A Proof of The Changepoint Detection Threshold Conjecture in Preferential Attachment Models

Jiaming Xu

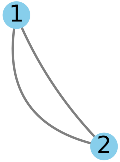
The Fuqua School of Business
Duke University

Joint work with
Hang Du (MIT) and Shuyang Gong (PKU)

Workshop on Statistical Network Analysis and Beyond
June 2, 2025

Preferential attachment models

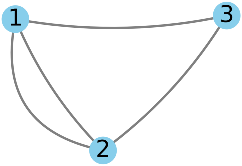
Initial graph G_2 consists of two vertices connected by m parallel edges



Preferential attachment models

At each time t, a new vertex t arrives and forms m edges, one at a time, to existing nodes $v \in [t-1]$:

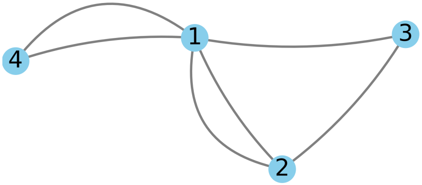
$$\mathbb{P}\{t \to v\} \propto \mathsf{deg}(v) + \delta_t$$



Preferential attachment models

At each time t, a new vertex t arrives and forms m edges, one at a time, to existing nodes $v \in [t-1]$:

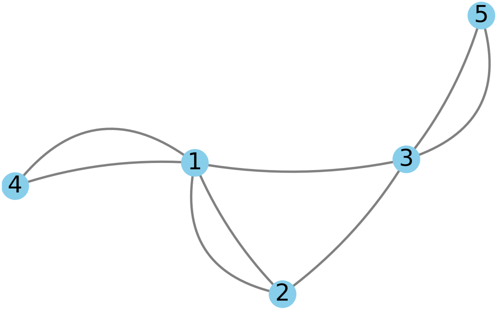
$$\mathbb{P}\{t \to v\} \propto \mathsf{deg}(v) + \delta_t$$



Preferential attachment models

At each time t, a new vertex t arrives and forms m edges, one at a time, to existing nodes $v \in [t-1]$:

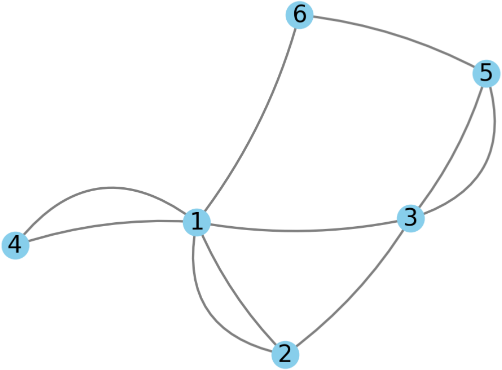
$$\mathbb{P}\{t \to v\} \propto \mathsf{deg}(v) + \delta_t$$



Preferential attachment models

At each time t, a new vertex t arrives and forms m edges, one at a time, to existing nodes $v \in [t-1]$:

$$\mathbb{P}\{t \to v\} \propto \deg(v) + \delta_t$$



Preferential attachment models

At each time t, a new vertex t arrives and forms m edges, one at a time, to existing nodes $v \in [t-1]$:

$$\mathbb{P}\{t \to v\} \propto \deg(v) + \delta_t,$$

- $\deg(v)$ is updated after each edge is added
- $\delta_t = \infty$: uniform attachment (ignore degrees)
- $\delta_t = 0$: Barabási-Albert model [Barabási-Albert '99]
- The smaller δ_t, the stronger preference for high-degree vertices
- A most popular dynamic graph model: various properties (e.g. limiting degree distribution) are well-understood [van der Hofstad '16 '24]

Changepoint detection problem

Definition

$$\mathbb{H}_0 : \delta_t = \delta$$

$$\mathbb{H}_1 : \delta_t = \delta \mathbf{1}_{t \le \tau_n} + \delta' \mathbf{1}_{\tau_n < t \le n}$$

Changepoint detection problem

Definition

$$\mathbb{H}_0 : \delta_t = \delta$$
$$\mathbb{H}_1 : \delta_t = \delta \mathbf{1}_{t \le \tau_n} + \delta' \mathbf{1}_{\tau_n < t \le n}$$

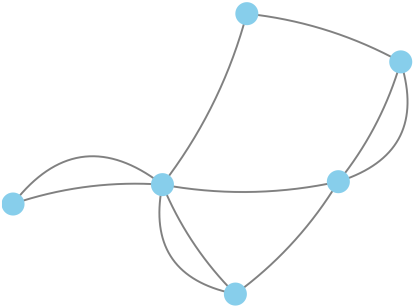
- $\delta \neq \delta' > -m$ are two fixed constants

Changepoint detection problem

Definition

$$\mathbb{H}_0 : \delta_t = \delta$$
$$\mathbb{H}_1 : \delta_t = \delta \mathbf{1}_{t \le \tau_n} + \delta' \mathbf{1}_{\tau_n < t \le n}$$

- $\delta \neq \delta' > -m$ are two fixed constants
- Only final network snapshot is observed (node arrival time unknown)



Changepoint detection problem

Definition

$$\mathbb{H}_0 : \delta_t = \delta$$
$$\mathbb{H}_1 : \delta_t = \delta \mathbf{1}_{t \le \tau_n} + \delta' \mathbf{1}_{\tau_n < t \le n}$$

- $\delta \neq \delta' > -m$ are two fixed constants
- Only final network snapshot is observed (node arrival time unknown)
- Problem gets harder with increasing τ_n: Quickest change detection

Changepoint detection problem

Definition

$$\mathbb{H}_0 : \delta_t = \delta$$
$$\mathbb{H}_1 : \delta_t = \delta \mathbf{1}_{t \le \tau_n} + \delta' \mathbf{1}_{\tau_n < t \le n}$$

- $\delta \neq \delta' > -m$ are two fixed constants
- Only final network snapshot is observed (node arrival time unknown)
- Problem gets harder with increasing τ_n: Quickest change detection
- Changepoint localization: estimate τ_n under $\mathbb{H}_1$ [Bhamidi-Jin-Nobel '18]

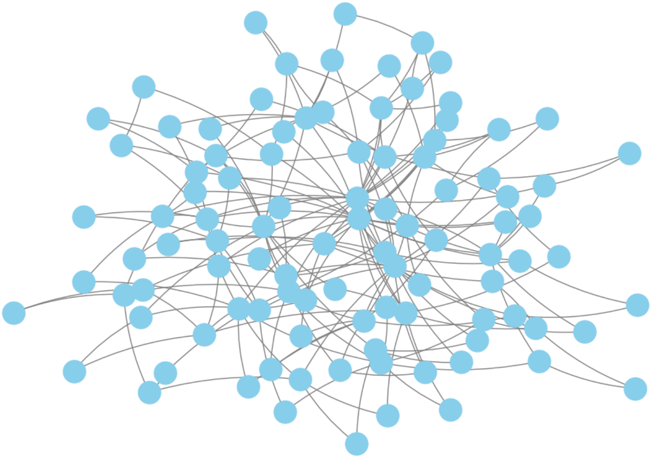
Changepoint detection problem

Definition

$$\mathbb{H}_0 : \delta_t = \delta$$
$$\mathbb{H}_1 : \delta_t = \delta \mathbf{1}_{t \le \tau_n} + \delta' \mathbf{1}_{\tau_n < t \le n}$$

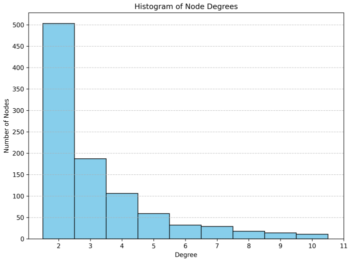
- $\delta \neq \delta' > -m$ are two fixed constants
- Only final network snapshot is observed (node arrival time unknown)
- Problem gets harder with increasing τ_n: Quickest change detection
- Changepoint localization: estimate τ_n under $\mathbb{H}_1$ [Bhamidi-Jin-Nobel '18]
- Applications: detect structural changes in various settings, such as communication networks, social networks, financial networks, and biological networks [Cirkovic-Wang-Zhang '24].

Looks like a daunting task

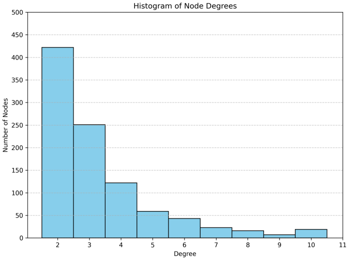


Change or no change?

A simple test based on minimum-degree



$n = 1000,\ m = 2,\ \delta(t) \equiv 0$



$n = 1000,\ m = 2,\ \delta(t) = 10 \cdot \mathbf{1}\left(t > n - n^{0.8}\right)$

A simple test based on minimum-degree

- Let $N_m(G_n)$ denote the number of degree-m vertices
- Let $p_m(\delta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left[N_m(G_n)\right]$ under $\mathcal{H}_0$
- Consider test $T(G_n) = N_m(G_n) - np_m(\delta)$

A simple test based on minimum-degree

- Let $N_m(G_n)$ denote the number of degree-m vertices
- Let $p_m(\delta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left[N_m(G_n)\right]$ under $\mathcal{H}_0$
- Consider test $T(G_n) = N_m(G_n) - np_m(\delta)$

Theorem (Bet-Bogerd-Castro-van der Hofstad '23)

Suppose $\tau_n = n - cn^\gamma$ for a constant c and $\gamma \in (0,1)$. If $\gamma > 1/2$, by choosing $\alpha_n/\sqrt{n}$ slowly tending to infinity,

$$\mathbb{P}_0\left\{|T(G_n)| \geq \alpha_n\right\} + \mathbb{P}_1\left\{|T(G_n)| \leq \alpha_n\right\} \to 0$$

A simple test based on minimum-degree

- Let $N_m(G_n)$ denote the number of degree-m vertices
- Let $p_m(\delta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left[N_m(G_n)\right]$ under $\mathcal{H}_0$
- Consider test $T(G_n) = N_m(G_n) - np_m(\delta)$

Theorem (Bet-Bogerd-Castro-van der Hofstad '23)

Suppose $\tau_n = n - cn^\gamma$ *for a constant* c *and* $\gamma \in (0,1)$. *If* $\gamma > 1/2$, *by choosing* $\alpha_n/\sqrt{n}$ *slowly tending to infinity,*

$$\mathbb{P}_0\left\{|T(G_n)| \geq \alpha_n\right\} + \mathbb{P}_1\left\{|T(G_n)| \leq \alpha_n\right\} \to 0$$

- Intuition: There are $\Theta(1)$ fraction of degree-m nodes $\Rightarrow$ probability of attaching to degree-m nodes changes by $\Theta(1)$ after $\tau_n \Rightarrow$ $\mathbb{E}_1[T] = \Theta(n^\gamma)$, while $\mathrm{Std}[T] = O(\sqrt{n})$

A simple test based on minimum-degree

- Let $N_m(G_n)$ denote the number of degree-m vertices
- Let $p_m(\delta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left[N_m(G_n)\right]$ under $\mathcal{H}_0$
- Consider test $T(G_n) = N_m(G_n) - np_m(\delta)$

Theorem (Bet-Bogerd-Castro-van der Hofstad '23)

Suppose $\tau_n = n - cn^\gamma$ *for a constant* c *and* $\gamma \in (0,1)$. *If* $\gamma > 1/2$, *by choosing* $\alpha_n/\sqrt{n}$ *slowly tending to infinity,*

$$\mathbb{P}_0\left\{|T(G_n)| \ge \alpha_n\right\} + \mathbb{P}_1\left\{|T(G_n)| \le \alpha_n\right\} \to 0$$

- Intuition: There are $\Theta(1)$ fraction of degree-m nodes $\Rightarrow$ probability of attaching to degree-m nodes changes by $\Theta(1)$ after $\tau_n \Rightarrow$ $\mathbb{E}_1[T] = \Theta(n^\gamma)$, while $\mathrm{Std}[T] = O(\sqrt{n})$
- If δ is unknown, can be replaced by a ML estimator

A simple test based on minimum-degree

- Let $N_m(G_n)$ denote the number of degree-m vertices
- Let $p_m(\delta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_0\left[N_m(G_n)\right]$ under $\mathcal{H}_0$
- Consider test $T(G_n) = N_m(G_n) - np_m(\delta)$

Theorem (Bet-Bogerd-Castro-van der Hofstad '23)

Suppose $\tau_n = n - cn^\gamma$ for a constant c and $\gamma \in (0,1)$. If $\gamma > 1/2$, by choosing $\alpha_n/\sqrt{n}$ slowly tending to infinity,

$$\mathbb{P}_0\left\{|T(G_n)| \ge \alpha_n\right\} + \mathbb{P}_1\left\{|T(G_n)| \le \alpha_n\right\} \to 0$$

- Intuition: There are $\Theta(1)$ fraction of degree-m nodes $\Rightarrow$ probability of attaching to degree-m nodes changes by $\Theta(1)$ after $\tau_n \Rightarrow$ $\mathbb{E}_1[T] = \Theta(n^\gamma)$, while $\mathrm{Std}[T] = O(\sqrt{n})$
- If δ is unknown, can be replaced by a ML estimator
- Can establish weak detection when $\gamma = 1/2$

Changepoint detection conjecture

Conjecture (Bet-Bogerd-Castro-van der Hofstad '23)

Suppose $\tau_n = n - cn^\gamma$ *for a constant* c *and* $\gamma < 1/2$.

1. *All tests based on vertex degrees are powerless.*
2. *All tests are powerless.*

- Part 2 of the conjecture is particularly striking, because, if true, neither degree information nor any higher-level graph structure is useful for detection when $\gamma < 1/2$

Significant progress

Theorem (Kaddouri-Naulet-Gassiat '24)

Suppose $\tau_n = n - \Delta$. *If* $\Delta = o(n^{1/3})$ *for* $\delta > 0$ *or* $\Delta = o(n^{1/3}/\log n)$ *for* $\delta = 0$, *then*

$$\mathbb{P}_0(A_n) \to 0 \Longrightarrow \mathbb{P}_1(A_n) \to 0, \quad \text{for all sequences of events } A_n$$

Significant progress

Theorem (Kaddouri-Naulet-Gassiat '24)

Suppose $\tau_n = n - \Delta$. *If* $\Delta = o(n^{1/3})$ *for* $\delta > 0$ *or* $\Delta = o(n^{1/3}/\log n)$ *for* $\delta = 0$, *then*

$$\mathbb{P}_0(A_n) \to 0 \Longrightarrow \mathbb{P}_1(A_n) \to 0, \quad \textit{for all sequences of events } A_n$$

- As a consequence, $\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) \leq 1 - \Omega(1) \Rightarrow$ strong detection is impossible
- Does not cover the entire regime $\Delta = o(\sqrt{n})$ and the regime $\delta < 0$
- Does not rule out the possibility of weak detection

Our resolution

Theorem (Du-Gong-X. '25)

Suppose $\tau_n = n - \Delta$. *If* $\Delta = o(n^{1/2})$, *then*

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) = o(1)$$

Our resolution

Theorem (Du-Gong-X. '25)

Suppose $\tau_n = n - \Delta$. *If* $\Delta = o(n^{1/2})$, *then*

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) = o(1)$$

- As a consequence, all tests are powerless $\Rightarrow$ resolves the changepoint detection conjecture [Bet-Bogerd-Castro-van der Hofstad '23] in positive
- We prove a stronger statement: all tests remain powerless even if, in addition to G_n, the entire network history were observed up to time $n - N$ for $\Delta^2 \ll N \ll n$
- As a corollary, we prove no estimator can locate τ_n within $o(\sqrt{n})$ with $\Omega(1)$ probability $\Rightarrow$ the estimator in [Bhamidi-Jin-Nobel'18], which achieves $|\hat{\tau}_n - \tau_n| = O_P(\sqrt{n})$, is order-optimal

Proof ideas

Challenge of directly bounding likelihood ratio

Define the likelihood ratio

$$L(G) \triangleq \frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)}$$

Then

$$\mathrm{Var}_{G_n \sim \mathbb{P}_0}\left[L(G_n)\right] = o(1) \implies \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_0) = o(1)$$

Challenge of directly bounding likelihood ratio

Define the likelihood ratio

$$L(G) \triangleq \frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)}$$

Then

$$\mathrm{Var}_{G_n \sim \mathbb{P}_0}\left[L(G_n)\right] = o(1) \implies \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_0) = o(1)$$

- Widely used to prove impossibility of detection in high-dimensional statistics and network analysis (e.g. community detection)

Challenge of directly bounding likelihood ratio

Define the likelihood ratio

$$L(G) \triangleq \frac{\mathbb{P}_1(G)}{\mathbb{P}_0(G)}$$

Then

$$\mathrm{Var}_{G_n \sim \mathbb{P}_0}\left[L(G_n)\right] = o(1) \implies \mathrm{TV}(\mathbb{P}_1, \mathbb{P}_0) = o(1)$$

- Widely used to prove impossibility of detection in high-dimensional statistics and network analysis (e.g. community detection)
- However, since only final network snapshot is observed, $L(G_n)$ involves an average over compatible network histories, making it hard to bound its variance directly

Consider an “easier” problem

- To simplify the likelihood ratio, one can make the problem “easier” by revealing network history
- However, revealing entire network history renders problem too easy...

Consider an "easier" problem

- To simplify the likelihood ratio, one can make the problem "easier" by revealing network history
- However, revealing entire network history renders problem too easy...

Theorem (Kaddouri-Naulet-Gassiat '24)

Denote $\overline{G}_n$ *as the entire network history and* $\overline{\mathbb{P}}_1, \overline{\mathbb{P}}_0$ *as its law under* $\mathcal{H}_0, \mathcal{H}_1$*, respectively. Then*

$$\mathrm{TV}(\overline{\mathbb{P}}_1, \overline{\mathbb{P}}_0) = 1 - o(1),$$

if and only if $\Delta \triangleq n - \tau_n \to \infty$.

Consider an "easier" problem

- To simplify the likelihood ratio, one can make the problem "easier" by revealing network history
- However, revealing entire network history renders problem too easy...

Theorem (Kaddouri-Naulet-Gassiat '24)

Denote $\overline{G}_n$ as the entire network history and $\overline{\mathbb{P}}_1, \overline{\mathbb{P}}_0$ as its law under $\mathcal{H}_0, \mathcal{H}_1$, respectively. Then

$$\mathrm{TV}(\overline{\mathbb{P}}_1, \overline{\mathbb{P}}_0) = 1 - o(1),$$

if and only if $\Delta \triangleq n - \tau_n \to \infty$.

- Reveal arrival times of all vertices, except for a carefully chosen subset of $n^{2/3}$ leaf vertices $\Rightarrow \Delta \ll n^{1/3}$

Consider an "easier" problem

- To simplify the likelihood ratio, one can make the problem "easier" by revealing network history
- However, revealing entire network history renders problem too easy...

Theorem (Kaddouri-Naulet-Gassiat '24)

Denote $\overline{G}_n$ as the entire network history and $\overline{\mathbb{P}}_1, \overline{\mathbb{P}}_0$ as its law under $\mathcal{H}_0, \mathcal{H}_1$, respectively. Then

$$\mathrm{TV}(\overline{\mathbb{P}}_1, \overline{\mathbb{P}}_0) = 1 - o(1),$$

if and only if $\Delta \triangleq n - \tau_n \to \infty$.

- Reveal arrival times of all vertices, except for a carefully chosen subset of $n^{2/3}$ leaf vertices $\Rightarrow \Delta \ll n^{1/3}$
- However, to prove the impossibility up to $\Delta = o(\sqrt{n})$, can only reveal network history up to $n - o(n)$

Our proof strategy

1. Interpolation: reduce to analyzing changepoint $\tau_n = n - 1$
2. Simplified model: reveal network history up to time $n - o(n)$
3. Derive the likelihood ratio
4. Bound its variance via Efron-Stein inequality and coupling

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

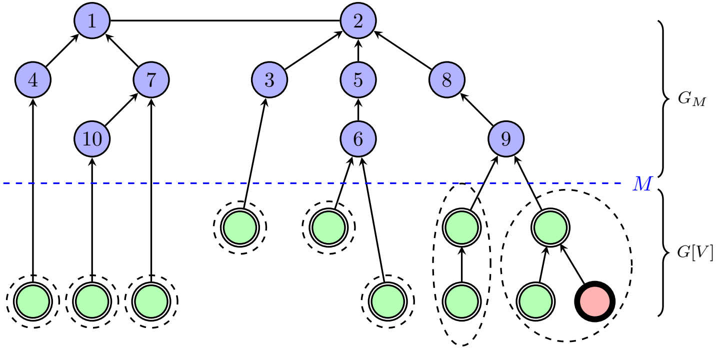
$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

- Applying triangle's and data-processing inequality, reduces to show

$$\mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-1}) = o\left(\frac{1}{\Delta}\right),$$

Step 2: Consider an "easier" model

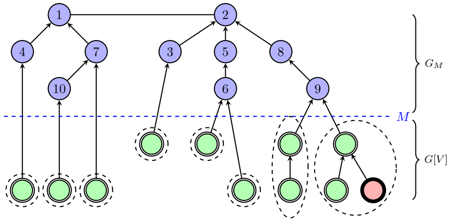
Reveal the network history up to time $M = n - N$ where $\Delta^2 \ll N \ll n$



$m = 1$ and $\tau_n = n - 1$: connected components are denoted by dashed ellipses

Step 3: Derive the likelihood ratio

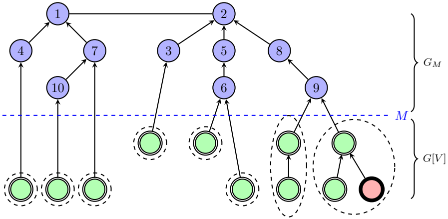
Let V denote the set of vertices arriving after time $M = n - N$. Consider the subgraph of G_n induced by V and let $\mathcal{C}(v)$ denote its connected component containing $v \in V$.



$m = 1$ and $\tau_n = n - 1$: connected components are denoted by dashed ellipses

Step 3: Derive the likelihood ratio

Let V denote the set of vertices arriving after time $M = n - N$. Consider the subgraph of G_n induced by V and let $\mathcal{C}(v)$ denote its connected component containing $v \in V$.

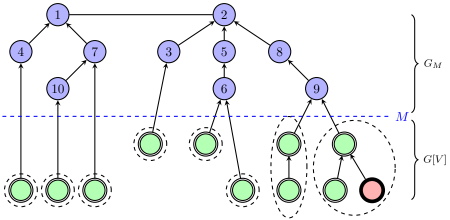


$m = 1$ and $\tau_n = n - 1$: connected components are denoted by dashed ellipses

Key: The connected components can arrive in any relative order

Step 3: Derive the likelihood ratio

Let V denote the set of vertices arriving after time $M = n - N$. Consider the subgraph of G_n induced by V and let $\mathcal{C}(v)$ denote its connected component containing $v \in V$.



$m = 1$ and $\tau_n = n - 1$: connected components are denoted by dashed ellipses

Then the likelihood ratio

$$L = \frac{C_1}{N} \sum_{v \in V} |\mathcal{C}(v)| \lambda_v X_v,$$

where C_1 is bounded constant, $\sum_{w \in \mathcal{C}(v)} \lambda_w = 1$, and $c_1 \leq X_v \leq c_2$.

Step 4: Efron-Stein inequality and coupling

- Encode the PA model using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n,1\leq i\leq m}$

Step 4: Efron-Stein inequality and coupling

- Encode the PA model using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\le n, 1\le i\le m}$
- Let $U = (U_{M+1,1}, \ldots, U_{t,i}, \ldots, U_{n,m})$ and $U^{(t,i)} = (U_{M+1,1}, \ldots, U'_{t,i}, \ldots, U_{n,m})$, where $U'_{t,i}$ is an independent copy of $U_{t,i}$. Write LRT L as $f(U)$ and apply Efron-Stein

$$\mathrm{Var}[L] \le \frac{1}{2} \sum_{M<t\le n} \sum_{1\le i\le m} \mathbb{E}\left[\left(f(U) - f(U^{(t,i)}) \right)^2 \right]$$

Step 4: Efron-Stein inequality and coupling

- Encode the PA model using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$
- Let $U = (U_{M+1,1}, \ldots, U_{t,i}, \ldots, U_{n,m})$ and $U^{(t,i)} = (U_{M+1,1}, \ldots, U'_{t,i}, \ldots, U_{n,m})$, where $U'_{t,i}$ is an independent copy of $U_{t,i}$. Write LRT L as $f(U)$ and apply Efron-Stein

$$\begin{aligned}\mathrm{Var}[L] &\leq \frac{1}{2} \sum_{M<t\leq n} \sum_{1\leq i\leq m} \mathbb{E}\left[\left(f(U) - f(U^{(t,i)})\right)^2\right] \\ &\leq O\left(\frac{1}{N}\right)\end{aligned}$$

Step 4: Efron-Stein inequality and coupling

- Encode the PA model using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$
- Let $U = (U_{M+1,1}, \ldots, U_{t,i}, \ldots, U_{n,m})$ and $U^{(t,i)} = (U_{M+1,1}, \ldots, U'_{t,i}, \ldots, U_{n,m})$, where $U'_{t,i}$ is an independent copy of $U_{t,i}$. Write LRT L as $f(U)$ and apply Efron-Stein

$$\begin{aligned}\mathrm{Var}[L] &\leq \frac{1}{2} \sum_{M<t\leq n} \sum_{1\leq i\leq m} \mathbb{E}\left[\left(f(U) - f(U^{(t,i)})\right)^2\right] \\ &\leq O\left(\frac{1}{N}\right)\end{aligned}$$

- Bound TV (recall $\Delta^2 \ll N \ll n$):

$$2\mathrm{TV} = \mathbb{E}\left[|L-1|\right] \leq \sqrt{\mathrm{Var}\left[L\right]} = O\left(\frac{1}{\sqrt{N}}\right) = o\left(\frac{1}{\Delta}\right)$$

Concluding remarks

- We show changepoint detection threshold is $\tau_n = n - o(\sqrt{n})$, confirming a conjecture of [Bet-Bogerd-Castro-van der Hofstad '23]
- As by-product, we show changepoint localization threshold is also $\tau_n = n - o(\sqrt{n})$, matching upper bound in [Bhamidi-Jin-Nobel '18]
- Key proof ideas: reduces to bounding TV when changepoint occurs at $n-1$, reveal network history up to $n - o(n)$, and bound the variance of likelihood ratio using Efron-Stein and coupling

Concluding remarks

- We show changepoint detection threshold is $\tau_n = n - o(\sqrt{n})$, confirming a conjecture of [Bet-Bogerd-Castro-van der Hofstad '23]
- As by-product, we show changepoint localization threshold is also $\tau_n = n - o(\sqrt{n})$, matching upper bound in [Bhamidi-Jin-Nobel '18]
- Key proof ideas: reduces to bounding TV when changepoint occurs at $n-1$, reveal network history up to $n - o(n)$, and bound the variance of likelihood ratio using Efron-Stein and coupling

Future directions

- General attachment rule: $\mathbb{P}(t \to v) \propto f(\deg(v))$ [Banerjee-Bhamidi-Carmichael '22]
- Changepoint detection in general dynamic graph models
- Other related reconstruction and estimation problems in PA graphs

Concluding remarks

- We show changepoint detection threshold is $\tau_n = n - o(\sqrt{n})$, confirming a conjecture of [Bet-Bogerd-Castro-van der Hofstad '23]
- As by-product, we show changepoint localization threshold is also $\tau_n = n - o(\sqrt{n})$, matching upper bound in [Bhamidi-Jin-Nobel '18]
- Key proof ideas: reduces to bounding TV when changepoint occurs at $n-1$, reveal network history up to $n - o(n)$, and bound the variance of likelihood ratio using Efron-Stein and coupling

Future directions

- General attachment rule: $\mathbb{P}(t \to v) \propto f(\text{deg}(v))$ [Banerjee-Bhamidi-Carmichael '22]
- Changepoint detection in general dynamic graph models
- Other related reconstruction and estimation problems in PA graphs

References

- Hang Du, Shuyang Gong, & Jiaming Xu. *A Proof of The Changepoint Detection Threshold Conjecture in Preferential Attachment Models*, arXiv:2502.00514, COLT 2025.

Backup slides

Limitation of previous strategy

- Reveal arrival times of all vertices, except for a carefully chosen subset $\mathcal{S}$ of leaf vertices (**bolded red vertices** shown below):

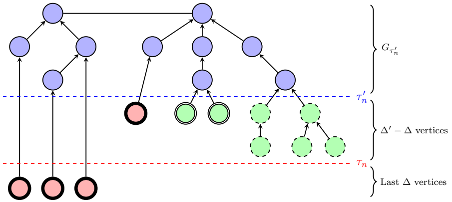


Figure credit [Kaddouri-Naulet-Gassiat '24]: $m = 1$

Limitation of previous strategy

- Reveal arrival times of all vertices, except for a carefully chosen subset $\mathcal{S}$ of leaf vertices (**bolded red vertices** shown below):

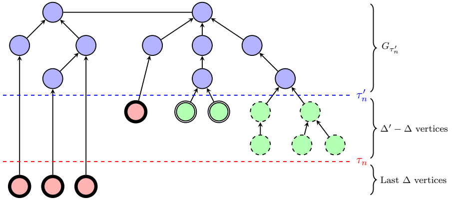


Figure credit [Kaddouri-Naulet-Gassiat '24]: $m = 1$

- $\mathcal{S}$ needs to contain all vertices arriving after τ_n, which happens w.p.

$$\approx \left(1 - \Delta'/n\right)^{\Delta} = 1 + o(1) \text{ when } \Delta'\Delta \ll n$$

Limitation of previous strategy

- Reveal arrival times of all vertices, except for a carefully chosen subset $\mathcal{S}$ of leaf vertices (**bolded red vertices** shown below):

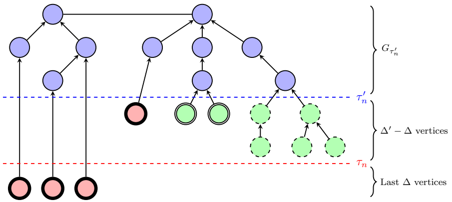


Figure credit [Kaddouri-Naulet-Gassiat '24]: $m = 1$

- $\mathcal{S}$ needs to contain all vertices arriving after τ_n, which happens w.p.

$$\approx \left(1 - \Delta'/n\right)^{\Delta} = 1 + o(1) \text{ when } \Delta'\Delta \ll n$$

- For detection to be impossible, also need $|\mathcal{S}| \asymp \Delta' \gg \Delta^2$
$\Rightarrow \Delta \ll n^{1/3}$

Challenge in the regime $n^{1/3} \lesssim \Delta \ll \sqrt{n}$

- To prove the impossibility up to $\Delta \leq o(\sqrt{n})$, can only reveal network history up to $\tau_n' = n - \Delta'$, where $\Delta^2 \ll \Delta' \ll n$

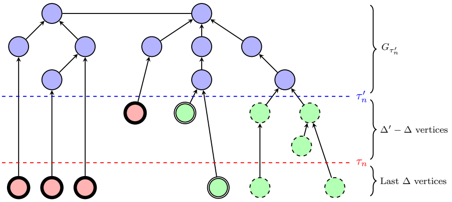


Figure credit [Kaddouri-Naulet-Gassiat '24]: $m = 1$

- Vertices arriving after τ_n may attach to vertices arrived in $[\tau_n' + 1, \tau_n]$

Our proof strategy

1. Interpolation: reduce to analyzing changepoint $\tau_n = n - 1$
2. Simplified model: reveal network history up to time $n - o(n)$
3. Bound TV by the second moment of likelihood ratio
4. Use Efron-Stein inequality and coupling

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) = \mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-\Delta})$$

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

$$\begin{aligned} \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-\Delta}) \\ &\le \sum_{k=1}^{\Delta} \mathrm{TV}(\mathbb{P}_{n,n-k+1}, \mathbb{P}_{n,n-k}) \quad \text{triangle's inequality} \end{aligned}$$

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

$$\begin{aligned} \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-\Delta}) \\ &\leq \sum_{k=1}^{\Delta} \mathrm{TV}(\mathbb{P}_{n,n-k+1}, \mathbb{P}_{n,n-k}) \quad \text{triangle's inequality} \\ &\overset{\mathrm{DP}}{\leq} \sum_{k=1}^{\Delta} \mathrm{TV}(\mathbb{P}_{n-k+1,n-k+1}, \mathbb{P}_{n-k+1,n-k}) \end{aligned}$$

Step 1: Interpolation

- $\mathbb{P}_{n,n-k}$: distribution of G_n with changepoint at time $n-k$

$$\mathbb{P}_0 = \mathbb{P}_{n,n} \to \mathbb{P}_{n,n-1} \to \mathbb{P}_{n,n-2} \to \cdots \to \mathbb{P}_{n,n-\Delta-1} \to \mathbb{P}_{n,n-\Delta} = \mathbb{P}_1$$

$$\begin{aligned} \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-\Delta}) \\ &\le \sum_{k=1}^{\Delta} \mathrm{TV}(\mathbb{P}_{n,n-k+1}, \mathbb{P}_{n,n-k}) \quad \text{triangle's inequality} \\ &\overset{\text{DP}}{\le} \sum_{k=1}^{\Delta} \mathrm{TV}(\mathbb{P}_{n-k+1,n-k+1}, \mathbb{P}_{n-k+1,n-k}) \end{aligned}$$

- Suffices to show

$$\mathrm{TV}(\mathbb{P}_{n',n'}, \mathbb{P}_{n',n'-1}) = o\left(\frac{1}{\Delta}\right), \quad \forall n' \in [n-\Delta+1, n]$$

WLOG, focus on $n' = n$ and $\tau_n = n-1$ henceforth

Step 2: Consider an "easier" model

- Reveal the network history up to time $M = n - N$, denoted by $\overline{G}_M$, where $\Delta^2 \ll N \ll n$

Step 2: Consider an "easier" model

- Reveal the network history up to time $M = n - N$, denoted by $\overline{G}_M$, where $\Delta^2 \ll N \ll n$
- Let $\mathcal{P}$ and $\mathcal{Q}$ denote the joint law of $\overline{G}_M$ and G_n, under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively

$$\mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-1}) = \mathrm{TV}(\mathcal{P}_{G_n}, \mathcal{Q}_{G_n})$$

Step 2: Consider an "easier" model

- Reveal the network history up to time $M = n - N$, denoted by $\overline{G}_M$, where $\Delta^2 \ll N \ll n$
- Let $\mathcal{P}$ and $\mathcal{Q}$ denote the joint law of $\overline{G}_M$ and G_n, under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively

$$\mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-1}) = \mathrm{TV}(\mathcal{P}_{G_n}, \mathcal{Q}_{G_n}) \overset{\mathrm{DP}}{\leq} \mathrm{TV}\left(\mathcal{P}_{G_n, \overline{G}_M}, \mathcal{Q}_{G_n, \overline{G}_M}\right)$$

Step 2: Consider an "easier" model

- Reveal the network history up to time $M = n - N$, denoted by $\overline{G}_M$, where $\Delta^2 \ll N \ll n$
- Let $\mathcal{P}$ and $\mathcal{Q}$ denote the joint law of $\overline{G}_M$ and G_n, under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively

$$
\begin{aligned}
\mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-1}) &= \mathrm{TV}(\mathcal{P}_{G_n}, \mathcal{Q}_{G_n}) \\
&\overset{\text{DP}}{\leq} \mathrm{TV}\left(\mathcal{P}_{G_n,\overline{G}_M}, \mathcal{Q}_{G_n,\overline{G}_M}\right) \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}_{\overline{G}_M \sim \mathcal{P}_{\overline{G}_M}}\left[\mathrm{TV}\left(\mathcal{P}_{G_n|\overline{G}_M}, \mathcal{Q}_{G_n|\overline{G}_M}\right)\right]
\end{aligned}
$$

Step 2: Consider an "easier" model

- Reveal the network history up to time $M = n - N$, denoted by $\overline{G}_M$, where $\Delta^2 \ll N \ll n$
- Let $\mathcal{P}$ and $\mathcal{Q}$ denote the joint law of $\overline{G}_M$ and G_n, under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively

$$\begin{aligned}
\mathrm{TV}(\mathbb{P}_{n,n}, \mathbb{P}_{n,n-1}) &= \mathrm{TV}(\mathcal{P}_{G_n}, \mathcal{Q}_{G_n}) \\
&\overset{\text{DP}}{\leq} \mathrm{TV}\left(\mathcal{P}_{G_n,\overline{G}_M}, \mathcal{Q}_{G_n,\overline{G}_M}\right) \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}_{\overline{G}_M \sim \mathcal{P}_{\overline{G}_M}}\left[\mathrm{TV}\left(\mathcal{P}_{G_n|\overline{G}_M}, \mathcal{Q}_{G_n|\overline{G}_M}\right)\right]
\end{aligned}$$

- Reduce to proving

$$\mathrm{TV}\left(\mathcal{P}_{G_n|\overline{G}_M}, \mathcal{Q}_{G_n|\overline{G}_M}\right) = o\left(\frac{1}{\Delta}\right), \quad \forall \overline{G}_M$$

Step 3: Bound the second moment

- Define likelihood ratio $L \triangleq \frac{\mathcal{Q}_{G_n|\overline{G}_M}}{\mathcal{P}_{G_n|\overline{G}_M}}$. Then

$$2\mathrm{TV}\left(\mathcal{P}_{G_n|\overline{G}_M}, \mathcal{Q}_{G_n|\overline{G}_M}\right) = \mathbb{E}_{\mathcal{P}_{G_n|\overline{G}_M}}\left[|L-1|\right] \le \sqrt{\mathrm{Var}_{\mathcal{P}_{G_n|\overline{G}_M}}\left[L\right]}$$

Step 3: Bound the second moment

- Define likelihood ratio $L \triangleq \frac{\mathcal{Q}_{G_n|\overline{G}_M}}{\mathcal{P}_{G_n|\overline{G}_M}}$. Then

$$2\mathrm{TV}\left(\mathcal{P}_{G_n|\overline{G}_M}, \mathcal{Q}_{G_n|\overline{G}_M}\right) = \mathbb{E}_{\mathcal{P}_{G_n|\overline{G}_M}}\left[|L-1|\right] \le \sqrt{\mathrm{Var}_{\mathcal{P}_{G_n|\overline{G}_M}}[L]}$$

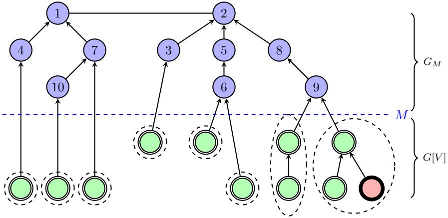
- Enough to show

$$\mathrm{Var}_{\mathcal{P}_{G_n|\overline{G}_M}}[L] = O(1/N),$$

where recall $M = n - N$ and $\Delta^2 \ll N \ll n$

Step 3: Bound the second moment

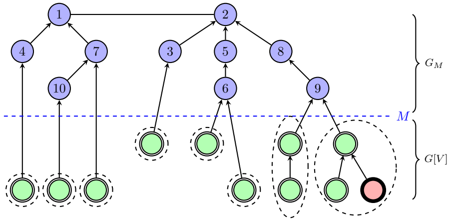
Let V denote the set of vertices arriving after time $M = n - N$. Consider the subgraph of G_n induced by V and let $\mathcal{C}(v)$ denote its connected component containing $v \in V$.



$m = 1$: connected components are denoted by dashed ellipses

Step 3: Bound the second moment

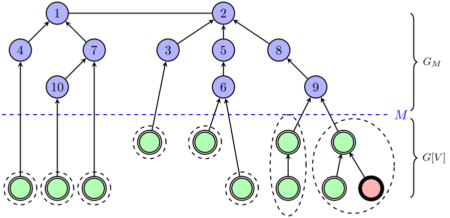
Let V denote the set of vertices arriving after time $M = n - N$. Consider the subgraph of G_n induced by V and let $\mathcal{C}(v)$ denote its connected component containing $v \in V$.



$m = 1$: connected components are denoted by dashed ellipses

Key: The connected components can arrive in any relative order

Let V denote the set of vertices arriving after time $M = n - N$. Consider the subgraph of G_n induced by V and let $\mathcal{C}(v)$ denote its connected component containing $v \in V$.



$m = 1$: connected components are denoted by dashed ellipses

Then

$$L \triangleq \frac{\mathcal{Q}_{G_n|\overline{G}_M}}{\mathcal{P}_{G_n|\overline{G}_M}} = \frac{C_1}{N} \sum_{v \in V} |\mathcal{C}(v)| \, \lambda_v X_v,$$

where C_1 is bounded constant, $\sum_{w \in \mathcal{C}(v)} \lambda_w = 1$, and $c_1 \leq X_v \leq c_2$.

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\le n, 1\le i\le m}$

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$

  e.g. for $m = 1$ and $\delta = 0$, recall at every time t,

$$\mathbb{P}\{t \to v\} \propto \mathsf{deg}(v)$$

  Equivalently, v is chosen by first sampling from all existing edges and then picking one of its two endpoints, uniformly at random

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\le n,1\le i\le m}$

  e.g. for $m=1$ and $\delta=0$, recall at every time t,

$$\mathbb{P}\{t \to v\} \propto \text{deg}(v)$$

  Equivalently, v is chosen by first sampling from all existing edges and then picking one of its two endpoints, uniformly at random $\Rightarrow \mathcal{P}_{G_n|\overline{G}_M}$ can be encoded by N independent uniform random variables supported over $[2(M-1)], [2M], \ldots, [2(n-2)]$, respectively

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$

  e.g. for $m=1$ and $\delta=0$, recall at every time t,

  $$\mathbb{P}\{t \to v\} \propto \text{deg}(v)$$

  Equivalently, v is chosen by first sampling from all existing edges and then picking one of its two endpoints, uniformly at random $\Rightarrow \mathcal{P}_{G_n|\overline{G}_M}$ can be encoded by N independent uniform random variables supported over $[2(M-1)], [2M], \ldots, [2(n-2)]$, respectively

  Similar encoding scheme extends to general $m \geq 1$ and $\delta > -m$

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n,1\leq i\leq m}$

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$
- Let $U = (U_{M+1,1}, \ldots, U_{t,i}, \ldots, U_{n,m})$ and $U^{(t,i)} = (U_{M+1,1}, \ldots, U'_{t,i}, \ldots, U_{n,m})$, where $U'_{t,i}$ is an independent copy of $U_{t,i}$. Write LRT L as $f(U)$ and apply Efron-Stein

$$\mathrm{Var}[L] \leq \frac{1}{2} \sum_{M<t\leq n} \sum_{1\leq i\leq m} \mathbb{E}\left[\left(f(U) - f(U^{(t,i)})\right)^2\right]$$

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$
- Let $U = (U_{M+1,1}, \ldots, U_{t,i}, \ldots, U_{n,m})$ and $U^{(t,i)} = (U_{M+1,1}, \ldots, U'_{t,i}, \ldots, U_{n,m})$, where $U'_{t,i}$ is an independent copy of $U_{t,i}$. Write LRT L as $f(U)$ and apply Efron-Stein

$$\mathrm{Var}[L] \leq \frac{1}{2} \sum_{M<t\leq n} \sum_{1\leq i\leq m} \mathbb{E}\left[\left(f(U) - f(U^{(t,i)})\right)^2\right]$$

- Our encoding scheme ensures that resampling $U_{t,i}$ can only affect $\mathcal{C}(t)$ (the component containing vertex arrived at time t), so

$$\left|f(U) - f(U^{(t,i)})\right| \leq O\left(\frac{|\mathcal{C}(t)| + |\mathcal{C}'(t)|}{N}\right).$$

Step 4: Efron-Stein inequality and coupling

- Encode $\mathcal{P}_{G_n|\overline{G}_M}$ using Nm ind. r.v.s $\{U_{t,i}\}_{M<t\leq n, 1\leq i\leq m}$
- Let $U = (U_{M+1,1}, \ldots, U_{t,i}, \ldots, U_{n,m})$ and $U^{(t,i)} = (U_{M+1,1}, \ldots, U'_{t,i}, \ldots, U_{n,m})$, where $U'_{t,i}$ is an independent copy of $U_{t,i}$. Write LRT L as $f(U)$ and apply Efron-Stein

$$\mathrm{Var}[L] \leq \frac{1}{2} \sum_{M<t\leq n} \sum_{1\leq i\leq m} \mathbb{E}\left[\left(f(U) - f(U^{(t,i)})\right)^2\right]$$

- Our encoding scheme ensures that resampling $U_{t,i}$ can only affect $\mathcal{C}(t)$ (the component containing vertex arrived at time t), so

$$\left|f(U) - f(U^{(t,i)})\right| \leq O\left(\frac{|\mathcal{C}(t)| + |\mathcal{C}'(t)|}{N}\right).$$

- Show the growth of $\mathcal{C}(t)$ is dominated by a sub-critical branching process to conclude $\mathbb{E}[|\mathcal{C}(t)|^2] = O(1)$