

# Лекция 7

## Алгоритмы на графах. Инструменты для анализа сетей

Анализ и разработка алгоритмов



УНИВЕРСИТЕТ ИТМО

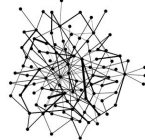
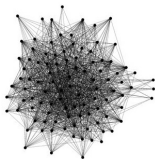
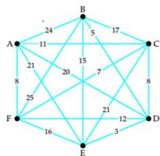
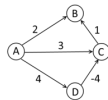
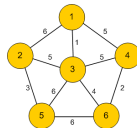
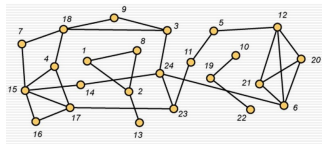
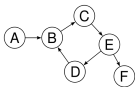
- 1 Основные и степенные меры
- 2 Меры расстояния
- 3 Мера плотности
- 4 Модулярность
- 5 Анализ сетей с помощью Gephi

# Постановка задачи

Как различать сети (графы)? Какие меры использовать для анализа сетей?

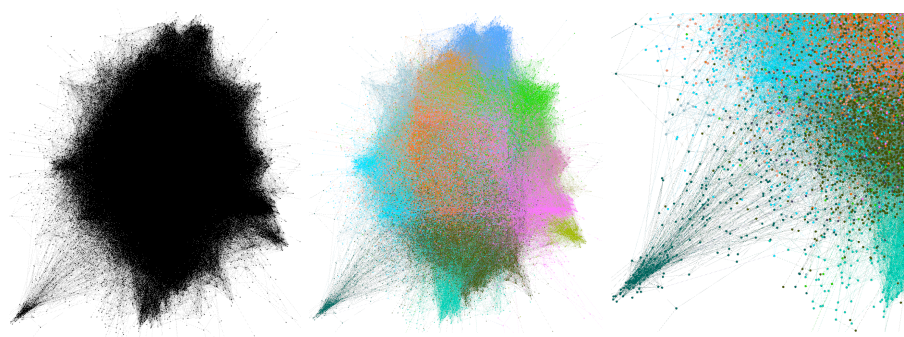
Какие графы называют

- не/ориентированными?
- не/взвешенными?
- не/полными?
- не/связными?
- плотными/разреженными?



# Сеть подписчиков группы банка (сеть банка)

Сеть подписчиков группы банка во ВКонтакте, представленная в виде невзвешенного неориентированного связного графа с 15,923 узлами (вершинами) и 200,633 связями (ребрами)



**Вопрос:** что должен включать анализ этой сети?

# Основные и степенные меры

## Основные меры:

$|V|$  — число вершин

$|E|$  — число ребер

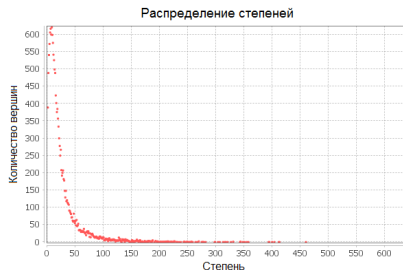
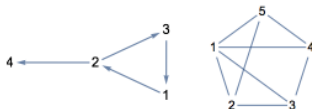
## Степенные меры:

$d(v)$ , **степень**  $v$ , — число входящих и исходящих ребер вершины  $v$

$d_{in}(v)$ , **полустепень захода**  $v$ , — число входящих ребер вершины  $v$

$d_{out}(v)$ , **полустепень исхода**  $v$ , — число исходящих ребер вершины  $v$

$\bar{d} = \frac{1}{|V|} \sum_{v \in V} d(v)$  — **средняя степень вершин**



## Значения для сети банка:

$|V| = 15,923$

$|E| = 200,633$

$\bar{d} = 25.20$

Большая часть вершин: малые  $d$

Малая часть вершин (*хабы*): большие  $d$

Гипотезы о распределении  $d$ ?

# Меры расстояния

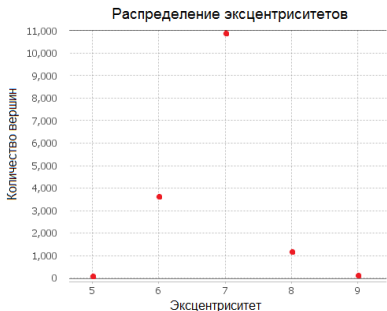
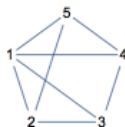
$\text{dist}(v, u)$  — расстояние (длина кратчайшего пути) между  $v$  и  $u$  ( $G$  — связный)

**Эксцентриситет**  $\epsilon(v) = \max_{u \in V} \text{dist}(v, u)$  — наибольшее расстояние между  $v$  и другими вершинами

**Радиус**  $r = \min_{v \in V} \epsilon(v)$  — минимальный эксцентриситет по всем вершинам

**Диаметр**  $D = \max_{v \in V} \epsilon(v)$  — максимальный эксцентриситет по всем вершинам, т.е. max расстояние между парой вершин

**Средняя длина пути**  $\ell = \frac{1}{|V| \cdot (|V|-1)} \sum_{v \neq u} \text{dist}(v, u)$   
 (“эффективность передачи информации по сети”)



**Значения для сети банка:**

$$r = 5$$

$$D = 9$$

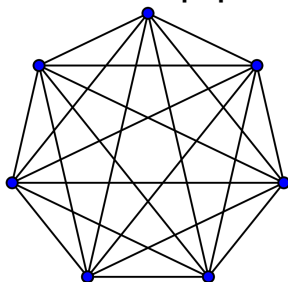
$$\ell = 3.48$$

# Мера плотности

**Плотность**  $\rho$  — частное  $|E|$  и числа возможных ребер с тем же  $|V|$ , т.е. числа ребер в полном графе с  $|V|$  вершинами:

$$\rho = \frac{2|E|}{|V|(|V| - 1)} \quad (\text{если } \rho \approx 0 \Rightarrow \text{граф разреженный})$$

Полный граф



$$|E| = \frac{|V|(|V|-1)}{2}$$

если  $|V| = 15,923$ , то  $|E| = 126,763,003$

**Значения для сети банка:**

$$|V| = 15,923$$

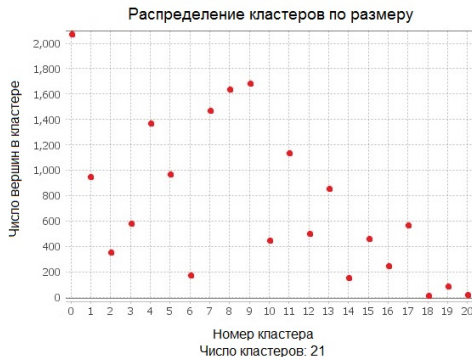
$$|E| = 200,633$$

$$\rho = 0.002$$

# Модулярность

**Модулярность  $Q$**  — мера разбиения графа на кластеры (подграфы, модули). Графы с высокой модулярностью  $Q > 0$  имеют плотные внутрикластерные связи и разреженные межкластерные связи.

$Q$  сравнивает количество ребер в кластерах исходного графа с количеством случайных ребер.



**Значение для сети банка:**  
 $Q = 0.463$

Далее  $G$  — неориентированный невзвешенный граф с матрицей смежности  $A$



# Случайное распределение ребер между вершинами

## Конфигурационная модель (КМ)

Для графа  $G$ , где каждая вершина  $v$  имеет степень  $d(v)$ , в КМ каждое ребро сначала разрезается на две части (каждая называется *обрубком*) и затем каждый обрубок случайно соединяется с другим обрубком в  $G$ . При этом  $d(v)$ -распределение сохраняется, но получается новый случайный граф  $\tilde{G}$ .

## Ожидаемое число ребер между $v, u \in \tilde{G}$

В графе  $\tilde{G}$  всего  $\sum_{w \in V} d(w) = 2|E|$  обрубков. Для  $i = 1, \dots, d(v)$  пусть  $I_i = 1$ , если  $i$ -й обрубок  $v$  соединен с одним из обрубков  $u$ , иначе  $I_i = 0$ . Поскольку  $i$ -й обрубок  $v$  может быть соединен с равной вероятностью с любым из  $2|E| - 1$  обрубков и поскольку  $u$  имеет  $d(u)$  обрубков,

$$\mathbb{E}[I_i] = \frac{d(u)}{2|E|-1}.$$

Между  $v$  и  $u$  всего  $J_{vu} = \sum_{i=1}^{d(v)} I_i$  ребер, так что

$$\mathbb{E}[J_{vu}] = \sum_{i=1}^{d(v)} \mathbb{E}[I_i] = \frac{d(v)d(u)}{2|E|-1} \approx \frac{d(v)d(u)}{2|E|} \quad (\text{для больших } |E|).$$

# Вычисление модулярности $Q$

Разность между числом ребер  $A_{uv}$  между  $v$  и  $u$  в исходном графе  $G$  (из матрицы  $A$ ) и ожидаемым числом ребер в случайном графе  $\tilde{G}$  равна

$$\Delta(u, v) := A_{vu} - \frac{d(v)d(u)}{2|E|}.$$

Пусть  $C$  — разбиение  $G$  на кластеры, а  $c(\cdot)$  обозначает кластер вершины  $\cdot$ .

- Если  $c(v) = c(u)$ , то  $Q$  должна возрасть при  $\Delta(u, v) > 0$  и убывать в противном случае.
- Если  $c(v) \neq c(u)$ , то  $Q$  не должна меняться.

После нормализации получаем:

Модулярность (Newman and Girvan, 2004; Newman, 2006)

$$Q(C) = \frac{1}{2|E|} \sum_{v, u \in V} \left( A_{vu} - \frac{d(v)d(u)}{2|E|} \right) \mathbb{1}(c(v) = c(u)),$$

$$Q = \max_C Q(C).$$

Для подсчета  $Q$  нужно найти разбиение  $C$ , доставляющее максимум  $Q(C)$ .  
Методы численной оптимизации в этой задаче: Fast Greedy, Louvain и др.


# Анализ сетей с помощью Gephi

“**Gephi** — ведущее ПО для визуализации и анализа всех видов графов и сетей. Gephi имеет открытый код и бесплатно.” — <https://gephi.org/>

- Выберите сеть из базы <https://snap.stanford.edu/data/>
- При необходимости измените формат данных на тот, с которым работает Gephi (.csv, .xls, etc.)

By Jure Leskovec

STANFORD UNIVERSITY



General Relativity and Quantum Cosmology collaboration network

Dataset Information

ArXiv GR-QC (General Relativity and Quantum Cosmology) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to General Relativity and Quantum Cosmology category. If an author  $i$  co-authored a paper with author  $j$ , the graph contains an undirected edge from  $i$  to  $j$ . If the paper is co-authored by  $k$  authors this generates a completely connected (sub)graph on  $k$  nodes.

The data covers papers in the period from January 1993 to April 2013 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its GR-QC section.

Dataset statistics	
Nodes	5242
Edges	14496
Nodes in largest WCC	4158 (0.793)
Edges in largest WCC	13428 (0.926)
Nodes in largest SCC	4158 (0.793)
Edges in largest SCC	13428 (0.926)
Average clustering coefficient	0.5296
Number of triangles	48260
Fraction of closed triangles	0.3619
Diameter (longest shortest path)	17
90-percentile effective diameter	7.6

Source (citation)

- J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

Files

File	Description
ca-GrQc.txt.gz	Collaboration network of ArXiv General Relativity category

SNAP for C++

SNAP for Python

SNAP Datasets

BIO2SNAP Datasets

What's new

People

Papers

Projects

Citing SNAP

Links

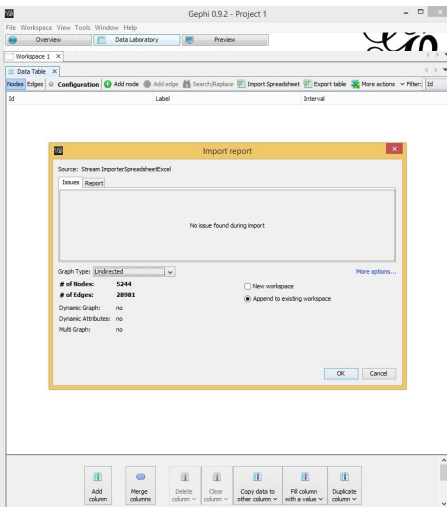
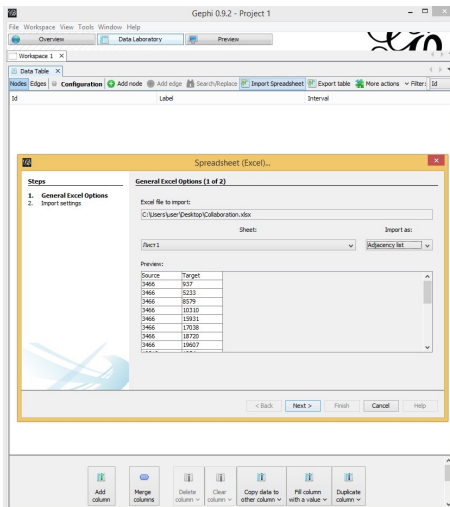
About

Contact us

Open positions

Open research positions in SNAP group are available at undergraduate, graduate and postdoctoral levels.

# Импорт данных



# Обработка данных

Gephi 0.9.2 - Project 1

File Workspace View Tools Window Help

Overview Data Laboratory Preview

Workspace 1 X

Data Table X

Nodes Edges Configuration Add node Add edge Search/Replace Import Spreadsheet Export table More actions Filter Source

Source	Target	Type	Id	Label	Interval	Weight
Source	Target	Undirected	0			1.0
3466	937	Undirected	1			1.0
3466	5233	Undirected	2			1.0
3466	8579	Undirected	3			1.0
3466	10310	Undirected	4			1.0
3466	15931	Undirected	5			1.0
3466	17038	Undirected	6			1.0
3466	18720	Undirected	7			1.0
3466	19607	Undirected	8			1.0
10310	1854	Undirected	9			1.0
10310	4983	Undirected	11			1.0
10310	5233	Undirected	12			1.0
10310	9572	Undirected	13			1.0
10310	10841	Undirected	14			1.0
10310	13056	Undirected	15			1.0
10310	14962	Undirected	16			1.0
10310	16310	Undirected	17			1.0
10310	19640	Undirected	18			1.0
10310	23855	Undirected	19			1.0
10310	24372	Undirected	20			1.0
10310	24814	Undirected	21			1.0
9052	899	Undirected	22			1.0
9052	1796	Undirected	23			1.0
9052	2287	Undirected	24			1.0
9052	3096	Undirected	25			1.0
9052	3386	Undirected	26			1.0
9052	4472	Undirected	27			1.0
9052	5346	Undirected	28			1.0
9052	5740	Undirected	29			1.0
9052	6094	Undirected	30			1.0
9052	6276	Undirected	31			1.0
9052	9124	Undirected	32			1.0
9052	10235	Undirected	33			1.0
9052	10427	Undirected	34			1.0
9052	10597	Undirected	35			1.0
9052	15159	Undirected	36			1.0

Add column Merge columns Delete column Clear column Copy data to other column Fill column with a value Duplicate column

Gephi 0.9.2 - Project 1

File Workspace View Tools Window Help

Overview Data Laboratory Preview

Workspace 1 X

Appearance X Graph X

Nodes Edges Unique Partition Ranking

Drapping (Configure)

Layout X

Choose a layout Run

<No Properties>

Projects... Reset

Context X

Nodes: 5244  
Edges: 14497  
Undirected Graph

Filters Statistics X

Settings

Network Overview

Average Degree Run

Avg. Weighted Degree Run

Network Diameter Run

Graph Density Run

HTTS Run

Modularity Run

PageRank Run

Connected Components Run

Girvan-Newman Clustering Run

Node Overview

Avg. Clustering Coefficient Run

Eigenvector Centrality Run

Edge Overview

Avg. Path Length Run

Dynamic

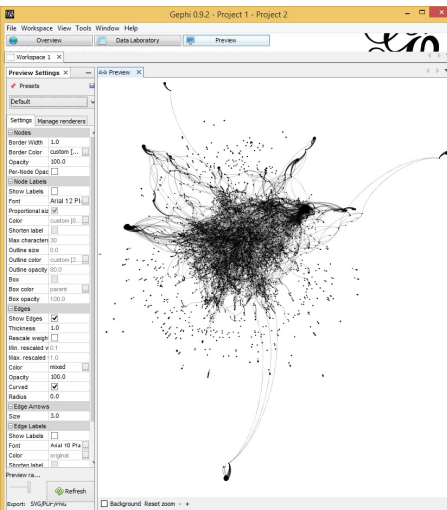
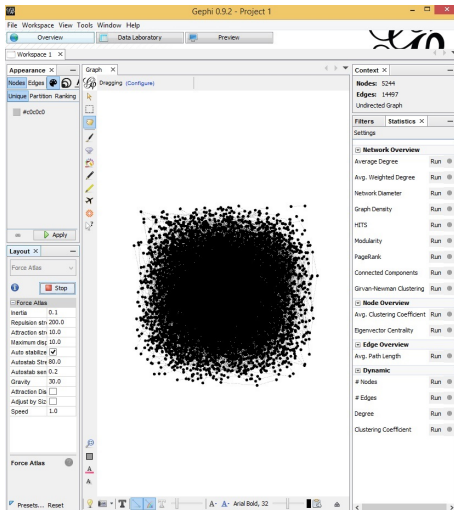
# Nodes Run

# Edges Run

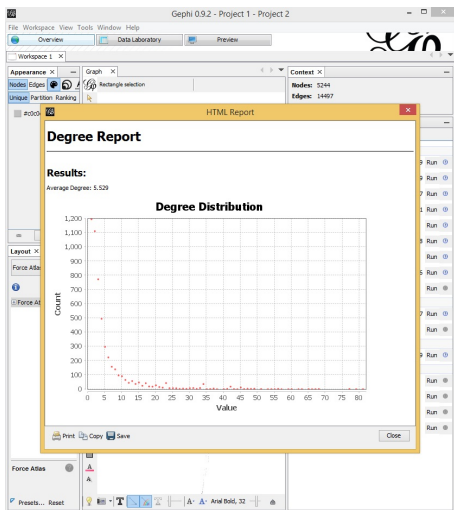
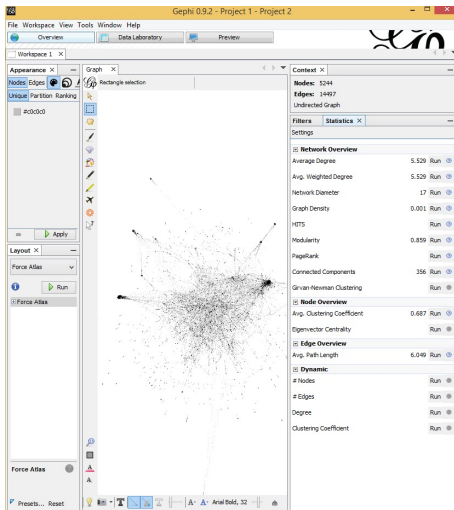
Degree Run

Clustering Coefficient Run

# Раскладка графа



# Подсчет мер



Спасибо за внимание!