Statistical Analysis of Stochastic Processes in Social Networks

Natalia Markovich*
V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia - nat.markovich@gmail.com

Abstract

Social networks are an important part of human activity. The statistical analysis and risk assessment in complex social networks has to take into account extremes arising as nodes with large degrees or large PageRanks. These large nodes surrounded by smaller nodes build clusters of interconnected objects. The cluster structure of the network is caused by dependence between nodes that have appeared due to social relationships and common interests. A study of the stochastic properties of these clusters may improve the dissemination of information and advertisements. Clusters may be determined as conglomerates of nodes whose degrees exceed a high threshold. To study clusters determined by the exceedances of the degree metric, we apply results of extreme value theory. Namely, the asymptotically equivalent distributions of cluster and inter-cluster sizes obtained in Markovich (2014) and the extremal index that determines the dependence measure of extremes are used. Sampling techniques like a PageRank random walk and Metropolis-Hastings Markov chains are often used to collect information about characteristics of nodes such as their influence and popularity. To compare sampling techniques, we propose to estimate the first hitting time, i.e. the minimal time to reach a large node. To this end, we use the author's recent results with regard to the hitting time of threshold exceedances and its asymptotically equivalent geometric-like distribution and the limit mean. Secondly, we look deeper into the PageRank random walk that can be considered as a specific branching process. Jelenković and Olvera-Cravioto (2010), and Volkovich, Litvak (2010) proved that the stationary distribution of the PageRank process is heavy-tailed, regularly varying. We study extremal properties of the PageRank process. Particularly, we represent the extremal index of the PageRank process obtained in Avrachenkov, Markovich and Sreedharan (2014) by means of copulas. The exposition is provided by illustrations arising from the statistical analysis of real data that have been obtained in social networks.

Keywords: Cluster of exceedances; extremal index; first hitting time; PageRank.

1. Introduction

In this paper we discuss our achievements regarding extremes in social networks elaborated in our recent work. These results can be extended to complex technical systems like Internet.
Chain-referable sampling is a powerful method to collect a public opinion within social networks. It can be also useful to distribute advertisements effectively. A sample of nodes may be collected effectively by a random walk. The sampling techniques like uniform and random walk sampling, PageRank (Avrachenkov et al. (2010)), non-backtracking random walk with re-weighting (NBRW) (Chul-Ho Lee et al. (2012)), the random walk Metropolis and Metropolis-Hastings algorithms (Metropolis (1953), Hastings, W.K. (1970)) are usually compared by means of the mixing time. It captures the notion of the speed of convergence of the Markov chain with some transition probability matrix to the stationary distribution (Chul-Ho Lee et al. (2012)) and a spectral gap $\delta = 1 - |\lambda_2|$, where λ_2 is the second eigenvalue of the transmission probability matrix associated with a random walk (Avrachenkov et al. (2010)).
Here, we propose to compare sampling techniques by their extremal properties. The latter include an extremal index and a first hitting time. The latter shows the minimal time to find a node which is influential in some sense. A small extremal index corresponds to a large first hitting time and the presence of clusters that are large in average and vice versa. Getting into the large cluster (a galaxy) implies the risk to stay there.
Social networks may contain "mice" and "elephants" nodes, that have small and large numbers of links called degrees, respectively. Hence, social networks contain clusters (clouds) of nodes centered in the "elephant" nodes whose degrees exceed sufficiently high thresholds and are surrounded by "mice" nodes.
The cluster structure of networks is caused by the dependence (social relationships) between nodes and heavy-tailed distributions of the node degrees. It is important to investigate the stochastic nature of such clusters

since it allows us to distribute advertisement or to collect opinions more effectively within the clusters.
To model clusters one has to obtain limit distributions of their characteristics such as a maximal node degree, a cluster size and an inter-cluster size of a node degree sequence. The node degree may be replaced by the PageRank or another characteristic of the node.
Let $\{X_n\}_{n\geq 1}$ be a stationary process with marginal distribution function $F(x)$ and $M_n = \max\{X_1, ..., X_n\}$. We assume $P\{X_1 = x_F\} = 0$, where the right end-point $x_F = \sup\{x : F(x) < 1\}$ of $F(x)$ can be finite or infinite. The sequence $\{X_n\}$ may be obtained by means of some sampling tool. The node degrees and PageRanks give examples of $\{X_n\}$. We study the events $\{X_1 > u_n\}$, where $\{u_n\}$ is a sequence of sufficiently high thresholds.
To study clusters determined by the exceedances of the degree metric, we apply results of extreme value theory. Namely, the asymptotically equivalent distributions of cluster and inter-cluster sizes obtained in Markovich (2014), (2015a) and the extremal index that determines the dependence measure of extremes are used. The distribution and the mean of the first hitting time derived in Markovich (2015b) aim to compare the sampling approaches.
In Avrachenkov et al. (2014) we model the process of the node sampling by Page Rank and its development proposed in Avrachenkov et al. (2010). We obtain copulas to model bivariate distributions arising from Markov chains corresponding to PageRank-like processes. We find the extremal index of corresponding sampling processes of node degrees by means of copulas according to Ferreira (2006), Ferreira & Ferreira (2007).
The paper is organized as follows. In Section 2 we describe some necessary results of the extreme value theory. In Section 3 some recent results regarding the first hitting time are given. The PageRank process and its extremal properties are discussed in Section 4. In Section 5 we provide some illustrations from real data gathered in social networks.

2. Extreme value theory

An extremal index θ, $0 < \theta \leq 1$ is the main characteristic of the clusters of exceedances over a threshold u.

Definition 1 *The stationary sequence $\{X_n\}_{n\geq 1}$ is said to have extremal index $\theta \in [0, 1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers $u_n = u_n(\tau)$ such that*

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

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

hold (Leadbetter et al. (1983), p.53).

The reciprocal $1/\theta$ approximates the mean cluster size for a sufficiently large sample size. At the same time, θ allows us to evaluate a limit distribution of the maximum of the node degree and the first hitting time. This is helpful to find quantiles of the maxima, namely, to show what is the probability to find a big "elephant" in the sample. More exactly, if the extremal index of the underlying node-degree process is known then by

$$P\{M_n \leq x\} = F^{n\theta}(x) + o(1), \qquad n \to \infty, \tag{3}$$

where $F(x) = P\{D_1 \leq x\}$ is the continuous distribution function of node degrees, one can approximate the $(1-\eta)$th quantile x_η of the maximal degree M_n based on the equation

$$F^{n\theta}(x_\eta) = P^{n\theta}\{D_1 \leq x_\eta\} = 1 - \eta,$$

i.e. it holds

$$x_\eta \approx F^{\leftarrow}\left((1-\eta)^{1/(n\theta)}\right). \tag{4}$$

The following dependence conditions relate to the extremal index.

Definition 2 *$D(u_n)$ is satisfied if for any $A \in \mathcal{I}_{1,l}(u_n)$ and $B \in \mathcal{I}_{l+s,n}(u_n)$, where $\mathcal{I}_{j,l}(u_n)$ is the set of all intersections of the events of the form $\{R_i \leq u_n\}$ for $j \leq i \leq l$, and for some positive integer sequence $\{s_n\}$ such that $s_n = o(n)$, $|P\{(A \bigcap B)\} - P\{A\}P\{B\}| \leq \alpha(n, s)$ holds and $\alpha(n, s_n) \to 0$ as $n \to \infty$.*

Definition 3 *The $D''(u_n)$-condition (Chernick et al. (1991), Leadbetter & Nandagopalan (1989)) states that if the stationary sequence $\{R_t\}$ satisfies the $D(u_n)$-condition with $u_n = a_n x + b_n$ and normalizing sequences $a_n > 0$ and $b_n \in R$ such that for all x there exists $\mu \in R$, $\sigma > 0$ and $\xi \in R$, such that*

$$n\,(1 - F(a_n x + b_n)) \to (1 + \xi(x - \mu)/\sigma)_+^{-1/\xi}, \qquad as \qquad n \to \infty,$$

holds, where $(x)_+ = \max(x, 0)$, then it follows

$$\lim_{n\to\infty} n \sum_{j=3}^{r_n} P\{R_1 > u_n \geq R_2, R_j > u_n\} \quad = \quad 0,$$

where $r_n = o(n)$, $s_n = o(n)$, $\alpha(n, s_n) \to 0$, $(n/r_n)\alpha(n, s_n) \to 0$ and $s_n/r_n \to 0$ as $n \to \infty$.
For the $D'(u_n)$-condition the latter equation should be substituted by

$$\lim_{n\to\infty} n \sum_{j=2}^{r_n} P\{R_1 > u_n, R_j > u_n\} = 0.$$

We also consider two characteristics of clusters of exceedances, namely, the inter-cluster size defined as

$$T_1(u) \quad = \quad \min\{j \geq 1 : M_{1,j} \leq u, R_{j+1} > u | R_1 > u\} \tag{5}$$

and the cluster size defined as

$$T_2(u) = \min\{j \geq 1 : L_{1,j} > u, R_{j+1} \leq u | R_1 \leq u\},$$

where $M_{1,j} = \max\{R_2, ..., R_j\}$, $M_{1,1} = -\infty$, $L_{1,j} = \min\{R_2, ..., R_j\}$, $L_{1,1} = +\infty$.
The following geometric-like asymptotically equivalent distributions of $T_1(u)$ and $T_2(u)$ derived in Markovich (2014) require a mixing condition that is based on the mixing coefficient $\alpha_{n,q}(u)$ used in Ferro & Segers (2003).

Definition 4 *For real u and integers $1 \leq k \leq l$, let $\mathcal{F}_{k,l}(u)$ be the σ-field generated by the events $\{R_i > u\}$, $k \leq i \leq l$. We define the mixing coefficients $\alpha_{n,q}(u)$,*

$$\alpha_{n,q}(u) = \max_{1\leq k\leq n-q} \sup |P(B|A) - P(B)|, \tag{6}$$

where the supremum is taken over all $A \in \mathcal{F}_{1,k}(u)$ with $P(A) > 0$ and $B \in \mathcal{F}_{k+q,n}(u)$ and k, q are positive integers.

The mixing condition used in Markovich (2014) has been simplified in Markovich (2015a).

Theorem 1 *(Markovich (2015a)) Let $\{R_n\}_{n\geq 1}$ be a stationary process with the extremal index θ. Let $\{x_{\rho_n}\}$ and $\{x_{\rho_n^*}\}$ be sequences of quantiles of R_1 of the levels $\{1 - \rho_n\}$ and $\{1 - \rho_n^*\}$, respectively,*[1] *those satisfy the conditions (1) and (2) if u_n is replaced by x_{ρ_n} or by $x_{\rho_n^*}$ and, $q_n = 1 - \rho_n$, $q_n^* = 1 - \rho_n^*$, $\rho_n^* = (1 - q_n^\theta)^{1/\theta}$. Let $\{p_{n,1}^*\}$ and $\{q_{n,1}^*\}$ be sequences of positive integers such that $p_{n,1}^* = o(j)$, $q_{n,1}^* = o(p_{n,1}^*)$ for $j \geq j_0(n)$, $j_0(n) \to \infty$ as $n \to \infty$ and*

$$\alpha_{j+1,q_{n,1}^*} \quad = \quad o(\rho_n) \tag{7}$$

holds as $n \to \infty$, where $\alpha_{n,q} = \alpha_{n,q}(x_{\rho_n})$ is determined by (6). Then it follows for $j \geq j_0(n)$

$$\lim_{n\to\infty} P\{T_1(x_{\rho_n}) = j\}/(\rho_n(1 - \rho_n)^{(j-1)\theta}) \quad = \quad 1, \tag{8}$$

$$\lim_{n\to\infty} P\{T_2(x_{\rho_n^*}) = j\}/(q_n^*\,(1 - q_n^*)^{(j-1)\theta}) \quad = \quad 1. \tag{9}$$

[1] $\overline{F}(x_{\rho_n}) = P\{R_1 > x_{\rho_n}\} = \rho_n$.

3. First hitting time

Let $T^*(u_n)$ denote a first hitting time to exceed the threshold u_n. By definition it holds

$$P\{T^*(u_n) = j+1\} \quad = \quad P\{M_j \leq u_n, X_{j+1} > u_n\}, \tag{10}$$

$j = 0, 1, 2, ..., M_0 = -\infty$.

In Markovich (2015b) the following results concerning the first hitting time are derived.

Theorem 2 *Let all conditions of Theorem 1 be satisfied. We assume*

$$\sup_n \rho_n E(T_1(x_{\rho_n})) < \infty. \tag{11}$$

Then we get

$$\lim_{n\to\infty} nP\{T^*(u_n) = n\} \quad = \quad e^{-\theta\tau}/\theta, \tag{12}$$

$$P\{T^*(x_{\rho_n}) = j\} \quad \sim \quad \frac{\rho_n}{\theta}(1-\theta\rho_n)^{j-1}. \tag{13}$$

for $j \geq j_0(n)$, $j_0(n) \to \infty$ *as* $n \to \infty$.

Lemma 1 *Let the conditions of Theorem 2 be satisfied. Then it follows*

$$\lim_{n\to\infty} \rho_n ET^*(x_{\rho_n}) = 1/\theta^3. \tag{14}$$

Lemma 1 specifies the rough representation

$$\lim_{n\to\infty} E(T^*(u_n)/n) = 1/\theta. \tag{15}$$

given in Roberts et al. (2006).

4. PageRank

Ranking pages on the World Wide Web became very important taking into consideration that the Web is an interconnected graph with billions of nodes (pages). The Google's PageRank algorithm proposed by Brin and Page (1998) remains to be the most popular method to evaluate the importance of the page relative to other pages. Originally, the Google's PageRank defines the rank of the page p_i as

$$R(p_i) = \frac{1-c}{n} + c\sum_{p_j \in M(p_i)} \frac{R(p_j)}{D_j},$$

where $M(p_i)$ is the set of pages that link to p_i (in-degree), D_j is the number of outgoing links of page p_j (out-degree), c is a damping factor, usually $c = 0.85$, an n is the total number of pages. Taking into account a set of absorbing nodes, called dangling nodes $\mathcal{D}$, the PageRank can be generalized as

$$R(p_i) = c\sum_{p_j \in M(p_i)} \frac{R(p_j)}{D_j} + \frac{c}{n}\sum_{j\in\mathcal{D}} R(p_j) + (1-c)T(i), \qquad i = 1, ..., n,$$

or as the corresponding scale-free PageRank

$$R_i = c\sum_{p_j \in M(p_i)} \frac{1}{D_j} R_j + \frac{c}{n}\sum_{j\in D} R_j + (1-c)nT(i), \qquad i = 1, ..., n, \tag{16}$$

denoting $R_j = nR(p_j)$, Volkovich and Litvak (2010). Here, $T(i)$ is the teleportation jump, i.e. the probability that the walk starts afresh from page i. We shall focus on the simplest case when $T(i) = 1/n$ holds for every $i = 1, ..., n$. There can be other representations of $T(i)$, Avrachenkov et al. (2010), Volkovich and Litvak

(2010).
Following Volkovich and Litvak (2010), Jelenković and Olvera-Cravioto (2010) to model the tail of the PageRank distribution, the stochastic equation

$$R =^D \sum_{j=1}^{N} A_j R_j + B$$

is analyzed and the power law behavior of PageRank

$$P\{R > x\} \quad \sim \quad Hx^{-\alpha} \tag{17}$$

is derived, where H is a positive constant and α is a tail index. The latter class of distributions is also called the regularly varying distributions and it is denoted by $RV_{-\alpha}$. Here $=^D$ stands for equality in distribution. It is derived that the tail distribution of R is insensitive to the distribution of the out-degree D_j and, hence, to $A_j = c/D_j$. It is dominated by the in-degree N distribution if its tail is heavier than the tail of B or vice versa.
In Jelenković and Olvera-Cravioto (2010), N, $\{A_j\}_{j\geq 1}$, $\{R_j\}_{j\geq 1}$, B are mutually independent nonnegative random variables, and R_j's are distributed as R. In Volkovich and Litvak (2010) all random variables are assumed to be positive, N and B may be dependent and the moment restriction $E(A) = (1-E(B))/E(N) < 1$ holds. Assuming that N and B are regularly varying distributed and using the theory of regular variation, both cases where N or B dominates are considered and corresponding constants H in (1) are obtained.
In Jelenković and Olvera-Cravioto (2010) the rank process is investigated as a weighted branching process

$$R_{n+1} = \sum_{j=1}^{N_n} A_{n,j} R_{n,j} + B_n \tag{18}$$

for $n \geq 0$, where N_n, B_n, $A_{n,j}$ are mutually independent iid r.v.s. The ranks $\{R_{n,j}\}_{j\geq 1}$ belong to the nth generation of nodes of the branching process. The $\{R_{n,j}\}$ within each generation are assumed to be iid copies of R_n. Then a power law of R is derived from the prospective of the renewal theorem of Goldie (1991).
An analytical form of the extremal index of the PageRank random walk remains an open problem. However, the extremal index may be calculated by means of copulas

$$\theta = C'(1,1) - 1, \tag{19}$$

where $C(u,u)$ denotes a bivariate copula, Ferreira (2006), Ferreira & Ferreira (2007). (19) is valid if the Markov property for the PageRank process is fulfilled. In Avrachenkov et al. (2014) the Markov property was substituted by the $D''(u_n)$-mixing condition that seems to be a weaker condition.

Proposition 1 *(Avrachenkov et al. (2014)) If the sampled sequence is stationary and satisfies conditions $D(u_n)$ and $D''(u_n)$, then the extremal index is given by (19).*

Since the PageRank is based on a Markov chain of the node selection and it is a measurable function of stationary Markov samples, then according to O'Brien (1987) a mixing condition $AIM(u_n)$ that is stronger than the $D(u_n)$ is satisfied. Hence, the extremal index of the PageRank can be calculated by formula

$$\lim_{n\to\infty} P\{M_{1,p_n} \leq u_n | R_1 > u_n\} = \theta,$$

where $\{p_n\}$ is an increasing sequence of positive integers, $p_n = o(n)$ as $n \to \infty$.
Considering the PageRank as a branching process the Markov property implies the independence of 'grand parents-grand children' generations. The clusters of nodes may be homogeneous with respect to the $D''(u_n)$ condition, i.e. all nodes in cloud clusters may contain only large PageRanks. In case that the $D''(u_n)$ condition fails, the clusters may contain also nodes with small PageRanks.
The validity of the $D''(u_n)$-mixing condition regarding the PageRank process and its modifications is an open problem.

Example 1 *Markovich (2015c) Let us consider a so called survival copula* $\widehat{C}(\cdot,\cdot)$ *which corresponds to the equation*

$$P\{D_1 > x, D_2 > x\} = \widehat{C}(\overline{F}(x), \overline{F}(x)),$$

where $\overline{F}(x) = 1 - F(x)$ *is the tail function. In case of the bivariate Pareto distribution*

$$\bar{F}(d_1, d_2) = \left(1 + \frac{d_1}{\sigma} + \frac{d_2}{\sigma}\right)^{-\alpha} \tag{20}$$

with $\sigma = 1$ *and a marginal Pareto tail distribution* $\overline{F}(x) = (1+x)^{-\alpha}$, $\alpha > 0$, *and taking* $\overline{F}(x) = u$, *we have* $\widehat{C}(u,u) = (2u^{-1/\alpha} - 1)^{-\alpha}$, $u \in (0,1)$, *Nelsen (2006), p.29. Since* $C(1-u, 1-u) = \widehat{C}(u,u) - 2u + 1$ *holds, we finally obtain*

$$C(v,v) = (2(1-v)^{-1/\alpha} - 1)^{-\alpha} + 2v - 1,$$

denoting $v = 1 - u$. *Thus, we get*

$$C'_v(v,v) = 2\left(1 - (2 - (1-v)^{1/\alpha})^{-\alpha-1}\right).$$

Then from (19) we get $\theta = C'_v(1,1) - 1 = 1 - 2^{-\alpha}$, *where* α *is the shape parameter of the Pareto distribution. For* $\alpha = 1$ *we get* $\theta = 1/2$.
From (4) and since $F^{\leftarrow}(u) = (1-u)^{-1/\alpha} - 1$ *holds, we get the* $(1-\eta)$*th quantile of the maxima* M_n

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

Particularly, for $\alpha = 1$ *it follows* $x_\eta \approx \left(1 - (1-\eta)^{2/n}\right)^{-1} - 1$.

5. Analysis of real data of social networks

In Markovich (2015d) the statistical analysis of extremes in real data gathered in social networks is provided. The Enron email[2] and the DBLP[3] networks taken from Leskovec and Krevl (2014) and corresponding to undirected graphs are investigated.
It was shown that the Enron email node degrees can be fitted by the power law model (17) with $\alpha = 1.337$ whereas the DBLP data cannot. The latter conclusion is supported by a study of the mean excess function $e(u) = E(X - u|X > u)$ that shows Pareto-like behavior for the Enron data and a mixture of exponential and Pareto distributions for the DBLP data. Since the tail index regarding the DBLP data is smaller, i.e. $\alpha = 1.028$, this means that the tail distribution is heavier than the one arising from the Enron data. It is shown that the polynomial convergence rate of the Metropolis random walk to the stationary distribution is slower for the DBLP due to the heavier tail.
The estimation of the extremal index by intervals estimator (see Ferro & Segers (2003)) gives $\theta = 0.22$ and $\theta = 0.15$ for the Enron and the DBLP data, respectively. This implies a smaller mean cluster size for the Enron and a larger one for the DBLP. Moreover, the first hitting time is smaller for the Enron data. This implies a shorter time to get into a large node by means of a random walk.

6. Conclusions

The theoretical results obtained by author regarding the asymptotically equivalent distributions of cluster and inter-cluster sizes and the first hitting time are discussed. The application of the results to social networks is considered. The extremal index of the PageRank process is discussed. We have compared sampling techniques in social networks using new measures such as the extremal index, the distribution of the first hitting time and its mean. Considering the real data of social networks we conclude that a heavier tail of the node degree distribution leads to (1) larger node clusters around the giant nodes, (2) slower convergence rate of the Metropolis random walk, and (3) a longer first hitting time to reach a large node.

References

[2] Email communication network from Enron.
[3] DBLP collaboration network

Avrachenkov, K., & Ribeiro, B., & Towsley, D. (2010). Improving Random Walk Estimation Accuracy with Uniform Restarts. ed: Kumar, Ravi and Sivakumar, Dandapani, Lecture Notes in Computer Science, 6516 Algorithms and Models for the Web-Graph, 98-109, Springer, Berlin Heidelberg.

Avrachenkov, K., & Markovich, N. M., & Sreedharan, J. K. (2014). Distribution and Dependence of Extremes in Network Sampling Processes. Workshop on Complex Networks and their Applications at the 10th IEEE International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2014) November 23-27, 2014, Marrakech, Morocco, 331-338.

Brin S. & Page L. (1998). The anatomy of a large-scale hypertextual Web search engine, Comput. Networks ISDN Systems, 30(1-7), 107-117.

Chernick, M.R. & Hsing, T. & McCormick, W.P. (1991). Calculating the extremal index for a class of stationary sequences. Adv. Appl. Prob. 23, 835-850.

Chul-Ho Lee, & Xin Xu, & Do Young Eun (2012). Beyond Random Walk and Metropolis-Hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling, CoRR.

Ferreira, H. (2006). A new dependence condition for time series and the extremal index of higher-order Markov chains. REVSTAT - Statistical Journal.

Ferreira, A., & Ferreira, H. (2007). Extremal functions, extremal index and Markov chains. Notas e comunicações CEAUL, December.

Ferro, C.A.T., & Segers, J. (2003). Inference for Clusters of Extreme Values. J. R. Statist. Soc. B., 65, 545-556.

Goldie, C.M. (1991). Implicit renewal theory and tails of solutions of random equations. Ann. Appl. Probab., 1(1):126166.
Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57(1), 97-109.

Leadbetter, M.R., & Lingren, G., & Rootzén, H. (1983), Extremes and Related Properties of Random Sequence and Processes. Springer.

Leadbetter, M.R., & Nandagopalan, L. (1989). On exceedance point processes for stationary sequences under mild oscillation restrictions. Lect. Notes Statist. 51, 69-80.

Leskovec, J. & Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection, http://snap.stanford.edu/data

Nelsen, R. B. (2006). An Introduction to Copulas. Springer, 2nd edition.

Markovich, N.M. (2014). Modeling clusters of extreme values. Extremes. March 2014, Volume 17, Issue 1, 97-125.

Markovich, N.M. (2015a). Clusters of extremes: models, estimation and applications (submitted).

Markovich, N.M. (2015b). Hitting times of threshold exceedances and their distributions, ICRA 2015, Barcelona (accepted), arXiv:1501.01561

Markovich N. M. (2015c). Distribution and dependence of extremes in pagerank-type processes. Proceedings of the X International conference System identification and control problems SICPRO-2015 Moscow January

26-29 , 2015 V.A. Trapeznikov Institute of Control Sciences p.1229-1238, ISBN 978-5-91450-162-1.

Markovich N. M. (2015d). Extremes Control of Complex Systems With Applications to Social Networks, the 15th IFAC Symposium on Information Control Problems in Manufacturing, INCOM 2015, Ottawa, Canada (submitted).

Metropolis, N., & Rosenbluth, A.W., & Rosenbluth, M. N., & Teller, A. H., & Teller,E. (1953). Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 21(6), 1087-1092.

O'Brien, G.L.: Extreme values for stationary and Markov sequences. Ann. Probab. 15, 1, 281-291 (1987)

Roberts, G. O., Rosenthal, J. S., Segers, J., Sousa B. (2006). Extremal Indices, Geometric Ergodicity of Markov Chains and MCMC Scandinavian Journal of Statistics, 9, 213-229.