Санкт–Петербургский Государственный Университет

Поляков Иван Михайлович

Отчёт по домашнему заданию №2

«Поиск структур в сети»
«Выделение сообществ в графе vk»
«Анализ графа IMDB»

Направление 01.04.02: «Прикладная математика и информатика»
Образовательная программа ВМ.5505.2021: «Математическое и информационное обеспечение экономической деятельности»

Руководитель:
кандидат физ.-мат. наук,
доцент Воронкова Ева Боруховна

Санкт-Петербург
2022 г.

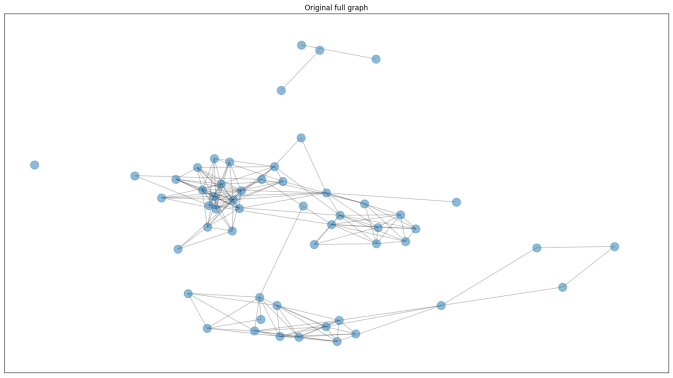
Постановка задания

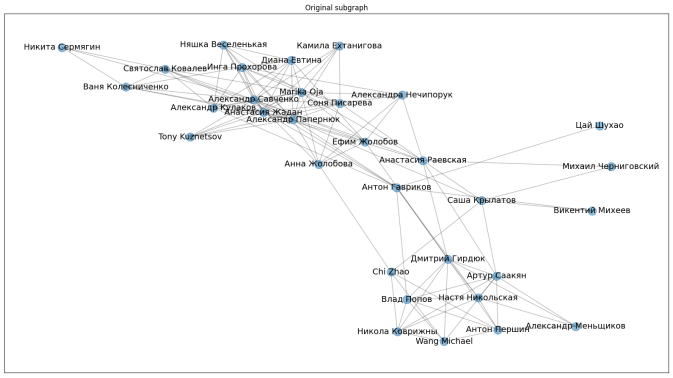
1. Поиск структур графа (для максимальной компоненты связности графа друзей ВКонтакте)
2. Выделение сообществ в графе ВКонтакте;
3. Анализа графа IMDB.

Поиск структур графа

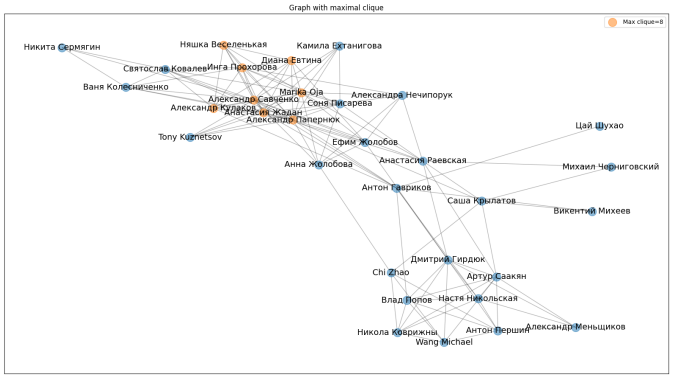
Задание 1

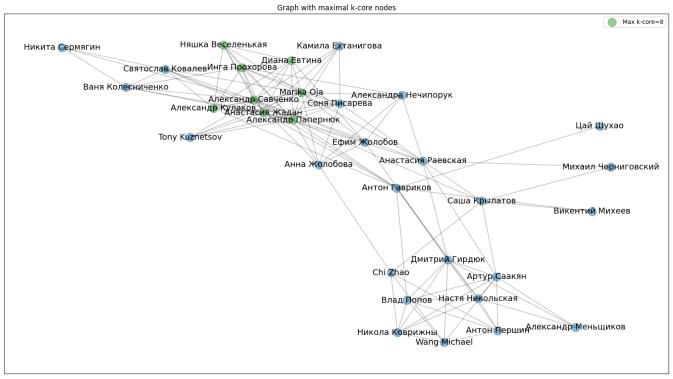
Граф *vk* и его максимальная компонента связности выглядят следующим образом:





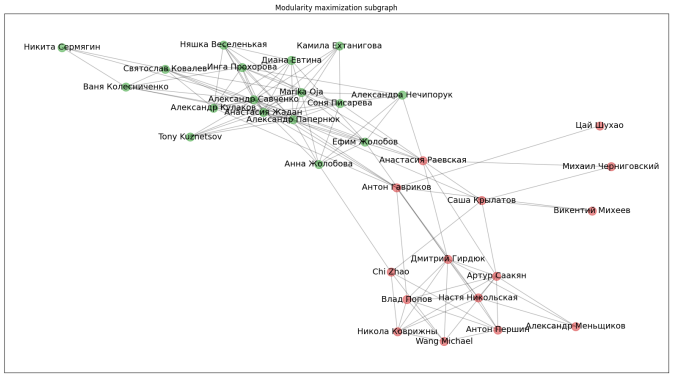
Полный подграф K_n (максимальная клика) и максимальное k-ядерное число совпадают - они равны 8.



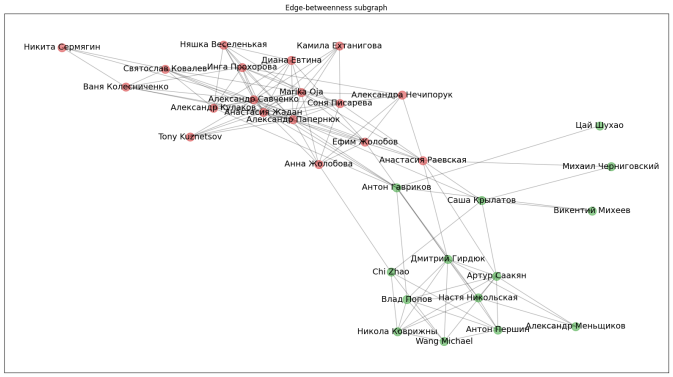


Задание 2

Для той же компоненты связности были найдены разбиения двумя способами. Разбиение, найденное с помощью максимизации меры модулярности и с помощью метода *Edge-Betweenness*:



Мера модулярности: 0.3672



Мера модулярности: 0.3534

Текущий граф был разбит на 2 сообщества – «Университет» и «Работа». Сами разбиения различаются лишь на одну вершину. Это объясняется тем, что полученные сообщества имеют много связей между собой – так сложилось, что всего 5 человек из этого графа не имеют отношения к Университету, но имеют отношение к Работе. В то же время, большая часть людей, имеющая отношение к Работе, училась (или продолжает учиться) в Университете.

Можно проводить улучшение разбиения на вершины, применяя алгоритмы выделения сообществ повторно к уже найденным. Однако с формальной точки зрения для данной конкретной сети это делать не имеет смысла, так как изначальные сообщества довольно тесно между собой связаны, то есть, оптимального разбиения существовать не будет.

Задание 3

В данном задании рассматривалось 2 графа IMDB: данный по заданию и скачанный с сайта https://www.imdb.com/interfaces/. Так как графы представляют собой большое количество данных, то анализ этих графов начинается с подграфов на 1000 вершин, которые были построены двумя методами: случайным и "снежным комом". Исходные полученные подграфы для обоих видов данных представлены ниже.

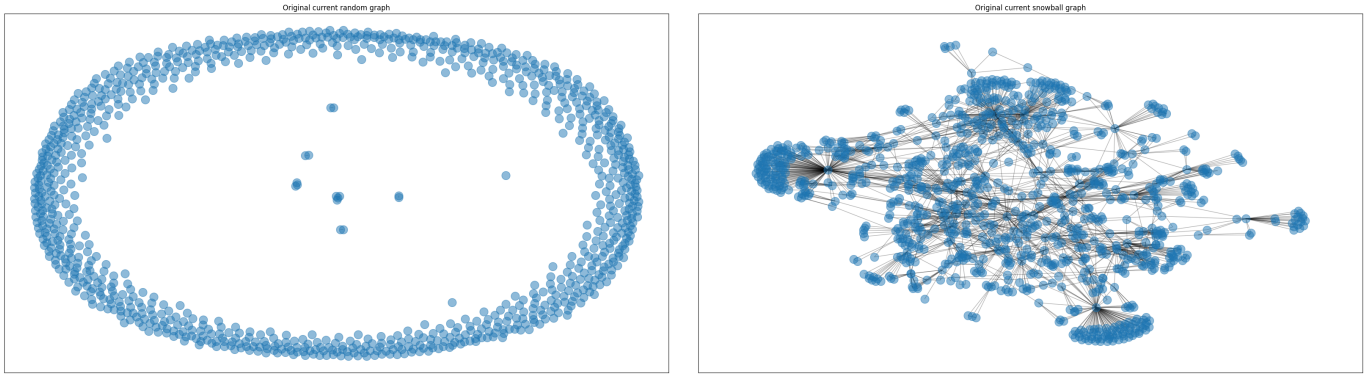


Рис. 1: actors_costar.edges

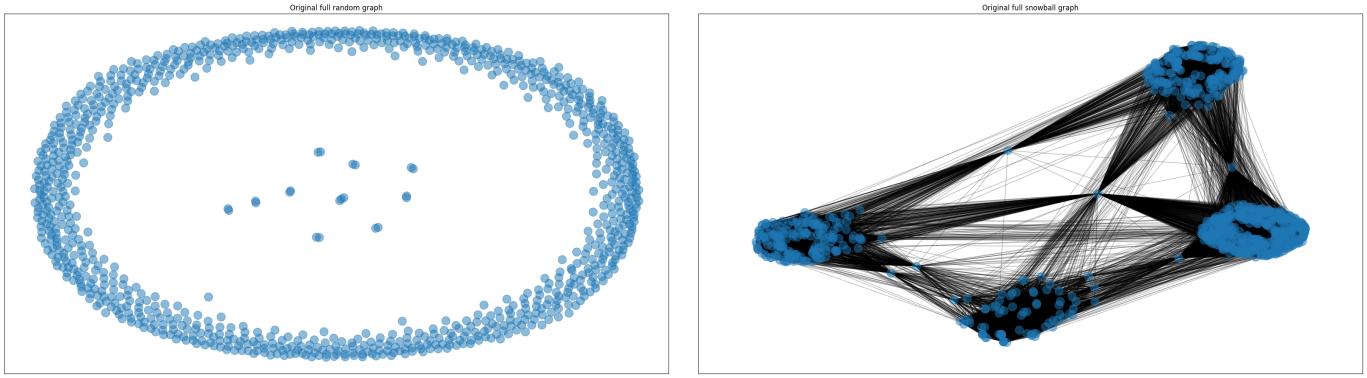
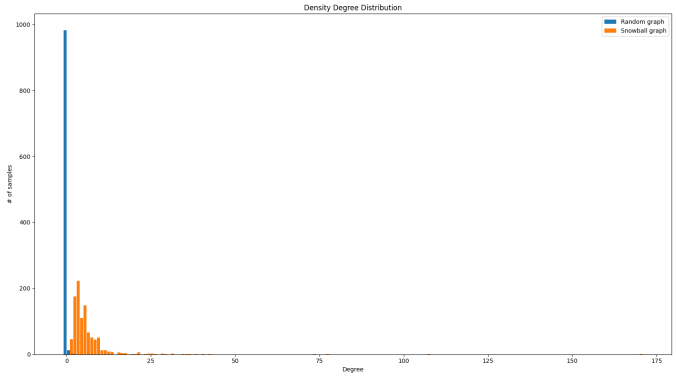


Рис. 2: name.basics.tsv.gz

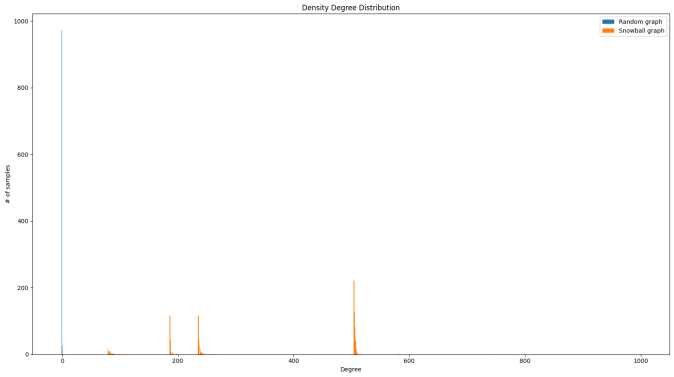
Как видно, случайные графы практически ничем не отличаются - крайне мало связей. Однако «снежный ком» отличается: в случае предоставленного графа имеются огромные разветвления, однако в данных на официальном сайте явно визуально выделяются 4 сообщества. Это обусловлено тем, что был выбран известный актёр (*Samuel L. Jackson*) и что во всём графе огромное число вершин (примерно 12млн). При составлении этого «снежного кома» не был сделан переход к соседям, так как число актёров хватило для составления данного графа. То есть, существует вершина со степенью 999, что и отражено в гистограммах распределения вероятностей степеней вершин. 4 выделенных сообщества - это фильмы.

Плотности полученных подграфов и средние длины путей отражены в следующей таблице:

Данные	Плотность	Средняя длина*
actors_costar.edges Random	0.00002202	0.0077
actors_costar.edges Snowball	0.00562162	4.2531
name.basics.tsv.gz Random	0.00002803	0.0135
name.basics.tsv.gz Snowball	0.35291491	1.6471



actors_costar.edges

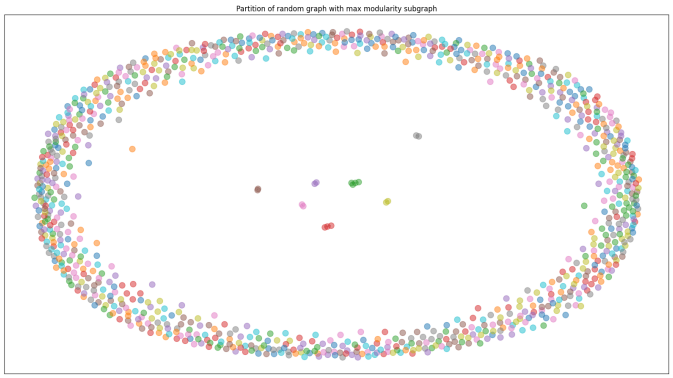


name.basics.tsv.gz

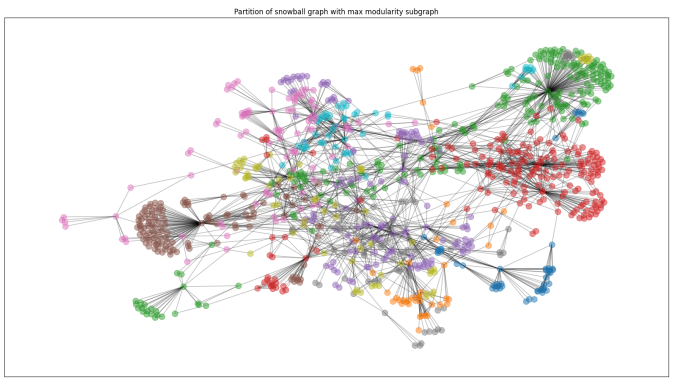
Рис. 3: Распределение вероятностей степеней вершин

Средняя длина в случайных графах рассчитывалась как среднее длин каждой компоненты связности.

В связи с большим числом рёбер в графе из данных *name.basics.tsv.gz*, то разбиение считается чересчур долго (3300 часов). Тем не менее, для данных *actors_costar.edges* для разбиения использовалась максимизация меры модулярности, результаты приведены ниже.



Случайный граф



Снежный ком

Рис. 4: Разбиение на сообщества