FIT3152 Data analytics – Lecture 5

Network Analysis

- Introduction: types of networks
- Network structure: elements
- Network statistics
- Centrality measures
- Using R for network analysis (igraph package)
- Examples

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
2/3/21	1	Intro to Data Science, review of basic statistics using R	...		
9/3/21	2	Exploring data using graphics in R	T1		
16/3/21	3	Data manipulation in R	T2	Released	
23/3/21	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
30/3/21	5	Network analysis	T4		
6/4/21		Mid-semester Break			
13/4/21	6	Regression modelling	T5		
20/4/21	7	Classification using decision trees	T6	Submitted	
27/4/21	8	Naïve Bayes, evaluating classifiers	T7		Released
4/5/21	9	Ensemble methods, artificial neural networks	T8		
11/5/21	10	Clustering	T9		
18/5/21	11	Text analysis	T10		Submitted
25/5/21	12	Review of course, Exam preparation	T11		

Assignment 1

Assignment 1

FIT3152 Data analytics: Assignment 1

This assignment is worth 20% of your final marks in FIT3152. Due: Friday 23rd April 2021.

Activity, language use and social interactions in an on-line community. Analyse the metadata and linguistic summary from a real on-line forum and submit a report of your findings. Do the following:

Assignment 1

a. Analyse activity and language on the forum over time. Some starting points:

- Describe your data: How active are participants, and are there periods where this increases or decreases? Is there a trend over time?
- Looking at the linguistic variables, do these change over time? Is there a relationship between variables?

b. Analyse the language used by groups. Some starting points:

- Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.
- By analysing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by different groups?
- Does the language used within threads (or between threads) change over time? How consistent or variable is the language used within threads?

Assignment 1

c. Challenge: Social networks online. We can think of participants posting to the same thread at similar times (for example during the same month) as forming a social network. When these participants also post to other threads over the same period, their social network extends.

- Can you define, graph and describe the social network that exists at a particular point in time, for example over one month? How does this change in the following months?
- Note: you only need to analyse a small portion of the social network over a short time period. We will cover social network analysis in Lecture 5.

d. Reflection on your investigation. What did you first investigate? How did you then modify your research based on the results of your first investigation?

- Using one of the data science methodologies in Lecture 4, illustrate your research process.

Assignment 1

Data

The data is contained in the file webforum.csv and consists of the metadata and linguistic analysis of posts over the years 2002 to 2011. You will each work with 20,000 posts, randomly selected from the original file. The linguistic analysis was conducted using Linguistic Inquiry and Word Count (LIWC), which assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication. See http://liwc.wpengine.com/ for more information, including the language manual http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf

Create your individual data as follows:

```
rm(list = ls())
set.seed(XXXXXXXX) # XXXXXXXX = your student ID
webforum <- read.csv("webforum.csv")
webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows
```

Assignment 1

ThreadID	AuthorID	Date	Time	WC	Analytic	Clout	Authentic	Tone	WPS	i	we	you	they	number	affect	posemo	negemo	anx
659289	193537	24/11/2009	5:36	53	82.26	71.43	25.14	25.77	26.5	0	1.89	0	3.77	3.77	3.77	1.89	1.89	0
432269	136196	26/11/2007	23:42	216	25.71	94.73	45.81	33.77	24	1.85	6.48	0.46	5.09	0.46	6.02	3.24	2.78	0
572531	170305	17/02/2009	7:31	136	31.61	67.04	28.81	79.41	13.6	3.68	0	5.15	2.94	0.74	9.56	5.88	2.94	0.74
230003	32359	7/09/2005	21:25	29	39.74	91.6	3.81	85.87	14.5	3.45	0	6.9	0	6.9	3.45	3.45	0	0
459059	47875	19/02/2008	5:23	108	80.75	60.95	23.51	88.52	13.5	2.78	0	0	0	0.93	9.26	6.48	2.78	0
635953	181593	28/09/2009	8:40	86	64.98	45.37	57.24	1	43	1.16	0	0	5.81	3.49	3.49	0	3.49	0
235116	51993	29/09/2005	15:59	49	33.33	20.71	13.15	25.77	16.33	6.12	0	0	2.04	0	8.16	4.08	4.08	0
593767	169459	23/04/2009	19:21	368	85.91	63.82	19.13	7.15	24.53	1.36	2.17	0	0.54	0.54	5.43	1.9	3.53	0.54
532649	248548	25/12/2011	8:28	13	92.84	50	1	25.77	13	0	0	0	0	61.54	0	0	0	0
517685	65	20/02/2005	10:50	65	91.21	62.1	33.6	81.28	13	7.69	0	0	0	0	9.23	6.15	3.08	0
588291	158329	23/04/2009	23:40	265	55.7	73.95	45.85	11.21	44.17	1.89	1.13	0.38	3.4	5.66	3.4	1.13	2.26	0
29936	194	25/07/2002	4:29	106	80.44	80.2	20.42	98.46	15.14	1.89	0	4.72	0	0.94	7.55	6.6	0.94	0.94
199787	47875	20/05/2005	16:48	160	94.48	73.4	2.07	5.64	22.86	1.25	0	0	0	5.62	8.12	3.12	5	1.88
545552	143229	24/11/2008	23:39	33	79.25	18.16	98.01	80.64	8.25	6.06	0	0	0	3.03	3.03	3.03	0	0
303058	88912	25/07/2006	23:57	244	44.21	65.92	33.49	7.09	27.11	2.87	0.82	0.41	4.51	1.64	6.56	2.46	4.1	0
772248	75628	16/01/2011	2:24	108	39.91	57.35	45.81	25.77	13.5	5.56	0	2.78	0	0.93	1.85	0.93	0.93	0
761807	227011	4/12/2010	23:48	104	73.9	57.63	74.76	62.24	34.67	0.96	0	2.88	3.85	2.88	5.77	3.85	1.92	0
110837	34501	24/01/2004	2:53	49	90.62	20.71	46.05	1	24.5	2.04	0	0	0	0	6.12	0	6.12	0
636255	180475	3/09/2009	22:25	2	92.84	99	1	99	2	0	0	0	0	0	50	50	0	0
178736	43291	18/01/2005	2:40	75	69.57	92.87	1	1	15	0	0	2.67	6.67	0	10.67	1.33	9.33	1.33
275754	-1	6/03/2006	18:01	56	92.84	70.4	41.07	6.15	18.67	1.79	0	1.79	0	1.79	1.79	0	1.79	0
833308	231141	21/09/2011	21:39	32	78.67	82.58	74.76	25.77	16	0	0	6.25	0	0	0	0	0	0
642657	180098	13/11/2009	16:34	13	92.84	6.21	99	1	13	23.08	0	0	0	0	7.69	0	7.69	7.69
365246	116735	17/02/2007	9:48	48	49.05	33.83	62.53	1	48	2.08	0	2.08	2.08	0	10.42	2.08	8.33	4.17
279233	84070	21/03/2006	1:59	51	77.76	50	66.34	25.77	51	3.92	0	1.96	0	1.96	7.84	3.92	3.92	0
300539	-1	8/06/2006	22:43	24	49.05	33.83	23.51	92.4	6	8.33	0	0	4.17	8.33	4.17	4.17	0	0
277955	32925	14/03/2006	23:45	87	55.99	78.96	62.98	3.63	43.5	0	0	1.15	4.6	2.3	2.3	0	2.3	1.15
90325	32485	25/09/2003	3:30	48	94.65	79.76	3.9	25.77	12	0	0	0	2.08	2.08	12.5	6.25	6.25	0
321495	90627	12/09/2006	1:40	42	40.66	68.29	37.24	70.57	21	4.76	4.76	2.38	2.38	0	2.38	2.38	0	0
281667	79878	28/03/2006	2:45	60	32.98	56.63	65.14	1.03	20	1.67	1.67	0	3.33	0	3.33	0	3.33	0
294983	75902	21/05/2006	0:07	60	56.15	25.24	32.84	25.77	60	3.33	0	0	0	0	6.67	3.33	3.33	0
397699	125170	21/06/2007	21:41	34	92.84	92.92	14.7	25.77	17	0	2.94	2.94	0	0	5.88	2.94	2.94	0
313191	101368	30/07/2006	17:53	25	81.4	2.31	43.37	25.77	25	0	0	0	0	12	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Assignment 1

Data fields are (see the language manual for more detail and examples):

Column	Brief Descriptor
ThreadID	Unique ID for each thread
AuthorID	Unique ID for each author
Date	Date
Time	Time
WC	Word count of the text of the post
Analytic	LIWC Summary (Analytical thinking)
Clout	LIWC Summary (Power, force, impact)
Authentic	LIWC Summary (Using an authentic tone of voice)
Tone	LIWC Summary (Emotional tone)
WPS	LIWC (Words per sentence)
i	LIWC ("I, me, mine" words) First person singular
we	LIWC ("We, us, our" words) First person plural
you	LIWC ("You" words) Second person
they	LIWC ("They" words) Third person plural
number	LIWC(Quantities and ranks)
affect	LIWC (Expressing sentiment)
posemo	LIWC (Positive emotions)
negemo	LIWC (Negative emotions)
anx	LIWC (Indicating anxiety)

Assignment 1

Submission. Due Friday 23rd April 2021 11:55pm GMT+10.
Suggested length: 6–8 A4 pages + appendix.
Submit the results of your analysis, answering the research questions and report anything else you discover of relevance. If you choose to analyse only a subset of your data, you should explain why.

You are expected to include at least one multivariate graphic summarising key results. You may also include simpler graphs and tables. Report any assumptions you've made in modelling, and include your R code as an appendix. Submit your report as a single PDF with the file name *FirstnameSecondnameID.pdf* on Moodle.

Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but please document these in your report and include in your R code.

Assignment 1

Assessment criteria will include:

The quality of your analysis and description of your analytical process; Graphics and tables supporting your analysis; The quality of graphics used in the report. Justification of your findings and the degree of proof you can offer (for example statistical tests); Readability and quality of your written report; Insights gained from the data; Novelty of your approach.

Factors you should consider (starting points, not a complete list):

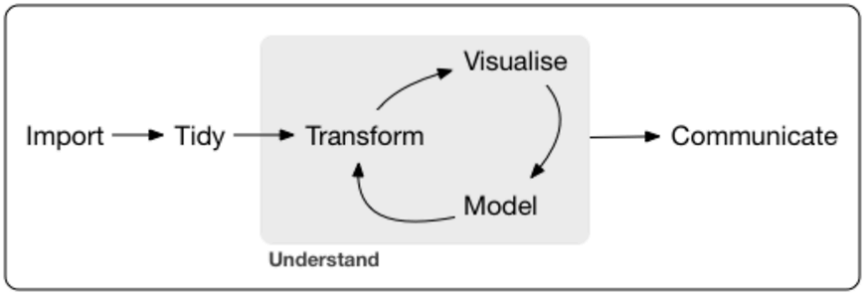
Techniques: summary/descriptive statistics, identification of important variables, networks, etc.
Major grouping variables: author, thread, date and/or time, or a combination of these.
Time window (days, weeks, months, years…); Subsets of the data to be analysed.
Graphics to communicate your analysis and insights (histograms, scatterplots, heat maps, time series are some basic starting points, but see https://datavizproject.com/ for inspiration.

Response to student questions

- John, do you have any advice with regards to dealing with a large dataset? The reason I ask because when I just plot raw data on a graph, it just results in a dump of black dots that simply cannot be analyzed.
  - > Plotting all the data together (any variable) will most likely be too noisy to tell you anything.
  - > Use grouping variables: time/threads (as per parts a and b). Look at individual threads. Look at individual LIWC variables.
  - > Previous lectures have shown you some manipulation techniques for grouping/summarising data.

Response to student questions

- Having difficulty starting? Consider the following:
  - > You will need to do some analysis to get a feel for the data before you decide on what to analyse.
  - > You may need to make some secondary tables of summaries (like we did for Dunnhumby problems).
  - > Expect to follow one of the KDD, SEMMA, CRISP-DM, R4DS models:



Review questions from last lecture

Please respond via Zoom chat if you want!

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

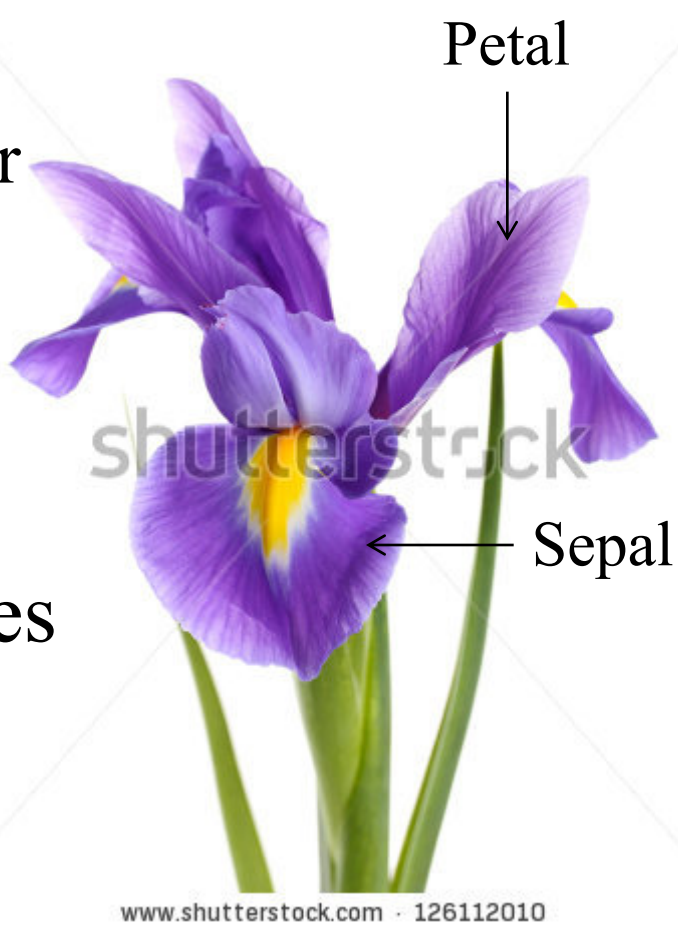
Four features measured:

- Sepal width and length
- Petal width and length

Is it possible to distinguish species using physical measurements?

- Data is packaged with R: "iris"

https://en.wikipedia.org/wiki/Iris_flower_data_set



Print

> iris *# = print(iris)*

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
...
```

Question 1

Predict the output from the following command:

```
> niris$Species = recode(niris$Species," 'versicolor' =
'0';'virginica' = '0';'setosa' = '1' ")
```

(a) Replace data in species column with 0 for I.versicolor and virginica, 1 for I.setosa.

(b) Add a new column of 0 and 1s

(c) Add a 0 or 1 to each species name

(d) Recode "setosa" = 1, leave others unchanged

Question 2

Which of the following is not a data science workflow methodology?

(a) CRISP-DM
(b) EDA
(c) KDD
(d) SEMMA

See (for amusement):
https://www.datasciencecentral.com/profiles/blogs/acronyms-of-big-data-analytics-from-a-to-z

Question 3

Which of the following data types is not dirty data – in its strictest sense?

(a) Duplicate data

(b) Business rule violation

(c) Inconsistent data

(d) Inexact data

Question 4

Which of the following is not true: Tidy data has:

(a) Each value in its own cell

(b) Each observation in its own row

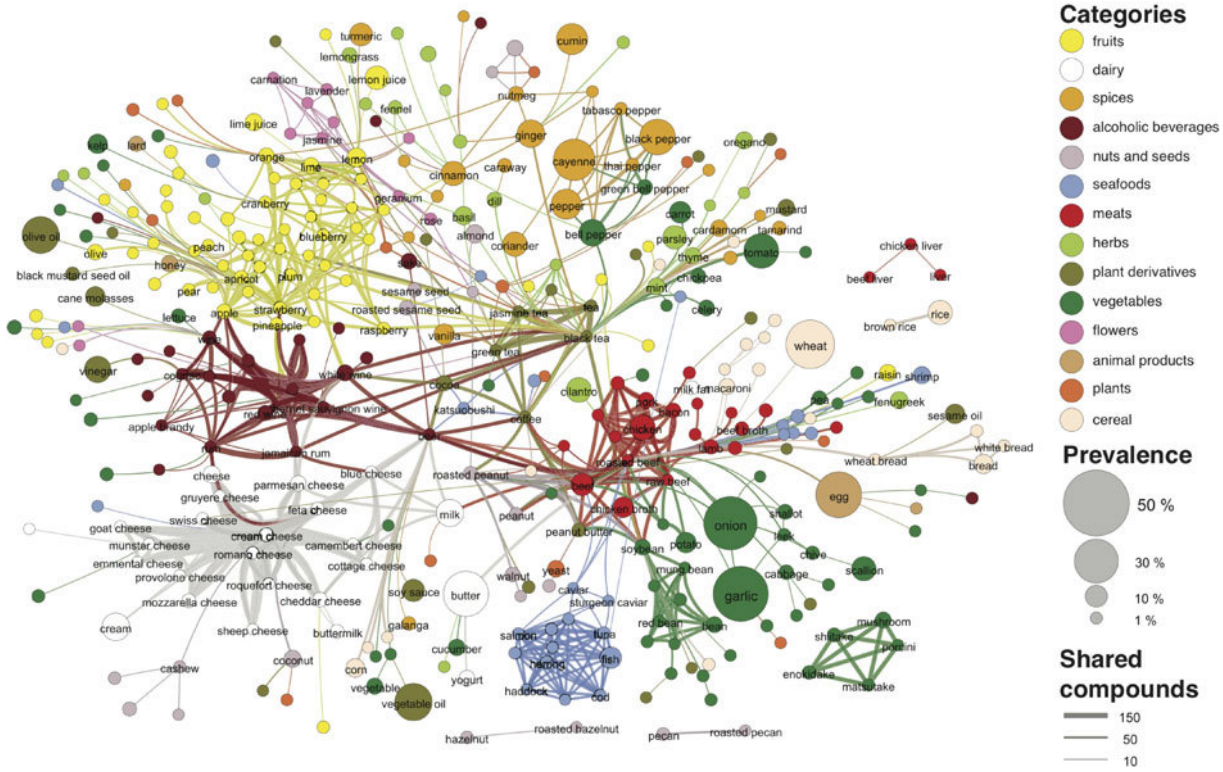
(c) Each factor level is a column index

(d) Each variable in its own column

Networks

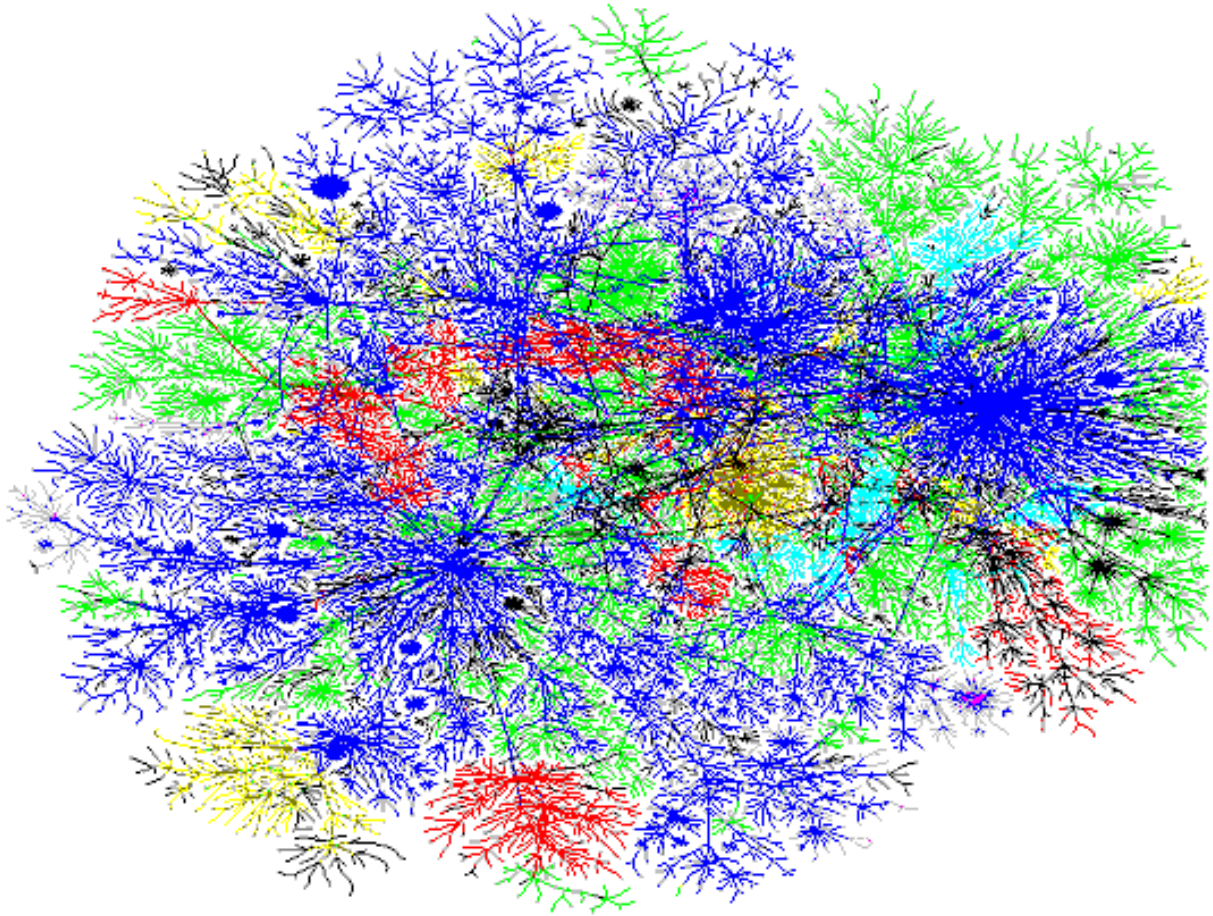
Flavor network

Flavor network and the principles of food pairing



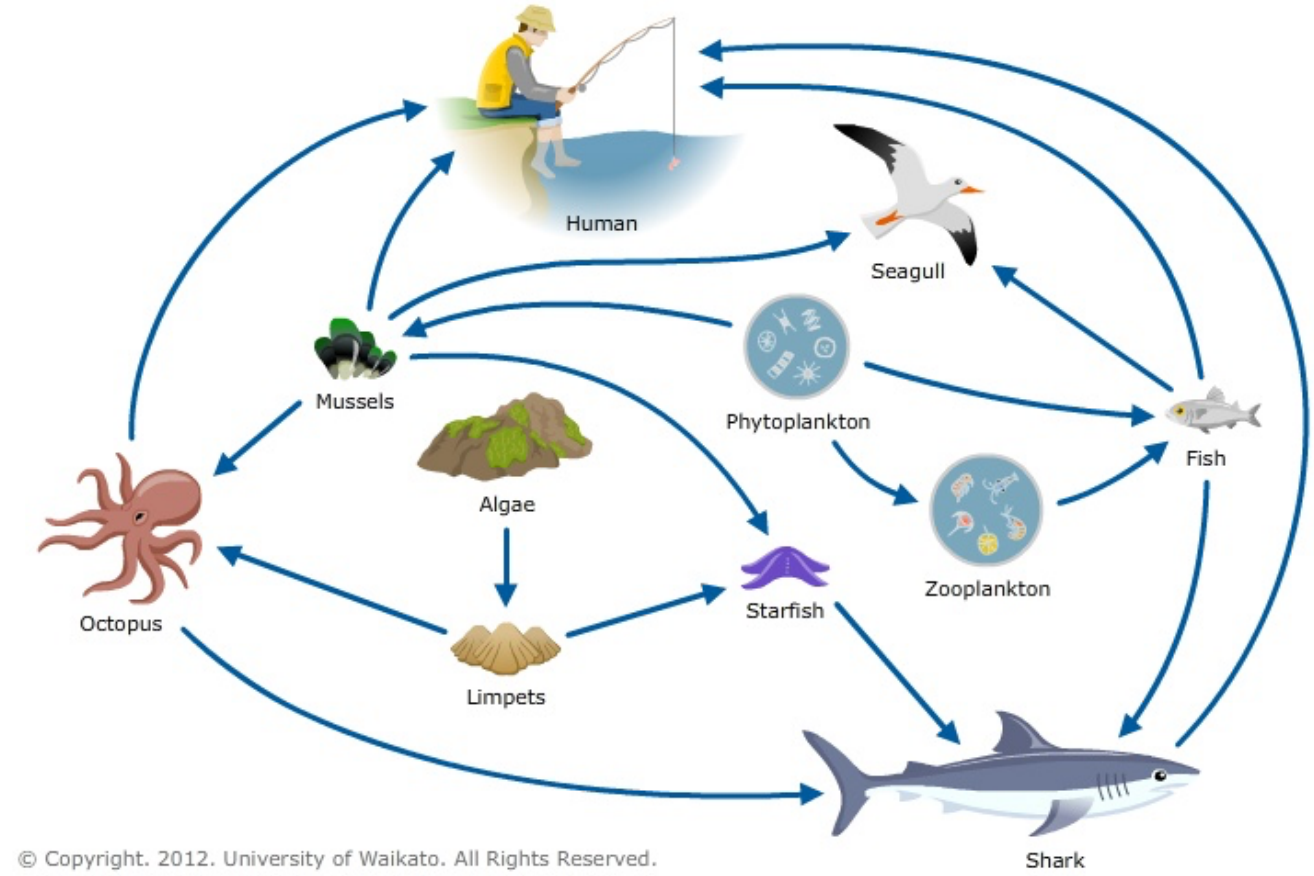
https://www.nature.com/articles/srep00196

The Internet



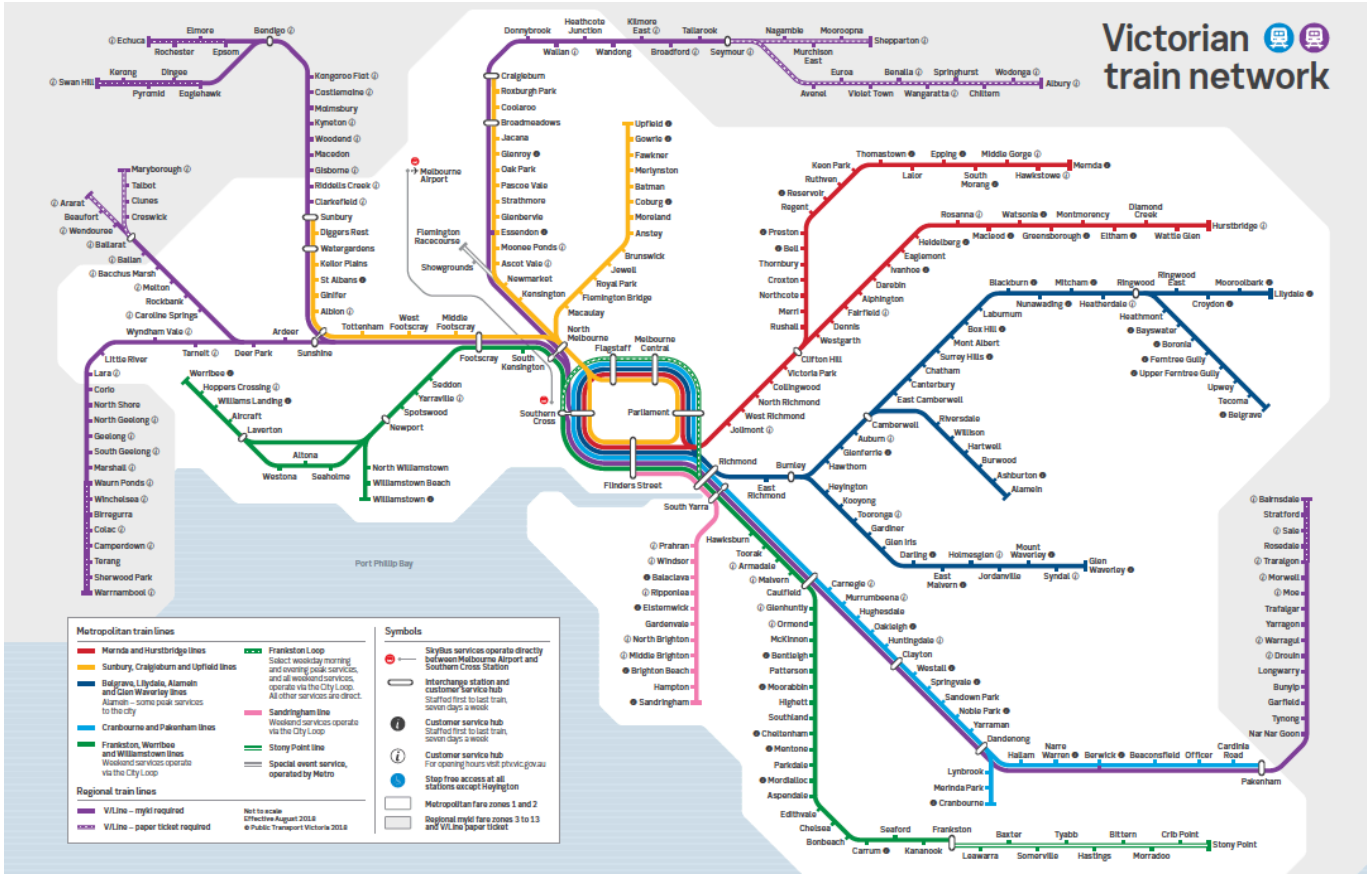
http://vv.arts.ucla.edu/thesis/cybergeog/atlas/topology.html

Food webs: predators and prey



https://www.sciencelearn.org.nz/resources/367-toxins-and-food-webs

Transport network

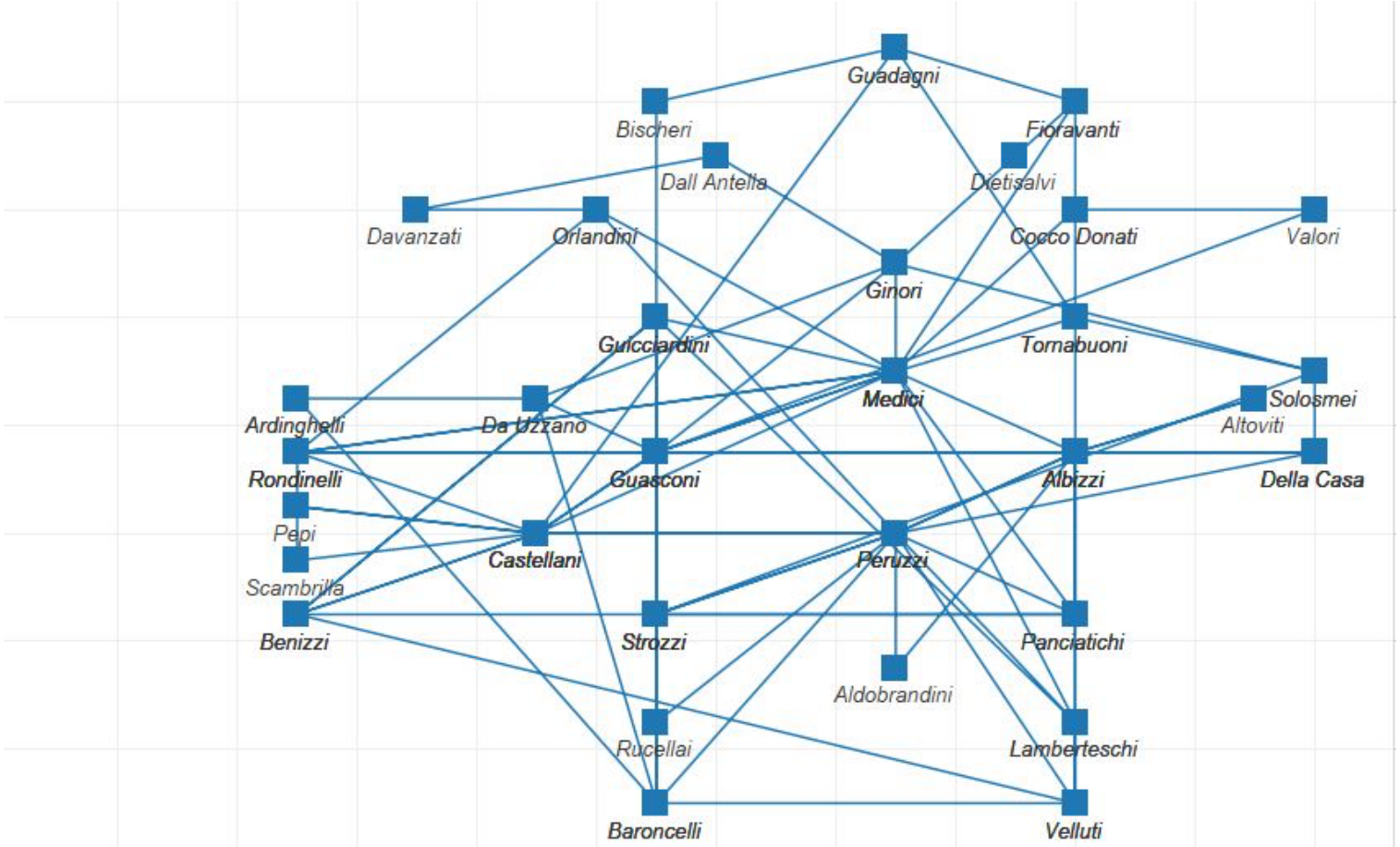


https://www.ptv.vic.gov.au/

Social networks

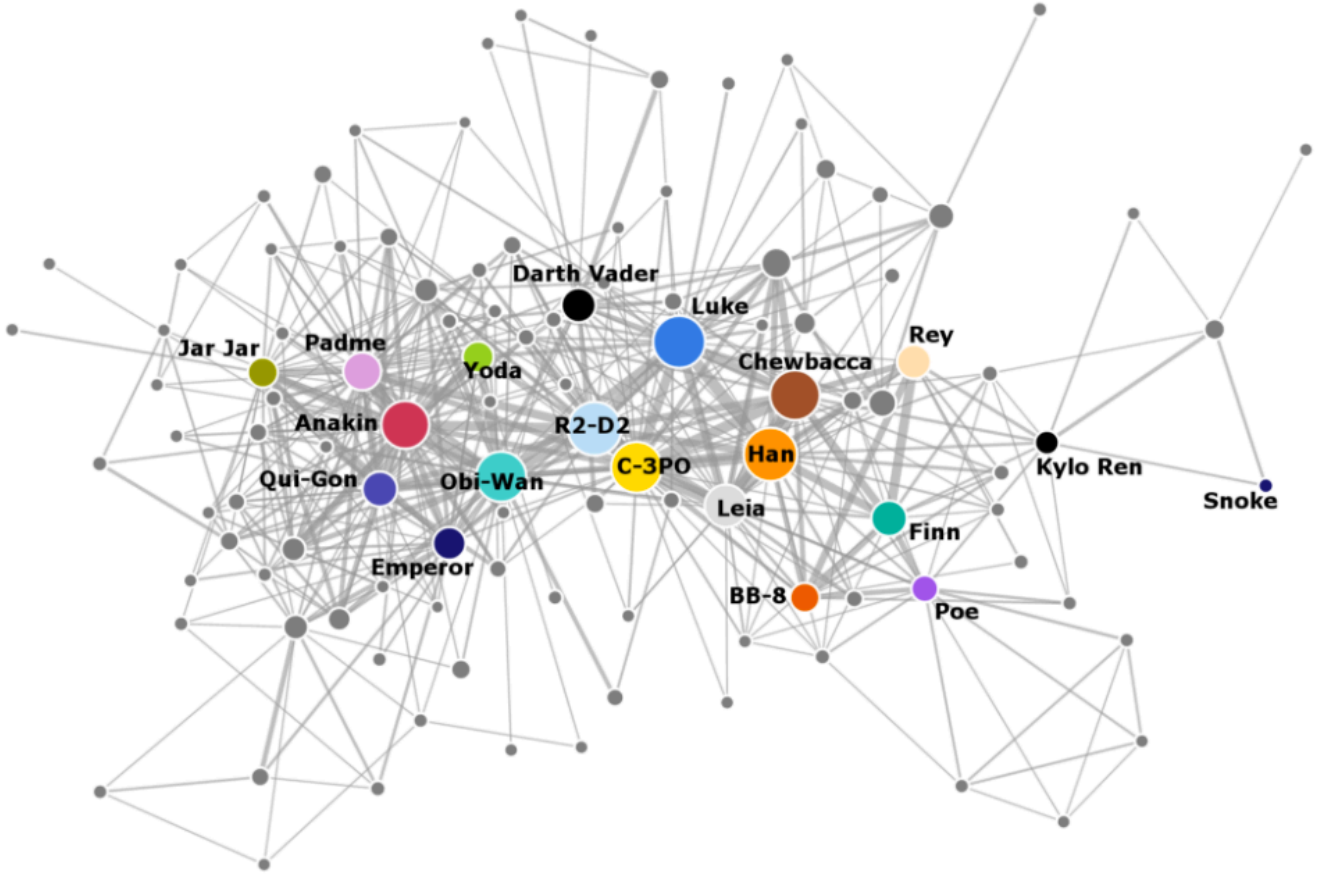
https://socialadr.com/features

Florentine social network: Medici



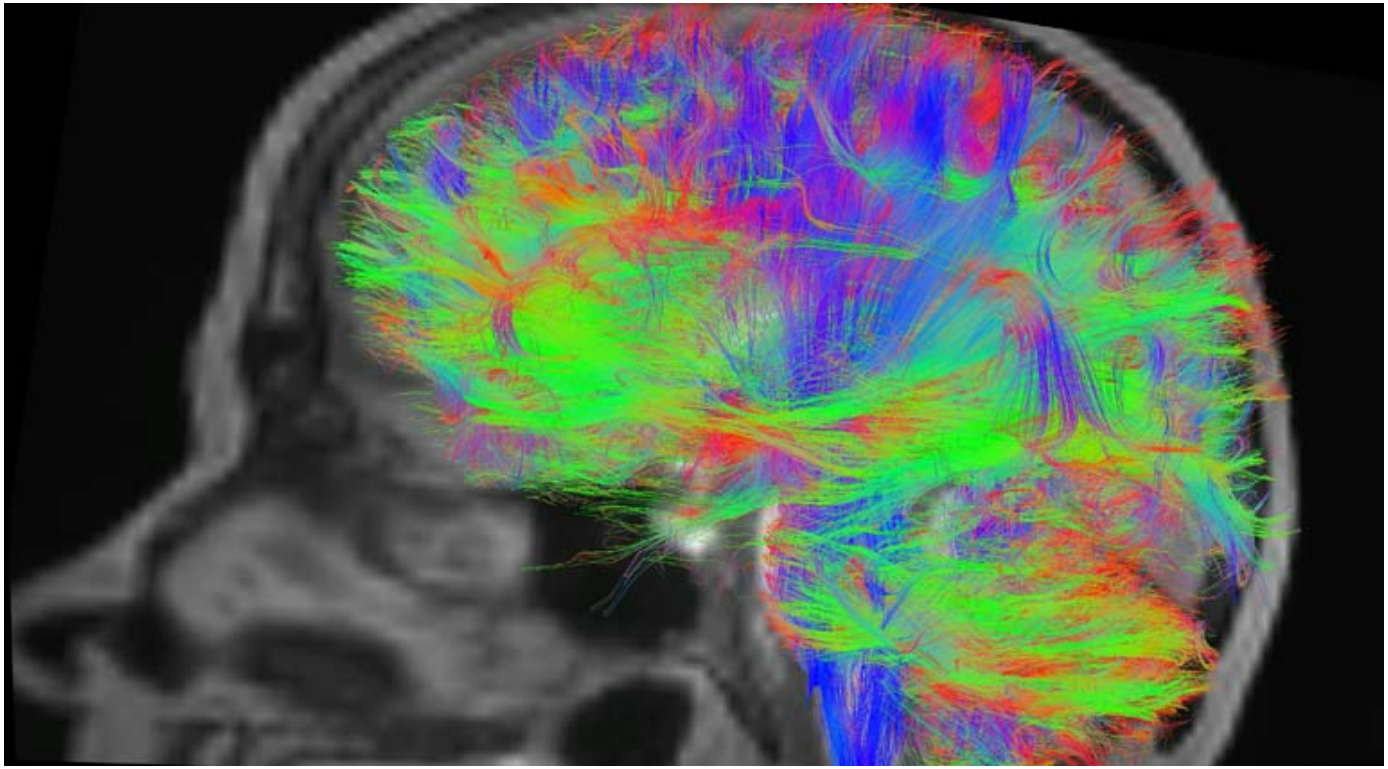
(inactive) http://blogs.teradata.com/international/tag/data-scientist/

Star Wars characters



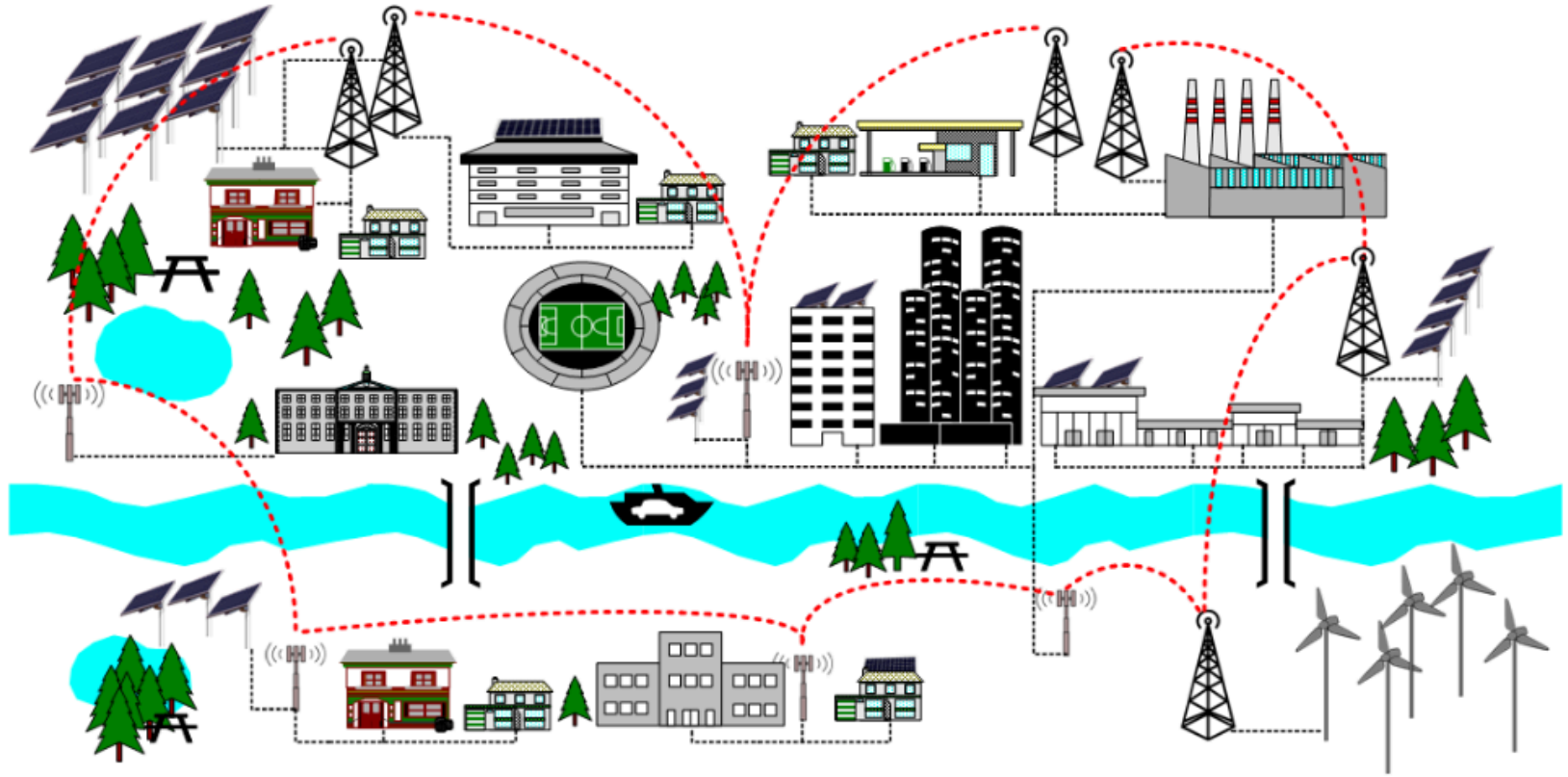
http://evelinag.com/blog/2016/01-25-social-network-force-awakens/#.WZwWUpOg8cl

Neural pathways



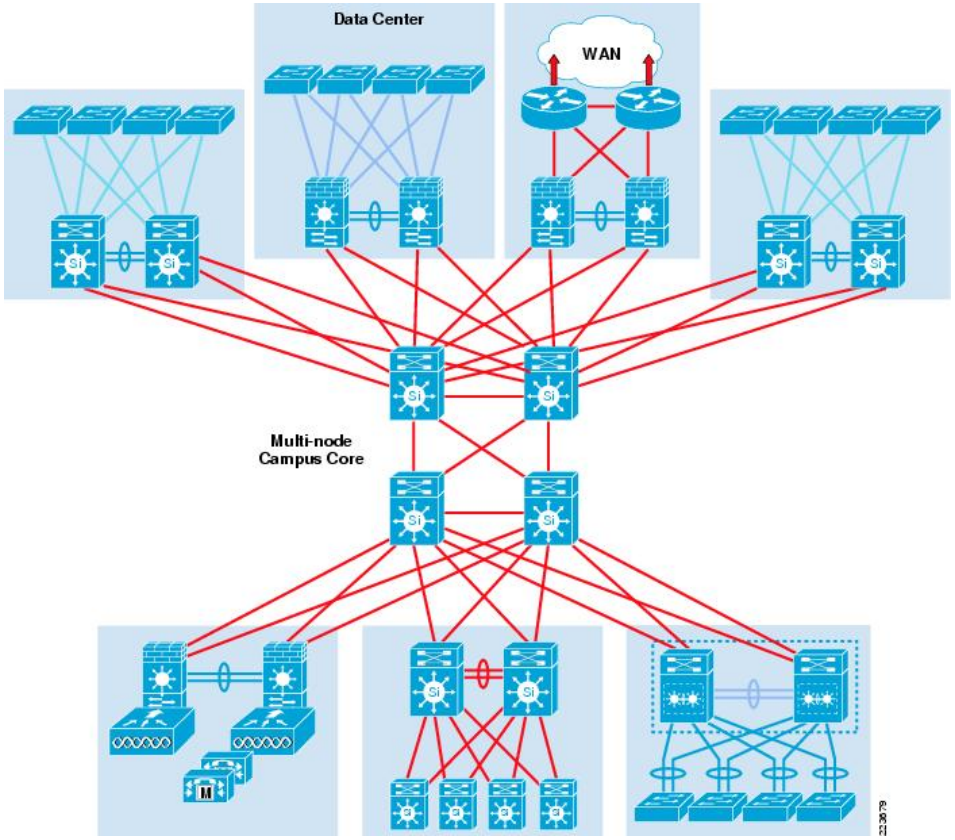
https://ki-galleries.mit.edu/2014/saygin-2

Power networks



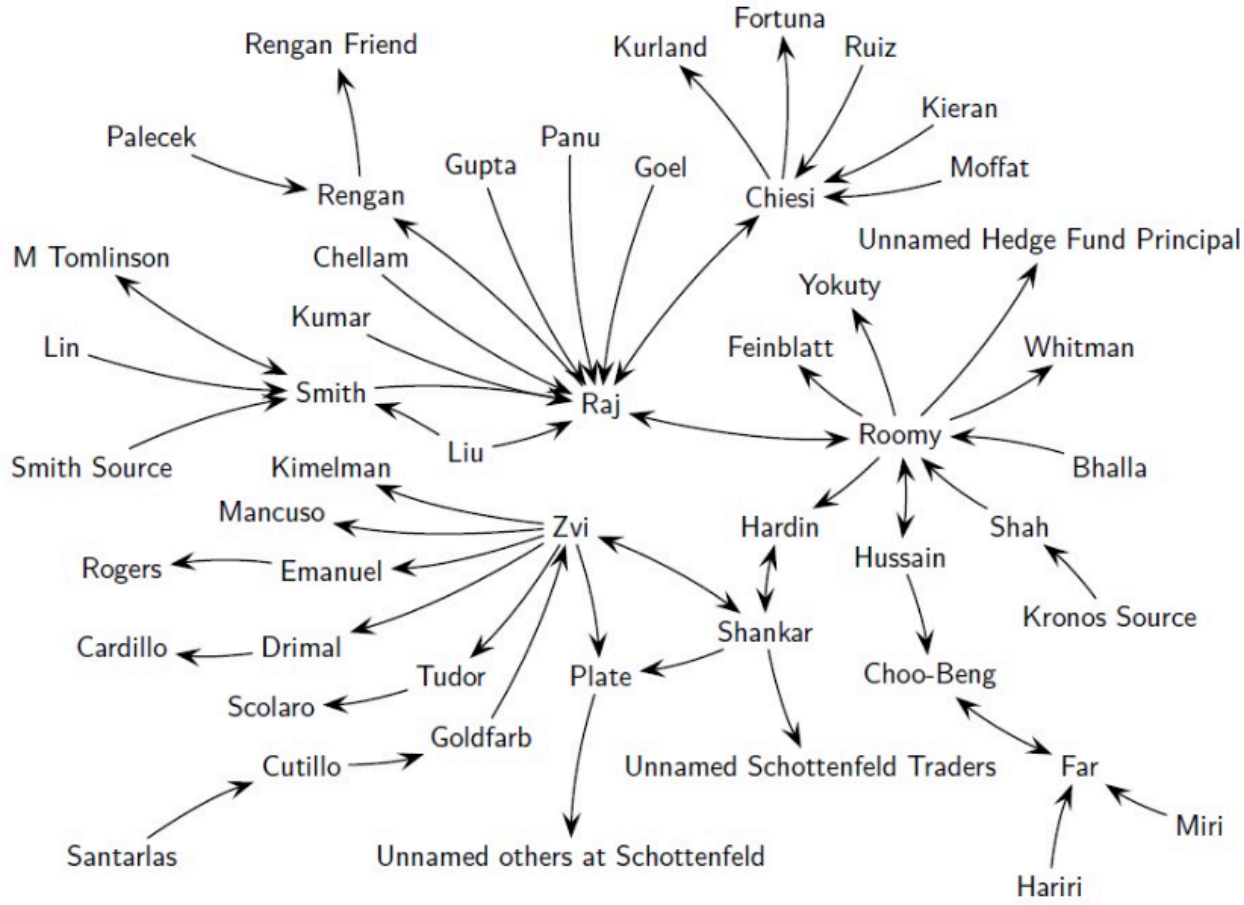
http://www.kios.ucy.ac.cy/ (inactive)

Enterprise systems



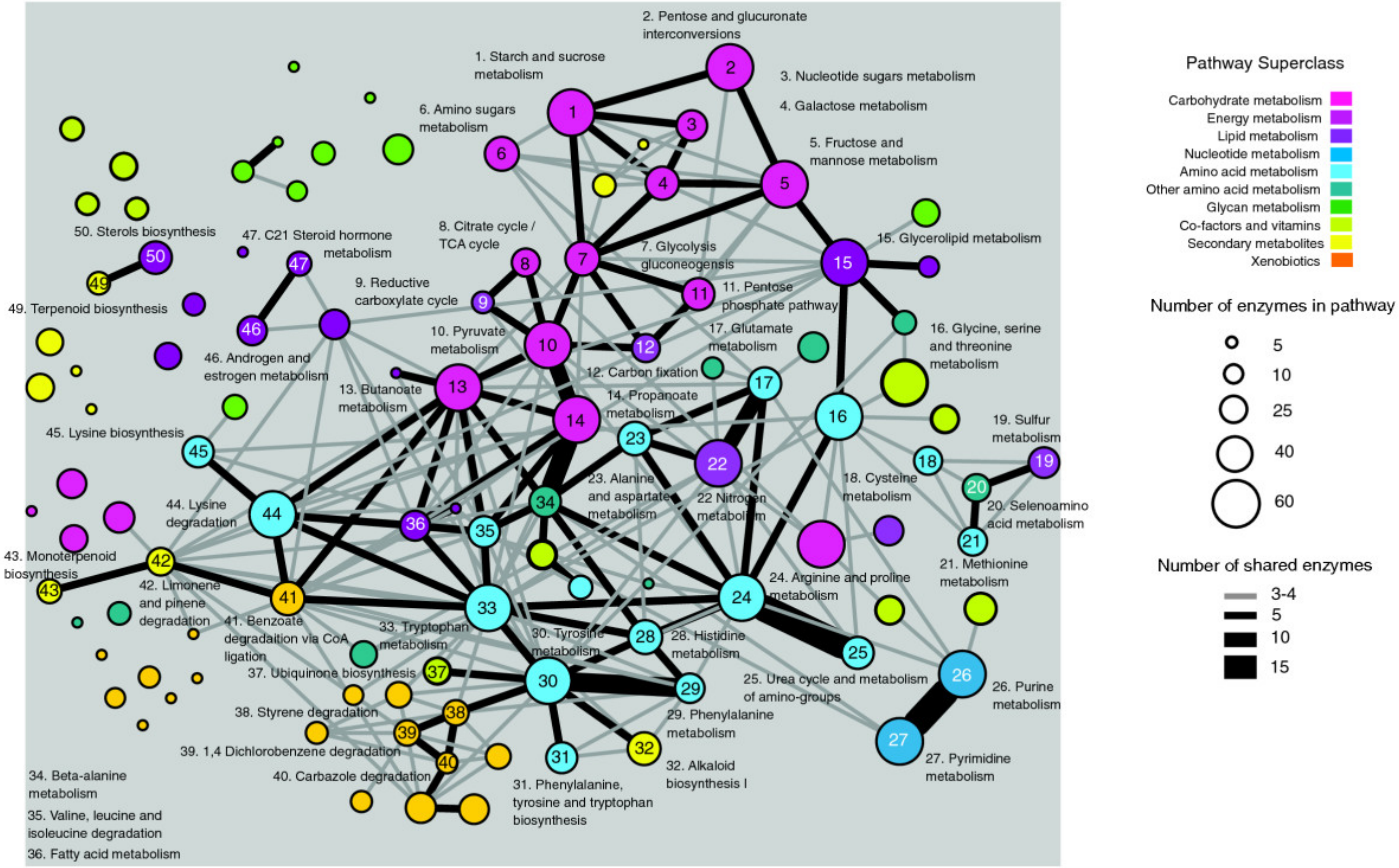
https://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Campus/campover.html

Information networks: insider trading



https://www.valuewalk.com/2015/03/information-networks-evidence-from-illegal-insider-trading-tips/

Metabolic pathways



https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-6-r63

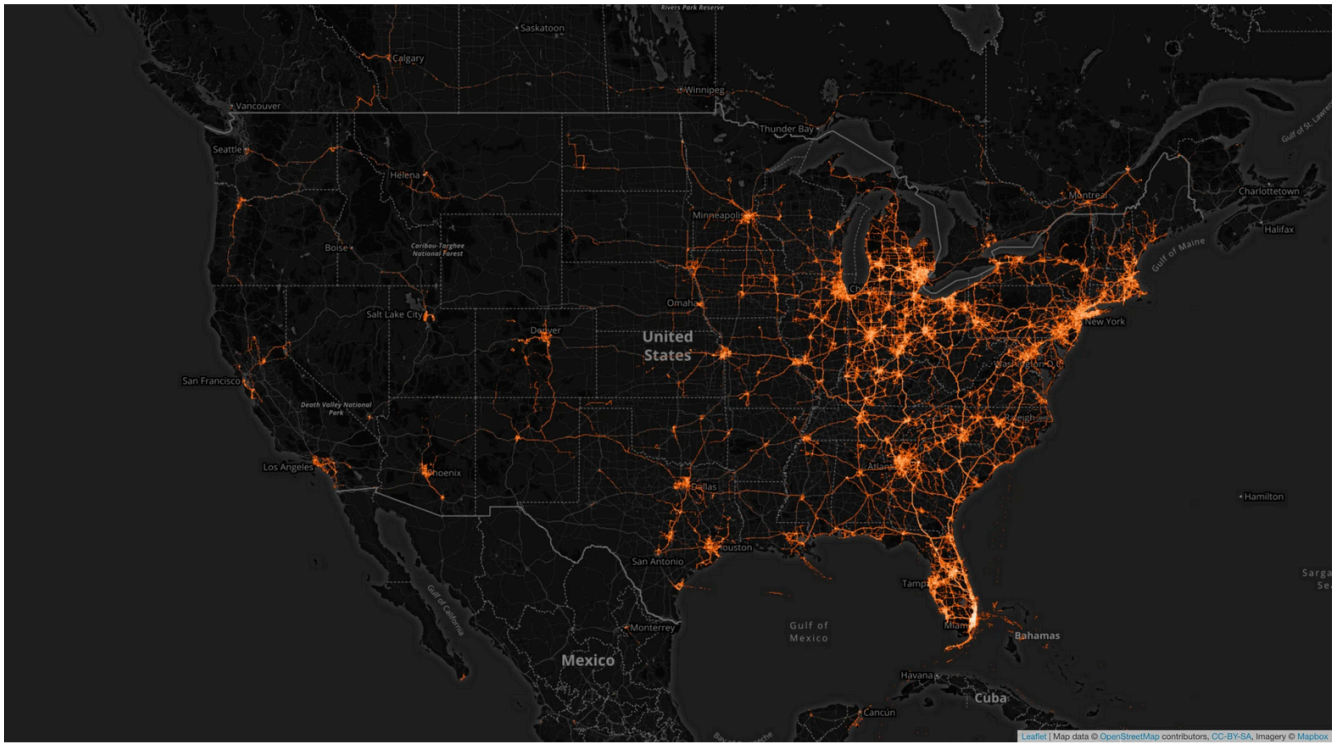
COVID-19 US social network

The Costly Toll of Not Shutting Down Spring Break Earlier

- People got sick — and some died — after attending crowded parties and theme parks in Florida as the coronavirus spread.
- A video by the data analytics company Tectonix showed how cellphones that were on one Fort Lauderdale beach at the beginning of March spread across the country over the next two weeks.

https://www.nytimes.com/2020/04/11/us/florida-spring-break-coronavirus.html

COVID-19 US social network



This image was generated by Tectonix GEO and X-Mode Social by analyzing secondary locations of anonymized mobile devices that were active at a single Fort Lauderdale beach during spring break. Courtesy Tectonix GEO

https://www.nytimes.com/2020/04/11/us/florida-spring-break-coronavirus.html

Why study networks?

Networks are everywhere:

- We all have a social network in the physical world, and an on-line network through Facebook, Instagram, …

Studying network structure lets us:

- Determine relative importance of key players in a social (or other) network.
- Understand the fundamental structure of natural or human networks. Reasons for fragility or robustness.

How to think about a network

For the previous examples, think about:

Elements in the network:

- Same?
- Different?

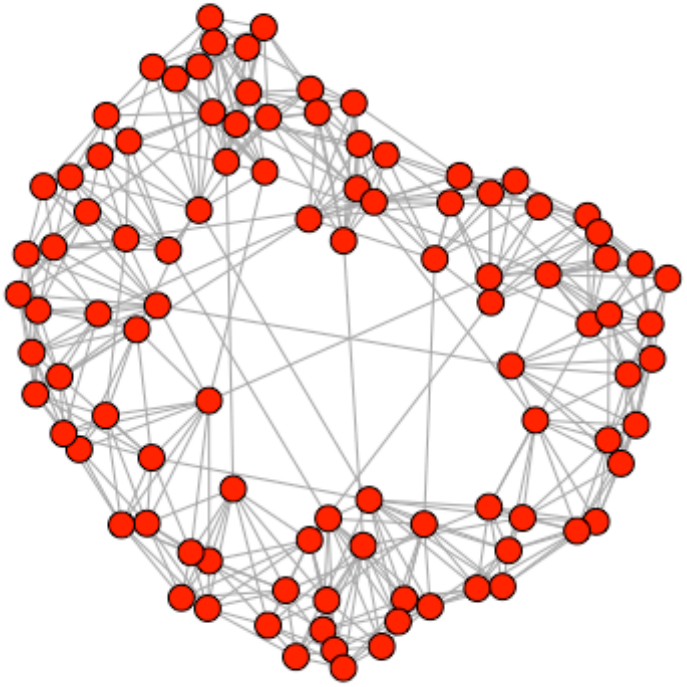
Connections between elements:

- One way?
- Two way?
- Varying strength?

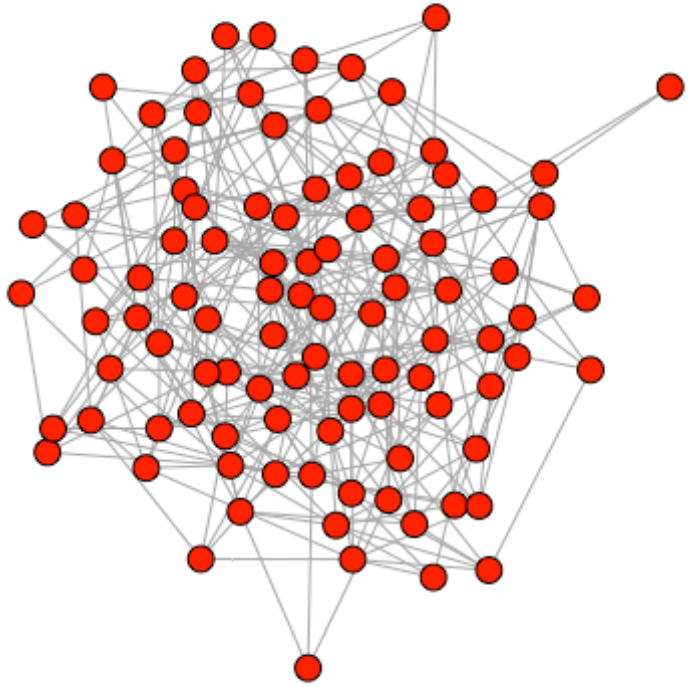
How to think about a network (a)

What is different about these networks?

(a)



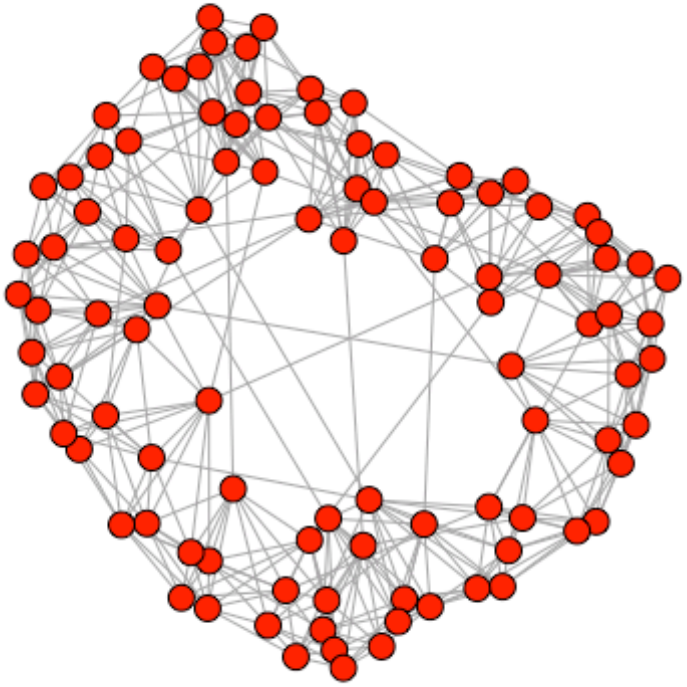
(b)



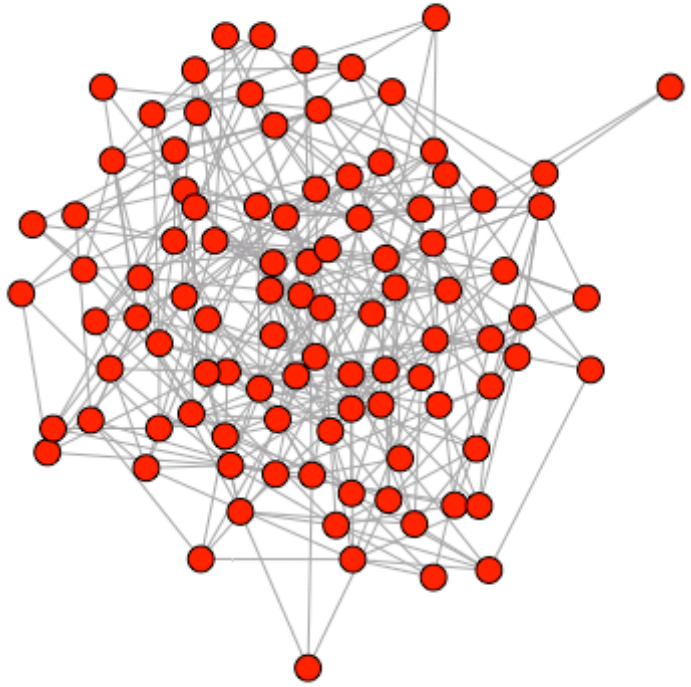
How to think about a network (b)

Which nodes are the most important?

(a)



(b)



A simple social network

Some of John's research collaborators:

- John collaborates with: Ana-Maria, RJ, Dilpreet
- + Ana-Maria also collaborates with: Matteo, RJ, Sue
- + Sue also collaborates with RJ.

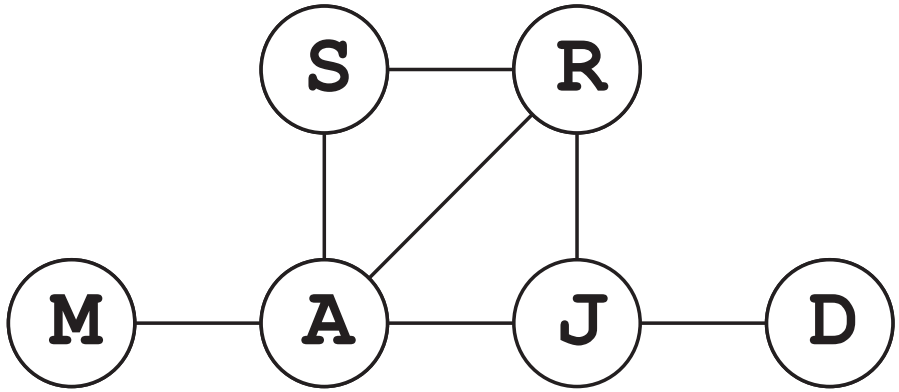
Questions to answer:

- Describe the network.
- Who is the most important/central/influential person in this network?

Network graph

Drawing the network graph:

- Vertices (nodes) indicate each person.
- Edges (lines/arcs) show that there is a connection.
- The graph below is one of many possible layouts.
- Assume relations are two-way.



Terminology: edges and vertices

- Vertices (or nodes) typically represent the entities in the network
- Edges (or arcs) represent connections between these entities
- Edges may be undirected, e.g. Friend A <==> Friend B or directed, e.g. Parent ==> Child etc.
- Edges may be weighted: to indicate the strength of a relationship or bond etc.
- See Newman (following) for vertex and edge names in some specific networks.

Terminology: edges and vertices

From Networks (Newman): An Introduction

Table 6.1: Vertices and edges in networks. Some examples of vertices and edges in particular networks.

Network	Vertex	Edge
Internet	Computer or router	Cable or wireless data connection
World Wide Web	Web page	Hyperlink
Citation network	Article, patent, or legal case	Citation
Power grid	Generating station or substation	Transmission line
Friendship network	Person	Friendship
Metabolic network	Metabolite	Metabolic reaction
Neural network	Neuron	Synapse
Food web	Species	Predation

Network structures

Particular link sequences have formal descriptions

- *Walk*: a sequence of links
- *Path*: a walk with no repeated vertices
- *Cycle*: a walk that begins and ends at the same vertex
- *Geodesic*: the shortest path between two vertices
- *Length*: the number of links in a walk or path
- *Connected*: there is a path between each pair of vertices
- *Directed graphs*: all definitions above apply but travel on each edge is permitted in one direction only.

Network structures cont…

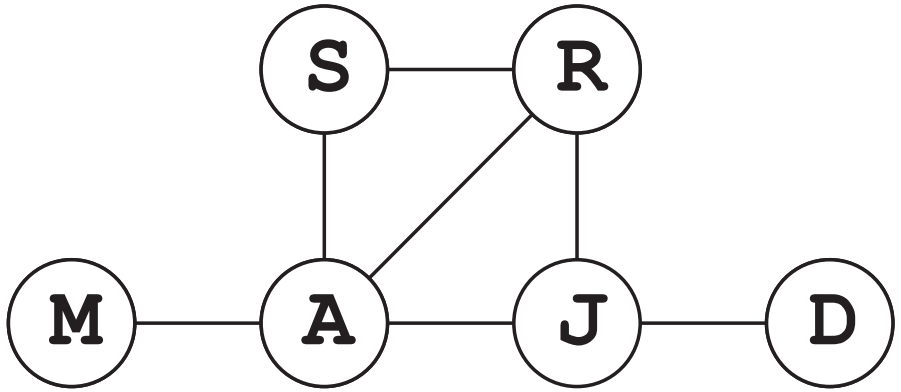
Particular link sequences have formal descriptions

- *Loop*: an edge from a vertex to itself
- *Complete*: a graph were every vertex is joined to every other vertex
- *Subgraph*: a subset of a graph
- *Clique*: a subgraph that is complete (every vertex joined to every other vertex)
- *Simple*: a graph with no loops or multi-edges (more than one edge between same pair of vertices) can be connected or disconnected

Network structures

Example from the research collaborators network:

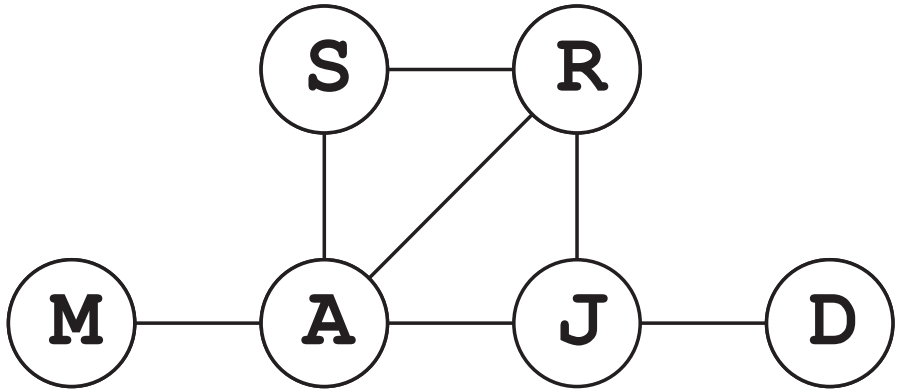
- *Walk*: (M, A, J, R, J, …)
- *Path*: (M, A, J, R)
- *Cycle*: (A, S, R, A)
- *Geodesic*: Geodesic(R, D) = (R, J, D)
- *Length*: dist(M,D) =3
- *Connected*: yes
- *Clique*: A, S, R
- *Simple*: yes.



Adjacency matrix

Summarizes the network by indicating the connections between individuals:

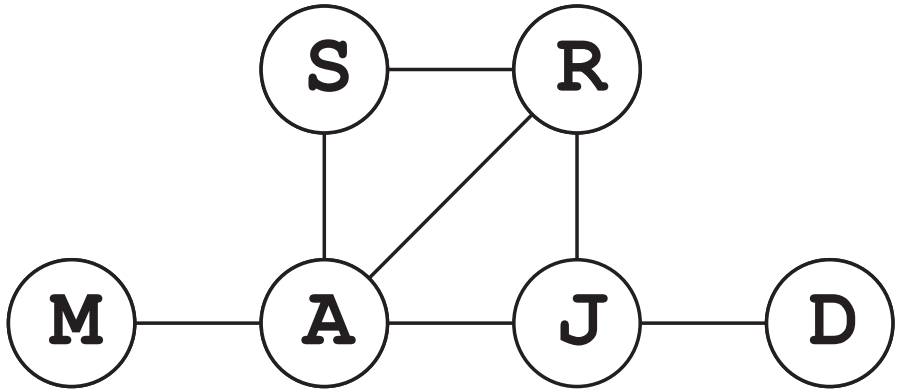
	J	A	S	D	R	M
J	0	1	0	1	1	0
A	1	0	1	0	1	1
S	0	1	0	0	1	0
D	1	0	0	0	0	0
R	1	1	1	0	0	0
M	0	1	0	0	0	0



Degree of a vertex

Arguably the single most important measure of a vertex's significance in a network:

- *Degree*: the number of edges connected to a vertex; the size of the vertex's *neighbourhood*.
- For directed graphs this is adapted to *in-degree* and *out-degree.*
- Example: $d_R = 3$.



Analysing the network (overall)

Statistics of the network as a whole include:

- *Diameter*: the longest geodesic between any two vertices.
- *Average path length*: average distance (geodesic) between any two vertices over the whole network.
- *Degree distribution*: the probability distribution describing the magnitude of vertices in the network.

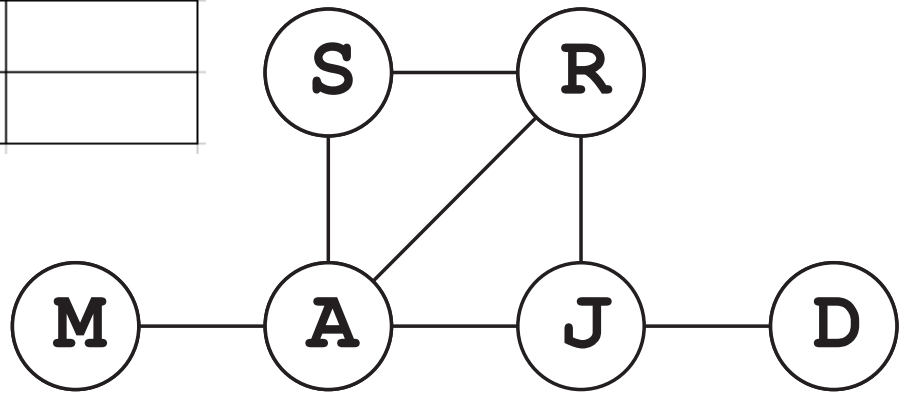
Analysis of these and other factors determine how *connected*, *robust* or *fragile* etc. a network is.

Analysing the network

Calculate degree distribution and distance matrices for the research collaborators network:

	J	A	S	D	R	M
J						
A						
S						
D						
R						
M	Distance Matrix					

Degree	N



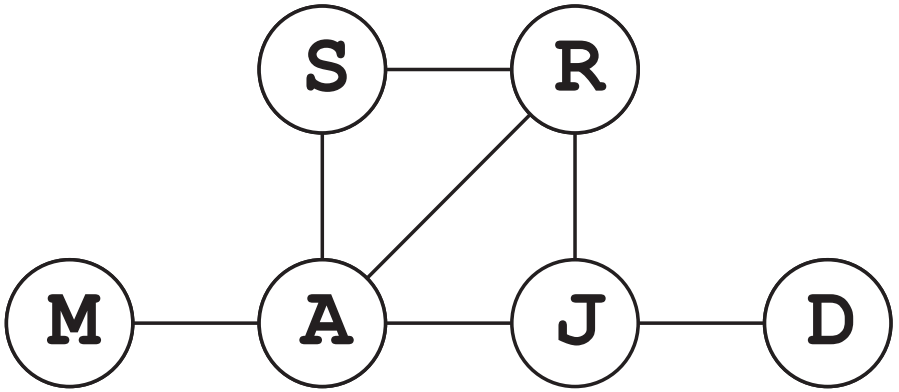
Analysing the network

Example from the research collaborators network:

- *Diameter*: = *max(dist(u,v)) = 3*
- *Degree distribution*: ⟶
- *Average path length*: *1.667*

Degree	N
0	0
1	2
2	1
3	2
4	1

	J	A	S	D	R	M
J		1	2	1	1	2
A			1	2	1	1
S				3	1	2
D					2	3
R						2
M	Distance Matrix					



Analysing the network

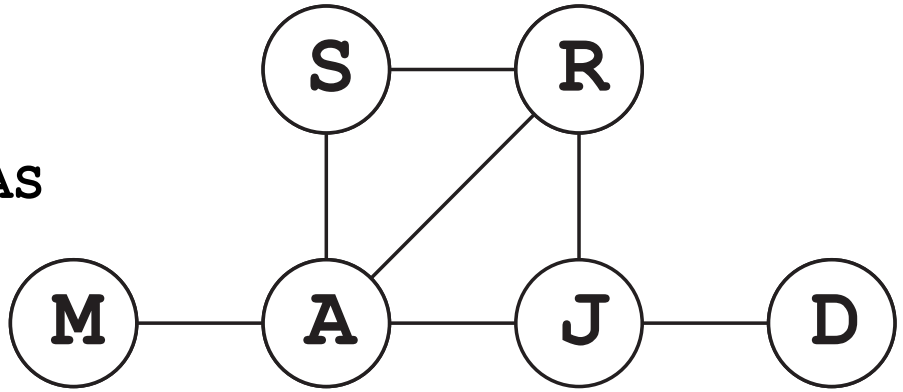
- *Density*: is the proportion of edges in a graph, relative to the maximum number possible.
- $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2}$
- where $|E_g|$ is number of edges, $|V_g|$ is number of vertices
- *Clustering coefficient*: is the proportion of triangles relative to the number of connected triples.
- $clt(g) = \frac{3\tau_{\Delta}(g)}{\tau 3(g)}$
- where $3\tau_{\Delta}(g)$ number of triangles, $\tau 3(g)$ is number of triples

Analysing the network

For the research collaborators:

- *Density*: $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2} = \frac{7}{(6\times5)/2} = 0.467$
- *Clustering coefficient*: $clt(g) = \frac{3\tau_{\Delta}(g)}{\tau 3(g)} = \frac{3\times2}{13} = 0.462$

(triples)
(M) MAS,MAJ,MAR,
(D) DJR,DJA,
(Square) ASR,SRJ,RJA,JAS
(Diag) ARJ,ARS,RAJ,RAS



Vertex importance

The importance of a vertex is based on two factors: Number of connections with other vertices

- *Degree*

Centrality of vertex within the network (strategic power to control information)

- *Betweenness*
- *Closeness*
- *Eigenvector*

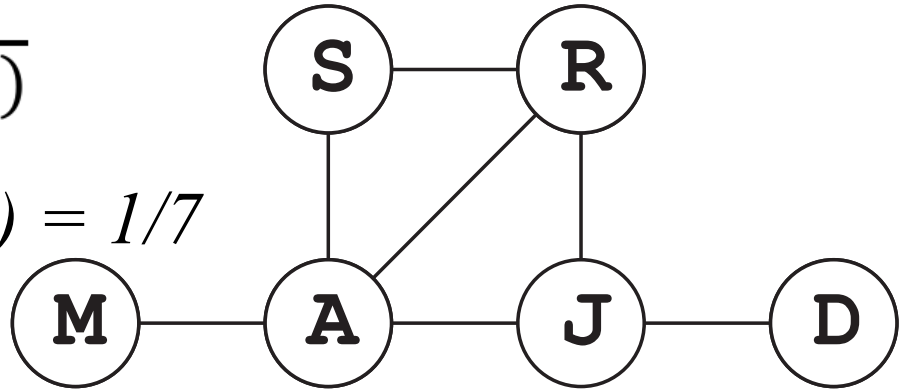
Closeness centrality

For this measure, a vertex is ‘close’ if there is a small total distance between it and all the other vertices in the network.

- Closeness centrality is the inverse of total distance between a vertex and the others.

$$c_{Cl}(v) = \frac{1}{\sum_{u \in V} dist(u, v)}$$

- $c_{Cl}(J) = 1/(1+1+2+1+2) = 1/7$

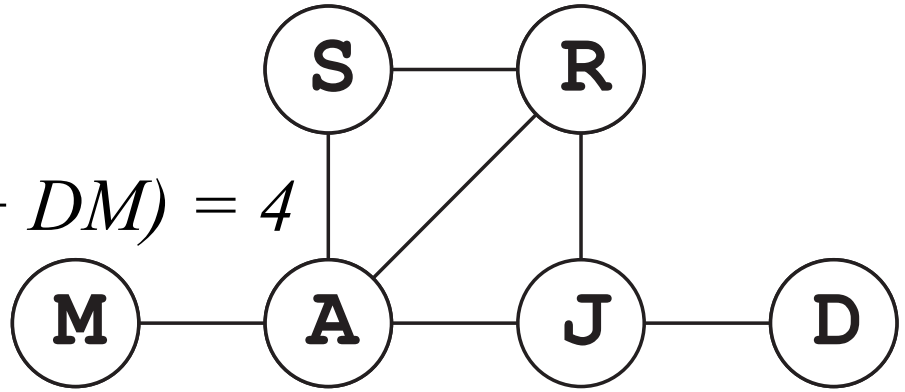


Betweenness centrality

This measure indicates the degree to which the vertex is 'between' other vertices.

- Betweenness centrality sums the number of shortest paths between s and t $\sigma(s,t)$ through vertex v (proportionally if more than one shortest path exists*).

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$



- $c_B(J) = (DR+DA+DS^*+ DM) = 4$
- $c_B(M) = 0$

Eigenvector centrality

Too difficult to calculate by hand but included for completeness…

- The basic idea is that it gives a higher weight to vertices with neighbours that are more central in the graph. (for example Google PageRank)

Which centrality measure?

In his blog, A Crazy Belief: Predicting Outcomes from Network Graphs, Fredembach discusses the reasons why the Medici family became the most prominent among the noble families of renaissance Florence.

Ideas of graph centrality are key to this. Essentially the Medici were the best-connected!

(inactive) http://blogs.teradata.com/international/a-crazy-belief-predicting-outcomes-from-network-graphs/

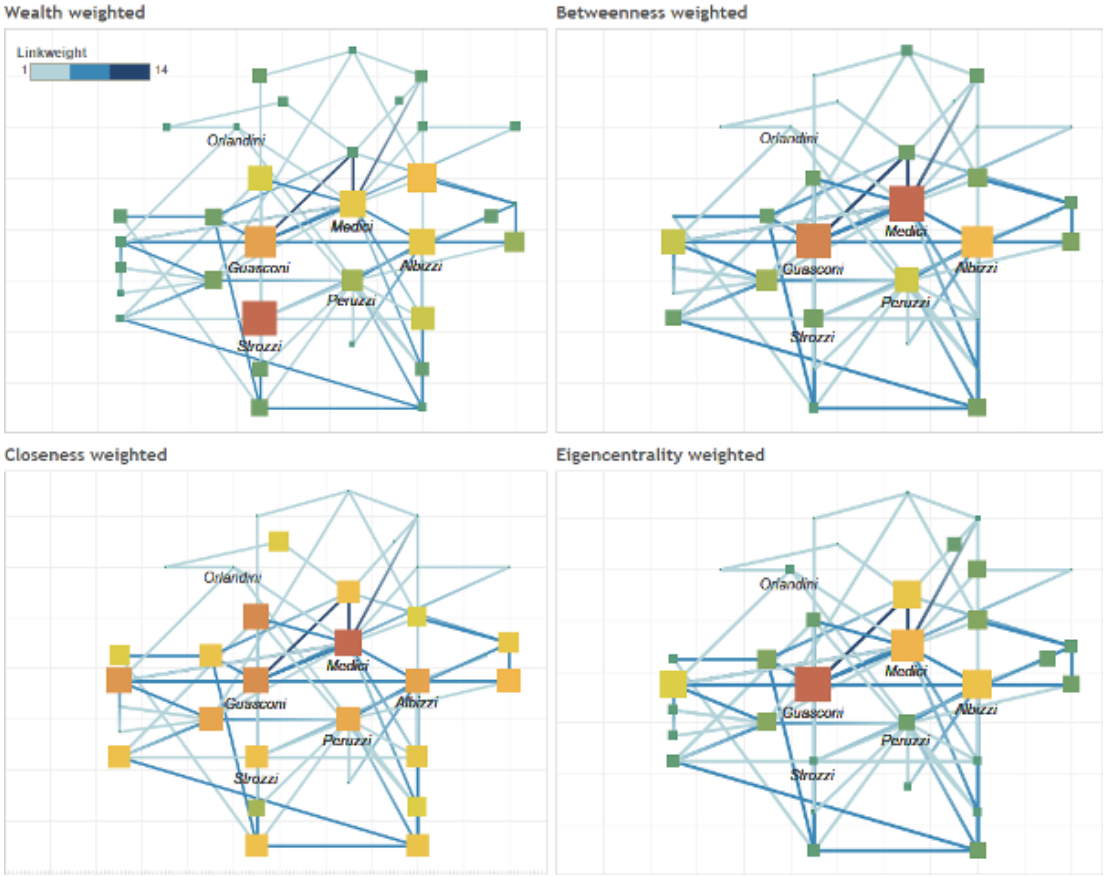
Which centrality measure?

Adapted from: Fredembach, C., A Crazy Belief: Predicting Outcomes from Network Graphs

- Betweeness centrality: measures the hub potential of a node. High BC nodes act as hubs/relays/bridges.
- Closeness centrality: measures how well a node is connected locally. High CC nodes are strong local influencers.
- Eignencentrality: weights a node according to the quality of its connections. Nodes connected to important nodes are ranked higher.

(inactive) http://blogs.teradata.com/international/a-crazy-belief-predicting-outcomes-from-network-graphs/

Which centrality measure?

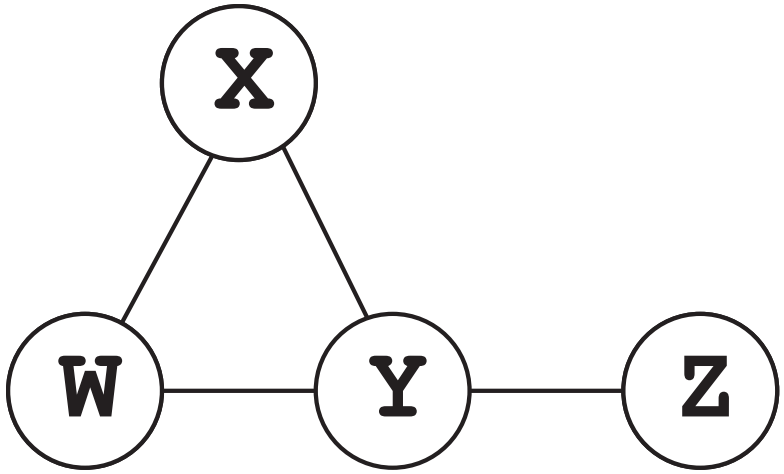


(inactive) http://blogs.teradata.com/international/a-crazy-belief-predicting-outcomes-from-network-graphs/

Class example

For the graph below calculate the following:

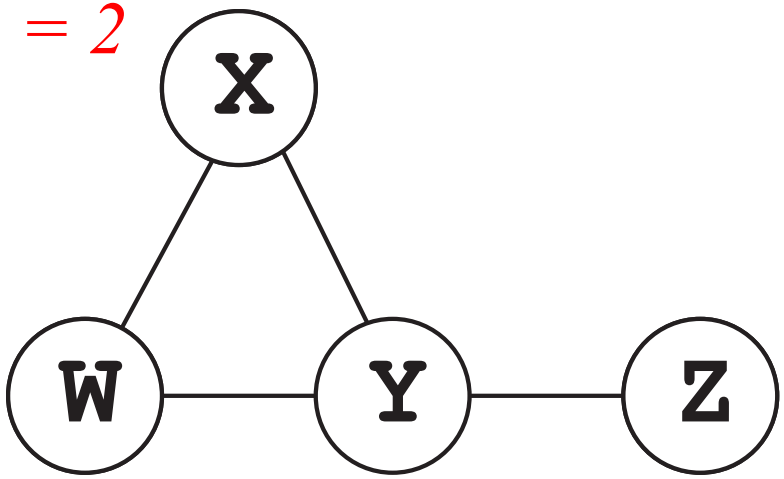
- *Average path length*
- *Diameter*
- d_Y
- *Degree distribution*
- $c_B(Y)$
- $c_{Cl}(Y)$
- *Cliques*



Class example – answers

For the graph below calculate the following:

- *Average path length ave(1, 1, 2, 1, 2, 1) = 1.333.*
- *Diameter for example d(W,Z) = 2*
- $d_Y = 3$
- *Degree distribution 3, 2, 2, 1*
- $c_B(Y)$ *W-Z, X-Z = 2*
- $c_{Cl}(Y)$ *1/(1+1+1) = 0.333.*
- *Cliques 2:W-X, W-Y, X-Y, Y-Z, 3:W-X-Y*



Analysing graphs using R

We will use the igraph (and igraphdata) package.

```
> install.packages(c("igraph", "igraphdata"))
> library(igraph)
> library(igraphdata)
```

There is a website devoted to igraph where you can download the documentation or use online reference:

https://igraph.org/r/

Alternatively search Stack Overflow…

Creating a graph

One way: create data frame and convert to graph object.

```
> graphdata <- data.frame(
  from = c("J", "J", "J", "A", "A", "A", "S"),
  to = c("D", "R", "A", "S", "R", "M", "R"),
  weight = c(1, 1, 1, 1, 1, 1, 1))
```

Then convert to graph object

```
> g <- graph.data.frame(graphdata, directed=FALSE)
```

Data frame

```
> graphdata
  from to weight
1    J  D      1
2    J  R      1
3    J  A      1
4    A  S      1
5    A  R      1
6    A  M      1
7    S  R      1
```

Graph summary

Graph object:

```
> g
IGRAPH UNW- 6 7 --
+ attr: name (v/c), weight (e/n)
+ edges (vertex names):
[1] J--D J--R J--A A--S A--R A--M S--R
```

Graph summary

Vertex and edge sequence:

```
> V(g)
+ 6/6 vertices, named:
[1] J A S D R M

> E(g)
+ 7/7 edges (vertex names):
[1] J--D J--R J--A A--S A--R A--M S--R
```

Graph summary

Count vertices, edges, test if simple graph:

```
> vcount(g)
[1] 6

> ecount(g)
[1] 7

> is.simple(g)
[1] TRUE
```

Graph summary

Diameter, average path length, clique size:

```
> diameter(g)
[1] 3

> average.path.length(g)
[1] 1.666667

> table(sapply(cliques(g),length))
1 2 3
6 7 2
```

Graph summary

Density and clustering coefficient:

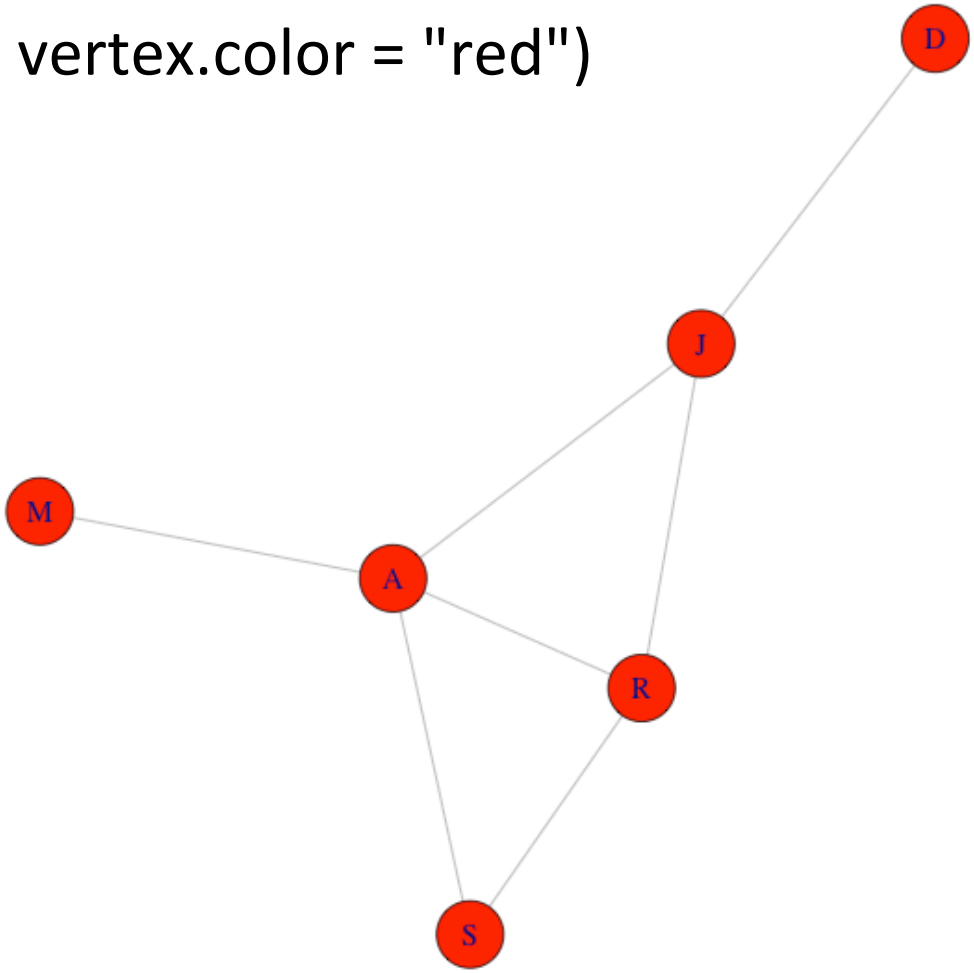
```
> graph.density(g)
[1] 0.467

> transitivity(g)
[1] 0.462
```

Adjacency matrix

```
> get.adjacency(g)
6 x 6 sparse Matrix of class "dgCMatrix"
  J A S D R M
J . 1 . 1 1 .
A 1 . 1 . 1 1
S . 1 . . 1 .
D 1 . . . . .
R 1 1 1 . . .
M . 1 . . . .
```

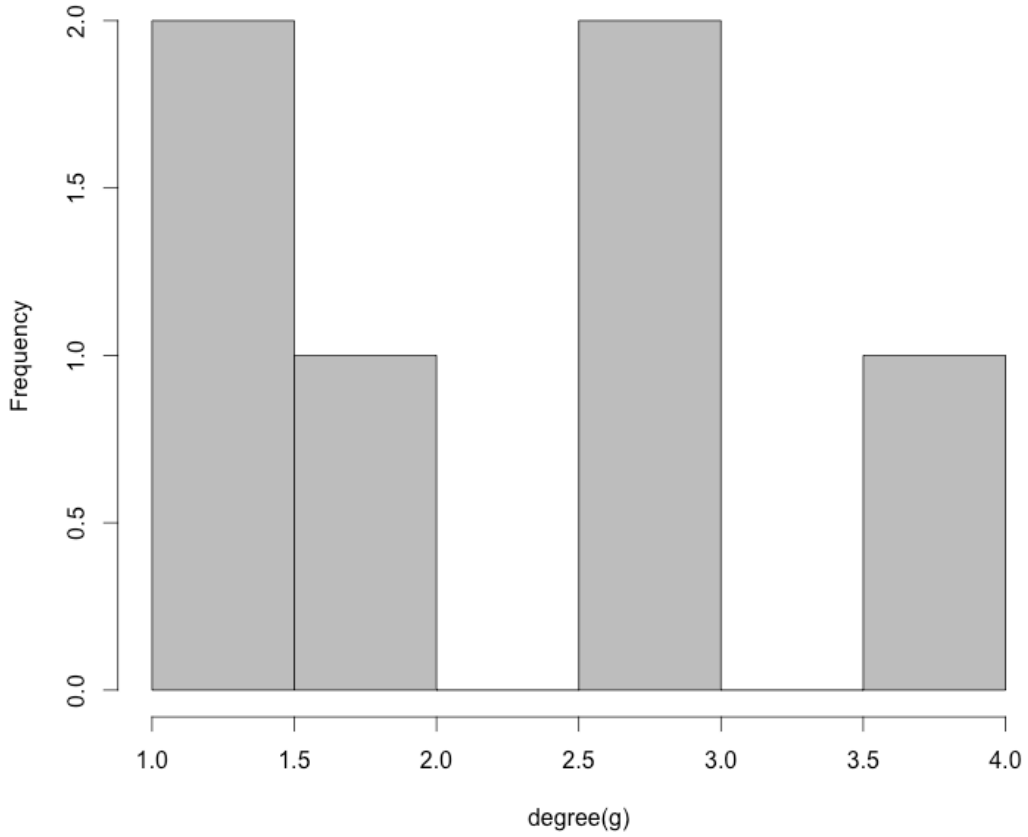
Basic plot

```
> plot(g, vertex.color = "red")
```



Degree distribution

```
> hist(degree(g), breaks = 5, col = "grey")
```



Vertex summary

Degree, betweenness and closeness centrality

> degree(g)

```
J A S D R M
3 4 2 1 3 1
```

> betweenness(g)

```
J A S D R M
4 5 0 0 1 0
```

> format(closeness(g), digits = 2)

```
      J       A       S       D       R       M
"0.143" "0.167" "0.111" "0.091" "0.143" "0.100"
```

Vertex summary cont…

Eigenvector centrality (first create the eigenvector and then format the output).

```
> e = evcent(g)
> format(e$vector, digits = 2)
     J      A      S      D      R      M
"0.80" "1.00" "0.69" "0.29" "0.90" "0.36"
```

What does this all mean?

Summary: who is the most important person in the network?

	J	A	S	D	R	M
Degree	3.00	4.00	2.00	1.00	3.00	1.00
Betweenness	4.00	5.00	0.00	0.00	1.00	0.00
Closeness	0.14	0.17	0.11	0.09	0.14	0.10
Eigenvector	0.80	1.00	0.69	0.29	0.90	0.36

. . .

Some extra R

- Alternate means of entering graph data
- Some important network topologies demonstrated using inbuilt graph models

Alternate means of entering data (a)

Simple graphs can be created using an edge list:

```
> g <- graph.formula(J-D, J-R, J-A, A-S, A-R, A-M, S-R)
```

This can be adapted to create directed graphs, for example:

```
> g <- graph.formula(J-+D, J++R, J+-A, A-+S, A++R, A-+M,
  S-+R)
```

Alternate means of entering data (b)

More complex graphs can be defined by adjacency matrix as csv file:

> gdata = read.csv(filename, header = TRUE, row.names = 1, check.names = FALSE)

> mdata = as.matrix(gdata) *# convert to matrix*

> g = graph.adjacency(mdata, mode = "undirected", weighted = NULL) *# create graph*

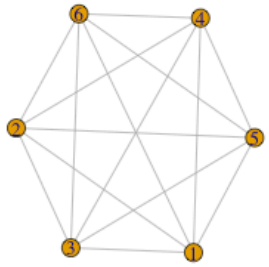
- Rows/columns may not be symmetrical.
- Weights can vary.

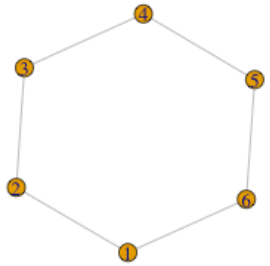
	J	A	S	D	R	M
J	0	1	0	1	1	0
A	1	0	1	0	1	1
S	0	1	0	0	1	0
D	1	0	0	0	0	0
R	1	1	1	0	0	0
M	0	1	0	0	0	0

Four important network topologies

Using R:

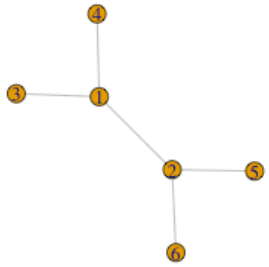
Complete





Ring

Tree



Star

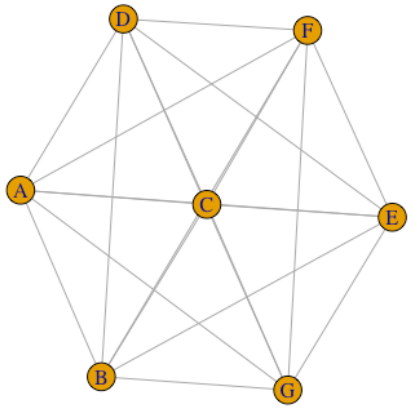
Four important network topologies

```
> # R code adapted from Kolaczyk
> g.full <- graph.full(6)
> g.ring <- graph.ring(6)
> g.tree <- graph.tree(6, children=3, mode="undirected")
> g.star <- graph.star(6, mode="undirected")
> par(mfrow=c(2, 2))
> plot(g.full)
> plot(g.ring)
> plot(g.tree)
> plot(g.star)
```

Creating a clique directly

You can create a clique (complete graph) directly by specifying the vertices as follows:

```
> # To make a complete graph from a set of vertices
> # https://igraph.org/r/doc/graph_from_literal.html
> gg = graph_from_literal(A:B:C:D:E:F:G --
  A:B:C:D:E:F:G)
> plot(gg)
```



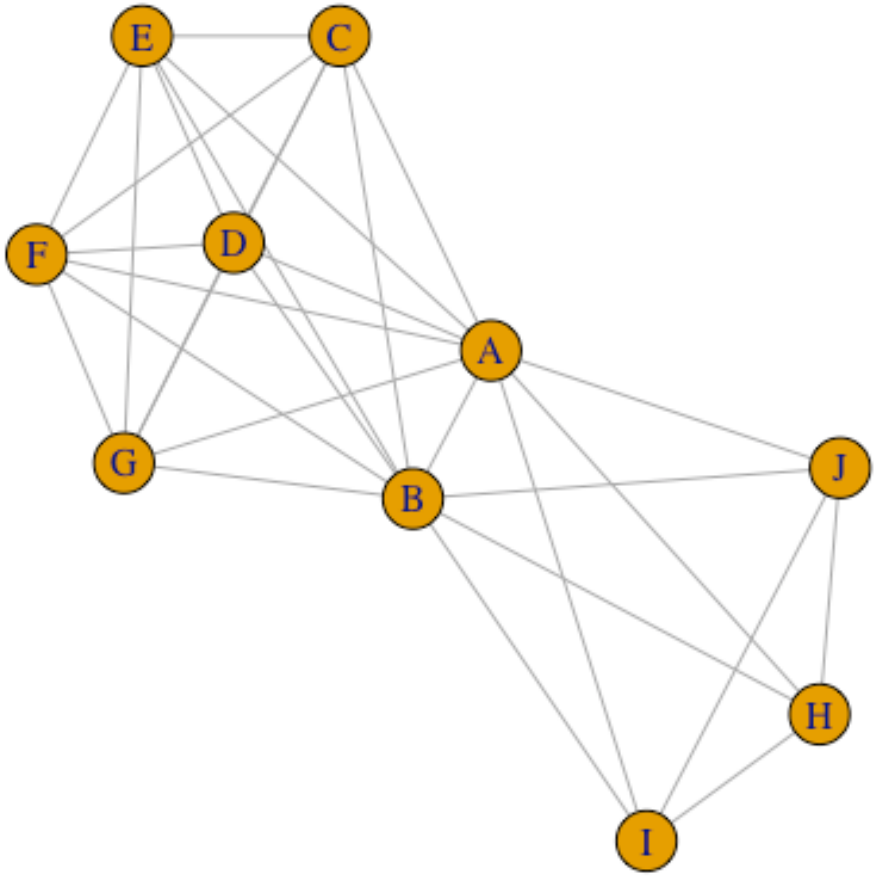
Merging two graphs

You can also merge graphs into a new graph:

```
> gg = graph_from_literal(A:B:C:D:E:F:G --
  A:B:C:D:E:F:G)
> # Make a second graph with some common vertices
> hh = graph_from_literal(A:B:H:I:J -- A:B:H:I:J)
> # now make a union
> ii = (gg %u% hh)
> # see https://igraph.org/r/doc/union.igraph.html
> plot(ii)
```

Merging two graphs

> plot(ii)



A bigger graph – karate club

In the 1970s the anthropologist W. W. Zachary studied the social network formed by members of a karate club.

- The club went through an interesting transformation which could potentially have been predicted from its network structure.
- W. Zachary, An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33 (4), 452–473 (1977)

Karate club – what happened

At the beginning of the study there was an incipient conflict between the club president, John A., and Mr. Hi over the price of karate lessons. Mr. Hi, who wished to raise prices, claimed the authority to set his own lesson fees, since he was the instructor. John A., who wished to stabilize prices, claimed the authority to set the lesson fees since he was the club's chief administrator.

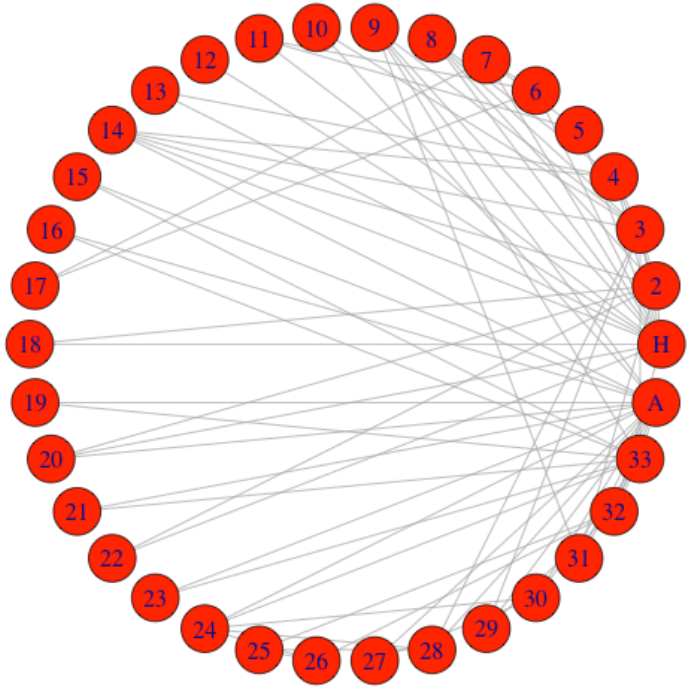
W. Zachary, An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33 (4), 452–473 (1977)

Karate club – looking at the data

```
> library(igraph)
> library(igraphdata)
> data(karate)
> diameter(karate)
[1] 13
> average.path.length(karate)
[1] 2.4082
> V(karate)
+ 34/34 vertices, named: ...
> E(karate)
+ 78/78 edges (vertex names): ...
```

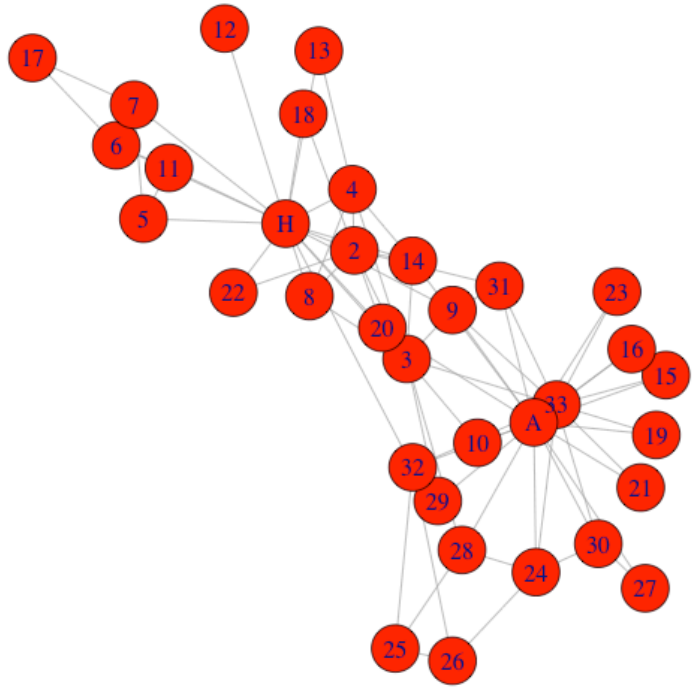
Karate club – circle plot

```
> plot(karate, layout = layout.circle, vertex.color = "red")
```



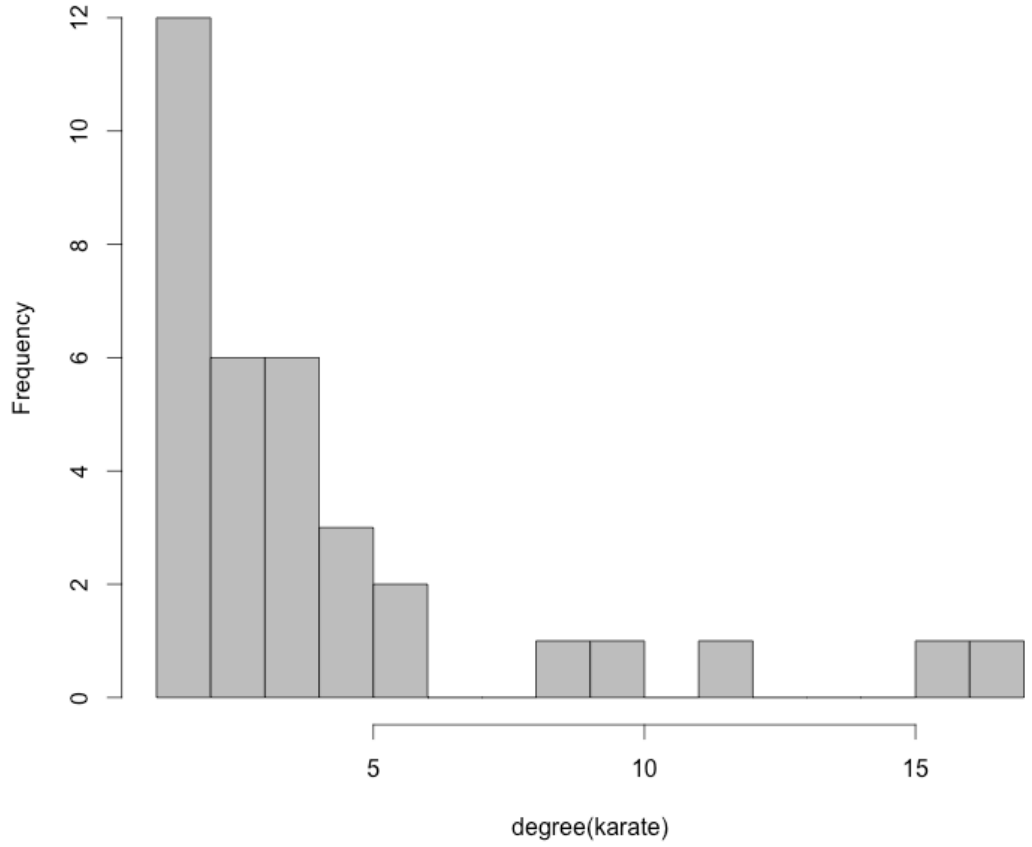
Karate club – force-directed plot

```
> plot(karate, layout = layout.fruchterman.reingold)
```



Karate club – degree distribution

```
> hist(degree(karate), breaks = 18, col = "grey")
```



Karate club – vertex statistics

Actor	Degree	Closeness	Betweenness
Mr Hi	16	0.0077	250.1
Actor 2	9	0.0061	33.8
Actor 3	10	0.0060	36.6
Actor 4	6	0.0053	1.3
Actor 5	3	0.0046	0.5
Actor 6	4	0.0046	15.5
Actor 7	4	0.0047	15.5
Actor 8	4	0.0055	0.0
Actor 9	5	0.0060	13.1
Actor 10	2	0.0058	7.3
Actor 11	3	0.0053	0.5
Actor 12	1	0.0044	0.0
Actor 13	2	0.0062	0.0
Actor 14	5	0.0058	1.2
Actor 15	2	0.0052	0.0
Actor 16	2	0.0042	0.0
Actor 17	2	0.0033	0.0

Actor	Degree	Closeness	Betweenness
Actor 18	2	0.0058	16.1
Actor 19	2	0.0057	3.0
Actor 20	3	0.0075	127.1
Actor 21	2	0.0062	0.0
Actor 22	2	0.0053	0.0
Actor 23	2	0.0048	0.0
Actor 24	5	0.0042	1.0
Actor 25	3	0.0048	33.8
Actor 26	3	0.0037	0.5
Actor 27	2	0.0051	0.0
Actor 28	4	0.0047	6.5
Actor 29	3	0.0061	10.1
Actor 30	4	0.0053	0.0
Actor 31	4	0.0053	3.0
Actor 32	6	0.0063	66.3
Actor 33	12	0.0061	38.1
John A	17	0.0076	209.5

What can we conclude about the structure of this network? Who are the most important people?

Karate club – cliques

```
> table(sapply(cliques(karate), length)) # Kolaczyk p.52
1  2  3  4  5
34 78 45 11  2

> cliques(karate)[sapply(cliques(karate), length) == 5]
[[1]]
+ 5/34 vertices, named:
[1] Mr Hi   Actor 2 Actor 3 Actor 4 Actor 8

[[2]]
+ 5/34 vertices, named:
[1] Mr Hi    Actor 2  Actor 3  Actor 4  Actor 14
```

Karate club – what happened

As time passed the entire club became divided over this issue, and the conflict became translated into ideological terms by most club members. The supporters of Mr. Hi saw him as a fatherly figure who was their spiritual and physical mentor, and who was only trying to meet his own physical needs after seeing to theirs. The supporters of John A. and the other officers saw Mr. Hi as a paid employee who was trying to coerce his way into a higher salary. After a series of increasingly sharp factional confrontations over the price of lessons, the officers, led by John A., fired Mr. Hi for attempting to raise lesson prices unilaterally. The supporters of Mr. Hi retaliated by resigning and forming a new organization headed by Mr. Hi, thus completing the fission of the club.

W. Zachary, An information flow model for conflict and fission in small groups. J. Anthropol. Res. 33 (4), 452–473 (1977)

Karate club – R code

```
> library(igraph)
> library(igraphdata)
> data(karate)
> plot(karate, layout = layout.circle, vertex.color = "red")
> plot(karate, layout = layout.fruchterman.reingold,
  vertex.color = "red")
> hist(degree(karate), breaks = 18, col = "grey")
> d = degree(karate)
> c = format(closeness(karate), digits = 2)
> b = format(betweenness(karate), digits = 2)
> ksum = as.data.frame(cbind(d, c, b))
> write.csv(ksum, "karatesum.csv")
```

A larger network – UKfaculty

Another igraphdata set is the friendship network from a UK university faculty: UKfaculty

Who are the key players in this network?

UKfaculty – overview

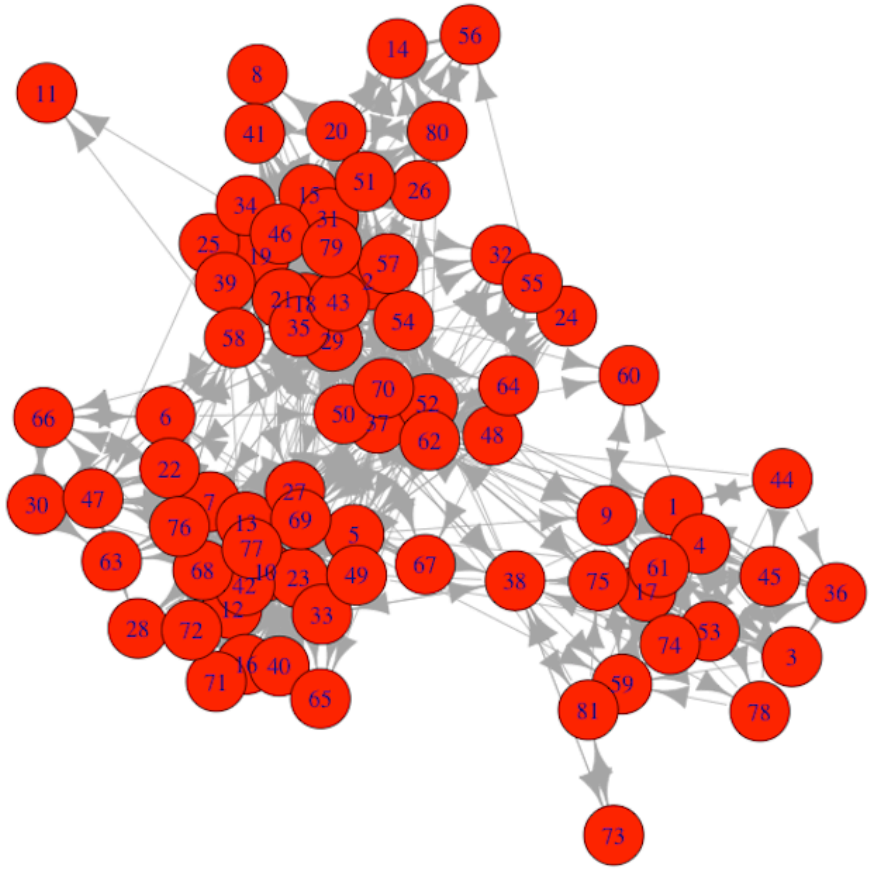
> data(UKfaculty)

> UKfaculty

```
IGRAPH D-W- 81 817 --
+ attr: Type (g/c), Date (g/c), Citation (g/c), Author
| (g/c), Group (v/n), weight (e/n)
+ edges:
 [1] 57->52 76->42 12->69 43->34 28->47 58->51  7->29 40->71
 [9]  5->37 48->55  6->58 21-> 8 28->69 43->21 67->58 65->42
[17]  5->67 52->75 37->64  4->36 12->49 19->46 37-> 9 74->36
[25] 62-> 1 15-> 2 72->49 46->62  2->29 40->12 22->29 71->69
[33]  4-> 3 37->69  5-> 6 77->13 23->49 52->35 20->14 62->70
[41] 34->35 76->72  7->42 37->42 51->80 38->45 62->64 36->53
[49] 62->77 17->61  7->68 46->29 44->53 18->58 12->16 72->42
+ ... omitted several edges
```

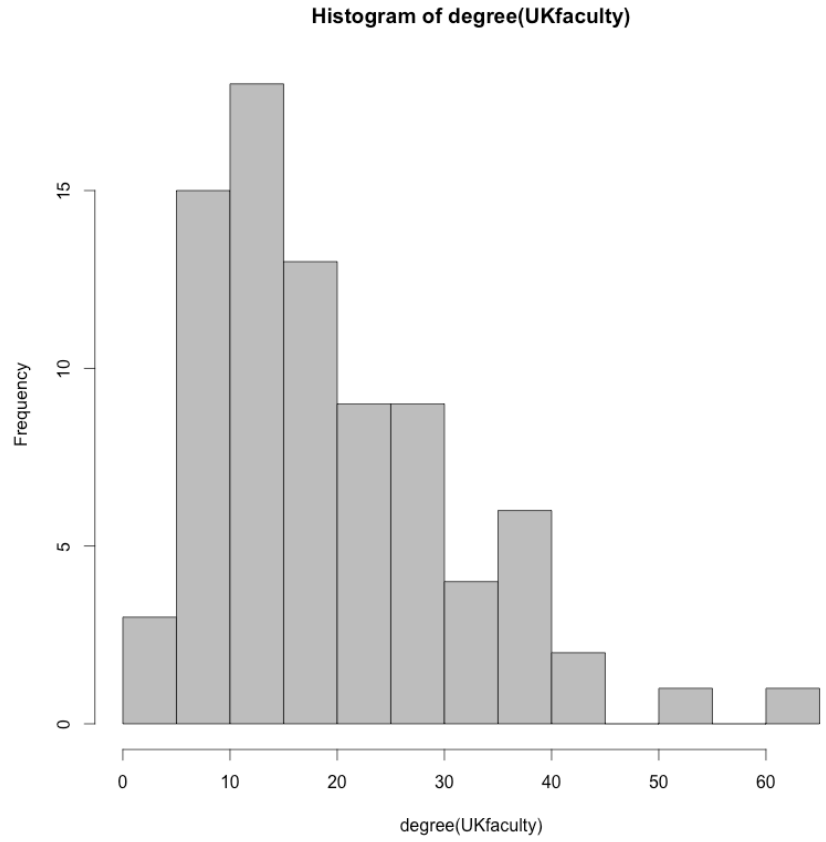
UKfaculty – plot

> plot(UKfaculty, vertex.color = "red")



UKfaculty – degree dist…

```
> hist(degree(UKfaculty), breaks = 14, col = "grey")
```



UKfaculty – data summary

```
> deg = as.table(degree(UKfaculty))
> bet = as.table(betweenness(UKfaculty))
> clo = as.table(closeness(UKfaculty))
> eig = as.table(evcent(UKfaculty)$vector)
> summ = as.data.frame(rbind(deg,bet,clo,eig))
> summ = t(summ)
> summ = as.data.frame(summ)
> head(summ)
  index deg   bet     clo     eig
A     1  15 347.1 0.00289 0.01137
B     2  36 711.3 0.00368 0.45216 ...
```

Notes for reading only.

UKfaculty – sorted by centrality

```
> hdeg = summ[order(-deg),]
> head(hdeg)
   index deg  bet     clo   eig
C1    29  62  684 0.00452 1.000
K1    37  54 1223 0.00450 0.317
Y2    77  44  350 0.00304 0.194
J2    62  43  966 0.00503 0.308
Z1    52  39  382 0.00427 0.190
E      5  38  727 0.00450 0.126
>
> hbet = summ[order(-bet),]
> head(hbet)
   index deg  bet     clo   eig
K1    37  54 1223 0.00450 0.317
J2    62  43  966 0.00503 0.308
E      5  38  727 0.00450 0.126
B      2  36  711 0.00368 0.452
C1    29  62  684 0.00452 1.000
L1    38  21  629 0.00398 0.057
```

```
> hclo = summ[order(-clo),]
> head(hclo)
   index deg  bet     clo   eig
J2    62  43  966 0.00503 0.308
C1    29  62  684 0.00452 1.000
E      5  38  727 0.00450 0.126
K1    37  54 1223 0.00450 0.317
Z1    52  39  382 0.00427 0.190
L1    38  21  629 0.00398 0.057
>
> heig = summ[order(-eig),]
> head(heig)
   index deg   bet     clo   eig
C1    29  62 684.2 0.00452 1.000
E1    31  35 109.8 0.00313 0.822
U     21  38 143.3 0.00254 0.740
A3    79  22  70.4 0.00281 0.590
I1    35  28 263.6 0.00287 0.587
S     19  24  16.5 0.00257 0.515
```

UKfaculty – cliques

```
> # model assumes graph undirected
> table(sapply(cliques(UKfaculty), length))

   1    2    3    4    5    6    7    8    9
  81  577 1626 2660 2732 1742  668  142   13
```

UKfaculty – largest cliques

```
> # looking at the largest cliques (abbreviate output)
> cliques(UKfaculty)[sapply(cliques(UKfaculty), length)
  == 9]
```

```
[1]  2 15 18 29 35 37 43 57 79
[1]  2 15 18 29 35 43 46 57 79
[1]  2 15 21 29 31 35 37 43 79
[1]  2 15 21 29 31 35 43 46 79
[1]  2 15 29 31 35 37 43 57 79
[1]  2 15 29 31 35 43 46 57 79
[1]  5  7 10 13 23 27 68 69 77
[1]  5  7 10 23 27 40 68 69 77
[1]  5  7 10 23 27 42 68 69 77
[1]  5 10 13 23 27 49 68 69 77
[1]  5 10 23 27 42 49 68 69 77
[1] 10 13 23 27 33 49 68 69 77
[1] 10 23 27 33 42 49 68 69 77
```

UKfaculty

So – who are the key players in this network?

Summary

Network Analysis

- Introduction: types of networks; network structure.
- Network statistics; node importance measures.
- Using R for network analysis (igraph package).
- Examples

Not covered, for you to follow up (see references)

- *Better graphics (put more dimension on to a plot).*
- *Deeper analysis, clustering for example.*

More theoretical…

If you're keen to pursue Network science in greater depth:

- Read Kolaczyk and Cśardi Chapter 5.
- Use R to generate some random networks for analysis and discovery based on the famous models:
- Random Graph Model (Erdós and Rényi),
- Small-World (Watts and Strogatz),
- Preferential Attachment (Barabási and Albert).

.

Review questions: answers

1. A
2. B
3. D
4. C

References

Two great textbooks, both available for free from the Monash Library

- *Statistical Analysis of Network Data with R*, Kolaczyk, E. D., Csárdi, G. Springer 2014. Chapters 1 – 4 used as the basis for much of this lecture.
- *A User's Guide to Network Analysis in R*, Luke, D. A. Springer 2015.

More resources

More theoretical resources:

- *Networks: An Introduction*, Newman, M., Oxford U. P. 2010 (full text available on line via library)
- The physics of networks, Newman, M., Physics Today, 2008. https://physicstoday.scitation.org/doi/10.1063/1.3027989

Great reference on graph design and layout

- Network visualization with R, PolNet 2018 Workshop

https://kateto.net/wp-content/uploads/2018/06/Polnet%202018%20R%20Network%20Visualization%20Workshop.pdf