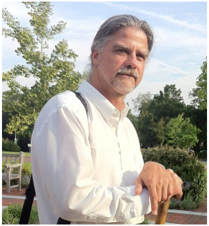
Maximum Likelihood Estimation and Graph Matching in Errorfully Observed Networks

Jesús Arroyo

September 20th, 2020

Joint work with:

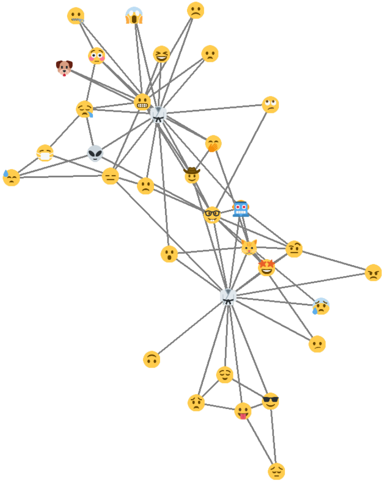
Daniel L. Sussman
(Boston University)



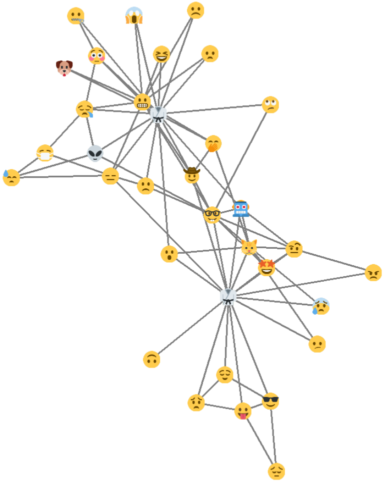
Carey E. Priebe
(Johns Hopkins University)

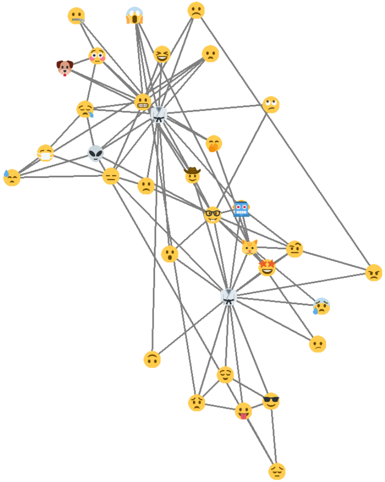
Vince Lyzinski
(University of Maryland)

Graph matching

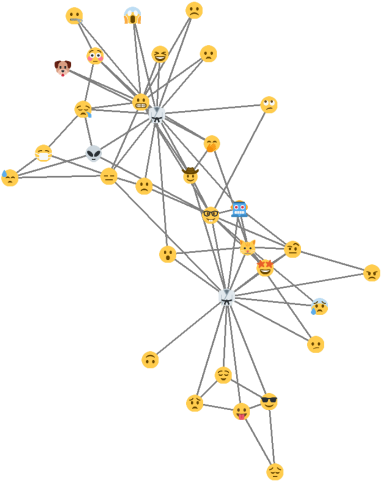


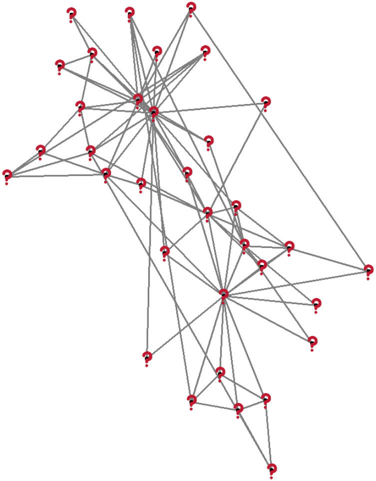
Graph matching



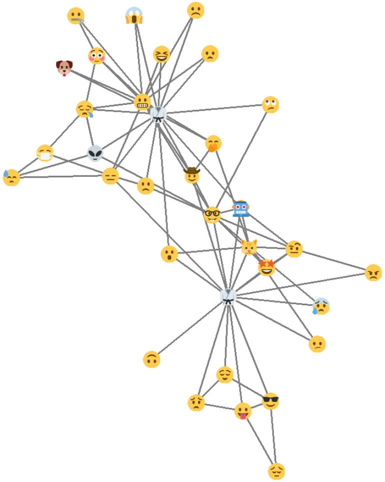


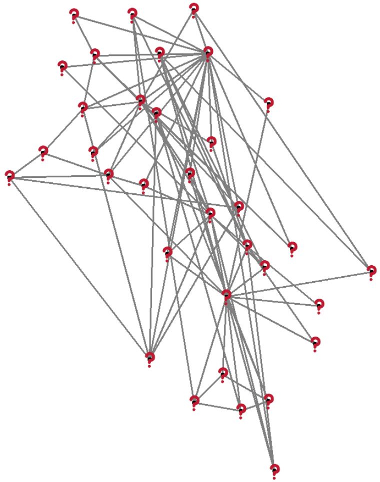
Graph matching



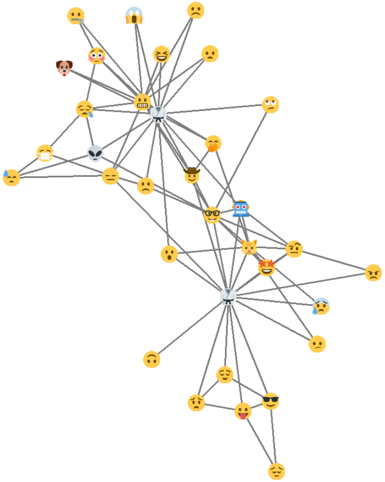


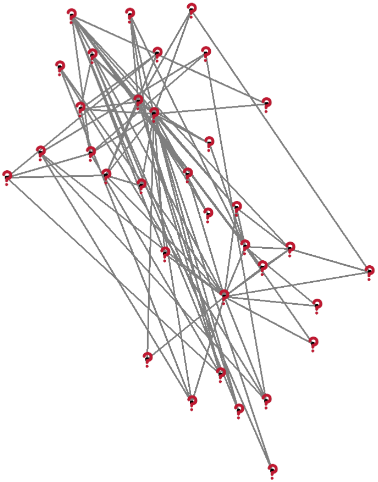
Graph matching



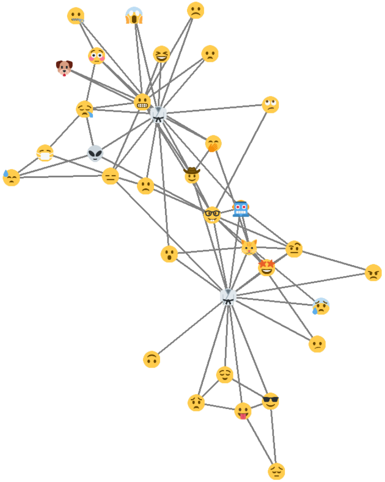


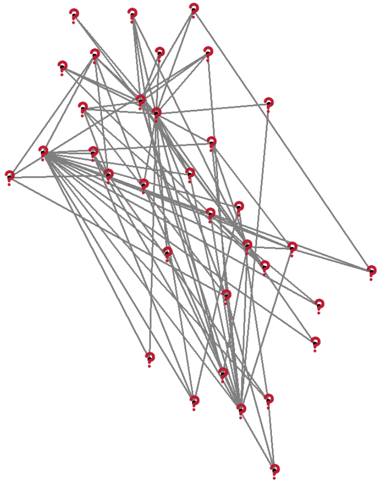
Graph matching



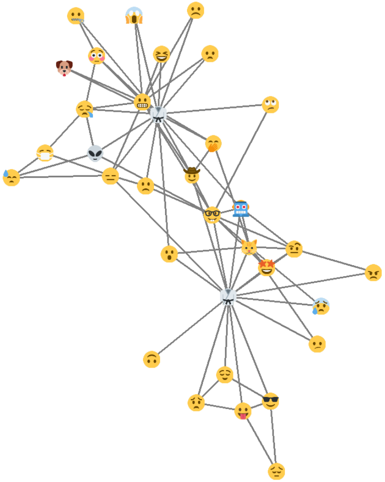


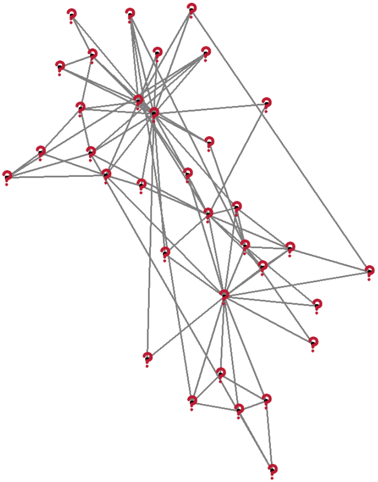
Graph matching



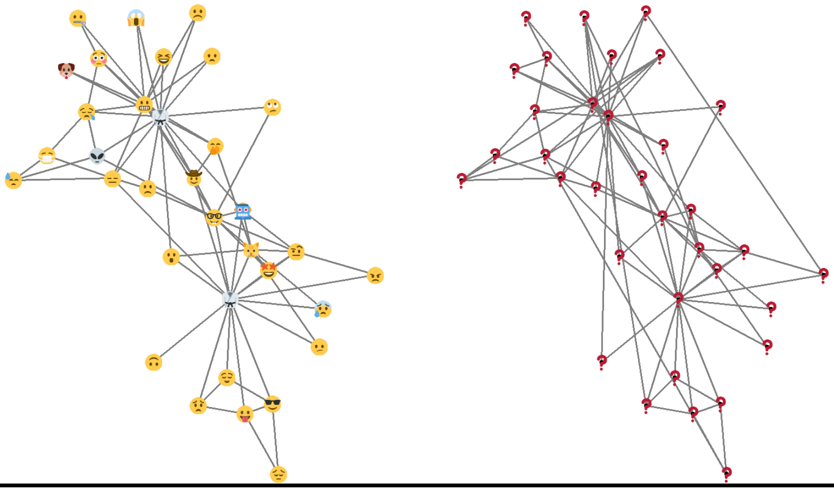


Graph matching





Graph matching



- How can we formulate a statistical model for graph matching?
- When is statistically feasible to solve the problem?

Graph matching problem

Goal: find a meaningful correspondence between the vertices of the graphs

- Typically defined as the permutation that minimizes the number of edge disagreements:

$$\begin{aligned}\widehat{Q}_{\mathsf{GM}} &= \operatorname*{argmin}_{Q \in \Pi_n} \|A - QBQ^T\|_F^2 \\ &= \operatorname*{argmin}_{Q \in \Pi_n} \sum_{i \neq j} \left(A_{ij} - B_{\sigma(i),\sigma(j)}\right)^2\end{aligned}$$

  - $A, B \in \{0,1\}^{n \times n}$ adjacency matrices
  - Q is a permutation matrix, σ is the associated permutation function

Graph matching problem

Goal: find a meaningful correspondence between the vertices of the graphs

- Typically defined as the permutation that minimizes the number of edge disagreements:

$$\begin{aligned}\widehat{Q}_{\mathsf{GM}} &= \underset{Q\in\Pi_n}{\operatorname{argmin}} \|A - QBQ^T\|_F^2 \\ &= \underset{Q\in\Pi_n}{\operatorname{argmin}} \sum_{i\neq j} \left(A_{ij} - B_{\sigma(i),\sigma(j)}\right)^2\end{aligned}$$

  - $A, B \in \{0,1\}^{n\times n}$ adjacency matrices
  - Q is a permutation matrix, σ is the associated permutation function

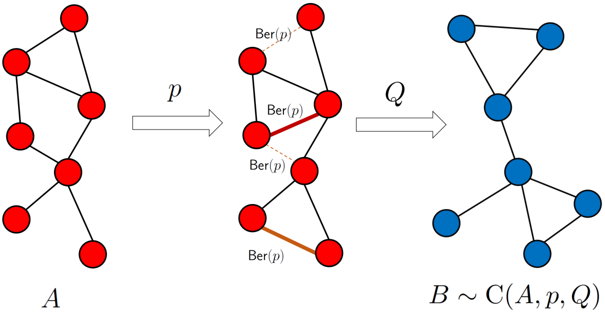
Some applications:

- Aligning biological networks
- De-anonymization in social networks
- Unsupervised word translation

Corrupting channel model

Model: B is an edge and vertex-label corrupted version of A

1. Edges and non-edges of A are independently flipped with probability p.
2. Vertices are shuffled with a permutation Q.



Maximum likelihood estimation

- Consider the maximum likelihood estimator for the parameters

$$(\widehat{p}_{\mathsf{MLE}}, \widehat{Q}_{\mathsf{MLE}}) := \underset{p,Q}{\operatorname{argmax}} \sum_{u>v} \log \mathbb{P}_p \left(A_{uv} = (QBQ^T)_{uv} \right).$$

- $\widehat{Q}_{\mathsf{MLE}}$ is the matching estimator via MLE.

Maximum likelihood estimation

- Consider the maximum likelihood estimator for the parameters

$$(\widehat{p}_{\mathsf{MLE}}, \widehat{Q}_{\mathsf{MLE}}) := \operatorname*{argmax}_{p,Q} \sum_{u>v} \log \mathbb{P}_p \left(A_{uv} = (QBQ^T)_{uv} \right).$$

- $\widehat{Q}_{\mathsf{MLE}}$ is the matching estimator via MLE.
- The MLE is equivalent to the typical graph matching formulation!

$$\widehat{Q}_{\mathsf{MLE}} = \operatorname*{argmin}_{Q \in \Pi_n} \|A - QBQ^T\|_F^2.$$

Maximum likelihood estimation

- Consider the maximum likelihood estimator for the parameters

$$(\widehat{p}_{\mathsf{MLE}}, \widehat{Q}_{\mathsf{MLE}}) := \underset{p,Q}{\operatorname{argmax}} \sum_{u>v} \log \mathbb{P}_p \left(A_{uv} = (QBQ^T)_{uv} \right).$$

- $\widehat{Q}_{\mathsf{MLE}}$ is the matching estimator via MLE.
- The MLE is equivalent to the typical graph matching formulation!

$$\widehat{Q}_{\mathsf{MLE}} = \underset{Q \in \Pi_n}{\operatorname{argmin}} \|A - QBQ^T\|_F^2.$$

- Extensions of the model to non-uniform corrupting probabilities
  - This result also holds with non-uniform corrupting probabilities.

When is the MLE correct?

- The difficulty of the problem depends on how different is A from QAQ^T for any $Q \neq I$, measured by

$$\|A - QAQ^T\|_F^2 = \sum_{i \neq j} (A_{ij} - A_{\sigma(i),\sigma(j)})\big)^2.$$

When is the MLE correct?

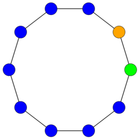
- The difficulty of the problem depends on how different is A from QAQ^T for any $Q \neq I$, measured by

$$\|A - QAQ^T\|_F^2 = \sum_{i \neq j} (A_{ij} - A_{\sigma(i),\sigma(j)}))^2.$$

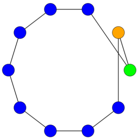
- $\Pi_{n,k}$ is the set of permutations that permute exactly k nodes.

Examples:

- $k = 2$
- $\|A - QAQ^T\|_F^2 = 4.$



A



QAQ^T

When is the MLE correct?

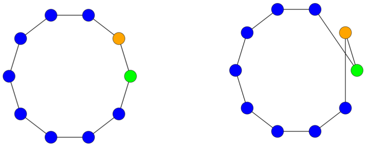
- The difficulty of the problem depends on how different is A from QAQ^T for any $Q \neq I$, measured by

$$\|A - QAQ^T\|_F^2 = \sum_{i \neq j} (A_{ij} - A_{\sigma(i),\sigma(j)}))^2.$$

- $\Pi_{n,k}$ is the set of permutations that permute exactly k nodes.

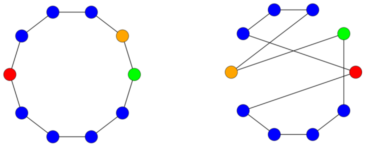
Examples:

- $k = 2$
- $\|A - QAQ^T\|_F^2 = 4.$



A QAQ^T

- $k = 3$
- $\|A - QAQ^T\|_F^2 = 10.$



A QAQ^T

Consistency of the MLE

Consider a sequence of networks $\{A_n\}$ with n vertices, and parameters $\{p_n, Q_n\}$

Definition

$\{\widehat{Q}_n\}$ is **consistent** if $\|\widehat{Q}_n - Q_n\|_F \xrightarrow{a.s.} 0$ as $n \to \infty$.

Consistency: all vertices are correctly matched in the limit.

Consistency of the MLE

Consider a sequence of networks $\{A_n\}$ with n vertices, and parameters $\{p_n, Q_n\}$

Definition

$\{\widehat{Q}_n\}$ is **consistent** if $\|\widehat{Q}_n - Q_n\|_F \xrightarrow{a.s.} 0$ as $n \to \infty$.

Consistency: all vertices are correctly matched in the limit.

Theorem

- $\widehat{Q}_{MLE}$ *is* **consistent** *if*

$$\min_{Q \in \Pi_{n,k}} \|A_n - QA_nQ^T\|_F^2 \geq \frac{6k \log n}{(1/2 - p_n)^2}, \qquad \forall k \geq 2.$$

Consistency of the MLE

Consider a sequence of networks $\{A_n\}$ with n vertices, and parameters $\{p_n, Q_n\}$

Definition

$\{\widehat{Q}_n\}$ is **consistent** if $\|\widehat{Q}_n - Q_n\|_F \xrightarrow{a.s.} 0$ as $n \to \infty$.

Consistency: all vertices are correctly matched in the limit.

Theorem

- $\widehat{Q}_{MLE}$ *is* **consistent** *if*

$$\min_{Q \in \Pi_{n,k}} \|A_n - QA_nQ^T\|_F^2 \geq \frac{6k \log n}{(1/2 - p_n)^2}, \qquad \forall k \geq 2.$$

- $\widehat{Q}_{MLE}$ *is* **not consistent** *if there exists* $m = \Omega(n)$ *disjoint permutations* $Q_1, \ldots, Q_m$

$$\max_{i \in [m]} \|A_n - Q_iA_nQ_i^T\|_F^2 = o\left(\frac{\log n}{(1/2 - p_n)^2}\right).$$

Relaxed consistency of the MLE

Matching all vertices correctly can be hard or impossible, for example:

- low-degree vertices
- automorphisms in the graph

Definition (Relaxed consistency)

$\{\widehat{Q}_n\}_{n=n_0}^{\infty}$ is $\{k_n\}$-**consistent** if $\frac{\|\widehat{Q}_n - Q_n\|_F}{k_n} \xrightarrow{a.s.} 0.$ as $n \to \infty$

Relaxed consistency: at most $o(k_n)$ misaligned vertices.

Relaxed consistency of the MLE

Matching all vertices correctly can be hard or impossible, for example:

- low-degree vertices
- automorphisms in the graph

Definition (Relaxed consistency)

$\{\widehat{Q}_n\}_{n=n_0}^{\infty}$ is $\{k_n\}$-**consistent** if $\frac{\|\widehat{Q}_n - Q_n\|_F}{k_n} \xrightarrow{a.s.} 0.$ as $n \to \infty$

Relaxed consistency: at most $o(k_n)$ misaligned vertices.

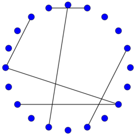
Theorem

$\widehat{Q}_{MLE}$ *is* k_n-**consistent** *if*

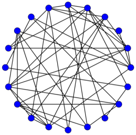
$$\min_{Q \in \Pi_{n,k}} \|A_n - QA_nQ^T\|_F^2 \geq \frac{6k \log n}{(1/2 - p_n)^2}, \qquad \forall k \geq k_n.$$

Consistency in random graph models

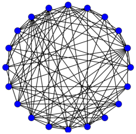
Erdös-Rényi graph model: $A_n \sim \mathrm{G}(n, \alpha_n)$



$\alpha = 0.05$



$\alpha = 0.2$



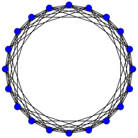
$\alpha = 0.4$

Corollary

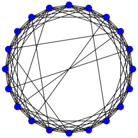
$\widehat{Q}_{\mathrm{MLE}}$ is **consistent** if

$$\alpha_n \geq \frac{c_1 \log n}{n(1/2 - p_n)^2}.$$

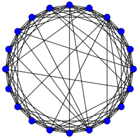
Newman-Watts small world graph model: $A_n \sim \mathrm{NW}(n, d_n, \beta_n)$,



$d = 4$,
$\beta = 0$

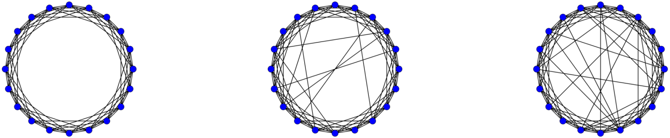


$d = 4$,
$\beta = 0.05$



$d = 4$,
$\beta = 0.2$

Newman-Watts small world graph model: $A_n \sim \mathrm{NW}(n, d_n, \beta_n)$,



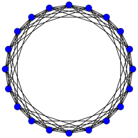
$d = 4$, $\beta = 0$ | $d = 4$, $\beta = 0.05$ | $d = 4$, $\beta = 0.2$

Corollary

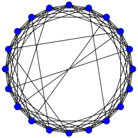
- For any $\{d_n\}$, $\widehat{Q}_{\mathsf{MLE}}$ is **not consistent** if $(1/2 - p_n)^2 = o(\sqrt{\log n / n})$ and

$$\beta_n = o\left(\frac{\log n}{(1/2 - p_n)^2 n}\right).$$

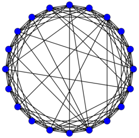
Newman-Watts small world graph model: $A_n \sim \mathrm{NW}(n, d_n, \beta_n)$,



$d = 4$,
$\beta = 0$



$d = 4$,
$\beta = 0.05$



$d = 4$,
$\beta = 0.2$

Corollary

- For any $\{d_n\}$, $\widehat{Q}_{\text{MLE}}$ is **not consistent** if $(1/2 - p_n)^2 = o(\sqrt{\log n / n})$ and

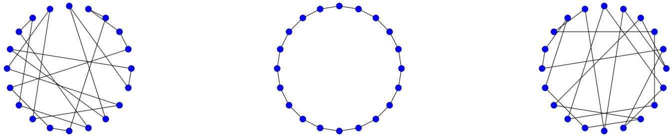
$$\beta_n = o\left(\frac{\log n}{(1/2 - p_n)^2 n}\right).$$

- $\widehat{Q}_{\mathrm{MLE}}$ is **consistent** if $d_n = o(\beta_n^2 n)$ and

$$\beta_n \geq C\sqrt{\frac{\log n}{n\,(1/2 - p_n)^2}}.$$

Relaxed consistency

Random regular graphs: $A \sim \mathrm{G}_{n,d_n}$



$d = 2$

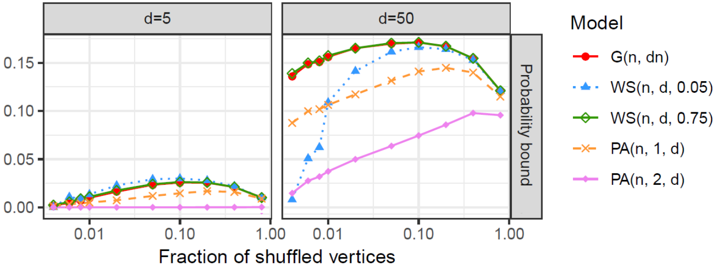
Corollary

If $d_n = c_1 n^{2/3+\varepsilon}$ and $(\frac{1}{2} - p_n)^2 \geq c_2 \frac{\log n}{d_n}$, then $\widehat{Q}_{\mathsf{MLE}}$ is $\{n^{2/3+\varepsilon}\}$-consistent.

A negligible fraction of the vertices ($o(n^{-(1/3-\epsilon)})$) might be incorrectly matched.

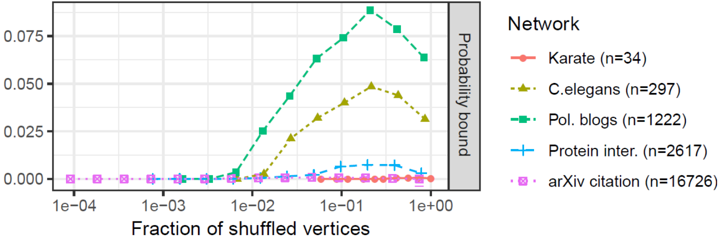
Matchability on random graphs

Measure matching feasibility: Upper bound for the noise probability tolerated by a graph based on the theory



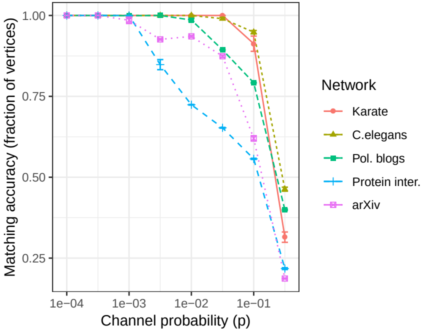
- Models considered
  - $G(n, dn)$ Erdös-Rényi
  - $WS(n, d, \beta)$: Watts-Strogatz
  - $PA(n, \gamma, d)$: preferential attachment

Matchability on real networks



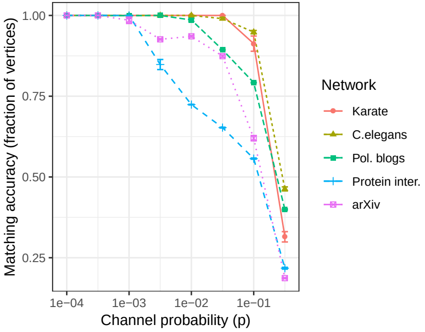
Matchability on real networks

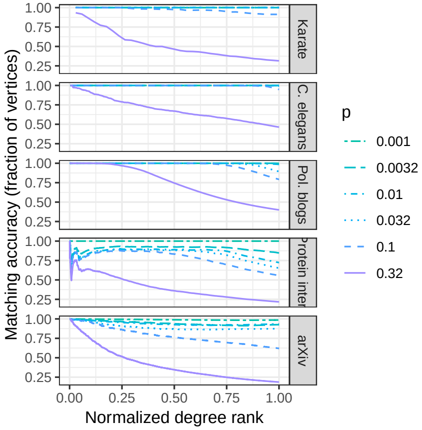
- Sample B from corrupting channel, and approximately compute the MLE
- Evaluate percentage of correctly matched vertices (left) and large degree vertices that are correctly matched (right).



Matchability on real networks

- Sample B from corrupting channel, and approximately compute the MLE
- Evaluate percentage of correctly matched vertices (left) and large degree vertices that are correctly matched (right).





Conclusion

- Graph matching on errorfully observed networks is *statistically* feasible when:
    - The topology of the network is sufficiently irregular
    - The noise is small enough
- What are the computational limits?
- Matching on other relational data structures? multilayer networks, vertex or edge attributes, bipartite graphs, etc.
    - *Graph matching between bipartite and unipartite networks:* arXiv:2002.01648.

Conclusion

- Graph matching on errorfully observed networks is *statistically* feasible when:
    - The topology of the network is sufficiently irregular
    - The noise is small enough
- What are the computational limits?
- Matching on other relational data structures? multilayer networks, vertex or edge attributes, bipartite graphs, etc.
    - *Graph matching between bipartite and unipartite networks:* arXiv:2002.01648.

Thank you!

Arroyo, J., Sussman, D. L., Priebe, C. E., Lyzinski, V. (2018). *Maximum Likelihood Estimation and Graph Matching in Errorfully Observed Networks*. arXiv:1812.10519.