МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М. В. Ломоносова

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

Групповое домашнее задание

по Сетевым моделям в экономике,

выполненное на основе набора данных «Political books»

Выполнили студенты ЭФ МГУ

Новиков Антон Денисович, группа, э306

Герцен Роман Андреевич, группа э304

Москва

2025

Для выполнения группового домашнего задания нами был выбрал набор данных «Political books», который первоначально был собран и проанализирован Валдисом Кребсом. Датасет представляет собой сеть совместных покупок книг о политике США на Amazon, которые были опубликованы во время президентских выборов 2004 года. Кребс выбрал политические книги из списка бестселлеров New York Times, нашел эти книги на Amazon и Barnes & Noble, а затем зафиксировал совместные покупки книг. Вершинами являются книги, каждой из них присвоена категориальная характеристика - идеологическая принадлежность. В данном датасете книги классифицированы по трём категориям: либеральные, консервативные и нейтральные. Рёбра представляют собой совместную покупку двух книг одним и тем же покупателем. Этот набор данных использовался в исследовании «M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. (2006)» для проверки работы алгоритма обнаружения сообществ на основе модулярности.

Наш исследовательский вопрос звучит следующим образом: «предсказание идеологической принадлежности книги на основе сетевых метрик». Помимо построения предсказательной модели нашей задачей также является выявить наличие и степень политической поляризации американского общества на основе совместных покупок политических книг.

Набор данных представляет собой ненагруженный, неориентированный, связный граф, имеющий **105** вершин и **441** ребро. Плотность графа составила **0,0404**, что является стандартным значением для сети такого типа. Средняя степень вершины составила **8,4**, т. е. книга в среднем связана совместной покупкой с 8 – 9 другими книгами. Диаметр графа равен **8**, т. е. максимальное «расстояние» между двумя книгами составляет 8 покупок. Средняя длина кратчайшего пути равняется около **2.92,** т. е. любые две книги в среднем отделены друг от друга примерно тремя покупками.

Оптимальным алгоритмом для визуализации данного графа является силовой алгоритм Fruchterman-Reingold. Обратимся к Рисунку 1. Можно заметить довольно четкое разделение множества книг на консервативный и либеральный кластеры, связь между ними слабая. Нейтральные книги расположены на периферии, а также выступают в роли «мостов» между сообществами.

Сравним фактические характеристики графа с модельными. Сравнение нашей сети со случайными графами Эрдеша – Реньи: средняя длина кратчайшего пути значительно ниже типичных значений, дизассортативность по степени значимо меньше нуля, локальная

транзитивность значимо более высокая, умеренная глобальная транзитивность. Сравнение нашей сети с моделью «Малого мира»: средняя длина пути значительно меньше модального значения, дизассортативность по степени значимо меньше нуля, локальная транзитивность соответствует модели «Малого мира», глобальная транзитивность значимо меньше. Таким образом, наш граф не относится к Случайным, но и сильно отличается от модели «Малого мира».

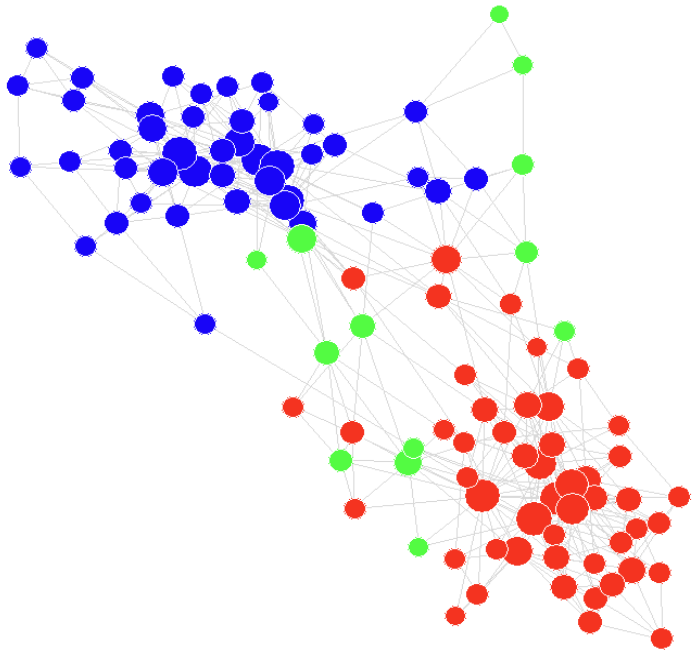


Рисунок 1. Визуализация графа "Political books". Красным цветом обозначены консервативные книги, синим - либеральные, зеленым - нейтральные. Построено авторами на основе набора данных Валдиса Кребса "Political books"

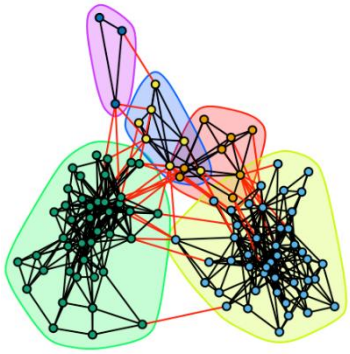
Либеральные книги обладают наибольшей центральностью по степени (среднее значение 8,84 против 8,69 у консерваторов) и собственному значению (среднее значение 0.395 против 0.126 у консерваторов). 7 из 10 самых влиятельных по собственному значению книги – либеральные. Либеральные книги образуют более сплоченную сеть. Однако консервативные книги занимают лидирующие позиции в абсолютном рейтинге центральности по степени («A National Party No More» - 25, «Off with Their Heads» - 25, «Losing Bin Laden» - 25). Наибольшей центральностью по близости и кратчайшему пути обладают книги с нейтральной идеологией: 63,9 в среднем («топ 1» - «Plan of Attack» с центральностью 366,58 – главный «мост» в сети). Таким образом, нейтральные книги играют немалую роль в связывании двух кластеров с противоположными идеологиями и выступают в качестве структурных «мостов» графа.

Было рассчитано 2 вида ассортативности: по степени вершины и по ее классу. Ассортативность по степени вершины составила -0.13. Наблюдается слабая дизассортативность, то есть узлы с большей степенью вершины чуть чаще соединяются с вершинами меньшей степени. Ассортативность по классу составила 0.72, что означает, что пользователи, склонные к определённым взглядам, покупают книги, принадлежащие преимущественно своей идеологической категории.

Для оценки гомофилии сети использовалось два показателя: диадичность и гетерофильность. В нашей сети диадичность оказалась равной 2, что говорит о том, что связей между вершинами одного типа примерно в 2 раза больше, чем это ожидалось случайно. Гетерофильность составила 0.41, что означает, что связей между разными идеологическими группами примерно в 2.5 раза меньше, чем ожидалось бы случайно.

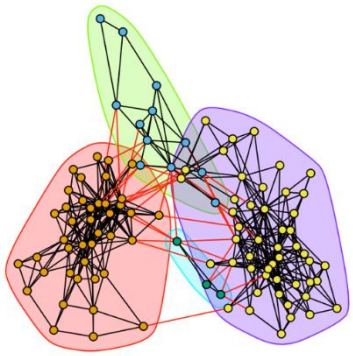
Для разбиения вершин на кластеры было использовано 3 подхода: Edge-betweenness, Fast-greedy и Block modelling.

Edge-betweenness



	conservative	liberal	neutral
1	4	0	4
2	42	0	3
3	1	39	2
4	2	3	2
5	0	1	2

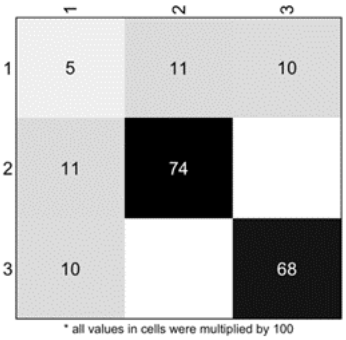
Fast-greedy



	conservative	liberal	neutral
1	1	38	2
2	3	5	4
3	2	0	1
4	43	0	6

Block modelling

Для анализа структуры взаимодействий между типами книг был применён логистический блок-моделинг. На рисунке показана матрица блоков, где строки и столбцы соответствуют трём типам книг. Она показывает, как часто внутри и между найденными блоками возникают связи. Значения в ячейках — вероятности (умноженные на 100) возникновения связей.



Главный результат — структура сети сильно сегментирована: блок 2 и блок 3 (либеральные и консервативные) имеют очень высокую внутригрупповую плотность (74 % и 68 %) связей внутри, а связи между блоками редки (10–11 %).

Построение предсказательной модели

Цель исследования — построить предсказательную модель, которая по структуре сети (совместным покупкам книг) определяет идеологическую принадлежность книги: либеральную, нейтральную или консервативную при малом числе размеченных данных.

Описание статьи (Kipf & Welling, 2016, “Semi-Supervised Classification with Graph Convolutional Networks”): в статье рассматривается схожая задача — классификация вершин в графе (например, документов) при малом числе размеченных данных. Авторы используют графовую структуру и информацию о соседях, чтобы распространять сигналы меток на неразмеченные узлы. Такой подход называется полусупервизируемым обучением (semi-supervised learning): часть данных имеет известные метки, а остальные используются без меток, но с учётом их связей в графе. Методы, применяемые в статье: графовые сверточные сети (Graph Convolutional Networks, GCN), которые обучаются на размеченных вершинах и распространяют признаки через ребра графа. В качестве бейзлайнов там использовались более простые модели, включая Label Propagation.

В качестве предсказательной модели мы использовали Label Propagation. Это простая полусупервизируемая модель, в которой метки известных узлов постепенно распространяются по графу через матрицу смежности. При каждой итерации вероятность принадлежности вершины к каждому классу обновляется на основе меток соседей. Метки размеченных узлов закрепляются и не изменяются. Алгоритм сходится, когда значения перестают меняться. Модель не требует обучения параметров и используется в статье как базовый метод.

Так как графовая структура должна сохраняться, случайное разбиение на независимые множества невозможно — удаление узлов разрушает связи. Поэтому в статье используется частичное сокрытие меток: все узлы остаются в графе, но у части вершин метки скрываются (им присваивается NA). Таким образом, сохраняется полная структура сети, а алгоритм обучается на известных узлах и метках и предсказывает метки скрытых. В нашем исследовании мы скрыли 80% меток ввиду предположения о том, что часто идеология книги неизвестна, хоть и на наших данных размечены все вершины.

План модели:

1. Реализуется функция label_propagation, которая принимает матрицу смежности A и вектор меток (функция была написана нейросетью, так как в новой версии библиотеки, где раньше была эта модель, используется более современная модель, а старой версии мы не нашли).
2. Выполняется разбиение на train и test с сохранением пропорции классов: 20 % меток остаются видимыми, 80 % скрываются.
3. Вся структура графа сохраняется, то есть все ребра остаются.
4. Выполняется финальное обучение на всём train и предсказание для test.

На наших данных модель показала: Macro-F1 = 0.749

Метрика качества Macro-F1: в многоклассовом случае F1-мера вычисляется для каждого класса отдельно, а затем усредняется с одинаковым весом классов (в отличие от взвешенного усреднения, где доля каждого класса пропорциональна его размеру): Macro-F1 показывает среднюю сбалансированную точность по всем классам, не давая преимуществ более частым категориям