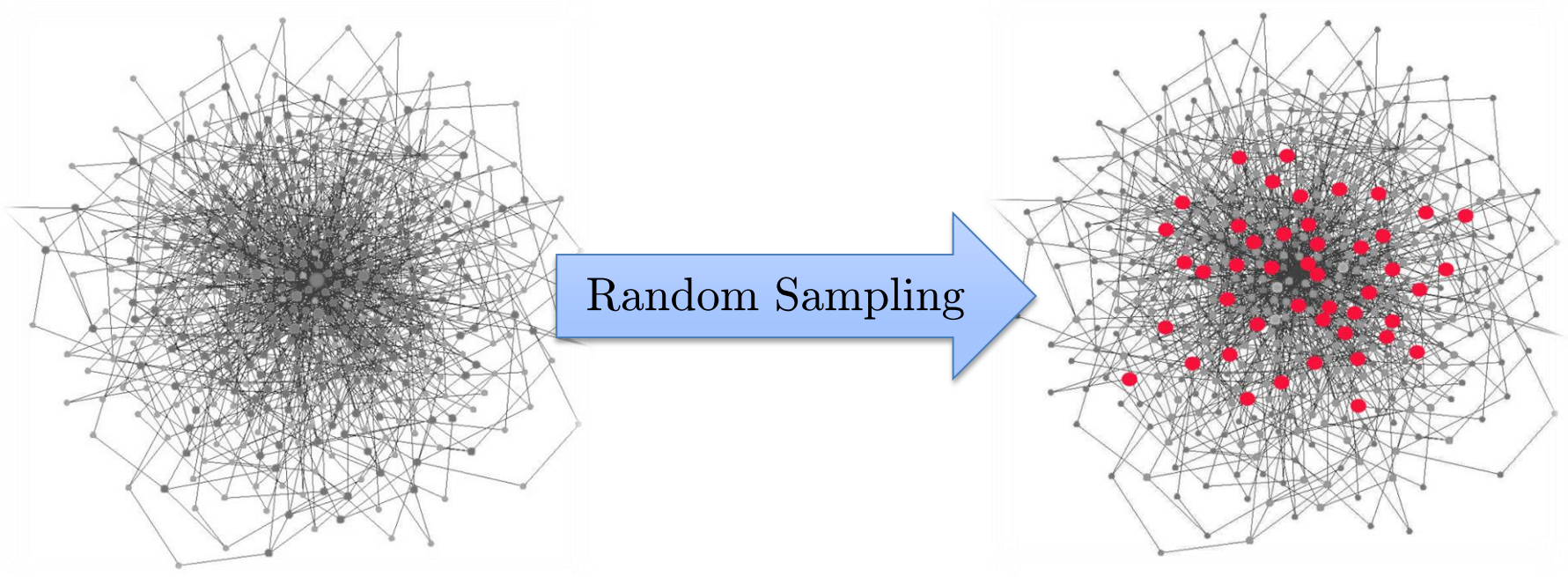
Distribution and Dependence of Extremes in Network Sampling Processes

Jithin K. Sreedharan*
with Konstantine Avrachenkov* and Natalia M. Markovich†

*INRIA Sophia Antipolis, France
†Institute of Control Sciences, Russian Academy of Sciences, Moscow

March 30, 2015



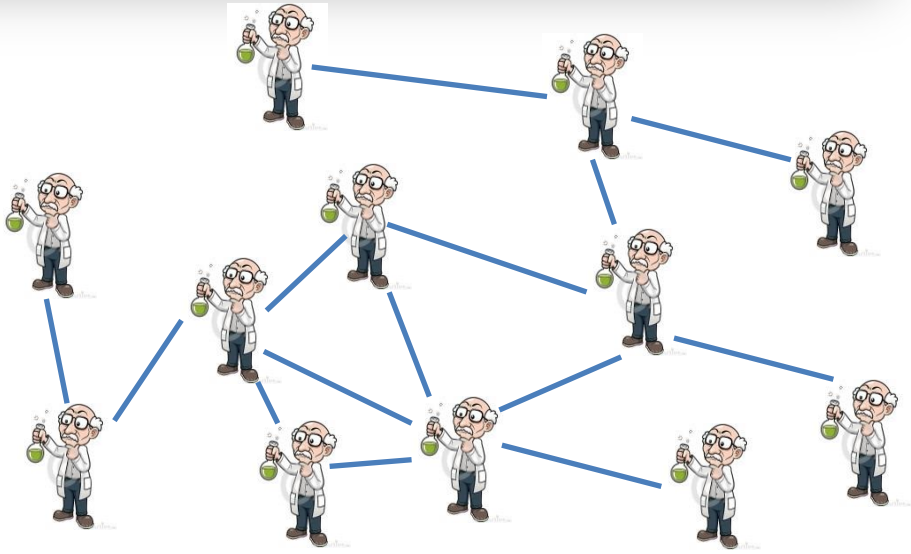
No complete picture a priori !

All we have: $X_1, X_2, \ldots, X_n$

Samples: any stationary (most likely dependent) sequence e.g. node ID's, degrees, number of followers or income of the nodes in OSN etc

Correlations in Graphs and Sampling

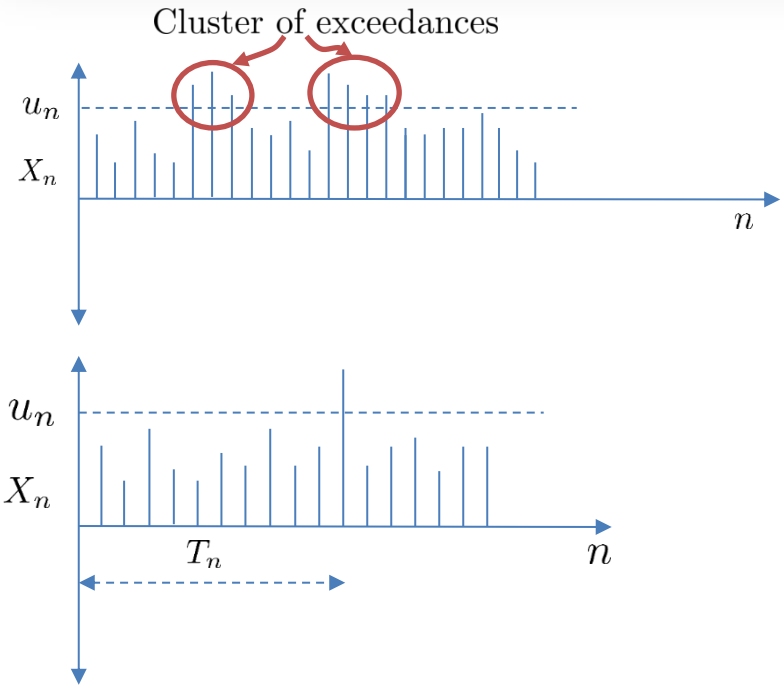
- Correlations in graph properties exist in real networks
  e.g: correlation in Coauthorship network
- Usually neglected in analysis of sampling algorithms



Effect of neglecting correlations:

- Assuming i.i.d. degrees, largest degree$\approx KN^{1/\gamma}$, N no. of nodes, γ tail index of Pareto distribution (N. Litvak et al, LNCS'12)
- Twitter graph (2012): N= 537M, $\gamma = 1.124$ for out-degree.
- Largest out-degree predicted is 59M. Actual largest out-degree is **22M**!

Questions We Address Here...



Statistical properties of clusters

First passage time

Kth largest value of samples and many more extremal properties

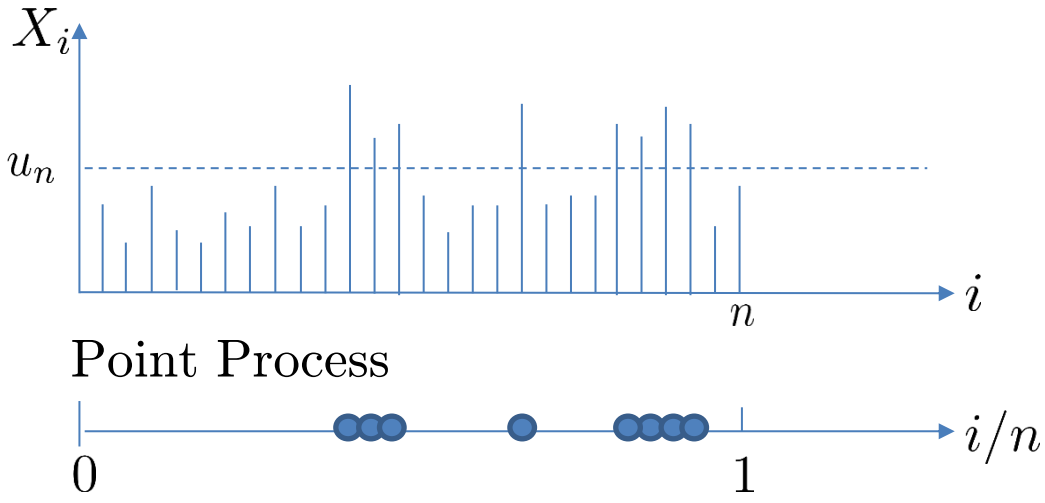
Is there a simple way to get information about many extremal properties? Ans: Extremal Index

Relation to Extreme Value Theory

Extremal Index (θ):

Defintion: If $\lim_{n\to\infty} E[\text{no of exceedances}] = \tau$,

$$P[\max(X_1, \ldots, X_n) \leq u_n] \to \exp(-\tau\theta)$$



Point process of exceedances →Compound poisson process (rate $\theta\tau$)
Tendency to form clusters

Extremal Index: Applications

$$P\{\max\{X_1, \ldots, X_n\} \leq x\} = F^{n\theta}(x) + o(1), n \to \infty$$

Gives maxima of the degree sequence with certain probability

Pareto case revisited:

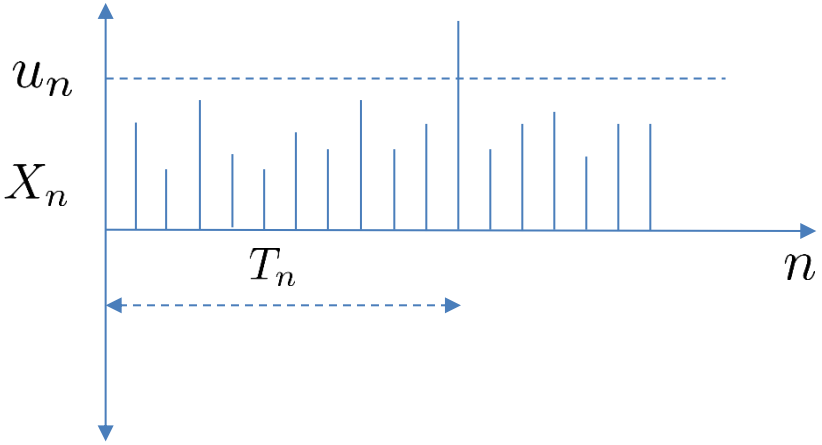
- i.i.d. degrees, largest degree$\approx KN^{1/\gamma}$, N no. of nodes, γ tail index of Pareto distribution (N. Litvak, LNCS'12)
- Stationary degree samples with EI, largest degree$\approx K(N\theta)^{1/\gamma}$

Extremal Index: Applications

First passage time:

$\frac{T_n}{n}$ asymptotically $\sim \text{Exp}(\theta\tau)$

$E(T_n) \approx n/(\theta\tau)$



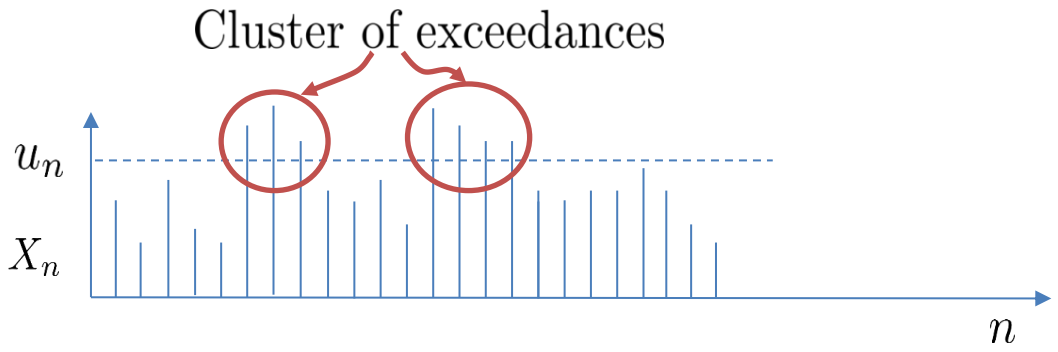
Lower the value of EI, more time to hit extreme levels

e.g. Pareto $P(X_i > u_n) = u_n^{-\alpha}, u_n = (n)^{1/\alpha}$ for $\tau = 1$

$$\implies E[T_n] \approx \frac{u_n^{\alpha}}{\theta}$$

Extremal Index: Applications

Relation to Mean Cluster Size:



$$\lim_{n\to\infty} E[\text{cluster size with } n \text{ samples}] = \frac{1}{\text{Extremal Index}}$$

Calculation of Extremal Index

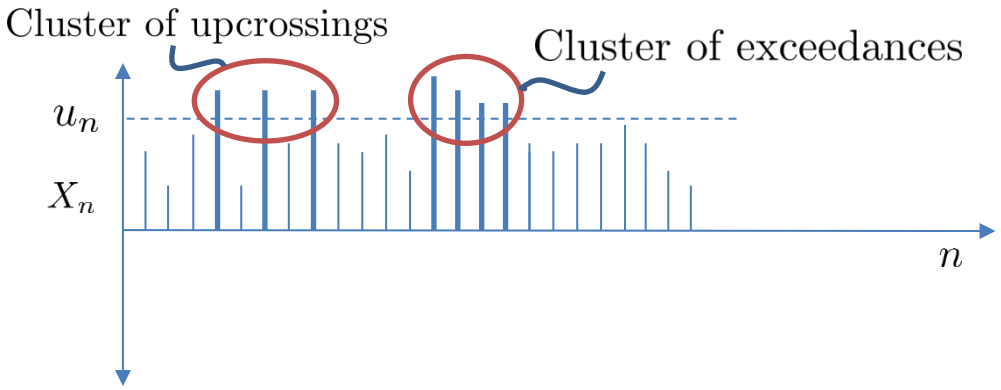
Two mixing conditions on the samples

Cond-1: Limits long range dependence

$$|P(\mathcal{AB}) - P(\mathcal{A})P(\mathcal{B})| \leq \alpha_n \quad \mathcal{A} \text{ and } \mathcal{B}\text{: events} \subset \{X_i \leq u_n\},\ l_n \text{ seperated}$$
$$l_n = o(n), \alpha_n \to 0$$

Stationary Markov samples or its measurable functions satisfy this

Cond-2:



$$\lim_{n\to\infty} P(\text{Cluster of upcrossings}) < \epsilon$$

$$\lim_{n\to\infty} P(\text{Cluster of exceedances}) \geq (1-\epsilon)$$

Proposition

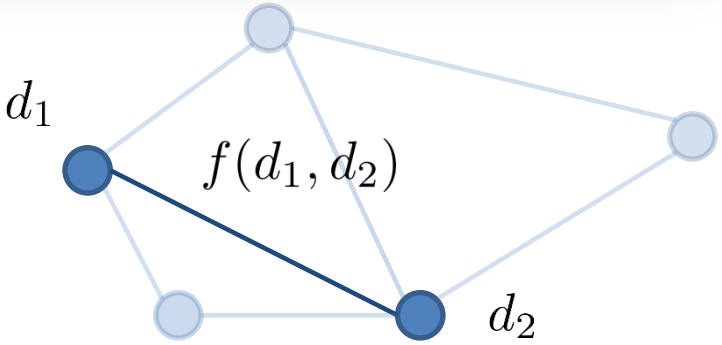
If the sampled sequence is stationary and satisfies mixing conditions, then Extremal Index

$$\theta =< \mathbf{1}, \nabla C > \Big|_{u=v=1} - 1,$$

$0 \leq \theta \leq 1$ and $C(u,v) = P(X_1 \leq F^{-1}(u), X_2 \leq F^{-1}(v))$ is the Copula.

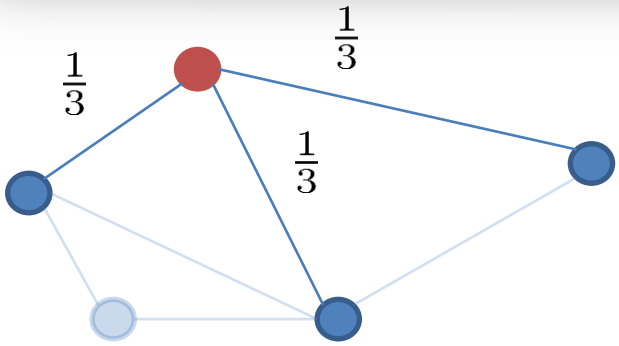
Degree Correlations

- Undirected and correlated
- $f(d_1, d_2)$ is enough to construct graph



- Crawling via Random Walks on vertices
- Degree sequence is a Hidden Markov chain
- What is the joint stationary distribution on degree state space?

Standard Random Walk



$$f_{RW}(d_{t+1}|d_t) \approx \frac{1}{d_t} \cdot \frac{E[D] f(d_t, d_{t+1})}{f_d(d_t)}$$

$$f_{RW}(d_{t+1}, d_t) \approx f(d_{t+1}, d_t)$$

Page Rank

$P(\text{head}) = c$

with c, follow RW
with $1 - c$, uniform node sampling

$$f_{PR}(d_{t+1}|d_t) \approx c f_{RW}(d_{t+1}|d_t) + (1-c) f_d(d_{t+1})$$

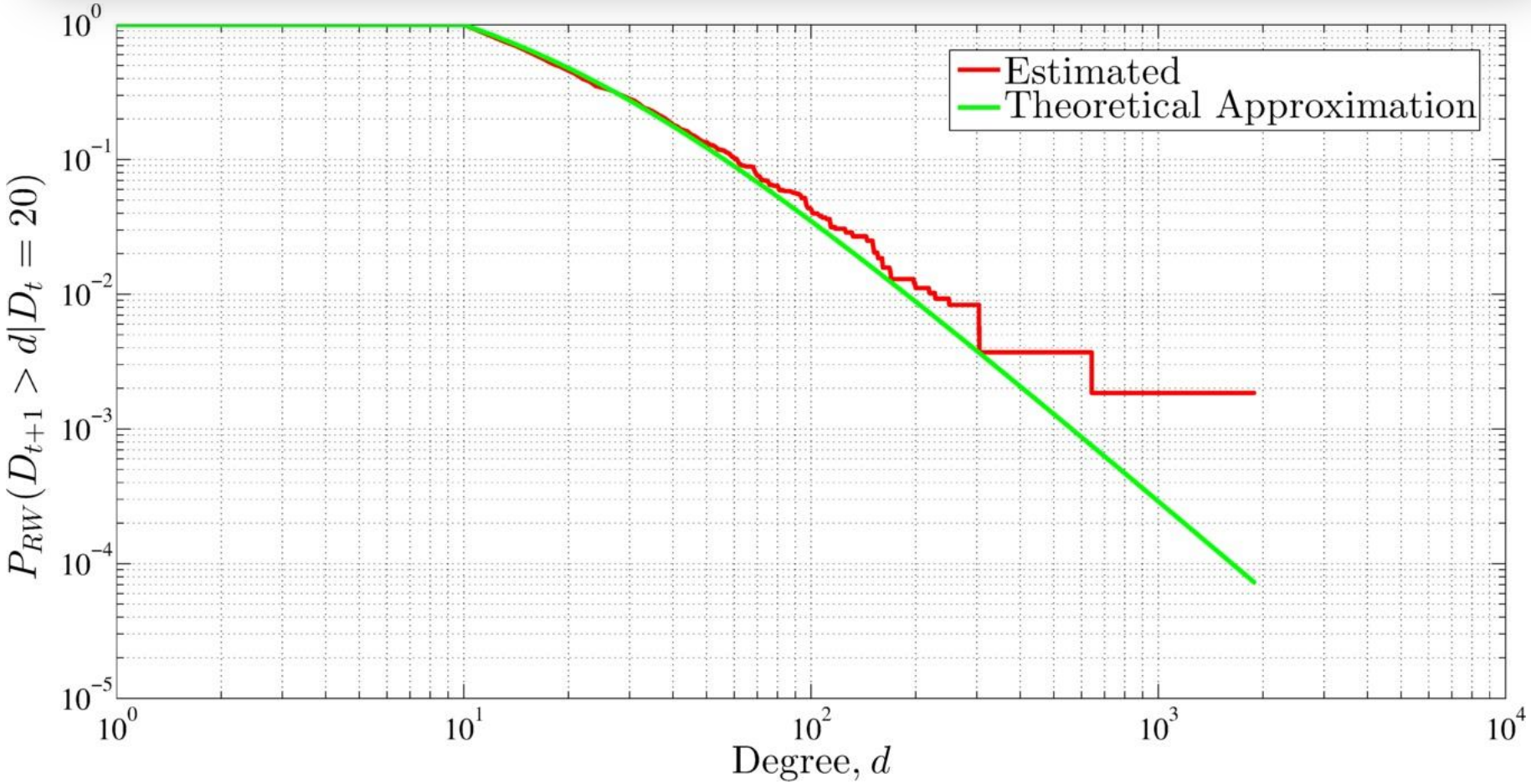
Random Walk with Jumps (RWJ)

$P(\text{head}) = c$

$$c = \frac{d_t}{d_t + \alpha}$$

$$f_{RWJ}(d_{t+1}, d_t) \approx \frac{E[D]}{E[D] + \alpha} f(d_{t+1}, d_t) + \frac{\alpha}{E[D] + \alpha} f_d(d_{t+1}) f_d(d_t)$$

Check of Meanfield Model in Random Walks



Degree correlation among neighbours, bivariate Pareto distributed

$$\bar{F}(d_1, d_2) = \left(1 + \frac{d_1 - \mu}{\sigma} + \frac{d_2 - \mu}{\sigma}\right)^{-\gamma} \quad \mu = 10, \sigma = 15, \gamma = 1.2$$

Extremal Index for Bivariate Pareto Model

$$\bar{F}(d_1, d_2) \sim \left(1 + \frac{d_1 - \mu}{\sigma} + \frac{d_2 - \mu}{\sigma}\right)^{-\gamma}$$

Random Walk: $\mathrm{EI} = 1 - 1/2^{\gamma}$

Random Walk with Jumps: $\mathrm{EI} = 1 - \dfrac{E[D]}{E[D] + \alpha} 2^{\gamma}$

PageRank: $\mathrm{EI} \geq (1 - c)$

(for any kind of degree correlations)

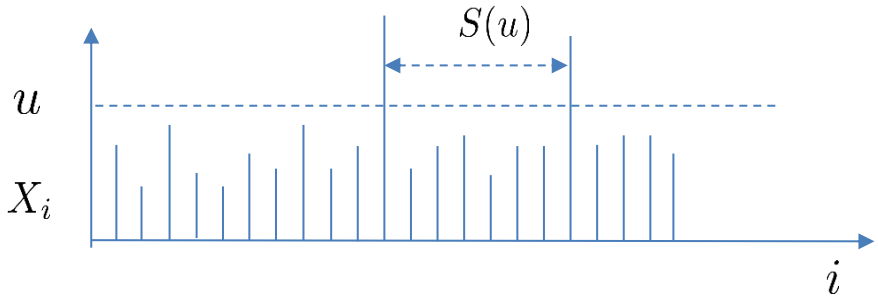
Estimation of Extremal Index

Empirical Copula based estimator:

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\frac{R_i^X}{n+1} \le u, \frac{R_i^Y}{n+1} \le v\right)$$

EI: slope at (1; 1),Linear least square fitting & numerical differentiation

Intervals Estimator:



Based on

$$\bar{F}(u)S(u) \xrightarrow{d} T_\theta \quad T_\theta \sim \begin{cases} 1-\theta & \text{when } \theta = 0 \\ \text{Exp}(\theta) & \text{when } \theta > 0 \end{cases}$$

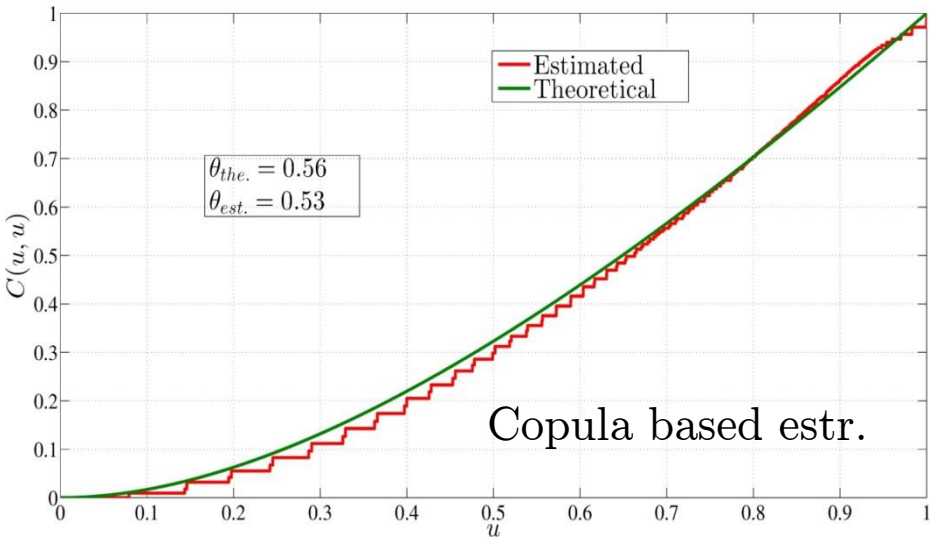
Use $E(T_\theta^2) = 2/\theta$ to obtain estimates

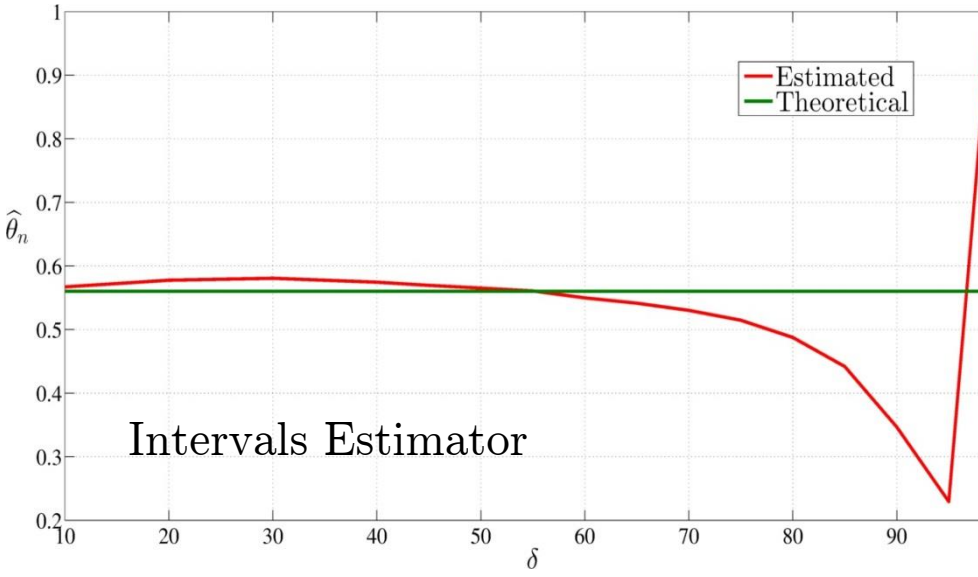
Numerical Results: Synthetic Graphs

Degree correlation between neighbours

$$\bar{F}(d_1, d_2) = \left(1 + \frac{d_1 - \mu}{\sigma} + \frac{d_2 - \mu}{\sigma}\right)^{-\gamma} \quad \mu = 10, \sigma = 15, \gamma = 1.2$$

EI	Analysis	Copula based estimator	Intervals Estimator
Synthetic graph (5K Nodes)	0.56	0.53	0.58





Numerical Results: Real Graphs

EI	Copula based estimator	Intervals Estimator
DBLP (32K Nodes,1.1M Edges)	0.29	0.25
Enron Email (37K Nodes,368K Edges)	0.61	0.62

Conclusions

- Associated Extremal Value Theory of stationary sequence to sampling of large graphs
- For any general stationary samples meeting two mixing conditions, knowledge of bivariate distribution or bivariate copula is sufficient to derive many extremal properties
- Extremal Index (EI) encapsulates this relation
- Applications of EI to many relevant extrems:
  - First hitting time
  - Order statistics
  - Mean cluster size
- Modeled correlation in degrees of adjacent nodes and random walk in degree state space
- Estimates of EI for synthetic graph with degree correlations and find a good match with theory
- Estimated EI for two real world networks

Thank You!