From trees to seeds:

on the inference of the seed from large random trees

Joint work with
Sébastien Bubeck, Ronen Eldan, and Elchanan Mossel

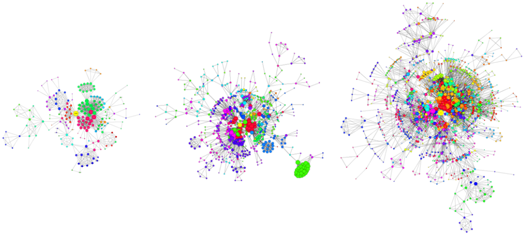
Miklós Z. Rácz

Microsoft Research

Banff Retreat
September 25, 2016.

Statistical inference in non-equilibrium networks

Apple's inventor network over a 6-year period. Source: Kenedict.



2007-2008 | 2009-2010 | 2011-2012

Given the current state of a network, what can we say about a previous state?

PNAS

Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network

Manuel Middendorf[†], Etay Ziv[‡], and Chris H. Wiggins[§¶‖]

[†]Department of Physics, [‡]College of Physicians and Surgeons, [§]Department of Applied Physics and Applied Mathematics, and [¶]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10027

Communicated by Barry H. Honig, Columbia University, New York, NY, December 20, 2004 (received for review September 7, 2004)

Naturally occurring networks exhibit quantitative features revealing underlying growth mechanisms. Numerous network mecha- method assesses significance of given subgraphs relative to an assumed null model, generated by Monte Carlo sampling of

PNAS

Recovering time-varying networks of dependencies in social and biological studies

Amr Ahmed and Eric P. Xing[1]

Language Technology Institute and Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved April 29, 2009 (received for review February 23, 2009)

A plausible representation of the relational information among entities in dynamic systems such as a living cell or a social community is a stochastic network that is topologically rewiring and semantically evolving over time. Although there is a rich literature cryptic liaisons and communities in society. A key technical hurdle that prevents us from an in-depth investigation of the mechanisms underlying this phenomena is the unavailability of serial snapshots of the rewiring network during the unfolding and progression of the

OPEN ACCESS Freely available online — PLoS COMPUTATIONAL BIOLOGY

Not All Scale-Free Networks Are Born Equal: The Role of the Seed Graph in PPI Network Evolution

Fereydoun Hormozdiari[1], Petra Berenbrink[1], Nataša Pržulj[2], S. Cenk Sahinalp[1*]

1 School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada, 2 Department of Computer Science, University of California Irvine, California, United States of America

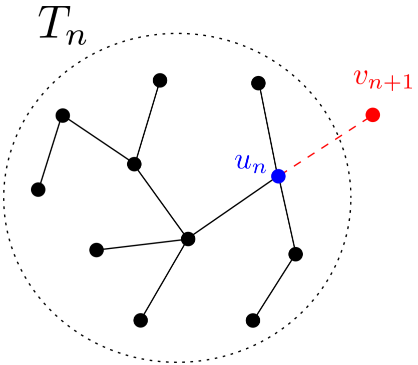
OPEN ACCESS Freely available online — PLoS COMPUTATIONAL BIOLOGY

Network Archaeology: Uncovering Ancient Networks from Present-Day Interactions

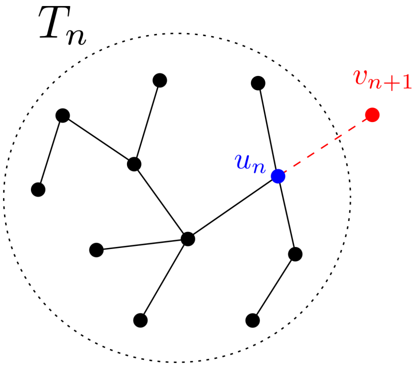
Saket Navlakha, Carl Kingsford*

Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, United States of America

Randomly growing trees



Randomly growing trees



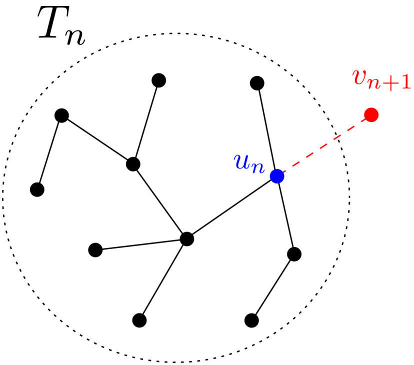
Preferential attachment:

$$\mathbb{P}\left(u_n = u\right) = \frac{d_{T_n}(u)}{2n-2}$$

Uniform attachment:

$$\mathbb{P}\left(u_n = u\right) = \frac{1}{n}$$

Randomly growing trees



Preferential attachment:

$$\mathbb{P}(u_n = u) = \frac{d_{T_n}(u)}{2n-2}$$

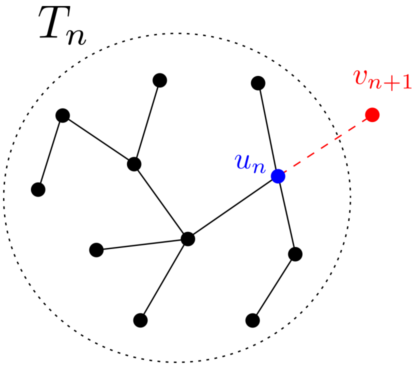
Uniform attachment:

$$\mathbb{P}(u_n = u) = \frac{1}{n}$$

In general:

$$\mathbb{P}(u_n = u) = \frac{(d_{T_n}(u))^\alpha}{Z}$$

Randomly growing trees



Preferential attachment:

$$\mathbb{P}(u_n = u) = \frac{d_{T_n}(u)}{2n-2}$$

Uniform attachment:

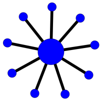
$$\mathbb{P}(u_n = u) = \frac{1}{n}$$

In general:

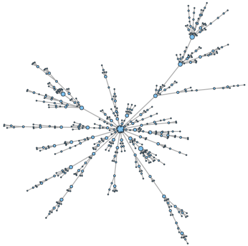
$$\mathbb{P}(u_n = u) = \frac{(d_{T_n}(u))^{\alpha}}{Z}$$

Many other tree growth models...

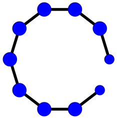
The influence of the seed — preferential attachment



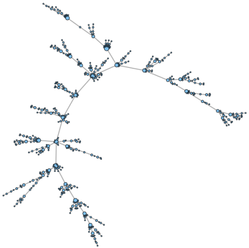
seed S_{10}



PA ($n = 500$, S_{10})

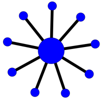


seed P_{10}

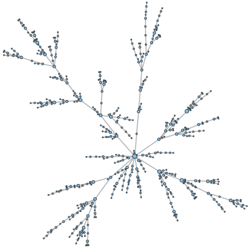


PA ($n = 500$, P_{10})

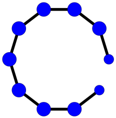
The influence of the seed — uniform attachment



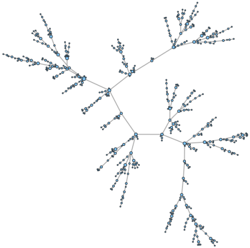
seed S_{10}



UA $(n = 500, S_{10})$



seed P_{10}



UA $(n = 500, P_{10})$

Measuring the influence of the seed

- A crude measure: limit as a countably infinite tree

Measuring the influence of the seed

- A crude measure: limit as a countably infinite tree

  $\rightsquigarrow$ seed has no influence for PA or UA

Measuring the influence of the seed

- A crude measure: limit as a countably infinite tree

  $\rightsquigarrow$ seed has no influence for PA or UA

But for superlinear attachment ($\alpha > 1$), see Oliveira, Spencer (2005)

Measuring the influence of the seed

- A crude measure: limit as a countably infinite tree

  $\rightsquigarrow$ seed has no influence for PA or UA

  But for superlinear attachment ($\alpha > 1$), see Oliveira, Spencer (2005)

- A finer measure: weak local limit (Benjamini-Schramm)

Measuring the influence of the seed

- A crude measure: limit as a countably infinite tree

  $\rightsquigarrow$ seed has no influence for PA or UA

  But for superlinear attachment ($\alpha > 1$), see Oliveira, Spencer (2005)

- A finer measure: weak local limit (Benjamini-Schramm)

  $\rightsquigarrow$ seed has no influence for PA or UA

Measuring the influence of the seed

- A crude measure: limit as a countably infinite tree

  $\rightsquigarrow$ seed has no influence for PA or UA

  But for superlinear attachment ($\alpha > 1$),
  see Oliveira, Spencer (2005)

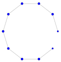
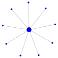
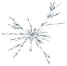
- A finer measure: weak local limit (Benjamini-Schramm)

  $\rightsquigarrow$ seed has no influence for PA or UA

  See Rudas, Tóth, Valkó (2007) (PA trees)
  and Berger, Borgs, Chayes, Saberi (2014) (in general)
  for weak local limits.

- A much finer measure: total variation distance

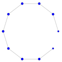
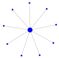
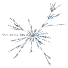
$$\delta_{\mathrm{PA}}(S, T) := \lim_{n \to \infty} \mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T))$$

$\delta_{\mathrm{PA}}($  $,$  $) = \lim_{n \to \infty} \mathrm{TV}($  $,$  $)$

Measuring the influence of the seed

- A much finer measure: total variation distance

$$\delta_{\mathrm{PA}}(S, T) := \lim_{n\to\infty} \mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T))$$

$\delta_{\mathrm{PA}}($  ,  $) = \lim_{n\to\infty} \mathrm{TV}($  ,  $)$

Hypothesis testing question:

$$H_0 : R \sim \mathrm{PA}(n, S), \qquad H_1 : R \sim \mathrm{PA}(n, T)$$

Q: test with asymptotically (in n) non-negligible power?

Main results

Preferential attachment:

Theorem (Bubeck, Mossel, R.)

If the degree profiles of S and T are different, and both have at least 3 vertices, then

$$\delta_{\mathrm{PA}}(S, T) > 0.$$

Theorem (Curien, Duquesne, Kortchemski, Manolescu)

If S and T are non-isomorphic and both have at least 3 vertices, then

$$\delta_{\mathrm{PA}}(S, T) > 0.$$

Uniform attachment:

Theorem (Bubeck, Eldan, Mossel, R.)

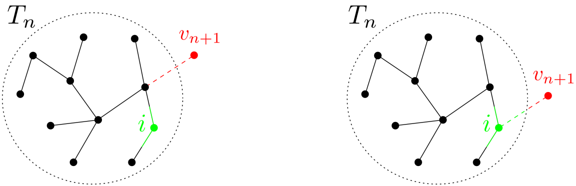
If S and T are non-isomorphic and both have at least 3 vertices, then

$$\delta_{\mathrm{UA}}(S, T) > 0.$$

PA heuristics: maximum degree

Degree evolution governed by Pólya urns

$$\left(2n - 2 - d_{\mathrm{PA}(n,S)}(i), d_{\mathrm{PA}(n,S)}(i)\right)$$

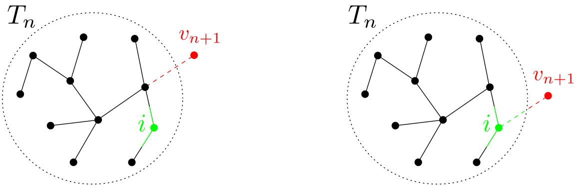


- Replacement matrix: $\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$; initial condition:
  - If $i \in S$ then $(2|S| - 2 - d_S(i), d_S(i))$;
  - If $i \notin S$ then $(2i - 3, 1)$.

PA heuristics: maximum degree

Degree evolution governed by Pólya urns

$$\left(2n-2-d_{\mathrm{PA}(n,S)}(i), d_{\mathrm{PA}(n,S)}(i)\right)$$



- Replacement matrix: $\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$; initial condition:
  - If $i \in S$ then $(2|S|-2-d_S(i), d_S(i))$;
  - If $i \notin S$ then $(2i-3, 1)$.

Rescaled degrees converge almost surely:

$$d_{\mathrm{PA}(n,S)}(i)/\sqrt{n} \xrightarrow{n\to\infty} D_i(S)$$

$$\Delta(\mathrm{PA}(n,S))/\sqrt{n} \xrightarrow{n\to\infty} D_{\max}(S)$$

$$D_{\max}(S) = \max_{i\geq 1} D_i(S)$$

See Móri (2005), Janson (2006), Peköz, Röllin, Ross (2013, 2014).

Influence of the seed on the maximum degree

Lemma (Tail behavior of the maximum degree)

Let S be a finite tree and let $m := |\{i \in \{1, \ldots, |S|\} : d_S(i) = \Delta(S)\}|$. Then

$$\mathbb{P}(D_{\max}(S) > t) \sim m \times c(|S|, \Delta(S))\, t^{1-2|S|+2\Delta(S)} \exp\left(-t^2/4\right)$$

as $t \to \infty$, where the constant c is explicit.

Influence of the seed on the maximum degree

Lemma (Tail behavior of the maximum degree)

Let S be a finite tree and let $m := |\{i \in \{1, \dots, |S|\} : d_S(i) = \Delta(S)\}|$. Then

$$\mathbb{P}(D_{\max}(S) > t) \sim m \times c(|S|, \Delta(S))\, t^{1-2|S|+2\Delta(S)} \exp(-t^2/4)$$

as $t \to \infty$, where the constant c is explicit.

$\rightsquigarrow$ the seed influences the polynomial factor!

Influence of the seed on the maximum degree

Lemma (Tail behavior of the maximum degree)

Let S be a finite tree and let $m := |\{i \in \{1, \ldots, |S|\} : d_S(i) = \Delta(S)\}|$. Then

$$\mathbb{P}(D_{\max}(S) > t) \sim m \times c(|S|, \Delta(S))\, t^{1-2|S|+2\Delta(S)} \exp\left(-t^2/4\right)$$

as $t \to \infty$, where the constant c is explicit.

$\rightsquigarrow$ the seed influences the polynomial factor!

Corollary (Distinguishing seeds)

If $|S| - \Delta(S) \neq |T| - \Delta(T)$, then

$$\delta_{\mathrm{PA}}(S, T) > 0.$$

Influence of the seed on the maximum degree

Lemma (Tail behavior of the maximum degree)

Let S be a finite tree and let $m := |\{i \in \{1, \dots, |S|\} : d_S(i) = \Delta(S)\}|$. Then

$$\mathbb{P}(D_{\max}(S) > t) \sim m \times c(|S|, \Delta(S))\, t^{1-2|S|+2\Delta(S)} \exp\left(-t^2/4\right)$$

as $t \to \infty$, where the constant c is explicit.

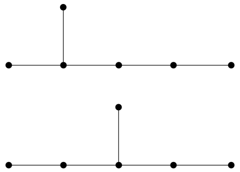
$\rightsquigarrow$ the seed influences the polynomial factor!

Corollary (Distinguishing seeds)

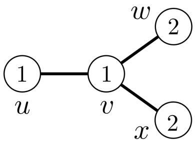
If $|S| - \Delta(S) \neq |T| - \Delta(T)$, then

$$\delta_{\mathrm{PA}}(S, T) > 0.$$

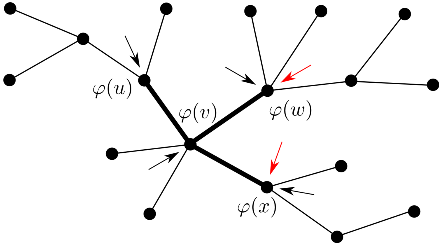
Two trees with the same degree profile:



The approach of Curien et al.



$$\underline{\tau} = (\tau, \ell)$$



$$D_{\underline{\tau}}(T) := \sum_{\varphi} \prod_{u \in \underline{\tau}} [d_T(\varphi(u))]_{\ell(u)}$$

Combinatorial interpretation: $D_{\underline{\tau}}(T) = \#$ decorated embeddings

Heuristic:

- large degree nodes contribute the most;
- captures geometric structure of large degree nodes.

The approach of Curien et al.

General framework:

- Construct a family of martingales using decorated embeddings:

$$M_{\underline{\tau}}^{(S)}(n) = \sum_{\underline{\tau}' \preccurlyeq \underline{\tau}} c_n(\underline{\tau}, \underline{\tau}') D_{\underline{\tau}'}(\mathrm{PA}(n, S)).$$

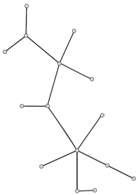
- For any S and T, there exists $\underline{\tau}$ and n such that

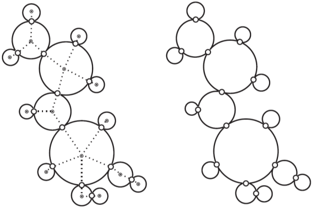
$$\mathbb{E}\left[M_{\underline{\tau}}^{(S)}(n)\right] \neq \mathbb{E}\left[M_{\underline{\tau}}^{(T)}(n)\right].$$

- Prove that the martingales are bounded in L^2.
- Conclude using the Cauchy-Schwarz inequality that

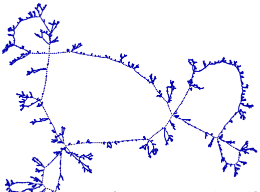
$$\delta_{\mathrm{PA}}(S, T) > 0.$$

The Brownian looptree





[Curien et al. (2014)]



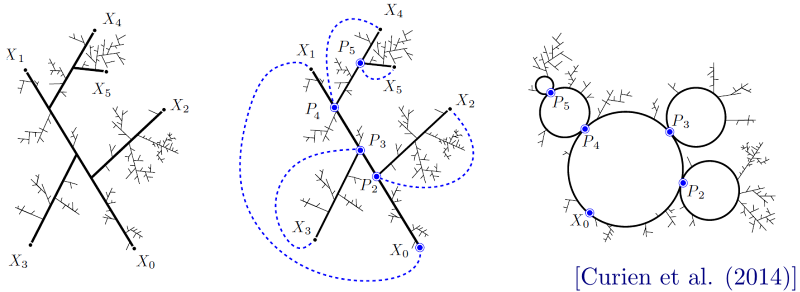
[Curien et al. (2014)]

Theorem (Curien, Duquesne, Kortchemski, Manolescu)

For any S there exists a random compact metric space $\mathcal{L}^{(\mathcal{S})}$ such that the following convergence holds a.s. in the Gromov-Hausdorff topology:

$$n^{-1/2} \cdot \mathrm{Loop}\left(\mathrm{PA}\left(n, S\right)\right) \xrightarrow{n \to \infty} 2\sqrt{2} \cdot \mathcal{L}^{(S)}.$$

The Brownian looptree



[Curien et al. (2014)]

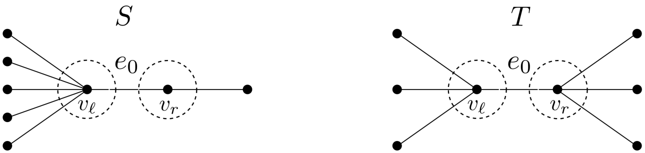
The metric space $\mathcal{L}$ is constructed as a quotient of Aldous's Brownian Continuum Random Tree.

Conjecture (Curien, Duquesne, Kortchemski, Manolescu)

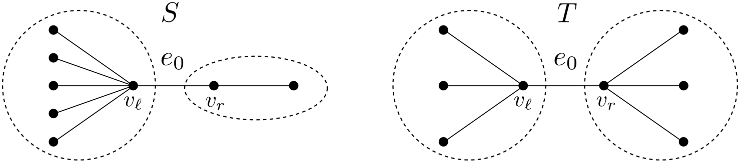
For any pair of seeds S and T,

$$\delta_{\mathrm{PA}}(S, T) = \mathrm{TV}\left(\mathcal{L}^{(S)}, \mathcal{L}^{(T)}\right).$$

Uniform attachment

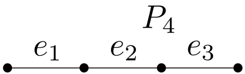


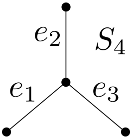
Preferential attachment: the degrees of v_ℓ and v_r are unbalanced in S but balanced in T, and this likely remains so throughout the process.



Uniform attachment: the subtree sizes under v_ℓ and v_r are unbalanced in S but balanced in T, and this likely remains so throughout the process.

An example: distinguishing P_4 and S_4

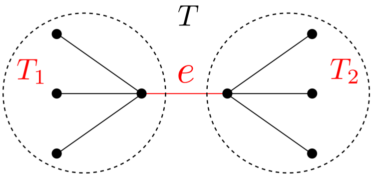




Measuring balancedness:

$$g(T,e) := \frac{|T_1|^2 \, |T_2|^2}{|T|^4}$$

$$G(T) := \sum_e g(T,e)$$



In order to show that $\delta_{\mathrm{UA}}(P_4, S_4) > 0$, it suffices to show that

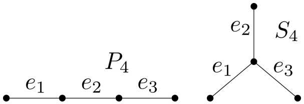
$$\liminf_{n\to\infty} \left|\mathbb{E}\left[G\left(\mathrm{UA}(n,P)\right)\right] - \mathbb{E}\left[G\left(\mathrm{UA}(n,S)\right)\right]\right| > 0$$

$$\limsup_{n\to\infty} \left(\mathrm{Var}[G(\mathrm{UA}(n,P))] + \mathrm{Var}[G(\mathrm{UA}(n,S))]\right) < \infty$$

An example: distinguishing P_4 and S_4

Let $\{e_j^P\}$ and $\{e_j^S\}$ denote the edges.
For every $j \geq 4$:

$$g\left(\mathrm{UA}\left(n, P\right), e_j^P\right) \stackrel{d}{=} g\left(\mathrm{UA}\left(n, S\right), e_j^S\right).$$



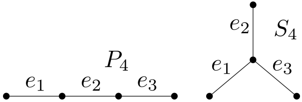
We also have this for $j = 1$ and $j = 3$, so

$$\mathbb{E}\left[G\left(\mathrm{UA}\left(n, P\right)\right)\right] - \mathbb{E}\left[G\left(\mathrm{UA}\left(n, S\right)\right)\right] = \mathbb{E}[g(\mathrm{UA}(n, P), e_2^P)] - \mathbb{E}[g(\mathrm{UA}(n, S), e_2^S)]$$
$$= \frac{2n^3 + 5n^2 + 8n + 5}{140n^3} \to \frac{1}{70}.$$

An example: distinguishing P_4 and S_4

Let $\{e_j^P\}$ and $\{e_j^S\}$ denote the edges.
For every $j \geq 4$:

$$g\left(\mathrm{UA}\,(n, P)\,, e_j^P\right) \stackrel{d}{=} g\left(\mathrm{UA}\,(n, S)\,, e_j^S\right).$$



We also have this for $j = 1$ and $j = 3$, so

$$\begin{aligned}\mathbb{E}\left[G\left(\mathrm{UA}\,(n, P)\right)\right] - \mathbb{E}\left[G\left(\mathrm{UA}\,(n, S)\right)\right] &= \mathbb{E}[g(\mathrm{UA}(n, P), e_2^P)] - \mathbb{E}[g(\mathrm{UA}(n, S), e_2^S)] \\ &= \frac{2n^3 + 5n^2 + 8n + 5}{140n^3} \to \frac{1}{70}.\end{aligned}$$

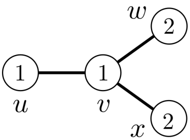
For the variance we use Cauchy-Schwarz:

$$\mathrm{Var}[G(\mathrm{UA}\,(n, S))] \leq \left(\sum_{j=1}^{n-1} \sqrt{\mathrm{Var}[g(\mathrm{UA}\,(n, S)\,, e_j)]}\right)^2,$$

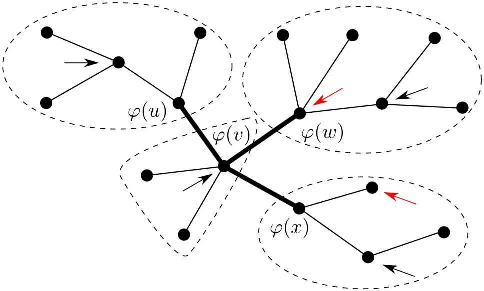
and estimates on moments of the beta-binomial distribution to give

$$\mathbb{E}[g(\mathrm{UA}\,(n, S)\,, e_j)^2] \leq C/j^4.$$

General statistics



$$\underline{\tau} = (\tau, \ell)$$



$$F_{\underline{\tau}}(T) := \sum_{\varphi} \prod_{u \in \underline{\tau}} \left[f_{\varphi(u)}(T) \right]_{\ell(u)}$$

Combinatorial interpretation: $F_{\underline{\tau}}(T) =$ # decorated embeddings

Heuristic:

- embeddings that are "central" contribute the most;
- captures global balancedness properties of the tree.

General framework

- Construct a family of martingales using decorated embeddings:

$$M_{\underline{\tau}}^{(S)}(n) = \sum_{\underline{\tau}' \preccurlyeq \underline{\tau}} c_n\left(\underline{\tau}, \underline{\tau}'\right) F_{\underline{\tau}'}\left(\mathrm{UA}\left(n, S\right)\right).$$

- For any S and T, there exists $\underline{\tau}$ and n such that

$$\mathbb{E}\left[M_{\underline{\tau}}^{(S)}(n)\right] \neq \mathbb{E}\left[M_{\underline{\tau}}^{(T)}(n)\right].$$

- Prove that the martingales are bounded in L^2.
- Conclude using the Cauchy-Schwarz inequality that

$$\delta_{\mathrm{UA}}(S, T) > 0.$$

Main technical issue: second moment

Lemma (First moment)

Let $\underline{\tau} \in \mathcal{D}_+$ be a decorated tree with positive labels and $|\underline{\tau}| \geq 2$, and let S be a seed tree. Then

$$n^{w(\underline{\tau})} \gtrsim \mathbb{E}\left[F_{\underline{\tau}}\left(\mathrm{UA}\left(n, S\right)\right)\right] \gtrsim n^{w(\underline{\tau})},$$

where $w\left(\underline{\tau}\right) = \sum_{u \in \tau} \ell\left(u\right)$.

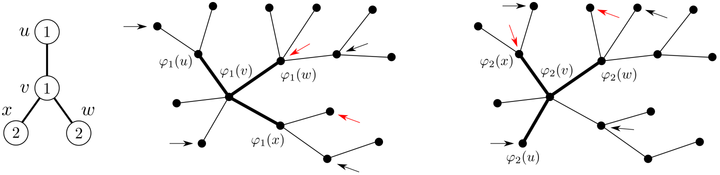
Lemma (Second moment)

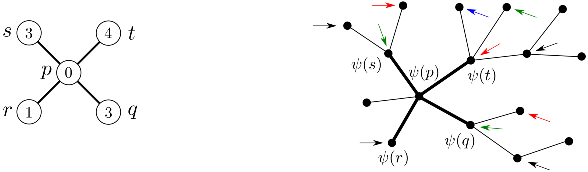
Let $\underline{\tau} \in \mathcal{D}_+$ be a decorated tree with positive labels and $|\underline{\tau}| \geq 2$, and let S be a seed tree. Then

(a) $$\mathbb{E}\left[F_{\underline{\tau}}\left(\mathrm{UA}\left(n, S\right)\right)^2\right] \gtrsim n^{2w(\underline{\tau})},$$

(b) $$\mathbb{E}\left[\left(F_{\underline{\tau}}\left(\mathrm{UA}\left(n+1, S\right)\right) - F_{\underline{\tau}}\left(\mathrm{UA}\left(n, S\right)\right)\right)^2\right] \gtrsim n^{2w(\underline{\tau})-2}.$$

Main technical issue: second moment



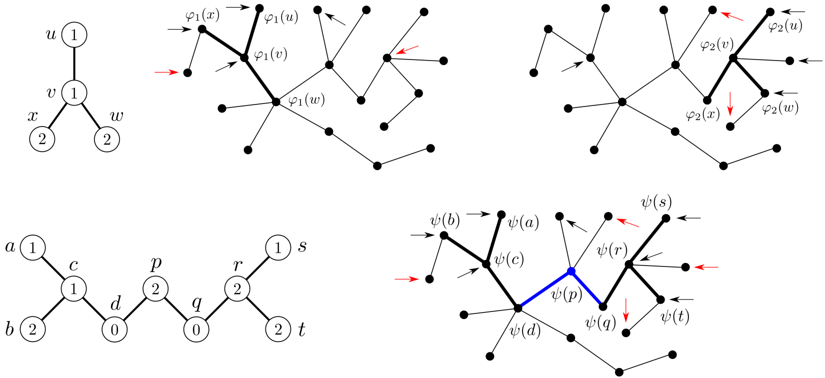


Top row: a decorated tree $\underline{\tau}$ and two decorated embeddings, $\underline{\varphi}_1$ and $\underline{\varphi}_2$, of it into a larger tree T.

Bottom row: an associated decorated tree $\underline{\sigma}$ and the decorated embedding $\underline{\psi}$ of it into T.

Note: $w(\underline{\sigma}) \leq 2w(\underline{\tau})$.

Main technical issue: second moment



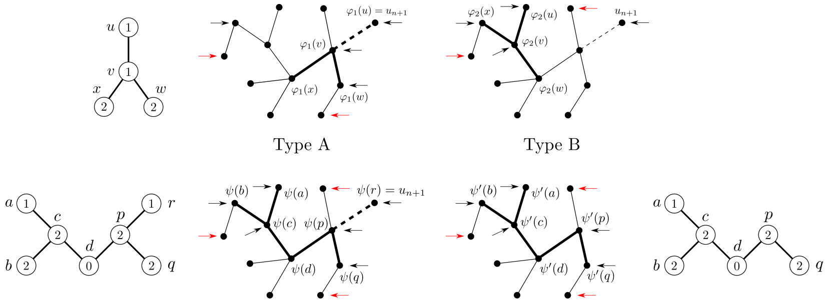
Top row: a decorated tree $\underline{\tau}$ and two decorated embeddings, $\underline{\varphi}_1$ and $\underline{\varphi}_2$, of it into a larger tree T.

Bottom row: an associated decorated tree $\underline{\sigma}$ and the decorated embedding $\underline{\psi}$ of it into T.

Note: $w(\underline{\sigma}) \leq 2w(\underline{\tau})$, but no a priori bound on $|\underline{\sigma}|$.

$\rightsquigarrow$ use the fact that $\mathrm{diam}(\mathrm{UA}(n, S)) = O(\log n)$ whp.

Main technical issue: second moment



Top row: There are two types of decorated embeddings that use the new vertex.

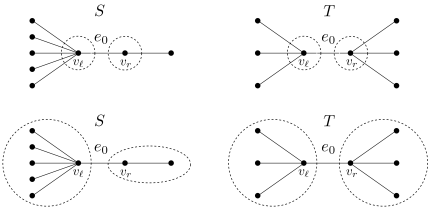
Bottom row: associated decorated trees and decorated embeddings.

Roughly speaking, the two arrows associated with the new vertex give the extra factor of n^{-2} required in the bound of (b).

Summary and open questions

Takeaways:

- Every seed has an influence, both in PA and in UA
- Degrees (PA) and balancedness (UA) are key statistics

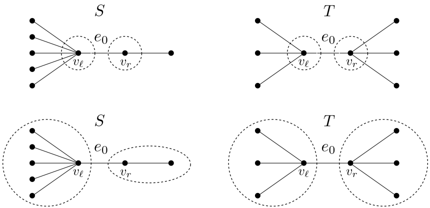


Open questions:

- Multiple edges added at each time step?
- Is $\delta_\alpha(S, T) > 0$ for $\alpha \in (0, 1)$? Is it monotone in α? Is it convex?
- Other models of randomly growing graphs.
- Estimation. Finding the seed. (Bubeck, Devroye, Lugosi)
- Hiding the seed. Rumor source obfuscation.

Summary and open questions

Takeaways:



- Every seed has an influence, both in PA and in UA
- Degrees (PA) and balancedness (UA) are key statistics

Open questions:

- Multiple edges added at each time step?
- Is $\delta_\alpha(S, T) > 0$ for $\alpha \in (0, 1)$?
  Is it monotone in α? Is it convex?
- Other models of randomly growing graphs.
- Estimation. Finding the seed. (Bubeck, Devroye, Lugosi)
- Hiding the seed. Rumor source obfuscation.

Thank you!