

Исследование научного сотрудничества в Новой Зеландии

Работу выполнили студенты группы Э301:
Кабанов Илья, Сухов Александр, Сысоев Никита

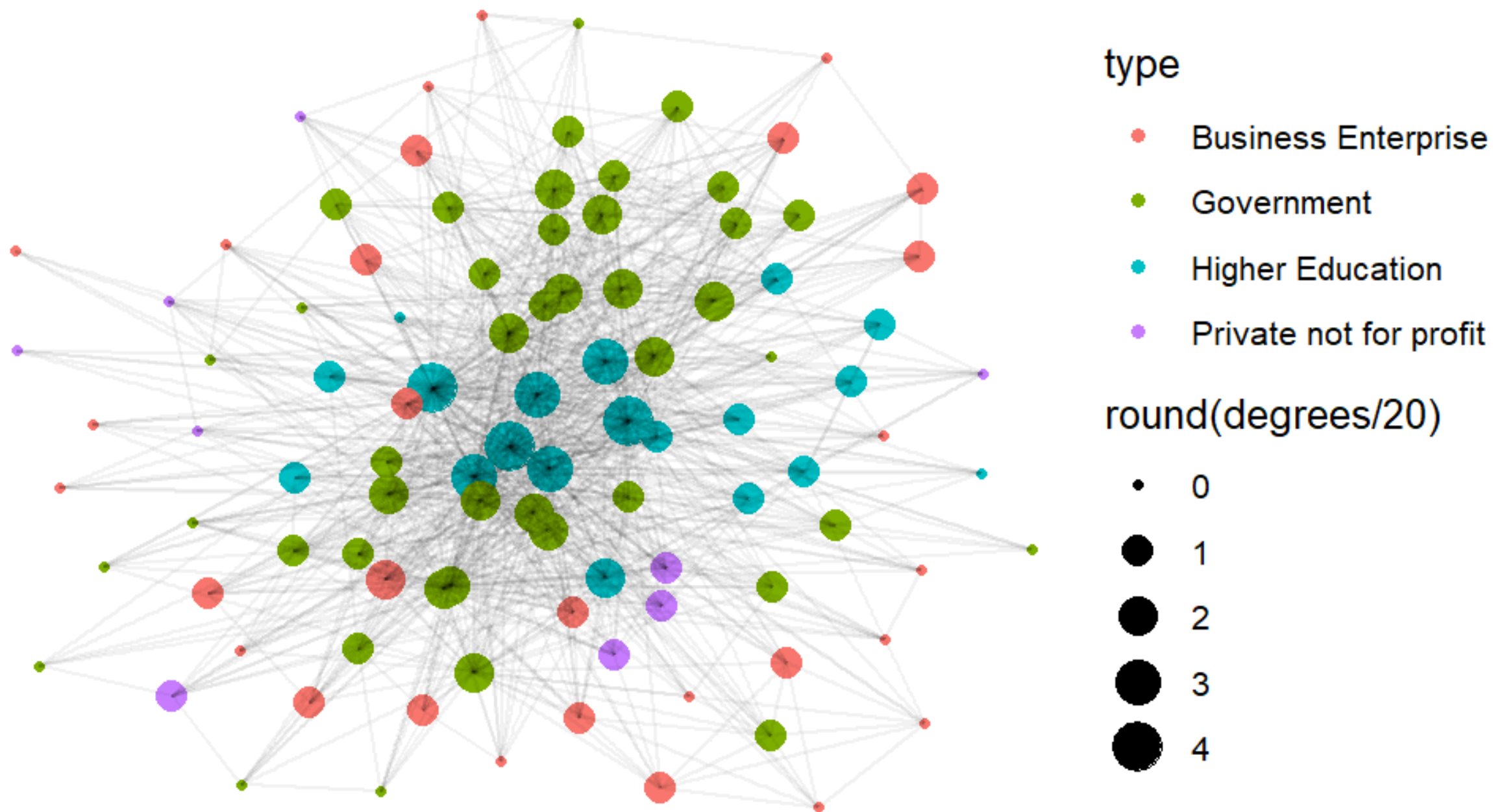
Описание данных

- Вершины - организации Новой Зеландии, которые занимаются написанием научных статей в качестве одного из видов деятельности.
- Качественной характеристикой вершин является вид организации (государственная организации, высшее учебное заведение, коммерческое предприятие и частное некоммерческое предприятие).
- Между i -ой и j -ой вершинами (организациями) существует ребро, если за период 2010-2015 гг. на Scopus была опубликована хотя бы одна статья, хотя бы один автор которой числится в организации i , а другой автор – в организации j .
- Вес ребра отображает силу связи, а именно количество совместных опубликованных научных работ.

Исследовательский вопрос

- Какие есть способы улучшения взаимодействия учреждений высшего образования Новой Зеландии (если это вообще требуется)?
- Наше исследование показывает, что возможно повысить эффективность небольших университетов посредством увеличения сотрудничества с крупными, которые в свою очередь имеют тесное сотрудничество с бизнесом. (Aref et al, 2017)

Network of New Zealand organisations

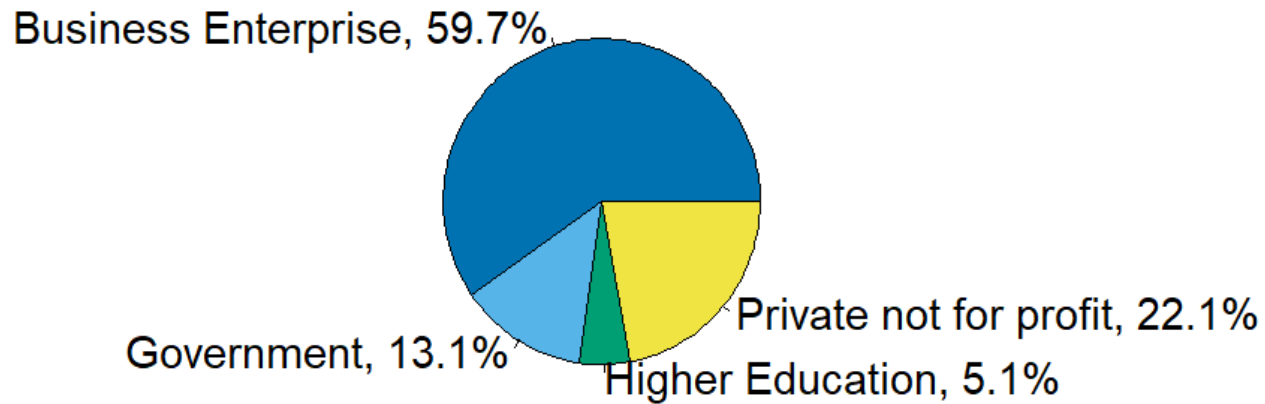


Основные характеристики графа

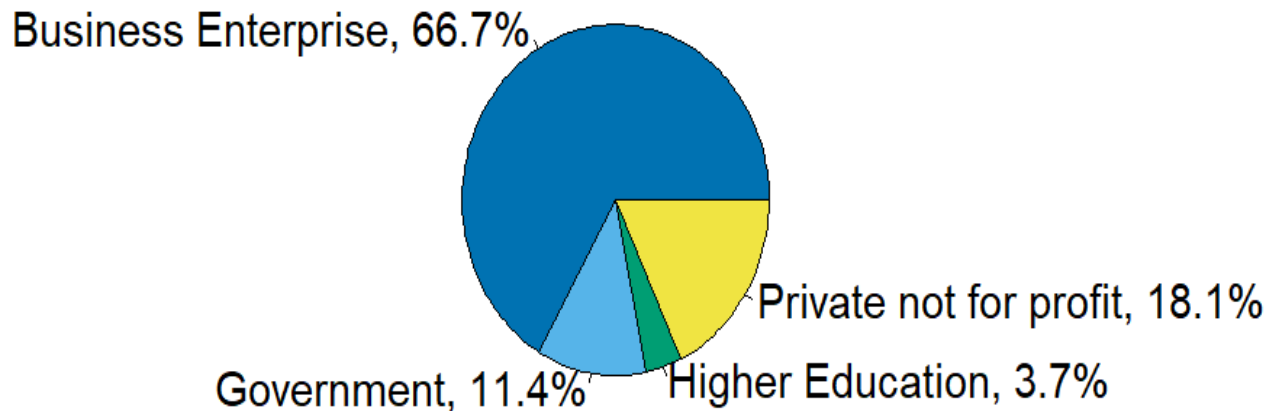
- Вершин: 1511
- Ребер: 4273
- Плотность: 0,0037
- Порог случайного графа: 0.0048
- Диаметр: 6, 4.01 (weighted)
- Компонент связности: 24, из которых 23 из 2-3 вершин
- Средняя длина пути: 2.8, 1.5 (weighted)
- Средняя центральность по степени: 5.65

Соседство Оклендского университета

Neighbors of University of Auckland



Two-step neighbors of University of Auckland

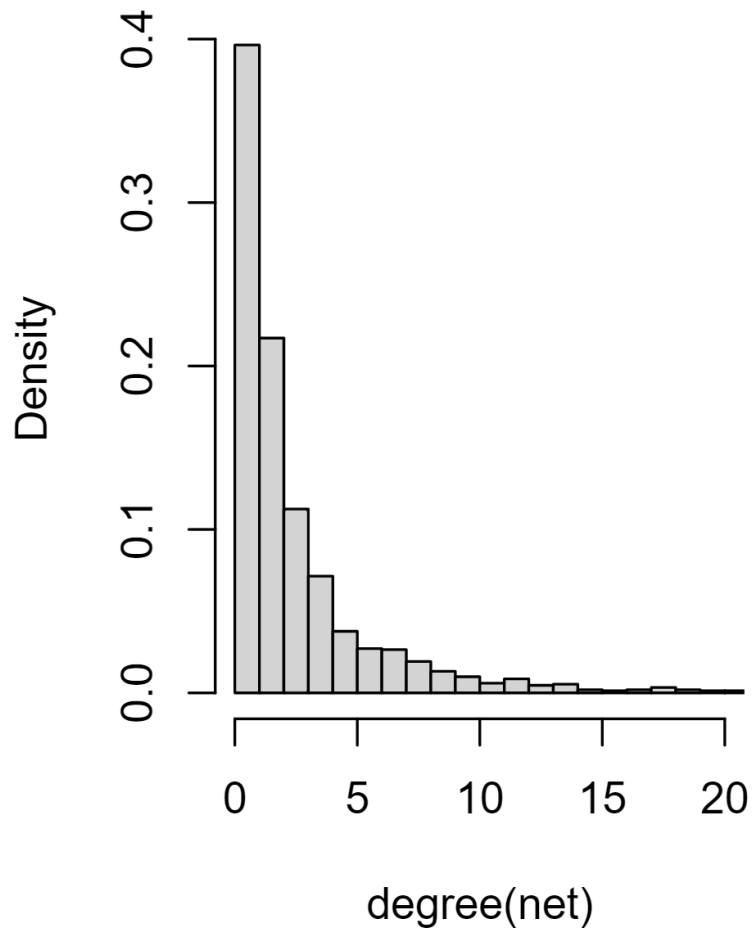


Задание 2

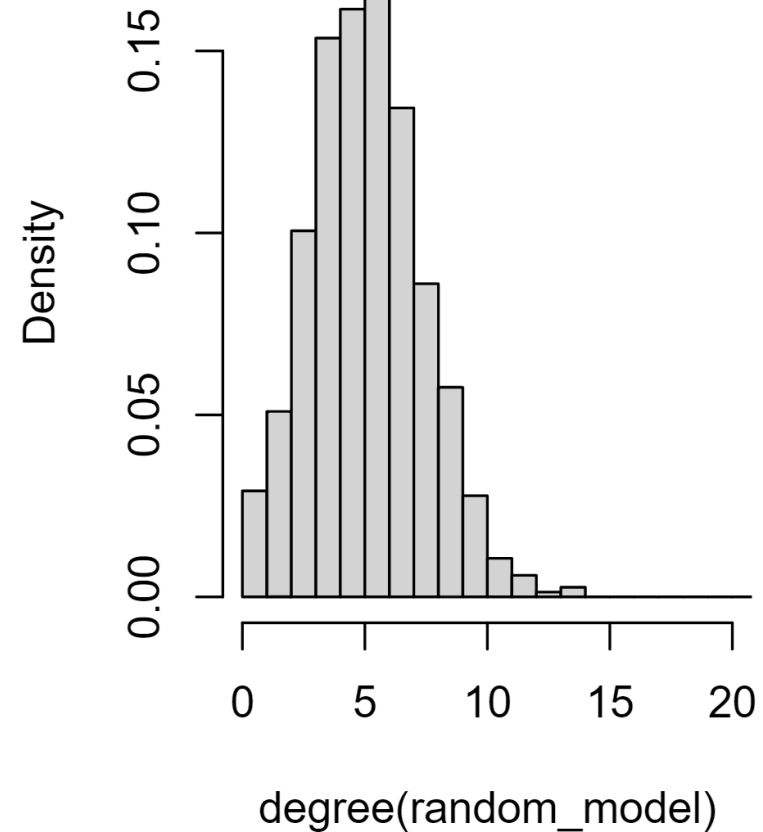
- Сравним нашу сеть с теоретическими моделями случайного графа на основе распределения числа их соседей, диаметра, среднего кратчайшего пути и числа компонент связности.

Модель Эрдеша-Реньи

Histogram of node degree

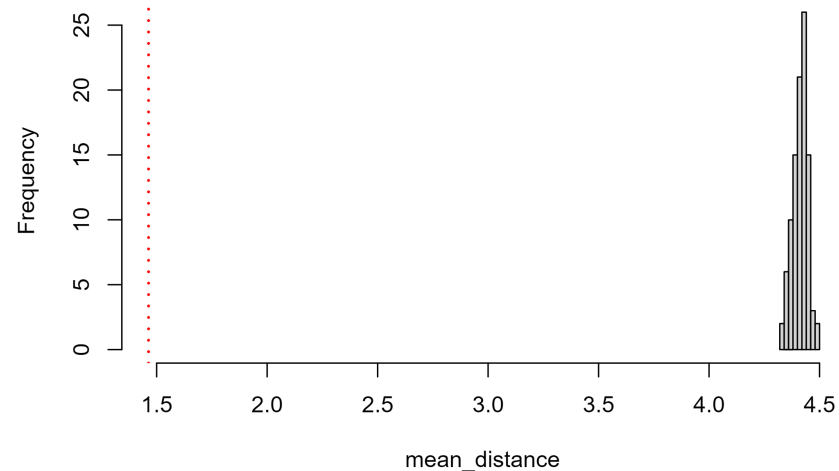


Histogram of node degree

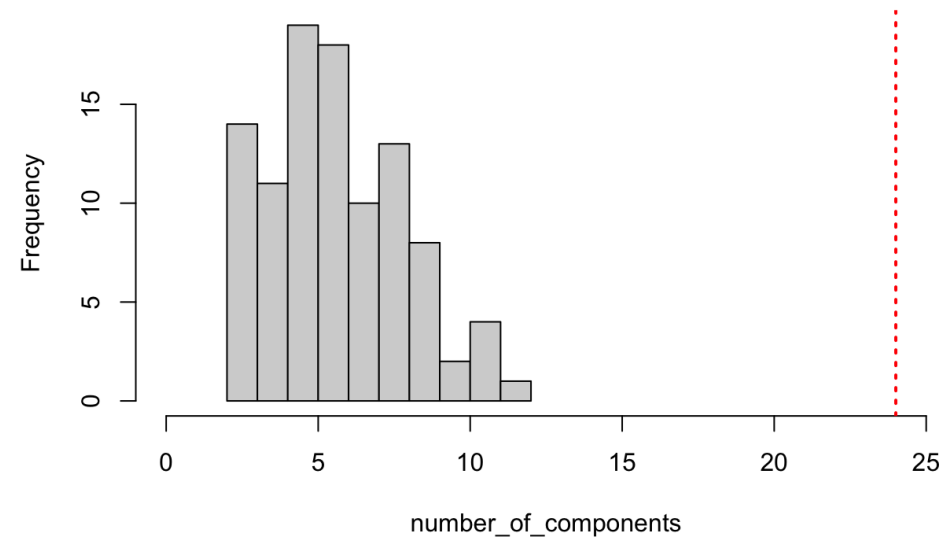


Модель Эрдеша-Реньи

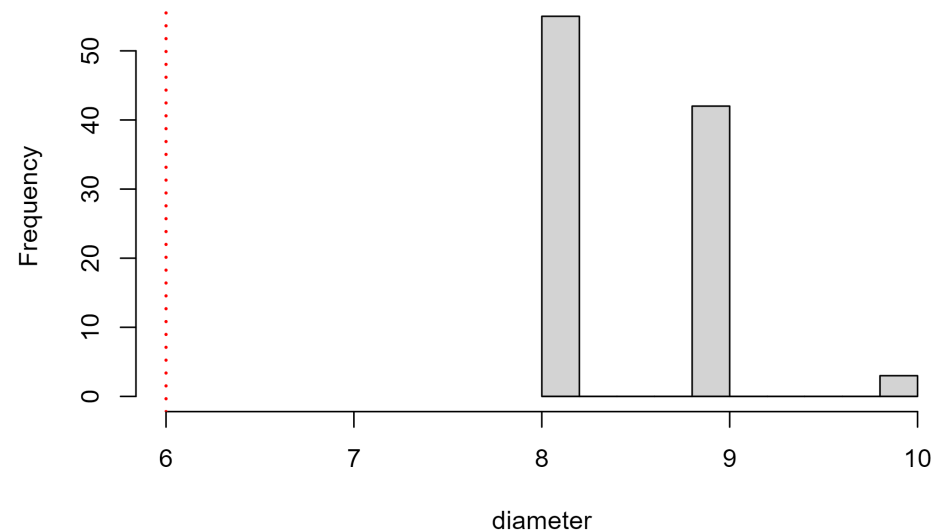
Histogram of mean_distance



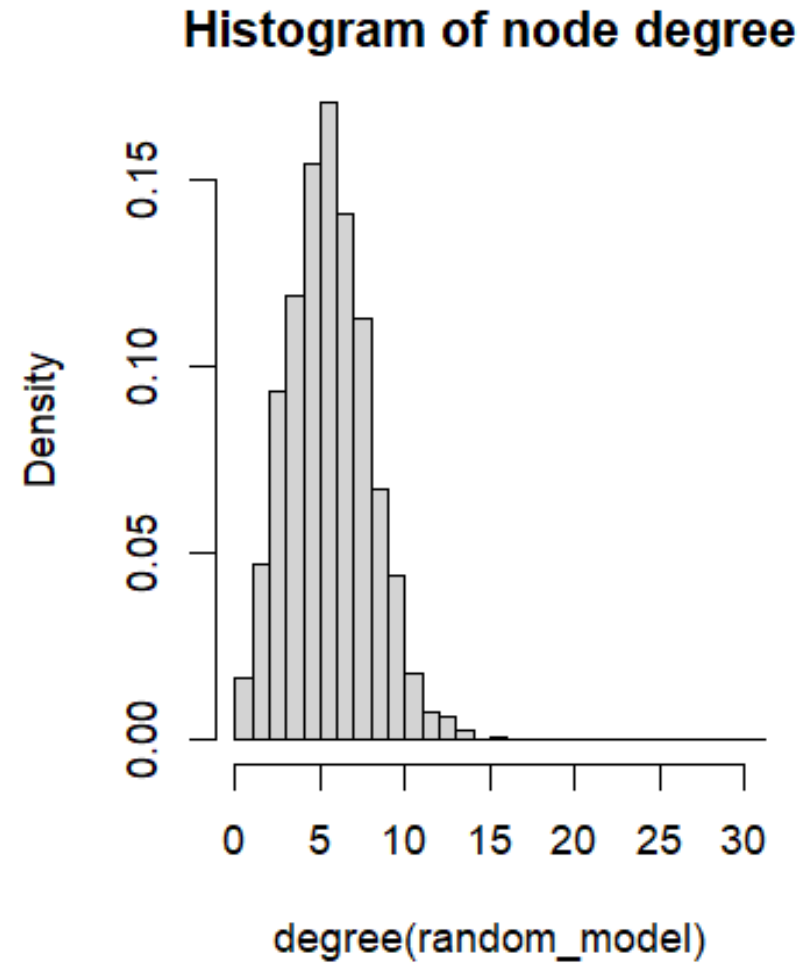
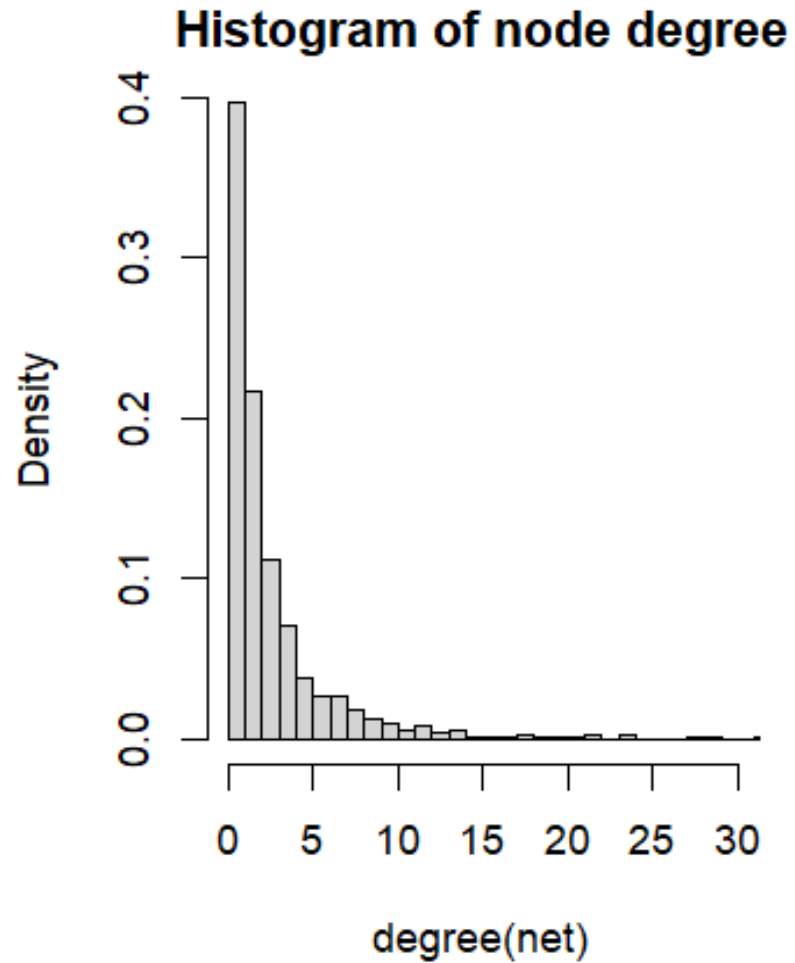
Histogram of number_of_components



Histogram of diameter

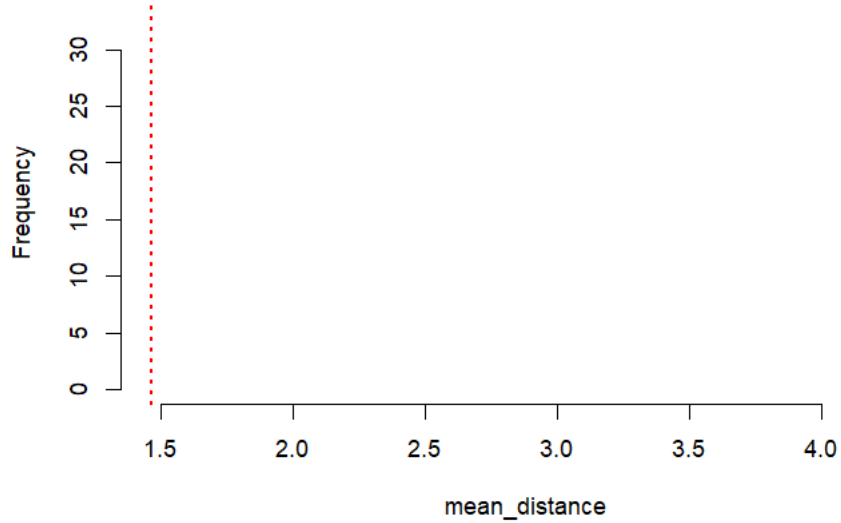


Модель малого мира

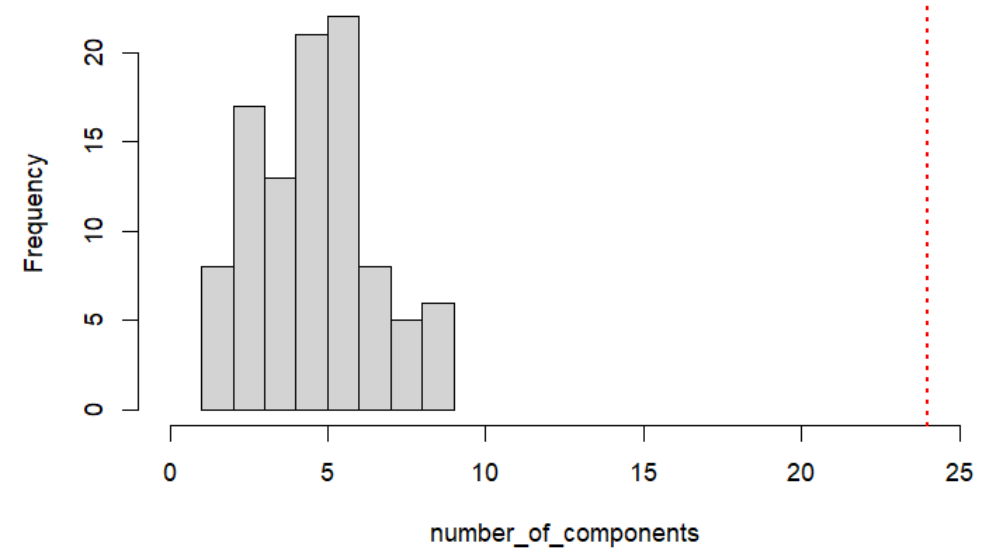


Модель малого мира

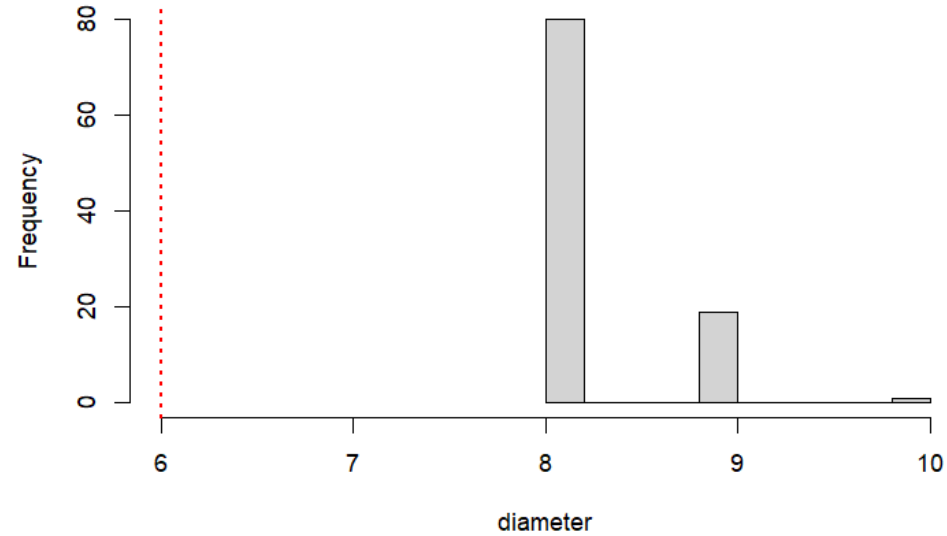
Histogram of mean_distance



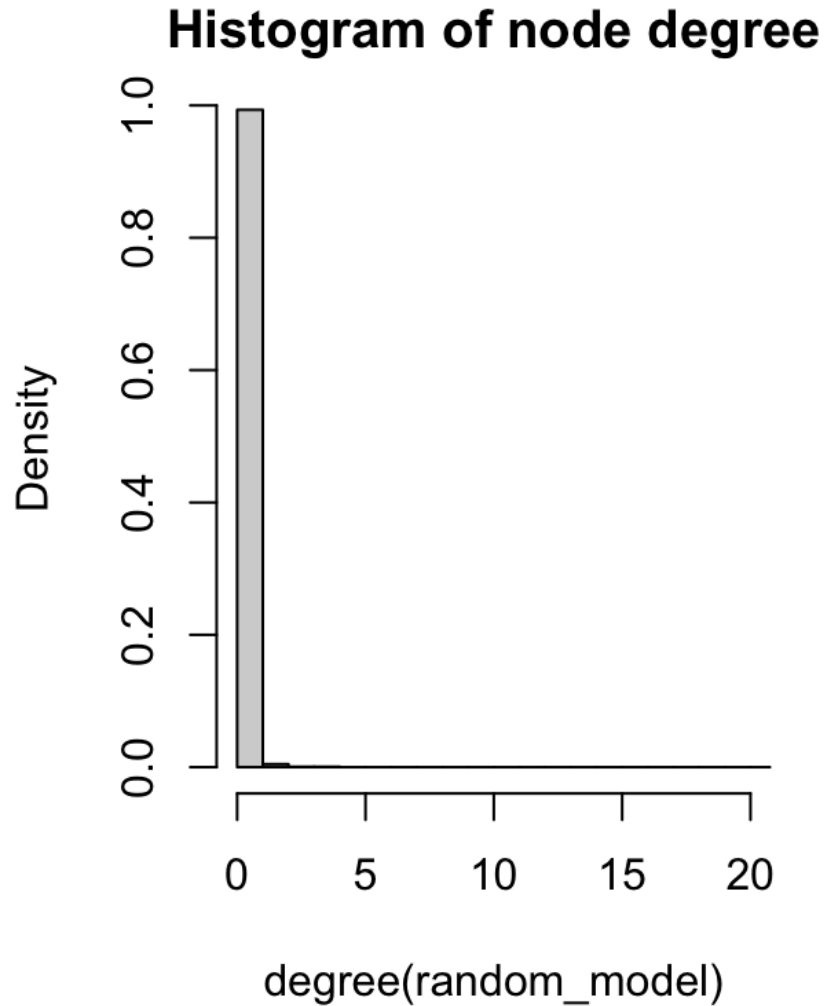
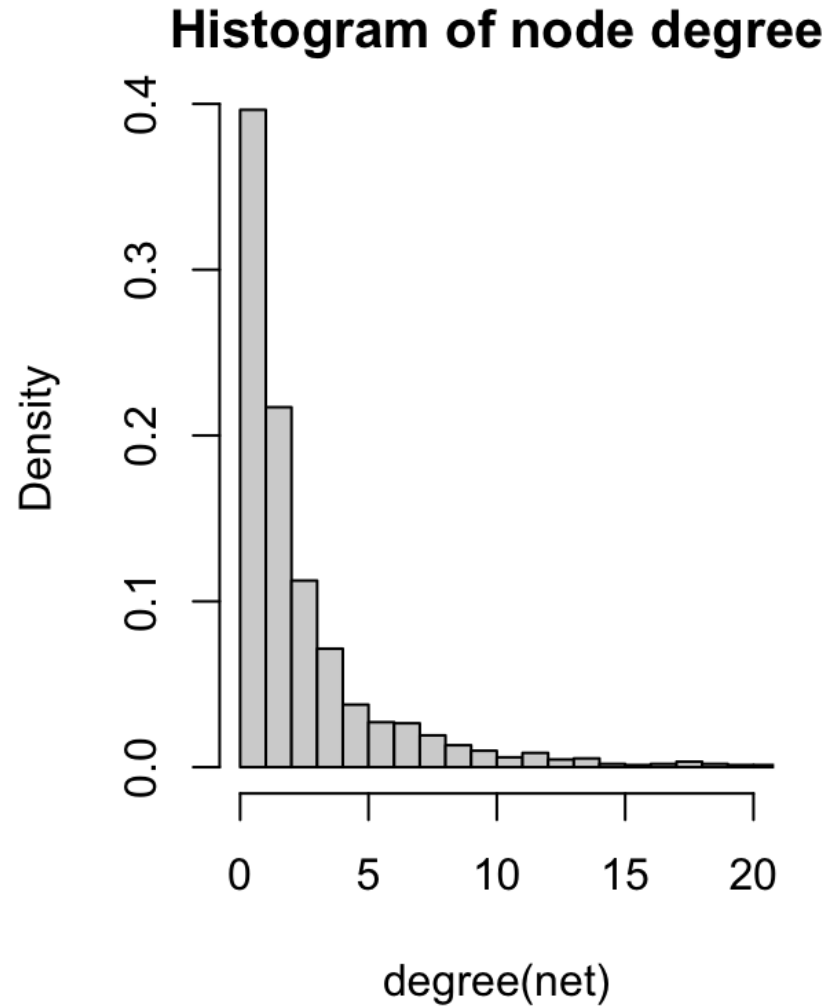
Histogram of number_of_components



Histogram of diameter

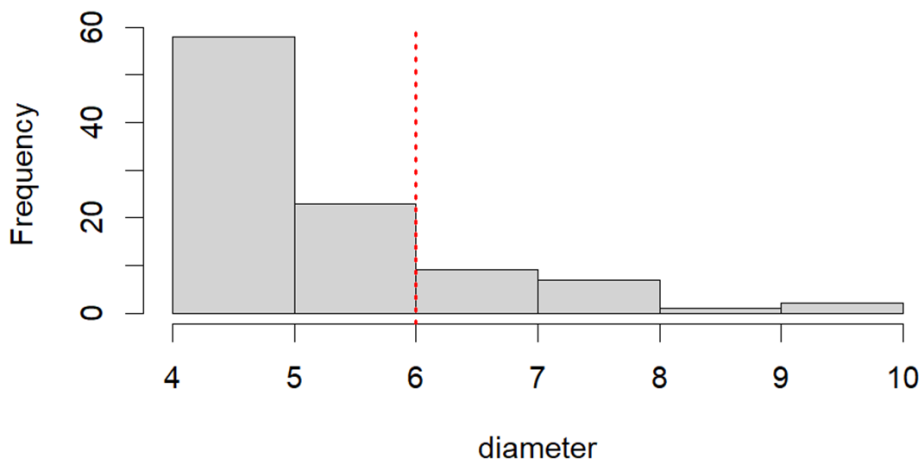


Модель Барабаши-Альберта

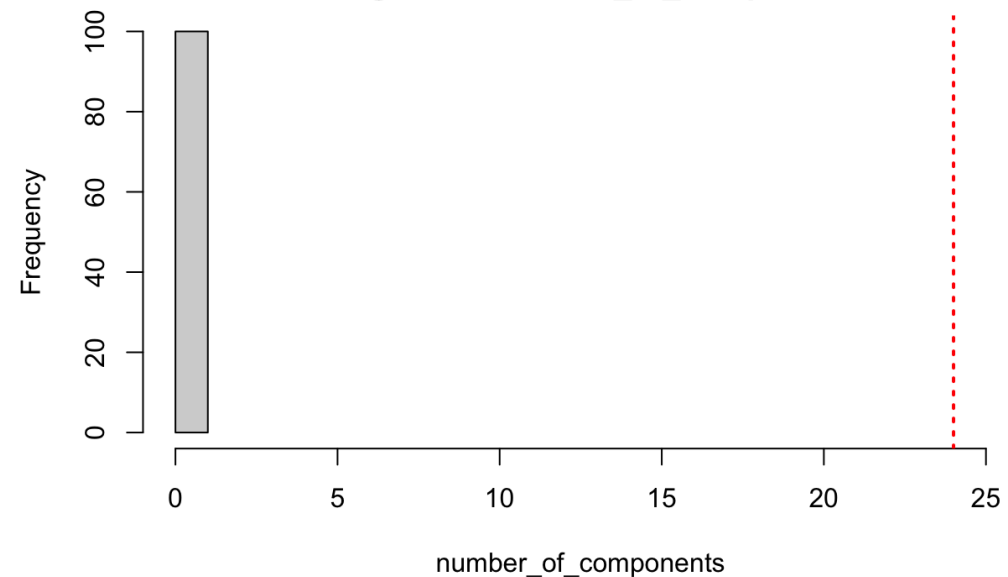


Модель Барабаши-Альберта

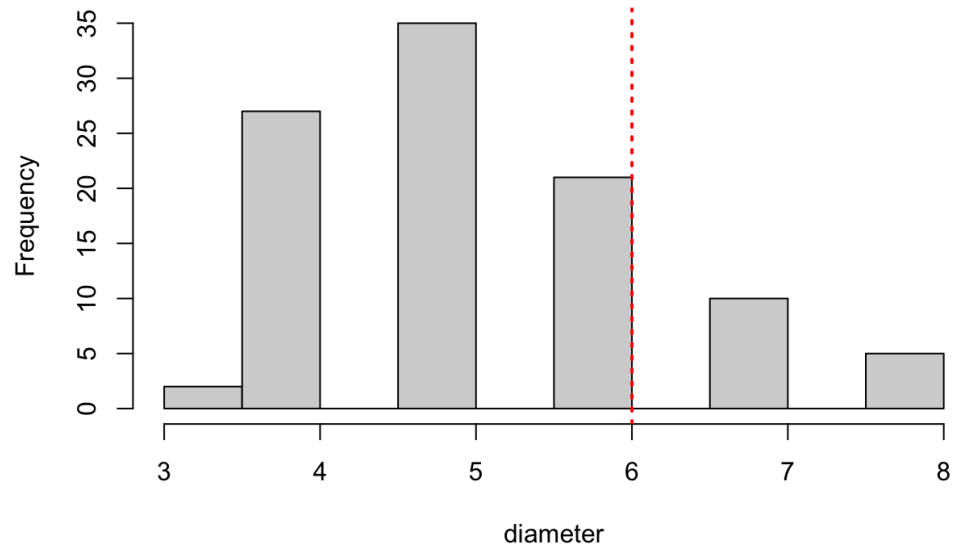
Histogram of diameter



Histogram of number_of_components



Histogram of diameter



Выводы по заданию 2

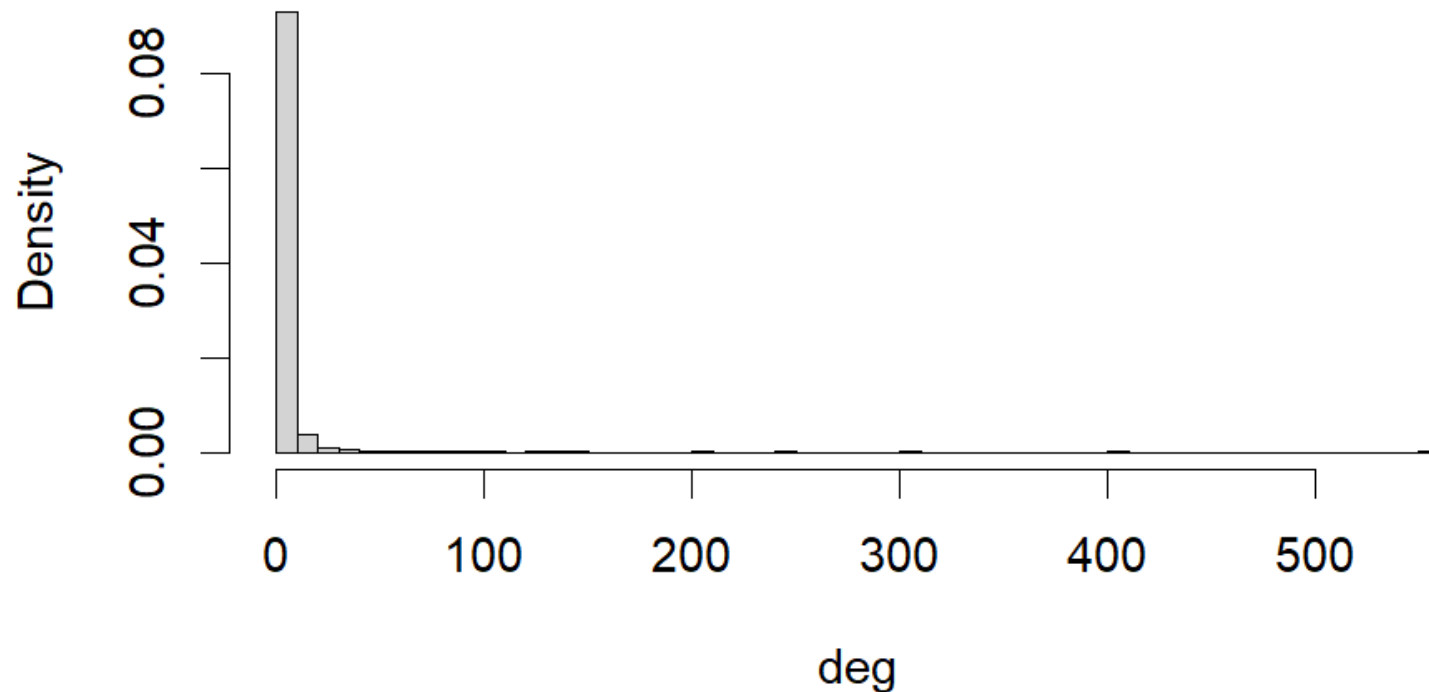
- Ни одна из теоретических моделей, известных нам, не описывает структуры нашего графа достаточно точно.

Задание 3

- Для анализа используем обратные веса, т.к. их использование делает осмысленным расчет центральности по близости и кратчайшему пути, но не искажает остальных мер центральности вершин.

Центральность по степени

Histogram of node degree



- В среднем степень вершины равняется 5,66.
- Модальная степень – 1, встречается 599 раз.
- Медианная степень – 2.

Самые центральные вершины

Canterbury District Health Board	AgResearch	University of Waikato
121	126	136
Auckland District Health Board	AUT University Victoria	University of Wellington
143	145	204
University of Canterbury	Massey University	University of Otago
241	310	406
University of Auckland		
551		

Центральность по кратчайшему пути и по собственному значению

- Значительное число кратчайших путей между любыми двумя вершинами проходит именно через университеты, входящие в топ (в том числе университет Окленда).
- Высокое собственное значение вершин университетов означает, что сами университеты связаны со многими вершинами, которые в свою очередь имеют также высокое значение центральности.

Корреляционные матрицы

	deg	clos1	betw1	eig1.vector
deg	1.00000000	-0.03452649	0.96148582	0.8926645
clos1	-0.03452649	1.00000000	-0.01592647	-0.1091064
betw1	0.96148582	-0.01592647	1.00000000	0.7596773
eig1.vector	0.89266448	-0.10910642	0.75967732	1.0000000

Без учета весов

	deg	clos2	betw2	eig2.vector
deg	1.00000000	-0.03349182	0.90929438	0.8418800
clos2	-0.03349182	1.00000000	-0.01198628	-0.1332681
betw2	0.90929438	-0.01198628	1.00000000	0.7737961
eig2.vector	0.84188004	-0.13326806	0.77379607	1.0000000

С учетом весов

Сравнение топов по центральностям

names.top_10_degree. <chr>	names.top_10_betw1. <chr>	names.top_10_eig1. <chr>
Canterbury District Health Board	Auckland District Health Board	AgResearch
AgResearch	Plant and Food Research	Auckland District Health Board
University of Waikato	AgResearch	Canterbury District Health Board
Auckland District Health Board	University of Waikato	University of Waikato
AUT University	AUT University	AUT University
Victoria University of Wellington	Victoria University of Wellington	Victoria University of Wellington
University of Canterbury	University of Canterbury	University of Canterbury
Massey University	Massey University	Massey University
University of Otago	University of Otago	University of Otago
University of Auckland	University of Auckland	University of Auckland

Выводы по заданию 3

- Сравнив топы по максимальным центральностям (за исключением топа центральностей по близости из-за малого количества соседей порядка 1), заметим, что по трем центральностям первую пятерку занимают одни и те же организации: университет Окленда, университет Отаго, университет Мэсси, университет Кантербери и Викторианский университет Веллингтона.

Задание 4

- Ассортативность в графе по типу вершин = -0.059 => мы не можем утверждать наличие предпочтительного присоединения узлов сети к своему типу.
- Ассортативность по степени вершины = 0.336 => предпочтительное присоединение узлов к вершинам с относительно большим количеством соседей.

Используемый метод

- Для выявления более точного эффекта предпочтительного присоединения среди университетов рассчитаем показатели гомофилии для университетов и других типов институтов.

Результаты расчётов

- Соберем полученные значения dyadicity и heterophilicity вместе:
 - Higher Education vs все остальные: dyadicity: 32.09, heterophilicity: 6.96
 - Higher Education vs Business Enterprise: dyadicity: 39.59, heterophilicity: 6.88
 - Higher Education vs Government: dyadicity: 2.84, heterophilicity: 0.49
- Мы видим, что dyadicity университетов очень высокая. При этом heterophilicity также довольно сильно превышает 1.

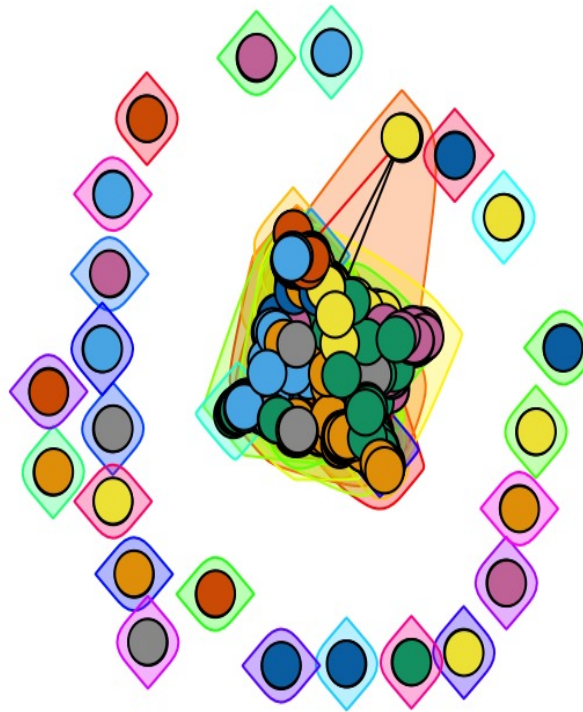
Выводы по заданию 4

- В чем может быть объяснение этой особенности данных?
“Functional dyadicty and heterophilicity of gene-gene interactions in statistical epistasis networks” -> университеты хорошо связаны друг с другом и с вершинами другого типа.
- Плотные связи: университеты – университеты и университеты – бизнес.

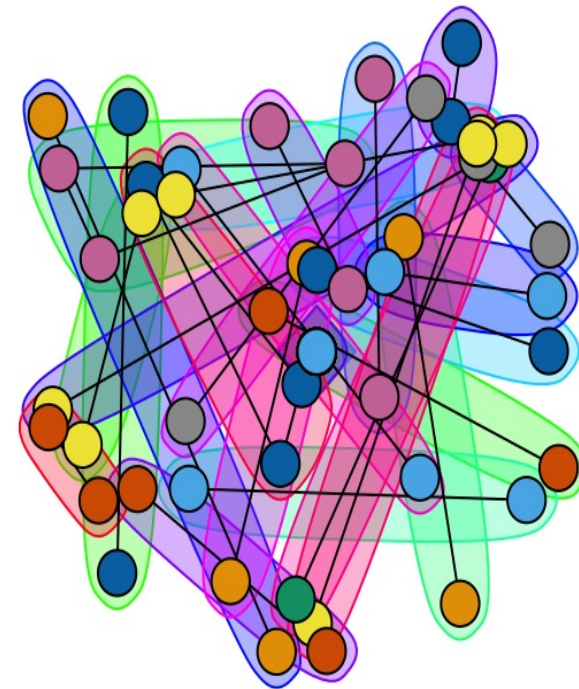
Задание 5

- Проведем процедуру кластеризации вершин при помощи Edge-betweenness и Fast-greedy методов и сравним полученные результаты с существующей классификацией.
- Обратные веса

Визуализация кластеризации методом Edge-betweenness

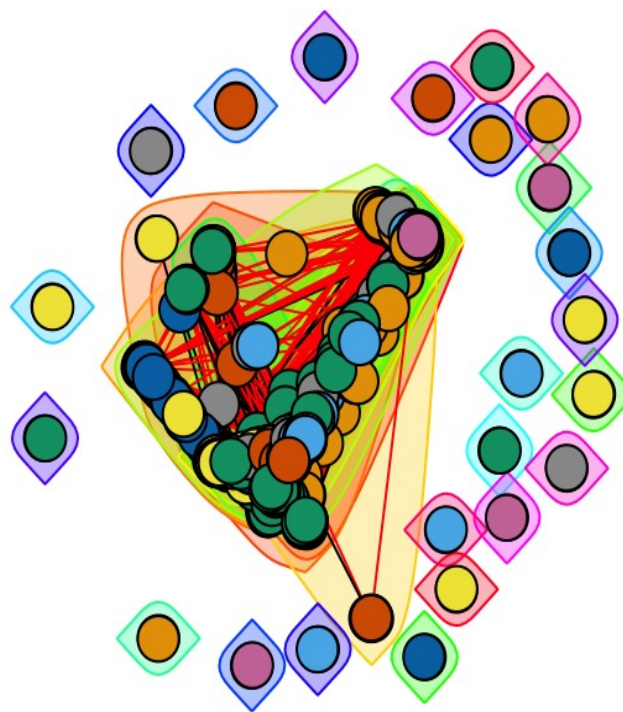


Метод Edge-betweenness

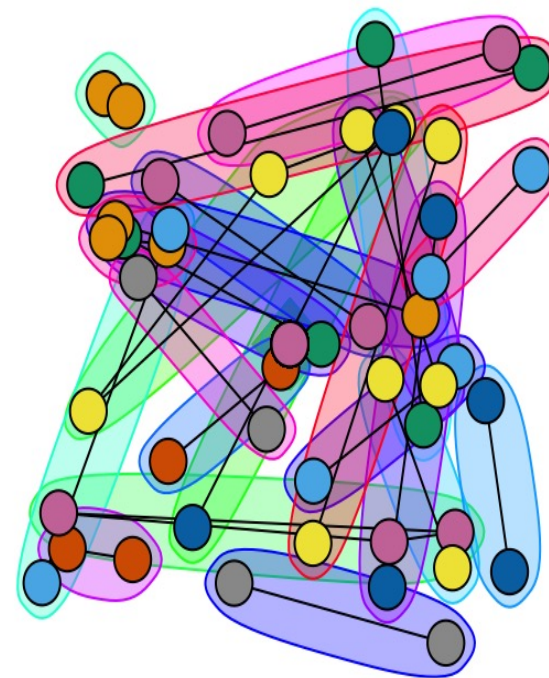


0,03

Визуализация кластеризации методом Fast-greedy



Метод Fast-greedy



0,03

Выводы по заданию 5

- Известные нам способы кластеризации (Edge-betweenness и Fast-greedy) не дали результата, похожего на имеющуюся в датасете классификацию.

Задание 6

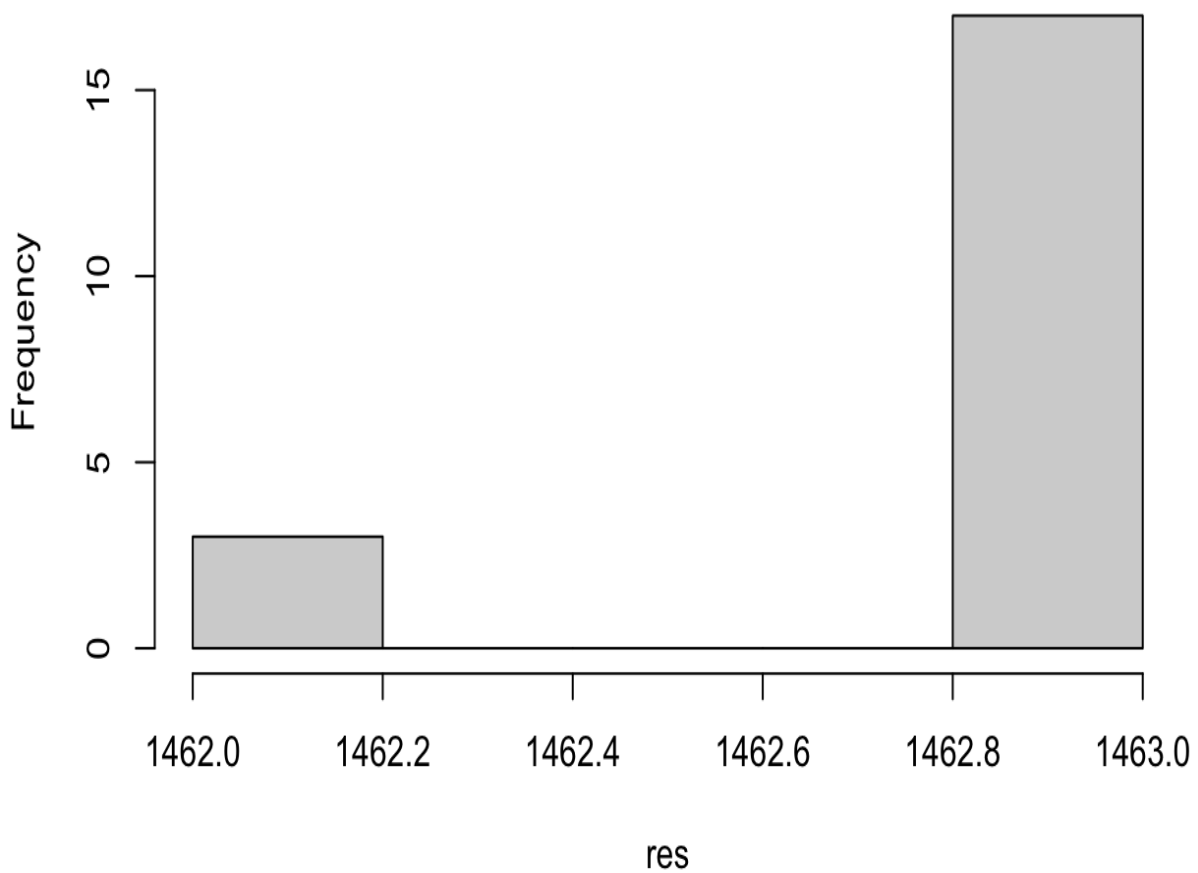
- Исследовательский вопрос: насколько активно при написании научных работ университеты Новой Зеландии взаимодействуют с другими типами организаций, нужно ли им как-то в этом способствовать извне.
- Для ответа на заданный вопрос нам может помочь модель предсказания класса вершин *Relational Neighbors Classifier*

Используемый метод

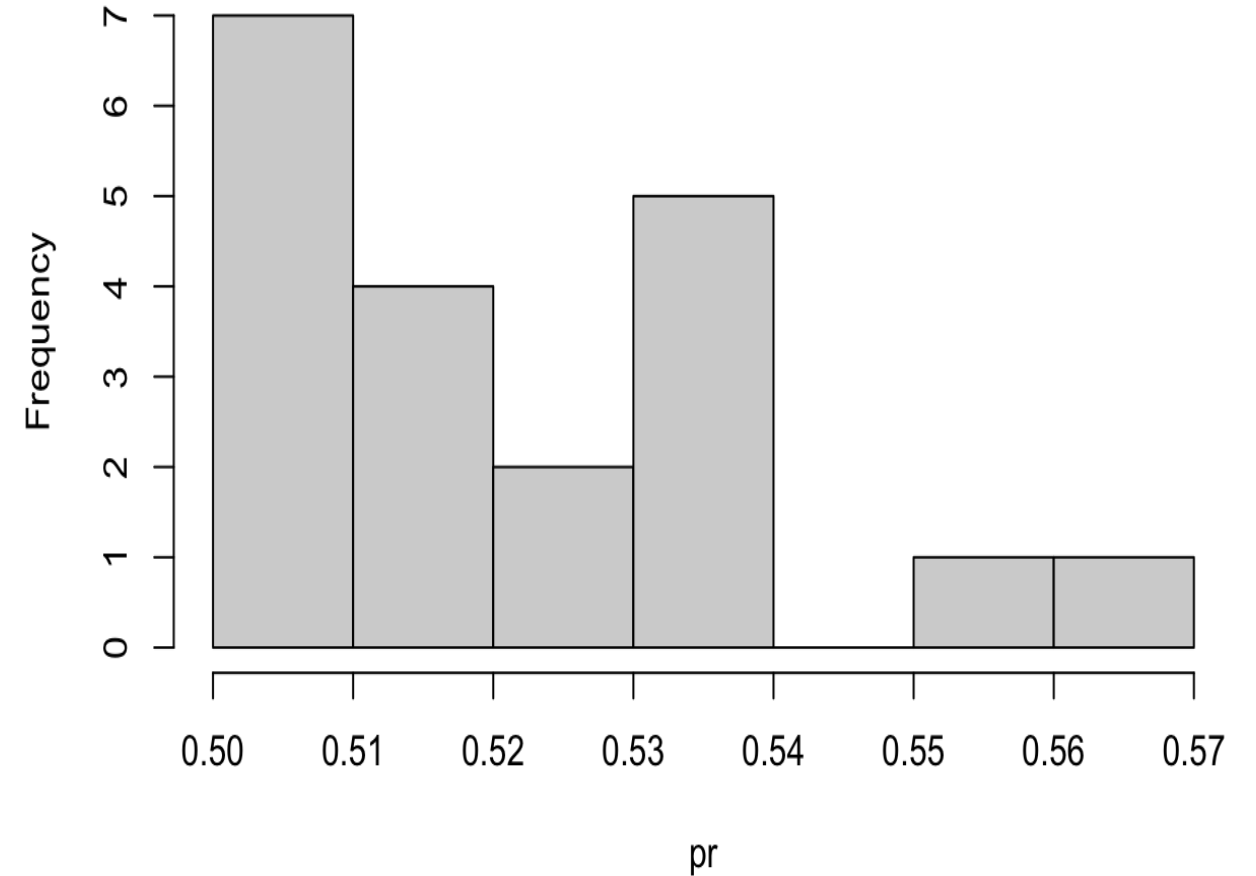
- Разделим все вершины на 2 класса: “1” - университет, “0” - иначе. Отберем из каждого класса по 10 вершин и положим вероятности их принадлежности первому классу соответственно 1 и 0. Для всех остальных вершин положим эту вероятность равной 0.5. При каждой итерации будем обновлять вероятности каждой вершины исходя из вероятностей её соседей. В итоге, после 10 таких итераций получим для каждой вершины вероятность её принадлежности классу “1”, и отберем те вершины, чья вероятность больше порога 0.5 в качестве принадлежащих классу “1”.
- Во избежание случайности повторим этот алгоритм 20 раз и построим гистограмму распределения числа предсказанных университетов, а также гистограмму распределения средних вероятностей принадлежности группе университетов.

Результаты алгоритма

Число предсказанных университетов



Средняя вероятность принадлежности классу 1



Выводы по заданию 6

- Как можно увидеть, почти во всех случаях абсолютное число вершин предсказано в качестве университетов. Получаем, что какие бы 10 университетов мы ни взяли, их класс дойдет до большинства вершин быстрее, чем противоположный класс 10 случайных вершин. Это распространяется и на случай, когда были выбраны наименее центральные университеты. Если бы университеты были плохо связаны с организациями другого типа, то их класс намного слабее распространялся бы на вершины другого типа, а это не так.

Выводы по заданию 6

- Значит, по крайней мере какие-то университеты (а именно наиболее центральные) довольно активно кооперируются с другими организациями (как, например, это делает Оклендский университет). Поэтому для таких университетов нет необходимости строить совместные с бизнесом и государством научные центры.
- Вспомним про высокий показатель dyadicity у ВУЗов: выходит, что он получается таким большим из-за высокой гомофилии более мелких учебных заведений. Тогда можно предположить, что именно они плохо взаимодействуют с бизнесом и гос. органами.

Ответ на исследовательский вопрос

- При моделировании класс ВУЗов распространялся по сети через их соседей — крупные университеты. В таком случае, более разумно строить научные центры на базе крупных университетов с активной вовлеченностью мелких, так как сами по себе они плохо связаны с организациями других типов.
- Замечание: Здесь мы предполагаем эталонность крупных университетов (например, Оклендского) с точки зрения эффективности взаимодействия, а также качества и количества научных работ, что продемонстрировано, например, в статье “Why Should University and Business Cooperate? A Discussion of Advantages and Disadvantages”