ALFRED NOBEL UNIVERSITY

SOCIAL NETWORK ANALYSIS

Author: V. Kosariev, Alfred Nobel University, PhD in Technical Sciences, Associate Professor.
Co-author: N. Rizun, PhD in Technical Sciences, Associate Professor.

Dnipro
2018

УНІВЕРСИТЕТ імені АЛЬФРЕДА НОБЕЛЯ

АНАЛІЗ СОЦІАЛЬНИХ МЕРЕЖ

НАВЧАЛЬНИЙ ПОСІБНИК

ЕЛЕКТРОННЕ ВИДАННЯ

В.М. Косарєв, кандидат технічних наук, доцент.
Н.О. Різун, кандидат технічних наук, доцент.

Дніпро
2018

УДК 004.7:316.4
К 71

Рекомендовано вченою радою
Університету імені Альфреда Нобеля
(протокол № 7 від 14.12.2017 р.)

Рецензенти:
І.В. Іщенко, доктор політичних наук, професор, завідувач кафедри міжнародних відносин Дніпровського національного університету імені Олеся Гончара;
В.Є. Момот, доктор економічних наук, професор, проректор з організації та розвитку наукової та міжнародної науково-освітньої діяльності Університету імені Альфреда Нобеля.

Косарєв В.М.
К 71 Social Network Analysis = Аналіз соціальних мереж: навчальний посібник [Електронний ресурс] / В.М. Косарєв, Н.О. Різун. – Дніпро: Університет імені Альфреда Нобеля, 2018. – 264 с.

ISBN 978-966-434-427-9

У навчальному посібнику висвітлюються питання, пов’язані з умовами і принципами успішної роботи з різноманітними сервісами, службами та послугами у популярних соціальних мережах. Завдяки пропонованому посібнику студенти мають можливість набути не лише систематизованих теоретичних знань, але і практичного досвіду професійної діяльності.

Навчальний посібник може становити інтерес для студентів вищих навчальних закладів різних спеціальностей.

УДК 004.7:316.4

ISBN 978-966-434-427-9

CONTENTS

SUMMARY

Social Network Analysis (SNA) has become the main method of research in contemporary sociology, anthropology, geography, social psychology, computer science and research organizations, as well as a widespread theme for research and discussion. Research in several academic areas has shown that social networks operate on many levels, ranging from families to nationalities, and play an important role in solving problems, working organizations, and succeeding on their own goals for the individual.

In western sociology, the network approach is used in an extremely wide range of studies. The phenomenon of social capital, the relationship between corporations and non-governmental organizations, the ability to self-organize within local communities, organized crime and labor migration, ethnic entrepreneurship and access to resources, gender social networks and health, computer networks and the global economic system – this is only an incomplete list of topics that are analyzed using the network theory.

With ready access to computing power, the popularity of social networking websites such as Facebook, and automated data collection techniques the demand for solid expertise in SNA has recently exploded.

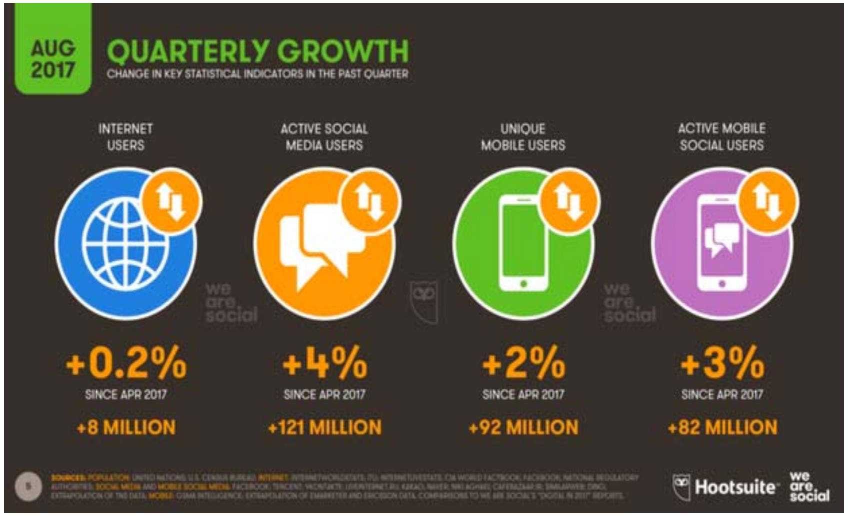
The number of **Facebook** users has exceeded **two** billion people. At the same time, the total number of users of social media in the world has reached a new mark in **three** billion people.

The latest study by We Are Social and Hootsuite indicates that the number of social media users in the world has reached a new level of 3 billion users.

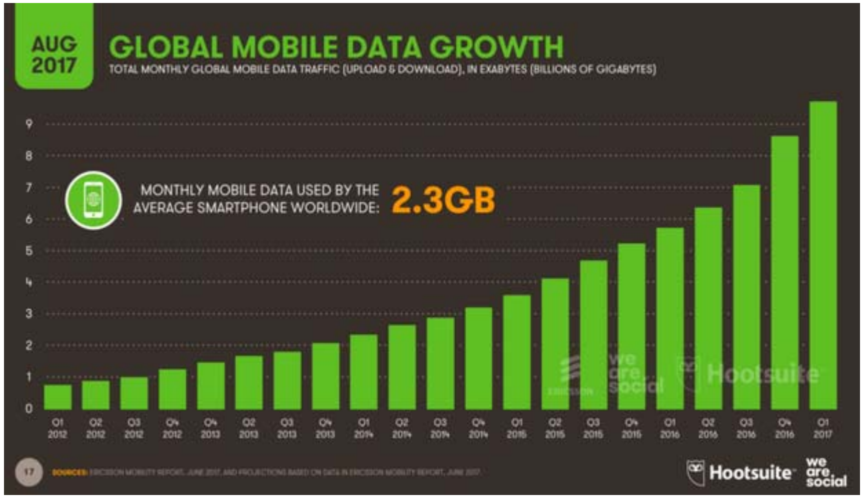
Over the past 4 months, Facebook showed significant growth and reached 2,046 billion monthly active users.

YouTube took the second place in terms of the number of active users. The third place was taken by the WhatsApp application. Popular instant messengers such as WhatsApp and Facebook Messenger have increased their audience, but currently, platforms do not provide data on individual markets.

The use of mobile devices continues to grow worldwide, and according to GSMA Intelligence, more than 650,000 users are growing every day. Moreover, average smartphone users now consume more than 2.3GB of data each month with the help of a device. New figures suggest that mobile phones are the most important device for two-thirds of the world's population, and three-quarters of all smartphones run on Android.



This interdisciplinary course introduces students to the basic concepts and analysis techniques in SNA. Students learn how to identify key individuals and groups in social systems, to detect and generate fundamental network structures, and to model growth and diffusion processes in networks [1].



The social network analysis process involves four basic steps as shown in the graph on below:

- Define a goal, question or task.
- Collect data.
- Analyze the data.
- Interpret the results in order to complete your goal, answer your question, or solve your task.

In this course we will start with basic statistical descriptions of networks, analyze network structure, roles and positions of nodes in networks, connectivity patterns and methods for community detection. In the second part of the course we will discuss processes on networks and practical methods of network visualization.

We conclude the course with examples from social media mining [2].

Learning Objectives

The learning objective of the course «Social Network Analysis» is to provide students with essential knowledge of network analysis applicable to real world data, with examples from today's most popular social networks.

Learning outcomes

After completing the study of the discipline «Social Network Analysis» the student should:

- Know basic notation and terminology used in network science.
- Be able to visualize, summarize and compare networks.
- Understand basic principles behind network analysis algorithms.
- Be capable of analyzing real work networks.

The class meeting time will be centered on lecture, but will also include a substantial amount of class discussion at times.

Unit 1
INTRODUCTION IN SOCIAL NETWORK ANALYSIS

Social Networks and Social Media in the real life

The basic idea of social networks is very simple. Under the social network refers to many actors (points, vertices, agents) that can interact with each other. From a formal point of view, such networks are convenient to represent in graphs and apply for their analysis developed mathematical models.

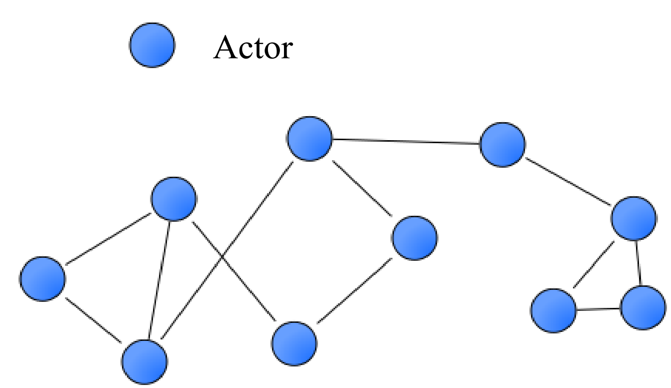
A ***social network*** is a social structure made up of individuals (or organizations) called «***nodes***», which are ***tied*** (connected) by one or more specific types of interdependency, such as: friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige.

The term was introduced in 1954 by a sociologist from the «Manchester School» James Barnes in the work «Classes and Meetings in the Norwegian Isle of Ward», which was included in the collection Human Relations. Before that, many thinkers about society expressed the idea of the importance of considering society as a complex interweaving relationship.

By this term he expressed the opinion that society is a complex interweaving of relations. Barnes investigated interconnections between people using visual diagrams in which individual individuals are represented by dots, and the links between them – lines.

Social network analysis views social relationships in terms of network theory consisting of *nodes* and *ties* (also called *edges*, *links*, or *connections*).

Nodes are the individual ***actors*** within the networks, and ***ties*** are the relationships between the actors. The resulting graph-based structures are often very complex. There can be many kinds of ties between the nodes.



Picture of the social network of Barnes

Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals.

Social network analysis (related to network theory) has emerged as a key technique in modern *sociology*. It has also gained a significant following in *anthropology*, *biology*, *communication* studies, *economics*, *geography*, *information* science, *organizational* studies, social *psychology*, and *sociolinguistics*, and has become a popular topic of speculation and study.

People have used the idea of «social network» loosely for over a century to connote complex sets of relationships between members of social systems at all scales, from interpersonal to international.

Social network analysis has now moved from being a suggestive metaphor to an analytic approach to a paradigm, with its own theoretical statements, methods, social network analysis software, and researchers.

Analysts reason from whole to part; from structure to relation to individual; from behavior to attitude. They typically either study whole networks (also known as complete networks), all of the ties containing specified relations in a defined population, or personal networks (also known as egocentric networks), the ties that specified people have, such as their «personal communities».

Several analytic tendencies distinguish social network analysis:

– There is no assumption that groups are the building blocks of society: the approach is open to studying less-bounded social systems, from nonlocal communities to links among websites.

– Rather than treating individuals (persons, organizations, states) as discrete units of analysis, it focuses on how the structure of ties affects individuals and their relationships.

– In contrast to analyses that assume that socialization into norms determines behavior, network analysis looks to see the extent to which the structure and composition of ties affect norms.

SNA differs from convention an approaches to business problems in one very important way: SNA **assumes** that people are all interdependent.

This **assumption** is radically different from traditional research approaches which assume that what people do, think, and feel is independent of who they know.

The focus on interdependence means that SNA can ask and answer questions such as:

• Is Sales effectively communicating with Marketing to share and coordinate information about the customer?

• When two companies or organizations merge, how can management use the informal network to spread important messages?

• Are decisions in a distributed software development team being made and carried out efficiently or are one or more people acting as bottlenecks?

• In R&D group, are there enough people bringing in ideas from outside and are those ideas being acted upon?

Many traditional statistical techniques are based on the assumption of independence. For this reason, traditional statistics, such as comparing the means of two groups, cannot be conducted on interdependent data.

To deal with this problem, SNA has developed a set of SNA-specific statistics such as centrality and density that provide measures of interdependence

History of social network analysis

1900 s. Precursors of social networks in the late 1800s include **Émile Durkheim** and **Ferdinand Tönnies**.

In the 1930s, **Jacob Levy Moreno** pioneered the systematic recording and analysis of social interaction in small groups, especially classrooms and work groups (*sociometry*).

In 1959 s. Hungarian mathematicians Paul Erdos and Alfred Renyi became interested in the principles of social networking. They wrote a number of articles based on their research. Duncan J. Watts and Steven H. Strogatz developed the theory of social networks and first proposed the concept of a clustering factor, i.e. the degree of closeness between heterogeneous groups (when a person expands the network of his connections at the expense of persons whom he does not knows personally, but know her acquaintances) [14].

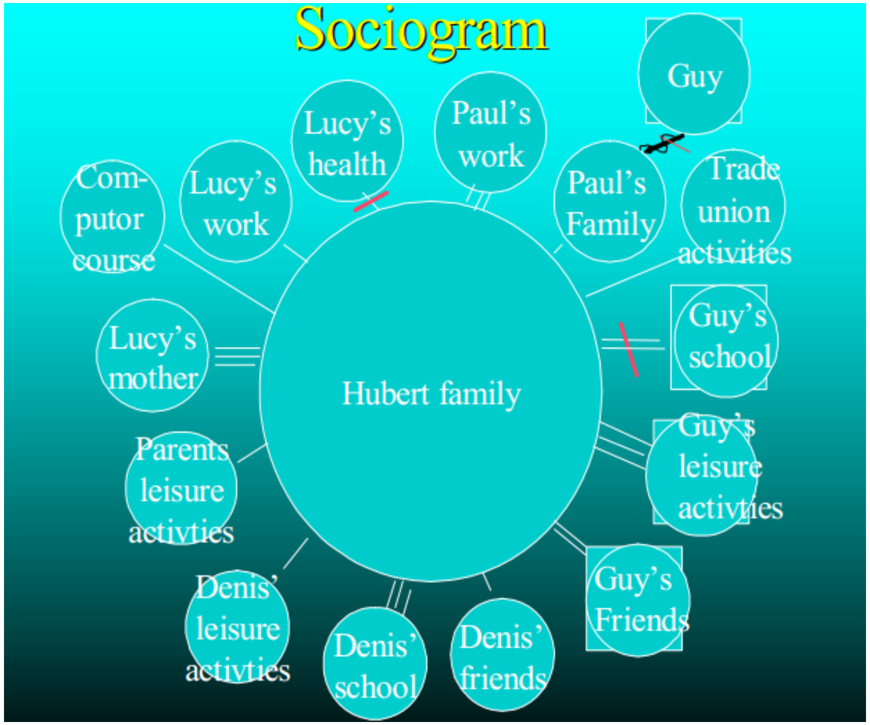
Another researcher, J. Scott, highlights the mathematical tradition of studying social networks as an independent direction within the framework of network analysis. However, the mathematical operationalization of the catego-ries of network analysis, mathematical modeling of processes in social networks, visualization of social networks with the use of mathematical methods is not so much an independent direction of network analysis, but as a method of analysis.

A social network is a community of people united by the same interests, preferences or those who have other reasons for direct communication between themselves.

The theory of social networks considers social relationships in terms of nodes and relationships. The nodes are isolated actors in the networks, and the links correspond to the relationships between the actors. There may be many types of links between nodes. In the simplest form, the social network is a reflection of all relationships that are relevant to the research between nodes. Networks can be used to establish the social capital of individual actors. These concepts are often reflected in the diagram of a social network, on which the nodes meet the points, and the links are lines [14].

In this graphic representation, as in the genogram, the intensity of the ties is indicated by a code of lines: a dotted line indicates a weak relationship, and as the dots get weaker, the relationship is weaker.

If the **dotted line is white**, the relationship is almost non-existent.



A **single** line shows a **good bond**, **two lines** a **closer bond** and **three lines** a relationship which is even **more important**, but also **conflicting**.

A line with a **slash** represents a **difficulty**.

A **broken** line or with **two slashes** signifies a rupture, while a **zigzag** line reveals a conflict.

An **arrow** indicates **unilateral** relationship, that is, the subject feels an attachment to the indicated person or group but this is not reciprocal, in other words, the indicated person or group reject them.

The arrow can also indicate a particular responsibility or inversely, for example, the case of a daughter who takes care of her mother.

The subject observed is indicated by a double circle, a c circle within a square or their name is written in letters of a different colour.

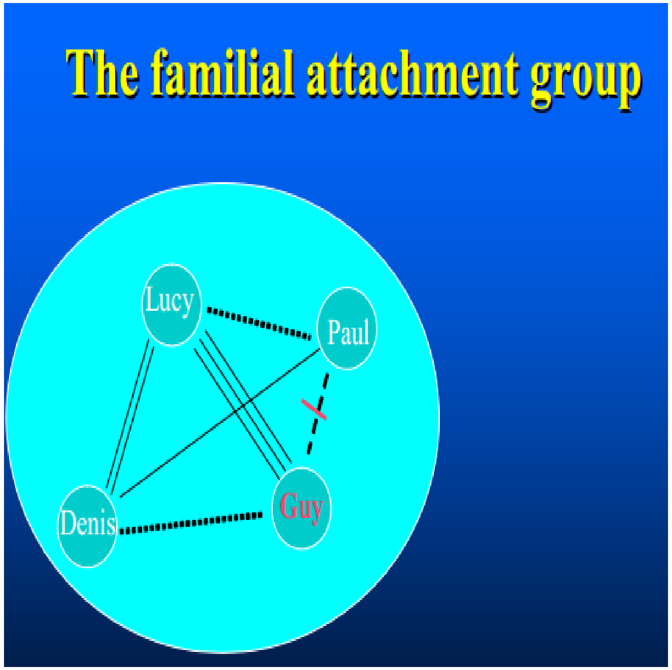
The dimensions depend on what one wants to show, for example, we might want to show the openness of a family system to exterior human relationships i.e. friendships, the enlarged family, leisure time activities or even the relationship of a single person to their work environment, e.g. a nurse with her work team or a student with a few of his classmates.

This visual tool complements the other two sociological tools: the genogram and the diagram of the attachment group, both of which favour the study of the family group.

The attached example shows the ensemble of the relationships of the Hubert family:

Paul, the father,
Lucy, the mother,
Guy and Denis, their two sons, with their in-laws,
the health and educational system,
the mother and father's professional activities
and their involvement in social organisations.

One can thus see that Paul is intensely involved with his work which signifies that he is somewhat absent at home especially since he is involved with trade union work. Moreover, his relationship with his family of origin is



The Hubert family is dysfunctional. The relationship between Lucy and Paul is not very good whereas the relationship between Lucy and her son Denis is very good and is confluent with her son Guy. The relationship between Paul and Guy is conflicting but is good with his other son, Denis. The relationship between the two brothers, Guy and Denis, is not very good.

not very strong. As far as Lucy is concerned, one notices that she works outside the home, is taking a computer course and that she has some health problems all of which suggest that this may be causing some difficulties. Moreover, she has a very strong, even confluent relationship with her mother and with Guy. There is not much evidence of leisure time activities shared by the entire family.

These observations indicate several factors of imbalance. As far as the children are concerned, Denis is a good student who is not too distracted by friends or his time the family since there is a conflicting activity; whereas his brother Guy has trouble in school but has a network of friends who occupy his leisure time which may be a cause of his school problems and a compensation for what is missing in relationship with his father. He is a problem child.

The social network form helps determine the extent to which its members are useful. Smaller, more connected networks may be less useful to their members than networks with many weak links with individuals outside the main network. More open networks, with many weak links and social relationships, are more likely to offer new ideas and opportunities to their members than closed networks with many redundant links. In other words, a group of friends who communicate with each other already have common knowledge and capabilities. A group of people with links to other social communities is likely to gain access to a wider range of information. To succeed, individuals are better off having more than one network connection than many connections within the same network. Similarly, individuals can influence or act as a broker in the middle of their social networks, connecting two networks that do not have direct links (called social holes).

Social networks have also been used to explore how companies interact, characterizing many informal ties that combine leadership, as well as associations and relationships between individual workers in different companies. These networks enable companies to gather information, keep up the competition, and even secretly negotiate pricing or policies.

There are four approaches to analyzing social networks:

1. ***Structural*** **–** focuses on the geometric form and intensity of interactions (the weight of the edges). All actors are considered as vertices of the graph, which affect the configuration of the edges and other network actors. Particular attention is paid to the mutual arrangement of vertices, centrality, transitivity of interactions. In order to interpret the results, in this case, structural theories and theory of network exchange [4] are used.

2. ***Dynamic*** – attention is focused on changes in the network structure over time. The causes of the disappearance and appearance of the network edges are studied; Change the structure of the network in external actions; Stationary configurations of the social network [4].

3. ***Normative*** – examines the level of trust between actors, as well as rules, rules and sanctions that affect the behavior of actors in the social network and the processes of their interactions. In this case, the social roles that are associated with the given edge of the network are analyzed, for example, the relationship between the leader and the subordinate, friends or family ties. The combination of individual and network resources actor with the rules and rules that operate in this social network, forms its «social capital» [4].

4. ***Resource*** – considers the ability of actors to attract individual and network resources to achieve a certain goal and differentiate actors who are in identical structural positions of the social network, their resources. Individual resources can be knowledge, prestige, wealth, ethnicity, gender (gender identity). Under network resources refers to the impact, status, information, capital [4].

In the social analysis, the following analytical methods are used:

1) the methods of the theory of graphs;

2) methods of finding local properties of subjects;

3) methods for determining equivalence actors, including their structural equivalence;

4) block models and roles algebra;
5) analysis of dyads and triads;
6) probabilistic models;
7) correspondent analysis and topological methods representing the network as a simplicial complex [4].

The peculiarity of the available social networks as sources of social data is that users are expanding their network fairly quickly and without additional cost to the social analyst. This leads to the large size of social graphs, which can be carried out extensive research of various properties of social networks.

Today, there are a lot of analysis of social data, used mainly by sociologists for research: UCINET, NetDraw, Pajek, Netminer, Visone, SNA / R, StOCNET, Negopy, InFlow, GUESS, NetworkX, prefuse, JUNG, BGL / Python and Others A detailed list of systems for analyzing social networks can be found on the ISNA International Social Analysis Network [12] and in the Survey [5].

One of the most famous is UCINET – a commercial product developed by the American company Analytic Technologies. It allows you to analyze social networks, using a wide variety of methods of analysis, export data to most popular formats, integrates with network visualization system NetDraw. One of the most important restrictions is the maximum size of the actors on the network 32767, but when processing data for 5,000–10,000 actors there are significant delays in the work. Slovenian developers Vladimir Batagelj and Andrej Mrvar system Pajek [13] created for processing large data. The processing of large social networks is achieved by clustering them to smaller ones and applying adapted algorithms.

Unlike the classical network and social group, say, scientists, engineers, doctors, social community operating in the Internet environment, that is, the Social Network, allows operative study, measurement and classification – through the integrated into the control program environment of the modules of statistics, analysis and forecasting.

Social networks can potentially become an instrument for transforming society, spreading scientific and technical knowledge, forming teams and civic movements, marketing research and promoting products and services.

In order for social networks to take their place in the scientific, technical and social spheres, they must go from the current initial state to maturity and become a common tool of social life and communication of the information society. Predicted path of development – the division of spheres of influence on the network of contacts, communication and professional networks.

Traditional Media

One-to-Many

Media (communication) tools used to store and deliver information or data

Traditional Media

One to One

Many to Many

Conversation
Entertainment
Relationship Building
Participation
Sharing
Rating/ Tagging
Publishing
Content Dissemination/ Distribution
SOCIAL MEDIA
(for business)
Virtual Worlds
MMO
Social Gaming
Microblogging
Social Networking
Discussion
Video Sharing
Livecasting
Photo Sharing
Content Publishing
Articles
Blogging
Wiki's
Lifestreaming
Reviews/ Recommendations
Collaborative Filtering
Amplifiers: Aggregation/Recombination, Tracking, Apps/Tools, Mobile

Various Forms of Social Media

Blog	Wordpress, Blogspot, LiveJournal, BlogCatalog
Forum	Yahoo! answers, Epinions
Media Sharing	Flickr, YouTube, Justin.tv, Ustream, Scribd
Microblogging	Twitter, foursquare, Google buzz
Social Networking	Facebook, MySpace, LinkedIn, Orkut, PatientsLikeMe
Social News	Digg, Reddit
Social Bookmarking	Del.icio.us, StumbleUpon, Diigo
Wikis	Wikipedia, Scholarpedia, ganfyd, AskDrWiki

40% of the top 20 websites are social media sites

Social Networks growth..

Social networking accounts for:

- One in every nine people on Earth is on Facebook
- Each Facebook user spends on average 15 hours and 33 minutes a month on the site
- 30 billion pieces of content is shared on Facebook each month
- 300,000 users helped translate Facebook into 70 languages
- People on Facebook install 20 million "Apps" every day

[http://www.jeffbullas.com/2011/09/02/20-stunning-social-media-tatistics/#q3eTJhr64rtD0tLF.99]

Social Networks Growth..

- YouTube has 490 million unique users who visit every month (02/2011)
- Users on YouTube spend a total of 2.9 billion hours per month (326,294 years)!
- Wikipedia hosts 17 million articles and has over 91,000 contributors
- People upload 3,000 images to Flickr every minute and hosts over 5 billion images!
- 190 million average Tweets per day occur on Twitter (May 2011)
- Twitter is handling 1.6 billion queries per day
- Google+ was the fastest social network to reach 10 million users at 16 days (Twitter took 780 days and Facebook 852 days)

[http://www.jeffbullas.com/2011/09/02/20-stunning-social-media-statistics/#q3eTJhr64rtD0tLF.99]

Characteristics of Social Media

I. User Based

- Users submit and organize information
- Direction of content can be determined by any user – no one person dictates the current topic
- Freeform/unstructured

Characteristics of Social Media

II. Interactive

- Not just a collection of chat rooms and forums
- Users can play games, take fun quizzes, share photos and ideas with friends
- A way to connect and have fun with friends

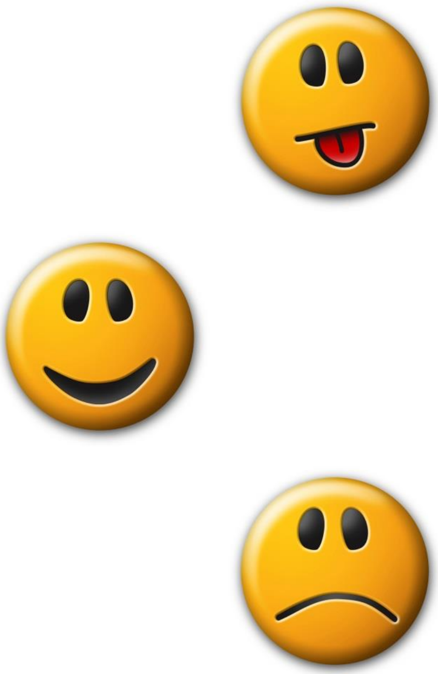
Characteristics of Social Media

III. Community Driven

- Members hold common beliefs or interests
- Can make new friends with people who say they share your interests or beliefs
- Can reconnect with old friends

Characteristics of Social Media

IV. Emotional Content



- In the past, web content was primarily information
- Social networks allow people to communicate needs within a community of friends and receive immediate responses

Characteristics of Social Media

V. “Consumers” become “Producers”

The consumer is:

- writing content for blogs, online news gathering orgs, Twitter, Facebook, self-published books;
- shooting and editing her own videos/films;
- creating, recording and producing her own music;
- shooting her own photographs
- etc.

Characteristics of Social Media

VI. Publish, then Filter

Clay Shirky popularised the elegant 'publish, then filter' formulation in his book *Here Comes Everybody*, published 2008, but had been using the phrase for many years before that. In 2002, he told an audience at the BBC:

> *"The order of things in broadcast is 'filter, then publish'. The order in communities is 'publish, then filter'. If you go to a dinner party, you don't submit your potential comments to the hosts, so that they can tell you which ones are good enough to air before the group, but this is how broadcast works every day. Writers submit their stories in advance, to be edited or rejected before the public ever sees them. Participants in a community, by contrast, say what they have to say, and the good is sorted from the mediocre after the fact.*
>
> *Media people often criticize the content on the internet for being unedited, because everywhere one looks, there is low quality – bad writing, ugly images, poor design. What they fail to understand is that the internet is strongly edited, but the editorial judgment is applied at the edges, not the center, and it is applied after the fact, not in advance. Google edits web pages by aggregating user judgment about them, Slashdot edits posts by letting readers rate them, and of course users edit all the time, by choosing what (and who) to read."*

'Time to review peer review', Andrew Pontzen notes that:

> *"These days most physicists now download papers from arxiv.org, a site which hosts papers regardless of their peer-review status. We skim through the new additions to this site pretty much every day, making our own judgements or talking to our colleagues about whether each paper is any good. Peer-review selection isn't a practical priority for a website like arxiv.org, because there is little cost associated with letting dross rot quietly in a forgotten corner of the site. Under a digital publication model, the real value that peer review could bring is expert opinion and debate; but at the moment, the opinion is hidden away or muddled up because we're stuck with the old-fashioned filtration model."*

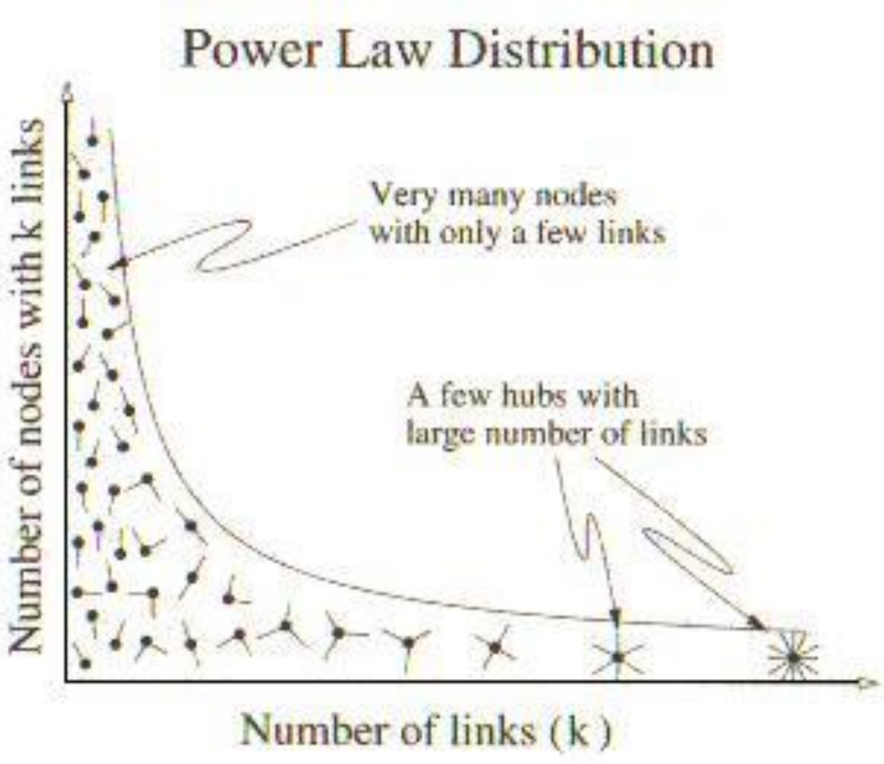
Top 20 Websites in the US

Rank	Site	Rank	Site
1	google.com	11	blogger.com
2	facebook.com	12	msn.com
3	yahoo.com	13	myspace.com
4	youtube.com	14	go.com
5	amazon.com	15	bing.com
6	wikipedia.org	16	aol.com
7	craigslist.org	17	linkedin.com
8	twitter.com	18	cnn.com
9	ebay.com	19	espn.go.com
10	live.com	20	wordpress.com

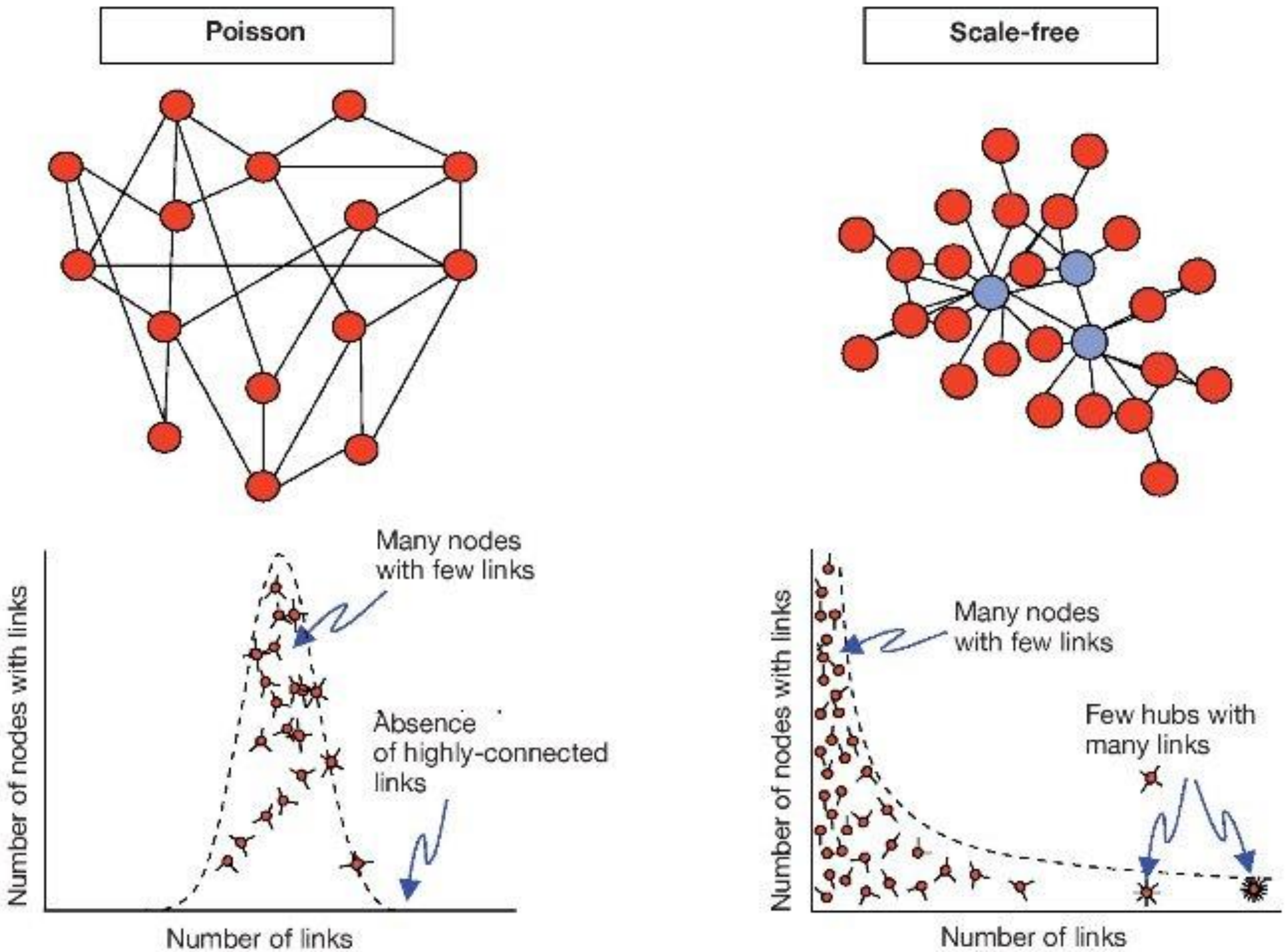
PROPERTIES OF LARGE-SCALE NETWORKS

I. Scale Free Distributions

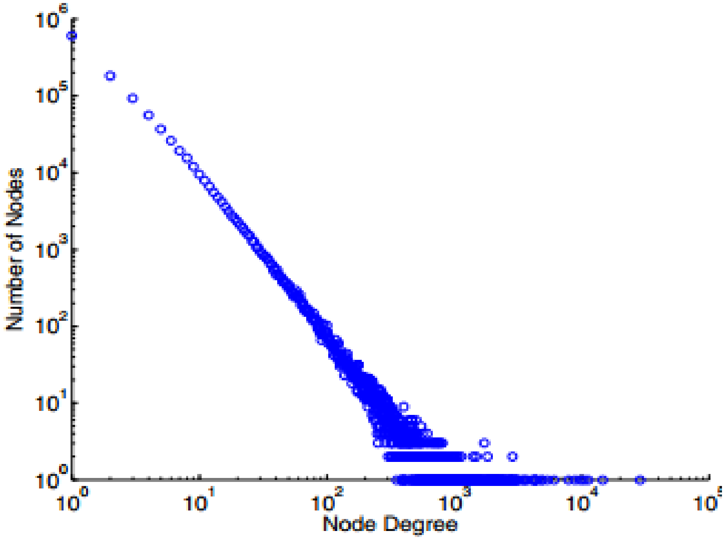
$$p(x) = Cx^{-\alpha} \qquad x \geq x_{min} \qquad \alpha > 1$$



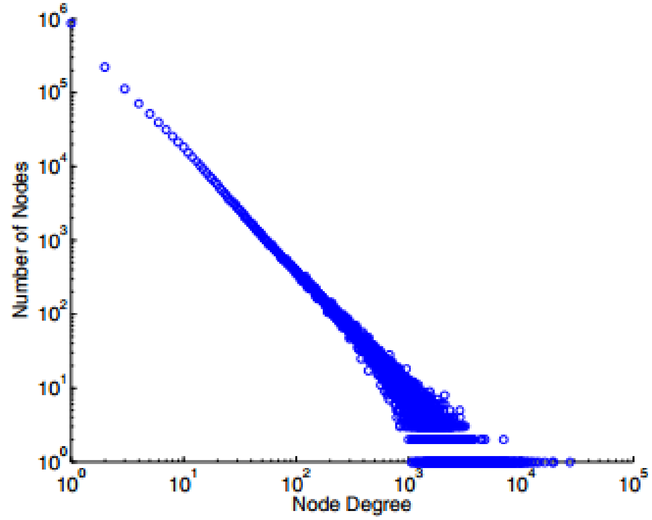
Scale free distribution



PLOT THE DEGREE DISTRIBUTION ON A **LOG-LOG SCALE**



Friendship Network in Flickr



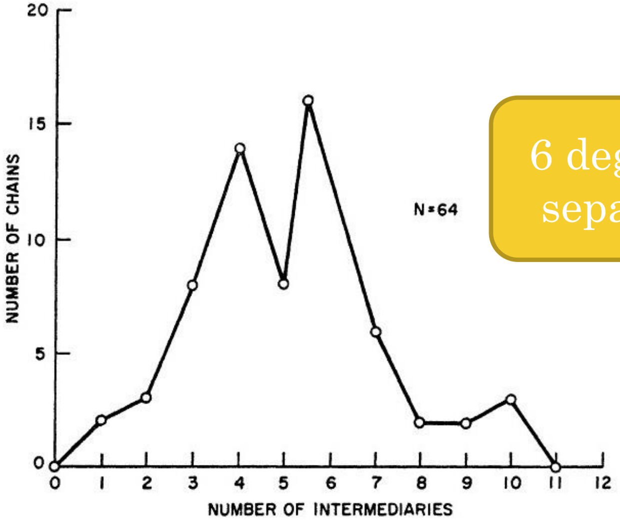
Friendship Network in YouTube

if we zoom into the tail (say, examine those nodes with degree>100), we will still see a power law distribution.
This self-similarity is independent of scales

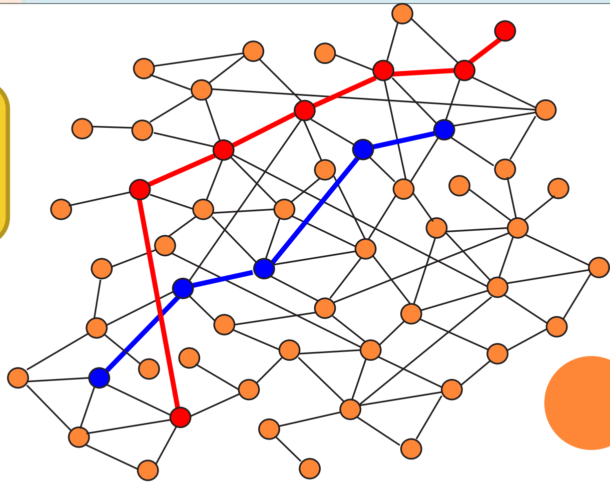
PROPERTIES OF LARGE-SCALE NETWORKS

II. Small World Effect

Stanley Milgram's experiment developed out of a desire to learn more about the probability that two randomly selected people would know each other.



6 degrees of separation

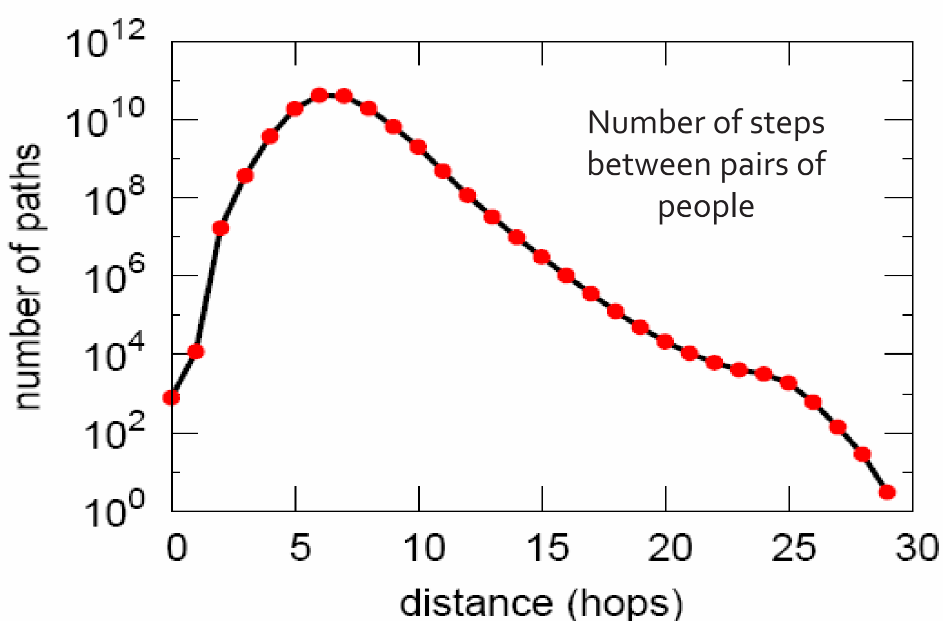


Network: Small world

- Messenger **social network** of the whole planet Earth
  - 240M people, 1.3B edges

Leskovec and Horvitz, 2008

MSN Messenger network



Avg. path length **6.6**
90% of the people can be reached in < 8 hops

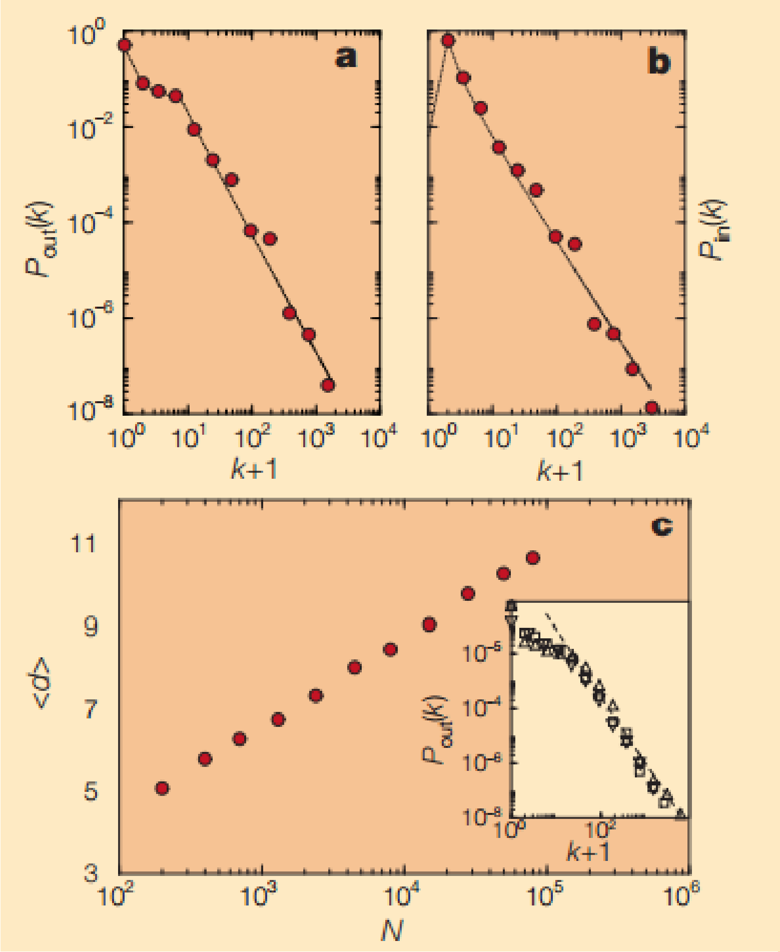
Hops	Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

Small World Effect

Albert, Jeong, and **Barabasi** (1999

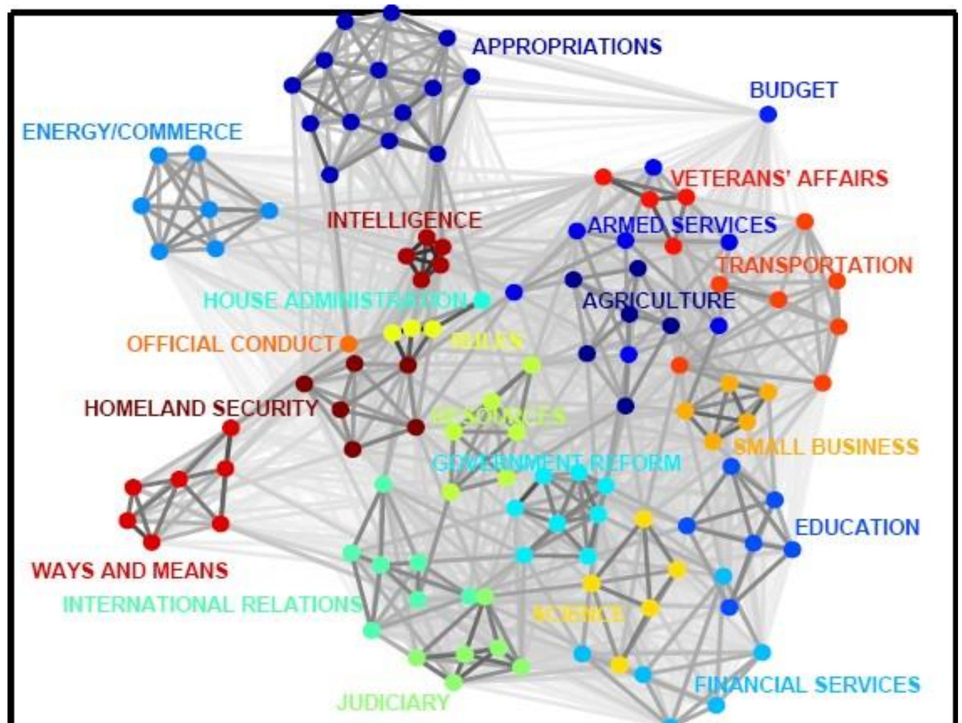
"Diameter of the World Wide Web"

Estimated that in 1998 it took on average **11** clicks to go from one random website to another (at the time there were 800 million websites) .



PROPERTIES OF LARGE-SCALE NETWORKS

III. Community Structure



http://graphstream-project.org/media/other/CSSS2012/gs-communities.html#/home

Community - people in a group interact with each other more frequently than those outside the group

CHALLENGES

Scalability

- The network presented in social media can be huge, often in a scale of millions of actors and hundreds of millions of connections, while traditional social network analysis normally deals with hundreds of subjects or fewer.
- Existing network analysis techniques might fail when applied directly to networks of this astronomical size

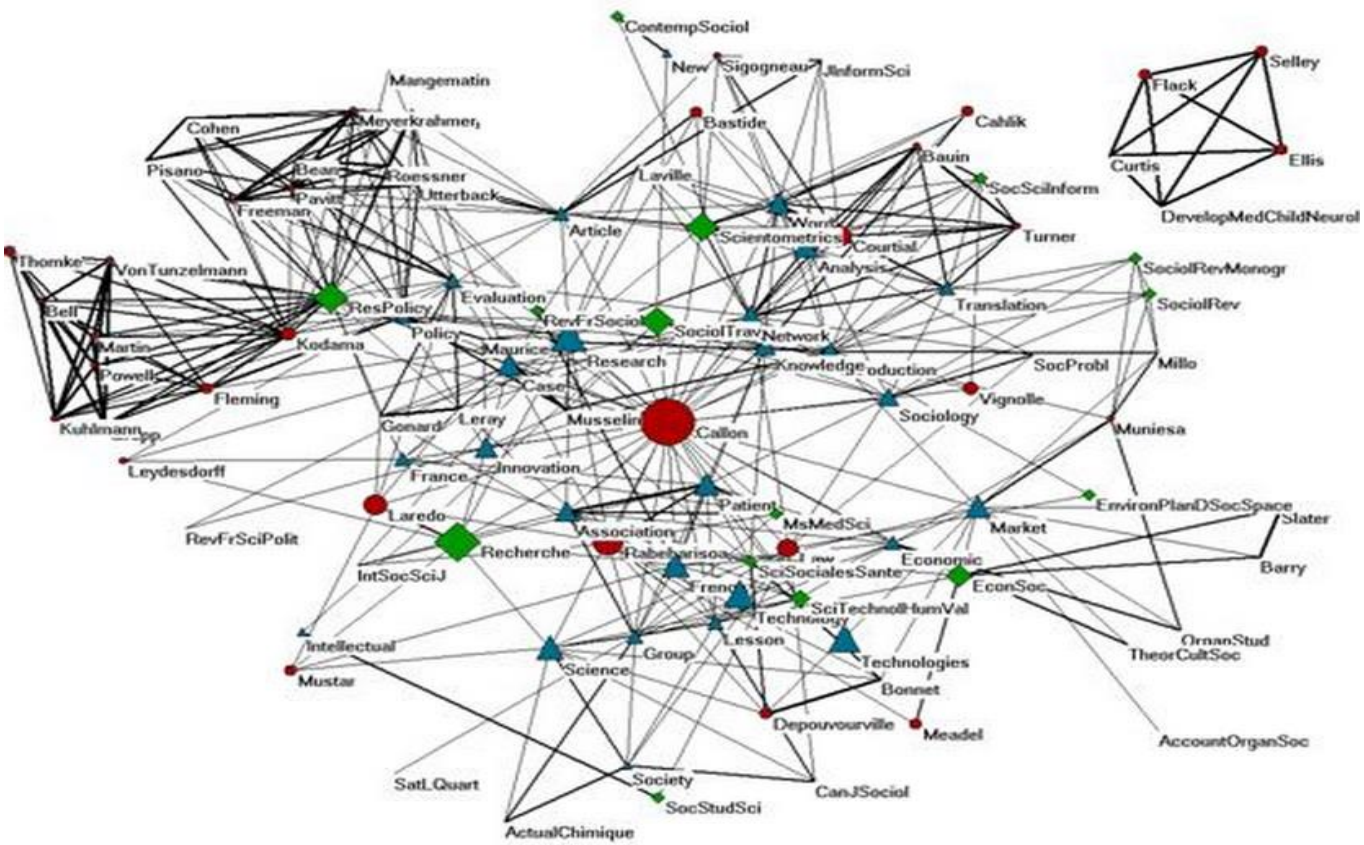
http://www.blogging4jobs.com/social-media/us-social-media-adoption-may-2012/#UKohGCTQdEep0OrB.97

CHALLENGES

Heterogeneity

- In reality, multiple relationships can exist between individuals. Two persons can be friends and colleagues at the same time.
- Thus, a variety of interactions exist between the same set of actors in a network. Multiple types of entities can also be involved in one network.
- For many social bookmarking and media sharing sites, users, tags and content are intertwined with each other, leading to heterogeneous entities in one network.
- Analysis of these heterogeneous networks involving heterogeneous entities or interactions requires new theories and tools.

INTEGRATED MAP OF 48 AUTHORS, 27 WORDS, AND 26 JOURNALS BASED ON 65 PUBLICATIONS OF MICHEL CALLON CONTAINED IN THE ISI DATABASE



http://www.leydesdorff.net/mcallon/index.htm

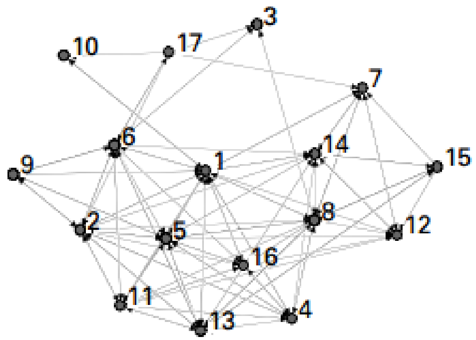
words: ▲; authors: ●; journals: ♦

CHALLENGES

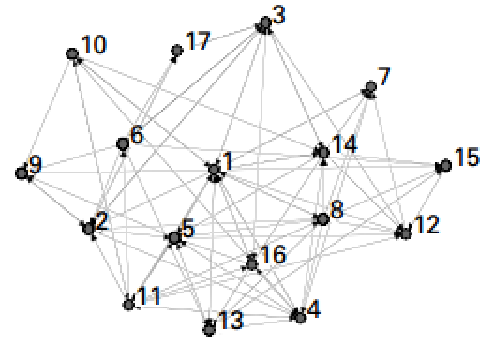
Evolution

- Social media emphasizes timeliness. For example, in content sharing sites and blogosphere, people quickly lose their interest in most shared contents and blog posts.
- This differs from classical web mining. New users join in, new connections establish between existing members, and senior users become dormant or simply leave.
- How can we capture the dynamics of individuals in networks?
- Can we find the die-hard members that are the backbone of communities?
- Can they determine the rise and fall of their communities?

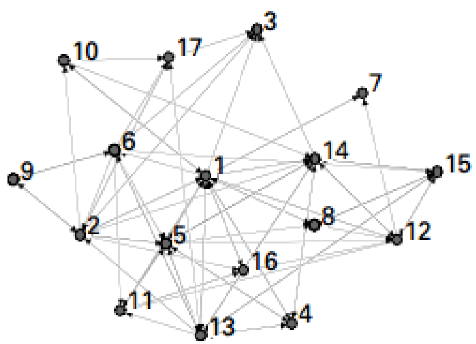
THE TRUST NETWORK AT FOUR POINTS OF TIME



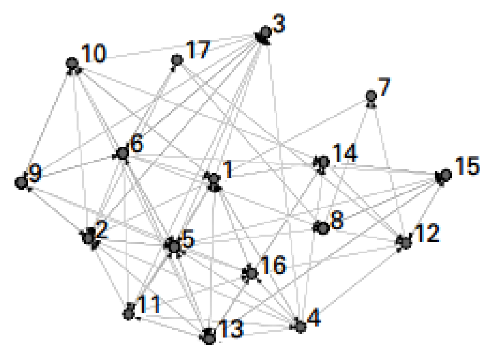
December 1995 (time = t_1)



July 1996 (time = t_2)



December 1996 (time = t_3)



July 1997 (time = t_4)

http://www.s3ri.soton.ac.uk/qmss/documents/Wittek_paper.pdf

Social Computing and Data Mining

Social computing is concerned with the study of Social computing is concerned with the study of social behavior and social context based on computational systems computational systems.

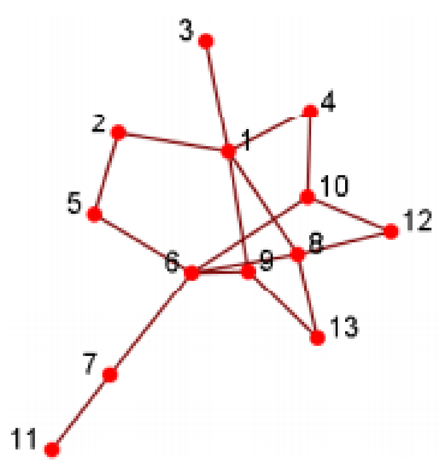
Data Mining Related Tasks:

- ***Centrality Analysis***
- ***Community Detection***
- *Classification*
- *Link Prediction*
- ***Viral Marketing***
- *Network Modeling*

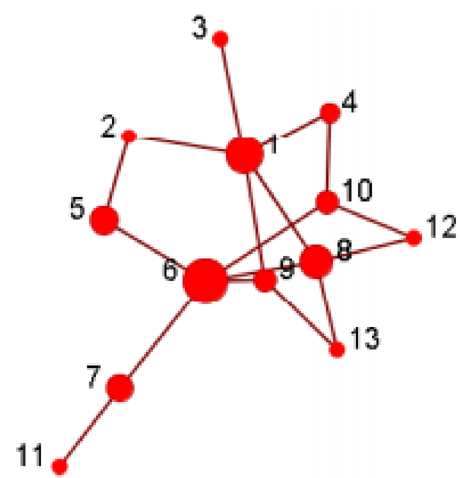
CENTRALITY ANALYSIS/INFLUENCE STUDY

Identify the most important actors in a social network

- **Given**: a social network
- **Output**: a list of top-ranking nodes



Top 5 important nodes: 6,1,8,5,10



Nodes resized by Importance

Community Detection

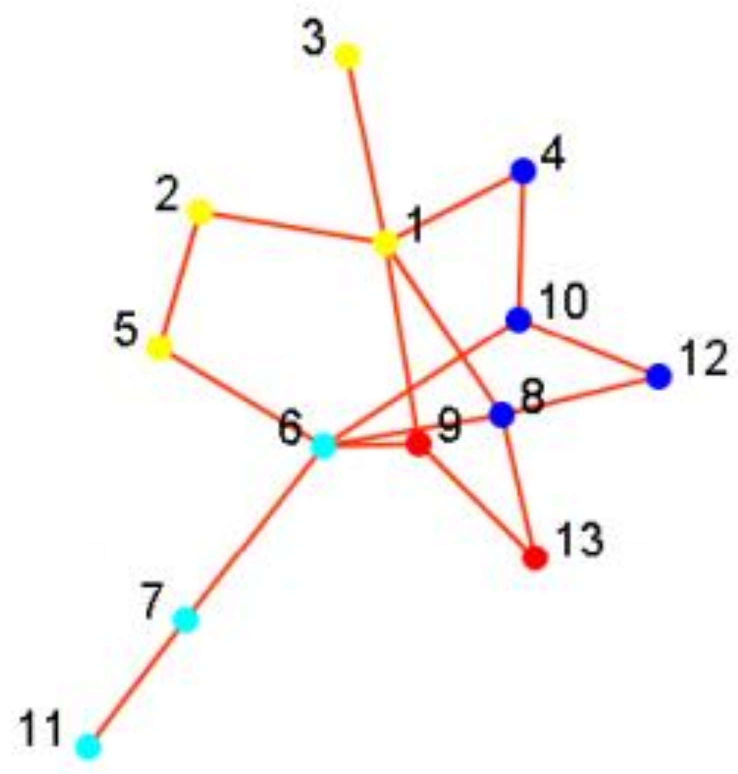
A community is a set of nodes between which the interactions are (relatively frequent

- **Given**: a social network
- **Output**: community membership of (some actors)

Applications

- Understanding the interactions between people
- Visualizing and navigating huge networks
- Forming the basis for other tasks such as data mining

Visualization after Grouping



Nodes colored by Community Membership

CLASSIFICATION

- Classification and recommendation tasks are common in social media applications.

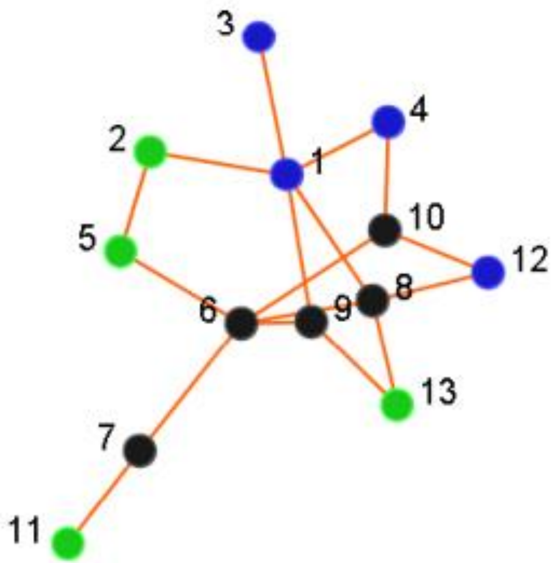
Given

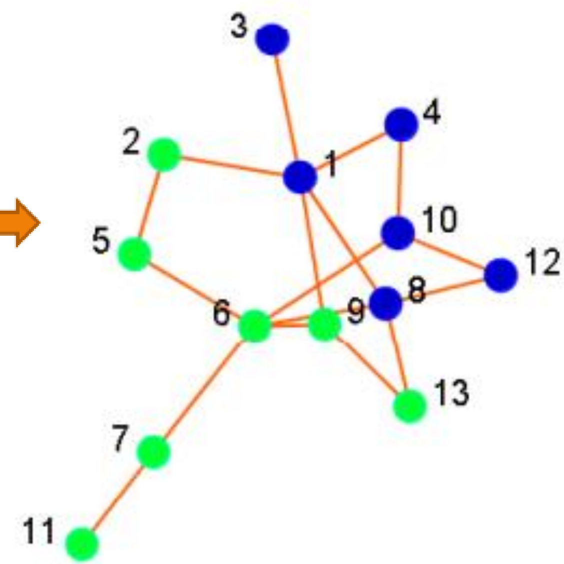
- A social network
- Labels of some actors in the network

Output

- friend recommendation that suggests a list of friends that a user might know
- link prediction (Liben-Nowell and Kleinberg, 2007)
- spam recognition

Visualization after Prediction



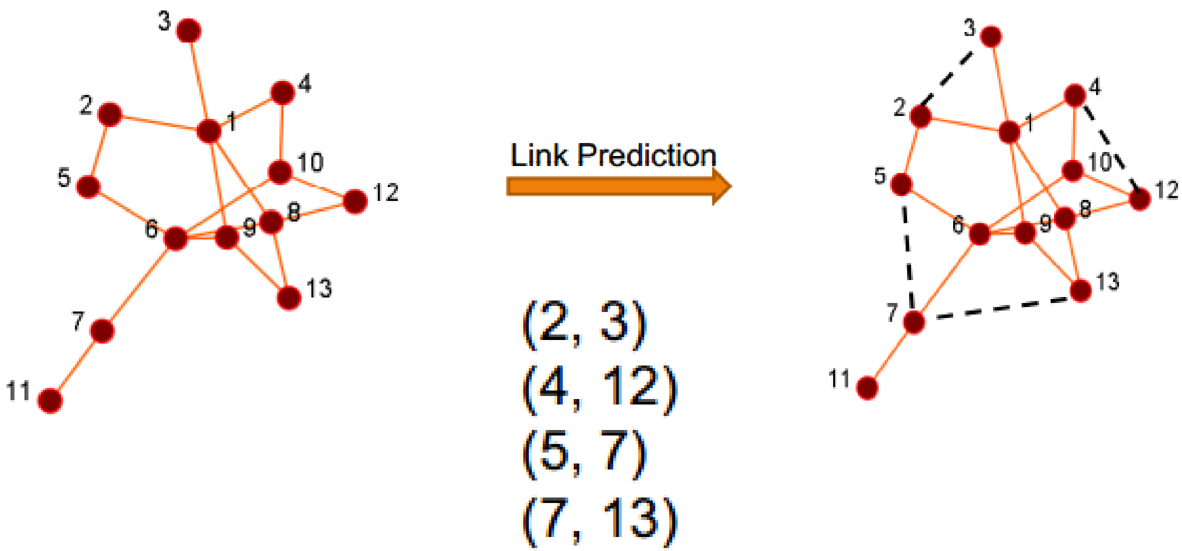


: Smoking
: Non-Smoking
: ? Unknown

Predictions
6: Non-Smoking
7: Non-Smoking
8: Smoking
9: Non-Smoking
10: Smoking

Link Prediction

- Given a social network, predict which nodes are likely to get connected
- **Output** a list of (ranked pairs of nodes)
- **Example**: Friend recommendation in Facebook



1.1. THE HIDDEN INFLUENCE OF SOCIAL NETWORKS

- Critically watch the following short documentaries video about hidden influence of social networks and prepare a 2-3 slides on which phenomena in your life are suitable to be investigated with network theory

http://www.ted.com/talks/nicholas_christakis_the_hidden_influence_of_social_networks?language=en

Viral Marketing/Outbreak Detection

Users have different social capital (or network values) □ within a social network, hence, how can one make best use of this information? use of this information?

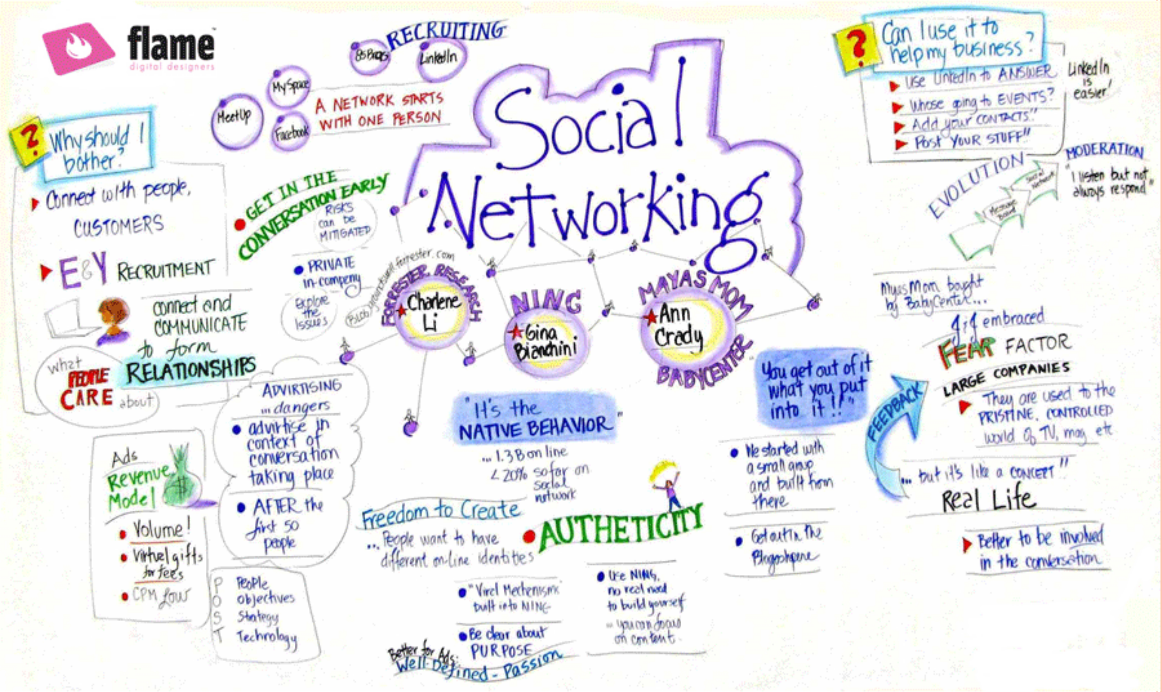
Viral Marketing: find out a set of users to provide coupons and promotions to influence other people in the network so my benefit is maximized

Outbreak Detection: monitor a set of nodes that can help detect outbreaks or interrupt the infection spreading (e.g., H1N1 flu)

□

Goal: given a limited budget, how to maximize the overall benefit?

Walks, Paths, and Trails

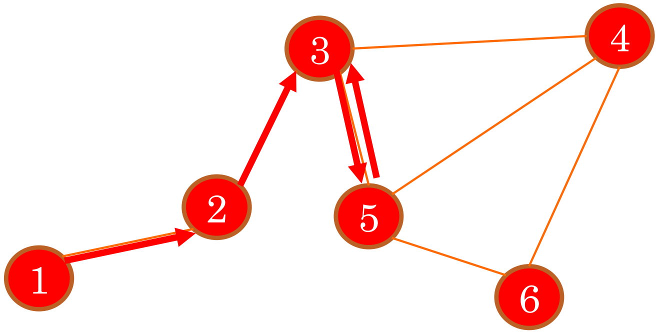


Walks, Trails and Paths

(Directed) WALK (W)	Sequence of nodes and lines starting and ending with different nodes (called origin and terminus). Nodes and lines can be included more than once
INVERSE of a (directed WALK (W^{-1})	Walk in opposite order
LENGTH of a walk	How many lines occur in the walk? (same line counts double, in weighted graphs add line weights)
(Directed) TRAIL	Is a walk in which all lines are distinct
(Directed) PATH	Walk in which all nodes and all lines are distinct

Every path is a trail and every trail is a walk

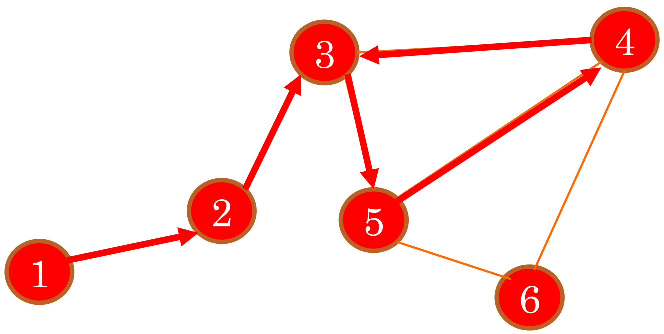
Walks, Trails and Paths



WALK

Sequence of nodes and lines starting and ending with (different nodes)

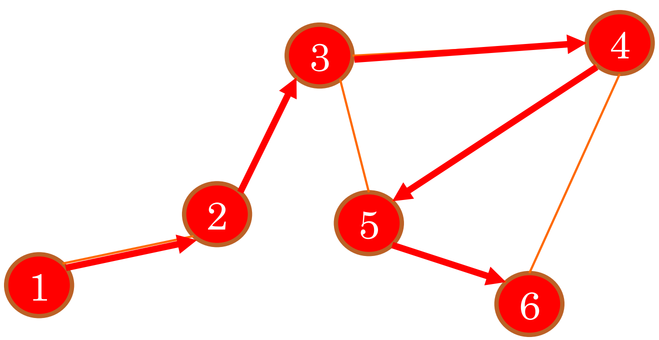
*Origin **n1** and terminus **n3***



TRAIL

Is a walk in which all lines are distinct

*Origin **n1** and terminus **n3***



PATH

Walk in which all nodes and all lines are distinct

*Origin **n1** and terminus **n6***

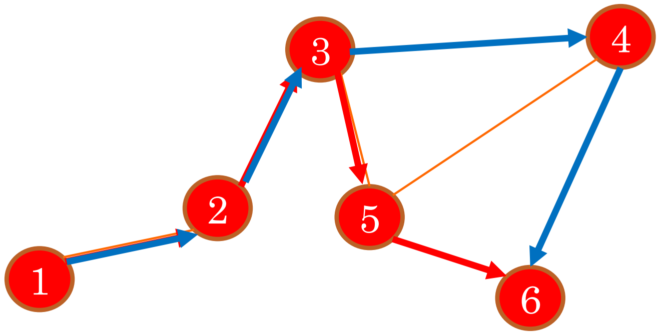
1.2. Topology of flow processes

- Present the examples of the different Topology of flow processes from the "Walks, Trails and Paths" point of view.

	Parallel duplication	**Serial duplication**	**Transfer**
Geodesics			
Path			
Trail			
Walks			

Reachability and Distances

- **REACHABILITY**
  - If there is a path between nodes $\mathbf{n_i}$ and $\mathbf{n_j}$
- **GEODESIC**
  - Shortest path between two nodes
- Geodesic **DISTANCE** $\mathbf{L(n_i,n_j)}$
  - Length of Geodesic (also called „degrees of separation“)



$\mathbf{L(n_1,n_6)=4}$

Shorter paths are desirable when speed of communication or exchange is desired (often the case in many studies, but sometimes not, e.g. in networks that spread disease)

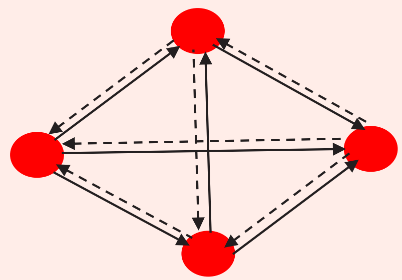
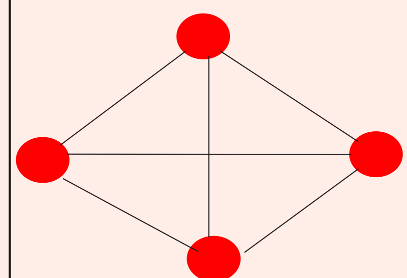
DIAMETER

The **diameter** of a network is the largest distance between any two nodes in the network:

$$\mathbf{Diameter = max\ L(n_i, n_j)}$$

It indicates how long it will take at most to reach any node in the network
(sparser networks will generally have greater diameters)

Graph density

Measure	Directed graph	Undirected graph
Total number of possible edges between all pairs of nodes	$e_{max} = g^*(g-1)$ Each of the **g** nodes can connect to (**g-1**) other nodes	$e_{max} = g^*(g-1)/2$ Since edges are undirected, count each one only once
Example		
Density	$\Delta = e / e_{max}$	

Average Path Length

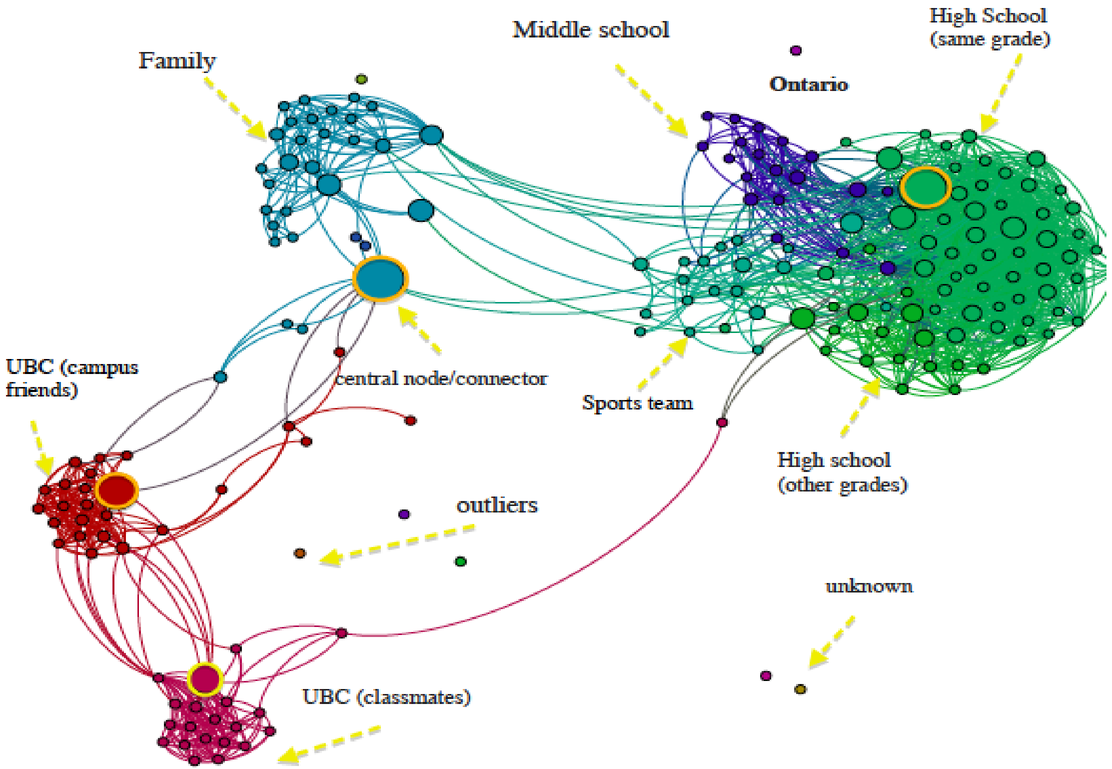
The **average path length** is the average distance between any two nodes in the network:

$$D_{avg}(n_i) = \frac{\sum_{n_i \leq n_j} L(n_i, n_j)}{g(g-1)} \qquad \textit{undirected graph}$$

$$D_{avg}(n_i) = \frac{\sum_{n_i \leq n_j} L(n_i, n_j)}{g(g-1)} \qquad \textit{directed graph}$$

- The *average of all shortest paths* in a network is interesting because it indicates how far apart any two nodes will be on average
- *Average path length* is bounded from above by the diameter; in some cases, it can be much shorter than the diameter.
- If the network is not connected, one often checks the diameter and the *average path length* in the largest component.

Centrality in Social Networks



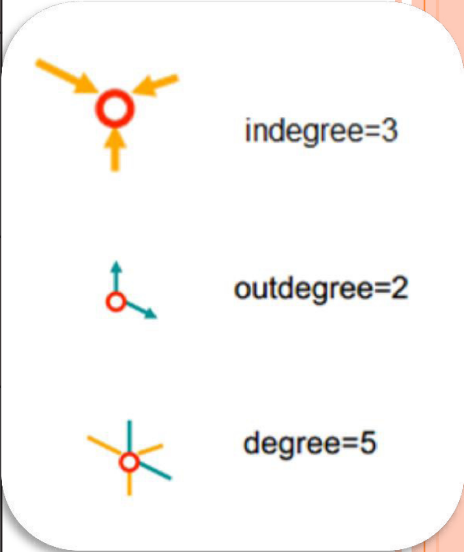
Centrality in Social Networks

In recent work, Borgatti (2003; 2005) discusses centrality in terms of two key dimensions:

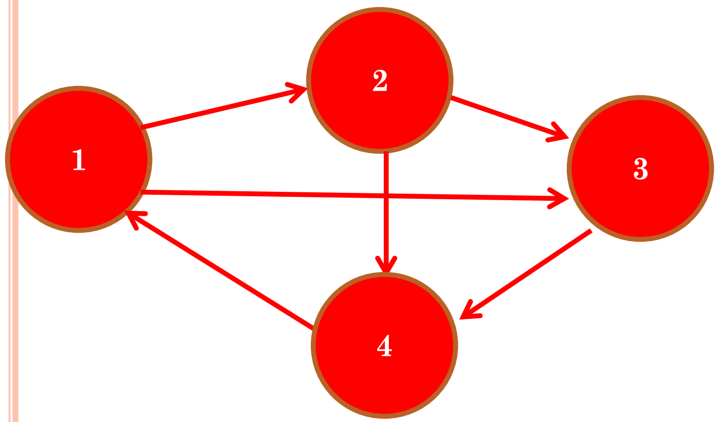
	Radial	Medial
Frequency	**Degree Centrality**	**Betweenness**
Distance	**Closeness Centrality**	(empty: but would be an interruption measure based on distance, but see Borgatti forthcoming)

DEGREE CENTRALITY

From immediate connections	
NEIGHBORHOOD	is the set of nodes that node is connected to
INDEGREE	how many directed edges (arcs are incident on a node)
OUTDEGREE	how many directed edges (arcs originate at a node)
DEGREE (in or out)	number of edges incident/originate node
From the entire graph	
CENTRALITY	betweenness, closeness



Node degree from matrix values



Adjacency matrix (A

Verte x	1	2	3	4
1	0	1	1	0
2	0	0	1	1
3	0	0	0	1
4	1	0	0	0

OUTDEGREE: $d_O(n_i) = \sum_{j=1}^{g} A_{ij}$

INDEGREE: $d_I(n_i) = \sum_{j=1}^{g} A_{ji}$

example: ***outdegree*** for node 3 is **1**, which we obtain by summing the number of non-zero entries in the 3rd row

example: the ***indegree*** for node 3 is **2**, which we obtain by summing the number of non-zero entries in the 3rd column

$$d_O(n_3) = \sum_{j=1}^{4} A_{3j} = 1$$

$$d_I(n_3) = \sum_{j=1}^{g} A_{j3} = 2$$

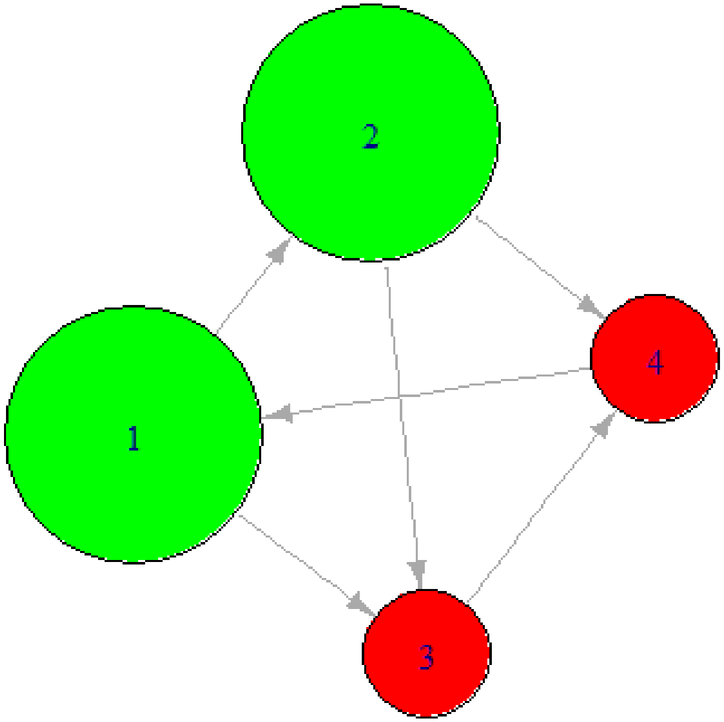
Expansiveness

Popularity, status, deference, degree prestige

Creating Network Graph

> library(igraph)

```
> g <- graph.formula(1-+2,1-+3,2-+3,2-+4,3-+4,4-+1)
> get.adjacency(g)
4 x 4 sparse Matrix of class "dgCMatrix"
  1 2 3 4
1 . 1 1 .
2 . . 1 1
3 . . . 1
4 1 . . .
> V(g)$label<-V(g)
> V(g)$size = degree(g,mode="out")*50
> V(g)$color[V(g)$size == 50] = "red"
> V(g)$color[V(g)$size == 100] = "green"
>  plot(g)
> degree(g)

> di<- degree(g, mode="in")
 1 2 3 4
 3 3 3 3
> do<-degree(g, mode="out")
1 2 3 4
1 1 2 2

1 2 3 4
2 2 1 1
```



DEGREE CENTRALITY

- Interpretation: opportunity to (be influenced)

- Classification of Nodes
  - **Isolates**

$$d_O(n_i) = d_I(n_i) = 0$$

  - **Sender**

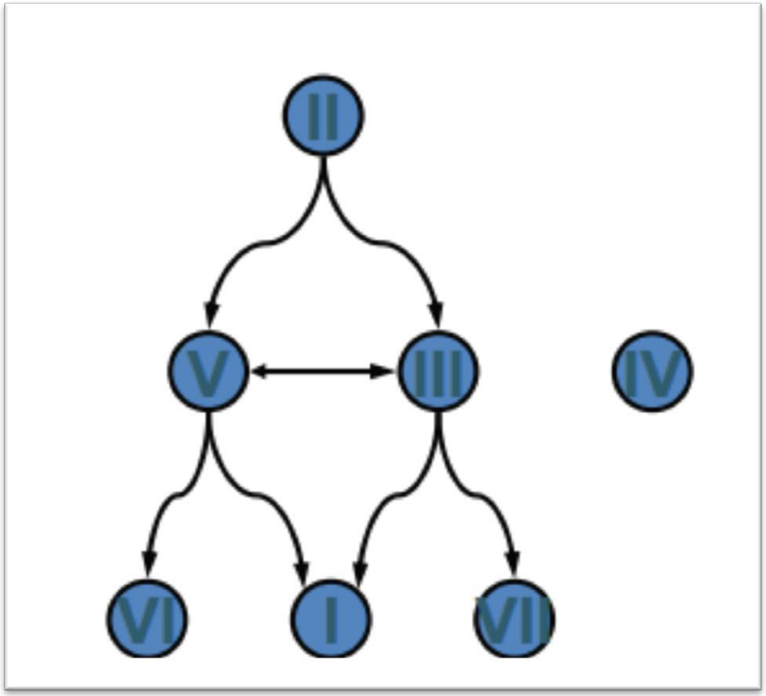
$$d_O(n_i) > 0 \; and \; d_I(n_i) = 0$$

  - **Receivers**

$$d_O(n_i) = 0 \; and \; d_I(n_i) > 0$$

  - **Ordinaries**

$$d_O(n_i) > 0 \; and \; d_I(n_i) > 0$$



50

Degree Centrality

- **Standardization** of d_{oi} and d_{Ii} to allow comparison of the nodes across networks of different sizes:

Directed graph

divide by the max. possible

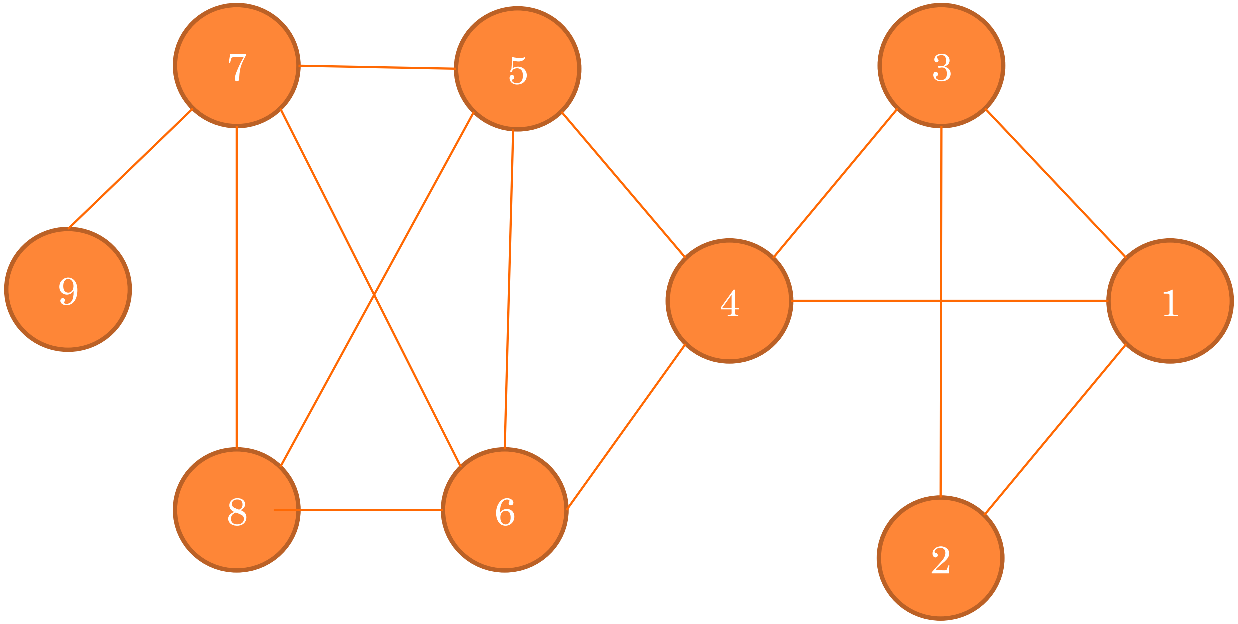
$$C_{DO}(n_i) = \frac{d_{Oi}}{g-1}$$

$$C_{DI}(n_i) = \frac{d_{Ii}}{g-1}$$

Undirected graph

$$C_D(n_i) = \frac{d_i}{g-1}$$

Degree Centrality



$$d_1 = \sum_{j=1}^{9} A_{1j} = 3$$

$$C_D(n_1) = \frac{d_1}{g-1} = \frac{3}{9-1} = \frac{3}{8}$$

DEGREE CENTRALIZATION: HOW EQUAL ARE THE NODES?

How much **variation** is there in the centrality scores among the nodes?

Freeman's general formula for centralization (can use other metrics, e.g. standard deviation :

Maximum value in the network

$$C_D = \frac{\sum_{i=1}^{g} \left(\boxed{C'_D(n^*)} - C'_D(n_i) \right)}{(g-1)\cdot(g-2)}$$

DEGREE CENTRALITY

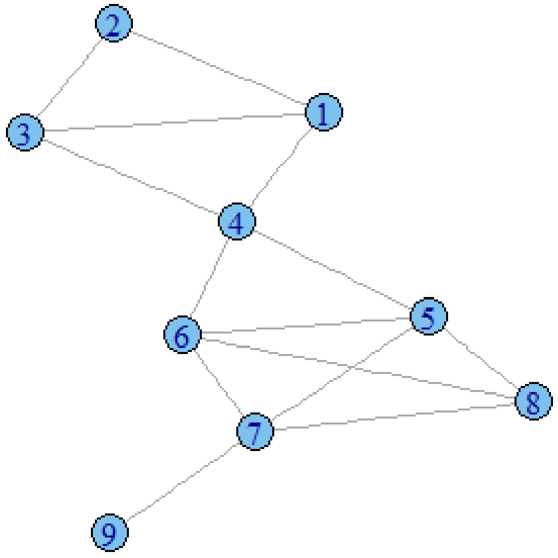
```
>  g <- graph.formula(1-2,2-3,1-3,3-4,4-1,4-5,4-6,5-7,5-6,5-8,6-7,6-8,7-8,7-9)
> d<- degree(g)
1 3 2 4 5 6 7 8 9
3 3 2 4 4 4 4 3 1
```

 # normalized vertex degree - degree centrality

```
> n <- vcount(g)
> n
[1] 9
> degrees <- d/(n-1)
> degrees
    1     3     2     4     5     6     7     8     9
0.375 0.375 0.250 0.500 0.500 0.500 0.500 0.375 0.125
```

 # degree centralization

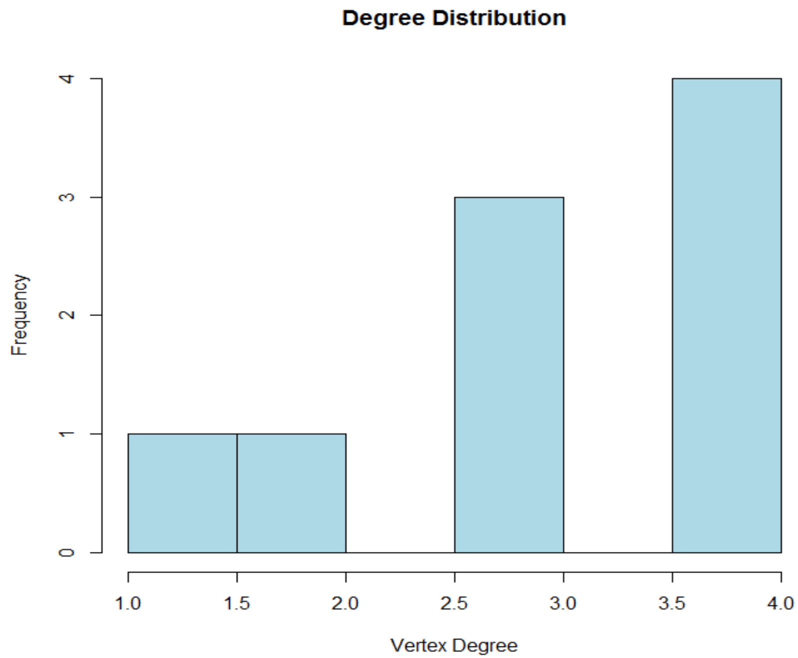
```
> numerator <- sum(max(degrees) - degrees)
> denominator <- (n-1)*(n-2)
> degreec<-numerator/denominator
> degreec
[1] 0.01785714
```



Shortest path

```
> shortest.paths(g)
  1 2 3 4 5 6 7 8 9
1 0 1 1 1 2 2 3 3 4
2 1 0 1 2 3 3 4 4 5
3 1 1 0 1 2 2 3 3 4
4 1 2 1 0 1 1 2 2 3
5 2 3 2 1 0 1 1 1 2
6 2 3 2 1 1 0 1 1 2
7 3 4 3 2 1 1 0 1 1
8 3 4 3 2 1 1 1 0 2
9 4 5 4 3 2 2 1 2 0
```

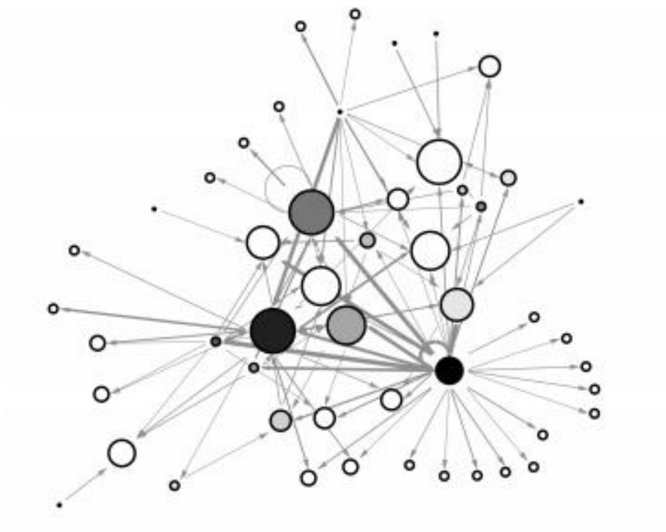
```
> f<-get.shortest.paths(g,2,5)
> f
$vpath
$vpath[[1]]
[1] 2 1 4 5
```




```
> hist(degree(g), col="lightblue",  xlim=c(min(degree(g)),max(degree(g))),
+ xlab="Vertex Degree", ylab="Frequency",  main="Degree Distribution")

> p<-graph.density(g)
>p
[1] 0.3888889
```

DEGREE CENTRALIZATION EXAMPLES

Example financial trading networks



high *centralization: one node trading with many others*

low *centralization: trades are more evenly distributed*

CLOSENESS CENTRALITY: DEFINITION

- **The closeness centrality** of a vertex is based on the total distance between one vertex and all other vertices, where larger distances yield lower closeness centrality scores
- The closer a vertex is to all other vertices, the easier information may reach it, the **higher** its **centrality**

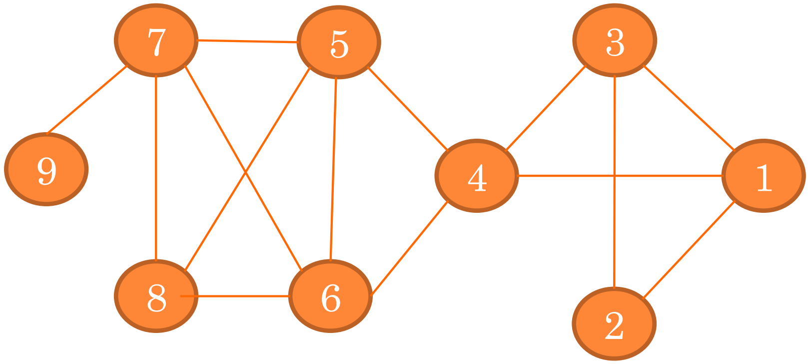
Closeness Centrality:
Index of expected arrival time

$$C_C(n_i) = \frac{1}{\sum_{i=1}^{g} L(n_i, n_j)}$$

Normalized Closeness Centrality

$$C'_C(n_i) = \frac{C_C(n_i)}{g-1}$$

CLOSENESS CENTRALITY: EXAMPLE



$$C_C(n_i) = \frac{1}{\sum_{i=1}^{g} L(n_i, n_j)}$$

$$C'_C(n_i) = \frac{C_C(n_i)}{g-1}$$

	1	2	3	4	5	6	7	8	9	Σ
1	-	1	1	1	2	2	3	3	4	17
2	1	-	1	2	3	3	4	4	5	23
3	1	1	-	1	2	2	3	3	4	17
4	1	2	1	-	1	1	2	2	3	13
5	2	3	2	1	-	1	1	1	2	13
6	2	3	2	1	1	-	1	1	2	13
7	3	4	3	2	1	1	-	1	1	16
8	3	4	3	2	1	1	1	-	2	17
9	4	5	4	3	2	2	1	2	-	17
	17	23	17	13	14	14	16	17	17	

Reciprocal of marginals of geodesic distance matrix

$$C'_C(n_3) = \frac{1}{17} \cdot 8 = 0.47$$

$$C'_C(n_4) = \frac{1}{13} \cdot 8 = 0.62$$

Closeness Centralization

Maximum value in the network

$$C_D = \frac{\sum_{i=1}^{g} \left(C'_C(n^*) - C'_C(n_i) \right)}{\boxed{(g-1)\cdot(g-2)\cdot(2g-3)}}$$

59

Closeness centrality: example

> cl <-closeness(g) * (vcount(g)-1)

```
> cl
        1         3         2         4         5         6         7         8 
0.4705882 0.4705882 0.3478261 0.6153846 0.6153846 0.6153846 0.5000000 0.4705882 
        9 
0.3478261 
```

normalized vertex closeness

> cln<- closeness(g)

```
>  cln
         1          3          2          4          5          6          7 
0.05882353 0.05882353 0.04347826 0.07692308 0.07692308 0.07692308 0.06250000 
         8          9 
0.05882353 0.04347826 
```

closeness centralization

> numerator <- sum(max(cln)-cln)

```
>  numerator
[1]  0.1356114
```

> denominator <- (n-2)*(n-1)/(2*n-3)

```
>  denominator
[1]  3.733
```

> clc<-numerator/denominator

```
> clc
[1]  0.03632447
```

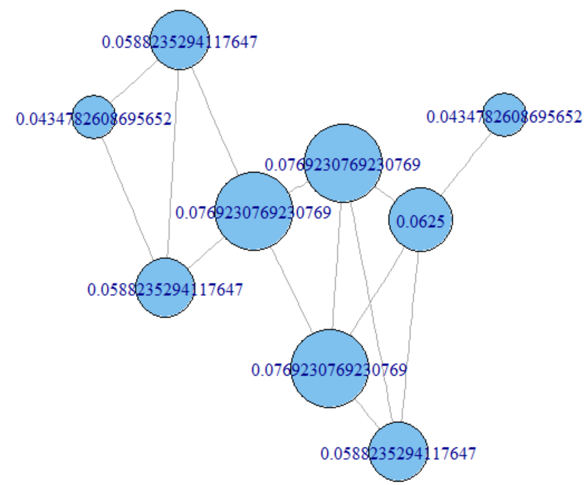
> V(g)$label<-closeness(g)

> V(g)$size = closeness(g)*500

> plot(g)

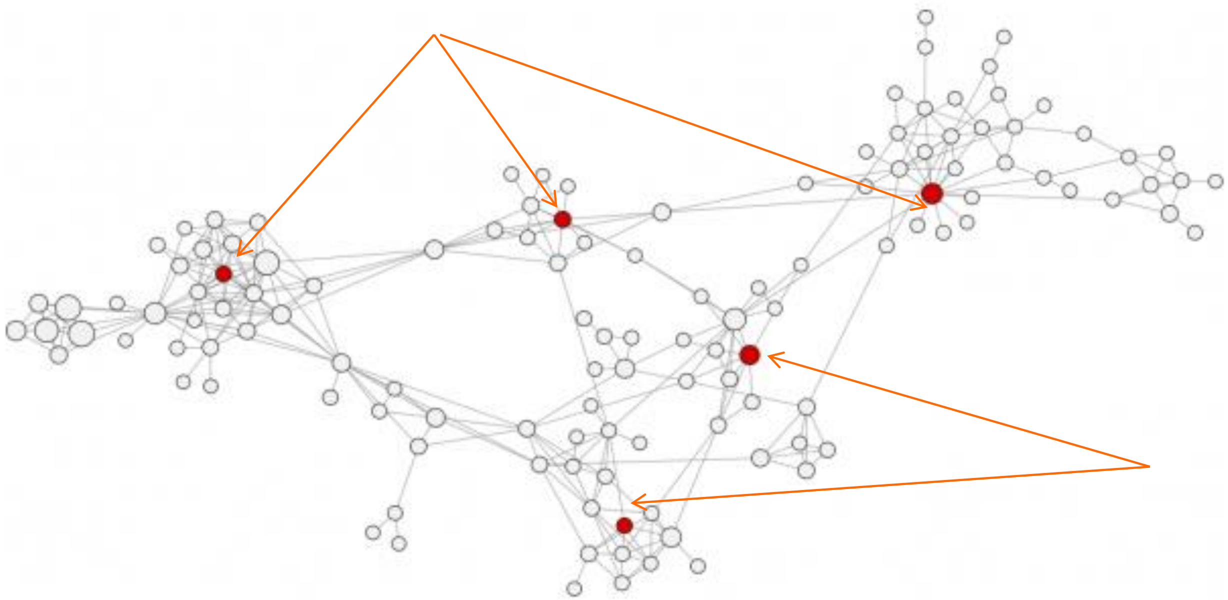
> graph.density(g)

[1] 0.3888889



APPLICATIONS

- High closeness centrality individuals tend to be important influencers within their local network community.
- They may often **not be public** figures to the entire network of a corporation or profession, but they are often respected locally and they occupy short paths for information spread within their network community.



Individuals who are highly connected to others within their own cluster will have a high closeness centrality.

BETWEENNESS CENTRALITY

Degree and ***closeness*** centrality are based on the reachability of a person within a network:

How easily can information reach a person?

A next approach to centrality and centralization rests on the idea that a person is more central if he or she is more important as an intermediary in the communication network:

1. ***How crucial is a person to the transmission of information through a network?***
2. ***How many flows of information are disrupted or must make longer detours if a person stops passing on information or disappears from the network?***
3. ***To what extent may a person control the flow of information due to his or her position in the communication network?***

Betweenness Centrality

How many *geodesic linkings* between two actors $\mathbf{n_j}$ and $\mathbf{n_k}$ contain actor $\mathbf{n_i}$?

Now let:

$\boldsymbol{g_{ik}}$ - the number of geodesics from $\mathbf{n_j}$ to $\mathbf{n_k}$ and

$\boldsymbol{g_{jk}(n_i)}$ - the number of geodesics that contain point $\mathbf{n_i}$ as an intermediary in the geodesics from $\mathbf{n_j}$ to $\mathbf{n_k}$, then:

$\frac{g_{jk}(n_j)}{g_{jk}}$ is the probability that distinct actor $\mathbf{n_i}$ „involved“ in communication between two actors $\mathbf{n_j}$ and $\mathbf{n_k}$

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_j)}{g_{jk}}$$

BETWEENNESS CENTRALITY

Usually normalized by:

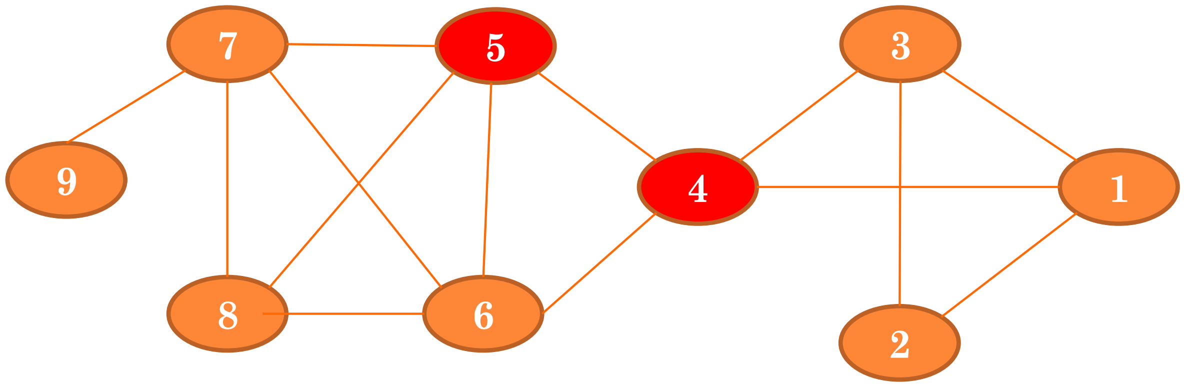
$$C'_B(n_i) = \frac{C_B(n_i)}{[\boxed{(g-1)(g-2)}/2]}$$

number of pairs of vertices excluding the vertex itself

Betweennes Centralization:

$$C_D = \frac{\sum_{i=1}^{g} C_B(n^*) - C_B(n_i)}{(g-1)\cdot(g-2)^2}$$

Betweenness Centrality



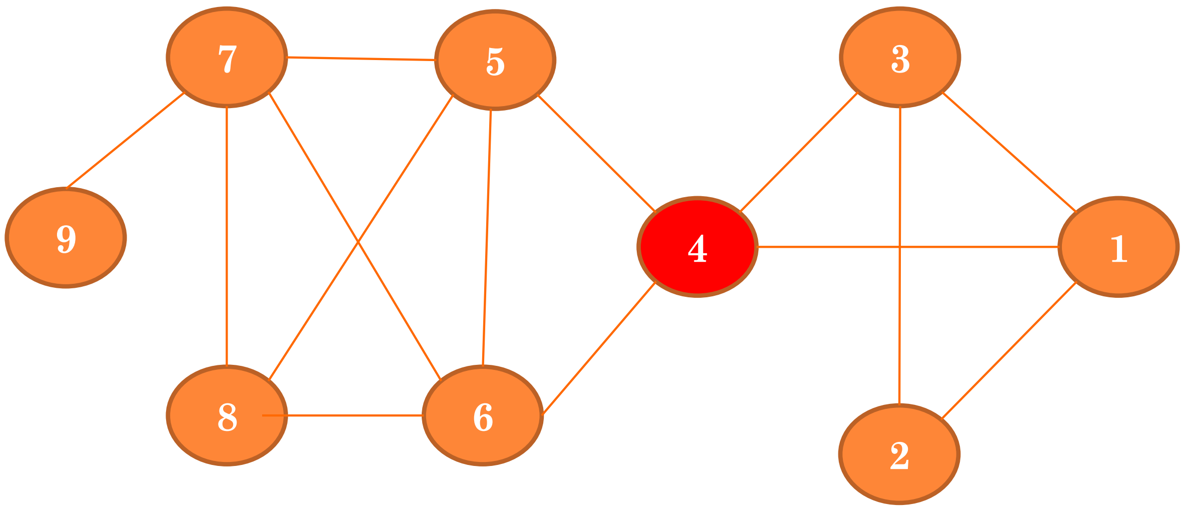
$$\sum \frac{g_{jk}(n_4)}{g_{jk}} = 15$$

$$\sum \frac{g_{jk}(n_5)}{g_{jk}} = 6$$

	k=1	*k=2*	*k=3*
j=5	1/1	2/2	1/1
j=6	1/1	2/2	1/1
j=7	2/2	4/4	2/2
j=8	2/2	4/4	2/2
j=9	2/2	4/4	2/2

	k=1	*k=2*	*k=3*
j=5	1/1	2/2	1/1
j=6	1/1	2/2	1/1
j=7	2/2	4/4	2/2
j=8	2/2	4/4	2/2
j=9	2/2	4/4	2/2

Betweenness Centrality



$C_B(n_i)$	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
	3	*0*	*3*	*15*	*6*	*6*	*7*	*0*	*0*

V(g)$label<-paste("Actor ",V(g)," (", betweenness(g),")")

BETWEENNESS CENTRALITY

- The more a person is a go-between, the more central his or her position in the network.
- If we consider the geodesics to be the most likely channels for transporting information between actors, an actor who is situated on the geodesics between many pairs of vertices is very important to the flow of information within the network.
- This actor is more central.

APPLICATIONS

- High betweenness individuals are often critical to collaboration across departments and to maintaining the spread of a new product through an entire network. Because of their locations between network communities, they are natural brokers of information and collaboration.
- One difference between high betweenness individuals in a network and **actual brokers** is the latter usually have a public profile as part of their business, whereas high betweenness individuals often are overlooked. This occurs because they are not central to any single social clique, and instead reside on the periphery of several such cliques each of which all engender more trust and admiration within rather than outside of the clique.

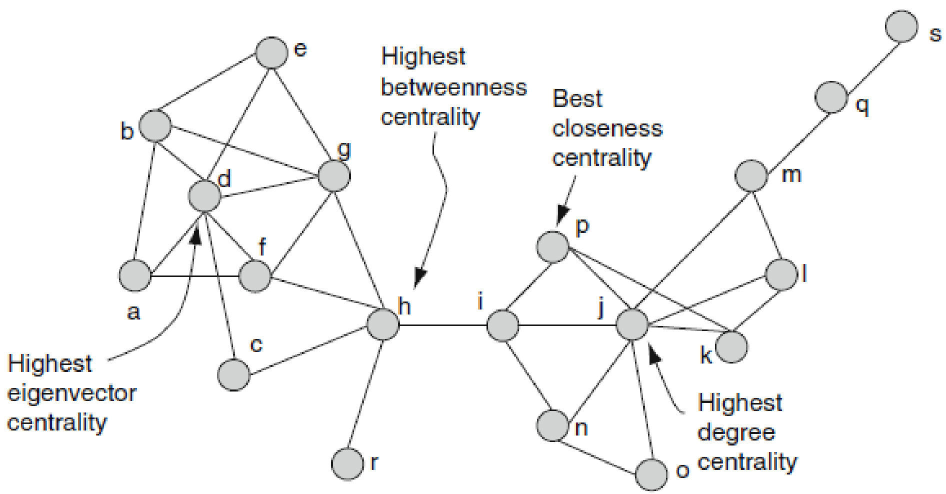
Those who act as bridges between clusters in the network have high betweenness centrality.

INTERPRETATION OF MEASURES

Centrality measure	Interpretation in social networks
DEGREE	How many people can this person reach directly?
CLOSENESS	Who has the shortest distance to the other actors?
BETWEENNESS	Who controls knowledge flows?

EIGENVECTOR CENTRALITY

- The underlying idea beneath eigenvector centrality is that one's importance is defined by his friends' importance.
- In other words, if one has many important friends, he should also be important.
- More precisely, **centrality of a node is proportional to the sum of scores of its neighbors**



EIGENVECTOR CENTRALITY

In particular:

$$\boxed{C_E(v_i) \alpha \sum_{v_j \in N_i} A_{ij} \cdot C_E(v_j)}$$

Let **x** denote the eigenvector centrality of node from $\boldsymbol{v_l}$ to $\boldsymbol{v_n}$. The above equation can be written as in a matrix form

$$x \,\alpha\, Ax$$

Equivalently, we can write $x = \frac{1}{\chi} Ax$, where λ is a constant. It follows that

$$Ax = \lambda x$$

*An **eigenvector** of a square **adjacency** matrix A is a non-zero vector x that, when the matrix multiplies x, yields a constant multiple of x, the latter multiplier being commonly denoted by λ.*

*The number λ is called the **eigenvalue** of corresponding to x*

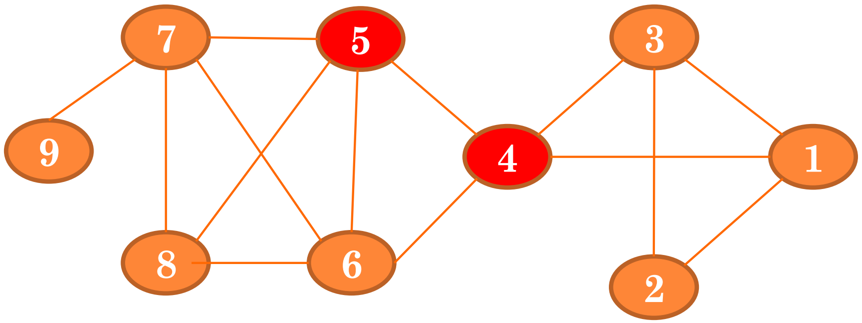
Eigenvector Centrality

Google's Pagerank (Page et al., 1999 is a variant of the eigenvector centrality. In Pagerank, a **transition** matrix is constructed based on the adjacency matrix by normalizing each column to a sum of 1:

$$\widetilde{A}_{ij} = {A_{ij}} \Big/ {\sum_i A_{ij}}$$

- In the transition matrix an entry $\widetilde{A}_{ij}$ denotes the **probability** of transition from node $\boldsymbol{v_j}$ to node $\boldsymbol{v_i}$.

Eigenvector Centrality



Column-Normalized Adjacency Matrix

	1	2	3	4	5	6	7	8	9
1	0	1/2	1/3	1/4	0	0	0	0	0
2	1/3	0	1/3	0	0	0	0	0	0
3	1/3	1/2	0	1/4	0	0	0	0	0
4	1/3	0	1/3	0	1/4	1/4	0	0	0
5	0	0	0	1/4	0	1/4	1/4	1/3	0
6	0	0	0	1/4	1/4	0	1/4	1/3	0
7	0	0	0	0	1/4	1/4	0	1/3	1
8	0	0	0	0	1/4	1/4	1/4	0	0
9	0	0	0	0	0	0	1/4	0	0

EIGENVECTOR CENTRALITY

- Pagerank scores correspond to the top eigenvector of the transition matrix $\tilde{A}_{ij}$
- It can be computed by the power method, i.e., repeatedly left-multiplying a non-negative vector ***x*** with $\tilde{A}_{ij}$.
- Suppose we start from $\boldsymbol{x^{(0)}} = \boldsymbol{1}$

Then

$$x^{(1)} \alpha \, \tilde{A}_{ij} \cdot x^{(0)}$$

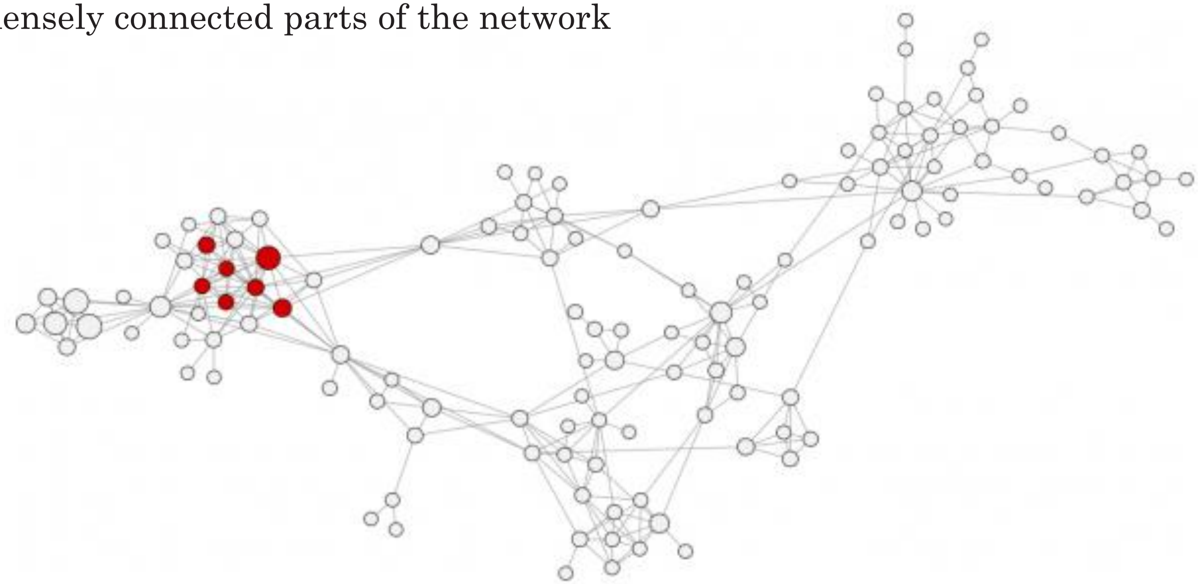
$$x^{(2)} \alpha \, \tilde{A}_{ij} \cdot x^{(1)}$$

....

evcent(g)$vector

APPLICATIONS

- High eigenvector centrality individuals are leaders of the network. They are **often public figures** with many connections to other high-profile individuals. Thus, they often play roles of key opinion leaders and shape public perception. A related example of this is Google's page rank algorithm, which is closely related to eigenvector centrality calculated on websites based on links to them.
- High eigenvector centrality individuals, however, cannot necessarily perform the roles of high closeness and betweenness. They do not always have the greatest local influence and may have limited brokering potential. Like an aloof king in his court or CEO in her boardroom, they may at times be isolated from peripheral individuals and smaller network communities that have limited connectivity with the most densely connected parts of the network



Highly connected individuals within highly interconnected clusters, or 'big fish in big ponds', have high eigenvector centrality

Large-scale networks

Centrality Measures	Time Complexity	
Closeness Centrality	**$O(n^2)$**	*Floyd-Warshall algorithm (Floyd,1962)*
	$O(n^2 \log n + nm)$	*Johnson's algorithm (Johnson, 1977)*
Betweenness Centrality	**$O(nm)$**	*Brandes, 2001*
Eigenvector Centrality	**$O(ml)$** *l is the number of iterations*	*Power method (Golub and Van Loan,1996)*

1.3. BASIC NETWORK MEASURES

1. Create the two random graphs: directed and undirected (not less than 10 nodes)
2. Apply the during class learned basic measures on the provided data set. (Please consider the following measures: number of vertices, number of edges, average path length, degree, closeness, betweenness, eigenvector). What can you say about the network? Please provide your results in a **table**.
3. Calculate the centralization measures for the **undirected** network (we discussed these measures during the lecture). Interpret your results.
4. Complete the task “Centrality: check your understanding” (see next slide)
5. Visualize the network by using one of the shown functions. Indicate the measures of each vertex by the size of the node.
6. Formulate not less then 3 questions for your neighbor about her/his results.
7. Make a discussion

CENTRALITY: CHECK YOUR UNDERSTANDING

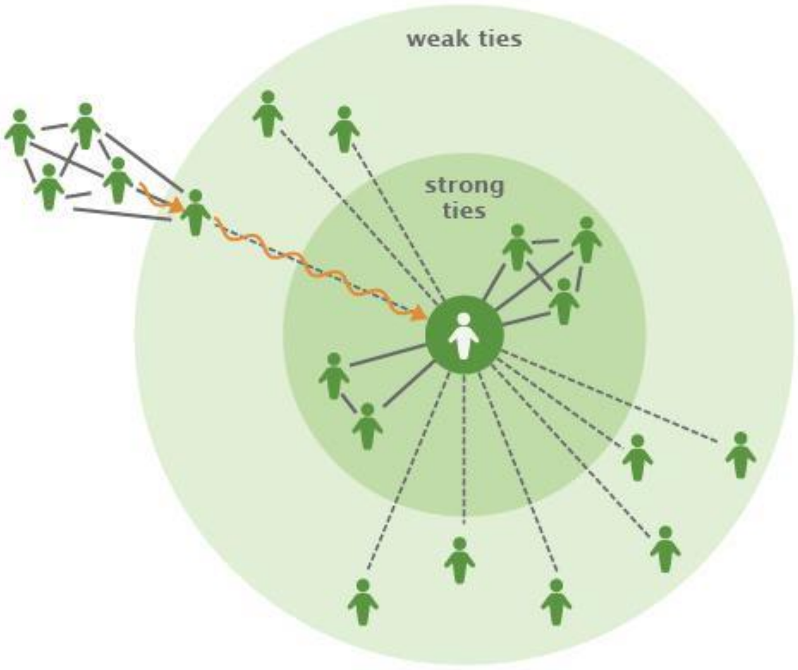
- generally different centrality metrics will be positively correlated
- when they are not, there is likely something interesting about the network
- suggest possible topologies and node positions to fit each square

	Low Degree	Low Closeness	Low Betweenness
High Degree	-		
High Closeness		-	
High Betweenness			-

Assistants to the assignment 1

Number of vertices in the graph	vcount(g)
Number of edges in the graph	ecount(g)
Graph Density	graph.density(g)
List of all nodes	V(g)
List of all vertex attributes names	list.vertex.attributes(g)
List of all edges/arcs	E(g)
Checks if the graph is directed	is.directed()
Shortest Path	shortest.paths(graph, v=V(graph))
Degree	degree(g)
Normalized Vertex Closeness Centrality	closeness(graph, v=V(graph), mode = "all") closeness(graph, mode="out") (for directed graph)
Betweenness Centrality	betweenness(graph, v=V(graph), directed = TRUE)
Sum, max	sum(), max()
Eigenvector Centrality	evcent(g)$vector

STRENGTHS OF TIES



As defined by Granovetter (1973) , "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) , and the reciprocal services which characterize the tie."

STRENGTHS OF TIES

Social Media allows users to connect to each other more easily than ever :

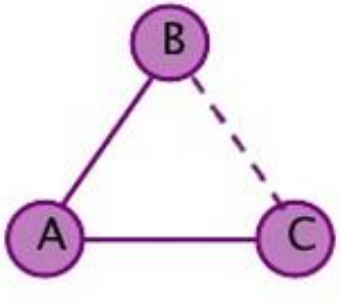
- One user might have thousands of friends online
- Who are the most important ones among your 300 Facebook friends?‘

Three chief approaches to **estimate the strengths of ties** for advanced analysis:

1) Analyzing network topology
2) Learning from user attributes
3) Learning from user activities

Learning from Network Topology

Probability of tie between B and C is positively related with tie strength of A-B and A-C.



- Frequency of Interaction (Homan)

More frequent interaction = more friendship

A and B are together 60% of the time

A and C are together 40% of the time

A, B and C will be together 24% (0.6 × 0.4) of the time

- Similarity (Precker, Newcomb and others)

Stronger tie = similarity between persons

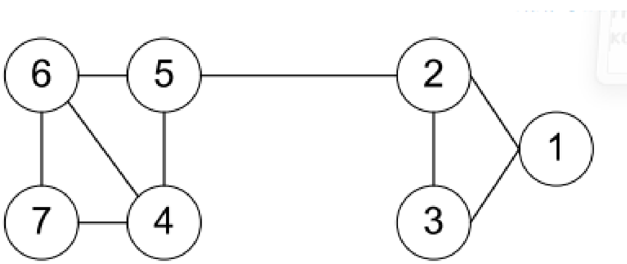
- Cognitive balance (Heider, Newcomb)

C and B want congruence of feelings with A

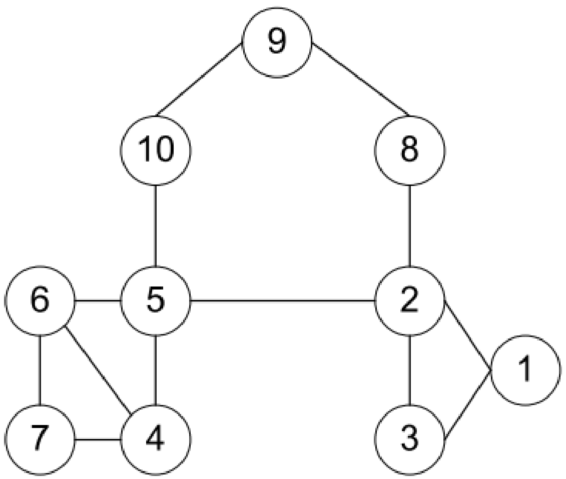
- If A-B and A-C are strong ties then by **transitivity** B and C have a weak tie at least. The Triad A-B, A-C with absent tie B-C is not to be expected.

Learning from Network Topology

- And edge that joins two nodes 2 and 5 in a graph is called a **bridge** if deleting the edge would cause 2 and 5 to lie in two different components
- 2 and 5 in a graph is a **local bridge** if its endpoints 2 and 5 have no friends in common — in other words, if deleting the edge would increase the distance between 2 and 5 to a value strictly more than two



After its removal, nodes 2 and 5 become disconnected



A network in which the removal of e(2,5) increases the geodesic distance between nodes 2 and 5 to 4.

Learning from Network Topology

> ggg <- graph.formula(1-2,1-3,2-3,4-5, 4-6,4-7,2-5,5-6,6-7,2-8,8-9,9-10,10-5)

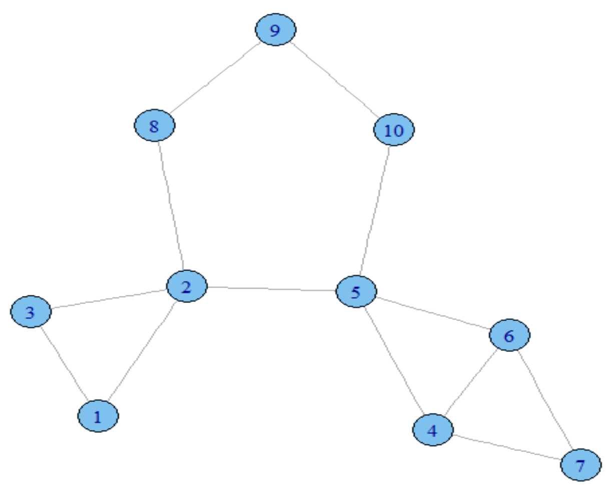
```
> shortest.paths(ggg)
   9 10  8  5  6  4  7  2  1  3
9  0  1  1  2  3  3  4  2  3  3
10 1  0  2  1  2  2  3  2  3  3
8  1  2  0  2  3  3  4  1  2  2
5  2  1  2  0  1  1  2  1  2  2
6  3  2  3  1  0  1  1  2  3  3
4  3  2  3  1  1  0  1  2  3  3
7  4  3  4  2  1  1  0  3  4  4
2  2  2  1  1  2  2  3  0  1  1
1  3  3  2  2  3  3  4  1  0  1
3  3  3  2  2  3  3  4  1  1  0
```

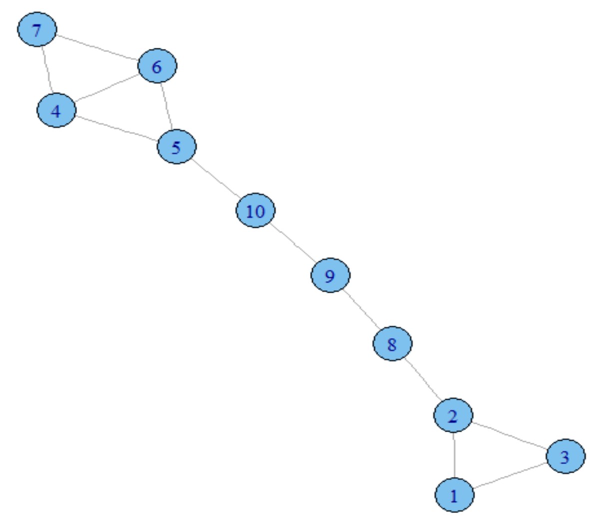
```
> E(ggg)
Edge sequence:

[1]  2  -- 1
[2]  3  -- 1
[3]  3  -- 2
[4]  5  -- 2
[5]  8  -- 2
[6]  10 -- 5
[7]  4  -- 5
[8]  6  -- 5
[9]  9  -- 8
[10] 10 -- 9
[11] 6  -- 4
[12] 7  -- 4
[13] 7  -- 6
```

> ggg1<-ggg – edges(4)

```
> shortest.paths(ggg1)
   1  2  3  5  8  9 10  4  6  7
1  0  1  1  5  2  3  4  6  6  7
2  1  0  1  4  1  2  3  5  5  6
3  1  1  0  5  2  3  4  6  6  7
5  5  4  5  0  3  2  1  1  1  2
8  2  1  2  3  0  1  2  4  4  5
9  3  2  3  2  1  0  1  3  3  4
10 4  3  4  1  2  1  0  2  2  3
4  6  5  6  1  4  3  2  0  1  1
6  6  5  6  1  4  3  2  1  0  1
7  7  6  7  2  5  4  3  1  1  0
```

```
> E(ggg1)
Edge sequence:

[1]  2  -- 1
[2]  3  -- 1
[3]  3  -- 2
[4]  8  -- 2
[5]  10 -- 5
[6]  4  -- 5
[7]  6  -- 5
[8]  9  -- 8
[9]  10 -- 9
[10] 6  -- 4
[11] 7  -- 4
[12] 7  -- 6
```





“SHORTCUT” BRIDGE

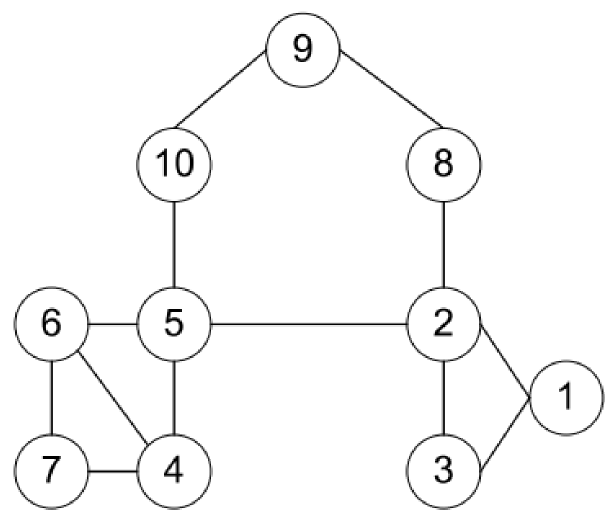
- The larger the distance, the weaker the tie is

- d(2,5) = 4 if e(2,5) is removed
- d(5,6) = 2 if e(5,6) is removed
- e(5,6) is a stronger tie than e(2,5)

> ggg2<-ggg-edges(8)

```
> E(ggg)
Edge sequence:

[1]  2  -- 1
[2]  3  -- 1
[3]  3  -- 2
[4]  5  -- 2
[5]  8  -- 2
[6]  10 -- 5
[7]  4  -- 5
[8]  6  -- 5
[9]  9  -- 8
[10] 10 -- 9
[11] 6  -- 4
[12] 7  -- 4
[13] 7  -- 6
```

```
> shortest.paths(ggg2)
   1 2 3 5 8 9 10 4 6 7
1  0 1 1 2 2 3  3 3 4 4
2  1 0 1 1 1 2  2 2 3 3
3  1 1 0 2 2 3  3 3 4 4
5  2 1 2 0 2 2  1 1 2 2
8  2 1 2 2 0 1  2 3 4 4
9  3 2 3 2 1 0  1 3 4 4
10 3 2 3 1 2 1  0 2 3 3
4  3 2 3 1 3 3  2 0 1 1
6  4 3 4 2 4 4  3 1 0 1
7  4 3 4 2 4 4  3 1 1 0
```



NEIGHBORHOOD OVERLAP

Tie Strength can be measured based on **neighborhood overlap**; the larger the overlap, the stronger the tie is (Onnela et al., 2007; Easley and Kleinberg, 2010).

Let N_i denote the friends of node v_i. Given a link $e(v_i,v_j)$, the neighborhood overlap is defined as:

$$overlap(v_i,v_j) = \frac{\text{number of nodes who are neighbors of both } v_i \text{ and } v_j}{\text{number of firends, who are neighbors of at least } v_i \text{ or } v_j} =$$

$$\frac{|N_i \cap N_j|}{|N_i \cup N_j| - 2}$$

We have -2 in the denominator just to exclude v_i and v_j from the set $N_i \cup N_j$

NEIGHBORHOOD OVERLAP

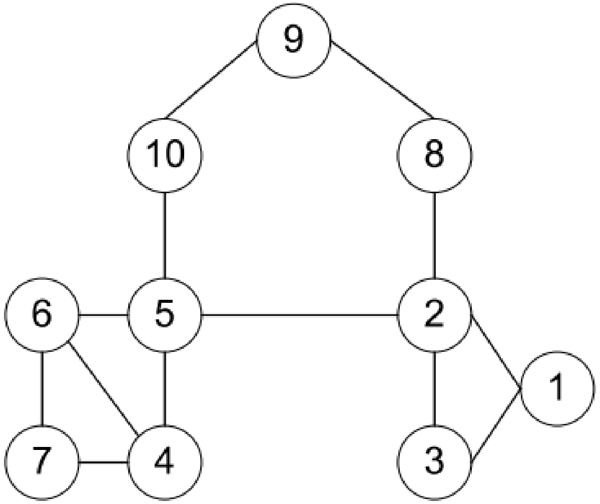
$N_2=\{1,3,5,8\}$, $N_5=\{2,4,6,10\}$.

Since $N_2 \cap N_5=\phi$, we have **overlap(2,5) =0**, indicating a weak tie between them.

On the contrary, the neighborhood **overlap between nodes 5 and 6** is

$$overlap(v_5, v_6) = \frac{|\{4\}|}{|\{2,4,5,6,7,10\}| - 2} = 1/4$$

Thus, edge ***e(5,6)*** is a stronger tie than ***e(2,5)***



NEIGHBORHOOD OVERLAP

```
>  ggg <- graph.formula(1-2,1-3,2-3,4-5,2-5,4-6, 4-7,5-6,6-7,2-8,8-9,9-10,10-5)
> plot(ggg)
> h <- neighbors(ggg,5)
> hh<- neighbors(ggg,6)
> h
```

[1] 2 4 6 10

```
> hh
```

[1] 4 5 7

```
> h1 <- induced.subgraph(ggg, h)
> hh1 <- induced.subgraph(ggg, hh)
> hhh<-h1+hh1
> V(hhh)
```

Vertex sequence:

[1] "2" "4" "6" "10" "5" "7"

```
> b=vcount(hhh)
> b
```

[1] 6

```
> i<-intersect(h, hh)
> i
```

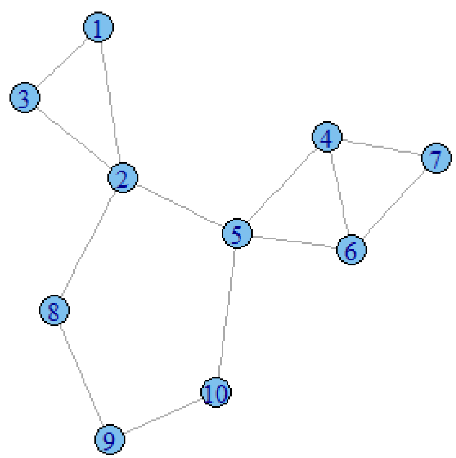
[1] 4

```
> t=length(i)
> t
```

[1] 1

```
> overlap56=t/(b-2)
> overlap56
```

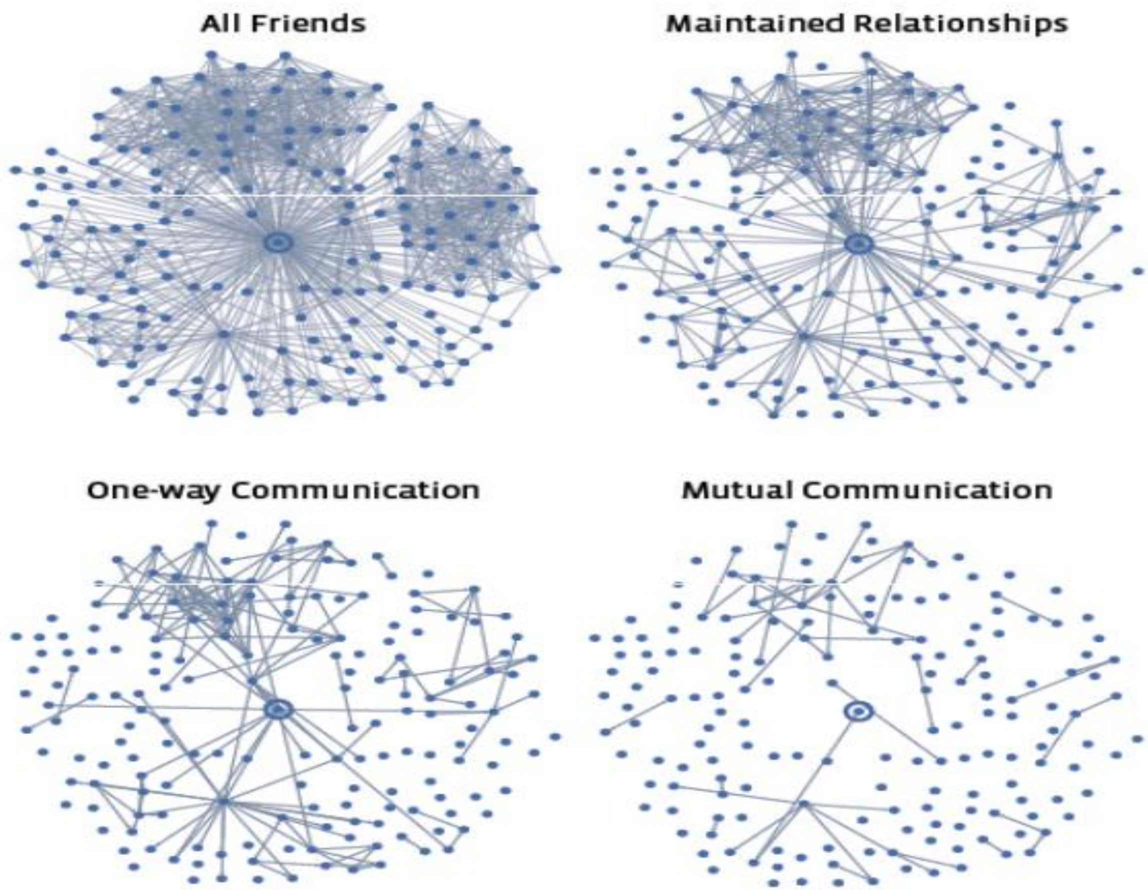
[1] 0.25



Tie Strength on Facebook

Tie Strength on Facebook

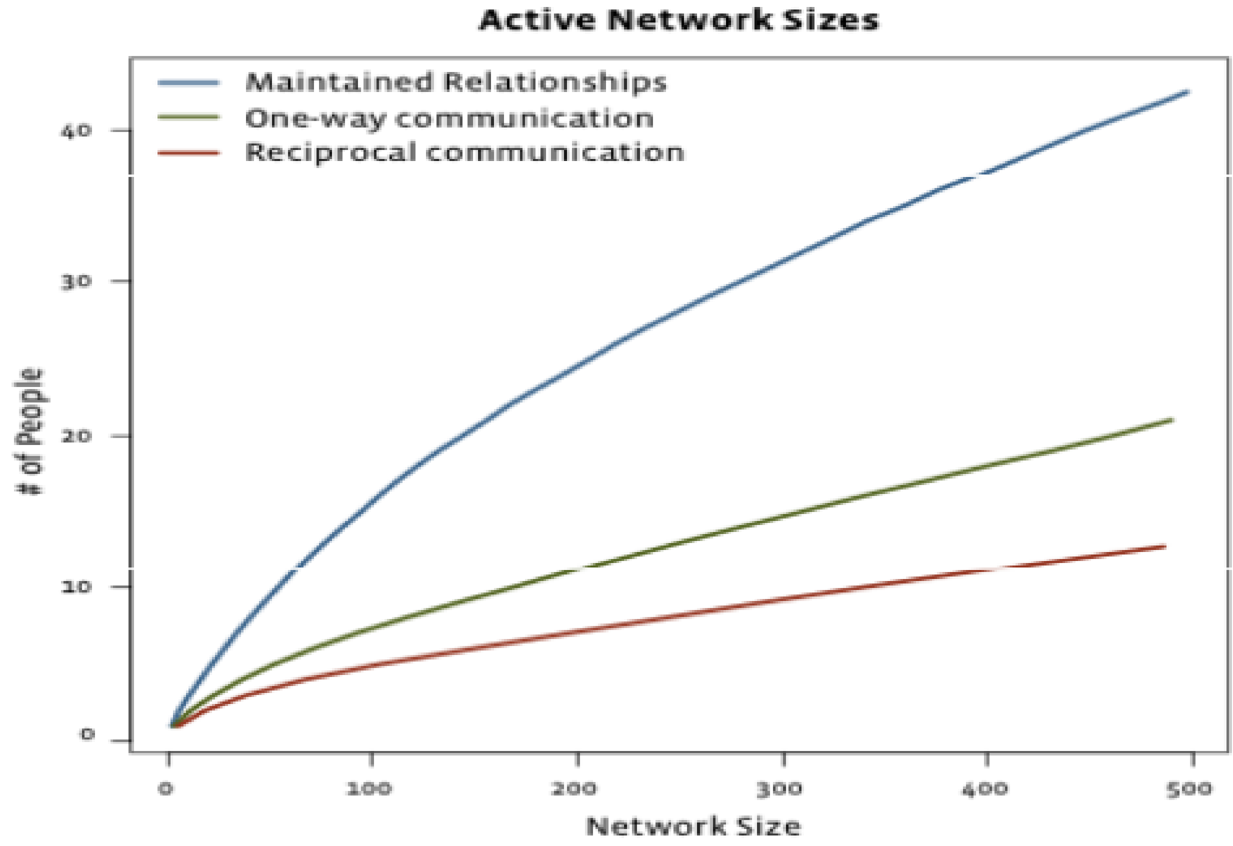
- A link represents **reciprocal (mutual) communication**, if the user both sent messages to the friend at the other end of the link, and also received messages from them during the observation period
- A link represents **one-way communication** if the user sent one or more messages to the friend at the other end of the link (whether or not these messages were reciprocated).
- A link represents a **maintained relationship** if the user followed information about the friend at the other end of the link, whether or not actual communication took place; “following information" here means either clicking on content via Facebook's News Feed service (providing information about the friend) or visiting the friend's profile more than once.



Networks, Crowds, and Markets: Reasoning About a Highly Connected World

Authors: David Easley,Jon Kleinberg

Tie Strength on Facebook



The number of links corresponding to maintained relationships, one-way communication, and reciprocal communication as a function of the total neighbohood size of users Facebook

LEARNING FROM USER ATTRIBUTES AND INTERACTIONS

- Twitter: **one can follow others without followee's confirmation**

"…even when using a very weak definition of "friend" (i.e. any one who a user has directed a post to at least twice) we find that Twitter users have a very small number of friends compared to the number of followers and followees they declare.

This implies the existence of two different networks:

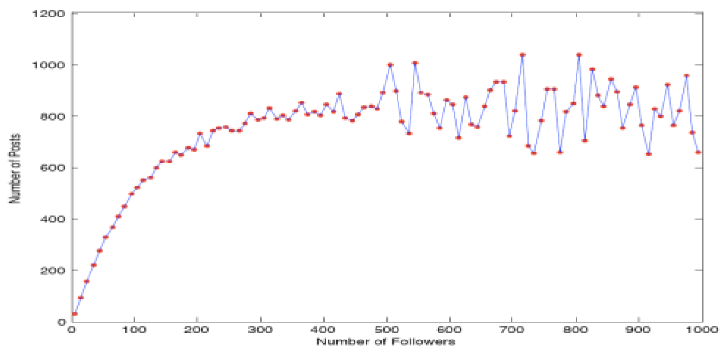
- a very dense one made up of followers and followees,

- and a sparser and simpler network of actual friends.

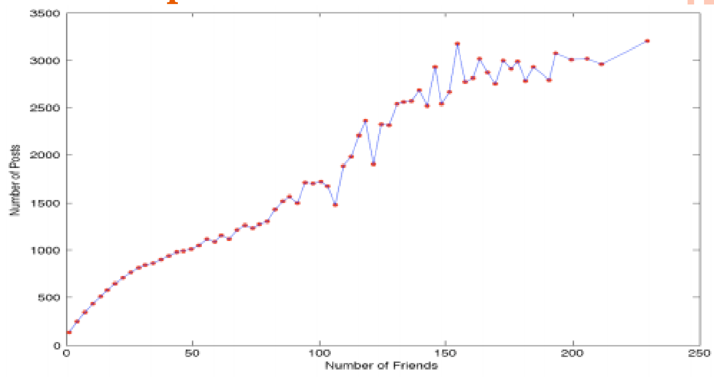
The latter proves to be a more influential network in driving Twitter usage since users with many actual friends tend to post more updates than users with few actual friends. On the other hand, users with many followers or followees post updates more infrequently than those with few followers or followees"

B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. First Monday, 14(1) , 2009

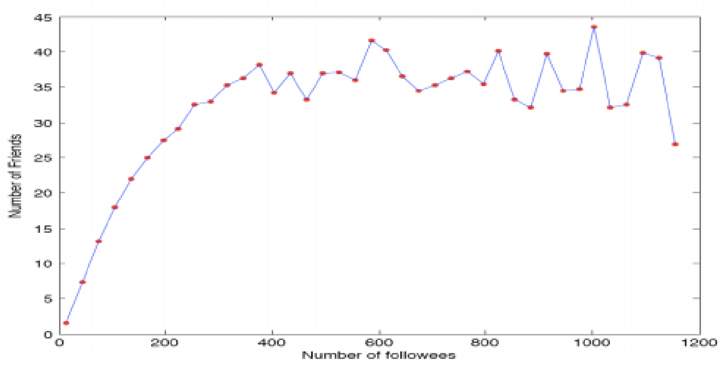
Twitter under the microscope



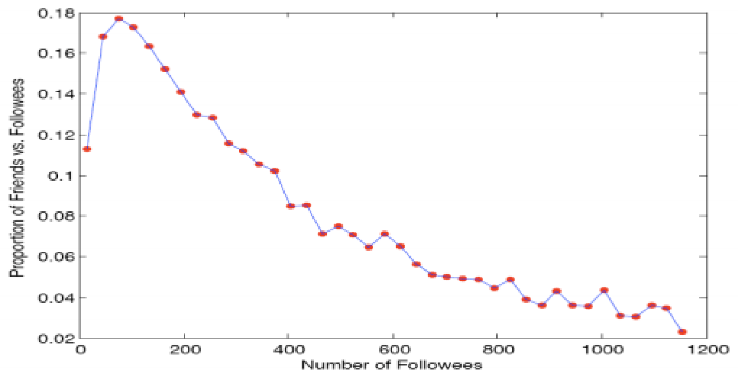
Number of posts as a function of the number of followers. The number of posts initially increases as the number of followers increases but it eventually saturates



Number of posts as a function of the number of friends. The number of posts increases as the number of friends increases, reaching 3200 without saturating



Number of friends as a function of the number of followees. The total number of friends saturates while the number of followees keeps growing due to the minimal effort required to add a followee



Proportion of friends vs. followees as a function of followers. It initially increases but rapidly approaches zero as the number of followees increases

UNIT 2: COMMUNITY DETECTION AND NETWORK VISUALIZATION

2.1. STRENGTH OF TIES.

NODE-CENTRIC MEASURES

1. Download the Karate-graph
   - *read.graph("path/karate.net", format="gml")*
   - *setwd('path/to/your_directory')*
2. Write the R-codes for finding:
   - Neighborhood Overlap between any two nodes
   - Cliques Quantity
   - Quantity of Cliques with k-length
   - Largest Cliques
   - Cores Quantity
   - Quantity of Cores with k-length
3. Place the quantitative results on the Table
4. Extract and Visualize the Cliques and Cores of the different size
5. Read and analyze the paper http://iussp2005.princeton.edu/papers/52438
6. Make the general conclusions about Node-Centric Measures specificity

ASSISTANTS TO THE ASSIGNMENT 2.1

Finding the Cliques	cliques(graph, min=k, max=k)
Finding the Neighbors	neighbors (graph, vertix)
Extracting the Subgraph	induced.subgraph(graph, subgraph)
Finding the Intersection of Graphs	intersect(graph, graph1)
Finding the Clique Number	clique.number(graph)
Finding the Largest Cliques	largest.cliques(graph)
Calculating the coreness for each vertex	graph.coreness(graph)
Forming the 3-columns Plot Area	par(mfrow=c(1,3))

THE NETWORK VISUALIZATION

SLIDES ARE MODIFIED BY DR. MEHMET GUNES

PRODUCTION ISSUES

Display

The more nodes there are the more pixels on the screen you will need.

800 x 600 = 480,000 pixels

1024 x 768 = 786,432 pixels

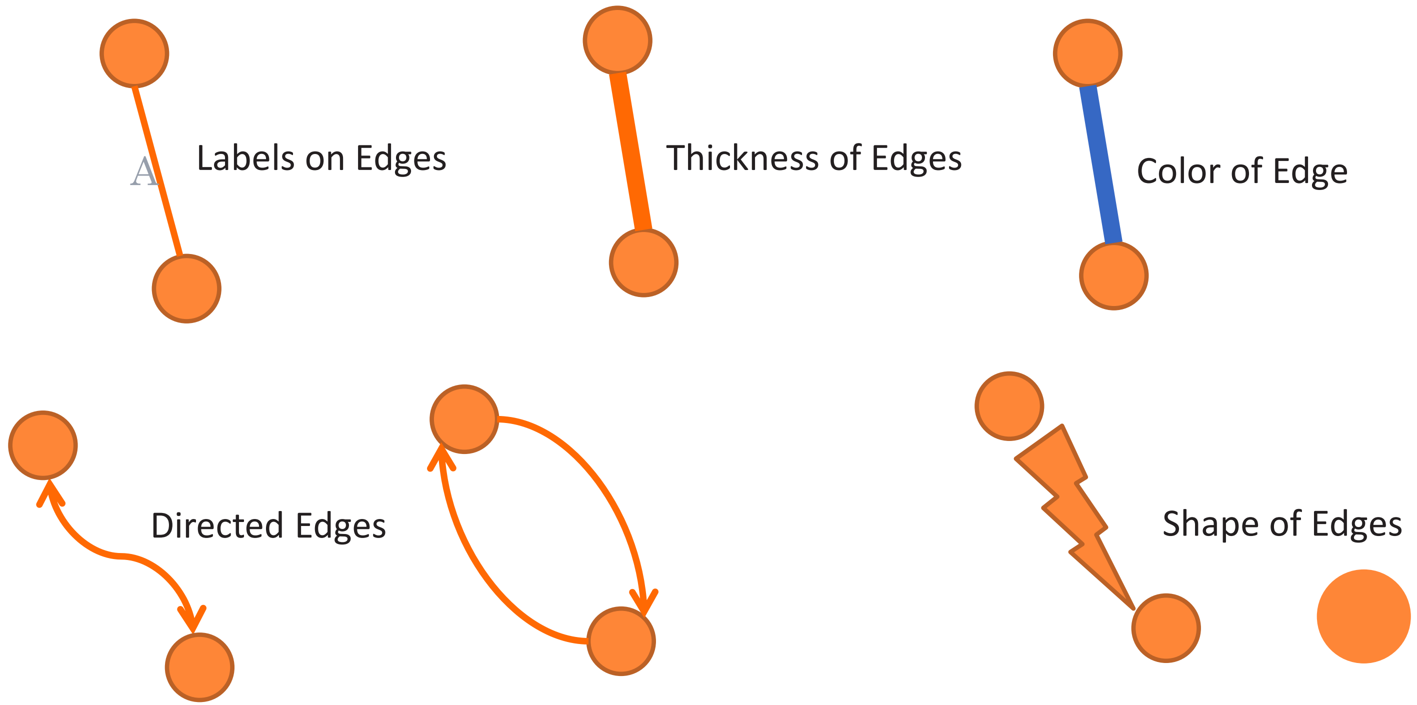
1920 x 1200 = 2,304,000 pixels

Not enough pixels to display all the nodes!!!

The more information that needs to be presented on the screen the more window design and window management become increasingly important.

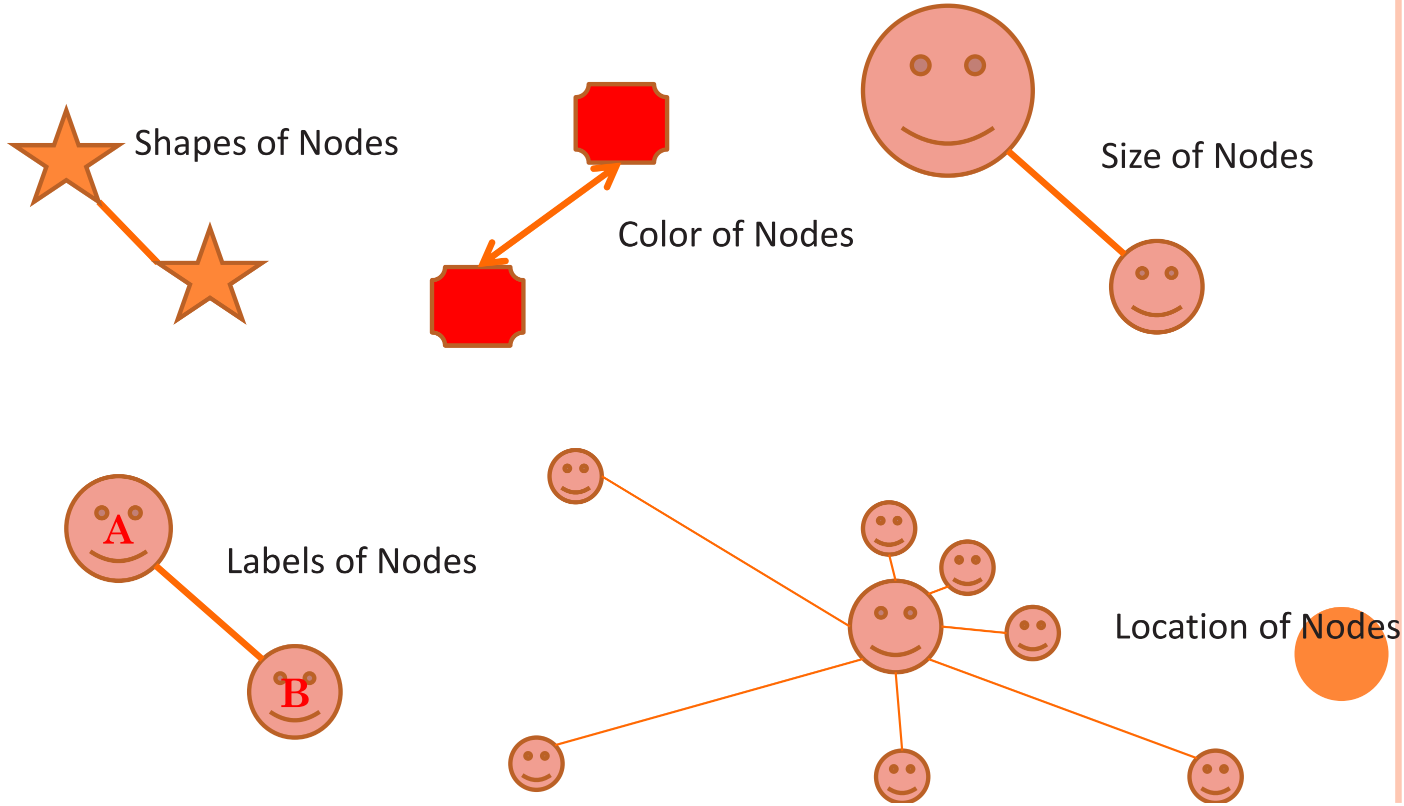
LAYOUT ISSUES

How to represent an edge?



LAYOUT ISSUES

How to represent a node?



Tips for effective visualizations

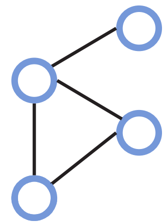
- *"The success of a visualization is based on deep knowledge and care about the substance, and the quality, relevance and integrity of the content.“*
  (Tufte, 1983)
  - know thy network!

- **Five Principles in the Theory of Graphic Display**
  - Above all else show the data
  - Maximize the data-ink ratio, within reason
  - Erase non-data ink, within reason
  - Erase redundant data-ink
  - Revise and edit

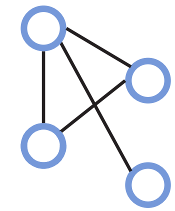
Source: http://www.edwardtufte.com/tufte/

Aesthetic criteria for network visualizations

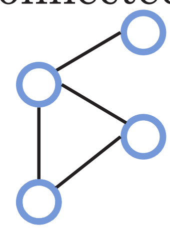
- minimize edge crossings



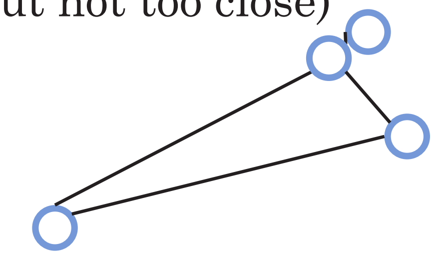
better than



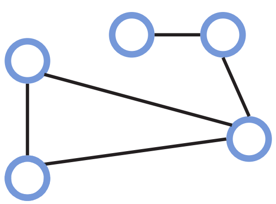
- uniform edge lengths
  - (connected nodes close together but not too close)



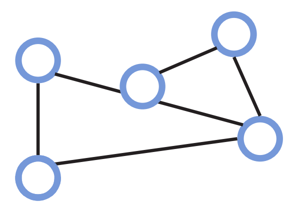
better than



- don’t allow nodes to overlap with edges that are not incident on them

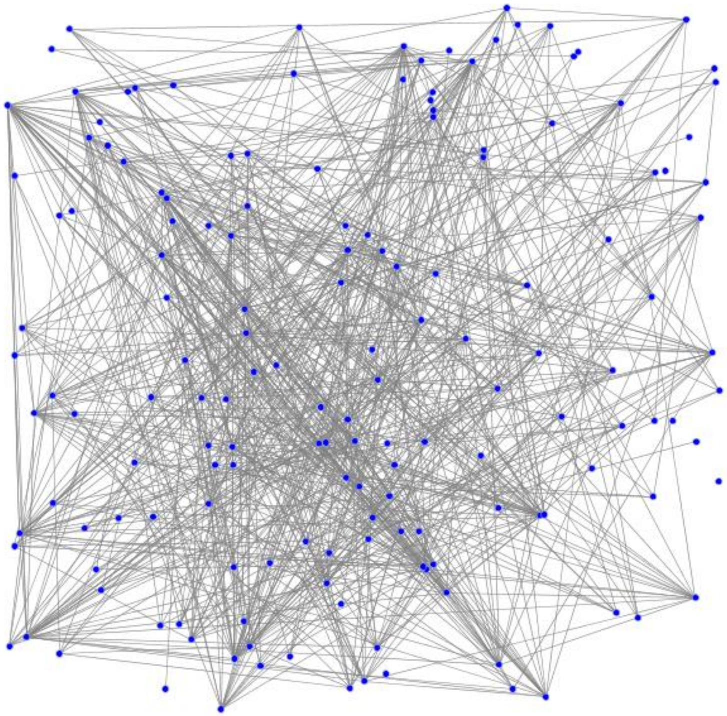


better than



Random Layout

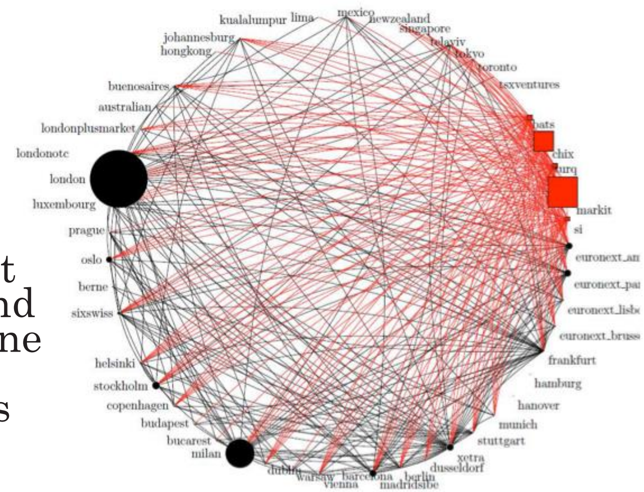
- Choose x & y coordinates at random
  - *advantage*: very fast
  - *disadvantage*: impossible to interpret

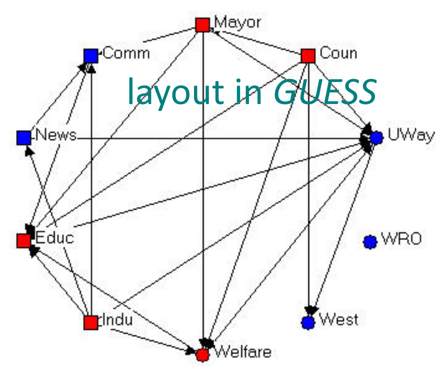


layout in *GUESS*

Circular layout

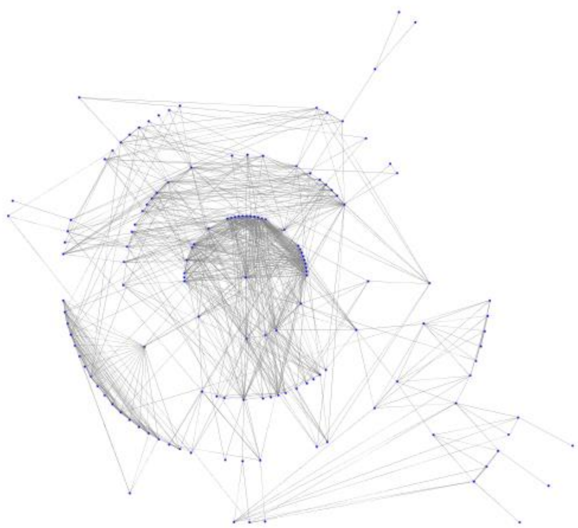
- Layout nodes along a circle and draw in all edges between them
- Advantages
  - neutrality:by placing all vertices at equal distances from each other and from the center of the drawing, none is given a privileged position, countering the tendency of viewers to perceive more centrally located nodes as being more important'
  - Very fast
- Disadvantages
  - difficult to interpret for large networks
    - many overlapping edges
    - many long edges
      - connected nodes need not be close together
    - clusters hard to identify

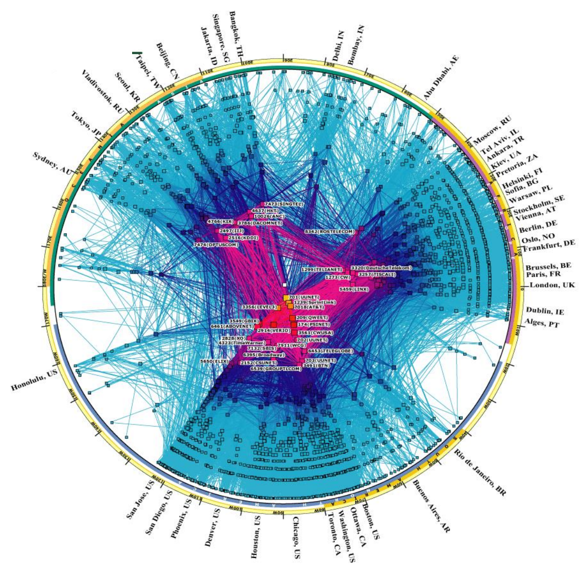




Radial Layout

- Start with one node, draw all other nodes in circular layers according to how many hops it takes to reach them
- Recall a tree representation



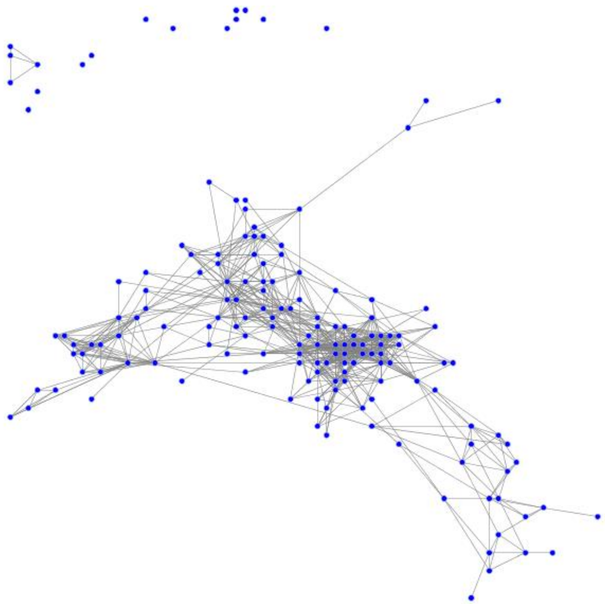


SPRING LAYOUT ALGORITHMS: FRUCHTERMAN AND REINGOLD

- Model roughly corresponds to electrostatic attraction between connected nodes

- Use adjacency matrix directly
- Iterative optimization
  - at each step, every node reacts to the pulls and pushes of the springs that tie it to all the other nodes

- Can be slow as the network grows

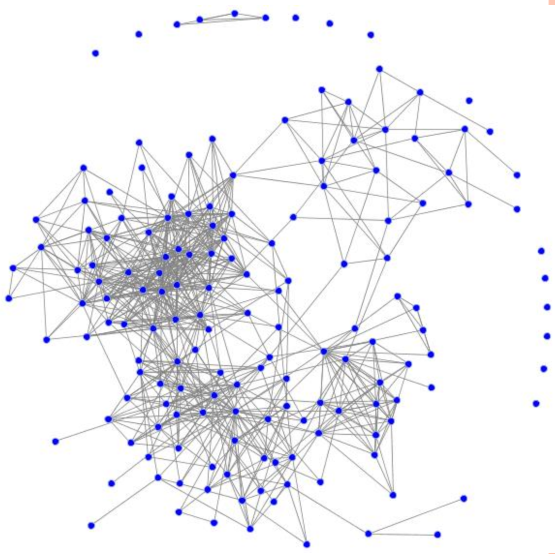


SPRING LAYOUT ALGORITHMS: KAMADA KAWAI

- All nodes are connected by springs with a resting length proportional to the length of the shortest path between them

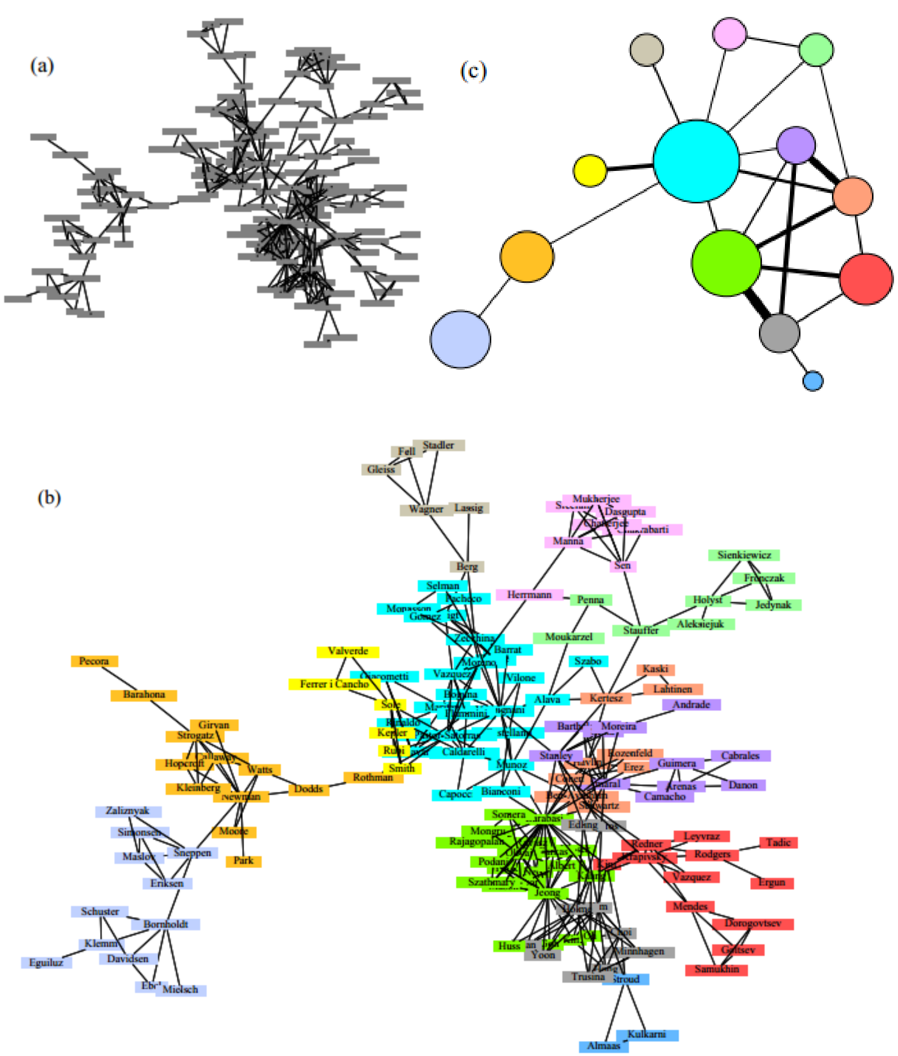
- Need to calculate all pairs shortest paths first
- Iterative optimization

- Advantage: can be used on edge-weighted graphs
- Can be slow as the network grows



Strategies for visualizing large graphs

- Reduce the number of nodes and edges
  - introduce thresholds
    - only authors who have written at least x papers
    - only edges with weight **> y**
    - only nodes with degree **> z** (e.g. removing leaf nodes)
  - show minimum spanning trees
    - can visualize all the nodes with a subset of the edges
  - use pathfinder network scaling (http://iv.slis.indiana.edu/sw/pfnet.html)
    - triangle inequality to eliminate redundant or counter-intuitive links
    - remaining edges are more representative of internode relationships than minimum spanning trees
  - collapse nodes into clusters
    - show multiple nodes as a single node
    - display connections between clusters



Example of coarsening network structure

Newman & Girvan 2004

(a) The initial network is a network of coauthorships between physicists who have published on topics related to networks.

(b) Application of the shortest-path betweenness version of the community structure algorithm produces the communities shown by the colors.

(c) A coarse-graining of the network in which each community is represented by a single node, with edges representing collaborations between communities. The thickness of the edges is proportional to the number of pairs of collaborators between communities.

Clearly panel (c) reveals much that is not easily seen in the original network of panel (a)

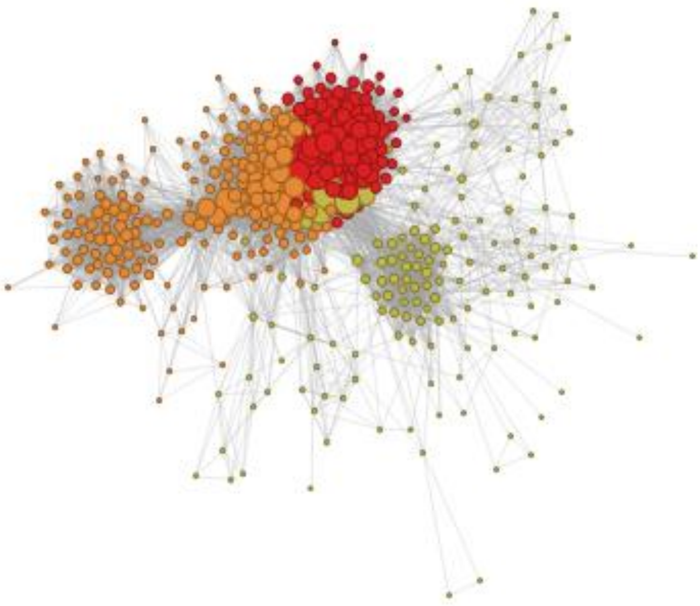
Source: Finding and Evaluating Community Structure in Netowrks, M. E. J. Newman and M. Girvan, http://link.aps.org/doi/10.1103/PhysRevE.69.026113 DOI: 10.1103/PhysRevE.69.026113

NETWORK GRAPHS

- **It does not exist n a priori right network representation**
- One has to try different ones, having in mind what it is needed to be visualized

- Many software:
  - **Gephi (https://gephi.org/)**
  - **Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/)**
  - **UCINET (https://sites.google.com/site/ucinetsoftware/home)**
  - **NetworkX with Python (http://networkx.github.io/)**
  - **R iGraph (http://igraph.org/r/)**
  - **NodeXL (http://nodexl.codeplex.com)**
  - **Graphviz (http://www.graphviz.org/)**
  - **Cuttlefish (http://cuttlefish.sourceforge.net/**

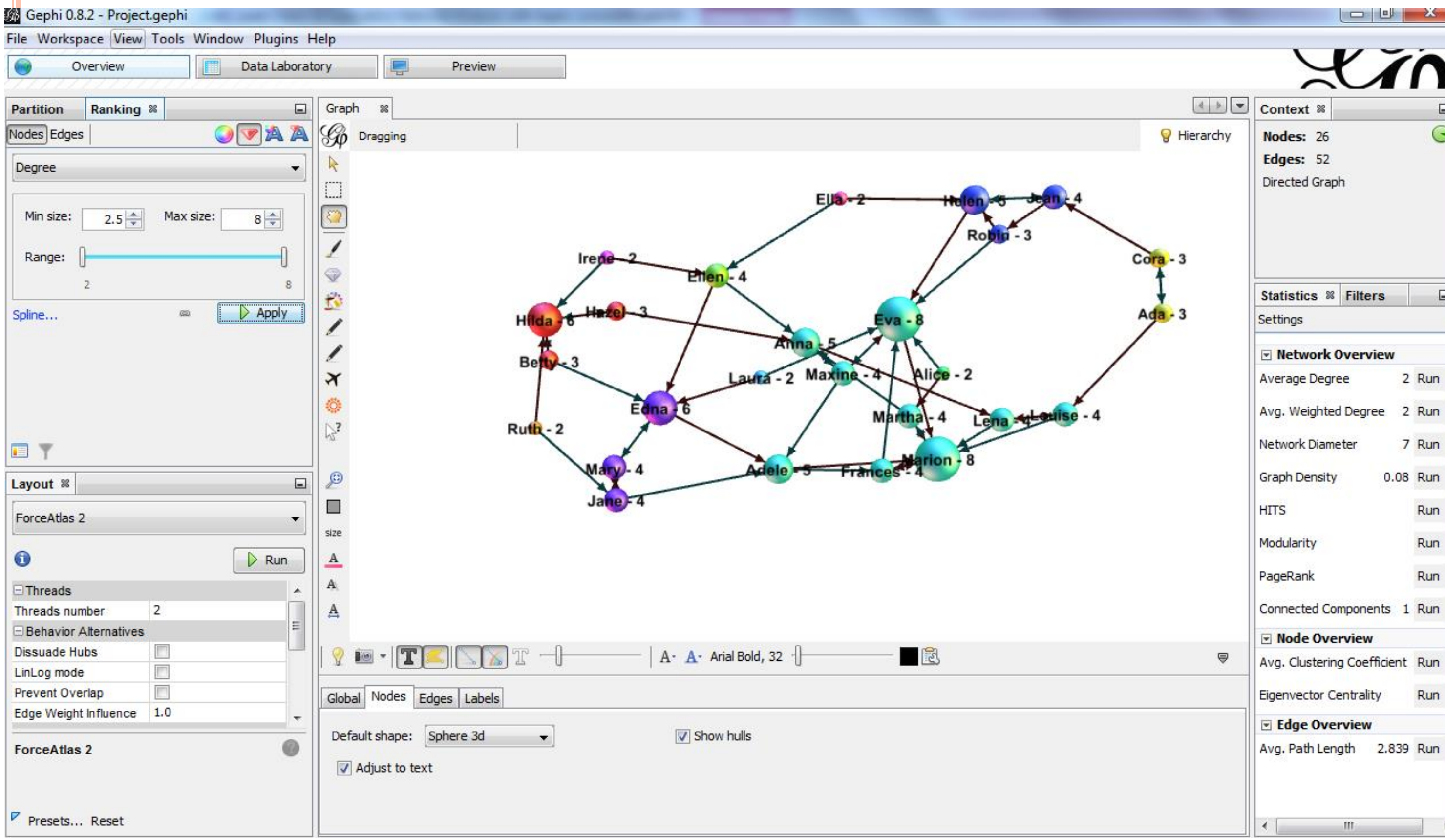
Overview of Network Analysis Tools



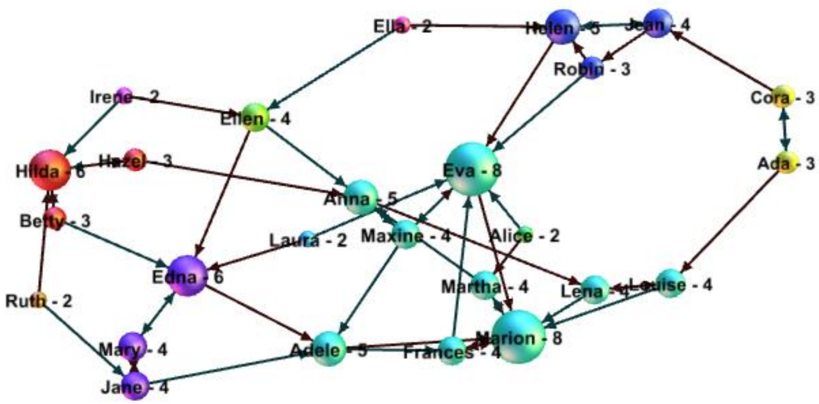
SOFTWARE TOOLS

Gephi

Graph exploration and manipulation software



Gephi



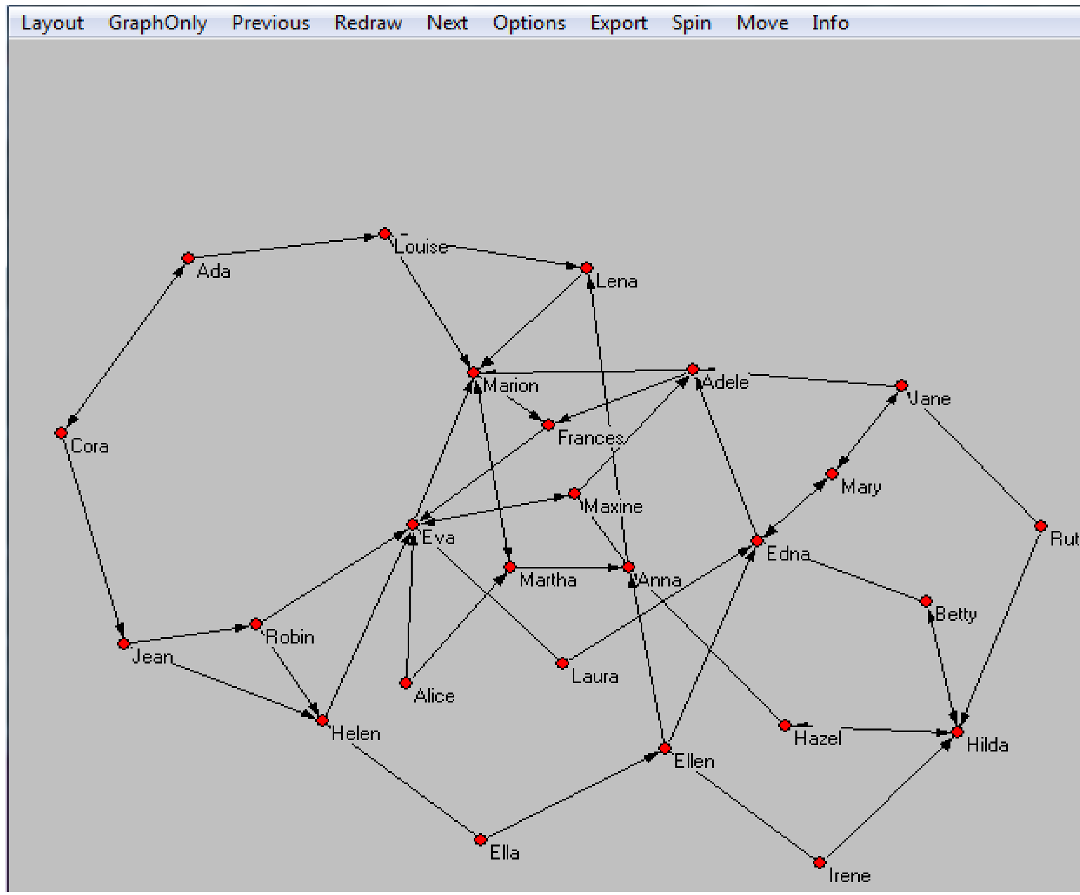
http://gephi.org/

INPUT FORMAT	OUTPUT FORMAT
GraphViz(.dot), Graphlet(.gml), GUESS(.gdf), LEDA(.gml), NetworkX(.graphml, .net), NodeXL(.graphml, .net), Pajek(.net, .gml), Sonivis(.graphml), Tulip(.tlp, .dot), UCINET(.dl), yEd(.gml), Gephi (.gexf), Edge list(.csv), databases	GUESS(.gdf), Gephi(.gexf), .svg, .png

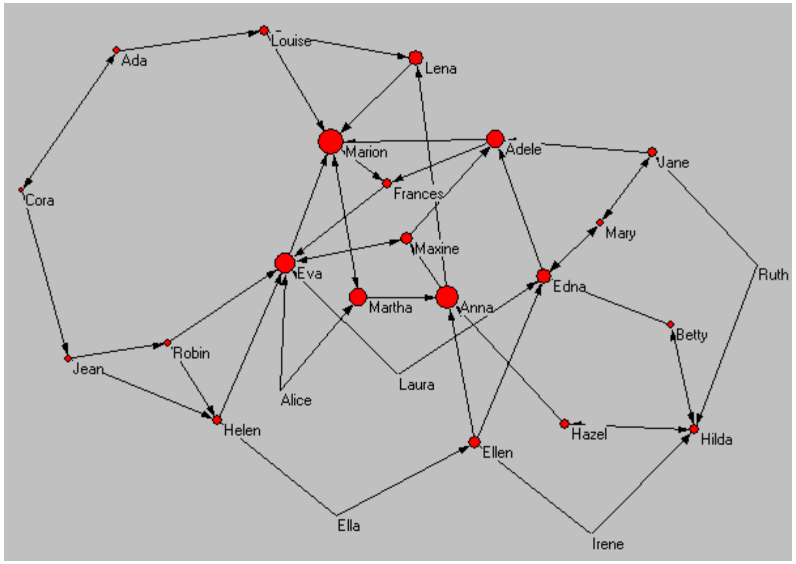
Software tools

Analysis and Visualization of Large Scale Network

Pajek



Pajek



http://pajek.imfm.si/doku.php?id=download

INPUT FORMAT	OUTPUT FORMAT
net, .paj, .dat(UCINET), .ged, .bs, .mac, .mol	.net, .paj, .dat(UCINET), .xml(graphML), .b

IGRAPH LIBRARY

http://cneurocvs.rmki.kfki.hu/igraph/index.html

- iGraph is a free software package for creating and manipulating undirected and directed graphs
- Contains the implementation of a lot graph algorithms
- Provides a platform for the developing and/or implementing graph algorithms

- Empty graphs

```
> e <- graph.empty()
> e <- graph.empty(n=10)
```

- Full graphs

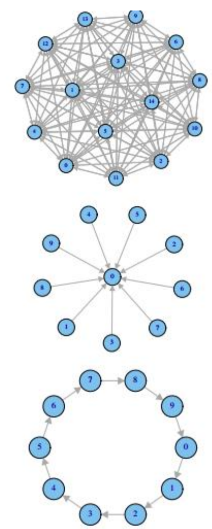
```
> f <- graph.full(15)
> f <- graph.full(15, directed=TRUE)
```

- Stars and Rings

```
> s <- graph.star(10, mode = "in")
> s <- graph.star(10, mode = "out")
> s <- graph.star(10, mode = "undirected")
> r <- graph.ring(10, directed=TRUE)
> r <- graph.ring(10, mutual=TRUE)
> r <- graph.ring(10, circular=TRUE)
```

- Other structures

```
> graph.lattice()
> graph.tree()
```



NetworkX with Python

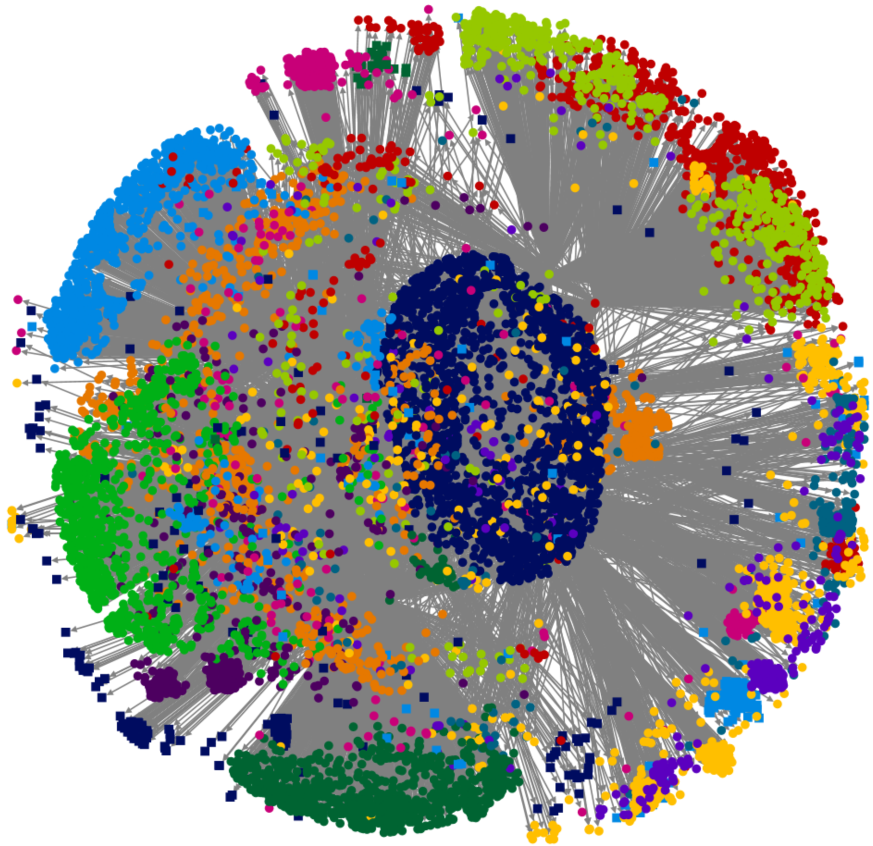
http://networkx.lanl.gov/

Software tools

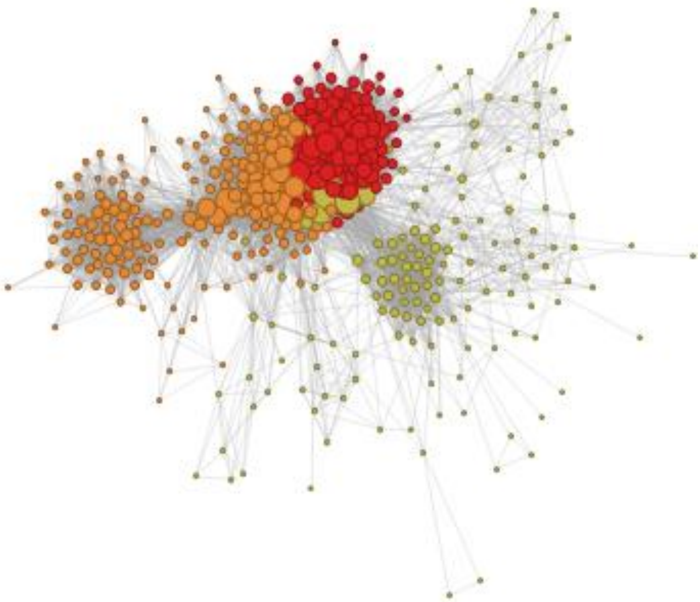
Network Graphs
The Social Media Research Foundation

SNA integrated into Excel

http://nodexl.codeplex.com/

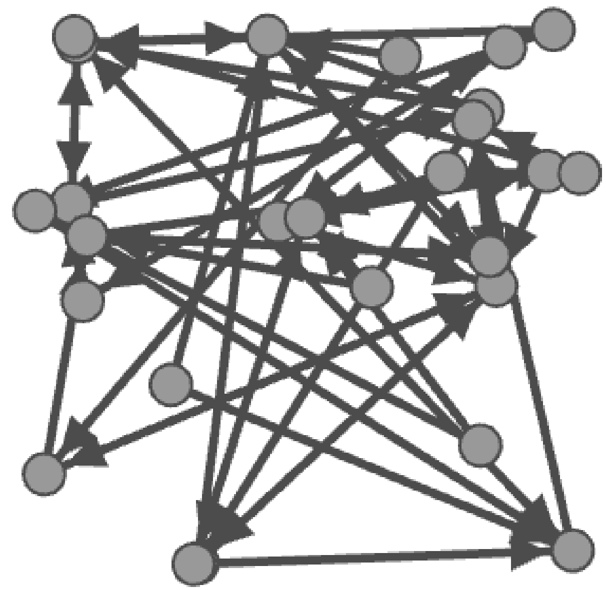


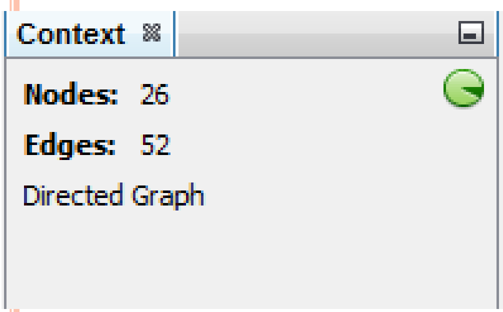
Using Gephi for Exploratory Social Network Analysis



INSTRUCTION FOR ANALYSIS OF THE GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

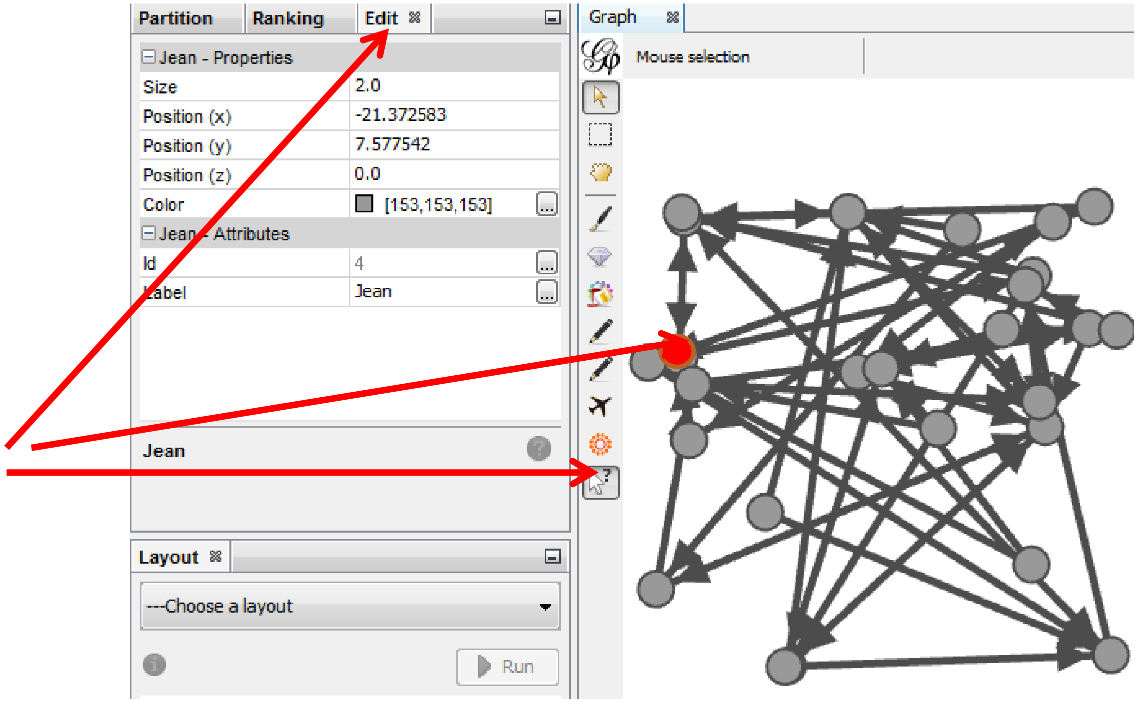
- Open file: **File / Open**





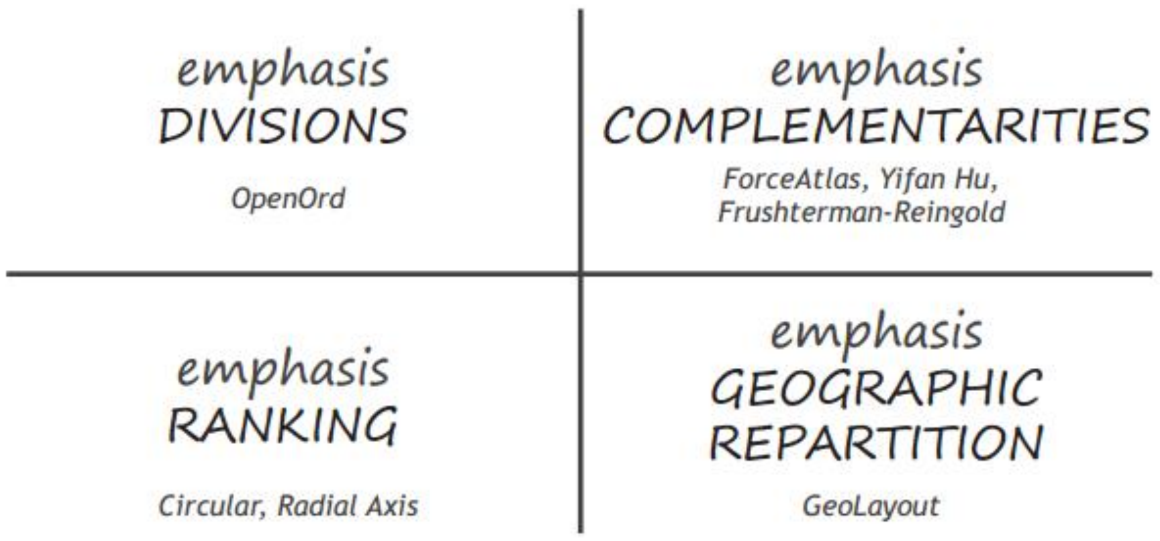
GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

- Check the **Properties** of the selected node



LAYOUT

- The purpose of layout properties is to let you control the algorithm in order to make a readable representation.



Force Atlas layout

- Home-brew layout of Gephi, it is made to spatialize Small-World / Scale-free networks.
- It is focused on quality (meaning "being useful to explore real data") to allow a rigorous interpretation of the graph (e.g. in SNA) with the fewest biases possible, and a good readability even if it is slow.

Author: Mathieu Jacomy

Date: 2007

Kind: Force-directed

Complexity: $O(N^2)$

Graph size: 1 to 10 000 nodes

Use edge weight: Yes

Run ForceAtlas

the layout by applying the following settings step by step:

Autostab strength = 2 000	Increase to move the nodes slowly
Repulsion strength = 1 000	How strongly does each node reject others
Attraction strength = 1	How strongly each pair of connected nodes attract each other
Gravity = 100	Attract all nodes to the center to avoid dispersion of disconnected components
Attraction Distrib. = checked	Push hubs (high number of output links) at the periphery and put authorities (high number of in put links) more central

FRUCHTERMAN-REINGOLD LAYOUT

It simulates the graph as a system of mass particles. The nodes are the mass particles and the edges are springs between the particles. The algorithms try to minimize the energy of this physical system. It has become a standard but remains very slow.

Author: Thomas Fruchterman & Edward Reingold

Date: 1991

Kind: Force-directed

Complexity: $O(N^2)$

Graph size: 1 to 1 000 nodes

Use edge weight: No

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11).

RUN FRUCHTERMAN-REINGOLD

the layout by applying the following settings step by step:

Area = 100	Graph size area
Area = 100 000	
Gravity = 1 000	Attract all nodes to the center to avoid dispersion of disconnected components
Gravity = 100	

Yifan Hu Multilevel layout

It is a very fast algorithm with a good quality on large graphs. It combines a force-directed model with a graph coarsening technique (multilevel algorithm) to reduce the complexity. The repulsive forces on one node from a cluster of distant nodes are approximated by a Barnes-Hut calculation, which treats them as one super-node. It stops automatically.

Author: Yifan Hu

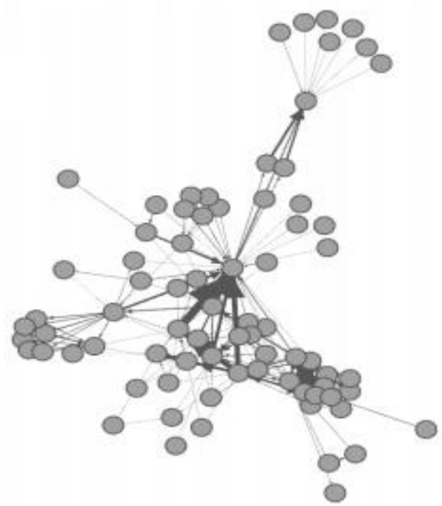
Date: 2005

Kind: Force-directed + multilevel

Complexity: O(N*log(N))

Graph size: 100 to 100 000 nodes

Use edge weight: No



Y. F. Hu, Efficient and high quality force-directed graph drawing. The Mathematica Journal, 10 (37-71), 2005

Run Yifan Hu Multilevel

- Launch the layout by applying the following settings step by step:

Step ratio = 0.99	Ratio used to update the step size. Increase it for a better quality (vs speed)
Optimal distance = 200	Natural length of the springs. Increase it to place nodes farther apart
Theta = 1.0	Approximation for Barnes-Hut calculation. Smaller values mean more accuracy

FORCEATLAS 2 LAYOUT

Improved version of the Force Atlas to handle large networks while keeping a very good quality. Nodes repulsion is approximated with a Barnes-Hut calculation, which therefore reduces the algorithm complexity. Replace the attraction" and "repulsion" forces by a "scaling" parameter.

Author: Mathieu Jacomy

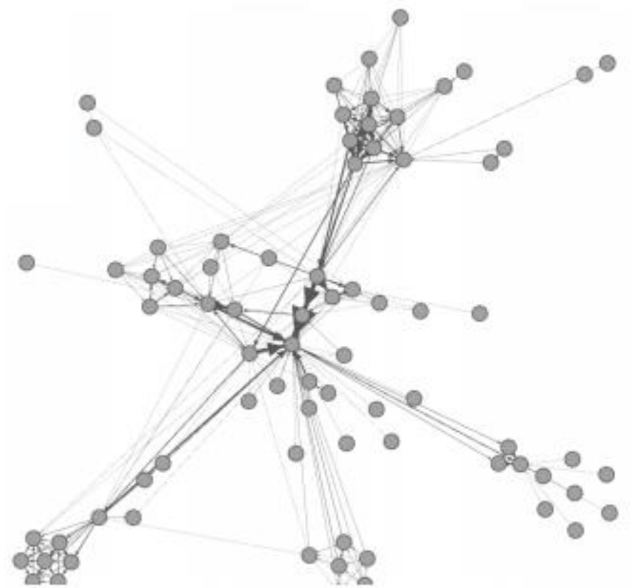
Date: 2011

Kind: Force-directed

Complexity: O(N*log(N))

Graph size: 1 to 1 000 000 nodes

Use edge weight: Yes



RUN FORCEATLAS 2

the layout by applying the following settings step by step:

<table>
<tr><td>LinLog mode = checked</td><td rowspan="2">Linear attraction & logarithmic repulsion (lin-lin by default), makes clusters tighter.</td></tr>
<tr><td>LinLog mode = unchecked</td></tr>
<tr><td>Scaling = 100</td><td>Increase to make the graph sparser</td></tr>
<tr><td>Edge weight influence = 0</td><td>From 0 (no influence) to 1 (normal). Set 0 to calculate forces without edge weight</td></tr>
</table>

Label Adjust layout

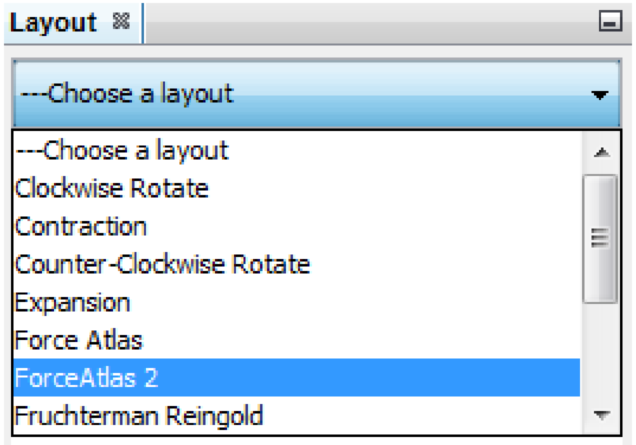
It works on text size to repulse nodes and therefore makes every label readable. It only runs on the visible nodes in the Visualization panel

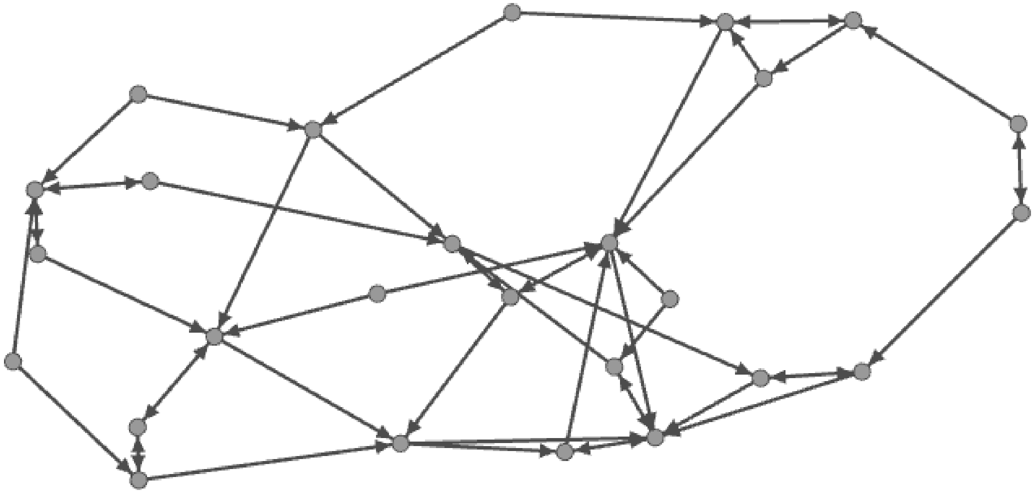
- Locate the Visualization settings
- Click on  to activate text display
- Increase the text size to the maximum
- Go to the Layout panel.
- Select the "Label Adjust" algorithm and run it until it stops

Instruction for analysis of the
GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

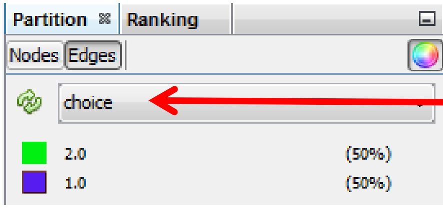
- Change the **Layout** of network: 

  Layout / Run



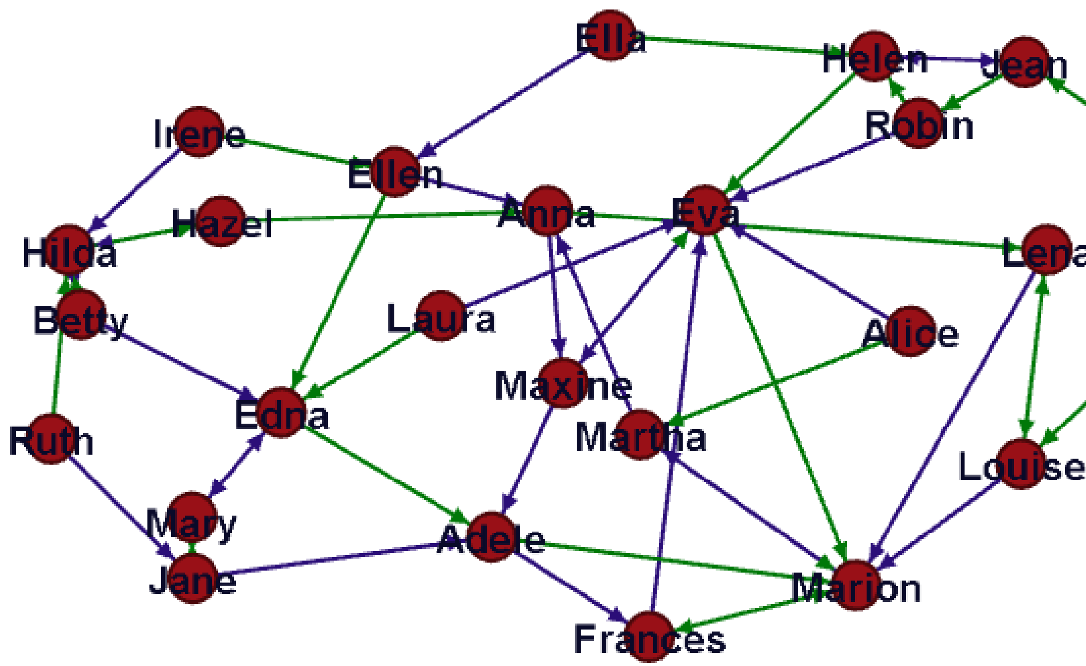


Instruction for analysis of the GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS



Define **first** and **second** choices:

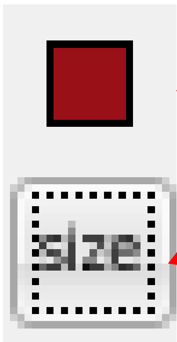
Partition / Edges / Choice

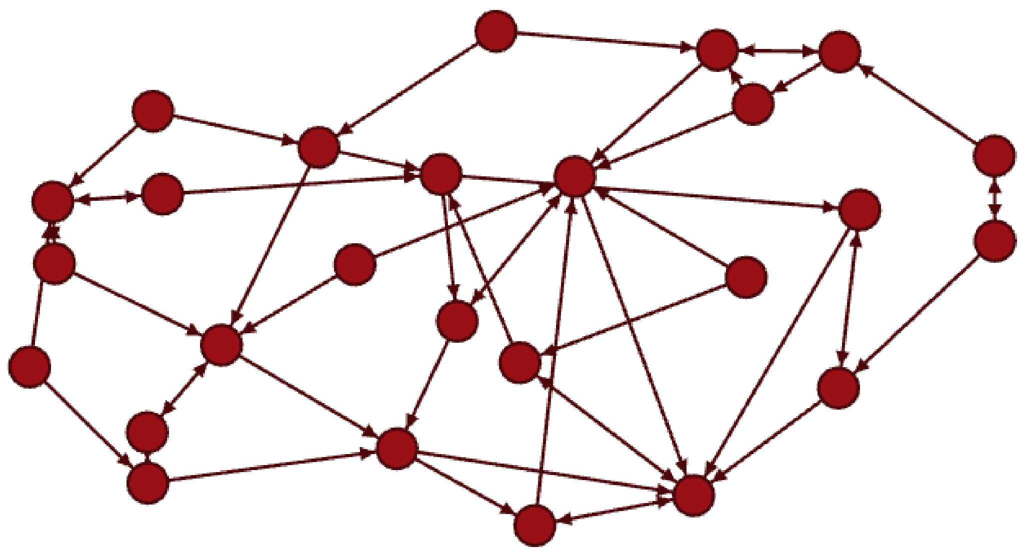


Instruction for analysis of the
GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

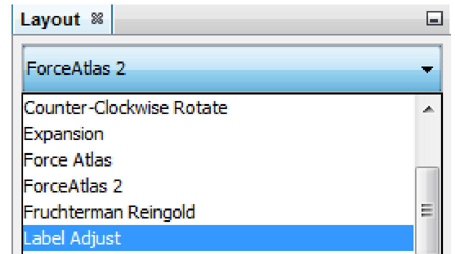
Right click / Change the Color of the nodes

Right click / Change the Size of the nodes



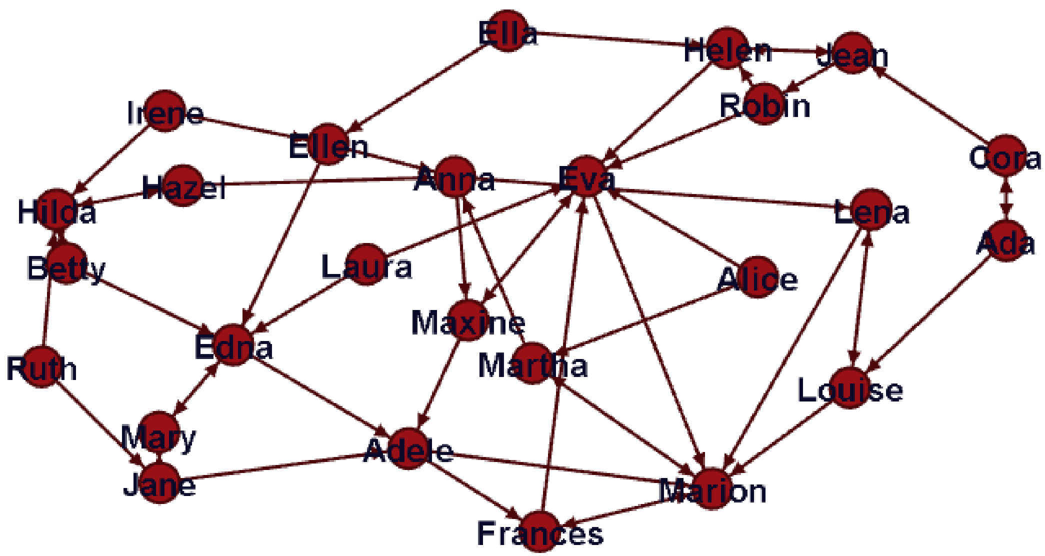


Instruction for analysis of the GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

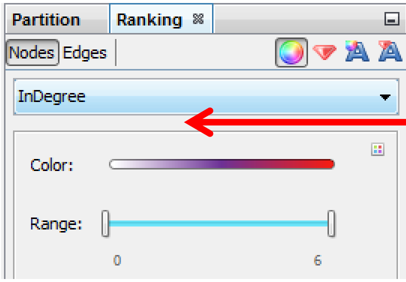


Show Labels

Adjust Labels

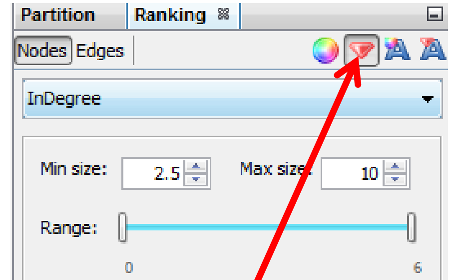


Instruction for analysis of the
GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

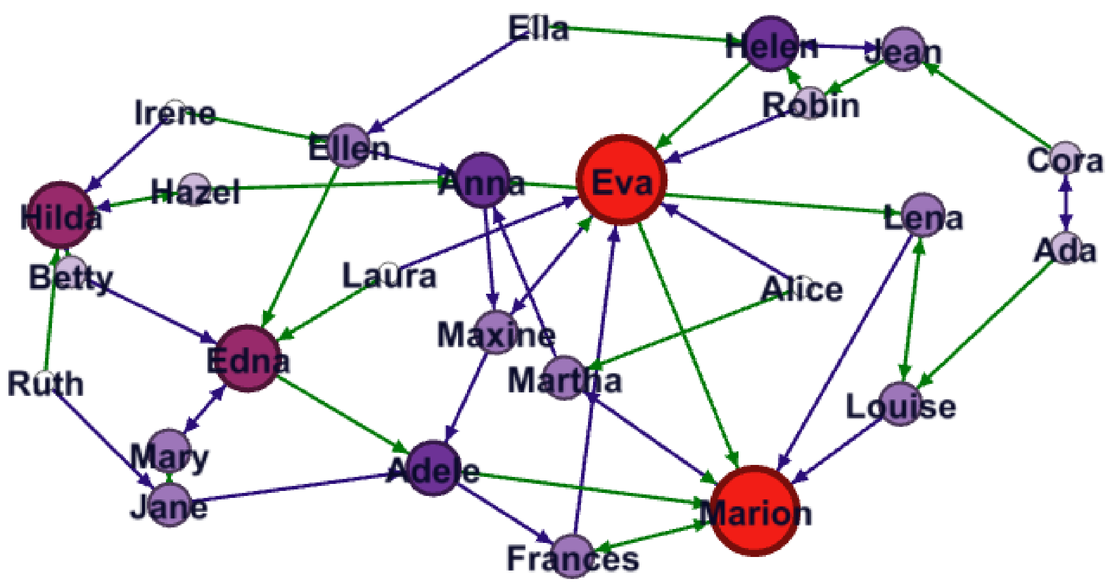


Define **Indegree** Centrality:
Partition / Nodes / Indegree

Apply



Define **Max** and **Min Size** of the Nodes



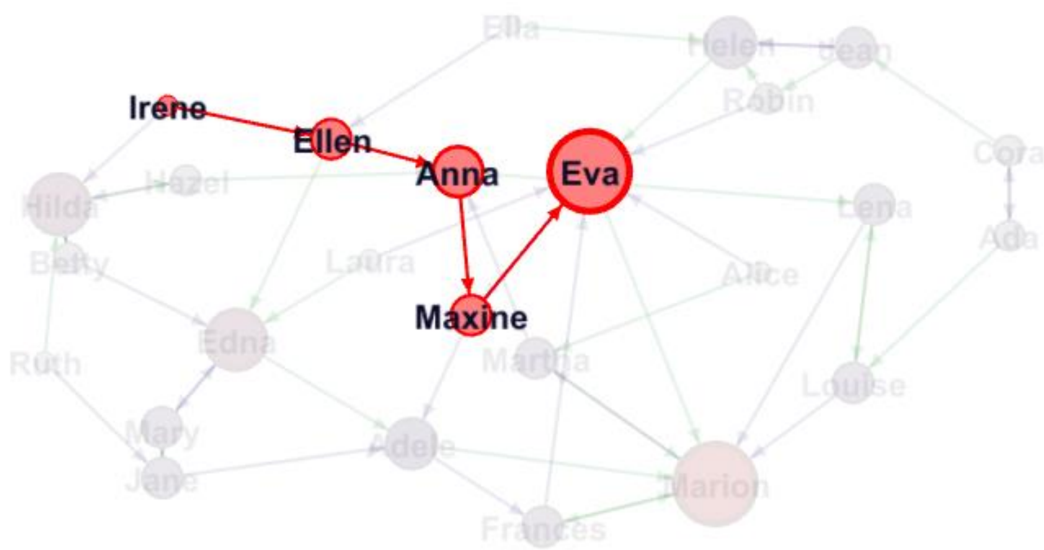
Instruction for analysis of the GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS

Display the **Shortest Path** if Exist between two clicked Nodes

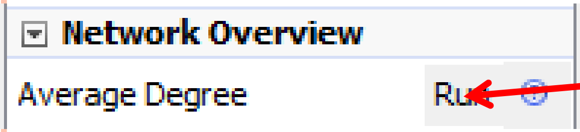
Select a source node

Select a target node

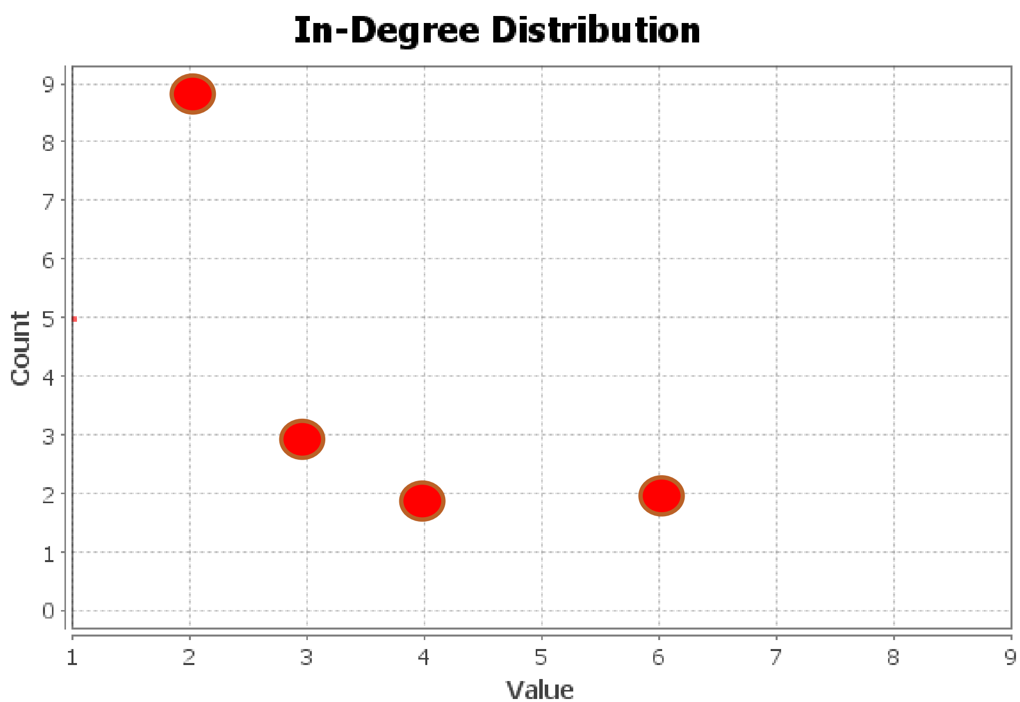
A 4 distant path has been found



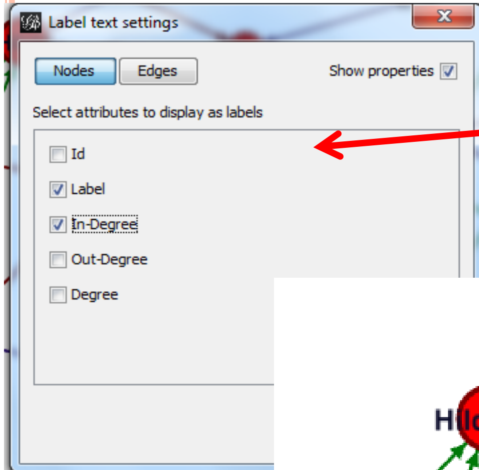
Instruction for analysis of the GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS



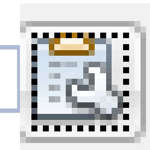
Define **Average Degree** of the Network

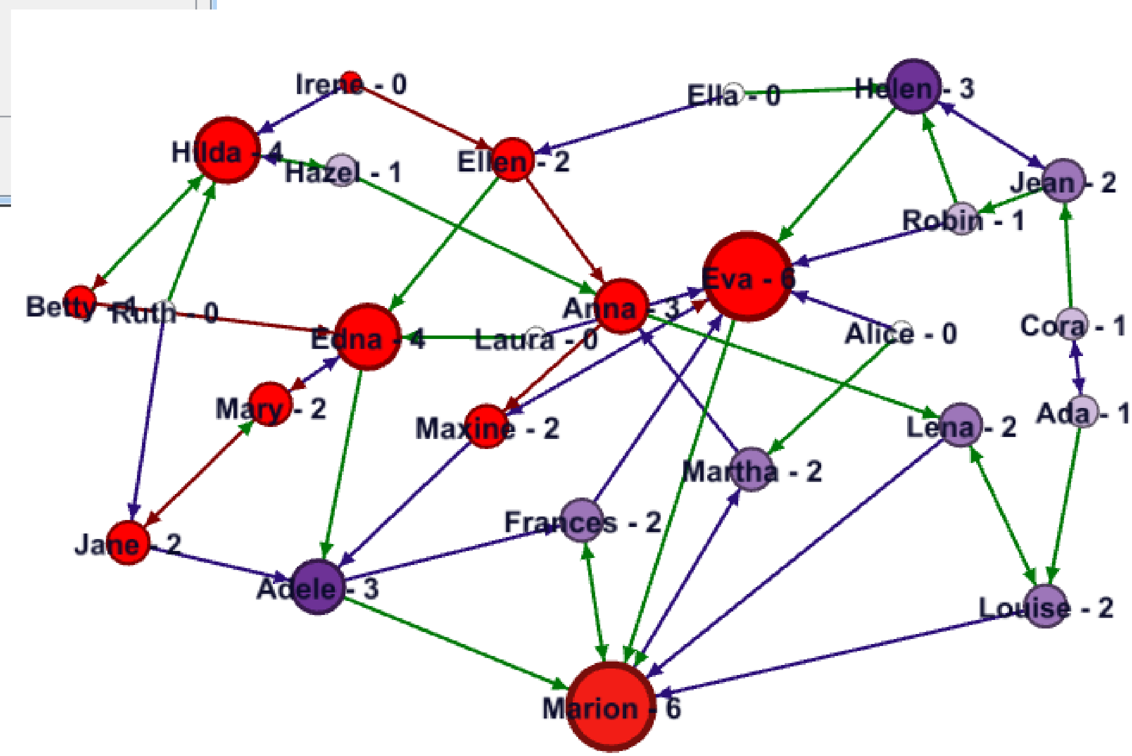


Instruction for analysis of the GIRLS' SCHOOL DORMITORY DINING-TABLE PARTNERS



Define **Labels settings**





2.2. Analyzing Facebook networks with Gephi (Netvizz app)

NETVIZZ

netvizz

1. Read the Paper "Studying Facebook via Data Extraction: The Netvizz Application" Bernhard Rieder

http://rieder.polsys.net/files/rieder_websci.pdf

- Using additional sources write a little review about exited tools for data extracting from Facebook

2. Sign into a Facebook account

- Search for "Netvizz" application
- Choose parameters you would like included in the file Personal network (extracts your friends and the friendship connections between them):

NETVIZZ

facebook

Buscar

netvizz v1.0

Netvizz is a tool that extracts data from different sections of the Facebook platform (personal profile, groups, pages) for research purposes. File outputs can be easily analyzed in standard software.

For questions, please consult the **FAQ** and **privacy** sections. Non-commercial use only.

Big networks may take some time to process. **Be patient and try not to reload!**

Developing and hosting netvizz costs time and money. If the tool is useful for you, please consider to **Donate**

The following modules are currently available:

personal network - extracts your friends and the friendship connections between them
personal like network - creates a network that combines your friends and the objects they liked in a bipartite graph
group data - creates networks and tabular files for both friendships and interactions in groups
page like network - creates a network of pages connected through the likes between them
page data - creates networks and tabular files for user activity around posts on pages

---

Choose option: **Personal Network** creating

Netvizz

facebook Buscar

netvizz v1.0

your personal friend network:

Creates a network file (gdf format) with all the friendship connections in your personal network, as well as a stat file (tsv format).

Select user data to include in the file (sex, interface language, and account age ranking are standard):

☐ friends' like and post count (public and visible to logged user, first 1000 only), includes counts for received likes and comments on posts, adds an additional ±6 seconds of waiting time per friend

start

and

netvizz v1.0

Your files files have already been generated for today:

Your **gdf file** (right click, save as...).

Your **tab file** (right click, save as...).

GOAL OF ASSIGNMENT:

- Work in the Group (not less then 3 students)
- Invent the “draft’ idea of **Media-Project**, which you are going to develop nearest time.

Your **goal**: Choose the potential group-owner of this Media-Project on the basis of the results of Social Network Analysis.

For this **goal**:

- Import your data from ***Facebook*** to ***Gephi***.
- Complete a research of the ***Facebook Networks via*** applying all Gephi’s tools and during class learned technique for comparing ***Group’s Facebook Networks.***
- Prove the advantages of chosen group-owner Network.
- Write the Report about results of Social Network Analysis and present it in the not less than 15 Slides.
- Report must contain:
  - Tables with statistical data (see next slides)
  - Visualizing the Networks using different known SNA measures and Filters tools
  - Detecting the Communities of the Network and there short characteristic (description)

For completing this assignment you should study

- *several examples of SNA research Reports. For example: http://www.slideshare.net/SandeepSharma65/social-media-analysis-project?qid=80dd2c41-2e22-4a42-8d96-dc082afb5ef1&v=default&b=&from_search=5)*
- Gephi guide: *For example:*
  http://www.clementlevallois.net/gephi/tuto/en/gephi_tutorial%20foundations_en.pdf)

Node Level Analysis (Assignment)

- You need to report the top 20 developers in terms of their
  - Degrees
  - Betweeness
  - Closeness
- Analyze and identify individuals (among top 20) who has uncorrelated centrality measures and explain why this may happen.
- Also please explain what these measures mean in the context of **Media-Project**

UNCORRELATED CENTRALITY MEASURES

	Low Degree	Low Closeness	Low Betweenness
High Degree	-		
High Closeness		-	
High Betweenness			-

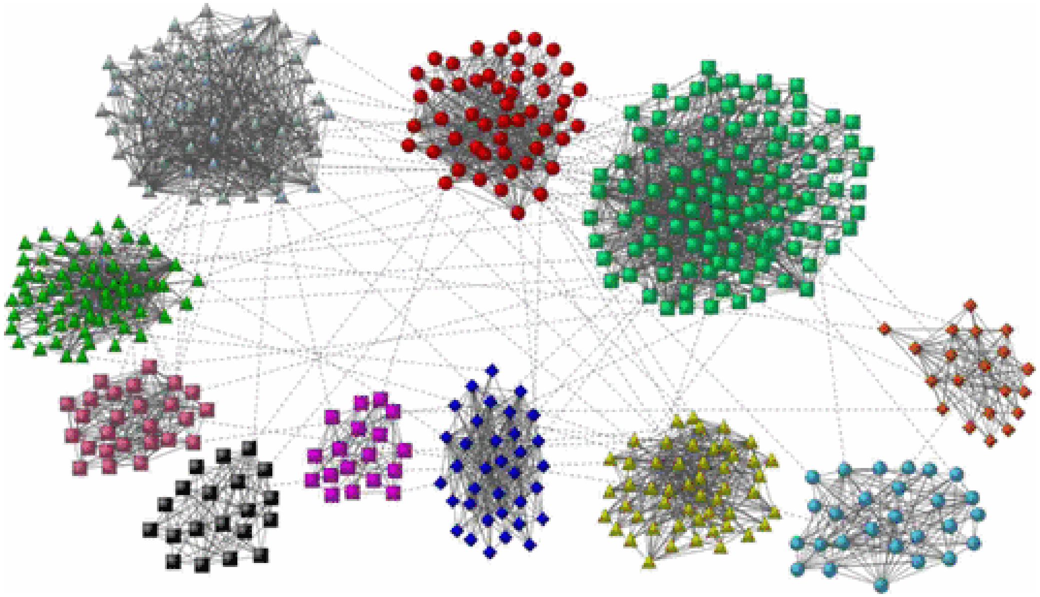
Group and Link Level Analysis (Assignment)

- Identify and visualize the largest component in the given network. Describe its size and other characteristics.
- You need to report
  - The number of communities
  - Size of the largest component
  - You also can use R

Network Level Analysis (Assignment)

- Analyze the whole network and calculate all the following network measures. Compare them with the Group's network topologies.
- You need to report
  - Size of the network
  - Centralization score (degree)
  - Average degree
  - Average path length
  - Clustering coefficient
  - Degree distribution

Community Detection and Evaluation



Community Detection and Evaluation

- **Community**: It is formed by individuals such that those within a group interact with each other more frequently than with those outside the group

group, *cluster*, *cohesive subgroup*, *module* in different contexts

- Community detection: discovering groups in a network where individuals' group memberships are not explicitly given

Why communities in social media?

- *Human beings are social*
- *Easy-to-use social media allows people to extend their social life in unprecedented ways*
- *Difficult to meet friends in the physical world, but much easier to find friend online with similar interests*
- *Interactions between nodes can help determine communities*

Community detection

- Community detection in graphs aims to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology.
- First attempt dates back to 1955 by Weiss and Jacobson searching for work groups within a government agency.

COMMUNITIES – APPLICATION DOMAINS

- Social communities have been studied for a long time (Coleman, 1964; Freeman, 2004; Kottak, 2004; Moody and White, 2003).
- In biology - protein-protein interaction networks, communities are likely to group proteins having the same specific function within the cell (Chen, 2006; Rives and Galitski 2003; Spirin and Mirny, 2003),
- World Wide Web: communities correspond to groups of pages dealing with the same or related topics (Dourisboure et al., 2007; Flake et al., 2002),
- Metabolic networks they may be related to functional modules such as cycles and pathways (Guimera and Amaral, 2005; Palla et al., 2005),
- In food webs they may identify compartments (Krause et al., 2003; Pimm, 1979)

Community Detection and Evaluation

Two types of groups in social media

– **Explicit Groups**: formed by user subscriptions

– **Implicit Groups**: implicitly formed by social interactions

Some social media sites allow people to join groups, is it necessary to extract groups based on network topology?

– Not all sites provide community platform

– Not all people want to make effort to join groups

– Groups can change dynamically

• Network interaction provides rich information about the relationship between users

– Can complement other kinds of information

– Help network visualization and navigation

– Provide basic information for other tasks

COMMUNITY DETECTION IN GRAPHS

- How can we extract the inherent communities of graphs?
- Typically, a two-step approach

1. Specify a quality measure (evaluation measure, objective function) that quantifies the desired properties of communities
2. Apply algorithmic techniques to assign the nodes of graph into communities, optimizing the objective function

- Several measures for quantifying the quality of communities have been proposed
- They mostly consider that communities are set of nodes with many edges between them and few connections with nodes of different communities
  - Many possible ways to formalize it

Taxonomy of Community Criteria

Roughly, community detection methods can be divided into 4 categories:

- **Node-Centric Community**

– Each node in a group satisfies certain properties

- **Group-Centric Community**

– Consider the connections within a group as a whole. The group has to satisfy certain properties without zooming into node-level

- **Network-Centric Community**

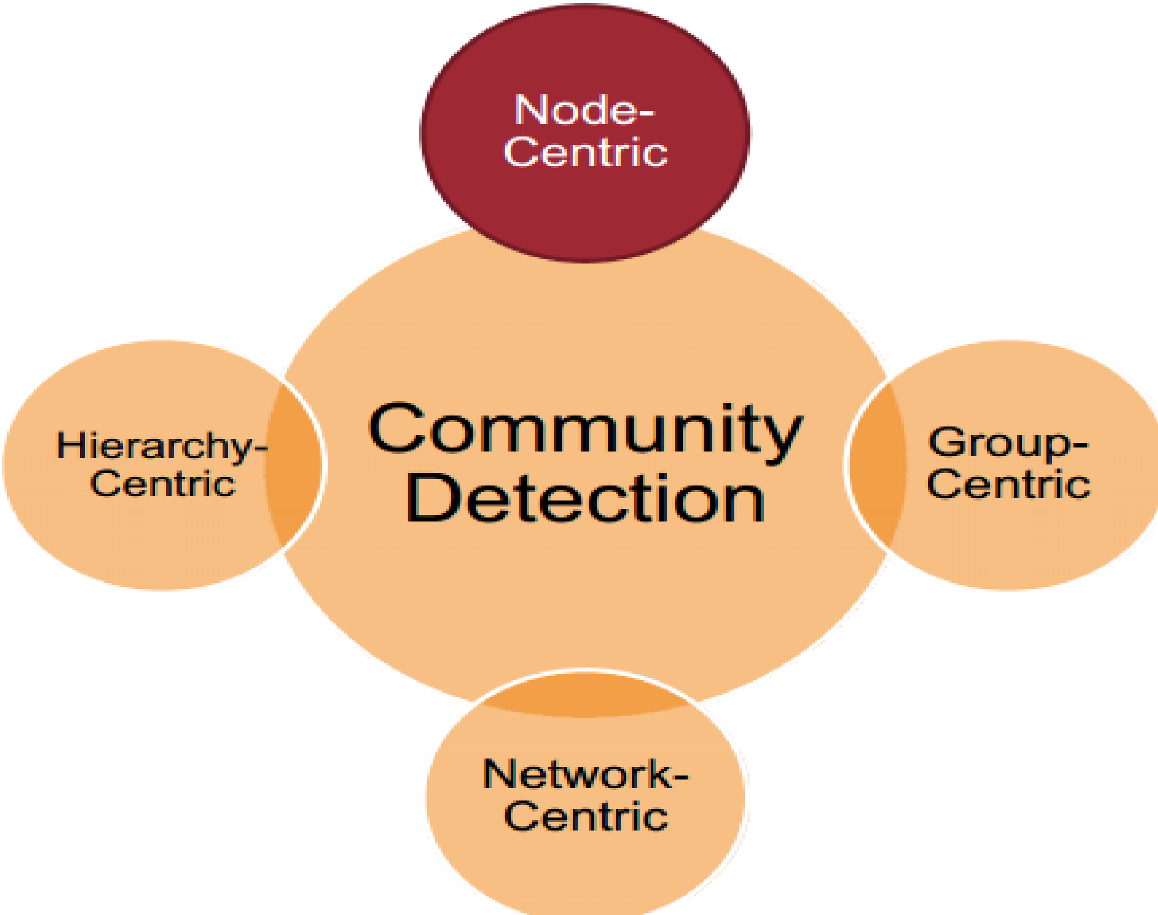
– Partition the whole network into several disjoint sets

- **Hierarchy-Centric Community**

– Construct a hierarchical structure of communities

Community Detection and Mining in Social Media. Lei Tang and Huan Liu, Morgan & Claypool? 2010

Node-Centric Community Detection



NODE-CENTRIC COMMUNITY DETECTION

Nodes satisfy different properties

- **Complete Mutuality**

cliques

- **Reachability of members**

k-clique, k-clan, k-club

- **Nodal degrees**

k-plex, k-core

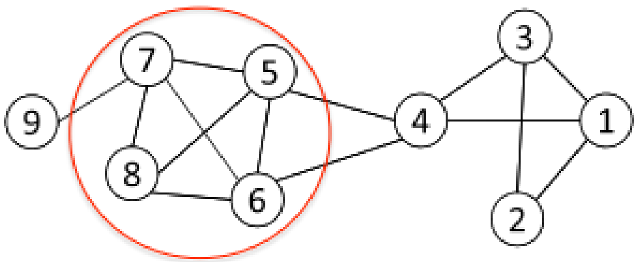
- **Relative frequency of Within-Outside Ties**

LS sets, Lambda sets

Commonly used in traditional social network analysis

Complete Mutuality: Clique

- **Clique**: A maximal complete subgraph of three or more nodes all of which are adjacent to each other



Nodes 5, 6, 7 and 8 form a clique

- computationally complex to find the maximal clique
- *Recursive pruning*:

To find a clique of size ***k***, remove those nodes with less than *k-1* degrees

- Very strict definition, unstable

FINDING THE MAXIMUM CLIQUE

For a clique of size k, each node maintains degree **>=** k**-1**

- Nodes with degree **<** k**-1** will not be included in the maximum clique
- Recursively apply the following **pruning** procedure:

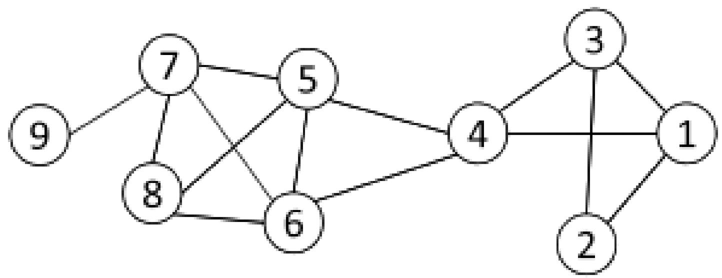
– Sample a sub-network from the given network, and find a clique in the sub-network

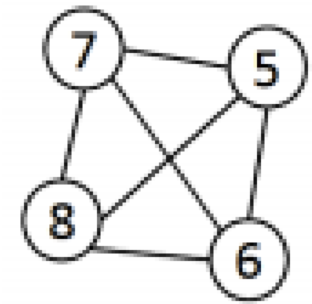
– The maximum clique found on the sub-network (say, it contains k nodes) serves as the lower bound for pruning. That is, the maximum clique in the original network should contain at least k members

– In order to find a clique of size larger thank, the nodes with degree less than or equal to $k-1$, in conjunction with their connections can be removed from future consideration.

- Repeat until the network is small enough
- As social media networks follow a power law distribution for node degrees, i.e., the majority of nodes have a low degree, this pruning strategy can reduce the network size significantly

Maximum Clique Example





- Suppose we sample a sub-network with nodes ***{1-5}*** and find a maximal clique ***{1, 2, 3}*** or ***{1,3,4}*** of size ***3***
- In order to find a clique ***>3***, remove all nodes with degree ***<=3-1=2***

– Remove nodes ***2*** and ***9***

– Remove nodes ***1*** and ***3***

– Remove node ***4***

Maximum Clique Example

> g <- graph.formula(1-1,1-3,3-2,4-1,4-5,4-6,5-7,5-6,5-8,6-7,6-8,7-8,7-9,1-2,3-4)

> plot(g)

> cliques(g)[sapply(cliques(g), length) == 3]

> clique.number(g)

[1] 4

> largest.cliques(g)

[[1]]

[1] 8 5 6 7

 # Defining the function

> c1 <- function(g){

\+ return(induced.subgraph(g, largest.cliques(g)[[1]]))

\+ }

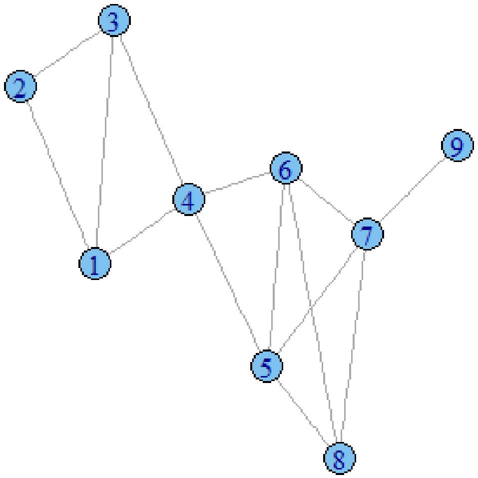
> kk<-c1(g)

> V(kk)

Vertex sequence:

[1] "5" "6" "7" "8"

> plot(kk, vertex.label=largest.cliques(g)[[1]])



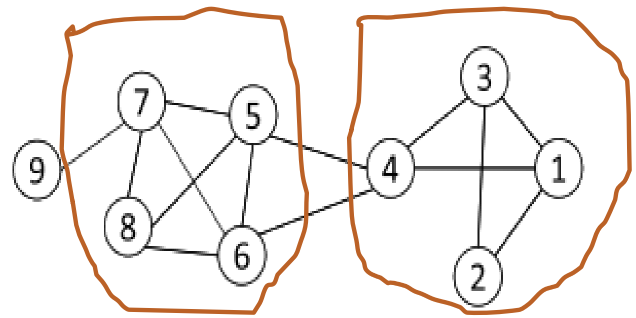
Clique Percolation Method (CPM)

Normally use cliques as **a core or seed** to explore larger communities

CPM is such a method to find **overlapping** communities *(Palla et al.,2005)*

- **Input**
  - A parameter ***k***, and a network
- **Procedure**
  - Find out all cliques of size ***k*** in a given network
  - Construct a clique graph. Two cliques are adjacent if they share ***k-1*** nodes
  - Each connected components in the clique graph form a community

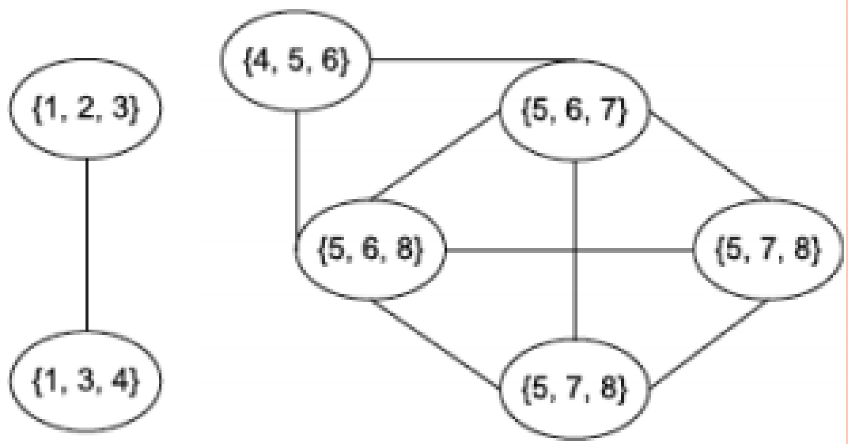
CPM EXAMPLE



For ***k=3***, we can identify all the cliques of size 3 as follows:

{1, 2, 3} {1, 3, 4} {4, 5, 6} {5, 6, 7} {5, 6, 8} {5, 7, 8} {6, 7, 8}

Two cliques are connected as long as they share k−1 (2 in our case) nodes



Communities:
{1, 2, 3, 4}
{4, 5, 6, 7, 8}

Reachability of members

This type of community considers the reachability among actors.

In the extreme case, two nodes can be considered as belonging to one community if there exists a path between the two nodes. Thus each connected component is a community.

The components can be efficiently identified in ***O(n+m)*** time (*Hopcroft and Tarjan,1973*), linear with respect to number of nodes and edges in a network.

Any node in a group should be reachable in ***k*** hops

REACHABILITY OF MEMBERS. K-CLIQUE, K-CLUB

k-clique: a maximal subgraph in which the largest geodesic distance between any nodes **<= k.**

The set is **maximal** in the sense that no other node in the graph is distance ***k*** or less from every other node in the subgraph.

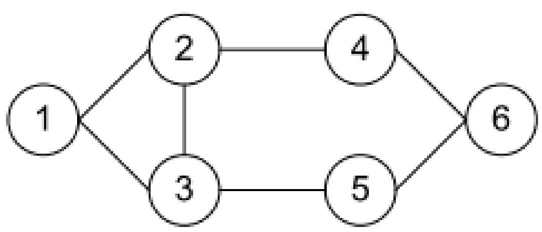
This corresponds to being "a friend of a friend"

Note that the geodesic distance is defined on the original network. Thus, the geodesic is not necessarily included in the group structure

A *k-clique* can have geodesic distance larger than ***k*** in the subgraph

e.g., 2-clique {1,2,3,4,5}, but geodesic distance between nodes 4 and 5 within the group is 3

k-club: a substructure of diameter **<= k**



Cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}
2-clubs: {1,2,3,4}, {1, 2, 3, 5}, {2, 3, 4, 5, 6}

Node-Centric Community Detection

NODAL DEGREES: K-PLEX

Each node should have a **certain number of connections** to nodes within the group

- **k-plex**: for a group with $\boldsymbol{n_s}$ nodes, each node should be adjacent no fewer than $\boldsymbol{n_s\text{-}k}$ in the group
  - *if A has ties with B and C, but not D; while both B and C have ties with D, all four actors could fall in clique under the K-Plex approach*
  - *Rather than the large and "stringy" groupings sometimes produced by n-clique analysis, k-plex analysis tends to find relatively large numbers of smaller groupings. This tends to focus attention on overlaps and co-presence (centralization) more than solidarity and reach*

Nodal Degrees: k-core

- **k-core**: a substructure that each node connects to at least ***k*** members within the group (each vertex has at least degree ***k***).
- Hence, every member of a 2-core is connected to at least 2 other members, and no node outside the 2-core is connected to 2 or more members of the core (otherwise it would not be maximal).

The definitions are complementary

A **k-core** is a **(n_s-k)-plex**

The **coreness** of a vertex is ***k*** if it belongs to the ***k***-core but not to the (*k*+*1*)-core.

K-CORE EXAMPLE

```
> cores<-graph.coreness(g)
> cores
1 2 3 4 5 6 7 8 9
2 2 2 2 3 3 3 3 1
```

 # This function calculates the coreness for each vertex

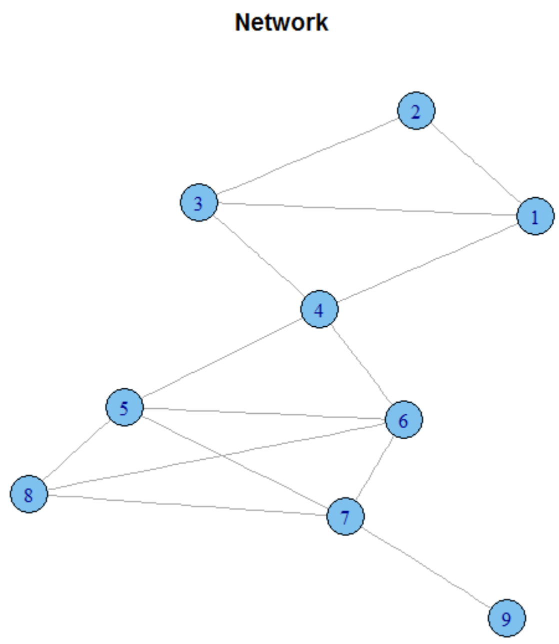
```
> maxCoreness <- max(cores)
> maxCoreness
[1] 3
```

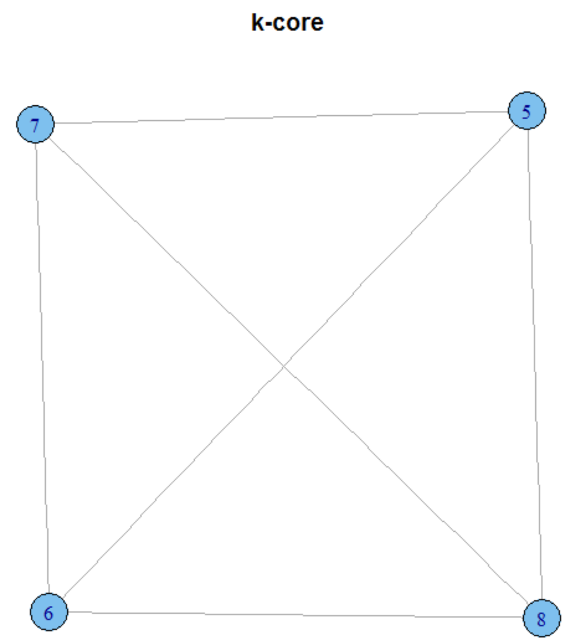
Getting k-core subgraph

```
> verticesHavingMaxCoreness <- which(cores == maxCoreness)
>  verticesHavingMaxCoreness
5 6 7 8
5 6 7 8

>  kcore <- induced.subgraph(graph=g,vids=verticesHavingMaxCoreness)
> str(kcore)
IGRAPH UN-- 4 6 --
+ attr: name (v/c)
+ edges (vertex names):
5 -- 6, 7, 8
6 -- 5, 7, 8
7 -- 5, 6, 8
8 -- 5, 6, 7
```

Extract the Maximal-core

```
> par(mfrow=c(1, 2))
> plot(kcore, main="k-core",
+ vertex.label=get.vertex.attribute(kcore,name='vert.names',index=V(kcore)))
> plot(g,main="Network")
```





Recap of Node-Centric Communities

- Each node has to satisfy **certain properties**

- Complete mutuality

- Reachability

- Nodal degrees

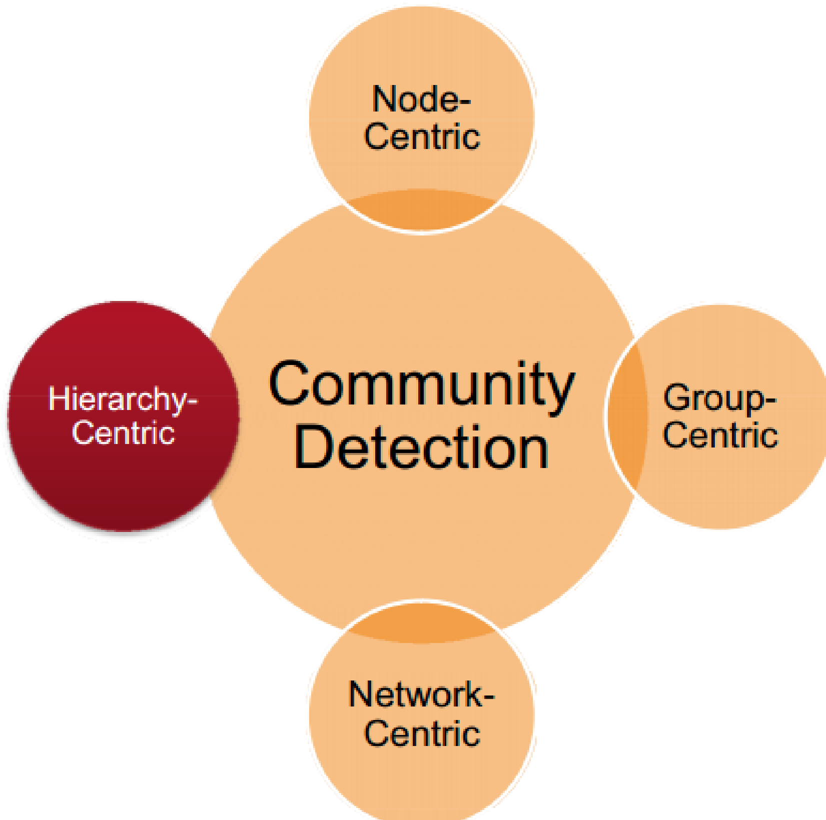
- **Limitations**:

- Too strict, but can be used as the core of a community

- Not scalable, commonly used in network analysis with small-size network

- Sometimes not consistent with property of large-scale networks - e.g., nodal degrees for scale-free networks

Hierarchy-Centric Community Detection



Hierarchy-Centric Community Detection

- **Goal**: build a hierarchical structure of communities based on network topology

- Allow the analysis of a network at **different resolutions**

- Representative **approaches**:
  - Divisive Hierarchical Clustering
  - Agglomerative Hierarchical clustering

Divisive Hierarchical Clustering

- Divisive clustering
  - Partition nodes into several sets
  - Each set is further divided into smaller ones
  - Network-centric partition can be applied for the partition

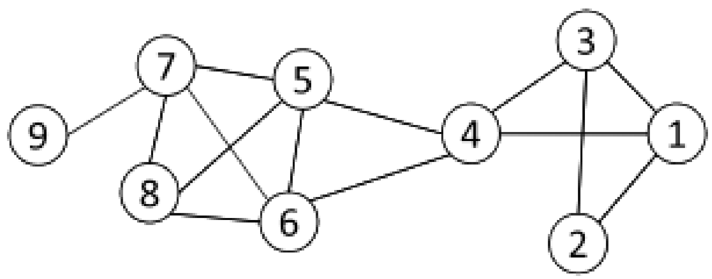
- One particular example: **recursively remove the "weakest" tie**
  - Find the edge with the least strength
  - Remove the edge and update the corresponding strength of each edge

- Recursively apply the above two steps until a network is discomposed into desired number of connected components.
- Each component forms a community

Edge Betweenness

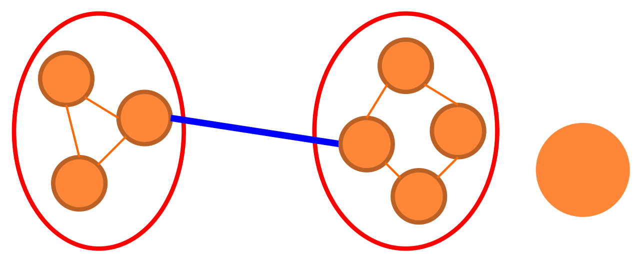
- The strength of a tie can be measured by edge betweenness
- **Edge betweenness** is a measure to count how many shortest paths between pair of nodes pass along the edge, and this number is expected to be large for those between-group edges (*Brandes, 2001*)

$$\text{edge-betweenness}(e) = \Sigma_{s<t} \frac{\sigma_{st}(e)}{\sigma_{s,t}}$$



The edge betweenness of e(1, 2) is 4.
Since all the shortest paths from node 2 to any node in {4, 5, 6, 7, 8, 9} has either to pass e(1, 2) or e(1, 3), leading to a weight of 6×1/2=3 for e(1,2).
e(1,2) is the shortest path between 1 and 2
Hence, the betweenness of e(1,2) is 3+1=4

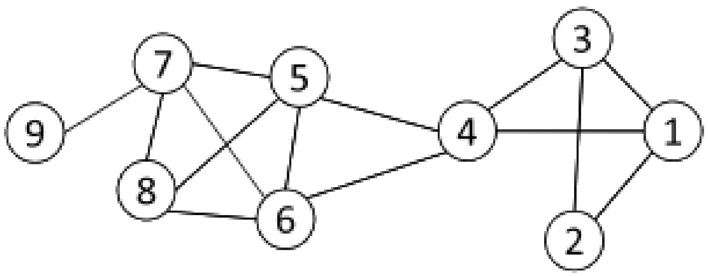
- The edge with higher betweenness tends to be the bridge between two communities.



Hierarchy-Centric Community Detection

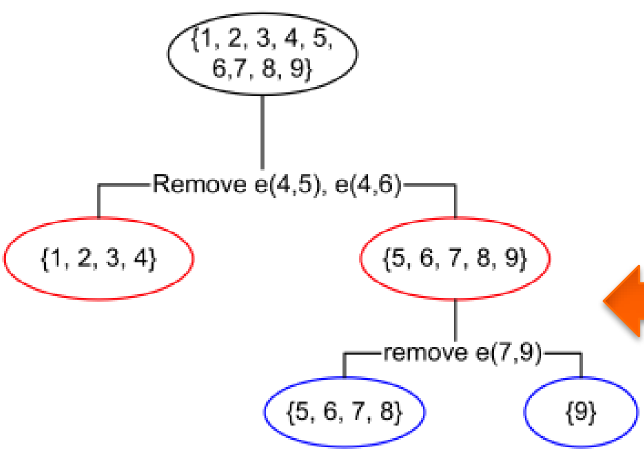
Divisive clustering based on edge betweenness

(Community structure in social and biological networks, Newman and Girvan, 2004)



Initial betweenness value

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0



After remove e(4,5), the betweenness of e(4, 6) becomes 20, which is the highest;

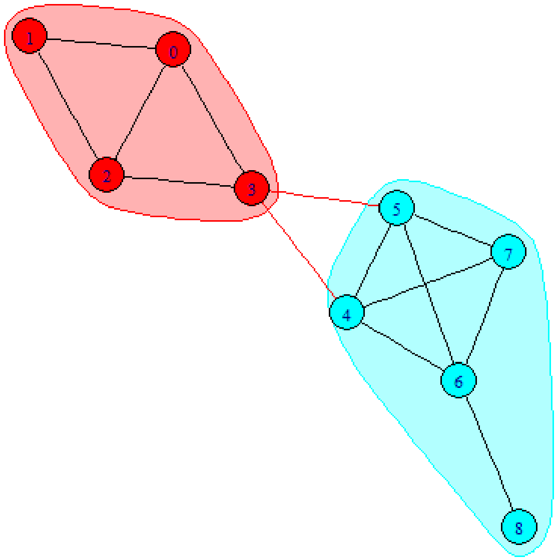
After remove e(4,6), the edge e(7,9) has the highest betweenness value 4, and should be removed.

Hierarchy-Centric Community Detection

Divisive Clustering Based on Edge Betweenness Example

edge.betweenness.community(graph)

 # returns a numeric vector, the edge betweenness value of the removed edges



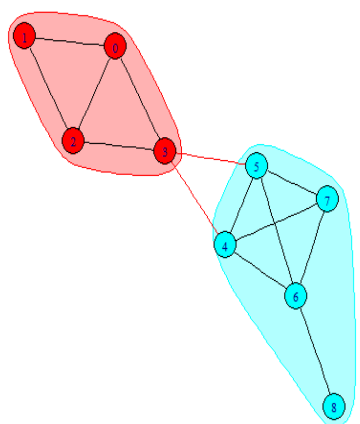
DIVISIVE CLUSTERING BASED ON EDGE BETWEENNESS EXAMPLE

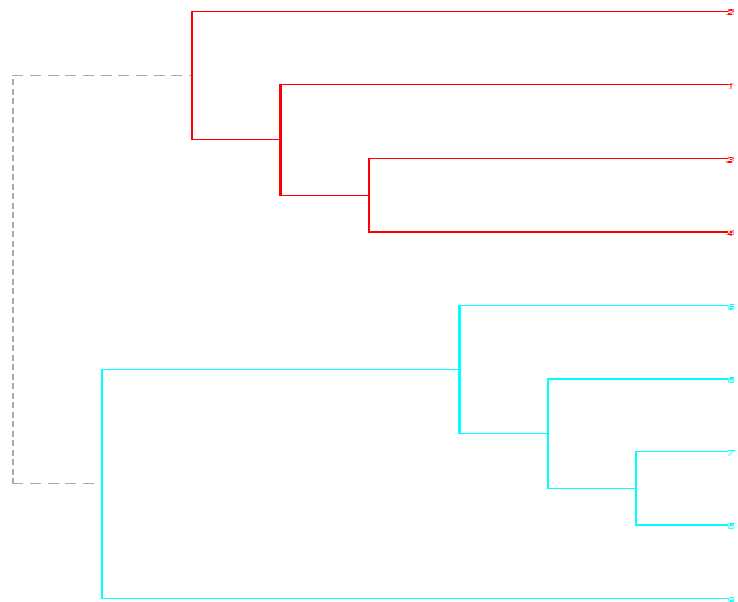
```
> bet<-edge.betweenness.community(g)
> system.time(bet<-edge.betweenness.community(g))
```

```
> bet$membership
[1] 1 1 1 1 2 2 2 2 2
```

```
> bet$removed.edges
 [1]  6  7 14  1  4  2  3  5  8  9 10 11 12 13
```

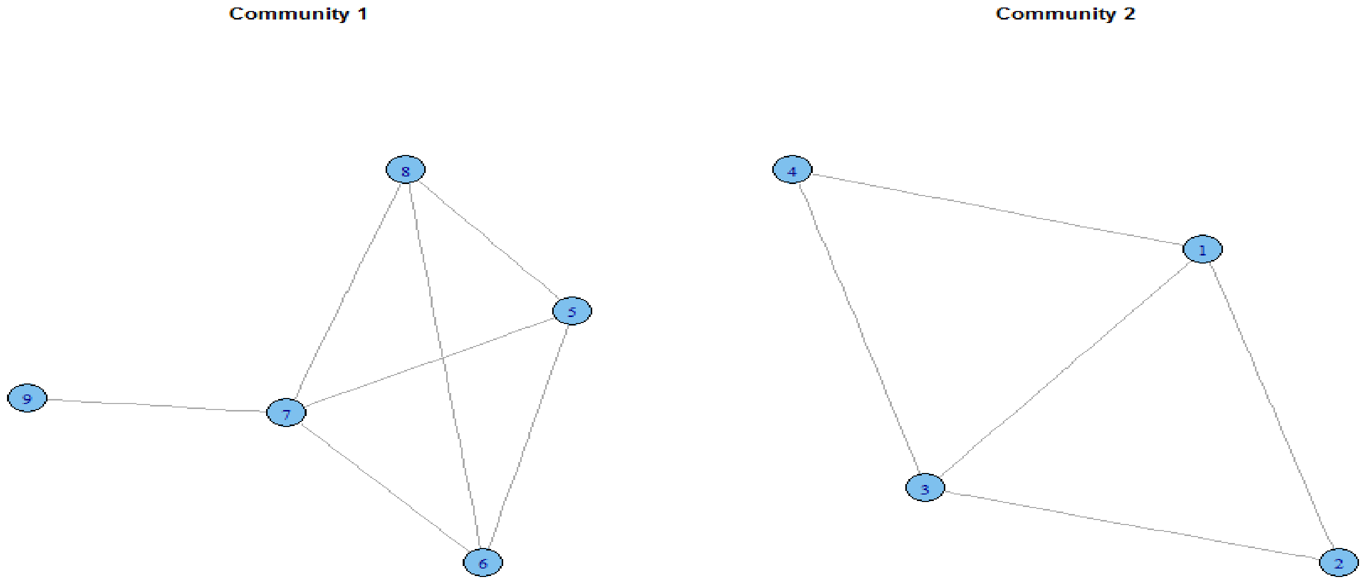
```
> plot(bet, g)
> library(ape)
> de
```





Divisive Clustering based on edge Betweenness Example

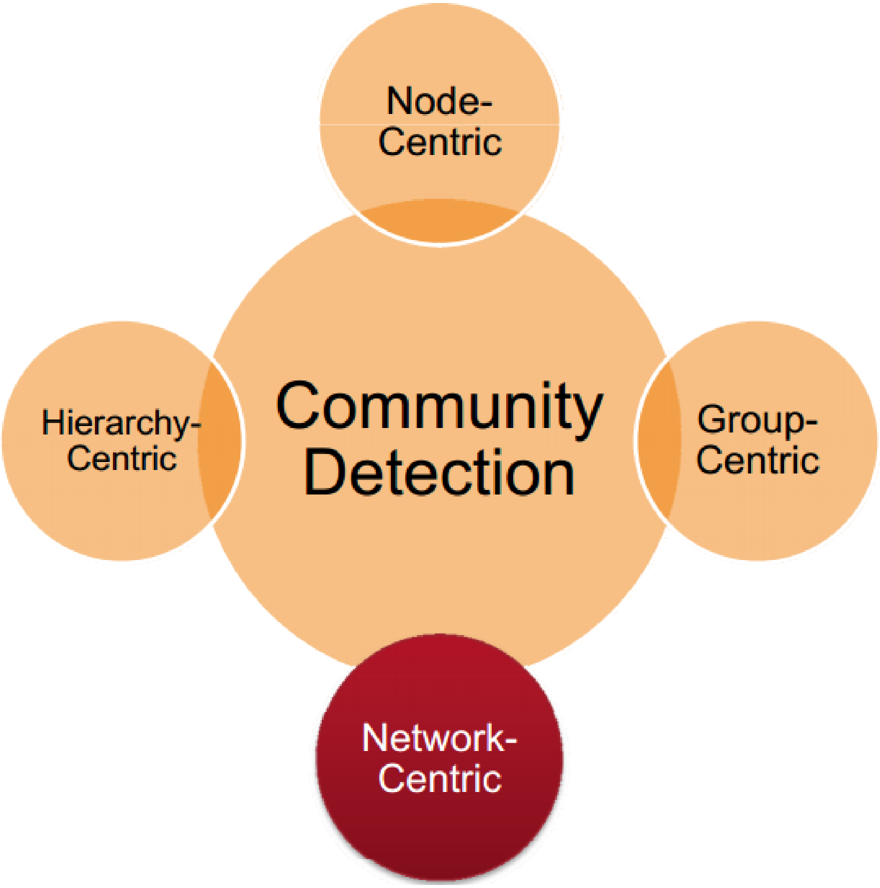
```
> kk<-induced.subgraph(g, bet$membership==1)
> kk1<-induced.subgraph(g, bet$membership==2)
> plot(kk1,vertex.label=V(g)[bet$membership==2])
> plot(kk,vertex.label=V(g)[bet$membership==1])
```



Recap of Hierarchical Clustering

- Most hierarchical clustering algorithm output a binary tree
  - Each node has two children nodes
  - Might be highly imbalanced
- Divisive clustering is more stable, but generally more computationally expensive

Network-Centric Community Detection



Network-Centric Community Detection

To form a group, we need to consider the **connections of the nodes globally**.

Goal: **partition the network into disjoint sets**

- Groups based on **Modularity Maximization**
- Groups based on **Node Similarity**
- Groups based on **Cut Minimization**

Modularity. Main Idea

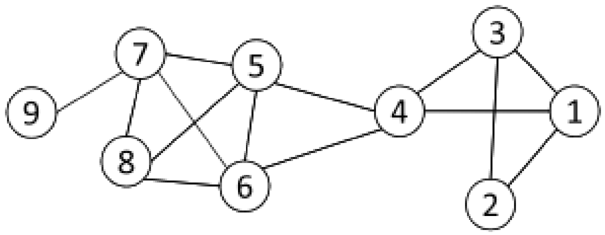
- Modularity (*Newman and Girvan 2004*; *Newman, 2006*) measures the strength of a community partition by taking into account the **degree distribution**

- Initially introduced as a measure for assessing the strength of communities

Q = (fraction of edges within communities) –
(expected number of edges within communities)

MODULARITY

- What is the **expected** number of edges?
  - **Random graphs** are not expected to present inherent community structure
- Consider a configuration model
  - Random graph (with *m* edges) model with the same degree distribution
  - Let $\mathbf{P_{ij}}$ = **probability** of an edge between nodes **i** and **j** with degrees $\mathbf{d_i}$ and $\mathbf{d_j}$ respectively
  - Then $\mathbf{P_{ij} = d_i\, d_j / 2m}$



The expected number of edges between nodes 1 and 2 is 3*2/ (2*14) = 3/14

- So $A_{ij} - d_i d_j / 2m$ measures how far the true network interaction between nodes *i* and *j* (A*ij*) deviates from the expected random connections.

MODULARITY

- Given a group of nodes ***C***, the strength of community effect is defined as

$$\sum_{i \in C, j \in C} A_{ij} - d_i d_j / 2m.$$

- To partition the group into multiple groups, we maximize

$$Q = \frac{1}{2m} \sum_{\ell=1}^{k} \sum_{i \in C_\ell, j \in C_\ell} \left(A_{ij} - d_i d_j / 2m \right)$$

where the coefficient **1/2m** is introduced to normalize the value between **-1** and **1**.

- **Modularity** calibrates the quality of community partitions thus can be used as an objective measure to maximize

Properties of modularity

- **Larger** modularity Q indicates **better** communities (more than random intra-cluster density)
- The community structure would be better if the number of internal edges exceed the expected number
- Modularity value is always **smaller than** 1
- It can also take **negative** values,
  - E.g., if each node is a community itself
  - No partitions with positive modularity → No community structure
  - Partitions with large negative modularity → Existence of subgraphs with small internal number of edges and large number of inter-community edges

[Newman and Girvan '04], [Newman '06], [Fortunato '10]

APPLICATIONS OF MODULARITY

- Modularity can be applied:
  - As **quality function** in clustering algorithms
  - As **evaluation measure** for comparison of different partitions or algorithms
  - As a community detection tool itself
    - **Modularity optimization**
  - As criterion for reducing the size of a graph (size reduction preserving modularity [Arenas et al. '07])

Modularity-based community detection

- Modularity was first applied as a **stopping criterion** in the Newman-Girvan algorithm
- **Basic steps**:

1. Compute betweenness centrality for all edges in the graph

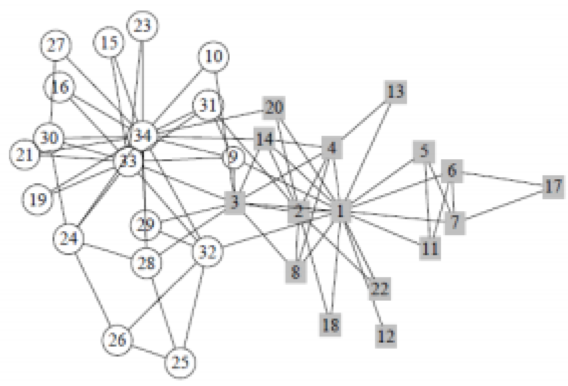
2. Find and remove the edge with the highest score

3. Recalculate betweenness centrality score for the remaining edges

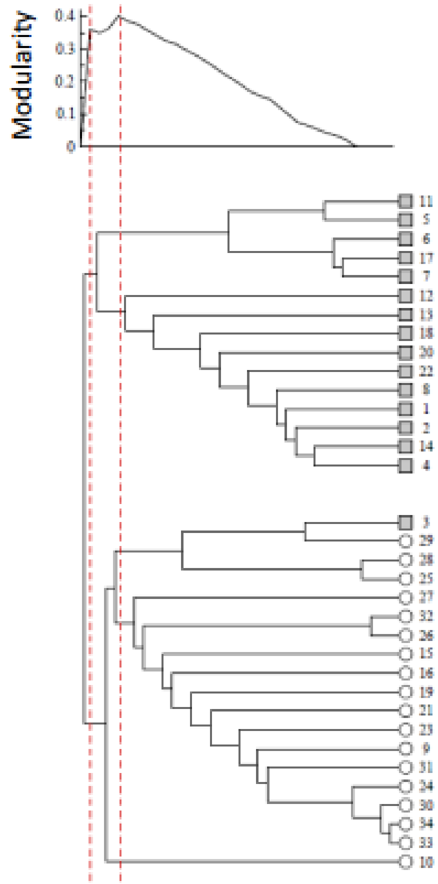
4. Go to step 2

- How do we know if the produced communities are **good ones** and stop the algorithm?
- The output of the algorithm is in the form of a **dendrogram**
- Use **modularity** as a criterion to cut the dendrogram and terminate the algorithm (Q~=0.3-0.7 indicates good partitions)
- Complexity: **$O(m^2n)$** (or **$O(n^3)$** on a sparse graph)

NEWMAN-GIRVAN ALGORITHM



Zachary's karate club



Community structure

[Newman and Girvan '04]

Modularity optimization

- High values of modularity indicate good quality of partitions
- **Goal**: find the partition that corresponds to the maximum value of modularity
- **Modularity maximization** problem:
  - Computational difficult problem [Brandes et al. '06]
  - Approximation techniques and heuristics
- Four main categories of techniques
  1. Greedy techniques
  2. Spectral optimization
  3. Simulated annealing
  4. Extremal optimization

[Fortunato '10]

Newman's algorithm (Greedy techniques 1)

Fast algorithm for detecting community structure in networks, Newman, 2004

- Newman's algorithm [Fast algorithm for detecting community structure in networks, Newman, 2004]
- Agglomerative (bottom-up) hierarchical clustering algorithm
- **Idea**: Repeatedly join pairs of communities that achieve the greatest increase of modularity (dendrogram representation)

  1. Initially, each node of the graph belongs on its own cluster (n)
  2. Repeatedly, join communities in pairs by adding edges:
     a. At each step, choose the pairs that achieve the **greatest increase** (or minimum decrease) of modularity
     b. Consider only pairs of communities **between which there exist edges** (merging communities that do not share edges, it can never improve modularity)
- Complexity: **O((m+n) n)** (or **O(n^2)** on a sparse graph)

Modularity Optimization Example

optimal.community (graph, weights = NULL)

 # This function calculates the optimal community structure of a graph, by maximizing the modularity measure over all possible partitions

```
> oc <- optimal.community(g)
> oc
Graph community structure calculated with the optimal algorithm
Number of communities: 2
Modularity: 0.3469388
Membership vector:
0 1 2 3 4 5 6 7 8
1 1 1 1 2 2 2 2 2
```

```
> system.time(bet<-edge.betweenness.community(g))
> system.time(oc <- optimal.community(g))
```

GREEDY TECHNIQUES (2)

- Can we improve the complexity of Newman's algorithm?
- Greedy optimization algorithm by Clauset, Newman and Moore [Clauset et al. '04]
- **Key point**: large graphs are **sparse**
- Exploit sparsity by using appropriate **data structures** for sparse graphs (e.g., max-heaps)
  - A sparse matrix for storing the variations of modularity $\Delta Q_{i,j}$ after joining two communities i, j (in the case they are connected by an edge)
  - A max-heap data structure for the largest element of each row of matrix $\Delta Q_{i,j}$ (fast update time and constant time for findmax() operation)
- Complexity: O($m\ d \log n$), d is the depth of the dendrogram describing the performed partitions (the community structure)

Greedy optimization algorithm

Finding community structure in very large networks, Aaron Clauset,1 M. E. J. Newman,2 and Cristopher Moore, 2004

fastgreedy.community(graph, merges=TRUE, modularity=TRUE, membership=TRUE, weights=E(graph)$weight)

 # This function tries to find dense subgraph, also called communities in graphs via directly optimizing a modularity score

Greedy Optimization Algorithm Example

> fcs <- fastgreedy.community(g)

```
List of 2
 $ merges     : num [1:8, 1:2] 7 2 3 12 5 6 10 16 9 1 ...
 $ modularity: num [1:9] -0.1224 -0.0612 -0.0051 0.0995 0.1607 ...
 - attr(*, "class")= chr "communities"
```

> str(fcs)

```
> fcs
Graph community structure calculated with the fast greedy algorithm
Number of communities (best split): 2
Modularity (best split): 0.3469388
Membership vector:
0 1 2 3 4 5 6 7 8
2 2 2 2 1 1 1 1 1
```

The merges performed by the algorithm will be stored here. Each merge is a row in a two-column matrix and contains the ids of the merged communities. Communities are numbered from zero. In each step a new community is created from two other communities and its id will be one larger than the largest community id so far. This means that before the first merge we have n communities.

```
> fcs$merges
     [,1] [,2]
[1,]    7    9
[2,]    2    1
[3,]    3   11
[4,]   12    4
[5,]    5    8
[6,]    6   14
[7,]   10   15
[8,]   16   13
```

> dendPlot(fcs, mode="phylo")

SPECTRAL OPTIMIZATION ALGORITHM

M.E.J. NEWMAN. FINDING COMMUNITY STRUCTURE USING THE EIGENVECTORS OF MATRICES

leading.eigenvector.community(graph, steps = -1, weights = NULL, start = NULL, options = igraph.arpack.default, callback = NULL, extra = NULL, env = parent.frame())

 # The function documented in these section implements the 'leading eigenvector' method developed by Mark Newman.

- The leading eigenvector method works by calculating the eigenvector of the modularity matrix for the largest positive eigenvalue and then separating vertices into two community based on the sign of the corresponding element in the eigenvector. If all elements in the eigenvector are of the same sign that means that the network has no underlying community structure.
    - Directed edges: FALSE
    - Weighted edges: FALSE
    - Handles multiple components: TRUE
    - Runtime: c|V|^2 + |E|

NODE SIMILARITY

- Node similarity is defined by how similar their interaction patterns are

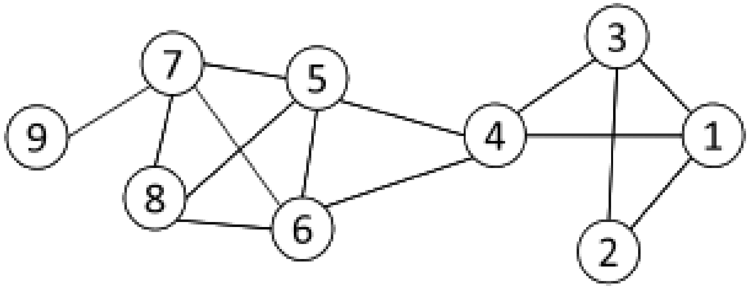
Or

- Node similarity is defined in terms of the similarity of their neighborhood

A key related concept is **structural equivalence:**

two nodes are structurally equivalent if they are connecting to **the same set of actors**

Nodes 1 and 3 are structurally equivalent; So are nodes 5 and 7.



Node Similarity

- Groups are defined over equivalent nodes

- Too strict
- Rarely occur in a large-scale
- Relaxed equivalence class is difficult to compute

- In practice, use **vector similarity**
- e.g.,
  - Cosine similarity (*Hopcroft et al.,2003*),
  - Jaccard similarity (*Gibson et al.,2005*)

VECTOR SIMILARITY

- For two nodes $\boldsymbol{v_i}$ and $\boldsymbol{v_j}$ in a network, the similarity between the two are defined as

$$Jaccard(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \qquad Cosine(v_i, v_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_j| \cdot |N_j|}}$$

$N_4=\{1,3,5,6\}$, and $N_6=\{4,5,7,8\}$

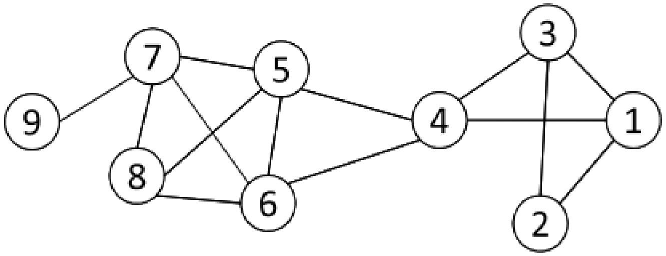
Thus, the similarity between the two nodes are:

$$Jaccard(4,6) = \frac{|\{5\}|}{|\{1,3,4,5,6,7,8\}|} = \frac{1}{7} \qquad Cosine(v_i, v_j) = \frac{|\{5\}|}{\sqrt{4 \cdot 4}} = \frac{1}{4}$$

But what about similarity of nodes 7 and 9?

Normal similarity-based methods have to compute the similarity for each pair of nodes, totaling $\boldsymbol{O(n^2)}$.

It is time-consuming when $\boldsymbol{n}$ is very large.

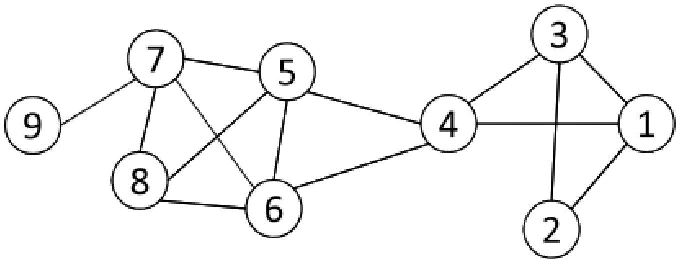


VECTOR SIMILARITY

$N_4=\{1,3,5,6\}$, and $N_6=\{4,5,7,8\}$

Thus, the similarity between the two nodes are:

$$Jaccard(4,6) = \frac{|\{5\}|}{|\{1,3,4,5,6,7,8\}|} = \frac{1}{7} \qquad Cosine(v_i,v_j) = \frac{|\{5\}|}{\sqrt{4 \cdot 4}} = \frac{1}{4}$$



But what about similarity of nodes 7 and 9?

Normal similarity-based methods have to compute the similarity for each pair of nodes, totaling ***O(n²)***.

It is time-consuming when ***n*** is very large.

CUT-MINIMIZATION

The objective function is modified so that the group sizes of communities are considered. Two commonly used variants are *ratio cut* and *normalized cut*.

- Let $\pi=(C_1, C_2, \cdots, C_k)$ be a graph partition such that $C_i \cap C_j = \phi$ and $\cup^k_{i=1} C_i = V$.

The ratio cut and the normalized cut are defined as:

$$\text{Ratio Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{|C_i|},$$

$$\text{Normalized Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{vol(C_i)}$$

C_i: a community
$|C_i|$: number of nodes in C_i
$vol(C_i)$: sum of degrees in C_i

Cut-Minimization

For partition in red: π_1

$C_1 = \{9\}$

$C_2 = \{1, 2, 3, 4, 5, 6, 7, 8\}$

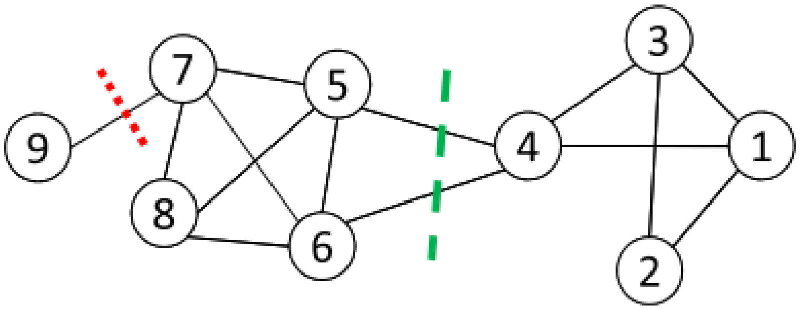
$cut(C_1, \bar{C}_1) = 1$

$|C_1| = 1, |C_2| = 8,$

$vol(C_1) = 1$, and $vol(C_2) = 27$

$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{27}\right) = 14/27 = 0.52$$



$$\text{Ratio Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{|C_i|},$$

$$\text{Normalized Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{vol(C_i)}$$

$C_{i,}$: a community

$|C_i|$: number of nodes in C_i

$vol(C_i)$: sum of degrees in C_i

For partition in green: π_2

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4} + \frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

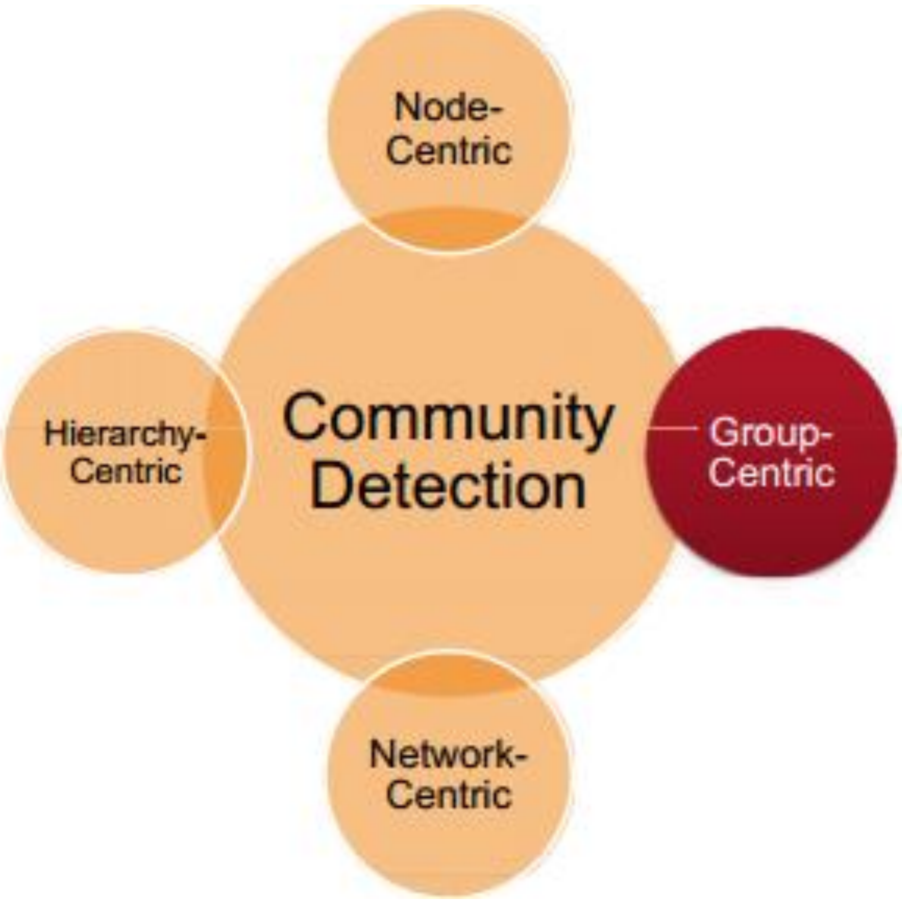
$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12} + \frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

Both ratio cut and normalized cut prefer a balanced partition $\pi 2$

Recap of Network-Centric Community

- **Network-Centric Community Detection**
  - Groups based on Modularity maximization
  - Groups based on Node Similarity
  - Groups based on Cut Minimization
- **Goal**: Partition network nodes into several disjoint sets
- **Limitation**: Require the user to specify the number of communities beforehand

Group-Centric Community Detection



Group-Centric Community Detection

- Consider the connections **within a group as whole** □
- □Some nodes to have low connectivity

One such example is density-based groups:

- A subgraph with V_s nodes and E_s edges is a γ-dense (also called **quasi-clique** *(Abello et al.,2002))* if

$$\frac{|E_s|}{|V_s|(|V_s|-1)/2} \geq \gamma$$

- A similar strategy to that of cliques can be used (Recursive pruning):
- **Local search**: Sample a subgraph, and find a maximal γ-dense quasi-clique (say, of size k)
  - the resultant size = k
- **Heuristic pruning**: Remove nodes that:
  - whose degree <$k\gamma$
  - all their neighbors with degree <$k\gamma$

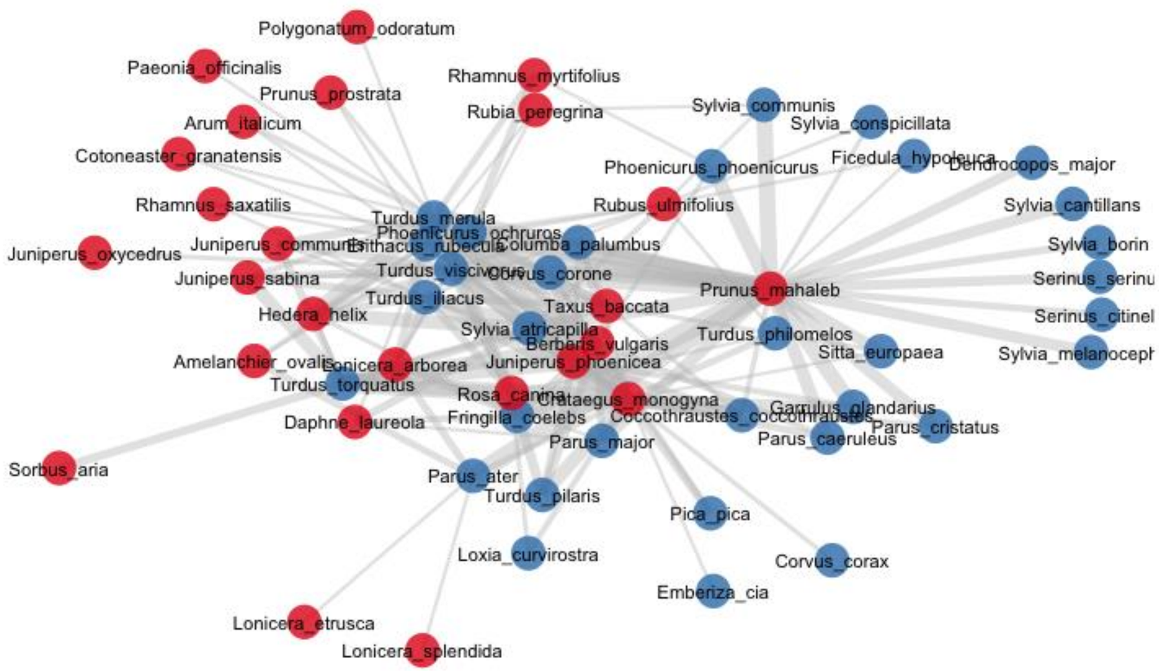
Group-Centric Community Detection

2.3. COMPARING THE COMMUNITY DETECTION ALGORITHMS IN IGRAPH

1. Export Graph file from Gephi to .graphml format and read it in R.
   - *read.graph("path/graph", format=*"graphml"*)*
2. Analyze the different community detection algorithms in IGraph .
3. Results present in the Graphical and Table form (fill the following table)

Function	Runtime	Weighted edges	Directed edges	Calculate the Modularity	Form the Membership List	Reference	Comments
walktrap.community						P. Pons, M. Latapy. Computing communities in large networks using random walks	
label.propagation.community						P. Pons, M. Latapy. Computing communities in large networks using random walks	
spinglass.community						J. Reichardt and S. Bornholdt. Statistical mechanics of community Detection	
leading.eigenvector.community						M.E.J. Newman. Finding community structure using the eigenvectors of matrices	
fastgreedy.community						A. Clauset, M.E.J. Newman, C. Moore. Finding community structure in very large networks	
edge.betweenness.community						M.E.J. Newman and M. Girvan. Finding and evaluating community structure in network	

BIPARTITE NETWORKS



BIPARTITE NETWORKS

A *complete bipartite* graph is a graph G whose vertex set V can be partitioned into two non empty sets:

$$V_1 \text{ and } V_2$$

in such a way that:

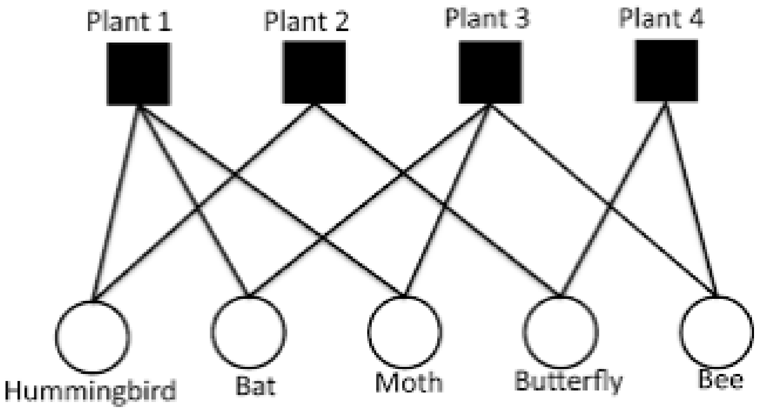
- every vertex in V_1 is adjacent to every vertex in V_2,
- no vertex in V_1 is adjacent to a vertex in V_1,
- and no vertex in V_2 is adjacent to a vertex in V_2.

If V_1 has r vertices and V_2 has s vertices then the complete bipartite graph is written as $K_{r,\,s}$.

BIPARTITE NETWORKS (FIRST SCENARIO)

- two types of actors,
- one type of actor can only interact with the other type of actor.

For example, picture a network of pollination interactions between plants and their pollinators. Pollinators only pollinate plants, so there are no edges between pollinators. Similarly, there are no edges between plants.



BIPARTITE NETWORKS (SECOND SCENARIO)

You have individuals belong to particular groups.
Edges represent membership of individuals in groups.
These are sometimes called ***affiliation*** networks.

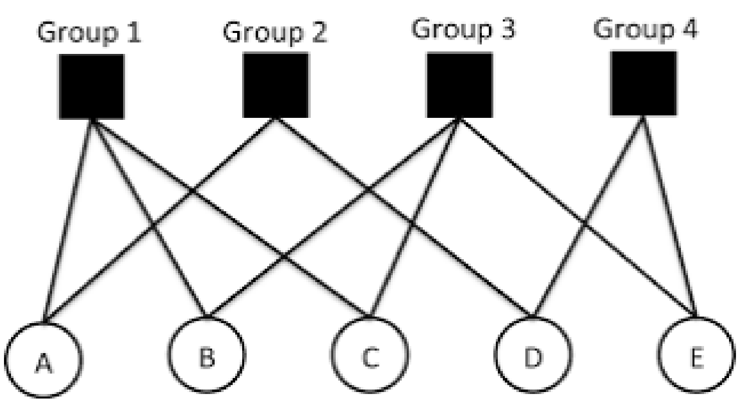
Here, you can imagine:

- 'individuals' being one set of nodes,
- and 'groups' being another set of nodes.

Individuals can only be connected to groups, and vice versa.

In this case, you might actually be interested in generating a social network of individuals based on co-membership. We can do this using what is called a *one-mode projection*, or *bipartite projection*

Examples:
Suppliers-Store (Transportation)
Machine-Jobs (Assignment)
Actors-Films
Authors-Articles
Directors-Boards
Readers-Newspapers
Researchers-Conferences



BIPARTITE NETWORKS

We'll use the example of 5 individuals belonging to 4 different groups.

```
> A=c(1,1,0,0)
> B=c(1,0,1,0)
> C=c(1,0,1,0)
> D=c(0,1,0,1)
> E=c(0,0,1,1)
> bm=matrix(c(A,B,C,D,E),nrow=5,byrow=TRUE)
> bm
     [,1] [,2] [,3] [,4]
[1,]    1    1    0    0
[2,]    1    0    1    0
[3,]    1    0    1    0
[4,]    0    1    0    1
[5,]    0    0    1    1
```

```
> dimnames(bm)=list(c("A","B","C","D","E"),c("Group1","Group2","Group3","Group4"))
```

This should create a matrix that looks like this:

```
> bm
  Group1 Group2 Group3 Group4
A      1      1      0      0
B      1      0      1      0
C      1      0      1      0
D      0      1      0      1
E      0      0      1      1
```

BIPARTITE NETWORKS

- You can convert this directly into an igraph object:

```
> bg=graph.incidence(bm)

> str(bg)
IGRAPH UN-B 9 10 --
+ attr: type (v/l), name (v/c)
+ edges (vertex names):
 [1] A--Group1 A--Group2 B--Group1 B--Group3 C--Group1 C--Group3 D--Group2 D--Group4
 [9] E--Group3 E--Group4
```

The first line always starts with IGRAPH, showing you that the object is an igraph graph. Then a four letter long code string is printed.

The first letter distinguishes between directed ('**D**') and undirected ('**U**') graphs.

The second letter is '**N**' for named graphs, i.e. graphs with the name vertex attribute set. The third letter is '**W**' for weighted graphs, i.e. graphs with the weight edge attribute set. The fourth letter is '**B**' for bipartite graphs, i.e. for graphs with the type vertex attribute set.

Then, after two dashes, the name of the graph is printed, if it has one, i.e. if the name graph attribute is set.

From the second line, the attributes of the graph are listed, separated by a comma. After the attribute names, the kind of the attribute – graph ('**g**'), vertex ('**v**') or edge ('**e**') – is denoted, and the type of the attribute as well, character ('**c**'), numeric ('**n**'), logical ('**l**'), or other ('**x**').

BIPARTITE NETWORKS

Check the attributes

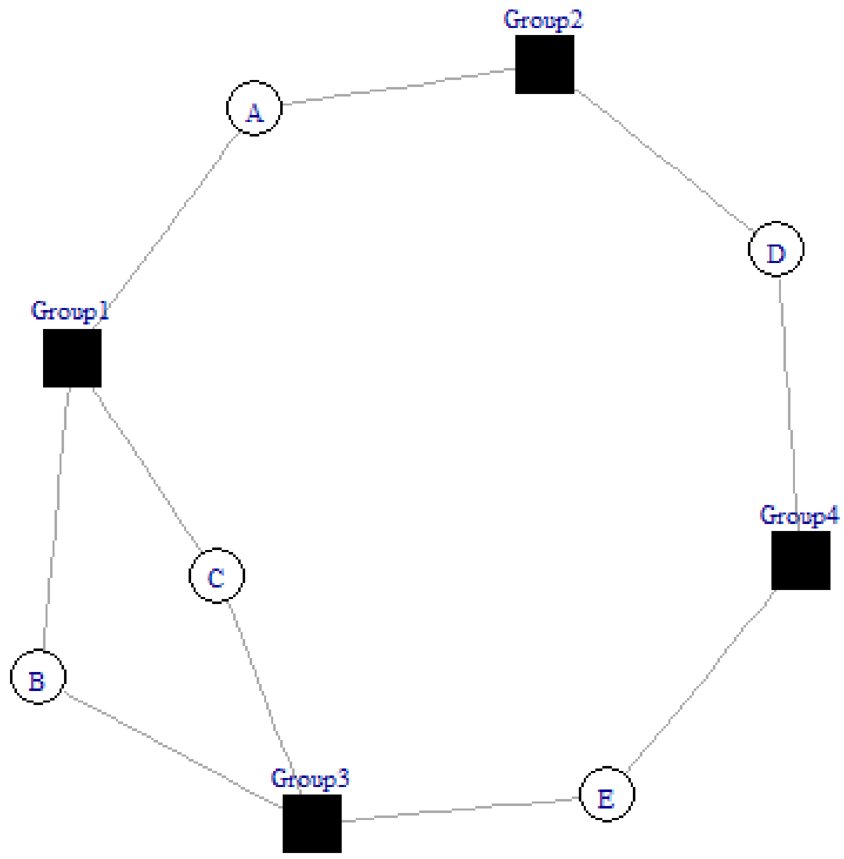
```
> V(bg)$type
[1] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE

> V(bg)$name
[1] "A"      "B"      "C"      "D"      "E"      "Group1" "Group2" "Group3" "Group4"
```

You see that igraph has arranged both individuals and groups as nodes, and then created the attribute 'type' to indicate that these are two distinct classes of nodes

Now let's see what this network looks like

```
> shapes=c("circle","circle","circle","circle","circle",
+ "square","square","square","square")
> labeldistances=c(0,0,0,0,0,0.6,0.6,0.6,0.6)

> plot(bg,vertex.shape=shapes,
+ vertex.label.dist=labeldistances,vertex.color=V(bg)$type)
```

BIPARTITE NETWORKS



Bipartite Projections

Let's say you are interested in the co-membership relations between the individuals.

What you need to do is to create a *one-mode projection* of the bipartite network.

Note that there are actually TWO ways to project a bipartite network:

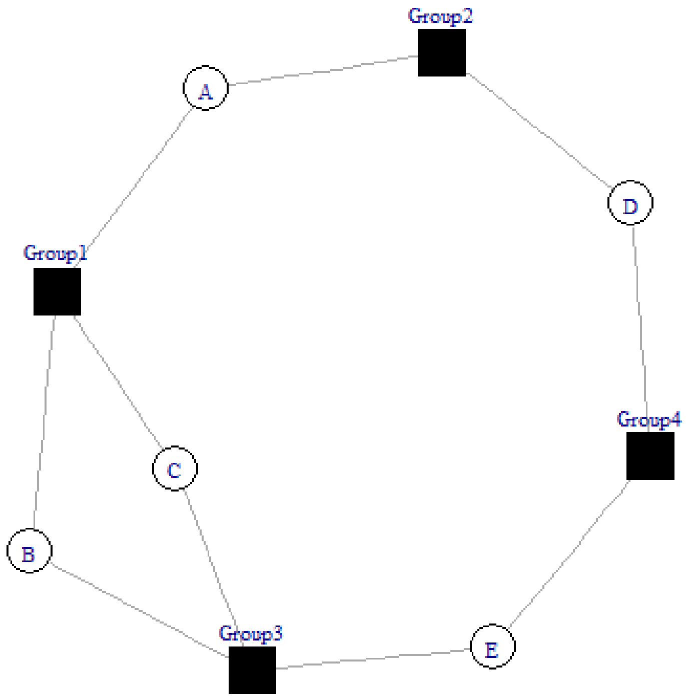
- you can make a co-membership network of nodes,
- or a network of groups that share members.

You can get both of these at once with this function:

```
> pr=bipartite.projection(bg)
> pr
$proj1
IGRAPH UNW- 5 7 --
+ attr: name (v/c), weight (e/n)

$proj2
IGRAPH UNW- 4 4 --
+ attr: name (v/c), weight (e/n)
```

you can see here that the bipartite projection has given us a list object with two graphs:
pr$proj1 (5 vertices and 6 edges)
pr$proj2 (4 vertices and 6 edges)

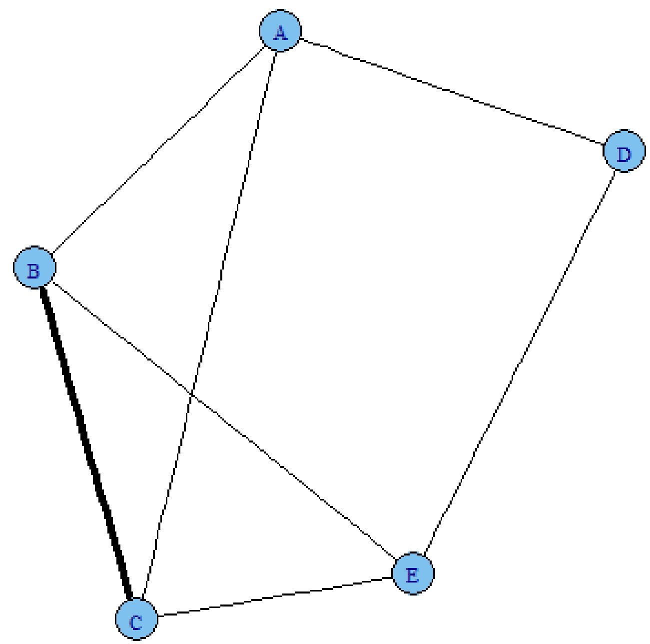
Bipartite Projections



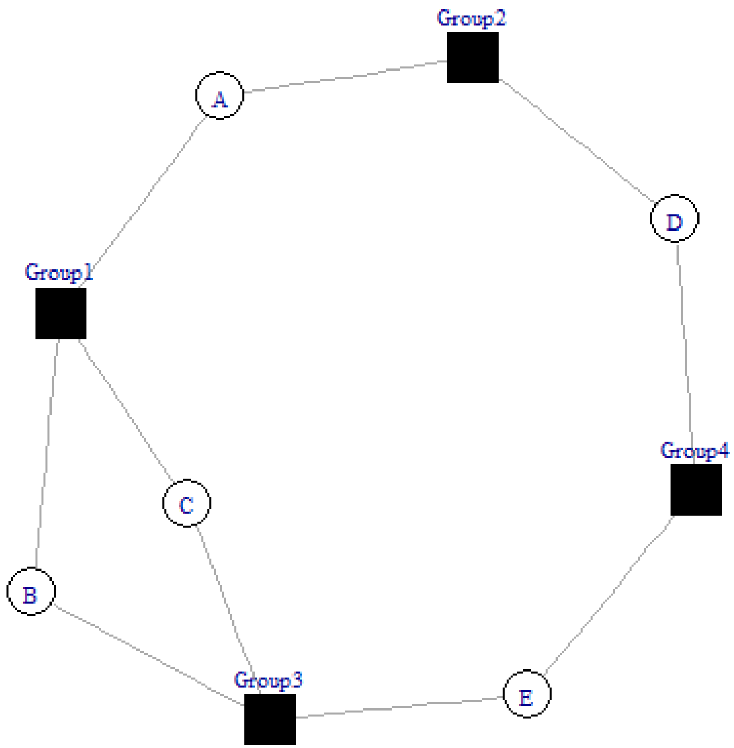
cency matrix of the **first**
dividuals)

```
attr="weight")
```

Co-membership network of nodes



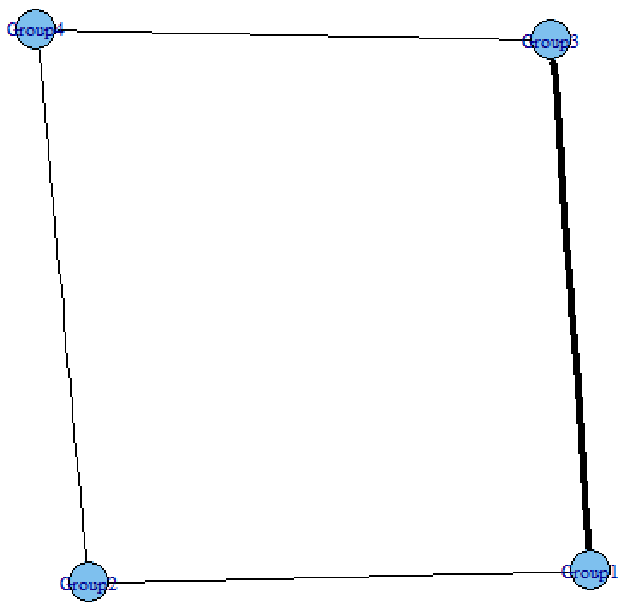
Now, you have an ***affiliation network*** of individuals based on co-membership in groups.



cy matrix of the second
ps)

ht")

Network of groups



Bipartite Projections

- … at this point, it is useful to know that bipartite projections are not magic. In fact, you can project the bipartite network manually by **multiplying the bipartite/affiliation matrix by its transpose (the same matrix flipped along its diagonal), then setting the diagonal to 0:**

```
> aff=bm%*%t(bm)
>  diag(aff)=0
> aff
  A B C D E
A 0 1 1 1 0
B 1 0 2 0 1
C 1 2 0 0 1
D 1 0 0 0 1
E 0 1 1 1 0
```

CONCLUSION

Our goal in preparing this book was to provide a very basic introduction to the core ideas of social network analysis, and how these ideas are implemented in the methodologies that many social network analysts use.

Social network analysis is a continuously and rapidly evolving field and is one branch of the broader study of networks and complex systems. The concepts and techniques of social network analysis are informed by, and inform the evolution of these broader fields. We hope that this tutorial will serve as a starting point for working in social networks.

Recommendation

The Encyclopedia of Social Network Analysis and Mining (ESNAM) is the first major reference work to integrate fundamental concepts and research directions in the areas of social networks and applications to data mining. While ESNAM reflects the state-of-the-art in social network research, the field had its start in the 1930s when fundamental issues in social network research were broadly defined. These communities were limited to relatively small numbers of nodes (actors) and links. More recently the advent of electronic communication, and in particular on-line communities, have created social networks of hitherto unimaginable sizes. People around the world are directly or indirectly connected by popular social networks established using web-based platforms rather than by physical proximity. Reflecting the interdisciplinary nature of this unique field, the essential contributions of diverse disciplines, from computer science, mathematics, and statistics to sociology and behavioral science, are described among the 300 authoritative yet highly readable entries. Students will find a world of information and insight behind the familiar facade of the social networks in which they participate. Researchers and practitioners will benefit from a comprehensive perspective on the methodologies for analysis of constructed networks, and the data mining and machine learning techniques that have proved attractive for sophisticated knowledge discovery in complex applications. Also addressed is the application of social network methodologies to other domains, such as web networks and biological networks.

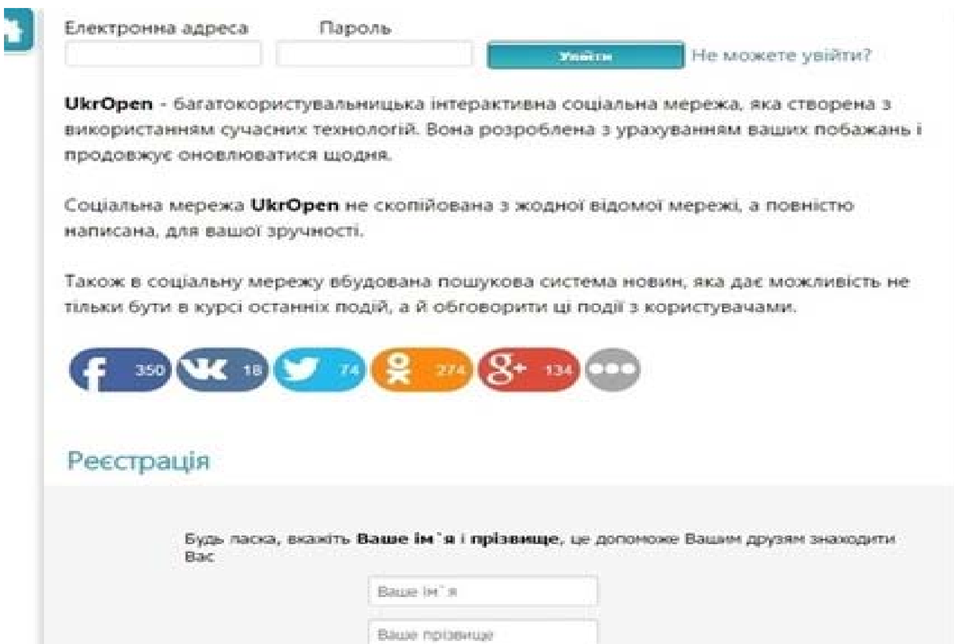
Network Data Collected via Web | Request PDF. Available from: https://www.researchgate.net/publication/272157724_Network_Data_Collected_via_Web [accessed Feb 11 2018].

ДОДАТОК

Найкращі українські соцмережі

Указом Президента України від 15 травня 2017 р. № 133 введено заборону інтернет-провайдерам на надання послуг з доступу користувачам мережі Інтернет до ресурсів сервісів «Mail.ru» (www.mail.ru) і соціально орієнтованих ресурсів «ВКонтакте» і «Одноклассники».

Також санкції введено щодо сервісів «Yandex» «Яндекс Авіаквитки», «auto.ru», «Яндекс аудиторії», «Яндекс Афіша», «Яндекс Гроші», «Яндекс Вебмайстер», «Яндекс відео», «Яндекс Час», «Яндекс директ», «Яндекс диск», «Яндекс карти» та ін.



20 українських соціальних мереж, якими можна замінити заблоковані «ВКонтакте» і «Одноклассники»

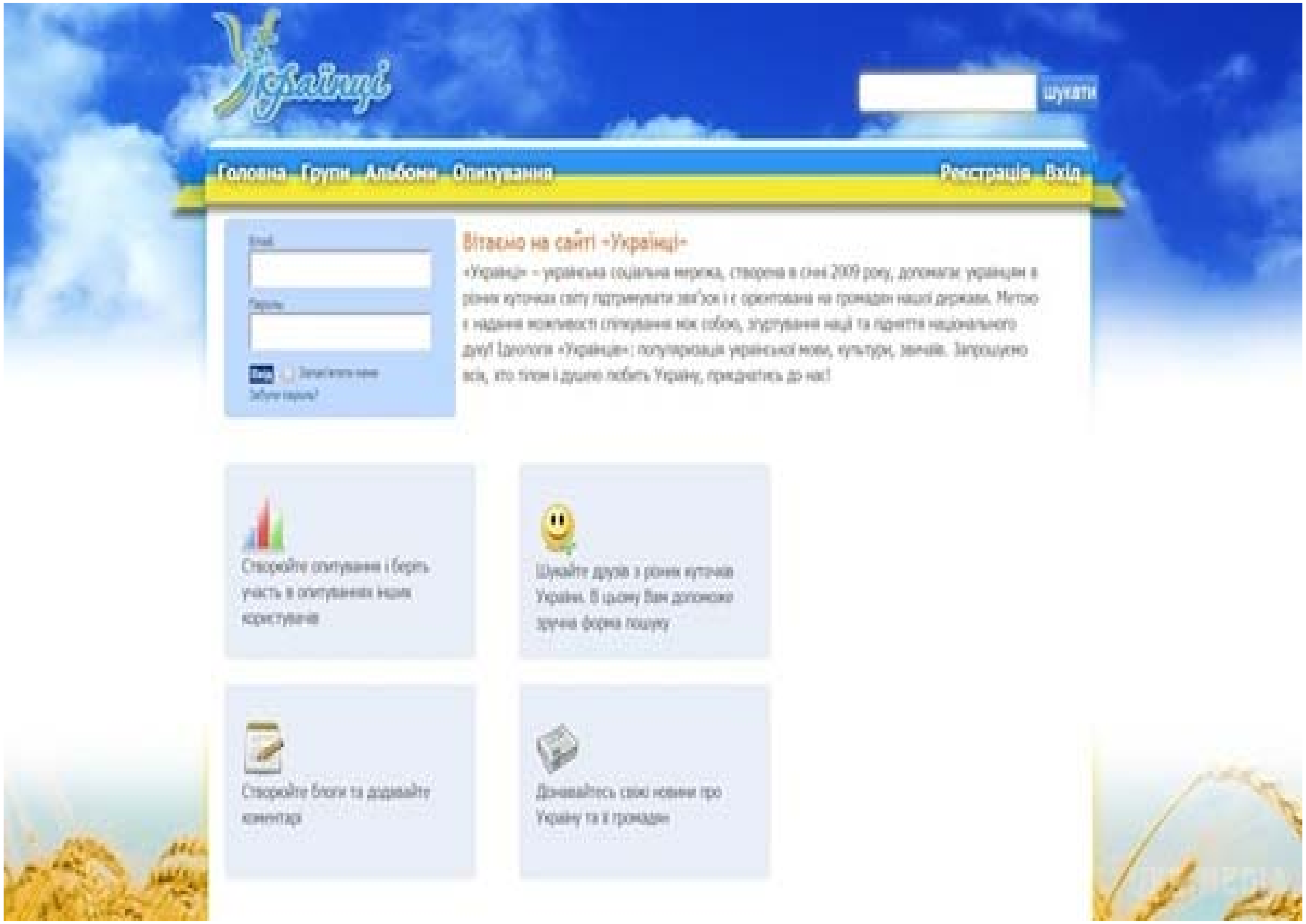
ON 28 ТРАВНЯ 2017.

1. **UkrOpen** – українська соціальна мережа, створена з нуля. У неї вбудована пошукова система, що дозволяє не тільки бути в курсі останніх подій, а й обговорювати їх з іншими користувачами. Реєстрація на ресурсі дозволена користувачам від 13 років.

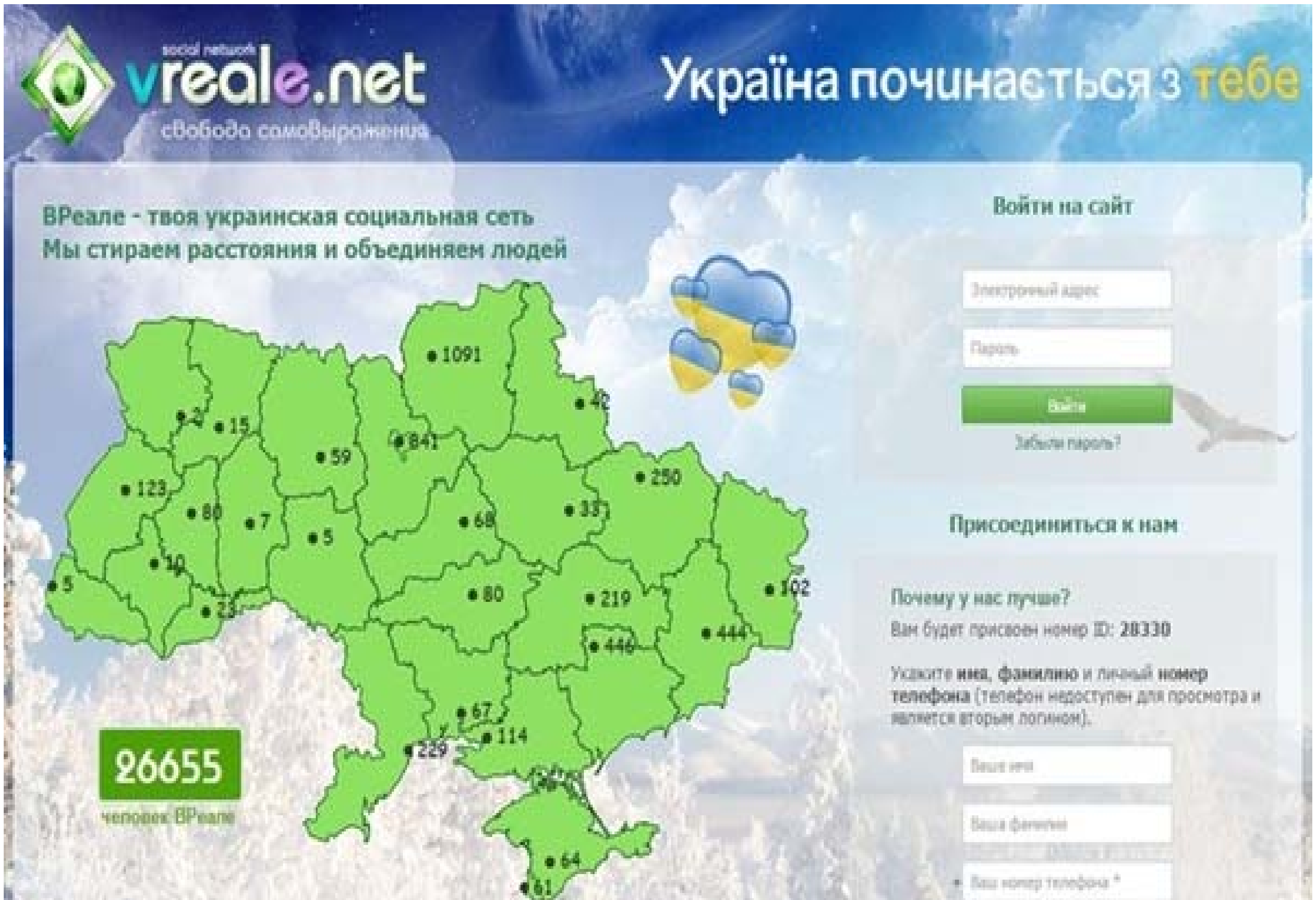
2. Ц. укр – виглядає як поєднання «Фейсбук» і «ВКонтакте». Назва походить від скорочення слів «Це – Україна». Ресурс позиціонує себе як молоду українську соцмережу; створений під керівництвом киянина Михайла Гриненка. Портал постійно проводить конкурси для своїх користувачів. Також сюди можна одним кліком перенести всю інформацію з «ВК».

3. <u>UKRFACE</u> являє собою певний гібрид «ВК» і «ФБ». Там точно не загубишся: є й особиста сторінка, і друзі, і музика, фото, відео та інші звичні пункти меню запису. Головна «фішка» – доступне спілкування кримсько-татарською мовою.

4. <u>Українці</u> – одна з перших вітчизняних соцмереж з патріотичним ухилом. Створена в січні 2009 р., проте користувачів там не так вже й багато, немає фото, відео або музики. Також немає і стрічки новин.

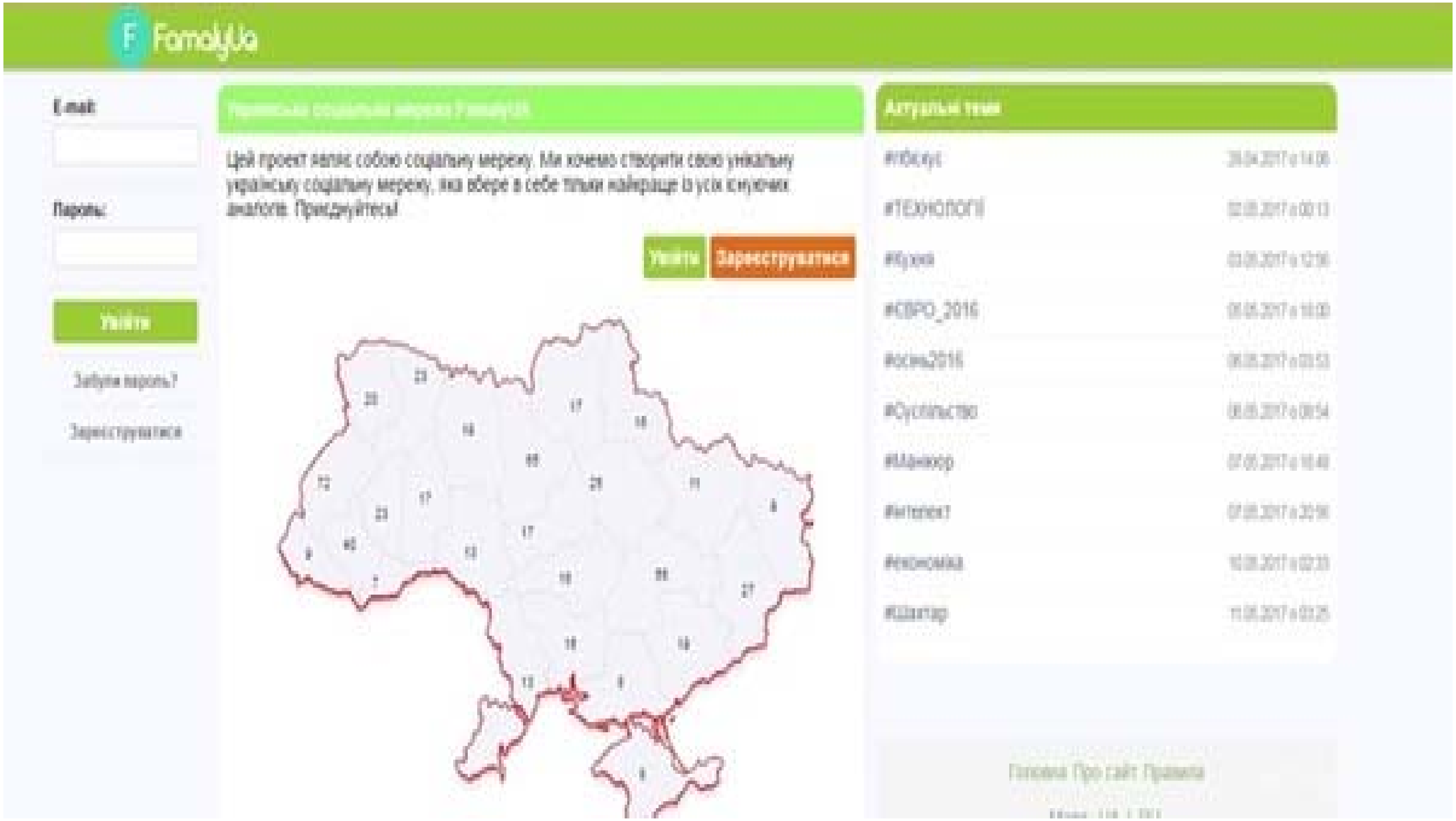


5. <u>Vreale.net</u> позиціонує себе як незалежне українське інтернет-співтовариство, яке надає актуальні та популярні інструменти для знайомства, спілкування, розваг, самоосвіти, проведення конкурсів і віртуальних змагань, організації груп спільних інтересів і захоплень, а також для організації особистих та корпоративних зустрічей у реальному житті.



6. **VsiTut.com** – соціальна мережа України, створена як волонтерський інтернет-проект і не належить до жодної партії чи спільноти. Як заявляють розробники, головною метою проекту є об’єднання українських патріотів незалежно від мови спілкування та місця проживання. Головною особливістю цієї соцмережі є щомісячне оновлення компонентів і модулів сайту, відкриті розділи для гостей сайту.

7. <u>FamalyUA</u> – проект, що являє собою соціальну мережу. Творці запевняють: вони взяли все найкраще від інших аналогів і об'єднали.



8. <u>Yachudo.com</u> – українська соціальна мережа, яка надає унікальні можливості спілкування та персоналізації свого цифрового простору. «***Я – чудо***» українська соцмережа, спрямована на українських і англомовних користувачів, адже російська мова у цій соцмережі відсутня. Дизайн відрізняється від дизайну інших соцмереж, також відсутній музичний розділ.

9. livebook – проект, що являє собою унікальну українську соціальну мережу, яка вбирає в себе тільки найкраще з усіх існуючих аналогів. Наразі перебуває в стадії розробки.

10. 1ua.com.ua – соціальна мережа, що об’єднує населені пункти, сприяє обміну інформацією, активний форум.

11. Haps – українська незалежна соціальна мережа, створена 1 січня 2014 р. Допомагає користувачам взаємодіяти один з одним. Живий чат з повідомленнями, коментарями, оцінками, обміном фотографіями, фотоальбомами, іграми, групами, сторінками організацій, музикою, відео, інформацією про життєві події тощо.

12. Катапульта – комунікаційний онлайн-майданчик з обміну текстовими і графічними даними між користувачами, що призначений виключно для соціальної комунікації користувачів.

13. SocialFace – все в одному. Тут поєднуються всі «фішки» відомих соціальних мереж Facebook, Twitter, «ВКонтакте» та навіть торговий майданчик Slando (згодом). Велика команда молодих перспективних юнаків та дівчат створила цей проект за спільною ідеєю. Усі сервери, на яких працює мережа, знаходяться в Україні і належать українцям. Ніхто не має впливу на соціальну мережу SocialFace. Проект не спрямований на отримання прибутку, мета соціальної мережі – об’єднати Україну в національній мережі.

14. Друзі – це соціальна мережа для швидкої і зручної комунікації між людьми по всьому світу. Завдання мережі «Друзі» – в кожен окремо взятий момент залишатися швидким і естетичним засобом спілкування в мережі. На 100% адаптується до будь-якого мобільного екрана, таблета або смарт-пристрою.

15. <u>Googoodoo.com</u> – соціальна мережа з приємним та інтуїтивно зрозумілим дизайном, призначена для спілкування з друзями і однодумцями. Користувачам пропонується безліч вбудованих модулів, які допоможуть організувати процес навчання, роботи та продажів товарів.

16. Ravlyk Link – це проект, що об’єднує людей з різними інтересами з усього світу.

17. Одногрупники – мережа, що нагадує «Одноклассники».

18. Сусіди – соціальна мережа, яку робили на противагу «ВКонтакте» і «Одноклассникам». Тим не менш, на головній сторінці сайту є можливість поділитися цим сайтом у конкуруючих соцмережах.

19. **hurtom.com** – портал для обміну матеріалами українською мовою. Є форум, де можна обговорювати фільми, книжки та комп'ютерні ігри.

20. uamodna – інформаційно-розважальна веб-платформа, що об'єднує людей з усього світу. Тут можна знайти статті, інтерв'ю, фото, відео, артдослідження, висвітлення соціальних питань, авторські погляди на актуальні події, творчі та наукові доробки сучасників. Інформація для реєстрації мінімальна, окрім цього на платформу можна увійти через *Facebook*.

Найпопулярніші соціальні мережі в країнах СНД і світі

Спілкування в соціальних мережах вже давно стало буденністю для сучасних користувачів мережі Інтернет. На просторах глобальної мережі люди вирішують безліч питань, діляться своїми переживаннями, демонструють моменти радісних подій життя, стежать за подіями, що відбуваються в житті рідних і близьких. Така затребуваність соціальних мереж не залишилася непоміченою розробниками програмних продуктів. Не дивно, адже кожен хоче просунути свій товар, отримати гідний прибуток і загальне визнання.

Наведемо рейтинг найбільш популярних соціальних мереж, що мають велику аудиторію користувачів у світі за 2017 р. (джерело: gs.statcounter.com, наводить статистику за такими сайтами, як: Google+, LinkedIn, Facebook, StumbleUpon, YouTube, Twitter, reddit, Digg, MySpace, NowPublic, iWiW, orkut, Fark, Delicious, ВКонтакте, Hi5, Yahoo! Buzz, Vimeo, Mixx, FriendFeed, Hyves, Bebo, Tuenti, Kaboodle, Одноклассники.

Рейтинг соціальних мереж в Україні

Визначимо тенденцію за двома графіками. Перший графік показує середню величину за останні 12 місяців (рис. 1).

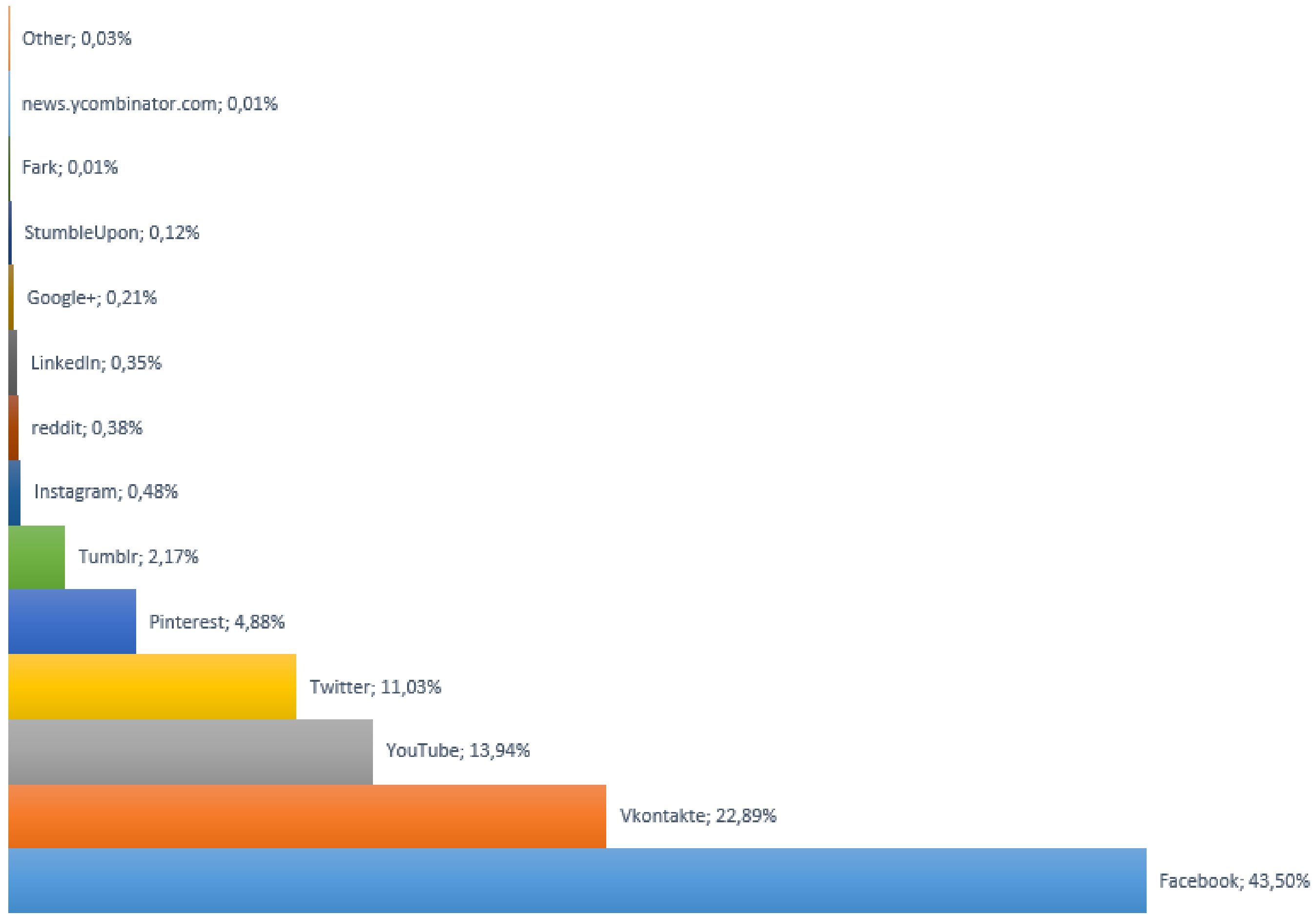


Рис. 1. Україна (середнє за 2016/2017 рр.)

Наступний графік соцмереж відображає популярність станом на червень 2017 р.

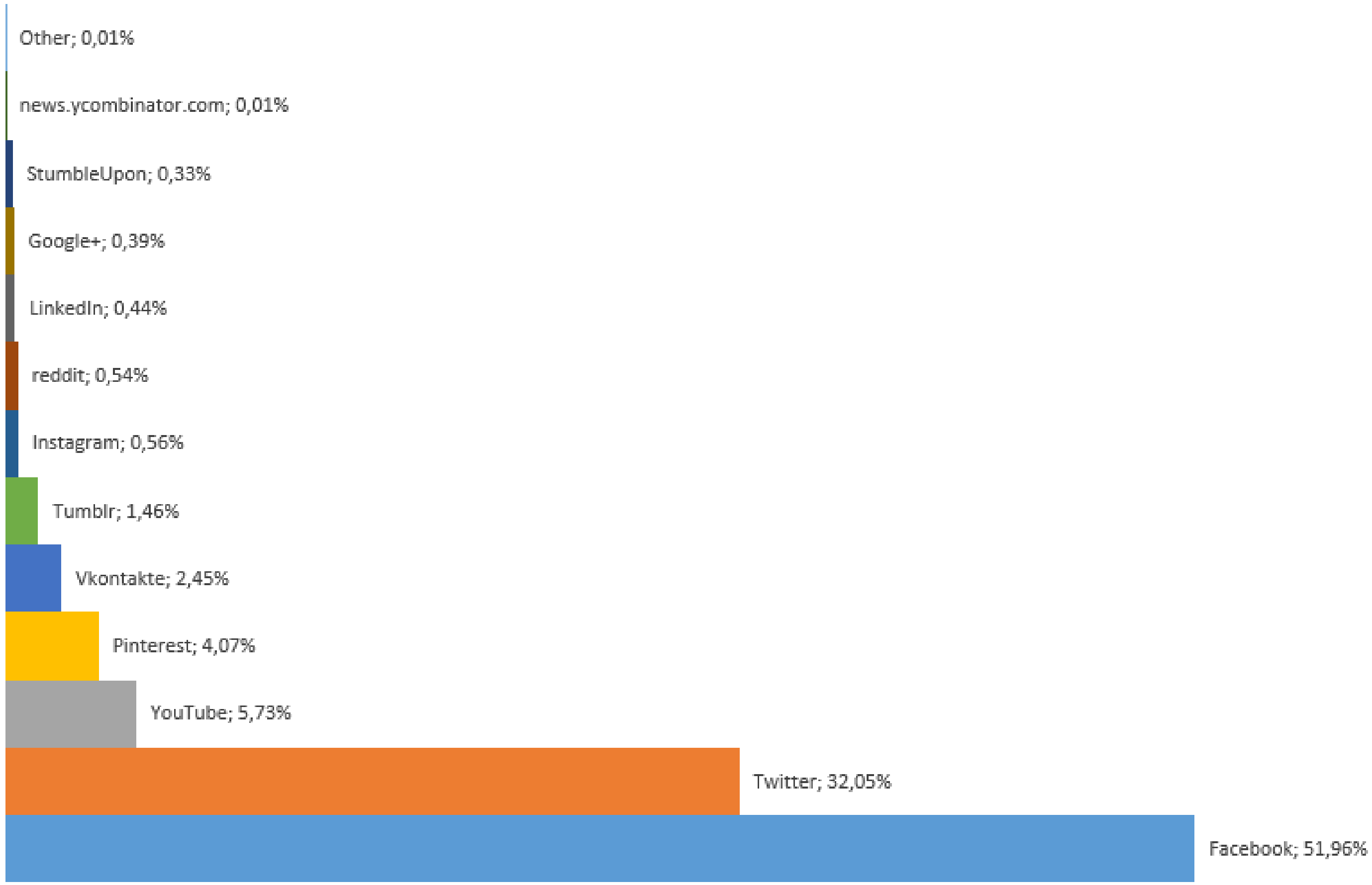


Рис. 2. Україна, червень 2017

Як видно з рис. 2, сильно «просів» показник *ВКонтакте*. Цікаво, що згідно з статистичними даними його місце зайняв *Твіттер*.

В Україні швидко зростає число нових соціальних мереж, серед яких три найвідоміших:

- Ukrainians.co
- Nimses.com
- WEUA.info

ТОП популярних соціальних платформ у Росії

Згідно з рейтингом gs.statcounter.com, в Росії середньорічні показники такі (рис. 3, 4):

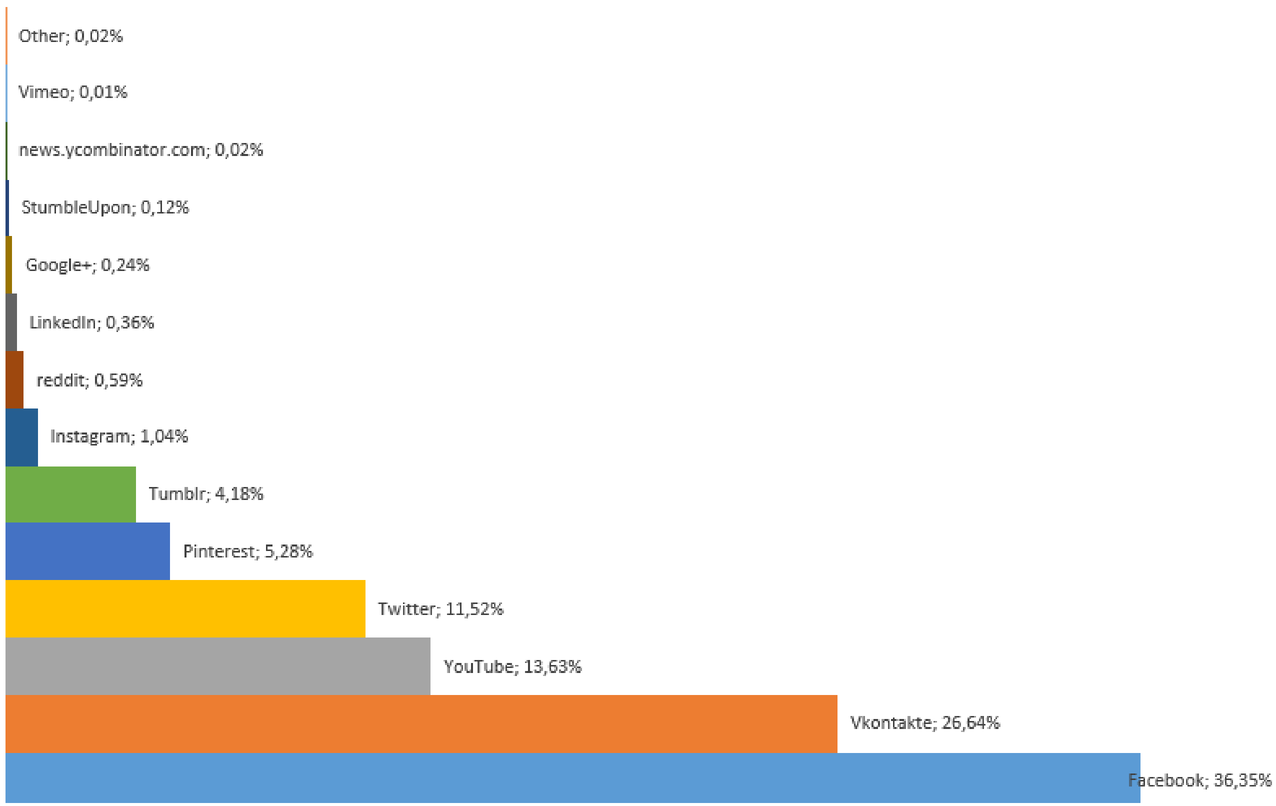


Рис. 3. Росія, середнє за 2016/2017 рр.

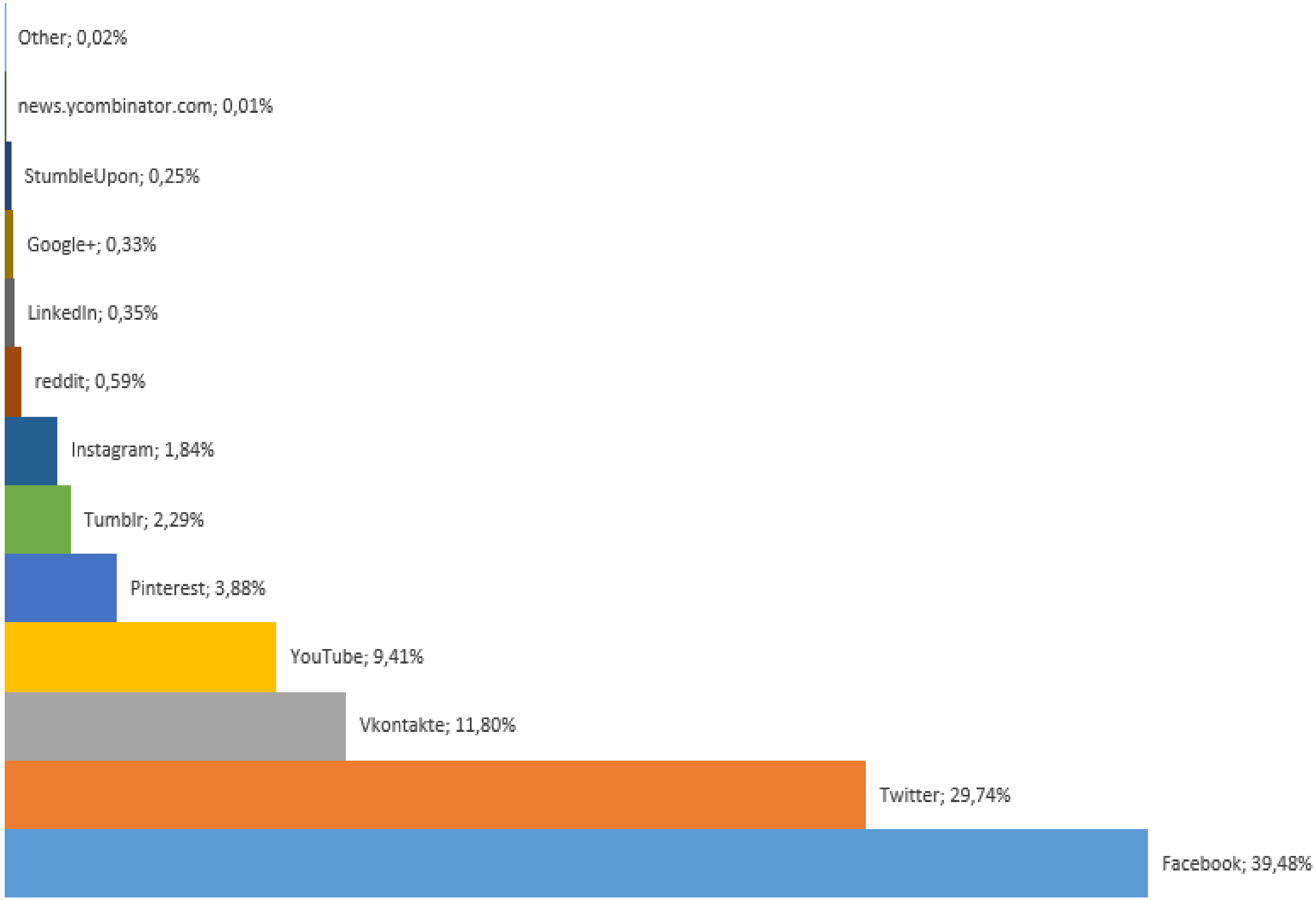


Рис. 4. Росія, червень 2017

Згідно з даними gs.seo-auditor.com.ru, методологія якого відрізняється, рейтинг популярності такий:

Соціальна мережа	**Популярність, %**
ВКонтакте	44,25%
Facebook	26,54%
Одноклассники	11,37%
Twitter	4,13%
YouTube	14,57%
LiveInternet	0,73%
Мой мир@Mail.ru	0,09%
Живой Журнал	0,95%
Blogger	0,49%
БэбиБлог	0,01%
Хабрахабр	0,03%
@дневники	0,07%
Google+	0,17%

Найпопулярніші соцмережі в Білорусі

У Білорусі також два лідери: *ВКонтакте* і *Фейсбук* (рис. 5).

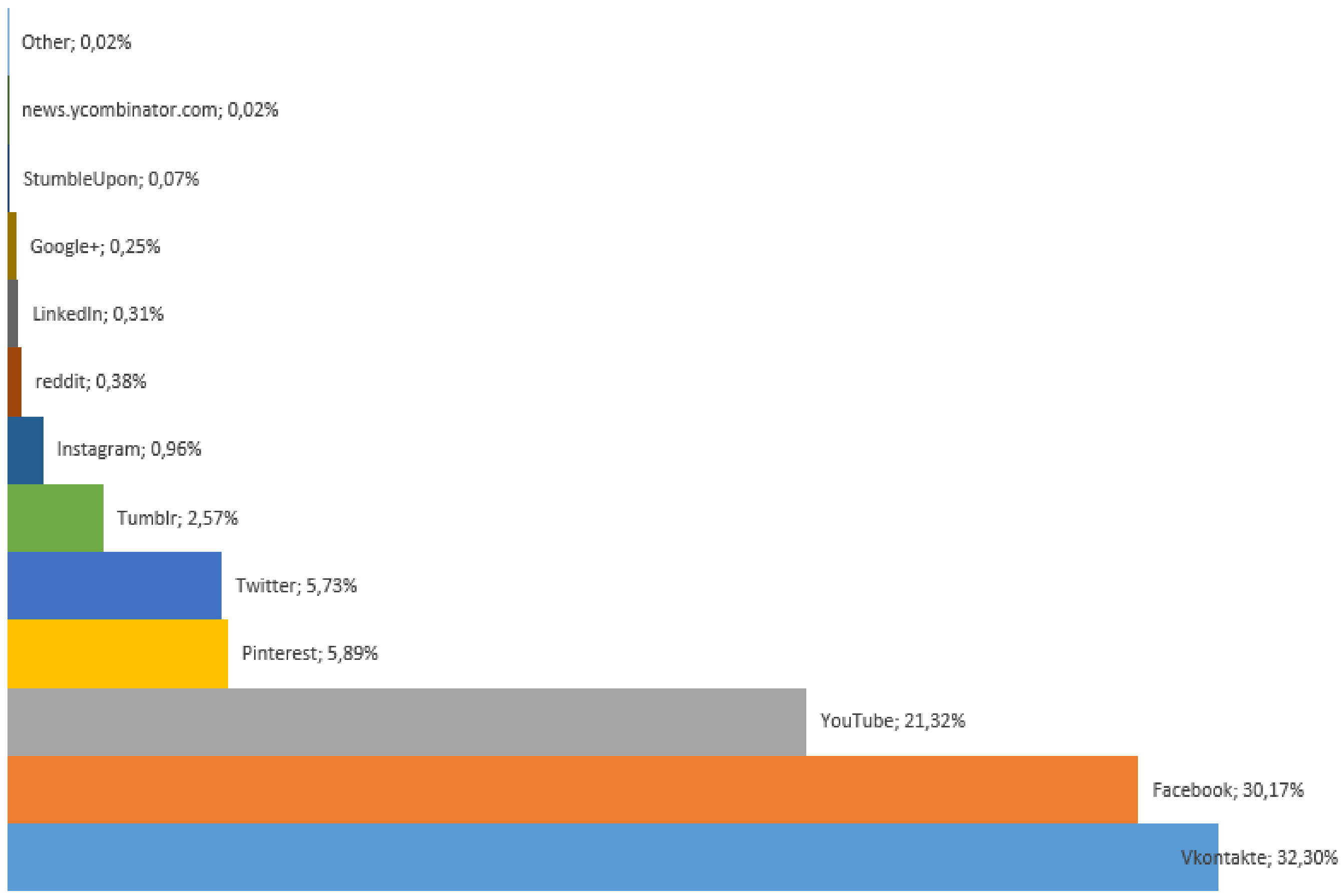


Рис. 5. Білорусь (середнє за 2016–2017 рр.)

Причому за підсумком 2017 р. *ВК* поступився *ФБ* і вони помінялися у рейтингу місцями (рис. 6).

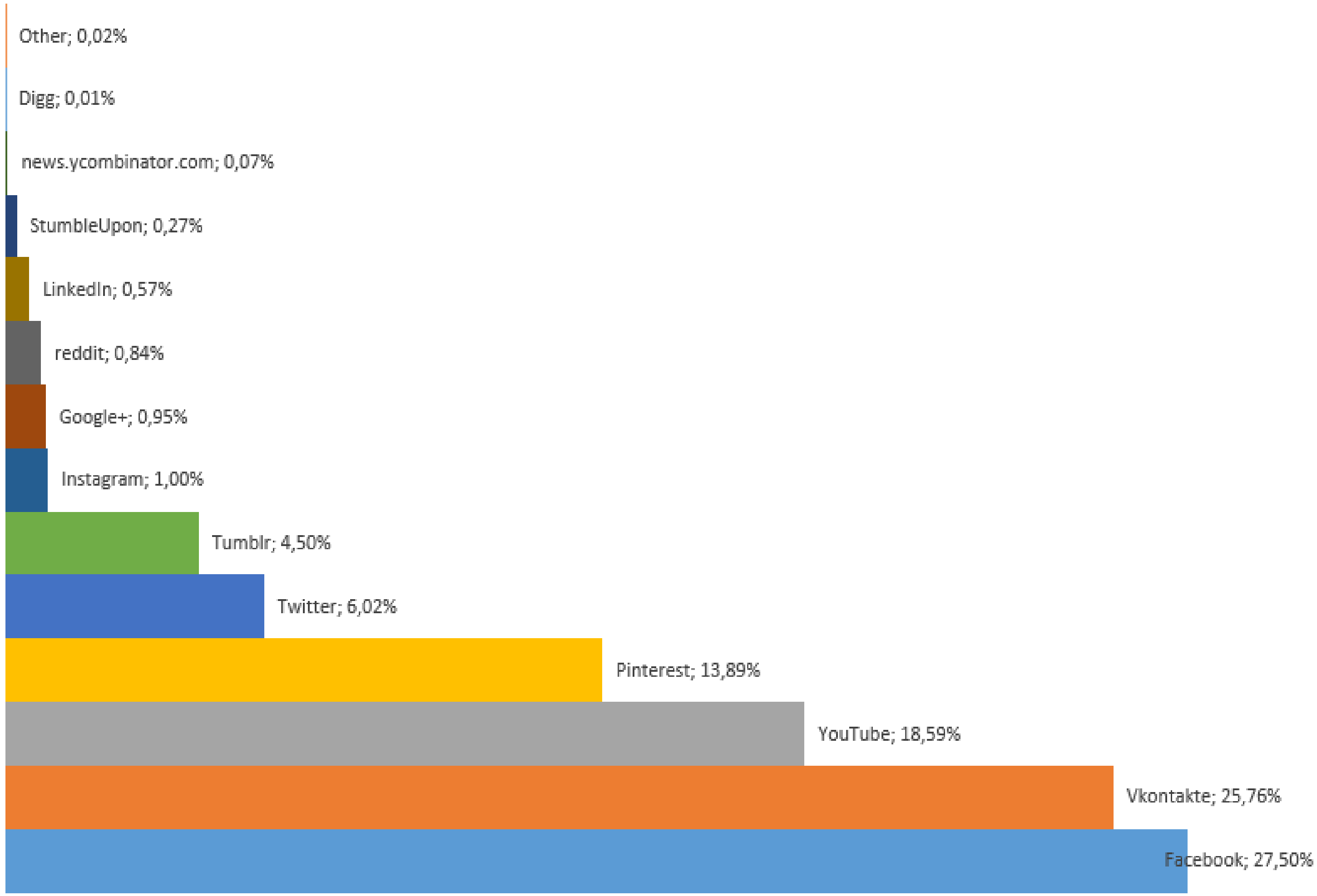


Рис. 6. Білорусь, червень 2017

Рейтинг популярності соціальних мереж у Казахстані

Користувачі з Казахстану віддають перевагу *Фейсбуку (рис. 7)*.

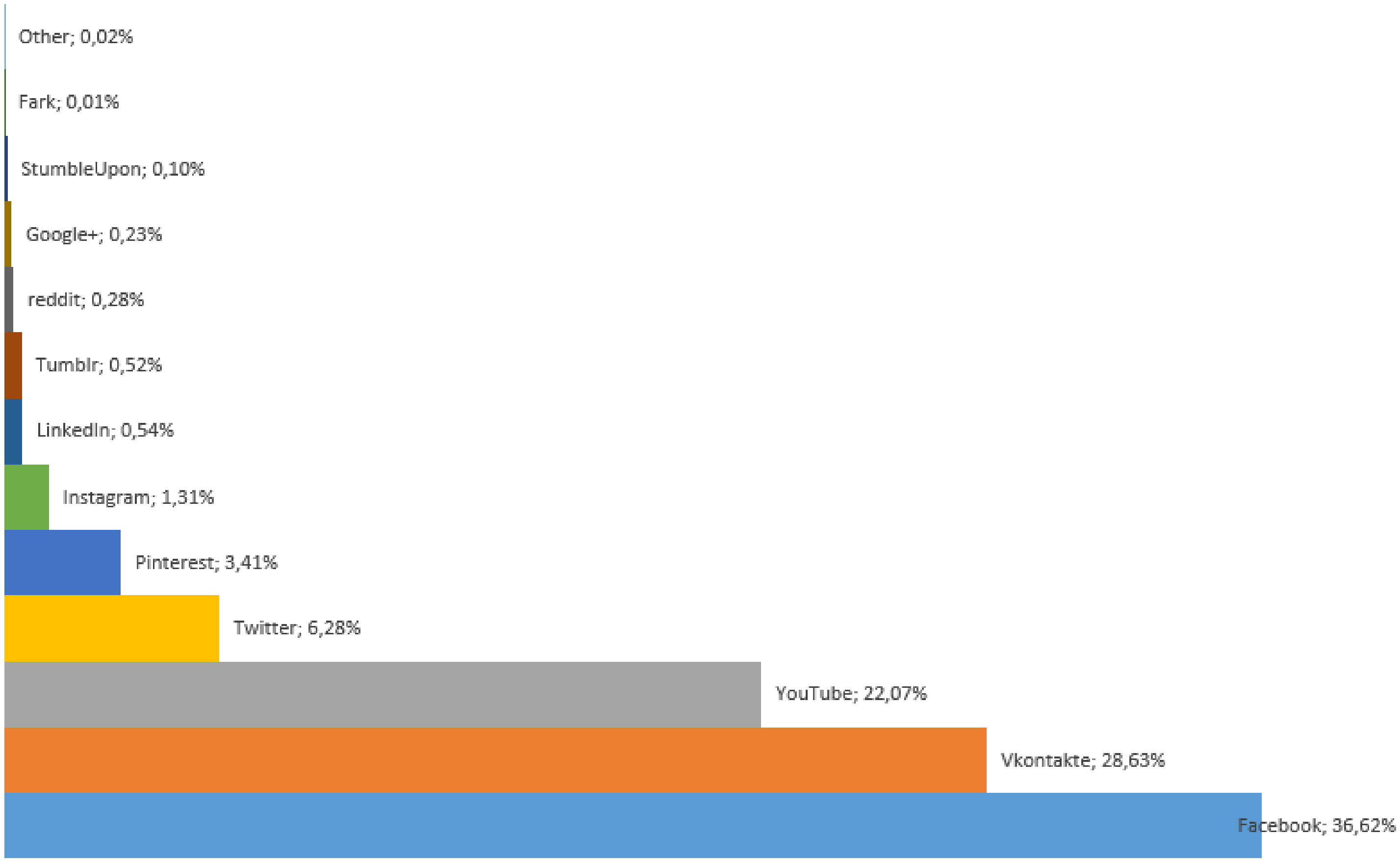


Рис. 7. Казахстан (середнє за 2016–2017 рр.)

Причому мережевий ресурс *ВКонтакте* у цій країні теж трохи втратив позиції за підсумком аналізованого року (рис. 8).

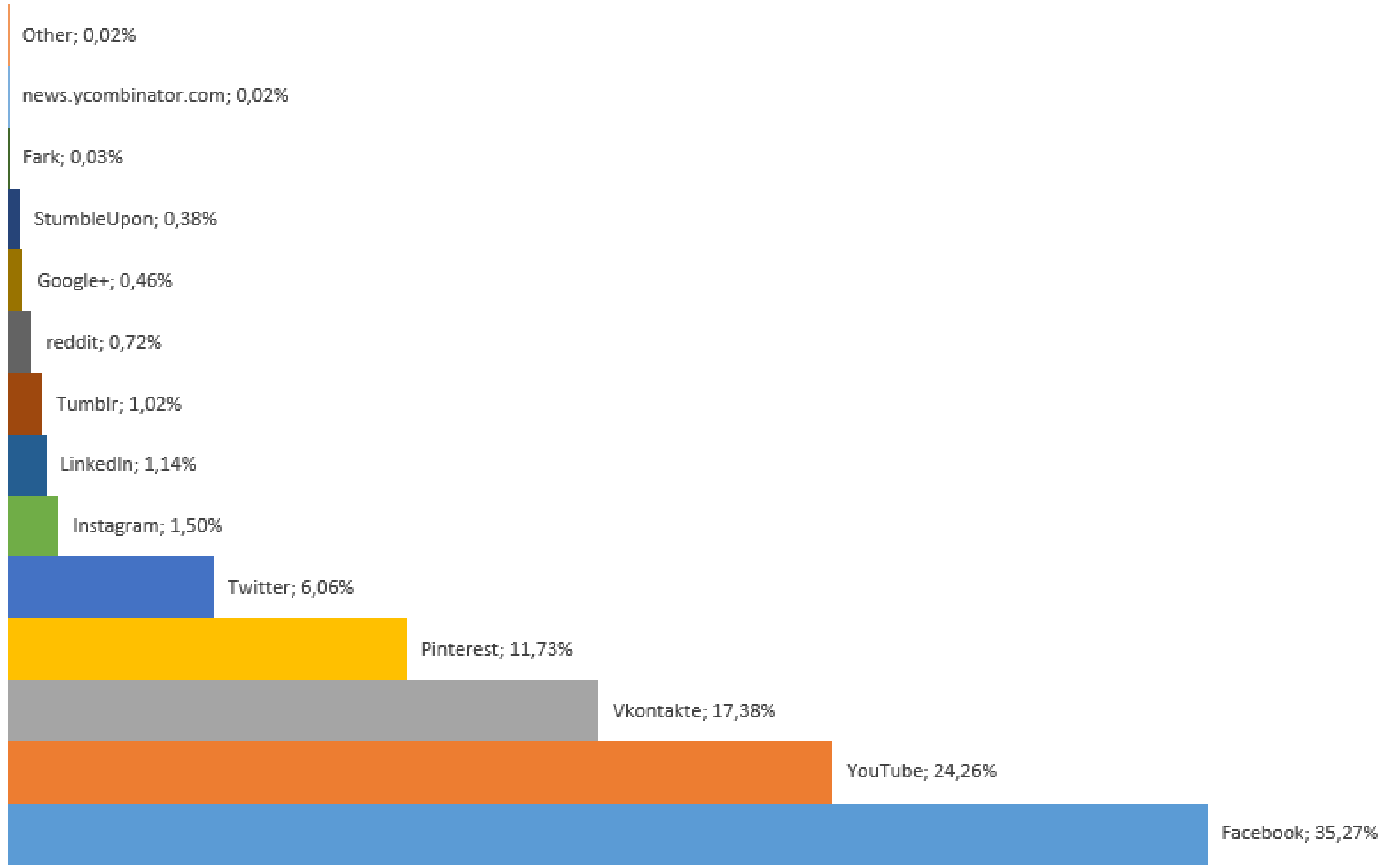


Рис. 8. Казахстан, червень 2017 р.

У казахських інтернет-користувачів стабільно популярний відеоагрегатор *Youtube*.

Тенденції популярності соцмереж у світовому інтернет-просторі

Незмінний лідер світу за статистикою gs.statcounter.com – *Facebook* (рис. 9).

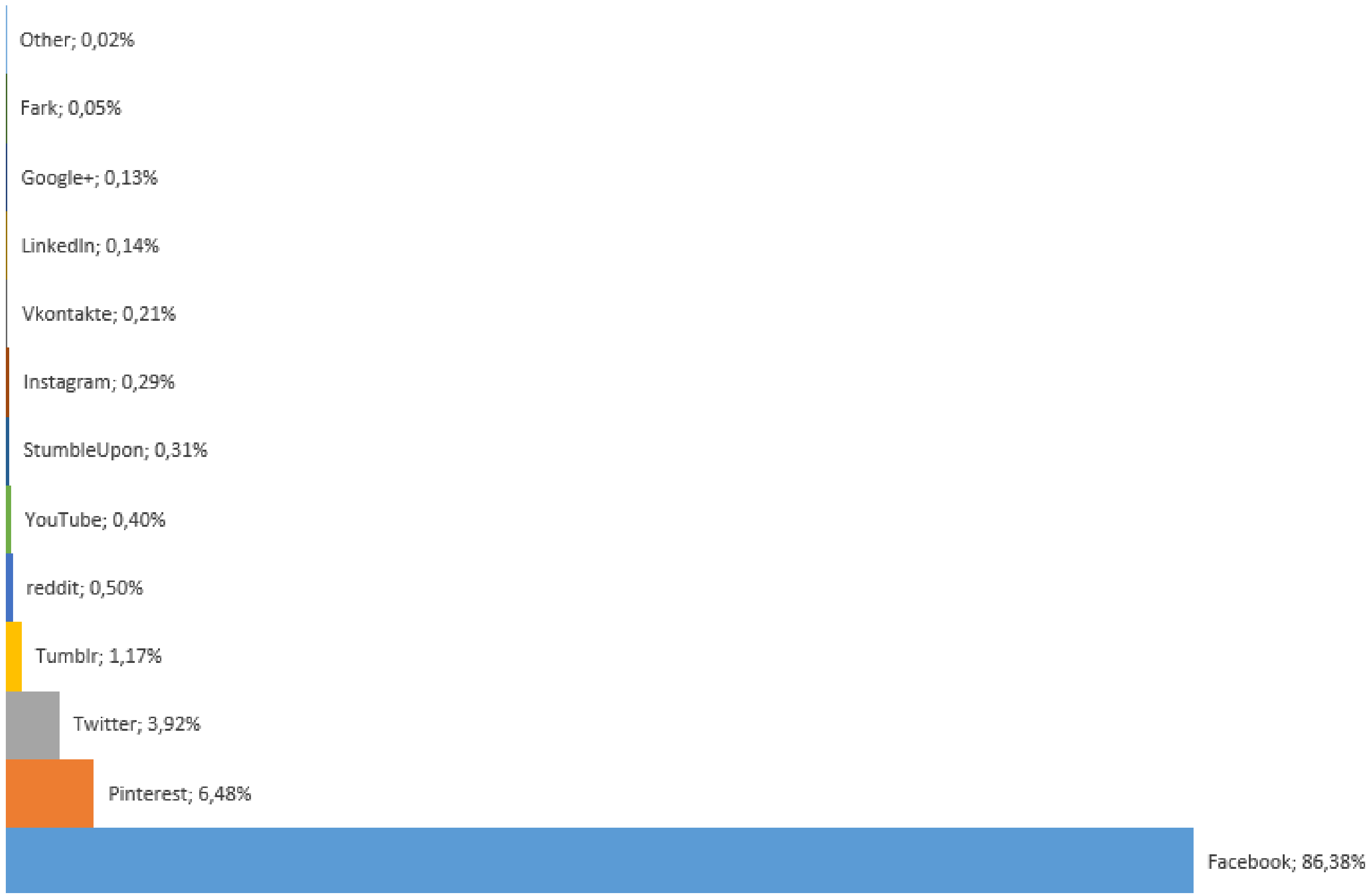


Рис. 9. Світ, середнє за 2016–2017 рр.

Pinterst займає друге місце і активно почав набувати популярності, в першу чергу, через свої нові можливості пошуку за картинками (рис. 10).

Рис. 10. Світ, червень 2017 р.

Twitter, як завжди, займає позитивно стабільні позиції. Хоча його новинна стрічка зараз набагато привабливіша.

Згідно з даними з dreamgrow.com кількість користувачів мереж по всьому світу в місяць становить:

Social network	**Monthly Visitors**
Facebook	2,000,000,000
YouTube	1,000,000,000
Instagram	700,000,000
Twitter	313,000,000
Reddit	250,000,000
Vine (In January 2017, The Vine became the Vine Camera)	200,000,000
Pinterest	150,000,000
Ask.fm	160,000,000
Tumblr	115,000,000
Flickr	112,000,000
Google+	111,000,000
LinkedIn	106,000,000
VK	90,000,000
ClassMates	57,000,000
Meetup	30,300,000

Отже, лідери мають великий відрив.

Найвідоміші соціальні мережі

Facebook

(Майже 2 млрд відвідувачів в місяць).

Соціальна мережа, заснована молодим бізнесменом ***Марком Цукербергом***, впевнено лідирує серед аналогічних сервісів. Щомісяця близько двох мільярдів чоловік з різних куточків планети користуються цим ресурсом з різною метою. Така затребуваність у населення планети зобов’язує Фейсбук підлаштовуватися під кожного користувача. Тому на сьогодні існує більше 60 мовних версій веб-ресурсу, щоб люди з різних держав мали можливість спілкуватися між собою. Незважаючи на жодні перипетії і фінансові кризи, *Facebook* успішно розвивається, залучаючи до себе все більшу кількість нових членів користувачів. А засновник інтернет-мережі, користуючись зростаючою популярністю свого дітища, не втомлюється жертвувати отримані грошові кошти на благодійність, тим самим ще більше завойовуючи любов інтернет-користувачів.

Twitter

(Більше 300 млн відвідувань на місяць).

Інтерес до *Твіттер* пояснюється простотою використання цього сервісу, а також швидкістю відображення повідомлень. Немає нічого простішого, ніж зареєструватися на сторінці соцмережі і відразу ж потрапити у вир подій, дізнаватися останні новини, ділитися своєю думкою з іншими людьми тощо. *Твіттер* був заснований на два роки пізніше, ніж *Фейсбук*, але вже впевнено наздоганяє визнаного лідера. Ця соціальна мережа популярна в багатьох країнах світу.

Linkedin

(Понад 100 млн користувачів в місяць).

Ця веб-система особливо затребувана серед бізнесменів. Така любов підприємців з усього світу пояснюється досить просто: використовуючи цей інтернет-ресурс, можна успішно просувати свій бізнес, заводити корисні знайомства, шукати співробітників у різних куточках земної кулі, а також отримувати замовлення від компаньйонів, які проживають далеко від них самих. Тут можна без великих зусиль розмістити своє резюме або оголошення про вакансії, знайти компанії, профіль роботи яких збігається з вашими інтересами, і багато іншого.

Pinterest

(Приблизно 150 000 000 відвідувань щомісяця).

У рейтингу світових лідерів *Pinterest* займає гідну позицію. Хоча за кількістю користувачів *Pinterest* на 7-й позиції, він швидко набирає популярності. Жителі України, Росії, Білорусі, Казахстану та інших країн колишнього СРСР незабаром можуть включити цю соціальну мережу в список лідерів, оскільки вона має достатній потенціал.

Instagram

(700 000 000 користувачів щомісяця відвідують *Інстаграм*).

Ресурс *Instagram* набув популярності завдяки продуманому менеджменту власників *Фейсбук*. Саме вони викупили права на цю соціальну мережу і вивели її в лідери.

Крім зазначених вище, наводимо й інші соціальні сервіси, які включені до деяких переліків найпопулярніших соціальних мереж. Йдеться про веб-ресурсі ***Youtube*** (1 млрд зареєстрованих користувачів), який займає друге місце. Крім того, варто згадати про ВКонтакте і Одноклассники, одні з найбільш улюблених платформ для спілкування, якими користуються жителі країн СНД. Тепер Україна не входить до списку цих країн, оскільки на території держави вони блокуються Указом Президента України.

Також варто згадати про ***Viber***, ***WhatsApps***, ***Telegram*** та інші соціальні ресурси, які присутні в різноманітних рейтингах.

Великий вибір всіляких соцмереж змушує їх власників постійно розвиватися, а користувачі тільки виграють від такого суперництва. Головне, щоб конкуренція була чесною і здоровою, з використанням тільки відкритих (і законних) методів. А споживачі будуть самі вирішувати, якій саме соціальній мережі віддавати перевагу.

Для того, щоб соціальні мережі зайняли своє місце в науковій, технічній та соціальній сферах, вони мають прямувати від нинішнього початкового стану до зрілості і стати спільним інструментом суспільного життя та спілкування інформаційного суспільства. Прогнозований шлях розвитку – поділ сфер впливу на мережу контактів, зв'язку та професійних мереж.

GLOSSARY

Ameoba – модель, яка показує, як природним шляхом відбувається ріст і розмноження амеб.

Cafun – вільно доступний пакет на основі технологій штучного життя.

Classmates.com – соціальна мережа, створена в 1995 р. Ренді Конрадсом, який заснував Classmates Online, Inc.

Connect.ua – українська соціальна мережа, створена в грудні 2007 р.

E-cexutive.ru – мережа менеджерів.

EVS-моделі – ансамблі зі змінною структурою.

Facebook – веб-сайт популярної соціальної мережі, що почав працювати 4 лютого 2004 р.

Hi5 – соціальний мережевий веб-вузол, який з 2007 р. був одним з 25 найбільш відвідуваних сайтів.

Insna.org – міжнародна мережа аналізу соціальних мереж.

It-n.ru – мережа творчих вчителів.

Last.fm – Інтернет-проект музичної тематики, основним сервісом якого є збір інформації про музику, що слухає користувач, і її каталогізація в індивідуальних і загальних чартах.

LinkedIn – соціальна мережа для пошуку і встановлення ділових контактів.

LovePlanet – мережа сайтів знайомств, що почала свою роботу в 2005 р.

MySpace – популярний соціально-мережевий веб-вузол, що пропонує орієнтовану на користувача мережу друзів, особистих профілів, блогів, груп, фотографій, музики і відео, став впливовою частиною сучасної популярної культури, особливо в країнах, де говорять англійською.

Orkut – соціальна мережа, створена Google, названа на честь творця, працівника Google.

Twitter – соціальна мережа, яка являє собою мережу мікроблогів, що дозволяє користувачам надсилати короткі текстові повідомлення, використовуючи SMS, служби миттєвих повідомлень і сторонні програми-клієнти.

Webby.ru – проект соціальної мережі професійних знайомств.

XML – мова розмітки, що фактично є зведенням загальних синтаксичних правил.

Актор – діючий суб’єкт; індивід, що чинить дію, яка спрямована на інших.

Аналіз соціальних мереж – напрям структурного підходу, основними цілями якого є дослідження взаємодій між соціальними об'єктами і виявлення умов виникнення цих взаємодій.

Афіліація – прагнення бути в товаристві інших людей, потреба людини у створенні теплих, емоційно значущих відносин з іншими людьми.

Блог – веб-сайт, головний зміст якого – тимчасові записи, зображення чи мультимедіа, що регулярно додаються.

Блогосфера – сукупність усіх блогів спільноти чи соціальної мережі.

Відкриті системи – такі системи, які підтримуються в певному стані за рахунок безперервного припливу ззовні речовин, енергії, інформації.

Відсутні зв'язки – такі зв'язки, які характеризуються браком емоційної складової, часу, довіри і взаємності.

Гіпертекст – текст для перегляду на комп'ютері, який містить зв'язки з іншими документами («гіперзв'язки» чи «гіперпосилання»).

Граф – сукупність непорожньої безлічі вершин і безлічі пар вершин.

Джеймс Барнс – соціолог, перший ввів поняття соціальної мережі.

Дисипативна система – відкрита нелінійна система, яка є далекою від стану термодинамічної рівноваги.

Доменне ім'я – унікальне алфавітно-цифрове ім'я, що ідентифікує конкретний вузол Інтернету.

Домінування – відхилення від рівномірного розподілу зв'язків між «центрами» і «кліками».

Електронна пошта – популярний сервіс в Інтернеті, що робить можливим обмін даними будь-якого змісту (текстові документи, аудіо, відеофайли, архіви, програми).

Ентропія – в термодинаміці міра енергії у термодинамічній системі, яка не може бути використана для виконання роботи. Вона також є мірою хаосу, присутнього в системі.

Емергентність – ступінь незвідності властивостей системи до властивостей окремих елементів.

«ЖЖ» – блог-платформа для ведення онлайн-щоденників (блогів), а також окремий персональний блог, розміщений на цій платформі.

Інформатика – наука про властивості, закони, процеси, методи та засоби формування, утворення та поширення інформації в природі і суспільстві, в тому числі за допомогою технічних систем.

Клік – підвищена мережева щільність.

Контроль – міст між великою кількістю вузлів, який контролює потоки інформації.

Маркетинг – управління створенням товарів і послуг та механізмами їх реалізації як єдиним комплексним процесом.

Математична модель – система математичних співвідношень, які описують досліджуваний процес або явище.

Модель – спрощене подання явищ або об'єктів дійсності, що належать до природи і суспільства, у вигляді схем, рисунків, описів, математичних формул, будь-якого реального предмета (явища або процесу), досліджуваного як їх аналог.

Пошукова система – онлайн-служба, що надає можливість пошуку інформації.

Предмет дослідження інформатики – властивості, закономірності, процеси, методи й засоби формалізації інформації (даних і знань), її подання, кількісної оцінки, зберігання, перетворення і поширення в природі і суспільстві, а також закони створення та використання для цих цілей відповідних систем.

Ранг вузла – найбільша кількість взаємопов'язаних вузлів.

Ранг міжособистісної взаємодії – поняття, яке дозволяє класифікувати відносини.

Сайт – сукупність електронних документів (файлів) приватної особи або організації у комп'ютерній мережі, об'єднана під однією адресою.

Синергетика – міждисциплінарна наука, що займається вивченням процесів самоорганізації і виникнення, підтримки стійкості і розпаду структур (систем) різної природи на основі методів математичної фізики.

Система масового обслуговування – система, яка виконує обслуговування вимог, що надходять до неї.

Системологія – науково-практична сфера діяльності, пов'язана з вивченням, опрацюванням знань та їх застосуванням у вивченні, проектуванні, створенні, управлінні системами, явищами системогенезу.

Слабкі зв'язки – це зв'язки з людьми, яких ми майже не знаємо, з якими не ділимося своїми переживаннями і не підтримуємо відносини.

Служба соціальних мереж – веб-сайт або інша служба у Web, яка дозволяє користувачам створювати публічну або напівпублічну анкету, складати список користувачів, з якими вони мають зв'язок, та переглядати власний список зв'язків і списки інших користувачів.

Соціабельність – товариськість, контактність.

Соціальна група – об'єднання людей, що мають загальну значущу соціальну ознаку, базовану на їхній участі в деякій діяльності, пов'язану системою відносин, які регулюються формальними або неформальними соціальними інститутами.

Соціальна мережа – соціальна структура, утворена індивідами або організаціями, відображає розмаїті зв'язки між ними через різноманітні соціальні взаємовідносини, починаючи з випадкових знайомств і закінчуючи тісними родинними узами.

Соціальна психологія – розділ, галузь психології, яка займається вивченням закономірностей діяльності людини в умовах взаємодії в соціальних групах.

Соціологія – наука про умови, хід співжиття людей та спостереження за ними.

Соціотехнічний об'єкт – об'єкт, в якому окремі учасники мережі – агенти за допомогою сукупності фізичної структури – комп'ютерів учасників мережі – серверів каналів зв'язку і концентраторів реалізують свої потреби в спілкуванні, встановленні контактів, пошуку інформації, роботи, вирішують життєві завдання і замислюються над загальнолюдськими проблемами.

Теорія імовірностей – розділ математики, що вивчає закономірності випадкових явищ: випадкові події, випадкові величини, їх функції, властивості і операції над ними.

Теорія перколяції – теорія, що описує виникнення нескінченних зв'язних структур (кластерів), що складаються з окремих елементів.

Уніформізм – історична геологічна гіпотеза, згідно з якою в геологічному минулому діяли ті ж сили і з тією ж інтенсивністю, що і в сучасну епоху, тому знання сучасних геологічних явищ можна без поправок поширювати на тлумачення геологічного минулого будь-якої давності.

Хабрахабр – багатофункціональний сайт, що являє собою поєднання соціальної мережі і колективного блогу, створений для публікації новин, аналітичних статей, думок, пов'язаних з високими технологіями, бізнесом та Інтернетом.

BIBLIOGRAPHY

1. Knoke, David and Song Yang (2008). Social Network Analysis (2nd Ed). Sage, 2008. – 132 p.

2. Song Yang, Franziska B. Keller, Lu Zheng (2016). Social Network Analysis: Methods and Examples. SAGE Publications, Incorporated, 2016. – 248 p.

3. Hanneman, Robert A. and Mark Riddle. Introduction to social network methods. – Riverside, CA: University of California, Riverside, 2005. – 322 p. (published in digital form at http://faculty.ucr.edu/~hanneman/)

4. Huisman M., Duijn M. Software for Social Network // Analysis Proceedings of the Sixth International Conf. on Logic and Methodology, August 17–20. – Amsterdam, The Netherlands, 2004. – Pp. 578–600.

6. Scott, John. Social Network Analysis (Second Edition). – London: Sage, 2000. – 220 p.

7. Ритцер Джордж. Современные социологические теории. 5-е изд. / Джордж Ритцер. – СПб.: Питер, 2002. – 688 с.

8. Губанов Д.А. Социальные сети: модели информационного влияния, управления и противоборства / Д.А. Губанов, Д.А. Новиков, А.Г. Чхартишвили. – М.: Изд-во физико-математической литературы, 2010. – 228 с.

9. Shantanu Ghosh. Top seven social media threats [Electronic resourse] / Shantanu Ghosh // Electronic data. – [Computerweekly, 2017]. Mode of access: World Wide Web: http://www.computerweekly.com/tip/Top-seven-social-media-threats (Accessed 07 March 2017).

10. Charu C. Social network data analytics / C. Charu // Springer Science & Business Media, 2012. – 486 p.

11. Огляд моделей аналізу соціальних мереж [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/279535422_Oglad_modelej_analizu _socialnih_merez (останній візит 11 лютого 2018).

12. http://uk.wikipedia.org

13. *Software*: The AutoMap and ORA software are freely available from www.casos.cs.cmu.edu.

14. *Software*: www.insna.org/software/index.html.

15. The Encyclopedia of Social Network Analysis and Mining (ESNAM)/ Network Data Collected via Web | Request PDF. Available from: https://www.researchgate.net/publication/272157724_Network_Data_Collected_via_Web (accessed Feb 11 2018).

Навчальне видання

Косарєв Вячеслав Михайлович
Різун Ніна Олегівна

АНАЛІЗ СОЦІАЛЬНИХ МЕРЕЖ

Навчальний посібник

(англійською та українською мовами)

Електронне видання

Редактор Л.В. Пилипчак
Комп'ютерна верстка Г.М. Хомич

ВНЗ «Університет імені Альфреда Нобеля».
49000, м. Дніпро, вул. Січеславська Набережна, 18.
Тел. (056) 778-58-66, e-mail: rio@duan.edu.ua
Свідоцтво ДК № 5309 від 20.03.2017 р.