

# Structural analysis of small streamer's chatters network


# Структурный анализ сети чаттерсов небольшого стримера

# Выбор сети для анализа

- Отсутствие опыта в работе с VK-API, но наличие такового для Twitch API
- Небольшое количество друзей в вк
- Собственноручно собранный датасет из 96245 сообщений в чате твича
- Личный интерес в структуре общения в чате

# Принцип построения сети

- Users  $\rightarrow$  nodes
- Дуга  $(i,j)$  образуется, когда юзер  $i$  “тэгает” юзера  $j$
- Вес дуги  $(i,j)$  равен количеству раз, когда юзер  $i$  “тэгнул” юзера  $j$
- Стример удаляется из сети (так как это по сути эго-сеть стримера)

 **Juliam7:** @xseix я никому ничего плохого не делаю, только доброту и отдаю безвозмездно

соответственно...

# Тип графа

- Ориентированный
- Гомогенный
- Взвешенный

У каждой вершины имеется атрибут 'label' - имя пользователя и атрибут 'size', равный количеству отправленных сообщений в чат

Атрибут рёбер 'weight' описан на предыдущем слайде

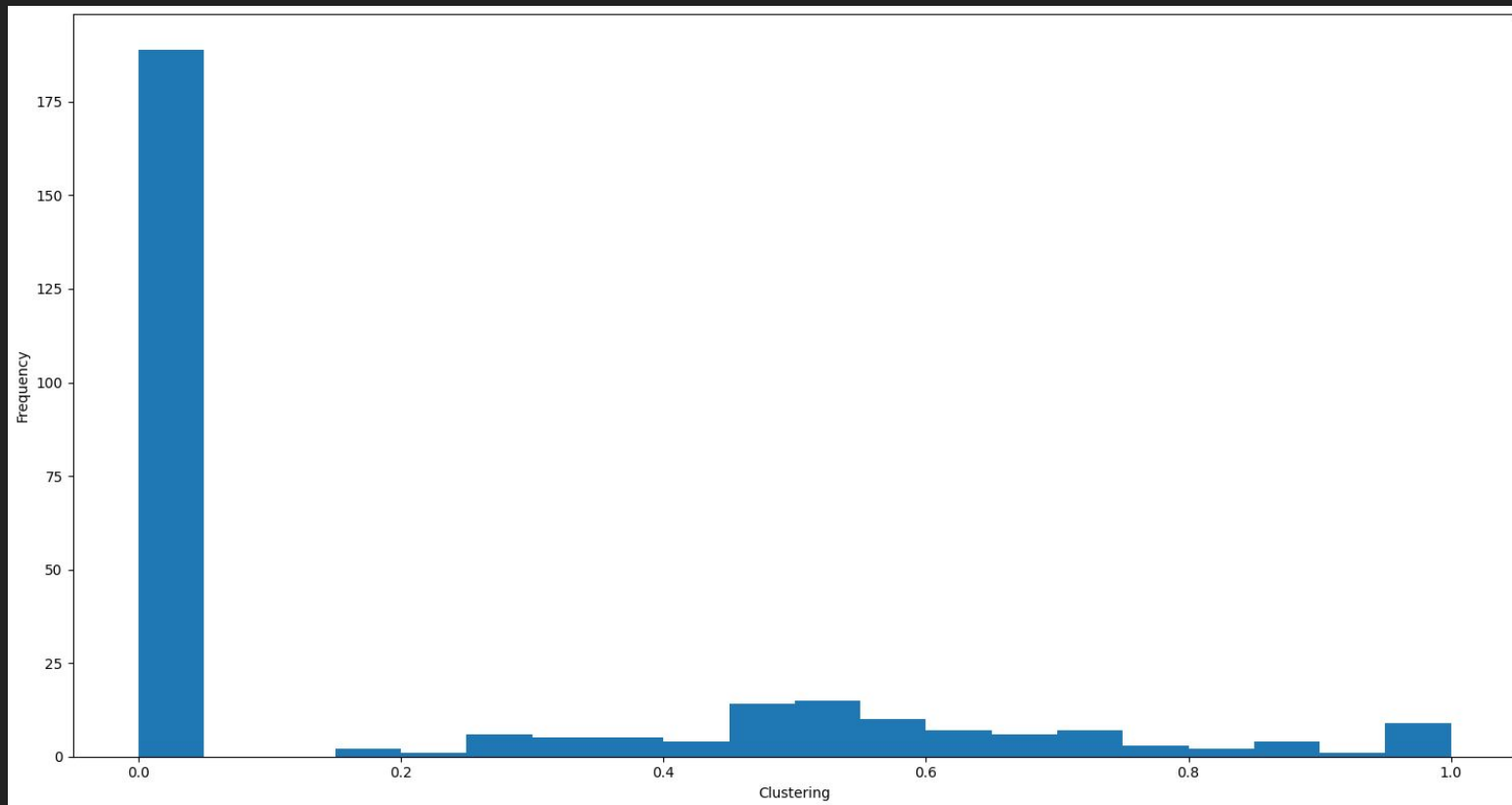
# Network summary

- Размер графа: 290 вершин, 1361 дуга
  - Размер слабосвязанной компоненты: 156 вершин, 1361 дуга
  - Размер сильносвязанной компоненты: 104 вершины, 1292 дуги
- Диаметр сильносвязанной компоненты: 5
- Радиус сильносвязанной компоненты: 3
- Средняя длина кратчайшего пути: 2.3
  - С учётом весов: 3.2
- Коэффициент кластеризации графа: 0.2
  - Коэффициент кластеризации слабосвязанной компоненты: 0.4
  - Коэффициент кластеризации сильносвязанной компоненты: 0.5

# Промежуточные выводы

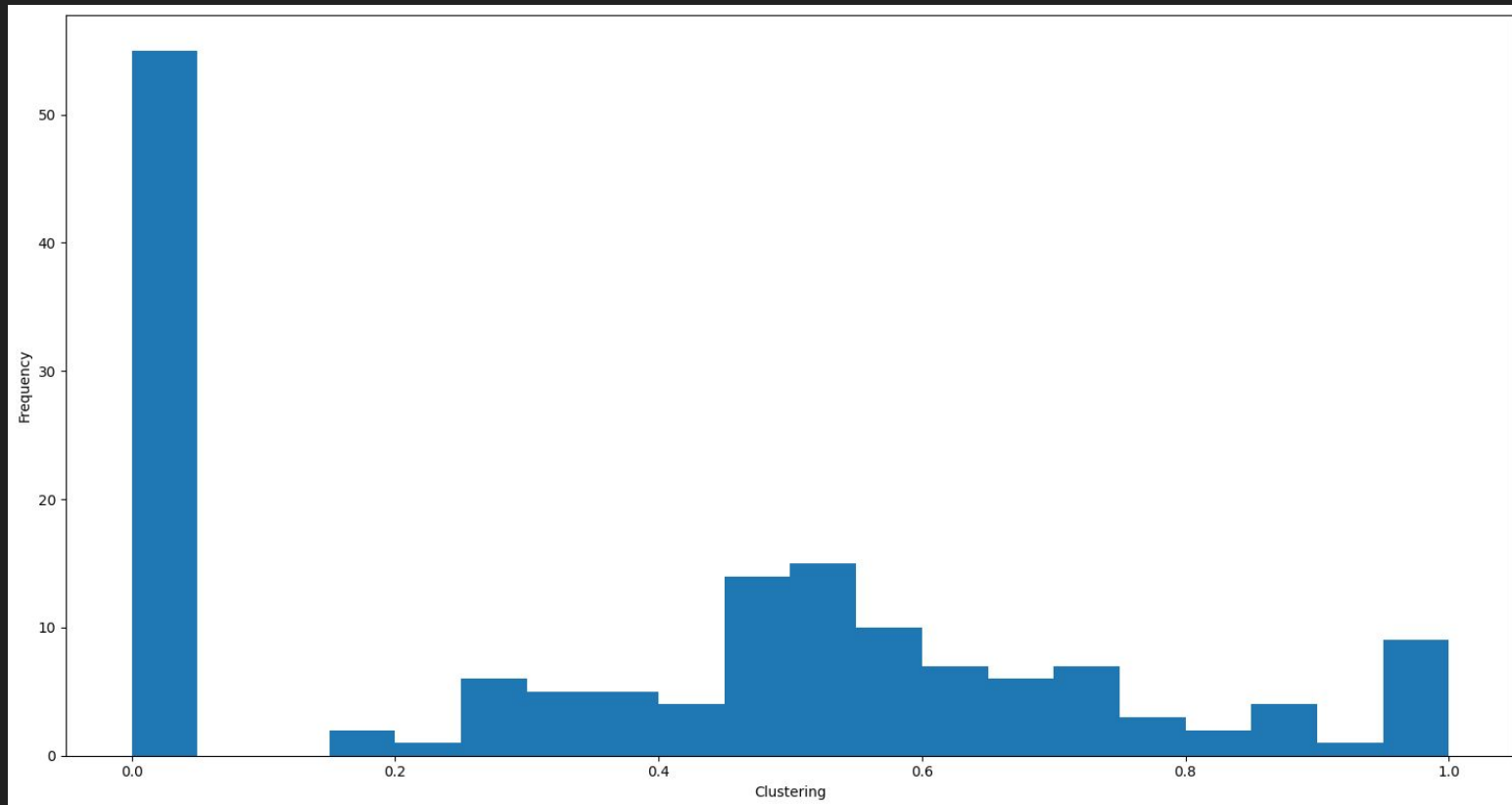
- Граф состоит из одной слабосвязанной компоненты и изолированных вершин - юзеров, которые не тэгают других. Все остальные так или иначе друг друга тэгают
- Средняя длина кратчайшего пути достаточно маленькая, следовательно связей много, то есть юзеры много друг друга тэгают
- Поэтому и коэффициент кластеризации компонент связности достаточно высок

# Гистограмма кластеризации

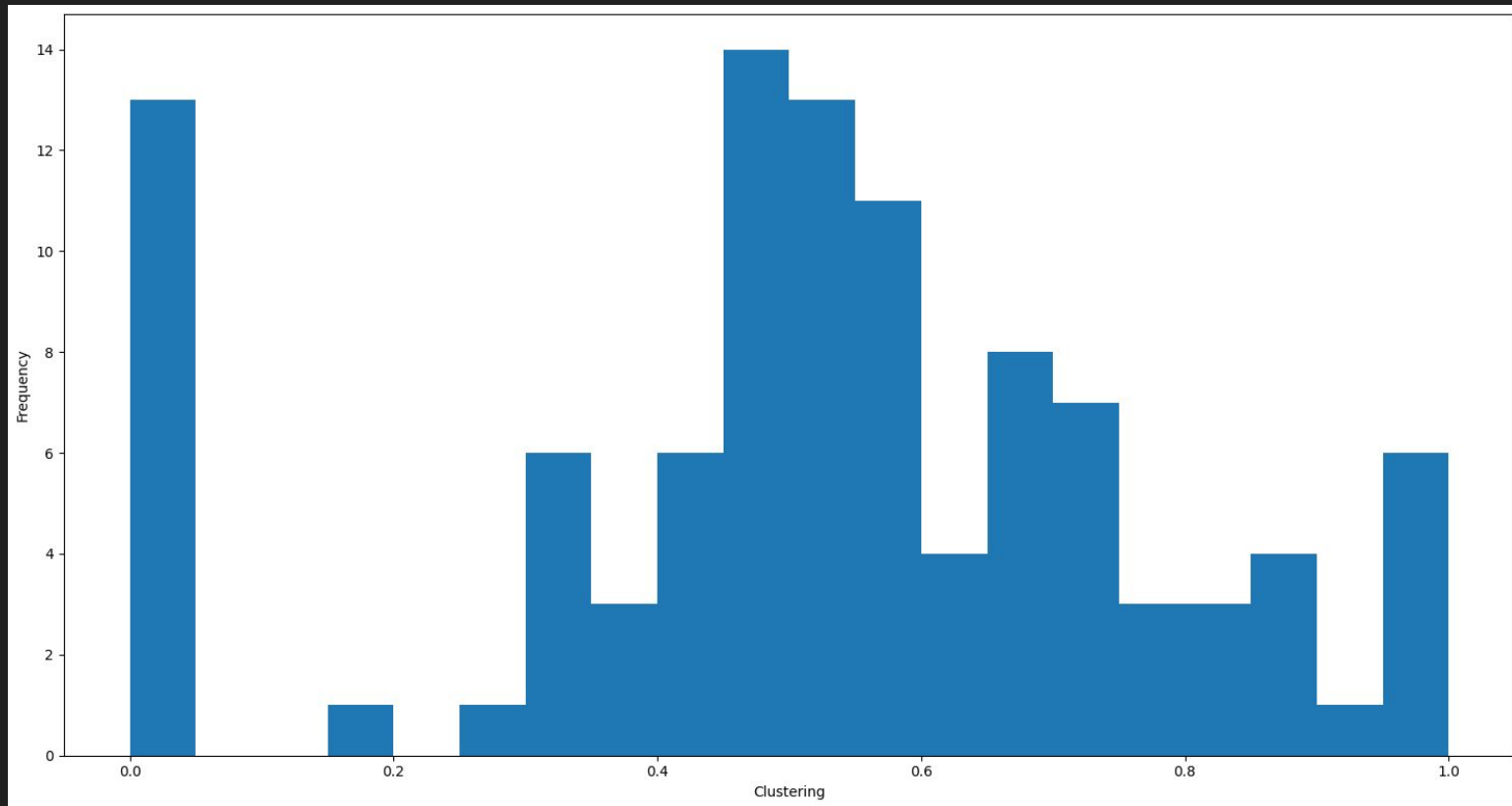




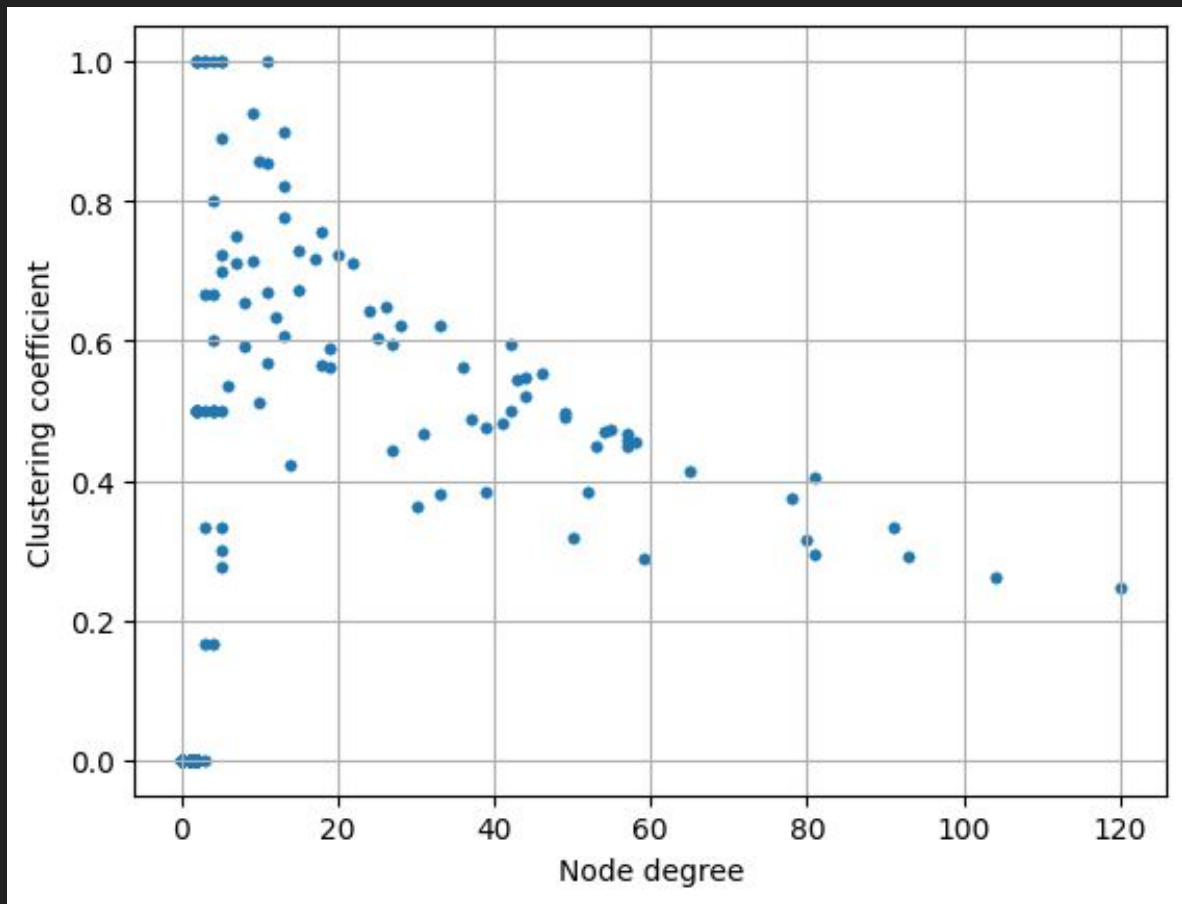
# Гистограмма кластеризации компоненты слабой связности



# Гистограмма кластеризации компоненты сильной связности



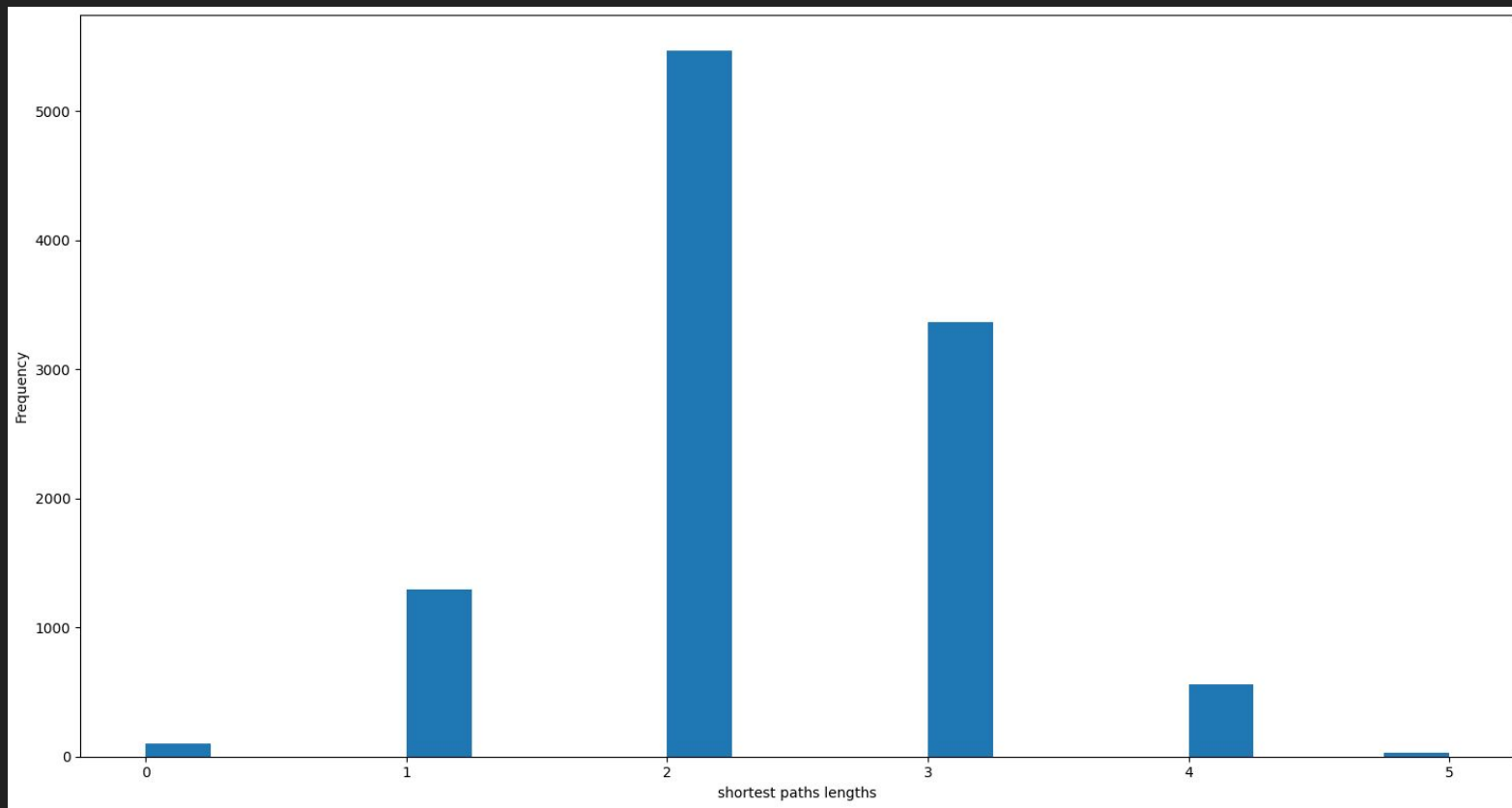
# Зависимость коэффициента кластеризации от степени вершины



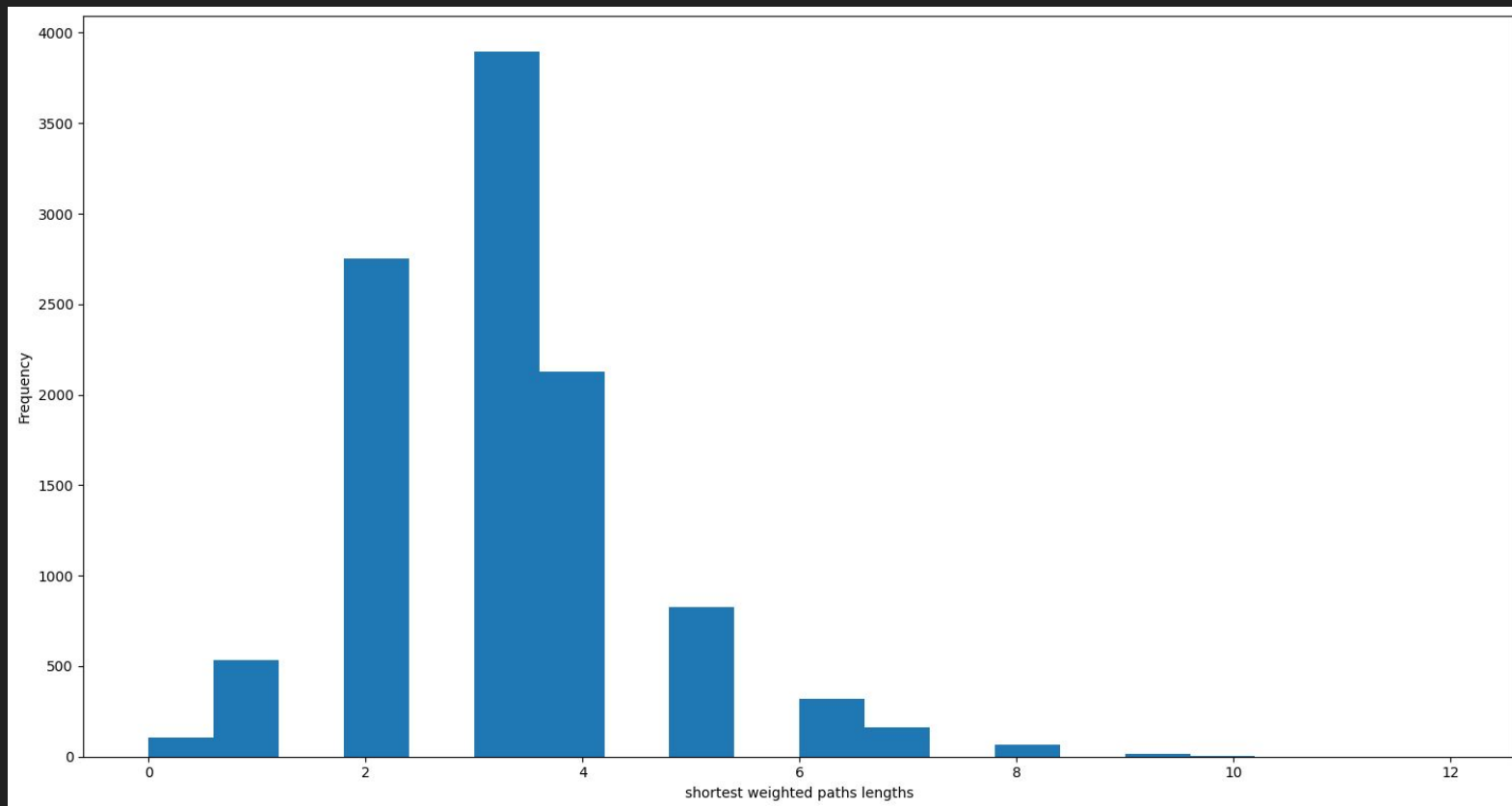
# Промежуточные выводы

- В графе много вершин, почти не связанных с основными вершинами, соответственно в них низкий коэффициент кластеризации - юзеры, не слишком часто пишущие в чат
- С ростом степени вершины слегка уменьшается средний коэффициент кластеризации

# Гистограмма длин кратчайшего пути



# Гистограмма взвешенных длин кратчайшего пути



# Промежуточные выводы

- В графе много связей, соответственно длина кратчайшего пути небольшая, причём у каждой вершины есть связь с другой небольшого веса, поэтому даже с учётом весов длина кратчайшего пути остаётся небольшой - значит чаттеры тэгают только тех, кого тэгают часто

# Визуализация

- Размер вершины зависит от количества сообщений пользователя
- Размер имени зависит от степени вершины
- Толщина связи зависит от веса
- Цвет вершины: чем краснее, тем больше исходящих тэгов, чем зеленее, тем больше входящих тэгов

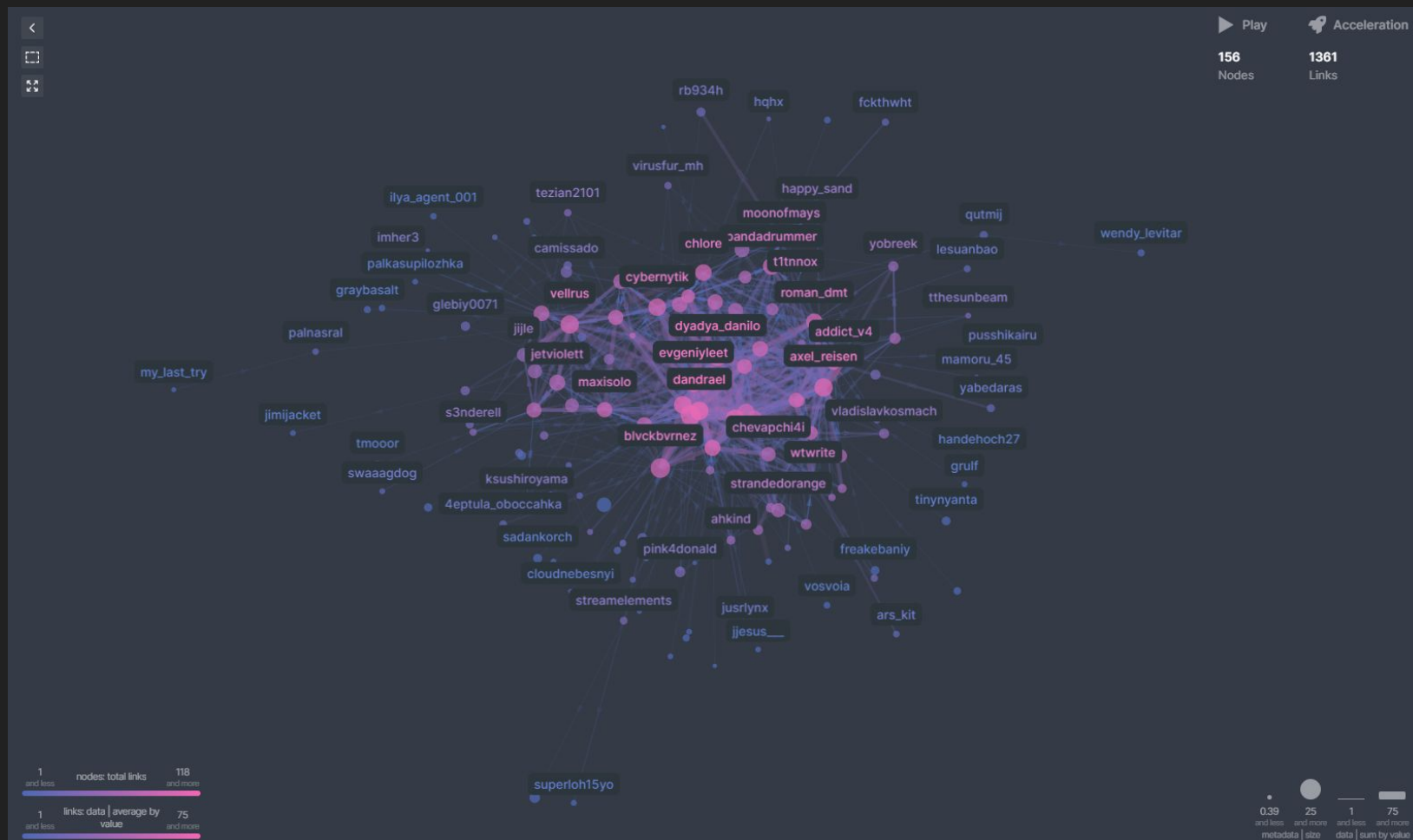




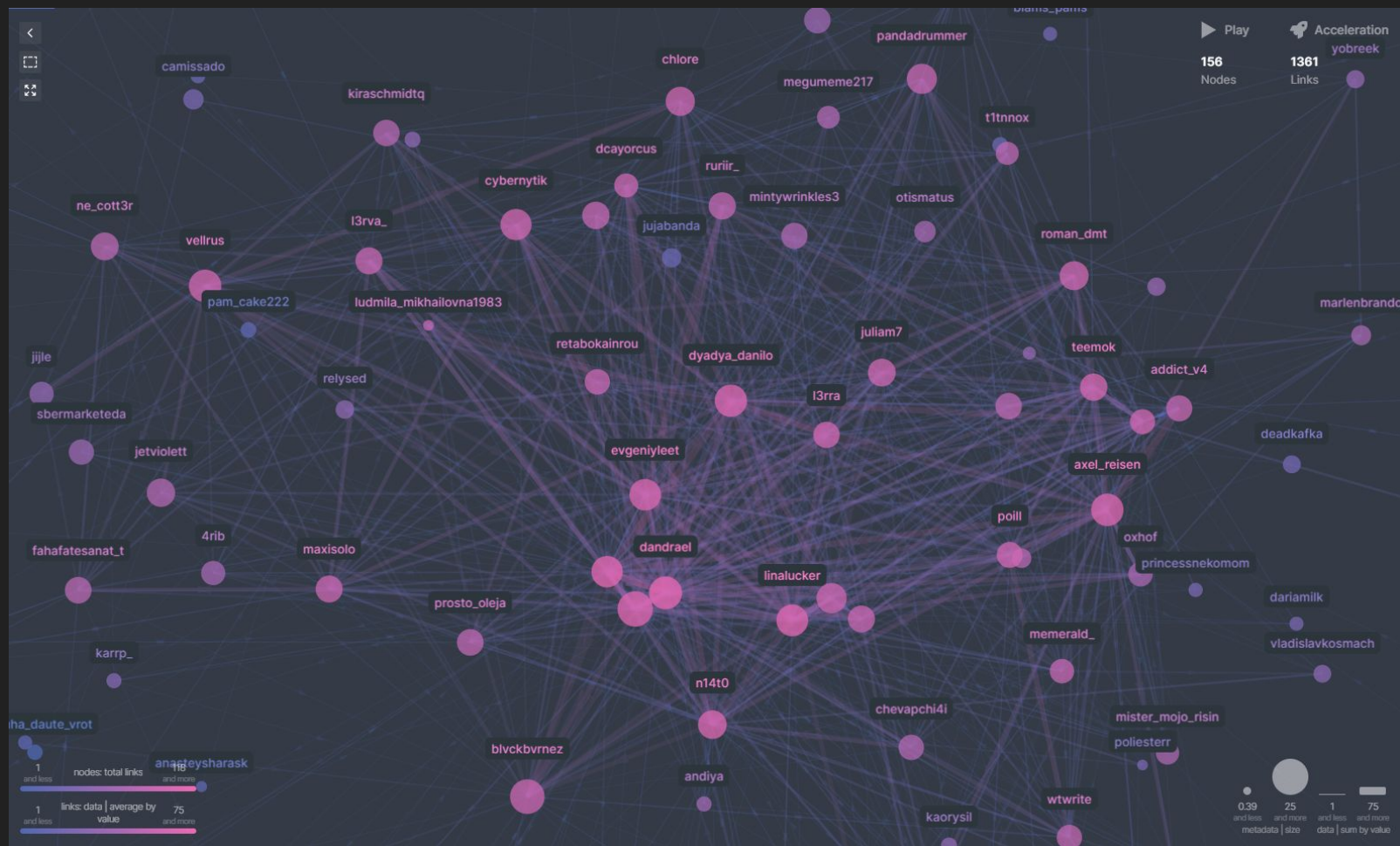
# Визуализация

- Размер вершины зависит от количества сообщений пользователя
- Толщина связи зависит от веса
- Цвет вершины: чем краснее, тем больше суммарная степень

# Визуализация компоненты слабой связности



# И центральная часть сети



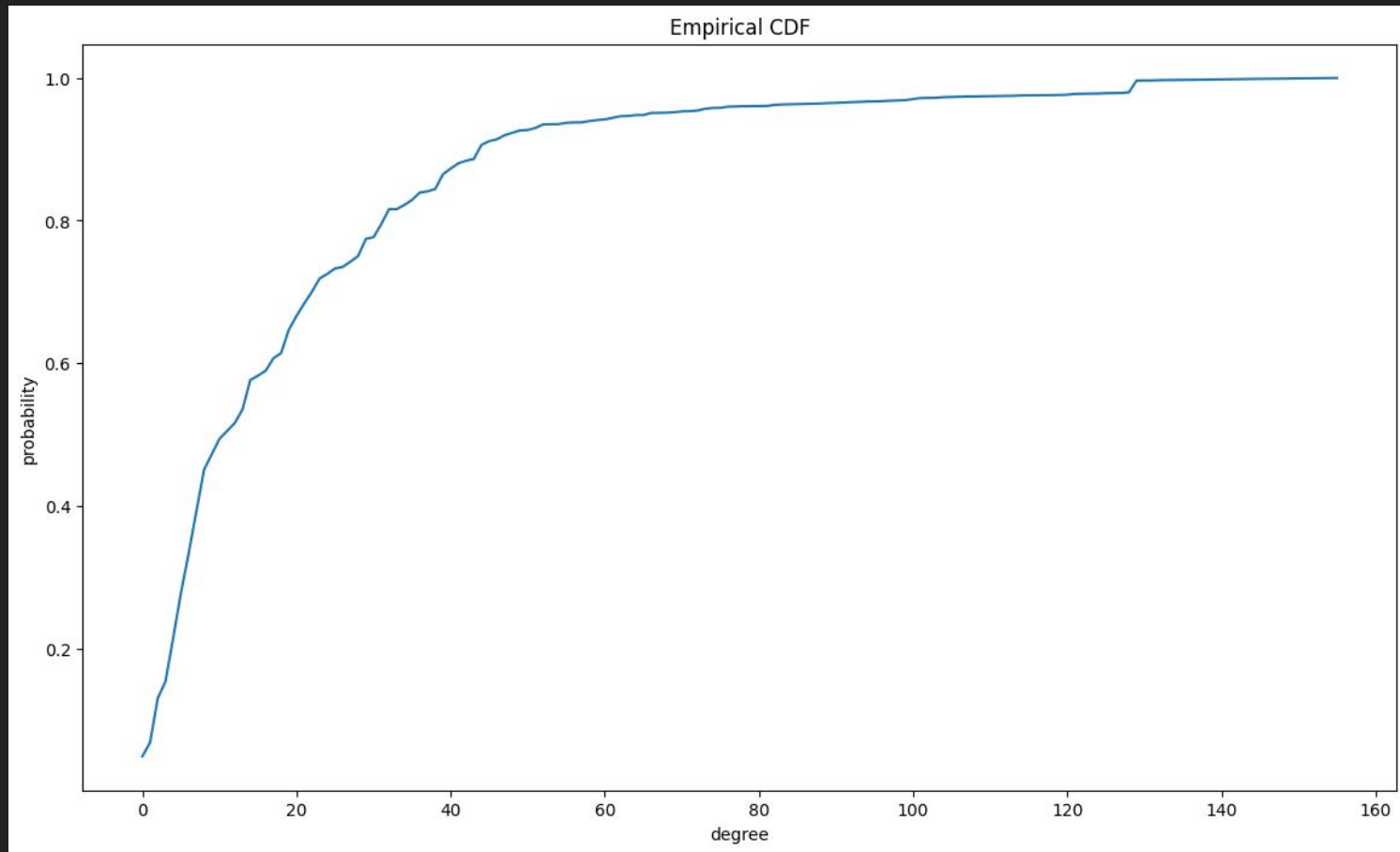
# Промежуточные выводы

- В графе выделяется “основная часть” - юзеры, которые много пишут в чат и часто друг друга тэгают
- Сложно заметить какую-то структуру, как не изображай граф, следовательно это не самая наглядная для анализа сеть, и автор презентации получит очень мало баллов за проект



Juliam7: надо что-то чтобы смеяться, а то я только плачу

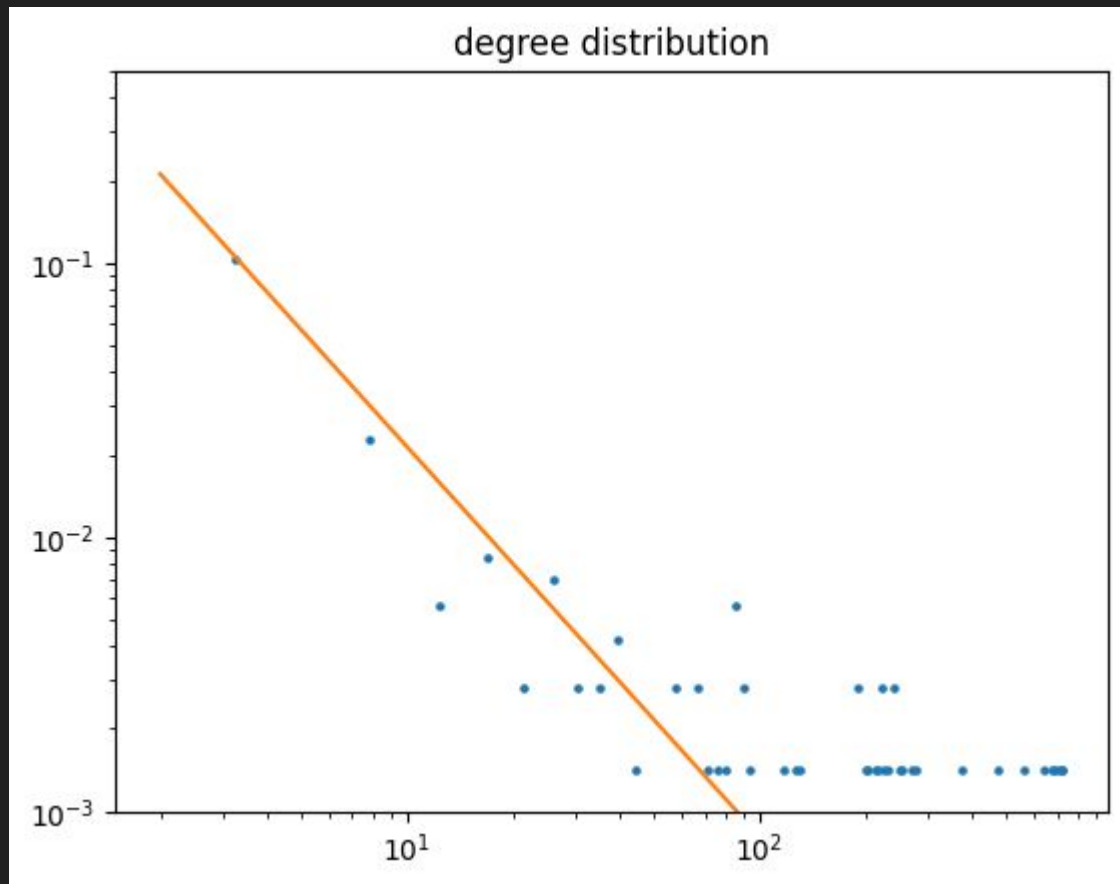
# Эмпирическая функция распределения степени вершин



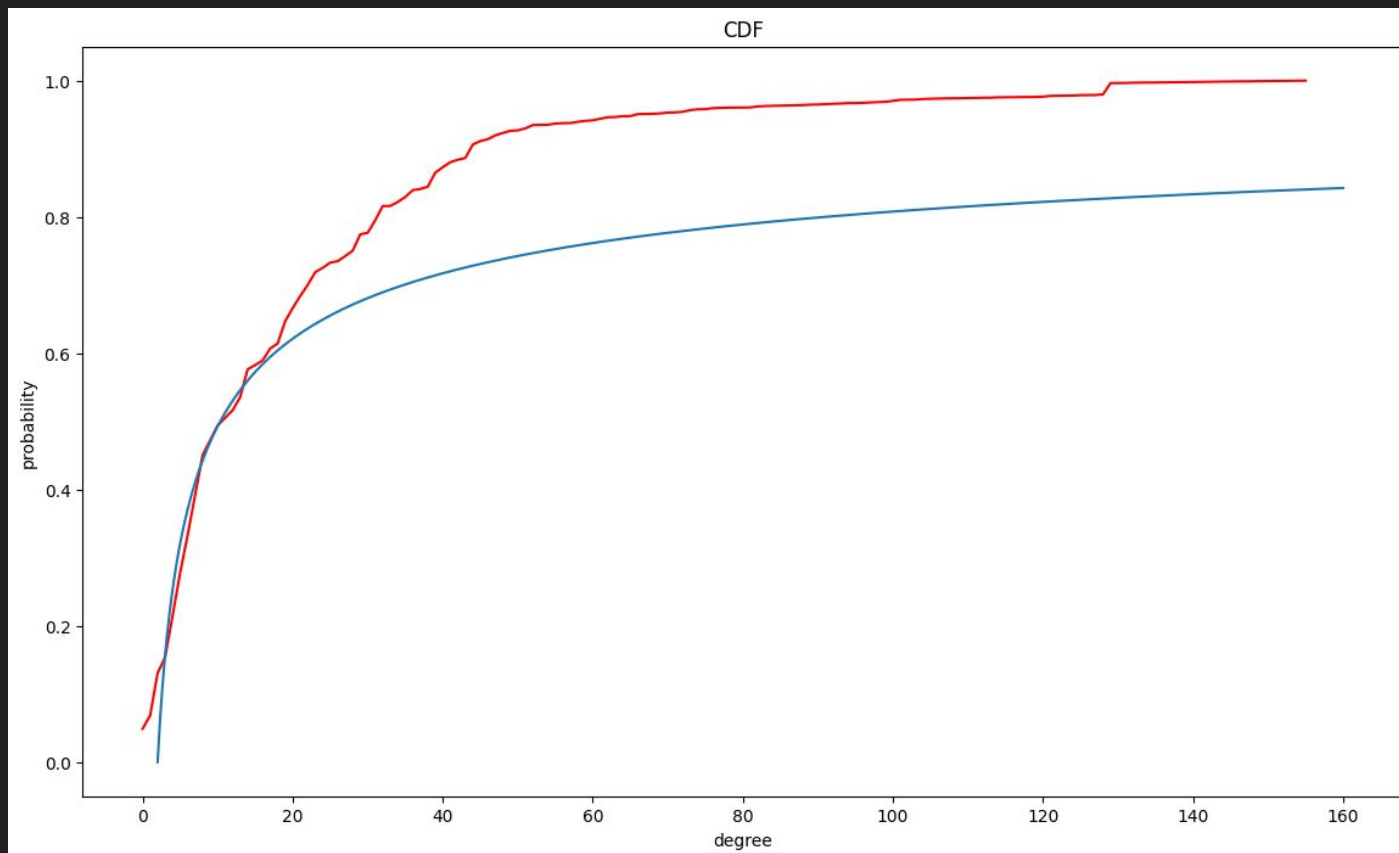
# Оценка максимального правдоподобия на параметры Power Law

$$\alpha = 1.42$$
$$x_{min} = 2$$

Не особо соответствует  
Power Law...



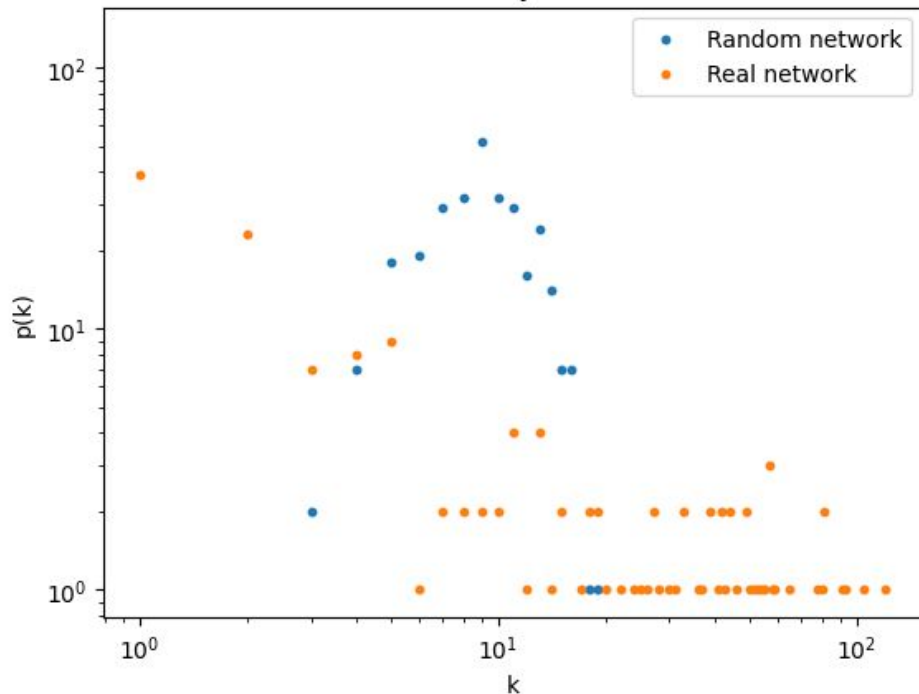
# Эмпирическая и теоретическая функции распределения степеней вершин



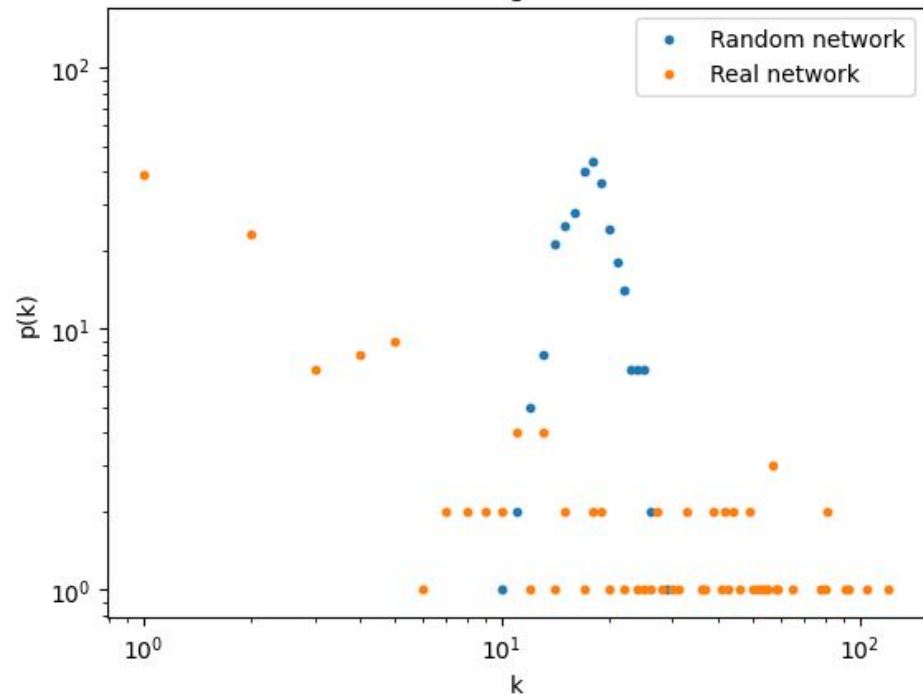


# Модели случайного графа

erdos-renyi model



watts-strogatz model



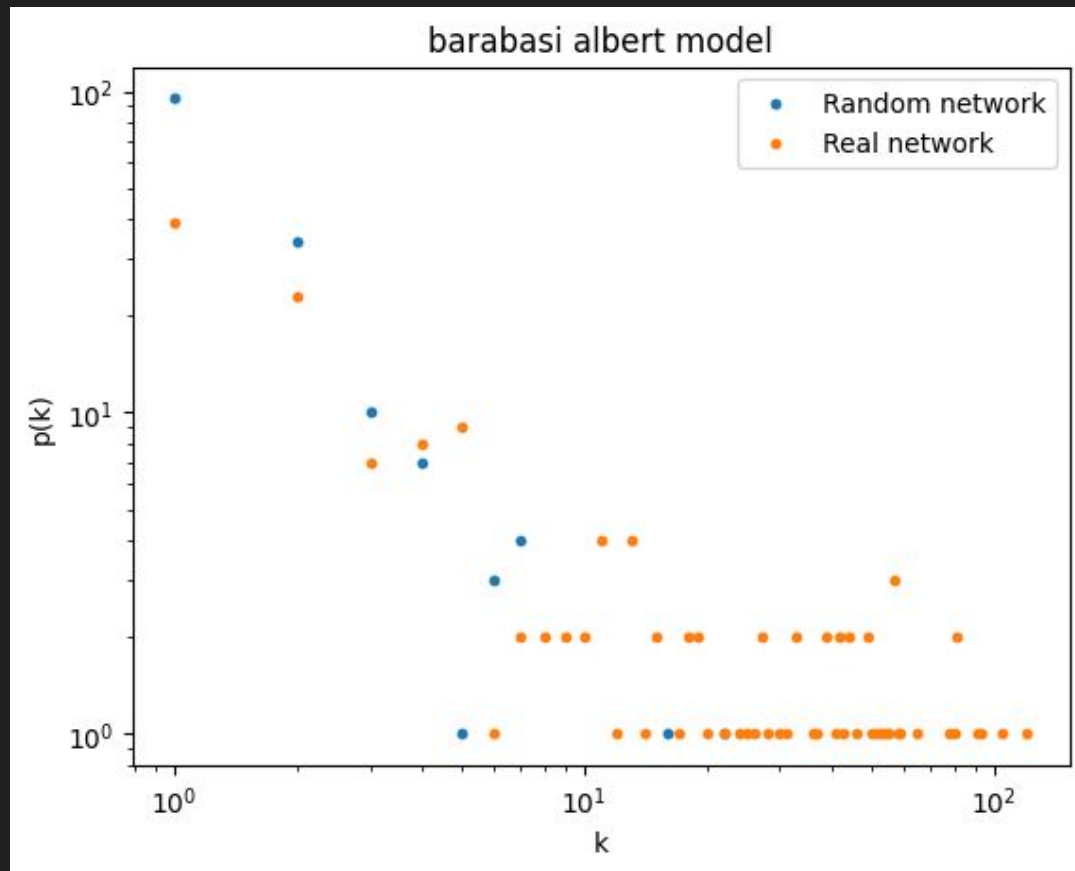
# Модели случайного графа

Радиус: 3 (3)

Диаметр: 5 (6)

Средний кратчайший путь: 2.8 (2.5)

Средний коэффициент  
кластеризации: 0.08 (0.44)



# Промежуточные выводы

- Распределение степеней вершин не особо соответствует Power Law
- Лучше всего на граф похожа модель Барабаши-Альберт, следовательно можно смоделировать эволюцию графа по этой модели, то есть новый зритель, заходя в чат, скорее всего будет тэгнут кем-то, кто больше всего пишет в чат и чаще тегают, и скорее всего тегнет его же

# Топ 10 пользователей по центральностям

## Degree centrality

	username	centrality
0	dandrael	0.415225
1	justreyb	0.359862
2	ceddisss	0.321799
3	linalucker	0.314879
4	dyadya_danilo	0.280277
5	n14t0	0.280277
6	axel_reisen	0.276817
7	evgeniyleet	0.269896
8	teemok	0.224913
9	vellrus	0.204152

## Closeness centrality

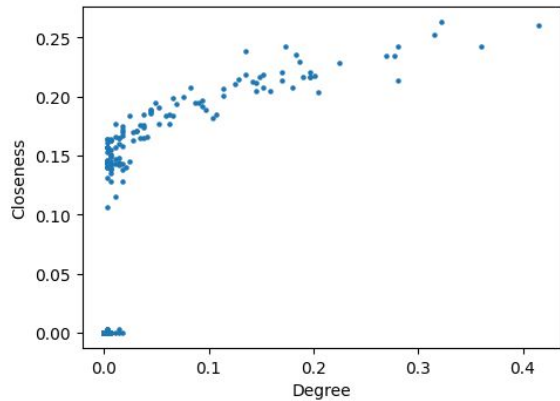
	username	centrality
0	ceddisss	0.263387
1	dandrael	0.260805
2	linalucker	0.252152
3	blvckbvnez	0.242941
4	dyadya_danilo	0.242941
5	justreyb	0.242941
6	ludmila_mikhailovna1983	0.238584
7	juliam7	0.235417
8	axel_reisen	0.234380
9	evgeniyleet	0.234380

## Betweenness centrality

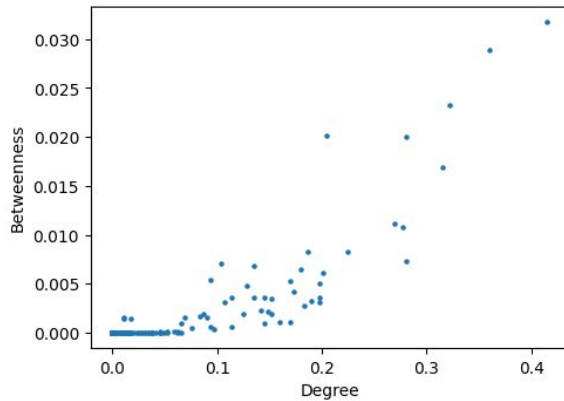
	username	centrality
0	dandrael	0.031758
1	justreyb	0.028898
2	ceddisss	0.023261
3	vellrus	0.020094
4	n14t0	0.020040
5	linalucker	0.016898
6	evgeniyleet	0.011141
7	axel_reisen	0.010768
8	roman_dmt	0.008331
9	teemok	0.008269

# Корреляции

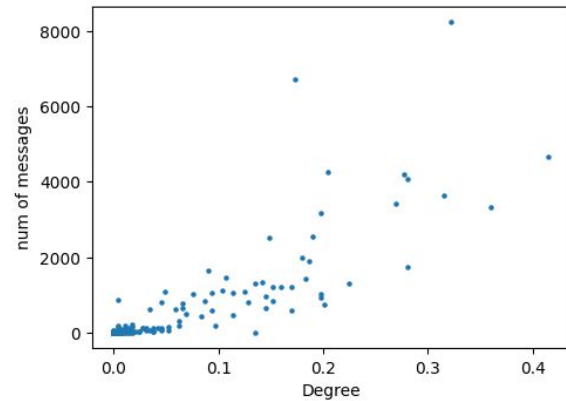
Correlation: 0.6801



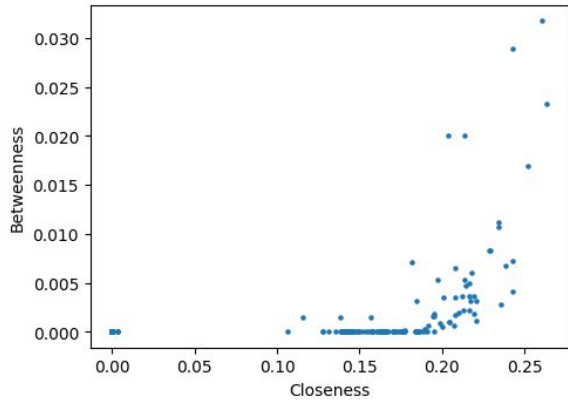
Correlation: 0.8477



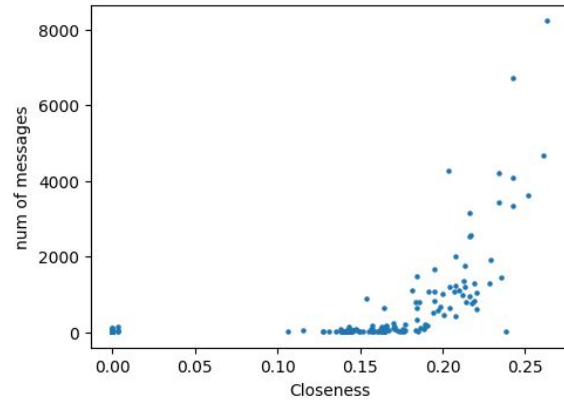
Correlation: 0.8425



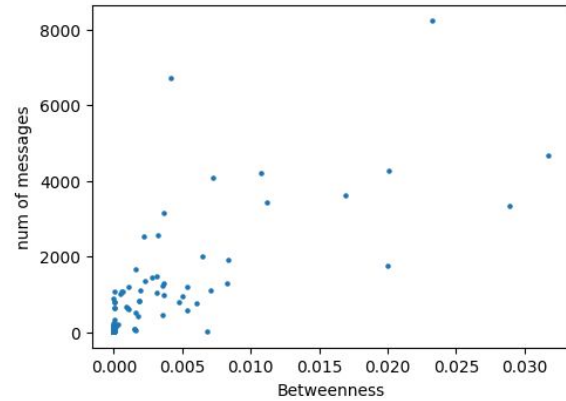
Correlation: 0.4425



Correlation: 0.5237



Correlation: 0.7819

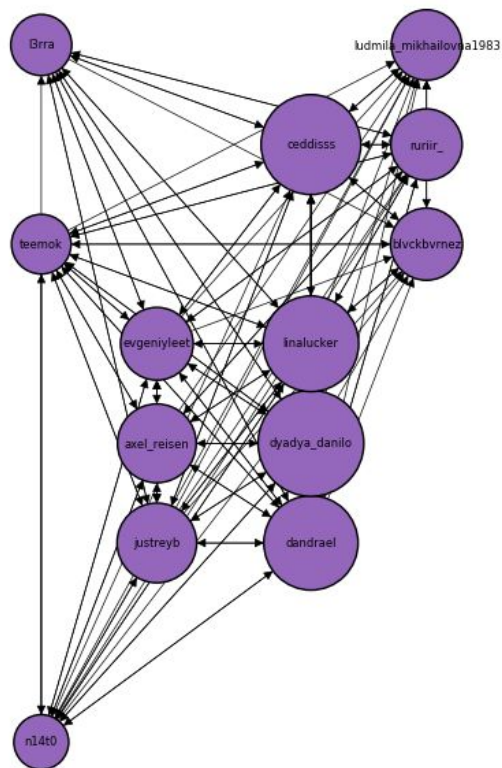


# Промежуточные выводы

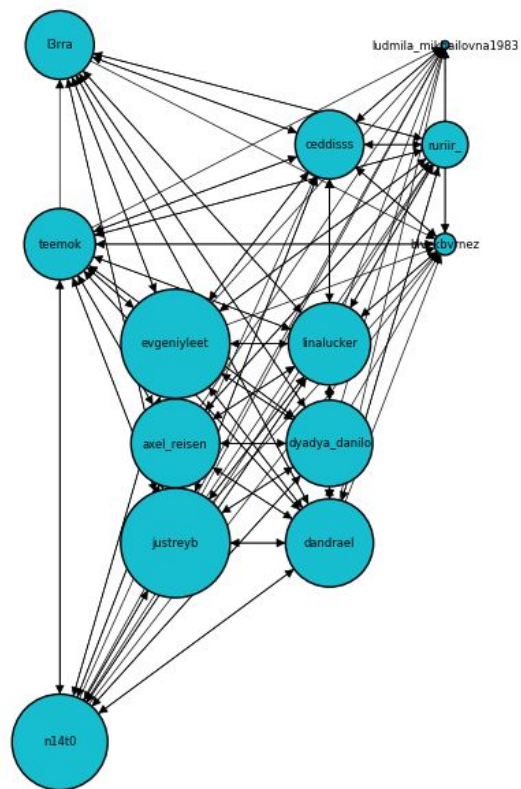
- Все топы содержат примерно одних и тех же людей - которые больше всего пишут в чат (а значит и чаще тэгают других), именно такой вывод можно сделать из корреляций между центральностями и количеством сообщений
- Меньше всего между собой коррелируют степени близости и посредничества, что может свидетельствовать о наличии юзеров, которые либо в основном тэгают других, либо их самих тэгают, следовательно у них повышается степень близости, но падает степень посредничества, наличие таковых проверим далее...

# HITS

Node size by authority score



Node size by hub score

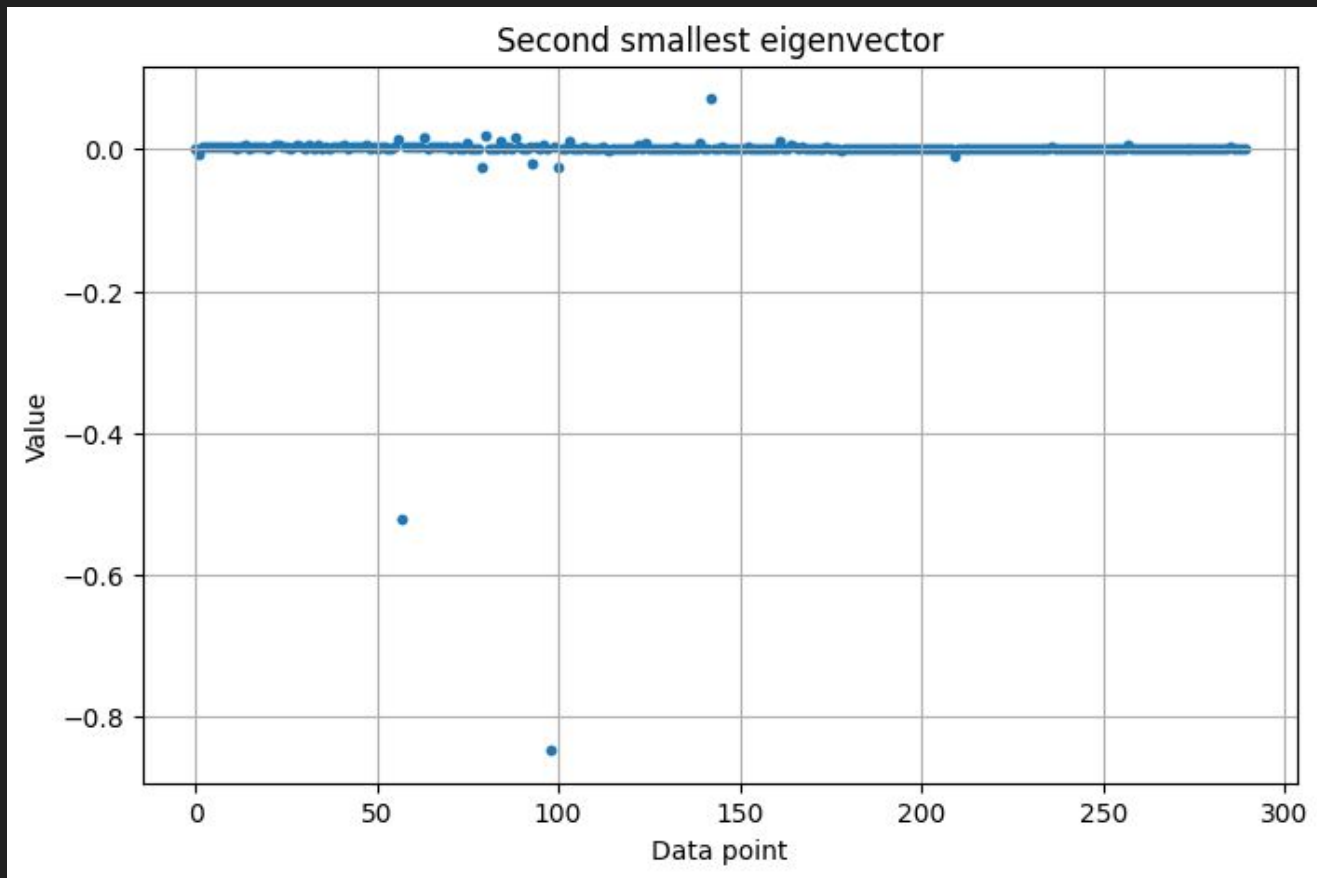


# Промежуточные выводы

- Самые авторитетные авторитеты зачастую являются также и одними из самых хабных хабов и наоборот
- Есть юзеры, чей авторитетный скор сильно превосходит хабный, но юзеры с высоким хаб скором обычно обладают и неплохим авторитет скором



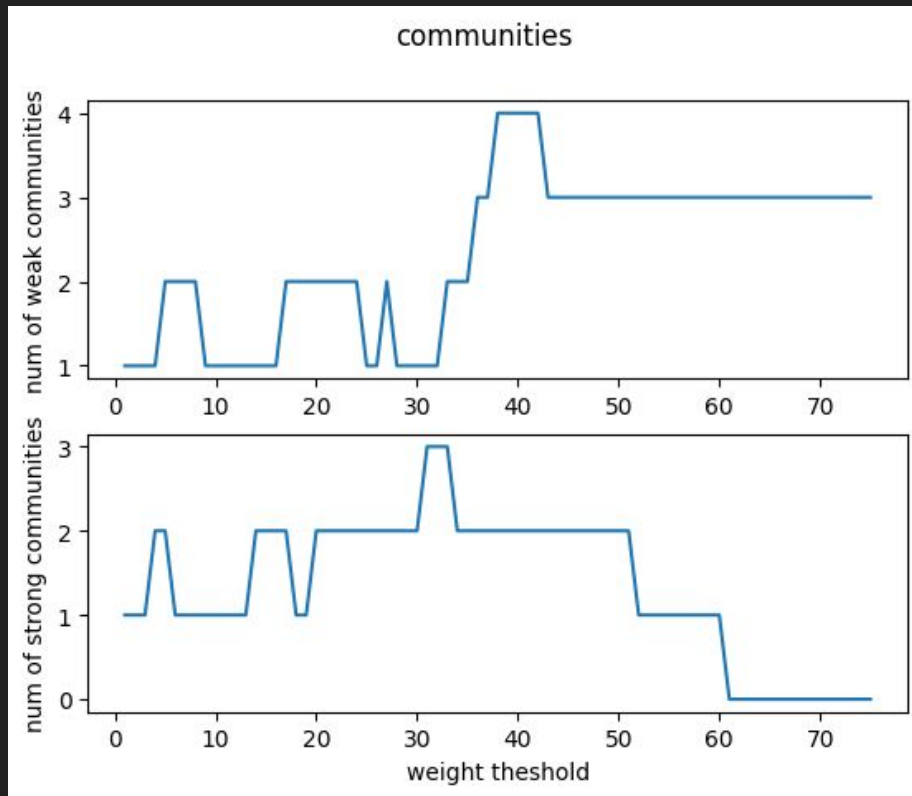
К сожалению спектральный анализ бесполезен...



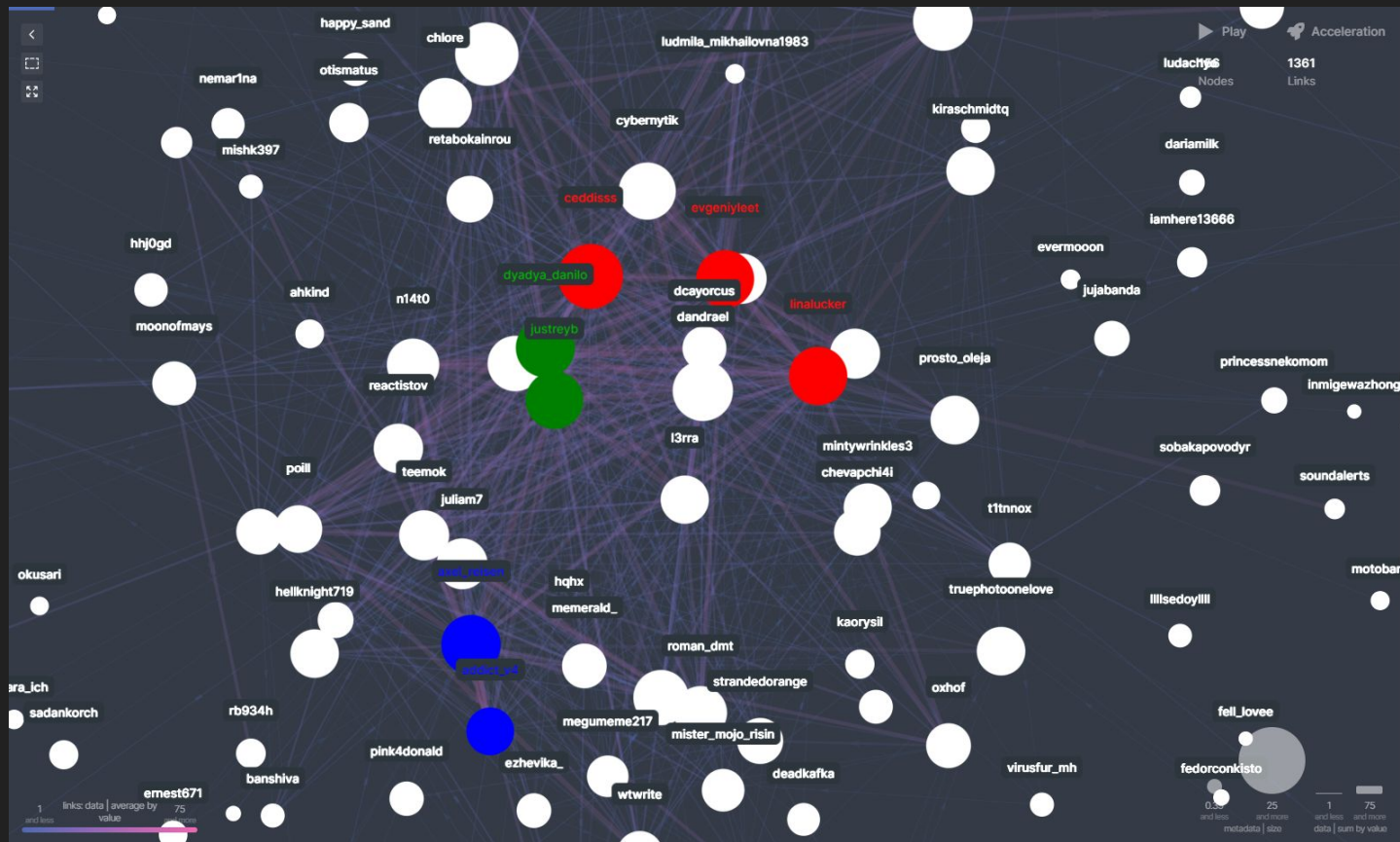
# Схема с весами для выявления сообществ

- Устанавливаем порог и удаляем рёбра, веса которого ниже порогового значения
- Повышаем порог, пока не возникнет наибольшее число компонент сильной связности с числом вершин больше одной, получаем некоторые “ядра” сообществ
- Вершины, не вошедшие в компоненты связности, последовательно добавляем в сообщество, с которым больше связь (суммарный нормированный вес рёбер)

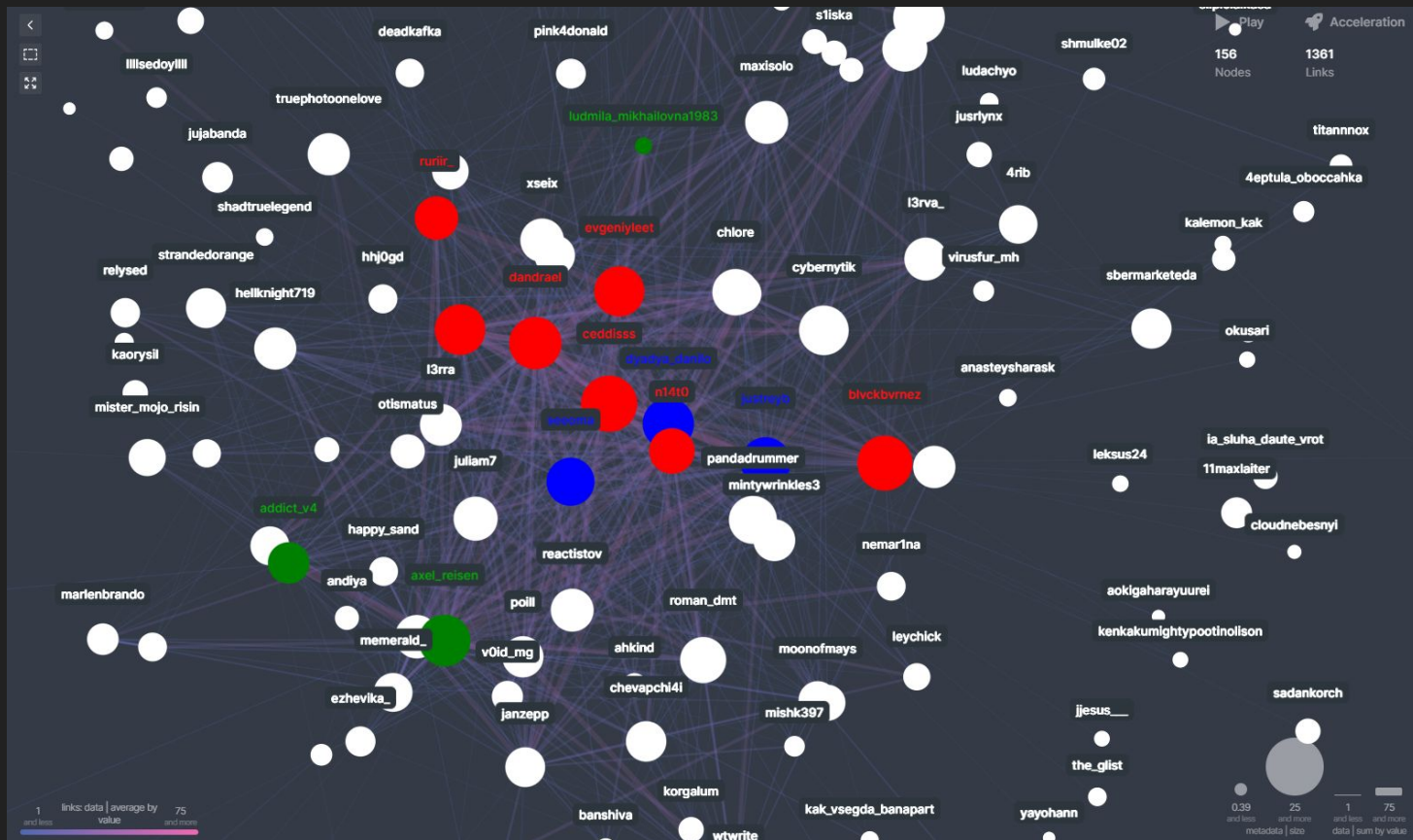
# Зависимость числа компонент связности от порога



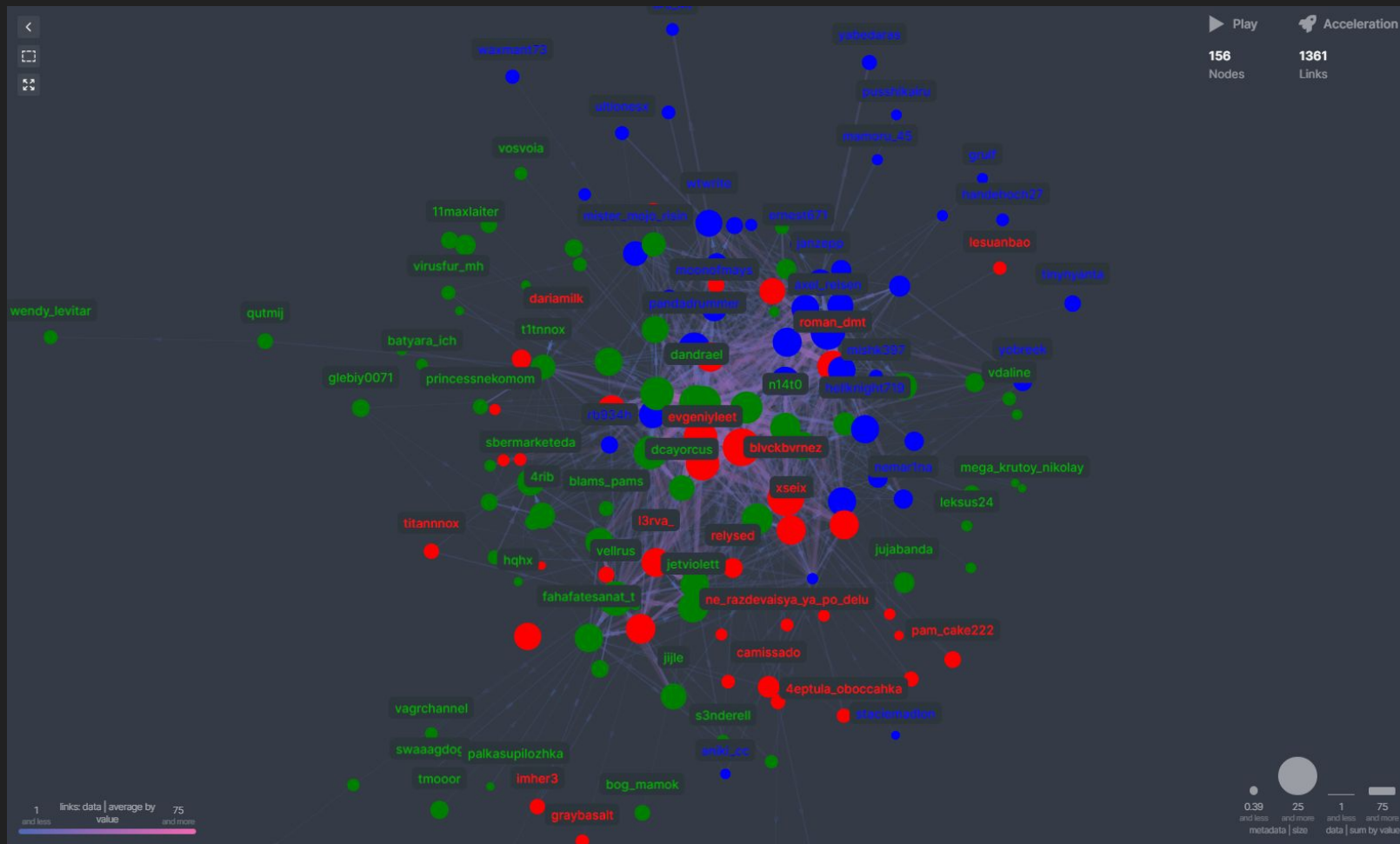
# Ядра сообществ (компоненты сильной связности)



# Ядра сообществ (компоненты слабой связности)



# Сообщества



# Сообщества

- 42 красные вершины
- 72 зелёные вершины
- 42 синие вершины

# Промежуточные выводы

- Выделилось три основные группы пользователей. По опыту пребывания в чате можно сказать, что:
  - Синяя группа состоит из тех, кто общается в основном со стримером на конкретную тему
  - Красная группа - из тех, кто в основном общается между собой на отвлечённые темы
  - Зелёная группа - что-то посередине

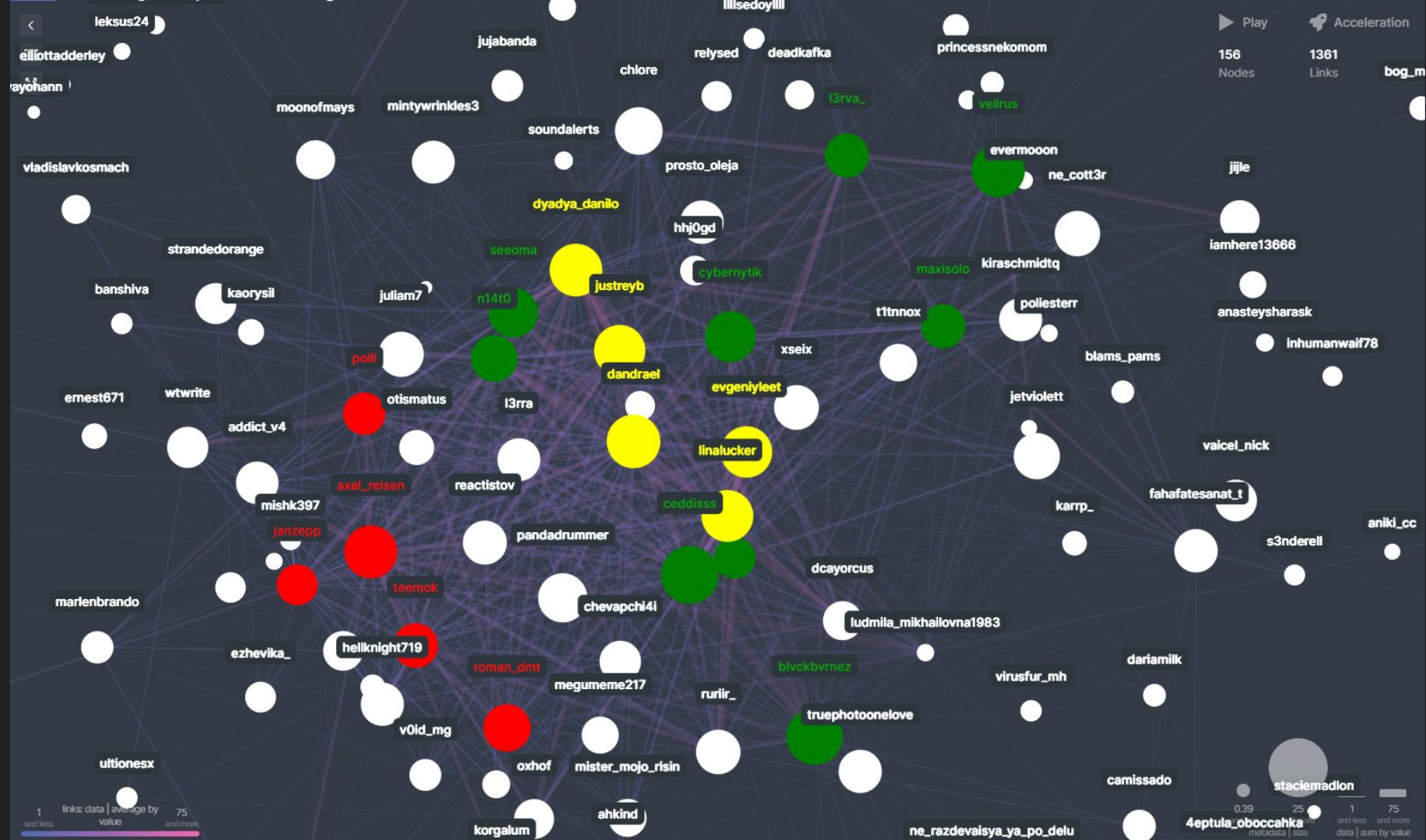


# Максимальные клики

- Слабая клика: размер 14, на визуализации зелёный цвет
- Сильная клика: размер 10, на визуализации красный цвет
- Вершины, вошедшие в обе клики имеют жёлтый цвет

\*Слабая клика - подграф, между любыми двумя вершинами которого есть дуга

\*Сильная клика - подграф, между двумя вершинами которого есть дуги в обе стороны



# Выводы

- Чаще друг друга тэгают те, кто больше всего пишет в чат, они же самые важные “авторитеты” и “хабы”, и именно эти люди являются самыми центральными
- Сеть тэгов в чате твича формируется похожим на модель Барабаши-Альберта образом
- В чате не очень популярного стримера все друг друга знают и часто друг друга тэгают
- Выделилось три основных сообщества чаттерсов, но эти сообщества не ярко-выраженные, они предположительно формируются темами, на которые люди общаются в чате

# Простите за

44 слайда....