Защита домашнего задания

по сетевым моделям в экономике,
выполненное на основе набора данных

«Political books»

Новиков Антон, э306

Герцен Роман, э304

Экономический факультет МГУ

Датасет

Набор данных **«Political books»** 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.

Вершинами являются книги, каждой из них присвоена категориальная характеристика - идеологическая принадлежность (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.

Исследовательский вопрос:

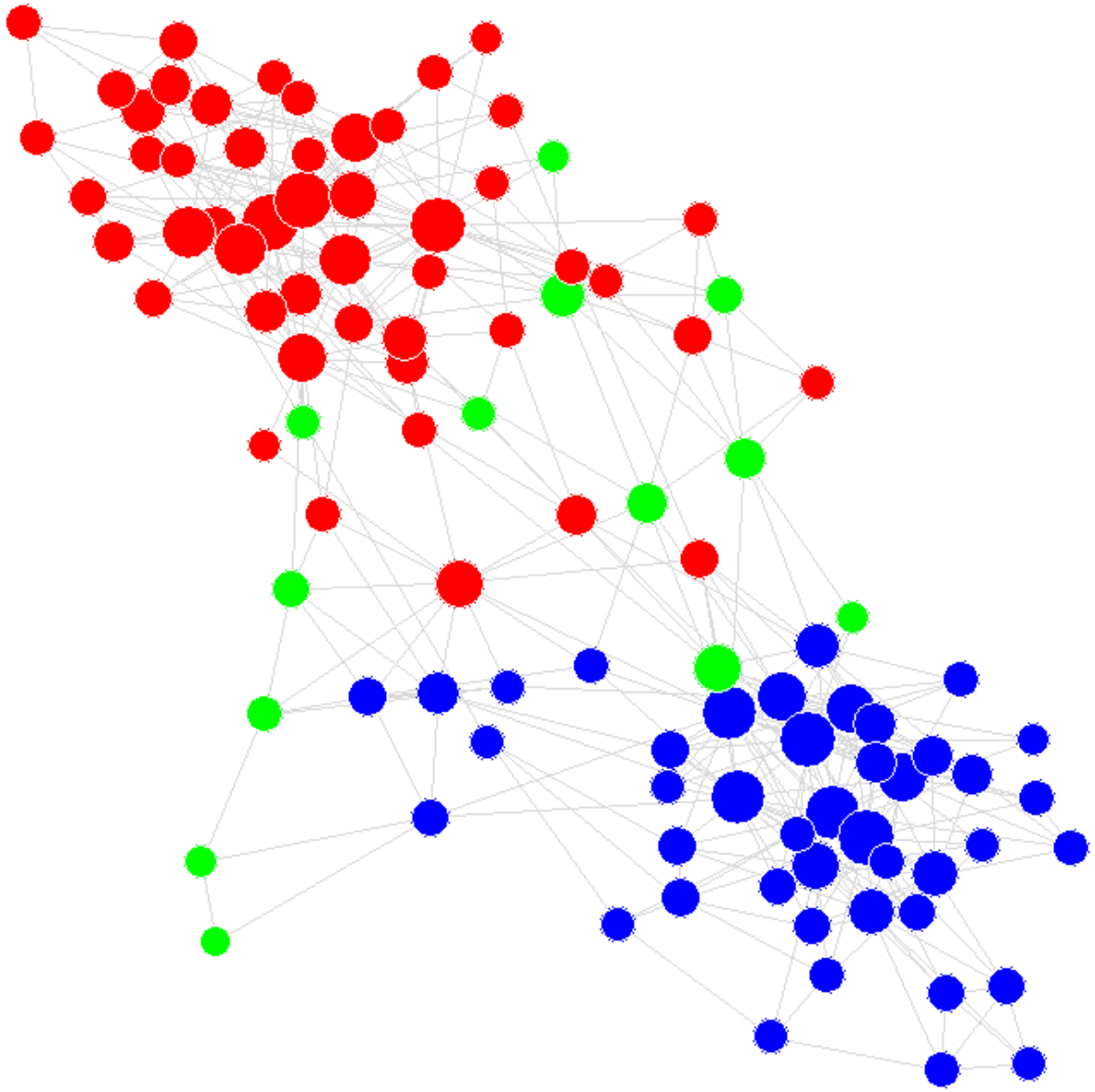
Предсказание идеологической принадлежности книги на основе сетевых метрик

Помимо построения предсказательной модели нашей задачей также является выявить наличие и степень политической поляризации американского общества на основе совместных покупок политических книг.

Базовые характеристики графа

- Ненагруженный, неориентированный, связный
- 105 узлов, 441 ребро
- Плотность: 0,0404
- Средняя степень вершины: 8,4
- Диаметр: 8
- Средняя длина кратчайшего пути: 2,92

Визуализация графа (алгоритм Fruchterman-Reingold)



Легенда

- Красные – консервативные книги
- Синие – либеральные книги
- Зеленые – нейтральные книги

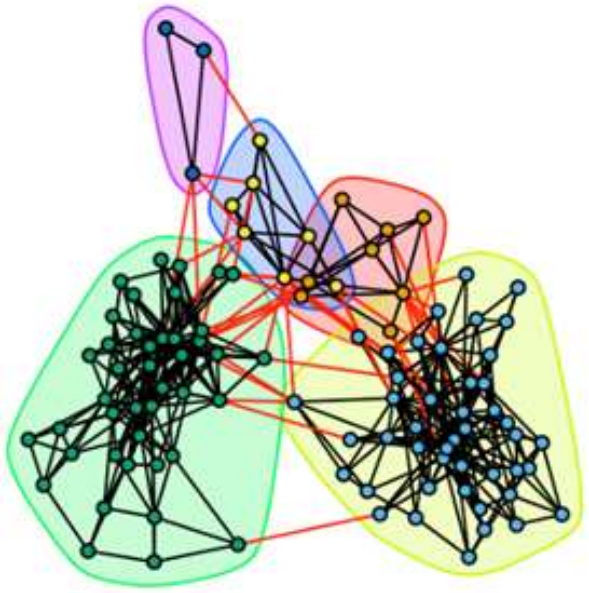
Расчет центральностей

- **Либеральные книги:** наибольшей средняя центральность по степени (среднее значение 8,84 против 8,69 у консерваторов) и собственному значению (среднее значение 0.395 против 0.126 у консерваторов). Образуют более сплоченную сеть.
- **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).
- Книги с **нейтральной идеологией** обладают наибольшей центральностью по близости и кратчайшему пути: 63,9 в среднем («Plan of Attack» с центральностью 366,58 – главный «мост» в сети). Играют важную роль в связывании двух кластеров с противоположными идеологиями и выступают в качестве структурных «мостов» графа.

Предпочтительное присоединение

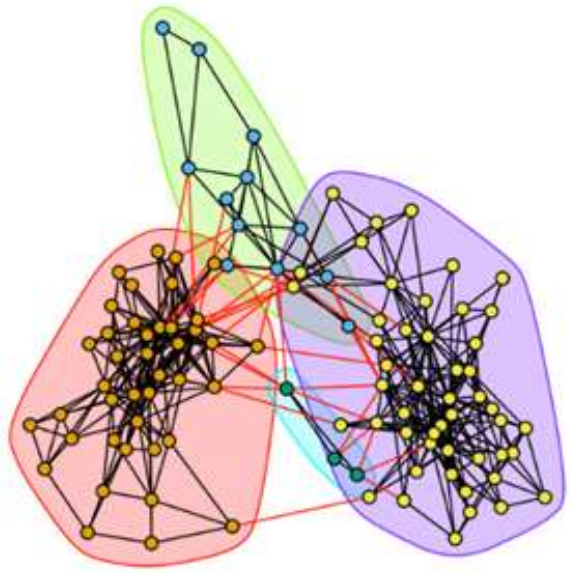
- *Ассортативность по степени вершины: -0.13.* Наблюдается слабая дизассортативность, то есть узлы с большей степенью вершины чуть чаще соединяются с вершинами меньшей степени.
- *Ассортативность по классу: 0.72.* Люди, склонные к определённым политическим взглядам, покупают книги, преимущественно относящиеся к их идеологической категории.
- *Dyadicity = 2.* Связей между книга одной идеологии примерно в 2 раза больше, чем это ожидалось случайно.
- *Heterophilicity = 0.41. Связей* между разными идеологическими группами примерно в 2.5 раза меньше, чем это ожидалось случайно.

Кластеризация



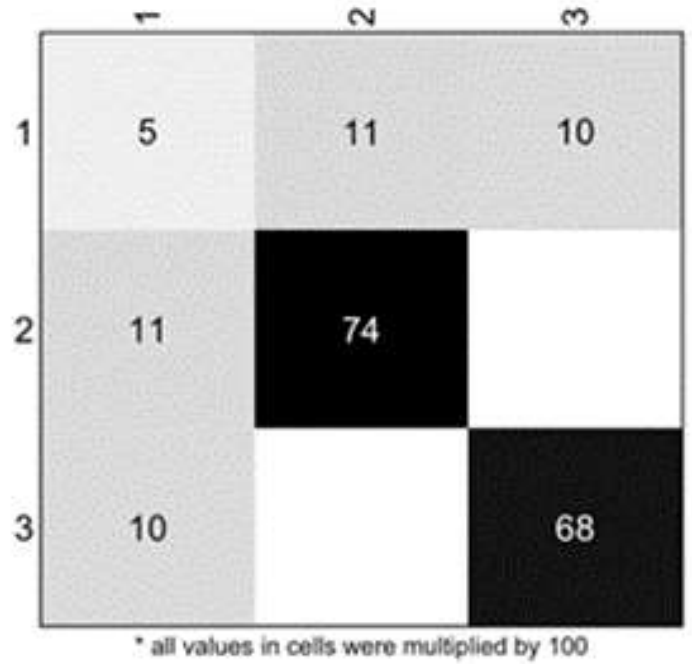
	conservative	liberal	neutral
1	4	0	4
2	42	0	3
3	1	39	2
4	2	3	2
5	0	1	2

Edge-betweenness



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

Fast-greedy



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%).

Block modelling

Предсказательная модель

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.

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.

Предсказательная модель. План

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

Спасибо за внимание!