



The Linear Preferential Attachment Model for Social Network Growth

Sidney Resnick

School of Operations Research and Information Engineering
Rhodes Hall, Cornell University
Ithaca NY 14853 USA

<http://people.orie.cornell.edu/sid>

<http://www.orie.cornell.edu/research/groups/multheavyltail>
sir1@cornell.edu

BIRS Oaxaca

June 15, 2018

MURI: R. Davis, P. Wan (Columbia); T. Wang, S. Resnick,
G. Samorodnitsky (Cornell)

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 1 of 34

Go Back

Full Screen

Close

Quit

1. Preferential Attachment Outline:

- Describe a 5 (really 4) parameter preferential attachment (PA) model.
- Highlight some relevant mathematical properties.
- Think about calibrating (fitting) the model to simulated and real data.
- Compare statistical approaches: MLE, SN and EV.
 - MLE: Is it naive or useful (or both) to fit a 5-parameter model to 1.5 million data?
 - How do asymptotic EVT or HT methods compare with SN or MLE:
 - * assuming the model is correct;
 - * assuming data corruption;
 - * assuming model error–data simulated from different model.
 - Conclusions: MLE efficient but usually more sensitive to model error or data corruption.
- What can we learn by trying to fit?
- Why does everyone use the Hill estimator?



<i>Plan</i>
<i>PA Model</i>
<i>MRV & PA</i>
<i>Calibration</i>
<i>Why Hill estimation?</i>
<i>Concluding</i>

[Title Page](#)

◀◀ ▶▶

◀ ▶

Page 2 of 34

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

2. A growing preferential attachment network model

See Bollobás, Borgs, Chayes, and Riordan (2003) and Krapivsky and Redner (2001).

2.1. Model description

- Model parameters:
 - Flip a **three-sided** coin corresponding to three *scenarios*: 1,3,2, with probabilities α, γ, β , with $\alpha + \beta + \gamma = 1$.
 - * Alternatively consider iid multinomial variables $\{J_n\}$ with cells 1,3,2.
 - * (Data suggests you may need a **5-sided** coin.)
 - $\delta_{in} \geq 0, \delta_{out} \geq 0$. (MLE inference requires these to be strictly positive.)
- $G(n) = (V_n, E_n)$ is a directed random graph with n edges.
- Node set of $G(n)$ is V_n ; so $|V_n| = N(n)$.
- Set of edges of $G(n)$ is $E_n = \{(u, v) \in V_n \times V_n : (u, v) \in E_n\}$.
- In-degree of v in $G(n)$ is $D_{in}^{(n)}(v)$; out-degree of v is $D_{out}^{(n)}(v)$.
- Obtain $G(n)$ from $G(n-1)$ in Markovian construction as follows:



- Plan
- PA Model
- MRV & PA
- Calibration
- Why Hill estimation?
- Concluding

Title Page

◀◀ ▶▶

◀ ▶

Page 3 of 34

Go Back

Full Screen

Close

Quit

1. With probability α , append to $G(n-1)$ a new node $v \notin V_{n-1}$ and create directed edge $v \mapsto w \in V_{n-1}$ with probability

$$\frac{D_{\text{in}}^{(n-1)}(w) + \delta_{\text{in}}}{n-1 + \delta_{\text{in}}N(n-1)}.$$

3. With probability γ , append to $G(n-1)$ a new node $v \notin V_{n-1}$ and create directed edge $w \in V_{n-1} \mapsto v \notin V_{n-1}$ with probability

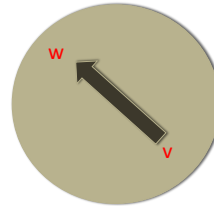
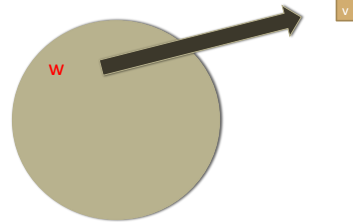
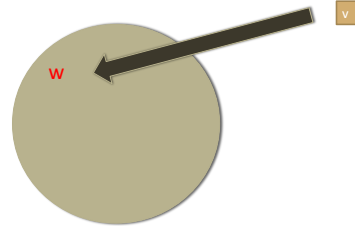
$$\frac{D_{\text{out}}^{(n-1)}(w) + \delta_{\text{out}}}{n-1 + \delta_{\text{out}}N(n-1)}.$$

2. With probability β , create new directed edge between existing nodes

$$v \in V_{n-1} \mapsto w \in V_{n-1}$$

with probability

$$\left(\frac{D_{\text{out}}^{(n-1)}(v) + \delta_{\text{out}}}{n-1 + \delta_{\text{out}}N(n-1)} \right) \left(\frac{D_{\text{in}}^{(n-1)}(w) + \delta_{\text{in}}}{n-1 + \delta_{\text{in}}N(n-1)} \right)$$



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 4 of 34

Go Back

Full Screen

Close

Quit

2.2. Background: What's known.

Notation:

$N(n)$ = # nodes in V_n . Binomial rv with success prob $\alpha + \gamma$.

n = # edges in E_n .

$N_{ij}(n)$ = # nodes with in-degree= i and out-degree= j in $G(n)$.

Then (eg, [Bollobás, Borgs, Chayes, and Riordan \(2003\)](#)) the limiting proportion of nodes with in-degree= i and out-degree= j is

$$\lim_{n \rightarrow \infty} \frac{N_{ij}(n)}{N(n)} = p_{ij} = \text{a prob mass function.}$$

Marginally, the limiting degree frequency (p_{ij}) has power-law tails: For some finite positive constants C_{in} and C_{out} ,

$$p_i(\text{in}) := \sum_{j=0}^{\infty} p_{ij} \sim C_{in} i^{-\iota_{in}} \quad \text{as } i \rightarrow \infty, \text{ as long as } \alpha \delta_{in} + \gamma > 0,$$

$$p_j(\text{out}) := \sum_{i=0}^{\infty} p_{ij} \sim C_{out} j^{-\iota_{out}} \quad \text{as } j \rightarrow \infty, \text{ as long as } \gamma \delta_{out} + \alpha > 0,$$

where

$$\iota_{in} = 1 + \frac{1 + \delta_{in}(\alpha + \gamma)}{\alpha + \beta}, \quad \iota_{out} = 1 + \frac{1 + \delta_{out}(\alpha + \gamma)}{\gamma + \beta}.$$

Conclude that $\iota_{in} > 1$, $\iota_{out} > 1$, and manufacture random pair (I, O) with

$$(I, 0) \sim \{p_{ij}\}.$$

Then

$$\begin{aligned} P[I = i] &\sim C_{in} i^{-\iota_{in}}, & i \rightarrow \infty; \\ P[O = j] &\sim C_{out} j^{-\iota_{out}}, & j \rightarrow \infty. \end{aligned}$$

So,

$$P[I > x] \sim k_{in} x^{-(\iota_{in}-1)}, \quad P[O > x] \sim k_{out} x^{-(\iota_{out}-1)}. \quad (x \rightarrow \infty).$$

Note: First we let $n \rightarrow \infty$ and then $i, j \rightarrow \infty$.

Question: Does (I, O) have a joint heavy tail? That is, is the distribution multivariate regularly varying?

Yes:

But₁: it's non-standard regular variation; in general $\iota_{in} \neq \iota_{out}$.

But₂: We have already let $n \rightarrow \infty$.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 6 of 34

Go Back

Full Screen

Close

Quit



3. Multivariate regular variation and preferential attachment

3.1. One answer: Regular variation of measures.

Samorodnitsky, Resnick, Towsley, Davis, Willis, and Wan (2016),
Resnick and Samorodnitsky (2015)

Theorem. Set $c_1 = 1/(\iota_{in} - 1)$, $c_2 = 1/(\iota_{out} - 1)$. The random vector (I, O) with joint mass function $\{p_{ij}\}$ satisfies as $t \rightarrow \infty$,

$$tP \left[\left(\frac{I}{t^{1/(\iota_{in}-1)}}, \frac{O}{t^{1/(\iota_{out}-1)}} \right) \in \cdot \right] \xrightarrow{v} \frac{\gamma}{\alpha + \gamma} \nu_1(\cdot) + \frac{\alpha}{\alpha + \gamma} \nu_2(\cdot),$$

vaguely in $M_+([0, \infty]^2 \setminus \{\mathbf{0}\})$ and ν_1 and ν_2 concentrate on $(0, \infty)^2$ and have Lebesgue densities f_1, f_2 given by,

$$f_1(x, y) = c_1^{-1} (\Gamma(\delta_{in} + 1) \Gamma(\delta_{out}))^{-1} x^{\delta_{in}} y^{\delta_{out}-1} \times \int_0^\infty z^{-(2+1/c_1+\delta_{in}+a\delta_{out})} e^{-(x/z+y/z^a)} dz,$$

and

$$f_2(x, y) = c_1^{-1} (\Gamma(\delta_{in}) \Gamma(\delta_{out} + 1))^{-1} x^{\delta_{in}-1} y^{\delta_{out}} \times \int_0^\infty z^{-(1+a+1/c_1+\delta_{in}+a\delta_{out})} e^{-(x/z+y/z^a)} dz.$$

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 7 of 34

Go Back

Full Screen

Close

Quit

3.2. Another answer: Regular variation of the mass function $\{p_{i,j}\}$.

Wang and Resnick (2016) Based on form of the generating function of $p_{i,j} = p(i, j)$ and recalling the representation of

$$(I, O) \sim p_{i,j}.$$

we get regular variation of the mass functions:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{p([n^{c_1}x], [n^{c_2}y])}{n^{-(1+c_1+c_2)}} &= \frac{\gamma}{\alpha + \gamma} f_1(x, y) + \frac{\alpha}{\alpha + \gamma} f_2(x, y) \\ &= \frac{\gamma}{\alpha + \gamma} \frac{x^{\delta_{in}} y^{\delta_{out} - 1}}{c_1 \Gamma(\delta_{in} + 1) \Gamma(\delta_{out})} \int_0^\infty z^{-(2+1/c_1 + \delta_{in} + a\delta_{out})} e^{-\left(\frac{x}{z} + \frac{y}{z^a}\right)} dz \\ &\quad + \frac{\alpha}{\alpha + \gamma} \frac{x^{\delta_{in} - 1} y^{\delta_{out}}}{c_1 \Gamma(\delta_{in}) \Gamma(\delta_{out} + 1)} \int_0^\infty z^{-(1+a+1/c_1 + \lambda + a\delta_{out})} e^{-\left(\frac{x}{z} + \frac{y}{z^a}\right)} dz. \end{aligned}$$

Note:

- Regular variation of the measure does not always imply regular variation of the mass function.
- More surprising: variation of the mass function does not always imply regular variation of the measure (though this is true in 1-dimension). Need a regularity condition.

4. Model Calibration/Fitting/Estimation

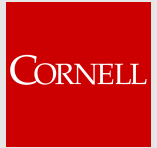
Wan, Wang, Davis, and Resnick (2017a,b)

It is ambitious to fit the model to real data (as opposed to simulated data)!

4.1. Issues, approaches, thoughts:

4.1.1. Asymptotic EVT methods vs MLE applied to fully parameterized model?

- Do we trust such a simple model?
 - Probably not. Rule of thumb: The larger the dataset, the more likely you reject a model.
 - Perhaps we learn something from discrepancies between the model and the data.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 9 of 34

Go Back

Full Screen

Close

Quit



- Should we use extreme value (EV) or heavy tail asymptotics to do estimation or full parametric MLE? For certain network estimation problems,
 - EV methods may be more robust against
 - * inevitable model error or,
 - * data corruption,
 - but definitely suffer in accuracy compared to model based estimation when the model is correct (ie simulated).
 - Both MLE and EV implemented.

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 10 of 34

Go Back

Full Screen

Close

Quit



- The limit density of regular variation $f(x, y; \boldsymbol{\theta})$ has parameter

$$\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta_{in}, \delta_{out}).$$

But recall $f(x, y; \boldsymbol{\theta})$ results from essentially a double limit:

- Taking $\lim_{n \rightarrow \infty} N_n(i, j)/N(n)$ to get $p(i, j)$.
- Letting $i \rightarrow \infty$ and $j \rightarrow \infty$ in a controlled way to get the limit density.

Hence, the asymptotic EV method requires two levels of pretend:

- Pretend $n = \infty$ or $N_n(i, j)/N(n) = p_{ij}$.
- Pretend i, j large so that $p_{ij} \approx f(x, y)$.
- To fit this particular five-parameter model, asymptotics philosophy can be implemented and requires using $f(x, y; \boldsymbol{\theta})$. Use Hill estimator to get

- l_{in} ;
- l_{out} ;

and then estimate remaining parameters from the profile likelihood corresponding to $f(x, y; \boldsymbol{\theta})$.

- The data does not result from repeated sampling. We followed tradition and avoided this issue. But some progress.

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 11 of 34

Go Back

Full Screen

Close

Quit

4.2. What data is available?

- Do we have the full history of edge creation with time stamps?
 - Available when simulate network (Atwood, Ribeiro, and Towsley (2015), J. Roy, P. Wan)
 - Sometimes available with real data (SNAP, KONECT); time stamps reliable? See Kunegis (2013).
 - Full MLE methodology implemented and works beautifully when model is correct (simulated).
 - * Construct the likelihood

$$L(\alpha, \beta, \delta_{in}, \delta_{out}; \underbrace{G(t), e_t, (\text{etc}), t = n_0, \dots, n}_{\text{observables}})$$

Express as product of observables and parameters.

- * Differentiate likelihood to get score function; check \exists unique max at “true” parameter values. Check MLE is strongly consistent.
- * Efficient: using MG CLT applied to the score function:

Let

$$\hat{\theta}_n^{MLE} = (\hat{\alpha}^{MLE}, \hat{\beta}^{MLE}, \hat{\delta}_{in}^{MLE}, \hat{\delta}_{out}^{MLE}).$$

be the MLE estimator for θ , the parameter vector of the



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 12 of 34

Go Back

Full Screen

Close

Quit



preferential attachment model. Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{\text{MLE}} - \boldsymbol{\theta}) \Rightarrow N(\mathbf{0}, \Sigma(\boldsymbol{\theta})),$$

where

$$\Sigma^{-1}(\boldsymbol{\theta}) = I(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1-\beta}{\alpha(1-\alpha-\beta)} & \frac{1}{1-\alpha-\beta} & 0 & 0 \\ \frac{1}{1-\alpha-\beta} & \frac{1-\alpha}{\beta(1-\alpha-\beta)} & 0 & 0 \\ 0 & 0 & I_{\text{in}} & 0 \\ 0 & 0 & 0 & I_{\text{out}} \end{bmatrix},$$

with

$$I_{\text{in}} = \sum_{i=0}^{\infty} \frac{p_{>i}^{\text{in}}}{(i + \delta_{\text{in}})^2} - \frac{\gamma}{\delta_{\text{in}}^2} - \frac{(\alpha + \beta)(1 - \beta)^2}{(1 + \delta_{\text{in}}(1 - \beta))^2}, \quad (1)$$
$$I_{\text{out}} = \sum_{j=0}^{\infty} \frac{p_{>j}^{\text{out}}}{(j + \delta_{\text{out}})^2} - \frac{\alpha}{\delta_{\text{out}}^2} - \frac{(\gamma + \beta)(1 - \beta)^2}{(1 + \delta_{\text{out}}(1 - \beta))^2}.$$
$$p_{>i}^{\text{in}} = \sum_{k>i, j} p_{kj}, \quad p_{>j}^{\text{out}} = \sum_{k, l>j} p_{kl}.$$

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 13 of 34

Go Back

Full Screen

Close

Quit

4.2.1. Data available? One snapshot (SN) method.

- Fixed time snapshot of the network; effectively observe at time n and NOT at times $1, \dots, n$.
 - MLE (approximate) still works well; estimators appear to be CAN but unsurprisingly there is noticeable loss of efficiency compared to MLE on full history.



- Plan*
- PA Model*
- MRV & PA*
- Calibration***
- Why Hill estimation?*
- Concluding*

Title Page

◀◀ ▶▶

◀ ▶

Page 14 of 34

Go Back

Full Screen

Close

Quit

4.3. EV method

- Use Hill (or comparable method) to estimate t_{in}, t_{out} based on the data consisting of degrees of each node.

Hill estimator Hill (1975): Suppose $X_n, n \geq 1$ are iid (!!!) rv's and

$$P[X_1 > x] \sim x^{-\alpha}.$$

Suppose order statistics in sample of size n is $X_{(1)} \geq \dots \geq X_{(n)}$. Then the Hill estimator of α based on k -upper order statistics is $\hat{\alpha} = H_{k,n}^{-1}$ where

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

This is consistent and *usually* asymptotically normal.

- If you trust the model, use asymptotic density to estimate remaining parameters by maximizing a profile likelihood assuming $\hat{t}_{in}, \hat{t}_{out}$.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 15 of 34

Go Back

Full Screen

Close

Quit

4.4. Example: MLE applied to Dutch Wiki talk network (KONECT)

KONECT: Kunegis (2013); from University of Koblenz-Landau.

- 225,749 nodes: registered users of Dutch Wikipedia.
- 1,554,699 edges from A to B means node A wrote a message on the “talk page” of user B.
- Edges recorded with time stamps.
- Group broadcasts exist and must be cleaned; deleted 37 senders of >40 messages on the assumption that this did not constitute normal social network behavior. Deleted dead nodes: nodes with in-degree 0. After cleaning, refer to the data as the *reduced* data.
- Additional edge formation scenarios in the data:
 - $J_n = 4$ with probability ξ means two new nodes (v, w) arrive simultaneously with an edge (v, w) .
 - $J_n = 5$ with probability ρ means new node arrives with self-loop.
- Estimate $\alpha, \beta, \gamma, \xi$ with scenario frequencies. Estimate $\delta_{in}, \delta_{out}$ with approximate MLE estimating equations (replace scenario probabilities by scenario frequencies).



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 16 of 34

Go Back

Full Screen

Close

Quit

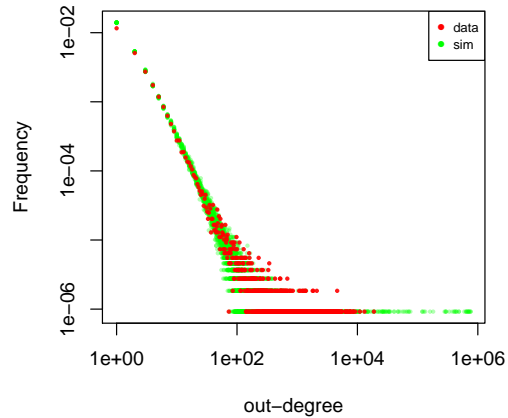
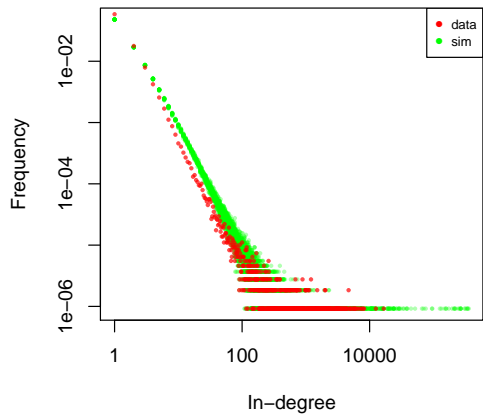


Figure 1: In- and out-degree frequencies of the reduced Wiki talk network (red) and 20 simulated fitted linear preferential attachment networks with constant parameters (green).

- Plan
- PA Model
- MRV & PA
- Calibration**
- Why Hill estimation?
- Concluding

Title Page

⏪ ⏩

◀ ▶

Page 17 of 34

Go Back

Full Screen

Close

Quit

Another issue:

- \exists evidence the parameter vector is not constant over time.
- Some success fitting using parameters that are piecewise constant over time. Need more systematic investigation of change points.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 18 of 34

Go Back

Full Screen

Close

Quit

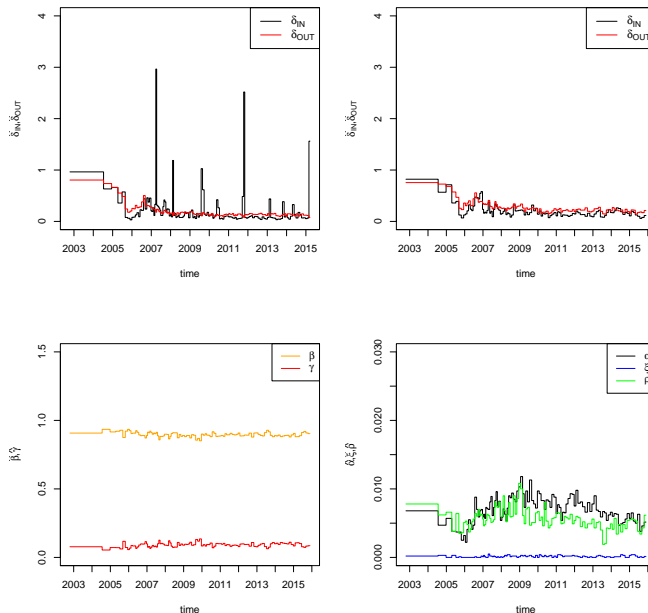


Figure 2: Local parameter estimates of linear PA model for the full and reduced Wiki talk network. Upper left: $(\hat{\delta}_{in}, \hat{\delta}_{out})$ for full network. Upper right, lower left, lower right: $(\hat{\delta}_{in}, \hat{\delta}_{out})$, $(\hat{\beta}, \hat{\gamma})$, $(\hat{\alpha}, \hat{\xi}, \hat{\rho})$ for the reduced network, respectively.

- Plan
- PA Model
- MRV & PA
- Calibration**
- Why Hill estimation?
- Concluding

Title Page

⏪ ⏩

◀ ▶

Page 19 of 34

Go Back

Full Screen

Close

Quit

4.5. Comparison of methods.

On simulated data where we know correct answers, compare:

- EV: extreme value asymptotic methods,
- SN: one snapshot method;
- MLE: Full MLE knowing history of edge formation.

For meaningful comparison of SN, EV, MLE, allow:

- Data corruption: simulated data with randomly added or deleted edges.
- Model error:
 - simulate superstar (Lady Gaga) model where fixed proportion of new nodes attach to the superstar; rest obey linear PA.
 - estimate parameters while pretending data from ordinary PA.

Broad conclusions:

- If there is either model error or data corruption, EV methods hold their own and are a useful supplementary technique.
- Unsurprisingly, if there is no model error or data corruption (hah!)

$$EV < SN=\text{one snapshot} < \text{full MLE.}$$



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 20 of 34

Go Back

Full Screen

Close

Quit



4.5.1. No model error or data corruption

- Hold $(\beta, \delta_{in}, \delta_{out}) = (0.4, 1, 1)$ fixed and $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$.
- Simulate 200 networks with 10^5 edges.
- Estimate α, l_{in}, l_{out} for each replication.

- Compute boxplots for **biases** of 200 estimates

$$(\hat{\alpha}^{EV}, \hat{l}_{in}^{EV}, \hat{l}_{out}^{EV}), (\hat{\alpha}^{SN}, \hat{l}_{in}^{SN}, \hat{l}_{out}^{SN}), (\hat{\alpha}^{MLE}, \hat{l}_{in}^{MLE}, \hat{l}_{out}^{MLE}).$$

- Conclusion: EV more biased and variable than SN or MLE for uncorrupted data from correct model.

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



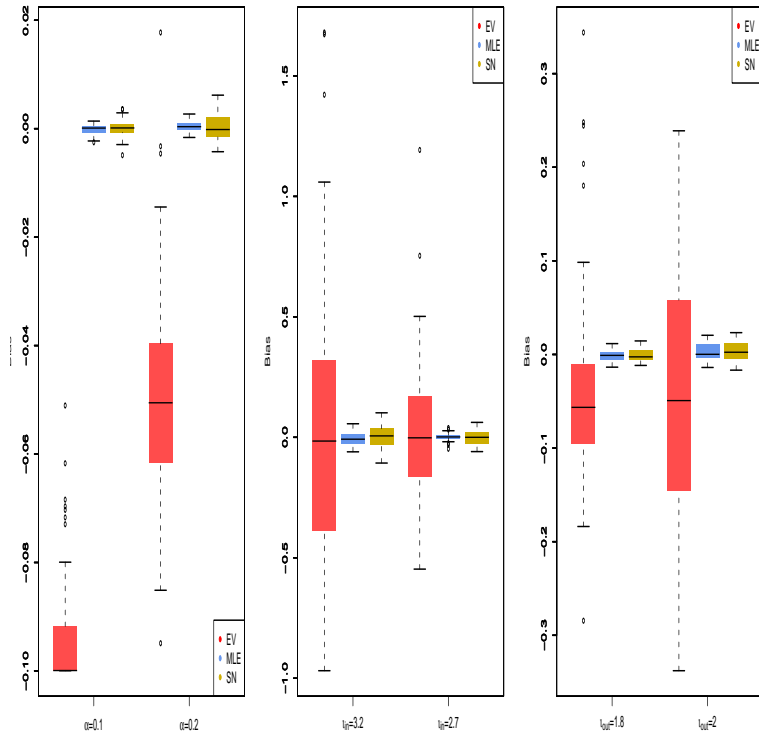
Page 21 of 34

Go Back

Full Screen

Close

Quit



$\alpha = 0.1, 0.2$. Fix $(\beta, \delta_{in}, \delta_{out}) = (0.4, 1, 1)$.

- Plan
- PA Model
- MRV & PA
- Calibration**
- Why Hill estimation?
- Concluding

Title Page

⏪ ⏩

◀ ▶

Page 22 of 34

Go Back

Full Screen

Close

Quit

4.5.2. Model error: random edge additions

- To get $G(n)$ from $G(n - 1)$,
 - Flip a coin; with prob p_a sequentially pick 2 nodes at random and wire an edge from first to 2nd.
 - Otherwise, with prob $1 - p_a$ construct using PA rules. (interpolate between PA and Erdos-Renyi.)

- Use this mechanism to generate 200 graphs with 10^5 edges and

$$(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}}) = (0.3, 0.4, 0.3, 1, 1),$$
$$p_a \in \{0.025, 0.05, 0.075, 0.1, 0.125, 0.15\}.$$

- Pretend data from linear PA.
- Display mean estimates and 2.5% and 97.5% empirical quantiles of (a) δ_{in} ; (b) δ_{out} ; (c) α ; (d) γ ; (e) ι_{in} ; (f) ι_{out} , using MLE (black), SN (red) and EV (blue) methods over 200 replications
- Dotted line=true values.
- For some parameters, EV much better and never worse.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 23 of 34

Go Back

Full Screen

Close

Quit



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page

◀◀

▶▶

◀

▶

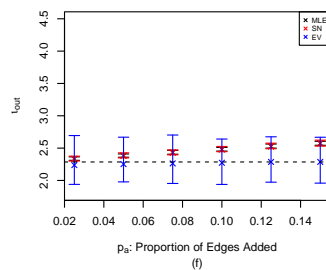
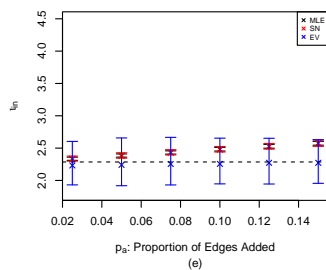
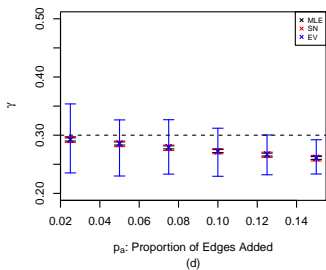
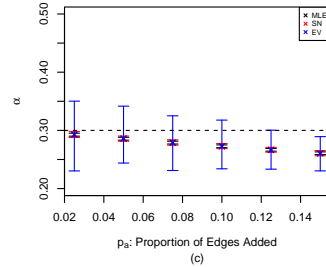
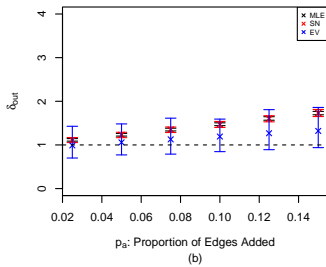
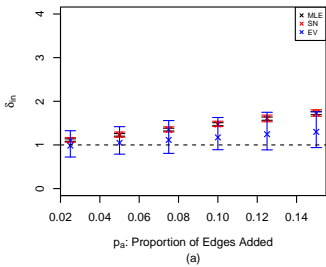
Page 24 of 34

Go Back

Full Screen

Close

Quit



4.5.3. Model error: Simulate superstar model but don't tell the statistician.

- Simulate the superstar model:
 - To get from $G(n-1)$ to $G(n)$, flip a coin: With prob p a new node appears and attaches to the superstar. Otherwise, with prob $1-p$ do linear PA.
 - Use:

$$(\alpha, \beta, \gamma, \delta_{\text{in}}, \delta_{\text{out}}, n, p) = (0.3, 0.4, 0.3, 1, 1, 10^5, 0.25).$$

- Don't tell the MLE estimator it's the wrong model.
- Compute empirical in- and out-degree frequencies from simulated data (green next page).
- Compute estimates $\hat{\Theta}^{EV}$, $\hat{\Theta}^{MLE}$ using EV, MLE applied to linear PA.
- Simulate 20 linear PA networks using $\hat{\Theta}^{EV}$ and 20 using $\hat{\Theta}^{MLE}$.
- Overlay empirical frequencies.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 25 of 34

Go Back

Full Screen

Close

Quit



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



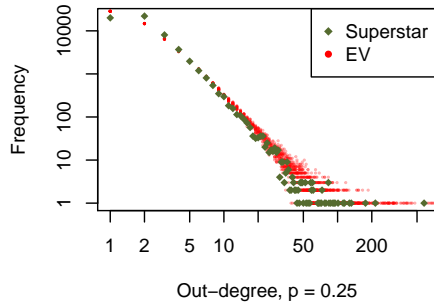
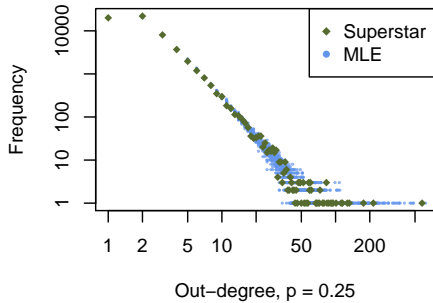
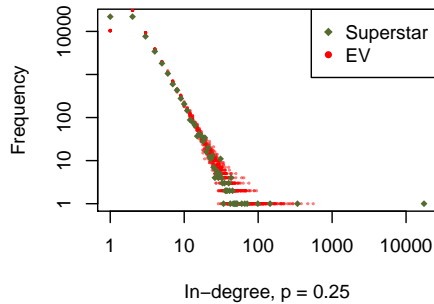
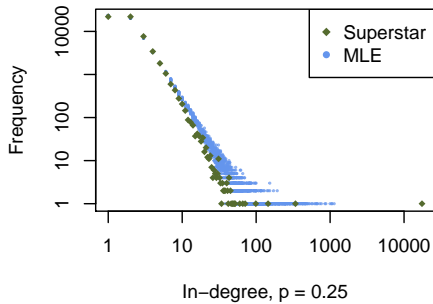
Page 26 of 34

Go Back


Full Screen

Close

Quit



5. Why Hill estimation?



KONECT

Home Networks Statistics Plots Search Downloads Software Publications License About Help

Google Custom Search

KONECT > Networks > DNC emails

DNC emails

About this network

This is the directed network of emails in the 2016 Democratic National Committee email leak. The Democratic National Committee (DNC) is the formal governing body for the United States Democratic Party. A dump of emails of the DNC was leaked in 2016. Nodes in the network correspond to persons in the dataset. A directed edge in the dataset denotes that a person has sent an email to another person. Since an email can have any number of recipients, a single email is mapped to multiple edges in this dataset, resulting in the number of edges in this network being about twice the number of emails in the dump.

Network info

Code	DNc
Category	● Communication
Data source	http://www.rene-pickhardt.de/extracting-2-social-network-graphs-from-the-democratic-national-committee-email-corpus-on-wikileaks/
Vertex type	Person
Edge type	Email
Format	📄 Directed
Edge weights	📊 Multiple unweighted
Metadata	🕒 Loop 🕒 Timestamps
Size	2,029 vertices (persons)
Volume	39,264 edges (emails)
Unique volume	5,598 edges (emails)
Average degree (overall)	38,703 edges / vertex
Fill	0.0015655 edges / vertex ²
Maximum degree	5,813 edges
Reciprocity	41.9%
Size of LCC	1,833 vertices
Size of LSCC	520 vertices
Wedge count	317,905
Claw count	59,899,010
Triangle count	9,431
Square count	209,206
4-tour count	2,954,036
Power law exponent (estimated) with d_{min}	2.0110 ($d_{min} = 1$) ←
Gini coefficient	91.1%
Relative edge distribution entropy	79.0%
Assortativity	-0.30655
Clustering coefficient	8.90%
Diameter	8 edges
90-percentile effective diameter	3.98 edges

CORNELL

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 28 of 34

Go Back

Full Screen

Close

Quit

Why does Hill work?

Wang and Resnick (2017), Wang and Resnick (201 ∞).

Progress: Cases where Hill is provably consistent:

- Undirected model Wang and Resnick (2017).
- Directed model with $\beta = 0$.
- Results constrained by the methods which are not sufficiently robust: Embed in birth or BI or switched BI processes.



Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding

Title Page



Page 29 of 34

Go Back

Full Screen

Close

Quit

6. Concluding remarks

6.1. What would we like to know?

- Reciprocity? Cliques? Neighborhoods? Nothing analytical to report.
- Change point methods? Can we identify regimes where the parameters change and hence uncover evidence of interference in normal social behavior (intrusion, bot or admin behavior).
- Fragility of methods.
- Allow models to have a node attaching simultaneously to multiple other nodes. Do this without introducing a gazillion parameters.
- Models that allow participants to leave the social network?

Contents

Plan

PA Model

MRV & PA

Calibration

Why Hill estimation?

Concluding



Title Page



Page 31 of 34

Go Back

Full Screen

Close

Quit

References

- J. Atwood, B. Ribeiro, and D. Towsley. Efficient network generation under general preferential attachment. *Computational Social Networks*, 2(1):7, 2015. ISSN 2197-4314. doi: 10.1186/s40649-015-0012-9. URL <http://dx.doi.org/10.1186/s40649-015-0012-9>.
- B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, 2003)*, pages 132–139, New York, 2003. ACM.
- B.M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3:1163–1174, 1975.
- P.L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123:1–14, 2001.
- J. Kunegis. Konect: the Koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350. ACM, 2013.
- S.I. Resnick and G. Samorodnitsky. Tauberian theory for multivariate regularly varying distributions with application to prefer-

ential attachment networks. *Extremes*, 18(3):349–367, 2015. doi: 10.1007/s10687-015-0216-2.

G. Samorodnitsky, S. Resnick, D. Towsley, R. Davis, A. Willis, and P. Wan. Nonstandard regular variation of in-degree and out-degree in the preferential attachment model. *Journal of Applied Probability*, 53(1):146–161, March 2016. doi: 10.1017/jpr.2015.15.

P. Wan, T. Wang, R. A. Davis, and S. I. Resnick. Fitting the linear preferential attachment model. *Electron. J. Statist.*, 11(2):3738–3780, 2017a. ISSN 1935-7524. doi: 10.1214/17-EJS1327.

P. Wan, T. Wang, R. A. Davis, and S. I. Resnick. Are extreme value estimation methods useful for network data? *ArXiv e-prints*, December 2017b.

T. Wang and S.I. Resnick. Multivariate regular variation of discrete mass functions with applications to preferential attachment networks. *Methodology and Computing in Applied Probability*, 2016. ISSN 1573-7713. doi: 10.1007/s11009-016-9503-x. URL <http://dx.doi.org/10.1007/s11009-016-9503-x>.

T. Wang and S.I. Resnick. Consistency of Hill estimators in a linear preferential attachment model. <https://arxiv.org/abs/1711.05911>, 2017. Under revision for *Extremes*.

T. Wang and S.I. Resnick. Degree growth rates and index estimation in a directed preferential attachment model. 201 ∞ . In preparation.



Title Page



Page 34 of 34

Go Back

Full Screen

Close

Quit