Shotgun Assembly of Labelled Graphs

Charles Bordenave[3], Uri Feige[3], **Elchanan Mossel**[1,2,3], Nathan Ross[1], Nike Sun[2]

[1]Shotgun assembly of Labelled Graphs (arxiv.org/abs/1504.07682)

[2]Shotgun Assembly of Random Regular Graphs, (arxiv.org/abs/1512.08473)

[3]Shotgun Assembly of Random Jigsaw Puzzles, in progress.

Simons Institute, Berkeley

Graph Shotgun Problem

- Can one reconstruct a graph from collection of subgraphs?
- Reconstruction Conjecture (Kelley, Harary 50s): Any two graphs on 3 or more vertices that have the same multi-set of vertex-deleted subgraphs are isomorphic.

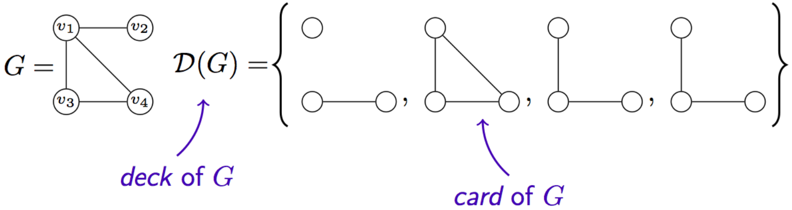


Figure: From Topology and Combinatorics Blog by Max F. Pitz

Graph Shotgun Problem

- Can one reconstruct a graph from collection of subgraphs?
- Reconstruction Conjecture (Kelley, Harary 50s): Any two graphs on 3 or more vertices that have the same multi-set of vertex-deleted subgraphs are isomorphic.
- Mossel-Ross-15: What if Graphs are Random or have random labels? (*easier*)
- And given only local neighborhoods of each vertex (*harder*)?

DNA Shotgun Sequencing

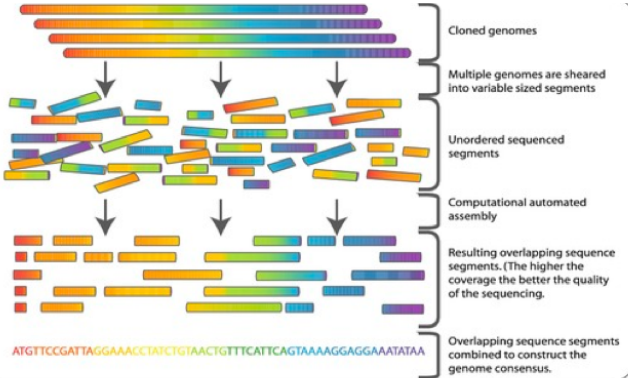


Figure: From "Whole genome shotgun sequencing versus Hierarchical shotgun sequencing" by Commins, Toft, and Fares (2009).

Q1: Deterministic

- Sequence of letters (A, C, G, T or other) of length N.
- All "reads" of length r are given.

Example: $N = 14$, $r = 3$:

$$ATGGGCACTGAGCC$$

Reads:

$$\{ATG, TGG, GGG, GGC, GCA, CAC, ACT, CTG, TGA, GAG, AGC, GCC\}$$

Combinatorial Question:

When does this multi-set uniquely determine the sequence?

Q1: Deterministic

Ans (Ukkonen-Pevzner):

Identifiability is possible **if and only** if none of the following blocking patterns appear:

- Rotation:

$$x\alpha y\beta x \iff y\beta x\alpha y$$

- Triple repeat:

$$\cdots x\alpha x\beta x\cdots \iff \cdots x\beta x\alpha x\cdots$$

- Interleaved repeat:

$$\cdots x\alpha y\cdots x\beta y\cdots \iff \cdots x\beta y\cdots x\alpha y\cdots$$

[x, y are $(r-1)$-tuples and α, β are non-equal strings]

Q1: Deterministic

Proof is based on creating a de Bruijn graph:

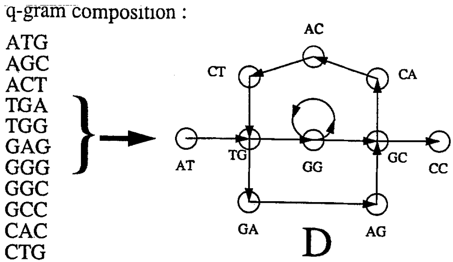


Figure: From "DNA Physical Mapping and Alternating Eulerian Cycles in Colored Graphs" by Pevzner (1996).

$$ATGGGCACTGAGCC$$

Q1: Deterministic

Proof is based on creating a de Bruijn graph:

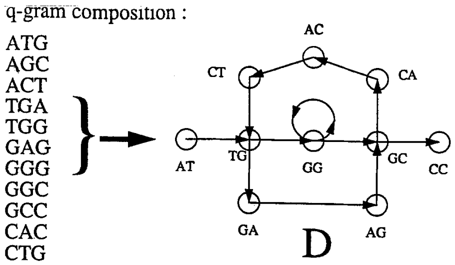


Figure: From "DNA Physical Mapping and Alternating Eulerian Cycles in Colored Graphs" by Pevzner (1996).

Identifiability is possible if and only if a unique Eulerian path (though not circuit).

Setup Q2: Randomized

Random sequence, entries independent and uniform on q letters.

- What is the probability of identifiability?
- Criteria on growth of $r = r_N$ as $N \to \infty$ such that the chance sequence is identifiable tends to zero or one?

Ukkonen-Pevzner useful – understand the probability of the appearance of the blocking patterns.

- If $r/\log(N) > 2/\log(q)$ eventually, then probability of identifiability tends to one.
- If $r/\log(N) < 2/\log(q)$ eventually, then probability of identifiability tends to zero.
- Dyer-Frieze-Suen-94,....
- Still active area of research: e.g.: reads with errors, e.g: Ganguly-M-Racz-16.

What about other Graphs??

Paninski et al. (2013) : How to reconstruct neural network from subnetworks?

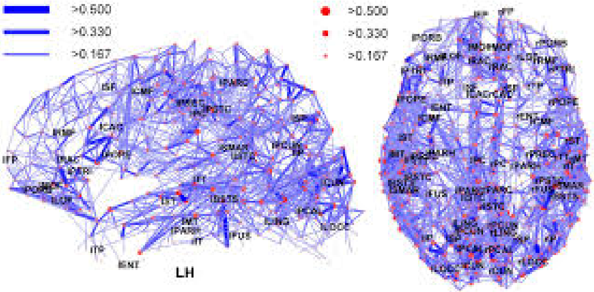


Figure: wiki commons

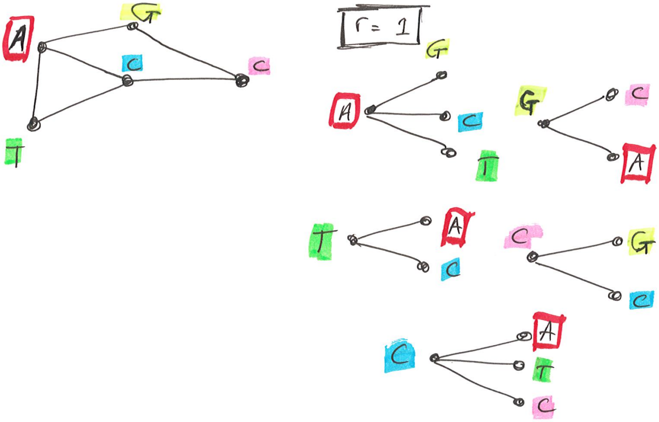
Random Puzzle Problem

Figure: wiki commons

Math Question: For an $n \times n$ puzzle with q types of random jigs, how large should $q(n)$ be so that the puzzle can be assembled uniquely??

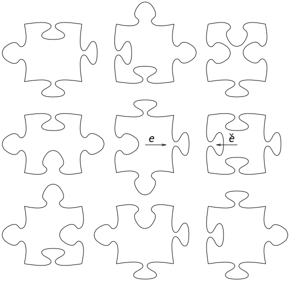
A general setup

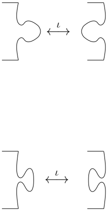
1. $\mathcal{G}$ is a (fixed or random) graph,
2. Possibly with random labeling of the vertices,
3. For each vertex v, given a rooted neighborhood $\mathcal{N}_r(v)$ of "radius" r.



Random jigsaw Puzzle

- Puzzle $= [n] \times [n]$ grid with uniform q-coloring of the edges of the grid.
- Piece = vertex along with 4 adjacent colored half edges.
- Given: n^2 pieces.
- Goal: Recover the puzzle.
- Assume pieces at the edges also have 4 colors (harder).
- For presentation purposes: colored edges vs.
- Real Puzzle: colored half edges and a compatibility involution.





The unique Assembly Question

- A *feasible assembly* is a permutation of the pieces such that adjacent two half-edges have the same color.
- A puzzle has unique vertex assembly (UVA) if (up to rotations) it has only one feasible assembly.
- A puzzle has unique edge assembly (UEA) if for every feasible assembly, every edge has the same color as in the planted solution (up to rotations).
- **Question:** How large should q be to ensure unique edge/vertex assembly with high probability ($\to 1$ as $n \to \infty$) ?

Bounds on puzzle assembly

From M-Ross:

- $q << n \implies P(UVA) \to 0$.

Bounds on puzzle assembly

From M-Ross:

- $q << n \implies P(UVA) \to 0$.
- $q << n^{2/3} \implies P(UEA) \to 0$.

Bounds on puzzle assembly

From M-Ross:

- $q << n \implies P(UVA) \to 0$.
- $q << n^{2/3} \implies P(UEA) \to 0$.
- $q >> n^2 \implies P(UVA) \to 1$.

Bounds on puzzle assembly

From M-Ross:

- $q << n \implies P(UVA) \to 0.$
- $q << n^{2/3} \implies P(UEA) \to 0.$
- $q >> n^2 \implies P(UVA) \to 1.$
- Intuition: use unique colors.

Bounds on puzzle assembly

From M-Ross:

- $q << n \implies P(UVA) \to 0.$
- $q << n^{2/3} \implies P(UEA) \to 0.$
- $q >> n^2 \implies P(UVA) \to 1.$
- Intuition: use unique colors.

Theorem (Bordenave-Feige-M)

For all $\varepsilon > 0$, If $q \geq n^{1+\varepsilon}$ then $P(UVA) \to 1$.

- Open Problem 1: Zoom in on threshold?
- Open Problem 2: Threshold for UEA.

Assembly algorithm

We use a simple assembly algorithm:

- A feasible k-neighborhood of piece v is map f from $[-k, k]^2 \to$ pieces such that $f(0) = v$ and if $x \sim y \in [-k, k]^2$ then the corresponding half-edges in $f(x)$ and $f(y)$ have the same color.
- Algorithm: find all feasible k-neighborhoods for each vertex v.
- Declare piece u to be a neighbor of v if it is its neighbor of v in each k-neighborhood.
- We take $k = O(1/\varepsilon)$.
- How to analyze?

Analysis 1

- Note: impossible to hope to recover k-neighborhood exactly, e.g - corners are often wrong.
- Fix $f : [-k, k]^2 \to [n]^2$ with $f(0) = v$. What is the probability that f is feasible?
  - If $f(x) = v + x$ then probability 1.
  - If f is *random* then probability $q^{-8k^2(1+o(1))}$.

Analysis 2

- Define a *tile* of f to be a connected component of $f([-k,k]^2)$.
- Let $v \in T_0, T_1, \ldots, T_r$ be the tiles of f.

Analysis 2

- Define a *tile* of f to be a connected component of $f([-k,k]^2)$.
- Let $v \in T_0, T_1, \ldots, T_r$ be the tiles of f.
- Then:

$$P[f \text{ feasible }] = q^{-\gamma}, \quad \gamma = \frac{1}{2}(\sum |\partial T_i| - 8k)$$

Analysis 2

- Define a *tile* of f to be a connected component of $f([-k,k]^2)$.
- Let $v \in T_0, T_1, \ldots, T_r$ be the tiles of f.
- Then:

$$P[f \text{ feasible }] = q^{-\gamma}, \quad \gamma = \frac{1}{2}(\sum |\partial T_i| - 8k)$$

- Isoperimetric lemma: If f separates v from its neighbors then:

$$n^2 n^{2r} q^{-\gamma} = n^2 n^{2r} n^{-\gamma(1+\varepsilon)} << 1$$

- E.g: many small tiles - each contributed at least 2 to γ.

Analysis 2

- Define a *tile* of f to be a connected component of $f([-k,k]^2)$.
- Let $v \in T_0, T_1, \ldots, T_r$ be the tiles of f.
- Then:

$$P[f \text{ feasible }] = q^{-\gamma}, \quad \gamma = \frac{1}{2}\left(\sum |\partial T_i| - 8k\right)$$

- Isoperimetric lemma: If f separates v from its neighbors then:

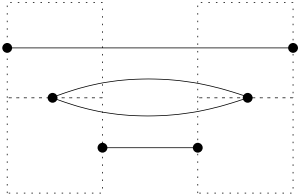
$$n^2 n^{2r} q^{-\gamma} = n^2 n^{2r} n^{-\gamma(1+\varepsilon)} << 1$$

- E.g: many small tiles - each contributed at least 2 to γ.
- Isoperimetric lemma + union bound $\implies$ proof.

Cheat and Punishment

Sadly boundary events are *not* independent.

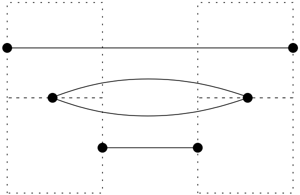
$(3,2)$	$(1,2)$
$(1,1)$	$(3,1)$



Cheat and Punishment

Sadly boundary events are *not* independent.

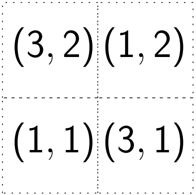
$(3,2)$	$(1,2)$
$(1,1)$	$(3,1)$

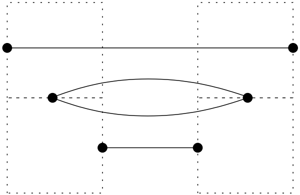


- Graph theoretic definition of $\gamma(f)$, the number of "unique constraints".

Cheat and Punishment

Sadly boundary events are *not* independent.



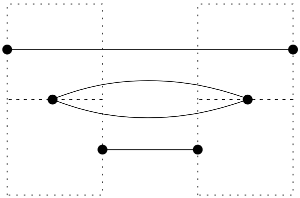


- Graph theoretic definition of $\gamma(f)$, the number of "unique constraints".
- Isoperimetric lemma to lower bound $\gamma(f)$.

Cheat and Punishment

Sadly boundary events are *not* independent.

$(3,2)$	$(1,2)$
$(1,1)$	$(3,1)$



- Graph theoretic definition of $\gamma(f)$, the number of "unique constraints".
- Isoperimetric lemma to lower bound $\gamma(f)$.
- Interesting: lower bound uses both $\sum |\partial T_i|$ and $\sum |\partial f(T_i)|$

Some Random Graph Examples

- We now look at some random graph examples.

Some Random Graph Examples

- We now look at some random graph examples.
- "Guiding principle" (M-Ross): Threshold for assembly

$$r = min(k : u \neq v \implies B_k(u) \not\sim B_k(v))(+1)$$

Some Random Graph Examples

- We now look at some random graph examples.
- "Guiding principle" (M-Ross): Threshold for assembly

$$r = min(k : u \neq v \implies B_k(u) \not\sim B_k(v))(+1)$$

- Easy direction: "name" vertex v by $B_k(v)$.

Some Random Graph Examples

- We now look at some random graph examples.
- "Guiding principle" (M-Ross): Threshold for assembly

$$r = min(k : u \neq v \implies B_k(u) \not\sim B_k(v))(+1)$$

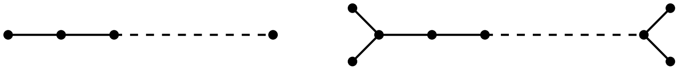
- Easy direction: "name" vertex v by $B_k(v)$.
- Other direction requires more work per-example.

Example: Sparse Erdős-Rényi random graph

Each edge present with probability $p_N = \lambda/N$ independently so Average degree is λ.

Example: Sparse Erdős-Rényi random graph

Each edge present with probability $p_N = \lambda/N$ independently so Average degree is λ.
Blocking configuration for r-neighborhoods (line graph has is of length $r+1$)



Since has same r-neighborhoods as

- if $r < \log N[\lambda - \log(\lambda)]^{-1}$, then the probability of identifiability tends to zero.

Example 1a: Sparse Erdős-Rényi random graph

Diameter

- For $\lambda \neq 1$, the diameter of the sparse Erdős-Rényi random graph is of order $\log(N)$ (different constants than that above).
- Corollary (Mossel-Ross-15): If $\lambda \neq 1$ then reconstruction threshold is $r = \Theta(\log N)$.

Example 1a: Sparse Erdős-Rényi random graph

Diameter

- For $\lambda \neq 1$, the diameter of the sparse Erdős-Rényi random graph is of order $\log(N)$ (different constants than that above).
- Corollary (Mossel-Ross-15): If $\lambda \neq 1$ then reconstruction threshold is $r = \Theta(\log N)$.
- Harder/Open: $r = C \log N(1 + o(1))$?

Example 1a: Sparse Erdős-Rényi random graph

Diameter

- For $\lambda \neq 1$, the diameter of the sparse Erdős-Rényi random graph is of order $\log(N)$ (different constants than that above).
- Corollary (Mossel-Ross-15): If $\lambda \neq 1$ then reconstruction threshold is $r = \Theta(\log N)$.
- Harder/Open: $r = C \log N(1 + o(1))$?
- Critical case?

Example 1b: Less sparse Erdős-Rényi random graph

Structure of the Erdős-Rényi graph depends on behavior of $N \times p_N$.

2. The Denser Case

- Assume $\underline{Np_N / \log(N)^2 \to \infty}$.

Example 1b: Less sparse Erdős-Rényi random graph

Structure of the Erdős-Rényi graph depends on behavior of $N \times p_N$.

2. The Denser Case

- Assume $\underline{Np_N/\log(N)^2 \to \infty}$.
- Mossel-Ross-15: If $r = 3$, then the probability of identifiability tends to one.
- multiset of degrees of neighbors of each vertex become unique.
- Allows to give distinct names to vertices.

Example 1b: Less sparse Erdős-Rényi random graph

Structure of the Erdős-Rényi graph depends on behavior of $N \times p_N$.

2. The Denser Case

- Assume $\underline{Np_N / \log(N)^2 \to \infty}$.
- Mossel-Ross-15: If $r = 3$, then the probability of identifiability tends to one.
- multiset of degrees of neighbors of each vertex become unique.
- Allows to give distinct names to vertices.
- Open: when is $r = 2$ enough?
- Distributed computing perspective: unique i.d's from local information.

Example 2: Random Regular Graphs

Theorem (M+Sun)

The threshold for assembly of random d regular graphs is

$$r = \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1).$$

Happy and Sad neighborhoods

Why?

- (Almost) all $0.5 \log_{d-1}(n)$ neighborhoods are happy trees.

Happy and Sad neighborhoods

Why?

- (Almost) all $0.5\log_{d-1}(n)$ neighborhoods are happy trees.
- Each $0.5(1+\epsilon)\log_{d-1}(n)$ neighborhoods is unhappy due a unique cycle structure.

The Upper Bound

Theorem (Bollobas 82)

For all $\varepsilon > 0$ *if* $r \geq (0.5 + \varepsilon)\log_{d-1} n$ *then for all* $u \neq v$ *it holds that* $(d_1(v), \ldots, d_r(v)) \neq (d_1(u), \ldots, d_r(u))$ *where* $d_i(v)$ *are the number of nodes at distance* i *from* v.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.
- Compact: only *polylog*(n) cycles.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.
- Compact: only $polylog(n)$ cycles.
- Show that each fixed cycle structure is obtained with probability $\leq n^{-100}$.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.
- Compact: only $polylog(n)$ cycles.
- Show that each fixed cycle structure is obtained with probability $\leq n^{-100}$.
- Cycle structures not independent.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.
- Compact: only $polylog(n)$ cycles.
- Show that each fixed cycle structure is obtained with probability $\leq n^{-100}$.
- Cycle structures not independent.
- Fix No. 1: For each v, for all $u \sim v$, look at cycle structure around u avoiding (v, u).

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.
- Compact: only $polylog(n)$ cycles.
- Show that each fixed cycle structure is obtained with probability $\leq n^{-100}$.
- Cycle structures not independent.
- Fix No. 1: For each v, for all $u \sim v$, look at cycle structure around u avoiding (v, u).
- Still every two cycle structures intersect a little bit.

Theorem (M-Sun)

For all $\varepsilon > 0$ *if* $r \geq \frac{\log n + \log\log n}{2\log(d-1)} + \Theta(1)$ *then for all* $u \neq v$ *it holds that* $B_r(v) \neq B_r(u)$.

Main ideas:

- Encode neighborhood by cycle structure.
- Compact: only $polylog(n)$ cycles.
- Show that each fixed cycle structure is obtained with probability $\leq n^{-100}$.
- Cycle structures not independent.
- Fix No. 1: For each v, for all $u \sim v$, look at cycle structure around u avoiding (v, u).
- Still every two cycle structures intersect a little bit.
- Fix No . 2: Define a metric on cycle structures and study corresponding measure metric space.

The lower bound

Find the following:

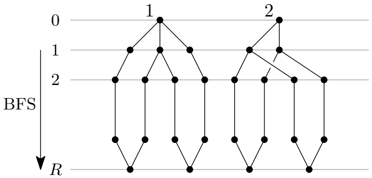


Figure: Two neighborhoods that are hard to distinguish

- Based on second moment argument.

The lower bound

Find the following:

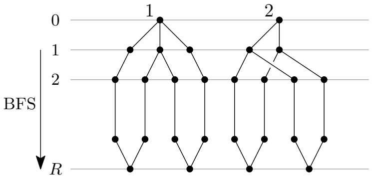


Figure: Two neighborhoods that are hard to distinguish

- Based on second moment argument.
- Need to consider cycle structures of 4 vertices.

The lower bound

Find the following:

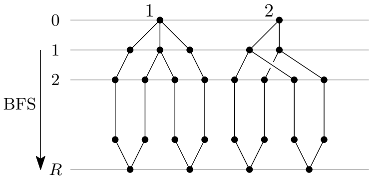


Figure: Two neighborhoods that are hard to distinguish

- Based on second moment argument.
- Need to consider cycle structures of 4 vertices.
- Uses metric-measure space on cycle structure.

Conclusion

- For your favorite generative model - when do we have unique asembly?

Conclusion

- For your favorite generative model - when do we have unique asembly?
- Are there computationally hard regimes? (note graph isomorphism is a module).

Conclusion

- For your favorite generative model - when do we have unique asembly?
- Are there computationally hard regimes? (note graph isomorphism is a module).
- Applications?

Conclusion

- For your favorite generative model - when do we have unique asembly?
- Are there computationally hard regimes? (note graph isomorphism is a module).
- Applications?
- Questions?