# Rank Aggregation from Pairwise Comparisons in the Presence of Adversarial Corruptions

Arpit Agarwal, Shivani Agarwal, Sanjeev Khanna, **Prathamesh Patil**

**ICML 2020**

Penn
UNIVERSITY *of* PENNSYLVANIA

# Rank Aggregation from Pairwise Comparisons

In many practical applications, the available data comes in the form of comparisons and choices.

Aggregating these partial preferences into a complete ordering is important in order to understand user behavior and predict future behavior.

Applications include e-commerce, recommendation systems, and information retrieval.

JOKER
≻
AVENGERS ENDGAME
≻
SONY
≻
Nikon
⋮

# Need for Robustness

Rank aggregation algorithms play a critical role in modern web applications.

Determining product placement,
Ordering search results,
Providing recommendations.

Their significant economic and societal impact provides strong incentives for malicious players to manipulate the comparison data in order to skew the outcome in their favor.

Voter fraud in elections,
Inflated purchases in e-commerce,
Click fraud in online advertising,

Designing rank aggregation algorithms that are robust to adversarial corruptions in input comparison data is a crucial challenge.

# Our Contribution

We initiate the study of robustness in rank aggregation from pairwise comparisons under the Bradley-Terry-Luce model.

We propose a powerful adversarial contamination model, under which

★ Given arbitrary comparison data, we exactly characterize the extent of contamination that can be tolerated up to which the true BTL model parameters are uniquely identifiable.

★ We show that robustness to adversarial contamination is a structural property of the comparison data itself. Not all data are created equal!

★ For a natural family of comparison data (Erdős-Rényi comparison graphs), we present a near-quadratic time algorithm (based on Linear Programming) for parameter recovery from comparison data containing a non-trivial fraction of contamination.

# Outline

**Preliminaries**

- Bradley-Terry-Luce Model
- Comparison Graphs

**Adversarial Contamination Model**

**Condition for Unique Identifiability**

- Robustness as a Structural Property

**Results for Erdős-Rényi Comparison Graphs**

- A Sharp Threshold Condition for Identifiability
- Algorithm for Parameter Recovery

# Outline

**Preliminaries**

- Bradley-Terry-Luce Model
- Comparison Graphs

**Adversarial Contamination Model**

**Condition for Unique Identifiability**

- Robustness as a Structural Property

**Results for Erdős-Rényi Comparison Graphs**

- A Sharp Threshold Condition for Identifiability
- Algorithm for Parameter Recovery

# The Bradley-Terry-Luce Model

[Zermelo, 1928; Bradley & Terry, 1952; Luce, 1959]

It is a comparison model used to explain outcomes of pairwise comparisons.

Given a universe of $n$ items/alternatives, associates a positive weight $w_i > 0$ with each item $i \in [n]$, and posits that for any pair $i, j \in [n] \times [n]$,

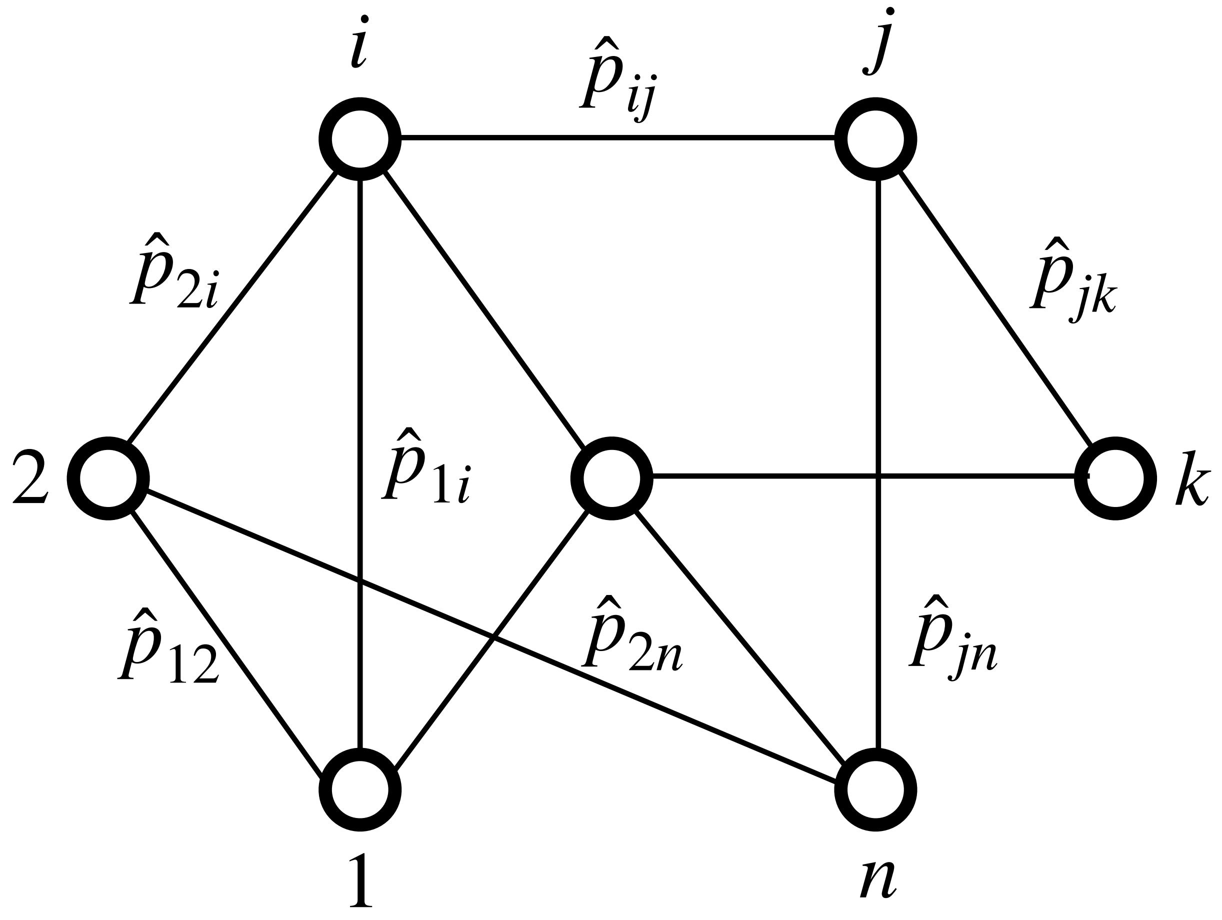
$$P(i \succ j) = \frac{w_i}{w_i + w_j}$$

Given data consisting of pairwise comparisons whose outcomes are assumed to be drawn according to the BTL model, the objective is typically to recover the underlying item weights $\mathbf{w}$ (up to multiplicative scaling).

# Comparison Data $\equiv$ Weighted Comparison Graph

Comparison data, which consists of pairs $\{i, j\}$ of items and the observed probability $\hat{p}_{ij}$ with which $i$ beats $j$ induces a weighted graph $G = (V, E)$, where

- The vertex set $V$ corresponds to the set of items $[n]$.
- An edge $\{i, j\} \in E$ iff items $\{i, j\}$ were compared.
- If an edge $\{i, j\} \in E$, then its weight is $\hat{p}_{ij}$.



# Outline

**Preliminaries**

- Bradley-Terry-Luce Model
- Comparison Graphs

**Adversarial Contamination Model**

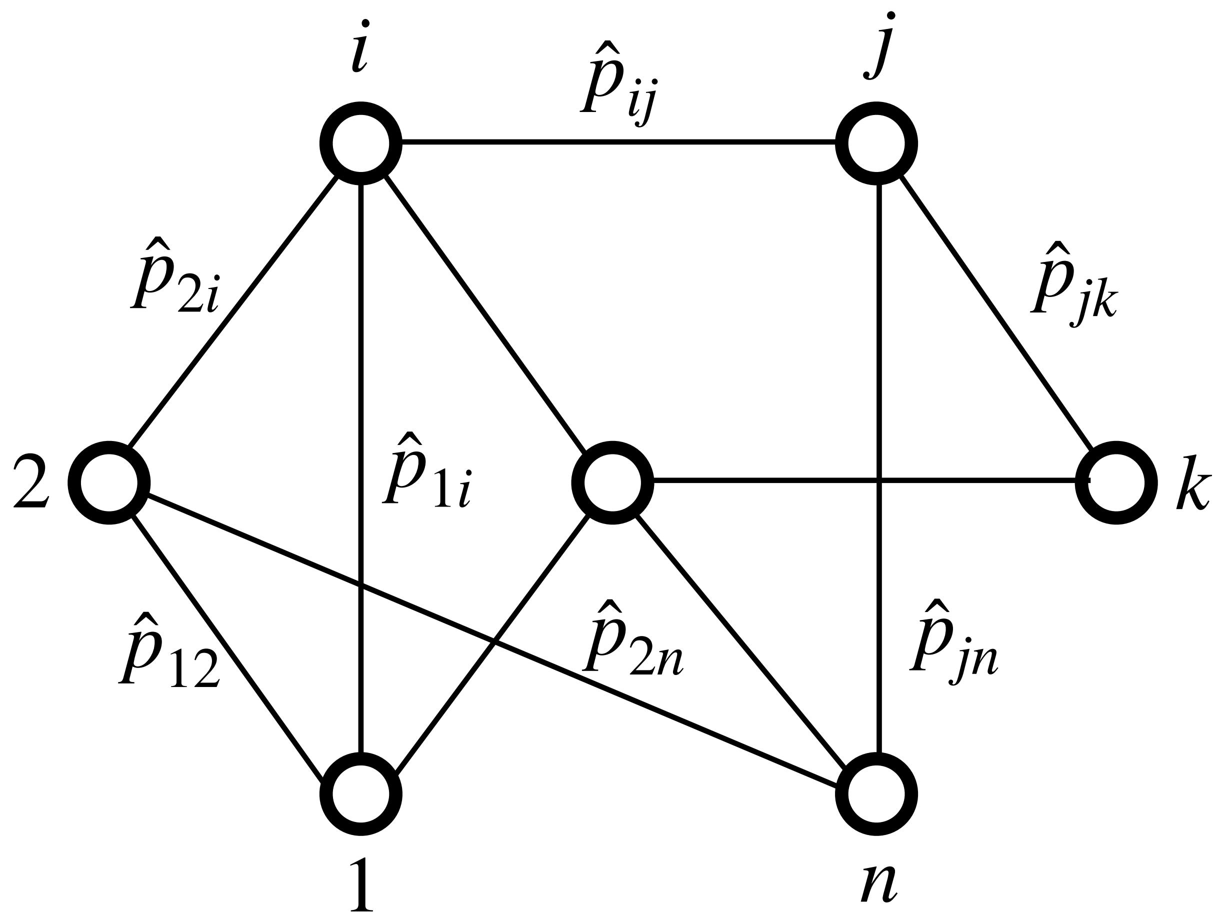
**Condition for Unique Identifiability**

- Robustness as a Structural Property

**Results for Erdős-Rényi Comparison Graphs**

- A Sharp Threshold Condition for Identifiability
- Algorithm for Parameter Recovery

# Adversarial Contamination Model



Nature generates a comparison graph $G^* = ([n], E^*)$

Each edge $\{i, j\} \in E^*$ is labeled with a truthful estimate $\hat{p}_{ij}$ consistent with a BTL model with (unknown) weights $\mathbf{w}^*$

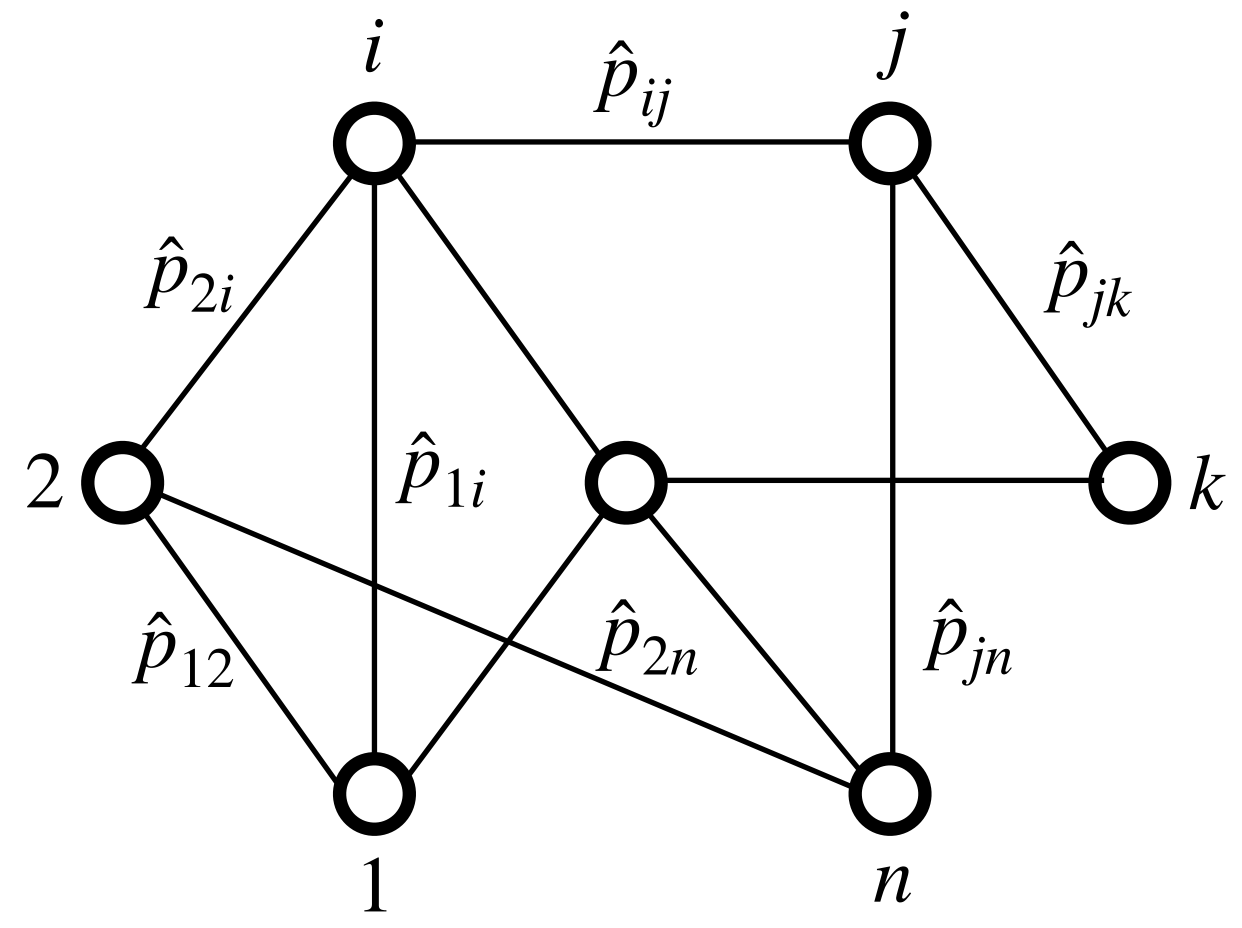
"Truthful Estimate" $\hat{p}_{ij}$ consistent with $\mathbf{w}^*$:

$\hat{p}_{ij}$ is a good approximation for the true probability

$$\hat{p}_{ij} \approx p^*_{ij} = \frac{w^*_i}{w^*_i + w^*_j}$$

Practical example: $\hat{p}_{ij}$ is the empirical fraction of times $i$ beats $j$ out of $L$ independent comparisons between them.

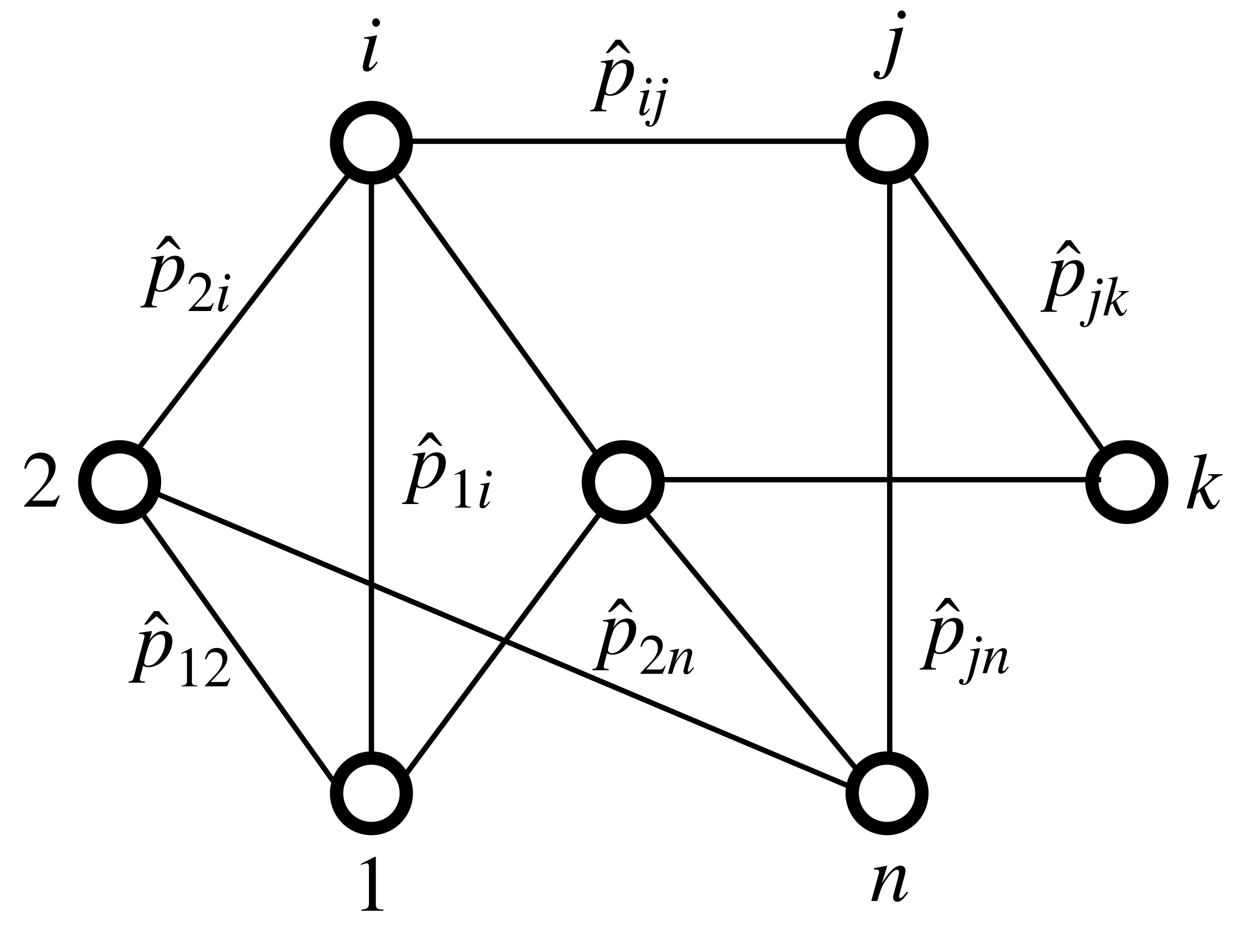
# Adversarial Contamination Model



Nature generates a comparison graph $G^* = ([n], E^*)$

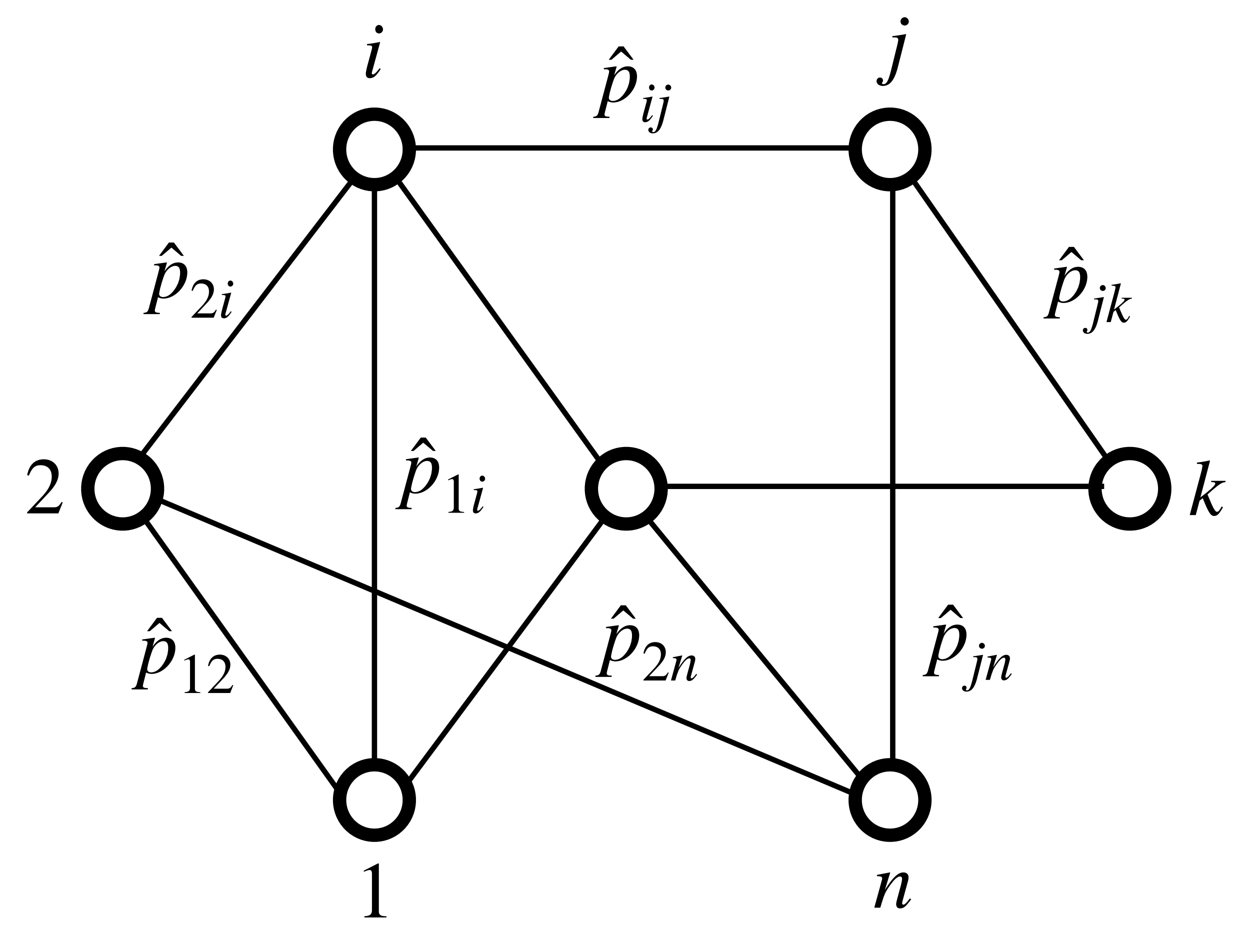
Each edge $\{i, j\} \in E^*$ is labeled with a truthful estimate $\hat{p}_{ij}$ consistent with a BTL model with (unknown) weights $\mathbf{w}^*$

$\xrightarrow{\textbf{Adversary}}$



Contaminated comparison graph $G = ([n], E)$

# Adversarial Contamination Model

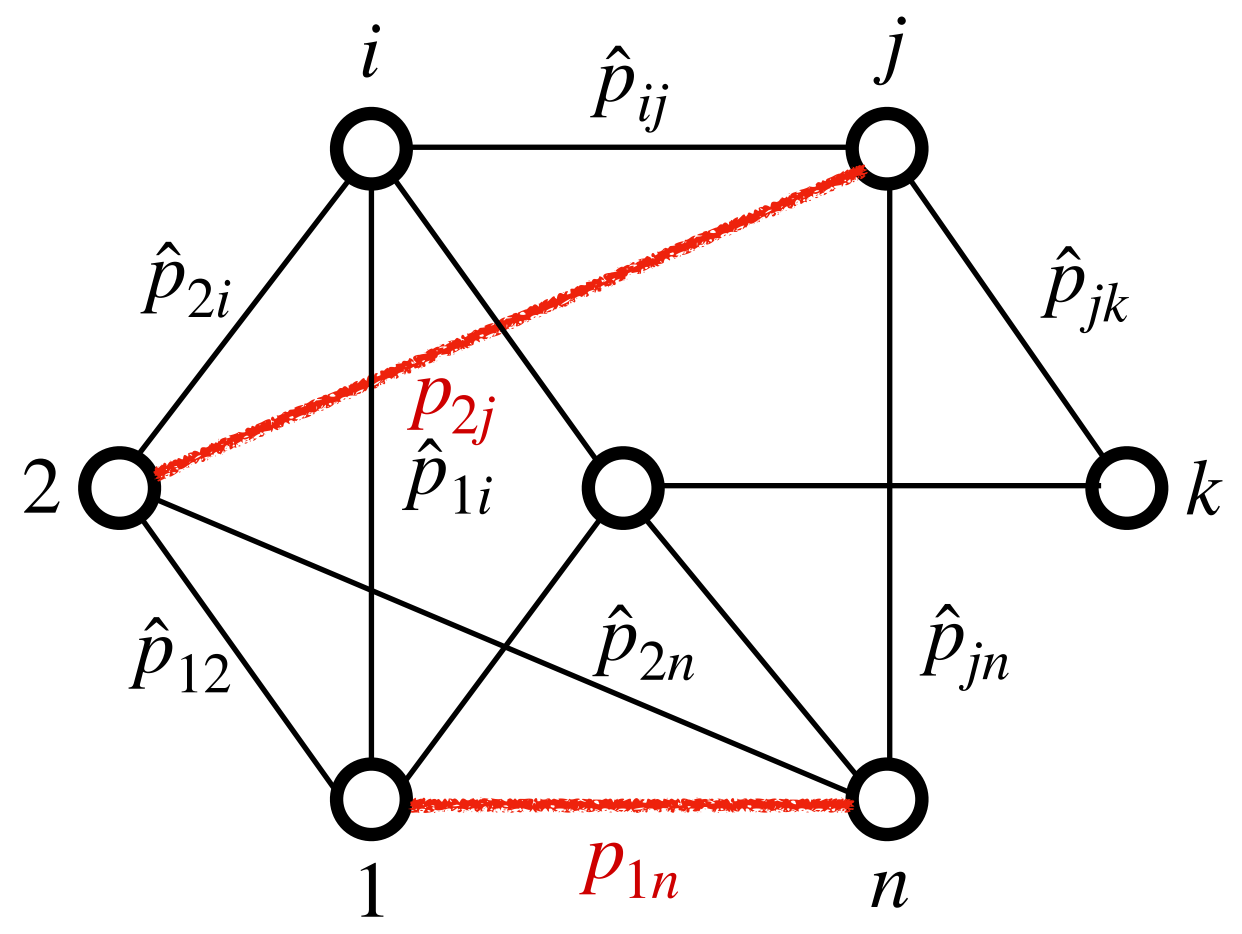


Nature generates a comparison graph
$G^* = ([n], E^*)$

Each edge $\{i, j\} \in E^*$ is labeled with a truthful estimate $\hat{p}_{ij}$ consistent with a BTL model with (unknown) weights $\mathbf{w}^*$

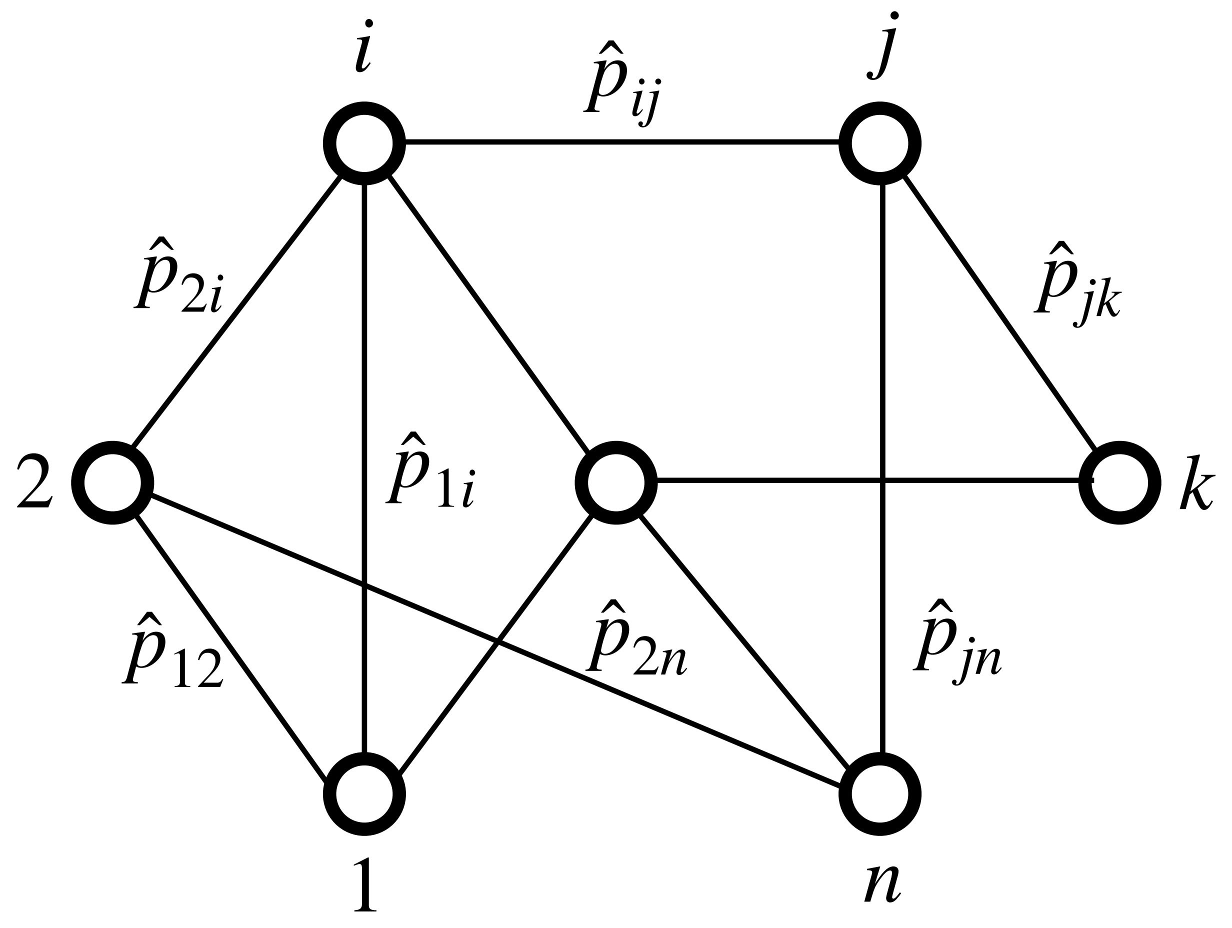
**Adversary**

Add edges with spurious labels



Contaminated comparison graph
$G = ([n], E)$

# Adversarial Contamination Model



Nature generates a comparison graph

$G^* = ([n], E^*)$

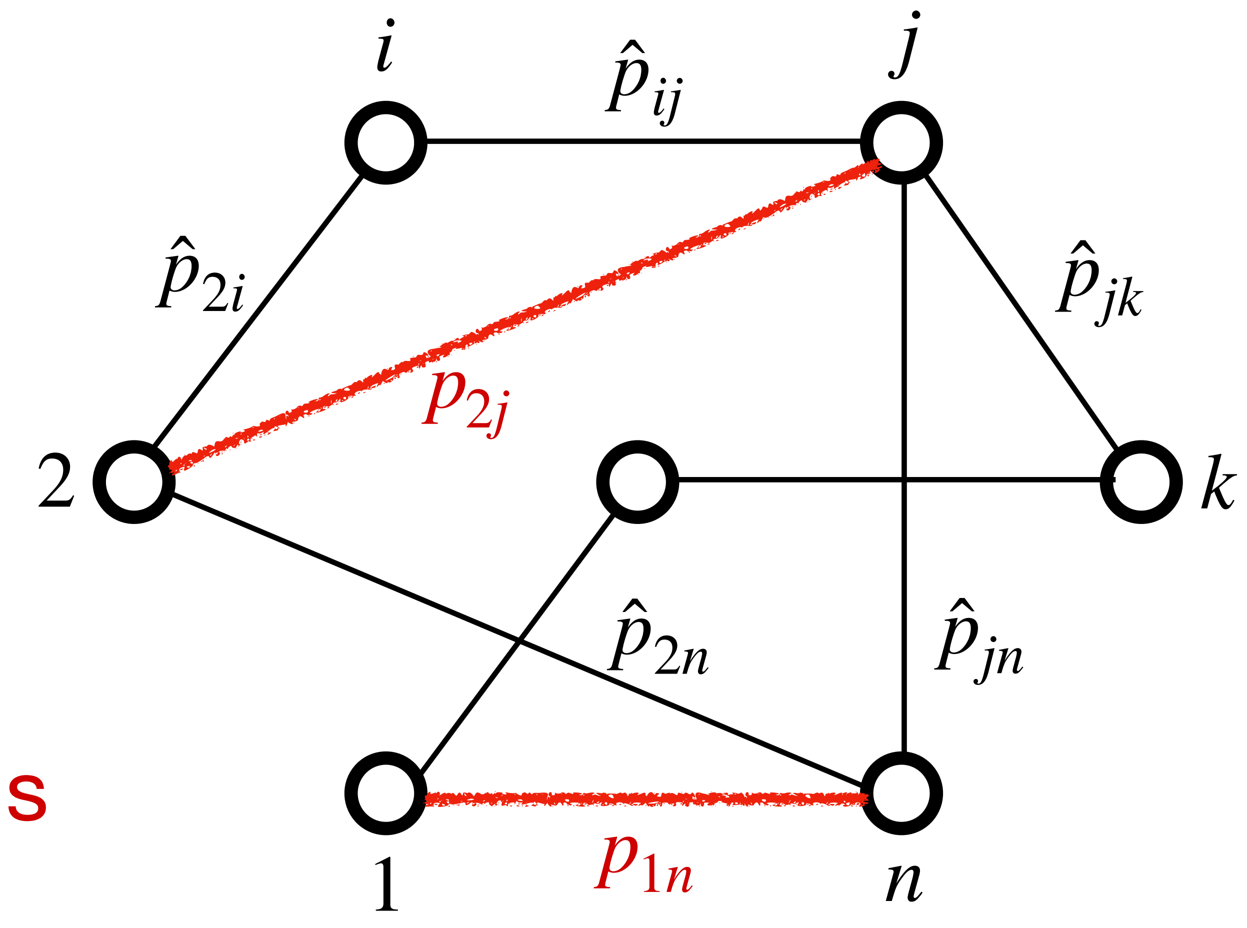
Each edge $\{i, j\} \in E^*$ is labeled with a truthful estimate $\hat{p}_{ij}$ consistent with a BTL model with (unknown) weights $\mathbf{w}^*$

**Adversary**

$\longrightarrow$

Add edges with spurious labels

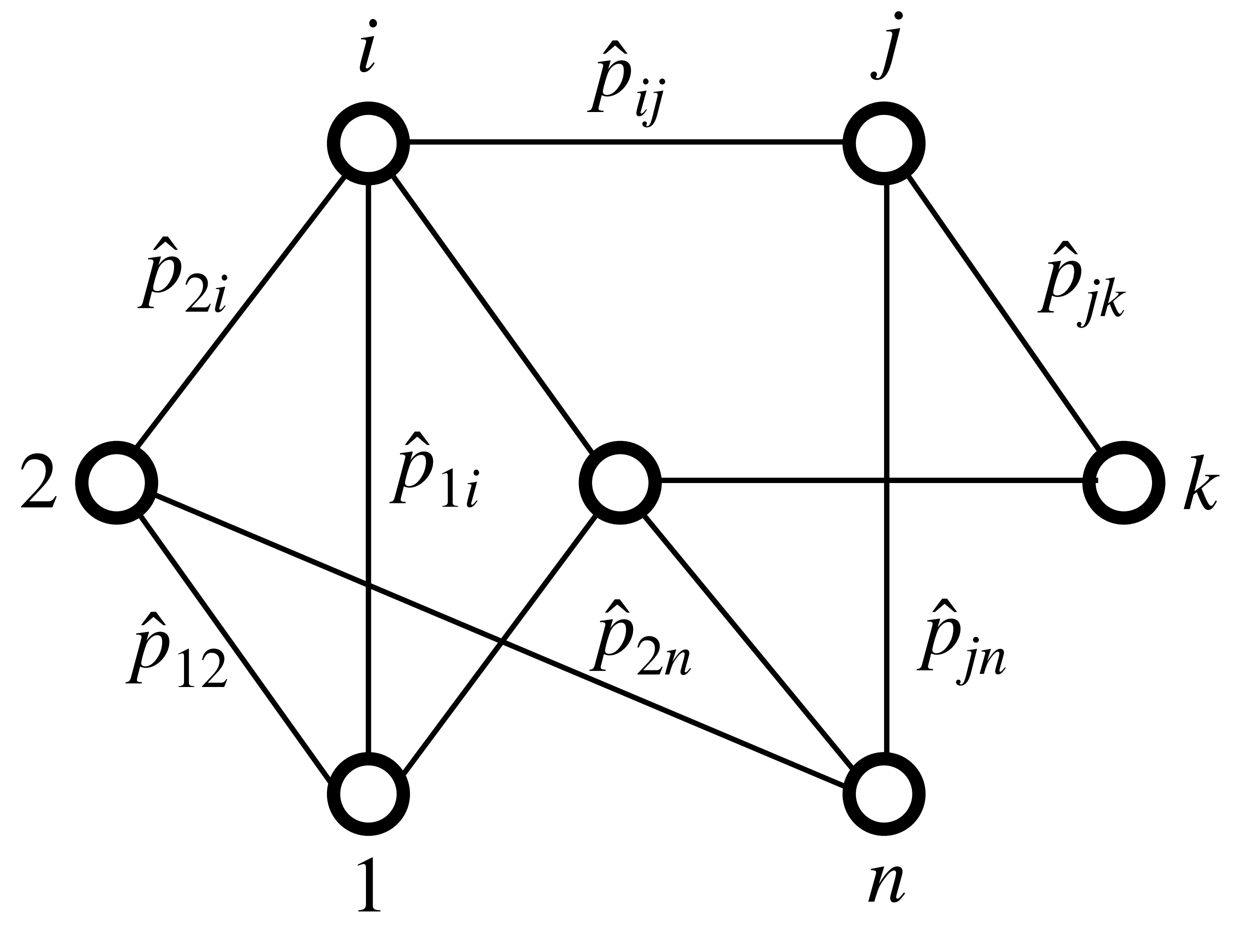
Delete existing edges and their labels



Contaminated comparison graph

$G = ([n], E)$

# Adversarial Contamination Model



Nature generates a comparison graph $G^* = ([n], E^*)$

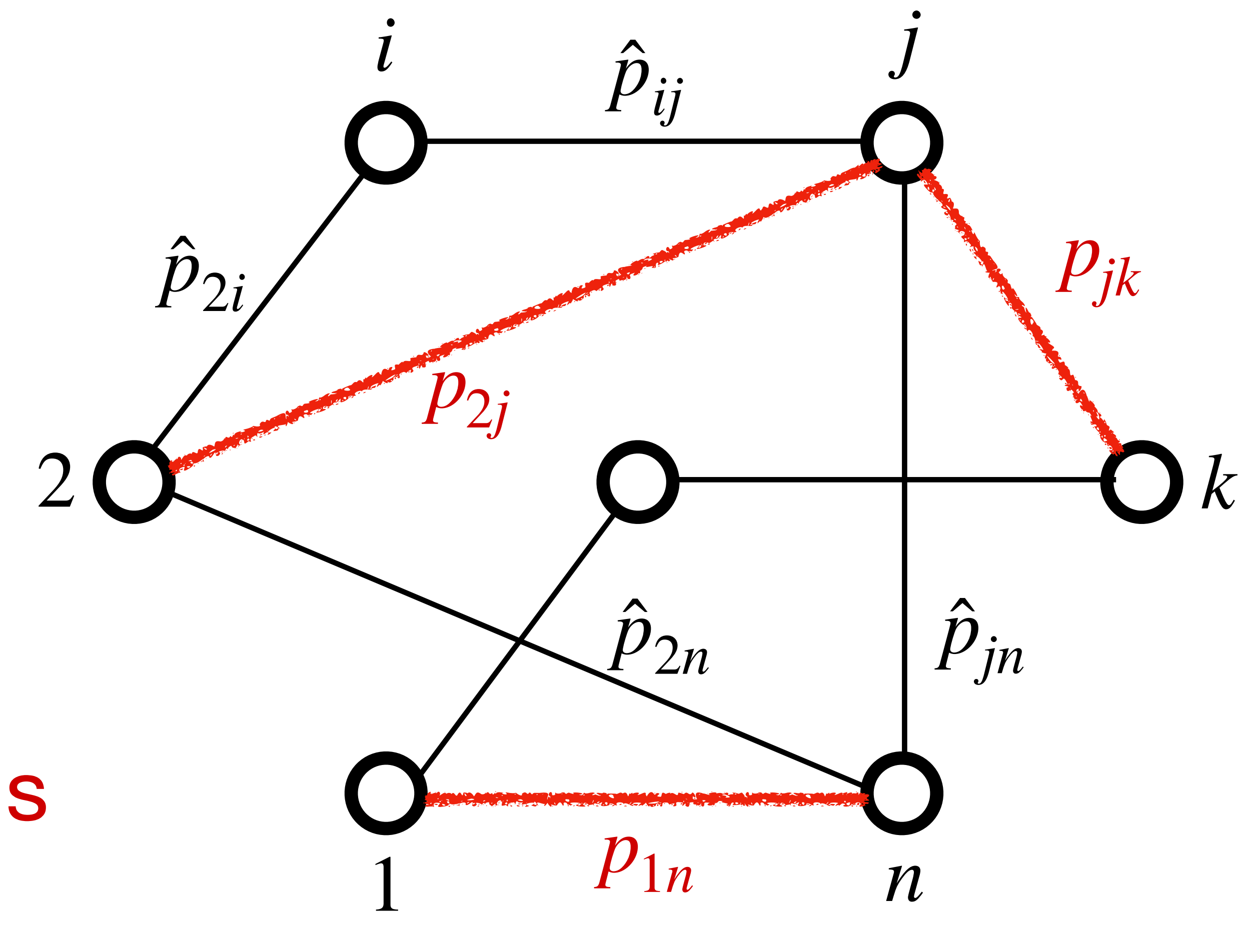
Each edge $\{i, j\} \in E^*$ is labeled with a truthful estimate $\hat{p}_{ij}$ consistent with a BTL model with (unknown) weights $\mathbf{w}^*$

**Adversary**

Add edges with spurious labels

Delete existing edges and their labels

Corrupt labels on existing edges



Contaminated comparison graph $G = ([n], E)$

**Received as Input**

# Existing Methods… Don’t Work

Parameter estimation under the (uncontaminated) BTL model has received a lot of attention in the ML community, and is a very well understood problem.

Negahban et al., 2012
Hajek et al., 2014
Chen and Suh., 2015
Maystre and Grossglauser, 2015
Shah et al., 2016
Agarwal et al., 2018
Hendrickx et al., 2019
Chen et al., 2019
⋮

Efficient, consistent algorithms for parameter estimation in the uncontaminated setting.

However… these are not robust.

Crucially rely on the assumption that input data is truthfully generated.

Their recovery guarantees do not hold in the presence of adversarial corruptions!

# Outline

**Preliminaries**

- Bradley-Terry-Luce Model
- Comparison Graphs

**Adversarial Contamination Model**

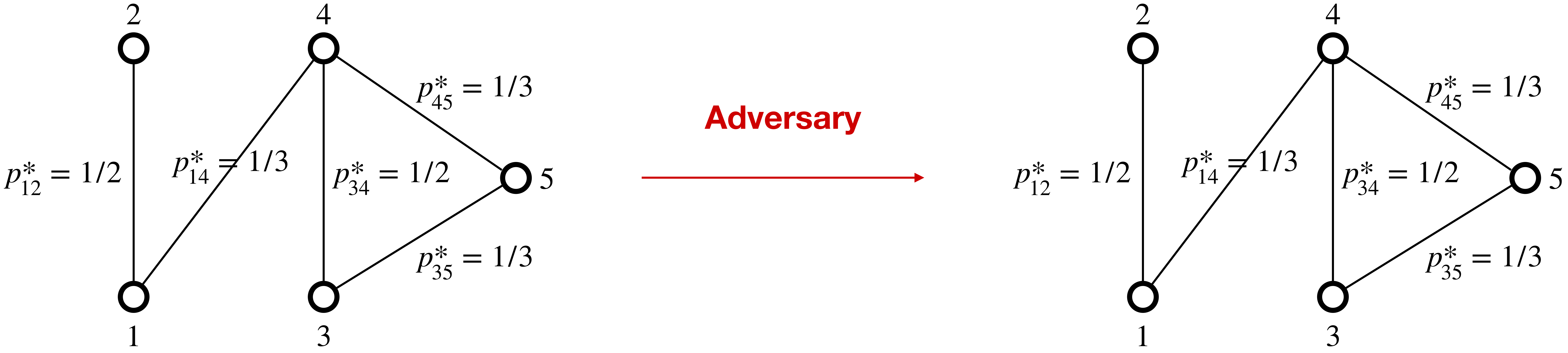
**Condition for Unique Identifiability**

- Robustness as a Structural Property

**Results for Erdős-Rényi Comparison Graphs**

- A Sharp Threshold Condition for Identifiability
- Algorithm for Parameter Recovery

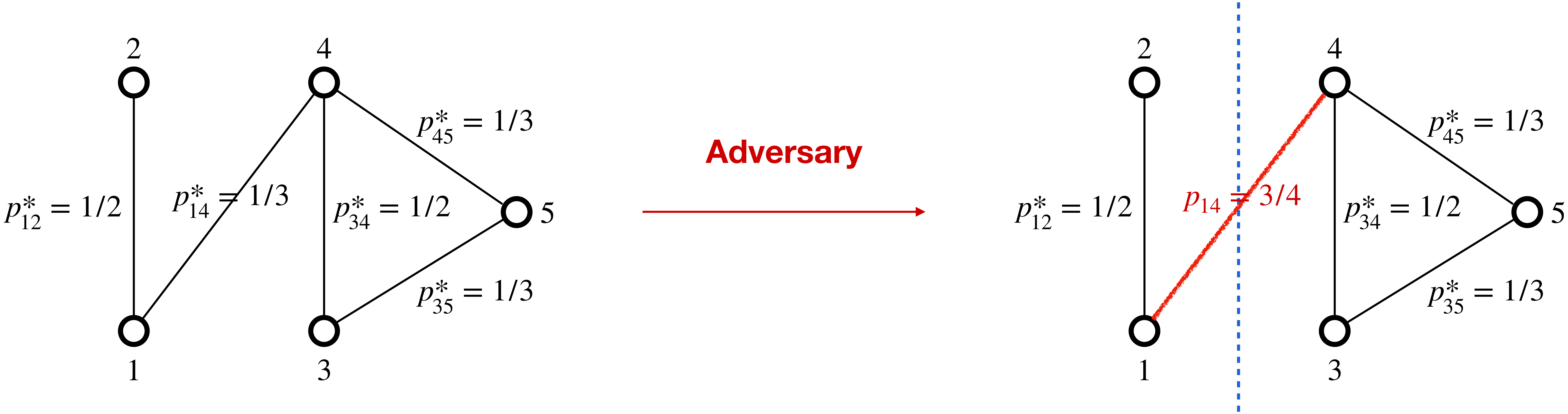
# A Challenging Example



Truthful comparison graph entirely consistent with

$$\mathbf{w}^* = (1,1,2,2,4)/10$$

# A Challenging Example



Truthful comparison graph entirely consistent with

$\mathbf{w}^* = (1,1,2,2,4)/10$

Contaminated graph entirely consistent with

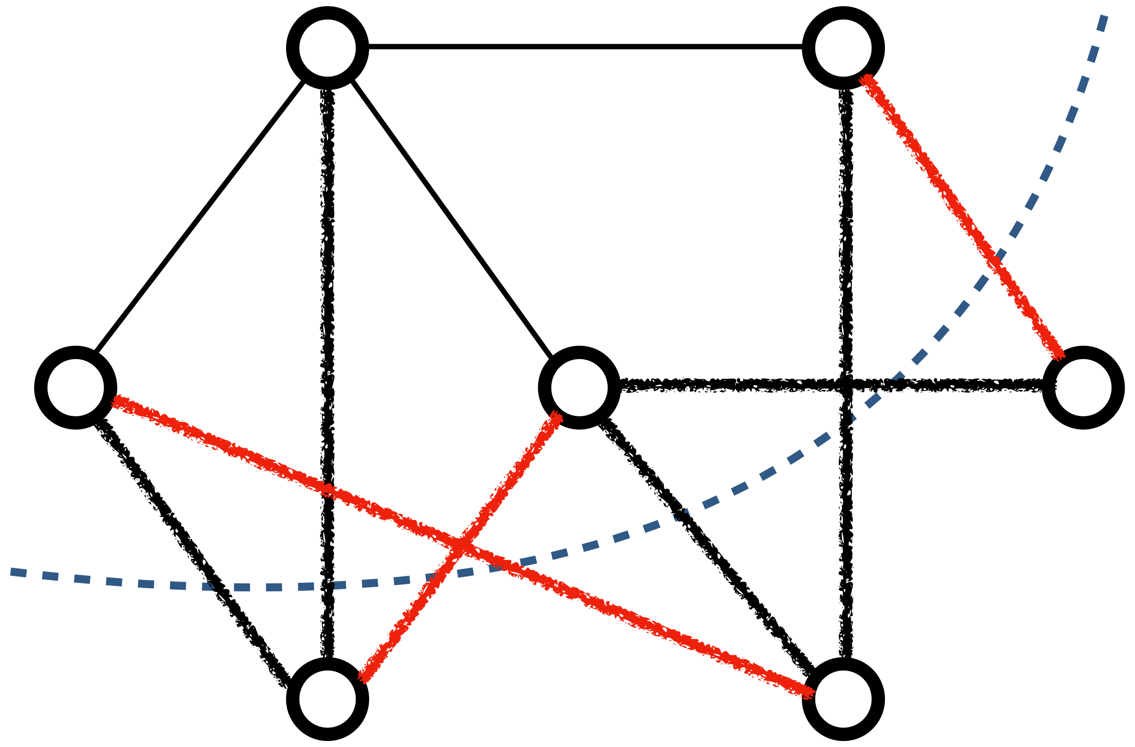
$\mathbf{w} = (3,3,1,1,2)/10$

No evidence of corruption in the contaminated graph!

Items with the lowest scores have highest scores post corruption!

# Exact Condition for Identifiability of $\mathbf{w}^*$

**Theorem 1. (Cut Majority Condition)**
*Given an arbitrary, contaminated comparison graph $G$, the true weights $\mathbf{w}^*$ are uniquely identifiable if and only if every cut in $G$ has strictly more uncorrupted edges than corrupted edges crossing the cut.*



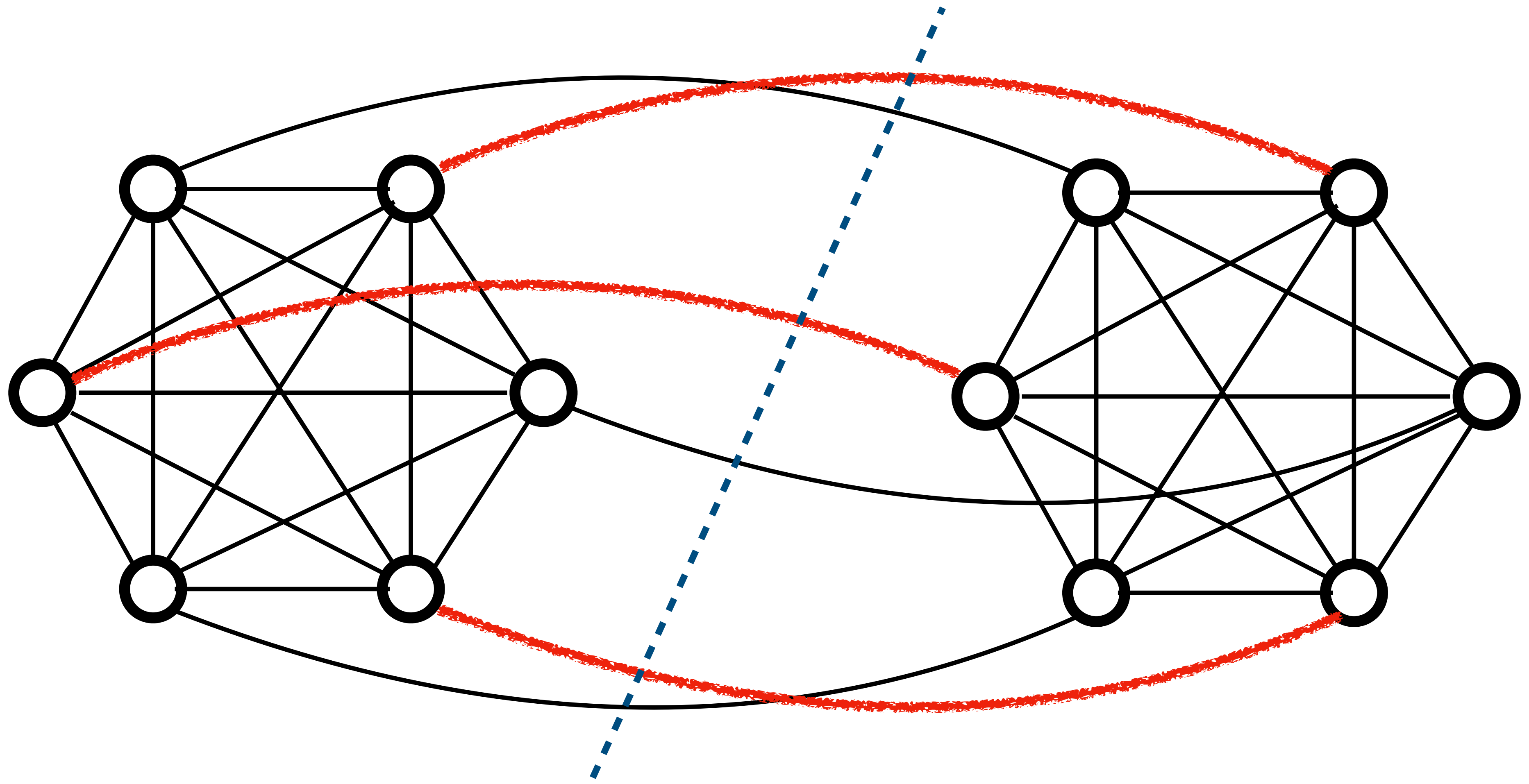
# Takeaway: Robustness is a Structural Property

The structure of the comparison graph plays a crucial role in determining resilience to adversarial corruption.

Bad news! Certain topologies are fundamentally vulnerable to adversarial contamination.

For such topologies, even a marginal amount of corruption can make parameter recovery fundamentally impossible.

Fraction of corrupted edges incident on any vertex is $\leq O(1/n)$, yet the cut majority condition fails.



Sparse cuts across dense subgraphs can easily be exploited, even by a limited budget adversary!

# Outline

**Preliminaries**

- Bradley-Terry-Luce Model
- Comparison Graphs

**Adversarial Contamination Model**

**Condition for Unique Identifiability**

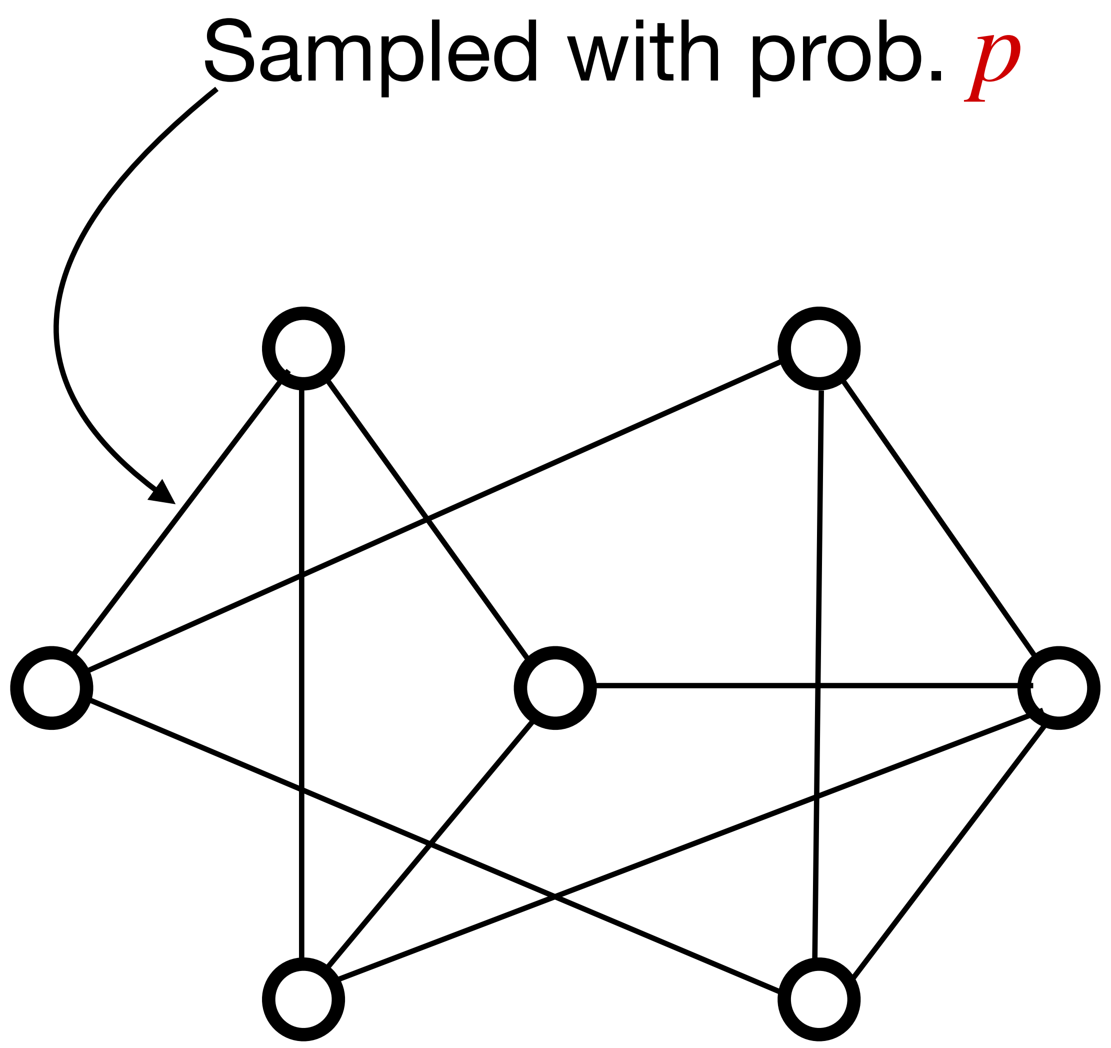
- Robustness as a Structural Property

**Results for Erdős-Rényi Comparison Graphs**

- A Sharp Threshold Condition for Identifiability
- Algorithm for Parameter Recovery

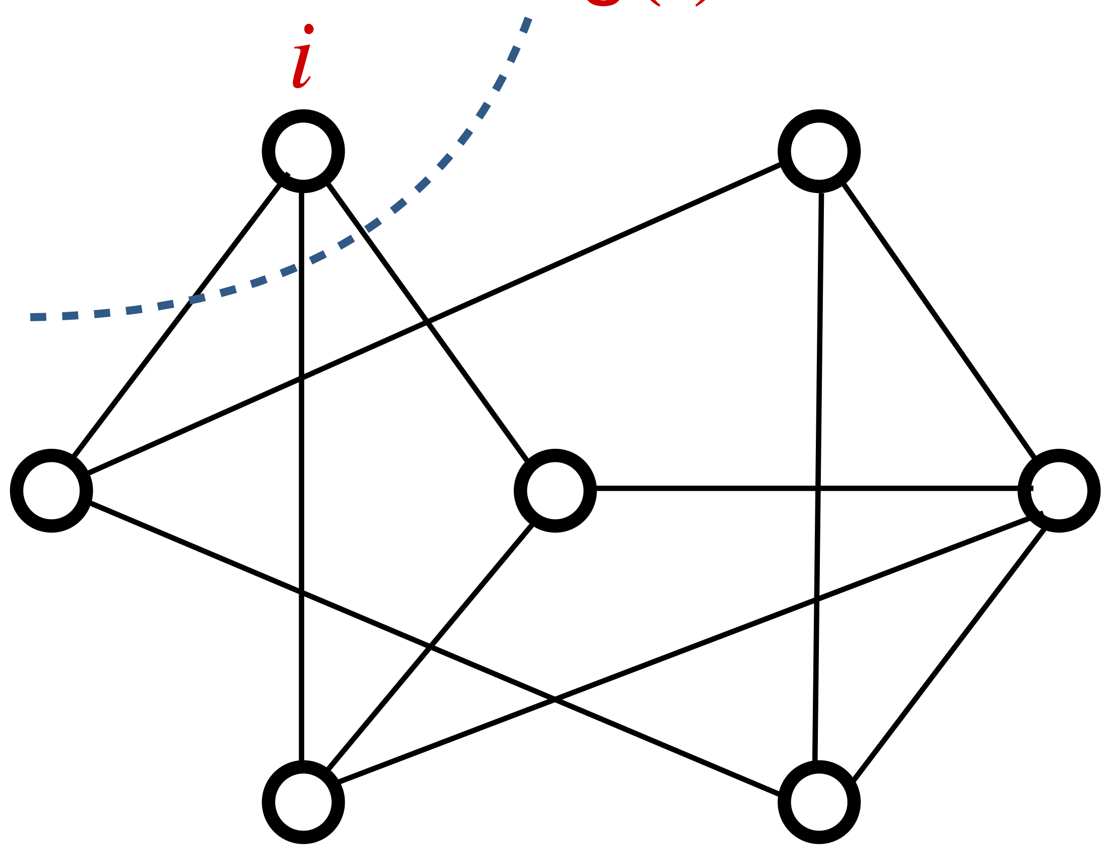
# Erdős-Rényi Graphs are Highly Robust to Contamination

Given a parameter $p \in [0,1]$, an Erdős-Rényi graph $G_{n,p}$ is a random graph over $n$ vertices, where each edge $\{i, j\}$ is sampled independently with probability $p$.



These graphs exhibit strong connectivity properties due to which they can tolerate large degrees of corruption.

# Budget Constrained Adversary

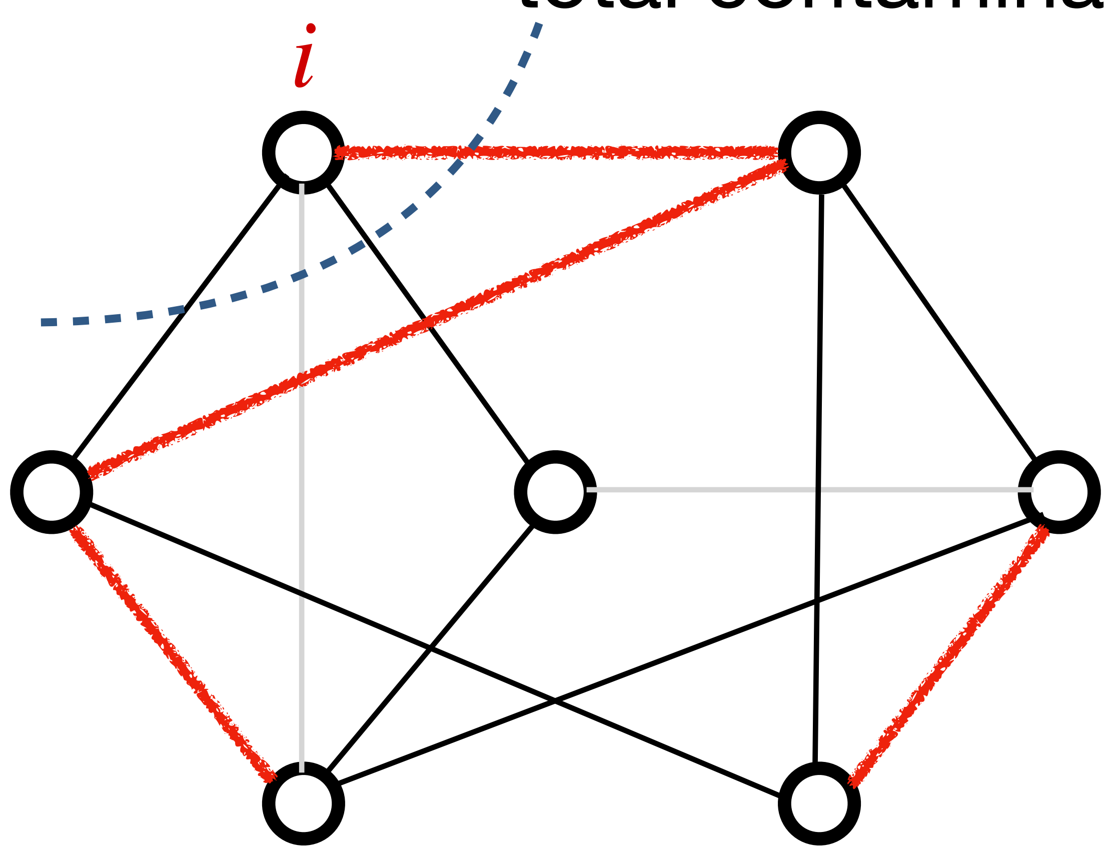


deg$(i)$ in $G^* \overset{\text{def}}{=} E^*(i)$

Nature draws an ER comparison graph $G^* \sim G_{n,p}$ with $p \geq (c \log n)/n$ for a sufficiently large constant $c$

**Adversary**

$$\xrightarrow[\phantom{xxxxxxxxxxxx}]{\text{“Budget” limited by } \gamma < 1}$$



total contamination $\leq \gamma\, E^*(i)$

Contaminated graph $G$ such that for any vertex $i$, |additions + deletions + corruptions| incident on $i$ in $G$ is $\leq \gamma\, E^*(i)$

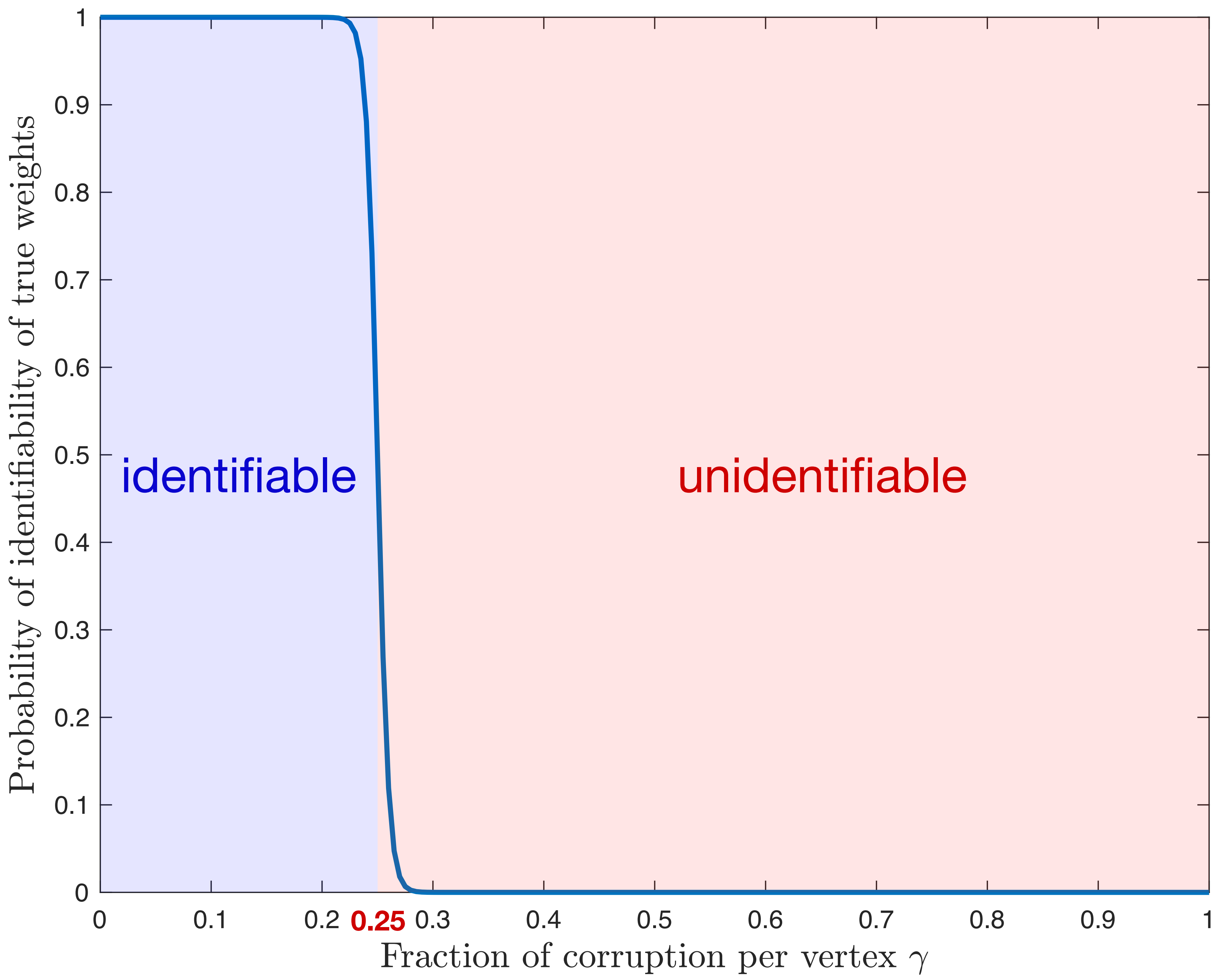
# Sharp Threshold Condition for Identifiability

**Theorem 2. (Vertex Majority Condition)** *For $G^* \sim G_{n,p}$ , with high probability over the generation of the graph,*

*The true weights are uniquely identifiable in the contaminated graph $G$ if the fraction of total contamination per vertex $\gamma < 1/4 - \epsilon$.*

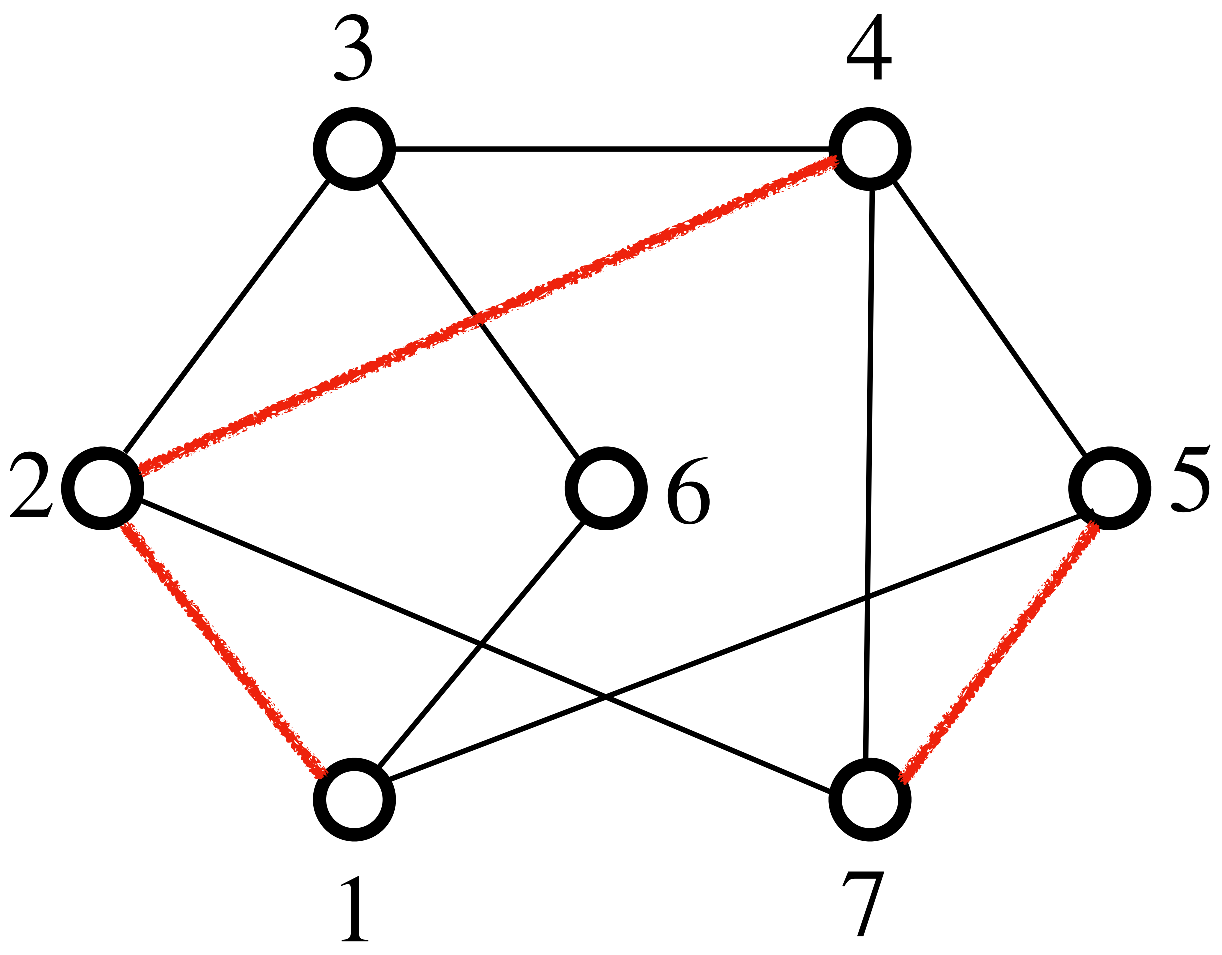
*Conversely, the true weights are are not uniquely identifiable in the contaminated graph $G$ if the fraction of total contamination per vertex $\gamma > 1/4 + \epsilon$,*

*where $\epsilon$ is any arbitrarily small positive constant.*



# Recovery Algorithm for Erdős-Rényi Graphs

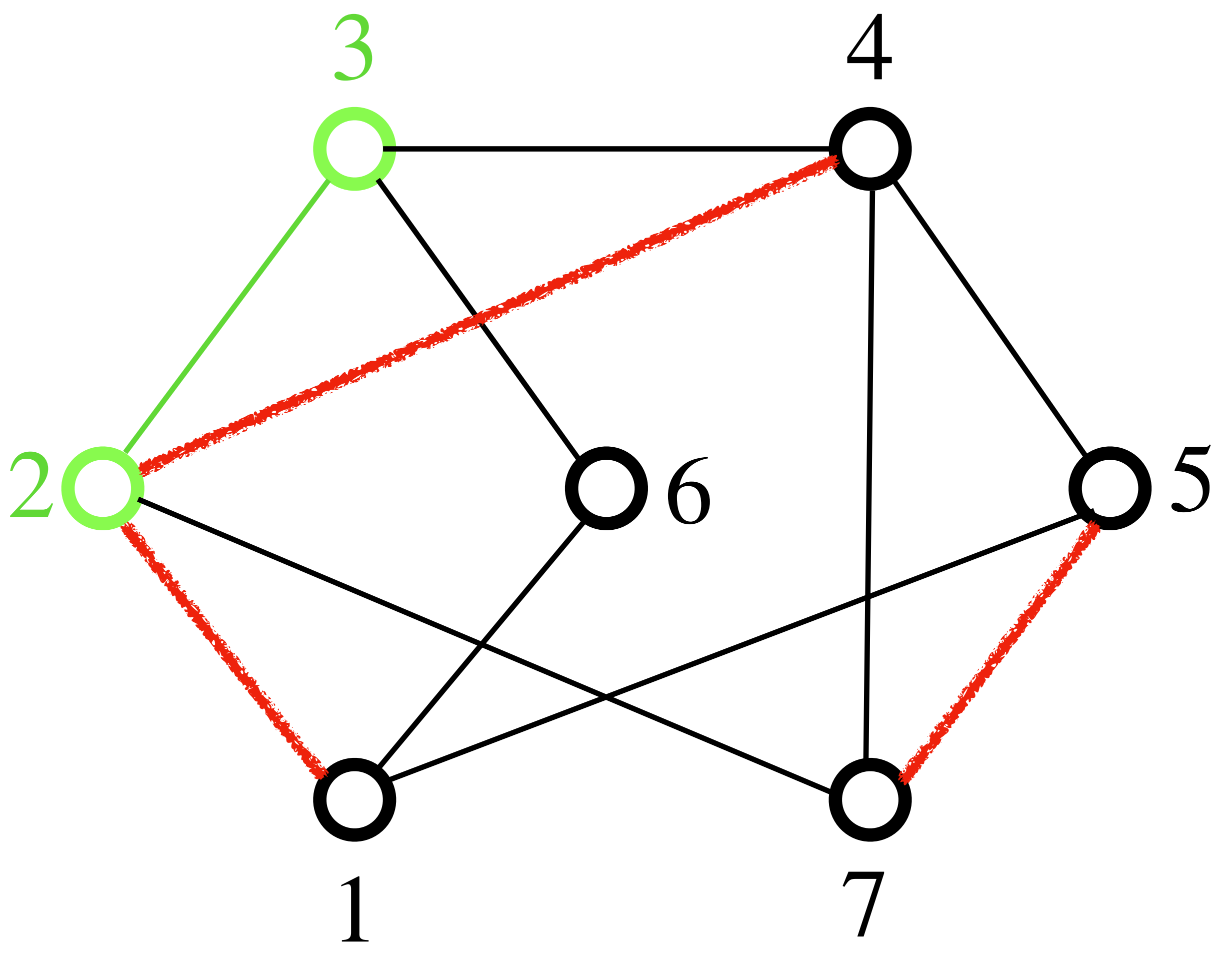
**Idea.** Corrupted edge labels are detectable.



# Recovery Algorithm for Erdős-Rényi Graphs

**Idea.** Corrupted edge labels are detectable.

The ratio of probabilities $p_{ij}/p_{ji}$ approximately determines the relative ratio of weights $w_i/w_j$.

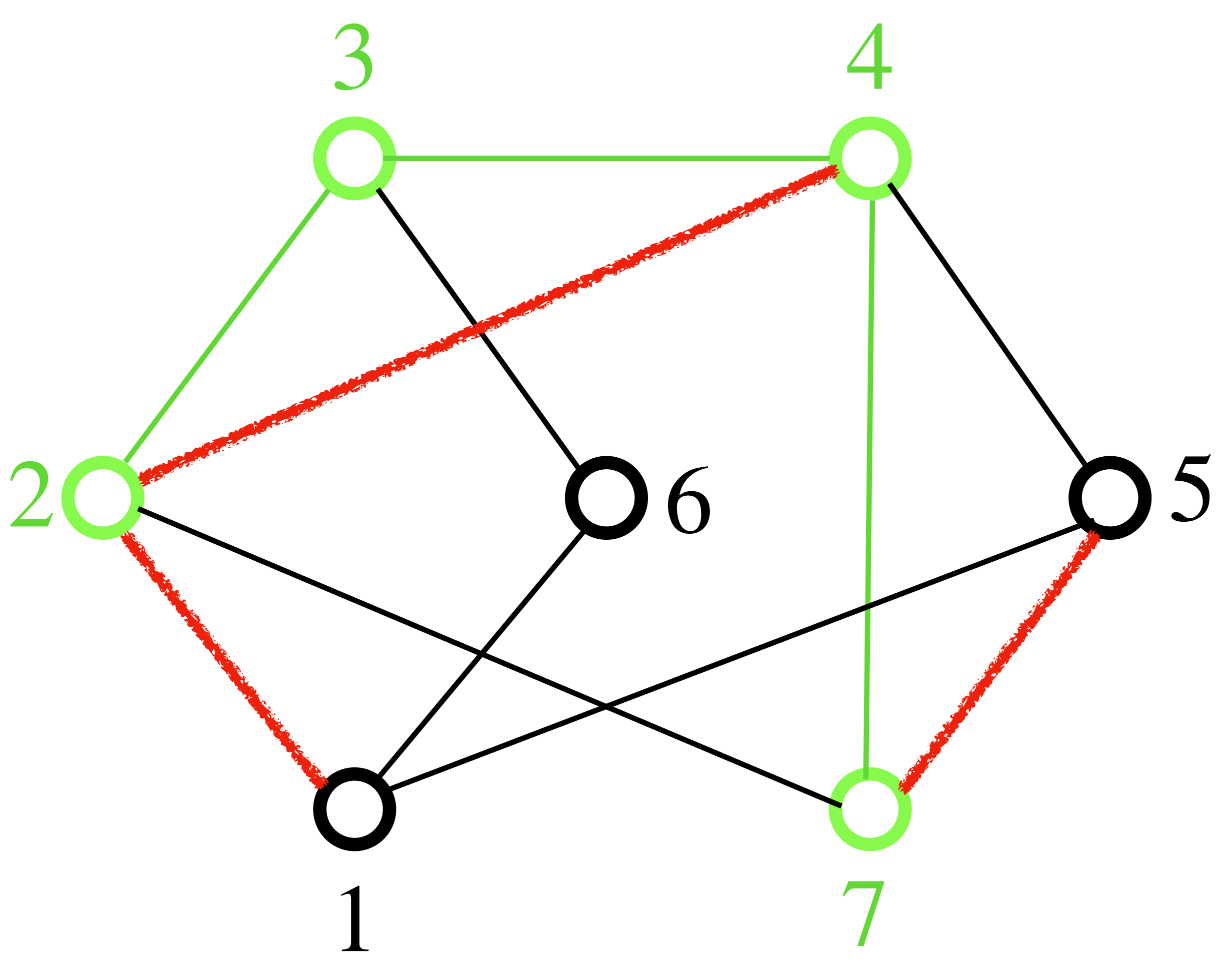


# Recovery Algorithm for Erdős-Rényi Graphs

**Idea.** Corrupted edge labels are detectable.

The ratio of probabilities $p_{ij}/p_{ji}$ approximately determines the relative ratio of weights $w_i/w_j$.

$\implies$ For a path $(i_1, \ldots, i_l)$, the product $\prod_{k=1}^{l} p_{i_k i_{k+1}}/p_{i_{k+1} i_k}$ approximately determines the relative ratio $w_{i_1}/w_{i_l}$.



# Recovery Algorithm for Erdős-Rényi Graphs

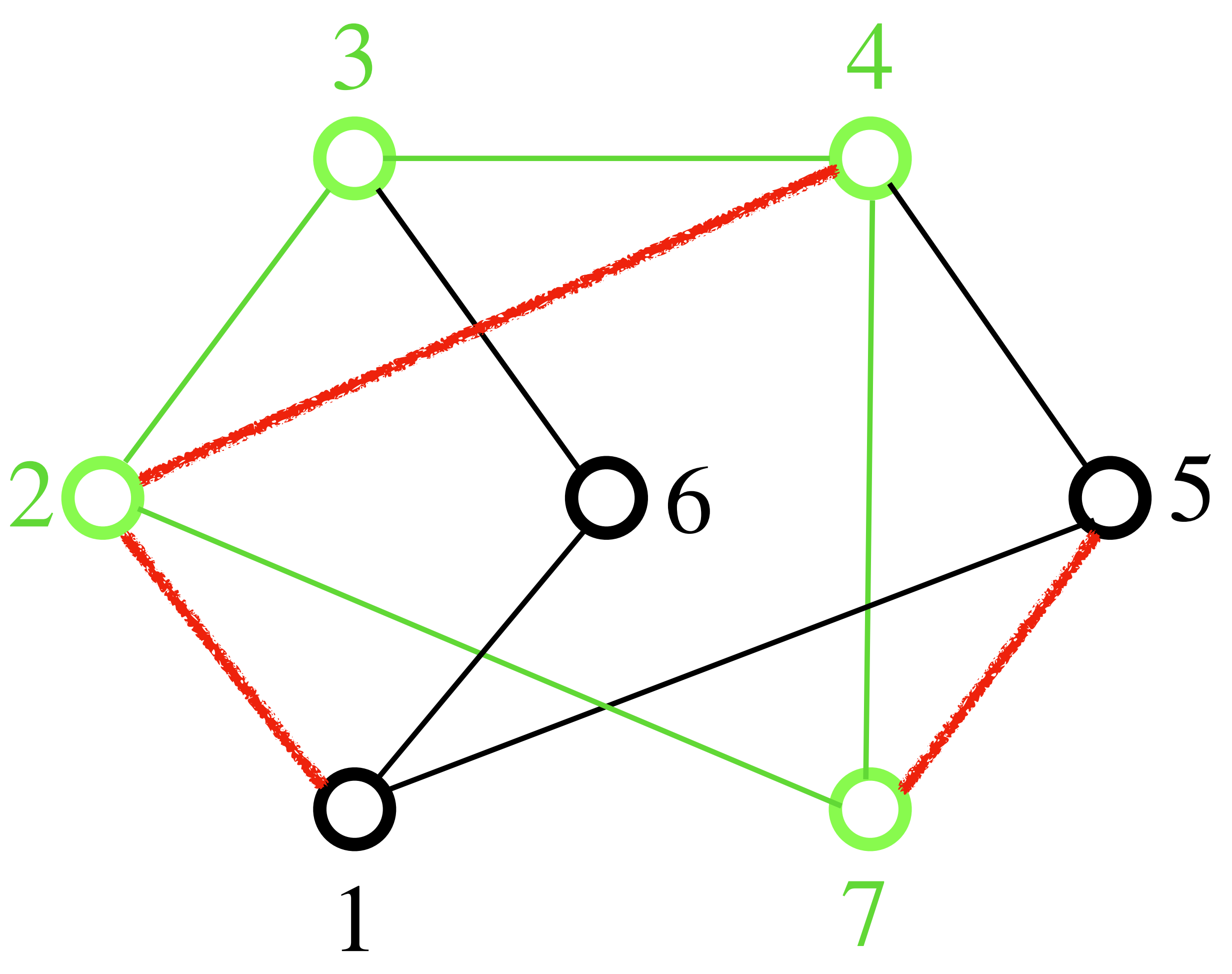
**Idea.** Corrupted edge labels are detectable.

The ratio of probabilities $p_{ij}/p_{ji}$ approximately determines the relative ratio of weights $w_i/w_j$.

$\Longrightarrow$ For a path $(i_1, \ldots, i_l)$, the product $\prod_{k=1}^{l} p_{i_k i_{k+1}}/p_{i_{k+1} i_k}$ approximately determines the relative ratio $w_{i_1}/w_{i_l}$.

In cycle $(2,3,4,7,2)$, which consists of only good edges,

$$\frac{p_{23}}{p_{32}}\frac{p_{34}}{p_{43}}\frac{p_{47}}{p_{74}}\frac{p_{72}}{p_{27}} \approx 1$$

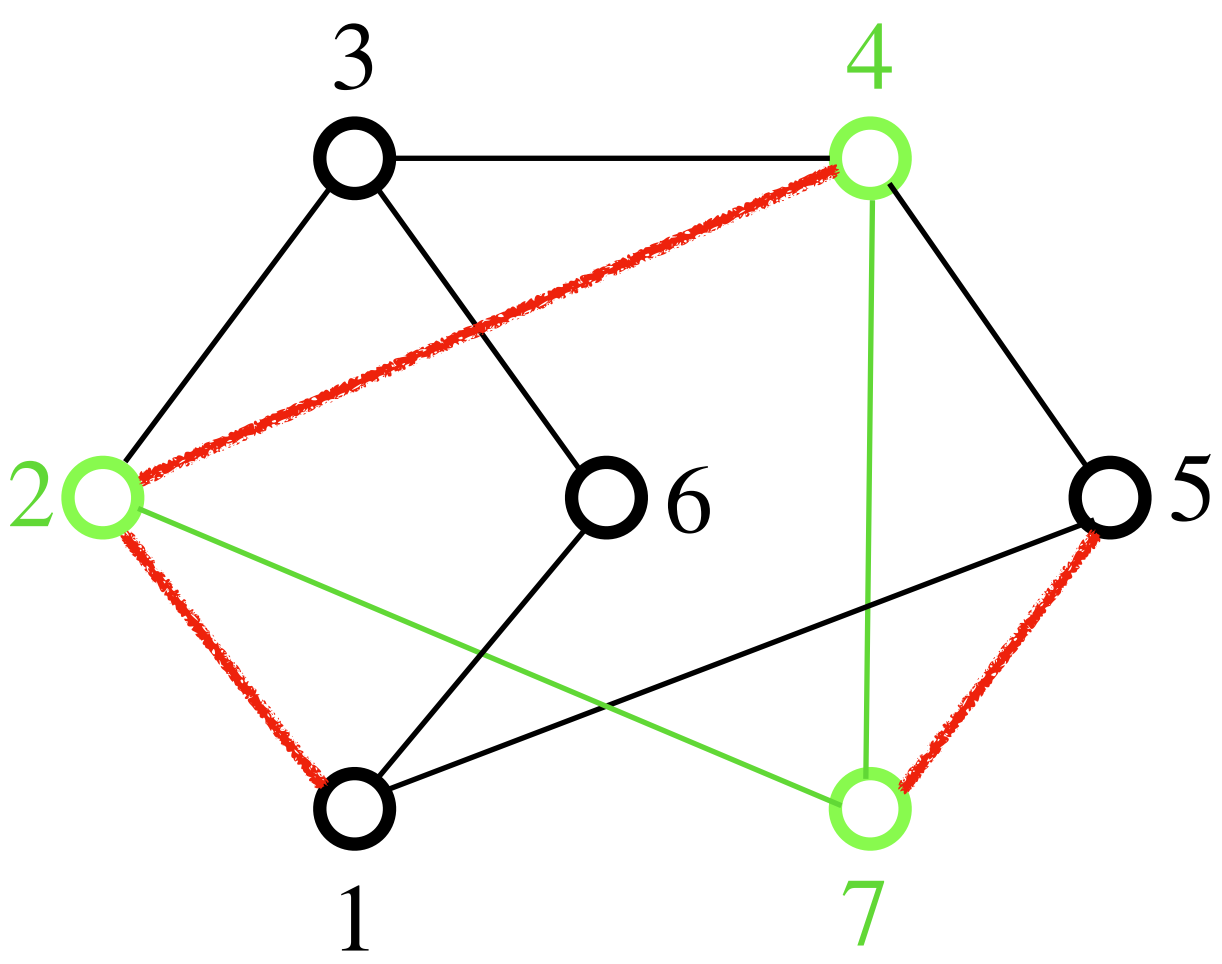


# Recovery Algorithm for Erdős-Rényi Graphs

**Idea.** Corrupted edge labels are detectable.

The ratio of probabilities $p_{ij}/p_{ji}$ approximately determines the relative ratio of weights $w_i/w_j$.

$\Longrightarrow$ For a path $(i_1, \ldots, i_l)$, the product $\prod_{k=1}^{l} p_{i_k i_{k+1}}/p_{i_{k+1} i_k}$ approximately determines the relative ratio $w_{i_1}/w_{i_l}$.



# Recovery Algorithm for Erdős-Rényi Graphs

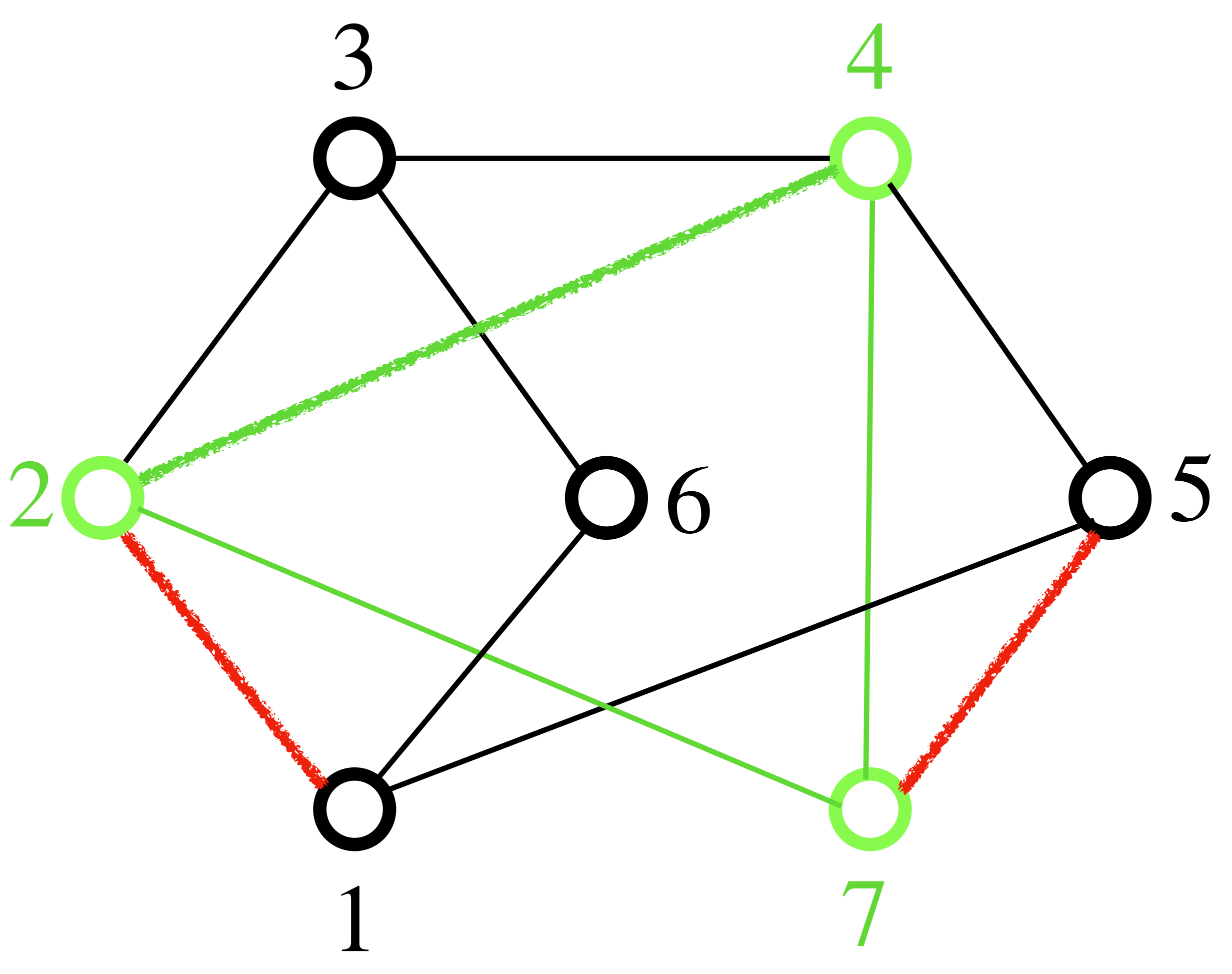
**Idea.** Corrupted edge labels are detectable.

The ratio of probabilities $p_{ij}/p_{ji}$ approximately determines the relative ratio of weights $w_i/w_j$.

$\implies$ For a path $(i_1, \ldots, i_l)$, the product $\prod_{k=1}^{l} p_{i_k i_{k+1}}/p_{i_{k+1} i_k}$ approximately determines the relative ratio $w_{i_1}/w_{i_l}$.

In cycle $(2,4,7,2)$, which contains a bad edge,

$\frac{p_{24}}{p_{42}}\frac{p_{47}}{p_{74}}\frac{p_{72}}{p_{27}}$ bounded away from $1$



**Inconsistent cycles provide evidence of corruption!**

# Linear Programming Relaxation

Given contaminated graph $G = ([n], E)$, decision variables $x(e)$ for every edge $e \in E$,

| | | | |
|---|---|---|---|
| Minimize | $\sum_{e \in E} x(e)$ | | Identify edges whose deletion leaves us with a consistent subgraph |
| Subject To | $\sum_{e \in C} x(e) \geq 1$ | $\forall C \in \mathbb{C}$ | Hitting set constraint for inconsistent cycles |
| | $\sum_{e \in E(v)} x(e) \leq \gamma \lvert E(v) \rvert$ | $\forall v \in V$ | Budget constraint for every vertex |
| | $0 \leq x(e) \leq 1$ | $\forall e \in E$ | Relaxation of integer constraint for edges |

Feasible and computationally efficient because of strong connectivity properties of Erdős-Rényi graphs. For sparse graphs, this can be done in near quadratic $O(n^{2+o(1)})$ time.

# Linear Programming Rounding

Given a feasible solution $\mathbf{x} \in [0,1]^{|E|}$ to the LP, discard any edge $e \in E$ with $x(e) > T$ for a suitably chosen threshold $T$, resulting in a pruned graph $\tilde{G} = ([n], \tilde{E})$.

The surviving graph $\tilde{G}$ is connected.
The labels on surviving edges $e \in \tilde{E}$ satisfy a uniform deviation bound.

This is non-trivial to prove. We prove a robust connectivity property of Erdős-Rényi graphs, which forms a central part of the proof of the above two claims.

Pass this pruned graph to any existing non-robust estimation algorithm.
We chose ASR *[Agarwal et al., 2018]*, and analyzed it to give guarantees for the recovered weights.

# Recovery Guarantees

**Theorem 3.** *When the initial truthful comparison graph* $G^* \sim G_{n,p}$ *, and the fraction of total contamination per-vertex in the contaminated graph* $\gamma = O\left(\frac{\log np}{\log n}\right)$*, the LP based algorithm given*

***perfect data*** ( $\hat{p}_{ij} = p^*_{ij}$ for $\{i, j\} \in G^*$ ) *recovers the true weights exactly.*

***sampled data*** ( $\hat{p}_{ij} \sim \text{Binomial}(p^*_{ij}, L)$ for $\{i, j\} \in G^*$ ) *returns an estimate* $\mathbf{w}$ *such that*

$$\|\mathbf{w} - \mathbf{w}^*\|_1 \leq O\left(\sqrt{\frac{\log n}{L}}\right).$$

The former guarantee is a special case of the latter, corresponding to the limit $L \to \infty$.

# Remarks

Sample complexity bounds match the best known bounds for the uncontaminated setting up to constants. Robustness to adversarial contamination comes with no statistical penalty!*

*How much contamination can we tolerate?

Sparse regime [$O(n \log n)$ edges] $\Longrightarrow$ $O(\log \log n / \log n)$ fraction of corrupted comparisons per item.

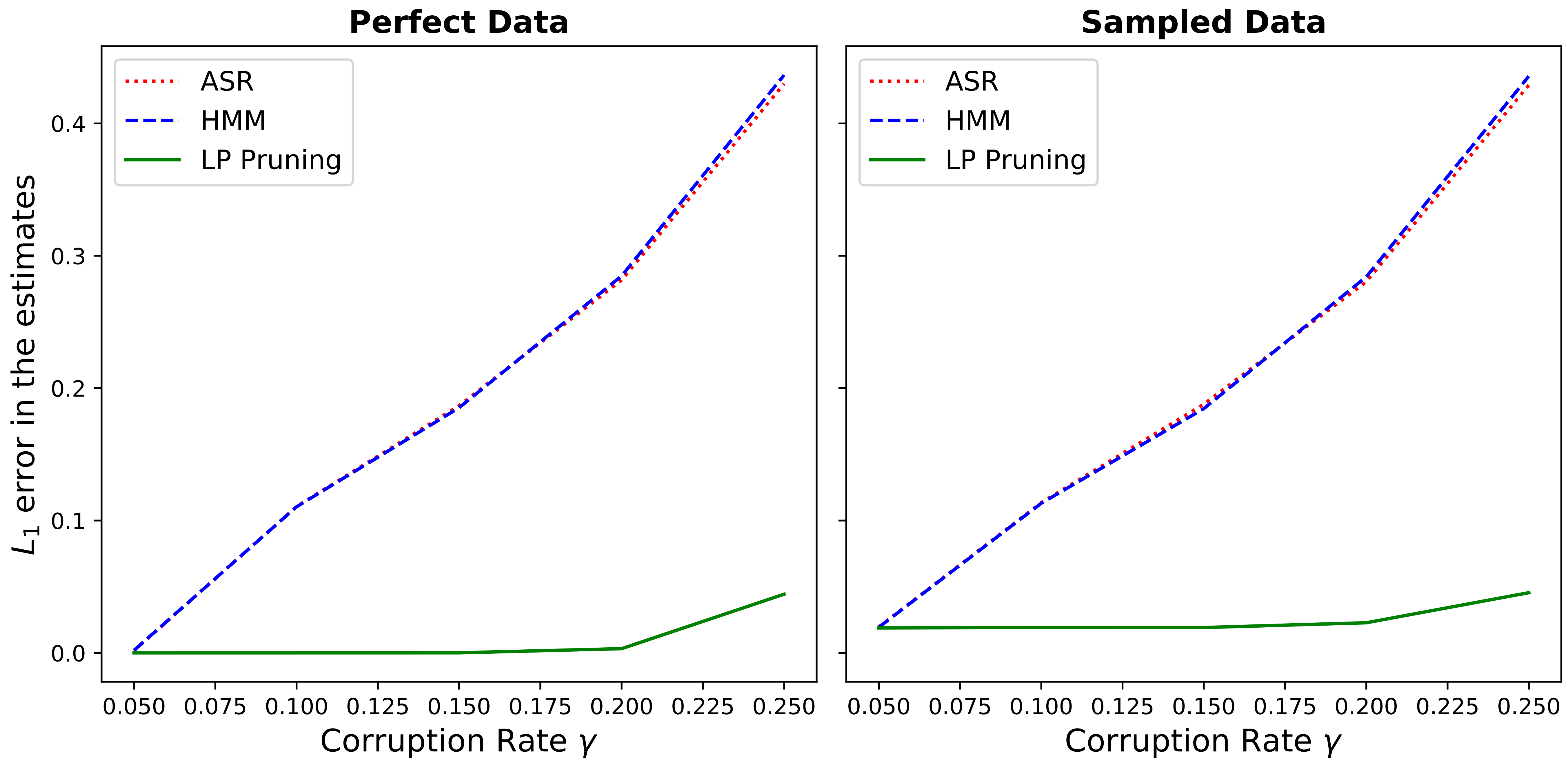
Dense regime [$O(n^{1+\epsilon})$ edges, for any constant $\epsilon$] $\Longrightarrow$ constant fraction of corrupted comparisons per item.

# Experiments

**Error in the returned estimates with an increasing corruption rate in synthetic data.**

ASR - Accelerated Spectral Ranking, (Spectral Method) [*Agarwal et al., 2018*]

HMM - Hunter's Minorization-Maximization, (Maximum Likelihood) [*Hunter, 2004*]



# Conclusion

We initiated the study of robustness in rank aggregation in the BTL model by introducing a powerful contamination model, under which

- ★ We characterized the exact necessary and sufficient condition for structural identifiability of the true BTL parameters for arbitrary comparison graphs.
  - ‣ Robustness is a structural property of comparison graphs. One cannot hope to be robust for arbitrary topologies.
- ★ For the family of Erdős-Rényi comparison graphs, we proved a simpler necessary and sufficient condition for identifiability.
  - ‣ Identifiability in Erdős-Rényi comparison graphs has a sharp threshold at ~25% corruption per item.
- ★ For Erdős-Rényi comparison graphs, we provided an efficient linear-programming based algorithm for parameter estimation that could tolerate up to $O(\log\log n / \log n)$ fraction corruption per item in the sparse regime, and constant fraction corruption per item in the dense regime.
  - ‣ Sample complexity bounds match the usual uncontaminated setting up to constants.

# Thank You!