

Lecture 9

- Assignment 2/Quiz and Practical Activity
- Introduction to Networks
- Network Analysis in R

Presenter: Dr John Betts

Week-by-week outline

Week Starting	Seminar	Topic	App Ses	A1	A2	Q/P	A3	Due Date
2/3/2026	1	Introduction to Data Science, R, review of basic statistics	-					
9/3/2026	2	Data visualisation	S1	■				
16/3/2026	3	Data manipulation	S2	■				
23/3/2026	4	Regression modelling	S3	■				
30/3/2026	5	Clustering	S4	■				
6/4/2026	-	Mid-semester Break		■	■	■	■	
13/4/2026	6	Classification using decision trees	S5	■	■			17/4/2026
20/4/2026	7	Improving and evaluating classifiers. Naïve Bayes classification	S6		■			
27/4/2026	8	Ensemble methods, Artificial Neural Networks	S7		■			
4/5/2026	9	Network analysis	S8		■			
11/5/2026	10	Introduction to text analysis	S9		■			15/5/2026
18/5/2026	11	Text analysis applications	Quiz/Prac			■		22/5/2026
25/5/2026	12	Text Network Analysis, Review of the unit, Assignment 3	S10,11,12				■	
1/6/2026		SWOT VAC	-				■	
8/6/2026		EXAM PERIOD	-				■	12/6/2026

SETU

Student Evaluation of Teaching and Units (SETU) has opened for Semester 1.

- All students are encouraged to participate. Your feedback in all units studied is very important.
- You will see a block in the side bar on Moodle linking you to the survey.

Assignment 2

FIT3152 Data analytics – S1 2026: Assignment 2

Your task	<ul style="list-style-type: none">• The objective of this assignment is to gain familiarity with classification models using R.• This is an individual assignment.
Value	<ul style="list-style-type: none">• This assignment is worth 20% of your total marks for the unit.• It has 40 marks in total.
Suggested Length	<ul style="list-style-type: none">• 8 – 10 A4 pages, approximately 1,000 words (for your report) + extra pages as appendix for your R script.• Font size 11 or 12pt, single spacing.
Due Date	11.55pm Friday 15th May 2026
Submission	<ul style="list-style-type: none">• Submit a single PDF file and single video presentation file on Moodle.• Use the naming convention: <i>FirstnameSecondnameID.{pdf, mp4, mov etc.}</i>• Turnitin will be used for similarity checking of all written submissions.
Generative AI Use	<ul style="list-style-type: none">• Statement on Generative AI required by the university: In this assessment, you can use generative artificial intelligence (AI) in order to search for R functions and examples to perform tasks that you specify only. Any use of generative AI must be appropriately acknowledged (see Learn HQ)..
Late Penalties	<ul style="list-style-type: none">• 5% (2 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 2

Instructions and data

The objective of this assignment is to gain familiarity with classification models using R. You will be using a modified version of the World Values Survey (WVS) data that was used for Assignment 1. We now want to create Machine Learning models to predict the level of confidence in different social organizations based on participant responses to other selected attributes. For Assignment 2, confidence in individual social organisations has been re-coded as binary class variables. You can read more on the World Values Survey here: <https://www.worldvaluessurvey.org/WVSContents.jsp>

There are two options for compiling your written report:

- (1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
- (2) As an R Markdown document that contains the R code with the discussion/text interleaved. Render this as an HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

Assignment 2

Creating your data set

Clear your workspace, set the number of significant digits to a sensible value. Download “WVSBinaryExtract.csv” and create your individual data using the following code. Your individual data will have Country, Wave, **three** different binary class variables (confidence in social organisations) and **30** coded participant responses (attributes), with **20,000** observations.

```
rm(list = ls())
set.seed(XXXXXXXX) # Your Student ID is the random seed
WD = read.csv("WVSBinaryExtract.csv")
selected_cols = c(sample(3:49, 30), sample(50:63, 3))
WD = WD[c(1:2, selected_cols)]
WD = WD[sample(nrow(WD), 20000, replace = FALSE),]
```

Consider saving your working file.

Use your individual data created above to answer all questions.

Assignment 2

Questions (18 Marks)

1. Explore the data: What is the proportion of “**High**” (Confidence = 1), to “**Low**” (Confidence = 0) for your three binary class variables? Obtain descriptions of the other predictor attributes. Is there anything noteworthy in the data? Are there any attributes you need to consider omitting from your analysis? **(2 Marks)**
2. Document any pre-processing required to make the data set suitable for the model fitting that follows. **(1 Mark)**
3. Divide your data into a 70% training and 30% test set by adapting the following code (written for the iris data). Use your student ID as the random seed.

```
set.seed(XXXXXXXX) #Student ID as random seed
train.row = sample(1:nrow(iris), 0.7*nrow(iris))
iris.train = iris[train.row,]
iris.test = iris[-train.row,]
```

Assignment 2

4. Using your training data and each of the techniques below, implement models to separately classify **each** of your three class variables. For this question you may use the R functions at their default settings if suitable. Note: Do not use any class variables as predictor attributes. **(5 Marks)**
- Decision Tree
 - Naïve Bayes
 - Bagging
 - Boosting
 - Random Forest

Assignment 2

5. Using the test data, classify each of the test cases as “High” (Confidence = 1) or “Low” (Confidence = 0) for each of your three class variables using each of the models implemented in Question 4. Create confusion matrices and report the accuracy, precision, recall, and F1-score of each model in classifying each class variable. Summarise your results in **three tables**, one for each class variable. **(3 Marks)**

6. Using the test data, calculate the confidence (probability) of predicting “High” (Confidence = 1) or “Low” (Confidence = 0) using each classifier, and construct ROC curves for each of your three class variables. Make a separate plot for each class variable with all the curves for that class on the same axis. Use a different colour for each classifier. This will give you **three plots** with **five curves** in each. Calculate and report the AUC for each classifier over each class variable. **(4 Marks)**

7. Compare the results in Questions 5 and 6 in relation to the performance of all classifiers across your three class variables. Is there a single “best” classifier for predicting each of the classes? Is there a “worst” **ensemble** classifier overall? Do the “best” or “worst” classifiers vary between your three class variables? Explain your reasoning. **(3 Marks)**

Assignment 2

Investigative Tasks (18 Marks)

8. Examining each of the **ensemble** models for each of your three class variables, determine the most important attributes in predicting confidence. Summarise your results in **three** tables for each class variable. Which attributes are important overall for predicting confidence in each class variable? Which attributes could be omitted from the data with very little effect on performance? Give reasons. **(5 Marks)**

9. Looking at the relative strengths and weaknesses of each model, and consulting relevant references, can you explain any difference in performance between the models across each of the class variables? That is, why do some of these models perform better than others on the type of data to be fitted? **(2 Marks)**

Assignment 2

10. Improve the worst performing **ensemble** model from Question 7. You may do this by adjusting parameters, attribute selection, cross-validation of the basic ensemble model and/or any other applicable process. Show that your new model is better using F1-scores, ROC curves and AUC. Describe how you created your improved model. What factors were important in your decision? State why you chose the attributes you used. **(4 Marks)**

11. Using the insights from your analysis so far, implement an Artificial Neural Network classifier to compare predictability over time for the country with the greatest number of observations in your data. Using the training data for that country, fit an Artificial Neural Network model to classify confidence in **one** of your three class variables. Describe the attributes used and data pre-processing for implementation. **(3 Marks)**

Assignment 2

12. Using the Artificial Neural Network model from Question 11, classify test instances of the selected country separately for the two **Waves** having the **greatest** number of observations in your test data. How does the model performance change over the two waves? Use F1-scores, ROC curves and AUC to compare. **(4 Marks)**

Assignment 2

Report and Video Presentation (4 Marks)

Write a brief report (suggested length 8 – 10 pages) summarizing your results. Include your R script, copied and pasted as machine readable text, as an appendix. Use commenting in your R script, where appropriate, to help a reader understand your code. Alternatively combine working, comments and reporting in R Markdown. **(2 Marks)**

Record a short presentation using your smart phone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings, as well as describing how you conducted your research, and any assumptions made. Pay particular emphasis to your results for the investigative tasks. **(Submission Hurdle and 2 Marks)**

Assignment 2

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 2

Description of the data

Most fields are on integer scales over varying ranges. The convention is that larger numbers generally indicate greater agreement with statement or frequency of occurrence. Fields in bold at the end indicate binary class variables giving the level of confidence each participant has on the relevant social organization. “High confidence” (Confidence = 1), to “Low confidence” (Confidence = 0).

You can access more detail on each field in the WVS-7 Master Questionnaire 2017-2020 English.pdf, linked from <https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp> Use the question ID given in the **WVS Wave 7 Reference** in the table below.

Column Name	Original Descriptor	WVS Wave 7 Reference
Country	ISO 3166-1 alpha-3 country code	B_COUNTRY_ALPHA
Wave	Chronology of EVS-WVS waves	A_WAVE
ILFam	Important in life: Family	Q1
ILFriends	Important in life: Friends	Q2
ILLeisure	Important in life: Leisure time	Q3
ILPolitics	Important in life: Politics	Q4

Quiz and Practical Activity

- This will run in Week 11 in your applied session.
- Instructions, sample questions and formulas available under the Week 9 tab.
- Solutions visible as a separate document!

Quiz and Practical Activity



Faculty of
Information
Technology

FIT3152 Data analytics – 2026 On-Campus Quiz and Practical Activity Sample Answers

Your task	<ul style="list-style-type: none">• You will answer a set of multiple choice and longer questions in the form of a Moodle Quiz, on campus, during your applied session in Week 11. The questions cover topics taught during Weeks 1 – 9.
Value and Structure	<ul style="list-style-type: none">• This assignment is worth 30% of your total marks for the unit.• It has 30 marks in total, comprised of• 6 multiple choice questions of 1 Mark each,• 3 free responses of 2 Marks each, and• 3 grouped free responses of 6 Marks each.
Time	<ul style="list-style-type: none">• You will have 1 Hour during tutorial time to complete the test.
Due Date	Your scheduled applied session during Week 11
Submission	<ul style="list-style-type: none">• Via Moodle Quiz

Quiz and Practical Activity

Generative AI Use	<ul style="list-style-type: none">• Statement on Generative AI required by the university: In this assessment, you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task. AI must not be used in any way for producing the assessment task which requires demonstration of individual human capacities/skills/knowledge to meet the related unit learning outcomes.
Late Penalties	<ul style="list-style-type: none">• This activity can only be deferred/re-scheduled on medical or other serious grounds with relevant documentation.
Instructions	<ul style="list-style-type: none">• Attend your scheduled applied session in Week 11.• Answer the questions on the Moodle Quiz.• The activity is closed book, therefore lecture and tutorial notes or online references are not permitted.• You may use any calculator (physical or digital).• You must keep your camera on if you are in an online tutorial.
NOTE	<p>You will be asked to stop this activity early and submit what you have done if:</p> <ul style="list-style-type: none">• You are found to be using any software other than that permitted.• You are found to be accessing web sites or online resources other than the Moodle Quiz.• You are found to be communicating with any other student.• You are found using online resources besides the Moodle Quiz.• You are found to be cheating in any way.

Q&P 1 Mark question

Multiple Choice (1 Mark)

The table below shows a classification model for 10 customers based on whether or not they did **buy** a new product (did buy = 1, did not buy = 0), and the **confidence** level of the prediction.

Customer	Confidence-buy	Did-buy
C01	0.8823	0
C02	0.5547	0
C03	0.6469	1
C04	0.1252	0
C05	0.7050	0
C06	0.7065	1
C07	0.1441	0
C08	0.7398	1
C09	0.7865	1
C10	0.4874	0

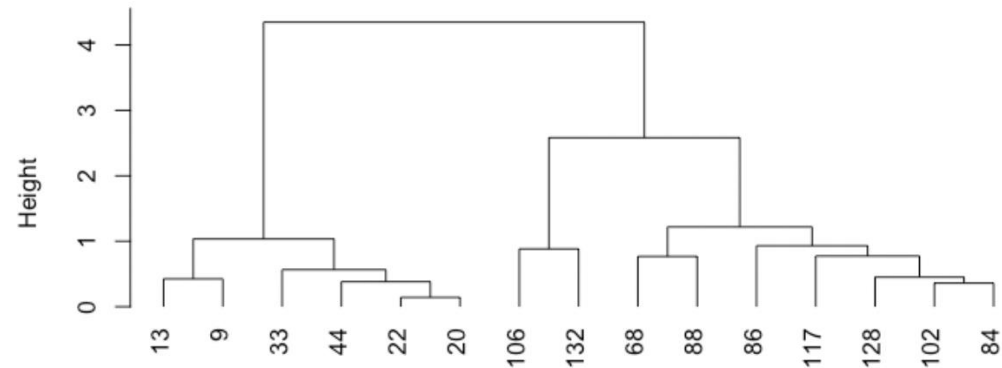
What is the lift value if you target the top 50% of customers that the classifier is most confident of?

- A. 0.2
- B. 0.5
- C. 1.5
- D. 2.0
- E. 2.5

Q&P 1 Mark question

Multiple Choice (1 Mark)

15 observations were sampled at random from the Iris data set. The dendrogram resulting from clustering, based on their sepal and petal measurements, is below.



What is the smallest number of clusters that would put all of species Setosa (observations 1:50) in a cluster their own.

- A. 1
- B. 2
- C. 3
- D. 5
- E. 15

Q&P 2 Mark question

Free Response (2 Marks)

Use the data below and Naïve Bayes classification to predict whether the following test instance will be happy or not.

Test instance: (Age Range = young, Occupation = professor, Gender = F, Happy = ?)

ID	Age Range	Occupation	Gender	Happy
1	Young	Tutor	F	Yes
2	Middle-aged	Professor	F	No
3	Old	Tutor	M	Yes
4	Middle-aged	professor	M	Yes
5	Old	Tutor	F	Yes
6	Young	Lecturer	M	No
7	Middle-aged	lecturer	F	No
8	Old	Tutor	F	No

Q&P 6 Mark question

Free Response (6 Marks)

The DunHumby (**DH**) data frame records the **Date** a **Customer** shops at a store, the number of **Days** since their last shopping visit, and amount **Spent** for 20 customers. The first 4 rows are shown below.

```
> head(DH)
  customer_id visit_date visit_delta visit_spend
      <int> <chr>         <int>      <dbl>
1          40 04-04-10         NA         44.8
2          40 06-04-10          2         69.7
3          40 19-04-10         13         44.6
4          40 01-05-10         12         30.4
```

The following R code is run:

```
rm(list = ls())
library(ggplot2)
DH <- read.csv("Dunnhumby1-20.csv")
head(DH)
attach(DH)
AS = as.table(by(visit_spend, customer_id, mean))
AS = sort(AS, increasing = TRUE)
AS = as.data.frame(tail(AS, 6))
colnames(AS) = c("Cust", "AveSp")
DH6 = DH[(DH$customer_id %in% AS$Cust),]
g = ggplot(DH6, aes(x = visit_delta, y = visit_spend))
g = g + geom_point() + facet_wrap(~customer_id, nrow = 2)
g
ggsave("Graph Data.pdf", g, width = 20, height = 12, unit = "cm")
```

Q&P 6 Mark question cont.

Describe the data contained in the data frame "AS." [2 Marks]

Describe the data contained in the data frame "DH6." [2 Marks]

Describe the contents of the graphic shown by plot "g." [2 Marks]

Q&P 6 Formulas A

The Visualization Zoo – Graphic Types

Time-Series Data

- Index Charts
- Stacked Graphs
- Small Multiples
- Horizon Graphs

Statistical Distributions

- Stem-and-Leaf Plots
- Q-Q Plots
- SPLOM
- Parallel Coordinates

Maps

- Flow Maps
- Choropleth Maps
- Graduated Symbol Maps
- Cartograms

Hierarchies

- Node-Link diagrams
- Adjacency Diagrams
- Enclosure Diagrams

Networks

- Force-Directed Layouts
- Arc Diagrams
- Matrix Views

Entropy

If S is an arbitrary collection of examples with a binary class attribute, then:

$$\begin{aligned} Entropy(S) &= -P_{C1} \log_2(P_{C1}) - P_{C2} \log_2(P_{C2}) \\ &= -\frac{N_{C1}}{N} \log_2\left(\frac{N_{C1}}{N}\right) - \frac{N_{C2}}{N} \log_2\left(\frac{N_{C2}}{N}\right) \end{aligned}$$

where $C1$ and $C2$ are the two classes.

P_{C1} and P_{C2} are the probability of being in Class 1 or Class 2 respectively. N_{C1} and N_{C2} are the number of examples in each class. N is the total number of examples.

$$\text{Note: } \log_2 x = \frac{\log_{10} x}{\log_{10} 2} = \frac{\log_{10} x}{0.301}$$

Information gain

The $Gain(S, A)$ of an attribute A relative to a collection of examples, S , with v groups having $|S_v|$ elements is:

$$Gain(S, A) = Entropy(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} * Entropy(S_v)$$

Q&P 6 Formulas B

Networking

Closeness Centrality: $C_{CL}(v) = \frac{1}{\sum_{u \in V} dist(u,v)}$

Betweenness Centrality: $C_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$,

where $\sigma(s,t)$ is the number of shortest paths between s and t .

$\sigma(s,t|v)$ is the number of shortest paths between s and t passing through v

Density: $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2}$,

where $|E_g|$ is number of edges, $|V_g|$ is number of vertices

Clustering coefficient: $clt(g) = \frac{3\tau_{\Delta}(g)}{\tau_3(g)}$,

where $3\tau_{\Delta}(g)$ is number of triangles, $\tau_3(g)$ is number of connected triples

Naïve Bayes'

For events A_1, A_2, \dots, A_n and event C , the classification probability is

$$P(C_j | A_1 \cap A_2 \dots \cap A_n) = \frac{P(C_j) \cdot P(A_1 \cap A_2 \dots \cap A_n | C_j)}{P(A_1 \cap A_2 \dots \cap A_n)}$$

For Bayesian classification, a new point is classified to C_j if $P(C_j) * P(A_1 | C_j) * P(A_2 | C_j) * \dots * P(A_n | C_j)$ is maximised.

Naïve Bayes assumes $P(A \cap B) = P(A) * P(B)$ etc.

Accuracy

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

ROC

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

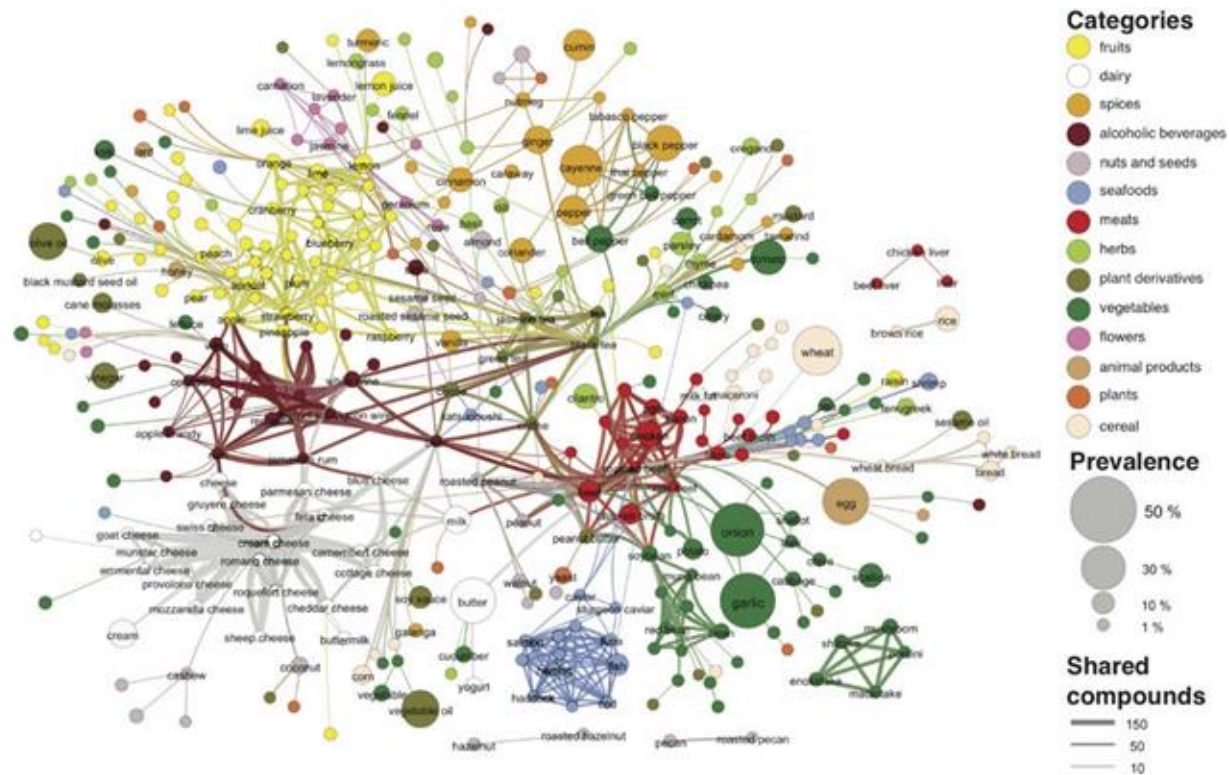
Introduction to networks

We'll cover:

- Introduction to networks
- Types of networks
- Network structure
- Network statistics
- Centrality measures
- Hand calculations

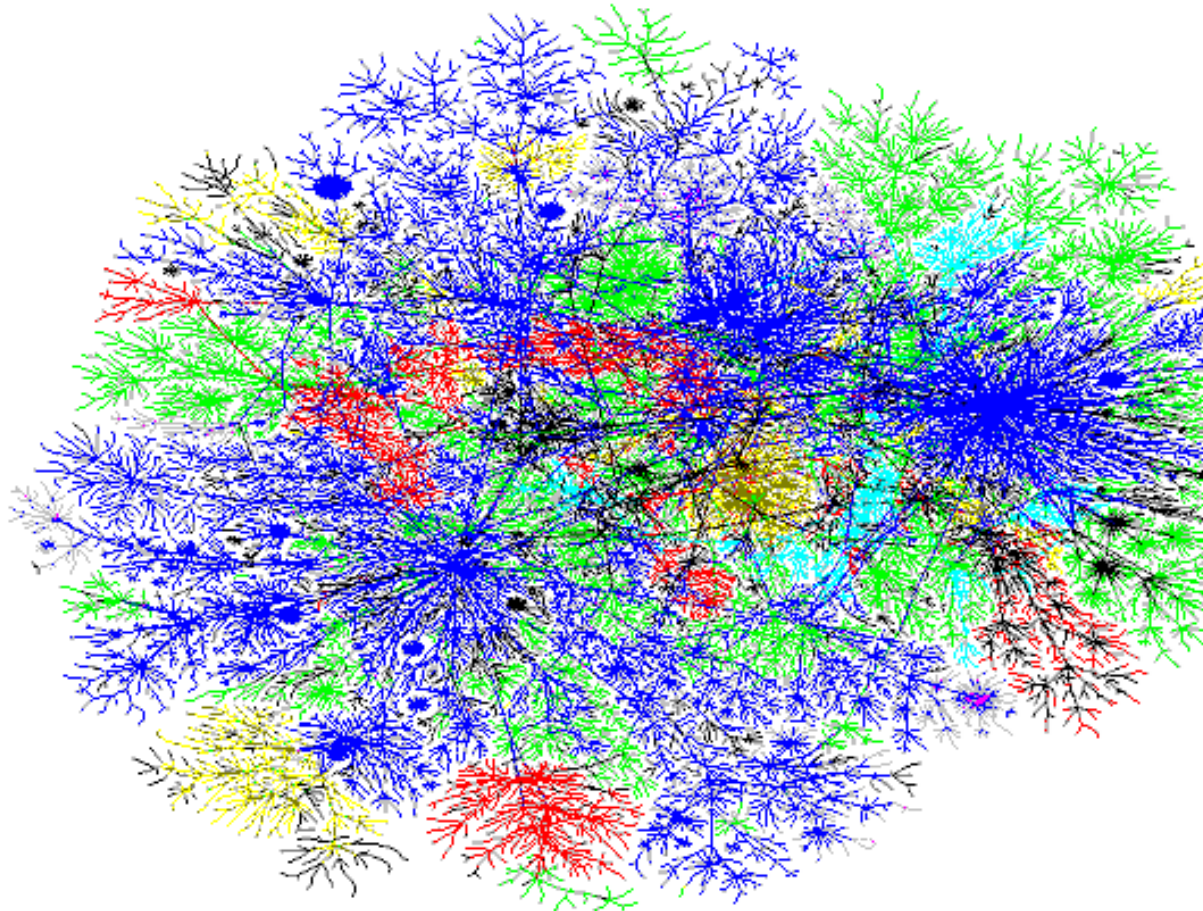
Flavor network

Flavor network and the principles of food pairing



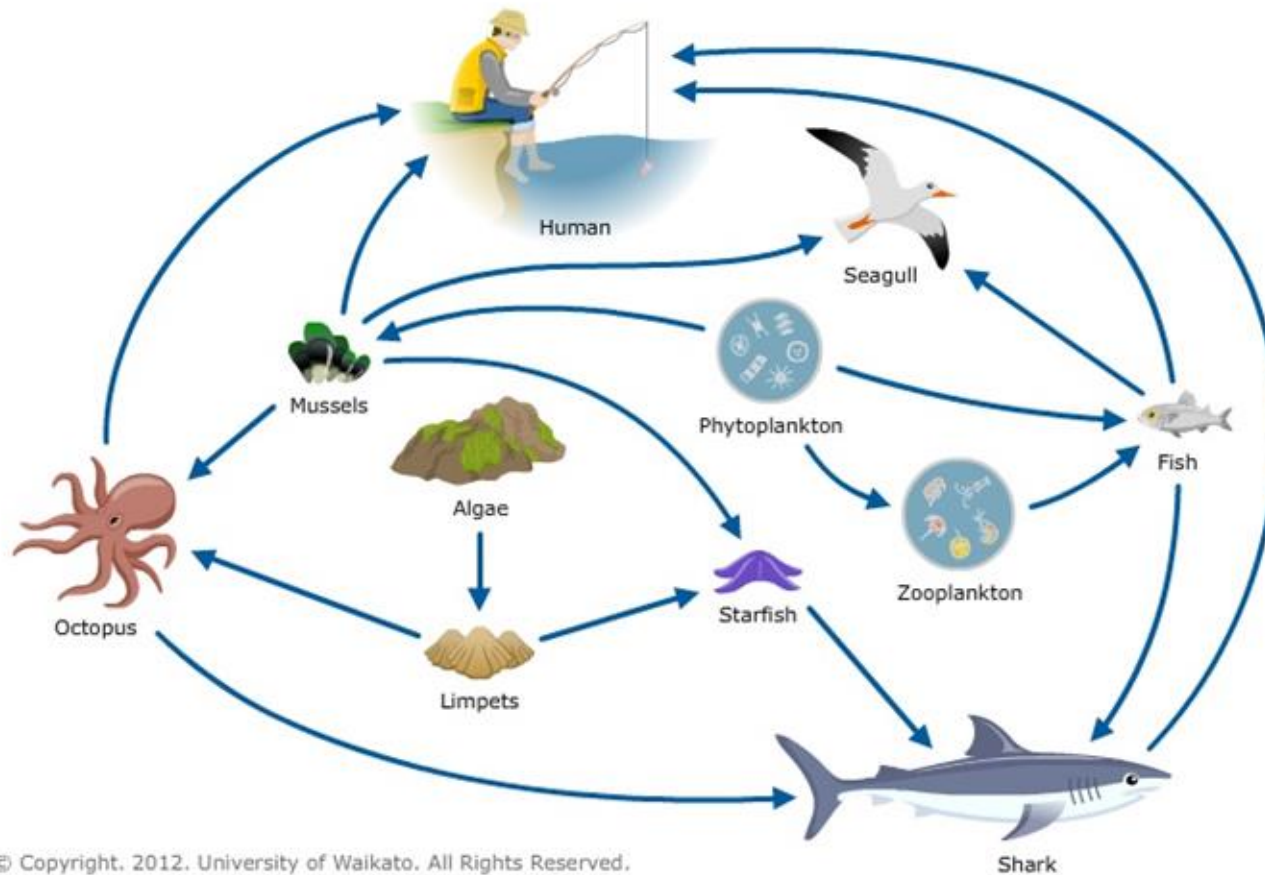
<https://www.nature.com/articles/srep00196>

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



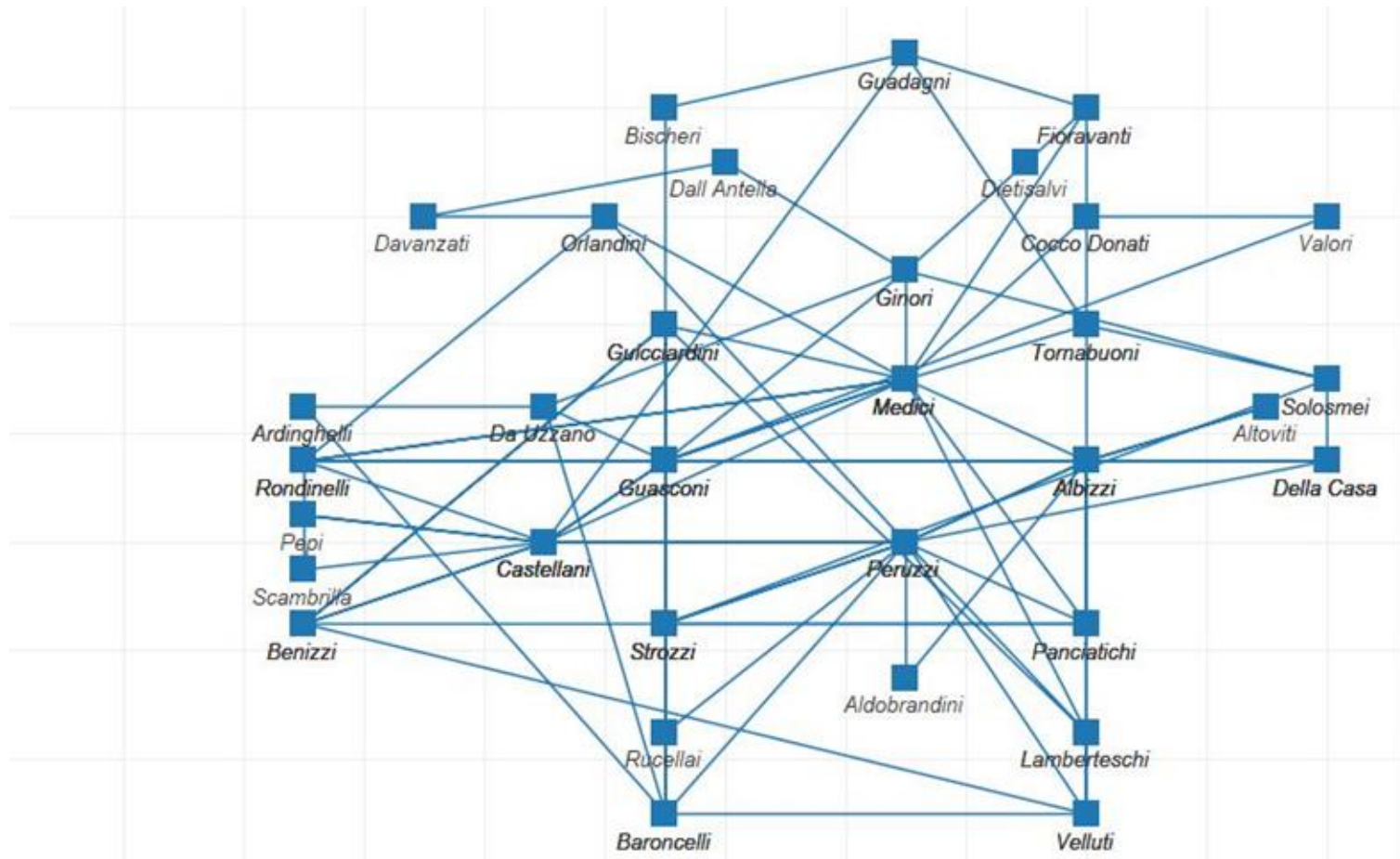
<https://www.ptv.vic.gov.au/>

Social networks



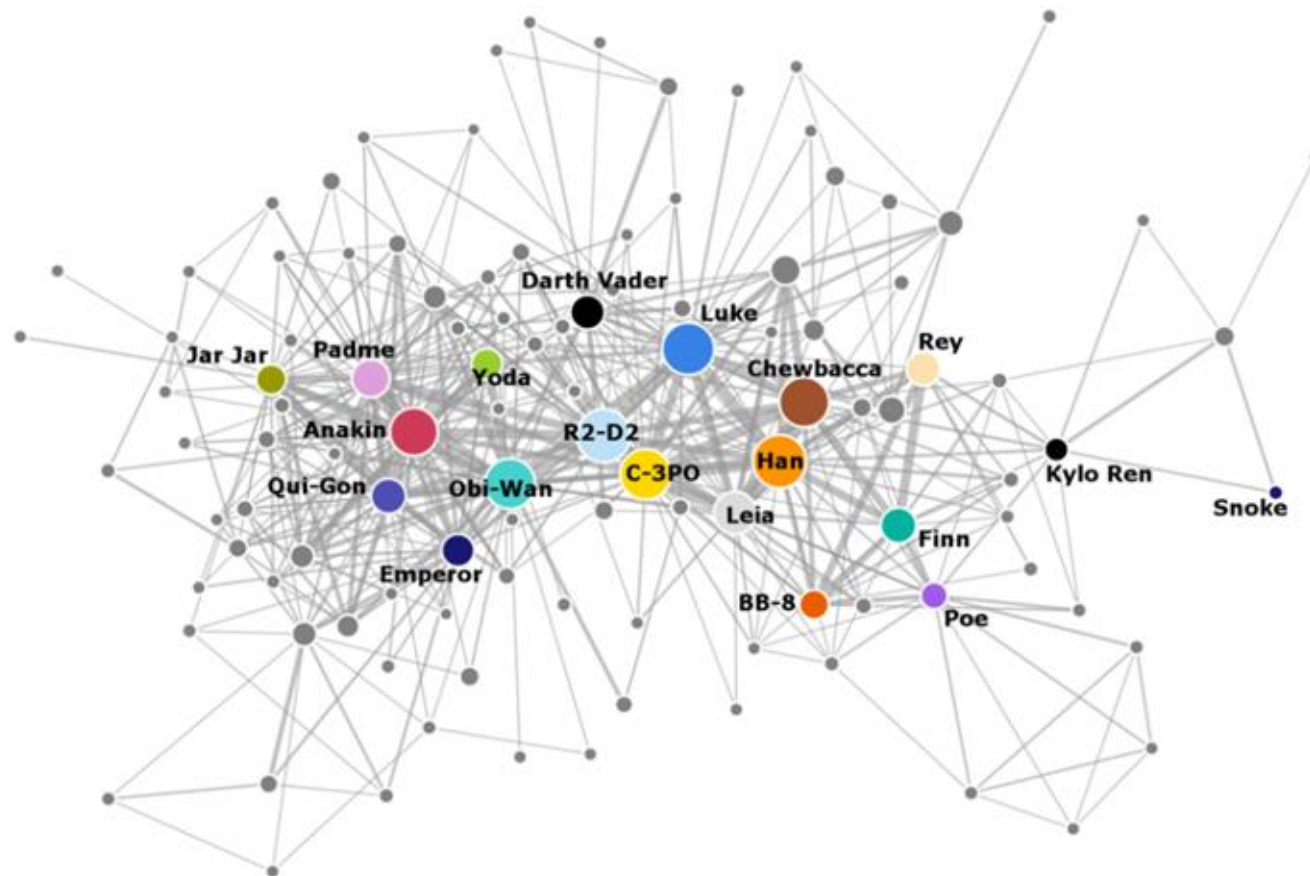
<https://socialadr.com/features>

Florentine social network: Medici



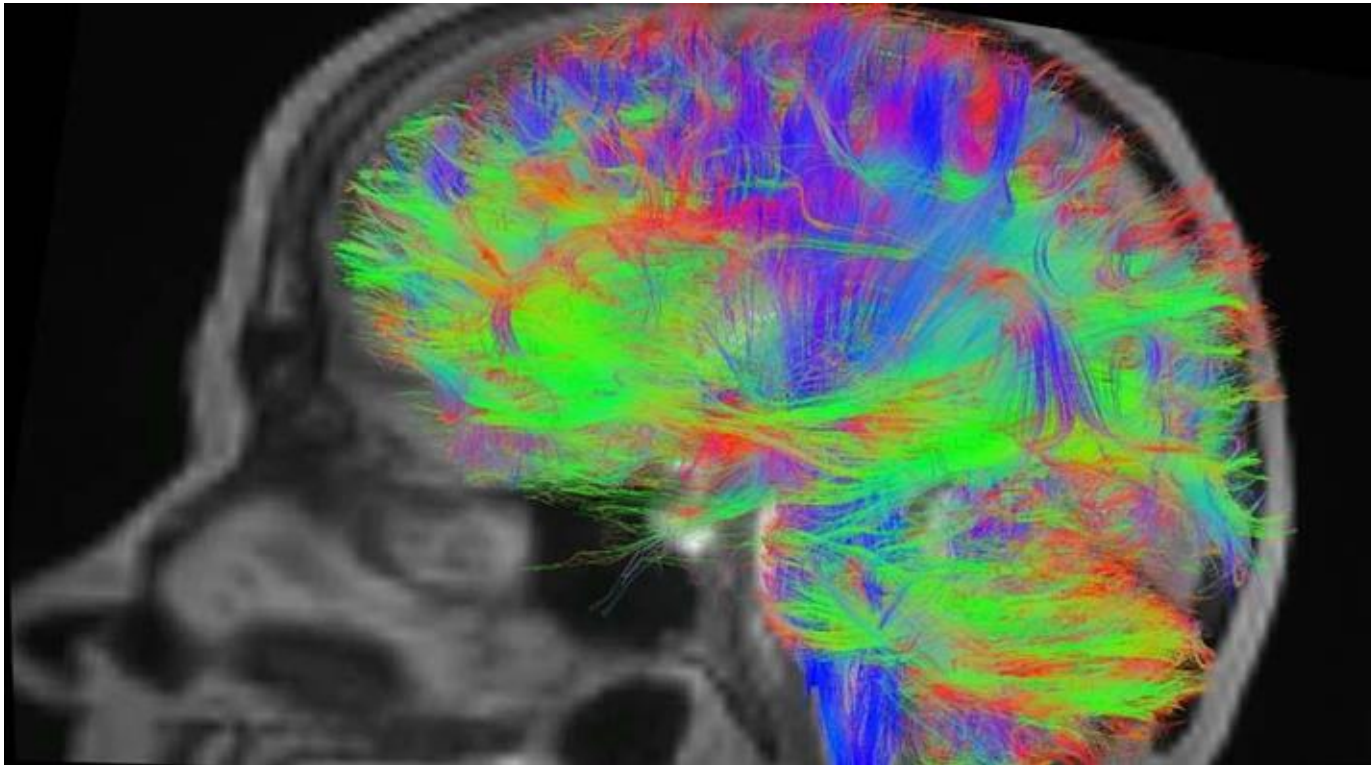
(inactive) <http://blogs.teradata.com/international/tag/data-scientist/>

Star Wars characters



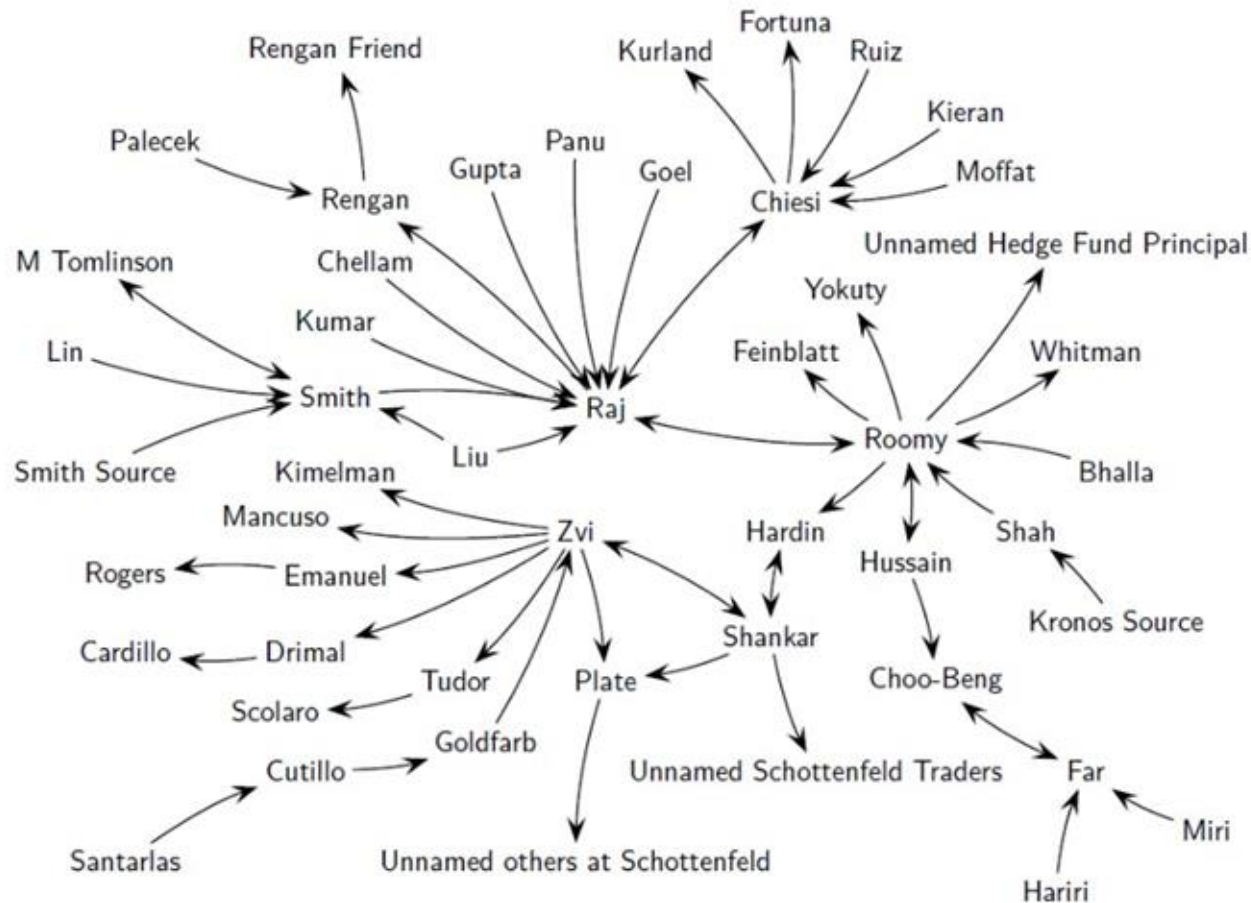
<http://evelinag.com/blog/2016/01-25-social-network-force-awakens/#.WZwWUpOg8cl>

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/>

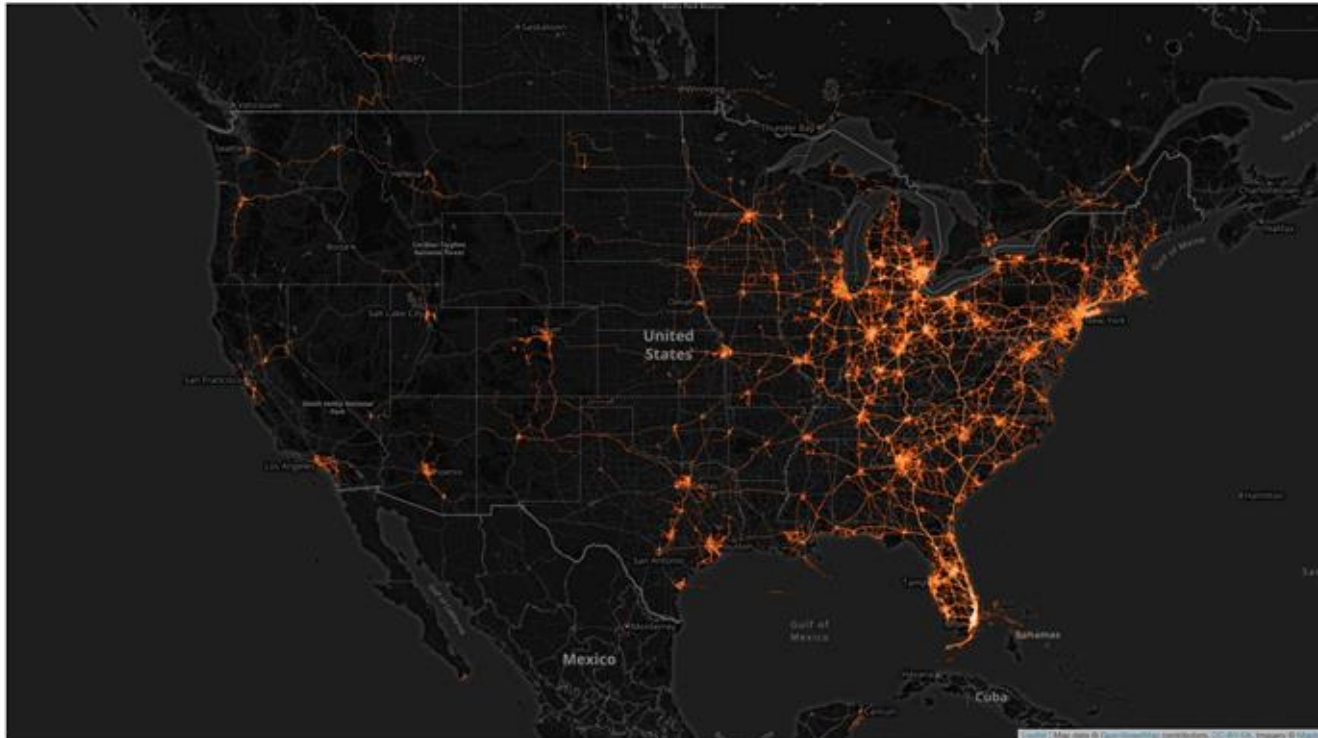
COVID-19 US social network

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>

You probably know 611 people...

You Probably Know 611 People. Here's How We Know.

Asking how many Michaels you know unlocks all sorts of knotty social and demographic puzzles



By *Josh Zumbrun* [Follow](#)

Nov. 16, 2023 9:00 pm ET



<https://www.wsj.com/us-news/you-probably-know-611-people-heres-how-we-know-88dd27d9>

How many people do you know? You've probably never counted. Well, now you don't have to. Tyler McCormick has worked it out: around 600.

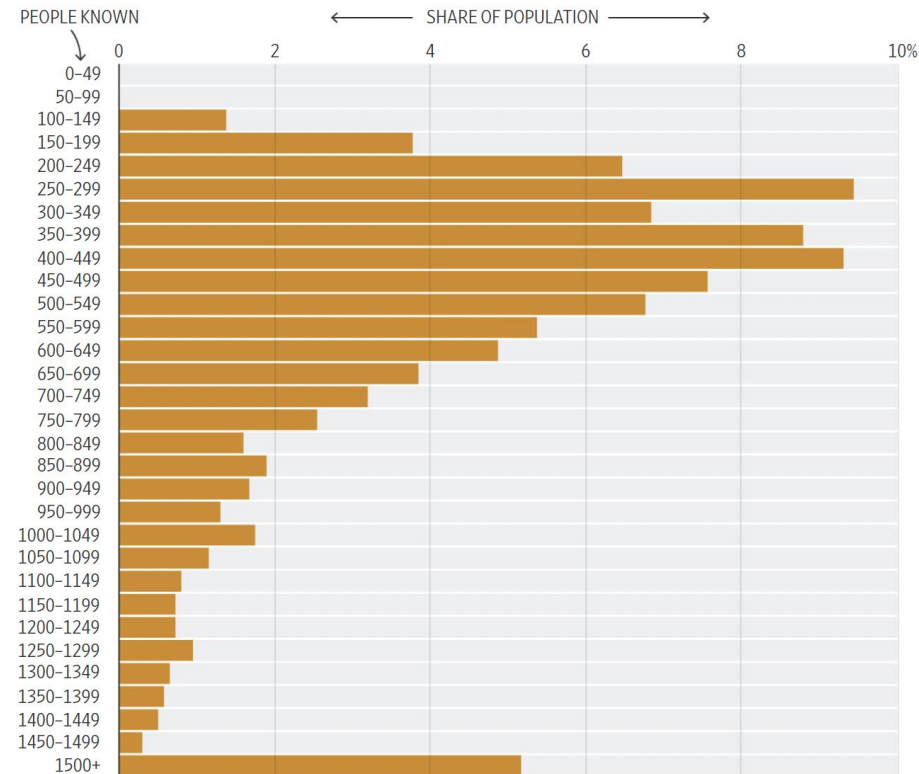
Or more precisely 611, according to estimates by McCormick, a professor in the statistics and sociology departments at the University of Washington. That's a national average, but McCormick can actually compute an estimate for you, or anyone. His technique is a fascinating illustration of the power of statistics to illuminate society: not just how many acquaintances the average person has, but the number of homeless and other hard-to-reach populations.

<https://www.wsj.com/us-news/you-probably-know-611-people-heres-how-we-know-88dd27d9>

You probably know 611 people...

People Who Know People

The median person knows 472 other people. On average, people know 611 others, driven in part by a small number of people with huge numbers of acquaintances.



<https://www.wsj.com/us-news/you-probably-know-611-people-heres-how-we-know-88dd27d9>

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. Why some networks are fragile (break easily) and others are robust.

How to think about a network

For the previous examples, think about:

Elements in the network:

- Same?
- Different?

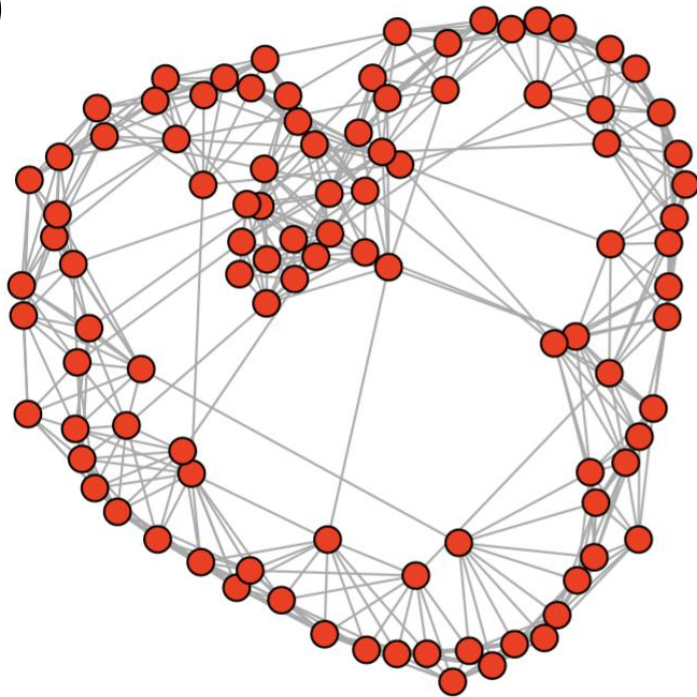
Connections between elements:

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

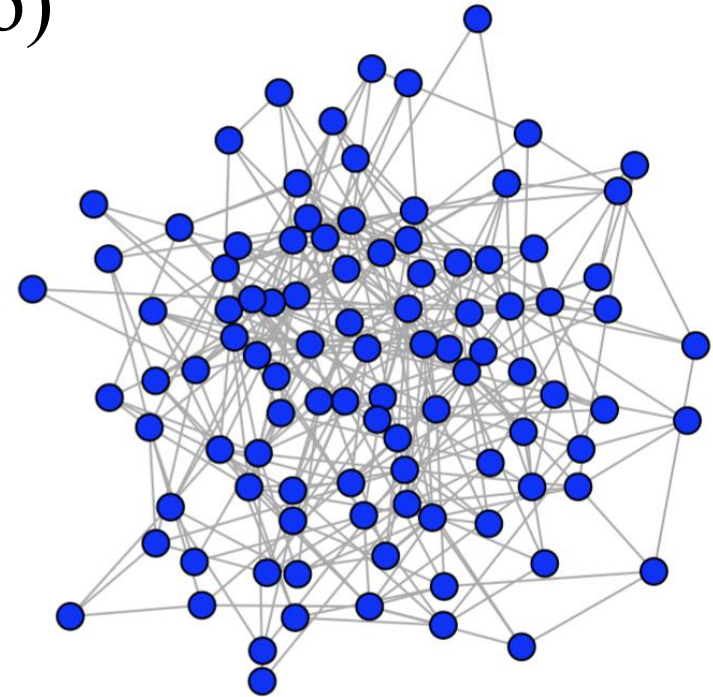
How to think about a network (a)

How is the structure of these networks different?

(a)



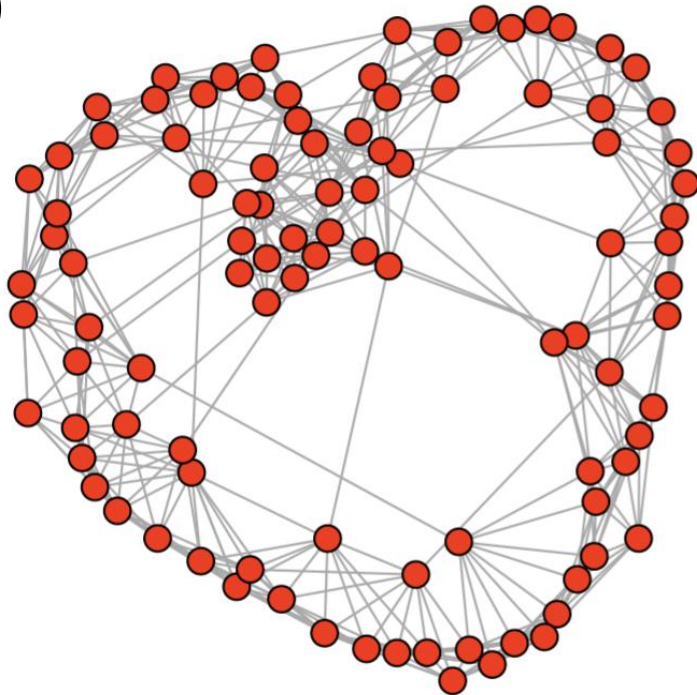
(b)



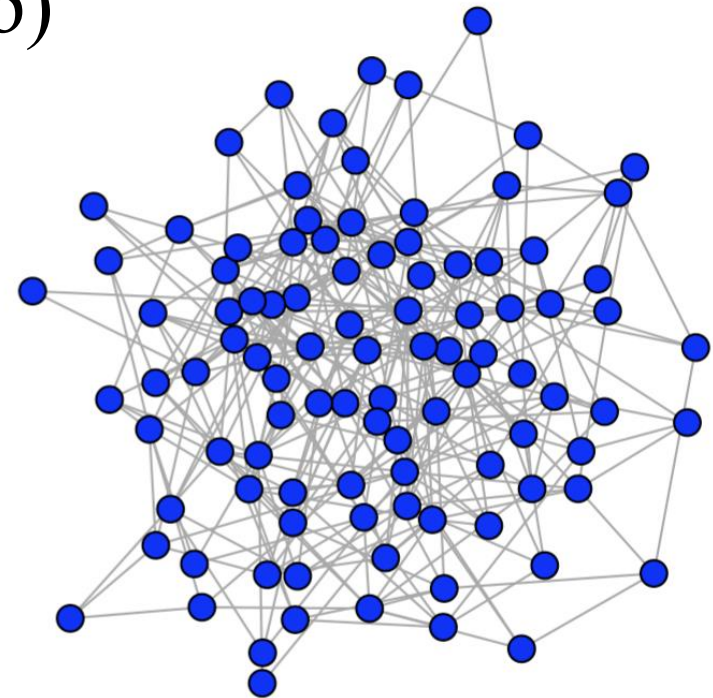
How to think about a network (b)

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.

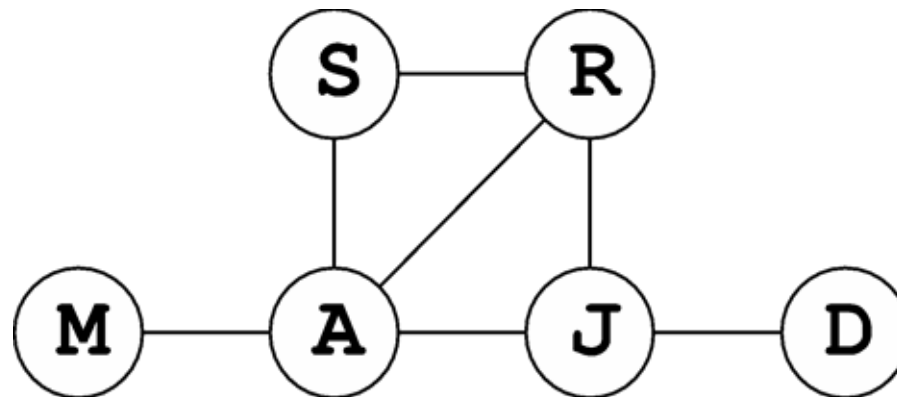
Questions to answer:

- 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 $\langle == \rangle$ Friend B or directed, e.g., Parent $== \rangle$ 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...

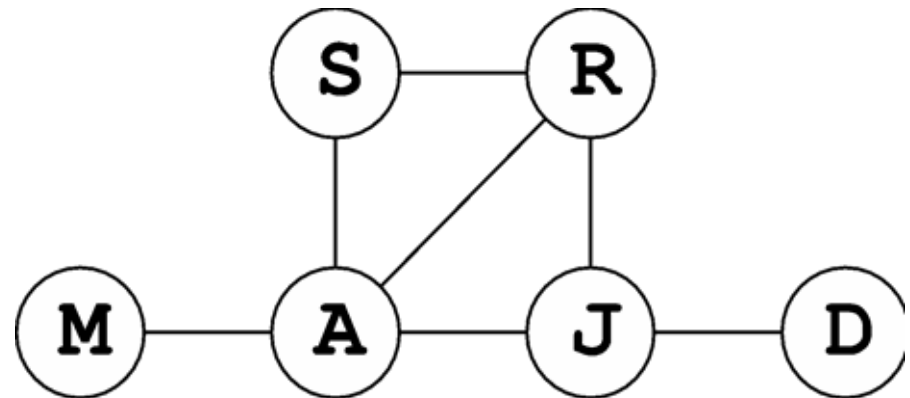
Particular link sequences have formal descriptions

- *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:

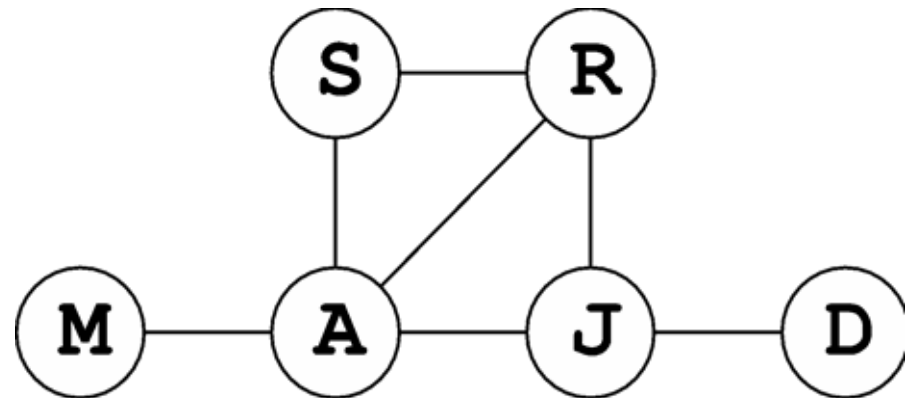
- *Walk*: (M, A, J, R, J, ...)
- *Path*: (M, A, J, R)
- *Cycle*: (A, S, R, A)
- *Geodesic*: $\text{Geodesic}(R, D) = (R, J, D)$
- *Length*: $\text{dist}(M, D) = 3$
- *Connected*: yes
- *Clique*: A, S, R
- *Simple*: yes.



Adjacency matrix

We can represent the network as a matrix indicating the connections between individuals:

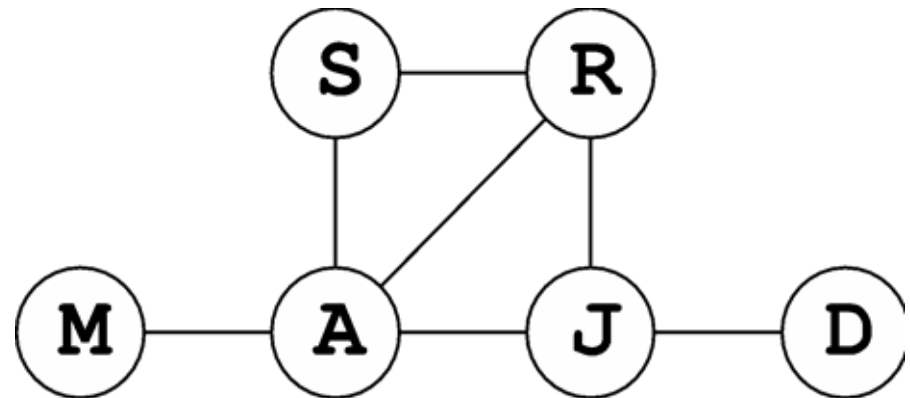
	J	A	S	D	R	M
J	0	1	0	1	1	0
A	1	0	1	0	1	1
S	0	1	0	0	1	0
D	1	0	0	0	0	0
R	1	1	1	0	0	0
M	0	1	0	0	0	0



Degree of a vertex

Arguably the single most important measure of the importance of a node (vertex) 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.

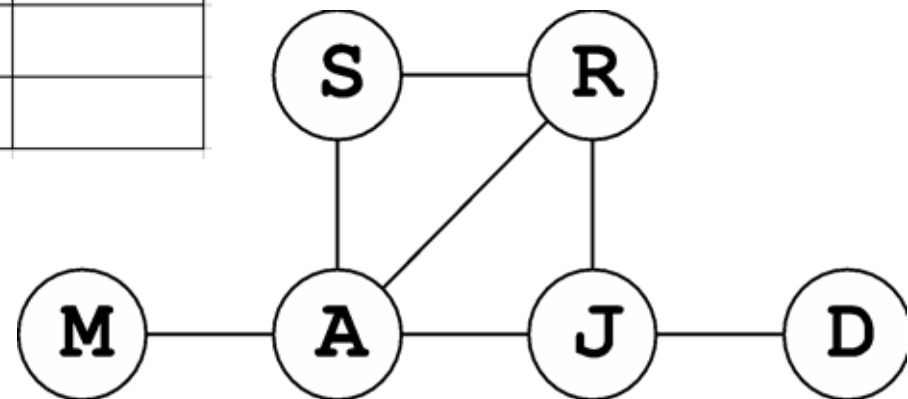
Analysis of these and other factors indicate how *connected, robust or fragile* etc. a network is.

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



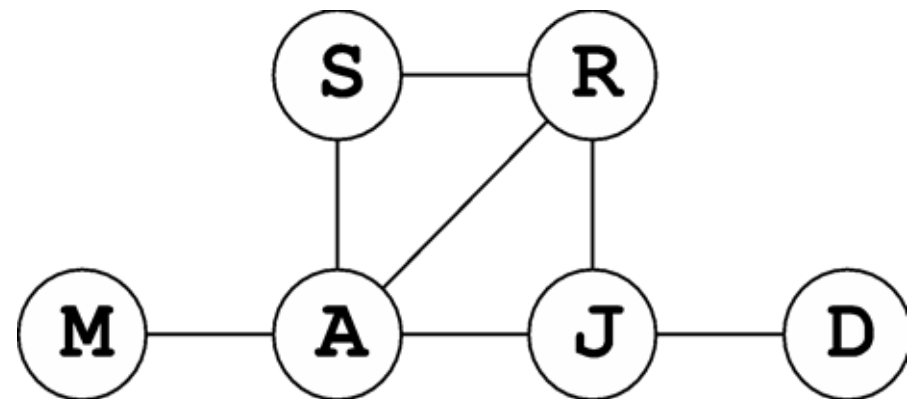
Analysing the network

Example from the research collaborators network:

- *Diameter*: $= \max(\text{dist}(u,v)) = 3$
- *Degree distribution*: \longrightarrow
- *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.

- $$clt(g) = \frac{3\tau_{\Delta}(g)}{\tau_3(g)} \quad \text{Also known as transitivity}$$

- where $\tau_{\Delta}(g)$ number of triangles, $\tau_3(g)$ is number of triples

Analysing the network

For the research collaborators:

- *Density*: $den(g) = \frac{|E_g|}{|V_g|(|V_g|-1)/2} = \frac{7}{(6 \times 5)/2} = 0.467$
- *Clustering coefficient*: $clt(g) = \frac{3\tau_{\Delta}(g)}{\tau_3(g)} = \frac{3 \times 2}{13} = 0.462$

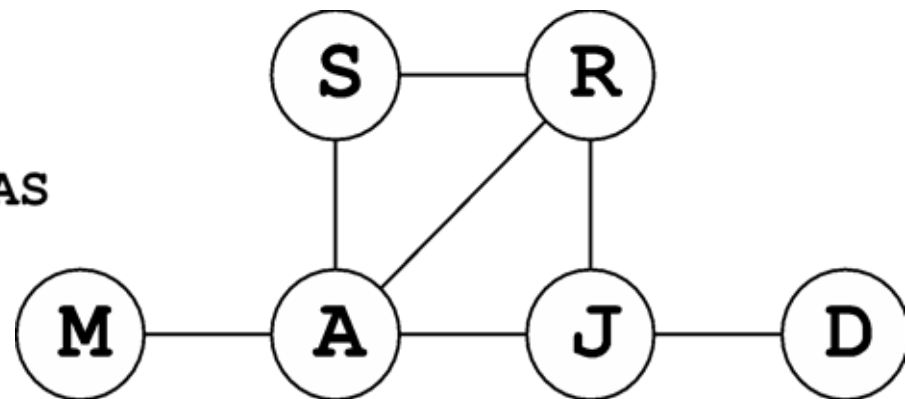
(triples)

(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*

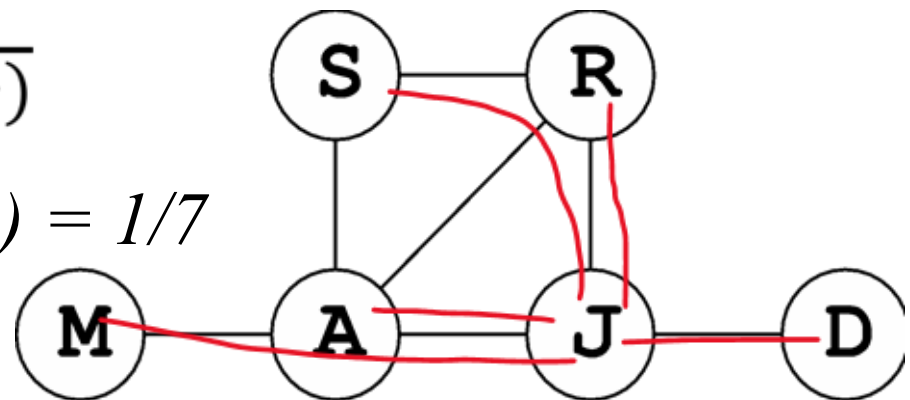
Closeness centrality

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 (shortest) distance between a vertex and the others.

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

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



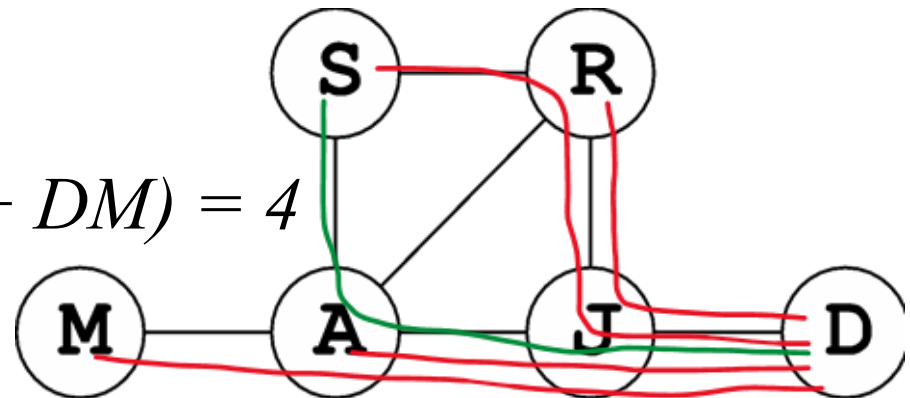
Betweenness centrality

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

- Betweenness centrality is the number of shortest paths between s and t $\sigma(s,t)$ that pass-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). See Kolaczyk and Csárdi (in references), Section 4.2.

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/>

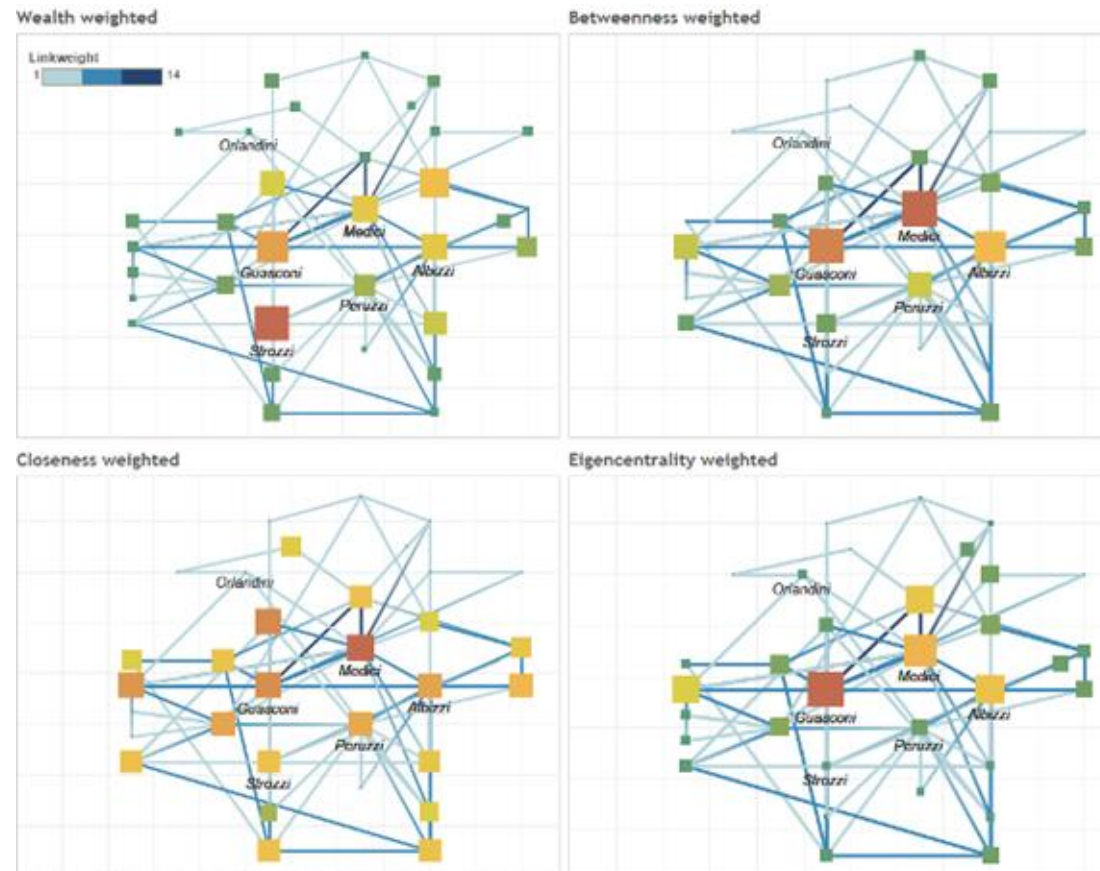
Which centrality measure?

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

- Betweenness 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.
- Eigencentrality: 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/>

SNA and the making of a new pope

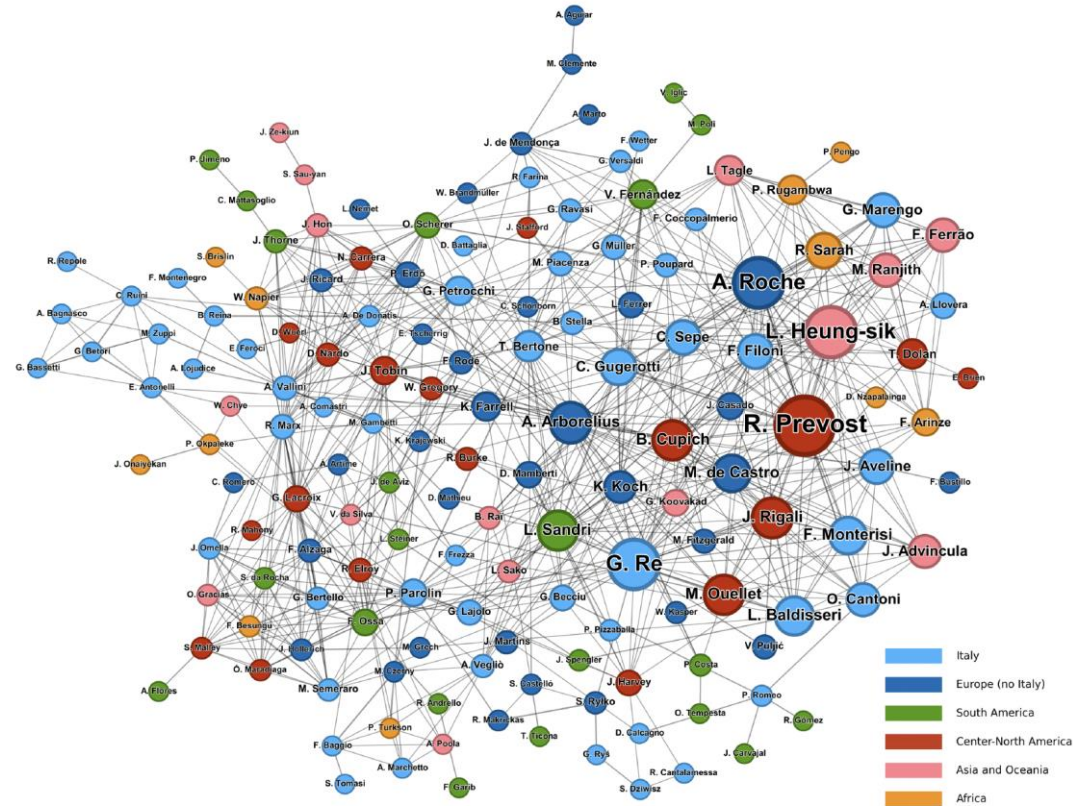


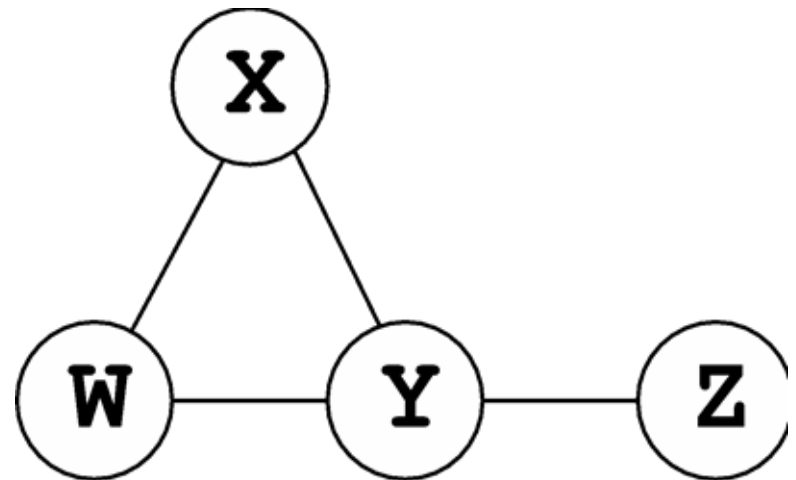
Figure A4: Multiplex College of Cardinals Network – Geographical Split

<https://arxiv.org/html/2505.17635v1>

Class example

For the graph below calculate the following:

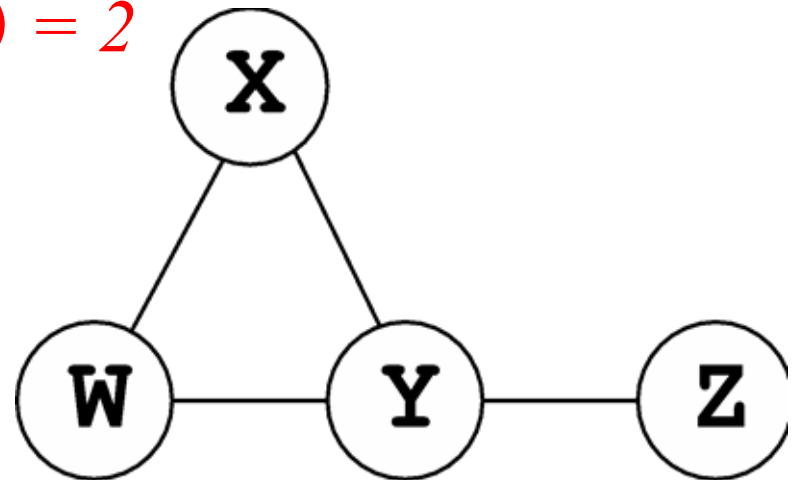
- *Average path length*
- *Diameter*
- d_Y
- *Degree distribution*
- $c_B(Y)$
- $c_{Cl}(Y)$
- *Cliques*



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* $2: W-X, W-Y, X-Y, Y-Z, 3: W-X-Y$



Network analysis in R

We'll cover:

- Creating a network in R
- Network statistics and analysis
- Community detection

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 or other online resources.

Creating a graph from data frame

Create a data frame and then convert it to a graph object. For John's research collaborators:

```
> 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))
```

Convert to a graph object:

```
> g <- graph_from_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 (mean distance),
clique size:

```
> diameter(g)
```

```
[1] 3
```

```
> mean_distance(g)
```

```
[1] 1.666667
```

```
> table(sapply(cliques(g),length))
```

```
1 2 3
```

```
6 7 2
```

Graph summary

Density and clustering coefficient:

```
> edge_density(g)
```

```
[1] 0.467
```

```
> transitivity(g)
```

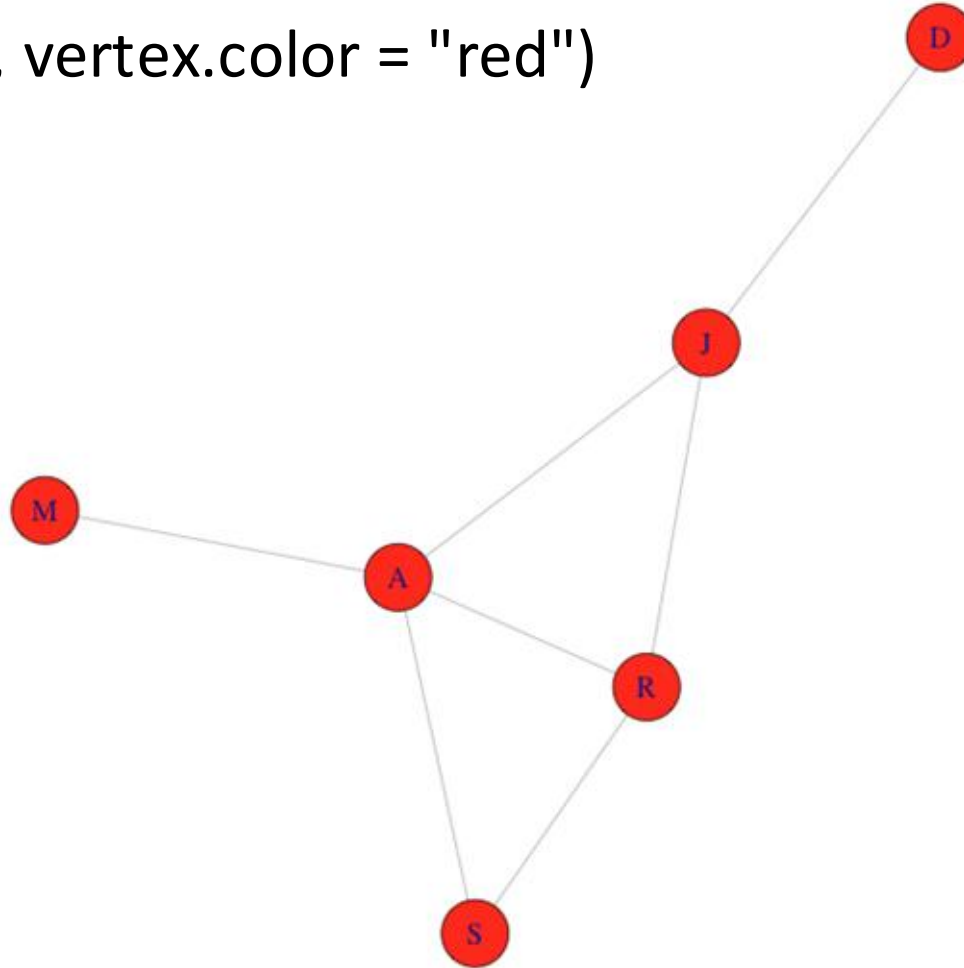
```
[1] 0.462
```

Adjacency matrix

```
> as_adjacency_matrix(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 . . . .
```

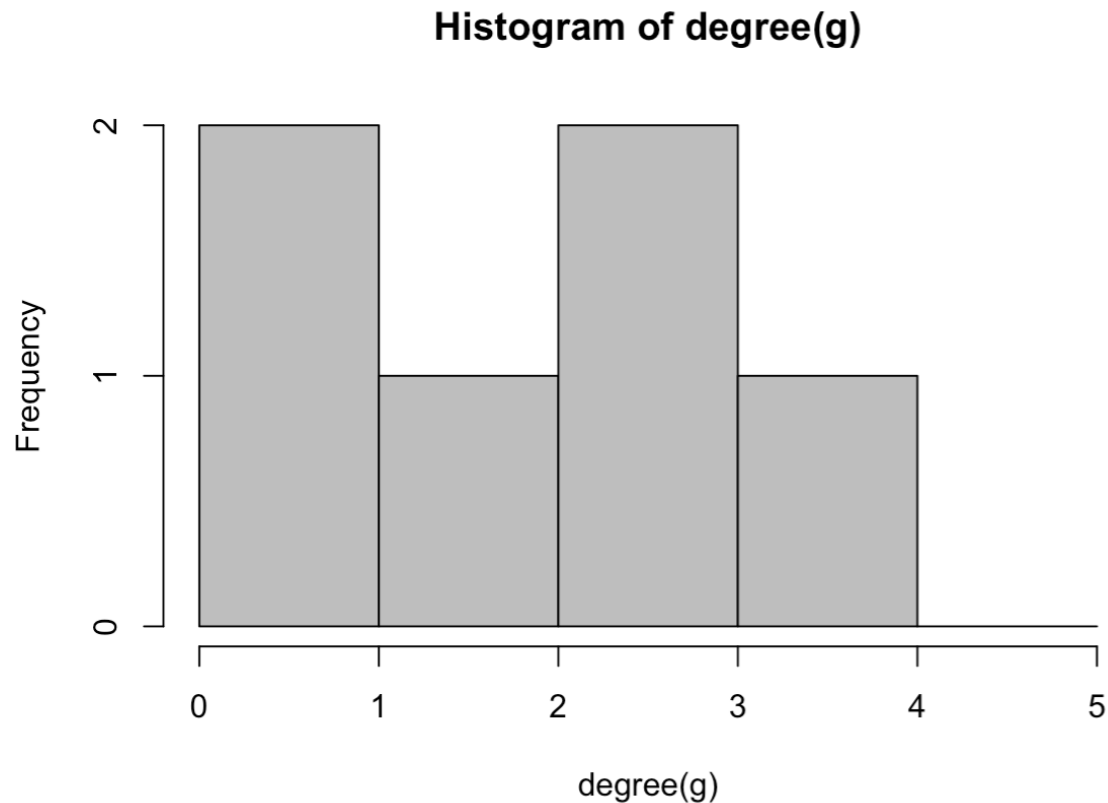
Basic plot

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



Degree distribution

> `hist(degree(g), breaks = c(0,1,2,3,4,5), 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 calculate the measures and then format the output.

```
> e = eigen_centrality(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 research collaborators 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.

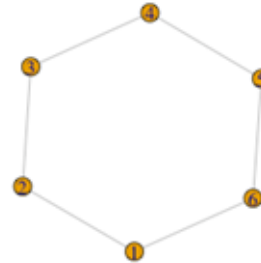
Four important network topologies

Using R:

Complete



Ring



Tree



Star



Four important network topologies

- > # Code for graphs on prev. slide 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_from_adjacency_matrix(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

For reading only



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.

Two-mode network



- The attendance of 18 women at social events in the 1930s is a famous two-mode network.

NAME OF PARTICIPANTS OF GROUP I	CODE NUMBERS AND DATES OF SOCIAL EVENTS REPORTED IN <i>Old City Herald</i>													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson.....	x	x	x	x	x	x	x	x	x					
2. Miss Laura Mandeville.....	x	x	x	x	x	x	x	x	x					
3. Miss Theresa Anderson.....		x	x	x	x	x	x	x	x					
4. Miss Brenda Rogers.....	x		x	x	x	x	x	x						
5. Miss Charlotte McDowd.....			x	x	x	x	x							
6. Miss Frances Anderson.....			x		x	x	x							
7. Miss Eleanor Nye.....					x	x	x							
8. Miss Pearl Oglethorpe.....						x	x	x	x					
9. Miss Ruth DeSand.....					x		x	x	x					
10. Miss Verne Sanderson.....							x	x	x			x		
11. Miss Myra Liddell.....							x	x	x	x		x		
12. Miss Katherine Rogers.....							x	x	x	x		x	x	x
13. Mrs. Sylvia Avondale.....							x	x	x	x		x	x	x
14. Mrs. Nora Fayette.....						x	x	x	x	x		x	x	x
15. Mrs. Helen Lloyd.....							x	x	x	x		x		
16. Mrs. Dorothy Murchison.....								x	x	x				
17. Mrs. Olivia Carleton.....									x	x		x		
18. Mrs. Flora Price.....										x				

- See Applied Session 9, Question 8.
- Ideas for creating the network in igraph follow.

Creating cliques 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



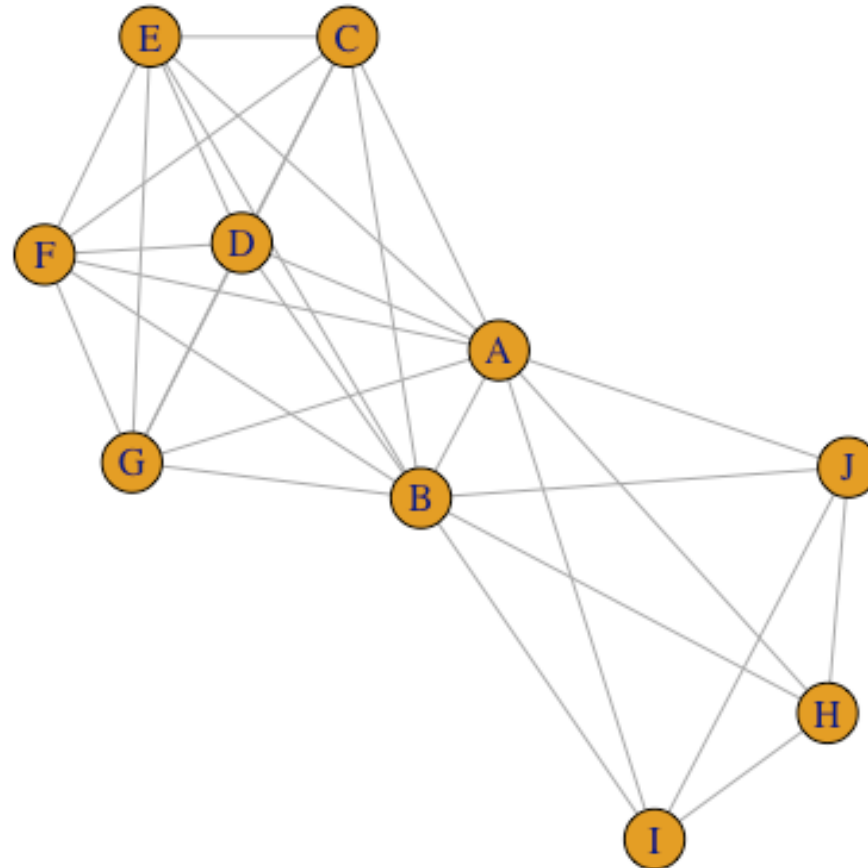
You can merge two graphs to form a new graph:

- > # 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:

	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

	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)

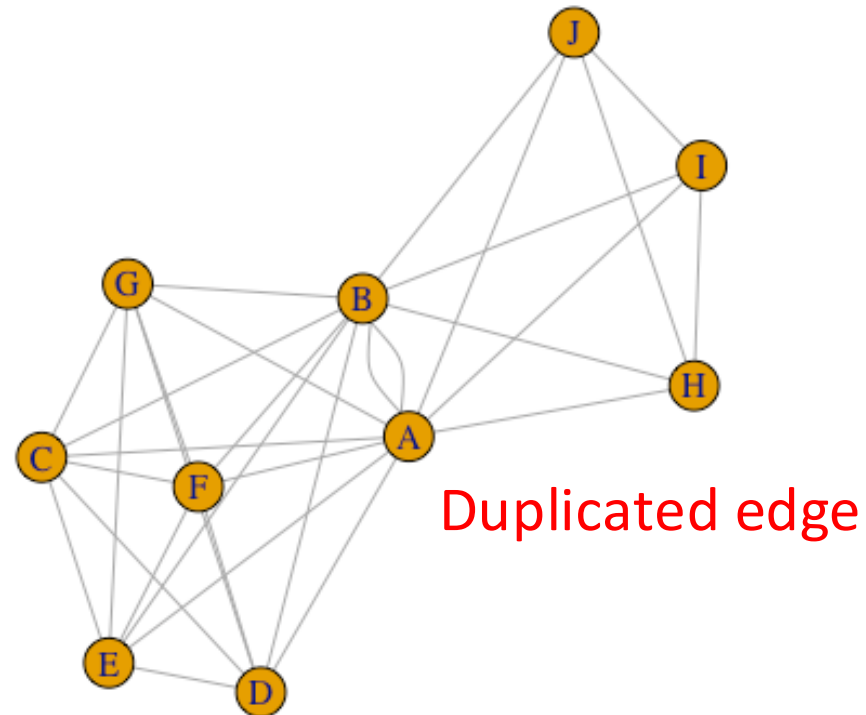
```
> # loop through each group
> for (k in unique(ClubData$Club)){
>   temp = ClubData[(ClubData$Club == k),]
>   # combine each pair of agents to make an edge list
>   Edgelist = as.data.frame(t(combn(temp$Person,2)))
>   colnames(Edgelist) = c("P1", "P2")
>   # loop through pairs of edges and add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

Coding directly



Plot, and simplify to remove duplicated edge

- > plot(g)
- > g = simplify(g)
- > plot(g)



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.

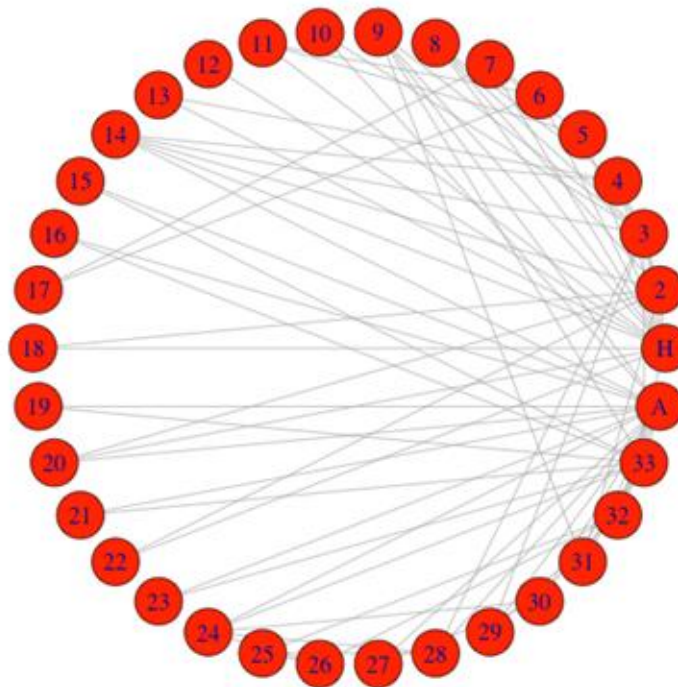
W. Zachary, An information flow model for conflict and fission in small groups.
J. Anthropol. Res. 33 (4), 452–473 (1977)

Karate club – looking at the data

- > library(igraph)
- > library(igraphdata)
- > **data(karate) # we use an inbuilt data set**
- > diameter(karate)
[1] 13
- > mean_distance(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")`



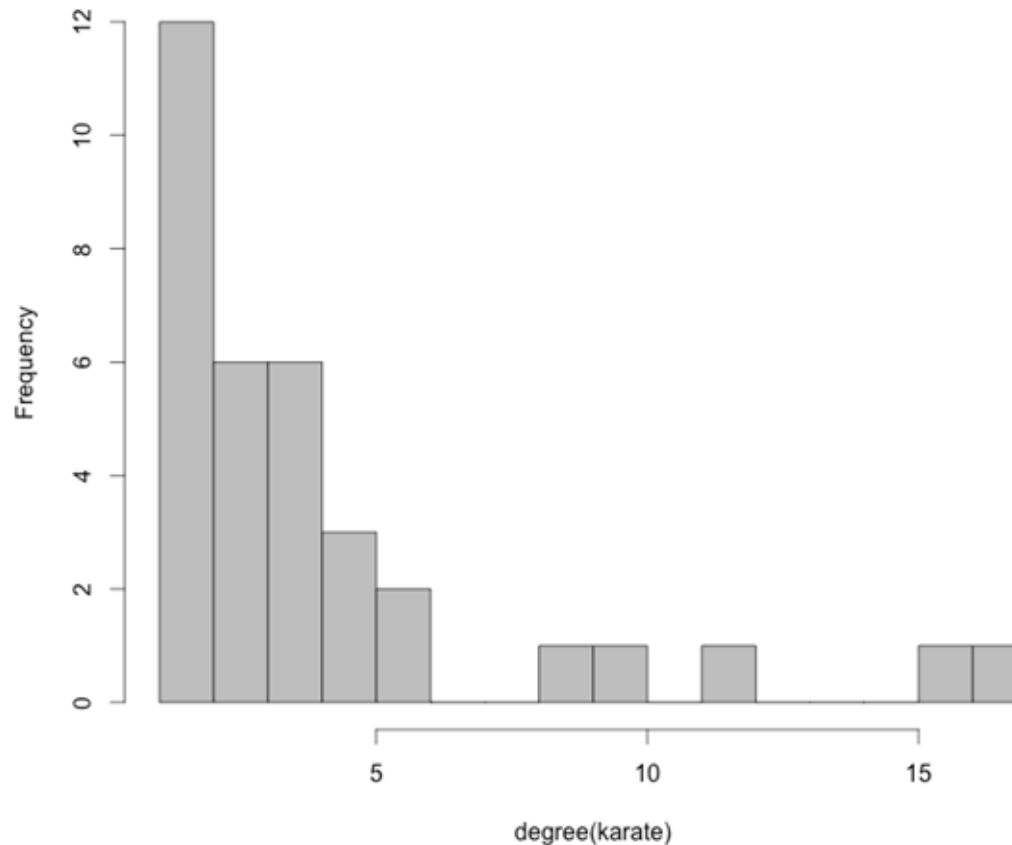
Karate club – force-directed plot

> `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

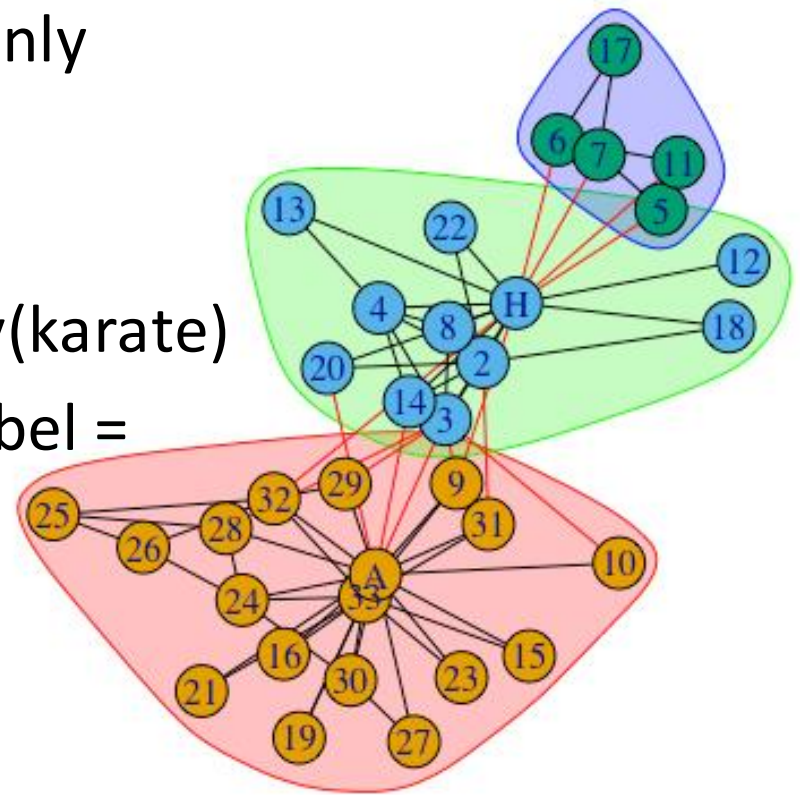
iGraph enables the detection of communities (or subgroups) in a social network.

- 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

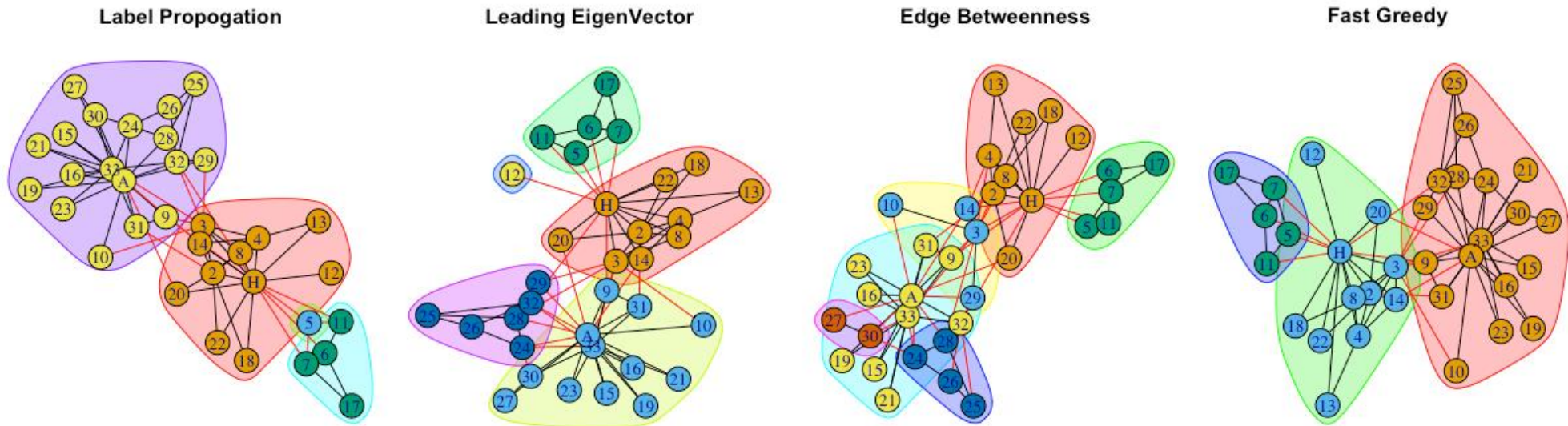
We can get another view of the club by clustering

- > # 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

- > #for prev. slides: 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 Propagation")`

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

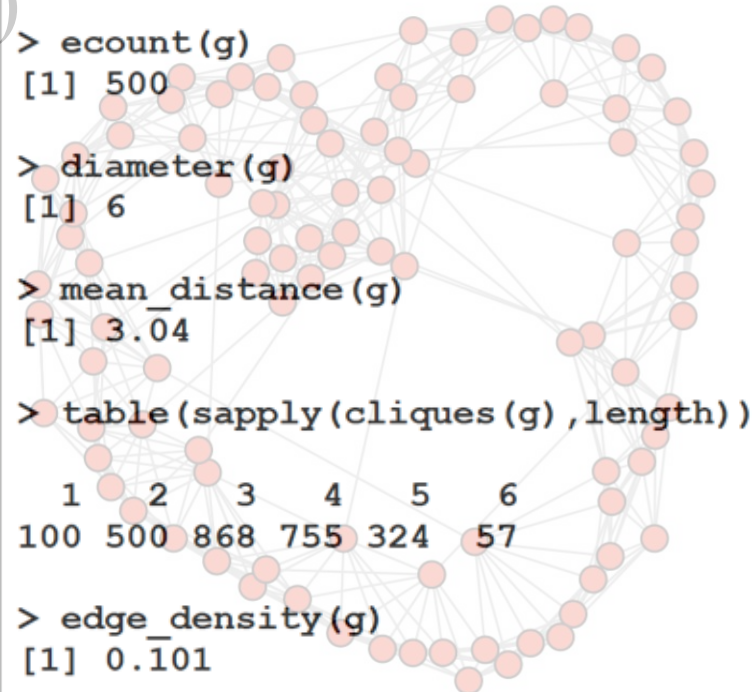
- > library(igraph)
- > library(igraphdata)
- > data(karate)
- > plot(karate, layout = layout.circle, vertex.color = "red")
- > plot(karate, layout = layout.fruchterman.reingold, vertex.color = "red")
- > hist(degree(karate), breaks = 18, col = "grey")
- > d = degree(karate)
- > c = format(closeness(karate),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

Recap: How to think about a network

How is the structure of these networks different?

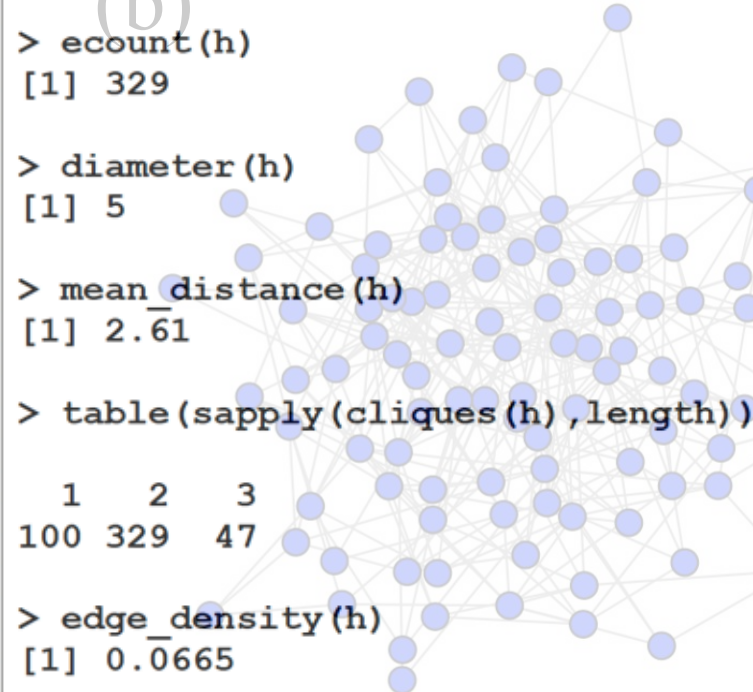
(a)

```
> # RED
> vcount(g)
[1] 100
> ecount(g)
[1] 500
> diameter(g)
[1] 6
> mean_distance(g)
[1] 3.04
> table(sapply(cliques(g), length))
  1   2   3   4   5   6
100 500 868 755 324  57
> edge_density(g)
[1] 0.101
> transitivity(g)
[1] 0.576
```



(b)

```
> # BLUE
> vcount(h)
[1] 100
> ecount(h)
[1] 329
> diameter(h)
[1] 5
> mean_distance(h)
[1] 2.61
> table(sapply(cliques(h), length))
  1   2   3
100 329  47
> edge_density(h)
[1] 0.0665
> transitivity(h)
[1] 0.0663
```



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, community detection, for example.*

More theoretical...

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

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

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 2020. Chapters 1 – 4 especially.
- *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>