Social Network Analysis

SICSS 2019 | Fundamentals | Dr. Oliver Posegga

What is network analysis?

What is network ~~analysis~~ theory?

- Definition by Borgatti et al. (2014):
  - Perhaps the most fundamental characteristic of network theory (...) is the focus on relationships among actors as an explanation of actor and network outcomes.
  - This is in contrast to traditional dispositional or individualist explanations that focus on attributes of actors that are treated as independent cases or replications.
- Example:
  - Rather than trying to model adoption of innovation solely in terms of characteristics of the adopter (e.g., age and personality type), network theorists posit interpersonal processes in which one person imitates, is influenced by, or is given an opportunity by another.

Network Theory vs. Theory of Networks

- Network Theory
  - Network theory refers to the mechanisms and processes that interact with network structures to yield certain outcomes for individuals and groups.
  - In the terminology of Brass (2002), network theory is about the consequences of network variables, such as having many ties or being centrally located.
- Theory of Networks
  - In contrast, theory of networks refers to the processes that determine why networks have the structures they do—the antecedents of network properties, in Brass's terms.
  - This includes models of who forms what kind of tie with whom, who becomes central, and what characteristics (e.g., centralization or small-worldness) the network as a whole will have.

Great introduction to the fundamental perspectives on networks, network analysis, and network theory

- Borgatti SP, Mehra A, Brass DJ, & Labianca G (2009) Network Analysis in the Social Sciences. *Science* 323 (5916)

REVIEW

CORRECTED 24 APRIL 2009; SEE LAST PAGE

Network Analysis in the Social Sciences

Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, Giuseppe Labianca

Over the past decade, there has been an explosion of interest in network research across the physical and social sciences. For social scientists, the theory of networks has been a gold mine, yielding explanations for social phenomena in a wide variety of disciplines from psychology to economics. Here, we review the kinds of things that social scientists have tried to explain using social network analysis and provide a nutshell description of the basic assumptions, goals, and explanatory mechanisms prevalent in the field. We hope to contribute to a dialogue among researchers from across the physical and social sciences who share a common interest in understanding the antecedents and consequences of network phenomena.

One of the most potent ideas in the social sciences is the notion that individuals are embedded in thick webs of social relations and interactions. Social network theory provides an answer to a question that has preoccupied social philosophy since the time of Plato, namely, the problem of social order: how autonomous individuals can combine to create enduring, functioning societies. Network theory also provides explanations for a myriad of social phenomena, from individual creativity to corporate profitability. Network research is "hot" today, with the number of articles in the Web of Science on the topic of "social networks" nearly tripling in the past decade. Readers of *Science* are already familiar with network research in physics and biology (*1*), but may be less familiar with what has been done in the social sciences (*2*).

History

In the fall of 1932, there was an epidemic of runaways at the Hudson School for Girls in upstate New York. In a period of just 2 weeks, 14 girls had run away— a rate 30 times higher than the norm. Jacob Moreno, a psychiatrist, suggested the reason for the spate of runaways had less to do with individual factors pertaining to the girls' personalities and motivations than with the positions of the runaways in an underlying social network (*3*). Moreno and his collaborator, Helen Jennings, had mapped the social network at Hudson using "sociometry," a technique for eliciting and graphically representing individuals' subjective feelings toward one another (Fig. 1). The links in this social network, Moreno argued, provided channels for the flow of social influence and ideas among the girls. In a way that even the girls themselves may not have been conscious of, it was their location in the social network that determined whether and when they ran away.

Moreno envisioned sociometry as a kind of physics, complete with its own "social atoms" and its laws of "social gravitation" (*3*). The idea of modeling the social sciences after the physical ones was not, of course, Moreno's invention. A hundred years before Moreno, the social philosopher Comte hoped to found a new field of "social physics." Fifty years after Comte, the French sociologist Durkheim had argued that human societies were like biological systems in that they were made up of interrelated components. As such, the reasons for social regularities were to be found not in the intentions of individuals but in the structure of the social environments in which they were embedded (*4*). Moreno's sociometry provided a way of making this abstract social structure tangible.

In the 1940s and 1950s, work in social networks advanced along several fronts. One front was the use of matrix algebra and graph theory to formalize fundamental social-psychological concepts such as groups and social circles in network terms, making it possible to objectively discover emergent groups in network data (*5*). Another front was the development of a program of laboratory experimentation on networks. Researchers at the Group Networks Laboratory at the Massachusetts Institute of Technology (MIT) began studying the effects of different communication network structures on the speed and accuracy with which a group could solve problems (Fig. 2). The more centralized structures, such as the star structure, outperformed decentralized structures, such as the circle, even though it could be shown mathematically that the circle structure had, in principle, the shortest minimum solution time (*6*). Why the discrepancy? Achieving the mathematically optimal solution would have required the nodes to execute a fairly complex sequence of information trades in which no single node served as integrator of the information. But the tendency in human networks seemed to be for the more peripheral members of a network (i.e., the nodes colored blue in the "Star," "Y," and "Chain" networks in Fig. 2) to channel information to the most central node (i.e., the nodes colored red in Fig. 2), who then decided what the correct answer was and sent this answer back out to the other nodes. The fastest performing network structures were those in which the distance of all nodes from the obvious integrator was the shortest (*7*).

The work done by Bavelas and his colleagues at MIT captured the imagination of researchers in a number of fields, including psychology, political science, and economics. In the 1950s, Kochen, a mathematician, and de Sola Pool, a political scientist, wrote a highly circulated paper, eventually published in 1978 (*8*), which tackled what is known today as the "small world" problem: If two persons are selected at random from a population, what are the chances that they would know each other, and, more generally, how long a chain of acquaintanceship would be required to link them? On the basis of mathematical models, they speculated that in a population like the United States, at least 50% of pairs could be linked by chains with no more than two intermediaries. Twenty years later, Stanley Milgram tested their propositions empirically, leading to the now popular notion of "six degrees of separation" (*9*).

During this period, network analysis was also used by sociologists interested in studying the changing social fabric of cities. The common conviction at the time was that urbanization destroyed community, and that cities played a central role in this drama. These sociologists saw concrete relations between people—love, hate, support, and so

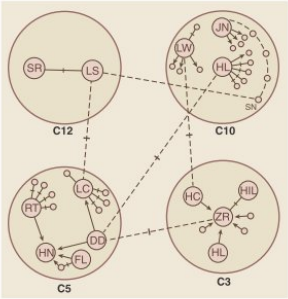


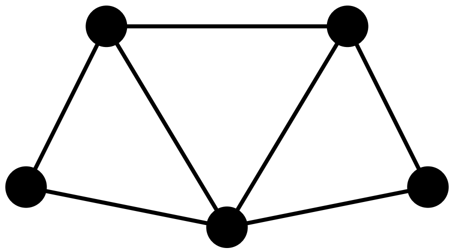
Fig. 1. Moreno's network of runaways. The four largest circles (C12, C10, C5, C3) represent cottages in which the girls lived. Each of the circles within the cottages represents an individual girl. The 14 runaways are identified by initials (e.g., SR). All nondirected lines between a pair of individuals represent feelings of mutual attraction. Directed lines represent one-way feelings of attraction.

LINKS Center for Network Research in Business, Gatton College of Business and Economics, University of Kentucky, Lexington, KY 40506–0034, USA. E-mail: sborgatti@uky.edu (S.P.B.), ajay.mehra@uky.edu (A.M.), dbrass@uky.edu (D.J.B.), and joe.labianca@uky.edu (G.L.)

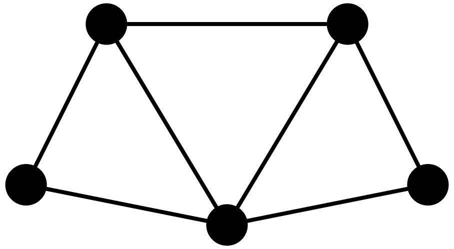
What is a network?

„A *network* is, in it simplest form, a collection of points, joined together in pairs by lines.“

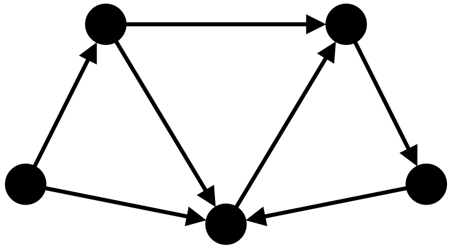
(Newman, 2009)



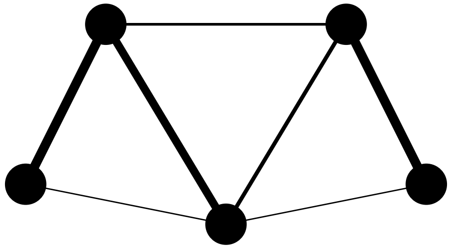
Despite the simple definition – there are many different ways of representing a network



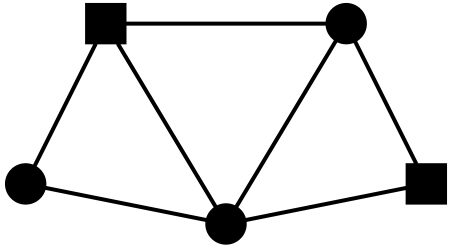
Undirected Network



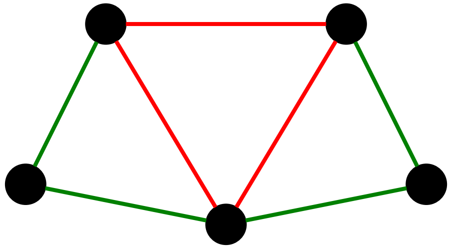
Directed Network



Weighted Network



Two-Mode Network



Signed Network

Notation: nodes and edges

Nodes

- Nodes can be individual people, firms, countries, or other organizations.
- They are also referred to as 'vertices', 'players', 'individuals' or 'actors', depending on the setting

$$N = \{1, \dots, n\}$$

- A network with one type of nodes is called *one-mode* network
- Networks with two types of nodes are referred to as *two-mode* networks
- If there are more than two types of nodes, we deal with a *multimodal* network

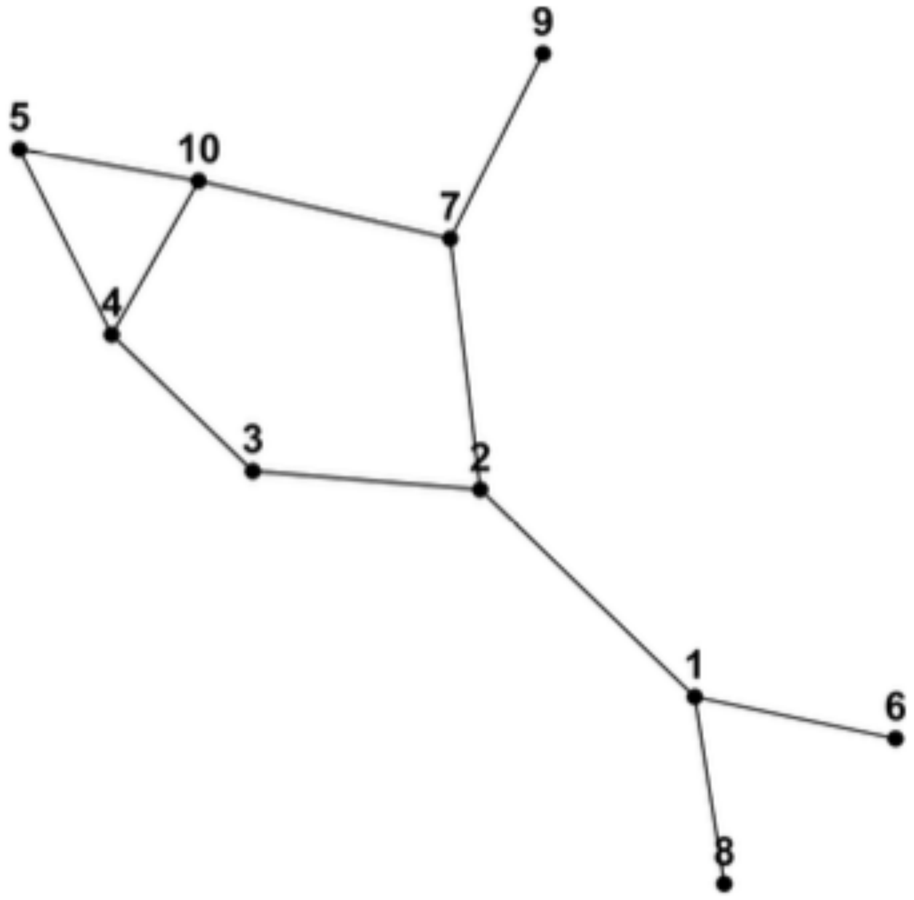
Edges

- Edges represent relationships between nodes and are also referred to as links or ties
- They are usually denoted as an *adjacency matrix* g, which is an $n \times n$ matrix
- Each element a_{ij} of g is 1, when there is an edge between node *i* and *j*, and 0 otherwise

$$g = \begin{pmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{ij} \end{pmatrix}$$

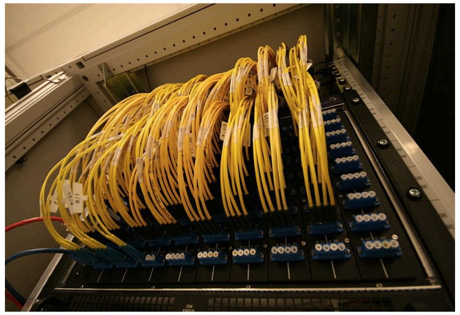
Example

- $N = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
- $g = \begin{pmatrix} 0&1&0&0&0&1&0&1&0&0 \\ 1&0&1&0&0&0&1&0&0&0 \\ 0&1&0&1&0&0&0&0&0&0 \\ 0&0&1&0&1&0&0&0&0&1 \\ 0&0&0&1&0&0&0&0&0&1 \\ 1&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&1&1 \\ 1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&1&1&0&1&0&0&0 \end{pmatrix}$



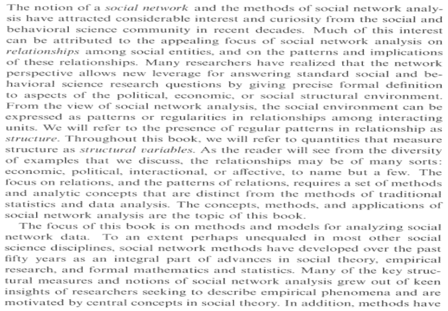
The fundamental characteristics of a network depend on the underlying system represented by the network

Social Networks



Technological Networks

Biological Networks

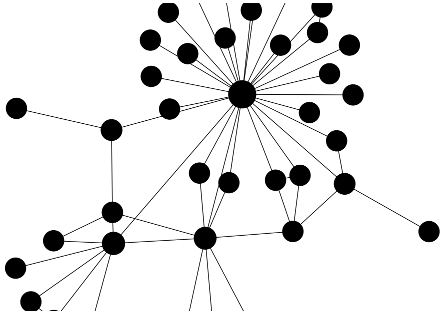


The notion of a *social network* and the methods of social network analysis have attracted considerable interest and curiosity from the social and behavioral science community in recent decades. Much of this interest can be attributed to the appealing focus of social network analysis on *relationships* among social entities, and on the patterns and implications of these relationships. Many researchers have realized that the network perspective allows new leverage for answering standard social and behavioral science research questions by giving precise formal definition to aspects of the political, economic, or social structural environment. From the view of social network analysis, the social environment can be expressed as patterns or regularities in relationships among interacting units. We will refer to the presence of regular patterns in relationship as *structure*. Throughout this book, we will refer to quantities that measure structure as *structural variables*. As the reader will see from the diversity of examples that we discuss, the relationships may be of many sorts: economic, political, interactional, or affective, to name but a few. The focus on relations, and the patterns of relations, requires a set of methods and analytic concepts that are distinct from the methods of traditional statistics and data analysis. The concepts, methods, and applications of social network analysis are the topic of this book.

The focus of this book is on methods and models for analyzing social network data. To an extent perhaps unequaled in most other social science disciplines, social network methods have developed over the past fifty years as an integral part of advances in social theory, empirical research, and formal mathematics and statistics. Many of the key structural measures and notions of social network analysis grew out of keen insights of researchers seeking to describe empirical phenomena and are motivated by central concepts in social theory. In addition, methods have

Information Networks

Transportation Networks



Others

Different types of relationships

Similarities			Social Relations				Interactions	Flows
Location	**Membership**	**Attribute**	**Kinship**	**Other role**	**Affective**	**Cognitive**	e.g.,	e.g.,
e.g.,	e.g.,	e.g.,	e.g.,	e.g.,	e.g.,	e.g.,	Sex with	Information
Same spatial and temporal space	Same clubs	Same gender	Mother of	Friend of	Likes	Knows	Talked to	Beliefs
	Same events	Same attitude	Sibling of	Boss of	Hates	Knows about	Advice to	Personnel
				Student of	etc.		Helped	Resources
	etc.	etc.		Competitor of		Sees as happy	Harmed	etc.
						etc.	etc.	

Fig. 3. A typology of ties studied in social network analysis.

(Borgatti et al., 2009)

Different types of flows

Table 1
Typology of flow processes

	Parallel duplication	Serial duplication	Transfer
Geodesics	<No process>	Mitotic reproduction	Package delivery
Paths	Internet name-server	Viral infection	Mooch
Trails	E-mail broadcast	Gossip	Used goods
Walks	Attitude influencing	Emotional support	Money exchange

(Borgatti et al., 2005)

Networks can be analyzed on different levels using multiple theoretical explanations for each level

Dyad level

- The unit of analysis is a set of two actors and their relationship
- e.g. strength, mutuality/reciprocity

Triad level

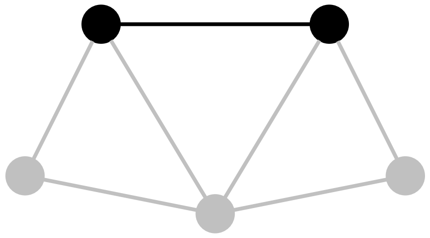
- Similar to the dyadic level, but with a set of three actors
- e.g. transitivity, cyclicality

Actor level

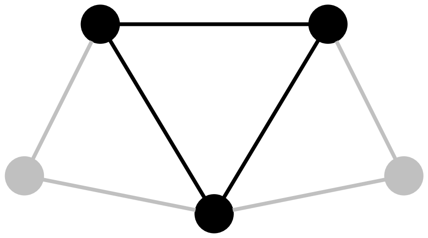
- The unit of analysis is a single actor and its direct neighborhood
- e.g. centralities

Global level

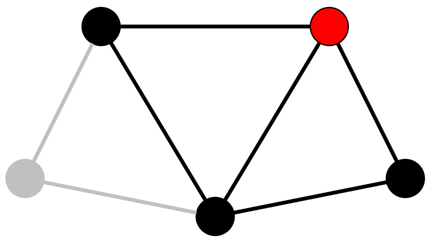
- Focus on the complete network
- e.g. density, centralization, clustering, distance and reachability



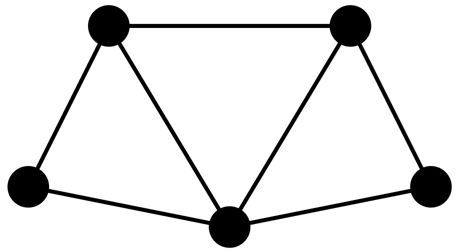
dyad



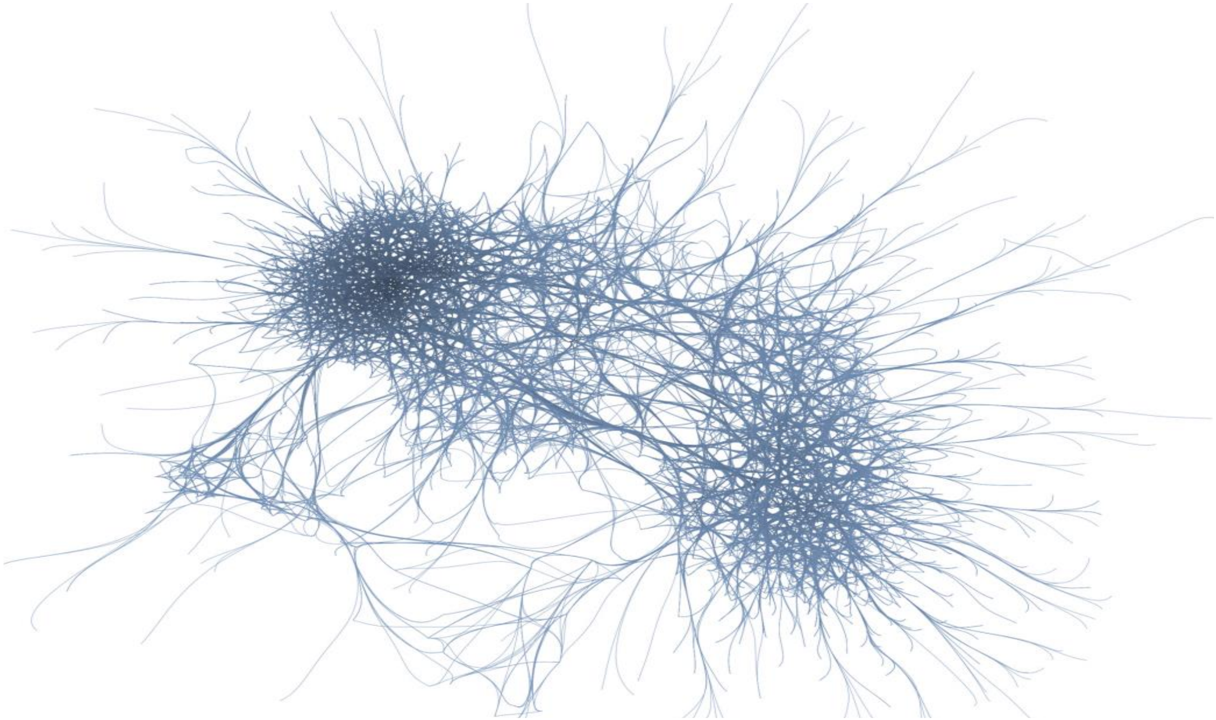
triad



ego network



Global network



Moving on to analyzing networks

Network data can be stored in various ways – different formats come with different (dis-) advantages

- The most common formats are simple plain text matrices, edge- and nodelists, e.g.:
  - CSV
  - NET Pajek
  - DL Ucinet
- Other based formats can be used to store more complex node, edge and network properties, e.g.:
  - GML
  - GraphML
  - GEXF
- Visit [gephi's wiki]() for a comprehensive overview

p2p-Gnutella08.txt — data

p2p-Gnutella08.txt

```
# Directed graph (each unordered pair of nodes is saved once): p2p-Gnutella08.txt
# Directed Gnutella P2P network from August 8 2002
# Nodes: 6301 Edges: 20777
# FromNodeId  ToNodeId
0 1
0 2
0 3
0 4
0 5
0 6
0 7
0 8
0 9
0 10
3 703
3 826
3 1097
3 1287
3 1591
3 1895
3 1896
3 1897
3 1898
3 1899
4 144
4 258
4 491
4 1021
4 1418
4 1669
4 1900
4 1901
4 1902
4 1903
5 121
5 127
5 128
5 179
5 247
5 249
```

„I would like to learn something about SNA – but where do I start and where do I find datasets?“

- A multitude of network datasets is available for free, for example:
- On websites and blogs of scholars and authors, e.g.:
  - MEJ Newman
  - AL Barabási
- On repositories of universities, software projects and other organizations, e.g.:
  - Pajek Datasets
  - UCINET Datasets
  - Gephi Datasets
  - SNAP - Stanford Large Network Dataset Collection

STANFORD UNIVERSITY

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks and metadata : Talk, editing and voting data from Wikipedia
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

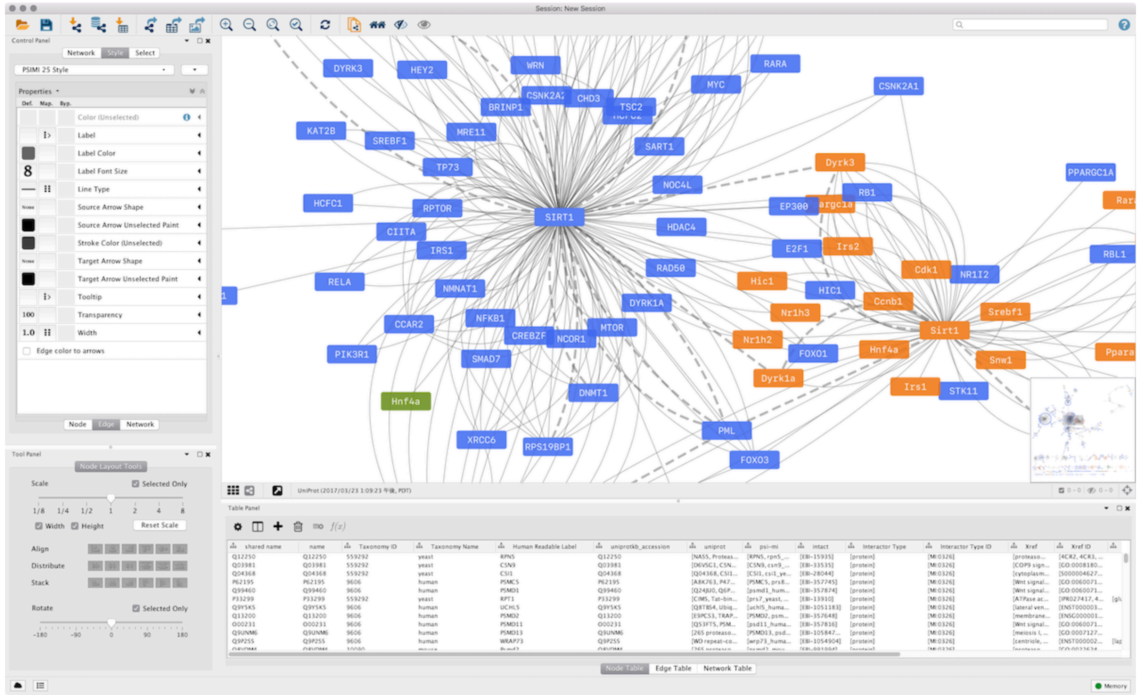
SNAP networks are also availalbe from UF Sparse Matrix collection. **Visualizations of SNAP networks** by Tim Davis.

Social networks

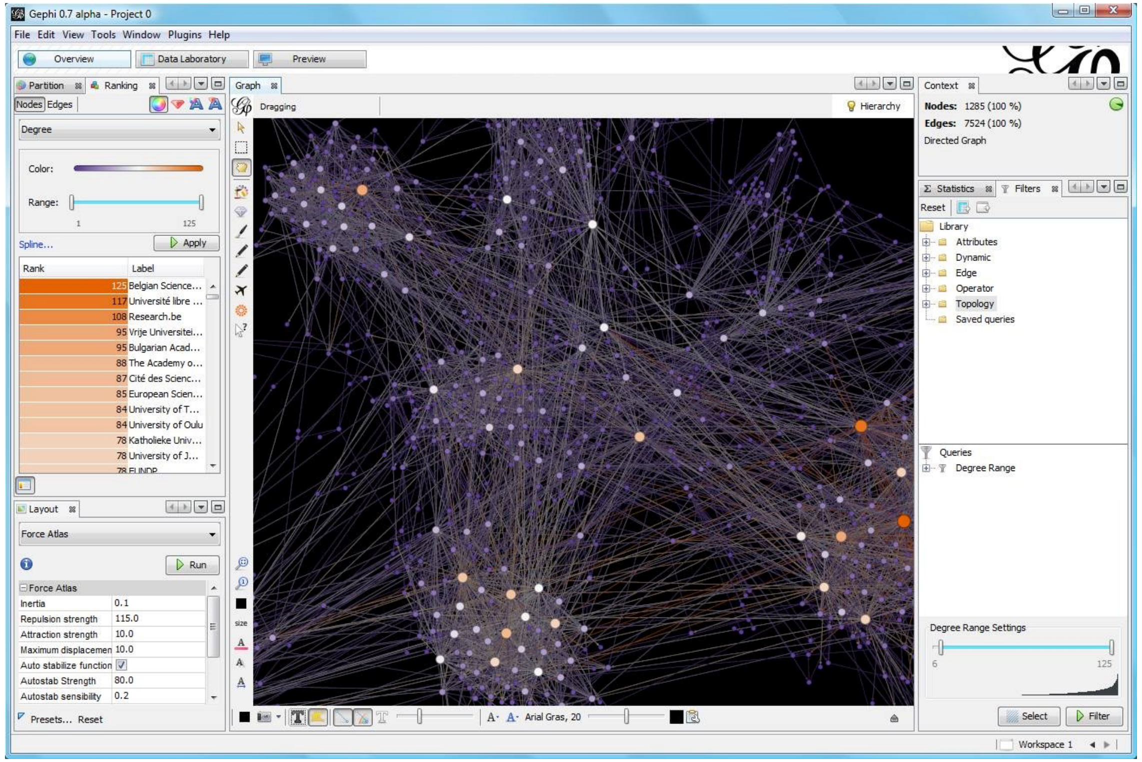
Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter
soc-Epinions1	Directed	75,879	508,837	Who-trusts-whom network of Epinions.com
soc-LiveJournal1	Directed	4,847,571	68,993,773	LiveJournal online social network
soc-Pokec	Directed	1,632,803	30,622,564	Pokec online social network
soc-Slashdot0811	Directed	77,360	905,468	Slashdot social network from November 2008
soc-Slashdot0922	Directed	82,168	948,464	Slashdot social network from February 2009
wiki-Vote	Directed	7,115	103,689	Wikipedia who-votes-on-whom network

Tools focused on visualization

Cytoscape

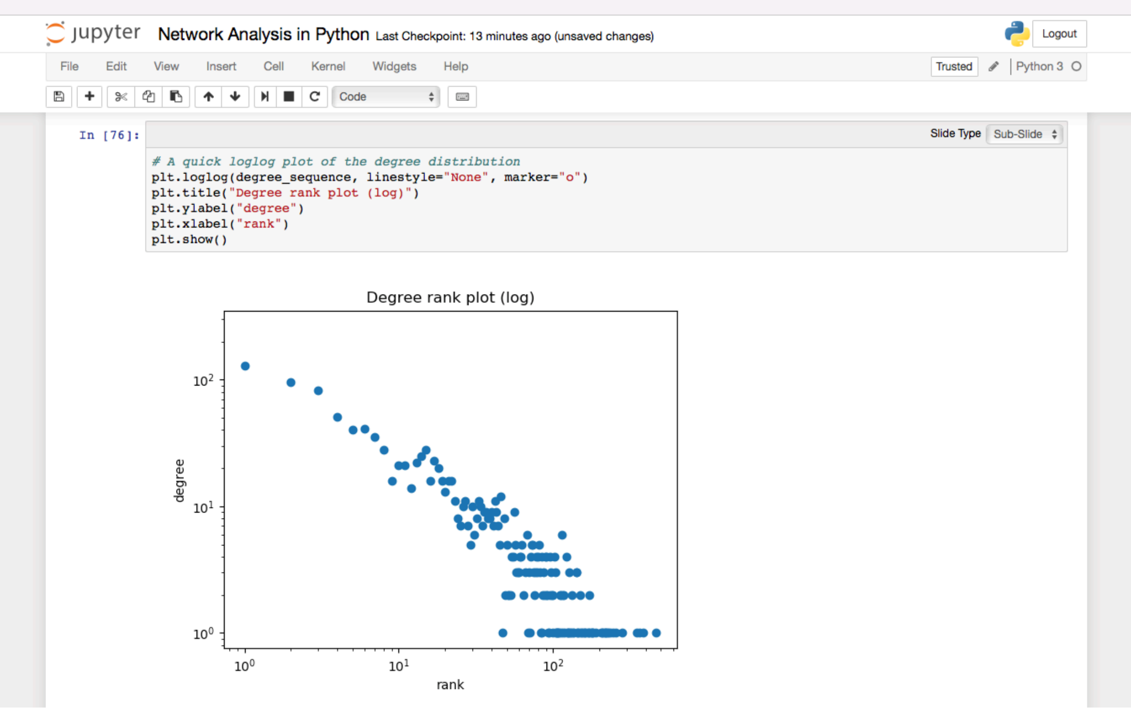


Gephi

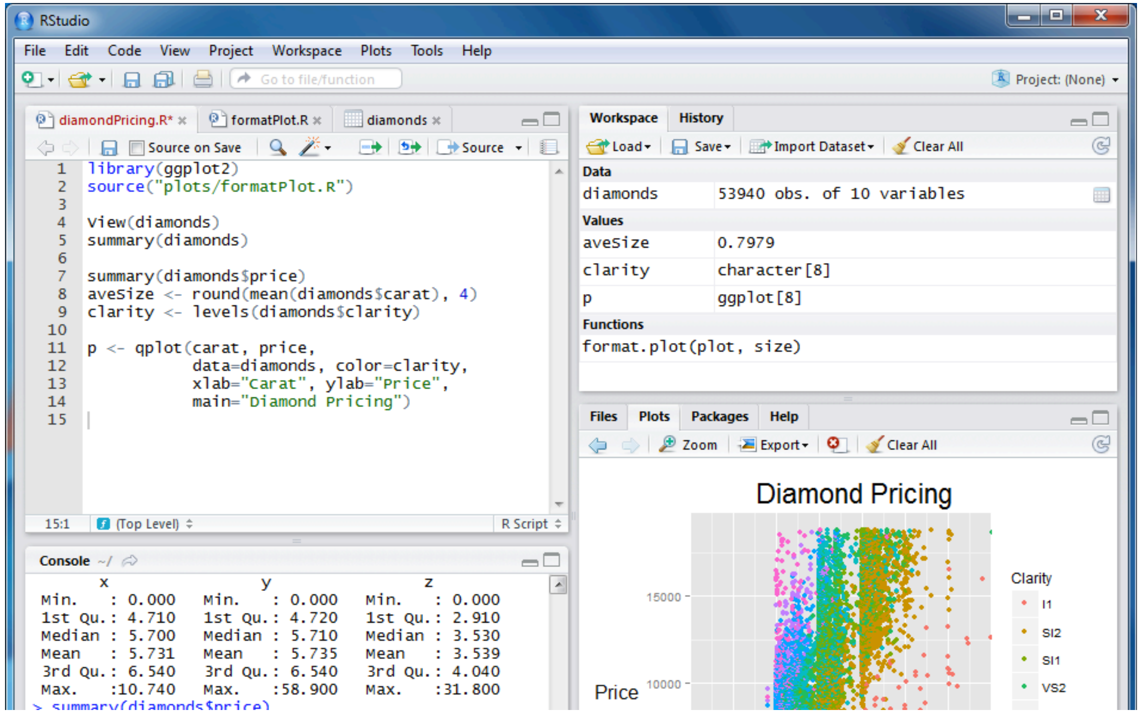


Tools focused on computation

Python

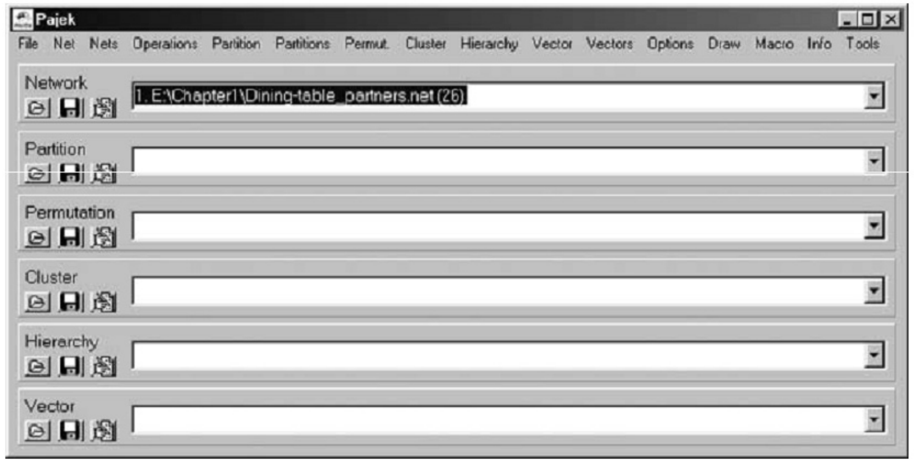


R

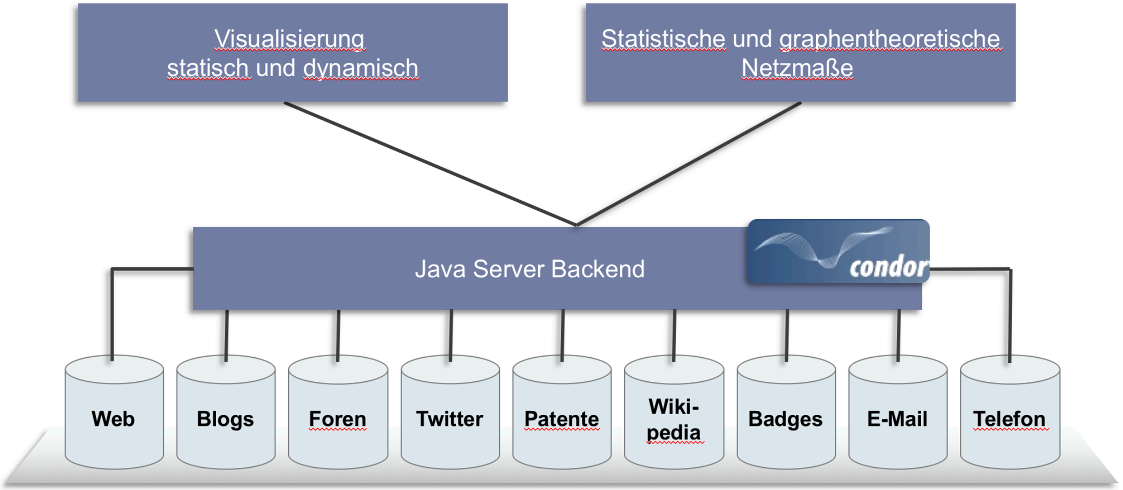


Hybrids and specialized tools

Pajek

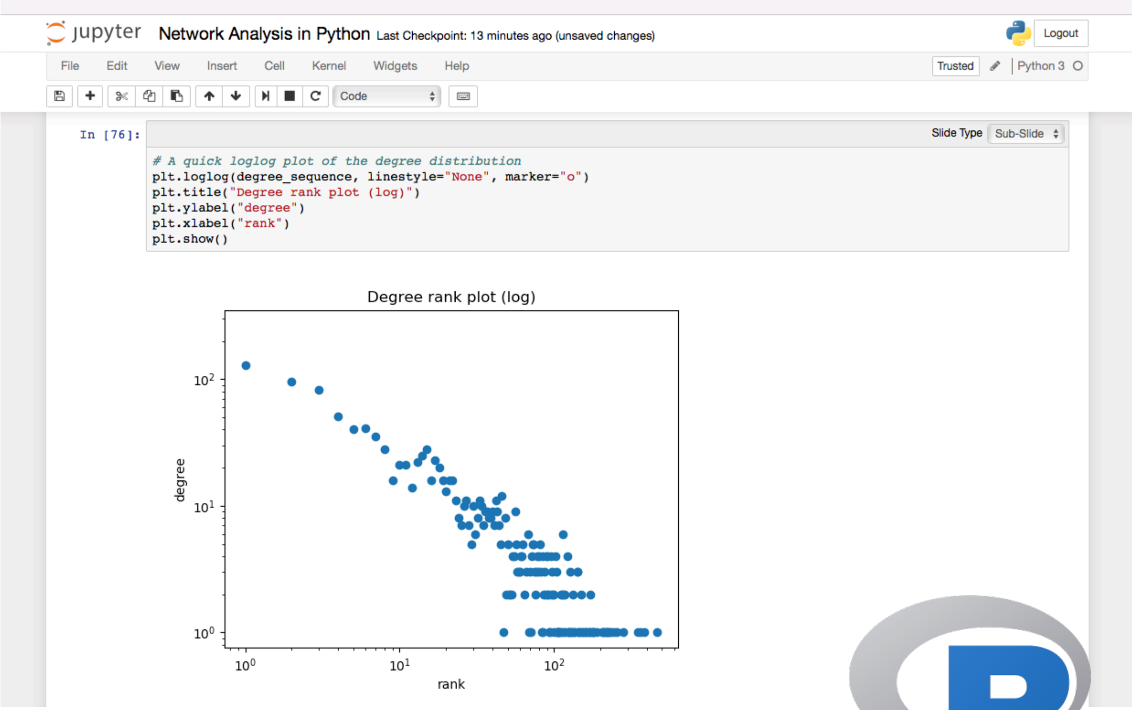


Condor



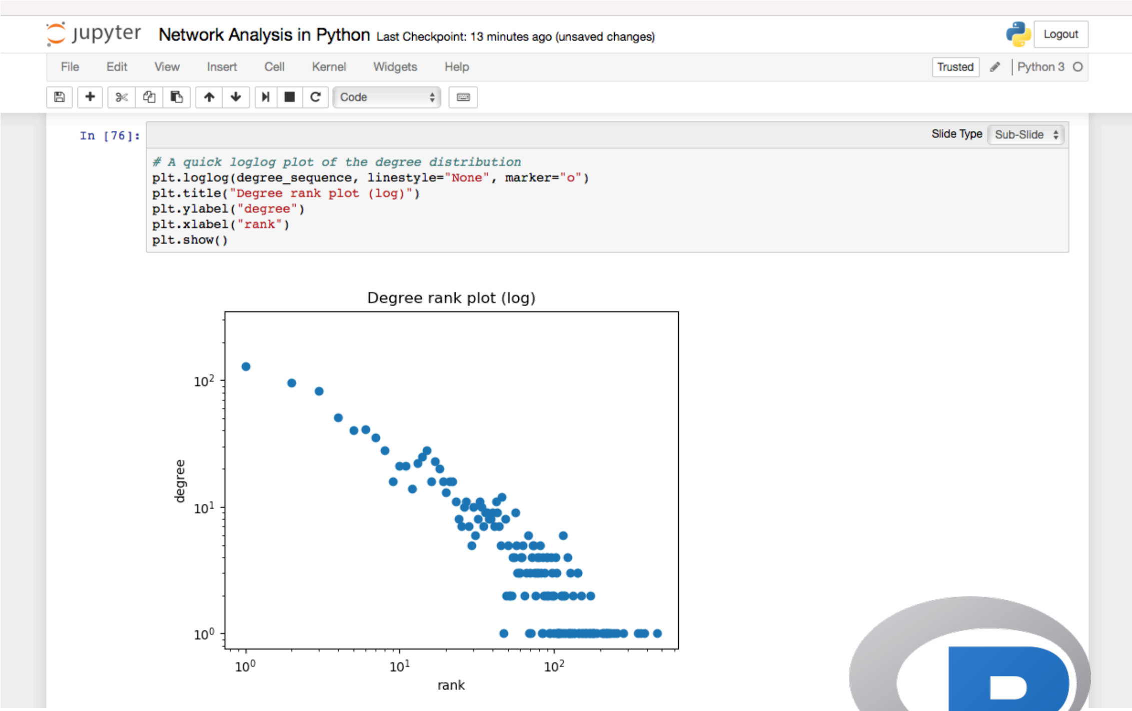
Why should you use R for network analysis?

- Clean and easy to learn syntax
- Availability of libraries
- Strong and growing ecosystem
- Quality and documentation of libraries
- Widespread use in the data science community
- Available in several IDEs (e.g. RStudio, Jupyter)



Why should you **not** use R for network analysis?

- Data collection and data processing can be troublesome in R – other languages, like Python, are more capable in this regard
- Visualization options are limited for very large networks
- For some computations and operations, R does not scale well and requires advanced tricks and techniques



Introduction: igraph

- Well established library comprising many network analysis tools
- Comprehensive librabry of algorithms
- Focus on efficiency and portability
- Open source
- Actively maintained
- Available in R, Python, Mathematica, and C/C++
- Behaves well for small to large networks

README.md

R/igraph

R/igraph is an R package of the igraph network analysis library.

Installation

You can install the stable version of R/igraph from CRAN:

```
install.packages("igraph")
```

For the development version, you can use Github, with the `devtools` package:

```
devtools::install_github("gaborcsardi/pkgconfig")
devtools::install_github("igraph/rigraph")
```

Documentation

See the igraph homepage for the complete manual.

Contributions

Please read our contribution guide.

License

GNU GPL version 2 or later