Network analysis

Understanding Society with Big Data: Computational Social Science (CSS)

Tod Van Gunten/Björn Ross

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

Challenge of The Course

- Every Week:
  Understanding complex social phenomena using big data
- This Week:
  Using data on social relationships and interactions

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

Overview

- Why study social networks?
- Key concepts:
  - Social influence
  - Network structure
  - Small worlds
  - Homophily
- Methods and approaches:
  - Centrality
  - Finding subgroups: community detection
  - How to do it: using Gephi

THE UNIVERSITY *of* EDINBURGH
informatics

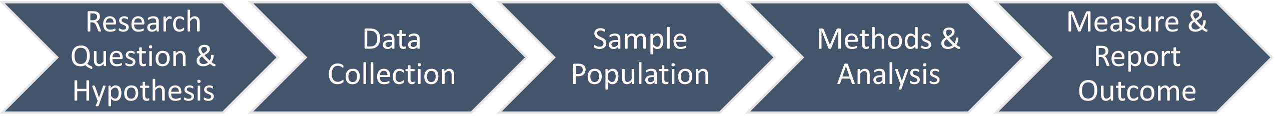
THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science

Basic Research Framework

- Research Question & Hypothesis
  - RQ: What makes persuasion effective?
  - Hypothesis: Evidence makes persuasion effective
- Data Collection
  - Download Posts & Replies from Change My View Subreddit
- Sample Population
  - Redditors in Change My View as proxy for “people engaging in persuasion online”
- Methods & Analysis
  - Identify “Evidence” in replies (e.g., search for “http” links)
- Measure & Report Outcome
  - Persuasion rate of replies with evidence vs. without evidence

THE UNIVERSITY of EDINBURGH
informatics

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science



Social networks and Covid-19

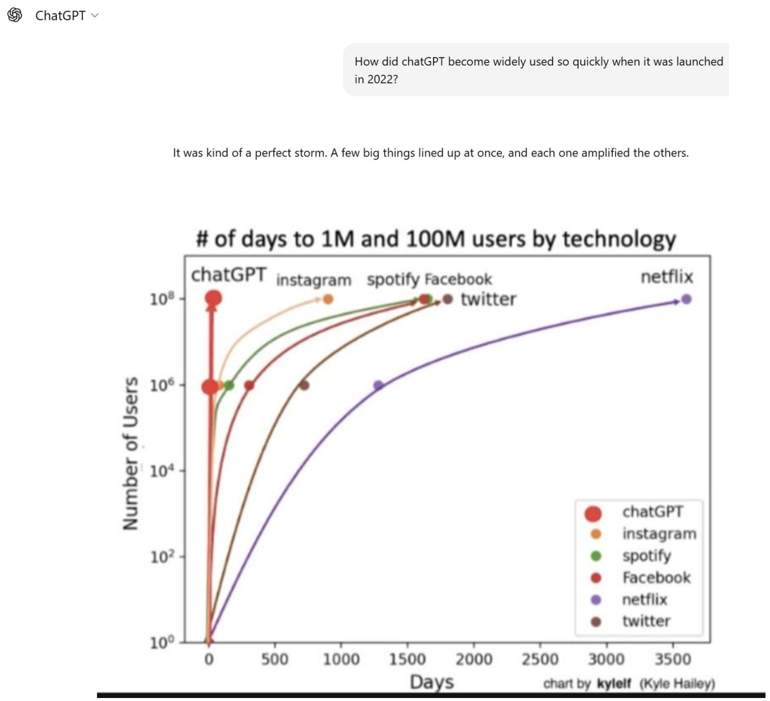
- Many viruses (including SARS-CoV2) spread through direct human contact
- You probably got Covid from family or friends
- The number of people you interact with on a daily basis affects your risk of exposure

Social networks and labour markets

- How do people get jobs? Networking!
- That’s why we have platforms like Linkedin →
- Which are more valuable for finding a job?
  - Your close friends?
  - Your acquaintances?

Social networks and new technologies

- Thank about the first time you ever used a large language model like ChatGPT
- Did you hear about it from:
  - General media, like the news?
  - Your friends and contacts, online or offline?
- Technological innovations often spread through social networks – like viruses



ChatGPT

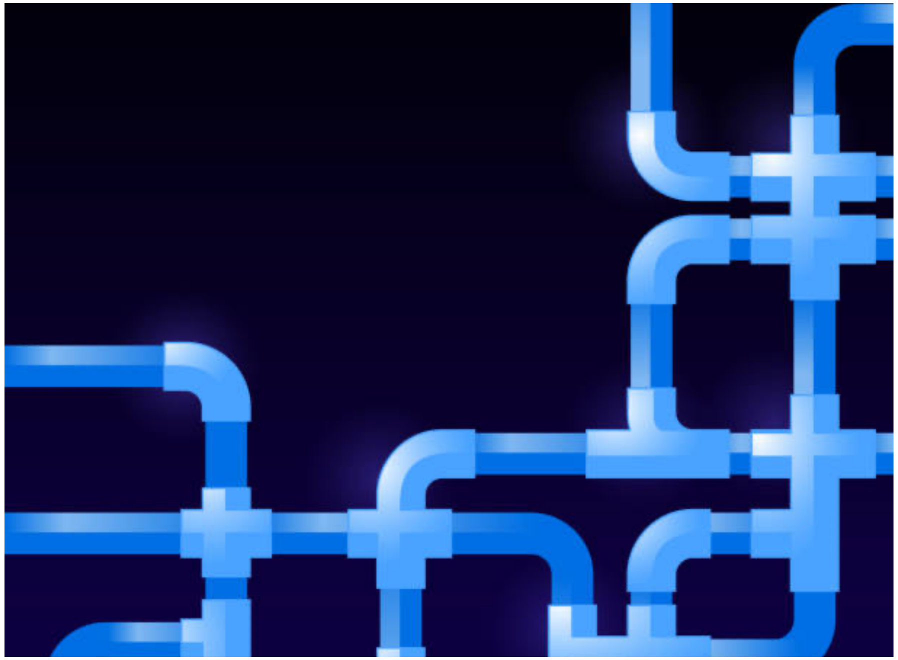
How did chatGPT become widely used so quickly when it was launched in 2022?

It was kind of a perfect storm. A few big things lined up at once, and each one amplified the others.

Graph: Kyle Hailey via John Nosta, medium

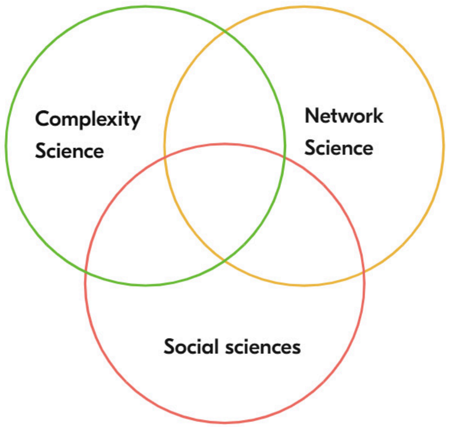
Common features: networks as plumbing

- Many social processes can be modelled as "something" flowing through a social network.
- Often that "something" is **information**:
  - Info about jobs
  - Info about new technologies
  - Memes, conspiracy theories, new songs, ideas, beliefs, attitudes…
- Information is "contagious"



Network science and social network analysis

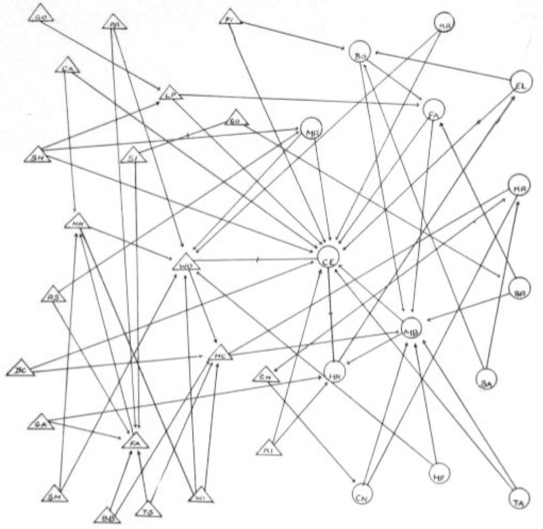
- Network science is a broad field spanning mathematics, physics, computer science, biology, complex systems and social sciences.
- Social network analysis (SNA) applies these tools to the social world



What are networks?

- A network is any set of units (nodes) connected by links (ties/edges)
- Units could be people
  - Or organisations, countries, concepts, neurons…
- Ties could be friendship
  - Or acquaintanceship, sharing information, economic exchange, similarity, murder…

EVOLUTION OF GROUPS



Class Structure, 1st Grade

21 boys and 14 girls. *Unchosen*, 18, GO, PR, CA, SH, FI, RS, DC, GA, SM, BB, TS, WI, KI, TA, HF, SA, SR, KR; *Pairs*, 3, EI-GO, WO-CE, CE-HN; *Stars*, 5, CE, WO, HC, FA, MB; *Chains*, 0; *Triangles*, 0; *Inter-sexual Attractions*, 22.

Jacob Moreno and Hellen Hall Jennings produced the earliest representations of social networks in 1934

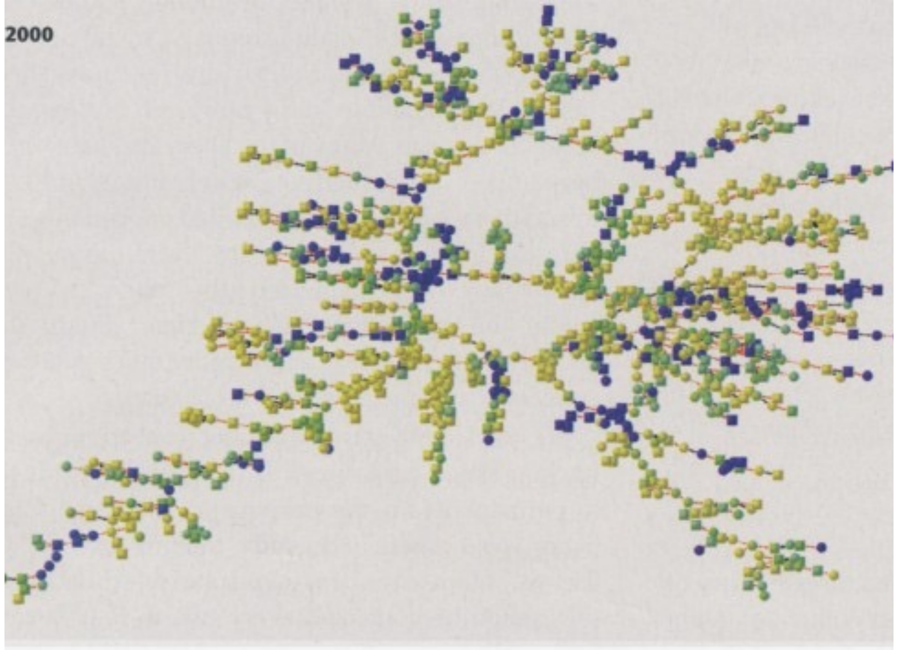
Key concepts: social influence and contagion

- Social influence: we are affected by our friends, peers and contacts.
- If a contact has "adopted" a technology, belief, behavior, attitude, etc, we become more likely to adopt it ourselves.
- Chains of this process result in social contagion or diffusion.

Image: Wikimedia

Example: is happiness contagious?

- Researchers found that those surrounded by more happy people were more likely to become happy themselves.
- Process of “emotional contagion”
- Similar findings for smoking and obesity
- Findings controversial for reasons we will come back to.



Fowler, and Christakis (2008) https://doi.org/10.1136/bmj.a2338.

Cultural tastes

- Does knowing that other people like a song make it more likely that you like it?
- Music Lab experiment provided experimental evidence that it does.
- Researchers constructed an online platform for music exchange – an influential online experiment.

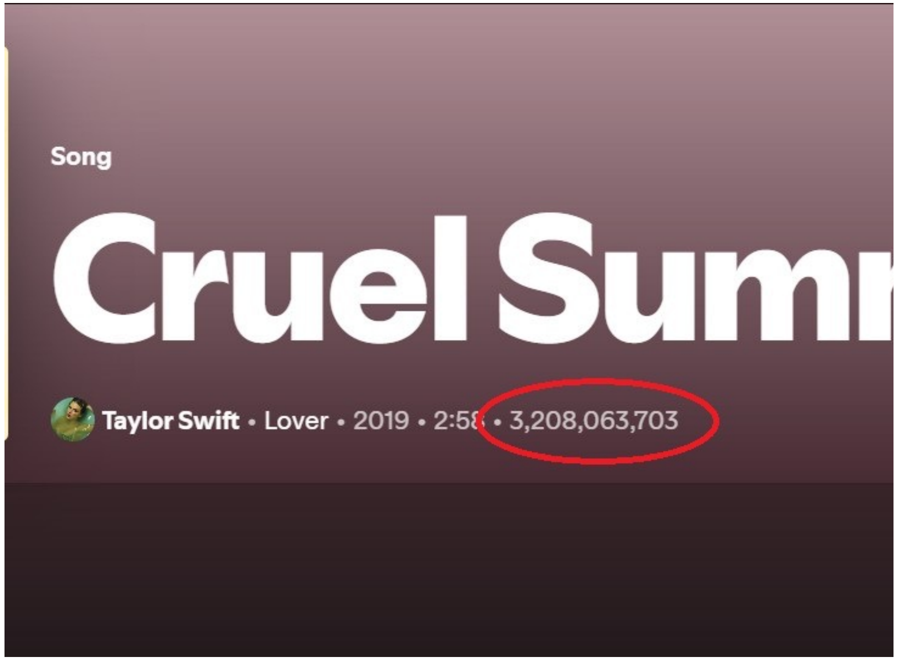


Image: Spotify

Social movements

- A classic study of participation in risky collective action found that contact with existing participants was the most important predictor of participation.
- Information about political actions can be contagious.

McAdam 1986; doi: 10.1086/228463; Image: https://www.freedomforum.org/freedom-summer-1964/

A core tension: influence or homophily?

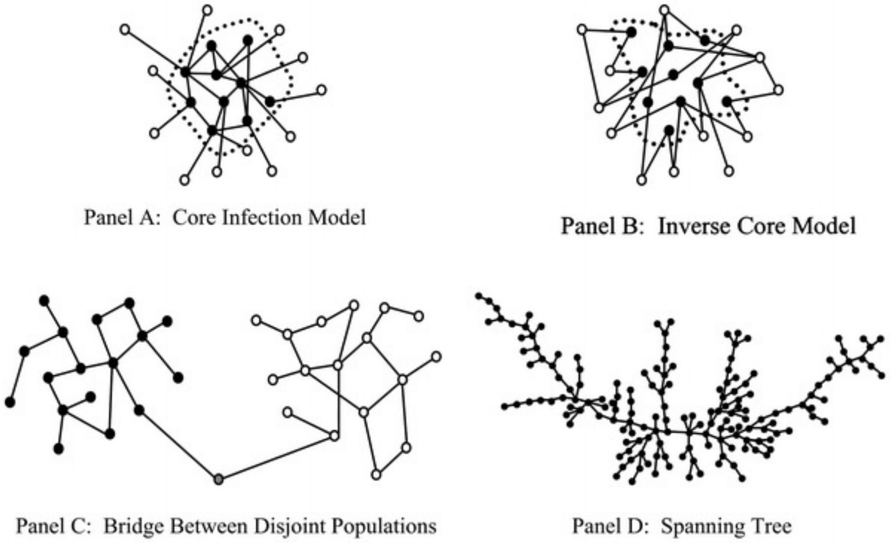
- Do you like heavy metal because your friends do, or did you choose your friends because they like the same music?
- Are you happy because your friends are, or did you choose happy friends?
- Are you going to the protest because your friends are, or because…

Homophily drives network formation

- The tendency to form relationships with similar others is widely documented.
- Similarity on demographic attributes, political beliefs, tastes, etc.
- It is extremely difficult to distinguish between influence and homophily.

Network structure matters

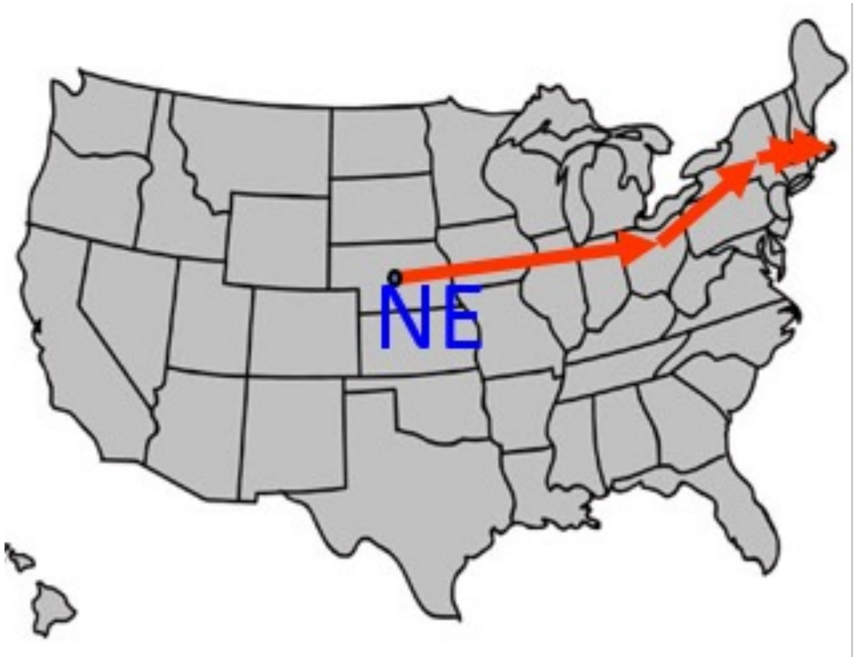
- The structure of network can inform us about processes occurring in the network.
- Levels of structure:
  - Dyad (pair): reciprocity, tie strength
  - Position: centrality, clustering
    - Local
    - Global
  - Whole network: density, average path length


Panel A: Core Infection Model
Panel B: Inverse Core Model
Panel C: Bridge Between Disjoint Populations
Panel D: Spanning Tree

Source: Bearman et. al. 2004: doi: 10.1086/386272

It’s a small world, after all

- Small world experience: meeting a stranger and discovering a surprising mutual friend/contact
- Chain letter experiment: how many steps would it take to send a letter by word of mouth from Nebraska to Massachusetts (in the 1960s)

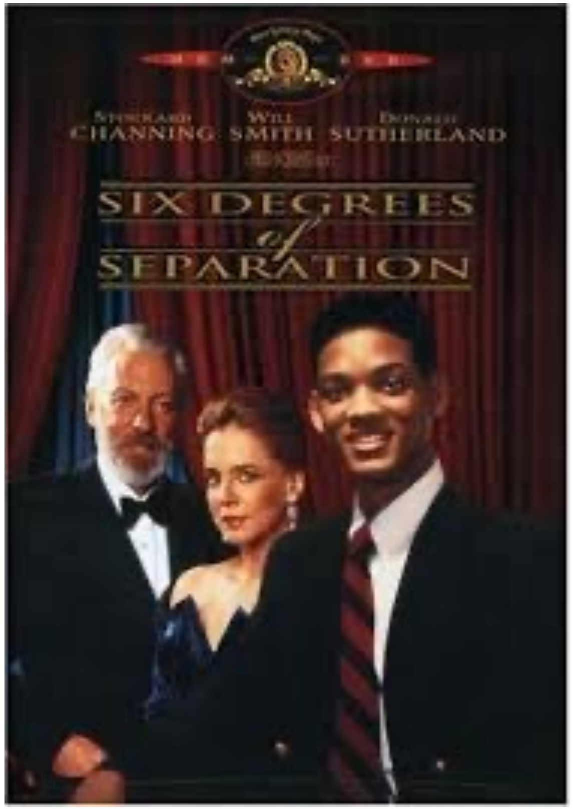


Milgram, S. (1967) “The small world problem,” *Psychology Today*, 1(1), pp. 61–67.

Six degrees of separation

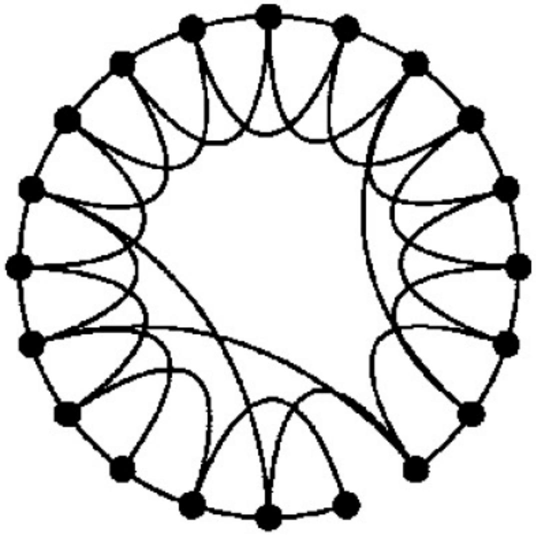
- Classic finding: an average of six steps in the chain from Nebraska to Massachussets
  - More recent email study: 7 steps
- This seems surprisingly close (or does it?)

For the more recent study, see Dodds (2003, doi: 10.1126/science.1081058



The small world model

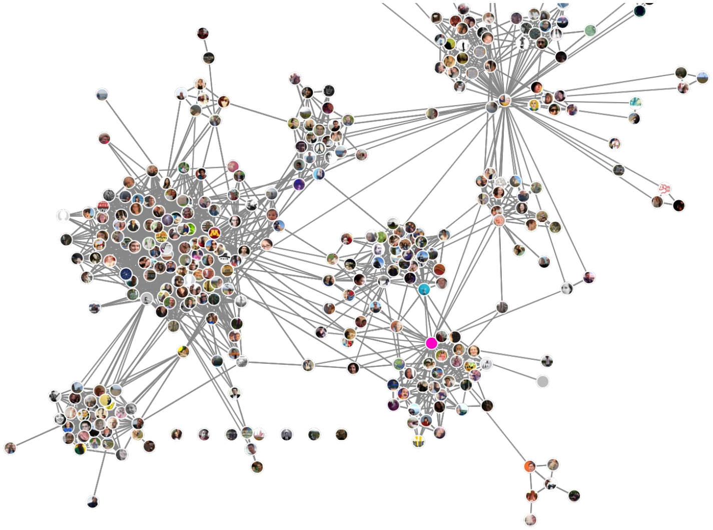
- The small world model is a formalization of this property of networks.
- Small world network have:
  - High levels of clustering
  - A small proportion of random, ‘long-distance’ ties
- Under these conditions, paths are short and it is only a few steps to nodes that seem “far away”



For an overview: Watt 2004; doi: 10.1146/annurev.soc.30.020404.104342

Are real worlds small worlds?

- Many real-world social (and physical) networks seem to have small world properties:
  - A lot of clustering: structures that look like groups
  - But also connections between these clusters



Tod Van Gunten's facebook network (a long time ago) via lost circles (now defunct)

Implications of the small world model

- If human social networks are small worlds, then viruses and information can move quickly
  - Small world model influential in epidemiological studies of contagion
- In cultural, political and other social processes, contagion may be slower/more unpredictable → complex contagions

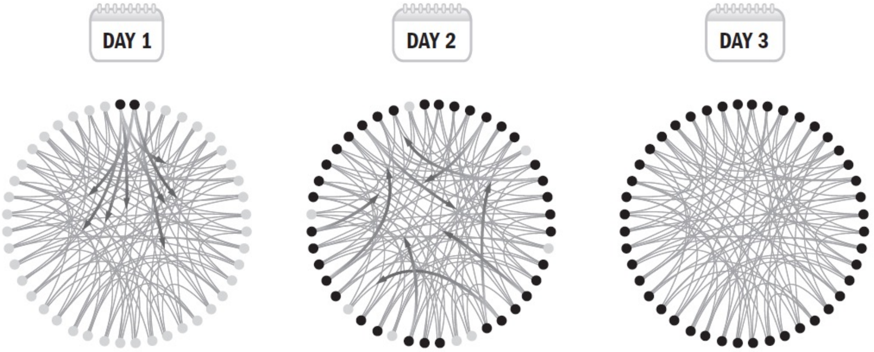


Figure 2.11 Diffusion in a Small World

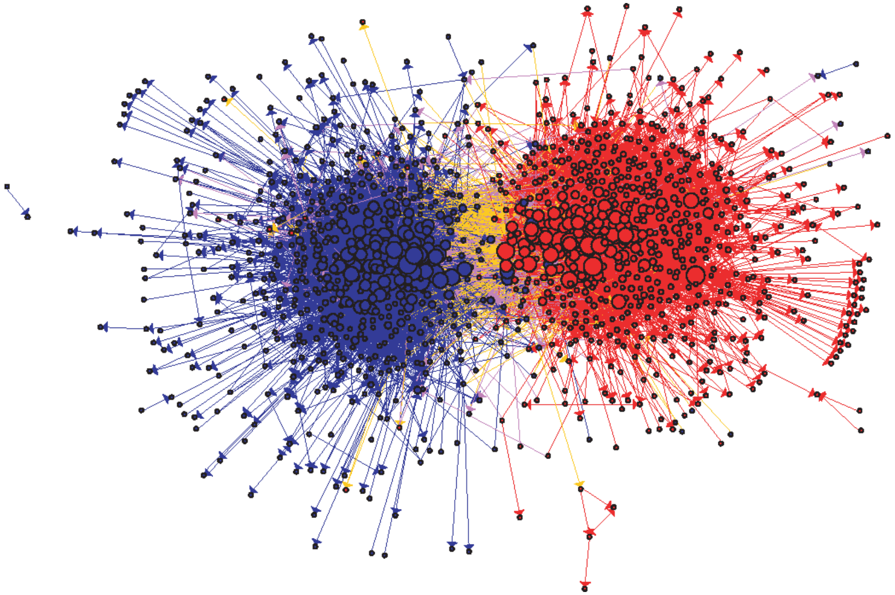
Source: Centola, D. (2018) *How behavior spreads: the science of complex contagions*. Princeton: Princeton University Press.

Implications for research

- Network analysts often measure structural properties of networks as a whole:
  - Density (proportion of connected nodes)
  - Transitivity/clustering (proportion of completely connected triangles)
  - Reciprocity (proportion of connections for which A → B then B → A
  - Modularity: tendency to form groups
  - Many others…

Structure can reveal interaction

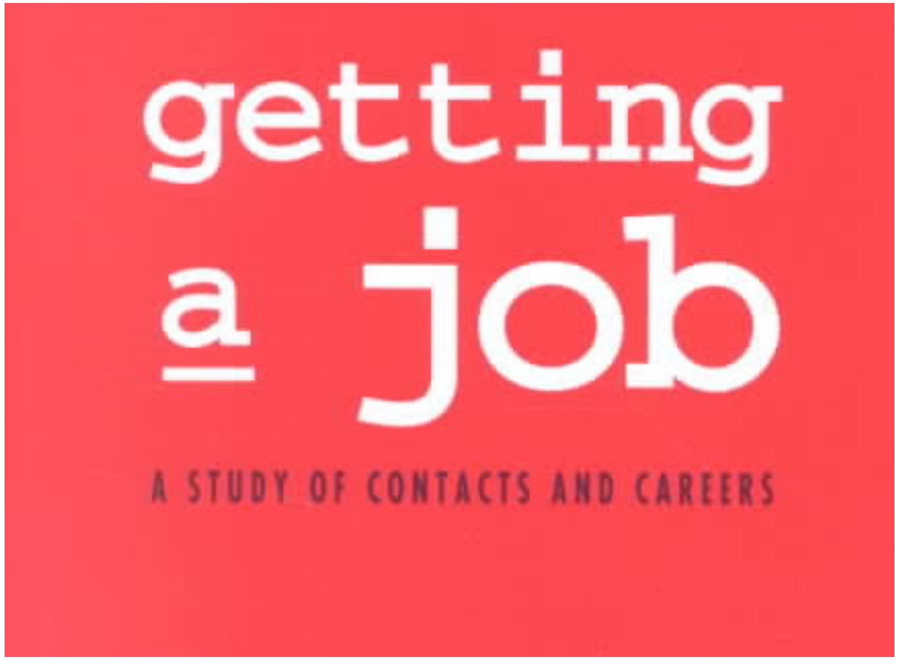
- Sometimes the structure tells us a lot about the interactions that generate a network.
- Network of hyperlinks between political blogs →
- Left links to left, right links to right
- One aspect of echo chambers and political polarization



Adamic and Adar (2005); doi: 10.1145/1134271.1134277

Another key idea: positions in networks

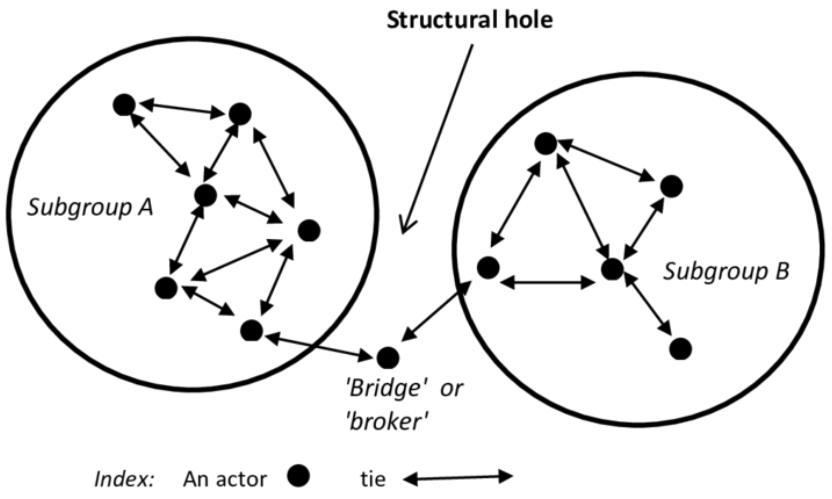
- Earlier example: getting a job through your social network.
- One idea: having more contacts is better
- A node’s count of network contacts is their network size or **degree centrality**



Granovetter, M. (1995) *Getting a job: A study of contacts and careers*. Second Ed. Chicago: University of Chicago Press.

Network size isn’t everything

- Theories of social capital suggest that the structure of a node’s network is important, in addition to (or more than) size.
- One aspect of structure: being a **bridge** between different subgroups or clusters in the network.
- Gap: **structural holes**



On structural holes, see Burt, R. (1992) *Structural holes: The social structure of competition*. Cambridge: Cambridge University Press.

Strength of weak ties

- **Strength of weak ties** theory suggests that weak ties (e.g. acquaintanceships) are better sources of valuable information
- This is because **weak ties tend to be bridging ties** giving access to distinctive (**non-redundant**) information

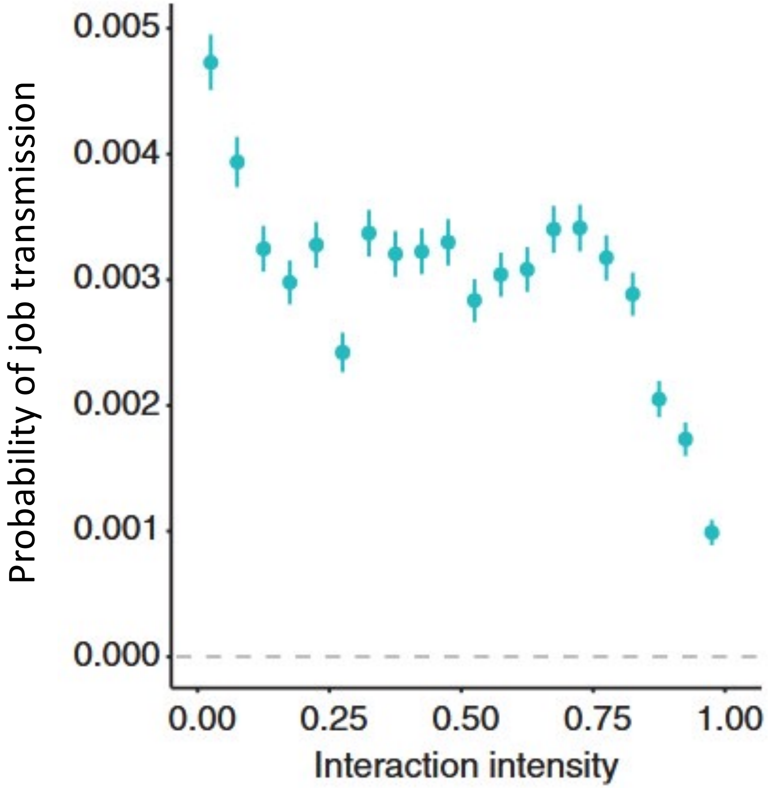


Image: Rajkumar et. al. (2022); 10.1126/science.abl4476

Centrality: measures of position

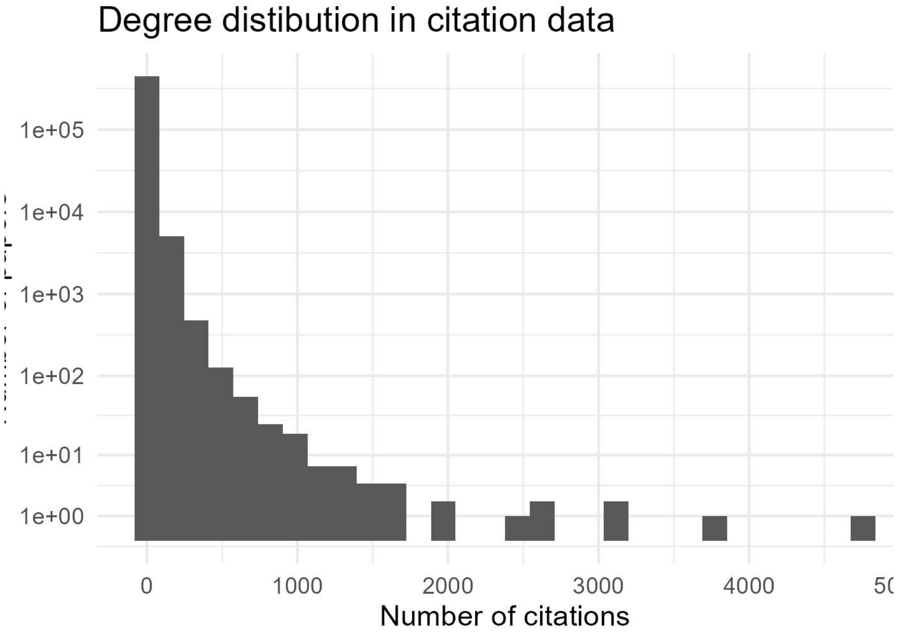
- Network analysts have developed many measures of centrality
- Depending on the network they capture:
  - Influence
  - Popularity
  - Power
  - Many other features

Saint-Denis
Gennevilliers
Le Bourget
Le Blanc Mesnil
Drancy
Aulnay sous-Bois
La Courneuve Aubervilliers
Les Grésillons
Stade de France Saint-Denis
La Plaine Stade de France
Saint-Ouen
Nanterre Université
Nanterre - Préfecture
La Défense
Pte de Clichy
Pereire Levallois
Rosa Parks
Pantin
Noisy-le-Sec
Bondy
Neuilly Pte Maillot
Gare du Nord
Haussmann St-Lazare
Magenta
Rosny Bois-Perrier
Avenue Foch
Ch. de Gaulle Etoile
Auber
Rosny-sous-Bois
Av. Henri Martin
Pont de l'Alma
Invalides
Châtelet-les-Halles
Boulainvilliers
Musée d'Orsay
Val de Fontenay
Av. du Pdt Kennedy
Ch. de Mars Tour Eiffel
St-Michel Notre-Dame
Gare de Lyon
Nation
Vincennes
Luxembourg
Javel
Port Royal
Gare d'Austerlitz
Fontenay-sous-Bois
Nogent Le Perreux
Nogent-sur-Marne
Boulevard Victor
Denfert Rochereau
Issy Val de Seine
Bibliothèque F. Mitterrand
Joinville le-Pont
Issy
Cité Universitaire
Gentilly
Ivry-sur-Seine
Laplace
Créteil
Meudon Val-Fleury
Maisons-Alfort Alfortville
Le Parc de St-Maur
Arcueil - Cachan

Image: Wikimedia

Degree distribution

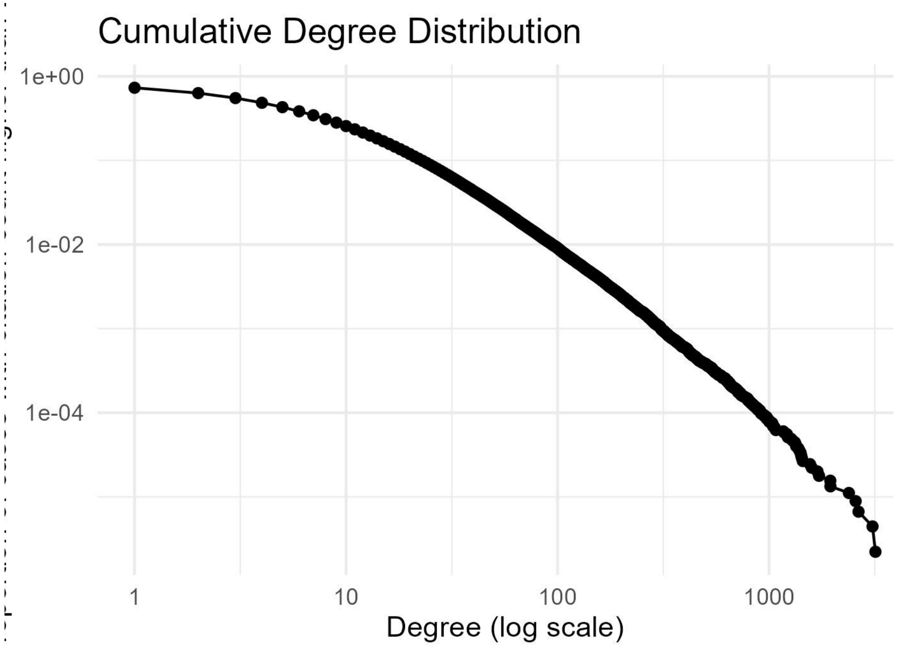
- Degree centrality is the simplest: the simple count of ties
- Degree distributions are typically right-skewed/heavy-detailed
- Most nodes have few contacts
- Some have many contacts



On degree distributions, power laws and preferential attachment, Barabási, A.-L. (2002) *Linked: the new science of networks*. Cambridge, Mass: Perseus Pub.

Degree distribution and network formation

- The shape of the degree distribution may be a clue about how networks form
- Some networks are like popularity contests: we want to be friends with people who already have a lot of friends
- This kind of network formation process will generate a degree distribution following a power law



Key concepts

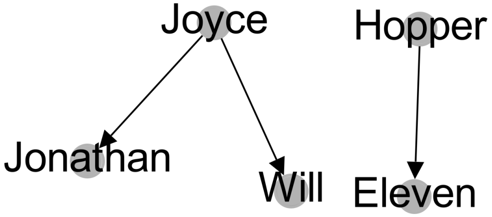
- Social influence and contagion
- Network structure
- Homophily
- Brokerage, structural holes, strength of weak ties
- Centrality
  - Degree centrality
  - Degre distribution, preferential attachment

Break

Recap: Network data as tabular data

Edge list

```
parent,child
Joyce,Will
Joyce,Jonathan
Hopper,Eleven
```



Adjacency matrix

```
  ,JC,WI,JN,HO,EL
JC, 0, 1, 1, 0, 0
WI, 0, 0, 0, 0, 0
JN, 0, 0, 0, 0, 0
HO, 0, 0, 0, 0, 1
EL, 0, 0, 0, 0, 0
```

- Each node is a row and a column
- “1” indicates a directed edge from row node to column node

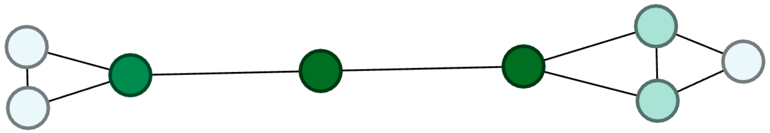
Incidence matrix

```
JC,-1,-1, 0
WI, 1, 0, 0
JN, 0, 1, 0
HO, 0, 0,-1
EL, 0, 0, 1
```

- Each node is a row
- Each edge is a column
- “-1” for outgoing edges, “1” for incoming edges

Centrality measures

- Degree centrality ("count of ties")
  - In-degree centrality: column sum in the adjacency matrix
  - Out-degree centrality: row sum in the adjacency matrix
- Betweenness centrality:
  - For each pair of nodes (s, t)
    - Find all shortest paths from s to t
    - Check if node i lies on these paths
    - Track the fraction of shortest paths through node i
  - For each node i, sum this fraction across all pairs
- PageRank

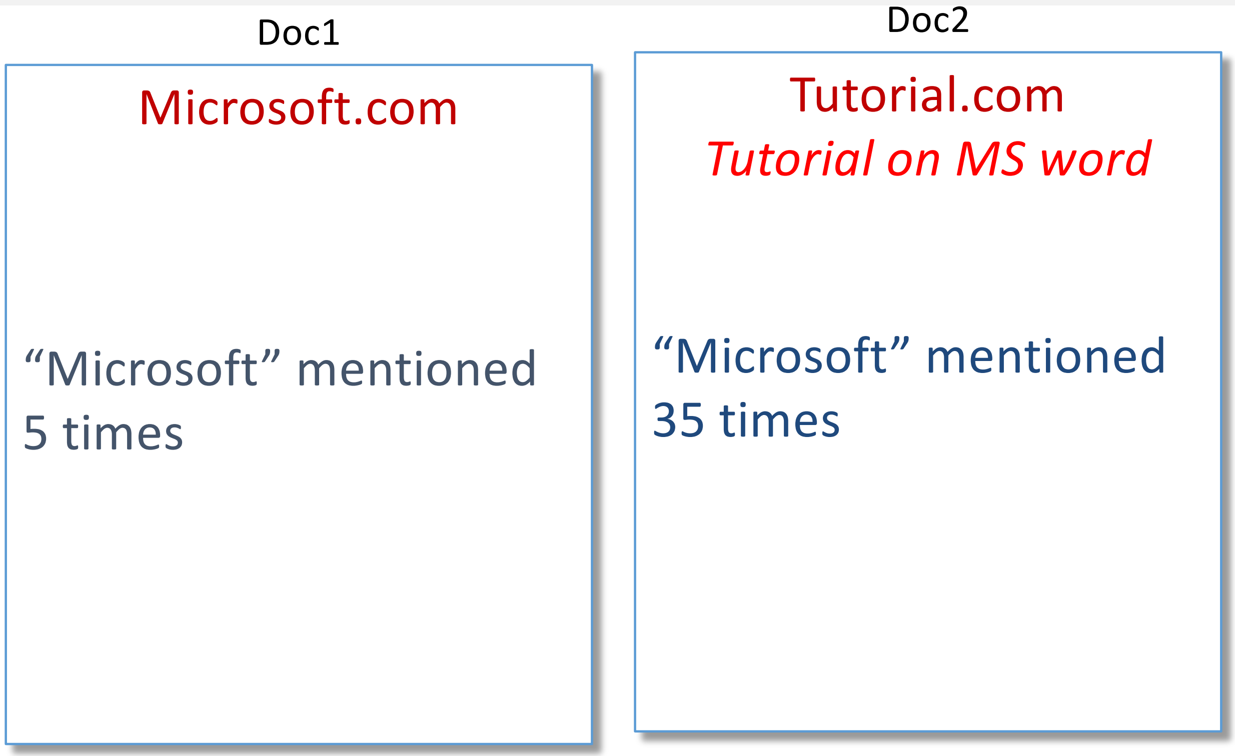


PageRank

A problem in search engines

Search:
‘Microsoft’

Which document is more relevant for the query?



Doc1

Microsoft.com

“Microsoft” mentioned 5 times

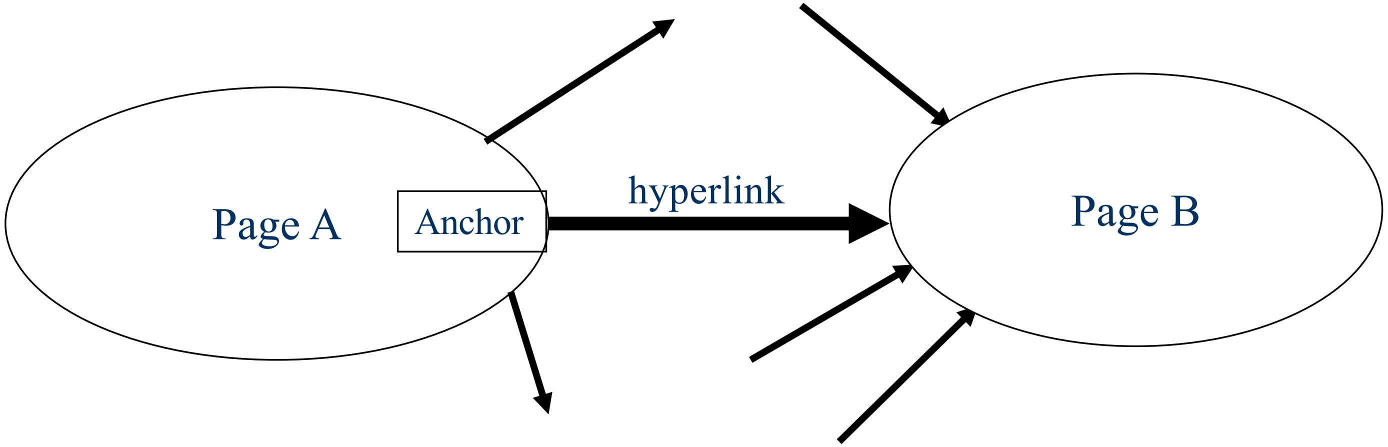
Doc2

Tutorial.com

Tutorial on MS word

“Microsoft” mentioned 35 times

The Web as a directed network

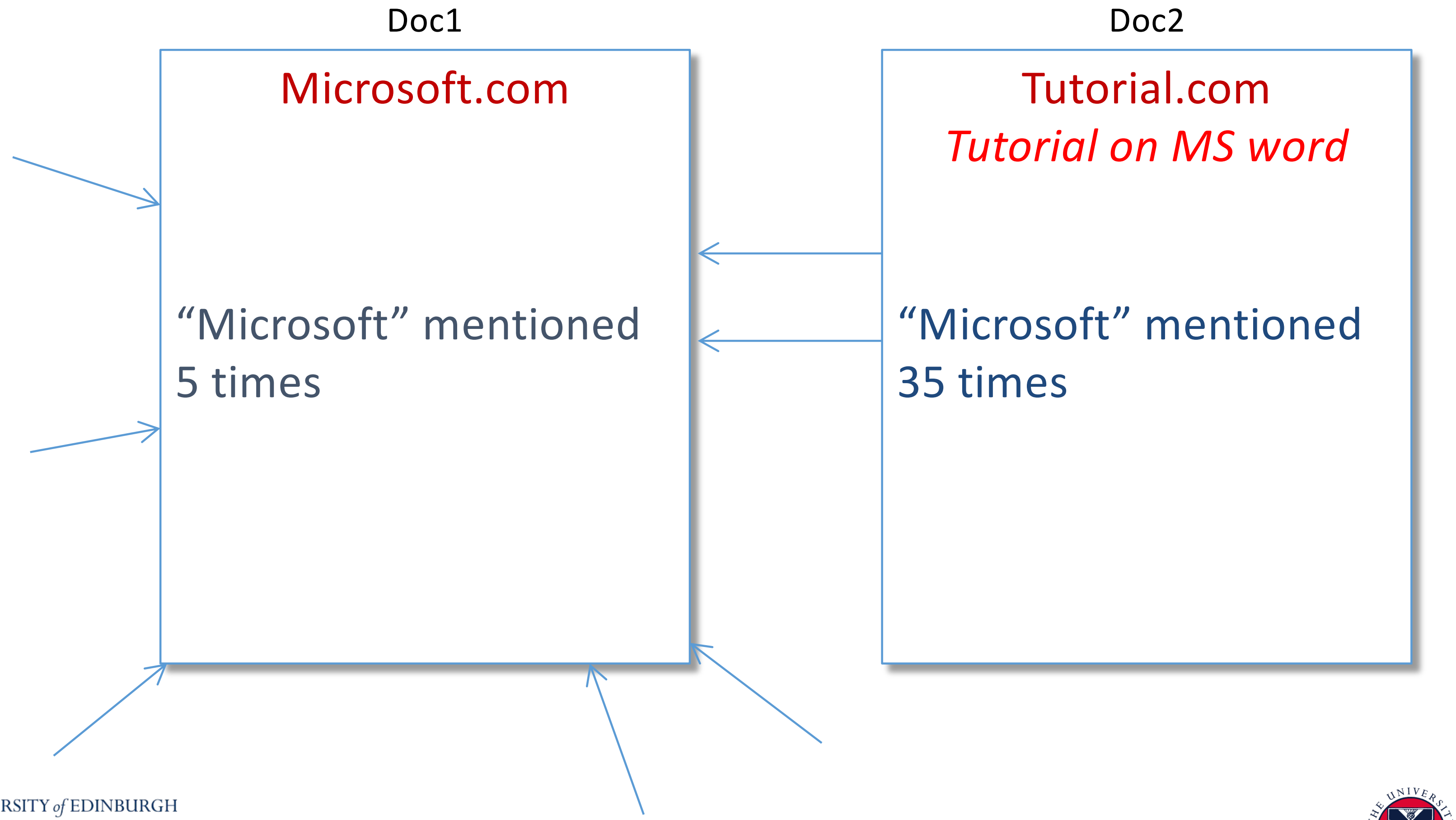


Assumption: A hyperlink between pages denotes perceived relevance (quality signal)

Links between pages

- Google's description of **PageRank**:
  - Relies on the "uniquely democratic" nature of the web
  - Interprets a link from page A to page B as "a vote"
- A link from to B means A thinks B is worth something
  - "wisdom of the crowds": many links means B must be good
  - Content-independent measure of quality of B
- Use as ranking feature, combined with content
  - But not all pages that link to B are of equal importance!
    - Importance of a link from BBC >>> link from blog page
- How many "good" pages link to B?

Search ‘Microsoft’



Doc1

Microsoft.com

“Microsoft” mentioned 5 times

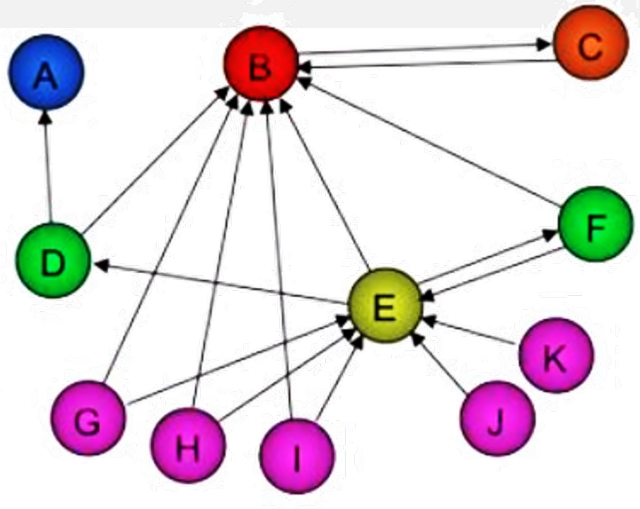
Doc2

Tutorial.com

Tutorial on MS word

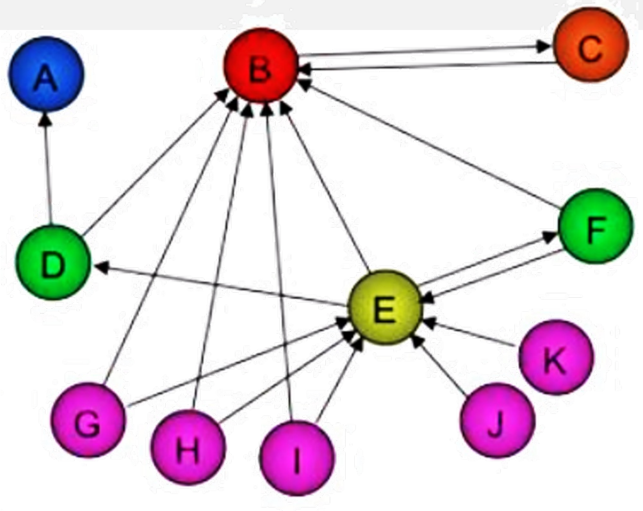
“Microsoft” mentioned 35 times

PageRank: A random surfer



- Analogy:
  - User starts browsing at a random page
  - Pick a random outgoing link → goes there → repeat forever
  - Example:
    G → E → F → E → D → B → C
  - With probability 1-λ jump to a random page
    - Otherwise, can get stuck forever A, or B ⟷ C
- **PageRank** of page x
  - Probability of being at page x at a random moment in time

PageRank: Algorithm



- Initialize $PR_0(x) = \frac{100\%}{N}$
  - N: total number of pages
  - $PR_0(A) = .. = PR_0(K) = \frac{100\%}{11} = 9.1\%$
- For every page x

$$PR_{t+1}(x) = \frac{1-\lambda}{N} + \lambda \sum_{y \longrightarrow x} \frac{PR_t(y)}{L_{out}(y)}$$

- $y \longrightarrow x$ contributes part of its PR to x
- Spread PR equally among out-links
- Iterate until convergence → PR scores should sum to 100%

PageRank: Example

Let λ = 0.82

$$PR(B) = \frac{0.18}{11} + 0.82\times[PR(C) + \frac{1}{2}PR(D) + \frac{1}{3}PR(E) + \frac{1}{2}PR(F) + \frac{1}{2}PR(G) + \frac{1}{2}PR(H) + \frac{1}{2}PR(I)$$

$$\approx 0.31 = 31\%$$

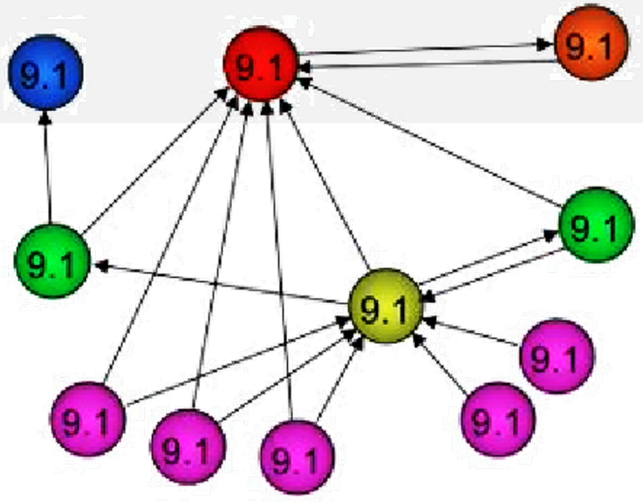
$$PR(C) = \frac{0.18}{11} + 0.82\times PR(B)$$

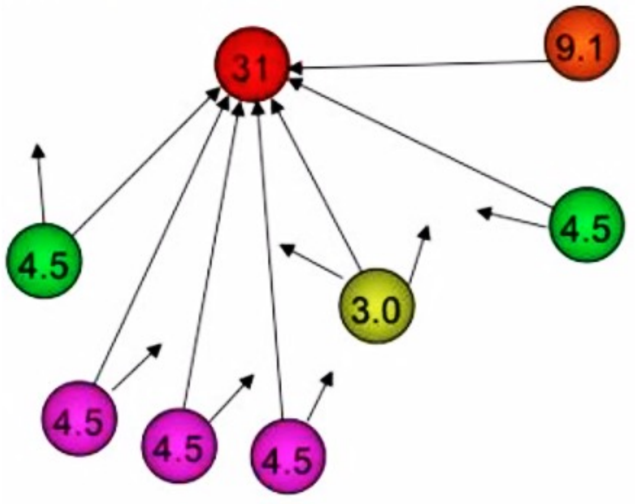
$$= 0.18\times 9.1\% + 0.82\times 9.1\%$$

$$= 9.1\%$$

$$PR_{t+1}(C) = 0.18\times 9.1\% + 0.82\times 31\%$$

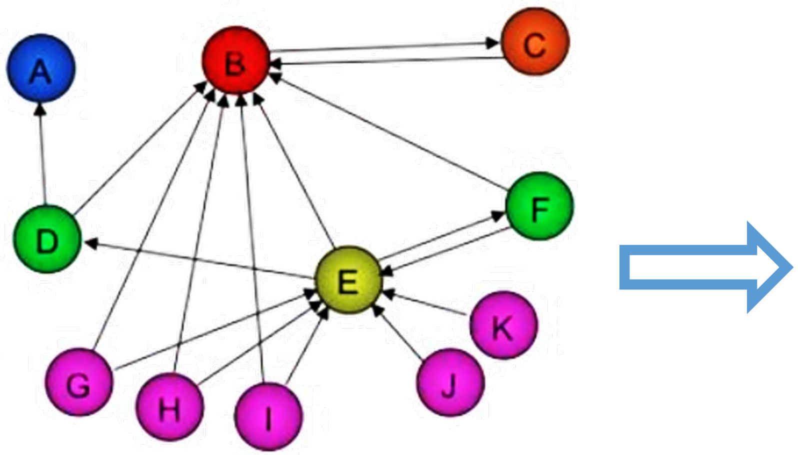
$$\approx 26\%$$

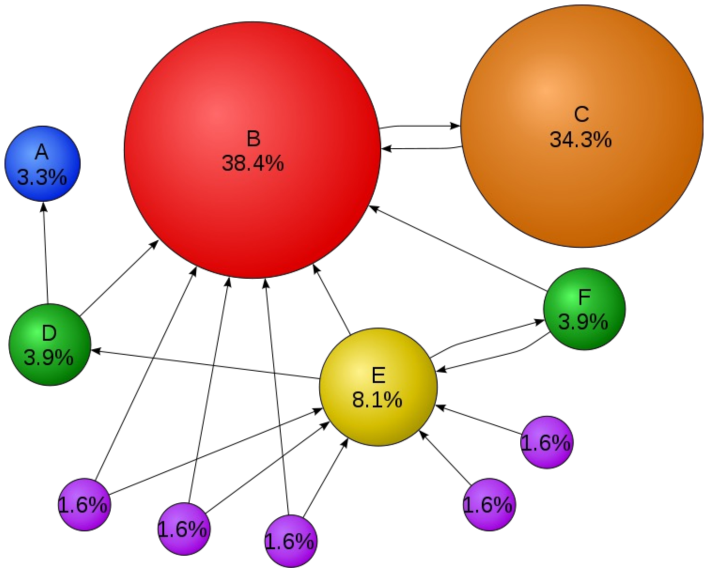




PageRank: Example result

- Algorithm converges after few iterations





- Observations
  - Pages with no inlinks: PR = $(1-\lambda)/N$= 0.18/11 = 1.6%
  - Same (or symmetric) inlinks → same PR (e.g. **D** and **F**)
  - One inlink from high PR >> many from low PR (e.g. **C** vs **E**)

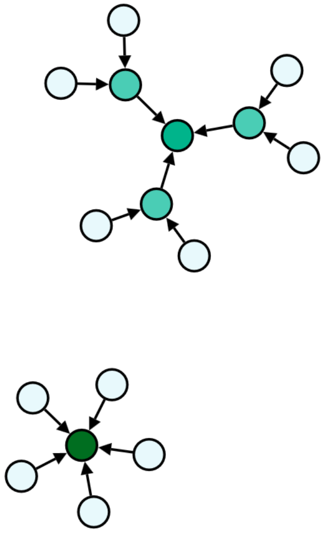
PageRank in retweet networks

- PageRank was initially developed for search engine, as a way of taking links between web pages into account in their ranking in search results.

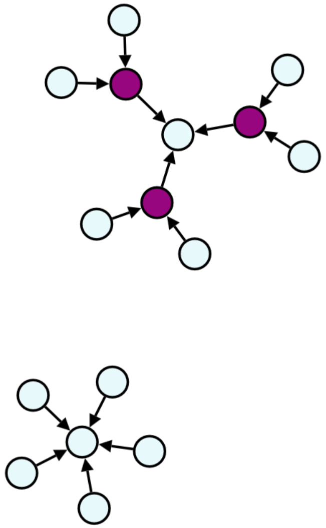
- *What interpretation does it have in retweet networks?*

Centrality measures: quiz

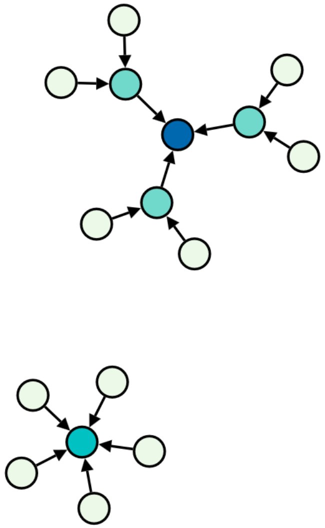
- Which figure shows which centrality measure out of betweenness centrality, in-degree centrality, and PageRank?



(a)



(b)



(c)

Community detection

Community detection

Some networks contain multiple distinct subgroups or **communities**.

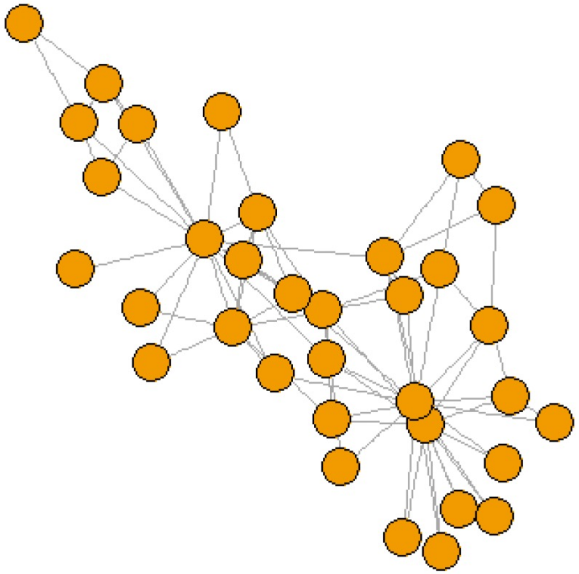
We might want to

- Identify these groups and interpret who belongs to which group
- Measure the extent to which the network forms distinct groups: how ‘groupy’ or ‘modular’ is the network?

Core intuition: groups are areas of a network that have many ties (high density) **within** the group but few ties (low density) **between** groups

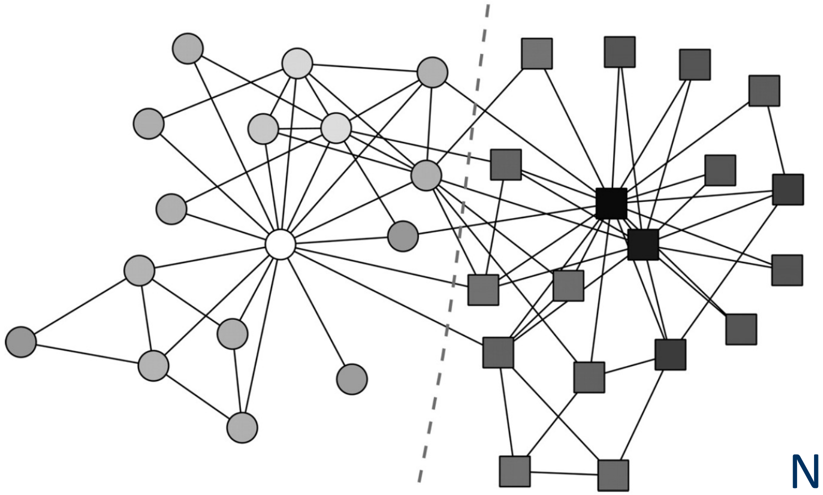
Detecting communities

Can we identify factions based on this network structure?



Detecting communities: modularity

Karate club divides into factions



Newman 2006

Modularity score

Modularity (Newman and Girvan 2004) compares the observed density to that in a null model

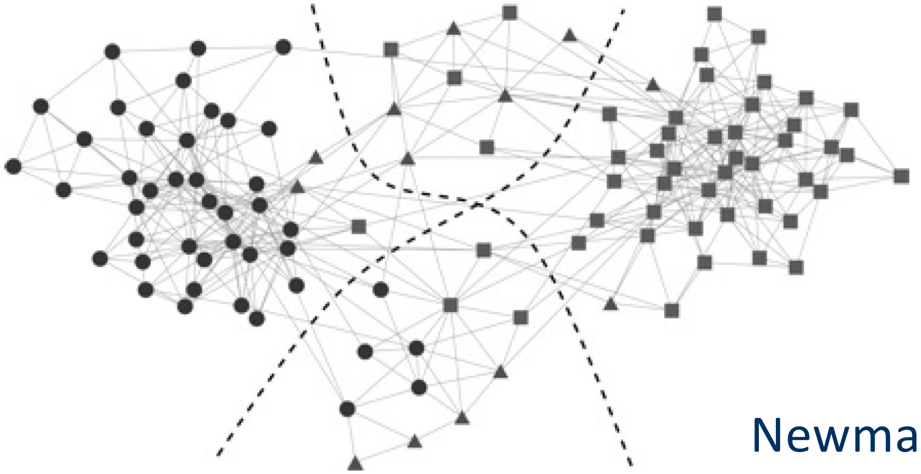
Null model: a random graph preserving some structural properties (e.g. size, density)

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j)$$

m is the number of ties
A is the adjacency matrix
P_{ij} is the expected number of edges between i and j under the null model
$\delta(C_i, C_j) = 1$ if i and j are part of the same community, otherwise 0

Detecting communities

Network of books about American politics
(ties are books frequently purchased by the same readers)



Newman 2006

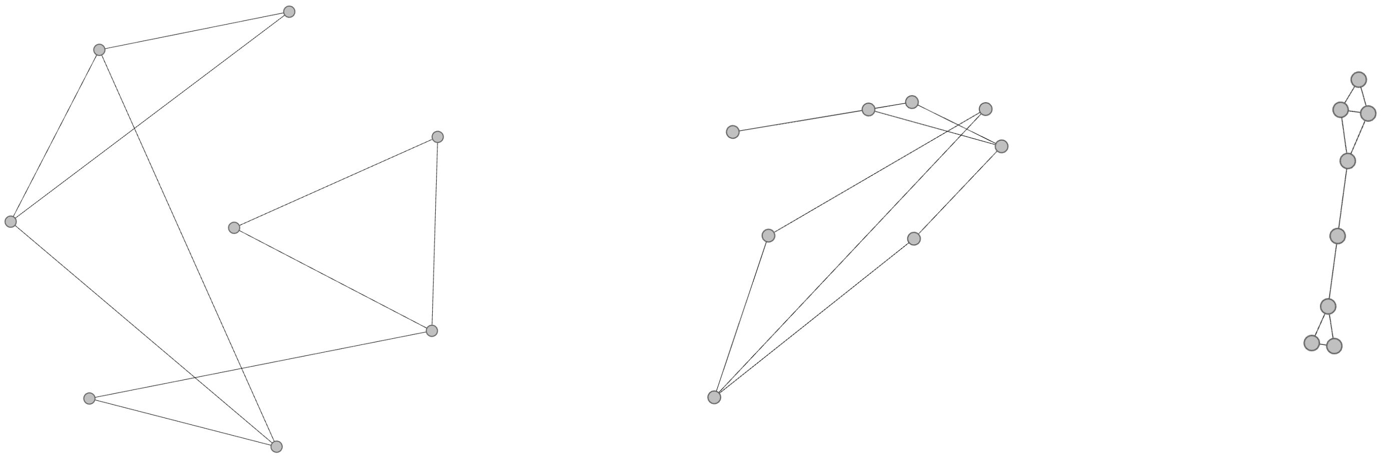
Uses of the modularity score

Two uses of the modularity score

1. We can infer network communities by using one of several community detection algorithms
   Example: Louvain algorithm to maximise modularity
   - Phase 1: Local optimization – move node to the neighbouring community that gives the largest increase in modularity
   - Phase 2: Network aggregation – collapse communities into super-nodes, creating a new network; then repeat Phase 1
2. If we have an exogenous indicator of group (e.g. party) we can ask how fragmented a network is by party

Network visualisation

What do all these networks have in common?



They all show the same network (same nodes, same edges) – they only differ in the position of the nodes!

Network visualisation

Network visualization relies on **layout algorithms** to position nodes in two-dimensional space

Most common approach are **force-directed** algorithms

We can also use node **colour**, **size** and **shape** to represent node attributes and positions

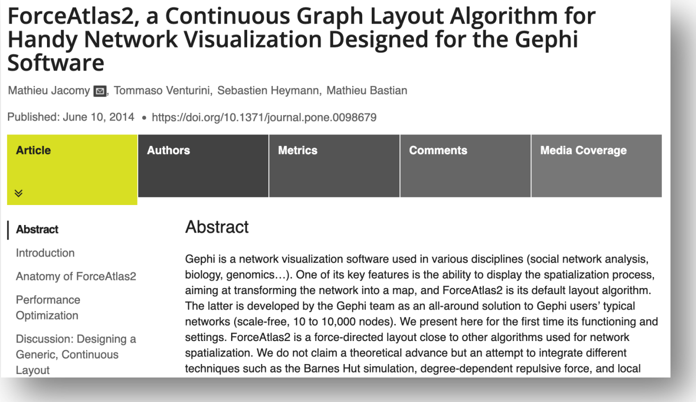
Force-directed layouts

Simulate a physical system. Basic principles:

- Nodes **repel** each other
- Edges **pull** nodes **together**
- Repeat until convergence (or for a set amount of time)

Other elements:

- Gravity
- Scaling
- Dissuading hubs



ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software

Mathieu Jacomy ✉, Tommaso Venturini, Sebastien Heymann, Mathieu Bastian

Published: June 10, 2014 • https://doi.org/10.1371/journal.pone.0098679

Abstract

Gephi is a network visualization software used in various disciplines (social network analysis, biology, genomics…). One of its key features is the ability to display the spatialization process, aiming at transforming the network into a map, and ForceAtlas2 is its default layout algorithm. The latter is developed by the Gephi team as an all-around solution to Gephi users' typical networks (scale-free, 10 to 10,000 nodes). We present here for the first time its functioning and settings. ForceAtlas2 is a force-directed layout close to other algorithms used for network spatialization. We do not claim a theoretical advance but an attempt to integrate different techniques such as the Barnes Hut simulation, degree-dependent repulsive force, and local

Jacomy et al. 2014

Gephi Demo

Questions?

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Social
& Political Science