Занятие 2: Описательная статистика социальных сетей

НИУ ВШЭ

Москва, 2017

Описательная статистика социальных сетей

Описательная статистика для социальных сетей разделяется на две категории индексов:

- Статистика для всей сети. Рассматриваются характеристики всего графа (плотность, взаимность, транзитивность/коэффициент кластеризации). Такие статистики имеет смысл сравнивать для разных графов.
- Статистика для элементов сети: вершин и ребер. Рассматриваются характеристики отдельных элементов графа (степени центральности, близости, посредничества и т.д.). Такие характеристики имеет смысл сравнивать для разных вершин/ребер одного графа.

Плотность и средняя степень

Отношение существующего в графе числа ребер ко всем возможным называется **плотностью графа**.

Для направленного графа плотность вычисляется так:

$$D=\frac{2E}{N(N-1)}$$

Для ненаправленного так:

$$D=\frac{E}{N(N-1)}$$

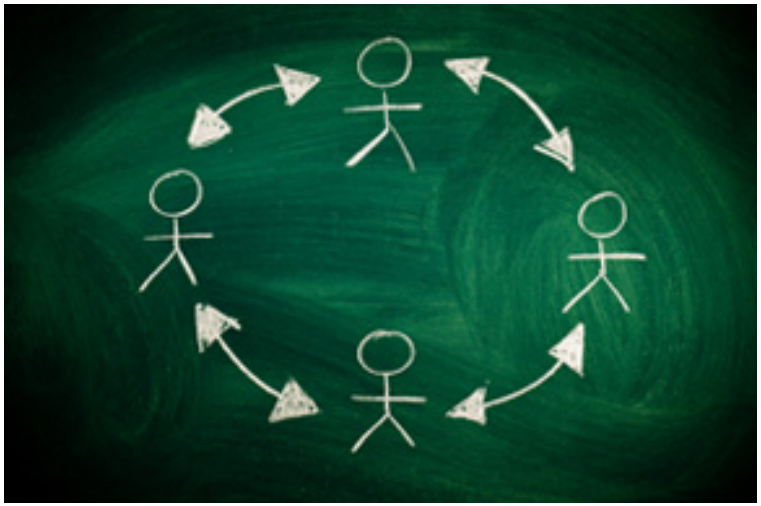
где D - плотность сети, E - число ребер, N - число вершин.

В подавляющем большинстве социальных сетей плотность очень низкая, так как число существующих ребер значительно меньше максимально возможного числа ребер.

Взаимность

Для направленных социальных сетей важным показателем является *взаимность (реципрокность)*, показывающая долю взаимных связей в графе

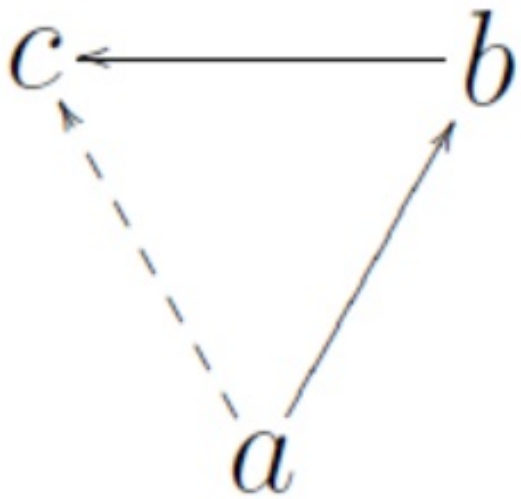
$$r = \frac{E_{\text{RECIP}}}{E}$$



В большинстве социальных сетей взаимность высока, часто близка к 50%.

Транзитивность и коэффициент кластеризации

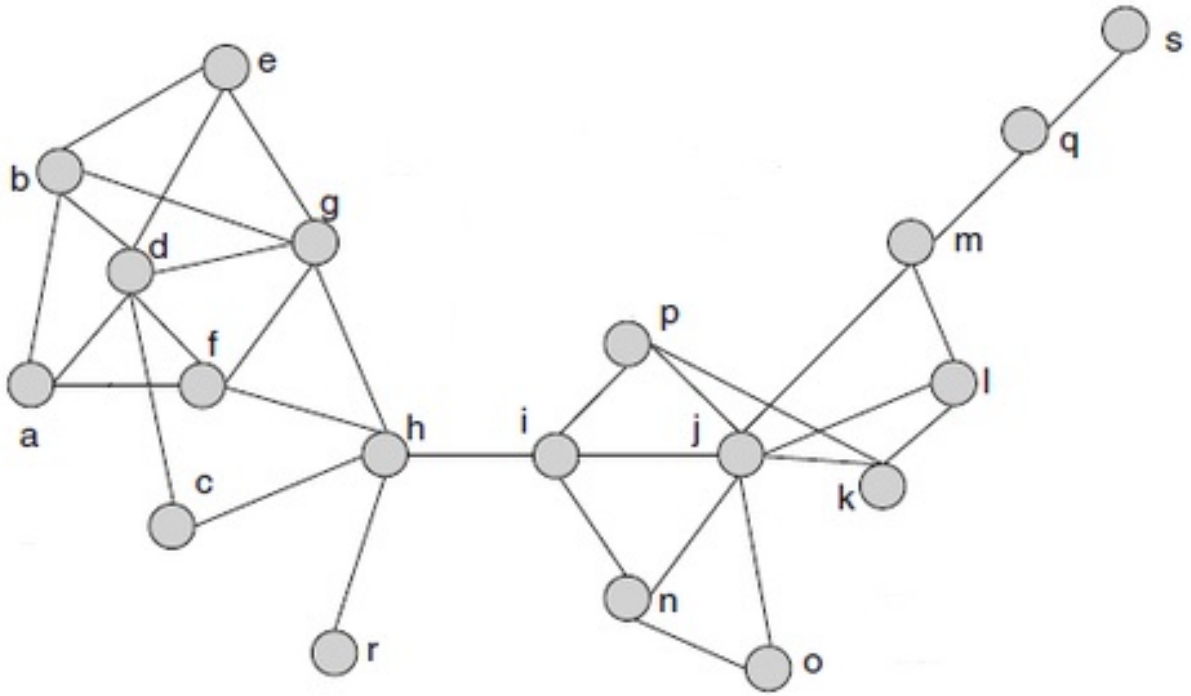
Транзитивность (коэффициент кластеризации) показывает, насколько справедливо выражение 'Друг моего дурга - мой друг'.
$C = \frac{3Xtriangle}{Triads}$



Для большинства социальных сетей характерен высокий коэффициент кластеризации.

Значимость вершин

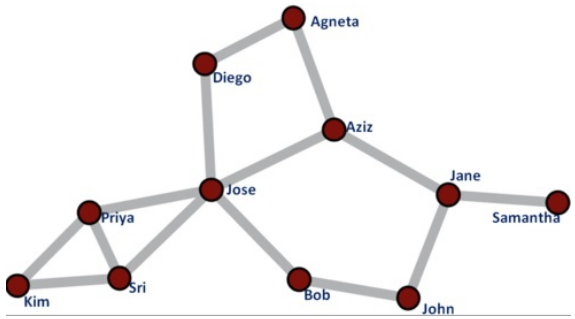
Какая из вершин в этой социальной сети наиболее значима?



На этот вопрос ответить не так просто. Рассмотрим, каким образом мы можем оценить значимость вершин.

Степень центральности (Degree Centrality)

Самый простой индекс, описывающий положение вершины в графе. Рассчитывается как число вершин, инцидентных данной (число вершин, связанных с целевой вершиной).



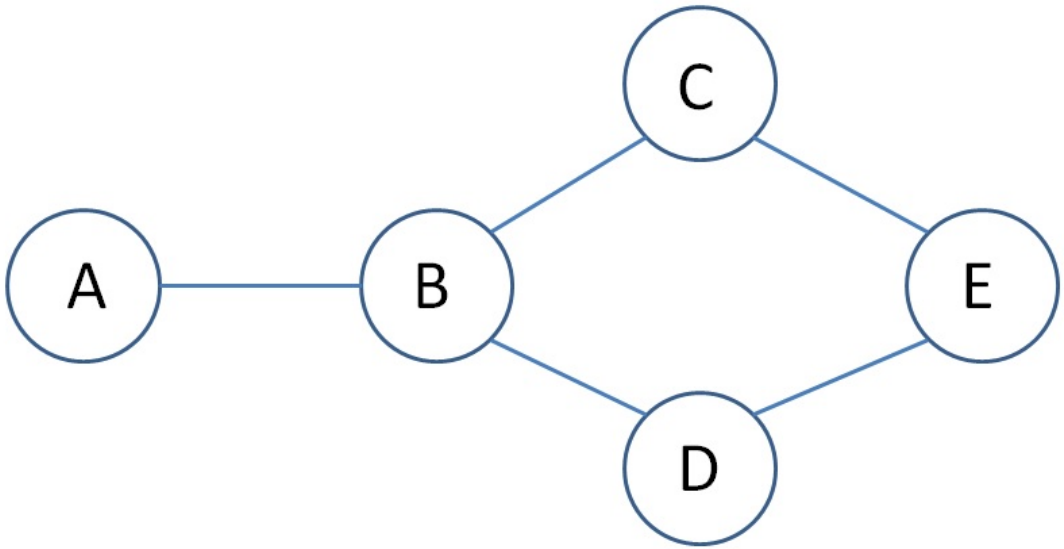
Рассчитывается как:

$C_d = \sum_i A_{ij}$

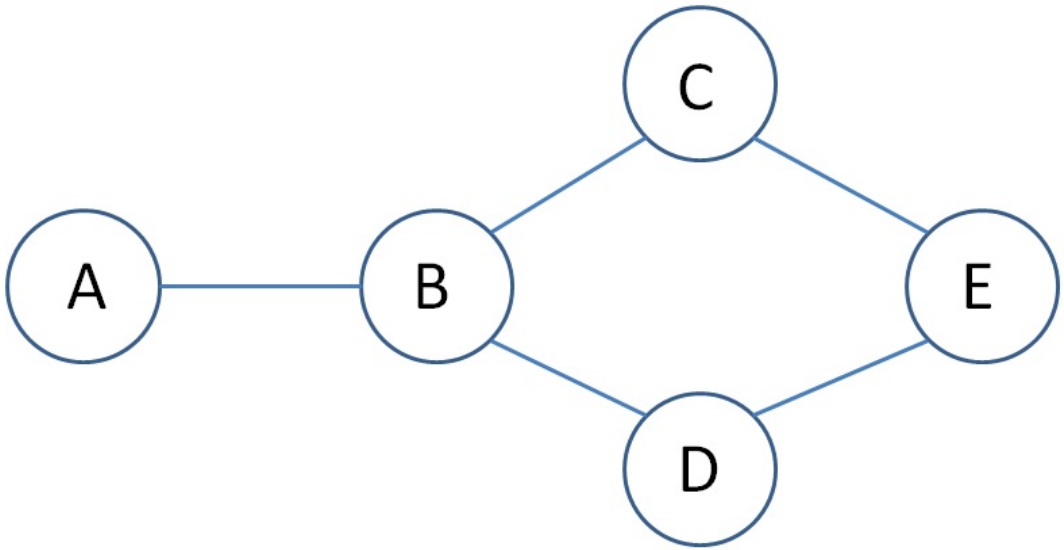
где C_d - степень центральности, а A_{ij} - матрица смежностей.

Описывает локальное положение вершины, но не характеризует ее глобальное положение в сети.

Степень центральности: Расчет на графе



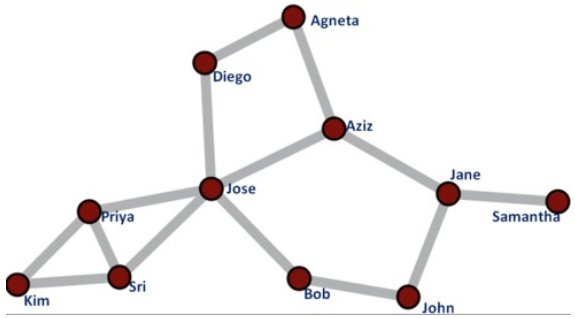
Степень центральности: Расчет на графе



$C_d(A)=1$
$C_d(B)=3$
$C_d(C)=2$
$C_d(D)=2$
$C_d(E)=2$

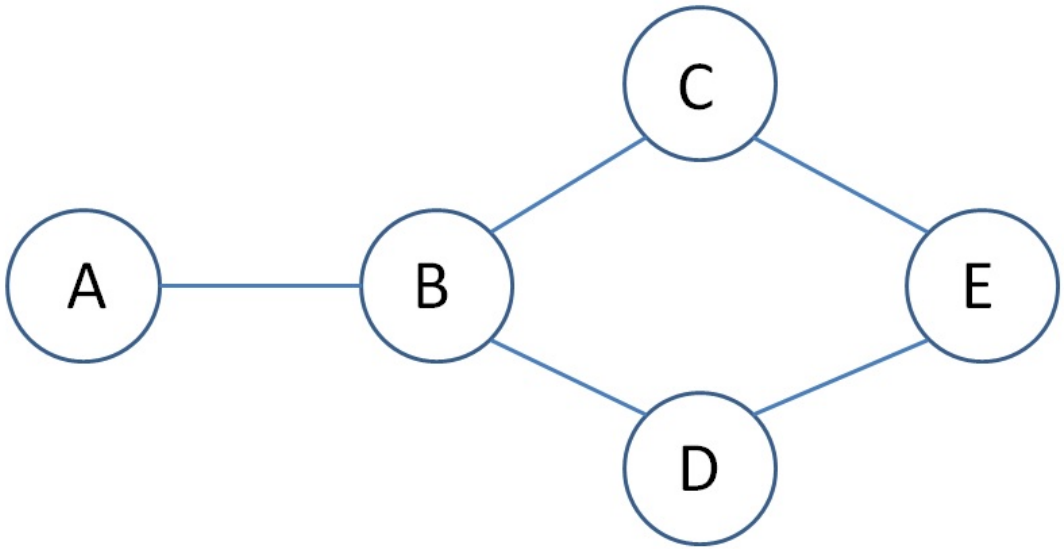
Степень близости (Closeness Centrality)

Индекс, который призван описать, насколько вершина близка к другим вершинам сети. Основная идея в том, что чем ближе актор к другим, тем проще ему наладить с ними взаимодействие (Wasserman and Faust, 1994).

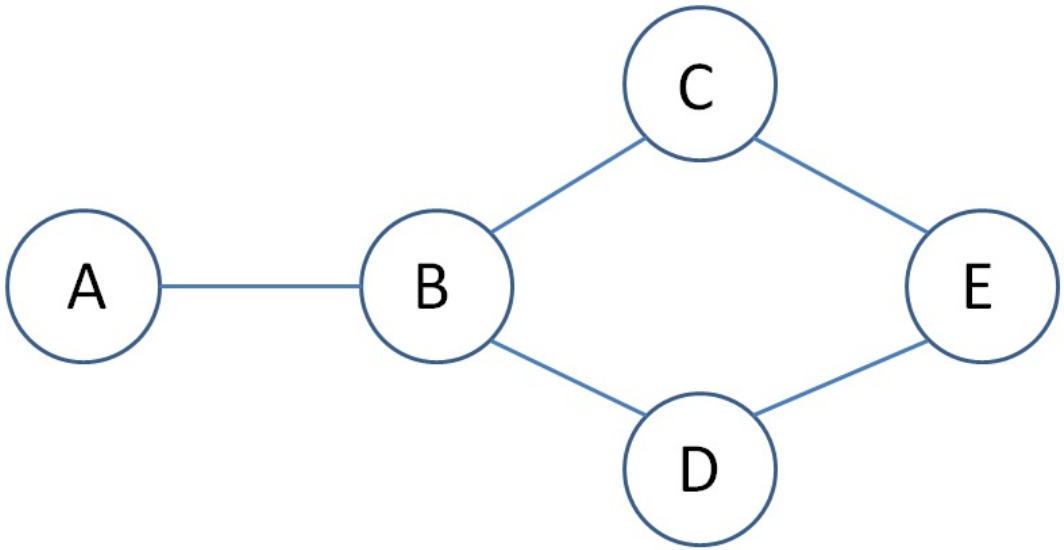


Рассчитывается как:
$C_{cl}=(\sum_i d(i,j))^{-1}$
где C_{cl} - степень близости, а $\sum_i d(i,j)$ - сумма путей от рассматриваемой вершины до всех остальных вершин графа.

Степень близости: Расчет на графе



Степень близости: Расчет на графе



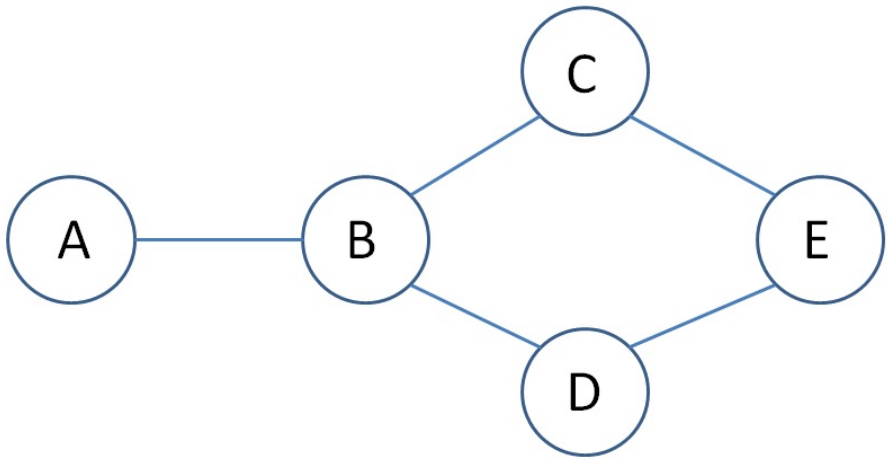
$C_d(A)=(1+2+3+2)^{-1}=1/8$
$C_d(B)=(1+1+1+2)^{-1}=1/5$
$C_d(C)=(1+1+2+2)^{-1}=1/6$
$C_d(D)=(1+1+2+2)^{-1}=1/6$
$C_d(E)=(1+1+3+2)^{-1}=1/7$

Степень близости: Расчет на графе

Иногда при расчетах степень близости стандартизируют на размер графа. В таком случае формула расчета выглядит следующим образом:

$C_{cl}=((\sum_i d(i,j))/(N-1))^{-1}$

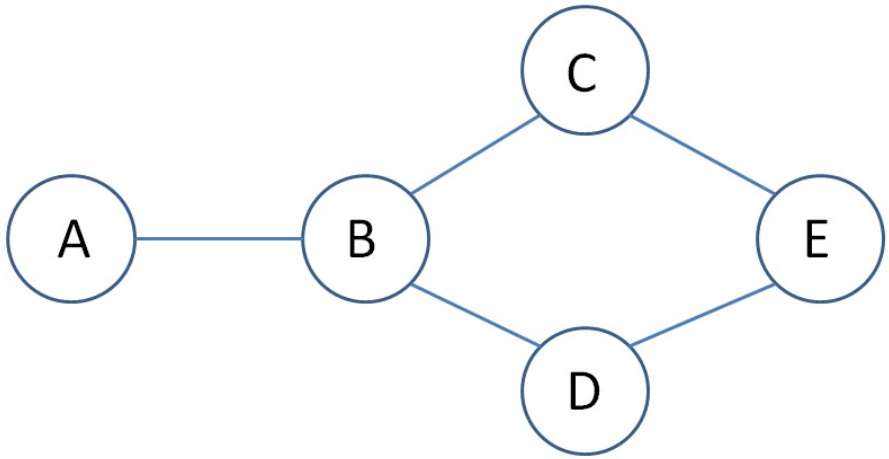
где C_{cl} - степень близости, а $\sum_i d(i,j)$ - сумма путей от рассматриваемой вершины до всех остальных вершин графа, а N - число вершин.



$C_{cl}(A)$=?; $C_{cl}(B)$=?; $C_{cl}(C)$=?; $C_{cl}(D)$=?; $C_{cl}(E)$=?

Стандартизированная степень близости

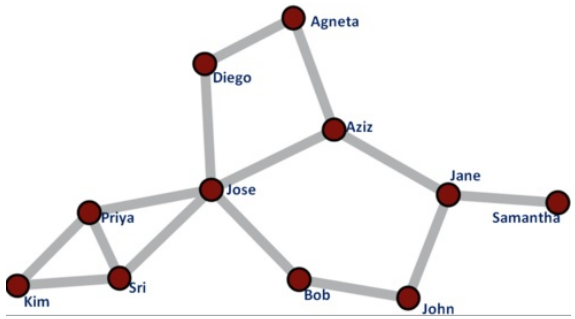
$C_{cl}=((\sum_i d(i,j))/(N-1))^{-1}$
где C_{cl} - степень близости, а $\sum_i d(i,j)$ - сумма путей от рассматриваемой вершины до всех остальных вершин графа, а N - число вершин.



$C_{cl}(A)=((1+2+3+2)/4)^{-1}=1/2$
$C_{cl}(B)=((1+1+1+2)/4)^{-1}=4/5$
$C_{cl}(C)=((1+1+2+2)/4)^{-1}=2/3$
$C_{cl}(D)=((1+1+2+2)/4)^{-1}=2/3$
$C_{cl}(E)=((1+1+3+2)/4)^{-1}=4/7$

Степень посредничества (Betweenness Centrality)

Показатель, характеризующий степень контроля над распространением информации (Wasserman and Faust, 1994). Основная идея в том, что чем больше кратчайших путей между вершинами контролирует актор, тем выше его контроль.



Рассчитывается так:

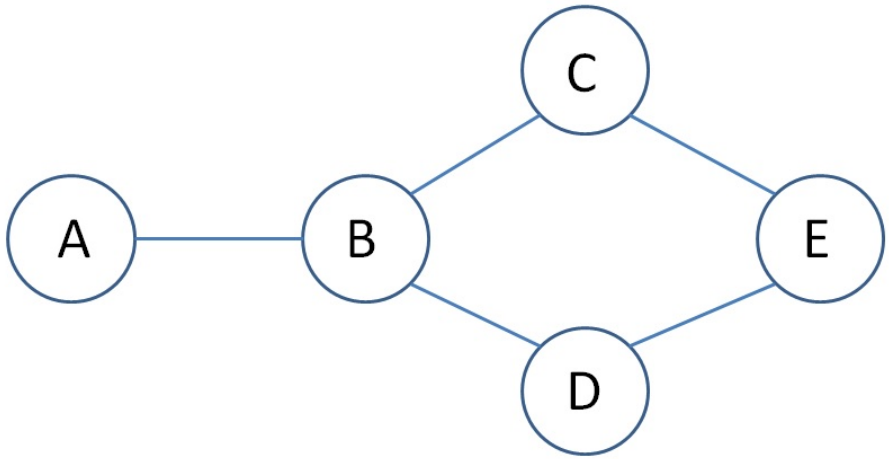
$C_b = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$

где C_b - степень посредничества, $\sigma_{st}(i)$ - число кратчайших путей между каждой парой вершин, на которых лежит вершина i, а σ_{st} - число кратчайших путей между каждой парой вершин.

Степень посредничества (Betweenness Centrality)

$C_b = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$

где C_b - степень посредничества, $\sigma_{st}(i)$ - число кратчайших путей между каждой парой вершин, на которых лежит вершина i, а σ_{st} - число кратчайших путей между каждой парой вершин.

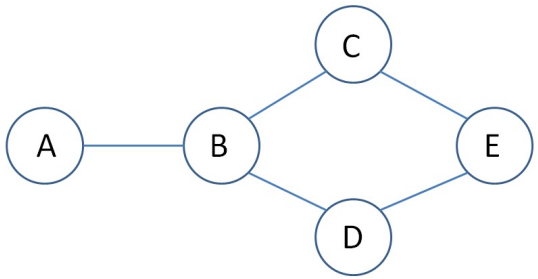


$C_b(A)=?;\ C_b(B)=?;\ C_b(C)=?;\ C_b(D)=?;\ C_b(E)=?$

Степень посредничества (Betweenness Centrality)

$C_b = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$

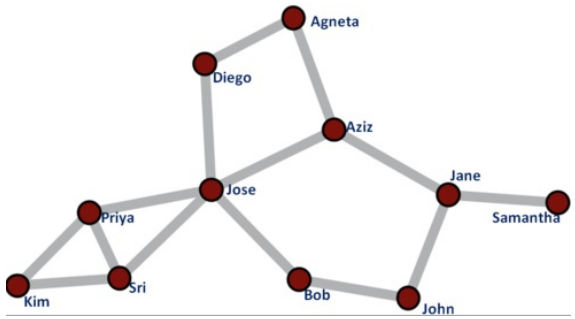
где C_b - степень посредничества, $\sigma_{st}(i)$ - число кратчайших путей между каждой парой вершин, на которых лежит вершина i, а σ_{st} - число кратчайших путей между каждой парой вершин.



$C_b(A)=0$
$C_b(B)=1$ (AC) $+1$ (AD) $+1$ (AE) $+0.5$ (CD)$=3.5$
$C_b(C)=0.5$ (AE) $+0.5$ (BE) $=1$
$C_b(D)=0.5$ (AE) $+0.5$ (BE) $=1$
$C_b(E)=0.5$ (CD)

Собственный вектор (Eigenvector centrality)

Мера влиятельности вершины в социальной сети.
'Важность вершины равна сумме важностей соседей данной вершины'
С чем более влиятельными вершинами соединена вершина, тем, соответственно, больше влияния у этой вершины.



$C_i = \frac{1}{k} * \sum_j A_{ij} c_j$
где C_i - собственный вектор (eigenvector) вершины i, k - нормировочный коэффициент, A_{ij} - матрица смежностей, C_j - собственный вектор вершины j.
Вопрос - для чего мы вводим нормировочный коэффициент? Какую проблему он призван решить?

Эйгенвектор (Eigenvector centrality)

Проблемой при вычислении эйгенвектора вершины является тот факт, что величина эйгенвектора вершины i зависит от величины эйгенвектора вершины j (рекурсивное определение).

$C_i = \frac{1}{k} * \sum_j A_{ij} c_j$

где C_i - эйгенвектор вершины i, k - нормировочный коэффициент, A_{ij} - матрица смежностей, C_j - эйгенвектор вершины j.

В матричной форме:

$AC = \lambda C$

где C - эйгенвектор (собственный вектор), λ - эйгензначение (собственное значение), A - матрица смежностей.

Вопрос: для каких сетей может быть использован собственный вектор?

Page Rank

Пейдж ранк (Brin and Page, 1998) - параметр, который отражает влиятельность (авторитетность) вершин. Этот алгоритм ранжирования вершин графа лежал в основе ранжирования **Google**. Изначально разработан и предложен для ***направленного графа***.
Вероятность того, что в результате случайного блуждания по страницам пользователь откроет определенную страницу, пропорциональна PageRank этой страницы.
$PR(A)=(1-d)+d(PR(T_1)/C(T_1)+...+PR(T_n/C(T_n)))$
где $PR(A)$ - пейдж ранк вершины A, $PR(T_1$ - пейдж ранк соседней с A вершины, $C(T_1$ - число соседей соседней с A вершины, d - вероятность попадания в вершину A в результате случайного блуждания.

Page Rank: Вычисление

Формирование матрицы вероятностей P. $P{=}D^{-1}A$,
где P - матрица вероятностей перехода, A - матрица смежностей.
На основании матрицы вероятностей мы строим стохастическую матрицу. Сумма элементов каждого из рядов такой матрицы должна равняться единице.
$P'{=}P + \frac{se}{n}$,
где P' - стохастическая матрица, P - матрица вероятностей перехода, s - вектор индикатор вершин, e - единичный вектор, n - число узлов сети.
Таким образом, PageRank может быть представлен в следующем виде:
$P''{=}\alpha P' + (1-\alpha)\frac{se}{n}$,
где P'' - PageRank, P' - стохастическая матрица, α - вероятность случайного перехода, s - вектор индикатор вершин, e - единичный вектор, n - число узлов сети.

Page Rank: Вариация

PageRank:
$P''=\alpha P' + (1-\alpha)\frac{ee}{n}$,
где P'' - PageRank, P' - стохастическая матрица, α - вероятность случайного перехода, s - вектор индикатор вершин, e - единичный вектор, n - число узлов сети.

- Какова должна быть мера случайностей при расчете PageRank?
- При каких α структура графа не оказывает влияние на переход?

HITS: Хабы и авторитетные вершины

Алгоритм HITS предложен Джоном Кляйнбергом для анализа сетей цитирования.
Основная идея в том, что в сети существуют два типа вершин:

- Авторитетные вершины (*authorities*) - популярные вершины, в контексте научных статей - статьи, на которые многие ссылаются;
- Активные вершины-хабы (*hubs*) - активные вершины, в контексте научных статей - статьи, которые ссылаются на нмогие работы.

Для каждой из вершин реальной социальной сети одновременно существуют два показателя - авторитетность и хабность.
$a_i = \sum_j A_{ji} h_j$
$h_i = \sum_j A_{ij} a_j$
Вопрос: для каких случаев нерелевантен расчет хабности и авторитетности? В каком случае они равны? С каким другим показателем степени центральности в таком случае они совпадают?

Значимость вершин

Возвращаемся к вопросу о том, какая из вершин в этой социальной сети наиболее значима?