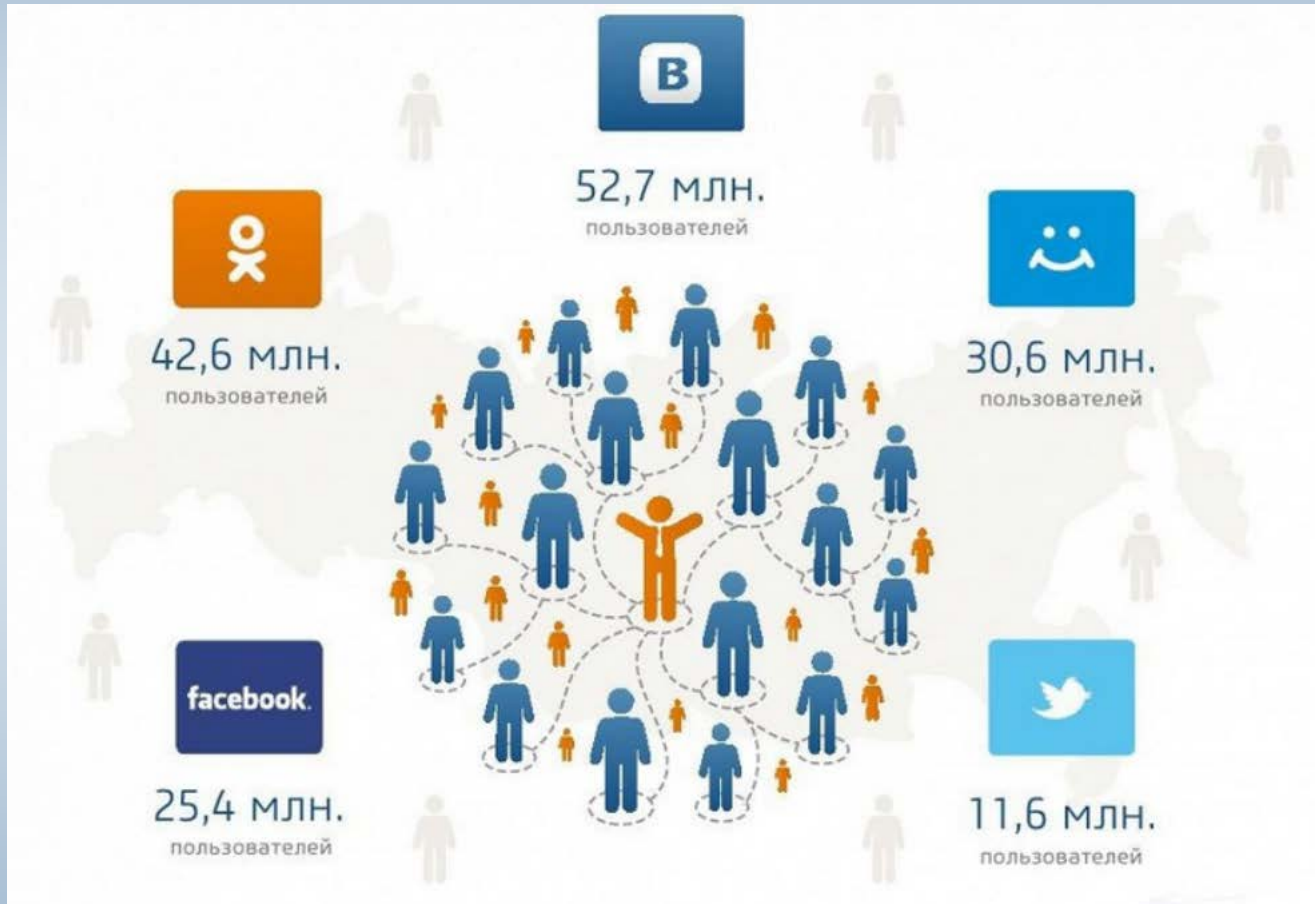


Network analysis



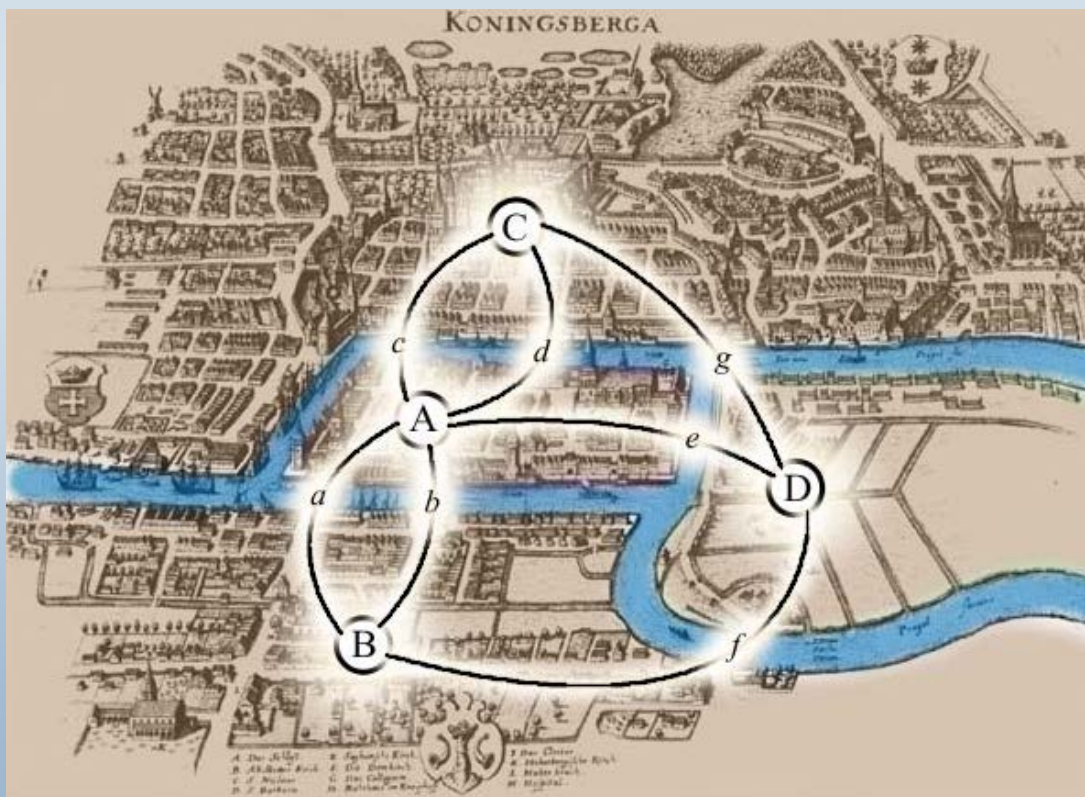
Кольцов С.Н

Исследование социальных сетей



Исследование сетей (Network analysis)

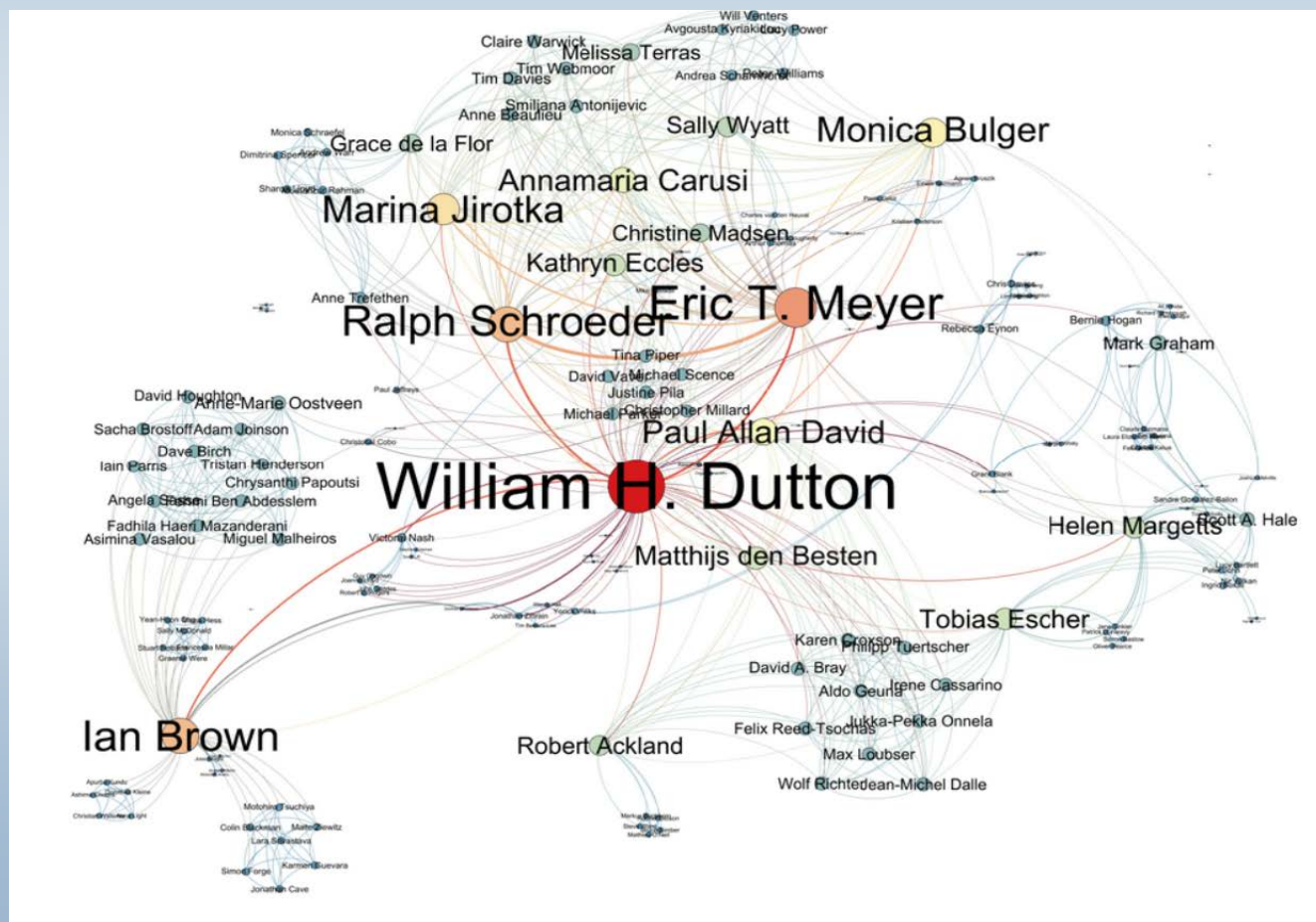
Теория графов — раздел дискретной математики, изучающий свойства графов. В общем смысле граф представляется как множество вершин (узлов), соединённых рёбрами.



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)

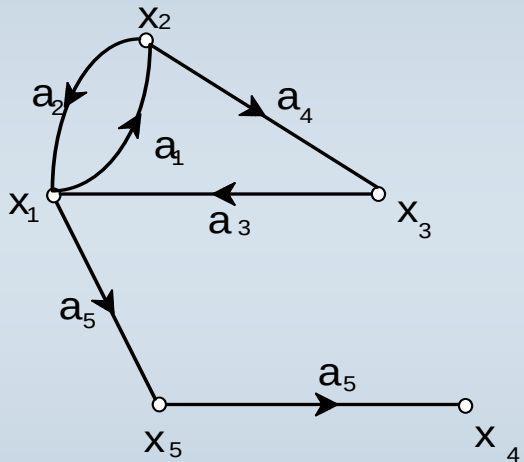
Теория шести рукопожатий — теория, согласно которой любые два человека на Земле разделены в среднем лишь пятью уровнями общих знакомых. Теория была выдвинута в 1969 году американскими психологами Стэнли Милгрэмом и Джеффри Трэверсом. Предложенная ими гипотеза заключалась в том, что каждый человек опосредованно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

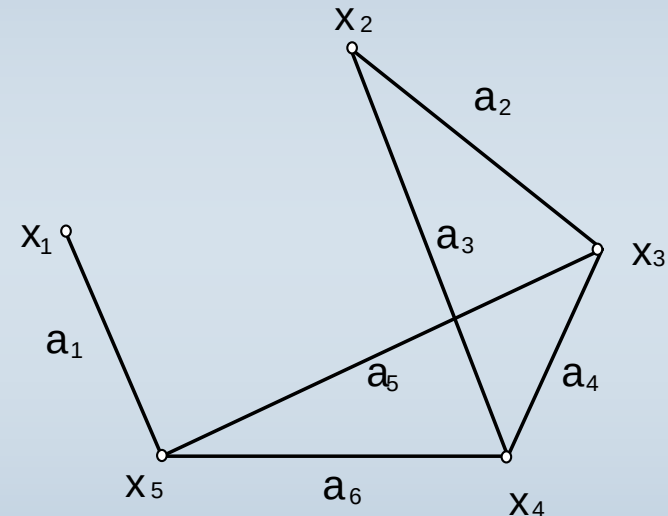
Миланский университет и социальная сеть Facebook установили, что двух любых пользователей Facebook отделяет 4,74 уровня связи. Для США количество звеньев составило 4,37.

Математическая основа теории графов

Ориентированный и не ориентированный виды графов



Ориентированный граф



Неориентированный граф

Примером неориентированного графа является карта дорог

Исследование социальных 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

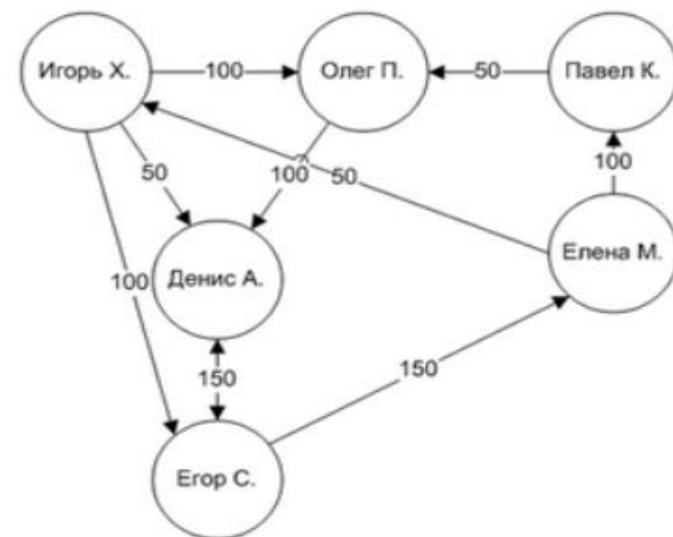
Прямоугольная матрица

	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 социальных сетей (Network analysis)

В простейшей диаграмме связей не учитывается ни направление, ни вес связи. В реальной жизни эти характеристики имеют критическое значение. Например, Сергей может знать Олега, но это не значит, что Олег знает Сергея. Или с партнером по бизнесу исследуемое лицо может иметь гораздо более тесные отношения, чем с клиентом своей компании. В этом случае матрица уже не будет симметричной относительно диагонали.

	Игорь	Олег	Павел	Денис	Елена	Егор
Игорь	0	100	0	50	0	100
Олег	0	0	0	100	0	0
Павел	0	50	0	0	0	0
Денис	0	0	0	0	0	150
Елена	50	0	100	0	0	0
Егор	0	0	0	150	150	0



Исследование социальных сетей

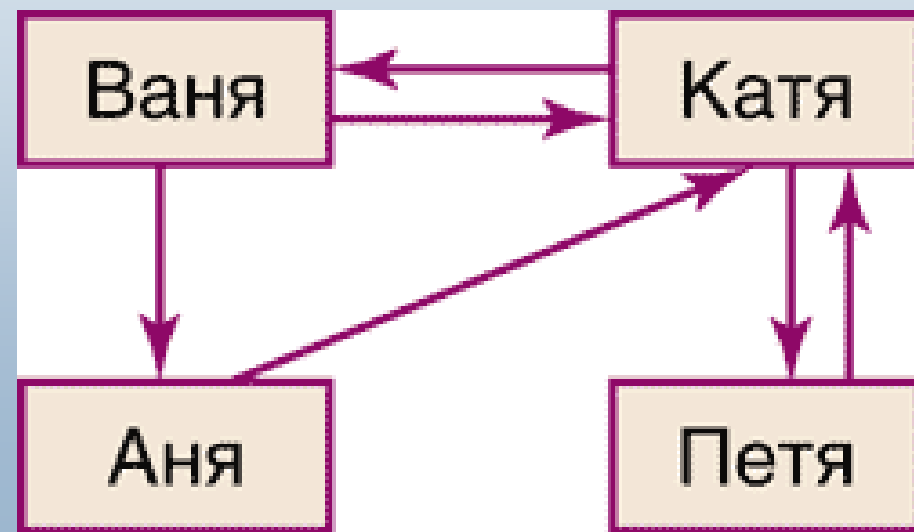
Параметры сети: Число вершин или ребер

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.

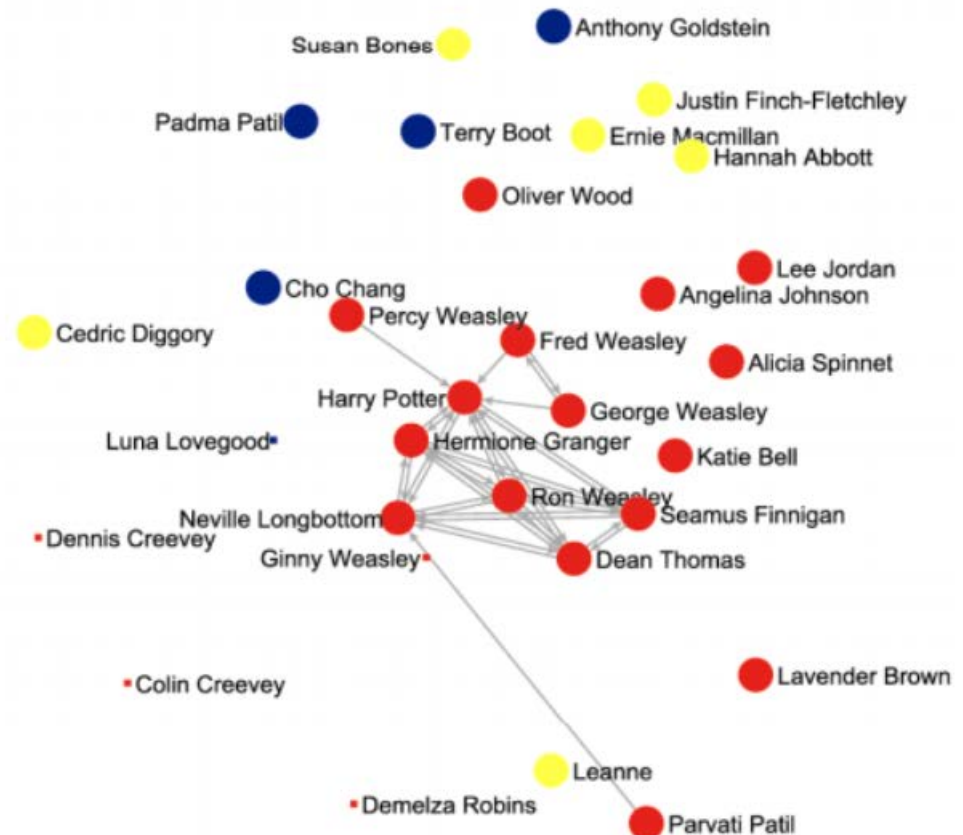
Диаметр социальной сети - параметр, который показывает, насколько велика сеть — это наибольшее геодезическое расстояние в социальной сети

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

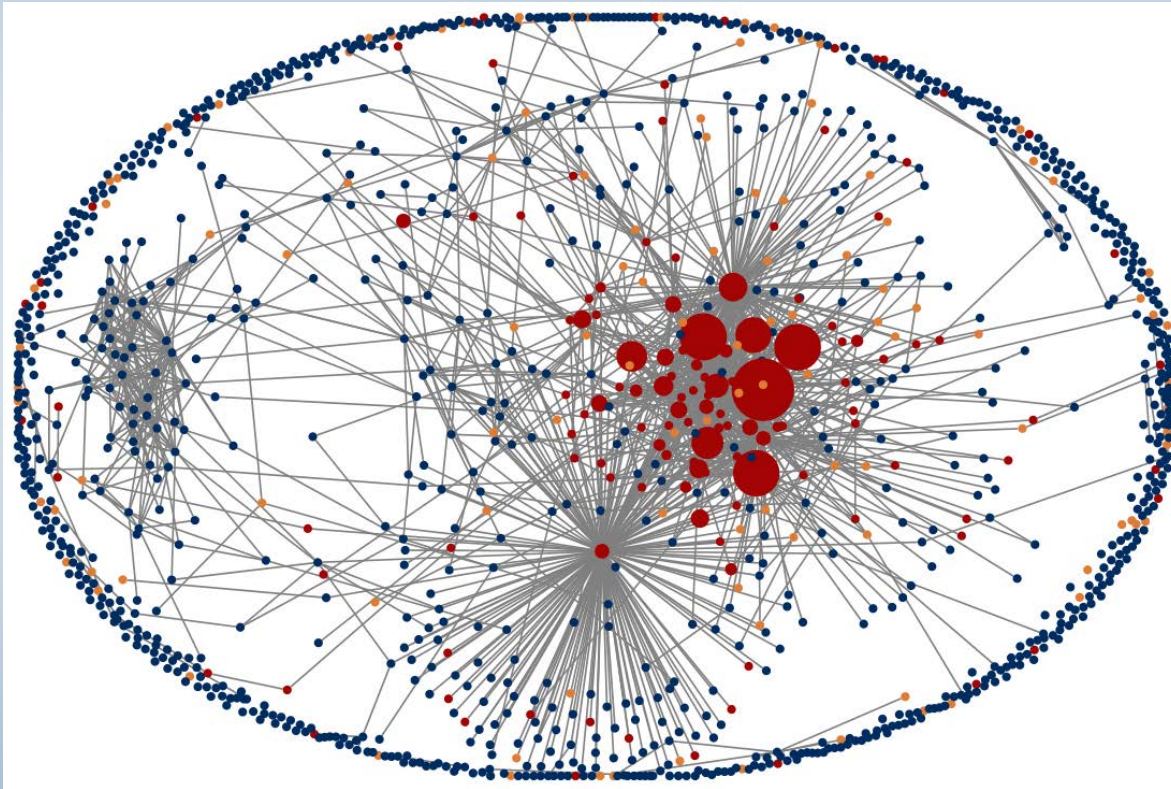


Исследование социальных сетей An Analysis of Friendship Networks in the Harry Potter

- (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.



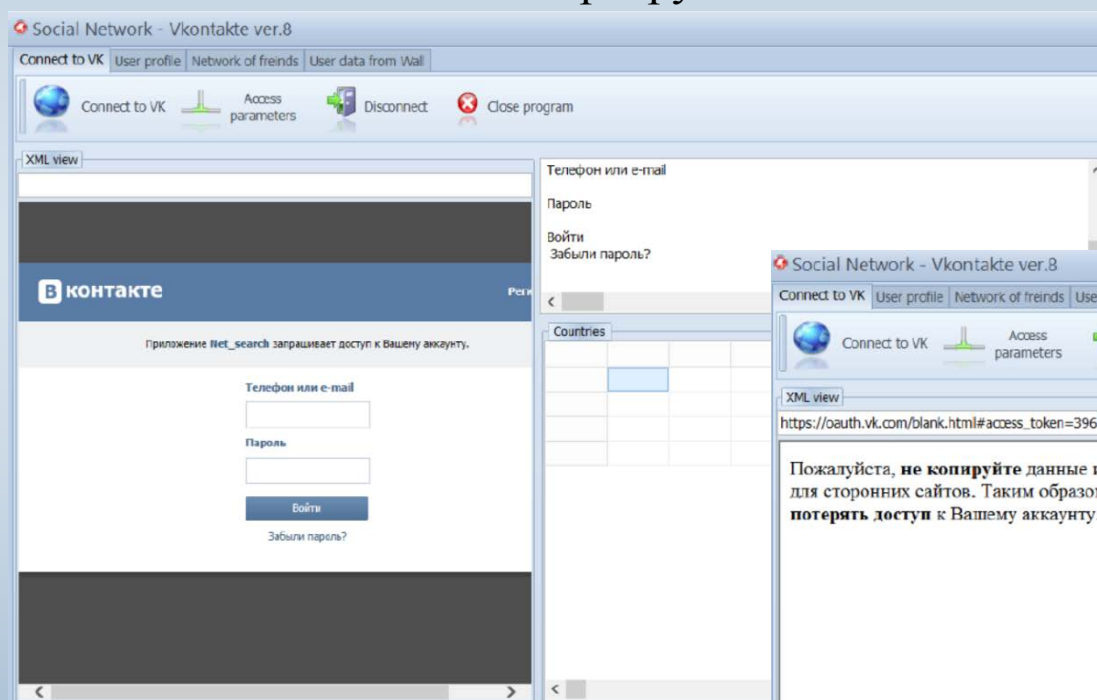
Исследование социальных сетей (Network analysis)



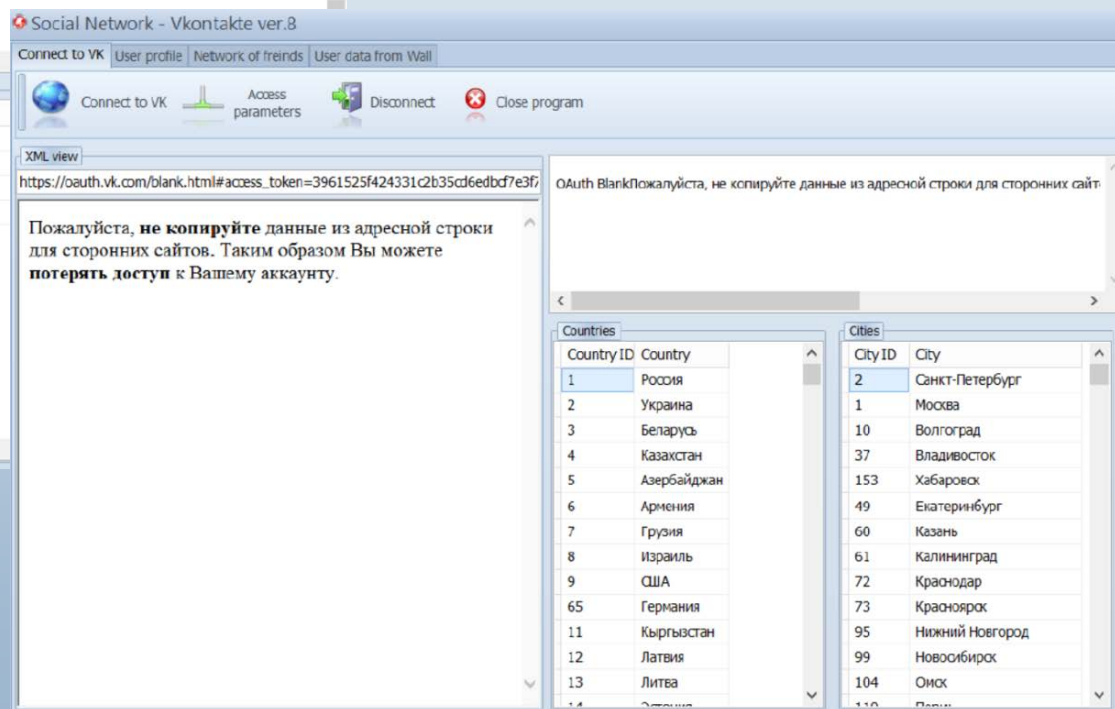
Сеть "дружбы" членов малой группы СПИД-диссидентов.
Размер вершин пропорционален внесенному вкладу в
создание контента.

Методы сбора данных из Вконтакте на примере программы VkMiner

1. этап. Подключение к серверу.

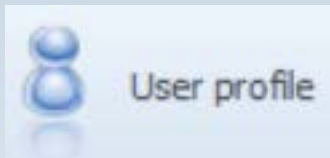


В результате подключения появится список городов и стран.



Методы сбора данных из Вконтакте на примере программы VkMiner

2. этап. Сбор данных.



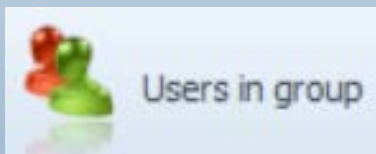
Сбор персональных данных (на основе ID пользователя).
Получаем персональные данные.



Сбор списка друзей пользователя. Получаем список ID друзей и их персональные данные



Сбор списка групп пользователя. Получаем список ID групп + наименование групп.



Сбор персональных данных пользователей из заданной группы.
Получаем список ID-ков пользователей и их персональные данные.

Методы сбора данных из Вконтакте на примере программы VkMiner

Social Network - Vkontakte ver.8

Connect to VK | Wall User Profile | **User profile** | Network of freinds | User data from Wall

User profile | User Freinds | Users' group | Users in group | To CSV | Save ID to TXT | Close program

Parameters for request

User ID: 1465880 Group ID: 33468641

Pause (ID group): 500 Pause (user info): 0

```
<first_name>Marc</first_name>
<last_name>Smith</last_name>
<sex>2</sex>
<nickname />
<screen_name>id259948270</screen_name>
<bdate>19.8.1965</bdate>
<city>208</city>
<country>9</country>
<photo>http://cs619216.vk.me/v619216270/11a00/51429of
<has_mobile>1</has_mobile>
<online>0</online>
<mobile_phone />
<home_phone />
<university>1064197</university>
<university_name>Drexel University</university_name>
<faculty>859987</faculty>
<faculty_name />
<graduation>1988</graduation>
<user_id>259948270</user_id>
</user>
</response>
```

Table grid

uid	first_name	last_name	screen_name	sex	birthdate
8733	Екатерина	Григина	e_grinina	1	
344291	Вероника	Богачек	id344291	1	2.11
407108	Ирина	Билик	irbilik	1	
512978	Евгений	Виноградов	id512978	2	3.5
897257	Андрей	Щербак	id897257	2	11.10
918070	Даша	Самарская	dash_x	1	2.10
965165	Эдуард	Писковацкий	repus	2	5.9.1971

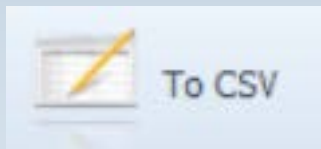
List of ID

53342500
58403904
82161314
86545793
132303479
225379983
259948270

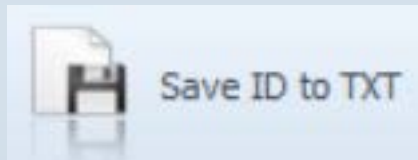
Statistics

Методы сбора данных из Вконтакте на примере программы VkMiner

3. этап. Выгрузка данных.



Выгрузка персональных данных в формате csv

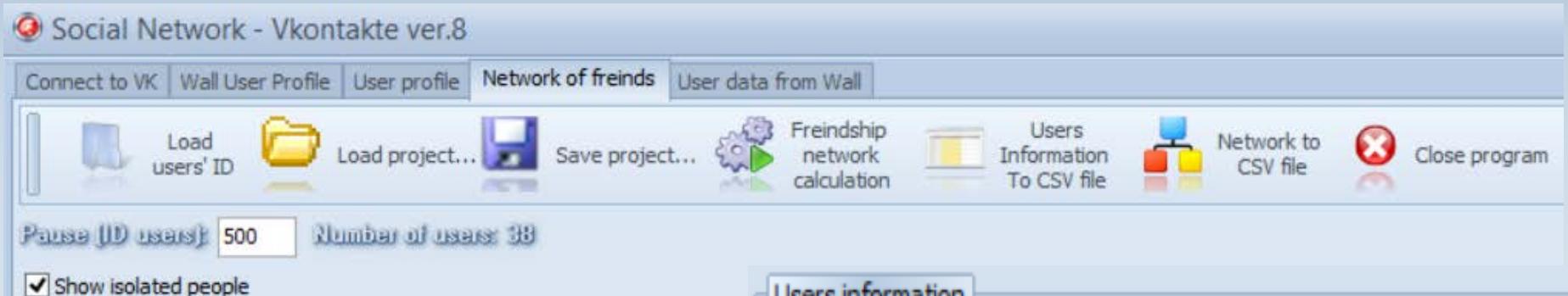


Выгрузка списка ID-ков в текстовом формате

	A	B	C	D	E	F	G	H	I
1	User ID	Edge	First name	Last name	Screen name	Sex	Birthday	City	Country
2	8733	4	Екатерина	Гринина	e_grinina	1		Санкт-Петербург	Россия
3	344291	0	Вероника	Богачек	id344291	1	2.11	Санкт-Петербург	Россия
4	407108	4	Ирина	Билик	irbilik	1		Санкт-Петербург	Россия
5	512978	3	Евгений	Виноград	id512978	2	3.5	Санкт-Петербург	Россия
6	897257	1	Андрей	Щербак	id897257	2	11.1	Санкт-Петербург	Россия
7	918070	8	Даша	Самарская	dash_x	1	2.1	Санкт-Петербург	Россия
8	965165	8	Эдуард	Писковац	repus	2	05.09.1971	Санкт-Петербург	Россия
9	1093666	4	Ростислав	Борисов	rrostik	2	24.01.1971	Санкт-Петербург	Россия

Расчет эго сети в VkMiner

Для того, что бы вычислить эго сеть, нужно иметь список ID-ков



Результат расчета:

Friendship network

	ID of users	ID of friends
1	8733	1786781
2	8733	2480879
3	8733	3207465
4	8733	9581896
5	407108	2123817
6	407108	2136355
7	407108	5413511

Users information

	User ID	Edge	First name	Last name	Screen nam
1	8733	4	Екатерина	Григина	e_grinina
2	344291	0	Вероника	Богачек	id344291
3	407108	4	Ирина	Билик	irbilik
4	512978	3	Евгений	Виноградов	id512978
5	897257	1	Андрей	Щербак	id897257
6	918070	8	Даша	Самарская	dash_x
7	965165	8	Эдуард	Писковацкий	repus
8	1093666	4	Ростислав	Борисов	rrostik
9	1291824	0	Анюта	Кобрина	id1291824
10	1786781	2	Сергей	Чечулин	s.chechulin

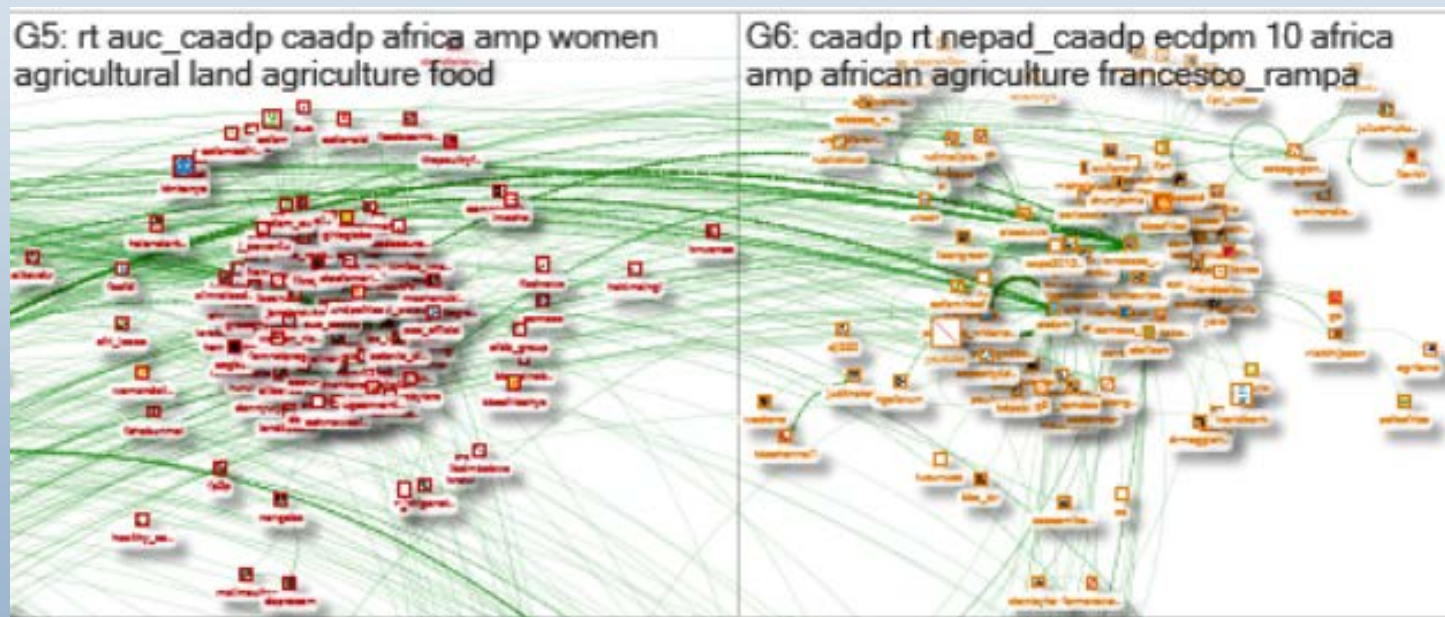
Сетевой анализ в NODEXL



NodeXL: Network Overview, Discovery and Exploration
for Excel

<http://nodexl.codeplex.com/>

NodeXL - программа-приложение, работающая на базе программы **Microsoft Excel**. По сути, это дополнительный набор функций, интегрированных в интерфейс Microsoft Excel, позволяющих визуализировать, анализировать, производить кластеризацию и иные вычислительные операции с графами. NodeXL - это бесплатное ПО, находящееся в открытом доступе.



Сетевой анализ в NODEXL

Загрузка данных.

	A	B
1	ID of users	ID of friends
2	8733	1786781
3	8733	2480879
4	8733	3207465
5	8733	9581896
6	407108	2123817
7	407108	2136355
8	407108	5413511
9	407108	58403904
10	512978	1093666
11	512978	9581896
12	512978	46179260
13	897257	5413511
14	918070	965165

Копируем
данные из
Экселя в шаблон
NODEXL



	A	B
1		
2	Vertex 1 ▾	Vertex 2 ▾
3	8733	1786781
4	8733	2480879
5	8733	3207465
6	8733	9581896
7	407108	2123817
8	407108	2136355
9	407108	5413511
10	407108	58403904
11	512978	1093666
12	512978	9581896
13	512978	46179260
14	897257	5413511

Сетевой анализ в NODEXL

В списке персональных данных (отсортированных по Edge) нужно удалить изолянтов. Изолянты – это пользователи у которых 0 связей.

A2		f _x		344291		
	A	B	C	D	E	F
1	User ID	Edge	First name	Last name	Screen name	Sex
2	344291	0	Вероника	Богачек	id344291	1
3	1291824	0	Анюта	Кобрина	id1291824	1
4	8199406	0	Влад	Дормидон	id8199406	2
5	8565247	0	Наталия	Синичкин	natali_sini	1
6	2.6E+08	0	Marc	Smith	id2599482	2
7	897257	1	Андрей	Щербак	id897257	2
8	6123691	1	Вероника	Итигилова	id6123691	1
9	18213157	1	Павел	Браславск	pbras	2
10	27260599	1	Лиза	Данилова	daniliz	1
11	46179260	1	Дмитрий	Лебедев	id4617926	2
12	1786781	2	Сергей	Чечулин	s.chechulir	2
13	1973164	2	Александр	Диденко	id1973164	2

Сетевой анализ в NODEXL

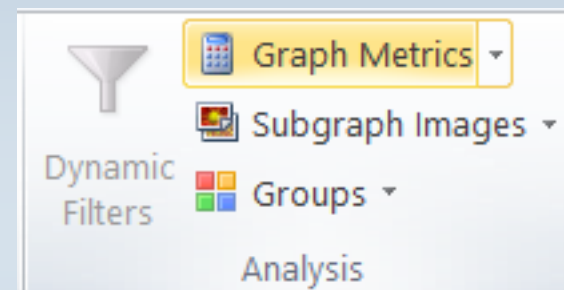
1. Список id-ков без изолянтов нужно скопировать на вкладку Vertices.

A3 fx 897257					
	A	B	C	D	E
1	Visual Properties				
2	Vertex	Color	Shape	Size	Opacity
3	897257				
4	612				
5	1821				
6	2726				
7	4617				
8	1786781				

Vertex Name
Enter the name of the vertex.

Calculate Metrics

2. Расчет метрик графа.



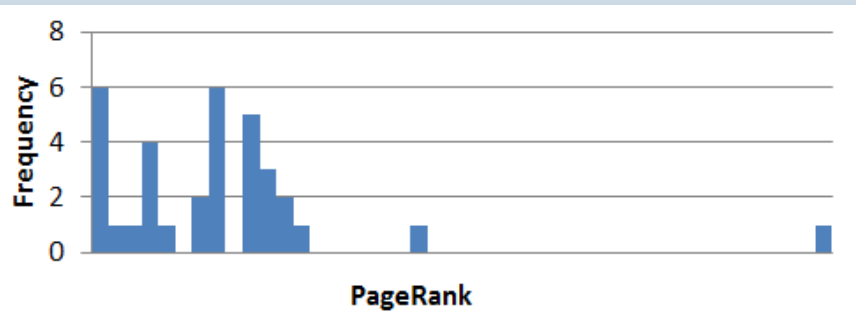
- ☒ Overall graph metrics
- ☐ Vertex degree (undirected graphs only)
- ☐ Vertex in-degree (directed graphs only)
- ☐ Vertex out-degree (directed graphs only)
- ☐ Vertex betweenness and closeness centralities
- ☐ Vertex eigenvector centrality
- ☒ Vertex PageRank
- ☒ Vertex clustering coefficient
- ☐ Vertex reciprocated vertex pair ratio (directed graphs only)
- ☐ Edge reciprocation (directed graphs only)
- ☒ Group metrics
- ☐ Words and word pairs
- ☐ Top items
- ☐ Twitter search network top items

Сетевой анализ в NODEXL

Метрики:

Degree (степень) - характеристика вершины, равная числу всех ребер, непосредственно соединяющих эту вершину с другими вершинами. Или другими словами, степень - это количество "соседних" вершин, непосредственно связанных с данной.

PageRank (индекс значимости узла) - является распределенной вероятностью и определяется не только за счет числа связей узла (вершины), но и за счет учета значимости самих связей. Принимается во внимание то, как много связей у каждой соседней вершины, связанной с данной.



Minimum PageRank	0.414
Maximum PageRank	3.729
Average PageRank	1.000
Median PageRank	0.980



Minimum Clustering Coefficient	0.000
Maximum Clustering Coefficient	1.000
Average Clustering Coefficient	0.608
Median Clustering Coefficient	0.732

Сетевой анализ в NODEXL

Метрики:

	A	B
1	Graph Metric ▾	Value ▾
2	Graph Type	Undirected
3		
4	Vertices	34
5		
6	Unique Edges	5
7	Edges With Duplicates	144
8	Total Edges	149

Graph Type: направленный или ненаправленный граф

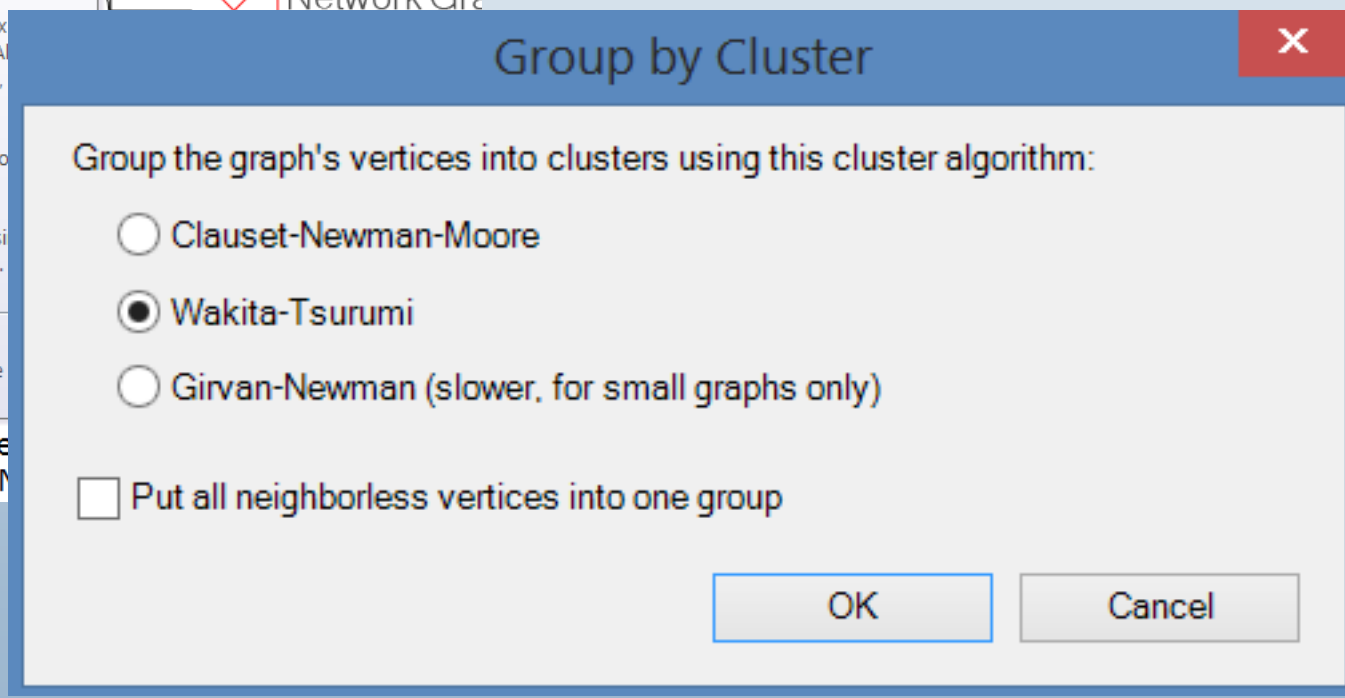
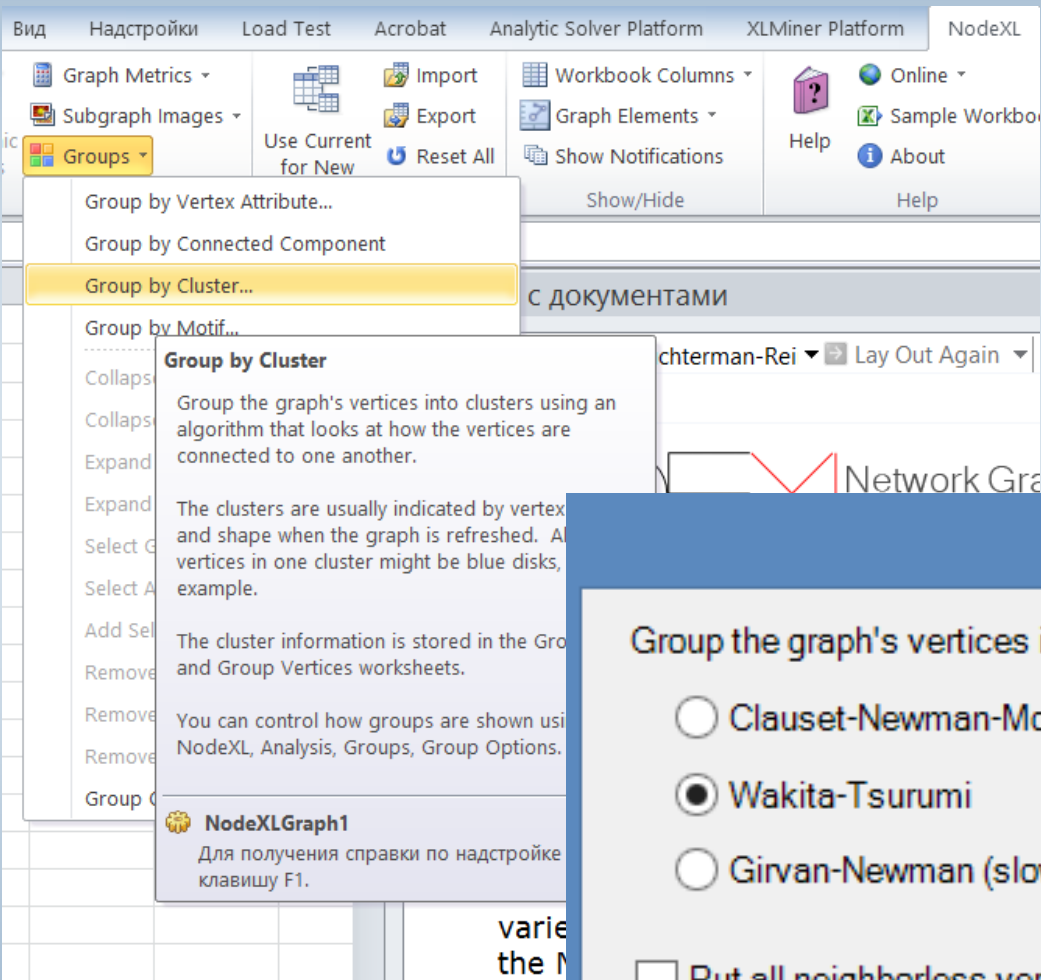
Vertices: Число уникальных узлов

Total Edges: общее число связей

"Maximum Geodesic Distance (Diameter)". Геодезическая дистанция - это кратчайший путь между двумя парами вершин. Этот показатель можно рассчитать для всех и каждой пары вершин в соединенном компоненте графа. Диаметр - это максимальное значение среди всех геодезических дистанций графа. Таким образом, диаметр - это показатель размеров графа.

Community detection в NODEXL

Группировка узлов по кластерам:
Группировка производится на основе
меры **Modularity**.



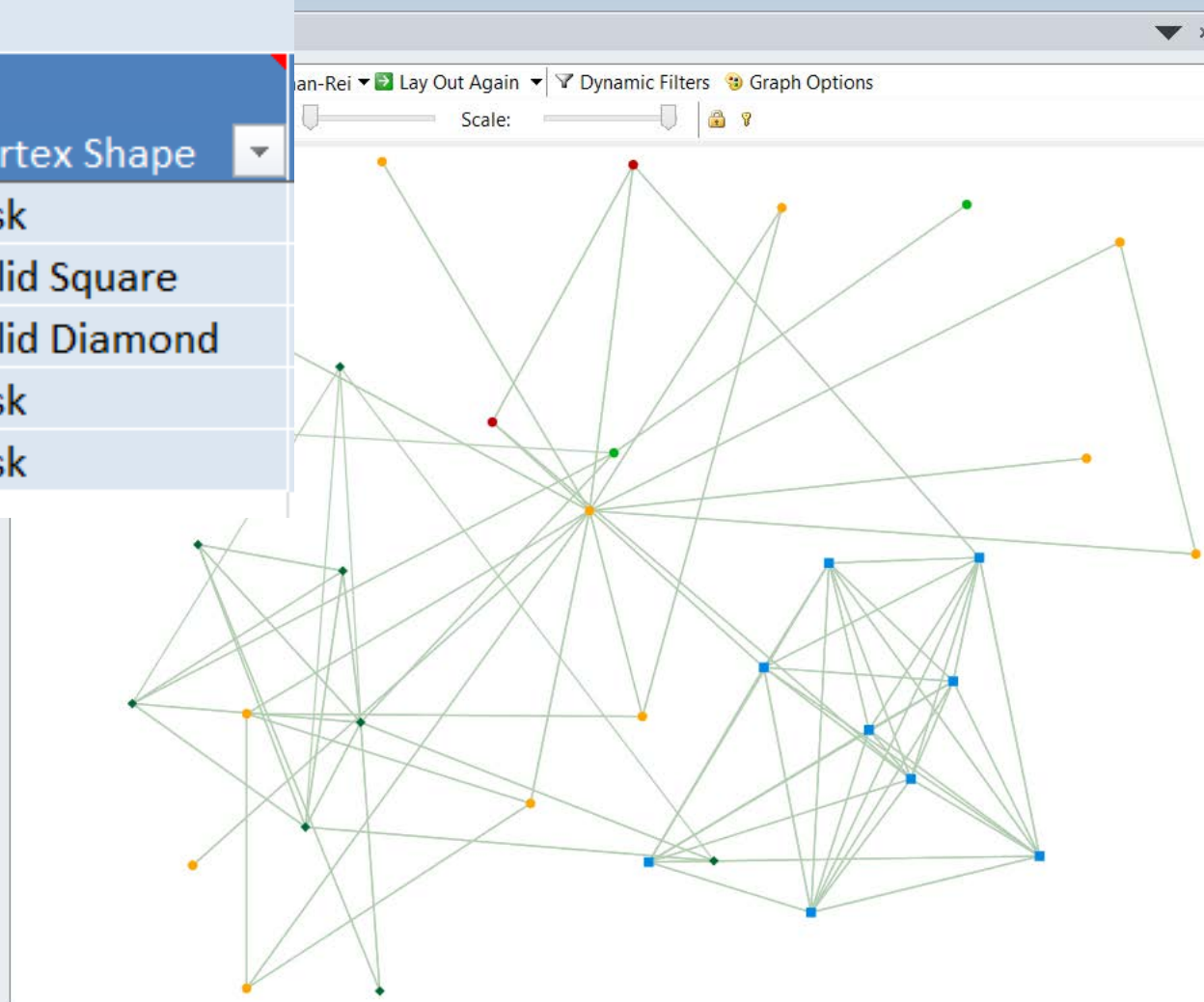
Сетевой анализ в NODEXL

В теории графов под понятием сообщества понимается сегмент графа (субграф), где фактическое количество ребер превышает ожидаемое количество ребер в случайном графе таких же размеров. Другими словами, **сообщество - это подмножество вершин, имеющих большую плотность связей, то есть связанных между собой большим количеством ребер, чем с "внешними" вершинами.** Визуально сообщества выглядят как "уплотнения" в полном графе, как сгустки узлов и связей. Именно поэтому выявление сообществ дает нам сведения о топологии сети, показывает на сколько и какие кластеры разбивается граф, какие узлы в них входят, какие кластеры с какими связываются и др.

Модульность (Modularity) - это отношение суммы разниц между фактическим и ожидаемым количеством ребер (в случайно сгенерированном графе таких же размеров), соединяющих пару вершин из одного кластера. Таким образом, модульность - это мера отклонения плотности связей в наблюдаемой сети от плотности, распределенной случайным образом.

Представление результатов

A	B	C
	Visual Properties	
Group	Vertex Color	Vertex Shape
G1	orange	Disk
G2	0, 136, 227	Solid Square
G3	0, 100, 50	Solid Diamond
G4	0, 176, 22	Disk
G5	191, 0, 0	Disk

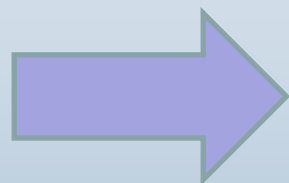


Представление результатов

Каждый узел (vertex) теперь имеет атрибут group (номер кластера)

	A	B	
1	Group ▼	Vertex ▼	V
2	G1	259948270	
3	G1	27260599	
4	G1	897257	
5	G1	18213157	
6	G1	2136355	
7	G1	2123817	
8	G1	407108	
9	G1	5413511	
10	G1	31309245	
11	G1	82161314	
12	G1	225379983	
13	G1	58403904	
14	G2	132303479	
15	G2	5483431	
16	G2	3829124	
17	G2	86545793	
18	G2	1887833	
19	G2	965165	

Эти две колонки
можно
скопировать в
эксель, где
находится
информация о
пользователях.



Таким образом
можно потом
сгруппировать
пользователей по
принадлежности к
кластерам.

C	D	E	F	G	H
		User ID	Edge	First name	Last name
8733	G3	8733	4	Екатерина	Григина
407108	G1	407108	4	Ирина	Билик
512978	G4	512978	3	Евгений	Виноградов
897257	G1	897257	1	Андрей	Щербак
918070	G2	918070	8	Даша	Самарская
965165	G2	965165	8	Эдуард	Писковацкий
1093666	G3	1093666	4	Ростислав	Борисов
1786781	G3	1786781	2	Сергей	Чечулин
1887833	G2	1887833	8	Александр	Ларюшин
1973164	G3	1973164	2	Александр	Диденко
2123817	G1	2123817	3	Надежда	Билик
2136355	G1	2136355	3	Виктория	Лейко
2480879	G3	2480879	4	Александр	Петров
3207465	G3	3207465	5	Денис	Евладов
3770247	G5	3770247	3	Галина	Хотина
3829124	G2	3829124	9	Михаил	Россет
5413511	G1	5413511	12	Олеся	Кольцова
5483431	G2	5483431	8	Нина	Рохлина
6123691	G4	6123691	1	Вероника	Итигилова

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