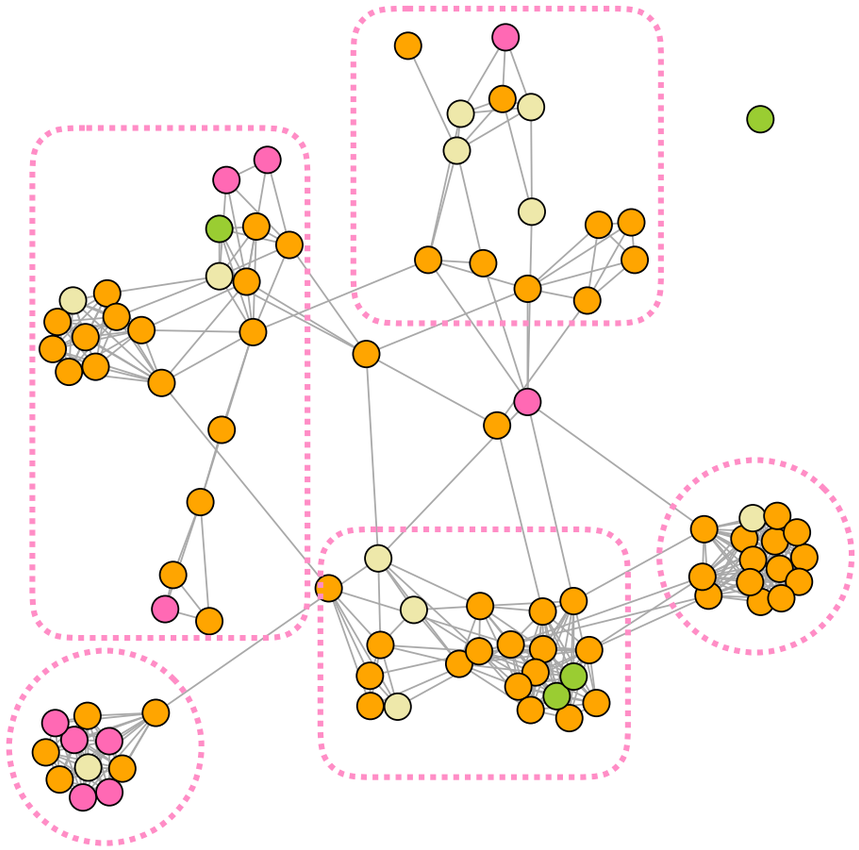
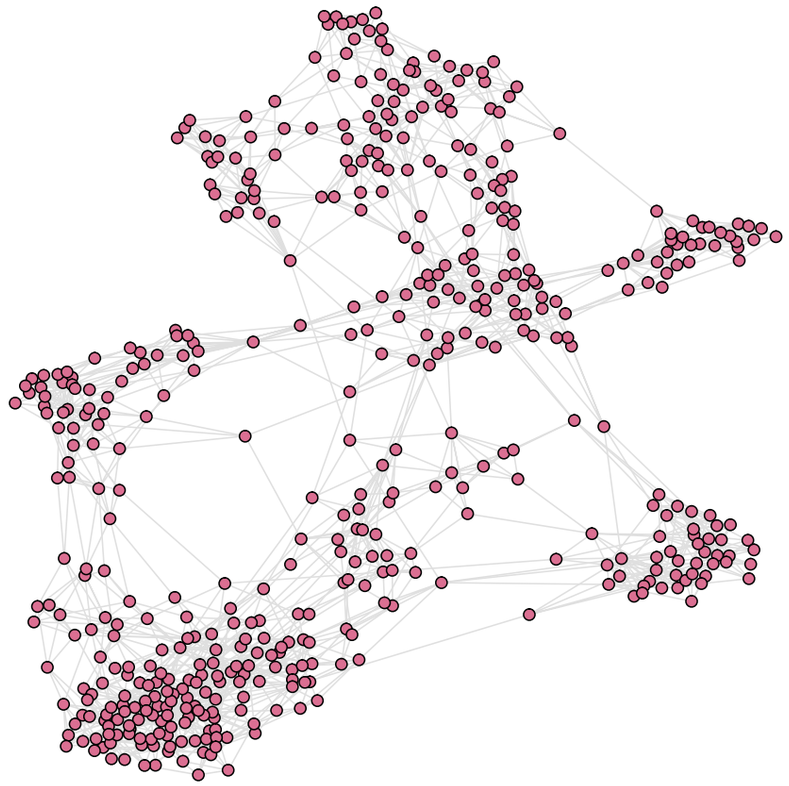




# Adapting Community Detection Approaches to Large, Attributed, and Multilayer Networks

Natalie Stanley
Advised by Professor Peter Mucha
March 19, 2018

# A quick intro to networks

---

A network is built of nodes and edges and encodes relational information between a set of objects.

**Nodes:** Authors

**Edges:** Whether the authors have written a paper together.

# Networks in Biology

## Microbiome

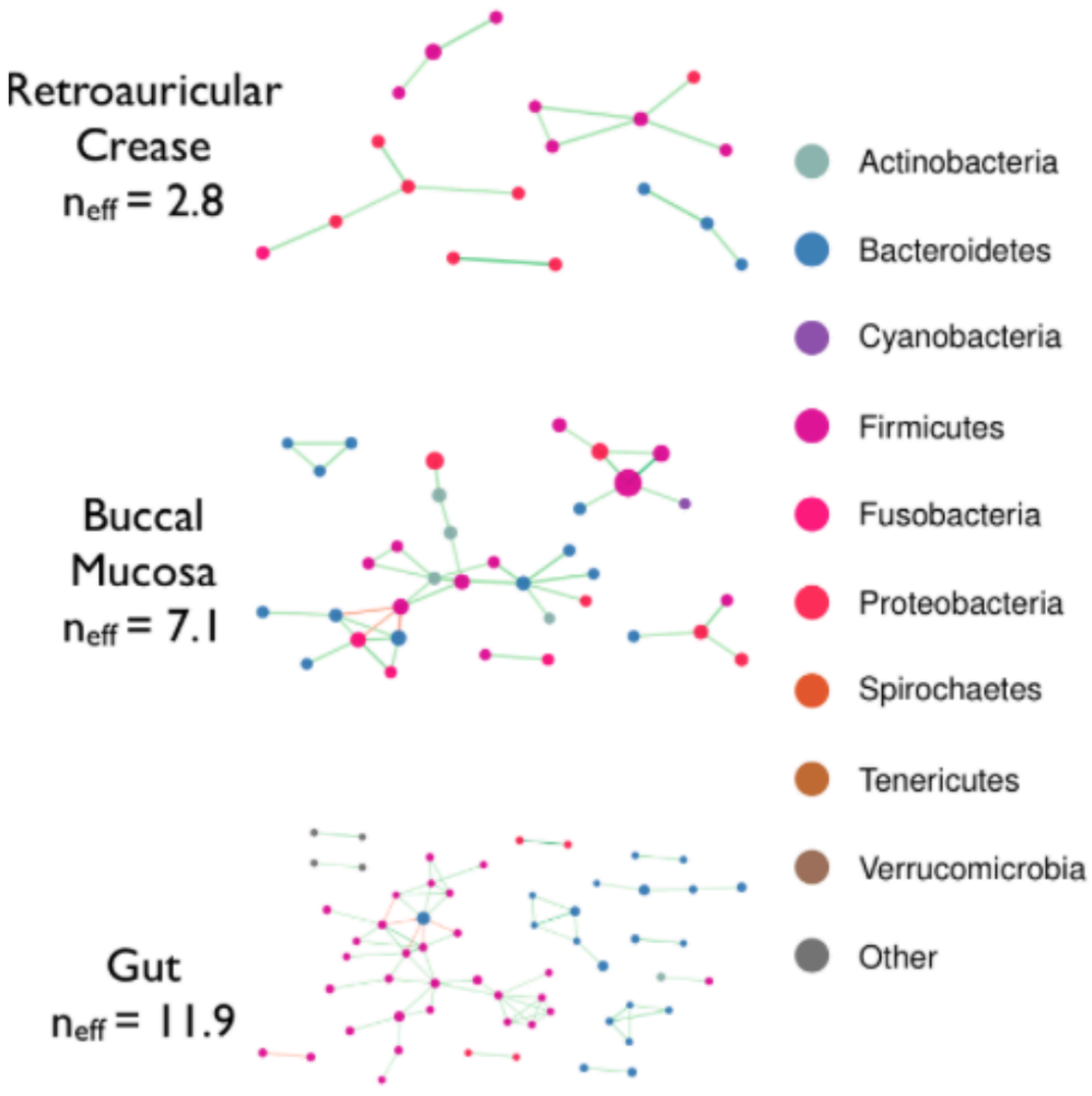



***Image from**: Inferring Correlation Networks from Genomic Survey Data. Friedman et al. 2012.*

## Protein Interaction Network

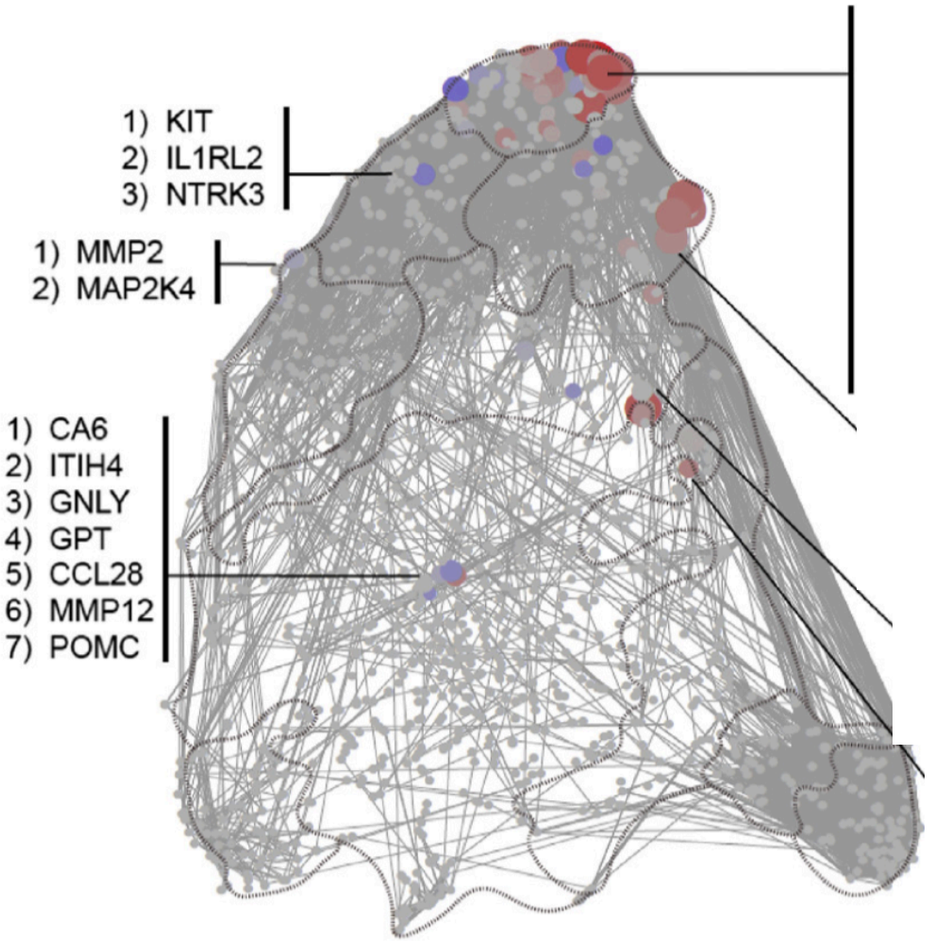
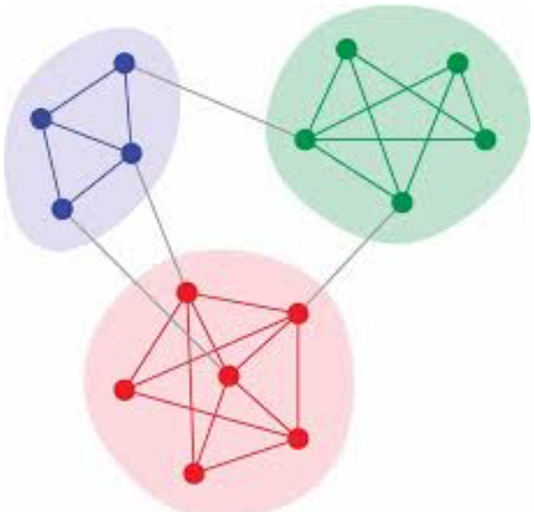



***Image from:** A Proteomic Clock of Human Pregnancy. Aghaeepour et al., 2017.*

# Network Analysis



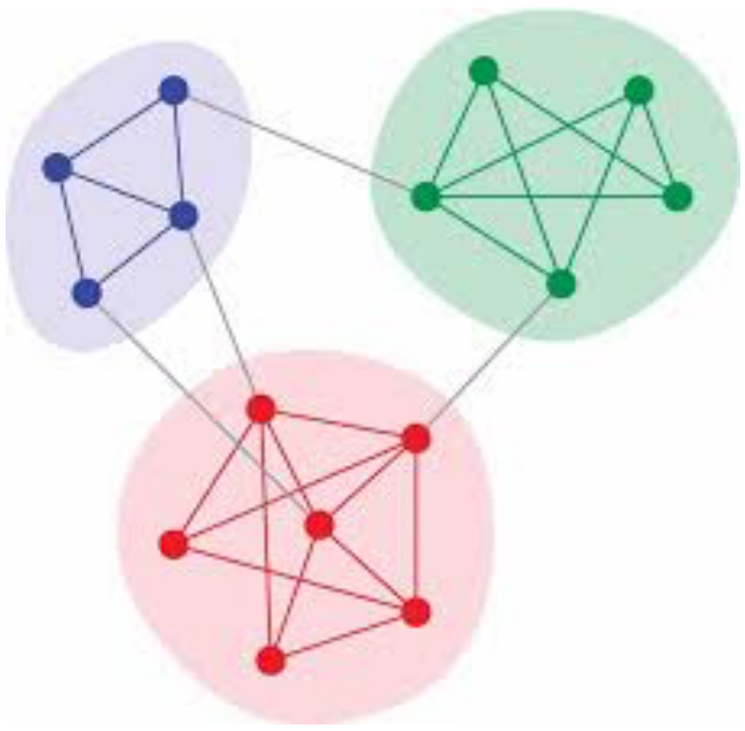
**Analysis of organizational patterns**

**Community Detection**

# Community Detection

---

**Objective:** Given a network with *N* nodes, we seek to partition these *N* nodes into *K* communities, such that members of a community interact with other communities in a stochastically equivalent way.



*You can think of community detection as clustering on a network*

Each colored set of nodes is a community.

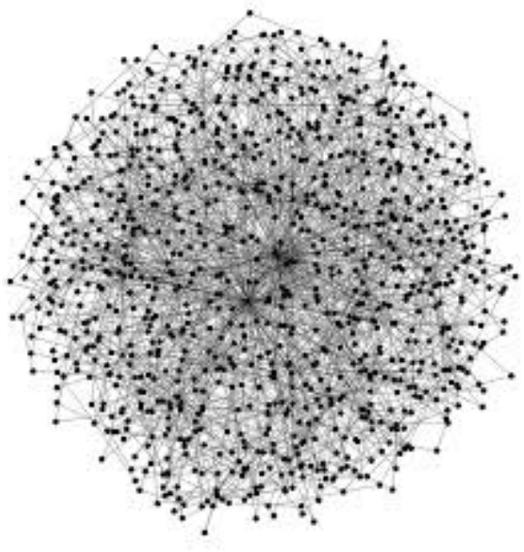
# Pick Your optimization problem

- Depending on what you like to do, you can cast this problem in a way that you like.
  - Probabilistic (solved through likelihood optimization) 
  - Quality function (write an objective and use heuristic) 
  - Spectral (use spectral properties of network adjacency matrix and graph Laplacian)
  - Deep learning 
  - k-Means (a more controversial suggestion- will not always work)

# Quality Function: Modularity

---

Intuition: Find the partition that is maximally different from the null model.



Null model and structureless

**Edge probability under null model based on # of neighbors**

**Edge: is it there?**

**Indicator for whether nodes i and j are in the same community**

$$Q = \frac{1}{2M} \sum_{i,j} [a_{ij} - \gamma \frac{k_i k_j}{2M}] \delta(z_i, z_j)$$

This has a straight-forward weighted analog

# A heuristic for maximizing modularity

---

## Louvain Algorithm

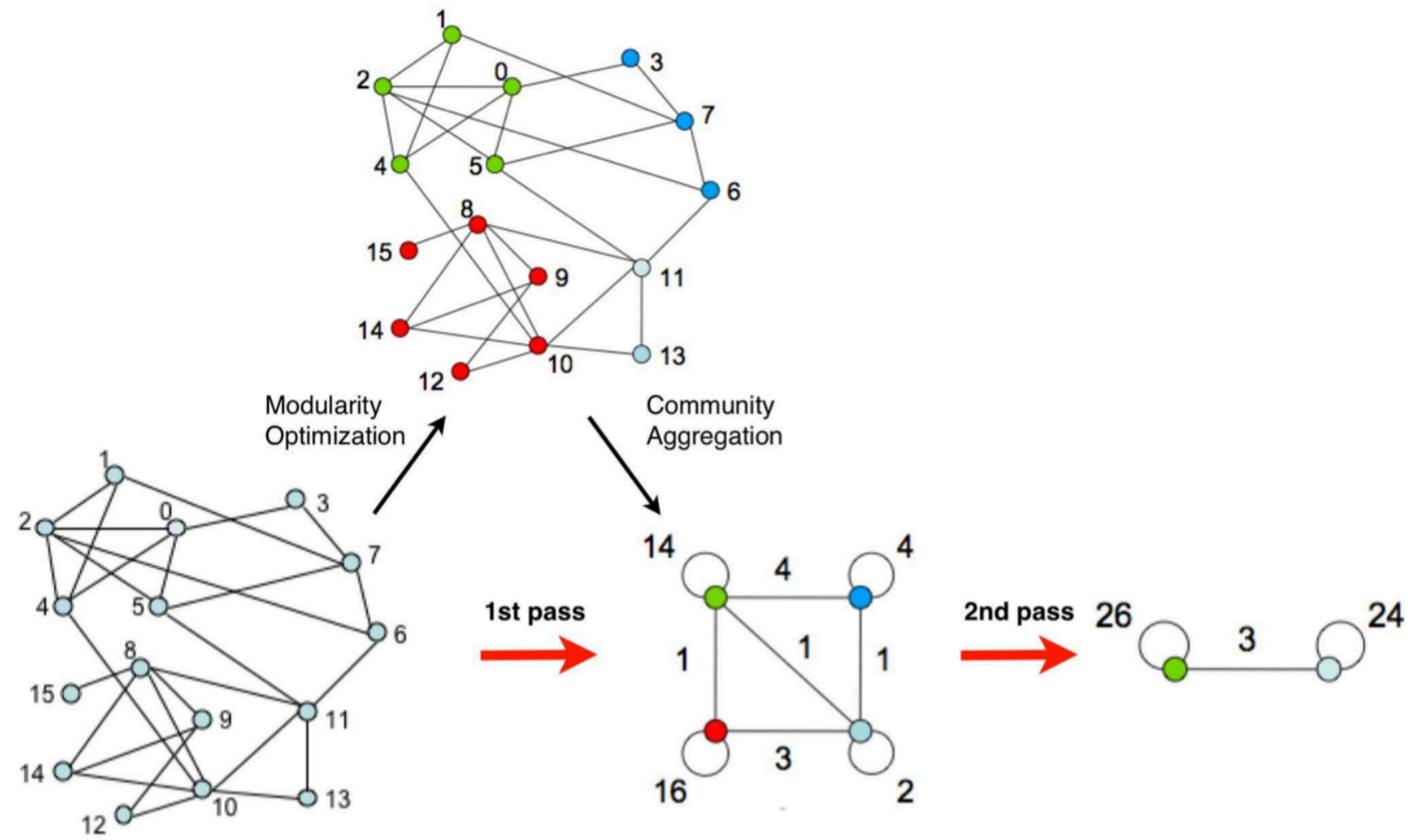


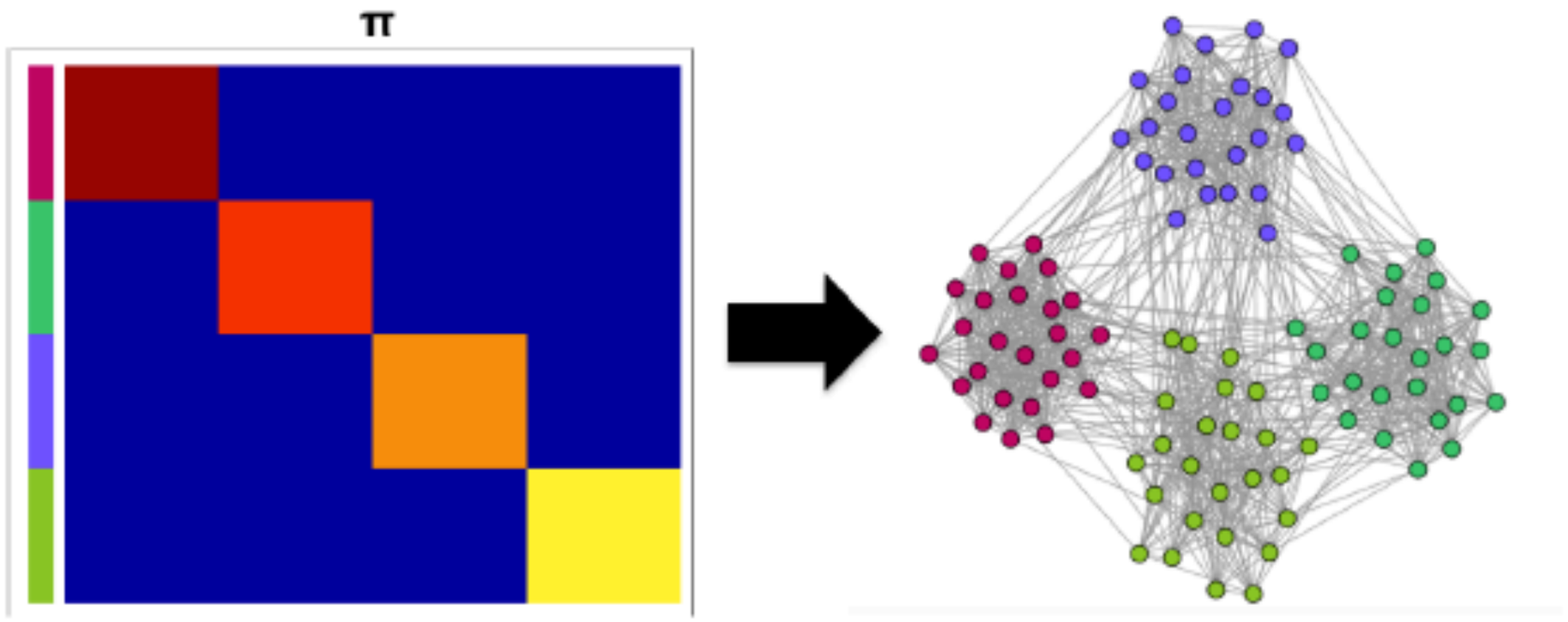

**Image From:** *Fast unfolding of communities in large networks. Blondel et al., 2008*

# Probabilistic: Stochastic Block Model

Intuition: Pairwise node connectivity is modeled based on the community assignments of the edge “stubs”.

Learning Objective:

- Node to community assignments
- SBM probability matrix




High connectivity probability

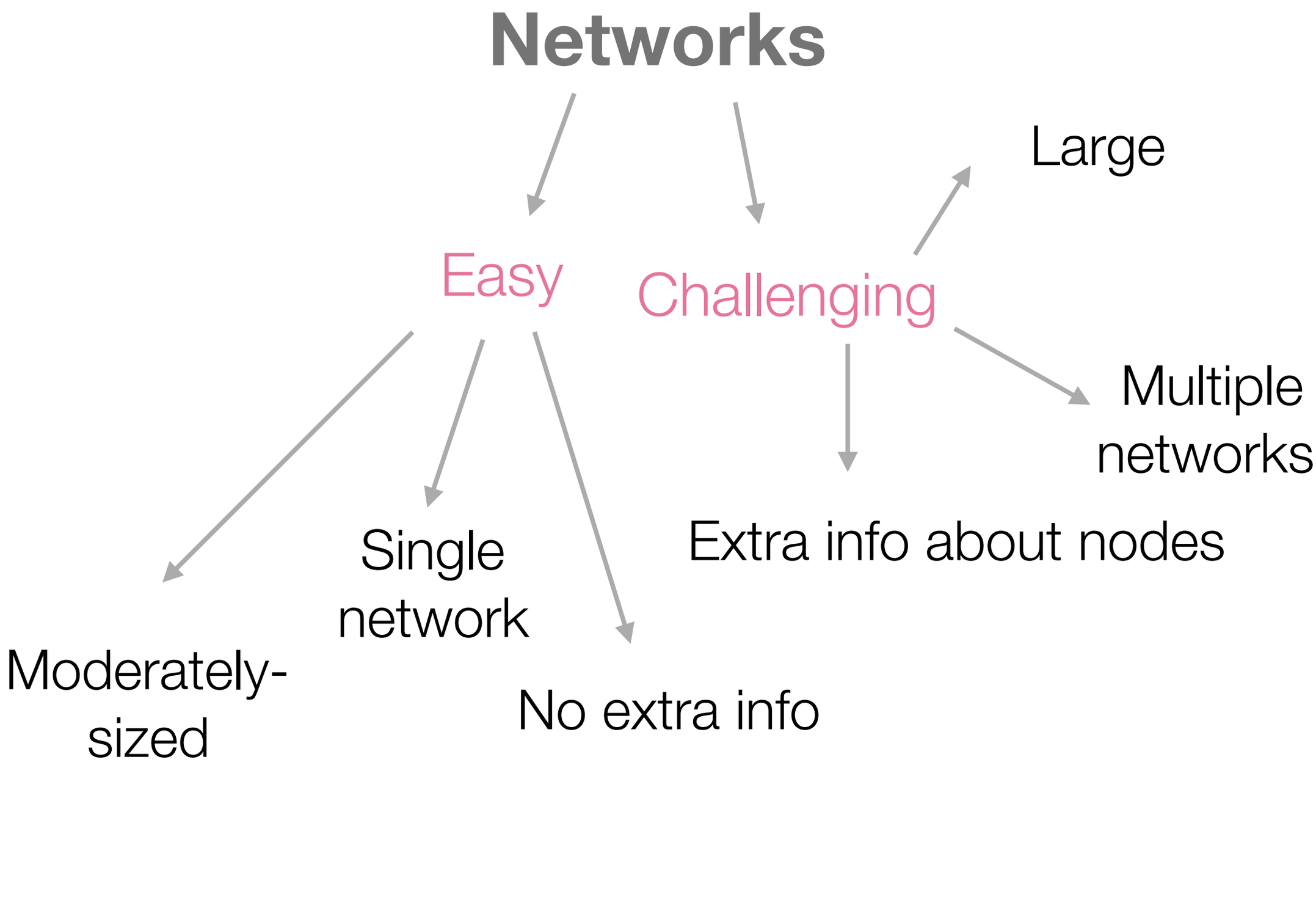
Low connectivity probability

# Fit an SBM by maximizing likelihood

$$\log(P(\mathbf{A}, \mathbf{Z})) = \sum_{i \neq j} \sum_{k<l} Z_{ik} Z_{il} [a_{ij} \log(\pi_{kl}) + (1 - a_{ij}) \log(1 - \pi_{kl})]$$

Find optimal value of parameters using EM and a factorized approximation of the posterior P(Z|X)

# Addressing the limitations of community detection



# Overview of community detection in challenging network types

---

- **Large**: When the network has > 10,000 nodes, community detection results become **inconsistent within and between algorithms, prohibitively slow, and uninterpretable.**

- **Multilayer:** When there exists multiple relational definitions between a set of nodes. How do you use the **connectivity collectively to identify communities?**

- **Attributed:** How do we **jointly integrate connectivity and node attributes/classifications** to assign nodes to communities?

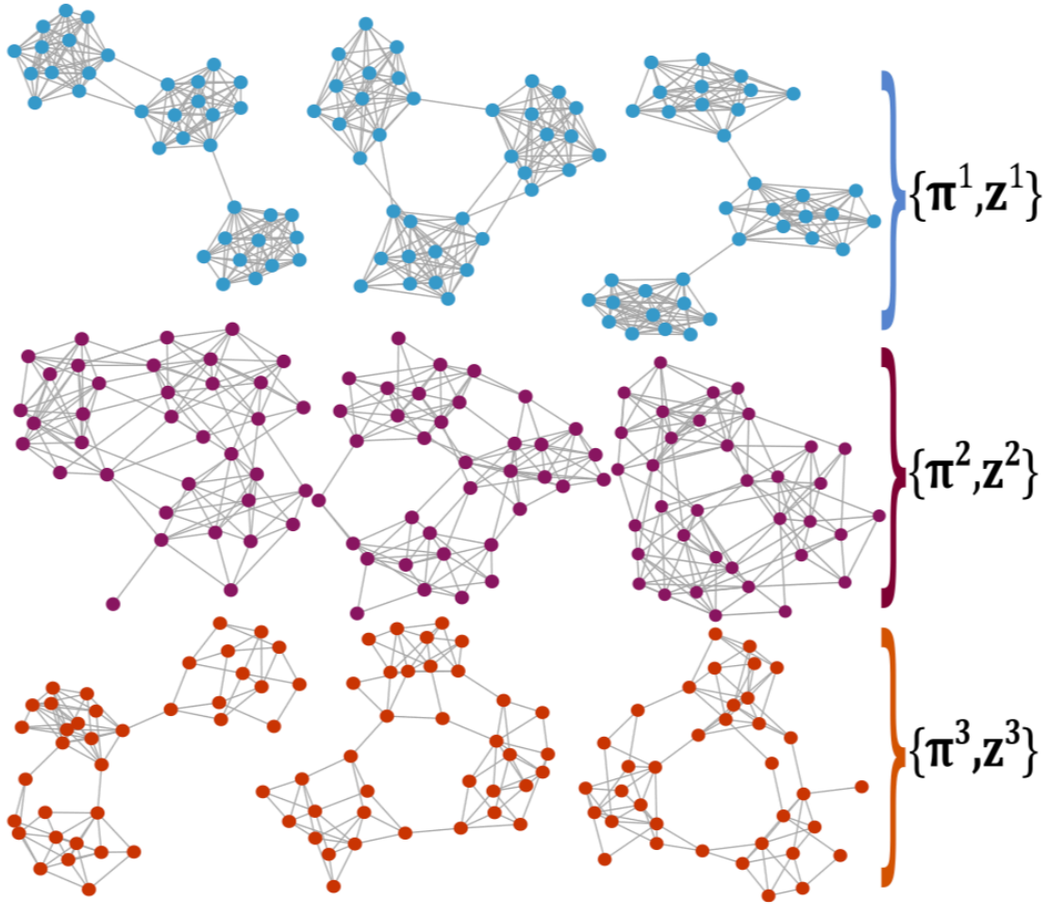
# Talk Outline

---

- Multilayer \\ sMLSBM
  - Brief overview
  - Application
- Large \\ SuperNodes
  - Objective & Method Overview
  - One result
- Attributed networks
  - Attribute SBM
  - Attribute Align

# Thesis Contribution 1: sMLSBM

- We developed an extension to the stochastic block model for multilayer networks (sMLSBM). This method learns a collection of SBMs to describe multilayer network.

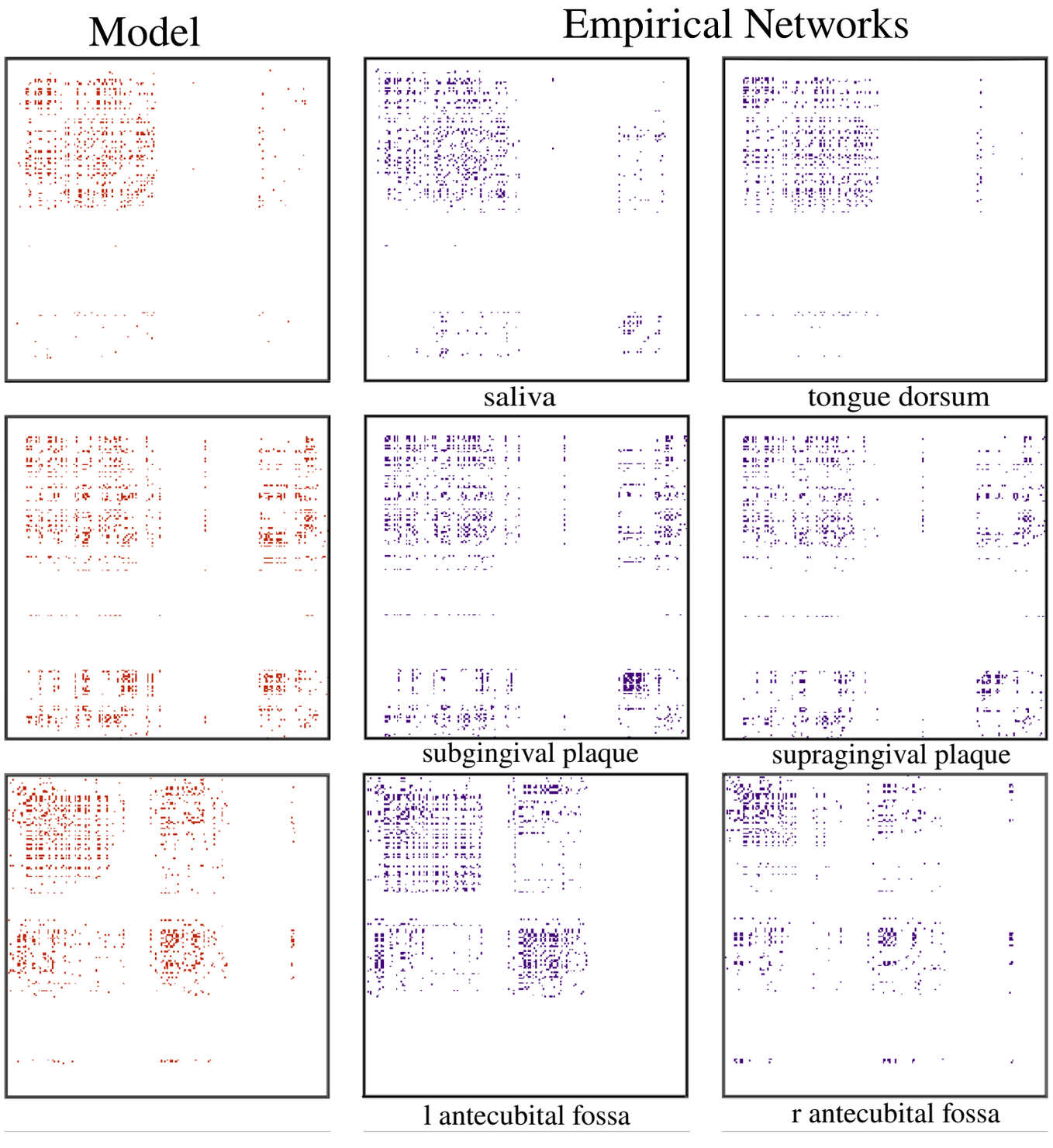



1.Clustering Network Layers With the Strata Multilayer Stochastic Block Model. N. Stanley, S. Shai, D. Taylor & P.J. Mucha. IEEE Transactions on Network Science and Engineering. 2016. paper

2.Enhanced detectability of community structure in multilayer networks through layer aggregation. D. Taylor, S Shai, N. Stanley, and P.J. Mucha.Physical Review Letters. 2016. paper

https://github.com/stanleyn/sMLSBM

# Application: Human Microbiome Project

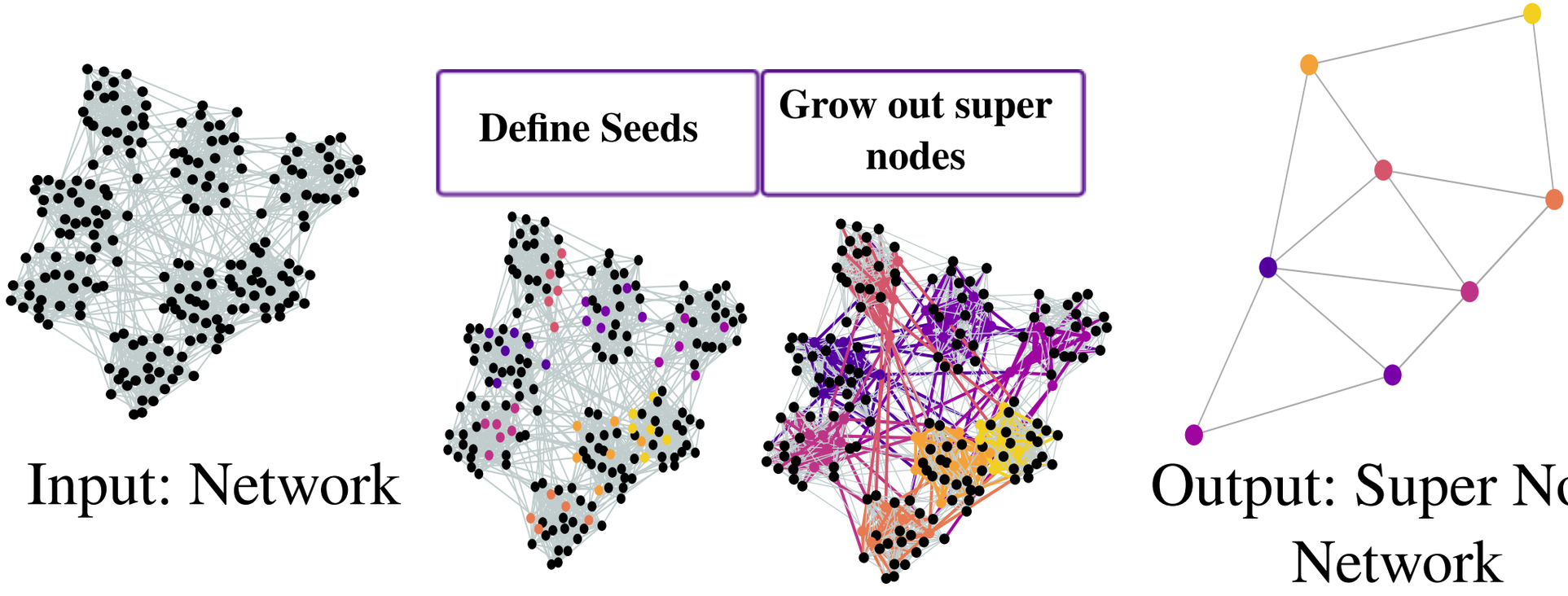



We learned the partition of these microbiome networks and the consensus SBM describing each cluster of networks.

# Thesis Contribution 2: Super Nodes

---

We developed an approach to pre-process a network into a smaller version before applying community detection. This resulted in less variability between partitions, faster run time, and higher label agreement within a graph region

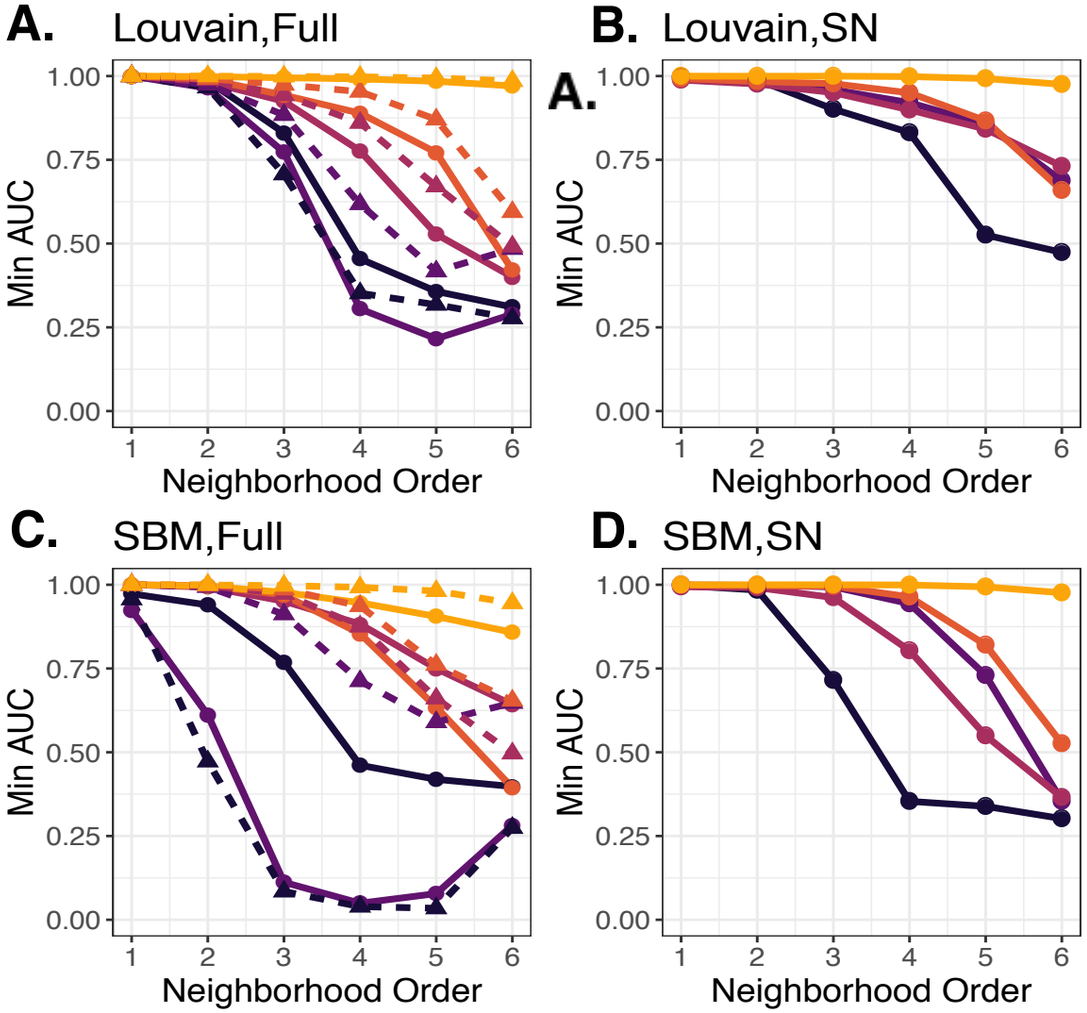


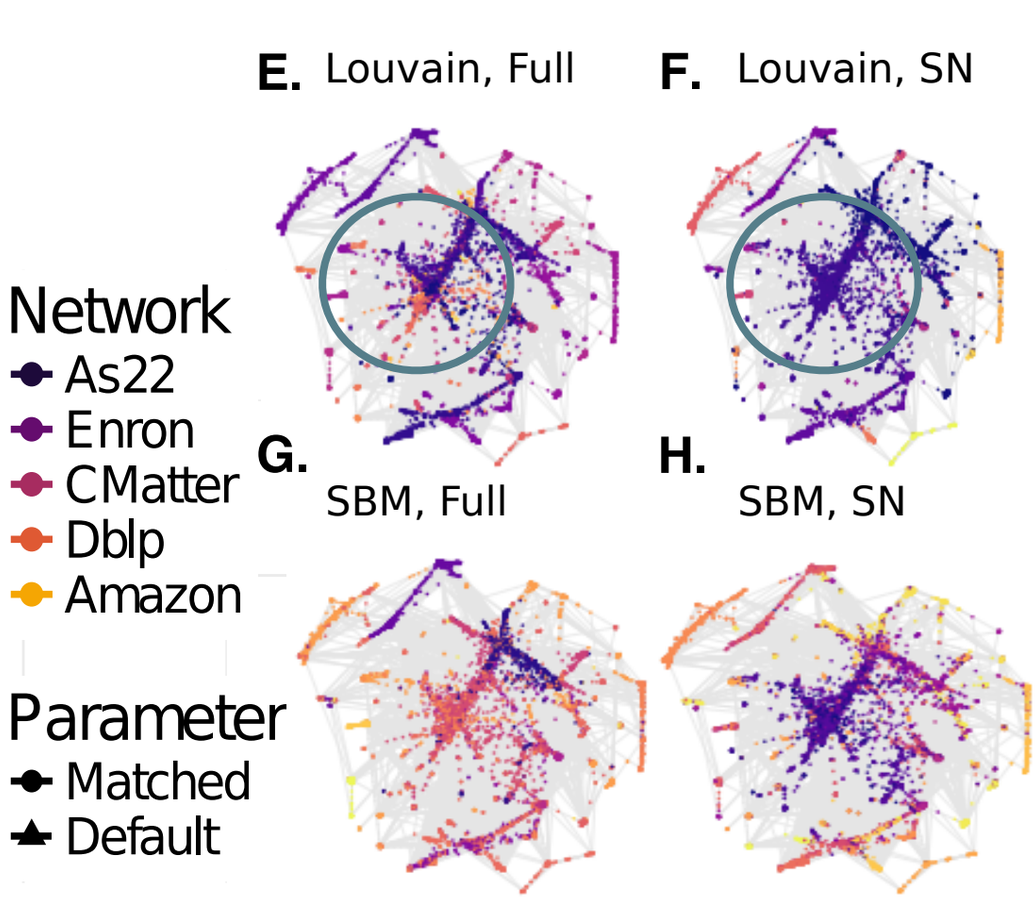

Compressing Networks with Super nodes. N. Stanley, R. Kwitt, M. Niethammer & P.J. Mucha. arxiv. Under Review

https://github.com/stanleyn/SuperNode

# Use of the pre-processed network leads to interpretable communities

Prediction Task: Estimate probability of a node belonging to each community based on neighbors. Generate ROC for each community by varying membership probability and report min AUC across communities



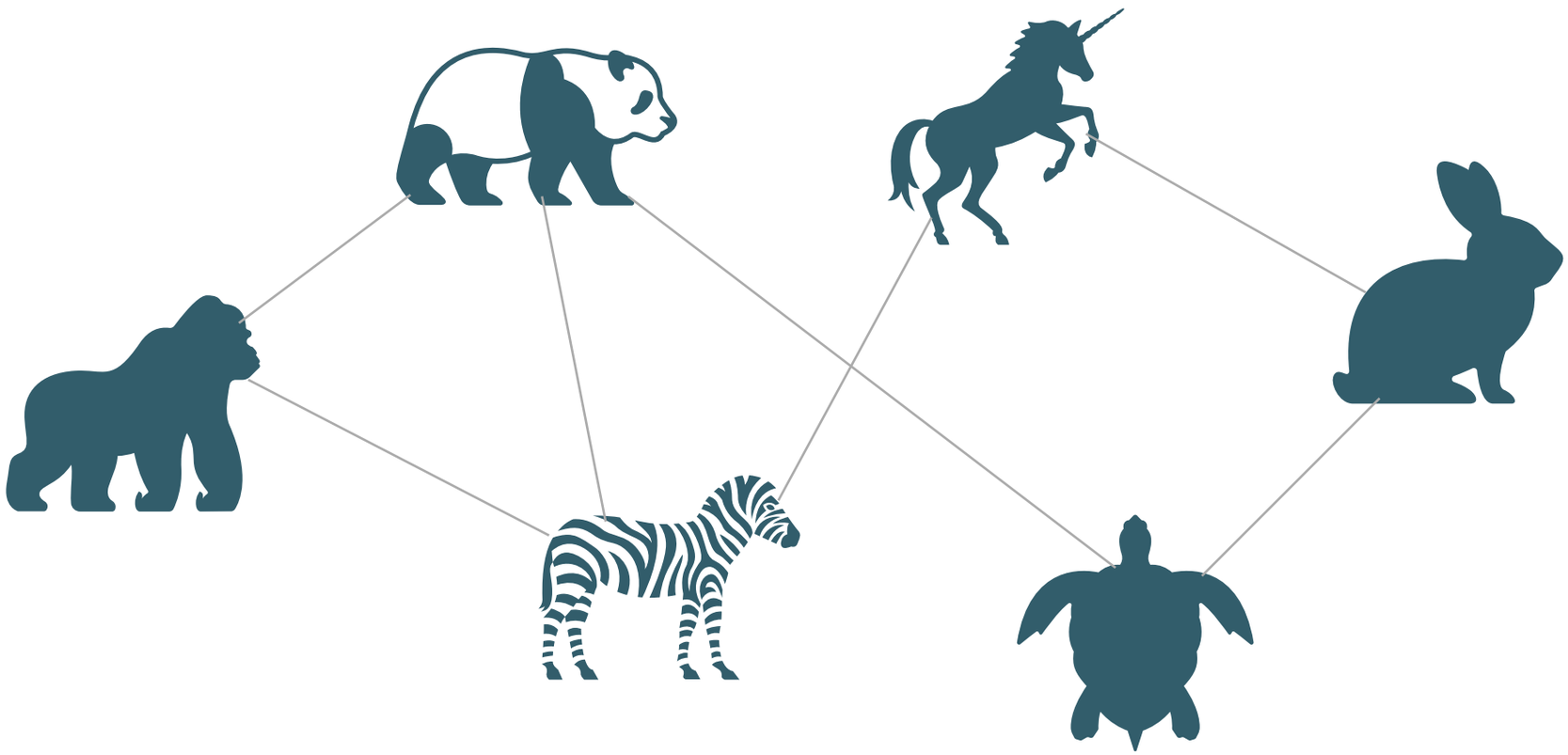



# Recap on sMLSBM and Super Nodes

- In sMLSBM, given a collection of multiple networks, we jointly learn node-to-community assignments and layer-to-strata assignments such that the likelihood over all layers is maximized. This is achieved through a modification of the stochastic block model.

- We can re-cast a large network into a smaller network of super nodes. Each super node contains one of more nodes of the original network. We showed how the super node representation leads to communities aligned with network structure.

# Attributed Networks



How do we integrate network connectivity and node attributes to best partition these animals?

| Eats leaves | Average life span | Average temperature for ideal life | Likes bananas |
|---|---|---|---|

# Our Objective

---

We seek to develop an attributed version of the stochastic block model that can incorporate multiple, continuous attributes in inferring node-to-community assignments.

Stochastic Block Models with Multiple Continuous Attributes. N. Stanley, T. Bonacci, R. Kwitt, M. Niethammer & P.J. Mucha. arxiv. Under Review.

https://github.com/stanleyn/AttributedSBM

# Related Work

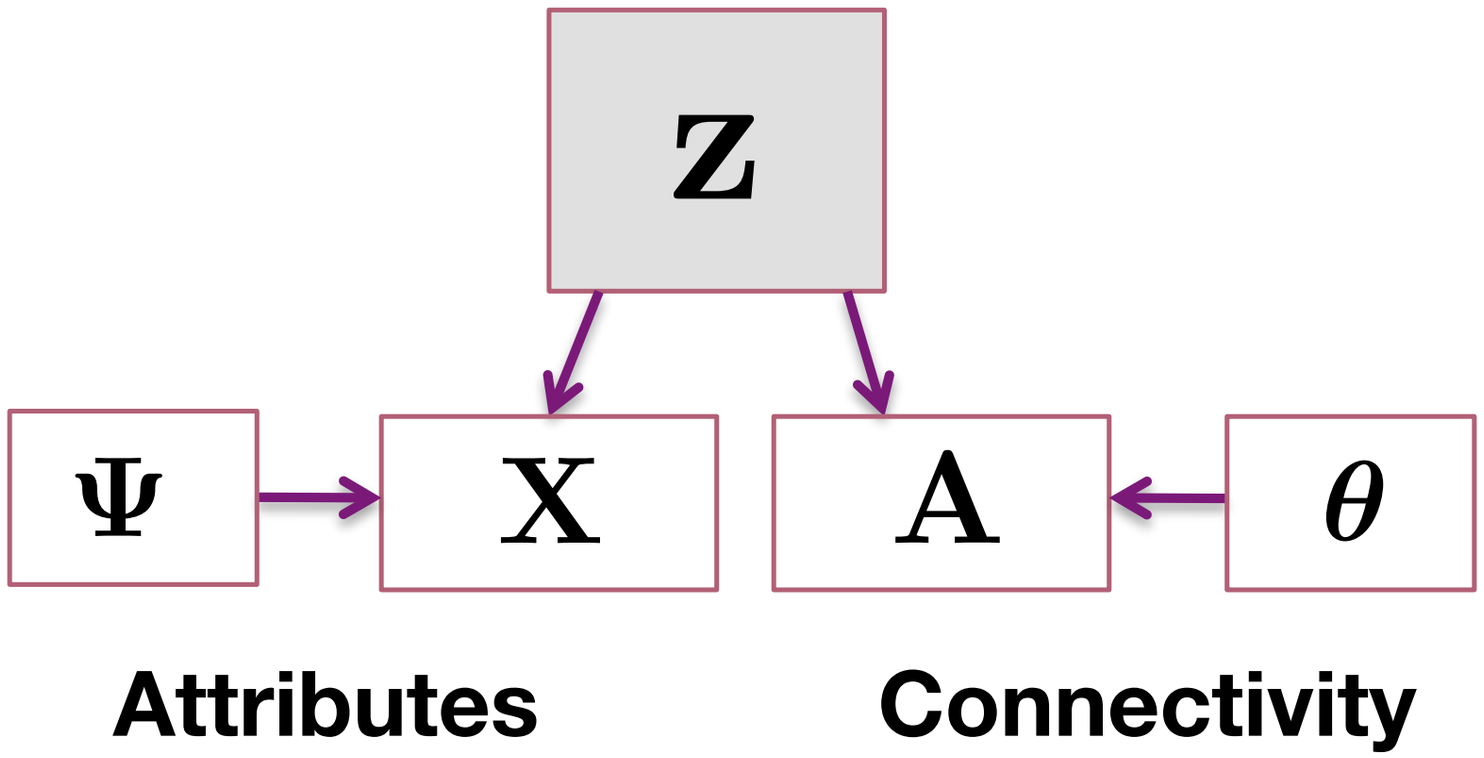
| | SBM | Multiple attributes | Attributes can be continuous |
|---|---|---|---|
| CESNA | | ✓ | |
| ILouvain | | ✓ | ✓ |
| Newman/ Clauset | ✓ | | ✓ |
| NeoSBM | ✓ | | |
| Our Attributed SBM | ✓ | ✓ | ✓ |

# Our attributed SBM

---

Connectivity— Stochastic Block Model
Attributes— Multivariate Gaussian for each community

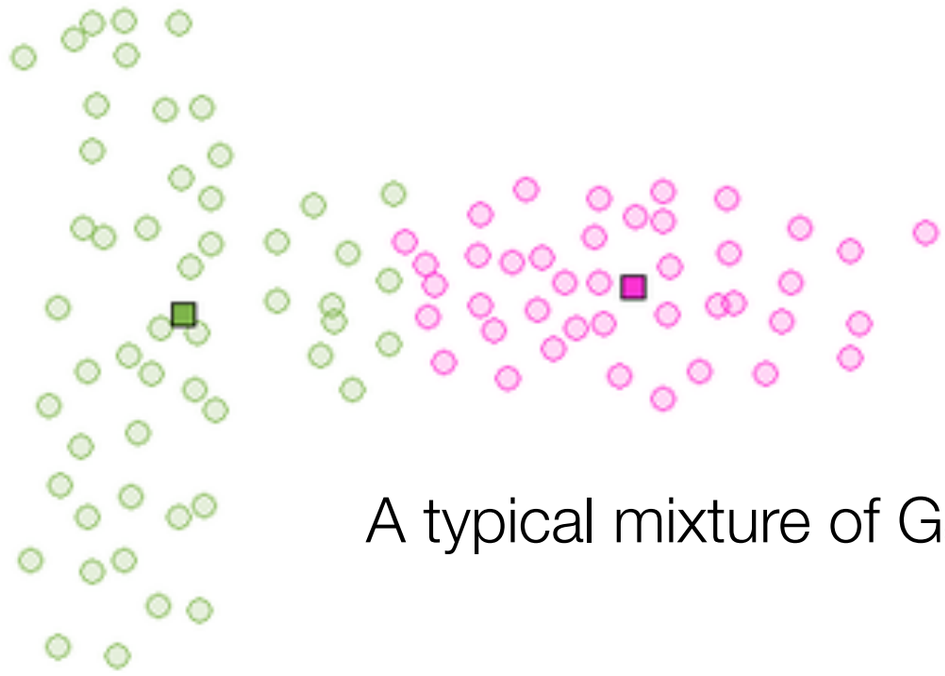
Attributes and connectivity are assumed to be conditionally independent, given community assignments ($\mathbf{z}$).



# Gaussian mixture model for the attributes

$$P(\mathbf{X} \mid \boldsymbol{\Psi}) = \sum_{i=1}^{N} \log\{\sum_{c=1}^{K} \pi_c \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\}$$

Collection of MV Gaussian parameters



A typical mixture of Gaussians

# Some of the details: EM algorithm

---

Because of conditional independence

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_X$$

Compute posterior in each E-step

$$\gamma(z_{ic}) = p(z_{ic} = 1 \mid \mathbf{x}_i, \mathbf{a}_i)$$

$$= \frac{p(\mathbf{x}_i \mid z_{ic} = 1) p(\mathbf{a}_i \mid z_{ic} = 1) \pi_c}{\sum_{c=1}^{K} p(\mathbf{x}_i \mid z_{ic} = 1) p(\mathbf{a}_i \mid z_{ic} = 1) \pi_c}$$

Updates in each M-step

SBM

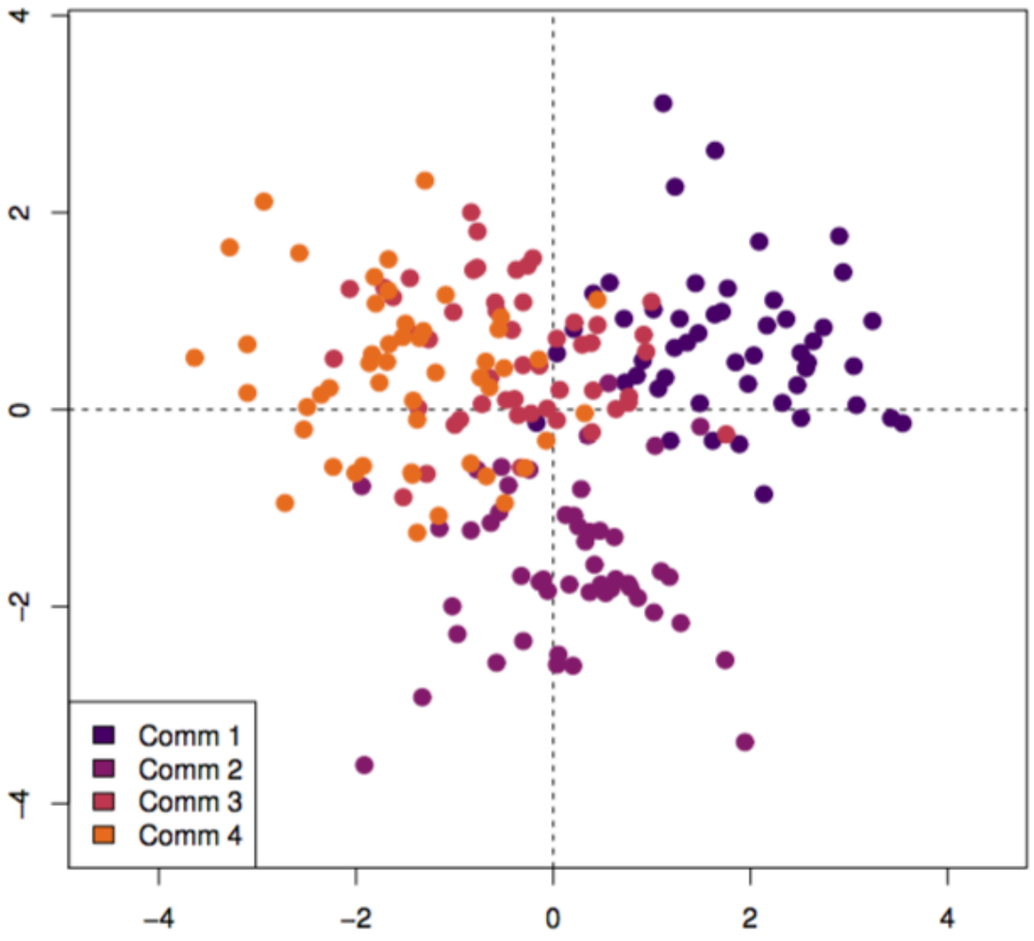
$$\boldsymbol{\theta}_{ql} = \frac{\sum_{i \neq j} \gamma(z_{iq}) \gamma(z_{jl}) x_{ij}}{\sum_{i \neq j} \gamma(z_{iq}) \gamma(z_{jl})}$$

MV Gauss in community *c*

$$\boldsymbol{\Sigma}_c = \frac{\sum_{i=1}^{N} \gamma(z_{ic})(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T}{\sum_{i=1}^{N} \gamma(z_{ic})} \quad \boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N} \gamma(z_{ic}) \mathbf{x}_i}{\sum_{i=1}^{N} \gamma(z_{ic})}$$

# Detectability: How do attributes influence the ability to identify a community?

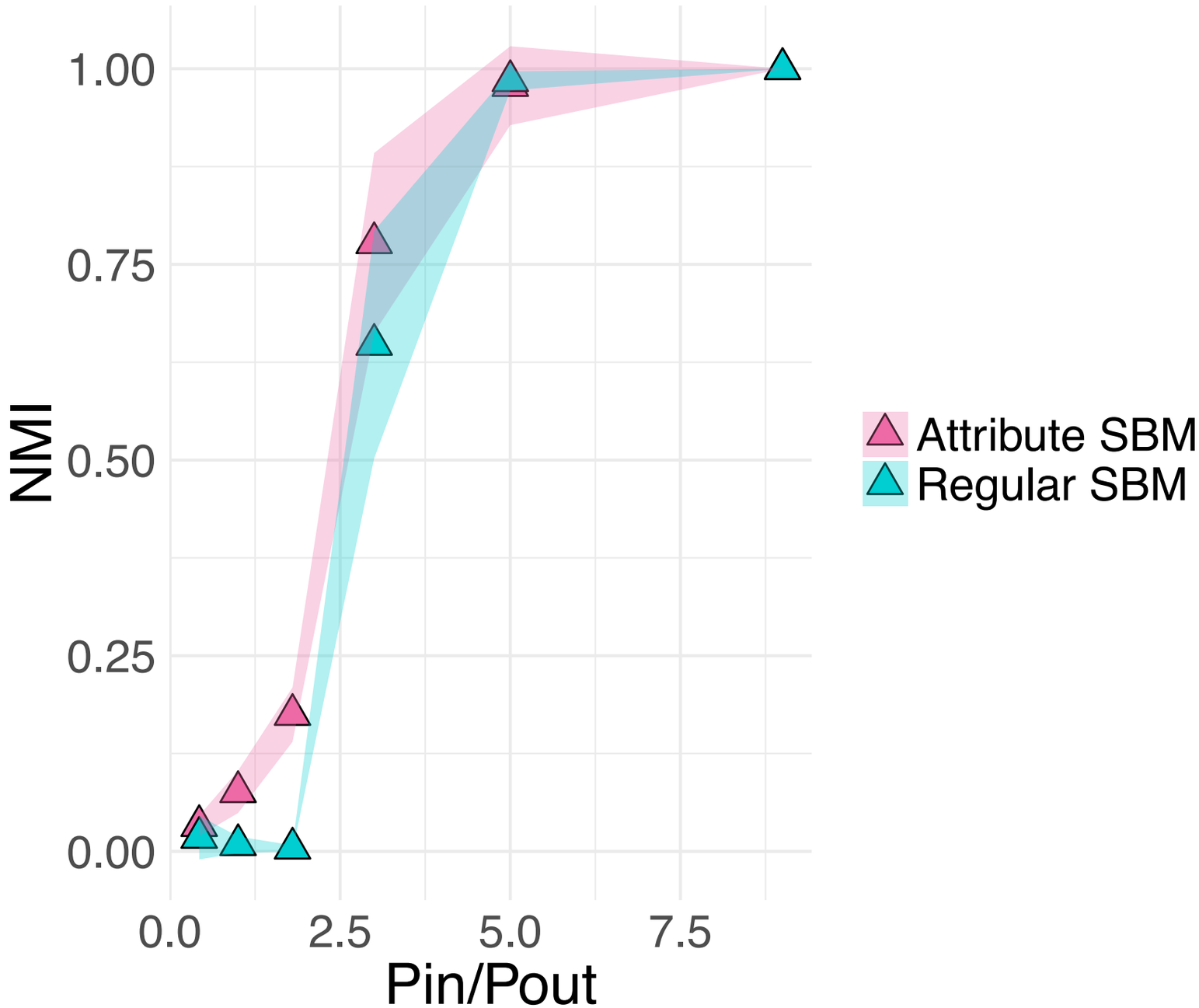
2-dimensional PCA projection of attributes




- **Experiment Varying Connectivity**: Vary the ratio of within-community (pin) to between-community probability (pout)

- Vary pin between 0.05 and 0.3 and fix mean degree.

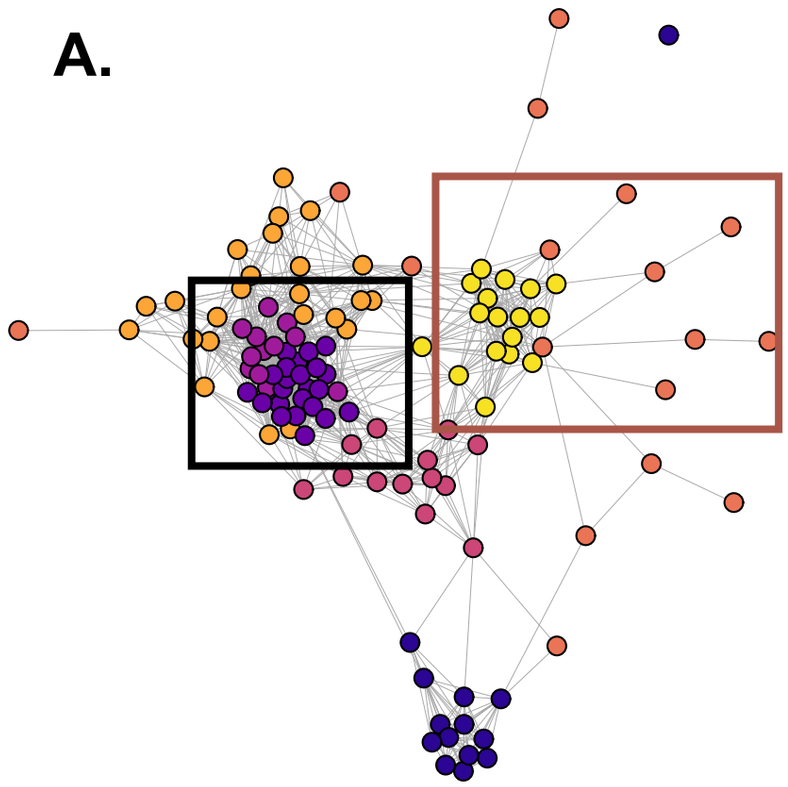
# Incorporating attributes moves the detectability limit



# Validation on biological networks (I of II)

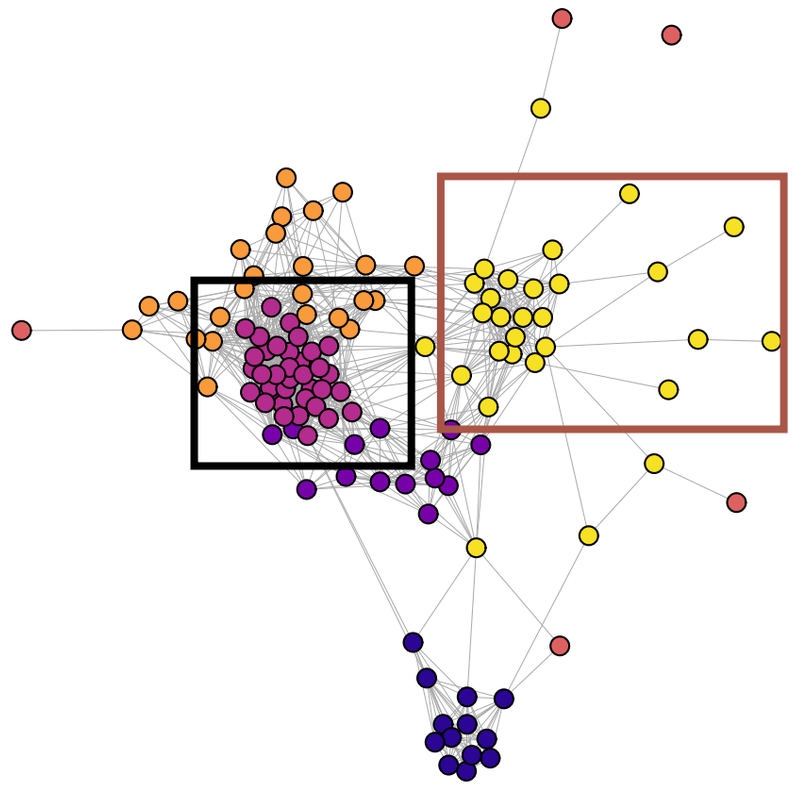
## Microbiome Subject Similarity Networks

**A.**



Nodes colored by SBM communities

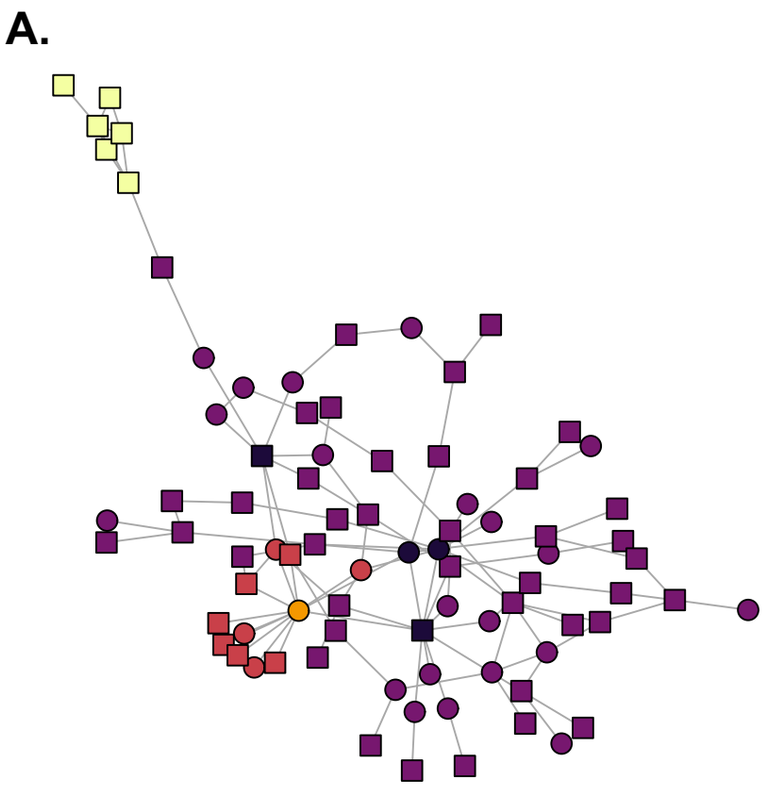
**B.**



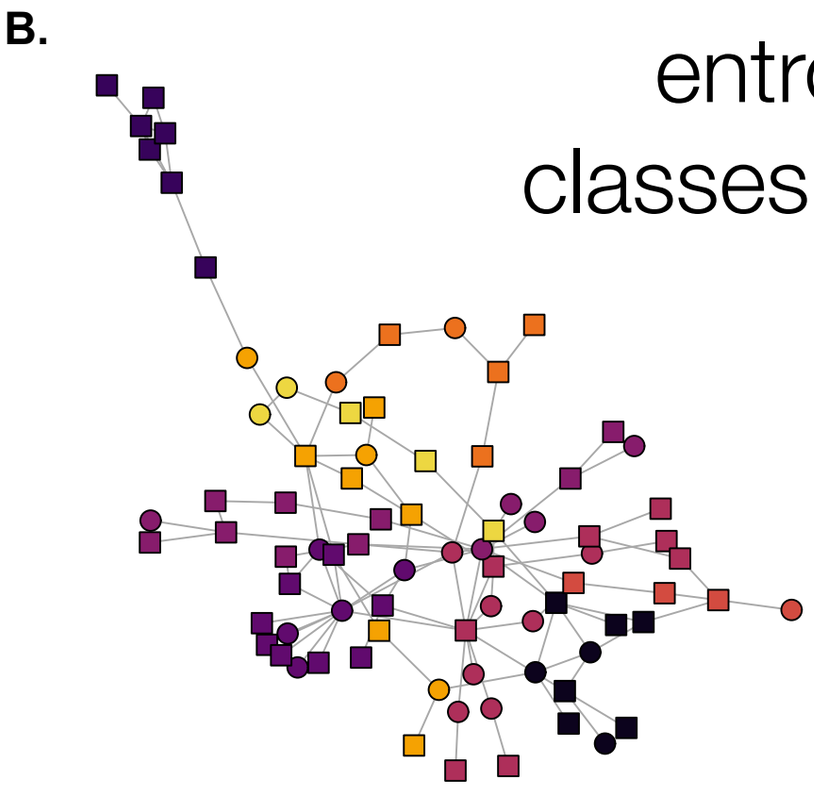
Nodes colored by attribute SBM

# Validation on biological networks (II of II)

## Protein Interaction Network



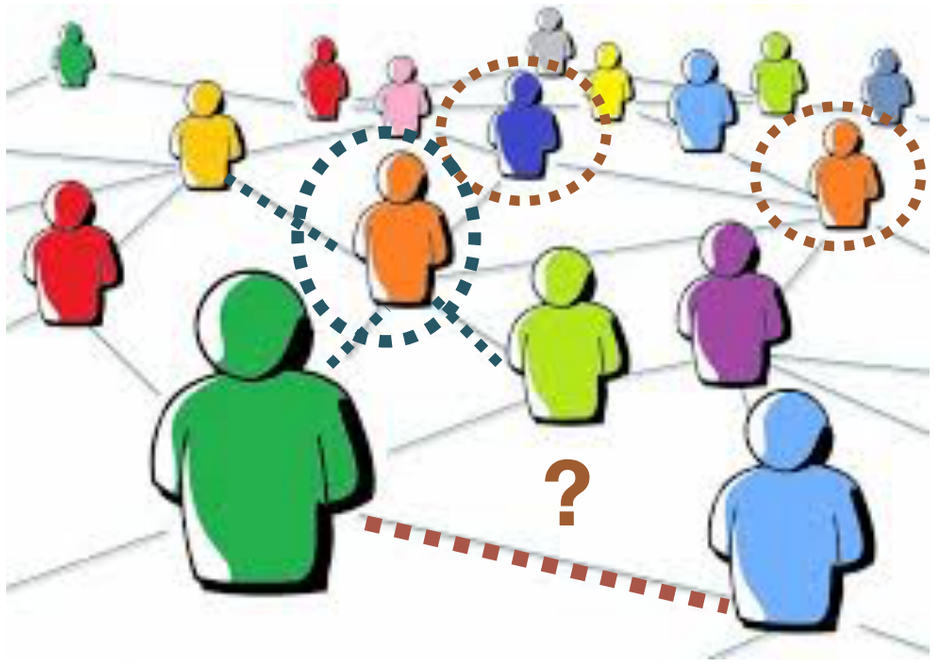

Nodes colored by SBM communities




Nodes colored by attribute SBM

Attributes correspond to functional classes. We show decreased entropy of these classes in communities

# Predicting complementary sources of information with the fitted model




1. Predict a link ..... between a pair based on learned model based on the community assignments of the closest neighbors in attribute space ◯ ◯
2. Predict an attribute ◯ based on neighbors in connectivity ..... ..... .....

# Validation with link prediction

---

For a pair of nodes, we seek to predict whether an edge exists between them.

**Jaccard**: $\text{Score}(m,n) = \frac{\Gamma(m) \cap \Gamma(n)}{\Gamma(m) \cup \Gamma(n)}$

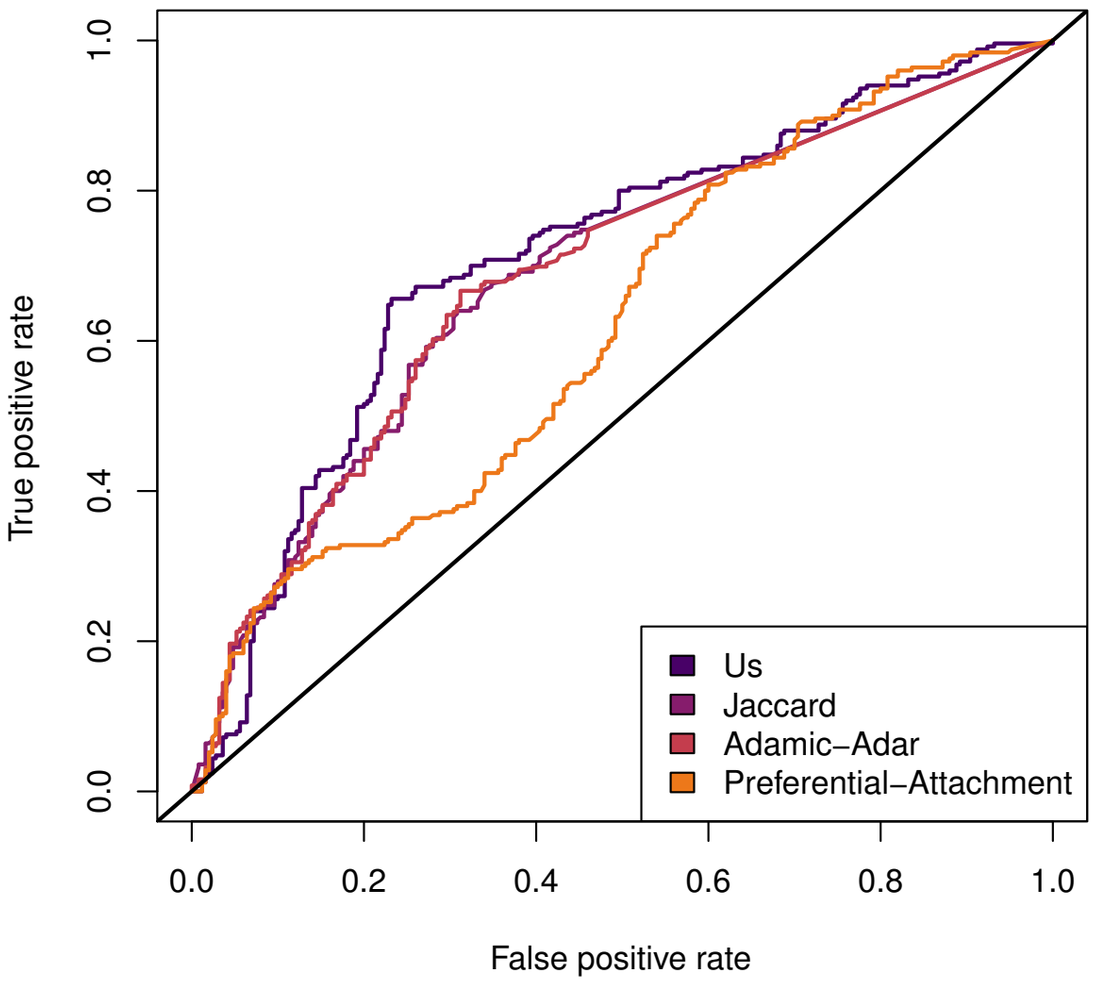
**Adamic Adar**: $\text{Score}(m,n) = \sum_{c \in \Gamma(m) \cap \Gamma(n)} \frac{1}{\log |\Gamma(c)|}$

**Preferential Attachment**: $\text{Score}(m,n) = |\Gamma(m)| \times |\Gamma(n)|$

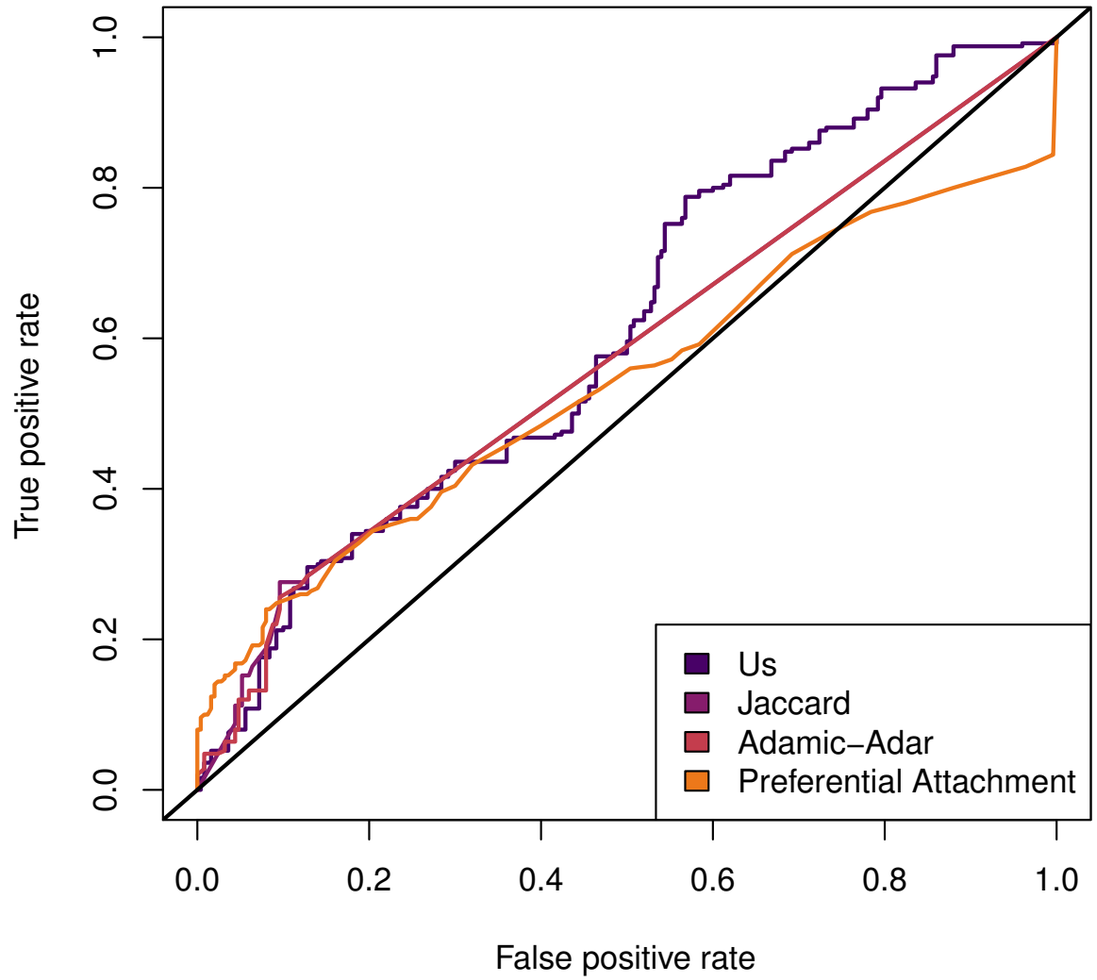
Score possible edges according to one of these methods and rank. Methods are generally based on comparing neighbors.

# Link prediction results

## Microbiome




## Protein




Prediction based on community assignments from nearest neighbors in attribute space and score is SBM edge probability based on those communities

# Validation with collaborative filtering

---

Given connectivity patterns, collaborate filtering seeks to predict node attributes.

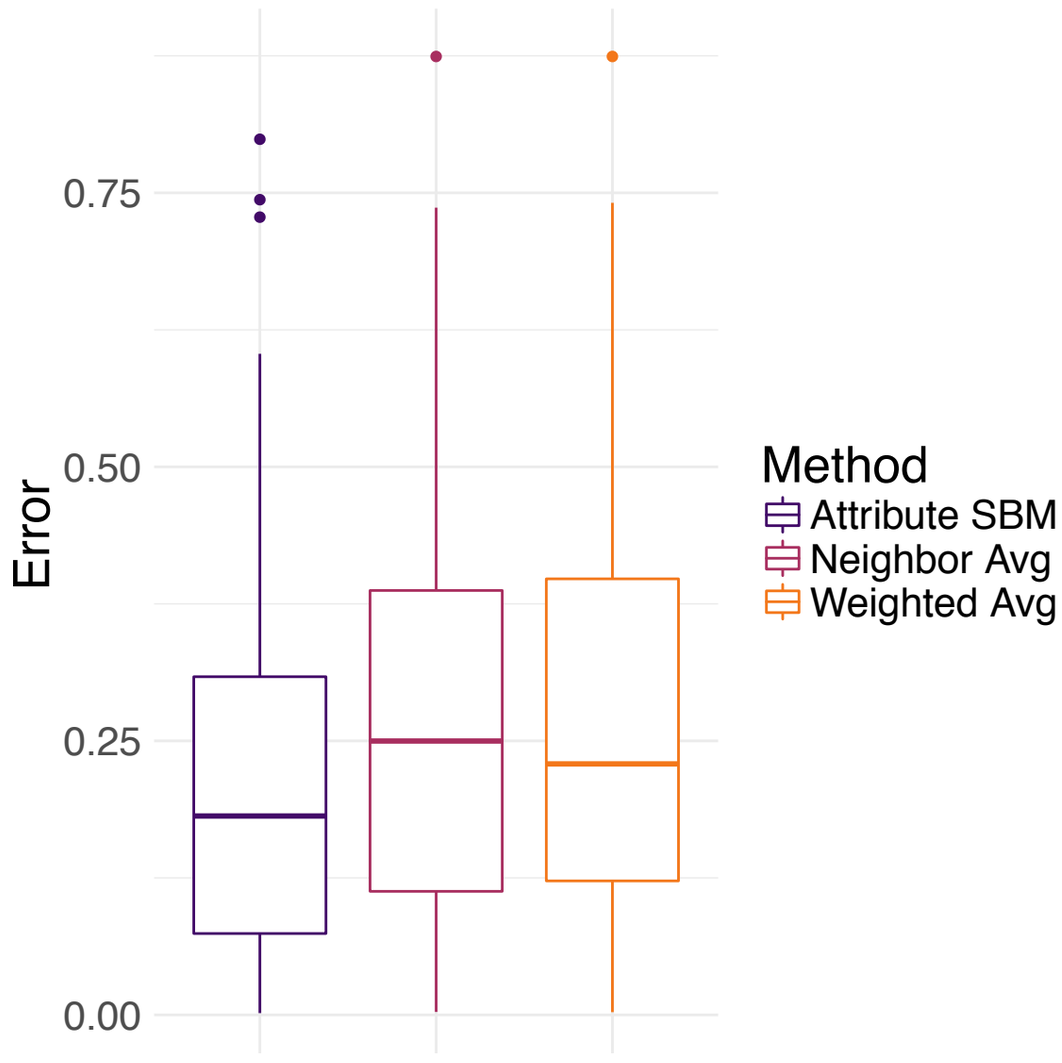
**Neighborhood Avg:** $\hat{\mathbf{x}}_i = \frac{1}{|\mathcal{N}^k(i)|} \sum_{j \in \mathcal{N}^k(i)} \mathbf{x}_j$

**Weighted Neighborhood Avg:**

$$\hat{\mathbf{x}}_i = \frac{1}{\sum_{j \in \mathcal{N}^k(i)} s_{ij}} \sum_{j \in \mathcal{N}^k(i)} s_{ij} \mathbf{x}_j$$

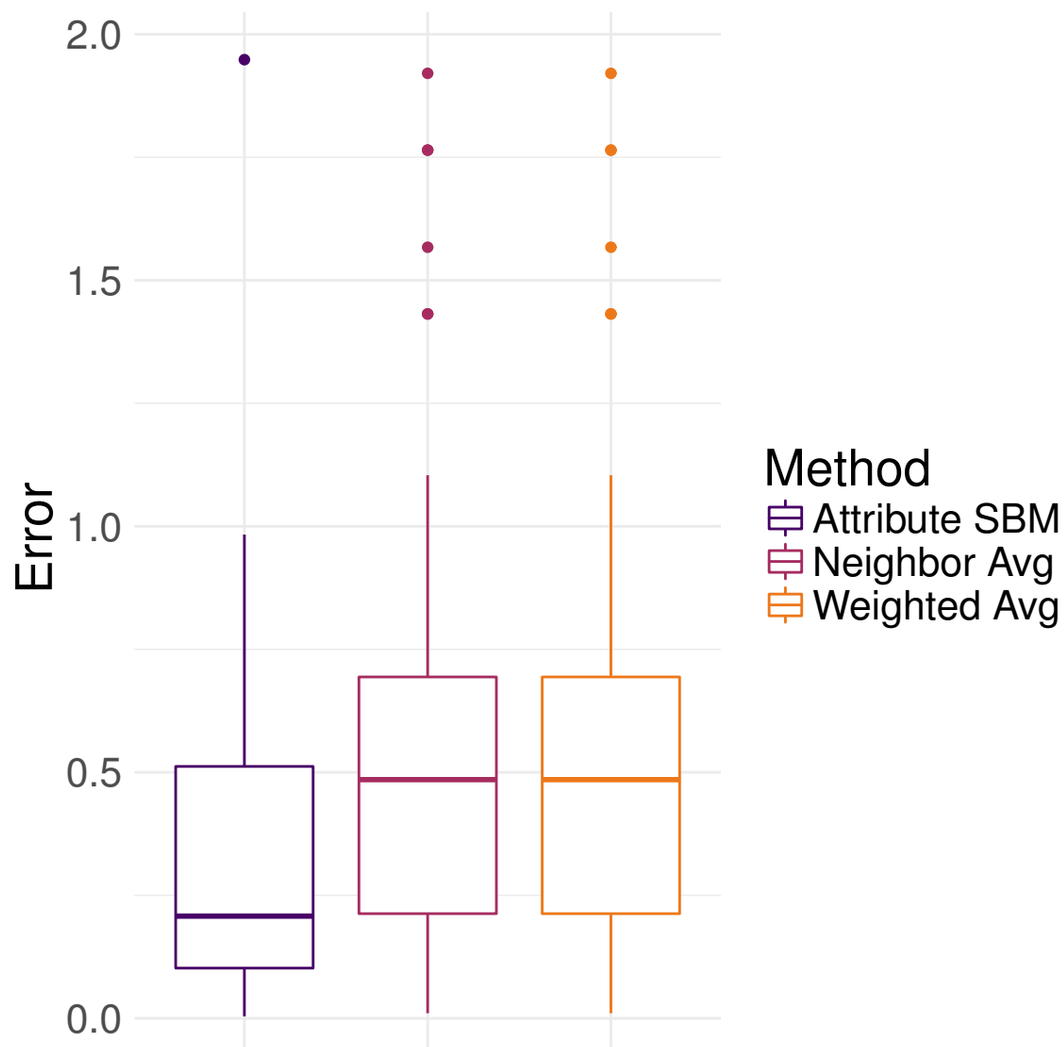
# Collaborative Filtering results

## Microbiome




Error $\mathcal{E} = \frac{||\hat{\mathbf{x}}_i - \mathbf{x}_i||_2}{||\mathbf{x}_i||_2}.$

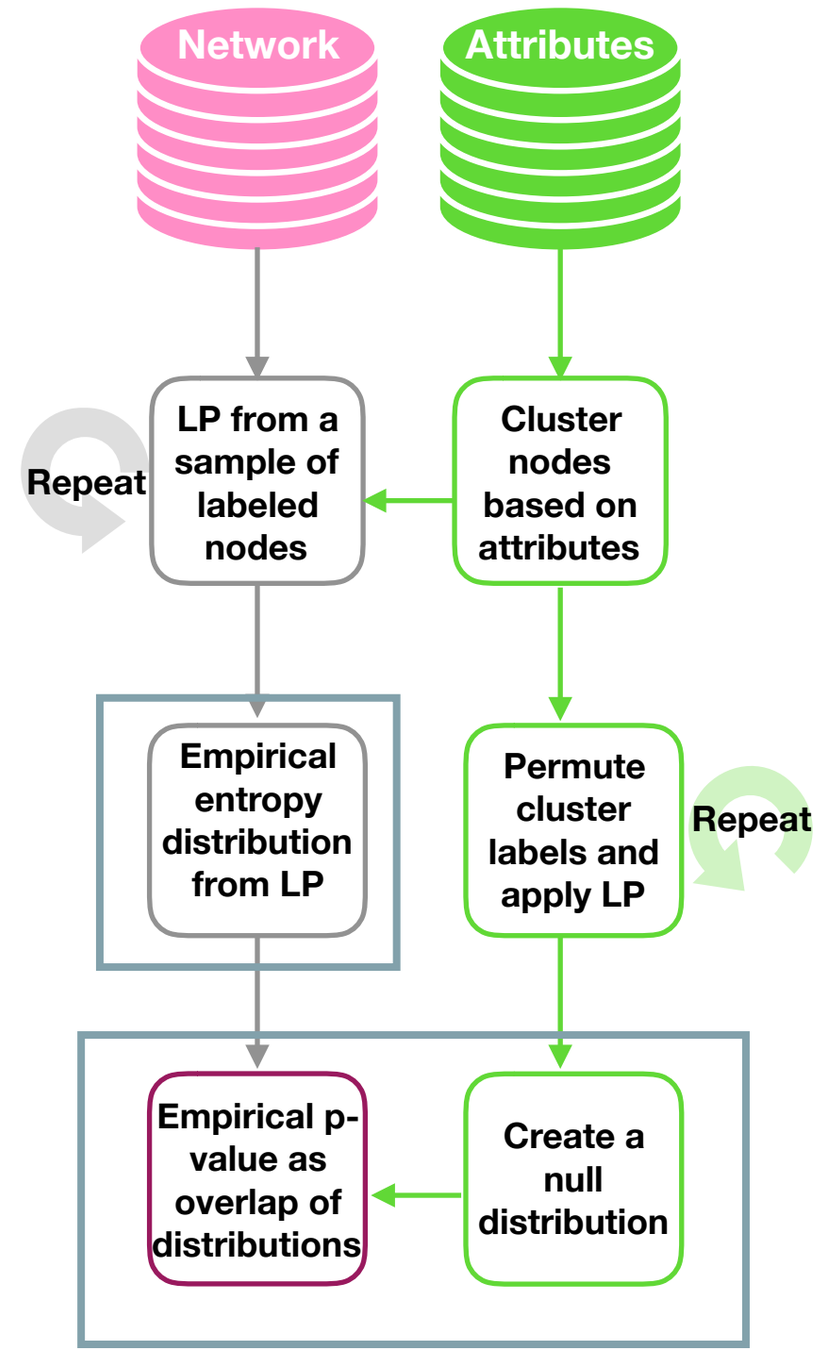
## Protein




Use the most common community assignment of neighbors in connectivity space and predict attribute as the mean of that community

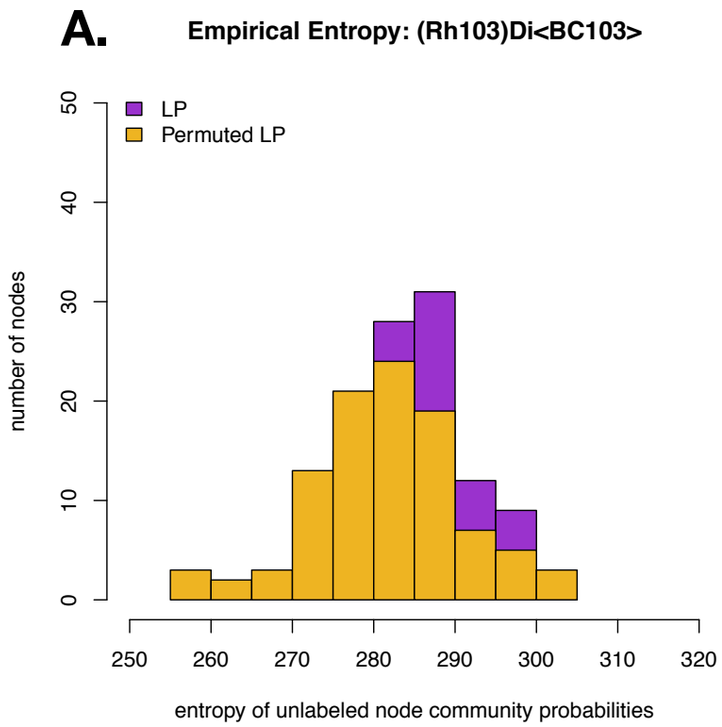
# A test for our assumption of alignment between attributes and connectivity.

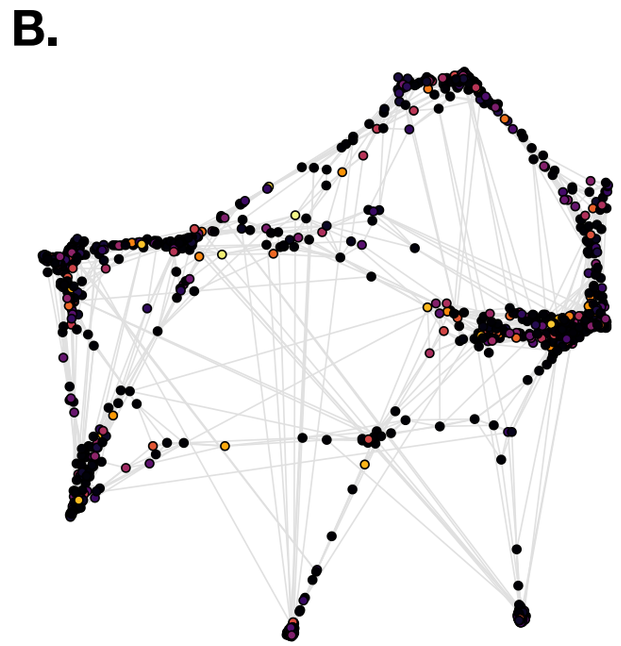
Use label propagation to compute the probability distribution over an unlabeled set of nodes. **The higher the entropy, the less attributes align with connectivity.**



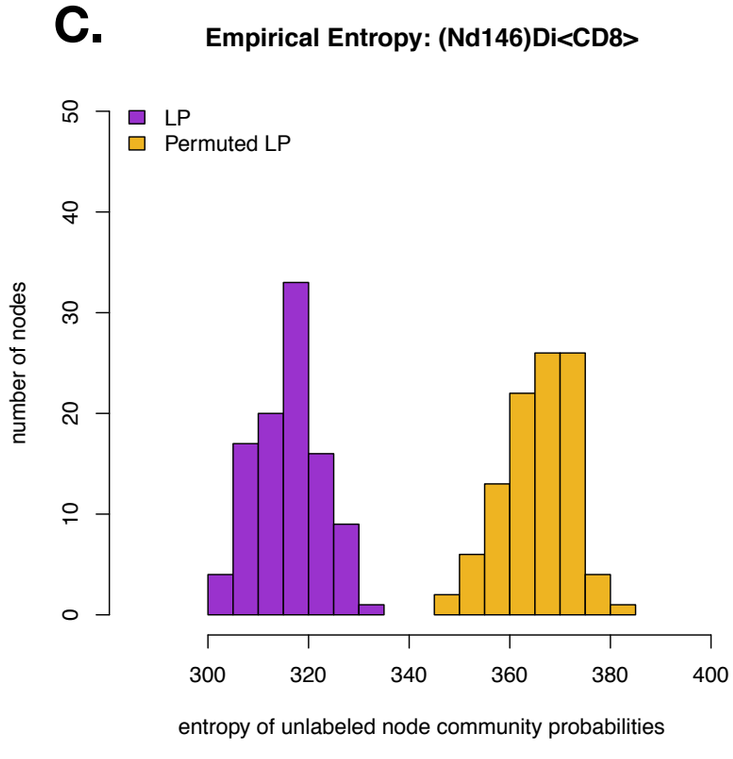
# Test on a single cell kNN

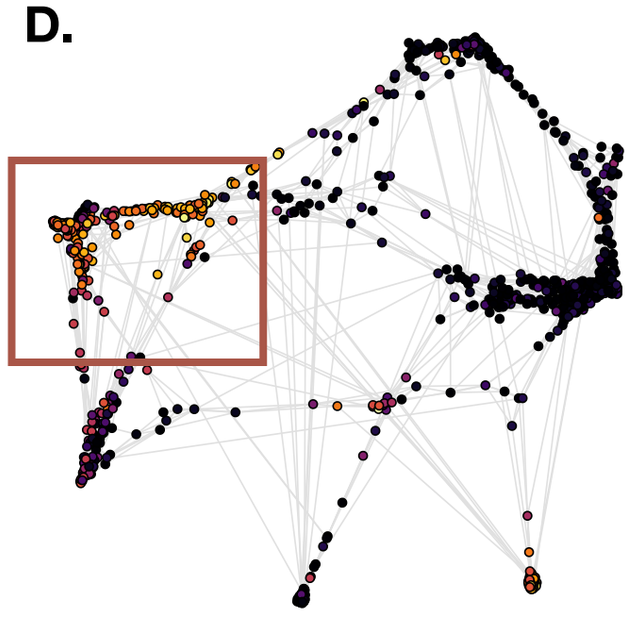
Poorly aligned attribute



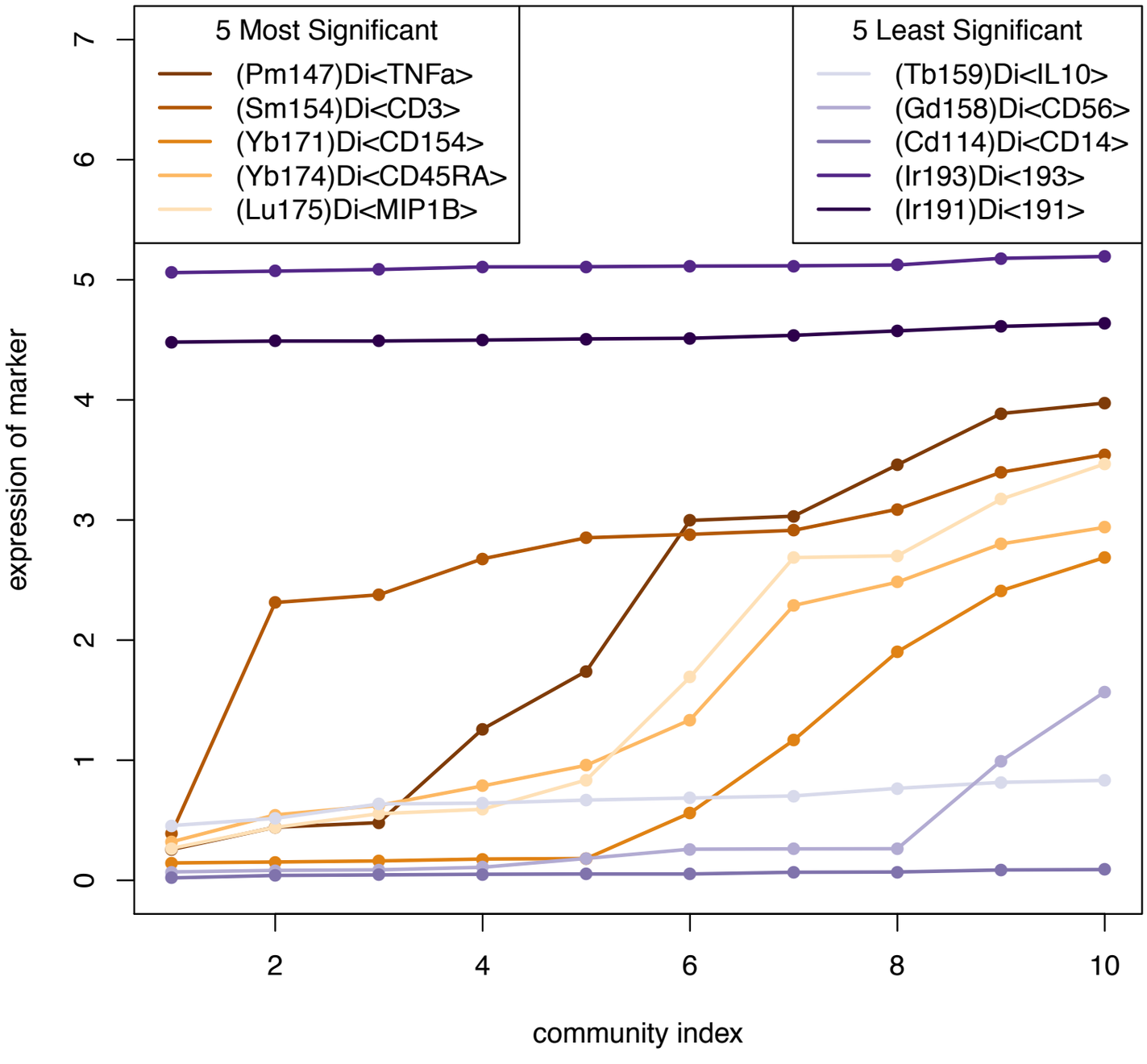




Well aligned attribute



# Significant attributes are discriminative between communities



# Recap on attribute work

- We developed an extension to the **stochastic block model** that can **incorporate multiple node attributes**.
    - Attributes are modeled with a multivariate Gaussian model based on community membership.
    - We validate our results on link prediction and collaborative filtering tasks.
- We developed a test to **quantify the relatedness between attributes and connectivity**.
    - We validate our results on a single cell kNN with single mass cytometry features as attributes.

# Future Work

- Better automated model selection
    - Model selection criteria for the number of communities.
- Adapting these approaches to weighted networks
    - Stochastic block models are particularly challenging in weighted networks

# Acknowledgements (Part I): Mentors

Pareto Scaling

Peter

Saray

Marc



Dane

# Acknowledgements (II): Committee


- Jeremy Purvis (who is also the chair of the committee)
- Tamara Berg
- David Gotz
- Laura Miller

# Acknowledgements (III). Good times in Chapel Hill