MURI Update Meeting
Multivariate Heavy Tail Phenomena: Modeling and Diagnostics
October 7, 2014,
Columbia University,

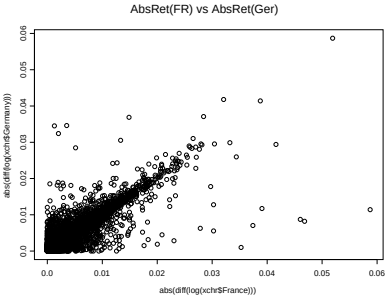
Sidney Resnick
School of Operations Research and Information Engineering
Rhodes Hall, Cornell University
Ithaca NY 14853 USA

http://people.orie.cornell.edu/∼sid
607 255 1210 sir1@cornell.edu

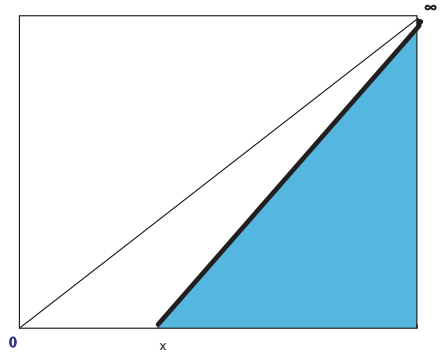
October 5, 2014

1. Mathematical framework

- Finished a survey giving proper mathematical framework in Lindskog, Resnick, and Roy (2014). Final submitted.
- Allows for flexible mathematical consideration of multivariate regular variation problems with multiple asymptotic regimes simultaneously existing.
- Ideal for handling risk contagion in presence of *asymptotic independence*. Also suitable for seeking multiple regular variation regimes in presence of *asymptotic full dependence*.



- Returns are multivariate heavy tailed.
- Limit measure concentrates on diagonal.
- Other examples
  - Xchr (Aus$, Chinese).
  - (Exxon, Chevron).
  - Tech sector returns? Lagged variables (Davis).



- If limit measure concentrates on the diagnonal, no mass in blue region

$$\{(u, v) : u > v + x\}.$$

- Wrong asymptotics?

Ongoing (Amy Willis, Bikram Das)

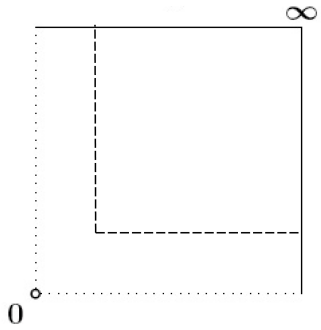
- Apply framework to cases other than asymptotic independence.
- Find data diagnostics.
- Dimension issues beyond 2.
- Discovery of regions where the limit measure has zero mass.

2. Model generation and diagnostic

2.1. A general construction of a standardized distribution

- On $\mathbb{R}^d_+$, delete closed cone F.
- Write
$$\aleph_F = \{\mathbf{x} : d(\mathbf{x}, F) = 1\}.$$
Take
$$\boldsymbol{\Theta} \text{ re in } \aleph_F, \quad R \sim \text{ Pareto}.$$
- Set
$$\boldsymbol{X} = R\boldsymbol{\Theta}$$
and $\boldsymbol{X} \in$ MRV on $R^d_+ \setminus F$.

- Can apply to multiple regimes by
  - first deleting F_1 (eg, origin)
  - then deleting $F_1 \cup F_2$; ie delete 2nd cone (eg, axes).



- How do the 2 regimes interact? Statistically identifiable? Das and Resnick (2014).

2.2. Model Detection Diagnostics

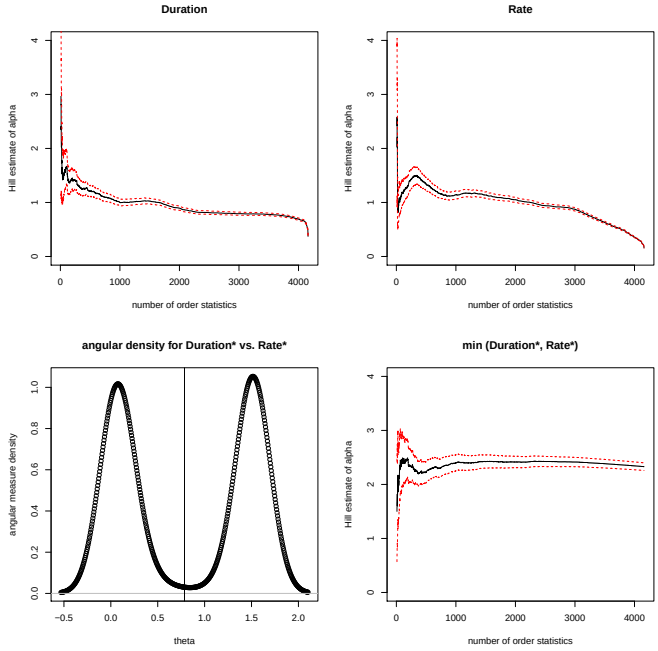
1. Reduction to one dimension:
   - $\boldsymbol{X} \in$ MRV iff $aX_1 \vee bX_2 \in RV(\alpha)$ for all $a \geq 0$, $b \geq 0$.
   - $\boldsymbol{X} \in$ HRV iff $aX_1 \wedge bX_2 \in RV(\alpha_0)$ for $a \wedge b > 0$.

   [Hint: Cannot check $\forall a, b$; works in higher dimensions.]
2. Use GPOLAR to convert to the CEV model and then use CEV diagnostics (Das and Resnick, 2011) using the *Hillish* and *Pickandsish* plots.
3. CEV model is of the form: Sppse $(\xi, \eta) \in \mathbb{R}_+ \times \mathbb{R}$ is a random vector and
   - $\exists b(t) \to \infty$,
   - $\exists$ limit measure μ

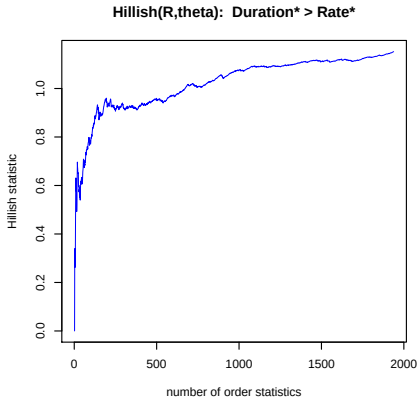
   such that
   $$tP\left[\left(\frac{\xi}{b(t)}, \eta\right) \in \cdot\right] \to \mu(\cdot).$$
   This is the form of convergence after coordinate transformation via GPOLAR.
4. Hillish statistic applied to both (ξ, η) and $(\xi, -\eta)$ detects this.

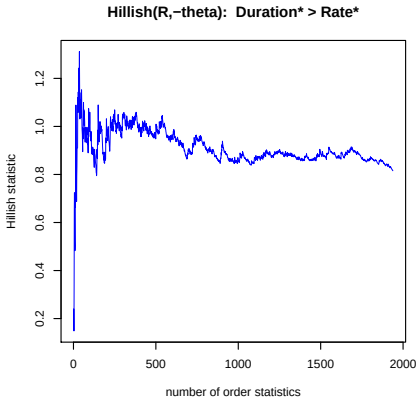
One dimensional analysis: BU (duration, rate).

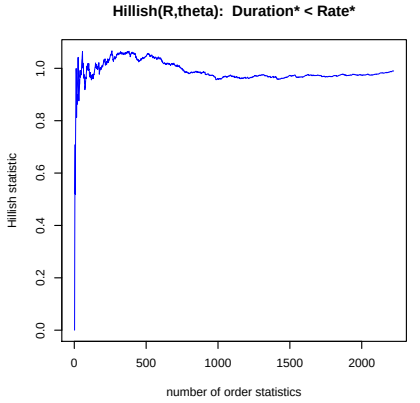
BU browser downloads; duration = duration of download; rate = file size/duration.

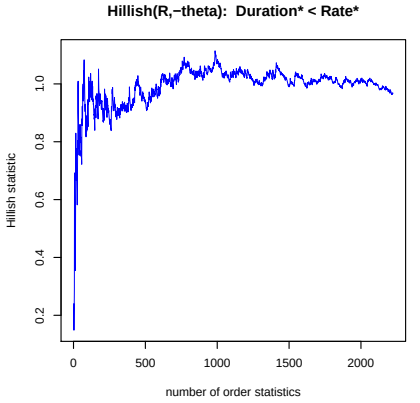


Hillish and CEV analysis for HRV.









Ongoing:

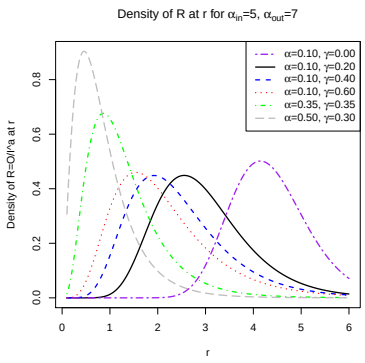
- Extensions to higher dimensions.
- Extensions to other deleted cones.
- More data examples.

3. Growing preferential attachment networks.

- Directed edge network grow according to preferential attachment postulates (Samorodnitsky, Resnick, Towsley, Davis, Willis, and Wan, 2014).
- Each node has (in-degree, out-degree) evolving with $n = 1, 2, \ldots$.
- At nth stage, the proportion of nodes with (in-degree, out-degree) $= (i, j)$ is

$$p_{ij}^{(n)} \overset{n\to\infty}{\to} p_{ij}.$$

- $\{p_{ij}\}$ is the mass function of a heavy tailed measure whose limit measure has a density that can be explicitly computed as a function of input parameters.
- General and powerful Tauberian theorem relating Laplace transform of (in-degree, out-degree) to the multivariate heavy tail (Resnick and Samorodnitsky, 2014).



- Algorithms (Atwood, Roy) to simulate the network growth.
- Provides test bed data for statistical methods where we know the answers (because we simulated the model). More from Joyjit.

Ongoing:

- Model error: Observe or infer frequencies $p_{ij}^{(n)}$ but try to infer something about heavy tailed structure of

$$\lim_{n\to\infty} p_{ij}^{(n)} = p_{ij}$$

  for $i \to \infty$ and $j \to \infty$.
- Theory for multivariate heavy tailed mass functions as opposed to their measures.
- Data is node based and not iid.
- Properties of estimators. Apply MG's and MG CLT.
- Simplify analysis using embedding in birth processes which are conditionally Poisson.
- Apply the Tauberian method to other models.
- Possibility of more than just 2 attributes per node (eg, friend, foe, strength of opinion).

- Other graceful methods to handle censoring.
- Neighborhood and community discovery (Davis, Wan, Zhang, Towsley, Jiang).

Contents

References

B. Das and S. Resnick. Generation and detection of multivariate regular variation and hidden regular variation. *ArXiv e-prints*, March 2014. URL http://adsabs.harvard.edu/abs/2014arXiv1403.5774D. Accepted pending revision in Stochastic Systems.

B. Das and S.I. Resnick. Detecting a conditional extreme value model. *Extremes*, 14(1):29–61, 2011.

F. Lindskog, S.I. Resnick, and J. Roy. Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. Technical report, School of ORIE, Cornell University, 2014. URL http://arxiv.org/abs/1307.5803. Accepted: Probability Surveys.

S.I. Resnick and G. Samorodnitsky. Tauberian Theory for Multivariate Regularly Varying Distributions with Application to Preferential Attachment Networks. *ArXiv e-prints*, June 2014. URL http://adsabs.harvard.edu/abs/2014arXiv1406.6395R.

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. *ArXiv e-prints*, May 2014. URL http://adsabs.harvard.edu/abs/2014arXiv1405.4882S.

A tale of tails!

Estimating power law indexes for network data

Joyjit Roy
In collaboration with Sidney I. Resnick

School of Operations Research and Information Engineering
Cornell University

October 07, 2014

Outline

Outline

Description of Datasets

Real-world network data

- Directed, signed social network data from technology news website `SLASHDOT`
- Dataset was provided to us by Zhi-li Zhang
- Conjectured to have "scale-free power-law like" behavior
- Main interest is to study multivariate tail behavior

Simulated data

- Preferential Attachment Model
- Yields directed network where asymptotically in-degrees and out-degrees exhibit power-law behavior with known tail-exponents
- Can be used to benchmark our methods
- We have developed a method to simulate from this model in $O(N)$ time where N is the number of nodes

Outline

Problems in tail-index estimation

- Usually power-law behavior is only asymptotic or may only be exhibited after a cutoff
- Forgoing the discrete assumption to fit continuous models lead to erros due to approximation
- Data might be also be censored as in the case of `SLASHDOT`

Problems in tail-index estimation

- Usually power-law behavior is only asymptotic or may only be exhibited after a cutoff
- Forgoing the discrete assumption to fit continuous models lead to erros due to approximation
- Data might be also be censored as in the case of `SLASHDOT`
- One popular package is `poweRlaw` package but the methods can be improved upon
- Slope based methods are inaccurate
- Other methods require tuning to yield credible estimates

Problems in tail-index estimation

- Usually power-law behavior is only asymptotic or may only be exhibited after a cutoff
- Forgoing the discrete assumption to fit continuous models lead to erros due to approximation
- Data might be also be censored as in the case of `SLASHDOT`
- One popular package is `poweRlaw` package but the methods can be improved upon
- Slope based methods are inaccurate
- Other methods require tuning to yield credible estimates

Our Goal

Find a statistical model for the data and an automated method to fit it that does not require manual threshold selection

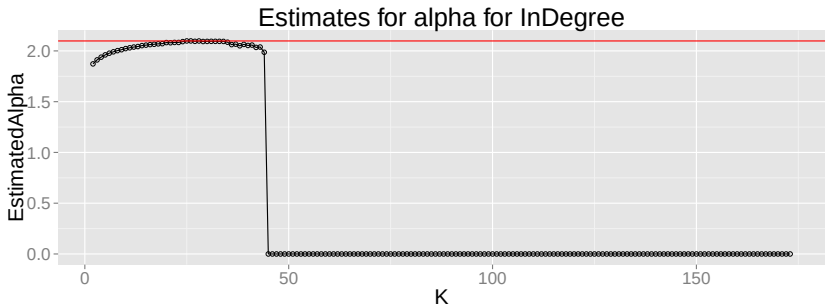
Outline

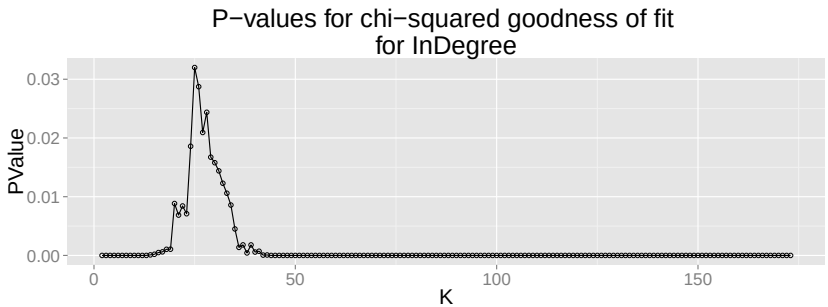
Using weighted linear regression

- Need to model dependence and the fact that the coefficients are constrained
- Let Y_i be the number of nodes with degree i. Assume
  - $Y_i \sim$ Multinomial(N, π) where N is the number of nodes
  - $\pi_i = Ci^{-\alpha}$ where α is the tail-index for i greater than threshold K
- Use the multivariate normal approximation to the multinomial and the delta method to obtain the approximate distribution for $\log Y_i$
- $\log Y_i \sim N\left(\log(1 - \sum_{i<K} \pi_i) - \log \zeta(K, \alpha) - \alpha \log i, \;\; \sigma_i(\alpha)\right)$ for $i \geq K$, where $\zeta(K, \alpha) = \sum_{i \geq K} i^{-\alpha}$
- Fit a model with parameters $(\pi_1, \ldots, \pi_{K-1}, \alpha)$ using iteratively reweighted least squares and choose K via a goodness of fit criterion
- Fast and reliable but iterative method to obtain estimate is unstable sometimes

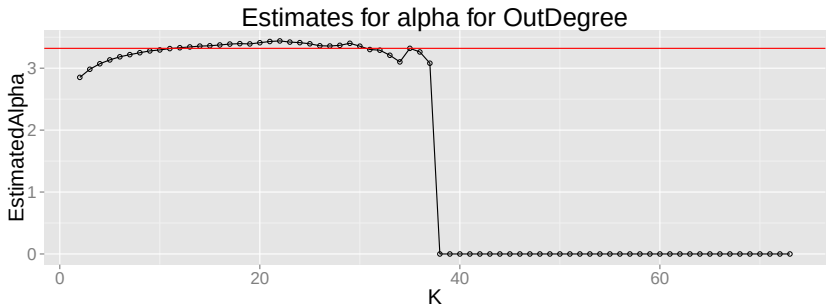
Comments

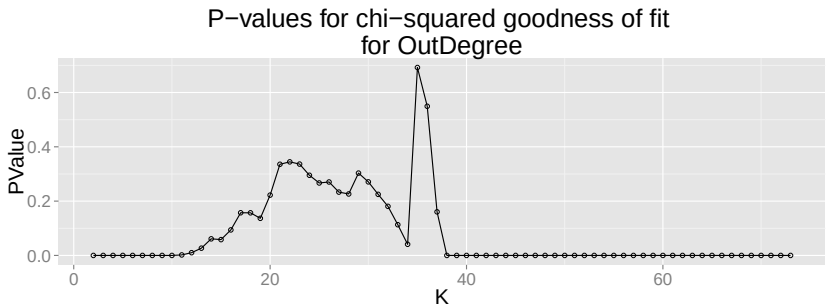
`SLASHDOT` data analysis of in-degree





Simulated data with tail-index of out-degree $= 3.67$





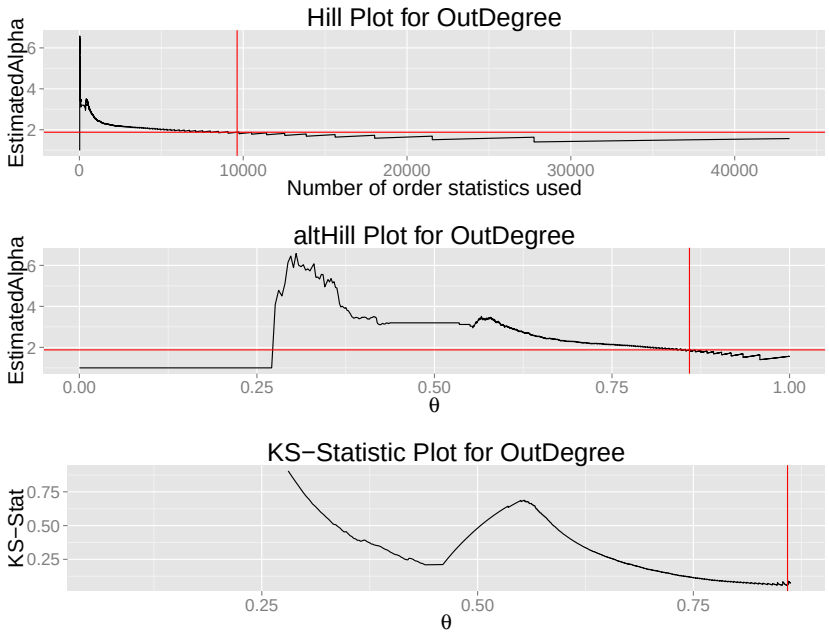
Outline

Using Likelihood based methods

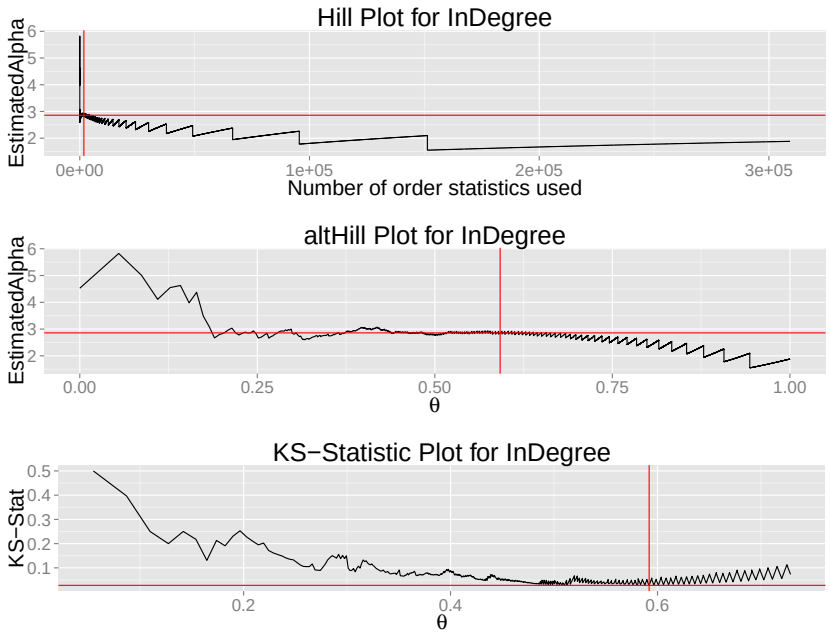
- Let $X_i, i = 1, 2, \ldots, N$ be the degrees of the N nodes in the graph. Assume X_i are iid and $\mathbb{P}[X = j | X > K] = \frac{j^{-\alpha}}{\zeta(K,\alpha)}$ for some threshold K. Recall $\zeta(K, \alpha) = \sum_{i \geq K} i^{-\alpha}$.
- Likelihood equation does not have a closed form solution for the maximum
- We also modify the standard likelihood to deal with censored data points
- For every K, we find a near optimal solution to the MLE, $\hat{\alpha}_K$, by carefully approximating the ζ-function
- Choose α_K which minimizes the Kolmogorov-Smirnov distance between the tail empirical distribution thresholded at K and our theoretical distribution
- This is nothing but a modified version of the Hill Estimator
- Method corresponds well to the traditional method of picking out approximate linear portion of Hill Plots
- Can iteratively improve estimate by improving approximation to ζ but the process is slow

Comments

SLASHDOT data analysis of out-degree



Simulated data with tail-index of in-degree = 3



Outline

Performance Comparison

- Both methods yield estimates which agree with each other
- When the true parameters are known, they outperform traditional methods
- Both methods yield ready made measures of uncertainty in estimation

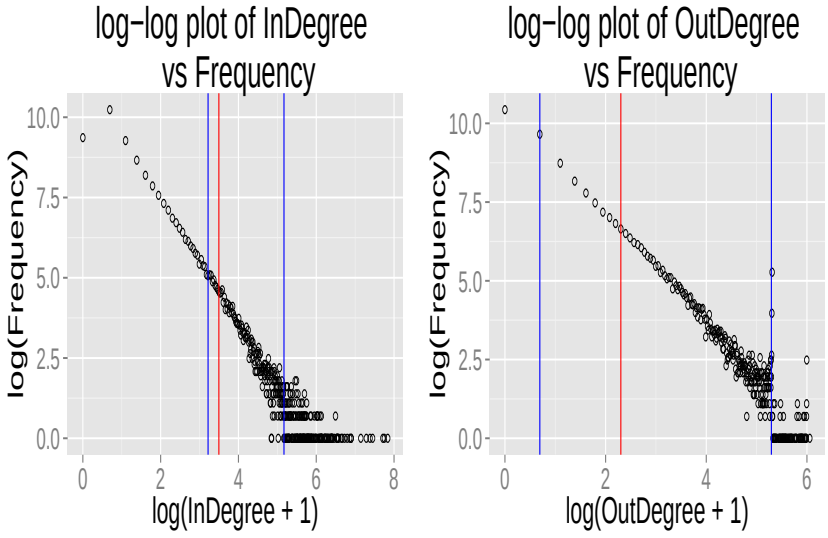
Outline

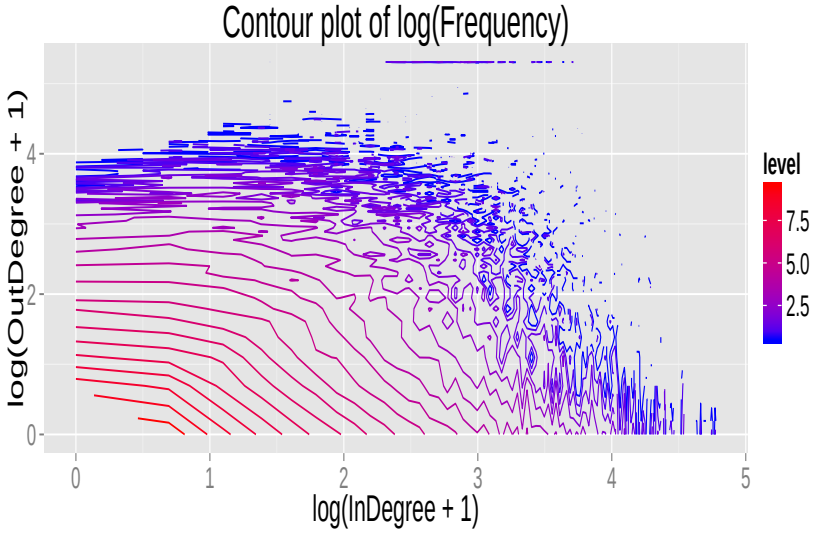
Towards Multivariate Methods

- No easily and reliably estimable statistical model is known
- Traditional methods only work when tail-indexes of both variables are the same which is rarely the case
- Extending the weighted regression method with terms to account for tail-dependence seem to work but still requires manual tuning
- Estimating measures like angular densities after standardizing marginals via rank-transforms also seem promising but again is not yet fully automated
- Both methods seem to be sensitive to model parameter selection
- Still qualitatively as well quantitatively, if we were to trust our not that robust multivariate methods, we can see distinct differences between the `SLASHDOT` dataset and simulated datasets from the model
- Simply cannot chalk-off these discrepancies to the fact that the `SLASHDOT` out-degrees are bounded

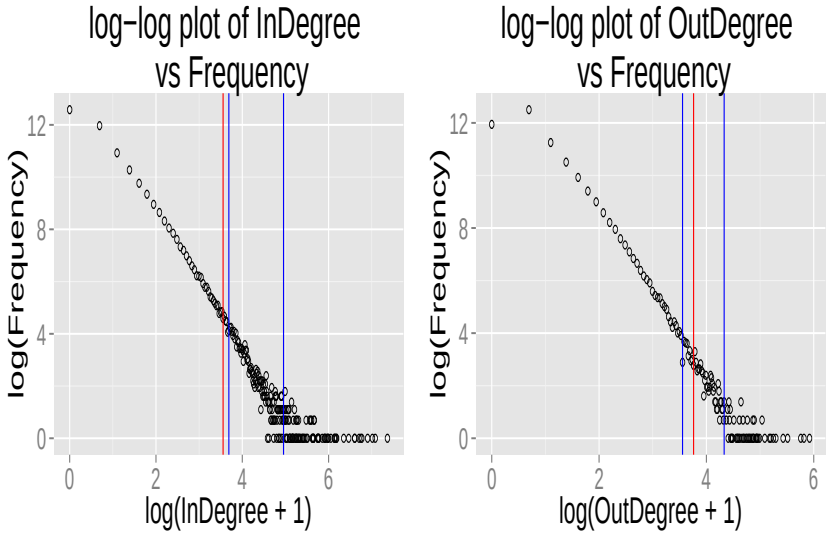
Caveat

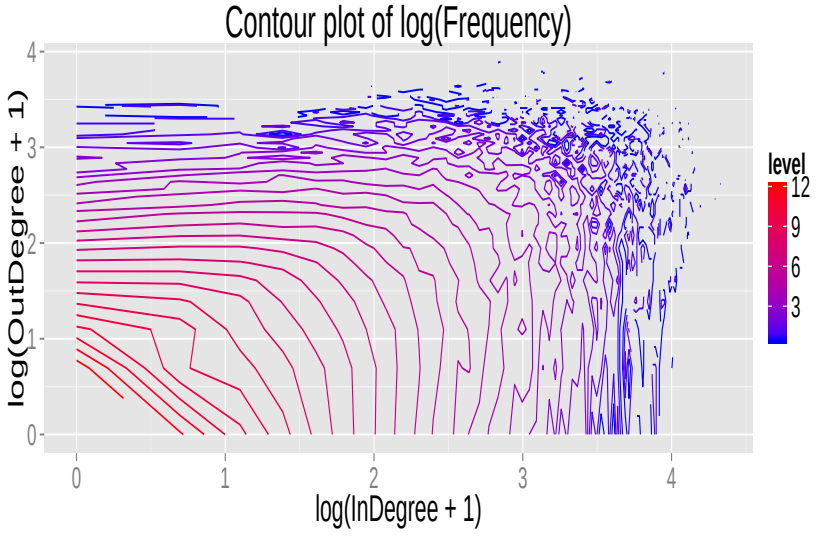
SLASHDOT data





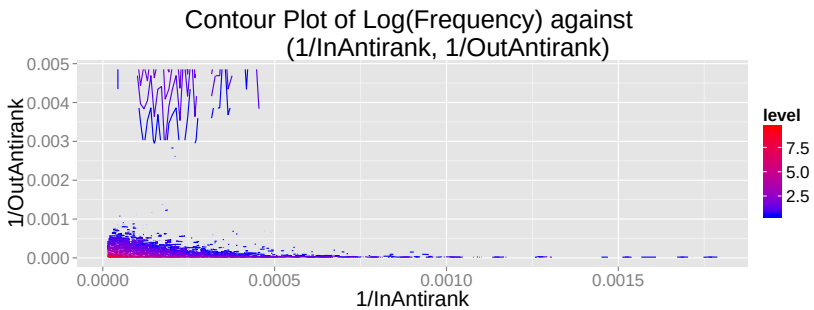
Simulated data

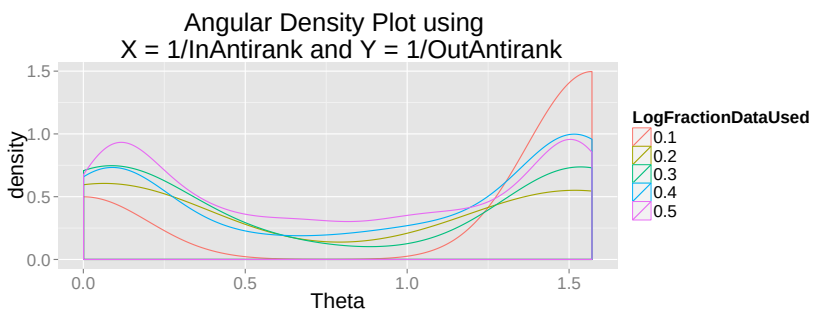




Caveat

SLASHDOT data





Simulated data