Тестирование стационарности сообществ в динамических реальных графах

Васильев Михаил Владимирович

11 мая 2024 г.

1 Введение

Рассматриваются эволюция реальных ненаправленных сетей и описывающих их графов. Обозначим граф на шаге эволюции t, как $G_t = (V_t, E_t)$, V_t- множество вершин, E_t- множество ребер. Эволюция, т.е. присоединение новых узлов новыми связями к существующим узлам, начинается с начального графа G_0 и имеет сложную природу. Существуют некотороые методы моделирования эволюции графа такие как предпочтительное присоединение (ПП, preferential attachment), кластерное присоединение (КП, clustering attachment) и их смеси. [9].

Изучаются методы разбиения реальных графов на сообщества, такие как алгоритм Leuven [8].

В работе данные взяты из открытого графового репозитория [2], а именно динамический граф fb-messages.csv. Граф составлен на основе сообщений студентов Калифорнийского университета в Ирвине в социальной сети, похожей на Facebook. В набор данных входят пользователи, которые отправили или получили хотя бы одно сообщение. Граф имеет три столбца: source, target, time. Количество вершин - 1900, количество рёбер - 61732.

Цель работы состоит в разработке и сравнении методов тестирования стационарности распределений сообществ при различных методах разбиения на сообщества.

2 Описательные характеристики

Зафиксируем состояние графа на последний момент времени, рассчитаем для каждого его узла PageRank и проведём исследование рас-

спределения значений PageRank. Исследование состоит в описании распределения и оценивании индекса экстримальной величины γ разными методами. Методы оценивания включают в себя: оценку Хилла (Hill), оценку отношения (Ratio), оценку моментов (Moment) [10] и смешанную оценку ментов (Mixed Moment) [11]. Так же интерес представляет исследование тяжести хвоста распределения.

2.1 Гистограмма распределения

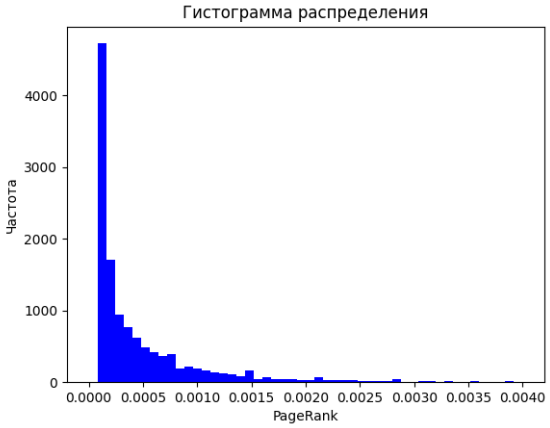

Рис. 1: Распределение PageRank.

2.2 Расчёт количества моментов распределения

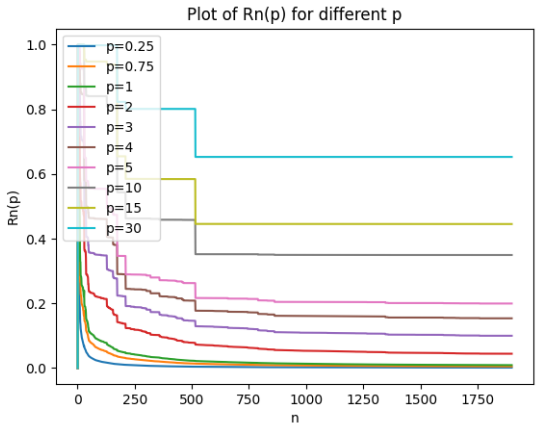

Рис. 2: Зависимость отношшения максимума к сумме элементов для различных значений хвостового индекса.

По итогам исследования можно сказать, Для $p \in \{0.25, 0.75, 1\}$ $R_n(p)$ по всей видимости стремиться к нулю при возрастании n. Для $p \in \{2, 3, 4, 5, 10, 15, 30\}$ $Rn(p)$ по всей видимости стремиться к положительной константе при возрастании n.

Вывод: $E|X|^p < \infty$ для $p \leqslant 1$ только, $E|X|^p = \infty$ для $p > 1$. Значит распределение имеет единственный момент.

2.3 Оценка распределения

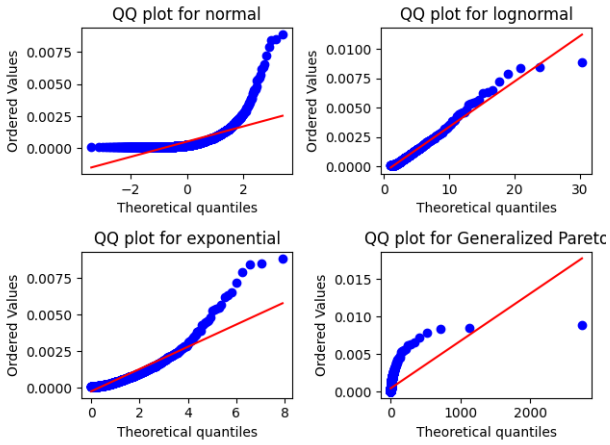

Рис. 3: QQ-plot для нормального распределения.

Как видно нормальное, экспоненциальное и Парето обобщённое распределения не подходят для выборки. В то время как логнормальное распределение относительно подходит для выборки.

2.4 Оценки индекса экстримального значения

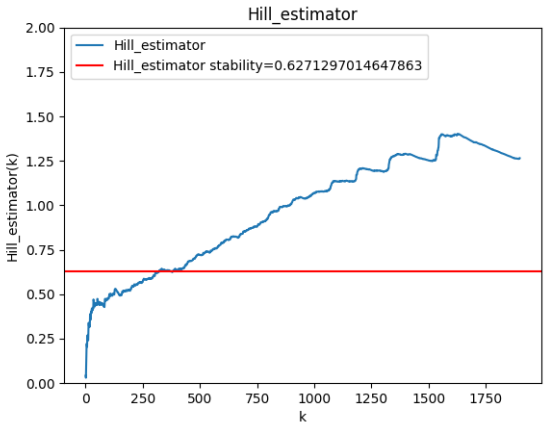

Рис. 4: Hills estimator plot.

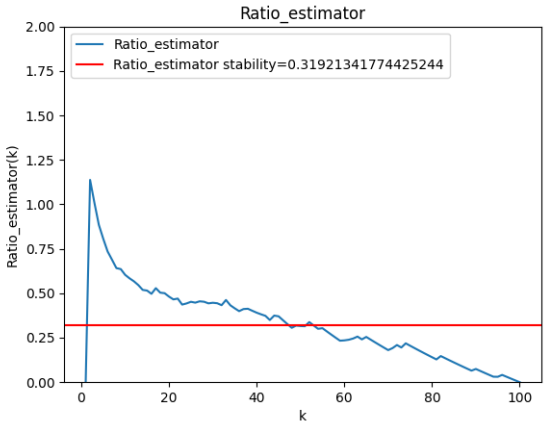

Рис. 5: Ratio estimator plot.

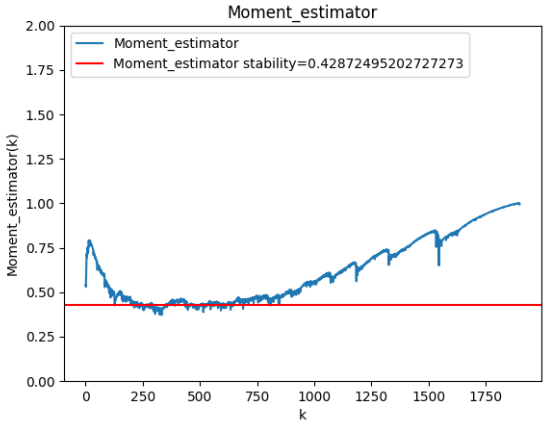

Рис. 6: Moment estimator plot.

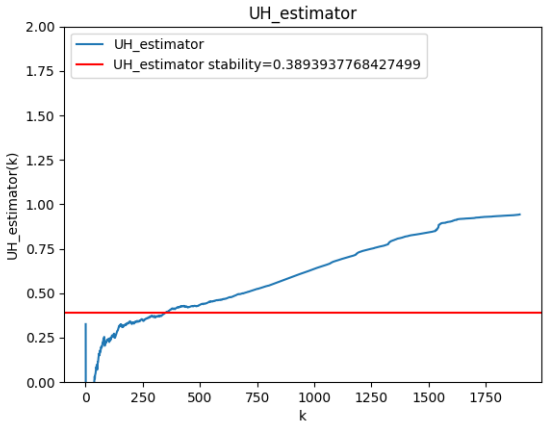

Рис. 7: UH estimator plot.

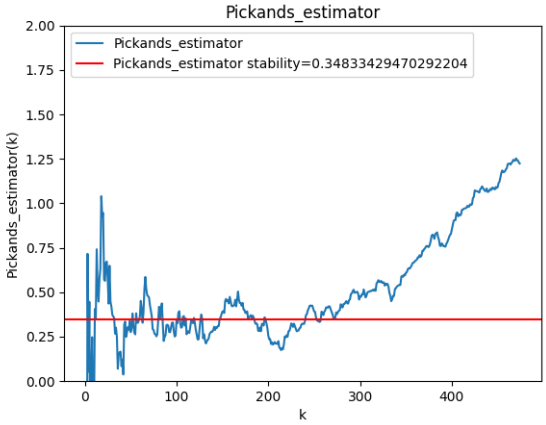

Рис. 8: Pickands estimator plot.

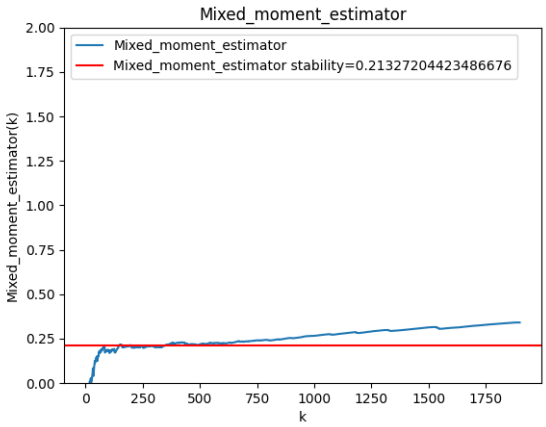

Рис. 9: Mixed estimator plot.

Из графиков видно, что оценки Hill, Ratio и UH на предоставленных данных работают плохо. Тогда как оценки Moment, Pickands и Mixed Moment имеют интервалы стабильности. На данный момент не понятно, какую оценку лучше взять. Изучению этого вопроса, следует уделить особое внимание.

3 Выявление сообществ в графе

Сообщество - это группа тесно взаимодействующих узлов, которые слабо связаны с остальным графом. Одним из методов разбиения стационарного графа на сообщества является алгоритм Leuven. [8]

Библиотека NetworkX позволяет разделить граф на сообщества посредствам этого алгоритма.

Номер сообщества	Количество вершин
0	125
1	2
2	236
3	367
4	122
5	41
6	331
7	248
8	29
9	287
10	107
11	2
12	2

Таблица 1: Таблица количества вершин в сообществах графа.

Выявлено 13 сообществ разного размера.

Для сообщества 2 проведён визуальный анализ с помощью пакета gephi.

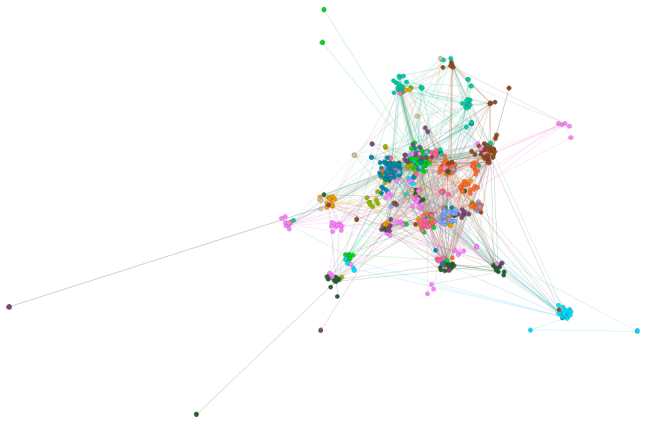

Рис. 10: Сообщество 2

4 Стационарность сообществ

4.1 Подход случайных последовательностей

Одним из подходов для тестирования сообществ на стационарность распределения может быть сведение к тестированию случайных последовательностей, возникающих на случайных графах. Для этого могут быть использованы случайные блуждания для сбора информации о характеристиках влиятельности (например число входящих связей узла или PageRank) узлов сообщества, которые определяют соответствующие случайные последовательности.

Введём определение стационарной последовательности, как последовательности в которой распределение значений $(\varepsilon_{j_1}, ..., \varepsilon_{j_n})$ и $(\varepsilon_{j_1+m}, ..., \varepsilon_{j_n+m})$ одинакого для любого выбора $n, j_1, ..., j_n$, и m [7].

4.2 Подход разбиения на подграфы

Ещё один подход - это разбиение сообщества на более мелкие подграфы, рассматривая их как блоки данных. В каждом блоке можно оценить

индекс экстремальной величины или его обратную величину, хвостовой индекс, и понять, насколько сильно меняется его величина.

4.3 Исследование временной динамики

Реализуем подход разбиения на подграфы и проведём проверку на реальных данных. Для этого поступим следующим образом. Рассмотрим состояние динамического графа на каждом из 23 эволюционных периодов. Для каждого из периодов рассчитаем PageRank каждой вершины, выделим 30 случайных последовательностей по 300 значений в каждой. Оценим для каждой последовательности индекс экстримальной величины γ. И усредним индексы по последовательностям, так же рассчитаем дисперсию оценки. На рис. 11 представлен результат.

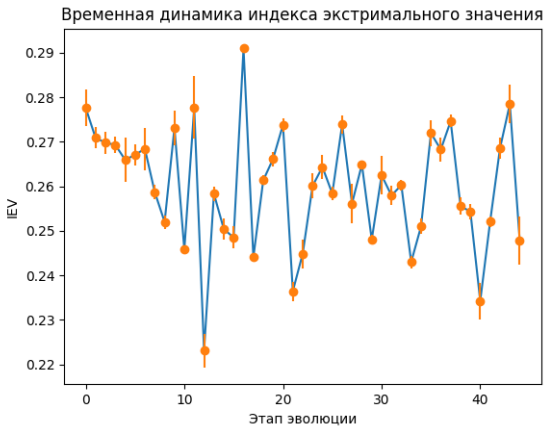

Рис. 11: Временная динамика усреднённого ИЭВ.

Интересно сопоставить временну динамику ИЭВ с динамикой количесва вершин на тех же этапах эволюции.

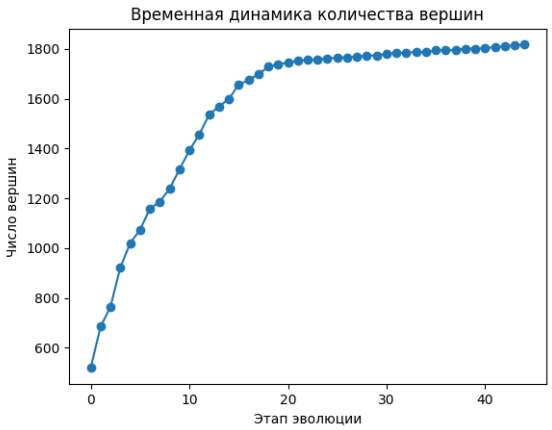

Рис. 12: Временная динамика количества вершин

5 Выводы

Произведено описание динамического реального графа. Приведены некоторые подходы для проверки стационарности сообществ. В дальнейшем эти подходы будут применены к реальным данным.

Список литературы

[1] Райгородский А. М. Модели случайных графов и их применения // Труды МФТИ. 2010.

[2] Dynamic Networks // Network Repository URL: https://networkrepository.com (дата обращения: 05.05.2024).

[3] Маркович Н.М., Вайсиулюс М.Р. Extreme Value Statistics for Evolving Random Networks // Mathematics. 2023. 11(9). С. 2171 https://www.mdpi.com/2227-7390/11/9/2171.

[4] Nicolas Dugué, Anthony Perez. Directed Louvain : maximizing modularity in directed networks. [Research Report] Université d'Orléans. 2015.

[5] BEIRLANT J., GOEGEBEUR Y., TEUGELS J., SEGERS J. Applications. – Chichester, West Sussex: Wiley, 2004 – 504 p.

[6] BAGROW J., BROCKMANN D. Natural Emergence of Clusters and Bursts in Network Evolution // Physical Review X. – 2012 –V. 3 – No. 2 –P. 21016

[7] Leadbetter, M.R., Lingren, G. Rootz´n, H. (1983). Extremes and Related Properties of Random Sequence and Processes. ch.3, New York: Springer.

[8] Nicolas Dugué, Anthony Perez. Directed Louvain : maximizing modularity in directed networks. [Research Report] Université d'Orléans. 2015.

[9] ARNOLD N.A., MONDRAG Likelihood-based approach to discriminate mixtures of network models that vary in time // Sci. Rep. –2021. –No.11 –P. 5205

[10] DEKKERS A. L. M., EINMAHL J. H. J., DE HAAN L. A Moment Estimator for the Index of an Extreme-Value Distribution // Ann. Statist. –1989. –No. 17 –P. 1833–1855

[11] FRAGA ALVES M.I., GOMES M.I., DE HAAN L. Mixed moment estimator and location invariant alternatives //Extremes. – 2009 – No. 12 – P. 149-–185.