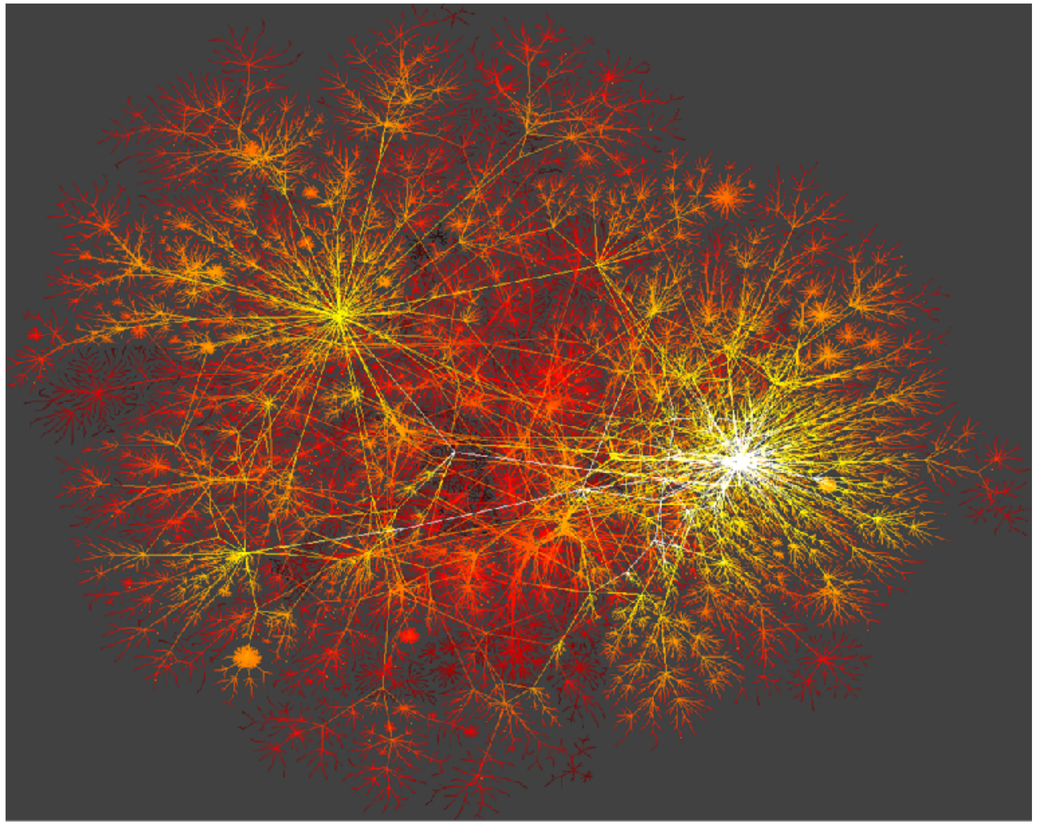
Class 21: Introduction to network science

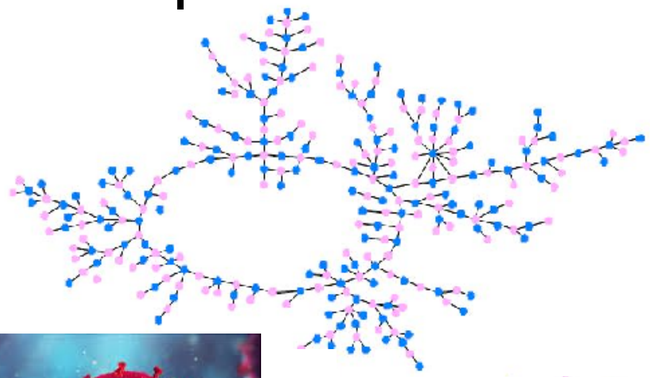
Instructor: Michael Szell
Nov 13, 2019

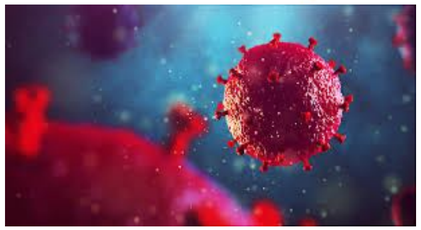


IT UNIVERSITY OF COPENHAGEN

Today you will learn about networks

Why networks are important





Mathematical concepts and data structures

$$I = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 3 \\ 4 \end{pmatrix} \quad J = \begin{pmatrix} 5 \\ 1 \\ 3 \\ 1 \\ 2 \\ 3 \end{pmatrix}$$

Organizing principles

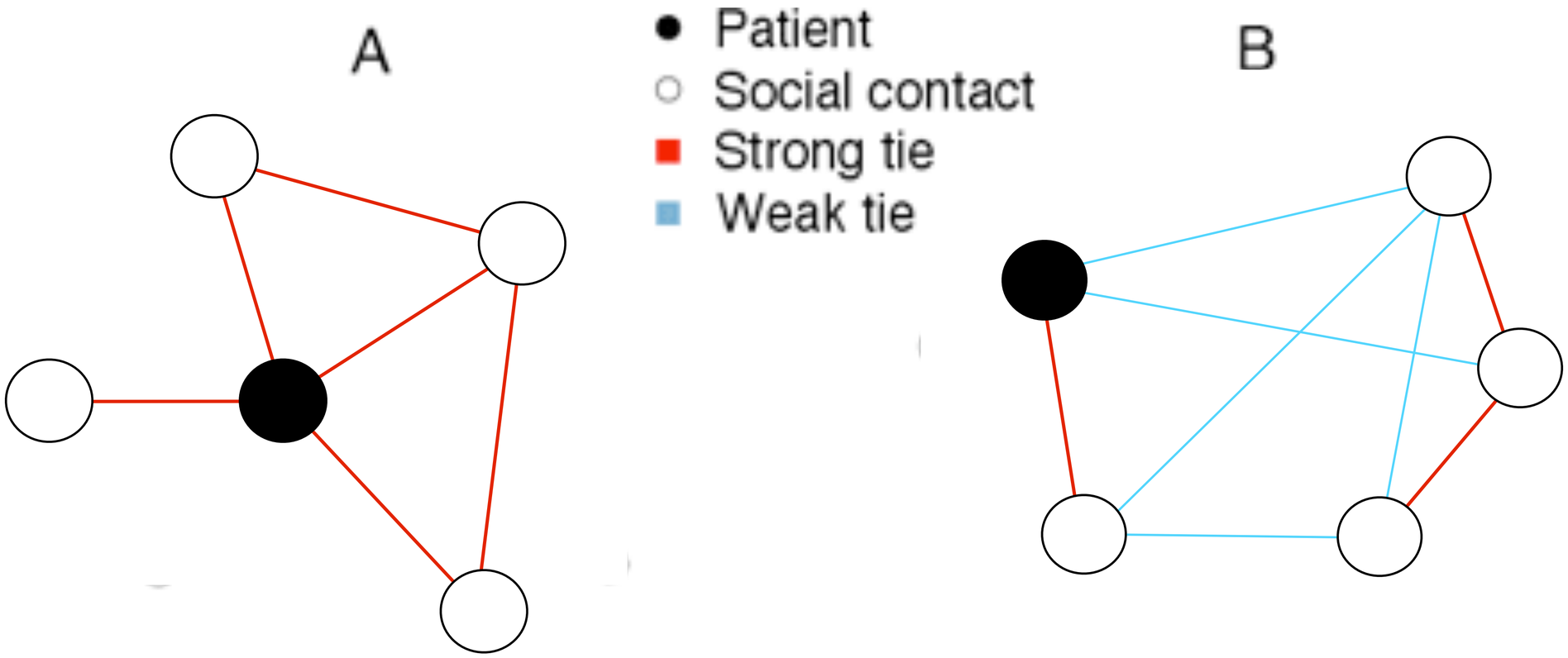
After a stroke, how fast do you arrive in the hospital?

© 2013 American Stroke Association

What matters most?

https://www.mentimeter.com/s/386e1866e0c5f9de383d0ed886ef98c0/771f2a4ba0a0

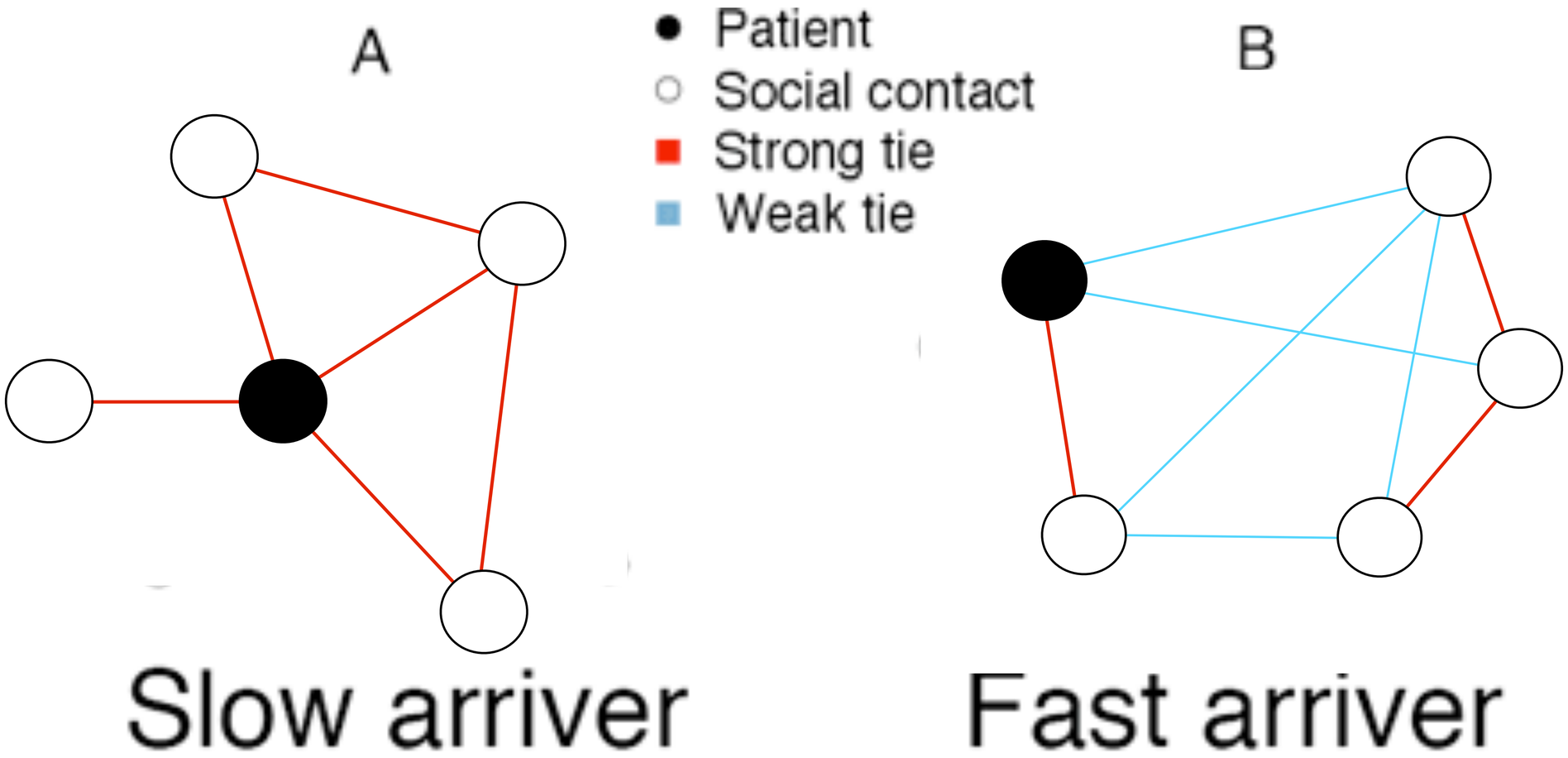
Your social network matters. But how?



Who arrives first in the hospital?

https://www.mentimeter.com/s/386e1866e0c5f9de383d0ed886ef98c0/771f2a4ba0a0

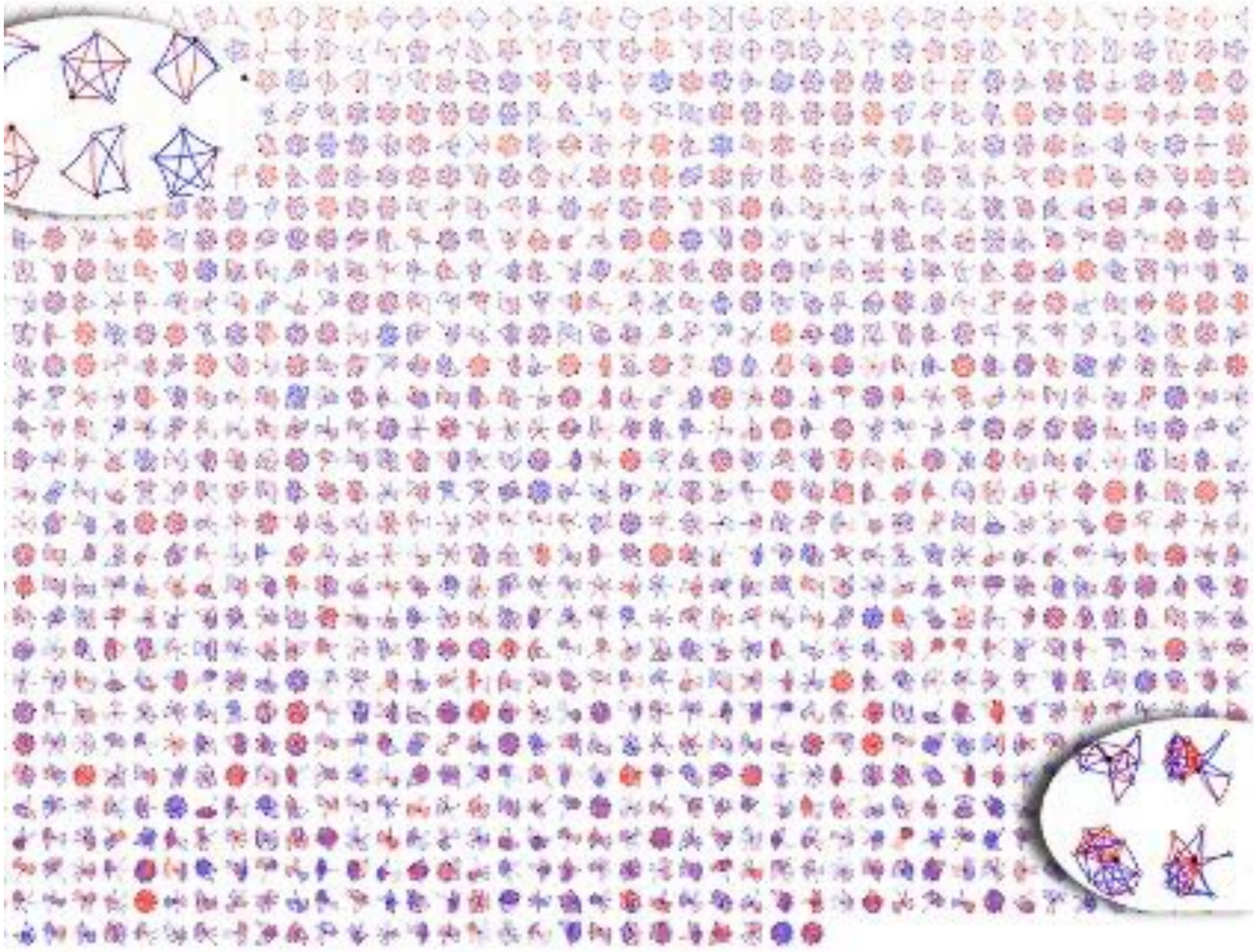
Your social network matters



Research by Amar Dhand, Northeastern University

Video
(removed for privacy reasons)

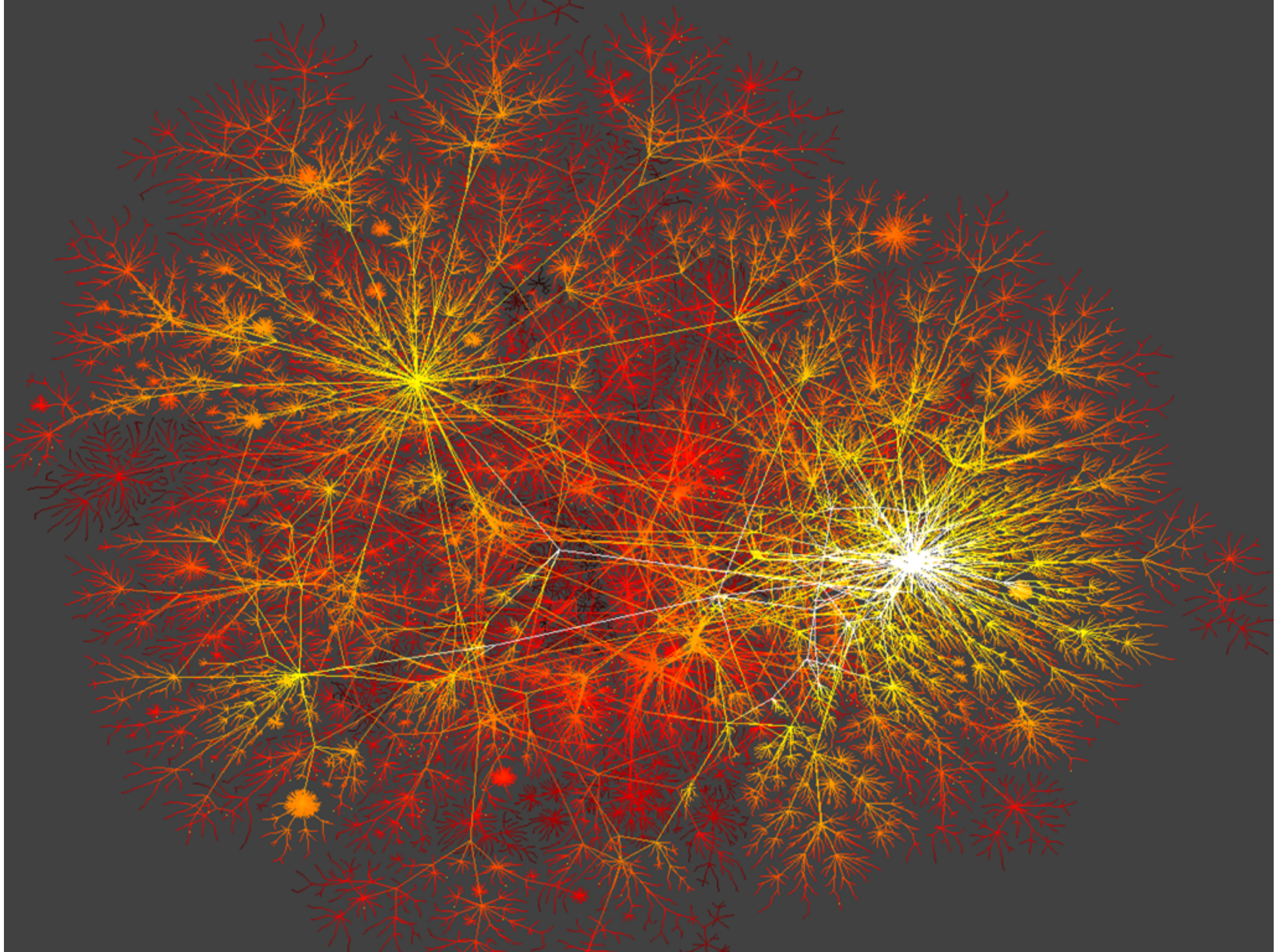
Your social network matters

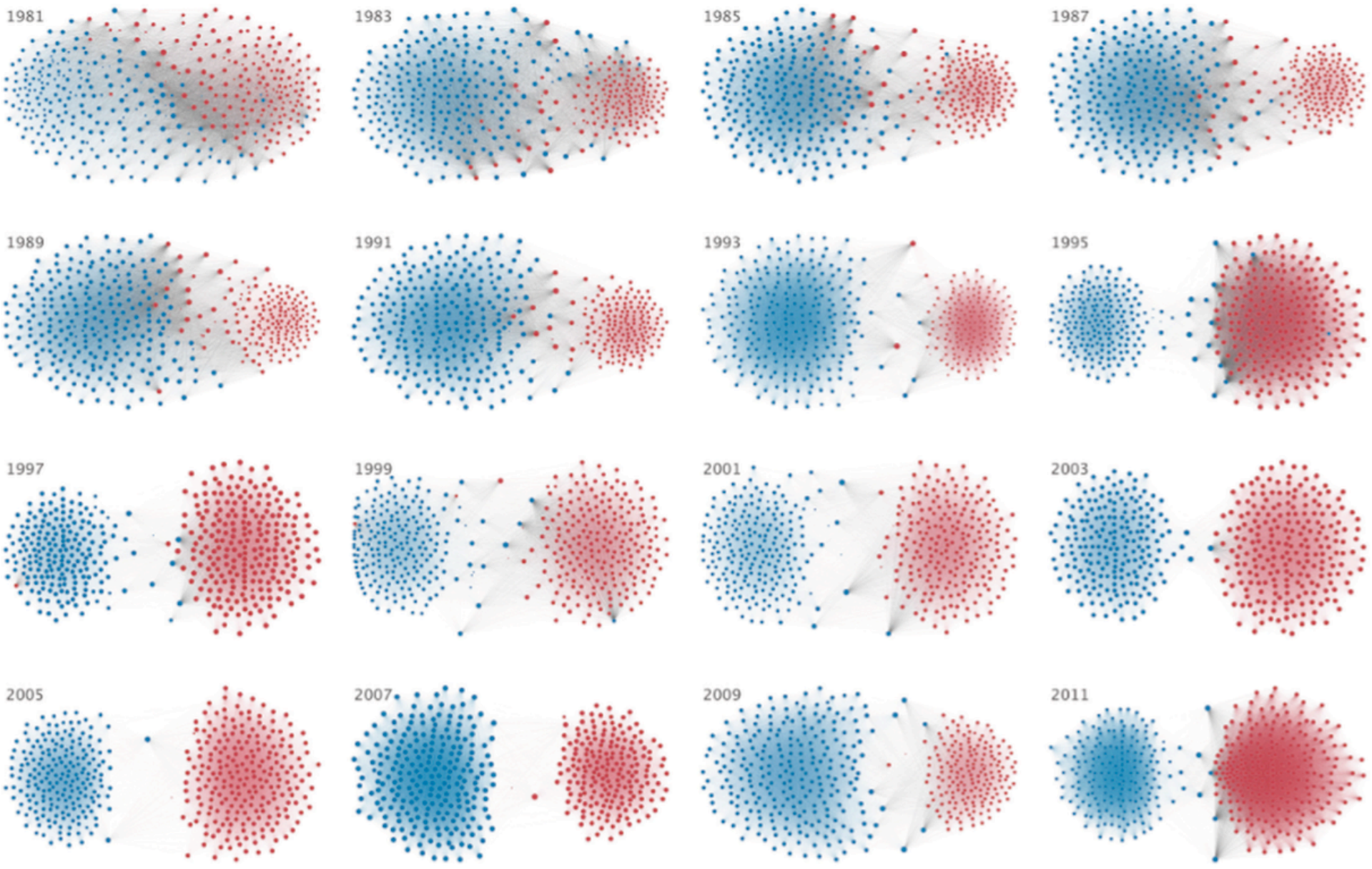


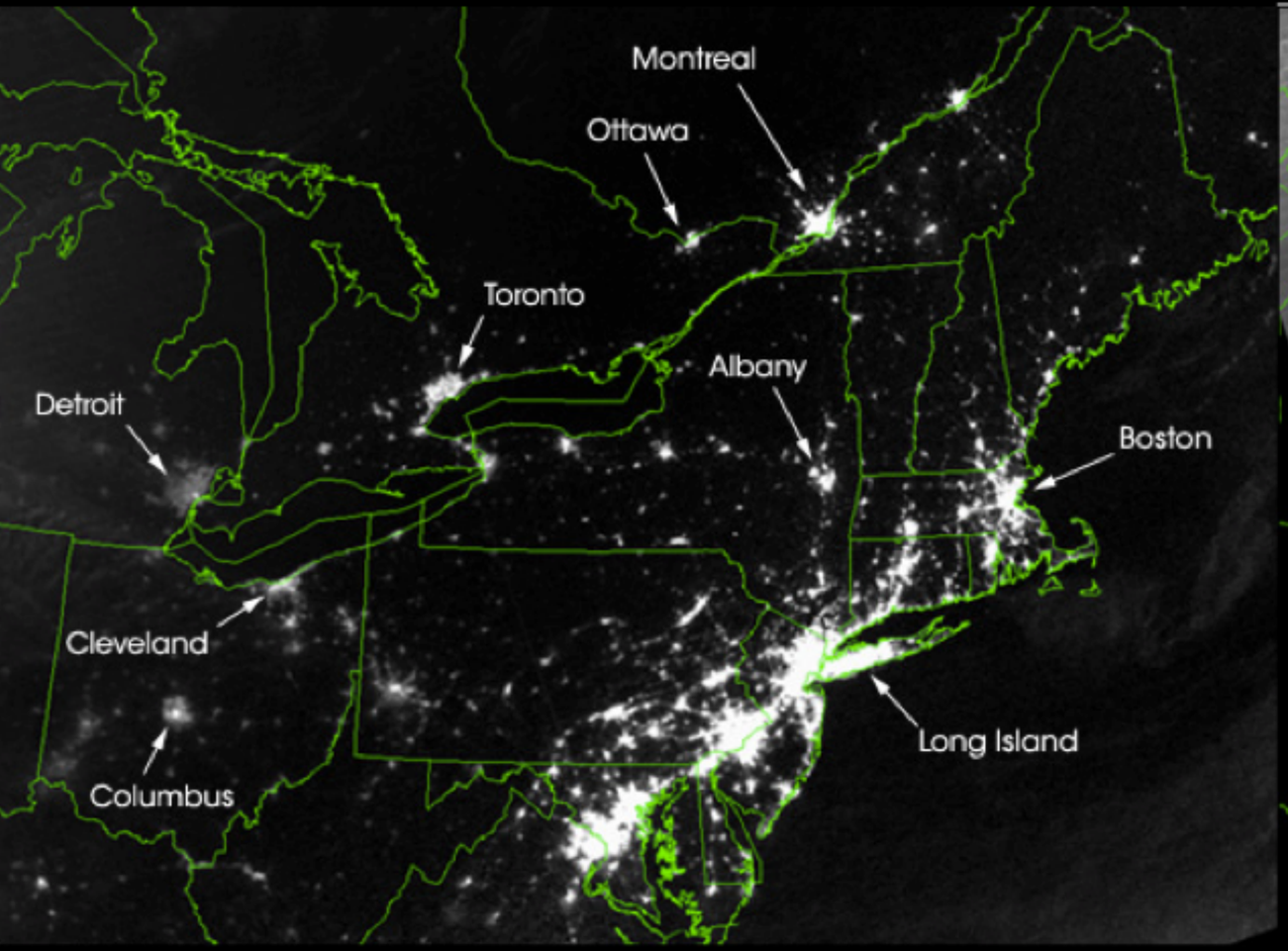
Research by Amar Dhand, Northeastern University

Networks describe interactions or relations

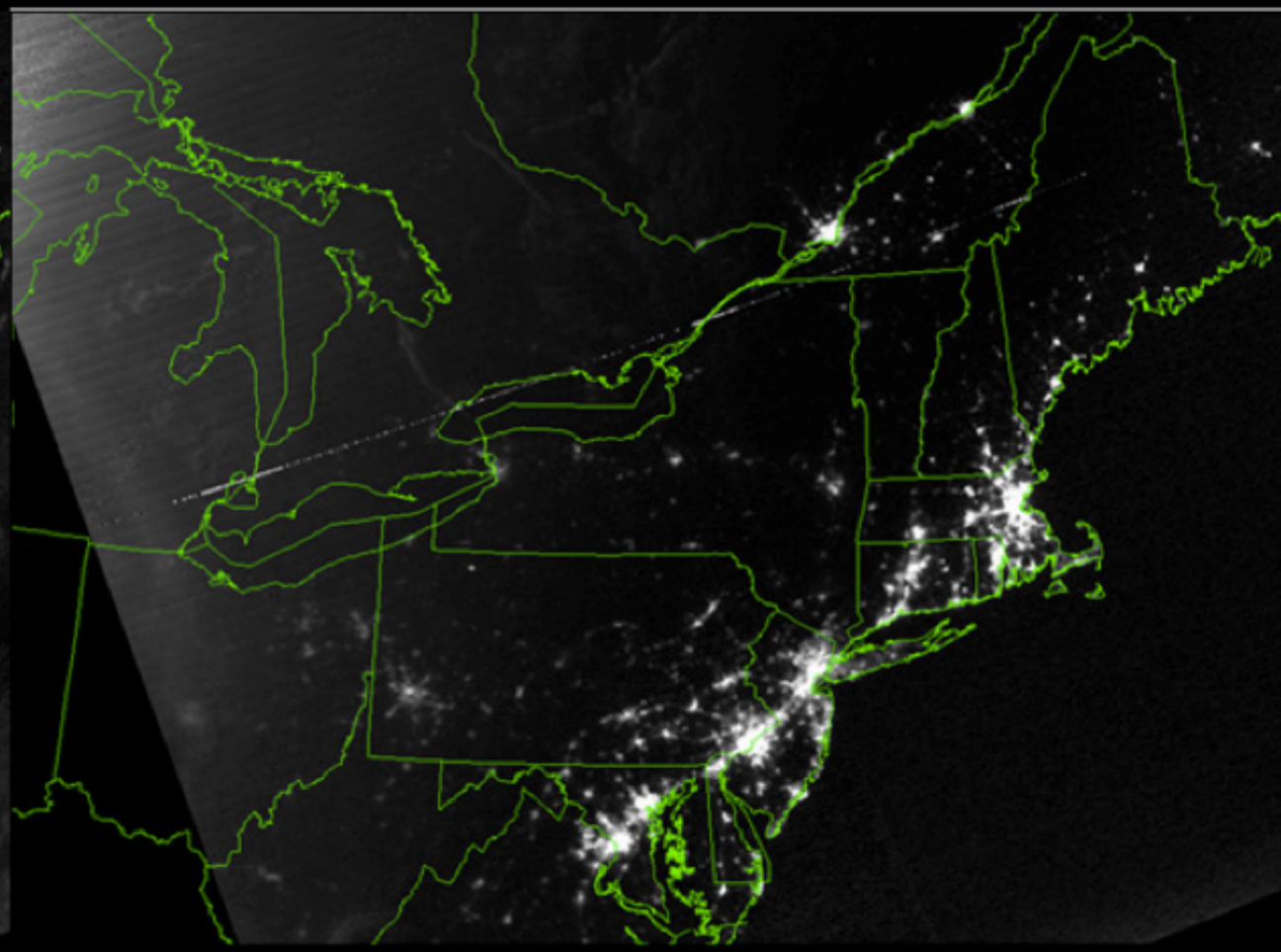
facebook
December 2010



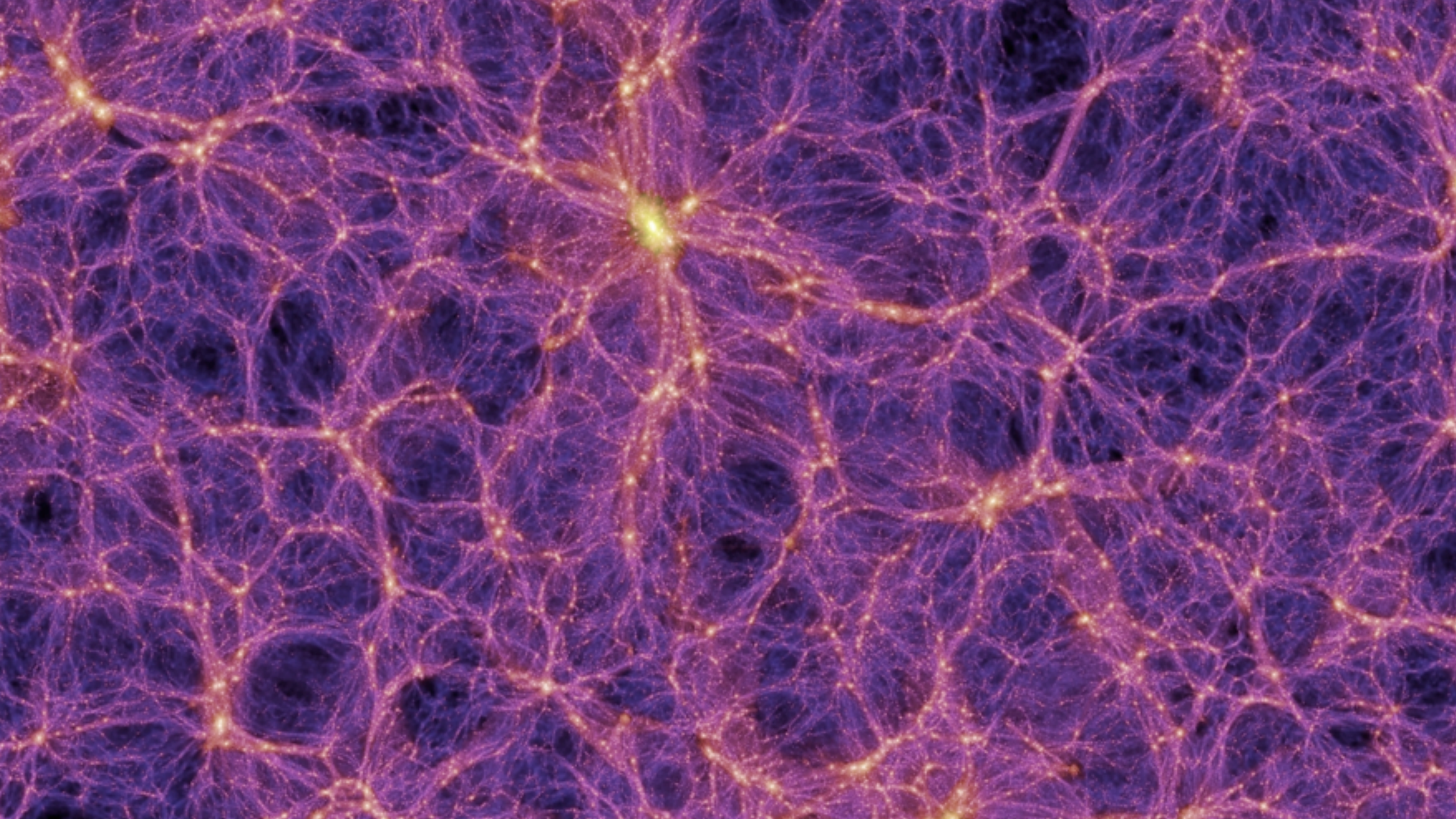

1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011

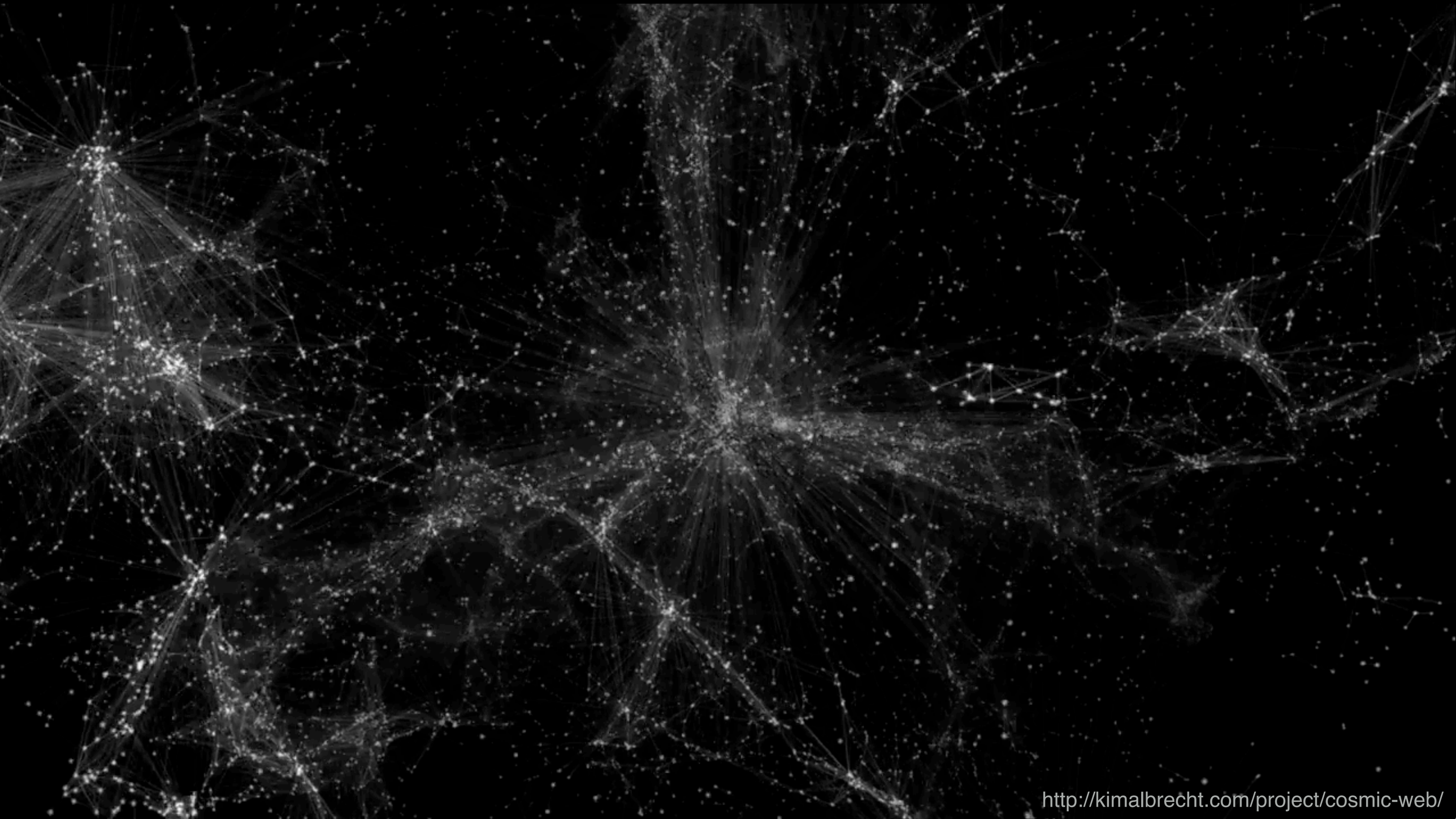


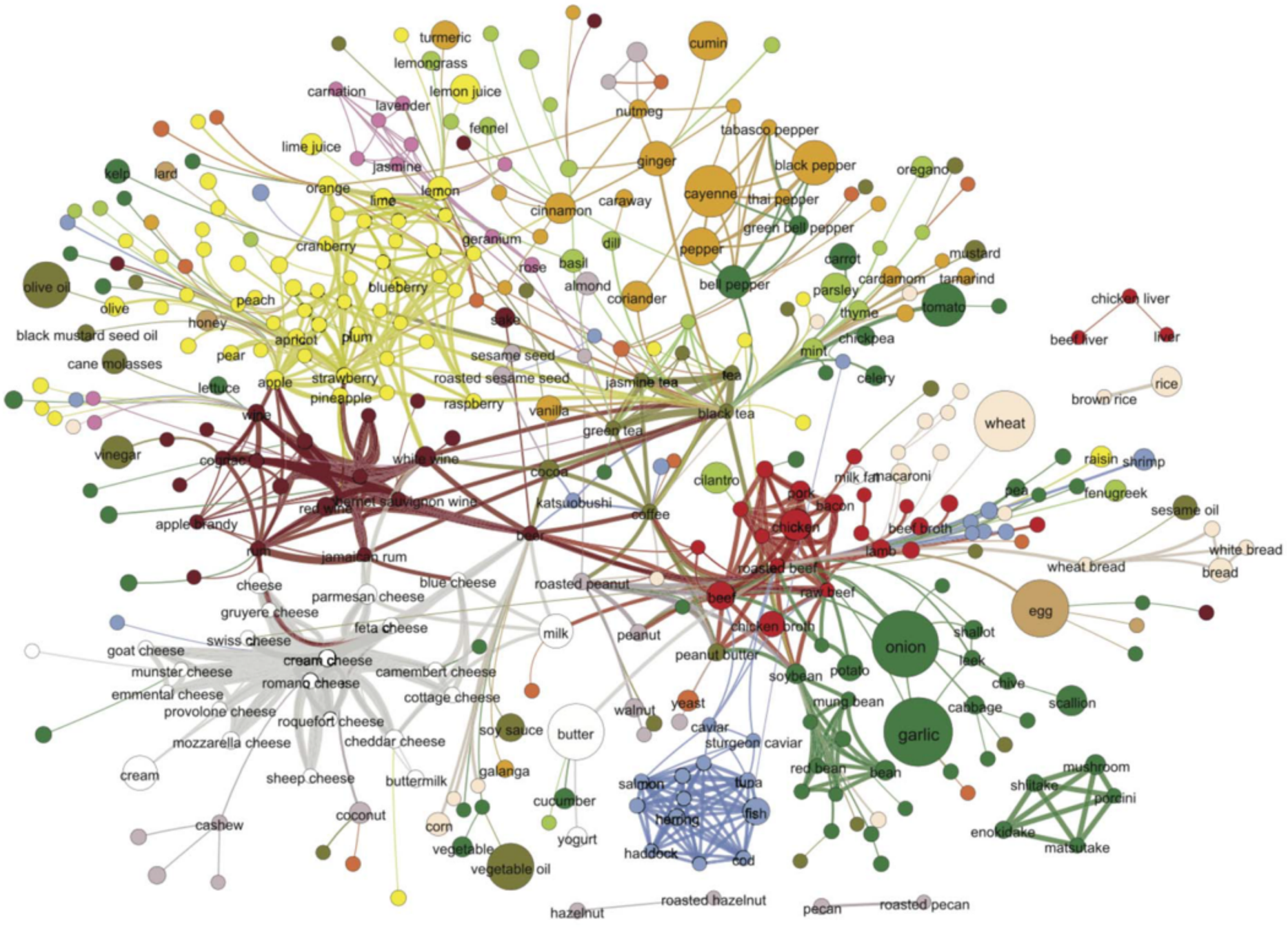
August 14, 2003: 9:29pm EDT
20 hours before

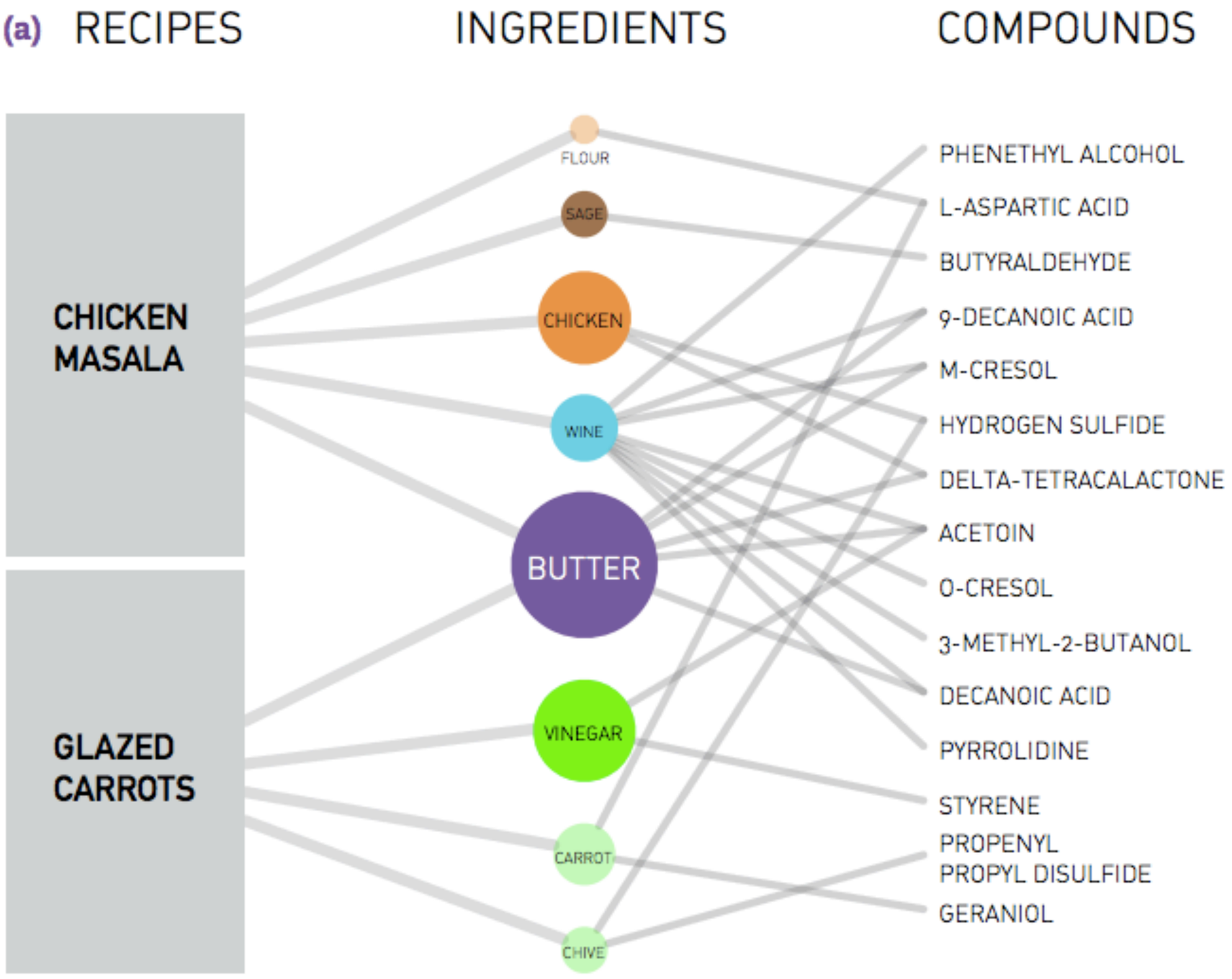


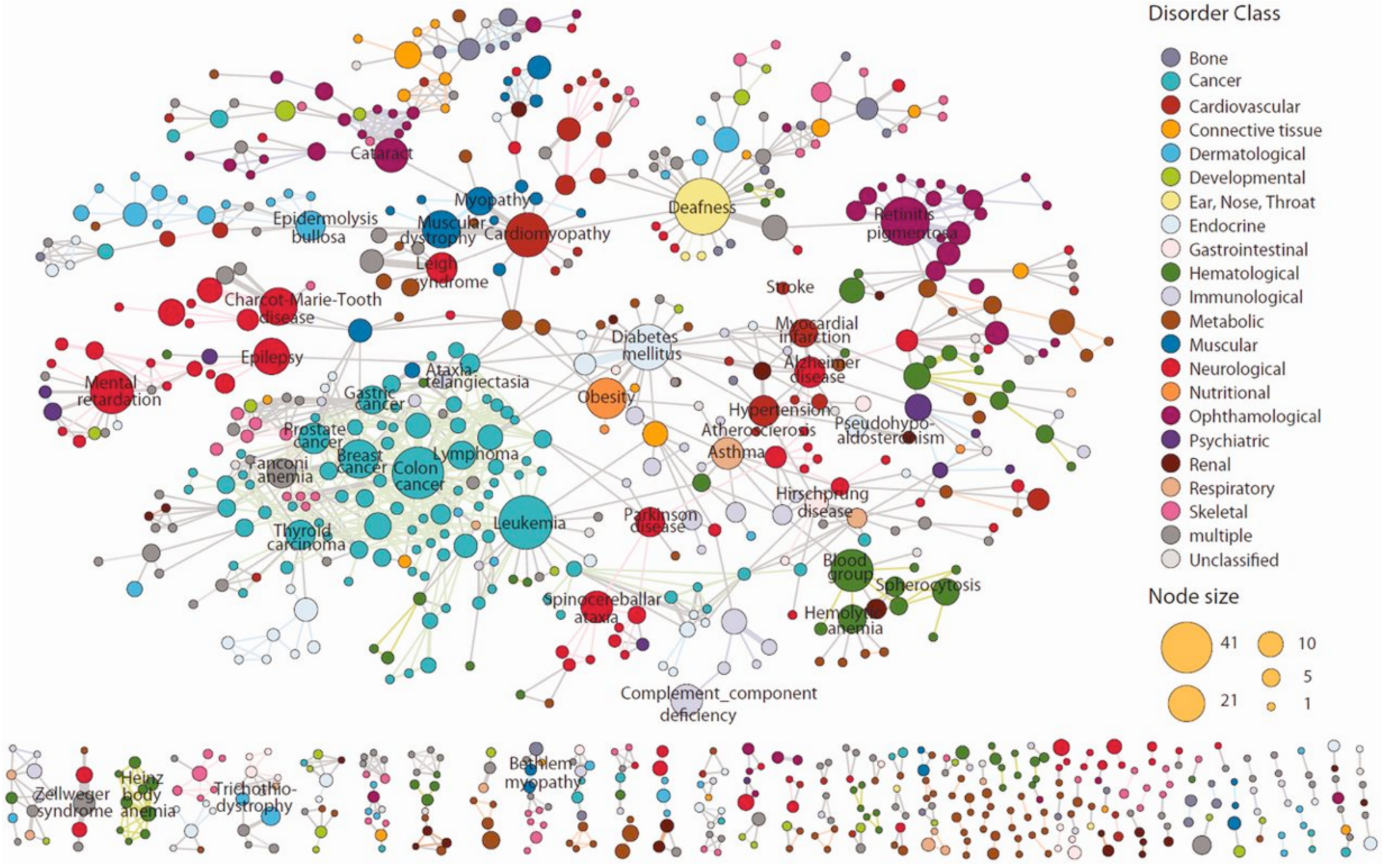
August 15, 2003: 9:14pm EDT
7 hours after

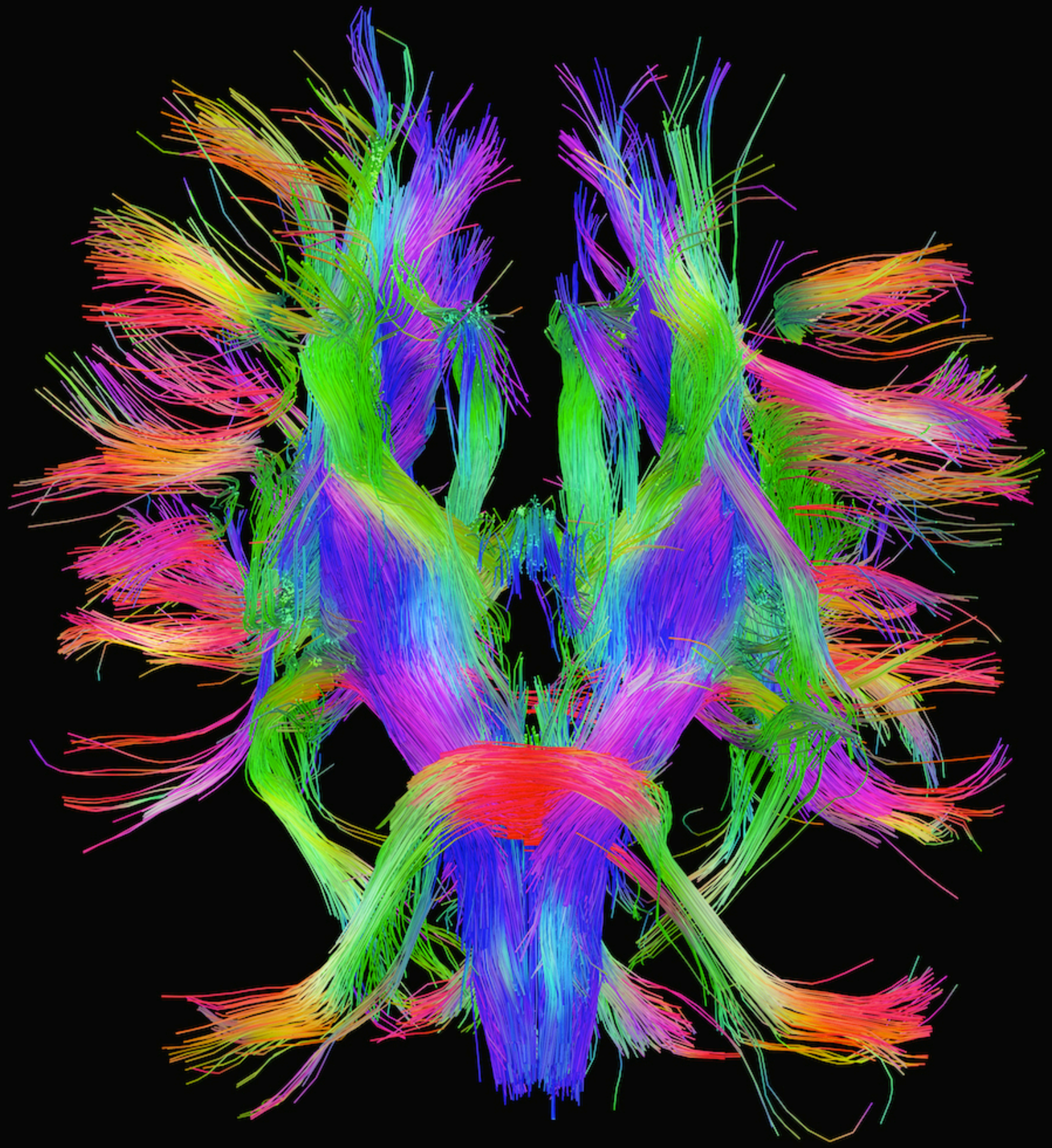


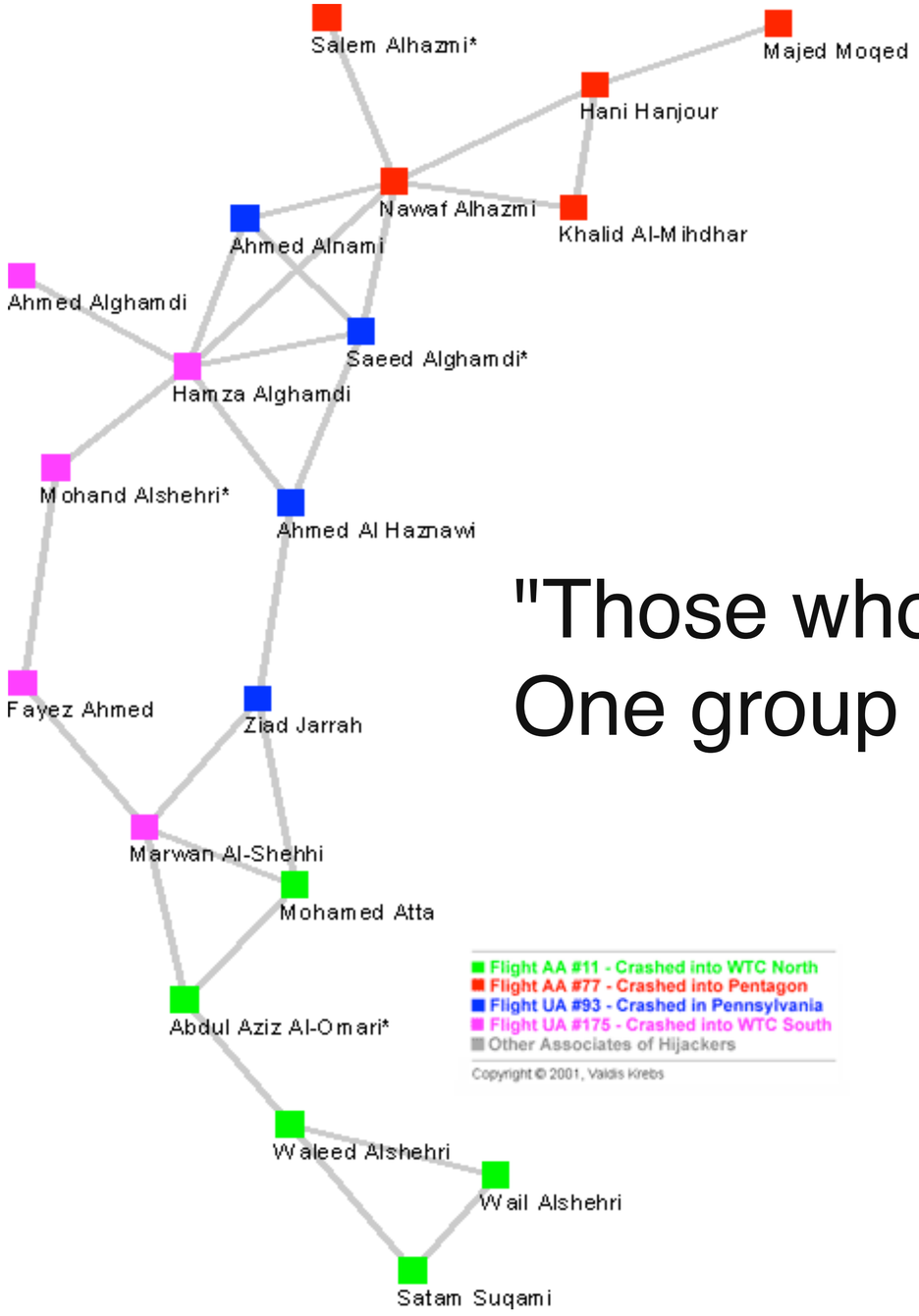

http://kimalbrecht.com/project/cosmic-web/


turmeric
lemongrass
cumin
carnation
lavender
lemon juice
nutmeg
fennel
tabasco pepper
lime juice
jasmine
ginger
black pepper
oregano
kelp
lard
orange
lemon
lime
cinnamon
caraway
cayenne
thai pepper
green bell pepper
cranberry
geranium
dill
pepper
carrot
mustard
rose
basil
olive oil
blueberry
almond
bell pepper
parsley
cardamom
tamarind
coriander
chicken liver
olive
peach
thyme
tomato
honey
sake
black mustard seed oil
apricot
plum
chickpea
beef liver
liver
pear
sesame seed
mint
cane molasses
strawberry
roasted sesame seed
jasmine tea
tea
celery
rice
lettuce
apple
pineapple
raspberry
vanilla
brown rice
wine
black tea
wheat
green tea
vinegar
cognac
white wine
raisin
shrimp
cocoa
cilantro
milk fat
macaroni
pea
fenugreek
red wine
garnet sauvignon wine
katsuobushi
pork
bacon
coffee
sesame oil
apple brandy
chicken
beef broth
beer
rum
jamaican rum
lamb
white bread
roasted beef
wheat bread
bread
cheese
blue cheese
roasted peanut
parmesan cheese
beef
raw beef
egg
gruyere cheese
feta cheese
milk
peanut
chicken broth
shallot
swiss cheese
onion
goat cheese
cream cheese
peanut butter
leek
munster cheese
camembert cheese
soybean
potato
romano cheese
chive
emmental cheese
cottage cheese
mung bean
cabbage
scallion
provolone cheese
walnut
yeast
roquefort cheese
caviar
garlic
soy sauce
butter
mozzarella cheese
cheddar cheese
sturgeon caviar
red bean
bean
mushroom
cream
sheep cheese
buttermilk
galanga
salmon
tuna
shiitake
porcini
cucumber
herring
fish
coconut
cashew
corn
yogurt
enokidake
vegetable
haddock
cod
matsutake
vegetable oil
hazelnut
roasted hazelnut
pecan
roasted pecan


(a)
RECIPES
INGREDIENTS
COMPOUNDS
CHICKEN MASALA
GLAZED CARROTS
FLOUR
SAGE
CHICKEN
WINE
BUTTER
VINEGAR
CARROT
CHIVE
PHENETHYL ALCOHOL
L-ASPARTIC ACID
BUTYRALDEHYDE
9-DECANOIC ACID
M-CRESOL
HYDROGEN SULFIDE
DELTA-TETRACALACTONE
ACETOIN
O-CRESOL
3-METHYL-2-BUTANOL
DECANOIC ACID
PYRROLIDINE
STYRENE
PROPENYL PROPYL DISULFIDE
GERANIOL


Disorder Class
Bone
Cancer
Cardiovascular
Connective tissue
Dermatological
Developmental
Ear, Nose, Throat
Endocrine
Gastrointestinal
Hematological
Immunological
Metabolic
Muscular
Neurological
Nutritional
Ophthamological
Psychiatric
Renal
Respiratory
Skeletal
multiple
Unclassified
Node size
41
21
10
5
1
Cataract
Myopathy
Epidermolysis bullosa
Muscular dystrophy
Cardiomyopathy
Leigh syndrome
Deafness
Retinitis pigmentosa
Charcot-Marie-Tooth disease
Stroke
Myocardial infarction
Diabetes mellitus
Epilepsy
Alzheimer disease
Ataxia-telangiectasia
Mental retardation
Gastric cancer
Obesity
Hypertension
Prostate cancer
Atherosclerosis
Pseudohypo-aldosteronism
Breast cancer
Lymphoma
Asthma
Fanconi anemia
Colon cancer
Hirschprung disease
Parkinson disease
Leukemia
Thyroid carcinoma
Blood group
Spherocytosis
Spinocerebellar ataxia
Hemolytic anemia
Complement_component deficiency
Zellweger syndrome
Heinz body anemia
Trichothio-dystrophy
Bethlem myopathy





"Those who were trained to fly didn't know the others. One group of people did not know the other group."

https://youtu.be/CtWqv0Z3ErM?t=767

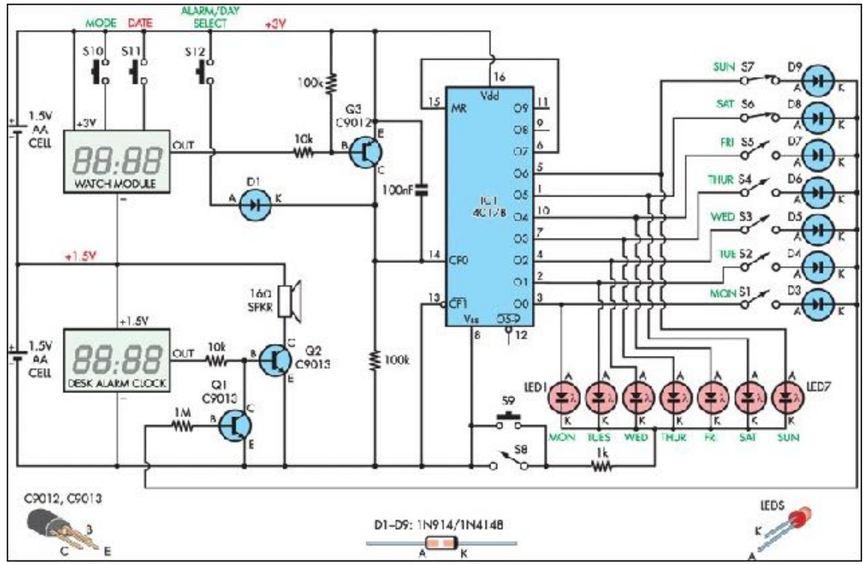
Networks are the backbones of complex systems

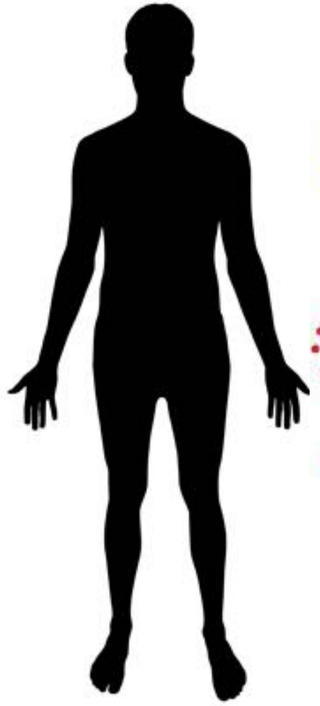
Many single units

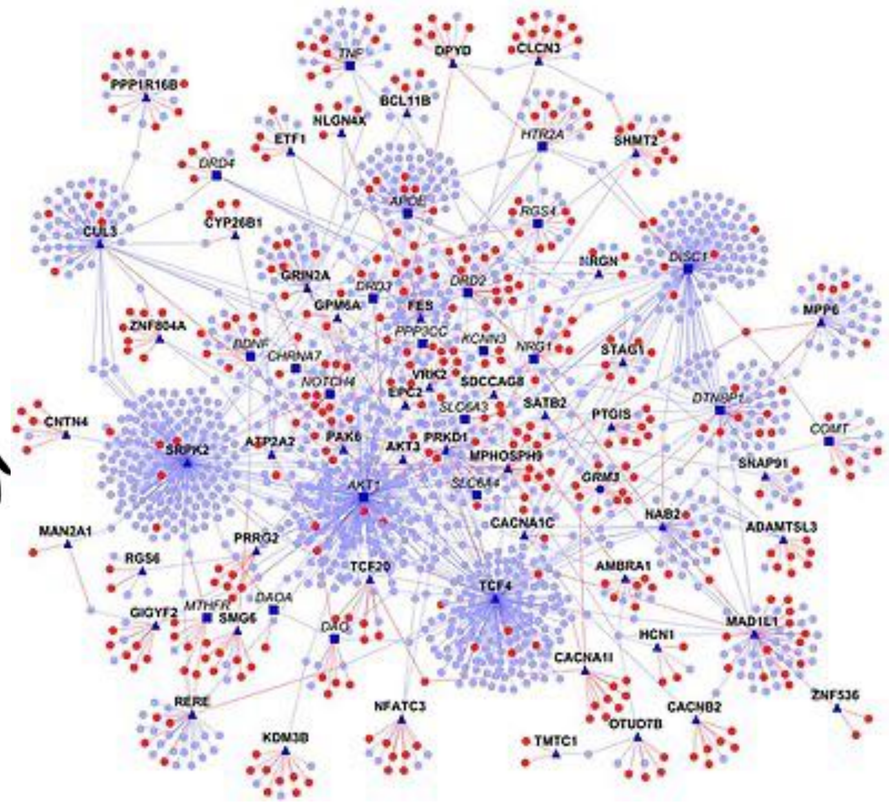
Strong, nonlinear interactions

Emergence of collective behavior

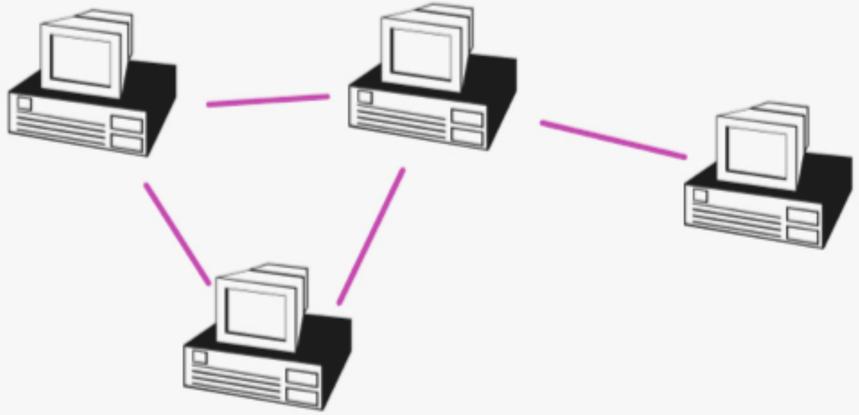
Networks are the maps of complex systems
We now have the data to see and study them

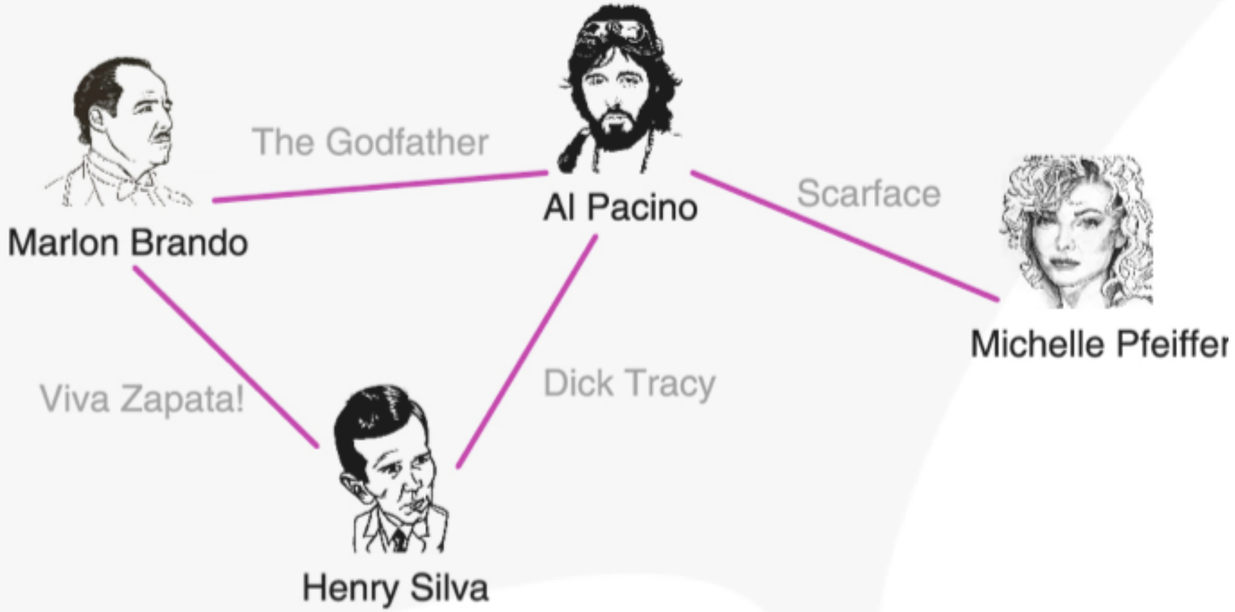


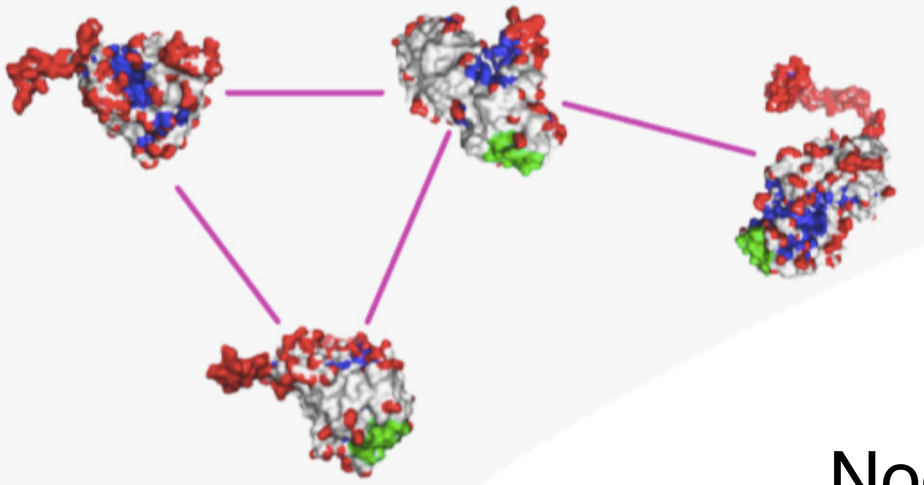


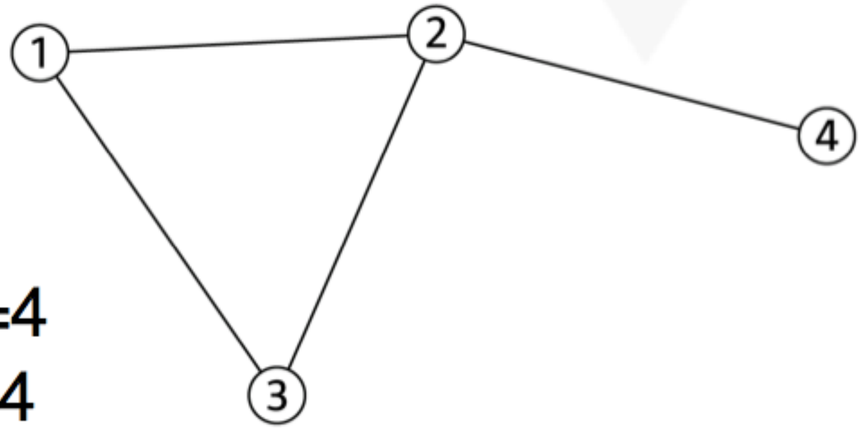


Networks are a common language for different applications



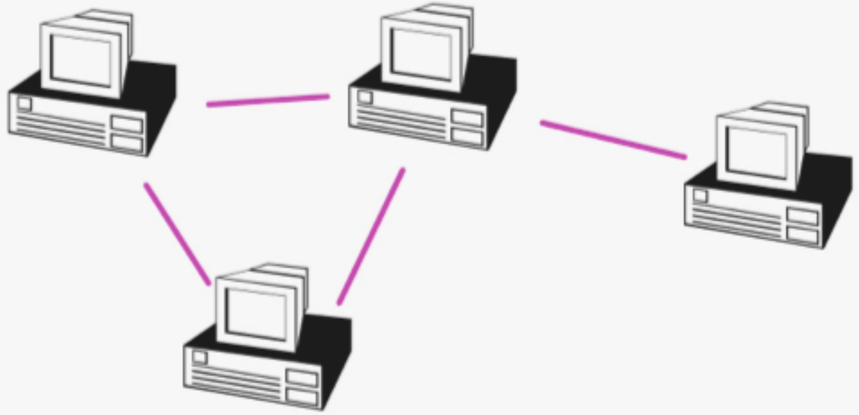


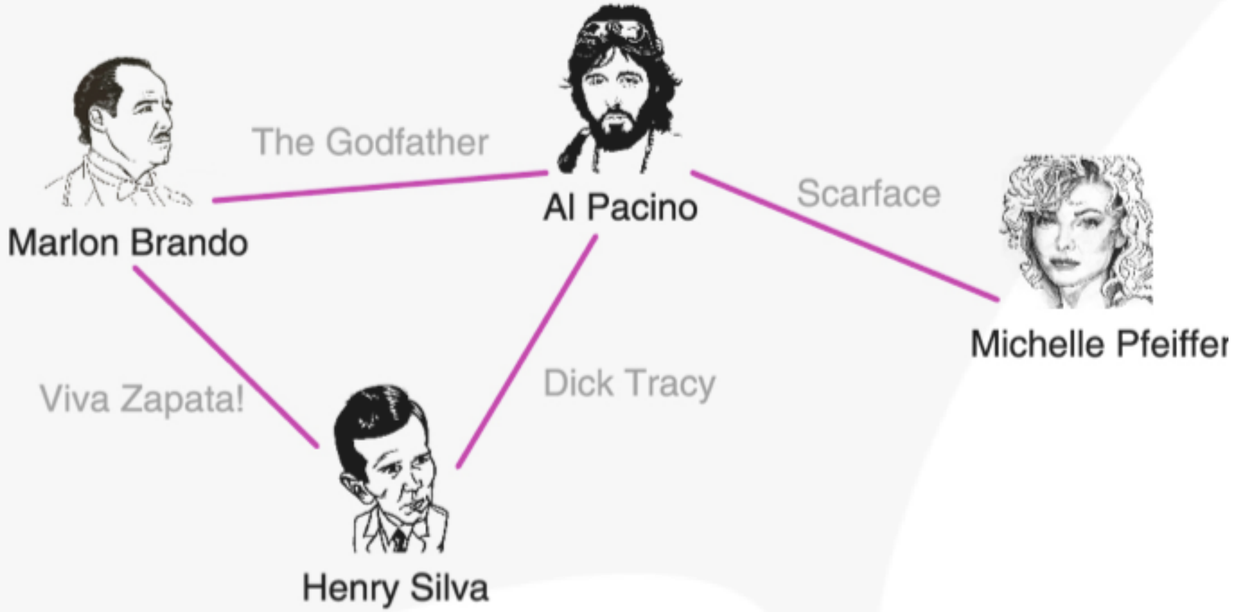


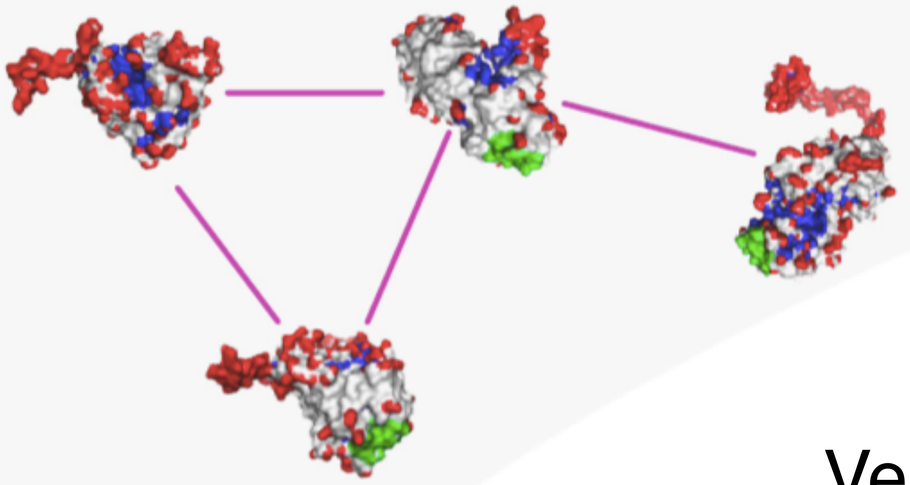


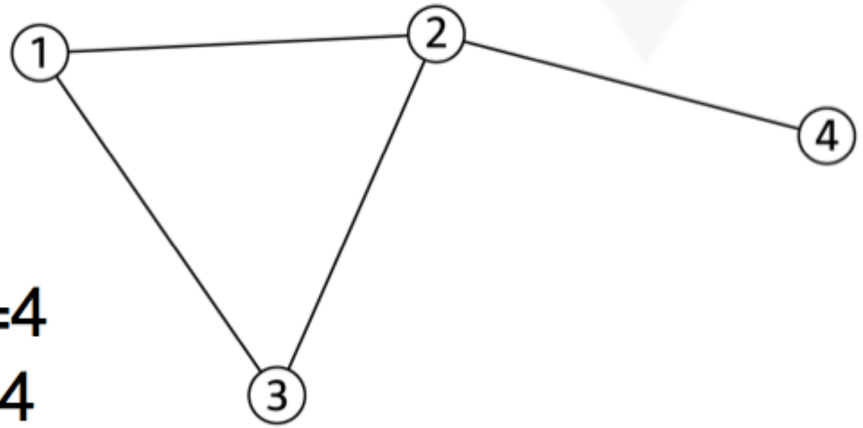
Nodes	N=4
Links	L=4
Graph	$\mathcal{G} = (\mathcal{N}, \mathcal{L})$

Network = Graph + real-world meaning







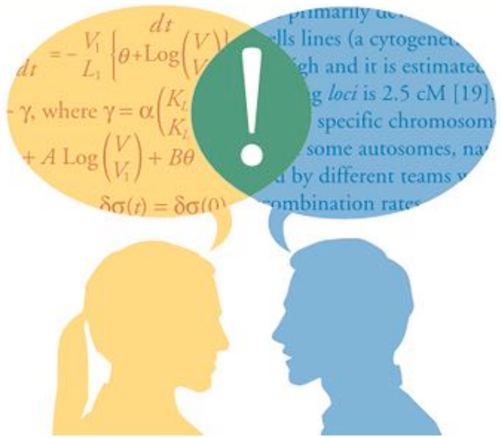


Vertices V=4

Edges E=4

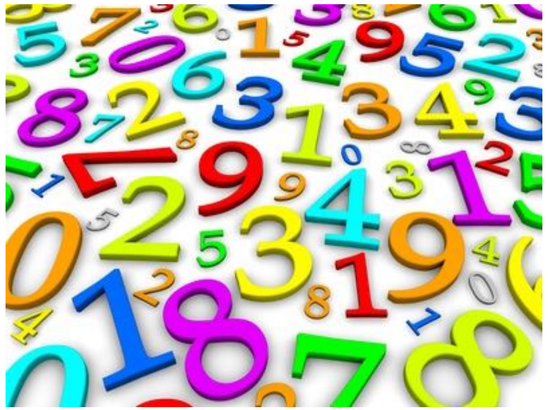
Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Network science is:



Interdisciplinary

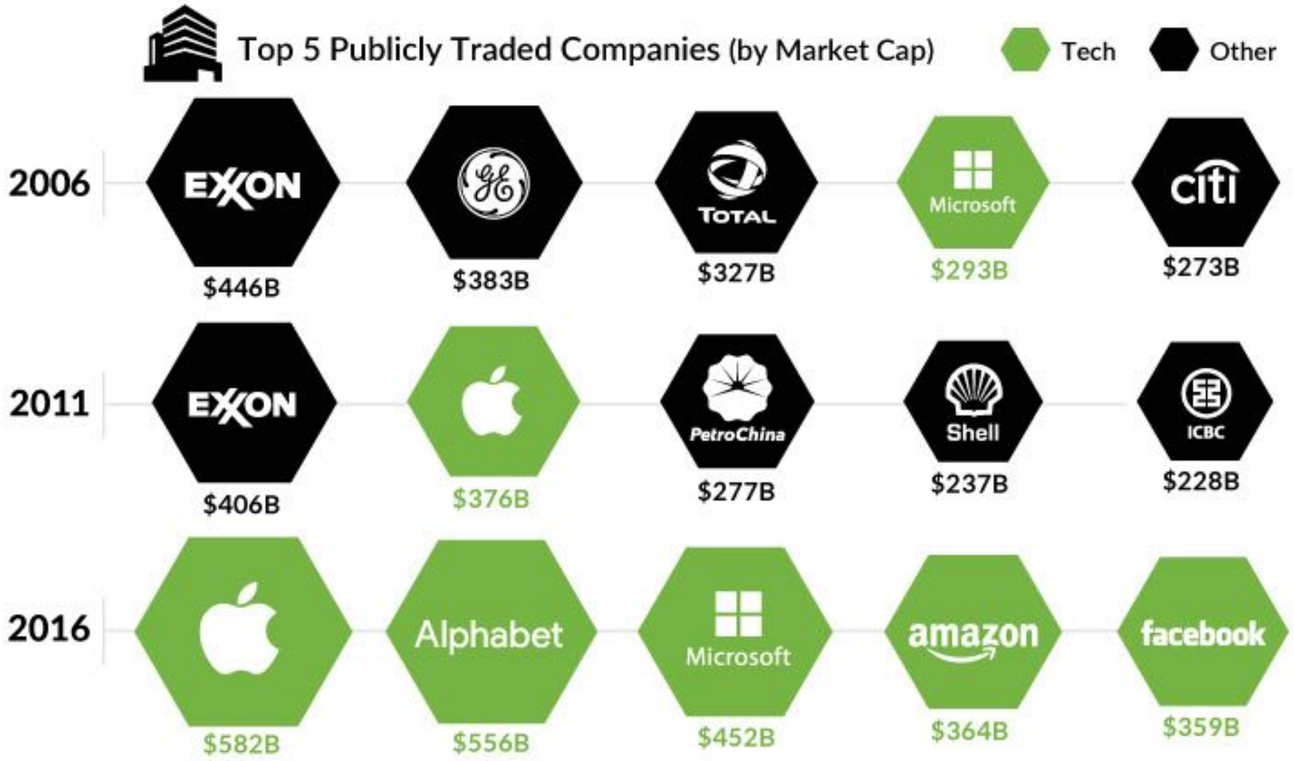
Empirical



Quantitative, mathematical

Computational

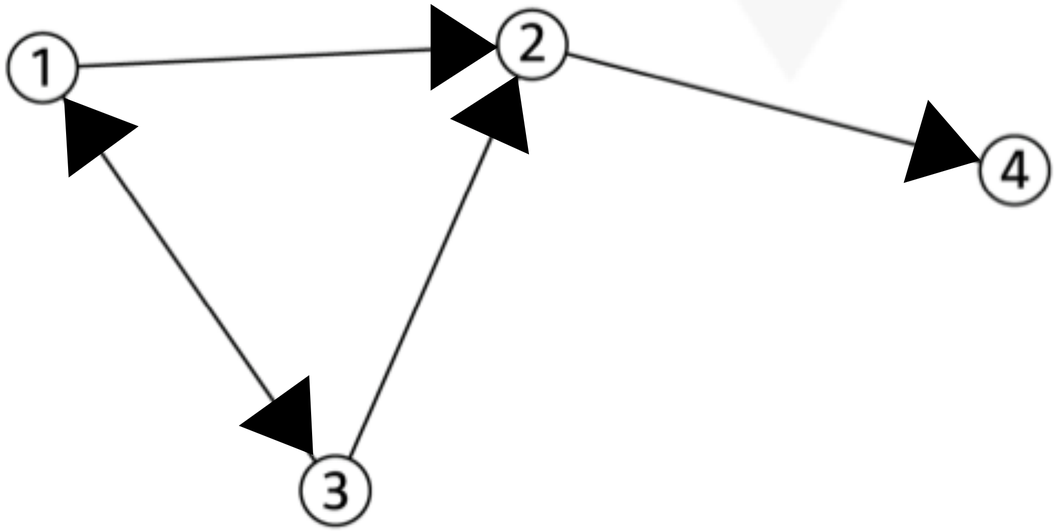
Networks have a huge economic impact



Data is the new oil

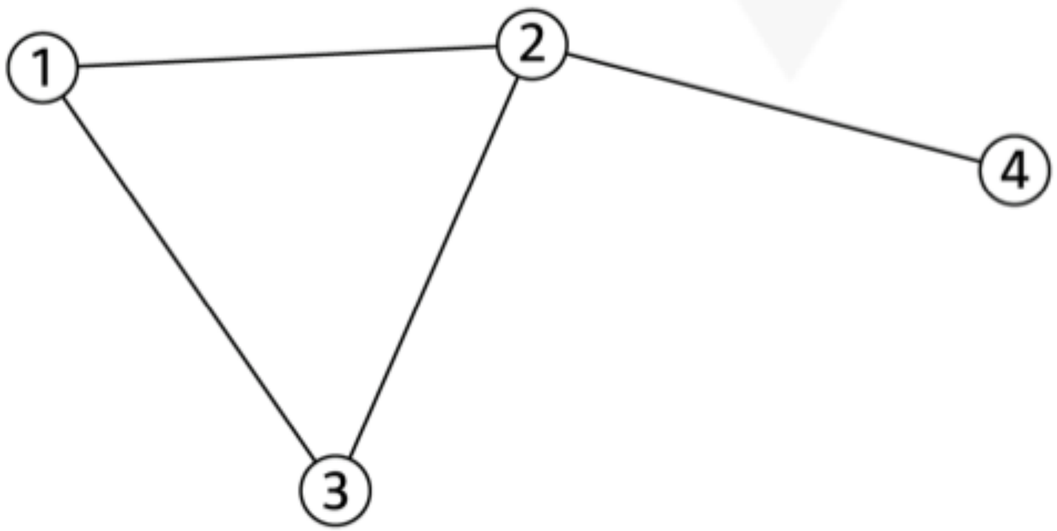
A directed graph (digraph) has links with a direction

Nodes N=4
Links (Arcs) L=5



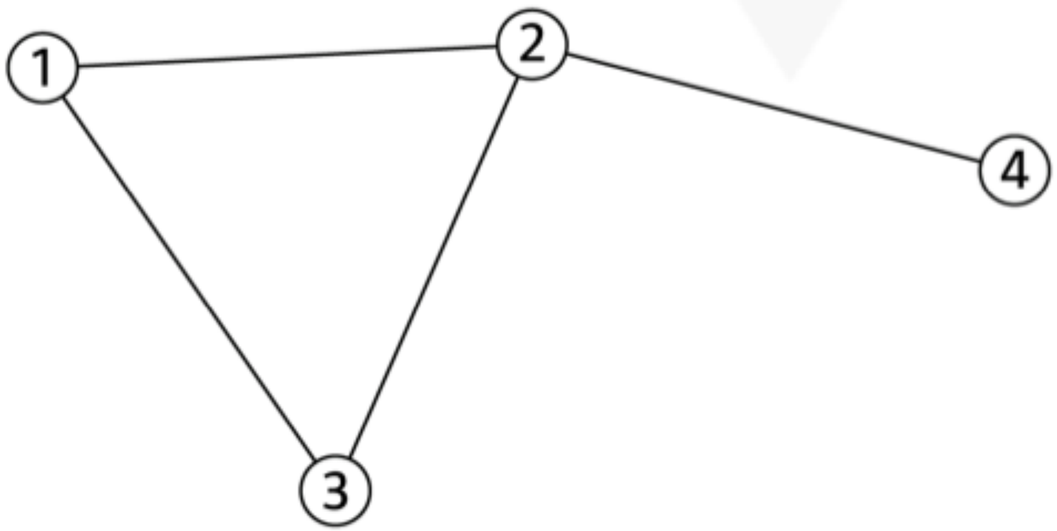
The degree k_i of a node i is the number of incident links

$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$



Every network has an average degree $\langle k \rangle$

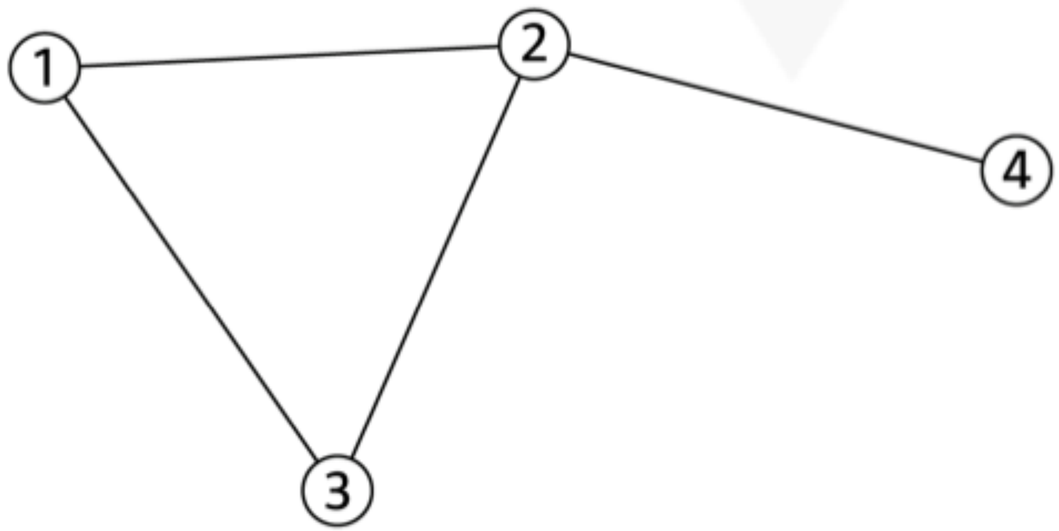
$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$

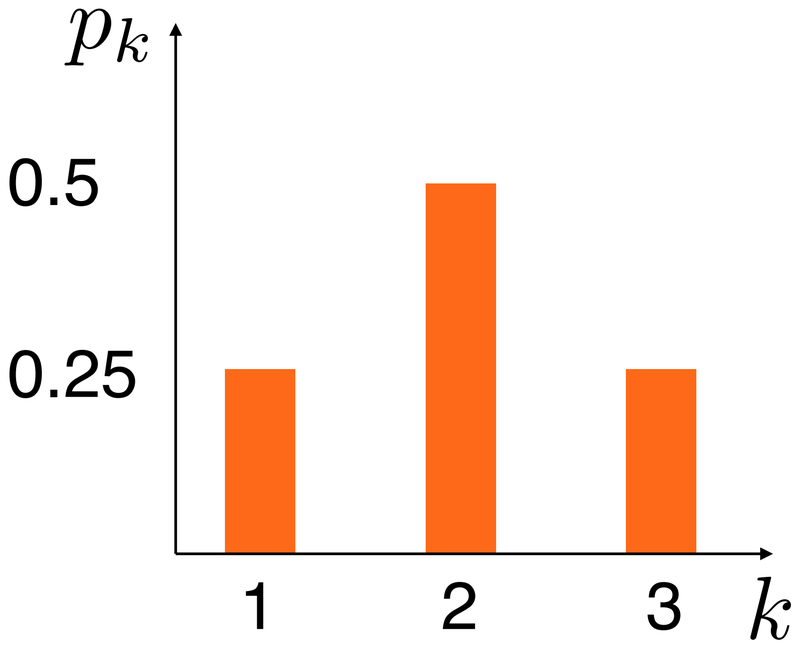


$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \qquad \langle k \rangle = \frac{2+3+2+1}{4} = 2$$

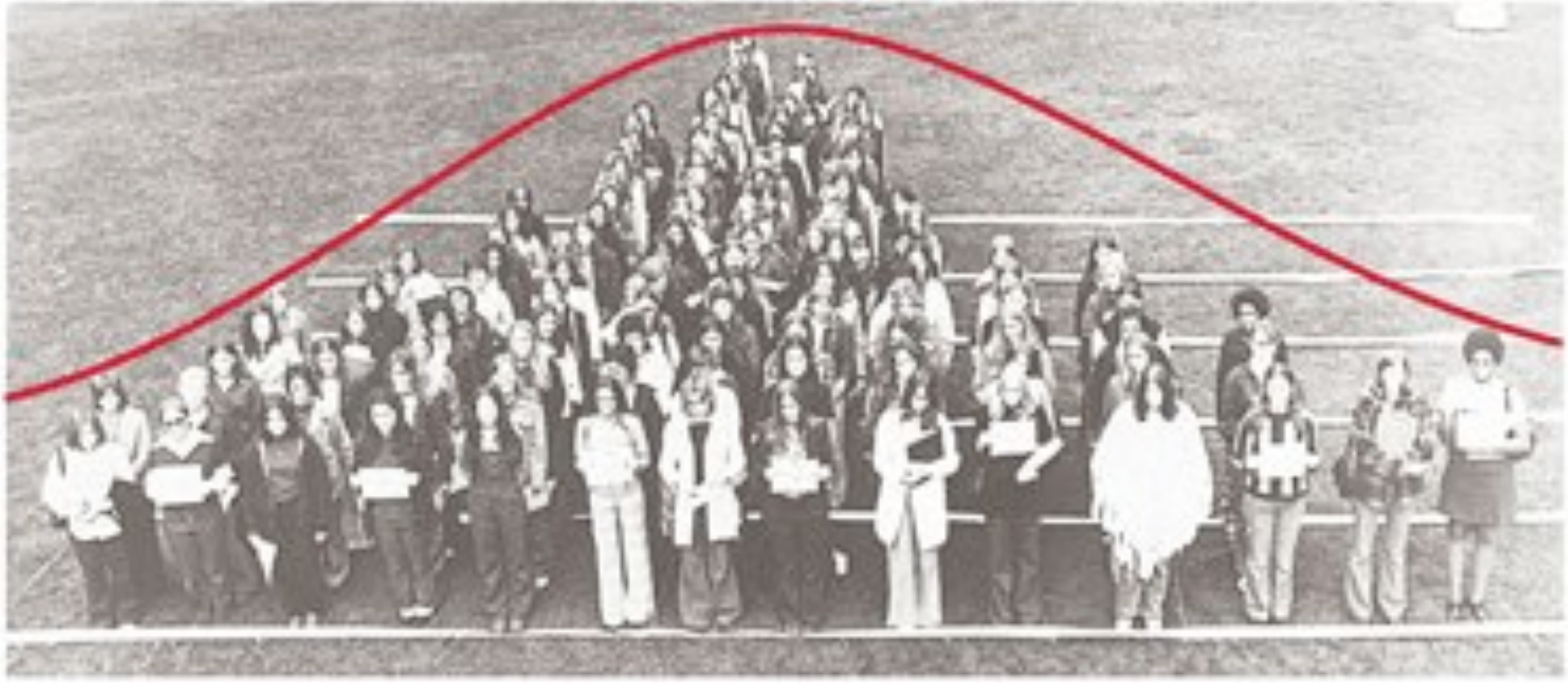
The degree distribution $p_k = N_k/N$ captures the probabilities that a node has a certain degree

$$k_1 = 2 \qquad k_3 = 2$$
$$k_2 = 3 \qquad k_4 = 1$$



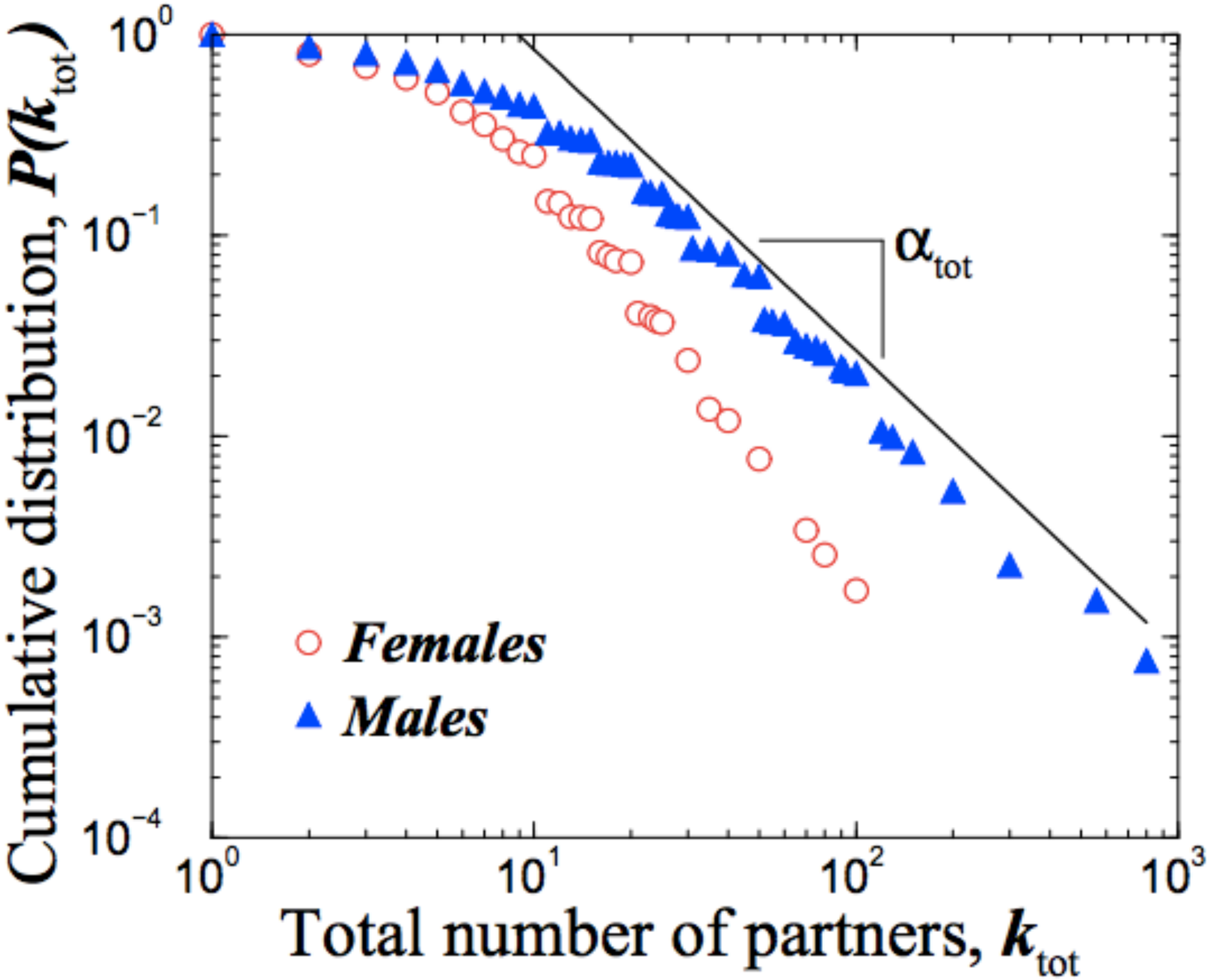


A network's degree distribution tells us something fundamental about how individuals in the system connect

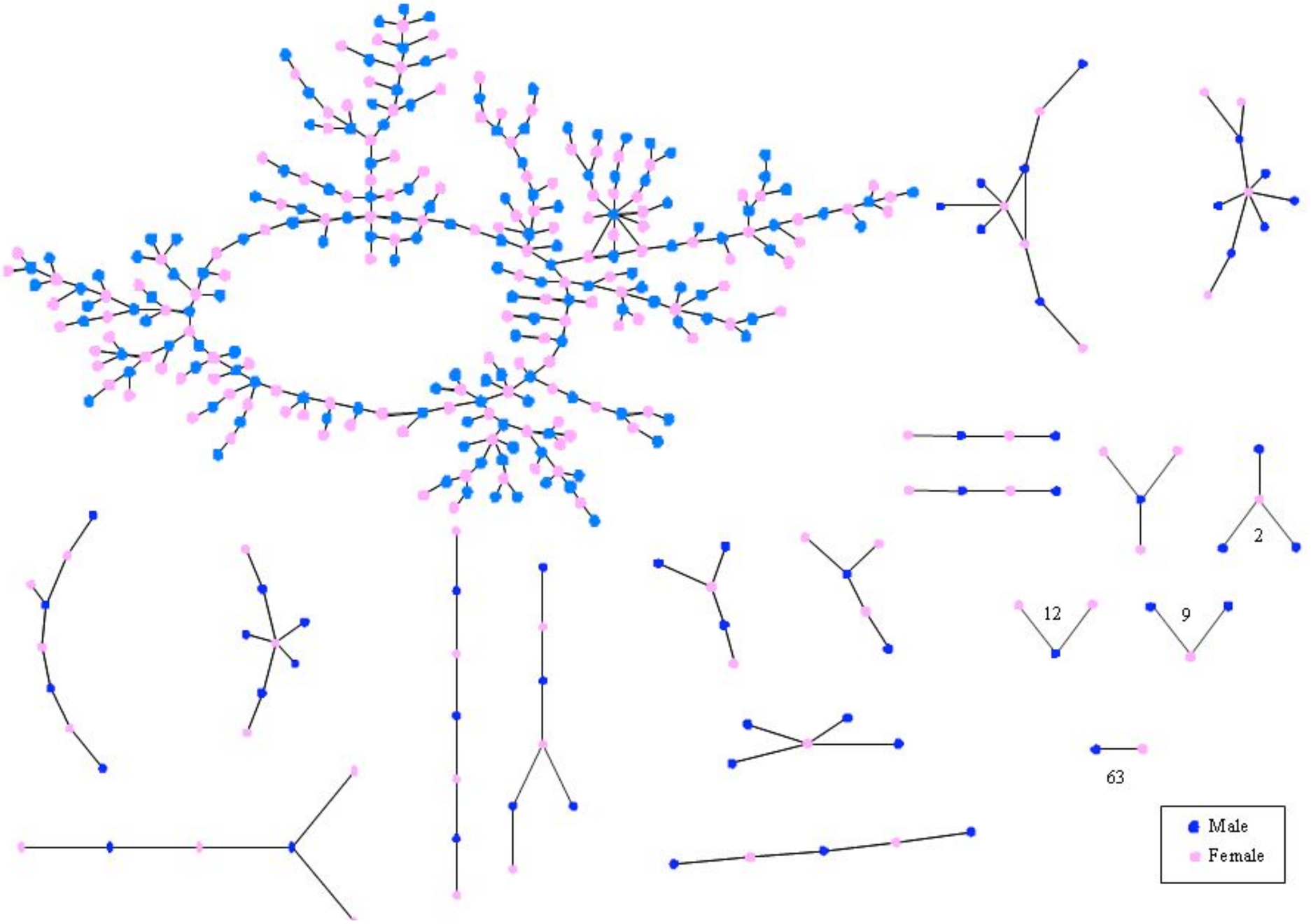


23/10 /2 * (62-15)*365 ~ **20000**

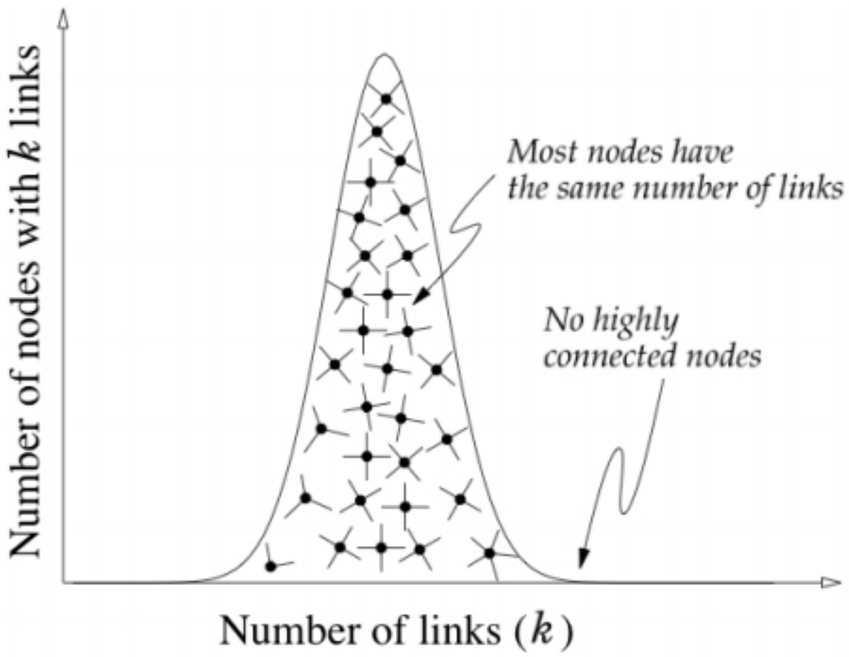
Sexual behavior is heavy-tailed

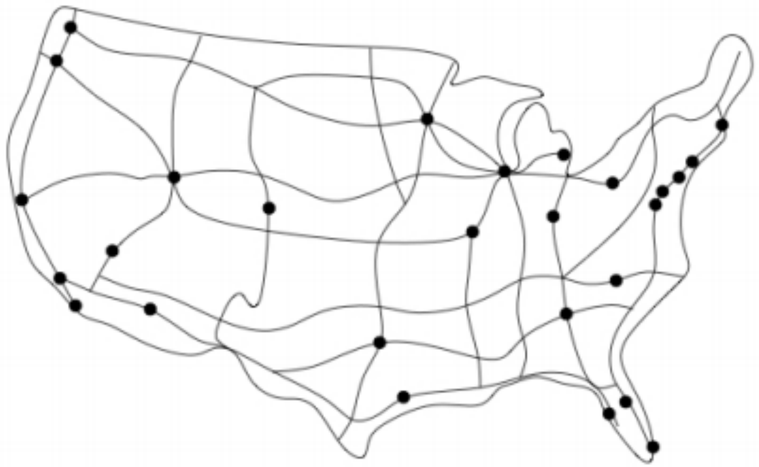


If our heights were like our sex lifes


2
12
9
63
Male
Female

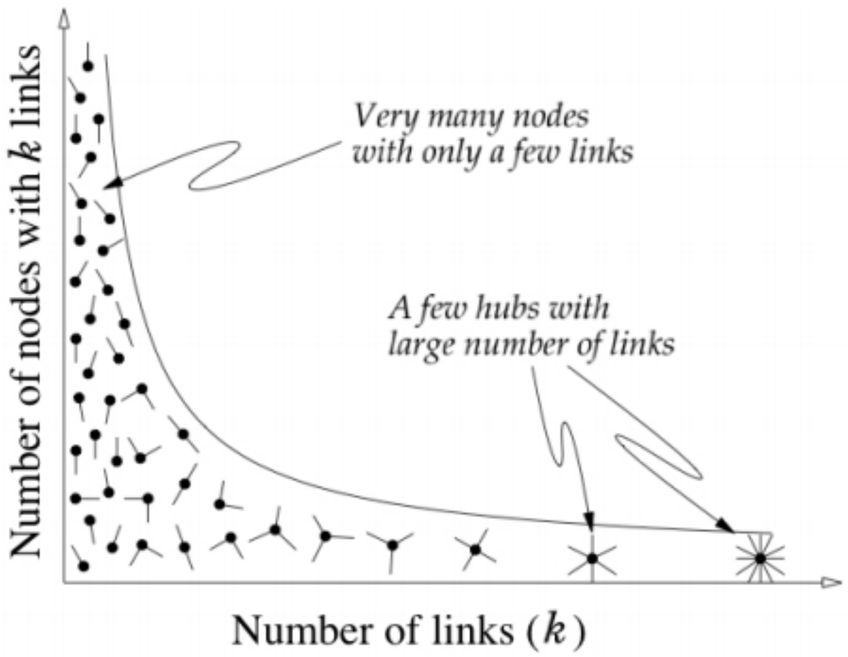
Thin-tailed network

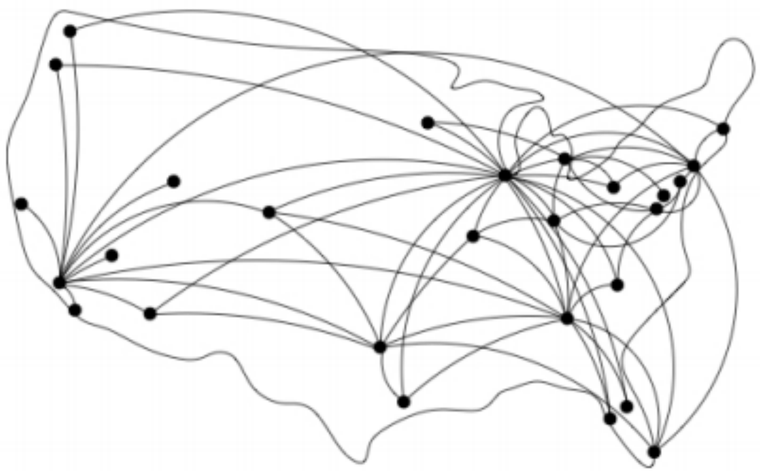




Streets

Heavy-tailed network

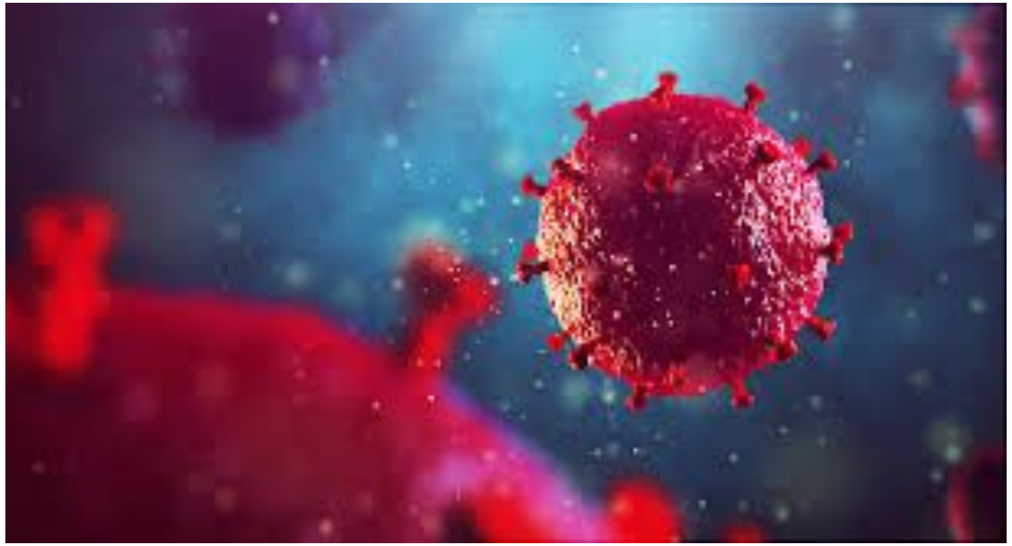




Airlines

A heavy-tailed degree distribution means we have hubs

Hubs completely change network processes
like epidemic spreading



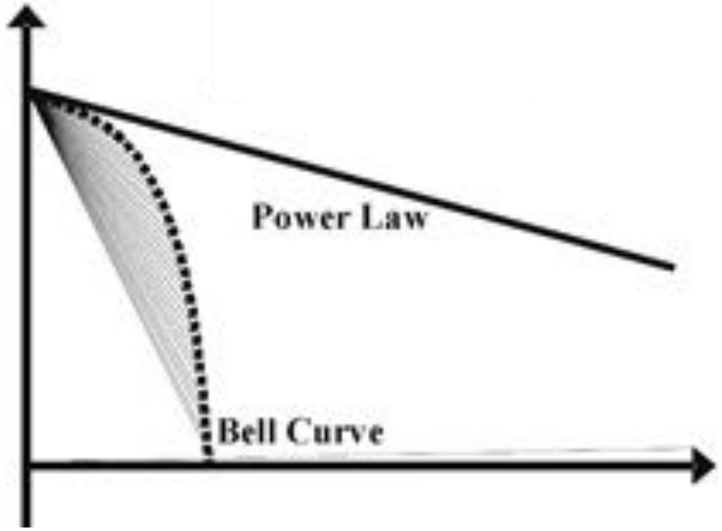
Heavy-tailed distributions govern the world

20th century statistics

Strange outliers

Focus on the head/center

21th century statistics add:



Not outliers but part of the system

Focus on the tail

Organizing principles of networks

Many networks are:

1) Heavy-tailed

2) Sparse

3) Small-world

4) Clustered

Organizing principles of networks

Many networks are:

1) Heavy-tailed

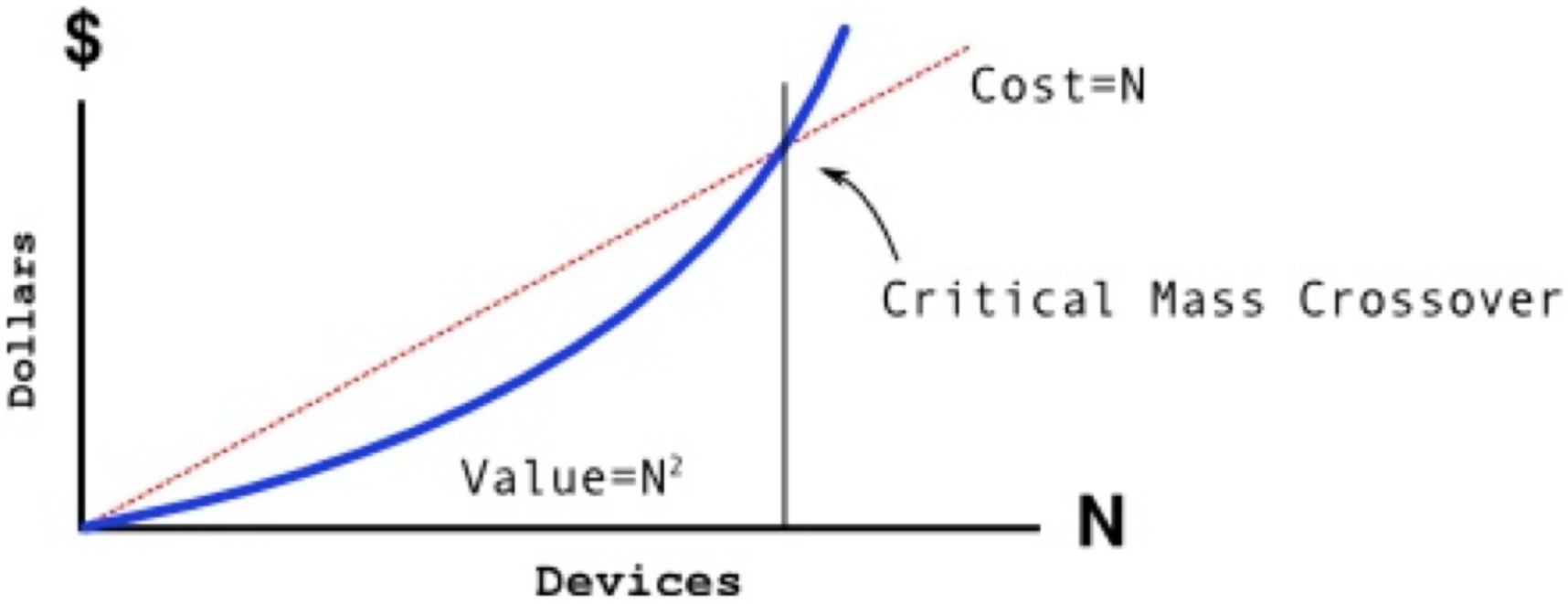
2) Sparse

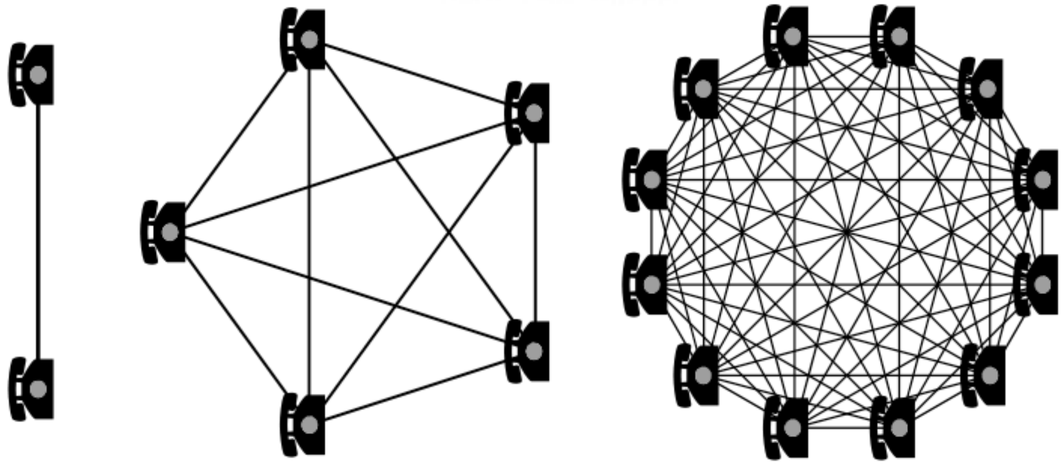
3) Small-world

4) Clustered

The person who bought the first fax, what was he/she thinking???

Metcalfe's law states that the value of a communication network increases with the square of users





$$L_{\max} = \frac{N(N-1)}{2} \sim N^2$$

Sparsity means: Although a lot of links are possible, only very few are actually there: $L \ll L_{\max}$

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

Sparsity means: Although a lot of links are possible, only very few are actually there: $L \ll L_{\max}$

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

Also: $\langle k \rangle \ll \langle k \rangle_{\max} = N - 1$

Organizing principles of networks

Many networks are:

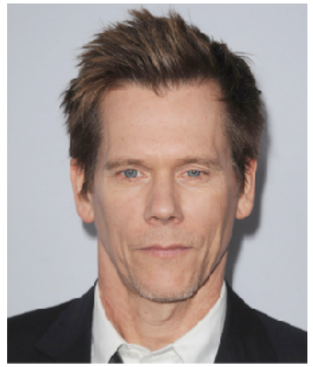
1) Heavy-tailed

2) Sparse

3) Small-world

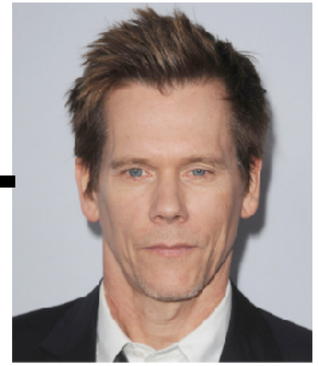
4) Clustered

The Kevin Bacon game



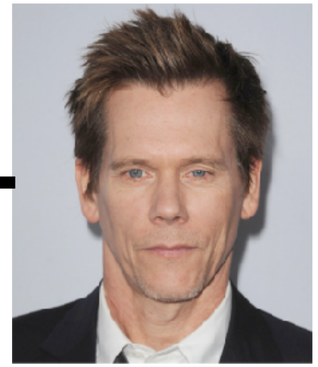
The Kevin Bacon game

A few good men



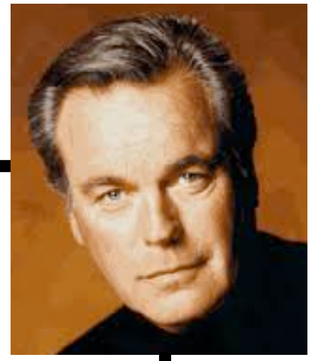
The Kevin Bacon game

A few good men

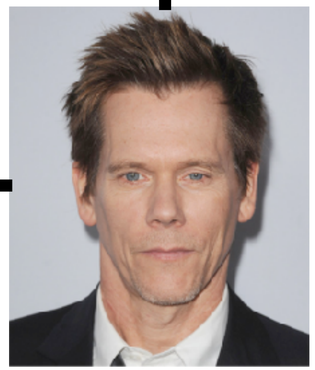


The Kevin Bacon game

Austin Powers

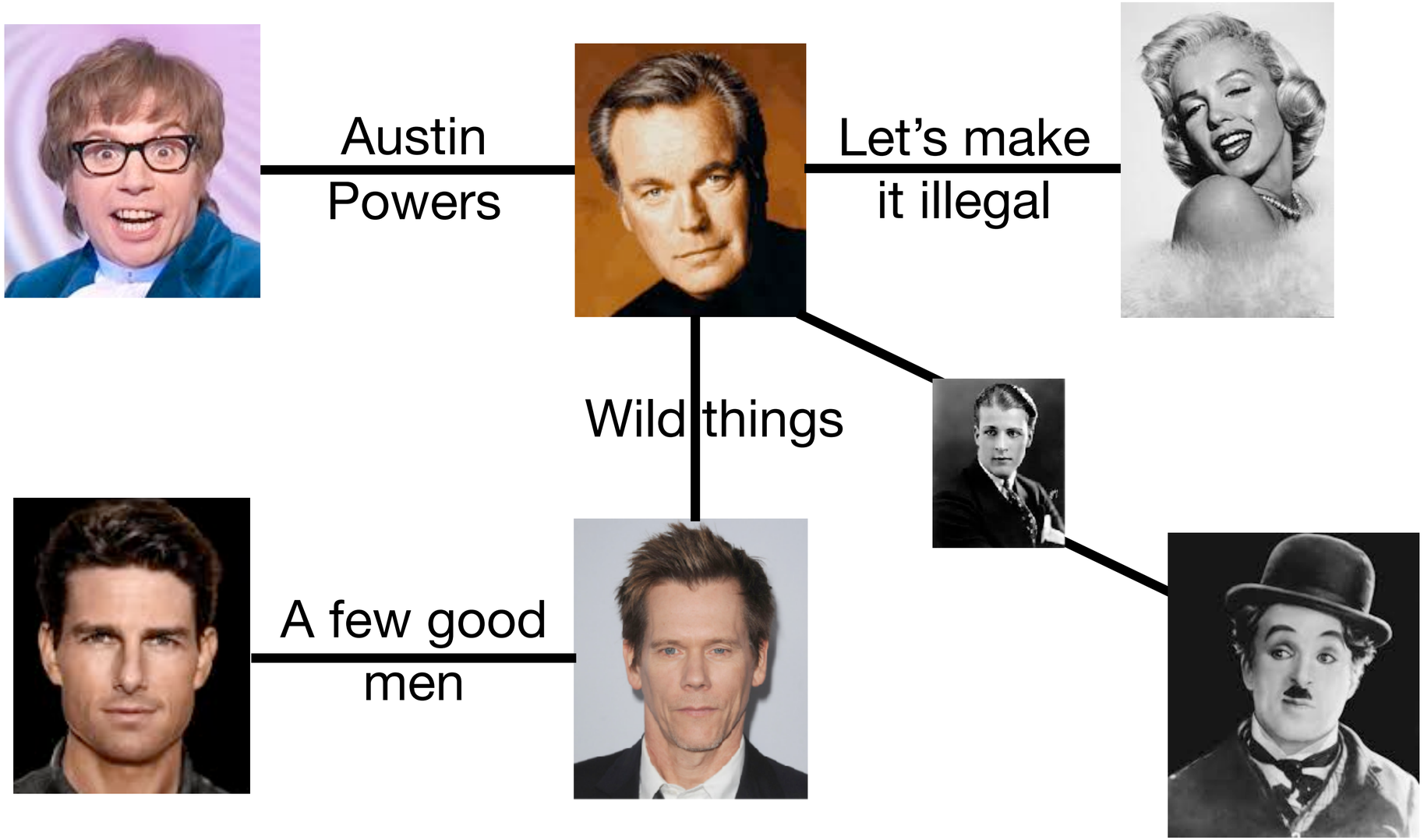


Wild things

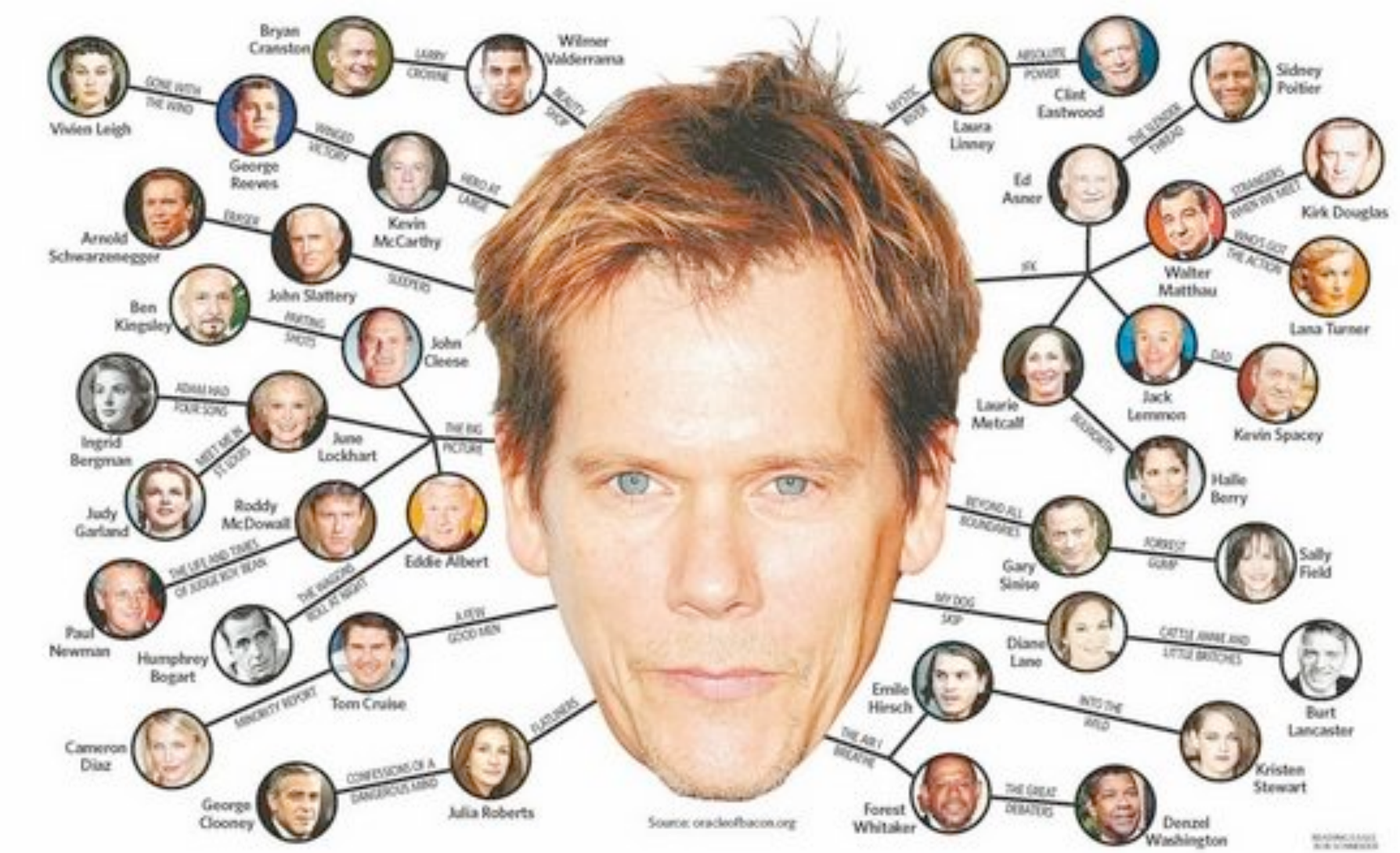


A few good men

The Kevin Bacon game



Six degrees of Kevin Bacon: Everyone is connected with Kevin Bacon, although he is not the most famous actor



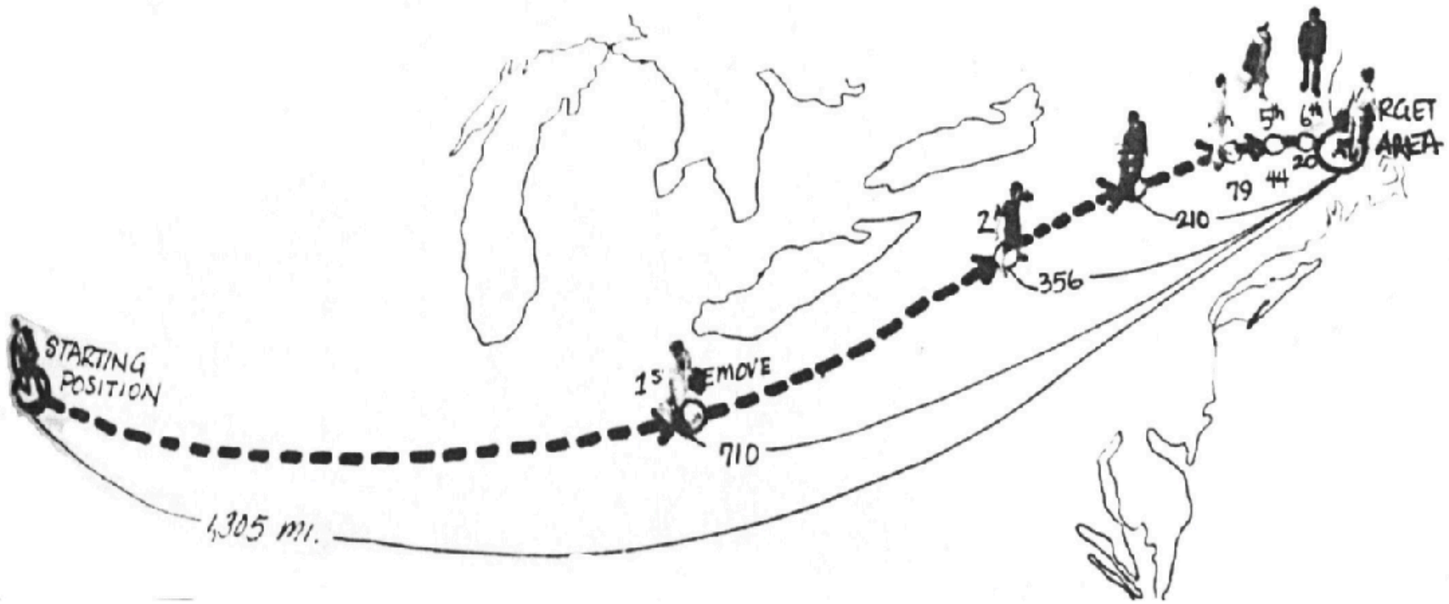
Six degrees of Erdős

MR Erdos Number = 5

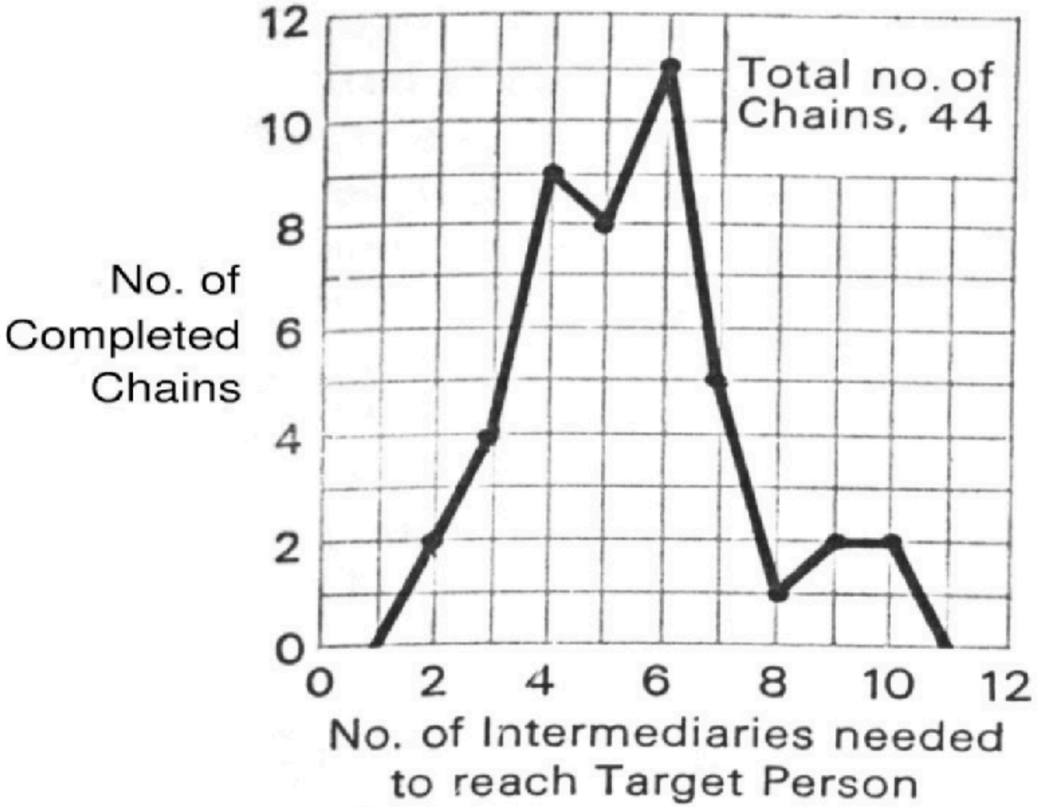
Michael Szell	coauthored with	Stefan Thurner[1]	MR2969349
Stefan Thurner[1]	coauthored with	Murray Gell-Mann	MR2793803
Murray Gell-Mann	coauthored with	Sheldon Lee Glashow	MR0134224
Sheldon Lee Glashow	coauthored with	Daniel J. Kleitman	MR0172630
Daniel J. Kleitman	coauthored with	Paul Erdős[1]	MR0228375

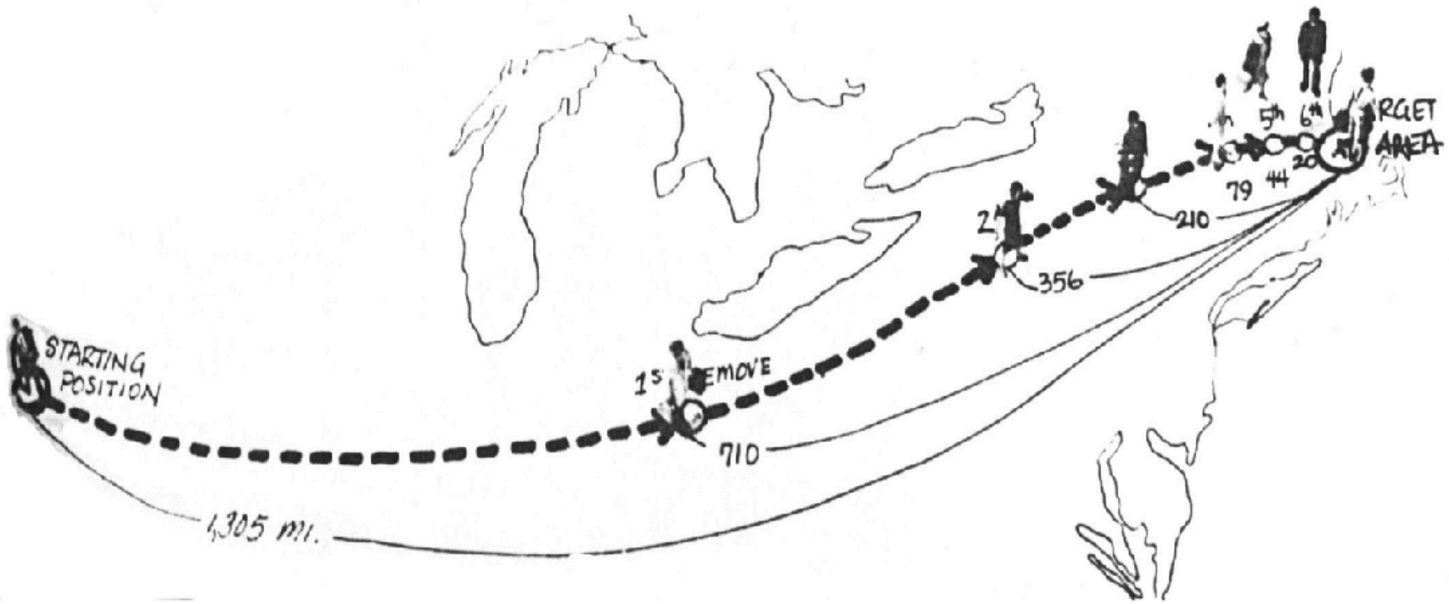
http://www.ams.org/mathscinet/collaborationDistance.html

Stanley Milgram, 1967: Deliver a letter to somebody you don't know

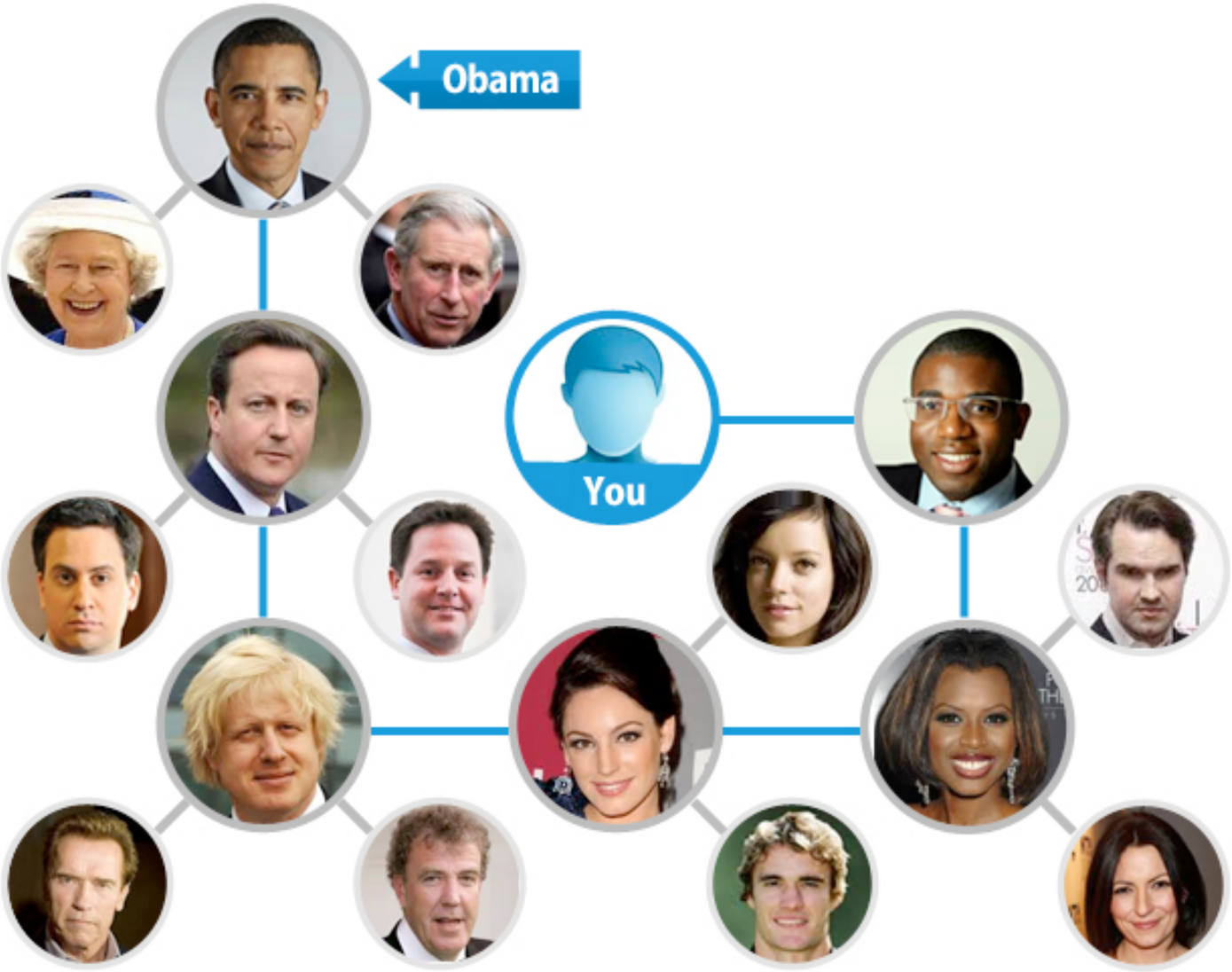


The letters which reached their destination had a median degree of separation of 6: The world is small





Six degrees of separation means you are connected to almost any person in the world through 6 steps



Six degrees also applies to knowledge (Wikipedia)

Toothpaste → William Shakespeare

Six degrees also applies to knowledge (Wikipedia)

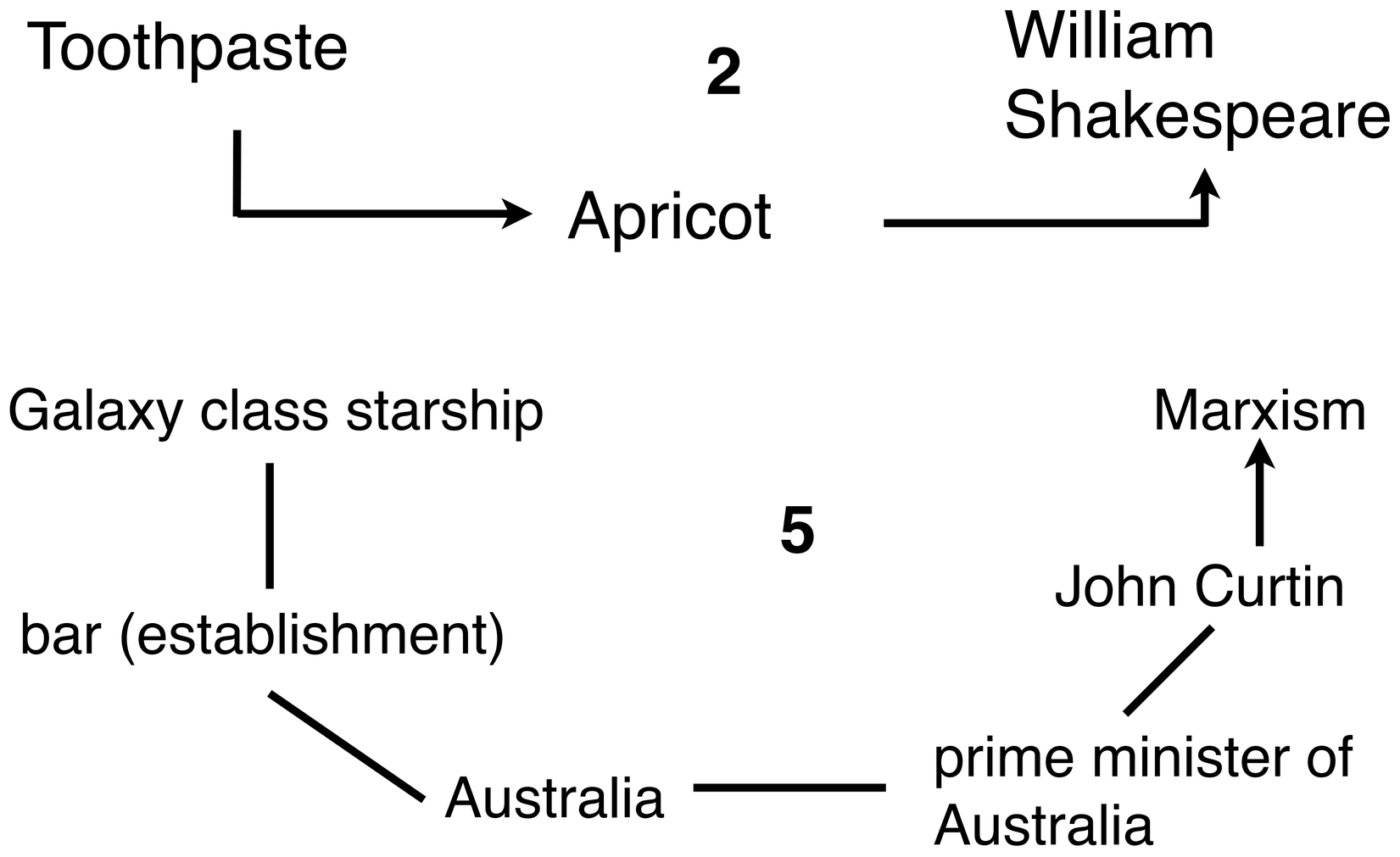
Six degrees also applies to knowledge (Wikipedia)

Toothpaste → Apricot → William Shakespeare

2

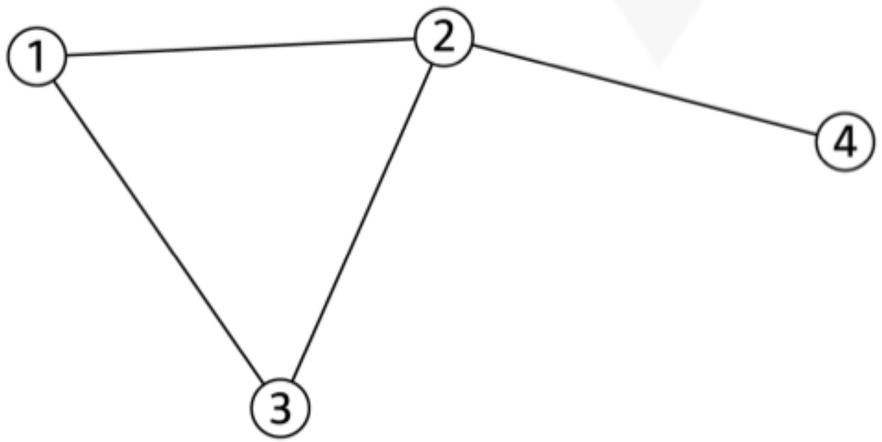
Galaxy class starship → Marxism

Six degrees also applies to knowledge (Wikipedia)



To formalize this idea, we need some definitions:

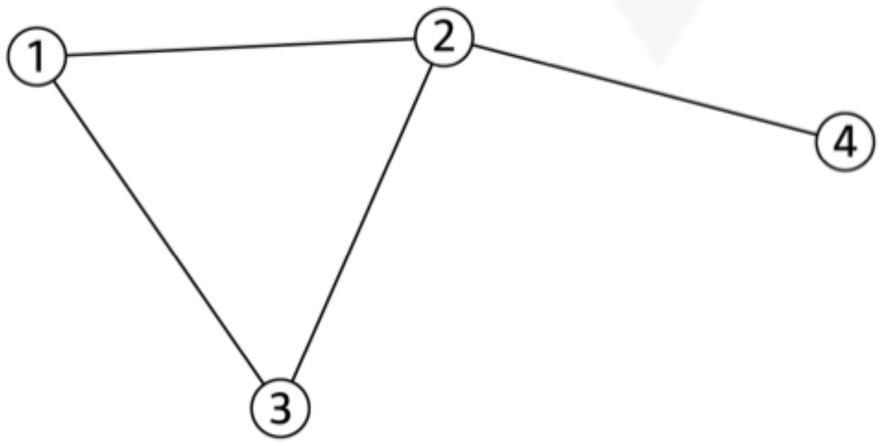
Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$



To formalize this idea, we need some definitions:

Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$

Path: A walk where no node is repeated $\{n_1, n_3, n_2, n_4\}$

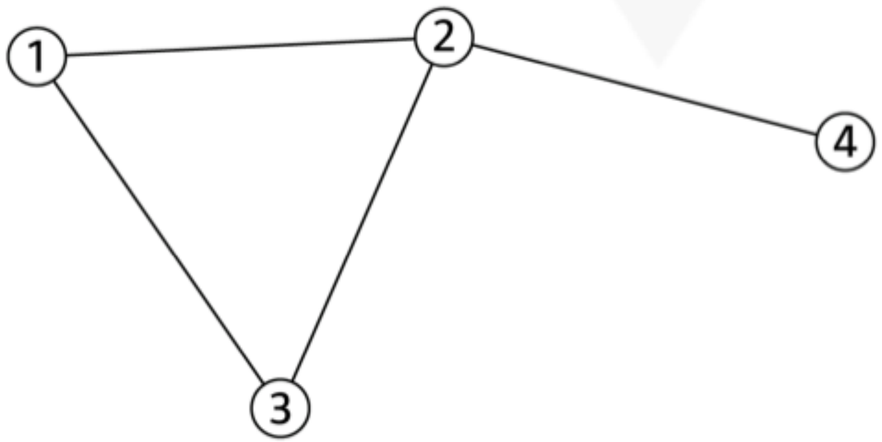


To formalize this idea, we need some definitions:

Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$

Path: A walk where no node is repeated $\{n_1, n_3, n_2, n_4\}$

Shortest path: A path of minimal length $\{n_1, n_2, n_4\}$



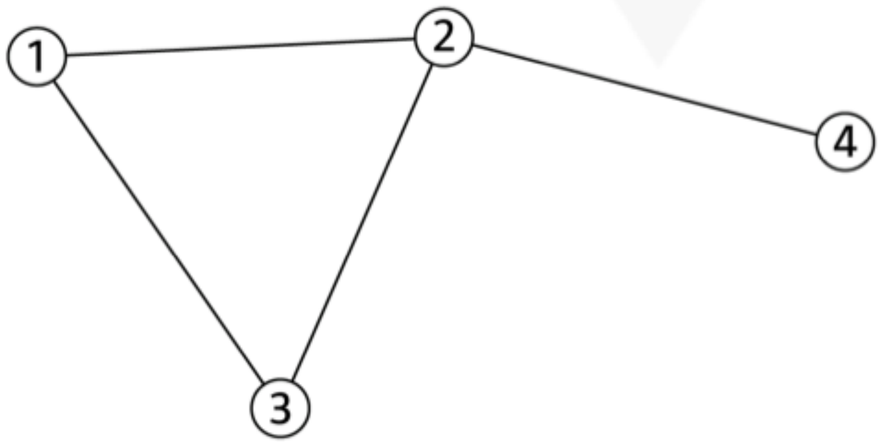
To formalize this idea, we need some definitions:

Walk: A sequence of neighboring nodes $\{n_1, n_2, n_1, n_3, n_2, n_4\}$

Path: A walk where no node is repeated $\{n_1, n_3, n_2, n_4\}$

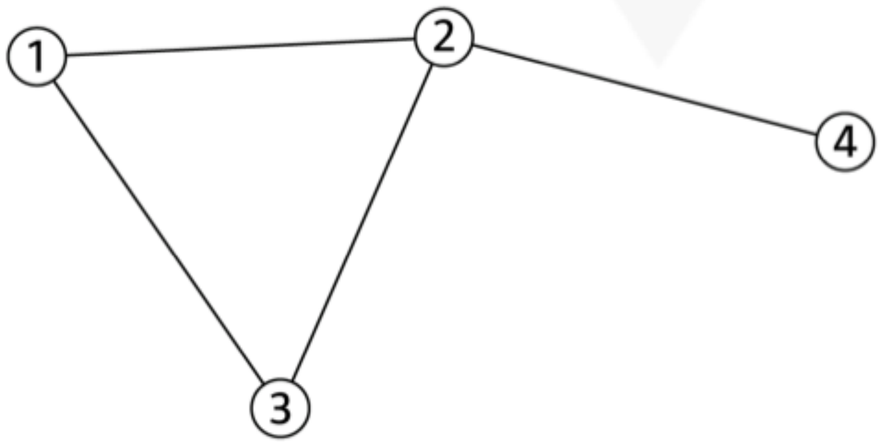
Shortest path: A path of minimal length $\{n_1, n_2, n_4\}$

Graph distance: Length of shortest path $d_{1,4} = 2$



The average path length ℓ is the mean graph distance over all pairs of nodes

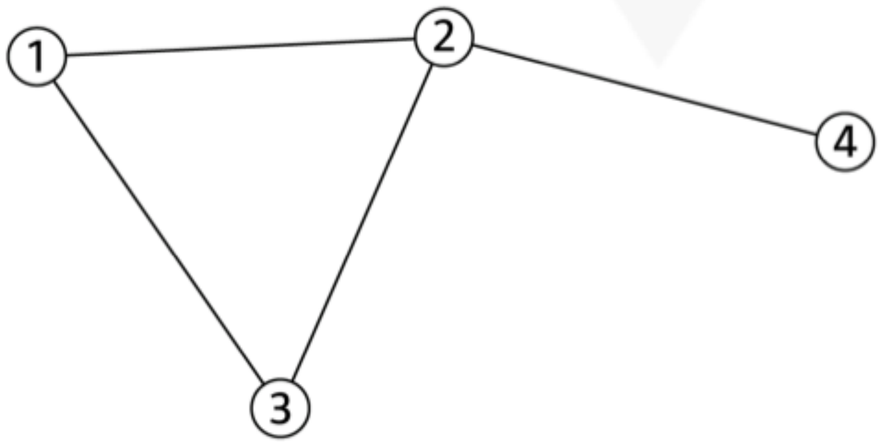
$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$



The average path length ℓ is the mean graph distance over all pairs of nodes

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

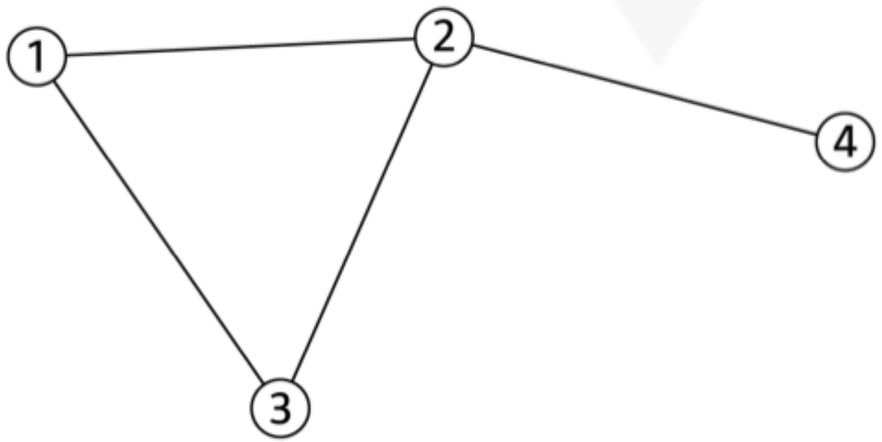
$$\ell = \frac{4+3+4+5}{4(4-1)} = \frac{4}{3}$$



The diameter D is the maximum length of shortest paths

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

$$\ell = \frac{4+3+4+5}{4(4-1)} = \frac{4}{3}$$

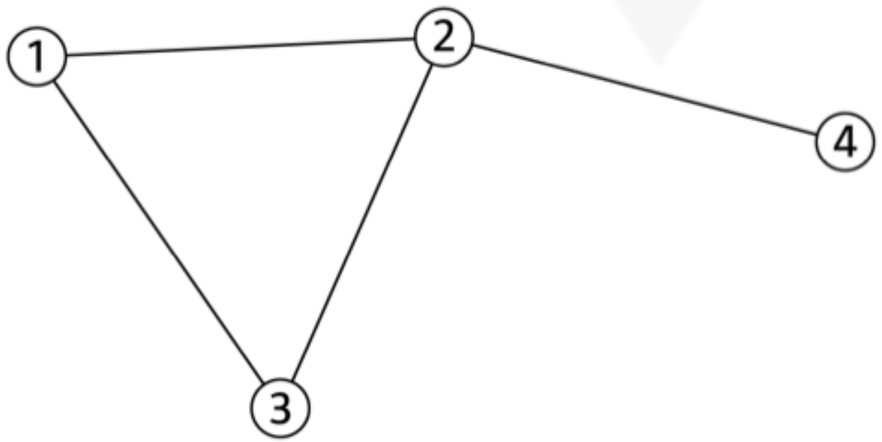


The diameter D is the maximum length of shortest paths

$$\ell = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{i,j}$$

$$\ell = \frac{4+3+4+5}{4(4-1)} = \frac{4}{3}$$

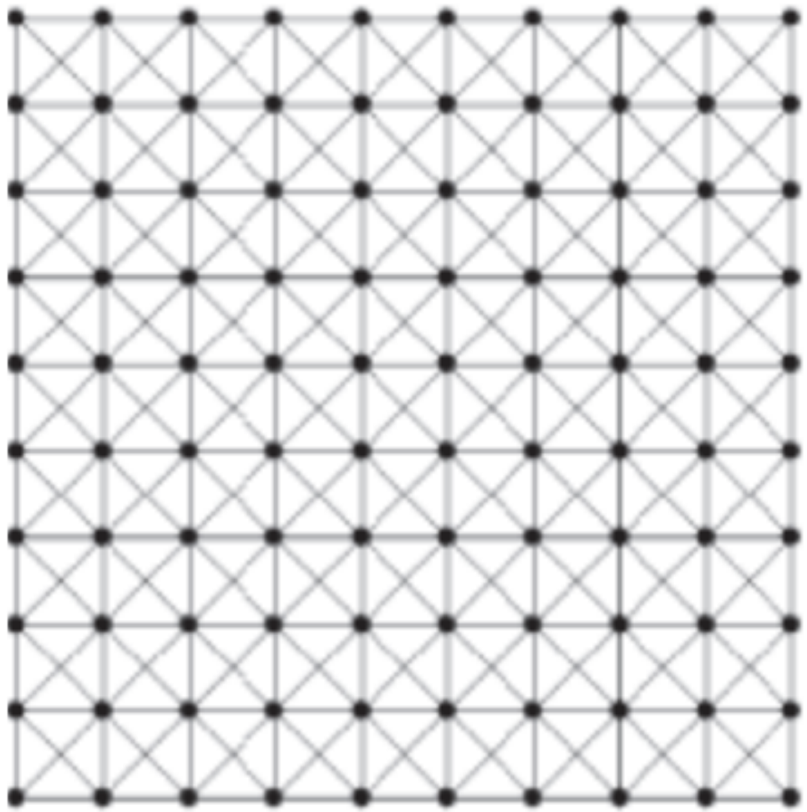
$$D = 2$$



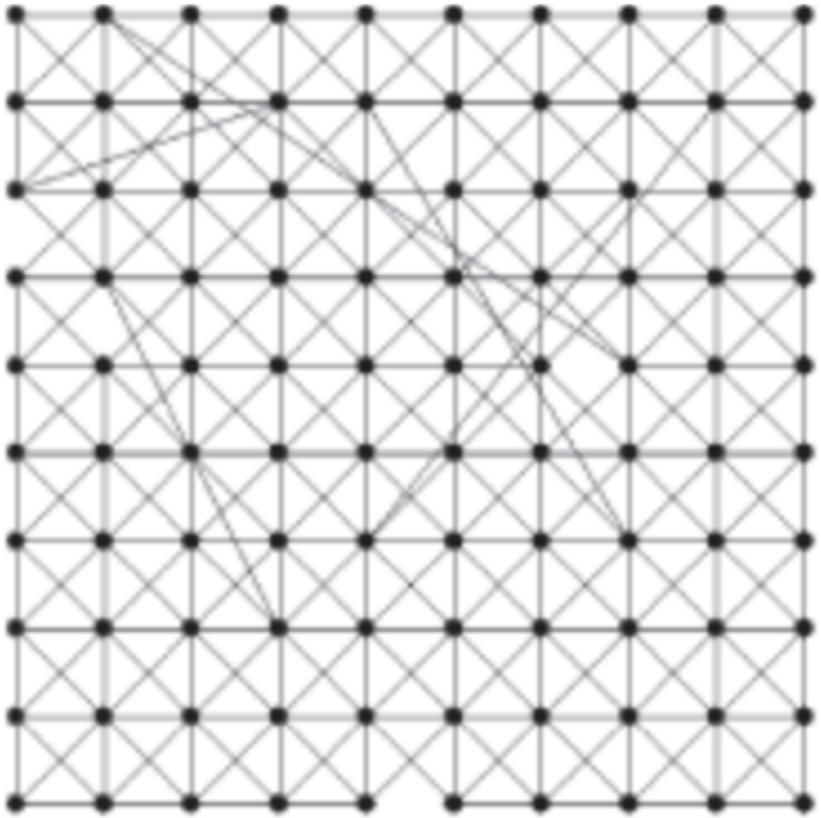
A small world network has short average path length

$$\ell \sim N$$

$$\ell \sim \log N$$



Regular network ϕ=0



Small-world network ϕ=0.01

Organizing principles of networks

Many networks are:

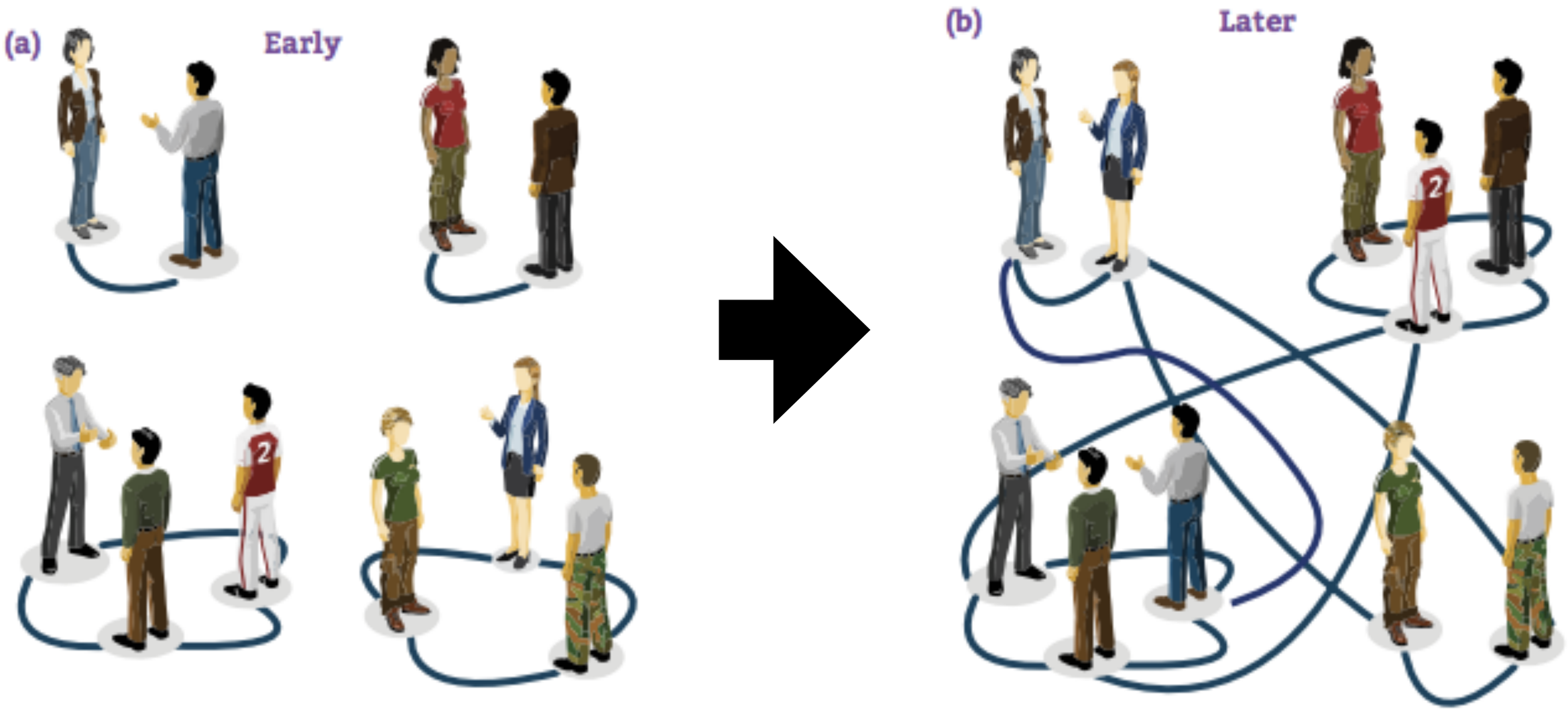
1) Heavy-tailed

2) Sparse

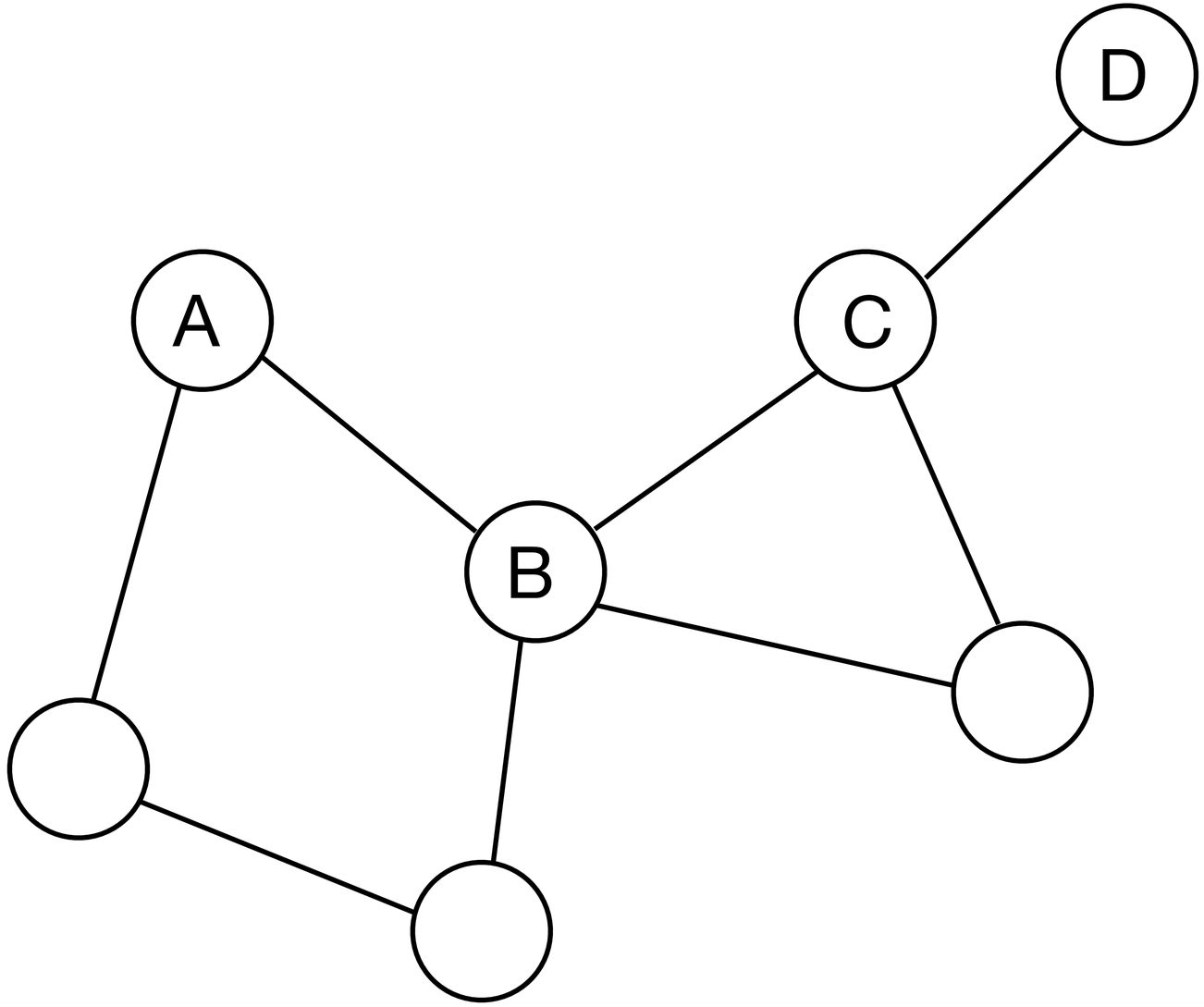
3) Small-world

4) Clustered

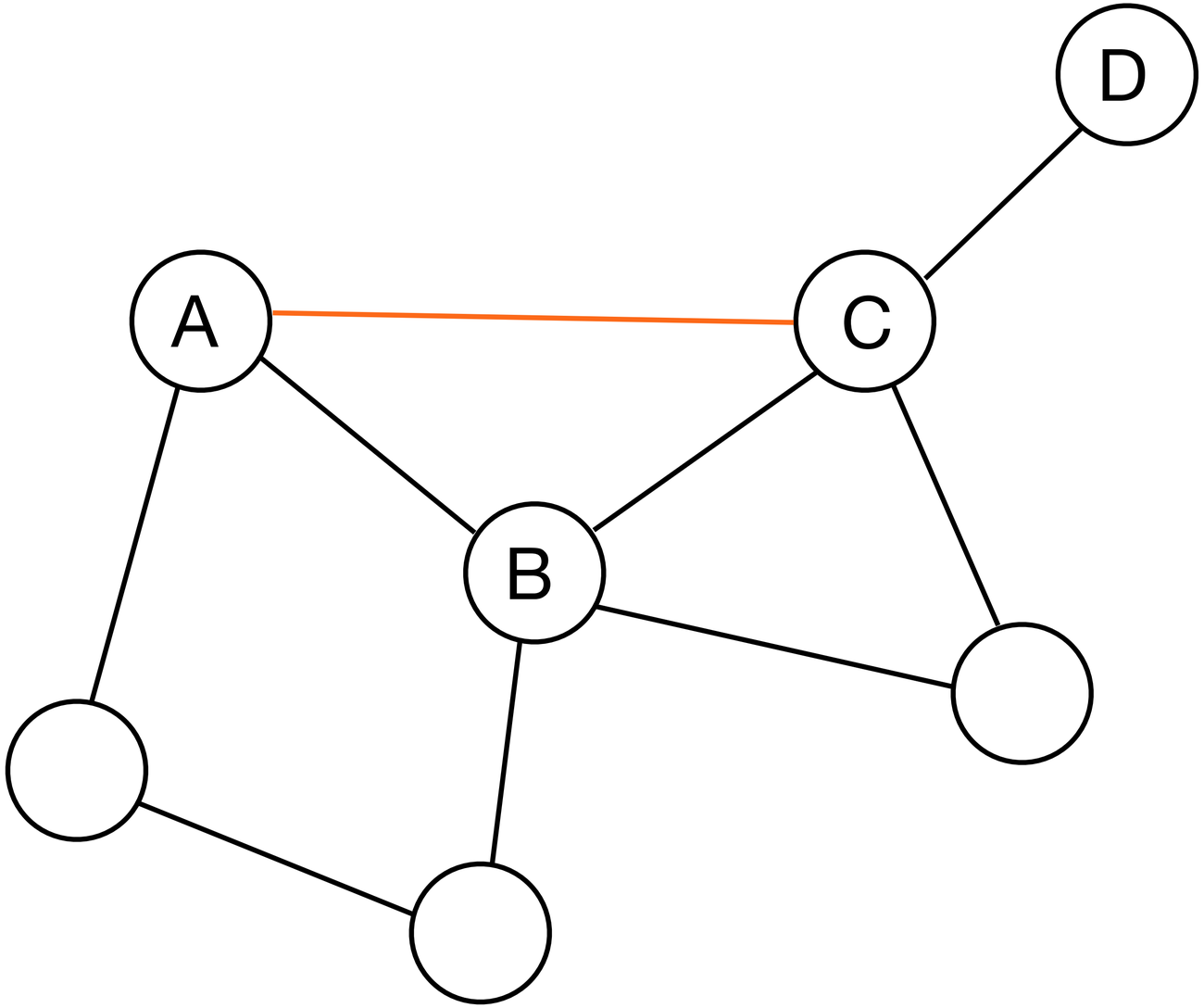
In a cocktail party you introduce each other to new people



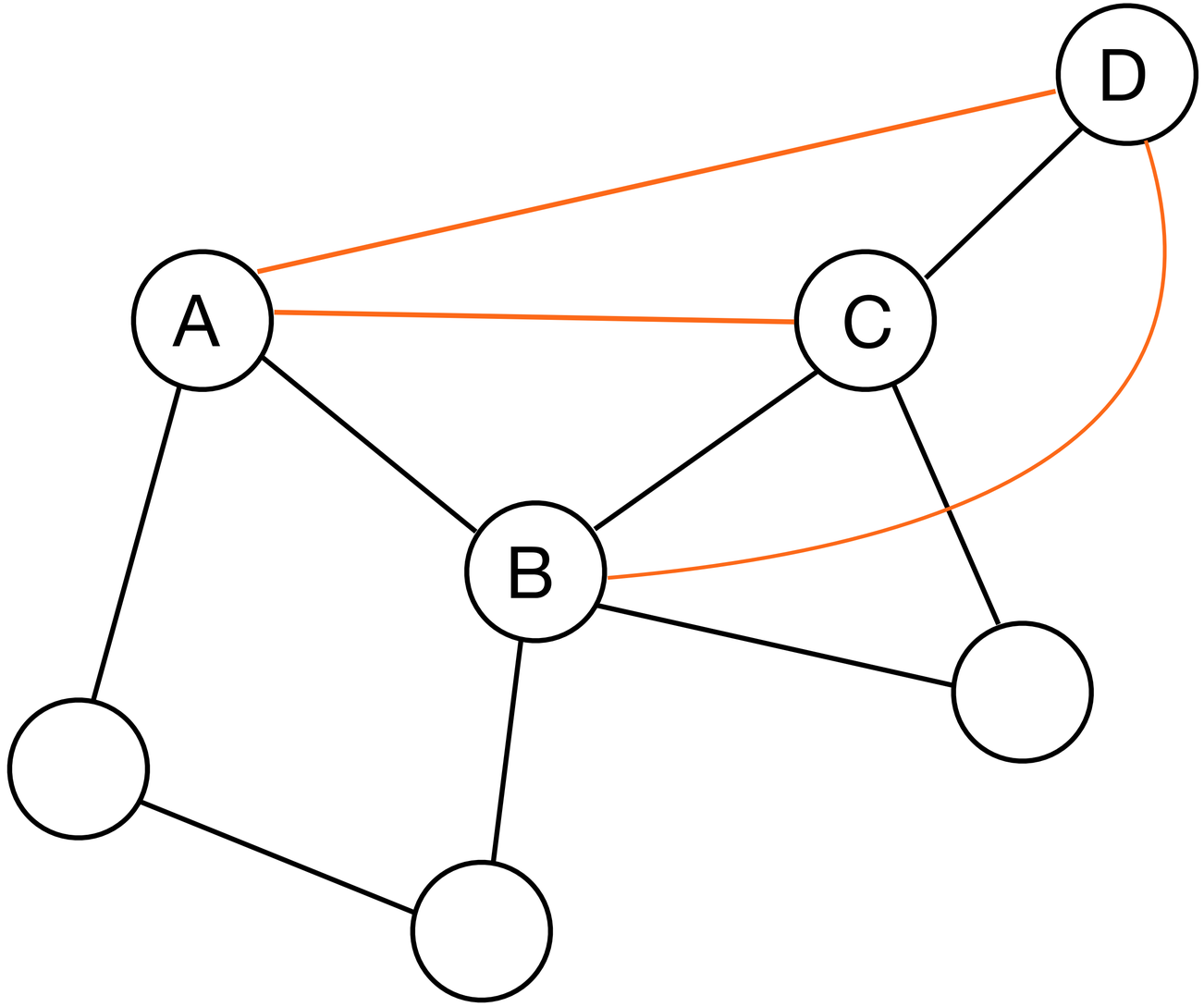
Who are more likely to connect: A—C or A—D?



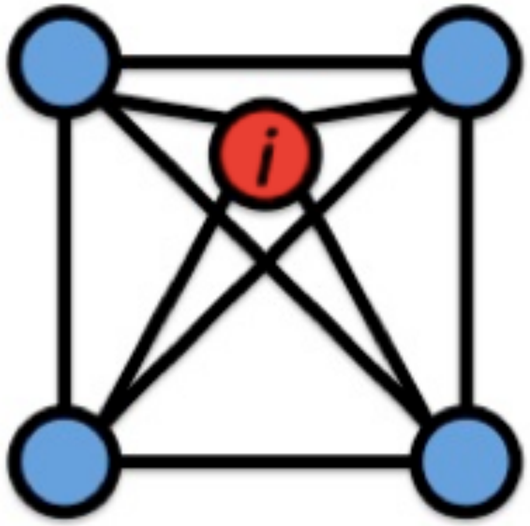
A and C already have a common friend, B



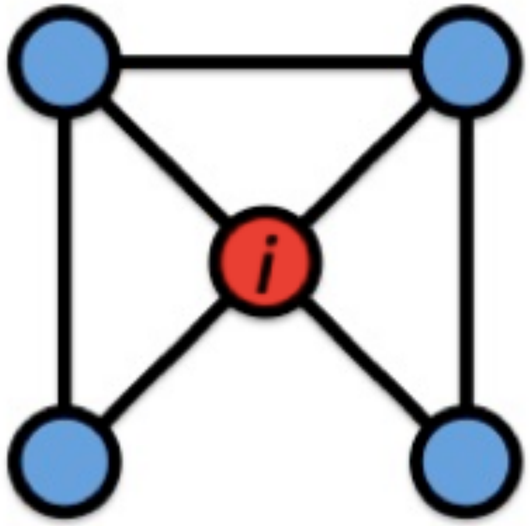
Closing open triangles is called triadic closure



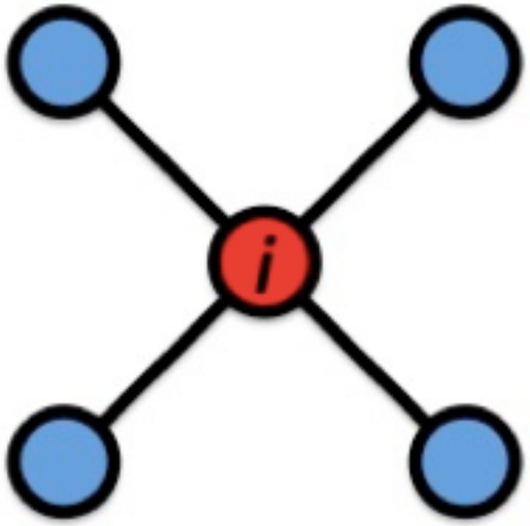
The clustering coefficient measures the fraction of your neighbor pairs who are linked



$$c_i = 1$$



$$c_i = 1/2$$



$$c_i = 0$$

By definition, $0 \leq c_i \leq 1$

Organizing principles of networks

Many networks are:

1) Heavy-tailed

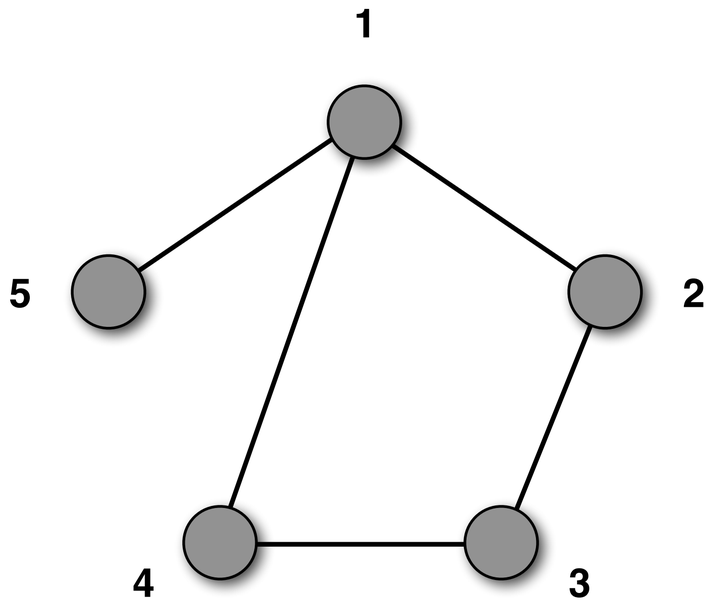
2) Sparse

3) Small-world

4) Clustered

Network data structures

The adjacency matrix stores all possible connections
Realized connections get a 1

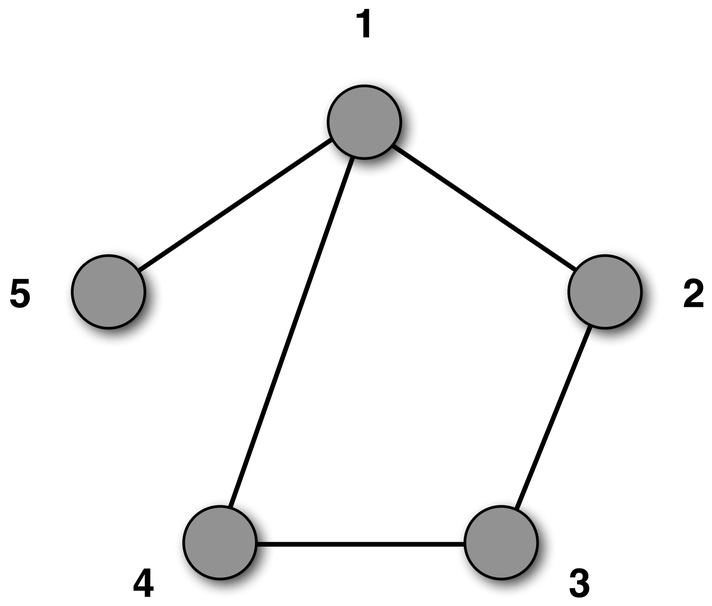


$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ A_{N1} & \dots & \dots & A_{NN} \end{pmatrix}$$

$A_{ij} = A_{ji} = 1$ If there is a link between node *i* and node *j*

$A_{ij} = A_{ji} = 0$ If node *i* and node *j* are not connected

The adjacency matrix stores all possible connections
Realized connections get a 1



$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$A_{ij} = A_{ji} = 1$ If there is a link between node *i* and node *j*

$A_{ij} = A_{ji} = 0$ If node *i* and node *j* are not connected

The adjacency matrix is good for dense networks, but not practical for real networks that are large and sparse

Easy to use analytical formulas

Easy to find/remove/add a link: $\mathcal{O}(1)$

Useful for dense networks

Needs a lot of memory: $\mathcal{O}(N^2)$

Inconvenient for many numerical calculations

The adjacency matrix is not practical for large networks

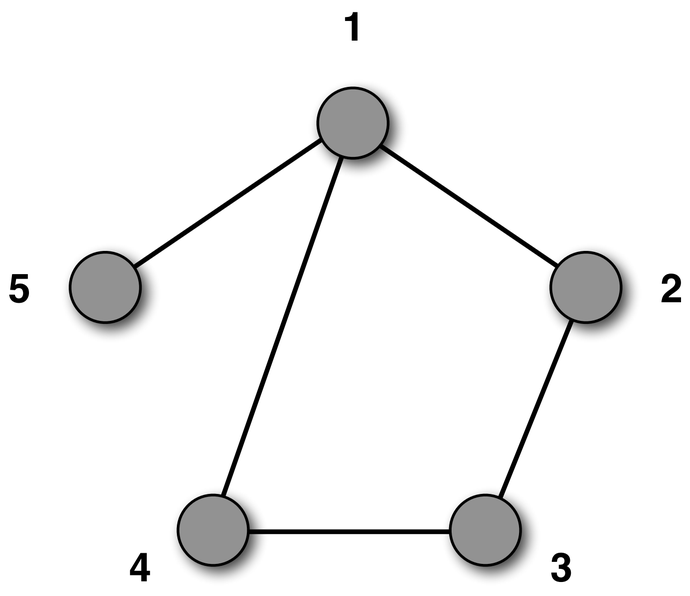
Needs a lot of memory: $\mathcal{O}(N^2)$

A matrix takes $4N^2$ bytes.
Assuming 10 GB = 10^{10} bytes of RAM, the largest possible network satisfies $4N^2 = 10^{10}$, which means $N = 50000$.

The edge list stores the node IDs of connected links

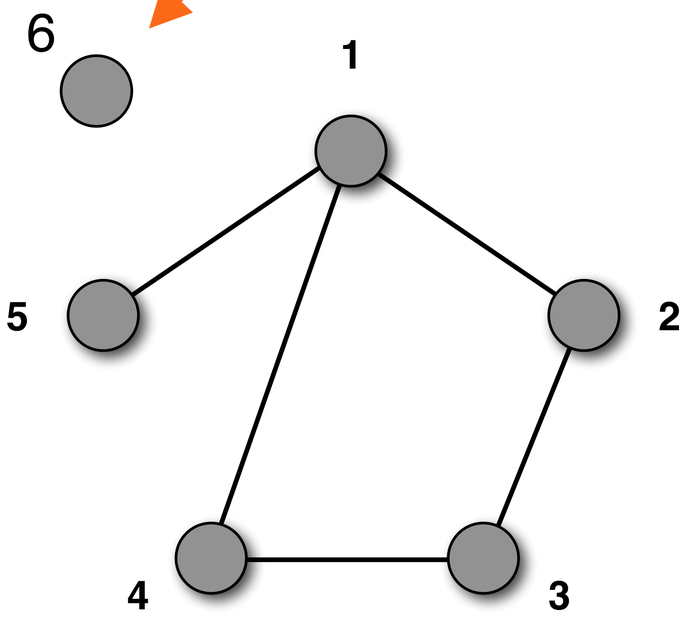
$$I = \begin{pmatrix} \cdots \\ i \\ \cdots \end{pmatrix} \quad J = \begin{pmatrix} \cdots \\ j \\ \cdots \end{pmatrix}$$

Only if $A_{ij} = 1$



$$I = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 3 \end{pmatrix} \quad J = \begin{pmatrix} 2 \\ 4 \\ 5 \\ 3 \\ 4 \end{pmatrix}$$

The edge list is missing isolated nodes



$$I = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 3 \end{pmatrix} \quad J = \begin{pmatrix} 2 \\ 4 \\ 5 \\ 3 \\ 4 \end{pmatrix}$$

The edge list needs less space, but has disadvantages for calculations

Needs less memory: $2L$

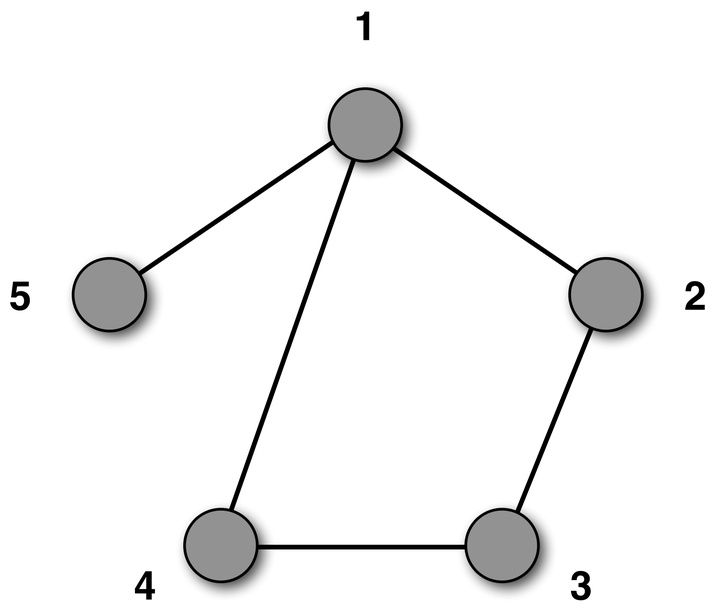
Not fast to find edges: $\mathcal{O}(L)$

Convenient for data collection

Inconvenient for calculations involving node neighbors

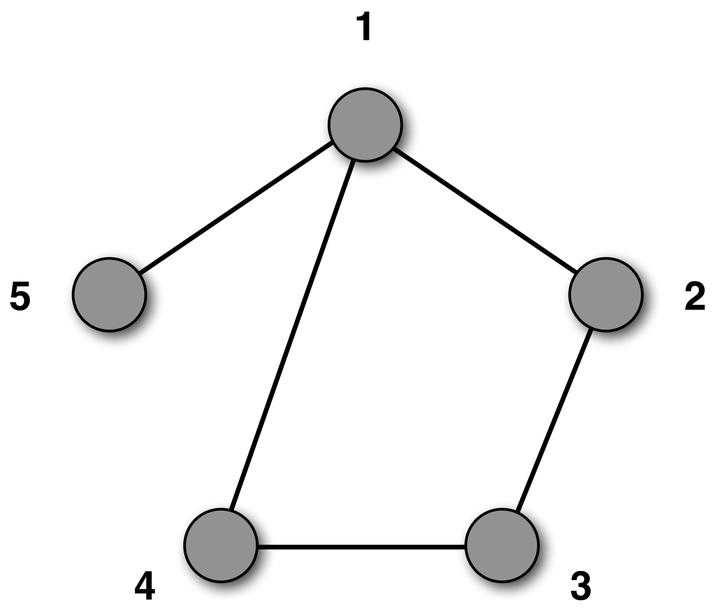
Convenient for data storage
(See: snap.stanford.edu/data/index.html)

The adjacency list stores nodes with their lists of neighbors



Node	Neighbors
1	$\{k_1 \text{ neighbors}\}$
$\dots$	
i	$\{k_i \text{ neighbors}\}$
$\dots$	
N	$\{k_N \text{ neighbors}\}$

The adjacency list stores nodes with their lists of neighbors



Node	Neighbors
1	2,4,5
2	1,3
3	2,4
4	1,3
5	1

The adjacency list needs little space and is convenient for many calculations

Needs little memory: $2L$

Convenient for many calculations involving neighbors (BFS, spreading processes,...)

Fast to add elements: $\mathcal{O}(1)$

Not convenient to find/remove elements: $\mathcal{O}(L/N)$

If finding/removing elements is important, consider the adjacency tree: like list but the set of neighbors is a tree

Needs little memory: $2L$

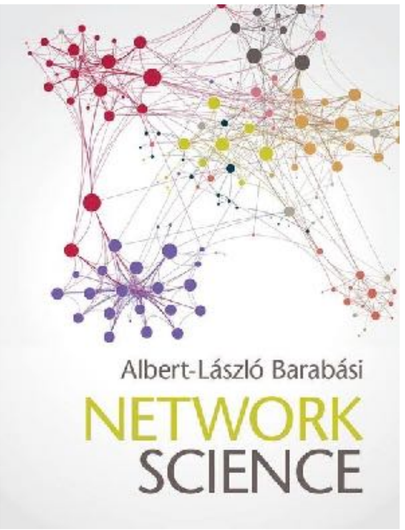
Additional complexity

Convenient for many calculations involving neighbors (BFS, spreading processes,...)

Convenient to find/remove elements: $\mathcal{O}(\log(L/N))$

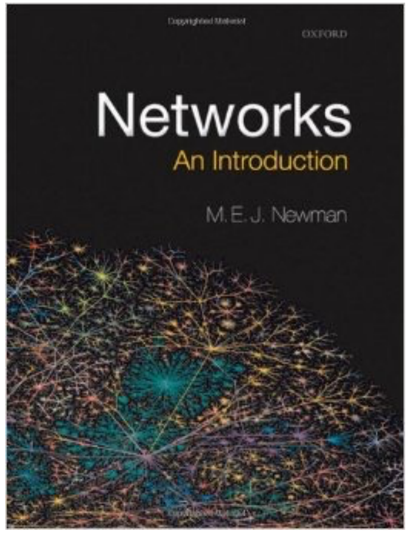
Jupyter

Sources and further materials for today's class



A.-L. Barabási.
Network Science.
Cambridge University Press (2016)

http://barabasi.com/networksciencebook/



M.E.J. Newman.
Networks: An Introduction.
Oxford University Press (2010)

Organizing principles of networks

Many networks are:

1) Heavy-tailed

2) Sparse

3) Small-world

4) Clustered