

# Системы разработки данных и машинного обучения

## Спектральная кластеризация для анализа

## Интернет-данных

Дмитрий Игнатов и Леонид Жуков

Национальный исследовательский университет Высшая школа экономики  
Кафедра анализа данных и искусственного интеллекта

28 января 2014

# План Доклада

- Социальные сети
  - ▶ Нахождение сообществ
- поисковая реклама
  - ▶ Сегментация рынка
- Интернет-радио
  - ▶ Рекомендательная система
- Математическая модель
  - ▶ Граф
  - ▶ Кластеризация (алгоритмы на графах)

# Социальные сети

- Социальная сеть (social network) — социальная структура, состоящая из группы узлов, которыми являются социальные объекты (люди или организации), и связей между ними (социальных взаимоотношений) - [Wikipedia](#)
- Интернет (2000 —)
  - ▶ FaceBook, (1.11 млрд. на март 2013), MySpace (25 млн. на июнь 2012), Friendster, ...
  - ▶ Одноклассники (205 миллионов на январь 2013 ), В контакте (>100 млн. на май 2013), Мой Круг ...
- Математическое представление – граф  $G( V, E )$ 
  - ▶ Множество вершин  $|V|$  – “люди”
  - ▶ Множество ребер  $|E|$  – “отношения”
  - ▶ Ориентированный / неориентированный

# Возможные исследования

- Анализ структуры
  - ▶ идентификации ролей пользователей
  - ▶ развитие и рост сети
  - ▶ нахождение сообществ
- Процессы в сети
  - ▶ распространение информации
  - ▶ распространение влияния
  - ▶ сетевая экономика
- Реклама и монетизация

# Социальная сеть Flickr

Photos: [Yours](#) · [Upload](#) · [Organize](#) · [Your Contacts](#) · [Explore](#)

**flickr** BETA

Hi leonid68!

- You have 4 new messages.
- [Choose your Flickr web address!](#)

**Printing? Can it be true?**  
Well, it's true if you're in the U.S., with more countries coming online soon! Get 10 free prints with your first order! [Click here to set yourself up for printing.](#)

**Flickr News**  
09 Feb 06 - Ladies and gentlemen, you'll notice a brand, spanking new line down in the footer. It's the Flickr Community Guidelines (applause, please)... [read more news](#)

[» Flickr Blog](#) Great photos & latest news, daily!

**Do more with your photos!** ENHANCED  
[!\[\]\(e236d1893811a6c5ad2d17439e3819c8\_img.jpg\) Poster Books](#) [!\[\]\(7e0d6a31a51eb3952a6a6daebf7e401c\_img.jpg\) DVDs](#)

Now there's even [more you can do](#) with your photos:

- [PhotoShow DVDs](#) NEW
- [Goog Calendars](#) NEW
- [Zazzle](#) U.S. postage stamps with your photos
- [Engage](#) back-up DVDs and Slideshows

Make your photos happy - do something with them!  
And don't forget to set your [printing preferences](#) so we

Photos: [Yours](#) · [Upload](#) · [Organize](#) · [Your Contacts](#) · [Explore](#)

**flickr** BETA

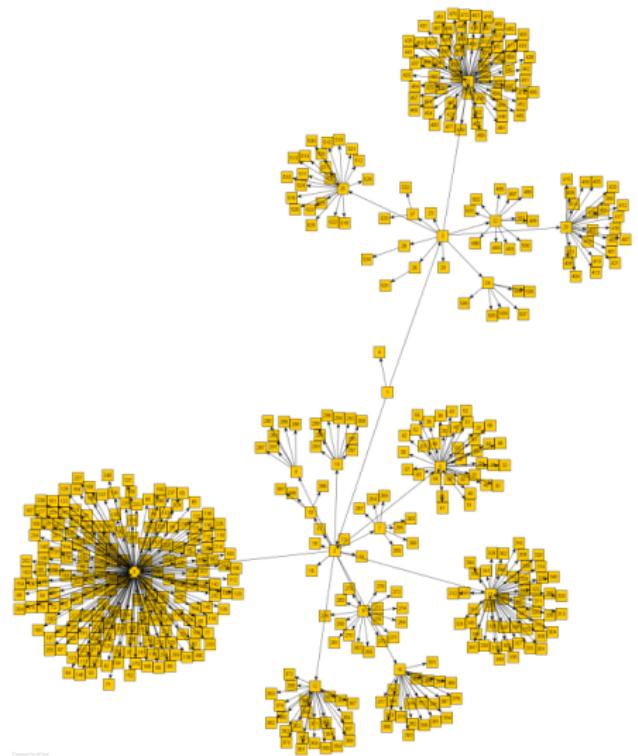
**Your contacts**

**Friends (5)**

  
Victoria Temple   stofit   latema 2006   Svetlana   mihajlo

Search for people    
(Or, try the [advanced search](#).)

• [Who counts you as a contact?](#)

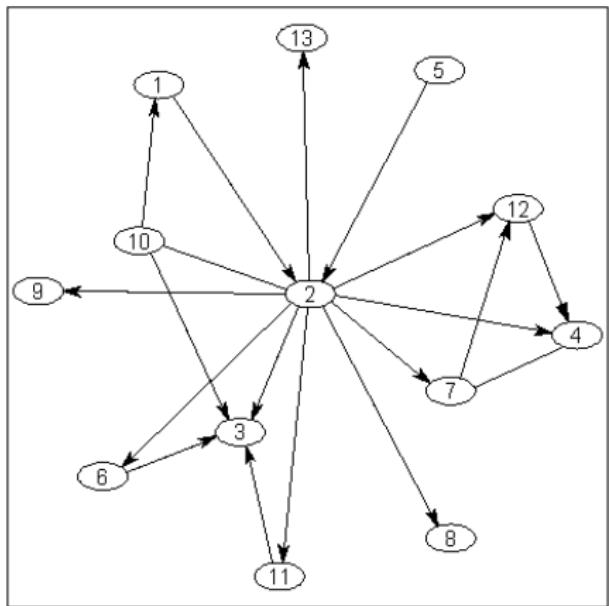


## Графическое представление



# Матрица смежности

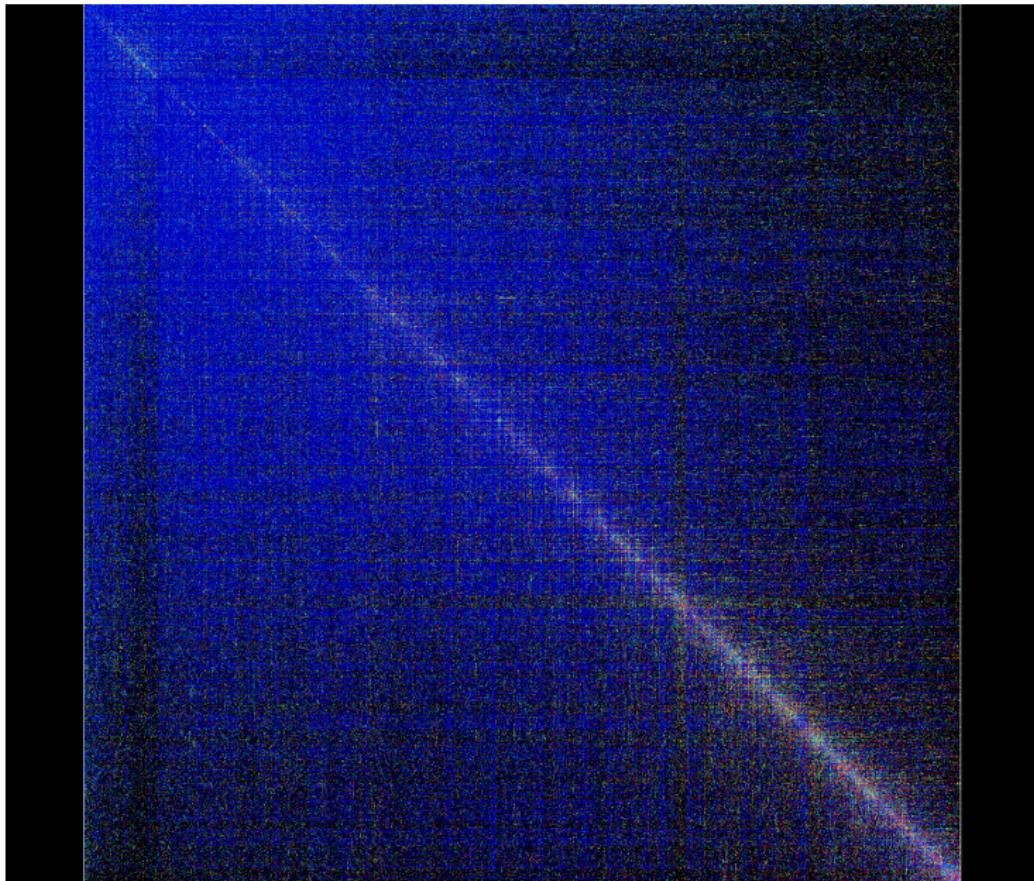
## Adjacency matrix



1	*													
2		*	*			*	*	*	*	*	*	*	*	*
3														
4														
5		*												
6			*											
7			*											*
8														
9														
10	*	*	*											
11			*											
12														
13														

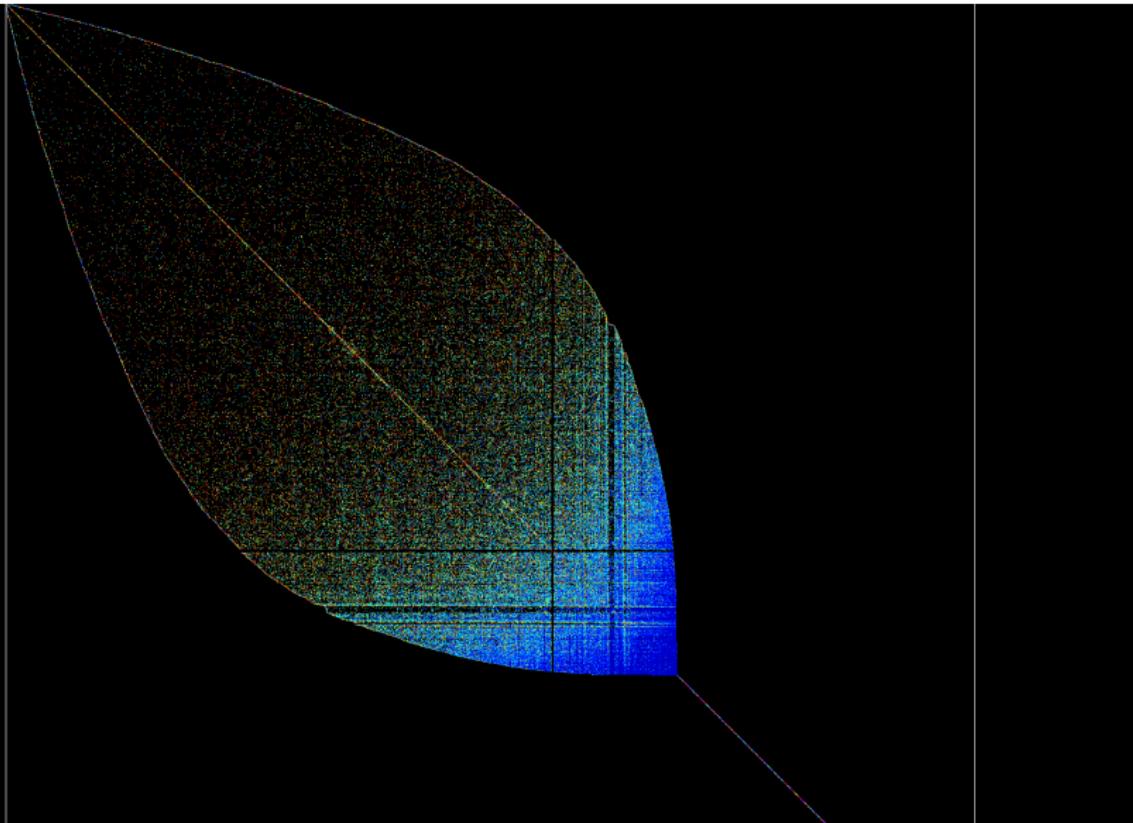
nz = 19

# Матрица смежности



Сортировка Cuthill-McKee

Reverse Cuthill-McKee ordering



## Flickr: статистика

- количество узлов (пользователей): 584207
- количество ребер (связей): 3555115
- максимальная входящая степень узла: 3531
- максимальная выходящая степень узла: 8976
- ср. входящая степень узла = ср. выходящая степень узла = 6
- диаметр графа: 18
- средняя длина пути: 5.3
- число сильно связанных компонент: 152324
- наибольшие сильно связанные компоненты: 274649, 374, 186, 155
- число связанных компонент: 43189
- наибольшие связанные компоненты: 404893, 378, 112, 108
- максимальное ядро (core number): 249 (size 668)

# Безмасштабные сети (scale-free networks)

- Степенной закон распределения степеней узлов (power law)  $P(k) = Ck^{-\gamma}$
- Медленно растущее среднее расстояние между узлами (small world)
- Высокий коэффициент кластеризации
- Наличие гигантской связанной компоненты

# Безмасштабные сети

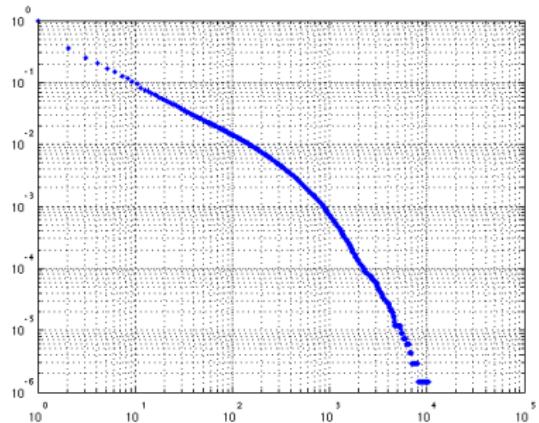


Рис.: Распределение степеней вершин

$$P(k) = Ck^{-\gamma}$$

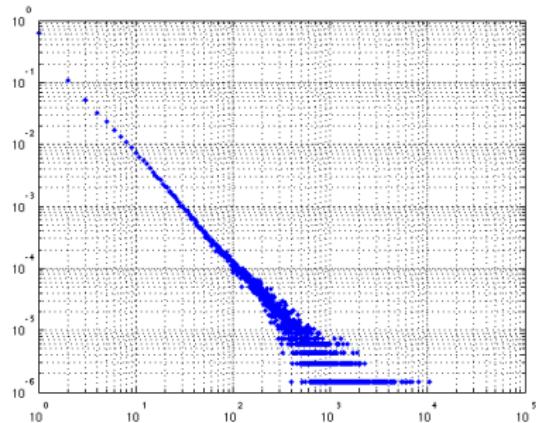


Рис.: Распределение степеней вершин

# Безмасштабные сети

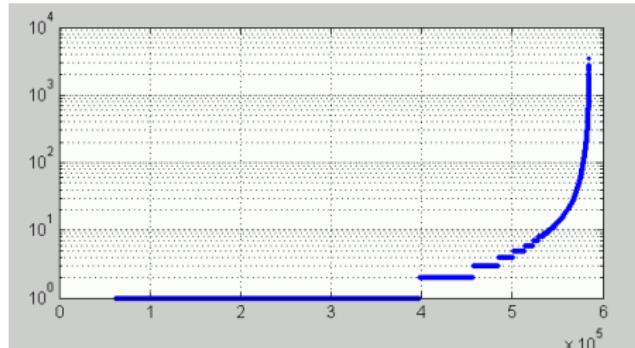


Рис.: Вершины отсортированы по входящей степени

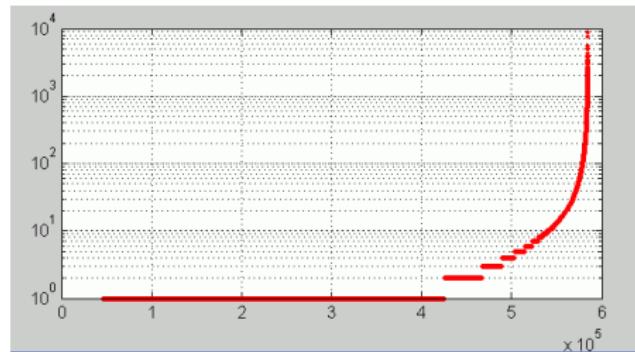


Рис.: Вершины отсортированы по исходящей степени

# $K$ – ядра графа (graph $k$ -cores)

V. Batagelj and M. Zaveršnik

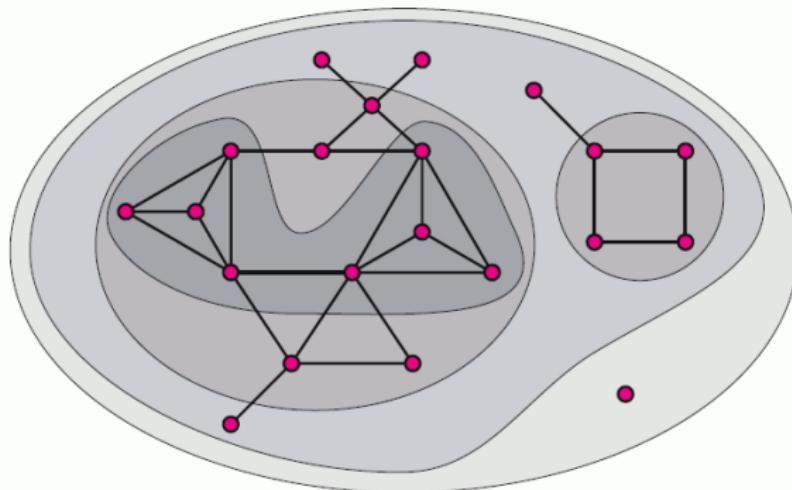
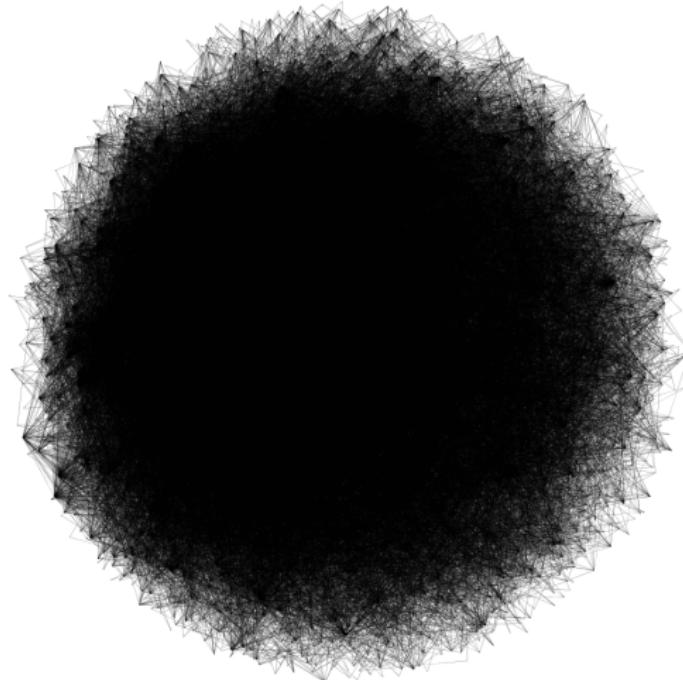
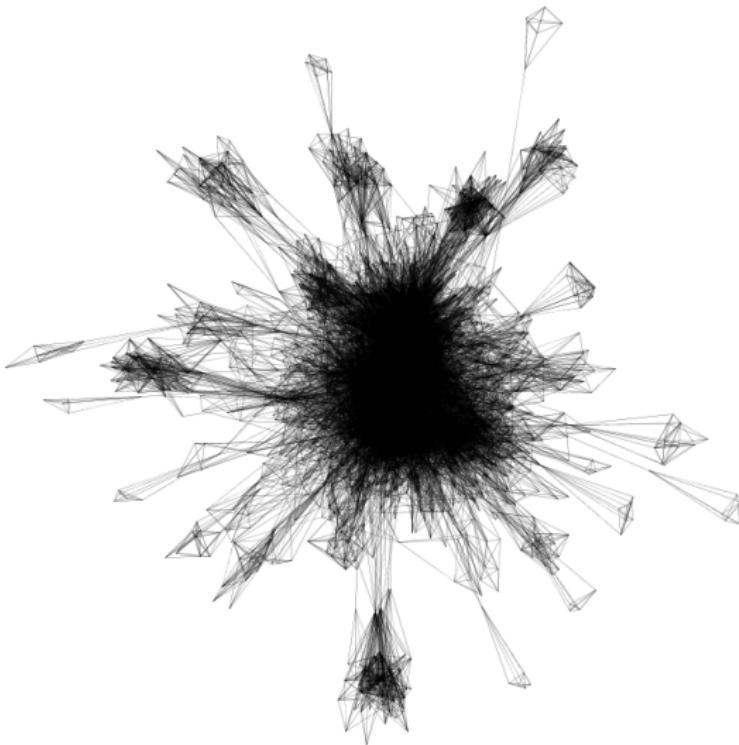


Рис.:  $k$ -ядро – максимальный подграф со степенями вершин  $\geq k$

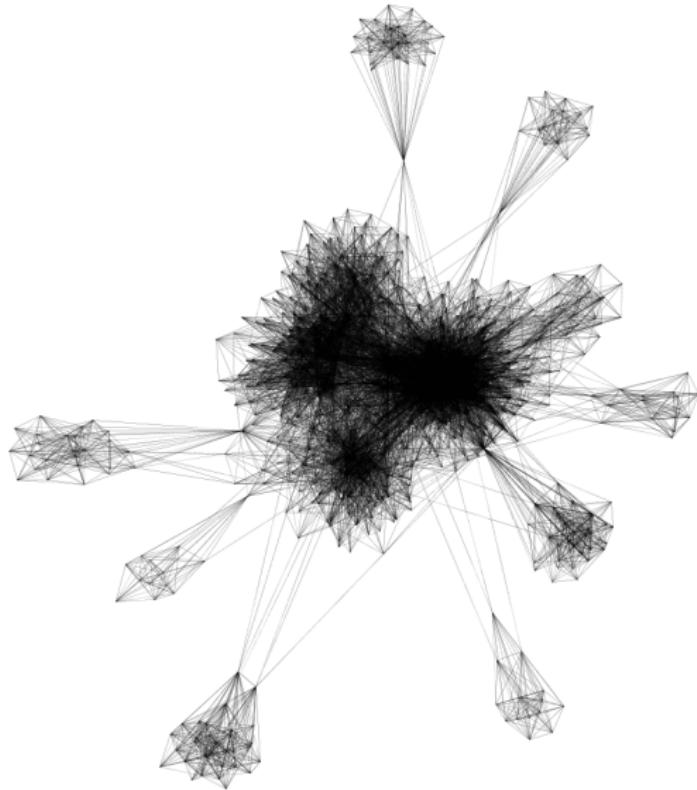
К - ядра. 2-core, 7815 вершин



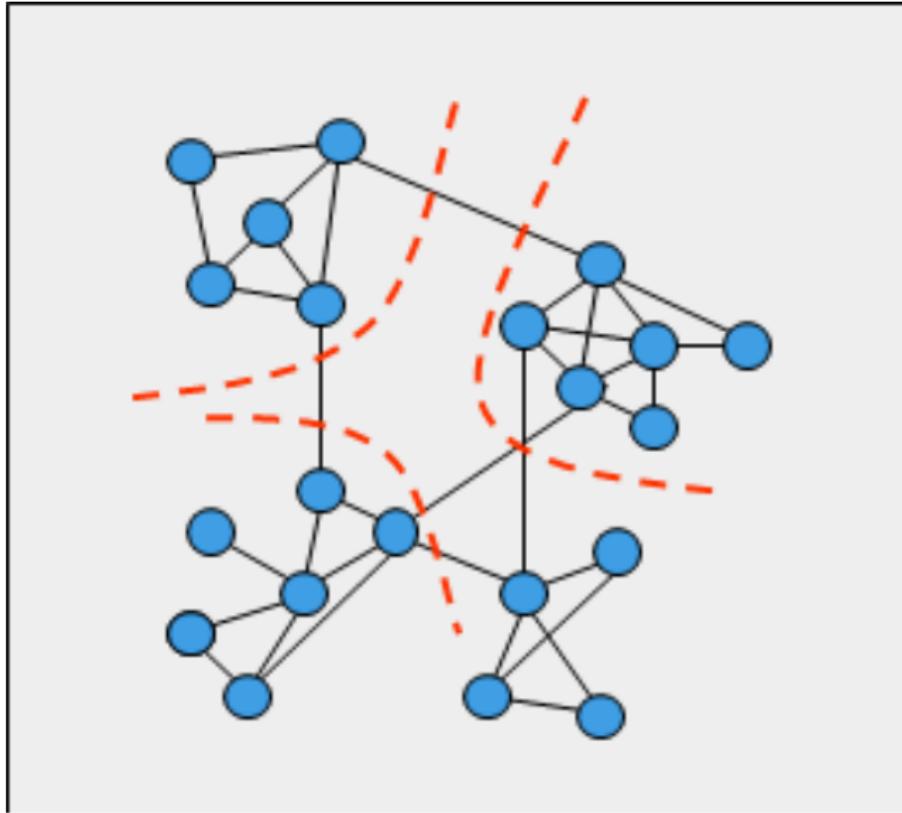
К - ядра. 5-core, 2233 вершин



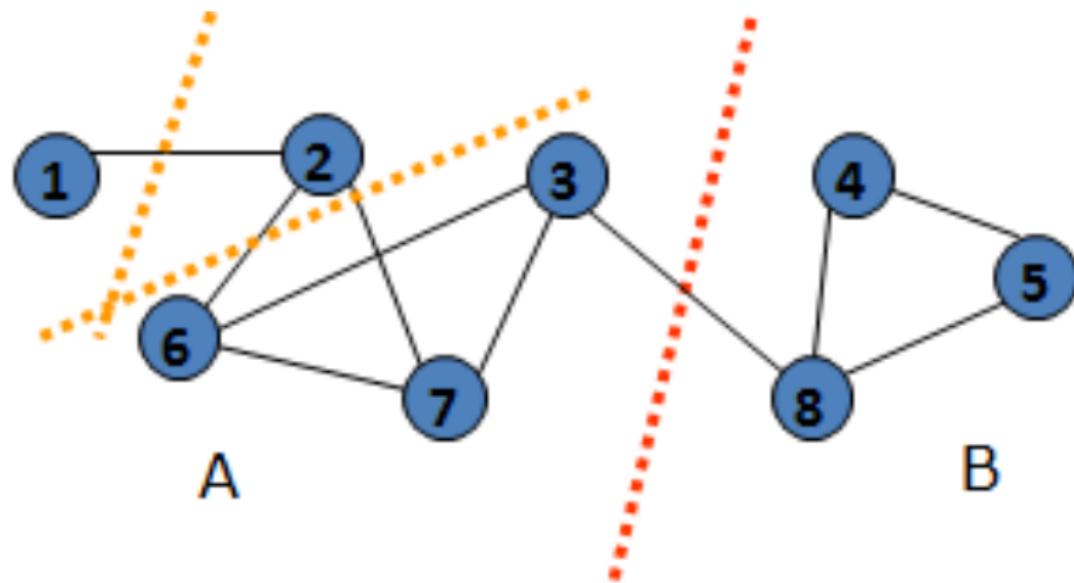
К - ядра. 10-core, 819 вершин



# Разделение графа (graph partitioning)



## Разделение графа

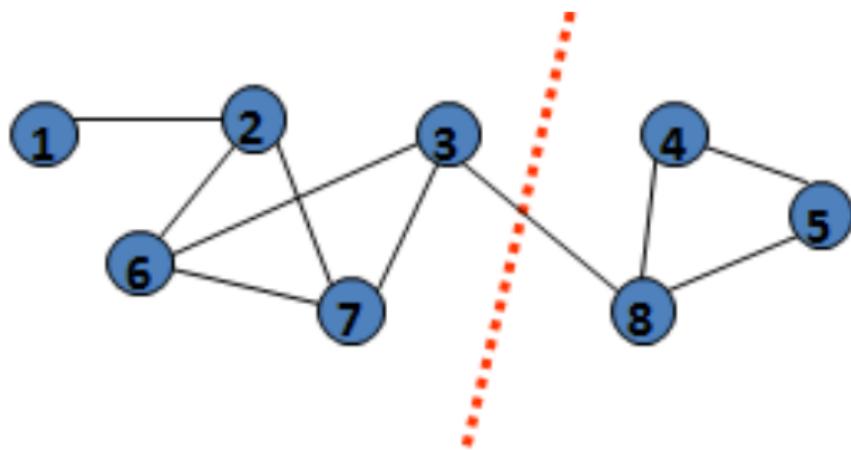


Разделение графа:  $cut(A, B) = \sum_{v_a \in A, v_b \in B} \omega(v_a, v_b)$

Нормированное разделение:  $NCut(A, B) = \frac{cut(A, B)}{\text{assoc}(A, V)} + \frac{cut(A, B)}{\text{assoc}(B, V)}$

J. Shi and J. Malik, 2000

## Спектральное разделение графа



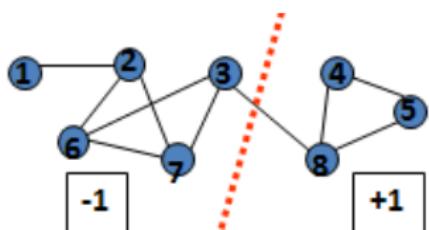
M. Fiedler, 1973

- Каждый узел характеризуется индикатором класса:  $p = \pm 1$ ,  
 $p = \{-1, -1, -1, +1, +1, -1, +1\}$
- Оптимальное разделение:  $cut = \frac{1}{4} \sum_{i>j} (p_i - p_j)^2 \omega_{ij}$
- Это задача комбинаторной оптимизации, она NP-сложна:  
 $p_i = \{-1, 1\}^N \Rightarrow x_i \in [-1, 1], x_i \in R^1$

# Решение

- Квадратичная оптимизация:  $E = \frac{1}{4} \sum_{i>j} (x_i - x_j)^2 \omega_{ij} = \frac{x^T L x}{4}$   
 $\sum_i x_i^2 = N, (x^T, e) = 0$
- Поиск собственных значений:  $Lx = \lambda x, \lambda_2, x_2$
- Округление:  $p_i = +1, if x_i > 0; p_i = -1, if x_i < 0$   
 $p = \{-1, -1, -1, +1, +1, -1, +1\}$

# Пример: нормированное разделение (normalized cut)



$$L = D - A$$
$$Lx = \lambda Dx$$
$$\lambda_2, x_2$$

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 3 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 2 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 2 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 & 0 & 3 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 3 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 3 \end{pmatrix}$$
$$x = \begin{pmatrix} -0.2641 \\ -0.2285 \\ -0.0341 \\ 0.3294 \\ 0.3294 \\ -0.1646 \\ -0.1646 \\ 0.2406 \end{pmatrix} \Rightarrow p = \begin{pmatrix} -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

# Спектральное разделение графа

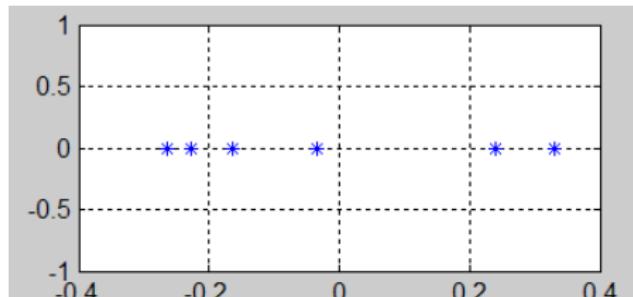


Рис.: Расположение узлов

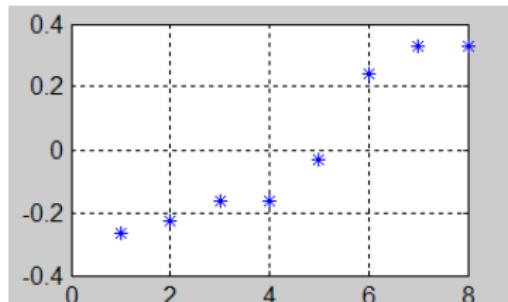


Рис.: Собственные векторы

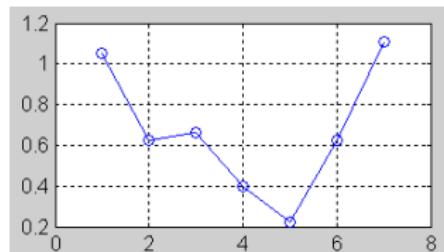
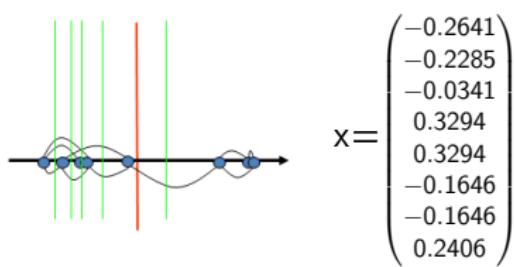
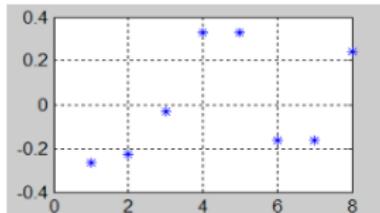


Рис.: Значения разделений (cut values)

# Спектральная сортировка

**Собственные вектора**

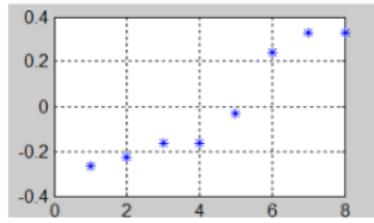
-0.2641  
-0.2285  
-0.0341  
0.3294  
0.3294  
-0.1646  
-0.1646  
0.2406



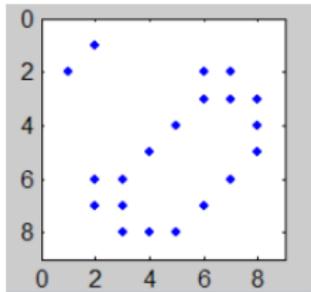
perm = [ 1 2 6 7 3 8 4 5 ]

**Собственные вектора - отсортированные**

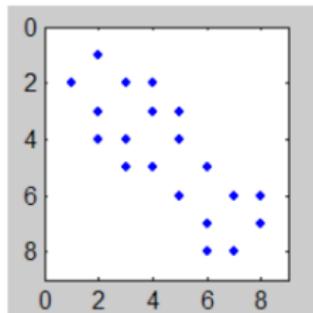
0.3294  
0.3294  
0.2406  
-0.0341  
-0.1646  
-0.1646  
-0.2285  
-0.2641



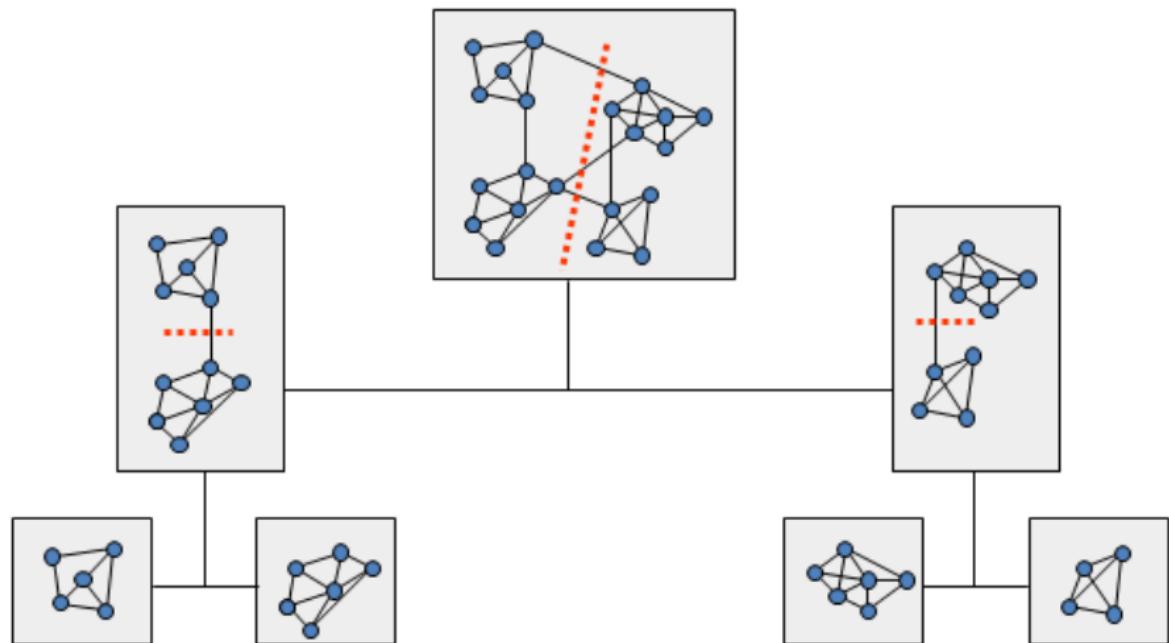
**Матрица смежности**



**Матрица смежности - переупорядоченная**



## Спектральная сортировка



# Спектральная сортировка

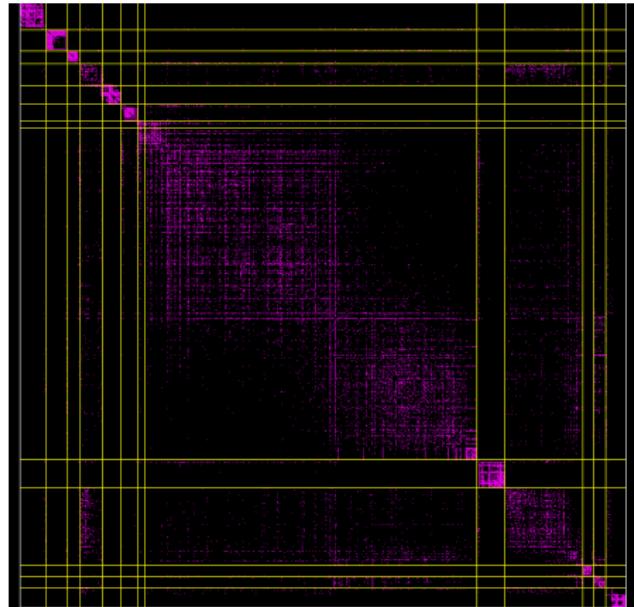


Рис.:  $2^N$  vs  $N$

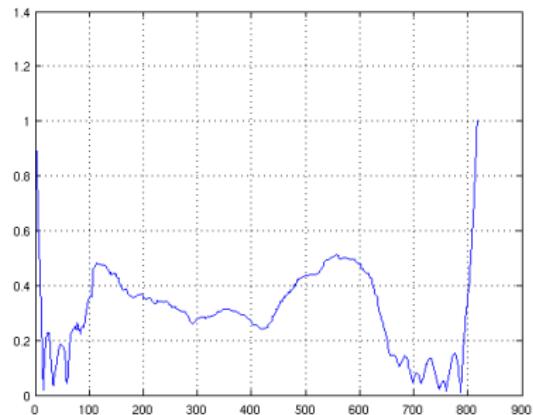
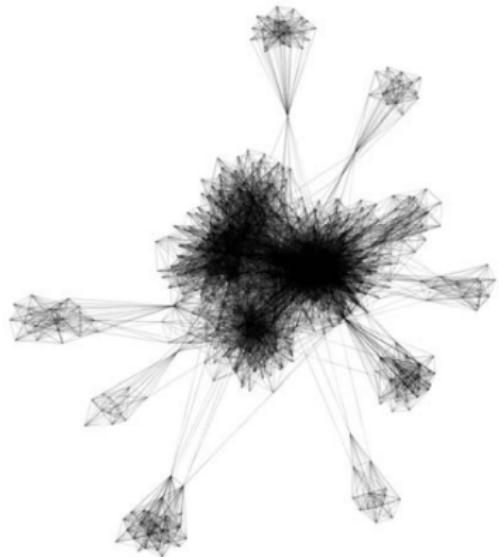
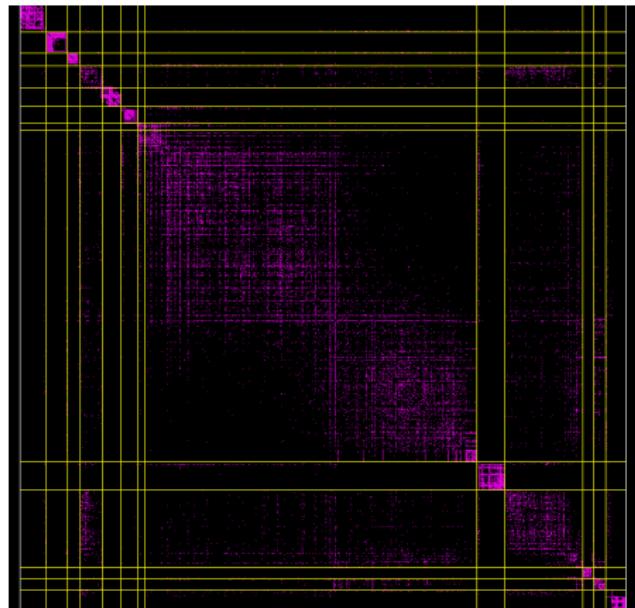


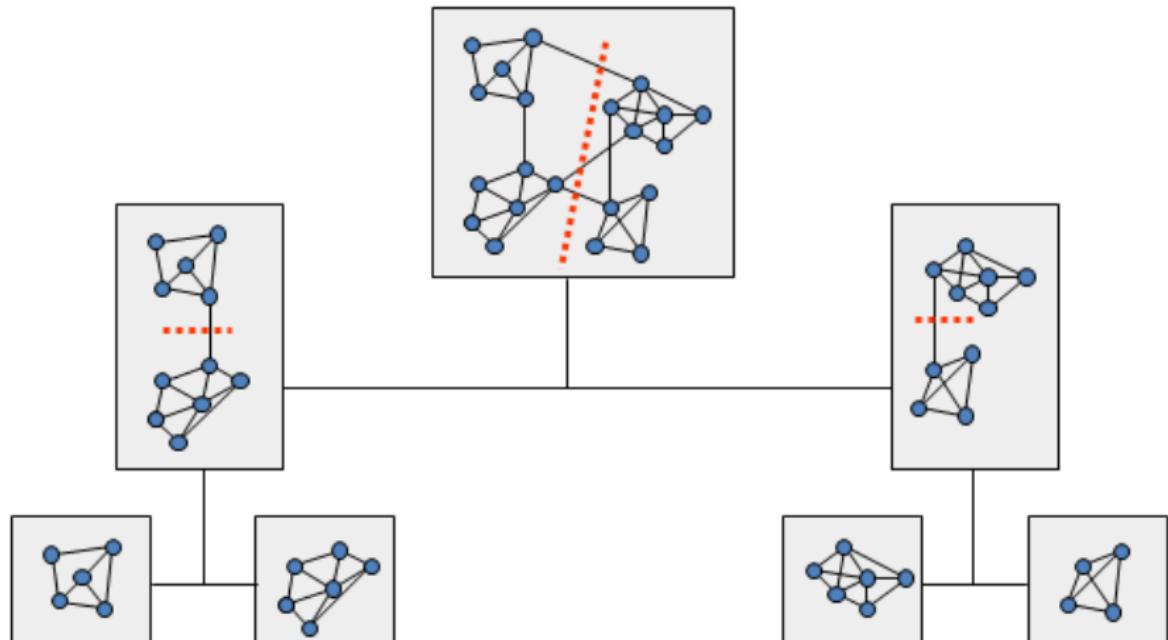
Рис.:  $cut = \frac{1}{4} \sum_{i>j} (p_i - p_j)^2 \omega_{ij}$

# Кластеризация

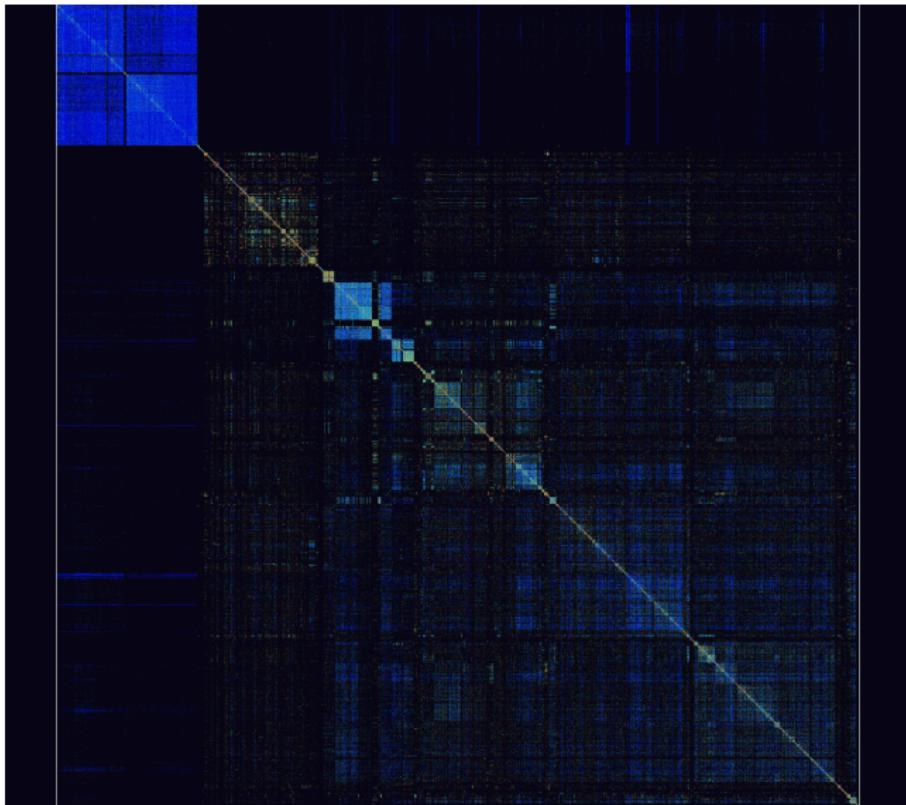


(c) 10-core, 819 nodes

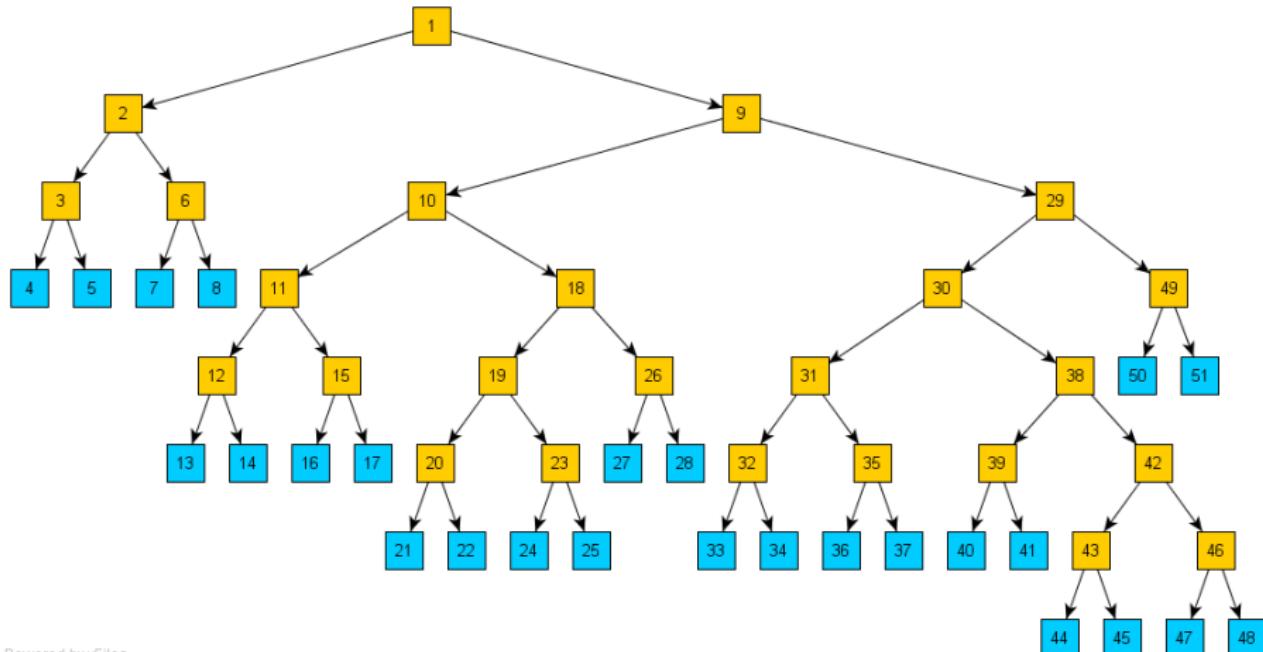
## Рекурсивное дерево



# Спектральная сортировка ядра Flickr



# Иерархическая кластеризация (таксономия)



Powered by yFiles

## Поисковая реклама

**YAHOO! SEARCH** epson

**My Web** [Beta](#)

[Web](#) | [Images](#) | [Video](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#)

[Subscriptions \(New\)](#) | [Shortcuts](#) | [Advanced Search](#) | [Preferences](#)

**Search Results** Results 1 - 10 of about 28,800,000 for **epson** - 0.02 sec. ([About this page](#))

Also try: [epson printers](#), [epson driver](#), [epson p-2000](#), [epson scanners](#) More...

**Epson Batteries and Chargers**  
- [eBatts](#)  
eBatts sells **Epson** batteries and chargers. Batteries and chargers for...  
[www.ebatts.com](#)

**Save on Epson Inks**  
Buy 2 **Epson** printer ink cartridges, get 1 cartridge free. Free 2 day...  
[www.clickinks.com](#)

**Epson Inkjet Cartridges - Save**  
Save up to 80% on **Epson** inkjet cartridges. We offer a 100% money back...  
[www.printhead.com](#)

**Epson Multimedia Projectors**  
Shop with us to save money on multimedia projection equipment.  
[www.projectorsforsale.com](#)

**Epson Compatible Ink**

**1. Epson**  
Go with an industry favorite in business printers-Hewlett-Packard. Reliable, high-quality laser, color and all-in-one printers. Compare HP printers head-to-head with **Epson**.  
[www.hp.com](#)

**2. Epson Compatible Ink Cartridges - \$6.95**  
High-quality **Epson** compatible inkjet cartridges from [Inkjetcartridge.com](#). Toll-free sales and support. Yahoo five-star award-winning service.  
[www.inkjetcartridge.com](#)

**3. 75% off Epson Ink - Free Shipping**  
Up to 75% off **Epson** ink and toner, plus a one-year quality guarantee and free shipping. Save an extra 5% with coupon code over15. Free promotional items with every purchase.  
[www.123inkjets.com](#)

# Граф аукционных заявок

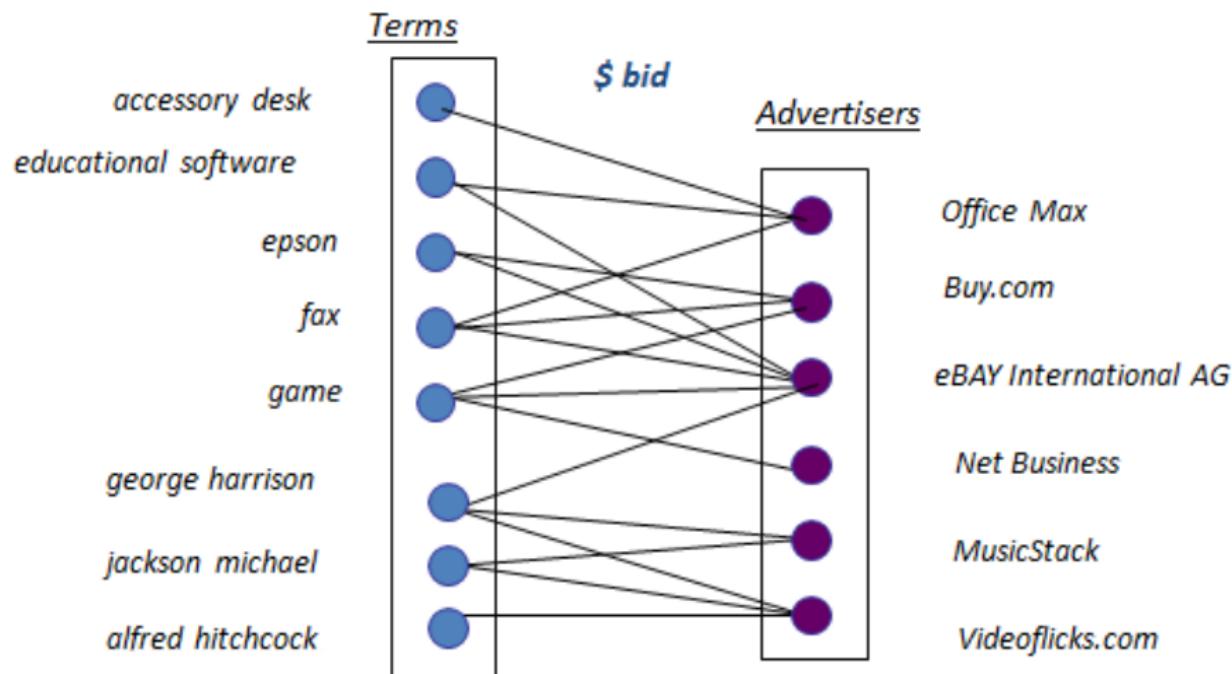
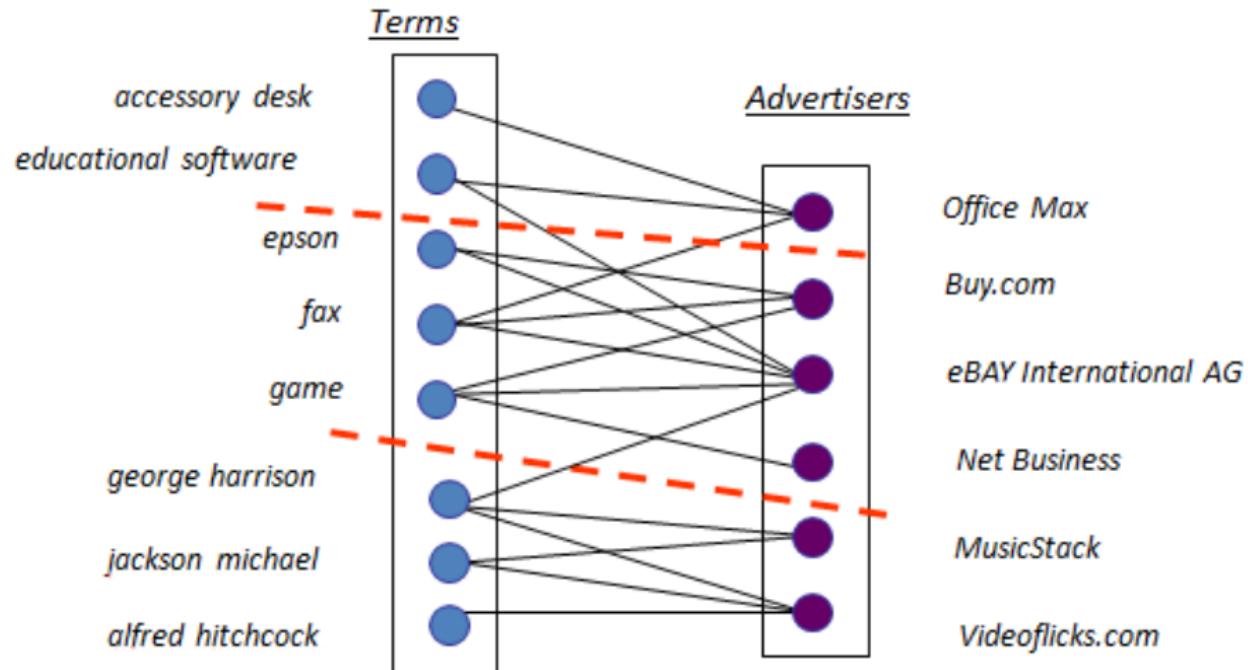


Рис.: Сегментация рынка

# Разделение графа



## Таблица аукционных заявок

	1	2	3	4	5	6	7	8
Office Max	■	■		■				
Net Business			■	■	■			
Buy.com			■	■	■			
eBAY	■	■	■	■	■	■		
MusicStack			■	■	■	■	■	
Videoflicks.com						■	■	■

① Accessory Desk

② Education Software

③ Epson

④ Fax

⑤ Game

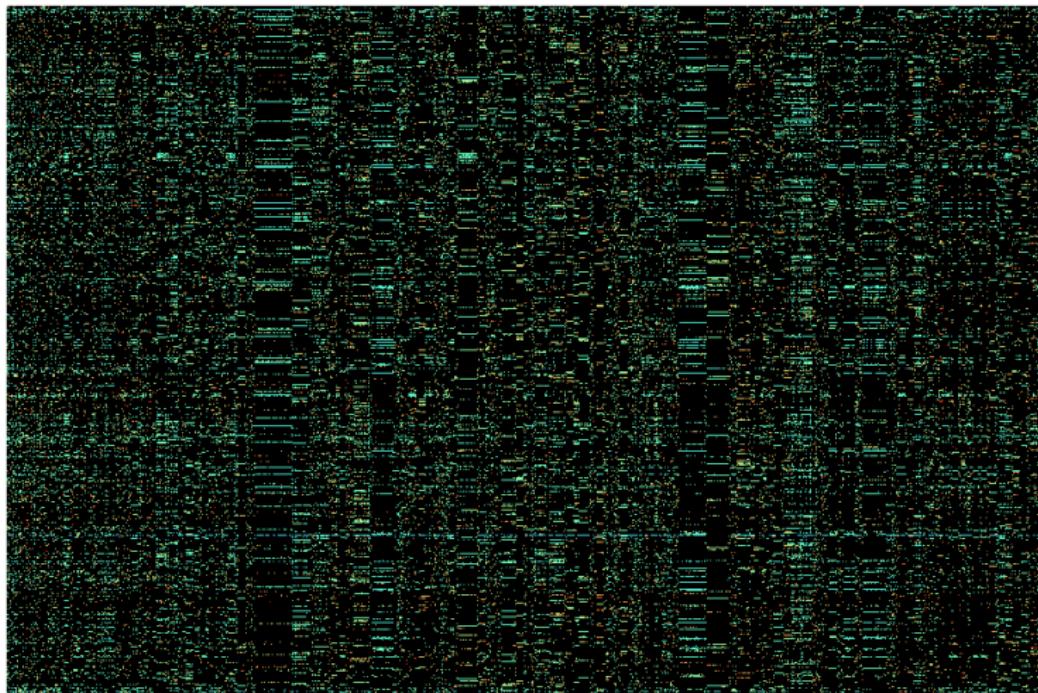
⑥ George Harrison

⑦ Jackson Michael

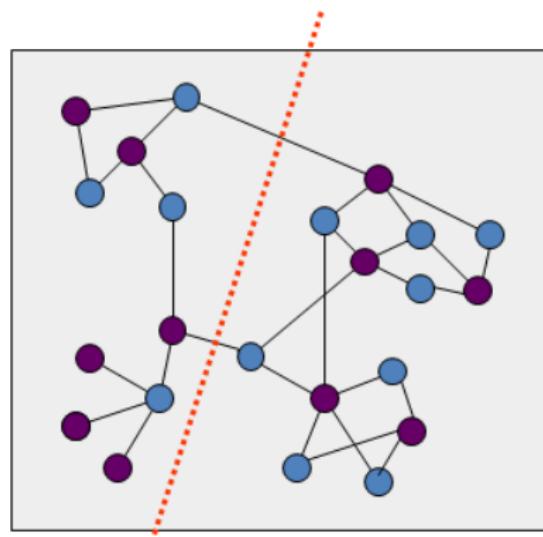
⑧ Alfred Hitchcock

# Таблица аукционных заявок

## Overture bid graph: real life



# Разделение двудольного графа Bi-partite graph partitioning



● Terms

● Advertisers

$$\widehat{A} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$$

# Формулировка для двудольного графа

- Уравнение для поиска собственных значений и векторов:

$$\begin{pmatrix} D_1 & -A \\ -A^T & D_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

- Сингулярное разложение матрицы (SVD):

$$A_n = D_1^{-\frac{1}{2}} A D_2^{-\frac{1}{2}}$$
$$A_n = u \sigma v^T, \sigma = 1 - \lambda$$

- Решение:  $x = D^{-\frac{1}{2}} u, y = D^{-\frac{1}{2}} v$

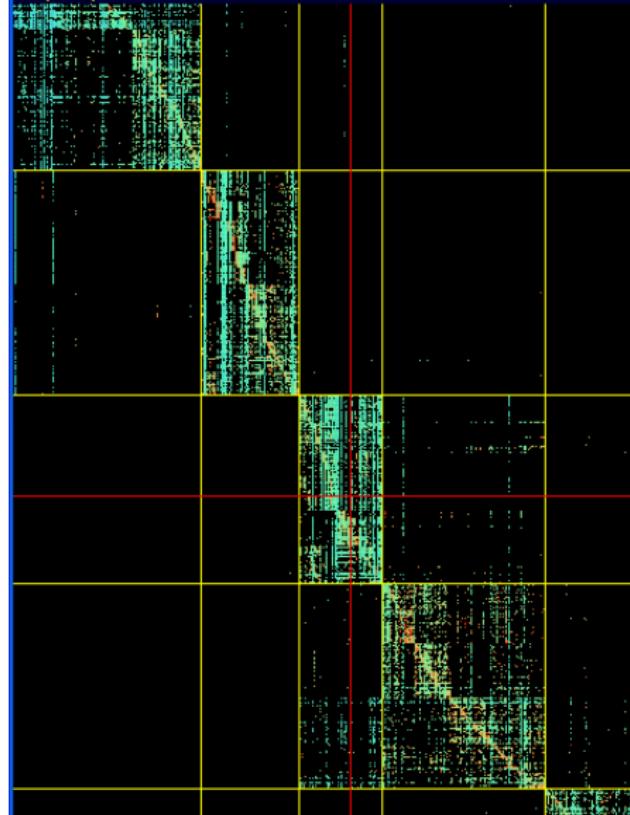
## Таблица аукционных заявок

This figure displays a 2D grid-based visualization of data distribution. The grid consists of numerous small squares arranged in a larger matrix. Most squares are either entirely black or contain a single point. Several distinct clusters of points are present, primarily in the upper-left, central, and lower-right regions. These clusters are composed of points in shades of red, orange, and yellow, suggesting higher values or densities in those areas.

# Примеры кластеров

row [1455]: diet pill tenuate

col [1277]: 3765ca18556



# Примеры кластеров

'A.S.C. Incorporated'  
'Atlantic Telecom'  
'AudioLink'  
'Cost Plus Electronics'  
'Headset Express'  
'Hello Direct'  
'PDA Mountain'  
'PK TECH INC'

.....

'accessory phone'  
'cordless headset'  
'cordless headset phone'  
'free hands'  
'headset'  
'headset microphone'  
'headset phone'  
'headset plantronics'  
'headset wireless'  
'isdn'  
'plantronics'

.....

'Berent Associates Center for  
Shyness and Social Therapy'  
'Dr. Puff'  
'New York Psychotherapy  
Collective'  
'Ruby Shoes'  
'Self-employed psychologist'  
'www.kabbalah.com'

.....

'health mental'  
'help self'  
'improvement self'  
'parenting'

# Интернет-радио Yahoo! Music — LAUNCHcast radio



Рис.: Плеер

Онлайн-рекомендация

My Station > Edit > Artists	
ARTISTS	1 - 20 (of 46 artists) YOUR RATING <a href="#">Why rate?</a>
Gipsy Kings	0 ★★★★
Tito Nieves	0 ★★★★
Ricardo Montaner	0 ★★★★
Eddie Palmieri	0 ★★★★
Tito Puente	0 ★★★★
Sade	0 ★★★★
Santana	0 ★★★★
Vanessa-Mae	0 ★★★★
Ruben Blades	0 ★★★★
Mary J. Blige	0 ★★★★

Рис.: Профиль пользователя

# Данные интернет-радио

<i>user ID</i>	<i>artist ID</i>	<i>ratings</i>
190389	1034047	100
190389	1034060	20
190389	1034129	50
190389	1034276	100
190389	1034388	80
190389	1034801	100
190389	1034831	100
190389	1034882	0
190389	1034883	40
190389	1035010	20
190389	1035473	60
190389	1035840	0
190389	1035926	30
190389	1036157	30
190389	1036454	60
190389	1036736	30
190389	1036737	100
190389	1036740	70
190389	1036746	80
190389	1036762	100

<i>artist ID</i>	<i>artist</i>
1037729	Coral
1037730	Rick Braun
1037731	Britney Spears
1037732	Gary Motley
1037733	Candido Fabre
1037734	K'Stalia Y Los Salchichas
1037735	Donny Hoffa
1037736	Rafael Mendez
1037737	Lithops
1037738	The Paris Ensemble
1037739	The Sunset Orchestra
1037740	The Mariachis Of Chiapas
1037741	Jenny Simpson
1037742	Doc West & The Yard Dogs
1037743	Geoff Bartley
1037744	Twilight Circus Dub Sound...
1037745	Tekneek

# Векторное представление данных

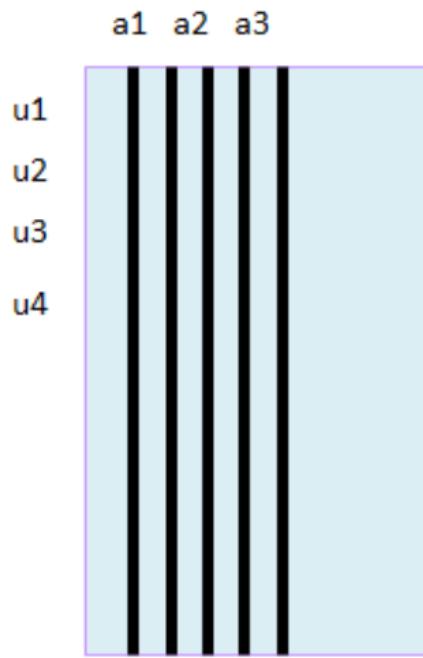


Рис.: ratings: users — artists

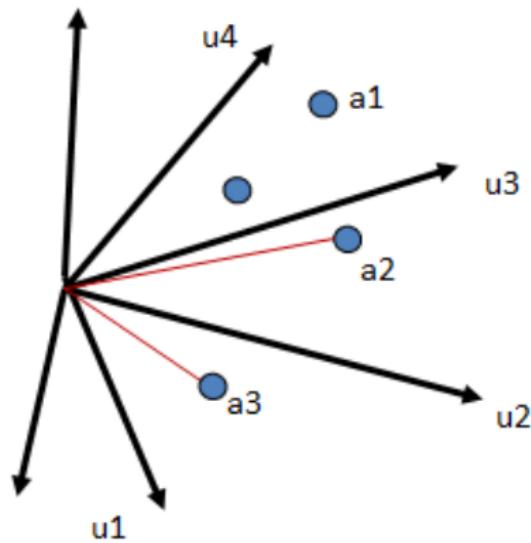
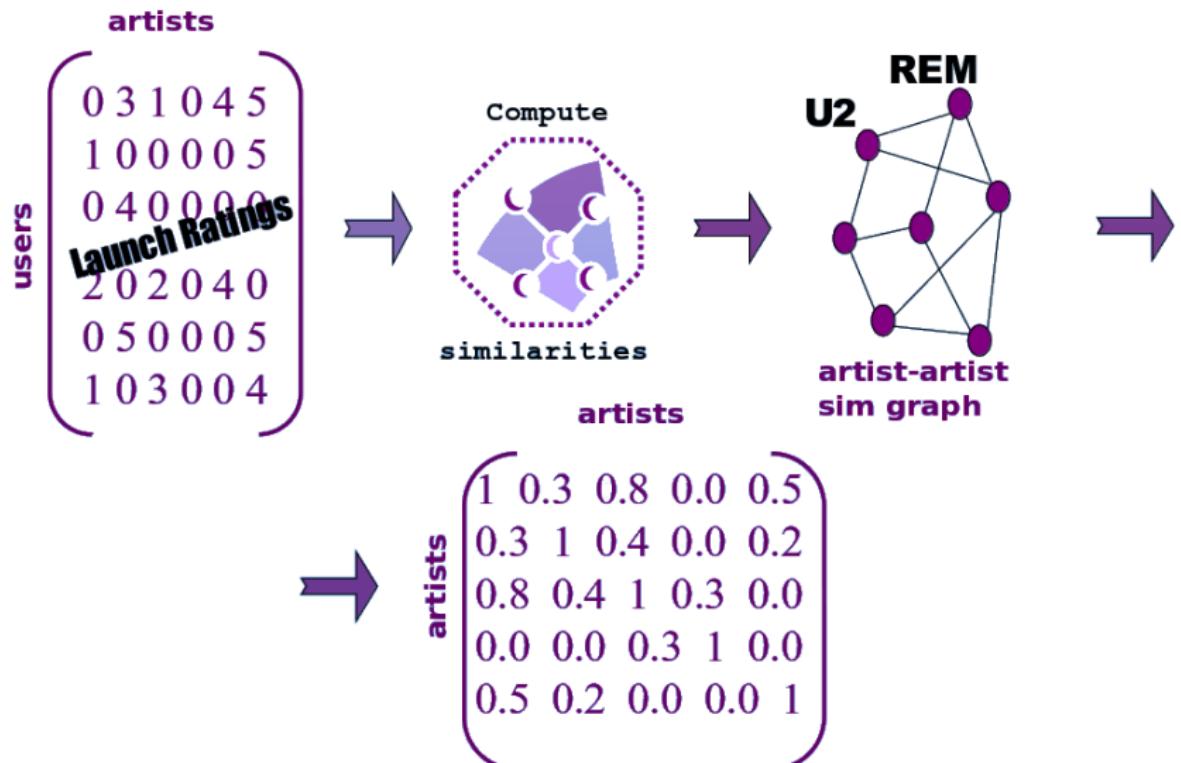
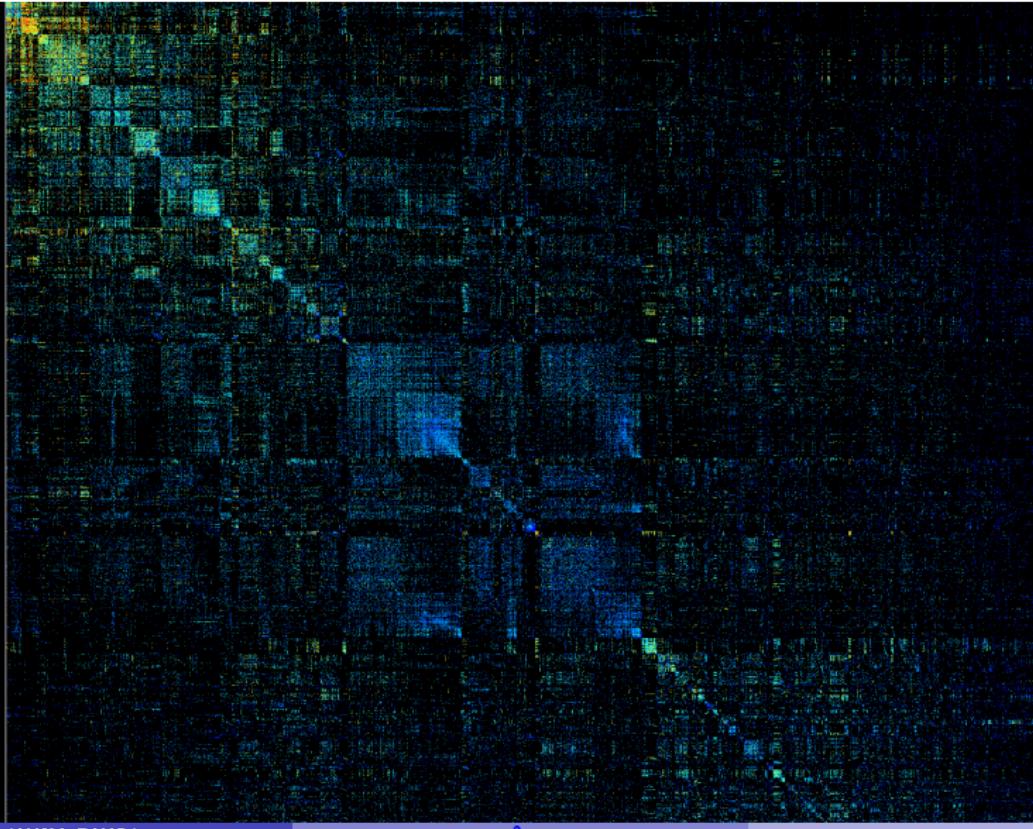


Рис.: Similarity in vector space model:  
 $\omega_{ij} = \cos(a_i, a_j) = \frac{a_i \cdot a_j}{\|a_i\| \cdot \|a_j\|}$

# Граф подобия

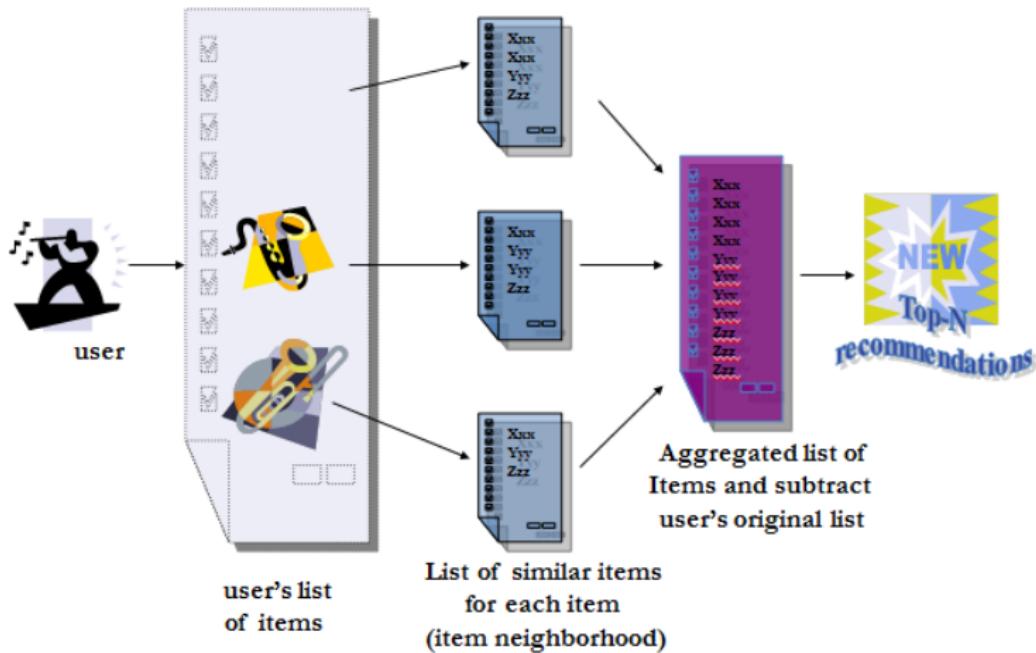


Матрица подобия, исполнитель – исполнитель  
Yahoo! Music, artist-artist

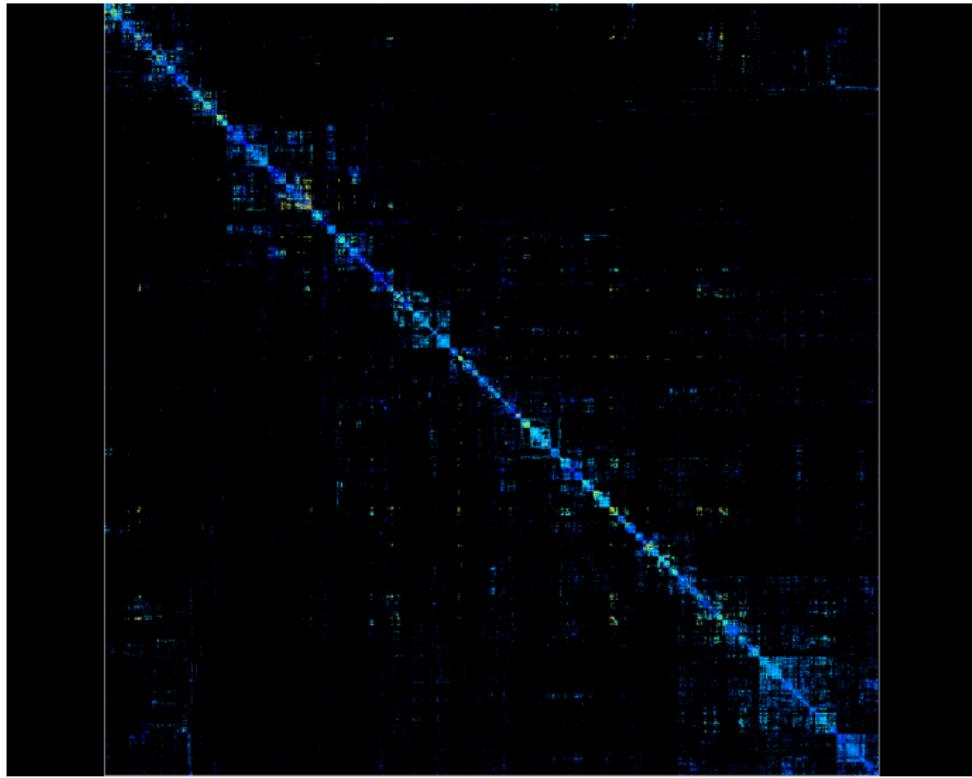


# Рекомендательная система

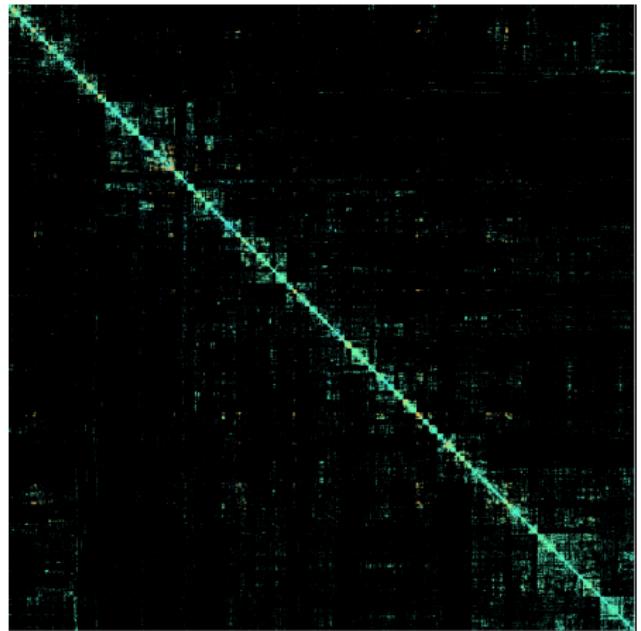
## Item based collaborative filtering



# Матрица подобия "исполнитель-исполнитель" Yahoo! Music, artist-artist



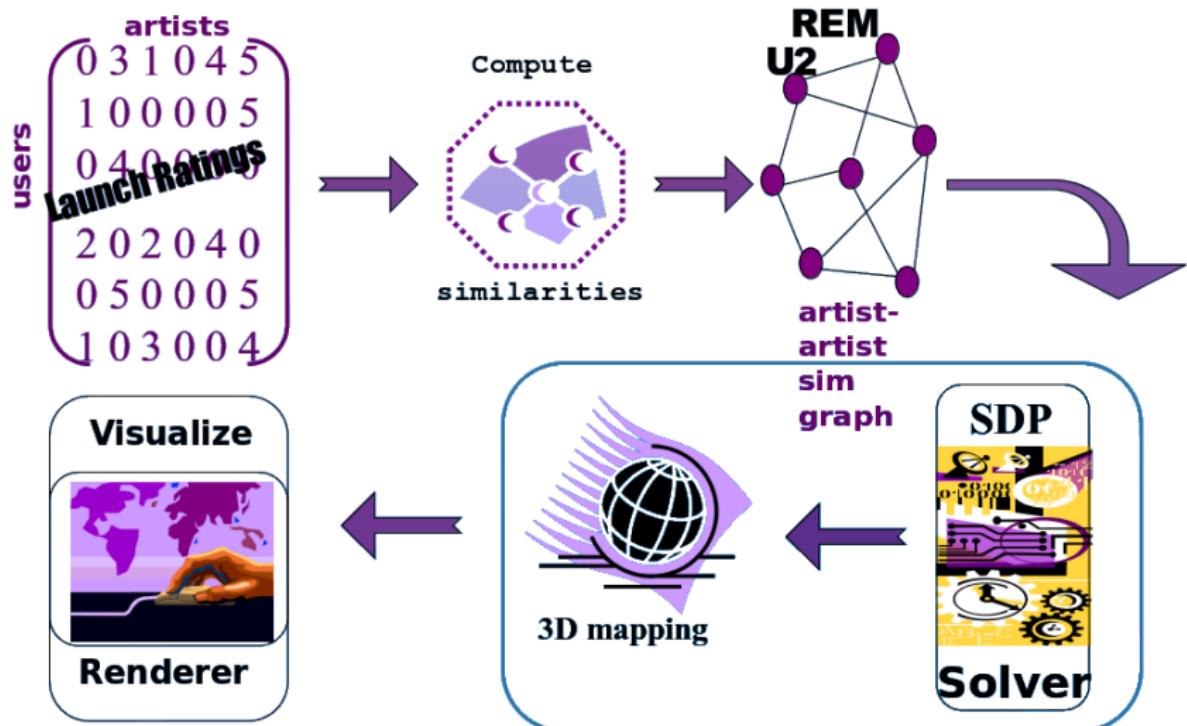
# Кластеры исполнителей



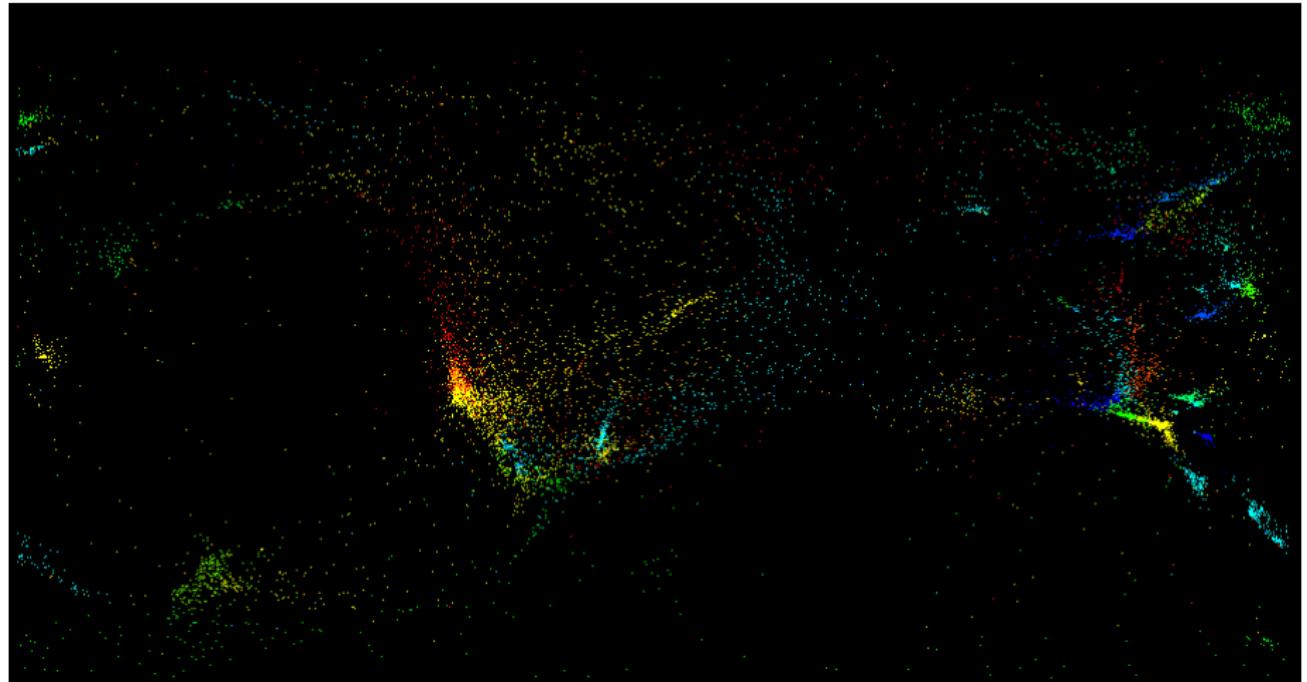
## Пример двух кластеров исполнителей

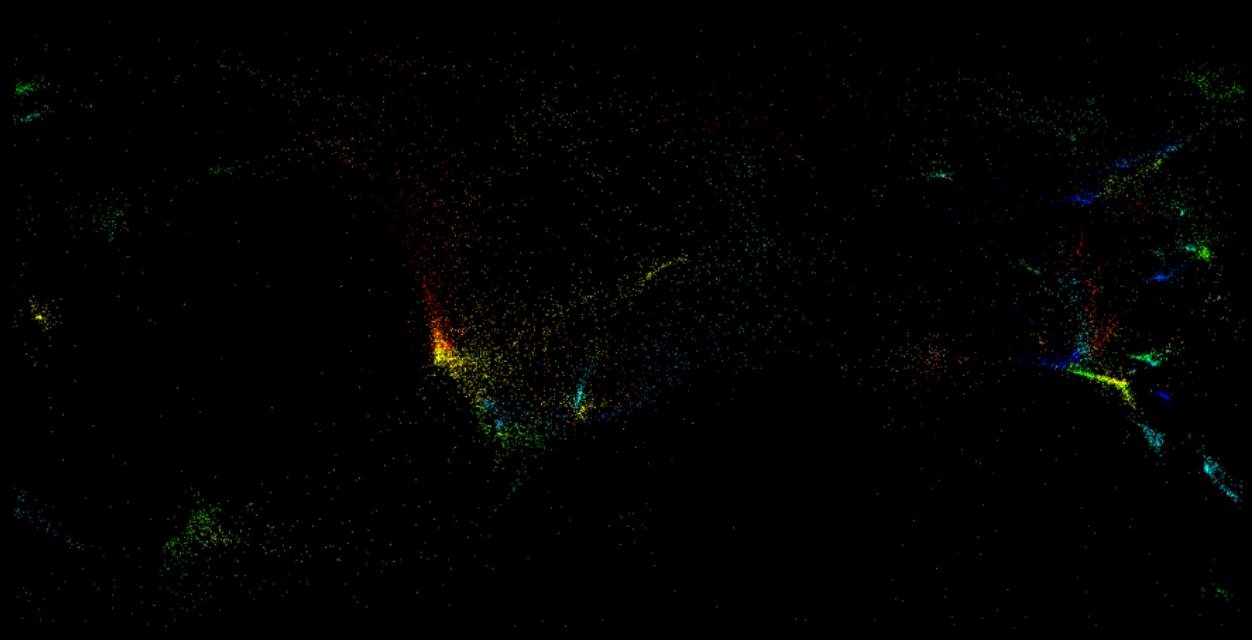
Rank	Label	Rank	Label
1	Mozart	1	Enrique Iglesias
2	Royal Philharmonic Orchestra	2	Avril Lavigne
3	Yo-Yo Ma	3	Backstreet Boys
4	Scottish Chamber Orchestra	4	Kylie Minogue
5	New York Philharmonic	5	Good Charlotte
6	Gilbert Johnson	6	Simple Plan
7	New York Pro Music Antiqua	7	T.A.T.U.
8	New Philharmonia Orchestra	8	Hilary Duff
9	Izzy	9	Paula Abdul
10	London Symphony Orchestra	10	No Doubt
11	Philadelphia Orchestra	11	Britney Spears
12	Roberto Michelucci	12	Westlife
13	Columbia Symphony Orchestra	13	O-Town
14	San Francisco Symphony	14	LFO (Lyte Funky Ones)
15	London Philharmonic Orchestra	15	98 Degrees
16	Berlin Philharmonic Orchestra	16	Shakira

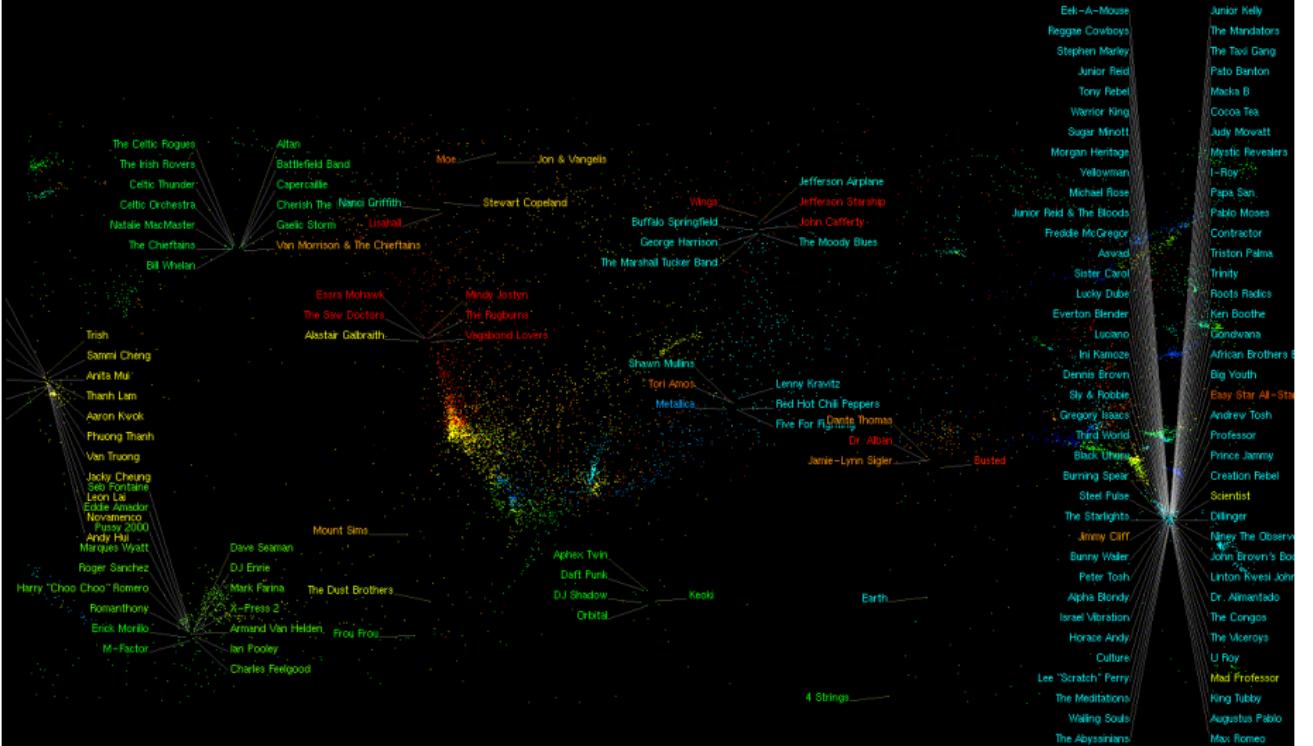
# Визуализация



# Двумерное представление графа

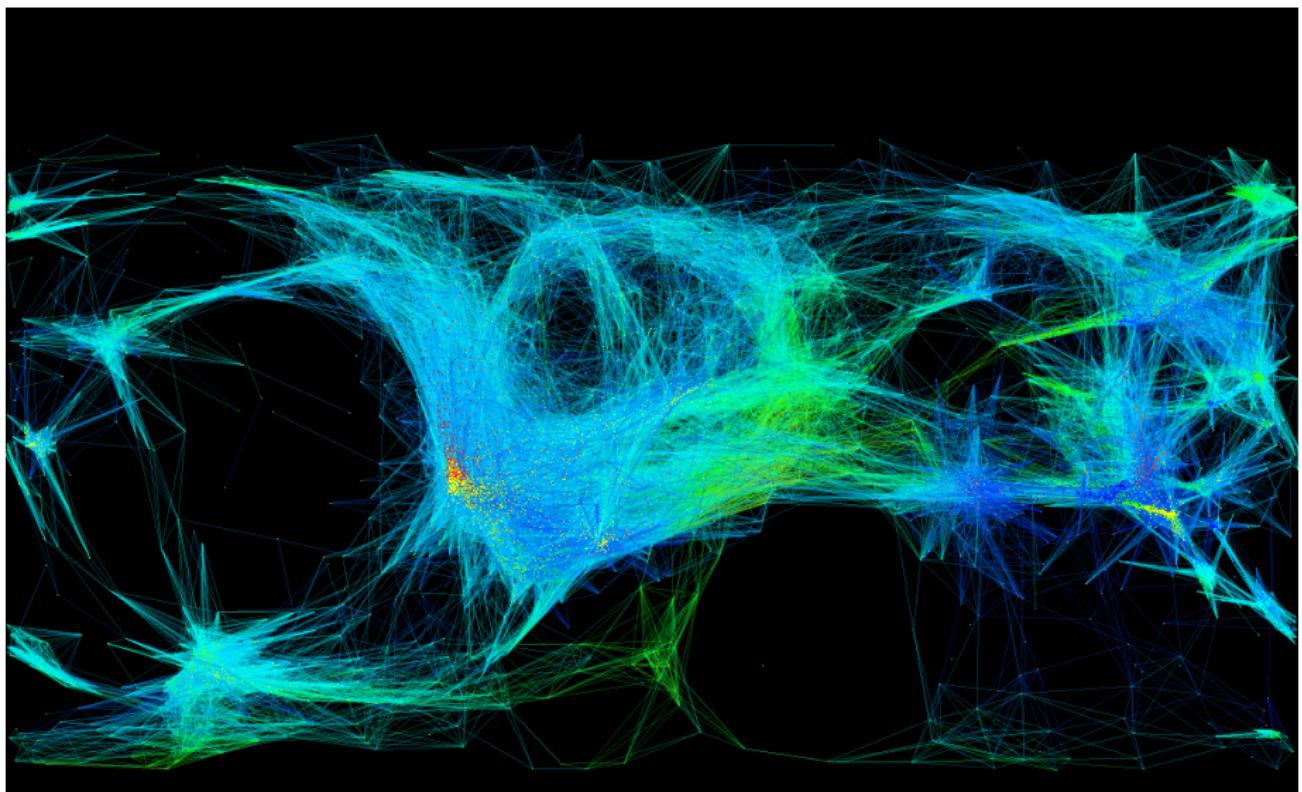






# Кластеры исполнителей





Arist name:

