Научно Исследовательский Семинар 2017

Математические модели в экономике. часть 2.

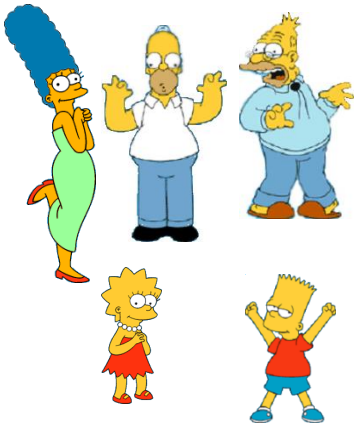
Sergei Koltcov skoltsov@hse.ru https://linis.hse.ru

Содержимое

4. Кластерный анализ. Задачи и методы кластерного анализа. Проблемы кластерного анализа. Применение кластерного анализа в Rstudio.
5. Сетевой анализ. Теория шести рукопожатий. Параметры в сетевом анализе. Кластерный анализ в исследовании сетей. Применение кластерного анализа в NodExl
6. Задачи линейного программирования. Математическая постановка. Графическое решение задачи.

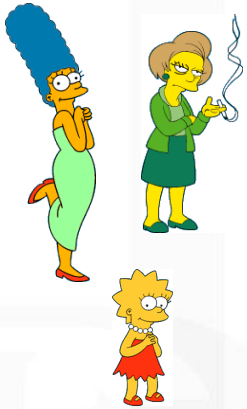
Задачи и методы кластер - анализа

Кластеризация – это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров.



Семья

Сотрудники



Женщины

Мужчины

Лейбелинг групп – то что нужно найти

Лейбелинг достаточно субъективен и зависит от цели пользователя

Задачи и методы кластер - анализа

Процедура кластеризации – зависит от меры сходства или не сходства. Такие меры выражаются виде функций расстояний, выраженных в виде той или иной функции.

Сходство тяжело определить

"*We know it when we see it*"

Задача определения сходства является задачей Machine learning.

Применение кластер - анализа

1. Статистика
1. Распознавание образов
2. Машинное обучение
3. Финансовая математика
4. Автоматическая классификация в различных областях науки (например, в археологи, биологии (кластеризация видов животных и растений))
5. Маркетинг. Маркетологи выделяют группы с целью оптимизации рекламной деятельности, оптимизации логистической деятельности.
6. Исследование свойств ДНК
7. Страхование (цель выделения групп населения и соотнесение групп с геогрф. расположением, заработком, семейным статусом и другой..)
8. Городское планирование.
9. Финансовое планирование города, района….
10. Социологические исследования.

Направления в кластер - анализе

Partitioning approach: плоская кластеризация - предполагает разделение объектов на кластеры сразу, причем один объект относится только к одному кластеру.

Typical methods**:** **K-means**, k-medoids, CLARANS

Fuzzy approach: Метод нечеткой кластеризации позволяет разбить имеющееся множество объектов p на заданное число нечетких множеств, то есть один и тот же объект может принадлежать разным классам. Принадлежность характеризуется степенью принадлежности, например вероятностью.

Typical methods:**C-means** (C-средних)

Hierarchical approach: Восходящая/нисходящая кластеризации: Иерархическая кластеризация (восходящая) - допускаем наличие подкластеров, осуществляется в несколько приемов, в результате образуется в иерархическое дерево (дендрограмму).

Typical methods: **Hierarchical**, Diana, Agnes, BIRCH, ROCK, CAMELEON

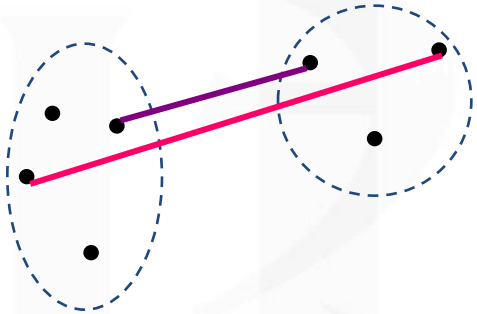
Density-based approach: Based on connectivity and density functions

Typical methods: DBSACN, OPTICS, DenClue

Меры близости

Евклидово расстояние - наиболее общий тип расстояния. Является геометрическим расстоянием между точками в многомерном пространстве:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]^{1/2}$$



где: $\mathbf{X_i}$, $\mathbf{X_j}$ - координаты **i**-го и **j**-го объектов в k-мерном пространстве;

$\mathbf{x_{il}}$ - $\mathbf{x_{jl}}$ - величина **l**-той компоненты у **i**-го (**j**-го) объекта (**l=1,2,...,k; i,j=1,2,...,n**).

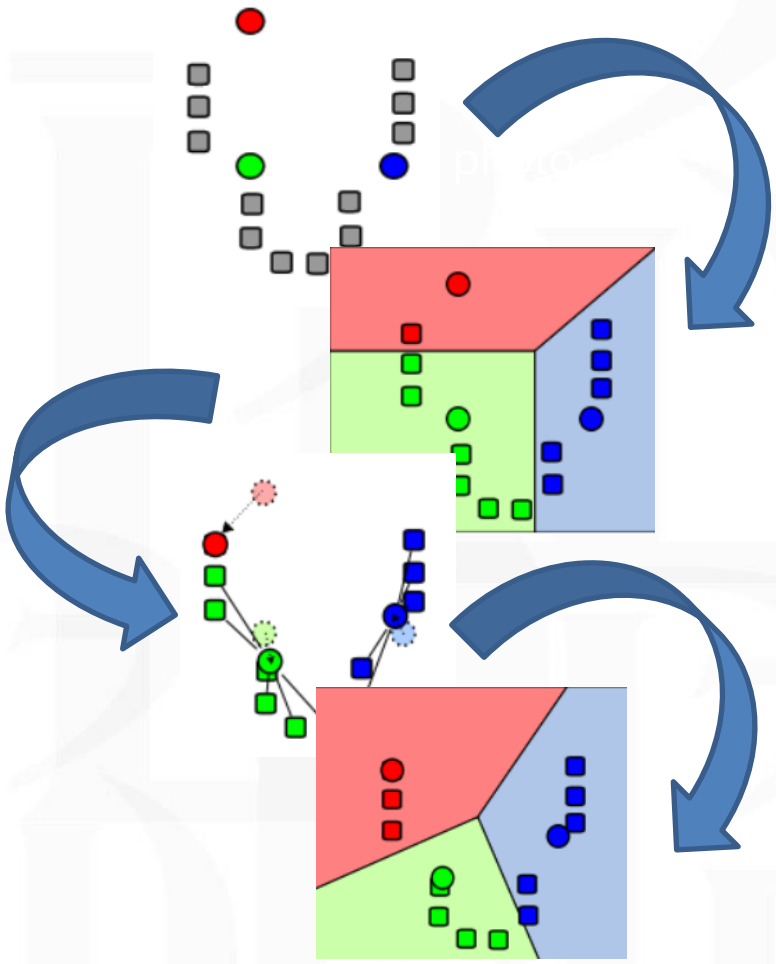
Квадрат евклидова расстояния - используется, чтобы придать большие веса более отдаленным друг от друга объектам:

$$\rho_{ij} = \left[\sum_k (x_{ik} - x_{jk})^2 \right]$$

Алгоритм K means

Основная суть кластеризации заключается в следующем:
Пусть у нас есть совокупность объектов.

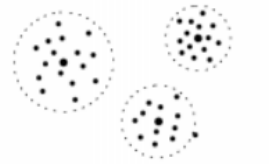
1. Выбираем начальные точки для кластеров.
2. Привязать ближайшие точки к центрам кластеров.
3. Пересчитать центры кластеров, исходя из того, что в кластер были добавлены новые объекты.
4. После того как нашли новые центры кластеризации, снова перераспределяем ближайшие точки по кластерам.



Виды кластеров

внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры



кластеры с центром

Разные виды кластеров ведут к проблеме выбора оптимального числа кластеров.

кластеры могут соединяться перемычками

кластеры могут накладываться на разрежен-ный фон из редко расположенных объектов

кластеры могут перекрываться

Реализация K means в R

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)

x - numeric matrix of data, (data frame with all numeric columns).

centers - number of clusters
iter.max - the maximum number of iterations allowed.

nstart - if centers is a number, how many random sets should be chosen?

Algorithm – name of algorithm.
Note that "Lloyd" and "Forgy" are alternative names for one algorithm.

method – type of method, for example "centers" causes fitted to return cluster centers.

trace - logical or integer number, currently only used in the default method ("Hartigan-Wong"): if positive (or true), tracing information on the progress of the algorithm is produced. Higher values may produce more tracing information.

Загружаем в R studio файл с данными ‘city_csv.csv’

city_csv

Filter

	city	density	density_per_km	average_salary	average.crime	average_edu
1	Moscow	11514	10588	38410	16	68
2	St. Petersburg	4848	3480	27189	13	72
3	Novosibirsk	1473	2947	23374	29	83
4	Yekaterinburg	1350	2489	23216	33	147
5	N. Novgorod	1250	3153	21821	35	84
6	Samara	1164	2152	20690	27	82
7	Omsk	1154	1923	19317	17	86
8	Kazan	1143	1865	19410	20	88
9	Chelyabinsk	1130	2258	20510	26	87
10	Rostov on Don	1064	3127	21053	19	78
11	Ufa	1089	1518	22089	21	93
12	Volgograd	1021	1791	18294	17	81
13	Perm	991	1239	22678	29	93
14	Krasnoyarsk	973	2754	25159	29	86
15	Voronezh	890	1633	18178	14	87
16	Saratov	837	2192	18107	18	155
17	Krasnodar	744	991	22587	18	103

Пример кластеризации в R

myCluster = kmeans(city_csv[, 2:6], 5, nstart = 100)

city_csv – наименование таблицы

2 - 6 – номера колонок в таблице, которые будут использованы в кластерном анализе

5 – число кластеров

nstart = 100 – число итераций

> **myCluster$cluster**

1 3 5 5 4 4 2 2 4 4 5 2 5 5 2 2 5

> myCluster$centers

	density	density_per_km	average_salary	average.crime	average_edu
1	11514.000	10588.000	38410.00	16.00	68.0000
2	1009.000	1880.800	18661.20	17.20	99.4000
3	4848.000	3480.000	27189.00	13.00	72.0000
4	1152.000	2672.500	21018.50	26.75	82.7500
5	1103.333	1989.667	23183.83	26.50	100.8333

Выбор числа кластеров – проблема остановки расчета

Если алгоритм кластеризации вычисляет центры кластеров μy, y ∈ Y , то можно определить функционалы, вычислительно более эффективные.
Сумма средних внутрикластерных расстояний должна быть как можно меньше:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\, y_i = y} \rho^2(x_i, \mu_y) \to \min,$$

где Ky = {xi ∈ Xℓ | yi = y} — кластер с номером y. В этой формуле можно было бы взять не квадраты расстояний, а сами расстояния. Однако, если ρ — евклидова метрика, то внутренняя сумма в Φ_0 приобретает физический смысл момента инерции кластера Ky относительно его центра масс, если рассматривать кластер как материальное тело, состоящее из |Ky| точек одинаковой массы.

Сумма межкластерных расстояний должна быть как можно больше:

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max,$$

Поведение кластерных решений при изменении числа кластеров

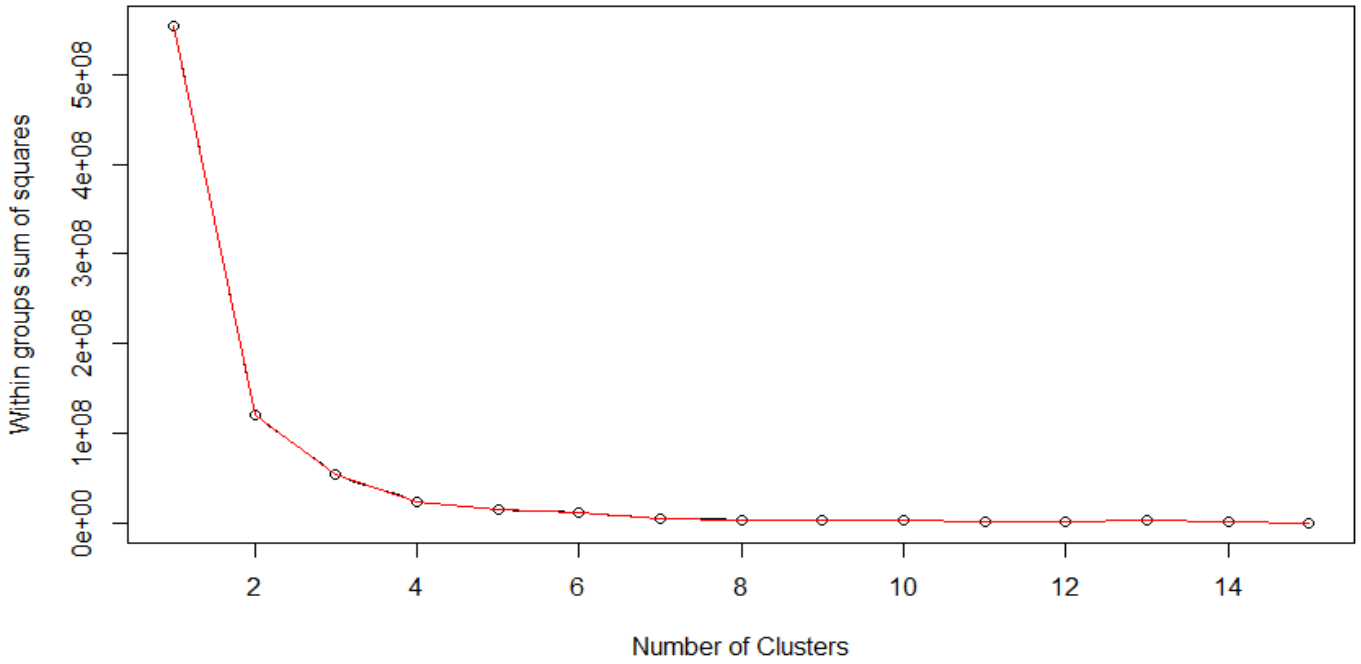
Рассмотрим как меняется величина внутри - кластерного расстояния (withinss) при изменении числа кластеров.

for (i in 1:15){ c[i] =sum(kmeans(x, centers=i)$withinss)}

Теперь выведем на график содержимое вектора c.

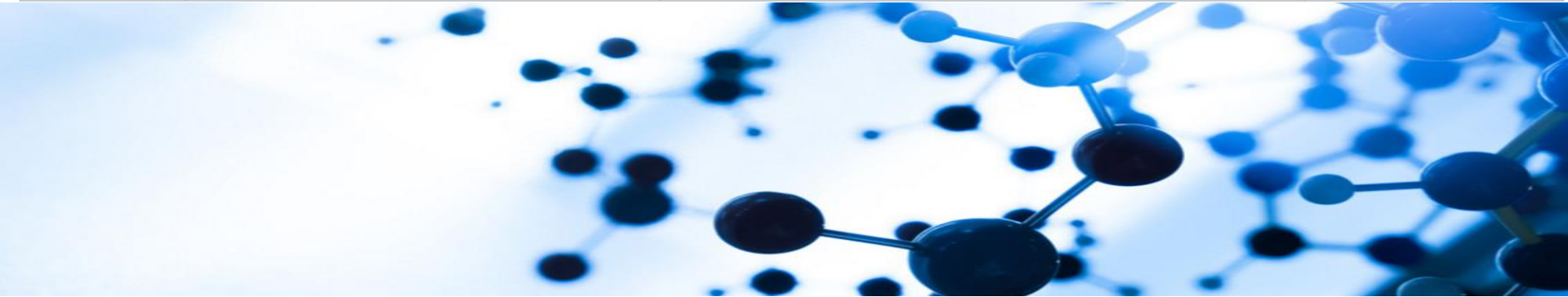
> **plot(c, type="b", xlab="Number of Clusters", + ylab="Within groups sum of squares")**

> **lines(c, col=2)**



Network analysis

Название	Количество зарегистрированных учётных записей	Количество активных учётных записей	Дата статистики	Дата запуска	Страна
Facebook	1.4 млрд.[1]	1 млрд.[2]	январь 2014	февраль 2004	
Google+	500+ млн.[3]	235 млн.[3]	декабрь 2012	июнь 2011	
Tumblr	220+ млн.[4]	100 млн.[5]	май 2013	февраль 2007	
Twitter	500+ млн.[6]	316+ млн.[7]	декабрь 2012	март 2006	
LinkedIn	200+ млн.[8]	160 млн.[8]	январь 2013	май 2003	
Tencent Qzone	623,3 млн.[1]	150 млн.[9][10]	январь 2014	2005	
Sina Weibo	500+ млн.[11]	100+ млн.[12][13]	февраль 2013	август 2009	
ВКонтакте	230+ млн.[1][14]	314.7 млн посетителей/месяц (статистика LiveInternet) 80+ млн.[15] активных учетных записей	январь 2014	сентябрь 2006	
Одноклассники	205+ млн.	148 млн.[16]	апрель 2013	март 2006	
Renren	160+ млн.[17]	45+ млн.[18]	август 2012	декабрь 2005	



Центральная предельная теорема

Цель выделить из больших данных связь между предикторами и зависимой переменной. Предполагается, что подобная связь позволит построить социологический, экономический, политологический анализ явления.

Регрессионные модели основаны на идеи, что предикторы должны быть нормально распределены.

Центральная предельная теорема

Если случайная величина X_i (i=1,2,…..,n) имеет конечные математическое ожидания M(X)=a и дисперсию σ, то распределение средней арифметической $x= (x_1+x_2+……+x_n)/n$, вычисленной по наблюдавшимся значениям случайной величины в n независимых испытаниях, при n->∞ приближается к нормальному закону с математическим ожиданием и дисперсией.

$$\frac{\sum_{j=1}^{n} X_j - na}{\sqrt{n}\,\sigma} \overset{d}{\sim} G,$$

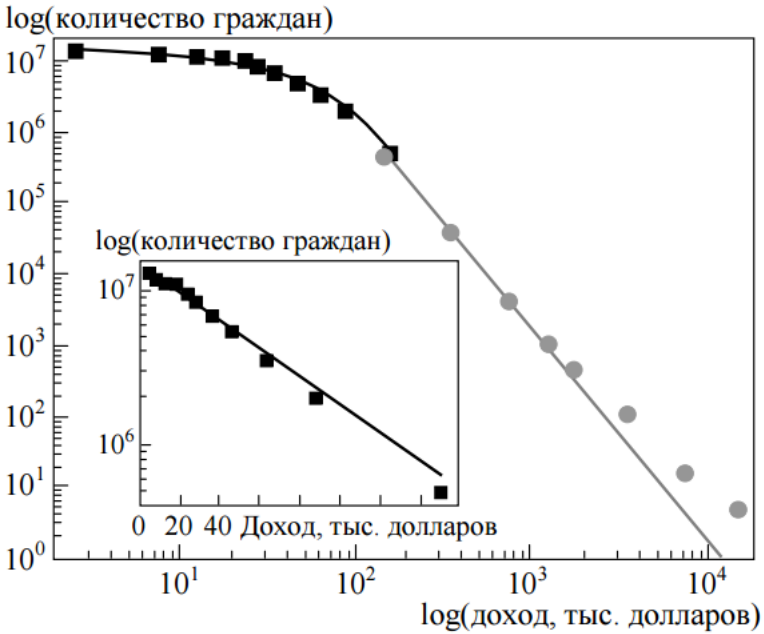
$$p_G(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

Отклонения от нормального закона в экономике

Индивидуальный доход граждан x распределен с плотностью вероятности f(x):

$$f(x) \sim Cx^{-\alpha}.$$

Pareto V. Cours d'Economie Politique. Lausanne and Paris, 1897.



Зависимость количества граждан от размера их индивидуального дохода в тыс. долларов в США в 2004 г

Галкин, С. А. **Экспоненциальные распределения индивидуальных доходов и расходов граждан: наблюдения и модели** / Труды института общей физики им. А. М. Прохорова, РАН. – 2009. – Т. 65. – С. 29–49.

Отклонения от нормального закона в экономике

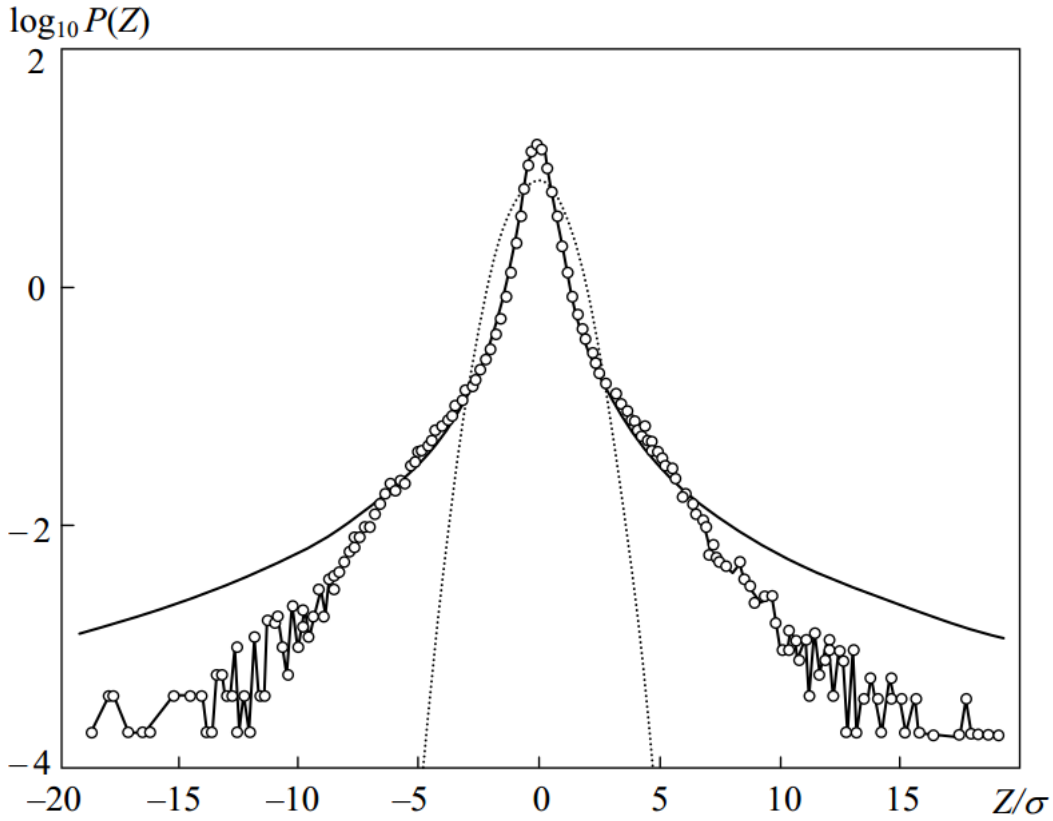
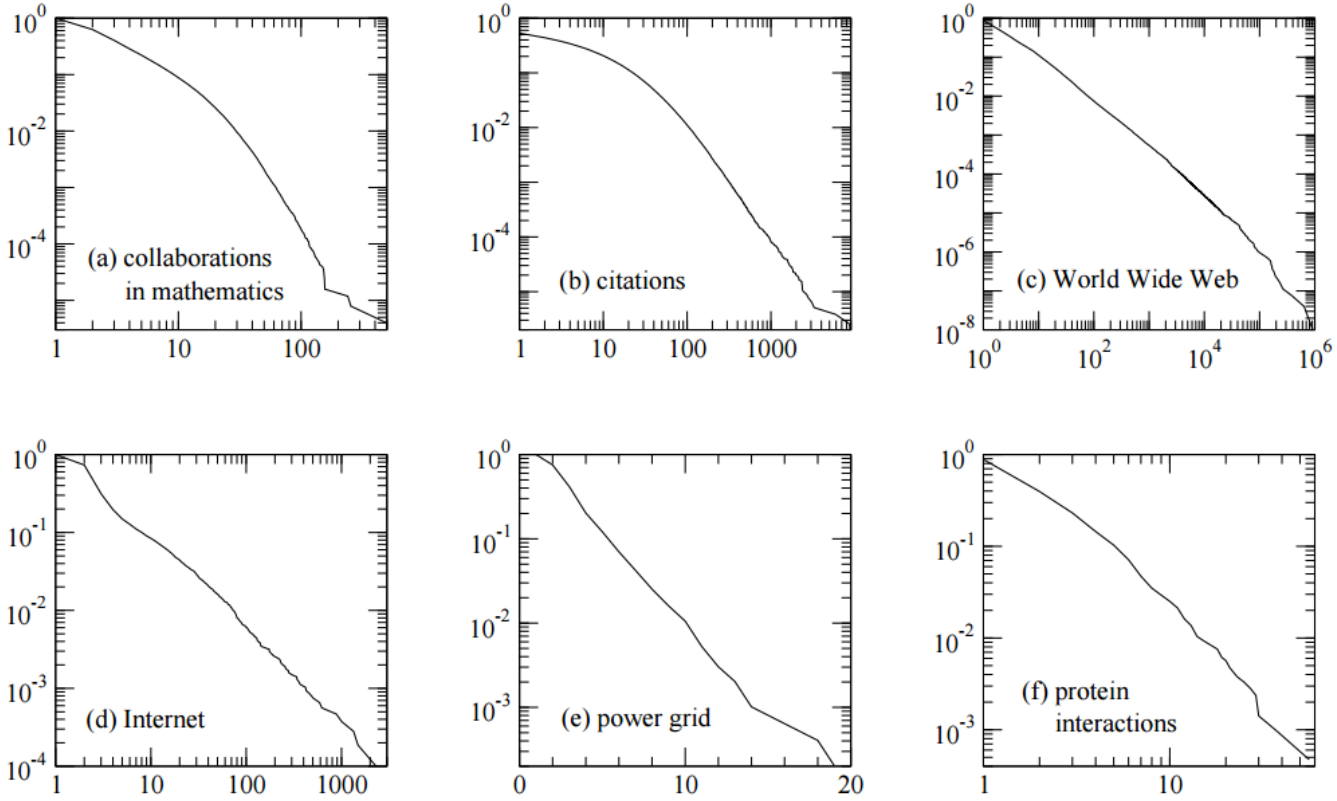


Рис. 1. Распределение флуктуаций $P(Z)$ индекса S&P500 в полулогарифмическом масштабе, измеренных на интервале времени $\Delta t = 1$ мин (○). Для сравнения приведено гауссово распределение (пунктирная кривая), а также центральное распределение Леви с $\alpha = 1.4$ (сплошная кривая) (данные [4])

Распределения в сети

Cumulative degree distributions for six different networks

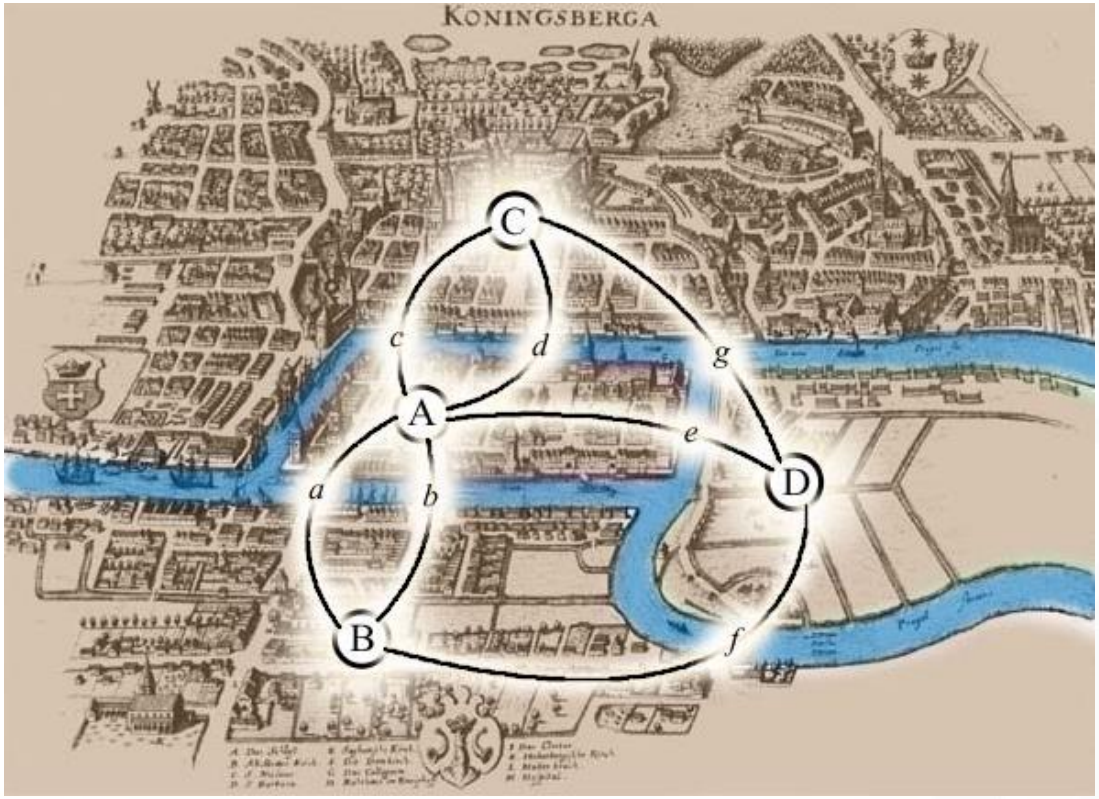


$$P_k \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)}.$$

Newman M.E.J. The structure and function of complex networks // SIAM Review. — 2003. — Vol. 45, N 2. — P. 167—256

Исследование сетей (Network analysis)

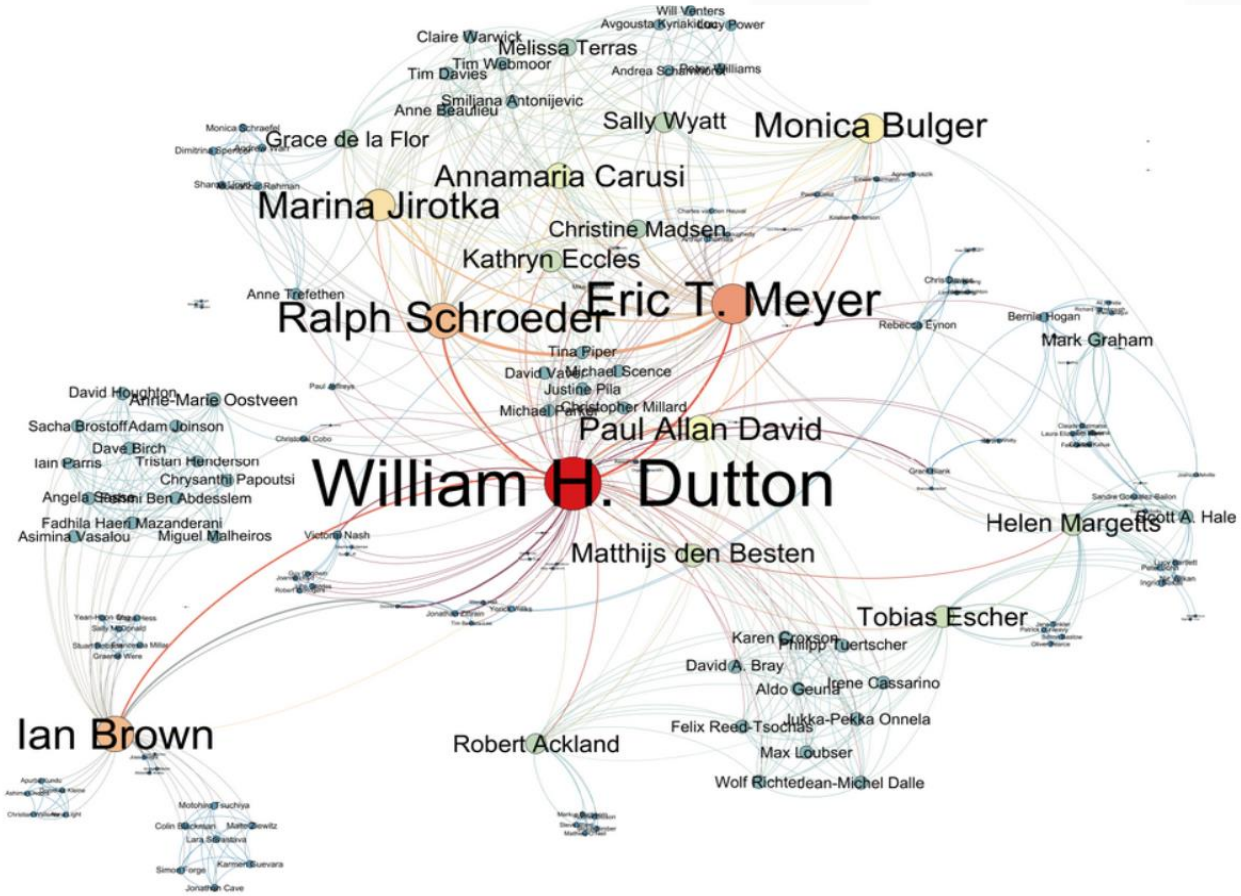
Теория графов — раздел дискретной математики, изучающий свойства графов. В общем смысле граф представляется как множество вершин (узлов), соединённых рёбрами.



Родоначальником теории графов считается Леонард Эйлер. В 1736 году в одном из своих писем он формулирует и предлагает решение задачи о семи кёнигсбергских мостах.

Исследование сетей (Network analysis)

Сеть – это набор узлов (таких как люди, организации, веб-страницы или государственные образования. Каждое отношение соединяет несколько узлов. Узлы соединяются между собой дугами (направленными или не направленными)



INTERNET STUDIES LAB

Исследование сетей (Network analysis)

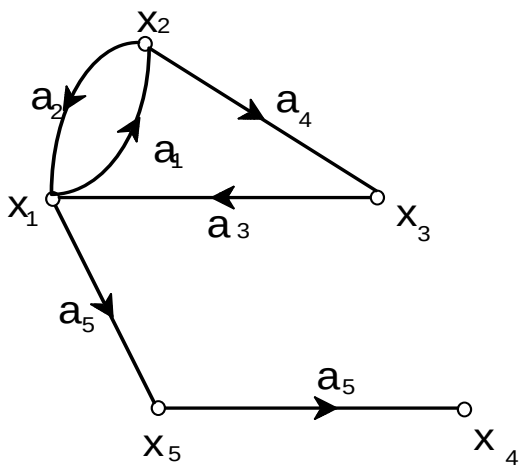
Теория шести рукопожатий — теория, согласно которой любые два человека на Земле разделены в среднем лишь пятью уровнями общих знакомых. Теория была выдвинута в 1969 году американскими психологами Стэнли Милгрэмом и Джеффри Трэверсом. Предложенная ими гипотеза заключалась в том, что каждый человек опосредованно знаком с любым другим жителем планеты через цепочку общих знакомых, в среднем состоящую из пяти человек.

Милгрэм опирался на данные эксперимента в двух американских городах. Жителям одного города было роздано 300 конвертов, которые надо было передать определённому человеку, живущему в другом городе. Конверты можно было передавать только через своих знакомых и родственников. До бостонского адресата дошло 60 конвертов. Произведя подсчеты, Милгрэм определил, что в среднем каждый конверт прошёл через пять человек.

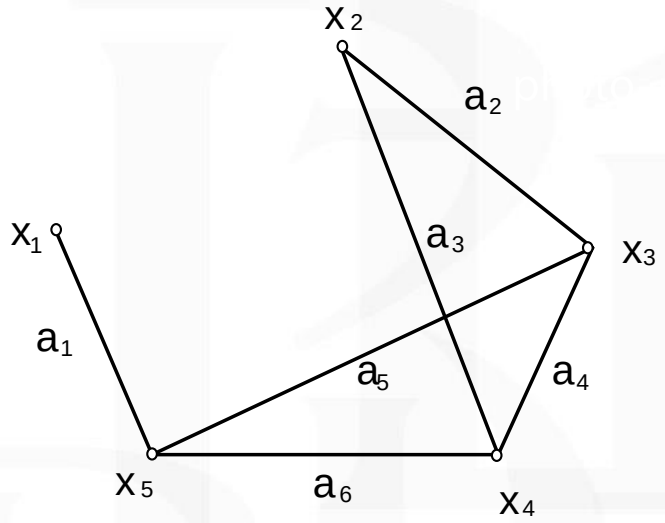
Миланский университет и социальная сеть Facebook установили, что двух любых пользователей Facebook отделяет 4,74 уровня связи. Для США количество звеньев составило 4,37.

Математическая основа теории графов

Ориентированный и не ориентированный виды графов



Ориентированный граф



Неориентированный граф

Примером неориентированного графа является карта дорог

Исследование социальных и экономических сетей

Матрица связей:
Матрицы могут быть квадратными, если анализируются однородные объекты (например, люди), и прямоугольными для анализа связей разнородных объектов (например, люди и организации). Принципы их построения идентичны в обоих случаях: наличие связи между объектами помечается выбранным символом в ячейке, лежащей на пересечении соответствующих строки и столбца.

Квадратная матрица

	Алексей	Сергей	Максим	Павел
Алексей	0	1	1	0
Сергей	1	0	1	1
Максим	1	1	0	0
Павел	0	1	0	0

Прямоугольная матрица

	Орг-ция 1	Орг-ция 2	Орг-ция 3	Орг-ция 4	Орг-ция 5
Алексей	1	0	1	0	0
Сергей	1	0	0	1	0
Максим	0	0	1	0	1
Павел	0	1	0	0	0

Математическая основа теории графов

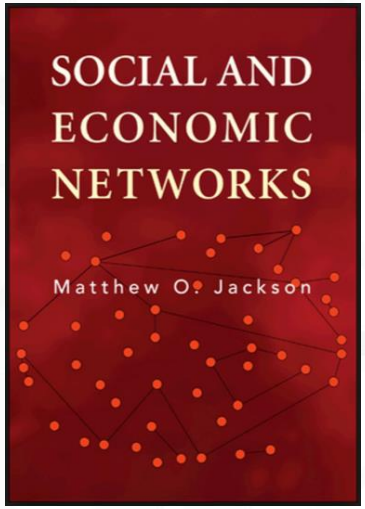
Параметры сети: Число вершин или ребер

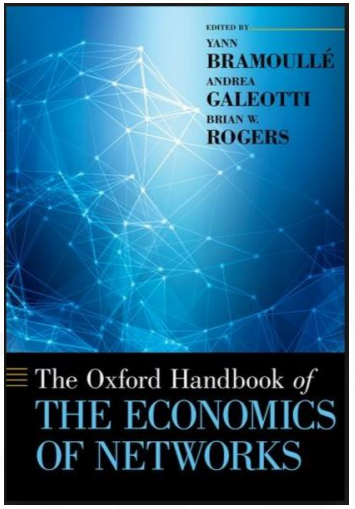
Вычисляемые параметры: ***Плотность*** - вычисляется как нормированное число ребер (отношение наличных связей в сети к возможному максимальному количеству связей в сети с данным количеством вершин)

Среднее расстояние от одной вершины до других - рассчитывается на основе минимальных расстояний от данной вершины до всех остальных.

Диаметр социальной сети - параметр, который показывает, насколько велика сеть – это наибольшее геодезическое расстояние в социальной сети

Центральность
(центральным агентом сети является тот, у кого больше всего связей)



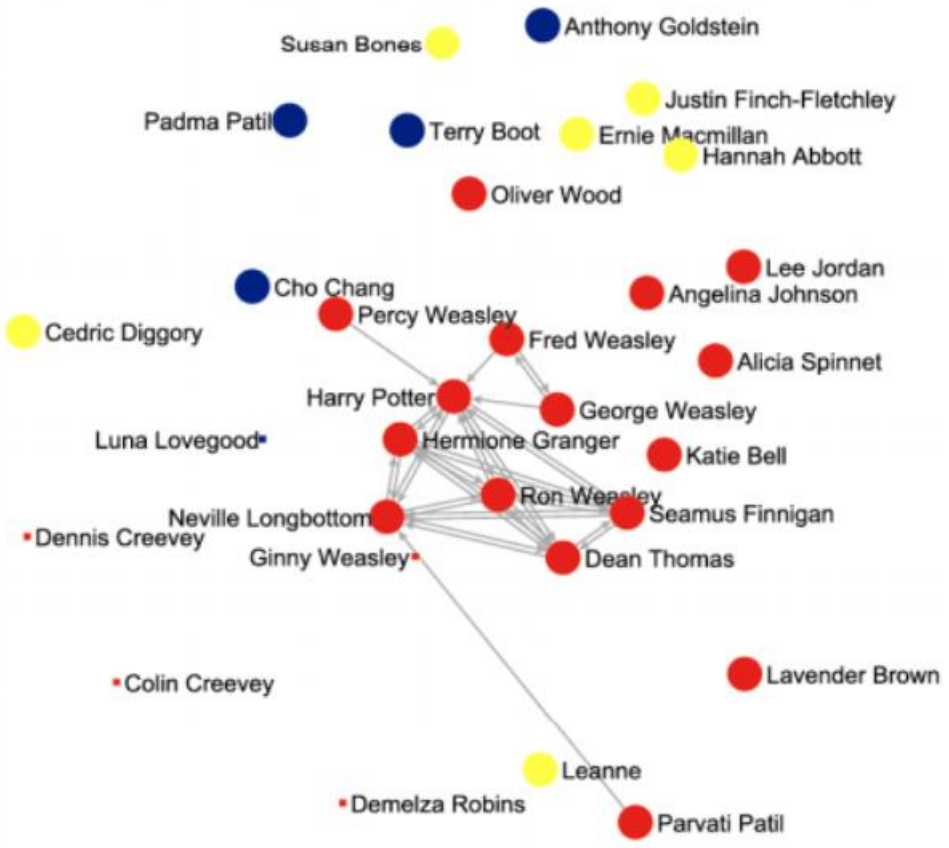


An Analysis of Friendship Networks in the Harry Potter

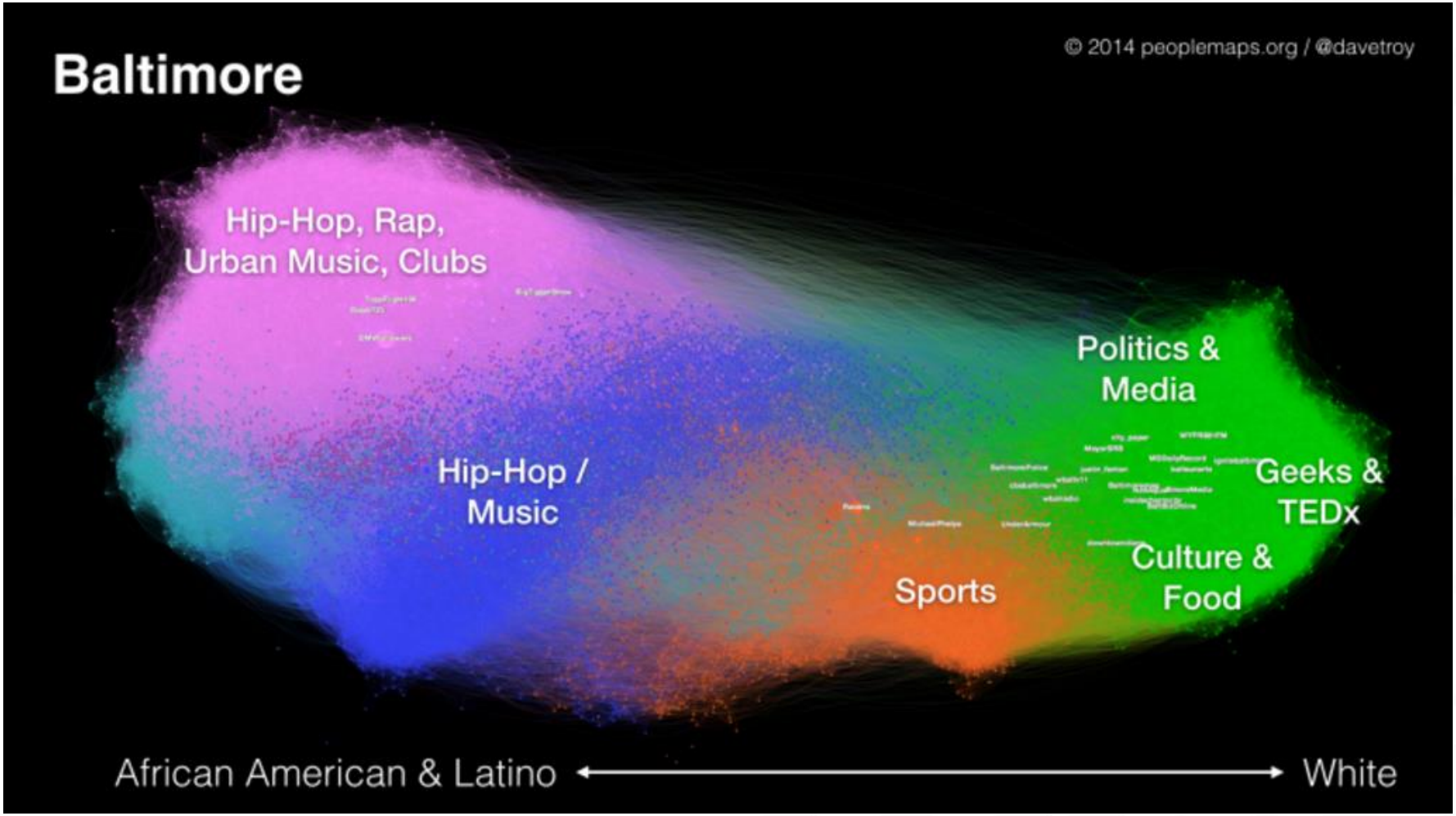
(1) **Student A supports student B emotionally**: Harry, Ron and Hermione assure Neville that he is definitely a Gryffindor when he doubts he is not brave enough to be in house.

(2) **Student A gives students B instrumental help**: Fred and George Weasley help Harry Potter to get his trunk into Hogwarts Express.

(3) **Student A gives student B certain information to help student**: Hermione Granger helps Harry Potter with his homework.

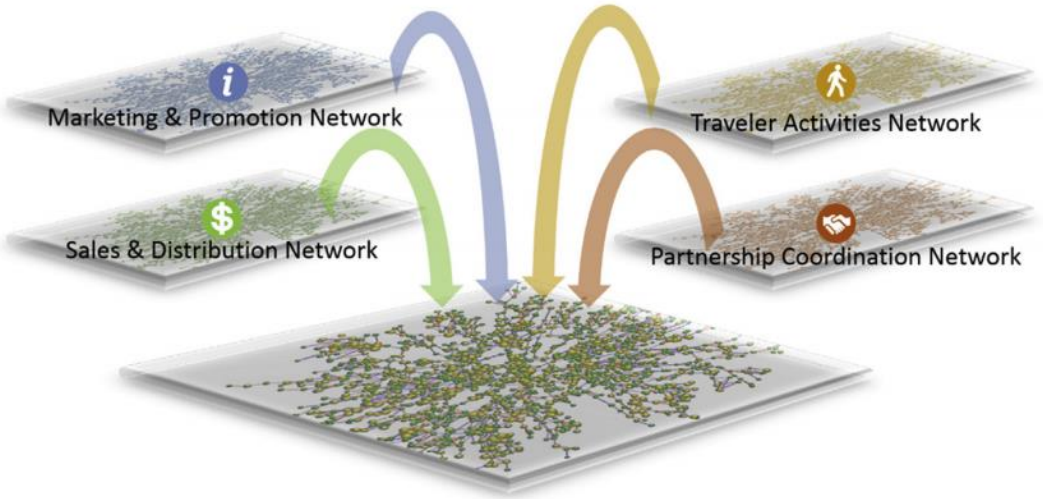


Baltimore: network analysis

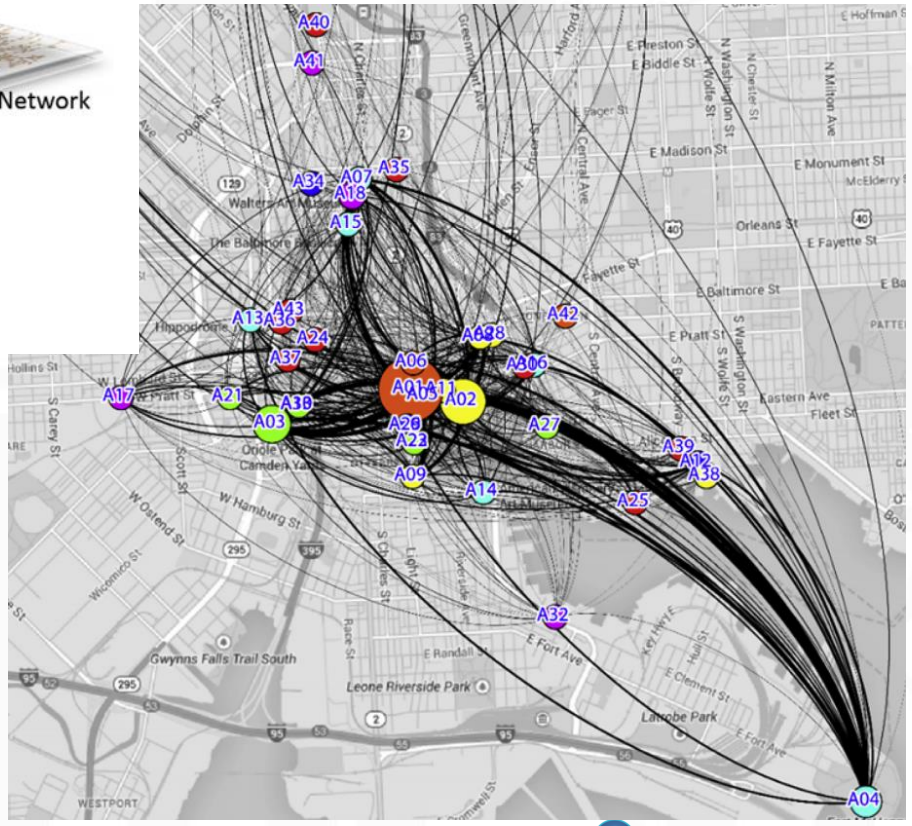


Estimating value in Baltimore, Maryland: An attractions network analysis

J.L. Stienmetz, D.R. Fesenmaier / Tourism Management 50 (2015) 238–252

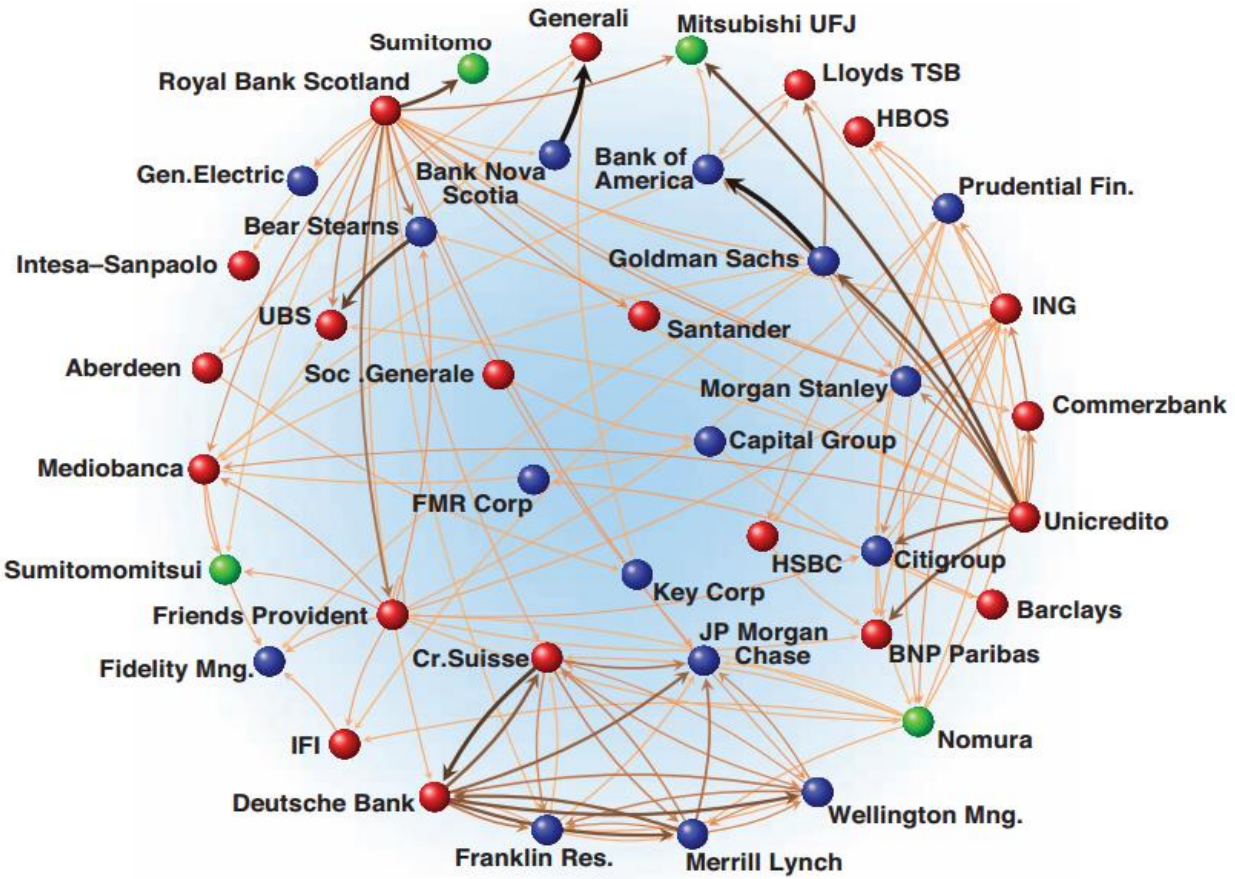


Jason L. Stienmetz* , Daniel R. Fesenmaier



INTERNET STUDIES LAB

A sample of the international financial network



where the nodes represent major financial institutions and the links are both directed and weighted and represent the strongest existing relations among them. Node colors express different geographical areas: European Union members (red), North America (blue), other countries (green).

Задачи линейного программирования

Задачи оптимального планирования, связанные с отысканием оптимума заданной целевой функции (линейной формы) при наличии ограничений в виде линейных уравнений или линейных неравенств относятся к задачам линейного программирования.

Линейное программирование - наиболее разработанный и широко применяемый раздел математического программирования. **Потому что**:

1. Математические модели очень большого числа экономических задач линейны относительно искомых переменных;
2. Для линейных задач разработаны специальные численные методы, с помощью которых эти задачи решаются, и соответствующие стандартные программы для их решения на ЭВМ;
3. Некоторые задачи, которые в первоначальной формулировке не являются линейными, после ряда дополнительных ограничений и допущений могут стать линейными или могут быть приведены к такой форме, что их можно решать методами линейного программирования.

Таким образом, ***Линейное программирование*** – это направление математического программирования, изучающее методы решения экстремальных задач, которые характеризуются линейной зависимостью между переменными и линейным критерием.

Математическая постановка задачи (ЗЛП).

Математическая задача в общем виде состоит в определении наибольшего или наименьшего значения целевой функции $F(x_1, x_2, x_3\ldots)$ при условии что $a_i(x1, x2, x3\ldots) <= b_i$, F, a_i – заданные функции, b_i - некоторые действительные числа.

Если функции F, a_i – линейные задача является задачей линейного программирования, а если не линейные (хотя бы из одна из функций), то данная задача является не линейной. Функция $F(x_1, x_2, x_3\ldots)$ - называется **целевой функцией** от переменных $x_1, x_2, x_3\ldots$.

$$F = c_1x_1 + c_2x_2 + c_3x_3 + \cdots + c_nx_n$$

Система прямых ограничений

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1j}x_j + \ldots + a_{1n}x_n \le (=, \ge) b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2j}x_j + \ldots + a_{2n}x_n \le (=, \ge) b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n \le (=, \ge) b_i, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mj}x_j + \ldots + a_{mn}x_n \le (=, \ge) b_m, \end{cases}$$

$$x_j \ge 0, \ j = 1, 2, \ldots, n.$$

Вектор X называется планом или допустимым решением ЗЛП.

Пример задачи.

Ресурс	Мужской костюм	Женский костюм	Ограничение ресурса
Труд (чел. - день)	1	1	150
Сырье 1 (метр ткани, шерсть)	3.5	1	350
Сырье 2 (метр ткани, лавсан)	0.5	2	240

Вопрос. Сколько нужно сшить костюмов, так что бы максимизировать прибыль? При том, что прибыль от одного женского костюма составляет 10$, а прибыль от мужского костюма – 20$. При этом нужно сшить не менее 60 мужских костюмов.

Пусть x_1 – число женских костюмов, x_2 – число мужских костюмов. Нам нужно максимизировать функцию:

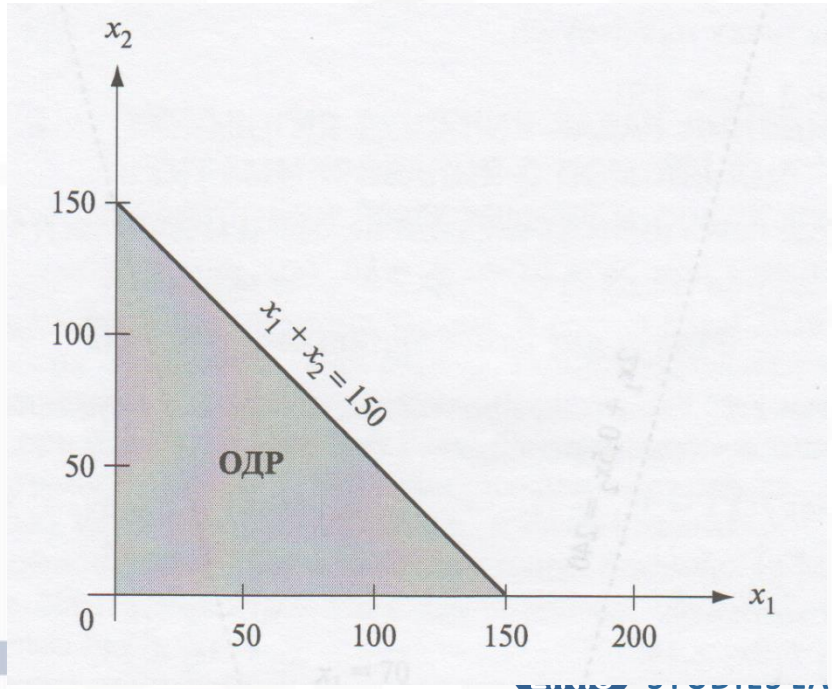
Экономико – математическая модель задачи

Ресурс	Мужской костюм	Женский костюм	Ограничение ресурса
Труд (чел. - день)	1	1	150
Сырье 1 (метр ткани, шерсть)	3.5	1	350
Сырье 2 (метр ткани, лавсан)	0.5	2	240

1. Первое ограничение по труду

$$x_1 + x_2 \le 150.$$

ОДР – область допустимых решений



Ограничение задачи

$$\begin{aligned} x_1 + x_2 &\le 150, \\ 2x_1 + 0{,}5x_2 &\le 240, \\ x_1 + 3{,}5x_2 &\le 350, \\ x_2 &\ge 60, \\ x_1 &\ge 0. \end{aligned}$$

Графическое решение задачи

1. Ограничение по лавсану

$$2x_1 + 0.5x_2 \leq 240$$

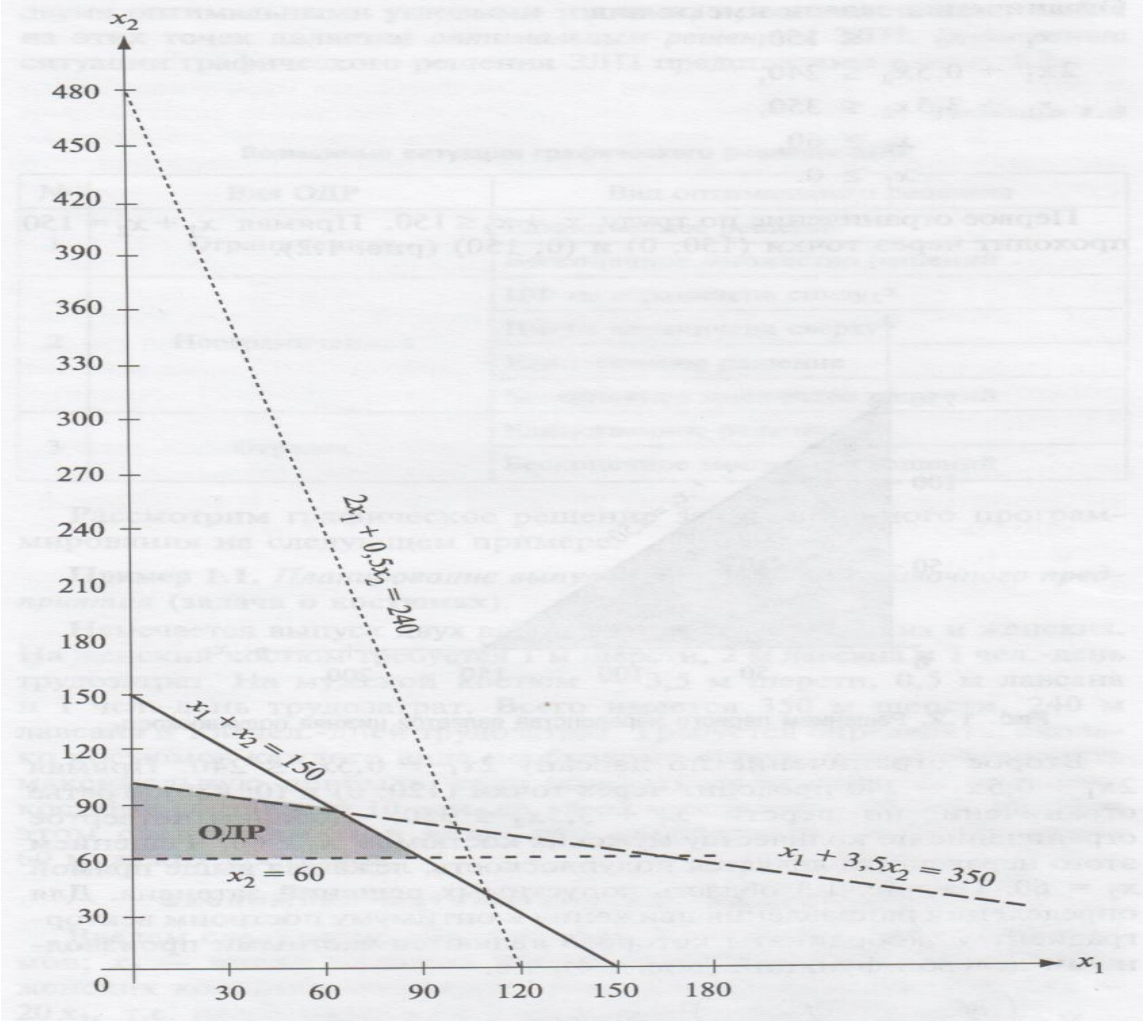
2. Ограничение по шерсти

$$x_1 + 3.5x_2 \leq 350$$

2. Ограничение по количеству мужских костюмов

$$x_1 \geq 60$$

Область допустимых значений при всех ограничениях – затемненная область.



NATIONAL RESEARCH UNIVERSITY

Thank you for your attention!

Room 216, building 2, 55 Sedova St., St.Petersburg, Russia
Laboratory for Internet Studies
www.linis.hse.ru