Avrachenkov *et al. Computational Social Networks* (2015) 2:12
DOI 10.1186/s40649-015-0018-3

Computational Social Networks
a SpringerOpen Journal

RESEARCH Open Access

Distribution and dependence of extremes in network sampling processes

Konstantin Avrachenkov[1], Natalia M. Markovich[2] and Jithin K. Sreedharan[1]*

*Correspondence:
jithin.sreedharan@inria.fr
[1] INRIA Sophia Antipolis 2004, route des Lucioles - BP 93, 06902 Sophia Antipolis Cedex, France
Full list of author information is available at the end of the article

Abstract

We explore the dependence structure in the sampled sequence of complex networks. We consider randomized algorithms to sample the nodes and study extremal properties in any associated stationary sequence of characteristics of interest like node degrees, number of followers, or income of the nodes in online social networks, which satisfy two mixing conditions. Several useful extremes of the sampled sequence like the *k*th largest value, clusters of exceedances over a threshold, and first hitting time of a large value are investigated. We abstract the dependence and the statistics of extremes into a single parameter that appears in extreme value theory called extremal index (EI). In this work, we derive this parameter analytically and also estimate it empirically. We propose the use of EI as a parameter to compare different sampling procedures. As a specific example, degree correlations between neighboring nodes are studied in detail with three prominent random walks as sampling techniques.

Keywords: Network sampling; Extreme value theory; Extremal index; Random walks on graph

Introduction

Data from real complex networks shows that correlations exist in various forms, for instance the existence of social relationships and interests in social networks. Degree correlations between neighbors, correlations in income, followers of users, and number of likes of specific pages in social networks are some examples, to name a few. These kind of correlations have several implications in network structure. For example, degree-degree correlation manifests itself in assortativity or disassortativity of the network [1].

We consider very large complex networks where it is impractical to have a complete picture *a priori*. Crawling or sampling techniques can be employed in practice to explore such networks by making the use of application programming interface (API) calls or HTML scrapping. We look into randomized sampling techniques which generate stationary samples. As an example, random walk-based algorithms are in use in many cases because of several advantages offered by them [2, 3].

We focus on the extremal properties in the correlated and stationary sequence of characteristics of interest $X_1, \ldots, X_n$ which is a function of the node sequence, the one actually generated by sampling algorithms. The characteristics of interest, for instance, can be node degrees, node income, number of followers of the node in online social networks (OSN), etc. Among the properties, clusters of exceedances of such sequences over high thresholds are studied in particular. The cluster of exceedances is roughly defined as the consecutive exceedances of $\{X_n\}$ over the threshold $\{u_n\}$ between two consecutive non-exceedances. For more rigorous definitions, see [4–6]. It is important to investigate stochastic nature of extremes since it allows us to collect statistics or opinions more effectively in the clustered (network sampling) process.

The dependence structure of sampled sequence exceeding sufficiently high thresholds is measured using a parameter called extremal index (EI), θ. It is defined in extremal value theory as follows.

Definition 1. *([7], p. 53)* The stationary sequence $\{X_n\}_{n\geq 1}$, with F as the marginal distribution function and $M_n = \max\{X_1, ..., X_n\}$, is said to have the extremal index $\theta \in [0,1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers (thresholds) $u_n = u_n(\tau)$ such that

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau \text{ and} \tag{1}$$

$$\lim_{n\to\infty} \mathrm{P}\{M_n \leq u_n\} = e^{-\theta\tau}. \tag{2}$$

The maxima M_n is related to EI more clearly as ([4], p. 381)[1]

$$\mathrm{P}\{M_n \leq u_n\} = F^{n\theta}(u_n) + o(1). \tag{3}$$

When $\{X_n\}_{n\geq 1}$ is independent and identically distributed (i.i.d.) (for instance, in uniform independent node sampling), $\theta = 1$ and point processes of exceedances over threshold u_n converges weakly to homogeneous Poisson process with rate τ as $n \to \infty$ ([4], chapter 5). But when $0 \leq \theta < 1$, point processes of exceedances converges weakly to compound Poisson process with rate $\theta\tau$ and this implies that exceedances of high threshold values u_n tend to occur in clusters for dependent data ([4], chapter 10).

EI has many useful interpretations and applications like

- Finding distribution of order statistics of the sampled sequence. These can be used to find quantiles and predict the kth largest value which arise with a certain probability. Specifically for the distribution of maxima, Eq. 3 is available and the quantile of maxima is proportional to EI. Hence in case of samples with lower EI, lower values of maxima can be expected. When sampled sequence is the sequence of node degrees, these give many useful results.
- Close relation to the distribution and expectation of the size of clusters of exceedances (see for e.g. [4, 6]).
- Characterization of the first hitting time of the sampled sequence to (u_n, ∞). Thus in case of applications where the aim is to detect large values of samples quickly, without actually employing sampling (which might be very costly), we can compare different sampling procedures by EI: smaller EI leads to longer waiting of the first hitting time.

These interpretations are explained later in the paper. The network topology as well as the sampling method determine the stationary distribution of the characteristics of interest under a sampling technique and is reflected on the EI.

Our contributions

The main contributions in this work are as follows. We associated extremal value theory of stationary sequences to sampling of large complex networks, and we study the extremal and clustering properties of the sampling process due to dependencies. In order to facilitate a painless future study of correlations and clusters of samples in large networks, we propose to abstract the extremal properties into a single and handy parameter, EI. For any general stationary samples meeting two mixing conditions, we find that knowledge of bivariate distribution or bivariate copula is sufficient to compute EI analytically and thereby deriving many extremal properties. Several useful applications of EI (first hitting time, order statistics, and mean cluster size) to analyze large graphs, known only through sampled sequences, are proposed. Degree correlations are explained in detail with a random graph model for which joint degree distribution exists for neighbor nodes. Three different random walk-based algorithms that are widely discussed in literature (see [2] and the references therein) are then revised for degree state space, and EI is calculated when the joint degree distribution is bivariate Pareto. We establish a general lower bound for EI in PageRank processes irrespective of the degree correlation model. Finally, using two estimation techniques, EI is numerically computed for a synthetic graph with neighbor degrees correlated and for two real networks (Enron email network and DBLP network).

The paper is organized as follows. In section “Calculation of extremal index (EI)”, methods to derive EI are presented. Section “Degree correlations” considers the case of degree correlations. In section “Description of the configuration model with degree-degree correlation”, the graph model and correlated graph generation technique are presented. Section “Description of random walk-based sampling processes” explains the different types of random walks studied and derives associated transition kernels and joint degree distributions. EI is calculated for different sampling techniques later in section “Extremal index for bivariate Pareto degree correlation”. In section “Applications of extremal index in network sampling processes”, we provide several applications of EI in graph sampling techniques. In section “Estimation of extremal index and numerical results”, we estimate EI and perform numerical comparisons. Finally, section “Conclusions” concludes the paper.

A shorter version of this work has appeared in [8].

Calculation of extremal index (EI)

We consider networks represented by an undirected graph G with N vertices and M edges. Since the networks under consideration are huge, we assume it is impossible to describe them completely, i.e., no adjacency matrix is given beforehand. Assume any randomized sampling procedure is employed and let the sampled sequence $\{X_i\}$ be any general sequence.

This section explains a way to calculate EI from the bivariate joint distribution if the sampled sequence admits two mixing conditions.

Condition $(D(u_n))$**.**

$$\begin{aligned}\Big|\mathrm{P}(X_{i_1} \le u_n, \ldots, X_{i_p} \le u_n, X_{j_1} \le u_n, \ldots, X_{j_q} \le u_n) \\ -\mathrm{P}(X_{i_1} \le u_n, \ldots, X_{i_p} \le u_n)\mathrm{P}(X_{j_1} \le u_n, \ldots, X_{j_q} \le u_n)\Big| \le \alpha_{n,l_n},\end{aligned}$$

where $\alpha_{n,l_n} \to 0$ for some sequence $l_n = o(n)$ as $n \to \infty$, for any integers $i_1 \le \ldots < i_p < j_1 < \ldots \le j_q$ with $j_1 - i_p > l_n$.

Condition $(D''(u_n))$**.**

$$\lim_{n\to\infty} n \sum_{m=3}^{r_n} \mathrm{P}(X_1 > u_n \geq X_2, X_m > u_n) = 0,$$

where $(n/r_n)\alpha_{n,l_n} \to 0$ and $l_n/r_n \to 0$ with α_{n,l_n}, l_n as in Condition $D(u_n)$ and r_n as $o(n)$.

Let $C(u,v)$ be a bivariate copula [9] ($[0,1]^2 \to [0,1]$) and $\underline{1}\cdot\nabla C(u,v)$ is its directional derivative along the direction $(1,1)$. Using Sklar's theorem ([9], p. 18), with F as the marginal stationary distribution function of the sampling process, the copula is given by

$$C(u,v) = \mathrm{P}(X_1 \leq F^{-1}(u), X_2 \leq F^{-1}(v)),$$

where F^{-1} denotes the inverse function of F. This representation is unique if the stationary distribution $F(x)$ is continuous.

Theorem 1. *If the sampled sequence is stationary and satisfies conditions $D(u_n)$ and $D''(u_n)$, and the limits in Eqs. 1 and 2 take place, then the extremal index is given by*

$$\theta = \underline{1}\cdot\nabla C(1,1) - 1, \tag{4}$$

and $0 \leq \theta \leq 1$.

Proof. For a stationary sequence $\{X_n\}$ holding conditions $D(u_n)$ and $D''(u_n)$, if the limits in Eqs. 1 and 2 take place, $\theta = \lim_{n\to\infty} \mathrm{P}(X_2 \leq u_n | X_1 > u_n)$ [10]. Then, we have

$$\begin{aligned}\theta &= \lim_{n\to\infty} \frac{\mathrm{P}(X_2 \leq u_n, X_1 > u_n)}{\mathrm{P}(X_1 > u_n)}\\ &= \lim_{n\to\infty} \frac{\mathrm{P}(X_2 \leq u_n) - \mathrm{P}(X_1 \leq u_n, X_2 \leq u_n)}{\mathrm{P}(X_1 > u_n)}\\ &= \lim_{n\to\infty} \frac{\mathrm{P}(X_2 \leq u_n) - C\big(\mathrm{P}(X_1 \leq u_n), \mathrm{P}(X_2 \leq u_n)\big)}{1 - \mathrm{P}(X_1 \leq u_n)}\\ &= \lim_{x\to 1} \frac{x - C(x,x)}{1-x}\\ &= \underline{1}\cdot\nabla C(1,1) - 1,\end{aligned}$$

which completes the proof. □

Remark 1. *Condition $D''(u_n)$ can be made weaker to $D^{(k)}(u_n)$ presented in [11],*

$$\lim_{n\to\infty} n\mathrm{P}\left(X_1 > u_n \geq \max_{2\leq i\leq k} X_i, \max_{k+1\leq j\leq r_n} X_j > u_n\right) = 0,$$

where r_n is defined as in $D''(u_n)$. For the stationary sequence, $D^{(2)}(u_n)$ is identical to $D''(u_n)$. If we assume $D^{(k)}$ is satisfied for some $k \geq 2$ along with $D(u_n)$, then following the proof of Theorem 1, EI can be derived as

$$\theta = \underline{1}\cdot\nabla C_k(1,\ldots,1) - \underline{1}\cdot\nabla C_{k-1}(1,\ldots,1),$$

where $C_k(x_1,\ldots,x_k)$ represents the copula of k-dimensional vector $(x_1,\ldots,x_k)$, C_{k-1} is its $(k-1)$th marginal, $C_{k-1}(x) = C_{k-1}(x_1,\ldots,x_{k-1},1)$, and $\underline{1}\cdot\nabla C_k(x_1,\ldots,x_k)$ denotes the directional derivative of $C_k(x_1,\ldots,x_k)$ along the k-dimensional vector $(1,1,\ldots,1)$.

In some cases, it is easy to work with the joint tail distribution. Survival copula $\widehat{C}(\cdot,\cdot)$ which corresponds to

$$\mathrm{P}(X_1 > x, X_2 > x) = \widehat{C}(\overline{F}(x), \overline{F}(x)),$$

with $\overline{F}(x) = 1 - F(x)$, can also be used to calculate θ. It is related to copula as $\widehat{C}(u,u) = C(1-u, 1-u) + 2u - 1$ ([9], p. 32). Hence, $\theta = \underline{1} \cdot \nabla C(1,1) - 1 = 1 - \underline{1} \cdot \nabla\, \widehat{C}(0,0)$.

Lower tail dependence function of survival copula is defined as [12]

$$\lambda(u_1,u_2) = \lim_{t \to 0^+} \frac{\widehat{C}(tu_1, tu_2)}{t}.$$

Hence, $\underline{1} \cdot \nabla\, \widehat{C}(0,0) = \lambda(1,1)$. λ can be calculated for different copula families. In particular, if $\widehat{C}$ is a bivariate Archimedean copula, then it can be represented as $\widehat{C}(u_1,u_2) = \psi\left(\psi^{-1}(u_1) + \psi^{-1}(u_2)\right)$, where ψ is the generator function and ψ^{-1} is its inverse with $\psi : [0,\infty] \to [0,1]$ meeting several other conditions. If ψ is a regularly varying distribution with index $-\beta$, $\beta > 0$, then $\lambda(x_1,x_2) = \left(x_1^{-\beta^{-1}} + x_2^{-\beta^{-1}}\right)^{-\beta}$ and (X_1, X_2) has a bivariate regularly varying distribution [12]. Therefore, for Archimedean copula family, EI is given by

$$\theta = 1 - 1/2^{\beta}. \tag{5}$$

As an example, bivariate Pareto distribution of the form $\mathrm{P}(X_1 > x_1, X_2 > x_2) = (1 + x_1 + x_2)^{-\gamma}$, $\gamma > 0$ has Archimedean copula with generator function $\psi(x) = (1+x)^{-\gamma}$. This gives $\theta = 1 - 1/2^{\gamma}$. Bivariate exponential distribution of the form

$$\mathrm{P}(X_1 > x_1, X_2 > x_2) = 1 - e^{-x_1} - e^{-x_2} + e^{-(x_1 + x_2 + \eta x_1 x_2)},$$

$0 \leq \eta \leq 1$, also admits Archimedean copula.

Check of conditions $D(u_n)$ and $D''(u_n)$ for functions of Markov samples

If the sampling technique is assumed to be based on a Markov chain and the sampled sequence is a measurable function of stationary Markov samples, then such a sequence is stationary and [13] proved that another mixing condition $\mathrm{AIM}(u_n)$ which implies $D(u_n)$ is satisfied. Condition $D''(u_n)$ allows clusters with consecutive exceedances and eliminates the possibility of clusters with upcrossing of the threshold u_n $(X_i \leq u_n < X_{i+1})$. Hence in those cases, where it is tedious to check the condition $D''(u_n)$ theoretically, we can use numerical procedures to measure ratio of number of consecutive exceedances to number of exceedances and the ratio of number of upcrossings to number of consecutive exceedances in small intervals. Such an example is provided in section "Extremal index for bivariate Pareto degree correlation".

Remark 2. *The EI derived in [14] has the same expression as in Eq. 4. But [14] assumes $\{X_n\}$ is sampled from a first-order Markov chain. We relax the Markov property requirement to D and D'' conditions, and the example below demonstrates a hidden Markov chain that can satisfy D and D''.*

Let us consider a hidden Markov chain with the observations $\{X_k\}_{k \geq 1}$ and the underlying homogeneous Markov chain as $\{Y_k\}_{k \geq 1}$ in stationarity. The underlying Markov chain is finite state space, but the conditional distributions of the observations $\mathrm{P}(X_k \leq x | Y_k = y) = F_y(x)$ have infinite support and condition Eq. 1 holds for $F_y(x)$.

Proposition 1. *When condition Eq. 1 holds for $F_y(x)$, the observation sequence $\{X_k\}_{k\geq 1}$ of the hidden Markov chain satisfies Condition D''.*

Proof. Let the transition probability matrix of $\{Y_k\}_{k\geq 1}$ be P (with $\mathrm{P}(Y_2 = j|Y_1 = i) = P_{ij}$) and the stationary distribution be π (with $\mathrm{P}(Y_1 = i) = \pi_i$). We have,

$$\begin{aligned}
&\mathrm{P}(X_1 > u_n \geq X_2, X_m > u_n)\\
&= \sum_{i,j,k} \mathrm{P}(Y_1 = i, Y_2 = j, Y_m = k)\mathrm{P}(X_1 > u_n \geq X_2, X_m > u_n|Y_1, Y_2, Y_m)\\
&= \sum_{i,j,k} \pi_i P_{ij} P_{jk}^{(m-2)} P_i(X_1 > u_n) P_j(X_2 \leq u_n) P_k(X_m > u_n)\\
&\sim \sum_{i,j,k} \pi_i P_{ij} P_{jk}^{(m-2)} \frac{\tau}{n}\left(1 - \frac{\tau}{n}\right)\frac{\tau}{n}, \quad n \to \infty.
\end{aligned}$$

Thus

$$\lim_{n\to\infty} n \sum_{m=3}^{r_n} \mathrm{P}(X_1 > u_n \geq X_2, X_m > u_n) = 0,$$

since $r_n = o(n)$, which completes the proof. □

Proposition 1 essentially tells that if the graph is explored by a Markov chain-based sampling algorithm and the samples are taken as any measurable functions of the underlying Markov chain, satisfying Condition (1) then Condition D'' holds. Measurable functions, for example, can represent various attributes of the nodes such as income or frequency of messages in social networks.

Degree correlations

The techniques established in section "Calculation of extremal index (EI)" are very general, applicable to any sampling techniques and any sequence of samples which satisfy certain conditions. In this section, we illustrate the calculation of EI for dependencies among degrees. We revise different sampling techniques. We denote the sampled sequence $\{X_i\}$ as $\{D_i\}$ in this section, since the sampled degree sequence will be a case study in this section.

Description of the configuration model with degree-degree correlation

To test the proposed approaches and the derived formulas, we use a synthetically generated configuration type random graph with a given joint degree-degree probability distribution, which takes into account correlation in degrees between neighbor nodes. The dependence structure in the graph is described by the joint degree-degree probability density function $f(d_1, d_2)$ with d_1 and d_2 indicating the degrees of adjacent nodes or equivalently by the corresponding tail distribution function $\overline{F}(d_1, d_2) = \mathrm{P}(D_1 \geq d_1, D_2 \geq d_2)$ with D_1 and D_2 representing the degree random variables (see e.g., [1, 15, 16]).

The probability that a randomly chosen edge has the end vertices with degrees $d_1 \leq d \leq d_1 + \Delta(d_1)$ and $d_2 \leq d \leq d_2 + \Delta(d_2)$ is $(2 - \delta_{d_1 d_2}) f(d_1, d_2)\Delta(d_1)\Delta(d_2)$. Here $\delta_{d_1 d_2} = 1$ if $d_1 = d_2$, otherwise $\delta_{d_1 d_2} = 0$. The multiplying factor 2 appears on the above expression when $d_1 \neq d_2$ because of the symmetry in $f(d_1, d_2)$, $f(d_1, d_2) = f(d_2, d_1)$ due to the undirected nature of the underlying graph and the fact that both $f(d_1, d_2)$ and $f(d_2, d_1)$ contribute to the edge probability under consideration.

The degree density $f_d(d_1)$ can be related to the marginal of $f(d_1, d_2)$ as follows:

$$f(d_1) = \int_{d_2} f(d_1, d_2) d(d_2) \approx \frac{d_1 f_d(d_1)}{\mathrm{E}[D]}, \tag{6}$$

where E[D] denotes the mean node degree,

$$\mathrm{E}[D] = \left[\int \int \left(\frac{f(d_1, d_2)}{d_1} \right) d(d_1) d(d_2) \right]^{-1}.$$

$f(.)$ can be interpreted as the degree density of a vertex reached by following a randomly chosen edge. The approximation for $f(d_1)$ is obtained as follows: in the right-hand side (R.H.S.) of Eq. 6, roughly, $d_1 f_d(d_1) N$ is the number of half edges from nodes with degree around d_1 and $\mathrm{E}[D] N$ is the total number of half edges. For discrete distributions, Eq. 6 becomes equality.

From the above description, it can be noted that the knowledge of $f(d_1, d_2)$ is sufficient to describe this random graph model and for its generation.

Most of the results in this paper are derived assuming continuous probability distributions for $f(d_1, d_2)$ and $f_d(d_1)$ because an easy and unique way to calculate EI exists for continuous distributions in our setup (more details in section "Calculation of extremal index (EI)"). Also the EI might not exist for many discrete valued distributions [7].

Random graph generation

A random graph with bivariate joint degree-degree distribution can be generated as follows ([17]):

1. Degree sequence is generated according to the degree distribution, $f_d(d) = \frac{f(d)E[D]}{d}$
2. An uncorrelated random graph is generated with the generated degree sequence using configuration model ([1, 18])
3. Metropolis dynamics is now applied on the generated graph: choose two edges randomly (denoted by the vertex pairs (v_1, w_1) and (v_2, w_2)) and measure the degrees, (j_1, k_1) and (j_2, k_2), that correspond to these vertex pairs and generated a random number, y, according to uniform distribution in $[0, 1]$. If $y \leq \min(1, \big(f(j_1, j_2) f(k_1, k_2)\big) / \big(f(j_1, k_1) f(j_2, k_2)\big))$, then remove the selected edges and construct news ones as (v_1, v_2) and (w_1, w_2). Otherwise, keep the selected edges intact. This dynamics will generate an instance of the random graph with the required joint degree-degree distribution. Run Metropolis dynamics well enough to mix the generating process.

As an example, we shall often use the following bivariate Pareto model for the joint degree-degree tail function of the graph,

$$\bar{F}(d_1, d_2) = \left(1 + \frac{d_1 - \mu}{\sigma} + \frac{d_2 - \mu}{\sigma} \right)^{-\gamma}, \tag{7}$$

where σ, μ, and γ are positive values. The use of the bivariate Pareto distribution can be justified by the statistical analysis in [19].

Description of random walk-based sampling processes

In this section, we explain three different random walk-based algorithms for exploring the network. They have been extensively studied in previous works [2, 3, 20] where they are formulated with vertex set as the state space of the underlying Markov chain on graph. The walker in these algorithms, after reaching each node, moves to another node randomly by following the transition kernel of the Markov chain. However, the quantity of interest is generally a measurable function of the Markov chain. As a case study, let us again take the degree sequence. We use $f_{\mathscr{X}}$ and $\mathrm{P}_{\mathscr{X}}$ to represent the probability density function and probability measure under the algorithm $\mathscr{X}$ with the exception that f_d represents the probability density function of degrees.

Random walk (RW)

In a random walk, the next node to visit is chosen uniformly among the neighbors of the current node. Let $V_1, V_2, \ldots$ be the nodes crawled by the RW and $D_1, D_2, \ldots$ be the degree sequence corresponding to the sequence $V_1, V_2, \ldots$.

Theorem 2. *The following relation holds in the stationary regime*

$$f_{\mathrm{RW}}(d_1, d_2) = f(d_1, d_2), \tag{8}$$

where $f(d_1, d_2)$ is the joint degree-degree distribution and $f_{\mathrm{RW}}(d_1, d_2)$ is the bi-variate joint distribution of the degree sequences generated by the standard random walk.

Proof. We note that the sequence $\{(V_i, V_{i+1})\}_{i \geq 1}$ also forms a Markov chain. With the assumption that the graph is connected, the ergodicity holds for any function g, i.e.,

$$\frac{1}{T}\sum_{i=1}^{T} g(V_i, V_{i+1}) \to \mathrm{E}_{\pi}\left[g(V_{\xi}, V_{\xi+1})\right], \quad T \to \infty,$$

where E_{π} is the expectation under stationary distribution π of $\{(V_i, V_{i+1})\}$ (which is uniform over edges) and $(V_{\xi}, V_{\xi+1})$ indicates a randomly picked edge. The ergodicity can then be extended to functions of the degree sequence $\{(D_i, D_{i+1})\}$ corresponding to $\{(V_i, V_{i+1})\}$, and in particular

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}\{D_i = d_1, D_{i+1} = d_2\} &\to \mathrm{E}_{\pi}\left[\mathbf{1}\{D_{\xi} = d_1, D_{\xi+1} = d_2\}\right], \quad T \to \infty \\ &= \frac{1}{M}\sum_{(p,q)\in E} \mathbf{1}\left\{D_p = d_1, D_q = d_2\right\} \\ &= f(d_1, d_2), \end{aligned} \tag{9}$$

where $\mathbf{1}\{\mathcal{A}\}$ denotes the indicator function for the event $\mathcal{A}$. L.H.S. of (9) is an estimator of $f_{\mathrm{RW}}(d_1, d_2)$. This means that when the RW is in stationary regime $\mathrm{E}[\,\mathbf{1}\{D_i = d_1, D_{i+1} = d_2\}] = \mathrm{E}_{\pi}[\,\mathbf{1}\{D_{\xi} = d_1, D_{\xi+1} = d_2\}]$ and hence Eq. 8 holds. □

PageRank (PR)

Using Eq. 6, we can approximate the degree sequence by a random walk on degree space with the following transition kernel:

$$f_{\mathrm{RW}}(d_{t+1}|d_t) = \frac{\mathrm{E}[D] f(d_t, d_{t+1})}{d_t f_d(d_t)}, \tag{10}$$

where the present node has degree d_t and the next node is with degree d_{t+1}. The above relation holds with equality for discrete degree distribution, but some care needs to be taken if one uses continuous version for the degree distributions.

If the standard random walk on the vertex set is in the stationary regime, its stationary distribution (probability of staying at a particular vertex i) is proportional to the degree (see e.g., [20]) and is given by $d_i/2M$, M being the number of edges. Then in the standard random walk on degree set, the stationary distribution of staying at any node with degree around d_1 can be approximated as $Nf_d(d_1)\ (d_1/2M)$, with N as the number of nodes. Thus

$$f_{\mathrm{RW}}(d_1) = \frac{d_1}{\mathrm{E}[D]} f_d(d_1).$$

Check of the approximation

We provide comparison of simulated values and theoretical values of transition kernel of RW in Fig. 1. To be specific, we use the bivariate Pareto distribution given (7). In the figure, N is 5,000. $\mu = 10$, $\gamma = 1.2$ and $\sigma = 15$. These choices of parameters provide $E[D] = 21.0052$. At each instant Metropolis dynamics will choose two edges and it has run 200,000 times (provides sufficient mixing). The figure shows satisfactory fitting of the approximation.

PageRank is a modification of the random walk which with a fixed probability $1-c$ samples a random node with uniform distribution and with a probability c, it follows the random walk transition [3]. Its evolution on degree state space can be described as follows:

$$\begin{aligned} f_{PR}(d_{t+1}|d_t) &= c\, f_{RW}(d_{t+1}|d_t) + (1-c)\frac{1}{N} N f_d(d_{t+1}) \\ &= c\, f_{RW}(d_{t+1}|d_t) + (1-c) f_d(d_{t+1}). \end{aligned} \tag{11}$$

Here the $1/N$ corresponds to the uniform sampling on vertex set and $\frac{1}{N}Nf_d(d_{t+1})$ indicates the net probability of jumping to all the nodes with degree around d_{t+1}.

Consistency with PageRank value distribution

We make a consistency check of the approximation derived for transition kernel by studying tail behavior of degree distribution and PageRank value distribution. It is known that

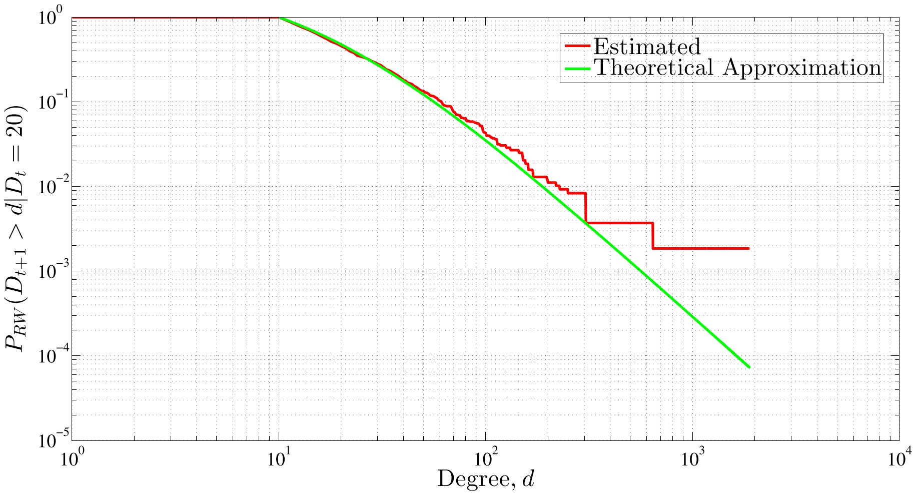


Fig. 1 Transition kernel comparison

under some strict conditions, for a directed graph, PageRank and Indegree have same tail exponents [21]. In our formulation in terms of degrees, for *uncorrelated* and undirected graph, PageRank for a given degree d, $\mathrm{PR}(d)$, can be approximated from the basic definition as,

$$\mathrm{PR}(d) = f_{\mathrm{PR}}(d) = c\, f_{\mathrm{RW}}(d) + (1-c)\, f_d(d).$$

This is a deterministic quantity. We are interested in the distribution of the random variable $\mathrm{PR}(D)$, PageRank of a randomly chosen degree class D. PageRank $\mathrm{PR}(d)$ is also the long term proportion or probability that PageRank process ends in a degree class with degree d. This can be scaled suitably to provide a rank-type information. Its tail distribution is

$$P(\mathrm{PR}(D) > x) = P\left(c.f_{\mathrm{RW}}(D) + (1-c).f_d(D) > x\right),$$

where $D \sim f_d(.)$. The PageRank of any vertex inside the degree class d is $\mathrm{PR}(d)/(Nf_d(d))$. The distribution of PageRank of a randomly chosen vertex i, $P(\mathrm{PR}(i) > x)$ after appropriate scaling for comparison with degree distribution is $P(N.\mathrm{PR}(i) > \hat{d})$, where $\hat{d} = Nx$. Now

$$\begin{aligned} P(N.\mathrm{PR}(i) > \hat{d}) &= P\left(N\frac{PR(D)}{Nf_d(D)} > \hat{d}\right) \\ &= P\left(D > \frac{E[D]}{c}\left[\hat{d} - (1-c)\right]\right). \end{aligned}$$

This of the form $P(D > A\hat{d} + B)$ with A and B as appropriate constants and hence will have the same exponent of degree distribution tail when the graph is *uncorrelated*.

There is no convenient expression for the stationary distribution of PageRank, to the best of our knowledge, and it is difficult to come up with an easy to handle expression for the joint distribution. Therefore, along with other advantages, we consider another modification of the standard random walk.

Random walk with jumps (RWJ)

RW sampling leads to many practical issues like the possibility to get stuck in a disconnected component, biased estimators etc. RWJ overcomes such problems [2].

In this algorithm, we follow random walk on a modified graph which is a superposition of the original graph and complete graph on same vertex set of the original graph with weight α/N on each artificially added edge, $\alpha \in [0,\infty]$ being a design parameter [2]. The algorithm can be shown to be equivalent to select $c = \alpha/(d_t + \alpha)$ in the PageRank algorithm, where d_t is the degree of the present node. The larger the node's degree, the less likely is the artificial jump of the process. This modification makes the underlying Markov chain time reversible, significantly reduces mixing time, improves estimation error, and leads to a closed form expression for stationary distribution.

Before proceeding to formulate the next theorem, we recall that the degree distribution $f_d(d_1)$ is different from the marginal of$f(d_1, d_2)$, $f(d_1)$.

Theorem 3. *The following relation holds in the stationary regime*

$$f_{\mathrm{RWJ}}(d_1, d_2) = \frac{\mathrm{E}[D]}{\mathrm{E}[D] + \alpha} f(d_1, d_2) + \frac{\alpha}{\mathrm{E}[D] + \alpha} f_d(d_1) f_d(d_2), \tag{12}$$

where $f(d_1, d_2)$ *is the joint degree-degree distribution,* $f_d(d_1)$ *is the degree distribution, and* $f_{\mathrm{RWJ}}(d_1, d_2)$ *is the bi-variate joint distribution of the degree sequences generated by the random walk with jumps.*

Proof. On the similar lines in the analysis of RW, $f_{\mathrm{RWJ}}(d_1, d_2)$ can be calculated as follows. The stationary distribution, $f_{\mathrm{RWJ}}(p)$, for node p (on the vertex set) is $(d_p+\alpha)/(2M+N\alpha)$. The transition probability from node p to node q, $f_{\mathrm{RWJ}}(q|p)$, is $(\alpha/N+1)/(d_p+\alpha)$ when there is a link from p to q, and when there is no link, it is $(\alpha/N)/(d_p+\alpha)$ [2]. Then, the joint distribution between nodes is given by

$$f_{\mathrm{RWJ}}(p,q) = f_{\mathrm{RWJ}}(q|p) f_{\mathrm{RWJ}}(p) = \begin{cases} \frac{\frac{\alpha}{N}+1}{2M+N\alpha} & \text{if } p \text{ has link to } q, \\ \frac{\frac{\alpha}{N}}{2M+N\alpha} & \text{if } p \text{ does not have link to } q. \end{cases}$$

Therefore

$$\begin{aligned} & f_{\mathrm{RWJ}}(d_1, d_2) \\ &= \mathrm{E}_\pi\left[\mathbf{1}\left\{D_\xi = d_1, D_{\xi+1} = d_2\right\}\right] \\ &\overset{(a)}{=} 2\,\frac{\frac{\alpha}{N}+1}{2M+N\alpha} \sum_{(p,q)\in E} \mathbf{1}\left\{D_p = d_1, D_q = d_2\right\} \\ &\qquad\qquad +2\,\frac{\frac{\alpha}{N}}{2M+N\alpha} \sum_{(p,q)\notin E} \mathbf{1}\left\{D_p = d_1, D_q = d_2\right\} \\ &\overset{(b)}{=} 2\,\frac{\frac{\alpha}{N}+1}{2M+N\alpha} M f(d_1, d_2) \\ &\qquad +2\,\frac{\frac{\alpha}{N}}{2M+N\alpha}\left(\frac{1}{2}\sum_{p\in V}\mathbf{1}\{D_p = d_1\}\sum_{q\in V}\mathbf{1}\{D_q = d_2\} - M f(d_1, d_2)\right) \\ &= \frac{\mathrm{E}[D]}{\mathrm{E}[D]+\alpha} f(d_1, d_2) + \frac{\alpha}{\mathrm{E}[D]+\alpha} f_d(d_1) f_d(d_2). \end{aligned}$$

Here $\mathrm{E}[D] = 2M/N$. The multiplying factor 2 is introduced in (a) because of the symmetry in the joint distribution $f_{\mathrm{RWJ}}(p,q)$ over the nodes, terms outside the summation in the R.H.S. The factor 1/2 in R.H.S. in (b) is to take into account the fact that only half of the combinations of (p,q) is needed. □

We also have the following. The stationary distribution on degree set by collecting all the nodes with same degree is

$$\begin{aligned} f_{\mathrm{RWJ}}(d_1) &= \left(\frac{d_1+\alpha}{2M+N\alpha}\right) N f_d(d_1) \\ &= \frac{(d_1+\alpha) f_d(d_1)}{\mathrm{E}[D]+\alpha}. \end{aligned} \tag{13}$$

Moreover, the associated tail distribution has a simple form,

$$f_{\mathrm{RWJ}}(D_{t+1} > d_{t+1}, D_t > d_t) = \frac{\mathrm{E}[D]\,\overline{F}(d_{t+1}, d_t) + \alpha \overline{F}_d(d_{t+1}) \overline{F}_d(d_t)}{\mathrm{E}[D]+\alpha}. \tag{14}$$

Remark 3. Characterizing Markov chain-based sampling in terms of degree evolution has some advantages.

- In the different random walk algorithms considered on the vertex set, all the nodes with same degree have same stationary distribution. This also implies that it is more natural to formulate the random walk evolution in terms of degree.
- For uncorrelated networks, $f_{\text{RW}}(d_1, d_2) = f_{\text{RW}}(d_1) f_{\text{RW}}(d_2)$, $f_{\text{PR}}(d_1, d_2) = f_{\text{PR}}(d_1) f_{\text{PR}}(d_2)$ and $f_{\text{RWJ}}(d_1, d_2) = f_{\text{RWJ}}(d_1) f_{\text{RWJ}}(d_2)$.

Extremal index for bivariate Pareto degree correlation

As explained in the "Introduction" section, EI is an important parameter in characterizing dependence and extremal properties in a stationary sequence. We assume that we have waited sufficiently long that the underlying Markov chain of the three different graph sampling algorithms are in stationary regime now. Here, we derive EI of RW and RWJ for the model with degree correlation among neighbors as bivariate Pareto (7).

The two mixing conditions $D(u_n)$ and $D''(u_n)$ introduced in section "Calculation of extremal index (EI)" are needed for our EI analysis. Condition $D(u_n)$ is satisfied as explained in section "Check of conditions $D(u_n)$ and $D''(u_n)$ for functions of Markov samples." An empirical evaluation of $D''(u_n)$ is provided in section "Check of condition D''".

EI for random walk sampling

We use the expression for EI given in Theorem 1. As $f_{\text{RW}}(x, y)$ is same as $f(x, y)$, we have,

$$\begin{aligned} \widehat{C}(u,u) &= \text{P}(D_1 > \bar{F}^{-1}(u), D_2 > \bar{F}^{-1}(u)) \\ &= \left(1 + 2(u^{-1/\gamma} - 1)\right)^{-\gamma} \\ \underline{1} \cdot \nabla\, \widehat{C}(u,u) &= 2(2 - u^{1/\gamma})^{-(\gamma+1)}. \end{aligned}$$

Thus $\theta = 1 - \underline{1} \cdot \nabla\, \widehat{C}(0,0) = 1 - 1/2^{\gamma}$. For $\gamma = 1$, we get $\theta = 1/2$. In this case, we can also use expression obtained in Eq. 5.

EI for random walk with jumps sampling

Although it is possible to derive EI as in RW case above, we provide an alternative way to avoid the calculation of tail distribution of degrees and inverse of RWJ marginal (with respect to the bivariate Pareto degree correlation). We assume the existence of EI in the following proposition.

Proposition 2. *When the bivariate joint degree distribution of neighboring nodes are Pareto distributed as given by Eq. 7 and random walk with jumps is employed for sampling, the EI is given by*

$$\theta = 1 - \frac{E[D]}{E[D] + \alpha} 2^{-\gamma}, \tag{15}$$

where $E[D]$ is the expected degree, α is the parameter of the random walk with jumps, and γ is the tail index of the bivariate Pareto distribution.

Proof. Under the assumption of D'',

$$\theta = \lim_{n\to\infty} \frac{P(D_2 \le u_n, D_1 > u_n)}{P(D_1 > u_n)} = \lim_{n\to\infty} \frac{P(D_1 \ge u_n) - P(D_2 \ge u_n, D_1 \ge u_n)}{P(D_1 > u_n)} \tag{16}$$

Now using the Condition 1 on the marginal and joint tail distribution of RWJ in Eq. 14, we can write[2]

$$\frac{P(D_1 \geq u_n) - P(D_2 \geq u_n, D_1 \geq u_n)}{P(D_1 > u_n)}$$
$$= \frac{\tau/n + o(1/n) - \frac{E[D]}{E[D]+\alpha} P_{\mathrm{RW}}(D_2 \geq u_n, D_1 \geq u_n) - \frac{\alpha}{E[D]+\alpha} O(\tau/n) O(\tau/n)}{\tau/n + o(1/n)}$$

The asymptotics in the last term of the numerator is due to the following:

$$\overline{F}_{\mathrm{RWJ}}(u_n) = \frac{E[D]}{E[D]+\alpha} \overline{F}(u_n) + \frac{\alpha}{E[D]+\alpha} \overline{F}_d(u_n) = \tau/n + o(1/n),$$

and hence $\overline{F}_d(u_n) = O(\tau/n)$. Therefore, Eq. 16 becomes

$$\theta = 1 - \frac{E[D]}{E[D]+\alpha} \lim_{n\to\infty} P_{\mathrm{RW}}(D_2 \geq u_n, D_1 \geq u_n) n/\tau$$

Then in the case of the bivariate Pareto distribution in Eq. 7, we obtain Eq. 15. □

Lower bound of EI of the PageRank

We obtain the following lower bound for EI in the PageRank processes.

Proposition 3. *For the stationary PageRank process on degree state space Eq. 10 with EI θ, irrespective of the degree correlation structure in the underlying graph, the EI is bounded by*

$$\theta \geq (1-c),$$

where c is the damping factor in the PageRank algorithm.

Proof. From [13], with another mixing condition AIM(u_n) which is satisfied for functions of stationary Markov samples (e.g., degree samples) the following representation of EI holds,

$$\lim_{n\to\infty} \mathrm{P}\{M_{1,p_n} \leq u_n | D_1 > u_n\} \leq \theta, \tag{17}$$

where $\{p_n\}$ is an increasing sequence of positive integers, $p_n = o(n)$ as $n \to \infty$ and $M_{1,p_n} = \max\{D_2, ..., D_{p_n}\}$. Let $\mathcal{A}$ be the event that the node corresponding to D_2 is selected uniformly among all the nodes, not following random walk from the node for D_1. Then, $\mathrm{P}_{\mathrm{PR}}(\mathcal{A}) = 1 - c$. Now, with Eq. 11,

$$\begin{aligned}
\mathrm{P}_{\mathrm{PR}}(M_{1,p_n} \leq u_n | D_1 > u_n) &\geq \mathrm{P}_{\mathrm{PR}}(M_{1,p_n} \leq u_n, \mathcal{A} | D_1 > u_n) \\
&= \mathrm{P}_{\mathrm{PR}}(\mathcal{A} | D_1 > u_n) \mathrm{P}_{\mathrm{PR}}(M_{1,p_n} \leq u_n | \mathcal{A}, D_1 > u_n) \\
&\overset{(i)}{=} (1-c) \mathrm{P}_{\mathrm{PR}}(M_{1,p_n} \leq u_n), \\
&\overset{(ii)}{=} (1-c) \mathrm{P}_{\mathrm{PR}}^{(p_n-1)\theta}(D_1 \leq u_n) + o(1) \\
&\geq (1-c) \mathrm{P}_{\mathrm{PR}}^{(p_n-1)}(D_1 \leq u_n) + o(1) \\
&\overset{(iii)}{\sim} (1-c)(1-\tau/n)^{p_n-1},
\end{aligned} \tag{18}$$

where $\{p_n\}$ is the same sequence as in Eq. 17 and (i) follows mainly from the observation that conditioned on $\mathcal{A}$, $\{M_{1,p_n} \leq u_n\}$ is independent of $\{D_1 > u_n\}$, and (ii) and (iii) result from the limits in Eqs. 3 and 1, respectively.

Assuming $p_n - 1 = n^{1/2}$ and since $(1 - \tau/n)^{p_n - 1} \sim e^{-\tau/\sqrt{n}} \to 1$ as $n \to \infty$, from Eqs. 17 and 18,

$$\theta \geq 1 - c.$$

The PageRank transition kernel (Eq. 11) on the degree state space does not depend upon the random graph model in section "Description of the configuration model with degree-degree correlation". Hence, the derived lower bound of EI is useful for any degree correlation model. □

Applications of extremal index in network sampling processes

This section provides several applications of EI in inferring the sampled sequence. This emphasizes that the analytical calculation and estimation of EI are practically relevant.

The limit of the point process of exceedances, $N_n(.)$, which counts the times, normalized by n, at which $\{X_i\}_{i=1}^n$ exceeds a threshold u_n provides many applications of EI. A cluster is considered to be formed by the exceedances in a block of size r_n ($r_n = o(n)$) in n with cluster size $\xi_n = \sum_{i=1}^{r_n} 1(X_i > u_n)$ when there is at least one exceedance within r_n. The point process N_n converges weakly to a compound poisson process (CP) with rate $\theta\tau$ and i.i.d. distribution as the limiting distribution of cluster size, under Condition 1 and a mixing condition, and the points of exceedances in CP correspond to the clusters (see [4], Section 10.3 for details). We also call this kind of clusters as blocks of exceedances.

The applications below require a choice of the threshold sequence $\{u_n\}$ satisfying Eq. 1. For practical purposes, if a single threshold u is demanded for the sampling budget B, we can fix $u = \max\{u_1, \ldots, u_B\}$.

The applications in this section are explained with the assumption that the sampled sequence is the sequence of node degrees. But the following techniques are very general and can be extended to any sampled sequence satisfying conditions $D(u_n)$ and $D''(u_n)$.

Order statistics of the sampled degrees

The order statistics $X_{n-k,n}$, $(n-k)$th maxima is related to $N_n(.)$ and thus to θ by

$$\mathrm{P}(X_{n-k,n} \leq u_n) = \mathrm{P}(N_n((0,1]\,) \leq k),$$

where we apply the result of convergence of N_n to CP ([4], Section 10.3.1).

Distribution of maxima

The distribution of the maxima of the sampled degree sequences can be derived as Eq. 3 when $n \to \infty$.

Hence if the EI of the underlying process is known then from Eq. 3, one can approximate the $(1-\eta)$th quantile x_η of the maximal degree M_n as

$$\mathrm{P}\{M_n \leq x_\eta\} = F^{n\theta}(x_\eta) = \mathrm{P}^{n\theta}\{X_1 \leq x_\eta\} = 1 - \eta,$$

i.e.,

$$x_\eta \approx F^{-1}\left((1-\eta)^{1/(n\theta)}\right). \tag{19}$$

In other words, quantiles can be used to find the maxima of the degree sequence with certain probability.

If the sampling procedures have same marginal distribution, with calculation of EI, it is possible to predict how much large values can be achieved. Lower EI indicates lower value for x_η and higher represents high x_η.

For the random walk example in section "EI for random walk sampling" for the degree correlation model, with the use of Eq. 19, we get the $(1-\eta)$th quantile of the maxima M_n

$$x_\eta \approx \mu + \sigma \left(\left(1 - (1-\eta)^{1/(n\theta)}\right)^{-1/\gamma} - 1 \right).$$

The following example demonstrates the effect of neglecting correlations on the prediction of the largest degree node. The largest degree, with the assumption of Pareto distribution for the degree distribution, can be approximated as $\mathrm{KN}^{1/\delta}$ with $K \approx 1$, N as the number of nodes and γ as the tail index of complementary distribution function of degrees [22]. For Twitter graph (recorded in 2012), $\delta = 1.124$ for out-degree distribution and $N = 537,523,432$ [23]. This gives the largest degree prediction as 59,453,030. But the actual largest out-degree is 22,717,037. This difference is because the analysis in [22] assumes i.i.d. samples and does not take into account the degree correlation. With the knowledge of EI, correlation can be taken into account as in Eq. 3. In the following section, we derive an expression for such a case.

Estimation of largest degree when the marginals are Pareto distributed

It is known that many social networks have the degree asymptotically distributed as Pareto [18]. We find that in these cases, the marginal distribution of degrees of the random walk based methods also follow Pareto distribution (though we have derived only for the model with degree correlations among neighbors, see section "Degree correlations".)

Proposition 4. *For any stationary sequence with marginal distribution following Pareto distribution $\bar{F}(x) = Cx^{-\delta}$, the largest value, approximated as the median of the extreme value distribution, is given by*

$$M_n \approx (n\theta)^{1/\delta} \left(\frac{C}{\log 2}\right)^{1/\delta}.$$

Proof. From extreme value theory [4], it is known that when $\{X_i, i \geq 1\}$ are i.i.d.,

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = H_\gamma(x), \tag{20}$$

where $H_\gamma(x)$ is the extreme value distribution with index γ and $\{a_n\}$ and $\{b_n\}$ are appropriately chosen deterministic sequences. When $\{X_i, i \geq 1\}$ are stationary with EI θ, the limiting distribution becomes $H'_{\gamma'}(x)$ and it differs from $H_\gamma(x)$ only through parameters. $H_\gamma(x) = \exp(-t(x))$ with $t(x) = \left(1 + \left(\frac{x-\mu}{\sigma}\right)\gamma\right)^{-1/\gamma}$. With the normalizing constants ($\mu = 0$ and $\sigma = 1$), $H'_{\gamma'}$ has the same shape as H_γ with parameters $\gamma' = \gamma$, $\sigma' = \theta^\gamma$ and $\mu' = (\theta^\gamma - 1)/\gamma$ ([4], Section 10.2.3).

For Pareto case, $\overline{F}(x) = Cx^{-\delta}$, $\gamma = 1/\delta$, $a_n = \gamma C^\gamma n^\gamma$, and $b_n = C^\gamma n^\gamma$. From Eq. 20, for large n, M_n is stochastically equivalent to $a_n\chi + b_n$, where χ is a random variable with distribution $H'_{\gamma'}$. It is observed in [22] that median of χ is an appropriate choice for the estimation of M_n. Median of $\chi = \mu' + \sigma' \left(\frac{(\log 2)^{-\gamma'} - 1}{\gamma'}\right) = (\theta^\gamma (\log 2)^{-\gamma} - 1)\gamma^{-1}$. Hence,

$$\begin{aligned} M_n &\approx a_n \left(\frac{\theta^\gamma (\log 2)^{-\gamma}}{\gamma} - 1 \right) + b_n \\ &= (n\theta)^{1/\delta} \left(\frac{C}{\log 2} \right)^{1/\delta} \end{aligned}$$

□

Relation to first hitting time and interpretations

Extremal index also gives information about the first time $\{X_n\}$ hits (u_n, ∞). Let T_n be this time epoch. As N_n converges to compound poisson process, it can be observed that T_n/n is asymptotically an exponential random variable with rate $\theta\tau$, i.e., $\lim_{n\to\infty} \mathrm{P}(T_n/n > x) = \exp(-\theta\tau x)$. Therefore, $\lim_{n\to\infty} \mathrm{E}(T_n/n) = 1/(\theta\tau)$. Thus, the smaller EI is, the longer it will take to hit the extreme levels as compared to independent sampling. This property is particularly useful to compare different sampling procedures. It can also be used in quick detection of high degree nodes [22, 24].

Relation to mean cluster size

If Condition $D''(u_n)$ is satisfied along with $D(u_n)$, asymptotically, a run of the consecutive exceedances following an upcrossing is observed, i.e., $\{X_n\}$ crosses the threshold u_n at a time epoch and stays above u_n for some more time before crossing u_n downwards and stays below it for some time until next upcrossing of u_n happens. This is called cluster of exceedances and is more practically relevant than blocks of exceedances at the starting of this section and is shown in [10] that these two definitions clusters are asymptotically equivalent resulting in similar cluster size distribution.

The expected value of cluster of exceedances converges to inverse of EI ([4], p. 384), i.e.,

$$\theta^{-1} = \lim_{n\to\infty} \sum_{j\geq 1} j\pi_n(j),$$

where $\{\pi_n(j), j \geq 1\}$ is the distribution of size of cluster of exceedances with n samples. Asymptotical cluster size distribution and its mean are derived in [6].

Estimation of extremal index and numerical results

This section introduces two estimators for EI. Two types of networks are presented: synthetic correlated graph and real networks (Enron email network and DBLP network (http://dblp.uni-trier.de/)). For the synthetic graph, we compare the estimated EI to its theoretical value. For the real network, we calculate EI using the two estimators.

We take $\{X_i\}$ as the degree sequence and use RW, PR, and RWJ as the sampling techniques. The methods mentioned in the following are general and are not specific to degree sequence or random walk technique.

Empirical copula-based estimator

We have tried different estimators for EI available in literature [4, 14] and found that the idea of estimating copula and then finding value of its derivative at $(1, 1)$ works without the need to choose and optimize several parameters found in other estimators. We assume that $\{X_i\}$ satisfies $D(u_n)$ and $D''(u_n)$, and we use Eq. 4 for calculation of EI. Copula $C(u, v)$ is estimated empirically by

$$C_n(u,v) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\left(\frac{R^X_{i_k}}{n+1} \le u, \frac{R^Y_{i_k}}{n+1} \le v\right),$$

with $R^X_{i_k}$ indicates rank of the element X_{i_k} in $\{X_{i_k}, 1 \le k \le n\}$ and $R^Y_{i_k}$ is defined respectively. The sequence $\{X_{i_k}\}$ is chosen from the original sequence $\{X_i\}$ in such a way that X_{i_k} and $X_{i_{k+1}}$ are sufficiently apart to make them independent to a certain extent and $Y_{i_k} = X_{i_k+1}$. The large sample distribution of $C_n(u,v)$ is normal and centered at copula $C(u,v)$. Now, to get θ, we use linear least squares error fitting to find the slope at $(1,1)$ or use cubic spline interpolation for better results.

Intervals estimator

This estimator does not assume any conditions on $\{X_i\}$ but has the parameter u to choose appropriately. Let $N = \sum_{i=1}^{n} 1(X_i > u)$ be the number of exceedances of u at time epochs $1 \le S_1 < \ldots < S_N \le n$ and let the interexceedance times be $T_i = S_{i+1} - S_i$. Then intervals estimator is defined as ([4], p. 391),

$$\hat{\theta}_n(u) = \begin{cases} \min(1, \hat{\theta}^1_n(u)), \text{if } \max T_i : 1 \le i \le N-1 \le 2, \\ \min(1, \hat{\theta}^2_n(u)), \text{if } \max T_i : 1 \le i \le N-1 > 2, \end{cases}$$

where

$$\hat{\theta}^1_n(u) = \frac{2\left(\sum_{i=1}^{N-1} T_i\right)^2}{(N-1)\sum_{i=1}^{N-1} T_i^2},$$

and

$$\hat{\theta}^2_n(u) = \frac{2\left(\sum_{i=1}^{N-1} (T_i - 1)\right)^2}{(N-1)\sum_{i=1}^{N-1} (T_i - 1)(T_i - 2)}.$$

We choose u as δ percentage quantile thresholds, i.e., δ percentage of $\{X_i, 1 \le i \le n\}$ falls below u,

$$k_\delta = \min\left\{k : \sum_{i=1}^{n} \frac{\mathbf{1}\{X_i \le X_k\}}{n} \ge \frac{\delta}{100}, 1 \le k \le n\right\}, \qquad u = X_{k_\delta}.$$

We plot θ_n vs δ for the intervals estimator in the following sections. The EI is usually selected as the value corresponding to the stability interval in this plot.

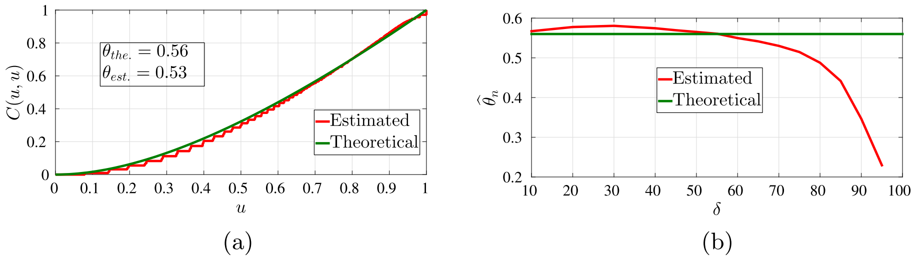


Fig. 2 RW sampling (synthetic graph). **a** Empirical and theoretical copulas. **b** Interval estimate and theoretical value

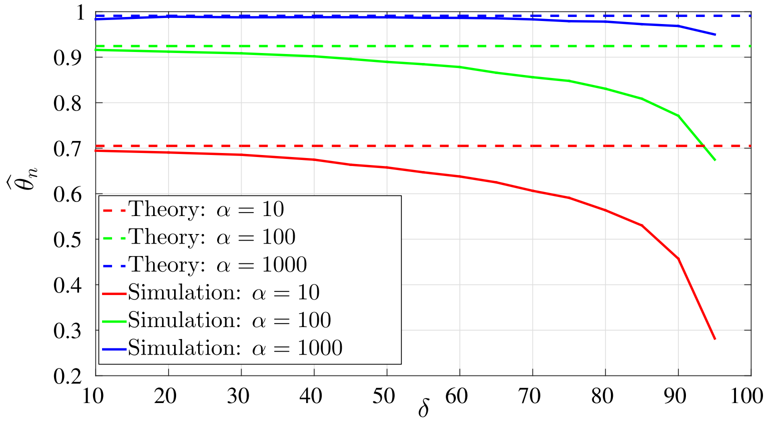


Fig. 3 RWJ sampling (synthetic graph): intervals estimate and theoretical value

Synthetic graph

The simulations in the section follow the bivariate Pareto model and parameters introduced in Eq. 7. We use the same set of parameters as for Fig. 1, and the graph is generated according to the Metropolis technique in section "Random graph generation".

For the RW case, Fig. 2a shows copula estimator, and theoretical copula-based on the continuous distribution in Eq. 7, and is given by

$$C(u, u) = \left(1 + 2((1-u)^{-1/\gamma} - 1)\right)^{-\gamma} + 2u - 1.$$

Though we take quantized values for degree sequence, it is found that the copula estimated matches with theoretical copula. The value of EI is then obtained after cubic interpolation and numerical differentiation of copula estimator at point (1, 1). For the theoretical copula, EI is $1 - 1/2^{\gamma}$, where $\gamma = 1.2$. Figure 2b displays the comparison between the theoretical value of EI and intervals estimate.

For the RWJ algorithm, Fig. 3 shows the interval estimate and theoretical value for different α. We used Eq. 15 for theoretical calculation. The small difference in theory and simulation results is due to the assumption of continuous degrees in the analysis, but the practical usage requires quantized version. Here $\alpha = 0$ case corresponds to RW sampling.

Figure 4 displays the interval estimate of EI with PR sampling. It can be seen that the lower bound proposed in Proposition 3 gets tighter as c decreases. When $c = 1$, PR sampling becomes RW sampling.

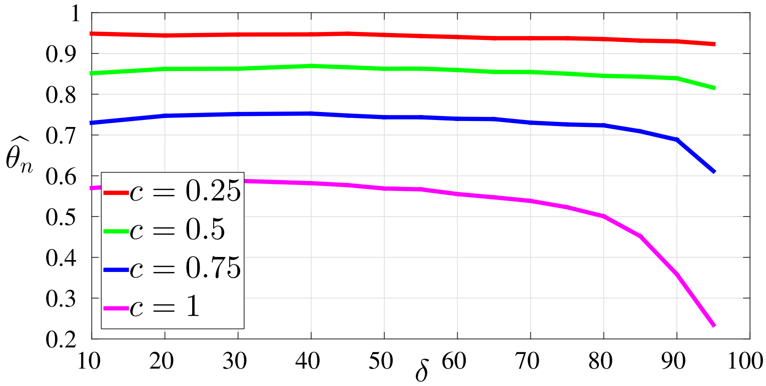


Fig. 4 PR sampling (synthetic graph): intervals estimate

Table 1 Test of Condition D'' in the synthetic graph

	r_{up} (%)	r_{cluster} (%)
RW	4	89
PR	7	91
RWJ	5	86

Check of condition D′′

The mixing conditions $D(u_n)$ and $D''(u_n)$ need to be satisfied for using the theory in section "Calculation of extremal index (EI)". Though intervals estimator does not require them, these conditions will provide the representation by Eq. 4. Condition $D(u_n)$ works in this case as explained in previous sections and for $D''(u_n)$, we do the following empirical test. We collect samples for each of the techniques RW, PR, and RWJ with parameters given in respective figures. Intervals are taken of duration 5, 10, 15, and 20 time samples. The ratio of number of upcrossings to number of exceedances r_{up} and ratio of number consecutive exceedances to number of exceedances r_{cluster} are calculated in Table 1. These proportions are averaged over 2000 occurrences of each of these intervals and over all the different intervals. The statistics in the table indicates strong occurrence of Condition $D''(u_n)$. We have also observed that the changes in the parameters does not affect this inference.

Real network

We consider two real-world networks: Enron email network and DBLP network. The data is collected from [25]. Both the networks satisfy the check for Condition $D''(u_n)$ reasonably well.

For the RW sampling, Fig. 5a shows the empirical copula, and it also mentions corresponding EI. Intervals estimator is presented in Fig. 5b. After observing plateaux in the plots, we took EI as 0.25 and 0.2 for DBLP and Enron email graphs, respectively.

In case of RWJ sampling, Fig. 6a, b presents the intervals estimator for email-Enron and DBLP graphs, respectively.

Conclusions

In this work, we have associated extreme value theory of stationary sequences to sampling of large networks. We show that for any general stationary samples (function of node samples) meeting two mixing conditions, the knowledge of bivariate distribution or bivariate copula is sufficient to derive many of its extremal properties. The parameter

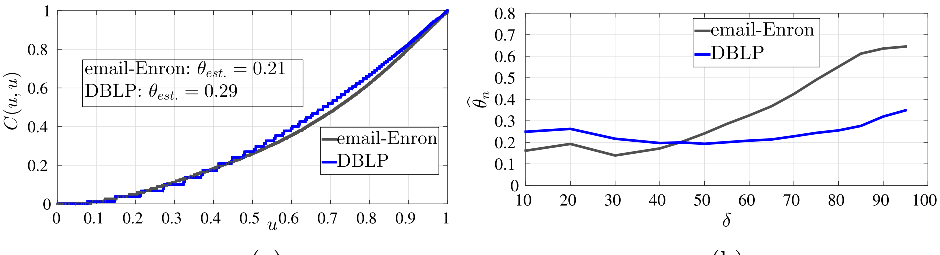
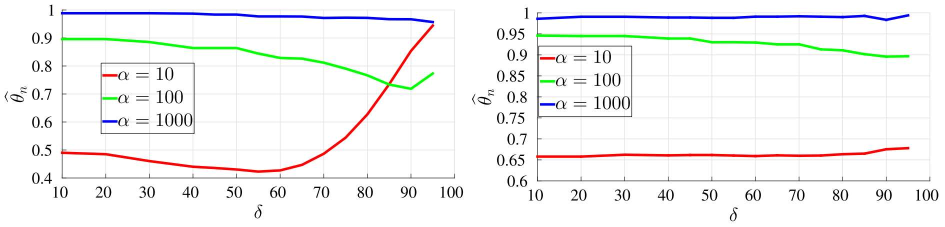


Fig. 5 RW sampling (real networks). **a** Empirical copulas. **b** Interval estimate



(a) (b)

Fig. 6 RWJ sampling (real networks). **a** Email-Enron. **b** DBLP

extremal index (EI) encapsulates this relation. We relate EI to many relevant extremes in networks like order statistics, first hitting time, and mean cluster size. In particular, we model dependence in degrees of adjacent nodes and examine random walk-based degree sampling. Finally, we have obtained estimates of EI for a synthetic graph with degree correlations and find a good match with the theory. We also calculate EI for two real-world networks. In future, we plan to investigate the relation between assortativity coefficient and EI and intends to study in detail the EI in real networks.

Endnotes

[1] $F^k(.)$ kth power of $F(.)$ throughout the paper except when $k = -1$ where it denotes the inverse function.

[2] $\sim$' stands for asymptotically equal, i.e., $f(x) \sim g(x) \Leftrightarrow f(x)/g(x) \to 1$ as $x \to a, x \in M$ where the functions $f(x)$ and $g(x)$ are defined on some set M, and a is a limit point of M. $f(x) = o(g(x))$ means $\lim_{x\to a} f(x)/g(x) = 0$. Also $f(x) = O(g(x))$ indicates that there exist $\delta > 0$ and $M > 0$ such that $|f(x)| \le M|g(x)|$ for $|x - a| < \delta$.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The contribution of all three authors to the manuscript is quite balanced. All authors read and approved the final manuscript.

Acknowledgements

The work of the first and third authors is partly supported by ADR "Network Science" of the Alcatel-Lucent Inria joint lab. The second author was partly supported by the Russian Foundation for Basic Research, grant 13-08-00744 A, and Campus France—Russian Embassy bilateral exchange programme.
The authors are thankful to Remco van der Hofstad for helpful suggestions during the preparation of this manuscript.

Author details

[1]INRIA Sophia Antipolis 2004, route des Lucioles - BP 93, 06902 Sophia Antipolis Cedex, France. [2]Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia.

Received: 19 February 2015 Accepted: 21 June 2015
Published online: 22 July 2015

References

1. Barrat, A, Barthelemy, M, Vespignani, A: Dynamical Processes on Complex Networks. Cambridge University Press, New York (2008)
2. Avrachenkov, K, Ribeiro, B, Towsley, D: Improving random walk estimation accuracy with uniform restarts. In: LNCS, pp. 98–109. Springer, Berlin Heidelberg, (2010)
3. Brin, S, Page, L: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1), 107–117 (1998)
4. Beirlant, J, Goegebeur, Y, Teugels, J, Segers, J: Statistics of Extremes: Theory and Applications. Wiley, Chichester, West Sussex (2004)
5. Ferro, CAT, Segers, J: Inference for clusters of extreme values. J. R. Stat. Soc. Ser. B. **65**, 545–556 (2003)
6. Markovich, NM: Modeling clusters of extreme values. Extremes. **17**(1), 97–125 (2014)
7. Leadbetter, MR, Lindgren, G, Rootzén, H: Extremes and Related Properties of Random Sequences and Processes, Vol. 21. Springer, New York (1983)

8. Avrachenkov, K, M. Markovich, N, Sreedharan, JK: Distribution and dependence of extremes in network sampling processes. In: Third International IEEE Workshop on Complex Networks and Their Applications. IEEE, Marrakesh, Morocco, (2014)
9. Nelsen, RB: An Introduction to Copulas. 2nd edn. Springer, New York (2007)
10. Leadbetter, MR, Nandagopalan, S: On exceedance point processes for stationary sequences under mild oscillation restrictions. In: Extreme Value Theory. Lecture Notes in Statistics, pp. 69–80. Springer, New York, (1989)
11. Chernick, MR, Hsing, T, McCormick, WP: Calculating the extremal index for a class of stationary sequences. Adv. Appl. Probab. **23**(4), 835–850 (1991)
12. Weng, C, Zhang, Y: Characterization of multivariate heavy-tailed distribution families via copula. J. Multivar. Anal. **106**(0), 178–186 (2012)
13. O'Brien, GL: Extreme values for stationary and Markov sequences. Ann. Probab. **15**(1), 281–291 (1987)
14. Ferreira, A, Ferreira, H: Extremal functions, extremal index and Markov chains. Technical report, Notas e comunicações CEAUL (December 2007)
15. Boguna, M, Pastor-Satorras, R, Vespignani, A: Epidemic spreading in complex networks with degree correlations. Stat. Mech. Complex Netw. Lect. Notes Physica. **625**, 127–147 (2003)
16. Goltsev, AV, Dorogovtsev, SN, Mendes, JFF: Percolation on correlated networks. Phys. Rev. E. **78**, 051105 (2008)
17. Newman, ME: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)
18. Van Der Hofstad, R: Random graphs and complex networks Vol. i. (2014). Available on http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf, accessed on 23 December 2014
19. Zhukovskiy, M, Vinogradov, D, Pritykin, Y, Ostroumova, L, Grechnikov, E, Gusev, G, Serdyukov, P, Raigorodskii, A: Empirical validation of the Buckley-Osthus model for the web host graph: degree and edge distributions. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1577–1581. ACM, Sheraton, Maui Hawaii, (2012)
20. Lovász, L: Random walks on graphs: a survey. Combinatorics, Paul erdos is eighty. **2**(1), 1–46 (1993)
21. Litvak, N, Scheinhardt, W. R, Volkovich, Y: In-degree and PageRank: why do they follow similar power laws?. Internet Math. **4**(2-3), 175–198 (2007)
22. Avrachenkov, K, Litvak, N, Sokol, M, Towsley, D: Quick detection of nodes with large degrees. In: Algorithms and Models for the Web Graph. Lecture Notes in Computer Science, pp. 54–65. Springer, Berlin Heidelberg, (2012)
23. Gabielkov, M, Rao, A, Legout, A: Studying social networks at scale: macroscopic anatomy of the twitter social graph. SIGMETRICS Perform. Eval. Rev. **42**(1), 277–288 (2014)
24. Avrachenkov, K, Litvak, N, Prokhorenkova, L. O, Suyargulova, E: Quick detection of high-degree entities in large directed networks. In: Proceedings of IEEE ICDM, (2014)
25. Stanford Large Network Dataset Collection. (2014). https://snap.stanford.edu/data/index.html, accessed on 11 December 2014