Математика для Data Science

Математические основы

Владимир Анатольевич Судаков

доктор технических наук,

Профессор кафедры 806 МАИ

sudakov@ws-dss.com

Telegram: @vladimir_255

Вероятность

Теория вероятностей обеспечивает формальную основу для рассуждений о вероятности событий.

Вероятность $p(s)$ результата s удовлетворяет:

$0 <= p(s) <= 1$

$$\sum_{s \in S} p(s) = 1$$

Эти базовые свойства часто нарушаются при случайном использовании слова «вероятность» в науке о данных.

Вероятность против статистики

- Вероятность занимается прогнозированием вероятности будущих событий, а статистика анализирует частоту прошлых событий.
- Вероятность — это теоретическая часть математики, посвященная следствиям определений, а статистика — это прикладная математика, пытающаяся осмыслить наблюдения из реального мира.

Сложные события и независимость

Предположим, половина моих студентов — девушки (событие А).

Уровень половины моих студентов выше среднего (событие B).

Какова вероятность того, что студент одновременно является А и B?

События А и B независимы, если

$$P(A \cap B) = P(A) \times P(B)$$

Независимость (нулевая корреляция) хороша для упрощения вычислений, но плоха для прогнозирования.

Условная вероятность

Условная вероятность P(A|B) определяется как:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

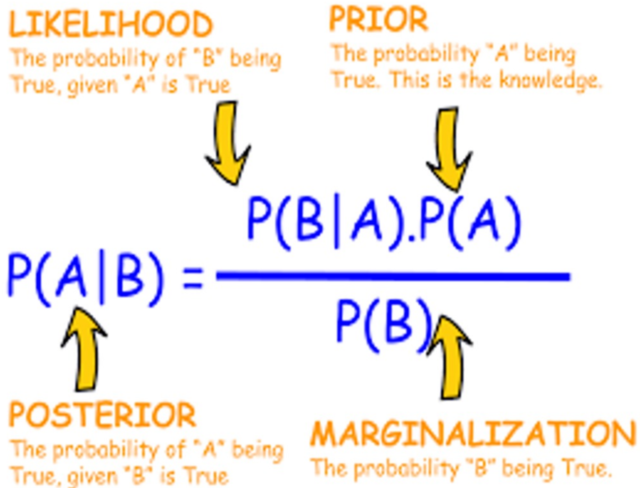
Условные вероятности становятся интересными только тогда, когда события **не** являются независимыми, иначе:

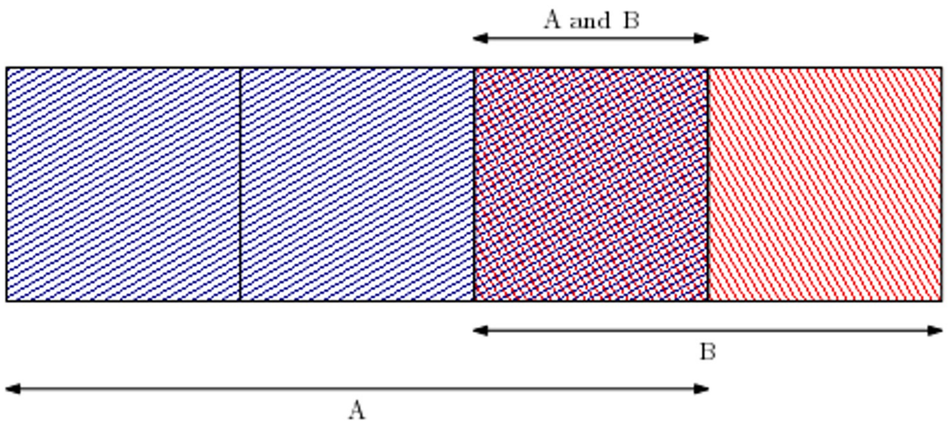
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Теорема Байеса

Теорема Байеса позволяет нам обновить нашу оценку правдоподобия A в ответ на знание B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$





Доказательство теоремы Байеса

The probability of two events A and B happening, P(A∩B), is the probability of A, P(A), times the probability of B given that A has occurred, P(B|A).

$$P(A \cap B) = P(A)P(B|A) \tag{1}$$

On the other hand, the probability of A and B is also equal to the probability of B times the probability of A given B.

$$P(A \cap B) = P(B)P(A|B) \tag{2}$$

Equating the two yields:

$$P(B)P(A|B) = P(A)P(B|A) \tag{3}$$

and thus

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)} \tag{4}$$

Типичные задачи статистики

- Определить закон распределения случайной величины (системы случайных величин)
- Проверить правдоподобие гипотез
- Найти неизвестные параметры распределения

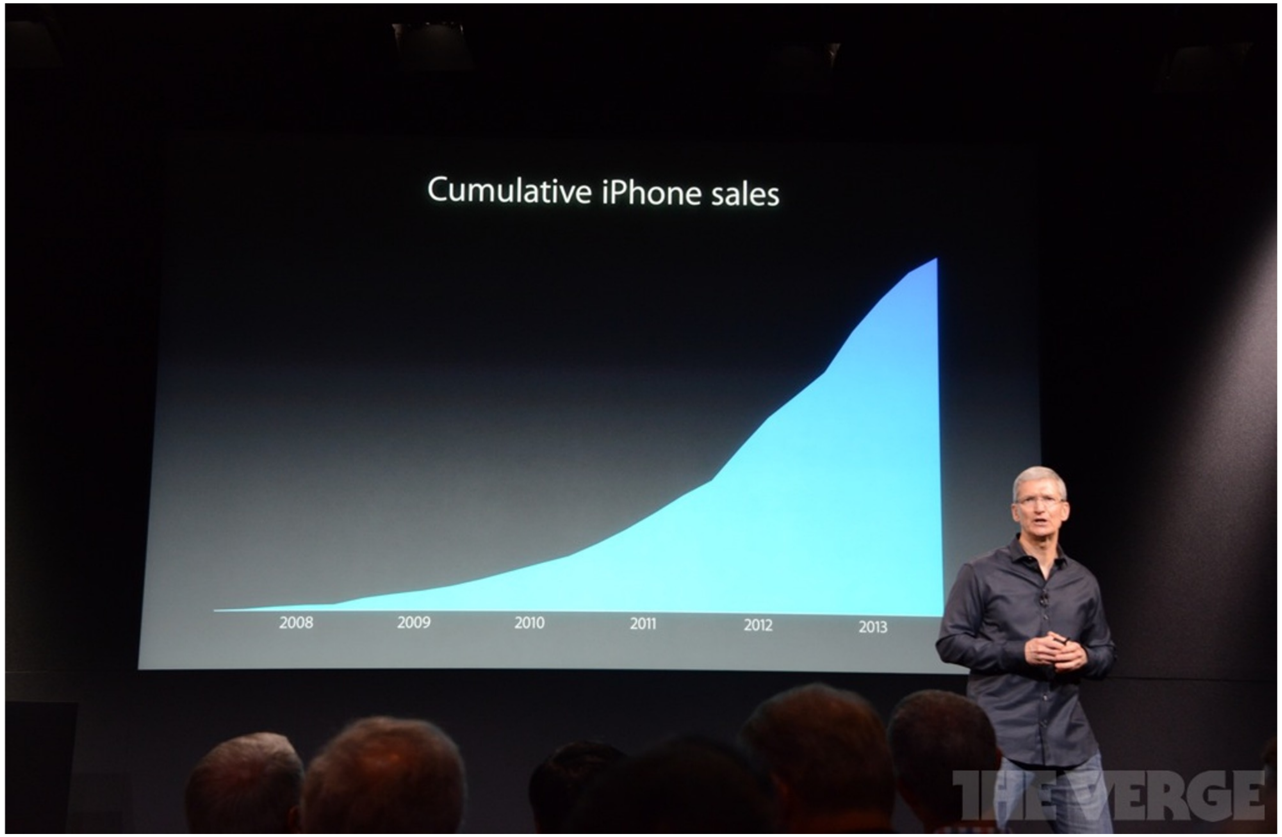
Простая статистическая совокупность

- Дана случайная величина X
- Совокупность наблюдаемых значений X - простой статистический ряд (простая статистическая совокупность)
- Статистическая функция распределения X:

$$F^*(x) = P^*(X<x)$$

Визуализация распределений

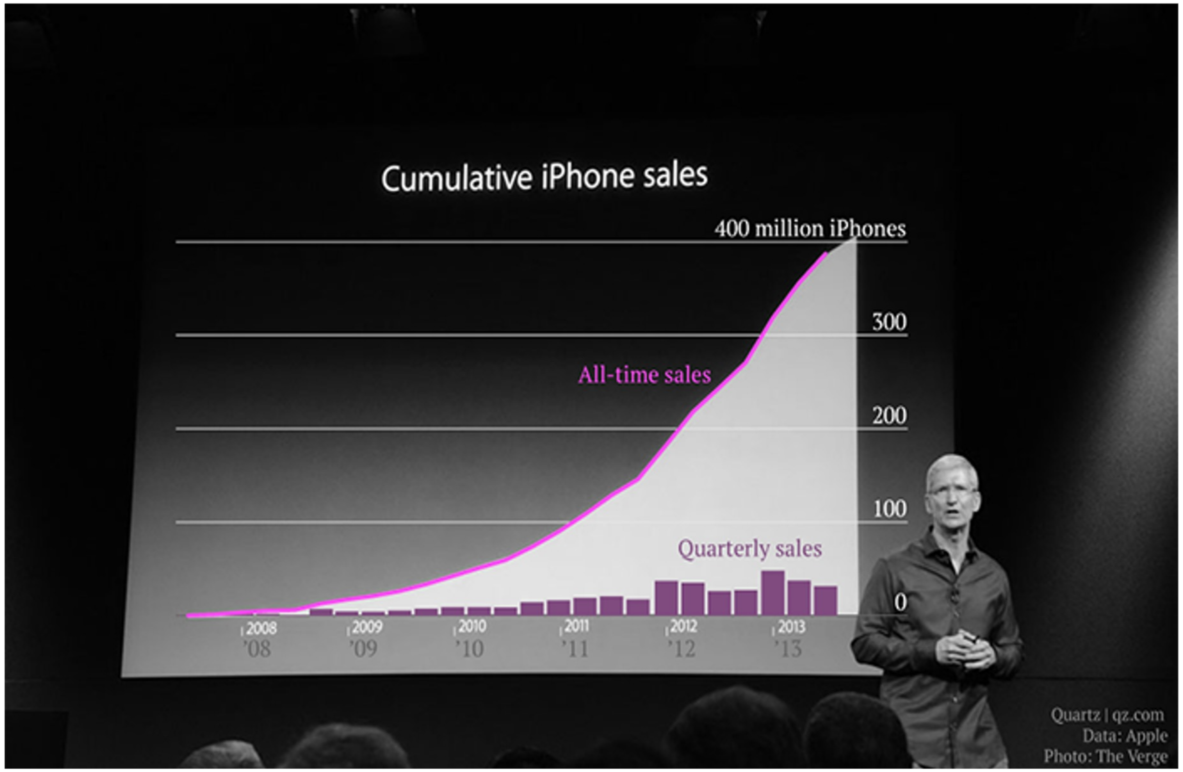
Продажи Apple iPhone стремительно растут, не так ли?



Насколько взрывным является этот рост на самом деле?

Кумулятивные распределения дают ошибочное представление о темпах роста.

Постепенное изменение является производной этой функции, которую трудно визуализировать.



Тренировка

- Дан ряд углов скольжения самолета в момент сбрасывания бомбы

-20,-60,-10, 30, 60, 70, -10,

-30,-120, -100, -80, 20, 40, -60,

-10, 20, 30, -80, 60, 70

- Построить статистическую функцию распределения
- У кого хорошее решение? Какие ошибки типичны на графике

Гистограмма

- Если данных много, то простой статистический ряд не удобен
- Разделим наблюдения на разряды и посчитаем частоты попадания:

$$p_i^* = \frac{m_i}{n}$$

- Таблица с интервалами разрядов и p_i^* называется статистическим рядом

I_i	$x_1;\ x_2$	$x_2;\ x_3$	...	$x_i;\ x_{i+1}$	...	$x_k;\ x_{k+1}$
p_i^*	p_1^*	p_2^*	...	p_i^*	...	p_k^*

Тренировка

- Давайте построим гистограмму по прошлым данным. У кого самая удобная получилась?

Построение статистической функции распределения

$$F^*(x_1) = 0;$$

$$F^*(x_2) = p_1^*;$$

$$F^*(x_3) = p_1^* + p_2^*;$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$F^*(x_k) = \sum_{i=1}^{k-1} p_i^*;$$

$$F^*(x_{k+1}) = \sum_{i=1}^{k} p_i = 1$$

Описательная статистика

Описательная статистика предоставляет способы фиксации свойств данного набора данных/выборки.

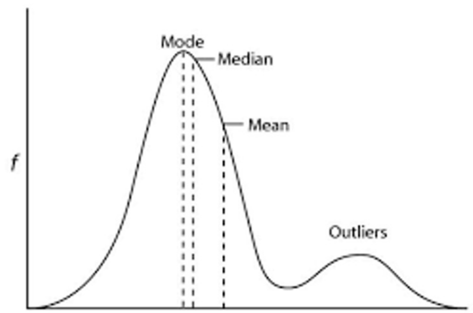
- Меры центральной тенденции описывают центр распределения данных.
- Меры вариации или изменчивости описывают разброс данных, т.е. насколько далеко измерения лежат от центра.

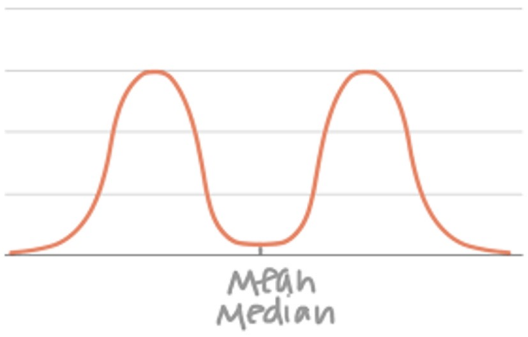
Мера центральности: среднее значение

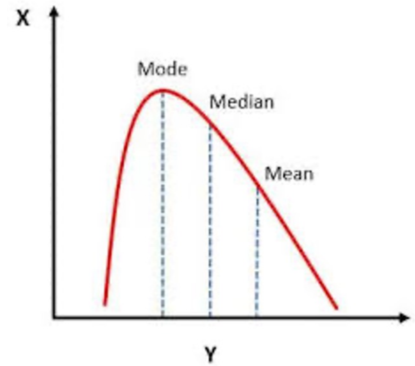
Чтобы вычислить среднее значение, просуммируйте значения и разделите их на количество наблюдений:

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Среднее значение имеет смысл для симметричных распределений без выбросов.







Другие меры центральности

Медиана представляет собой «серединное» значение.

Среднее геометрическое — это корень n-й степени из произведения n значений:

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Среднее геометрическое всегда <= среднее арифметическое и более чувствительно к значениям, близким к нулю.

Геометрические средние имеют смысл с соотношениями:

1/2 и 2/1 должны в среднем давать 1.

Какая мера лучше всего?

Среднее значение имеет смысл для симметричных распределений без выбросов: например. рост и вес.

Медиана лучше подходит для асимметричных распределений или данных с выбросами: например, богатство и доход.

Билл Гейтс добавляет 250 долларов к среднему доходу на душу населения, но ничего не добавляет к медиане.

Показатель отклонения: стандартное отклонение

Дисперсия представляет собой квадрат сигмы стандартного отклонения.

Мы делим на n или n-1?

$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(x_i - \overline{x})^2}{n-1}}$$

СКО генеральной совокупности делится на n, СКО выборки на n-1, но для больших n n ~ (n-1), так что это не имеет особого значения.

Интерпретация дисперсии (фондовый рынок)

Отношение «сигнал/шум» измерить сложно, поскольку многое из того, что вы видите, — это всего лишь дисперсия.

Рассмотрите возможность измерения относительного «навыка» различных инвесторов фондового рынка.

Ежегодные колебания эффективности фондов таковы, что результаты деятельности инвесторов случайны, а это означает, что реальная разница в навыках незначительна.

Интерпретация дисперсии (много моделей)

Обычно для каждой задачи мы разрабатываем несколько моделей, от очень простых до сложных.

Некоторая разница в производительности будет объяснена простой дисперсией: какие пары обучения/оценки были выбраны, насколько хорошо были оптимизированы параметры и т. д.

Небольшой выигрыш в производительности является аргументом в пользу более простых моделей.

Методы уменьшения дисперсии

Хотя идти на занятия пешком медленнее, чем ехать на автобусе, разница во времени прибытия меньше.

Повторение эксперимента несколько раз уменьшает дисперсию (перекрестная проверка в k-кратном размере).

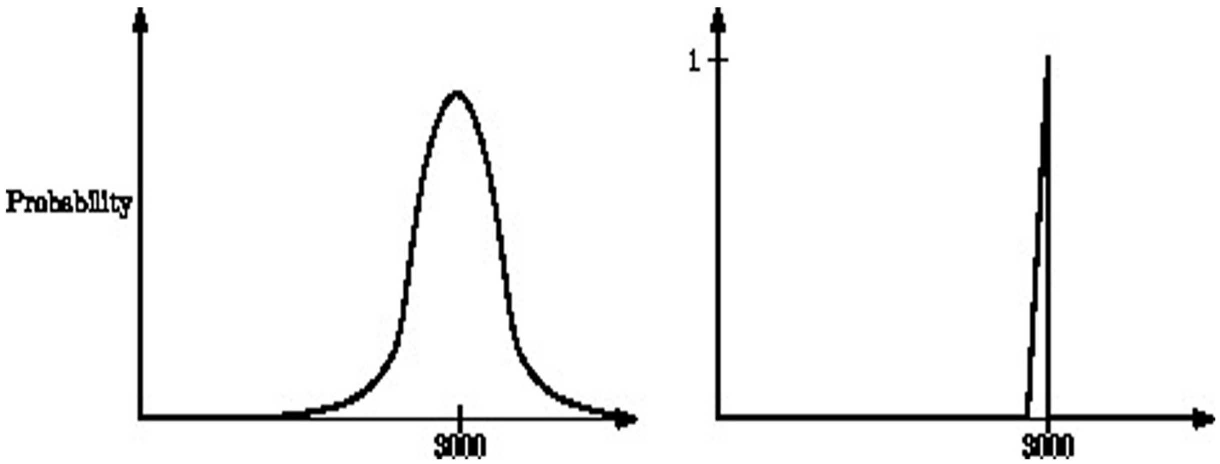
То же самое относится и к правильной случайной и детерминированной выборке.

Устранение выбросов (если это оправдано) уменьшает дисперсию.

Распределение срока службы картриджей принтера

Распределения с одинаковым средним значением могут выглядеть очень по-разному.

Но вместе среднее и стандартное отклонение довольно хорошо характеризуют любое распределение.



Центральность в графе

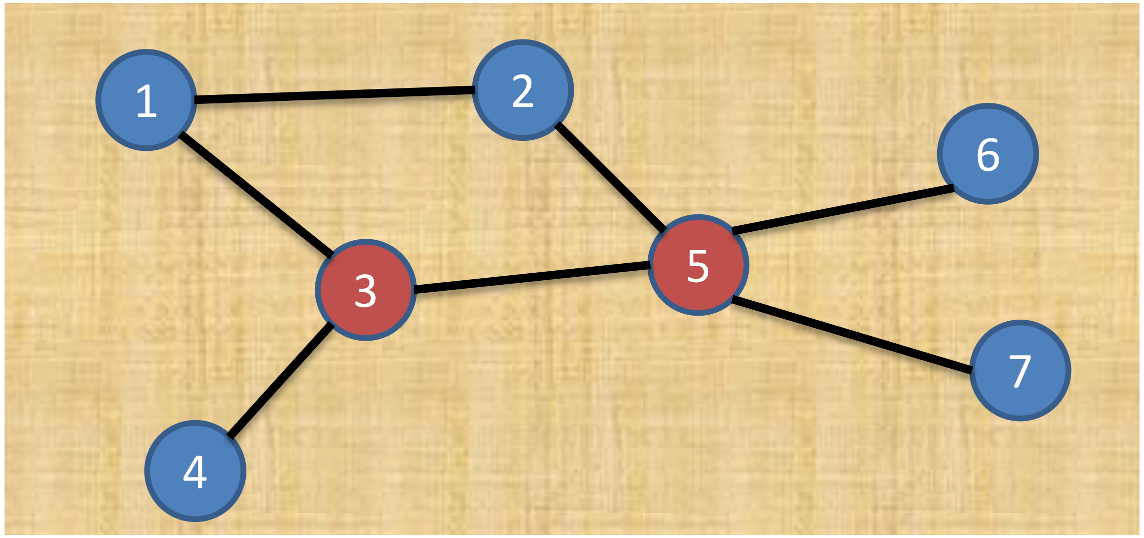
Центральность вершин в графе – это вектор, сопоставляющей каждой вершине графа некоторое число (индекс).

Наиболее распространенные индексы:

- Степенная центральность (degree centrality);
- Центральность по близости (closeness centrality);
- Центральность по посредничеству (betweenneess centrality);
- Центральность по собственному вектору (eigenvector centrality);
- Центральность PageRank.

Центральность по близости

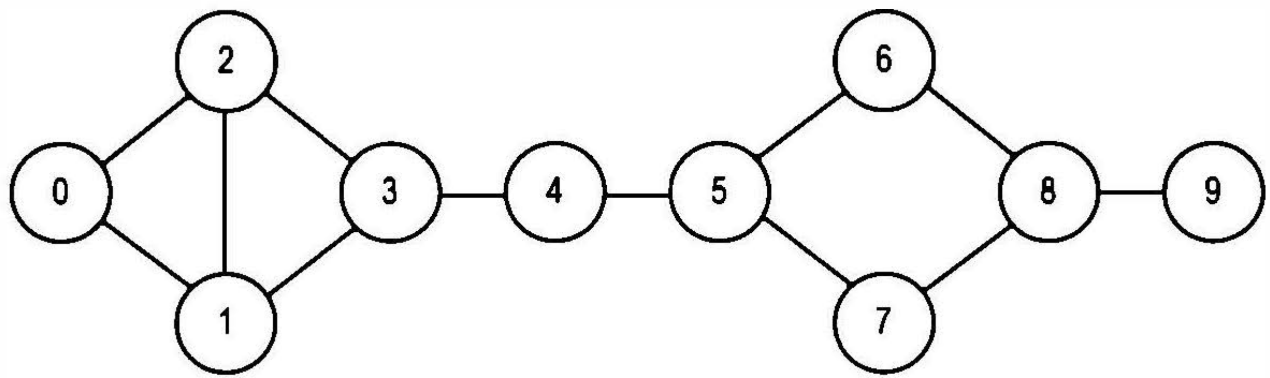
Вершина, находящаяся ближе всех к другим вершинам сети, является наиболее центральной



$$C_i = \frac{1}{\sum_j d_{ij}} \quad C_i = \sum_j \frac{1}{d_{ij}}$$

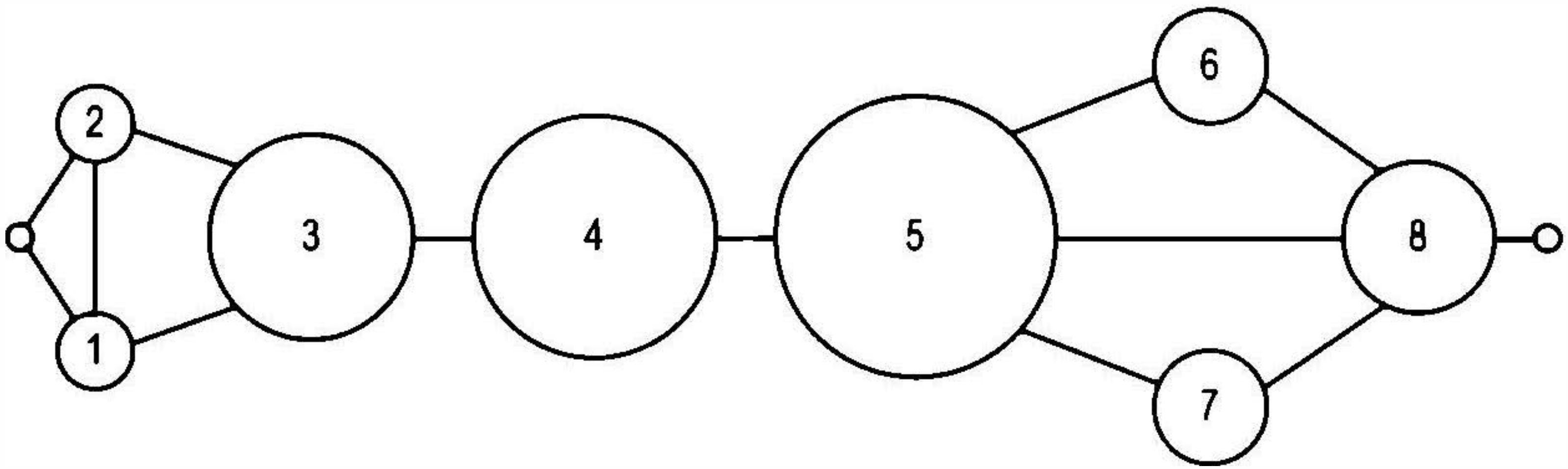
Центральность по посредничеству

Вершина, через которую проходит наибольшее число кратчайших путей, является наиболее центральной.



$$C_i = \sum_{jk} \frac{w_{jk}(i)}{w_{jk}}$$

Центральность по посредничеству



Центральность по собственному значению

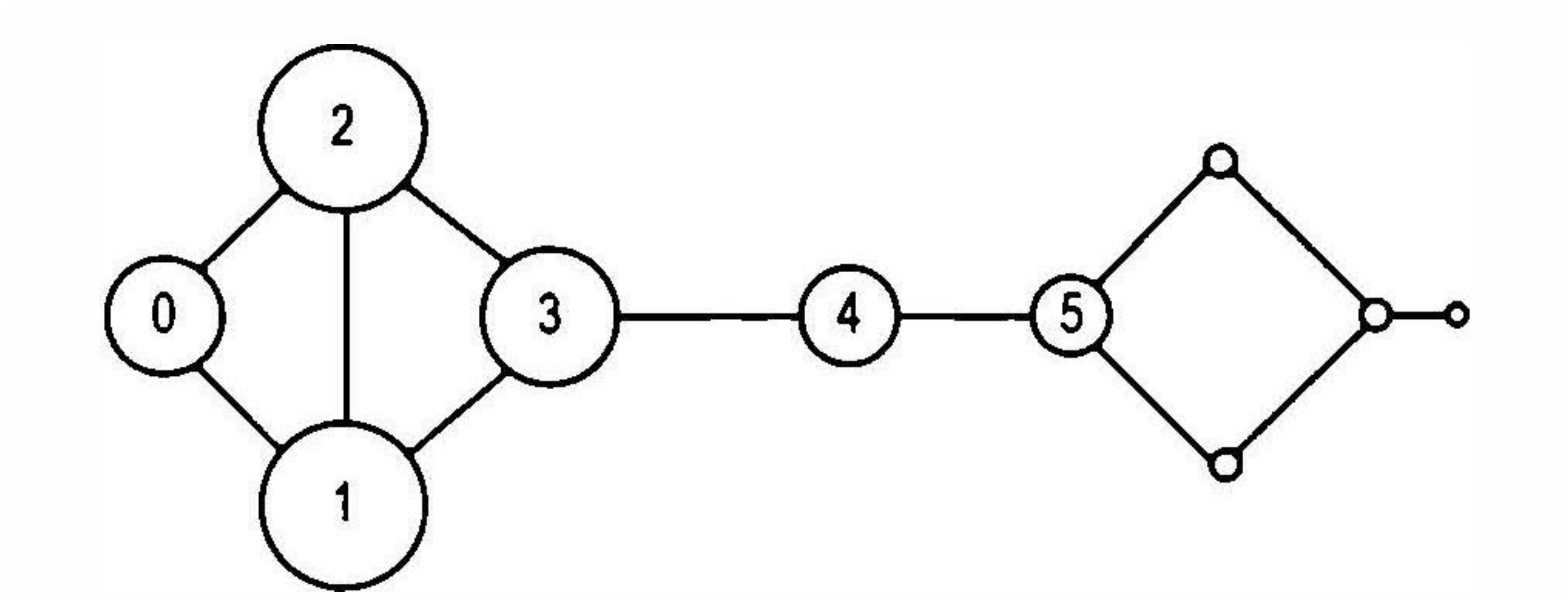
Центральность вершины i зависит от центральностей соседей вершины i.

$$x_i = \frac{1}{\lambda} \sum_{j \in F_i} x_j = \frac{1}{\lambda} \sum_j a_{ij} x_j$$

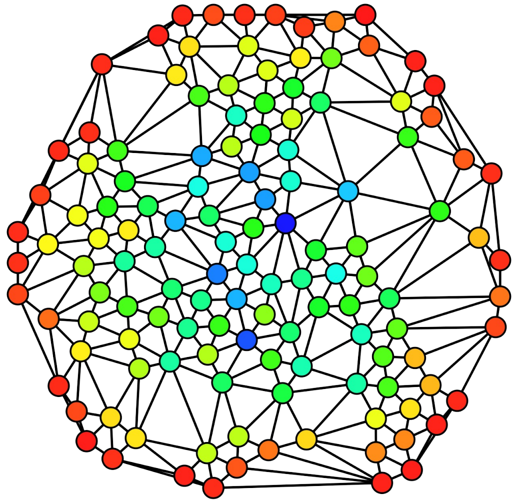
$$\lambda x = Ax$$

- Выбирается собственный вектор, соответствующий максимальному собственному значению.
- Данная центральность учитывает дальние взаимодействия.
- Наиболее центральными считаются вершины, которые сами указывают на сильные вершины.

Центральность по собственному значению



Задача



- Давайте соберем информацию о друзьях и друзьях Ваших друзей из VK по всем членам Вашей группы
- Построить граф дружбы. Все отношения дружбы между найденными профилями должны быть в графе
- Оценить центральность членов Вашей группы: по посредничеству, по близости, собственного вектора
- Визуализировать граф, по возможности красиво и информативно
- Вывести имена людей с максимальными центральностями из Вашей группы.