Nonparametric Analysis of Extremes on Web Graphs: PageRank Versus Max-Linear Model

Natalia M. Markovich[1], Maxim Ryzhov[1], and Udo R. Krieger[2(✉)]

[1] V.A. Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Profsoyuznaya Str. 65, 117997 Moscow, Russia
markovic@ipu.rssi.ru
[2] Fakultät WIAI, Otto-Friedrich-Universität,
An der Weberei 5, 96047 Bamberg, Germany
udo.krieger@ieee.org

Abstract. We analyze the cluster structure in large networks by means of clusters of exceedances regarding the influence characteristics of nodes. As the latter characteristics we use PageRank and the Max-Linear model and compare their distributions and dependence structure. Due to the heaviness of tail and dependence of PageRank and Max-Linear model observations, the influence indices appear by clusters or conglomerates of nodes grouped around influential nodes. The mean size of such clusters is determined by a so called extremal index. It is related to the tail index that indicates the heaviness of the distribution tail. We consider graphs of Web pages and partition them into clusters of nodes by their influence.

Keywords: Web graph · PageRank · Max-Linear model · Extremal index · Tail index

1 Introduction

The evaluation of the influence of nodes in a Web graph $G = (V, E)$ is an important problem of Web identification. PageRank (PR) and in-degree are the most popular indices of such influence. By Google's definition [2] PR is the rank of a Web page p_i. It is determined by

$$R(p_i) = c \sum_{p_j \in N(p_i)} \frac{R(p_j)}{D_j} + (1-c)\, q_i, \qquad i = 1, ..., n, \tag{1}$$

where $N(p_i)$ 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), and $c \in (0, 1)$ is a damping factor. $q = (q_1, q_2, ..., q_n)$ is a personalization probability vector such that $q_i \geq 0$ and $\sum_{i=1}^{n} q_i = 1$ holds, e.g. a uniform distribution $q_i = 1/n$, and n is the total number of pages p_i or corresponding nodes $i \in V$ of the Web graph G. The definition is simplified omitting the term relating to dangling nodes.

On the other hand, PR of a random page $i \in V$ can be considered as a weighted branching process

$$R_i = \sum_{j=1}^{N_i} A_j \; R_i^{(j)} + Q_i, \quad i = 1, ..., n, \tag{2}$$

denoting $R_i = R(p_i)$, $A_j =^d c/D_j$, $Q_i = (1-c)\; q_i$, [8,17]. Here, $\{R_i^{(j)}, j = 1, \ldots, N_i\}$ denotes the ranks of N_i nodes j with links outgoing to the node i. '$=^d$' denotes the equality in distribution. Moreover, PR was considered in [14] as an autoregressive process with random coefficients $\{A_j\}$ and a random depth N_i of dependence.

As an alternative to PR we use a Max-Linear Model (MLM), [5]. The MLM may be determined by the substitution of sums in (2) by maxima

$$R_i = \bigvee_{j=1}^{N_i} A_j R_i^{(j)} \vee Q_i, \quad i = 1, ..., n.$$

Such a model is practically useful when a largest rank of the most influential follower of a node is only available. In this case, (2) is not applicable.

Our first objective is to compare PR and the MLM by the tail and extremal indices. The tail index shows the heaviness of the distribution tail of the rank variable R_i. The reciprocal of the extremal index approximates the mean cluster size of ranks. We determine the *cluster* around a node of interest as a conglomerate of nodes connected to this node such that at least one node in the conglomerate has a rank that exceeds a sufficiently high threshold u.

Our second objective is a clustering of networks by evaluating the extremal indices of nodes. A node $i \in V = \{1, ..., n\}$ is considered as a root of a branching tree and its extremal index θ_i is estimated by samples of ranks of its followers. Since θ_i is the dependence measure around that node, the visualization of clusters of the network may be done by circles with diameter $1/\theta_i$ around each node. In [4,10] the clustering of nodes or the associated graph partition is proposed in terms of disconnected or weakly connected communities of nodes using samples of node indices. In this paper we develop a corresponding stochastic approach.

Usually, in- and out-degrees of nodes, i.e. the number of incoming and outgoing links of a node, are measured. They can be modelled by regularly varying distributions. The distribution function $F(x)$ is called regularly varying of tail index $\alpha > 0$ if $1 - F(x) \sim x^{-\alpha}\ell(x)$ as $x \to \infty$, where $\ell(x)$ is a slowly varying function, i.e. $\lim_{x\to\infty} \ell(tx)/\ell(x) = 1$, $\forall t > 0$, holds. In real-world networks $\alpha \in (1, 3)$ is observed, [3].

A term of the PR process that dominates its tail (i.e., one that has a smallest tail index α_{min}) may determine the cluster structure of the network controlled by the extremal index θ. We compare nonparametric estimates of α_{min} and θ for PR and the MLM by a study of a real network.

The paper is organized as follows. In Sect. 2 a theoretical basis of our study is given. We propose an adaptation of the blocks estimator of the extremal

index to a Web graph modelled by Thorny Branching Trees. In Sect. 3 we then compare the tail and extremal indices of PR and MLM by a study of a Web graph sample. Finally, we present some conclusions on our new nonparametric analysis approach.

2 Theoretical Foundation of Web Graph Modeling

We consider a Web graph $G = (V, E)$ with a sample $\{R_n, n \geq 1\}$ of a random rank variable, [8,13–15,17].

Definition 1 *([9], p. 53). The stationary sequence $\{R_n, n \geq 1\}$ 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 it holds*

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau, \qquad \lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta}, \tag{3}$$

where $M_n = \max\{R_1, ..., R_n\} = \bigvee_{j=1}^{n} R_j$ is used.

For independent r.v.s R_j $\theta = 1$ holds, but the converse is not true. $\theta \approx 0$ implies a strong dependence. As $1/\theta$ approximates the mean cluster size, $\theta = 0$ implies that the maximum M_n likely does not exceed a sufficiently high threshold u.

The practical significance of θ is that it determines the distribution of a first hitting time. This is the minimal time required to reach a sufficiently important node with a high rank [13]. The extremal index evaluates the mean first hitting time to find a subset of nodes with highest ranks in a network. This result helps to compare sampling random walks that are used to gather information about nodes and ranking algorithms like PR and the MLM.

It is a problem to get analytical formulae for θ of a PR process when the distributions of its components and their dependence are unknown.

For a given personalization vector $q_i = 1/n, 1 \leq i \leq n =\mid V \mid$, the scale-free PR $R_i^{(n)} = nR_i$ of a node i can be computed iteratively [17] by

$$\widehat{R}_i^{(n,0)} = 1, \quad \widehat{R}_i^{(n,k)} = \sum_{j\to i} \frac{c}{D_j} \widehat{R}_j^{(n,k-1)} + (1 - c), \quad k > 0, \tag{4}$$

until the difference between two consecutive iterations will be small enough. Here, $j \to i$ implies that node j links to node i, i.e. $(j, i) \in E$. To calculate the corresponding MLM values $\{X_i\}$ one can insert ranks obtained by (4) into

$$X_i = \bigvee_{j=1}^{N_i} \frac{c}{D_j} R_i^{(j)} \vee (1 - c), \quad i = 1, ..., n. \tag{5}$$

The stationary regularly varying distribution of PR R_i is derived in [8,17] under slightly different assumptions. Considering (2) and assuming that all r.v.s in the triple $(N_i, A_j R_i^{(j)}, Q_i)$ are mutually independent and that $\{N_i\}$, $\{A_j R_i^{(j)}\}$, $\{Q_i\}$ are sequences of iid regularly varying r.v.s, it is derived that the stationary

distribution of R_i is regularly varying with $\alpha = \min\{\alpha_N, \alpha_{AR}, \alpha_Q\}$, i.e. with the minimal tail index among the tail indices of all components in the triple. The same is proved in [14] under more relaxed conditions, i.e. $\{A_j R_i^{(j)}\}$ are assumed to be iid regularly varying r.v.s. It is derived therein that PR and MLM have the same tail and extremal indices.

An open question is whether the same tail and extremal indices are preserved for PR and the MLM in case that the ranks of followers of a node are dependent due to possible links among those followers. We check it for Web graphs by a nonparametric estimation of the tail and extremal indices.

2.1 Tail Index Estimation

Let $\{R_n, n \geq 1\}$ be a stationary sequence of r.v.s. of node ranks. To estimate the tail index α of these ranks, we use Hill's estimator [7] and the SRCEN estimator [16]. Hill's estimator is determined by

$$\widehat{\alpha}(n,k) = \left(\frac{1}{k} \sum_{i=1}^{k} \ln R_{(n-i+1)} - \ln R_{(n-k)} \right)^{-1}, \tag{6}$$

where $k \in \mathbb{N}$, $1 \leq k < n$, is the number of largest order statistics of $\{R_n\}$. It is the most popular estimator and it may be applied for $\alpha > 0$ and iid data.

The SRCEN estimator may be applied for $0 < \alpha < 2$. It is determined by

$$\widehat{\alpha}(n,b) = 2[n/b^2] \ln(b) / \sum_{i=1}^{[n/b^2]} \xi_i(b) \tag{7}$$

where $\xi_i(b) = \ln\left(\sum_{j=(i-1)b^2+1}^{ib^2} R_j^2\right) - 1/b \sum_{k=1}^{b} \ln\left(\sum_{j=(k-1)b^2+(k-1)b+1}^{(k-1)b^2+kb} R_j^2\right)$, and $[\cdot]$ denotes the integer part. We chop the data $\{R_1, ..., R_n\}$ into non-overlapping blocks of size b^2, e.g., $b = [n^{1/3}]$.

By a simulation it was shown that Hill is better than SRCEN for many cases of iid series, whereas SRCEN overcomes Hill for dependent data, [16].

2.2 Extremal Index Estimation by the Blocks Estimator

Regarding a graph structure we will use the blocks estimator [1] as the most appropriate one. Then a cluster is defined as a block of data $\{R_i\}$, where at least one observation exceeds a threshold u. The estimator states as follows

$$\widehat{\theta} = \frac{n \sum_{j=1}^{k} 1\left(M_{(j-1)r,jr} > u\right)}{rk \sum_{i=1}^{n} 1\left(R_i > u\right)}, \tag{8}$$

where $M_{i,j} = \max\{R_{i+1}, ..., R_j\}$, k is the number of blocks, $r = [n/k]$ is the number observations in the block, and $1(\cdot)$ is the indicator of an event.

In [15] a modification of the blocks estimator is proposed for Web graphs, where generations of followers of a root node in the branching tree are considered

(a) (b)

Fig. 1. Generations of followers of the root node in the branching tree used as blocks (1(a)), and graph of the Berkeley-Stanford dataset (1(b)).

as blocks (see Fig. 1(a)). The intuition behind (8) is that the blocks should not be overlapping. Typically, a node may be present in several generations due to loops, see Fig. 1(b). As the blocks of the generations are not equal-sized, one can use for a given threshold level u the ratio

$$\widehat{\theta}(u) = C(u)/N(u) \tag{9}$$

instead of (8), where $N(u)$ is the number of exceedances over u and $C(u)$ is the number of clusters. u is the most sensitive parameter of (9). It may be found visually corresponding to a stability interval of the plot $(u, \widehat{\theta}(u))$. For big data such as Web graphs we may select u by bootstrap methods, [12].

The same argument concerns the Hill' and SRCEN tools, where we have to select the number of the largest order statistics k and the block size b, respectively. For this purpose one may also use bootstrap methods, [12].

2.3 Bootstrap Method

In the following we briefly describe the bootstrap method to evaluate k in (6), [12].

Algorithm 1.

1. *Generate B re-samples $\{R_1^*, ..., R_{n_1}^*\}$ of size $n_1 < n$ with replacement from the original observations $\{R_i, 1 \le i \le n\}$, where n_1 is defined as*

$$n_1 = n^{\beta_b}, \qquad 0 < \beta_b < 1.$$

The number of the largest order statistics $k_1 \in \{1, ..., n_1 - 1\}$ corresponding to any re-sample relates to k and n by

$$k = k_1 \left(\frac{n}{n_1}\right)^{\alpha_b}, \quad 0 < \alpha_b < 1. \tag{10}$$

2. *Estimate B values $\widehat{\alpha}_{BS}(n_1, k_1, b)$ of the tail index by each $b \in \{1, \ldots, B\}$ of these B re-samples.*
3. *Calculate the mean squared error (MSE) by these re-samples,*

$$MSE(n_1, k_1) = (bias(n_1, k_1))^2 + \widehat{var}(n_1, k_1), \tag{11}$$

where the bias and the variance are determined by the following terms

$$bias(n_1, k_1) = \widehat{\alpha}_{BS}(n_1, k_1) - \widehat{\alpha}(n, k) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\alpha}_{BS}(n_1, k_1, b) - \widehat{\alpha}(n, k),$$

$$\widehat{var}(n_1, k_1) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\frac{1}{B}\sum_{b=1}^{B} \widehat{\alpha}_{BS}(n_1, k_1, b) - \widehat{\alpha}_{BS}(n_1, k_1, b)\right)^2,$$

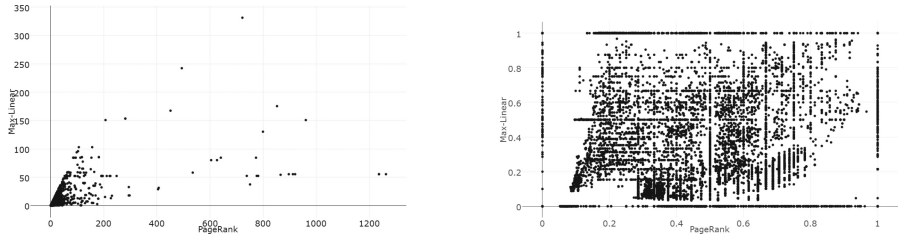
for a tail index estimate $\widehat{\alpha}(n, k)$ in (6) and find a minimal $MSE(n_1, k_1)$ among different $k_1 \in \{1, ..., n_1 - 1\}$.
4. *Using the obtained k_1, find the optimal k by (10) and then the corresponding estimate $\widehat{\alpha}(n, k)$ by (6).*

Replacing k and k_1 in Algorithm 1 by b and b_1, respectively, one can estimate the parameter b in (7) in the same way as the parameter k. In [6] it is recommended to choose $\alpha_b = 2/3$ and $\beta_b = 1/2$ for Hill's estimator. This selection leads to a bootstrap estimate of the MSE that is asymptotically close to the real MSE. To our best knowledge, the optimal values of the bootstrap parameters α_b and β_b are not obtained yet regarding SRCEN. But in this case we shall use the same values, too.

The same bootstrap algorithm can be applied to estimate u in (9), where k and k_1 in (10) may be interpreted as the total numbers of exceedances in the sample and in the re-sample, respectively. Then one can find u corresponding to the selected k and determine the estimate of the extremal index $\widehat{\theta}(u)$. In this case the values α_b and β_b are not precisely known due to the lack of theory and we may take $\alpha_b = 2/3$ and $\beta_b = 1/2$ as well. It is a subject of our future research to derive these values α_b and β_b by theoretical arguments.

3 Comparison of PageRank and the Max-Linear Model

We study the Web graph of the Berkeley-Stanford dataset in which nodes represent Web pages and edges represent hyperlinks between those pages, [11]. The graph contains 685230 nodes and 7600595 edges, [10]. We calculate PR and the MLM of each node by (4) and (5) with $c = 0.85$. The scatter plot in Fig. 2(a) shows the presence of outliers and, hence, the heavy-tailed distributions of PR and MLM.

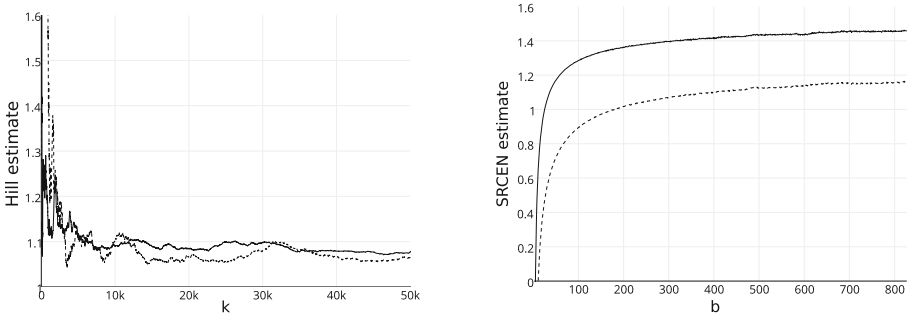


(a) (b)

Fig. 2. Scatter plots of the MLM versus PR for 150000 nodes (2(a)) and extremal indices of PR versus the MLM for 685230 nodes (2(b)).

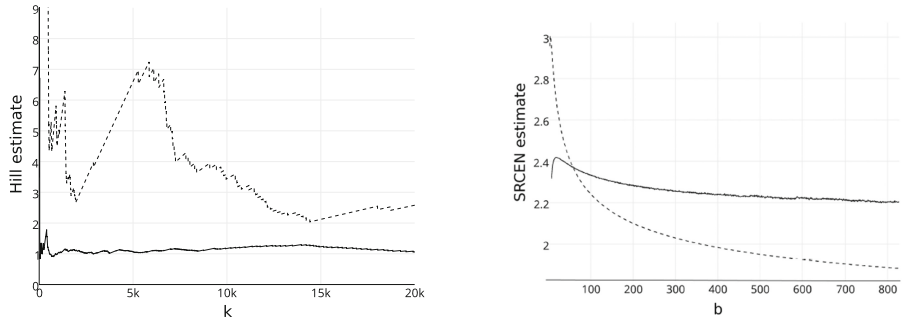
3.1 Tail Index Estimation

We estimate the tail index by PR and the MLM values that are obtained from the underlying datasets by the estimators (6) and (7) (see Fig. 3). Usually, the tail index value is taken according to a stability interval of the Hill's plot $(k, \widehat{\alpha}(n,k))$ regarding k. In the same way one can find the stability interval of the plot $(b, \widehat{\alpha}(n,b))$ of the SCREN estimator regarding b. Since the plots may have several stability intervals, we apply the bootstrap method with the number of bootstrap re-samples $B = 300$ and obtain the Hill's estimate equal to 1.081 and 1.052, and the SRCEN estimate equal to 1.3 and 1 for PR and MLM, respectively. Similar values can be obtained considering the first stability intervals from the left of the plots. Regarding the MLM, the values are closer for both estimators since the block-maxima used for the estimation in this case belong to the distribution tail in the same way as for the Hill' estimator that uses only the largest order statistics. As the tail index of PR and MLM are close to 1, this outcome implies that their distributions are likely regularly varying with infinite variance.



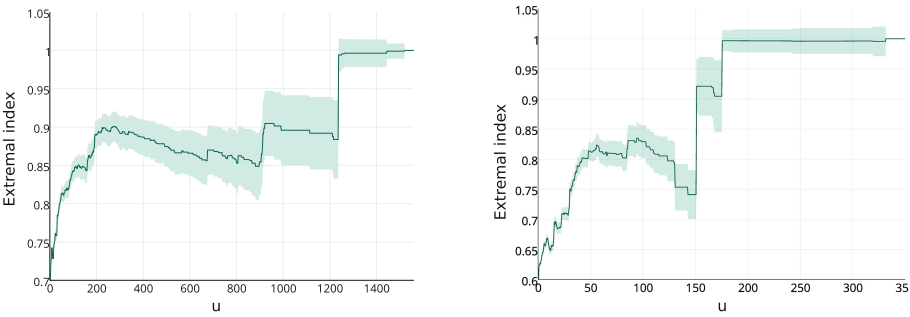
(a) (b)

Fig. 3. Tail index estimation by Hill's estimator (3(a)), and the SRCEN estimator (3(b)): PR (solid line), MLM (dashed line).



(a) (b)

Fig. 4. Tail index estimation of the in- and out-degrees by Hill's estimator (4(a)), and the SRCEN estimator (4(b)): in-degree (solid line), out-degree (dashed line).



(a) (b)

Fig. 5. Extremal index estimates by (9) for PR (5(a)) and the MLM (5(b)).

Moreover, we estimate the tail indices of the in- and out-degrees (see Fig. 4) and calculate their bootstrap values 1.026 and 2.730 for Hill's estimate and 2.2373 and 2.2650 for the SCREN estimate, respectively. The tail index of the in-degree is close to one which is a similarity regarding the tail index of PR and the MLM. This result implies that the distribution of the in-degree has a heavier tail than the distribution of the out-degree. Hence, the in-degree determines the heaviness of tail of PR and the MLM. This outcome is in the agreement with the results of [14,17].

3.2 Extremal Index Estimation of All Nodes in a Graph

To estimate the extremal index θ of the whole dataset, (then $1/\theta$ implies the mean cluster size over the whole network,) we select first generations of followers of each node as blocks. To avoid the overlapping of blocks, we copy the same sample 300 times and select blocks in such a way that each node belongs to only one generation. In Fig. 5 the blocks estimates of PR and the MLM averaged over

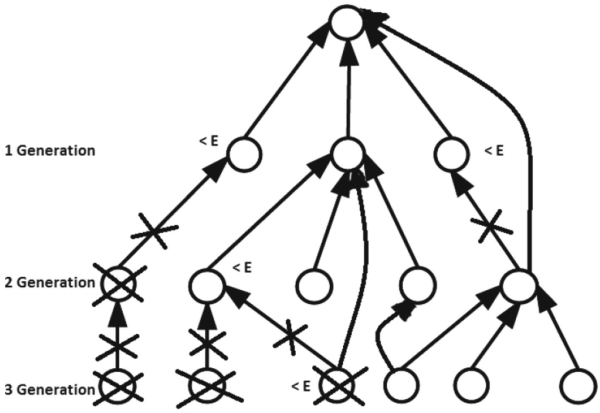


Fig. 6. Example of the truncation of the branching tree of a Web graph for a given ε.

300 samples are shown with the standard deviations. From the stability intervals we select an extremal index of about $\widehat{\theta} \in [0.85, 0.9]$ for PR and $\widehat{\theta} \in [0.8, 0.85]$ for MLM.

3.3 Extremal Index Estimation of an Individual Node

We estimate also the extremal index θ of each node. We consider a node as a root of the corresponding branching tree and generations of its followers as blocks. As the branching tree of a node can be very large, we propose the following truncation of the tree, see Fig. 6. Starting from the root, we take into consideration only a limited number of descendants using the following rule. If

$$\frac{c^{k-1}\widehat{R}_{j_k}}{\prod_{m=1}^{k-1} D_{j_m}\widehat{R}_i} < \varepsilon, \qquad 0 < \varepsilon < 1, \tag{12}$$

then the kth node will be included in the truncated graph but not its descendants. As a node may belong to different generations due to loops, some descendants may be preserved in the truncated graph as a member of the generations nearest to the root. The term on the left-hand side of (12) is arising by recursive replacements in (4) instead of $\widehat{R}_j^{(n,k-1)}$:

$$\widehat{R}_i = \sum_{j_1 \to i} \sum_{j_2 \to j_1} \dots \sum_{j_k \to j_{k-1}} \frac{c}{D_{j_k}} \cdot \frac{c^{k-1}}{D_{j_{k-1}} \cdot \ldots \cdot D_{j_1}} \widehat{R}_{j_k}$$
$$+ (1-c) \sum_{j_2 \to j_1} \dots \sum_{j_{k-1} \to j_{k-2}} \frac{c^{k-1}}{D_{j_{k-1}} \cdot \ldots \cdot D_{j_1}} + \ldots + (1-c) \sum_{j_1 \to i} \frac{c}{D_{j_1}} + (1-c).$$

Hereby, it is the intuition of this rule to exclude those nodes from the tree whose influence on PR of the root is weaker in the sense of (12). Then the extremal index is estimated by (9) using only generations of nodes of the truncated tree.

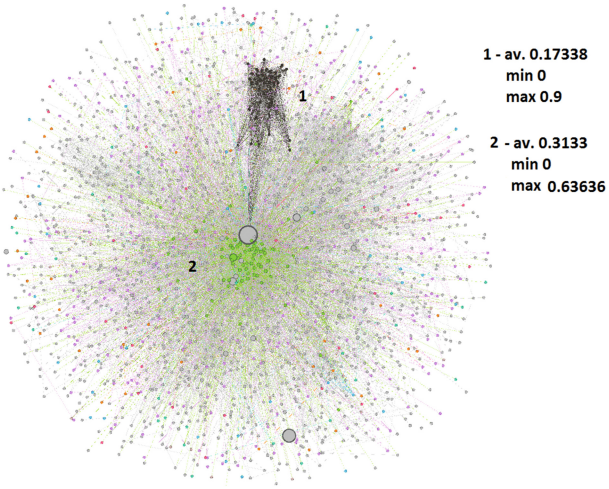


Fig. 7. Average, minimum and maximum of extremal indices of PR of nodes from two classes with MLM equal to 0.22 and 52.15 and colored by black - (1) and green - (2). (Color figure online)

To detect clusters we fix the values of u corresponding to the stability intervals of the plots in Fig. 5, i.e. $u_{PR} \approx 600$ and $u_{ML} \approx 75$ for PR and MLM, respectively. The scatter plot of the extremal indices of PR versus MLM is built for $\varepsilon = 0.01$. It shows diagonal trends which mean a similarity of the extremal indices of PR and MLM, see Fig. 2(b).

Figure 7 shows θ of PR for two classes with equal values of the MLM. Branching trees of depth equal to 7 associated with a node used to estimate θ may contain nodes lying outside these classes. The minimal index equal to zero is caused by the lack of exceedances over u w.r.t. PR of some nodes. The most valuable class with $MLM \approx 52.15$ has on average a mean cluster size approximately equal to $1/\theta = 1/0.313 \approx 3.195$, i.e. it includes at least 3 nodes with PR exceeding $u = 600$, and the class with $MLM \approx 0.22$ has a mean cluster size equal to 5.78.

In order to investigate the impact of ε we estimate the extremal index of PR of a triple of individual nodes for different values of ε by the blocks estimator (9) in Table 1. The threshold u corresponding to each estimate $\widehat{\theta}(u)$ is calculated by

Table 1. Blocks estimates of the extremal index of PR regarding three nodes in a Web graph with corresponding bootstrap estimates of u for different values of ε.

ε	"Black" node $PR = 6.48$			"Green" node $PR = 201.72$			"Grey" node $PR = 5031.31$		
	N	$\widehat{\theta}(u)$	u	N	$\widehat{\theta}$	u	N	$\widehat{\theta}$	u
0.01	20931	0.5	45	296146	0.86	1150	154449	0.77	1460
0.05	589	1	6.5	84372	0.65	1460	105386	0.68	1460

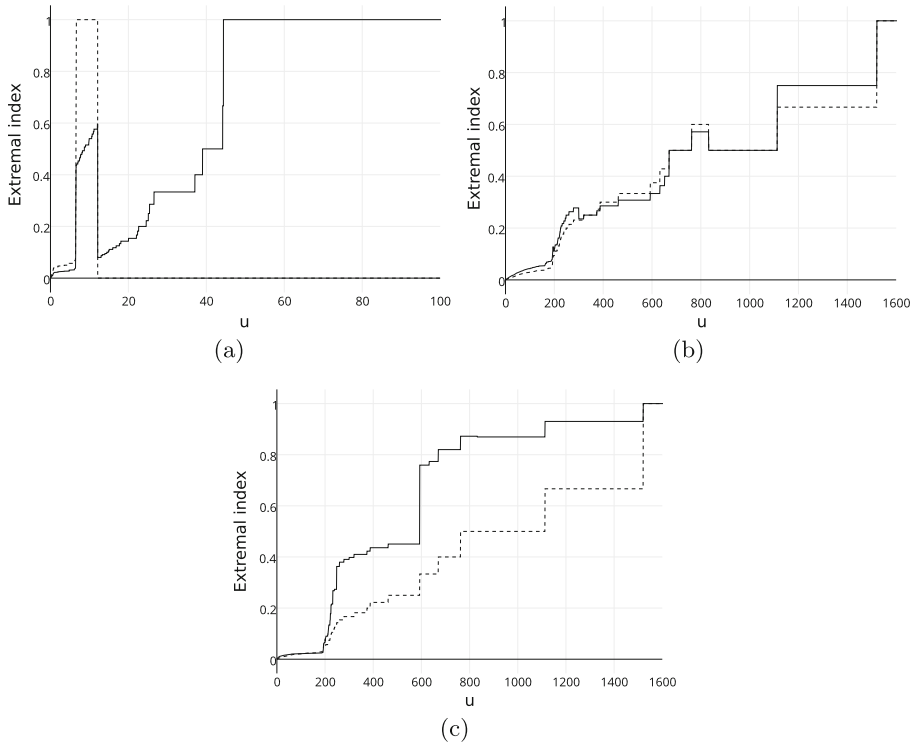


Fig. 8. Extremal index estimates by (9) for PR of the node from the black class (8(a)), of the node from the grey class (8(b)) and of the green node (8(c)) for $\varepsilon = 0.01$ (solid line) and $\varepsilon = 0.05$ (dashed line). (Color figure online)

the bootstrap algorithm described in Sect. 2.3. We select three nodes in Fig. 7: one is taken from the "black" class, one from the "green" class and one is a "grey" node located in the middle of the "green" class that has a large PR. The latter node does not belong to the "green" class but it has links to almost all nodes from the underlying network. The corresponding blocks estimates of the PRs of these nodes against the threshold u are shown in Fig. 8 for different values of ε. One may observe that the smaller the value of ε is the larger is the number N of the selected nodes in the truncated graph. This strongly impacts on the estimation of the extremal index. In order to calculate the blocks estimate well enough, we select a value u corresponding to the minimum of the bootstrap MSE (11), see Fig. 9. Then one can see in Fig. 9 the following tendency: the smaller ε corresponds to the larger optimal value of u. This outcome is achieved because the truncated branching tree contains a larger number of nodes in this case. Moreover, the "grey" node with the largest PR among all three nodes has the highest optimal u. One can select the following $u \approx 45, 1460, 1150$ from Fig. 9(a), (b) and (c) for $\varepsilon = 0.01$, respectively, for "black", "grey" and "green" nodes.

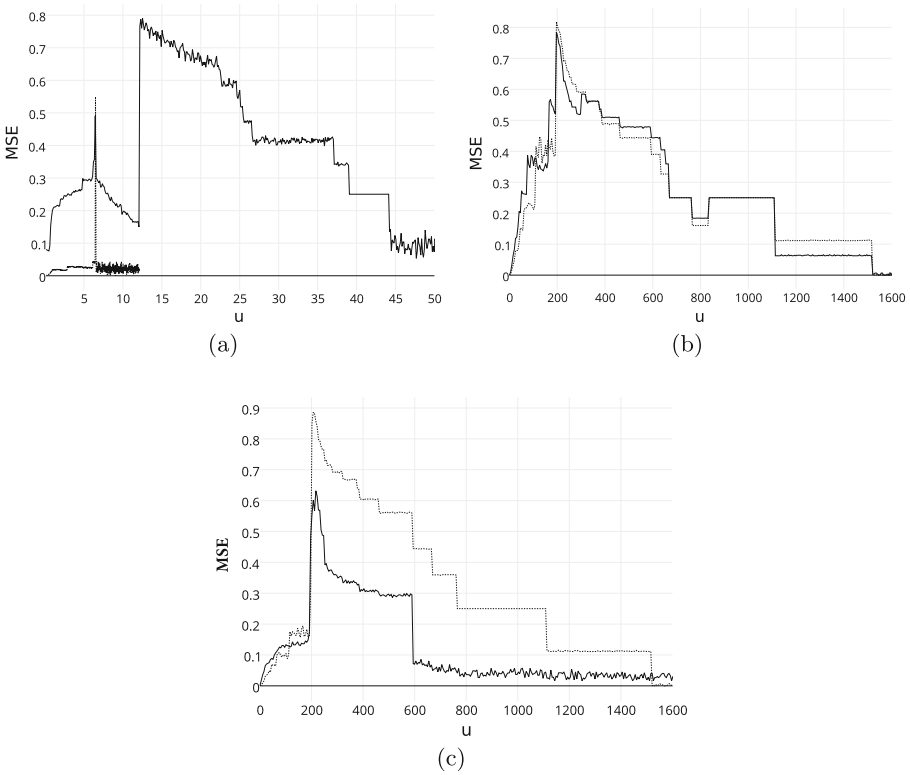


Fig. 9. The bootstrap estimate of the MSE against threshold u for three selected nodes and ε equal to 0.01 and 0.05, respectively: "black" node (9(a)), "grey" node (9(b)) and "green" node (9(c)) for $\varepsilon = 0.01$ (solid line) and $\varepsilon = 0.05$ (dashed line). (Color figure online)

This value u is the smallest threshold corresponding to the stability interval of smaller MSE. Hence, we obtain $\theta \approx 0.5, 0.86, 0.77$ for these nodes from Fig. 8 for $\varepsilon = 0.01$. Despite the PR of the "grey" node is the largest among all considered three nodes, its extremal index is the closest to one. This result implies that connections of this node are all arbitrary and independent. In other words, this node does not belong to a stable community with highly dependent links.

4 Conclusions

The paper is devoted to the stochastic analysis of a Web graph. Two characteristics of the node influence are considered, namely PageRank and a Max-Linear model. They are compared with regard to features of their underlying distributions and dependence structure. The latter dependence measure is represented by the extremal index of samples of the page rank variable.

Considering a Web graph, we propose a new clustering procedure of nodes by means of their extremal index values. Such clustering reflects the changes with regard to the extremal dependence structure of the Web graph. It may be an alternative to the clustering of nodes by the most distinct communities of nodes with a small number of edges between them. In our stochastic analysis approach we partition a real Web graph into clusters according to the extremal index values of the PageRank for equal MLM classes of nodes.

From the statistical point of view, the well-known nonparametric blocks estimator of the extremal index is modified in our paper with respect to random graphs. Considering the PageRank process corresponding to each node as an individual Thorny Branching Tree, we propose to utilize the new generations in such a tree as data blocks that are used by the blocks estimator. Due to loops in the graph such blocks may have common nodes. As the critical parameter of the blocks estimator is a threshold level, our next theoretical achievement is given by the proposal and the empirical study of a new bootstrap method to estimate this level. Due to the complexity of Web graphs several proposals to simplify the calculations of the extremal indices have been made. They include the truncation of the individual branching tree to calculate the extremal index of an individual node and replicating the same sample to select non-overlapping first generations as blocks to calculate the extremal index by the whole dataset of the node characteristics like PageRank or the Max-Linear model.

Our study of real Web graph data shows that PR and the MLM have similar tail and extremal indices. This result is in the agreement with our theoretical results [14]. It demonstrates the negligible impact of the dependence among generations of the branching trees associated with the nodes. The PR and MLM distributions are shown to be heavy tailed with an infinite variance.

Our future investigations will concern a theoretical study of the bootstrap procedure regarding the extremal index and a further study on the clustering of random graphs.

References

1. Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Comput. Netw. ISDN Syst. **30**(1), 107–117 (1998)
3. Chen, N., Litvak, N., Olvera-Cravioto, M.: PageRank in scale-free random graphs. In: Bonato, A., Graham, F.C., Prałat, P. (eds.) WAW 2014. LNCS, vol. 8882, pp. 120–131. Springer, Cham (2014). doi:10.1007/978-3-319-13123-8_10
4. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
5. Gissibl, N., Klüppelberg, C.: Max-Linear models on directed acyclic graphs. arXiv:1512.07522v1, pp. 1–33 (2015)
6. Hall, P.: Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. J. Multivar. Anal. **32**, 177–203 (1990)
7. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)

8. Jelenkovic, P.R., Olvera-Cravioto, M.: Information ranking and power laws on trees. Adv. Appl. Probab. **42**(4), 1057–1093 (2010)
9. Leadbetter, M.R.: Probability theory and related fields. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65**(2), 291–306 (1983)
10. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. eprint arxiv:0810.1355 (2008)
11. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection (2014). http://snap.stanford.edu/data
12. Markovich, N.M.: Nonparametric Analysis of Univariate Heavy-Tailed Data. Wiley, Chichester (2007)
13. Markovich, N.M.: Clustering and hitting times of threshold exceedances and applications. Int. J. Data Anal. Tech. Strat. 1–18 (2017, to appear)
14. Markovich, N.M.: Extremes in random graphs models of complex networks. arXiv:1704.01302v1, 5 April 2017 (2017)
15. Markovich, N.M.: Analysis of clusters in network graphs for personalized web search. In: IFAC 2017 World Congress, Toulouse, France, 7–14 July 2017 (2017, to appear)
16. McElroy, T., Politis, D.N.: Moment-based tail index estimation. J. Statist. Plan. Infer. **137**(4), 1389–1406 (2007)
17. Volkovich, Y., Litvak, N.: On the exceedance point process for a stationary sequence. Adv. Appl. Probab. **42**(2), 577–604 (2010)