Cosistency of Hill Estimators in a Linear Preferential Attachment Model

Tiandong Wang
Joint Work with **S.I. Resnick**

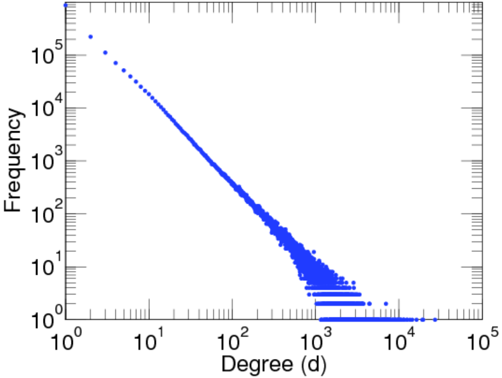
School of Operations Research and Information Engineering,
Cornell University

June 21st, 2018

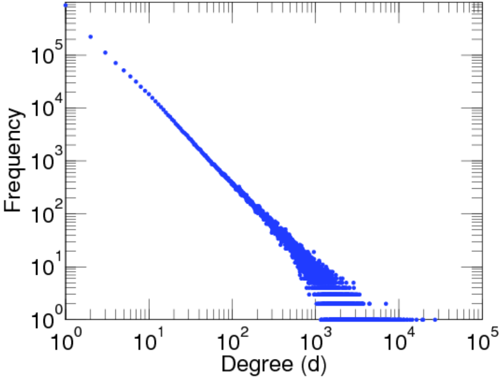
Flickr Links Data:

(cf. KONECT: http://konect.uni-koblenz.de/networks/flickr-links)

Undirected network of Flickr users and their connections.



Flickr Links Data:
(cf. KONECT: http://konect.uni-koblenz.de/networks/flickr-links)
Undirected network of Flickr users and their connections.

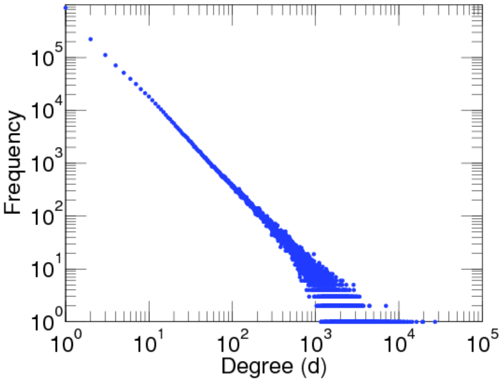


- Nodes: Users.
- Edges: Connections.

For a node v, $\mathbb{P}(\text{Degree}(v) = k) \sim k^{-1-\alpha}$, for k large.

Flickr Links Data:

(cf. KONECT: http://konect.uni-koblenz.de/networks/flickr-links)
Undirected network of Flickr users and their connections.



- Nodes: Users.
- Edges: Connections.

For a node v, $\mathbb{P}(\text{Degree}(v) = k) \sim k^{-1-\alpha}$, for k large.

Goal:
Estimate the power-law index α.

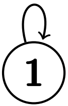
Square count	428,143,604,855
4-tour count	3,483,825,597,342
Power law exponent (estimated) with d_{min}	1.7310 (d_{min} = 8)
Gini coefficient	88.2%
Relative edge distribution entropy	82.2%
Assortativity	-0.015282

Undirected Preferential Attachment Model

Notations:

- $G(n) :=$ the random graph after n-steps.
- $[n] := \{1, 2, \ldots, n\}$, set of nodes in $G(n)$.
- $D_i(n) :=$ Degree of node $i \in [n]$.
- $\delta > -1$, parameter.

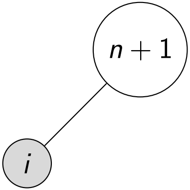
Initialize with a single node having a self loop.



This node is considered as having degree 2, i.e.

$$D_1(1) = 2.$$

From $G(n)$ to $G(n+1)$, assuming **linear** preferential attachment function: $f(i) = i + \delta$:



The new node $n+1$ attaches to node $i \in [n]$ with probability

$$\frac{D_i(n) + \delta}{(2+\delta)n},$$

and $D_{n+1}(n+1) = 1$.

Define $N_k(n) := \sum_{j=1}^{n} \mathbf{1}_{\{D_j(n)=k\}}$, then as $n \to \infty$,

$$N_k(n)/n \xrightarrow{a.s.} p_k \sim C(\delta) k^{-3-\delta} =: C(\delta) k^{-1-\alpha} \quad \text{for } k \to \infty.$$

Define $N_k(n) := \sum_{j=1}^{n} \mathbf{1}_{\{D_j(n)=k\}}$, then as $n \to \infty$,

$$N_k(n)/n \xrightarrow{a.s.} p_k \sim C(\delta)k^{-3-\delta} =: C(\delta)k^{-1-\alpha} \quad \text{for } k \to \infty.$$

How to estimate the power-law index α?

Option 1:

- Find MLE of δ, $\hat{\delta}^{MLE}$ (cf. Gao and van der Vaart (2017)).
- Plugging $\hat{\delta}^{MLE}$ into the theoretical value of α gives

$$\hat{\alpha}^{MLE} = 2 + \hat{\delta}^{MLE}.$$

Define $N_k(n) := \sum_{j=1}^{n} \mathbf{1}_{\{D_j(n)=k\}}$, then as $n \to \infty$,

$$N_k(n)/n \xrightarrow{a.s.} p_k \sim C(\delta)k^{-3-\delta} =: C(\delta)k^{-1-\alpha} \quad \text{for } k \to \infty.$$

How to estimate the power-law index α?

Option 1:

- Find MLE of δ, $\hat{\delta}^{MLE}$ (cf. Gao and van der Vaart (2017)).
- Plugging $\hat{\delta}^{MLE}$ into the theoretical value of α gives

$$\hat{\alpha}^{MLE} = 2 + \hat{\delta}^{MLE}.$$

However, the MLE approach is not **ROBUST** against modeling error, compared to the extreme value estimation approach (cf. Wan, Wang, Davis and Resnick (2017)).

Option 2: Hill estimator.

Let $X_{(1)} \geq \ldots \geq X_{(n)}$ be order statistics of $\{X_i : 1 \leq i \leq n\}$, then the Hill estimator $H_{k,n}$ based on k upper order statistics of $\{X_i : 1 \leq i \leq n\}$ is defined as (cf. Hill (1975))

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(1)}}{X_{(k+1)}}.$$

Option 2: Hill estimator.

Let $X_{(1)} \geq \ldots \geq X_{(n)}$ be order statistics of $\{X_i : 1 \leq i \leq n\}$, then the Hill estimator $H_{k,n}$ based on k upper order statistics of $\{X_i : 1 \leq i \leq n\}$ is defined as (cf. Hill (1975))

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(1)}}{X_{(k+1)}}.$$

ROBUST and widely used in practice, e.g. KONECT:

Option 2: Hill estimator.

Let $X_{(1)} \geq \ldots \geq X_{(n)}$ be order statistics of $\{X_i : 1 \leq i \leq n\}$, then the Hill estimator $H_{k,n}$ based on k upper order statistics of $\{X_i : 1 \leq i \leq n\}$ is defined as (cf. Hill (1975))

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(1)}}{X_{(k+1)}}.$$

ROBUST and widely used in practice, e.g. KONECT:

Power law exponent

The **power law exponent** is a number that characterizes the degrees of the nodes in the network. In many circumstances, networks are modeled to follow a degree distribution power law, i.e., the number of nodes with degree n is taken to be proportional to the power $n^{-\gamma}$, for a constant γ larger than one [1]. This constant γ is called the power law exponent. Given a network, its degree distribution can be used to estimate a value γ. There are multiple ways of estimating γ, and thus a network does not have a single definite value of it. In KONECT, we estimate γ using the robust method given in [2]

$$\gamma = 1 + n \left(\sum_{u \in V} \ln \frac{d(u)}{d_{\min}} \right)^{-1},$$

in which $d_{\min}$ is the minimal degree.

[1] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Phys.*, 46(5):323–351, 2006.

[2] Albert-LÃ¡szlÃ³ BarabÃ¡si and RÃ©ka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Consistency of Hill Estimators

Suppose that $\{X_i : 1 \leq i \leq n\}$ **iid** and non-negative with common regularly varying distribution tail $\overline{F} \in RV_{-\alpha}$, $\alpha > 0$, then:

- There exists a sequence $\{b(n)\}$ such that

$$\sum_{i=1}^{n} \epsilon_{X_i/b(n)} \Rightarrow PRM(\nu_\alpha) \qquad \text{in } M_p((0,\infty]),$$

  with $\nu_\alpha(y,\infty] = y^{-\alpha}$, $y > 0$.

- For some intermediate sequence $k_n \to \infty$, $k_n/n \to 0$ as $n \to \infty$:

$$\frac{1}{k_n}\sum_{i=1}^{n} \epsilon_{X_i/b(n/k_n)} \Rightarrow \nu_\alpha \qquad \text{in } M_+((0,\infty]).$$

- The Hill estimator is consistent:

$$H_{k_n,n} \xrightarrow{P} 1/\alpha.$$

Network data is **NOT** iid!!! Will $H_{k_n,n}$ still be consistent?

Network data is **NOT** iid!!! Will $H_{k_n,n}$ still be consistent?

Drawing analogies to the iid case, we want to show:

- The degree sequence has empirical measure

$$\sum_{i=1}^{n} \epsilon_{D_i(n)/n^{1/(2+\delta)}}$$

  converging weakly to some random limit point measure in $M_p((0,\infty])$.

Network data is **NOT** iid!!! Will $H_{k_n,n}$ still be consistent?

Drawing analogies to the iid case, we want to show:

- The degree sequence has empirical measure

$$\sum_{i=1}^{n} \epsilon_{D_i(n)/n^{1/(2+\delta)}}$$

  converging weakly to some random limit point measure in $M_p((0,\infty])$.

- For some intermediate sequence k_n and some function $b(\cdot)$:

$$\frac{1}{k_n}\sum_{i=1}^{n} \epsilon_{D_i(n)/b(n/k_n)} \Rightarrow \nu_{2+\delta}, \qquad \text{in } M_+((0,\infty]).$$

This would facilitate proving consistency of the Hill estimator.

Embedding

Idea: Embed the degree sequence $(D_1(n), \ldots, D_n(n), 0, \ldots)$ into a sequence of **birth immigration processes** (B.I. processes).

Embedding

Idea: Embed the degree sequence $(D_1(n), \ldots, D_n(n), 0, \ldots)$ into a sequence of **birth immigration processes** (B.I. processes).

Preliminaries:

- A **linear birth process** $\{\zeta(t) : t \geq 0\}$ is a continuous time Markov process taking values in the set $\mathbb{N}^+ = \{1, 2, \ldots\}$ and having a transition rate $q_{i,i+1} = \lambda i$, $i \in \mathbb{N}^+$, $\lambda > 0$.

Embedding

Idea: Embed the degree sequence $(D_1(n), \ldots, D_n(n), 0, \ldots)$ into a sequence of **birth immigration processes** (B.I. processes).

Preliminaries:

- A **linear birth process** $\{\zeta(t) : t \geq 0\}$ is a continuous time Markov process taking values in the set $\mathbb{N}^+ = \{1, 2, \ldots\}$ and having a transition rate $q_{i,i+1} = \lambda i$, $i \in \mathbb{N}^+$, $\lambda > 0$.
- The linear birth process $\{\zeta(t) : t \geq 0\}$ is a mixed Poisson process, i.e. with $\zeta(0) = 1$, we have

$$\zeta(t) = 1 + N_0\big(W(e^{\lambda t} - 1)\big),\ t \geq 0,$$

  where $\{N_0(t) : t \geq 0\}$ is a unit rate homogeneous Poisson on $\mathbb{R}_+$ with $N_0(0) = 0$ and W is a unit exponential random variable independent of N_0.

Embedding

Idea: Embed the degree sequence $(D_1(n), \ldots, D_n(n), 0, \ldots)$ into a sequence of **birth immigration processes** (B.I. processes).

Preliminaries:

- A **linear birth process** $\{\zeta(t) : t \geq 0\}$ is a continuous time Markov process taking values in the set $\mathbb{N}^+ = \{1, 2, \ldots\}$ and having a transition rate $q_{i,i+1} = \lambda i$, $i \in \mathbb{N}^+$, $\lambda > 0$.
- The linear birth process $\{\zeta(t) : t \geq 0\}$ is a mixed Poisson process, i.e. with $\zeta(0) = 1$, we have

$$\zeta(t) = 1 + N_0\big(W(e^{\lambda t} - 1)\big), \; t \geq 0,$$

  where $\{N_0(t) : t \geq 0\}$ is a unit rate homogeneous Poisson on $\mathbb{R}_+$ with $N_0(0) = 0$ and W is a unit exponential random variable independent of N_0.
- For $\zeta(0) = 1$, $e^{-\lambda t}\zeta(t) \xrightarrow{a.s.} W \sim \textbf{Exp}(1)$.

The linear birth process with *immigration* (**B.I. process**), $\{BI(t) : t \geq 0\}$, having lifetime parameter $\lambda > 0$ and immigration parameter $\theta \geq 0$ is a continuous time Markov process with state space $\mathbb{N} = \{0, 1, 2, \dots\}$ and transition rate $q_{i,i+1} = \lambda i + \theta$.

The linear birth process with *immigration* (**B.I. process**), $\{BI(t) : t \geq 0\}$, having lifetime parameter $\lambda > 0$ and immigration parameter $\theta \geq 0$ is a continuous time Markov process with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$ and transition rate $q_{i,i+1} = \lambda i + \theta$.

- Suppose that $N_\theta(t)$ is the counting function of homogeneous Poisson points $0 < \tau_1 < \tau_2 < \ldots$ with rate θ.
- Independent of $N_\theta(\cdot)$, we have independent copies of a linear birth process $\{\zeta_i(t) : t \geq 0\}_{i \geq 1}$ with parameter $\lambda > 0$ and $\zeta_i(0) = 1$ for $i \geq 1$.

The linear birth process with *immigration* (**B.I. process**), $\{BI(t) : t \geq 0\}$, having lifetime parameter $\lambda > 0$ and immigration parameter $\theta \geq 0$ is a continuous time Markov process with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$ and transition rate $q_{i,i+1} = \lambda i + \theta$.

- Suppose that $N_\theta(t)$ is the counting function of homogeneous Poisson points $0 < \tau_1 < \tau_2 < \ldots$ with rate θ.
- Independent of $N_\theta(\cdot)$, we have independent copies of a linear birth process $\{\zeta_i(t) : t \geq 0\}_{i \geq 1}$ with parameter $\lambda > 0$ and $\zeta_i(0) = 1$ for $i \geq 1$.
- Let $BI(0) = 0$, then the B.I. process is a shot noise process with form

$$BI(t) := \sum_{i=1}^{\infty} \zeta_i(t - \tau_i) \mathbf{1}_{\{t \geq \tau_i\}} = \sum_{i=1}^{N_\theta(t)} \zeta_i(t - \tau_i).$$

The linear birth process with *immigration* (**B.I. process**), $\{BI(t) : t \geq 0\}$, having lifetime parameter $\lambda > 0$ and immigration parameter $\theta \geq 0$ is a continuous time Markov process with state space $\mathbb{N} = \{0, 1, 2, \ldots\}$ and transition rate $q_{i,i+1} = \lambda i + \theta$.

- Suppose that $N_\theta(t)$ is the counting function of homogeneous Poisson points $0 < \tau_1 < \tau_2 < \ldots$ with rate θ.
- Independent of $N_\theta(\cdot)$, we have independent copies of a linear birth process $\{\zeta_i(t) : t \geq 0\}_{i \geq 1}$ with parameter $\lambda > 0$ and $\zeta_i(0) = 1$ for $i \geq 1$.
- Let $BI(0) = 0$, then the B.I. process is a shot noise process with form

$$BI(t) := \sum_{i=1}^{\infty} \zeta_i(t - \tau_i) \mathbf{1}_{\{t \geq \tau_i\}} = \sum_{i=1}^{N_\theta(t)} \zeta_i(t - \tau_i).$$

- For $BI(0) = k \geq 0$,

$$e^{-\lambda t} BI(t) \xrightarrow{a.s.} \sigma \sim \mathbf{Gamma}(k + \theta/\lambda, 1).$$

B.I. Process Setup Let $\{BI_i(t) : t \geq 0\}_{i \geq 1}$ be independent B.I. processes such that

$$BI_1(0) = 2, \quad BI_i(0) = 1, \quad \forall i \geq 2.$$

Each has transition rate is $q_{j,j+1} = j + \delta$, $\delta > -1$.

B.I. Process Setup Let $\{BI_i(t) : t \geq 0\}_{i \geq 1}$ be independent B.I. processes such that

$$BI_1(0) = 2, \quad BI_i(0) = 1, \quad \forall i \geq 2.$$

Each has transition rate is $q_{j,j+1} = j + \delta$, $\delta > -1$.

- Set $T_1 = 0$ and relative to $BI_1(\cdot)$ define T_2 be the first time that $BI_1(\cdot)$ jumps.
- Start the new B.I. process $\{BI_2(t - T_2) : t \geq T_2\}$ at T_2.

B.I. Process Setup Let $\{BI_i(t) : t \geq 0\}_{i \geq 1}$ be independent B.I. processes such that

$$BI_1(0) = 2, \quad BI_i(0) = 1, \quad \forall i \geq 2.$$

Each has transition rate is $q_{j,j+1} = j + \delta$, $\delta > -1$.

- Set $T_1 = 0$ and relative to $BI_1(\cdot)$ define T_2 be the first time that $BI_1(\cdot)$ jumps.
- Start the new B.I. process $\{BI_2(t - T_2) : t \geq T_2\}$ at T_2.
- Let T_3 be the first time after T_2 that either $BI_1(\cdot)$ or $BI_2(\cdot)$ jumps.
- Start a new, independent B.I. process $\{BI_3(t - T_3)\}_{t \geq T_3}$ at T_3.
- Continue in this way.

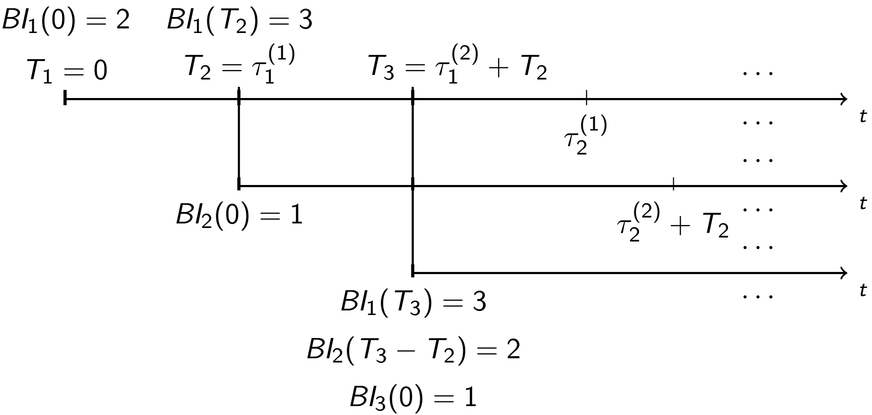


Figure 5.1: Embedding procedure for Model A assuming $\tau_1^{(2)} + T_2 < \tau_2^{(1)}$.

Embedding Results:

For each n, let $\mathbf{D}(n) := (D_1(n), D_2(n), \ldots, D_n(n), 0, \ldots)$ and $\widetilde{\mathbf{D}}(n) := (BI_1(T_n), BI_2(T_n - T_2), \ldots, BI_n(0), 0, \ldots)$. Then $\mathbf{D}(n)$ and $\widetilde{\mathbf{D}}(n)$ have the same distribution in $\mathbb{R}^\infty$.

Embedding Results:

For each n, let $\mathbf{D}(n) := (D_1(n), D_2(n), \ldots, D_n(n), 0, \ldots)$ and $\widetilde{\mathbf{D}}(n) := (BI_1(T_n), BI_2(T_n - T_2), \ldots, BI_n(0), 0, \ldots)$. Then $\mathbf{D}(n)$ and $\widetilde{\mathbf{D}}(n)$ have the same distribution in $\mathbb{R}^\infty$.

$$\text{Degree Sequence} \qquad \Rightarrow \qquad \text{B.I. Processes.}$$

Embedding Results:

For each n, let $\mathbf{D}(n) := (D_1(n), D_2(n), \ldots, D_n(n), 0, \ldots)$ and $\widetilde{\mathbf{D}}(n) := (BI_1(T_n), BI_2(T_n - T_2), \ldots, BI_n(0), 0, \ldots)$. Then $\mathbf{D}(n)$ and $\widetilde{\mathbf{D}}(n)$ have the same distribution in $\mathbb{R}^\infty$.

$$\text{Degree Sequence} \quad \Rightarrow \quad \text{B.I. Processes.}$$

Convergence of $\{T_n\}$:

The counting process $N(t) := \frac{1}{2}\sum_{i=1}^{\infty} BI_i(t - T_i)\mathbf{1}_{\{t \geq T_i\}}$ is a pure birth process with transition rate $q_{i,i+1} = (2+\delta)i$. Also,

$$\frac{n}{e^{(2+\delta)T_n}} \xrightarrow{\text{a.s.}} W,$$

where W is an exponential random variable with unit mean.

Convergence of the Degree for a Fixed Node:

(i) Suppose that $\{\sigma_i\}_{i\geq 1}$ is a sequence of independent Gamma random variables with

$$\sigma_1 \sim \mathrm{Gamma}(2+\delta, 1), \quad \text{and} \quad \sigma_i \sim \mathrm{Gamma}(1+\delta, 1), \quad i \geq 2,$$

then

$$\frac{D_i(n)}{n^{1/(2+\delta)}} \Rightarrow W^{-1/(2+\delta)} \sigma_i e^{-T_i}.$$

Convergence of the Degree for a Fixed Node:
(i) Suppose that $\{\sigma_i\}_{i\geq 1}$ is a sequence of independent Gamma random variables with

$$\sigma_1 \sim \text{Gamma}(2+\delta, 1), \quad \text{and} \quad \sigma_i \sim \text{Gamma}(1+\delta, 1), \quad i \geq 2,$$

then

$$\frac{D_i(n)}{n^{1/(2+\delta)}} \Rightarrow W^{-1/(2+\delta)} \sigma_i e^{-T_i}.$$

(ii) Set $D_i(n) := 0$ for all $i \geq n+1$. For $\delta > -1$,

$$\max_{i\geq 1} \frac{D_i(n)}{n^{1/(2+\delta)}} \Rightarrow W^{-1/(2+\delta)} \max_{i\geq 1} \sigma_i e^{-T_i},$$

Convergence of the Degree for a Fixed Node:
(i) Suppose that $\{\sigma_i\}_{i\geq 1}$ is a sequence of independent Gamma random variables with

$$\sigma_1 \sim \text{Gamma}(2+\delta, 1), \quad \text{and} \quad \sigma_i \sim \text{Gamma}(1+\delta, 1), \quad i \geq 2,$$

then

$$\frac{D_i(n)}{n^{1/(2+\delta)}} \Rightarrow W^{-1/(2+\delta)} \sigma_i e^{-T_i}.$$

(ii) Set $D_i(n) := 0$ for all $i \geq n+1$. For $\delta > -1$,

$$\max_{i\geq 1} \frac{D_i(n)}{n^{1/(2+\delta)}} \Rightarrow W^{-1/(2+\delta)} \max_{i\geq 1} \sigma_i e^{-T_i},$$

Convergence of the Empirical Measure:
In $M_p((0,\infty])$, we have for $\delta \geq 0$,

$$\sum_{i=1}^{n} \epsilon_{D_i(n)/n^{1/(2+\delta)}}(\cdot) \Rightarrow \sum_{i=1}^{\infty} \epsilon_{\sigma_i e^{-T_i}/W^{1/(2+\delta)}}(\cdot).$$

Consistency of Hill Estimators: Heuristics

From the limit measure:

- Gamma random variables σ_i have light tailed distributions.

Consistency of Hill Estimators: Heuristics

From the limit measure:

- Gamma random variables σ_i have light tailed distributions.
  $\Rightarrow \{\sigma_i : i \geq 1\}$ may not distort the consistency result.

Consistency of Hill Estimators: Heuristics

From the limit measure:

- Gamma random variables σ_i have light tailed distributions. $\Rightarrow \{\sigma_i : i \geq 1\}$ may not distort the consistency result.
- Set $Y_i := e^{-T_i}/W^{1/(2+\delta)}$ and apply the Hill estimator to the $Y's$:

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{Y_i}{Y_{k+1}}\right) = \frac{1}{k}\sum_{i=1}^{k} (T_{k+1} - T_i).$$

Consistency of Hill Estimators: Heuristics

From the limit measure:

- Gamma random variables σ_i have light tailed distributions.
  $\Rightarrow \{\sigma_i : i \geq 1\}$ may not distort the consistency result.
- Set $Y_i := e^{-T_i}/W^{1/(2+\delta)}$ and apply the Hill estimator to the $Y's$:

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{Y_i}{Y_{k+1}}\right) = \frac{1}{k}\sum_{i=1}^{k}(T_{k+1} - T_i).$$

- By the B.I. process construction, we have

$$T_{n+1} - T_n \stackrel{d}{=} E_n/(n(2+\delta)),$$

where $E_n, n \geq 1$ are iid unit exponential random variables.

Consistency of Hill Estimators: Heuristics

From the limit measure:

- Gamma random variables σ_i have light tailed distributions.
  $\Rightarrow \{\sigma_i : i \geq 1\}$ may not distort the consistency result.
- Set $Y_i := e^{-T_i}/W^{1/(2+\delta)}$ and apply the Hill estimator to the $Y's$:

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{Y_i}{Y_{k+1}}\right) = \frac{1}{k}\sum_{i=1}^{k}(T_{k+1} - T_i).$$

- By the B.I. process construction, we have

$$T_{n+1} - T_n \stackrel{d}{=} E_n/(n(2+\delta)),$$

  where $E_n, n \geq 1$ are iid unit exponential random variables.
- Provided that $k \to \infty$, we have

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\sum_{l=i}^{k}(T_{l+1} - T_l) = \frac{1}{k}\sum_{l=1}^{k}\frac{E_l}{2+\delta} \xrightarrow{a.s.} \frac{1}{2+\delta}.$$

For rigorous justifications we need:

For some function $b(\cdot)$ and some intermediate sequence $\{k_n\}$ wich $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$,

$$\frac{1}{k_n}\sum_{i=1}^{n} \epsilon_{D_i(n)/b(n/k_n)} \Rightarrow \nu_{2+\delta}, \quad \text{in } M_+((0,\infty]).$$

For rigorous justifications we need:

For some function $b(\cdot)$ and some intermediate sequence $\{k_n\}$ wich $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$,

$$\frac{1}{k_n} \sum_{i=1}^{n} \epsilon_{D_i(n)/b(n/k_n)} \Rightarrow \nu_{2+\delta}, \quad \text{in } M_+((0,\infty]).$$

Note that for any $y > 0$,

$$\frac{1}{k_n} \sum_{i=1}^{n} \epsilon_{D_i(n)/b(n/k_n)}(y,\infty] = \frac{1}{k_n} N_{>b(n/k_n)y}(n).$$

For rigorous justifications we need:

For some function $b(\cdot)$ and some intermediate sequence $\{k_n\}$ wich $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$,

$$\frac{1}{k_n}\sum_{i=1}^{n} \epsilon_{D_i(n)/b(n/k_n)} \Rightarrow \nu_{2+\delta}, \quad \text{in } M_+((0,\infty]).$$

Note that for any $y > 0$,

$$\frac{1}{k_n}\sum_{i=1}^{n} \epsilon_{D_i(n)/b(n/k_n)}(y,\infty] = \frac{1}{k_n} N_{>b(n/k_n)y}(n).$$

Hence, we need to control:

(i) Bias: $\left|N_{>b(n/k_n)y} - \mathbb{E}(N_{>b(n/k_n)y}(n))\right|$.

(ii) Concentration of $\mathbb{E}(N_{>b(n/k_n)y}(n))/n$ on $p_{>b(n/k_n)y}$: $\left|\mathbb{E}(N_{>b(n/k_n)y}(n)) - np_{>b(n/k_n)y}\right|$.

(iii) Difference between $\frac{n}{k_n} p_{>b(n/k_n)y}$ and $y^{-(2+\delta)}$.

We need the following: as $n \to \infty$,

(i) $\frac{1}{k_n}\left|N_{>b(n/k_n)y} - \mathbb{E}(N_{>b(n/k_n)y}(n))\right| \xrightarrow{P} 0.$

(ii) $\frac{1}{k_n}\left|\mathbb{E}(N_{>b(n/k_n)y}(n)) - np_{>b(n/k_n)y}\right| \longrightarrow 0.$

(iii) $\left|\frac{n}{k_n}p_{>b(n/k_n)y} - y^{-(2+\delta)}\right| \longrightarrow 0.$

The third part can be justified using Stirling's formula. We prove (i) and (ii) by establishing the following concentration results:

We need the following: as $n \to \infty$,

(i) $\frac{1}{k_n}\left|N_{>b(n/k_n)y} - \mathbb{E}(N_{>b(n/k_n)y}(n))\right| \xrightarrow{P} 0.$

(ii) $\frac{1}{k_n}\left|\mathbb{E}(N_{>b(n/k_n)y}(n)) - np_{>b(n/k_n)y}\right| \longrightarrow 0.$

(iii) $\left|\frac{n}{k_n}p_{>b(n/k_n)y} - y^{-(2+\delta)}\right| \longrightarrow 0.$

The third part can be justified using Stirling's formula. We prove (i) and (ii) by establishing the following concentration results:

Concentration of the Degree Sequence:

For $\delta > -1$ there exists a constant $C > 2\sqrt{2}$, such that as $n \to \infty$,

$$\mathbb{P}\left(\max_k |N_{>k}(n) - np_{>k}| \geq C(1 + \sqrt{n \log n})\right) = o(1).$$

We need the following: as $n \to \infty$,

(i) $\frac{1}{k_n}\left|N_{>b(n/k_n)y} - \mathbb{E}(N_{>b(n/k_n)y}(n))\right| \xrightarrow{P} 0.$

(ii) $\frac{1}{k_n}\left|\mathbb{E}(N_{>b(n/k_n)y}(n)) - np_{>b(n/k_n)y}\right| \longrightarrow 0.$

(iii) $\left|\frac{n}{k_n}p_{>b(n/k_n)y} - y^{-(2+\delta)}\right| \longrightarrow 0.$

The third part can be justified using Stirling's formula. We prove (i) and (ii) by establishing the following concentration results:

Concentration of the Degree Sequence:

For $\delta > -1$ there exists a constant $C > 2\sqrt{2}$, such that as $n \to \infty$,

$$\mathbb{P}\left(\max_k |N_{>k}(n) - np_{>k}| \geq C(1 + \sqrt{n \log n})\right) = o(1).$$

Such concentration results restrict the choice of k_n, since:

$$\mathbb{P}\left(\left|N_{>[b(n/k_n)y]}(n) - \mathbb{E}(N_{>[b(n/k_n)y]}(n))\right| > \epsilon k_n\right)$$
$$\leq \mathbb{P}\left(\max_k |N_{>k}(n) - \mathbb{E}(N_{>k}(n))| \geq \epsilon k_n\right).$$

Hence, the intermediate sequence k_n must be large enough so that

$$\mathbb{P}\left(\max_k |N_{>k}(n) - \mathbb{E}(N_{>k}(n))| \geq \epsilon k_n\right) = o(1).$$

Hence, the intermediate sequence k_n must be large enough so that

$$\mathbb{P}\left(\max_k |N_{>k}(n) - \mathbb{E}(N_{>k}(n))| \geq \epsilon k_n\right) = o(1).$$

Sufficient condition:

$$\liminf_{n\to\infty} k_n/(n \log n)^{1/2} > 0.$$

Hence, the intermediate sequence k_n must be large enough so that

$$\mathbb{P}\left(\max_k |N_{>k}(n) - \mathbb{E}(N_{>k}(n))| \geq \epsilon k_n\right) = o(1).$$

Sufficient condition:

$$\liminf_{n\to\infty} k_n/(n \log n)^{1/2} > 0.$$

Convergence of the Tail Empirical Measure:
Let $D_{(1)}(n) \geq D_{(2)}(n) \geq \cdots \geq D_{(n)}(n)$ be the order statistics of the degree sequence. Suppose that $\{k_n\}$ is some intermediate sequence satisfying

$$\liminf_{n\to\infty} k_n/(n \log n)^{1/2} > 0 \quad \text{and} \quad k_n/n \to 0 \quad \text{as} \quad n \to \infty,$$

then

$$\frac{1}{k_n} \sum_{i=1}^{n} \epsilon_{D_i(n)/D_{(k_n)}(n)}(\cdot) \Rightarrow \nu_{2+\delta},$$

in $M_+((0, \infty])$.

Consistency of the Hill Estimator:

Define the Hill estimator as

$$H_{k_n,n} = \frac{1}{k_n} \sum_{i=1}^{k_n} \log \frac{D_{(i)}(n)}{D_{(k_n+1)}(n)}.$$

Let $\{k_n\}$ be an intermediate sequence satisfying

$$\liminf_{n\to\infty} k_n/(n \log n)^{1/2} > 0 \quad \text{and} \quad k_n/n \to 0 \quad \text{as} \quad n \to \infty.$$

Then

$$H_{k_n,n} \xrightarrow{P} \frac{1}{2+\delta}.$$

Consistency of the Hill Estimator:

Define the Hill estimator as

$$H_{k_n,n} = \frac{1}{k_n}\sum_{i=1}^{k_n} \log \frac{D_{(i)}(n)}{D_{(k_n+1)}(n)}.$$

Let $\{k_n\}$ be an intermediate sequence satisfying

$$\liminf_{n\to\infty} k_n/(n\log n)^{1/2} > 0 \quad \text{and} \quad k_n/n \to 0 \quad \text{as} \quad n\to\infty.$$

Then

$$H_{k_n,n} \xrightarrow{P} \frac{1}{2+\delta}.$$

Proof idea: Write the Hill estimator as $H_{k_n,n} = \int_1^\infty \hat{\nu}_n(y,\infty]\frac{\mathrm{d}y}{y} =: T(\hat{\nu}_n)$, and justify the the continuity of the mapping T at $\nu_{2+\delta}$ so that

$$H_{k_n,n} = \int_1^\infty \hat{\nu}_n(y,\infty]\frac{\mathrm{d}y}{y} \xrightarrow{P} \int_1^\infty \nu_{2+\delta}(y,\infty]\frac{\mathrm{d}y}{y} = \frac{1}{2+\delta}.$$

Wrap-Up

- Undirected linear preferential attachment model is widely used to model social networks.
    - Generates power laws.

Wrap-Up

- Undirected linear preferential attachment model is widely used to model social networks.
    - Generates power laws.
- Practical issue: estimate the power-law exponent.
    - Hill estimator $\Rightarrow$ More **ROBUST**.

Wrap-Up

- Undirected linear preferential attachment model is widely used to model social networks.
  - Generates power laws.
- Practical issue: estimate the power-law exponent.
  - Hill estimator $\Rightarrow$ More **ROBUST**.
- Consistency of Hill estimator for network data:
  - Embedding technique:
    Degree sequence $\mapsto$ A sequence of birth immigration processes.
  - Convergence of the tail empirical measure.
  - Convergence of Hill.

Selected References

- T. Wang and S.I. Resnick. Consistency of Hill estimators in a linear preferential attachment model. ArXiv e-prints, 2017. Submitted to Extremes, under revision.
- F. Gao and A. van der Vaart (2017). On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees. Stoch. Process. Appl. 127.11, pp. 3754–3775.
- B.M. Hill. A simple general approach to inference about the tail of a distribution. Ann. Statist., 3:1163–1174, 1975.