The ground truth about metadata and community detection in networks

Leto Peel, **Daniel B. Larremore**, Aaron Clauset

Jason Cory Brunson

Center for Quantitative Medicine, UConn Health

November 12, 2019

Abstract

1. Metadata *are not* ground truth
2. Community detection *is not* uniquely solvable
3. Metadata–community interactions *can* be measured

Evaluating community detection methods

Community detection

- Analog of clustering for network (relational) data
- Diverse applications
- Diverse meanings of “community”

Ground truth

- Useful (vital?) to evaluate & compare methods
- Known for generative simulation-based models
- Epistemically questionable for empirical models

Metadata

- Categories or classifications
  - sex, ethnicity, ZIP, primary diagnosis
- Often substituted for ground truth
- Simulations may not reflect real-world processes

The trouble with metadata and community detection

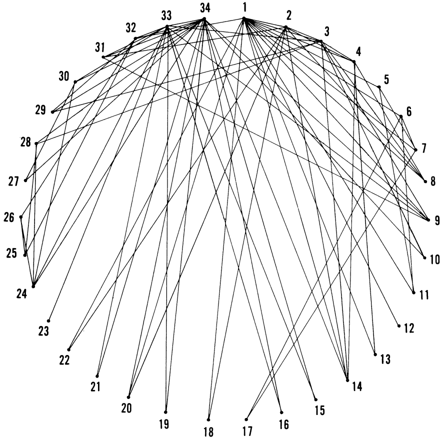
Dilemma

- High metadata–community correlation indicates that metadata are important to network generation
- Low correlation may arise from
  1. irrelevance of metadata to structure
  2. indirect relationship between metadata and structure
  3. absence of community structure
  4. failure of community detection method

Possible implications

- Over-reporting of poor performance by community detection methods
- Under-reporting of patterns uncorrelated with metadata

Illustration: Zachary's Karate Club



Zachary • 1977 • *J. Anthropol. Res.*

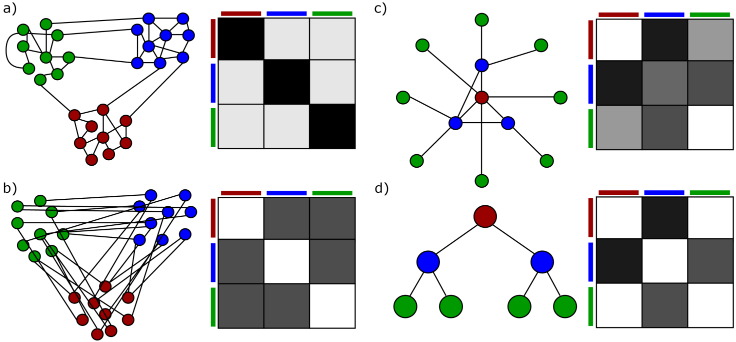
Epistemological status

- Heterogeneous, weighted links
  - university classes
  - karate workouts
  - rathskeller
  - nearby bar
  - tournaments
- Multiple metadata attributes
  - political leaning
  - faction joined
- Erroneous datum

Illustration: Zachary’s Karate Club

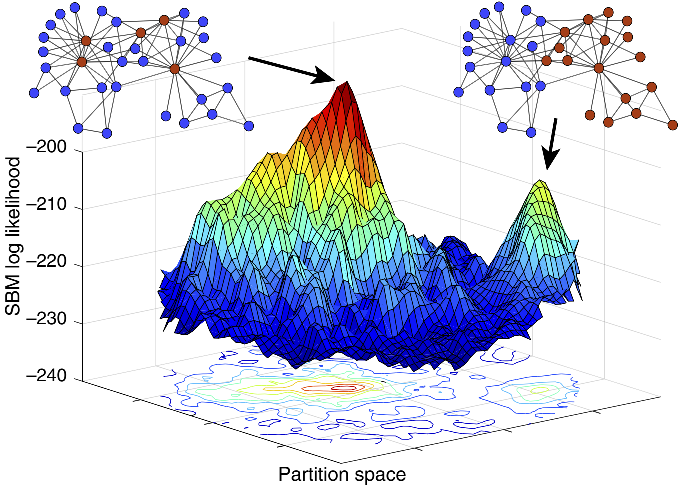
Specificity / well-definedness

- Embed the space of partitions in $\mathbb{R}^2$
- Graph log-likelihoods under the stochastic blockmodel



Funke & Becker • 2019 • *PLOS ONE*

Illustration: Zachary's Karate Club



The ground truth community detection problem

- $\mathcal{G}$ a network
  - generated by a process g
  - from a ground truth partition $\mathcal{T}$
- $\mathcal{C}$ be a partition of $\mathcal{G}$
  - obtained by a community detection method f
- d be a measure of distance between partitions of $\mathcal{G}$

Inverse Problem

$$f^*(\mathcal{G}) = \operatorname*{argmin}_f d\left(\mathcal{T}, f(\mathcal{G})\right)$$

Universal Solution

$$\exists f^*, \quad \forall \{g, \mathcal{T}\}, \quad \operatorname*{argmin}_f d\left(\mathcal{T}, f(g(\mathcal{T}))\right)$$

Ground-truth community detection is an ill-posed inverse problem

Well-posedness

i. A solution exists
ii. The solution is unique
iii. The solution changes continuously with initial conditions

Ground-truth community detection is an ill-posed inverse problem

Well-posedness

i. A solution exists
ii. The solution is unique
iii. The solution changes continuously with initial conditions

Theorem
For a fixed network $\mathcal{G}$, the solution to the ground truth community detection problem is not unique.

Ground-truth community detection is an ill-posed inverse problem

Well-posedness

i. A solution exists
ii. The solution is unique
iii. The solution changes continuously with initial conditions

Theorem

For a fixed network $\mathcal{G}$, the solution to the ground truth community detection problem is not unique.

Proof.

Any graph $\mathcal{G}$ can be produced with positive probability by both

- $\mathcal{T}$ = coarsest partition; g = Erdős-Rényi model
- $\mathcal{T}$ = finest partition; g = deterministic model that recovers $\mathcal{G}$

No Free Lunch for community detection

NFL for machine learning

For supervised learning problems, the expected misclassification rate across all possible data sets is independent of the algorithm.

NFL for community detection

1. Translate the community detection problem into the Extended Bayesian Framework (EBF)
2. Choose a suitable loss function ℓ with total error $L(\ell)$
3. Prove NFL:

$$\forall f, \quad \sum_{g,\mathcal{T}} \ell(\mathcal{T}, f(g(\mathcal{T}))) = L(\ell)$$

Community detection in the EBF

Supervised EBF (classification)

Posit:

- a countable input space X, $|X| = n$
- a countable output space Y, $|Y| = r$
- the density function $\sigma_X = P(x \mid \sigma)$
- the conditional distribution $\gamma = pY \mid X$
- a training set d of samples (x_i, y_i), $Y_i \sim \gamma(X_i)$

Compute:

- for each test case $x \in X$, a hypothesis $h \in Y$
- model (algorithm) $P(h \mid d, x)$ combining priors and data

Community detection in the EBF

Supervised EBF (classification)

Posit:

- a countable input space X, $|X| = n$
- a countable output space Y, $|Y| = r$
- the density function $\sigma_X = P(x \mid \sigma)$
- the conditional distribution $\gamma = pY \mid X$
- a training set d of samples (x_i, y_i), $Y_i \sim \gamma(X_i)$

Compute:

- for each test case $x \in X$, a hypothesis $h \in Y$
- model (algorithm) $P(h \mid d, x)$ combining priors and data

Unsupervised EBF (clustering and community detection)

- $d = \varnothing$
- $P(h)$ encodes priors (assumptions about γ) only

Loss functions

Supervised EBF (classification)

- error random variable $C \sim P(c \mid h, \gamma, d)$
- expected error $E(C \mid h, \gamma, d)$
- typical loss functions ℓ
  - misclassification rate
  - normalized mutual information

Loss functions

Supervised EBF (classification)

- error random variable $C \sim P(c \mid h, \gamma, d)$
- expected error $E(C \mid h, \gamma, d)$
- typical loss functions ℓ
  - misclassification rate
  - normalized mutual information

Unsupervised EBF (clustering and community detection)

Group labels:

- matter to classification problems
- *don't* matter to clustering problems

Normalized mutual information

- N objects
- partition $u \in \mathcal{P}(N)$ of objects into K_u groups
- proportional sizes $p_i = |u_i|/N$

Entropy of u:

$$H(u) = -\sum_{i=1}^{K_u} p_i \log(p_i)$$

Mutual information between u, v:

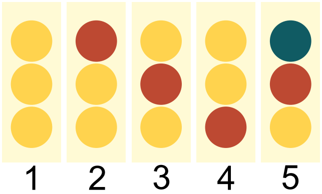
$$I(u,v) = \sum_{i=1}^{K_u} \sum_{j=1}^{K_v} p_{ij} \log\left(\frac{p_{ij}}{p_i p_j}\right)$$

Normalized mutual information between u, v:

$$\mathrm{NMI}(u,v) = \frac{I(u,v)}{\sqrt{H(u)H(v)}}$$

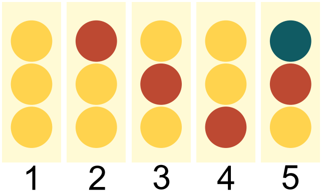
Loss functions and *a priori* superiority

Typical loss functions imply *a priori* superiority of some algorithms based on labeling schemes



Loss functions and *a priori* superiority

Typical loss functions imply *a priori* superiority of some algorithms based on labeling schemes



NMI on $\mathcal{P}(3)$:

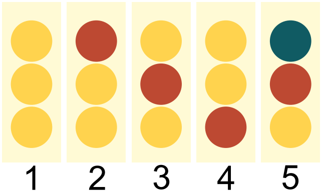
	Partition 2				
Partition 1	**1**	**2**	**3**	**4**	**5**
1	1	0	0	0	0
2	0	1	0.27	0.27	0.76
3	0	0.27	1	0.27	0.76
4	0	0.27	0.27	1	0.76
5	0	0.76	0.76	0.76	1
$\mathbb{E}$[NMI]	0.20	0.46	0.46	0.46	0.66

Adjusted MI (AMI) on $\mathcal{P}(3)$:

	Partition 2				
Partition 1	**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

Loss functions and *a priori* superiority

Typical loss functions imply *a priori* superiority of some algorithms based on labeling schemes



NMI on $\mathcal{P}(3)$:

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

Adjusted MI (AMI) on $\mathcal{P}(3)$:

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

Homogeneity:

- $\lim_{N\to\infty}$ AMI(u) = 0 (superexponentially)
- the space defined by AMI is “geometry-free”

Lemma and theorems

Lemma

AMI is a homogeneous loss function over the interior of $\mathcal{P}(N)$. Including boundary partitions, AMI is homogeneous within ${\mathcal{B}_N}^{-1}$.

Theorem

For a homogeneous loss function ℓ, the uniform average of $P(c \mid \gamma, d)$ over γ is $L(c)/r$.

Theorem

For the community detection problem with the AMI loss function, the uniform average of $P(c \mid \gamma)$ over γ equals $L(c)/r$.

Implications

- Any subset of problems for which an algorithm over-performs others is balanced by another subset for which is over-performed by others.
- A non-uniform subset of problems may have an algorithm that over-performs another.

Relating metadata and structure

Complementary roles

- Metadata describe the nodes (individually)
- Communities describe how the nodes interact

Proposed hypothesis tests

1. blockmodel entropy significance test (BESTest)
    - test whether metadata and communities are related
    - case (i)
2. neo–stochastic blockmodel (neoSBM)
    - test whether metadata represent the same or different aspects as communities
    - case (ii)

Testing for a relationship btw metadata and structure

Blockmodel entropy significance test (BESTest)

- Assumptions
  - network $\mathcal{G}$ generated via SBM with partition $\mathcal{C}$
  - metadata partition π
- Hypotheses
  - H_0: π is irrelevant to $\mathcal{C}$
  - H_A: π is relevant to $\mathcal{C}$
- Test statistic
  - SBM with MLE parameters $\hat{\omega}_{rs} = \dfrac{m_n}{n_r n_s}$
  - entropy $H_{\text{SBM}}(\mathcal{G};\pi)$
- Estimation
  - sample entropies $H(\mathcal{G};\tilde{\pi})$ over random permutations $\tilde{\pi}$
  - simplification (Bernoulli SBM) or first-order approximation (sparse networks) of $H(\mathcal{G})$
- p-value
  - $p = \Pr(H(\mathcal{G};\tilde{\pi}) \leq H_{\text{SBM}}(\mathcal{G};\pi))$

Sensitivity of the BESTest p-value

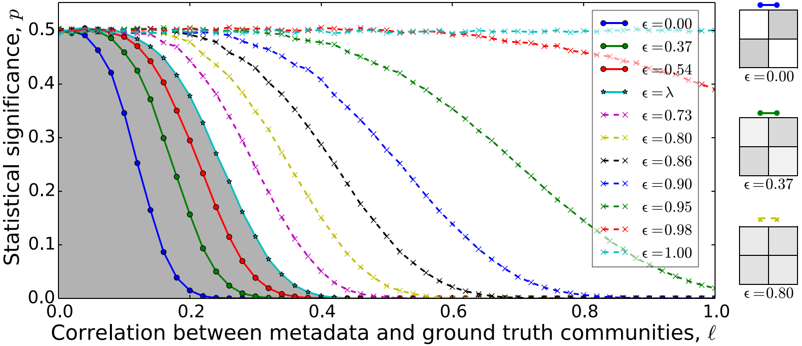
Synthetic networks

- $N = 1000$ nodes
- two planted communities r, s
  - nodes allocated with uniform probability to r, s
- community strength $\epsilon = \dfrac{\omega_{rs}}{\omega_{rr}}$
  - low ϵ: strongly assortative communities
  - value & constancy of density unclear
- nodes labeled correctly with probability $\ell \in [0, 1]$
  - otherwise randomly labeled
  - Pr(metadata matches community) = $\dfrac{1 + \ell}{2}$

Sensitivity of the BESTest p-value

- community strength $\epsilon = \frac{\omega_{rs}}{\omega_{rr}}$
- nodes labeled correctly with probability $\ell \in [0, 1]$
- detectability regime $\epsilon < \lambda$

Decelle, Krzakala, Moore, Zdeborova • 2011 • *Phys. Rev. Lett.*



Demonstrations of BESTest on real-world networks

Lazega Lawyers

- 71 attorneys
- 3 link types (friendship, advice, cases)
- 5 metadata variables (status, gender, location, practice, school)

Table 1. BESTest *P* values for Lazega Lawyers.

	Metadata attribute				
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

Demonstrations of BESTest on real-world networks

Lazega Lawyers

- 71 attorneys
- 3 link types (friendship, advice, cases)
- 5 metadata variables (status, gender, location, practice, school)

Table 1. BESTest *P* values for Lazega Lawyers.

	Metadata attribute				
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

Malaria parasite genes

- 307 gene sequences
- 9 layers (genetic substring–sharing networks)
- 3 metadata variables (upstream promoter, cysteine / PoLV group, parasite origin)
  Bull, Kyes, Buckee, &al • 2007 • *Mol. Biochem. Parasitol.*

Table 2. BESTest *P* values for malaria *var* genes.

	var gene network number								
	1	2	3	4	5	6	7	8	9
Genome	0.566	0.064	0.536	0.588	0.382	0.275	0.020	0.464	0.115

Diagnosing the structural aspects captured by both

neo-stochastic blockmodel (neoSBM)

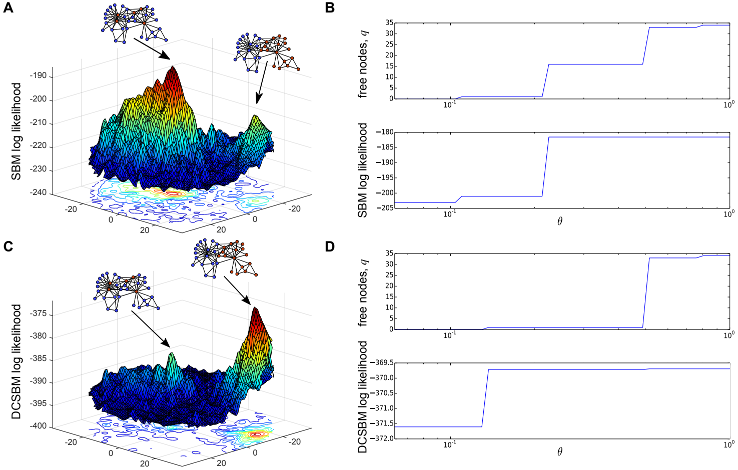
- Assumptions
  - network $\mathcal{G}$, $|G| = N$, optimal SBM partition $\mathcal{C}$
  - metadata partition π
  - latent node states $z_i \in \{b, r\}$; $q = |\{i \mid z_i = r\}|$
  - uniform prior probability $\theta = \Pr(z_i = r)$
- Likelihood
  - cost of freedom $\psi(\theta) = \dfrac{1}{N\theta} \sum_i \delta_{z_i r} \left(\log \dfrac{\theta}{1-\theta} \right)$
  - log-likelihood $\mathcal{L}_{\text{neo}}(\mathcal{G}; \pi, z) = \mathcal{L}_{\text{SBM}}(\mathcal{G}; \pi) + q\psi(\theta)$
- Estimation
  - necessarily $\mathcal{L}_{\text{SBM}}(\mathcal{G}; \pi) \leq \mathcal{L}_{\text{SBM}}(\mathcal{G}; \mathcal{C})$
  - optimize $\mathcal{L}_{\text{SBM}}$ when $\hat{q} = \sum_i 1 - \delta_{\pi_i, \mathcal{C}_i}$

Idea

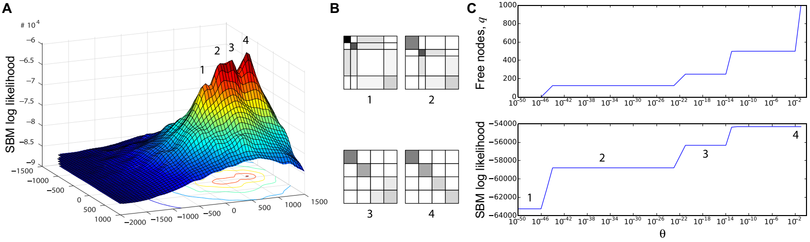
Interpolate through $\mathcal{P}(N)$ from π to $\mathcal{C}$ and monitor improvement in $\mathcal{L}_{\text{SBM}}$.

Demonstration of neoSBM on the Karate Club network

neoSBM versus neoDCSBM



Demonstration of neoSBM on a synthetic network

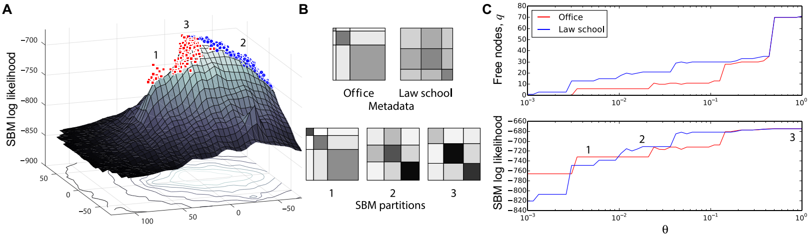


Observations

- transition from lowest local maximum π to highest $\mathcal{C}$
  - core–periphery structure at π
  - assortative group structure at $\mathcal{C}$

Demonstration of neoSBM on the Lazenga Lawyers

- office location π_1 and law school π_2 metadata partitions
- friendship network structure with global SBM optimum $\mathcal{C}$



Observations

- no intermediate local optima encountered from π_2
- one intermediate local optimum encountered from π_1

Discussion

There is no universally accepted definition of community structure, nor should there be.

Outlook

- trade-off between general and specialized community detection methods
  - **general**: perform reasonably well in many settings
  - **specialized**: perform very well in tailored settings
- most work to date is on general methods
- need to better understand general–specific trade-offs
  - measure errors obtained in domain-agnostic applications
  - incorporate metadata into the inference process

Fin