Hidden Hamiltonian Cycle Recovery via Linear Programming

Jiaming Xu

The Fuqua School of Business
Duke University

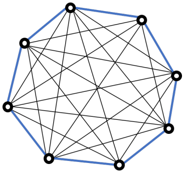
Joint work with
Vivek Bagaria (Stanford), Jian Ding (Penn), David Tse (Stanford)
and Yihong Wu (Yale)

Vilnius, July 5, 2018

Mathematical problem: Hidden Hamiltonian cycle model

- Given a weighted undirected complete graph on n vertices
- Latent: a Hamiltonian cycle C^*
- Edge weight

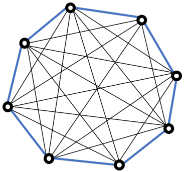
$$W_e \overset{\text{ind.}}{\sim} \begin{cases} P & e \in C^* \\ Q & e \notin C^* \end{cases}$$



Mathematical problem: Hidden Hamiltonian cycle model

- Given a weighted undirected complete graph on n vertices
- Latent: a Hamiltonian cycle C^*
- Edge weight

$$W_e \overset{\text{ind.}}{\sim} \begin{cases} P & e \in C^* \\ Q & e \notin C^* \end{cases}$$

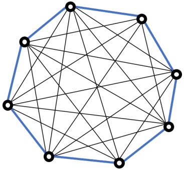


- Goal: observe W, recover C^* with high probability

Mathematical problem: Hidden Hamiltonian cycle model

- Given a weighted undirected complete graph on n vertices
- Latent: a Hamiltonian cycle C^*
- Edge weight

$$W_e \overset{\text{ind.}}{\sim} \begin{cases} P & e \in C^* \\ Q & e \notin C^* \end{cases}$$



- Goal: observe W, recover C^* with high probability

Remarks:

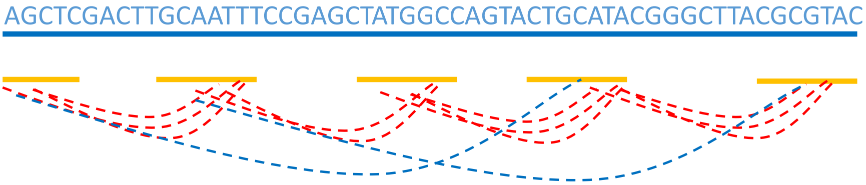
- For this talk, $Q = N(0,1)$ and $P = N(\mu, 1)$, so that

$$W = \mu \cdot \underbrace{\text{adj matrix of } C^*}_{\text{“signal”}} + \text{noise}$$

- Hidden Hamiltonian cycle planted in Erdös-Rényi graph [Broder-Frieze-Shamir '94]

Motivation: Link information in Chicago datasets

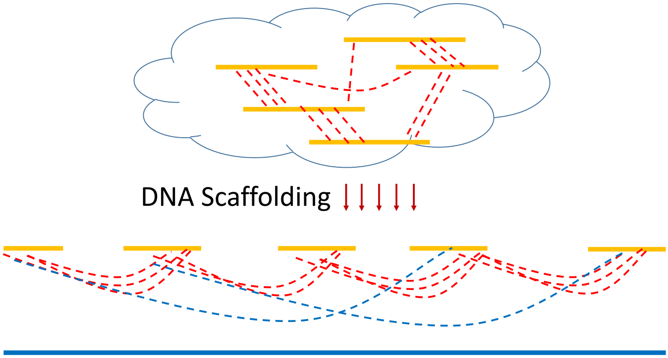
1. Reconstitute chromatin in vitro upon naked DNA
2. Produce cross-links by fixing chromatin with formaldehyde



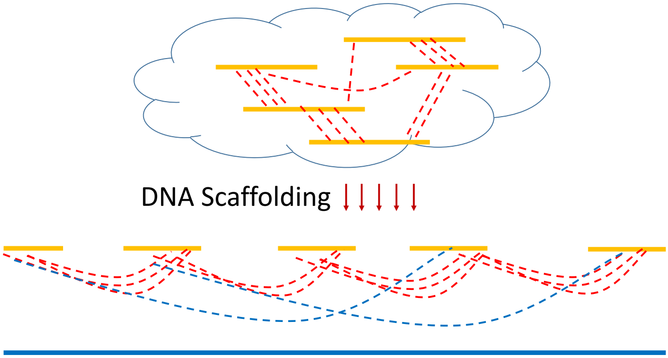
Chicago datasets generate cross-links among contigs [Putnam et al. '16]

On average more cross-links exist between adjacent contigs

Ordering DNA contigs with Chicago cross-links



Ordering DNA contigs with Chicago cross-links



Reduces to traveling salesman problem (TSP)

Find a path (tour) that visits every contig exactly once with the maximum number of cross-links

Key challenges for DNA scaffolding with Chicago data

- Computational: TSP is NP-hard in the worst-case
- Statistical: spurious cross-links between contigs that are far apart

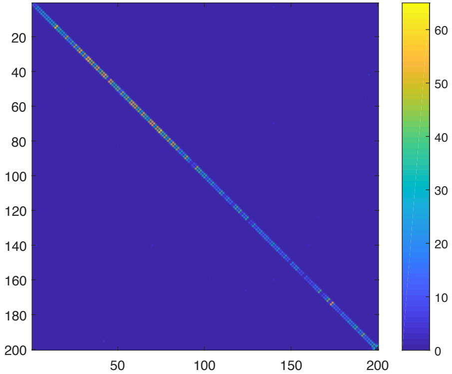
Key challenges for DNA scaffolding with Chicago data

- Computational: TSP is NP-hard in the worst-case
- Statistical: spurious cross-links between contigs that are far apart

Key questions:

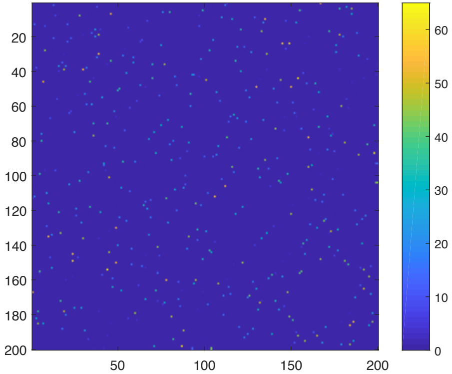
- How to efficiently order hundreds of thousands of contigs?
- How much noise can be tolerated for accurate DNA scaffolding?

Mathematical model for DNA scaffolding



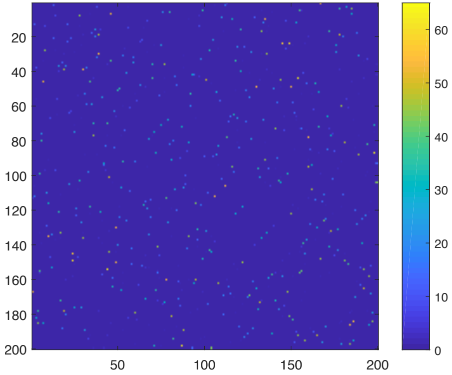
Chicago dataset [Putnam et al. '16]

Mathematical model for DNA scaffolding

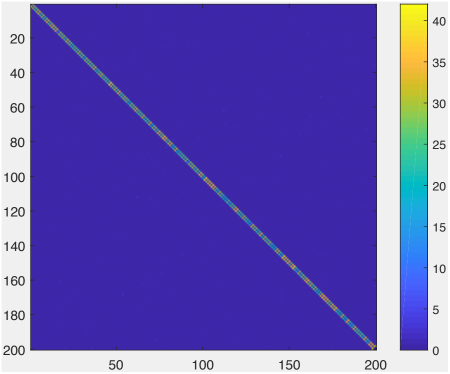


Chicago dataset [Putnam et al. '16]

Mathematical model for DNA scaffolding

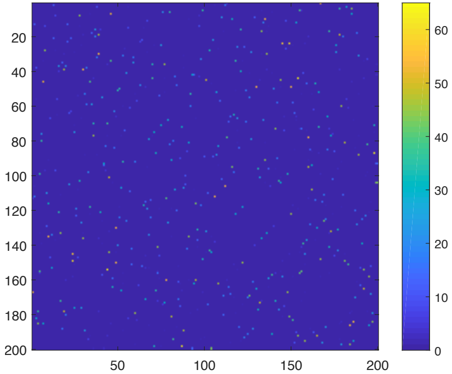


Chicago dataset [Putnam et al. '16]

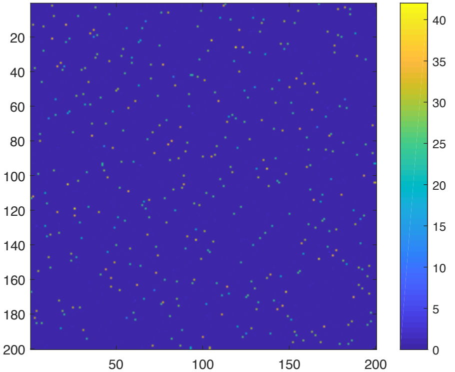


$P = \mathrm{Pois}(\lambda_1)$, $Q = \mathrm{Pois}(\lambda_2)$

Mathematical model for DNA scaffolding



Chicago dataset [Putnam et al. '16]



$P = \mathrm{Pois}(\lambda_1)$, $Q = \mathrm{Pois}(\lambda_2)$

What is known information-theoretically

Maximum likelihood estimator reduces to TSP

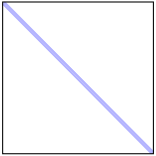
$$\widehat{X}_{\rm TSP} = \arg\max_{X} \ \langle W, X \rangle$$

$$\text{s.t.} \quad X \text{ is the adjacency matrix of some Hamiltonian cycle}$$

Theorem (Sharp threshold)

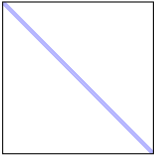
If $\mu^2 < 4 \log n$, exact recovery is information-theoretically impossible
If $\mu^2 > 4 \log n$, MLE succeeds in exact recovery

- Spectral methods fail miserably:
  - $\mu \gg n^{2.5}$ (spectral gap of cycle is too small)

$W = \mu$  $+$ Gaussian noise

What is known algorithmically

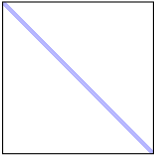
- Spectral methods fail miserably:
  - $\mu \gg n^{2.5}$ (spectral gap of cycle is too small)

$W = \mu$  + Gaussian noise

- Thresholding method: $\mu > \sqrt{8 \log n}$

What is known algorithmically

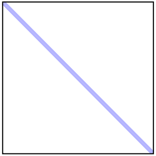
- Spectral methods fail miserably:
  - $\mu \gg n^{2.5}$ (spectral gap of cycle is too small)

$W = \mu$  + Gaussian noise

- Thresholding method: $\mu > \sqrt{8 \log n}$
- Greedy merging [Motahari-Bresler-Tse '13]: $\mu > \sqrt{6 \log n}$

What is known algorithmically

- Spectral methods fail miserably:
  - $\mu \gg n^{2.5}$ (spectral gap of cycle is too small)

$W = \mu$  + Gaussian noise

- Thresholding method: $\mu > \sqrt{8 \log n}$
- Greedy merging [Motahari-Bresler-Tse '13]: $\mu > \sqrt{6 \log n}$
- This talk: **linear programming** achieves sharp threshold

$$\frac{\mu^2}{\log n} > 4: \quad \text{LP succeeds}$$
$$\frac{\mu^2}{\log n} < 4: \quad \text{Everything fails}$$

In general

Threshold determined by Battacharyya distance (a.k.a. Rényi divergence of order $\frac{1}{2}$):

$$B(P,Q) \triangleq -2\log \int \sqrt{\mathrm{d}P\mathrm{d}Q}$$

In general

Threshold determined by Battacharyya distance (a.k.a. Rényi divergence of order $\frac{1}{2}$):

$$B(P,Q) \triangleq -2\log \int \sqrt{\mathrm{d}P\mathrm{d}Q}$$

LP succeeds when

$$\boxed{B(P,Q) - \log n \to +\infty}$$

optimal under mild assumptions

Convex relaxations of TSP

Integer Linear Programming reformulation of TSP

$$
\begin{aligned}
\widehat{X}_{\rm TSP} = \arg\max_{X} \ & \langle W, X \rangle \\
\text{s.t.} \ & \sum_j X_{ij} = 2, \ \forall i \\
& X_{ij} \in \{0, 1\} \\
& \sum_{i \in I, j \notin I} X_{ij} \geq 2, \ \forall \emptyset \neq I \subset [n]
\end{aligned}
$$

Integer Linear Programming reformulation of TSP

$$
\begin{aligned}
\widehat{X}_{\rm TSP} = \arg\max_X \; & \langle W, X \rangle \\
\text{s.t.} \; & \sum_j X_{ij} = 2, \; \forall i \\
& X_{ij} \in \{0, 1\} \\
& \sum_{i \in I, j \notin I} X_{ij} \ge 2, \; \forall \emptyset \neq I \subset [n]
\end{aligned}
$$

- The last constraint: subtour elimination

Subtour LP

$$\widehat{X}_{\rm SUB} = \arg\max_{X} \; \langle W, X \rangle$$

$$\text{s.t.} \quad \sum_{j} X_{ij} = 2, \quad \forall i$$

$$X_{ij} \in [0, 1]$$

$$\sum_{i \in I, j \notin I} X_{ij} \geq 2, \; \forall \emptyset \neq I \subset [n]$$

Subtour LP

$$\widehat{X}_{\mathrm{SUB}} = \arg\max_{X} \langle W, X \rangle$$
$$\text{s.t.} \quad \sum_j X_{ij} = 2, \quad \forall i$$
$$X_{ij} \in [0, 1]$$
$$\sum_{i \in I, j \notin I} X_{ij} \geq 2, \ \forall \emptyset \neq I \subset [n]$$

- Replacing the integrality constraint with box constraint: SUBTOUR LP relaxation [Dantzig-Fulkerson-Johnson '54, Held-Karp '70]
- Exponentially many linear constraints, nevertheless solvable using interior point method

F2F LP

$$\widehat{X}_{\mathrm{F2F}} = \arg\max_{X} \langle W, X \rangle$$
$$\text{s.t.} \quad \sum_j X_{ij} = 2, \quad \forall i$$
$$X_{ij} \in [0, 1]$$

- Further dropping subtour elimination constraints $\Longrightarrow$ Fractional 2-factor (F2F) LP

F2F LP

$$\widehat{X}_{\mathrm{F2F}} = \arg\max_{X} \; \langle W, X \rangle$$
$$\text{s.t.} \quad \sum_j X_{ij} = 2, \quad \forall i$$
$$X_{ij} \in [0,1]$$

- Further dropping subtour elimination constraints $\Longrightarrow$ Fractional 2-factor (F2F) LP
- Extensively studied in worst case [Boyd-Carr '99, Schalekamp-Williamson-van Zuylen '14]

F2F LP

$$\widehat{X}_{\mathrm{F2F}} = \arg\max_{X} \ \langle W, X \rangle$$
$$\text{s.t.} \quad \sum_j X_{ij} = 2, \quad \forall i$$
$$X_{ij} \in [0,1]$$

- Further dropping subtour elimination constraints $\Longrightarrow$ Fractional 2-factor (F2F) LP
- Extensively studied in worst case [Boyd-Carr '99, Schalekamp-Williamson-van Zuylen '14]
- How it performs in our random instance?

Optimality of Fractional 2-Factor LP

Theorem (Bagaria-Ding-Tse-Wu-X. '18)

If $\mu^2 - 4\log n \to \infty$, *then* $\widehat{X}_{\mathrm{F2F}} = X^*$ *with high probability.*

Optimality of Fractional 2-Factor LP

Theorem (Bagaria-Ding-Tse-Wu-X. '18)

If $\mu^2 - 4\log n \to \infty$, *then* $\widehat{X}_{\mathrm{F2F}} = X^*$ *with high probability.*

Remarks

- Achieving the IT-limit $\mu^2 = 4\log n$

Belief propagation

Max-Product Belief Propagation

$$m_{i\to j}(t) = w_{ij} - \underset{\ell\neq j}{\text{2nd}\max}\left\{m_{\ell\to i}(t-1)\right\}$$
$$m_{i\to j}(0) = w_{ij}$$

After T iterations, for each vertex i, keep the two largest incoming messages $m_{\ell\to i}(T)$ and delete the rest.

Belief propagation

Max-Product Belief Propagation

$$m_{i\to j}(t) = w_{ij} - \operatorname{2nd\,max}_{\ell\neq j}\{m_{\ell\to i}(t-1)\}$$

$$m_{i\to j}(0) = w_{ij}$$

After T iterations, for each vertex i, keep the two largest incoming messages $m_{\ell\to i}(T)$ and delete the rest.

- BP is exact provided the optimal solution of F2F is integral [Bayati-Borgs-Chayes-Zecchina '11]
- It can be shown that $T = O(n^2 \log n)$ whp

Theoretical analysis of convex relaxation

Primal approach vs Dual approach: high level

- Dual argument:
  - Construct dual witness that certify the ground truth whp (KKT conditions)

Primal approach vs Dual approach: high level

- Dual argument:
  - Construct dual witness that certify the ground truth whp (KKT conditions)
  - Successful in proving SDP relaxation attaining sharp threshold for graph partitions: community detection, densest subgraph, etc [Abbe-Bandeira-Hall '14, Hajek-Wu-X. '14,'15, Bandeira '15, Perry-Wein '15]

Primal approach vs Dual approach: high level

- Dual argument:
  - Construct dual witness that certify the ground truth whp (KKT conditions)
  - Successful in proving SDP relaxation attaining sharp threshold for graph partitions: community detection, densest subgraph, etc [Abbe-Bandeira-Hall '14, Hajek-Wu-X. '14,'15, Bandeira '15, Perry-Wein '15]
  - Limitations: construction is ad hoc

Primal approach vs Dual approach: high level

- Dual argument:
  - Construct dual witness that certify the ground truth whp (KKT conditions)
  - Successful in proving SDP relaxation attaining sharp threshold for graph partitions: community detection, densest subgraph, etc [Abbe-Bandeira-Hall '14, Hajek-Wu-X. '14,'15, Bandeira '15, Perry-Wein '15]
  - Limitations: construction is ad hoc

- Primal argument:
  - No feasible solution other than the ground truth has a better objective value whp

Primal approach vs Dual approach: high level

- Dual argument:
  - Construct dual witness that certify the ground truth whp (KKT conditions)
  - Successful in proving SDP relaxation attaining sharp threshold for graph partitions: community detection, densest subgraph, etc [Abbe-Bandeira-Hall '14, Hajek-Wu-X. '14,'15, Bandeira '15, Perry-Wein '15]
  - Limitations: construction is ad hoc
- Primal argument:
  - No feasible solution other than the ground truth has a better objective value whp
  - Key: for LP, can restrict to **extremal points** (vertices of the feasible polytope)

Dual approach

- KKT conditions (Farkas' lemma): $\widehat{X}_{\mathrm{F2F}} = X^* \iff \exists u \in \mathbb{R}^n$ (dual certificate):

$$\begin{aligned} u_i + u_j &\le W_{ij}, \quad \text{for } i \sim j \text{ in } C^* \\ u_i + u_j &\ge W_{ij}, \quad \text{for } i \not\sim j \text{ in } C^* \end{aligned}$$

Dual approach

- KKT conditions (Farkas' lemma): $\widehat{X}_{\mathrm{F2F}} = X^* \iff \exists u \in \mathbb{R}^n$ (dual certificate):

$$\begin{aligned} u_i + u_j &\le W_{ij}, \quad \text{for } i \sim j \text{ in } C^* \\ u_i + u_j &\ge W_{ij}, \quad \text{for } i \not\sim j \text{ in } C^* \end{aligned}$$

- One feasible choice of dual:

$$u_i = \frac{1}{2} \min\{W_{ij} : j \sim i\}$$

Dual approach

- KKT conditions (Farkas' lemma): $\widehat{X}_{\mathrm{F2F}} = X^* \iff \exists u \in \mathbb{R}^n$ (dual certificate):

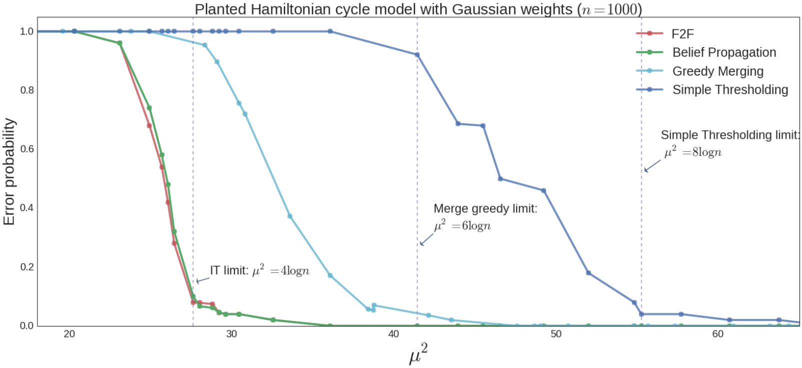
$$\begin{aligned} u_i + u_j &\le W_{ij}, \quad \text{for } i \sim j \text{ in } C^* \\ u_i + u_j &\ge W_{ij}, \quad \text{for } i \not\sim j \text{ in } C^* \end{aligned}$$

- One feasible choice of dual:

$$u_i = \frac{1}{2} \min\{W_{ij} : j \sim i\}$$

- This certificate shows correctness if $\mu^2 > 6 \log n$ (same as greedy merging)

Synthetic data experiment



Primal approach

- Show whp for all extremal points $X \neq X^*$:

$$\langle W, X \rangle < \langle W, X^* \rangle$$

- F2F polytope:

$$\left\{ X \in [0,1]^{n \times n} : \sum_{j=1}^{n} X_{ij} = 2 \right\}$$

- The proof heavily exploits the characterization of extremal points

Primal approach

- Show whp for all extremal points $X \neq X^*$:

$$\langle W, X \rangle < \langle W, X^* \rangle$$

- F2F polytope:

$$\left\{ X \in [0,1]^{n \times n} : \sum_{j=1}^{n} X_{ij} = 2 \right\}$$

- The proof heavily exploits the characterization of extremal points
  - F2F polytope is not integral: fractional vertices exist

Primal approach

- Show whp for all extremal points $X \neq X^*$:

$$\langle W, X \rangle < \langle W, X^* \rangle$$

- F2F polytope:

$$\left\{ X \in [0,1]^{n \times n} : \sum_{j=1}^{n} X_{ij} = 2 \right\}$$

- The proof heavily exploits the characterization of extremal points
  - F2F polytope is not integral: fractional vertices exist
  - Characterization [Balinski '65]: for any vertex X of F2F polytope
    - Half integrality

$$X_{ij} \in \{0, 1/2, 1\}$$

Primal approach

- Show whp for all extremal points $X \neq X^*$:

$$\langle W, X \rangle < \langle W, X^* \rangle$$

- F2F polytope:

$$\left\{ X \in [0,1]^{n \times n} : \sum_{j=1}^{n} X_{ij} = 2 \right\}$$

- The proof heavily exploits the characterization of extremal points
  - F2F polytope is not integral: fractional vertices exist
  - Characterization [Balinski '65]: for any vertex X of F2F polytope
    - Half integrality

$$X_{ij} \in \{0, 1/2, 1\}$$

    - $1/2$'s form disjoint odd cycle connected by path of 1's.

Primal approach

- Show whp for all extremal points $X \neq X^*$:

$$\langle W, X \rangle < \langle W, X^* \rangle$$

- F2F polytope:

$$\left\{ X \in [0,1]^{n \times n} : \sum_{j=1}^{n} X_{ij} = 2 \right\}$$

- The proof heavily exploits the characterization of extremal points
  - F2F polytope is not integral: fractional vertices exist
  - Characterization [Balinski '65]: for any vertex X of F2F polytope
    - Half integrality

$$X_{ij} \in \{0, 1/2, 1\}$$

    - $1/2$'s form disjoint odd cycle connected by path of 1's.

Proof of correctness for F2F LP

Proof Outline

1. Encode the solution: for any extremal point X, represent $2(X - X^*)$ as a bicolored multigraph G_X

$$w(G_X) = \langle W, 2(X - X^*) \rangle$$

Proof Outline

1. Encode the solution: for any extremal point X, represent $2(X - X^*)$ as a bicolored multigraph G_X

$$w(G_X) = \langle W, 2(X - X^*) \rangle$$

2. Divide and conquer: decompose G_X as edge-disjoint union of graphs in some family $\mathcal{F}$

$$w(G_X) = \sum_i w(F_i), \quad F_i \in \mathcal{F}$$

Proof Outline

1. Encode the solution: for any extremal point X, represent $2(X - X^*)$ as a bicolored multigraph G_X

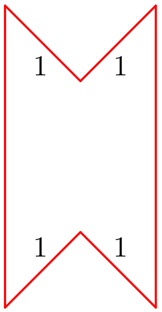
$$w(G_X) = \langle W, 2(X - X^*) \rangle$$

2. Divide and conquer: decompose G_X as edge-disjoint union of graphs in some family $\mathcal{F}$

$$w(G_X) = \sum_i w(F_i), \quad F_i \in \mathcal{F}$$

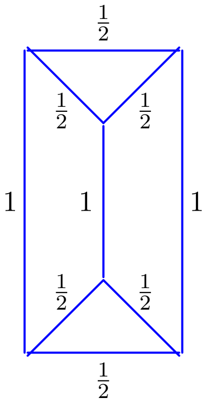
3. Counting: Show that whp $w(F) < 0$ for all $F \in \mathcal{F}$

Step 1: Bicolored multigraph representation



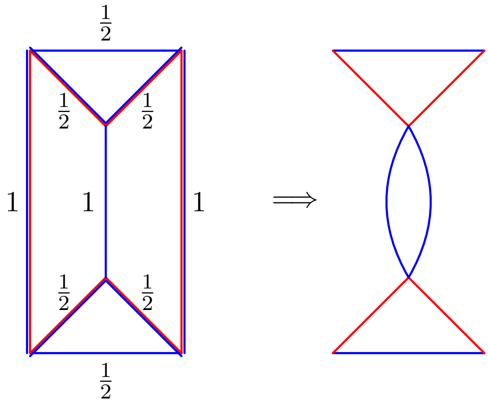
X^*: true cycle

Step 1: Bicolored multigraph representation



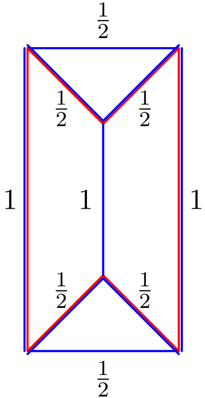
X: extremal solution

Step 1: Bicolored multigraph representation

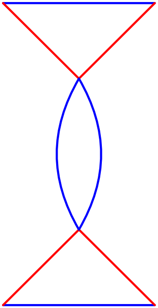


X: extremal solution $\qquad\qquad$ G_X

Step 1: Bicolored multigraph representation



$\Longrightarrow$



X: extremal solution

G_X

key observation

G_X is always balanced: red degree $=$ blue degree

Step 2: Edge decomposition

Theorem (Kotzig '68)

Every connected balanced bicolored multigraph has an alternating Eulerian circuit.

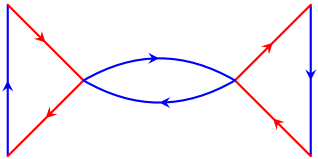
Step 2: Edge decomposition

Theorem (Kotzig '68)

Every connected balanced bicolored multigraph has an alternating Eulerian circuit.

Remarks

- An Eulerian circuit may traverse a double edge twice

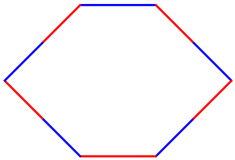


"Dumbbell" structure

Step 2: Edge decomposition

$\mathcal{U}$: collection of graphs recursively constructed

1. Start with an even cycle in alternating colors
2. **Blossoming procedure**: At each step, contract an edge in any cycle and attach a **flower** (alternating path of double edges followed by an alternating odd cycle)

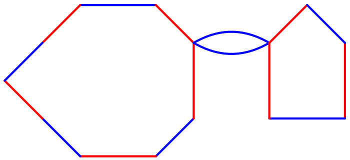


Obtained by starting with an 10-cycle and blossoming 3 times

Step 2: Edge decomposition

$\mathcal{U}$: collection of graphs recursively constructed

1. Start with an even cycle in alternating colors
2. **Blossoming procedure**: At each step, contract an edge in any cycle and attach a **flower** (alternating path of double edges followed by an alternating odd cycle)

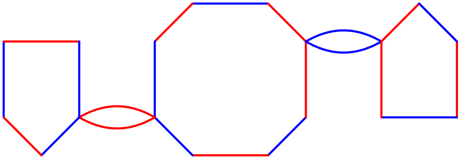


Obtained by starting with an 10-cycle and blossoming 3 times

Step 2: Edge decomposition

$\mathcal{U}$: collection of graphs recursively constructed

1. Start with an even cycle in alternating colors
2. **Blossoming procedure**: At each step, contract an edge in any cycle and attach a **flower** (alternating path of double edges followed by an alternating odd cycle)

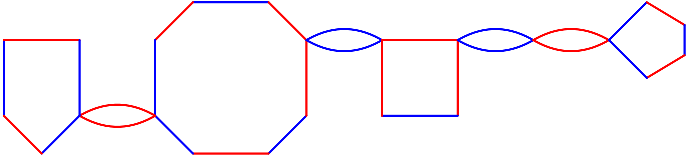


Obtained by starting with an 10-cycle and blossoming 3 times

Step 2: Edge decomposition

$\mathcal{U}$: collection of graphs recursively constructed

1. Start with an even cycle in alternating colors
2. **Blossoming procedure**: At each step, contract an edge in any cycle and attach a **flower** (alternating path of double edges followed by an alternating odd cycle)

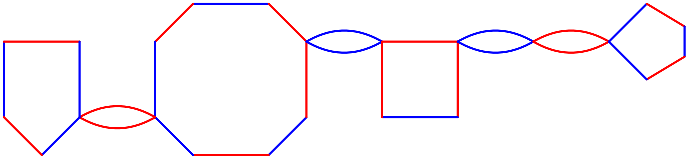


Obtained by starting with an 10-cycle and blossoming 3 times

Step 2: Edge decomposition

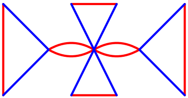
$\mathcal{U}$: collection of graphs recursively constructed

1. Start with an even cycle in alternating colors
2. **Blossoming procedure**: At each step, contract an edge in any cycle and attach a **flower** (alternating path of double edges followed by an alternating odd cycle)

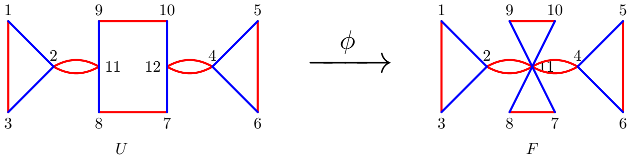


Obtained by starting with an 10-cycle and blossoming 3 times

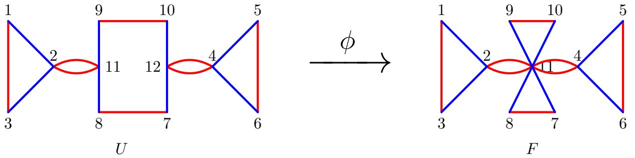
However, not every G_X is of this form...



- Graph homomorphism $\phi : U \to F$ is a vertex map that preserves edges and edge multiplicity



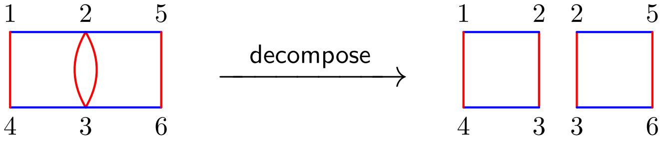
- Graph homomorphism $\phi : U \to F$ is a vertex map that preserves edges and edge multiplicity



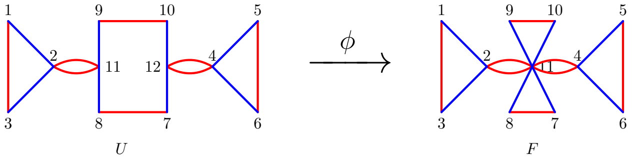
Lemma (Decomposition)

Every balanced bicolored multigraph G with edge multiplicity at most 2 can be decomposed as an edge-disjoint union of graphs in

$$\mathcal{F} = \{F : U \to F \text{ for some } U \in \mathcal{U}\}$$



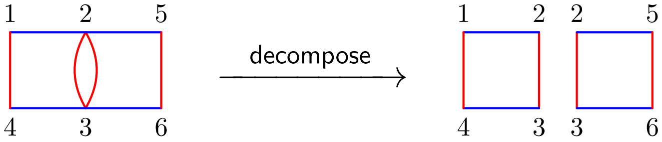
- Graph homomorphism $\phi : U \to F$ is a vertex map that preserves edges and edge multiplicity



Lemma (Decomposition)

Every balanced bicolored multigraph G with edge multiplicity at most 2 can be decomposed as an edge-disjoint union of graphs in

$$\mathcal{F} = \{F : U \to F \text{ for some } U \in \mathcal{U}\}$$



- It remains to show $\min_{F \in \mathcal{F}} w(F) < 0$ whp

Step 3: Counting and large deviation arguments

$\mathcal{F}_{k,\ell} = \{F \in \mathcal{F} : E(F)$ consists of k double edges and ℓ single edges $\}$

Lemma

For any $k \geq 0$ and $\ell \geq 3$. With probability at least $1 - n^{-\Theta(k+\ell)}$,

$$\max_{F \in \mathcal{F}_{k,\ell}} \left(w(F) - \mathbb{E}\left[w(F)\right]\right) \leq (1+\epsilon)\,(2k+\ell)\,\sqrt{\log n}$$

Step 3: Counting and large deviation arguments

$$\mathcal{F}_{k,\ell} = \{F \in \mathcal{F} : E(F) \text{ consists of } k \text{ double edges and } \ell \text{ single edges }\}$$

Lemma

For any $k \geq 0$ and $\ell \geq 3$. With probability at least $1 - n^{-\Theta(k+\ell)}$,

$$\max_{F \in \mathcal{F}_{k,\ell}} \left(w(F) - \mathbb{E}\left[w(F)\right]\right) \leq (1+\epsilon)\,(2k+\ell)\sqrt{\log n}$$

Remarks

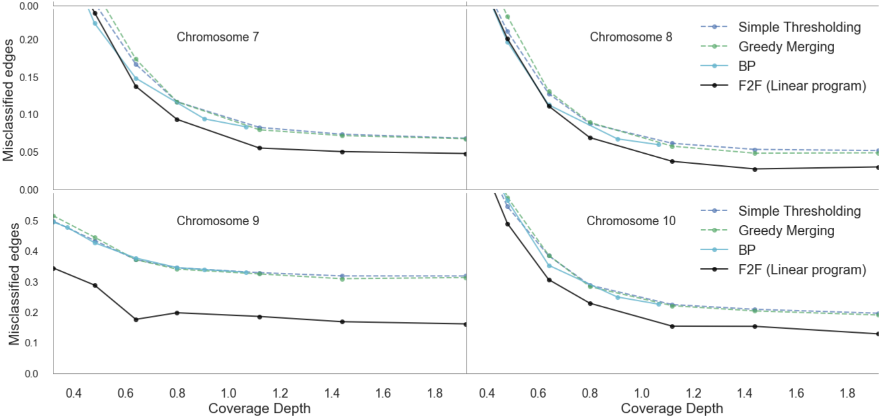
- Total: $2k + \ell$ edges, half red half blue.
- Weights on red edges $\sim \mathcal{N}(-\mu, 1)$; weights on blue edges $\sim \mathcal{N}(0, 1)$

$$w(F) \sim \mathcal{N}\left(-\frac{2k+\ell}{2}\mu,\ \mathbf{4}k + \ell\right)$$

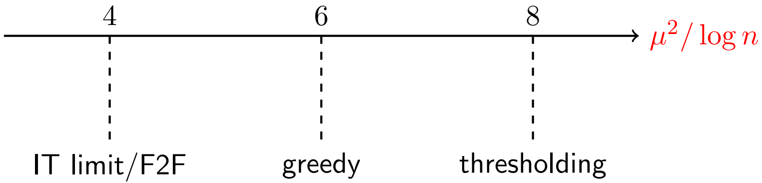
Real-data experiment

- 1000 DNA contigs of size 100 kbps
- 0.45 million Chicago cross-links
- Subsample each cross-link with probability p

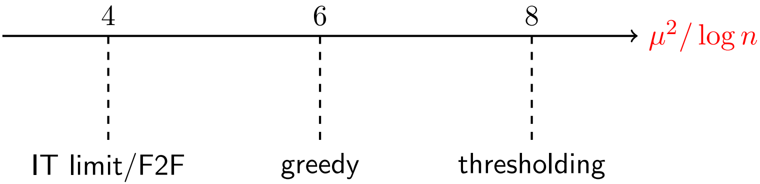
Homosapiens [Putnam et al 16, Genome Research]



Conclusion and remarks



Conclusion and remarks



References

- Vivek Bagaria, Jian Ding, David Tse, Yihong Wu & X. (2018). *Hidden Hamiltonian Cycle Recovery via Linear Programming*, https://arxiv.org/abs/1804.05436