УДК 519.178

РАЗРАБОТКА ТЕХНОЛОГИИ ВЫДЕЛЕНИЯ СООБЩЕСТВ В СОЦИАЛЬНОМ ГРАФЕ

Никифоров Г.А.
Магистр 2 курса
Национальный исследовательский ядерный университет "МИФИ"
Москва, Россия

Коревых М.А.
Магистр 2 курса
Национальный исследовательский ядерный университет "МИФИ"
Москва, Россия

Низаметдинов Ш.У.
Кандидат технических наук, научный руководитель
Национальный исследовательский ядерный университет "МИФИ"
Москва, Россия

Аннотация.
В данной статье систематизируются методы выделения сообществ социального графа, описываются технология разработки соответствующего инструментального средств, приводятся результаты экспериментального исследования качества и быстродействия основных методов выделения сообществ в социальных графах, основанных на данных групп социальной сети "Вконтакте".
Ключевые слова: выделение сообществ графа, кластеризация графа, Python, социальный граф, социальная сеть, визуализация графа, Вконтакте.

DEVELOPMENT OF TECHNOLOGY FOR COMMUNITY DETECTION IN SOCIAL GRAPH

Nikiforov G.A.
2nd year master
National research nuclear university "MEPhI"
Moscow, Russia

Korevykh M.A.
2nd year master
National research nuclear university "MEPhI"
Moscow, Russia

Nizametdinov Sh.U.
Candidate of Technical Sciences
National research nuclear university "MEPhI"
Moscow, Russia

Annotation.
In this article there were systematized methods for detecting community structure in social graph, was described the technology for developing corresponding tool, were presented the results of an experimental study of the quality and performance of the main methods for detecting communities in social graphs based on data from the Vkontakte social network.

Key words: community detection in graph, graph clustering, Python, social graph, social network, graph visualization, Vkontakte

Социальные сети как общественное явление появились довольно давно. Анализ социальных сетей используется для исследования взаимодействий между участниками сети, прогнозирования их поведения, классификации, моделирования информационных потоков в сетях.

Исследования социальных групп Эмиля Дюркгейма и Фердинанда Тённиса положили начало анализа социальных сетей. Эти исследования они проводили в течение второй половины XIX века. В первой половине XX века многие учёные из различных областей науки проводили работы в данном направлении независимо. В частности, собиралась и анализировалась информация о взаимоотношениях в малых группах. Во второй половине XX века теории, методы и подходы из разных наук начали формировать новое течение в науке. В XXI веке благодаря распространению интернета, сетевых социальных сервисов и ростом вычислительных мощностей начался рост количества информации, пригодной для анализа и машинной обработки. В связи с этим интерес к этой области повысился.

В настоящее время с развитием компьютерных технологий у людей появилась возможность общаться виртуально при помощи компьютерных

социальных сетей. Поэтому анализ именно компьютерных социальных сетей вызывает большой интерес у современных исследователей.

Социальные сети насчитывают сотни миллионов пользователей, которые объединяются в различные группы по профессиональным и личным интересам. Эти группы могут быть могут быть формально выделены или образовываться с течением времени. Также в них могут образовываться более мелкие подгруппы с абсолютно отличной структурой.

1 Выделение сообществ и выделение значимых вершин социального графа

В анализе социальных графов можно выделить четыре основных направления исследований: структурное, ресурсное, нормативное и динамическое [1]. В каждом из них решается довольно большой круг задач и применяются методы из различных областей знаний. В данной работе рассматриваются только структурные методы исследования социального графа.

При структурном анализе и анализе поведения связей используются методы статистического анализа, определения сообществ, алгоритмы классификации. Изучается поведение вершин в процессе кластеризации и типичных временных характеристик социальных сетей. Например, как меняется структура сети в процессе роста или как меняются поведение и распределение связных компонентов графа.

Большое значение придается определению сообществ в социальных сетях. Цель – попытаться определить регионы сети, внутри которых происходит активное взаимодействие участников. Алгоритмически эту задачу можно отнести к задаче о разделении графов [2]. Необходимо разделить сеть на плотные регионы на основе поведения связей между вершинами. Компьютерные социальные сети динамичны, что приводит к затруднениям с точки зрения выявления сообществ. В некоторых случаях удается интегрировать

информационное содержимое сети в процесс определения сообществ. Тогда контент является вспомогательным средством для выявления групп участников с похожими интересами.

1.1 Социальный граф

Социальный граф — граф, узлы которого представлены социальными объектами, такими как пользовательские профили с различными атрибутами (например, имя, день рождения, родной город), а рёбра — социальными связями между ними. [2]

В социальных графах существуют дополнительные определения, которые могут сильно влиять на финальную структуру графа.

Чаще всего построение социального графа строится на основе данных, полученных в результате парсинга веб-служб поставщиков социальных сетей.

В данной работе вершиной графа принимается аккаунт пользователя со своим уникальным идентификатором, при этом между двумя вершинами есть ребро, если оба соответствующих им аккаунта находятся друг у друга в списке друзей.

1.2 Выделение сообществ социальных графов

Процесс обнаружения сообществ схож с процессом кластеризации в интеллектуальном анализе данных и также называется обучением без учителя, потому что формирование сообществ может быть выполнено без предварительного знания кластеров. Выявление сообществ применяется в следующий задачах: классификация людей, а именно, влиятельных лиц, выявление взаимосвязанных структур в межбелковых сетях взаимодействия, географическая группировка веб-клиентов для повышения производительности, анализ социальных сетей, таких как Facebook, Twitter, определение веб-страниц той же тематической категории в сети гиперссылок, для выявления схемы связи

между научными работами в сетях цитирования, автоматизированная рекомендация по продуктам для сайтов розничной торговли, прогнозирование политических выборов на основе обсуждения определенных тем в Twitter и т. д. [3]

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных групп должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма. [4]

Применение кластерного анализа в общем виде сводится к следующим этапам:

1. Отбор выборки объектов для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.
3. Вычисление значений меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).
5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Кластеризация графов — это задача разбиения множества вершин графа на группы, называемые кластерами. Внутри каждой группы должны оказаться «наиболее близкие» вершины, а объекты разных группы должны быть как можно более отличны.

На концептуальном уровне сообществом называется такая группа вершин, что внутригрупповые связи гораздо плотнее межгрупповых [2]. Обычно также делают предположение, что сообщество содержится только в одной компоненте связности. В противном случае его можно разделить на несколько более мелких сообществ.

В данной работе используется предположение о структуре сообществ в графе, согласно которому каждая вершина графа входит в одно и только одно сообщество. Случай перекрывающихся сообществ, часто встречающихся в реальных данных, не рассматривается.

Отметим, что из предположения сразу следует, что сообщества полностью покрывают граф. Тогда можно говорить о поиске сообществ в графе как о задаче поиска разбиения множества вершин на подмножества, которое минимизирует некоторый функционал.

Есть другой взгляд на исходную проблему как на задачу кластеризации. Действительно, если ввести некоторую меру расстояния на вершинах графа с учетом структурной информации, можно решать задачу кластеризации. Такой подход также позволяет учитывать дополнительную информацию. Далее будем говорить о задаче кластеризации или поиска разбиения, имея ввиду исходную задачу выделения сообществ в графе.

Самой популярной и общепризнанной мерой качества для данной задачи является значение модулярности (modularity). Функционал был предложен Ньюманом и Гирваном в ходе разработки алгоритма кластеризации вершин графа [5]. Модулярность – это скалярная величина из отрезка [−1, 1], которая количественно описывает структуру сообществ:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta(C_i, C_j), \quad (1)$$

где A — матрица смежности графа, A_{ij} – (i,j) элемент матрицы, d_i — степень i вершины графа, C_i – метка вершины (номер сообщества, к которому относится вершина), m – общее количество ребер в графе. $\delta(C_i, C_j)$ – дельта-функция: равна единице, если $C_i = C_j$, иначе нулю.

Чаще всего задача поиска выделения сообществ в графе сводится к поиску таких, которые максимизируют значение модулярности.

Модулярность достаточно просто интерпретируется. Ее значение равно разности между долей ребер внутри сообщества и ожидаемой доли связей, если бы ребра были размещены случайно.

В данной работе рассматриваются следующие методы выделения сообществ в социальных графах:

- Fast-greedy [6]
- Walktrap [5]
- Label Propagation [7]
- Multilevel [8]
- Eigenvector [9]
- Infomap [10, 11]

2 Разработка инструментального средства выделения сообществ социальных графов

Было разработано инструментальное средство кластеризации социальных графов. Источником данных была выбрана социальная сеть «Вконтакте». Она была выбрана как самая популярная социальная сеть на постсоветском пространстве: на сайте зарегистрировано более 450 млн пользователей.

Для получения возможности скачивания данных из социальной сети «Вконтакте» необходимо иметь ключ доступа к её API, по которому сервер социальной сети сможет идентифицировать отправителя запросов. Цель идентификации отправителя запросов - предотвращение действий злоумышленников. Для этого нужно создать приложение. После этого будет получен секретный ключ, который будет использован для работы инструментального средства.

Разработка Инструментального средства осуществлялась в ОС Windows 8.1. Поскольку библиотека graph-tool для языка Python не поддерживается на ОС WIndows, было принято решение о разворачивании контейнера Docker с ОС Linux.

Для работы с данными социальной сети «Вконтакте» её разработчиками было создано API «Вконтакте». Это API позволяет получать данные из социальной сети в формате .json. Для работы с этим типом данных была использована python библиотека json.

Для обращения к «Вконтакте» требуется составить запрос в формате ссылки, в которой указывается исполняемый метод и требуемые параметры. Результатом этого запроса будет json объект, в котором будут запрашиваемые данные.

Хранение данных о ребрах графа в формате матрицы инцидентности (связей) для больших графов требует большой объем оперативной памяти, поэтому было принято решение использовать список рёбер графа.

После выполнения метода сбора данных был получен список участников и их друзей, а также список связей. На их основе был создан объект класса graph-tool.Graph.

После этого были удалены вершины, количество связей которых равнялось одному. Таким образом были убраны друзья участников групп, которые не являлись связующими нескольких участников группы, а только лишь увеличивали объём обрабатываемых данных.

Затем были удалены вершины, количество связей которых равнялось нулю, как не представляющие интерес в текущей задаче работы с социальными графами.

Был разработан скрипт на языке Python для применения методов кластеризации и визуализации графов.

Результатом предобработки данных на этапе 2.5 является объект класса graph-tool.Graph, содержащий данные об анализируемом графе. Для возможности применения методов кластеризации из библиотеки igraph был разработан метод, получающий на входе объект класса graph-tool.Graph, и выполняющий следующую последовательность действий:

1. Инициализация объекта класса igraph.Graph с таким же набором данных.
2. Применение заданного метода кластеризации на графе.
3. Передача полученных меток кластеров объекту класса graph-tool.Graph в качестве свойства объекта.

На выходе метод возвращает объект класса graph-tool.Graph с метками кластеров для каждой вершины.

В программе использованы следующие методы: community_fastgreedy, community_infomap, community_leading_eigenvector, community_multilevel, community_label_propagation, community_walktrap.

Результаты применения методов кластеризации вершин были визуализированы при помощи методов библиотеки graph-tool. Различным кластерам соответствуют различные цвета вершин.

3 Апробация инструментального средства на реальных данных

Для апробации созданного инструментального средства были выбраны две группы социальной сети «Вконтакте». При помощи разработанного инструментального средства был сформирован социальный граф участников групп и их друзей, содержащий 598 вершин и 1 879 рёбер. Методы выделения сообществ и значимых вершин были апробированы на полученном графе (рис.1).

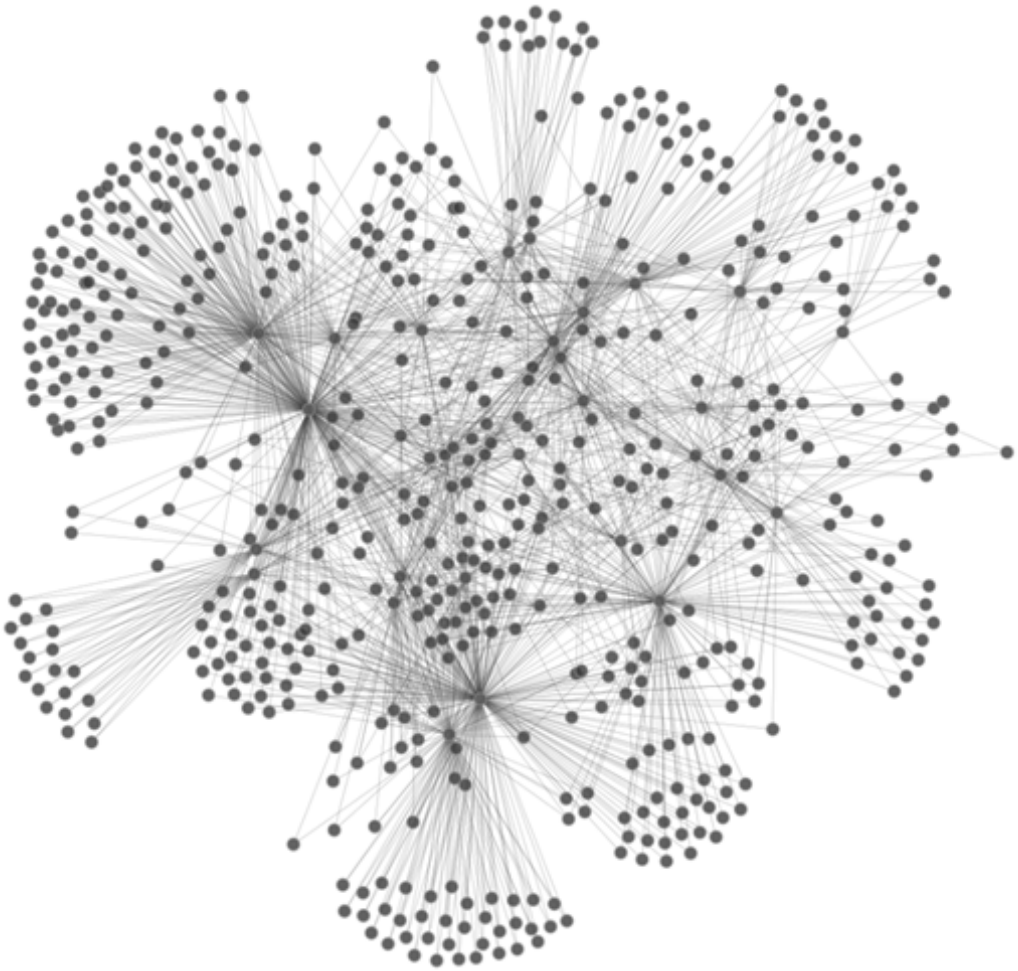

Рис.1 - Визуализация графа

На данном графе были апробированы методы выделения сообществ графа, полученные результаты были визуализированы. При этом вершины, принадлежащие одному сообществу, были закрашены одним цветом.

Была проведена апробация метода выделения сообществ графа Infomap на социальном графе А (рис.2).

Рис.2 - Результат выделения сообществ в рассматриваемом графе методом Infomap

В рассматриваемом социальном графе при помощи метода Infomap было выделено 11 сообществ, причем размеры сообществ различаются на порядок.

Была проведена апробация метода выделения сообществ графа Eigenvector на рассматриваемом графе (рис.3).

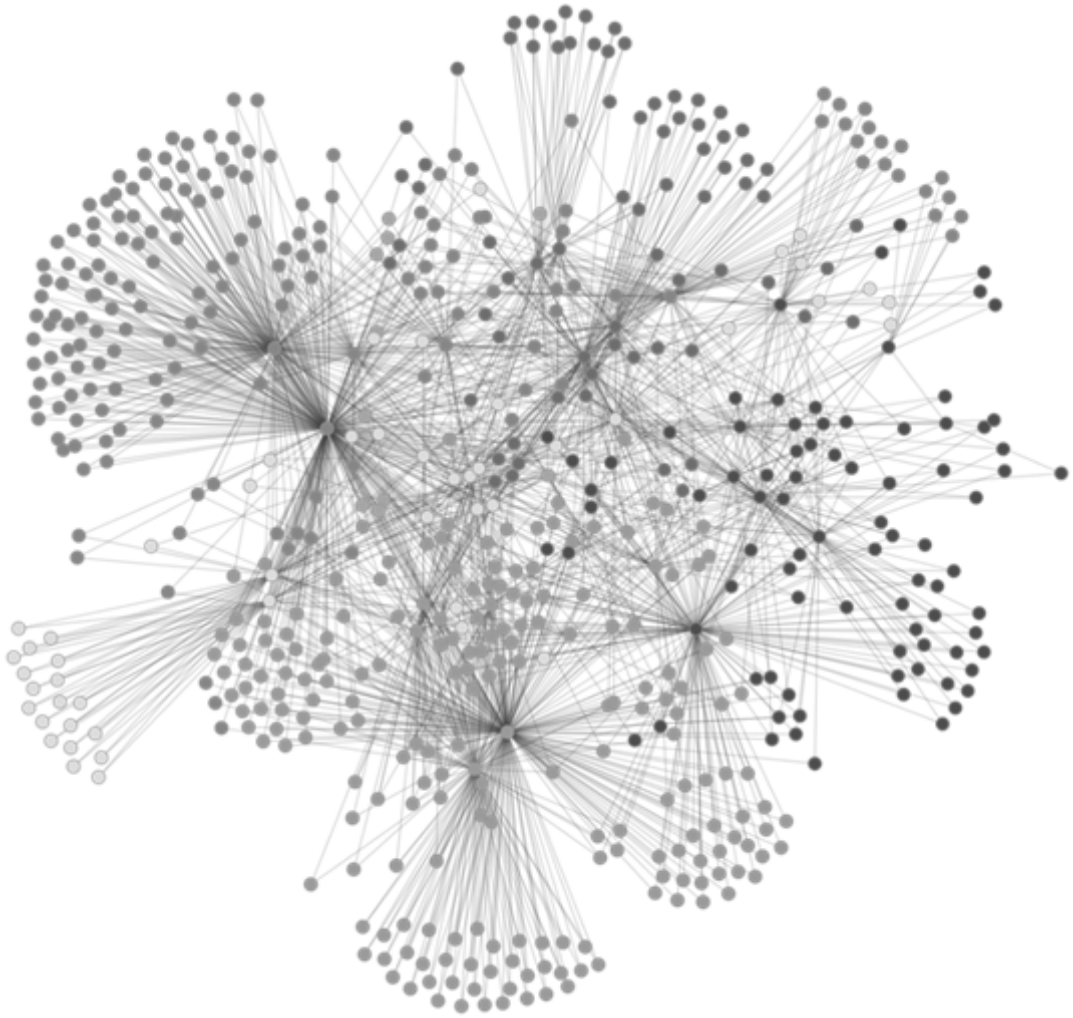

Рис.3 - Результат выделения сообществ в рассматриваемом графе методом Eigenvector

В рассматриваемом графе методом Eigenvector было выделено 6 сообществ, размеры которых различаются на порядок.

Была проведена апробация метода выделения сообществ графа Label Propagation на рассматриваемом графе (рис.4).

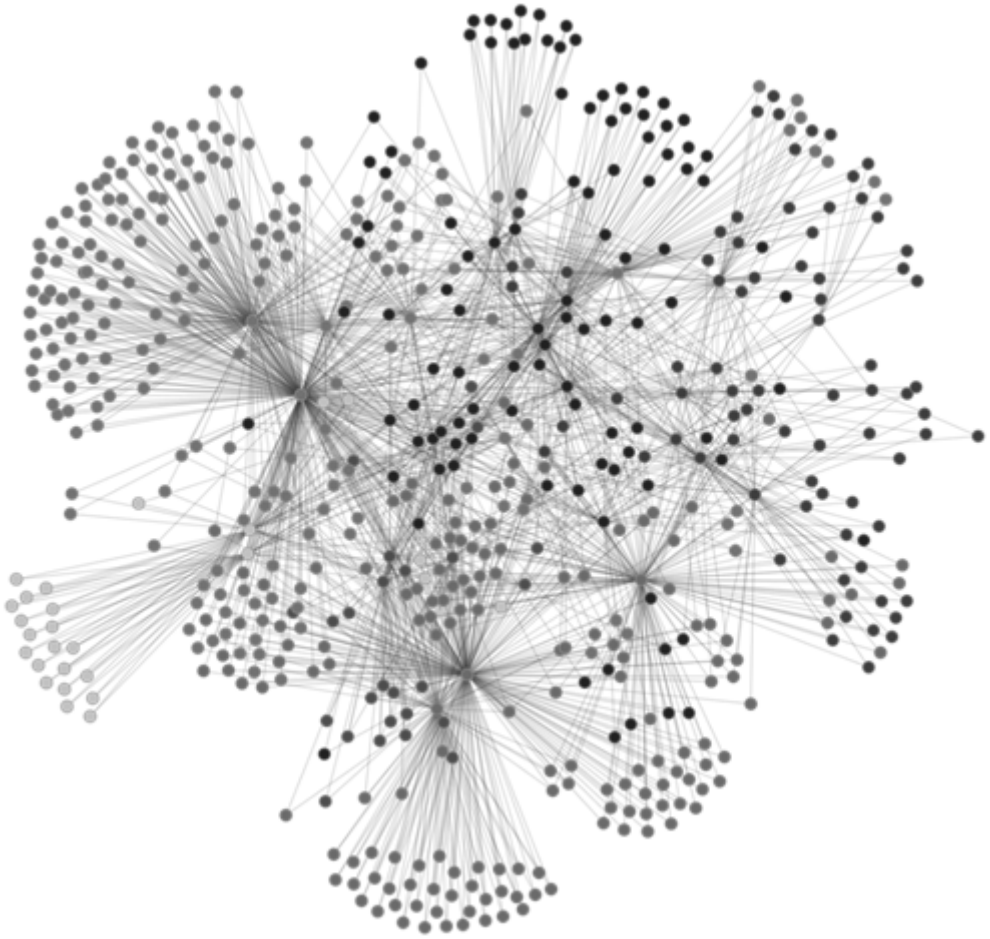

Рис.4 - Результат выделения сообществ в рассматриваемом графе методом Label Propagation

В рассматриваемом графе методом Label Propagation было выделено 5 сообществ.

Была проведена апробация метода выделения сообществ графа Multilevel на рассматриваемом графе (рис.5).

Рис.5 - Результат выделения сообществ в рассматриваемом графе методом Multilevel

В рассматриваемом графе методом Multilevel было выделено 6 сообществ.

Была проведена апробация метода выделения сообществ графа Walktrap на рассматриваемом графе (рис.6).

Рис.6 - Результат выделения сообществ в рассматриваемом графе методом Walktrap

В рассматриваемом графе методом Walktrap было выделено 11 сообществ.

Была проведена апробация метода выделения сообществ графа Fastgreedy на рассматриваемом графе (рис.7).

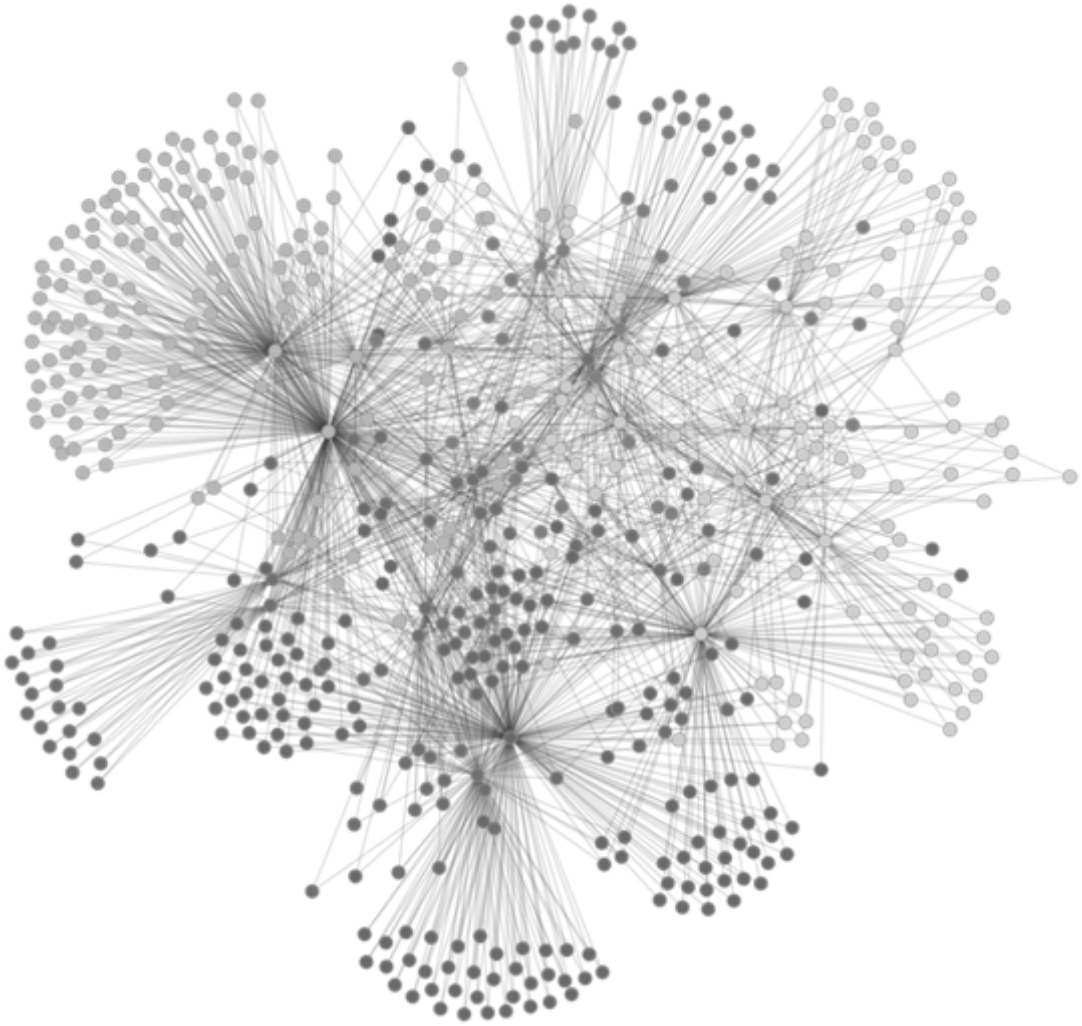

Рис.7 - Результат выделения сообществ в рассматриваемом графе методом Fastgreedy

В рассматриваемом графе методом Fastgreedy было выделено 5 сообществ.

Была проведена апробация метода выделения сообществ графа Edge Betweenness на рассматриваемом графе (рис.8).

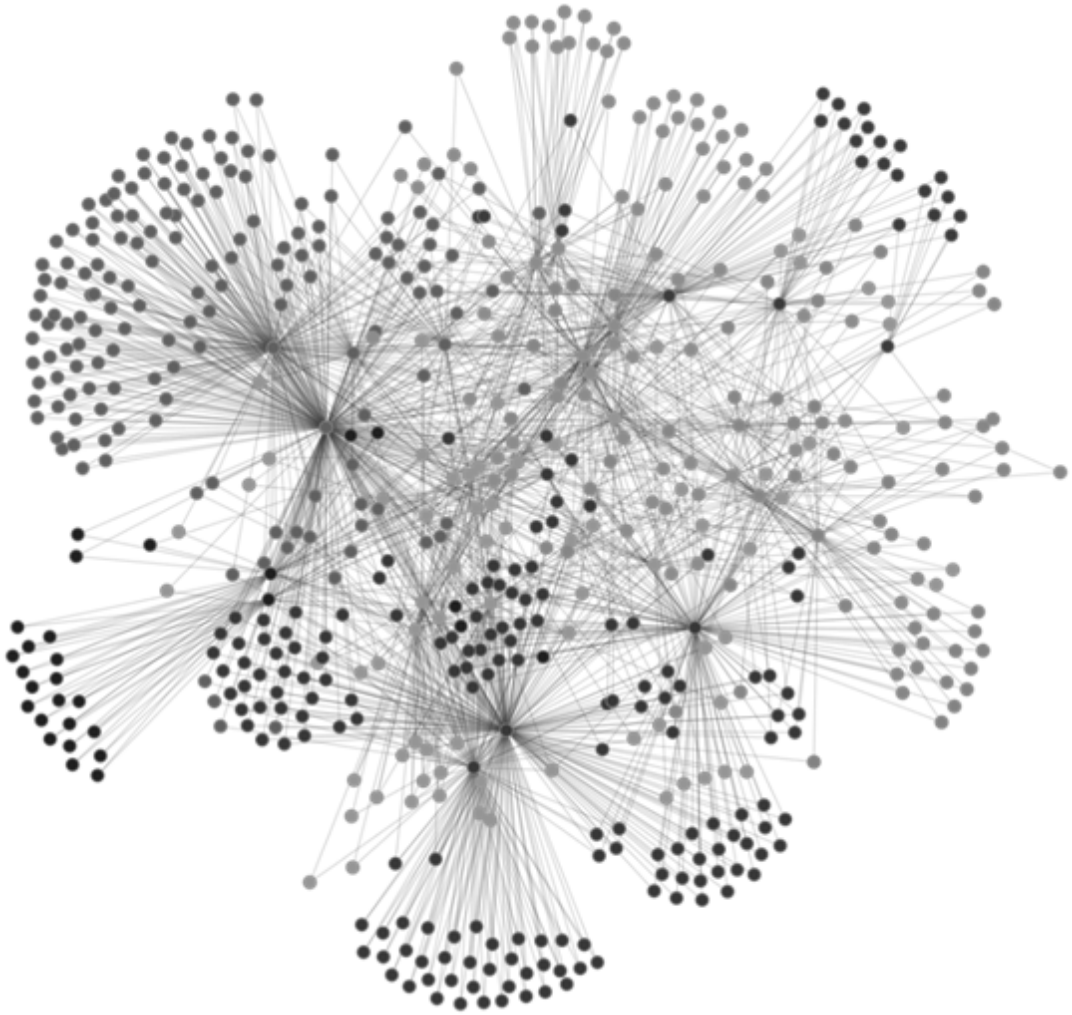

Рис.8 - Результат выделения сообществ в рассматриваемом графе методом Edge Betweenness

В рассматриваемом графе методом Fastgreedy было выделено 8 сообществ.

4 Оценка результатов апробации инструментального средства

В ходе апробации разработанного инструментального средства были получены оценки времени работы методов выделения сообществ в социальных графах (таблица 1). Здесь E - количество вершин в рассматриваемом графе, V - количество связей. Все вычисления производились на базе процессора Intel(R) Core(TM) i5-8250U с 8 Гб оперативной памяти.

Таблица 1 - Оценка времени работы методов на реальных данных и в общем случае.

Название метода	Время выполнения	Оценка в общем случае
Infomap	1.96 s	-
Leading_eigenvector	710 ms	O((E+V)V)
Label_propagation	508 ms	O(E+V)
Multilevel	510 ms	O(E+V)
Walktrap	595 ms	O(EV^2) в худшем случае и O(V^2 log V) в большинстве реальных случаев
Fastgreedy	544 ms	O(Ed log V) где d - это глубина дендрограммы, описывающей структуру графа
Edge_betweenness	1 min 32 s	O((V+E)EV^2)

Методы Infomap, edge_betweenness, fastgreedy, walktrap выполняются на порядок дольше методов label_propagation, multilevel и leading_eigenvector, что окажется существенно на больших наборах данных. В связи с этим рекомендуется использовать эти методы на небольших наборах.

В ходе работы была произведена оценка модулярности графов после кластеризации для апробированного набора данных (таблица 2).

Таблица 2 - Значения модулярности графа в результате кластеризации различными методами

Название метода	Модулярность
Infomap	0.464858
Leading_eigenvector	0.436747
Label_propagation	Метод является нестабильным, зависящим от начального разбиения. Значение модулярности различается на каждом запуске алгоритма.

Название метода	Модулярность
Multilevel	0.467066
Walktrap	0.421945
Fastgreedy	0.434820
Edge_betweenness	0.450384

Таким образом, на данном наборе данных метод Walktrap показал наименьшее значение модулярности, а метод Infomap - наибольшее. Однако, стоит заметить, что все значения модулярности отличаются от среднего значения не более чем на 6%. В связи с этим при выборе методов кластеризации графов для работы с небольшим набором данных следует ориентироваться на время выполнения методов, а также на ожидаемое количество кластеров.

ЗАКЛЮЧЕНИЕ

В ходе данной работы было разработано инструментальное средство кластеризации социального графа, созданного на основе заданного набора групп социальной сети «Вконтакте». Были реализованы следующие методы кластеризации: Infomap, Leading_eigenvector, Label_propagation, Multilevel, Walktrap, Fastgreedy и Edge_betweenness.

Разработанное инструментальное средство было апробировано на наборе данных, сформированном на основе двух групп «Вконтакте». В результате апробации были получены оценки времени выполнения методов на рассматриваемом наборе данных, а также рассчитаны значения модулярности кластеризованного графа для каждого метода.

На основе представленных оценок, а также оценок сложности алгоритмов в общем случае были даны рекомендации по выбору метода для работы с социальным графом, сформированном на основе данных из социальной сети «Вконтакте».

Библиографический список:

1. Чураков А.Н. Анализ социальных сетей // Социологические исследования. 2001. № 1. С. 109-121.
2. Люк Д.А. Анализ сетей (графов) в среде R. - М: ДМК Пресс, 2017.
3. S Rao Chintalapudi , M. H. M. Krishna Prasad Community Detection in Large-Scale Social Networks : A Survey // Graph Theoretic Approaches for Analyzing Large-Scale Social Networks - 2018 - p 189-206.
4. Прикладная статистика и снижение размерности / Айвазян С.А.,. Бухштабер В.М, Енюков И.С., Мешалкин Л. Д, Под ред. Айвазяна С.А. - М: Финансы и статистика, 1989 - С. 641.
5. Pascal Pons, Matthieu Latapy Computing communities in large networks using random walks // Physics and Society. - 2005.
6. Clauset A., Newman M.E.J., Moore C. Finding community structure in very large networks // Physical Review E. - 2004. - №70.
7. Raghavan U.N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks // Physical Review E. - 2007. - №76.
8. Blondel V.D., Guillaume J-L, Lambiotte R., Lefebvre E. Fast unfolding of community hierarchies in large networks // Physical Review E. - 2008.
9. Newman M.E.J. Finding community structure in networks using the eigenvectors of matrices // Physical Review E. - 2006. - №74.
10. Rosvall M., Bergstrom C. T. Maps of information flow reveal community structure in complex networks // Proceedings of the National Academy of Sciences. - 2008. - №105(4).
11. Rosvall M., Axelsson D., Bergstrom C. T. The map equation // The European Physical Journal Special Topics – 2009. - 178, 13.

Оригинальность 73%