Evsey V. Morozov
Alexander S. Rumyantsev
Oleg V. Lukashenko (Eds.)

Proceedings
of the Third International
Workshop **SMARTY'22**

SMARTY 22
Stochastic Modeling & Applied Research of TechnologY
Third International Workshop
Petrozavodsk, Karelia
August 21-25, 2022

Markovich, N.

Extremal properties of maxima and sums in evolving networks

Stochastic Modeling and Applied Research of Technology, Vol. 3, Pp. 30-35.
DOI: 10.57753/SMARTY.2023.82.79.004

Citation: Markovich, N. (2023). Extremal properties of maxima and sums in evolving networks. Stochastic Modeling and Applied Research of Technology, 3, 30-35. https://doi.org/10.57753/SMARTY.2023.82.79.004

Extremal properties of maxima and sums in evolving networks *

Natalia Markovich

V.A. Trapeznikov Institute of Control Sciences Russian Academy of Sciences,
Profsoyuznaya 65, Moscow 117997, Russia,
nat.markovich@gmail.com

Abstract. Real-world networks display a strong heterogeneity that is reflected in a heavy-tailed distribution of node influence indices. The PageRank and the Max-Linear Model may be used as node influence indices of random graphs. The present paper aims to summarize shortly some recent author's results with regard to extremal properties of maxima and sums of non-stationary random length sequences and their application to evolving networks. Under the extremal properties we understand the tail and extremal indices. The evolution of the random network by the preferential attachment is considered since it allows us to model heavy-tailed distributed node influence indices.

Keywords: Tail index · Extremal index · Sums · Maxima · PageRank · Max-Linear Model · Network evolution

1 Introduction

Real-world networks display a strong heterogeneity that is reflected in a heavy-tailed distribution of node influence indices. The in-degree, the PageRank (PR) and the Max-Linear Model (MLM) may be used as node influence indices of random graphs. The PRs and MLMs are proved to be regularly varying heavy-tailed distributed random variables (r.v.s) [4,5,14].

The present paper aims to summarize shortly some recent author's results with regard to extremal properties of maxima and sums of non-stationary random length sequences and their application to evolving networks. Under the extremal properties we understand the tail and extremal indices. The tail index (TI) and extremal index (EI) of the PRs and MLMs of newly appended nodes in evolving random networks are proposed to be predicted in [10]. The TI shows a heaviness of the distribution tail. The EI reflects a cluster structure of the stochastic process or its local dependence. It is known that the reciprocal of the EI approximates a mean cluster size of the stochastic process, i.e. the mean number of exceedances over a threshold per cluster [6]. The cluster of exceedances of the random sequences may imply a block of data with at least

* The reported study was funded by the Russian Science Foundation RSF, project number 22-21-00177.

one extreme observation (an exceedance) over a sufficiently high threshold [1], or a set of consecutive exceedances of the process over a sufficiently high threshold [3,12]. Clusters can also be defined as data blocks separated by some number of observations running under the high threshold [1].

The clustering structure may arise in random networks (and in the random graphs modeling them) and it can be increase rapidly over time. In fact, the clusters are built around giant nodes with a large number of links to the rest of the nodes. A network evolution means that a graph grows by adding new edges between existing nodes and newly appended nodes or between existing nodes only at discrete time steps. At each such a step, a new node may be either added or not. The evolution is often modeled by preferential attachment (PA) tools (see [2,13,15] among others) to explain power-law degree distributions observed in real-world networks [2]. Then a newly appended node prefers to attach to existing nodes with large node degrees (in the simplest case to one of the existing nodes). To model the evolution of directed random graphs, the $\alpha-$, $\beta-$ and $\gamma-$ PA schemes are proposed in [15] since they may create graphs with multiple edges between nodes and self-loops. The survey of results concerning the evolution of random networks and related extreme value statistics can be found in [11]. The evolution with a deletion of existing nodes or edges at each step is not enough studied yet, and this gap is partially filled by the simulation study in [8].

The aim of the paper [10] was to find the TI and EI of the PR and MLM in the PA-evolved random graphs those are dynamically changing in time. To this end, results of extreme value theory obtained in [7,9] are applied to random networks. Namely, the TI and EI of sums and maxima of weighted non-stationary random length sequences of regularly varying r.v.s are derived in [7,9]. In [7], conditions were found such that the sums and maxima have the same TI and EI. The latter indices are equal to ones of the most heavy-tailed term in the sum or maximum if such term is unique.

If there is a random number d of the most heavy-tailed dependent terms, then an additional condition regarding the mutual dependence between terms is required to make a conclusion about the TIs and the EIs of the sums and maxima [9]. Particularly, if the most heavy-tailed terms with the minimum TI are independent, then the sums and maxima have the same minimum TI. Their EI exists and it can be simply calculated as a linear combination of the EIs of the most heavy-tailed terms. The independence condition can be weakened assuming that the tail function of the sum of such terms is asymptotically equivalent to the tail of one of such terms. The latter assumption allows us to turn back to the case of the unique term with the minimum TI. To determine the EI of the sums and maxima in the case of the d most heavy-tailed terms we need to assume the asymptotic equivalence of the maxima over d dependent "columns" and the maxima over one of these "columns" [9]. Under the "column" sequence we understand the observations of a random term.

The contribution of the application paper [10] is as follows. Starting with a set of weakly connected stationary seed communities as a hot spot and ranking

them with regard to the TIs of their node PRs, the TI and EI of the PRs and MLMs of new nodes appended to the network may be determined by the most heavy-tailed community. This procedure allows us to predict a temporal network evolution in terms of the TI and EI. The exposition in [10] is provided by algorithms, examples and a study of simulated and real evolved random graphs.

In the next section we give a more detailed description of main results.

2 Main results

In [7], 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, is considered. The length N_n of "row" sequences $\{Y_{n,i} : i \geq 1\}$ for each n is generally random and $\{N_n : n \geq 1\}$ 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 EI θ_i and having a regularly varying distribution tail $P\{Y_{n,i} > x\} = \ell_i(x)x^{-k_i}$, where $k_i > 0$ is the TI and $\ell_i(x)$ is a slowly varying function. There are no assumptions on the dependence structure in i. Assuming that there is a unique "column" sequence with a minimum TI k_1, and N_n has a lighter tail than $Y_{n,i}$, it was found in [7] that the TI and EIs of the weighted sums and maxima

$$\begin{aligned} Y_n^*(z, N_n) &= \max(z_1 Y_{n,1}, ..., z_{N_n} Y_{n,N_n}), \\ Y_n(z, N_n) &= z_1 Y_{n,1} + ... + z_{N_n} Y_{n,N_n}, \end{aligned} \tag{1}$$

with positive constants $z_1, z_2, ...,$ are equal to k_1 and θ_1, respectively. If there is a random number d of such "column" sequences with TI k_1, then the sums and maxima may also have the same TI k_1 and the same EI, but it requires additional dependence conditions on the latter "column" sequences [9].

These results are applied in [10] to random graphs. Let us explain the main idea of this application. A seed graph from which the evolution starts is divided into communities. The PRs of nodes in the communities are calculated. The communities are ranked by their TI estimates assuming that the PRs of each community are stationary distributed with a regularly varying tail. The communities are considered as the "column" sequences, where N_n is the random number of communities. We call the community as "dominating" one if its TI is minimal and thus, the distribution of its PRs is the most heavy-tailed.

Let a set of new nodes be appended within a fixed time interval of the evolution. A new node can be considered as a root of the tree. Then N_n is at the same time an in-degree of the root node n that is the number of its nearest neighbors with incoming links to the root. The new nodes are divided into classes. If a new node has at least one link to the community with the minimum TI k_1, then the node is related to Class 1 with the TI k_1 and the EI θ_1 if such community is unique. Indeed, the TIs may only be estimated and the TI estimates are likely different. Hence, we may assume that there is such unique community. Those nodes which have no links to the "dominating" community may have links to

the second "dominating" one that has the second minimum TI $k_2 > k_1$, and we say that such nodes are related to Class 2, etc. Theorem 1 proved in [10] states that the TI and EI of the newly appended nodes are determined by the TI and EI of the communities of the seed graph. If nevertheless the TIs estimates of the "dominating" communities are close, then one has to investigate the mutual dependence between these communities as it follows from [9].

Let us explain more precisely how (1) may relate to the PRs and the MLMs. The PR R of a randomly chosen Web page (a node in the Web graph) is viewed as a r.v.. This R has been considered in [4,14] as the solution to the fixed-point problem

$$R =^D \sum_{j=1}^{N} A_j R_j + Q, \tag{2}$$

where $=^D$ denotes equality in distribution. In the same way, a MLM is considered as the 'minimal/endogeneous' solution of the equation [5]

$$R =^D \left(\bigvee_{j=1}^{N} A_j R_j \right) \vee Q. \tag{3}$$

One can rewrite the right-hand sides of (2) and (3) as, respectively,

$$Y_i(c, N_i) = c \sum_{j=1}^{N_i} Y_{i,j} + Q_i, \qquad Y_i^*(c, N_i) = c \bigvee_{j=1}^{N_i} Y_{i,j} \vee Q_i, \quad i \in \{1, ..., n\}.$$

$Y_i(c, N_i)$ relates to the definition of Google's PR with a damping factor $c > 0$ and Q is a personalization value of the node [14]. The following recursions

$$Y_{i,j}^{(m)} = c \sum_{s=j}^{N_i} Y_{i,s}^{(m-1)} + Q_i, \tag{4}$$

$$X_{i,j}^{(m)} = \left(c \bigvee_{s=j}^{N_i} X_{i,s}^{(m-1)} \right) \vee Q_i, \quad \{X_{i,j}^{(0)}\} \equiv \{Y_{i,j}^{(0)}\}, \tag{5}$$

where $m, i, j \geq 1$, m is connected with the time, were considered in [10].

In [10], matrices related to the scheme of series $\{Y_{n,i}^{(0)} : n, i \geq 1\}$ and the corresponding TI and EI $(k_i^{(0)}, \theta_i^{(0)})$ were considered:

$$A^{(0)} = \begin{pmatrix} Y_{1,1}^{(0)} & Y_{1,2}^{(0)} & Y_{1,3}^{(0)} & ... & 0 & Q_1 \\ Y_{2,1}^{(0)} & 0 & Y_{2,3}^{(0)} & ... & Y_{2,N_2}^{(0)} & Q_2 \\ ... & ... & ... & ... & ... & ... \\ Y_{n,1}^{(0)} & Y_{n,2}^{(0)} & Y_{n,3}^{(0)} & ... & Y_{n,N_n}^{(0)} & Q_n \end{pmatrix}, \tag{6}$$

$$\begin{pmatrix} k_1^{(0)} & k_2^{(0)} & k_3^{(0)} & ... & k_N^{(0)} & k_{N+1}^{(0)} \\ \theta_1^{(0)} & \theta_2^{(0)} & \theta_3^{(0)} & ... & \theta_N^{(0)} & 1 \end{pmatrix}.$$

Here, $\{Q_i\}$ is a sequence of independent identically distributed r.v.s. Due to independence, its EI is equal to 1.

The network communities may be interpreted as the columns of matrix $A^{(0)}$. A zero jth element in the ith row $Y_{i,j}^{(0)}$ of $A^{(0)}$ means that the ith root node has no followers in the jth 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. If books are cited by papers from the "dominating" community, then the TI and EI of PRs of the books are determined by ones of the latter papers.

The matrix $A^{(0)}$ corresponds to an initial (un)directed graph (a seed network) that is used to start an evolution of the graph in time by means of some attachment tool. The jth column $\{Y_{i,j}^{(m)}\}_{i\geq 1}$ (or $\{X_{i,j}^{(m)}\}_{i\geq 1}$) of the matrix $A^{(m)}$ is defined by (4) (or (5)) 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 $A^{(m-1)}$. The evolution looks like the "domino principle" [10]. The principle implies that the first column of $A^{(m)}$, $m \geq 1$, calculated as sums (or maxima) over the rows of $A^{(m-1)}$, has the minimum TI among all columns of $A^{(m-1)}$. The second column of $A^{(m)}$ is calculated by the same row elements apart of the elements related to the first column of $A^{(m-1)}$. Hence, the TI of the second column is equal to the second minimum among TIs of $A^{(m-1)}$, etc. The "domino principle" is generalized in [10] to the case when random numbers of columns have the minimum, the second minimum of the TI, etc, by [9]. It is important that for each row at least one element corresponding to the "column" sequences with the minimum TI has to be non-zero. Otherwise, the sums and maxima over rows may be non-stationary distributed with different TIs. It is proposed in [10] to permute the rows of $A^{(0)}$ to have blocks of rows with non-zero elements at least in one of the most heavy-tailed distributed column in the block.

It is derived in [10] that $\{Y_{i,j}^{(m)}\}_{i\geq 1}$ and $\{X_{i,j}^{(m)}\}_{i\geq 1}$ calculated by (4) and (5) have the same TI $k_j^{(0)} < k$, where $k := \lim_{n\to\infty} \inf_{j+d_j\leq i\leq l_n} k_i^{(0)}$ and the same EI $\theta_j^{(0)}$ if $d_j = 1$ for any $1 \leq j \leq l_n - 1$ for any $m \geq 1$.

3 Discussion and open problems

Since the nodes of the graph cannot be definitely enumerated, the definition and testing of the stationarity in the graphs or their communities remain an open problem. 'One can determine that a graph is stationary if for all finite sets of vertices with the same adjacency matrices the joint distributions of their in- and out-degrees are the same' [8]. The dependence detection between communities constitutes another open problem. Its solution can be based on the detection of the dependence of vectors, for instance, by the distance correlation, that in case

of graphs requires a proposal of a permutation test. Testing of stationarity and dependence of random graphs is discussed in [11].

The TIs and EIs of the PRs and MLMs of the sequence of the root nodes in a graph as well as the classification of newly appended nodes during evolution by their TIs constitute novelty, see [10]. The latter results serve as a motivation of the theoretical achievements of papers [7] and [9] that relate to the TIs and EIs of non-stationary random length sequences of r.v.s.

References

1. Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J.: Statistics of Extremes: Theory and Applications. Chichester, West Sussex: Wiley (2004)
2. Bollobás, B., Borgs, C., Chayes, J., Riordan, O.: Directed Scale-Free Graphs. Society for Industrial and Applied Mathematics, USA, SODA'03, 132–139 (2003).
3. Ferro, C.A.T., Segers, J.: Inference for Clusters of Extreme Values. J. R. Statist. Soc. B. **65**, 545–556 (2003).
4. Jelenkovic, P. R., Olvera-Cravioto, M.: Information ranking and power laws on trees. Adv. Appl. Prob. **42(4)**, 1057–1093 (2010).
5. Jelenkovic, P. R., Olvera-Cravioto, M.: Maximums on trees. Stoch. Process. Appl. **125**, 217–232 (2015).
6. Leadbetter, M.R., Lingren, G., Rootzén, H.: Extremes and Related Properties of Random Sequence and Processes. Ch.3, Springer: New York (1983)
7. Markovich, N.M., Rodionov, I.V.: Maxima and sums of non-stationary random length sequences. Extremes **23(3)**, 451–464 (2020).
8. Markovich, N.M., Ryzhov, M.S., Vaičiulis, M.: Tail Index Estimation of PageRanks in Evolving Random Graphs. Mathematics **10(16)**, 3026 (2022).
9. Markovich, N.M.: Weighted maxima and sums of non-stationary random length sequences in heavy-tailed models. Submitted paper (2023).
10. Markovich, N.M.: Extremal properties of evolving networks: local dependence and heavy tails. Annals of Operation Research **4**, 1–32 (2023). `https://doi.org/10.1007/s10479-023-05175-y` arXiv:2211.13574 [math.ST] 24 Nov 2022.
11. Markovich, N.M., Vaičiulis, M.: Extreme Value Statistics for Evolving Random Networks. Mathematics **11(9)**, 2171 (2023).
12. Markovich, N.M., Rodionov, I.V.: Threshold selection for extremal index estimation. Journal of Nonparametric Statistics [To appear] `https://doi.org/10.1080/10485252.2023.2266050` arXiv:2009.02318v1
13. Norros, I., Reittu, H.: On a conditionally poissonian graph process. Adv. Appl. Prob. (SGSA) **38**, 59–75 (2006).
14. Volkovich, Y., Litvak, N.: Asymptotic analysis for personalized web search. Adv. Appl. Prob. **42(2)**, 577–604 (2010).
15. Wan, P., Wang, T., Davis, R. A., Resnick, S.I.: Are extreme value estimation methods useful for network data? Extremes **23**, 171–195 (2020).