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)
- Community detection
- Examples

Advertising: Deepneuron

https://fb.me/e/1cT94iYlR

Consultations on Zoom

Clayton consultations have commenced:

- Any student can attend any consultation.
- Schedule on Moodle, https://lms.monash.edu/
- Current days/times:
- Monday 9:30-10:30AM, 2:00-3:00PM, 6:00-7:00PM,
- Tuesday 9:00-10:00AM, 12:00PM-1:00PM,
- Wednesday 10:00AM-11:00, 11:00-12:00PM,
- Thursday 1:00PM-02:00PM, 6:00PM-7:00PM.
- Please check the schedule for any changes.

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
28/2/22	1	Intro to Data Science, review of basic statistics using R	...		
7/3/22	2	Exploring data using graphics in R	T1		
14/3/22	3	Data manipulation in R	T2	Released	
21/3/22	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
28/3/22	5	Network analysis	T4		
4/4/22	6	Regression modelling	T5		
11/4/22	7	Classification using decision trees	T6		
		Mid-semester Break		Submitted	
25/4/22	8	Naïve Bayes, evaluating classifiers	T7		Released
2/5/22	9	Ensemble methods, artificial neural networks	T8		
9/5/22	10	Clustering	T9		
16/5/22	11	Text analysis	T10		Submitted
23/5/22	12	Review of course, Exam preparation	T11		

Assignment 1

Assignment 1: Summary

FIT3152 Data analytics – 2022: Assignment 1

Your task	• Analyse the activity, language use and social interactions of an on-line community using metadata and linguistic summary from a real on-line forum and submit a report of your findings. • This is an individual assignment.
Value	• This assignment is worth **20%** of your total marks for the unit. • It has 30 marks in total.
Suggested Length	• 6 – 8 A4 pages (for your report) + extra pages as appendix (for your code) • Font size 11 or 12pt, single spacing
Due Date	**11.55pm Friday 22nd April 2022**
Submission	• PDF file only. Naming convention: *FirstnameSecondnameID.pdf* • Via Moodle Assignment Submission. • Turnitin will be used for similarity checking of all submissions.
Late Penalties	• 10% (3 mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Assignment 1: Instructions

Instructions

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.

There are two options for compiling your report:
(1) You can submit a single pdf with R code pasted in as machine-readable text as an appendix, or
(2) As an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and print off as a pdf and submit.

Regardless of which method you choose, you will submit a single pdf, and your R code will be machine readable text. We need to conform to this format as the university now requires all student submission to be processed by plagiarism detection software.

Submit your report as a single PDF with the file name ***FirstnameSecondnameID.pdf*** on Moodle.

Assignment 1: Software

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: Questions a & b

Questions

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:

(a) Analyse activity and language on the forum over time:

1. How active are participants over the longer term (that is, over months and/or years)? Are there periods where activity increases or decreases? Is there a trend over time? **(3 Marks)**
2. Looking at the linguistic variables, do the levels of these change over the duration of the forum? Is there a relationship between linguistic variables over the longer term? **(3 Marks)**

(b) Analyse the language used by threads:

We can think of threads as groups of participants posting on the same topic.

1. Using the relevant linguistic variables, is it possible to see whether or not particular threads are happier or more optimistic than other threads, or the forum in general, at different periods in time. **(3 Marks)**

Assignment 1: Question c

(c) Analyse social networks online:

We can think of authors posting to the same thread at similar times (for example during the same month) as having a connection to each other, forming a social network. This is called a two-mode network. When an author posts to more than one network during the same time period their social network extends to include authors from both networks, and so on. We will cover social network analysis in Lecture 5.

1. Create a non-trivial social network of all authors who are posting over a particular time period. For example, over one month. To create this, your social network should include at least 30 authors, some of whom will have posted to multiple (2 or more) threads during this period. Your social network should be connected, although some authors may be disconnected from the main group. Present your result as a network graph. **(3 Marks)**
2. Identify the most important author in the social network you created. Looking at the language they use, can you observe any difference between them and other members of their social network? **(3 Marks)**

Assignment 1: Overall considerations

(d) Overall considerations:

- The quality and clarity of your reasoning and assumptions. **(3 Marks)**
- The strength of support for your findings. **(3 Marks)**
- The quality of your writing in general and communication of results. **(3 Marks)**
- The quality of your graphics throughout, including at least one high-quality multivariate graphic. **(3 Marks)**
- The quality of your R coding. **(3 Marks)**

Assignment 1: Data generation

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: Data fields

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

Column	Brief Descriptor	Column	Brief Descriptor
ThreadID	Unique ID for each thread	we	"We, us, our" words
AuthorID	Unique ID for each author	you	"You" words
Date	Date	shehe	"She, her "him words
Time	Time	they	"They" words
WC	Word count of the text of the post	posemo	Expressing positive emotions
Analytic	Summary: Analytical thinking	negemo	Expressing negative emotions
Clout	Summary: Power, force, impact	anx	Indicating anxiety
Authentic	Summary: Authentic tone of voice	anger	Indicating anger
Tone	Summary: Emotional tone	sad	Indicating sadness
ppron	"I, we, you" words	focuspast	Expressing a focus on the past
i	"I, me, mine" words	focuspresent	Expressing a focus on the present
focusfuture	Expressing a focus on the future	focusfuture	Expressing a focus on the future

Assignment 1: Data extract

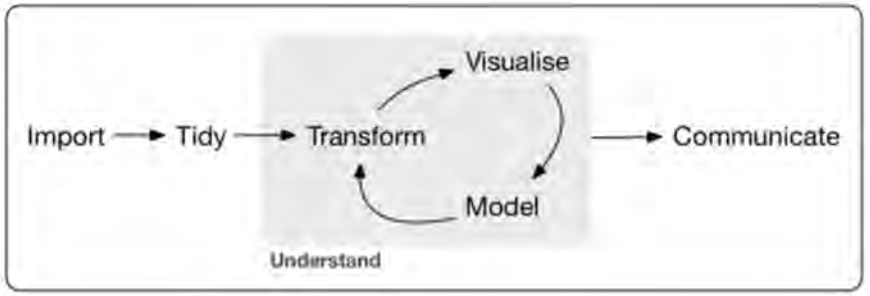
ThreadID	AuthorID	Date	Time	WC	Analytic	Clout	Authentic	Tone	ppron	i	we	you	shehe	they	...
144564	41084	9/8/04	4:46	134	55.23	69.94	63.91	68.05	7.46	2.99	2.24	1.49	0	0.75	...
404119	128515	21/7/07	22:27	12	1	79.76	74.76	25.77	33.33	8.33	0	0	0	25	...
395992	93243	19/6/07	1:02	28	13.85	76.25	1.06	99	7.14	3.57	0	3.57	0	0	...
405421	99958	24/7/07	1:40	16	84.57	89.42	35.37	1	6.25	0	0	6.25	0	0	...
662470	185647	5/12/09	16:05	37	32.06	79.13	21.26	75.85	18.92	8.11	0	0	5.41	5.41	...
420058	53655	13/9/07	22:59	17	26.21	3.89	99	1	11.76	5.88	0	0	0	5.88	...
13933	1740	9/3/02	2:01	61	22.35	37.15	72.51	25.77	11.48	6.56	1.64	0	0	3.28	...
245087	80190	9/11/05	15:06	94	82.45	66.48	44.79	25.77	4.26	2.13	1.06	0	0	1.06	...
442550	47686	6/12/07	5:06	80	61.95	54.96	59.88	96.76	7.5	5	0	1.25	0	1.25	...
352716	26979	5/1/07	21:33	10	8.19	84.14	1	25.77	0	0	0	0	0	0	...
463617	104430	29/2/08	8:02	249	98.57	78.92	15.3	83.06	3.61	0.8	1.61	0	0.8	0.4	...
363541	-1	15/2/07	11:30	26	53.63	87.57	38.39	99	11.54	3.85	0	7.69	0	0	...
258941	44297	1/1/06	13:47	59	94.34	91.23	10.76	6.73	8.47	1.69	1.69	5.08	0	0	...
765163	54960	17/12/10	21:06	139	26.01	58.53	13.52	66.61	7.91	1.44	0.72	2.88	0	2.88	...
263152	79878	18/1/06	7:34	114	48.42	73.03	9.58	1	10.53	4.39	0	2.63	0	3.51	...
228773	166362	6/9/09	4:52	14	13.85	98.33	89.63	25.77	14.29	0	0	14.29	0	0	...
254482	83344	6/1/06	0:17	107	80.6	77.26	24.3	1	2.8	0.93	0	0.93	0	0.93	...
255544	81721	17/12/05	21:46	166	98.84	45.21	34.91	17.07	1.2	0	0.6	0.6	0	0	...
218880	22130	18/7/05	5:07	11	12.85	81.84	99	1	18.18	9.09	0	9.09	0	0	...
244912	41084	8/11/05	2:46	35	99	38.74	13.15	98.56	0	0	0	0	0	0	...
273089	-1	25/2/06	4:22	92	90.46	58.59	68.63	11.64	8.7	2.17	1.09	0	5.43	0	...
265715	38794	2/2/06	0:57	275	81.4	69.47	29.78	20.28	6.55	2.91	0.73	0.73	1.09	1.09	...
198321	21367	17/4/05	22:23	110	54.02	89.83	14.1	94.75	10.91	5.45	0	1.82	0.91	2.73	...
45244	13359	21/12/02	18:01	45	92.84	81.29	10.08	67.75	8.89	4.44	0	0	0	4.44	...
233103	70832	1/10/05	9:19	77	95.05	69.84	65.41	97.38	2.6	0	0	1.3	0	1.3	...
566748	109818	25/3/09	5:25	77	89.94	74.2	9.09	99	2.6	0	1.3	0	0	1.3	...
146671	116703	24/1/07	7:25	38	33.88	1.81	98.54	74.74	7.89	7.89	0	0	0	0	...
745917	105443	1/11/10	6:46	242	27.37	38.61	93.65	6.99	12.81	8.26	1.24	2.48	0	0.83	...
618782	165386	11/7/09	2:46	119	55.71	50	10.42	1	3.36	0.84	0	0	0.84	1.68	...
55689	19796	10/2/03	2:07	12	1	20.24	98.01	25.77	16.67	16.67	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Response to student questions

- How many graphs do you recommend for us to produce to answer each question for Assignment 1? Thank you!
  - > I suggest one graph per question (for each or most of 5 main questions) plus one or two more if there is something you discover that you want to include. You can have some graphs in the appendix but not too many please. You should be looking at ways you can increase the amount of information each graph shows - for example multiple variables can be plotted in a single time series graph or histogram or boxplot. You can also use faceting to show multiple factor levels in a single plot.

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

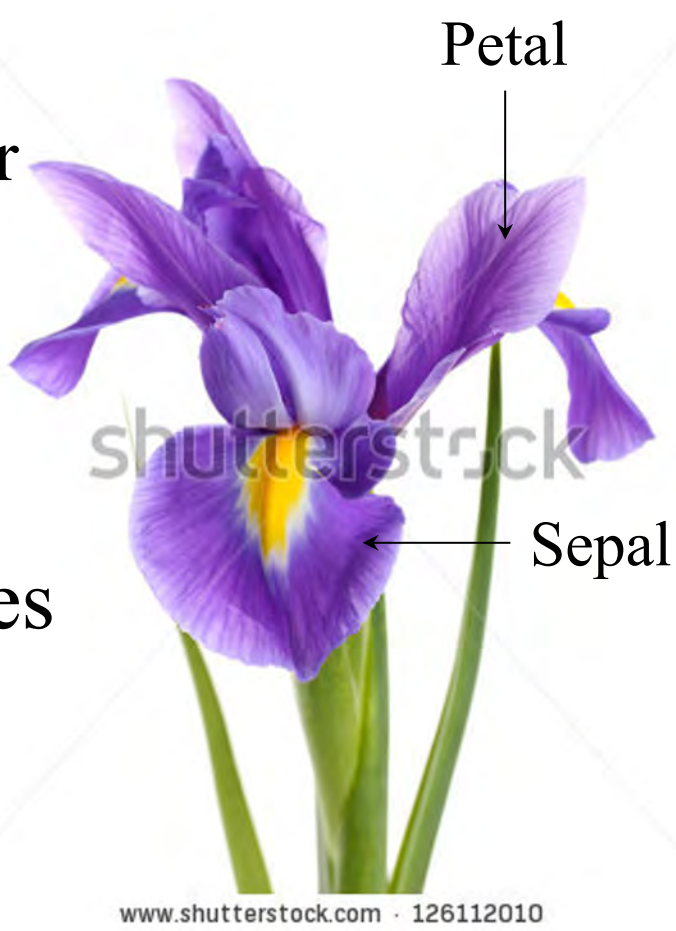
Four features measured:

- 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

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

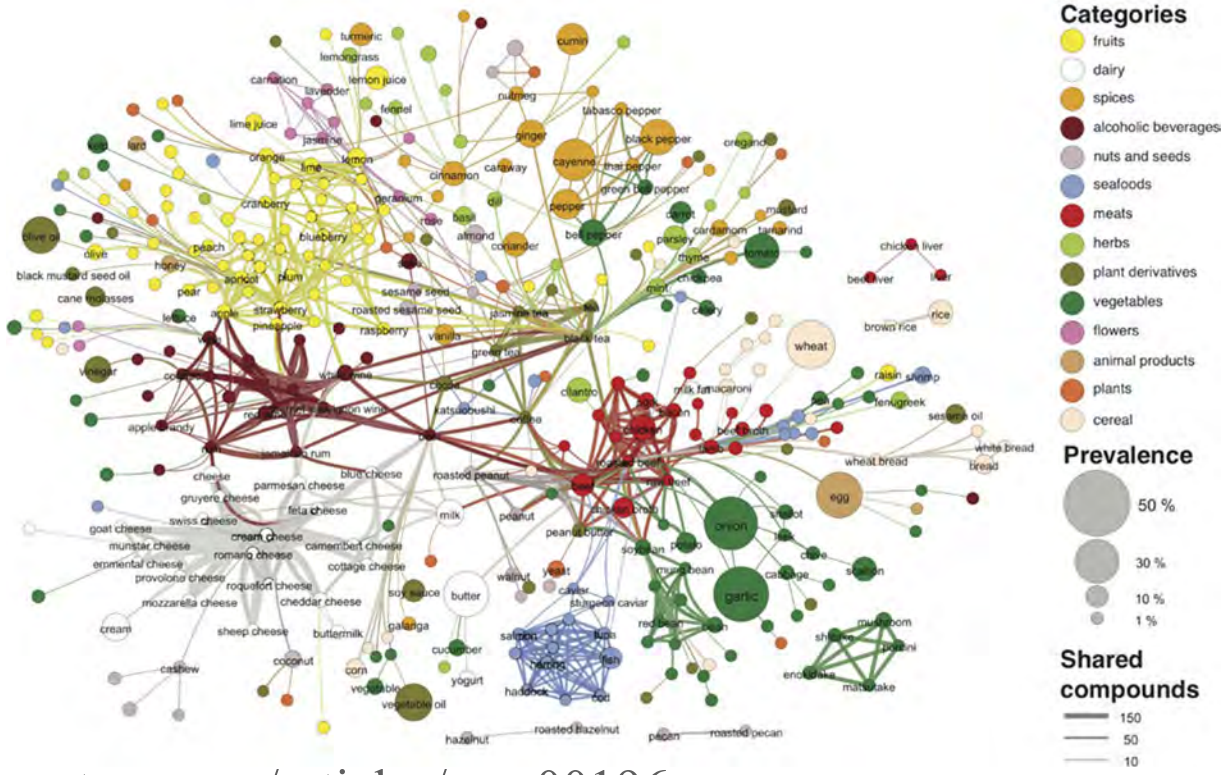
Which of the following is not true: Tidy data has:

```
(a)   Each value in its own cell
(b)   Each observation in its own row
(c)   Each factor level in its own column
(d)   Each variable in its own column
```

Networks

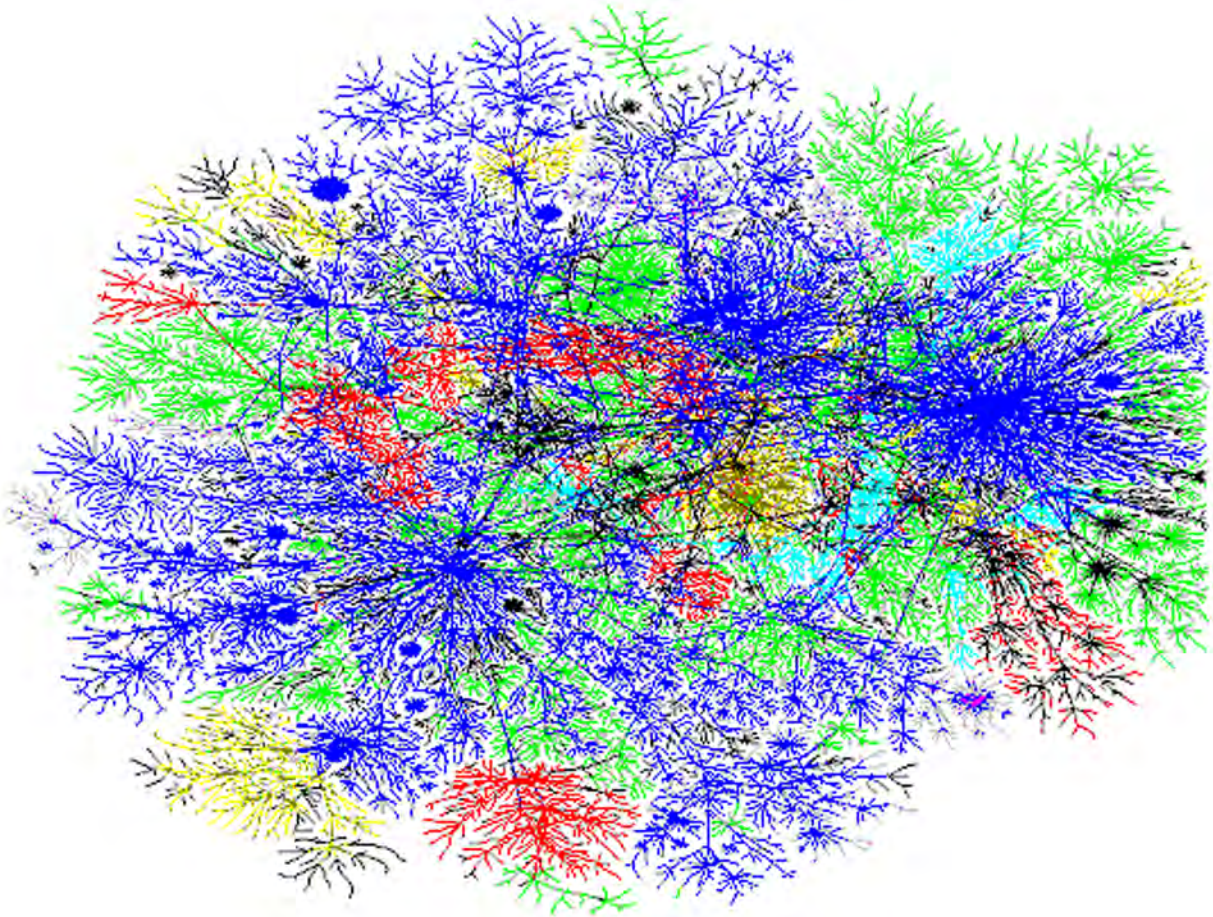
Flavor network

Flavor network and the principles of food pairing



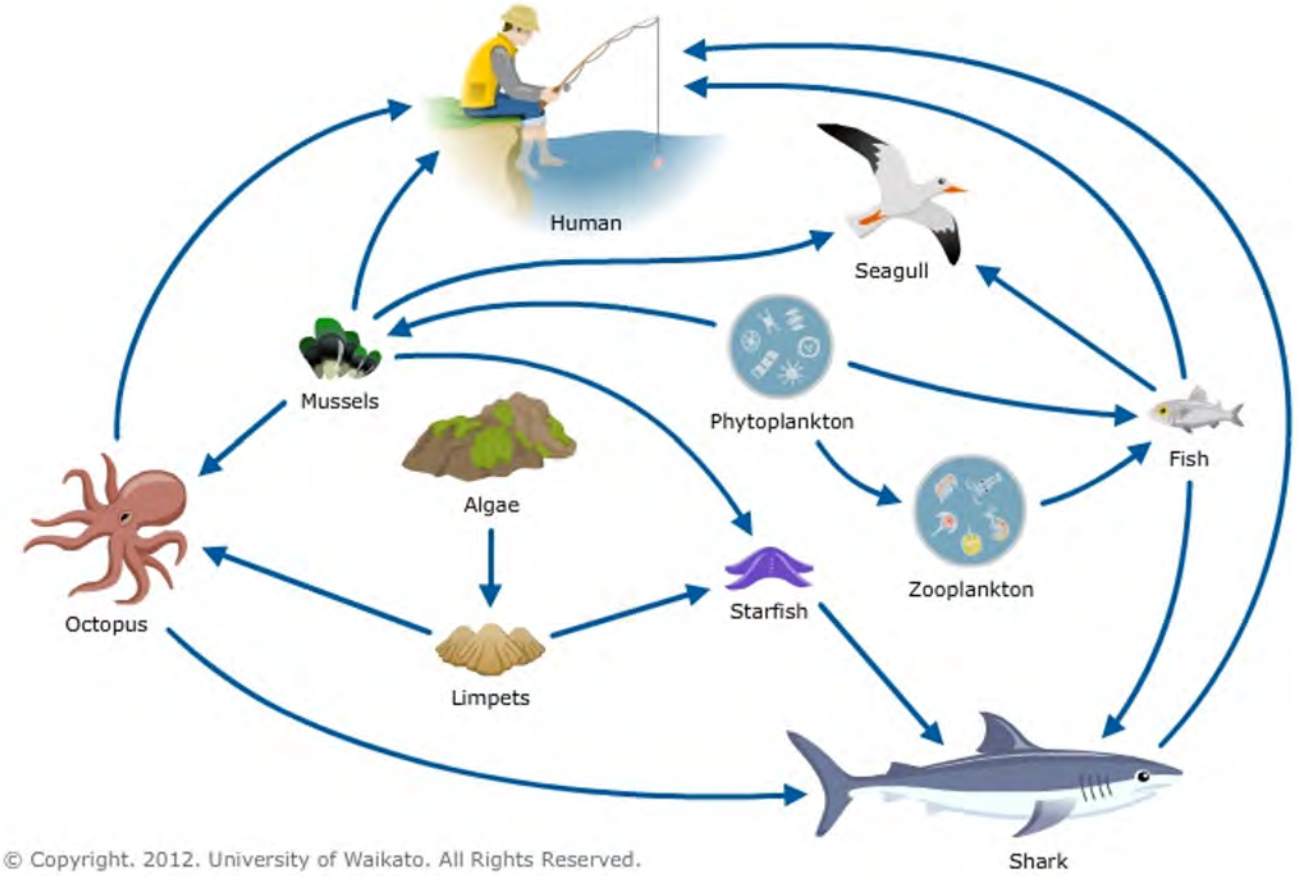
https://www.nature.com/articles/srep00196

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

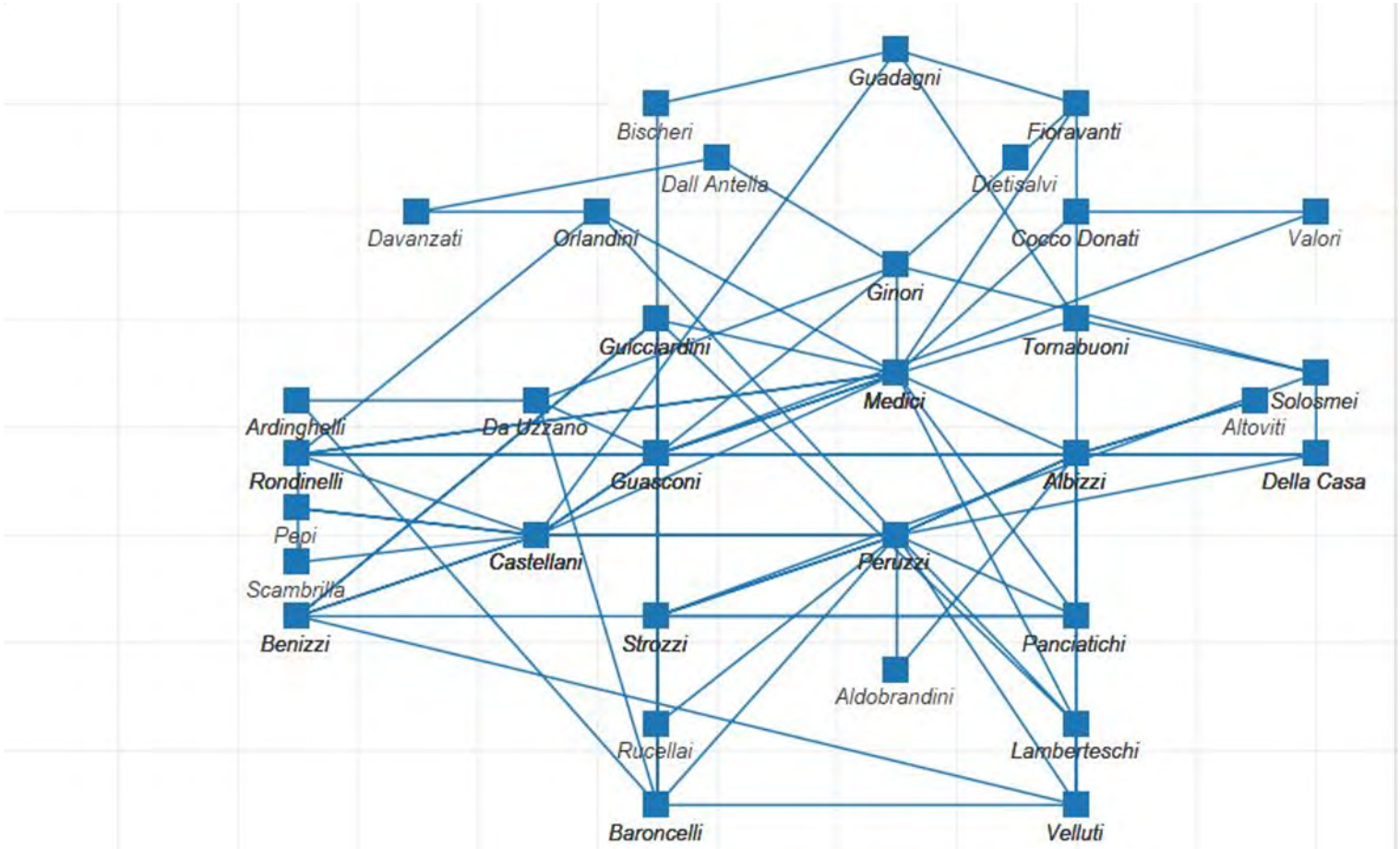
Transport network

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

Social networks

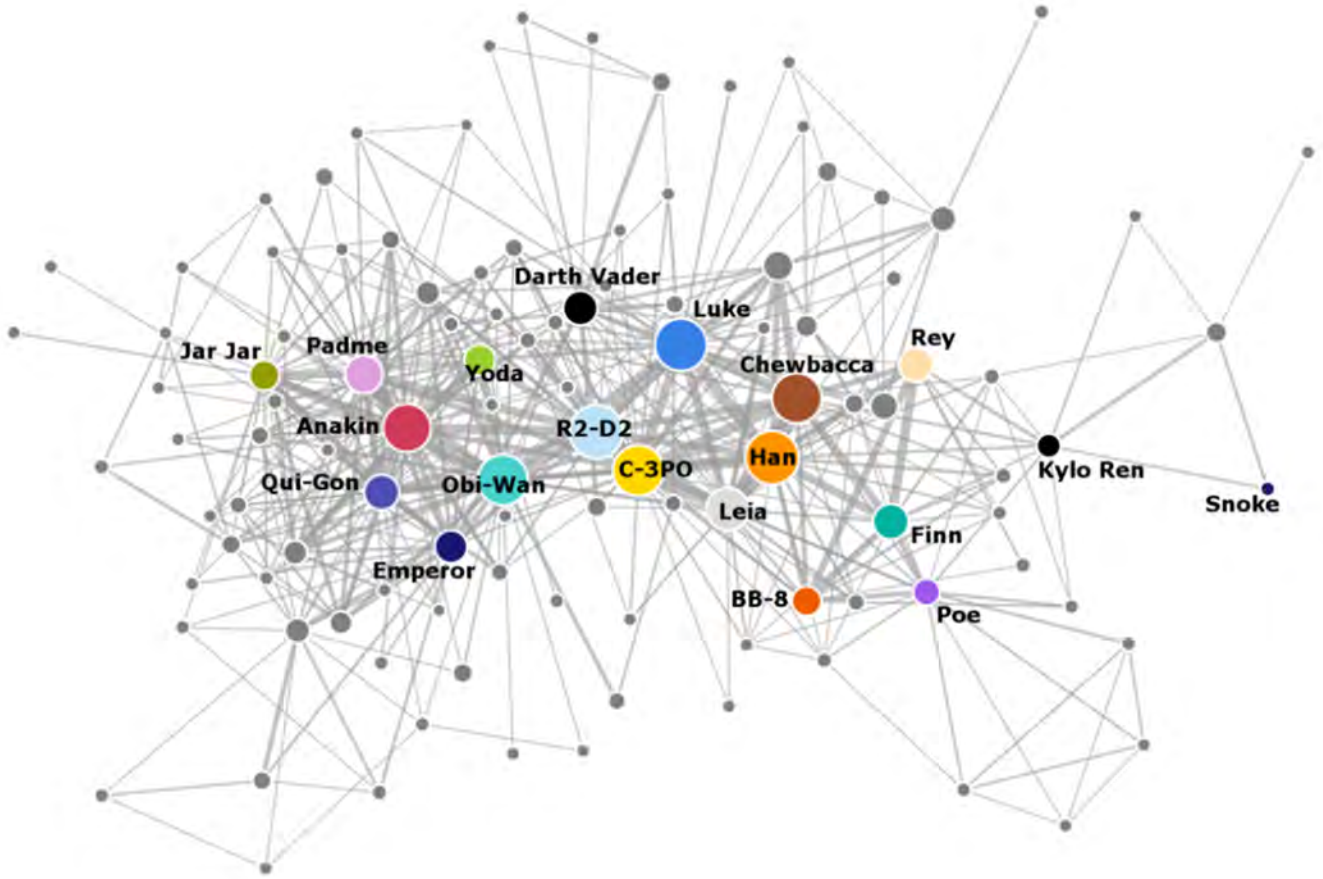
https://socialadr.com/features

Florentine social network: Medici



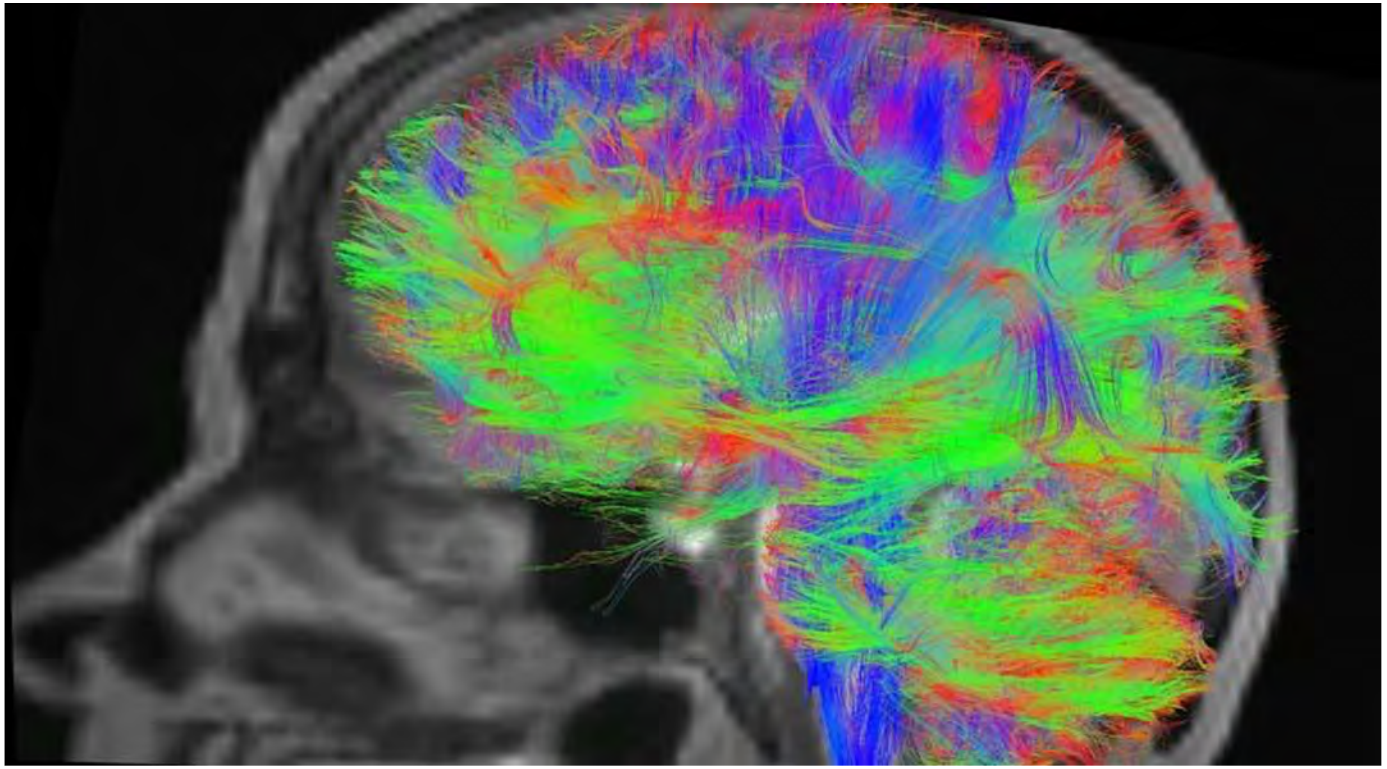
(inactive) http://blogs.teradata.com/international/tag/data-scientist/

Star Wars characters



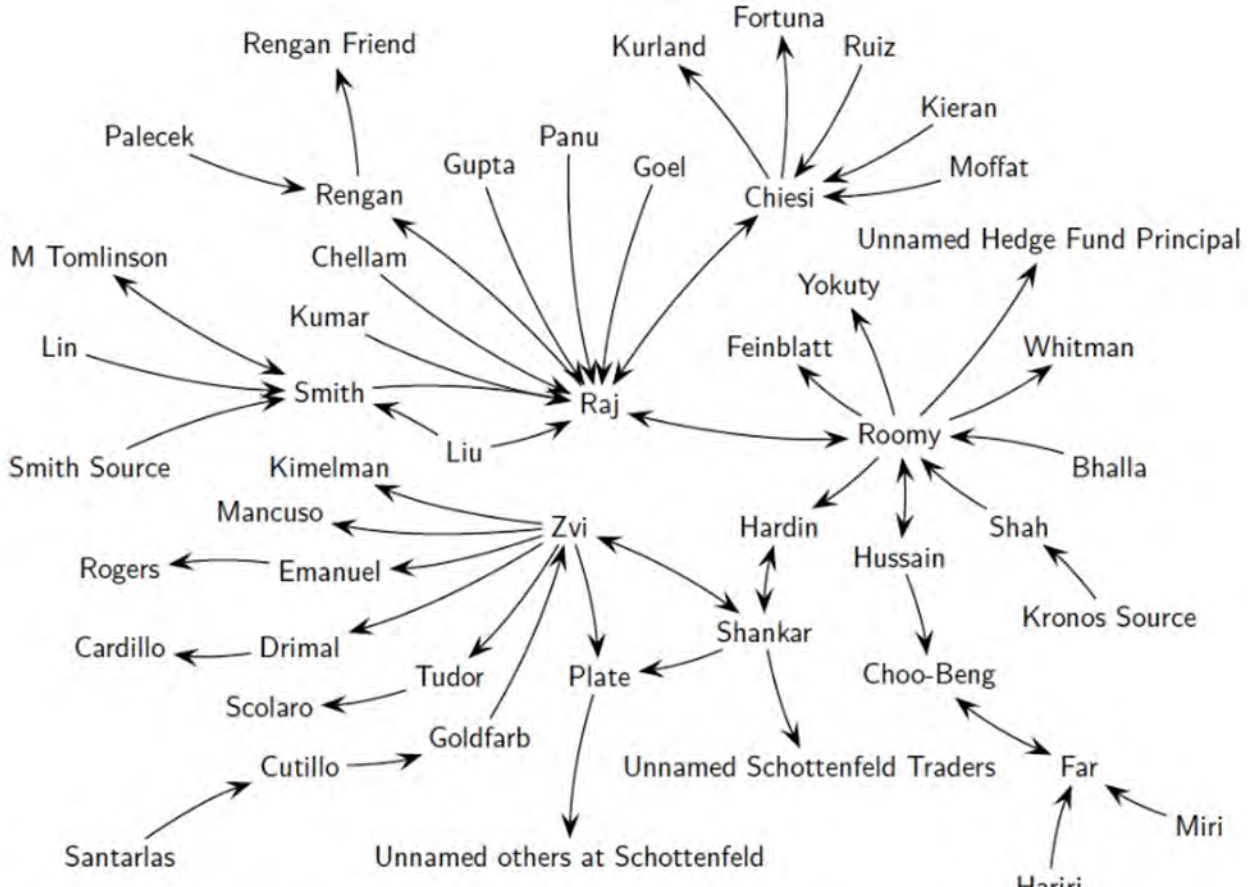
http://evelinag.com/blog/2016/01-25-social-network-force-awakens/#.WZwWUpOg8cl

Neural pathways



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

Information networks: insider trading



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

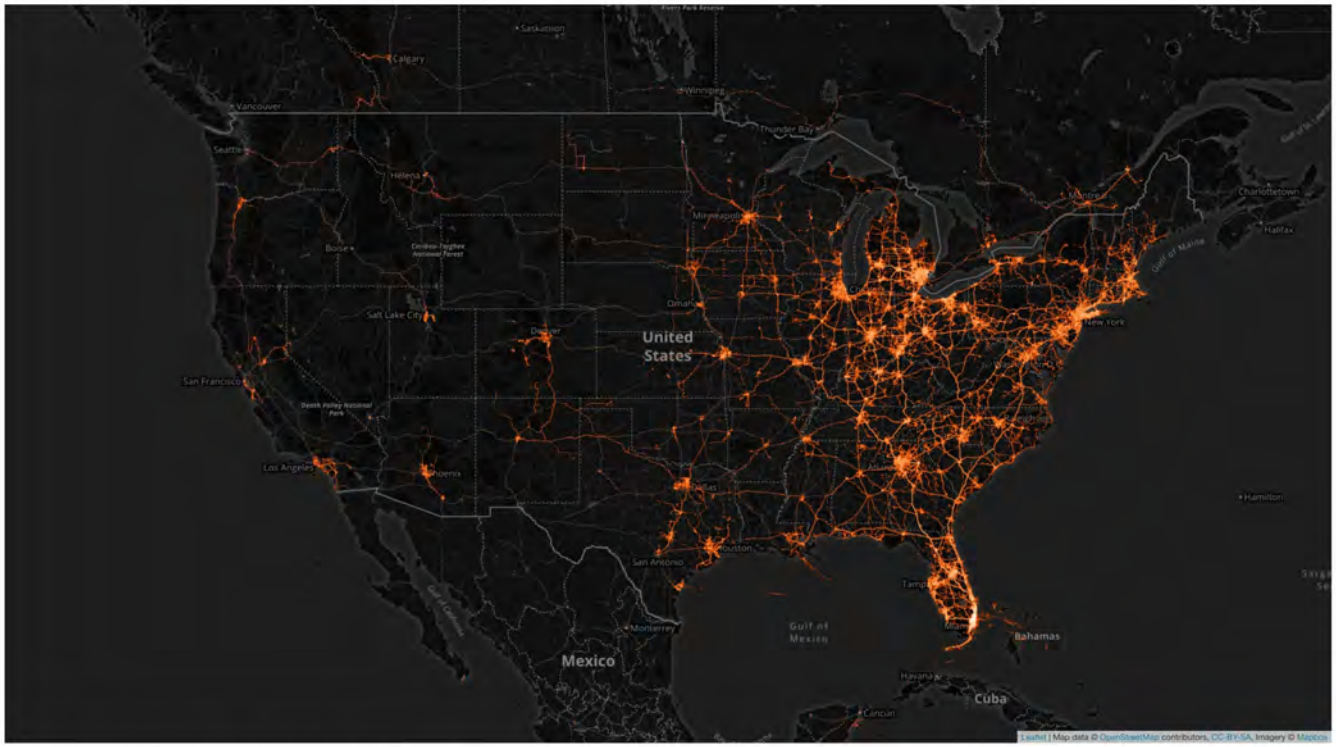
COVID-19 US social network

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?

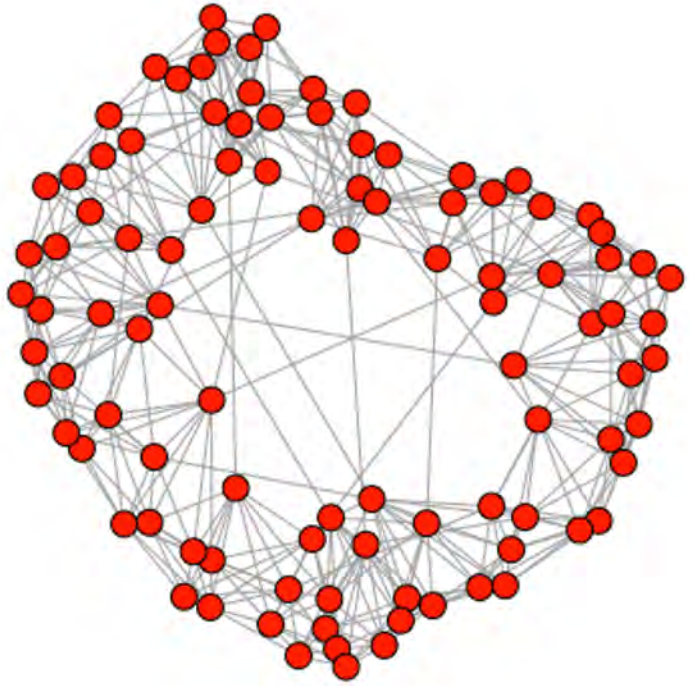
Connections between elements:

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

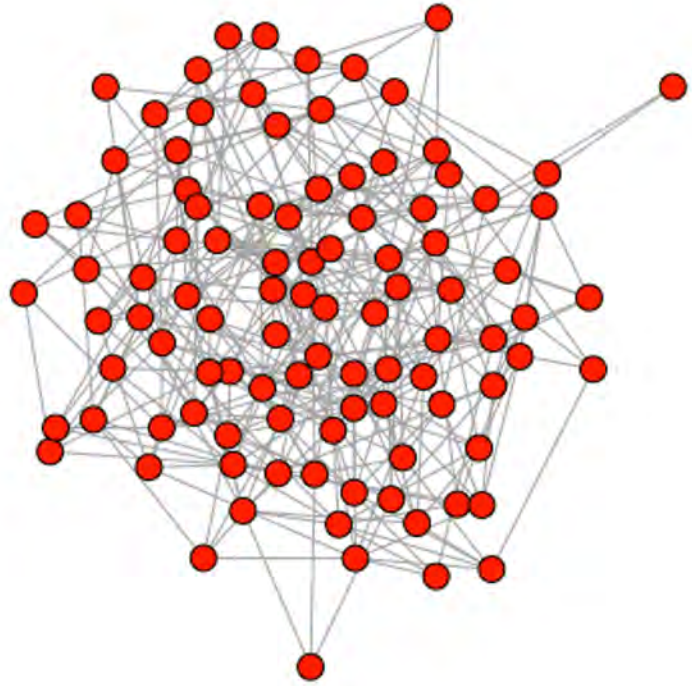
How to think about a network (a)

What is different about these networks?

(a)



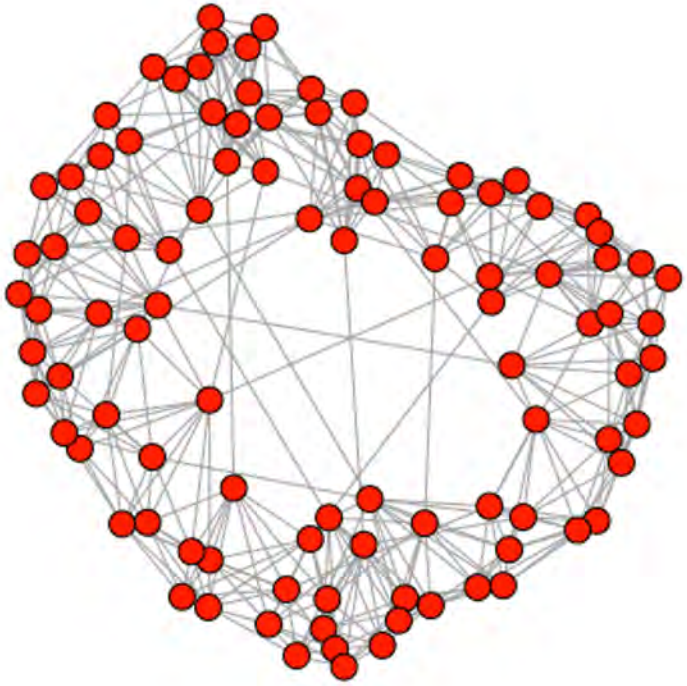
(b)



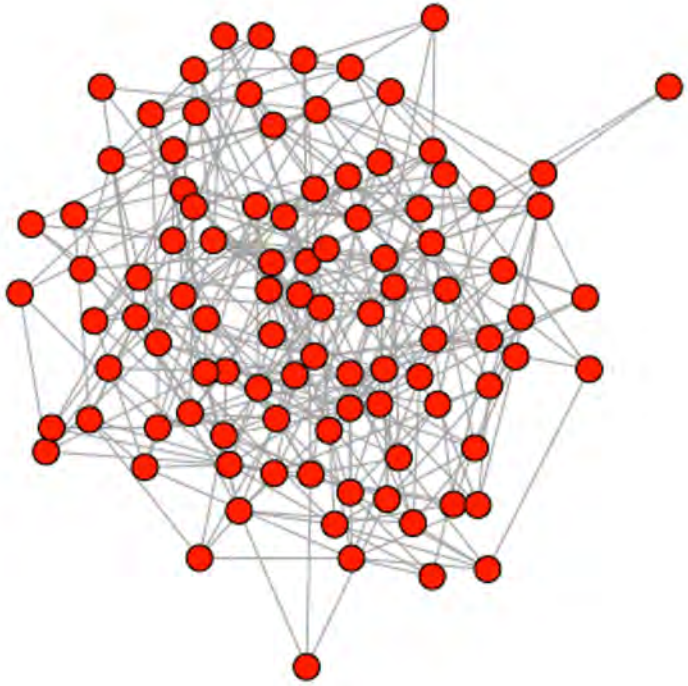
How to think about a network (b)

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.

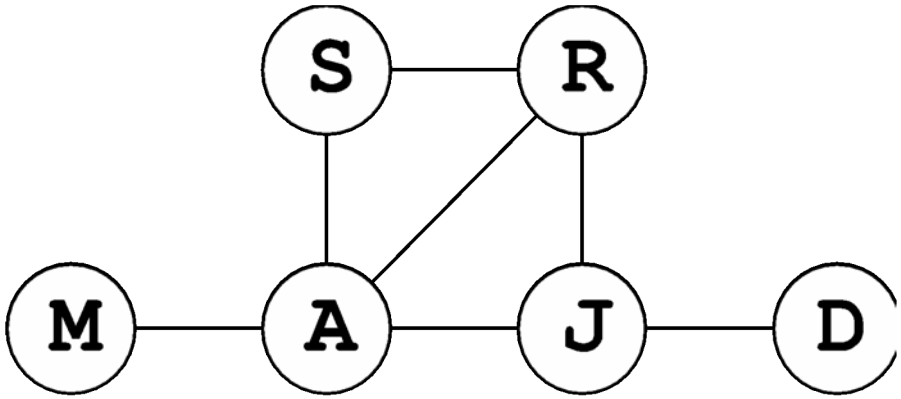
Questions to answer:

- 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…

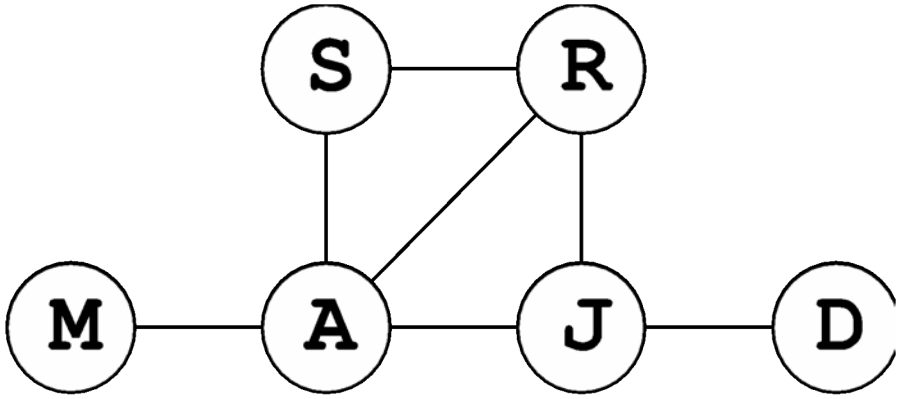
Particular link sequences have formal descriptions

- *Loop*: an edge from a vertex to itself
- *Complete*: a graph where 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:

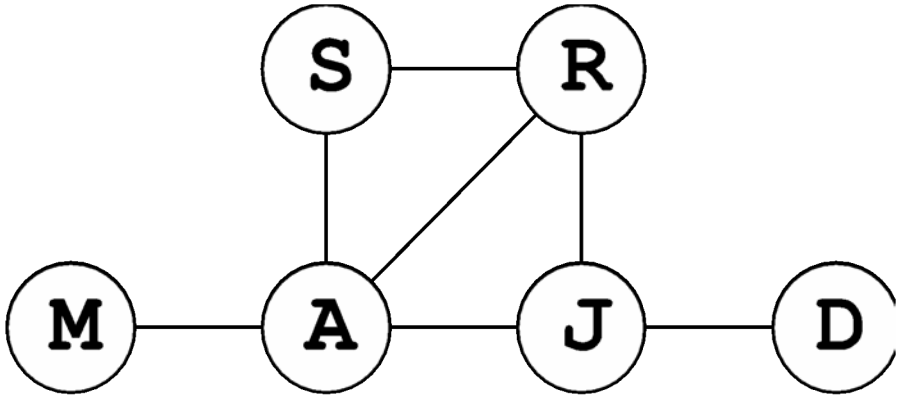
- *Walk*: (M, A, J, R, J, …)
- *Path*: (M, A, J, R)
- *Cycle*: (A, S, R, A)
- *Geodesic*: Geodesic(R, D) = (R, J, D)
- *Length*: dist(M,D) =3
- *Connected*: yes
- *Clique*: A, S, R
- *Simple*: yes.



Adjacency matrix

Summarizes the network by indicating the connections between individuals:

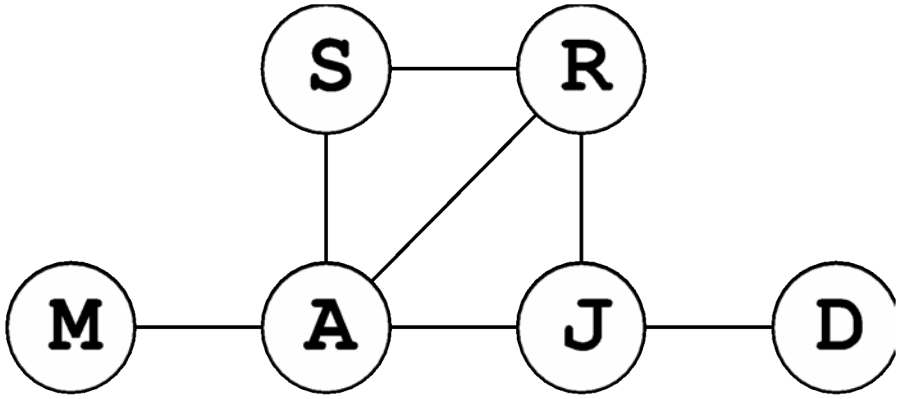
	J	A	S	D	R	M
J	0	1	0	1	1	0
A	1	0	1	0	1	1
S	0	1	0	0	1	0
D	1	0	0	0	0	0
R	1	1	1	0	0	0
M	0	1	0	0	0	0



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.

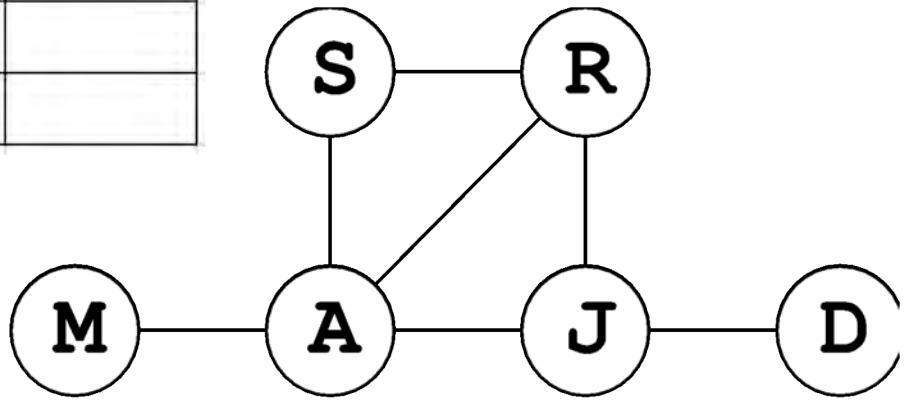
Analysis of these and other factors determine how *connected*, *robust* or *fragile* etc. a network is.

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



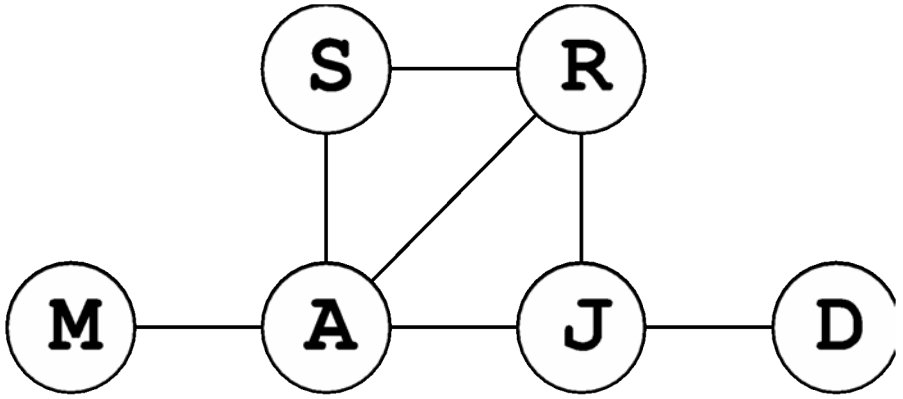
Analysing the network

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

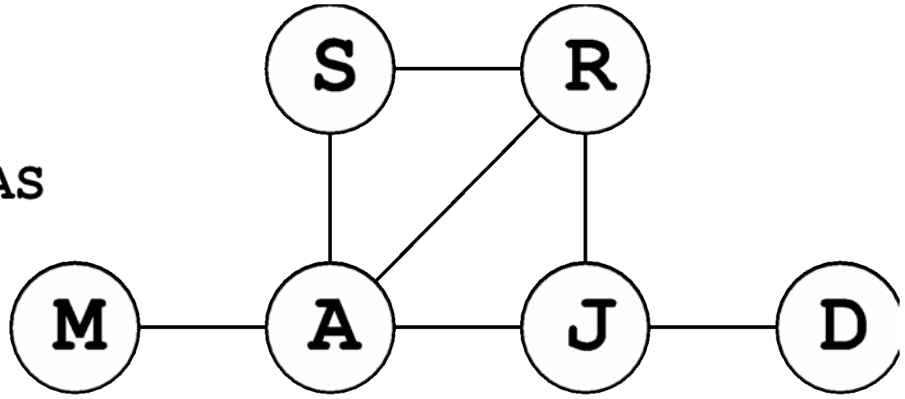
- *Density*: is the proportion of edges in a graph, relative to the maximum number possible.
- $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2}$
- where $|E_g|$ is number of edges, $|V_g|$ is number of vertices
- *Clustering coefficient*: is the proportion of triangles relative to the number of connected triples.
- $clt(g) = \frac{3\tau_\Delta(g)}{\tau 3(g)}$ Also known as transitivity
- 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*: $clt(g) = \frac{3\tau_{\triangle}(g)}{\tau 3(g)} = \frac{3\times2}{13} = 0.462$

(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*

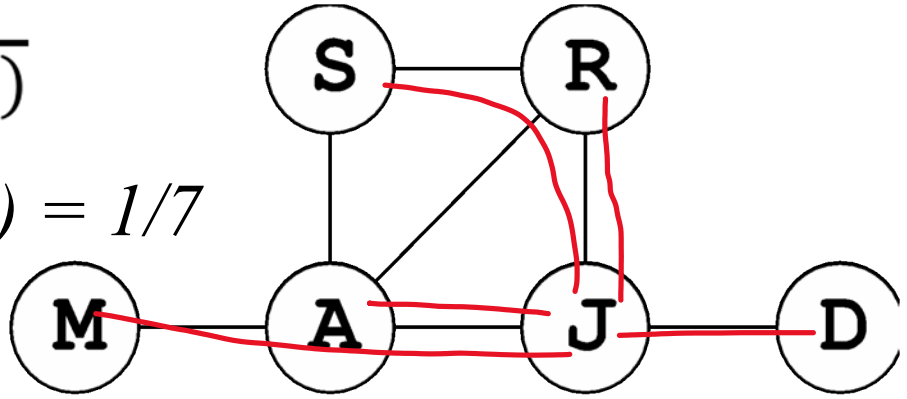
Closeness centrality

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.

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

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



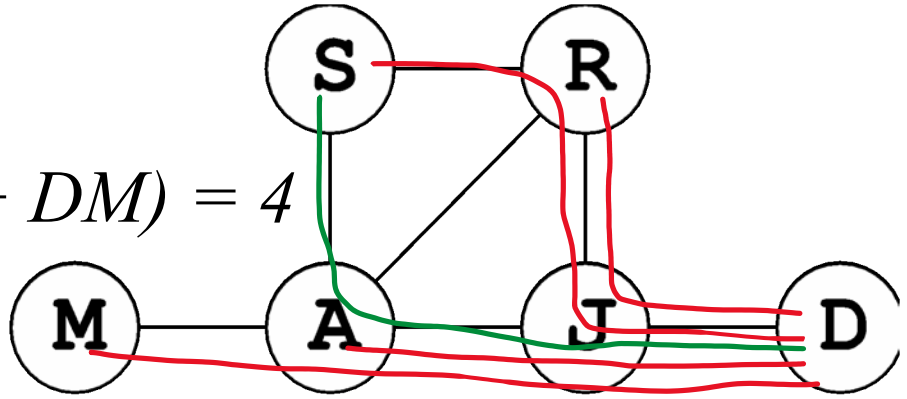
Betweenness centrality

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/

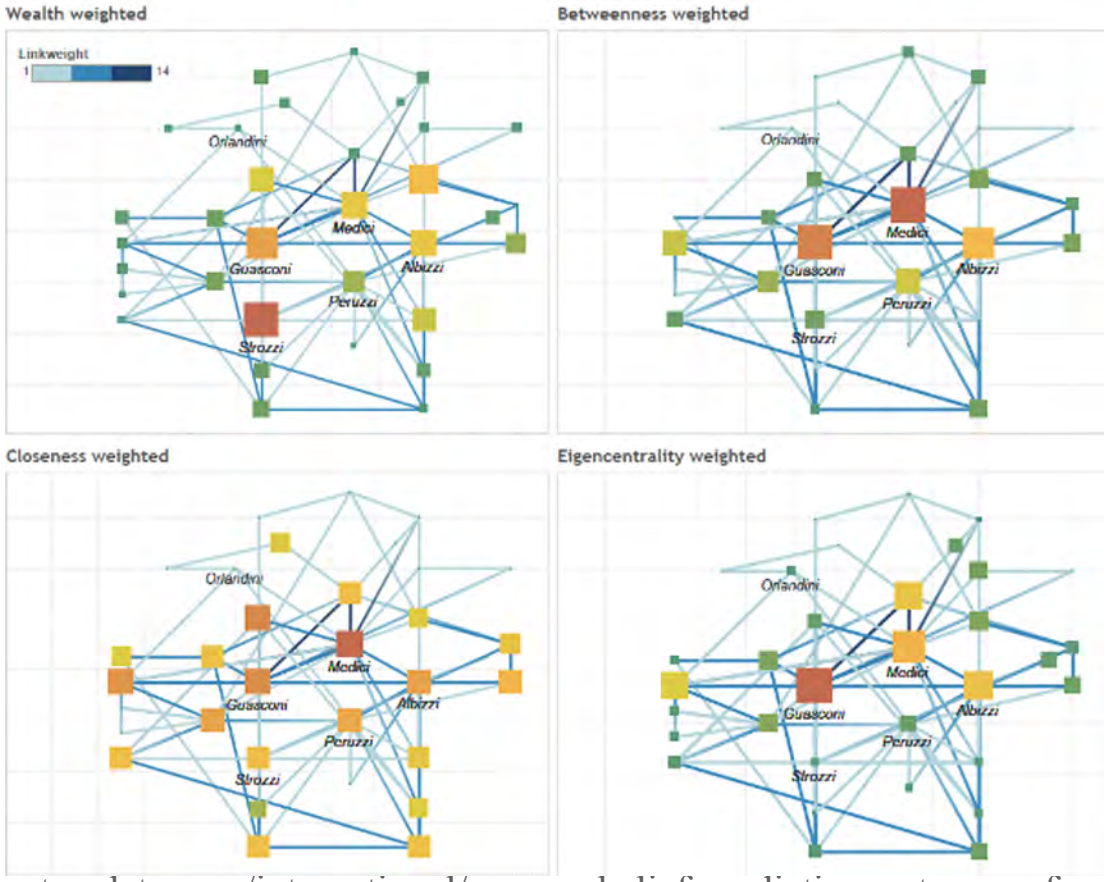
Which centrality measure?

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?

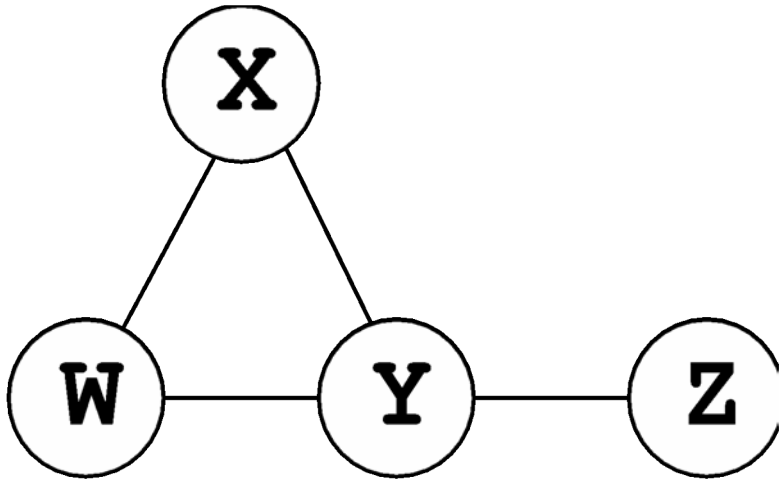


(inactive) http://blogs.teradata.com/international/a-crazy-belief-predicting-outcomes-from-network-graphs/

Class example

For the graph below calculate the following:

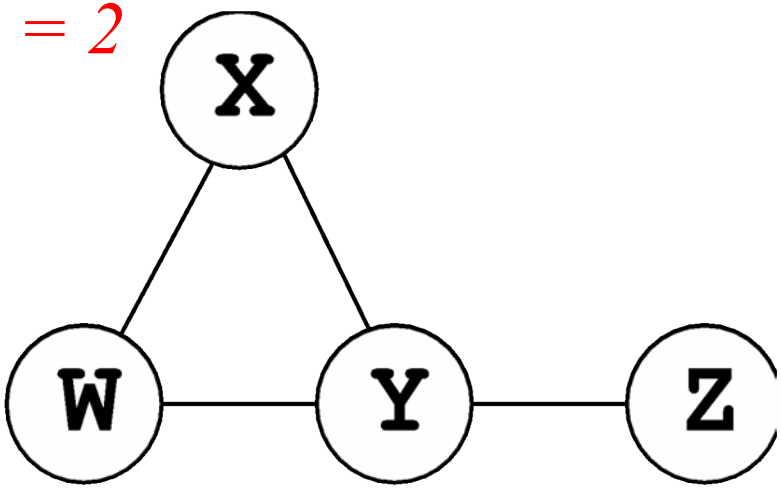
- *Average path length*
- *Diameter*
- d_Y
- *Degree distribution*
- $c_B(Y)$
- $c_{Cl}(Y)$
- *Cliques*



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 directly from df

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:

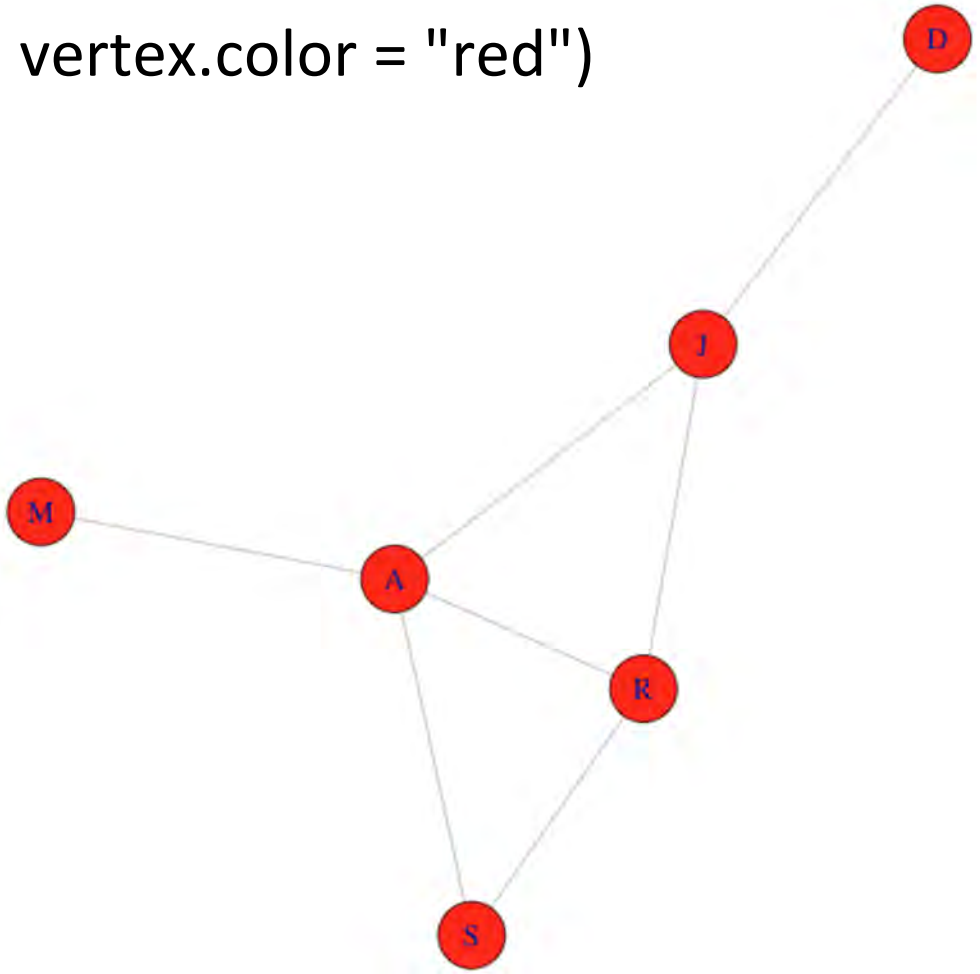
```
> graph.density(g)
[1] 0.467

> 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 . . . .
```

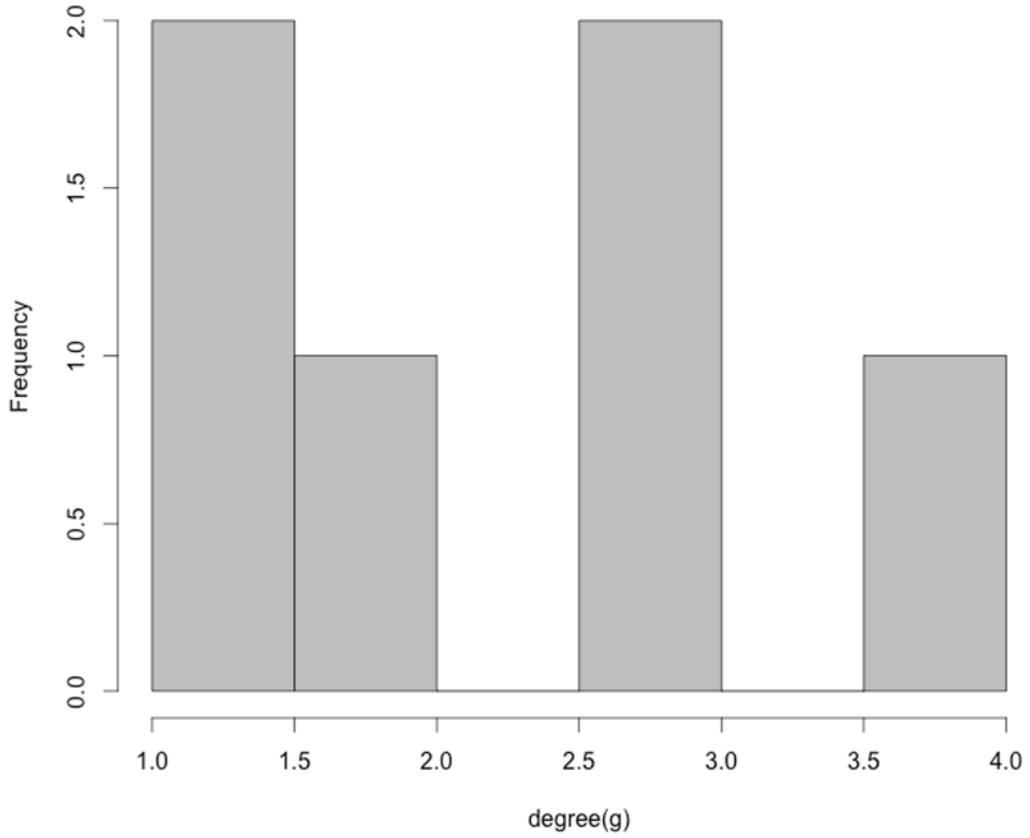
Basic plot

```
> 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

More on creating graphs

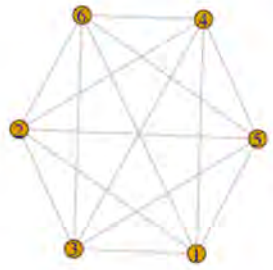
Other methods of creating graphs:

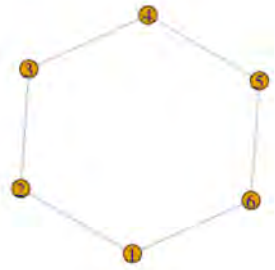
- Inbuilt graph models based on fundamental topologies.
- Direct creation of graphs.
- Graphs from a csv file
- Creating cliques directly and merging two graphs

Four important network topologies

Using R:

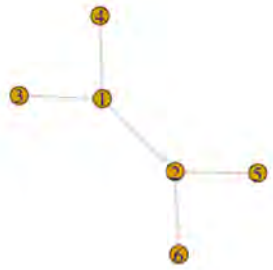
Complete





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)
```

Direct creation of graphs

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)
```

Graphs from a csv file

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

Two-mode network

Two-Mode networks arise when

- Entities can be considered connected via an association to a third party.
- For example, Person A and Person B are both members of a club. Therefore, they can be considered as having a social connection.

The following slides show:

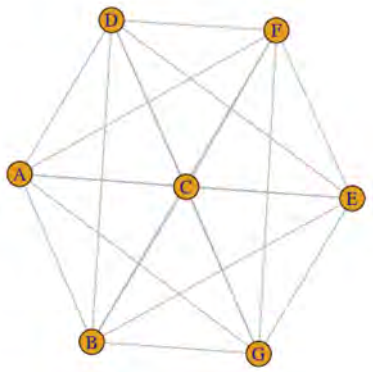
- Creation by hand creation of cliques and merging,
- Then in code (for reading/personal work only).

Creating a clique directly

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

- For a club having members: A, B, C, D, E, F, G:

```
> # 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)
```



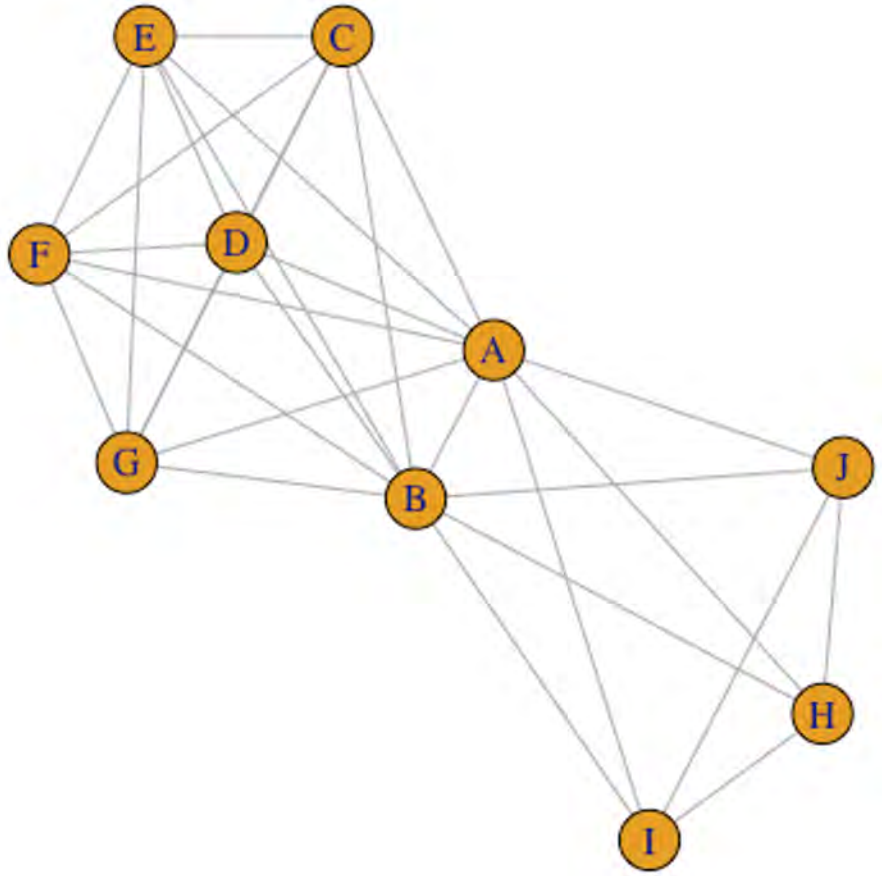
Merging two graphs

You can merge two graphs to form a new graph:

```
> # first group
> gg = graph_from_literal(A:B:C:D:E:F:G -- A:B:C:D:E:F:G)
> # second group from a different club with some
  members in common. Members A, B, H, I, J:
> 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)



Coding directly (make data frames)

Data frames of people and club membership

```
> rm(list = ls()); library(ggplot2); library(igraph)
> # make a data frame of people and club membership
> Person = as.data.frame((c("A", "B", "C", "D", "E", "F",
  "G", "H", "I", "J", "A", "B")))
> Club = as.data.frame((c("X", "X", "X", "X", "X", "X", "X",
  "Y", "Y", "Y", "Y", "Y")))
> ClubData = cbind(Person, Club)
> colnames(ClubData) = c("Person", "Club")
> UniquePerson = unique(Person)
> colnames(UniquePerson) = "Person"
```

Coding directly (make data frames)

The Data frames:

ClubData | Make two r

Filter

	Person	Club
1	A	X
2	B	X
3	C	X
4	D	X
5	E	X
6	F	X
7	G	X
8	H	Y
9	I	Y
10	J	Y
11	A	Y
12	B	Y

UniquePerson

Filte

	Person
1	A
2	B
3	C
4	D
5	E
6	F
7	G
8	H
9	I
10	J

Coding directly (create graph)

Make an empty graph and add vertices

```
> g <- make_empty_graph(directed = FALSE)


> # add vertices using “for loop”
> for (i in 1 : nrow(UniquePerson)) {
>   g <- add_vertices(g, 1, name =
  as.character(UniquePerson$Person[i]))
> }
```

Coding directly

Make complete graph for each group, add to g (two loops)

```
> # loop through each group
> for (k in unique(ClubData$Club)){
> temp = ClubData[(ClubData$Club == k),]
> # combine each pair of agents to make an edge list
> Edgelist = as.data.frame(t(combn(temp$Person,2)))
> colnames(Edgelist) = c("P1", "P2")
> # loop through pairs of edges and add
> for (i in 1 : nrow(Edgelist)) {
> g <- add_edges(g,
  c(as.character(Edgelist$P1[i]),as.character(Edgelist$P2[i])))
> }
> }
```

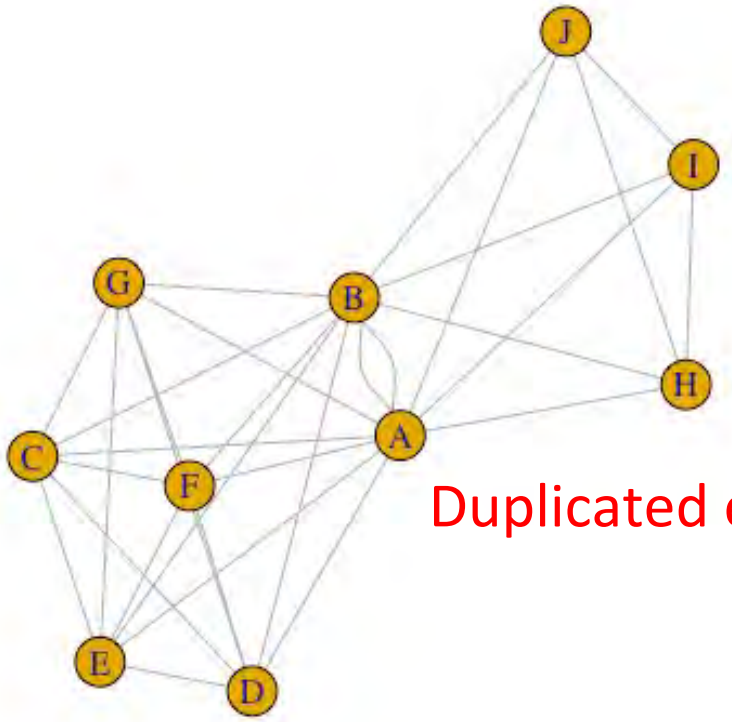
Coding directly

Plot, and simplify to remove duplicated edge

> plot(g)

> g = simplify(g)

> plot(g)



Duplicated edge

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

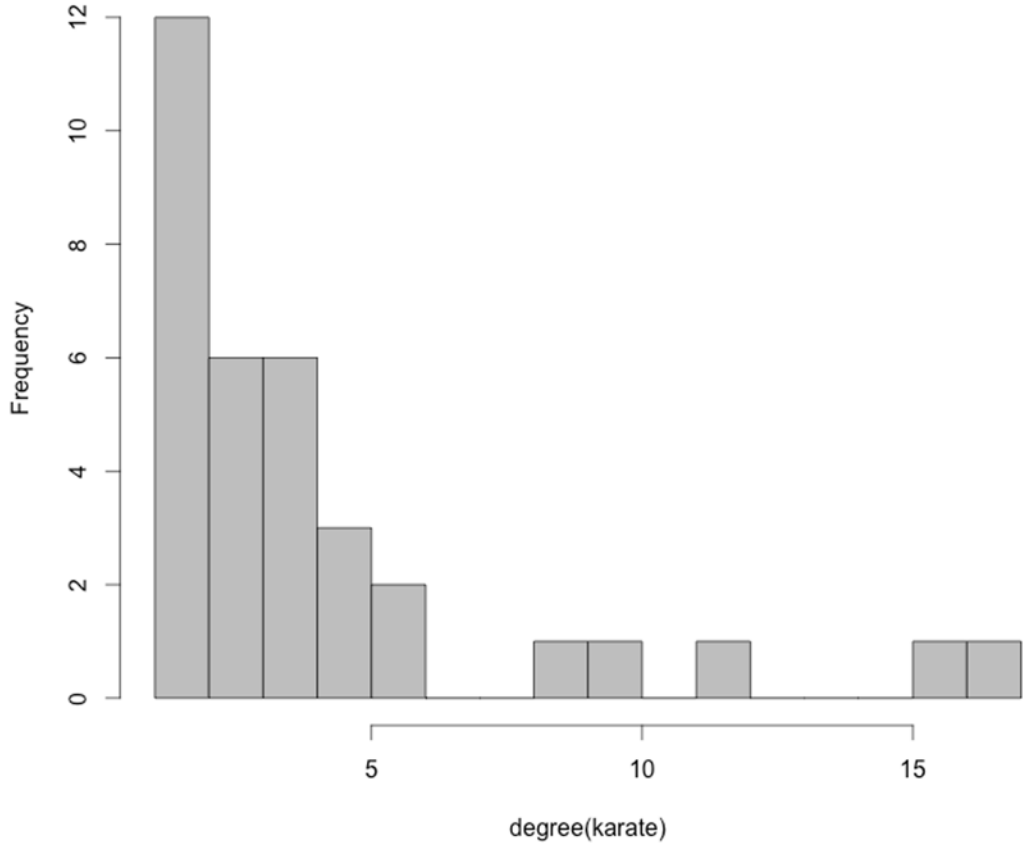
```
> plot(karate, layout = layout.circle, vertex.color = "red")
```

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
```

Community detection

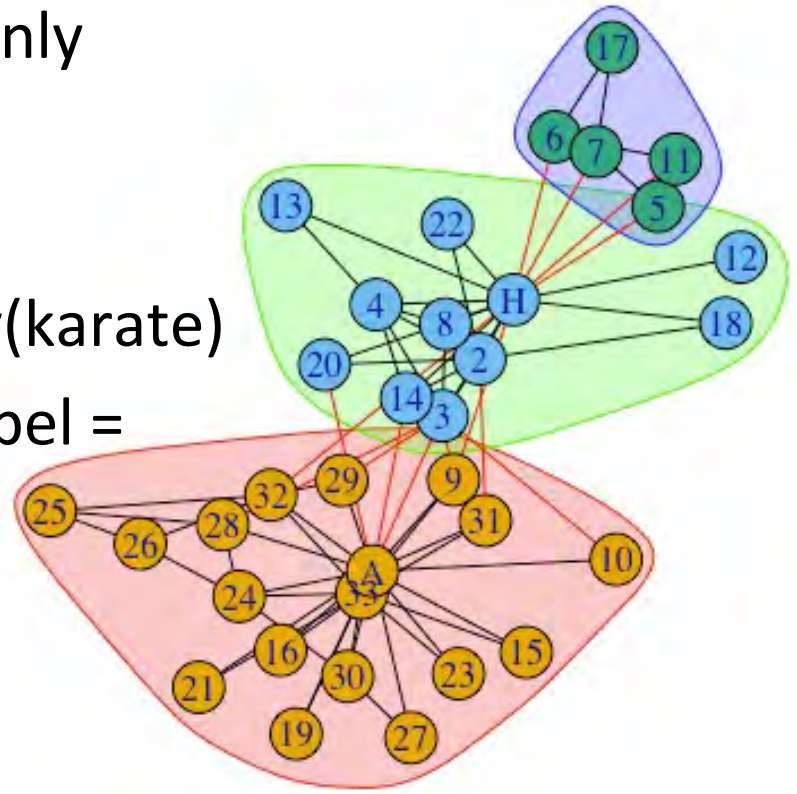
iGraph enables the detection of communities (or subgroups)in a social network.

- To do this, fit a community detection algorithm and store the results as a “communities” class object.
- The network can then be plotted to show the communities.
- There are many detection algorithms, see *A User’s Guide to Network Analysis in R*, pages 118-119. (References on last slides)

Karate club – community detection

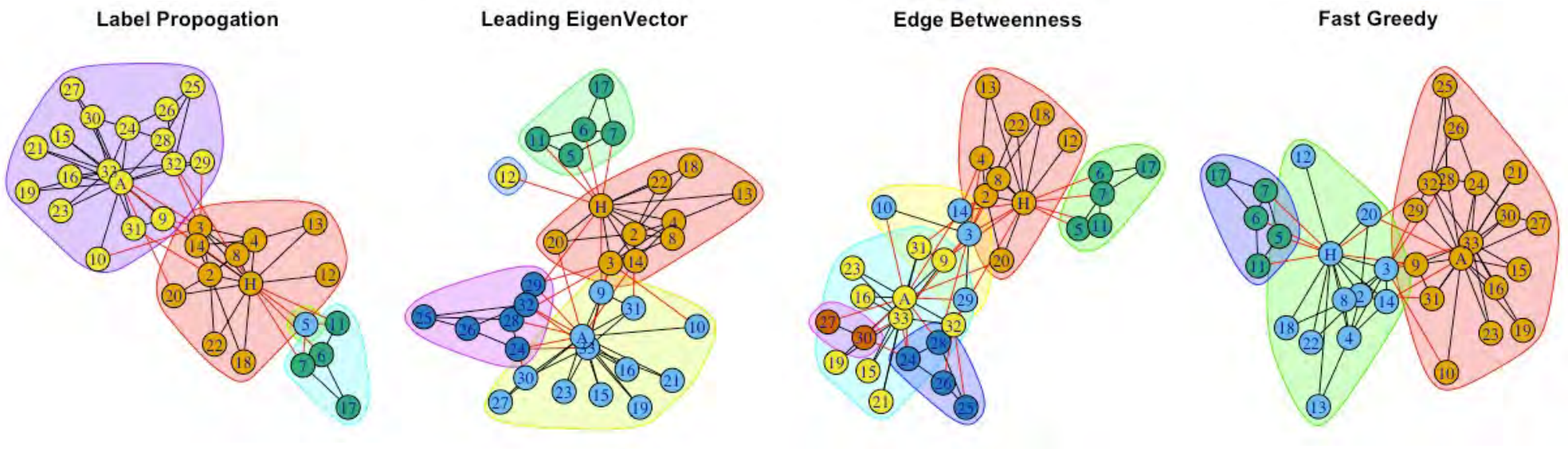
We can get another view of the club by clustering

```
> # fast greedy clustering only
> data("karate")
> # detect communities
> cfb = cluster_fast_greedy(karate)
> plot(cfb, karate,vertex.label =
  V(karate)$role, main
  ="Fast Greedy")
```



Karate club – community detection

Different algorithms give different communities



Something for you to investigate!

Karate club – community detection

Different methods give different clusters

```
> #create the clusters/community groupings
> ceb = cluster_edge_betweenness(karate)
> cfb = cluster_fast_greedy(karate)
> cle = cluster_leading_eigen(karate)
> clp = cluster_label_prop(karate)

> #create community plots in the karate network
> plot(ceb, karate,vertex.label=V(karate)$role,main="Edge Betweenness")
> plot(cfb, karate,vertex.label=V(karate)$role,main="Fast Greedy")
> plot(cle, karate,vertex.label=V(karate)$role,main="Leading EigenVector")
> plot(clp, karate,vertex.label=V(karate)$role,main="Label Propogation")
```

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")
```

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 online 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