Distribution and dependence of extremes in PageRank-type processes

Jithin K. Sreedharan

INRIA Sophia Antipolis, Project MAESTRO

May 15, 2014

- Correlations in the degree sequence generated from sampling algorithms
- Clusters of nodes in terms of degrees due to dependence between nodes
- Stochastic nature of clusters
- Abstract correlation and cluster statistics into a single parameter

Extremal index θ

- Characterizes dependence in a stationary sequence
- $\{X_n\}_{n\geq 1}$ i.i.d. with d.f. F, $M_n = \max\{X_1, \ldots, X_n\}$
  - Let $0 < \tau < \infty$, $u_n = u_n(\tau)$ be a sequence of real numbers such that $\lim_{n\to\infty} n(1 - F(u_n)) = \tau$, then

$$P(M_n \leq u_n) \to \exp(-\tau)$$

  - Limit of the point processes of exceedances is homogeneous Poisson process
- $\{X_n\}_{n\geq 1}$ stationary,
  - $P\{M_n \leq u_n\} \to \exp(-\tau\theta)$ for $0 \leq \theta \leq 1$.
  - Limit of the point processes of exceedances is compound Poisson $\implies$ Exceedances of high threshold values u_n tend to occur in clusters for dependent data.

Model

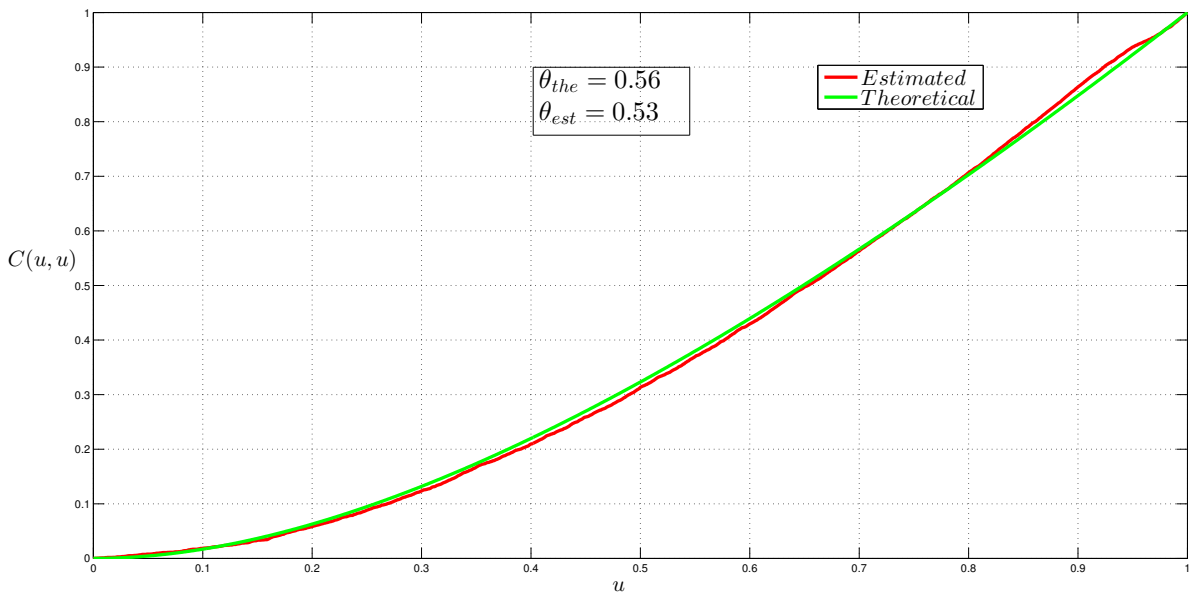
- Undirected and correlated
- Joint degree-degree p.d.f. $f(d_1, d_2)$
- Pr (chosen edge has the end degrees $d_1 \leq d \leq d_1 + \Delta(d_1)$ and $d_2 \leq d' \leq d_2 + \Delta(d_2)$)
$$= (2 - \delta_{d_1 d_2}) f(d_1, d_2) \Delta(d_1) \Delta(d_2)$$
- Degree distribution $f_d(d_1)$ from the marginal
- Network is unknown: only API requests are allowed

Random Walk based graph exploration algorithms

- Transition kernels defined for degree state space (not for vertex set)
- Standard Random Walk: Next node to visit is chosen uniformly among the neighbours.
  - $f_{RW}(d_{t+1}|d_t) = \frac{E[D] f(d_t, d_{t+1})}{d_t f_d(d_t)}$, $f_{RW}(d_{t+1}, d_t) = f(d_{t+1}, d_t)$
- PageRank: With $1 - c$, samples random node with uniform distribution and with c, follows standard Random walk.
  - $f_{PR}(d_{t+1}|d_t) = c\ f_{RW}(d_{t+1}|d_t) + (1 - c) f_d(d_{t+1})$
- Randomwalk with Jumps: $c = d_t/(d_t + \alpha)$
  - $f_{JP}(d_{t+1}, d_t) = \frac{E[D] f(d_{t+1}, d_t) + \alpha f_d(d_{t+1}) f_d(d_t)}{E[D] + \alpha}$

Calculation of θ

- $\theta = \lim_{x \to 1} \frac{x - C(x,x)}{1-x} = C'(1,1) - 1$
  $C(x_1, \ldots, x_d)$ is Copula function ($[0,1]^d \to [0,1]$)
- $C(u,u) = P_{\mathscr{X}}(D_1 \leq F_{\mathscr{X}}^{-1}(u), D_2 \leq F_{\mathscr{X}}^{-1}(u))$, $\mathscr{X}$ is RW, PR or RWJ, $F_{\mathscr{X}}^{-1}()$ is the inverse of stationary distribution function.
- $\bar{F}(d_1, d_2) = \left(1 + \frac{d_1 - \mu_1}{\sigma_1} + \frac{d_2 - \mu_2}{\sigma_2}\right)^{-\gamma}$
  RW: $\theta = 1 - 1/2^{\gamma}$, RWJ: $\theta = 1 - 1/2^{\gamma}$ for $\alpha = [0, \infty)$,$= 1$ for $\alpha = \infty$



Applications

- Relation to the mean cluster size: Cluster size $\xi = \sum_{i=1}^{r_n} 1(X_i > u_n)$ is no. of exceedances in a block of size r_n ($r_n = o(n)$) in n when there is atleast one exceedance. $\theta = (E\xi)^{-1}$.
- $P\{M_n \leq x\} = F^{n\theta}(x) + o(1), n \to \infty$. Helpful to find the highest degree with certain probability.

Thank You!