Lecture 7
Algorithms on graphs. Tools for network analysis

Analysis and Development of Algorithms

УНИВЕРСИТЕТ ИТМО

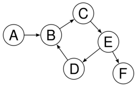
Overview

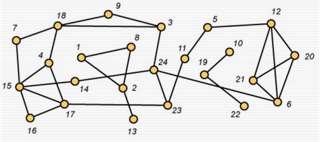
Problem statement

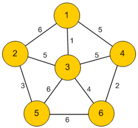
How to distinguish networks (graphs)? Which concepts and measures should be used for network analysis?

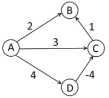
Recall the definition of

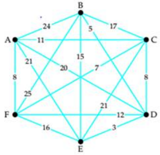
- directed and undirected graphs;
- weighted and unweighted graphs;
- complete and non-complete graphs;
- connected and disconnected graphs;
- dense and sparse graphs.

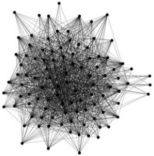


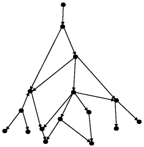






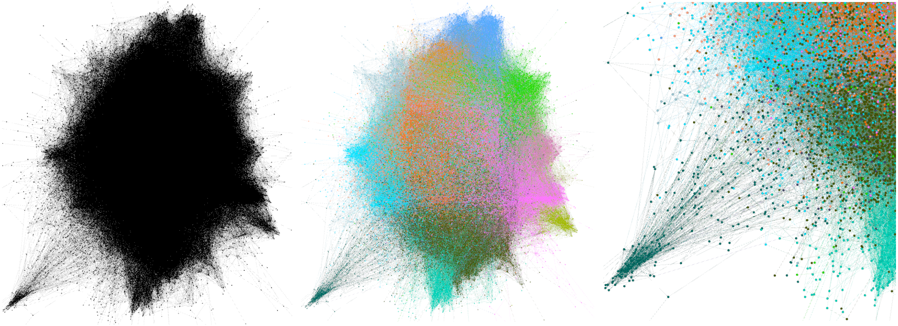






Bank group subscribers (BGS) network

A network of VK bank group subscribers represented by an unweighted undirected connected graph with 15,923 nodes and 200,633 edges



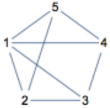
Question: How can we analyse this network?

Basic and degree measures

Basic measures:

$|V|$, the number of vertices

$|E|$, the number of edges



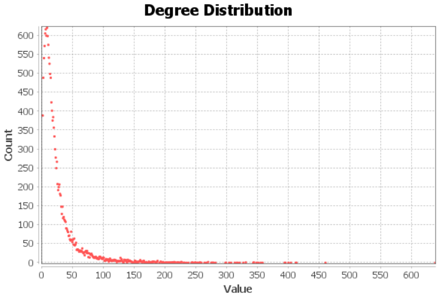
Degree measures:

$d(v)$, **degree of** v, i.e. the number of edges for vertex v

$d_{\text{in}}(v)$, **in-degree of** v, i.e. the number of in-edges for vertex v

$d_{\text{out}}(v)$, **out-degree of** v, i.e. the number of out-edges for vertex v

$\overline{d} = \frac{1}{|V|} \sum_{v \in V} d(v)$, **average degree** over all vertices



Values for BGS:

$|V| = 15,923$

$|E| = 200,633$

$\overline{d} = 25.20$

Large part: low d

Small part (aka *hubs*): high d

Statistical hypotheses about the distribution?

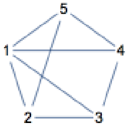
Distance measures

Given a connected G, $\text{dist}(v, u)$ is the distance (shortest path length) between v and u

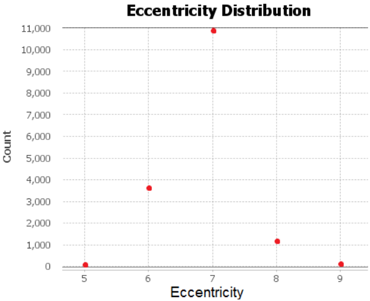
The **eccentricity** $\epsilon(v)$ of v is the greatest distance between v and any other vertex: $\epsilon(v) = \max_{u \in V} \text{dist}(v, u)$ ("how far a node is from the node most distant from it").

The **radius** r is the minimum eccentricity of any vertex: $r = \min_{v \in V} \epsilon(v) = \min_{v \in V} \max_{u \in V} \text{dist}(v, u)$.

The **diameter** D is the maximum eccentricity of any vertex, i.e. the greatest distance between any pair of vertices: $D = \max_{v \in V} \epsilon(v)$.



The **average path length** $\ell = \frac{1}{|V| \cdot (|V|-1)} \sum_{v \neq u} \text{dist}(v, u)$ ("the efficiency of information or mass transport on a network").



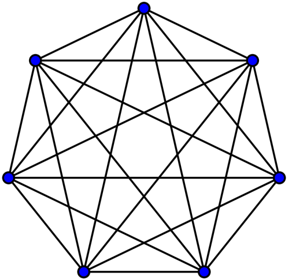
Values for BGS:
$r = 5$
$D = 9$
$\ell = 3.48$

Density measures

The **density** ρ of an undirected G is the ratio of $|E|$ and the number of possible edges, i.e. the number of edges in the complete graph with the same $|V|$:

$$\rho = \frac{2|E|}{|V|(|V|-1)} \qquad \text{(if } \rho \approx 0 \Rightarrow \text{graph is sparse)}$$

Complete graph



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

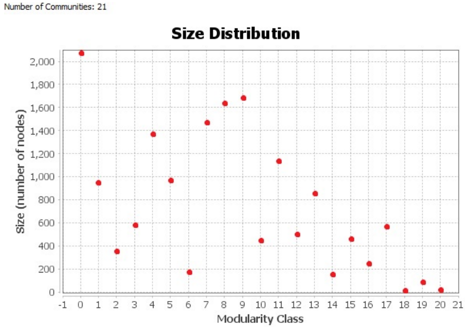
if $|V| = 15,923$, then $|E| = 126,763,003$

Values for BGS:
$|V| = 15,923$
$|E| = 200,633$
$\rho = 0.002$

Modularity

Modularity Q measures the strength of division of a graph into clusters (subgraphs, modules). Graphs with high $Q > 0$ have dense connections between the vertices within clusters but sparse between those in different clusters.

Q compares the number of edges within clusters in G with **the expected number of edges in a random graph** regardless of clusters.



Values for BGS:
$Q = 0.463$

For simplicity: an undirected unweighted graph G with the adjacency matrix A

Random distribution of edges between all vertices

Configuration model (CM)

For G with v having $d(v)$, CM cuts each edge into halves (each called a *stub*), and then each stub is rewired randomly with any other stub in G (except itself). Thus, $d(v)$-distribution remains the same but CM results in a new random $\tilde{G}$.

The expected number of edges between $v, u \in \tilde{G}$

The total number of stubs in $\tilde{G}$ is $\sum_{w \in V} d(w) = 2|E|$. For $i = 1, \ldots, d(v)$, let $I_i = 1$ if the i-th stub of v connects to one of stubs of u and $I_i = 0$, otherwise. Since the i-th stub of v can connect to any of the $2|E| - 1$ remaining stubs with equal probability and since there are $d(u)$ stubs of u,

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

The total number of edges between v and u is $J_{vu} = \sum_{i=1}^{d(v)} I_i$, so

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

Calculating Modularity Q

The difference between the actual number A_{vu} of edges between v and u (from the adjacency matrix A) and the expected number of edges between them is

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

Let C be a division of G into clusters and $c(v)$ denote the cluster of v.

- If $c(v) = c(u)$, Q should increase if $\Delta(u, v) > 0$ and decrease otherwise.
- If $c(v) \neq c(u)$, Q should not change.

Thus, after normalization,

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

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

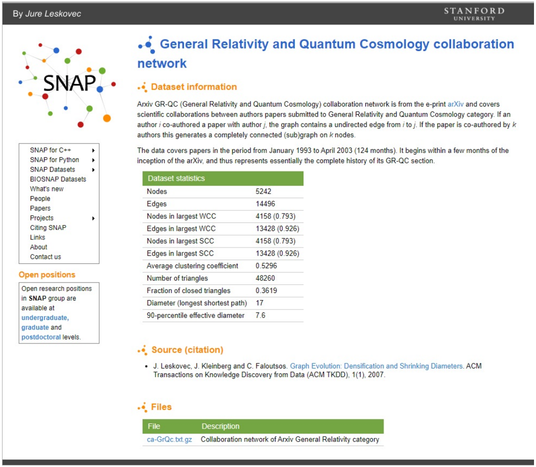
$$Q = \max_{C} Q(C).$$

To calculate Q, one has to find divisions C of G providing maximal $Q(C)$. It can be approximately done via numerical optimization (Fast Greedy, Louvain, etc.).

Network analysis with Gephi

“**Gephi** is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.” — https://gephi.org/

- Choose a graph from https://snap.stanford.edu/data/
- Change the format for that accepted by Gephi (.csv, .xls, etc.), if necessary



By *Jure Leskovec* STANFORD UNIVERSITY

SNAP

SNAP for C++
SNAP for Python
SNAP Datasets
BIOSNAP 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.

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 a 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 2003 (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

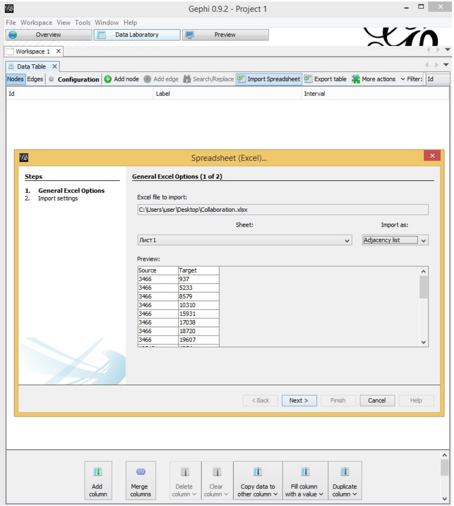
Source (citation)

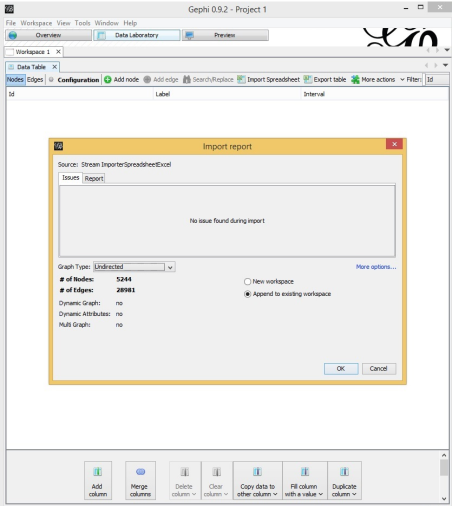
- J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

Files

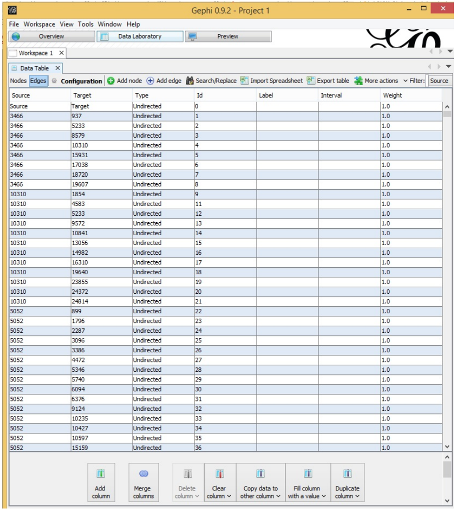
File	Description
ca-GrQc.txt.gz	Collaboration network of Arxiv General Relativity category

Import data

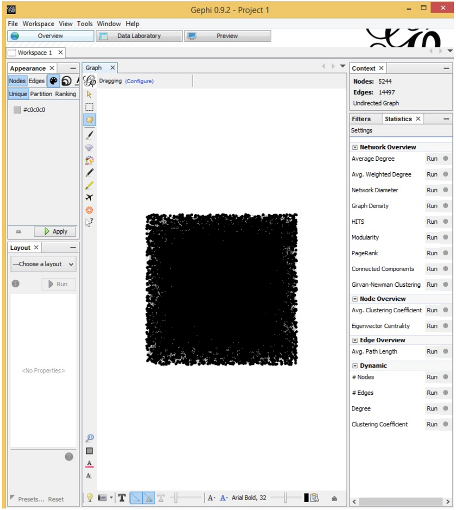

Gephi 0.9.2 - Project 1
Spreadsheet (Excel)...
Steps
1. General Excel Options
2. Import settings
General Excel Options (1 of 2)
Excel file to import:
C:\Users\user\Desktop\Collaboration.xlsx
Sheet:
Import as:
Лист1
Adjacency list
Preview:
< Back
Next >
Finish
Cancel
Help


Gephi 0.9.2 - Project 1
Import report
Source: Stream ImporterSpreadsheetExcel
No issue found during import
Graph Type: Undirected
# of Nodes: 5244
# of Edges: 28981
Dynamic Graph: no
Dynamic Attributes: no
Multi Graph: no
New workspace
Append to existing workspace
OK
Cancel

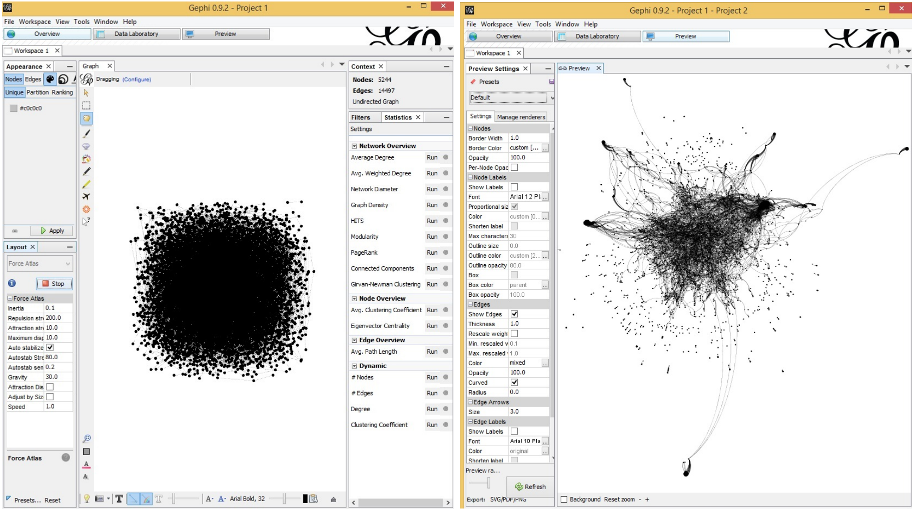
Process data



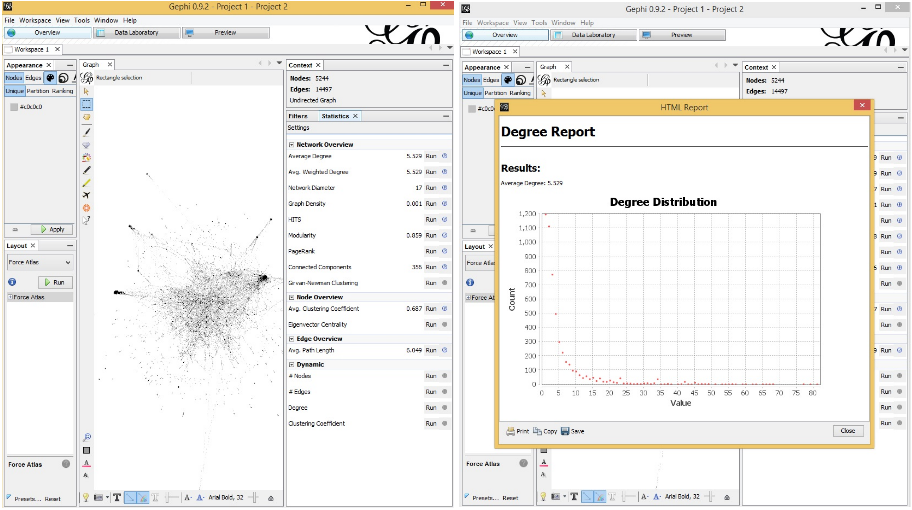
Source	Target	Type	Id	Label	Interval	Weight
Source	Target	Undirected	0			1.0
3466	937	Undirected	1			1.0
3466	5233	Undirected	2			1.0
3466	8579	Undirected	3			1.0
3466	10310	Undirected	4			1.0
3466	15931	Undirected	5			1.0
3466	17038	Undirected	6			1.0
3466	18720	Undirected	7			1.0
3466	19607	Undirected	8			1.0
10310	1854	Undirected	9			1.0
10310	4583	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	14982	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
5052	899	Undirected	22			1.0
5052	1796	Undirected	23			1.0
5052	2287	Undirected	24			1.0
5052	3096	Undirected	25			1.0
5052	3386	Undirected	26			1.0
5052	4472	Undirected	27			1.0
5052	5346	Undirected	28			1.0
5052	5740	Undirected	29			1.0
5052	6094	Undirected	30			1.0
5052	6376	Undirected	31			1.0
5052	9124	Undirected	32			1.0
5052	10235	Undirected	33			1.0
5052	10427	Undirected	34			1.0
5052	10597	Undirected	35			1.0
5052	15159	Undirected	36			1.0



Graph layout



Calculate measures



Thank you for your attention!