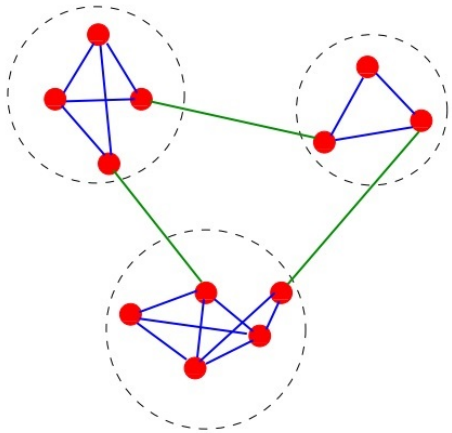
Занятие 3: Сообщества в социальных сетях

НИУ ВШЭ

Москва, 2017

Сообщества в социальных сетях

Распределение узлов в социальных сетях неравномерно. Между некоторыми вершинами оказывается значительно больше связей, между некоторыми - значительно меньше.



Подобная картина характерна для подавляющего большинства социальных сетей. В таком случае мы говорим о наличии *сообществ* (или *кластеров*) внутри социальных сетей. При этом четкого математического определения сообщества в сети пока не предложено.

Разделение графа и выделение сообществ в сети

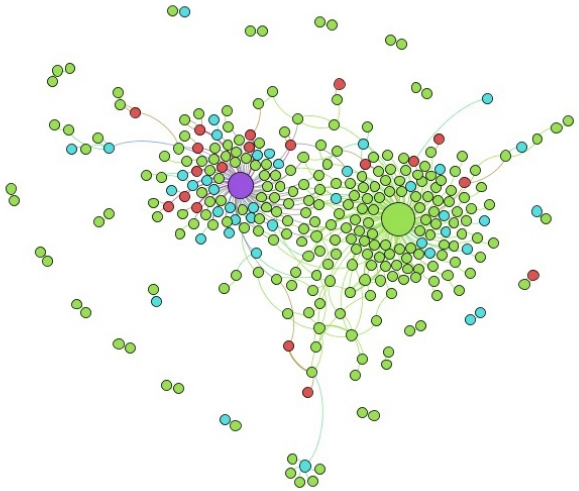
- Разбиение графа (graph partition) и выявление сообществ в социальных сетях (community detection in social networks) принципиально разные содержательные задачи.
- Разбиение графа обычно используется при решении технических задач (распределенные вычисления).
- Выявление сообществ в социальной сети призвано выявить естественным образом сформированные группы людей/организаций, которые связаны друг с другом более тесно, чем с другими акторами социальной сети. Сообщества могут быть разного размера.

Выявление сообществ в сетях

Интуитивно мы можем говорить о том, что число связей внутри сообщества должно быть значительно больше, чем число связей, выходящих вовне.

В каких социальных сетях могут быть выделены сообщества?

- Сеть должна быть разреженной (иметь низкую плотность). Иначе будет сложно выявить различия в плотностях между связными комп
- Сеть должна быть связной. Нет смысла выявлять сообщества, если между компоненты не соединены друг с другом.



Методы разделения сети на сообщества

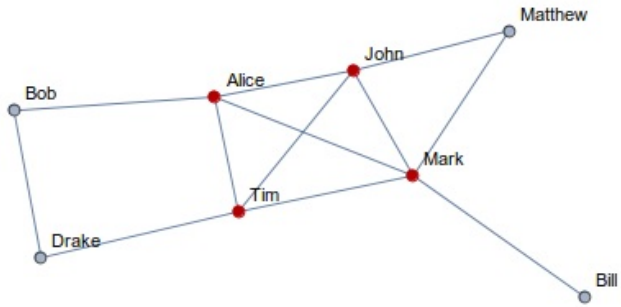
Выделения в социальных сетях сообществ может происходить на различных основаниях

- На основании числа связей вершин (k-core, k-plex)
- На основании достижимости вершин (n-cliques)
- На основании структурной схожести (structural equivalence)
- Label Propagation
- На основании степени посредничества ребер
- Максимизация модулярности - индекса, характеризующего *'качество разбиения на сообщества'* (метод Левена).
- Спектральные методы

В каждом из этих случаев операционализация понятия *сообщество* производится индивидуально. Таким образом, при выявлении сообществ в социальных сетях необходимо понимать, какая задача (оптмизация каких показателей) должна производиться.

Клики

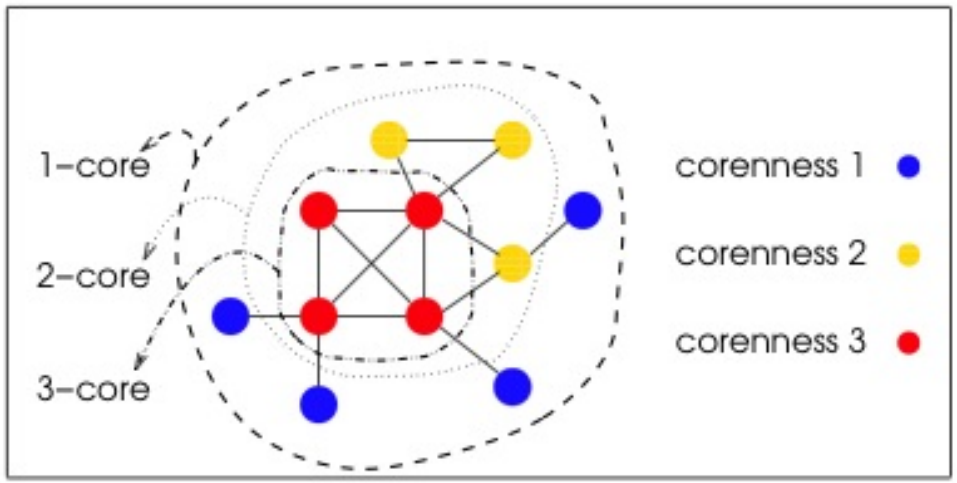
Клика - это полный субграф из трех или более вершин. Клика состоит из вершин, которые напрямую соединены со всеми остальными вершинами клики.
Клика - это **очень строгое** определение сообщества. В случае, если число исходящих связей от акторов ограничено k, то размеры клики не могут превышать k вершин.



Перечислите клики в данном графе. Какие размеры у клик в графе?
Использование клик целесообразно при анализе социальных сетей небольшого размера и/или для выявления очень тесно связанных групп. В большинстве случаев клик слишком мало и они слишком малы.

K-core

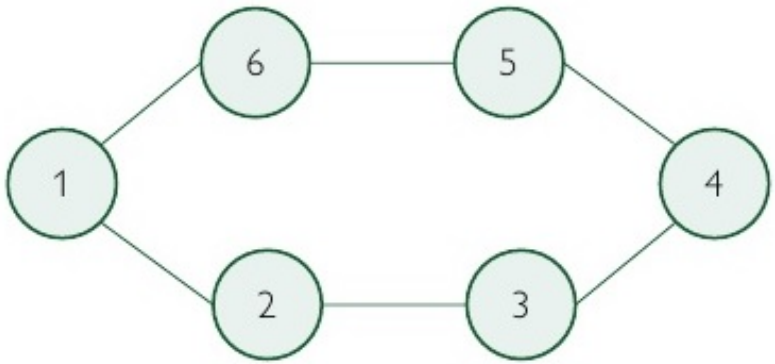
K-core - это субграф, в котором каждая из вершин имеет связи как минимум с k вершинами субграфа (Seidman, 1983).
$d_s(i) \geq k$ для всех $n_i \ni N_s$



В случае k-core речь идет о *минимальной степени центральности в субграфе*.

K-plex

K-plex - это субграф из g_s вершин, в котором каждая из вершин соединена не менее чем с $g_s - k$ вершинами субграфа.
$d_s(i) \geq (g_s - k)$



Диаметр K-plex меньше или равен 2 при условии $k < (g_s + 2)/2$

N-cliques

k-core и k-plex можно назвать показателями, базирующимися на степени центральности. В противовес им n-cliques основаны на дистанции между вершинами.
Предполагается, что дистанция между вершинами, принадлежащими одному сообществу, мала. Таким образом, вершины могут обмениваться друг с другом информацией/ресурсами с высокой скоростью.
n-clique - это субграф, в котором максимальная дистанция между любыми двумя вершинами не превышает значения n.
$d(i,j) \leq n$ для всех $n_i, n_j \ni N_s$

- В каком случае n-clique является кликой?
- Какой порог n целесообразен?

Структурная схожесть

Вершина, занимающие схожее структурное положение с высокой долей вероятности будут являться частями одного структурного сообщества в сети.

Для операционализации понятия **структурная схожесть (structural equivalence)** мы рассчитываем дистанцию между двумя вершинами:

$d_{AB}^{E} = \sum_{k=1}^{n} \sqrt{(a_k - b_k)^2}$

где $A = (a_1, a_2, ..., a_n)$, $B = (b_1, b_2, ..., b_n)$ - показатели, описывающие вершины A и B.

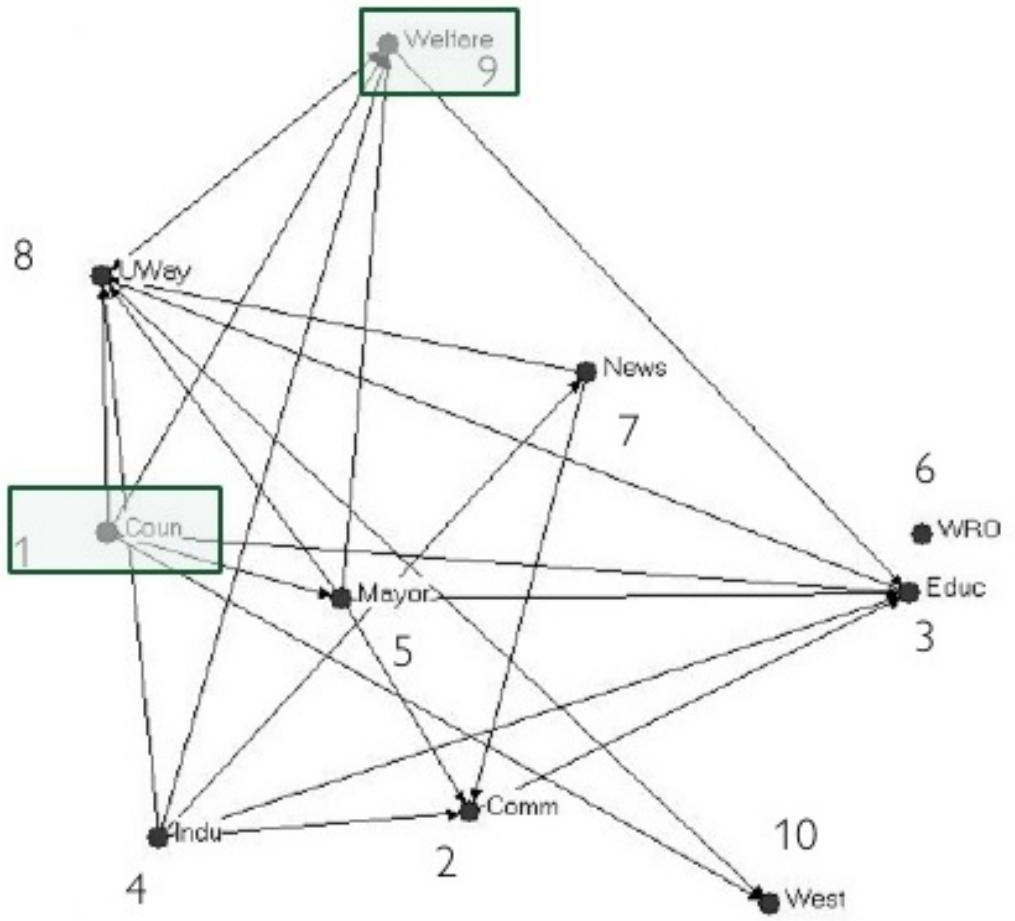
В случае, если мы рассматриваем только матрицу смежностей, то имеет смысл рассчитывать структурную схожесть следующим образом:

$d_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2}$

где A - матрица смежностей.

При каких d вершины имеют одинаковых соседей?

Структурная схожесть: Пример



Меры структурной схожести

Коэффициент Жаккара
$Score(x, y) = \frac{|N(x) \bigcap N(y)|}{|N(x) \bigcup N(y)|}$ (отношение общего числа друзей к суммарному числу друзей)
Адамик-Адар
$Score(x, y) = \sum_{z \in N(x) \bigcap N(y)} \frac{1}{log(|N(z)|)}$
Предпочтительное присоединение
$Score(x, y) = |N(x)| * |N(y)|$

Критерии выявления сообществ

Многие алгоритмы кластеризации сетей основаны на максимизации разницы между **числом связей внутри кластера** и **числом связей вершин вне кластера**.

$\delta_{int}(C) = \frac{internal}{n_c(n_c-1)/2}$

$\delta_{ext}(C) = \frac{inter-cluster}{n_c(n-n_c)}$

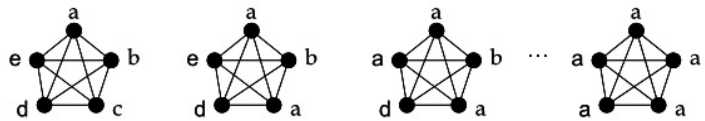
где n_c - число вершин кластера, а n - общее число вершин.

Максимизация $\delta_{int}(C) - \delta_{ext}(C)$ для всех кластеров социальной сети.

Label Propagation

Метод Label Propagation (Raghavan et al., 2007) также один из часто используемых и быстрых методов для выявления сообществ.

- На первом этапе каждой из вершин приписывается определенное сообщество (label). (Каждая из вершин при этом пронумерована).
- Каждая вершина меняет свое сообщество (label) на такое сообщество, к которому принадлежит большинство ее соседей.
- Рассчитываем, действительно ли у каждой из вершин сети такое же сообщество, как и у большинства ее соседей. Если да, то процесс разбиения на сообщества завершается. Если нет, то возвращаемся к предыдущему шагу.

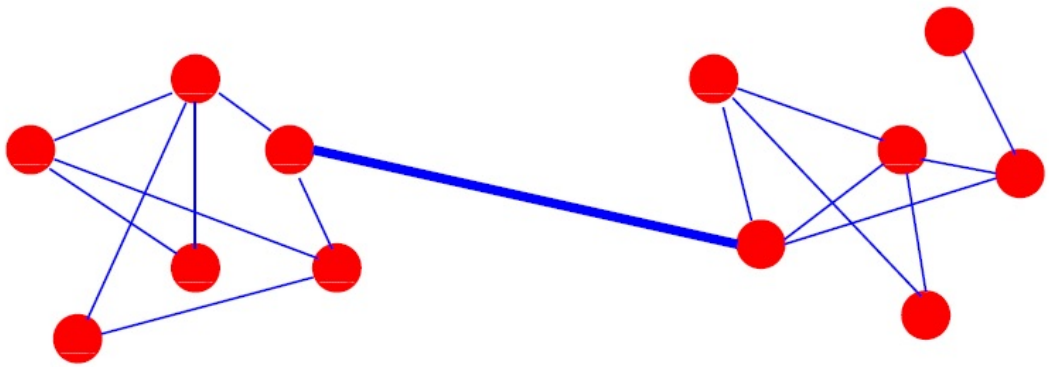


Отличительная особенность - при разделении с label propagation могут быть разные результаты разделения на сообщества после нескольких итераций.

Метод разделения на сообщества Ньюмана-Гирван

Одним из наиболее востребованных и популярных методов разбиения на сообщества является метод Ньюмана-Гирван (Newman and Girvan, 2004).
Алгоритм метода:

- Расчет степени посредничества (betweenness) для всех ребер
- Удаление ребра с наибольшей степенью посредничества
- Пересчет степени посредничества для всех узлов



Модулярность

Модулярность - это показатель, который характеризует *качество* разделения на сообщества (Newman, 2004).
Идея модулярности заключается в том, что в реальной социальной сети, в отличие от случайной (например сети Эрдоша-Реньи) выделяются четкие кластеры. Таким образом, необходимо сравнить реальную сеть со случайной.
$Q = \frac{1}{2m} \sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j)$
Или
$Q = \frac{1}{2m} \sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j)$
где Q - модулярность, m - число ребер в графе, A - матрица смежностей, P - матрица смежностей для случайной сети (сеть Эрдоша-Реньи), k_i - степень вершины i, $\delta(C_i, C_j)$ - дельта Кронекера.

Модулярность

Модулярность варьируется от -1 до 1.

$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j)$

В случае идеального разделения на сообщества модулярность 1.

Обычно модулярность варьируется в интервале от 0.3 до 0.7.

Ассортативность и гомофилия

Схожесть между вершинами может быть рассчитана и по аттрибутам вершин.
В анализе социальных сетей часто используется концепт **гомофилии**, по которому схожие вершины склонны формировать связи между собой (McPherson et al., 2001). Для оценки того, насколько схожие по какому-то параметру вершины склонны формировать друг с другом связи, используется индекс *ассортативности*.

$$r = \frac{\sum_{ij}((A_{ij} - d(i)d(j)/2m)x_i x_j}{\sum_{ij}((d(i)\delta_{ij} - d(i)d(j))/2m)x_i x_j}$$

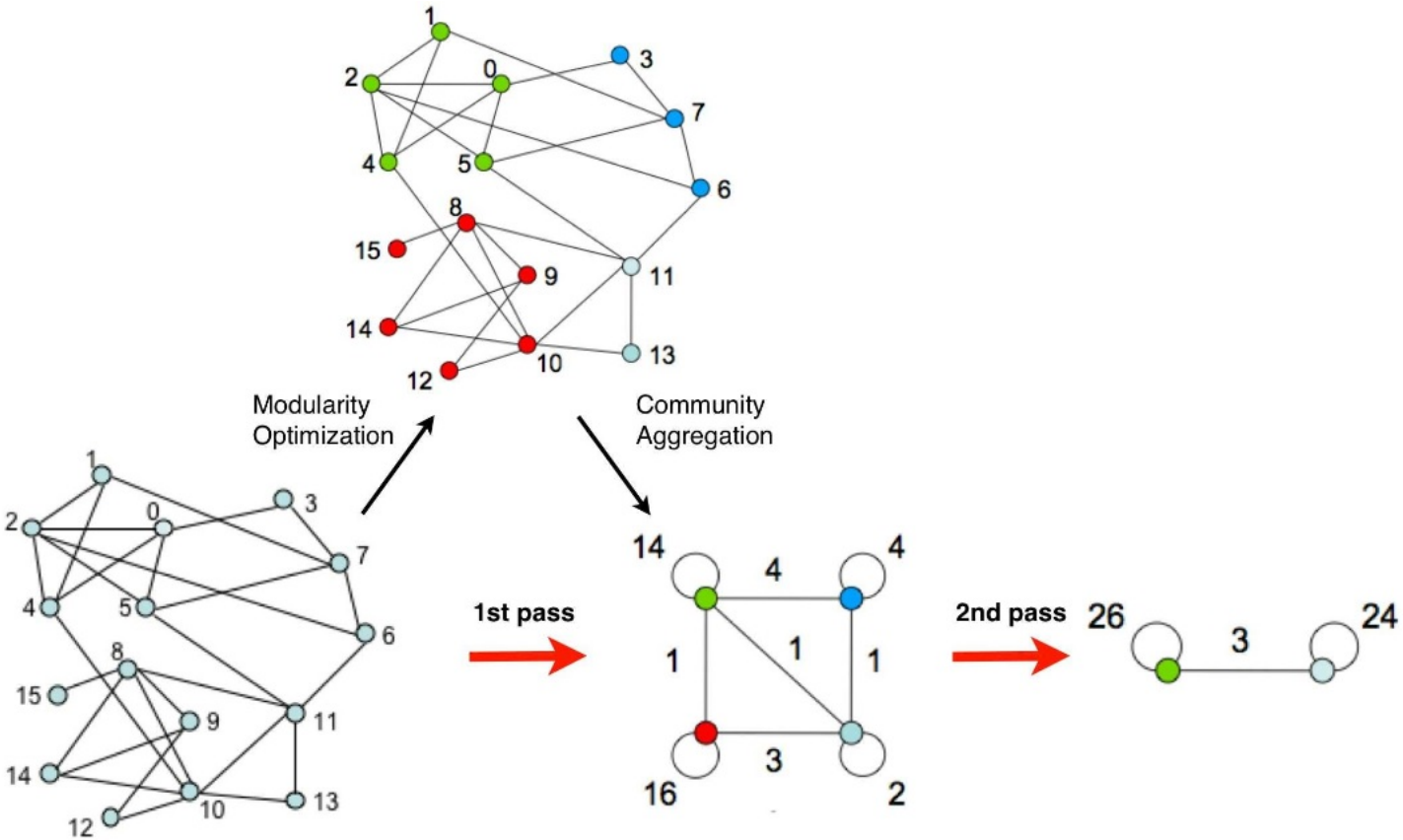
- Как варьируется ассортативность? При каких показателях ассортативности мы можем говорить о гомофилии?
- Какие показатели ассортативности по степени центральности характерны для **модели предпочтительного присоединения** (Барабаши-Альберт)?

Метод Левена

Метод Левена основан на оптимизации модулярности.

- На первом этапе каждая из вершин социальной сети приписывается к отдельному сообществу. Таким образом, изначально число сообществ оказывается равным числу вершин социальной сети.
- Затем мы оцениваем, насколько будет выгодно с точки зрения модулярности, если вершина i и вершина j будут приписаны к одному сообществу. Если модулярность растет, то мы оставляем i и j в одном сообществе. Если выигрыша с точки зрения модулярности нет, то каждая i остается в своем сообществе.
- Первый этап завершен в тот момент, когда достигнут локальный максимум модулярности.
- На втором этапе все вершины, принадлежащие одному сообществу, 'схлопываются' в одну вершину. Число связей внутри сообщества становится весом петли (self-loop), связи с вершинами из других сообществ сохраняются.
- Возвращение к шагу 1.

Метод Левена: Схема



Спектральные методы разделения на сообщества

Спектральные методы кластеризации основаны на использовании собственной матрицы (eigenmatrix) и эйгенвекторов (eigenvectors) или их производных.

Алгоритм:

- Вершины графа переводятся в точки с координатами-элементами эйгенвекторов
- Вершины графа могут также переводиться в Лапласиан - матрицу L=D-A, где L - лапласиан, D - это матрица степени центральностей, в которой по диагонали центральности вершин, а A - матрица смежностей.
- Затем уже точки с заданными координатами кластеризуются стандартными алгоритмыми кластеризации (например, методом k-средних).