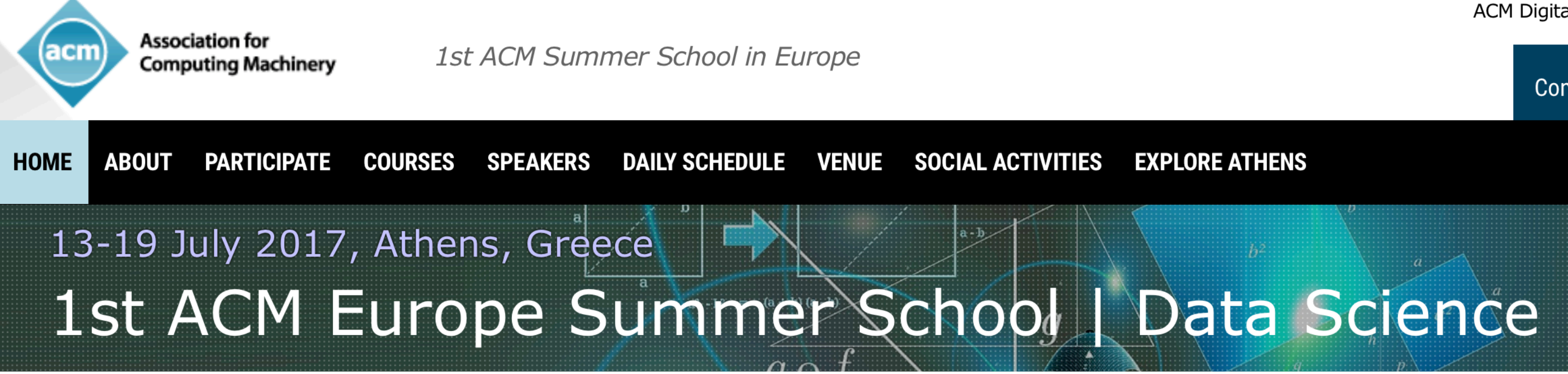


Social Network Analysis

Dino Pedreschi

KDD Lab

Università di Pisa & ISTI-CNR

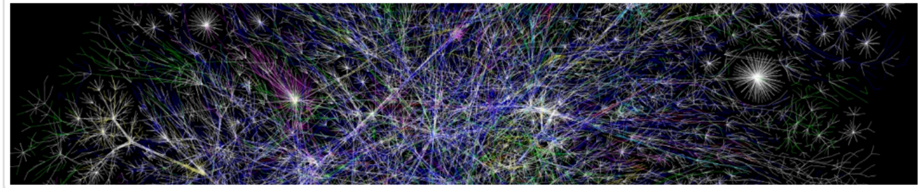
UNIVERSITÀ DI PISA

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

http://kdd.isti.cnr.it

Home Research Projects People Resources Blog

Mobility Data Mining for Science of Cities



Social Network Analysis and Visual Analytics

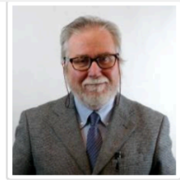
Ethical Data Mining

Analytical Platforms and Infrastructures for Social Mining

KDD Lab

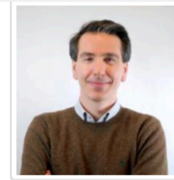
http://kdd.isti.cnr.it/people

Dino Pedreschi
Full Professor
pedre@di.unipi.it

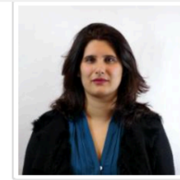


Franco Turini
Full Professor
turini@di.unipi.it

Fosca Giannotti
Director of Research
fosca.giannotti@isti.cnr.it

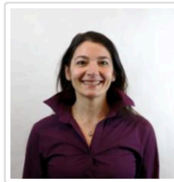


Salvatore Ruggieri
Full Professor
ruggieri@di.unipi.it

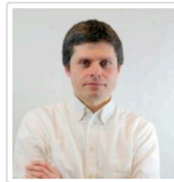


Anna Monreale
Assistant Professor
annam@di.unipi.it

Alina Sîrbu
Assistant Professor
alina.sirbu@unipi.it



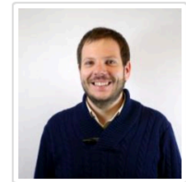
Barbara Furletti
Researcher
barbara.furletti@isti.cnr.it



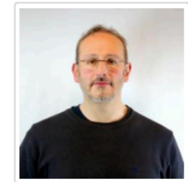
Mirco Nanni
Researcher
mirco.nanni@isti.cnr.it

Salvatore Rinzivillo
Researcher
rinzivillo@isti.cnr.it

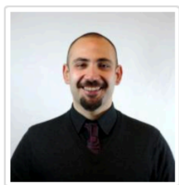
Roberto Trasarti
Researcher
roberto.trasarti@isti.cnr.it



Paolo Cintia
Post Doc
paolo.cintia@isti.cnr.it



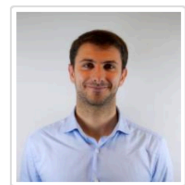
Valerio Grossi
Post Doc
vgrossi@di.unipi.it



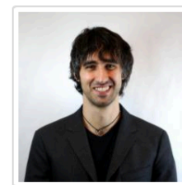
Luca Pappalardo
Post Doc
luca.pappalardo@isti.cnr.it

Giulio Rossetti
Post Doc
giulio.rossetti@isti.cnr.it

Alessandro Baroni
Ph.D. Student
baroni@di.unipi.it



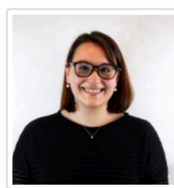
Lorenzo Gabrielli
Ph.D. Student
lorenzo.gabrielli@isti.cnr.it



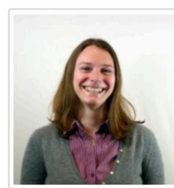
Riccardo Guidotti
Ph.D. Student
riccardo.guidotti@isti.cnr.it

Ioanna Miliou
Ph.D. Student
ioanna.miliou@for.unipi.it

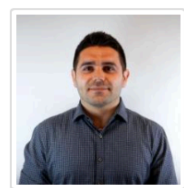
Letizia Milli
Ph.D. Student
letizia.milli@isti.cnr.it



Laura Pollacci
Ph.D. Student
laura.pollacci@di.unipi.it

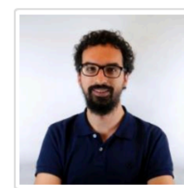


Francesca Pratesi
Ph.D. Student
francesca.pratesi@isti.cnr.it

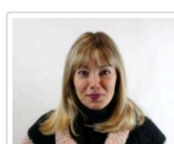


Farzad Vaziri
Ph.D. Student
vaziri@di.unipi.it

Viola Bachini
Collaborator
viola.bachini@isti.cnr.it



Daniele Fadda
Collaborator
daniele.fadda@isti.cnr.it



Knowledge Discovery & Data Mining Lab
http://kdd.isti.cnr.it

SoBigData Research Infrastructure

Social Mining & Big Data Analytics

H2020 - www.sobigdata.eu

September 2015- August 2019

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Istituto di Informatica e Telematica del CNR

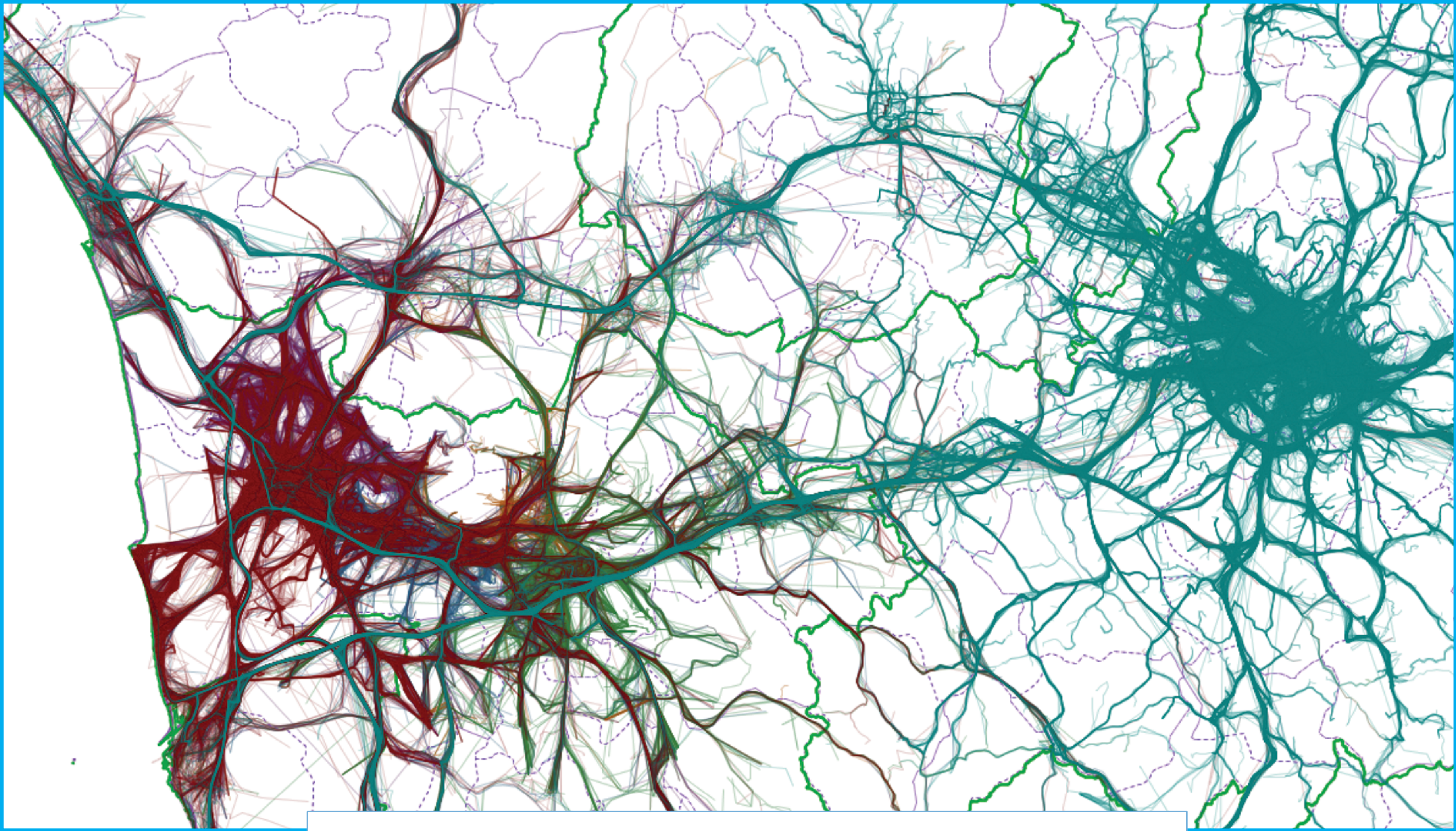
UNIVERSITÀ DI PISA

SCUOLA NORMALE SUPERIORE

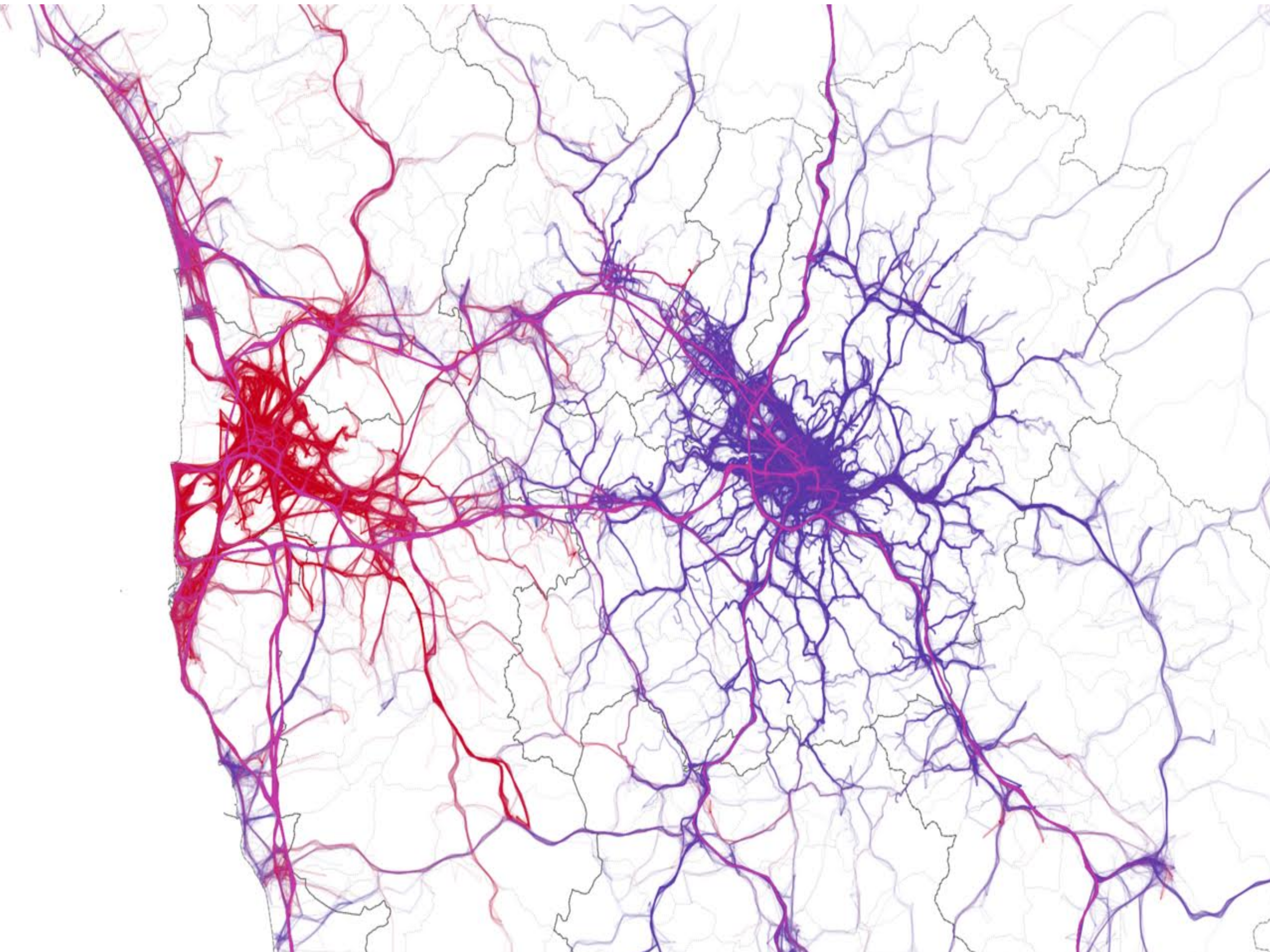
Institutions Markets Technologies IMT INSTITUTE FOR ADVANCED STUDIES LUCCA

Exploratory:

Big Data for City of Citizens



Personal Mobility, Social + Mobility, Personal Sensing

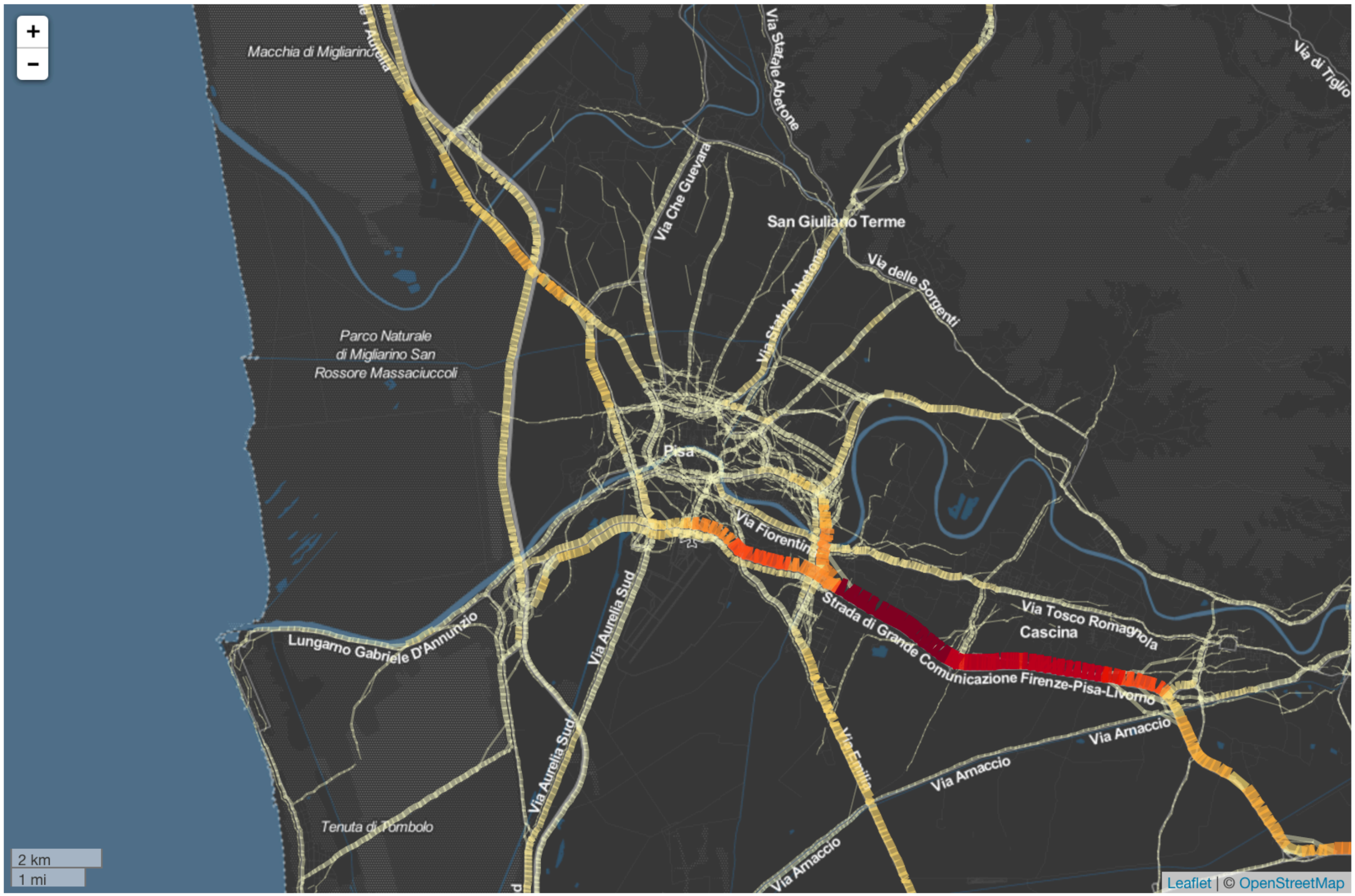


UMA Urban Mobility Atlas Select a city: Pisa | In | Out | Self | Occasional | Systematic | All

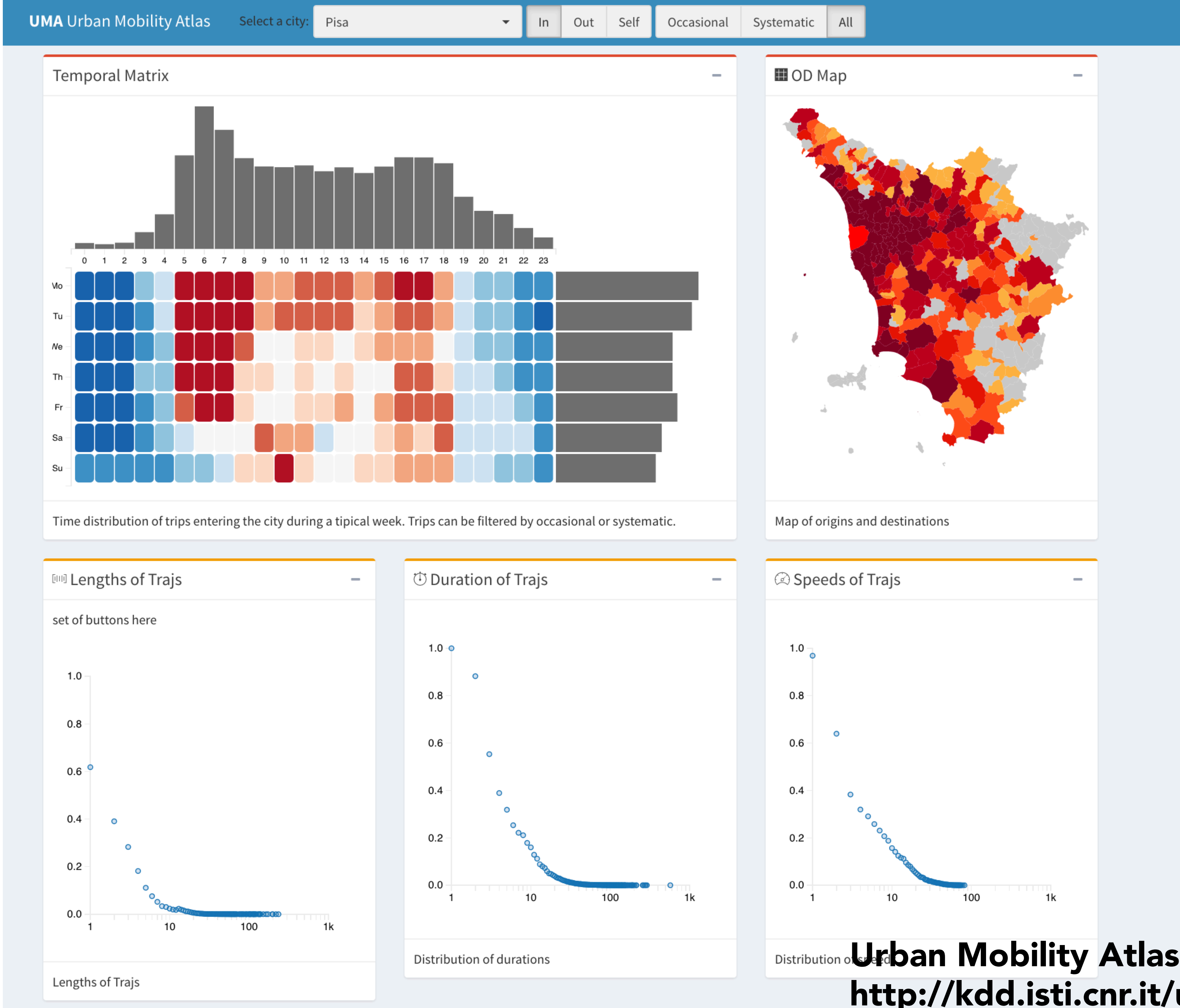
Pisa Pisa

NUMBER OF VEHICLES	RESIDENTS	DAYS	SYSTEMATICS TRAJS
5,615	89,694	31	38.88%
Source: OCTO Telemati...	Source: census 2001	May 1-31, 2011	

Flows and Traffic



Flows of traffic exiting from the city.



Urban Mobility Atlas

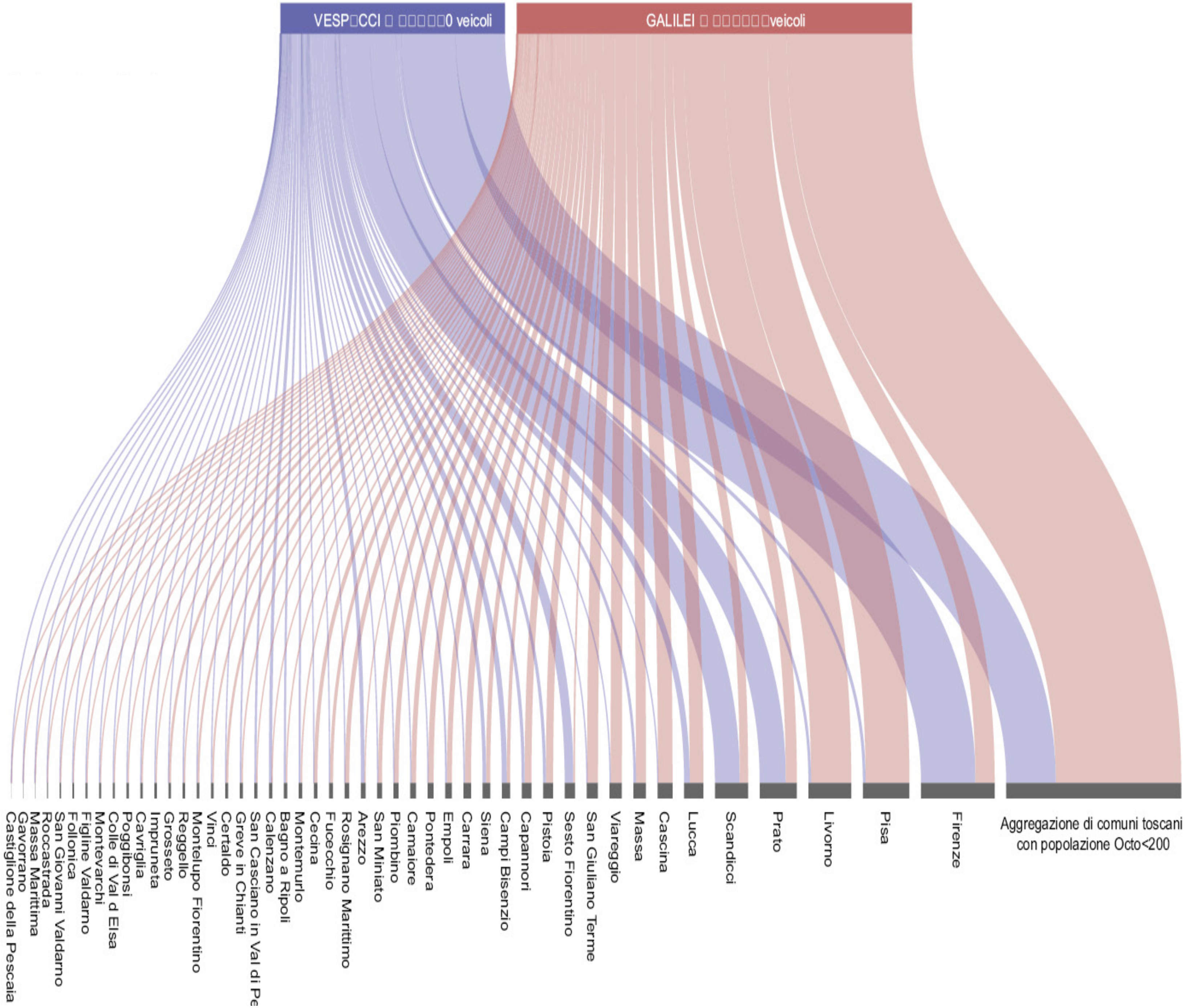
http://kdd.isti.cnr.it/uma2/

Analizzare l'influenza dei grandi attrattori sulla mobilità dei territori circostanti

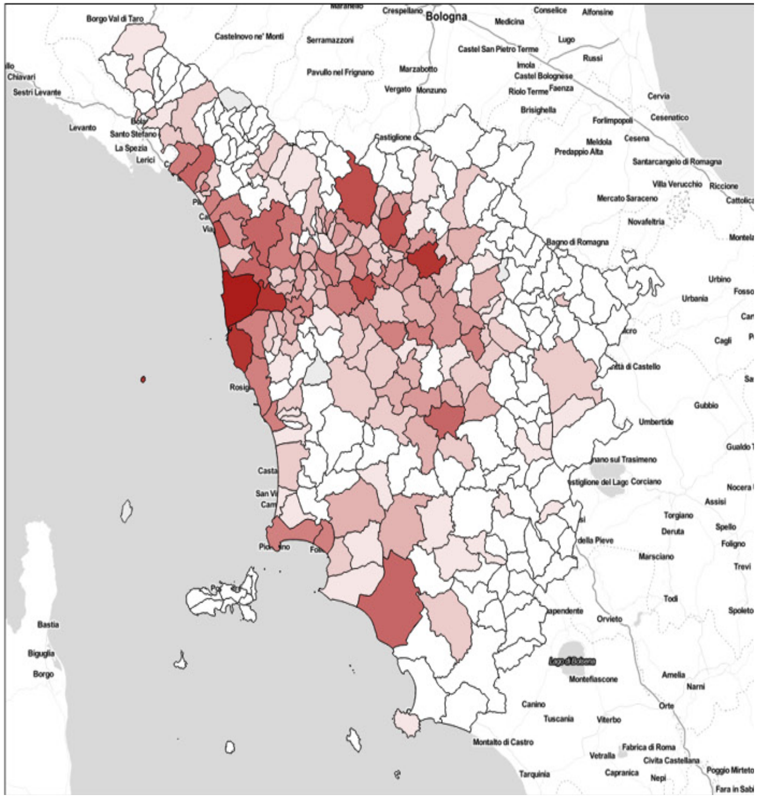
CASO STUDIO:

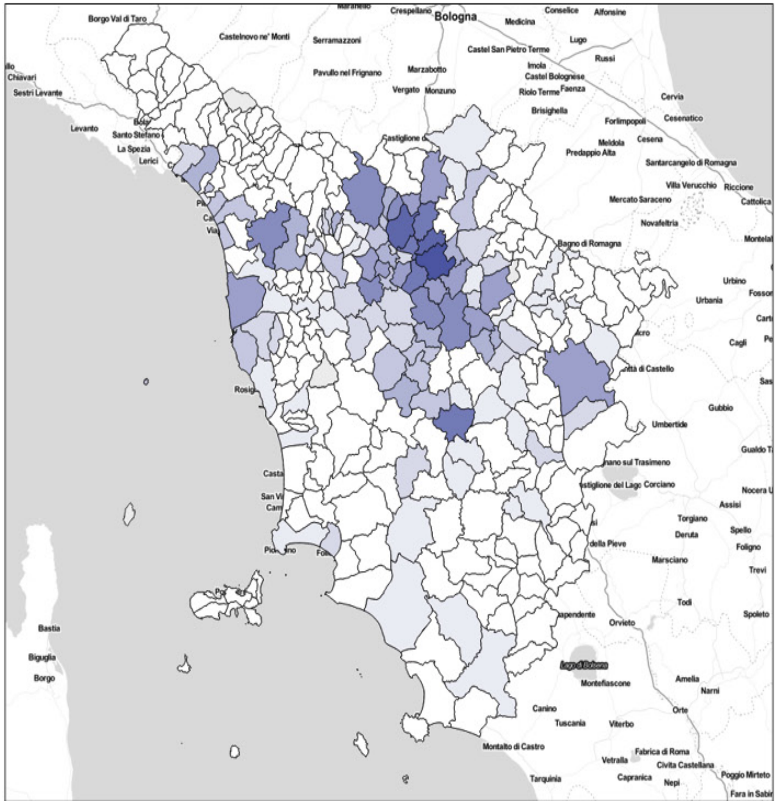
gli aeroporti di Firenze e Pisa e la propensione dei residenti toscani all'uso delle due infrastrutture.


ACCESS POINT PISA
principali access point alla citta di Pisa
degli utenti che si recano all'aeroporto Galilei
usando l'automobile
Access point
Volume auto osservato
Cluster di traiettorie
per route similarity
Migliarino - Vecchiano
Statale Abetone San Giuliano
Aurelia Nord e A12
FIPILI (Livorno) e A12-sud
Aurelia sud
FIPILI (Firenze)


VESP CCI 0 veicoli
GALILEI veicoli
Castiglione della Pescaia
Gavorrano
Massa Marittima
Roccastrada
San Giovanni Valdarno
Follonica
Figline Valdarno
Montevarchi
Colle di Val d Elsa
Poggibonsi
Cavriglia
Impruneta
Grosseto
Reggello
Montelupo Fiorentino
Vinci
Certaldo
Greve in Chianti
San Casciano in Val di Pe
Calenzano
Bagno a Ripoli
Montemurlo
Cecina
Fucecchio
Rosignano Marittimo
Arezzo
San Miniato
Piombino
Camaiore
Pontedera
Empoli
Carrara
Siena
Campi Bisenzio
Capannori
Pistoia
Sesto Fiorentino
San Giuliano Terme
Viareggio
Massa
Cascina
Lucca
Scandicci
Prato
Livorno
Pisa
Firenze
Aggregazione di comuni toscani
con popolazione Octo<200

Attractiveness of Galilei vs. Vespucci





Modeling Investments and Attractiveness on Tuscan Airports

An intertwined system based on investment and attractiveness

$$\frac{\mathrm{d}}{\mathrm{d}t}A = s(mF - (k+e)A),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}F = -rF + re\frac{bA}{1+bhA}.$$

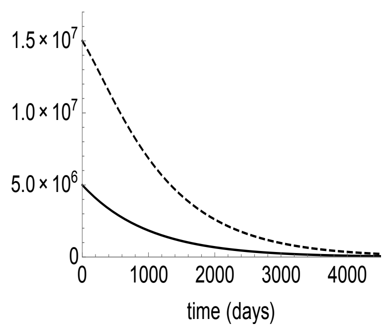
A - > Attractiveness of airport

F -> Number of passengers served

Attractiveness is proportional to the cost of operating the airport (k) and the extra investments (e)

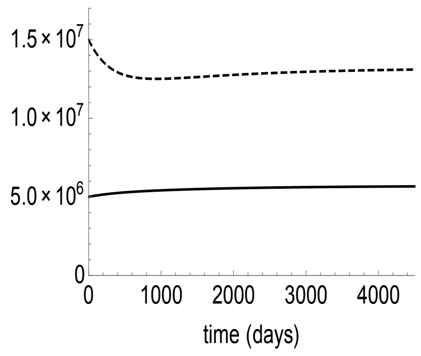
Simple case: non spatial model

No extra investments (e=0)



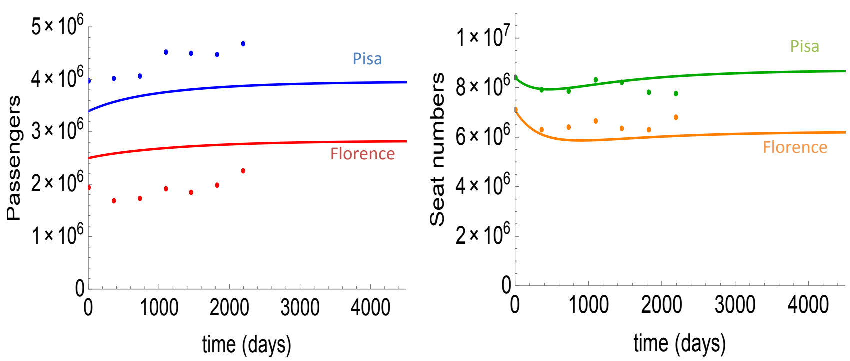
(a) $e = 0$

Structural investments (e=0.05)



(b) $e = 0.05$

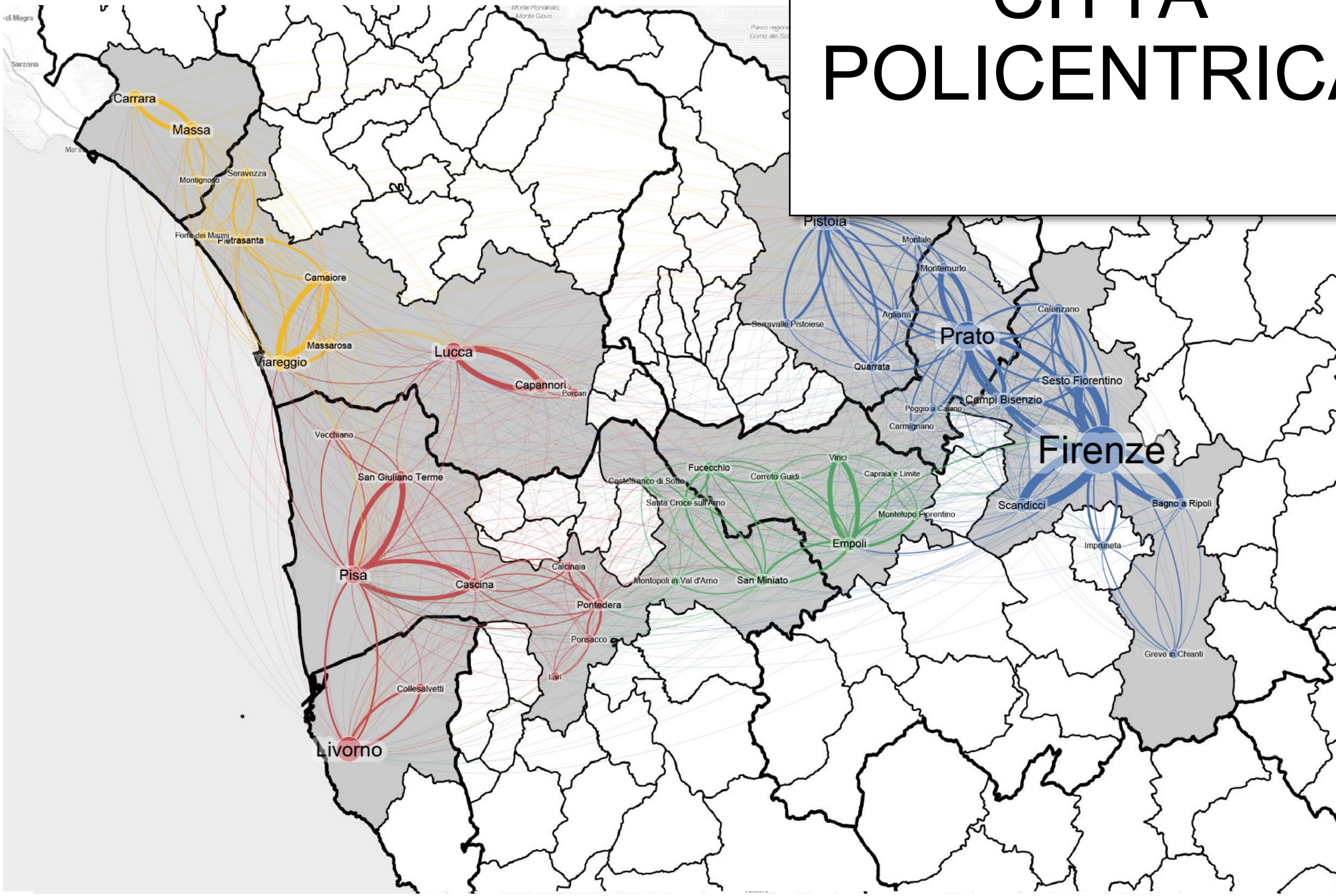
Spatial model: two airports, two populations

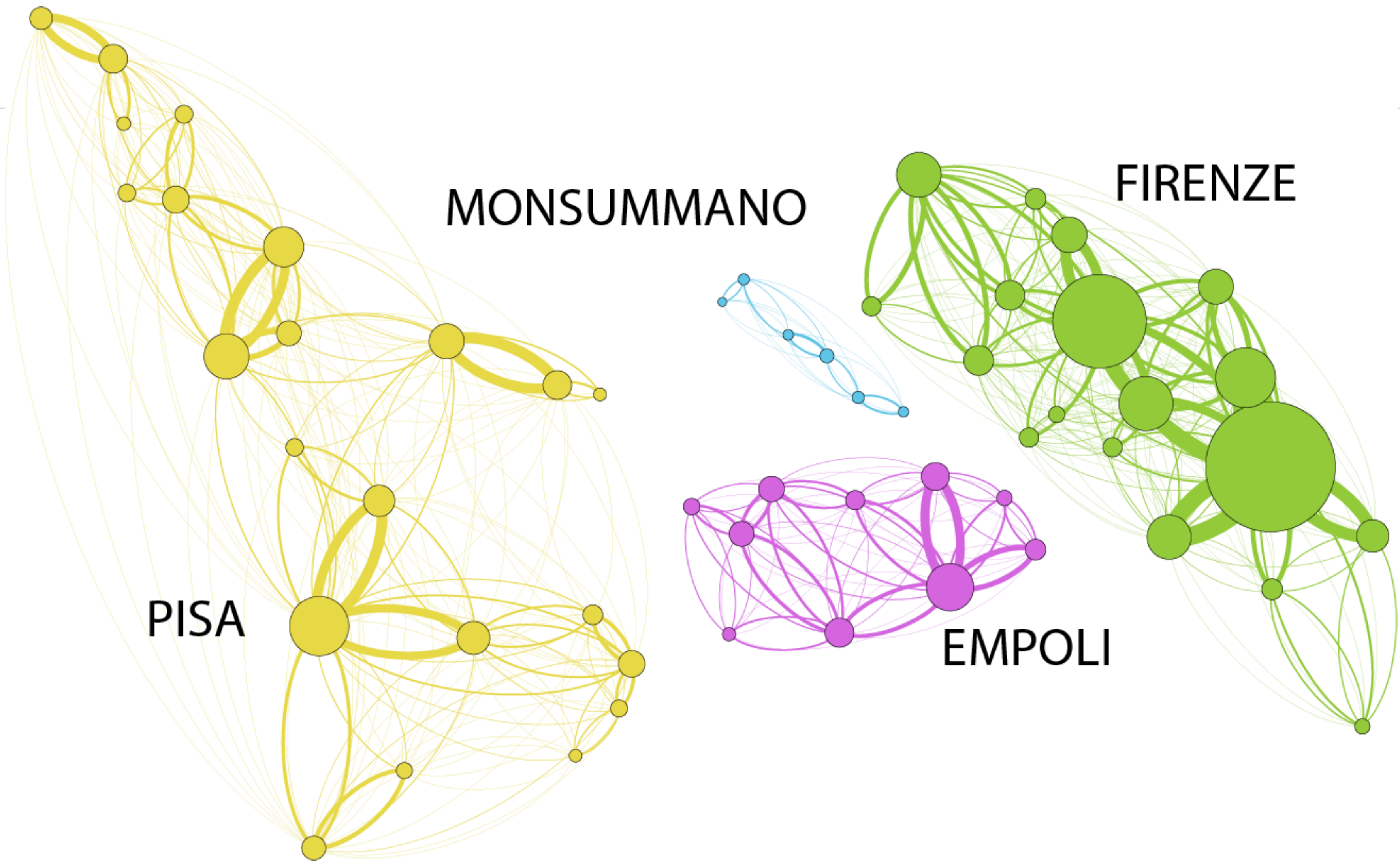


The two airports reach an equilibrium: neither of the two is overwhelming the other

Ioanna Miliou, Diana Knipl, Salvatore Rinzivillo, Seven R. Bishop, Dino Pedreschi.

CITTÀ POLICENTRICA




MONSUMMANO
FIRENZE
PISA
EMPOLI

Exploratory:

Big Data for Societal Debates

Polarization, controversy and topic trends on societal debates through social media

HOW BRITAIN TALKED BREXIT
NHS
IMMIGRATION
TRONGER IN
BREAKING POINT
We send the EU £350 million a week
let's fund our NHS instead
Vote Leave
ECONOMY
LAW
SOVEREIG

3 Million Brexit Tweets Reveal Leave Voters Talked About Immigration More Than Anything Else

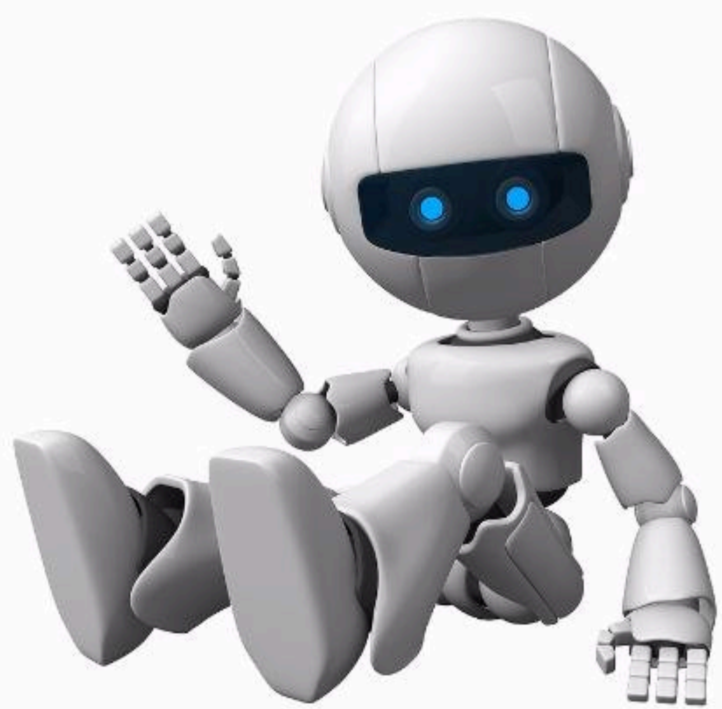
Groundbreaking analysis shows immigration, not sovereignty or the NHS, dominated the conversation – and making British judges responsible for British law was a key theme for Leave supporters.

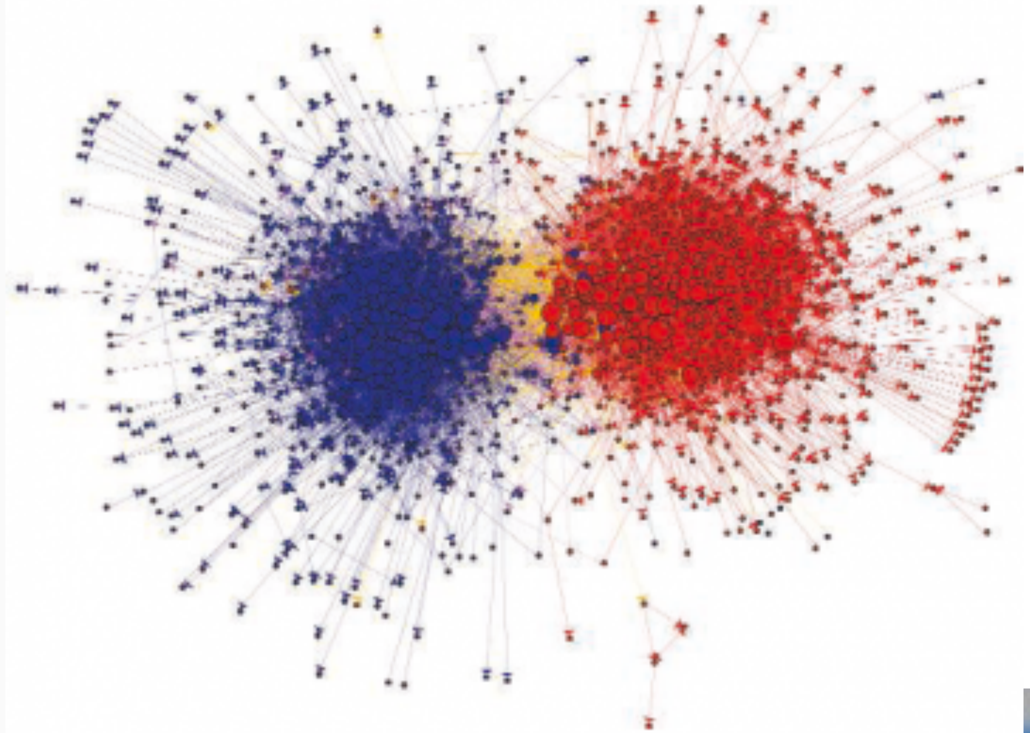
James Ball
BuzzFeed Special Correspondent

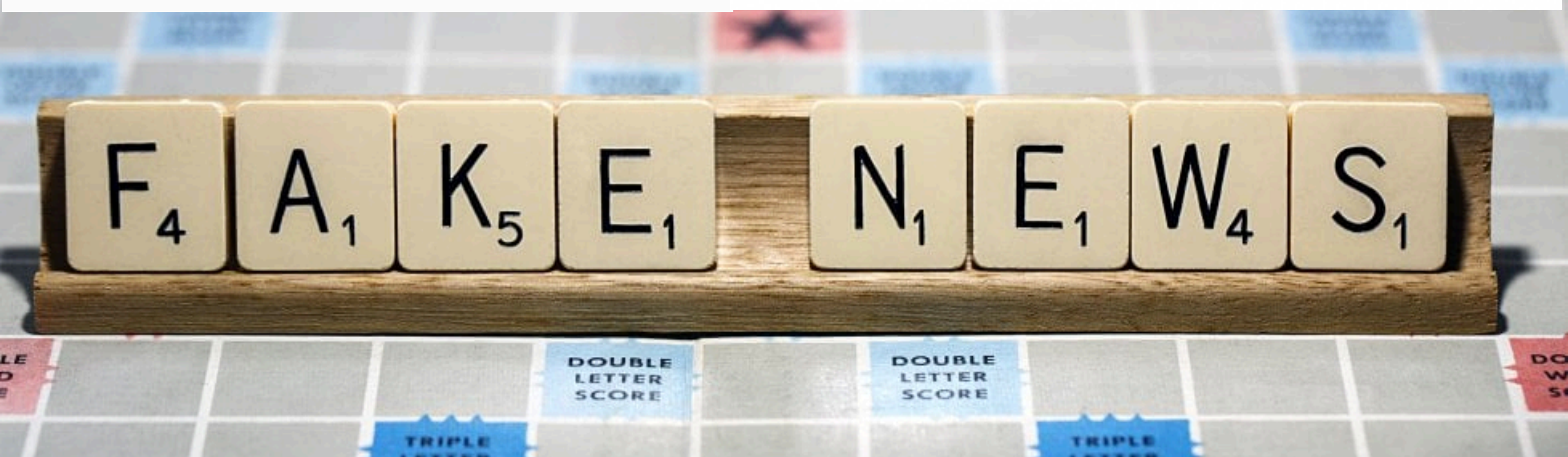
Chris Applegate
Editorial Developer, UK

posted on Dec. 9, 2016, at 2:03 p.m.

https://www.buzzfeed.com/jamesball/3-million-brexit-tweets-reveal-leave-voters-talked-about-imm?utm_term=.jmDQE9JNR#.fuOOrb145

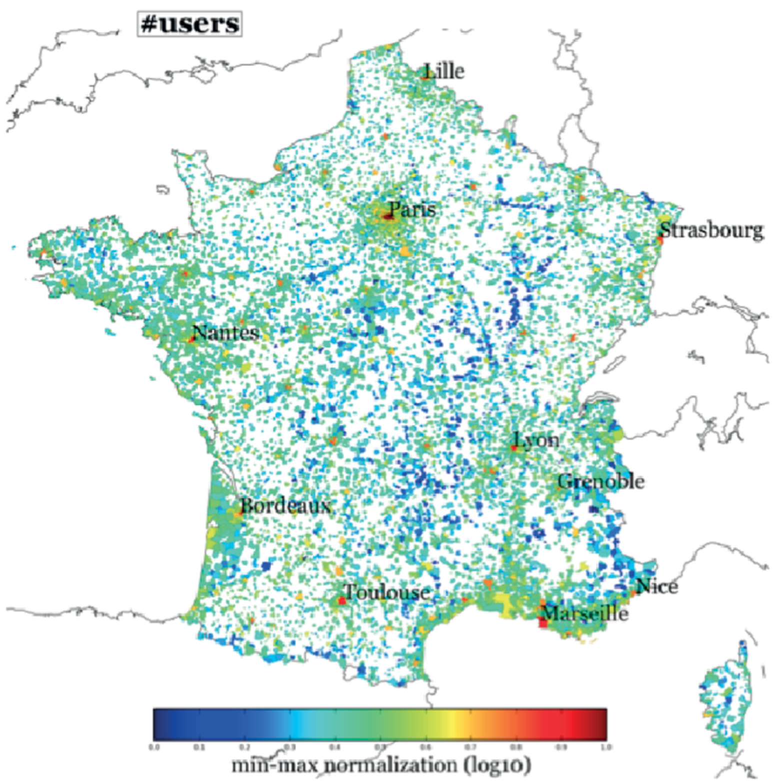


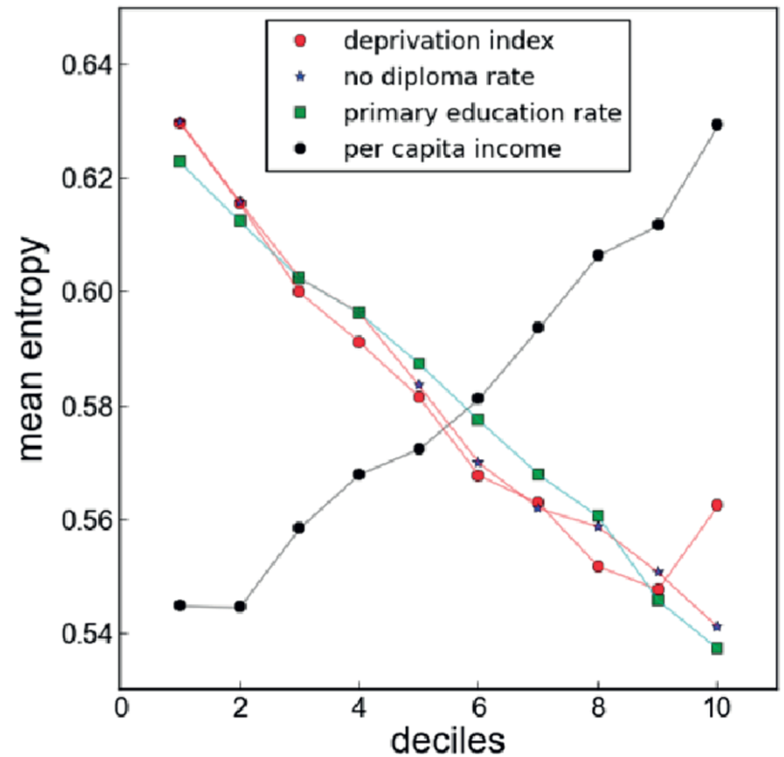


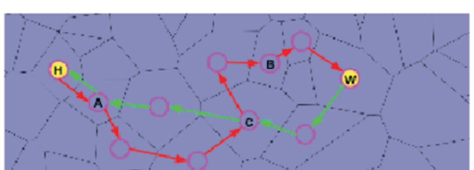

F A K E N E W S
DOUBLE LETTER SCORE
DOUBLE LETTER SCORE
TRIPLE
TRIPLE

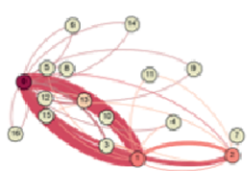
Exploratory:

Big Data for Well Being and Economic Performance









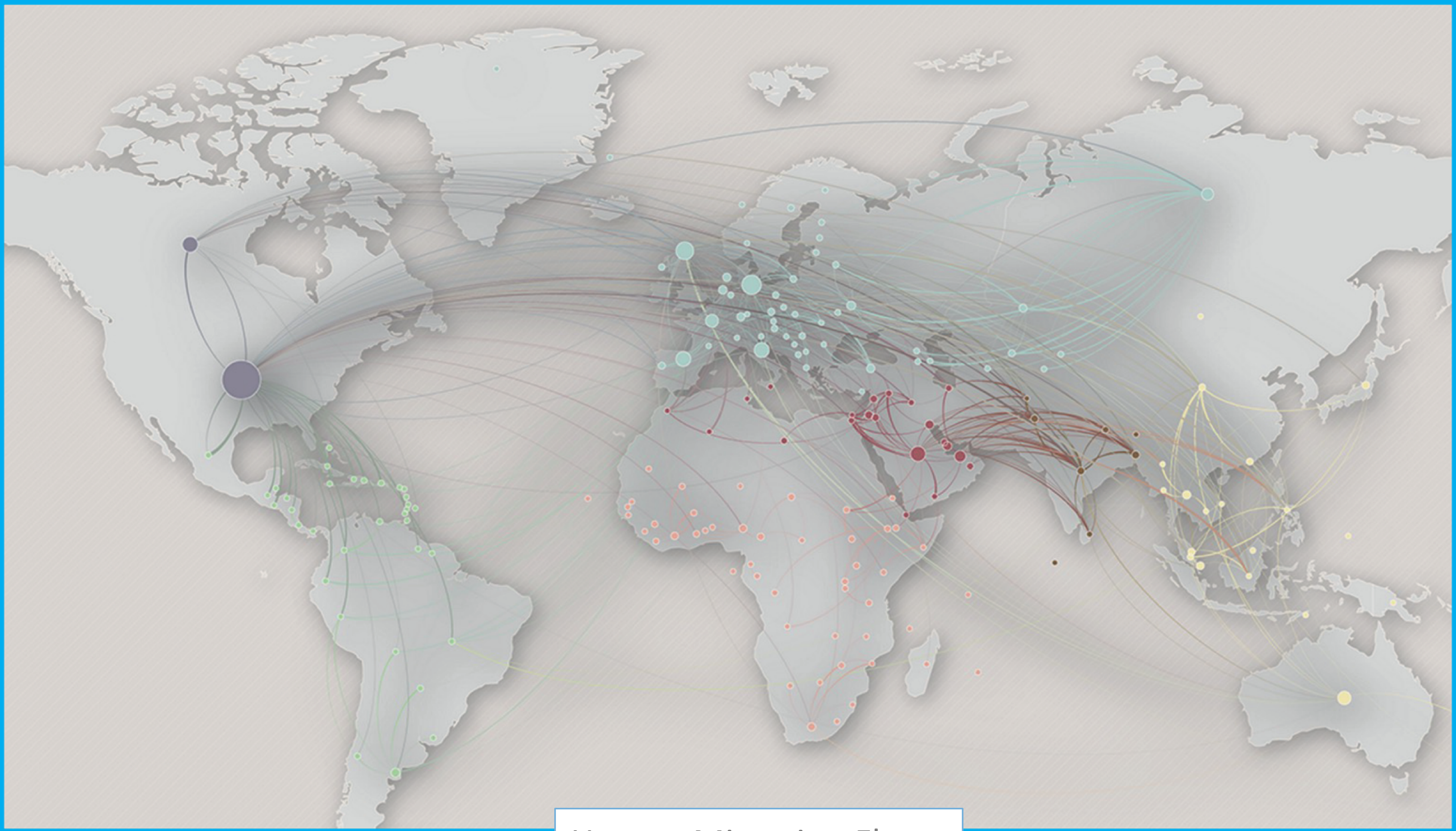
$$d_i^{(n)} = \sum_{j=1}^{|V|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i$$

$$p_j^{(n)} = \sum_{i=1}^{|U|} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j$$

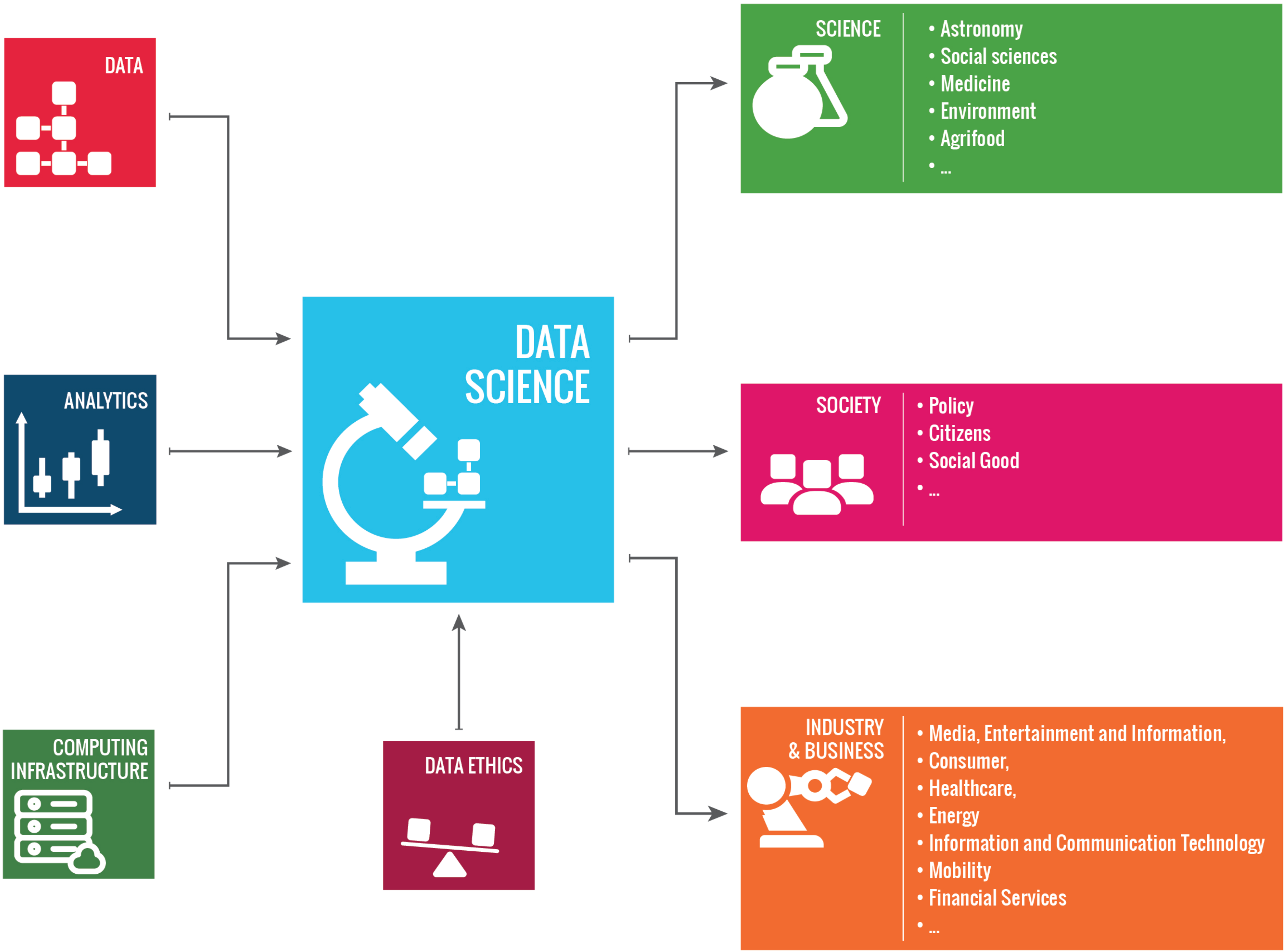
Deprivation Index (in France) predicted with Mobile Phone traces

Next Exploratory:

Big Data for Migration Studies



Human Migration Flows


DATA
ANALYTICS
COMPUTING INFRASTRUCTURE
DATA SCIENCE
DATA ETHICS
SCIENCE
• Astronomy
• Social sciences
• Medicine
• Environment
• Agrifood
• ...
SOCIETY
• Policy
• Citizens
• Social Good
• ...
INDUSTRY & BUSINESS
• Media, Entertainment and Information,
• Consumer,
• Healthcare,
• Energy
• Information and Communication Technology
• Mobility
• Financial Services
• ...

New economic growth:
the role of science, technology, innovation and infrastructure

Maryse Lassonde
Royal Society of Canada

Sébastien Candel
Académie des sciences

Jörg Hacker
Leopoldina Nationale Akademie der Wissenschaften

Alberto Quadrio-Curzio
Accademia Nazionale dei Lincei

Takashi Onishi
Science Council of Japan

Venki Ramakrishnan
Royal Society

Marcia McNutt
National Academy of Sciences

Policy recommendations

G7 Academies of Science urge governments to:

i. expand investment and capabilities in science and pre-competitive technologies;

Attention should be given to emerging technologies in light of their potential to impact virtually all economic activities:

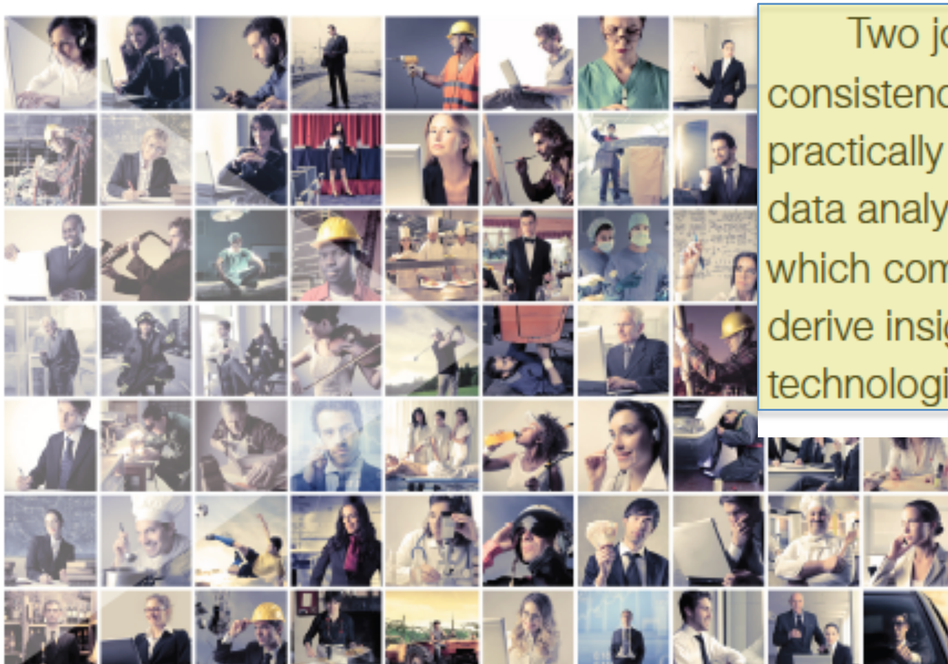
- Data Science, thanks to the ability to extract new knowledge and policy capability by the integrated algorithmic analysis of highly diverse data generated today at exponentially growing pace.

Global Challenge Insight Report

The Future of Jobs

Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution

January 2016



New and Emerging Roles

Our research also explicitly asked respondents about new and emerging job categories and functions that they expect to become critically important to their industry by the year 2020, and where within their global operations they would expect to locate such roles.

Two job types stand out due to the frequency and consistency with which they were mentioned across practically all industries and geographies. The first are data analysts, as already frequently mentioned above, which companies expect will help them make sense and derive insights from the torrent of data generated by the technological disruptions referenced above. The second

http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

Corso di laurea Magistrale in Data Science and Business Informatics

Dipartimento di Informatica

Università di Pisa

HOME DIDATTICA RICERCA DOTTORATO PERSONE ORGANIZZAZIONE EVENTI CERCA...

Home ▸ Didattica ▸ Laurea Magistrale in Data Science and Business Informatics ▸ Didattica ▸ Erasmus

Login

ENROLLMENT FOR FOREIGN STUDENTS

Magistrale in Data Science and Business Informatics

- Insegnamenti
- Docenti
- Orario
- Calendario AA 2016/2017
- Calendario AA 2017/2018
- Calendario appelli
- Piani di studio
- Servizio di tutorato
- Progetto formativo
- Lauree
- Internazionale
- Valutazione

Presentazione

Il Corso di Laurea Magistrale in **Data Science and Business Informatics** (fino all'A.A. 2016/17, Business Informatics) è progettato, a partire dal 2002, per preparare laureati magistrali in grado di padroneggiare sia le tecnologie informatiche che di

NOTIZIE DIPARTIMENTO

- 12 PhD positions in Computer Science at University of Pisa

https://www.di.unipi.it/it/didattica/wds-lm

Founded 2002

Big Data Analytics & Social Mining

SoBigData Master

Febbraio 01 2017

Master Big Data 2017 Inaugurazione

http://masterbigdata.it

Dal 2014

SCUOLA NORMALE SUPERIORE

Sant'Anna
Scuola Universitaria Superiore Pisa

CLOUD4WI

sky

extra
Open Solutions, Smart Integration

GENERALI

KPMG

NAVIONICS

net7

nòva 24

OCTO

TAN
SOFTWARE SOLUTION

Trust-IT Services
Communicating ICT to markets

coop
Unicoop Tirreno

UnipolSai
ASSICURAZIONI

WIND
Più vicini.

esa

IRPET

Istat

Ph.D. in Data Science

Start: academic year 2017-2018

http://phd.sns.it/data-science/

SCUOLA NORMALE SUPERIORE

SCUOLA SUPERIORE SANT'ANNA PISA

IMT SCUOLA ALTI STUDI LUCCA

Consiglio Nazionale delle Ricerche

SoBigData Research Infrastructure

www.sobigdata.eu

H2020 excellent science research infrastructure

SCUOLA NORMALE SUPERIORE

IN SUPREMÆ DIGNITATIS 1343

SCUOLA SUPERIORE SANT'ANNA PISA

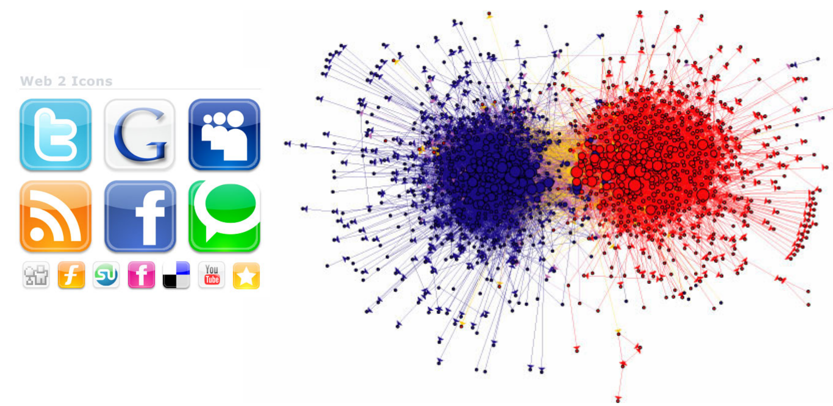
IMT SCUOLA ALTI STUDI LUCCA

Consiglio Nazionale delle Ricerche

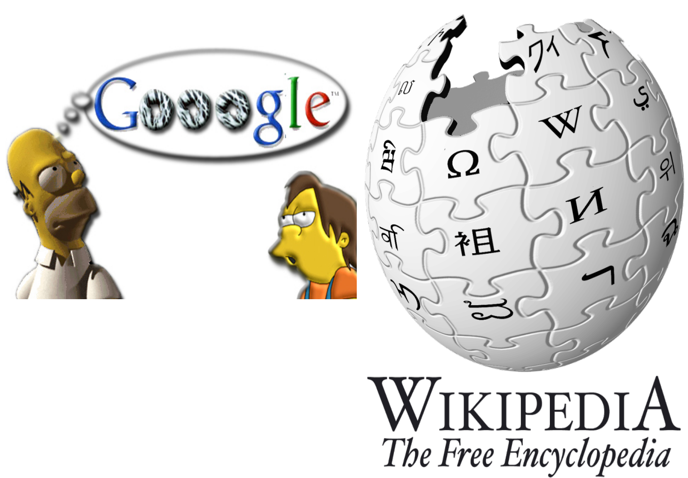
Big data proxies of social life

Shopping patterns & lyfestyle

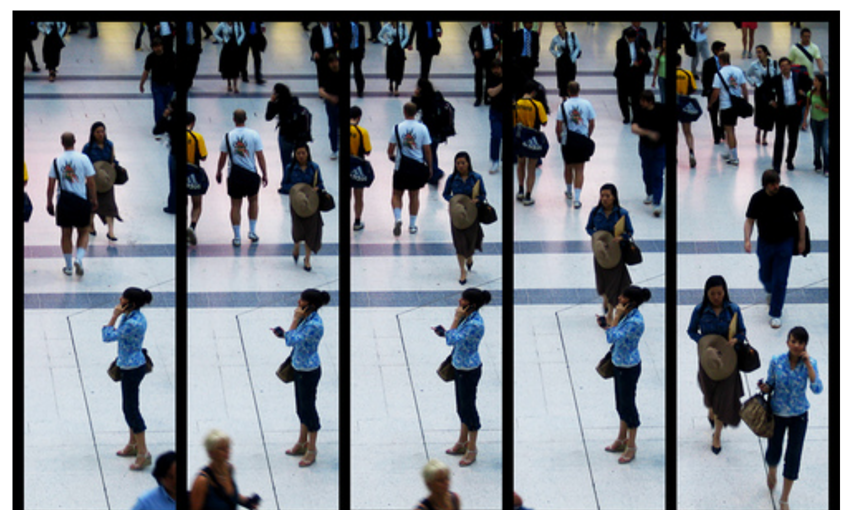
RELATIONSHIPS & SOCIAL TIES

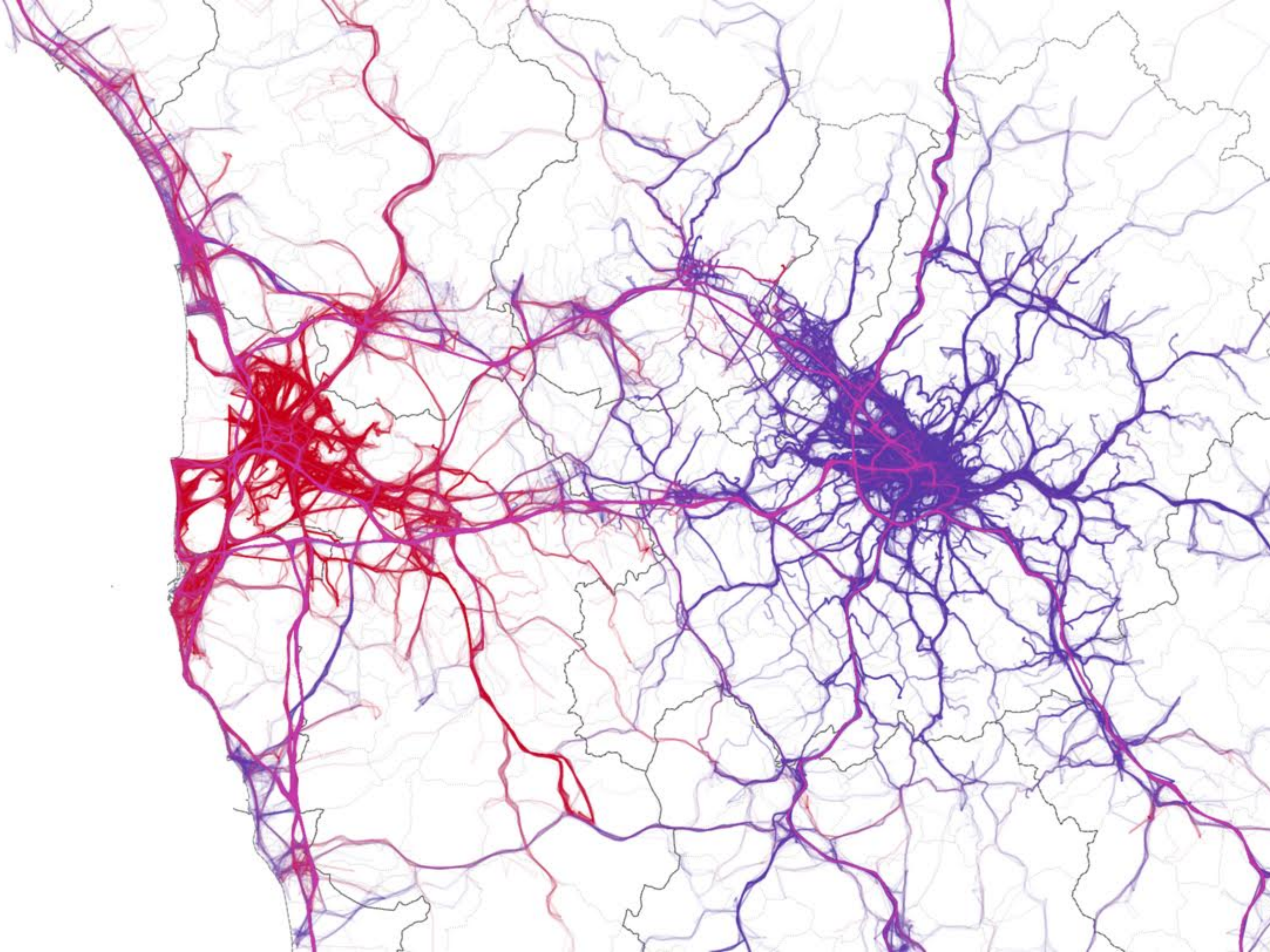


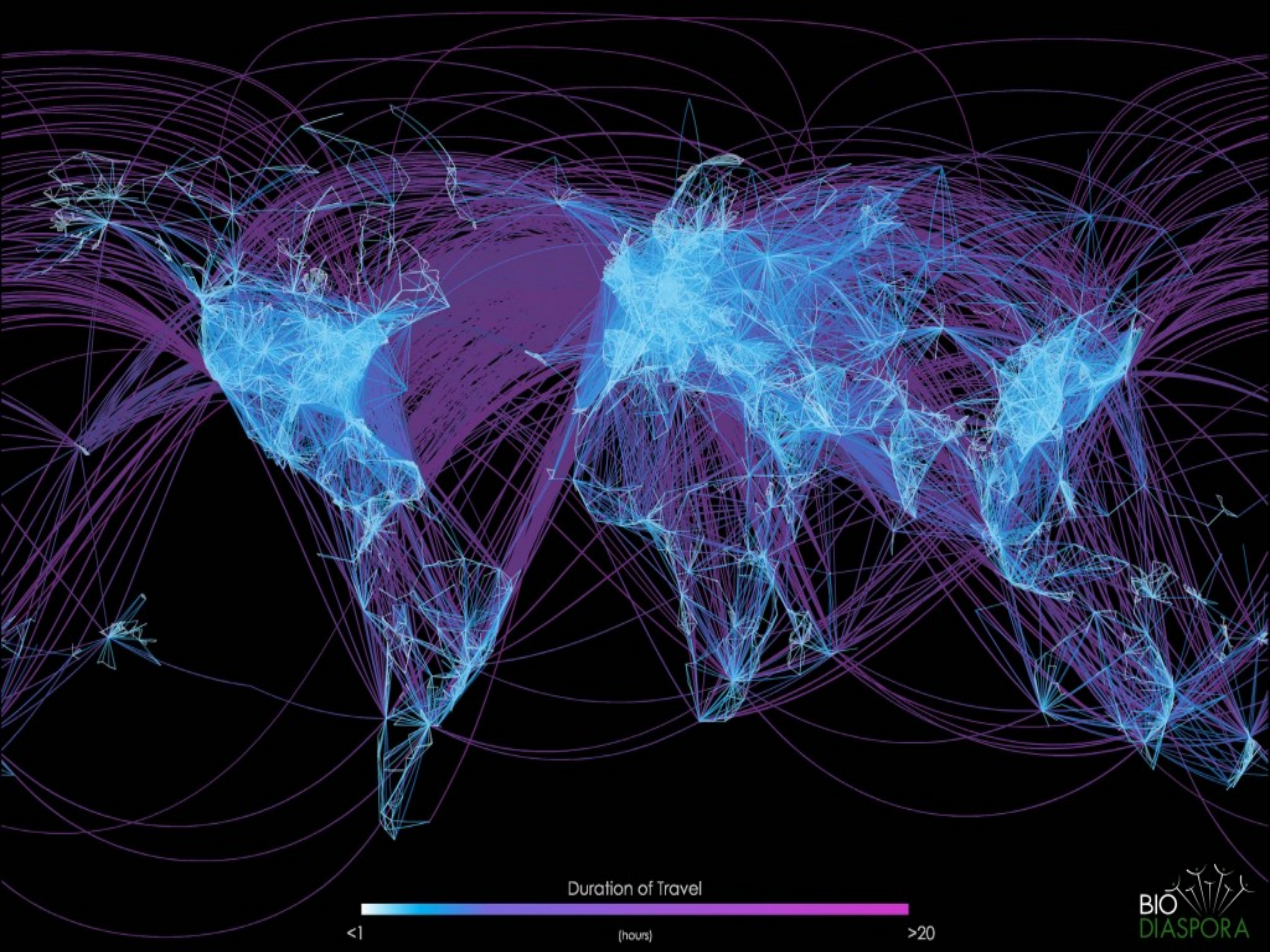
DESIRES, OPINIONS, SENTIMENts



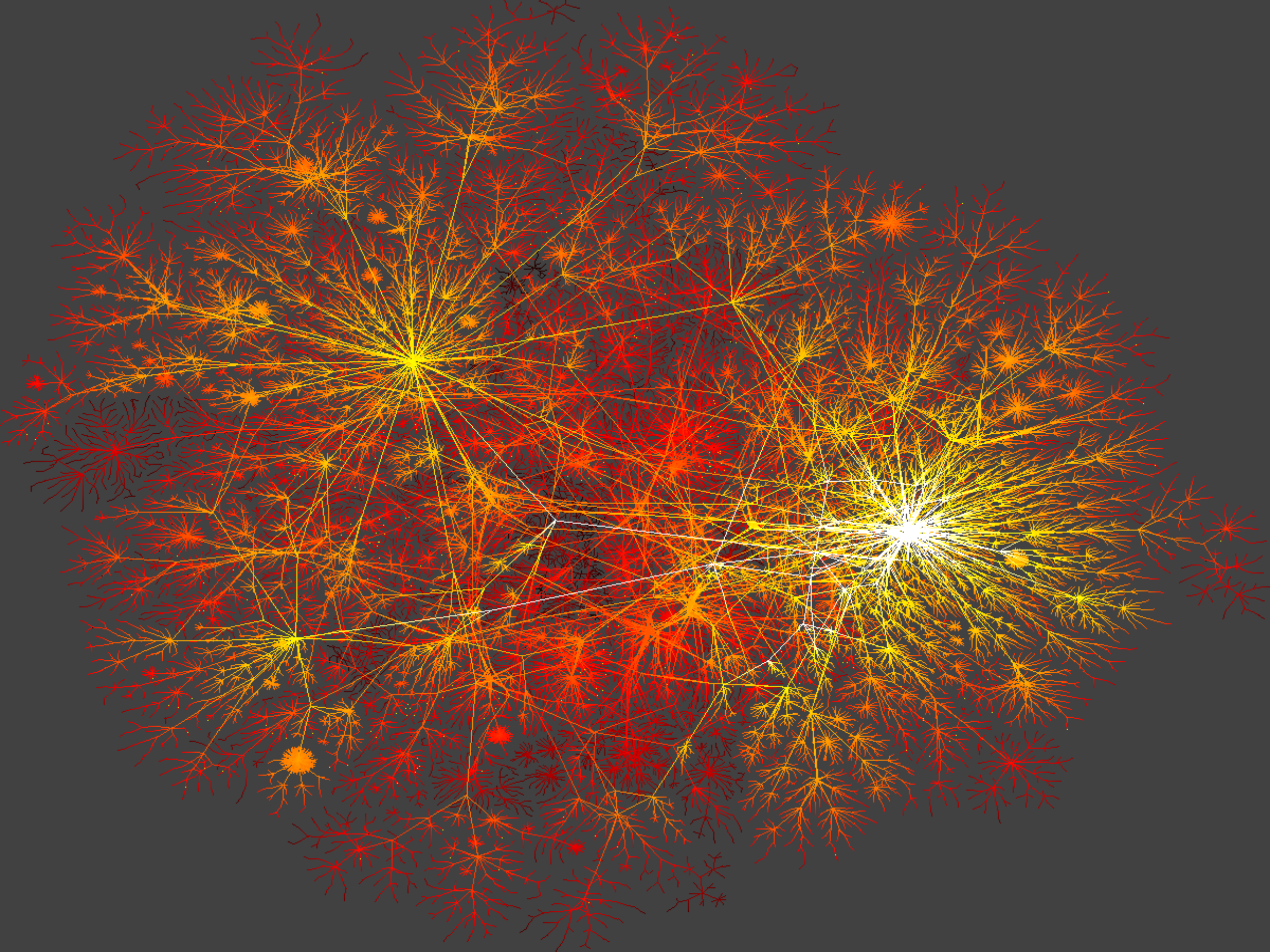
MOVEMENTS

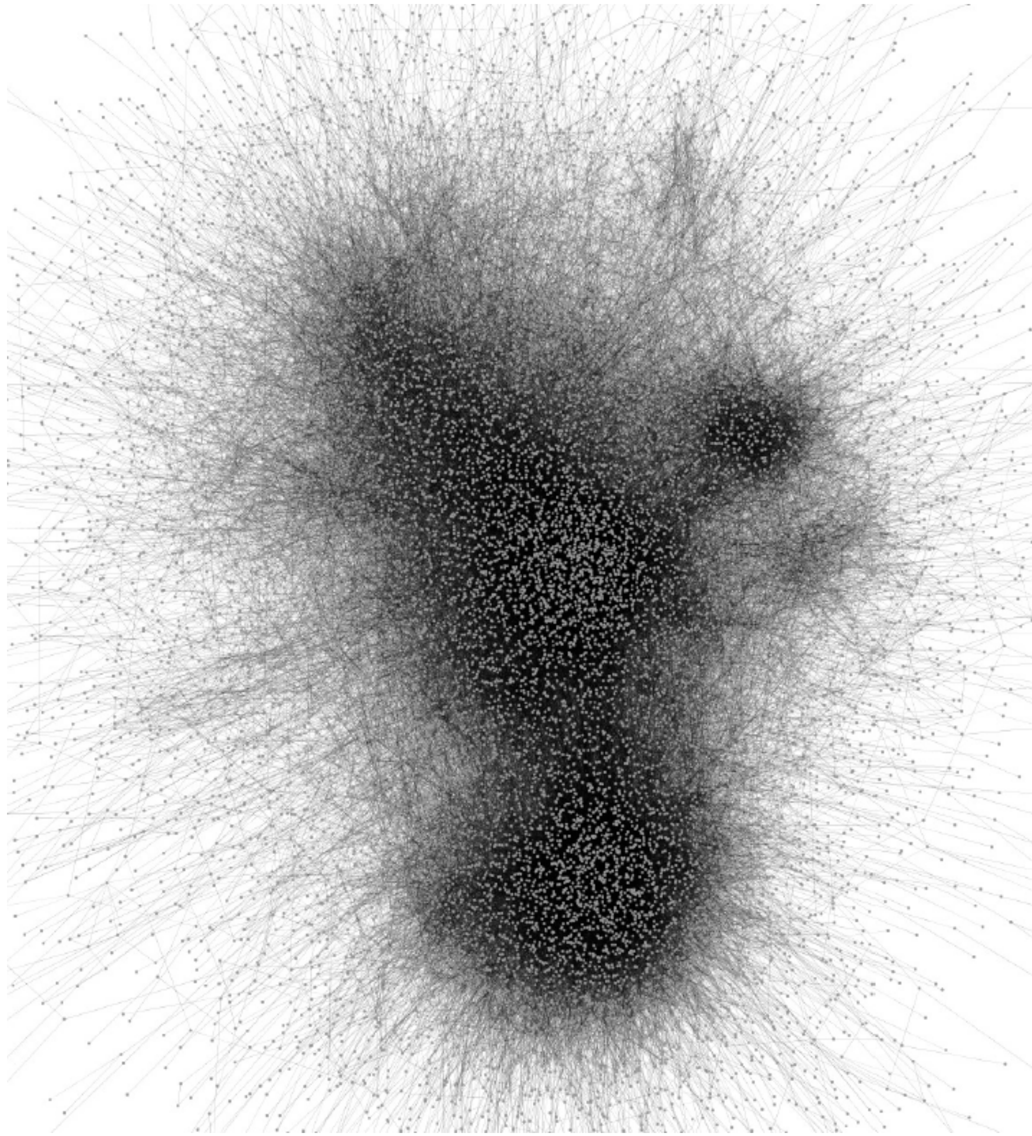


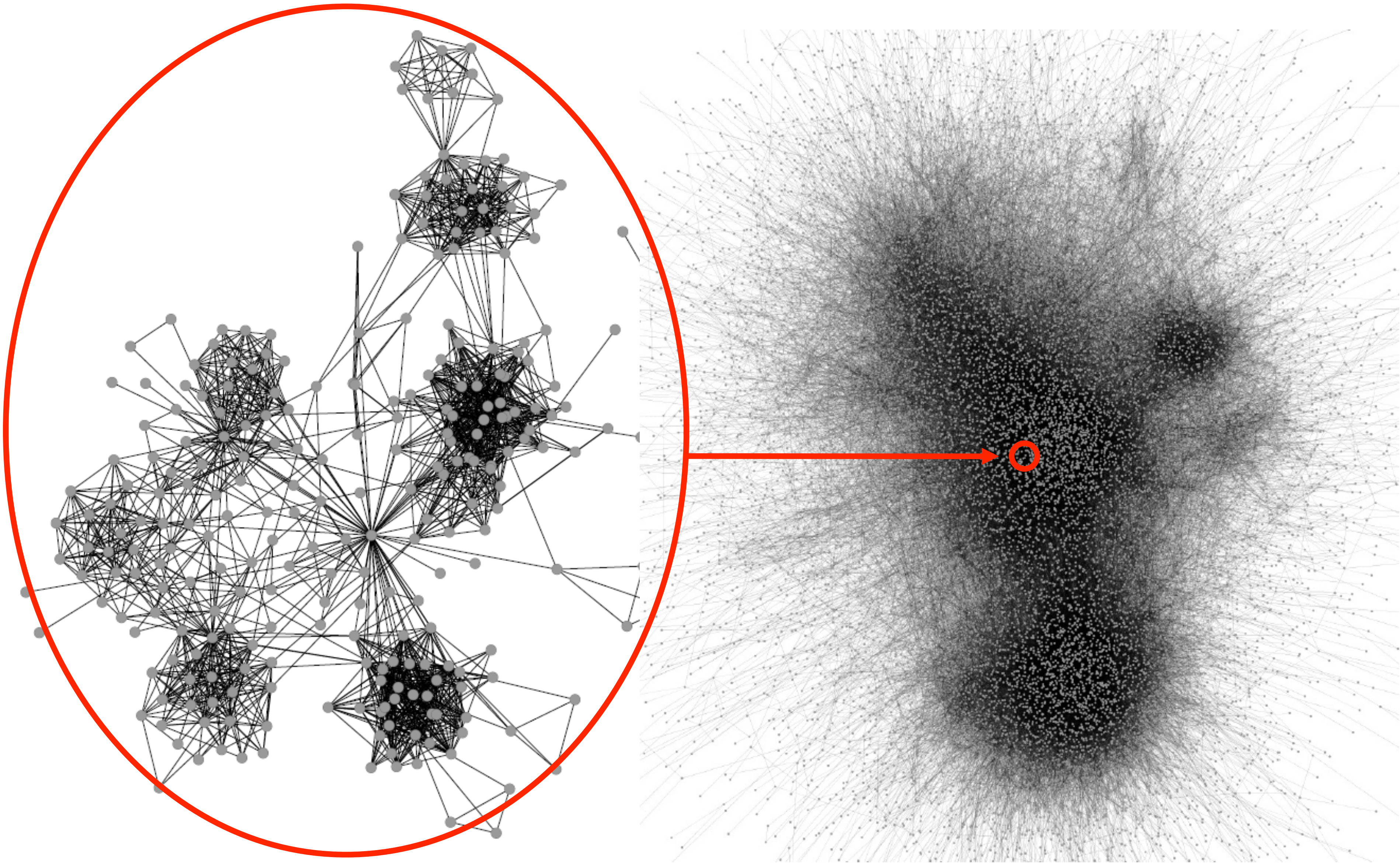




Duration of Travel
<1
(hours)
>20
BIO
DIASPORA







Complex (Social) Networks

- Big graph data and social, information, biological and technological networks
- The architecture of complexity and how real networks differ from random networks:
  - node degree and long tails,
  - social distance and small worlds,
  - clustering and triadic closure.
- Comparing real networks and random graphs.
- The main models of network science: small world and preferential attachment.

Complex (Social) Networks

- Strong and weak ties, community structure and long-range bridges.
- Robustness of networks to failures and attacks.
- Cascades and spreading. Network models for diffusion and epidemics. The strength of weak ties for the diffusion of information. The strength of strong ties for the diffusion of innovation.

Complex (Social) Networks

- Textbooks
  - Albert-Laszlo Barabasi. *Network Science* (2016)
  - http://barabasi.com/book/network-science
  - Easley, Kleinberg: *Networks, Crowds, and Markets* (2010)
  - http://www.cs.cornell.edu/home/kleinber/networks-book/
- Network Analytics Software:
  - Cytoscape: http://www.cytoscape.org/
  - Gephi: http://gephi.github.io/
- Network dynamics simulation :
  - NetLogo: https://ccl.northwestern.edu/netlogo/
- Network Data Repository
  - http://networkrepository.com/

Wiki of the course

- http://didawiki.di.unipi.it/doku.php/wma/acm-athens-july2017

- Special thanks to
  - Fosca Giannotti, ISTI-CNR Pisa
  - Albert-Laszlo Barabasi, Northeastern Univ. Boston
  - Giulio Rossetti, University of Pisa
  - Jure Leskovec, Stanford Univ.

The architecture of complexity

Lecture 1

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]

–adjective

1.
composed of many interconnected parts; compound; composite: a complex highway system.

2.
characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3.
so complicated or intricate as to be hard to understand or deal with: a complex problem.

Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as **emergent behaviour**, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

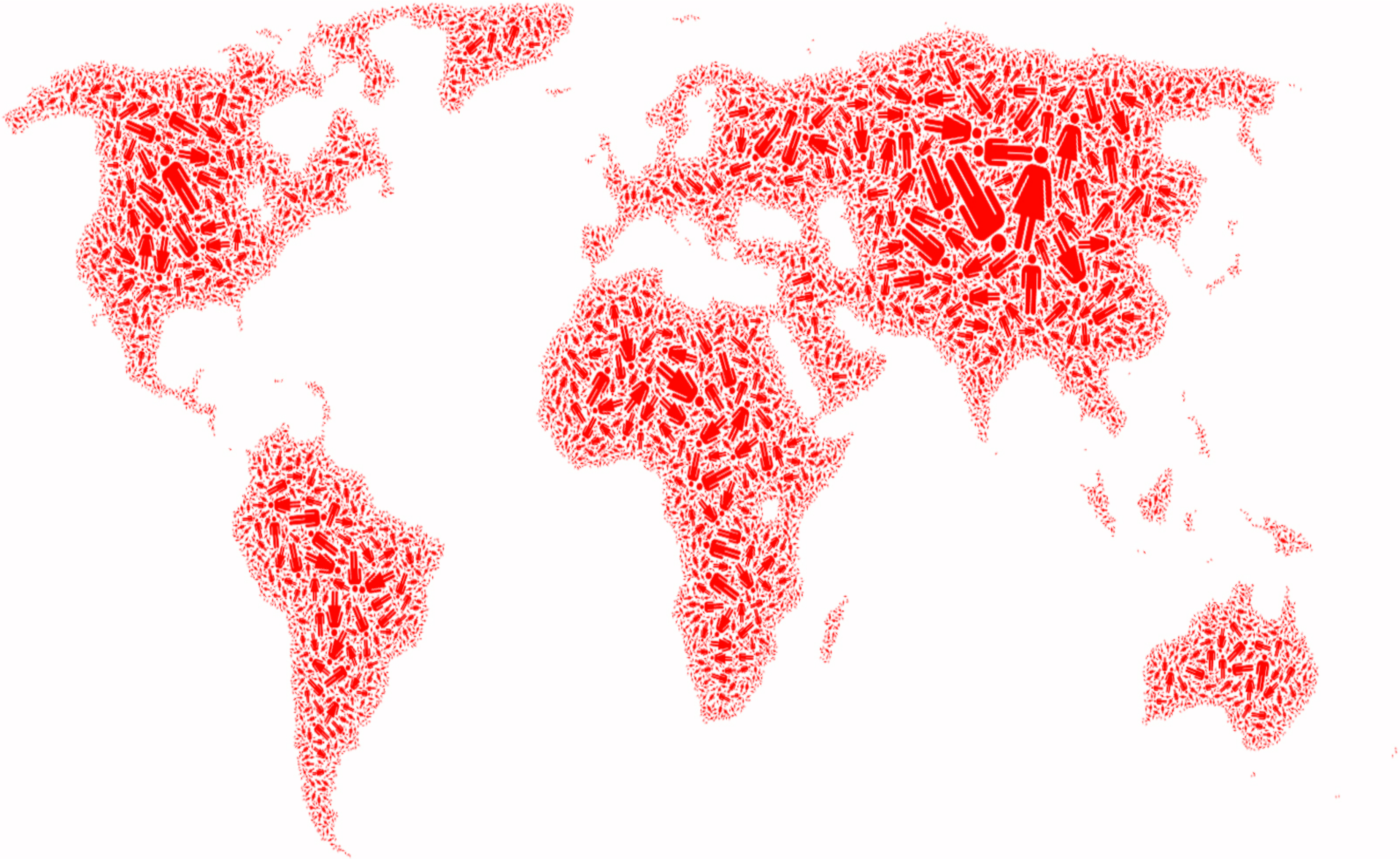
Source: John L. Casti, Encyclopædia Britannica

Complexity

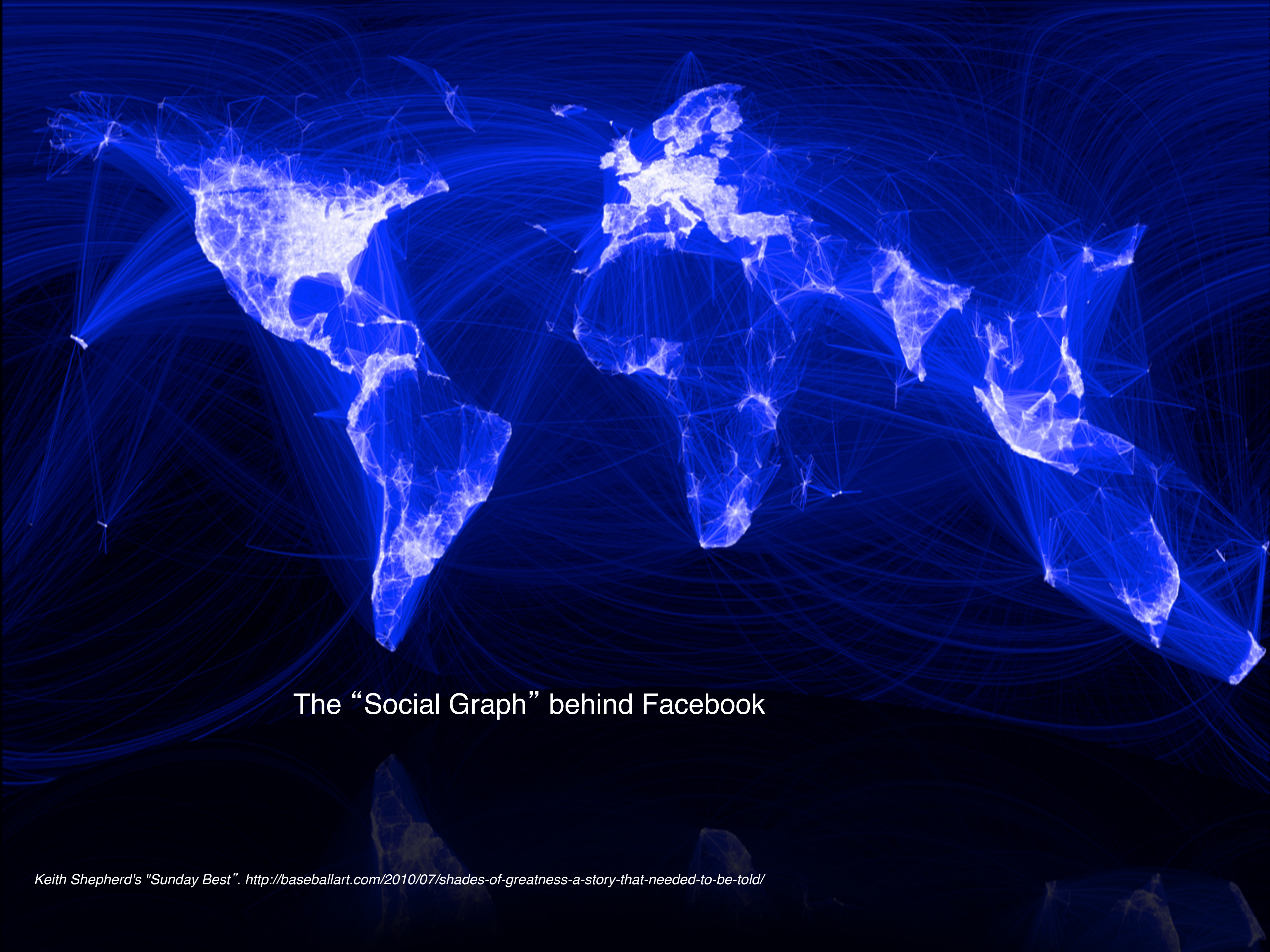
Emergent behavior: segregation

Behind each complex system there is a **network**, that defines the interactions between the components.

Social, informational, technological, biological networks

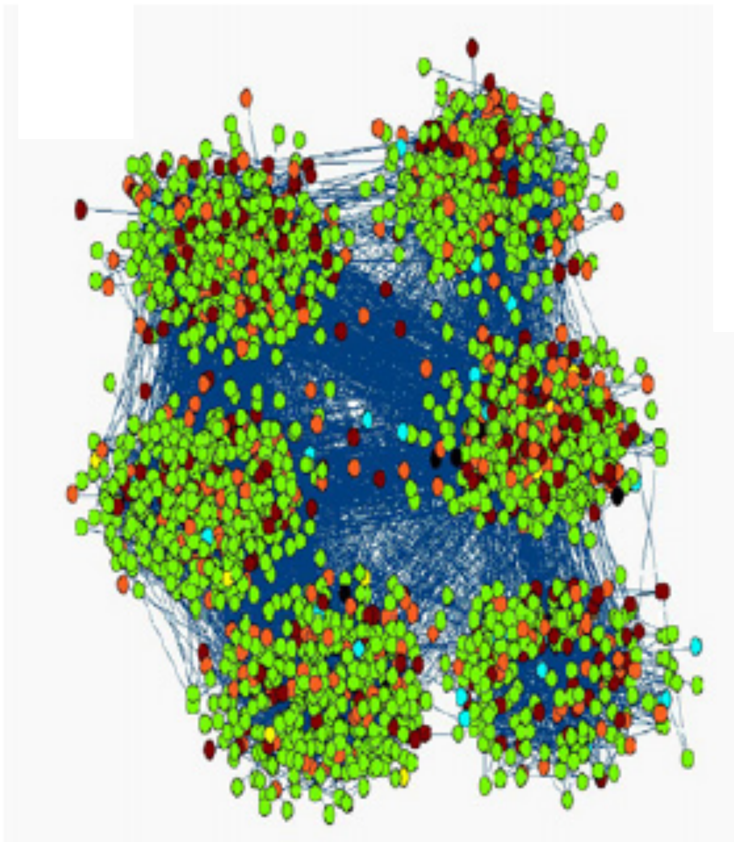
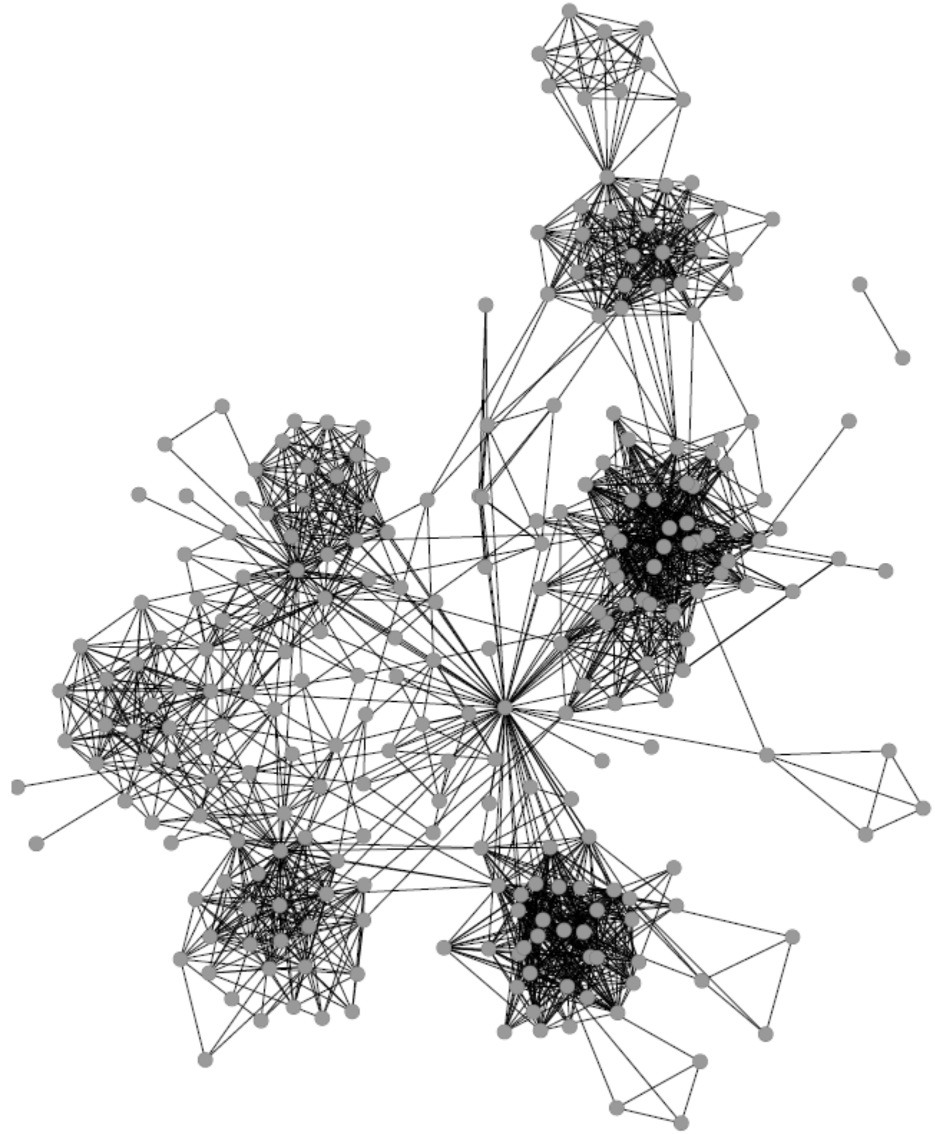


The "Day of 7 Billion“ has been in October 2011

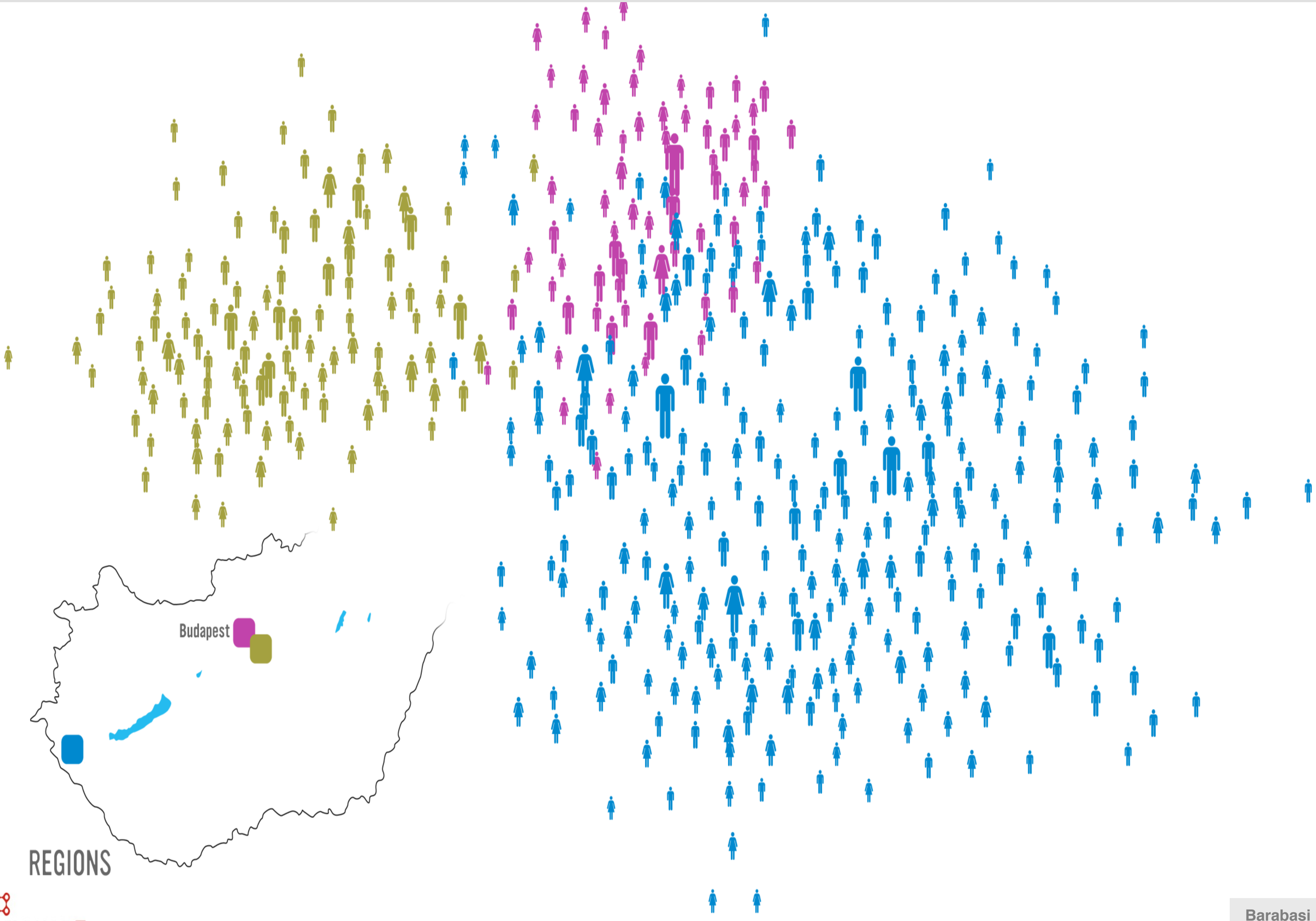


The “Social Graph” behind Facebook

Keith Shepherd's "Sunday Best”. http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

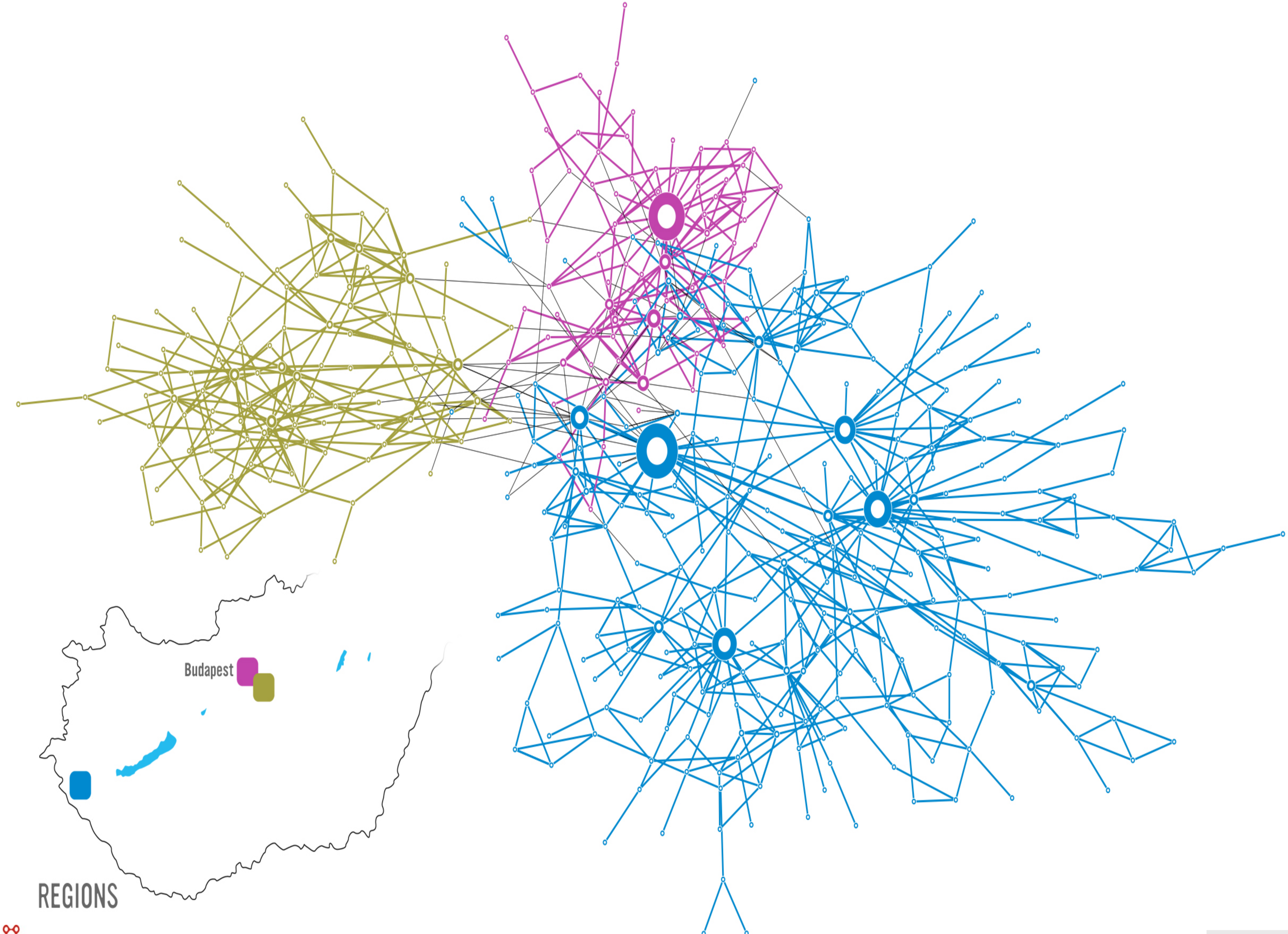



Mapping Organizations



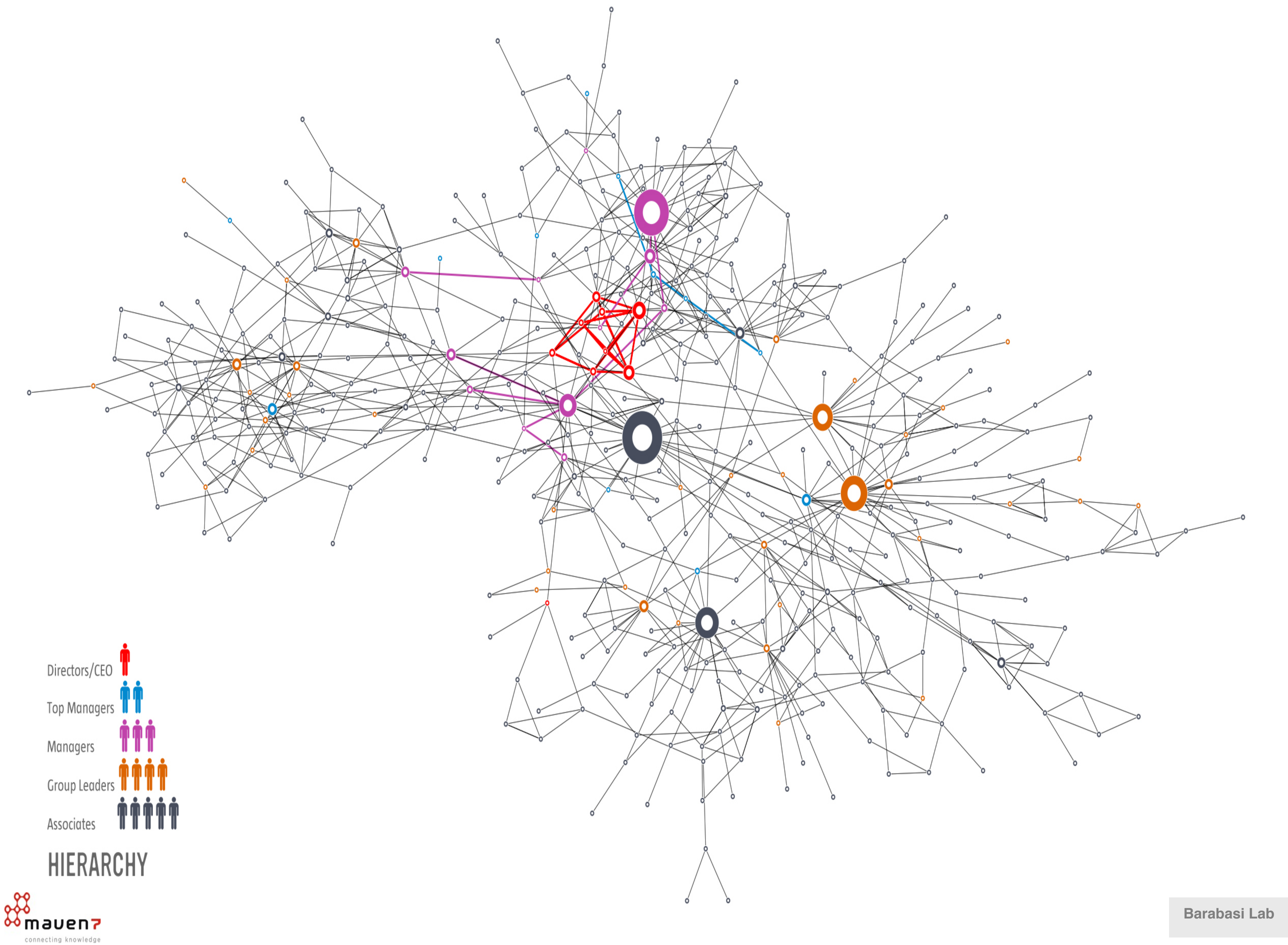
maven7
connecting knowledge

Barabasi Lab

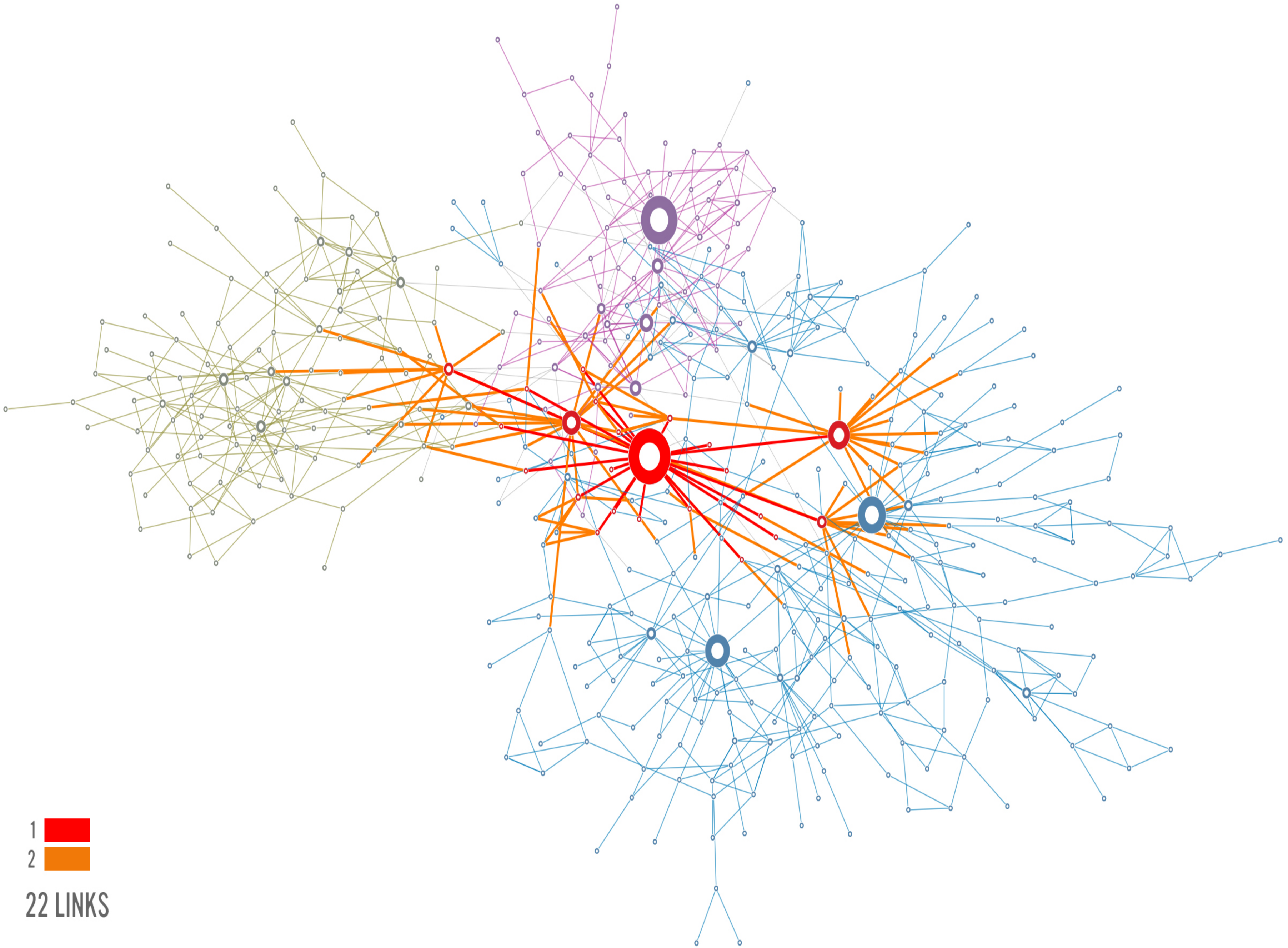


maven7
connecting knowledge

Barabasi Lab



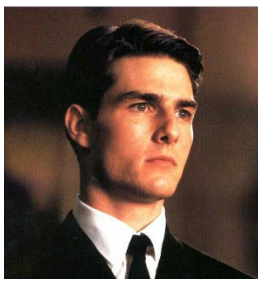
Directors/CEO
Top Managers
Managers
Group Leaders
Associates
HIERARCHY
maven7
connecting knowledge
Barabasi Lab


1
2
22 LINKS

COLLABORATION NETWORKS: ACTOR NETWORK

Nodes: actors
Links: cast jointly

IMDb Internet Movie Database

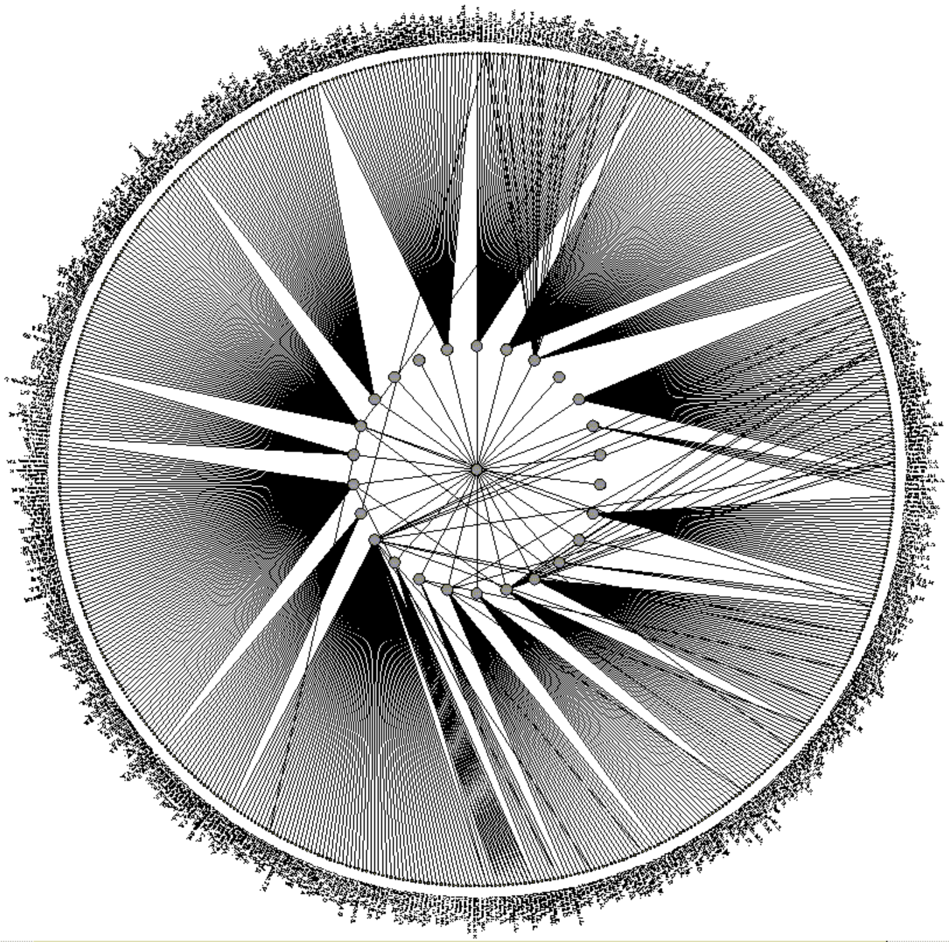


Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)

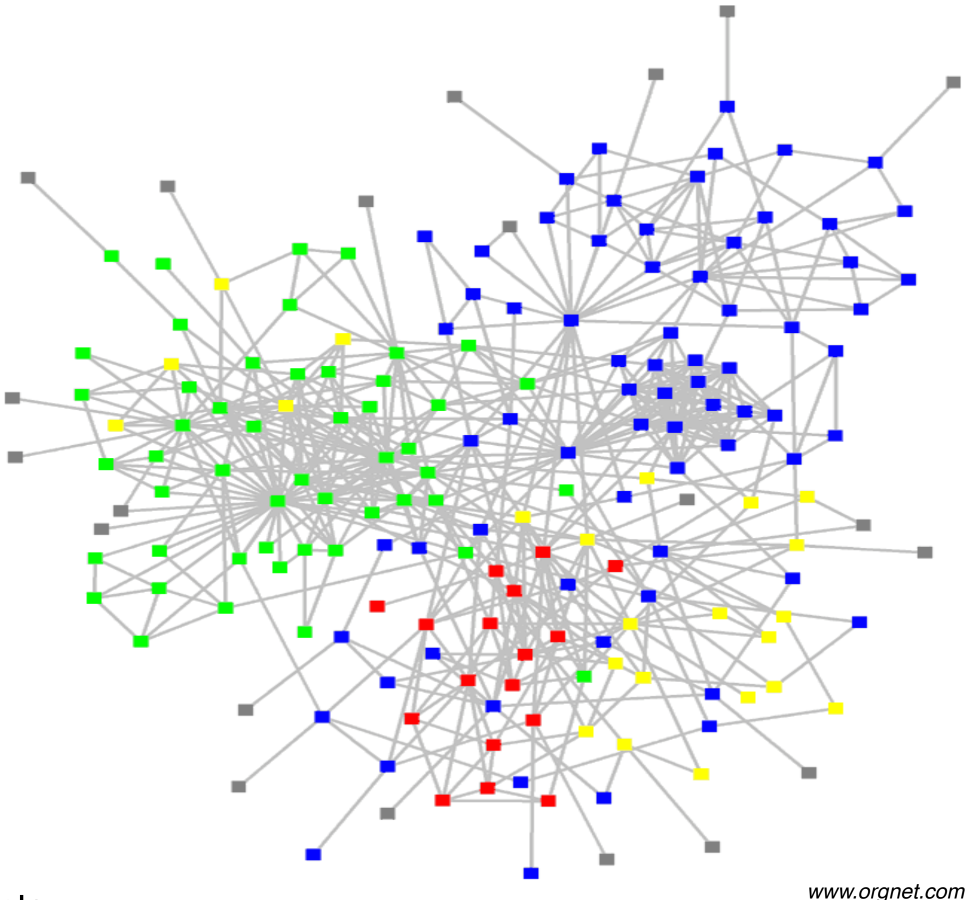
N = 212,250 actors $\langle k \rangle$ =28.78

COLLABORATION NETWORKS: SCIENCE CO-AUTHORSHIP

Nodes: scientist (authors)
Links: write paper together



STRUCTURE OF AN ORGANIZATION



: departments

: consultants

: external experts

BUSINESS TIES IN US BIOTECH-INDUSTRY

Nodes:

Companies

Investment

Pharma

Research Labs

Public

Biotechnology

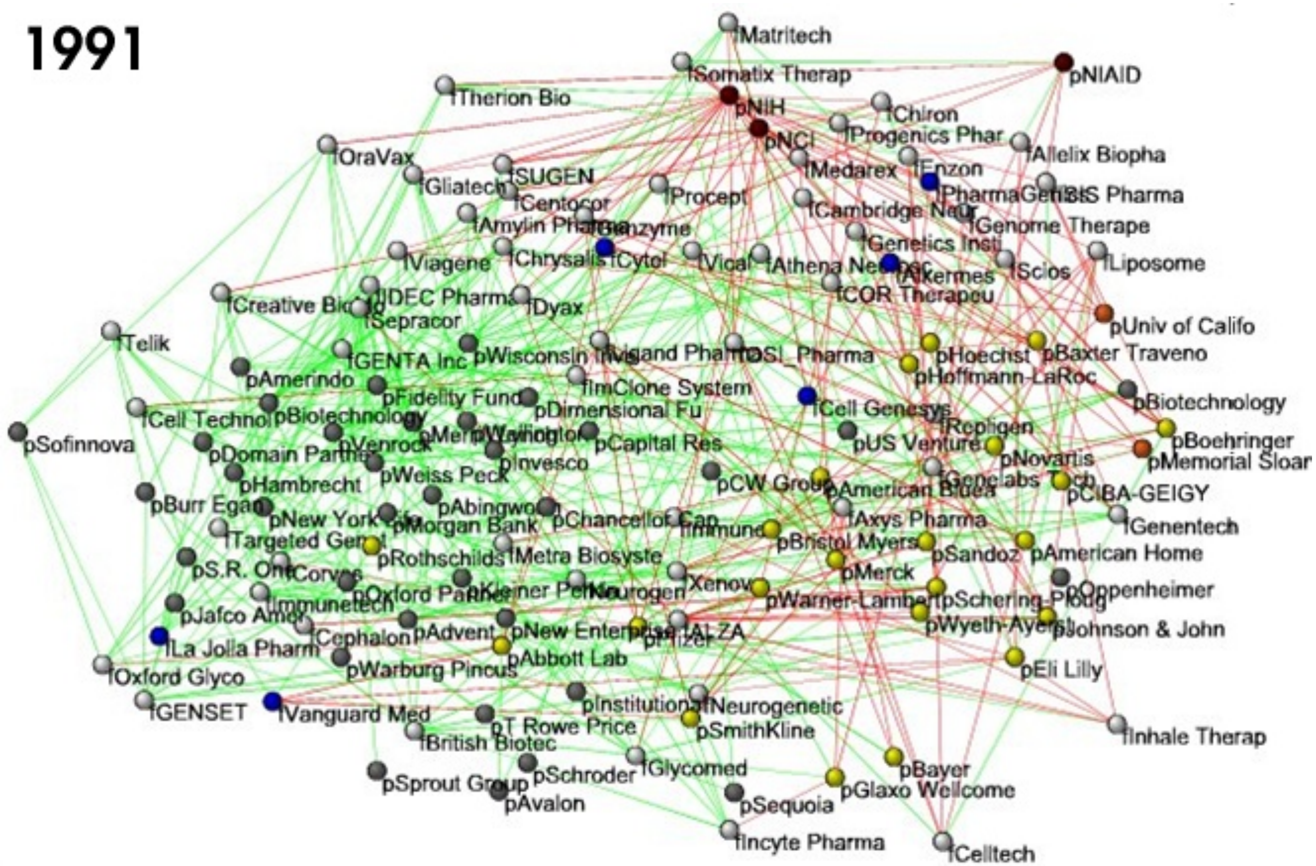
Links:

Collaborations

Financial

R&D

1991



http://ecclectic.ss.uci.edu/~drwhite/Movie

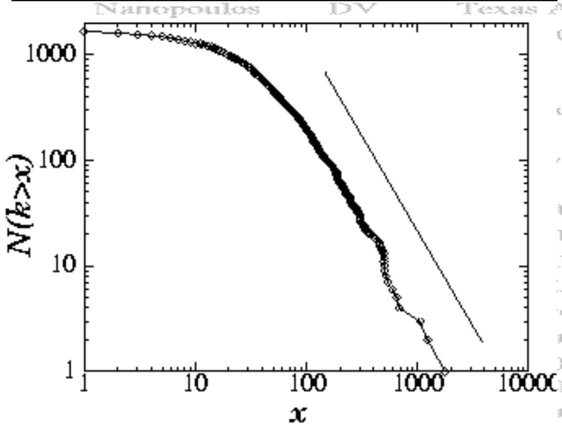
Information networks: the Web and Science Citation Indexes

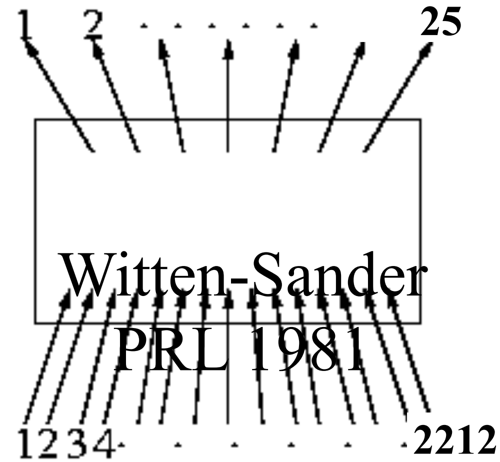
1,000 Most Cited Physicists

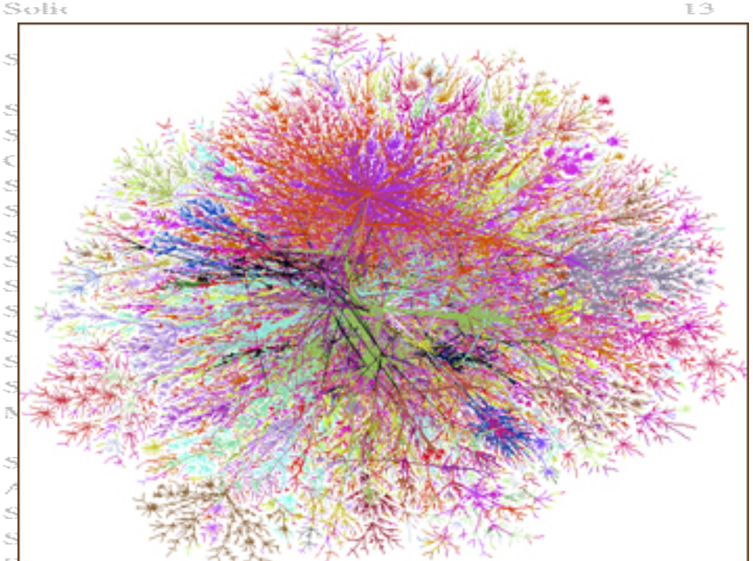
* citation total may be skewed because of multiple authors with the same name

Nodes: papers
Links: citations

1736 PRL papers (1988)

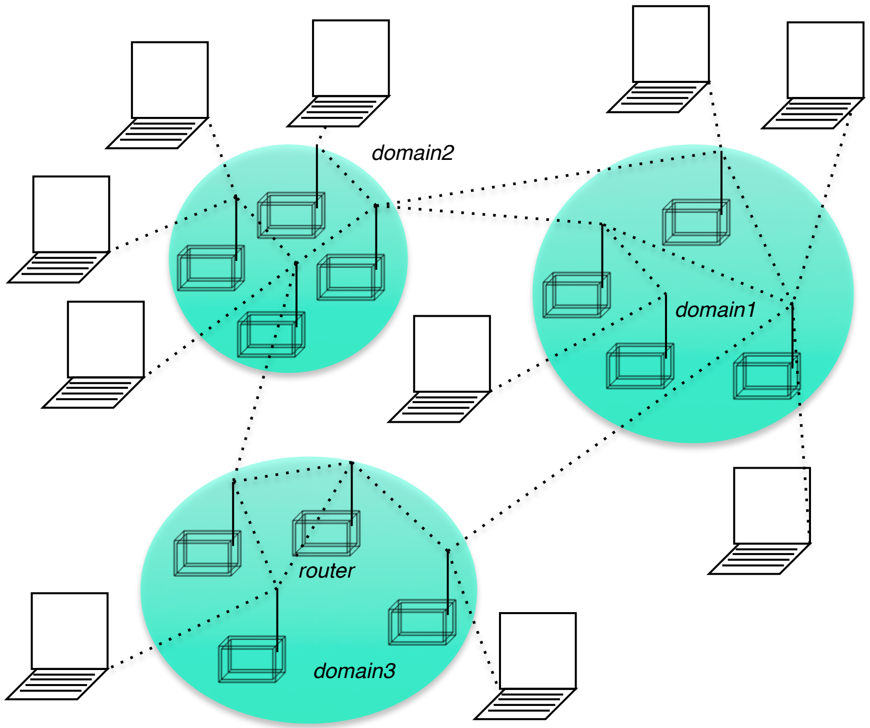


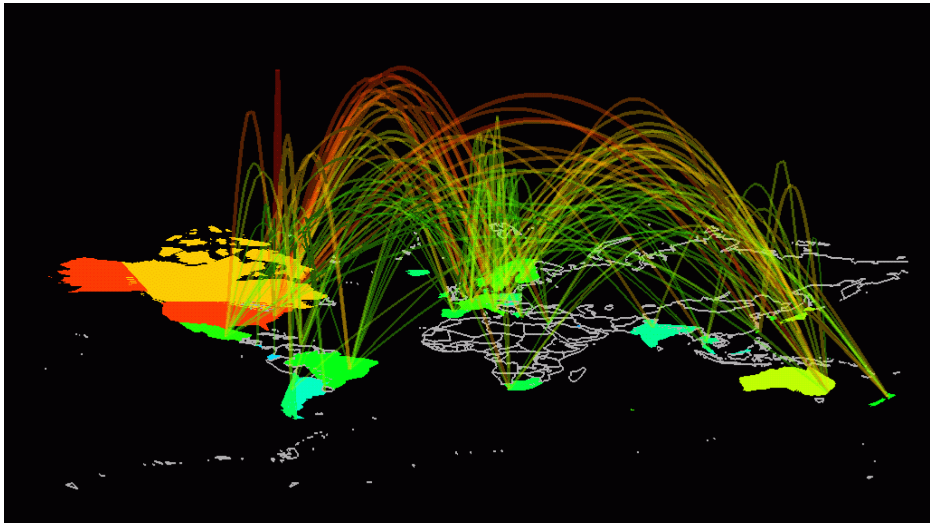


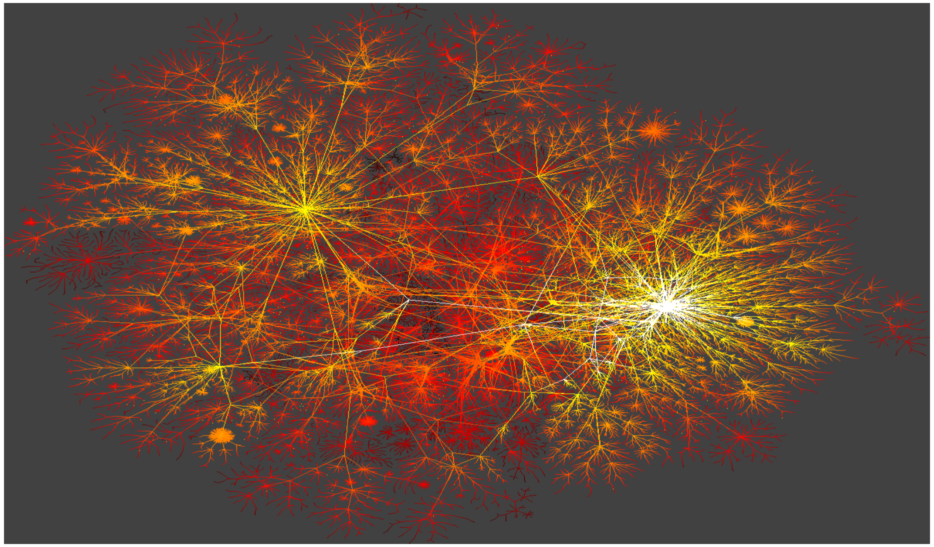


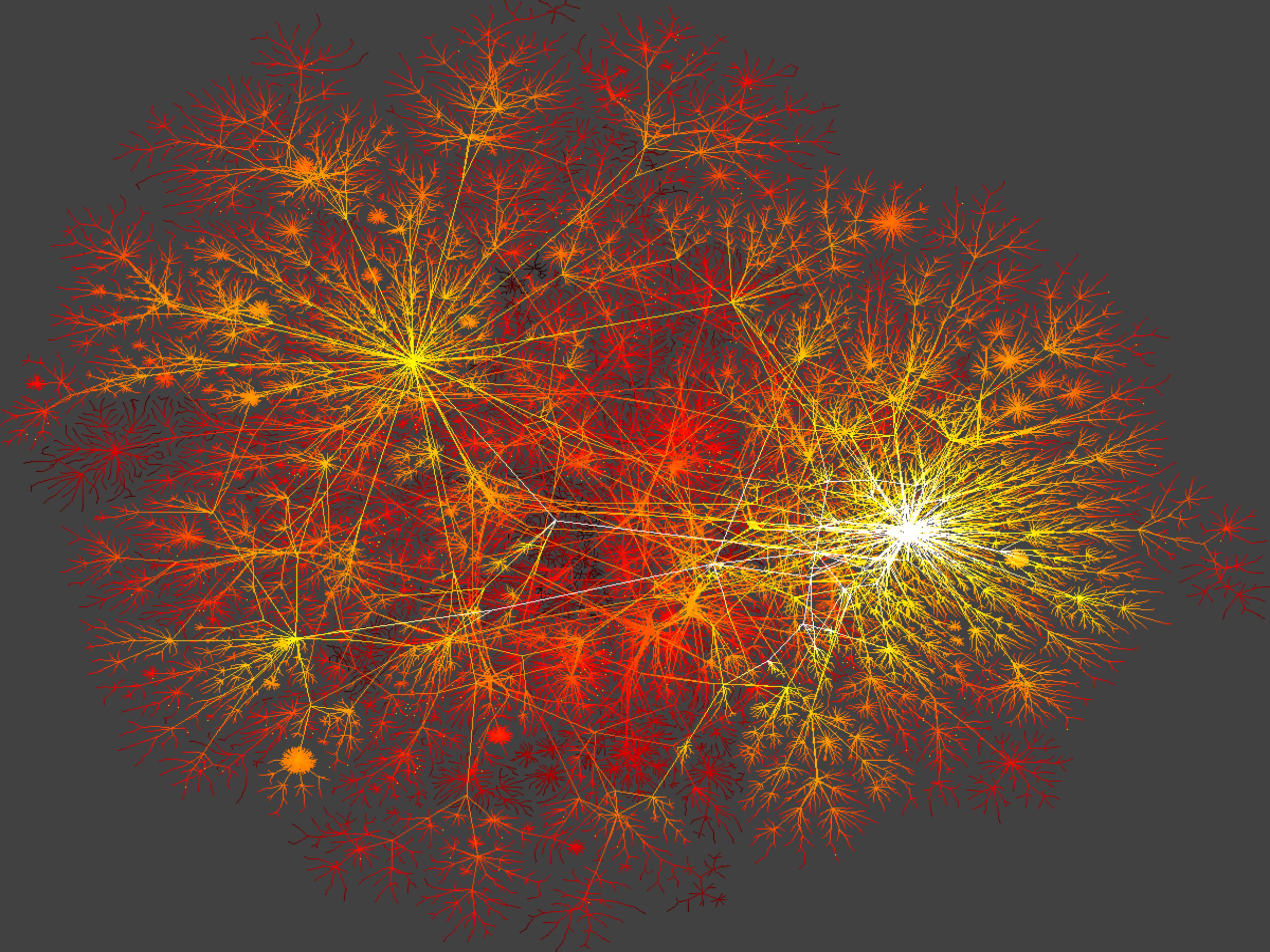
Nodes: web pages
Links: ditto ;-)

INTERNET

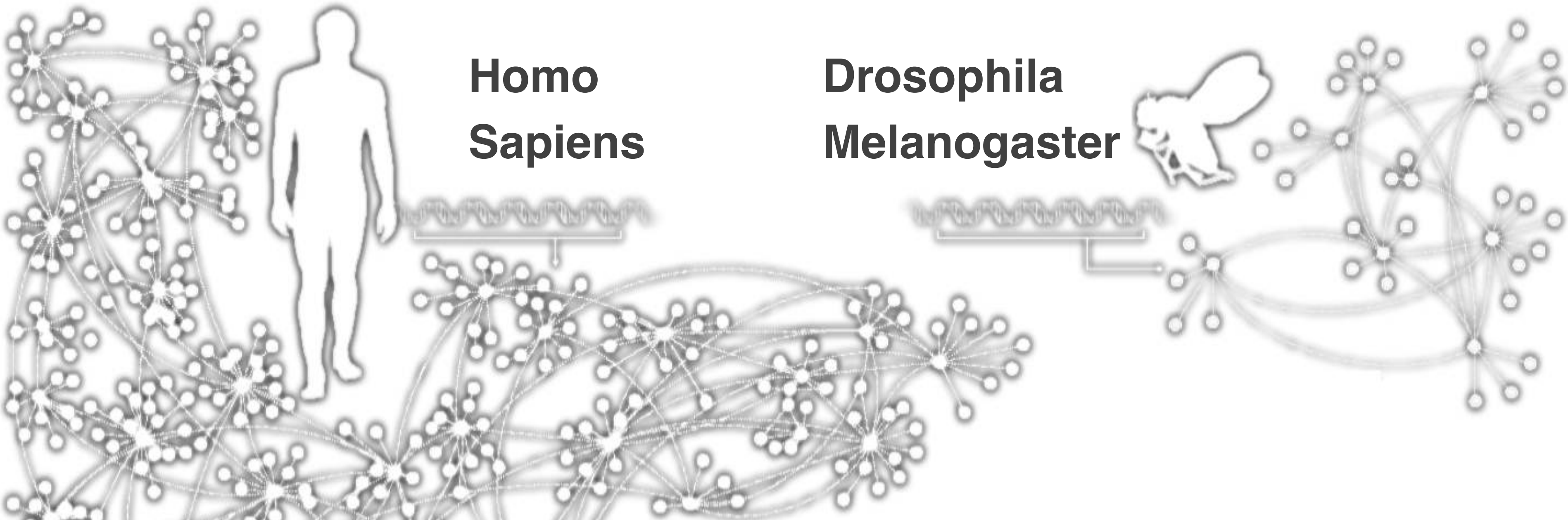


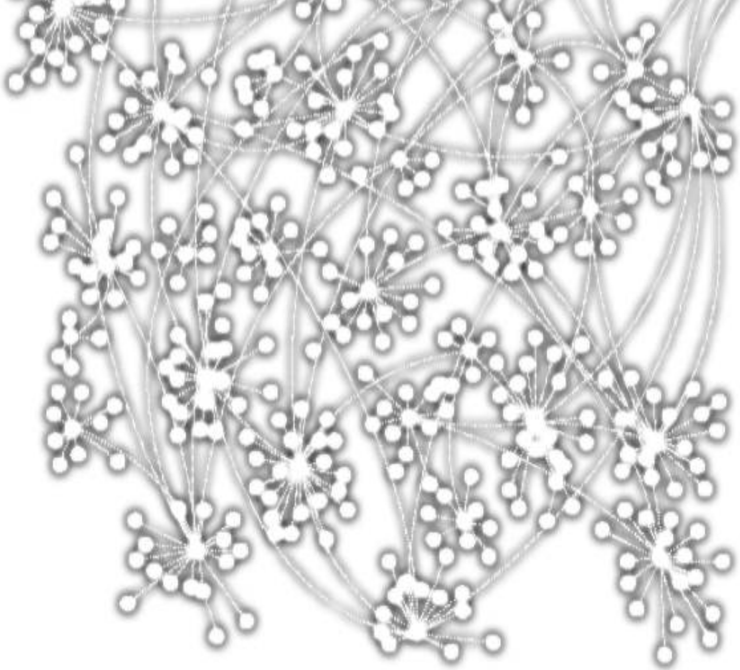






HUMANS GENES



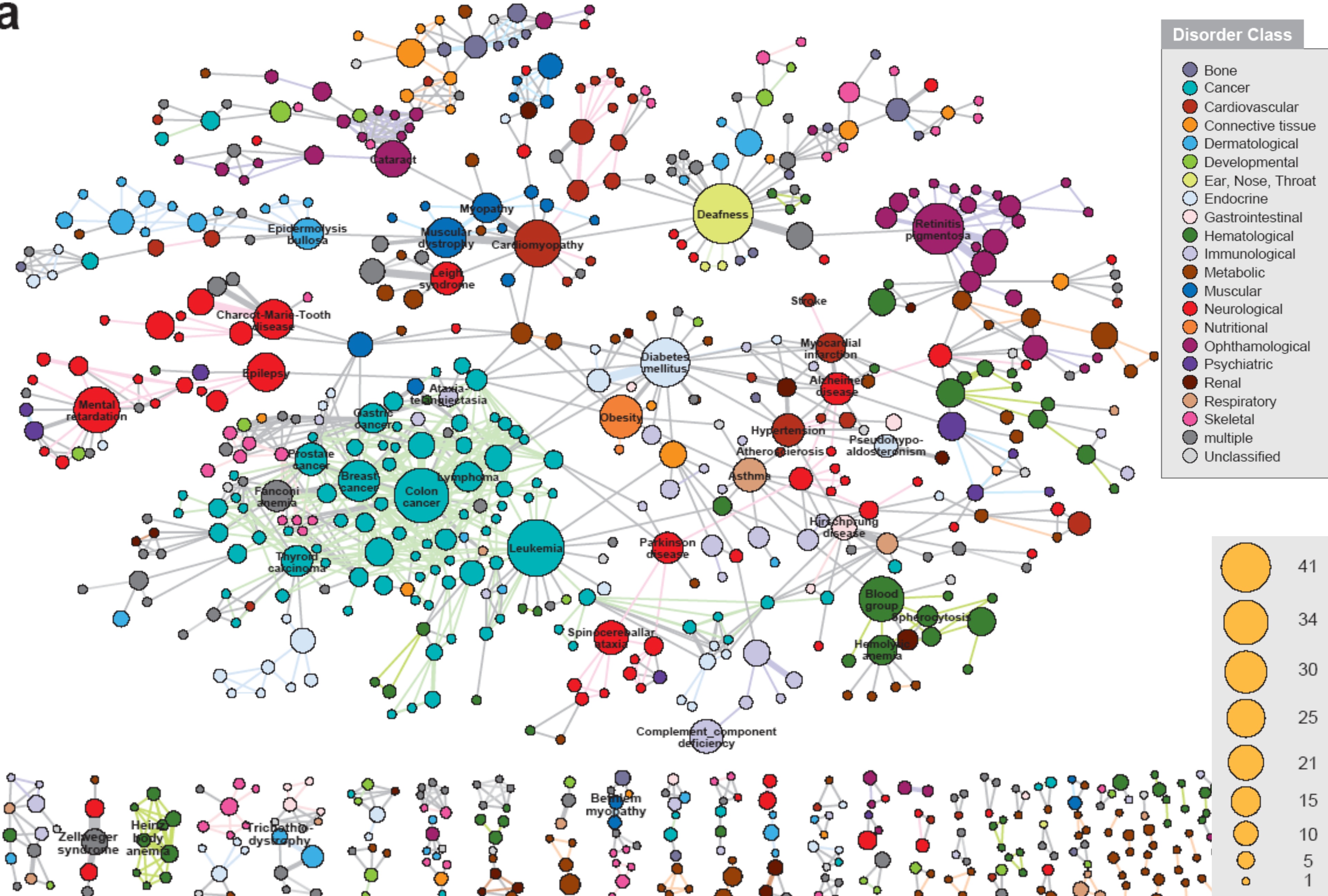


Complex systems

Made of many non-identical **elements** connected by diverse **interactions**.

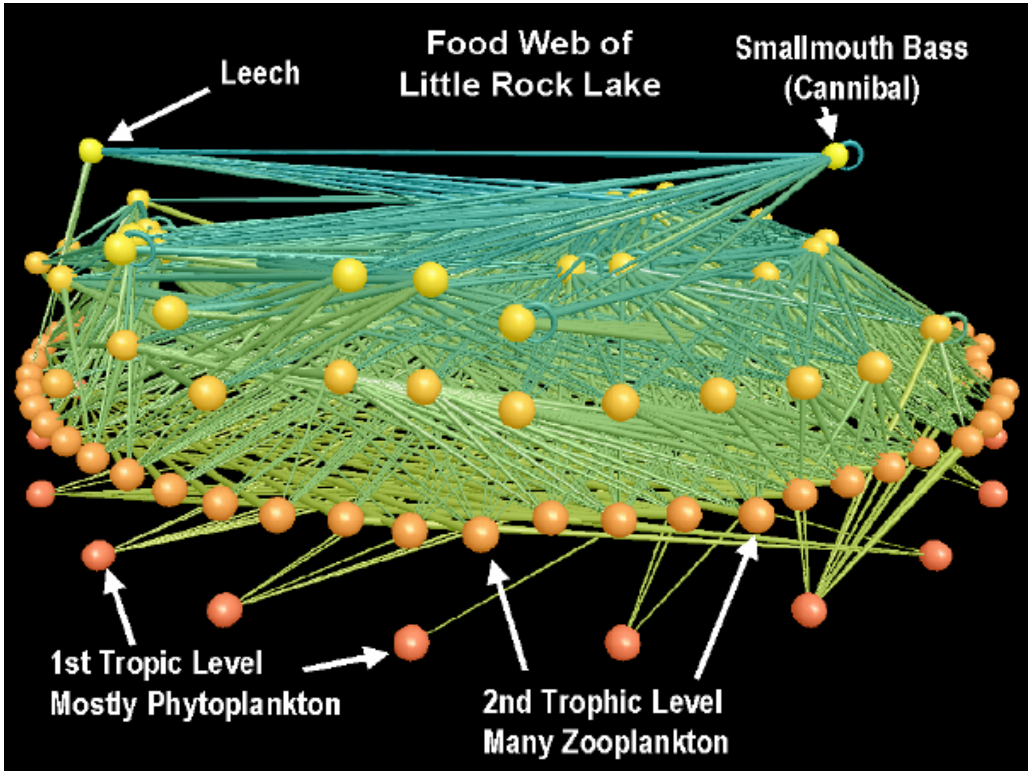
NETWORK

HUMAN DISEASE NETWORK



Biological networks: Food Web

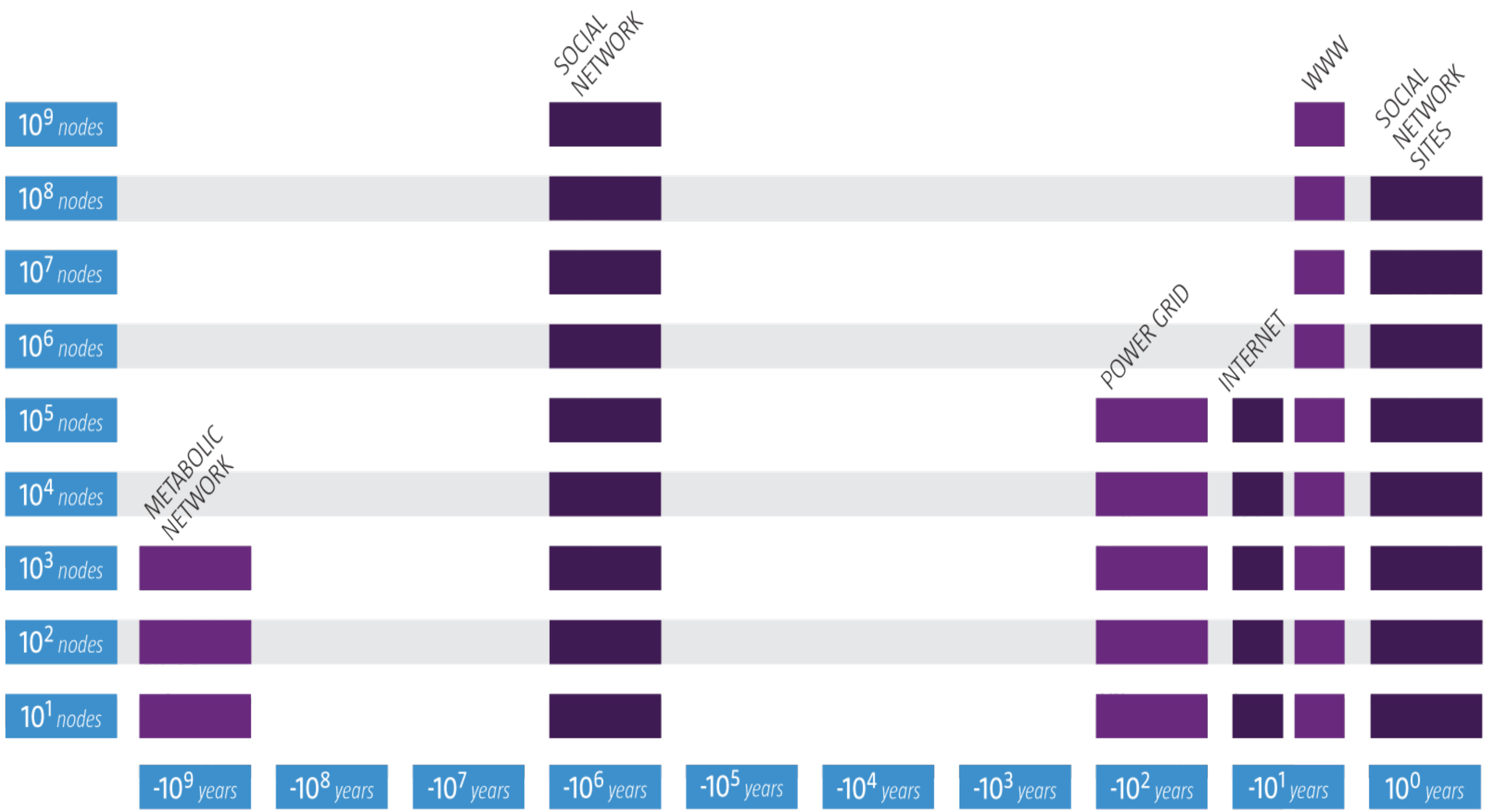
Nodes: species
Links: trophic interactions



R. Sole (cond-mat/0011195)

R.J. Williams, N.D. Martinez *Nature* (2000)

THE LIFE OF NETWORKS



THE EMERGENCE OF NETWORK SCIENCE

Data Availability:

Movie Actor Network, 1998;
World Wide Web, 1999.
C elegans neural wiring diagram 1990
Citation Network, 1998
Metabolic Network, 2000;
PPI network, 2001

Universality:

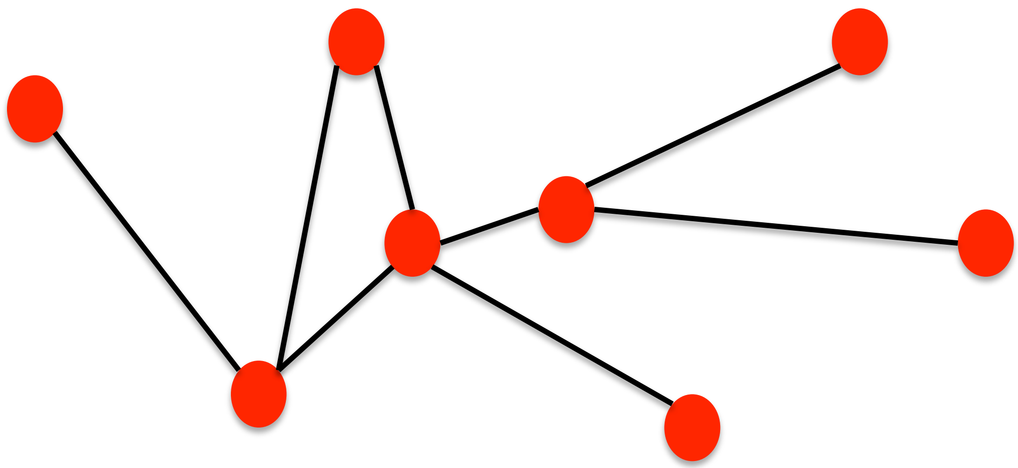
The architecture of networks emerging in various domains of science, nature, and technology are more similar to each other than one would have expected.

The (urgent) need to understand complexity:

Despite the challenges complex systems offer us, we cannot afford to not address their behavior, a view increasingly shared both by scientists and policy makers. Networks are not only essential for this journey, but during the past decade some of the most important advances towards understanding complexity were provided in context of network theory.

Networks and graphs

COMPONENTS OF A COMPLEX SYSTEM



- **components**: nodes, vertices — N
- **interactions**: links, edges — L
- **system**: network, graph — (N,L)

NETWORKS OR GRAPHS?

network often refers to real systems

- •www,
- •social network
- •metabolic network.

Language: (Network, node, link)

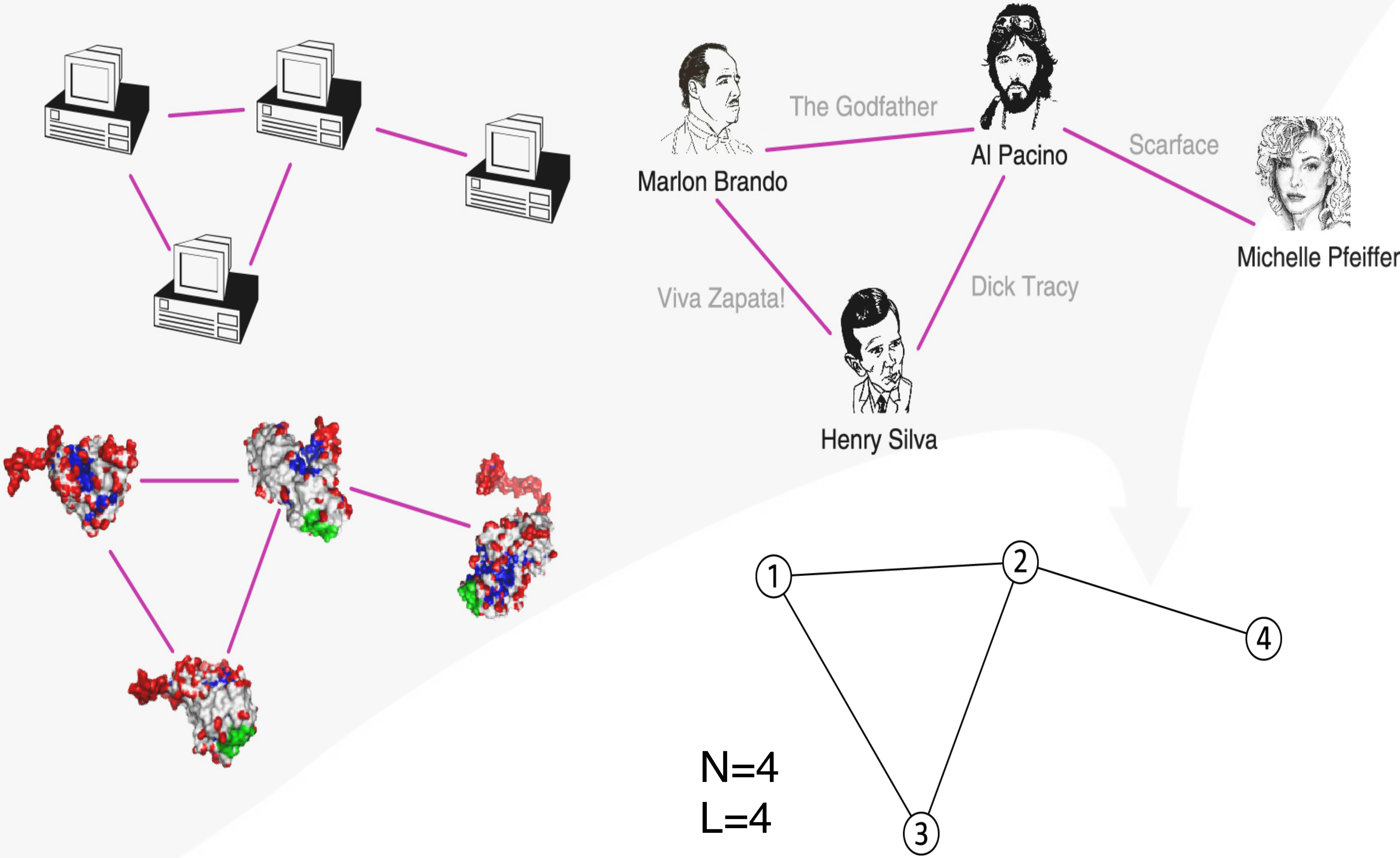
graph: mathematical representation of a network

- •web graph,
- •social graph (a Facebook term)

Language: (Graph, vertex, edge)

We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably.

A COMMON LANGUAGE

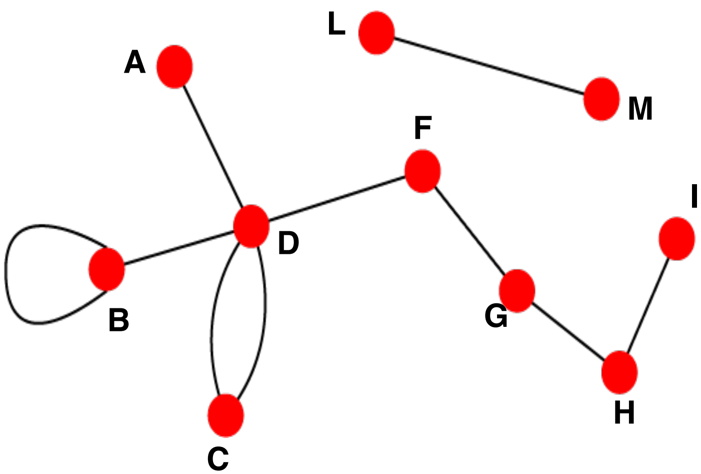


UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:

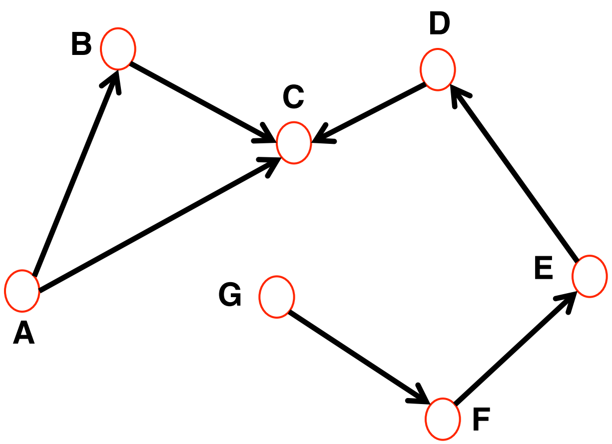


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



An undirected link is the superposition of two opposite directed links.

Directed links :
URLs on the www
phone calls
metabolic reactions

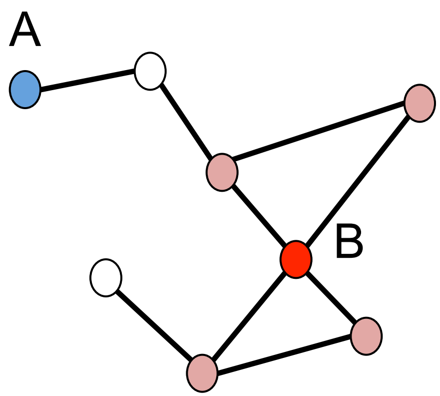
Reference Networks

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

Degree, Average Degree and Degree Distribution

NODE DEGREES

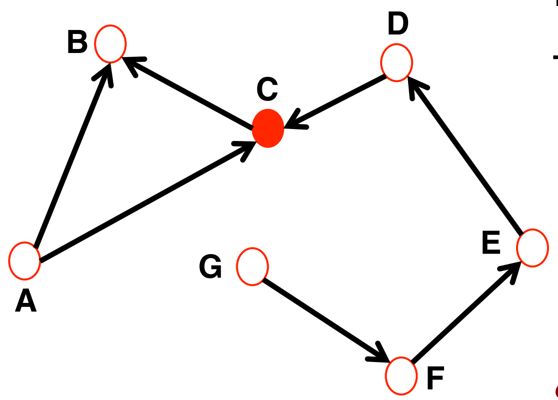
Undirected



Node degree: the number of links connected to the node.

$$k_A = 1 \qquad k_B = 4$$

Directed



In *directed networks* we can define an in-degree and out-degree.

The (total) degree is the sum of in- and out-degree.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

Source: a node with $k^{in} = 0$; Sink: a node with $k^{out} = 0$.

A BIT OF STATISTICS

BRIEF STATISTICS REVIEW

Four key quantities characterize a sample of N values $x_1, \dots, x_N$:

Average (mean):

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The n^{th} moment:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

Standard deviation:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

Distribution of x:

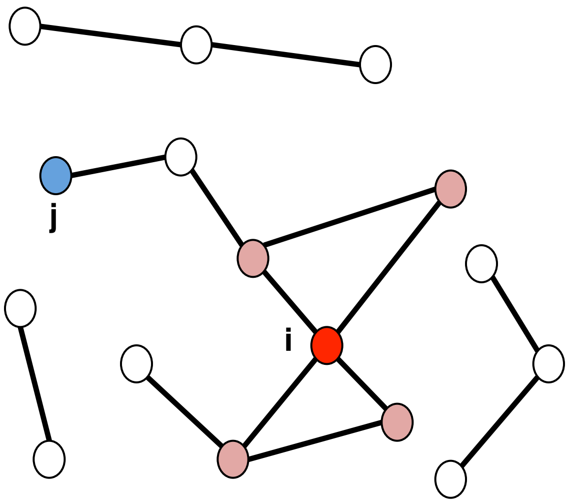
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

where p_x follows

$$\sum_i p_x = 1 \left(\int p_x \, dx = 1\right)$$

AVERAGE DEGREE

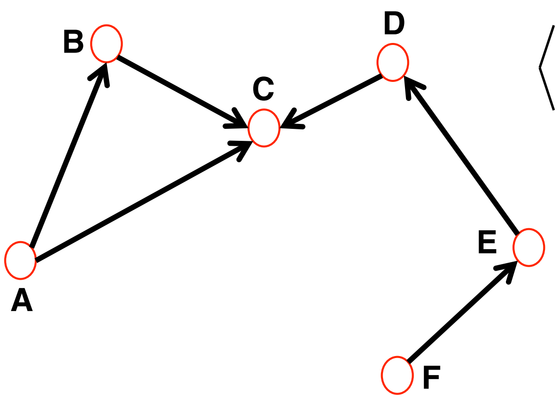
Undirected



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – the number of nodes in the graph

Directed



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

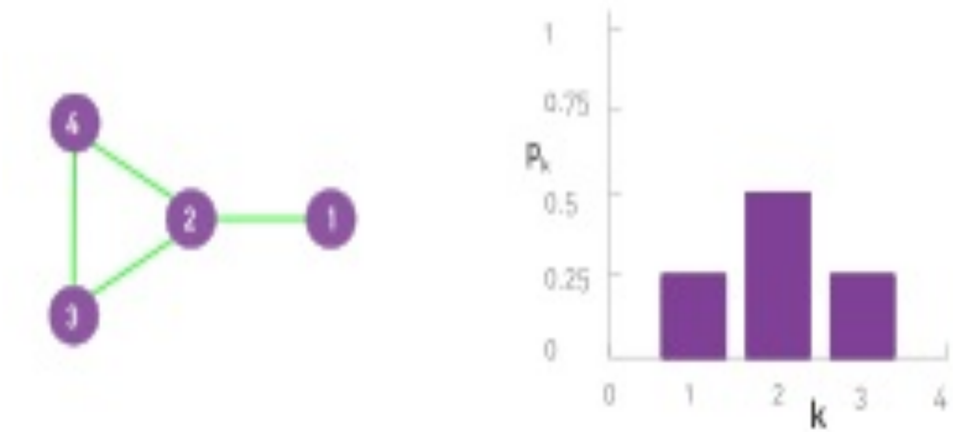
$$\langle k \rangle \equiv \frac{L}{N}$$

Average Degree

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

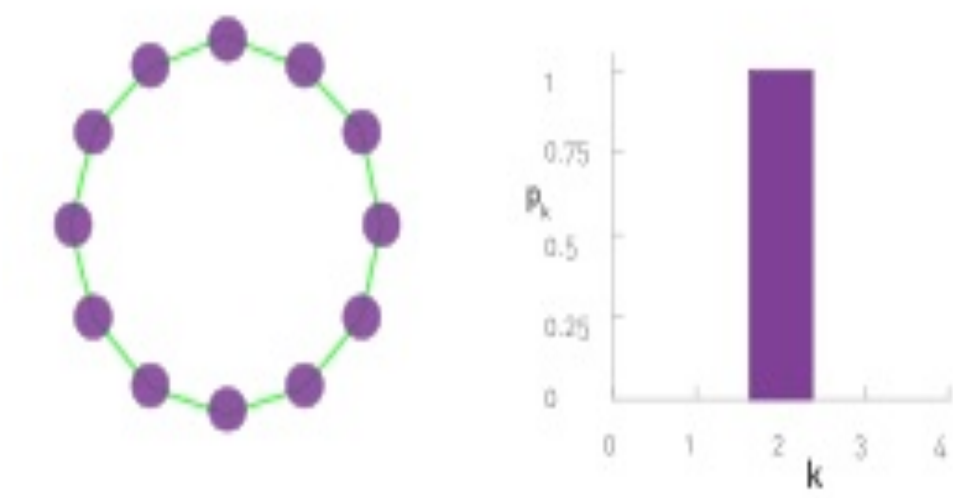
Degree distribution

P(k): probability that a randomly chosen node has degree *k*



N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot



DEGREE DISTRIBUTION

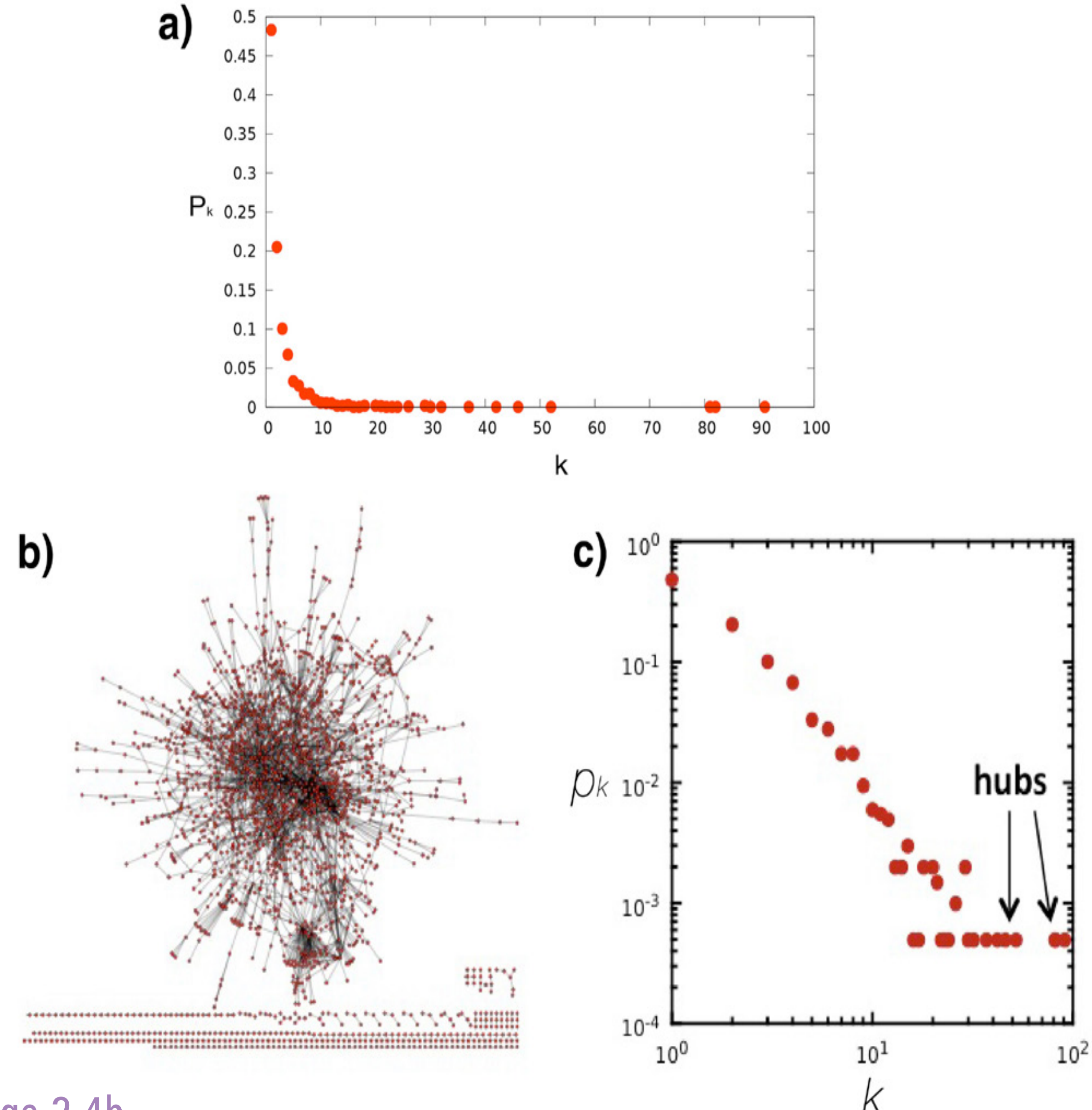
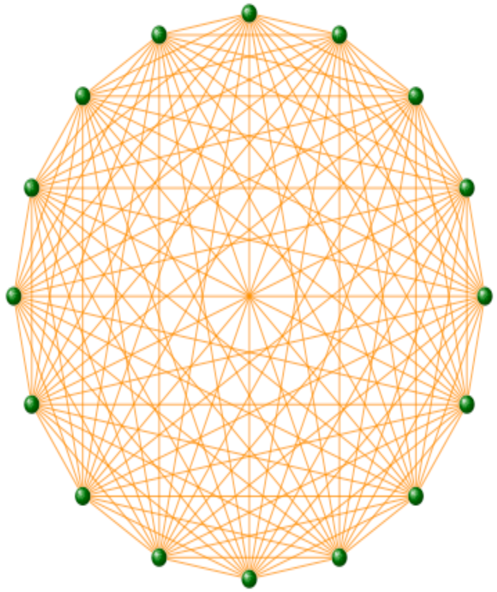


Image 2.4b

Real networks are sparse

COMPLETE GRAPH

The maximum number of links a network of N nodes can have is: $L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$



A graph with degree $L=L_{max}$ is called a complete graph, and its average degree is **<k>=N-1**

Most networks observed in real systems are sparse:

$$L \ll L_{max}$$

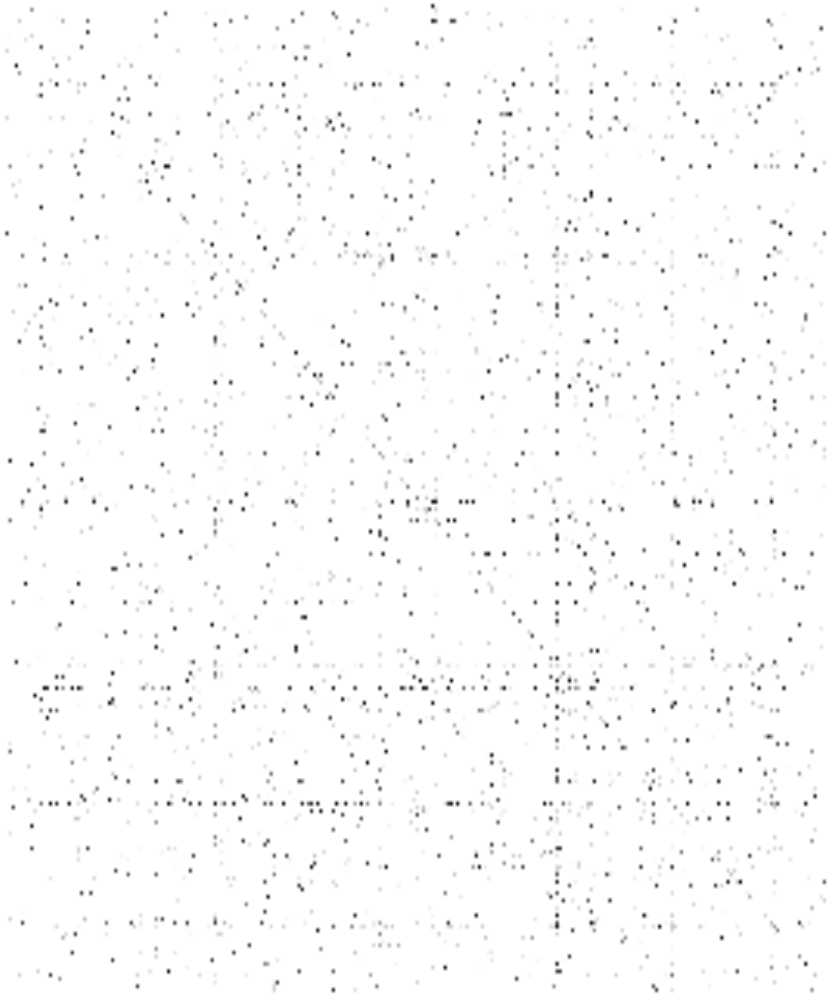
or

$$\langle k \rangle \ll N-1.$$

WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	$\langle k \rangle=4.51$
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^7$	$\langle k \rangle=2.39$
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	$\langle k \rangle=3.9$
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	$\langle k \rangle=28.78$

(Source: Albert, Barabasi, RMP2002)

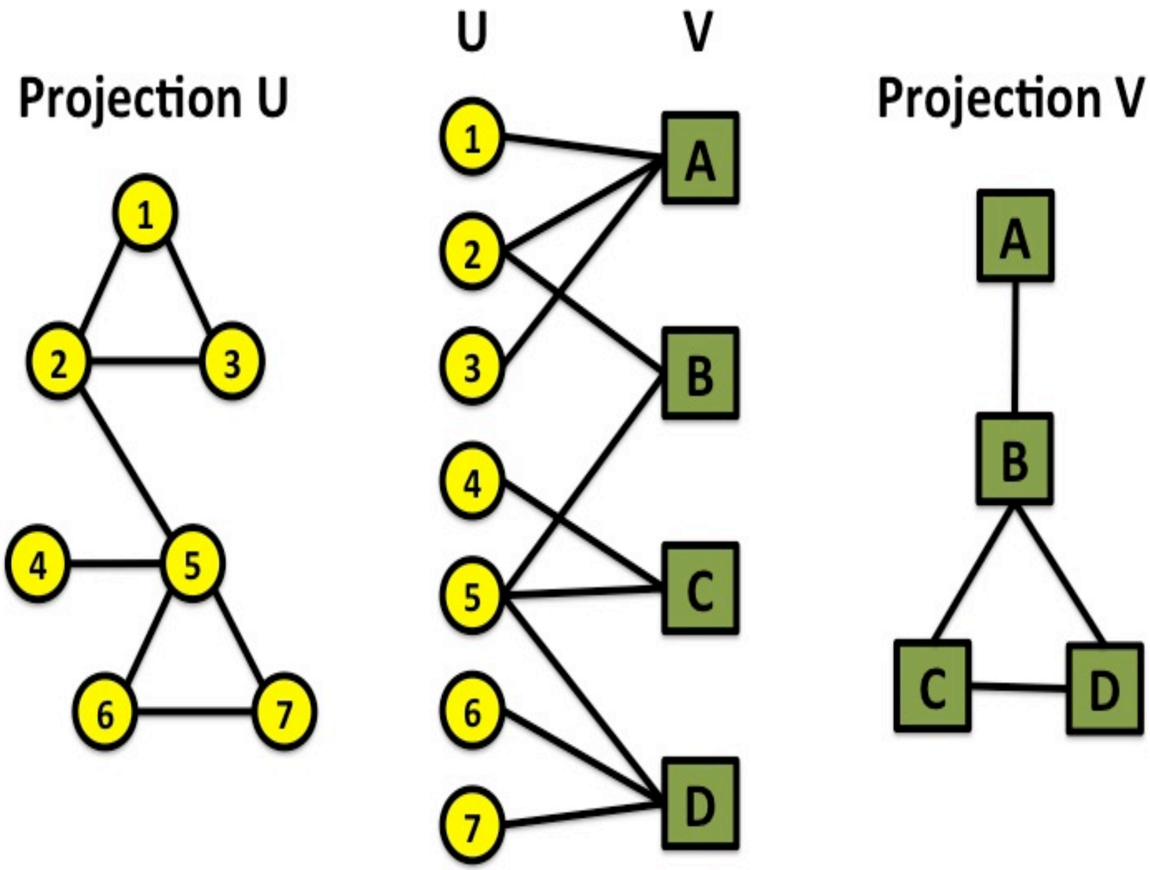
ADJACENCY MATRICES ARE SPARSE



BIPARTITE NETWORKS

BIPARTITE GRAPHS

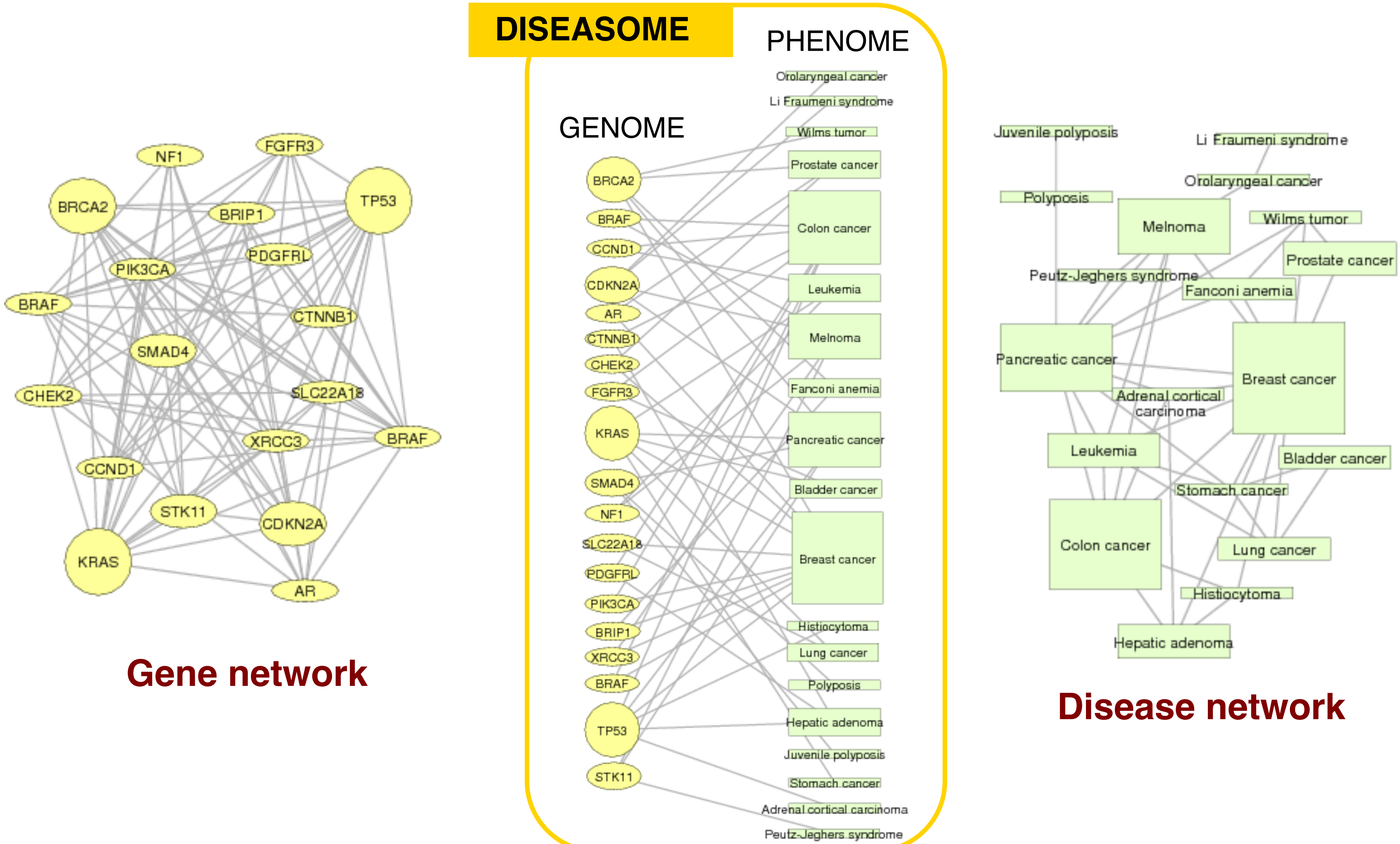
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



Examples:

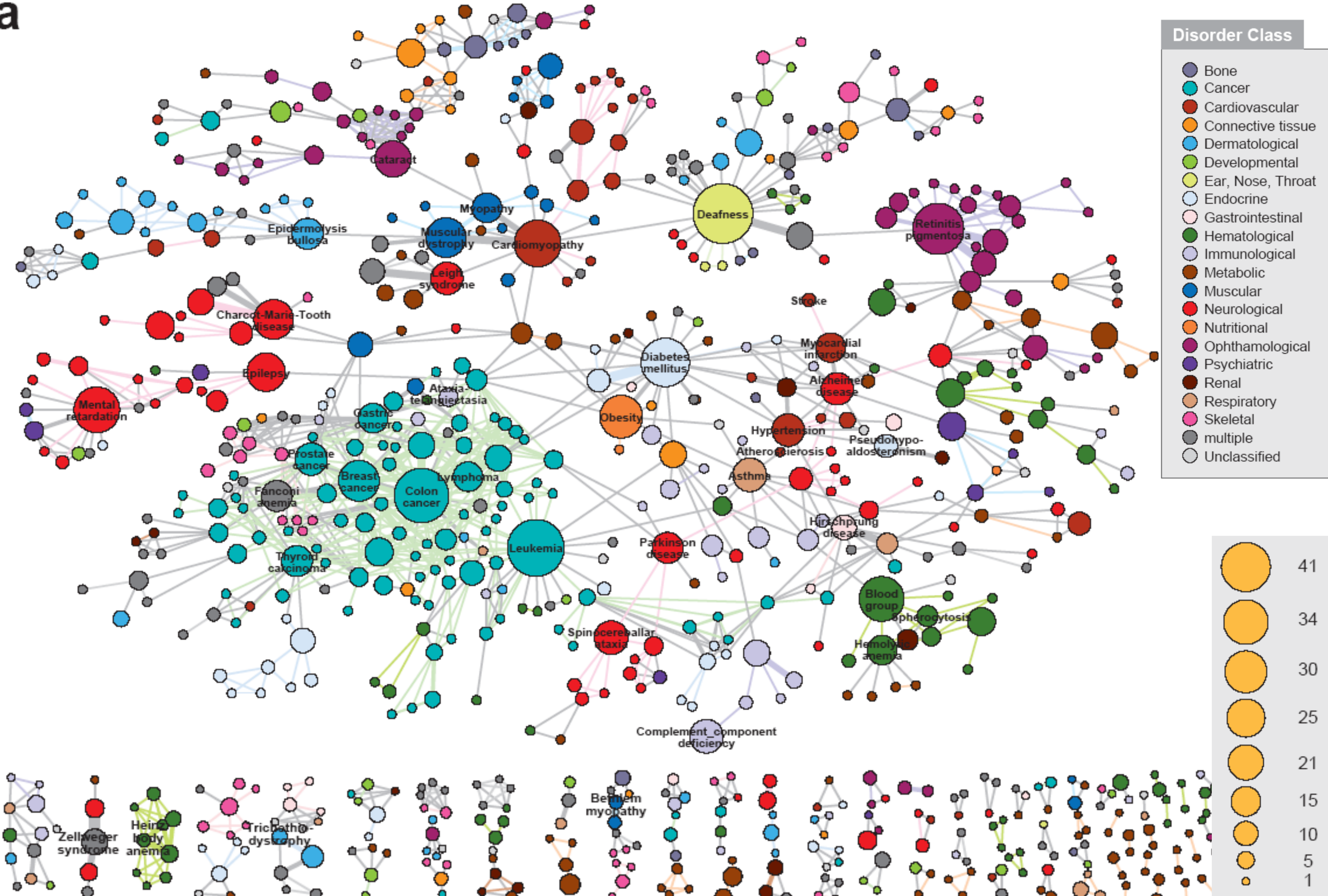
Hollywood actor network
Collaboration networks
Disease network (diseasome)

GENE NETWORK – DISEASE NETWORK

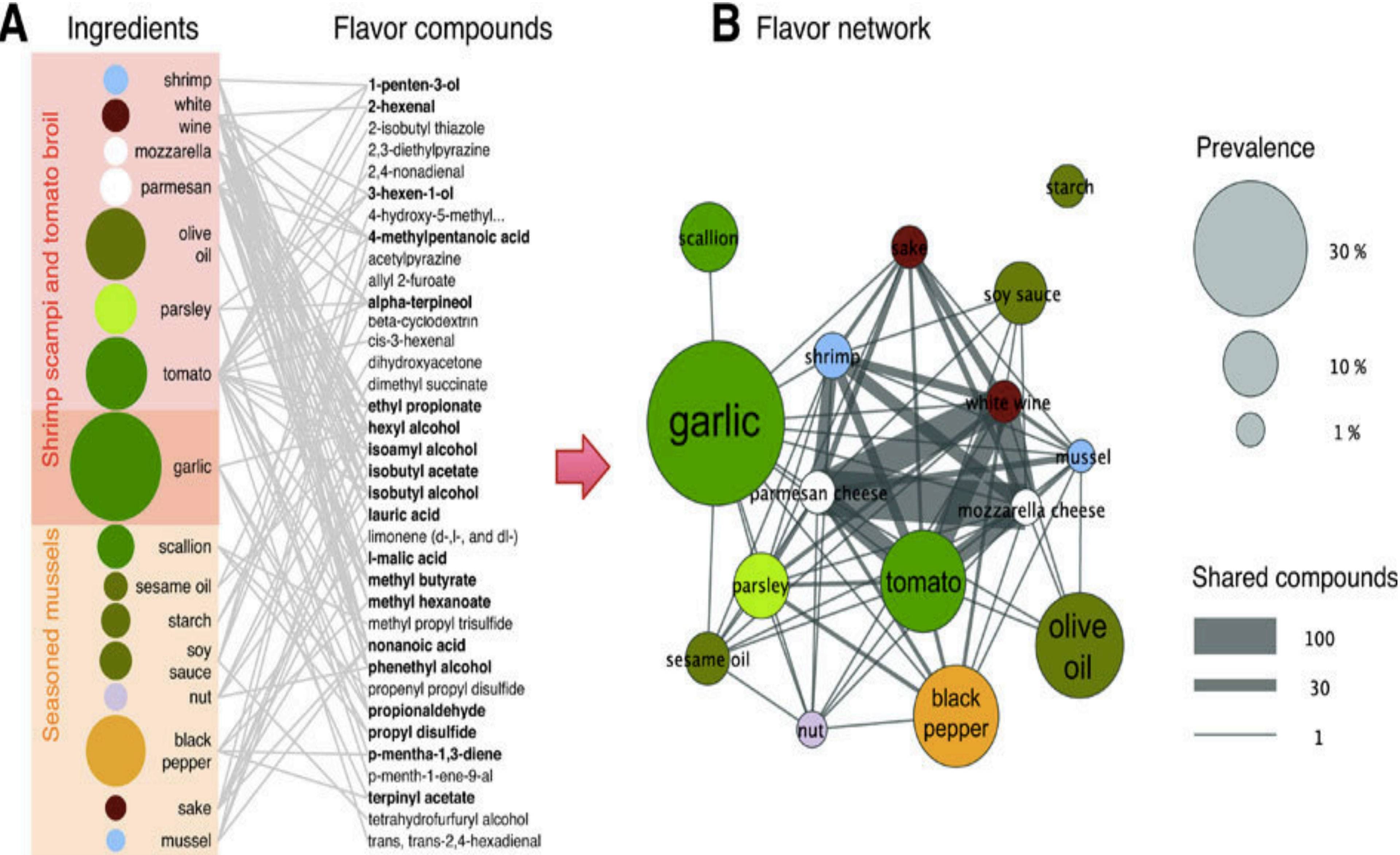


Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

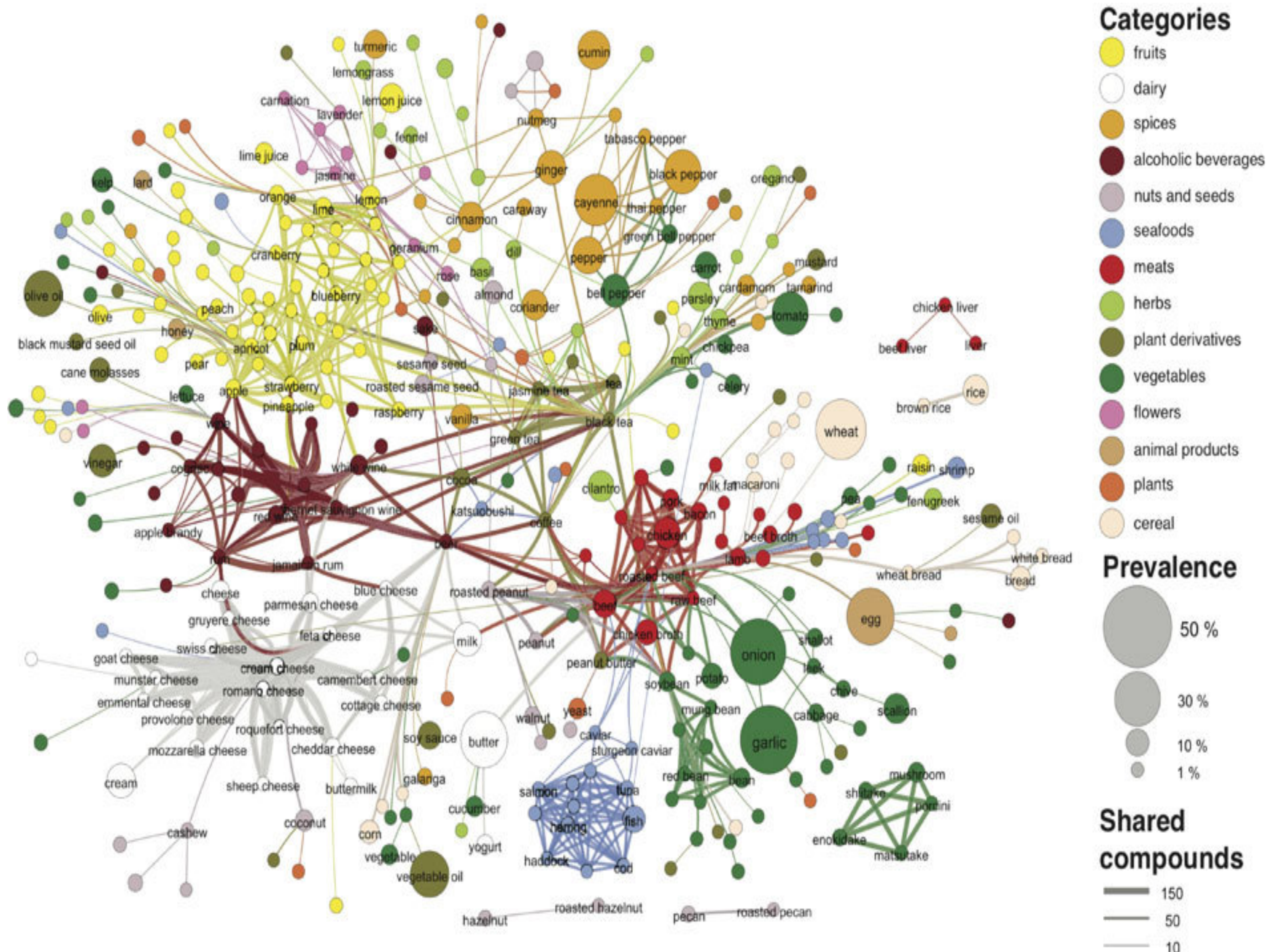
HUMAN DISEASE NETWORK



Ingredient-Flavor Bipartite Network



Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási
Flavor network and the principles of food pairing , Scientific Reports 196, (2011).


Categories
fruits
dairy
spices
alcoholic beverages
nuts and seeds
seafoods
meats
herbs
plant derivatives
vegetables
flowers
animal products
plants
cereal
Prevalence
50 %
30 %
10 %
1 %
Shared compounds
150
50
10
turmeric
lemongrass
cumin
carnation
lemon juice
lavender
nutmeg
fennel
tabasco pepper
lime juice
ginger
black pepper
kelp
lard
jasmine
oregano
orange
lemon
cayenne
lime
cinnamon
caraway
thai pepper
green bell pepper
cranberry
geranium
pepper
dill
mustard
carrot
rose
basil
cardamom
tamarind
olive oil
blueberry
almond
bell pepper
parsley
peach
coriander
tomato
olive
thyme
chicken liver
honey
black mustard seed oil
apricot
plum
chickpea
beef liver
liver
sesame seed
mint
cane molasses
pear
roasted sesame seed
jasmine tea
tea
celery
lettuce
apple
strawberry
rice
brown rice
pineapple
raspberry
vanilla
black tea
wine
wheat
green tea
vinegar
cognac
white wine
raisin
shrimp
cocoa
milk fat
macaroni
cilantro
pork
pea
fenugreek
red wine
cabernet sauvignon wine
katsuobushi
bacon
sesame oil
chicken
coffee
apple brandy
beef broth
white bread
beer
lamb
rum
jamaican rum
wheat bread
bread
roasted beef
cheese
blue cheese
roasted peanut
parmesan cheese
beef
raw beef
egg
gruyere cheese
feta cheese
milk
chicken broth
shallot
swiss cheese
peanut
onion
goat cheese
cream cheese
peanut butter
leek
munster cheese
camembert cheese
soybean
potato
chive
emmental cheese
romano cheese
cottage cheese
mung bean
scallion
walnut
yeast
cabbage
provolone cheese
roquefort cheese
soy sauce
butter
caviar
garlic
mozzarella cheese
cheddar cheese
sturgeon caviar
red bean
bean
mushroom
cream
galanga
sheep cheese
buttermilk
salmon
tuna
shiitake
porcini
cucumber
fish
cashew
coconut
herring
enokidake
corn
yogurt
matsutake
vegetable
haddock
cod
vegetable oil
roasted hazelnut
roasted pecan
hazelnut
pecan

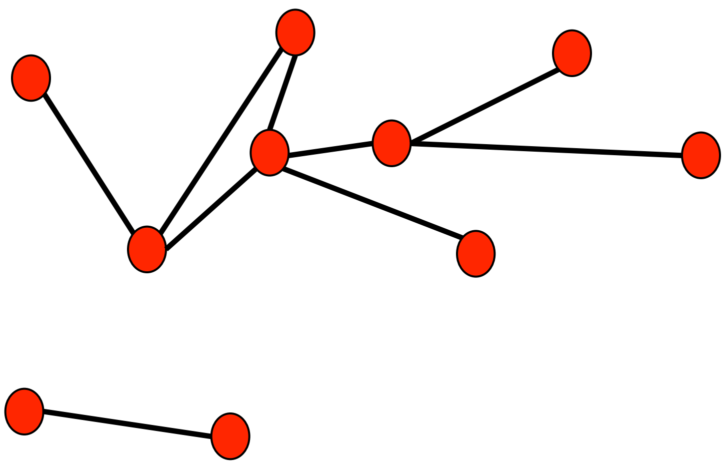
Basic network measures

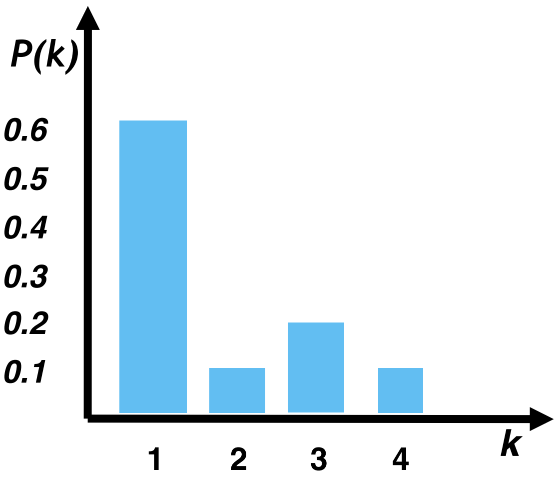
Degree of a node
Distance between two nodes
Clustering among three nodes

Degree distribution P(k): probability that a randomly chosen vertex has degree k

N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot



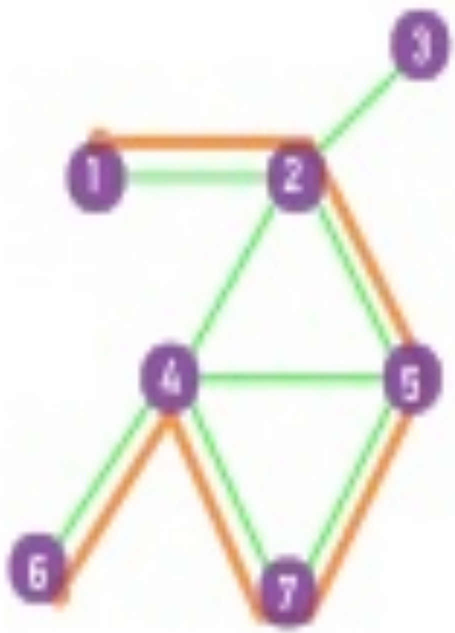


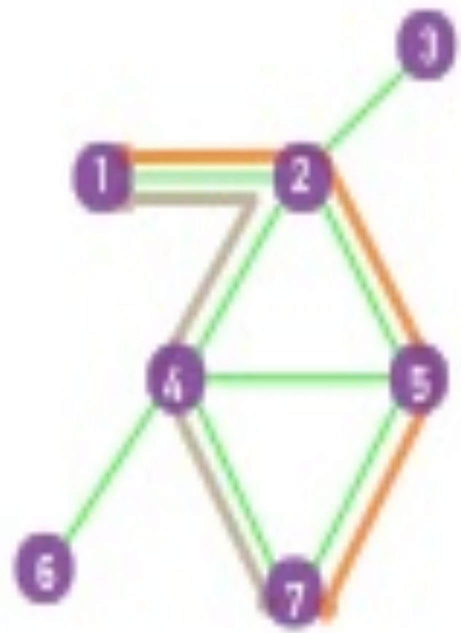
PATHS

A *path is* a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

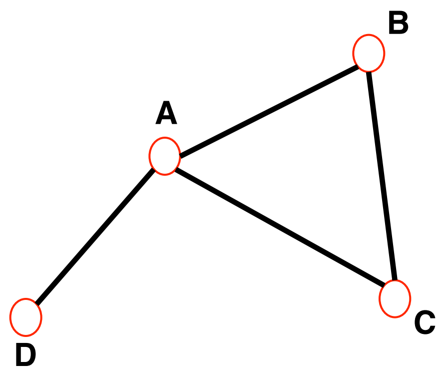
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$





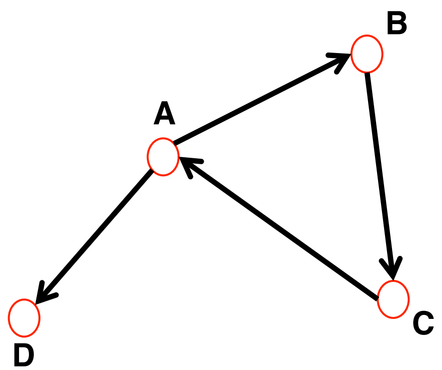
- In a directed network, the path can follow only the direction of an arrow.

Shortest Path, Geodesic Path



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

NETWORK DIAMETER AND AVERAGE DISTANCE

Diameter: $\boldsymbol{d_{max}}$ the maximum distance between any pair of nodes in the graph.

Average path length/distance, <d>, for a connected graph:

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

where d_{ij} is the distance from node i to node j

In an *undirected graph* $d_{ij} = d_{ji}$ *, so* we only need to count them once:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j > i} d_{ij}$$

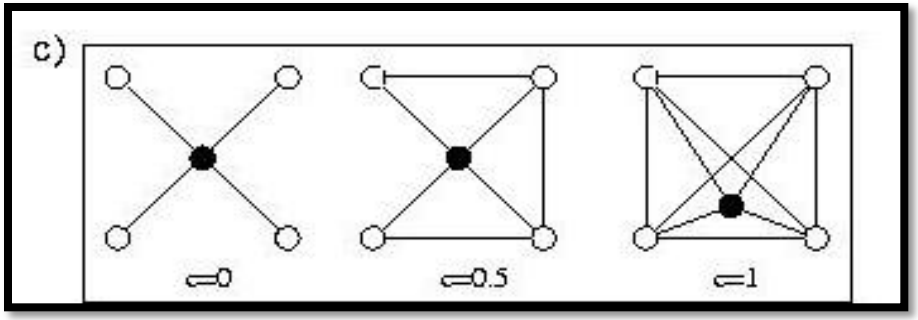
CLUSTERING COEFFICIENT

Clustering coefficient:

what portion of your neighbors are connected?

- Node i with degree k_i
- C_i in [0,1]

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$



KEY MEASURES

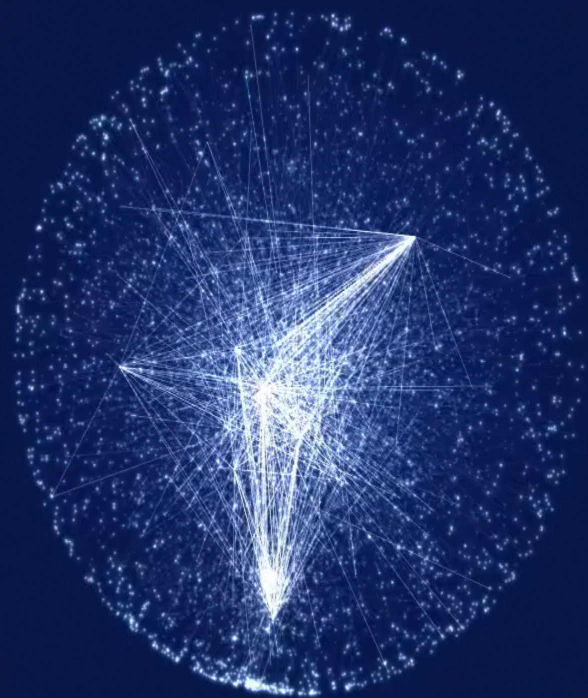
Degree distribution: **P(k)**

Path length: ***l***

Clustering coefficient:

$$\boxed{C_i = \frac{2e_i}{k_i(k_i - 1)}}$$

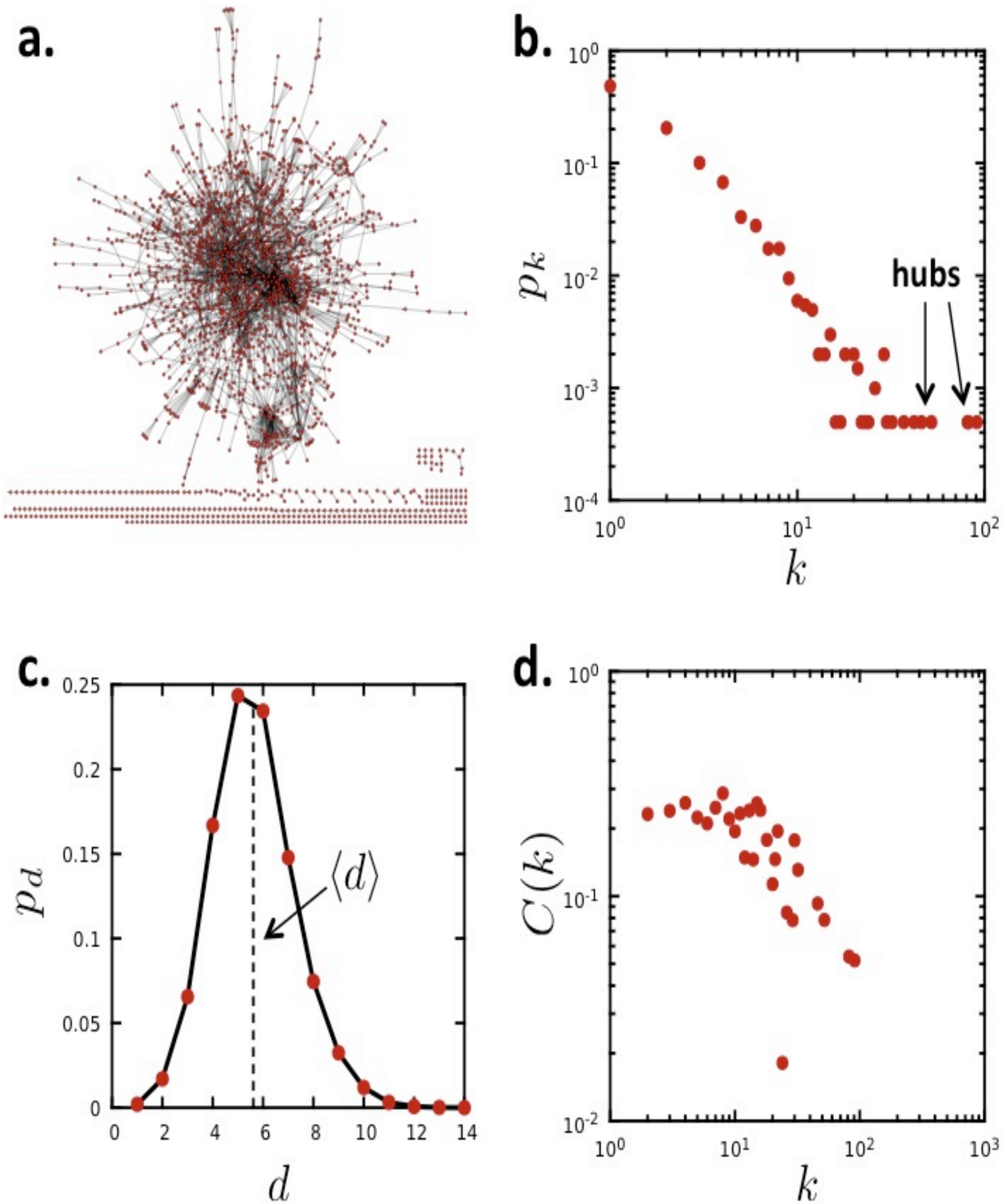
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Undirected network
N=2,018 proteins as nodes
L=2,930 binding interactions as links.
Average degree $\langle k \rangle$=2.90.

Not connected: 185 components
the largest (giant component)
1,647 nodes

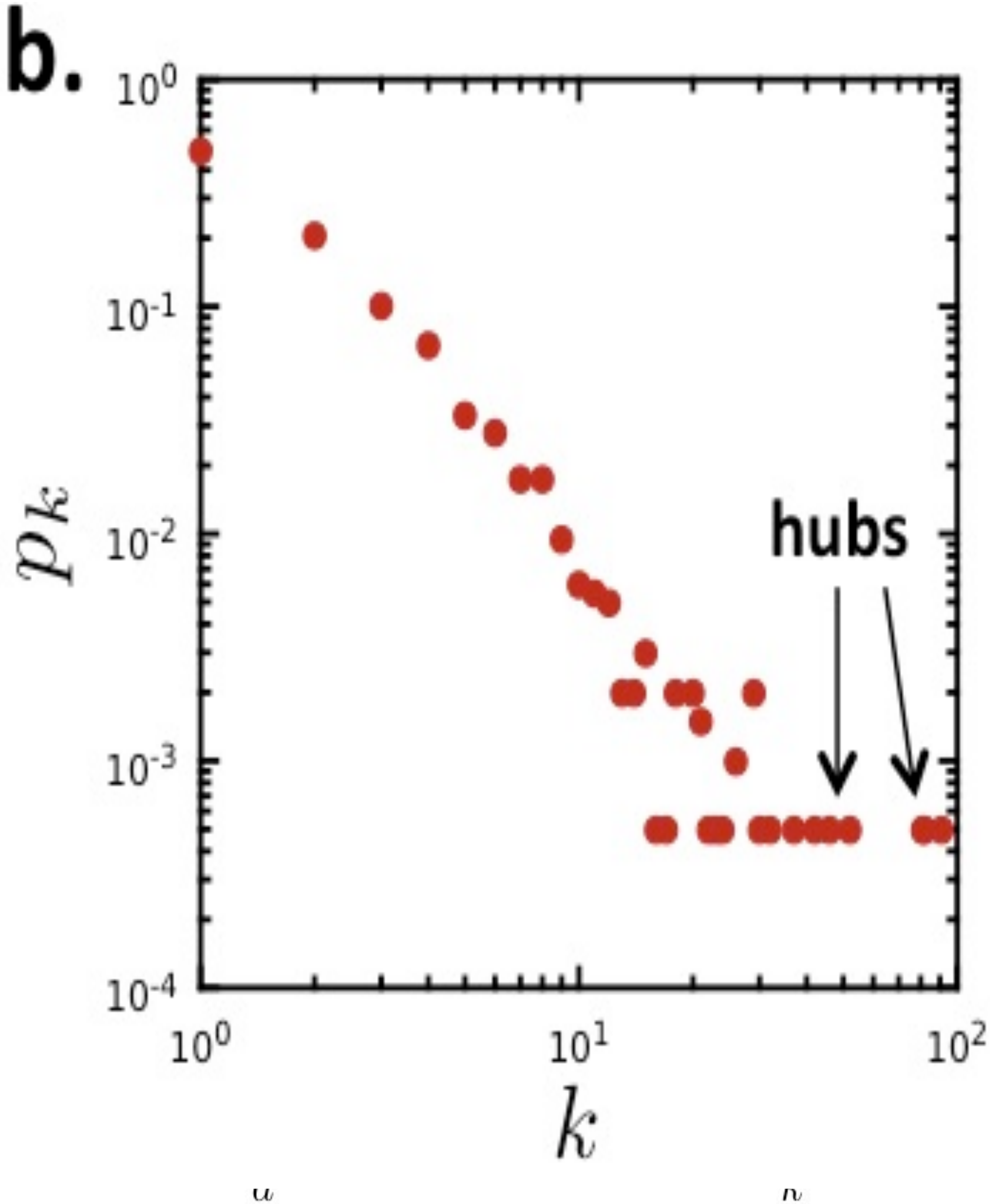
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK



Undirected network
N=2,018 proteins as nodes
L=2,930 binding interactions as links.
Average degree <k>=2.90.

Not connected: 185 components
the largest (giant component)
1,647 nodes

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

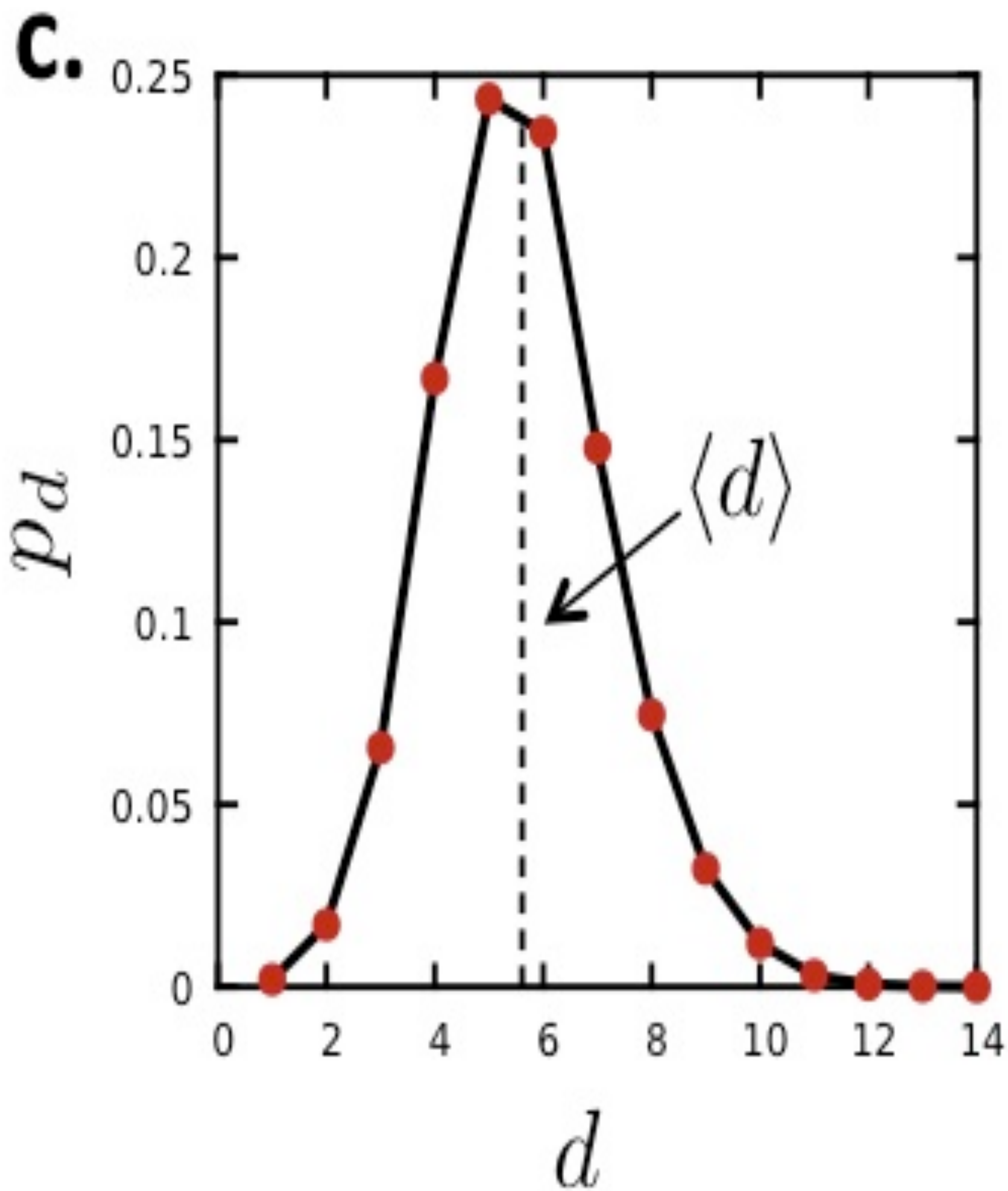


$\mathbf{p_k}$ is the probability that a node has degree **k**.

N_k = # nodes with degree k

$p_k = N_k / N$

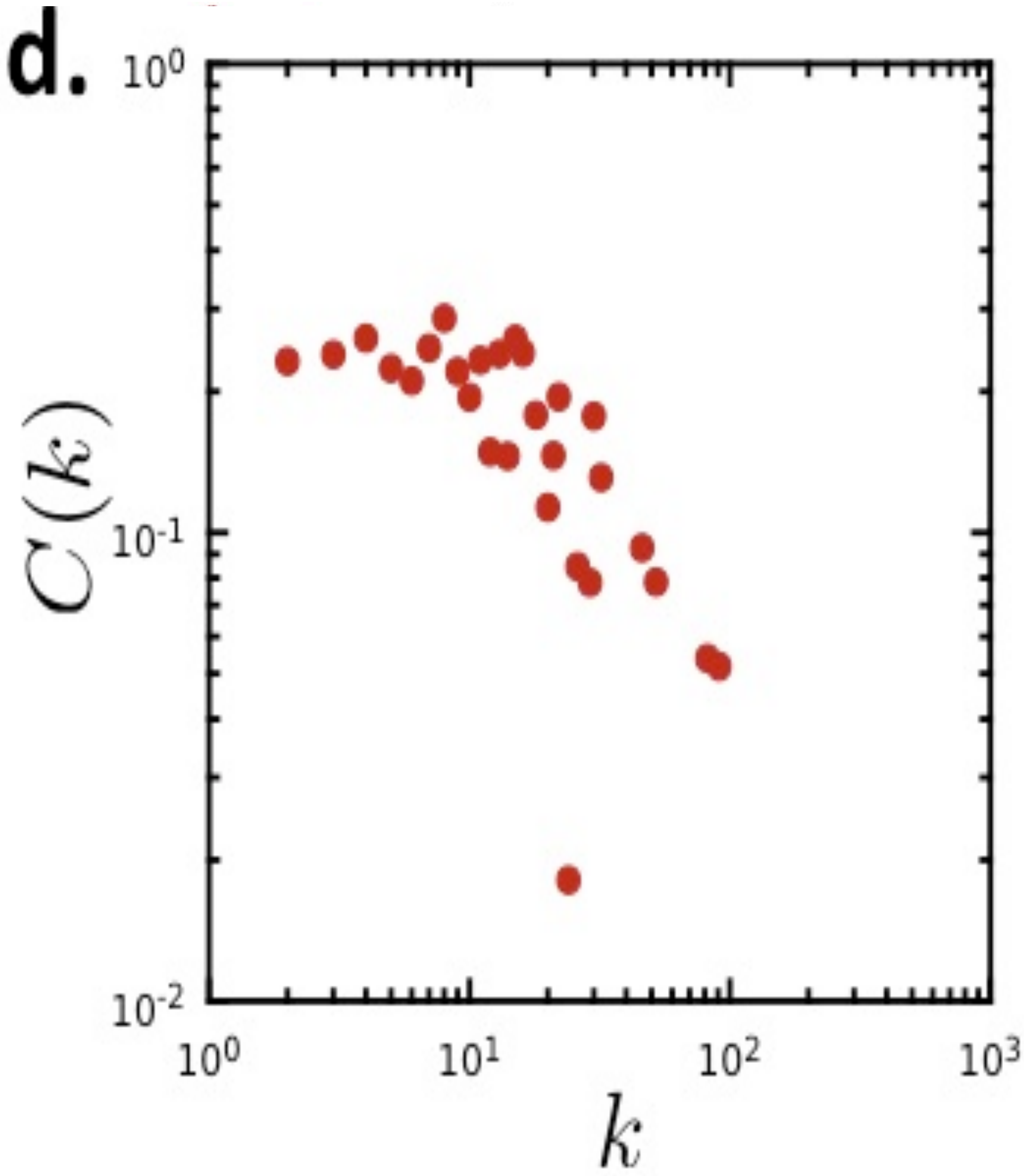
A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK

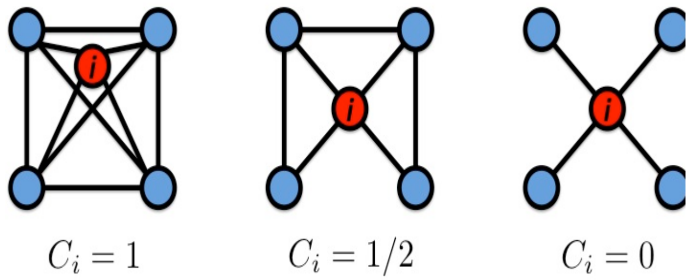


d_{max}=14

<d>=5.61

A CASE STUDY: PROTEIN-PROTEIN INTERACTION NETWORK





$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

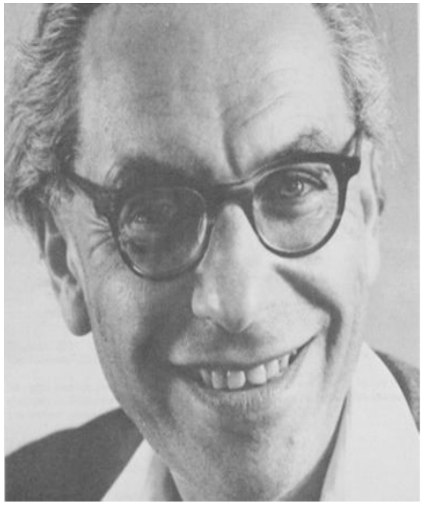
<C>=0.12

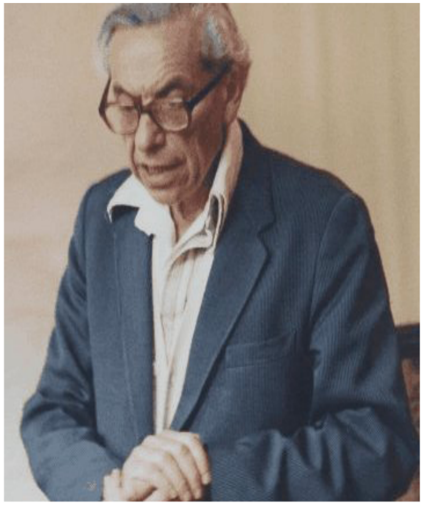
Random graphs

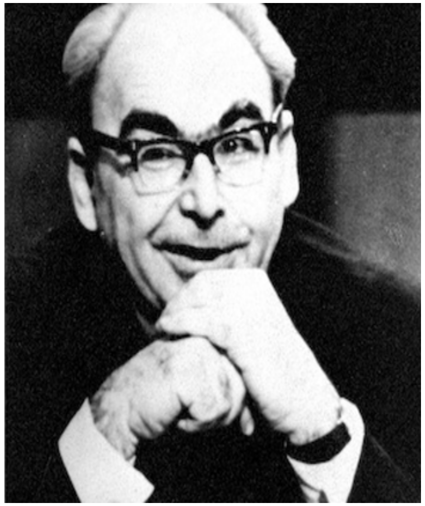
What are the expected basic measures emerging from random?

RANDOM NETWORK MODEL

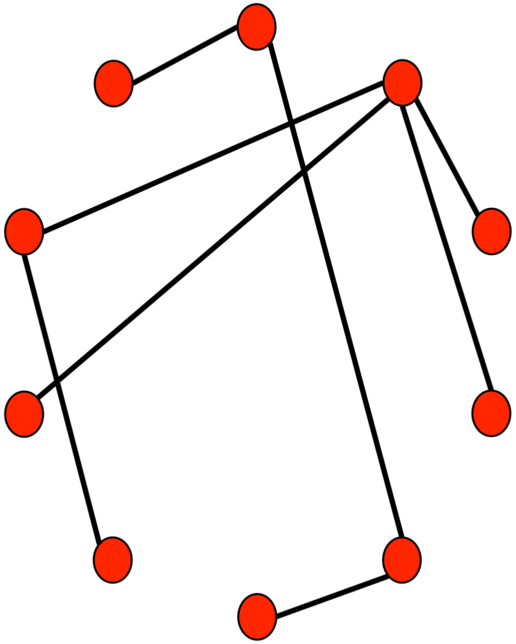
Pául Erdös
(1913-1996)







Erdös-Rényi model (1960)



Connect with probability p

p=1/6 N=10

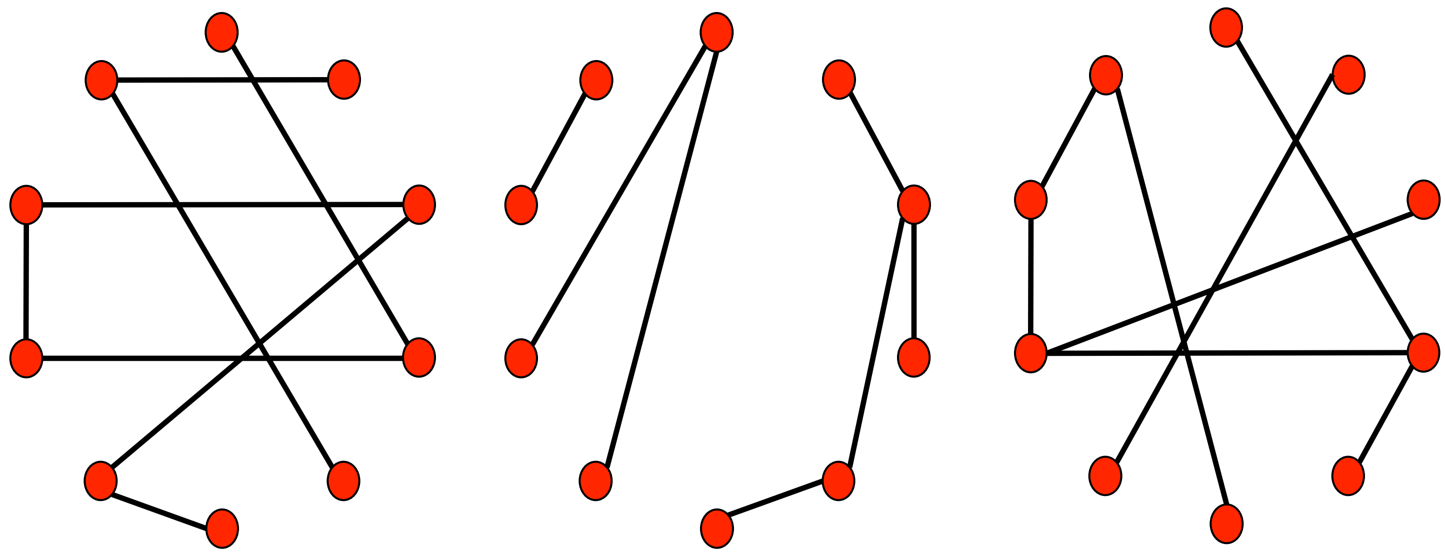
$\langle k \rangle \sim 1.5$

RANDOM NETWORK MODEL

Definition: A **random graph** is a graph of N labeled nodes where each pair of nodes is connected by a preset probability **p**.

RANDOM NETWORK MODEL

N and ***p*** do not uniquely define the network– we can have many different realizations of it. **How many?**



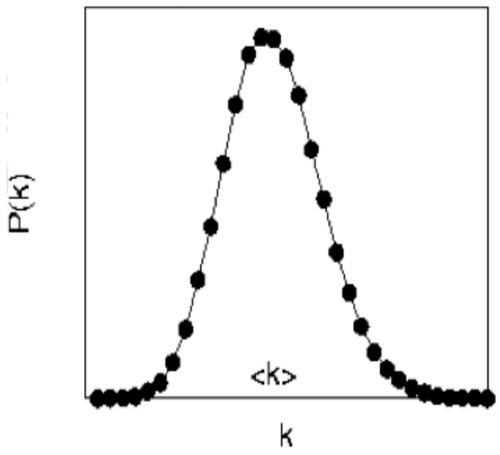
N=10
p=1/6

The probability to form a *particular* graph **G(N,L)** is

$$P(G(N,L)) = p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

That is, each graph **G(N,L)** appears with probability **P(G(N,L))**.

DEGREE DISTRIBUTION OF A RANDOM GRAPH



$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

Select k
nodes from N-1

probability of
having *k* edges

probability of
missing N-1-k
edges

$$<k> = p(N-1) \qquad \sigma_k^2 = p(1-p)(N-1)$$

$$\frac{\sigma_k}{<k>} = \left[\frac{1-p}{p}\frac{1}{(N-1)}\right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

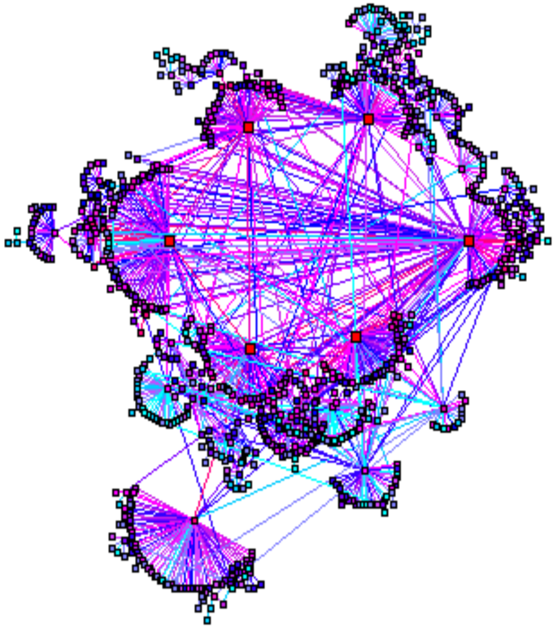
As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of <k>.

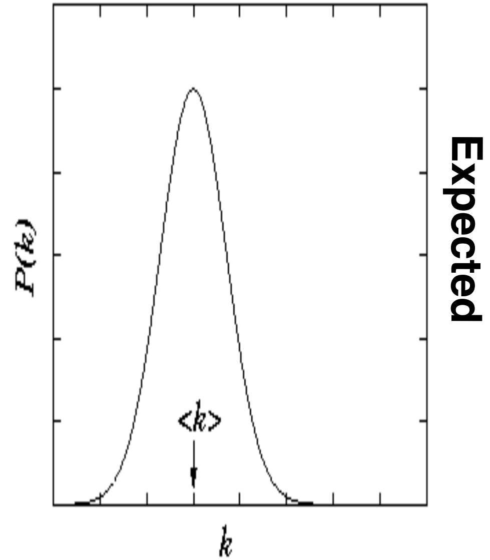
WORLD WIDE WEB

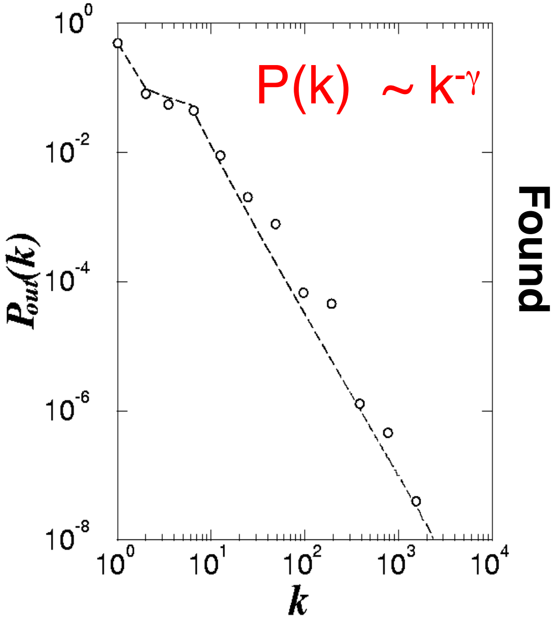
Nodes: **WWW documents**
Links: **URL links**

Over 3 billion documents

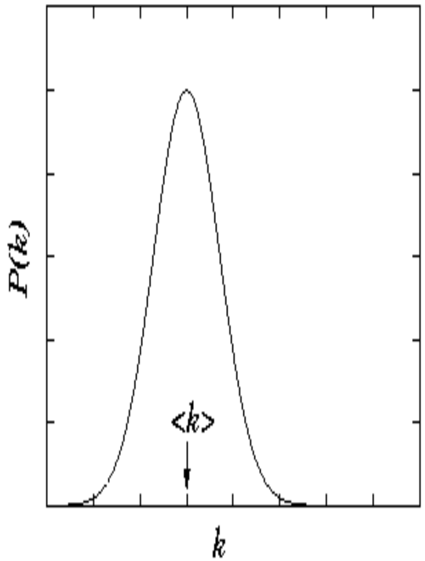
ROBOT: collects all URL's found in a document and follows them recursively



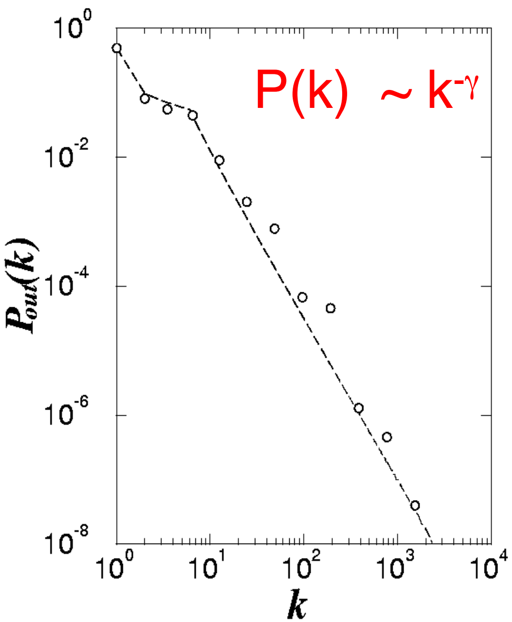




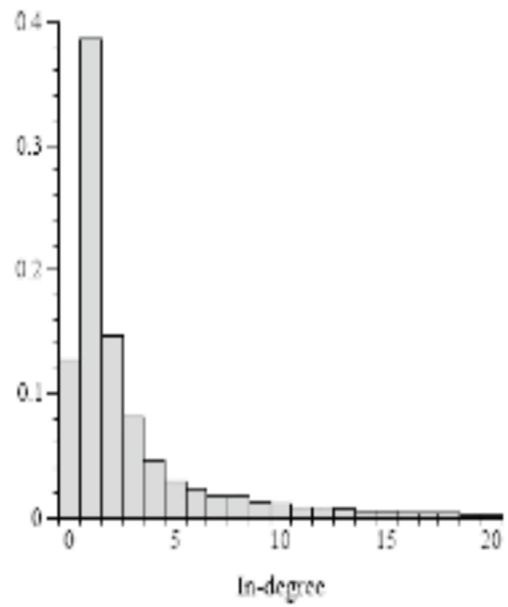
Degree distribution of the WWW

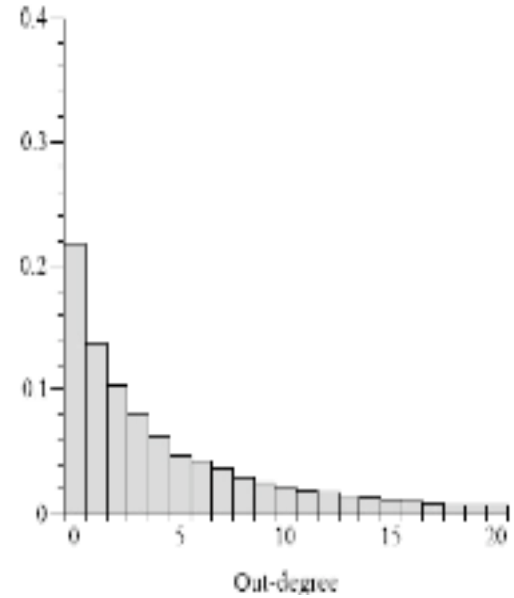


Expected



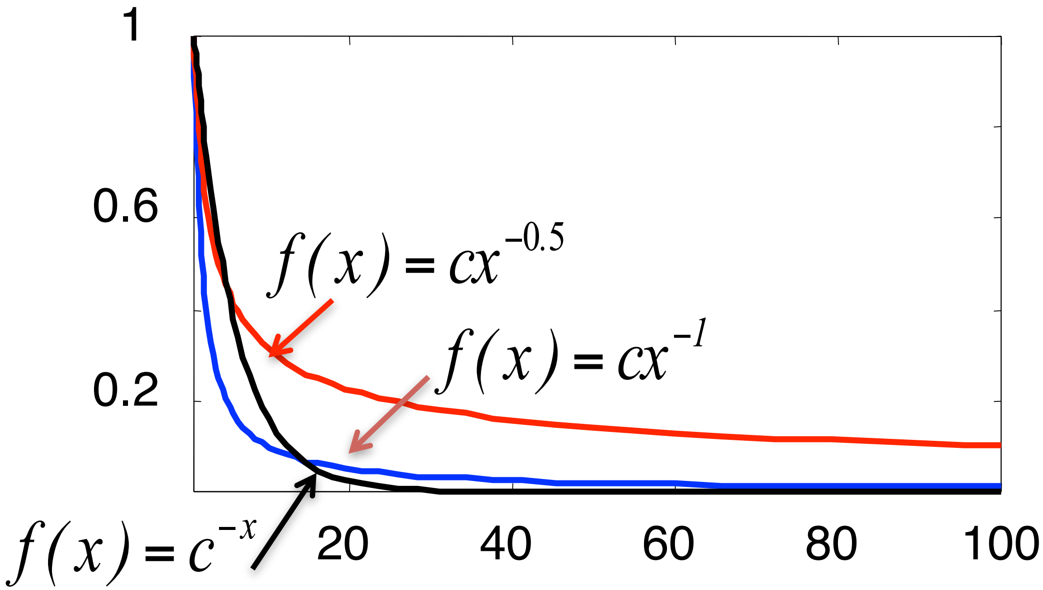
Found





R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

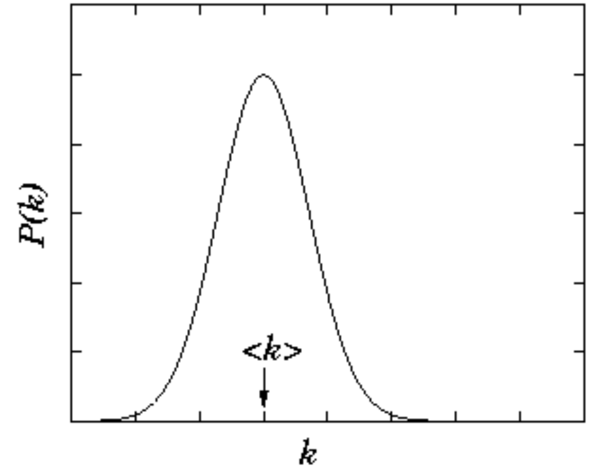
The difference between a power law and an exponential distribution



Above a certain x value, the power law is always higher than the exponential.

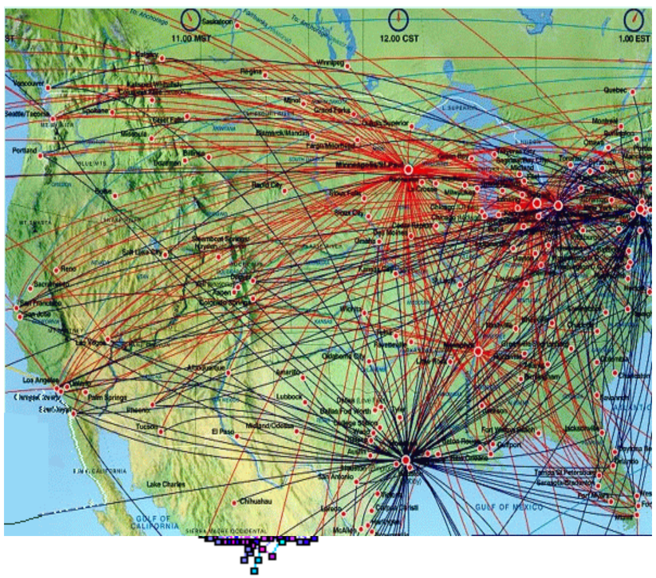
What does the difference mean? Visual representation.

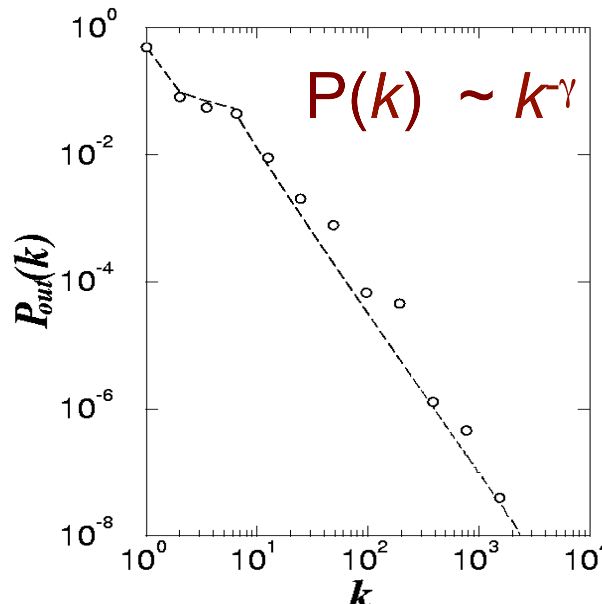
Exponential Network



Expected

Scale-free Network

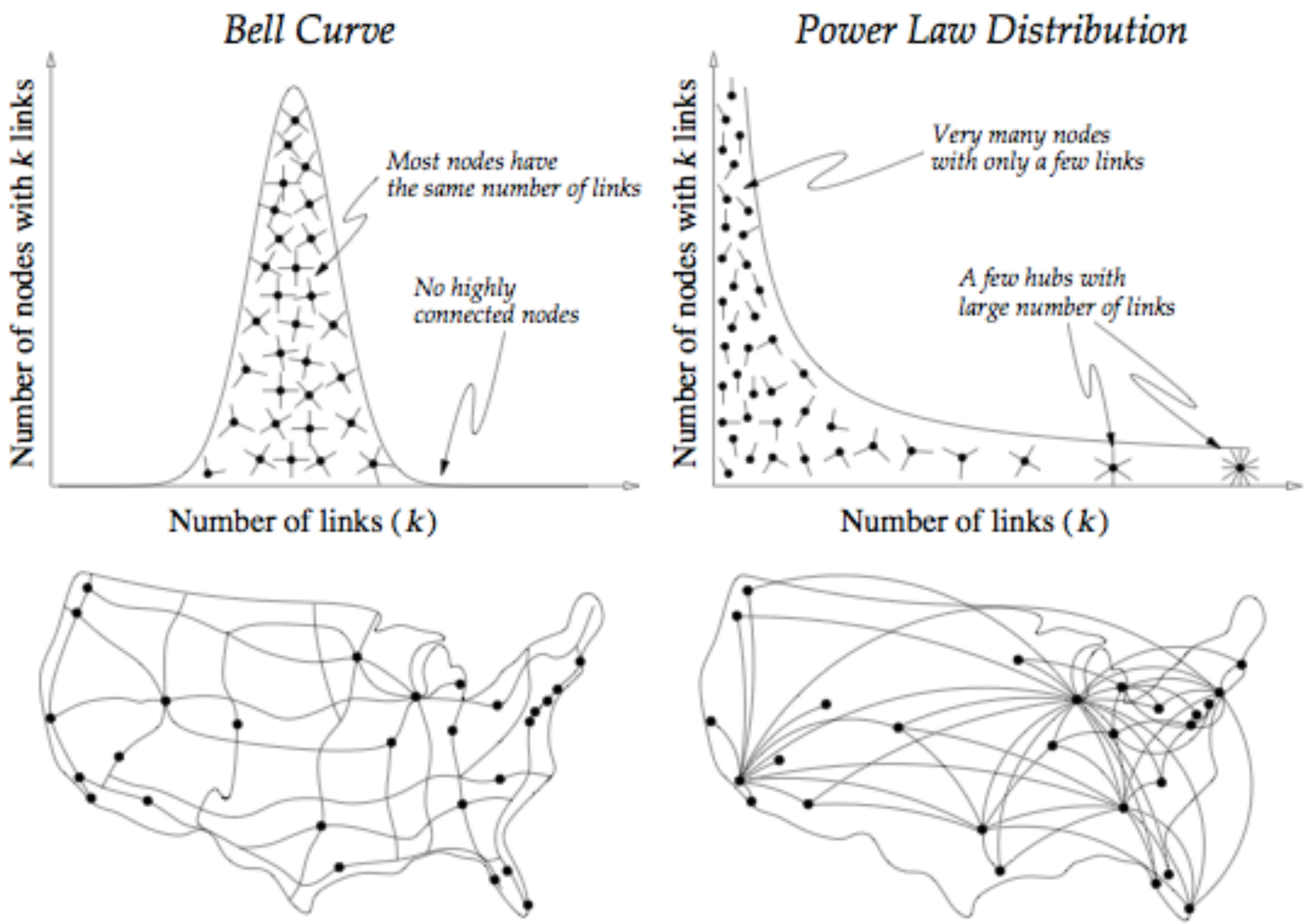




Found

R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

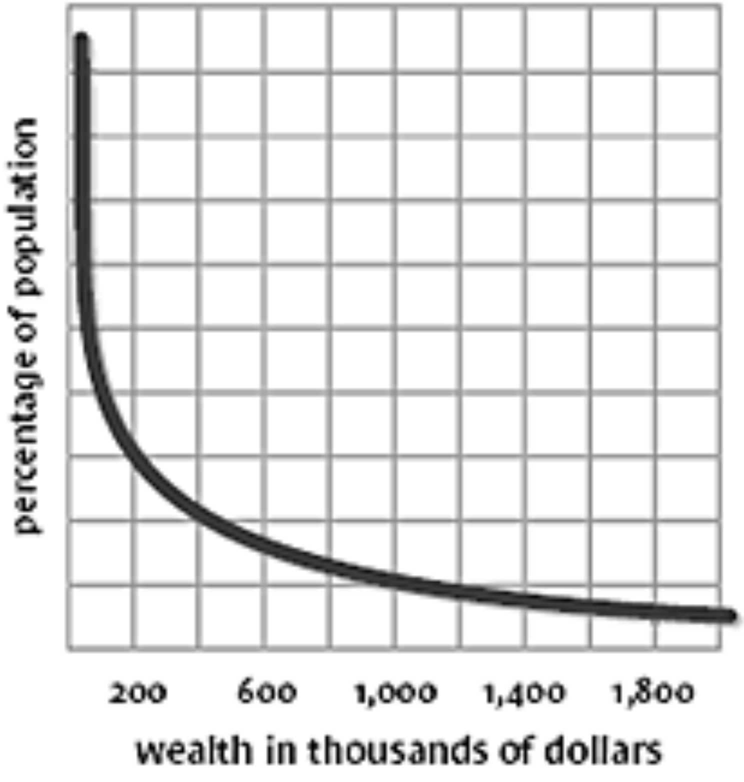
WORLD WIDE WEB



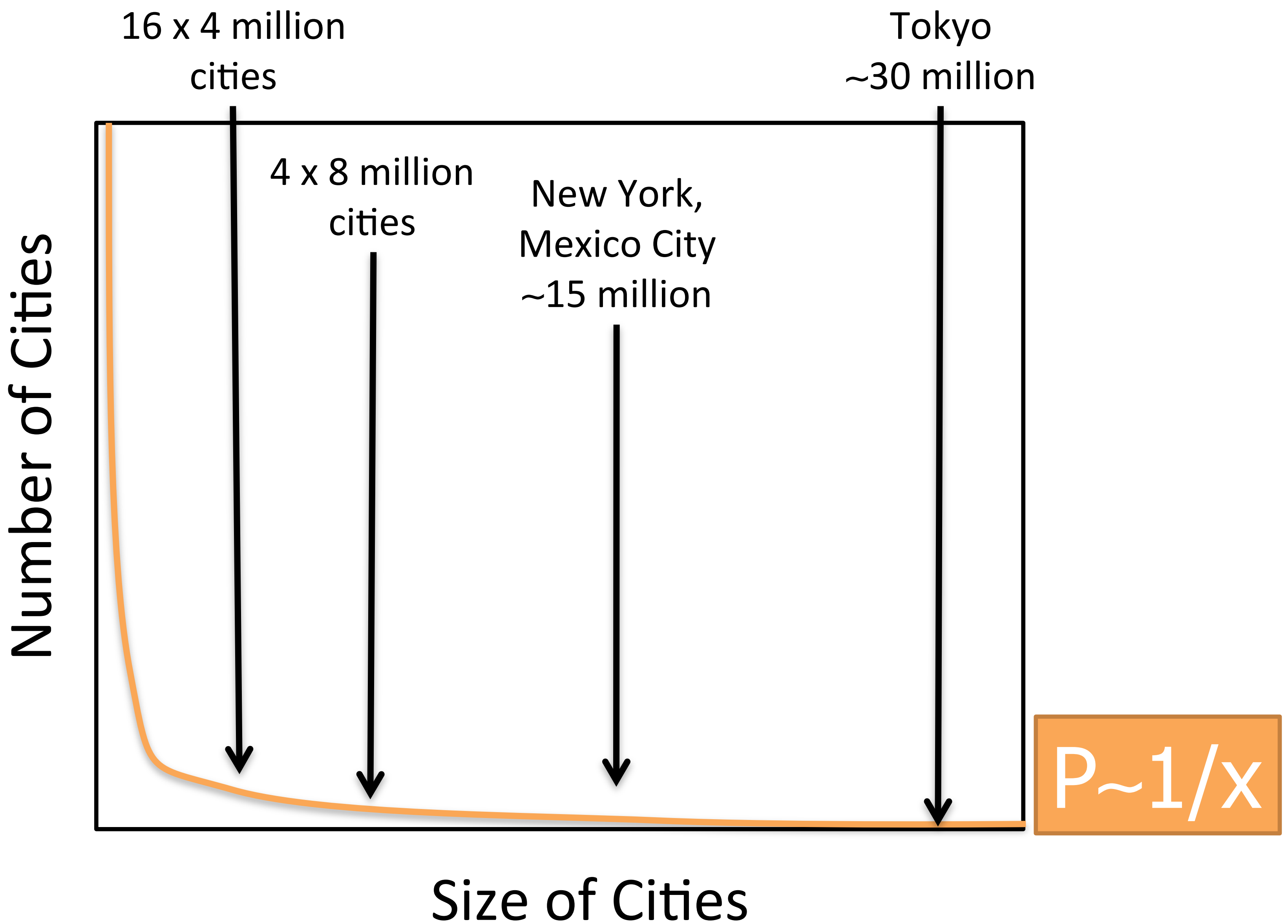
PARETO DISTRIBUTION OF WEALTH

Vilfredo Pareto (1848-1923)

Rich and Poor in America



This plot of household wealth in the United States, taken from 1998 census figures, clearly shows a distribution of rich and poor forming a Pareto curve. The highest percentage of households fall at the lower levels of wealth, but at the higher end, the curve drops off relatively slowly, displaying Pareto's "fat-tailed" pattern.


16 x 4 million cities
Tokyo ~30 million
4 x 8 million cities
New York, Mexico City ~15 million
Number of Cities
Size of Cities
P~1/x

NO OUTLIERS IN A RANDOM SOCIETY

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

→The most connected individual has degree k_{max}~1,185
→The least connected individual has degree k_{min} ~ 816

The probability to find an individual with degree k>2,000 is 10^{-27}. Hence the chance of finding an individual with 2,000 acquaintances is so tiny that such nodes are virtually inexistent in a random society.

→a random society would consist of mainly average individuals, with everyone with roughly the same number of friends.
→It would lack outliers, individuals that are either highly popular or recluse.

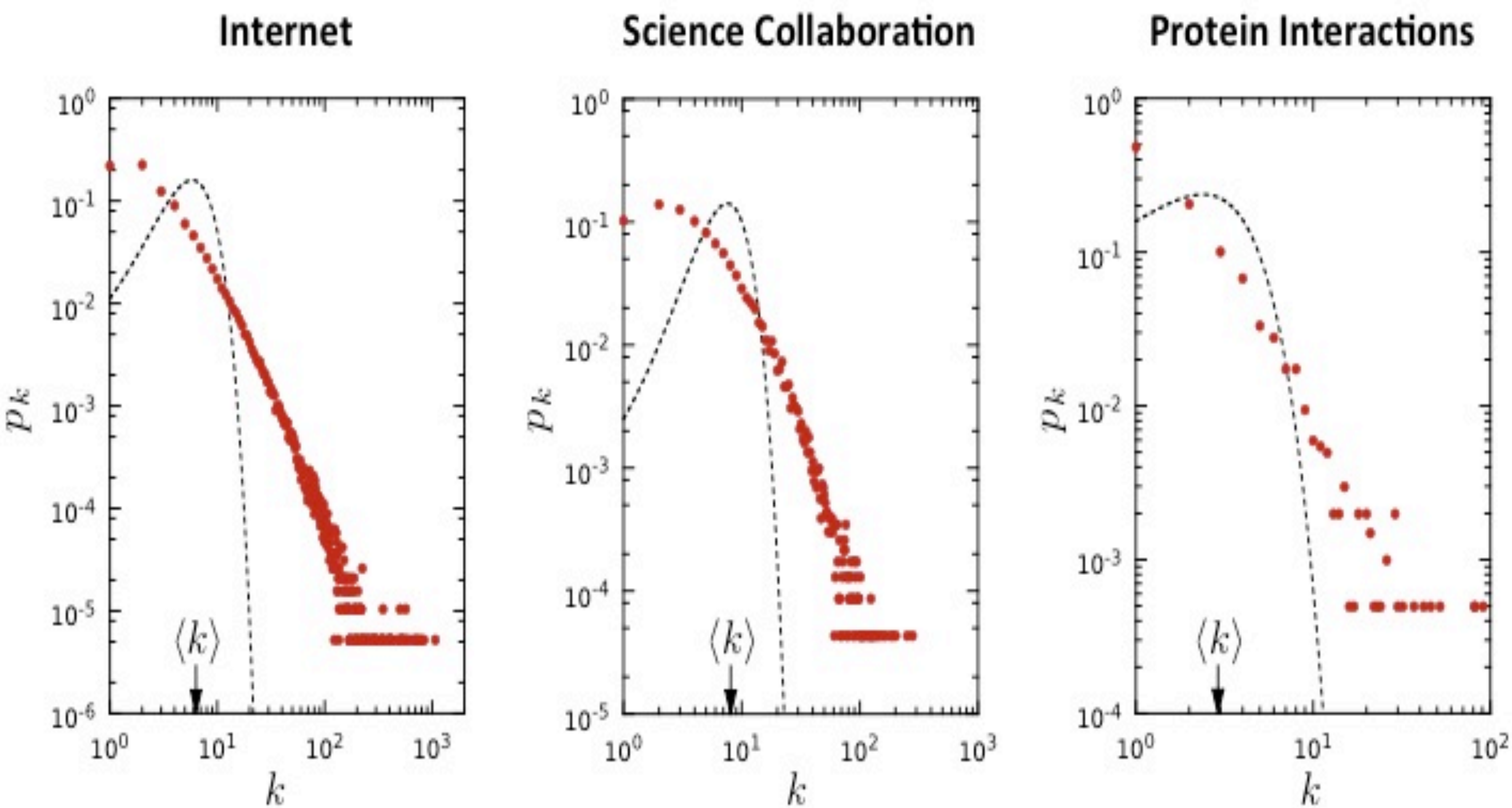
After Bill enters the arena the average wealth of the public ~ $1,000,000

~ $100 billion

10^5 people, 10^5 $ average wealth per capita

FACING REALITY: Degree distribution of real networks

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$



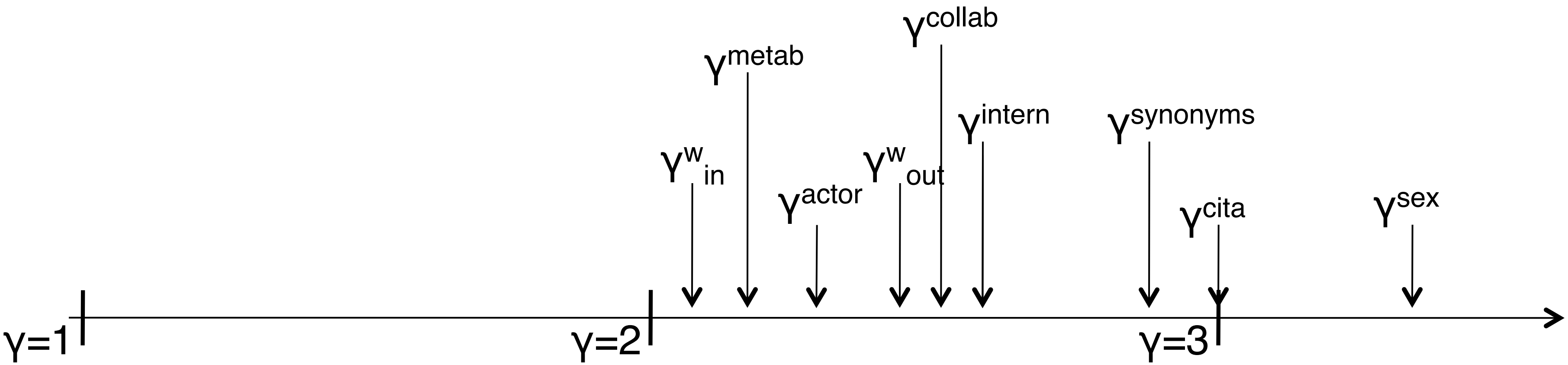
UNIVERSALITY

Network	Size	$\langle k \rangle$	κ	γ_{out}	γ_{in}
WWW	325 729	4.51	900	2.45	2.1
WWW	4×10^7	7		2.38	2.1
WWW	2×10^8	7.5	4000	2.72	2.1
WWW, site	260 000				1.94
Internet, domain*	3015–4389	3.42–3.76	30–40	2.1–2.2	2.1–2.2
Internet, router*	3888	2.57	30	2.48	2.48
Internet, router*	150 000	2.66	60	2.4	2.4
Movie actors*	212 250	28.78	900	2.3	2.3
Co-authors, SPIRES*	56 627	173	1100	1.2	1.2
Co-authors, neuro.*	209 293	11.54	400	2.1	2.1
Co-authors, math.*	70 975	3.9	120	2.5	2.5
Sexual contacts*	2810			3.4	3.4
Metabolic, *E. coli*	778	7.4	110	2.2	2.2
Protein, *S. cerev.**	1870	2.39		2.4	2.4
Ythan estuary*	134	8.7	35	1.05	1.05
Silwood Park*	154	4.75	27	1.13	1.13
Citation	783 339	8.57			3
Phone call	53×10^6	3.16		2.1	2.1
Words, co-occurrence*	460 902	70.13		2.7	2.7
Words, synonyms*	22 311	13.48		2.8	2.8

Networks:
The exponents vary from system to system.
Most are between 2 and 3

Universality:
the emergence of common features across different networks. Like the scale-free property.

VARIANCE DIVERGES!



$\langle k^2 \rangle$ diverges	$\langle k^2 \rangle$ finite

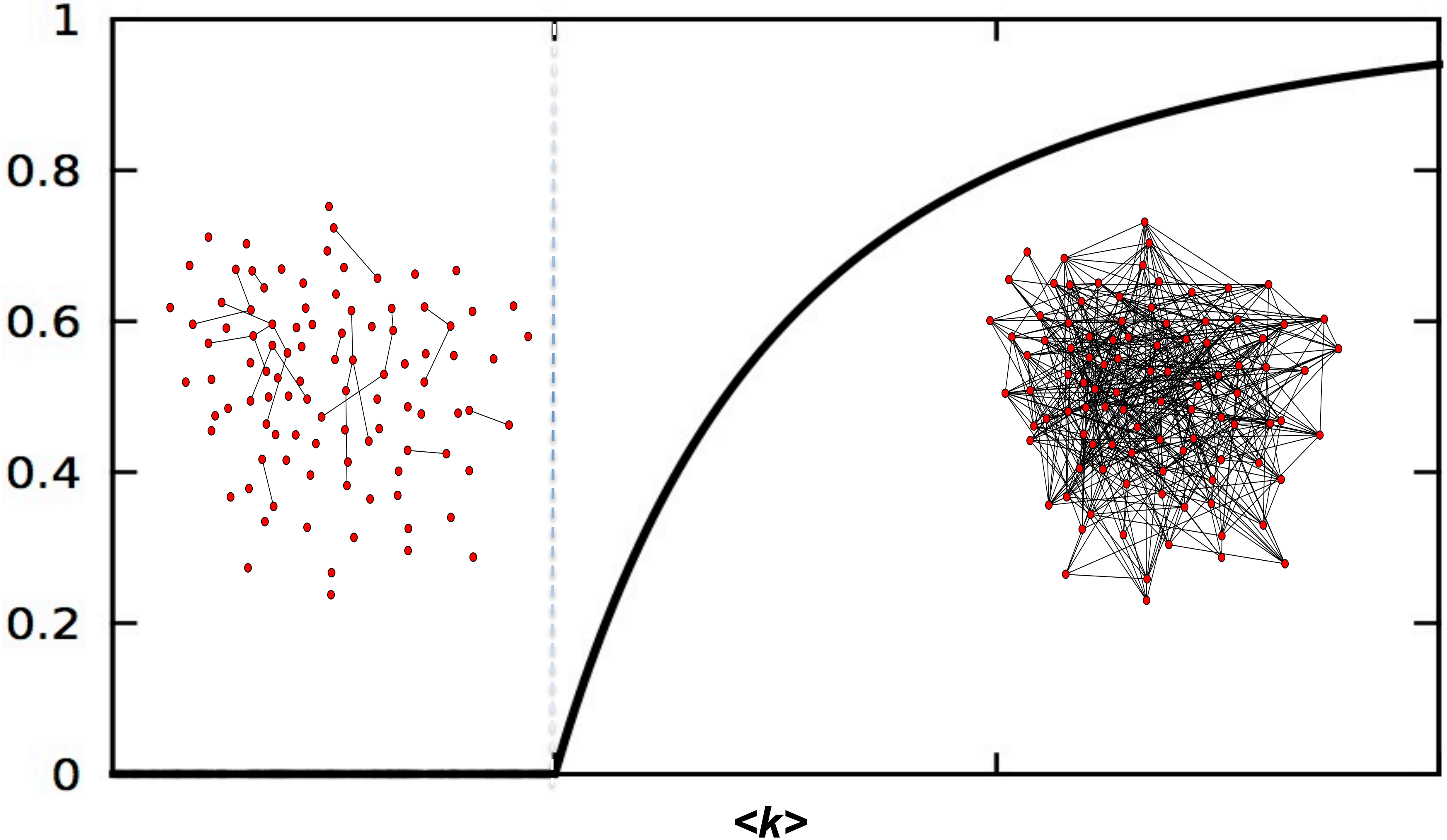
Regime full of anomalies…	The scale-free behavior is relevant	Behaves like a random network

Why are most exponents in this regime?

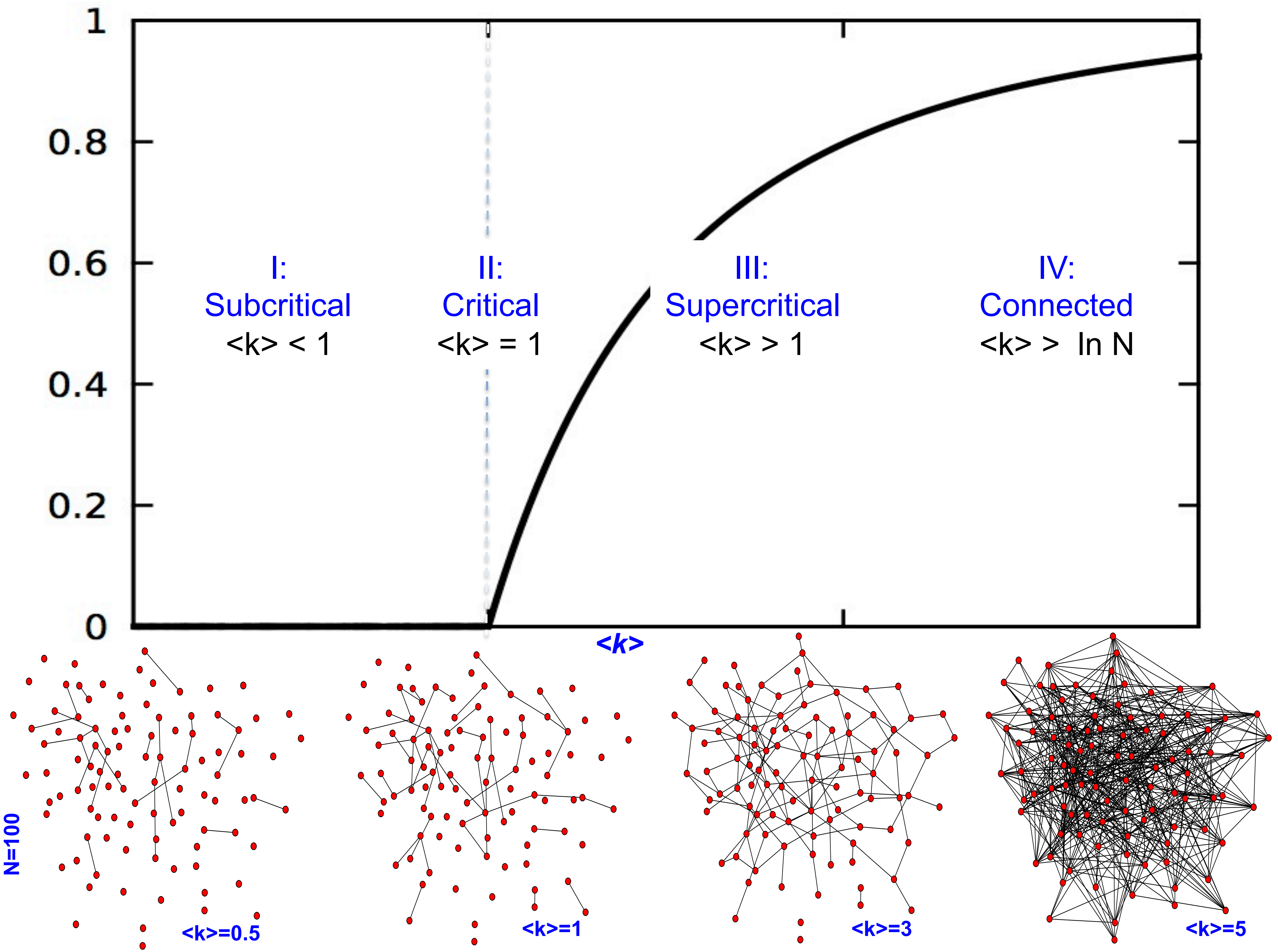
The evolution of a random network

EVOLUTION OF A RANDOM NETWORK

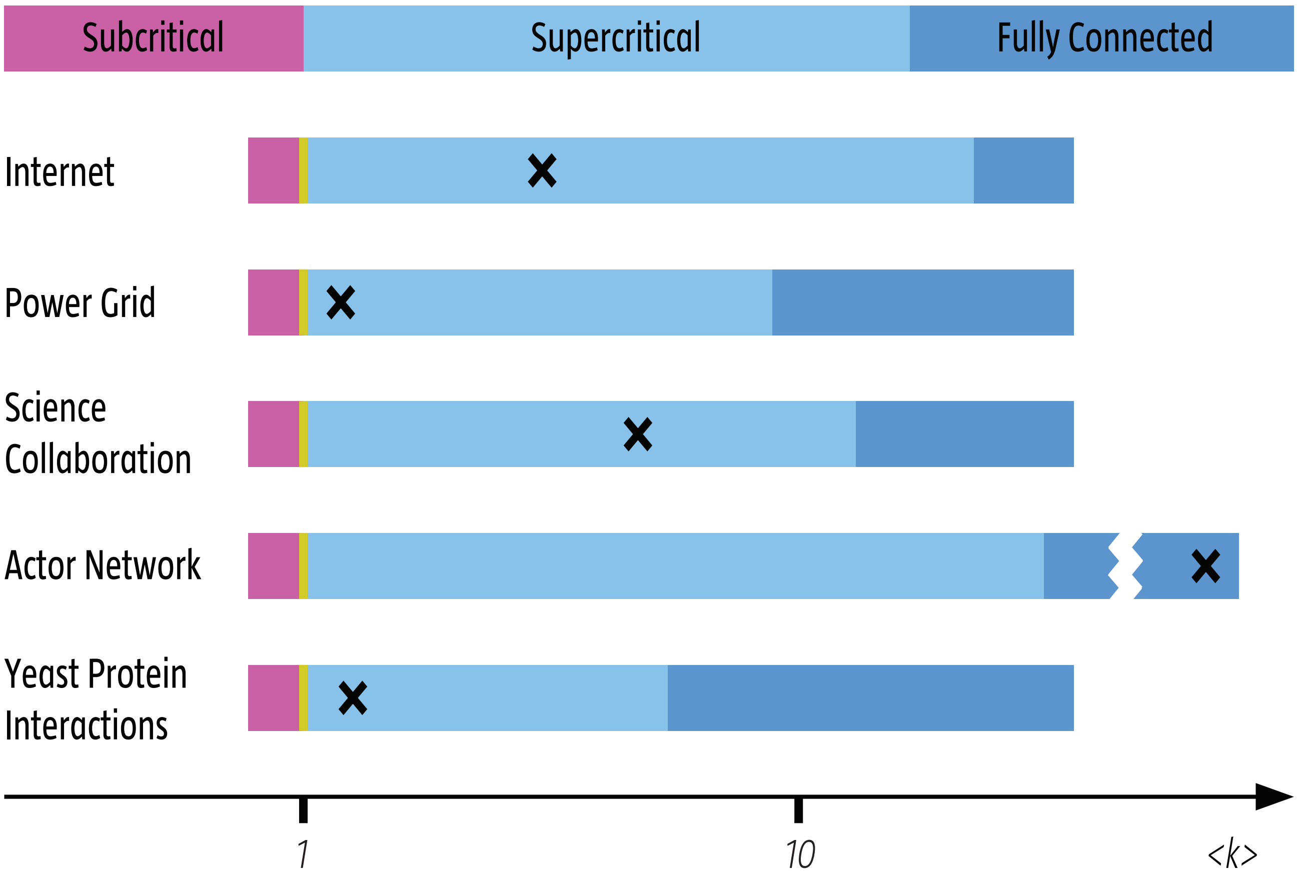
disconnected nodes ➔ **NETWORK.**



How does this transition happen?


1
0.8
0.6
0.4
0.2
0
I:
Subcritical
<k> < 1
II:
Critical
<k> = 1
III:
Supercritical
<k> > 1
IV:
Connected
<k> > ln N
<k>
N=100
<k>=0.5
<k>=1
<k>=3
<k>=5

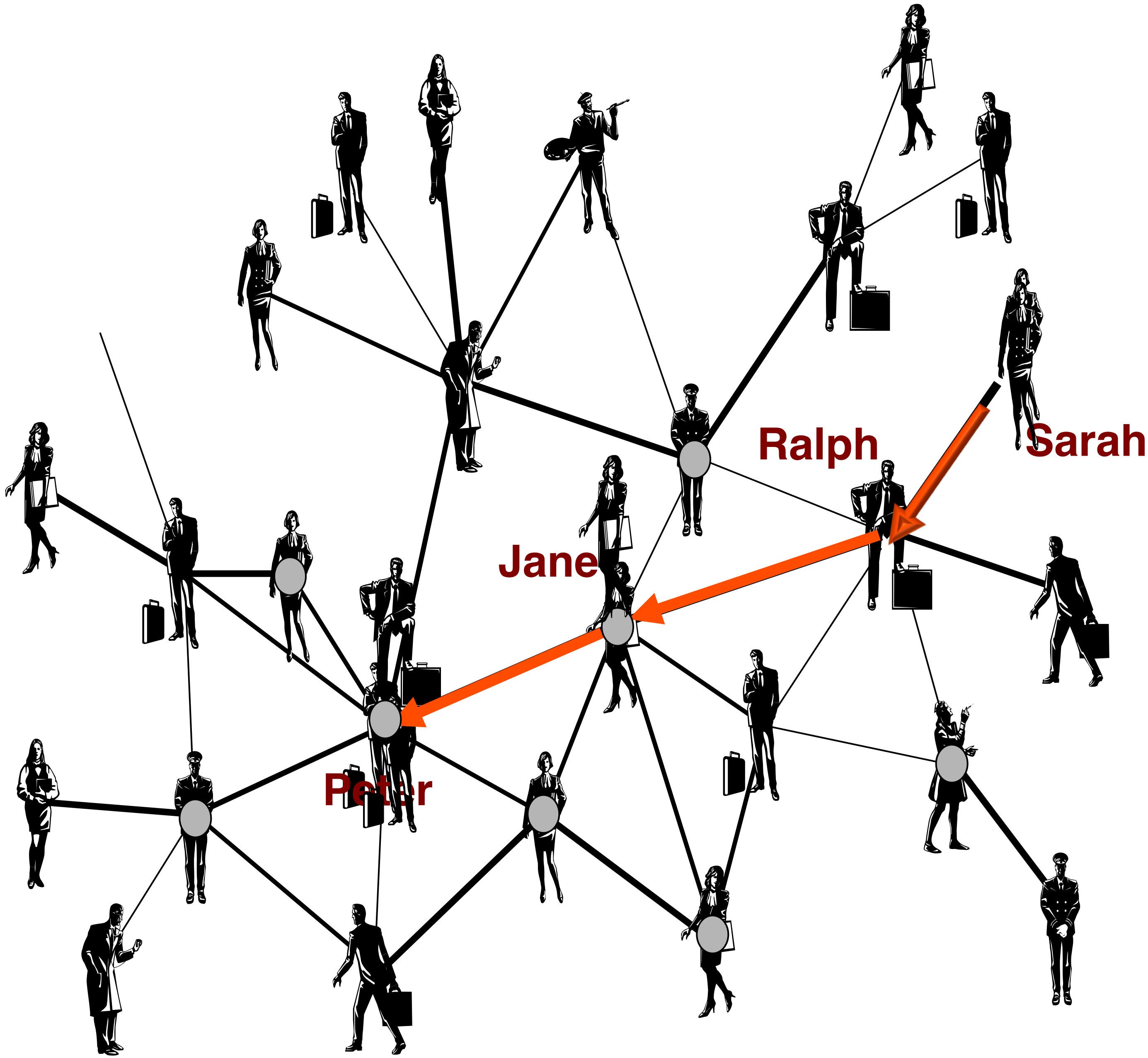
Real networks are supercritical



Network	N	L	$\langle k \rangle$	$\ln N$
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	186,936	8.08	10.04
Actor Network	212,250	3,054,278	28.78	12.27
Yeast Protein Interactions	2,018	2,930	2.90	7.61

Small world property

small worlds



Frigyes Karinthy, 1929
Stanley Milgram, 1967

1929: Frigyes Kartinthy

Frigyes Karinthy (1887-1938)
Hungarian Writer

1929: *Minden másképpen van* (Everything is Different)
Láncszemek (Chains)

“Look, Selma Lagerlöf just won the Nobel Prize for Literature, thus she is bound to know King Gustav of Sweden, after all he is the one who handed her the Prize, as required by tradition. King Gustav, to be sure, is a passionate tennis player, who always participates in international tournaments. He is known to have played Mr. Kehrling, whom he must therefore know for sure, and as it happens I myself know Mr. Kehrling quite well.”

"The worker knows the manager in the shop, who knows Ford; Ford is on friendly terms with the general director of Hearst Publications, who last year became good friends with Arpad Pasztor, someone I not only know, but to the best of my knowledge a good friend of mine. So I could easily ask him to send a telegram via the general director telling Ford that he should talk to the manager and have the worker in the shop quickly hammer together a car for me, as I happen to need one."

1967: Stanley Milgram

HOW TO TAKE PART IN THIS STUDY

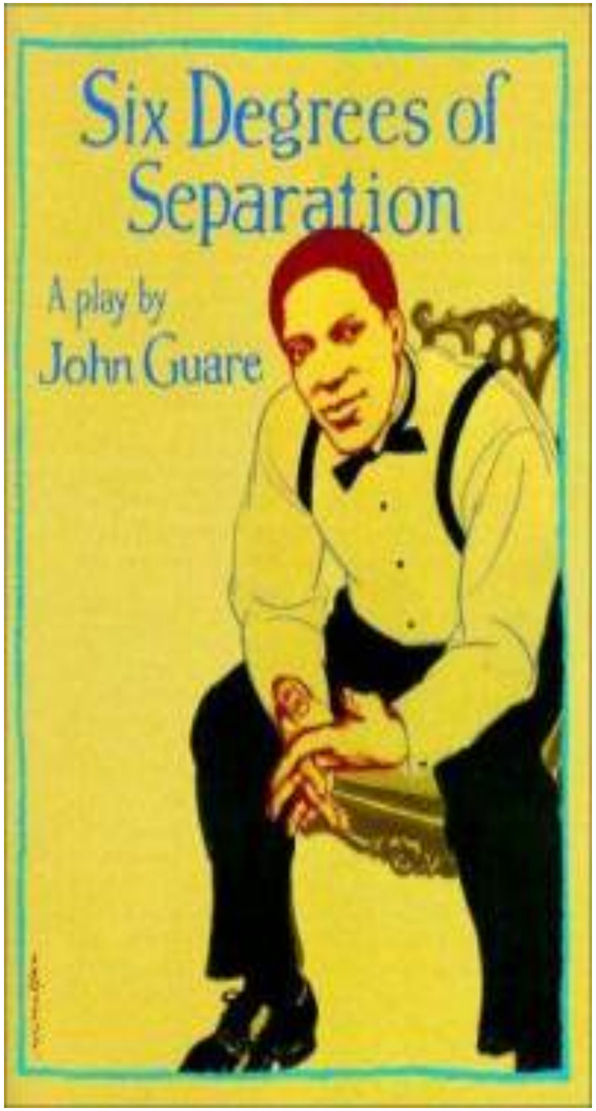
1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

2. DETACH ONE POSTCARD. FILL IT AND RETURN IT TO HARVARD UNIVERSITY. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.

3. IF YOU KNOW THE TARGET PERSON ON A PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIM (HER). Do this only if you have previously met the target person and know each other on a first name basis.

4. IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder to a friend, relative or acquaintance, but it must be someone you know on a first name basis.

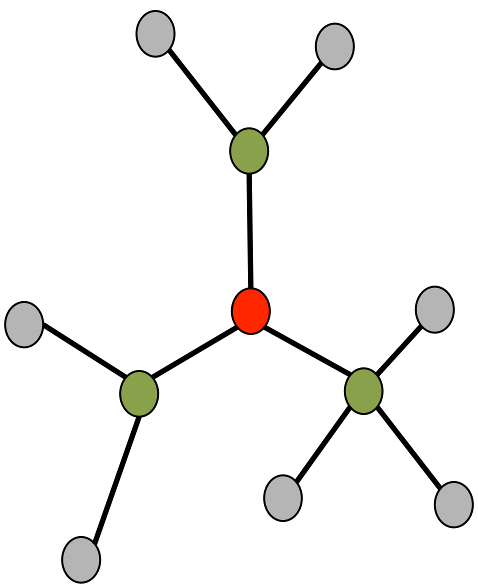
1991: John Guare



"Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice…. It's not just the big names. It's anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am bound to everyone on this planet by a trail of six people. It's a profound thought. How every person is a new door, opening up into other worlds."

DISTANCES IN RANDOM GRAPHS

Random graphs tend to have a tree-like topology with almost constant node degrees.



- nr. of first neighbors: $N_1 \cong \langle k \rangle$
- nr. of second neighbors: $N_2 \cong \langle k \rangle^2$
- nr. of neighbours at distance d: $N_d \cong \langle k \rangle^d$
- estimate maximum distance:

$$1 + \sum_{l=1}^{l_{\max}} \langle k \rangle^i = N \quad \Rightarrow \quad l_{\max} = \frac{\log N}{\log \langle k \rangle}$$

DISTANCES IN RANDOM GRAPHS — compare with real data

$$l_{\max} = \frac{\log N}{\log \langle k \rangle}$$

Network	Size	⟨k⟩	l	l_{rand}	C	C_{rand}	Reference	Nr
www, site level, undir	153127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015-6209	3.52-4.11	3.7-3.76	6.36-6.18	0.18-0.3	0.001	Yook e al., 2001a, Pastor-Satorras et al., 2001	2
Movie actors	225226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz,1998	3
LANL co-authorship	52909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE eo-authorship	1520251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabasi et al, 2001	8
Neurosci. co-authorship	209293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabasi et al, 2001	9
E. coli, sustrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Sole, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Sole, 2000	13
Words, co-occurrence	460902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Sole, 2001	14
Words, synonyms	22311	13.48	4.5	3.84	0.7	0.0006	Yook et al. 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C.Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Given the huge differences in scope, size, and average degree, the agreement is excellent.

CLUSTERING COEFFICIENT

$$C_i \equiv \frac{2n_i}{k_i(k_i - 1)}$$

Since edges are independent and have the same probability *p*,

$$n_i \cong p\frac{k_i(k_i - 1)}{2} \quad \Rightarrow \quad C \cong p = \frac{<k>}{N}$$

The clustering coefficient of random graphs is small.

For fixed degree C decreases with the system size N.

13.47 from Newman 2010

CLUSTERING IN RANDOM GRAPHS

compare with real data

Network	Size	⟨k⟩	l	l_{rand}	C	C_{rand}	Reference	Nr
www, site level, undir	153127	35.21	3.1	3.35	0.1078	0.00023	Adamic, 1999	1
Internet, domain level	3015-6209	3.52-4.11	3.7-3.76	6.36-6.18	0.18-0.3	0.001	Yook e al., 2001a, Pastor-Satorras et al., 2001	2
Movie actors	225226	61	3.65	2.99	0.79	0.00027	Watts and Strogatz,1998	3
LANL co-authorship	52909	9.7	5.9	4.79	0.43	1.8×10^{-4}	Newman, 2001a, 2001b, 2001c	4
MEDLINE eo-authorship	1520251	18.1	4.6	4.91	0.066	1.1×10^{-5}	Newman, 2001a, 2001b, 2001c	5
SPIRES co-authorship	56627	173	4.0	2.12	0.726	0.003	Newman, 2001a, 2001b, 2001c	6
NCSTRL co-authorship	11994	3.59	9.7	7.34	0.496	3×10^{-4}	Newman, 2001a, 2001b, 2001c	7
Math. co-authorship	70975	3.9	9.5	8.2	0.59	5.4×10^{-5}	Barabasi et al, 2001	8
Neurosci. co-authorship	209293	11.5	6	5.01	0.76	5.5×10^{-5}	Barabasi et al, 2001	9
E. coli, sustrate graph	282	7.35	2.9	3.04	0.32	0.026	Wagner and Fell, 2000	10
E. coli, reaction graph	315	28.3	2.62	1.98	0.59	0.09	Wagner and Fell, 2000	11
Ythan estuary food web	134	8.7	2.43	2.26	0.22	0.06	Montoya and Sole, 2000	12
Silwood Park food web	154	4.75	3.40	3.23	0.15	0.03	Montoya and Sole, 2000	13
Words, co-occurrence	460902	70.13	2.67	3.03	0.437	0.0001	Ferrer i Cancho and Sole, 2001	14
Words, synonyms	22311	13.48	4.5	3.84	0.7	0.0006	Yook et al. 2001b	15
Power grid	4941	2.67	18.7	12.4	0.08	0.005	Watts and Strogatz, 1998	16
C.Elegans	282	14	2.65	2.25	0.28	0.05	Watts and Strogatz, 1998	17

Erdös-Rényi MODEL (1960)

- **Degree distribution**
  Binomial, Poisson (exponential tails)
- **Clustering coefficient**
  Vanishing for large network sizes
- **Average distance among nodes**
  Logarithmically small

Are real networks like random graphs?
NO!

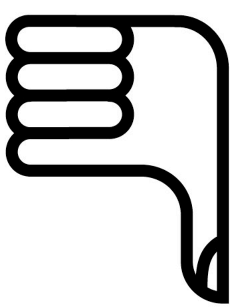
THE DEGREE DISTRIBUTION

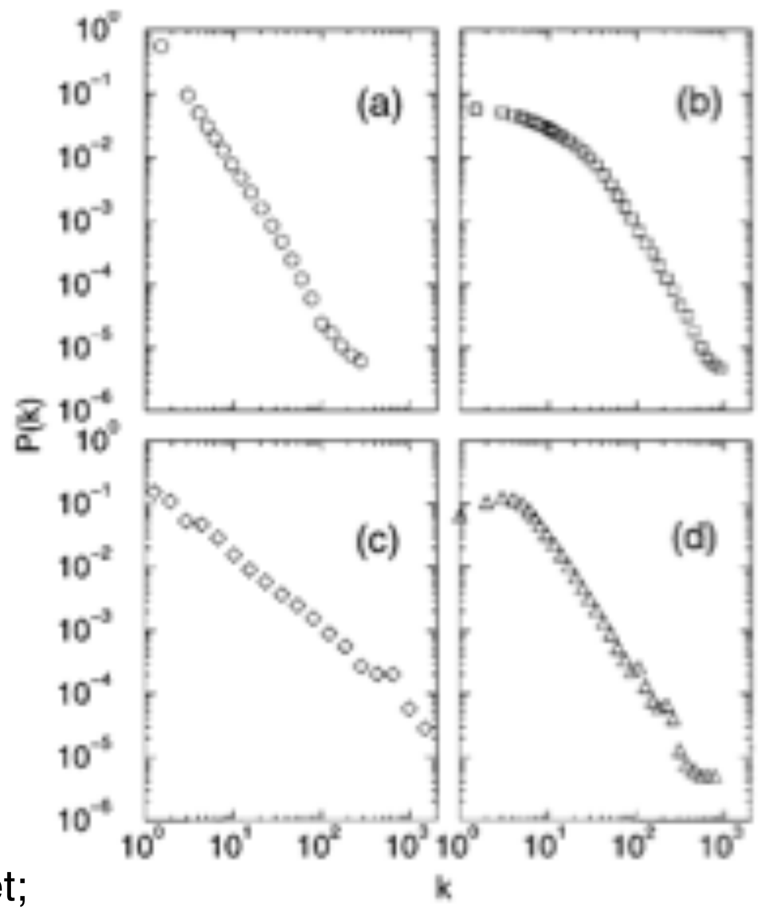
Prediction:

$$P_{rand}(k) \cong C_{N-1}^{k} p^{k} (1-p)^{N-1-k}$$

Data:

$$P(k) \approx k^{-\gamma}$$





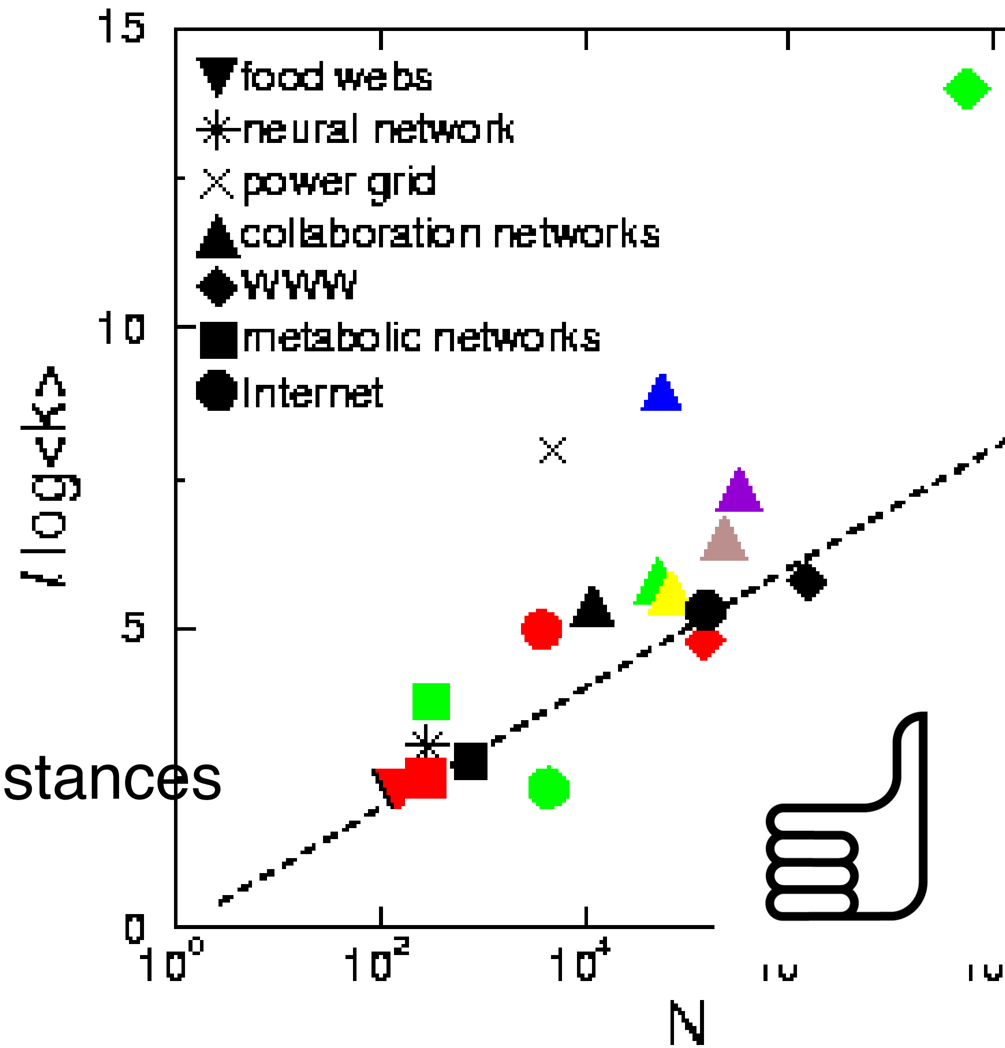
(a) Internet;
(b) Movie Actors;
(c) Coauthorship, high energy physics;
(d) Coauthorship, neuroscience

PATH LENGTHS IN REAL NETWORKS

Prediction:

$$l_{rand} = \frac{\log N}{\log \langle k \rangle}$$

Data:



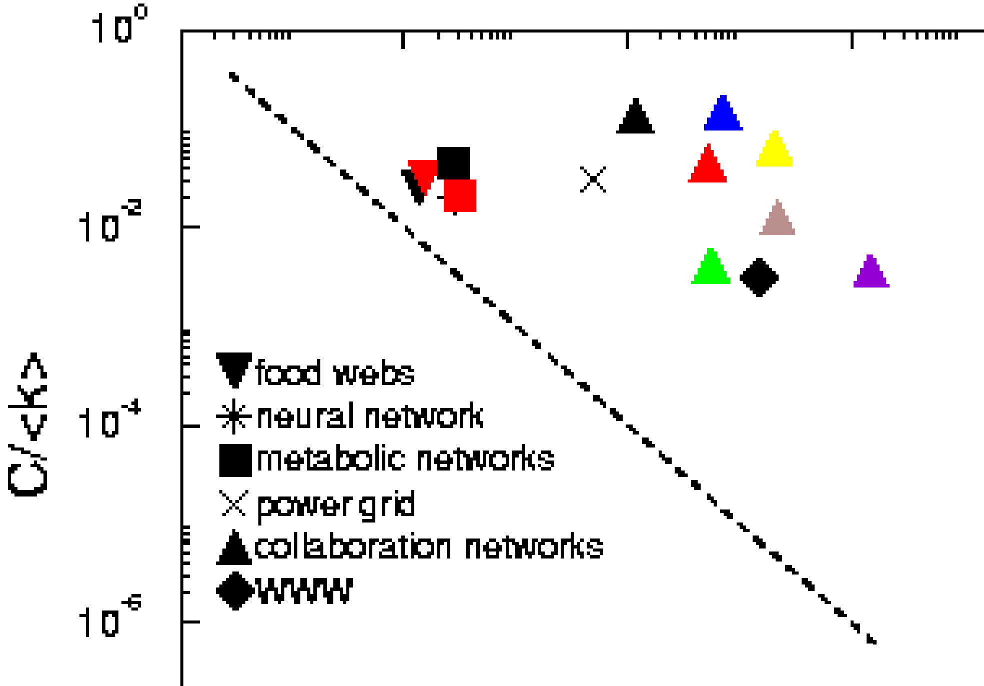
Real networks have short distances like random graphs.

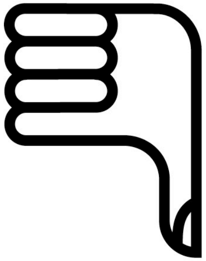
CLUSTERING COEFFICIENT

Prediction:

$$C_{rand} = \frac{\langle k \rangle}{N}$$

Data:

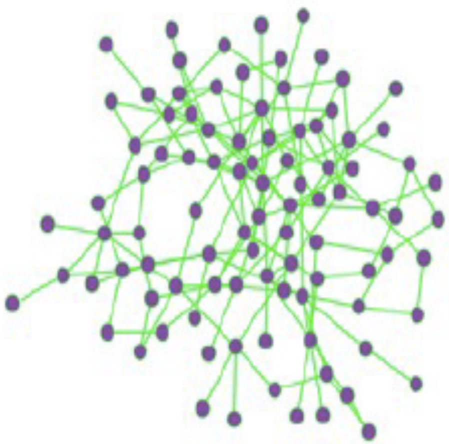


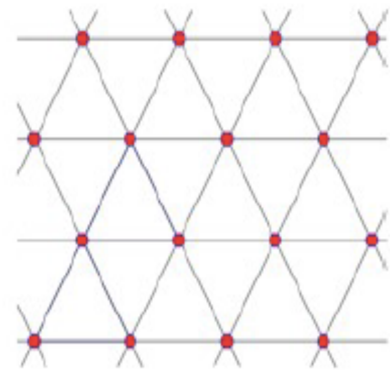


***C_{rand}* underestimates with orders of magnitudes the clustering coefficient of real networks.**

The small-world model

Real networks are between random networks and lattices





Real networks are somewhere here

Watts-Strogatz model

Duncan Watts

Steve Strogatz

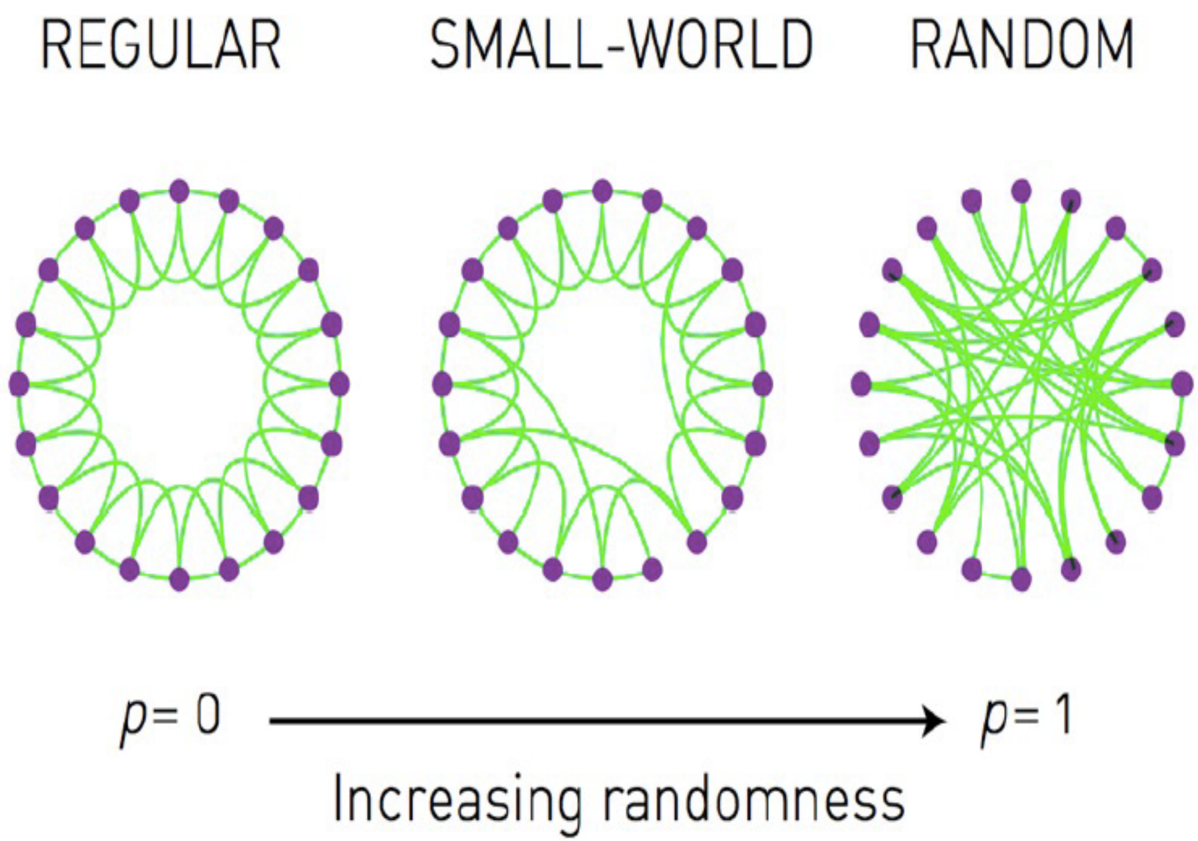
NATURE | VOL 393 | 4 JUNE 1998

Collective dynamics of 'small-world' networks

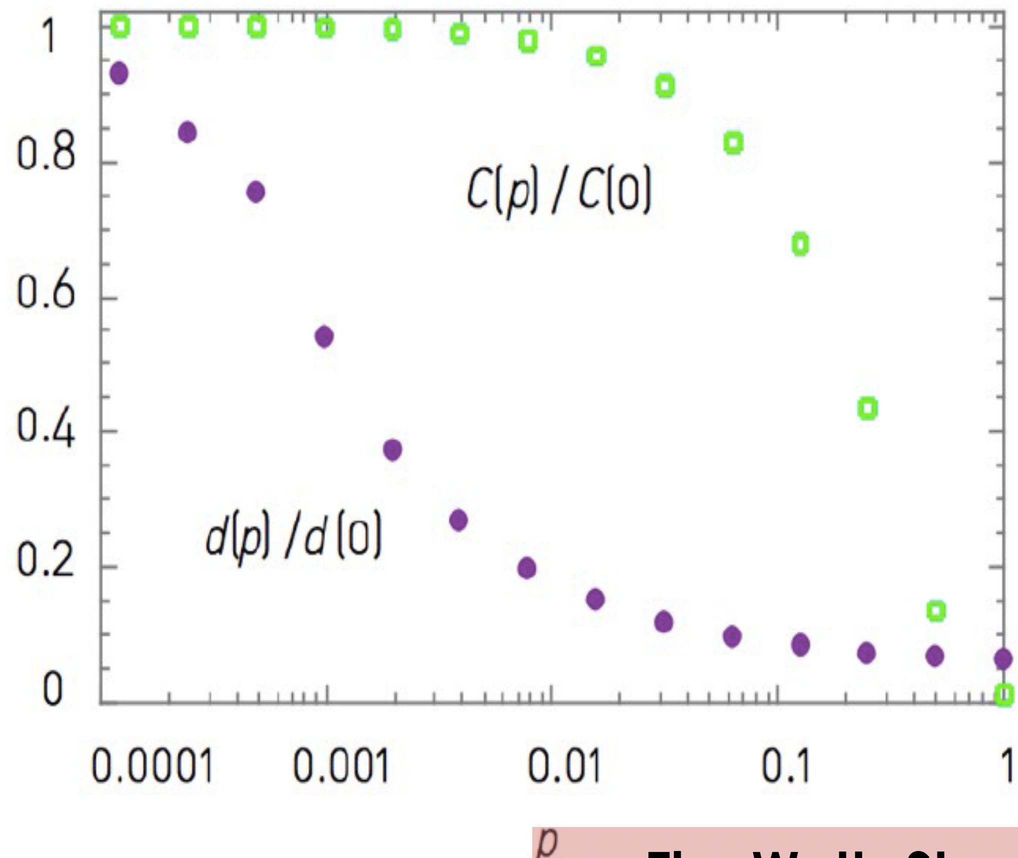
Duncan J. Watts* & Steven H. Strogatz

Department of Theoretical and Applied Mechanics, Kimball Hall, Cornell University, Ithaca, New York 14853, USA

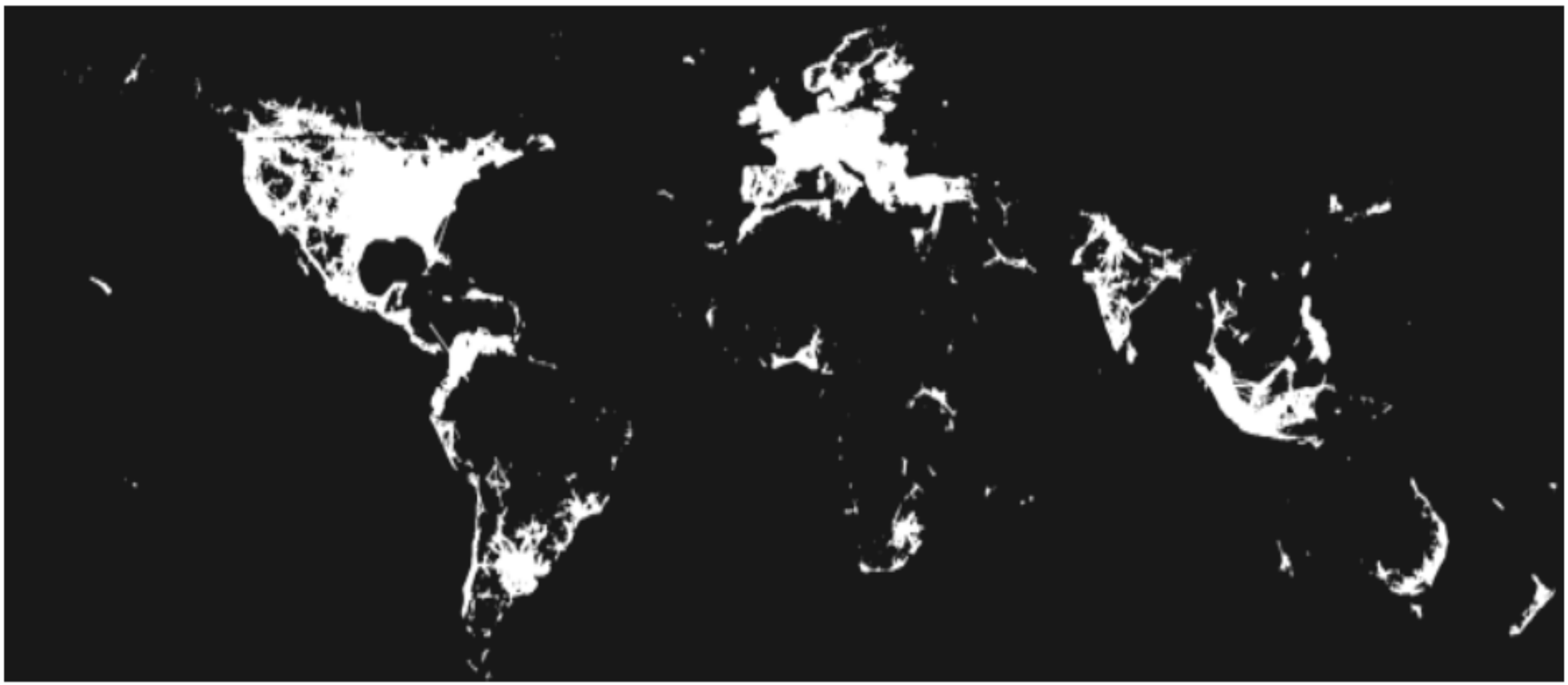
Networks of coupled dynamical systems have been used to model biological oscillators[1–4], Josephson junction arrays[5,6], excitable media[7], neural networks[8–10], spatial games[11], genetic control networks[12] and many other self-organizing systems. Ordinarily, the connection topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes.


REGULAR
SMALL-WORLD
RANDOM
p= 0
p= 1
Increasing randomness

Average path length vs. clustering coefficient



The Watts Strogatz Model:
It takes a lot of randomness to ruin the clustering, but a very small amount to overcome locality



Hubs represent the most striking difference between a random and a scale-free network. Their emergence in many real systems raises several fundamental questions:

•Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?

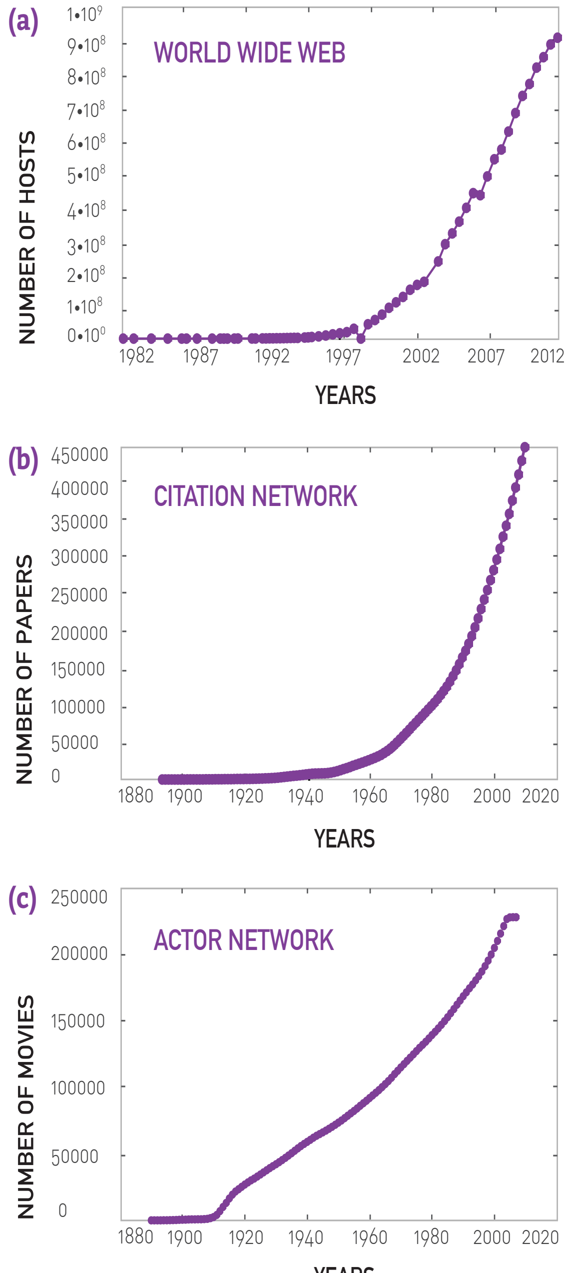
• Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

Growth and preferential attachment

BA MODEL: Growth

ER model:
the number of nodes, N, is fixed (static models)

networks expand through the addition of new nodes



Barabási & Albert, *Science* **286,** 509 (1999)

BA MODEL: Preferential attachment

ER model: links are added randomly to the network

New nodes prefer to connect to the more connected nodes

Barabási & Albert, *Science* **286,** 509 (1999)

Growth and Preferential Sttachment

The random network model differs from real networks in two important characteristics:

Growth: While the random network model assumes that the number of nodes is fixed (time invariant), real networks are the result of a growth process that continuously increases.

Preferential Attachment: While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.

Barabási & Albert, *Science* **286,** 509 (1999)

The Barabási-Albert model

Origin of SF networks: Growth and preferential attachment

(1) Networks continuously expand by the addition of new nodes

WWW : addition of new documents

(2) New nodes prefer to link to highly connected nodes.

WWW : linking to well known sites

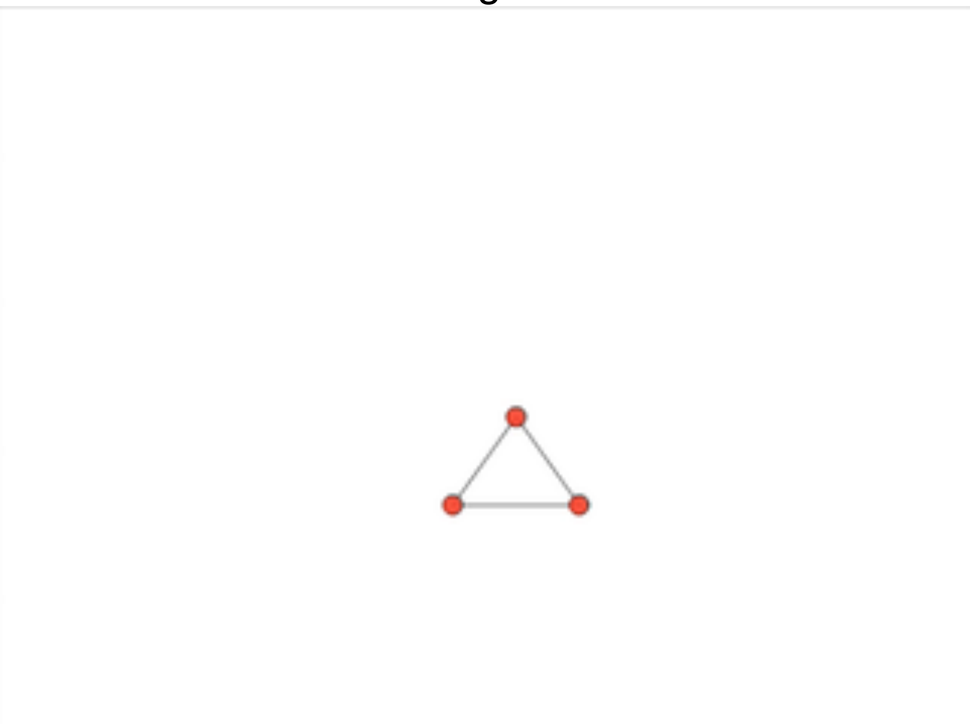
GROWTH:

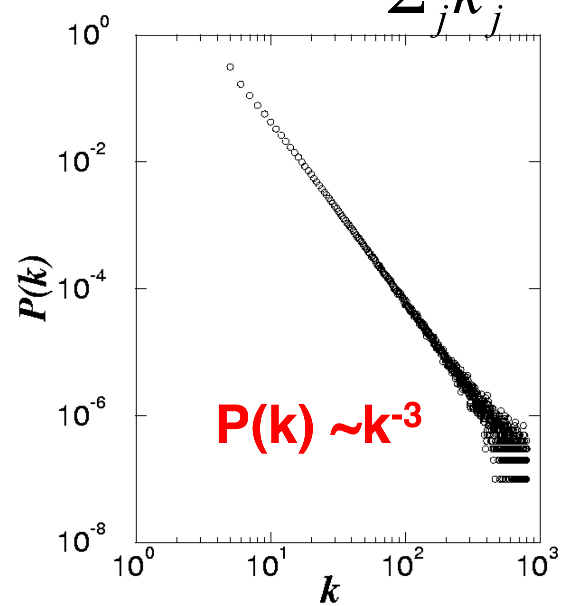
add a new node with m links

PREFERENTIAL ATTACHMENT:

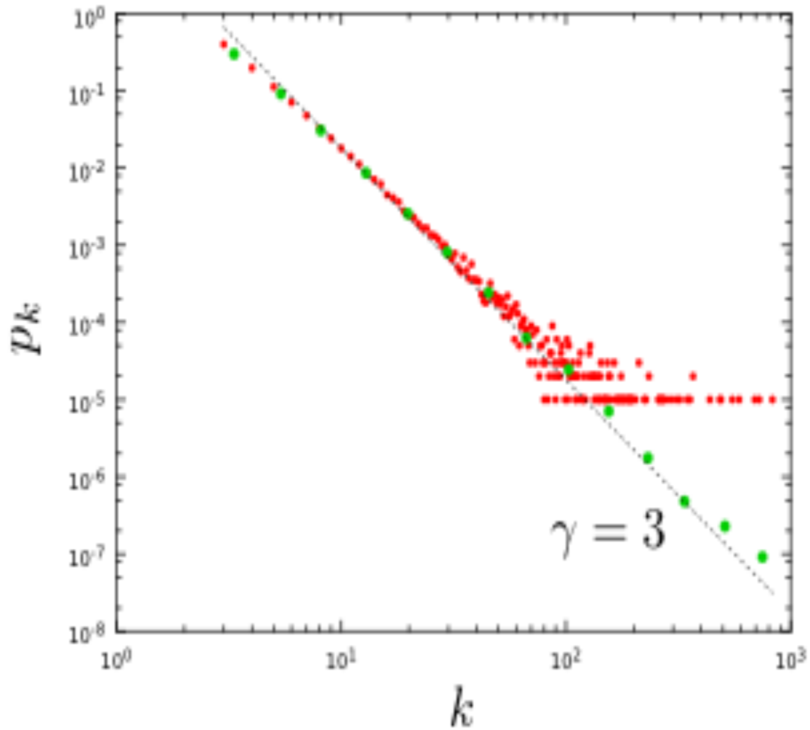
the probability that a node connects to a node with k links is proportional to k.

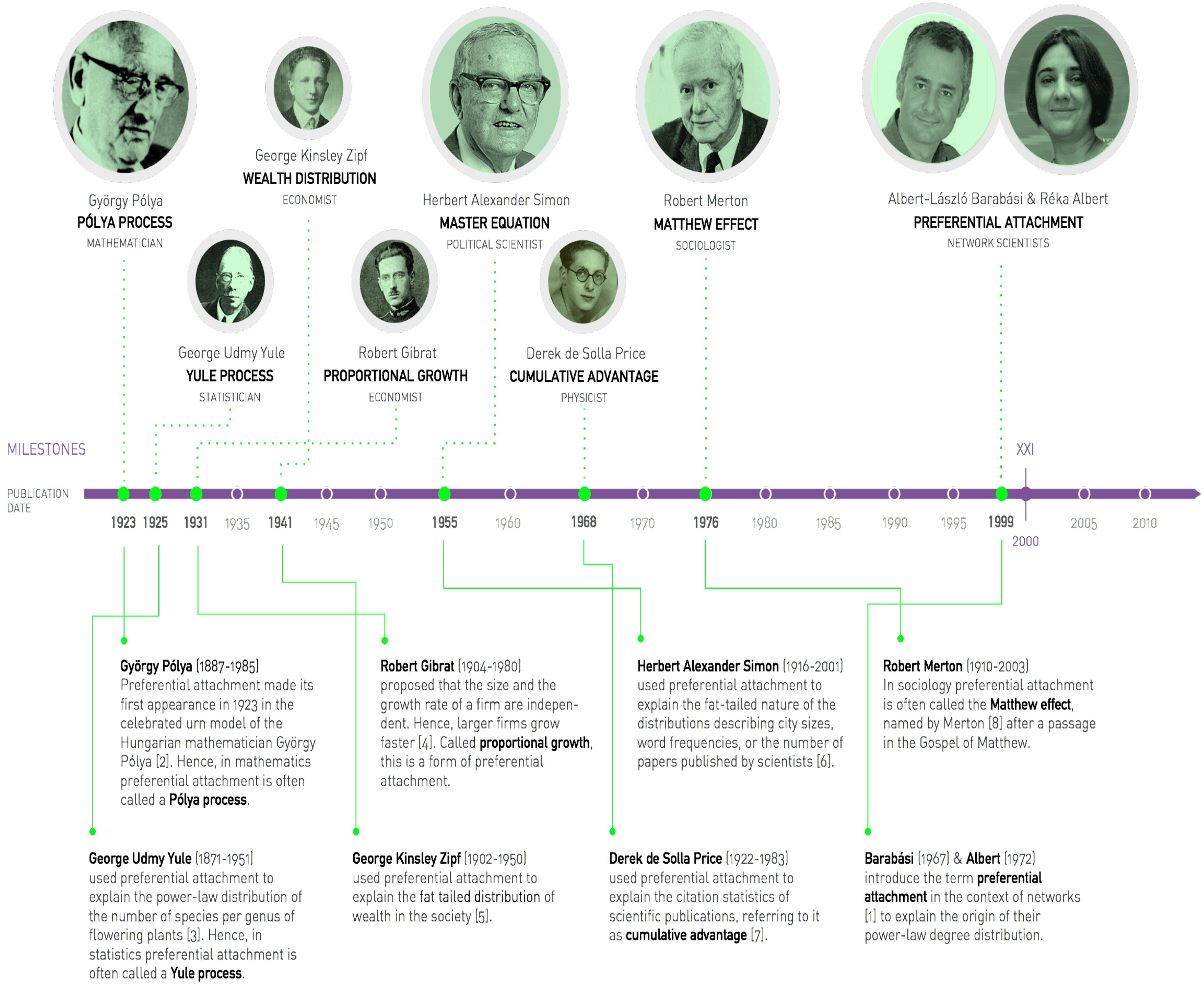
$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$





Barabási & Albert, *Science* **286,** 509 (1999)


$\gamma = 3$
p_k
k



György Pólya
PÓLYA PROCESS
MATHEMATICIAN
George Kinsley Zipf
WEALTH DISTRIBUTION
ECONOMIST
Herbert Alexander Simon
MASTER EQUATION
POLITICAL SCIENTIST
Robert Merton
MATTHEW EFFECT
SOCIOLOGIST
Albert-László Barabási & Réka Albert
PREFERENTIAL ATTACHMENT
NETWORK SCIENTISTS
George Udmy Yule
YULE PROCESS
STATISTICIAN
Robert Gibrat
PROPORTIONAL GROWTH
ECONOMIST
Derek de Solla Price
CUMULATIVE ADVANTAGE
PHYSICIST
MILESTONES
XXI
PUBLICATION DATE
1923 1925 1931 1935 1941 1945 1950 1955 1960 1968 1970 1976 1980 1985 1990 1995 1999 2000 2005 2010
György Pólya (1887-1985) Preferential attachment made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya [2]. Hence, in mathematics preferential attachment is often called a **Pólya process**.
Robert Gibrat (1904-1980) proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called **proportional growth**, this is a form of preferential attachment.
Herbert Alexander Simon (1916-2001) used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].
Robert Merton (1910-2003) In sociology preferential attachment is often called the **Matthew effect**, named by Merton [8] after a passage in the Gospel of Matthew.
George Udmy Yule (1871-1951) used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a **Yule process**.
George Kinsley Zipf (1902-1950) used preferential attachment to explain the fat tailed distribution of wealth in the society [5].
Derek de Solla Price (1922-1983) used preferential attachment to explain the citation statistics of scientific publications, referring to it as **cumulative advantage** [7].
Barabási (1967) & **Albert** (1972) introduce the term **preferential attachment** in the context of networks [1] to explain the origin of their power-law degree distribution.