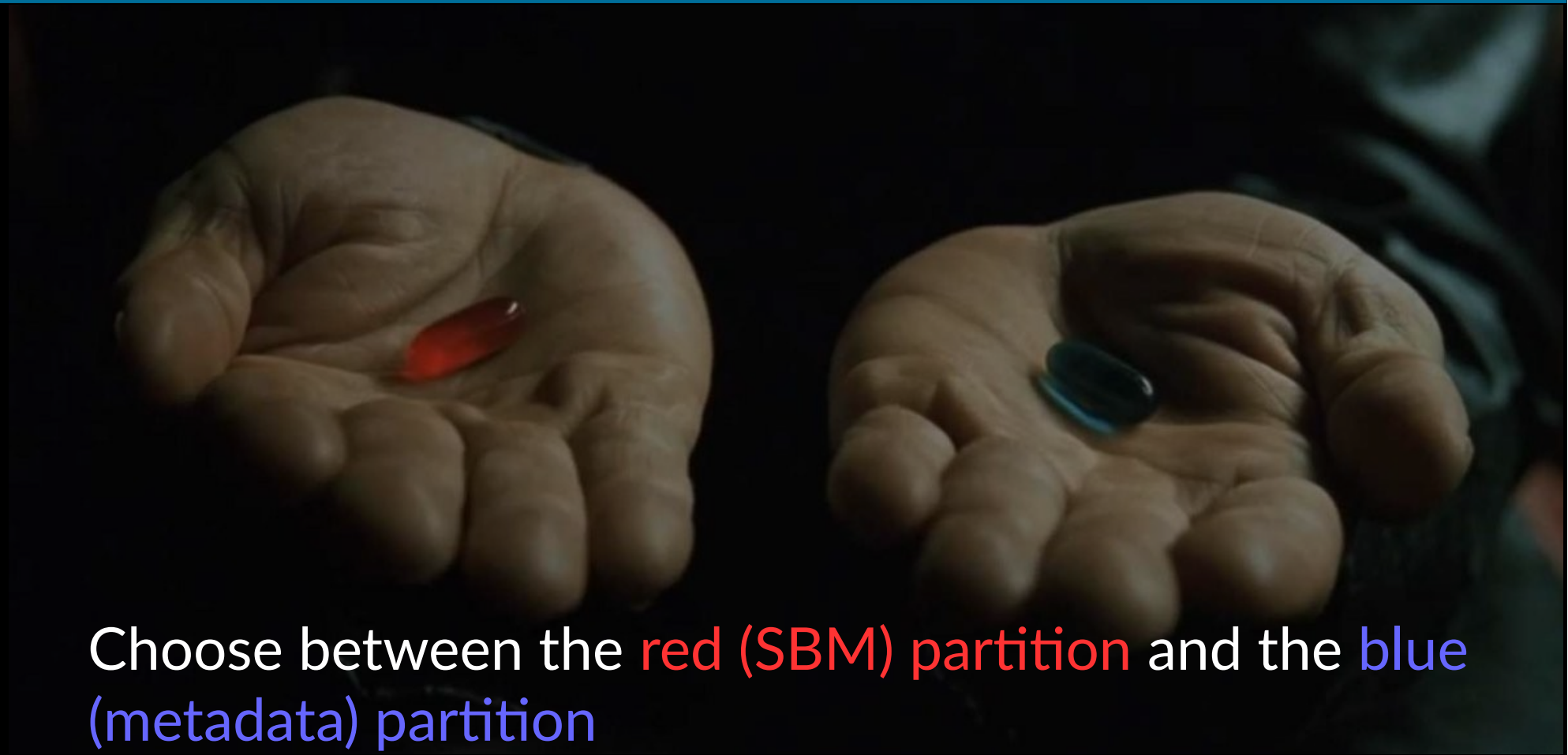
(i) the metadata do not relate to the network structure,

Do metadata and detected communities capture different aspects network structure?

neoSBM

(ii) communities and metadata capture different aspects network structure,

Do metadata and detected communities capture different aspects of the network?



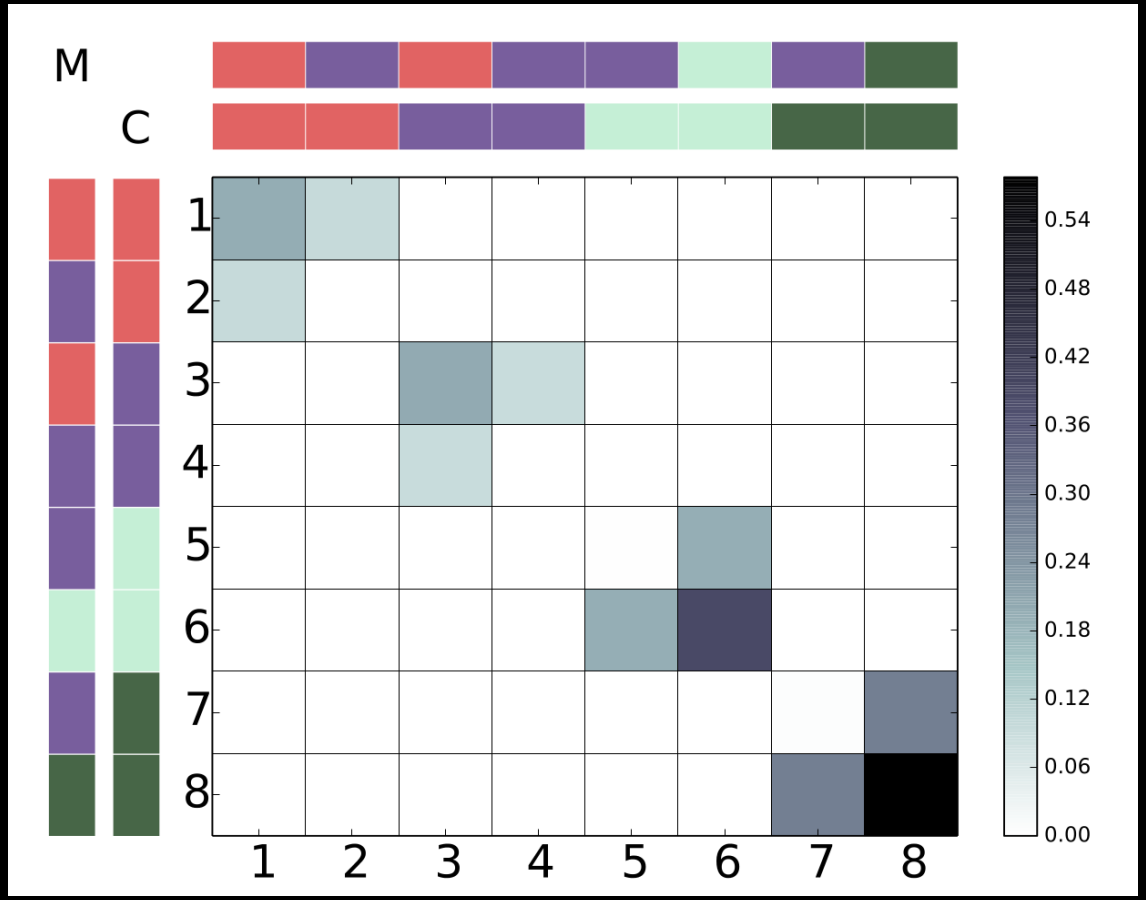
Choose between the red (SBM) partition and the blue (metadata) partition

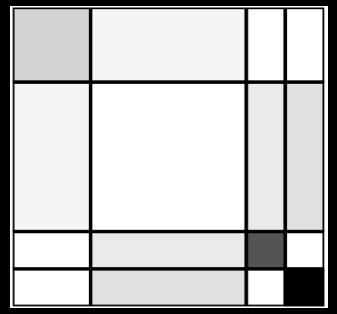
NEOSBM

arXiv:1608.05878

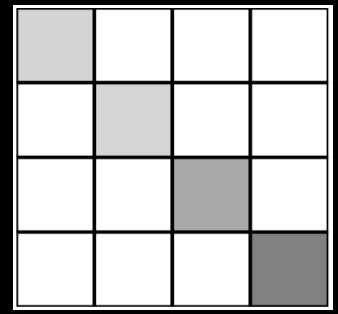
$$\underbrace{\mathcal{L}_{\text{neoSBM}}}_{\substack{\text{neoSBM}\\ \text{log likelihood}}} = \underbrace{\mathcal{L}_{\text{SBM}}}_{\substack{\text{SBM}\\ \text{log likelihood}}} + \underbrace{f(\theta)}_{\text{cost}}$$

Network with multiple 4-group optima





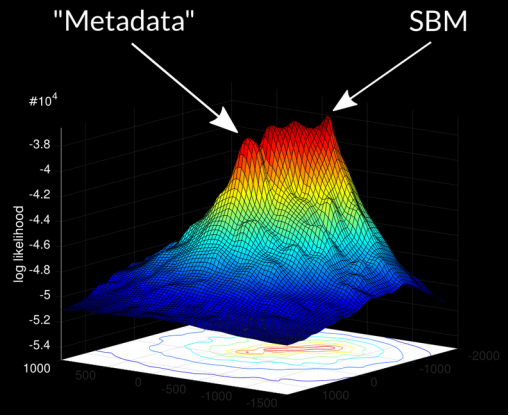
core-periphery
("metadata", M)

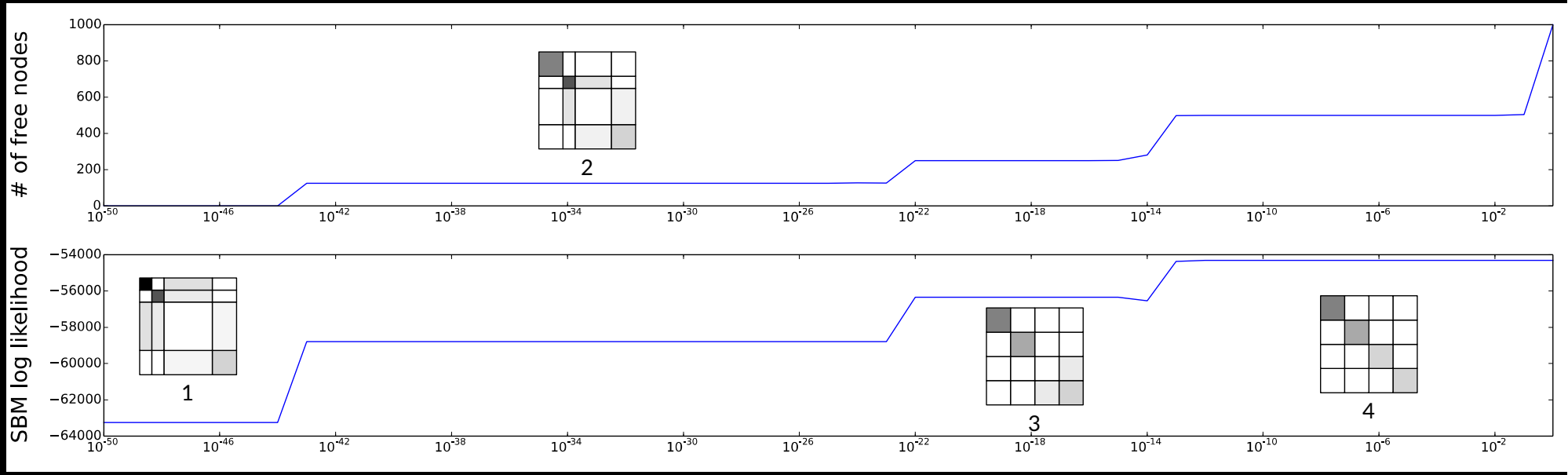


assortative
(SBM comms., C)

arXiv:1608.05878

As θ increases the cost of freeing a node decreases





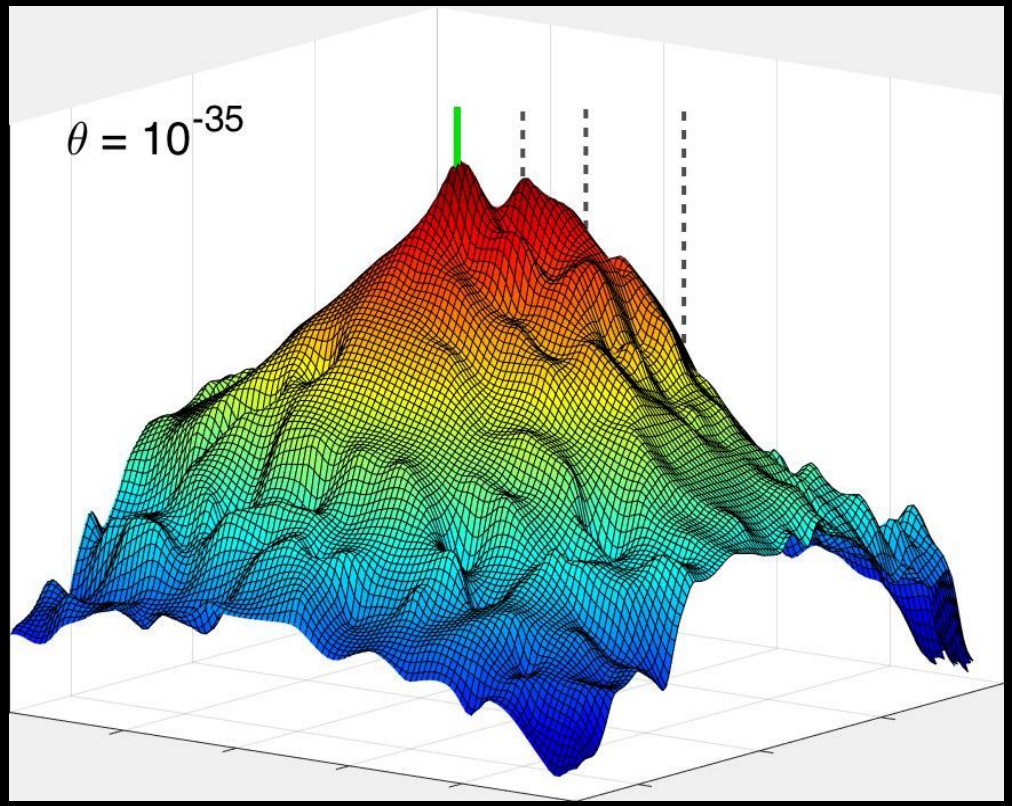
Metadata partition

θ

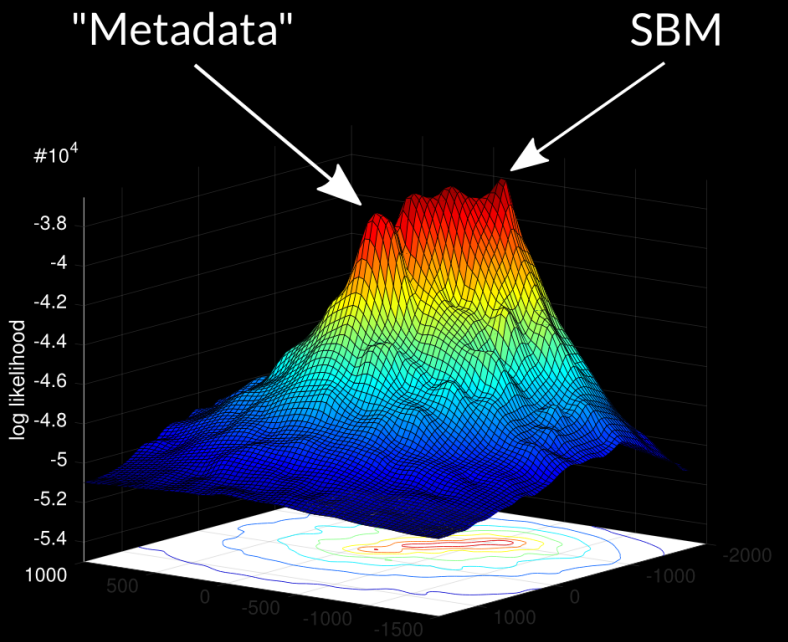
SBM partition

arXiv:1608.05878

neoSBM log likelihood



SBM log likelihood

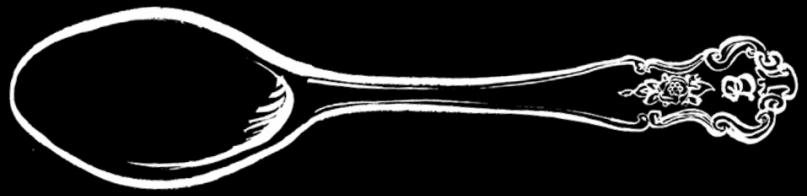


As θ increases the cost of freeing a node decreases

arXiv:1608.05878

arXiv:1608.05878

There is no ground truth



The future of community detection

"I don't know the future. I didn't come here to tell you how this is going to end. I came here to tell you how it's going to begin... Where we go from there is a choice I leave to you."

– Neo, The Matrix

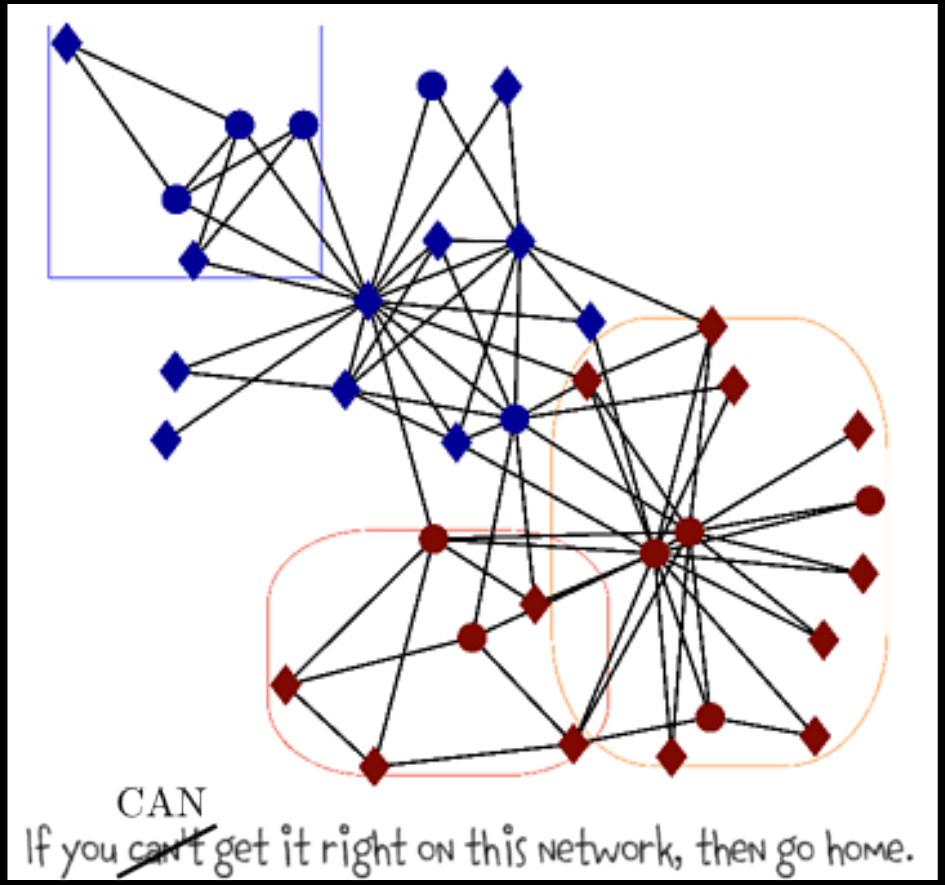
In colloboration with...

arXiv:1608.05878

Dan Larremore

Aaron Clauset

Questions?



arXiv:1608.05878