Shotgun threshold for sparse Erdős-Rényi graphs

Heng Ma

Peking Univertisy

Based on the joint work with
Jian Ding (Peking University)

Yiyang Jiang (Peking University)

Identifying graphs

- **Reconstruction Conjecture** (Kelly, Ulam, Harary' 57): Any two graphs on 3 or more vertices that have the same multi-set of vertex-deleted subgraphs are isomorphic.

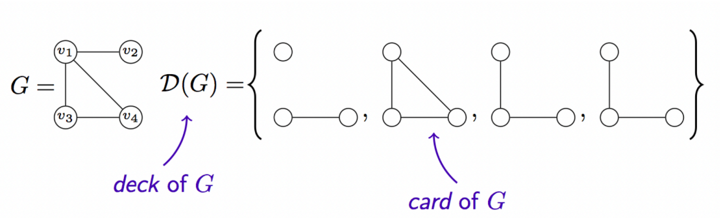
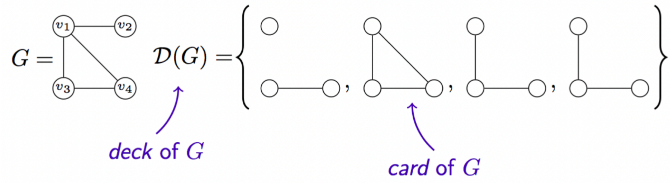
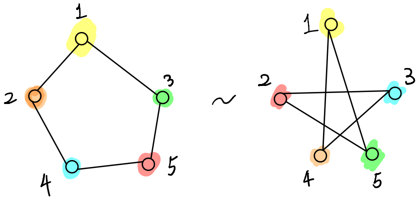
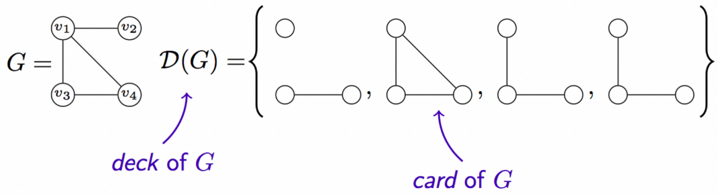


Figure 1: From Topology and Combinatorics Blog by Max F. Pitz

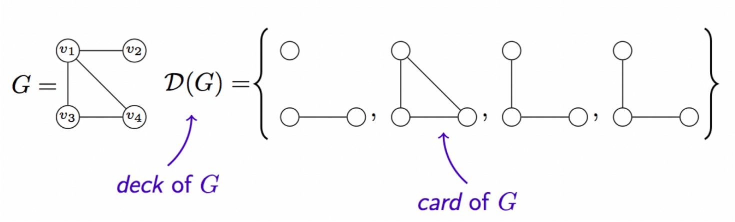
- **Reconstruction Conjecture**: Any two graphs on 3 or more vertices that have the same multi-set of vertex-deleted subgraphs are isomorphic.





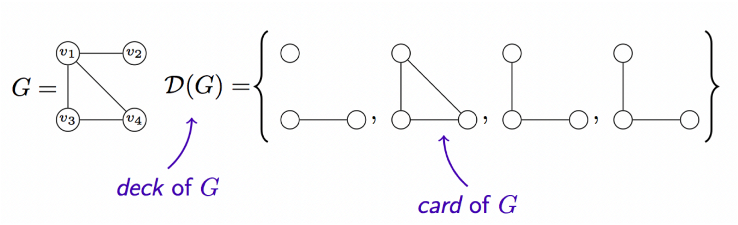


What if the underlying Graphs are random?



What if the underlying Graphs are random?

Bollobs'90: almost all graphs can be reconstructed from any 3 vertex-deleted subgraphs.



What if the underlying Graphs are random?

Bollobs'90: almost all graphs can be reconstructed from any 3 vertex-deleted subgraphs.

If the graph is random, but we are only given the very local information of each vertex, can we still identify the graph?

Motivating examples

- DNA shotgun sequencing: Reconstruct a DNA sequence from "shotgunned stretches of the sequence.

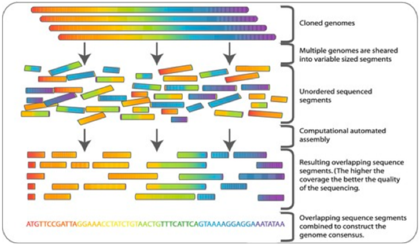


Figure 2: DNA shotgun sequencing by Commins, Toft, and Fares

Motivating examples

- DNA shotgun sequencing: Reconstruct a DNA sequence from “shotgunned stretches of the sequence.

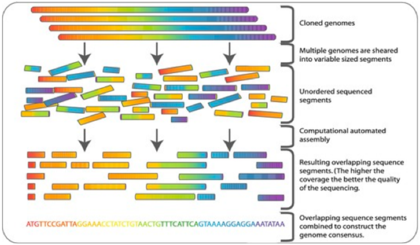


Figure 2: DNA shotgun sequencing by Commins, Toft, and Fares

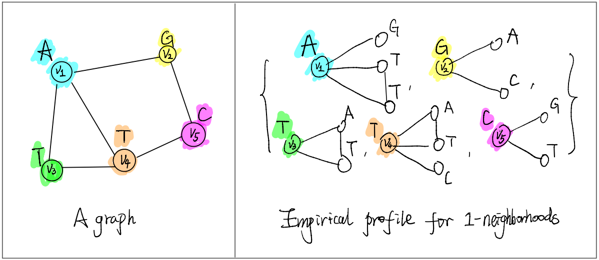
- **Paninski et al’15**: Reconstruct a big neural network from very local subnetworks that are observed in experiments.

Mathematical framework by Mossel-Ross'19

- Model: $\mathcal{G}$ is a (fixed or random) graph, possibly with random labeling of the vertices.
- Observation: For each vertex v, we are given its local r-neighborhood $\mathsf{N}_r(v)$: the subgraph induced by the vertices (forgetting their names) at distance no greater than r from v.

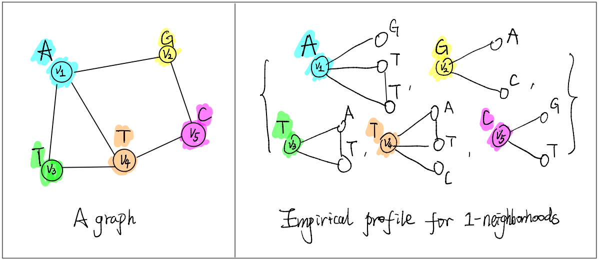
Mathematical framework by Mossel-Ross'19

- Model: $\mathcal{G}$ is a (fixed or random) graph, possibly with random labeling of the vertices.
- Observation: For each vertex v, we are given its local r-neighborhood $\mathsf{N}_r(v)$: the subgraph induced by the vertices (forgetting their names) at distance no greater than r from v.



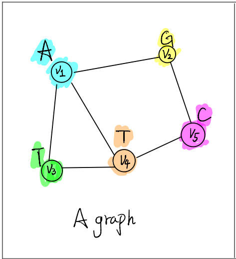
Mathematical framework by Mossel-Ross'19

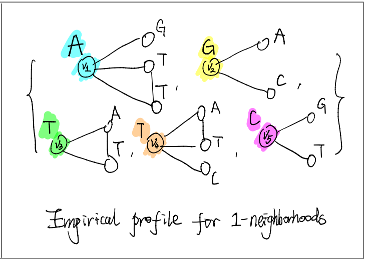
- Model: $\mathcal{G}$ is a (fixed or random) graph, possibly with random labeling of the vertices.
- Observation: For each vertex v, we are given its local r-neighborhood $\mathsf{N}_r(v)$: the subgraph induced by the vertices (forgetting their names) at distance no greater than r from v.
- Question: Can we identify $\mathcal{G}$ (up to isomorphism) from the empirical profile for r-neighborhoods $\{\mathsf{N}_r(v) : v \in \mathcal{G}\}$?



Mathematical framework by Mossel-Ross'19

- Model: $\mathcal{G}$ is a (fixed or random) graph, possibly with random labeling of the vertices.
- Observation: For each vertex v, we are given its local r-neighborhood $\mathsf{N}_r(v)$: the subgraph induced by the vertices (forgetting their names) at distance no greater than r from v.
- Question: Can we identify $\mathcal{G}$ (up to isomorphism) from the empirical profile for r-neighborhoods $\{\mathsf{N}_r(v) : v \in \mathcal{G}\}$?
- There is a shotgun (assembly) threshold r_* for the radius r since the monotonicity.

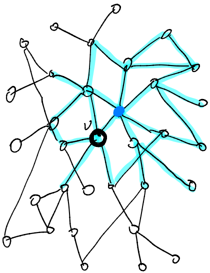




Mossel-Ross'19:

- Identifiability: Uniqueness of overlaps

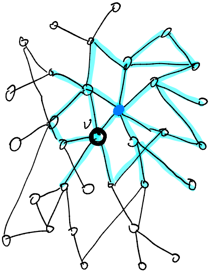
$$r_* \leq \min\{k : N_k(u) \neq N_k(v)\ \forall\, u, v\} + 1$$



Mossel-Ross'19:

- Identifiability: Uniqueness of overlaps

$$r_* \leq \min\{k : N_k(u) \neq N_k(v) \; \forall\, u, v\} + 1$$



- Non-identifiability: Blocking configurations.

Labeled lattice models

Graph: d-dimensional box of side length n, denoted as Λ_n.

Labels: i.i.d. uniform vertex labels from $\{1, \cdots, q\}$.

Observations: vertex labeling configurations for each r-box contained in Λ_n.

Labeled lattice models

Graph: d-dimensional box of side length n, denoted as Λ_n.

Labels: i.i.d. uniform vertex labels from $\{1, \cdots, q\}$.

Observations: vertex labeling configurations for each r-box contained in Λ_n.

Mossel-Ross'19: For any $\epsilon > 0$, as $n \to \infty$, with probability tending to 1

$$(1-\epsilon)\frac{d/2^{d-1}}{\log q}\log n \le (r_*)^d \le (1+\epsilon)\frac{2d}{\log q}\log n \,.$$

Labeled lattice models

Graph: d-dimensional box of side length n, denoted as Λ_n.

Labels: i.i.d. uniform vertex labels from $\{1, \cdots, q\}$.

Observations: vertex labeling configurations for each r-box contained in Λ_n.

Mossel-Ross'19: For any $\epsilon > 0$, as $n \to \infty$, with probability tending to 1

$$(1-\epsilon)\frac{d/2^{d-1}}{\log q}\log n \le (r_*)^d \le (1+\epsilon)\frac{2d}{\log q}\log n\,.$$

Ding-Liu'22+:

$$(1-\epsilon)\frac{2}{\log q}\log n \le r_* \le (1+\epsilon)\frac{2}{\log q}\log n \quad \text{when } d = 1;$$

$$(1-\epsilon)\frac{d}{\log q}\log n \le (r_*)^d \le (1+\epsilon)\frac{d}{\log q}\log n \quad \text{when } d \ge 2.$$

Random d-regular graphs

Graph: We randomly and uniformly choose a graph from all the d-regular graphs with n vertices. Fix $d \geq 3$.

Random d-regular graphs

Graph: We randomly and uniformly choose a graph from all the d-regular graphs with n vertices. Fix $d \geq 3$.

Bollobs'82: For every $\epsilon > 0$, as $n \to \infty$, w.h.p.,

$$r_* \leq (1+\epsilon)\frac{\log n}{2\log(d-1)}$$

Random d-regular graphs

Graph: We randomly and uniformly choose a graph from all the d-regular graphs with n vertices. Fix $d \geq 3$.

Bollobs'82: For every $\epsilon > 0$, as $n \to \infty$, w.h.p.,

$$r_* \leq (1+\epsilon)\frac{\log n}{2\log(d-1)}$$

Indeed, if $r \geq (0.5+\varepsilon)\log_{d-1} n$ then for all $u \neq v$, $(d_1(v), \ldots, d_r(v)) \neq (d_1(u), \ldots, d_r(u))$ where $d_i(v)$ are the number of nodes at distance i from v.

Random d-regular graphs

Graph: We randomly and uniformly choose a graph from all the d-regular graphs with n vertices. Fix $d \geq 3$.

Bollobs'82: $r_* \leq (1+\epsilon)\frac{\log n}{2\log(d-1)}$.

Mossel-Sun'15+:

$$r_* = \frac{\log n + \log\log n}{2\log(d-1)} + O(1)$$

Random d-regular graphs

Graph: We randomly and uniformly choose a graph from all the d-regular graphs with n vertices. Fix $d \geq 3$.

Bollobs'82: $r_* \leq (1+\epsilon)\frac{\log n}{2\log(d-1)}$.

Mossel-Sun'15+:

$$r_* = \frac{\log n + \log\log n}{2\log(d-1)} + O(1)$$

- (Almost) all $0.5\log_{d-1}(n)$ neighborhoods are trees.
- However, each $0.5(1+\epsilon)\log_{d-1}(n)$ neighborhoods is encoded by a unique cycle structure.

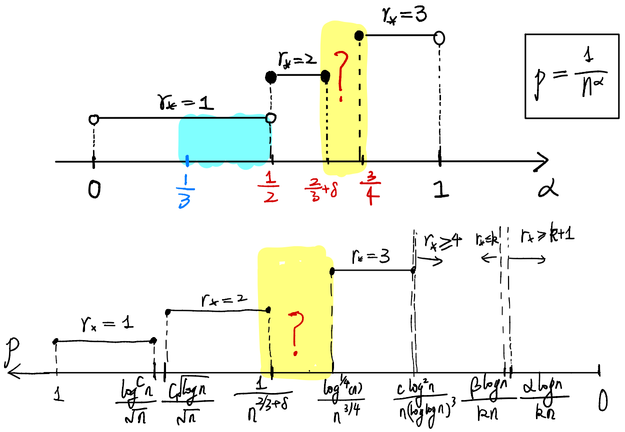
Erdős-Rényi graphs (dense regime)

ErdősRényi graph $\mathcal{G}_{n,p}$: each pair of vertices is connected with probability p independently.

Erdős-Rényi graphs (dense regime)

ErdősRényi graph $\mathcal{G}_{n,p}$: each pair of vertices is connected with probability p independently.

Mossel-Ross'19, Gaudio-Mossel'20, Huang-Tikhomirov'21+, Johnston-Kronenberg-Roberts-Scott'22+:



Sparse Erdős-Rényi graphs

Now we focus on $\mathcal{G}_{n,p}$ with $p = \frac{\lambda}{n}$ with fixed constant $\lambda > 0$.

Sparse Erdős-Rényi graphs

Now we focus on $\mathcal{G}_{n,p}$ with $p = \frac{\lambda}{n}$ with fixed constant $\lambda > 0$.

Mossel-Ross'19: For $\lambda \neq 1$, there exists a constant C_λ (with precise formula) such that for every $\epsilon > 0$, w.h.p.

$$\frac{1}{2(\lambda - \log \lambda)} \log n \leq r_* \leq C_\lambda \log n \, .$$

Sparse Erdős-Rényi graphs

Now we focus on $\mathcal{G}_{n,p}$ with $p = \frac{\lambda}{n}$ with fixed constant $\lambda > 0$.

Mossel-Ross'19: For $\lambda \neq 1$, there exists a constant C_λ (with precise formula) such that for every $\epsilon > 0$, w.h.p.

$$\frac{1}{2(\lambda - \log \lambda)} \log n \leq r_* \leq C_\lambda \log n \,.$$

- The lower bound also HOLDs for $\lambda = 1$.
- $C_\lambda = \frac{1}{\log(\lambda^{-1})}$ when $\lambda < 1$ and $C_\lambda = \frac{1}{\log(\lambda)} + \frac{2}{\log(1/\lambda_*)}$ when $\lambda > 1$, where $\lambda_* < 1$ satisfies $\lambda e^{-\lambda} = \lambda_* e^{-\lambda_*}$.

Shotgun threshold for $\mathcal{G}_{n,\frac{\lambda}{n}}$

Mossel-Ross'19: For $\mathcal{G}_{n,\frac{\lambda}{n}}$, $\frac{1}{2(\lambda-\log\lambda)}\log n \le r_* \le C_\lambda \log n$.

Shotgun threshold for $\mathcal{G}_{n,\frac{\lambda}{n}}$

Mossel-Ross'19: For $\mathcal{G}_{n,\frac{\lambda}{n}}$, $\frac{1}{2(\lambda-\log\lambda)}\log n \le r_* \le C_\lambda \log n$.

Theorem (Ding-Jiang-**M**. 22+)

Fix $\lambda \in (0,\infty)$. Let $\gamma_\lambda = \mathbb{P}(\mathbf{T} \sim \mathbf{T}')$, where $\mathbf{T}, \mathbf{T}'$ are two independent PGW(λ) trees. Let $\mathcal{G} \sim \mathcal{G}\left(n, \frac{\lambda}{n}\right)$. Then for any $\epsilon > 0$, w.h.p.,

$$(1-\epsilon)\frac{1}{\log\left(\lambda^2\gamma_\lambda\right)^{-1}}\log n \le r_* \le (1+\epsilon)\frac{1}{\log\left(\lambda^2\gamma_\lambda\right)^{-1}}\log n\,.$$

Shotgun threshold for $\mathcal{G}_{n,\frac{\lambda}{n}}$

Mossel-Ross'19: For $\mathcal{G}_{n,\frac{\lambda}{n}}$, $\frac{1}{2(\lambda-\log\lambda)}\log n \le r_* \le C_\lambda \log n$.

Theorem (Ding-Jiang-**M**. 22+)

Fix $\lambda \in (0,\infty)$. *Let* $\gamma_\lambda = \mathbb{P}(\mathbf{T} \sim \mathbf{T}')$, *where* $\mathbf{T}, \mathbf{T}'$ *are two independent PGW*(λ) *trees. Let* $\mathcal{G} \sim \mathcal{G}\left(n, \frac{\lambda}{n}\right)$. *Then for any* $\epsilon > 0$, *w.h.p.,*

$$(1-\epsilon)\frac{1}{\log\left(\lambda^2\gamma_\lambda\right)^{-1}}\log n \le r_* \le (1+\epsilon)\frac{1}{\log\left(\lambda^2\gamma_\lambda\right)^{-1}}\log n\,.$$

- Indeed there is a power series A with non-negative coefficients such that $\lambda^2\gamma_\lambda = A(\lambda e^{-\lambda})$.

Shotgun threshold for $\mathcal{G}_{n,\frac{\lambda}{n}}$

Mossel-Ross'19: For $\mathcal{G}_{n,\frac{\lambda}{n}}$, $\frac{1}{2(\lambda-\log\lambda)}\log n \le r_* \le C_\lambda \log n$.

Theorem (Ding-Jiang-**M**. 22+)

Fix $\lambda \in (0,\infty)$. Let $\gamma_\lambda = \mathbb{P}(\mathbf{T} \sim \mathbf{T}')$, where $\mathbf{T}, \mathbf{T}'$ are two independent PGW(λ) trees. Let $\mathcal{G} \sim \mathcal{G}\left(n, \frac{\lambda}{n}\right)$. Then for any $\epsilon > 0$, w.h.p.,

$$(1-\epsilon)\frac{1}{\log\left(\lambda^2\gamma_\lambda\right)^{-1}}\log n \le r_* \le (1+\epsilon)\frac{1}{\log\left(\lambda^2\gamma_\lambda\right)^{-1}}\log n\,.$$

- Indeed there is a power series A with non-negative coefficients such that $\lambda^2\gamma_\lambda = A(\lambda e^{-\lambda})$.
- We also give an algorithm with polynomial running time for reconstructing the original graph.

Non-identifiability: blocking configuration

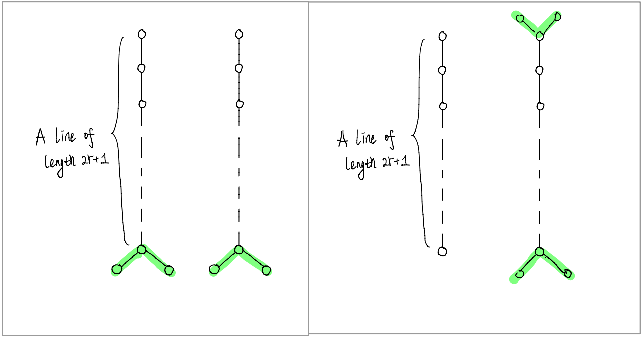


Figure 3: Blocking subgraph by Mossel and Ross

Non-identifiability: blocking configuration

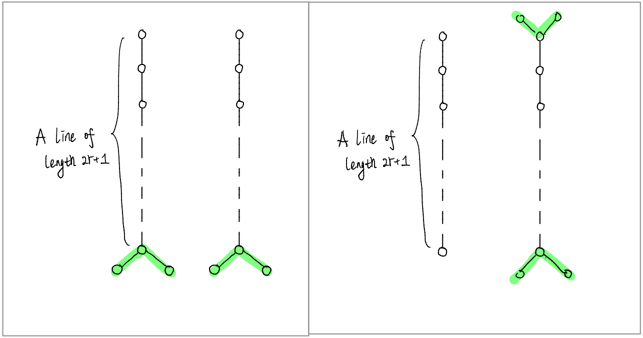


Figure 3: Blocking subgraph by Mossel and Ross

Letting the expactation of the number of such blocking subgraphs $n^2 \times (\lambda e^{-\lambda})^{2r} \times (\lambda e^{-\lambda})^{2r} \geq 1$, we have $r \leq \frac{1}{2(\lambda - \log \lambda)} \log n$.

Non-identifiability: blocking configuration

Key: The middle part (2r levels) are isomorphic; in addition removing red vertices results in small bushes

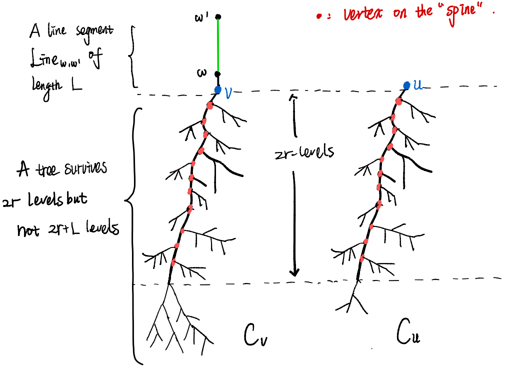
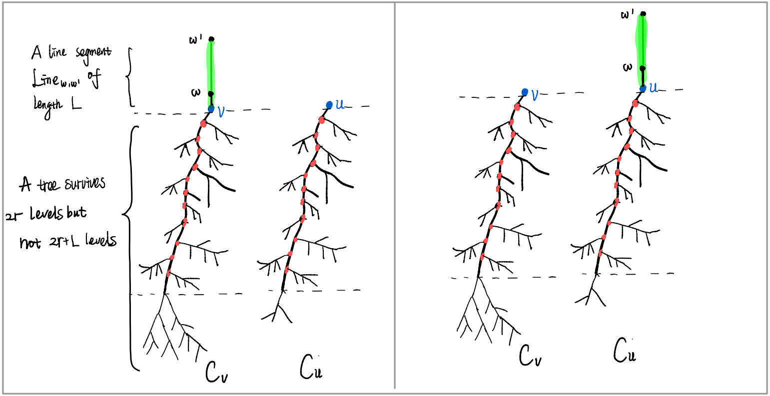


Figure 4: Improved blocking subgraph

Non-identifiability: blocking configuration



The expectation of the number of our blocking configuration is $n^2 \times \mathbb{P}(\mathbf{T} \sim_{2r} \mathbf{T}')$.

The expectation of the number of our blocking configuration is $n^2 \times \mathbb{P}(\mathbf{T} \sim_{2r} \mathbf{T}')$.

Lemma (Ding-Jiang-M. 22+)

$\mathbb{P}(\mathbf{T} \sim_{2r} \mathbf{T}') \asymp \alpha_\lambda^{2r}$ *where* $\alpha_\lambda := \lambda^2 \gamma_\lambda < 1$.

Letting $n^2 \times \alpha_\lambda^{2r} \geq 1$, we need $r \leq \frac{1}{\log(\alpha_\lambda^{-1})} \log n$.

Brief discussion on identifiability

Mossel-Ross'19: For $\lambda \neq 1$, $r_* \leq C_\lambda \log n$.

Brief discussion on identifiability

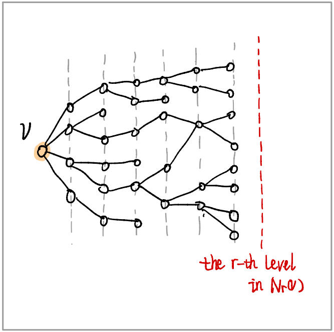
Mossel-Ross'19: For $\lambda \neq 1$, $r_* \leq C_\lambda \log n$.

- **uczak'98, Riordan-Wormald'10**: each connected component has diameter less than $C_\lambda \log n$ when $\lambda \neq 1$.

Brief discussion on identifiability

Mossel-Ross'19: For $\lambda \neq 1$, $r_* \leq C_\lambda \log n$.

- **uczak'98, Riordan-Wormald'10**: each connected component has diameter less than $C_\lambda \log n$ when $\lambda \neq 1$.



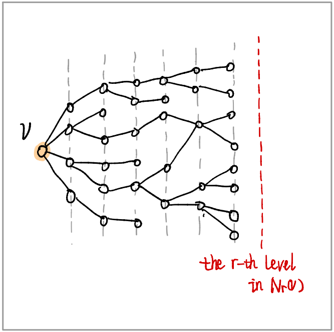
If $S_r(v) := \{u : dist(u, v) = r\} = \phi$

We know that $N_r(v)$ is a component!

Brief discussion on identifiability

Mossel-Ross'19: For $\lambda \neq 1$, $r_* \leq C_\lambda \log n$.

- **uczak'98, Riordan-Wormald'10**: each connected component has diameter less than $C_\lambda \log n$ when $\lambda \neq 1$.
- **Nachmias-Peres'08**: the diameter is of order $n^{1/3}$ for $\lambda = 1$.

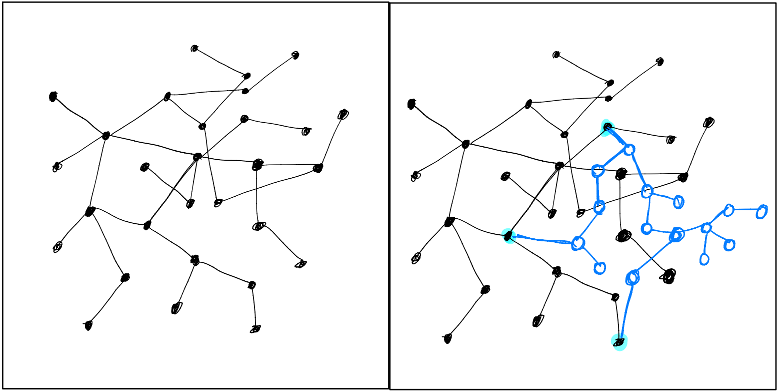


If $S_r(v) := \{u : dist(u, v) = r\} = \phi$

We know that $N_r(v)$ is a component!

Our Intuitions

Our key intuition is to recover bad components from good vertices.



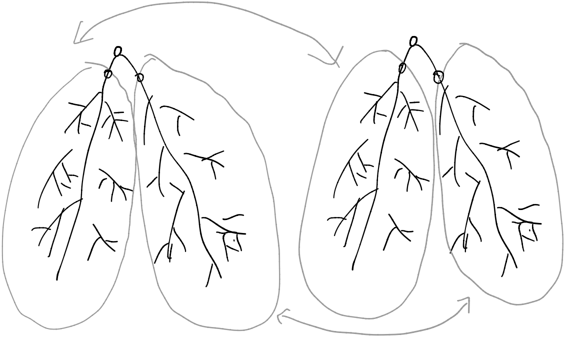
● : good veritices ○ : bad vertices

Brief discussion on identifiability

Assume that $r = \frac{1+\epsilon}{\log(\alpha_\lambda^{-1})} \log n$.

Key observation 1: Vertices which have **two disjoint r-arms** in their r-neighborhood are good:

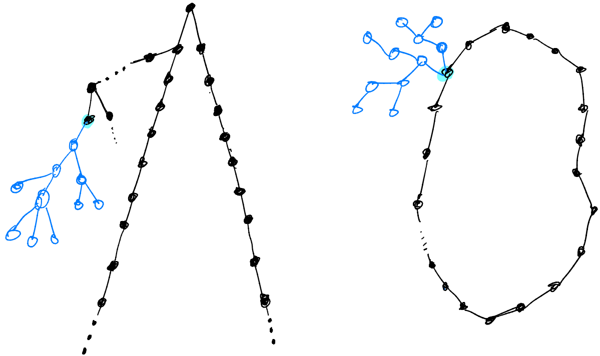
- "Essentially", their $(r-1)$-neighborhood is unique, since $n^2 \times \alpha_\lambda^{2r} \ll 1$.



Brief discussion on identifiability

Key observation 1: Vertices which have **two disjoint r-arms** in their r-neighborhood are good.

- Vertices without two r-arms can be identified from the r-neighborhood of some good vertex (or it is in a small component).



Brief discussion on identifiability

Caveat: we have **ignored cycles** in the graph in our analysis above, and this incurs serious challenge.

Key observation 2: Vertex which is contained by some cycle but doesn't have unique $(r-1)$ neighborhoods are very rare.

Thanks for your attention!