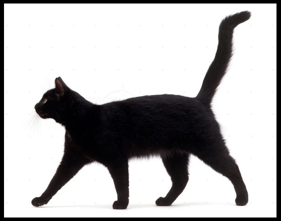
The ground truth about metadata and community detection in networks

Leto Peel
Université catholique de Louvain

Networks can have *metadata* attributes 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.

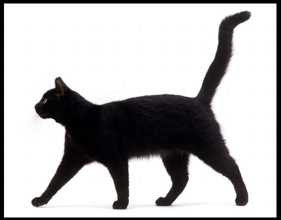
metadata *M* is often used to evaluate the accuracy of community detection algs.



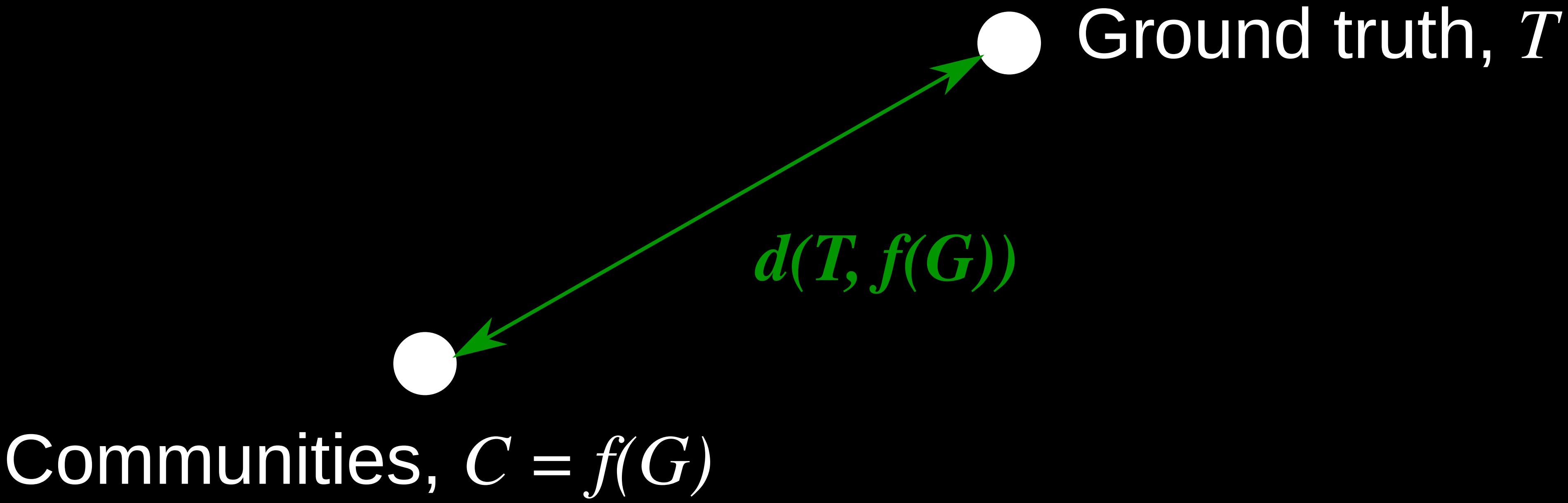
Networks can have *metadata* attributes that describe the nodes

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.

metadata *M* is often used to evaluate the accuracy of community detection algs.



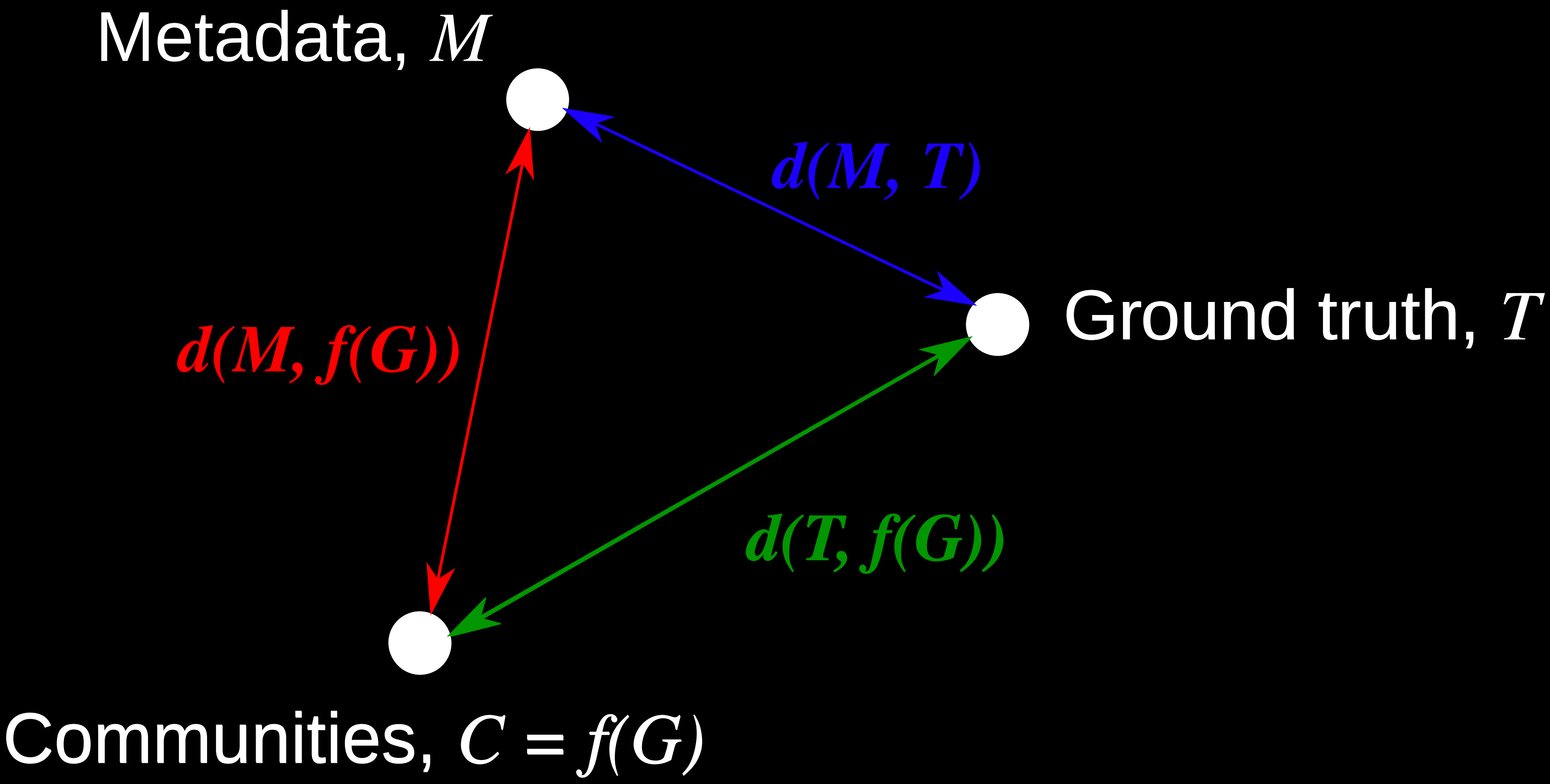
Do you think thats ground truth you're detecting?



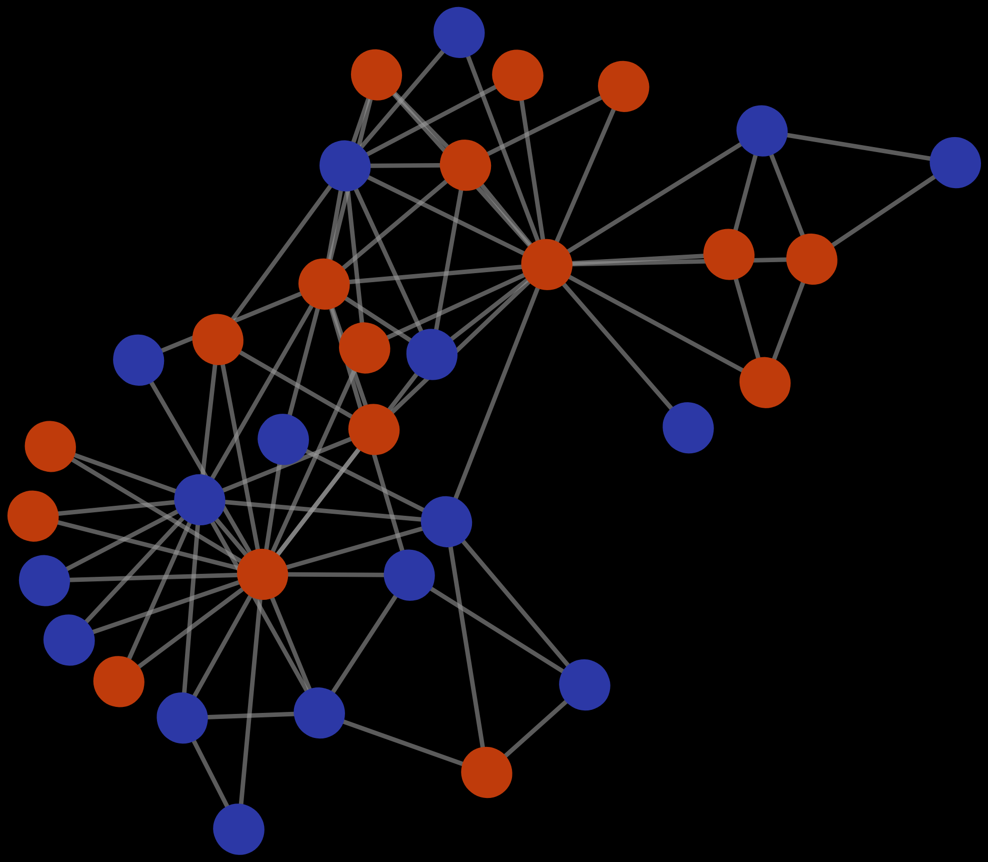
Ground truth, T

$d(T, f(G))$

Communities, $C = f(G)$

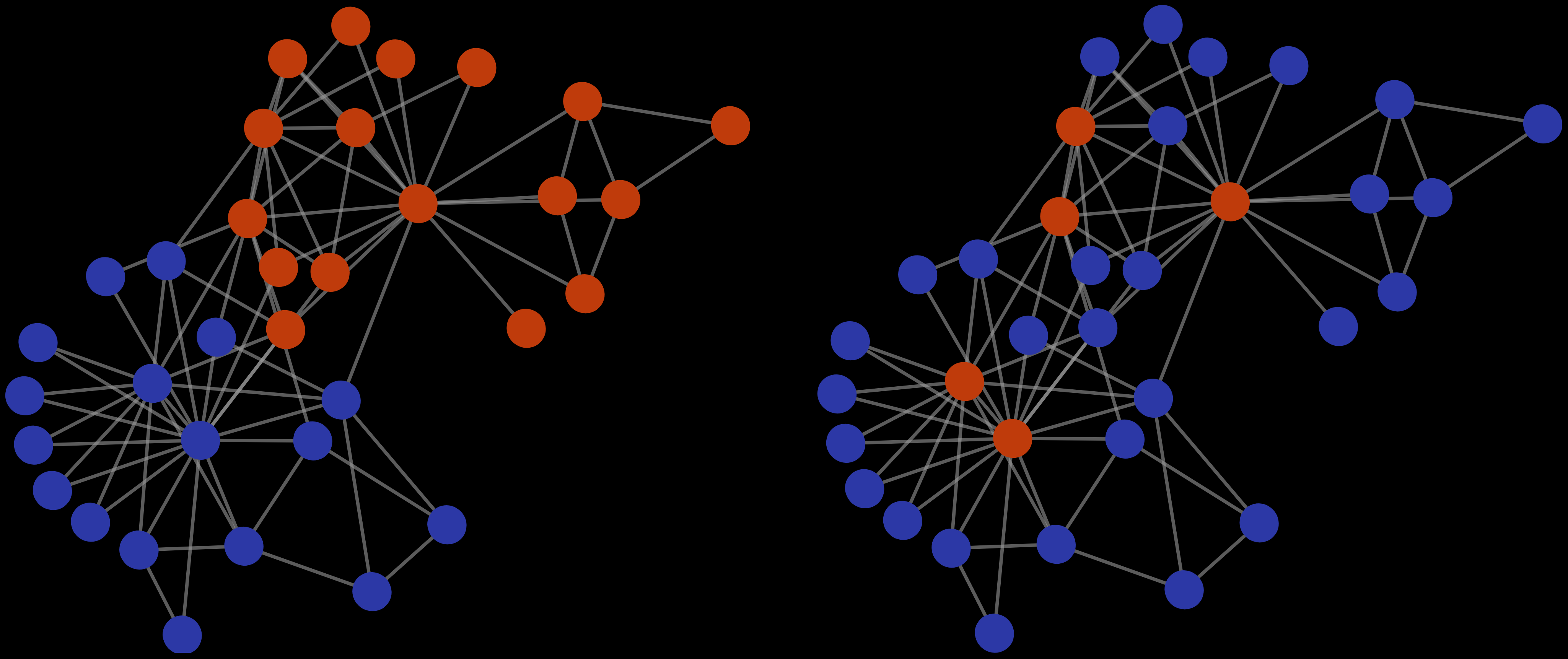

Metadata, M
$d(M, T)$
Ground truth, T
$d(M, f(G))$
$d(T, f(G))$
Communities, $C = f(G)$

When communities ≠ metadata...



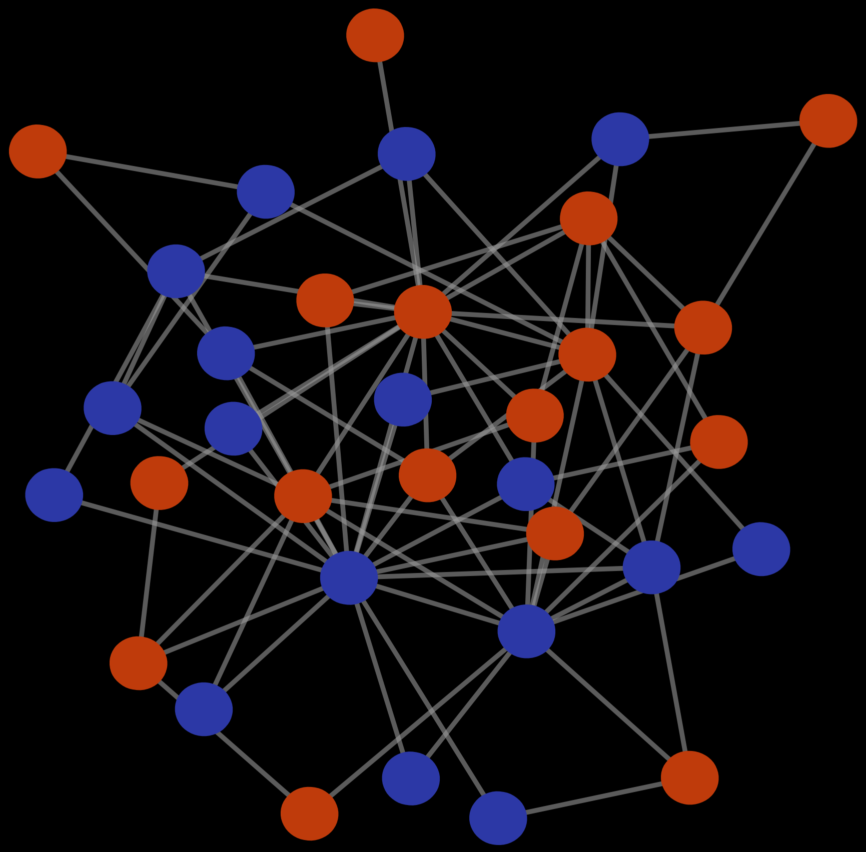
(i) the metadata do not relate to the network structure,

When communities ≠ metadata...



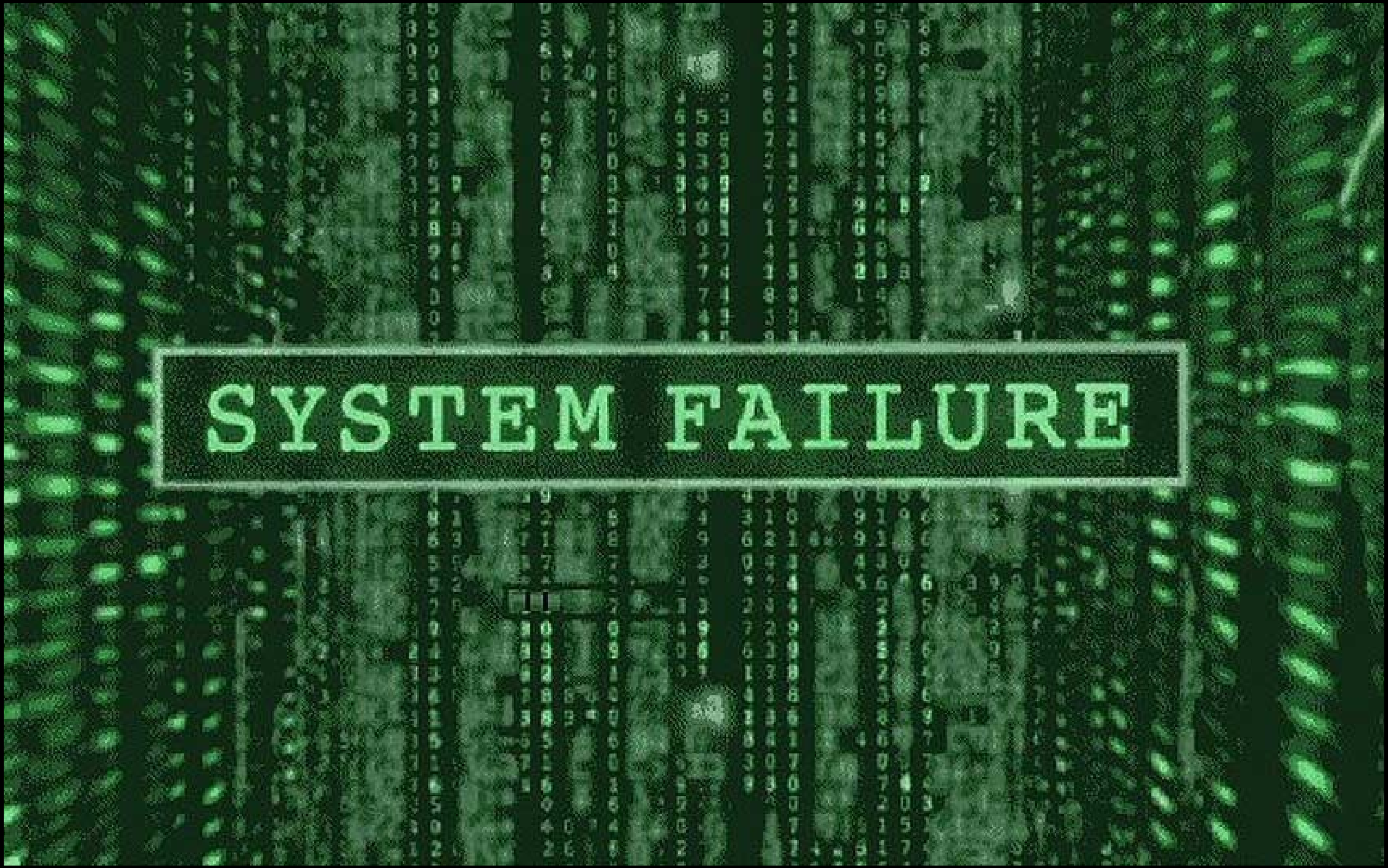
(ii) the detected communities and the metadata capture different aspects of the network's structure,

When communities ≠ metadata...



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

When communities ≠ metadata...



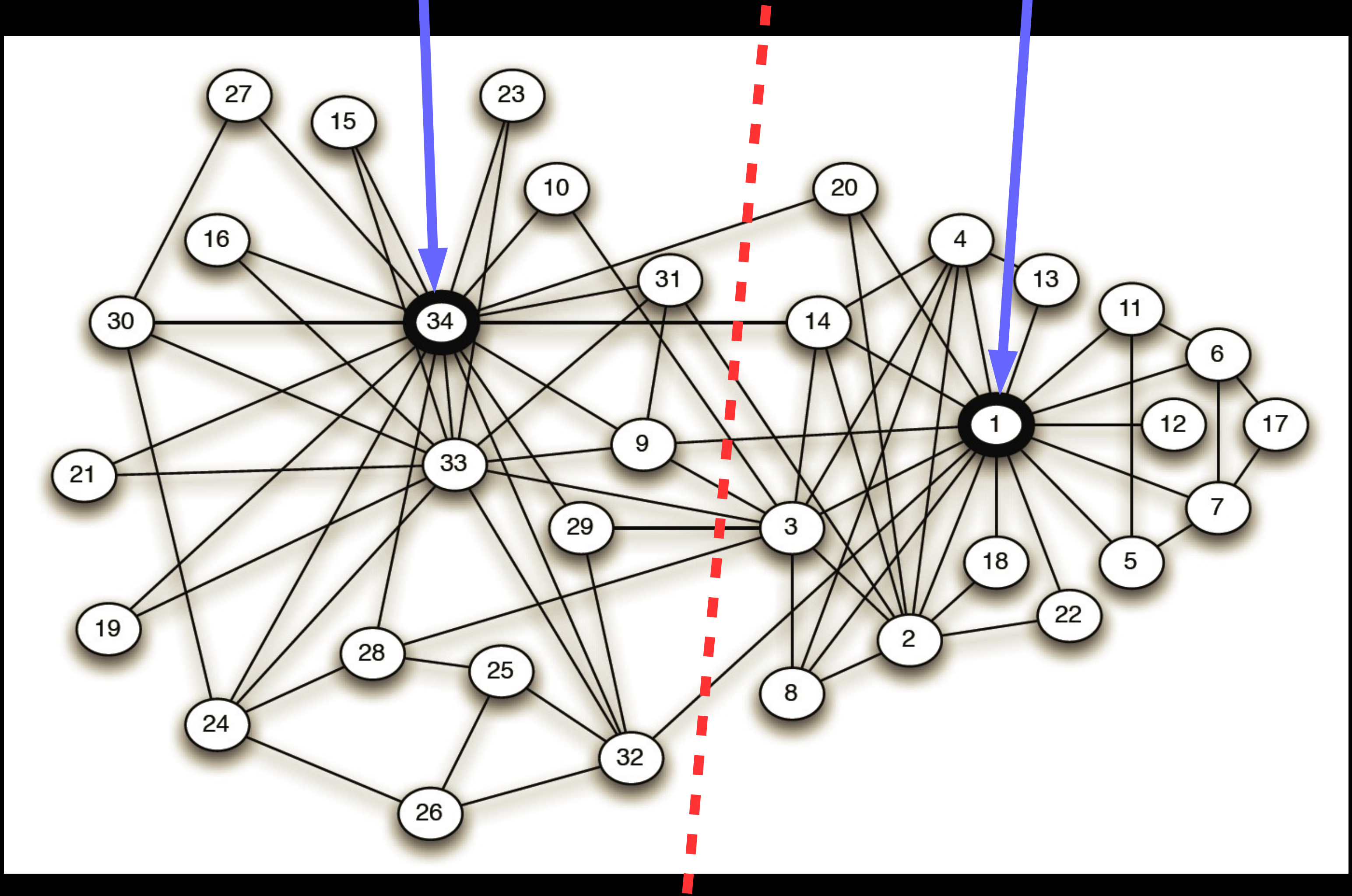
(iv) the community detection algorithm does not perform well.

Typically we assume this is the only possible cause

The Karate Club network

President

Instructor

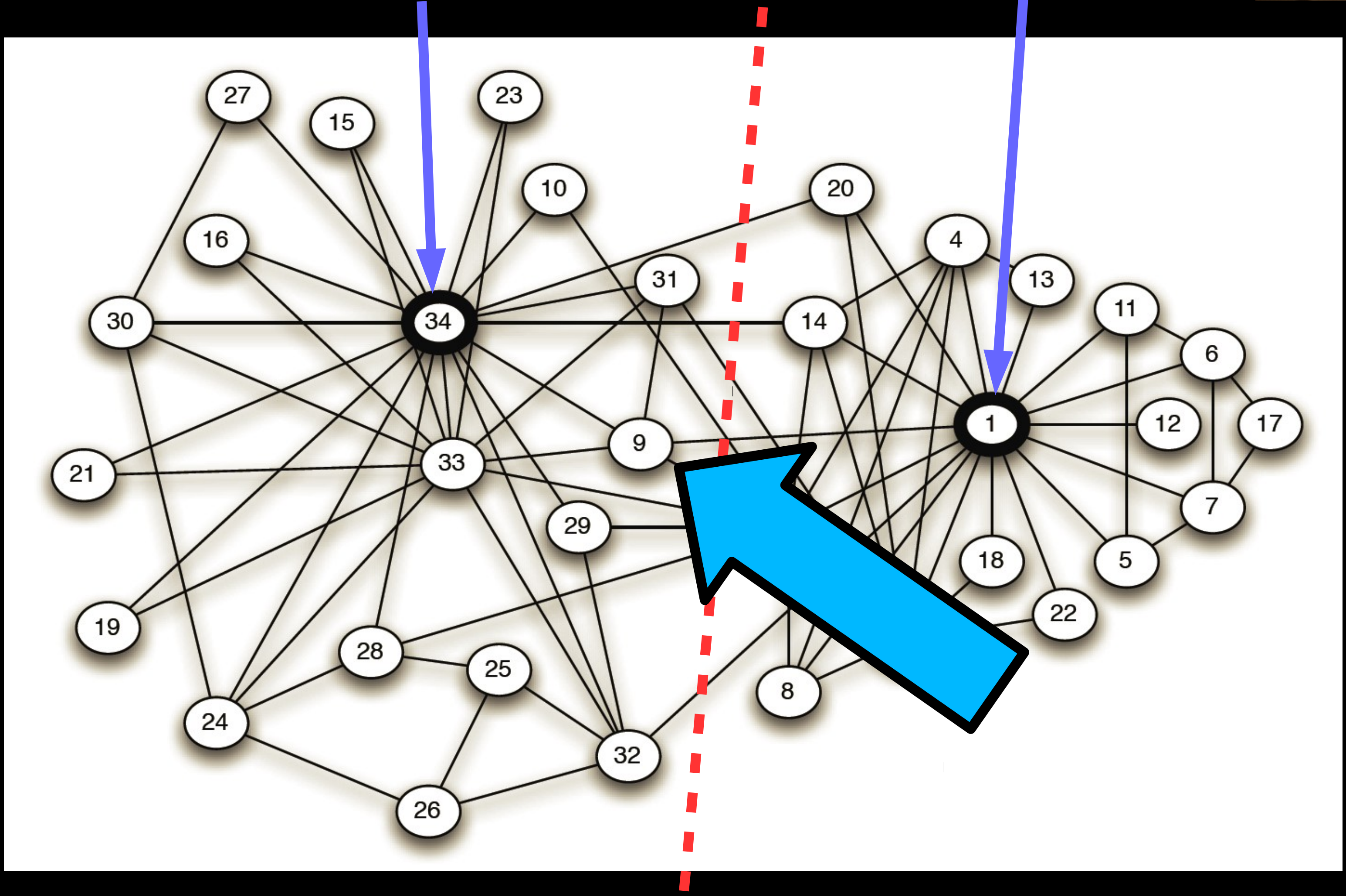


Split into factions

The Karate Club network

President

Instructor

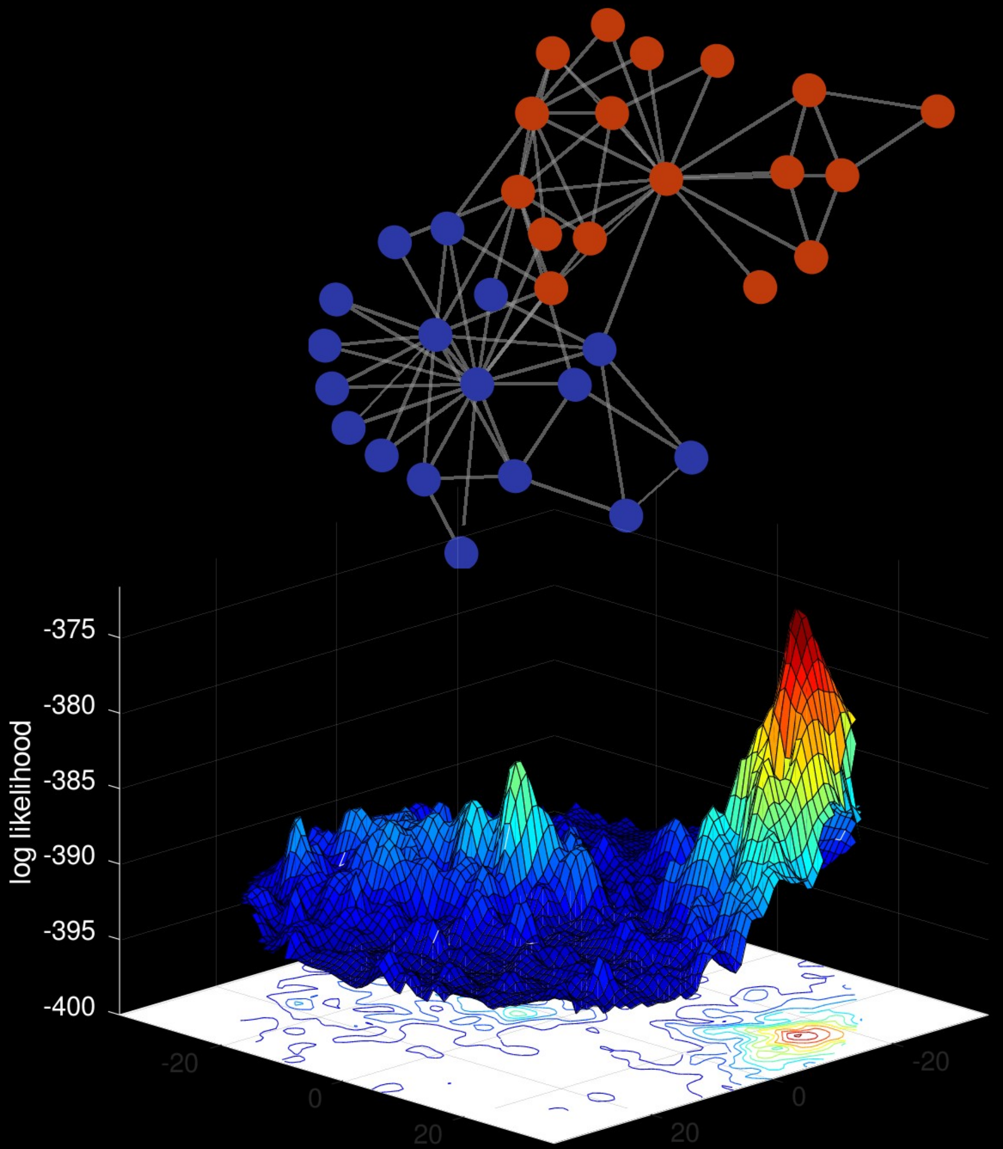


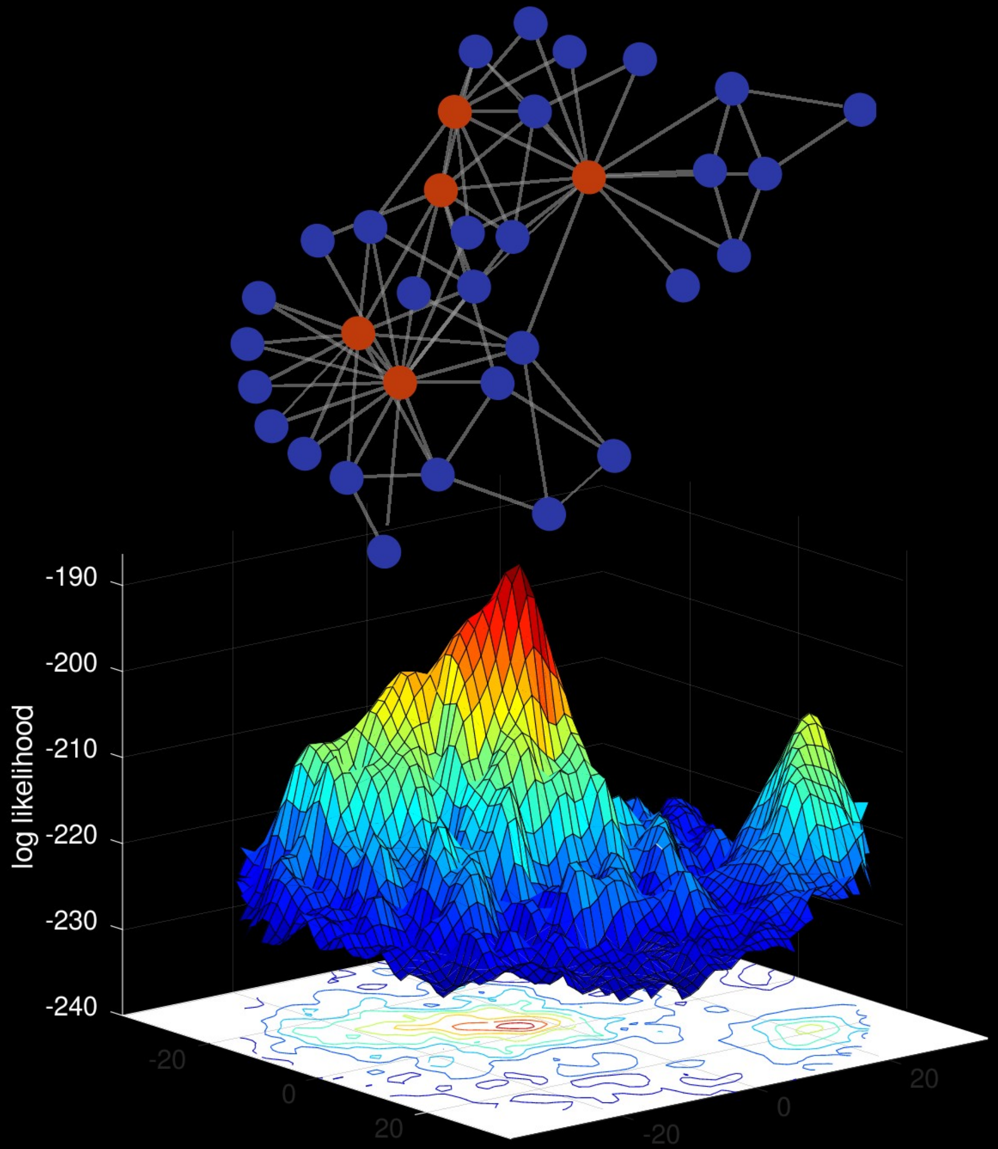
Split into factions

‘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’

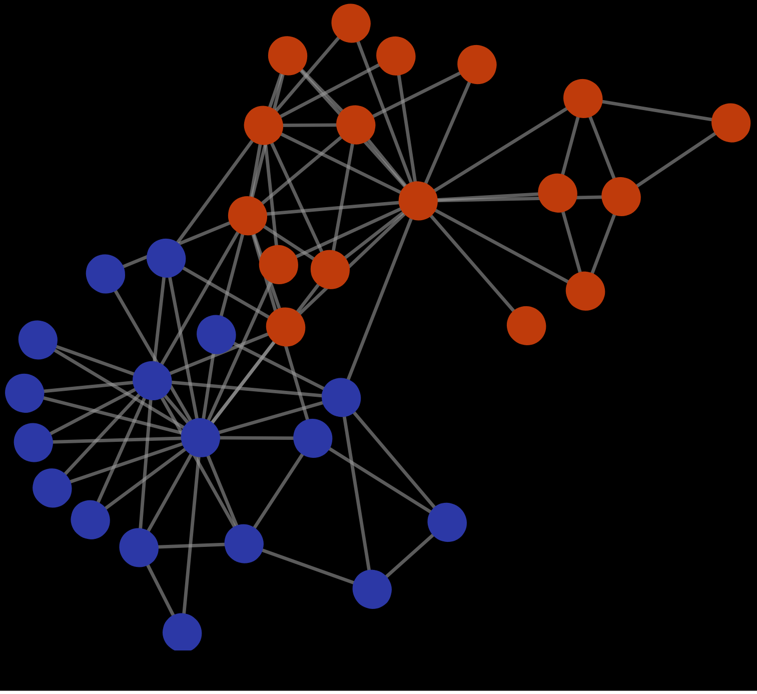
- Zachary 1977

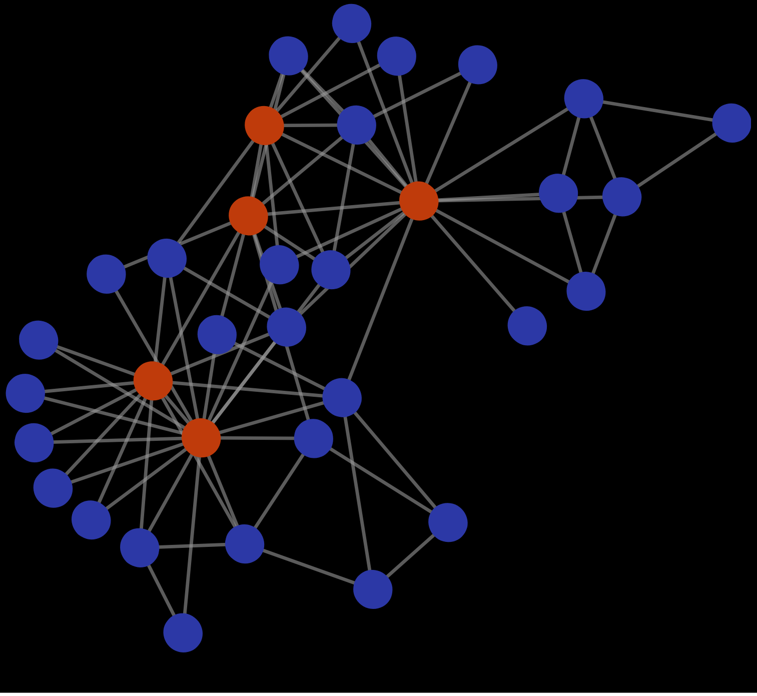
Different generative processes imply different community structures

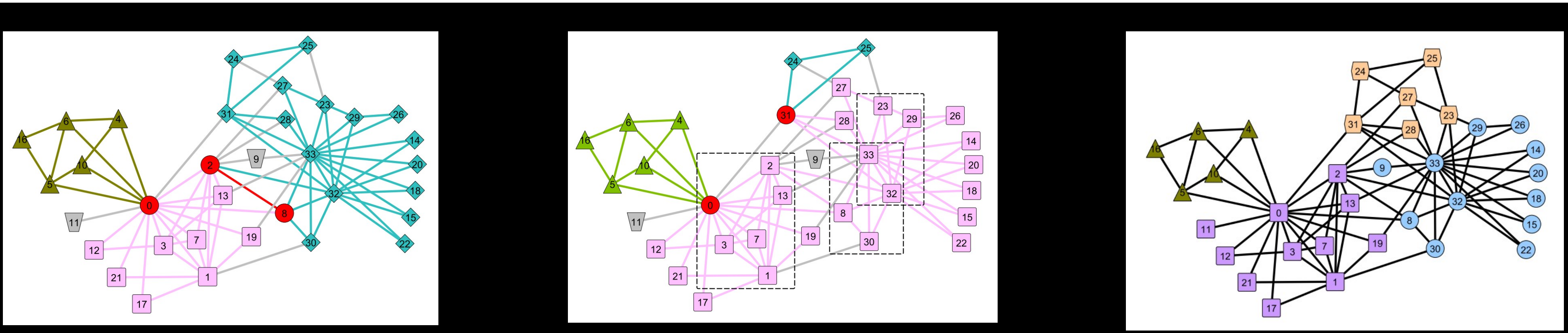




Many good partitions...







Evans 2010

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

So, is metadata useful?

Metadata = types of nodes

Communities = how nodes interact

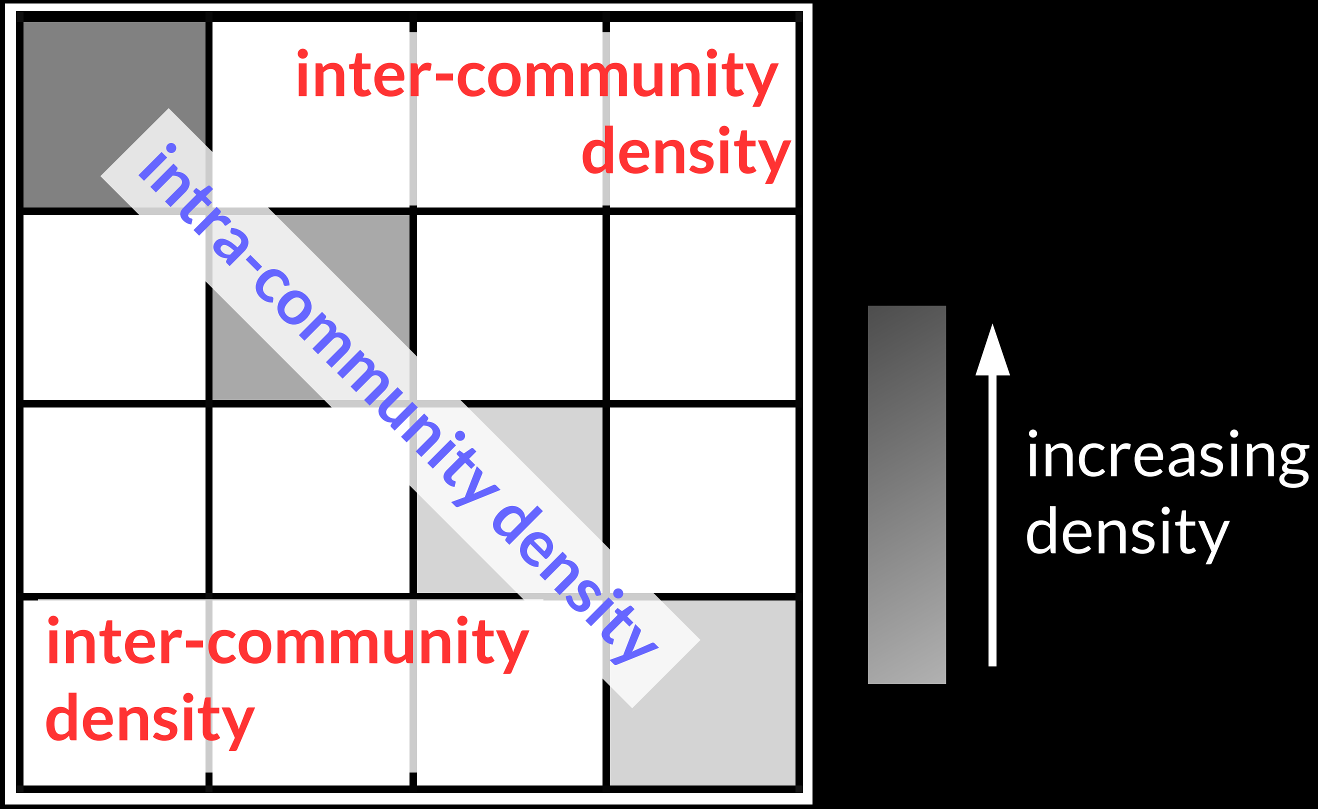
Metadata + Communities = how different types of nodes interact with each other

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

Stochastic Blockmodel

Edges are conditionally independent given community membership

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

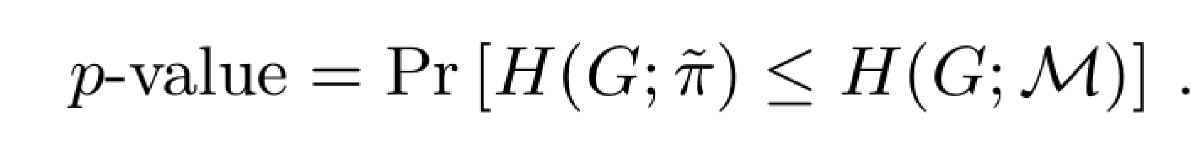


Is the metadata irrelevant to the network structure?

The *blockmodel entropy significance test:*

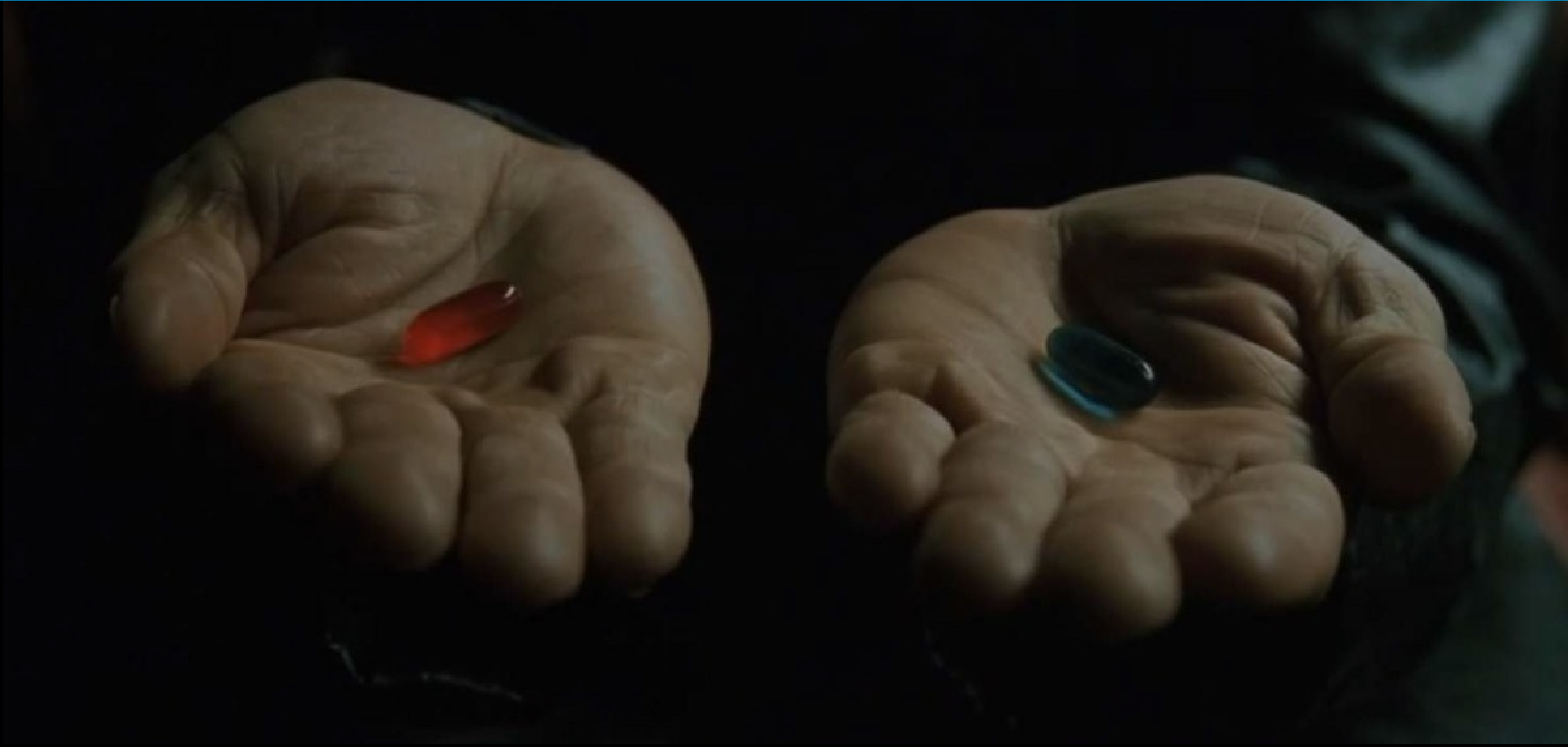
Entropy as a test statistic: the number of bits it takes to describe the network given the model and the metadata partition

Compare entropy of the observed network and metadata with the entropy of random permutations $\{\pi\}$ of the metadata labels



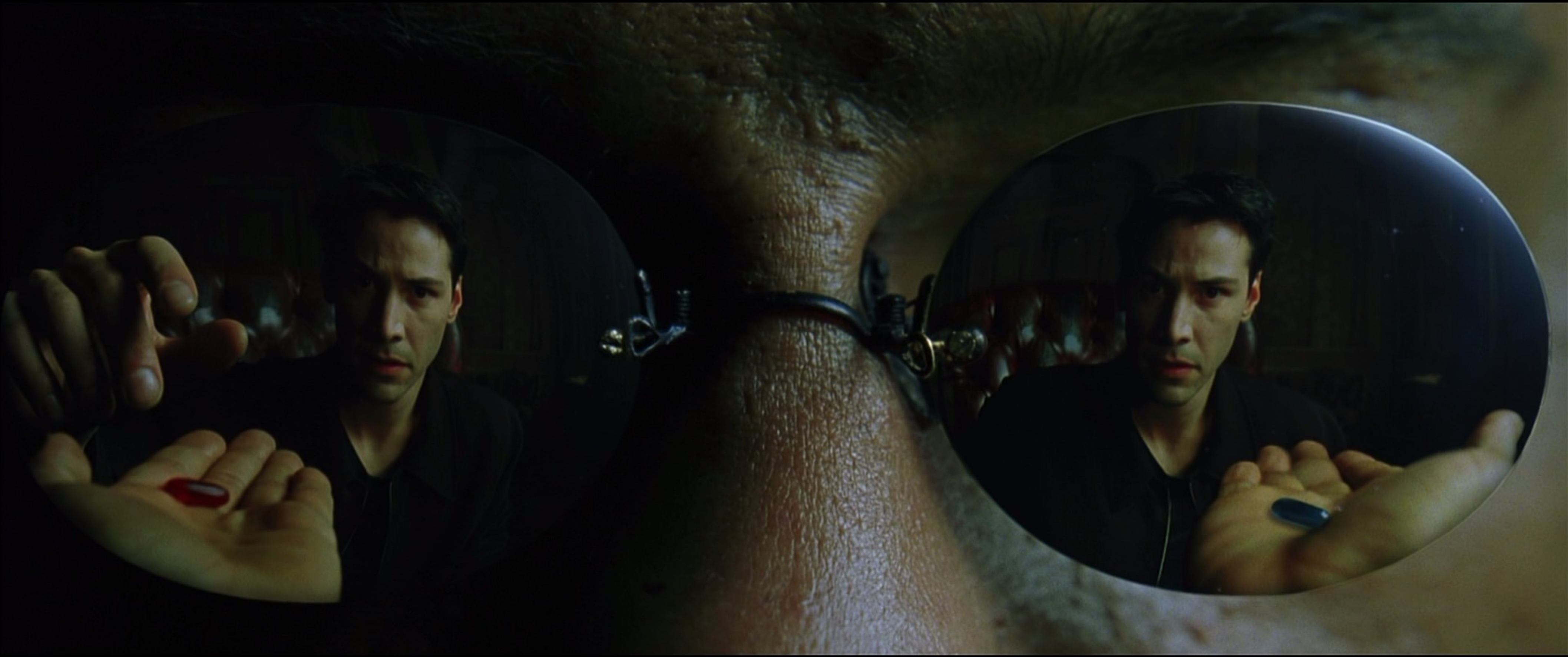
$$p\text{-value} = \Pr\left[H(G;\tilde{\pi}) \leq H(G;\mathcal{M})\right] .$$

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



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

NEOSBM



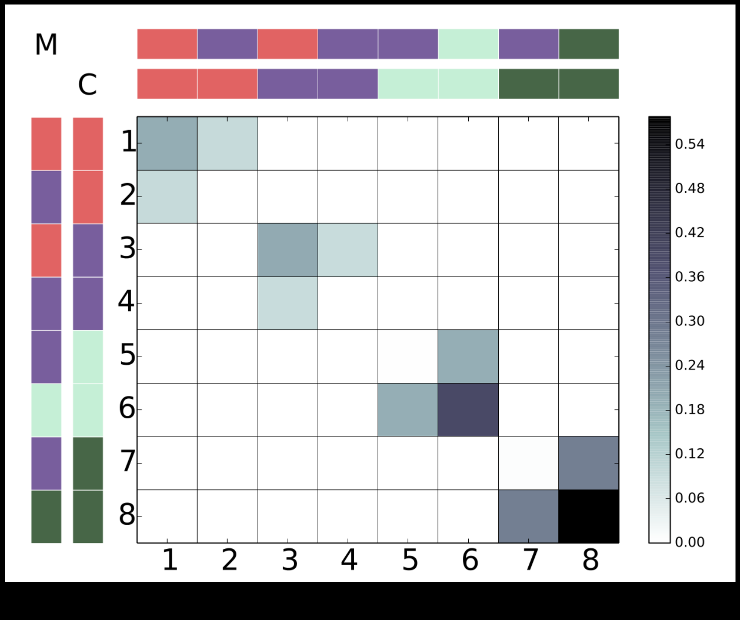
$$\mathcal{L}_{\text{neoSBM}} = \mathcal{L}_{\text{SBM}} + f(\theta)$$

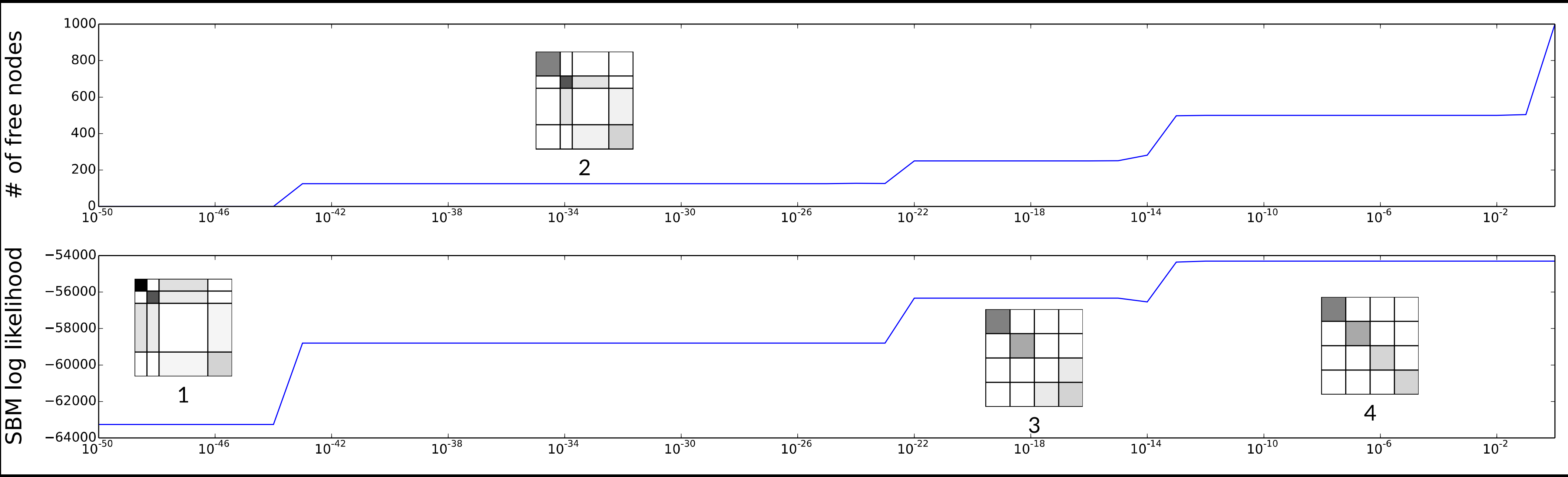
neoSBM log likelihood

SBM log likelihood

cost

Network with multiple 4-group optima



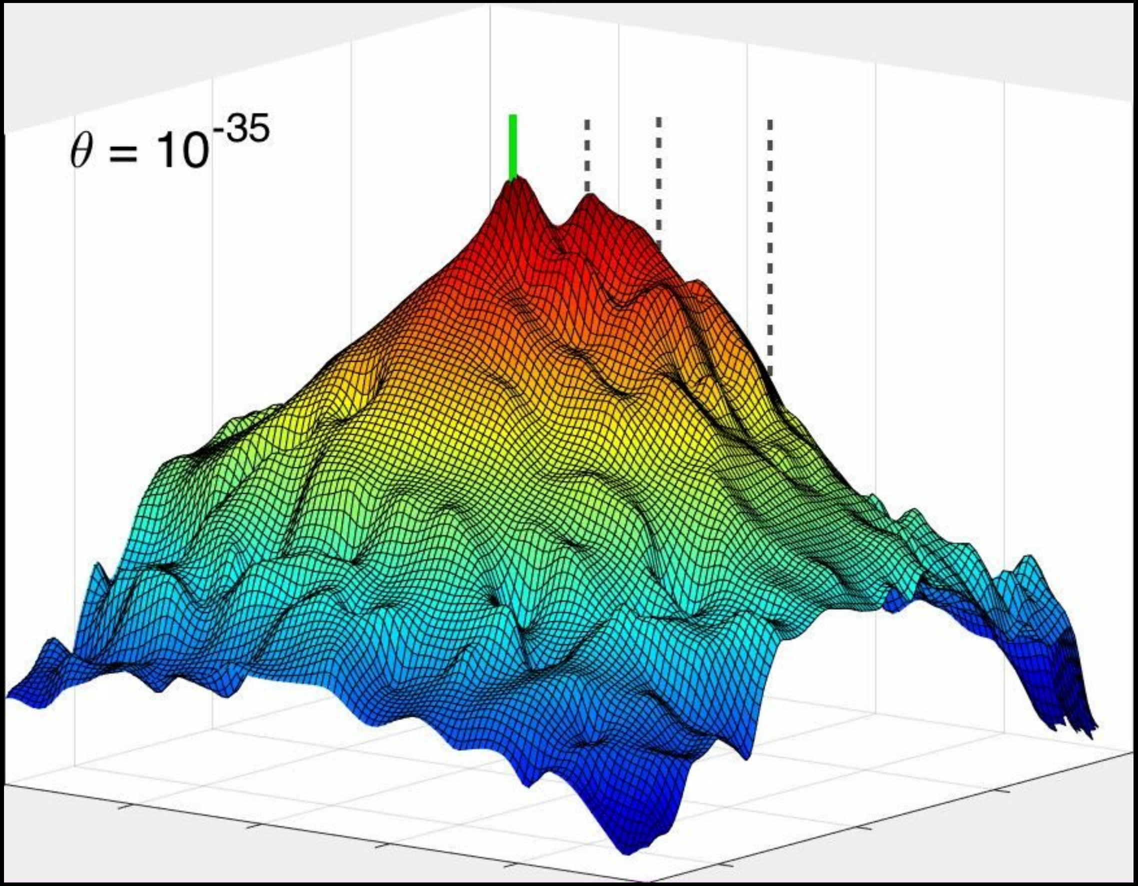


← θ →

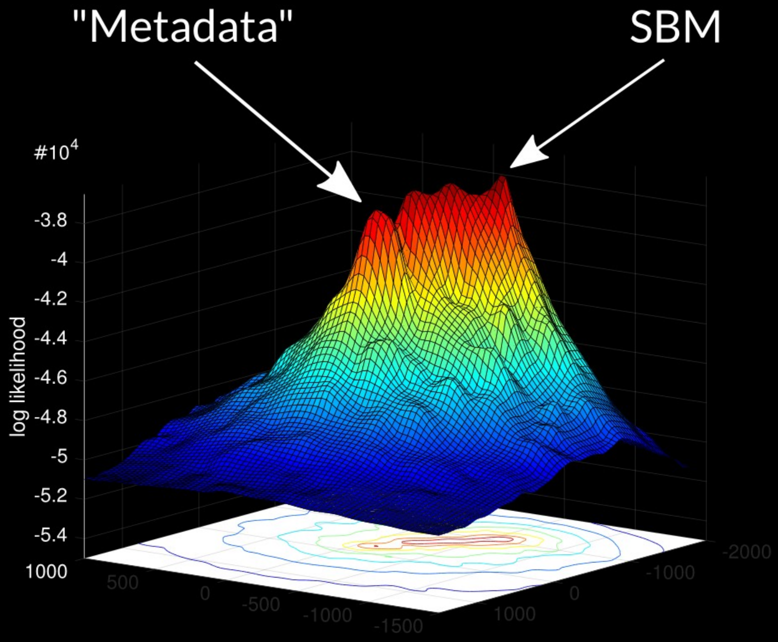
Metadata partition

SBM partition

neoSBM log likelihood



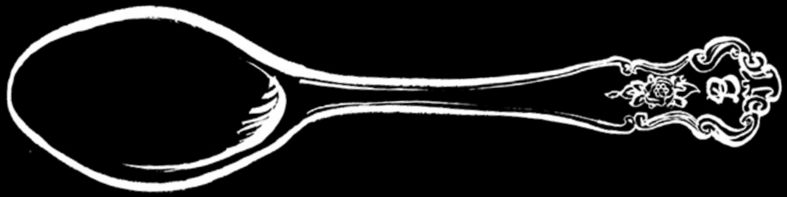
SBM log likelihood


"Metadata"

SBM

As θ increases the cost of freeing a node decreases

There is no ground truth



In colloboration with...

Dan Larremore

Aaron Clauset

Coarse-graining of Complex Systems

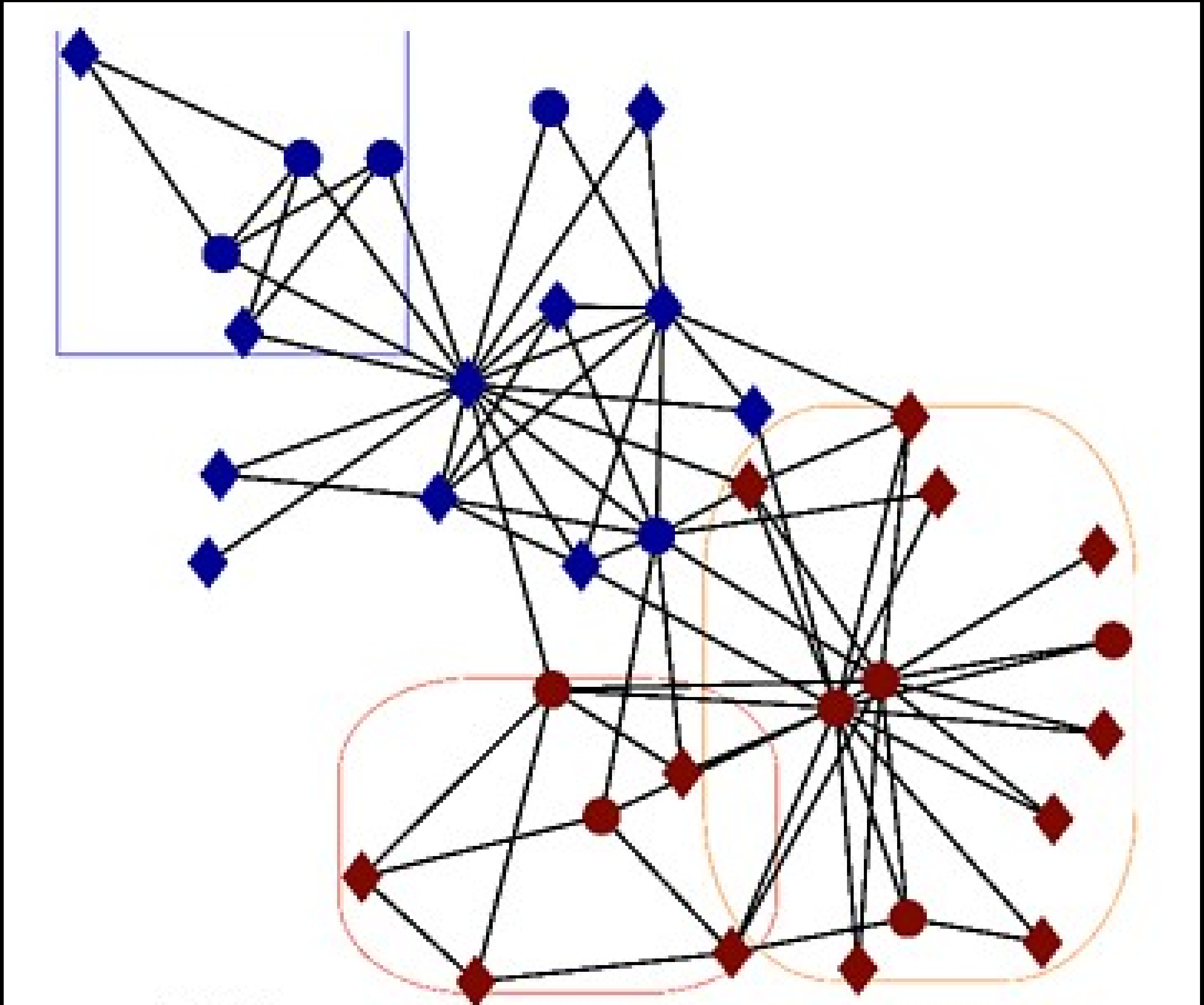
CCS@CCS'16

with Mauro Faccin, Renaud Lambiotte and Michael Schaub

Call for abstracts deadline: June 24

http://michaelschaub.github.io/ccs_at_ccs_2016/

Questions?



If you ~~can't~~ CAN get it right on this network, then go home.