Summer Institute in Computational Social Science

The University of Edinburgh

Day 3: Social Network Analysis

14 June 2023

Tod Van Gunten (tvangun@ed.ac.uk)

The plan for today

1. Introduction to the network perspective
2. Break
3. Network visualization exercise
4. A bit more lecture: social media networks (twitter)
5. Lunch
6. Twitter network exercise

Topical overview

1. Motivations: why networks?
2. A core tension: social influence and homophily
3. Social contagion and diffusion:
   - Why network structure matters
   - The small world model
   - Compex contagions
4. SNA applied to social media
   - Centrality and influence
   - Subgroups/community detection

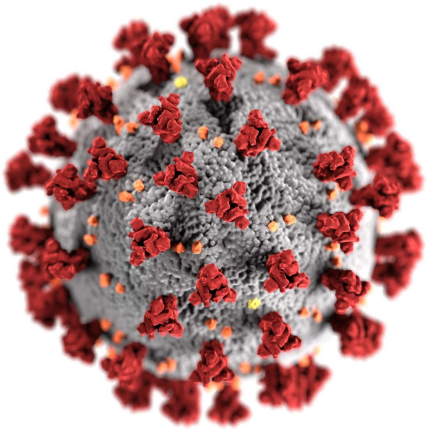
I assume no prior detailed knowledge of social network theory, but hope that those with some background will still be interested!

Topics not covered

The list is endless but some key topics omitted or glossed over due to time constraints:

- Theories of network formation (e.g. homophily, preferential attachment, social foci)
- Social capital theories (strength of weak ties, structural holes)
- State of the art network models (ERGM, Siena)

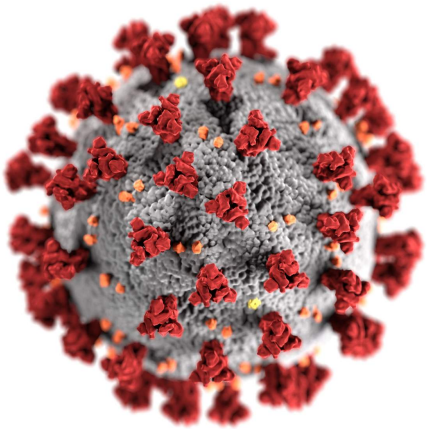
What does Covid-19 have to do with networks?



- SARS-CoV2 spreads primarily through sustained human contact.
- Much virus transmission occurs through pre-existing **social relationships** (family, friends, neighbours, co-workers, etc.)
- This fact is what makes **contact tracing** possible.
- The **number of people with whom we interact** every day is a key factor in how quickly the disease spreads (which is why we had lockdowns...)

Image: World Health Organization

What does Covid-19 have to do with networks?



- Most people interact with relatively few others but some people interact with many others (a reason why there are **super-spreading events**...)
- The study of variables such as the number of contacts people is central to **social network analysis**

Image: World Health Organization

What does Covid-19 have to do with networks?

- At larger scales, transportation links and flows of people between neighborhoods, cities, regions and countries permit a local epidemic to become a global pandemic.
- Networks are not just connections between physical individuals

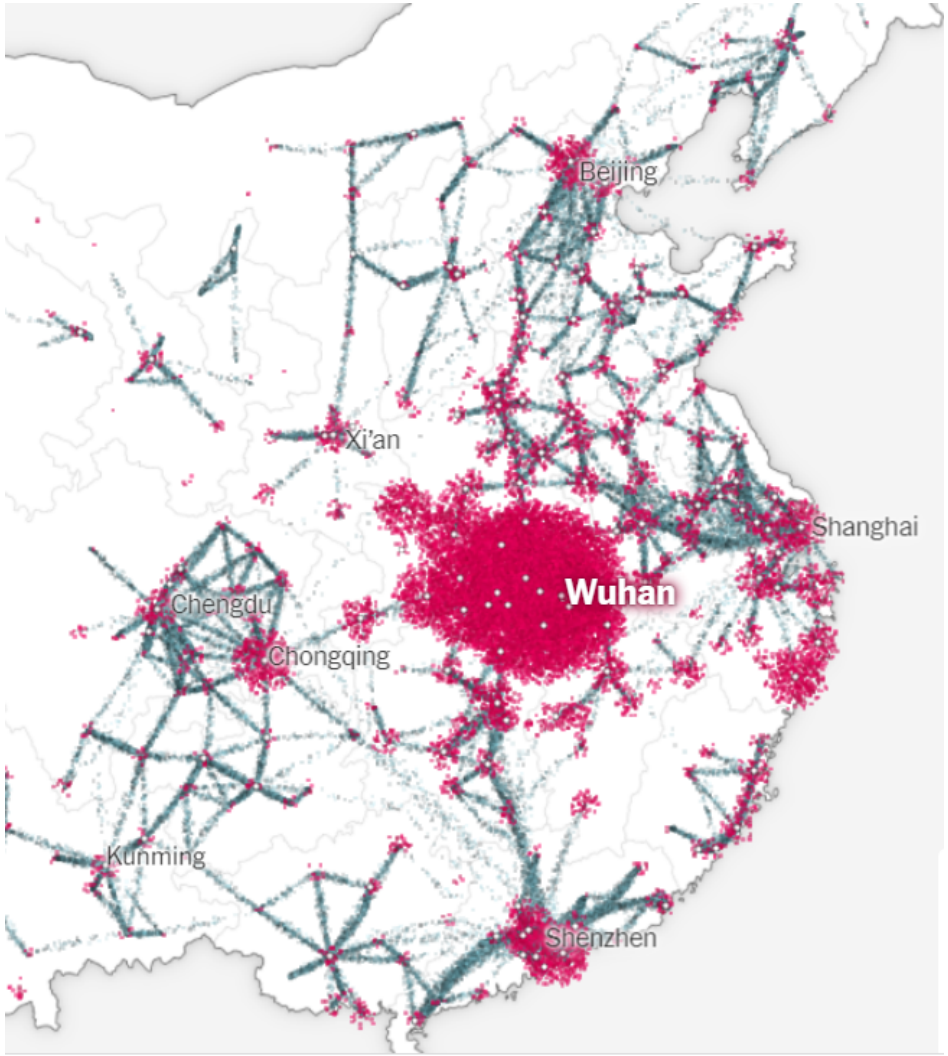
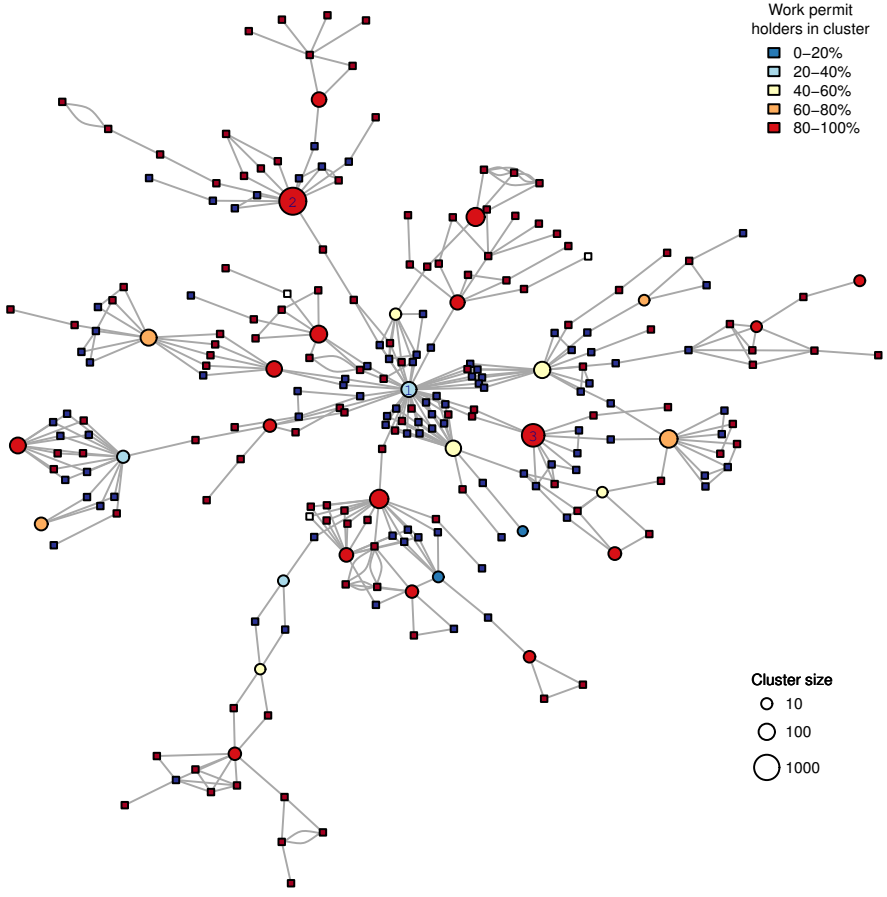


Image:
https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html

What does Covid-19 have to do with networks?



The **meso level**: in between **micro** (individuals) and **macro** (cities, countries), there are networks between neighbourhoods, communities, social groups

Source: Van Gunten (2021):
https://journals.sagepub.com/doi/10.1177/23780231211000171

What does Covid-19 have to do with networks?

- Not only viruses move through social networks: misinformation and conspiracy theories spread through social media.
- Online social networks also facilitate new forms of **collective action**

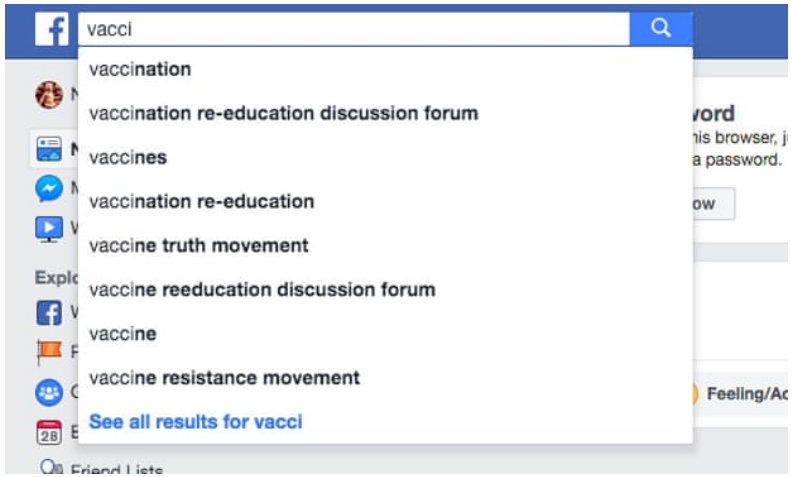


Image: https://www.theguardian.com/media/2019/feb/01/facebook-youtube-anti-vaccination-misinformation-social-media

What is social network analysis?

Social network analysis (SNA) is the study of patterns of relations between any set of social entities (individuals, organizations, communities, cities, countries, concepts)

Interested in the causes and consequences of social relations, structures of relations, positions in structures of relations.

Overlaps with broader **network science** field combining mathematics, physics, biology, epidemiology, etc. that draw on similar models and methods.

Basic terminology

Nodes (vertices): units that may or may not be connected.

- Often individual humans, but could be organizations, countries, animals, neurons...

Ties (edges, arcs): specific relations among nodes.

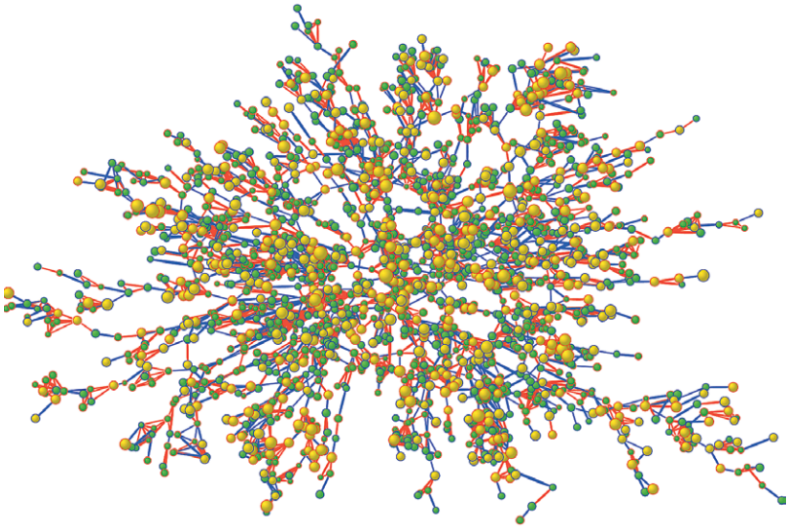
- Friendship, trust, alliance, exchange, competition, like, dislike...
- Directed (asymetric): I follow you but you don't follow me.
- Undirected (symmetric): on Facebook, users are only connected if both agree to the (electronic) relationship.

Core concept: Social influence

What leads individuals to adopt new technologies (e.g., a new social media platform)?

- Often, people adopt technologies because people they know already use the technology.
- General implication: innovations can **spread** through social networks (much like viruses...)
- Understanding the structure of a network can tell us how quickly and other implications.

Is obesity contagious?



Yellow = obese (BMI $>$ 30), Green = non-obese)

- Researchers found that having social contacts who became obese increased the likelihood of becoming obese
- e.g. a friend becoming obsese increased likelihood of obesity by 57%.
- Creates clusters of obese and non-obese individuals in the social network

Social influence leads to **social contagion** or **diffusion** of different social objects.

Source: Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." New England Journal of Medicine 357(4):370–79.

From technology and health to social movements

Similar social processes are involved in participation in **collective action**

The Freedom Summer movement for civil rights in the United States.

- Why do people participate in protest movements, especially risky forms?
- Classic finding: social connections to people already participating is a key factor in decisions to participate (McAdam 1986)
- Can contribute to sudden, surprising protest outbreaks.

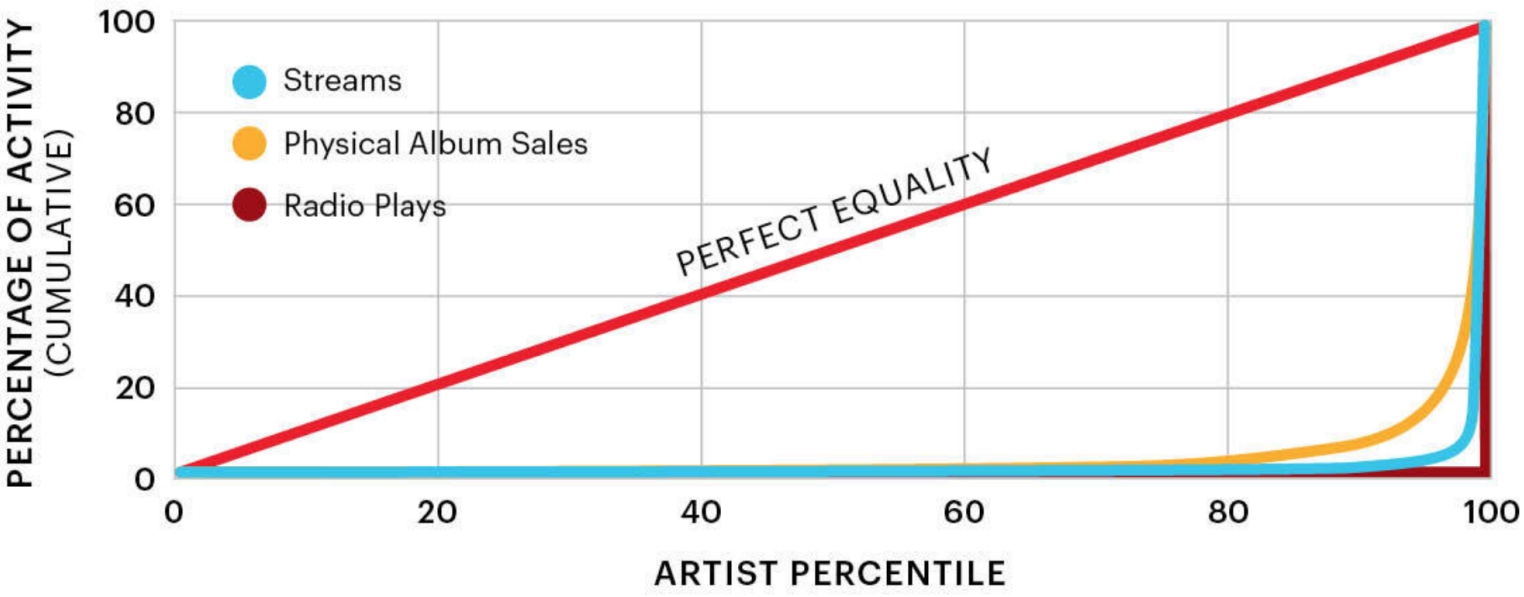
see McAdam, Doug. 1986. "Recruitment to High-Risk Activism: The Case of Freedom Summer." American Journal of Sociology 92(1):64–90

Social influence and culture

Does knowing that other people like a song make you more likely to like it?

- Researchers created a platform that allowed users to rate songs (Salganik et. al. 2006).
- When other users were able to view song ratings, the inequality in the number of streams increased.
- In other words, the most popular songs became even more popular

Inequality of attention

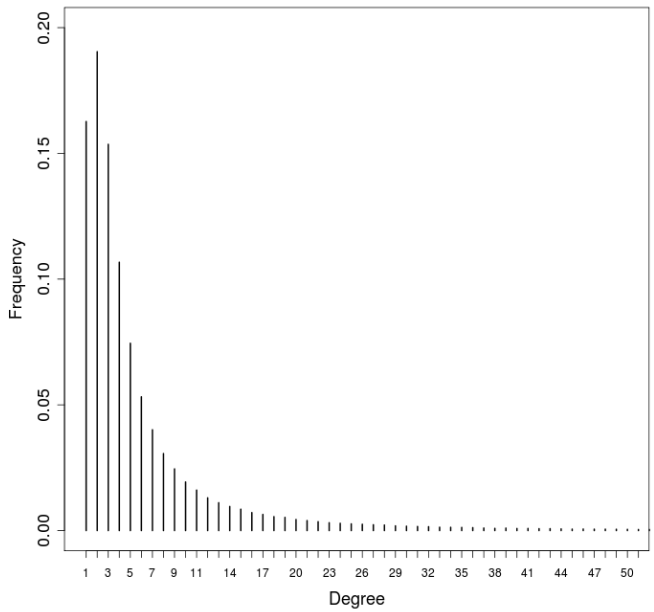


https://www.rollingstone.com/pro/news/top-1-percent-streaming-1055005/

Social influence and 'going viral'

- Markets and platforms in which social influence drives success can combine both extreme inequality and unpredictability of outcomes
- Some cultural objects (songs, videos, memes...) gather extreme attention while most go ignored: this is the essence of 'going viral'
- Very basic social process generating status hierarchies through **cumulative advantage** driven by social influence processes.

Degree distributions



Highly skewed **degree distributions** are a common feature of social networks.

- Related to viral super-spreading
- Sometimes referred to as 'scale free'

One process generating this property is **preferential attachment**: those who already have many ties are likely to form still more ties:

- e.g. friendship popularity contests
- e.g. citing papers that are already highly cited

Figure: https://tinyurl.com/2n524bjx

A core tension: Homophily versus influence

Do you like heavy metal music because your friends like it, or did you choose friends who like heavy metal because you already liked it?

Did you go to the protest because your friends did, or did you chose friends who share the same political beliefs?

Did you decide to smoke because your friends do, or did you choose friends who like to smoke because you already did?

A core tension: homophily versus influence

Homophily: the tendency of individuals to form social networks with others with whom they share common attributes (demographic characteristics, political beliefs, cultural tastes, etc.)

- Distinguishing peer influence from homophily is very difficult in practice.
- Example: controversy over findings about 'contagion' of obesity and smoking
- Homophily turns network question on its head: from **effects** of networks to **causes** of networks.

Key concepts so far

- Social influence
- Diffusion/social contagion
- Degree distribution
- Preferential attachment
- Homophily

Coming up: Why structure matters, small worlds, complex contagions

Why network structure matters

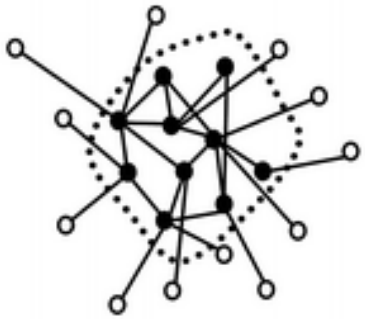
How do contagious diseases spread?

Sexually transmitted diseases (STDs) spread through direct contact, i.e. sexual/relationship networks.

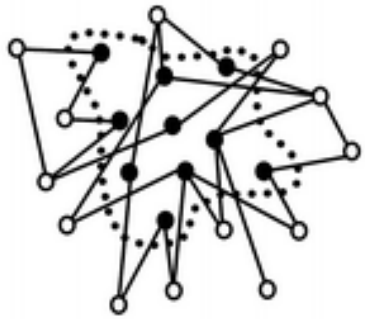
How can we prevent them from spreading?

Efficiency of **interventions** to interrupt contagion depends on the structure of relations: if you want to stop the spread of disease, who should you 'remove' (or change their behaviour) in order to stop the spread of disease?

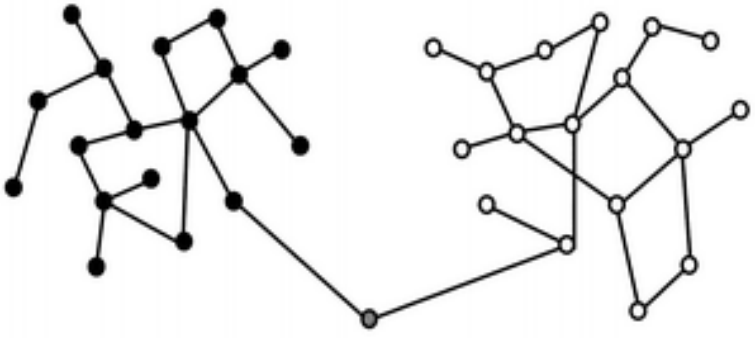
Models of STD transmission networks



Panel A: Core Infection Model



Panel B: Inverse Core Model

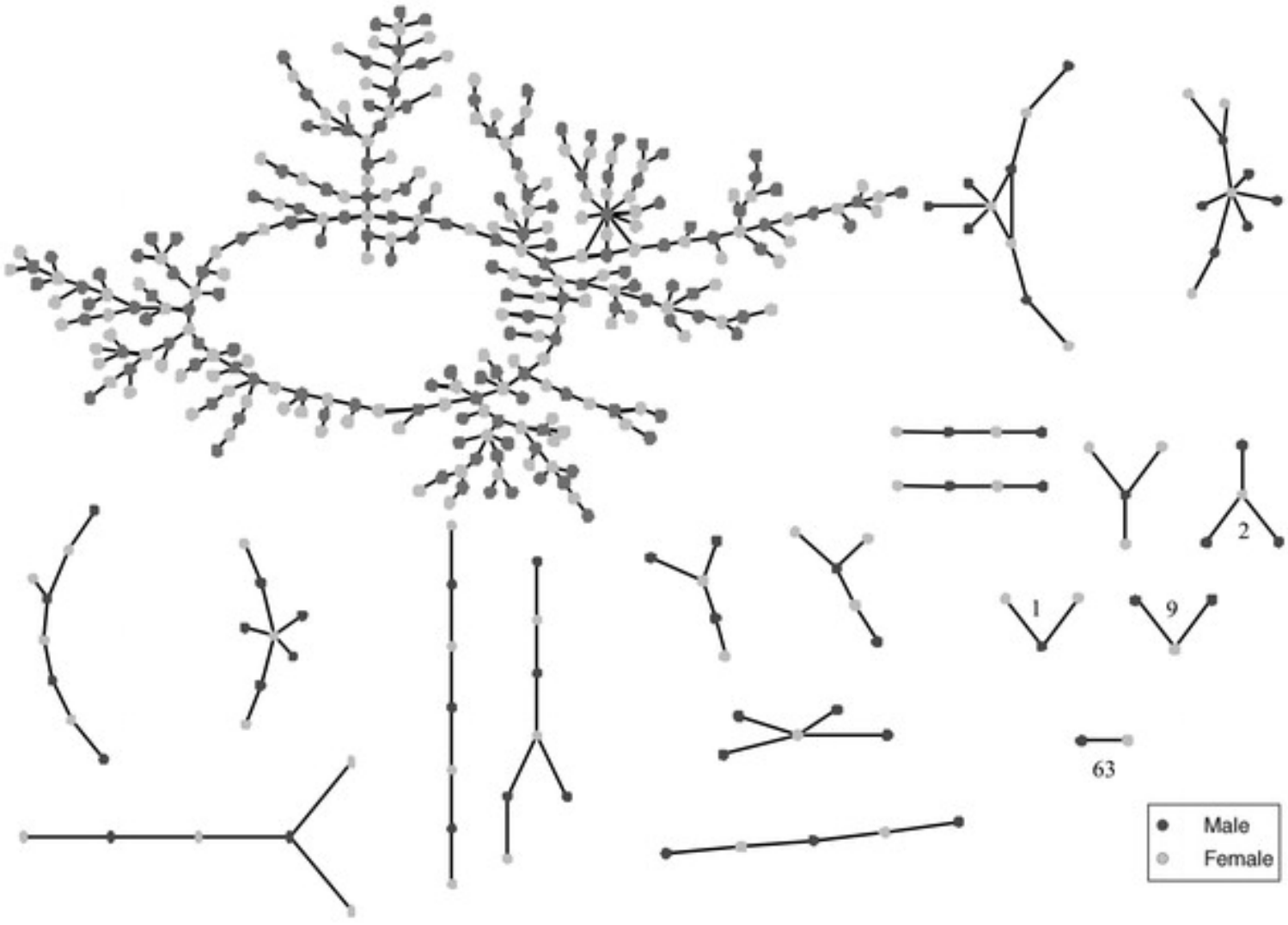


Panel C: Bridge Between Disjoint Populations

Panel D: Spanning Tree

Bearman et al. 2004, Chains of affection

Actual structure of adolescent sexual networks



Bearman et al. 2004, 'Chains of affection'

Implications of this structure

STDs move relatively slowly through this network (long paths)

But, there are no ‘core’ or ‘bridges’ to strategically target to prevent disease transmission.

Policies to prevent STDs have to be directed at the entire population, rather than some specific subset of individuals.

For discussion

What social processes might generate the network structure observed among adolescent romantic relationships?

- Why are there few triads (clusters of three nodes)?
- Why are there few four-step cycles?

Diffusion/social contagion: from disease to information

Information diffusion: analogy to disease

Private, valuable or dangerous information:

- Gossip
- Information about behavior: "I voted"
- Participation in risky/sanctioned activities (e.g. protest)
- Negotiation/bargaining "If you do X, I'll do Y"

Key assumption: such information is not 'broadcast' and spreads only through word of mouth.

Spreads only through relationships based on **trust**.

Network 'flow': networks as structures shaping diffusion

Diffusion/Social contagion: Networks are 'pipes' for the spread of diseases, innovations, ideas, information, etc.

Network structures affect the rate at which information spreads:

- The longer the **path length** between nodes A and B, the longer it will take for information to spread from A to B.
- If networks are highly **clustered**, information may become 'trapped' and diffusion can fail.

How structure matters: the small world problem

Small world experience: meeting a stranger and discovering a surprising mutual acquaintance.

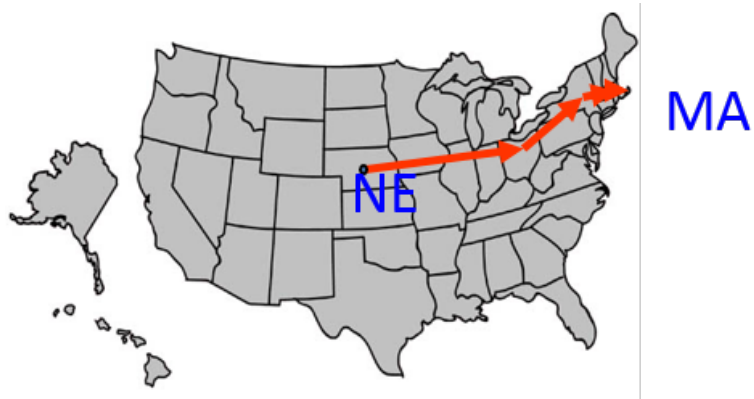
Implications:

- How 'far away' are any two randomly selected individuals?
- How hard is it to send information by 'word of mouth' between any two people (in a country, in the world...)?

The Milgram chain letter study

The experiment:

- Send letters to a few hundred randomly selected individuals in Omaha, Nebraska
- The goal: deliver letters to a target person in Boston, MA
- The rules: participants could only send letters on to their personal contacts (first-name basis)



Travers and Milgram (1969), An experimental study of the small world problem

6 degrees of separation

Key finding: mean path distance of 6 steps.***

***However, this is only among succesful paths (letters that reached the destination)

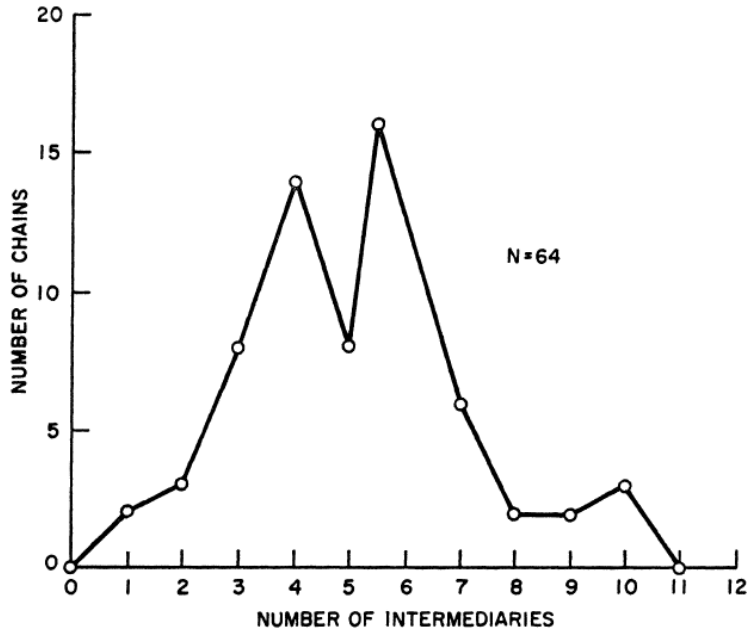


FIGURE 1

Lengths of Completed Chains

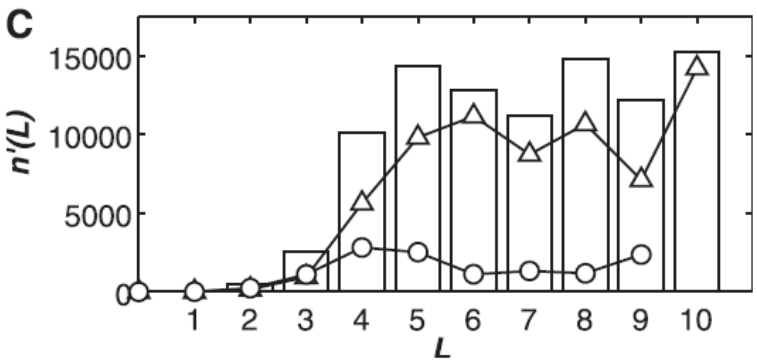
Travers and Milgram (1969), An experimental study of the small world problem

7 degrees of separation?

Email-based study (13 countries, 60,000+ participants)

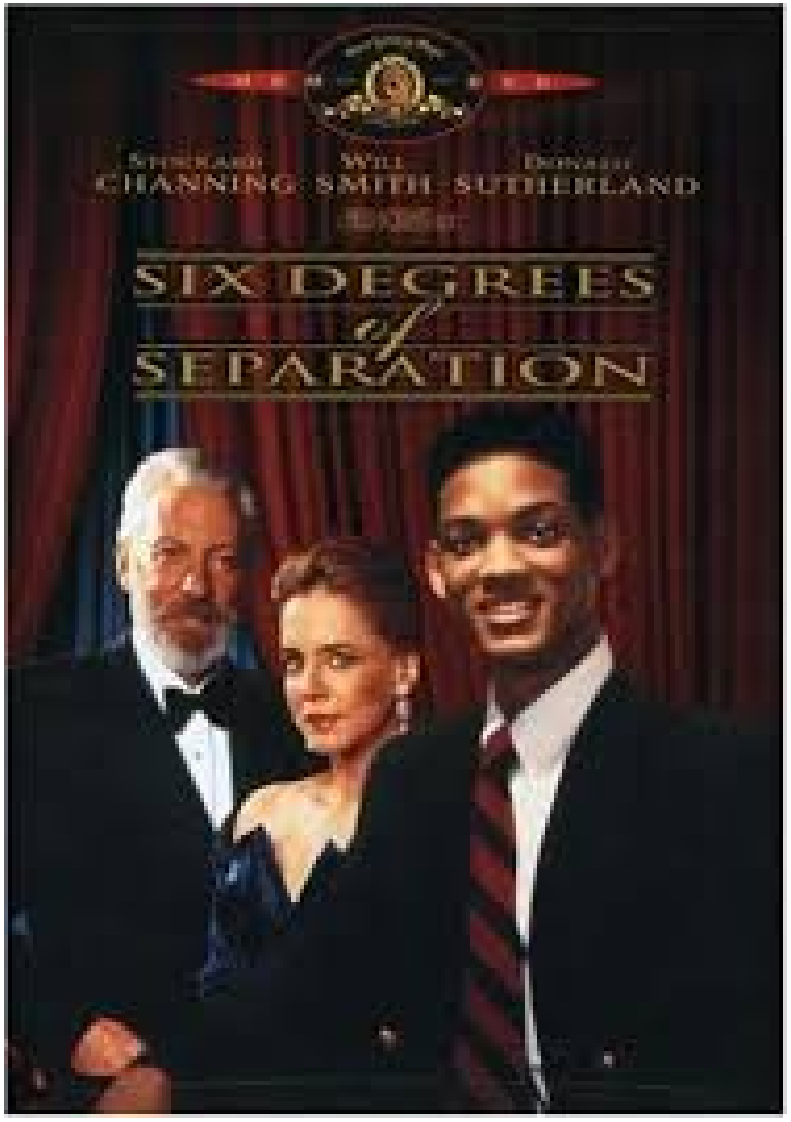
Adjustment for attrition (letters that stopped circulating)

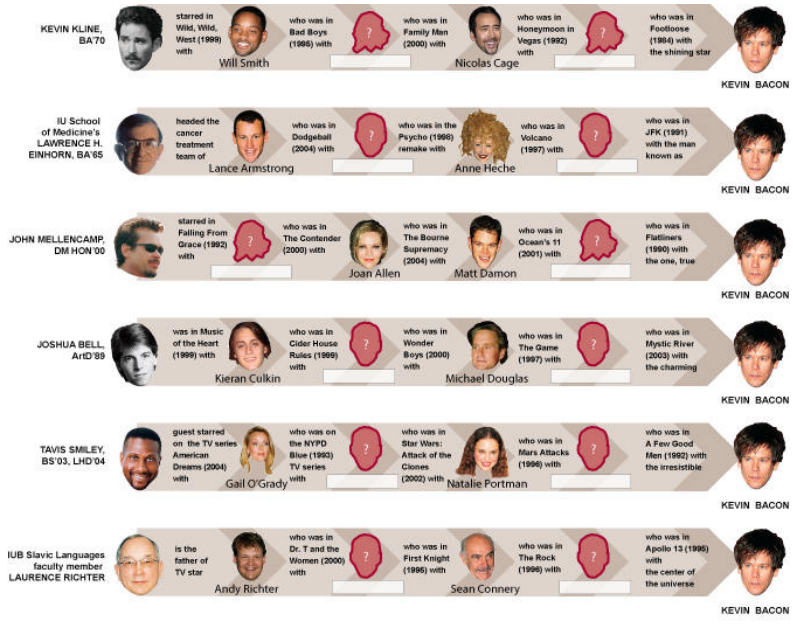
Median of 7 steps



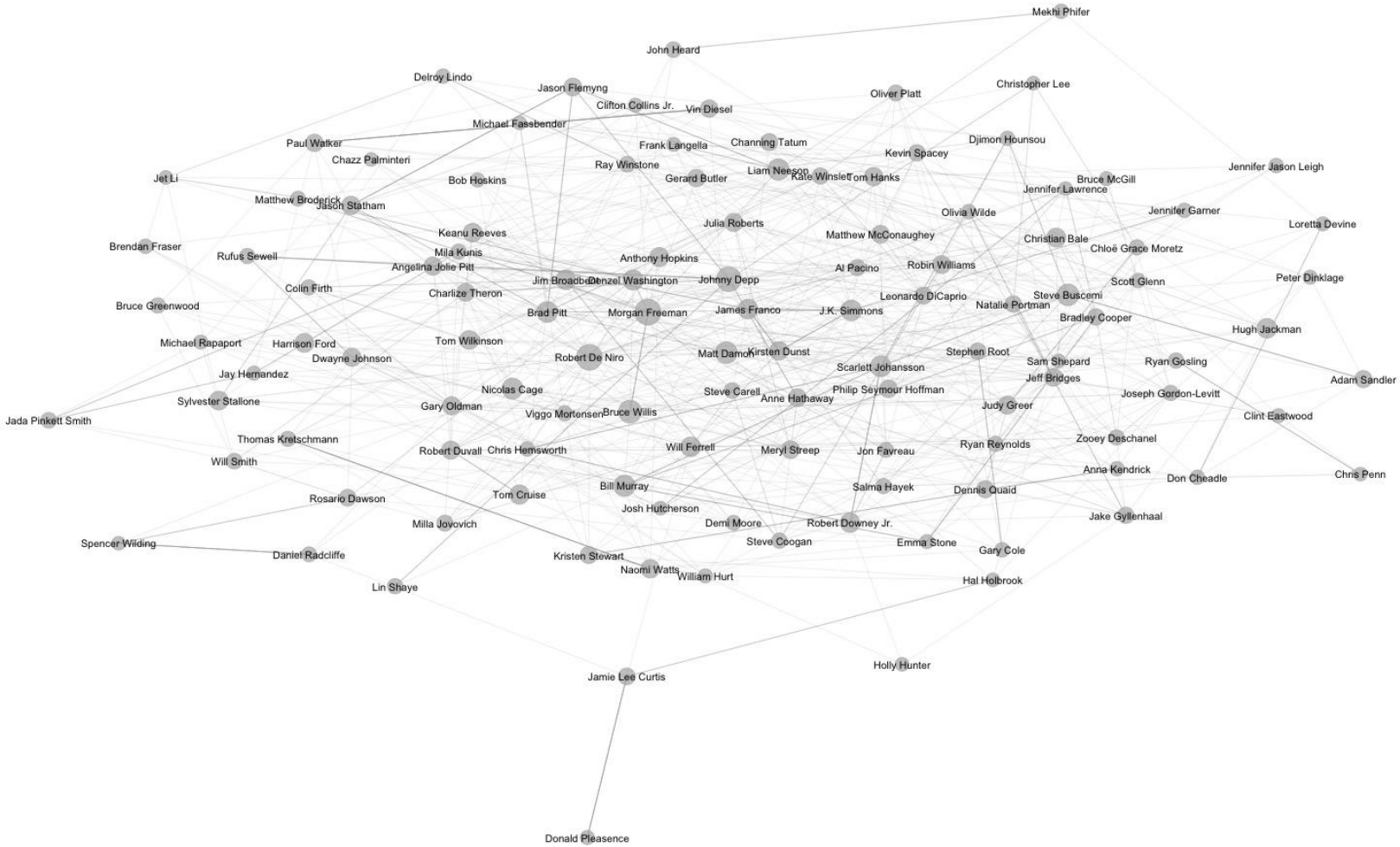
Dodds et. al. (2003), 'An experimental study of search in global social networks'

6 degrees of separation





Hollywood graph



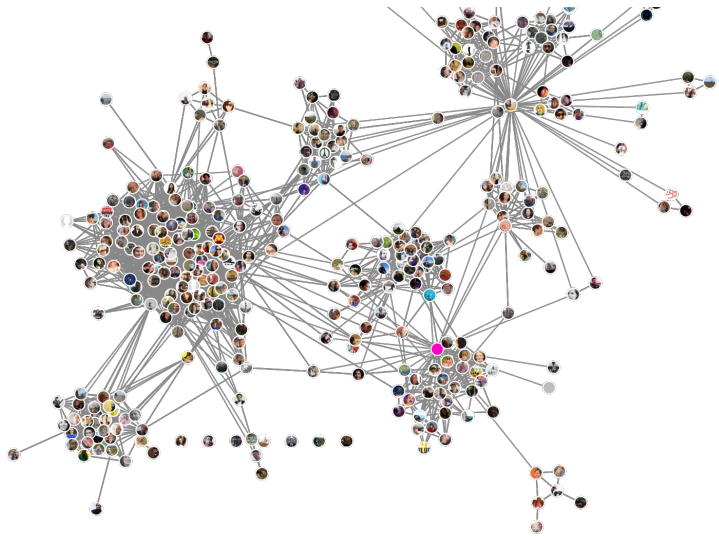
Why is the small world effect surprising?

Social networks are typically:

- Large: 7.8 billion and counting...
- Sparse: most people do **not** know one another.
- Decentralized: there is no human hub connected to a large proportion of the human population.
- Clustered: networks tend to form subgroups, most of the friends of our friends are also our friends.

In a large, sparse, decentralized and (importantly) highly clustered network, it ought to be difficult to reach many others in a few steps.

Clustering in social networks



My facebook network, retrieved using `https://lostcircles.com/` (apparently no longer workds)

Social relationships tend to form triads (triangles): friends of friends are also friends.

These triads group together to form **clusters** or subgroups.

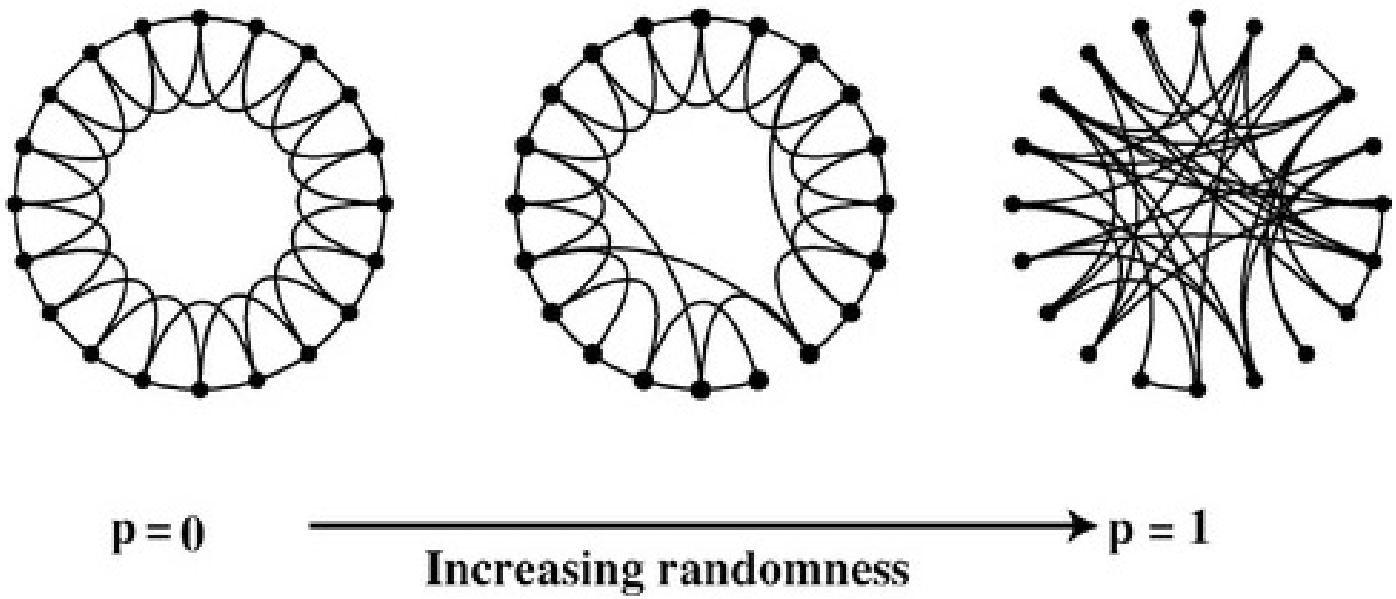
Homophily is one social process that can generate clustering (though there are others).

Community detection methods seek to identify subgroups in networks (discussed later).

Modeling small worlds

Networks are a combination of **order** and **randomness**:

- Assume an ordered local structure (a lattice): this is a simplified model of the tendency of networks to **cluster**
- 'Rewire' this network randomly: some ties span clusters or groups



Source: Watts (2004), 'The New Science of Networks'

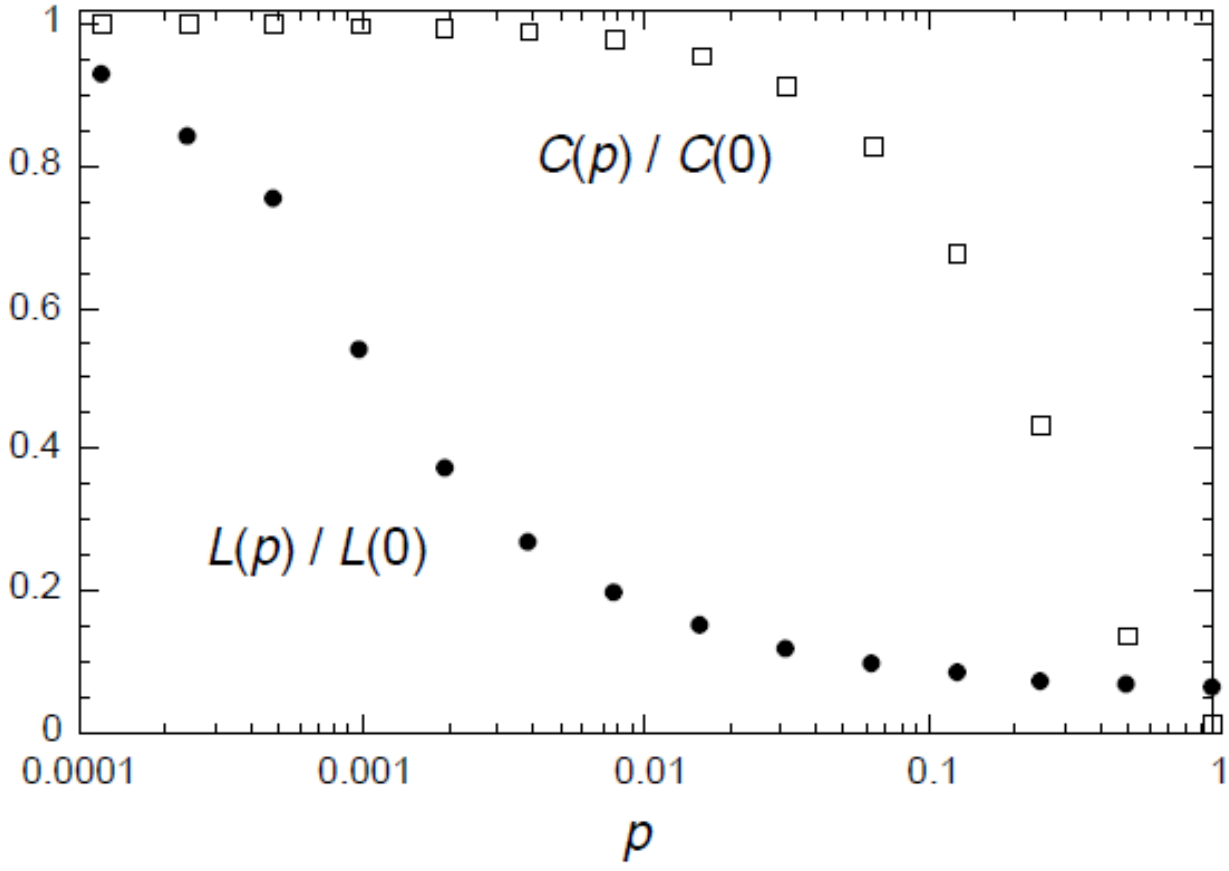
Modelling small worlds

We quantify the properties of the network with two measures:

1. C: the clustering coefficient (the extent to which friends of ego are also friends)
2. L: average path length (the average number of links in the shortest path between each ego and alter)

Source: Watts (2004), 'The New Science of Networks'

Modelling small worlds



C(p)/C(0) = standardized clustering (relative to clustering when p = 0)
L(p)/L(0) = standardized mean path length (relative to when p = 0)

Source: Watts (2004), 'The New Science of Networks'

Why 'rewiring?' weak ties

One way of interpreting the 'rewired' ties in the small world is as **weak ties**.

Granovetter (1973) argued that 'weak' ties are likely to provide more valuable information than strong ties, especially in a job-seeking context. Why?

1. Ties to socially 'distant' clusters are likely to bring non-redundant information, because close ties already have all the same information.
2. Those 'long' ties are likely to be weak in the sense of involving relatively infrequent or less emotionally intimate interaction.

Weak ties are 'bridges' between clusters: the concept of betweenness centrality.

Implications of the small world phenomenon

Many observed networks have been argued to have small world properties (e.g. the 'Hollywood graph', citation networks).

Any two randomly selected individuals are fairly 'close' in social networks.

Information, diseases, beliefs can spread quickly (even in highly clustered networks).

- Speed of disease contagion varies with p (the frequency of 'weak ties')

But, is disease contagion a good model for sociology?

Complex contagions

Sociologically interesting diffusion processes are often unlike diseases.

- Adoption of innovations, collective action, fashion

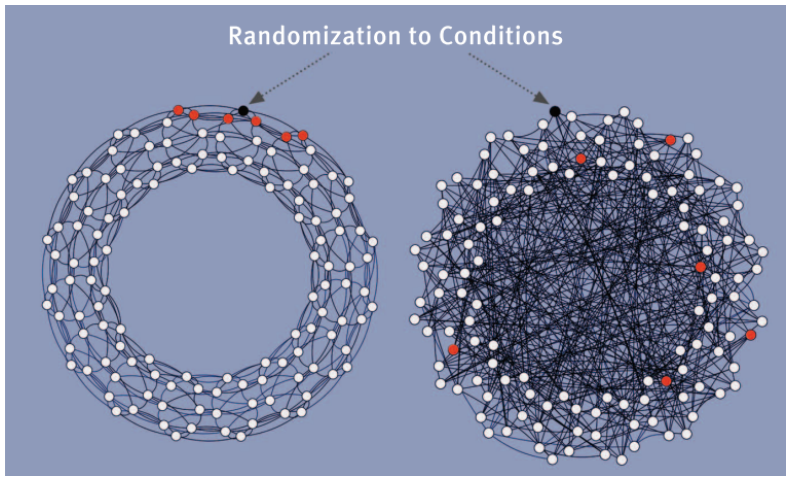
Complex contagions require exposure to **multiple sources**: thresholds for adoption.

- Did you sign up for facebook after hearing about it once, or once many of your friends had already signed up?
- Your social network includes both adopters and **non-adopters**

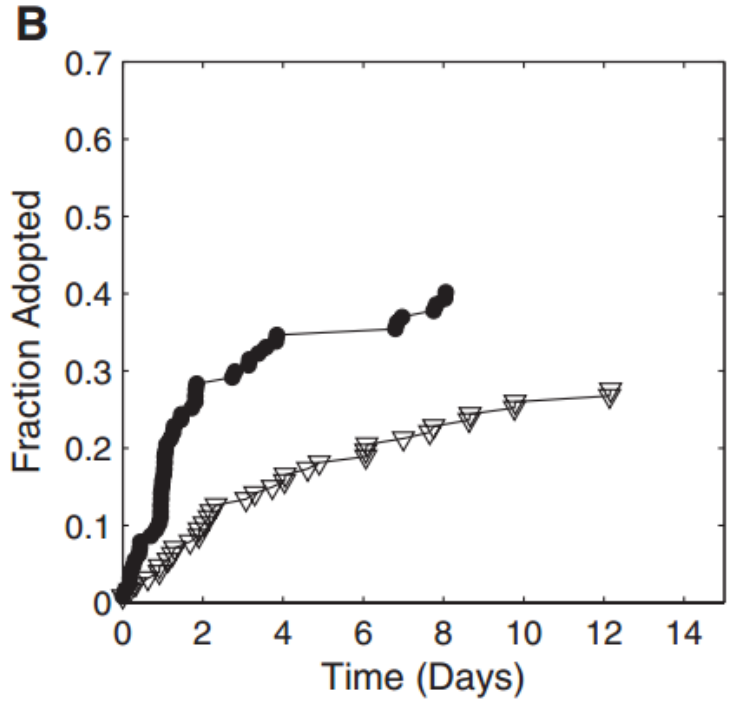
Key implication: highly clustered networks (networks with a clear community structure) can spread memes more efficiently than 'small worlds'.

Source: Centola and Macy (2007), 'Complex contagions and the weakness of long ties'

Centola (2010) created an online community in which participants were assigned to a **highly clustered** OR **randomly rewired** network (see lecture 2 on small worlds)



Left: highly clustered; Right: randomly rewired



Black circles: highly clustered; Open triangles: random

Diffusion was faster in highly clustered networks (contrary to small world theory)

Centola, Damon. 2010. "The Spread of Behavior in an Online Social Network Experiment." 329:5. See also Centola (2018), previous slide.

Collective action as contagion

Sociologists want to know: why do people participate in risky collective action (e.g. protests that may turn violent)?

- Imagine an authoritarian regime with an incipient pro-democracy movement.
- Citizens face classic collective action problem: participate in a risky protest or not.
- Information about others joining encourages participation, but sharing information is itself risky.
- Information travels only through relationships of trust.
- Individuals probably require multiple exposures to information and have varying thresholds for participation.
- What is the likelihood that protest participation will cascade and reach a tipping point?

Small worlds: summary

1. The Milgram study popularized the notion of 'six degrees of separation'
2. Formal models suggest that in the presence of high clustering plus some weak ties, social networks are 'small'
3. Small world model suggests rapid contagion, e.g. of diseases
4. Research frontier: more complex/realistic models of diffusion in social contexts. See:
   - Guilbeault et al (2018) Complex contagions: A Decade in Review.
   - Goldberg and Stein (2018) Beyond Social Contagion: Associative Diffusion and the Emergence of Cultural Variation.

Exercise I: Network visualization

Data: frequency and type (online/in person) of interaction among students on my social networks course this semester.

Data in edgelist format:

Ego	**Alter**
Maria	Satoshi
Maria	Janet
Janet	Pierre

This is a standard and usually efficient format for network data.

Node **attributes**: degree programme studied, gender.

Exercise I: Network visualization

Research question: to what extent does the degree programme that students are enrolled on structure their online and offline interactions? More frequent or less frequent interactions?

Note: degree programmes are an example of social foci (Feld 1981): institutional structures of interaction that give rise to social relationships.

Part II: Networks on social media

Why study social media? Why use SNA?

How do social influence processes shape cultural tastes and success in cultural markets (e.g. hit songs, 'going viral', etc.)?

How to network formation patterns such as homophily influence the spread of political information (e.g. 'fake news')?

Does social media enable collective action (aka 'did twitter cause the Arab spring?')?

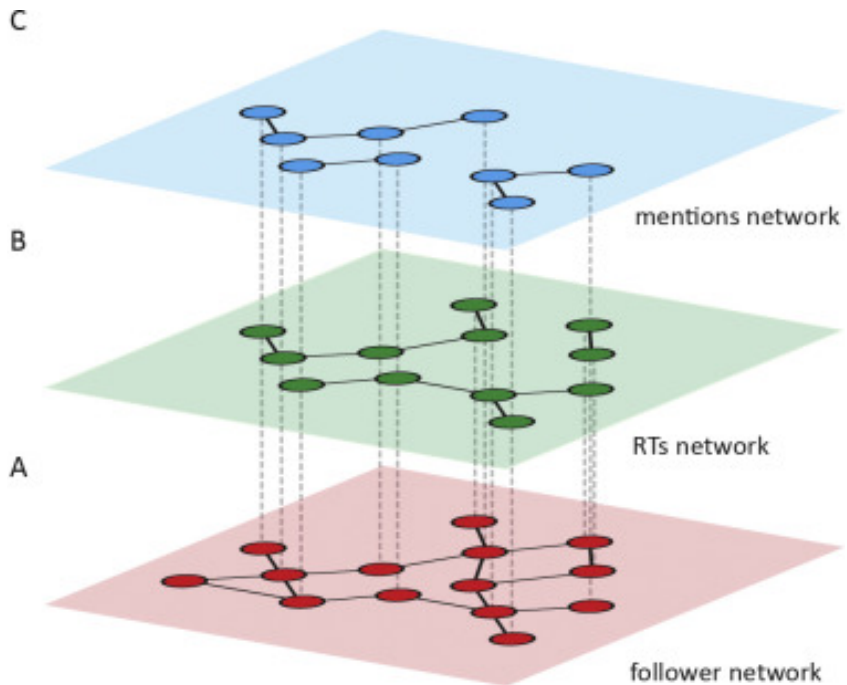
Twitter as multilayer network

Retweet networks capture information sharing (though not complete chains).

Retweets and mentions also reflect strategic behavior (e.g. 'hashjacking') native to twitter.

Following networks can be used to proxy identity (e.g. political partisanship).

Reminder that twitter $\neq$ real life.



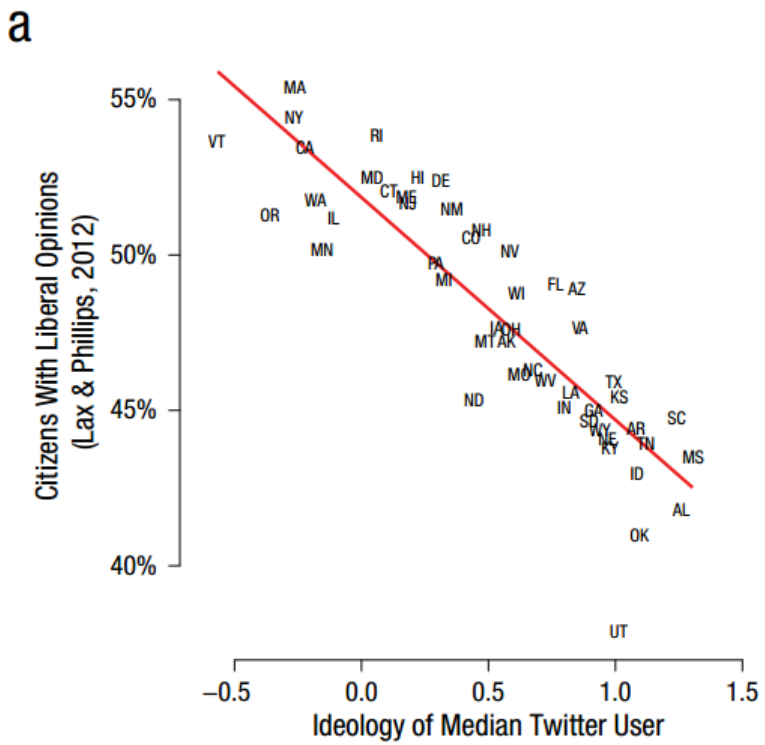
Echo chamber effect on Twitter

Do twitter users disproportionately share (retweet) information from ideologically similar others?

How can we identify twitter users' political ideology using networks?

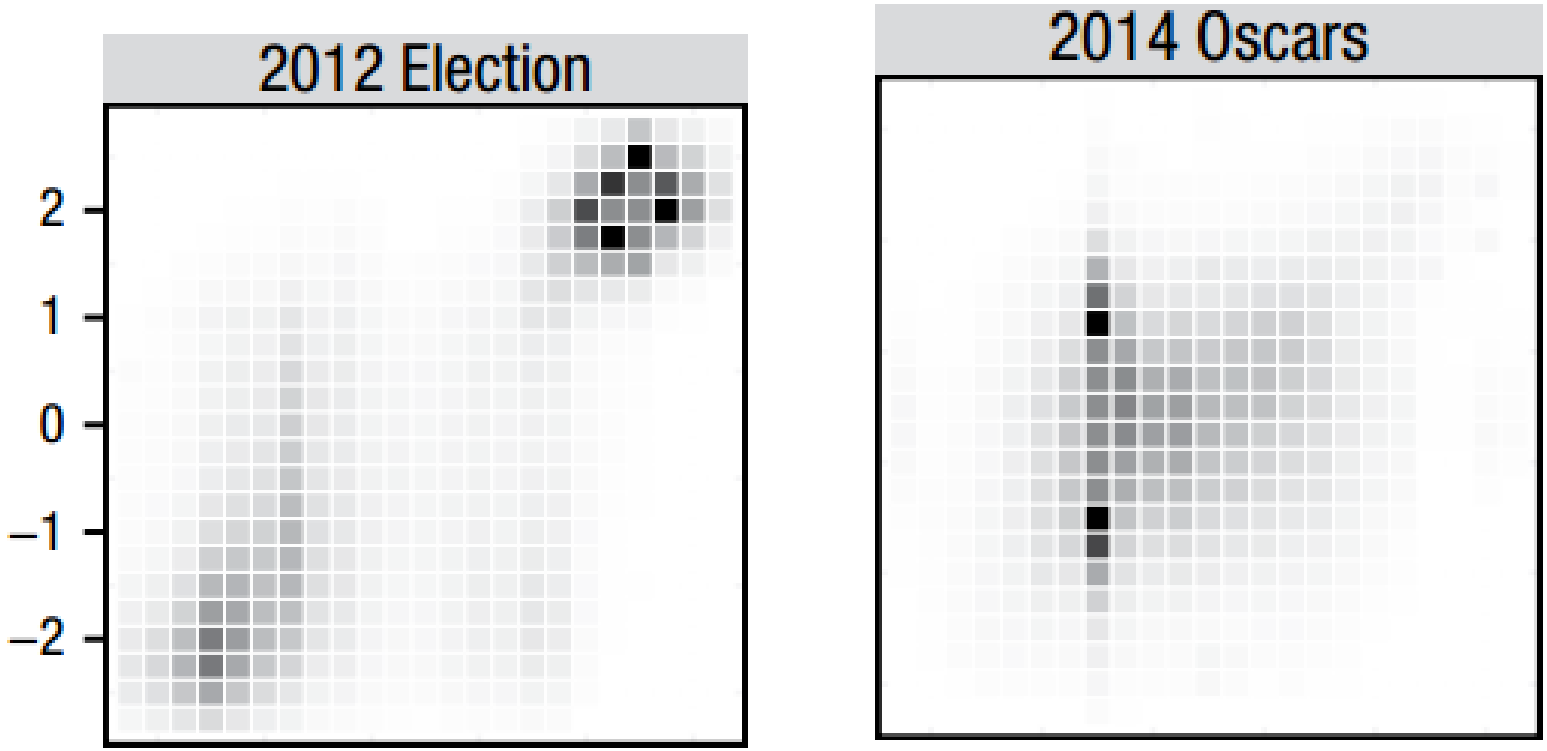
Barbera et. al. (2015) used twitter users' lists of political accounts followed to estimate ideological position.

This appears to produce a reliable measure of ideology (see figure right)



Source: Barberá, Pablo et. al. 2015. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" Psychological Science 26(10):1531–42.

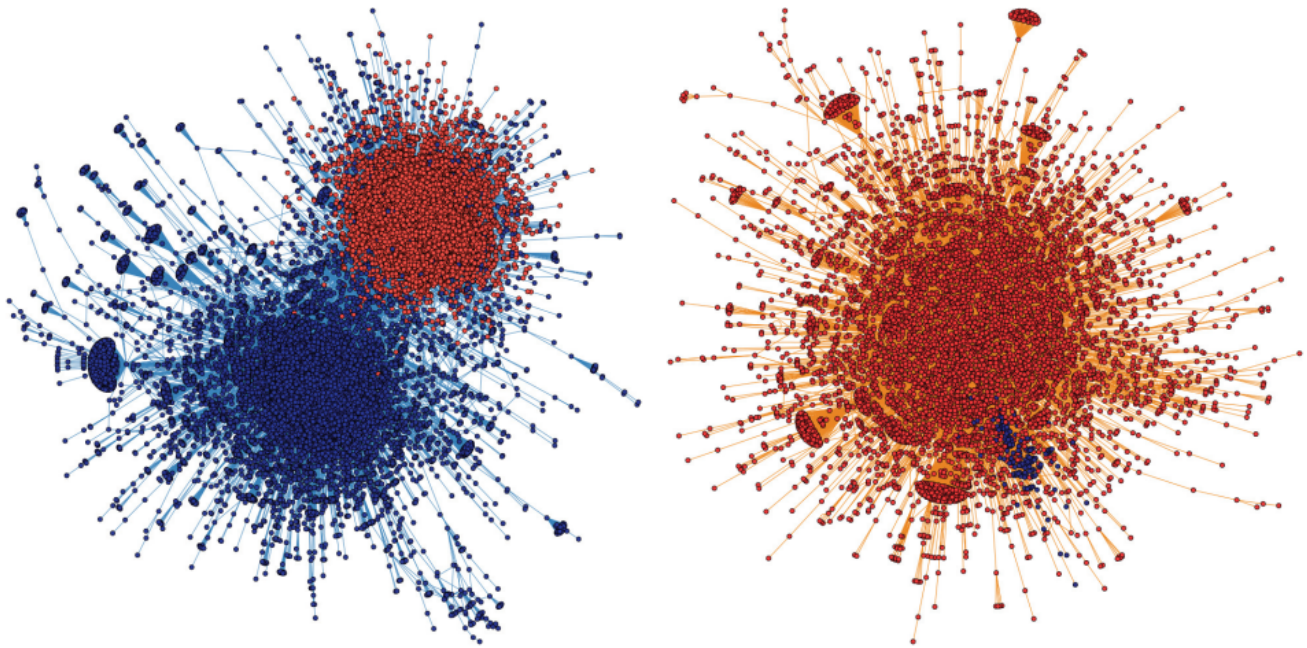
Key finding: twitter users disproportionately retweet ideologically similar others on political, but **not** non-political topics.



Authors argue that this is **not** consistent with echo chamber, because information segmentation is limited to political topics.

Political polarization, network structure and strategic behavior

Retweet (but not mention) networks form distinct communities in a political context:



Conover, M. D., J. Ratkiewicz, M. Francisco, B. Goncalves, A. Flammini, and F. Menczer. 2011. "Political Polarization on Twitter." P. 8 in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Network communities and polarization

Communities or subgroups are (roughly speaking) sets of nodes that share common ties while less connected to other sets of nodes.

- Limiting case: cliques (subgraphs in which every node is connected to ever other node)
- A more modern approach is **modularity**: communities are sets with high within-group density and low between-group density.

Community detection is an active area of reseach in network science; many available algorithms but in practice a combination of art + science.

A network consisting of discrete subgroups (between which information or opinion are unlikely to flow) is one interpretation of 'echo chamber'

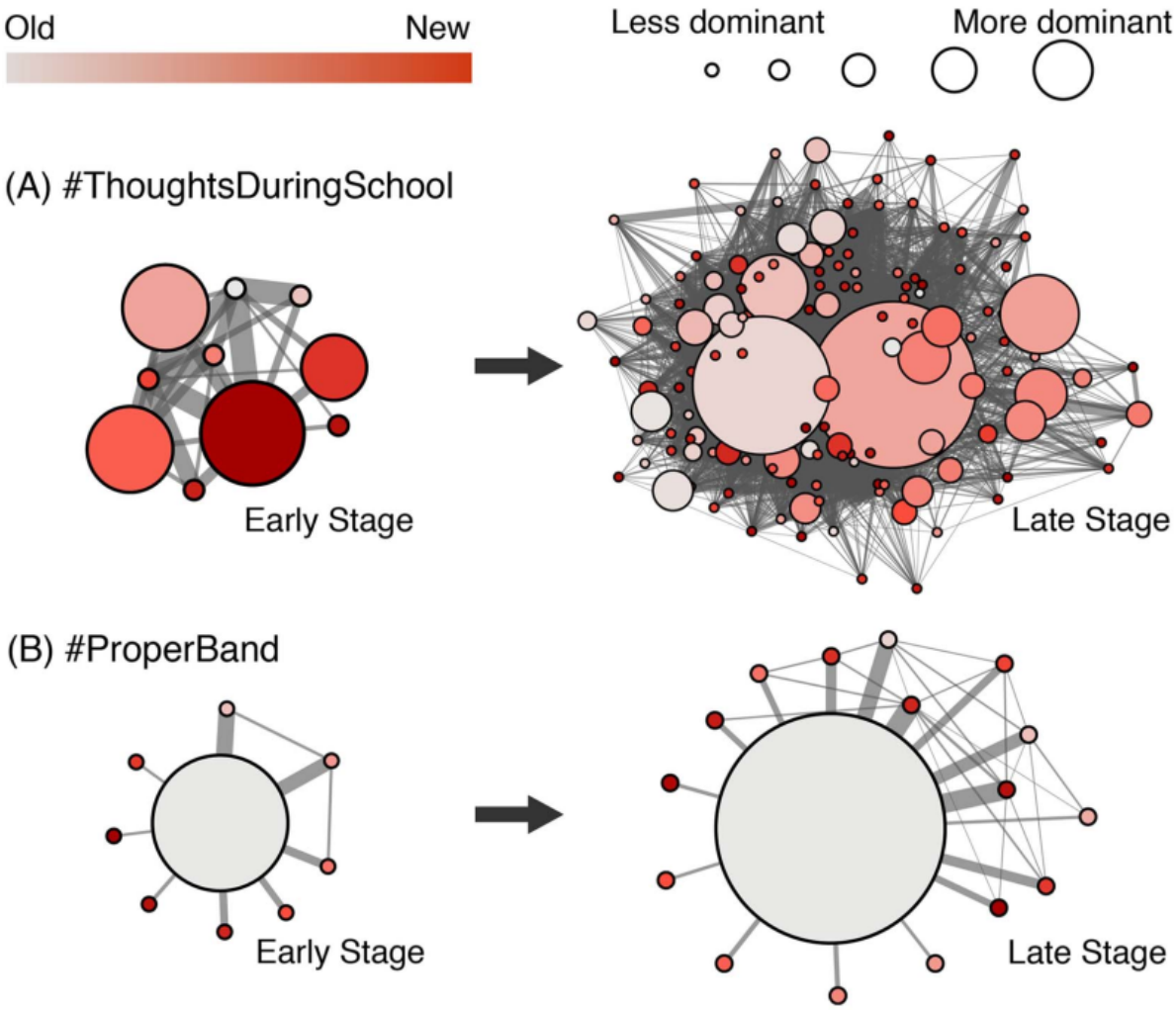
Community structure and viral diffusion

If memes spread through complex contagions, then only memes capable of jumping across communities will 'go viral.'

Memes that are highly concentrated within communities early in the diffusion process are less likely to 'go viral' (reach a large proportion of the network).

If this is right, then you can predict which memes will go viral (and make a lot of money).

Source: Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. 2013. "Virality Prediction and Community Structure in Social Networks." Scientific Reports 3(1):2522. doi: 10.1038/srep02522.



Note: nodes are communities

Source: Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. 2013. "Virality Prediction and Community Structure in Social Networks." Scientific Reports 3(1):2522. doi: 10.1038/srep02522. Figure 4.

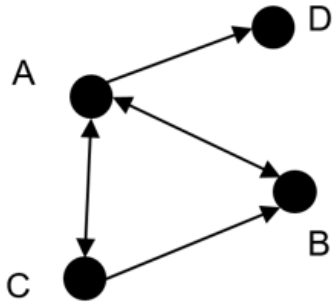
Centrality, influence and beyond...

Centrality can be used to measure influence, information brokerage, etc.

Classic measures:

- Degree centrality: each node's count of ties. In directed networks:
  - In degree: N of ties pointing at a node: intuitive measure of influence on social media.
  - Out degree: N of ties pointing way from a node.
  - Note that what counts as 'in' or 'out' depends on how you set up networks; some may approach this the opposite way.
- Closeness centrality: (inverse of) how many steps does it take (on average) to reach other nodes?
- Betweenness centrality: To what extent do paths through the network pass through a focal node?

Degree centrality



Adjacency matrix:

$$A_{ij} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} - & 1 & 1 & 0 \\ 1 & - & 1 & 0 \\ 1 & 0 & - & 0 \\ 1 & 0 & 0 & - \end{pmatrix} \end{array}$$

In degree $= \sum_{i=1}^{n} A_{ij}$ (column sum of adjacency matrix)

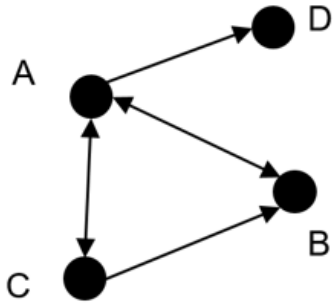
Out degree $= \sum_{j=1}^{n} A_{ij}$ (row sum of adjacency matrix)

Total centrality: sum of in + out

In some cases, we standardize by the maximum possible degree, which equals n - 1

Easily calculated in R with degree function in igraph and network packages.

Closeness centrality



Distance matrix:

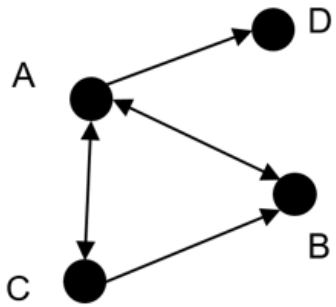
$$D_{ij} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} - & 1 & 1 & 1 \\ 1 & - & 2 & 2 \\ 1 & 1 & - & 2 \\ \infty & \infty & \infty & - \end{pmatrix} \end{array}$$

Closeness $= \frac{1}{\sum_{j=1}^{n} D_{ij}}$ (1 over row sum of distance matrix)

Important: this is undefined unless we account for disconnected nodes (infinite paths). Different packages address this in different ways.

Easily calculated with closeness function.

Betweenness centrality



g_{ij} = the number of geodesics between i and j.

$g_{ij}(p_k)$ = the number of geodesics between i and j that include k.

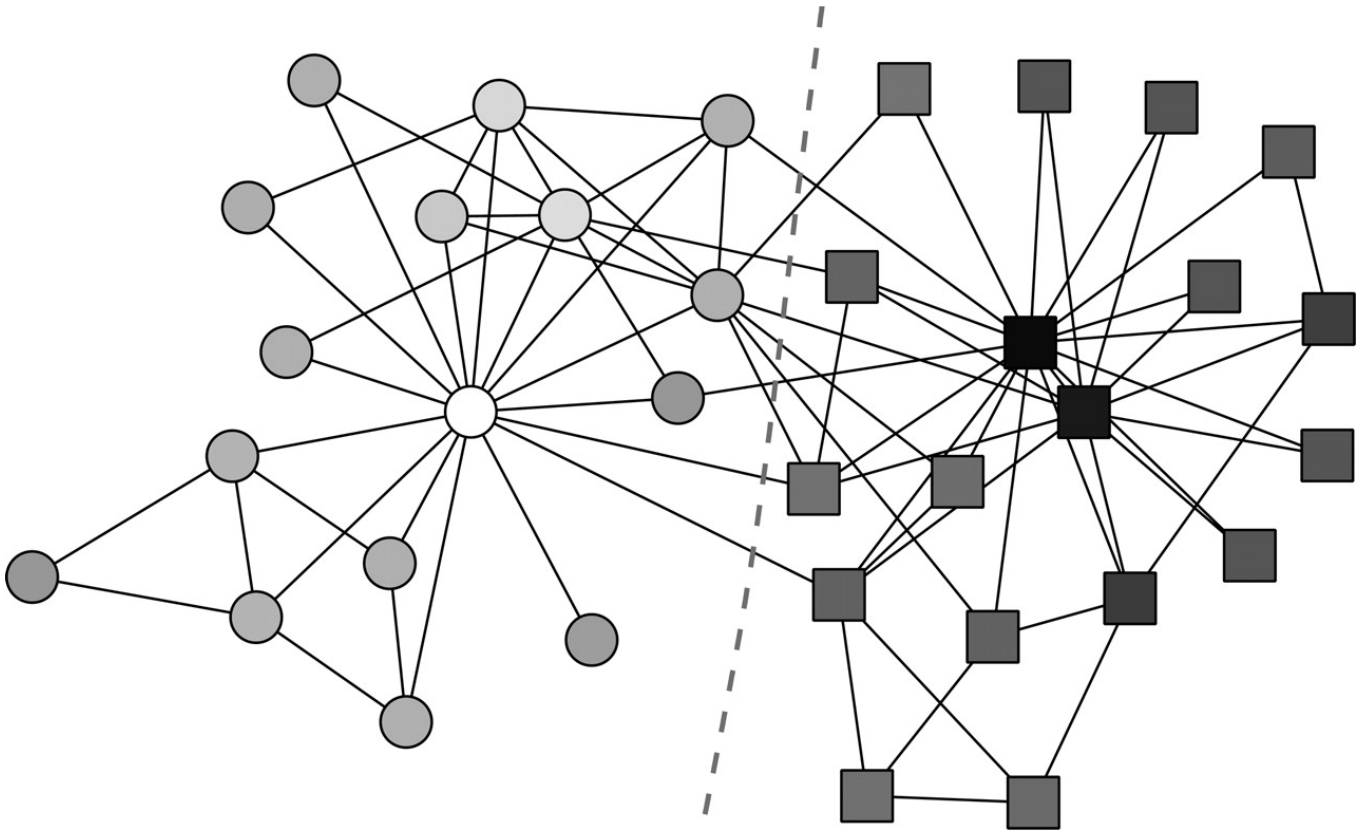
$$\text{Betw(k)} = \sum_{i \neq j \neq k} \frac{g_{ij}(p_k)}{g_{ij}}$$

Intuitively: the percentage of paths between i and j which pass through k.

Paths can be treated as direct or undirected.

Detecting communities: modularity

Karate club divided 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

Twitter network exercise

Data: 1K tweets by UK MPs between 2017-2019

- Includes much of the period of Brexit negotiation; Teresa May resigned July 2019.
- Also used tomorrow for word embedding exercise.

Research questions:

- How polarized is UK MP twitter – is it just an 'echo chamber'? How does party affiliation structure twitter behavior? Is there any time trend in the level of polarization?
- Which MPs were most influential, e.g. retweeted by most others MPs?
- Which MPs were 'opinion brokers' between parties?

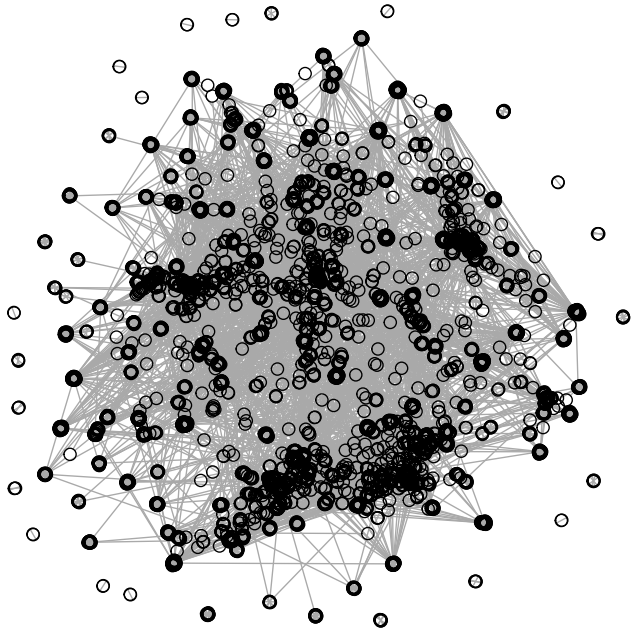
Overview of analytical approach

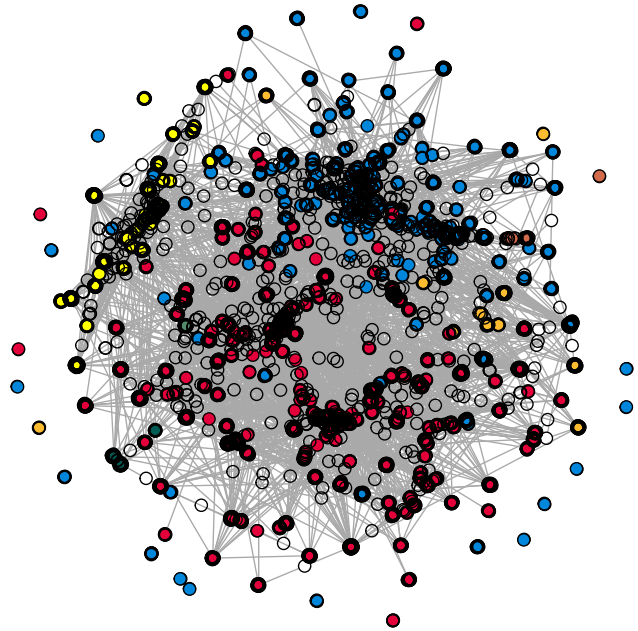
Data structure:

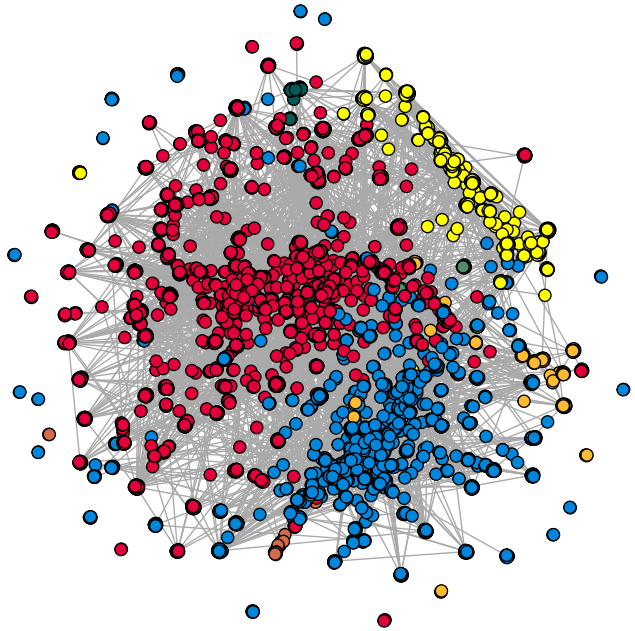
- Retweets: ties pointing **from** original tweet account to re-tweeting account.
- Mentions: ties pointing **from** account mentioning to account mentioned.

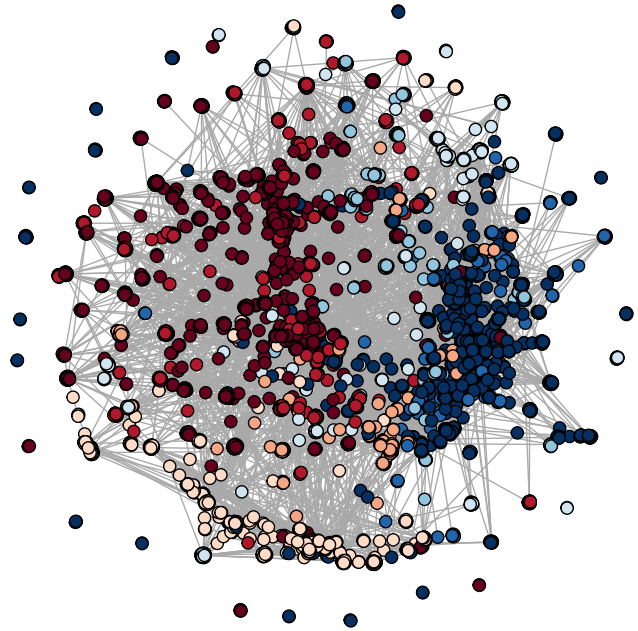
Measures:

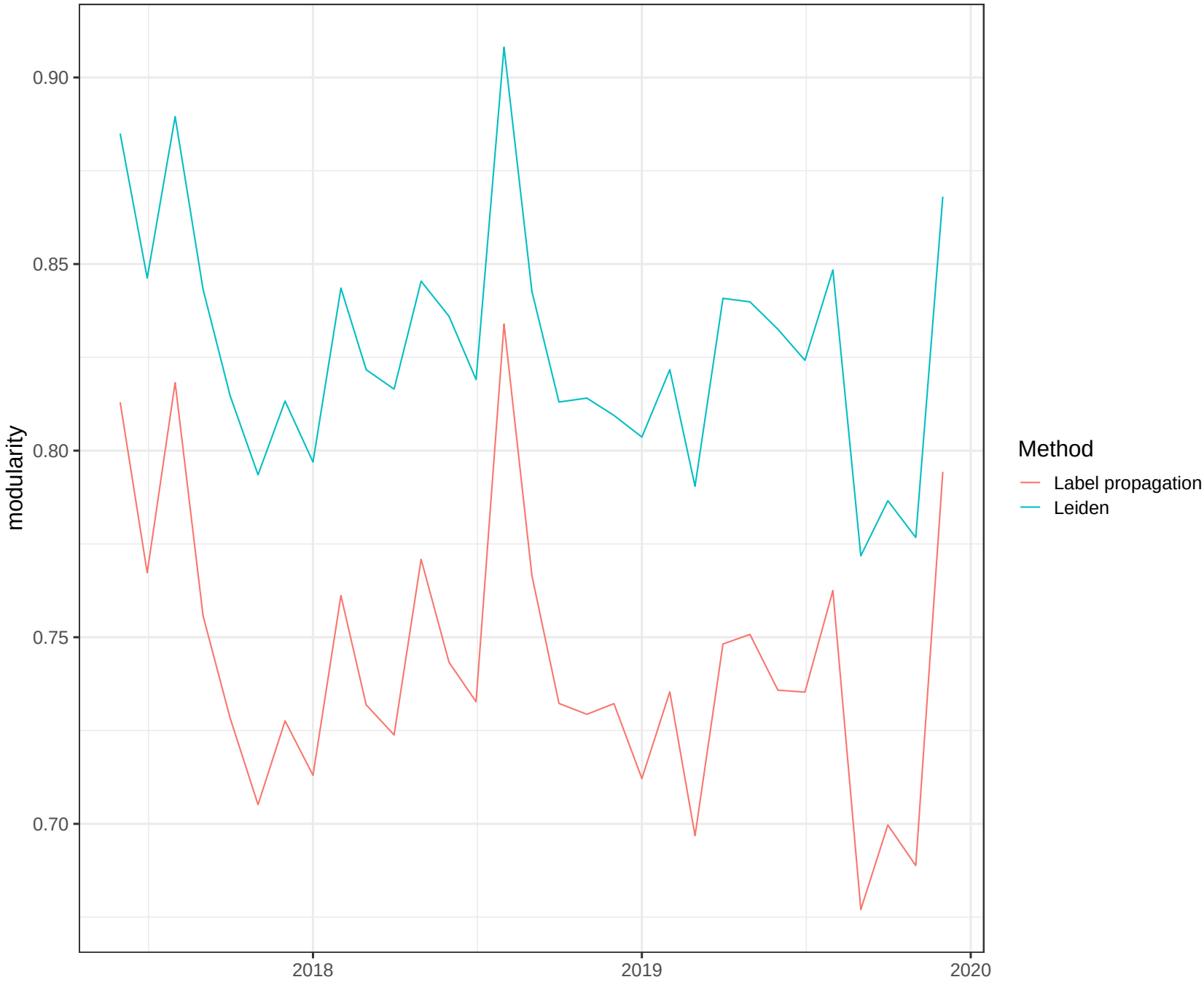
- In-degree as measure of influence
- Betweenness centrality as measure of information/opinion brokerage.
- Visualization, community detection and modularity score as tools to assess polarization.










modularity
0.90
0.85
0.80
0.75
0.70
2018
2019
2020
Method
Label propagation
Leiden