Reconstruction performance of the stochastic block model (SBM) in empirical networks

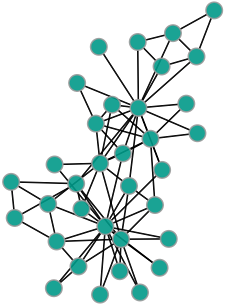
Felipe Vaca-Ramírez & Tiago P. Peixoto

Central European University
Vienna, Austria

NetSci, July 2023

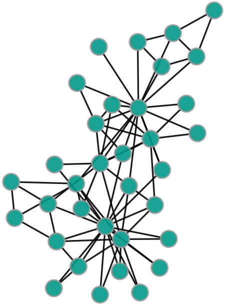
Network data are noisy

Many times we don't observe the true network $\boldsymbol{A}$,

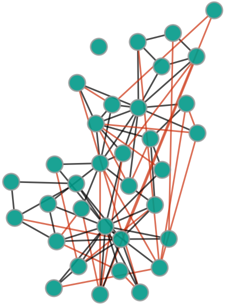


Network data are noisy

Many times we don't observe the true network $\boldsymbol{A}$,



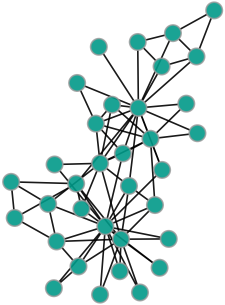
but noisy or incomplete data $\boldsymbol{D}$



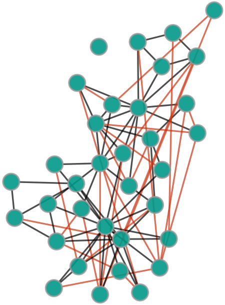
The measurement process is often overlooked: $P(\boldsymbol{D}|\boldsymbol{A})$

Network data are noisy

Many times we don't observe the true network $\boldsymbol{A}$,



but noisy or incomplete data $\boldsymbol{D}$

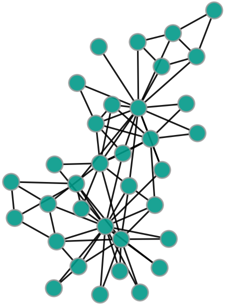


The measurement process is often overlooked: $P(\boldsymbol{D}|\boldsymbol{A})$

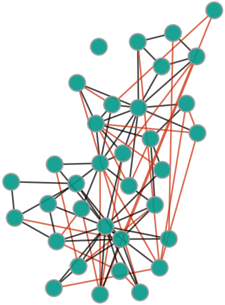
What can we do? Reconstruct the original network: $P(\boldsymbol{A}|\boldsymbol{D})$

Network data are noisy

Many times we don't observe the true network $\boldsymbol{A}$,



but noisy or incomplete data $\boldsymbol{D}$



The measurement process is often overlooked: $P(\boldsymbol{D}|\boldsymbol{A})$

What can we do? Reconstruct the original network: $P(\boldsymbol{A}|\boldsymbol{D})$

How? Generative models (e.g., SBM).

We can use the SBM to:

- Generate Networks

$$P(\boldsymbol{A}|\boldsymbol{\omega}, \boldsymbol{b}) = \prod_{i<j} \omega_{b_i,b_j}^{A_{ij}} \left(1 - \omega_{b_i,b_j}\right)^{1-A_{ij}} \tag{1}$$

- Infer node partitions of networks

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})} \tag{2}$$

We can use the SBM to:

- Generate Networks

$$P(\boldsymbol{A}|\boldsymbol{\omega}, \boldsymbol{b}) = \prod_{i<j} \omega_{b_i,b_j}^{A_{ij}} \left(1 - \omega_{b_i,b_j}\right)^{1-A_{ij}} \tag{1}$$

- Infer node partitions of networks

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})} \tag{2}$$

- Reconstruct Networks

$$P(\boldsymbol{A}, \boldsymbol{b}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{A})P(\boldsymbol{A}, \boldsymbol{b})}{P(\boldsymbol{D})} \tag{3}$$

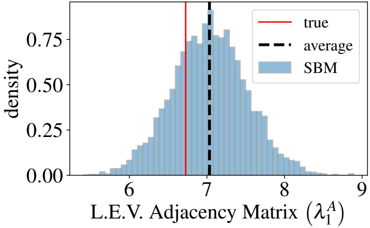
with $P(\boldsymbol{D}|\boldsymbol{A})$ being the model of the measurement process,

The SBM can be tested in this framework!

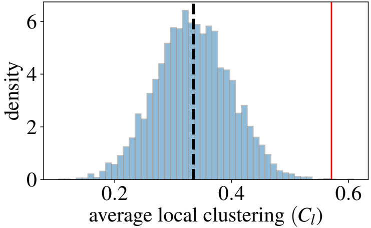
T. P. Peixoto, Physical Review X 8, 041011 (2018).

How to assess? Absolute assessment

Higher Accuracy
($error = 0.046$)

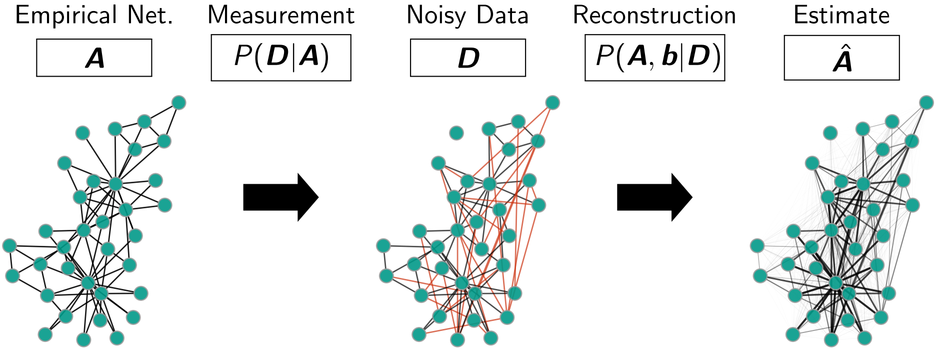


Lower Accuracy
($error = 0.236$)



How accurate is the SBM in estimating relevant features of empirical networks?

Experimental Setup and the *karate club*



Empirical Net.

$\boldsymbol{A}$

Measurement

$P(\boldsymbol{D}|\boldsymbol{A})$

Noisy Data

$\boldsymbol{D}$

Reconstruction

$P(\boldsymbol{A},\boldsymbol{b}|\boldsymbol{D})$

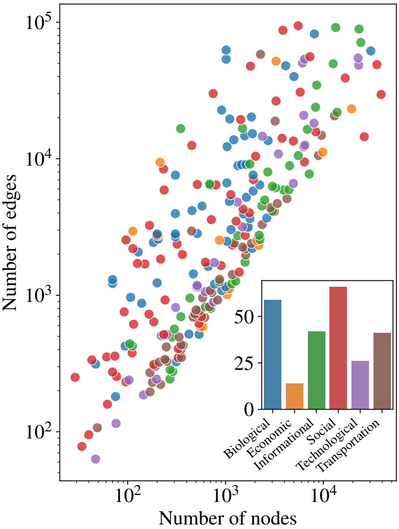
Estimate

$\hat{\boldsymbol{A}}$

How to generate noisy data?

- flip a coin on every edge and remove it with probability p.
- flip a coin on every non-edge and add a spurious edge with probability q.
- preserve density: $q = pE/\left(\binom{N}{2} - E\right)$.

Network Corpus and Descriptors

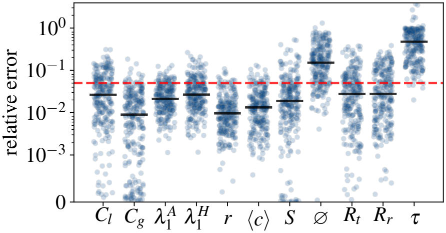


248 real-world networks

Symbol	Descriptor
r	Degree assortativity
$\langle c \rangle$	Mean k-core value
C_l	Mean local clustering coefficient
C_g	Global clustering coefficient
$\varnothing$	Pseudo-diameter
S	Fraction of nodes in the largest component
λ_1^A	Leading eigenvalue of the adjacency matrix
λ_1^H	Leading eigenvalue of the Hashimoto matrix
τ	Characteristic time of a random walk
R_r	Node percolation profile (random removal)
R_t	Node percolation profile (degree-targeted removal)

Assessing Performance ($p = 0.1$)

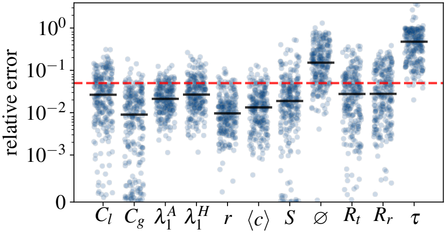
(a) Accuracy



(Distribution of reconstruction errors and 0.05-threshold. Median in black.)

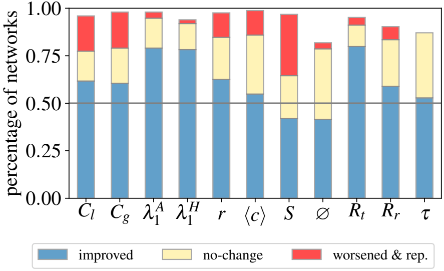
Assessing Performance ($p = 0.1$)

(a) Accuracy



(Distribution of reconstruction errors and 0.05-threshold. Median in black.)

(b) Improvement



(Error *after* reconstruction *vs.* error *before* reconstruction.)

Average local clustering C_l

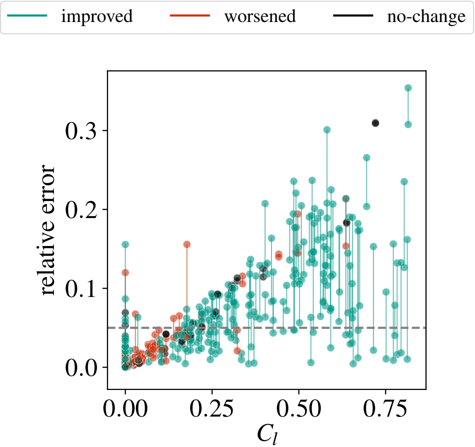


Figure 1: Relative error before and after reconstruction (joined by a line segment) as a function of the original value of the descriptor. The color indicates if the error after reconstruction is smaller than before doing it (i.e., there is improvement) or not. Noise level $p = 0.1$.

Diameter $\varnothing$

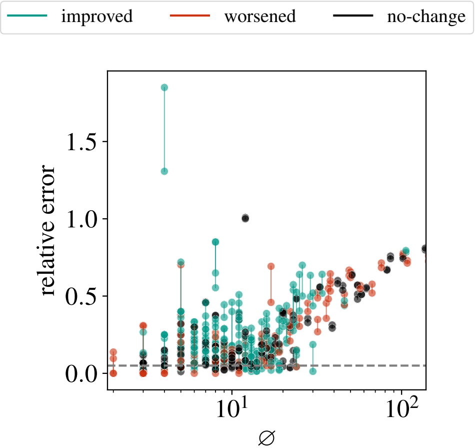
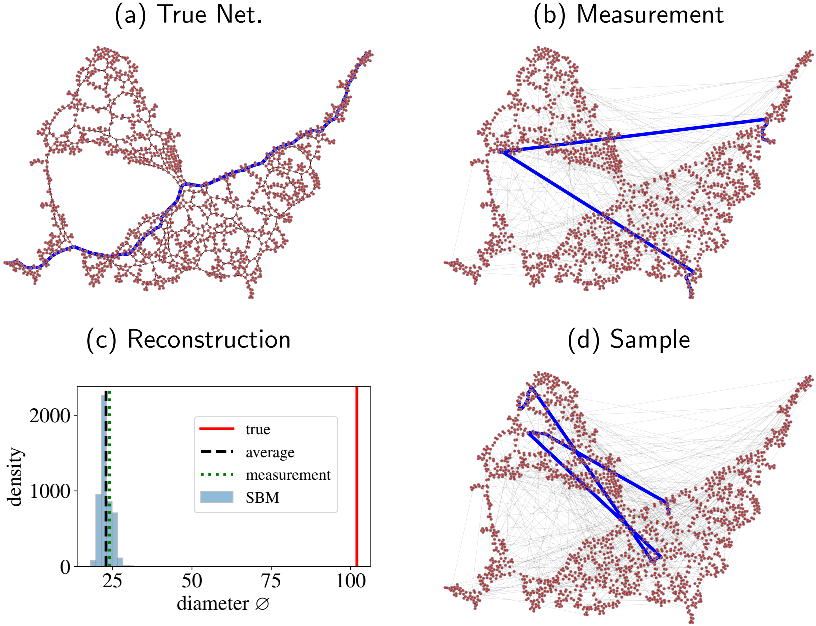


Figure 2: Relative error before and after reconstruction (joined by a line segment) as a function of the original value of the descriptor. The color indicates if the error after reconstruction is smaller than before doing it (i.e., there is improvement) or not. Noise level $p = 0.1$.

Estimation of diameter ∅ in Venice street network



(a) True Net.

(b) Measurement

(c) Reconstruction

(d) Sample

Network Domains

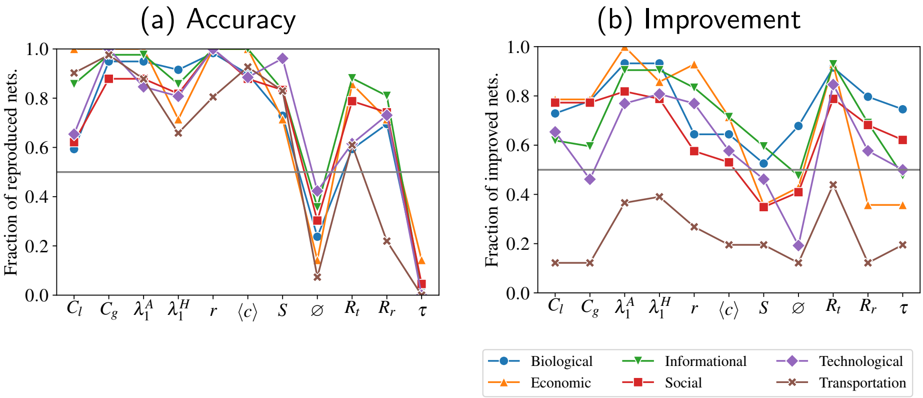


Figure 3: (a) Percentage of accurately estimated networks. (b) Average difference between the error before reconstruction and after reconstruction for each descriptor; in both cases by domain. Noise level $p = 0.1$.

Final remarks

- Does SBM provide accurate estimations of relevant features of empirical networks? Overall, yes...
- Do we gain from reconstruction? Overall, yes...

Final remarks

- Does SBM provide accurate estimations of relevant features of empirical networks? Overall, yes...
- Do we gain from reconstruction? Overall, yes...
- Exceptions: large diameter and slow-mixing random walks

Final remarks

- Does SBM provide accurate estimations of relevant features of empirical networks? Overall, yes...
- Do we gain from reconstruction? Overall, yes...
- Exceptions: large diameter and slow-mixing random walks
- Next steps:
  - other models of noise (e.g., on hubs) and structure.
  - other parameters (e.g., upper bound for noise).
  - other networks, descriptors (e.g., dynamics on multilayer networks)

Final remarks

- Does SBM provide accurate estimations of relevant features of empirical networks? Overall, yes...
- Do we gain from reconstruction? Overall, yes...
- Exceptions: large diameter and slow-mixing random walks
- Next steps:
  - other models of noise (e.g., on hubs) and structure.
  - other parameters (e.g., upper bound for noise).
  - other networks, descriptors (e.g., dynamics on multilayer networks)

THANK YOU!

Appendix: Dealing with error

What do we gain from doing more measurements n?

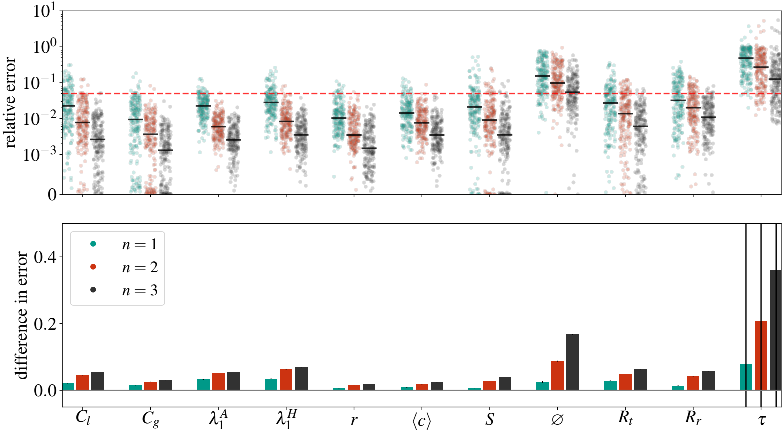


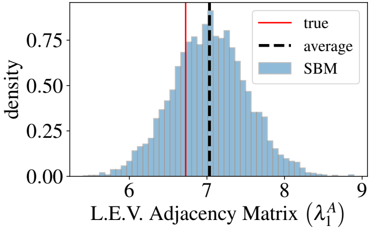
Figure 4: (Top) Percentage of accurately estimated networks. (Bottom) Average difference between the error before reconstruction and after reconstruction for each descriptor. The color maps to the number of times a node pair was measured (n). Noise level $p = 0.1$.

How to assess the Performance of Reconstruction?

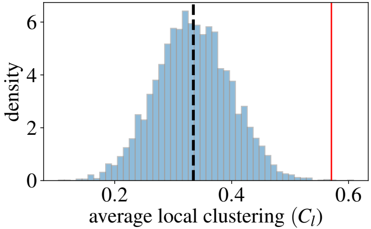
For each descriptor y, we also get a distribution...
and we can compute its average $\hat{y}$.

Error after reconstruction: $|(y(\boldsymbol{A}) - \hat{y})/y(\boldsymbol{A})|$,

Higher Accuracy
($error = 0.046$)



Lower Accuracy
($error = 0.236$)



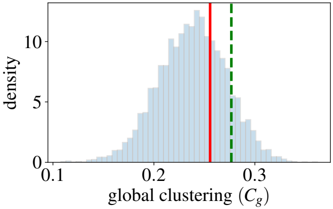
How accurate is the SBM in estimating relevant features of empirical networks?

How to assess the Performance of Reconstruction?

We need a base line, e.g., the **Error before reconstruction**...

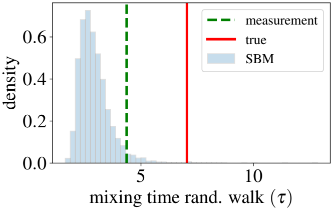
Less sensitive

($error = 0.021$)



More sensitive

($error = 0.381$)



What do we gain from reconstruction?