INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG

Научно Исследовательский Семинар 2016

Математические модели в экономике

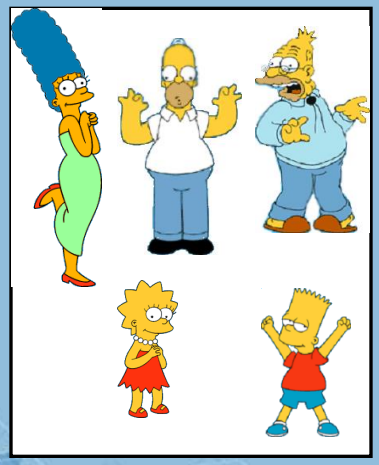
Sergei Koltcov
skoltsov@hse.ru
https://linis.hse.ru

Содержимое

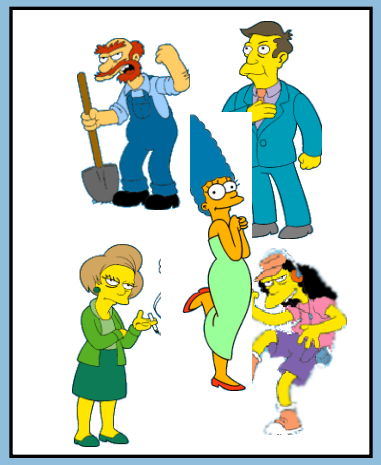
4. Кластерный анализ. Задачи и методы кластерного анализа. Проблемы кластерного анализа. Применение кластерного анализа в Orange.
5. Сетевой анализ. Теория шести рукопожатий. Параметры в сетевом анализе. Кластерный анализ в исследовании сетей. Применение кластерного анализа в NodExl
6. Задачи линейного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

Задачи и методы кластер - анализа

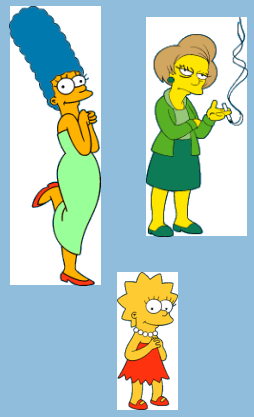
Кластеризация – это разбиение элементов некоторого множества на группы на основе их схожести. Задача кластеризации состоит в разбиении объектов из X на несколько подмножеств (кластеров), в которых объекты более схожи между собой, чем с объектами из других кластеров.



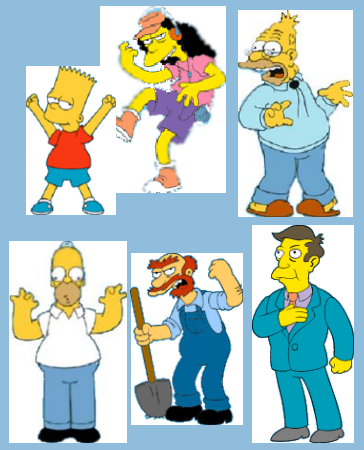
Семья



Сотрудники



Женщины



Мужчины

Лейбелинг групп – то что нужно найти

Кластеризация достаточно субъективна 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

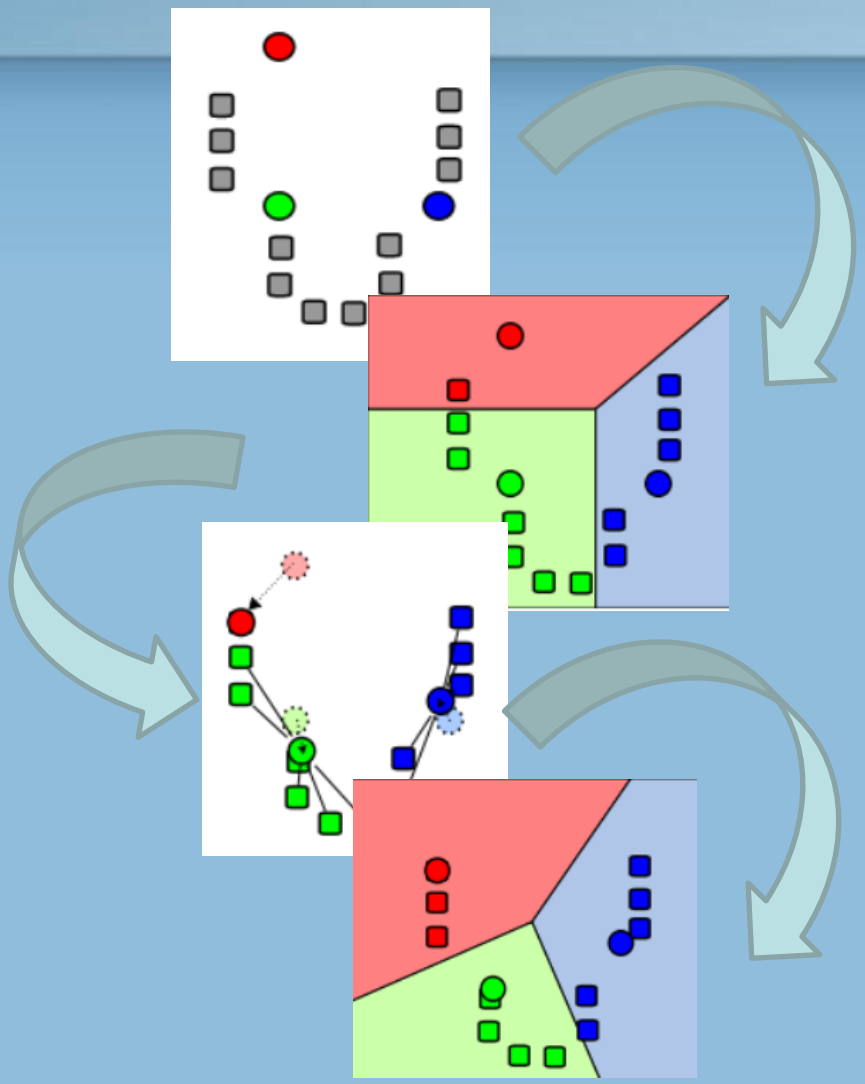
“*We know it when we see it*”

Задача определения сходства является задачей Machine learning.

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.



Процесс повторяется до тех пор пока центры кластеров перестанут изменятся.

Метод K means. Меры близости

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

$$\rho_e(X_i , X_j) = (\sum_l (x_{il} - x_{jl})^2)^{1/2} ,$$

где: X_i , X_j **-** координаты **i**-го и **j**-го объектов в k-мерном пространстве;

$x_{il} - x_{jl}$ **-** 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:

$$\rho_{ке}(X_i , X_j) = \sum_l (x_{il} - x_{jl})^2$$

Выбор числа кластеров – проблема остановки расчета

Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров Yi объектам Xi , чтобы значение выбранного функционала качества приняло наилучшее значение. Существует много разновидностей функционалов качества кластеризации, но нет «самого правильно го» функционала

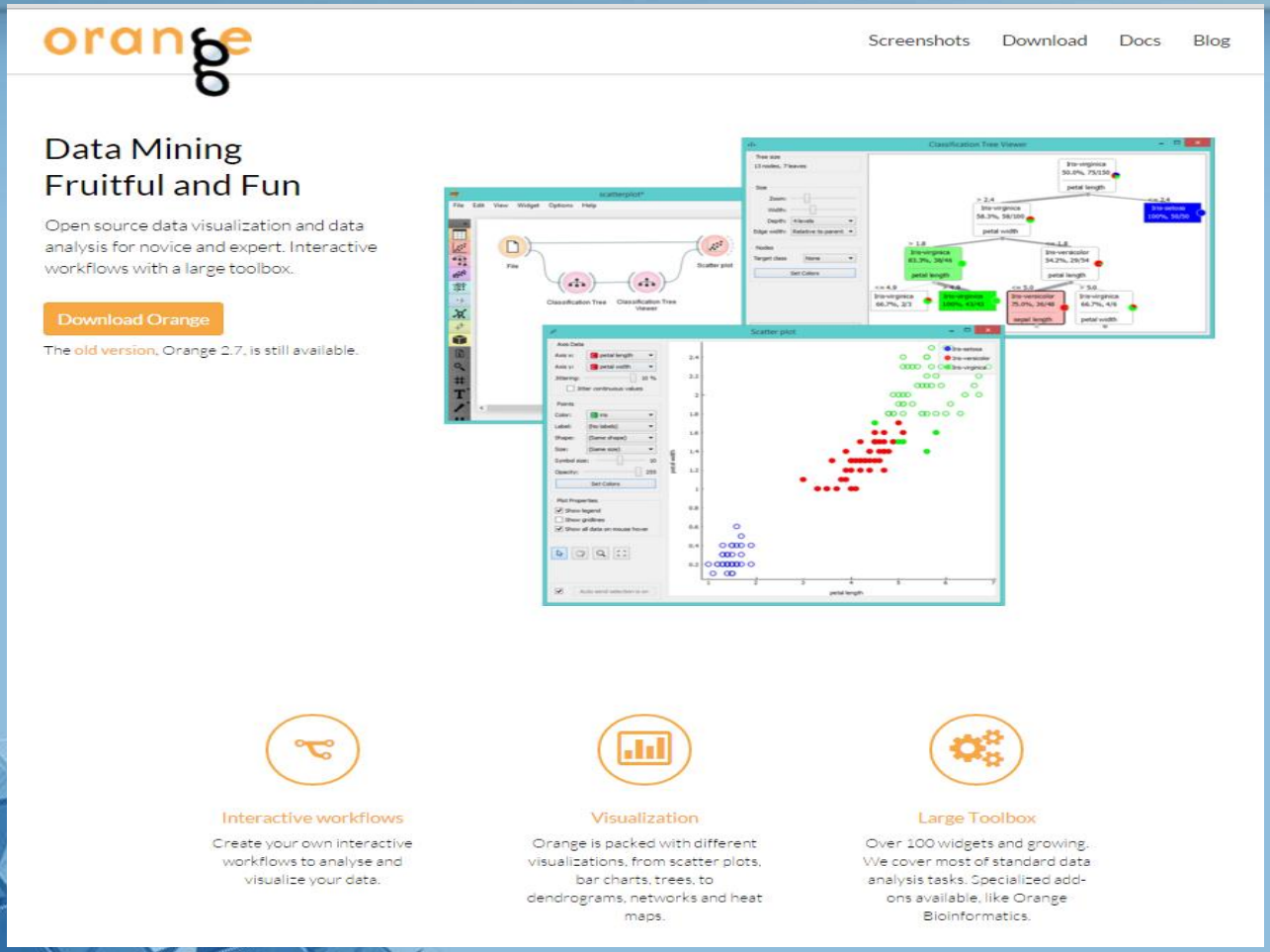
Среднее внутрикластерное расстояние должно быть как можно меньше:

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min .$$

Среднее межкластерное расстояние должно быть как можно больше:

$$F_1 = \frac{\sum_{i<j}[y_i \neq y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i \neq y_j]} \to \max .$$

Orange – Open source software for data visualization and data analysis

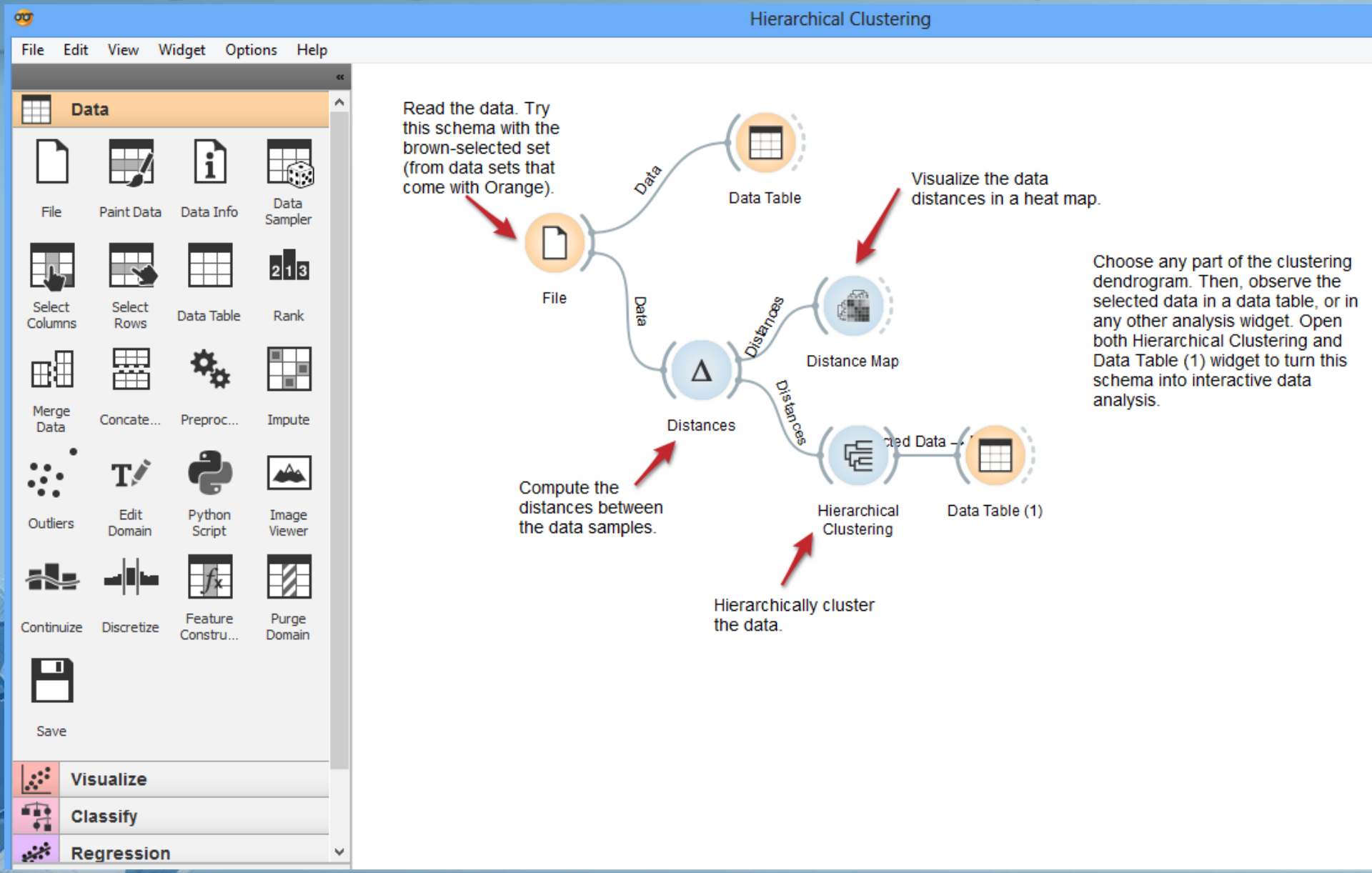


Можно скачать по адресу:
http://orange.biolab.si/

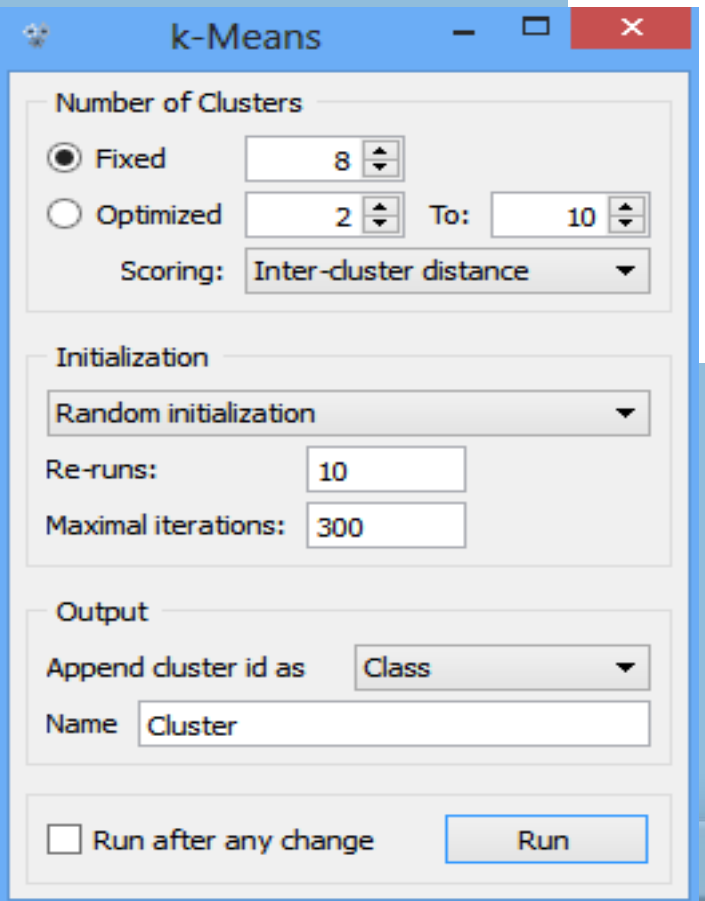
Faculty of Computer and Information Science
University of Ljubljana

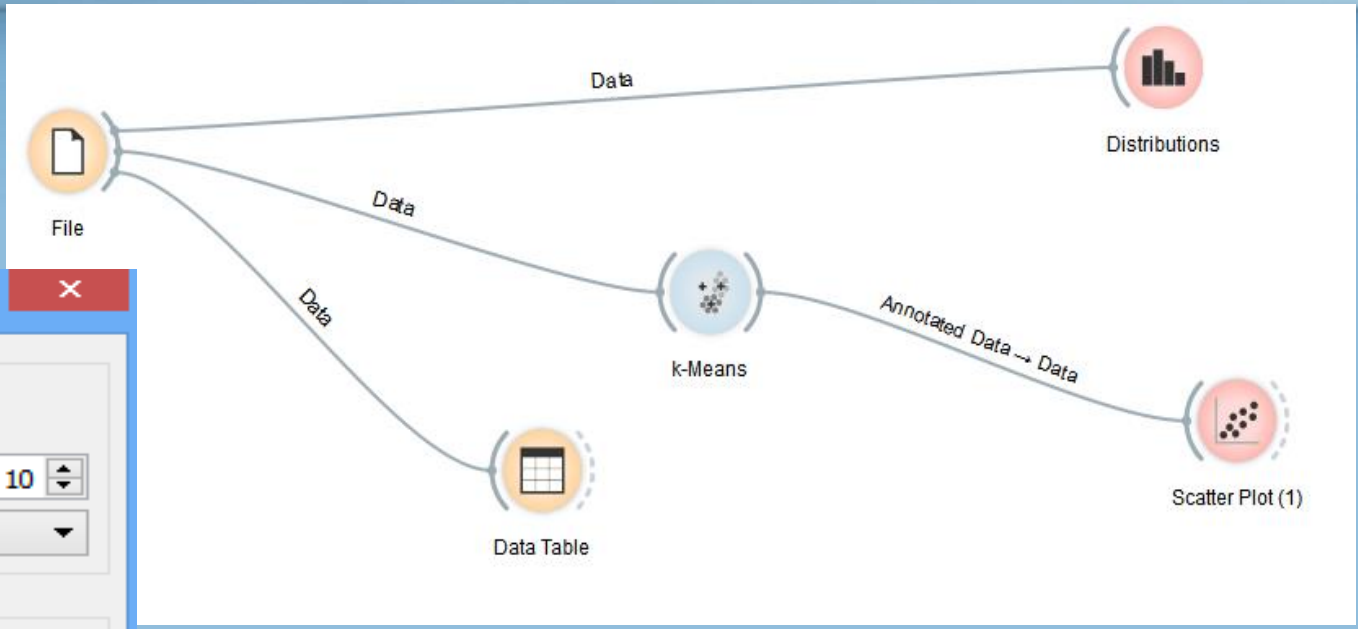
Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14(Aug):2349−2353.

Orange consists of a canvas interface onto which the user places widgets and creates a data analysis workflow.



K means - Orange

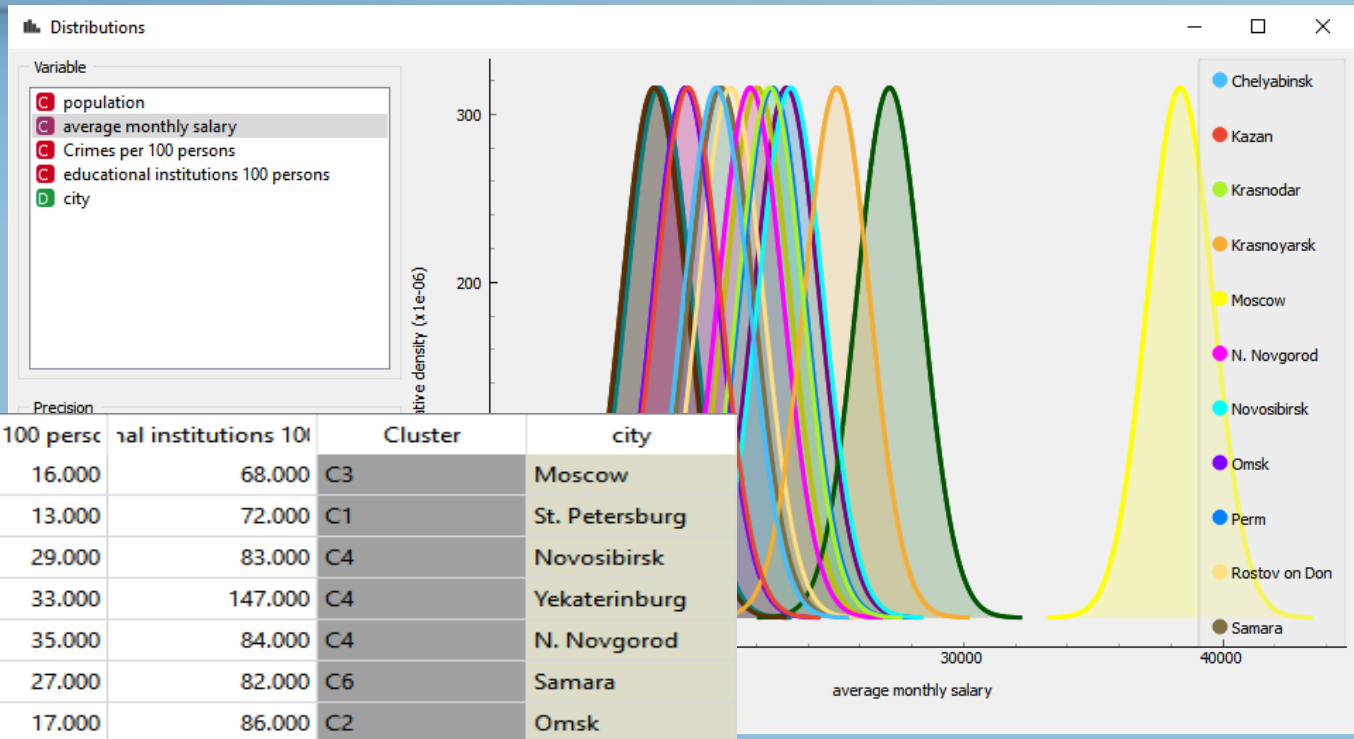




Виджет K-means осуществляет кластеризацию этим методом.
Параметры: Number of clusters.
Scoring

- Silhouette (contrasts average distance to elements in the same cluster with the average distance to elements in other clusters)
- Inter-cluster distance (measures distances between clusters)
- Distance to centroids (measures distances to the arithmetic means of clusters)

K-means in Orange.



	population	erage monthly sal	mes per 100 persc	nal institutions 10	Cluster	city
1	11514.000	38410.000	16.000	68.000	C3	Moscow
2	4848.000	27189.000	13.000	72.000	C1	St. Petersburg
3	1473.000	23374.000	29.000	83.000	C4	Novosibirsk
4	1350.000	23216.000	33.000	147.000	C4	Yekaterinburg
5	1250.000	21821.000	35.000	84.000	C4	N. Novgorod
6	1164.000	20690.000	27.000	82.000	C6	Samara
7	1154.000	19317.000	17.000	86.000	C2	Omsk
8	1143.000	19410.000	20.000	88.000	C2	Kazan
9	1130.000	20510.000	26.000	87.000	C6	Chelyabinsk
10	1064.000	21053.000	19.000	78.000	C6	Rostov on Don
11	1089.000	22089.000	21.000	93.000	C4	Ufa
12	1021.000	18294.000	17.000	81.000	C2	Volgograd
13	991.000	22678.000	29.000	93.000	C4	Perm
14	973.000	25159.000	29.000	86.000	C5	Krasnoyarsk
15	890.000	18178.000	14.000	87.000	C2	Voronezh
16	837.000	18107.000	18.000	155.000	C2	Saratov
17	744.000	22587.000	18.000	103.000	C4	Krasnodar

Проблемы K means.

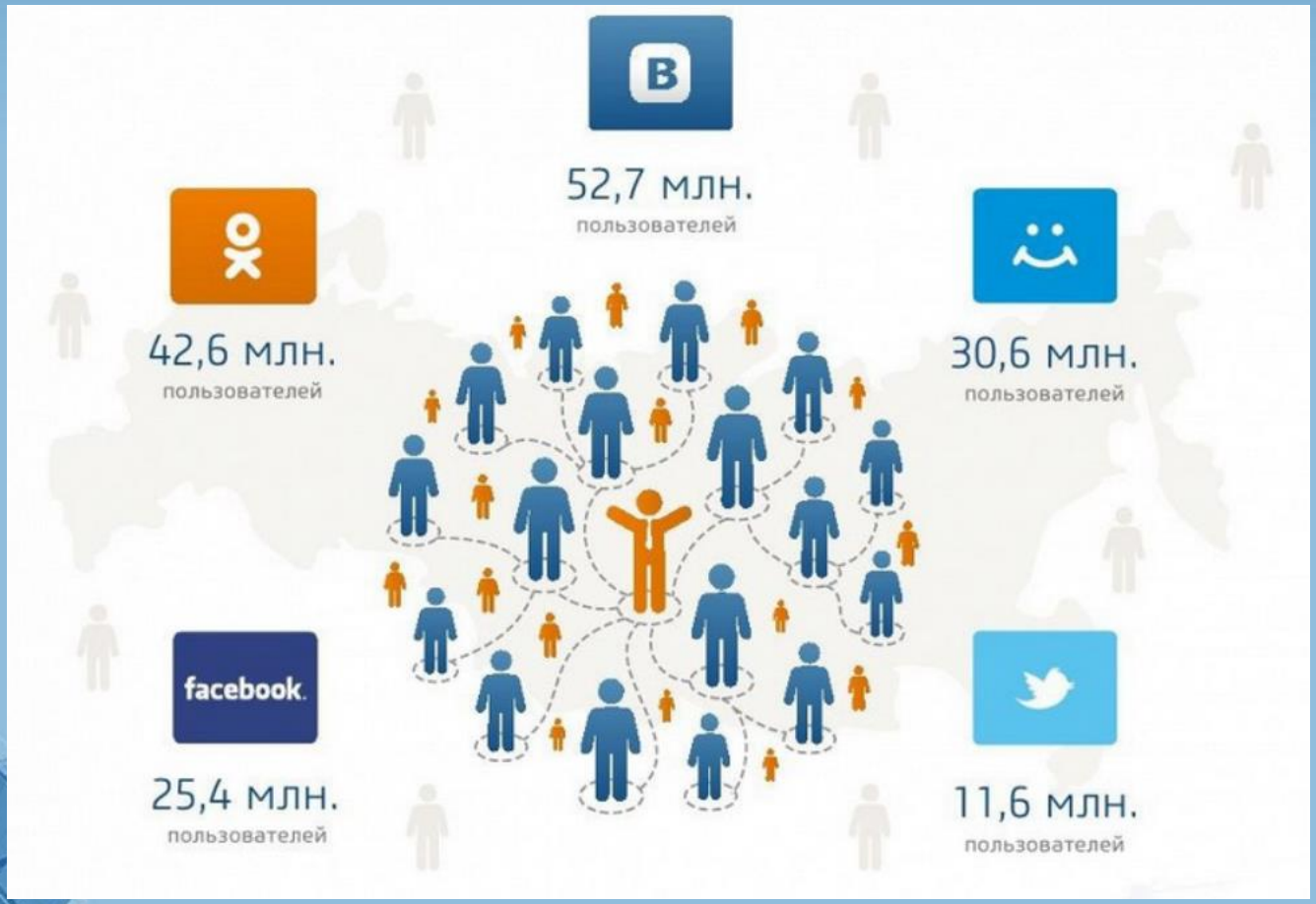
1. Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *V*, а только одного из локальных минимумов.
2. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
3. Число кластеров надо знать заранее.

Как можно преодолеть эти проблемы?

1. Запускать алгоритм много раз (с разными центрами кластеров), после чего выбрать результат с минимальной величиной ошибки.

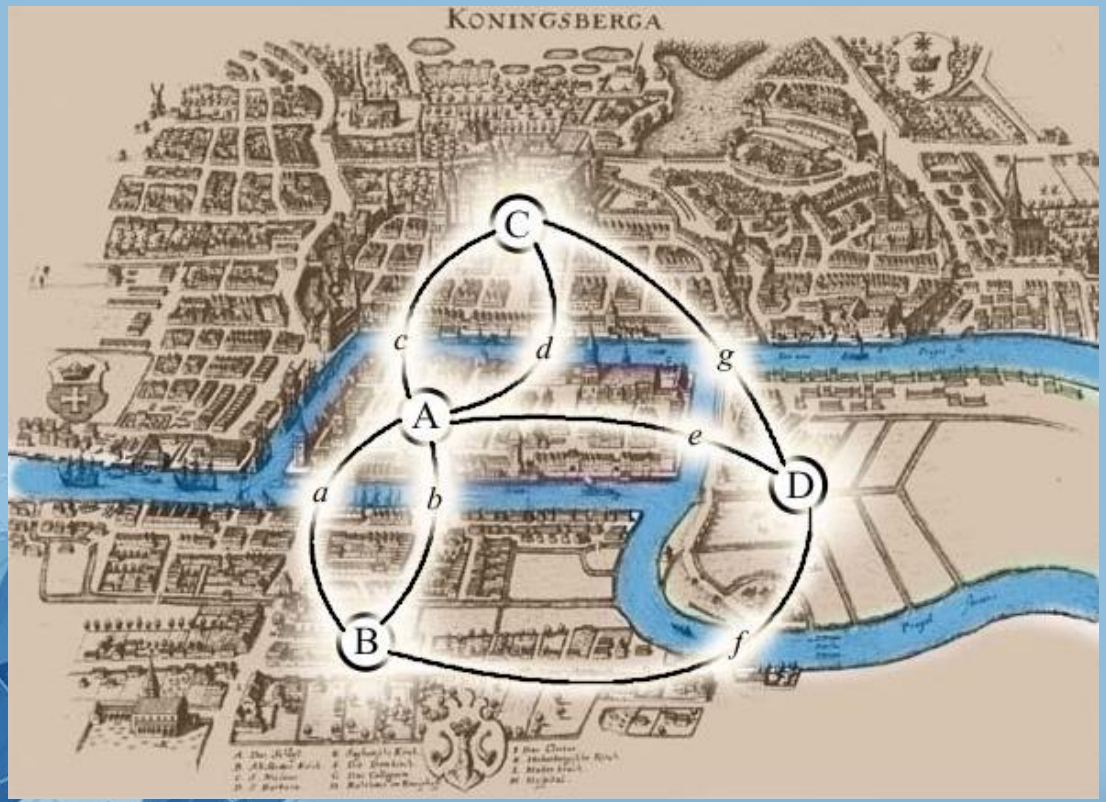
2. Использовать дополнительные модели дл оценки количества кластеров.

Network analysis



Исследование сетей (Network analysis)

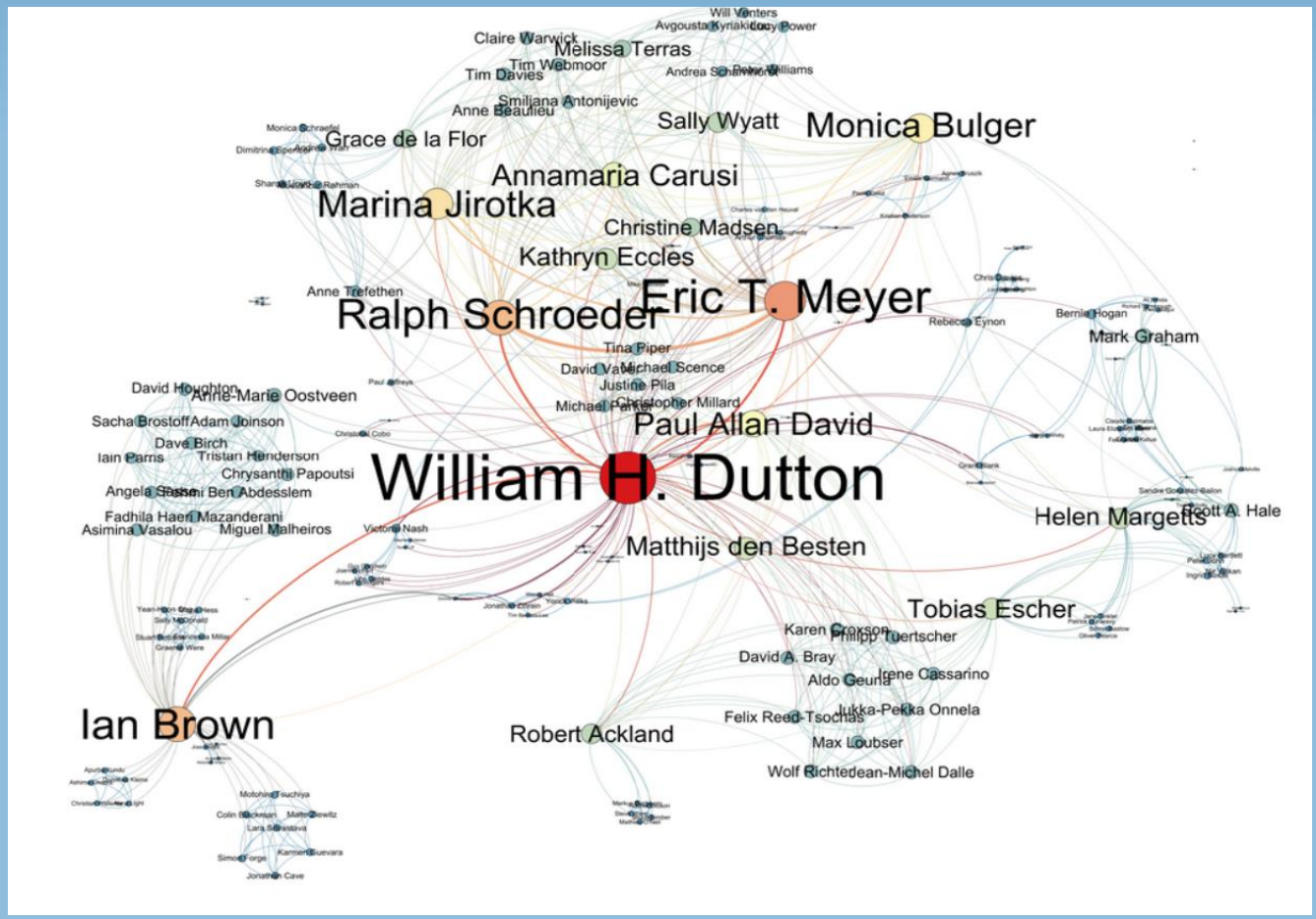
Теория графов — раздел дискретной математики, изучающий свойства графов. В общем смысле граф представляется как множество вершин (узлов), соединённых рёбрами.



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)



Исследование сетей (Network analysis)

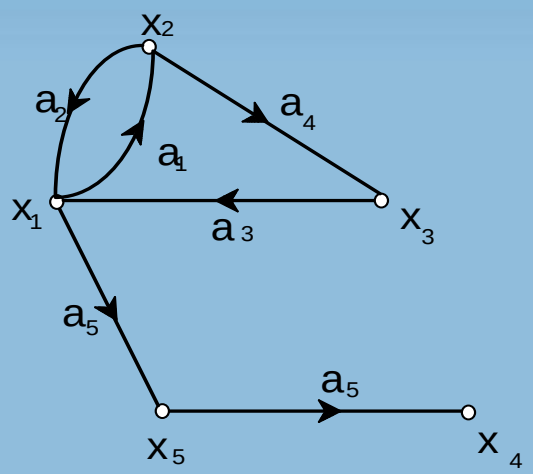
Теория шести рукопожатий — теория, согласно которой любые два человека на Земле разделены в среднем лишь пятью уровнями общих знакомых. Теория была выдвинута в 1969 году американскими психологами Стэнли Милгрэмом и Джеффри Трэверсом. Предложенная ими гипотеза заключалась в том, что каждый человек опосредованно 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.

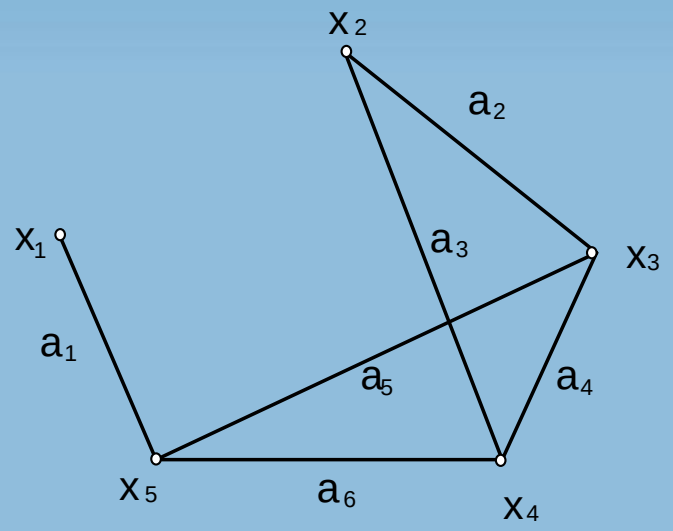
Миланский университет и социальная сеть Facebook установили, что двух любых пользователей Facebook отделяет 4,74 уровня связи. Для США количество звеньев составило 4,37.

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 (Network analysis)

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

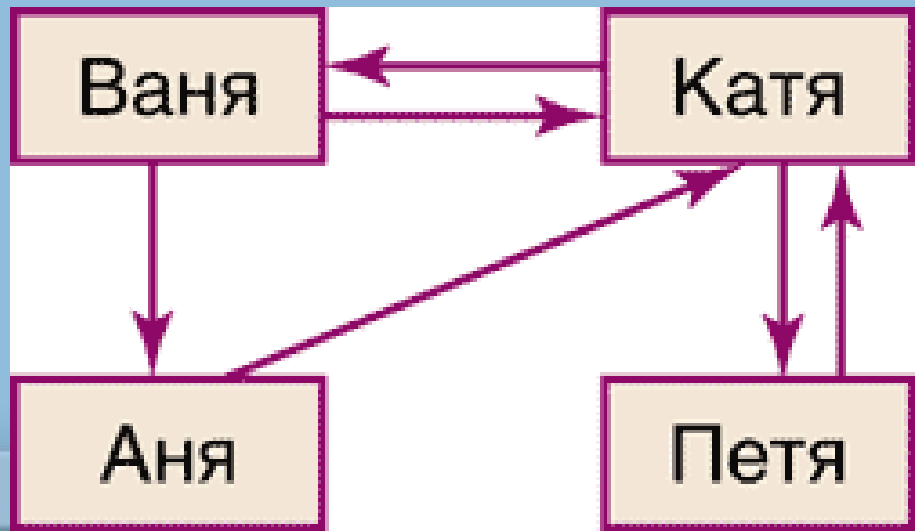
Исследование сетей

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)

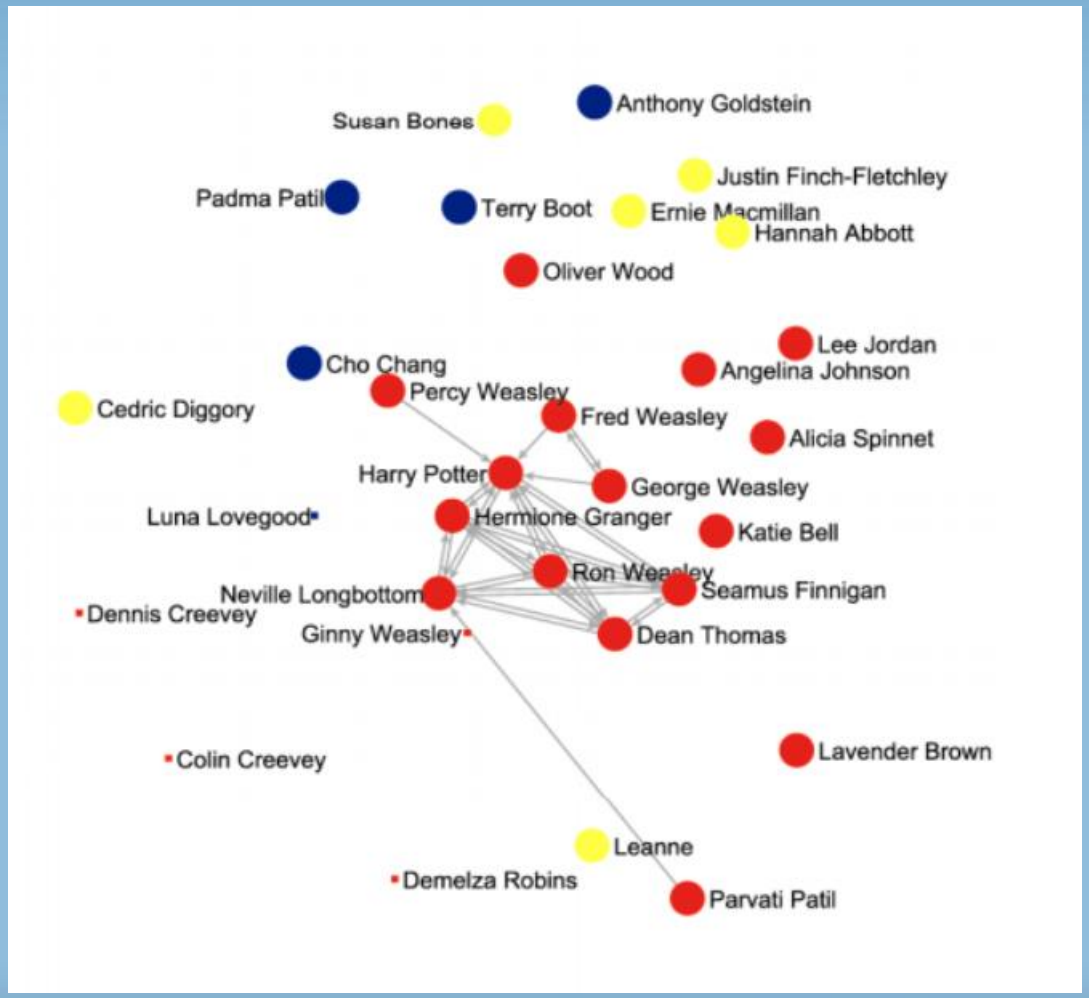


Исследование социальных сетей
An Analysis of Friendship Networks in the Harry Potter

(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 h**elp: 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.



NodExl: http://nodexl.codeplex.com/

CodePlex Project Hosting for Open Source Software

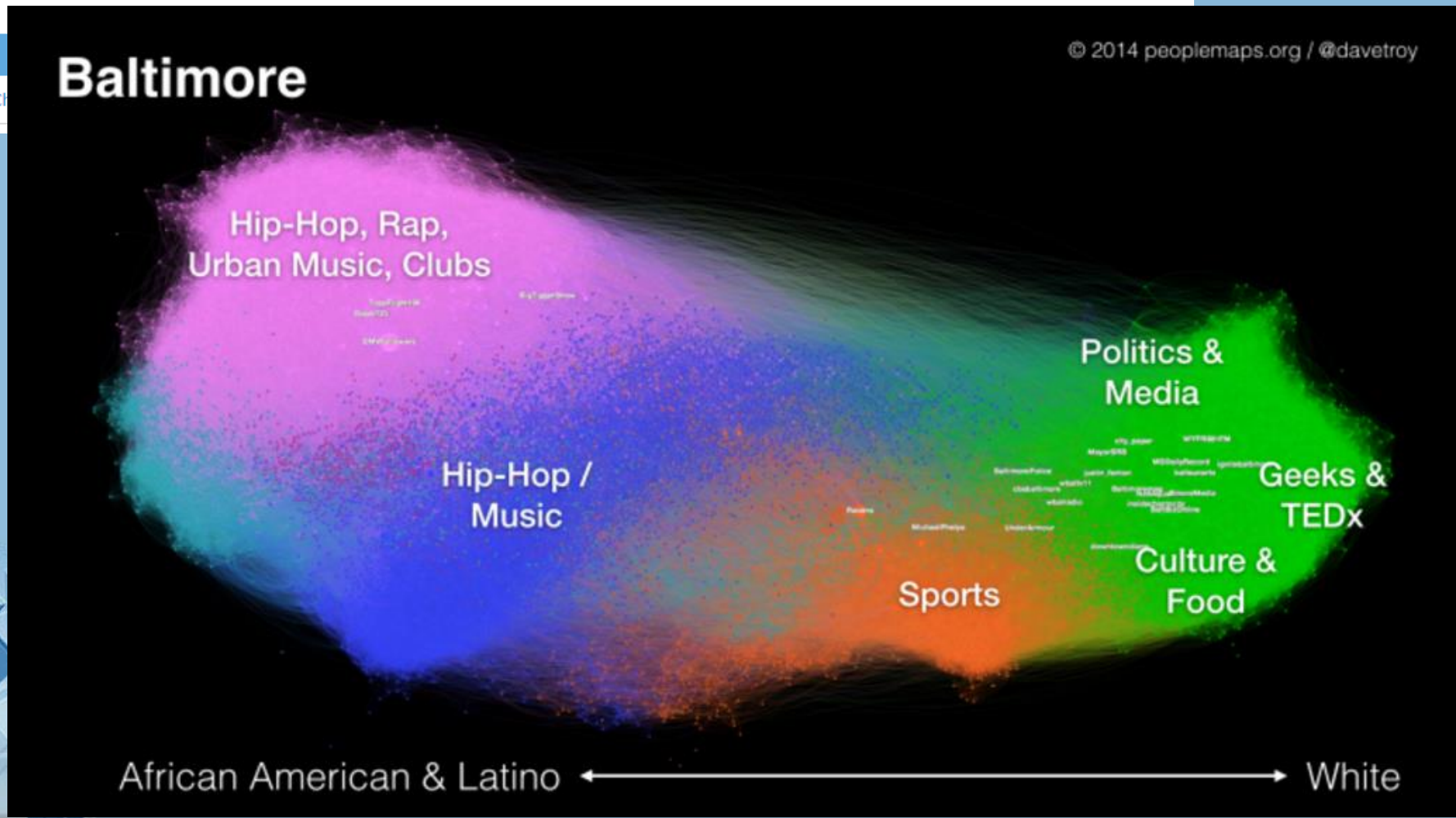
Register | Sign In | Search all projects

NODEXL Network Graphs
The Social Media Research Foundation

NodeXL: Network Overview, Discovery and Exploration for Excel

HOME

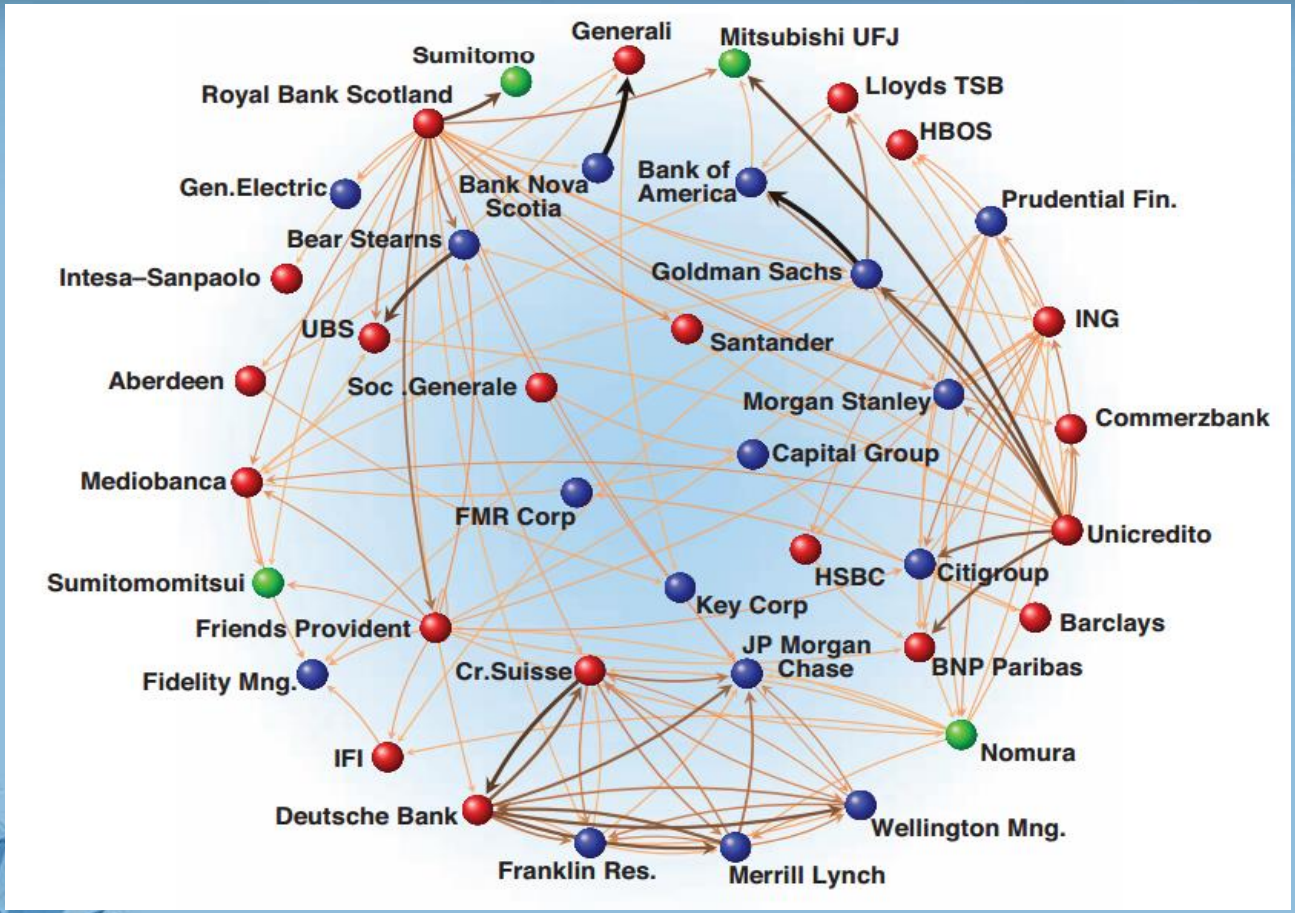
Page Info | Cl



Результат кластеризации сетевых сообществ

	A	B	C	D	E	F	G	H	I	J	K
1	User ID	Edge	First name	Last name	Screen	Sex	Birthday	City	Country	Universit	Group
2	407108	4	Ирина	Билик	irbilik	1		Санкт-Петербург	Россия		G1
3	897257	1	Андрей	Щербак	id897257	2	11.1	Санкт-Петербург	Россия	ЕУ СПб	G1
4	2123817	3	Надежда	Билик	id2123817	1	25.11	Санкт-Петербург	Россия		G1
5	2136355	3	Виктория	Лейко	id2136355	1	17.7	Санкт-Петербург	Россия	СПбГТИ (ТУ	G1
6	5413511	12	Олеся	Кольцова	id5413511	1	30.04.1973	Санкт-Петербург	Россия	СПбГУ	G1
7	18213157	1	Павел	Браславский	pbras	2	11.12.1973	Екатеринбург	Россия	УрФУ (ране	G1
8	27260599	1	Лиза	Данилова	daniliz	1	11.01.1982	Санкт-Петербург	Россия	СЗИП СПГУ	G1
9	31309245	2	Екатерина	Трушкова	id3130924	1	18.9	Москва	Россия	НИУ СГУ им	G1
10	58403904	3	Оксана	Лейко	id5840390	1		Санкт-Петербург	Россия	СПбГУ	G1
11	82161314	2	Наталия	Огнева	id8216131	1	05.08.1950	125	Россия		G1
12	225379983	2	Анна	Кольцова	id2253799	1	29.12	Санкт-Петербург	Россия		G1
13	918070	8	Даша	Самарская	dash_x	1	2.1	Санкт-Петербург	Россия		G2
14	965165	8	Эдуард	Писковацкий	repus	2	05.09.1971	Санкт-Петербург	Россия		G2
15	1887833	8	Александр	Ларюшин	id1887833	2	06.11.1970	Санкт-Петербург	Россия		G2
16	3829124	9	Михаил	Россет	mrosset	2	20.5	Санкт-Петербург	Россия		G2
17	5483431	8	Нина	Рохлина	id5483431	1	30.01.1951		Россия		G2
18	11154926	7	Кирилл	Михайлов	id1115492	2	21.04.1971	Санкт-Петербург	Россия		G2
19	27334684	7	Аня	Добровольская	dobrovolsl	1	08.08.1971	Санкт-Петербург	Россия		G2
20	86545793	8	Александр	Орлов	id8654579	2	03.05.1971	Санкт-Петербург	Россия	СПбГЭТУ (Л	G2
21	132303479	8	Марина	Сиверьянова	id1323034	1	22.05.1971	Санкт-Петербург	Россия		G2
22	8733	4	Екатерина	Гринина	e_grinina	1		Санкт-Петербург	Россия		G3
23	1093666	4	Ростислав	Борисов	rrostik	2	24.01.1971	Санкт-Петербург	Россия	СПбГУ	G3
24	1786781	2	Сергей	Чечулин	s.chechulir	2	14.02.1970	Санкт-Петербург	Россия		G3
25	1973164	2	Александр	Диденко	id1973164	2	04.12.1970	Санкт-Петербург	Россия	СПбГУ	G3
26	2480879	4	Александр	Петров	sasapetrof	2	22.03.1971	Санкт-Петербург	Россия	СПбГУ	G3
27	3207465	5	Денис	Евладов	id3207465	2		Екатеринбург	Россия	СПбГУ	G3
28	9581896	7	Иван	Круглов	id9581896	2		Санкт-Петербург	Россия		G3
29	53342500	4	Евгений	Джуринский		2					G3
30	512978	3	Евгений	Виноградов	id512978	2	3.5	Санкт-Петербург	Россия		G4
31	6123691	1	Вероника	Итигилова	id6123691	1	18.8	Санкт-Петербург	Россия		G4

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

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

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

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:

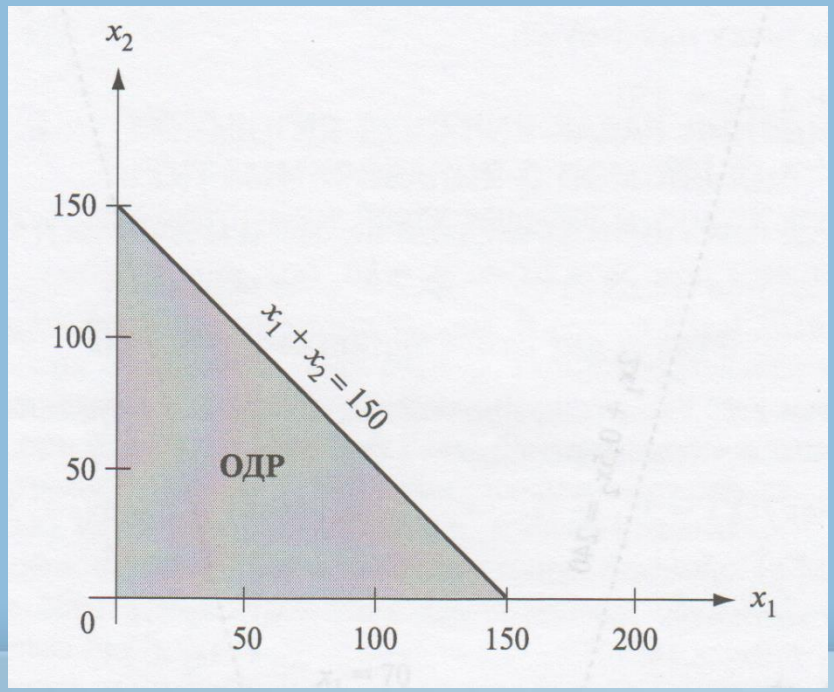
Экономико – математическая модель задачи

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.