Disentangling homophily, community structure and triadic closure in networks

Tiago P. Peixoto

Department of Network and Data Science
Central European University, Vienna

Austin, April 2023

T.P.P., Phys. Rev. X 12, 011004 (2022)

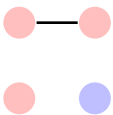
Networks as relational big data

- Often large
- High-dimensional
- Sparse
- Heterogeneously structured
- Richly annotated
- Often dynamic

Often the outcome of multiple generative mechanisms: how do we disentangle them?

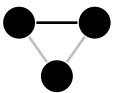
Homophily vs. triadic closure

Homophily



Tendency of an edge being placed between nodes of the same kind (e.g. age, race, location, etc.)

Triadic closure



Tendency of an edge being placed between nodes if they share a common neighbor.

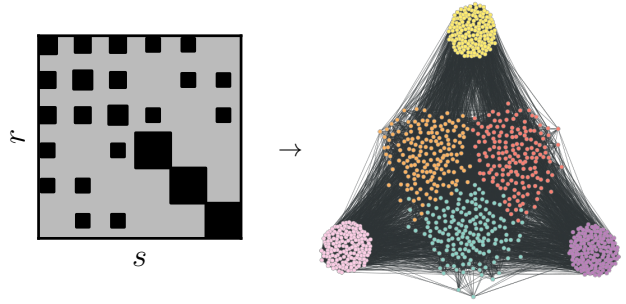
Both processes induce the formation of **triangles** and **community structure**.

How do we model homophily?

The stochastic block model (SBM)

Planted partition: N nodes divided into B groups.

Parameters: $b_i \to$ group membership of node i
$\lambda_{rs} \to$ edge probability from group r to s.

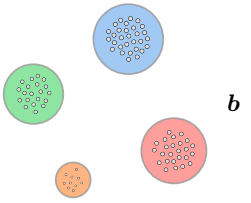


Degree-corrected: Arbitrary degree sequence: $\{k_i\}$

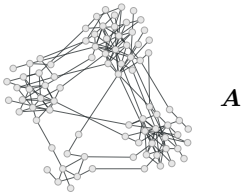
- Not restricted to homophily; can accommodate arbitrary mixing.

How do we detect homophily?

Generative model



Statistical inference



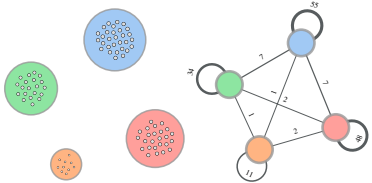
Bayes' rule:

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

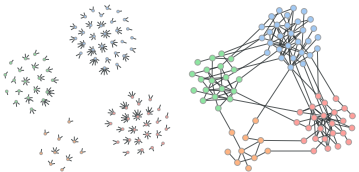
✓ **Avoids overfitting**
Incorporates Occam's razor; can be formally mapped into *data compression*. Can also be used to find the number of groups.

✓ **Efficient**
Scales to large networks.

Microcanonical degree-corrected stochastic block model (DC-SBM)



$P(\boldsymbol{b})$ $\quad$ $P(\boldsymbol{e}|\boldsymbol{b})$



$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})$ $\quad$ $P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})$

Microcanonical priors and likelihood

$$P(\boldsymbol{b}) = \frac{\prod_r n_r!}{N!} \times \binom{N-1}{B-1}^{-1} \times \frac{1}{N} \qquad P(\boldsymbol{e}|\boldsymbol{b}) = \left(\!\!\left(\!\!\binom{\left(\!\binom{B}{2}\!\right)}{E}\!\!\right)\!\!\right)^{-1}$$

$$P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b}) = \prod_r \left(\!\!\binom{n_r}{e_r}\!\!\right)^{-1} \qquad P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \frac{\prod_{i<j} A_{ij}! \prod_{ii} A_{ii}!! \prod_r e_r!}{\prod_{r<s} e_{rs}! \prod_r e_{rr}!! \prod_i k_i!}$$

Nonparametric posterior distribution:

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{k}|\boldsymbol{e},\boldsymbol{b})P(\boldsymbol{e}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{Data}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{Model}}$$

Inference $\leftrightarrow$ compression

Minimum description length (MDL)

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

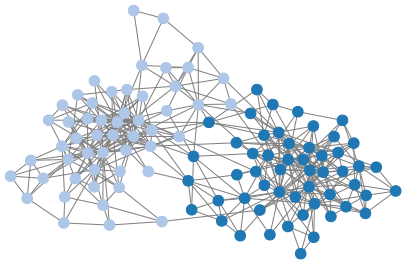
$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$

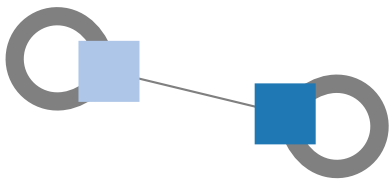
Minimum description length (MDL)

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$





$B = 2, \mathcal{S} \approx 1395.8$ bits

Model, $\mathcal{L} \approx 581.6$ bits

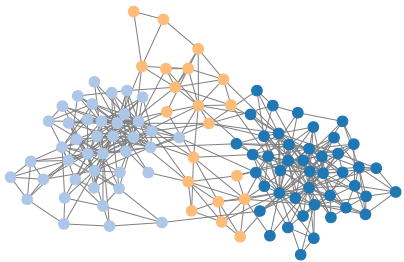
$\Sigma \approx 1977.5$ bits

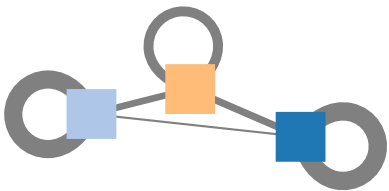
Minimum description length (MDL)

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$





$B = 3$, $\mathcal{S} \approx 1285.7$ bits

Model, $\mathcal{L} \approx 654.1$ bits

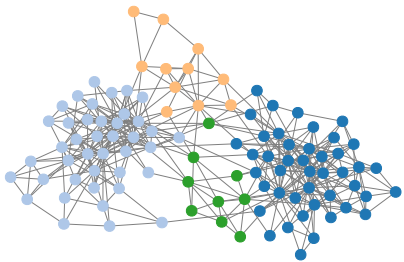
$\Sigma \approx 1939.7$ bits

Minimum description length (MDL)

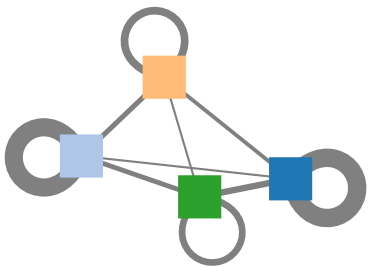
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 4$, $\mathcal{S} \approx 1246.9$ bits



Model, $\mathcal{L} \approx 693.2$ bits

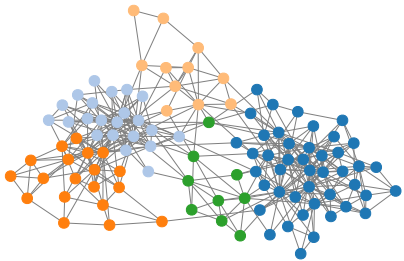
$\Sigma \approx 1940.2$ bits

Minimum description length (MDL)

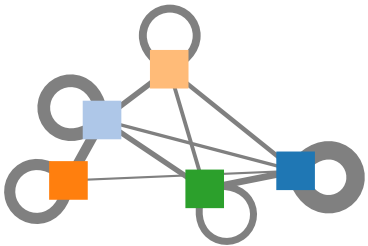
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 5$, $\mathcal{S} \approx 1195.2$ bits



Model, $\mathcal{L} \approx 747.9$ bits

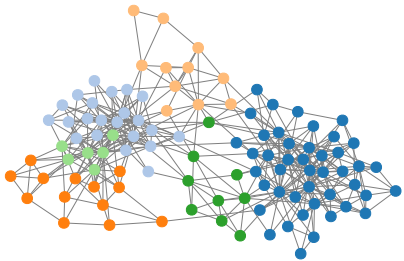
$\Sigma \approx 1943.2$ bits

Minimum description length (MDL)

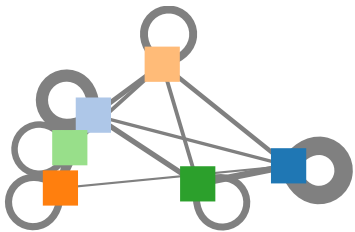
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 6$, $\mathcal{S} \approx 1165.1$ bits



Model, $\mathcal{L} \approx 780.7$ bits

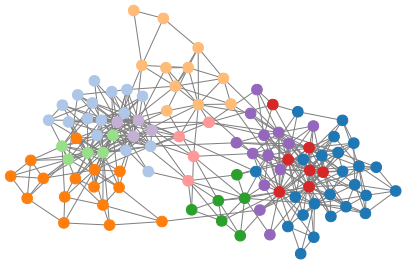
$\Sigma \approx 1951.9$ bits

Minimum description length (MDL)

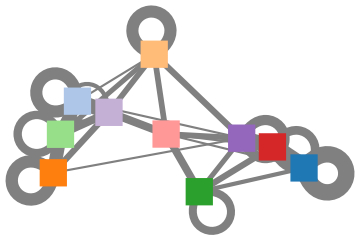
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 10$, $\mathcal{S} \approx 1028.4$ bits



Model, $\mathcal{L} \approx 942.7$ bits

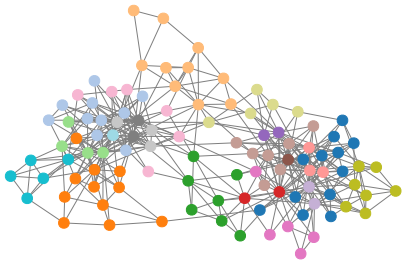
$\Sigma \approx 1971.1$ bits

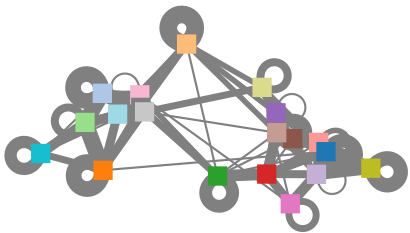
Minimum description length (MDL)

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$





$B = 20$, $\mathcal{S} \approx 808.9$ bits

Model, $\mathcal{L} \approx 1280.5$ bits

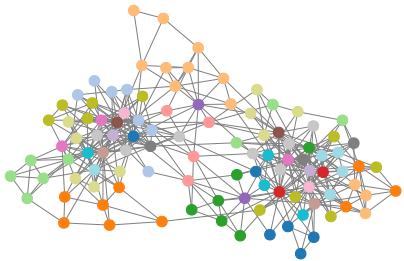
$\Sigma \approx 2089.5$ bits

Minimum description length (MDL)

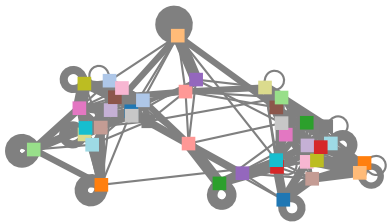
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 40$, $\mathcal{S} \approx 497.1$ bits



Model, $\mathcal{L} \approx 1748.4$ bits

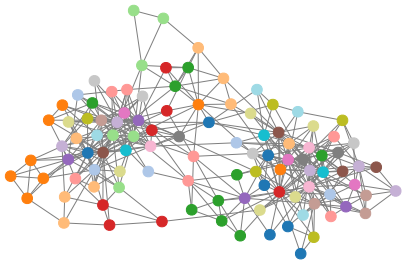
$\Sigma \approx 2245.6$ bits

Minimum description length (MDL)

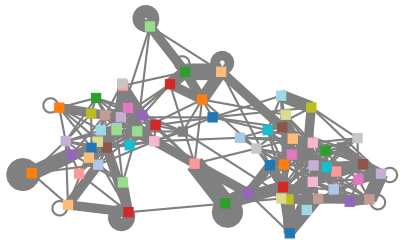
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = 70$, $\mathcal{S} \approx 178.5$ bits



Model, $\mathcal{L} \approx 2272.4$ bits

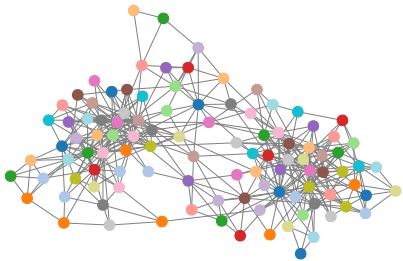
$\Sigma \approx 2450.9$ bits

Minimum description length (MDL)

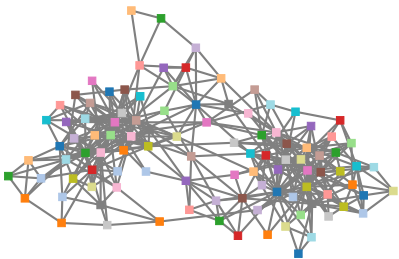
$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



$B = N$, $\mathcal{S} = 0$ bits



Model, $\mathcal{L} \approx 2568.2$ bits

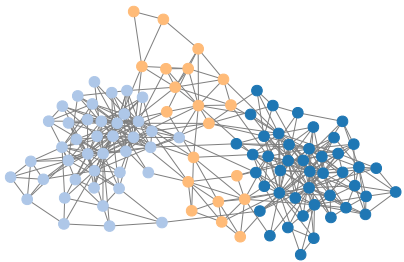
$\Sigma \approx 2568.2$ bits

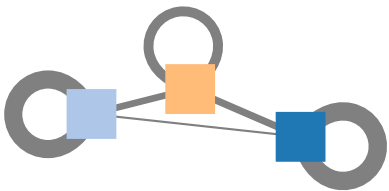
Minimum description length (MDL)

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}{P(\boldsymbol{A})} = \frac{2^{-\Sigma}}{P(\boldsymbol{A})}$$

Description length:

$$\Sigma = -\log_2 P(\boldsymbol{A},\boldsymbol{k},\boldsymbol{e},\boldsymbol{b}) = \underbrace{-\log_2 P(\boldsymbol{A}|\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{data|model, } \mathcal{S}} \underbrace{-\log_2 P(\boldsymbol{k},\boldsymbol{e},\boldsymbol{b})}_{\text{model, } \mathcal{L}}$$



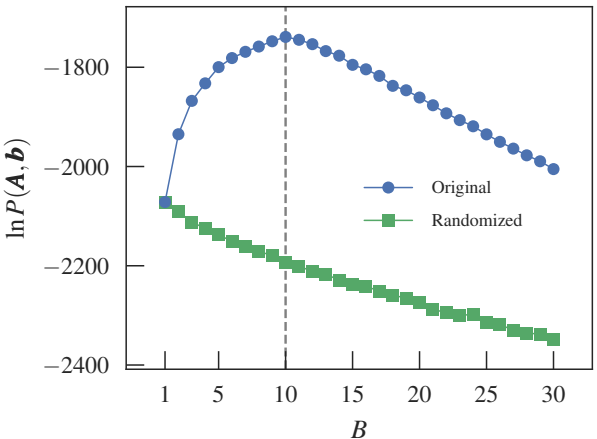


$B = 3$, $\mathcal{S} \approx 1285.7$ bits

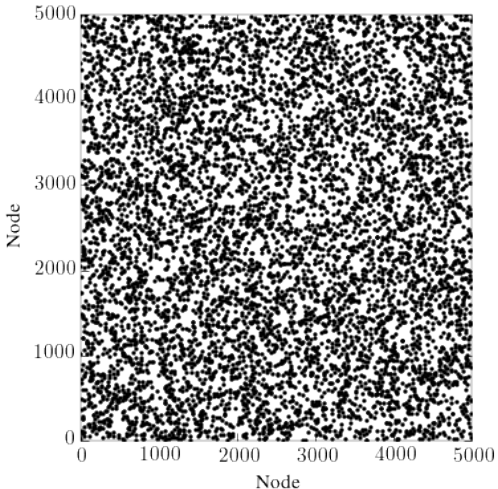
Model, $\mathcal{L} \approx 654.1$ bits

$\Sigma \approx 1939.7$ bits

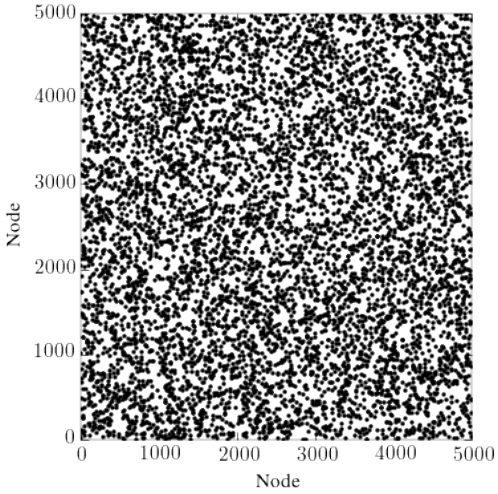
Example: American college football teams

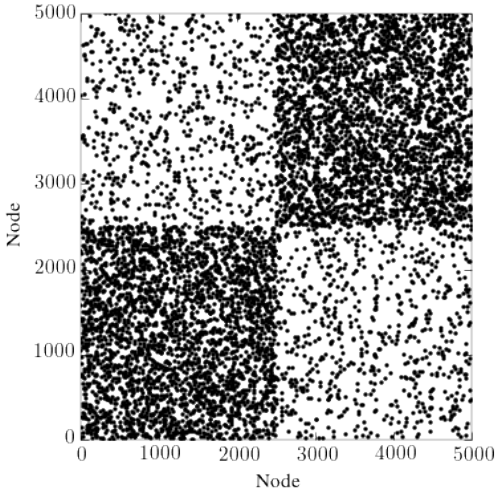


Structure vs. Randomness

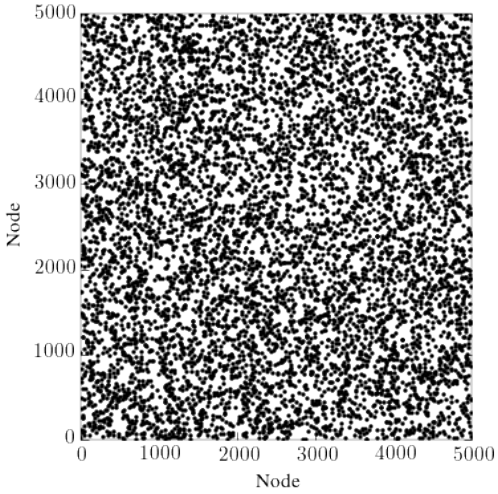


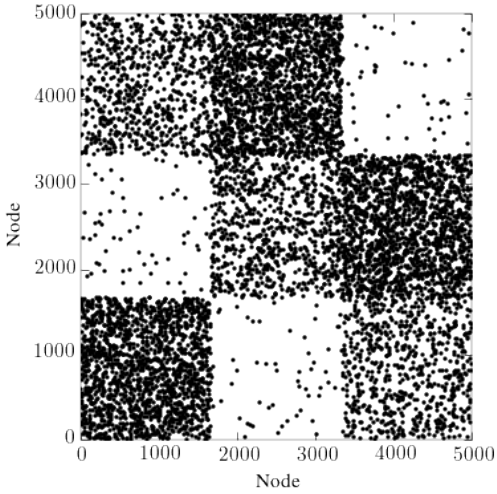
Structure vs. Randomness



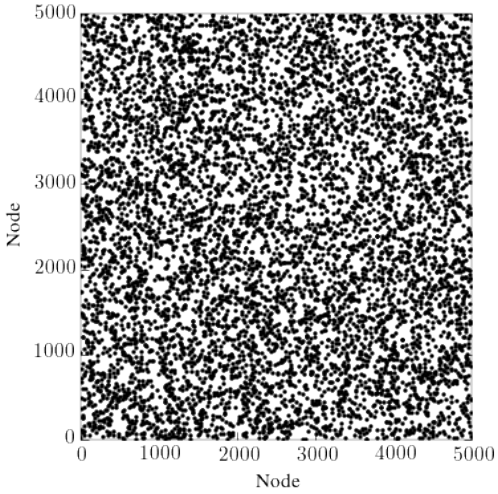


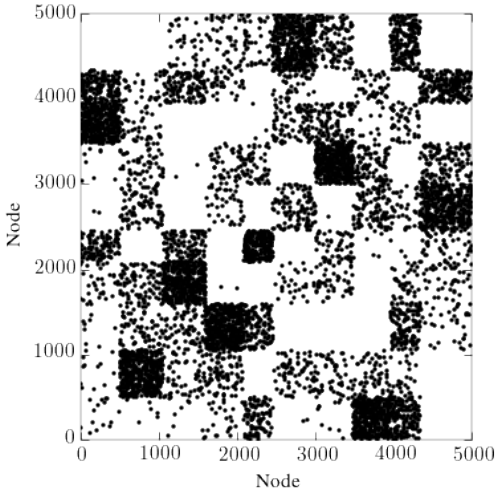
Structure vs. Randomness



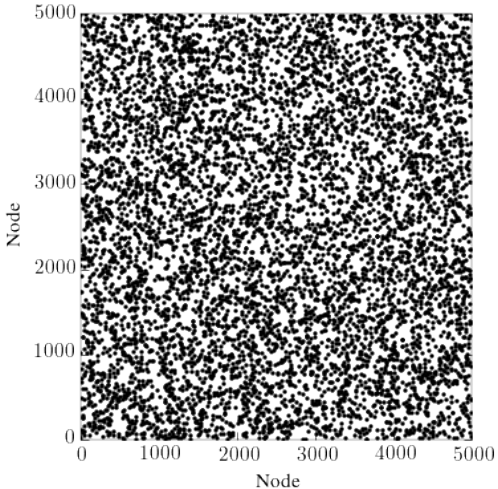


Structure vs. Randomness

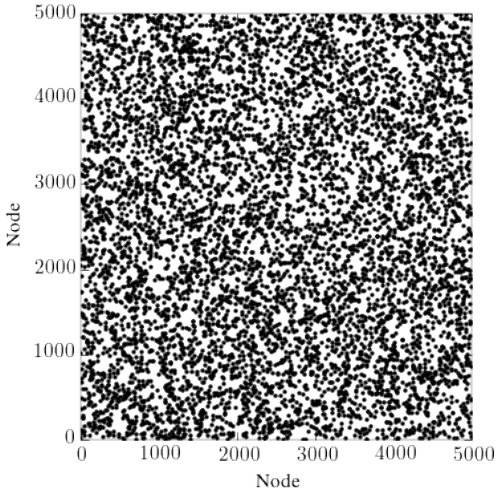




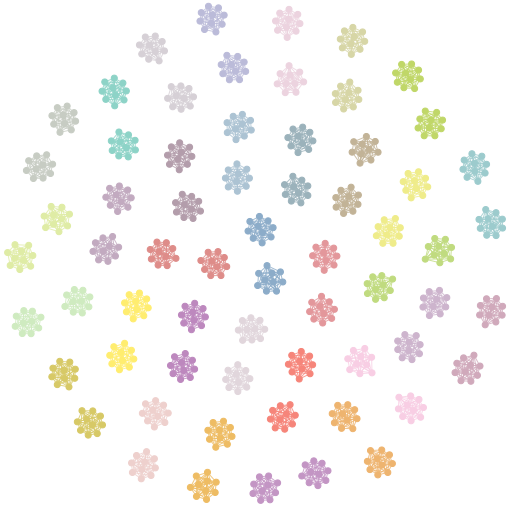
Structure vs. Randomness



Structure vs. Randomness



Homophily can induce triangles



$C \in [0,1] \to$ clustering coefficient (fraction of connected triads that form a triangle).

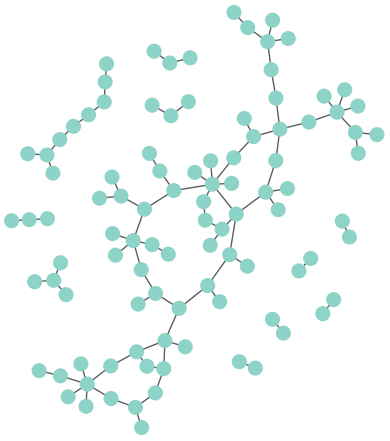
Assuming uniform group size and density:

$$C = O\left(\frac{\langle k \rangle B}{N}\right)$$

Large values of C can be obtained for globally sparse networks, with a sufficiently large B.

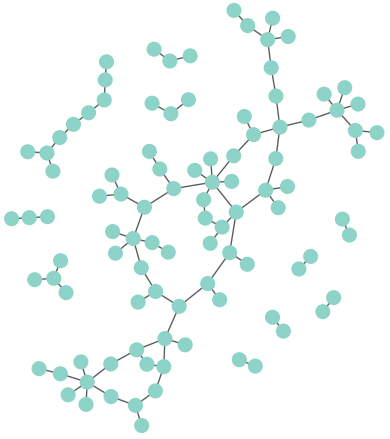
However, for B and k fixed, and $N \to \infty$, we have $C \to 0$.
(Typical real networks do not look like this.)

Triadic closure "induces" community structure

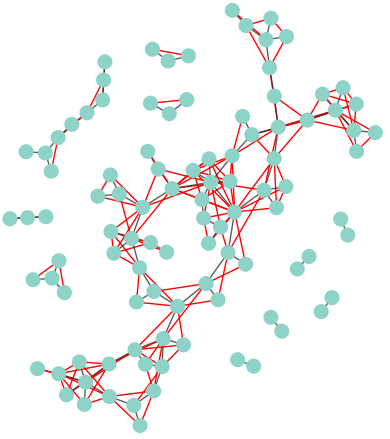


(a) Fully random network (no homophily)

Triadic closure "induces" community structure



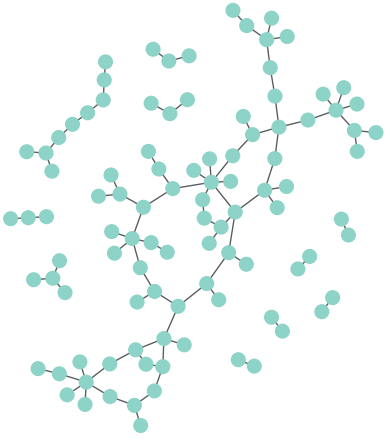
(a) Fully random network (no homophily)



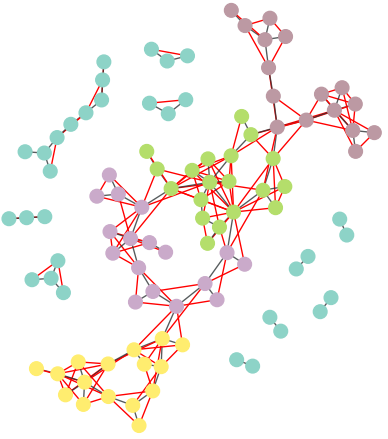
(b) Triadic closure edges

Foster *et al* PRE (2010), PRE (2011), Bianconi et al PRE (2014), Wharrie et al PRE (2019), Lopez and Coolen, arXiv:2008.11002

Triadic closure "induces" community structure



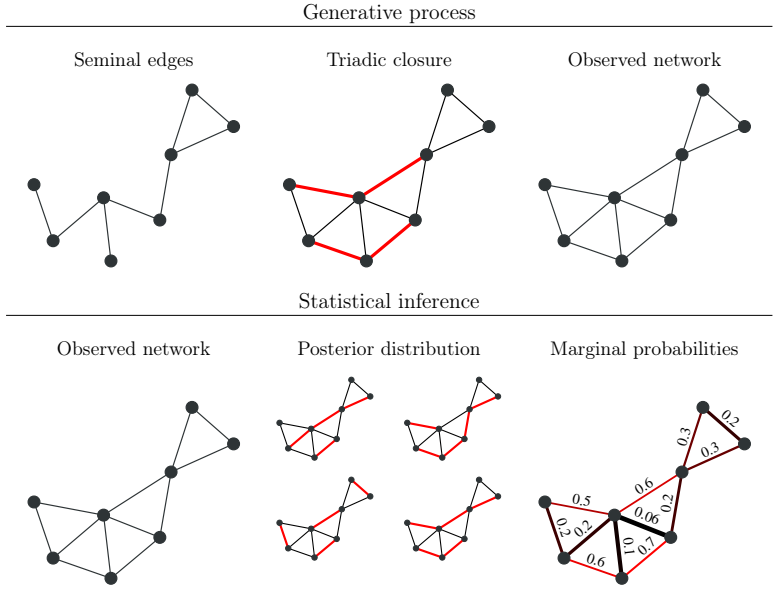
(a) Fully random network (no homophily)



(b) Triadic closure edges

Foster *et al* PRE (2010), PRE (2011), Bianconi et al PRE (2014), Wharrie et al PRE (2019), Lopez and Coolen, arXiv:2008.11002

A latent triadic closure model



SBM + triadic closure (SBM/TC)

Generative process

$$\underbrace{\boxed{P(\boldsymbol{A}|\boldsymbol{b})}}_{\text{SBM}} \longrightarrow \underbrace{\boxed{P(\boldsymbol{g}|\boldsymbol{A})}}_{\text{triadic closure}} \longrightarrow \underbrace{\boxed{P(\boldsymbol{G}|\boldsymbol{A},\boldsymbol{g})}}_{\text{final network}}$$

$\boldsymbol{b} \rightarrow$ node partition

$\boldsymbol{A} \rightarrow$ seminal edges

$\boldsymbol{g} \rightarrow$ triadic closure edges

$\boldsymbol{G} \rightarrow$ final network

Bayesian inference

$$P(\boldsymbol{g},\boldsymbol{A},\boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}|\boldsymbol{A},\boldsymbol{g})P(\boldsymbol{g}|\boldsymbol{A})P(\boldsymbol{A}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{G})}$$

Model selection & description length

Pure SBM

$$\begin{aligned}\Sigma_{\text{SBM}} &= -\ln P(\boldsymbol{G},\boldsymbol{b}) \\ &= \underbrace{-\ln P(\boldsymbol{G}|\boldsymbol{b})}_{\text{Data}}\underbrace{-\ln P(\boldsymbol{b})}_{\text{Model}}\end{aligned}$$

SBM/TC

$$\begin{aligned}\Sigma_{\text{SBM/TC}} &= -\ln P(\boldsymbol{G},\boldsymbol{A},\boldsymbol{g},\boldsymbol{b}) \\ &= -\ln P(\boldsymbol{G}|\boldsymbol{A},\boldsymbol{g}) - \ln P(\boldsymbol{g}|\boldsymbol{A}) \\ &\quad - \ln P(\boldsymbol{A}|\boldsymbol{b}) - \ln P(\boldsymbol{b})\end{aligned}$$

How do we close triangles?

Given a network $\boldsymbol{A}$ sampled from the SBM, with probability $P(\boldsymbol{A}|\boldsymbol{b})$.

To each node u we consider its "ego" graph $\boldsymbol{g}(u)$ with adjacency matrix $g_{ij}(u)$.

An open triad (i, u, j) exists in $\boldsymbol{A}$ iff $m_{ij}(u) = 1$, with

$$m_{ij}(u) = A_{ui}A_{uj}(1 - A_{ij}).$$

Each open triad is closed with probability p_u, leading to

$$P(\boldsymbol{g}(u)|\boldsymbol{A}, p_u) = \prod_{i<j} [p_u m_{ij}(u)]^{g_{ij}(u)} \left[1 - p_u m_{ij}(u)\right]^{1-g_{ij}(u)}.$$

We then erase the possible multiple edges from the union of all $\boldsymbol{g}(u)$, yielding a final graph $\boldsymbol{G}$:

$$G_{ij}(\boldsymbol{A}, \boldsymbol{g}) = \begin{cases} 1, & \text{if } A_{ij} + \sum_u g_{ij}(u) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

For all ego graphs we can compute the marginal,

$$\begin{aligned} P(\boldsymbol{g}|\boldsymbol{A}) &= \prod_u \int_0^1 P(\boldsymbol{g}(u)|\boldsymbol{A}, p) P(p)\, \mathrm{d}p \\ &= \prod_u \left[\binom{\sum_{i<j} m_{ij}(u)}{\sum_{i<j} g_{ij}(u)}^{-1} \frac{1}{1 + \sum_{i<j} m_{ij}(u)} \right], \end{aligned}$$

which give us the posterior distribution:

$$\boxed{P(\boldsymbol{g}, \boldsymbol{A}, \boldsymbol{b}|\boldsymbol{G}) = \frac{P(\boldsymbol{G}|\boldsymbol{A}, \boldsymbol{g}) P(\boldsymbol{g}|\boldsymbol{A}) P(\boldsymbol{A}|\boldsymbol{b}) P(\boldsymbol{b})}{P(\boldsymbol{g})}}$$

Given a final graph $\boldsymbol{G}$, we can then recover the seminal network $\boldsymbol{A}$, its partition $\boldsymbol{b}$, as well as the ego graphs $\boldsymbol{g}$ which contain the triadic closure edges.

Iterated triadic closures

An edge added to close an open triad may **create more open triads.**

We can then iterate triadic closures at successive generations l,

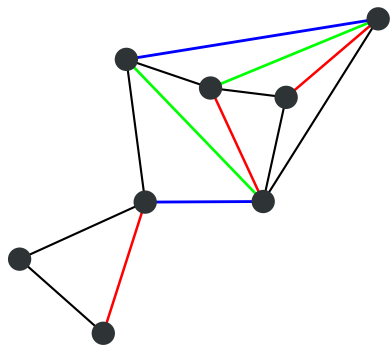
$$P(\boldsymbol{g}^{(l)}|\boldsymbol{A}^{(l-1)},\boldsymbol{g}^{(l-1)}) = \prod_u \left[\binom{\sum_{i<j} m_{ij}^{(l)}(u)}{\sum_{i<j} g_{ij}^{(l)}(u)}^{-1} \frac{1}{1+\sum_{i<j} m_{ij}^{(l)}(u)} \right].$$

This gives us a flexible model that can capture large triangle densities.

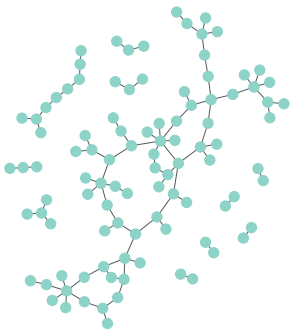
In this way, we have a joint posterior for all triadic closure generations

$$P(\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\},\boldsymbol{b}|\boldsymbol{g}) = \frac{P(\boldsymbol{g},\{\boldsymbol{g}^{(l)}\},\{\boldsymbol{A}^{(l)}\}|\boldsymbol{b})P(\boldsymbol{b})}{P(\boldsymbol{g})}.$$

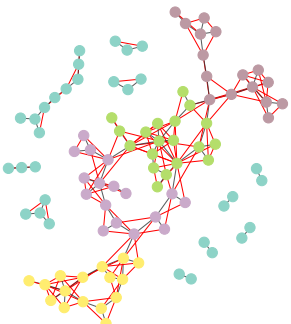
To facilitate identifiability, a new generation can only close a triad that involves at least one edge of the previous generation:



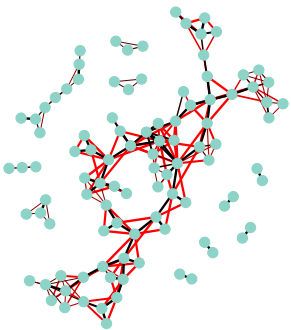
Distinguishing triadic closure from community structure



(a) Random seminal edges

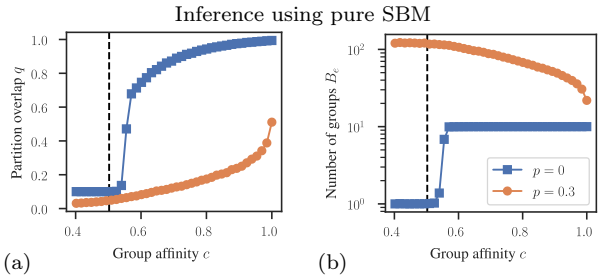


(b) Triadic closure edges and spurious communities found with SBM ($\Sigma_{\mathrm{SBM}} = 801.7$ nats)



(c) Inference of the SBM/TC model ($\Sigma_{\mathrm{SBM/TC}} = 590.6$ nats)

Artificial networks



Planted Partition model with triadic closure

The seminal graph $\boldsymbol{A}$ is generated with edge counts

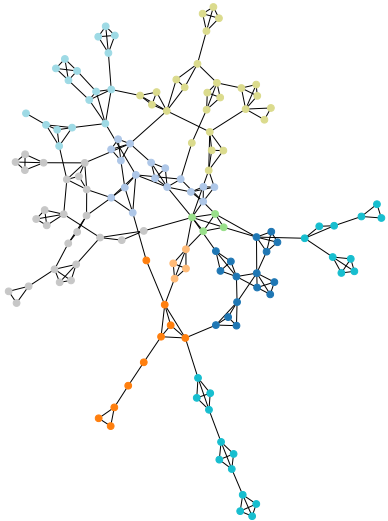
$$e_{rs} = 2E\left[\frac{c}{B}\delta_{rs} + \frac{1-c}{B(B-1)}(1-\delta_{rs})\right],$$

where $c \in [0, 1]$ is the group affinity.

The final graph $\boldsymbol{G}$ is obtained with one round of triadic closures on $\boldsymbol{A}$ with $p_u = p$.

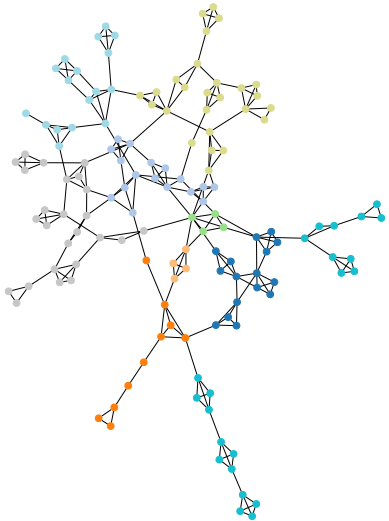
Left: $N = 10^4$, $\langle k \rangle = 5$, $B = 10$

Empirical network: Student cooperations

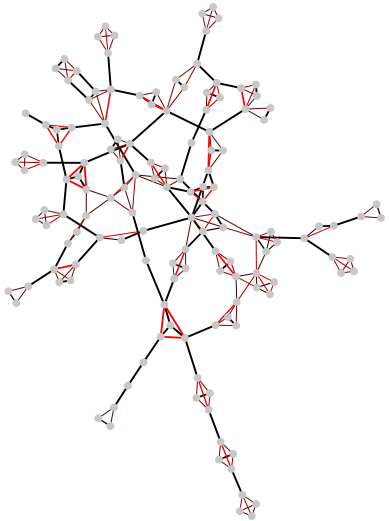


(a) SBM, $B = 9$, $\Sigma_{\text{SBM}} = 1145.6$ nats

Empirical network: Student cooperations

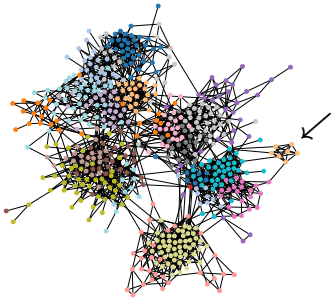


(a) SBM, $B = 9$, $\Sigma_{\mathrm{SBM}} = 1145.6$ nats



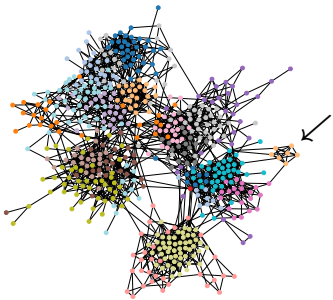
(b) SBM/TC, $\Sigma_{\mathrm{SBM/TC}} = 935.1$ nats

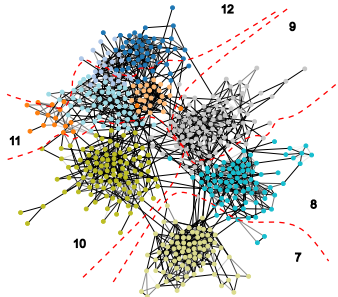
High-school student friendships



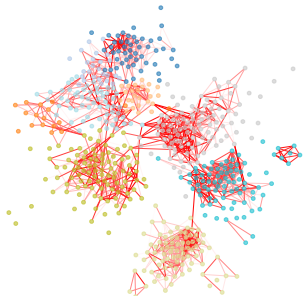
(a) SBM, $B = 26$, $\Sigma_{\text{SBM}} = 8757.0$ nats

High-school student friendships





Seminal edges

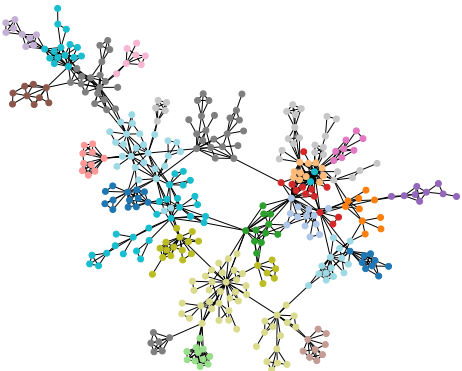


Triadic closure edges

(a) SBM, $B = 26$, $\Sigma_{\text{SBM}} = 8757.0$ nats

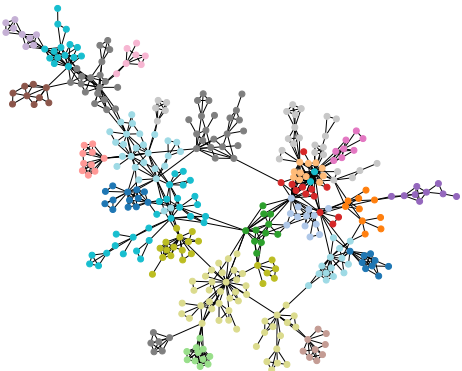
(b) SBM/TC, $B = 9$, $\Sigma_{\text{SBM/TC}} = 8456.3$ nats

Collaborations between network scientists

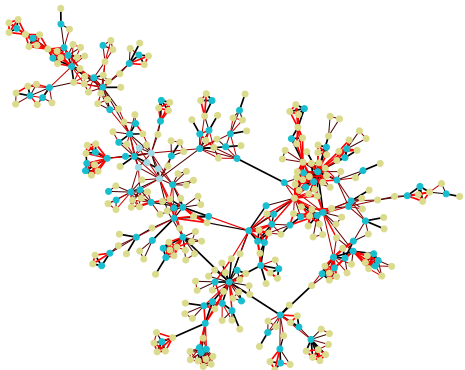


(c) SBM, $B = 27$, $\Sigma_{\text{SBM}} = 3816.3$ nats

Collaborations between network scientists

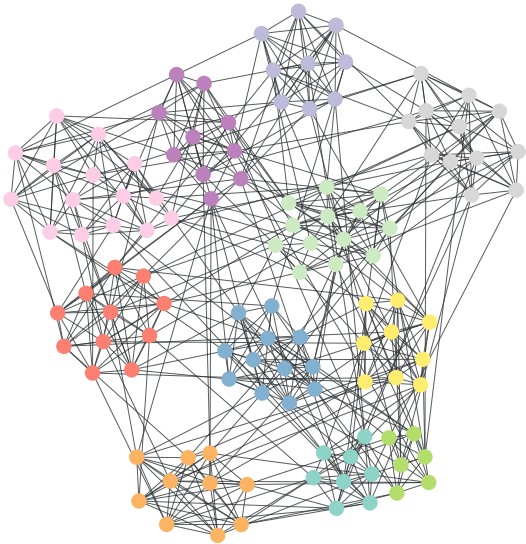


(c) SBM, $B = 27$, $\Sigma_{\text{SBM}} = 3816.3$ nats



(d) SBM/TC, $B = 3$, $\Sigma_{\text{SBM/TC}} = 3009.9$ nats

Games between American football teams



$B = 11$

$\Sigma_{\text{SBM}} = 1761.1$ nats

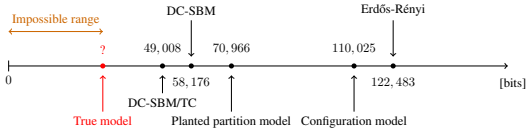
$\Sigma_{\text{SBM/TC}} = 1767.6$ nats

$C = 0.41$

Model selection and finding the "truth"

Protein-Protein interaction (wild turkey)

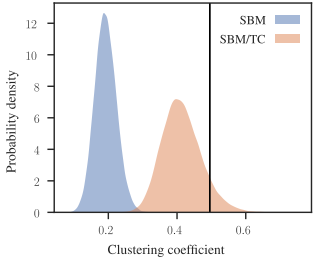
Description length



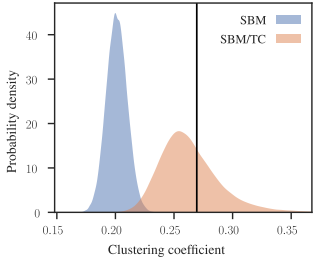
Compression points *toward* what is detectable of the true generative process.

SBM/TC yields plausible clustering values

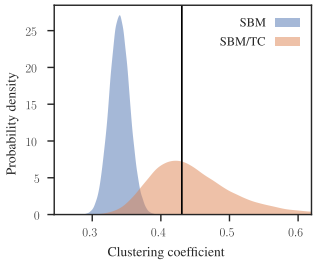
Posterior predictive checks: we sample networks from the fitted model and compare to the data.



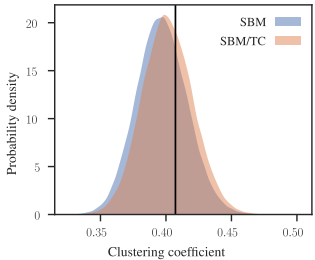
(a) Cooperation between students



(b) Adolescent health (comm26)

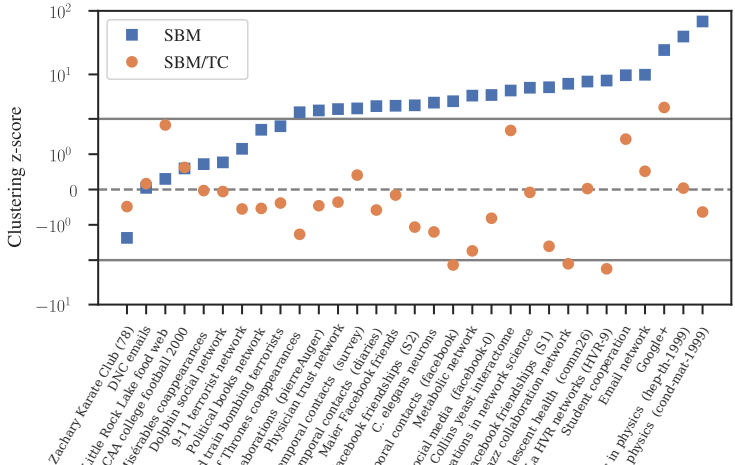


(c) Scientific collaborations in Network Science

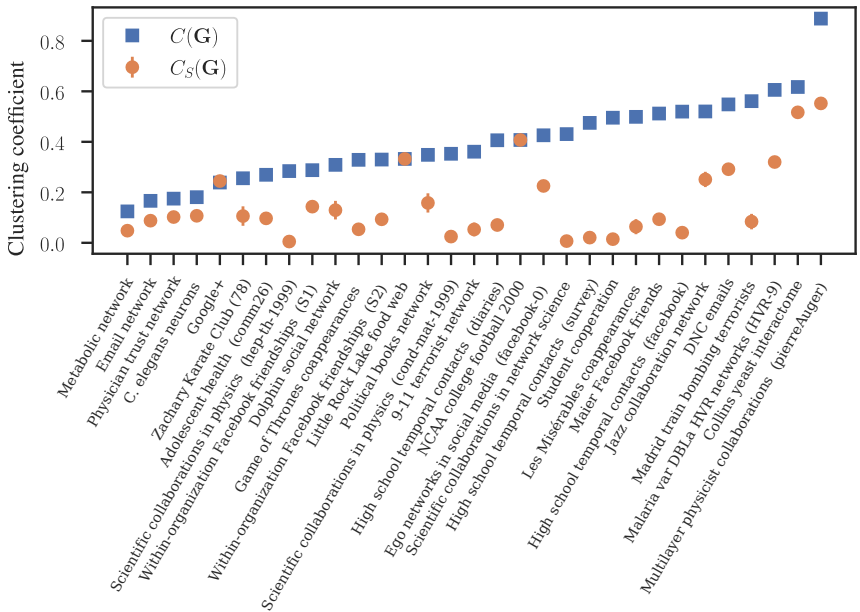


(d) NCAA college football 2000

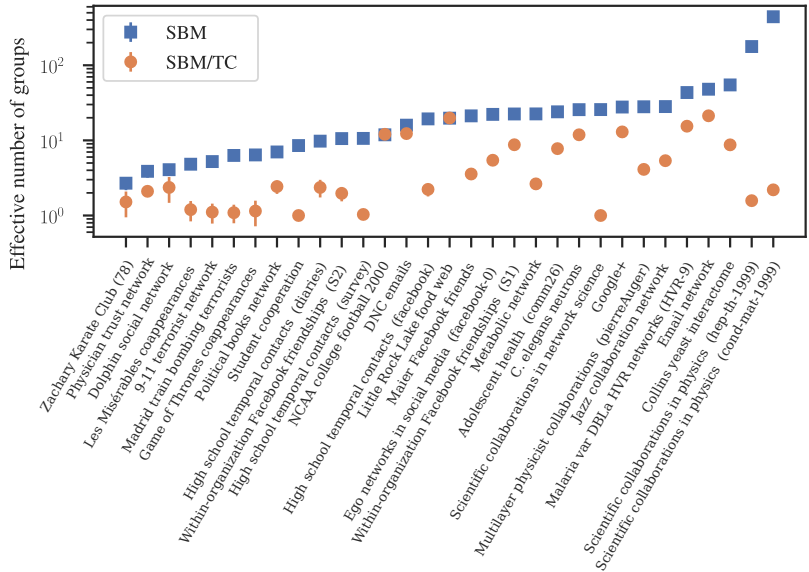
Capturing the clustering coefficient



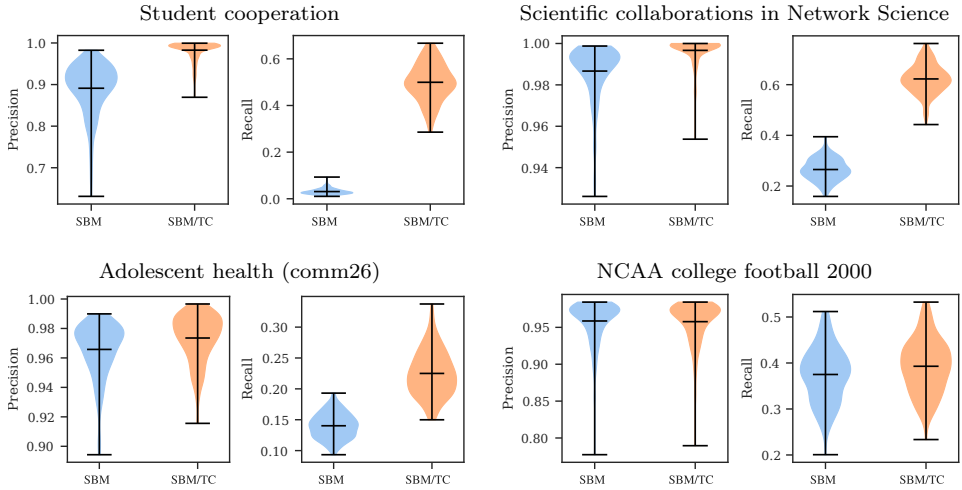
Decomposing the clustering coefficient



Community structure vs. triadic closure



SBM/TC is more predictive of missing edges



Summary

The SBM + triadic closure model allows us to

- Separate the effect of homophily from triadic closure
- Identify triadic closure edges
- Identify communities that *cannot* be explained by triadic closure
- Improve compression
- Improve link prediction

Paper: T.P.P., Phys. Rev. X 12, 011004 (2022)

Code: `https://graph-tool.skewed.de`