Выявление неявных сообществ в социальных сетях

Максим Гончаров
maxgon@microsoft.com

В статье рассмотрена задача выявления сообществ в социальных сетях, т.е. групп узлов тесно связанных между собой и слабо связанных с остальными узлами. Выведена мера качества разделения сети на сообщества – еще один вариант модулярности. Приведены два алгоритма отыскания тесно связанных сообществ.

Введение

Большое количество интересных для исследования социальных систем можно представить в виде графа. Например:

- Сеть сотрудников компании, где узлами служат люди, а дугами соединены те из них, которые обмениваются электронными сообщениями с определенной интенсивностью в течение какого-то отрезка времени.
- Сеть страниц в интернете, где дугами служат ссылки.
- Блоги Livejournal и др., в которых узлами служат пользователи, а дугами соединены взаимные «друзья» или пользователи, обменивающиеся комментариями с определенной частотой.
- Социальные сети Facebook, «ВКонтакте» и проч.

Несмотря на то, что попытки исследования социальных сетей с использованием математических методов предпринимаются давно, в последние два десятилетия интерес к этой теме значительно возрос. Это вызвано с одной стороны доступностью большого объёма точных данных из реальных социальных систем (Livejournal, Facebook, Twitter, Flickr и проч.), а с другой – доступностью и дешевизной мощных вычислительных ресурсов. Статистический анализ социальных сетей выявил ряд интересных и иногда неочевидных свойств:

1. «Маленький диаметр» или «феномен шести рукопожатий». Если мы определим диаметр социальной сети как среднюю длину пути между двумя ее членами (узлами) или как 90%-квантиль расстояния между ними, то в большинстве социальных сетей в интернете эта величина

равна приблизительно 6-7, даже при миллионах пользователей и миллиардах связей. Это связано прежде всего с тем, что большинство пользователей не желает быть на периферии сообщества, а стремится приблизиться как можно ближе к «авторитетам» и «хабам».

2. «Сужающийся диаметр». В момент образования социальная сеть представляет собой большой набор мало связанных сообществ. В процессе развития эти сообщества соединяются друг с другом, а затем образуется гигантский связанный кластер с малым диаметром, а также небольшое количество «маргинальных» сообществ. В дальнейшем диаметр сети продолжает уменьшаться, так как новые члены присоединяются к основному сообществу. Неочевидным является тот факт, что величина второго и третьего по величине сообщества не растет даже линейно по мере роста общего числа членов, присоединяющихся к гигантскому кластеру.
3. «Транзитивность сетей»: если два пользователя имеют общего «друга», то с большей, чем ожидаемо вероятностью они также являются (или станут) взаимными друзьями. Т.е. в социальных сетях возникает большое количество «треугольников».
4. «Гравитация»: существует корреляция между весом соседних узлов: «авторитетные» члены сообщества имеют гораздо больший вес связи друг с другом, чем ожидаемый. Веса связей определяются числом взаимных комментариев и прочими характеристиками, описывающими интенсивность взаимодействия.
5. «Heavy tails». Распределение связей внутри социальной сети имеет «тяжелые хвосты»: существует немного «авторитетов» с огромным количеством связей и большое количество «обычных» пользователей с почти нормальным распределением связей между собой.

Большое количество исследований в последнее время посвящено теме выявления неявных сообществ в социальных сетях. Неявных, то есть не обозначенных общими тегами и не заявленных членством в тематических группах. Под сообществом мы интуитивно понимаем группу узлов, тесно связанных друг с другом и слабо связанных с узлами вне этого сообщества. Т.е. задача обнаружения неявных сообществ заключается в разбиении графа социальной сети на непересекающиеся группы узлов, внутри каждой из которых число связей между узлами значительно больше, чем число связей, соединяющих узлы из разных групп. Таким образом, с математической точки зрения речь идет о решении задачи кластеризации графа, где каждый кластер характеризуется повышенной плотностью находящихся внутри него дуг. Способность выделять такие плотно связанные группы может иметь значительный практический интерес. Например, выделение связанных групп веб-страниц помогает определить тематически однородные источники информации, выявление групп в социальных сетях может помочь обнаружить недекларируемые сообщества, связанные общими интересами и мнениями. Интересно также сопоставить таким образом найденные сообщества с сообществами, членами которых пользователи объявляют себя сами.

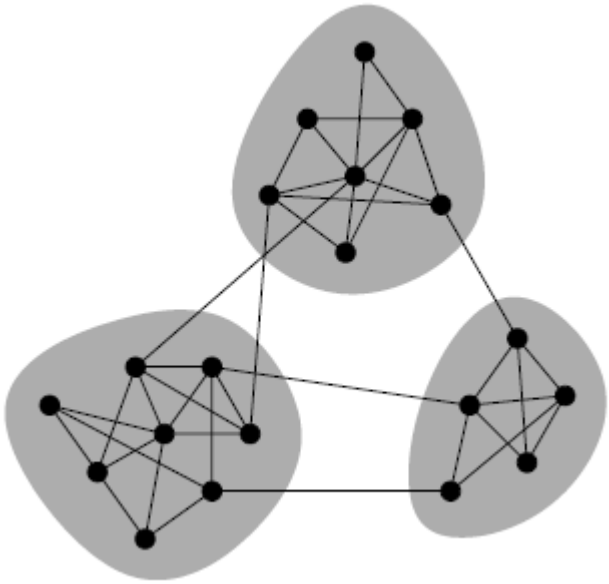


Рис 1. *Разбиение сети на группы с б**о**льшей плотностью связей внутри, чем между собой.*

Обозначения

В качестве социальной сети мы будем рассматривать ненаправленный граф G=(N, E), где N – множество узлов с числом элементов n, а E – множество (ненаправленных) дуг, с числом m. Каждая дуга e из E характеризуется парой инцидентных ей узлов из N. Чтобы подчеркнуть ненаправленность дуги мы будем обозначать ее как *множество* из пары вершин $e = \{i, j\} \subset N$, подчеркивая таким образом, что дуги $\{i, j\}$ и $\{j, i\}$ эквивалентны. Мы допускаем возможность, что дуга может связывать вершину с собой, т.е. дуги вида $e = \{i\} \subset N$ допустимы. Таким образом, дуги – это подмножества из N, состоящие из одного или двух элементов.

Степенью узла i мы будем называть число инцидентных ему дуг, и обозначать d_i. Сумма степеней всех узлов графа в два раза больше общего числа дуг, так как дуга, инцидентная узлам i и j вносит вклад в степень обоих этих узлов:

$$\sum_{i=1}^{n} d_i = 2m \qquad \textbf{(1)}$$

Степенью подмножества узлов $V \subset N$ мы будем называть сумму степеней узлов, входящих в это подмножество: $d(V) := \sum_{i \in V} d_i$.

Под разбиением графа на k сообществ мы будем понимать разбиение множества узлов N на k непересекающихся множеств $N = \bigcup_{i=1}^{k} N_i : N_i \bigcap N_j = \varnothing, i \neq j$. Число элементов в группе N_i будем обозначать n_i. Далее обозначим множество дуг с узлами, находящимися в i-ой группе как E_i:

$E_i = \{e = \{k,l\} \in E | k,l \in N_i\}$, а $E_{i,j}$ – множество дуг, соединяющих две *разные* группы i и j: $E_{i,j} = \{e = \{k,l\} \in E | k \in N_i \text{ и } l \in N_j\}$. Число элементов в группах E_i и $E_{i,j}$ будем обозначать как m_i и $m_{i,j}$ соответственно.

Модулярность разбиения

Чтобы оценить качество разбиения графа на сообщества, введем ([1], [2], [3]) понятие модулярности, описывающее, насколько при заданном разбиении графа на группы плотность внутригрупповых связей больше плотности межгрупповых связей. Если мы будем просто максимизировать число связей внутри групп, то оптимальной структурой будет единственная группа, содержащая все узлы и все связи между ними, что не дает нам никакой информации. Поэтому в качестве метрики имеет смысл использовать не величину, описывающую насколько для данного разбиения внутригрупповые связи более плотные, чем межгрупповые, а насколько они более плотные по сравнению с некой *ожидаемой* величиной. Эта ожидаемая величина соответствует нулевой гипотезе, заключающейся в том, что дуги распределены между группами случайно, т.е. никакой закономерности в распределении плотности дуг внутри групп нет.

Введем понятие случайного эквивалентного графа $G' = (N, E')$. Для этого будем использовать понятие конфигурационной модели ([5]). Согласно этой модели для того, чтобы сконструировать случайный эквивалентный граф с тем же множеством вершин и с теми же степенями всех вершин (а, значит, и с тем же количеством дуг), мы выполняем следующие действия:

1. «Разрываем» все дуги графа на две половинки – полу-дуги:

Рис. 2 – *Исходный граф и граф с дугами, разорванными на полу-дуги*

В результате мы получаем структуру с тем же набором вершин, причем с каждой вершиной связано столько полу-дуг, сколько с ней было связано дуг в исходном графе.

2. Случайным образом соединяем пары полу-дуг друг с другом, чтобы получить «полноценные» дуги и, таким образом, новый случайный эквивалентный граф $G' = (N, E')$. В новом графе значения степеней всех узлов сохраняются, но дуги соединяют уже, возможно, другие пары узлов. Под случайным соединением полу-дуг в дуги мы понимаем то, что для случайно выбранной дуги вероятность того, что она была получена соединением данных двух полу-дуг одинакова для всех пар полу-дуг. При этом возможны дуги с обеими вершинами в одном узле и несколько дуг, соединяющих одну пару узлов.

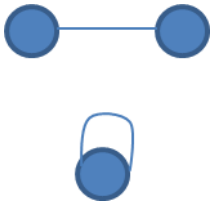


Рис. 3 – *Полу-дуги соединяются в пары случайным образом, и структура преобразуется в новый граф с теми же степенями узлов.*

Выберем из графа $G' = (N, E')$ случайным образом дугу e. Эта дуга была образована какой-то парой полу-дуг. Число способов выбора пары полу-дуг среди всех 2m полу-дуг для соединения в дугу e равняется:

$$\binom{2m}{2} = \frac{1}{2}(2m)(2m-1) \quad \textbf{(2)}$$

Для того, чтобы дуга $e \in E'$ соединяла узлы i и j, необходимо, чтобы она получилась объединением двух полу-дуг, связанных с этими узлами. Если узлы i и j различные, то существует $d_i d_j$ способов соединения дугой узла i с узлом j посредством объединения одной из d_i полу-дуг узла i с одной из d_j полу-дуг узла j. Если дуга соединяет один узел с самим собой, то она получена соединением двух полу-дуг, выходящих из узла i. Число способов выбора пары полу-дуг, исходящих из узла i, среди общего числа d_i полу-дуг равно $\binom{d_i}{2} = \frac{d_i(d_i-1)}{2}$. Т.е. число возможных соединений полу-дуг в дугу e, при которых эта дуга соединяет узлы i и j равно:

$$\begin{cases} d_i d_j, i \neq j \\ \frac{d_i(d_i-1)}{2}, i = j \end{cases} \quad \textbf{(3)}$$

Чтобы получить число возможных соединений полу-дуг в дугу e, при которых эта дуга соединяет узлы из группы N_l, просуммируем (3) по всем парам узлов $i \in N_l$ и $j \in N_l, i \neq j$ с множителем ½ (чтобы не суммировать каждую пару дважды) и по всем узлам $i \in N_l$ (чтобы учитывать возможность инцидентности дуги единственному узлу):

$$\frac{1}{2}\sum_{i\in N_l}\sum_{j\in N_l, j\neq i} d_i d_j + \sum_{i\in N_l}\frac{d_i(d_i-1)}{2} = \frac{1}{2}\left[\left(\sum_{i\in N_l}\sum_{j\in N_l} d_i d_j - \sum_{i\in N_l} d_i^2\right) + \sum_{i\in N_l} d_i^2 - \sum_{i\in N_l} d_i\right] =$$

$$= \frac{1}{2}\left[\left(\sum_{i\in N_l} d_i\right)^2 - \sum_{i\in N_l} d_i\right] = \frac{1}{2}\left[\left(d(N_l)\right)^2 - d(N_l)\right] = \frac{1}{2} d(N_l)\left(d(N_l)-1\right) \quad \textbf{(4)}$$

Вероятность того, что случайно выбранная дуга была образована объединением определенной пары полу-дуг одинакова для всех пар полу-дуг. Следовательно, вероятность того, что случайно выбранная дуга $e \in E'$ соединяет узлы из группы N_l равна отношению числа возможных соединений полу-дуг в дугу e, при которых она связывает узлы из N_l (4) к общему числу возможных соединений пар полу-дуг в дугу e (2):

$$P\left(e \in E', e \in N_l\right) = \frac{\frac{1}{2} d\left(N_l\right)\left(d\left(N_l\right)-1\right)}{\frac{1}{2}(2m)(2m-1)} = \frac{d\left(N_l\right)\left(d\left(N_l\right)-1\right)}{(2m)(2m-1)} \quad \textbf{(5)}$$

Таким образом, вероятность того, что случайно выбранная из графа $G' = \left(N, E'\right)$ дуга лежит целиком в одной какой-то группе (инцидента узлам из одной группы) равна сумме (5) по всем группам:

$$P\left(e \in E',\ \text{вершины } e \text{ находятся в одной группе}\right) =$$
$$= \sum_{l=1}^{k} P\left(e \in E',\ \text{вершины } e \text{ находятся в группе } N_l\right) = \quad \textbf{(6)}$$
$$= \sum_{l=1}^{k} \frac{d\left(N_l\right)\left(d\left(N_l\right)-1\right)}{(2m)(2m-1)}$$

Теперь рассмотрим исходный граф $G = \left(N, E\right)$. Число дуг, обе вершины которых находятся в группе N_l, равно m_l, следовательно, вероятность того, что случайно выбранная дуга из $G = \left(N, E\right)$ будет находиться в E_l равна: $P\left(e \in E,\ \text{вершины } e \text{ находятся в } N_l\right) = \frac{m_l}{m}$. Следовательно, вероятность того, что случайно выбранная дуга из $G = \left(N, E\right)$ будет находиться в одной какой-либо группе будет равна сумме:

$$P\left(e \in E,\ \text{вершины } e \text{ находятся в одной группе}\right) =$$
$$= \sum_{l=1}^{k} P\left(e \in E,\ \text{вершины } e \text{ находятся в группе } N_l\right) = \sum_{l=1}^{k} \frac{m_l}{m} \quad \textbf{(7)}$$

Вычитая (6) из (7), получаем метрику, равную вероятности того, что случайно выбранная дуга из исходного графа находится внутри одной группы, минус вероятность того же события, но для графа со случайным распределением дуг. Эта метрика дает представление о том, в какой степени разбиение на группы узлов исходного графа увеличивает вероятность попадания случайной дуги внутрь какой-то группы по сравнению с графом со случайной структурой.

$$Q := \sum_{l=1}^{k}\left[\frac{m_l}{m} - \frac{d\left(N_l\right)\left(d\left(N_l\right)-1\right)}{(2m)(2m-1)}\right] = \frac{1}{m}\sum_{l=1}^{k}\left[m_l - \frac{d\left(N_l\right)\left(d\left(N_l\right)-1\right)}{2(2m-1)}\right] \quad \textbf{(8)}$$

Метрика (8) служит для оценки качества разбиения графа на группы.

Если число дуг велико, то $2m \approx 2m-1$, а если группы имеют большую степень, то $d\left(N_l\right) \approx d\left(N_l\right)-1$, следовательно, (8) принимает приблизительный вид:

$$Q = \frac{1}{m}\sum_{l=1}^{k}\left[m_l - \frac{d\left(N_l\right)\left(d\left(N_l\right)-1\right)}{2(2m-1)}\right] \approx \frac{1}{m}\sum_{l=1}^{k}\left[m_l - \frac{\left(d\left(N_l\right)\right)^2}{4m}\right] \quad \textbf{(8')}$$

Приблизительное равенство (8') совпадает с определениями в [1], [2], [3].

Альтернативный подход к вычислению модулярности

Выражение ($8'$) можно также получить, используя другие рассуждения. Под событием A_i, будем понимать, что у случайно выбранной дуги e из G=(N, E) случайно выбранная вершина находится в группе N_i. При помощи $A_{i,j}$ мы будем обозначать событие, что у случайно выбранной дуги случайно выбранная вершина находится в группе N_i, а вторая вершина находится в группе N_j. Будем обозначать случайно выбранную вершину у случайно выбранной дуги n_1, а вторую вершину – n_2.

Если выбранная дуга находится в E_i, то обе ее вершины лежат в N_i, а значит, случайно выбранная вершина будет находиться в N_i с вероятностью 1. Если дуга находится в $E_{i,j}$, то случайно выбранная вершина будет находиться в N_i с вероятностью ½. Во всех остальных случаях дуга e не инцидента вершинам из N_i. Поэтому

$$P\left(A_i\right)=P\left(n_1\in N_i\right)=P\left(e\in E_i,n_1\in N_i\right)+\sum_{j=1,\,j\neq i}^{k}P\left(e\in E_{i,j},n_1\in N_i\right)=$$
$$=\underbrace{P\left(e\in E_i\right)}_{=m_i/m}\underbrace{P\left(n_1\in N_i\,\middle|\,e\in E_i\right)}_{=1}+\sum_{j=1,\,j\neq i}^{k}\underbrace{P\left(e\in E_{i,j}\right)}_{=m_{i,j}/m}\underbrace{P\left(n_1\in N_i\,\middle|\,e\in E_{i,j}\right)}_{=1/2}=\frac{m_i}{m}+\frac{1}{2}\sum_{j=1,\,j\neq i}^{k}\frac{m_{i,j}}{m} \quad \textbf{(9)}$$

С другой стороны, если мы просуммируем степени узлов в группе N_i, т.е. число дуг, инцидентных каждому узлу из N_i, то каждая дуга внутри группы N_i будет посчитаны два раза (потому что оба инцидентных ей узла находятся в группе N_i), а дуги с одной вершиной в другой группе будут посчитаны только один раз. Получаем:

$d\left(N_i\right)=2m_i+\sum_{j=1,\,j\neq i}^{k}m_{i,j}$, следовательно, с учетом (9):

$$P\left(A_i\right)=\frac{m_i}{m}+\frac{1}{2}\sum_{j=1,\,j\neq i}^{k}\frac{m_{i,j}}{m}=\frac{1}{2m}\left(2m_i+\sum_{j=1,\,j\neq i}^{k}m_{i,j}\right)=\frac{d\left(N_i\right)}{2m} \quad \textbf{(10)}$$

Если дуга находится в E_i, то вероятность того, что случайно выбранная вершина находится в N_i, а другая вершина также находится в N_i равна 1. Если дуга находится в $E_{i,j}$, то вероятность того, что случайно выбранная вершина находится в N_i, а другая вершина находится в N_j равна ½. Т.е.

$$P\left(A_{i,i}\right)=P\left(n_1,n_2\in N_i\right)=P\left(e\in E_i;n_1,n_2\in N_i\right)=$$
$$=\underbrace{P\left(e\in E_i\right)}_{=m_i/m}\underbrace{P\left(n_1\in N_i\,\middle|\,e\in E_i\right)}_{=1}\underbrace{P\left(n_2\in N_i\,\middle|\,e\in E_i,n_1\in N_i\right)}_{=1}=\frac{m_i}{m}$$
$$P\left(A_{i,j}\right)=P\left(n_1\in N_i,n_2\in N_j\right)=P\left(e\in E_{i,j};n_1\in N_i,n_2\in N_j\right)= \quad \textbf{(11)}$$
$$=\underbrace{P\left(e\in E_{i,j}\right)}_{=m_{i,j}/m}\underbrace{P\left(n_1\in N_i\,\middle|\,e\in E_{i,j}\right)}_{=1/2}\underbrace{P\left(n_2\in N_j\,\middle|\,e\in E_{i,j},n_1\in N_i\right)}_{=1}=\frac{m_{i,j}}{2m}$$

Определим случайный эквивалентный граф как граф, в котором принадлежность одной вершины дуги к определенной группе не дает информации о принадлежности к какой-либо группе другой его вершины, т.е.

$$P\left(A_{i,j}\right)=P\left(n_1\in N_i,n_2\in N_j\right)=P\left(n_1\in N_i\right)P\left(n_2\in N_j\right) \quad \textbf{(12)}$$

Вероятность того, что вторая вершина, инцидентная дуге, после случайного выбора первой вершины, будет находиться в группе N_j равна вероятности того, что случайно выбранная вершина будет находиться в N_j. Формально:

$$P\left(n_2 \in N_j\right) = \sum_{i=1}^{k} P\left(n_1 \in N_i, n_2 \in N_j\right) = P\left(n_1, n_2 \in N_j\right) + \sum_{i=1, i\neq j}^{k} P\left(n_1 \in N_i, n_2 \in N_j\right) =$$

$$= P\left(A_{i,i}\right) + \sum_{i=1, i\neq j}^{k} P\left(A_{i,j}\right) \underset{(11)}{=} \frac{m_i}{m} + \frac{1}{2}\sum_{i=1, i\neq j}^{k} \frac{m_{i,j}}{m} \underset{(9)}{=} P\left(n_1 \in N_j\right) \quad \textbf{(13)}$$

Из (12) и (13) получаем для случайного графа:

$$P\left(A_{i,j}\right) = P\left(n_1 \in N_i, n_2 \in N_j\right) = P\left(n_1 \in N_i\right) P\left(n_1 \in N_j\right) = P\left(A_i\right) P\left(A_j\right) \quad \textbf{(14)}$$

Вероятность принадлежности к i-ой группе обеих вершин случайно выбранной дуги для исходного графа равна: $P\left(A_{i,i}\right)$. А вероятность принадлежности к i-ой группе обеих вершин случайно выбранной дуги для случайного графа с учетом (14) примет вид: $P\left(A_{i,i}\right) = \left(P\left(A_i\right)\right)^2$.

Тогда вероятность принадлежности к одной какой-то группе обеих вершин случайно выбранной дуги для *исходного* графа минус вероятность того же события для *случайного* графа будет равна:

$$Q = \sum_{i=1}^{k}\left(P\left(A_{i,i}\right) - \left(P\left(A_i\right)\right)^2\right) \underset{(10),(11)}{=} \sum_{i=1}^{k}\left(\frac{m_i}{m} - \left(\frac{d\left(N_i\right)}{2m}\right)^2\right) \quad \textbf{(15)}$$

Таким образом, мы получили выражение для модулярности, совпадающее с (8′).

Жадный алгоритм максимизации модулярности

Так как модулярность (8’) описывает качество разделения графа на группы, то можно подойти к решению задачи отыскания оптимального разбиения графа, решая задачу максимизации модулярности. Однако решить эту задачу простым перебором практически невозможно, потому что число вариантов разделения n узлов на k групп растет, как минимум экспоненциально с ростом n. В [2] предложен жадный алгоритм оптимизации функции модулярности, основанный на пошаговом объединении двух групп, дающих наибольший прирост модулярности.

Рассмотрим некое разбиение узлов из N на k групп. Функция модулярности (8’) будет равна

$$Q_1 = \frac{1}{m}\sum_{l=1, l\neq i, l\neq j}^{k}\left[m_l - \frac{\left(d\left(N_l\right)\right)^2}{4m}\right] + \frac{1}{m}\left[m_i + m_j - \frac{\left(d\left(N_i\right)\right)^2 + \left(d\left(N_j\right)\right)^2}{4m}\right] \quad \textbf{(16)}$$

Теперь объединим группы i и j в одну, которую обозначим как $N_{i\cup j} = N_i \bigcup N_j$. Функция модулярности для нового графа примет вид:

$$Q_2 = \frac{1}{m}\sum_{l=1, l\neq i, l\neq j}^{k}\left[m_l - \frac{\left(d\left(N_l\right)\right)^2}{4m}\right] + \frac{1}{m}\left[m_{i\cup j} - \frac{\left(d\left(N_{i\cup j}\right)\right)^2}{4m}\right] \quad \textbf{(17)}$$

Число дуг внутри группы $N_{i\cup j}$ равно сумме дуг внутри групп N_i и N_j плюс число дуг между ними, т.е.

$$m_{i\cup j} = m_i + m_j + m_{i,j}$$

Степень объединённой группы $N_{i\cup j}$ равна сумме степеней групп N_i и N_j.

$$d\left(N_{i\cup j}\right) = d\left(N_i\right) + d\left(N_j\right) \Rightarrow \left(d\left(N_{i\cup j}\right)\right)^2 = \left(d\left(N_i\right)\right)^2 + \left(d\left(N_j\right)\right)^2 + 2d\left(N_i\right)d\left(N_j\right)$$

С учетом этого получаем из (16) и (17):

$$\Delta Q = Q_2 - Q_1 = \frac{1}{m}\left[m_{i,j} - \frac{2d\left(N_i\right)d\left(N_j\right)}{4m}\right] = \frac{1}{m}\left[m_{i,j} - \frac{d\left(N_i\right)d\left(N_j\right)}{2m}\right] \quad \textbf{(18)}$$

Из (18) следует, что наибольший рост модулярности происходит при объединении таких групп N_i и N_j, для которых величина

$$\Delta\left(N_i, N_j\right) := m_{i,j} - \frac{d\left(N_i\right)d\left(N_j\right)}{2m} \text{ максимальна.} \quad \textbf{(19)}$$

Из (19) видно, что объединение групп, между которыми нет дуг ($m_{i,j}=0$), не может дать увеличения модулярности.

Итак, **алгоритм последовательного объединения групп**:

1. Разделяем все множество узлов на n групп, где в каждую группу входит только один узел. Вычисляем модулярность разбиения.
2. Среди всех пар групп, между которыми существуют дуги, находим ту пару групп N_i и N_j, объединение которых в одну группу даст наибольший рост модулярности, что соответствует максимальному значению $\Delta\left(N_i, N_j\right) = m_{i,j} - \frac{d\left(N_i\right)d\left(N_j\right)}{2m}$. Эта величина может быть отрицательной, если лучшего разбиения нет, т.е. на определенных шагах качество разбиения может ухудшаться. Сохраняем значение модулярности для нового разбиения.
3. Если число групп больше одной, и существует хотя бы одна пара групп, между которыми есть дуги, переходим к шагу 2.
4. Находим разбиение, соответствующее максимальному значению модулярности.

Так как на каждом шаге 2 мы ищем только среди тех пар групп, между которыми есть дуги, то число рассматриваемых пар всегда ограничено сверху числом дуг, т.е. мы ищем среди O(m) пар.

Алгоритм, основанный на последовательном удалении «нагруженных» дуг

Алгоритм, основанный на последовательной максимизации функции модулярности, имеет недостаток, заключающийся в том, что на каждом шаге при принятии решения об объединении двух групп мы используем только локальную информацию: число дуг между этими двумя группами и их степени. В работе [1] был предложен алгоритм, использующий на каждом шаге глобальную информацию обо всей сети при помощи меры «промежуточности» (betweenness), определяемой для каждой дуги. Промежуточность дуги является мерой того, насколько часто она входит в кратчайшие

пути между различными парами узлов. Интуитивно понятно, что если два сообщества соединены небольшим числом дуг, то все пути между узлами из одного сообщества к узлам из другого сообщества должны будут проходить через эти несколько дуг. Подсчитывая количество раз, когда каждая дуга входит в кратчайший путь между всеми парами узлов графа, мы получаем меру промежуточности этой дуги.

Промежуточность считается следующим образом: для каждого узла S строится граф достижимости из узла S всех остальных узлов, как показано на рис. 4:

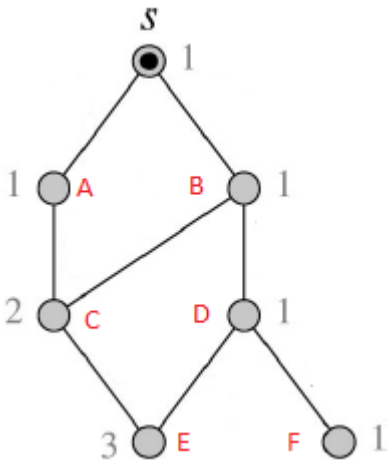
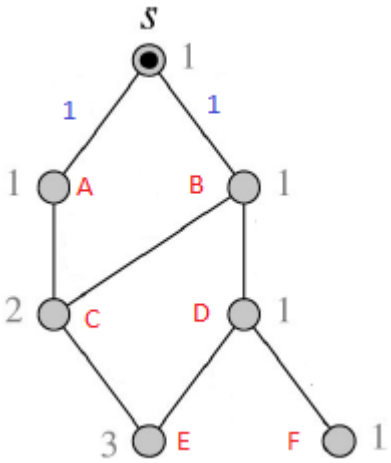
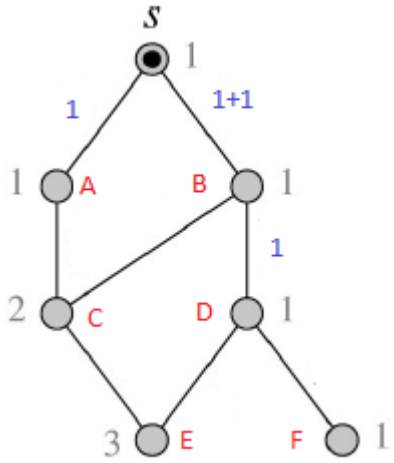


Рис. 4 - *Граф кратчайших путей из узла S.*

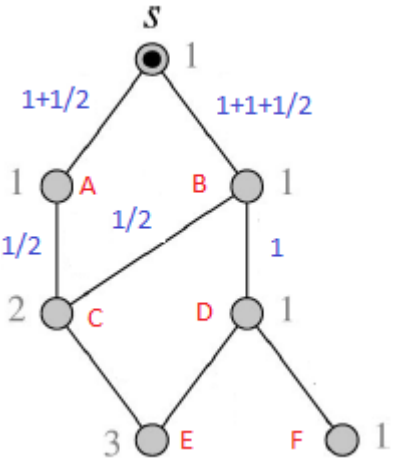
Каждый узел помечается числом кратчайших путей, при помощи которых в него можно попасть из узла S. Мы представляем себе, что из узла S надо доставить в каждый достижимый узел 1 единицу некой жидкости. Если существует несколько кратчайших путей, то поток жидкости разделяется. При передаче 1 единицы жидкости из узла S в узлы A и B по дугам (S,A) и (S,B) проходит 1 единица жидкости.



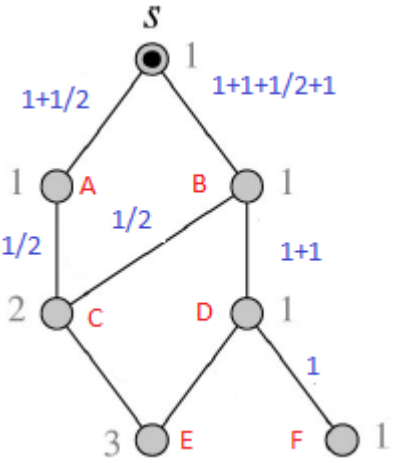
При передаче 1 единицы жидкости из узла S в узел D дуги (S,B) и (B,D) пропускают через себя по единице жидкости.



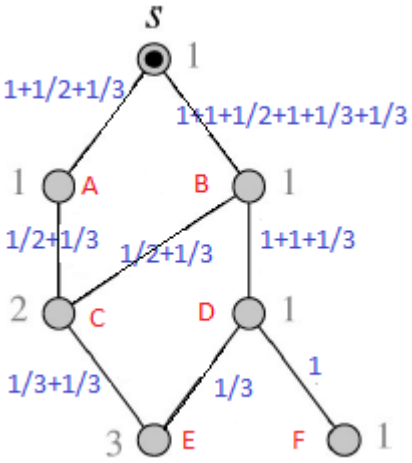
Между узлами S и C существуют два пути: S-A-C и S-B-C. Поэтому дуги, входящие в эти пути, при передаче 1 единицы жидкости между S и C пропускают дополнительный поток, равный 1/2.



Между S и F существует только один путь, поэтому дуги, в него входящие, т.е. S-B-D-F, получают дополнительный поток, равный 1 единице при передаче жидкости из S в F.



Между S и E существуют 3 пути: S-A-C-E, S-B-C-E, S-B-D-E. Поэтому добавляем 1/3 к каждой дуге, входящей в эти пути.



Получаем:

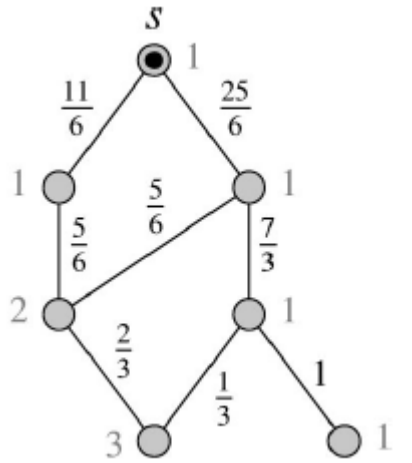


Рис. 5 – *Меры промежуточности для каждого пути из S*

Затем меры промежуточности для каждого пути из узла S суммируются по всем узлам S.

Идея алгоритма заключается в том, что мы последовательно удаляем из графа дуги с максимальной промежуточностью, при этом после каждого удаления снова пересчитываем промежуточность для всех дуг. Как только граф распадается на несвязные компоненты, мы получаем разбиение по этим компонентам связности. Для того, чтобы оценить качество этого разбиения, мы используем меру модулярности.

Алгоритм, основанный на последовательном удалении дуг с большой мерой промежуточности:

1. Рассчитать меры промежуточности для каждой дуги в графе. Рассчитать меру модулярности разбиения, состоящего из одной группы.
2. Выбрать дугу с наибольшей мерой промежуточности и удалить ее. Если таких дуг несколько – выбрать одну случайно.
3. Если после удаления дуги граф распался на несвязанные компоненты, то вычислить меру модулярности для разбиения, соответствующего этим компонентам.

4. Если в графе еще остались дуги, то пересчитать меры промежуточности для каждой оставшейся дуги. Переходим к шагу 2.
5. Выбрать разбиение, соответствующее максимуму меры модулярности.

Литература

1. Newman, Girvan “Finding and evaluating community structure in networks”, 2004
2. Newman, “Fast algorithm for detecting community structure in networks”, 2004
3. Newman, “Modularity and community structure in networks”, 2006
4. Aggarwal, “Social Network Data Analytics”, 2011
5. van der Hofstad, “Random Graphs and Complex Networks”, 2009