Extremal properties of evolving networks: local dependence and heavy tails

Natalia Markovich

Received: date / Accepted: date

Abstract A network evolution with predicted tail and extremal indices of PageRank and the Max-Linear Model used as node influence indices in random graphs is considered. The tail index shows a heaviness of the distribution tail. The extremal index is a measure of clustering (or local dependence) of the stochastic process. The cluster implies a set of consecutive exceedances of the process over a sufficiently high threshold. Our recent results concerning sums and maxima of non-stationary random length sequences of regularly varying random variables are extended to random graphs. Starting with a set of connected stationary seed communities as a hot spot and ranking them with regard to their tail indices, the tail and extremal indices of new nodes that are appended to the network may be determined. This procedure allows us to predict a temporal network evolution in terms of tail and extremal indices. The extremal index determines limiting distributions of a maximum of the PageRank and the Max-Linear Model of newly attached nodes. The exposition is provided by algorithms and examples. To validate our theoretical results, our simulation and real data study concerning a linear preferential attachment as a tool for network growth are provided.

Keywords Network evolution · Tail index · Extremal index · PageRank · Max-Linear Model · Preferential attachment

1 Introduction

Extreme value theory concerning the sums and maxima of random sequences attracts a lot of interest due to numerous applications (see, Asmussen & Foss, 2018; Jessen & Mikosch, 2006; Lebedev, 2015; Markovich & Rodionov, 2020a;
Olvera-Cravioto, 2012; Robert & Segers, 2008; Tillier & Wintenberger, 2018). It has progressed in recent years from finite to random length sequences, particularly with regard to an application in random graphs and networks (Jelenkovic & Olvera-Cravioto, 2010; Jelenkovic & Olvera-Cravioto, 2015; Garavaglia et al., 2020; Volkovich & Litvak, 2010). Tail and extremal indices of the sums and maxima of non-stationary sequences of regularly varying random variables (r.v.s) with finite and random lengths were obtained in Goldaeva (2013), Markovich and Rodionov (2020a), Markovich (2021, 2022). While the tail index shows the heaviness of the distribution tail, the extremal index is a local dependence measure that shows the cluster structure of a stationary distributed sequence.

The distribution tail of a non-negative r.v. \( X \) is called regularly varying RV\(^{-\alpha}\) with the tail index \( \alpha \) if it holds

\[
F(x) = P\{X > x\} = x^{-\alpha}\ell(x),
\]

where the function \( \ell(x) \) is slowly varying, i.e. \( \lim_{x \to \infty} \ell(tx)/\ell(x) = 1 \) holds for any \( t > 0 \).

Let \( X^n = \{X_i\}_{i=1}^n \) be a sample of r.v.s with cumulative distribution function (cdf) \( F(x) \). By Leadbetter et al. (1983, p. 67) 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 it holds

\[
\lim_{n \to \infty} n(1 - F(u_n)) = \tau, \quad \lim_{n \to \infty} P\{M_n \leq u_n\} = e^{-\tau \theta}, \tag{1}
\]

where \( M_n = \max\{X_1, ..., X_n\} \). Particularly, the non-stationarity of \( \{X_n\} \) causes the non-existence of the extremal index. In case the extremal index exists, it allows us to obtain a limiting distribution of \( M_n \), namely, \( P\{M_n \leq u_n\} = F^\theta(u_n) + o(1) \) holds. For independent r.v.s \( \{X_1, ..., X_n\} \) it holds \( \theta = 1 \). The reciprocal of \( \theta \) approximates the mean cluster size. In this sense, it measures a local dependence. Throughout the article the cluster of exceedances defines a set of consecutive observations exceeding a threshold between two consecutive non-exceedances. Since such clusters may cause destructive events, the extremal index plays an important theoretical and practical role.

A network evolution arises in many applications like the World Wide Web, urban transport networks, citations between scientific articles, percolation theory to site and bond percolation (Bollobás & Riordan, 2006; Newman, 2018), information spreading (Censor-Hiller & Shachnai, 2010; Mosk-Aoyama & Shah, 2006), economic networks of trades (da Cruz & Lind, 2013) and many others. The evolution is of main interest, particularly with regard to the brain neurological networks (Bagrow & Brockmann, 2013; McCormick & Contreras, 2001), the infection spreading (Holme & Litvak, 2017), and the popularity of Web pages. The appending of new nodes and edges may be modelled by a preferential attachment (PA) (Ghoshal et al., 2013; Krapivsky & Redner, 2001; Newman, 2018; Norros & Reittu, 2006; Samorodnitsky et al., 2016; Wan et al., 2020; da Cruz & Lind, 2013) or the attachment depending on the clustering coefficient (Bagrow & Brockmann, 2013; Schroeder et al., 2022). The choice
of a seed network as a hot spot is also important for future attachments since
the network may be non-homogeneous.

Our objective is to obtain an evolved network with predicted tail and extremal indices of the PageRank (PR) and the Max-Linear Model (MLM) that are used as influence indices of the nodes. We apply the results obtained in Markovich and Rodionov (2020a), Markovich (2021, 2022) to the graph enlargement and assume that the node PRs of a seed network are regularly varying distributed r.v.s (see Appendices A and B for details).

We begin the attachment from a seed network consisting of stationary communities of nodes that may be connected by a few edges. The community consists of sets of nodes that are strongly connected with each other and weakly connected with nodes from other communities (Fortunato, 2010; Leskovec et al., 2009; Mester et al., 2021). A Directed Louvain’s Algorithm is a powerful tool to divide a graph into non-overlapping and weakly connected communities, Dugue & Perez (2015) (see Markovich et al. (2022) for an implementation). The definition and testing of the stationarity in the graphs remain an open problem. One can determine that a graph is stationary if for all finite subsets of vertices with the same adjacency matrices the joint distributions of their in- and out-degrees are the same (Markovich et al., 2022).

The communities can be ranked regarding their tail indices. The community with the minimum tail index determines the tail index of PRs and the MLMs of newly appended nodes that have each at least one edge with this “dominating” community. Tail indices of node’s in- and out-degrees for their power-law distributions were obtained in Samorodnitsky et al. (2016), Wan et al. (2020) depending on parameters of linear PA tools used for growing networks. In Banerjee and Olvera-Cravioto (2021) the tail behavior of the power law distribution of the PR of a uniformly chosen vertex in a directed preferential attachment (DPA) graph is obtained. It is shown that this power law is heavier than the tail of the limiting in-degree distribution. The DPA has a limited application since it assigns to each vertex a deterministic out-degree and it produces graphs without directed cycles. The PA schemes by Samorodnitsky et al. (2016), Wan et al. (2020) used here are free from these restrictions. To our best knowledge, the evolution of the tail and extremal indices of the PR and the MLM is considered here at the first time.

A bridge between the sums and maxima of random sequences of random lengths on one side and the PR and the MLM on the other side was given in Markovich (2022) by finding the conditions when the tail and extremal indices of the sum and maximum in the right-hand sides of equations (16) and (17) given in Appendix A are the same. This property was proved under practically plausible assumptions (see Appendix B). In this paper this approach is extended further to random graphs.

We determine the mean size of the cluster of exceedances in the graph from perspectives of extreme value theory. It is not evident how to identify clusters of high level exceedances in graphs and to calculate the mean size of the clusters as it is done for sequences of r.v.s due to an arbitrary enumeration of nodes in the graphs. Considering PRs and the MLMs of a set of root nodes as
sequences of sums and maxima of PRs of their nearest neighbors with in-coming links to the roots, one may determine the extremal index of PRs and the MLMs of the roots. We call this value the extremal index of the (sub)graph. It follows from Markovich (2021, 2022) that the extremal index of the subgraph (if it does exist) is determined by the extremal index of the most heavy-tailed ("dominating") sets of nodes within the subgraph, see Fig. 1(a). These sets can be obtained in the same way as communities. We assume that each "dominating" community contains a stationary regularly varying distributed set of node PRs with a minimum tail index.

Fig. 1  The set of root nodes (open circles) and their nearest neighbors (filled circles) where the most heavy-tailed community is marked by a rectangle with a dotted line (Fig. 1(a)); the creation of the enumerated columns of the next iteration matrix by the initial matrix and its submatrices by the "domino principle" (Fig. 1(b)).

The extremal index depends strongly on the mutual dependence between "dominating" communities. The latest results in Markovich (2022) contains conditions to obtain the extremal index of the PRs and MLMs of the root nodes, see Appendix B.

The paper is organized as follows. A problem description is given in Sect. 2.1. Theoretical constraints which are important for graphs are mentioned in Sect. 2.2. In Sect. 3 the results obtained in Markovich and Rodionov (2020a), Markovich (2021, 2022) are further developed to obtain the tail and extremal indices of PR and the MLM in an enlarged network started from a seed subgraph. Sect. 3 includes estimation methods and algorithms to implement the
ideas to simulated and real networks. Sect. 4 contains some conclusions. The paper is finalized by proofs in the Appendix, where necessary theoretical results concerning sums and maxima of random length sequences of regularly varying distributed r.v.s as well as methods relating to graphs are recalled.

2 Preliminaries

2.1 Problem Description

As in Markovich and Rodionov (2020a), Markovich (2021, 2022) we focus on a doubly-indexed array \( \{Y_{n,i} : n, i \geq 1 \} \) of nonnegative r.v.s in which the "row index" \( n \) corresponds to time, and the "column index" \( i \) enumerates the series. The length \( N_n \) of "row" sequences \( \{Y_{n,i} : i \geq 1 \} \) for each \( n \) is generally random. Namely, \( \{N_n : n \geq 1 \} \) is a sequence of non-negative integer-valued r.v.s. For each \( i \) the "column" sequence \( \{Y_{n,i} : n \geq 1 \} \) is assumed to be strict-sense stationary with extremal index \( \theta_i \) having a regularly varying distribution tail

\[
P\{Y_{n,i} > x\} = \ell_i(x) x^{-k_i}
\]

with tail index \( k_i > 0 \) and a slowly varying function \( \ell_i(x) \). There are no assumptions on the dependence structure in \( i \). Following Markovich and Rodionov (2020a), Markovich (2022) we consider the weighted sums and maxima

\[
Y_n^*(z, N_n) = \max(z_1 Y_{n,1}, ..., z_{N_n} Y_{n,N_n}),
\]

\[
Y_n^*(z, N_n) = z_1 Y_{n,1} + ... + z_{N_n} Y_{n,N_n}
\]

for positive constants \( \{z_i\} \), \( z_i > 0 \), \( i = 1, 2, ... \).

Let \( G_n = (V_n, E_n) \) be a directed graph with a set of vertices \( V_n = \{1, ..., n\} \), and a set of directed edges \( E_n \). The \( Y_n^*(z, N_n) \) and \( Y_n^*(z, N_n) \) may be interpreted as sums and maxima at the right-hand side of (16) and (17) (see Appendix A). Each sequence \( z_1 Y_{n,1}, ..., z_{N_n} Y_{n,N_n}, n \geq 1 \) represents the weighted influence indices of nodes in the one-link neighborhood from the root node \( n \). These neighbor nodes are marked by filled circles in Fig. 1(a). \( N_n \) denotes an in-degree of the root node \( n \) that is the number of its nearest neighbors with in-coming links to the root.

In Markovich (2021) \( A_j R_j, j \in \{1, ..., N_i\} \) in (16) and (17) is denoted as \( z_j Y_{i,j} \) with \( z_j = c \). By the definition of Google’s PR it follows that \( A_j = c/D_j \) holds, where \( D_j \) is the out-degree of the node \( j \) and \( c > 0 \) is a damping factor, the only parameter of the Personalized PR (Volkovich & Litvak, 2010). By Lemma A.1 (iii) in Volkovich and Litvak (2010) \( Y_{i,j} = R_j/D_j \) has the same tail index as \( R_j \) since \( R_j \) and \( D_j \) are assumed to be mutually independent and \( E(1/D_j) < 1 \) holds. Hence, the tail of \( Y_{i,j} \) is dominated by the tail of \( R_j \). One can rewrite the right-hand sides of (16) and (17) as

\[
Y_i(c, N_i) = c \sum_{j=1}^{N_i} Y_{i,j} + Q_i, \quad Y_i^*(c, N_i) = c \bigvee_{j=1}^{N_i} Y_{i,j} \lor Q_i, \quad i \in \{1, ..., n\}. (4)
\]
In Markovich (2022) the conditions were found when $Y_i(c, N_i)$ and $Y_i^*(c, N_i)$ have the same tail and extremal indices (see Appendix B). Each node in a random network is considered as a root of some directed graph of its followers which may contain cycles. As in Markovich (2021) we consider graph communities as the "column" series.

We aim to get the tail and extremal indices of an evolved graph starting from a seed set of nodes with known tail and extremal indices.

2.2 Important constraints

Let us mention the constraints of Theorem 2 (Appendix B) that are important for graphs.

1. The stationarity of the node's in-degrees $\{N_n\}$ is not assumed, but $N_i$'s have the same tail index and their distribution tail has to be lighter than the tail of the node PRs.
2. The in-degree $\{N_i\}$ of the ith node and the PRs $\{Y_{i,j}\}$ of the jth nodes that link to node i in (4) are independent.
3. The mutual pair-wise dependence between elements of the stationary d "column" sequences with minimum tail index has to be the same. Otherwise, the sequences of sums and maxima over d "row" elements corresponding to these columns are non-stationary and thus, the extremal indices of such sequences do not exist.
4. For each row at least one element corresponding to the "column" sequences with the minimum tail index has to be non-zero. Since the tail index has to be estimated one may deal with a unique community with a minimum tail index in a graph.
5. Elements of the "column" sequences with non-minimum tail index may be arbitrarily dependent. They may have different tail indices larger than the minimum tail index and hence, these "column" sequences may be non-stationary and have no extremal indices.
6. In terms of graphs, the communities may be considered as "column" sequences and only the d "dominating" communities are required to be stationary distributed. Note that d is generally a random variable.

3 Main results

3.1 Iterations

We focus on a directed graph $G_n$ with $n$ nodes. Theorem 2 (Appendix B) relates to a single iteration by ranks of the one-link neighbors of root nodes. It states that the tail index of PRs and the MLMs of the roots is determined by the tail index of their most heavy-tailed nearest neighbors. PR and the MLM of the root nodes have the same tail index and in some cases the same extremal index.
The connection between the PR of a node and the solution of (16) is proved by convergence in distribution of the $m$th iteration

$$R^{(m)} = \sum_{j=1}^{N} A_j R^{(m-1)}_j + Q, \quad m \geq 1,$$

corresponding to the Galton-Watson tree to $R$ as $m \to \infty$ starting from an initial distribution $R^{(0)}$. The r.v.s $\{R^{(m-1)}_j\}$ are assumed to be independent identically distributed (iid) copies of $R^{(m-1)}$. The r.v. with the heaviest tail among $N$ and $Q$. The initial distribution of $R^{(0)}$ is assumed to have a lighter tail than $N$ or $Q$ which both are regularly varying r.v.s. i.e. the iterations may start with $R^{(0)} \equiv 1$.

The tail behavior of $R^{(m)}$ is proved in Theorem 3.2 (Appendix B) omitting the independence Assumptions A (Appendix A) and assuming that the r.v.s $\{A_j R^{(m-1)}_j\}$ are non-stationary regularly varying distributed and a random number of the most heavy-tailed r.v.s $\{A_j R^{(m-1)}_j\}$ are independent or weakly dependent (conditions (A1) or (A2) in Appendix B). The statement of Theorem 2 (Appendix B) is similar to Proposition 3.1 in Volkovich and Litvak (2010), where $R^{(0)}$ is assumed to be a regularly varying r.v. with tail index $\alpha_R > 0$. If $P\{N > x\} = o(P\{R^{(0)} > x\})$ and $P\{Q > x\} = o(P\{R^{(0)} > x\})$, then $P\{R^{(m)} > x\} \sim C_R^{(m)} P\{R^{(0)} > x\}$ for all $m \geq 1$ as $x \to \infty$ is stated. $C_R^{(m)}$ is a constant depending on $m, \alpha_R$ and $E(N)$. The distribution of $R^{(\infty)}$, the unique nontrivial solution of (16), does not depend on the distribution of $R^{(0)}$, assuming that $E(R^{(0)}) = 1, E(A) = (1 - E(Q))/E(N) < 1$ and Assumptions A hold (Volkovich & Litvak, 2010, Theorem 3.1).

The convergence of the maximum recursion

$$R^{(m)} = \left( \bigvee_{j=1}^{N} A_j R^{(m-1)}_j \right) \lor Q,$$

to the solution $R$ of (17) under the Assumptions A as $m \to \infty$ and provided that the initial values corresponding to leafs possess a moment condition is derived in Jelenkovic and Olvera-Cravioto (2015). $\{R^{(m-1)}_j\}$ are independent copies of $R^{(m-1)}$ corresponding to the tree starting with an individual node $j$ in the first generation and ending at the $m$th generation.

3.2 Evolution from a seed subgraph (the "domino principle")

Let us describe the evolution process of the network by means of changing of matrices corresponding to the doubly-indexed array $\{Y_{n,i}\}$ determined in
Then \( m \) below is connected with the time. Using the notations of Sect. 2.1 we deal with recursions

\[
Y_{i,j}^{(m)} = c \sum_{s=j}^{N_i} Y_{i,s}^{(m-1)} + Q_i, \quad (5)
\]

\[
X_{i,j}^{(m)} = \left( c \bigvee_{s=j}^{N_i} X_{i,s}^{(m-1)} \right) \vee Q_i, \quad \{X_{i,j}^{(0)}\} \equiv \{Y_{i,j}^{(0)}\}, \quad (6)
\]

\( m, i, j \geq 1 \). Let us consider matrices related to the scheme of series \( \{Y_{n,i}^{(0)} : n, i \geq 1\} \) and corresponding tail and extremal indices \( (k_1^{(0)}, \theta_1^{(0)}) \):

\[
A^{(0)} = \begin{pmatrix}
Y_{1,1}^{(0)} & Y_{1,2}^{(0)} & Y_{1,3}^{(0)} & \cdots & 0 & Q_1 \\
Y_{2,1}^{(0)} & 0 & Y_{2,3}^{(0)} & \cdots & Y_{2,N}^{(0)} & Q_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
Y_{n,1}^{(0)} & Y_{n,2}^{(0)} & Y_{n,3}^{(0)} & \cdots & Y_{n,N}^{(0)} & Q_n \\
\end{pmatrix}, \quad (7)
\]

\[
\begin{pmatrix}
k_1^{(0)} \\
k_2^{(0)} \\
k_3^{(0)} \\
\vdots \\
k_N^{(0)} \\
\theta_1^{(0)} \\
\theta_2^{(0)} \\
\theta_3^{(0)} \\
\vdots \\
\theta_N^{(0)} \\
1 \\
\end{pmatrix}. \quad (8)
\]

\( \{Q_i\} \) is a sequence of iid r.v.s and thus, its extremal index is equal to 1. We start the evolution with a seed network that can be represented by a matrix \( A^{(0)} \). Network communities may be interpreted as columns of \( A^{(0)} \). A zero \( s \)th element in the \( i \)th row \( Y_{i,s}^{(0)} \), \( s \geq 1 \) of \( A^{(0)} \) means that the \( i \)th root node has no followers in the \( s \)th community or there is no link between them. For instance, if a row corresponds to a set of papers citing a book, then zero implies that the book is not cited by a paper from the corresponding community.

Without loss of generality we assume that \( k_1^{(0)} \leq k_2^{(0)} \leq k_3^{(0)} \leq \ldots \) holds. We may assume that the first \( d_1^{(0)} \) columns of \( A^{(0)} \) are the most heavy-tailed distributed with the minimum tail index \( k_1^{(0)} \), the next \( d_2^{(0)} \) columns have the second minimum tail index \( k_2^{(0)} \) etc. Generally, \( \{d_i^{(0)}\} \) are r.v.s.

We further consider the evolution iterations as follows: starting with this seed network, each time appending a set of nodes which transform the matrix \( A^{(m)} \). The nodes are appended by some attachment tool. The \( j \)th column \( \{Y_{i,j}^{(m)}\}_{i \geq 1} \) (or \( \{X_{i,j}^{(m)}\}_{i \geq 1} \)) of the matrix \( A^{(m)} \) is defined by (5) (or (6)) using the submatrix \( \{Y_{n,i}^{(m-1)} : n \geq 1, i \geq j\} \) (or \( \{X_{n,i}^{(m-1)} : n \geq 1, i \geq j\} \)) of the matrix \( A^{(m-1)} \). It is assumed that the newly attached nodes that built the matrix \( A^{(m)} \) have edges only with nodes of the subgraph corresponding to the matrix \( A^{(m-1)} \). The evolution looks like the ”domino principle” as \( j \) in (5) or (6) increases (Fig. 1(b)). What would be the extremal and tail indices of the ”column” sequences for the next iteration matrices \( \{A^{(m)}\}, m \geq 1? \)
Returning to the citation example, the "domino principle" means that some books may be cited by representative papers from all considered communities including the most influential ones or by papers from parts of the communities which are not so distinguish as the first ones. For instance, \( Y_{i,1}^{(1)} \) is calculated by all elements of the \( i \)th row of \( A^{(0)} \), as far as \( Y_{i,2}^{(1)} \) by the same row elements starting from the second one etc.

In Item (ii) of the next theorem, conditions (A1) – (A4) (Appendix B) with regard to elements of matrix \( A^{(0)} \) are assumed where \( d \) is replaced by \([d_n - 1]\), where \( d_n = \min(C, l_n)\), \( C > 1 \) and \( l_n \) satisfies (21), (22).

**Theorem 1** Let the conditions of Theorem 1 (Appendix B) with regard to \( \{Y_{n,i}^{(0)} : n, i \geq 1\} \) be fulfilled and at least one element in each row of \( \{Y_{n,i}^{(0)}\}_{i \geq 1} \) among the columns with tail index \( k_j^{(0)} \), i.e. \( \{Y_{n,i}^{(0)} : n \geq 1, d_j^{(0)} + 1 \leq i \leq d_j^{(0)} + d_j^{(0)}\}, d_j^{(0)} = 0, j \in \{1, 2, \ldots\} \) for each \( n \) be non-zero. Assume that \( d_j^{(0)} \) and \( \{Y_{n,i}^{(0)}\}_{i \geq 1} \) are independent for each \( j \geq 1 \).

(i) If \( d_j^{(0)} = 1, j \in \{1, 2, \ldots\} \) almost surely (a.s.), then \( \{Y_{i,j}^{(m)}\}_{i \geq 1} \) and \( \{X_{i,j}^{(m)}\}_{i \geq 1} \) calculated by (5) and (6) have the same tail index \( k_j^{(0)} \) and the same extremal index \( \theta_j^{(0)} \) for any \( m \geq 1 \);

(ii) Let \( \{d_j^{(0)}\}, j \in \{1, 2, \ldots\} \) be bounded discrete r.v.s such that \( 1 < d_j^{(0)} < d_n = \min(C, l_n), C > 1 \) holds.

(a) If (A1) or (A2) for any \( d_j^{(0)} \in \{2, 3, \ldots, d_n - 1\}, j \in \{1, 2, \ldots\} \) holds and \( N_n \) and \( \{Y_{n,i}^{(0)}\}_{i \geq 1} \) are independent, then \( \{Y_{i,j}^{(m)}\}_{i \geq 1} \) and \( \{X_{i,j}^{(m)}\}_{i \geq 1} \) have the tail index \( k_j^{(0)} \) for any \( j \geq 1 \) and \( m \geq 1 \).

(b) If (A4) where in (22) \( d_j^{(0)} \), \( j \in \{1, 2, \ldots\} \) is replaced by \([d_n - 1]\) holds, then \( \{X_{i,s}^{(m)}\}_{i \geq 1} \) has the extremal index \( \theta_j^{(0)} \) for any \( s \geq 1 \) and \( m \geq 1 \).

If, in addition, (A1) (or (A2)) for any \( d_j^{(0)} \in \{2, 3, \ldots, d_n - 1\} \) holds, then \( \{Y_{i,s}^{(m)}\}_{i \geq 1} \) has the same extremal index for any \( s \geq 1 \) and \( m \geq 1 \).

Theorem 1 is valid assuming that each row of \( A^{(0)} \) contains at least one non-zero element in the most heavy-tailed columns, i.e. the columns with the minimum tail index. Otherwise, the sums and maxima over rows may be non-stationary distributed with different tail indices. The elements of the matrices \( \{A^{(m)}\}, m \geq 1 \) of the next iterations are non-zero by their construction as row sums or maxima. Theorem 1 states that the limit distributions of recursions (5) and (6) depend on distributions of columns \( \{Y_{n,i}^{(0)} : n \geq 1, i \geq 1\} \).

Due to the complex nature of real-world networks the "non-zero assumption" may be rather restrictive when the number \( d \) of the most heavy-tailed columns is small. To overcome the problem, one can permute the rows of \( A^{(0)} \) to have blocks of rows with non-zero elements at least in one of the most heavy-tailed columns in the block.
Example 1 An example of such permutation is given by matrix $A_s^{(0)}$:

$$
A_s^{(0)} = \begin{pmatrix}
Y_{1,1}^{(0)} & Y_{1,2}^{(0)} & Y_{1,3}^{(0)} & \cdots & 0 & Q_1 \\
Y_{2,1}^{(0)} & Y_{2,2}^{(0)} & Y_{2,3}^{(0)} & \cdots & \cdots & Q_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
Y_{n_1,1}^{(0)} & Y_{n_1,2}^{(0)} & 0 & \cdots & 0 & Q_{n_1} \\
0 & Y_{n_1+1,2}^{(0)} & Y_{n_1+1,3}^{(0)} & \cdots & 0 & Q_{n_1+1} \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & Y_{n_2,2}^{(0)} & Y_{n_2,3}^{(0)} & \cdots & Y_{n_2,N_2}^{(0)} & Q_{n_2} \\
0 & 0 & Y_{n_2+1,3}^{(0)} & \cdots & 0 & Q_{n_2+1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & Y_{n_3,3}^{(0)} & \cdots & Y_{n_3,N_3}^{(0)} & Q_{n_3}
\end{pmatrix},
$$

where

$$
A_s^{(1)} = \begin{pmatrix}
Y_{1,1}^{(1)} & Y_{1,2}^{(1)} & Y_{1,3}^{(1)} & \cdots & Q_1 \\
Y_{2,1}^{(1)} & Y_{2,2}^{(1)} & Y_{2,3}^{(1)} & \cdots & Q_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
Y_{n_1,1}^{(1)} & Y_{n_1,2}^{(1)} & Y_{n_1,3}^{(1)} & \cdots & Q_{n_1} \\
0 & Y_{n_1+1,2}^{(1)} & Y_{n_1+1,3}^{(1)} & \cdots & Q_{n_1+1} \\
0 & \cdots & \cdots & \cdots & \cdots \\
0 & Y_{n_2,2}^{(1)} & Y_{n_2,3}^{(1)} & \cdots & Q_{n_2} \\
0 & 0 & Y_{n_2+1,3}^{(1)} & \cdots & Q_{n_2+1} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & Y_{n_3,3}^{(1)} & \cdots & Q_{n_3}
\end{pmatrix},
$$

is the matrix of the next iteration, $n_i = o(n)$, $i \in \{1, 2, \ldots\}$ as $n \to \infty$. The "column" series are assumed to be stationary regularly varying distributed and their tail indices $k_1^{(0)} < k_2^{(0)} < k_3^{(0)} < \ldots$ are increasing. We assume for simplicity that there is a unique column with a minimum tail index in each block. Zeroes in $A_s^{(0)}$ imply that the corresponding node has no out-going links to other nodes, particularly, to newly appearing ones. By Theorem 4 in Markovich & Rodionov (2020) the sums and maxima over each of the first $n_1$ rows in the first block have the tail index $k_1^{(0)}$, and over rows with numbers $\{n_1 + 1, \ldots, n_2\}$ in the second block the tail index $k_2^{(0)}$ etc. as $n = \sum_i n_i \to \infty$. Elements of the $i$th block of $A_s^{(0)}$ may contain zeroes apart of the $i$th "dominating" column, $i \in \{1, 2, \ldots\}$, but they do not impact the tail index of the $i$th column of $A_s^{(1)}$ since they have non-minimum tail indices. In terms of citations it means that a set of books cited by the most influential community inherits its tail and extremal indices. The part of the $i$th column within the $i$th block of $A_s^{(1)}$ corresponding to rows $\{n_{i-1} + 1, \ldots, n_i\}$, $i \geq 1$ with $n_0 = 0$ has the tail index $k_i^{(0)}$ as $n \to \infty$. The non-zero rest of the $i$th column in $A_s^{(1)}$
corresponding to rows 1, 2, ..., \( \sum_{j=1}^{i-1} n_j \) may be non-stationary distributed by Remark 2 (Appendix B).

In case the first \( d > 1 \) columns of \( A_s^{(0)} \) have the minimum tail index, e.g., \( k_1^{(0)} = ... = k_d^{(0)} < k_{d+1}^{(0)} < ... \), the first rows in \( A_s^{(0)} \) providing non-zero elements in the first \( d \) column are selected into one block. \( A_s^{(1)} \) can be partitioned into blocks with non-zero elements and used further for the next iterations.

**Example 2** Let us consider a citation network (Krapivsky & Redner, 2001; Newman, 2018). Each newly appearing paper cites a list of selected papers. In terms of networks the number of cited papers implies the out-degree of the new node (the paper), Fig. 2. Since papers are published continuously, the citation to some paper can appear in papers published later. PRs of the newly appearing papers in a unit time build a row in matrix \( A_s^{(0)} \). The sum and maxima of these PRs over the row provide the PR and the MLM of the paper published earlier. At time \( n \) one can build \( n \) rows of the matrix \( A^{(0)} \). A column of the matrix \( A_s^{(1)} \) of influence measures of previously published cited papers is built by \( A_s^{(0)} \). One has to detect the columns of \( A_s^{(0)} \) with the minimum tail index. It is assumed that each column of newly appearing papers in \( A_s^{(0)} \) with a minimum tail index is stationary distributed. According to Corollary 1 (Appendix B) other columns may be non-stationary distributed and include r.v.s with different tail indices. This allows us to obtain the tail and extremal indices of the sequence of previously published cited papers, i.e. the first "column" of \( A_s^{(1)} \). To predict the tail and extremal indices of other columns, the rest of the columns of \( A_s^{(0)} \) has to be stationary distributed.

The number of "dominating" communities (the columns of \( A_s^{(0)} \) with the minimum tail index) is plausible to be a bounded r.v.. There exists a "top" community among the latter ones such that its maximum PR is the largest. Then the MLMs of cited papers published earlier have the extremal and tail indices of the "top" community. If the communities with the minimum tail index are independent, then the PRs of cited papers have the same extremal and tail indices. This follows by Markovich (2022), see Appendix B. It implies that the citation by the "top dominating" community of newly appearing papers impacts the influence of the cited papers.

The network can be re-directed such that the rows of the matrix \( A_s^{(0)} \) contain previously published cited papers and the columns of \( A_s^{(1)} \) include the PR and the MLM of the newly appearing papers, see Fig. 3.

The evolution of the citation network is determined here by the dynamics of the tail and extremal indices which may predict the long-term citation impact of a set of publications. The mutual pair-wise dependence between r.v.s in the most heavy-tailed columns has to be the same and a part of each row related to the latter columns has to contain at least one non-zero element. Otherwise, sums \( Y_n(z, N_n) \) and maxima \( Y^*_n(z, N_n) \) are not stationary distributed, and the extremal index does not exist, see Example 1 in Markovich (2021).

**Remark 1** Theorem 1 is based on the assumption that node PRs of the initial seed graph corresponding to (7) are regularly varying distributed. This is sup-
Fig. 2 The scheme of a citation network with newly appearing papers marked by squares and cited papers marked by black circles; cited papers refer to papers published earlier. The scheme can be re-directed to the "ancestor" nodes taking a root of the tree as a newly appearing paper.

ported by empirical studies where the in-degree and PR of the Web are shown to follow a power law with the same exponent (Litvak et al., 2007; Pandurangan et al., 2002; Volkovich & Litvak, 2010). Marginal degree power laws for growing random networks were established in Krapivsky and Redner (2001). In Samorodnitsky et al. (2016), the joint distribution of in- and out-degrees in networks growing by a linear PA model is proved to have regularly varying tails. To obtain regularly varying PRs, one can consider nodes with PRs in as roots of classical branching trees whose leafs have PRs equal to 1 as in Chen (2014), Volkovich and Litvak (2010).

3.3 Selection of the seed matrix \( A^{(0)} \)

Next, we study how to define elements of the matrix \( A^{(0)} \) in and to find the tail and extremal indices of its columns. One can collect nodes by an attachment tool and calculate their PRs. Then, one can partition the obtained seed graph into communities (Clauset et al., 2004; Coscia, 2011; Dugné & Perez, 2015). The node’s PRs of the communities can be considered as the columns of \( A^{(0)} \).

For a given personalization vector \( Q_i = 1/n, 1 \leq i \leq n = |V_n| \), the scale-free PR \( R^{(n)}_i = nR_i \) of a node \( i \) can be computed iteratively (Chen et al., 2014) by

\[
\hat{R}^{(n,0)}_i = 1, \quad \hat{R}^{(n,k)}_i = \sum_{j \rightarrow i} \frac{c}{D_j} \hat{R}^{(n,k-1)}_j + (1 - c), \quad k > 0,
\]

until the difference between two consecutive iterations will be small enough. Here, \( j \rightarrow i \) implies that node \( j \) links to node \( i \), i.e. \((j, i) \in E_n\).
The mean field analysis is based on an aggregation of Web pages in classes according to pairs $k = (k_{\text{in}}, k_{\text{out}})$ of their in- and out-degrees and using averages of PRs within each $k$-degree class to calculate the PR (Fortunato et al., 2011).

The estimation of the tail index does not require an enumeration of elements of the sample. Thus, it can be estimated by one of the nonparametric estimators. The extremal index determines roughly the inverse of the expected cluster length. Its estimation may depend on the node enumeration. We propose a specific intervals estimator in Section 3.4.1 that allows us to avoid the node enumeration and select only a threshold $u$ as one parameter.

3.4 Empirical estimation

The following problems have to be solved regarding the estimation of the model parameters: (i) stationarity testing of communities; (ii) detection of pair-wise dependence between the elements of the most heavy-tailed communities that has to be the same. Stationarity of a community is required for the tail and extremal indices to exist according to their definitions. Condition (ii) has to be checked since the sequences of row sums and maxima have to be stationary distributed, otherwise their extremal index does not exist. Testing (ii) is limited due to the complexity of real-world networks. If the most heavy-tailed community is unique, then (ii) is naturally omitted.

The well-known nonparametric estimators of the extremal index for random sequences like blocks and runs estimators require usually the choice of a threshold $u$ and/or a declustering parameter, e.g. the block size (Beirlant et al., 2004). The intervals estimator (Ferro & Segers, 2003), estimators introduced in Robert (2009) and the $K$-gaps estimator (Süveges & Davison, 2010)) are threshold-based, i.e. they require a choice of $u$ as a single parameter. Since nodes belonging to a graph community cannot be enumerated, the estimators used for random sequences require a modification.

The tail index of a community of nodes may be estimated by one of the nonparametric methods based on the upper order statistics of the sample which is more appropriate for dependent data, e.g. the ratio estimator in Novak (2002), Resnick and Stărică (1999) or the SRCEN estimator in McElroy and Politis (2007). The Hill’s estimator

$$\hat{\alpha}(n, k) = \left( \frac{1}{k} \sum_{j=1}^{k} \log \left( \frac{X_{(n-j+1)}}{X_{(n-k)}} \right) \right)^{-1}$$

based on the $k$ upper order statistics $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ still works in practice, despite the network data may not be iid (Wang and Resnick, 2019, 2020; Wan et al., 2020). The value $k$ can be selected by minimizing the bootstrap mean squared error (MSE) (Markovich, 2007; Markovich et al., 2017) or by minimizing the Kolmogorov-Smirnov distance (Clauset et al., 2009; Drees et al., 2020; Wan et al., 2020).
We aim to predict the tail and extremal indices of the PR and the MLM of newly attached nodes of an evolving graph. The main restriction is that each new node is considered as a root of one-link graph of its neighbors and must have at least one neighbor with the minimum tail index of its PR. We propose first to partition a seed network into stationary distributed communities, e.g. by maximizing the modularity (Clauset et al., 2004; Newman, 2018). This can be done efficiently by a directed Louvain Method (Dugué & Perez, 2015). The stationarity of communities can be roughly checked by the mean excess function. Our idea is to find a set of “dominating” communities in the graph with a minimum tail index and start an attachment of new nodes to an existing graph to be sure that each newly appearing node has at least one link to the “dominating” communities. We allow a newly attached node to link to nodes of any community of the seed network and to assign the new node to class \(i\) if it has at least one directed link to nodes of the \(i\)th community. The classes may correspond to in- or out-degree. The \(i\)th community corresponds to the \(i\)th column of matrix \(A^{(0)}\) in Example 1. For instance, if \(i = 1\) holds, then Class 1 includes a set of nodes with PRs calculated by the upper block of \(A^{(0)}\).

Beforehand, we assign the code 00...0 of the length \(N_C\) to each new node. \(N_C\) is a number of classes equal to the number of communities. The communities are numbered in ascending order of their tail indices. Once a new node has an edge to the \(i\)th community, its code is replaced by 00...\(i\)...\(J\)...\(J\), \(J \in \{0, 1, 2, \ldots, N_C\}\), where \(i\) stands at the \(i\)th position. After \(n\) nodes are appended, we obtain \(n\) codes and classify the nodes. For example, for \(N_C = 3\) we assign the nodes with codes 123, 103, 120, 100 to Class 1, with codes 023, 020 to Class 2 and with code 003 to Class 3. One can distinguish classes regarding in-coming or out-going edges to the \(i\)th community. Since an attachment tool establishes an appearance time of new nodes, one can estimate the extremal index by an estimator that is determined for sequences, e.g., by the intervals estimator. The classification that is described in Algorithm 1 is supported by Theorem 1. It is explored in Sections 3.5, 3.6. Item 9(a) is justified by Theorem 4 in Markovich & Rodionov (2020), and Item 9(b) by Theorem 2.

3.4.1 Modification of the intervals estimator for graphs

If the extremal index of the community with enumerated nodes exists, then it can be estimated by the intervals estimator (Ferro & Segers, 2003)

\[
\hat{\theta}_n(u) = \begin{cases} 
\min(1, \hat{\theta}^1_n(u)), & \text{if } \max\{(T(u))_i : 1 \leq i \leq L - 1\} \leq 2, \\
\min(1, \hat{\theta}^2_n(u)), & \text{if } \max\{(T(u))_i : 1 \leq i \leq L - 1\} > 2,
\end{cases}
\]

where

\[
\hat{\theta}^1_n(u) = \frac{2(\sum_{i=1}^{L-1} (T(u))_i)^2}{(L - 1) \sum_{i=1}^{L-1} (T(u))_i^2},
\]

(10)
Algorithm 1 Classification of newly appended nodes
1: Select an initial directed graph with \( n > 1 \) nodes as a seed network and calculate the PRs of its nodes.
2: Partition the seed network into \( N_C \) communities, where \( N_C \) is predefined beforehand.
3: Check the stationarity of the communities and in case of non-stationarity select another seed network.
4: Estimate the tail index of each community, e.g. by the Hill estimator \( \hat{\theta}_J \), and rank communities in ascending order of their tail indices.
5: Attach \( N_0 \) new nodes and corresponding new edges by the PA schemes (see Appendix D) to the nodes of the communities.
6: Encode each newly appended node according to its edges to the \( i \)th community.
7: Assign each new node to one of \( N_C \) classes according to its code of the length \( N_C \): codes 1\( J \ldots J \) imply Class 1 and codes 00\( \ldots \)0\( J \ldots J \), \( J \in \{0, 1, 2, \ldots, N_C\} \) with \( i - 1 \) zeroes to Class \( i \), \( 2 \leq i \leq N_C \).
8: Estimate the extremal index of each community, e.g., by a modified intervals estimator presented in Algorithm 2.
9: Assign the minimum tail index to PRs and the MLMs of the new nodes from Class 1 and calculate their extremal index: (a) if the community with minimum tail index \( k_1 \) is unique, then the extremal index is equal to the extremal index \( \hat{\theta}_1 \) of the "dominating" community; (b) if there is a random number of "dominating" communities, then the MLMs of the new nodes from Class 1 have the extremal index of the "dominating" community with the maximum PR, and if, in addition, arbitrary enumerated sequences of node PRs of the "dominating" communities satisfy conditions (A1) or (A2) (see, Appendix B), then the PRs of the new nodes from Class 1 have the same extremal index as MLMs; the conditions (A1) or (A2) for the "dominating" communities provide the same minimum tail index \( k_1 \) for the PRs and the MLMs of the new nodes from Class 1.
10: Class 2 obtains the second minimum tail index corresponding to the next set of communities in the range and the respective extremal index as in item 9 and in the same way classes with numbers \( i > 2 \) obtain their tail and extremal indices.

\[
\hat{\theta}_J^J(u) = \frac{2\sum_{i=1}^{L-1}((T(u))_i - 1))^2}{(L - 1)\sum_{i=1}^{L-1}((T(u))_i - 1)((T(u))_i - 2)}, \quad (12)
\]

\( L - 1 \) is the random number of the inter-exceedance times \( \{(T(u))_i\} \). \( T(u) \) denotes a r.v. equal in distribution to r.v. 

\[
\min\{j \geq 1 : X_{j+1} > u\} \text{ given } X_1 > u,
\]

or it holds

\[
P[T(u) = n] = P[M_{1,n} \leq u, X_{n+1} > u|X_1 > u] \text{ for } n \geq 1,
\]
where $M_{i,j} = \max\{X_{i+1}, ..., X_j\}$, $M_{i,1} = -\infty$, for the underlying sequence $\{X_n\}$ with the cdf $F(x)$. For exceedance times $1 \leq S_1 < ... < S_L \leq n$ it follows

$$T(u)_i = S_{i+1} - S_i, \quad i \in \{1, ..., L - 1\}.$$  \hspace{1cm} (13)

$T(u)$ implies the number of observations running under $u$ between two consecutive exceedances. $T(u_n)$ normalized by the tail function $\{Y = T(u_n)T(u_n)\}$ is derived to be asymptotically exponentially distributed with a weight $\theta$ and with an atom at zero with a weight $1 - \theta$ (Ferro & Segers, 2003). Here, $\theta$ is the extremal index. The advantage of the intervals estimator is that it requires only the threshold value $u$ as a parameter. $u$ can be found as a high quantile of $\{X_n\}$ since then the inter-exceedance times are approximately independent.

The event $\{T(u) = 1\}$ corresponds to neighbor exceedances. Such $\{T(u)\}$ have to be excluded from consideration (S¨ uveges & Davison, 2010).

The intervals estimator was proposed for random sequences. Let us give an intuition, why the application of the intervals estimator to random graphs is plausible. Clusters of exceedances form asymptotically a Poisson process with a weight $\tau\theta$ where $\tau$ is taken from (1) and $\theta$ is the extremal index of the underlying process $X_i$ (Beirlant et al., 2004; Rootzén, 1988).

A naive approach is to determine locations of nodes with PRs consistently exceeding a sufficiently high $u$ by a Poisson process $P_n$ of rate $\tau\theta$. Roughly speaking, we superimpose the $P_n$ on the graph. The weight of an edge of the graph defined on a Poisson number of points may be taken equal to the inter-exceedance times. The latter may be measured as the number of nodes on the path (not equal to one) between a pair of nodes with exceedances of PRs. A rigorous proof is out of scope of the paper.

Nodes in random graphs can be arbitrarily enumerated. Thus, the intervals estimator has to be modified to avoid an enumeration of nodes (Algorithm 2).

Regarding a graph community, $T(u)$ can be taken equal to the length of the path expressed in edges between two nodes whose influence indices exceed the threshold $u$. All internal nodes along the path should have the influence indices less than $u$.

Regarding a simple example, the chain graph $G = (V, E)$ with edges $E = \{\{1, 2\}, \{2, 3\}, ..., \{m - 1, m\}\}, m = 6$ is shown in Fig. 3(a). We have $T(u)_1 = 3$ and $T(u)_2 = 2$. For the 2–barbell graph in Fig. 3(b) we get $\{T(u)\} = \{2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4\}$ excluding single edges. The number of inter-

---

**Algorithm 2** Modified intervals estimator

1: Let $\{X_i\}$ be influence node indices, e.g., PRs.
2: Take a high quantile of $\{X_i\}$ as threshold $u^*$.
3: Find nodes with exceedances, i.e., such that $X_i > u^*$ holds.
4: Set sequences $X_{xy} = \{X_x, X_{x_1}, X_{x_2}, ..., X_{x_m}, X_y\}, m \geq 1$ corresponding to paths $x \to \ldots \to y$ from a node $x$ to each node $y$.
5: Define $\{T(u^*)\}, i \in \{1, 2, ..., L\}$, by (13) for all sequences $X_{xy}$ such that influence indices $X_x$ and $X_y$ exceed $u^*$ but $X_{x_1}, X_{x_2}, ..., X_{x_m}$ do not, where $L$ is a total number of inter-exceedance times.
6: Estimate the extremal index $\hat{\theta} = \theta(u^*)$ of the graph by (10) (12).
Fig. 4 The $TBT_1$ (black circles), $TBT_2$ (grey circles) and $TBT_3$ (white circles) before (Fig. 4(a)) and after the preferential attachment $PA(0.4;0.2;0.4)$ with parameters $\delta_{in} = \delta_{out} = 1$ of 5493 (Fig. 4(b)) and $10^4$ (Fig. 4(c)) new nodes, where new nodes are shown by grey circles and “old” nodes by black ones; the circle sizes are proportional to their PR values.

3.5 Inference and simulation

We provide simulation results to support our theoretical conclusions. Our simulation concerns PRs of newly appended nodes only, but not their MLMs since the latter have the same tail and extremal indices by Markovich and Rodionov (2020a), Markovich (2021) and Theorem 1.

We generate three Thorny Branching Tree (TBT) graphs. The TBT is a variation of a branching tree where each node has an edge pointing to its parent. But it also has a certain number of unpaired outbound links that are pointing outside of the tree (Chen et al., 2014a). Algorithm 1 proposed in Chen et al. (2014b) generates a bi-degree sequence in such a way that the in- and out-degrees of nodes follow closely the desired regularly varying distributions and that the sums of in- and out-degrees are the same. The bi-degree sequences are later used to construct random graphs using the configuration model and TBT simultaneously. Our $TBT_1$, $TBT_2$, $TBT_3$ have power-law distributed in- and out-degrees with tail indices $(\iota_{in}, \iota_{out})$ equal to $(3.8, 2.0)$, $(2.5, 2.5)$, $(3.0, 4.5)$, respectively, Tab. 1. Due to the condition $\sum_{s=1}^{n} i_{in,s} = \sum_{s=1}^{n} i_{out,s}$, the tail indices of simulated in- and out-degrees may slightly differ from the initial values $(\iota_{in}, \iota_{out})$.

The TBTs are further connected by 100 additional edges to simulate the connection between TBTs, see Fig. 4(a). Each TBT contains 800 nodes. To append new nodes and edges, $\alpha-\beta-\gamma-$ schemes of the linear $PA(\alpha, \beta, \gamma)$ (see (25)-(27) in Appendix D) are used. The number of attached nodes $N_0$ is taken equal to 5493 and $10^4$, Fig. 4(b) and 4(c).

We evaluate the tail index of the PRs by Hill’s estimator where $k$ is selected by a bootstrap method. The consistency of Hill’s estimator for the
PR tail index has not been justified rigorously. The proof of the consistency is out of scope of this paper. The numerical consistency indirectly follows from Tab. 4 since Hill’s estimates decrease as the number of appending nodes \( N_0 \) increases. This reflects the appearance of giant nodes with a large node degree as the number of newly created edges to the TBTs increases.

We consider the TBTs (“old” nodes) before and after the PA of new nodes and edges, and classes of newly appended nodes to the existing ones by the PA schemes (“new” nodes). Bootstrap confidence intervals for the Hill’s estimates are calculated by 500 bootstrap resamples (Markovich, 2007).

The classes of “new” nodes are built by an encoding of nodes as in Algorithm 1. The “out-degree-classes” correspond to out-going edges from “new” nodes to “old” ones and vice versa for the “in-degree-classes”. For our TBT1-TBT3, a new node \( v \) may be encoded by one of the codes 123, 103, 120, 100 and related to “in-degree - class1” (or to “out-degree - class1”), if the edge \((v, w)\) leads from an “old” node \( w \) belonging to TBT1 to \( v \) (or vice versa). Class4 contains new nodes that have no links to “old” nodes from the TBTs but only to previously appended new nodes. The number of newly appended nodes \( N_0 \) (and thus, the cumulative size of Class1-Class4) is random. It may be less or equal to the number of evolution steps due to the \( \beta \)-scheme. The latter creates a new edge between two existing nodes and no new node is added. The PA-scheme is selected by means of a trinomial r.v. (see Appendix D). Class4 can be re-encoded and divided into classes with regard to their edges to Class1-Class3. The latter are considered further as seed communities.

Analyzing the tail indices of PRs of “old” nodes in Tab. 4 one can conclude that the TBT1 is the “dominating” community since it has the minimum tail index. The Hill’s estimates \( \hat{\alpha}_1(n, k) \) and \( \hat{\alpha}_3(n, k) \) as well as \( \hat{\alpha}_2(n, k) \) and \( \hat{\alpha}_4(n, k) \) of the corresponding TBTi and in(out)-degree - classi, \( i \in \{1, 2, 3\} \), are close, see Tab. 4 and Fig. 5. The closeness grows up as the number of appended nodes increases. The PR tail indices of the “old” nodes of TBTs decrease due to the appending of “new” nodes. Hence, their distribution tails become heavier. This property implies the appearance of giant components in

---

1 Since PRs are regularly varying distributed, the smaller positive tail index implies the heavier tail by Breiman’s theorem (Jessen & Mikosch, 2006; Markovich, 2007).
Table 1 The Hill’s estimates \( \hat{\alpha}_1(n, k) \) and \( \hat{\alpha}_2(n, k) \) of the PR tail index of "old" nodes in the \( TBT_1 - TBT_3 \) before and after the attachment of \( N_0 \) new nodes by the \( PA(0.4, 0.2, 0.4) \) with \( \delta_n = \delta_{out} = 1 \), \( \hat{\alpha}_3(n, k) \) and \( \hat{\alpha}_4(n, k) \) correspond to the "in-degree-classes" and "out-degree-classes" of sizes \( n \) with 97.5\% bootstrap confidence intervals \((u_1, u_2)\).

| Class \((i_{in}, i_{out})\) | \( \hat{\alpha}_1(n, k) \) | \( \hat{\alpha}_2(n, k) \) | \( n \) | \( \hat{\alpha}_3(n, k) \) | \( \hat{\alpha}_4(n, k) \) |
|-------------------------|-----------------|-----------------|------|-----------------|-----------------|
| \( TBT_1 \) (3.8, 2.0) | 3.3142 | 2.2937 | Class 1 | 967 | 3.1469 | 788 |
|                         | (2.53, 3.043) | (1.82, 3.041) | (2.29, 3.041) | (2.01, 3.772) |
| \( TBT_2 \) (2.5, 2.5) | 3.3513 | 2.4600 | Class 2 | 824 | 3.9242 | 728 |
|                         | (2.53, 2.89) | (1.83, 3.18) | (2.86, 4.23) | (2.26, 4.03) |
| \( TBT_3 \) (3.0, 4.5) | 3.4697 | 2.5518 | Class 3 | 608 | 3.8494 | 557 |
|                         | (1.84, 3.18) | (1.99, 3.51) | (3.23, 4.62) | (2.12, 4.03) |

with \( \delta_{in} = 1, \delta_{out} = 1 \).

\( \hat{\alpha}_3(n, k) \) and \( \hat{\alpha}_4(n, k) \) correspond to the "in-degree-classes" and "out-degree-classes" of sizes \( n \) with 97.5\% bootstrap confidence intervals \((u_1, u_2)\).

Table 2 Intervals and \( K \)-gaps estimates of PR extremal index of "old" nodes in the \( TBT_1 - TBT_3 \) before and after the \( PA(0.4, 0.2, 0.4) \) with \( \delta_n = \delta_{out} = 1 \) of 5493 new nodes and of the new nodes in the "in-degree-classes" and "out-degree-classes".

| \( gT_{IA} \) | \( g_{dis} \) | \( g_{KIMT} \) | \( gT_{IA} \) | \( g_{dis} \) | \( g_{KIMT} \) |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( TBT_1 \) 0.9167 | 0.8143 | 0.8282 | Class 1 | 0.9952 | 0.9479 | 0.9551 |
| (0.9167)(0.9311) | (1) | (1) | \( gT_{IA} \) | \( g_{dis} \) | \( g_{KIMT} \) |
| \( TBT_2 \) 0.8632 | 0.8098 | 0.8157 | Class 2 | 0.9977 | 0.9861 | 0.9879 |
| (0.9987)(0.9445) | (1) | (1) | \( TBT_3 \) 0.9804 | 0.9783 | 0.9808 | Class 3 | 0.9173 | 0.9356 | 0.9466 |
| (0.8506)(0.8579) | (0.9453)(0.9494) | \( TBT_1 \) 1 | 0.9345 | 0.9377 | 0.9525 | Class 1 | 0.2772 | 0.3130 | 0.3121 |
| (1) | (1) | (0.1961)(0.1883) | \( TBT_2 \) 0.3098 | 0.3175 | 0.3389 | Class 2 | 0.2715 | 0.2974 | 0.2818 |
| (0.3257)(0.3333) | (0.1140)(0.1091) | \( TBT_3 \) 0.9618 | 0.9070 | 0.9211 | 0.8529 | Class 3 | 0.2421 | 0.3970 | 0.4276 |
| (0.8732)(0.8817) | (0.1987)(0.1870) | \( N_0 \) = 5493 | \( N_0 \) = 10000 | \( N_0 \) = 5493 | \( N_0 \) = 10000 |

The same tail index drop is shown in Fig. 6, where the impact of the ratio \( |E|_{\text{added}}/|E|_{\text{init}} \) is represented. In total, 20000 edges and 16083 nodes were

the TBTs. It is in agreement with the conclusion that the PA leads to giant nodes, since each new vertex tends to get connected to big others, rather than to small ones (Krapivsky & Redner, 2001; Norros & Reittu, 2006). The tail index drop is stabilized as the number of appended nodes increases, since the newly appearing nodes gradually lose connection with the "old" ones. The increasing size of \( Class_4 \) reflects this effect also.
added. It is shown that the tail indices of both in- and out-degree classes tend to the tail indices of the TBTs before the evolution as the number of attached edges grows up. This observation is in agreement with the item (i) of Theorem 1 where the TBTs play the role of the columns of the matrix $A(0)$.

The values are also in agreement with formulae (2.9) in Wan et al. (2009) for tail indices of the in- and out-degrees: $\hat{\alpha}_{in} = (1 + \delta_{in}(\alpha + \gamma))/\alpha$ and $\hat{\alpha}_{out} = (1 + \delta_{out}(\alpha + \gamma))/\beta$. Regarding our model $PA(0.4, 0.2, 0.4)$ with $\delta_{in} = \delta_{out} = 1$ one obtains $\hat{\alpha}_{in} = \hat{\alpha}_{out} = 3$. The tail indices of the PR and the in-degree are comparable (Litvak et al., 2007).

The TBTs before the PA are stationary distributed by the simulation. We check the stationarity of PRs in the TBTs before the PA and in the classes of “new” nodes. To this end, we check the mean excess function $e(u) = E[X - u|X > u]$ and calculate its sample analogue $e_n(u) = \sum_{i=1}^n (X_i - u)1(X_i > u)/\sum_{i=1}^n 1(X_i > u)$, see Fig.7. The increasing, the decreasing and a constant value of $e(u)$ imply heavy-, light-tailed and exponential distributions, respectively (Markovich, 2007). The linear increase of $e(u)$ implies Pareto-like distributions of the underlying random sequences. Since for Pareto distribution

$$e(u) = (1 + \gamma u)(1 - \gamma), \quad \gamma < 1,$$

holds, where $\gamma = 1/\alpha$ is the reciprocal of the tail index, one can see that the lower curves at the plot of the mean excess function correspond to the larger $\alpha$. Fig.7 is in the agreement with Tab.1.

We estimate the extremal index of the TBTs before and after the PA as well as classes of the newly appearing nodes. The intervals and $K$-gaps estimators are used. The $K$-gaps estimator proposed in Süveges and Davison (2010) may show better accuracy than the intervals estimator according to a simulation study in Ferreira (2018), Markovich and Rodionov (2020b). It has the following form

$$\hat{\theta}_K^* = 0.5 \left( (a + b)/c + 1 - \sqrt{(a + b)/c + 1)^2 - 4b/c} \right),$$
with $a = L - N_C$, $b = 2N_C$, $c = \sum_{i=1}^{L} T(u_n)S(u_n)^{(K)}$. $N_C$ is the number of non-zero $K$-gaps. The $K$-gaps are determined by

$$S(u)^{(K)} = (\max (T(u) - K, 0)), \quad K = 0, 1, 2, ...$$

where $T(u)$ is the same r.v. as for the intervals estimator. Both the intervals and $K$-gaps estimators require sufficiently large samples since they are based on inter-exceedance times $\{T(u)_i\}$ which can constitute a very moderate sample. In order to estimate the threshold $u$ that is a single parameter of these estimators we use the $\omega^2$–discrepancy method proposed in Markovich and Rodionov (2020b). The latter estimates are denoted as $\hat{\theta}_{\text{Idis}}(n, k)$ and $\hat{\theta}_{K\text{dis}}(n, k)$. We take $k = [\hat{\theta}_0 L]$, where $\hat{\theta}_0$ is a pilot intervals estimator.

Using the algorithm in Markovich and Rodionov (2020b, Sect. 4), we find $u$ as solution of the discrepancy inequality in formula (12) by Markovich and Rodionov (2020b) and calculate

$$\hat{\theta}_1 = \frac{1}{l} \sum_{i=1}^{l} \hat{\theta}(u_i), \quad \hat{\theta}_2 = \hat{\theta}(u_{min})$$

as resulting estimates of the extremal index, where $u_1, ..., u_l$ are possible solutions of the discrepancy inequality, $l$ is their random number and $u_{min} = \min\{u_1, ..., u_l\}$. $\hat{\theta}_1$ and $\hat{\theta}_2$ are shown in Tab. 2. $\hat{\theta}_2$ is shown in brackets.

To find $u$ for the intervals estimator, we apply also the plateau-finding Algorithm 1 proposed in Ferreira (2018) denoted as $\hat{\theta}_{\text{IA}1}(n, k)$. To find an optimal pair $(u, K)$ in the $K$–gaps estimator we use the IMT method proposed in Fukutome et al. (2015). This estimate is denoted as $\hat{\theta}_{K\text{IMT}}(n, k)$. All estimates provide similar results, see Tab. 2. The ”out-degree-classes” and the $TBT_2$ after the PA demonstrate small estimates of the extremal indices except $\hat{\theta}_{K\text{IMT}}(n, k)$. This property may imply strong local dependence in these data sets. The rest of the TBTs and the ”in-degree-classes” have extremal indices close to 1 that may mean nearly independence. The closeness of the extremal indices of the ”in-degree-classes” and the TBTs before the PA is in the agreement with the item (ii) of Theorem 1. Really, the PRs of ”new” nodes belonging to an ”in-degree-class” are obtained as sums of the PRs of ”old” nodes (of the corresponding TBTs) with in-coming links to ”new” nodes.
Fig. 8 Examples of subnetworks of smaller sizes (Fig. 8(a)) and larger sizes (Fig. 8(b)) divided into communities \( \{ C_i \} \) received with Directed Louvain’s algorithm by the Berkeley-Stanford data, where point sizes are proportional to the nodes’ PRs.

3.6 Real data analysis

We investigate the Berkeley-Stanford web graph from 2002 with 685230 nodes and 7600595 edges (see snap.stanford.edu), which represents pages from the berkely.edu and stanford.edu domains that are connected in a union network by directed edges as hyperlinks between them (Leskovec et al., 2009). Within this network we select two examples of subnetworks of smaller and larger sizes (Fig. 8) and partition each of them into \( m \) communities by means of the Directed Louvain’s algorithm, based on modularity maximization (Dugue & Perez, 2015). PRs of nodes are calculated by (8).

These communities are used as seed networks. Starting from the seeds we apply the \( \beta^- \) and \( \gamma^- \)-PA schemes to evolve graphs and to obtain ”in-degree-classes” \( \text{Class}_1 - \text{Class}_m \) by Algorithm 1 (the \( \alpha^- \) and \( \beta^- \)-schemes lead to "out-degree-classes"). The \( \text{Class}_{m+1} \) contains new nodes which are appended to previously appearing new nodes but not to the seed communities. The latter class can be further partition into subclasses by Algorithm 1 with regard to their links to \( \text{Class}_1 - \text{Class}_m \) taken now as the seed network.

We check the stationarity of node PRs of the communities and the ”in-degree-classes” by the estimation of the mean excess function, Fig. 9. One may conclude that distributions of all communities and classes belong to the Pareto-type due to the linear increasing of the mean excess functions. The latter show the priority of the communities and classes by an ascending of their tail indices due to (14). Particularly, the community \( C_1 \) has the largest \( \gamma \) and thus, the smallest tail index \( \alpha \) as one can see in Tab. 3 too.

Fig. 10 shows that the cluster structure of PRs before and after the attachment of new edges is different. The PRs are enumerated with regard to the node appearance in the dataset and the sequence of further attachment. The
Fig. 9 Mean excess functions of the communities and the "in-degree-classes" of smaller sizes (upper line) and larger sizes (lower line) with appended edges obtained by 5000 and $10^4$ newly appearing nodes, respectively.

Fig. 10 The PRs of the "large" communities 1 – 4 before (upper line from left to right) and after (lower line from left to right) the $PA(0.4, 0.2, 0.4)$ with $\delta_{in} = \delta_{out} = 1$ of $10^4$ new nodes.
appearance of higher peaks after the PA corresponds to giant nodes with large PRs. The tail indices of the PRs after the PA may only decrease.

In Tab. 3 the Hill’s estimates $\hat{\alpha}(n, k_b)$ and $\hat{\alpha}^D(n, k_b)$ of the PR tail indices of "old" nodes in two sets of communities with appended new edges to the $N_0$ PA-appended "new" nodes and of new nodes in "in-degree-classes" with the 95% bootstrap confidence intervals $(u_1, u_2)$ obtained by 5000 bootstrap resamples, where the $PA(0.4, 0.2, 0.4)$ with $\delta_{in} = \delta_{out} = 1$ is used.

| Community n | $\alpha(n, k_b)$ | $\alpha'(n, k_b)$ | $\alpha''(n, k_b)$ | Class n | $\alpha(n, k_b)$ | $\alpha'(n, k_b)$ | $\alpha''(n, k_b)$ |
|-------------|-----------------|------------------|------------------|---------|-----------------|------------------|------------------|
| 1           | 1.1782          | 1.2677           |                   | Class n | 1.1146          | 1.5520           |                   |
|             | (0.3647, 1.9481)| (0.3719, 1.9921) |                  |         | (0.7713, 1.7581)| (0.7708, 1.7741) |                  |
| 2           | 1.2582          | 1.3020           |                   |         | 1.2280          | 1.4946           |                   |
|             | (0.3886, 2.1817)| (0.3776, 2.1500) |                  |         | (0.7804, 1.8732)| (0.7613, 1.8124) |                  |
| 3           | 1.8228          | 2.6509           |                   |         | 1.6247          | 1.4708           |                   |
|             | (0.6234, 9.2239)| (0.6965, 8.7533) |                  |         | (1.3935, 2.7271)| (0.7959, 3.0236) |                  |
| 4           | 1.2493          | 1.0752           |                   |         | 1.8196          | 1.7822           |                   |
|             | (0.5661, 5.0906)| (0.5725, 5.0560) |                  |         | (0.7677, 2.9515)| (0.7671, 3.0409) |                  |
| 5           | 1.9378          | 1.4255           |                   |         | 2.8095          | 3.6141           |                   |
|             | (0.5846, 14.1712)| (0.5855, 12.4387)|                  |         | (2.2747, 3.3840)| (1.6424, 4.2651) |                  |

In Tab. 4 the Hill’s estimates $\hat{\alpha}(n, k_b)$ and $\hat{\alpha}^D(n, k_b)$ of the PRs of the communities of both smaller and larger sizes after the appending of $N_0$ new nodes and the "in-degree-classes" of appended new nodes are shown. Due to a possible non-stationarity the tail index of the rest Class$_{m+1}$ is not estimated. The number of largest order statistics $k_b$ of $\hat{\alpha}(n, k_b)$ is obtained by a bootstrap method and of $\hat{\alpha}^D(n, k_b)$ by a double bootstrap method (see Markovich (2007) for details). The tail indices of the in-degree-classes are larger than ones of the evolved communities and correspond to the tail indices of the communities before the evolution. The drop and stabilization of the tail indices of the communities and "in-degree-classes" is shown in Fig. 11 when the number of new edges increases. In Tab. 4 estimates (15) of the extremal index are shown when the evolution starts from the "large" communities by the $PA(0.4, 0.2, 0.4)$ with $\delta_{in} = \delta_{out} = 1$. The same estimators of the extremal index as in Tab. 2 are used. The evolution from the "small" communities is not considered due to a lack of data. The extremal indices of the "in-degree-classes" are close to those of the communities before the PA. After the PA of new edges the extremal indices of the communities are decreased which implies the increasing of their clustering.
Extremal properties of evolving networks: local dependence and heavy tails

Fig. 11 The Hill’s estimates \( \hat{\theta}(n, k) \) of PR of the large size communities (Fig. 11(a)) and in-degree classes (Fig. 11(b)) with \( k \) selected by bootstrap method versus \( |E_{\text{added}}|/|E_{\text{init}}| \), where \( |E_{\text{added}}| \) is the number of edges added to the communities during the PA and the initial number of edges \( |E_{\text{init}}| \) in the communities.

Table 4 Estimation of the PR extremal index of the "old" nodes in the "large" communities and the "in-degree-classes" of \( N_0 \) appended new nodes.

| Community | \( \hat{\theta}^{IA} \) | \( \hat{\theta}^{Ids} \) | \( \hat{\theta}^{KID} \) | \( \hat{\theta}^{KIMT} \) |
|-----------|----------------|----------------|----------------|----------------|
| Before PA |                |                |                |                |
| 1         | 0.9983 0.9781 0.9842 0.8991 | 1 0.9356 0.9419 0.9475 0.8851 | (0.9449)(0.9455) | (0.8033)(0.8065) |
| 2         | 0.9919 0.9874 0.9912 0.9124 | 2 0.8924 0.9245 0.9284 0.8951 | (0.9249)(0.9324) | (0.8595)(0.8643) |
| 3         | 1 0.9988 1 0.9278 | 3 0.9583 0.9737 0.9745 0.9379 | (1) (1) | (1) (1) |
| 4         | 0.6138 0.7001 0.7081 0.8399 | 4 0.8982 0.9357 0.9424 1 | (0.6179)(0.6218) | (1) (1) |
| After PA of \( N_0 = 10^4 \) nodes |                |                |                |                |
| 1         | 0.4877 0.4756 0.4684 0.6823 | 0.4817(0.4785) | 0.4817(0.4785) | (0.4817) (0.4785) |
| 2         | 0.2538 0.2409 0.3528 0.5523 | 0.3462(0.3322) | 0.3462(0.3322) | (0.3462) (0.3322) |
| 3         | 0.8638 0.8957 0.9039 0.7484 | 0.6771(0.6947) | 0.6771(0.6947) | (0.6771) (0.6947) |
| 4         | 0.5185 0.6716 0.6769 0.7364 | 0.5985(0.6193) | 0.5985(0.6193) | (0.5985) (0.6193) |

4 Conclusion

The prediction of the tail and extremal indices of node influence characteristics of an evolving network is studied. Assuming that the network can be partitioned into stationary distributed communities of nodes, we classify the newly appended nodes according to their edges to the communities. Ranking the communities by their tail indices in ascending order, one can select the most heavy-tailed ("dominating") community with a minimum tail index. We assign a set of new nodes to the first class with the tail and extremal indices of
the latter community if each node of this set has at least one edge to nodes of the "dominating" community and the latter is unique. In the next step, we repeat the procedure with the rest of the communities finding the "dominating" one among them and classifying the rest of the newly appearing nodes. Clearly, the procedure can be done in discrete time moments. The same procedure can be applied if there are a random number of "dominating" communities since the communities of the seed network are independent or weak dependent due to a few links between them.

The assumption regarding the uniqueness of the community with a minimum tail index is plausible and not restrictive for graphs since one has anyway to estimate the tail indices of PRs. The tail index estimates are likely different. This uniqueness property simplifies the analysis. Since the most heavy-tailed "column" series is likely unique, the checking of the homogeneous pair-wise dependence between the components of the most heavy-tailed communities required for an application of Theorems 1 and 2 and thus, for a prediction of the tail and extremal indices of new classes can be omitted. In case, the community with the minimum tail index is not unique, the pair-wise dependence can be investigated as in Appendix C. In fact, the "dominating" community with the largest maximum PR determines the extremal index of the PR and MLM of the set of newly appended nodes.

There are some problems of the graph analysis that make it complicated and rough: (a) nodes are not enumerated; (b) the dependence between nodes may be complex and non-homogeneous; (c) the node characteristics in communities may be non-stationary distributed. The estimation of the tail index is based on the largest order statistics of the sample and hence, it does not require any enumeration of the nodes. In contrast, the estimation of the extremal index depends on the node enumeration.

To estimate the extremal index we apply the intervals and $K$-gaps estimators. To modify the intervals estimator for graphs, we propose to take the number of nodes at the paths between the nodes with exceedances of some feature (e.g., PR) over a sufficiently high threshold as the inter-exceedance times. The node attachment like the preferential attachment provides a natural enumeration of nodes. Thus, one can use the intervals estimator for random sequences of the PRs or the MLMs for newly appended classes of nodes of evolving networks.

The stationarity of the communities is proposed to be checked by the mean excess function that does not require the enumeration of nodes.

### A The PageRank and Max-Linear Model

PR (Langville & Meyer, 2006) and the MLM may be considered as node influence characteristics (Gissibl & Klüppelberg, 2018; Markovich et al., 2017). The PR $R$ of a randomly chosen Web page (a node in the Web graph) is viewed as a r.v.. It was considered as the...
solution to the fixed-point problem

\[ R = D \sum_{j=1}^{N} A_j R_j + Q \]  

in Jelenkovic and Olvera-Cravioto (2010), Volkovich and Litvak (2010). \( R \overset{D}{=} \) denotes equality in distribution. The r.v.s \( \{R_j\} \) are assumed to be iid copies of \( R \) and \( E(Q) < 1 \) holds. \( Q, N, \{A_j\} \) is a real-valued vector, \( \{A_j\} \) are independent non-negative r.v.s distributed as some r.v. \( A \) with \( E(A) < 1 \). \( N \) denotes the in-degree of a node, \( Q \) is a personalization value of the vertex. Under the assumptions (we shall call them Assumptions A) that \( \{R_j\} \) are regularly varying iid and independent of \( \{Q, N, \{A_j\}\} \) with \( \{A_j\} \) independent of \( \{N, Q\} \), \( N \) is regularly varying r.v., and that \( N \) and \( Q \) are allowed to be dependent, it is stated in Jelenkovic and Olvera-Cravioto (2010), Volkovich and Litvak (2010) that the stationary distribution of \( R \) in (16) is regularly varying and its tail index is determined by the most heavy-tailed distributed term in the pair \( \{N, Q\} \). The approach implicitly assumes that the underlying graph is an infinite tree, an assumption that is not plausible in real-world networks. In Chen et al. (2014), the behavior of the PR is considered on a directed configuration model, which is a tree-like graph in a sense that the first loop is observed at a distance of order \( \log n \), where \( n \) is the size of the graph. It is derived that the PR in the latter model is well approximated by the PR of the root node of a suitably constructed tree as \( n \to \infty \).

In the same way, a MLM is considered as the ‘minimal/endogeneous’ solution of the equation

\[ R = D \left( \bigvee_{j=1}^{N} A_j R_j \right) \lor Q, \]  

(Jelenkovic & Olvera-Cravioto, 2015). Assuming that all r.v.s in the triple \( \{R_j, Q, N\} \) are regularly varying and mutually independent and \( \{R_j\} \) are iid, \( PR R \) was proved to have a regularly varying tail in Jelenkovic and Olvera-Cravioto (2010), Volkovich and Litvak (2010) at the directed configuration model. A similar statement was proved in Jelenkovic and Olvera-Cravioto (2015) with regard to the MLM.

B Important results from extreme value analysis

We recall the theorems derived in Markovich (2022) that are important for the prediction of the tail and extremal indices of evolving random graphs. These theorems generalize Theorems 3 and 4 in Markovich and Rodionov (2020a). The latter state the conditions when the sequences of sums \( Y_n(z, N_n) \) and maxima \( Y^*_n(z, N_n) \) (see (3)) have the same tail and extremal indices. There are the following constrains for these statements. The slowly varying functions \( \{\ell_i(z)\} \) in (2) are uniformly upper bounded in \( i \) by a polynomial function, i.e. for all constants \( A > 1, \delta > 0 \) there exists \( x_0(A, \delta) \) such that for all \( i \geq 1 \)

\[ \ell_i(x) \leq Ax^\delta, \quad x > x_0(A, \delta) \]  

holds. Despite \( N_n \) is integer-valued, one can accept a distribution with regularly varying tail with tail index \( \alpha > 0 \) as a relevant model for \( N_n \), i.e. it holds

\[ P(N_n > x) = x^{-\alpha} \ell_n(x), \]  

\( \ell_n(x) \) is a slowly varying function. This model is motivated in several papers, see Jessen and Mikosch (2006), Robert and Segers (2008) and Volkovich and Litvak (2010) among them.

In Markovich and Rodionov (2020a) it is assumed that there is a unique "column" sequence with a minimum tail index \( k_1 < k \), \( k := \lim_{n \to \infty} \inf_{1 \leq i \leq l_n} k_i \), and \( N_n \) has a lighter tail than \( Y_{n,i} \), i.e. it holds

\[ P(N_n > l_n) = o \left( P(Y_{n,1} > u_n) \right), \quad n \to \infty, \]  

(20)
where the sequence of thresholds \( u_n \) is taken as \( u_n = yn^{1/k_1} \ell_1^d(n) \), \( y > 0 \), \( \ell(x) \) is the de Brujin conjugate of \( \ell(x) \), the sequence \( l_n \) satisfies

\[
  l_n = \lfloor nx \rfloor,
\]

and \( \chi \) satisfies

\[
  0 < \chi < \chi_0, \quad \chi_0 = \frac{k - k_1}{k_1(k + 1)}.
\]

An arbitrary dependence between "column" sequences and between \( \{ Y_{n,i} \} \) and \( \{ N_n \} \) is allowed. In Theorem 4 in Markovich (2022) recalled here in Theorem 2 the number \( d \) of "column" sequences with a minimum tail index is allowed to be random. This is realistic for random graphs since a random number of communities considering as the "column" sequences may have a minimum tail index (or a tail index close to that).

Let us recall the following conditions for a fixed \( d > 1 \) proposed in Markovich (2022).

(A1) The stationary sequences \( \{ Y_{n,i} \}_{n \geq 1}, i \in \{ 1, \ldots, d \} \) are mutually independent, and independent of the sequences \( \{ Y_{n,i} \}_{n \geq 1}, i \in \{ d + 1, \ldots, l_n \} \).

(A2) Assume \( \{ Y_{n,i} \}_{n \geq 1}, i \in \{ 1, \ldots, d \} \) satisfy the following conditions as \( x \to \infty \)

\[
  \frac{P(Y_{n,i} > x)}{\binom{x}{k_1} \ell_1(x)} \to c_i, \quad i \in \{ 1, \ldots, d \},
\]

for some non-negative numbers \( c_i \),

\[
  \frac{P(Y_{n,i} > x, Y_{n,j} > x)}{\binom{x}{k_1} \ell_1(x)} \to 0, \quad i \neq j, \quad i, j \in \{ 1, \ldots, d \}.
\]

(A3) Assume that for each \( n \geq 1 \) there exists \( i \in \{ 1, \ldots, d \} \) such that

\[
  P\left( \max_{1 \leq j \leq d, j \neq i} (z_j Y_{n,j}) > x, z_i Y_{n,i} \leq x \right) = o(P(z_i Y_{n,i} > x)), \quad x \to \infty
\]

holds.

(A4) Assume that there exists \( i \in \{ 1, \ldots, d \} \) such that it holds

\[
  P\left( \max_{1 \leq j \leq d, j \neq i} (z_j M_n^{(j)}) > u_n, z_i M_n^{(i)} \leq u_n \right) = o(1), \quad n \to \infty.
\]

Let us denote \( M_n^{(i)} = \max\{Y_{1,i}, Y_{2,i}, \ldots, Y_{l_n,i}\}, i \in \{ 1, \ldots, l_n \} \).

**Theorem 2** (Markovich, 2022) Let the sets of slowly varying functions \( \{ \ell(x) \}_{x \geq 1} \) in \( 1 \) and \( \{ \ell_i(x) \}_{x \geq 1} \) in \( 2 \) satisfy the condition \( 13 \), and \( 22 \), \( 23 \), \( 24 \) hold. Assume that \( d \) and \( \{ Y_{n,i} \} \) are independent.

(i) Let \( d \) be a bounded discrete r.v. such that \( 1 < d < d_n = \min(C, l_n), C > 1 \) holds.

(a) If (A1) or (A2) for any \( d \in \{ 2, 3, \ldots, [d_n - 1] \} \) holds and \( N_n \) and \( \{ Y_{n,i} \} \) are independent, then \( Y_n(z, N_n) \) and \( Y_n^*(z, N_n) \) have the tail index \( k_1 \). If, instead of (A1) and (A2), (A3) holds, then \( Y_n^*(z, N_n) \) has the same extremal index.

(b) If (A4) where in \( 24 \) \( d \) is replaced by \( [d_n - 1] \) holds, then \( Y_n^*(z, N_n) \) has the extremal index \( \theta \). If, in addition, (A1) (or (A2)) for any \( d \in \{ 2, 3, \ldots, [d_n - 1] \} \) holds, then \( Y_n(z, N_n) \) has the same extremal index.

(ii) Suppose that \( d > 1 \) is a bounded discrete r.v. equal to a positive integer a.s.. Then all statements of Item (i) are fulfilled.

It follows by Example 2 in Markovich (2022) that \( 23 \) is valid for all \( d \) "column" sequences such that

\[
  M_n^{(1)} \leq M_n^{(2)} \leq M_n^{(3)} \leq \ldots \leq M_n^{(d)}
\]
holds. Theorem 2 means that if there are a random number $d$ of "column" sequences with a minimum tail index, then $Y_{n,i}^*(z, N_n)$ has the extremal index $\theta_i$ of the $i$th "column" satisfying (23), $1 \leq i \leq d$. If the latter "column" sequences are independent (see, (A1)) or weakly dependent (see, (A2)), then $Y_{n}(z, N_n)$ has the same extremal index. For random networks this implies that the extremal index of the MLMs of the newly appended nodes is equal to the extremal index of the community with the minimum tail index that has a largest maximum PR among $d$ dominating communities. The extremal index of the PRs of newly appended nodes is the same if the dominating communities satisfy (A1) or (A2). Since the communities are not enumerated, their maxima can be reordered as (24). The statements of Theorem 2 are asymptotic. Thus, the approximation can be applied for sufficiently large size communities.

In case of different pair-wise dependency among elements of the $d$ "column" series with the minimum tail index, the extremal index of the maxima $Y_{n,i}^*(z, N_n)$ and sums $Y_{n}(z, N_n)$ may not exist due to a non-stationarity of these sequences.

**Remark 2** The theorems in Markovich and Rodionov (2020a), Markovich (2022) are valid if there are non-zero elements in each row corresponding to the "column" sequences $\{Y_{n,i} : n \geq 1\}$ with minimum tail index. If the most heavy-tailed column is unique and at least one element in the latter sequence is equal to zero, the sequences of the sums and maxima of the "row" elements are not stationary. This feature plays a role for graphs.

**Corollary 1** The statements of Theorems 2 remain valid if the tail indices $\{k_{n,i}\}$ of the elements in the "columns" $\{Y_{n,i} : n \geq 1\}$ are different, apart of those columns with the minimum tail index.

Corollary 1 is very important for practice. It implies that the columns with non-minimum tail indices may be non-stationary distributed and hence, their extremal indices may not exist. Its proof follows from the proofs of Theorem 3 in Markovich and Rodionov (2020a) and Theorems 3 and 4 in Markovich (2022). The columns with the minimum tail index impact on the distribution and dependence structure of the sequence of sums and maxima over rows.

### C Dependence structures in graphs

We have to investigate the dependence of PRs of two communities. One of the approaches is to consider Pearson’s correlation of two r.v.s belonging to two graphs. Each r.v. shows whether there is an edge between two nodes in a graph or not. Each edge may be sampled iid from a Bernoulli distribution with some parameter $p$ (Xiong et al., 2020). A distance correlation is an extension of Pearson’s correlation both to linear and nonlinear associations between two r.v.s or random vectors (Shen et al., 2020). It takes values in $[0, 1]$. The distance correlation equal to zero does imply independence. Since nodes can be enumerated arbitrarily, the distance correlation has to be combined with a permutation test to check the dependence hypothesis. The distance correlation is calculated first for an original pair of vectors. It is compared with those ones calculated by shuffles of these vectors.

In contrast to Shen et al. (2020), in our setting pairs of observations relating to two stationary distributed communities can be dependent and not necessarily identically distributed. One can use the distance correlation and the permutation test with regard to the row-column pairs. The $p$-value of the permutation test is the proportion of the number of the correlation measures from the samples with permuted pairs that are larger than the distance correlation that was calculated from the original data.

To measure dependencies in heavy tailed graph data using statistical inference for multivariate regular variation one can apply the polar coordinate transform to the examined random vectors $\{X_i\}$ and $\{Y_i\}$, $i = 1, \ldots, n$ (Resnick & Štúrca, 1999; Volkovich et al., 2008; Wan et al., 2020). One can estimate the empirical distribution function (edf) of the angular coordinates for the $k$ largest values of the radial coordinate. The total dependence (or total independence) corresponds to the concentration of the edf to $\pi/4$ (or, to 0 or $\pi/2$). A Starica plot can be used to find a suitable value of $k$. 
D Preferential attachment

Let us consider a network growth where each node is attached to a small seed network at a unit time. The well-known tool is a linear PA. A node $i$ can be attached randomly to existing nodes according to a probability $P_{PA}(i) = d_i / \sum_{j=1}^{N} d_j$ proportional to its degree $d_i$, or the number of its neighbors, where $N$ is the number of nodes. Nodes $i$ and $j$ may be connected with probability $d_id_j / \sum_{x=1}^{m} d_x$ (Norros & Reittu, 2006). A kind of PA with a Poisson random number of new edges to the new vertex is proposed in Norros and Reittu (2006). The PA provides the "rich-get-richer" mechanism since earlier appearing nodes may increase their numbers of edges longer. This property leads to a power-law degree distribution $P(i) \sim i^{-(1+\alpha)}$ of node degrees (Newman, 2018; Wan et al., 2020). In Wan et al. (2020) it is derived that the linear and superstar linear PA models on directed graphs lead to networks with power-law distributed in- and out-degrees.

The $\alpha-$, $\beta-$ and $\gamma-$schemes of the linear PA provide proportions of new nodes with incoming (outgoing) links to (from) existing nodes (scheme $\alpha-$ ($\gamma-$)) or the directed edges between pairs of existing nodes (scheme $\beta-$) (Samorodnitsky et al., 2016; Wang et al., 2020). Let $I_n(v)$ and $O_n(v)$ be in- and out-degree of vertex $v \in V_n$ in a graph $G_n$, $n$ and $N(n-1)$ denote the numbers of edges and nodes in $G_n$, respectively. Appending a new node $v$ to an existing graph $G_{n-1}$, one can select one of three scenarios by generating an iid sequence of trinomial r.v.s with cells marked $1, 2, 3$ with probabilities $\alpha, \beta, \gamma$. The probability to generate the edge $v \rightarrow w$ from $v$ to an existing node $w$ is given by

$$P\{\text{choose } w \in V(n-1)\} = \frac{I_{n-1}(w) + \delta_n}{n-1 + \delta_n N(n-1)}$$

(25)

by the $\alpha-$scheme; between the existing nodes $v$ and $w$

$$P\{\text{choose } (v, w)\} = \left(\frac{I_{n-1}(w) + \delta_n}{n-1 + \delta_n N(n-1)}\right) \left(\frac{O_{n-1}(w) + \delta_{out}}{n-1 + \delta_{out} N(n-1)}\right)$$

(26)

by the $\beta-$scheme; from the existing node $w$ to $v$

$$P\{\text{choose } w \in V(n-1)\} = \frac{O_{n-1}(w) + \delta_{out}}{n-1 + \delta_{out} N(n-1)}$$

(27)

by the $\gamma-$scheme, where $\delta_n$ and $\delta_{out}$ are parameters of the PA method. The latter may be estimated by the semi-parametric extreme value method based on the maximum-likelihood method (Wan et al., 2020).

E Proof of Theorem

Proof Part (i) follows by Theorem 4 in Markovich & Rodionov (2020). We start the induction with $n = 1$. The columns of matrix $A^{(1)}$ of the first iteration are obtained using submatrices of $A^{(0)}$ for different $j$ by recursions (5) and (6). By the latter theorem the $j$th columns of $\{Y_{i,j}^{(1)}\}$ and $\{X_{i,j}^{(1)}\}$ have the same tail indices $\{k_{i,j}^{(0)}\}$ and the same extremal indices $\{\theta_{i,j}^{(0)}\}$, $j \geq 1$. Getting matrices $A^{(2)}, A^{(3)}, ...$ for the next iterations both for sums and maxima similarly we obtain the same pairs of indices $\{k_{i,j}^{(0)}\}$, $\{\theta_{i,j}^{(0)}\}$ for their $j$th columns.

Part (ii), (a). The independence condition (A1) for $A^{(0)}$ is valid since communities (the "column" series of $A^{(0)}$) are nearly disconnected. The condition (A2) follows by (A1). The random number of communities $N_w$ is evidently independent on the PRs of nodes within the communities. Then the first $d_1^{(0)}$ "column" series of matrix $A^{(1)}$ have the tail index $k_{i,j}^{(0)}$ by Theorem 2. The next $d_2^{(0)}$ columns have the tail index $k_{i,j}^{(0)} > k_{i,j}^{(1)}$, etc. The columns of $A^{(3)}$ have the same tail indices as $A^{(0)}$.

Note that the "column" series $\{Y_{i,j}^{(m)}\}$ and $\{X_{i,j}^{(m)}\}$ of $A^{(m)}$, $m \geq 1$ are dependent due to
their definition as partial sums and maxima of row elements of \( A^{(m-1)} \) by the "domino" principle. Hence, we have \( Y_{1,2}^{(m)} \geq Y_{1,3}^{(m)} \geq \cdots \geq Y_{1,N}^{(m)} \) and \( X_{1,2}^{(m)} \geq X_{1,3}^{(m)} \geq \cdots \geq X_{1,N}^{(m)} \). Elements of matrices \( A^{(m)} \), \( m \geq 1 \) may be represented by elements of \( A^{(0)} \). Really, for \( m = 2 \) we have

\[
Y_{n,i}^{(2)} = \sum_{j=1}^{N_n} Y_{n,j}^{(1)} = \sum_{j=1}^{N_n} j Y_{n,j}^{(0)}, \quad i \geq 1,
\]

and similarly for \( X^{(2)} \). Considering weights \( z_j = j \) in (28) as in (1), we obtain by Theorem 2 that the first \( d^{(0)} \) sequences \( \{Y_{n,j}^{(2)}\} \) and \( \{X_{n,j}^{(2)}\} \) have the tail index \( k_{d_1}^{(0)} \), the next \( d_2^{(0)} \) ones - \( k_{d_2}^{(0)} \), etc. in the same way as "column" series of matrix \( A^{(1)} \). The same is valid for \( A^{(m)} \) with \( m \geq 2 \) by induction.

Part (ii), (b). In order to find the extremal indices of the "column" series of \( A^{(1)} \), let us enumerate the \( d^{(0)} \) columns (the communities) in an descending order of the PR maxima of \( A^{(0)} \) over columns, i.e. \( M^{(1)}_1 \geq M^{(1)}_2 \geq \cdots \geq M^{(1)}_{d_0} \) for each fixed value of r.v. \( d^{(0)} \in \{2, 3, \ldots, |d_n - 1| \} \) and \( j \in \{1, 2, \ldots, \} \). Then (A4) is fulfilled consistently for \( i \in \{1, 2, \ldots, d^{(0)} - 1\} \)

By Theorem 2 the first \( d^{(0)} \) \( j \geq 1 \) "column" series \( \{X_{i,j}^{(1)}\}_{i \geq 1} \) of \( A^{(1)} \) have the extremal indices \( \theta^{(0)}_{d^{(0)} - 1 + 1} \), \( \theta^{(0)}_{d^{(0)} - 1 + 2} \), \ldots, \( \theta^{(0)}_{d^{(0)} - 1 + d^{(0)}} \) for any values of \( d^{(0)} \in \{2, 3, \ldots, |d_n - 1| \} \), \( d_0^{(0)} = 0 \). Since, in addition, (A1) (or (A2)) for \( A^{(0)} \) and any \( d^{(0)} \in \{2, 3, \ldots, |d_n - 1| \} \) holds, then \( \{Y_{i,j}^{(1)}\}_{i \geq 1} \) have the same extremal indices. Since elements of \( A^{(m)} \) may be represented as weighted sums or maxima of elements of \( A^{(0)} \), the extremal indices of the "column" series of \( A^{(m)} \), \( m \geq 1 \) are the same as ones of \( A^{(0)} \).

References

1. Asmussen, S., Foss, S. (2018). Regular variation in a fixed-point problem for single-and multi-class balking processes and queues. Branching Processes and Applied Probability. Papers in Honour of Peter Jagers. Adv. Appl. Prob., 50A, 47-61. DOI:10.1017/apr.2018.69
2. Bagrow, J. P., Brockmann, D. (2013). Natural Emergence of Clusters and Bursts in Network Evolution. Physical Review X, 021016. DOI: 10.1103/PhysRevX.3.021016
3. Banerjee, S., Olera-Cravito, M. (2021). Pagerank asymptotics on directed preferential attachment networks. arXiv:2102.08894v1
4. Beirlant, J., Goekeler, Y., Teugels, J. & Segers, J. (2004). Statistics of Extremes: Theory and Applications. Chichester, West Sussex: Wiley.
5. Bollobás, B., Riordan, O. (2006). Percolation. Cambridge University Press, Cambridge.
6. Censor-Hillel K., Shachnai H. (2010). Partial Information Spreading with Application to Distributed Maximum Coverage. In Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing (PODC ’10) (pp. 161-170) ACM. New York, USA. https://doi.org/10.1145/1835698.1835739
7. Chen, N., Litvak, N., Olera-Cravito, M. (2014a). Ranking Algorithms on Directed Configuration Networks. arXiv:1409.7443v2
8. Chen, N., Litvak, N., Olera-Cravito, M. (2014b). PageRank in Scale-Free Random Graphs. WAW 2014, LNCS 8882, ed. A. Bonato et al. (pp. 120-131). Switzerland: Springer. DOI: 10.1007/978-3-319-13123-8
9. Clauset, A., Newman, M. E. J., Moore, C. (2004). Finding community structure in very large networks. Physical Review E 70(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111
10. Clauset, A., Shalizi, K.R., Newman, M. E. J. (2009). Power-law distributions in empirical data. SIAM Rev., 51 (4), 661-703. https://doi.org/10.1137/070710111
11. Coscia, M., Giannotti, F., Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4 (5), 512-546. https://doi.org/10.1002/sam.10133
12. da Cruz, J.P., Lind, P.G. (2013). The bounds of heavy-tailed return distributions in evolving complex networks. *Physics Letters A*, 377, 189-194.
13. Drees, H., Jansen, A., Resnick, S. I., Wang, T. (2020). On a minimum distance procedure for threshold selection in tail analysis. *SIAM J. Math. Data Sci.*, 2(1), 75-102. DOI: 10.1137/19M1260463
14. Duguet N., Perez A. (2015). Directed Louvain: maximizing modularity in directed networks. [Research Report] Université d’Orléans. hal-01231784.
15. Ferro, C.A.T., Segers, J. (2003). Inference for Clusters of Extreme Values. *J. R. Statist. Soc. B.*, 65, 545-556.
16. Ferreira, M. (2018). Heuristic tools for the estimation of the extremal index: a comparison of methods. *REVSTAT - Statistical Journal*, 16(1), 115-136.
17. Fortunato S., Boguna M., Flammini A. & Menczer F. (2011). On Local Estimations of PageRank: A Mean Field Approach. *Internet Mathematics*, 4(2-3), 245-266.
18. Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3), 75–174.
19. Fukutome, S., Liniger, M.A. & Süveges, M. (2015). Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland. *Theor. Appl. Climatol.*, 120, 403-416. DOI: https://doi.org/10.1007/s00704-014-1180-5
20. Jelenkovic, P. R., Olvera-Cravioto, M. (2010). Information ranking and power laws on trees. *Adv. Appl. Prob.*, 42(4), 1057-1093. DOI:10.1239/aap/1293113151
21. Jelenkovic, P. R., Olvera-Cravioto, M. (2015). Maximums on trees. *Stoch. Process. Appl.*, 125, 217-232. DOI: 10.1016/j.spa.2014.09.004
22. Jessen, A. H., Mikosch, T. (2006). Regularly varying functions. *Publ. Inst. Math. (Beograd) (N.S.)*, 80, 171-192. DOI:10.2298/PIM0694171H
23. Garavaglia, A., van der Hofstad, R. & Litvak, N. (2020). Local weak convergence for PageRank. *Ann. Appl. Prob.*, 30(1), 40-79. https://doi.org/10.1214/19-AAP1494
24. Gissibl, N., Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A), 2693-2720.
25. Ghoshal, G., Chi, L., Barabási, A.L. (2013). Uncovering the role of elementary processes in network evolution. *Scientific Reports*, 3, 29, 2020.
26. Holme P., Litvak N. (2017). Cost-efficient vaccination protocols for network epidemiology. *PLoS Comput Biol.*, 13(9) https://doi.org/10.1371/journal.pcbi.1005696
27. Krapivsky P. L., Redner S. (2001). Organization of growing random networks. *Phys. Rev., E 63*, 066123. DOI: 10.1103/PhysRevE.63.066123
28. Langville, A. N., Meyer, C. D. (2006). *Google’s PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press.
29. Leadbetter, M.R., Lingren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequence and Processes*. ch.3, New York: Springer.
30. Lebedev, A. V. (2015). Activity maxima in some models of information networks with random weights and heavy tails. *Problems of Information Transmission*, 51(1), 66-74.
31. Leskovec J., Lang K., Dasgupta A., Mahoney M. (2009). Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics*, 6 (1), 29-123. DOI: 10.1080/15427951.2009.10129177
32. Litvak, N., Scheinhardt, W. R. W. & Volkovich, Y. (2007). In-Degree and PageRank: Why Do They Follow Similar Power Laws? *Internet Mathematics*, 4(2-3), 175-198. DOI: 10.1080/15427951.2007.10129293
33. Markovich, N.M. (2007). *Nonparametric Analysis of Univariate Heavy–Tailed data: Research and Practice*. Chichester, West Sussex: Wiley.
34. Markovich, N.M. (2017). Clustering and Hitting Times of Threshold Exceedances and Applications. *Int. J. Data Analysis Techniques and Strategies*, 9(4), 331-347. DOI: 10.1504/IJDATS.2017.088360
35. Markovich, N.M., Rodionov I.V. (2020a). Maxima and sums of non-stationary random length sequences. *Extremes*, 23(3), 451-464. DOI: 10.1007/s10687-020-00372-5
36. Markovich, N.M., Rodionov I.V. (2020b). Threshold selection for extremal index estimation. arXiv:2009.02918
37. Markovich, N.M., Ryzhov, M. & Krieger, U.R. (2017). Nonparametric analysis of extremes on web graphs: pagerank versus max-linear model. *Communications in Computer and Information Science*, 700, 13-26.

38. Markovich, N.M. (2021). Extremes of Sums and Maxima with Application to Random Networks. In Proceedings 5th International Conference on Stochastic Methods 2020 ICSM5 November 23-27, 2020 Moscow, Russia, pp. 107-120 arXiv:2110.04120

39. Markovich, N.M., Ryzhov, M. & Vaičiulis, M. (2022). Tail Index Estimation of PageRanks in Evolving Random Graphs. *Mathematics*, 10(16), 3026.

40. Markovich, N.M. (2022). Weighted maxima and sums of non-stationary random length sequences in heavy-tailed models. arXiv: 2209.08485v [math.ST] 18 Sep 2022.

41. McCormick, D. A., Contreras, D. (2001). On the Cellular and Network Bases of Epileptic Seizures. *Annu. Rev. Physiol.*, 63, 815. DOI: 10.1146/annurev.physiol.63.1.815

42. McElroy, T., Politis, D.N. (2007). Moment-Based Tail Index Estimation. *J. Statist. Plan. Infer.*, 137, 1389-1406. DOI: 10.1016/j.jspi.2006.04.002

43. Mester, A., Pop, A., Mursa, B.-E.-M., Grebla, H., Diosan, L., Chira, C. (2021). Network Analysis Based on Important Node Selection and Community Detection. *Mathematics*, 9, 2294.

44. Mosh-Aoyama D., Shah D. (2006). Computing separable functions via gossip. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing (PODC ’06).* (pp. 113-122). ACM. New York, USA.

45. Newman M. E. J. (2018). *Networks: An Introduction*. (2nd ed.). Oxford: Oxford University Press.

46. Norros I., Reittu H. (2006). On a conditionally poissonian graph process. *Adv. Appl. Prob.* (SGSA), 38, 59-75. DOI: 10.1239/aap/1143906140

47. Novak, S. Y. (2002). Inference of heavy tails from dependent data. *Siberian Advances in Mathematics*, 12(2), 73-96. https://doi.org/10.1016/C2015-0-01492-7

48. Olvera-Cravioto, M. (2012). Asymptotics for weighted random sums. *Adv. Appl. Prob.*, 44(4), 1142-1172. DOI:10.1239/aap/1354716592

49. Resnick, S. I., Stáríča, C. (1999). Smoothing the Moment Estimate of the Extreme Value Parameter. *Extremes*, 1(3), 263-294. https://doi.org/10.1023/A:1009925716617

50. Robert, C.Y. (2008). Inference for the limiting cluster size distribution of extreme values. *Ann. Statist.*, 37, 271-310. DOI: 1214/07-AOS501

51. Robert, C.Y., Segers, J. (2008). Tails of random sums of a heavy-tailed number of light-tailed terms. *Insurance: Mathematics and Economics*, 43, 85-92. https://doi.org/10.1016/j.insmatheco.2007.10.001

52. Rootzén, H. (1988). Maxima and exceedances of stationary Markov chains. *Adv. Appl. Prob.*, 20, 371-390. DOI:10.2307/1427395

53. Pandurangan, G., Raghavan, P. & Upfal, E. (2002). Using PageRank to Characterize Web Structure. In *Ibarra O.H., Zhang L. (eds) Computing and Combinatorics. COCOON 2002. LNCS 2387.* (pp. 330-339). Springer, Berlin, Heidelberg.

54. Samorodnitsky, G., Resnick, S., Towsley, D., Davis, R., Willis, A. & Wan, P. (2016). Nonstandard regular variation of in-degree and out-degree in the preferential attachment model. *J. Appl. Prob.*, 53(1), 146-161. DOI: https://doi.org/10.1017/jpr.2015.15

55. Shen, C., Priebe, C.E. & Vogelstein, J.T. (2020). From Distance Correlation to Multiscale Graph Correlation. *Journal of the American Statistical Association*, 115:529, 280-291. https://doi.org/10.1080/01621459.2018.1543125

56. Schroeder, D.T., Langguth, J., Burchard, L., Pogorelov K. & Lind, P.G. (2022). The connectivity network underlying the German’s Twittersphere: a testbed for investigating information spreading phenomena. *Scientific Reports*, 12, 4085. https://doi.org/10.1038/s41598-022-07961-3

57. Søveges, M., Davison, A.C. (2010). Model misspecification in peaks over threshold analysis. *Ann. Appl. Statist.*, 4(1), 203-221. DOI: 10.1214/09-AOAS292

58. Tillier, C., Wintenberger, O. (2018). Regular variation of a random length sequence of random variables and application to risk assessment. *Extremes*, 21, 27-56. https://doi.org/10.1007/s10687-017-0297-1

59. Volkovich, Y. V., Litvak, N. (2010). Asymptotic analysis for personalized web search. *Adv. Appl. Prob.*, 42(2), 577-604. DOI: https://doi.org/10.1239/aap/1275055243
60. Volkovich Y., Litvak N., Zwart B. (2008). Measuring extremal dependencies in Web graphs. *WWW ’08: Proceedings of the 17th international conference on World Wide Web*. April (pp. 1113-1114). https://doi.org/10.1145/1367497.1367682

61. Xiong, J., Shen, C., Arroyo, J. & Vogelstein, J. (2020). Graph Independence Testing. *arXiv: 1906.03661*

62. Wan, P., Wang, T., Davis, R. A. & Resnick, S.I. (2020). Are extreme value estimation methods useful for network data? *Extremes*, 23, 171-195. DOI: https://doi.org/10.1007/s10687-019-00359-x

63. Wang, T., Resnick, S.I. (2019). Consistency of Hill estimators in a linear preferential attachment model. *Extremes*, 22, 1-28. DOI: 10.1007/s10687-018-0335-7.

64. Wang, T., Resnick, S.I. (2020). Degree growth rates and index estimation in a directed preferential attachment model. *Stochastic Processes and their Applications*, 130(2), 878-906. DOI: https://doi.org/10.1016/j.spa.2019.03.021

**Acknowledgements** The author was supported by the Russian Science Foundation (grant No. 22-21-00177). The author would like to thank anonymous reviewers for useful comments.