COMP30120

Introduction to Social Network Analysis

Derek Greene

School of Computer Science and Informatics

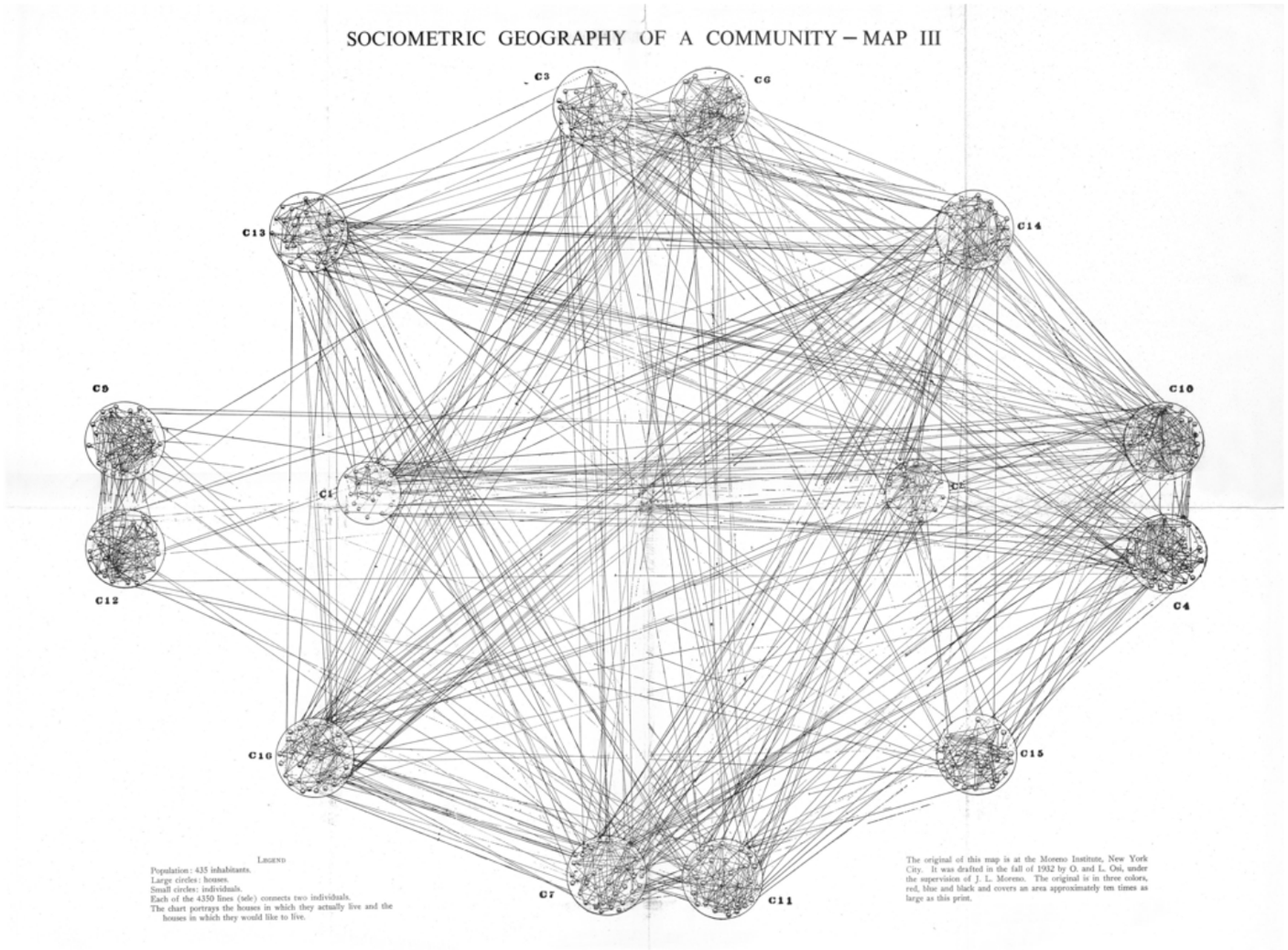
Autumn 2015

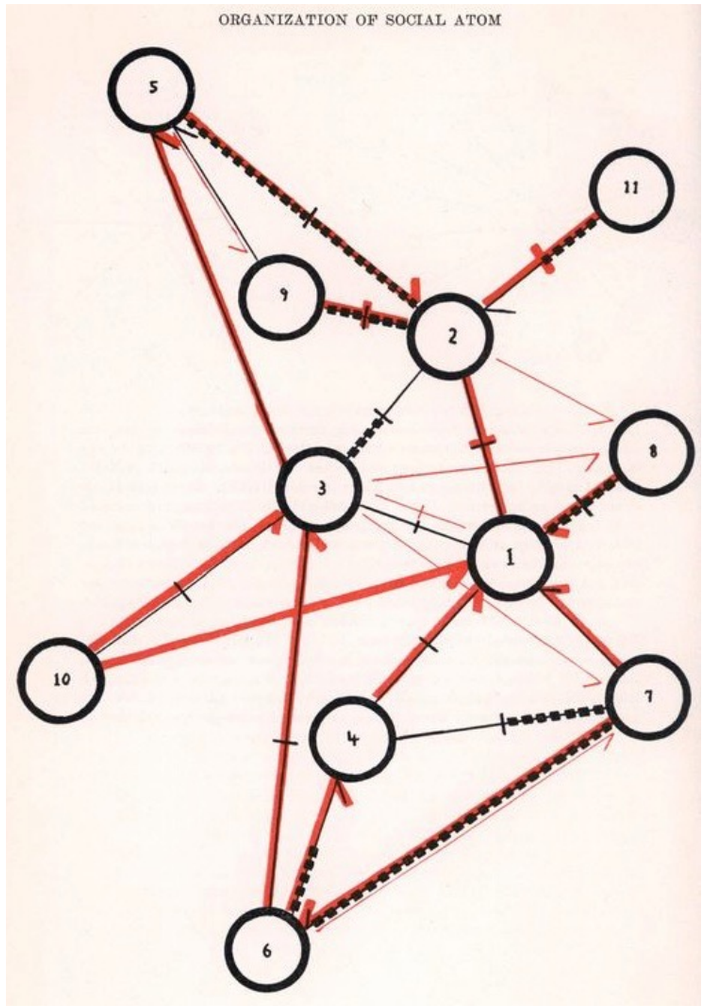
Overview

- Social Network Analysis
- Graphs - Basic Concepts
- Characterising Graphs
  - Density
  - Connectivity
- Measures of Centrality
- Homophily
- Clustering Coefficient
- Ego Networks

Social Network Analysis

Social network analysis - an old field, rediscovered...





J. Moreno. "Who shall survive?: A new approach to the problem of human interrelations". Nervous and Mental Disease Publishing Co., 1934

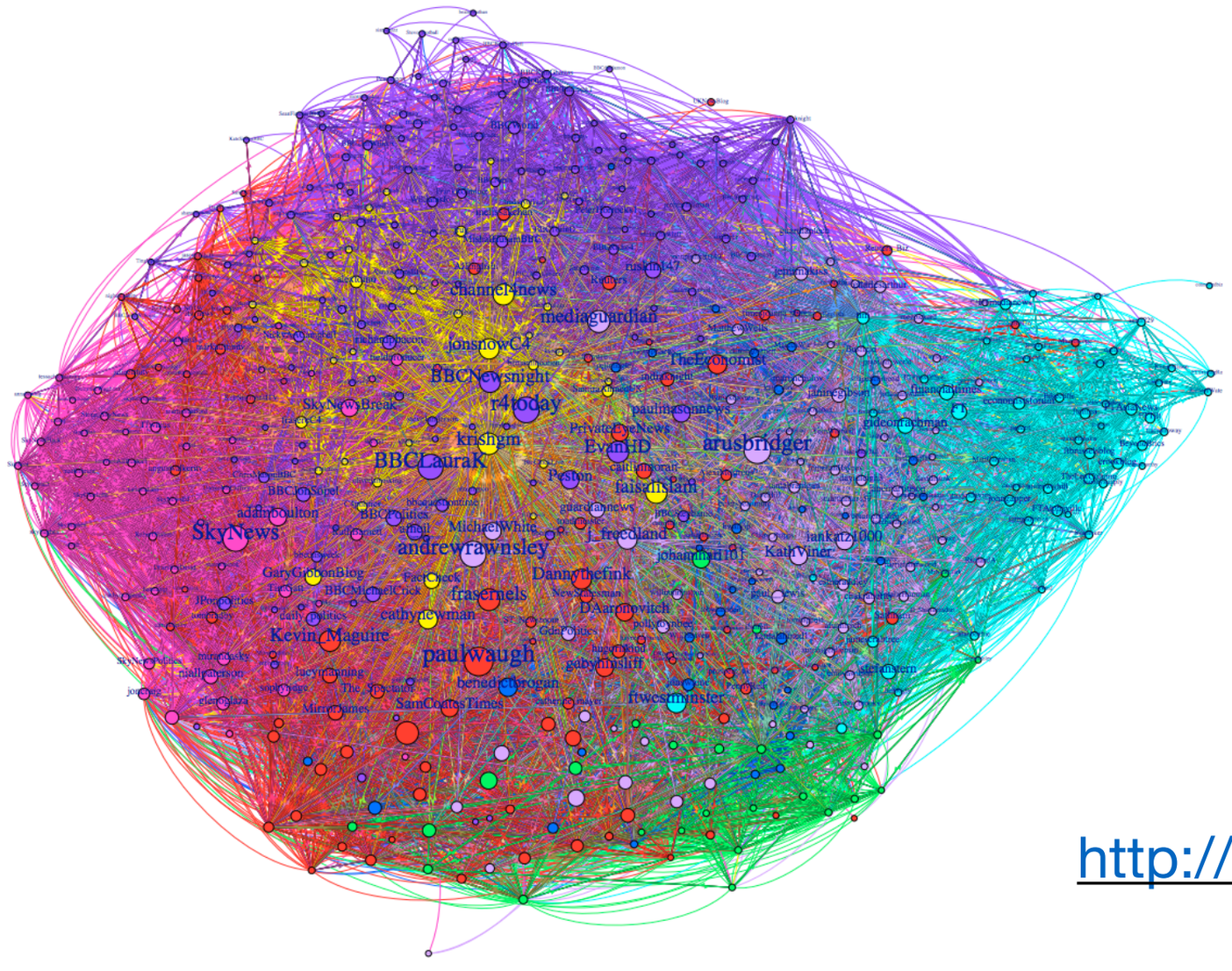
Social Network Analysis

We now have the computational resources to perform network analysis on a large scale...

http://www.facebook.com/note.php?note_id=469716398919

Social Network Analysis

Example: Network of 524 UK political & current affairs journalists on Twitter, produced by The Guardian in 2011.

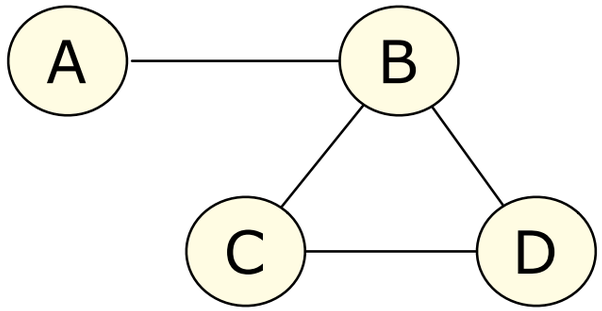


http://gu.com/p/2zbq7/stw

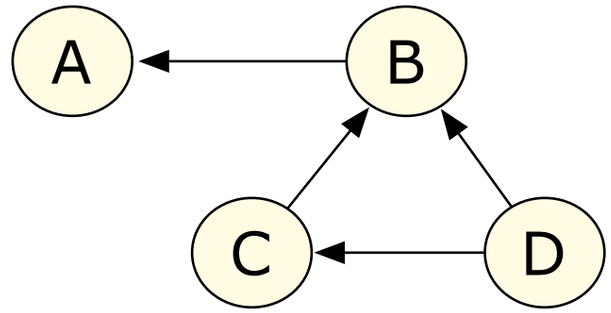
Q. How can we characterise a social network with ~500 users? What about a network with ~500m users?

Basic Concepts

- Graph: a way of representing relations among a collection of items.
- Consists of a set of items, called nodes, with certain pairs of items connected to one another by relations called edges.



Undirected graph

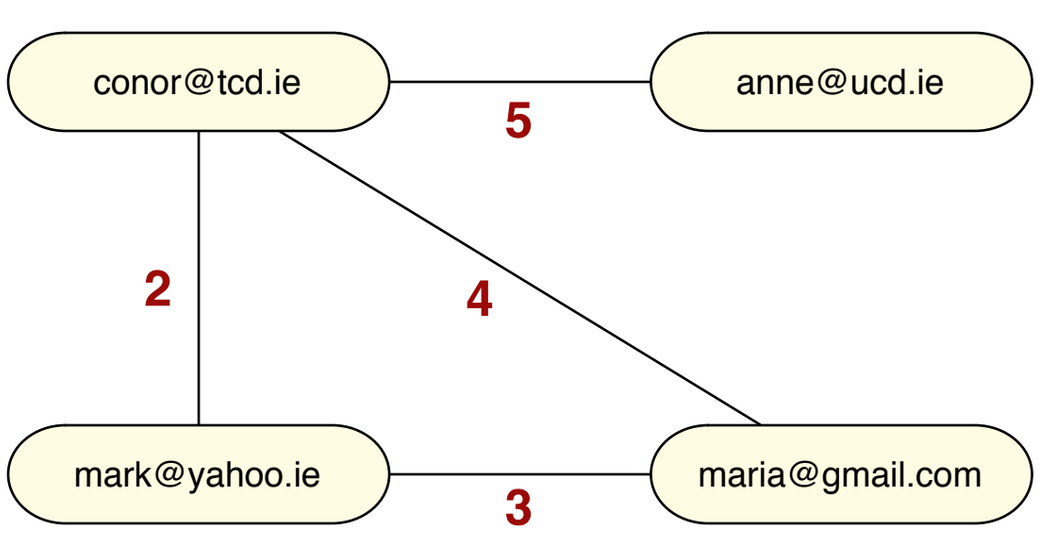


Directed graph

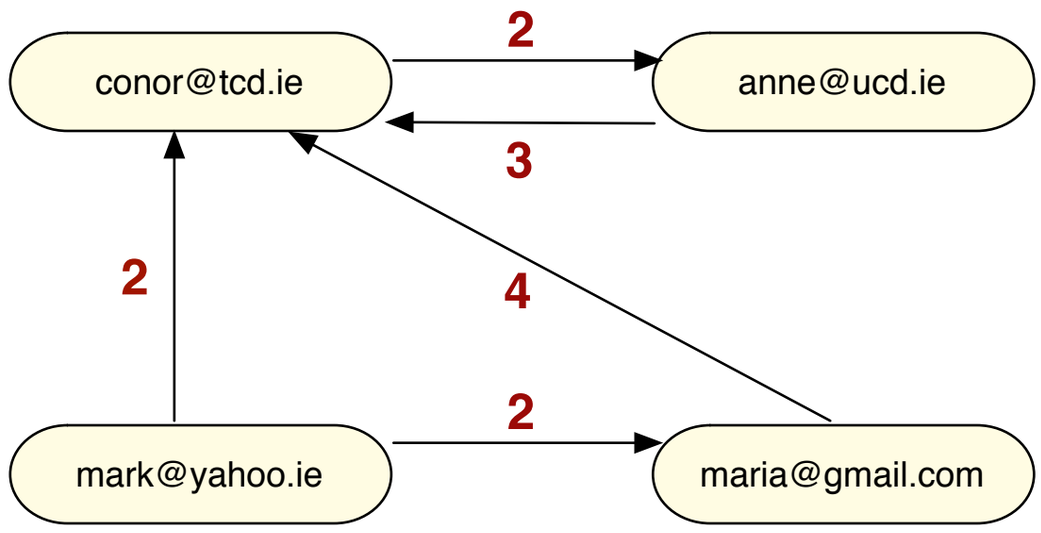
- Two nodes are neighbours if they are connected by an edge.
- Degree of a node is the number of edges ending at that node.
- For a directed graph, the in-degree and out-degree of a node refer to numbers of edges incoming to or outgoing from the node.

Weighted Graphs

- Weighted graph: numeric value is associated with each edge.
- Edge weights may represent a concept such as similarity, frequency, or strength of association.



Undirected weighted graph



Directed weighted graph

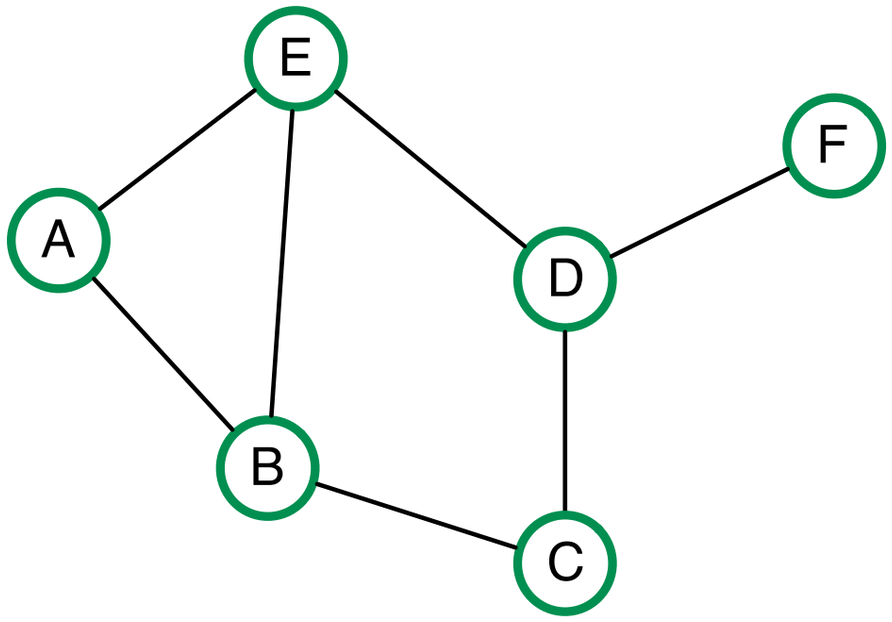
Representing Undirected Graphs

Many different ways to represent same graph

- Network visualisation
- Edge List
- Adjacency Matrix

Edge List with 7 edges

(A, B)
(A, E)
(D, E)
(B, E)
(B, C)
(C, D)
(D, F)



Undirected Graph
6 nodes, 7 edges

	A	B	C	D	E	F
A	-	1	0	0	1	0
B	1	-	1	0	1	0
C	0	1	-	1	0	0
D	0	0	1	-	1	1
E	1	1	0	1	-	0
F	0	0	0	1	0	-

6x6 Symmetric Adjacency Matrix
for Undirected Graph

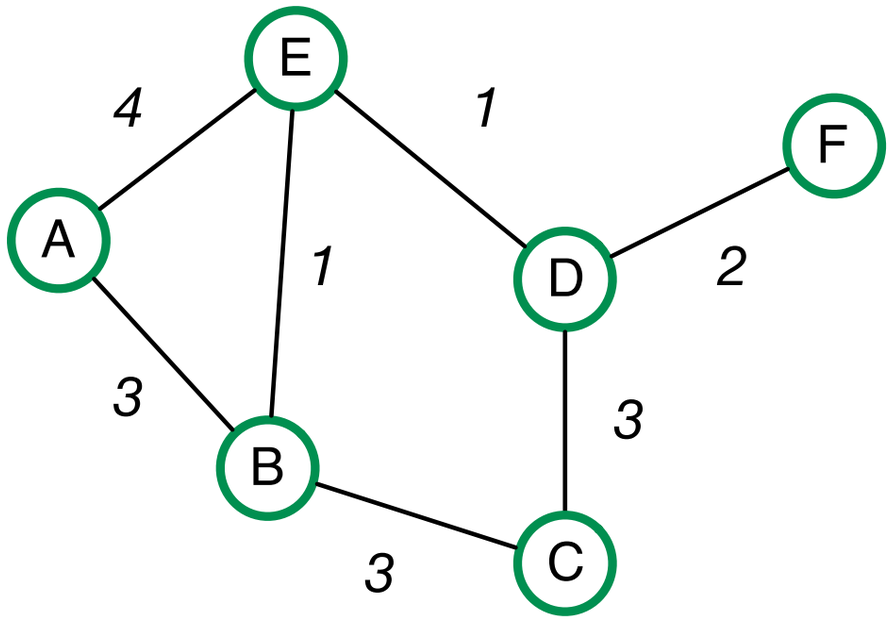
Representing Weighted Graphs

Many different ways to represent same graph

- Network visualisation
- Edge List
- Adjacency Matrix

Weighted Edge List with 7 edges

(A, B, 3)
(A, E, 4)
(D, E, 1)
(B, E, 1)
(B, C, 3)
(C, D, 3)
(D, F, 2)



Weighted Undirected Graph
6 nodes, 7 edges

	A	B	C	D	E	F
A	-	3	0	0	4	0
B	3	-	3	0	1	0
C	0	3	-	3	0	0
D	0	0	3	-	1	2
E	4	1	0	1	-	0
F	0	0	0	2	0	-

6x6 Symmetric Adjacency Matrix
for Weighted Undirected Graph

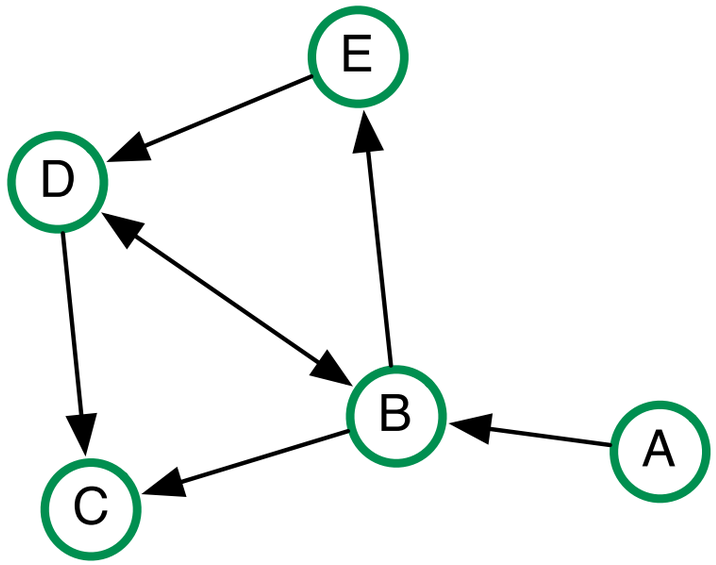
Representing Directed Graphs

Many different ways to represent same graph

- Network visualisation
- Edge List
- Adjacency Matrix

Directed Edge List
with 7 edges

(A, B)
(B, C)
(D, C)
(E, D)
(B, E)
(D, B)
(B, D)



Directed Graph
5 nodes, 7 edges
(1 reciprocal pair)

	A	B	C	D	E
A	-	1	0	0	0
B	0	-	1	1	1
C	0	0	-	0	0
D	0	1	1	-	0
E	0	0	0	1	-

5x5 Asymmetric Adjacency Matrix
for Directed Graph

Graph Density

- A dense graph is a graph in which the number of edges is close to the maximum possible number of edges.
- Measure graph density as the number of edges over number of possible edges:

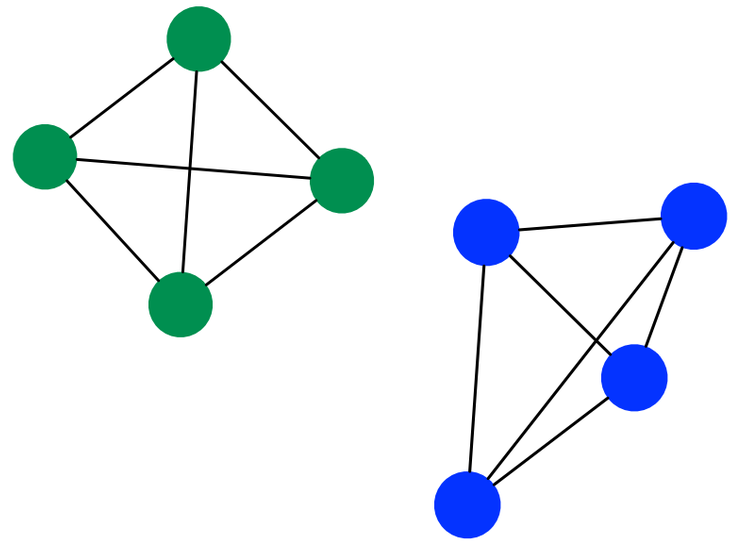
Undirected graph with *n* nodes and *m* edges

$$\frac{2 \times m}{n \times (n-1)}$$

Directed graph with *n* nodes and *m* edges

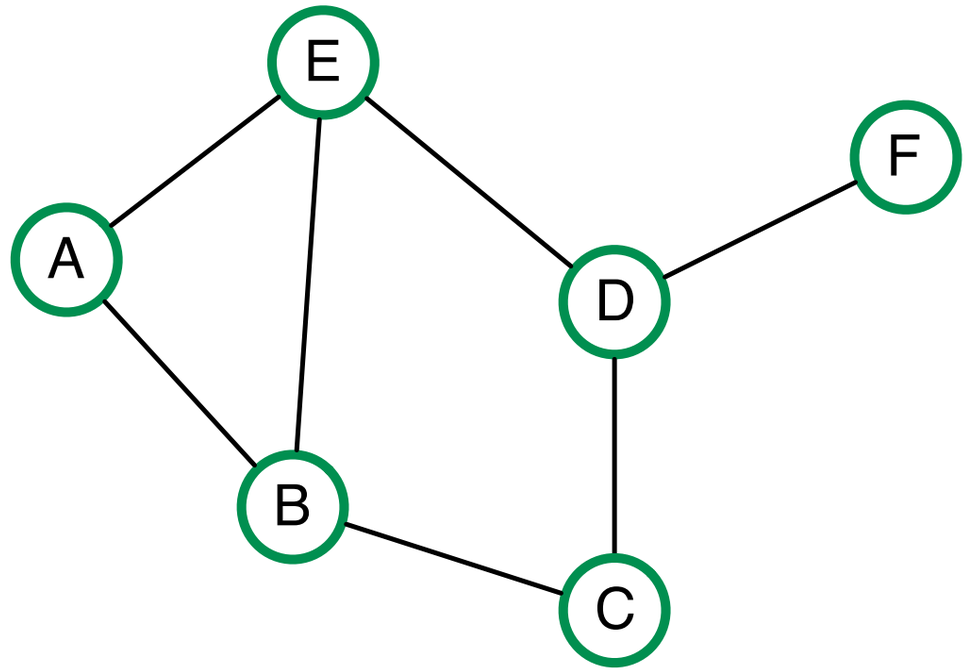
$$\frac{m}{n \times (n-1)}$$

- Most large real-world social networks will be highly-sparse i.e. they have very low density.
- In a fully-connected graph, an edge exists between every pair of nodes in the graph
  ⇒ Graph density = 1
- Clique: a fully-connected sub-graph within a larger graph.



Graph Connectivity: Paths

- Walk: A sequence of nodes in a graph, where each node is adjacent to the next (i.e. there is an edge between them).
- Path: A walk, where each edge is never crossed more than once. The path length is the number of edges.



Paths from A to C:
[A, B, C]
[A, B, E, D, C]
[A, E, B, C]
[A, E, D, C]

Paths from A to E:
[A, B, C, D, E]
[A, B, E]
[A, E]

- Shortest Path: A path between two nodes in the graph with minimal cost (i.e. minimum number of edges).
- Graph's diameter: longest shortest path over all pairs of nodes.

Shortest path from A to C: [A, B, C]

Shortest path from A to E: [A, E]

Diameter of graph: 3

Small World Networks

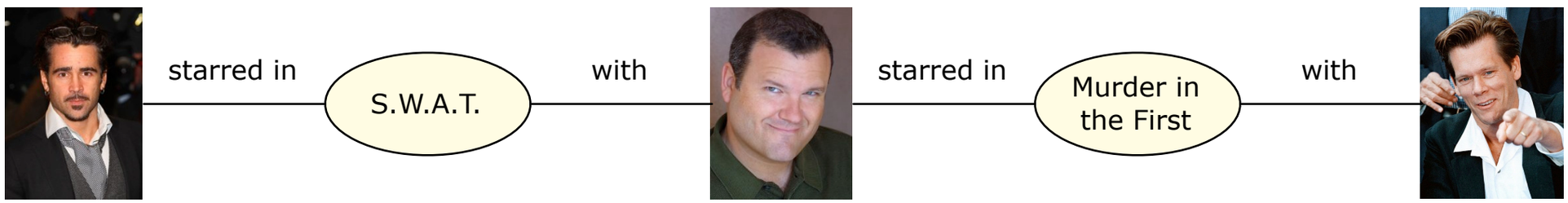
Milgram's Small World Experiment:

- Route a package to a stockbroker in Boston by sending them to random people in Nebraska and requesting them to forward to someone who might know the stockbroker.
- Although most nodes are not directly connected, each node can be reached from another via a relatively small number of hops.

http://en.wikipedia.org/wiki/Small-world_experiment

Six Degrees of Kevin Bacon

- Examine the actor-actor "co-starred" graph from IMDB.
- The Bacon Number of an actor is the number of degrees of separation he/she has from Kevin Bacon, via the shortest path.



$\Rightarrow$ Bacon Number $= 2$

http://oracleofbacon.org

Small World Networks

- Degree of Separation: Average value of the shortest-path length for all pairs of nodes.
- Small World: Sparse graph in which most nodes are not neighbours of one another. But most nodes can be reached from every other node by a small number of hops.
- **Example:** In 2012 Suzumura et al crawled “complete” Twitter follower network of 826 million nodes and 29 billion edges.

Degree of Separation		Diameter	
2010	2012	2010	2012
4.50	4.59	26	71

Far less than 6 degrees of separation! How is this possible?

Twitter users			Followers
1		KATY PERRY @katyperry	59,448,740
2		Justin Bieber @justinbieber	56,435,312
3		Barack Obama @BarackObama	49,220,306
4		Taylor Swift @taylorswift13	46,210,940
5		YouTube @YouTube	46,273,775

Many “hubs” on Twitter mediating the short path lengths.

Small World Networks

Small world phenomenon is common across many types of networks…

NEWS TECHNOLOGY

Home | World | UK | England | N. Ireland | Scotland | Wales | Business | Politics | Health | Education | Sci/Environ

23 November 2011 Last updated at 08:54

Facebook users average 3.74 degrees of separation

There are on average 3.74 degrees of separation between any one Facebook user and another, a study suggests.

The number of degrees represents the number of people in a friendship chain, excluding the people at either end.

Or, as the authors put it: "When considering another person in the world, a friend of your friend knows a friend of their friend."

Facebook founder Mark Zuckerberg's public not reveal how many friends he has

The study was carried out in May and involved all of the social network's active members.

Facebook defines a user as active if they have logged on at least once over the past 28 days.

Four Degrees of Separation

Lars Backstrom* Paolo Boldi† Marco Rosa† Johan Ugander* Sebastiano Vigna†

January 6, 2012

Abstract

Frigyes Karinthy, in his 1929 short story "Láncszemek" ("Chains") suggested that any two persons are distanced by at most six friendship links.[1] Stanley Milgram in his famous experiment [20, 23] challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. The average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen.

We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (≈ 721 million users, ≈ 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", showing that the world is even smaller than we expected, and prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time.

The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

1 Introduction

studying the distance distribution of very large graphs: HyperANF [3]. Building on previous graph compression [4] work and on the idea of diffusive computation pioneered in [21], the new tool made it possible to accurately study the distance distribution of graphs orders of magnitude larger than it was previously possible.

One of the goals in studying the distance distribution is the identification of interesting statistical parameters that can be used to tell proper social networks from other complex networks, such as web graphs. More generally, the distance distribution is one interesting *global* feature that makes it possible to reject probabilistic models even when they match local features such as the in-degree distribution.

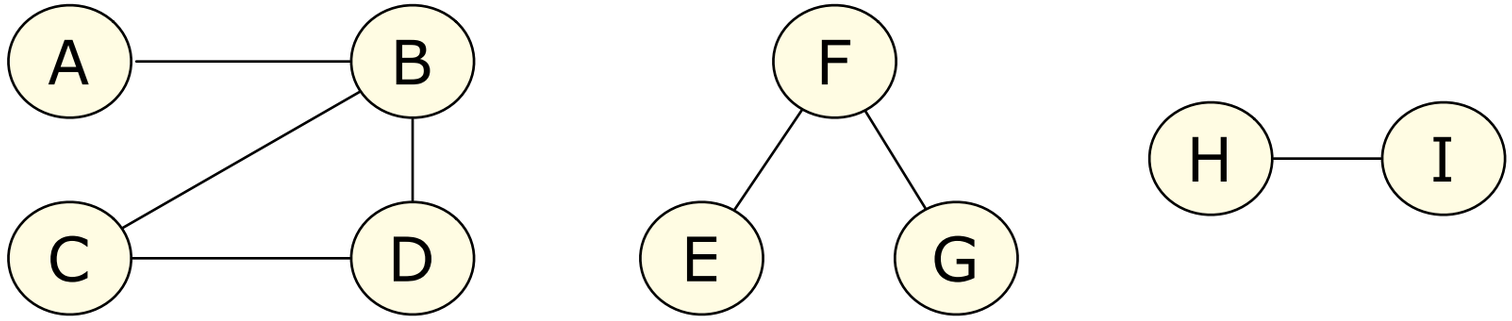
In particular, earlier work had shown that the $spid^2$, which measures the *dispersion* of the distance distribution, appeared to be smaller than 1 (underdispersion) for social networks, but larger than one (overdispersion) for web graphs [3]. Hence, during the talk, one of the main open questions was "What is the spid of Facebook?".

Lars Backstrom happened to listen to the talk, and suggested a collaboration studying the Facebook graph. This was of course an extremely intriguing possibility: beside testing the "spid hypothesis", computing the distance distribution of the Facebook graph would have been the largest Milgram-like [20] experiment ever performed, orders of magnitudes larger than previous attempts (during our experiments Facebook has ≈ 721 million active users and ≈ 69 billion friendship links).

http://arxiv.org/abs/1111.4570

Graph Connectivity: Components

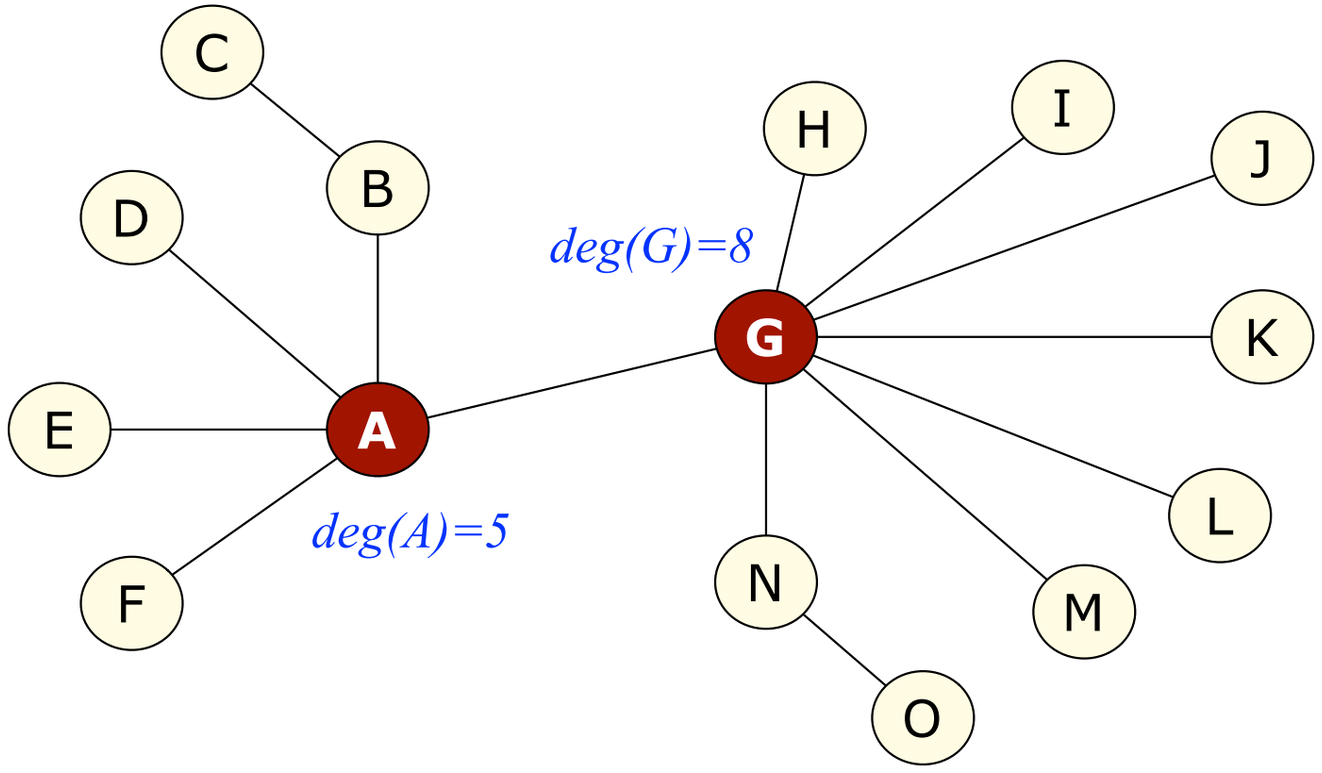
- A graph is connected if there is a path between every pair of nodes in the graph.
- A connected component is a subset of the nodes where:
  1. A path exists between every pair in the subset.
  2. The subset is not part of a larger set with the above property.



- In many real-world social networks, a larger proportion of all nodes will belong to a single giant component - the graph's "core".

Measures of Centrality

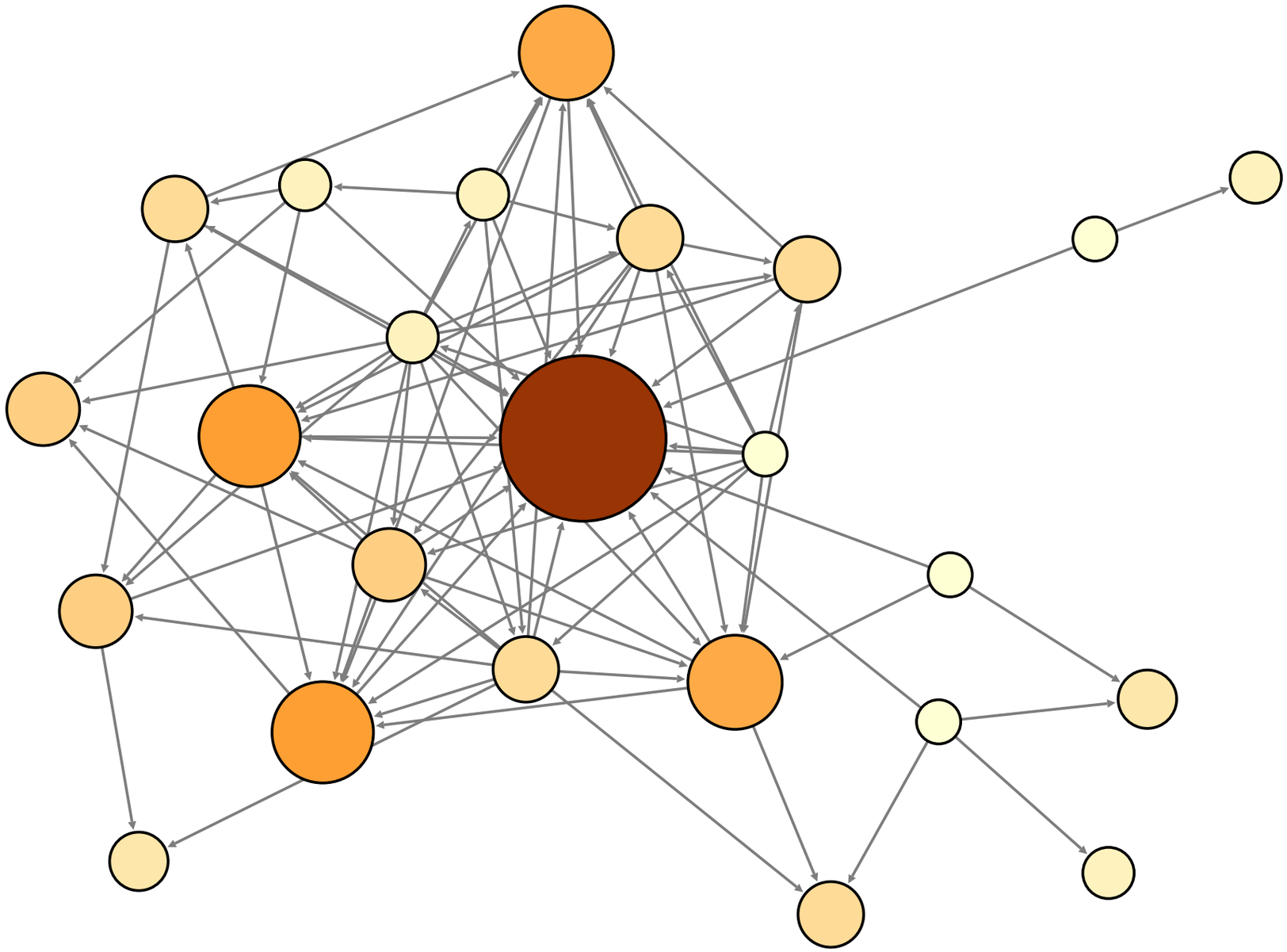
- A variety of measures exist to measure the importance, influence, popularity, or social capital of a node in a social network.
- Degree centrality focuses on individual nodes - it simply counts the number of edges that a node has.
- Hub nodes with high degree usually play an important role in a network.



Degree Centrality

- For directed networks, in-degree centrality is often used as a proxy for popularity - i.e. number of incoming edges for a node.

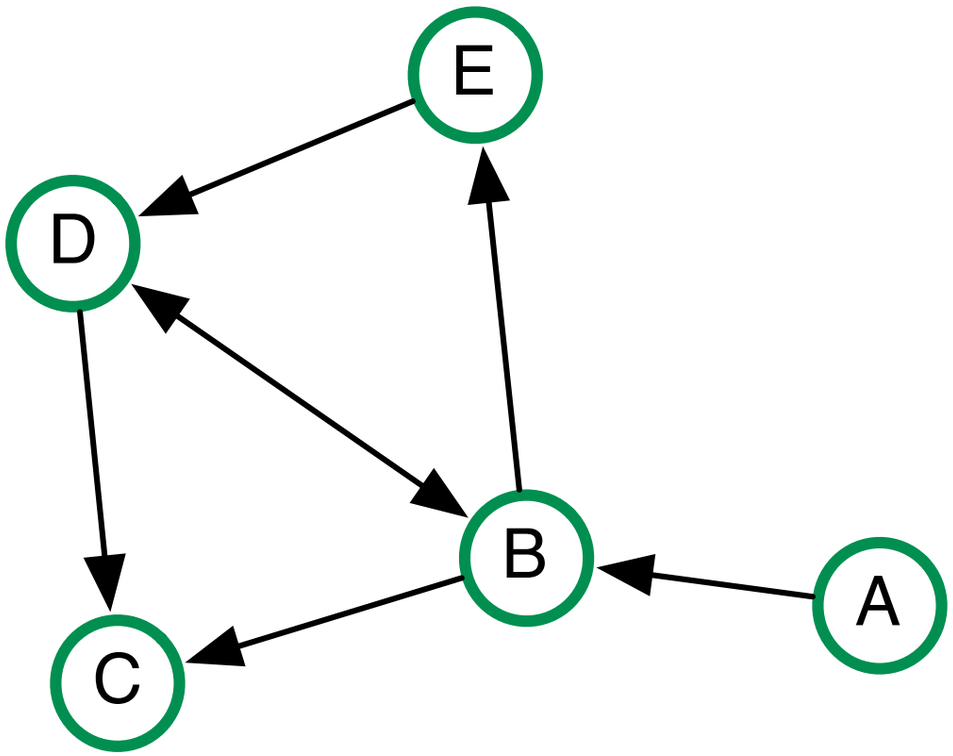
Directed “follower network” for Twitter users



Nodes scaled and coloured based on in-degree centrality

Degree Centrality

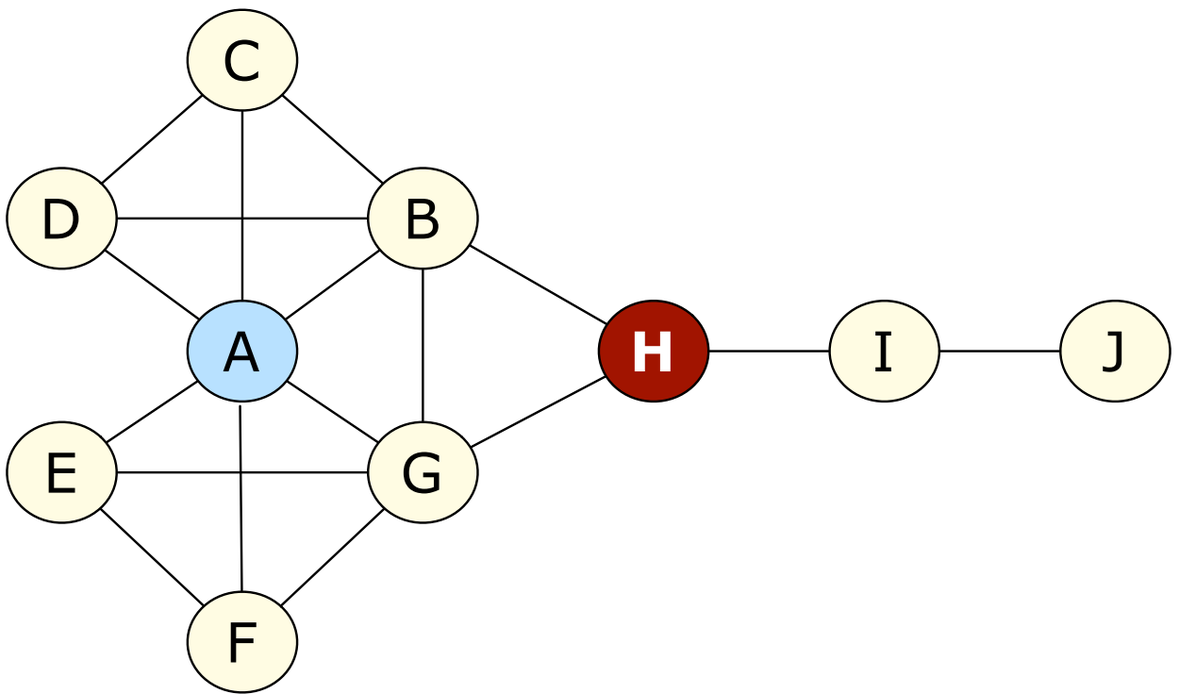
- For directed networks, in-degree centrality is often used as a proxy for popularity - i.e. number of incoming edges for a node.
- For certain directed networks, we might also look at out-degree centrality - i.e. number of out-going edges for a node.



Node	In-Degree	Out-Degree
A	0	1
B	2	3
C	2	0
D	2	2
E	1	1

Betweenness Centrality

- Identifies strategic linkages, “brokers” or “bridging nodes” in a graph.
- Nodes that occur on many shortest paths between other nodes in the graph have a high betweenness centrality score.
- Nodes with a high degree of betweenness centrality can be thought of as gatekeepers that control the flow of information between different parts of the graph.

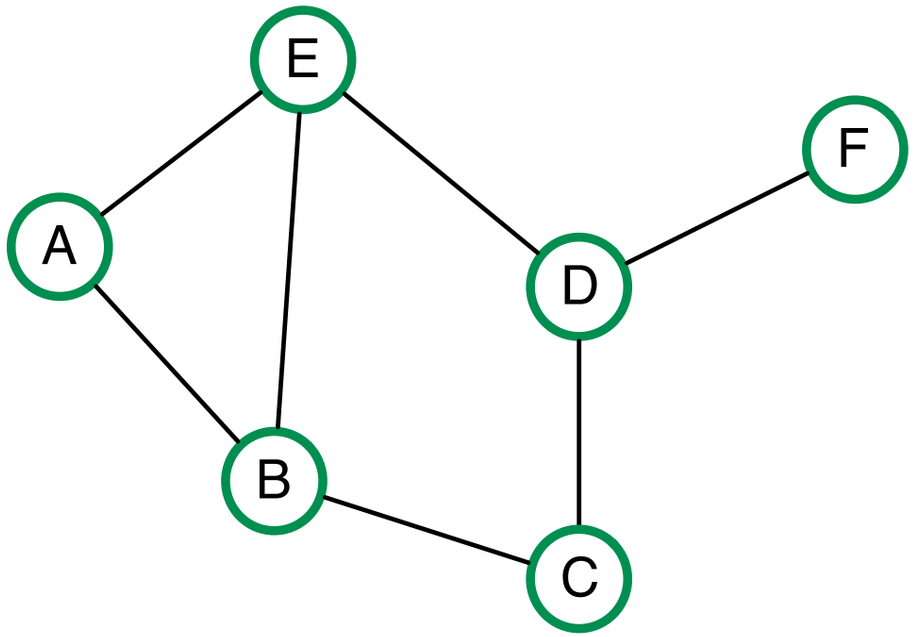


Node "A" has higher degree centrality than "B", as "B" has few direct connections.

Node "H" has higher betweenness centrality, as "H" plays a broker role in the graph.

Closeness Centrality

- Closeness centrality measures the extent to which a node is close to all other nodes in a network, either directly or indirectly.
- Reflects the node's ability to access information through the "grapevine" of network members.
- Calculated as inverse of the distance from a node to all other nodes.



Distances from node "A" to others:

D(A, B) = 1
D(A, C) = 2
D(A, D) = 2
D(A, E) = 1
D(A, F) = 3

Mean distance = (1+2+2+1+3)/5 = 1.8
Closeness(A) = 1/1.8 = 0.56

- Values are typically normalised to the range [0,1] for comparison across different networks.

Eigenvector Centrality

- The eigenvector centrality of a node proportional to the sum of the centrality scores of its neighbours.

➡ A node is important if it connected to other important nodes.

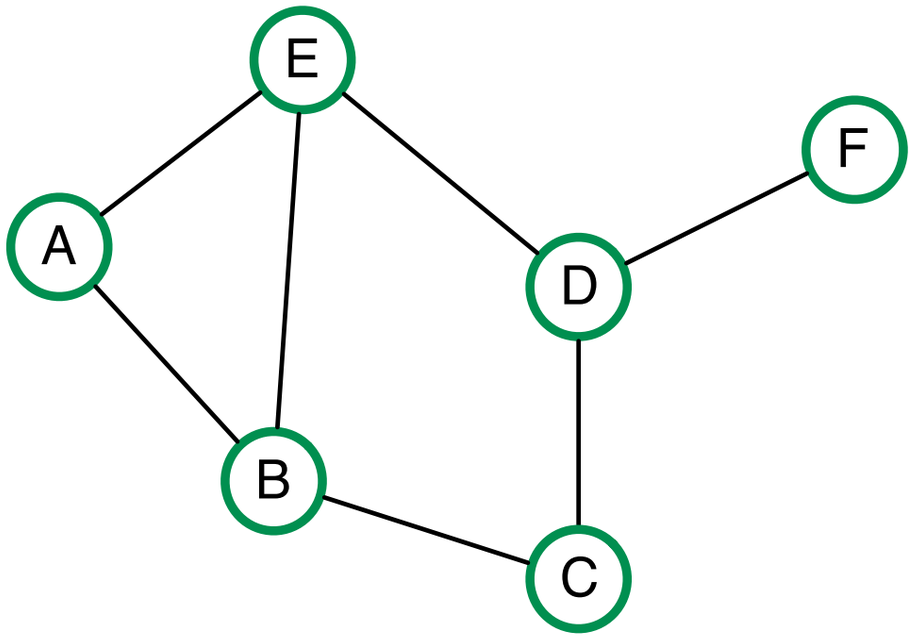
➡ A node with a small number of influential contacts may outrank one with a larger number of mediocre contacts.

- Computation:
  1. Calculate the eigendecomposition of the pairwise adjacency matrix of the graph.
  2. Select the eigenvector associated with largest eigenvalue.
  3. Element *i* in the eigenvector gives the centrality of the *i*-th node.

- Values are typically normalised to the range [0,1] for comparison across different networks.

Centrality Measures

- Depending on the semantics of the network and the task involved, different centrality measures will be appropriate.

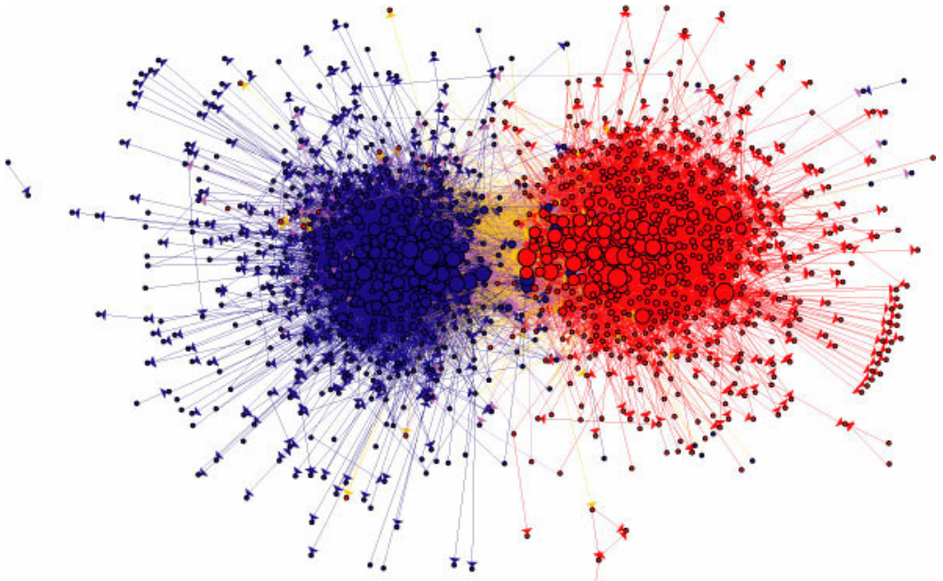


Node	Degree	Between	Close	Eigen
A	0.40	0.00	0.56	0.40
B	0.60	0.15	0.62	0.50
C	0.40	0.10	0.62	0.36
D	0.60	0.45	0.71	0.41
E	0.60	0.30	0.71	0.52
F	0.20	0.00	0.45	0.16

- Degree centrality: when number of connections is important
- Betweenness centrality: when control over transmission is important
- Closeness centrality: when time taken to reach nodes is important
- Eigenvector centrality: when importance of neighbours is important

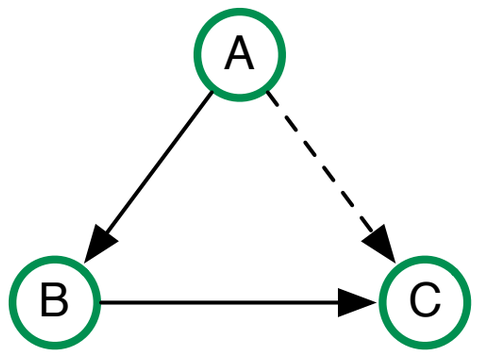
Homophily

- Homophily: tendency to relate to people with similar characteristics (e.g. demographics, tastes, opinions, etc.)
- Leads to homogeneous clusters where forming relations is easier. Extreme homogenisation can act counter to innovation and idea generation - the “filter bubble”.



(Adamic & Glance, 2005)

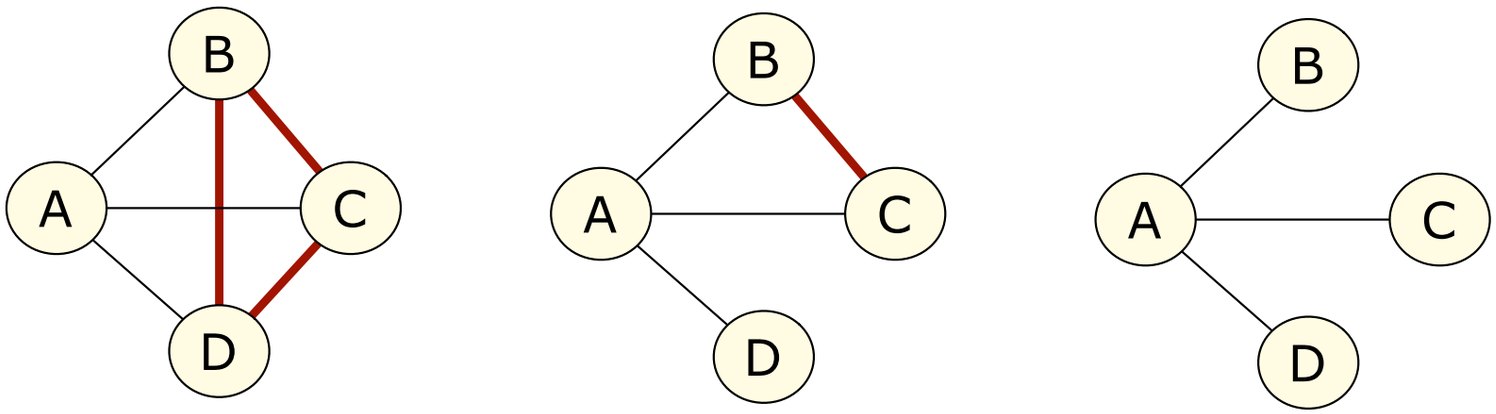
- Transitivity: If there is a tie between A and B, and a tie between B and C, then in a transitive network A and C are also likely to be connected.
- “If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future” - Granovetter
- Transitivity and homophily together lead to the formation of cliques.



Transitivity in triads in a directed network

Clustering Coefficient

- Measures have been proposed to capture prevalence of triadic closure.
- The neighbourhood of a node is set of nodes connected to it by an edge, not including itself.
- The clustering coefficient of a node is the fraction of pairs of its neighbours that have edges between one another.

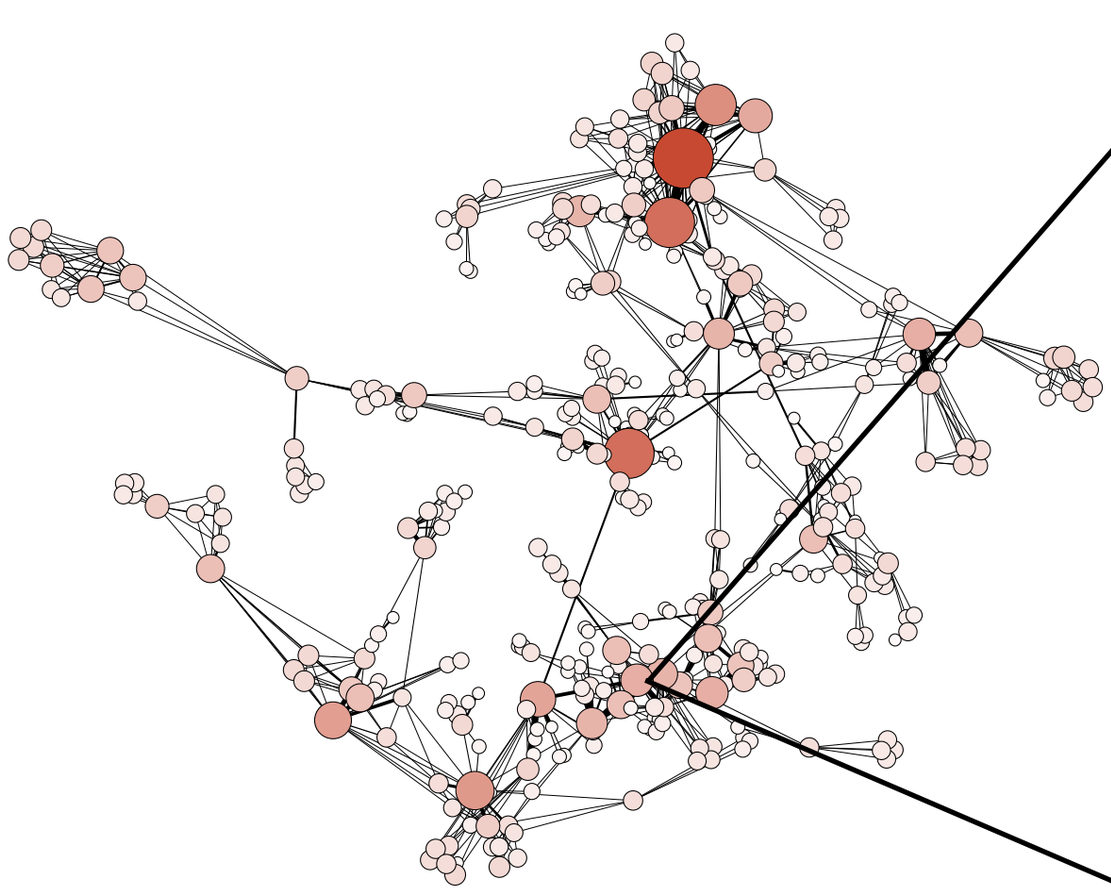


Node A: $CC = \frac{3}{3}$ $CC = \frac{1}{3}$ $CC = \frac{0}{3}$

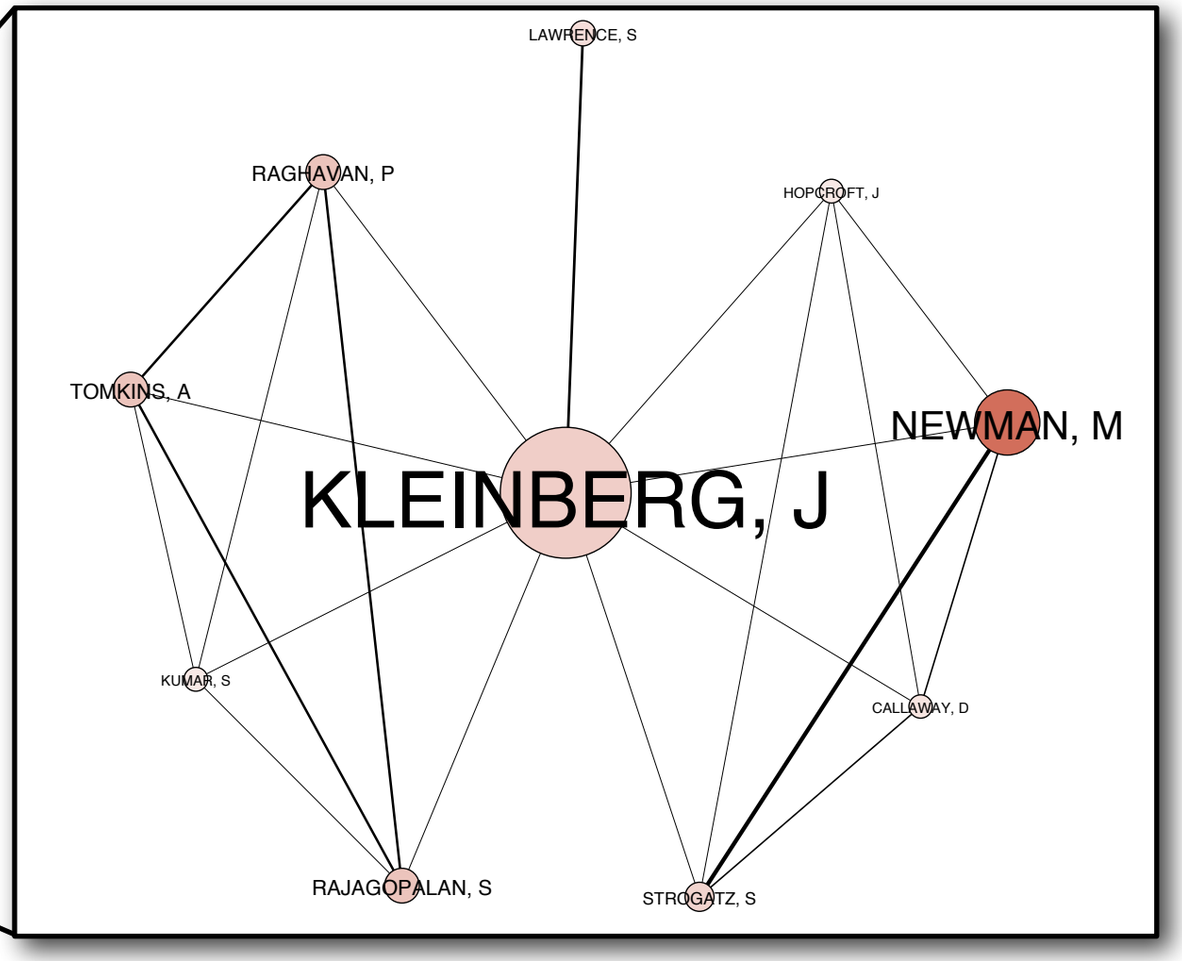
- Locally indicates how concentrated the neighbourhood of a node is, globally indicates level of clustering in a graph.
- Global score is average over all nodes: $\bar{CC} = \frac{1}{n}\sum_{i=1}^{n} CC(v_i)$

Ego Networks

- Ego-centric methods really focus on the individual, rather than on network as a whole.
- By collecting information on the connections among the modes connected to a focal ego, we can build a picture of the local network of the individual.



(Newman,2006)



Summary

- Social Network Analysis
- Graphs - Basic Concepts
- Characterising Graphs
  - Density
  - Connectivity
- Measures of Centrality
  - In-Degree, Out-Degree, Betweenness, Closeness, Eigenvector centrality
- Homophily, Clustering Coefficient
- Ego Networks