УДК 519.24

МОДЕЛИРОВАНИЕ И ОЦЕНИВАНИЕ КЛАСТЕРОВ ЭКСТРЕМУМОВ СЛУЧАЙНЫХ ГРАФОВ

Н.М. Маркович

Институт проблем управления им. В.А. Трапезникова РАН

Россия, 117997, Москва, Профсоюзная ул., 65

E-mail: nat.markovich@gmail.com

Ключевые слова: случайные графы, кластеры превышения уровня, экстремальный индекс, влиятельности узлов, декластеризация сети, распределения с тяжелыми хвостами, ПейджРанг, Max-Linear Model.

Аннотация: В работе дается обзор результатов автора для моделирования и статистического оценивания кластеров случайных графов, описывающих связи узлов сложных динамических сетей таких, как Интернет, социальные, физические, транспортные и энергетические системы. Под кластерами понимаются блоки узлов, характеристики которых превышают по величине достаточно высокий уровень. В качестве характеристик узлов рассматриваются PageRank и его аналог в смысле замены всех сумм на максимумы - Max-Linear Model. Доказывается, что эти характеристики имеют тяжелые хвосты распределений, и тяжесть хвоста определяет значение экстремального индекса. Идея разбиения сети на сообщества основана на оценивании экстремального индекса каждого узла сети, рассматриваемого как корень дерева. Приводятся теоретические и прикладные результаты.

1. Введение

Объектом исследования работы являются сложные сети и соответствующие их структуре случайные графы. Графы называются случайными из-за связей (ребер) между узлами, случайно возникающих в какие-то моменты времени. Это порождает необходимость исследования характеристик узлов таких сетей, как случайных процессов.

Важными задачами статистического анализа сложных сетей являются декластеризация (т.е. поиск сообществ сети) и наиболее быстрый поиск топ-листа наиболее влиятельных узлов. В работах автора решать эти задачи предлагается, оценивая экстремальные индексы узлов сети. Экстремальный индекс является, наряду с хвостовым индексом, ключевым параметром в теории экстремальных величин, поскольку показывает меру зависимости максимумов. Его обратная величина аппроксимирует средний размер кластера превышения уровня наблюдаемой случайной последовательностью, а также определяет наименьшее время достижения (first hitting time) случайным блужданием наиболее влиятельного узла, чья характеристика превышает достаточно высокий уровень.

Доступная статистика, собираемая об узлах сети случайным блужданием, состоит из числа входящих в узел и выходящих из узла связей (in-degree и out-degree, соответственно). На основании этих данных можно восстановить графовую структуру сети и оценить с помощью известных алгоритмов характеристики влиятельности узлов. В качестве таких характеристик будем рассматривать PageRank и Max-Linear Model.

В ряде работ PageRank и Max-Linear Model случайно выбранного узла рассматривается как решение стохастического уравнения, где в правой части стоит сумма случайного числа возможно зависимых случайных величин [2,5,6,14,16]. Таким образом, теоретическое исследование распределения характеристики влиятельности, его хвостовой и экстремальный индекс, сводится к получению распределений таких сумм. Исследование этих сумм имеет широкое применение в системах массового обслуживания, финансовых и биологических приложениях, помимо случайных графов.

Освещаются следующие, полученные автором, результаты: (1) точные выражения и связь экстремального и хвостового индексов сумм случайного числа случайных величин, их интерпретация для PageRank и Max-Linear Model; (2) распределение наименьшего времени достижения первого влиятельного узла; (3) алгоритмы оценивания экстремального индекса узлов сети непараметрическими методами; (4) алгоритмы декластеризации сети с помощью оцененных значений экстремального индекса узлов сети; (5) применение алгоритмов к реальным данным.

2. Необходимые факты

Пусть $X^n = \{X_1, ..., X_n\}$ выборка случайных величин с функцией распределения $F(x)$.

Определение 1. *([7], стр. 53) Говорят, что стационарная последовательность $\{X_n\}_{n\geqslant 1}$ имеет экстремальный индекс $\theta \in [0,1]$, если для любого $0 < \tau < \infty$ существует последовательность действительных чисел $u_n = u_n(\tau)$ таких, что выполнено*

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

$$\lim_{n\to\infty} P\{M_n \leqslant u_n\} = e^{-\tau\theta},$$

где $M_n = \max\{X_1, ..., X_n\}$.

Наиболее популярными оценками экстремального индекса являются оценки блоков (blocks estimator), пробежек (runs estimator), интервалов (intervals estimator), а также их модификации, такие как например, оценка скользящих блоков (sliding blocks estimator), направленные на уменьшение смещения или дисперсии оценки, [3, 11–13, 15].

Согласно определению Google [1] ПейджРанг Веб-страницы p_i определяется как

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

где $N(p_i)$ - число страниц, имеющих входящую связь в p_i (in-degree), D_j - число выходящих связей из p_j (out-degree), $c \in (0,1)$ - коэффициент демпфирования. $q = (q_1, q_2, ..., q_n)$ - вектор телепортаций такой, что $q_i \geqslant 0$ и $\sum_{i=1}^{n} q_i = 1$, например, $q_i = 1/n$ соответствует равномерно-распределенному выбору произвольного узла сети. n - общее число страниц или узлов Веб графа. Для простоты в определении опущено слагаемое, соответствующее висячим вершинам.

В качестве процедур оценивания ПейджРанга известен ряд рекуррентных методов, среди них [2]. Разработка новых методов оценивания ПейджРанга не является целью нашей работы.

Альтернативой ПейджРангу является Модель Линейного Максимума (Max-Linear Model (MLM)), [4]

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

Эта модель получена заменой суммы на максимум в (1) и может быть полезна, если доступна статистика только о наибольших рангах влиятельности узлов.

В [2, 5, 6, 14, 16] показано, что PageRank и Max-Linear Model случайного узла Веб графа, рассматриваемого как корень дерева Гальтона-Ватсона, могут рассматриваться как случайные величины, решения стохастических уравнений

$$R =^D \sum_{j=1}^{N} A_j R_j + Q, \tag{3}$$

$$R =^D \left(\bigvee_{j=1}^{N} A_j R_j \right) \vee Q, \tag{4}$$

аналогов (1) и (2), где равенство $=^d$ понимается по распределению. Связь между (1) и (3), (2) и (4) доказана через сходимость рекуррентных оценок ПейджРанга и Max-Linear Model к соответствующим решениям стохастических уравнений при следующих условиях: $\{R_j\}$ являются независимыми, одинаково распределенными копиями R, не зависящими от $(Q, N, \{A_j\})$ с $\{A_j\}$, независимыми от (N, Q). Эти условия были ослаблены в [8].

3. Теория

Пусть $\{Y_n^{(1)}, Y_n^{(2)}, ..., Y_n^{(l)}\}$, $n \geqslant 1$, $l \geqslant 1$, последовательности случайных величин, имеющих стационарные распределения с хвостовыми индексами $\{k_1, ..., k_l\}$ и экстремальные индексы $\{\theta_1, ..., \theta_l\}$, т.е.

$$P\{Y_n^{(i)} > x\} \sim c^{(i)} x^{-k_i} \qquad \text{as} \qquad x \to \infty, \tag{5}$$

где $c^{(i)}$ положительные вещественные константы.

В [8] были доказаны теоремы, относящиеся к максимумам и суммам многомерных последовательностей случайной длины. Пусть $\{N_n\}$, $n \geqslant 1$, последовательность

целочисленных случайных величин таких, что $N_n \to^P +\infty$, $n \to \infty$. Предполагается, что N_n имеет распределение (5) с индексом $\alpha > 0$. Пусть $\{Y_n^{(j)}, n \geqslant 1, j = 1, 2, ...\}$ и $\{Q_n\}$ последовательности случайных величин с стационарными распределениями (5) и с положительными хвостовыми индексами $\{k_1, k_2, ...\}$ и β соответственно. Пусть $\{Y_n^{(i)}\}$ не зависят от $\{Q_n\}$, $\{Q_n\}$ - последовательность независимых, одинаково распределенных случайных величин. Пусть $\{Y_n^{(1)}, Y_n^{(2)}, ...\}$ имеют экстремальные индексы $\{\theta_1, \theta_2, ...\}$, соответственно. Обозначим

$$Y_n^*(z, N_n) = \max(z_1 Y_n^{(1)}, .., z_{N_n} Y_n^{(N_n)}, Q_n),$$

$$Y_n(z, N_n) = z_1 Y_n^{(1)} + .. + z_{N_n} Y_n^{(N_n)} + Q_n.$$

Теорема 1. *Пусть* $k_1 < k_2 \leqslant k_3 \leqslant ....$ *Тогда последовательности* $Y_n^*(z, N_n)$ *и* $Y_n(z, N_n)$ *имеют распределение (5) с тем же хвостовым индексом* $k^*(z) = \min(\alpha \cdot \chi \cdot k(z), k_1, \beta)$, *где* $k(z) = \min(\alpha \cdot \chi, k_1, \beta)$, χ *определяется как* $0 < \chi \leqslant 1/(1 + k_1(k_1 + 2))$ *и тем же экстремальным индексом* $\theta(z)$ *таким, что* $\theta(z) = \{\theta_1, 1, 0\}$ *при* $k(z) = \{k_1, \beta, \alpha\chi\}$, *соответственно, если* $k^*(z)/k(z) = 1$.

Следующая теорема доказана в предположениях Теоремы 1, но теперь $\{Y_n^{(j)}, n \geqslant 1, j = 1, 2, ...\}$ предполагаются независимыми с тем же положительным хвостовым индексом k, и $\{N_n\}$ не зависит от $\{Y_n^{(j)}\}$ и $\{Q_n\}$.

Теорема 2. *Последовательности* $Y_n^*(z, N_n)$ *и* $Y_n(z, N_n)$ *имеют распределение (5) с тем же хвостовым индексом* $k(z) = \min(k, \beta)$ *и тем же экстремальным индексом* $\theta(z)$ *таким, что* $\theta(z) = \{E(d_{N_n}(z, \theta))/E(c_{N_n}(z)), 1\}$ *при* $k(z) = \{k, \beta\}$, *соответственно, где*

$$c_{N_n}(z) = \sum_{j=1}^{N_n} c^{(j)} z_j^k, \quad d_{N_n}(z, \theta) = \sum_{j=1}^{N_n} c^{(j)} z_j^k \theta_j.$$

Эти теоремы интерпретированы для (3) и (4) в [8].

4. Практика

В [9] и [10] предлагаются алгоритмы оценивания экстремального индекса на графах с помощью модифицированной оценки блоков, а также кластеризации графа по значениям экстремального индекса ПейджРангов и Max-Linear Model узлов. Работа алгоритмов демонстрируется на данных реальной сети. При этом теоретические результаты об идентичности экстремального индекса для ПейджРангов и Max-Linear Model находят подтверждение на реальных измерениях сети.

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (19-01-00090а).

Список литературы

1. Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine // Computer Networks and ISDN Systems. 1998. Vol. 30, No. 1, P. 107-117.

2. Chen N., Litvak N. and Olvera-Cravioto M. PageRank in Scale-Free Random Graphs // WAW 2014. LNCS 8882 / Ed. A. Bonato et al. Springer, 2014. P. 120-131.
3. Drees H. Bias correction for estimators of the extremal index // Preprint, arXiv: 1107.0935. 2011.
4. Gissibl N., Klüppelberg C. Max-linear models on directed acyclic graphs. arXiv:1512.07522v1[math.PR] 2015. P. 1-33.
5. Jelenkovic P. R., Olvera-Cravioto M. Information ranking and power laws on trees // Adv. Appl. Prob. 2010. Vol. 42, No. 4. P. 1057-1093.
6. Jelenkovic P. R., Olvera-Cravioto M. (2015). Maximums on trees // Stoch. Process. Appl. Vol. 125. P. 217-232.
7. Leadbetter M. R., Lingren G. and Rootzen H. Extremes and Related Properties of Random Sequence and Processes. Ch.3. Springer, 1983.
8. Markovich N.M. Extremes of random length sequences with application to random networks // Submitted. 2018.
9. Markovich N.M., Ryzhov M.S., Krieger U.R. Nonparametric Analysis of Extremes on Web Graphs: PageRank versus Max-Linear Model // Proceedings of the IEEE 19th Distributed computer and communication networks: control, computation, communications (DCCN-2017). Moscow, 25-29 September 2017. Communications in Computer and Information Science (20th International Conference DCCN2017 Moscow). CCIS. 2017. Vol. 700. P. 13-26
10. Markovich N.M., Ryzhov M.S., Krieger U.R. Statistical Clustering of a Random Network by Extremal Properties // Proceedings of the IEEE 20th Distributed computer and communication networks: control, computation, communications (DCCN-2018). Moscow, 17-21 September 2018. Communications in Computer and Information Science (21th International Conference DCCN2018 Moscow). CCIS. 2018. Vol. 919. P. 71-82.
11. Northrop P.J. An efficient semiparametric maxima estimator of the extremal index // Extremes. 2015. Vol. 18, No. 4. P. 585-603.
12. Robert C.Y. Asymptotic distributions for the intervals estimators of the extremal index and the cluster-size probabilities // Journal of Statistical Planing and Inference. 2009. Vol. 139. P. 3288-3309.
13. Robert C.Y., Segers J., Ferro C.A.T. A sliding blocks estimator for the extremal index // Electronic Journal of Statistics. 2009. Vol. 3. P. 993-1020.
14. Olvera-Cravioto M. Asymptotics for weighted random sums // Adv. Appl. Prob. 2012. Vol. 44, No. 4. P. 1142–1172.
15. Sun J., Samorodnitsky G. Multiple thresholds in extremal parameter estimation. Extremes. 2018. https://doi.org/10.1007/s10687-018-0337-5.
16. Volkovich Y. V., Litvak N. (2010). Asymptotic analysis for personalized web search // Adv. Appl. Prob. Vol. 42, No. 2. P. 577-604.