

Научно Исследовательский Семинар

1. Кластерный анализ

2. Сетевой анализ

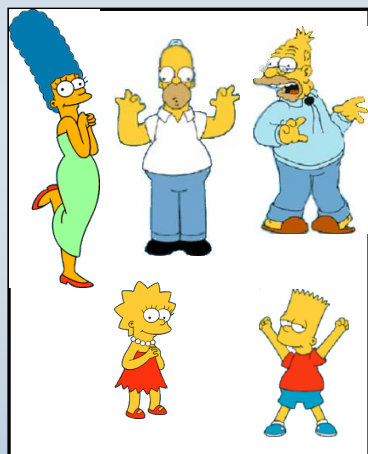
3. Задачи линейного программирования.

Кольцов С.Н

www.linis.ru

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.



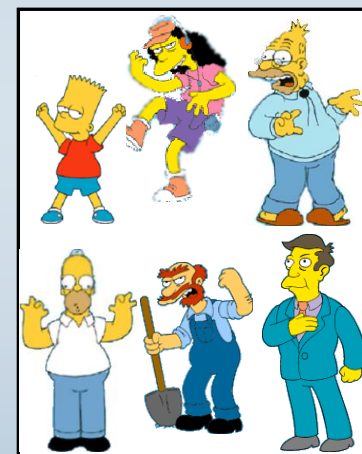
Семья



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”

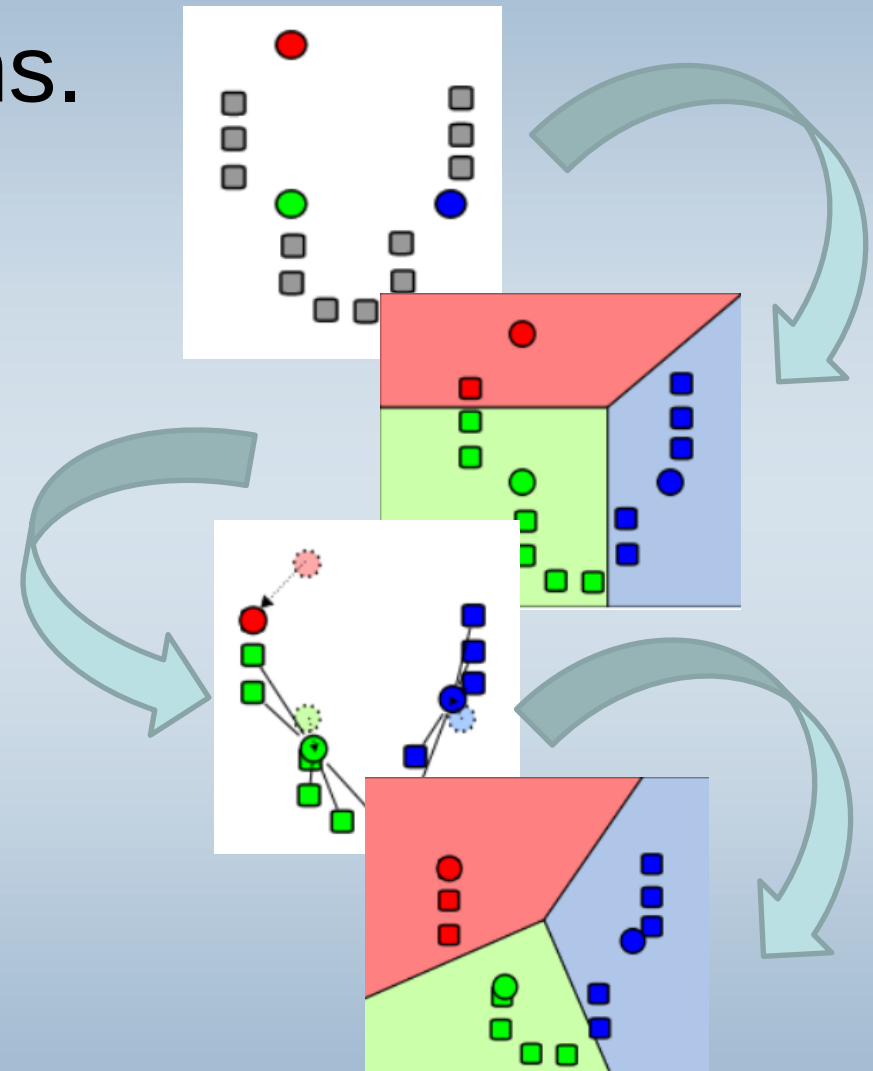
Задача определения сходства является задачей Machine learning.



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 изменяться.

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 ($l=1,2,\dots,k$; $i,j=1,2,\dots,n$).

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

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

Выбор числа кластеров – проблема остановки расчета

Задачу кластеризации можно ставить как задачу дискретной оптимизации: необходимо так приписать номера кластеров Y_i объектам X_i , 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:

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

Среднее межкластерное расстояние должно быть как можно больше:

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

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?

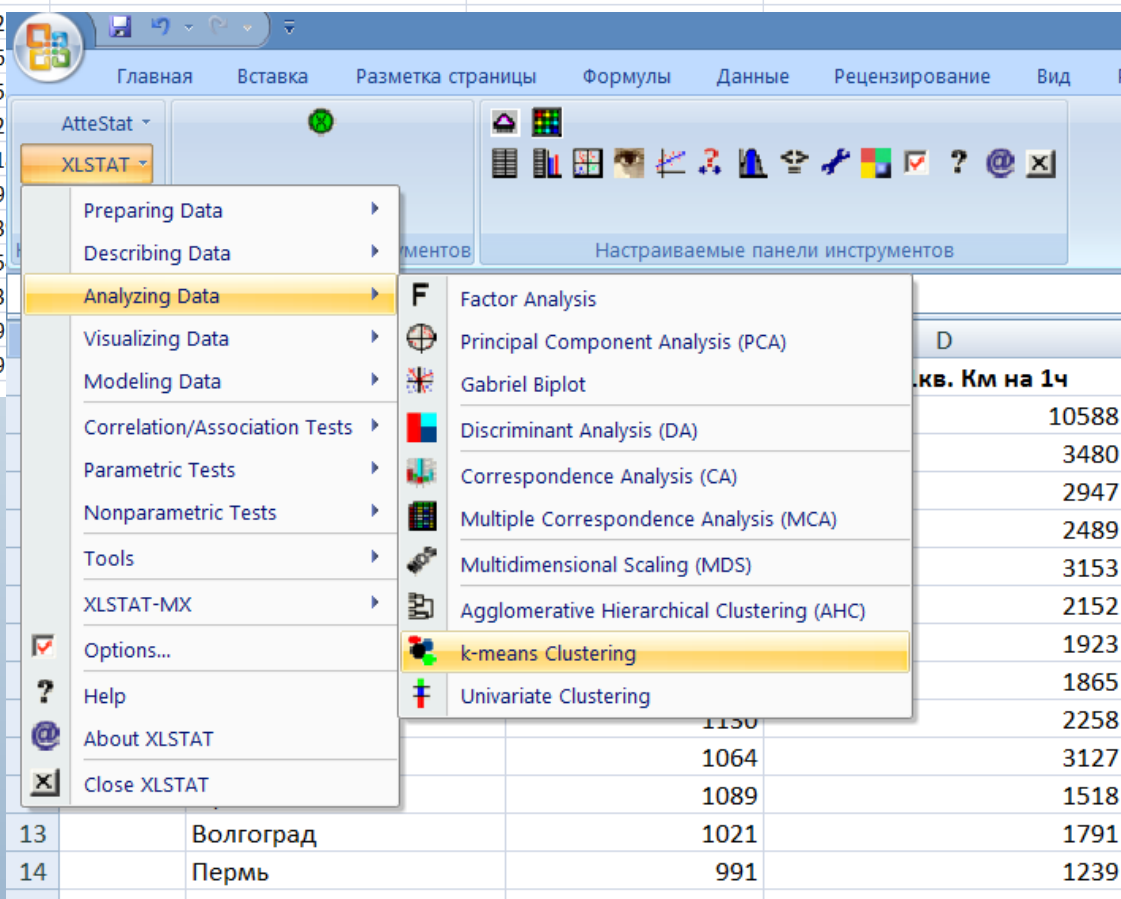
1. Запускать алгоритм много раз (с разными центрами кластеров), после чего выбрать результат с минимальной величиной ошибки.
2. Использовать дополнительные модели для оценки количества кластеров.



Кластеризация в пакете XLStat

1	2	3	4	5	6	7
	города	Население т.ч	Плотность 1 кв. Км на 1ч	среднемесячная заработная плата руб	кол. Преступлений на 100 чел.	кол. Образов. Учржд. На 100 чел.
	Москва	11514.00	10588.00	38410.00	16.00	68.00
	Санкт-Петербург	4848.00	3480.00	27189.00	13.00	72.00
	Новосибирск	1473.00	2947.00	23374.00	29.00	83.00
	Екатеринбург	1350.00	2489.00	23216.00	33.00	147.00
	Н. Новгород	1250.00	3153.00	21821.00	35.00	84.00
	Самара	1164.00	2152.00	20690.00	27.00	82.00
	Омск	1154.00	192			
	Казань	1143.00	186			
	Челябинск	1130.00	225			
	Ростов на дону	1064.00	312			
	Уфа	1089.00	151			
	Волгоград	1021.00	179			
	Пермь	991.00	123			
	Красноярск	973.00	275			
	Воронеж	890.00	163			
	Саратов	837.00	219			
	Краснодар	744.00	99			

13	Волгоград	1021	1791
14	Пермь	991	1239



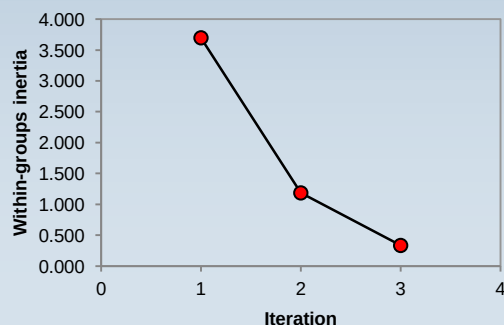
The screenshot shows the XLStat software interface. The 'Analyzing Data' menu is open, and 'k-means Clustering' is selected. The menu options are:

- Preparing Data
- Describing Data
- Analyzing Data**
 - Factor Analysis
 - Principal Component Analysis (PCA)
 - Gabriel Biplot
 - Discriminant Analysis (DA)
 - Correspondence Analysis (CA)
 - Multiple Correspondence Analysis (MCA)
 - Multidimensional Scaling (MDS)
 - Agglomerative Hierarchical Clustering (AHC)
 - k-means Clustering**
 - Univariate Clustering
- Visualizing Data
- Modeling Data
- Correlation/Association Tests
- Parametric Tests
- Nonparametric Tests
- Tools
- XLSTAT-MX
- Options...
- Help
- About XLSTAT
- Close XLSTAT

Результаты кластеризации в пакете XLStat

Сумма квадратов отклонений объектов от центра кластеров

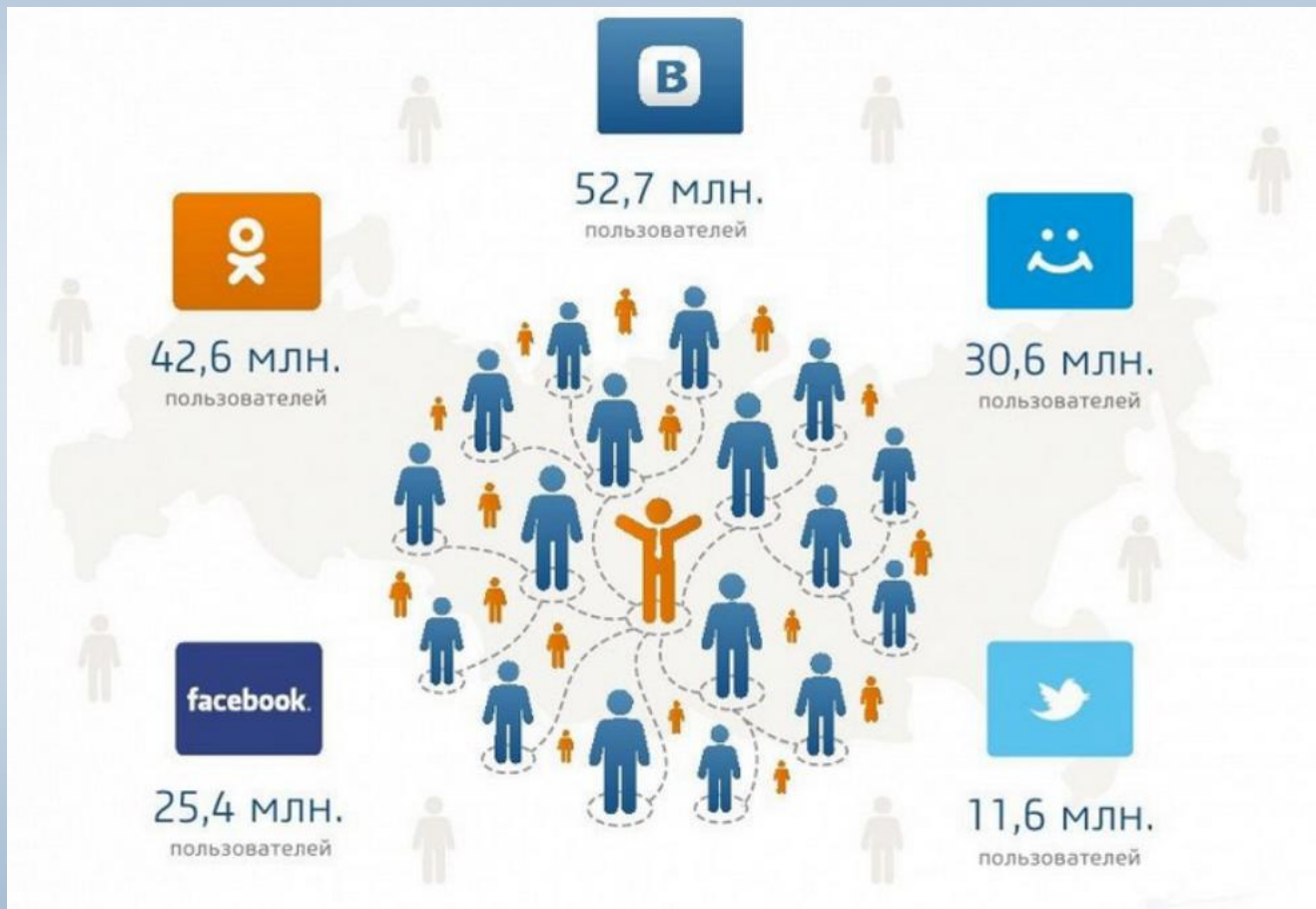
Changes in within-groups inertia



Clusters composition:

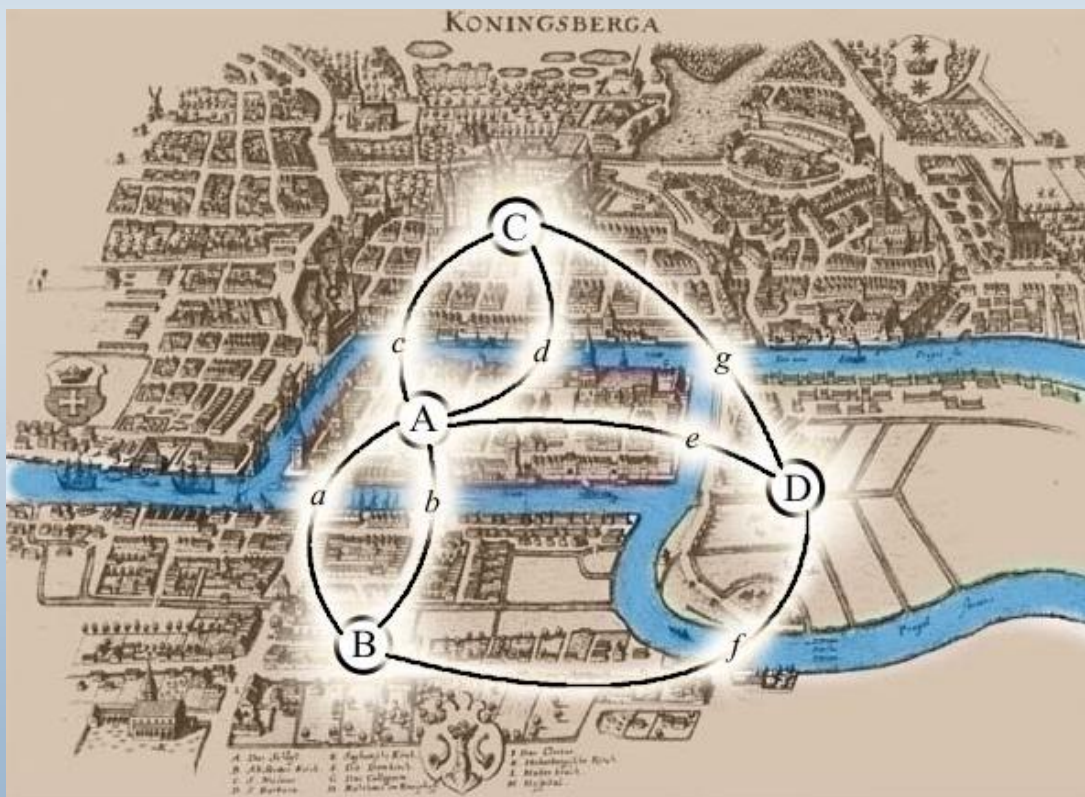
Cluster	1	2	3	4	5	6
Within-groups inertia	6073.500	27813.000	22409.000	0.000	0.000	0.000
Minimum distance from centroid	55.107	117.926	105.851	0.000	0.000	0.000
Average distance from centroid	55.107	117.926	105.851	0.000	0.000	0.000
Maximum distance from centroid	55.107	117.926	105.851	0.000	0.000	0.000
Size	2	2	2	1	1	1
	Омск Казань	Волгоград Воронеж	Самара Челябинск	Ростов на дону	Новосибирск	Уфа

Network analysis



Исследование сетей (Network analysis)

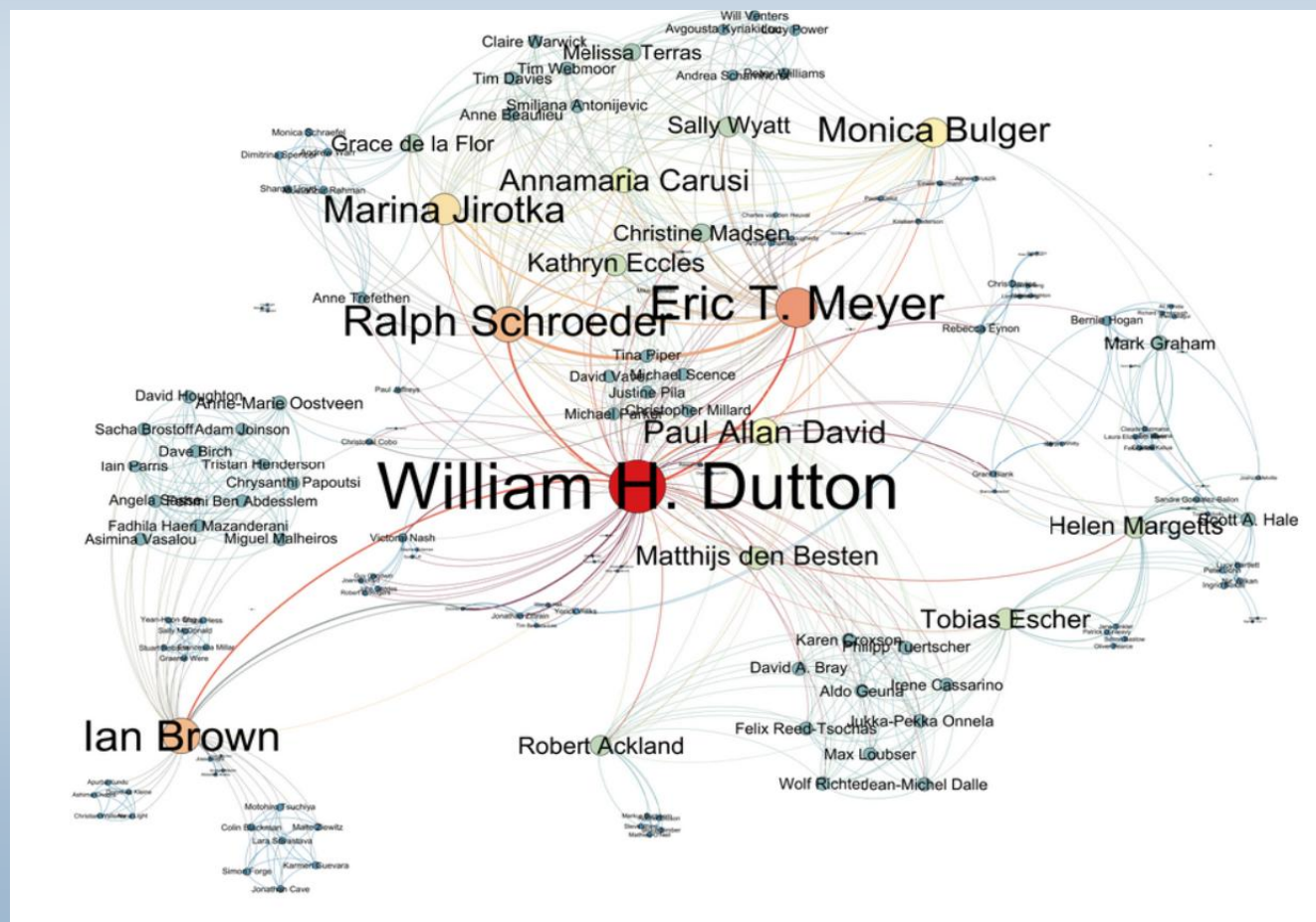
Теория графов — раздел дискретной математики, изучающий свойства графов. В общем смысле граф представляется как множество вершин (узлов), соединённых рёбрами.



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)

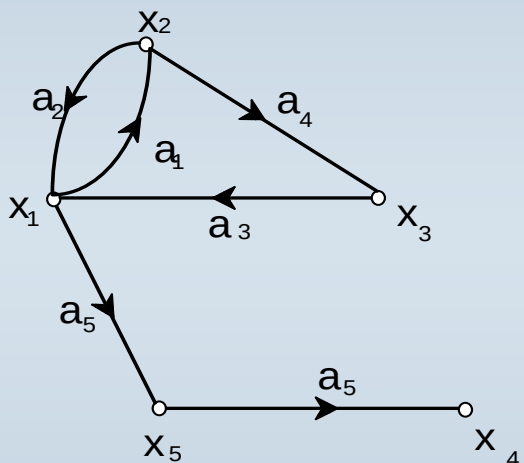
Теория шести рукопожатий — теория, согласно которой любые два человека на Земле разделены в среднем лишь пятью уровнями общих знакомых. Теория была выдвинута в 1969 году американскими психологами Стэнли Милгрэмом и Джеффри Трэверсом. Предложенная ими гипотеза заключалась в том, что каждый человек опосредованно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

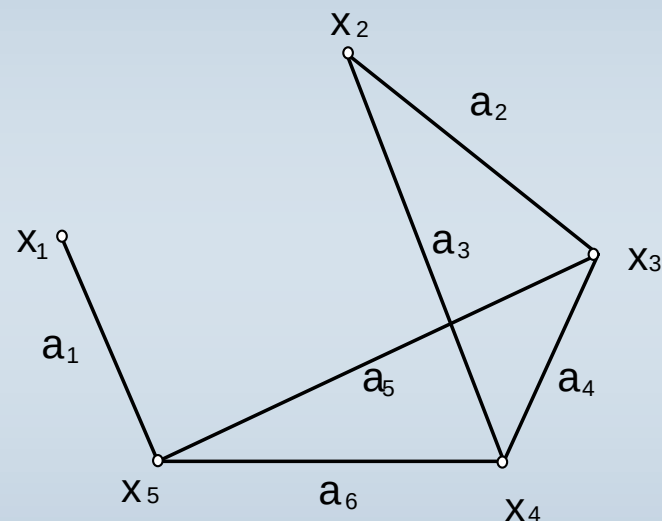
Миланский университет и социальная сеть Facebook установили, что двух любых пользователей Facebook отделяет 4,74 уровня связи. Для США количество звеньев составило 4,37.

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

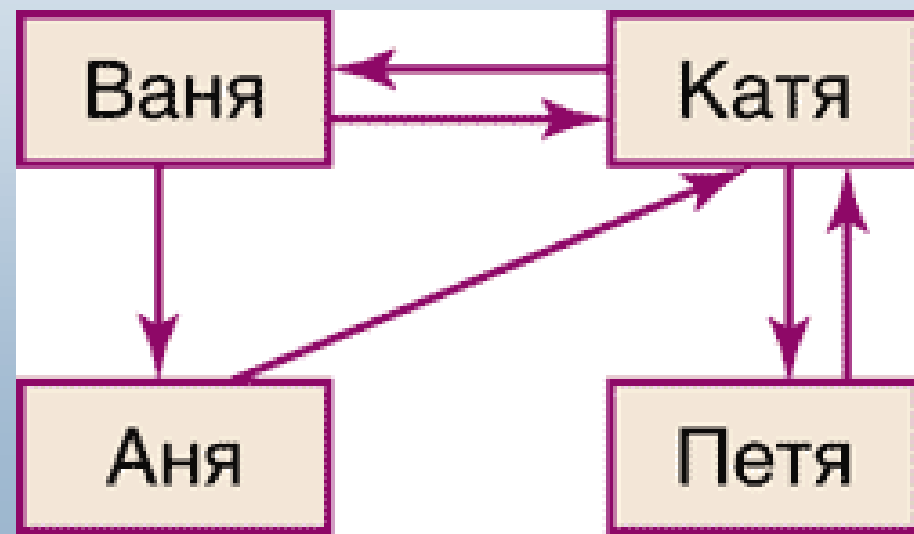
Параметры сети: Число вершин или ребер

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

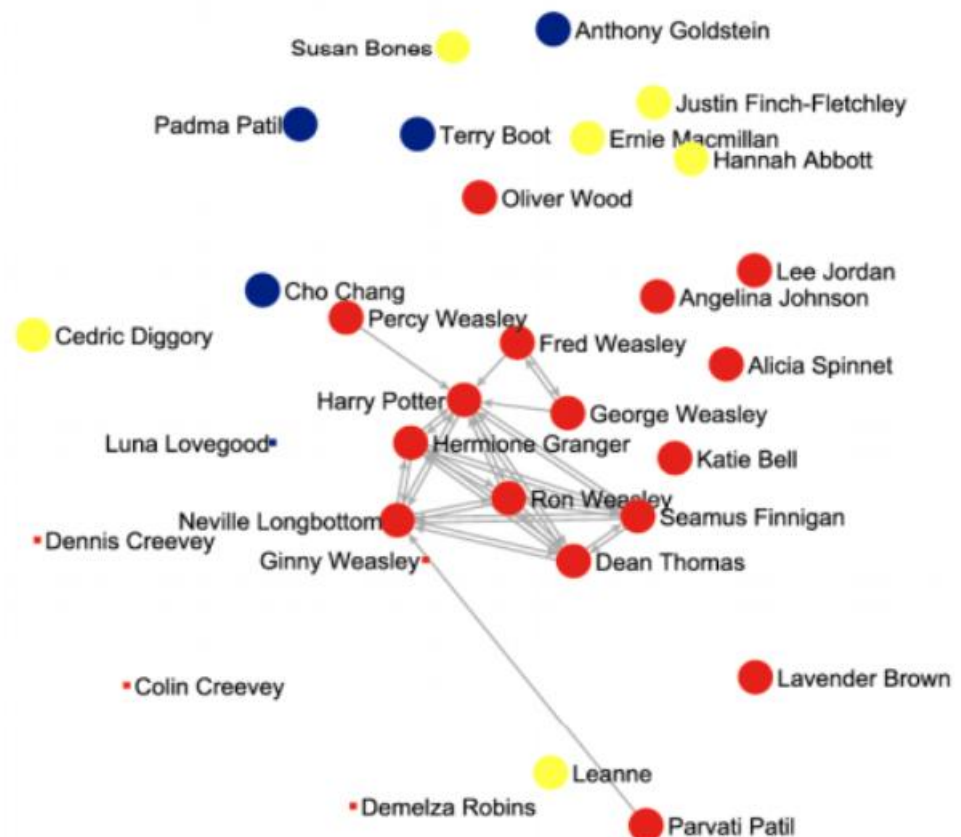
Центральность (центральным агентом сети является тот, у кого больше всего связей)



Исследование социальных сетей

An Analysis of Friendship Networks in the Harry Potter

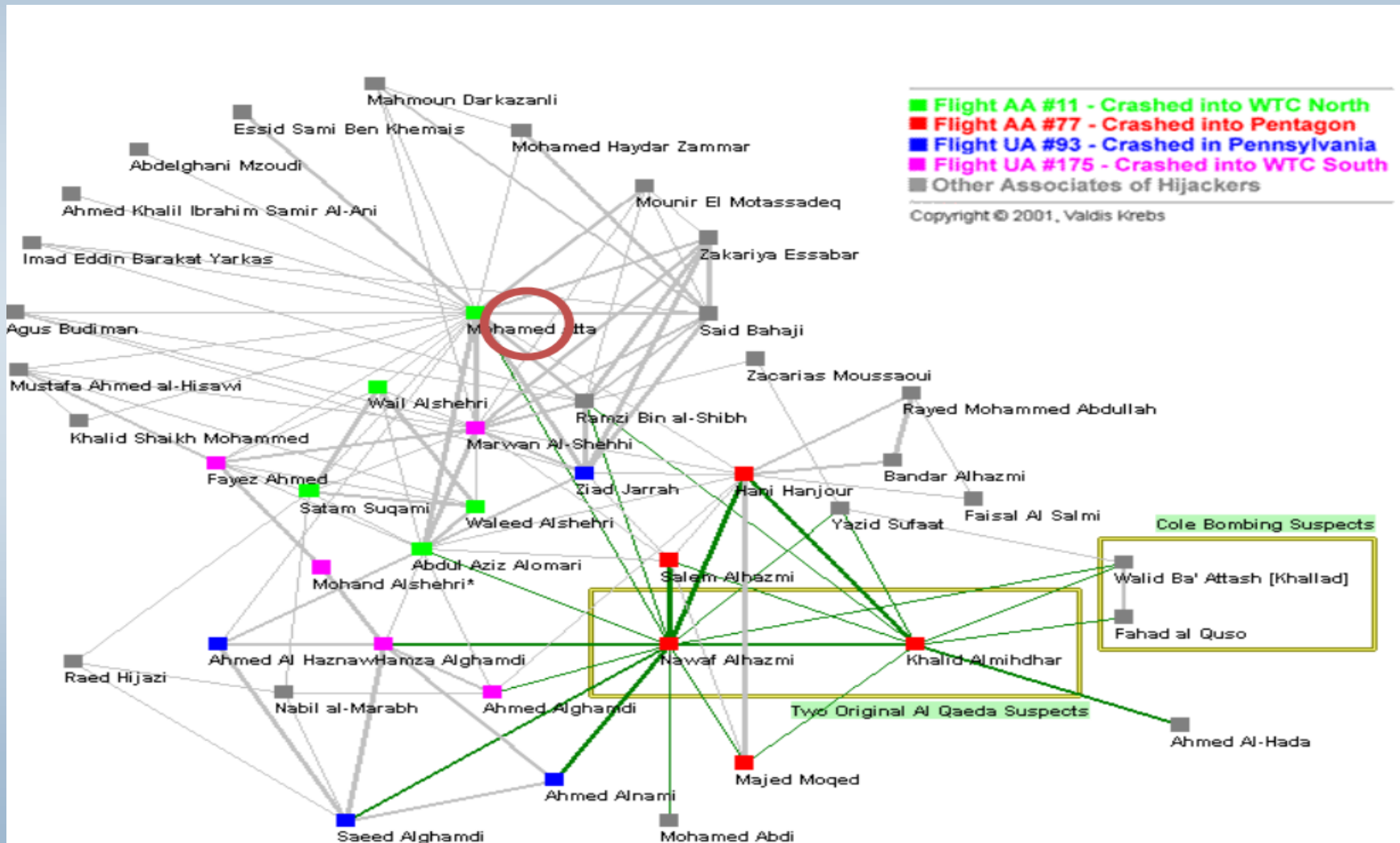
- (1) **Student A supports student B emotionally:** Harry, Ron and Hermione assure Neville that he is definitely a Gryffindor when he doubts he is not brave enough to be in house.
- (2) **Student A gives students B instrumental help:** Fred and George Weasley help Harry Potter to get his trunk into Hogwarts Express.
- (3) **Student A gives student B certain information to help student:** Hermione Granger helps Harry Potter with his homework.



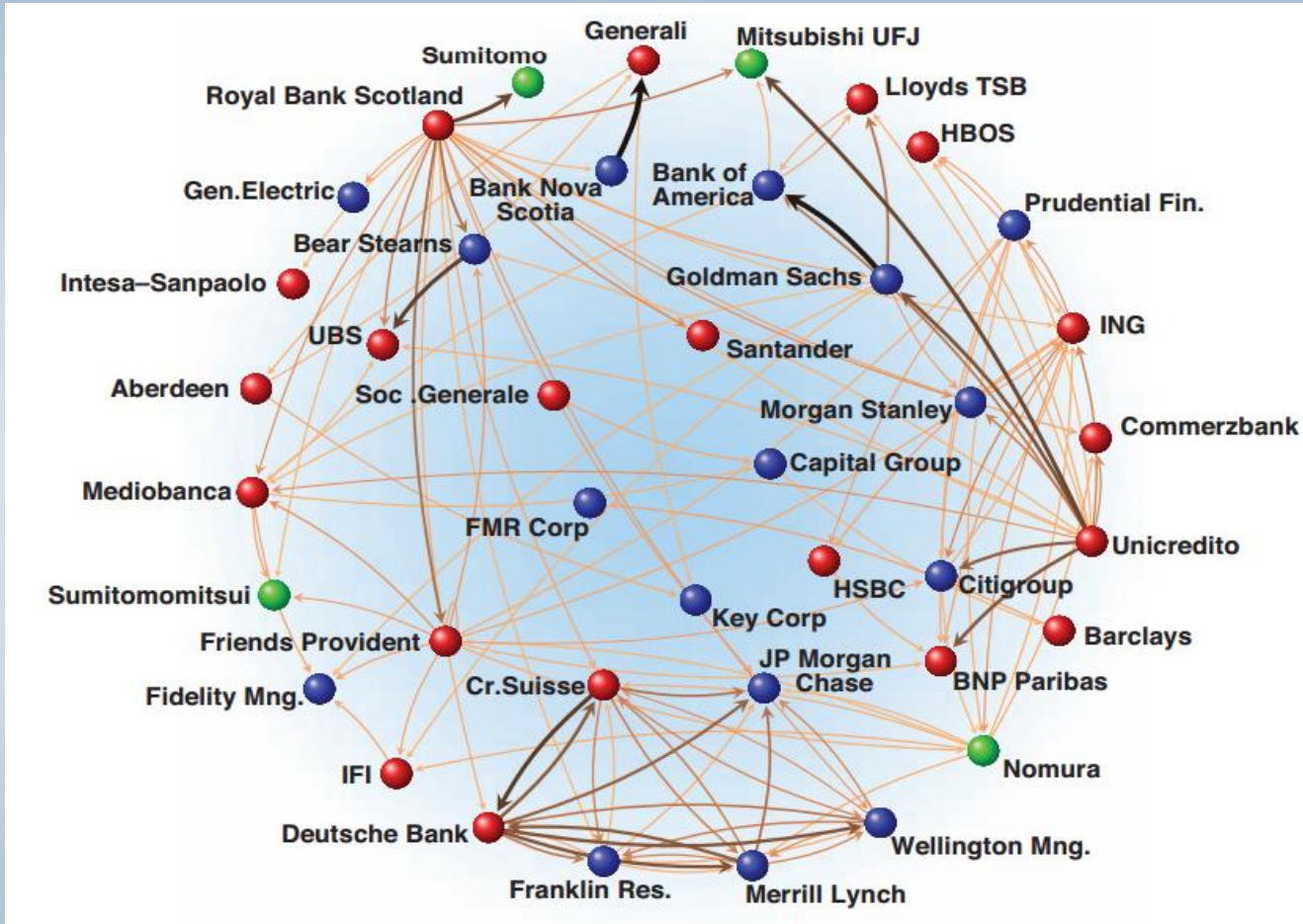
Результат кластеризации сетевых сообществ

	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 СПб	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	Наталья	Огнева	id8216131	1	05.08.1950		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	dobrovols	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.chechulin	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

Результат анализа сети участников терактов 2003г в США



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.

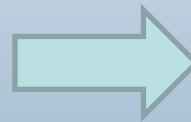
(ЗЛП).

Математическая задача в общем виде состоит в определении наибольшего или наименьшего значения целевой функции $F(x_1, x_2, x_3 \dots)$ при условии что $a_i(x_1, x_2, x_3 \dots) \leq b_i$, 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 $F(x_1, x_2, x_3...)$ - называется **целевой функцией** от переменных $x_1, x_2, x_3....$

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

Система прямых ограничений



Вектор X называется планом или допустимым решением ЗЛП.

[illegible]

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:

$$F = c_1 x_1 + c_2 x_2 \sim \max$$

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

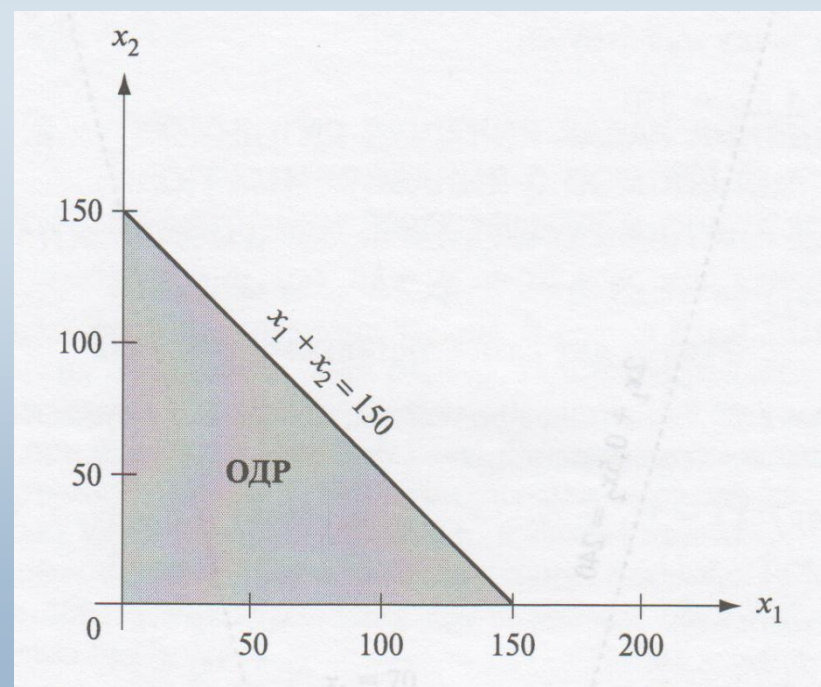
Ограничение задачи

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

1. Первое ограничение по труду

$$x_1 + x_2 \leq 150.$$

ОДР – область допустимых решений



Графическое решение задачи

1. Ограничение по лавсану

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

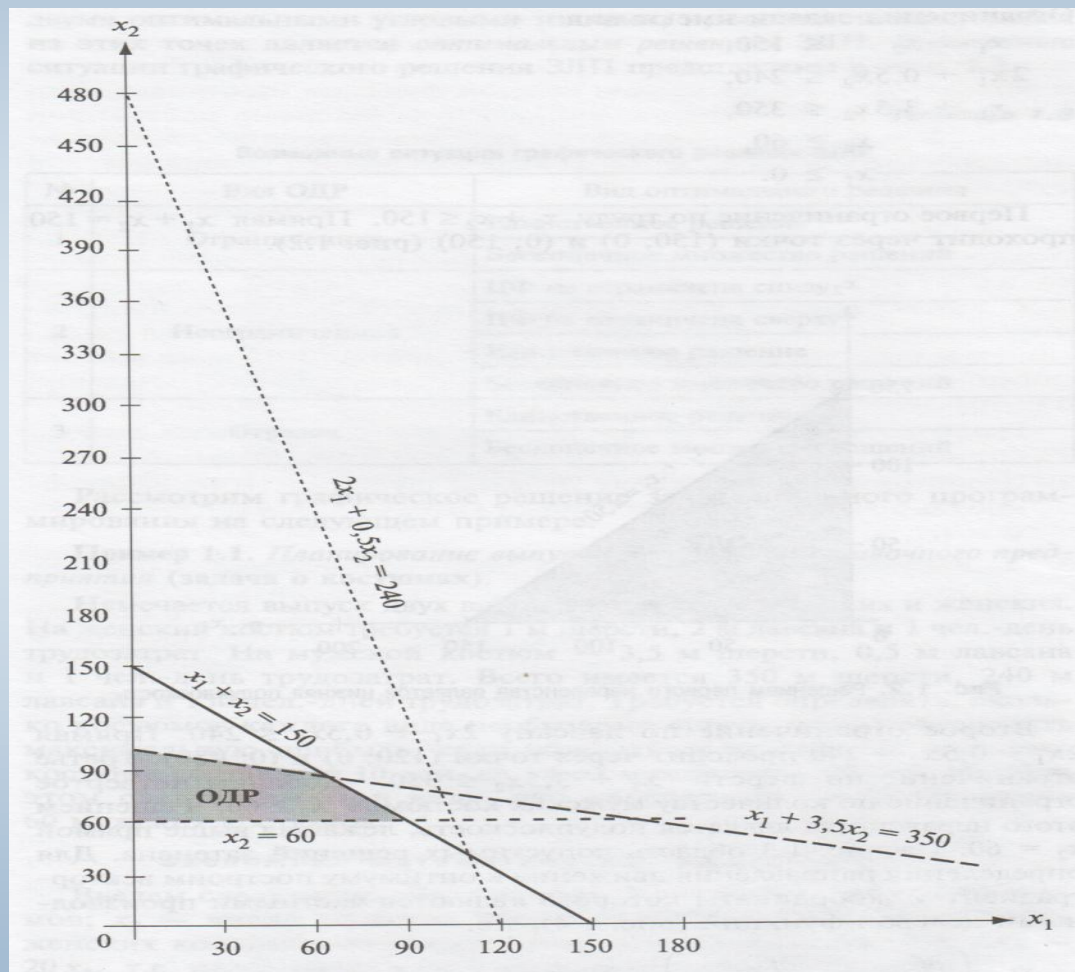
2. Ограничение по шерсти

$$x_1 + 3.5x_2 \leq 350$$

2. Ограничение по количеству мужских костюмов

$$x_1 \geq 60$$

Область допустимых значений при всех ограничениях – затемненная область.



Графическое решение задачи

С геометрической точки зрения в задаче линейного программирования ищется такая угловая точка (или набор точек) из допустимого множества решений, на котором достигается самая верхняя точка (ну или нижняя точка) в направлении наискорейшего роста.

Для нахождения экстремального значения целевой функции используется вектор – градиент, координаты которого являются производными целевой функции.

$$\nabla = \left(\frac{\partial f}{\partial x_1} = c_1, \frac{\partial f}{\partial x_2} = c_2 \right).$$



$$\nabla = \left(\frac{\partial f}{\partial x_1} = c_1, \frac{\partial f}{\partial x_2} = c_2 \right) = (10; 20).$$

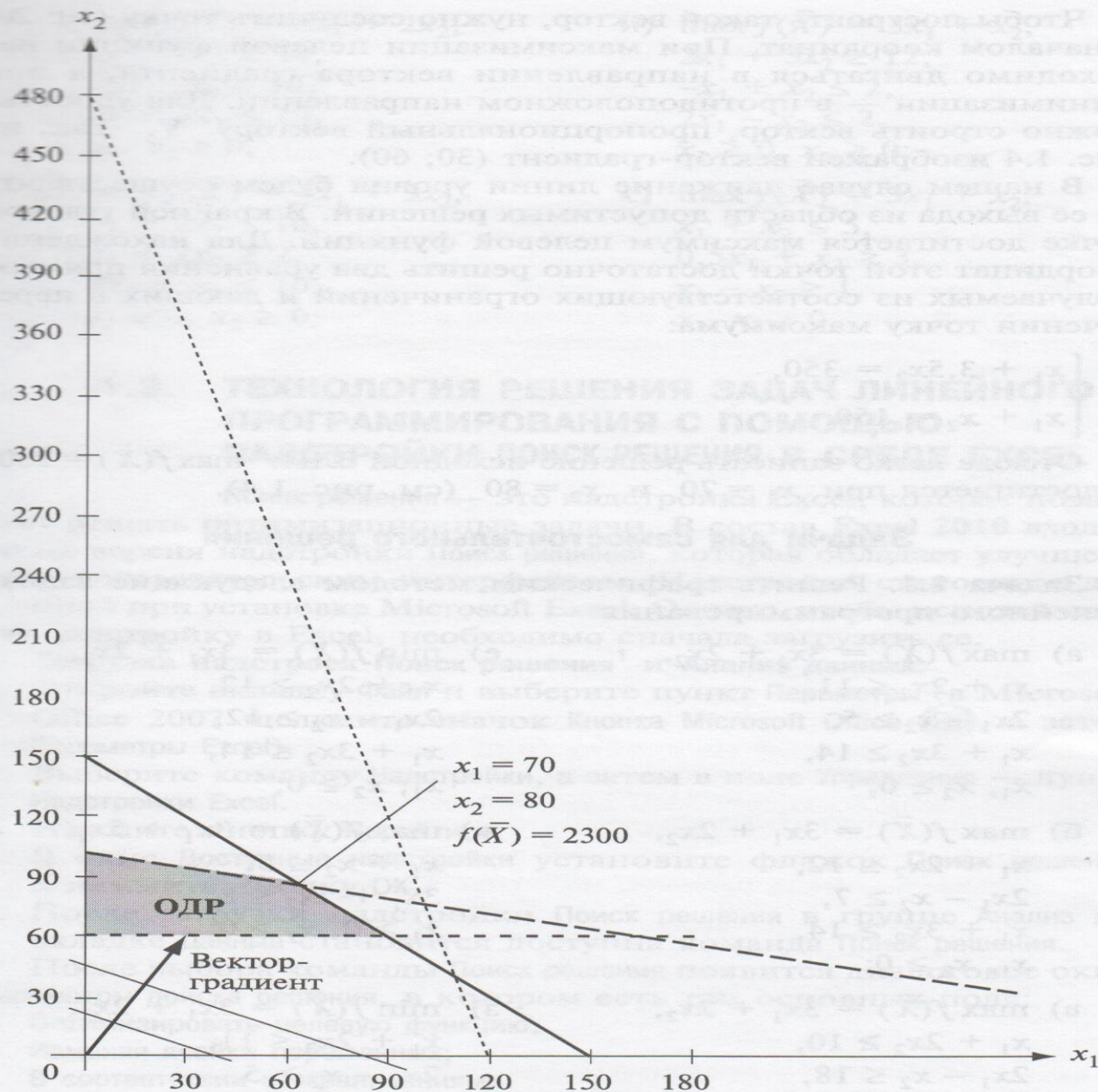
Нужно строить вектор с координатами: 0,0 – 10,20

Итоговое решение
Это точка с
координатами:

$$x_1 = 70$$

$$x_2 = 80$$

Максимум целевой
Функции: $F=2300$



THANK YOU!

← → ↻ 🏠 linis.hse.ru/o-nas

🔍 ☆ Ⓢ ☰



→ [НИУ ВШЭ в Санкт-Петербурге](#) → [Лаборатория интернет-исследований](#) → О нас

RU EN



О нас >

Мероприятия >

Исследования >

Ссылки >

Программное
обеспечение >

Публикации >

Материалы >

Работа и
стажировки в
ЛИНИС >

О нас

Лаборатория Интернет-исследований — это...

— это [команда исследователей](#), протянувших друг другу руки через дисциплинарные границы, чтобы получать новое знание о таких явлениях, которые не видны из узких колодцев отдельных наук. Миссия лаборатории — изучать интернет как уникальный феномен, имеющий социально-политическое, экономическое, лингвистическое и техническое измерения. Мы также стремимся узнать больше об обществе, для которого интернет стал естественной «средой обитания». Для этого мы разрабатываем и адаптируем математические методы, алгоритмы и [программное обеспечение](#). Все наши [проекты](#) коллективны и включают людей с разными компетенциями и разного уровня — от студентов-бакалавров и аспирантов до опытных исследователей, что обеспечивает взаимообогащение и обмен идеями. Мы поддерживаем креативную обстановку и равноправную коммуникацию в лаборатории, приветствуем новые инициативы и сотрудничество с бизнесом и государственными органами, а также проводим множество [мероприятий](#). Результаты наших усилий можно увидеть в наших [публикациях](#).