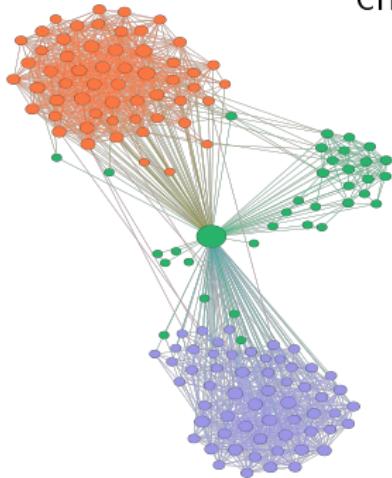


Учебная практика

«Анализ социальных графов»

Бронников Егор

СПбГЭУ



1 Введение

2 Сбор данных и визуализация

3 Характеристики сети

4 Генетический алгоритм

5 Алгоритм Гирван-Ньюмена

6 Заключение

Введение

Практическая работа посвящена построению эго-графов пользователей в социальной сети ВКонтакте, вычислению основных показателей сети и выделение пользователей в группы (сообщества).

Цель исследования

Приобретение навыков научно-исследовательской работы, посвящённой анализу эго-графов пользователей социальной сети ВКонтакте и разбиению пользователей сети на группы.

Объект исследования

Социальные графы и эго-графы пользователей социальной сети ВКонтакте.

Предмет исследования

Характеристики и метрики для сети и алгоритмы разбиения пользователей в группы.

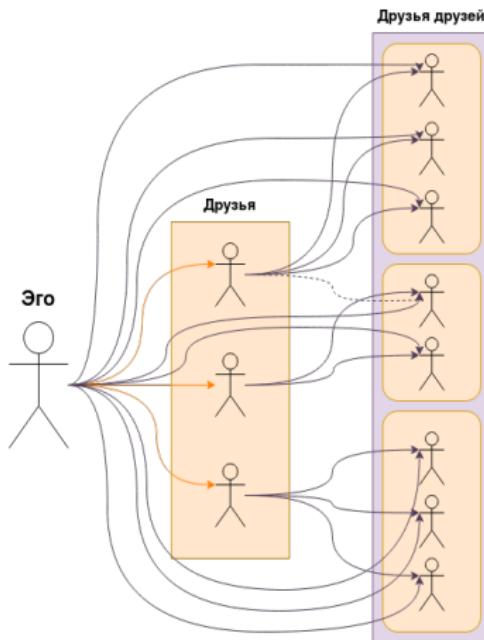
Задачи работы

- ❶ изучить общий подход анализа социальных графов и эго-графов;
- ❷ собрать данные для построения сети с помощью VK API;
- ❸ визуализировать получившуюся сеть;
- ❹ рассчитать основные метрики и характеристики сети;
- ❺ реализовать алгоритмы для выделения пользователей в сообщества.

Результатом работы является программные решения для анализа социальных графов и эго-графов.

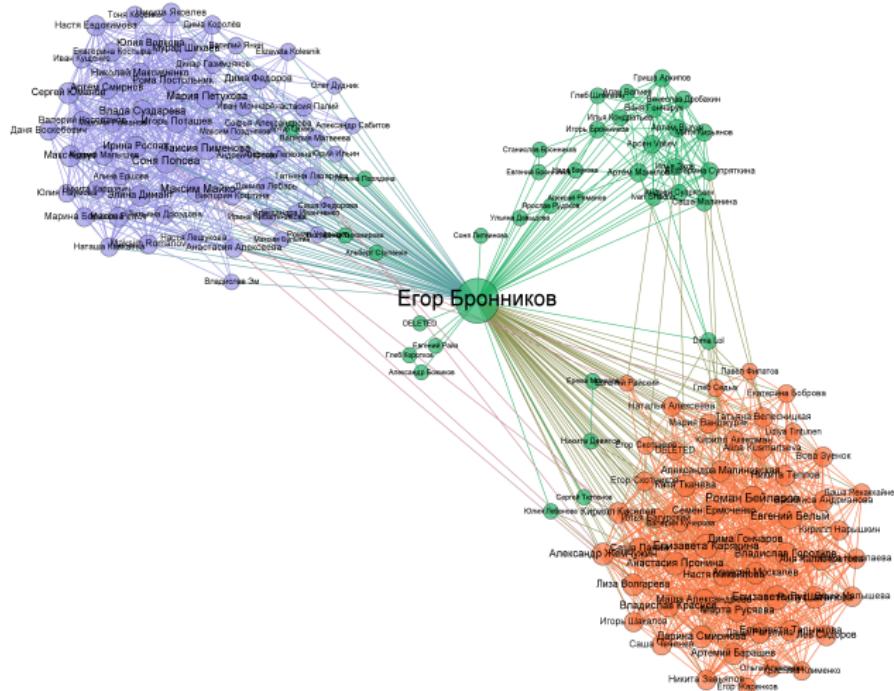
Сбор данных и визуализация

Формируется список друзей пользователя, а затем список друзей друзей, но с которыми знаком исходный пользователь.



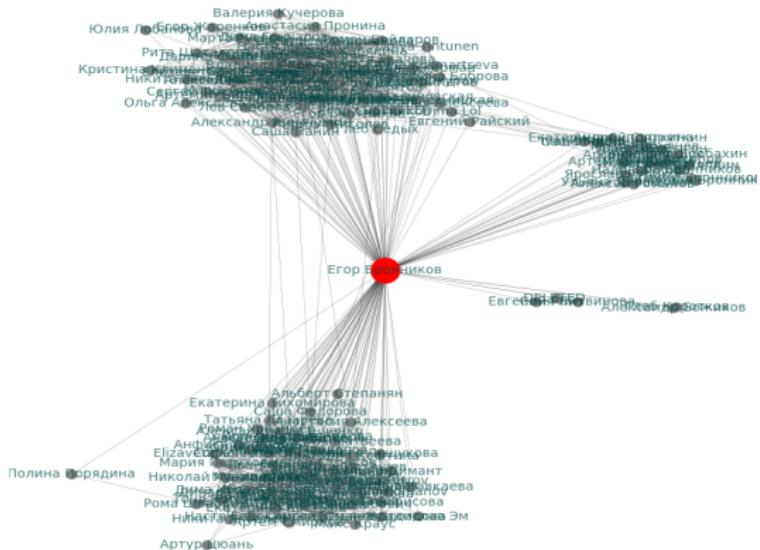
Сбор данных и визуализация

Визуализация данных с помощью инструмента *Gephi*.



Сбор данных и визуализация

Визуализация данных с помощью модуля *NetworkX*.



Степень связности

В метрике степени связности важность вершины определяется тем, с каким количеством смежных вершин она связана.

$$\deg(i) = \sum_{j \in V} m_{ij}$$

Степень близости у другим узлам

Степень близости к другим узлам можно определить как то, насколько близко к определённому субъекту находятся другие субъекты сети.

$$C(i) = \sum_{j \in V} d_{ij}$$

где d_{ij} — количество звеньев в кратчайшем пути от вершины i до вершины j .

Степень связности

В метрике степени связности важность вершины определяется тем, с каким количеством смежных вершин она связана.

$$\deg(i) = \sum_{j \in V} m_{ij}$$

Степень близости у другим узлам

Степень близости к другим узлам можно определить как то, насколько близко к определённому субъекту находятся другие субъекты сети.

$$C(i) = \sum_{j \in V} d_{ij}$$

где d_{ij} – количество звеньев в кратчайшем пути от вершины i до вершины j .

Степень посредничества

Степень посредничества представляет собой количество раз, которое участник должен пройти через данный узел, чтобы достичь другого участника сети.

$$b(i) = \sum_{j,k \in V} \frac{g_{ijk}}{g_{jk}}$$

где g_{jk} — это количество кратчайших путей от вершины j до вершины k ;

g_{ijk} — это количество кратчайших путей от вершины j до вершины k проходящих через вершину i .

Эксцентриситет

$$E(i) = \max_{j \in V} d_{ij}$$

где d_{ij} — количество звеньев в кратчайшем пути от вершины i до вершины j .

Характеристики сети

Степень посредничества

Степень посредничества представляет собой количество раз, которое участник должен пройти через данный узел, чтобы достичь другого участника сети.

$$b(i) = \sum_{j,k \in V} \frac{g_{ijk}}{g_{jk}}$$

где g_{jk} — это количество кратчайших путей от вершины j до вершины k ;

g_{ijk} — это количество кратчайших путей от вершины j до вершины k проходящих через вершину i .

Эксцентриситет

$$E(i) = \max_{j \in V} d_{ij}$$

где d_{ij} — количество звеньев в кратчайшем пути от вершины i до вершины j .

Пусть $S = (I, J)$ подматрица матрицы A , где I – это подмножество строк $X = \{I_1, \dots, I_N\}$ матрицы A и J – это подмножество столбцов $Y = \{J_1, \dots, J_N\}$ матрицы A .

Постановка задачи

Найти разбиение матрицы смежности A на k подматриц, которые максимизируют сумму плотностей подматриц.

Среднее значение строки

Пусть a_{iJ} означает среднее значение i -й строки S :

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}$$

Среднее значение столбца

Пусть a_{Ij} означает среднее значение j -й столбца S :

$$a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$$

Объём матрицы

$$v_S = \sum_{i \in I, j \in J} a_{ij}$$

Среднее значение мощности подматрицы

Среднее значение мощности подматрицы S порядка r , обозначаемое как $M(S)$, рассчитывается по следующей формуле:

$$M(S) = \frac{\sum_{i \in I} (a_{iJ})^r}{|I|}$$

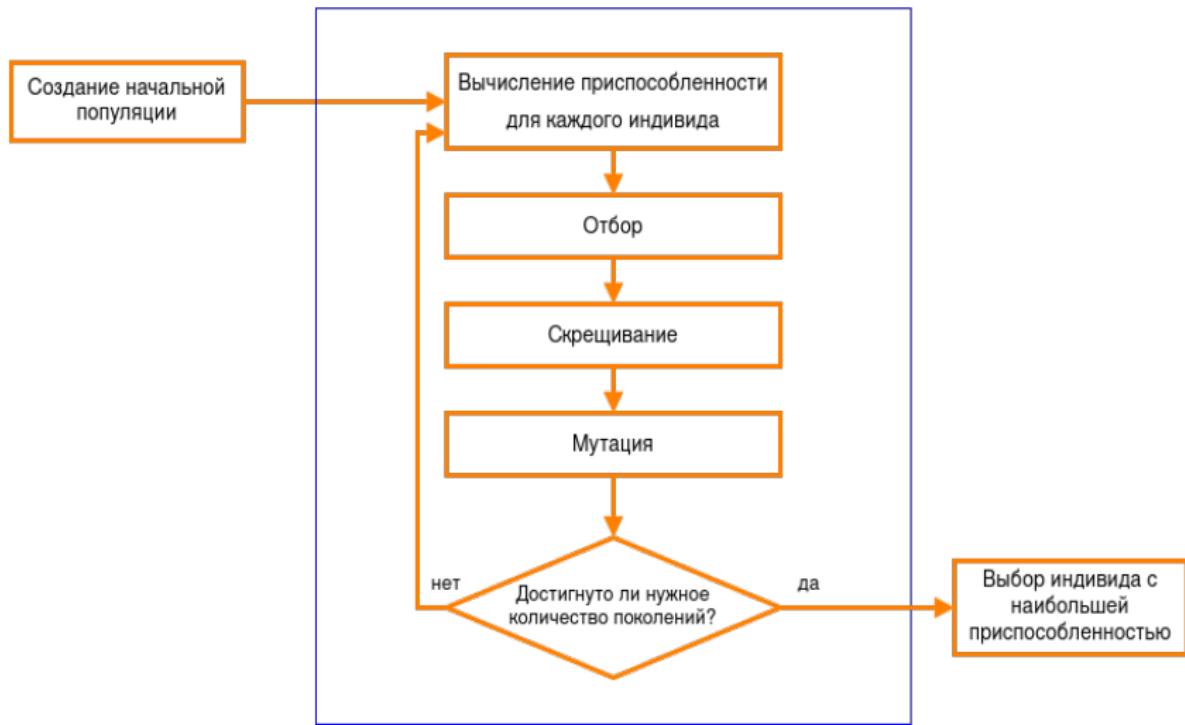
Оценка множества разбиений

Оценка множества разбиений $\{S_1, \dots, S_k\}$ матрицы A определяется следующим образом:

$$CS = \sum_{i=1}^k Q(S_i) \longrightarrow \max$$

где оценка подматрицы S_i определяется как: $Q(S_i) = M(S_i) \times v_{S_i}$

Имитация эволюционного процесса



Инициализация

Инициализация индивидов популяции осуществляется следующим образом:

Вершины графа: 1, 2, 3, ..., N
Индивид: [45, 123, 4, ..., g_N]

Создание решений задачи

- 1 генерируется исходный список подмножеств индивида как пара $\{\{i, g_i\} \mid \forall i \in \{1, \dots, N\}\}$;
- 2 два подмножества объединяются если существует элемент из пересечения этих разбиений.

Инициализация

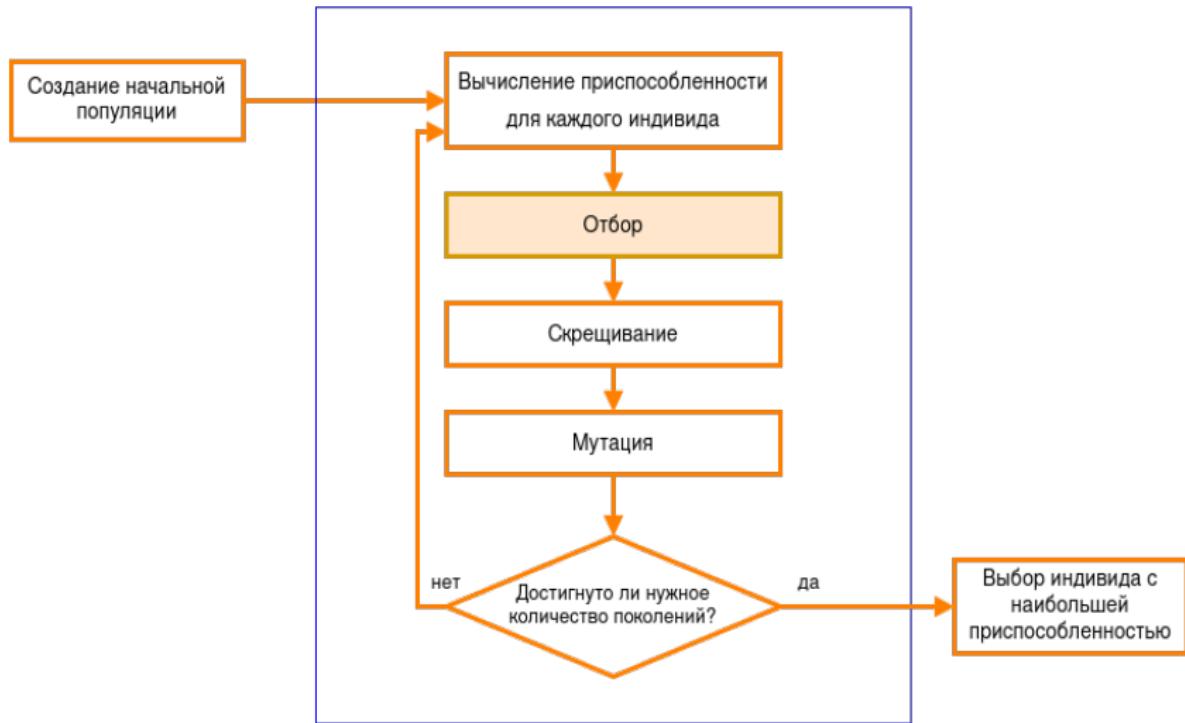
Инициализация индивидов популяции осуществляется следующим образом:

Вершины графа: 1, 2, 3, ..., N
Индивид: [45, 123, 4, ..., g_N]

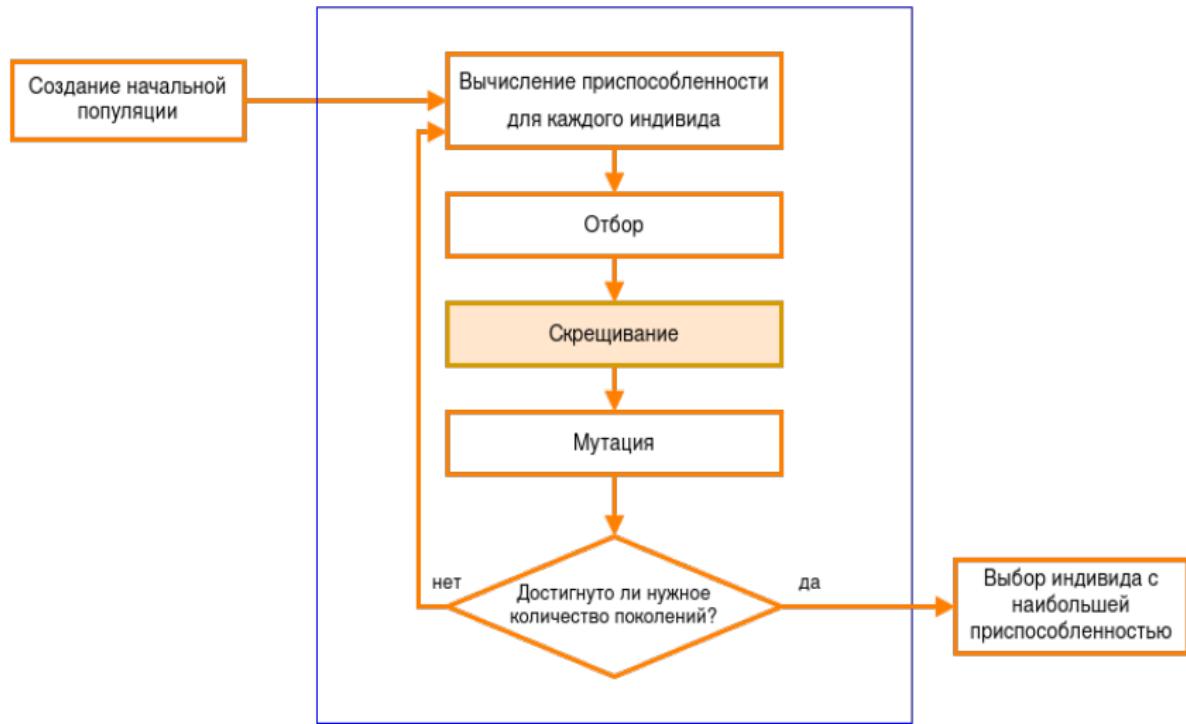
Создание решений задачи

- 1 генерируется исходный список подмножеств индивида как пара $\{\{i, g_i\} \mid \forall i \in \{1, \dots, N\}\}$;
- 2 два подмножества объединяются если существует элемент из пересечения этих разбиений.

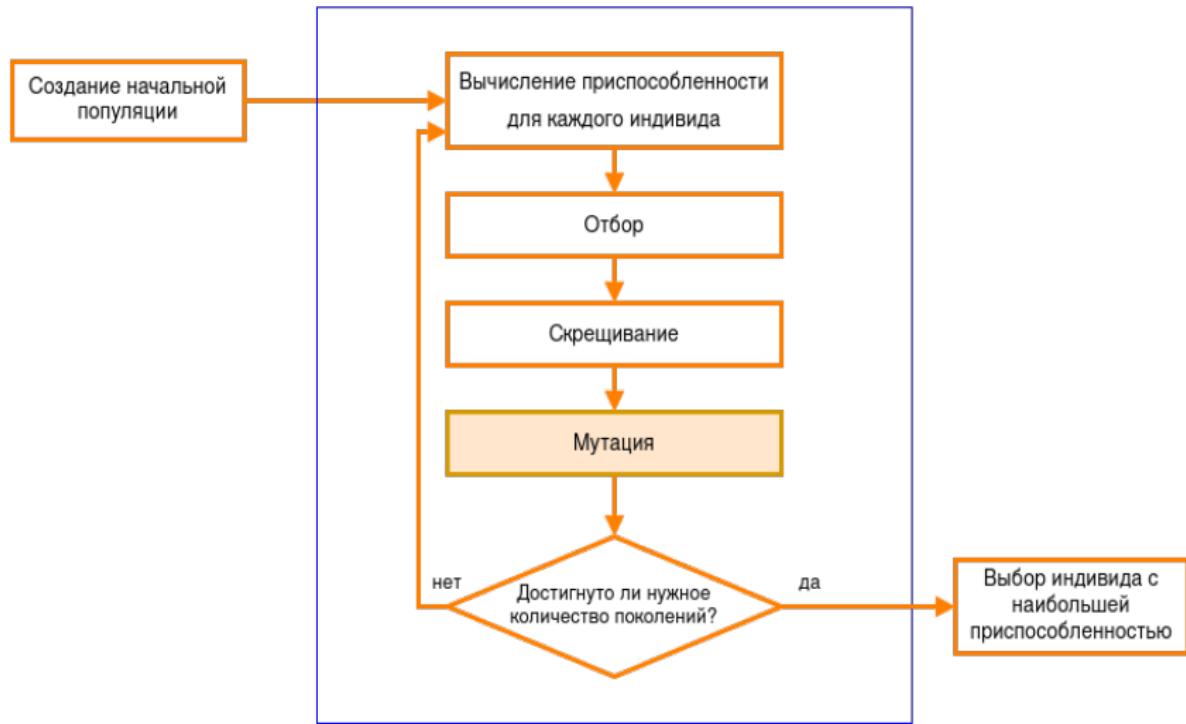
Имитация эволюционного процесса



Имитация эволюционного процесса

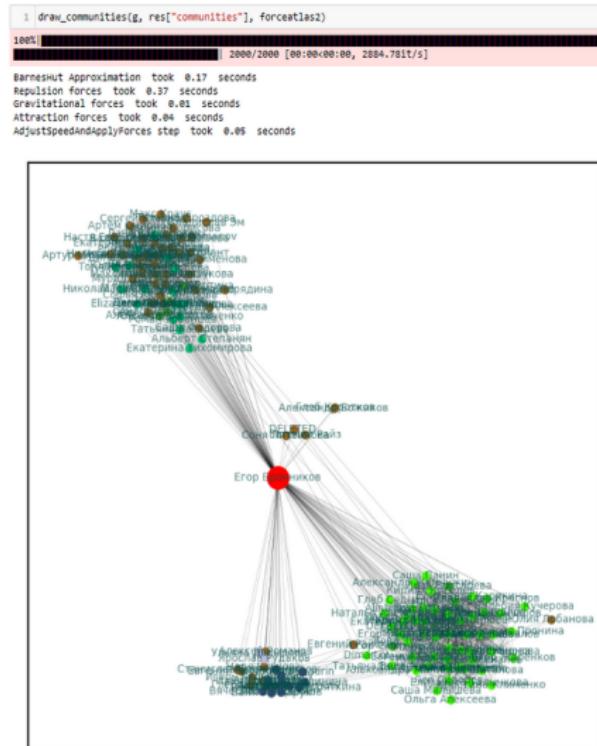


Имитация эволюционного процесса



Генетический алгоритм

Визуализация полученного результата:



Степень посредничества ребра

$$c_B(e) = \sum_{s \in V, t \in V} \frac{\sigma(s, t|e)}{\sigma(s, t)}$$

где $\sigma(s, t)$ – количество кратчайших путей между вершинами s и t ;
 $\sigma(s, t|e)$ – количество кратчайших путей между вершинами s и t , которые проходят через ребро e

Модулярность

$$Q = \sum_{c=1}^n \left[\frac{L_c}{m} - \gamma \left(\frac{k_c}{2m} \right)^2 \right]$$

где m – количество рёбер;

c – сообщество;

L_c – количество внутренних рёбер сообщества c ;

k_c – сумма степеней вершин в сообществе c ;

γ – параметр разрешения.

Алгоритм Гирван-Ньюмена

Степень посредничества ребра

$$c_B(e) = \sum_{s \in V, t \in V} \frac{\sigma(s, t|e)}{\sigma(s, t)}$$

где $\sigma(s, t)$ – количество кратчайших путей между вершинами s и t ;
 $\sigma(s, t|e)$ – количество кратчайших путей между вершинами s и t , которые проходят через ребро e

Модулярность

$$Q = \sum_{c=1}^n \left[\frac{L_c}{m} - \gamma \left(\frac{k_c}{2m} \right)^2 \right]$$

где m – количество рёбер;

c – сообщество;

L_c – количество внутренних рёбер сообщества c ;

k_c – сумма степеней вершин в сообществе c ;

γ – параметр разрешения.

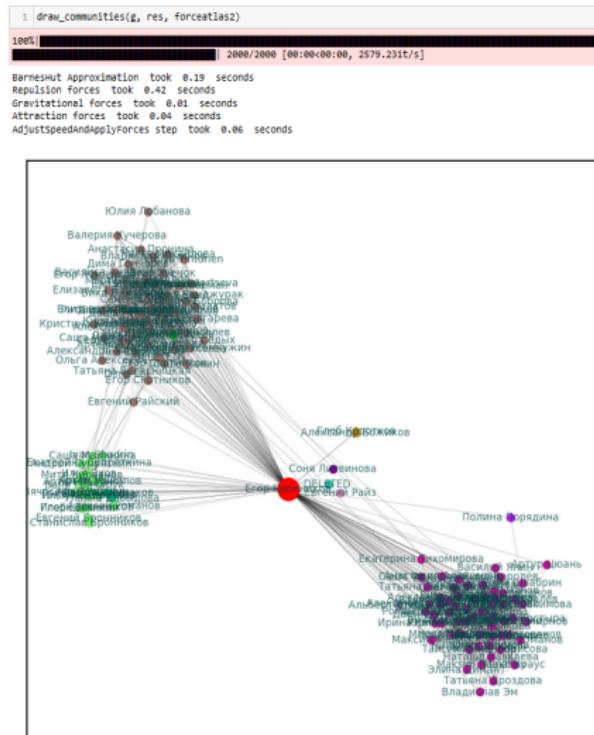
Описание алгоритма

Данный алгоритм может быть сформулирован несколькими следующими этапами:

- ① вычисляются степени посредничества всех рёбер;
- ② ребро с наибольшей степенью посредничества удаляется;
- ③ степени посредничества всех затронутых рёбер вычисляются заново;
- ④ шаги 2 и 3 повторяются до тех пор, пока уровень модулярности не станет убывать.

Алгоритм Гирван-Ньюмена

Визуализация полученного результата:



В ходе прохождения учебной практики, посвящённой приобретению навыков научно-исследовательской работы были достигнуты следующие результаты:

- проанализированы метрики и характеристики социальных графов;
- реализованы два алгоритма нахождения сообществ;
- изучены вспомогательные средства для визуализации графов.

В рамках данной работы разработано программное средство для выделения сообществ в социальных графах.