Computational Lower Bounds for Community Detection on Random Graphs

B. Hajek, Y. Wu, J. Xu

http://proceedings.mlr.press/v40/Hajek15.pdf

2015

Presented by Neven Villani

2024-02-26

Introduction and problem statement

Context

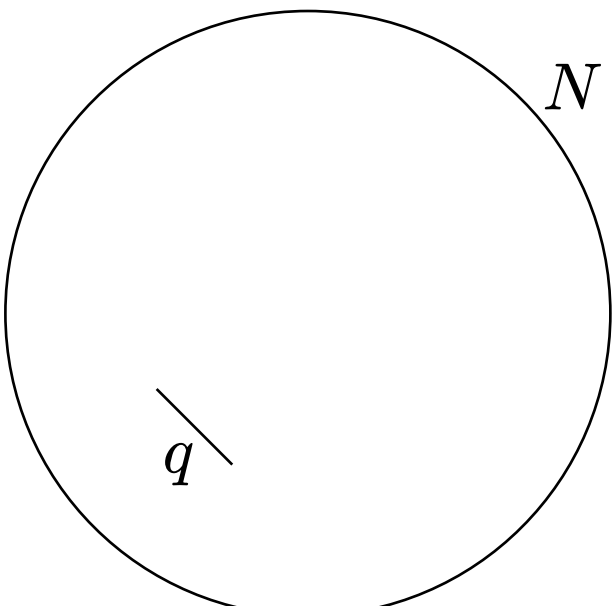
Erdös-Rényi graph $\mathcal{G}(N, q)$:

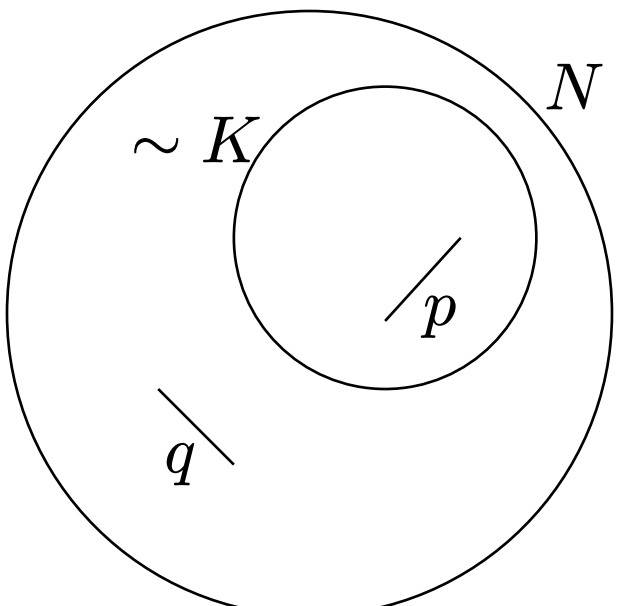
- N vertices
- each pair connected independently with probability q

Community:

- subset of the vertices
- higher connectivity

Erdös-Rényi with Planted Dense Subgraph (PDS)





Erdös-Rényi with Planted Dense Subgraph (PDS)

$\mathcal{G}(N, K, p, q)$ defined as

- N vertices
- of which a subset S is distinguished
  $n \in [N]$ belongs to S independently with probability $\frac{K}{N}$
- $n, n' \in [N]$ are connected independently of other pairs
  - with probability $p > q$ if both are in S
  - with probability q otherwise

Remarks and hypotheses

- N, K, p, q are known
- $p = cq$ for some known fixed $c > 1$
- $\mathbb{E}(|S|) = K$
- $\mathbb{E}(|V|) = \binom{N}{2} q + \binom{K}{2}(p - q)$

Community detection

Test φ s.t.

Input: a random graph $G \sim \mathcal{G}(N, q)$ (H_0) or $G \sim \mathcal{G}(N, K, p, q)$ (H_1)

Output: 0 or 1

Goal: minimize the error

$$\mathbb{P}_{(H_0)}(\varphi(G) = 1) + \mathbb{P}_{(H_1)}(\varphi(G) = 0)$$

Statistical tests

Linear test

Count the total number of edges

$$T_{\text{lin}} := \sum_{i<j} A_{i,j}$$

Which is expected to be

- $\binom{N}{2} q$ under (H_0)
- $\binom{N}{2} q + \binom{K}{2}(p-q)$ under (H_1)

Linear test

Count the total number of edges

$$T_{\text{lin}} := \sum_{i<j} A_{i,j}$$

Answer 1 iff

$$T_{\text{lin}} > \tau_{\text{lin}} := \binom{N}{2} q + \binom{K}{2} \frac{p-q}{2}$$

Scan test

Count the edges in the densest K-subgraph

$$T_{\text{scan}} := \max_{S' :\, |S'| = K} \sum_{i<j \in S'} A_{i,j}$$

Which is expected to be

- $\binom{K}{2} q$ under (H_0)
- $\binom{K}{2} p$ under (H_1)

Scan test

Count the edges in the densest K-subgraph

$$T_{\text{scan}} := \max_{S' \colon |S'| = K} \sum_{i<j \in S'} A_{i,j}$$

Answer 1 iff

$$T_{\text{scan}} > \tau_{\text{scan}} := \binom{K}{2} \frac{p+q}{2}$$

Success conditions

It is possible to…

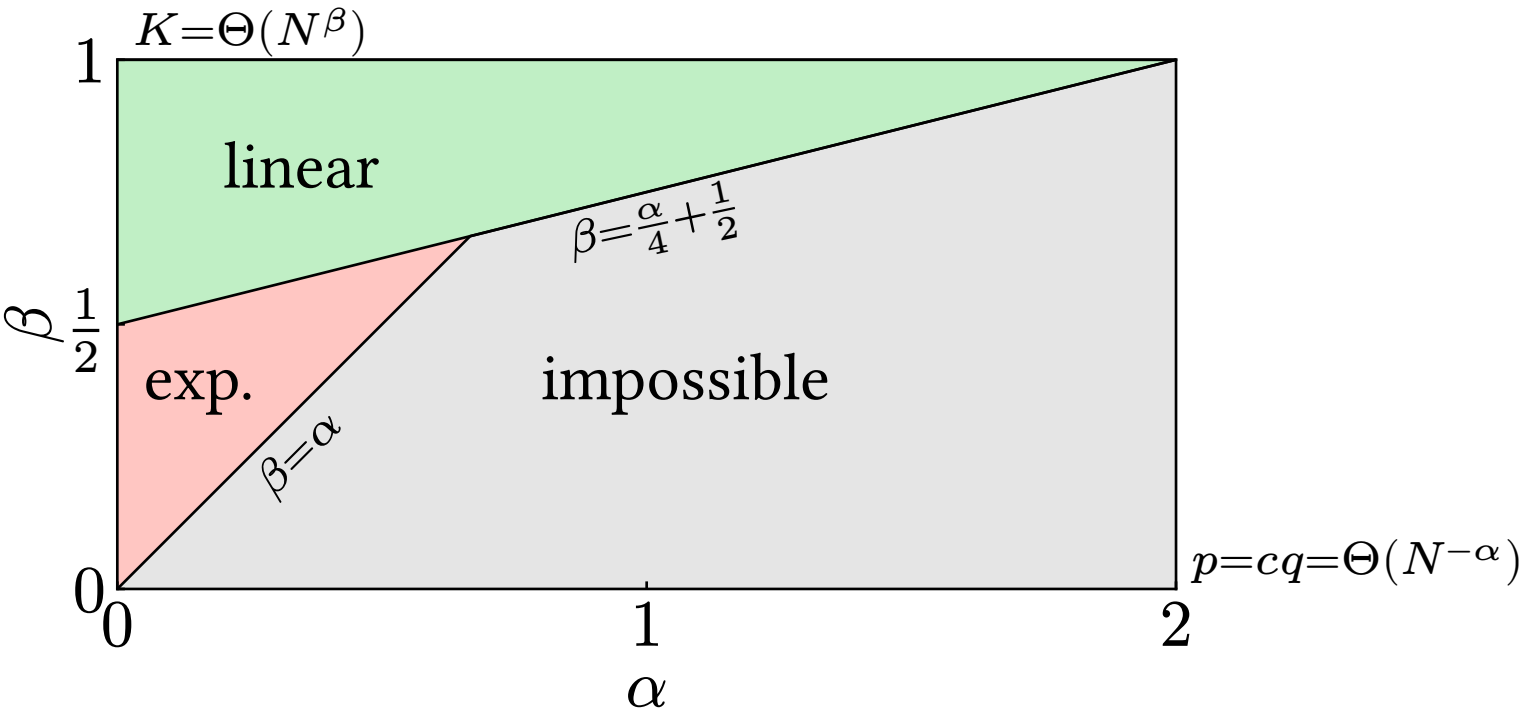
- upper bound $\mathbb{P}_{(H_0)}(T_{\text{lin}} > \tau_{\text{lin}}) + \mathbb{P}_{(H_1)}(T_{\text{lin}} \leq \tau_{\text{lin}})$ by a term that decreases exponentially with $\frac{K^4 q}{N^2}$
- upper bound $\mathbb{P}_{(H_0)}(T_{\text{scan}} > \tau_{\text{scan}}) + \mathbb{P}_{(H_1)}(T_{\text{scan}} \leq \tau_{\text{scan}})$ by a term that decreases exponentially with $K \log \frac{Ne}{K} - K^2 q$
- lower bound $\min_{\varphi} \mathbb{P}_{(H_0)}(\varphi(G) = 1) + \mathbb{P}_{(H_1)}(\varphi(G) = 0)$ close to 1 when $q = O\Big(\min\Big(\frac{1}{K} \log \frac{Ne}{K}, \frac{N^2}{K^4}\Big)\Big)$

Success condition under asymptotic regime

$$q = \frac{1}{c}p = N^{-\alpha},\ K = \Theta\left(N^{\beta}\right),\ N \to +\infty$$

- plugging the above in $\frac{K^4 q}{N^2} \to +\infty$ gives as a condition $\beta > \frac{\alpha}{4} + \frac{1}{2}$
- while $K \log \frac{Ne}{K} - K^2 q \to -\infty$ requires $\beta > \alpha$
- the impossibility lower bound is thus tight in this regime

Landscape (asymptotic regime)



A hardness lower bound through a reduction

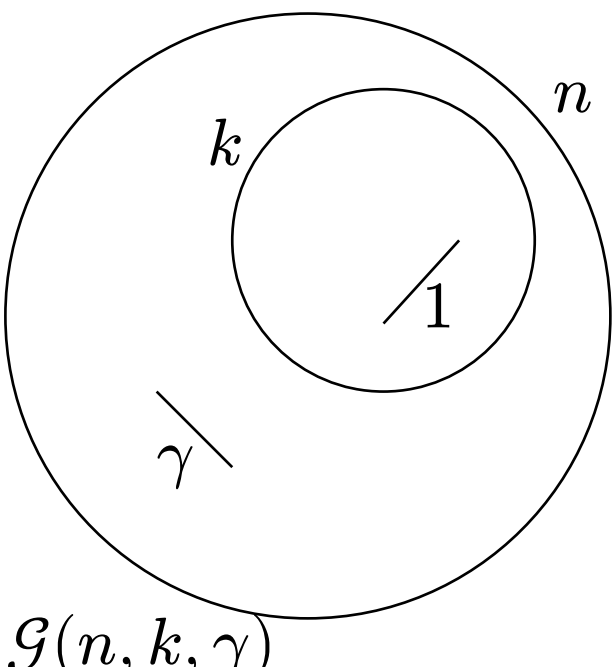
Planted Clique (PC)

$\mathcal{G}(n, k, \gamma)$

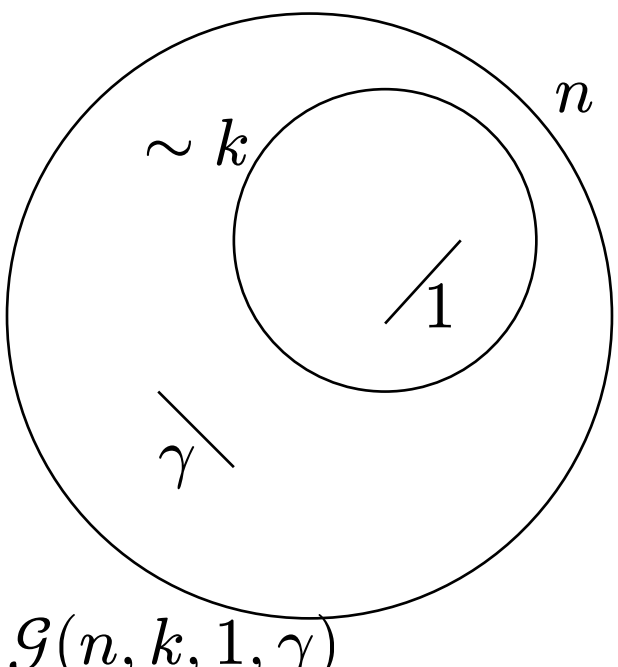
- Erdös-Rényi graph $\mathcal{G}(n, \gamma)$
- of which a subgraph S of size **exactly** k is randomly chosen
- S is turned into a clique

PC vs PDS

Remark: $\mathcal{G}(n, k, \gamma)$ is *almost* $\mathcal{G}(n, k, 1, \gamma)$



$\mathcal{G}(n, k, \gamma)$



$\mathcal{G}(n, k, 1, \gamma)$

PC Hypothesis

(parameterized by $0 < \gamma \leq \frac{1}{2}$)

Conjecture:

There is no polynomial test that can distinguish between

- (H_0): $\mathcal{G}(n, \gamma)$, and
- (H_1): $\mathcal{G}(n, k, \gamma)$

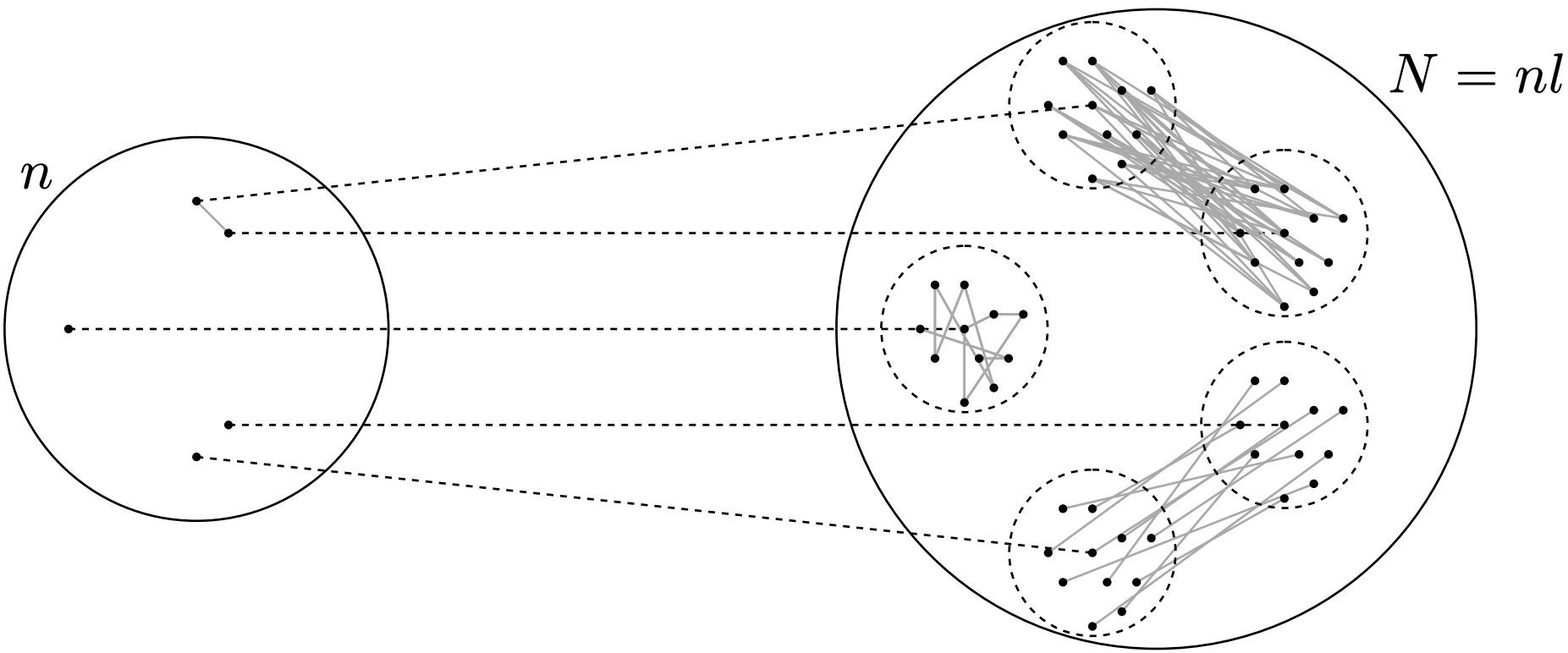
when $k = o(\sqrt{n})$

The reduction

Known: $n, k, \gamma, N = ln, K = lk, q, p = cq$
Input: adjacency matrix A of a graph that is either $\mathcal{G}(n, \gamma)$ or $\mathcal{G}(n, k, \gamma)$
Output: $\tilde{A}$ close in distribution to $\mathcal{G}(N, q)$ or $\mathcal{G}(N, K, p, q)$ resp.

Remark: impossible to map *exactly* $\mathcal{G}(n, k, \gamma)$ to $\mathcal{G}(N, K, p, q)$, but we can get close enough when averaging over a certain set.

Intuition



Main steps

- randomly partition $[n]$ into sets $\{V_i\}_{i\in[n]}$
- within V_i add $\mathrm{Binom}\left(\binom{|V_i|}{2}, q\right)$ edges
- between V_i and V_j add
  - $P_{i,j}$ edges if i and j are connected
  - $Q_{i,j}$ edges if they are not

  for well-chosen distributions $P_{i,j}$ and $Q_{i,j}$ s.t.
  - $(1-\gamma)Q_{i,j} + \gamma P_{i,j}$ is exactly $\mathrm{Binom}\big(|V_i|\ |V_j|, q\big)$
  - $P_{i,j}$ is close to $\mathrm{Binom}\big(|V_i|\ |V_j|, p\big)$

Key properties

- under the null hypothesis of PC, the resulting graph is distributed according to $\mathcal{G}(N, q)$
- under the alternative distribution of PC, the output is indistinguishable from $\mathcal{G}(N, K, p, q)$ when averaging over the random partition

$\rightarrow$ a polynomial test for PDS would give a polynomial test for PC.

Conclusion

Conclusion

- Community detection modeled as distinguishing $\mathcal{G}(N, q)$ from $\mathcal{G}(N, K, p, q)$
- easy when community is large or graph is dense
- provably impossible when graph is sparse and community is small
- combinatorial algorithm when community is small and graph is dense
  - optimal under a conjectured hardness result of a similar well-studied problem