# Сетевые сообщества

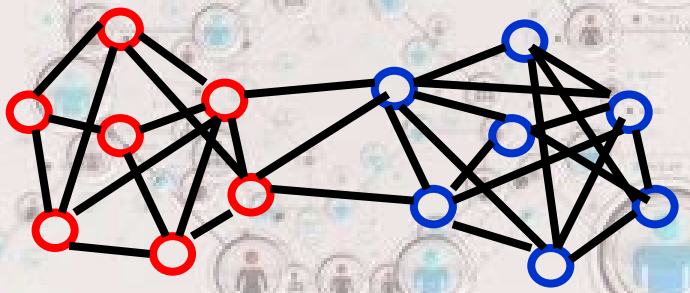
Николай Скворцов
nskv@mail.ru

# План лекции

- Определение сообщества
- Клики
- Иерархические алгоритмы
- Спектральная кластеризация

# Сетевые сообщества

- Сообщества (сплочённые подгруппы) – это группы вершин социальной сети, плотность связей в которых существенно выше, чем плотность связей между группами
  - Неформальное определение



- Представление
  - Сеть G: n вершин, m рёбер
  - Сообщество C: $n_C$ вершин, $m_C$ рёбер
  - $m_{ext}$ внешних рёбер

# Зачем искать сообщества?

- Параметры и поведение сети могут сильно отличаться от сообщества к сообществу
- Исследование влиятельных участников подгруппы
- Исследование общих характеристик агентов в подгруппе
- Рекомендация недостающих связей внутри сообщества
- Типизация вершин
- Вложенные сообщества
- Перекрывающиеся сообщества (мосты, перетекание информации)
- Исследование изоляции подгрупп (слабое распространение информации)
- Локализация хранения и обработки информации о взаимодействии участников

# Базовый подход к выявлению сообществ

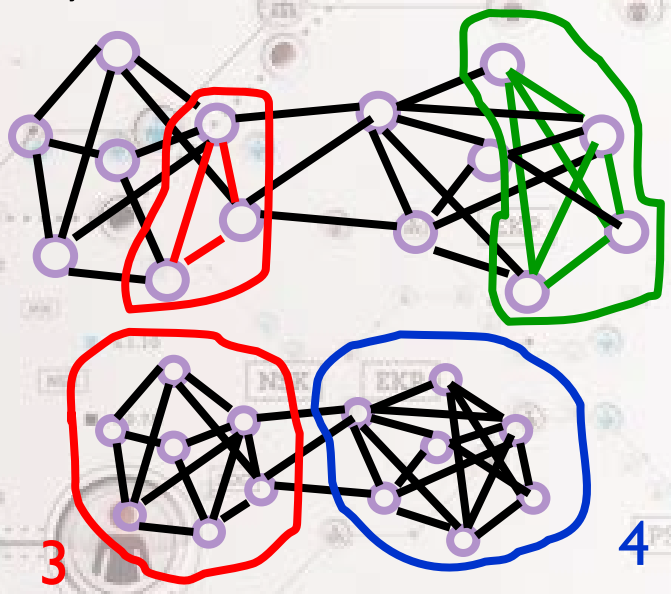
- Плотность графа $\rho = 2m / (n (n - 1))$
- Плотность внутренних связей сообщества
  $\delta_{int}(C) = 2m_C / (n_C (n_C - 1))$
  $\delta_{int}(C) > \rho$
- Плотность внешних связей
  $\delta_{ext}(C) = m_{ext} / (n_C (n - n_C))$
  $\delta_{ext}(C) < \rho$
- Задача обнаружения сообществ состоит в максимизации суммы разностей плотности ($\delta_{int} - \delta_{ext}$) по всем сообществам сети

# Возможные критерии определения сообществ

- Взаимность связей множества агентов
- Частота связей
- Близость и достижимость вершин
- Относительная частота связей между членами и нечленами сообщества
- Другие критерии

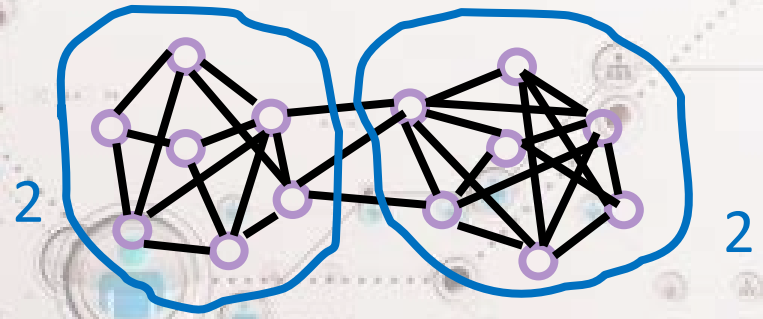
# Локальные подходы к обнаружению сообществ

- Локальные подходы принимают во внимание структуры, обнаруживаемые в графах
- Клика
  - подграф, в котором каждая вершина связана со всеми остальными в подграфе (связь каждый с каждым)
  - Неприменимо центральное расположение вершин
- k-ядро
  - подграф, в котором каждая вершина связана минимум с k другими вершинами того же подграфа

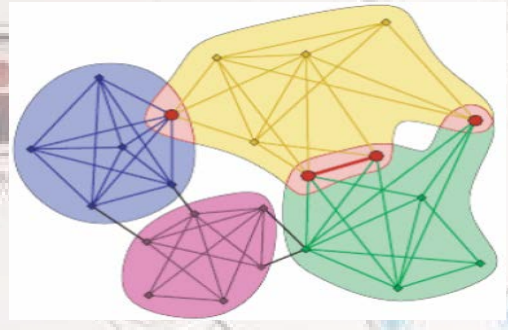


# Локальные подходы к обнаружению сообществ (2)

- n-клика
  - подграф, в котором максимальное расстояние между любыми двумя вершинами не больше n




- p-клика
  - подграф, в котором вершины имеют не менее p-той части соседних вершин в том же подграфе
- Перколяция клик
  - максимальное объединение k-клик, любые две из которых соединены набором смежных k-клик



- Все перечисленные подходы – NP-полные задачи

# Глобальные подходы к обнаружению сообществ

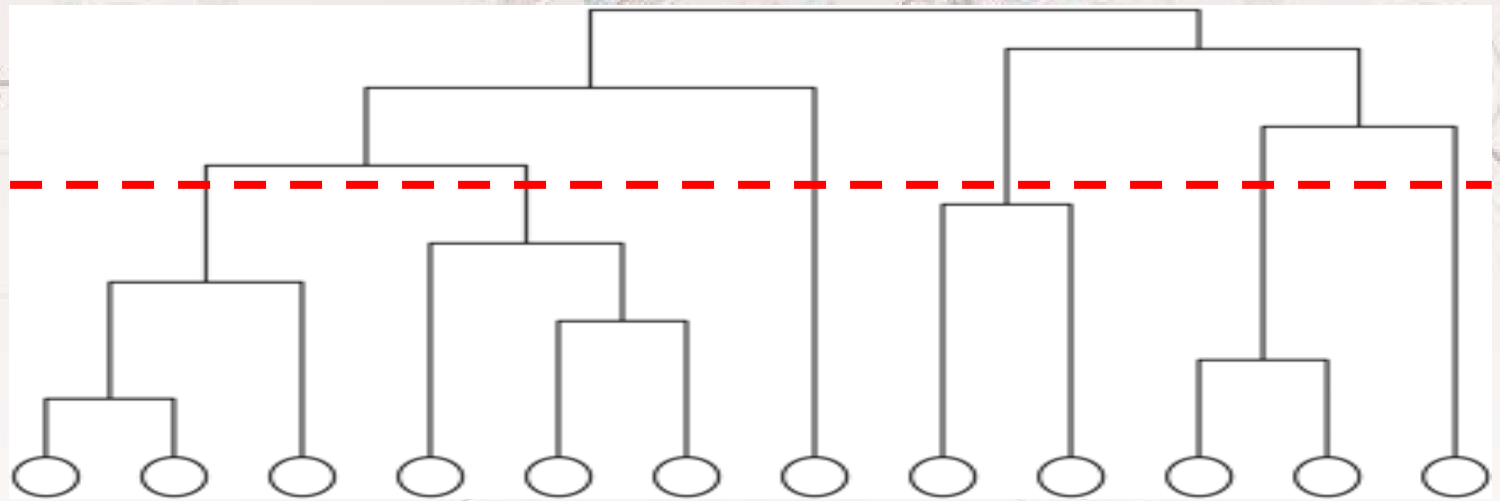
- Применение глобальных свойств сетей в качестве критериев сообществ
- Используются различные критерии сходства вершин
  - Случайное блуждание – количество посещений вершины при случайных переходах по графу
  - Центральность по посредничеству – количество кратчайших путей, проходящих через вершину
  - Модулярность вершин – характеристика, основанная на разности между фактическим количеством рёбер подграфа и ожидаемым количеством рёбер, если бы это была случайная сеть

# Метод распространения ярлыков

- первоначально каждая вершина сети имеет собственный ярлык
- на каждой итерации часть вершин принимает ярлыки своих «соседей»
- ярлык, который более представлен среди соседей данной вершины
- Критерий остановки: каждая вершина имеет по меньшей мере столько же соседей внутри своего сообщества, сколько она имеет соседей из других сообществ

# Иерархические алгоритмы

- Расчёт веса для всех пар вершин
  - По различным критериям (например, улавливание сообществом случайного блуждания)
  - случайные блуждания оказываются «пойманы» в той части графа, внутри которого связи плотнее, чем окружающие
  - Вычислительная сложность $O(n^2 \log n)$ для редких сетей
- Инициация: n несвязанных вершин
- Добавление рёбер между парами в порядке убывания веса
- Получаем дерево вложенных компонентов
- Уровень среза определяет детализацию сообществ

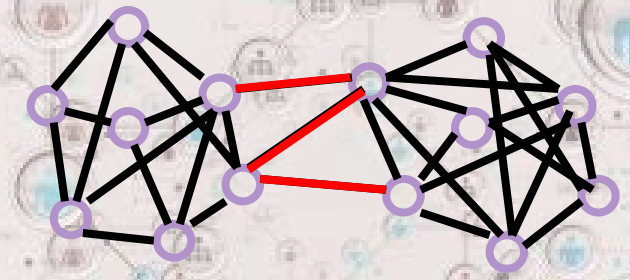


# Иерархические алгоритмы (2)

- Направление алгоритма
  - Агломеративные подходы – итеративное слияние подгрупп вершин, сходных по разным критериям
  - Разделяющие подходы – итеративное удаление рёбер между слабо связанными вершинами

# По посредничеству рёбер

- Посредничество
  - количество кратчайших путей графа, проходящих через ребро
- Иерархический алгоритм по посредничеству рёбер удаляет рёбра с высоким показателем посредничества (мосты), разбивая сеть на компоненты



- Эти компоненты являются сообществами полной сети
- Пока посредничество любого ребра больше порогового значения
  - Удалить ребро с наибольшим посредничеством
  - Пересчитать посредничество
- Плохо масштабируется
  - Удаление дуги влияет на посредничество других рёбер (расчёт кратчайших путей всех пар за один проход – $O(n^3)$)

# Модулярность подгруппы

- Рассмотрим рёбра в пределах сообщества или между сообществом и остальной частью сети
- Функция модулярности (М. Ньюман, М. Гирван)

$$Q = \frac{1}{2m} \sum_{ij} \left[ A_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j)$$

  - $A$ – матрица смежности
  - $d_i d_j / 2m$ – вероятность ребра между двумя вершинами, пропорциональная их степеням
  - $\delta$ – находятся ли обе вершины в одном сообществе
- Подграф представляет собой сообщество, если число ребер внутри подграфа превышает ожидаемое число внутренних ребер, которое этот подграф имел бы в нулевой модели (случайной сети)
- Модулярность является «одновременно глобальным критерием определения сообщества и функцией качества

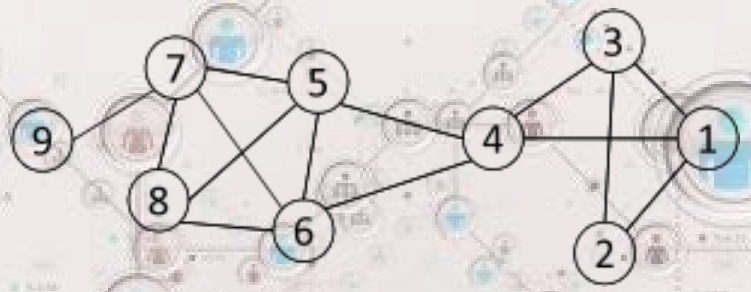
# Кластеризация по модулярности подгруппы

- Агломеративный подход
  - Инициализация: каждая вершина является сообществом
  - Жадная стратегия: последовательно соединяются сообщества с наибольшим увеличением $\Delta Q$
  - До тех пор, пока $\Delta Q <= 0$ от при соединении любых сообществ
- Сложность для редких сетей $O(n \log^2 n)$

# Спектральная кластеризация

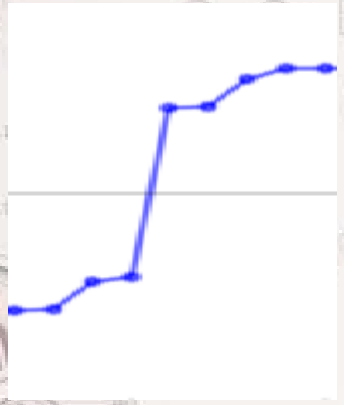
- Каждую вершину графа описываем точкой в многомерном вещественном пространстве
- Строится лапласиан $\tilde{L} = D - A$
  - D - диагональная матрица степеней $D = diag(d_1, d_2, \cdots, d_n)$
  - A – матрица смежности графа
- Находятся собственные числа и векторы матрицы
- Оптимальное решение – минимизация собственных чисел
- Первое собственное число пропускается, так как выделяет всю сеть как одно сообщество

# Спектральная кластеризация (2)




$$\widetilde{L} = D - A = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 4 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

$$S = \begin{bmatrix} 0.33 & -0.38 \\ 0.33 & -0.48 \\ 0.33 & -0.38 \\ 0.33 & -0.12 \\ 0.33 & 0.16 \\ 0.33 & 0.16 \\ 0.33 & 0.30 \\ 0.33 & 0.24 \\ 0.33 & 0.51 \end{bmatrix}$$



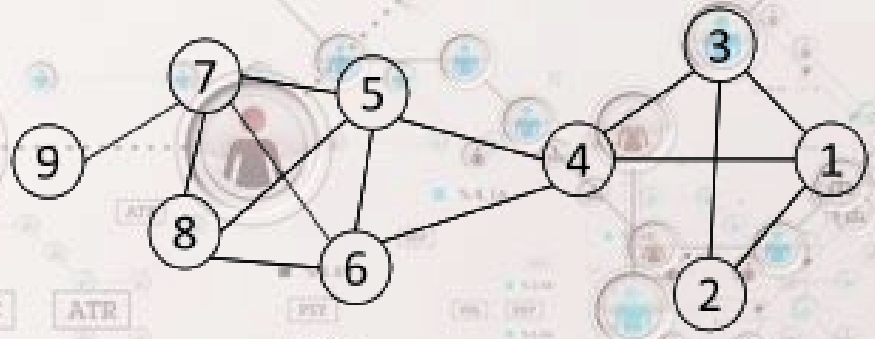
Сообщества (k-means):

{1, 2, 3, 4}

{5, 6, 7, 8, 9}

# Спектральная кластеризация по модулярности

- Матрица модулярности $B = A - \mathbf{d}\mathbf{d}^T/2m$




$$B = \begin{bmatrix} -0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\ 0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\ -0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\ -0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04 \end{bmatrix}$$

$$S = \begin{bmatrix} 0.44 & -0.00 \\ 0.38 & 0.23 \\ 0.44 & -0.00 \\ 0.17 & -0.48 \\ -0.29 & -0.32 \\ -0.29 & -0.32 \\ -0.38 & 0.34 \\ -0.34 & -0.08 \\ -0.14 & 0.63 \end{bmatrix}$$

Сообщества
{1, 2, 3, 4}
{5, 6, 7, 8, 9}

# Итоги

- Определение сообщества вершины
  - *Клика, k-клика*
- *Определение сообщества группы*
  - Клика, перколяция клик
- Определение сообществ сети в целом
  - *Агломеративные и разделяющие иерархические подходы*
  - Спектральная кластеризация
- Методы на основе сходства вершин