

# Social Media Analysis

**Ashish Kumar Jha**

# Agenda

**Graph Theory**

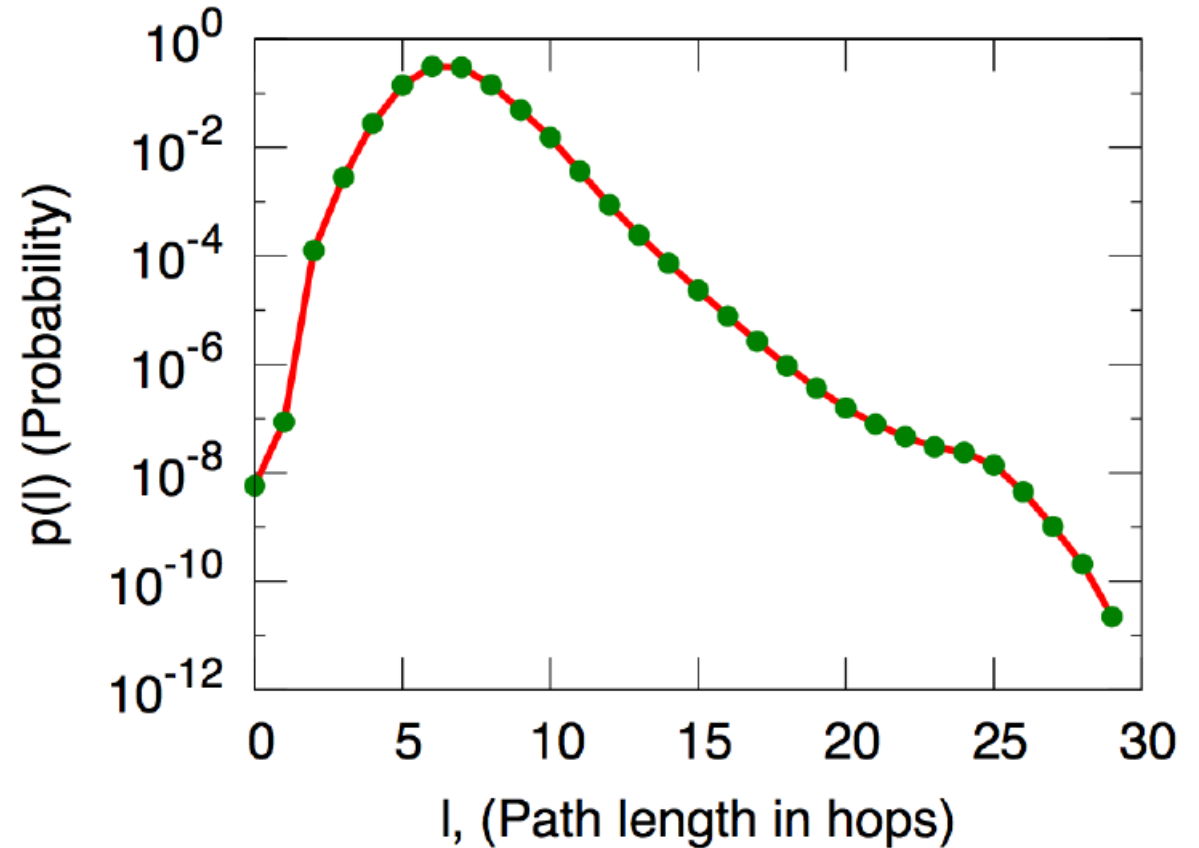
**Matrix Algebra**

**Nodes and networks**

**Network centrality**

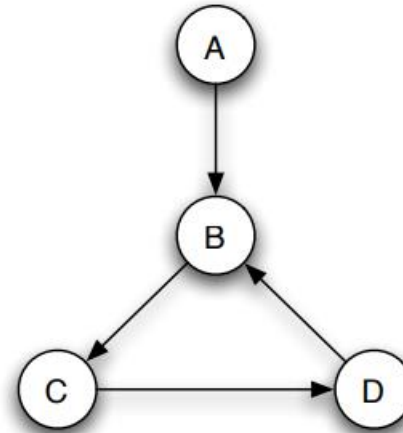
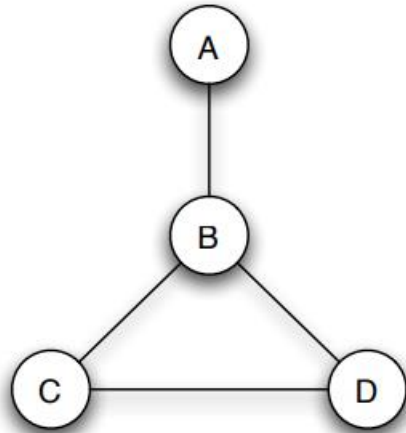
# Small World Phenomenon

- Social networks tend to have very short paths between essentially arbitrary pairs of people
- The world looks “small” when you think of how short a path of friends it takes to get from you to almost anyone else
- Erdos Number

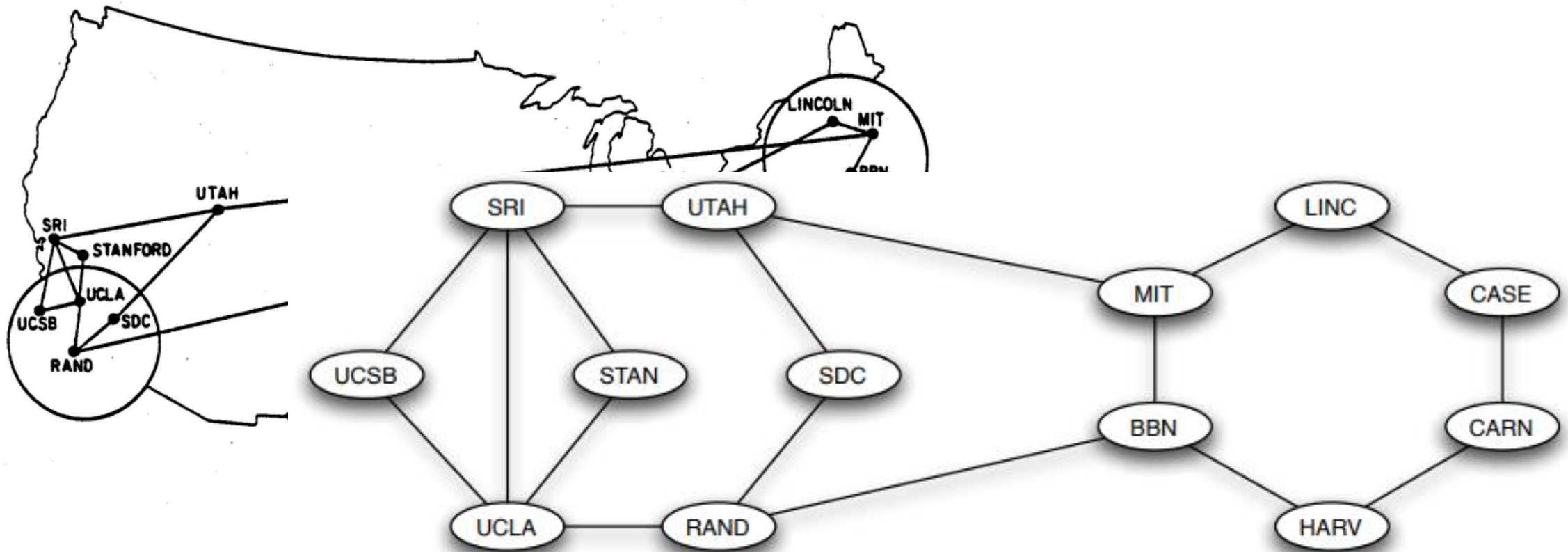


# Graphs

- A graph is a way of specifying relationships among a collection of items.
- A graph consists of a set of objects, called nodes, with certain pairs of these objects connected by links called edges



# Creating Graphs

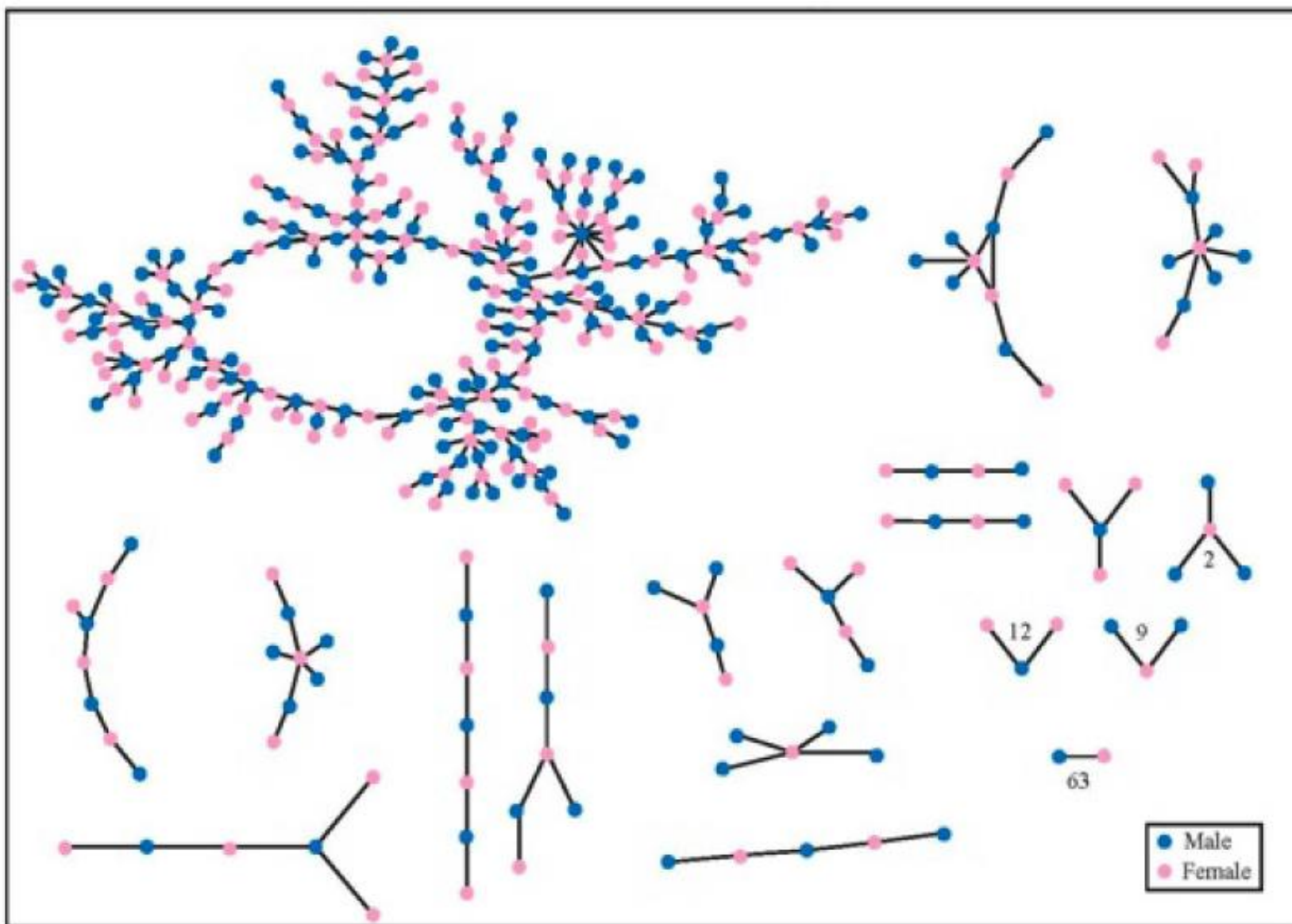


# Important Terms

- **Path :** A path is simply a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge
  - A simple path does not have repeat nodes
  - mit, bbn, rand, ucla is a simple path
- **Cycle:** A particularly important kind of non-simple path is a cycle, which informally is a “ring” structure
  - linc, case, carn, harv, bbn, mit, linc
- **Connectivity:** A graph is connected if for every pair of nodes, there is a path between them

# Important Terms

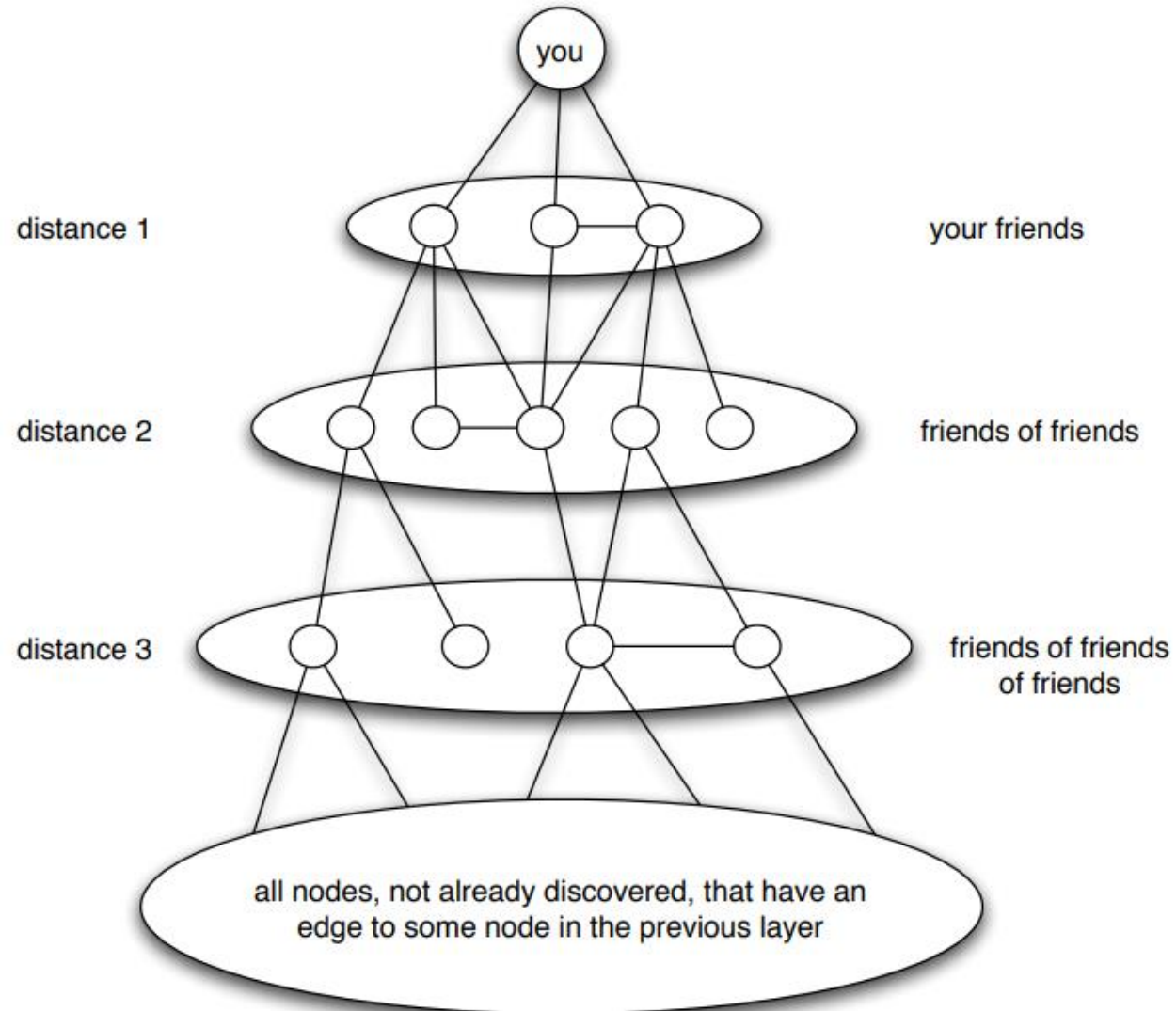
- **Connected component** (often shortened to “component”) is a subset of the nodes such that:
  - every node in the subset has a path to every other; and
  - the subset is not part of some larger set with the property that every node can reach every other.
- It really is a free-standing “piece” of the graph
- **Giant component:** A deliberately informal term for a connected component that contains a significant fraction of all the nodes





# Distance in a graph

- **Breadth first  
Search**



# Graph Concepts

- **Homophily:** the principle that we tend to be similar to our friends
- Social network on a set of people, in which everyone knows everyone else - so we have an edge joining each pair of nodes. Such a network is **called a clique, or a complete graph**

# Isolates and components

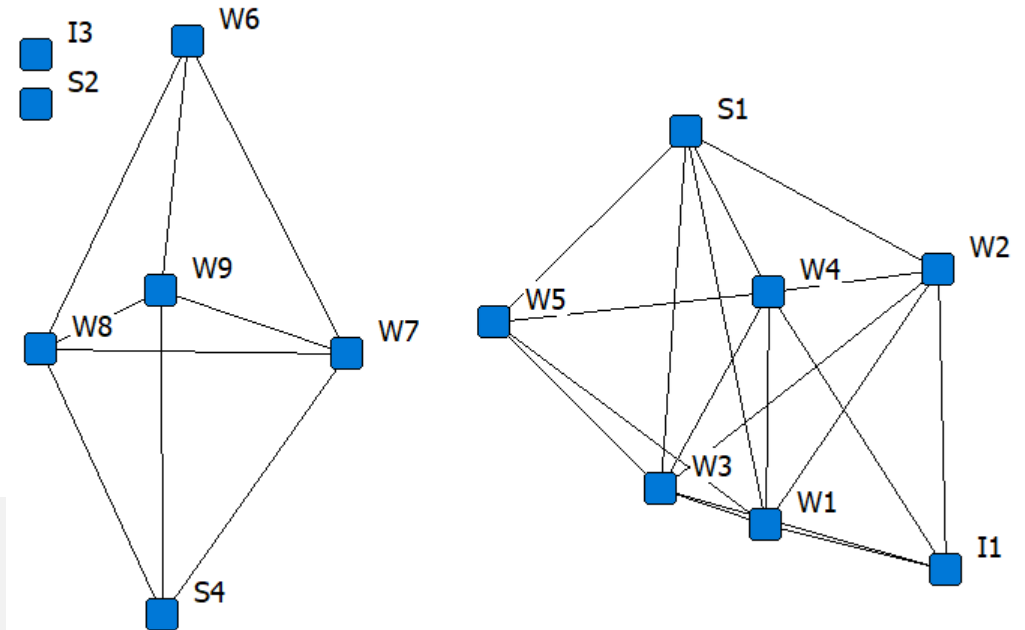
**A component is a fragment of a network with no ties to any other component**

- A better definition is forthcoming ...

**An isolate is a node with no ties**

- Every isolate is a component

A network with 4 components, including two isolates



# Graph Notations

- Graph is represented as  $G(V,E)$  where  $V$  and  $E$  are vectors of vertices and Edges respectively
- $V$  is represented as a set of vertices  $V=\{A,B,C,D,E\}$
- $E$  is represented as a set of ordered pairs such as  $E= \{(A,B), (A,D), (B,D), (C,D), (D,E)\}$
- Presence of dual edges are sometimes an indirect reference to the graph being undirected
- $(u,v) \in E(G)$  also known as  $uGv$

# Adjacency Matrix for Graphs

		$\mathcal{M}$			
		Mr. Jones	Ms. Smith	Ms. Davis	Mr. White
$\mathcal{N}$	Allison	1	0	0	0
	Drew	0	1	0	0
	Eliot	0	0	1	0
	Keith	0	0	0	1
	Ross	0	0	1	0
	Sarah	0	1	0	0

the receiving actor set. We will denote the sociomatrices by using their sending and receiving actor sets, so, for example, the sociomatrix  $\mathbf{X}^{\mathcal{N}\mathcal{M}}$  contains measurements on a relation defined from actors in  $\mathcal{N}$  to actors in  $\mathcal{M}$ . These sociomatrices and their sizes are:

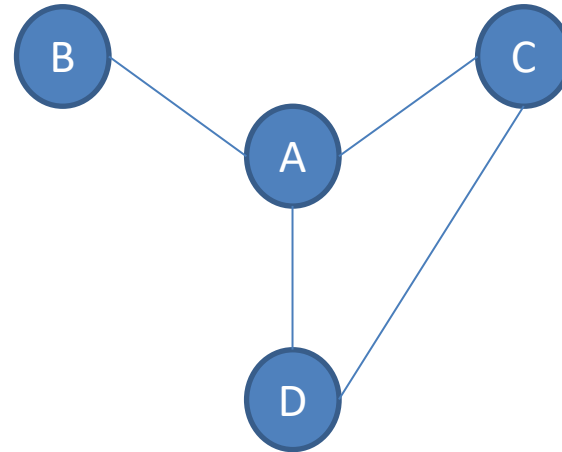
- $\mathbf{X}_r^{\mathcal{N}}$ , dimensions =  $g \times g$
- $\mathbf{X}_r^{\mathcal{M}}$ , dimensions =  $h \times h$
- $\mathbf{X}_r^{\mathcal{N}\mathcal{M}}$ , dimensions =  $g \times h$
- $\mathbf{X}_r^{\mathcal{M}\mathcal{N}}$ , dimensions =  $h \times g$

	a	b	c	d	e	f
a	0	1	0	0	0	0
b	1	0	1	1	0	0
c	0	1	0	1	0	0
d	0	1	1	0	1	0
e	0	0	0	1	0	1
f	0	0	0	0	1	0

# Graph Notations

	A	B	C	D
A	0	1	1	1
B	1	0	0	0
C	1	0	0	1
D	1	0	1	0
Adjacency Matrix				

	A	B	C	D
A	0	1	1	1
B	1	0	2	2
C	1	2	0	1
D	1	2	1	0
Distance Matrix				

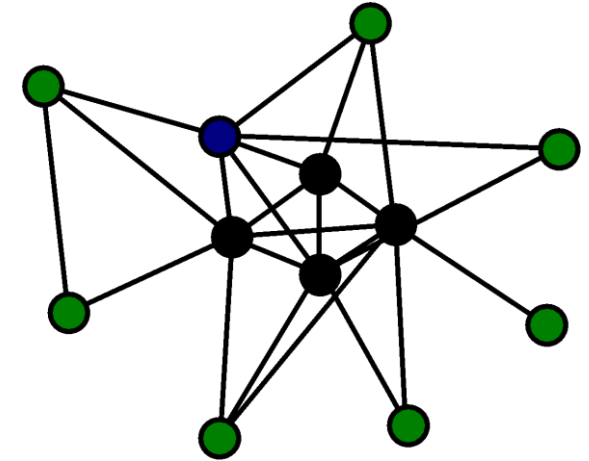


# Directed graphs (digraphs)

## Graphs can be directed or undirected

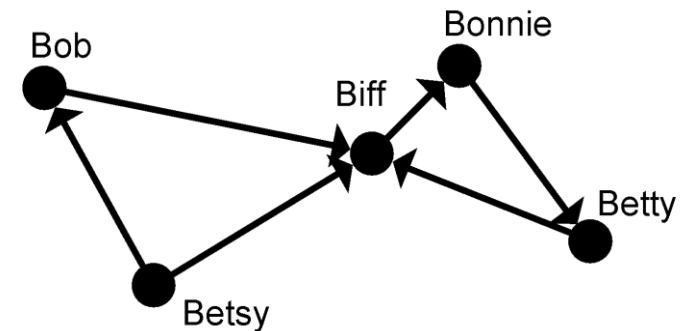
### Undirected

- Edges don't have direction
- Suppose  $G(V, M)$  indicates marriage ties
  - If  $xMy$  then  $yMx$



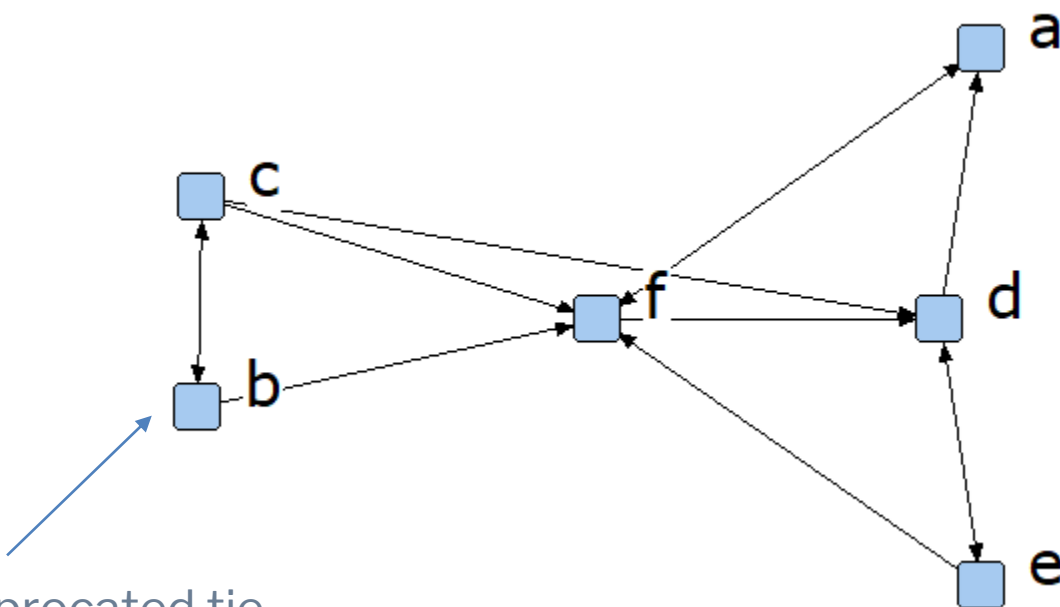
### Directed

- Ties ('arcs') have direction
- Suppose  $G(V, A)$  indicates advice-seeking ties
- Then  $uAv$  and  $vAu$  mean different things
  - $uAv$  means that  $u$  seeks advice from  $v$
  - $vAu$  means that  $v$  seeks advice from  $u$



# Adjacency matrix of directed graph

Adjacency matrix is (usually) not symmetric



Reciprocated tie –  
technically 2 ties, one in  
each direction

	a	b	c	d	e	f
a	0	0	0	0	0	1
b	0	0	1	0	0	1
c	0	1	0	1	0	1
d	1	0	0	0	1	0
e	0	0	0	1	0	1
f	1	0	0	1	0	0

Consider “likes” and “seeks advice  
from”



# Matrix Algebra

- **Matrix elements.** Consider the matrix below, in which matrix elements are represented entirely by symbols.

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \end{bmatrix}$$

By convention, first subscript refers to the row number; and the second subscript, to the column number.

Thus, the first element in the first row is represented by  $A_{11}$ . The second element in the first row is represented by  $A_{12}$ . And so on, until we reach the fourth element in the second row, which is represented by  $A_{24}$ .

- **Matrices.** There are several ways to represent a matrix symbolically. The simplest is to use a boldface letter, such as **A**, **B**, or **C**. Thus, **A** might represent a 2 x 4 matrix, as illustrated below.

$$\mathbf{A} = \begin{bmatrix} 11 & 62 & 33 & 93 \\ 44 & 95 & 66 & 13 \end{bmatrix}$$

Another approach for representing matrix **A** is:

$$\mathbf{A} = [A_{ij}] \text{ where } i = 1, 2 \text{ and } j = 1, 2, 3, 4$$

# Matrix Algebra

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 6 & 7 \\ 8 & 9 \\ 10 & 11 \end{bmatrix}$$

Let  $\mathbf{AB} = \mathbf{C}$ . Because  $\mathbf{A}$  has 2 rows, we know that  $\mathbf{C}$  will have two rows; and because  $\mathbf{B}$  has 2 columns, we know that  $\mathbf{C}$  will have 2 columns. To compute the value of every element in the 2 x 2 matrix  $\mathbf{C}$ , we use the formula  $C_{ik} = \sum_j A_{ij}B_{jk}$ , as shown below.

- $C_{11} = \sum A_{1j}B_{j1} = 0*6 + 1*8 + 2*10 = 0 + 8 + 20 = 28$
- $C_{12} = \sum A_{1j}B_{j2} = 0*7 + 1*9 + 2*11 = 0 + 9 + 22 = 31$
- $C_{21} = \sum A_{2j}B_{j1} = 3*6 + 4*8 + 5*10 = 18 + 32 + 50 = 100$
- $C_{22} = \sum A_{2j}B_{j2} = 3*7 + 4*9 + 5*11 = 21 + 36 + 55 = 112$

Based on the above calculations, we can say

$$\mathbf{AB} = \mathbf{C} = \begin{bmatrix} 28 & 31 \\ 100 & 112 \end{bmatrix}$$

# Matrix Multiplication Application

- Used to create compound graphs and social networks
- $F \rightarrow$  Adjacency matrix representing “Friend of” relation
- $E \rightarrow$  Adjacency matrix representing “Enemy of” relation
- Then  $C=FE$  represents “Enemy of Friend of” relation
- $FE(i,j)$  gives number of  $i$ ’s friends with  $j$  as enemy
- Remember
  - Matrix multiplication is non-commutative,  $FE$  is not equal to  $EF$
  - Self- multiplication is possible i.e.  $FF$  can exist
  - $F^k$  gives the number of paths of  $k$  length that start at row node and end at column node

# Matrix Multiplication Application

- **Self- multiplication is possible i.e. FF can exist**
- $F^k$  gives the number of paths of k length that start at row node and end at column node
- $F^3(i,j)=2$  implies presence of 2 paths of length 3 starting at i and ending at j
- E.g. i-k-m-j and i-k-i-j

# Matrix Multiplication Application

- Using  $E$  and  $F$  as enemy and friend matrices respectively, we get
  - $FF(i,j) > 0$  implies  $i$  has at least 1 friend who considers  $j$  as friend
  - Other relations
    - $F=FF$  Friend of my friend is my friend
    - $E=FE$  Friend of my enemy is my enemy
    - $E=EF$  Enemy of my friend is my enemy
    - $F=EE$  Enemy of my enemy is my friend

# Transposition

->A = copy(katz)  
->A<sup>t</sup> =  
transpose(A)

**Suppose F is the “father of” relation.**

- uFv means that u is the father of v

**We use F’ to indicate the converse of F, which is the reciprocal role: “child of”**

- If uFv, then vF’u -- v is the child of u

**If we represent a relation as an adjacency matrix, the converse is represented by the transpose of the matrix**

- Get the transpose by writing each row as a column, or vice-versa

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

matrix A  
“x advises  
y”

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

matrix A’  
“x is advised by  
y”



**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

Trinity Business School

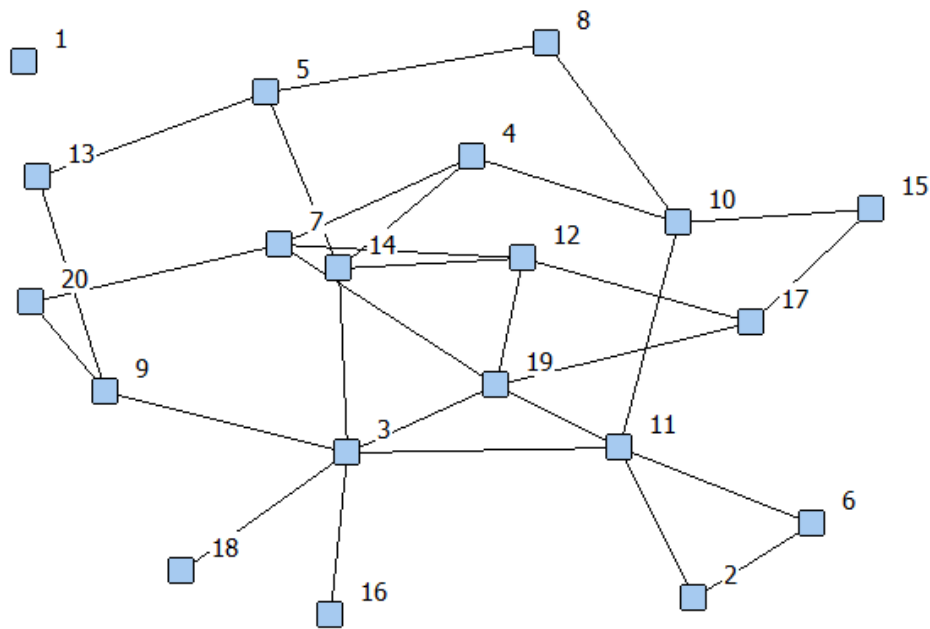
# Centrality Measures



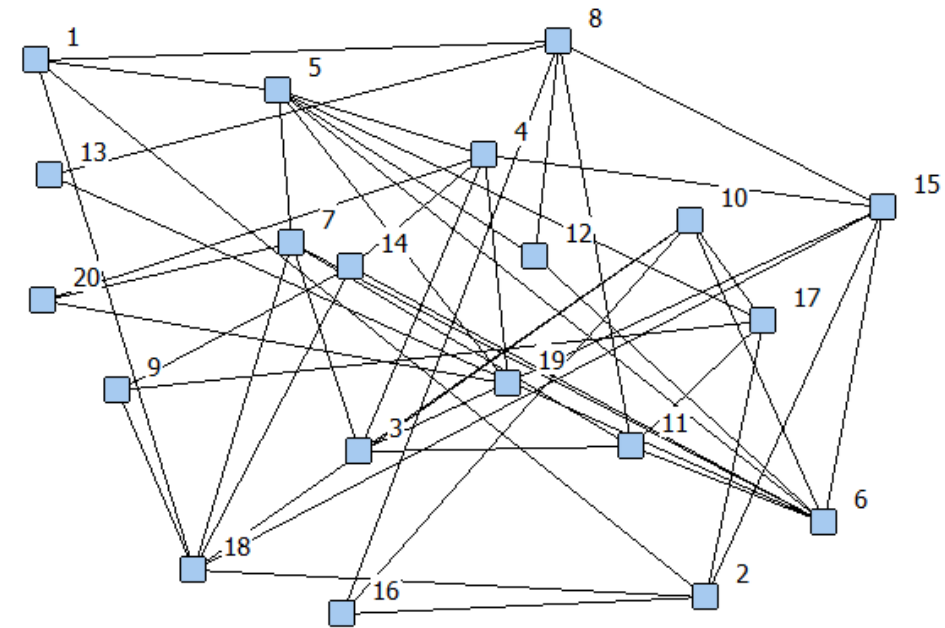


# Density

## Number of ties, expressed as proportion of # possible



Density =  
0.15



Density =  
0.25



# Density - formulas

**Density = # of ties / # possible**

**Equivalently, density = average of the adjacency matrix**

	Reflexive	Non-Reflexive
Undirected	$= \frac{T}{n^2 / 2}$	$= \frac{T}{n(n-1) / 2}$
Directed	$= \frac{T}{n^2}$	$= \frac{T}{n(n-1)}$

T = number of ties in network  
n = number of nodes

**Density is also the prob that a randomly chosen pair of nodes has a tie**

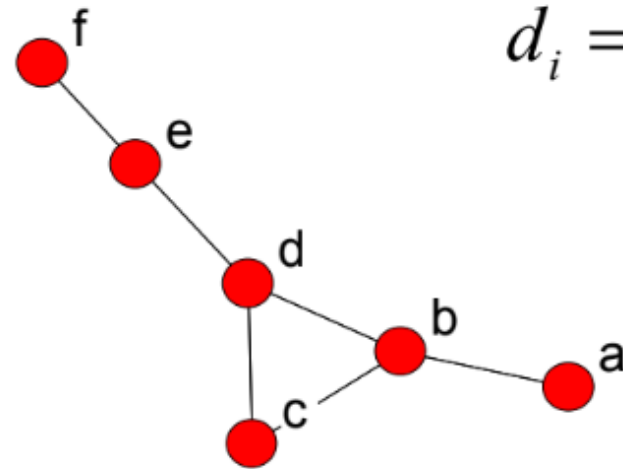
# Centrality

- **Measure of a node's position in a network**
- **4 major centrality**
  - Degree
  - Betweenness
  - Closeness
  - Eigenvalue
- **Other measures also exist, but are not used as frequently**

# Degree Centrality

- In undirected graph- easy to compute
  - No of nodes a node is connected to/ no. of edges a node has
  - Direct measure of influence

	a	b	c	d	e	f	Sum	Avg
a	0	1	0	0	0	0	1	.2
b	1	0	1	1	0	0	3	.6
c	0	1	0	1	0	0	2	.4
d	0	1	1	0	1	0	3	.6
e	0	0	0	1	0	1	2	.4
f	0	0	0	0	1	0	1	.2



$$d_i = \sum_j a_{ij}$$

# Average Degree

**Average number of links per person**

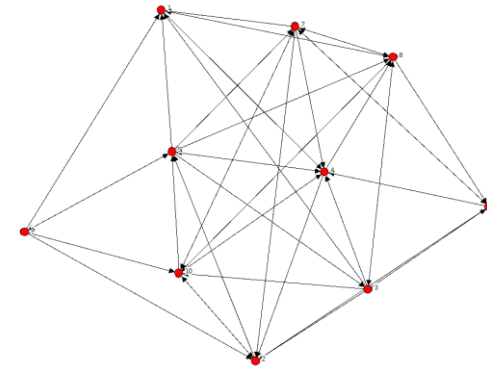
**Is same as  $\text{density} \times (n-1)$ , where  $n$  is size of network**

- Density is just normalized avg degree – divide by max possible

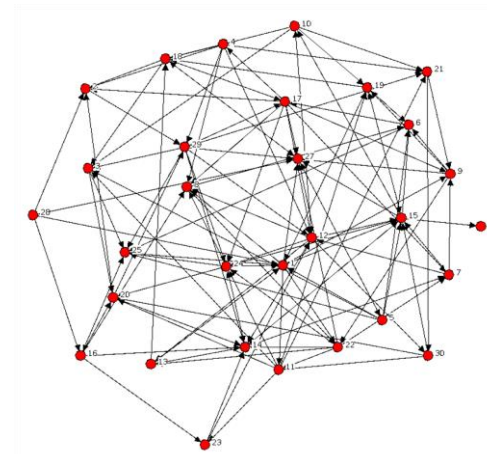
**Often more intuitive than density**

**But density has one very nice feature:**

- It is the probability that a randomly chosen pair of nodes have a tie



Density  
0.47  
Avg Deg 4



Density  
0.14  
Avg Deg 4

# Turbo-charging degree- Iterated degree

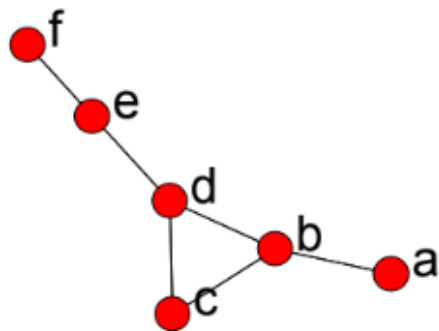
- Degree is a count of the number of nodes you are connected to
  - Treats all nodes equally
- What if you wanted to weight the nodes by how many nodes they were connected to?

$$t_i = \sum_j a_{ij} d_j$$

- But why stop there? Can keep iterating ...

## Iterated Degree

	a	b	c	d	e	f	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
a	0	1	0	0	0	0	1	3	6	16	35	86	195	465	1071	2524
b	1	0	1	1	0	0	3	6	16	35	86	195	465	1071	2524	5854
c	0	1	0	1	0	0	2	6	13	32	73	173	401	940	2190	5117
d	0	1	1	0	1	0	3	7	16	38	87	206	475	1119	2593	6086
e	0	0	0	1	0	1	2	4	9	20	47	107	253	582	1372	3175
f	0	0	0	0	1	0	1	2	4	9	20	47	107	253	582	1372
							12	28	64	150	348	814	1896	4430	10332	24128



	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
A	8.3	10.7	9.4	10.7	10.1	10.6	10.3	10.5	10.4	10.5
B	25.0	21.4	25.0	23.3	24.7	24.0	24.5	24.2	24.4	24.3
C	16.7	21.4	20.3	21.3	21.0	21.3	21.1	21.2	21.2	21.2
D	25.0	25.0	25.0	25.3	25.0	25.3	25.1	25.3	25.1	25.2
E	16.7	14.3	14.1	13.3	13.5	13.1	13.3	13.1	13.3	13.2
F	8.3	7.1	6.3	6.0	5.7	5.8	5.6	5.7	5.6	5.7

# Degree Centrality

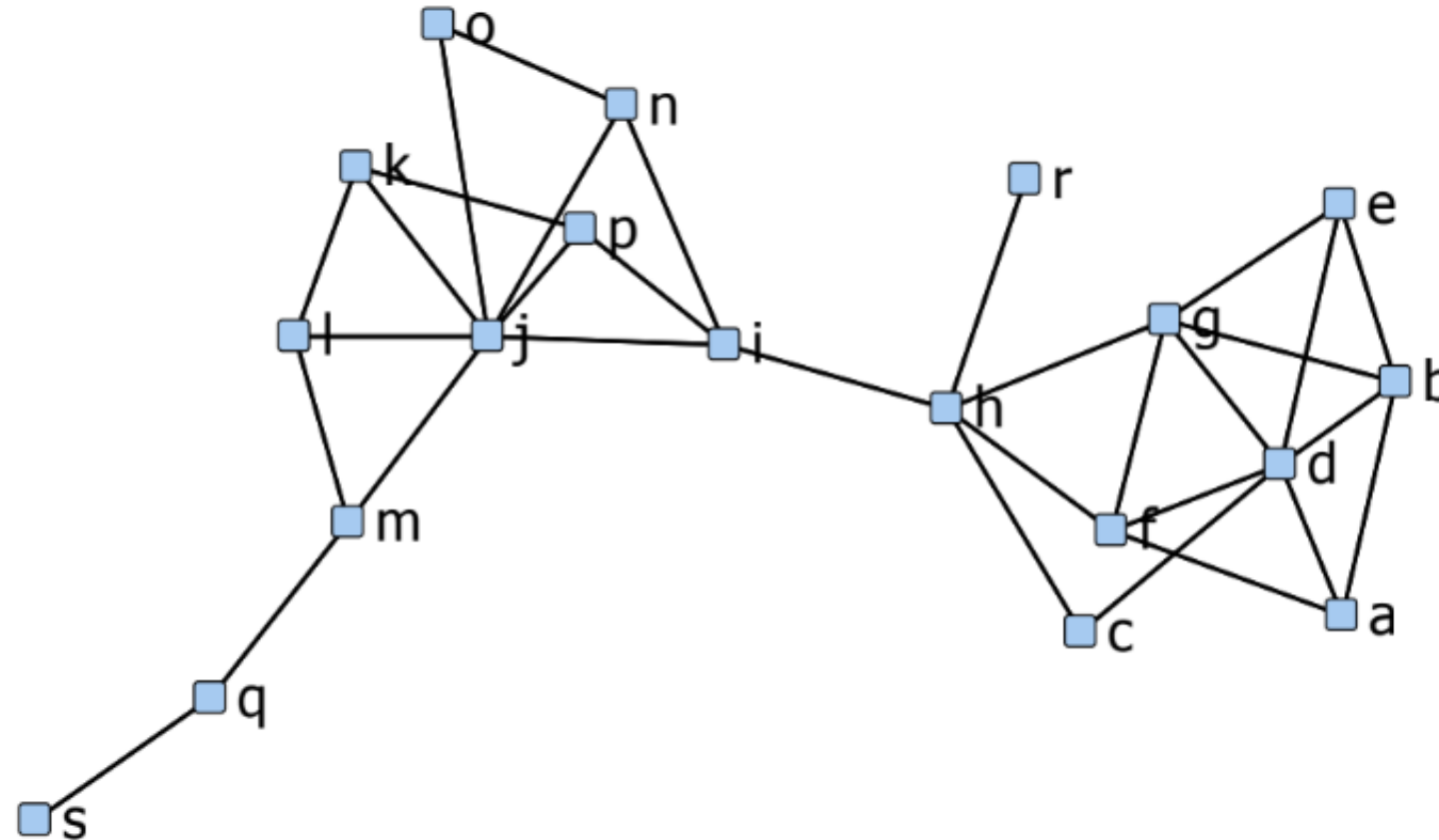
- **For directed graph**
- **In Degree- No. of incoming connections**
  - Used for popularity, prestige, influence etc.
  - How many people seek help from you
  - Is a sum of columns in adjacency matrix
- **Out Degree- N. of outgoing connections**
  - Example of that an individual can have or spread
  - Can be a measure of gregariousness
  - E.g. how many people you seek help from
  - Is a sum of rows in directed adjacency matrix

# Eigenvector

- Measure of the influence of a Node
- It is the principal eigenvector (with largest eigenvalue) of network adjacency matrix
- $Av = \lambda v \quad v_i = \frac{1}{\lambda} \sum_j a_{ij} v_j$
- $V$  is the eigenvector  $\lambda$  is the associated eigenvalue
- A node has high eigenvector score to the extent it is connected to many nodes who themselves have high scores
- Often interpreted as popularity or status – have ties not just to many others but many well-connected others



- Node **d** has the highest eigenvector centrality in the land



# Eigencentrality

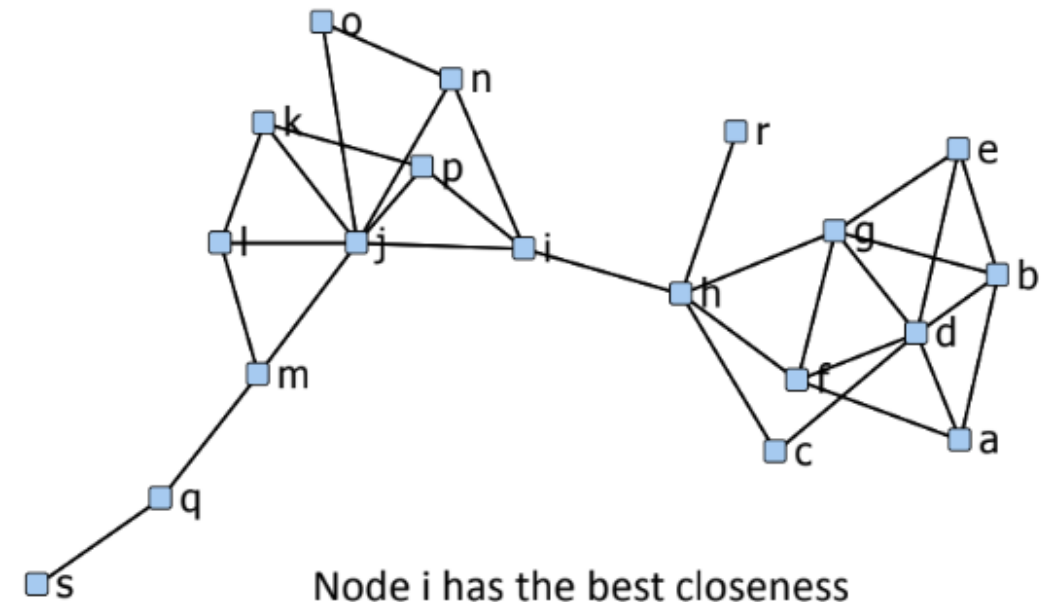
- Value does not exist for disconnected networks
- It is a measure of clique membership
- In clumpy networks, it favors the nodes in the larger cliques
  - Fails as a measure of risk/exposure because it doesn't take into account the fact that an alter's high degree might be because of ties with nodes that ego is already connected to

# Closeness Centrality

- Sum of distances from node to all others
- Often interpreted as index of time-until-arrival of stuff flowing through network
- In gossip network, persons strong in closeness centrality hear things early

# Closeness as marginals of distance matrix

ID	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	sum
a	0	1	2	1	2	1	2	2	3	4	5	5	5	4	5	4	6	3	7	62
b	1	0	2	1	1	2	1	2	3	4	5	5	5	4	5	4	6	3	7	61
c	2	2	0	1	2	2	2	1	2	3	4	4	4	3	4	3	5	2	6	52
d	1	1	1	0	1	1	1	2	3	4	5	5	5	4	5	4	6	3	7	59
e	2	1	2	1	0	2	1	2	3	4	5	5	5	4	5	4	6	3	7	62
f	1	2	2	1	2	0	1	1	2	3	4	4	4	3	4	3	5	2	6	50
g	2	1	2	1	1	1	0	1	2	3	4	4	4	3	4	3	5	2	6	49
h	2	2	1	2	2	1	1	0	1	2	3	3	3	2	3	2	4	1	5	40
i	3	3	2	3	3	2	2	1	0	1	2	2	2	1	2	1	3	2	4	39
j	4	4	3	4	4	3	3	2	1	0	1	1	1	1	1	1	2	3	3	42
k	5	5	4	5	5	4	4	3	2	1	0	1	2	2	2	1	3	4	4	57
l	5	5	4	5	5	4	4	3	2	1	1	0	1	2	2	2	2	4	3	55
m	5	5	4	5	5	4	4	3	2	1	2	1	0	2	2	2	1	4	2	54
n	4	4	3	4	4	3	3	2	1	1	2	2	2	0	1	2	3	3	4	48
o	5	5	4	5	5	4	4	3	2	1	2	2	2	1	0	2	3	4	4	58
p	4	4	3	4	4	3	3	2	1	1	1	2	2	2	2	0	3	3	4	48
q	6	6	5	6	6	5	5	4	3	2	3	2	1	3	3	3	0	5	1	69
r	3	3	2	3	3	2	2	1	2	3	4	4	4	3	4	3	5	0	6	57
s	7	7	6	7	7	6	6	5	4	3	4	3	2	4	4	4	1	6	0	86
sum	62	61	52	59	62	50	49	40	39	42	57	55	54	48	58	48	69	57	86	1048



Average distance would be more interpretable

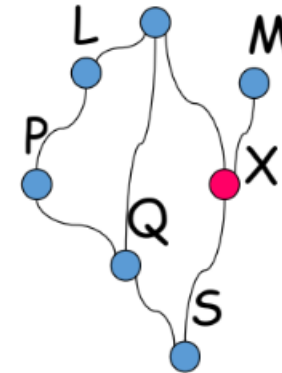
# Betweenness Centrality

- Measure of how centrally a node is located
- Often interpreted as control over flows (gatekeeping), correlated with power
- Also seen as index of frequency something reaches node

$$b_k = \sum_{i,j} \frac{g_{ikj}}{g_{ij}}$$

$g_{ij}$  is number of geodesic paths from  $i$  to  $j$

$g_{ikj}$  is number of geodesics from  $i$  to  $j$  that pass through  $k$



- More correctly,  $b_k$  is the share of geodesics between pairs of nodes that pass through  $k$

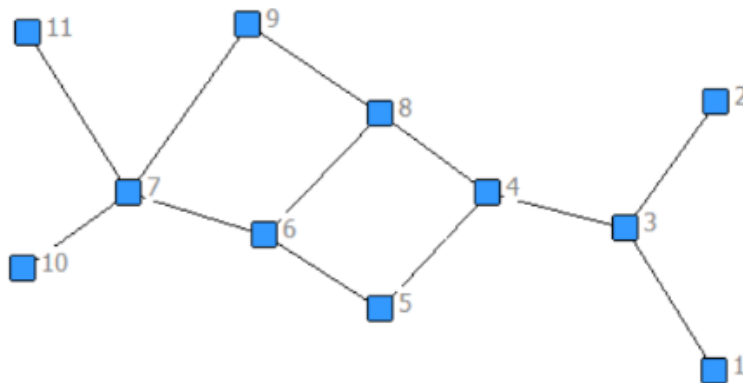
# Betweenness Features

- High degree is visible, high betweenness is not
- High betweenness nodes may be overlooked
- These nodes have high gatekeeping power- **Secretarial Power**
- **Cohesion of a network is dependent on these nodes\**
  - Networks with high betweenness are also very brittle

# Duality of closeness & betweenness

- Dependency matrix D, where  $d_{ij}$  = number of times\* that i needs to go through j to reach someone via a shortest path
- Column totals of D equal betweenness times 2
- Row totals of D equal closeness minus  $n-1$

	Closen	Between
1	36.000	0.000
2	36.000	0.000
3	27.000	17.000
4	22.000	21.833
5	23.000	6.000
6	22.000	13.667
7	25.000	17.833
8	21.000	15.167
9	24.000	5.500
10	34.000	0.000
11	34.000	0.000



	1	2	3	4	5	6	7	8	9	10	11	Clo
1		0.00	9.00	7.00	1.50	2.00	2.00	3.50	1.00	0.00	0.00	26.00
2	0.00		9.00	7.00	1.50	2.00	2.00	3.50	1.00	0.00	0.00	26.00
3	0.00	0.00		7.00	1.50	2.00	2.00	3.50	1.00	0.00	0.00	17.00
4	0.00	0.00	2.00		1.50	2.00	2.00	3.50	1.00	0.00	0.00	12.00
5	0.00	0.00	2.00	3.83		4.17	2.33	0.67	0.00	0.00	0.00	13.00
6	0.00	0.00	2.00	3.00	2.00		2.50	2.50	0.00	0.00	0.00	12.00
7	0.00	0.00	2.00	3.00	1.33	4.17		2.67	1.83	0.00	0.00	15.00
8	0.00	0.00	2.00	3.50	0.00	2.00	2.00		1.50	0.00	0.00	11.00
9	0.00	0.00	2.00	3.33	0.00	0.67	2.83	5.17		0.00	0.00	14.00
10	0.00	0.00	2.00	3.00	1.33	4.17	9.00	2.67	1.83		0.00	24.00
11	0.00	0.00	2.00	3.00	1.33	4.17	9.00	2.67	1.83	0.00		24.00

# Connectedness

**Proportion of pairs of nodes that can reach each other by some path, no matter how long**

$$C = \frac{\sum_{i \neq j} r_{ij}}{n(n-1)}$$

$r_{ij} = 1$  if node  $i$  can reach node  $j$  by a path of any length  
 $r_{ij} = 0$  otherwise

**Fragmentation = 1 – C**

**Connectedness is important for a group to coordinate actions, maintain a single culture**



# KeyPlayer application

## **Suppose you want to disrupt a network**

- E.g., stop epidemic by immunizing/quarantining an affordable # of people
- Disrupt terrorist group's ability to coordinate

## **You have the resources to isolate just 3 nodes. Which do you pick?**

- Suppose you pick the 3 nodes with the most ties

## **Two problems**

- Removing those nodes doesn't necessarily disconnect the network
- Picking an optimal set of  $k$  nodes is not the same thing as picking the  $k$  nodes that are individually most effective

## **Need combinatorial optimization algorithm to solve this**

# The Design Issue

**By standard off-the-shelf measures of node centrality, node 1 is the most important player, but deleting it ...**

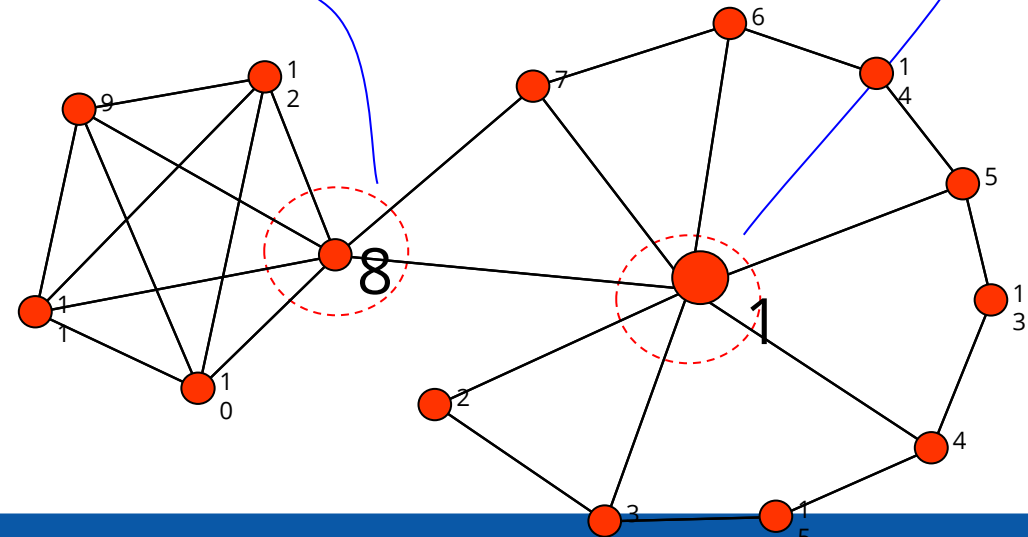
- does not disconnect the network

**In contrast, deleting node 8 breaks network into two components**

- Yet node 8 is not highest in centrality

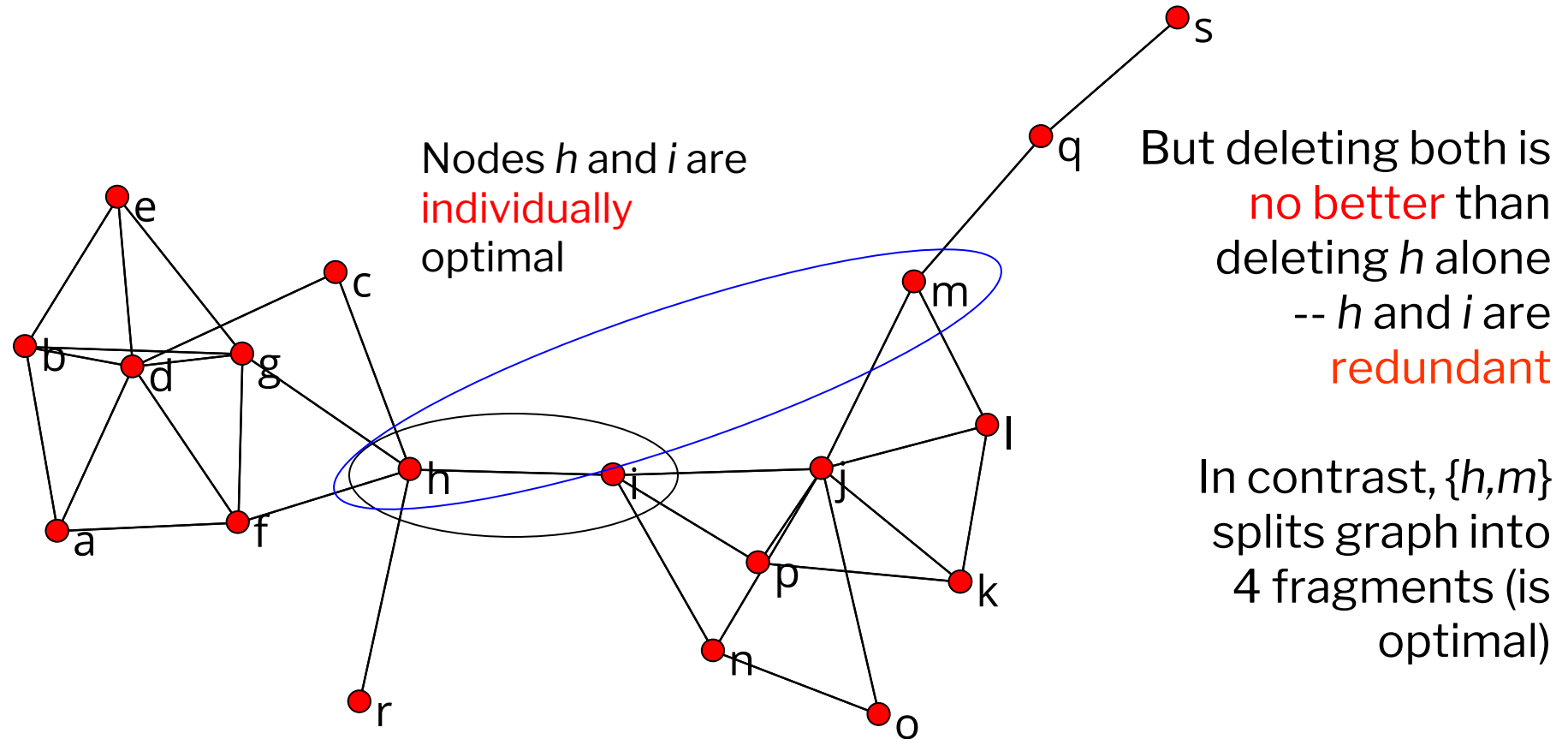
**Standard off-the-shelf centrality measures not optimal for the purpose of disrupting networks**

- Nor many other specific purposes



# The Ensemble Issue

Structural redundancy creates need for choosing complementary nodes

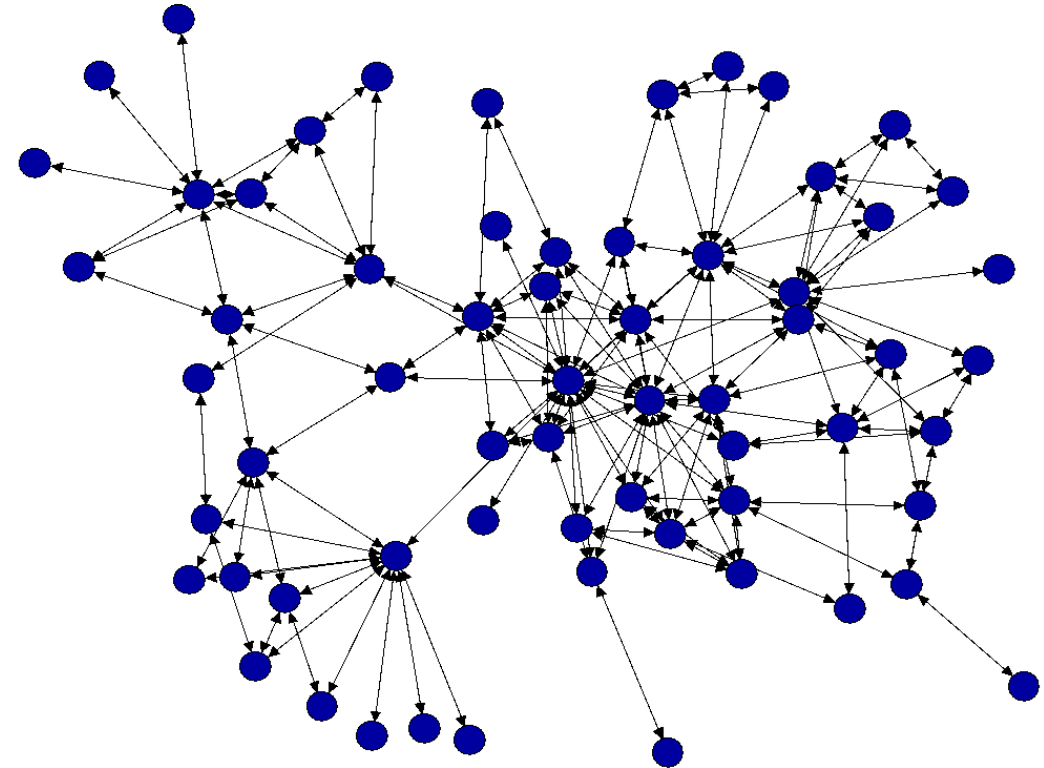


# Empirical Example #1

## Disrupt Terrorist Network<sup>k</sup>

DISRUPTION

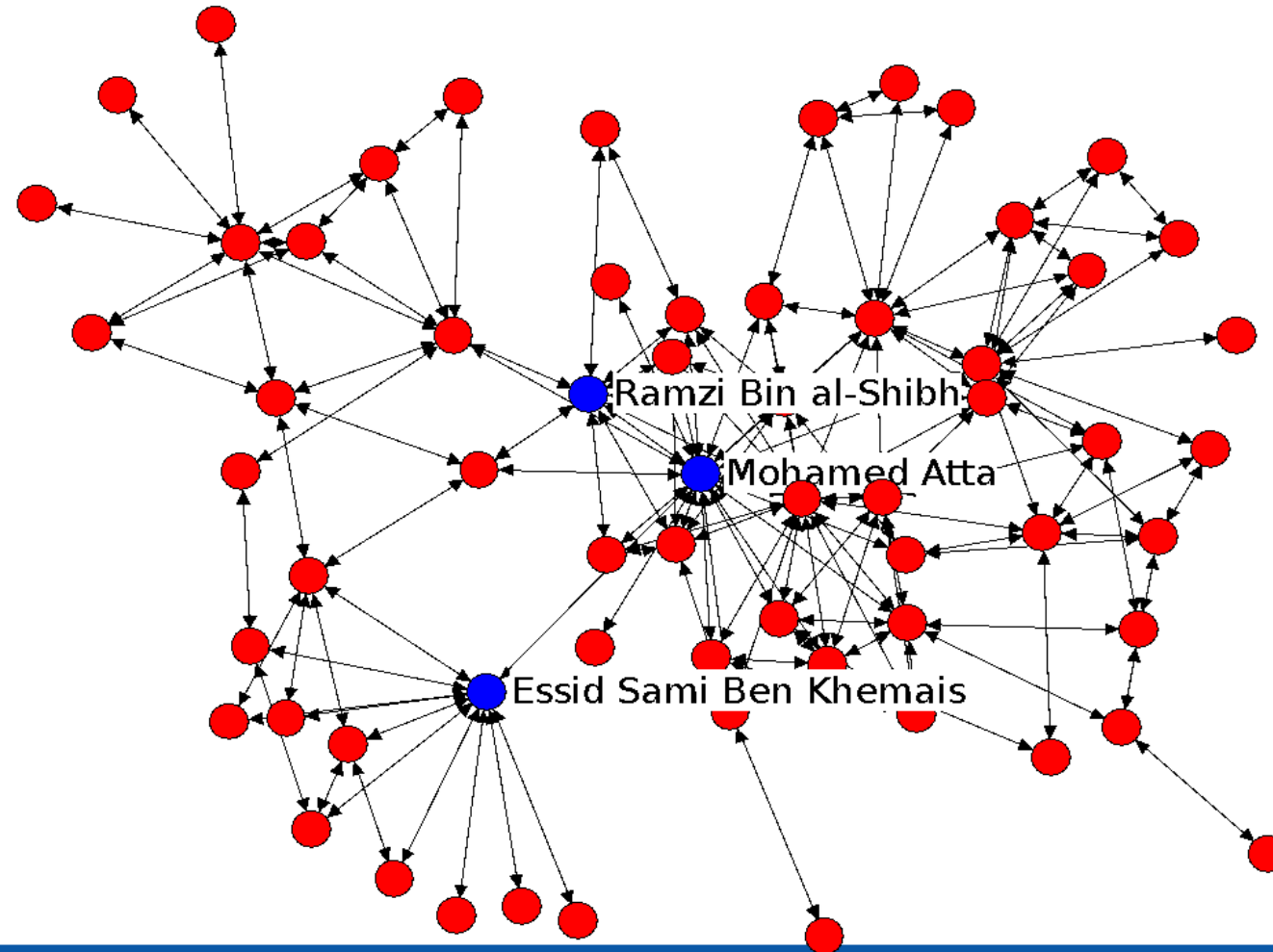
Which three nodes should be isolated in order to maximally disrupt the network?



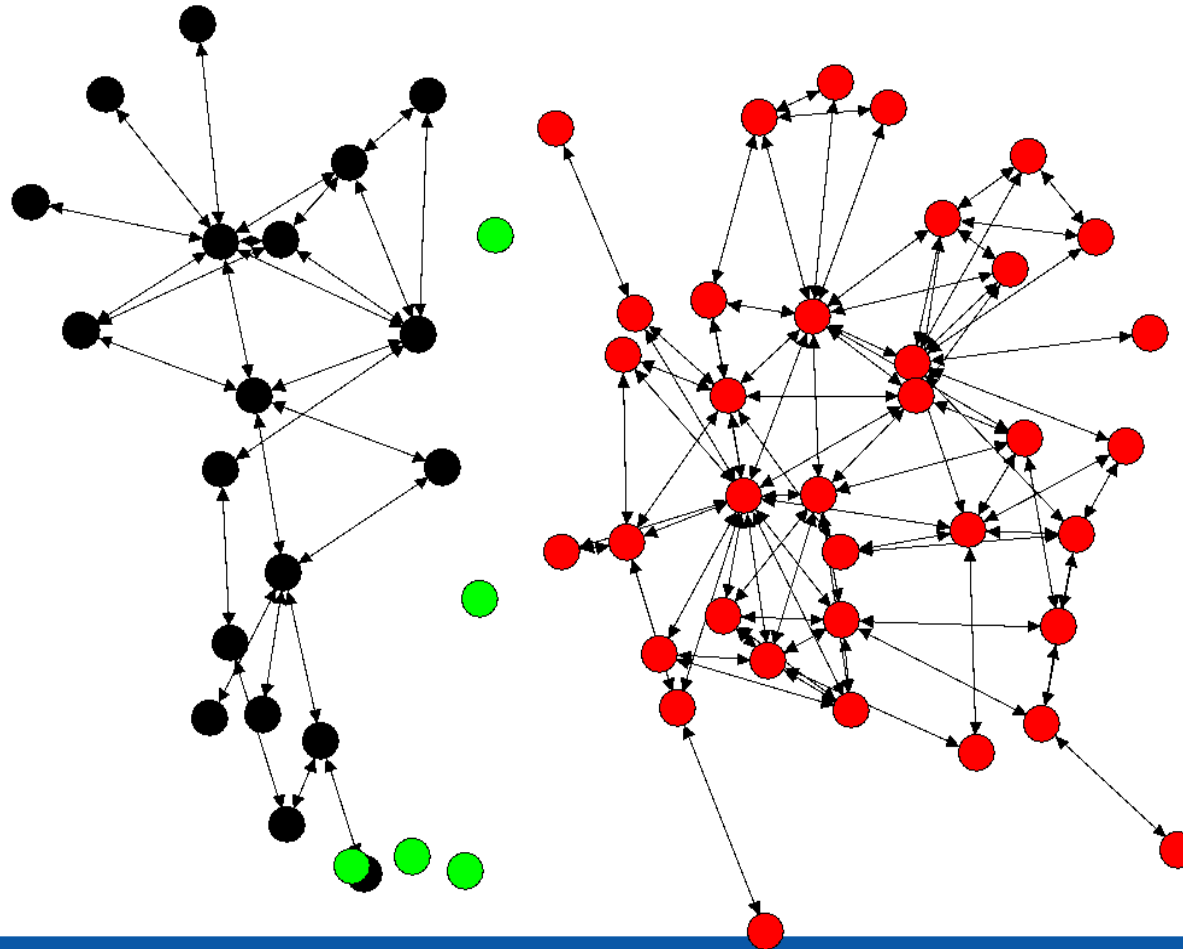
Data from: Krebs, V. 2002. Uncloaking terrorist networks.

First Monday 7(4): April. [http://www.firstmonday.dk/issues/issue7\\_4/krebs/index.html](http://www.firstmonday.dk/issues/issue7_4/krebs/index.html)

# KeyPlayer Solution



# KeyPlayer Solution (key players removed)



# Why do we want to know who the key players are?

<b>DISRUPT</b>	We want to <b>remove them</b> – to maximally <b>disrupt</b> the network
<b>ENHANCE</b>	We want to <b>help</b> them – in order to make network as a whole <b>function better</b> (diffuse info; coordinate well)
<b>INFLUENCE</b>	We want to identify <b>key opinion</b> leaders – to <b>influence</b> the network
<b>LEARN</b>	We want to know who <b>is in the know</b> – so we can question or <b>surveil</b> them
<b>REDIRECT</b>	We want to remove/prune them – to <b>redirect flows</b> in the network toward our preferred players





**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

Trinity Business School

# Network Data sources





# Research Design

- **Network data can be collected from experiment design or field study**
  - Causality vs. confoundedness
  - Experiments are truly causal but good experiment designs should have either pre-post or post only study
  - Field study do not allow same level of control on variables
  - Field study may also be longitudinal
- **Data can be secondary or primary**
  - Social media data is secondary

# Research Design

Paper	Study type	Design details	Independent Variable	Dependent Variable
Rand et al. 2011	Experiment	Links are created in 4 different groups a) randomly, b) fixed c) viscous d) fluid	Experimental network	Evolution of cooperation
Soyez et al., 2006	Quasi-experiment	Groups of substance abuse patients were taken and were given normal treatment and treatment plus network intervention	Social network intervention (ego)	Substance abuse treatment retention
Johnson et al. 2003	Prospective field study	Study of group dynamics in polar research station	Core-periphery structure	Individual morale
Burt 1995	Cross-sectional	Actor's position in a network effects bonus and evaluation	Individual social capital	Performance evaluation
Padgett and Ansell (1993)	Retrospective field study	Use data from historical records	Structural holes in marriage relation	Financial gain and power

# Network Data: Samples and Bounds

- **Bounds need to be created for study, networks may be unbounded**
  - For invite networks, move to egocentric networks
  - Sample initial set of individuals randomly
- **Standard sample types for data collection**
  - Random samples
  - Snowball samples
  - Census

# Data Collection

- **Design questionnaire for field study**
  - Formulate question in concert with those being studied
  - Spend time to understand the field and the relationships
  - Understand temporal component
  - Understand open-ended vs close-ended study design
    - In close ended study, you fix the participants/nodes at start of study
    - Open ended network data collection has recall problem
    - Open ended data collection also has infinite size problem



**Trinity College Dublin**  
Coláiste na Tríonóide, Baile Átha Cliath  
The University of Dublin

Trinity Business School

# Software Installation



# Gephi

- Installation of software

# Gephi

- **Understand gephi**
- **Install Plugins**
  - **GeoLayout**
  - **NoverlapLayout**
  - **Multimode Networks Transformation**
  - **GraphStreaming**
  - **TwitterStreamingImporter**
  - **EventGraphLayout**
  - **Vector Statistics**
  - **Minimum Spanning tree**

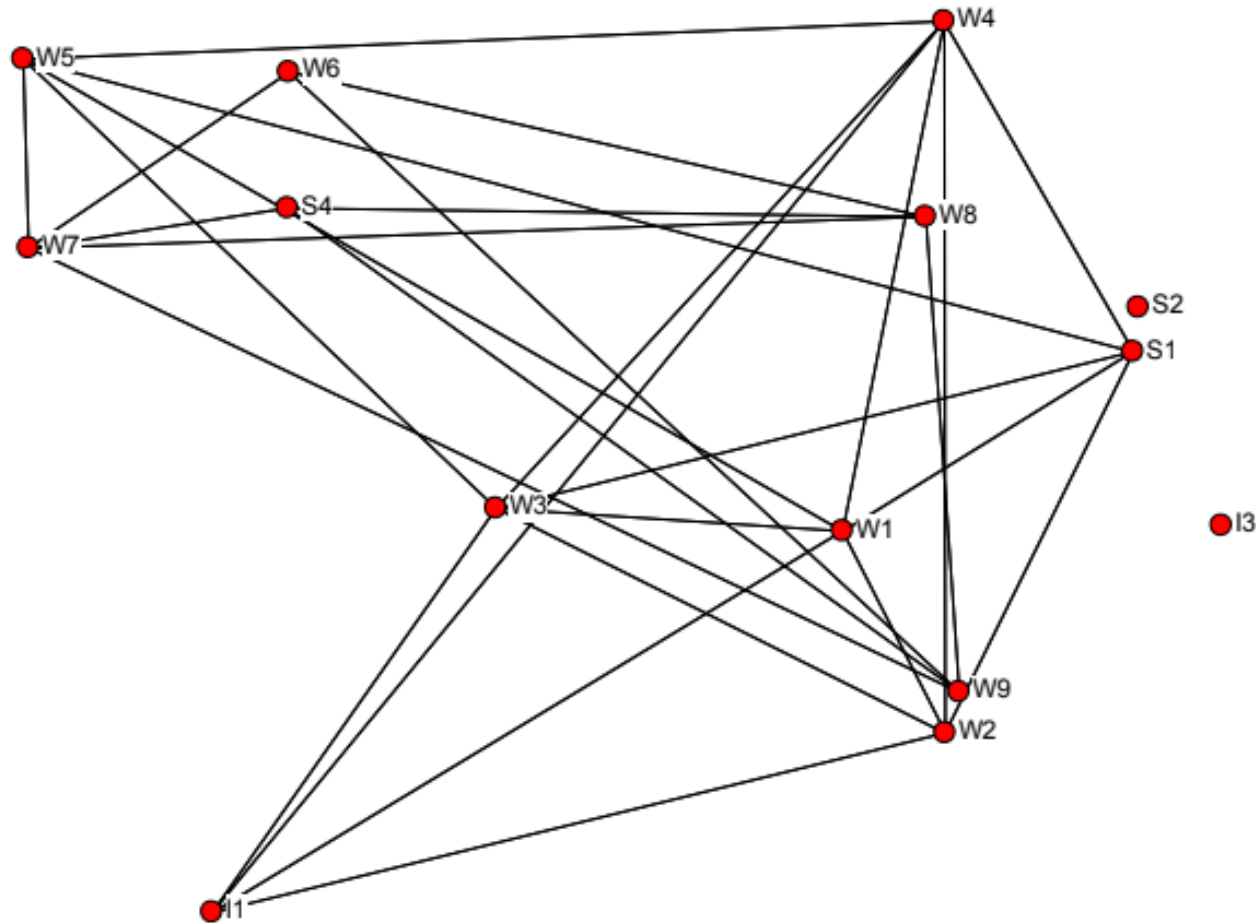
# Gephi

- **Creating graph**

- Setting rank of nodes
- and edges
- Be careful of spline in node size management

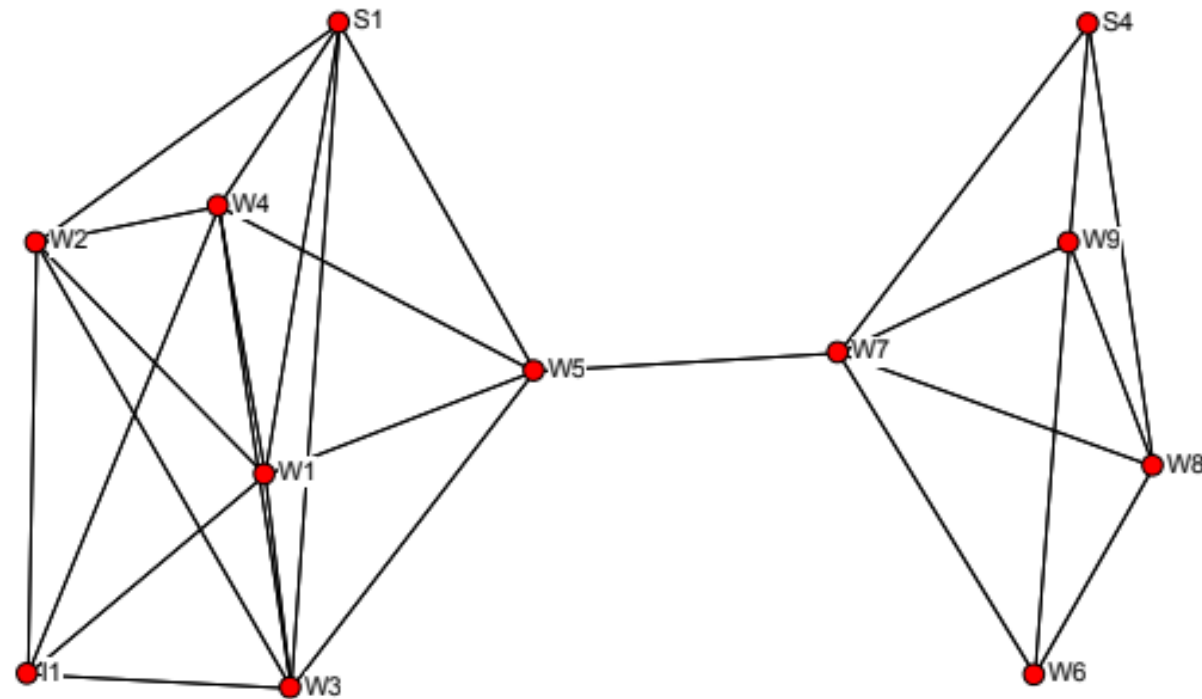


# Random Layout



# Better Layout

● I3  
● S2



# Visualizing

- **Central to understanding graphs**
- **Three major approaches**
  - Attribute based scatter plots
    - Similar nodes closer
    - Used to show connections between attributes and network
  - Ordination (Multidimensional scaling MDS)
    - Uses geodesic distances
    - Good for social sciences
  - Graph layout algorithm
    - Edges pull nodes together
    - Nodes repel each other

# Visualization

- Use line thickness
- Remove weak edges
- Use graph algorithms for cut off

# Graph layout Algorithm

- Very important for visualization
- Based on force directed methods
- Fundamental principle that all edges should be equitable in length with fewest overlapping edges
- Core algorithm
  - Hooke's law for outside edges- to attract them closer
  - Coulomb's law for inside edges- to open them up

# Sample algorithms

- The sum of the force vectors determines which direction a node should move. The step width, which is a constant determines how far a node moves in a single step. When the energy of the system is minimized, the nodes stop moving and the system reaches it's equilibrium state.

## So how to choose a layout?

In general, select one according to the feature of the topology you want to highlight:



# Force Atlas

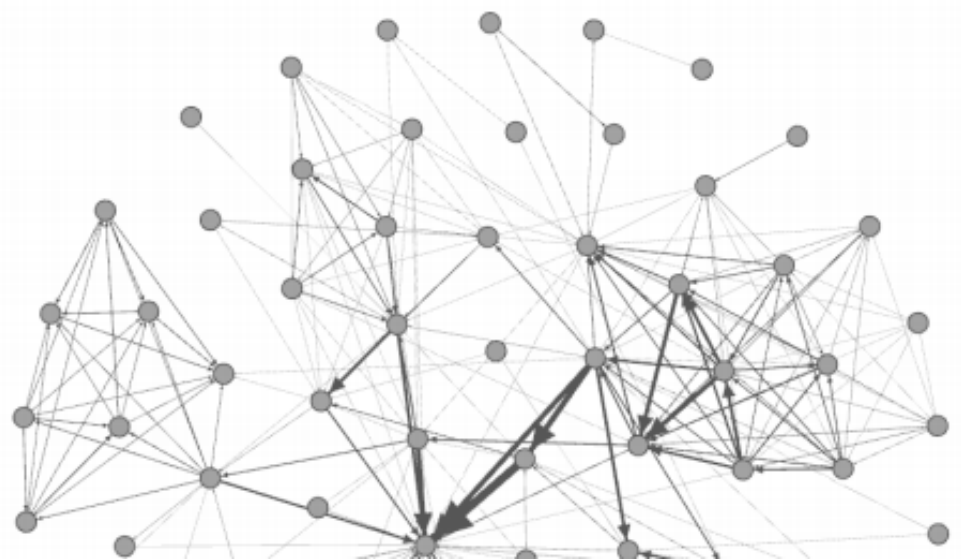
- **Repulsion strength** – How strongly nodes repel each other
- **Autostab strength**- How slowly nodes move
- **Attraction**- How much nodes attract each other
- **Gravity**- How strongly you want nodes to have a center
- **Attract Distrib** – Output links at periphery and input links at center



# Fruchterman-Reingold layout

It simulates the graph as a system of mass particles. The nodes are the mass particles and the edges are springs between the particles. The algorithms try to minimize the energy of this physical system. It has become a standard but remains very slow.

Author:	Thomas Fruchterman & Edward Reingold <sup>1</sup>
Date:	1991
Kind:	Force-directed
Complexity:	$O(N^2)$
Graph size:	1 to 1 000 nodes
Use edge weight:	No



# Run Fruchterman-Reingold



the layout by applying the following settings step by step:

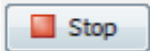
- Area = 100
- Area = 100 000

Graph size area.

- Gravity = 1 000
- Gravity = 100

Attract all nodes to the center to avoid dispersion of disconnected components.

And now

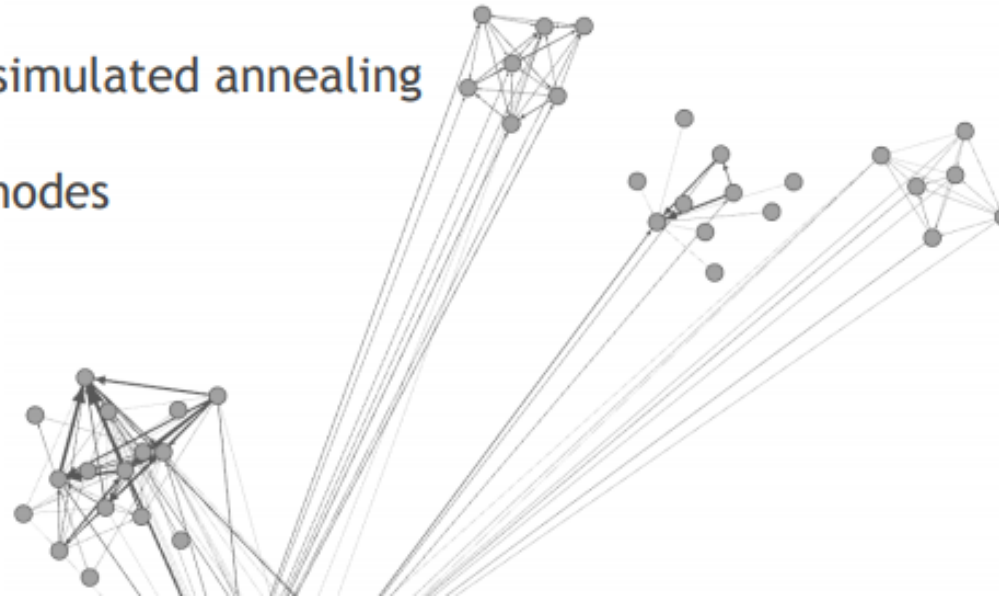


the algorithm.

# OpenOrd layout

It expects undirected weighted graphs and aims to better distinguish **clusters**. It can be run in parallel to speed up computing, and stops automatically. The algorithm is originally based on Fruchterman-Reingold and works with a fixed number of iterations controlled via a simulated annealing type schedule (liquid, expansion, cool-down, crunch, and simmer). Long edges are cut to allow clusters to separate.

Author:	S. Martin, W. M. Brown, R. Klavans, and K. Boyack <sup>1</sup>
Date:	2010 (VxOrd)
Kind:	Force-directed + simulated annealing
Complexity:	$O(N \cdot \log(N))$
Graph size:	100 to 1 000 000 nodes
Use edge weight:	Yes



# Run OpenOrd

Launch the layout by applying the following settings step by step:

- Edge cut = 0.95



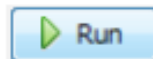
From 0 (standard Fruchterman-Reingold) to 1. Percentage of the greatest distance between two nodes in the drawing. A higher cutting means a more clustered result.

- Num iterations = 100
- Num iterations = 850



Contract the clusters.  
Expand the clusters.

- Random seed =  
-6308261588084905834



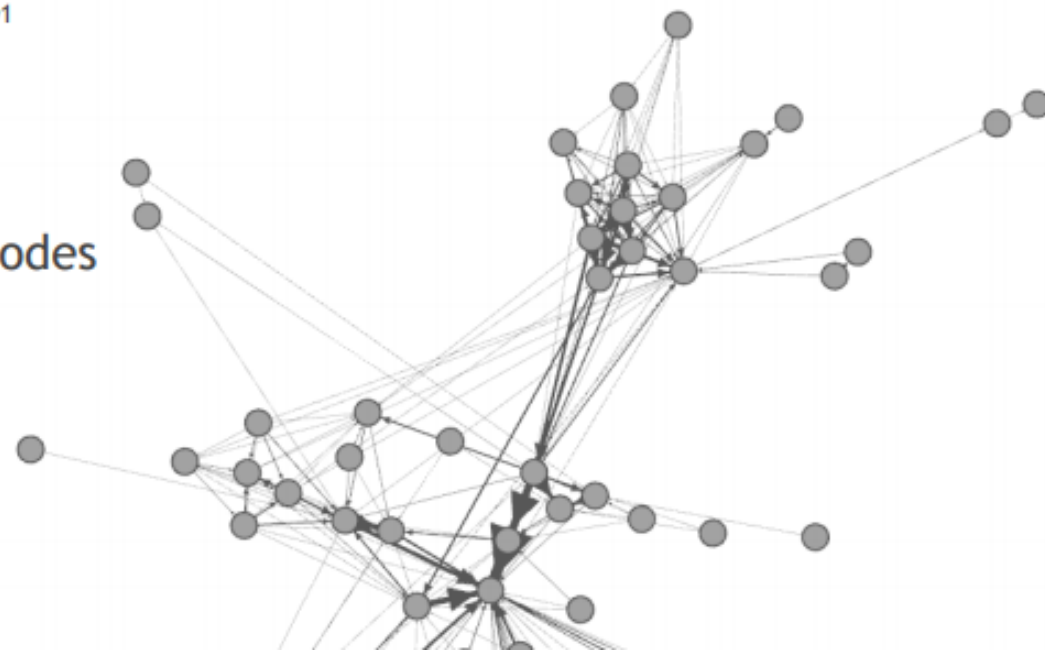
Use this value to produce exactly the same shape as shown before.




# ForceAtlas 2 layout

Improved version of the Force Atlas to handle large networks while keeping a very good quality. Nodes repulsion is approximated with a Barnes-Hut calculation, which therefore reduces the algorithm complexity. Replace the “attraction” and “repulsion” forces by a “scaling” parameter.

Author:	Mathieu Jacomy <sup>1</sup>
Date:	2011
Kind:	Force-directed
Complexity:	$O(N \cdot \log(N))$
Graph size:	1 to 1 000 000 nodes
Use edge weight:	Yes



# Run ForceAtlas 2

 the layout by applying the following settings step by step:

- LinLog mode = checked
- LinLog mode = unchecked


Linear attraction & logarithmic repulsion (lin-lin by default), makes clusters tighter.

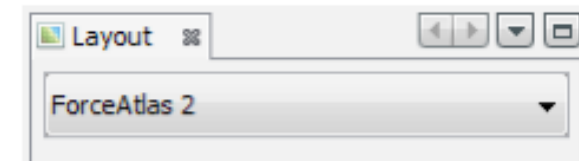
- Scaling = 100

Increase to make the graph sparser.

- Edge weight influence = 0

From 0 (no influence) to 1 (normal). Set 0 to calculate forces without edge weight.

And now  the algorithm.

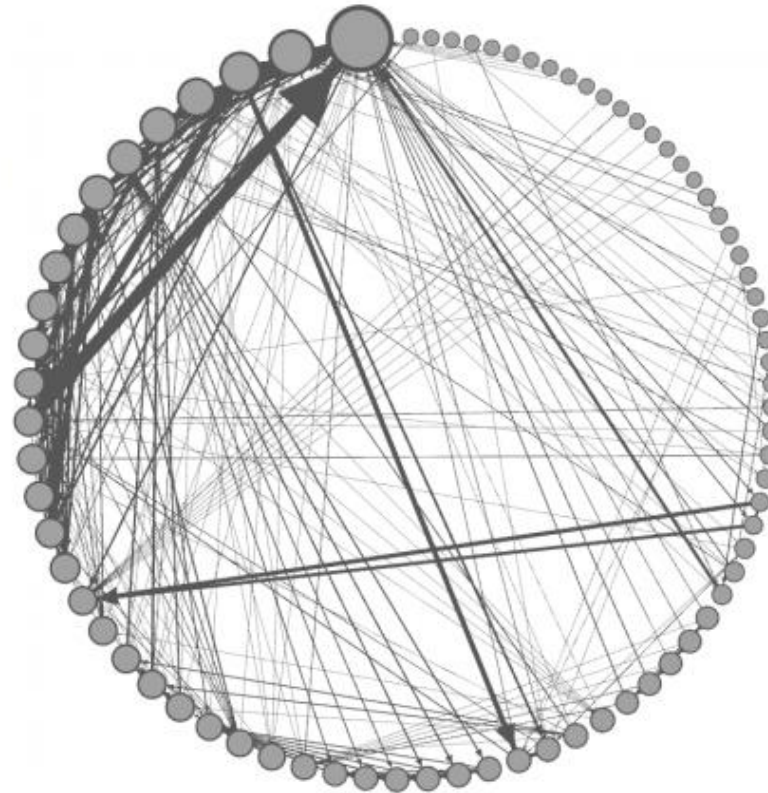





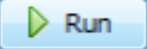
# Circular layout

It draws nodes in a circle ordered by ID, a metric (degree, betweenness centrality...) or by an attribute. Use it to show a distribution of nodes with their links.

Author:	Matt Groeninger <sup>1</sup>
Date:	2010
Kind:	Circular
Complexity:	$O(N)$
Graph size:	1 to 1 000 000 nodes



# Run Circular Layout

- Select the “Circular Layout” in the  Layout panel.
- Set the “Order nodes by” setting to “Degree”.
-  the layout.

