УДК 519.248
AMS MSC2020: 60K40

Статистический анализ случайных графов для задачи распространения информации[1]

Маркович Н. М., Рыжов М. С.

Институт проблем управления им. В. А. Трапезникова РАН

Аннотация. Работа посвящена распространению сообщений в случайных графах. Рассматривается задача передачи сообщения каким-то узлом графа другим узлам в графе. Для этой цели среди узлов графа находятся лидирующие узлы, которые наиболее быстро распространяют информацию, а также лидирующие сообщества, к которым такие узлы относятся. С помощью статистических методов, оценивая экстремальные и хвостовые индексы сообществ, проводится исследование фиксированных и динамически меняющихся графов, в которых распределения числа входящих и выходящих связей между узлами задается степенным законом с известными параметрами.

Ключевые слова: случайный граф, распространение сообщений, сообщество, хвостовой индекс, экстремальный индекс.

Введение

Распространение сообщений в случайных графах является важной задачей с приложением в различных областях, как распределенные вычисления [2, 11] и социальные сети. Например, время распространения инфекции в контактных сетях [5] может влиять на эффективность вакцинации.

В прошлых работах авторов [7, 8] были исследованы распределения ПейджРангов узлов и свойства сообществ узлов посредством экстремального и хвостового индексов. В качестве сообщества рассматривалась группа узлов, которые связаны между собой большим

[1]Работа выполнена при финансовой поддержке РФФИ (грант 19-01-00090)

числом связей и мало связаны с остальными узлами графа. В настоящей работе рассматриваются результаты исследования экстремального и хвостового индексов для задачи распространения сообщений в случайном графе.

Анализируя фиксированные ненаправленные графы (Секция 2) и эволюцию во времени направленных графов (Секция 3), в статье приводятся обнаруженные зависимости между экстремальным и хвостовым индексами, оцениваемыми по множествам характеристик узлов сообществ, и временем распространения сообщения.

1. Основные определения

1.1. Экстремальный индекс

Для стационарной последовательности $\{X_n\}_{n\geqslant 1}$ с функцией распределения (ф. р.) $F(x)$ и максимумом $M_n = \max\limits_{1\leqslant j\leqslant n} X_j$ существует экстремальный индекс (ЭИ) $\theta \in [0,1]$, если для любого $0 < \tau < \infty$ найдется вещественная последовательность $u_n = u_n(\tau)$ удовлетворяющая соотношениям

$$\lim_{n\to\infty} n(1-F(u_n)) = \tau \qquad \text{и} \qquad \lim_{n\to\infty} P\{M_n \leqslant u_n\} = e^{-\tau\theta},$$

в том смысле, что M_n остается ограниченным при $n \to \infty$ ([6], p. 63).

Для независимых случайных величин ЭИ равен единице. Обратное утверждение неверно. Чем ближе θ к нулю, тем сильнее степень локальной зависимости (кластерности). Обратная величина $1/\theta$ аппроксимирует средний размер кластера, то есть среднее число превышений на кластер. В классической теории кластер может определяться как блок данных с хотя бы одним превышением уровня u. В [3] кластер определяется, как число $T(u)$ превышений между двумя последовательными не превышениями

$$T(u) = \min\{t \geqslant 1 : X_{j+t} > u\} \quad \text{при} \quad X_j > u. \tag{1}$$

Значение ЭИ может быть оценено с помощью интервальной оценки [3]

$$\widehat{\theta}(u) = \min(1, \theta^*), \tag{2}$$

$$\theta^* = \begin{cases} \dfrac{2\Big(\sum\limits_{i=1}^{N-1} T(u)_i - 1\Big)^2}{(N-1)\sum\limits_{i=1}^{N-1}(T(u)_i - 1)(T(u)_i - 2)}, & \max\{T(u)_i\} > 2, \\ \dfrac{2\Big(\sum\limits_{i=1}^{N-1} T(u)_i\Big)^2}{(N-1)\sum\limits_{i=1}^{N-1}(T^2(u)_i)}, & \text{иначе,} \end{cases}$$

где $N = N(u) = \sum\limits_{i=1}^{n} \mathbb{1}(X_i > u)$. Чтобы ввести интервальную оценку для графа, предлагается определить $T(u)$ как количество ребер кратчайшего пути между узлами, характеристики которых больше, чем u [9]. Такое предположение помогает рассматривать ЭИ сообществ узлов, используя известные результаты, полученные для последовательностей.

1.2. Хвостовой индекс

Пусть $\{X_n\}_{n\geqslant 1}$ — стационарная последовательность независимых одинаково распределенных (н. о. р) случайных величин (с. в.) с ф. р. $F(x)$. Параметр α_{TI} называется хвостовым индексом (ХИ). Это может быть оценено с помощью оценки Хилла [4]

$$\widehat{\alpha}^H(k) = \left(\frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(n-i+1)}}{X_{(n-k)}}\right)\right)^{-1}, \tag{3}$$

где $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ порядковые статистики, соответствующие выборке. k — это число наибольших порядковых статистик, оптимальное значение которого выбирается методом бутстреп (bootstrap) [4].

Для исследования распределения характеристик узлов в случайном графе требуется полагать, что характеристики является

независимыми друг от друга, что в общем случае не является верным. Предполагая это условие выполненным или проверяя его с помощью статистических тестов, можно исследовать ХИ характеристик сообществ.

2. Распространение сообщений в ненаправленном графе

Опишем алгоритм распространения SPREAD, предложенный в [11] для неориентированного графа $G = (V, E)$. Здесь V и E — наборы вершин и ребер графа соответственно. Рассмотрим асинхронную модель времени, где узел может инициировать связь по тикам глобальных часов, которые моделируются как процесс Пуассона с параметром $n = |V|$, [2, 11]. Пусть $k \geqslant 0$ обозначает номер тика часов или шаг алгоритма. При каждом k только один узел i, выбранный равномерно среди всех, может отправить все свои имеющиеся сообщения, связываясь с другим узлом j с вероятностью $P_{ij} = 1/D_i$, где D_i — степень или количество связей узла i.

Выберем узел x с характеристикой X_x и сообщество узлов S, к которому он может принадлежать. Чтобы определить ЭИ сообщества S (предполагая, что ЭИ существует), берем высокий квинтиль $F(x)$ распределения выбранной характеристики в качестве порогового значения u^*. Для оценивания необходимо определить множество кратчайших путей $\{T(u^*)_i\}$ для всех возможных пар $(x, y) \in S$ [9]. Событие $\{T(u^*)_i = m\}$ означает, что характеристики $X_{i_1}, X_{i_2}, \ldots, X_{i_m}$ в последовательности X_{xy} меньше, чем u^*, но X_x и X_y превышают u^*.

Моделирование показало, что узлы с высокими значениями близости (closeness centrality)

$$C_x = \frac{n-1}{\sum_{y, y \neq x} d(x, y)}, 0 < C_x \leqslant 1,$$

[12], где $d(x, y)$ — длина кратчайшего пути между узлами x и y, распространяют информацию быстрее прочих узлов [9]. Такие узлы являются лидирующими узлами для задачи распространения информации. Также верно, что узлы с большим количеством связей D_x быстрее распространяют сообщения. Также были определены

лидирующие сообщества, содержащие наилучших узлов распространителей. Такие сообщества имеют такие же значения ЭИ, что и весь граф, рассмотренный как отдельное сообщество.

3. Распространение сообщений в направленном динамически меняющемся графе

Метод эволюции линейного предпочтительного присоединения (ПА) (Preferential Attachment) [1,13] начинает работу с начального ориентированного графа $G(k_0)$ с хотя бы одним узлом и k_0 ребрами. Для неотрицательных параметров α, β, γ, таких как $\alpha+\beta+\gamma=1$, и $\Delta_{in}, \Delta_{out}$, ПА строит растущую последовательность направленных случайных графов $G(k) = (V(k), E(k))$. Граф $G(k)$ создается из $G(k-1)$ путем добавления нового направленного ребра. Обозначим число узлов на шаге k через $N(k)$, а число входящих (in-degree) и выходящих связей (out-degree) узла w в графе $G(k)$ с ребрами k как $I_k(w)$ и $O_k(w)$. В [1,13] предложены три сценария создания ребра. На каждом шаге алгоритма путем подбрасывания 3-сторонней монеты с вероятностями α, β и γ выбирается один из сценариев:

- В соответствии с α-схемой добавляется новый узел w_{new} и ребро $(w_{new} \to w)$ с вероятностью α. Существующий узел $w \in V(k-1)$ выбирается с вероятностью

$$P(w \in V(k-1)) = \frac{I_{k-1}(w) + \Delta_{in}}{k-1+\Delta_{in}N(k-1)}.$$

- В соответствии с β-схемой добавляется новое ребро $(w_1 \to w_2)$ с вероятностью β, где оба существующих узла w_1 и w_2 выбираются независимо и с вероятностью

$$P(w_1 \to w_2) = \frac{O_{k-1}(w_1) + \Delta_{out}}{k-1+\Delta_{out}N(k-1)} \cdot \frac{I_{k-1}(w_2) + \Delta_{in}}{k-1+\Delta_{in}N(k-1)}.$$

- В соответствии с γ-схемой добавляется новый узел w_{new} и ребро $(w \to w_{new})$ с вероятностью γ, $w \in V(k-1)$ выбирается с вероятностью

$$P(w \in V(k-1)) = \frac{O_{k-1}(w) + \Delta_{out}}{k-1+\Delta_{out}N(k-1)}.$$

Это означает, что $N(k) = N(k-1)$ для β-схемы и $N(k) = N(k-1)+1$ для остальных.

Несмотря на то, что линейный ПА используется для эволюции ориентированных графов, он так может быть использован как модель для распространения информации [10]. Предполагая, что сообщение, находящееся в одном из узлов, распространяется среди фиксированного числа n узлов, на каждом шаге ПА с предопределенными значениями параметров сообщение может быть доставлено от узла i в узел j только, если создано направленное ребро $(i \to j)$. Такое ребро может быть добавлено к сети только с помощью $\gamma-$ или $\beta-$ схем. Если узел i не имеет сообщения, то ребро $(i \to j)$ не распространяет сообщение дальше на узел j. Схема α увеличивает число узлов без сообщения.

Для анализа скорости распространения сообщений с помощью ПА модели было проведено ее сравнение с алгоритмом SPREAD при $P_{ij} = 1/O_i$ и $P_{ij} = 1/(O_i + I_i)$. Здесь P_{ij} обозначает вероятность, что узел i выберет узел j для передачи сообщения. При исследовании некоторых модельных графов было показано, что ПА может быстрее распространять информацию для некоторых наборов параметров (α, β, γ) при $\Delta_{in} = \Delta_{out} = 1$, чем SPREAD [10].

Также было проведено моделирование неоднородных графов, состоящих из сообществ узлов с разными распределениями числа входящих и выходящих связей. Эти распределения, как показано многими авторами, имеют правильно меняющиеся хвосты. Было обнаружено, что узлы из сообществ с наименьшим ХИ для распределения out-degree, то есть с распределением, имеющим наиболее тяжелый хвост, распространяют свое сообщение быстрее, чем узлы из прочих сообществ [10]. Такие сообщества могут быть названы лидирующими.

Заключение

В работе исследовались фиксированные ненаправленные графы (Секция 2) и динамически развивающиеся направленные графы (Секция 3). Используя непараметрические оценки экстремального и хвостового индексов сообщества узлов графа, было обнаружено существование лидирующих сообществ, то есть групп, связанных между собой большим числом ребер и содержащих узлы — лидеры

по скорости распространения информации. Моделирование на ряде примеров графов в работах [9] и [10] показало следующие результаты. Для фиксированных ненаправленных графов лидирующие сообщества имеют то же значение экстремального индекса, что и весь граф. Для динамических направленных графов, эволюционирующих во времени, лидирующие сообщества обладают наименьшим хвостовым индексом для числа выходящих связей (out-degree) по сравнению с прочими сообществами.

Список литературы

[1] *Bollobás, B.* Directed scale-free graphs / B. Bollobás, C. Borgs, J. Chayes, O. Riordan // In Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (SODA '03). — Philadelphia, Pennsylvania : Society for Industrial and Applied Mathematics, 2003. — P. 132–139.

[2] *Censor-Hillel, K.* Partial Information Spreading with Application to Distributed Maximum Coverage / K. Censor-Hillel, H. Shachnai // In Proceedings of the 29th ACM symposium on Principles of distributed computing (PODC '10). — New York, N. Y.: ACM, 2010. — P. 161–170.

[3] *Ferro, C.* Inference for clusters of extreme values /C. Ferro, J. Segers // Journal of the Royal Statistical Society. Series B (Statistical Methodology). — 2003. — Vol. 65, №2. — P. 545–556.

[4] *Hall, P.* Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems // Journal of Multivariatte Analysis. — 1990. — Vol. 32. — P. 177–203.

[5] *Holme, P.* Cost-efficient vaccination protocols for network epidemiology / P. Holme, N. Litvak // PLoS Computational Biology. — 2017. — Vol. 13, №9. — e1005696.

[6] *Leadbetter, M. R.* Extremes and local dependence in stationary sequences // Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete. — 1983. — Vol. 65. — P. 291–306.

[7] *Markovich, N. M.* Nonparametric Analysis of Extremes on Web Graphs: PageRank versus Max-Linear Model / N. M. Markovich, M. S. Ryzhov, U. R. Krieger // CCIS-2017. — Vol. 700. — P. 13–26.

[8] *Markovich, N. M.* Statistical Clustering of a Random Network by Extremal Properties / N. M. Markovich, M. S. Ryzhov, U. R. Krieger // CCIS-2018. — Vol. 919. — P. 71–82.

[9] *Markovich, N. M.* Leader Nodes in Communities for Information Spreading / N. M. Markovich, M. S. Ryzhov // LNCS. — 2020. — Vol. 12563. — P. 475–484.

[10] *Markovich, N. M.* Information Spreading with Application to Non-homogeneous Evolving Networks / N. M. Markovich, M. S. Ryzhov // DCCN 2021 (Принято к публикации).

[11] *Mosk-Aoyama, D.* Computing separable functions via gossip / D. Mosk-Aoyama, D. Shah // In Proceedings of the 25th ACM symposium on Principles of distributed computing (PODC '06). — New York, N. Y.: ACM, 2006. — P. 113–122.

[12] *Newman, M. E. J.* Networks: An Introduction. — 2nd ed. — Oxford : Oxford University Press, 2018. — 800 p.

[13] *Wan P.* Are extreme value estimation methods useful for network data? / P. Wan, T. Wang, R. A. Davis, S. I. Resnick // Extremes. — 2020. — Vol. 23. — P. 171–195.

Библиографическая ссылка

Маркович, Н. М. Статистический анализ случайных графов для задачи распространения информации / Н. М. Маркович, М. С. Рыжов // Всероссийская научная конференция «Математические основы информатики и информационно-коммуникационных систем». Сборник трудов. — Тверь : ТвГУ, 2021. — С. 204–212.
https://doi.org/10.26456/mfcsics-21-30

Сведения об авторах

1. Наталья Михайловна Маркович
Институт проблем управления им. В. А. Трапезникова РАН. Главный научный сотрудник

Россия, 117997, Москва, ул. Профсоюзная, 65
E-mail: markovic@ipu.rssi.ru , nat.markovich@gmail.com

2. Максим Сергеевич Рыжов
Институт проблем управления им. В. А. Трапезникова РАН. Младший научный сотрудник

Россия, 117997, Москва, ул. Профсоюзная, 65
E-mail: maksim.ryzhov@frtk.ru