Topology Discovery Using Few Participants

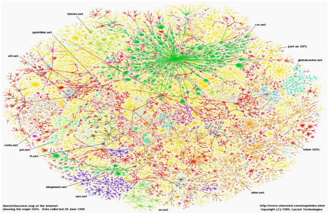
Anima Anandkumar

U.C. Irvine

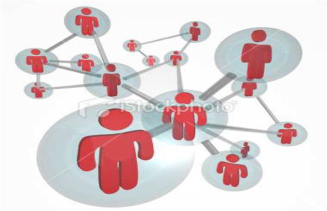
Joint work with Avinatan Hassidim and Jonathan Kelner.

Topology Discovery in Large Networks

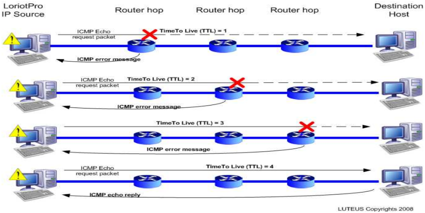
Internet Mapping



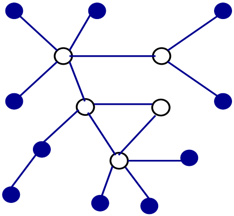
Social Network Mapping



Traceroute



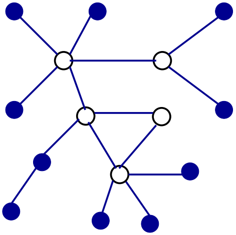
Tomography



Analysis of Network Tomography Approaches

Problem Formulation

- End-to-end measurements between uniformly chosen participants
  - For example, (random) delay measurements
- Unknown delay distribution, number of hidden nodes and topology.
- Topology is Erdős-Rényi random graph $G_n \sim \mathcal{G}(n, c/n)$: each edge has probability c/n.

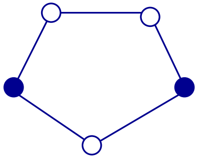


How many participants needed to reconstruct efficiently?

Problem Formulation

Two Scenarios for Path Measurements

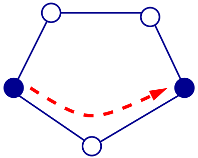
- Scenario 1: shortest-path delays among participants
- Scenario 2: delays along shortest paths and second shortest paths



Problem Formulation

Two Scenarios for Path Measurements

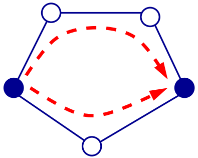
- Scenario 1: shortest-path delays among participants
- Scenario 2: delays along shortest paths and second shortest paths



Problem Formulation

Two Scenarios for Path Measurements

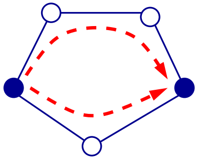
- Scenario 1: shortest-path delays among participants
- Scenario 2: delays along shortest paths and second shortest paths



Problem Formulation

Two Scenarios for Path Measurements

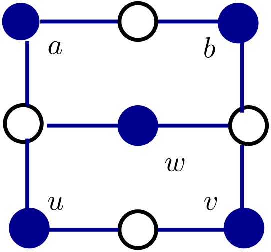
- Scenario 1: shortest-path delays among participants
- Scenario 2: delays along shortest paths and second shortest paths

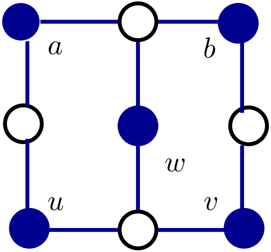


Reconstruction of Minimal Representation

Best possible reconstruction using any algorithm

Not All Graphs are Discoverable





Summary of Results

Topology discovery of $G_n \sim \mathcal{G}(n, c/n)$.

- For scenario 1, a sub-linear edit distance achieved with a sub-linear number of participants
  - $n^{0.75}$ participants for homogeneous setting (identical link delay distributions)

Summary of Results

Topology discovery of $G_n \sim \mathcal{G}(n, c/n)$.

- For scenario 1, a sub-linear edit distance achieved with a sub-linear number of participants
  - $n^{0.75}$ participants for homogeneous setting (identical link delay distributions)
- For scenario 2, consistent reconstruction is possible using a sub-linear number of participants
  - $n^{0.875}$ nodes for homogeneous setting

Summary of Results

Topology discovery of $G_n \sim \mathcal{G}(n, c/n)$.

- For scenario 1, a sub-linear edit distance achieved with a sub-linear number of participants
  - $n^{0.75}$ participants for homogeneous setting (identical link delay distributions)
- For scenario 2, consistent reconstruction is possible using a sub-linear number of participants
  - $n^{0.875}$ nodes for homogeneous setting
- The above results achieved when number of delay samples is $\Omega(\text{poly} \log n)$

Efficient Reconstruction Using Few Participants

Summary of Results

Topology discovery of $G_n \sim \mathcal{G}(n, c/n)$.

- For scenario 1, a sub-linear edit distance achieved with a sub-linear number of participants
  - $n^{0.75}$ participants for homogeneous setting (identical link delay distributions)
- For scenario 2, consistent reconstruction is possible using a sub-linear number of participants
  - $n^{0.875}$ nodes for homogeneous setting
- The above results achieved when number of delay samples is $\Omega(\text{poly} \log n)$

Efficient Reconstruction Using Few Participants

- Lower bound on graph reconstruction
  - $n^{0.5}$ nodes needed for reconstruction up to certain edit distance

Related Work

Practice: Mapping Internet/Social Networks

Eriksson et. al ('07), Gomez-Rodriguez et. al ('10)

Related Work

Practice: Mapping Internet/Social Networks

Eriksson et. al ('07), Gomez-Rodriguez et. al ('10)

Theory: Query-based Analysis

- Different kinds of queries: Shortest paths, distances, edges etc.
- Assume labels of all nodes and (mostly) unweighted graphs. Provide approximation guarantees.

Related Work

Practice: Mapping Internet/Social Networks

Eriksson et. al ('07), Gomez-Rodriguez et. al ('10)

Theory: Query-based Analysis

- Different kinds of queries: Shortest paths, distances, edges etc.
- Assume labels of all nodes and (mostly) unweighted graphs. Provide approximation guarantees.

Theory: Tree Reconstruction

- Reconstruction of a tree using end-to-end measurements among the leaves (Ni et. al, Shih & Hero).
- Assumes no information about hidden nodes.
- Not applicable for loopy graphs.

Outline

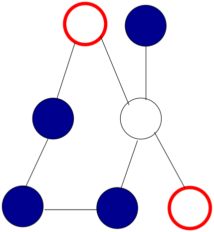
1. Introduction
2. Algorithms for Topology Discovery
   - Setup
   - Recap of Tree Reconstruction
   - Proposed Algorithms and Reconstruction Guarantees
   - Lower Bound on Topology Discovery
3. Conclusion

Outline

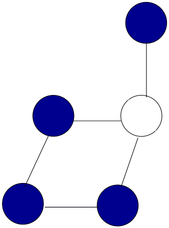
1. Introduction
2. Algorithms for Topology Discovery
   - Setup
   - Recap of Tree Reconstruction
   - Proposed Algorithms and Reconstruction Guarantees
   - Lower Bound on Topology Discovery
3. Conclusion

Minimal Representation for Graph Reconstruction

Identifiability of the graph given participants



A non-minimal graph

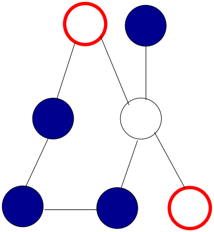


Minimal representation

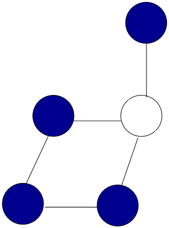
Reconstruction of minimal representation

Minimal Representation for Graph Reconstruction

Identifiability of the graph given participants



A non-minimal graph



Minimal representation

Reconstruction of minimal representation

- Assumption can be removed if degree of all nodes have degree 3 or higher (random regular family, degree distribution graphs)
- Original graph can be obtained from minimal representation with additional information

Delay Moments as Edge Lengths

- D_e : random delay along a link $e \in G_n$
- Delays along any two links are independent.
- Delays are additive along any route

$$D_{i,j} = \sum_{(k,l) \in \mathrm{Path}(i,j)} D_{k,l}$$

- Bounded moments of some fixed order, e.g., bounded variances $0 < f \leq l(e) \leq g < \infty$, where $l(e) = \mathsf{Var}(D_e)$.

Delay Moments as Edge Lengths

- D_e : random delay along a link $e \in G_n$
- Delays along any two links are independent.
- Delays are additive along any route

$$D_{i,j} = \sum_{(k,l) \in \mathrm{Path}(i,j)} D_{k,l}$$

- Bounded moments of some fixed order, e.g., bounded variances
  $0 < f \leq l(e) \leq g < \infty$, where $l(e) = \mathsf{Var}(D_e)$.

Moments of Delay Distribution Form an Additive Metric on Graph

Delay Moments as Edge Lengths

- D_e : random delay along a link $e \in G_n$
- Delays along any two links are independent.
- Delays are additive along any route

$$D_{i,j} = \sum_{(k,l) \in \mathrm{Path}(i,j)} D_{k,l}$$

- Bounded moments of some fixed order, e.g., bounded variances $0 < f \leq l(e) \leq g < \infty$, where $l(e) = \mathsf{Var}(D_e)$.

Moments of Delay Distribution Form an Additive Metric on Graph

Moment Estimator

$\widehat{l}^m(i,j) := \frac{1}{m-1} \sum_{k=1}^{m} (D_{i,j}(k) - \bar{D}^m_{i,j})^2$, where $\bar{D}^m_{i,j}$ is the sample mean

Topology Discovery Based on Distance Estimates

In this talk, analysis when exact statistics are available

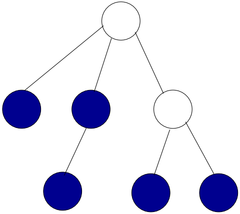
Outline

1. Introduction

2. Algorithms for Topology Discovery
   - Setup
   - Recap of Tree Reconstruction
   - Proposed Algorithms and Reconstruction Guarantees
   - Lower Bound on Topology Discovery

3. Conclusion

Reconstruction of Trees with Hidden Nodes



Setup

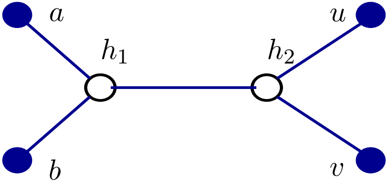
- Topology is a tree
- No knowledge about hidden nodes

Distance-Based Methods

- End-to-end measurements between observed nodes
- Additive metric on the tree

M.J. Choi, V. Tan, A. Anandkumar & A. Willsky, "Learning Latent Tree Graphical Models," *J. of Machine Learning Research*, volume 12, pp. 1771-1812, May 2011.

Quartet Tests



Quartet $Q(ab|uv)$

Quartet or Four-Point Condition

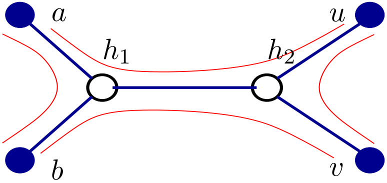
The pairwise distances $\{l(i,j)\}_{i,j \in \{a,b,u,v\}}$ satisfy

$$l(a,b) + l(u,v) < \min(l(a,u) + l(b,v), l(b,u) + l(a,v)).$$

Inference of Internal Distances

- 6 distances, 1 equality constraint and 5 unknowns
- Internal distances can be determined.

Quartet Tests



Quartet $Q(ab|uv)$

Quartet or Four-Point Condition

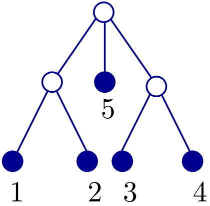
The pairwise distances $\{l(i,j)\}_{i,j\in\{a,b,u,v\}}$ satisfy

$$l(a,b)+l(u,v)<\min(l(a,u)+l(b,v),l(b,u)+l(a,v)).$$

Inference of Internal Distances

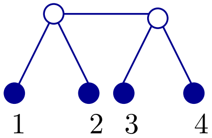
- 6 distances, 1 equality constraint and 5 unknowns
- Internal distances can be determined.

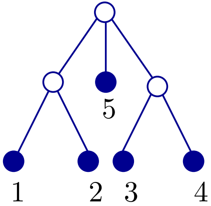
Merging Quartets for Tree Discovery (Pearl)



Unknown Tree

Merging Quartets for Tree Discovery (Pearl)

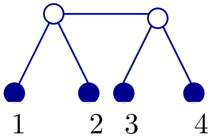


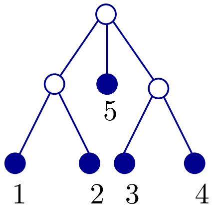


Unknown Tree

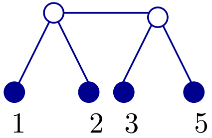
Set of Quartets

Merging Quartets for Tree Discovery (Pearl)



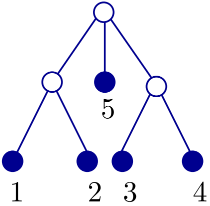


Unknown Tree

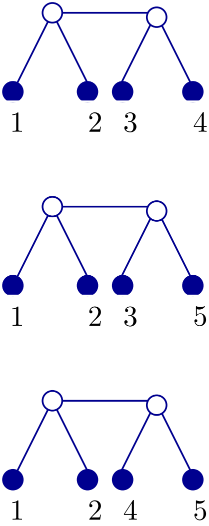


Set of Quartets

Merging Quartets for Tree Discovery (Pearl)

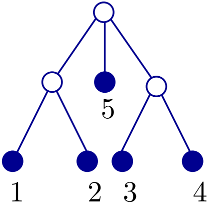


Unknown Tree

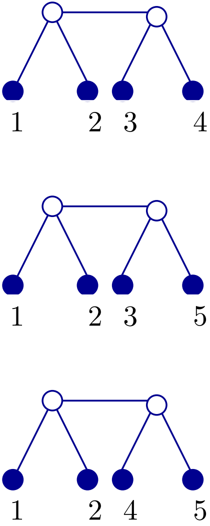


Set of Quartets

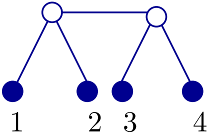
Merging Quartets for Tree Discovery (Pearl)



Unknown Tree

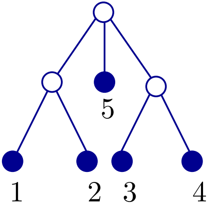


Set of Quartets

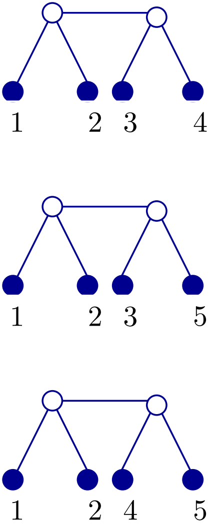


Quartet Merging

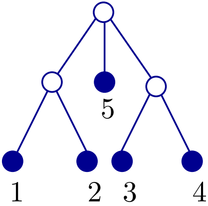
Merging Quartets for Tree Discovery (Pearl)



Unknown Tree



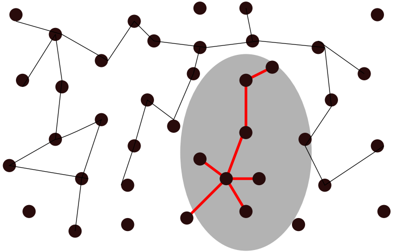
Set of Quartets



Quartet Merging

From Trees to Random Graphs

Random Graphs are Locally Tree-Like

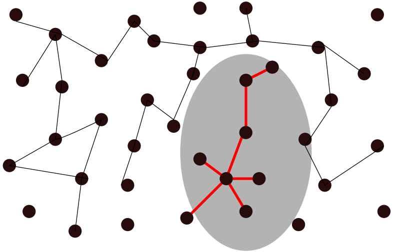


As # of nodes $p \to \infty$,

- Typical nbd. (up to $O(\log p)$) has no cycles
- Constant # of short cycles
- Short cycles do not overlap

From Trees to Random Graphs

Random Graphs are Locally Tree-Like



As # of nodes $p \to \infty$,

- Typical nbd. (up to $O(\log p)$) has no cycles
- Constant # of short cycles
- Short cycles do not overlap

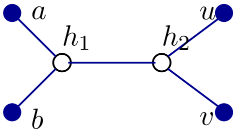
Direct application of Quartet merging not possible

Outline

Algorithm Under Scenario 1

Scenario 1

Shortest-path delays among participants



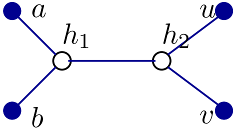
Short Quartet

- Test for four-point condition only when all distances less than $Rg + \tau$, where g is upper bound on edge lengths
- Merge quartets upon testing

Algorithm Under Scenario 1

Scenario 1

Shortest-path delays among participants

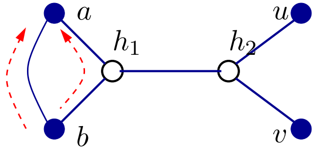


Short Quartet

- Test for four-point condition only when all distances less than $Rg + \tau$, where g is upper bound on edge lengths
- Merge quartets upon testing

Sources of errors

- Absence of short quartets: no close participants
- Presence of short cycles: Small (constant) number of short cycles in random graphs



Reconstruction Guarantee for Algorithm RGD1

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths

Reconstruction Guarantee for Algorithm RGD1

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths

Assumptions

$$R \leq \frac{f}{g}\left(2 + 2\frac{\log(n^{0.75}/3)}{\log c}\right)$$

Reconstruction Guarantee for Algorithm RGD1

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths

Assumptions

$R \leq \frac{f}{g}\left(2 + 2\frac{\log(n^{0.75}/3)}{\log c}\right)$ and $\rho c^{\frac{R}{2}} = \omega(1)$ or $\gamma > 2\beta$

Reconstruction Guarantee for Algorithm RGD1

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths

Assumptions

$R \leq \frac{f}{g}\left(2 + 2\frac{\log(n^{0.75}/3)}{\log c}\right)$ and $\rho c^{\frac{R}{2}} = \omega(1)$ or $\gamma > 2\beta$

Theorem: Edit Distance Guarantee

The algorithm RGD1 recovers the minimal representation $\widetilde{G}_n$ of the giant component of a.e. graph $G_n \sim \mathcal{G}(n, c/n)$ with edit distance

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{4\gamma g/f - 4\beta}).}$$

Reconstruction Guarantee for Algorithm RGD1

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths

Assumptions

$R \leq \frac{f}{g}\left(2 + 2\frac{\log(n^{0.75}/3)}{\log c}\right)$ and $\rho c^{\frac{R}{2}} = \omega(1)$ or $\gamma > 2\beta$

Theorem: Edit Distance Guarantee

The algorithm RGD1 recovers the minimal representation $\widetilde{G}_n$ of the giant component of a.e. graph $G_n \sim \mathcal{G}(n, c/n)$ with edit distance

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{4\gamma g/f - 4\beta}).}$$

$n^{0.75}$ nodes needed for sublinear edit distance for homogeneous case

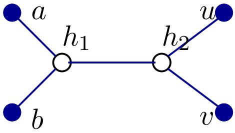
Algorithm Under Scenario 2

Scenario 2

Delays along shortest and second shortest paths

Short Quartet

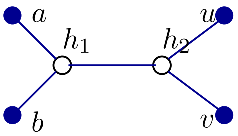
- Consider shortest-path and second shortest distances less than $Rg + \tau$
- Test for four-point condition for different combinations
- Merge quartets upon testing



Algorithm Under Scenario 2

Scenario 2

Delays along shortest and second shortest paths

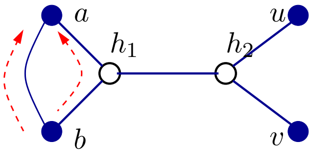


Short Quartet

- Consider shortest-path and second shortest distances less than $Rg + \tau$
- Test for four-point condition for different combinations
- Merge quartets upon testing

Source of Errors

- Absence of short quartets: no close by participants
- Presence of overlapping short cycles: No overlapping short cycles in random graphs



Reconstruction Guarantee for Algorithm RGD2

Notation

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths
- Same assumptions as before

Reconstruction Guarantee for Algorithm RGD2

Notation

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths
- Same assumptions as before

Theorem: Edit Distance Under RGD2

The algorithm RGD2 recovers the minimal representation $\widetilde{G}_n$ of the giant component of a.e. graph $G_n \sim \mathcal{G}(n, c/n)$ with edit distance

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{6\gamma g/f - 4\beta - 1}).}$$

Reconstruction Guarantee for Algorithm RGD2

Notation

- $R = \dfrac{\gamma \log n}{\log c}$: Parameter for short quartet
- $\rho_n = n^{-\beta}$: Fraction of participating nodes
- f and g: lower and upper bounds on edge lengths
- Same assumptions as before

Theorem: Edit Distance Under RGD2

The algorithm RGD2 recovers the minimal representation $\widetilde{G}_n$ of the giant component of a.e. graph $G_n \sim \mathcal{G}(n, c/n)$ with edit distance

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{6\gamma g/f - 4\beta - 1}).}$$

Compare with edit distance under RGD1:

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{4\gamma g/f - 4\beta}).}$$

Reconstruction Guarantee for Algorithm RGD2 contd.

Edit distance guarantee under RGD2

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{6\gamma g/f - 4\beta - 1}).}$$

Reconstruction Guarantee for Algorithm RGD2 contd.

Edit distance guarantee under RGD2

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{6\gamma g/f - 4\beta - 1}).}$$

Corollary: Consistency Under RGD2

The algorithm RGD2 consistently recovers the minimal representation

$$\boxed{c^{\frac{6Rg}{f}}\rho^4 = o(n), \quad c^{\frac{R}{2}}\rho = \omega(1).}$$

Reconstruction Guarantee for Algorithm RGD2 contd.

Edit distance guarantee under RGD2

$$\boxed{\Delta(\widehat{G}_n, \widetilde{G}_n; V_n) = \tilde{O}(n^{6\gamma g/f - 4\beta - 1}).}$$

Corollary: Consistency Under RGD2

The algorithm RGD2 consistently recovers the minimal representation

$$\boxed{c^{\frac{6Rg}{f}} \rho^4 = o(n), \quad c^{\frac{R}{2}} \rho = \omega(1).}$$

- When $f = g$ (homogeneous edge lengths), $n^{0.875}$ nodes suffice for consistent reconstruction

Efficient discovery using few participants

Outline

Lower Bound on Topology Discovery

Lower Bound on Edit Distance for Random Graphs

Almost every random graph $G_n \sim \mathcal{G}(n, c/n)$ has an edit distance at least $(0.5c - 1)n$ from any given graph F_n.

Lower Bound on Topology Discovery

Lower Bound on Edit Distance for Random Graphs

Almost every random graph $G_n \sim \mathcal{G}(n, c/n)$ has an edit distance at least $(0.5c - 1)n$ from any given graph F_n.

Theorem: Lower Bound for Graph Reconstruction

For $G_n \sim \mathcal{G}(n, c/n)$, any set of participants V_n and any graph estimator $\widehat{G}_n$, the edit distance $\Delta(\widehat{G}_n, G_n; V)$ satisfies

$$\mathbb{P}[\Delta(\widehat{G}_n, G_n; V) > \delta n] \to 1, \text{ when } |V|^2 < Mn(0.5c - \delta - 1)\frac{\log n}{\log \log n},$$

for a small enough constant $M > 0$ and any $\delta < (0.5c - 1)$.

Information-theoretic Covering Argument

Outline

Conclusion

Summary

- Considered network tomography with few participants
- Efficient reconstruction guarantees for random graph models
- Information-theoretic lower bound on graph reconstruction
- Infeasibility of topology discovery for general graphs

Outlook

- Other random graph models (with clustering)
- Other sampling techniques (non-uniform, adaptive)
- Other measurements (e.g., random walk measurements, Ising models)

---

http://newport.eecs.uci.edu/anandkumar