Тестирование стационарности сообществ

Васильев Михаил Владимирович
Студент 5 курса факультета ФРКТ

(Московский физико-технический институт)
(Dated: 11 мая 2024 г.)

I. ВВЕДЕНИЕ

Постановка задачи состоит в исследовании стационарности динамического графа во времени. Для этого используется датесет графов из открытых источников таких как (1). Одним из подходов исследования стационарности сообществ является сведение к тестированию случайных последовательностей, возникающих на графе. Для этого используются случайные блуждания для сбора информации о характеристиках влиятельности (PageRank), узлов сообщества, которые определяют соответствующие случайные последовательности. Далее производится вычисление хвостового индекса на элементах этих последовательностей и посредствам визуального анализа делается вывод о стационарности графа во времени. Ключевым критерием анализа является значительное изменение хвостового индекса между соседними временными промежутками.

II. ДАННЫЕ

Данные взяты из открытого графового репозитория (1), а именно динамический граф fb-messages.csv. Описание: "Социальная сеть, похожая на Facebook, создана на основе онлайн-сообщества студентов Калифорнийского университета в Ирвине. 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.

III. ОПИСАТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ

Рассмотрим граф как статитеческий на последний момент времени, рассчитаем для каждого его узла PageRank и проведём исследование расспределения. Исследование состоит в оценивании индекса экстримального значения γ разными методами и изучении хвоста распределения. Методы оценивания включают в себя: оценку Хилла (Hill), оценку моментов (moment), смешанную оценку ментов (Mixed Moment) и оценку отношения (Ratio).

A. Гистограмма распределения

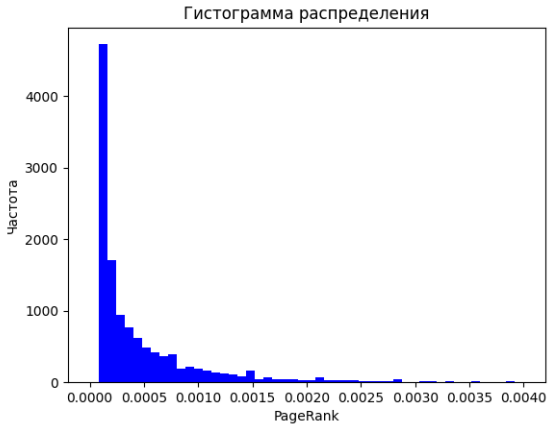

Рис. 1: Распределение PageRank

B. Расчёт количества моментов распределения

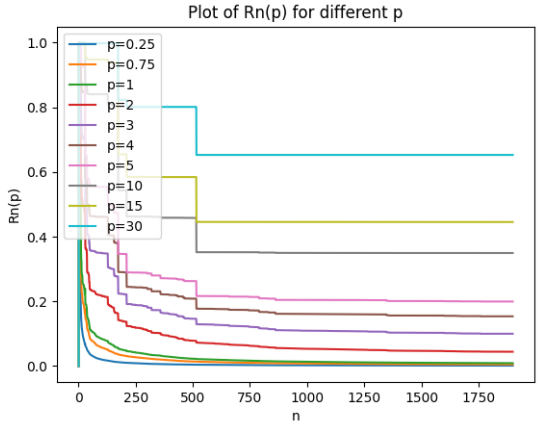

Рис. 2: Зависимость отношшения максимума к сумме элементов для различных значений хвостового индекса

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.

C. Оценка распределения

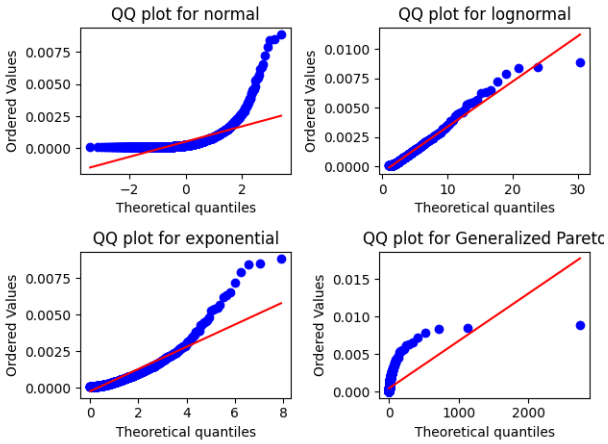

Рис. 3: QQ-plot для нормального распределения

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

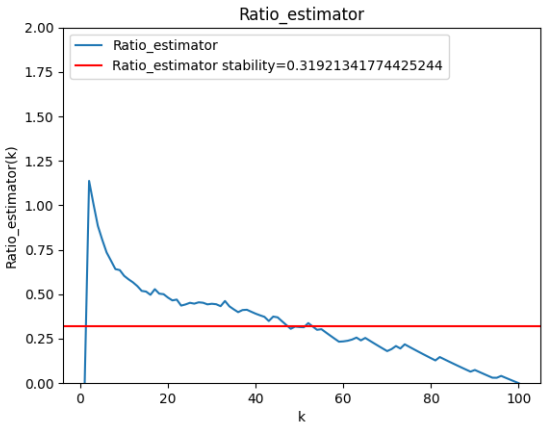

Рис. 5: Ratio estimator plot

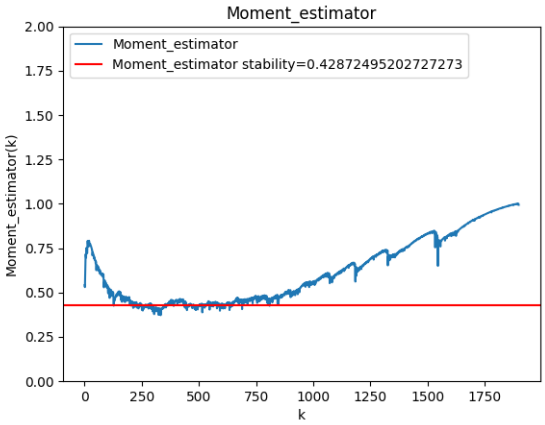

Рис. 6: Moment estimator plot

D. Оценки индекса экстримального значения

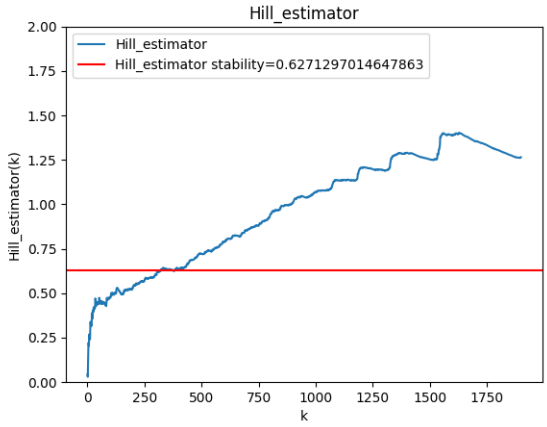

Рис. 4: Hills estimator plot

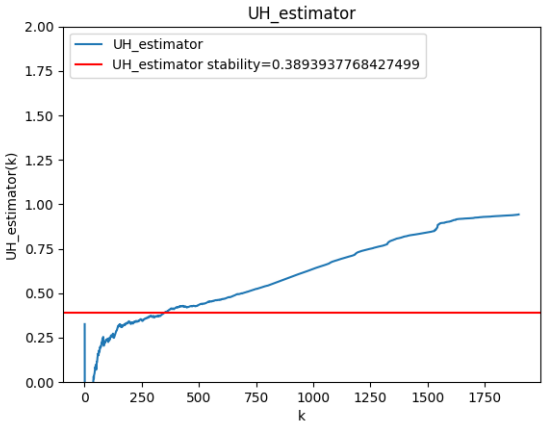

Рис. 7: UH estimator plot

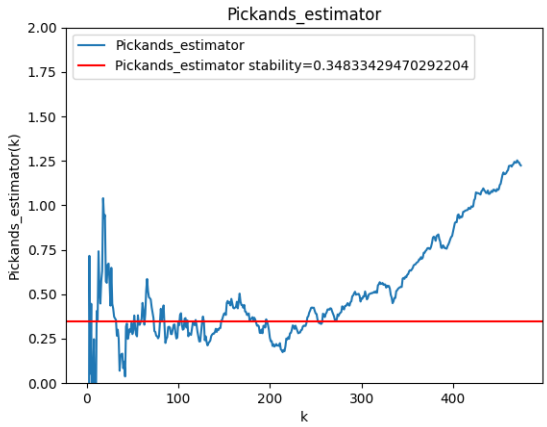

Рис. 8: Pickands estimator plot

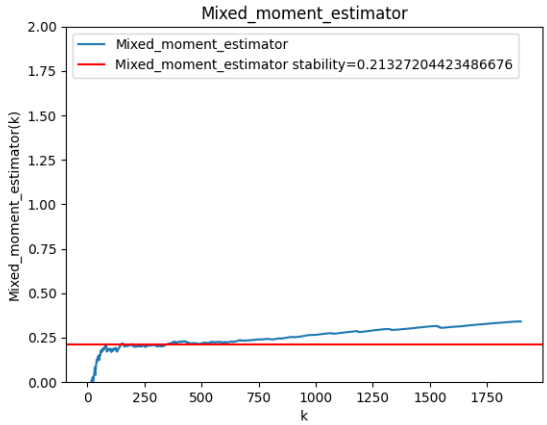

Рис. 9: Mixed estimator plot

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. При взятии среднего по этим оценкам по интервалу стабильности получаются значения 0.42, 0.34 и 0.21 соответственно. Для дальнейших вычислений хвостового индекса будем использовать Mixed Moment estimator в связи с относительной стабильностью оценки на всей последовательности.

Вывод: по всей видимости хвостовой индекс можно положить примерно равным $\gamma \approx 0.21$.

IV. РАЗБИЕНИЕ НА СООБЩЕСТВА

Библиотека NetworkX позволяет разделить граф на сообщества посредствам Louvain Community Detection Algorithm. Выявлено 13 сообществ разного размера.

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 I: Таблица количества вершин в сообществах графа

V. ТЕСТИРОВАНИЕ СТАЦИОНАРНОСТИ

Тестирование стационарности ссобществ производится методом сведения к тестированию случайных последовательностей, возникающих на графе. Для этого рассматриваются состояния динамического графа на выбранных временных отрезках. Для каждого узла рассчитывается его PageRank. Случайным образом выбираются вершины графа для формирования последовательности с фиксированной шириной окна. После чего производится оценка индекса экстримального значения методом смешанных моментов для элементов последовательности. По итогам рассчётов получены хвостовые индексы случайных последовательностей во времени. На рис. 10 представлена зависимость хвостового индекса динамического графа во времени.

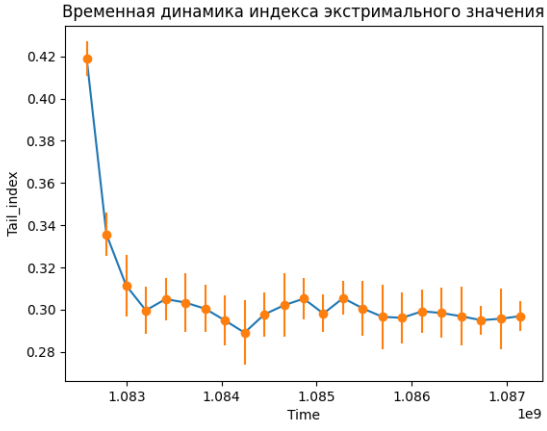

Рис. 10: Временная динамика хвостового индекса графа

На рис. 11 представлена зацвисимость числа вершин графа от времени. Шкала OY логарифмическая.

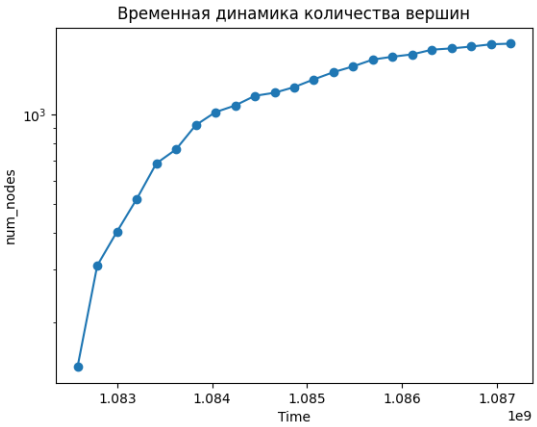

Рис. 11: Временная динамика числа вершин графа

Исходя из графиков видно, что в первые моменты времени граф нестационарен, так как происходит быстрое увеличение числа узлов и рёбер графа. Тогда как в дальнейшем хвостовой индекс находится в определённом диапозоне значений и меняется слабо.

VI. ВЫВОДЫ

К настоящемму моменту времени проведено исследование индекса экстримального значения динамического графа. Получен промежуток нестационарности графа, приведены возможные причины явления. В дальнейшем будут проведены исследования разных методов подбора последовательности, такие как алгоритм Метрополиса — Гастингса, будет изучены методы разбиения графа на сообщества.

VII. ЛИТЕРАТУРА

1) https://networkrepository.com 2)