Домашнее задание 1 по спецкурсу "Большие графы и модели сложных сетей"

Сюй Минчуань, группа 517, факультет ВМК МГУ

1 апреля 2022 г.

1 Теоретическая часть

1. Приведите пример последовательности плотных (неразреженных) графов, степени вершин которых не подчиняются степенному закону. Приведите пример последовательности разреженных графов, степени вершин которых не подчиняются степенному закону.

Решение. Пусть $\{G_n = \{K_2, K_3, K_4, K_5\}_{i=1}^n\}$ - последовательность полных графов с растущим числом индекса n. Каждый G_n содержит n четвёрок полных графов (от K_2 до K_5), и между полными графами нет никаих связей ребер. (как отдельные компоненты).
Например, $G_3 = \{K_2, K_3, K_4, K_5, K_2, K_3, K_4, K_5, K_2, K_3, K_4, K_5\}$

Компоненты	**Степень** G_n	**Степень** G'_n	#
K_2	1	1	2
K_3	2	2	3
K_4	3	3	4
K_5	4	4n	5

Таблица 1: компоненты графов задачи 1

Рассмотрим свойства последовательности графов G_n:

- Это **разреженные графы**, так как $|V_n| = (2+3+4+5)*n = 14n, |E_n| = (1+3+6+10)*n = 20n$ при каждом n. Тогда, если положить $m_1 = 1, m_2 = 2$, то выполняется такое неравеноство:

$$m_1|V_n| \leqslant |E_n| \leqslant m_2|V_n|, \quad \forall i \geqslant 1 \tag{1}$$

  что и означает разреженность графов.

- Их степени вершин **не подчиняются степенному закону**, поскольку количество вершин большой степени увеличивает с ростом степени, то есть обладает положительно линейной зависимостью.

Теперь мы немножко переопределим G_n, что в K_5 допустимы кратные ребра, причем кратность каждого ребра равна n (обозначим как K'_5), получим новые графы $\{G'_n = \{K_2, K_3, K_4, K'_5\}_{i=1}^n\}$

- G'_n не будет разреженным, то есть уже является **плотным**. Объясним: $|V'_n| = (2+3+4+5)*n, |E'_n| = (1+3+6+10n)*n = 10n^2+10n$. Таким образом, не найдутся m_1, m_2 такие, что неравенство (1) может выполняться.

- степени вершин G'_n останется **не подчиняться степенному закону**, поскольку добавление кратных ребер просто увеличивает степень вершин исходной $K_{5,5}$, количество вершин такой степени не изменится.

2. Приведите пример последовательности неориентированных графов (кратные рёбра и петли разрешаются), степени вершин которых подчиняются какому-нибудь степенному закону. Обязательно доказательство степенного закона.

Решение. Пусть x - степень вершины, $y(x)$ - число вершин степени x. Предполагается, что подчиняться степенному закону с $\gamma = 1$:

$$y = \frac{2^{10}}{x}$$

x	1	2	4	8	16	32	64	128	256	512
y	1024	512	256	128	64	32	16	8	4	2
кратность	1	2	4	8	16	32	64	128	256	512
# пар	512	256	128	64	32	16	8	4	2	1

Таблица 2: значения функции $y = \frac{2^{10}}{x}$ и компоненты графа G

Будем построить графы, подчиненные такому степенному закону следующим образом:

- $G_n = G$ - то есть граф G_n постоянный при изменении n.
- $G = \{\{g_1\}_{i=1}^{1024}, \{g_2\}_{i=1}^{512}, \ldots, \{g_{2^m}\}_{i=1}^{2^{10-m}}, \ldots, \{g_{512}\}_{i=1}^{2}\}$, $\quad m = 0, 1, 2, ..., 9$. g_{2^m} - это граф с 2 вершинами, соединенными ребром кратности 2^m, а 2^{10-m} - их 2^{10-m} штук.

Таким образом, степени вершин графа G подчиняются степенному закону.

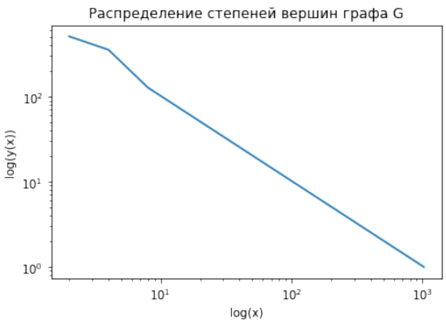


Рис. 1: степенной закон задачи 2

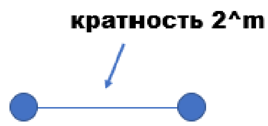
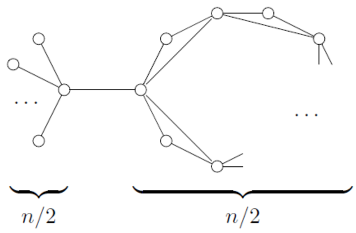


Рис. 2: граф g_{2^m} задачи 2

3. Приведите пример последовательности связных графов, которые не уязвимы к атакам на хабы.

Решение. Пусть $\{G_n\}$ - последовательность полных графов с растущим по времени число индекса n. Графы не уязвимы к атакам на хабы, поскольку какую бы вершину (и соответствующие ребра) удалить из графа, оставшаяся часть - полный граф и сам является гиганской компонентой.

4. Рассмотрим при растущих значениях n, кратных четырём, последовательность графов следующего вида (см. граф). Покажите, что при $n \to +\infty$ средний локальный кластерный коэффициент такого графа можно ограничить снизу положительной константой, а глобальный кластерный коэффициент стремится к 0.



Решение. Сначала перечисляем формулы: пусть $G = (V, E), v \in V$, N_v - множество всех соседей вершины v в G, $n_v = |N_v|$. Будем рассматривать только вершины с $n_v \geqslant 2$.

$$C_v = \frac{|\{\{x, y\} \in E : x, y \in N_v\}|}{C^2_{n_v}}, \quad C(G) = \frac{1}{|V|}\sum_{v \in V} C_v, \quad T(G) = \frac{3\#(K_3, G)}{\#(P_2, G)}$$

Посчитаем C_v, предпологая что n достаточно большое (т.к. мы хотим рассматривать случай в ∞):

- Для левой части: $n_v = 1$ (те, у которых deg$v = 1$) или $C_v = 0$ (та, соединяющая с правой частью).
- Для входной вершины правой части (той вершины, которая соединена с левой): $C_v = \frac{2}{C^2_5} = 1/5$.
- Для вершины правой части первого типа (той вершины, которая соединена с 2 другими вершинами): $C_v = \frac{2}{C^2_4} = 1/3$. Такие вершины $n/4 - 1$ штук.
- Для вершины правой части второго типа (той вершины, которая соединена с 4 другими вершинами): $C_v = \frac{1}{C^2_2} = 1$. Такие вершины $n/4$ штук.

Таким образом, средний локальный кластерный коэффициент графа G:

$$C(G) = \frac{1}{\frac{n}{2}+1} \times (\frac{1}{5} + \frac{1}{3} \times (\frac{n}{4} - 1) + 1 \times \frac{n}{4}) = \frac{2(5n-2)}{15(n+2)} \overset{n\to\infty}{\to} \frac{2}{3}$$

Поэтому, найдется такой положительный констант‌ c, для которого выполняется $C(G) \geqslant c$ при достаточном большом n.

Посчитаем $T(G)$:

$$T(G) = \frac{3\#(K_3, G)}{\#(P_2, G)} = \frac{3 \cdot \frac{1}{4}n}{C^2_{n/2} + C^2_5 + \frac{n}{4}C^3_4 + (\frac{n}{4} - 1)} = \frac{\frac{3}{4}n}{\frac{1}{8}n^2 + \frac{3}{2}n + 9} \overset{n\to\infty}{\to} 0$$

5. Докажите, что при вероятности ребра p такой, что $pn^{2/5} \to 0$ при $n \to +\infty$ в случайном графе $G(n, p)$ в модели Эрдеша-Реньи асимптотически почти наверное нет клик на 6 вершинах, а при вероятности ребра такой, что $pn^{2/5} \to +\infty$ при $n \to +\infty$ в случайном графе $G(n, p)$ асимптотически почти наверное есть клика на 6 вершинах.

Решение.

- Докажем, что а.п.н. нет клик на 6 вершинах при $n \to \infty$. Пусть $\xi = \xi(G)$ - с.в. числа клик на 6 вершинах. По неравенству Маркова:

  $$P(\xi \geqslant 1) \leqslant \mathbb{E}\xi$$

  Рассмотрим все шестёрки вершин $t_1, \ldots, t_{C^6_n}$. Пусть T_i - событие, что шестёрка t_i образует клик на 6 вершинах, $i = 1, \ldots, C^6_n$, и $\mathbb{I}_i$ - индикатор события i. Тогда

  $$\mathbb{E}\xi = \mathbb{E}(\mathbb{I}_{T_1} + \ldots + \mathbb{I}_{T_{C^6_n}}) = \sum_{i=1}^{C^6_n} \mathbb{E}\mathbb{I}_{T_i} = C^6_n p^{15} \sim \frac{n^6p^{15}}{6!} = \frac{((pn)^{2/5})^{15}}{6!} \overset{n\to\infty}{\to} 0$$

  Что и означает $P(\xi \geqslant 1) \to 0$ при $n \to \infty$.

- Докажем, что а.п.н. есть клик на 6 вершинах при $n \to \infty$. Пусть $\xi = \xi(G)$ - с.в. числа клик на 6 вершинах. Заметим, что по неравенству Чебышева выполнено

  $$P(|\xi - \mathbb{E}\xi| \geqslant \varepsilon) \leqslant \frac{\mathbb{D}\xi}{\varepsilon^2} = \frac{\mathbb{E}\xi^2 - (\mathbb{E}\xi)^2}{\varepsilon^2}$$

  Заменив ε на $\mathbb{E}\xi$, получаем

  $$P(\xi \geqslant 1) = 1 - P(\xi \leqslant 0) = 1 - P(-\xi \geqslant 0) = 1 - P(\mathbb{E}\xi - \xi \geqslant \mathbb{E}\xi) \geqslant$$
  $$1 - P(|\xi - \mathbb{E}\xi| \geqslant \mathbb{E}\xi) \geqslant 2 - \frac{\mathbb{E}\xi^2}{(\mathbb{E}\xi)^2}$$

Теперь докажем, что $\frac{\mathbb{E}\xi^2}{(\mathbb{E}\xi)^2} \to 1$ при $n \to \infty$. $\mathbb{E}\xi = (C_n^6 p^{15})^2 = (C_n^6)^2 p^{30}$, а $\mathbb{E}\xi^2$ вычисляется так: пусть опять T_i - событие, что шестёрка t_i образует клик на 6 вершинах, $i = 1, \ldots, C_n^6$, и $\mathbb{I}_i$ - индикатор события i. Тогда

$$\mathbb{E}\xi^2 = \mathbb{E}(\mathbb{I}_{T_1} + \ldots + \mathbb{I}_{T_{C_n^6}})^2 = \mathbb{E}(\sum_{i=1}^{C_n^6} \mathbb{I}_{T_i}^2 + \sum_{i \neq j} \mathbb{I}_{T_i}\mathbb{I}_{T_j}) = \sum_{i=1}^{C_n^6} \mathbb{E}\mathbb{I}_{T_i} + \sum_{i \neq j} \mathbb{E}\mathbb{I}_{T_i}\mathbb{I}_{T_j}$$

Здесь $\mathbb{I}_{T_i}\mathbb{I}_{T_j}$ - событие того, что клик i и клик j появляются одновременно. Все возможные типы комбинаций и их верояности и количества можно смотреть следующую таблицу:

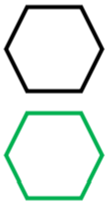
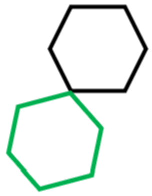
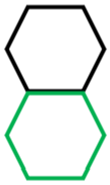
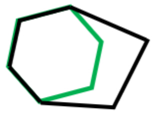
Возможные типы комбинаций шестёрок						
Вероятность	**p^30**	**p^30**	**p^29**	**p^28**	**p^27**	**p^26**
Количество	$C_n^6 \cdot C_{n-6}^6$	$C_n^6 \cdot C_6^1 \cdot C_{n-6}^5$	$C_n^6 \cdot C_6^2 \cdot C_{n-6}^4$	$C_n^6 \cdot C_6^3 \cdot C_{n-6}^3$	$C_n^6 \cdot C_6^4 \cdot C_{n-6}^2$	$C_n^6 \cdot C_6^5 \cdot C_{n-6}^1$

Рис. 3: Иллюстрация к задачи 5

Поэтому, итоговая формула для $\mathbb{E}\xi^2$ такая:

$$\mathbb{E}\xi^2 = C_n^6 \cdot (C_{n-6}^6 \cdot p^{30} + C_6^1 \cdot C_{n-6}^5 \cdot p^{30} + \sum_{i=2}^{5} C_6^i \cdot C_{n-6}^{6-i} \cdot p^{31-i}) \sim (C_n^6)^2 (p^{30} + \ldots + p^{26}), \text{при } pn^{2/5} \to \infty, n \to \infty$$

И того, мы получаем, что

$$\frac{\mathbb{E}\xi^2}{(\mathbb{E}\xi)^2} \sim \frac{(C_n^6)^2 (p^{30} + \ldots + p^{26})}{(C_n^6)^2 p^{30}} \sim 1, \quad \text{при } pn^{2/5} \to \infty, n \to \infty$$

$$P(\xi \geqslant 1) \geqslant 2 - \frac{\mathbb{E}\xi^2}{(\mathbb{E}\xi)^2} \to 1, \quad \text{при } pn^{2/5} \to \infty, n \to \infty$$

6. Докажите, что при $p = \frac{c}{n}$, где $c < 1$, в случайном графе $G(n, p)$ в модели Эрдеша-Реньи асимптотически почти наверное каждая компонента связности содержит не более одного цикла.

Решение.

2 Практическая часть

В данном пдф-ке представлены результаты практической части. Детали (коды программы, графики, визуализация итд) см. репозиторию

https://github.com/mmmiiinnnggg/BigGraph-Social-Network-Analysis

Задача 7. В файле **graph.txt** находится фрагмент веб-графа 2002 года, заданный списком ориентированных рёбер, представляющих гиперссылки между страницами. В графе 875 тыс. вершин и 5.1 млн. рёбер. Постройте график кумулятивного распределения полных степеней вершин этого графа в логарифмических координатах. Аппроксимируйте «основную» часть графика степенным законом $\frac{c}{d^\gamma}$ Постройте на том же графике полученную функцию.

Решение. Основные коды для кумулятивного распределения и аппроксимации степенным законом:

```
from collections import defaultdict
degrees = defaultdict(int)

for v in G.nodes():
    d = G.degree(v)
    degrees[d] += 1

# cumulative distribution
cd = defaultdict(int)
s = 0
for d in sorted(degrees.keys(), reverse = True):
    s += degrees[d]
    cd[d] = s

# main part of data
x_main = [i for i in x if i >= 1e1 and i <= 1e3]
y_main = [cd[d] for d in x_main]

def power_law(d, c, gamma):
    return c / np.power(d,gamma)

# approximation
popt, _ = curve_fit(power_law, x_main , y_main)
```

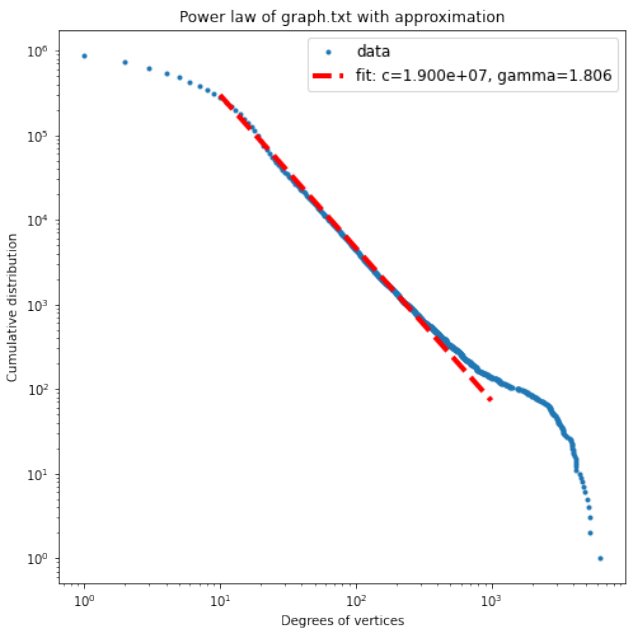


Рис. 4: Степенной закон в задачи 7

Задача 8. Коэффициент Жаккара для двух множеств A и B определяется как $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. Определим величину $J(e)$ для ребра $e = (u, v)$ графа G как $J(N_u, N_v)$, где за N_x обозначено множество соседей вершины x графа G. Далее, определим величину $J(G)$ как среднее значение $J(e)$ по всем ребрам графа G. При $n = 300$ сгенерируйте при разных значениях $p \in \{0, 0.05, 0.1, 0.15, ..., 1.0\}$ граф $G(n, p)$ в модели Эрдёша–Реньи и постройте график $J(G(n, p))$ в зависимости от p. Найдите теоретическую оценку для среднего значения этой величины и постройте её на том же графике.

Решение. Коэффициент Жаккера для графа модели Эрдёша-Реньи:

$$J(G) = \frac{1}{|E|} \sum_{e=(u,v) \in E} J(e) = \frac{1}{|E|} \sum_{e=(u,v) \in E} \frac{|N_u \cap N_v|}{|N_u \cup N_v|}$$

коды для "jaccard_coefficient":

```python
def jac_coeffi(G, u, v):
    union_size = len(set(G[u]) | set(G[v]))
    if union_size == 0:
        return 0
    return len(list(nx.common_neighbors(G, u, v))) / union_size

def jac_coeffi_graph(G):
    jc_g = 0
    for edge in list(G.edges()):
        u, v = edge
        jc_g += jac_coeffi(G, u, v)
    if len(G.edges()) == 0:
        return 0
    else:
        return jc_g / len(G.edges())
```

Теоретическая оценка для среднего значения этого коэффициента:

$$\mathbb{E}J(G) = \frac{p^2(n-2)}{p(2-p)(n-2)+2}$$

Из картинки видно, что у нас теоретическая оценка и результат эксперимента почти равны везде.

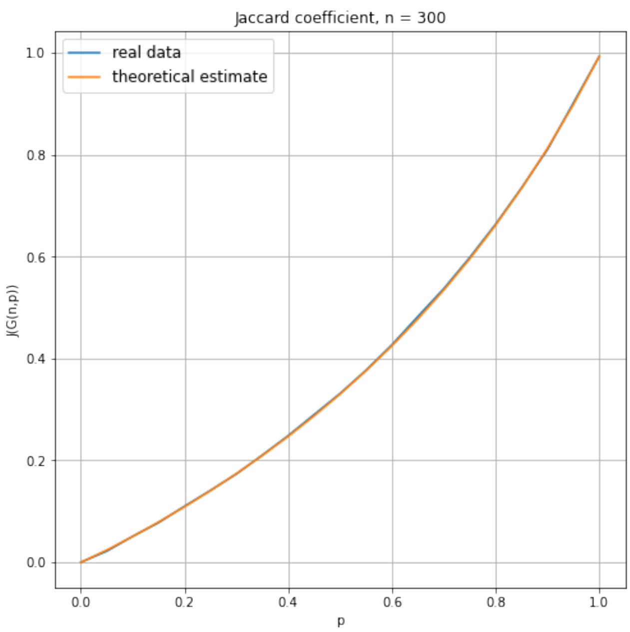


Рис. 5: Эксперимент и теоретическая оценка в задачи 8

Задача 9. Выберите LiveJournal какого-нибудь блогера. По адресу

http://www.livejournal.com/misc/fdata.bml?user=<username>

можно скачать список пользователей, которые являются друзьями или находятся в друзьях у пользователя <username>. Для каждого из друзей выбранного блогера также скачайте список его друзей, в итоге получится ориентированный граф друзей пользователя. Найдите локальный кластерный коэффициент выбранного блогера в графе социальной сети. Визуализируйте полученный фрагмент графа так, чтобы размер вершин и подписей к ним были пропорциональны полной степени. Вершины должны быть подписаны именами соответствующих пользователей.

Решение. В этой задаче берется “username = genbu_no_mike24”. Результаты визуализации можно смотреть на следующих картинах. Давайте пока считаем локальный кластерный коэффициент для данного пользователя:

```
count = 0
with open(file_name, 'r') as f:
    reader = csv.reader(f)
    headers = next(reader)
    for row in reader:
        if row[0] in user1_name_list and row[1] in user1_name_list:
            count += 1

num_poss_edges = len(user1_name_list) * (len(user1_name_list) - 1) / 2
print("The local cluster coefficient of '{}' = {}".format(user_one ,count / num_poss_edges))
```

The local cluster coefficient of 'genbu_no_miko24' = 0.10849963045084997

Рис. 6: Визуализация итоговой соцсети

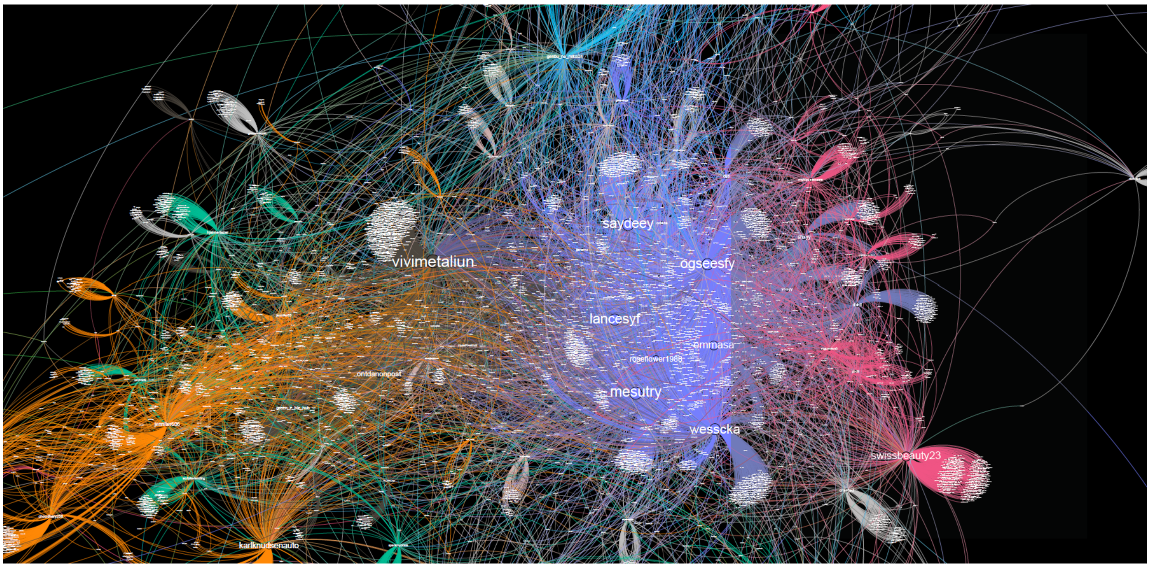


Рис. 7: Zoom-in на ключевых пользователей

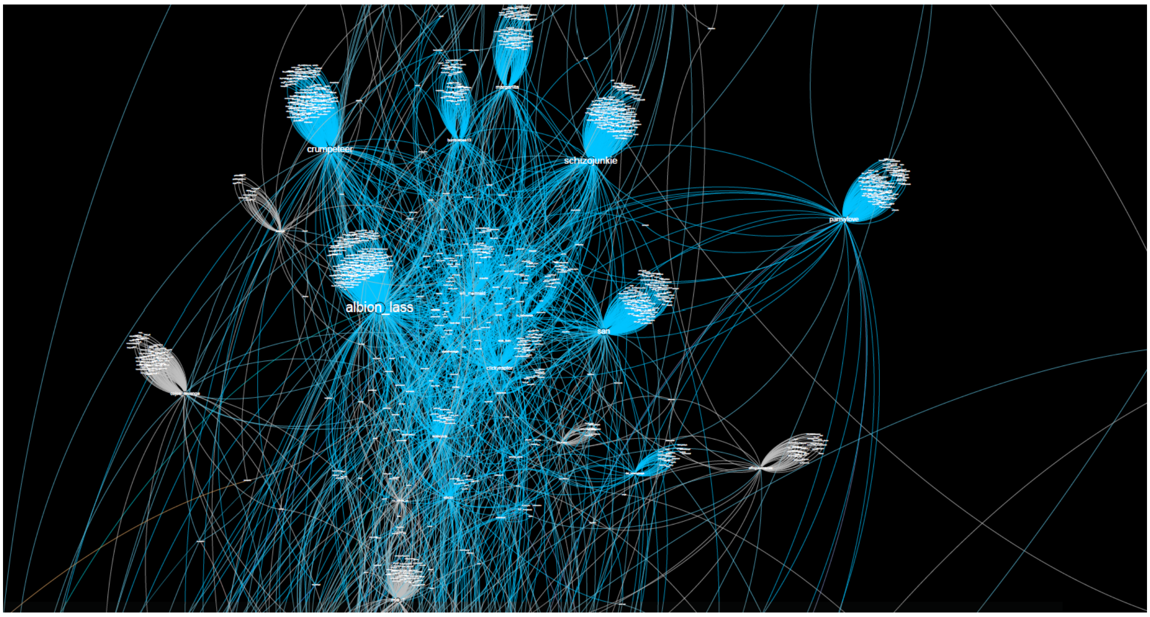


Рис. 8: Zoom-in для другого сообщества