On distribution and dependence of extremes in PageRank-type processes

Konstantin Avrachenkov*, Natalia M. Markovich**, Jithin K. Sreedharan*

* INRIA, Sophia Antipolis, France
k.avratchenkov@inria.fr,
jithin.sreedharan@inria.fr
**Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
markovic@ipu.rssi.ru

1. Introduction

Social networks contain clusters of nodes centered at high-degree nodes and surrounded by low-degree nodes. Such a cluster structure of the networks is caused by the dependence (social relationships and interests) between nodes and possibly by heavy-tailed distributions of the node degrees. We consider the degree sequences generated by PageRank type sampling processes and investigate clusters of exceedances of such sequences over large degrees. It is important to investigate its stochastic nature since it allows us to disseminate advertisement or collect opinions more effectively within the clusters.

The dependence structure of sampled degrees exceeding sufficiently high thresholds is measured using a parameter called extremal index, θ. It is defined as follows.

Definition 1 *[2] The stationary sequence $\{X_n\}_{n\geq 1}$, with $M_n = \max\{X_1, \ldots, X_n\}$ and F as the marginal distribution function, is said to have the extremal index $\theta \in [0,1]$ if for each $0 < \tau < \infty$ there is a sequence of real numbers (thresholds) $u_n = u_n(\tau)$ such that*

$$\begin{aligned} \lim_{n\to\infty} n(1-F(u_n)) &= \tau \text{ and} \\ \lim_{n\to\infty} P\{M_n \leq u_n\} &= e^{-\tau\theta} \end{aligned} \tag{1}$$

hold.

Some of the interpretations of θ are [2] :

- Relation to the mean cluster size : A cluster is considered to be formed by the exceedances in a block of size r_n ($r_n = o(n)$) in n with cluster size $\xi = \sum_{i=1}^{r_n} 1(X_i > u_n)$ when there is at least one exceedance within r_n. The point process of exceedances which counts the number of exceedances, normalized by n, of $\{X_i\}_{i=1}^n$ over threshold $\{u_n\}$ converge weakly to a compound Poisson process (CP) with rate $\theta\tau$ under condition (1) and a mixing condition, and the points of exceedances in CP correspond to the clusters . Then $\theta = (E\xi)^{-1}$.
- We also have $P\{M_n \leq x\} = F^{n\theta}(x) + o(1), n \to \infty$. Hence θ allows us to evaluate a limit distribution of the maximum of the node degree and it is also helpful to find quantiles of the maxima. These quantiles show the large degrees in the network which arise with a certain probability.

The main contributions in this work are as follows. We study the extremal and clustering properties due to degree correlations in large graphs. Since the network under consideration is known only through Application Programming Interfaces (API), different graph exploring or sampling algorithms are employed to get the samples. The considered algorithms are Random Walk based that are widely discussed in the literature (see [1] and the references therein). In order to facilitate a painless future study of correlations and clusters of degrees in large networks, we propose to abstract the cluster statistics to a single and handy parameter, θ. We derive analytical expressions of θ for different sampling techniques. Finally different estimators of θ are tried out and several other applications are proposed.

2. Model and Algorithms

We consider networks represented by an undirected graph G with N vertices and M edges. In accordance with the data from most of the real networks, we assume the network is not known completely (with N and M unknown) and also assume correlation in degrees between neighbor nodes. The dependence structure in the graph is described by the joint degree-degree probability density function $f(d_1, d_2)$ (see e.g., [3]). The probability that a randomly chosen edge has the end vertices with degrees $d_1 \leq d \leq d_1 + \Delta(d_1)$ and $d_2 \leq d \leq d_2 + \Delta(d_2)$ is $(2 - \delta_{d_1 d_2}) f(d_1, d_2)\Delta(d_1)\Delta(d_2)$. Here $\delta_{d_1 d_2} = 1$ if $d_1 = d_2$, zero otherwise. The degree distribution $f_d(d_1)$ can be calculated from the marginal of $f(d_1, d_2)$ as

$$f(d_1) = \sum_{d_2} f(d_1, d_2) = \frac{d_1}{E[D]} f_d(d_1),$$

where E[D] denotes the mean node degree. Then $E[D] = \left[\int\int (f(d_1, d_2)/d_1)\, d(d_1) d(d_2)\right]^{-1}$. Most of the results in this paper are derived assuming continuous probability distributions for $f(d_1, d_2)$ and $f_d(d_1)$ due to the ease in calculating θ for continuous case. In particular, for analytical tractability and with the sup-

port of empirical evidences, we assume the bivariate Pareto model for the joint degree-degree tail function.

2.1. Description of random walks

The different graph exploration algorithms considered in the paper are Random Walk based and transition kernels are defined for degree state space unlike in previous works where they were defined and well studied for vertex set (see [1] and the references therein). We use $f_{\mathscr{X}}$ to represent the probability density function under the algorithm $\mathscr{X}$.

2.1.1. Standard random walk (RW)

In a standard random walk, the next node to visit is chosen uniformly among the neighbours of the current node. Using "mean field" arguments, the joint density function of the standard random walk is derived as $f_{RW}(d_{t+1}, d_t) = f(d_{t+1}, d_t)$.

2.1.2. PageRank (PR)

PageRank is a modification of the random walk which with a fixed probability $1 - c$ samples random node with uniform distribution and with a probability c, it follows the standard Random walk transition. Its evolution can be described as

$$f_{PR}(d_{t+1}|d_t) = c f_{RW}(d_{t+1}|d_t) + (1-c) f_d(d_{t+1}).$$

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

2.1.3. Random walk with jumps (RWJ)

This algorithm follows a standard Random walk edge with probability $d_t/(d_t+\alpha)$ and jumps to an arbitrary node uniformly with probability $\alpha/(d_t+\alpha)$, where d_t is the degree of current node and $\alpha \in [0, \infty]$ is a design parameter ([1]). This modification makes the underlying Markov Chain time reversible, significantly reduces mixing time, improves estimation error and leads to a closed form expression for stationary distribution. The joint density for the random walk with jumps is derived as

$$f_{JP}(d_{t+1}, d_t) = \frac{E[D] f(d_{t+1}, d_t) + \alpha f_d(d_{t+1}) f_d(d_t)}{E[D] + \alpha}.$$

2.2. Calculation of θ

The extremal index of the random walk based sampling algorithms can be calculated by means of copula density as

$$\theta = \lim_{x \to 1} \frac{x - C(x,x)}{1-x} = C'(1,1) - 1,$$

where $C(u,u)$ is the Copula function ($[0,1]^2 \to [0,1]$) and C' is its derivative [4]. With the help of Sklar's theorem, $C(u,u) = P_{\mathscr{X}}(D_1 \le F^{-1}_{\mathscr{X}}(u), D_2 \le F^{-1}_{\mathscr{X}}(u))$ where $\mathscr{X}$ can be RW, PR or RWJ with $F^{-1}_{\mathscr{X}}()$ is the inverse of the stationary distribution function of the corresponding random walk.

2.3. Results

The extremal index is calculated analytically for the considered algorithms. Closed form expressions for RW and RWJ are obtained and a lower bound of θ for PR is derived.

As for numerical experiments, a random graph is generated as follows : $f(d_1, d_2)$ is taken as a bivariate Pareto distribution with suitable parameters and the degree distribution is calculated accordingly from (2). Then an uncorrelated graph is generated using the configuration model with this degree distribution. Finally the Metropolis algorithm is applied to rearrange the edges in order to get the desired joint degree-degree distribution.

The mean field argument for the transition kernel of the RW is checked with the generated graph and found to yield a reasonable match. θ is estimated for the different algorithms using the rank estimator of Copula function. Finally, for the application of cluster classification, we derive the number of nodes to be sampled to achieve a particular mean number of clusters.

References

1. K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *Lecture Notes in Computer Science*, v.6516, pages 98–109, 2010.
2. J. Beirlant, Y. Goegebeur, J.Teugels, and J. Segers. *Statistics of Extremes : Theory and Applications.* Wiley, Chichester, West Sussex, 2004.
3. M. Boguna, R. Pastor-Satorras, and A. Vespignani. Epidemic spreading in complex networks with degree correlations. *Statistical Mechanics of Complex Networks. Lecture Notes in Physica* v.625, pages 127–147, 2003.
4. A. Ferreira and H. Ferreira. Extremal functions, extremal index and Markov chains. Technical report, Notas e comunicações CEAUL, 12 2007.