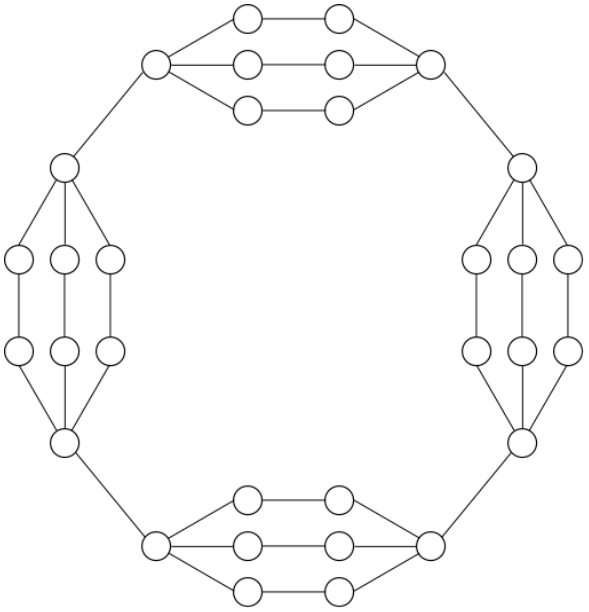
Beyond networks:
Incorporating node metadata into network analysis

Leto Peel
Université catholique de Louvain
@PiratePeel

Here is a network G=(V,E)



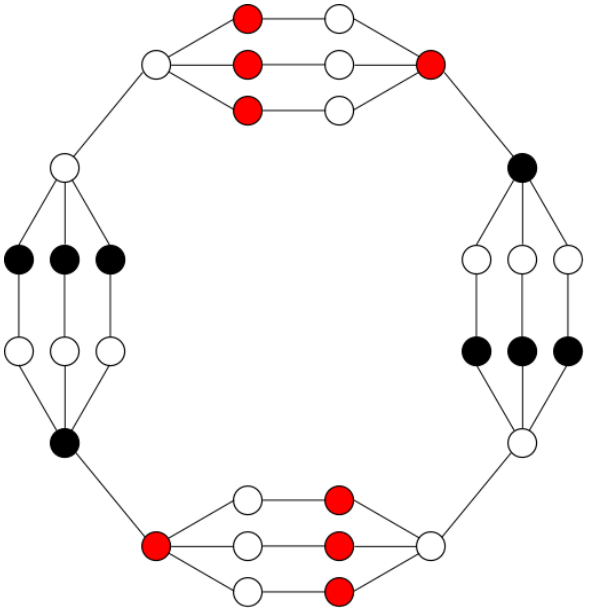
social networks

food webs

internet

protein interactions

Network nodes can have properties or attributes (metadata)



Metadata (M) values

Metadata (M) unknown

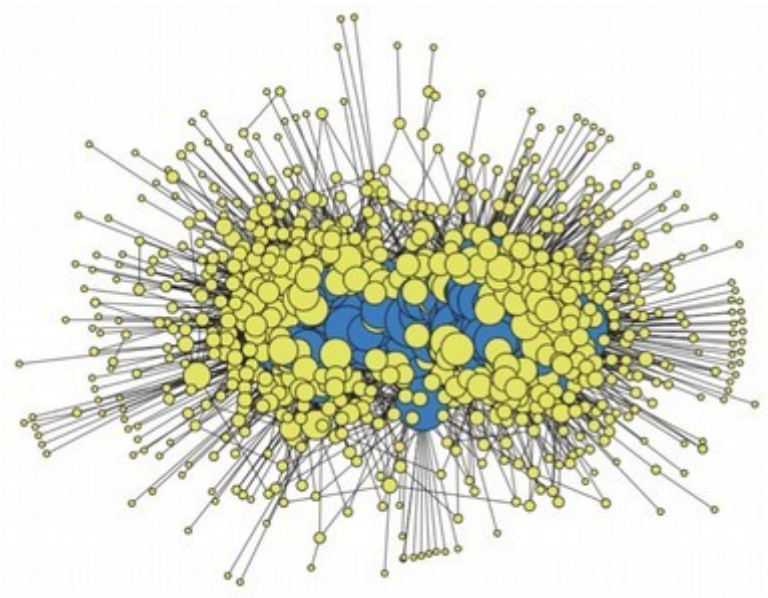
social networks *age, sex, ethnicity, race, etc.*

food webs *feeding mode, species body mass, etc.*

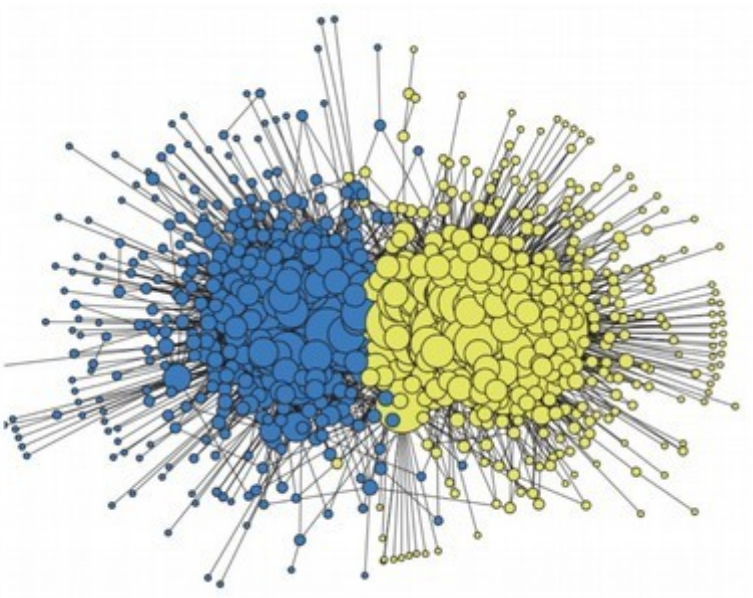
internet *data capacity, physical location, etc.*

protein interactions *molecular weight, association with cancer, etc.*

Recovering metadata implies sensible methods



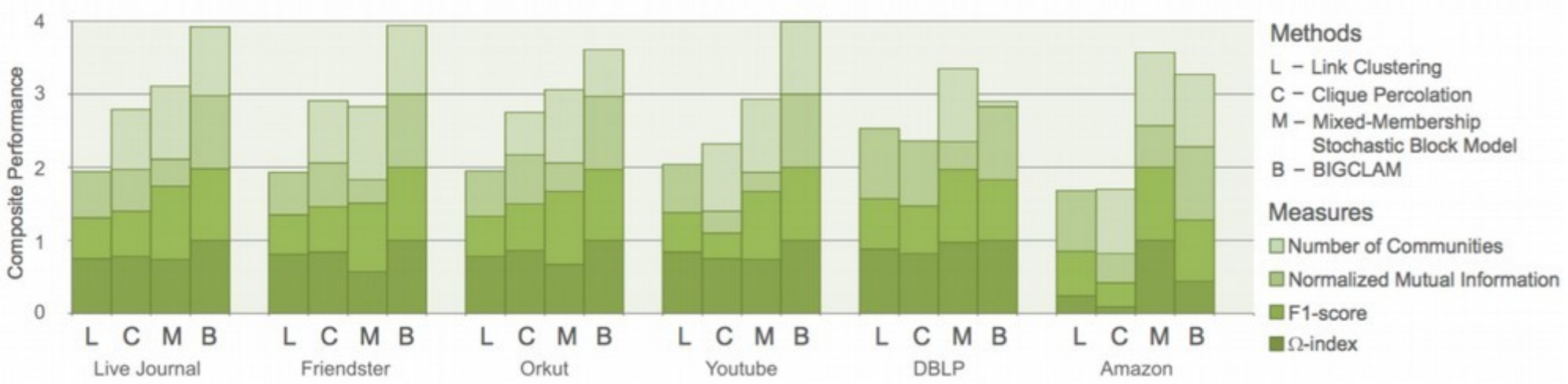
stochastic block model



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).

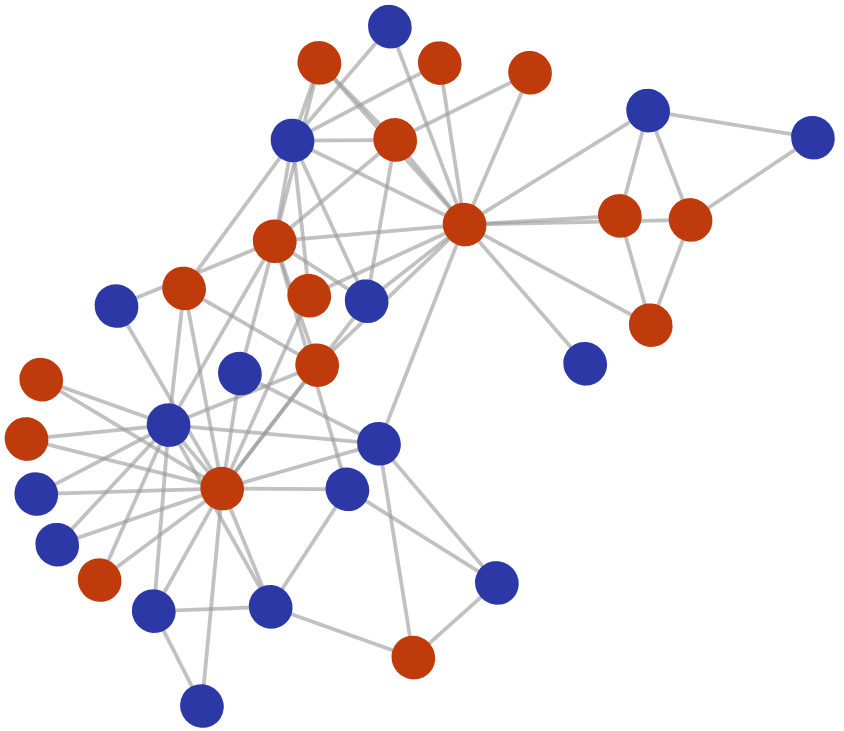
Metadata is often treated as *ground truth*



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


You know, I know these
communities aren't real...

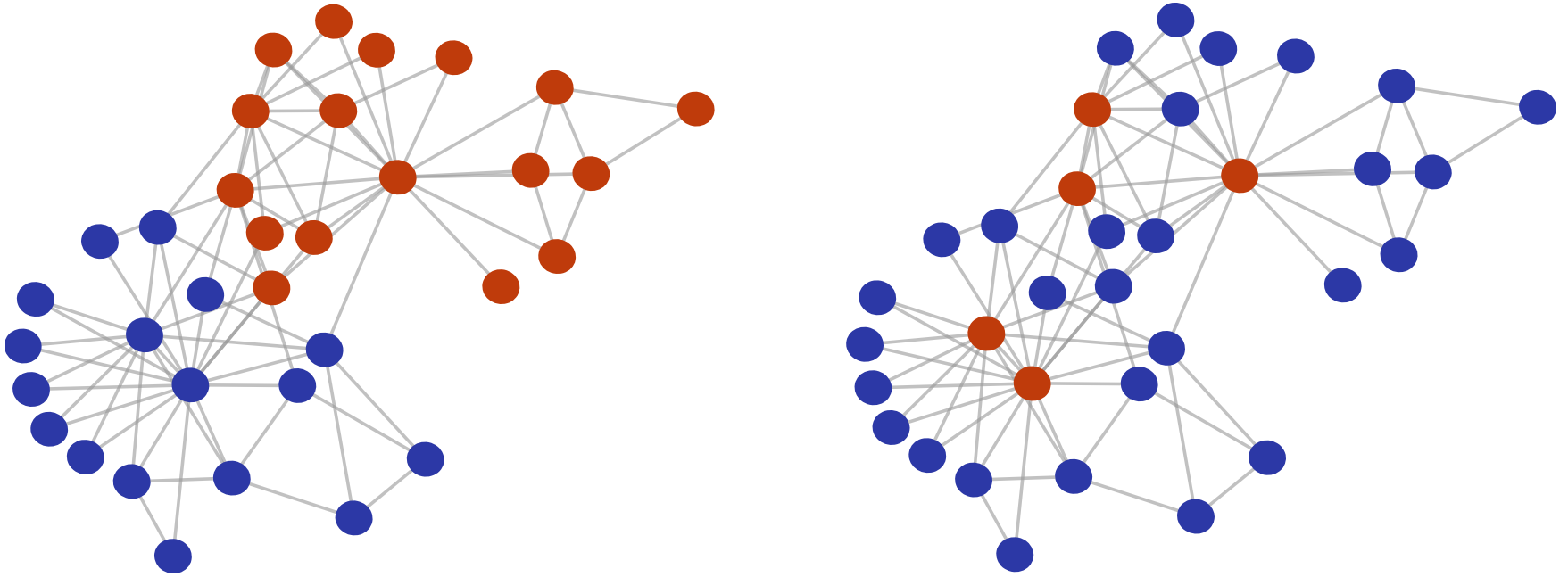
When communities ≠ metadata...



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

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

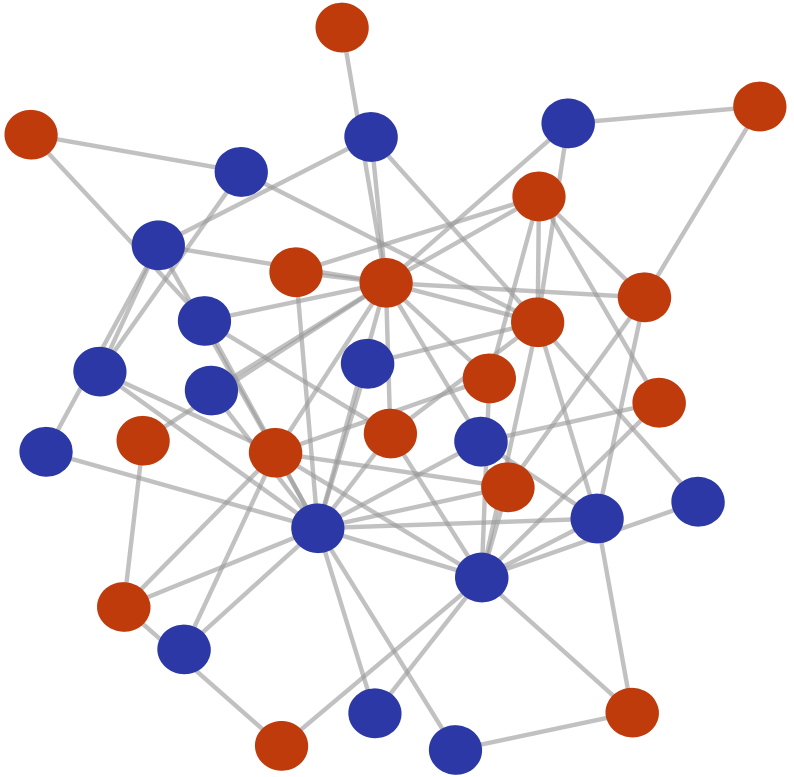
When communities ≠ metadata...



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

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

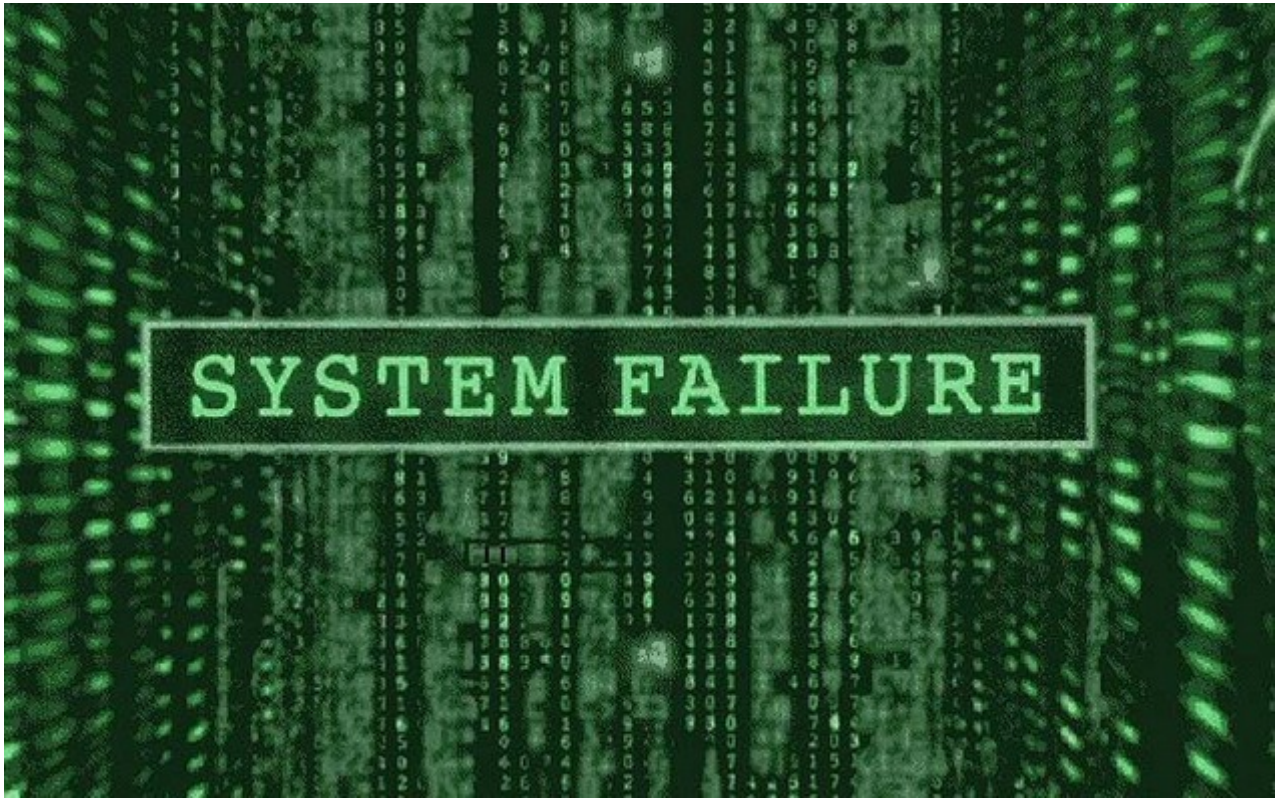
When communities ≠ metadata...



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

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

When communities ≠ metadata...



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

Typically we assume this is the only possible cause

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

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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?

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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.

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

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.

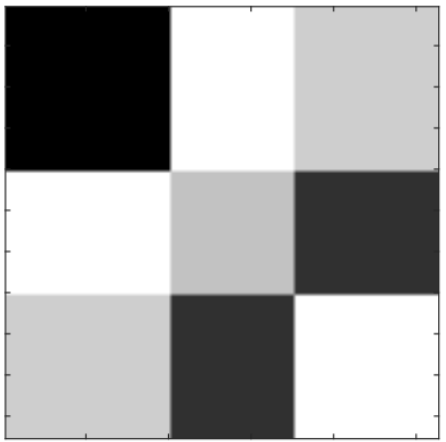
Do we expect all networks across all domains to have the same relationship with communities?

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)


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

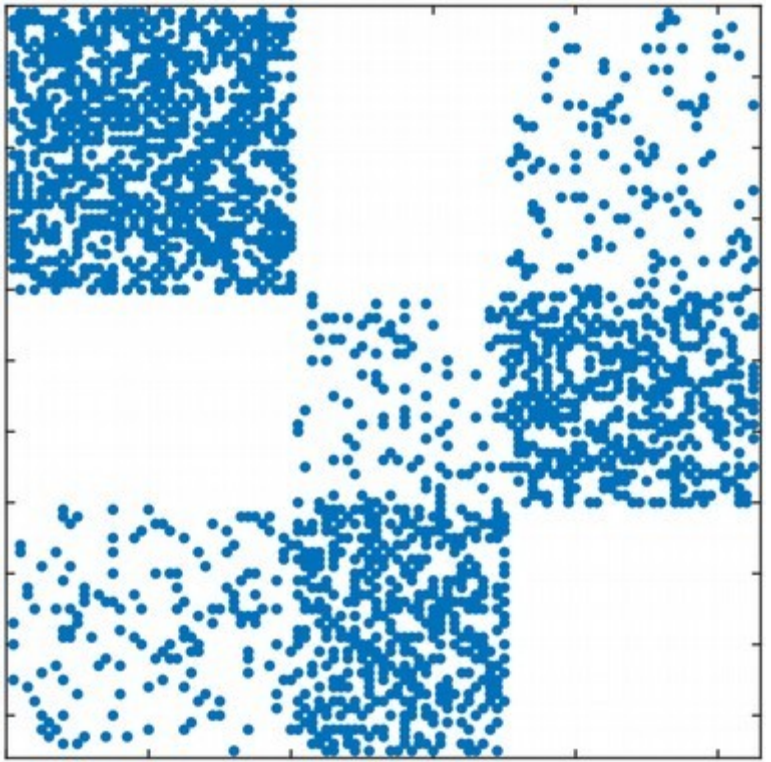
How should we use metadata?

The Stochastic Blockmodel (SBM)



generation

inference



Mixing Matrix

Adjacency Matrix

The Stochastic Blockmodel (SBM)

When communities have equal densities,

inference in SBM = modularity maximisation

Newman, Equivalence between modularity optimization and maximum likelihood methods for community detection. Phys. Rev. E 94, 052315 (2017)

Blockmodel Entropy Significance Test

How well do the metadata explain the network?

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

Blockmodel Entropy Significance Test

How well do the metadata explain the network?

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.

Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

Blockmodel Entropy Significance Test

How well do the metadata explain the network?

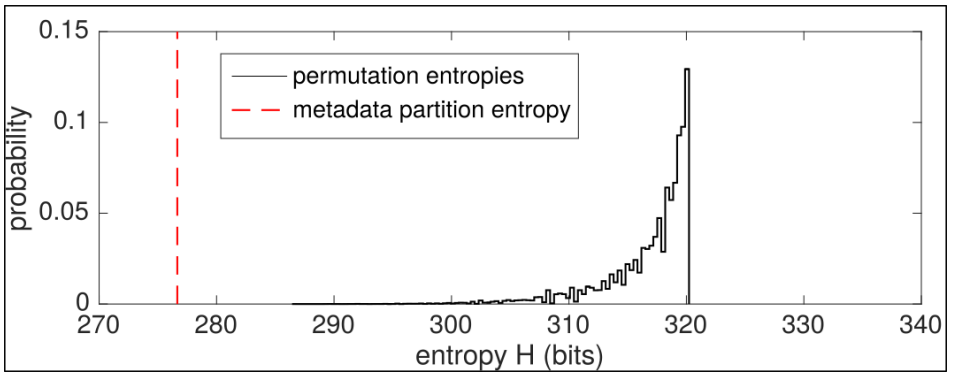
1. Divide the network G into groups according to metadata labels M.

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**



Peel, Larremore, Clauset, "The ground truth about metadata and community detection in networks". Science Advances 3 (5), e1602548 (2017)

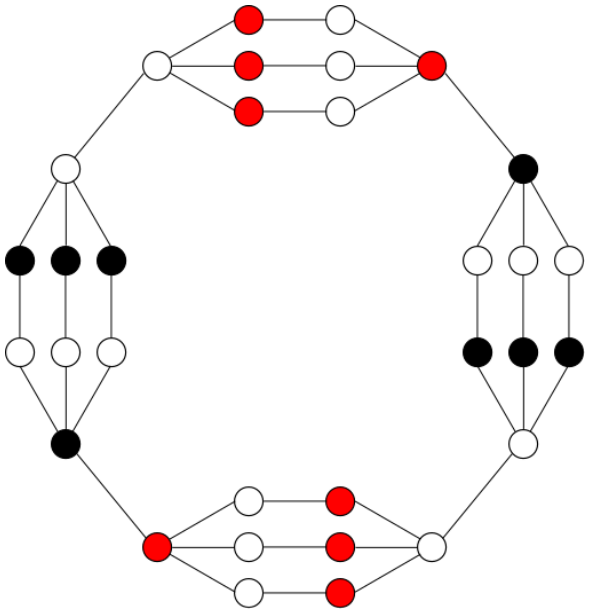
Multiple networks; multiple metadata attributes

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

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

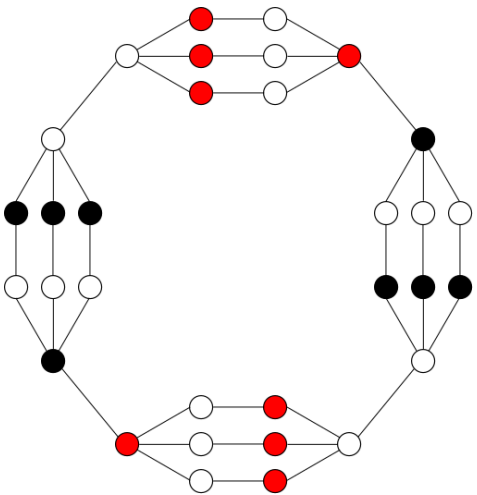
Lazega, The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership, Oxford University Press (2001).

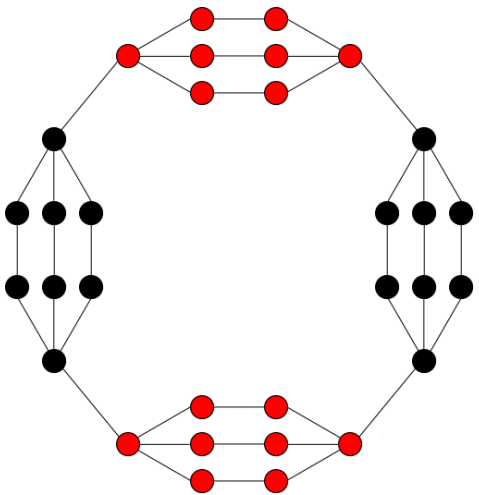
Can we predict unknown metadata values?



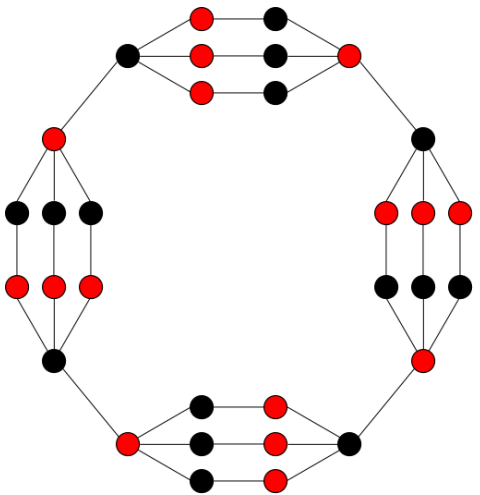
Metadata values

Metadata unknown

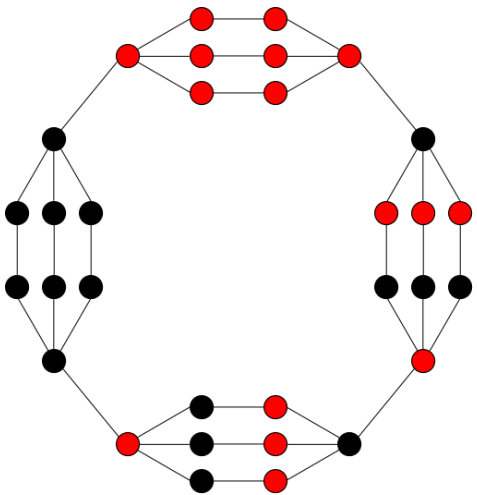




assortative

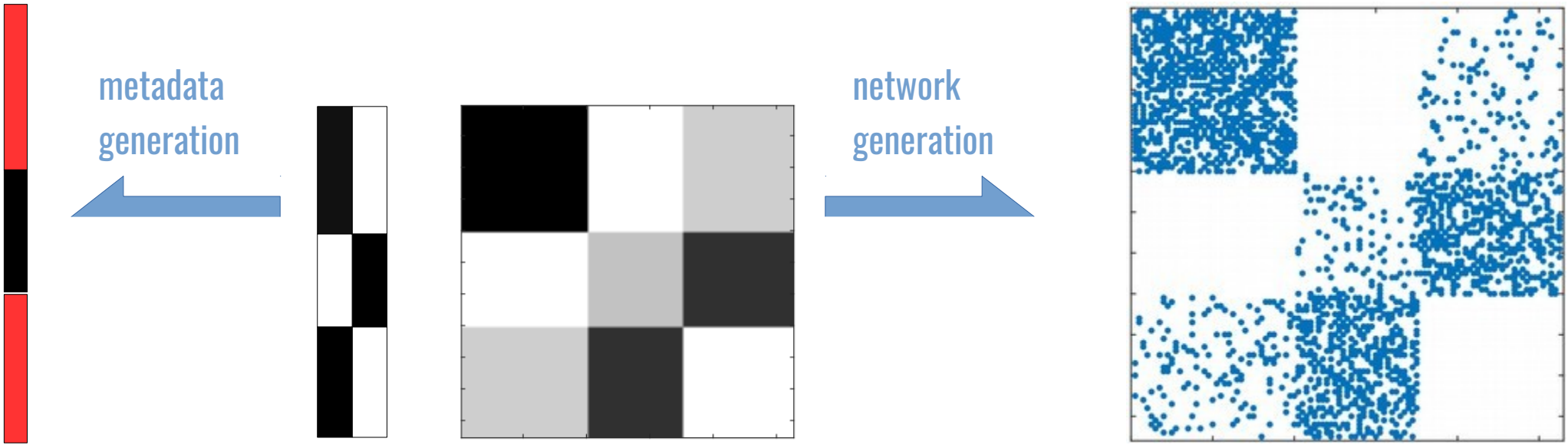


disassortative



mixed

A stochastic block model with metadata



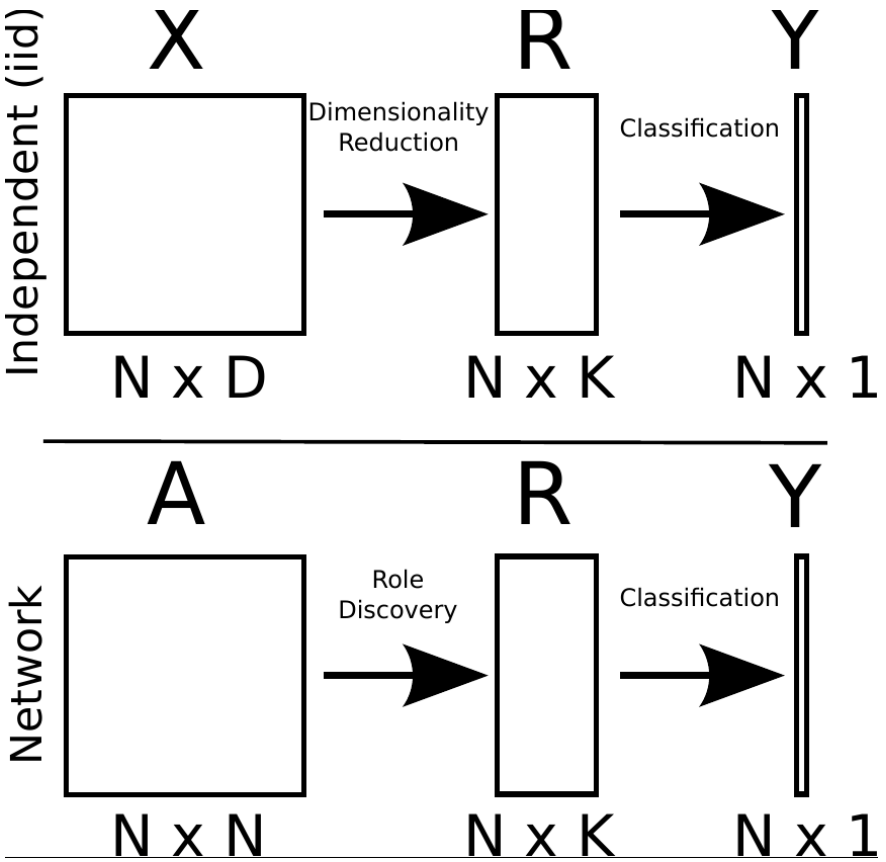
Community –
Metadata Matrix

Mixing Matrix

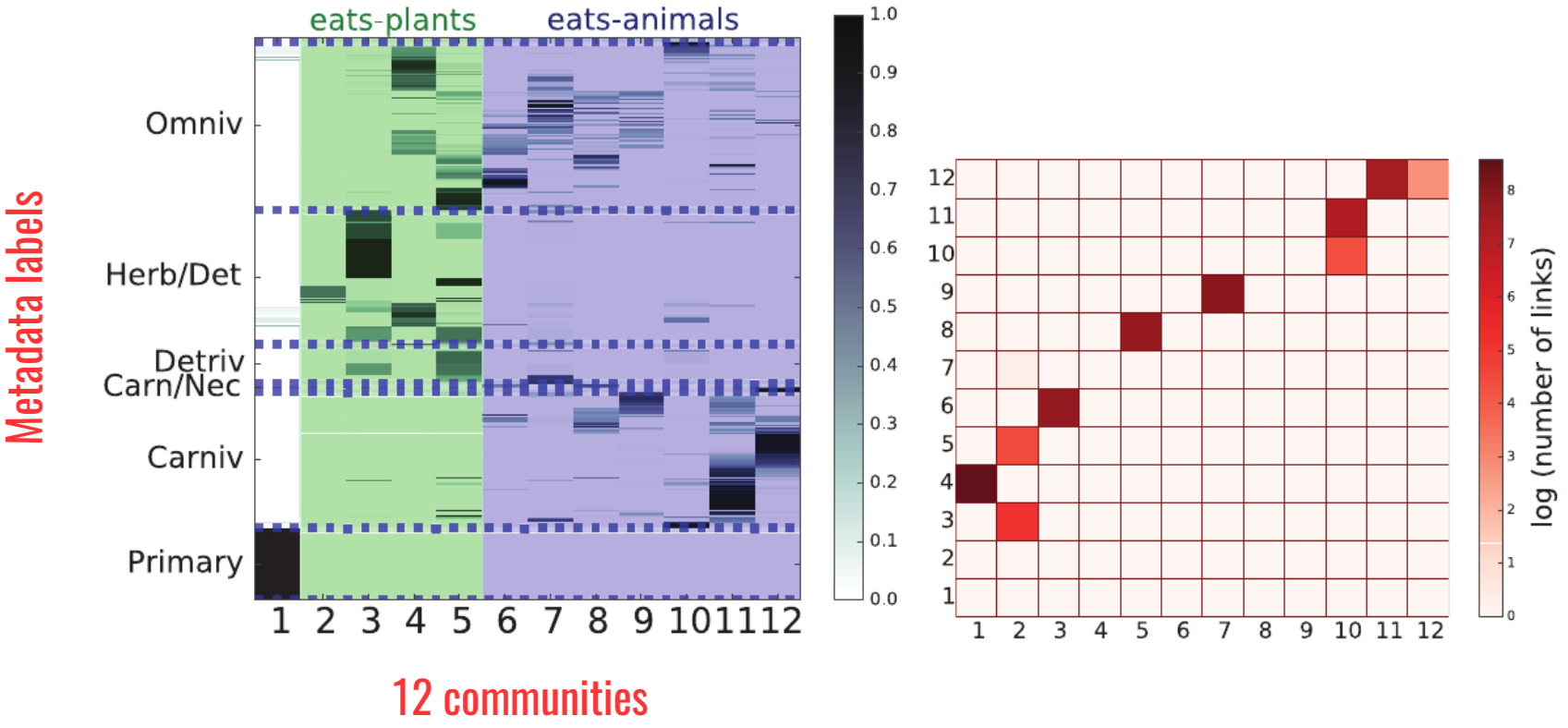
Adjacency Matrix

Peel, L., Topological Feature Based Classification 14th International Conference on Information Fusion (FUSION) 2011
Peel, L., Supervised Blockmodelling ECML/PKDD Workshop on Collective Learning and Inference on Structured Data (CoLISD) 2012

Dimensionality reduction + classification

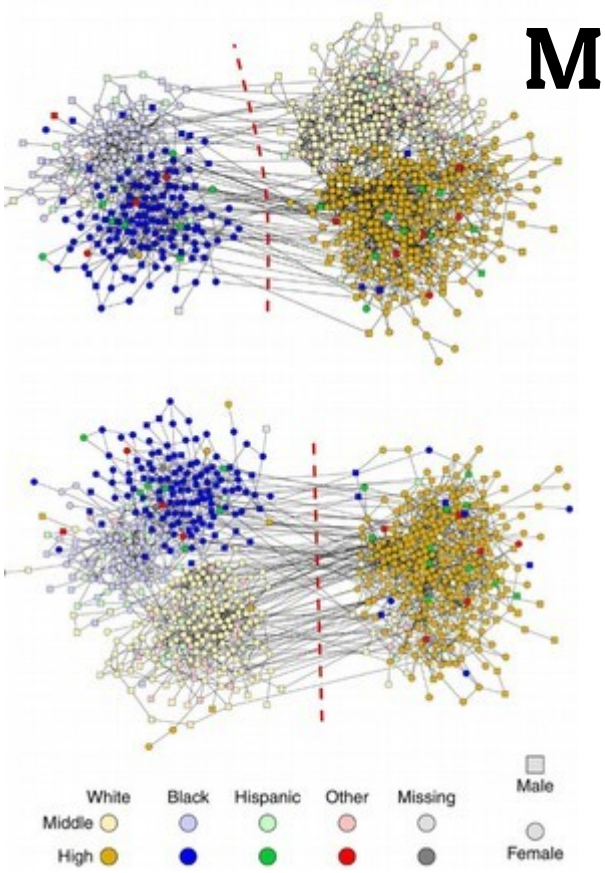


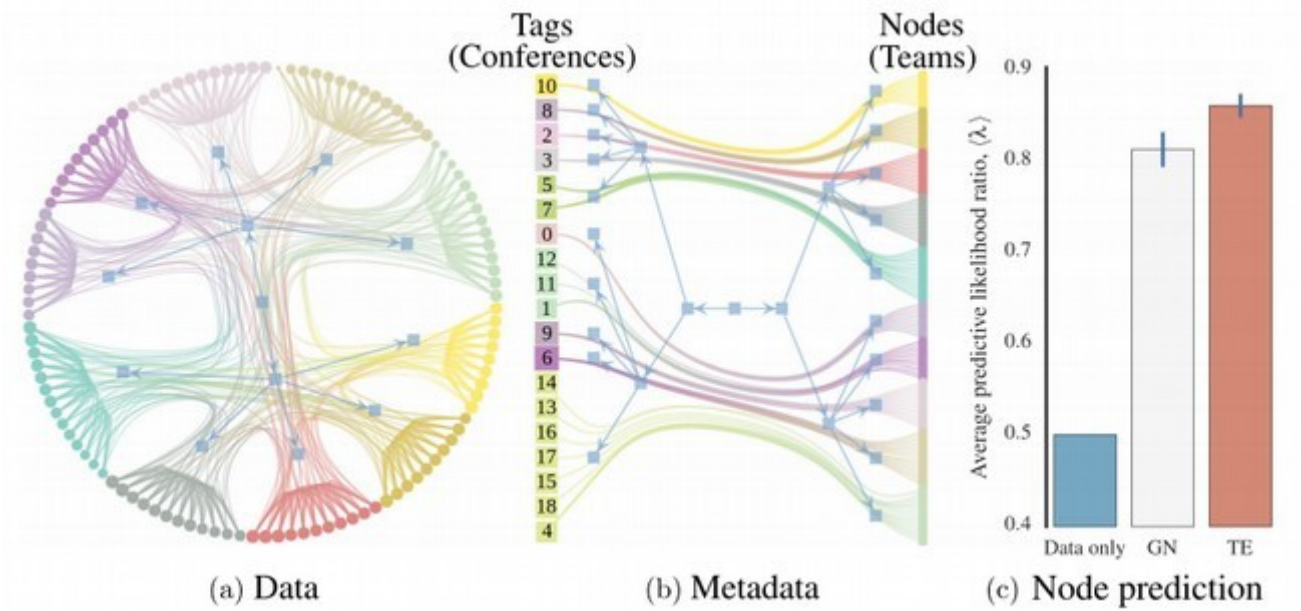
Peel, L., Active Discovery of Network Roles for Predicting the Classes of Network Nodes Journal of Complex Networks 3 (3): 431-449, 2015



Peel, L., Active Discovery of Network Roles for Predicting the Classes of Network Nodes Journal of Complex Networks 3 (3): 431-449, 2015

More SBMs + metadata

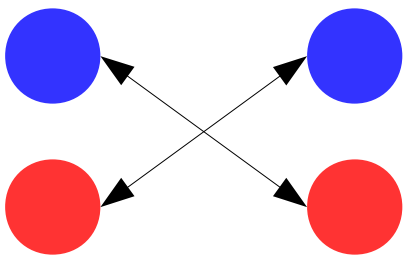




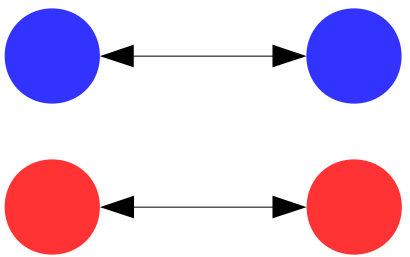
Newman, Clauset. "Structure and inference in annotated networks." Nat. Comms. 7 (2016).
Hric, Peixoto, Fortunato. "Network structure, metadata, and the prediction of missing nodes and annotations." Phys. Rev. X 6.3: 031038 (2016)

Mixing patterns in networks

$$r_{\text{global}} = \frac{\sum_g e_{gg} - \sum_g a_g b_g}{1 - \sum_g a_g b_g}$$



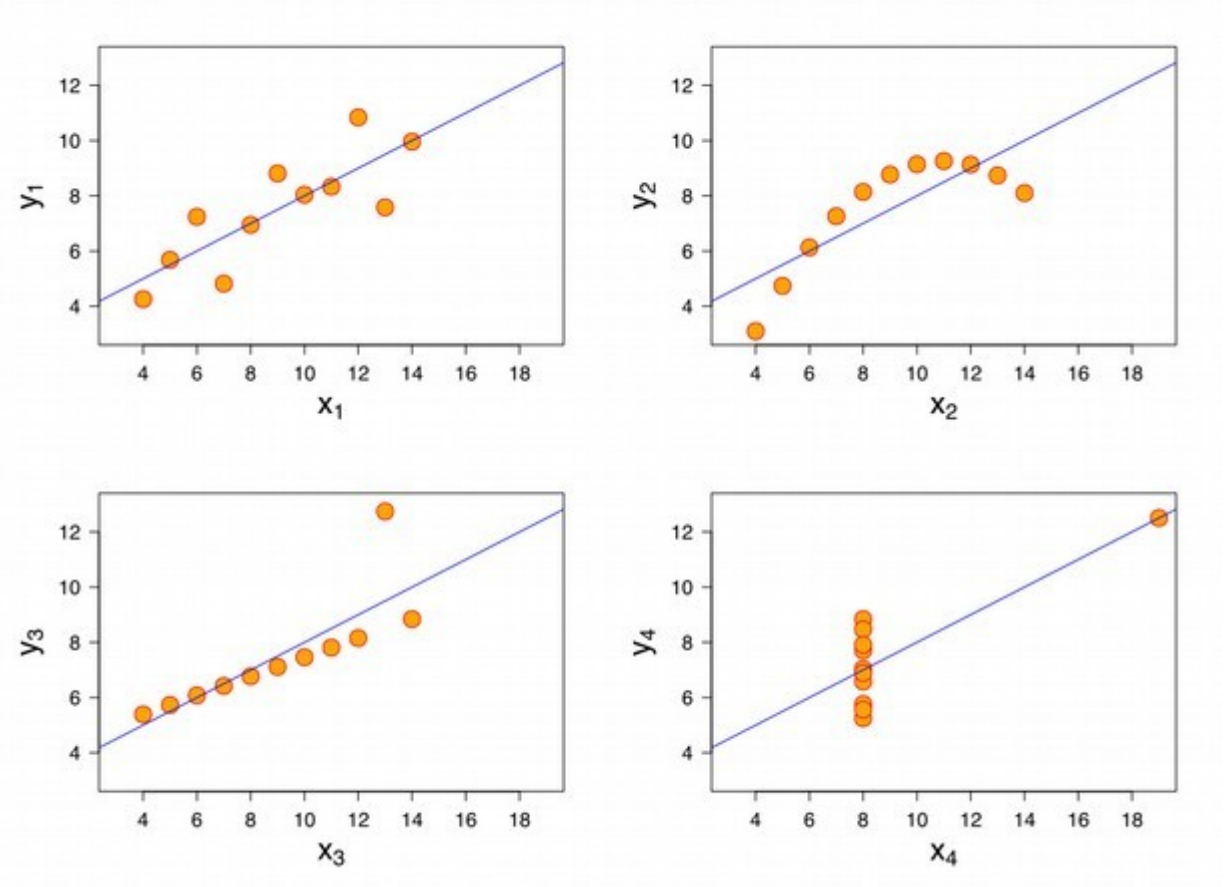
disassortative



assortative

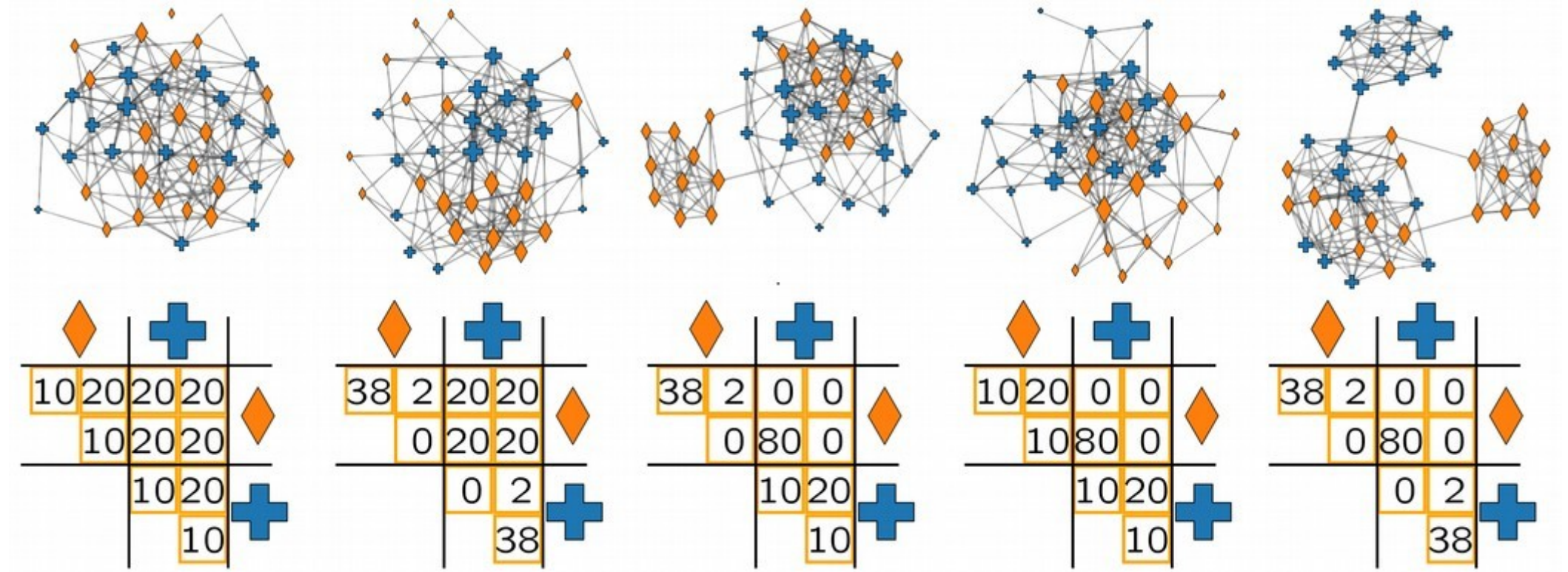
Newman “Mixing patterns in networks” Phys. Rev. E (2003)

Assortativity is correlation across edges



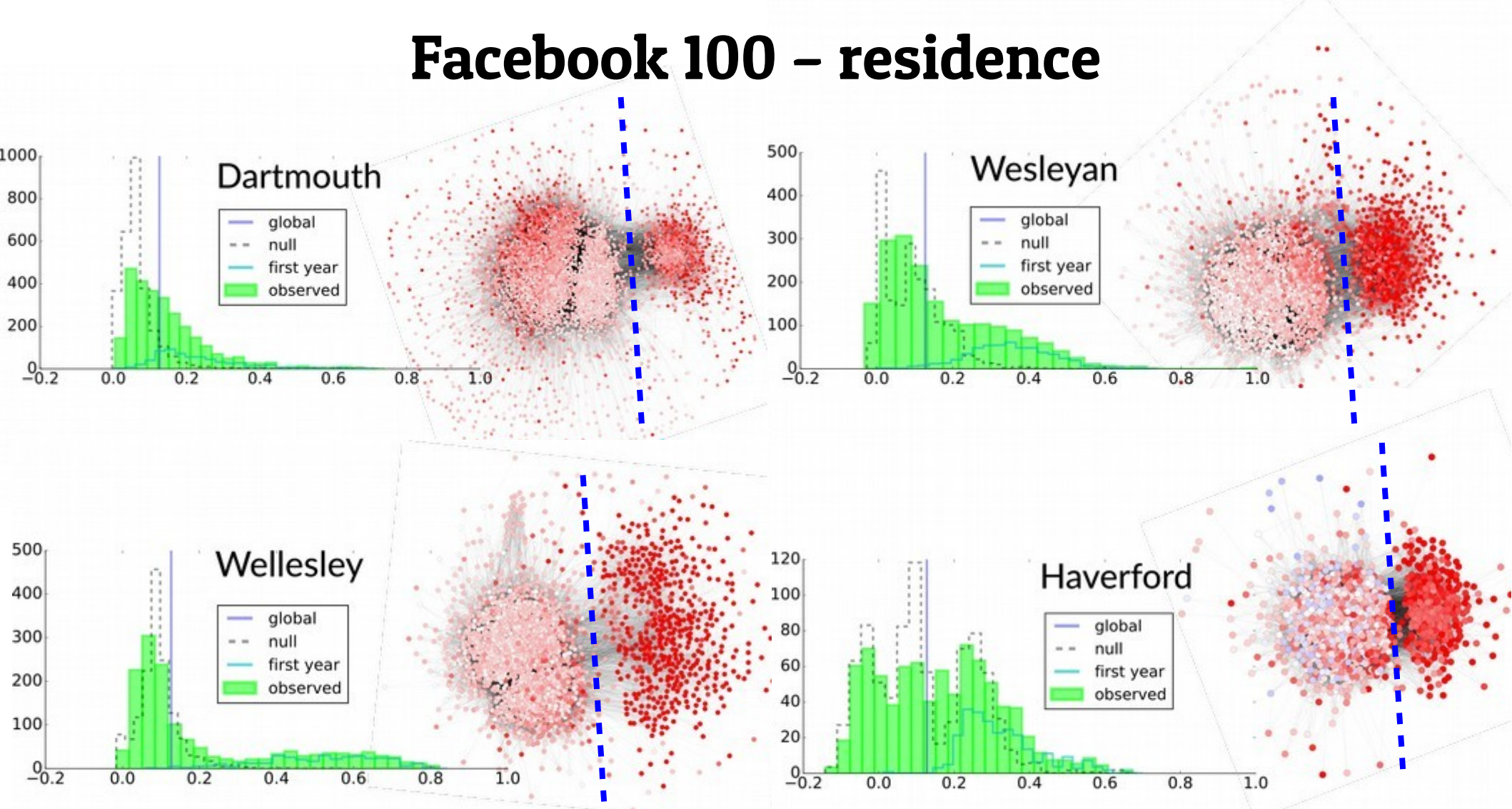
Anscombe, "Graphs in Statistical Analysis". American Statistician (1973)

All these networks have assortativity r=0



Peel, Delvenne, Lambiotte, "Multiscale mixing patterns in networks". ArXiv:1708.01236 (2017)

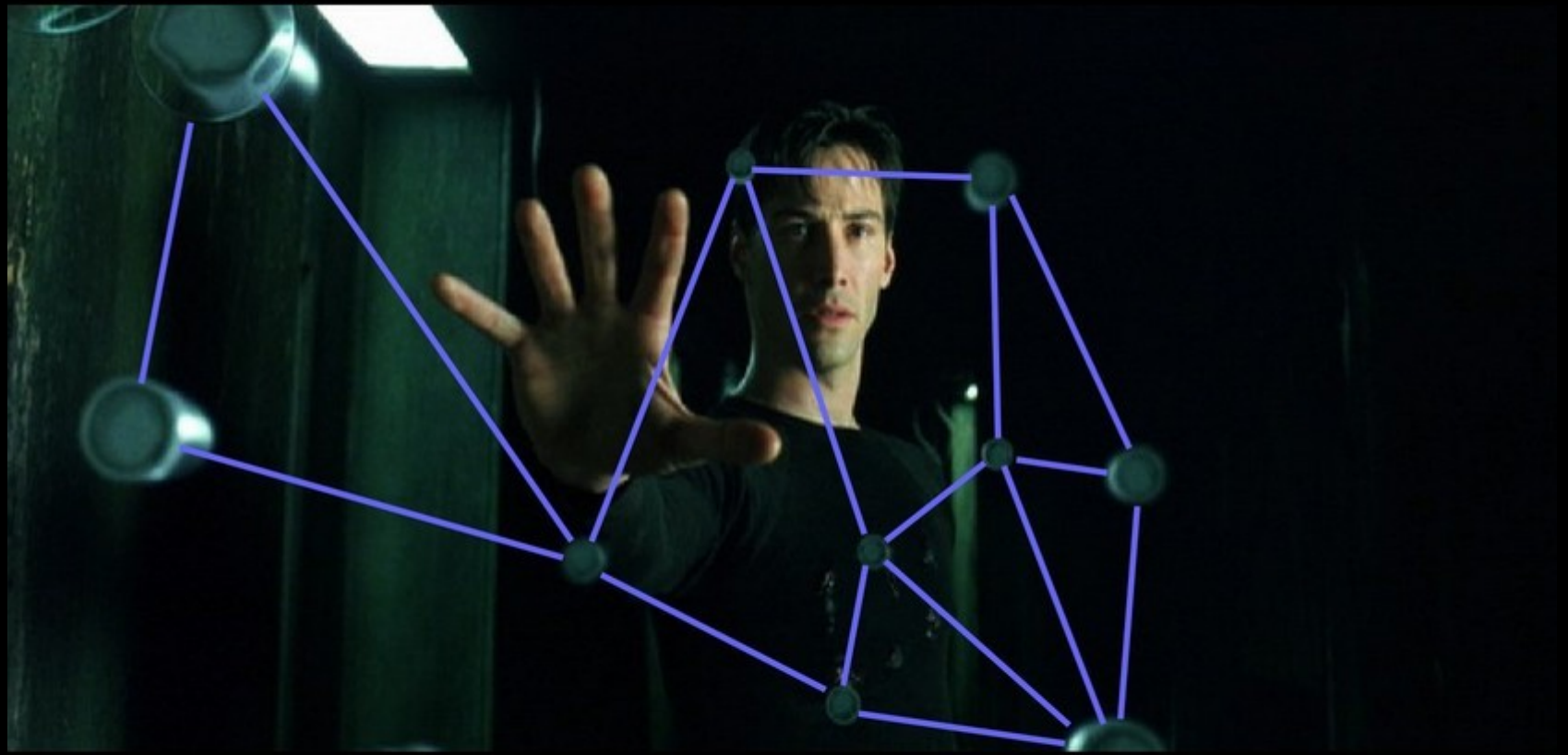
Facebook 100 – residence



Peel, Delvenne, Lambiotte, "Multiscale mixing patterns in networks". ArXiv:1708.01236 (2017)

Final thoughts...

Am I telling you that you can detect ground truth communities?



No, I'm telling you that when you're ready, you won't have to.

Final thoughts...

The relationship between network structure and metadata is important:

- Determine if the relationship is significant
- Predict missing values
- Understand how assortativity varies over a network

References & collaborators...

Peel, L., Topological Feature Based Classification 14th International Conference on Information Fusion (FUSION) 2011

Peel, L., Supervised Blockmodelling ECML/PKDD Workshop on Collective Learning and Inference on Structured Data (CoLISD) 2012

Peel, L., Active Discovery of Network Roles for Predicting the Classes of Network Nodes Journal of Complex Networks 3 (3): 431-449, 2015

Peel, L., Graph-based semi-supervised learning for relational networks SIAM International Conference on Data Mining (SDM) 2017

Daniel B. Larremore

Aaron Clauset

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†]

Jean-Charles Delvenne

Renaud Lambiotte

Multiscale mixing patterns in networks

Leto Peel,[1,2,*] Jean-Charles Delvenne,[1,3,†] and Renaud Lambiotte[2,4,‡]

[1] *ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, B-1348, Belgium*
[2] *naXys, Universit de Namur, Namur, B-5000, Belgium*
[3] *CORE, Université catholique de Louvain, Louvain-la-Neuve, B-1348, Belgium*
[4] *Mathematical Institute, University of Oxford, Oxford, UK*

pre-print arXiv:1708.01236

Contact: leto.peel@uclouvain.be

@PiratePeel