Detection and Reconstruction of a Random Hypergraph from its Noisy Graph Projection

Zhangsong Li

School of Mathemaitcal Sciences, Peking University

joint with

Shuyang Gong (PKU)

Qiheng Xu (Chicago)

Higher order interactions

- Networks or graphs focus on pairwise interactions.

Higher order interactions

- Networks or graphs focus on pairwise interactions.
- Not sufficient to describe the nature of complex interactions:
  - Social networks: triadic and larger groups;
  - Scientific co-authorship;
  - Interactions between neurons in brain networks ...

Higher order interactions

- Networks or graphs focus on pairwise interactions.
- Not sufficient to describe the nature of complex interactions:
  - Social networks: triadic and larger groups;
  - Scientific co-authorship;
  - Interactions between neurons in brain networks ...
- Hypergraphs are natural generalization of graphs that are able to capture higher-order interactions in real networks.
  - A hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is defined as a set of nodes $\mathcal{V} \neq \emptyset$ and a set of hyperedges $\mathcal{E}$.
  - Each hyperedge is a non-empty collection of m distinct nodes $(2 \leq m \leq |\mathcal{V}|)$.
  - We say a hypergraph is d-uniform if $|E| = d$ for all $E \in \mathcal{E}$.

Motivating example

- A general problem: reconstruct higher-order interactions from (related) pairwise interactions.

Motivating example

- A general problem: reconstruct higher-order interactions from (related) pairwise interactions.
- Example: a large team collaborating on several different projects;
  - The ground higher-order interaction (hypergraph) is defined by project membership;
  - [Klimt-Yang '04]: suppose we only observe email exchanges between pairs of members $\Longrightarrow$ simplified to a graph where an edge connects any two individuals who communicated.

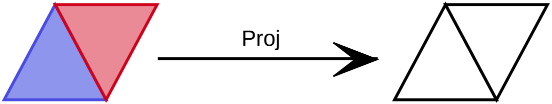
Motivating example

- A general problem: reconstruct higher-order interactions from (related) pairwise interactions.
- Example: a large team collaborating on several different projects;
  - The ground higher-order interaction (hypergraph) is defined by project membership;
  - [Klimt-Yang '04]: suppose we only observe email exchanges between pairs of members $\Longrightarrow$ simplified to a graph where an edge connects any two individuals who communicated.
- **Question:** what can we learn about the ground hypergraph after observing the graph of email exchange?

Graph projection

Suppose $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is a d-uniform hypergraph.

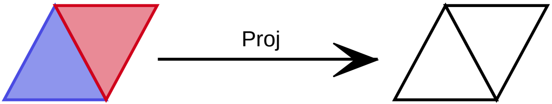
- $\Psi = \text{Proj}(\mathcal{H})$ (projected graph): same vertices and an edge between any two vertices in the same hyperedge.



Graph projection

Suppose $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is a d-uniform hypergraph.

- $\Psi = \text{Proj}(\mathcal{H})$ (projected graph): same vertices and an edge between any two vertices in the same hyperedge.

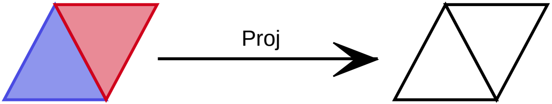


- **Stochastic model:** each hyperedge appears in $\mathcal{H}$ independently with probability $r = n^{-(d-1)+\delta+o(1)}$.

Graph projection

Suppose $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is a d-uniform hypergraph.

- $\Psi = \mathrm{Proj}(\mathcal{H})$ (projected graph): same vertices and an edge between any two vertices in the same hyperedge.



- **Stochastic model:** each hyperedge appears in $\mathcal{H}$ independently with probability $r = n^{-(d-1)+\delta+o(1)}$.
- **Noisy perturbation**: Observation is a graph G that edges in Ψ are kept with probability p, and edges in Ψ^c are added with probability q.

Inference tasks: detection and recovery

The prediction algorithm $\mathcal{A}$ takes the noisy projection G as an input.
Three natural tasks:

- **Detection**: distinguish the graph G from Erdős-Rényi graph $\mathcal{G}(n, q)$.
- **Partial recovery:** recover a positive fraction of the hyperedges in $\mathcal{H}$.
- **Almost-exact recovery:** recover $1 - o(1)$ fraction of the hyperedges in $\mathcal{H}$.

Loss function:

$$\frac{\mathbb{E}[|\mathcal{A} \triangle \mathcal{E}(\mathcal{H})|]}{\mathbb{E}[|\mathcal{E}(\mathcal{H})|]} = \frac{\mathbb{E}[|\mathcal{A} \triangle \mathcal{E}(\mathcal{H})|]}{r\binom{n}{d}}.$$

The noiseless case

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.
In [Bresler-Guo-Polyanskiy '24, Bresler-Guo-Polyanskiy-Yao '25], the noiseless case $p = 1, q = 0$ is settled.

Theorem (Bresler-Guo-Polyanskiy-Yao '25)

Suppose that d is a constant.

(1) *When $\delta > \frac{d-1}{d+1}$, partial and almost-exact recovery are information theoretically impossible;*

(2) *When $\delta < \frac{d-1}{d+1}$, almost-exact recovery is possible in poly-time.*

- Asked what happens in the general noisy case.

Our main result

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

Theorem (Gong-L.-Xu '26)

Suppose that d is a constant.

(1) *When $\beta > 2(\delta+\alpha)-1$, detection is information theoretically impossible; when $\beta < 2(\delta+\alpha)-1$, detection is possible in poly-time.*

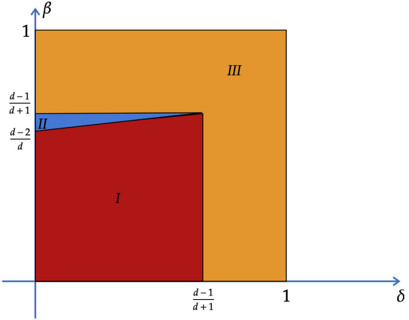
(2) *When either $p = o(1)$, $\delta > \frac{d-1}{d+1}$ or $\beta > \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$, partial recovery is information theoretically impossible; when $p = \Omega(1)$, $\delta < \frac{d-1}{d+1}$ and $\beta < \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$, partial recovery is possible in poly-time.*

(3) *Same results hold for almost-exact recovery with $p = o(1)/p = \Omega(1)$ replaced by $p = 1-\Omega(1)/p = 1-o(1)$.*

Detection-recovery gap

- Even when $p = 1$, the threshold for detection $\beta = 2\delta + 1$ and recovery $\beta < \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$ do not match! $\Longrightarrow$ a Detection-recovery gap occurs.

- Even when $p = 1$, the threshold for detection $\beta = 2\delta + 1$ and recovery $\beta < \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$ do not match! $\implies$ a Detection-recovery gap occurs.
- Cannot simply closing this gap by choosing a better null distribution $\mathcal{G}(n, \widetilde{q})$ to match the edge-density. Phase diagram: (III: recovery possible; II: detection with $\mathcal{G}(n, \widetilde{q})$ possible, recovery impossible; I: detection with $\mathcal{G}(n, q)$ possible, detection with $\mathcal{G}(n, \widetilde{q})$ impossible)



Proof ideas: detection

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

Proof ideas: detection

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \mathrm{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

- Upper bound when $\beta > 2(\alpha+\delta)-1$: a simple edge counting statistic $f(G) = \sum_{(i,j)} \mathbf{1}_{(i,j)\in E(G)}$ distinguishes $\mathbb{P}$ and $\mathbb{Q}$.

Proof ideas: detection

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

- Upper bound when $\beta > 2(\alpha+\delta)-1$: a simple edge counting statistic $f(G) = \sum_{(i,j)} \mathbf{1}_{(i,j)\in E(G)}$ distinguishes $\mathbb{P}$ and $\mathbb{Q}$.
- Lower bound when $\beta < 2(\alpha+\delta)-1$: bounding the chi-square divergence $\chi^2(\mathbb{P};\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(G))^2]$.

Proof ideas: detection

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

- Upper bound when $\beta > 2(\alpha+\delta)-1$: a simple edge counting statistic $f(G) = \sum_{(i,j)} \mathbf{1}_{(i,j)\in E(G)}$ distinguishes $\mathbb{P}$ and $\mathbb{Q}$.
- Lower bound when $\beta < 2(\alpha+\delta)-1$: bounding the chi-square divergence $\chi^2(\mathbb{P};\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(G))^2]$.
  - Reduce to bounding the exponential of "replica overlap": $\mathbb{E}[\exp(|E(\text{Proj}(\mathcal{H})) \cap E(\text{Proj}(\mathcal{H}'))|)]$, where $\mathcal{H}, \mathcal{H}'$ are independent random hypergraphs.

Proof ideas: detection

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

- Upper bound when $\beta > 2(\alpha+\delta)-1$: a simple edge counting statistic $f(G) = \sum_{(i,j)} \mathbf{1}_{(i,j)\in E(G)}$ distinguishes $\mathbb{P}$ and $\mathbb{Q}$.
- Lower bound when $\beta < 2(\alpha+\delta)-1$: bounding the chi-square divergence $\chi^2(\mathbb{P};\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(G))^2]$.
    - Reduce to bounding the exponential of "replica overlap": $\mathbb{E}[\exp(|E(\text{Proj}(\mathcal{H})) \cap E(\text{Proj}(\mathcal{H}'))|)]$, where $\mathcal{H}, \mathcal{H}'$ are independent random hypergraphs.
    - Key idea: use Talagrand's concentration inequality to show that $|E(\text{Proj}(\mathcal{H})) \cap E(\text{Proj}(\mathcal{H}'))|$ tightly concentrate around its mean.

Proof ideas: recovery

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

Proof ideas: recovery

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

- Upper bound (motivated by [Bresler-Guo-Polyanskiy-Yao '25]): consider the clique statistic $\widehat{\mathcal{H}} = \{E \in \binom{[n]}{d} : \text{Proj}(\mathcal{H}) \subset G\}$.

Proof ideas: recovery

$r = n^{-(d-1)+\delta+o(1)}$: probability that $E \in \mathcal{H}$;
$p = n^{-1+\alpha+o(1)}$: the probability of an edge in $\Psi = \text{proj}(\mathcal{H})$ is kept;
$q = n^{-1+\beta+o(1)}$: the probability that an edge in Ψ^c is added.

- Upper bound (motivated by [Bresler-Guo-Polyanskiy-Yao '25]): consider the clique statistic $\widehat{\mathcal{H}} = \{E \in \binom{[n]}{d} : \text{Proj}(\mathcal{H}) \subset G\}$.
- Lower bound: note that the problem becomes easier when p increase and q decrease;
  - For partial recovery, suffices to show the following two cases: (1) $p = o(1), q = 0$; (2) $p = 1, \beta < \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.
  - For almost-exact recovery, suffices to show the following two cases: (1) $p = 1 - \Omega(1), q = 0$; (2) $p = 1, \beta < \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.

Proof ideas: recovery

- **Case 1:** $p = o(1), q = 0$.
- Idea: use a modified version of Fano's inequality.

Lemma (Banarjee-Hedge-Massoulié '2012)

Let $Z \longrightarrow X \longrightarrow \widehat{Z}$ constitutes a Markov chain, and let d to be a loss function. For any z, we further define $B_d(z) = \{z' : d(z, z') \leq d\}$ and we let $M_d = \max_z\{|B_d(z)|\}$. Then we have

$$\mathbb{P}\big(d(Z, \widehat{Z}) \geq d\big) \geq 1 - \frac{I(Z; X) + 1}{\log(M/M_d)}$$

where $I(Z; X)$ is the mutual information between Z and X.

- Apply this to the Markov chain $\mathcal{H} \to G \to \widehat{\mathcal{H}}$ yields the desired lower bound.

Proof ideas: recovery

- **Case 2:** $p = 1, \delta > \frac{d-1}{d+1}$ and $q = n^{-1+\beta+o(1)}$ where $\beta > \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.

Proof ideas: recovery

- **Case 2:** $p = 1, \delta > \frac{d-1}{d+1}$ and $q = n^{-1+\beta+o(1)}$ where $\beta > \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.
- Idea [Ding-Wu-Xu-Yang '23]: suffices to prove that if we sample $\mathcal{H}'$ from the posterior distribution, then $|\mathcal{H}' \cap \mathcal{H}| = o(1)\binom{n}{d}$ w.h.p.

Proof ideas: recovery

- **Case 2:** $p = 1, \delta > \frac{d-1}{d+1}$ and $q = n^{-1+\beta+o(1)}$ where $\beta > \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.
- Idea [Ding-Wu-Xu-Yang '23]: suffices to prove that if we sample $\mathcal{H}'$ from the posterior distribution, then $|\mathcal{H}' \cap \mathcal{H}| = o(1)\binom{n}{d}$ w.h.p.
- Boils down to bounding the replica representation

$$\mathbb{E}_{\mathcal{H},\mathcal{H}' \overset{iid}{\sim} \mu} \left\{ \mathbb{E}_{A\sim\mathbb{Q}} \left[L_{\mathcal{H}}(A) L_{\mathcal{H}'}(A) \right] \mathbf{1}_{|\mathcal{H}' \cap \mathcal{H}| \geq \Omega(1)\binom{n}{d}} \right\}.$$

Proof ideas: recovery

- **Case 2:** $p = 1, \delta > \frac{d-1}{d+1}$ and $q = n^{-1+\beta+o(1)}$ where $\beta > \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.
- Idea [Ding-Wu-Xu-Yang '23]: suffices to prove that if we sample $\mathcal{H}'$ from the posterior distribution, then $|\mathcal{H}' \cap \mathcal{H}| = o(1)\binom{n}{d}$ w.h.p.
- Boils down to bounding the replica representation

$$\mathbb{E}_{\mathcal{H},\mathcal{H}' \overset{iid}{\sim} \mu} \left\{ \mathbb{E}_{A\sim\mathbb{Q}} \left[L_{\mathcal{H}}(A) L_{\mathcal{H}'}(A) \right] \mathbf{1}_{|\mathcal{H}'\cap\mathcal{H}| \geq \Omega(1)\binom{n}{d}} \right\}.$$

- $(*)$ reduce to bounding the exponential of "replica overlap": $\mathbb{E}[\exp(|E(\mathrm{Proj}(\mathcal{H})) \cap E(\mathrm{Proj}(\mathcal{H}'))|)]$, where $\mathcal{H}, \mathcal{H}'$ are independent random hypergraphs.

Proof ideas: recovery

- **Case 2:** $p = 1, \delta > \frac{d-1}{d+1}$ and $q = n^{-1+\beta+o(1)}$ where $\beta > \frac{d-2}{d} + \frac{2\delta}{d(d-2)}$.
- Idea [Ding-Wu-Xu-Yang '23]: suffices to prove that if we sample $\mathcal{H}'$ from the posterior distribution, then $|\mathcal{H}' \cap \mathcal{H}| = o(1)\binom{n}{d}$ w.h.p.
- Boils down to bounding the replica representation

$$\mathbb{E}_{\mathcal{H},\mathcal{H}' \overset{iid}{\sim} \mu} \left\{ \mathbb{E}_{A\sim\mathbb{Q}} \left[L_{\mathcal{H}}(A) L_{\mathcal{H}'}(A) \right] \mathbf{1}_{|\mathcal{H}' \cap \mathcal{H}| \geq \Omega(1)\binom{n}{d}} \right\}.$$

- $(*)$ reduce to bounding the exponential of "replica overlap": $\mathbb{E}[\exp(|E(\mathrm{Proj}(\mathcal{H})) \cap E(\mathrm{Proj}(\mathcal{H}'))|)]$, where $\mathcal{H}, \mathcal{H}'$ are independent random hypergraphs.
- Key idea: use Talagrand's concentration inequality to show that $|E(\mathrm{Proj}(\mathcal{H})) \cap E(\mathrm{Proj}(\mathcal{H}'))|$ tightly concentrate around its mean.

Conclusion and future perspectives

Take-home message:

- We determine the threshold for detecting and reconstructing a random d-uniform hypergraph from a noisy perturbation of its projected graph, provided that d is a constant.

Conclusion and future perspectives

Take-home message:

- We determine the threshold for detecting and reconstructing a random d-uniform hypergraph from a noisy perturbation of its projected graph, provided that d is a constant.

Open problems:

- Growing order d (especially $d = n^{\Omega(1)}$);
- Random hypergraphs with additional structures (e.g., hypergraph stochastic block model, ...).

Conclusion and future perspectives

Take-home message:

- We determine the threshold for detecting and reconstructing a random d-uniform hypergraph from a noisy perturbation of its projected graph, provided that d is a constant.

Open problems:

- Growing order d (especially $d = n^{\Omega(1)}$);
- Random hypergraphs with additional structures (e.g., hypergraph stochastic block model, ...).

Reference:

Shuyang Gong, Zhangsong Li, and Qiheng Xu.
Detection and Reconstruction of a Random Hypergraph from its Noisy Graph Projection. Preprint, arXiv:2506.17527.