DATA SCIENCE IN A NETWORKED ERA

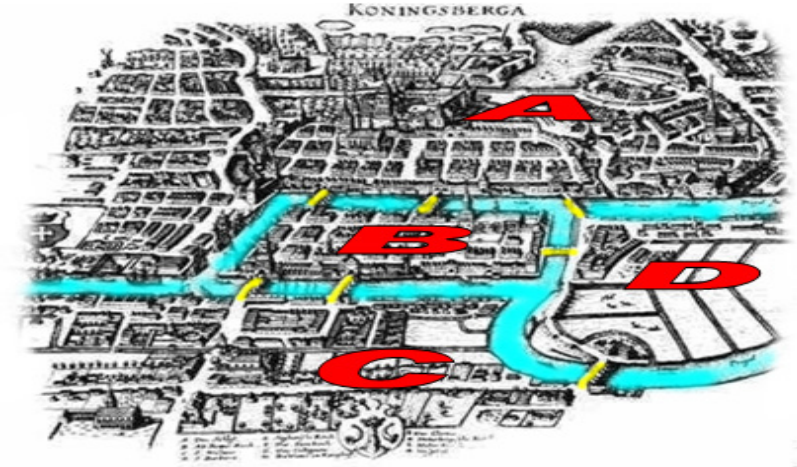
Noemi Derzsy

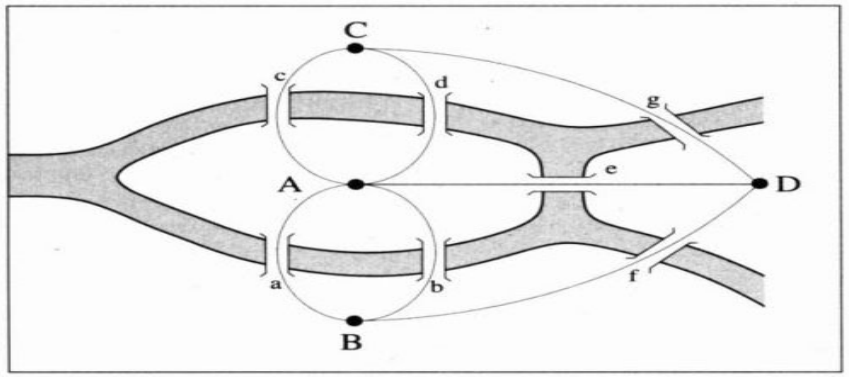
GRAPH THEORY

- Leonhard Euler: Seven bridges of Konigsberg (1735)

The problem: find a path to walk over the 7 bridges in Köngisberg

(each bridge can be passed only once!)





Conditions: 1. the number of bridges touching the islands must be even

2. none or two nodes of odd degree

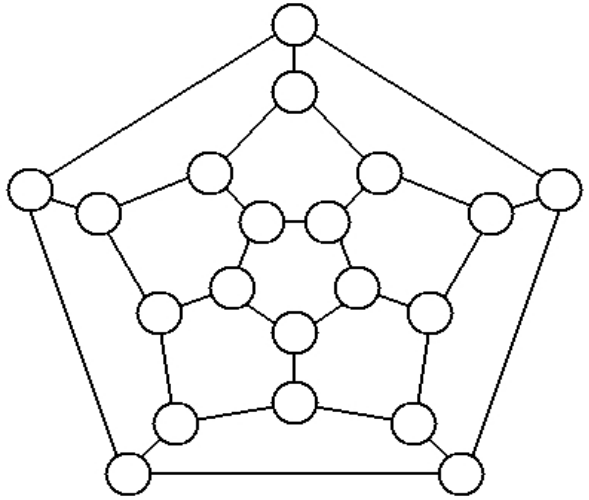
→ Depends on the nodes' degree (number of edges one node has) → NO solution!

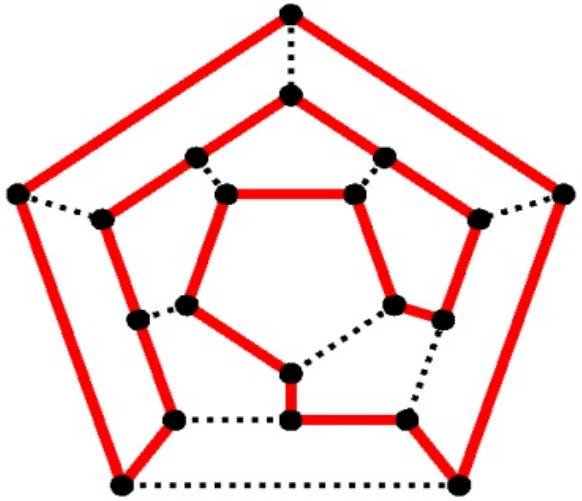
The theorem in the Königsberg bridges problem is the first graph theorem . The most important information: the number of edges and the nodes that are connected. This observation founded the development of mathematical topology.

GRAPH THEORY

William Rowen Hamilton: Icosian game (1856)

The problem: walk over the entire dodecahedron, in a cycle, touching each node only once having the ending node in the starting one





The solution: is a path with twenty edges

Hamiltonian path

The Icosian game defined a very important research topic in graph theory: finding a route in a graph that walks over the system in the most optimal way (passing through each node only once).

NETWORKS

- A graph is an abstract representation of a set of objects connected by links $G = \{P,E\}$.

networks = large graphs

- *Complex systems* can be studied through their *underlying network structure*: nodes (elements), links (interaction)
- The way we assign the nodes and links, defines the problem and questions we can study

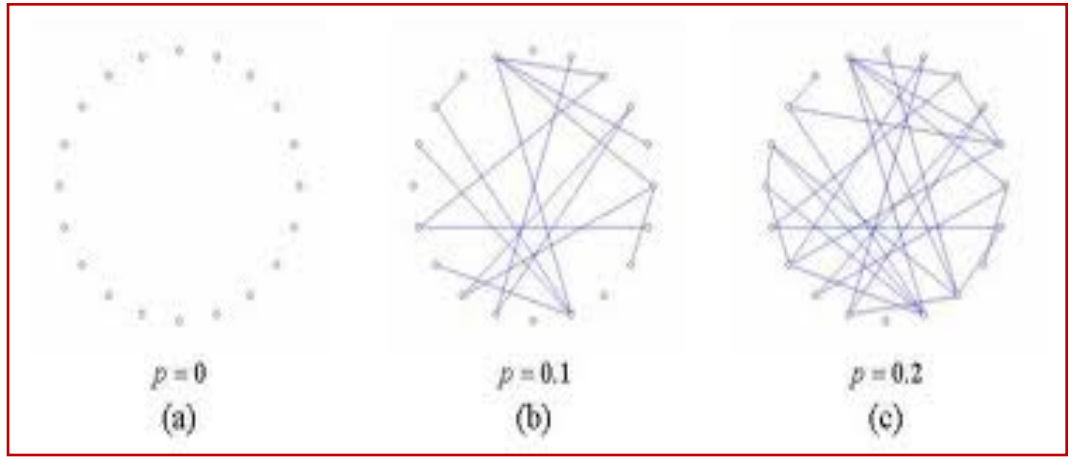
Basic structural properties

- Degree distribution P(k): the probability that a randomly selected node in the network has ***k*** connections
- Clustering coefficient: $C = \frac{3 \times triangles}{triples}$
- Average path length: the average of the smallest distance between two randomly selected nodes
- Diameter: the maximal distance between two elements in a graph
- Centrality: the importance of a node relying on its position, the level of contribution an agent has in a network

And much more….

RANDOM (ERDOS-RENYI) NETWORKS

Model proposed by
P. Erdős and A. Rényi



- *Degree distribution: Poisson distribution*
- *Small clustering coefficient*
- *High average path length*

Problems of the model

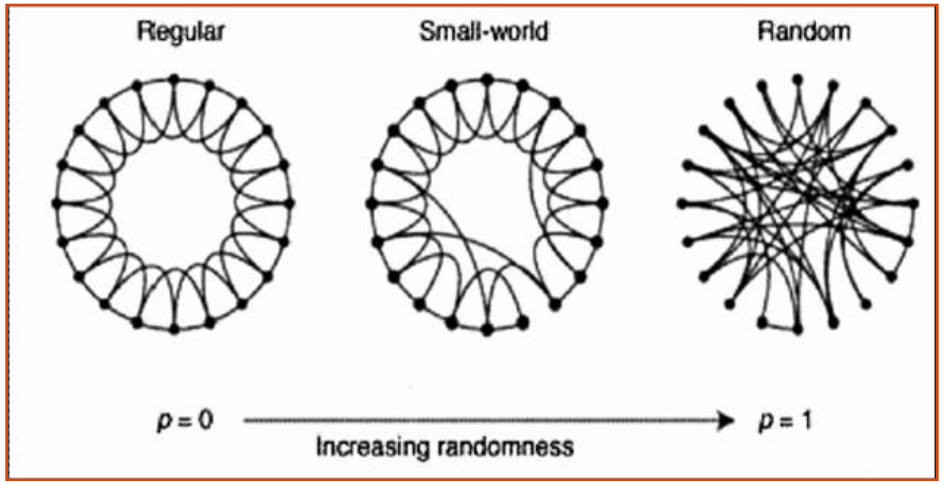
- ❑ Studies showed that real complex systems: do not have Poisson distribution!
- ❑ Several clusters were observed in real complex systems that are not reproduced by this model
- ❑ Presence of hubs in real complex systems

→ **A DIFFERENT MODEL IS NEEDED!**

P. Erdős, A. Rényi. (1959). On random graphs, I., Publ. Math. Debrecen 6.

SMALL-WORLD (WATTS-STROGATZ) NETWORKS

Model proposed by
D.J. Watts and S.H. Strogatz



- *Degree distribution: exponential*
- *High clustering coefficient*
- *Small average path length*

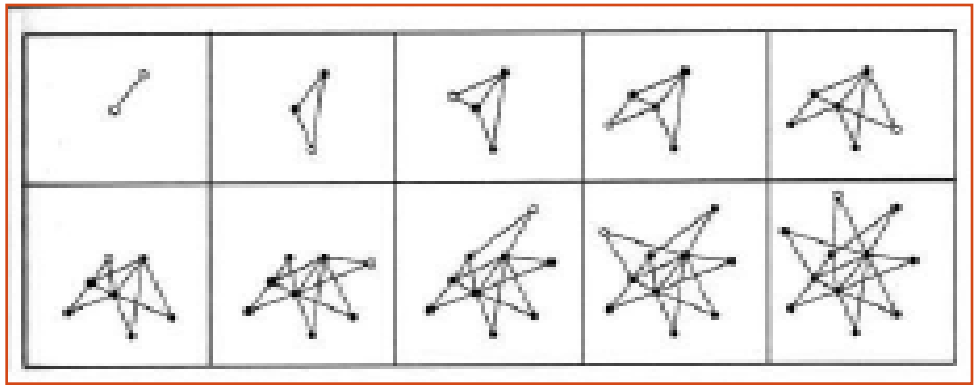
Problems solved

- ✓ Studies showed that real complex systems: do not have Poisson distribution!
- ✓ Several clusters were observed in real complex systems that are not reproduced by this model
- ❑ Presence of hubs in real complex systems → **The model can't reproduce it!**

D. J.Watts, S. H. Strogatz.(1998). Nature (London) 393, 440

SCALE-FREE (BARABASI-ALBERT) NETWORKS

Model proposed by
A.-L. Barabasi and R. Albert



- *Degree distribution: Poisson distribution*
- *Small clustering coefficient*
- *High average path length*

Problems solved

- ✓ Studies showed that real complex systems: do not have Poisson distribution!
- ✓ Several clusters were observed in real complex systems that are not reproduced by this model
- ✓ Presence of hubs in real complex systems

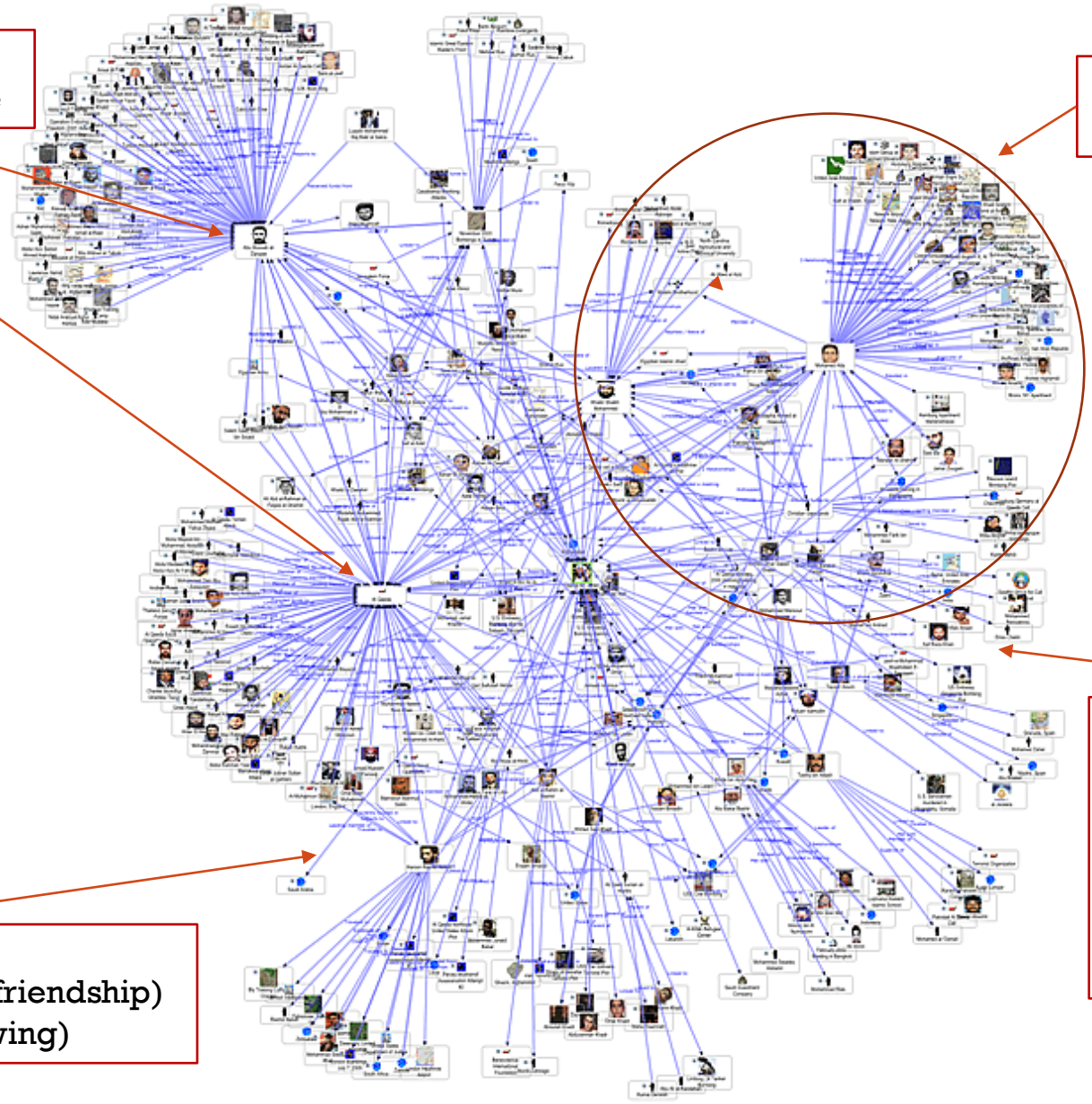
→ ***preferential attachment*** (the model's key ingredient)

R. Albert, A.-L. Barabási. (2002). Statistical mechanics of complex networks, Reviews of Modern Physics 74, 47-97.

Social Media Friendship Network

Hub: a node that has significantly more connections than the average

Cluster: densely connected community



Link direction:

- undirected (Facebook friendship)
- directed (Twitter following)

Link weight:

- unweighted, single connection (either friend or not)
- weighted, multiple connections (number of exchanged e-mails)

SOCIAL NETWORKS

- Friendship
  - Facebook
  - Twitter
  - LinkedIn
  - Instagram
- Collaboration
  - Actors
  - Scientists
  - Co-authorship
  - Student scholarships
  - Employer network
- Criminal organizations
  - Terrorist groups
  - Organized crime
- Sexual relationships
- Matching problem
- Medieval societies

BIOLOGICAL NETWORKS

- Epidemics
- Brain neural network
- Protein interaction
- Species network
- Medicine/treatment
- Cancer research

INFRASTRUCTURE NETWORKS

- Power grid systems
- Computer networks
- World Wide Web (WWW)
- Road traffic
- Airline traffic

FINANCIAL NETWORKS

- Trade market
- Financial transactions
- Lender/borrower system
- Business relationships

COMMUNICATION NETWORKS

- Mobile communication
- E-mail communication

OTHER APPLICATIONS

- Politics
  - Elections: opinion influencing
  - Foreign relations
- Ecology
  - Food chain
  - species
- Environment
- Literature
  - Word relationships
  - Genre networks
  - Novel character network
- News relationships
- Or any other complex system, where you can define a relationship (link) between elements…

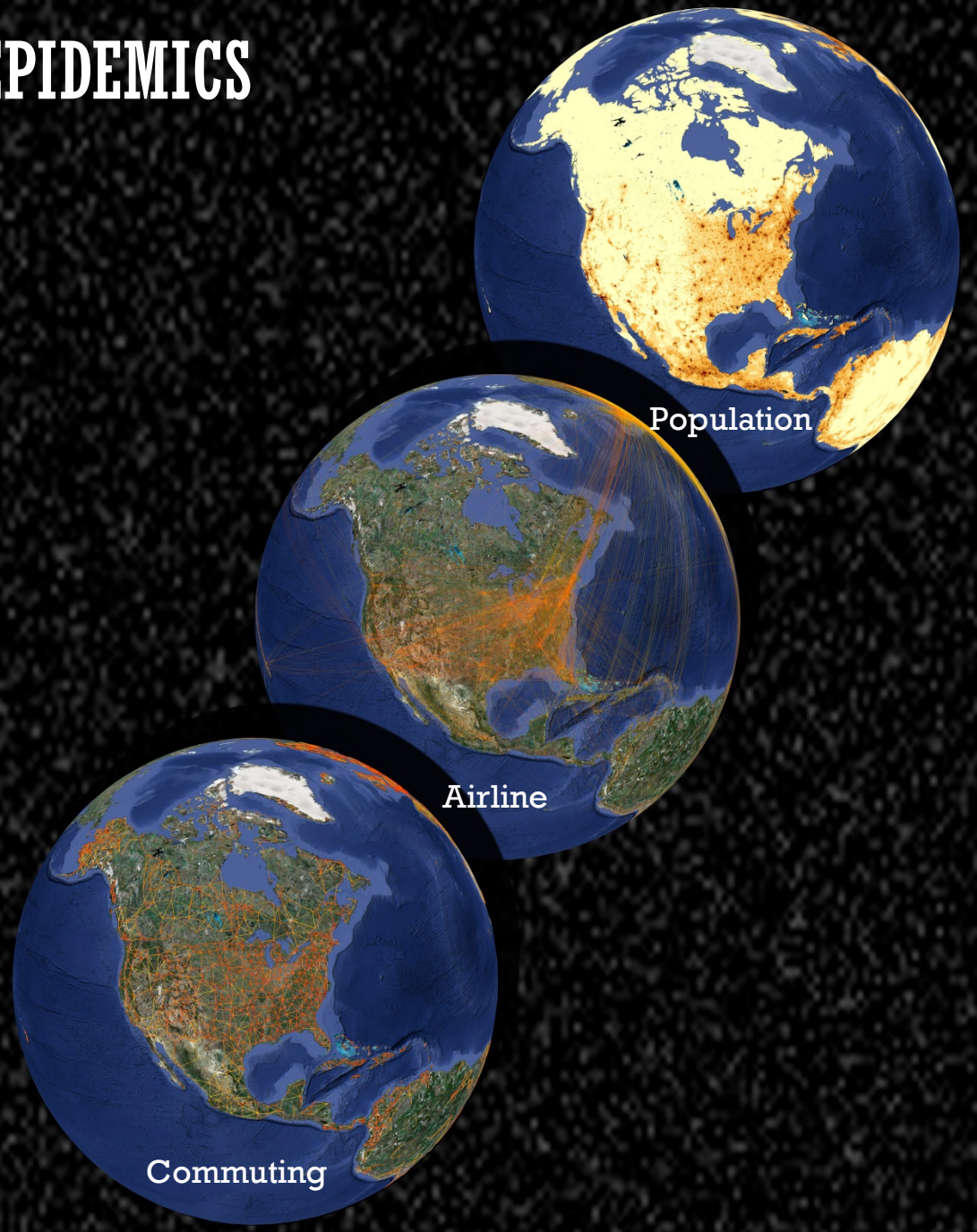
DISEASE OUTBREAK, EPIDEMICS

Data: airline traffic, census data, transportation data
Goal: model and predict the spread of epidemics

The blue network represents short-range commuting flows by car, train and other means of transportation and transport infrastructures. Yellow-to-red lines denote airline flows for a few selected cities; red corresponds to greater traffic intensity. Population density is identified on the grey/white colour scale, with white corresponding to areas of higher density. All features in this map were obtained from real data.

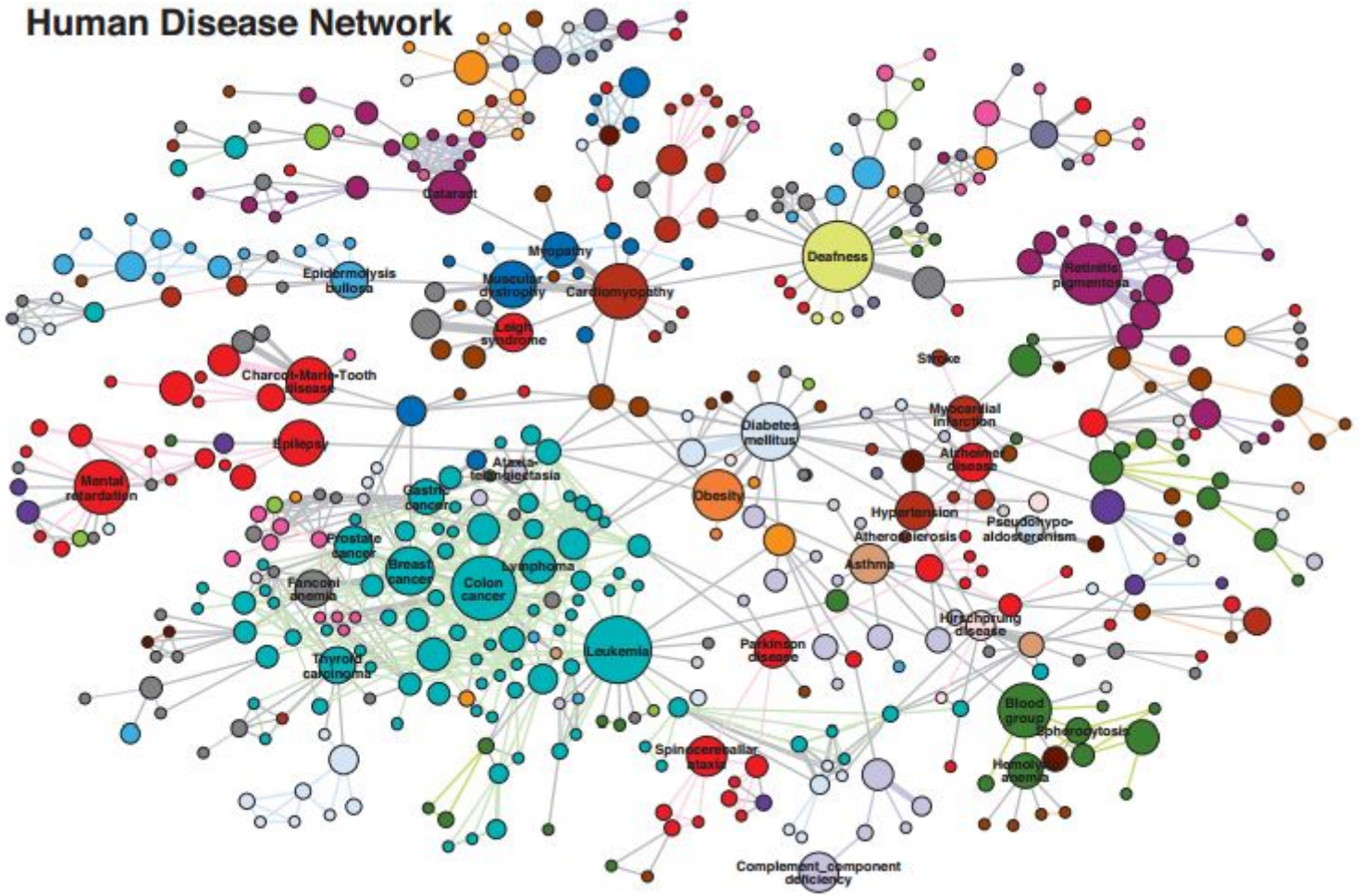
Real-time forecasting of epidemic spreading at the global scale

GLEaMViz
Software system for the simulation of infectious diseases on a global scale.
http://www.gleamviz.org/simulator/

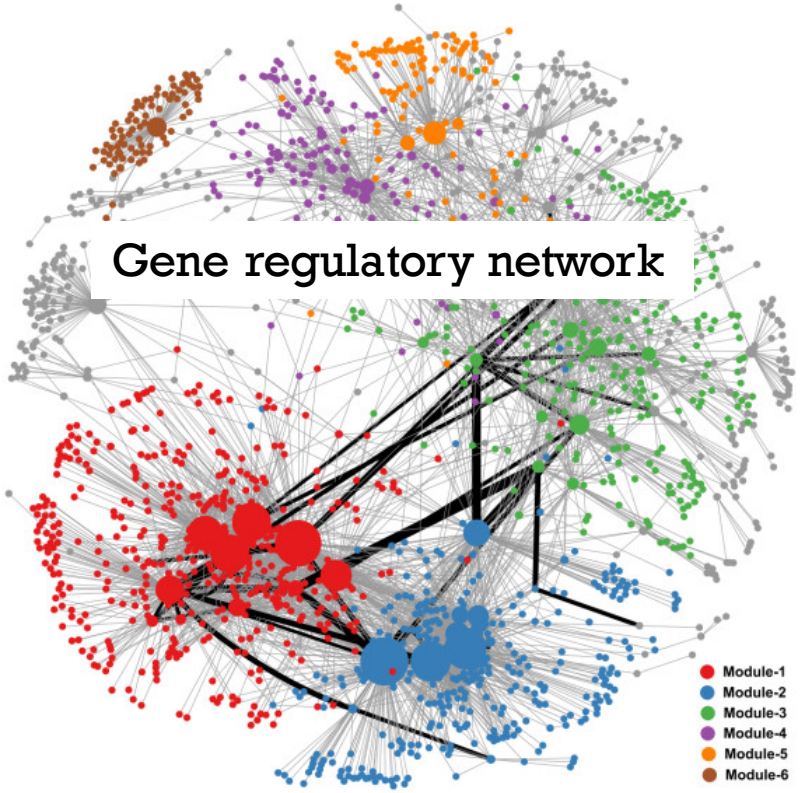


by Bruno Gonçalves and Alessandro Vespignani

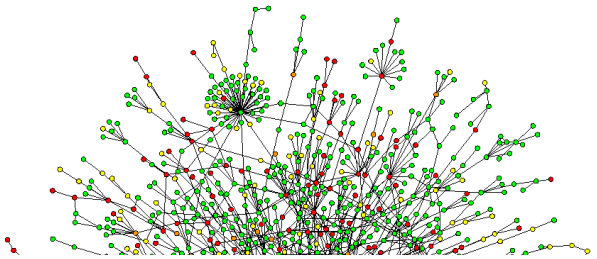
BIOLOGICAL NETWORKS



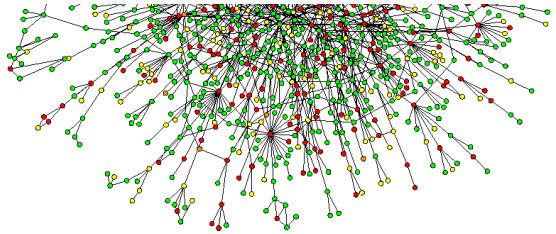
The human disease network
Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási
PNAS 2007



Gene regulatory network

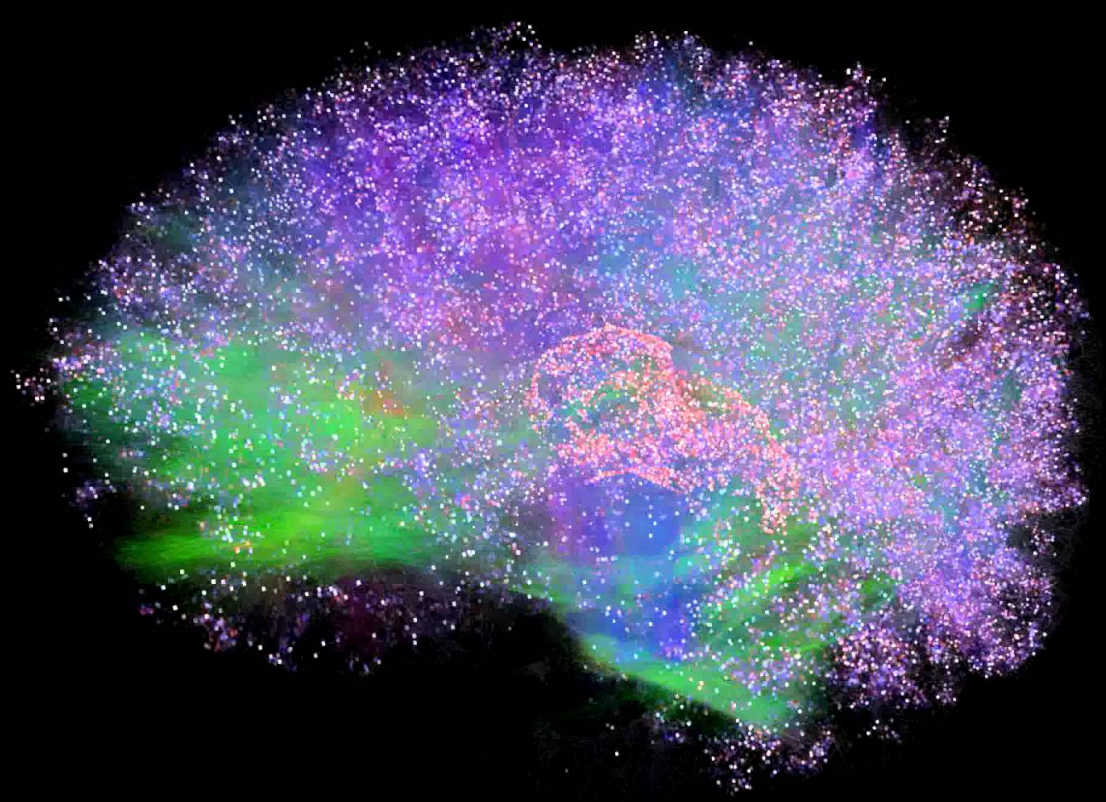


Protein interaction network



BRAIN NETWORKS

Data: brain fMRI
Goal: study brain structure and function, diseases (Alzheimer, dementia, etc.)



Simulated Thalamocortical Brain Network 6 - 3D View, 3 Million Neurons, 476M Synapses

A "connectome," or map of neural pathways and wires, of a human brain(Credit: Human Connectome Project)

POWERGRID NETWORK

Data: UCTE (Union of the Coordination of the Transmission of Electricity) European powergrid data
Goal: study cascading failures and mitigation strategies

A cascading failure is a failure in a system of interconnected parts in which the failure of a part can trigger the failure of successive parts.
Important vulnerability of networked systems.
Can occur in:

- Biological systems
- Infrastructure networks
- Financial systems
- Power grids:
  - → one element fails
  - → shifts its load to its neighboring elements; those elements are then pushed beyond their capacity
  - → some become overloaded and fail
  - → their loads are shifted onto their neighbors

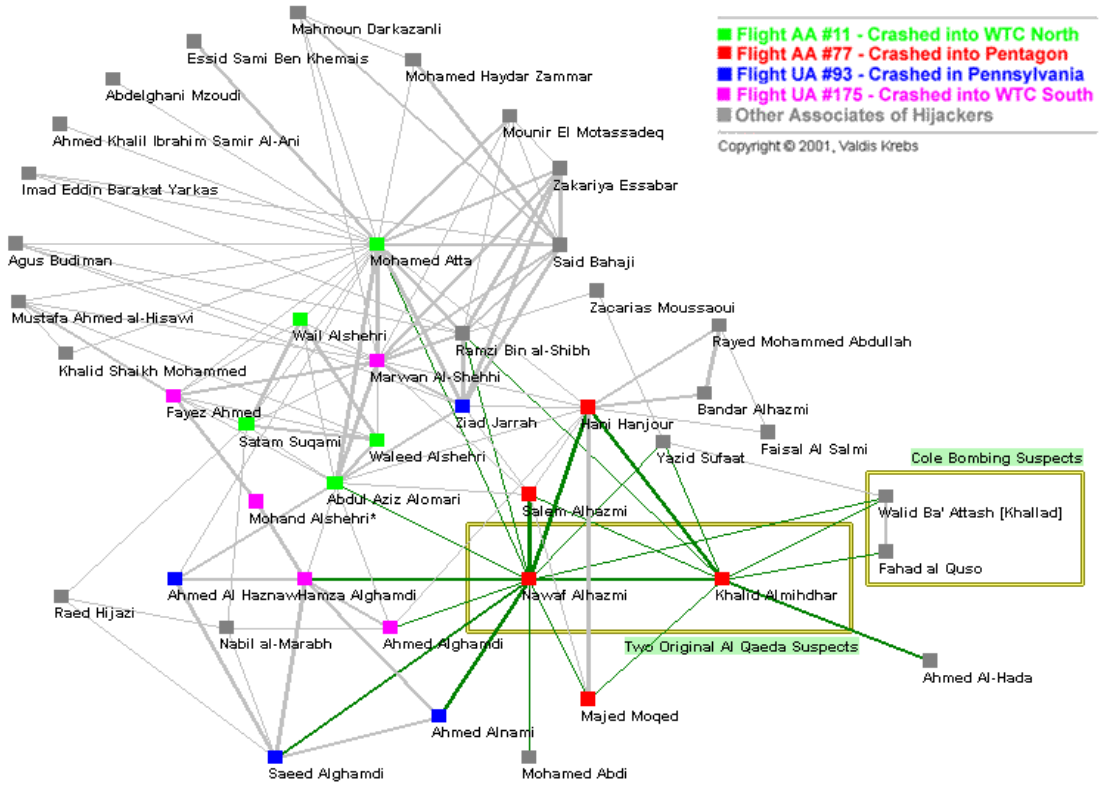
Power blackout in New York City (2012)

Blackout simulation in the European powergrid

Cascading Failures in Spatially-Embedded Random Networks
A. Asztalos, S. Sreenivasan , B. K. Szymanski, G. Korniss
PLoS ONE 9(1): e84563 (2014)

TERRORISM, ORGANIZED CRIME NETWORKS

9/11 Terrorist Group



- *All* 19 hijackers were within 2 steps of the two original suspects uncovered in 2000
- Social network metrics reveal Mohammed Atta *emerging* as the local leader

Uncloaking Terrorist Networks, Valdis Krebs
http://firstmonday.org/ojs/index.php/fm/article/view/941/863
http://www.orgnet.com/prevent.html

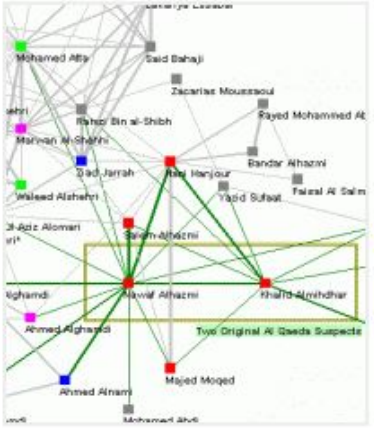
DigitalTonto
At the Crossroads of Media, Marketing and Technology...
Home

How The NSA Uses Social Network Analysis To Map Terrorist Networks

2013 JUNE 12

by Greg Satell

tags: Duncan Watts, Network Theory, Privacy, Social Network Analysis, Social Networks



Ever since *The Guardian* reported that the National Security Agency (NSA) has been collecting the phone record metadata of millions of Americans, the cable talk circuit has been ablaze with pundits demanding answers to what should be obvious questions.

Who knew about the program to collect data? (Apparently, all three branches of government). Who else has been supplying data? (Just about everybody, according to the *Washington Post*). What is metadata? (It's data about data).

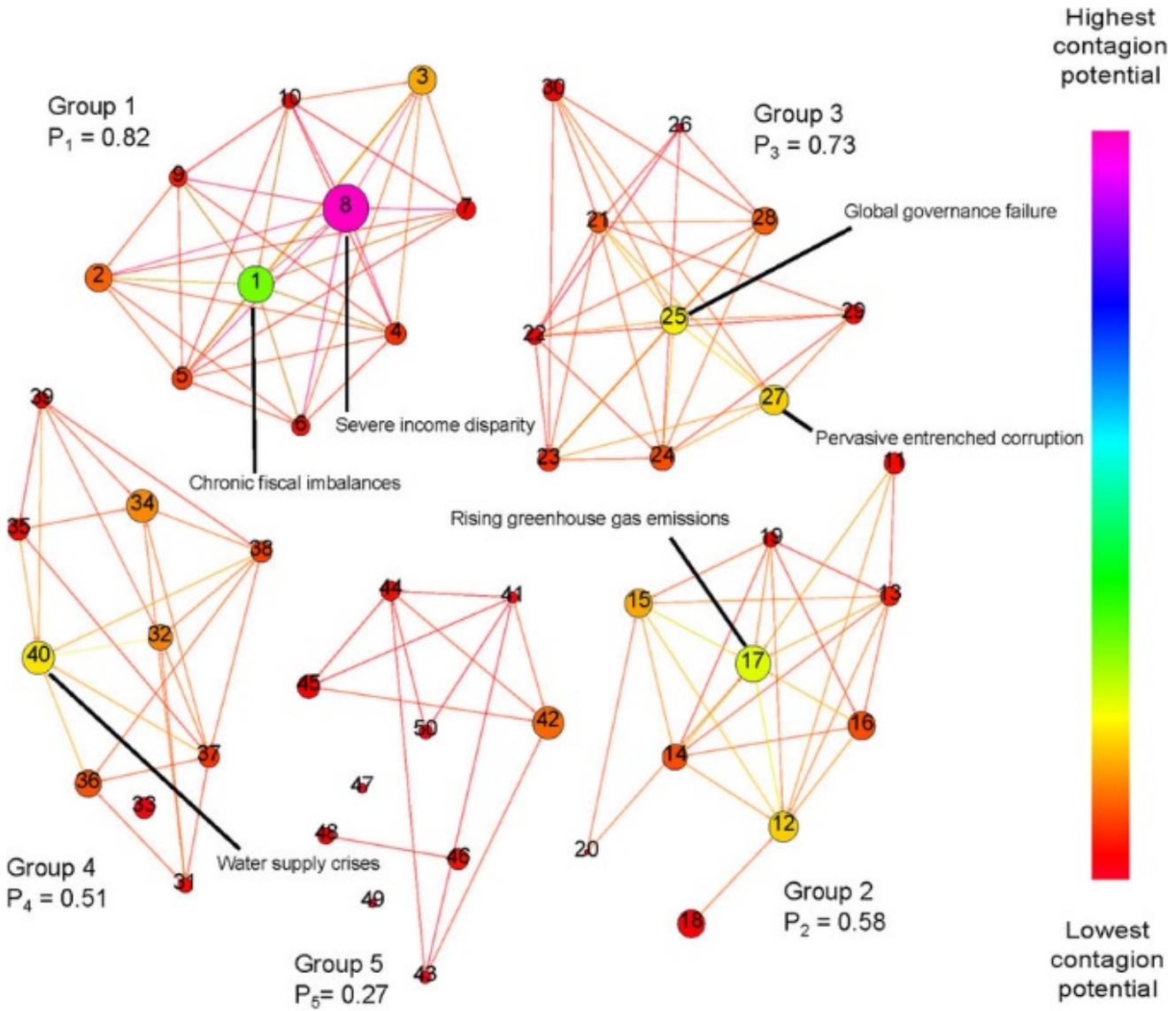
http://www.digitaltonto.com/2013/how-the-nsa-uses-social-network-analysis-to-map-terrorist-networks/

GLOBAL RISK NETWORK

Risks threatening modern societies form an intricately interconnected network that often underlies crisis situations.

Goal: study how risk materializations in distinct domains influence each other.

Data: WEF Report on Global Risks



Each node is sized proportionally to its internal failure probability while node color corresponds to its total contagion potential. The number of edges in each group shows the intra-group connectivity. The nodes with the highest congestion potential are identified by name.

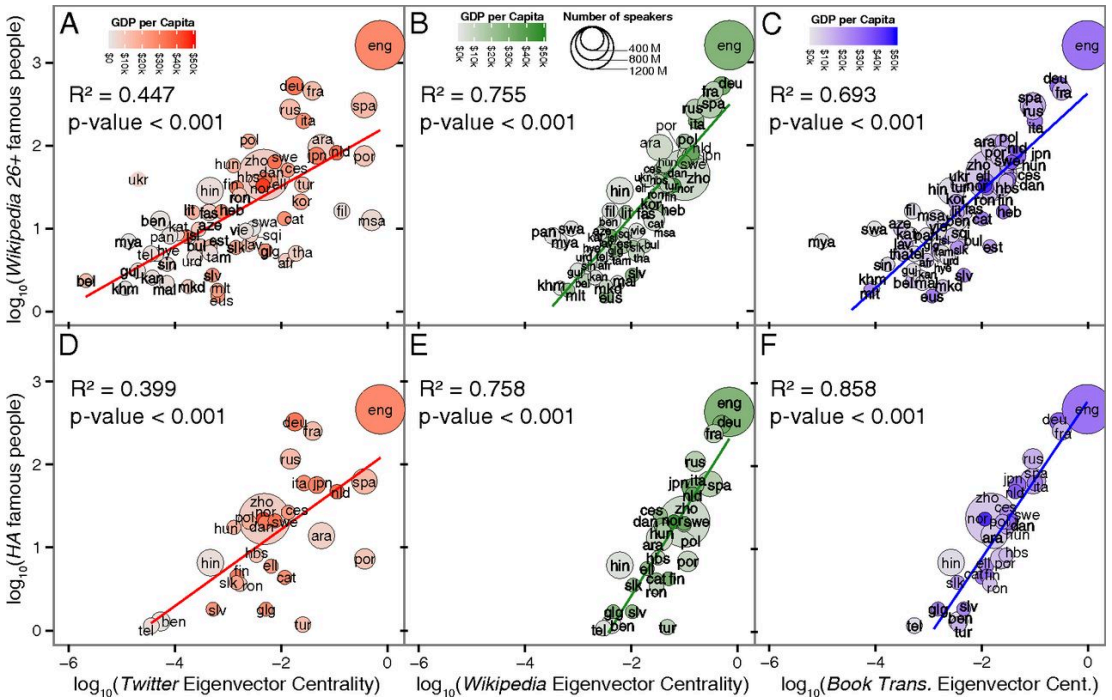
Failure dynamics of the global risk network
B. K. Szymanski, X. Lin, A. Asztalos & S. Sreenivasan
Scientific Reports 5, 10998

GLOBAL LANGUAGE NETWORK (GLN)

Data: Twitter, Wikipedia, UNESCO book translations

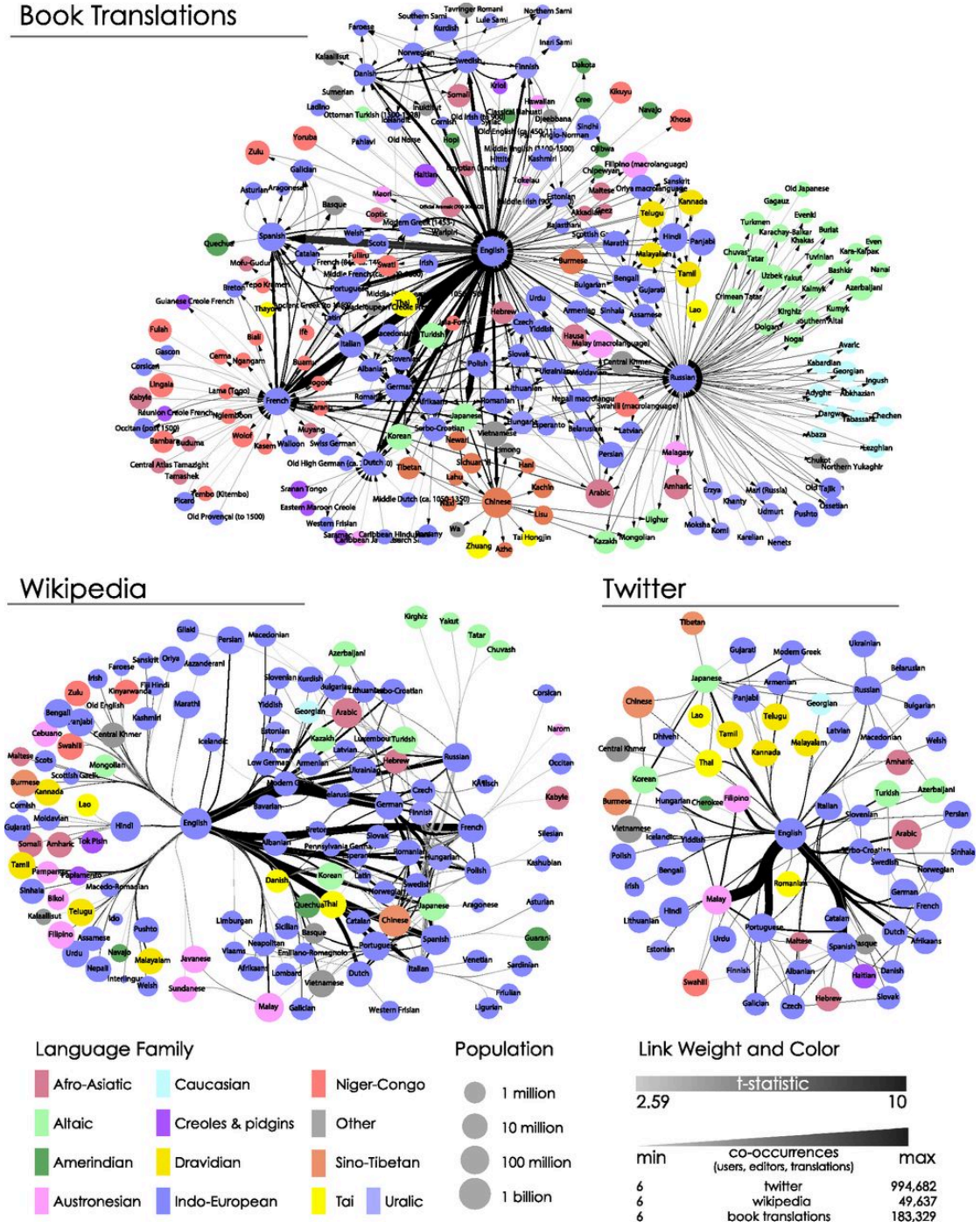
Goal: study the importance of languages and their impact

The position of a language in the GLN and the global impact of its speakers



The position of a language in the GLN contributes to the visibility of its speakers and the global popularity of the cultural content they produce.

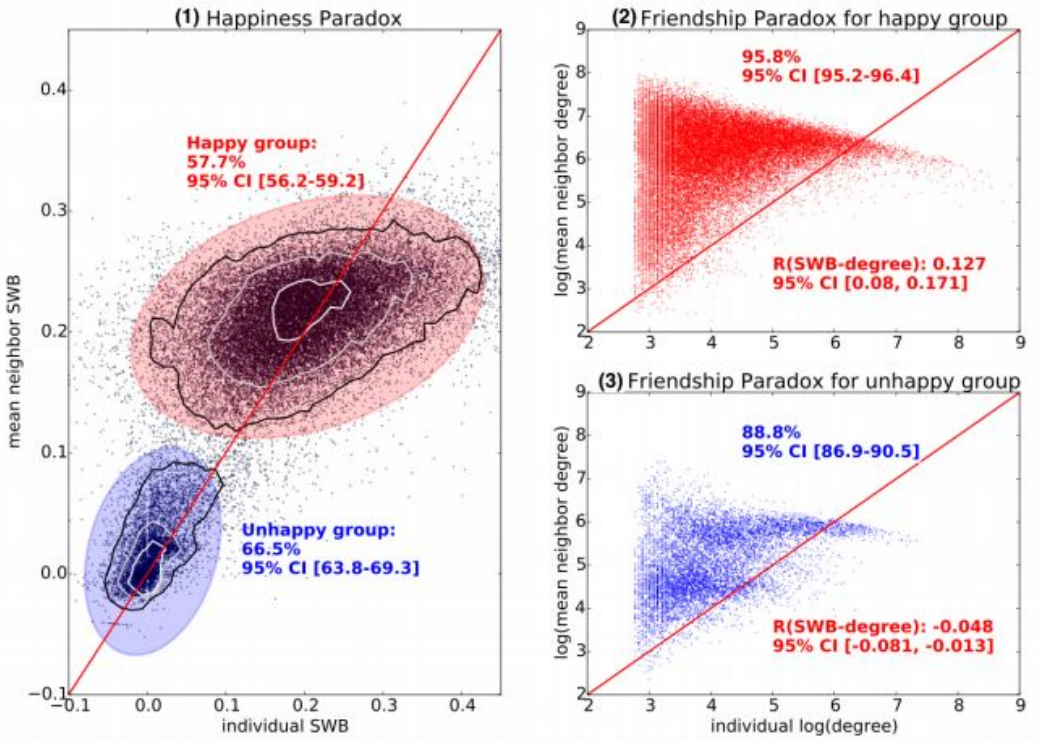
Links that speak: The global language network and its association with global fame.
S. Ronen, B. Gonçalves, K. Z. Hu, A. Vespignani, S. Pinker, and C. A. Hidalgo
PNAS 2014;111:E5616-E5622



FRIENDSHIP NETWORKS

Data: Twitter

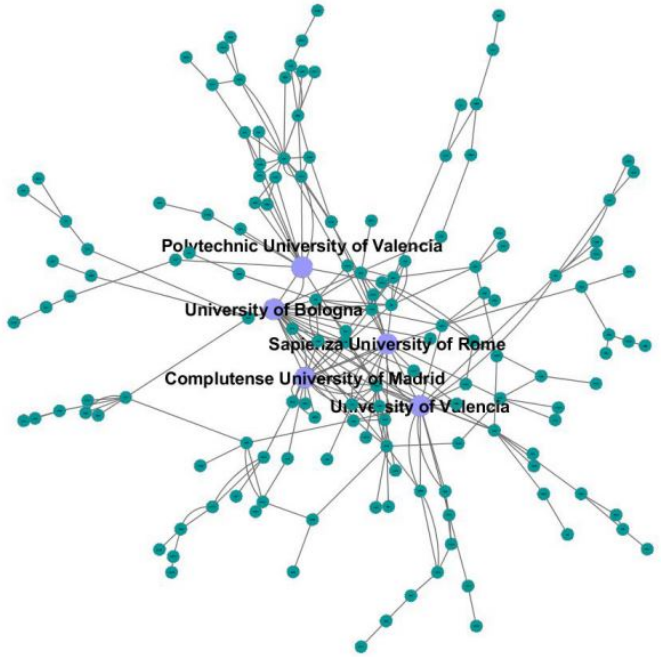
Goal: study the Happiness Paradox (most individuals in social networks experience a so-called Friendship Paradox: they are less popular than their friends on average)



Popular individuals are indeed happier and a majority of individuals experience a significant Happiness paradox.

The happiness paradox: your friends are happier than you
J. Bolen, B. Gonçalves, I. van de Leemput, G. Ruan
arXiv: 1602.02665

COLLABORATION NETWORKS

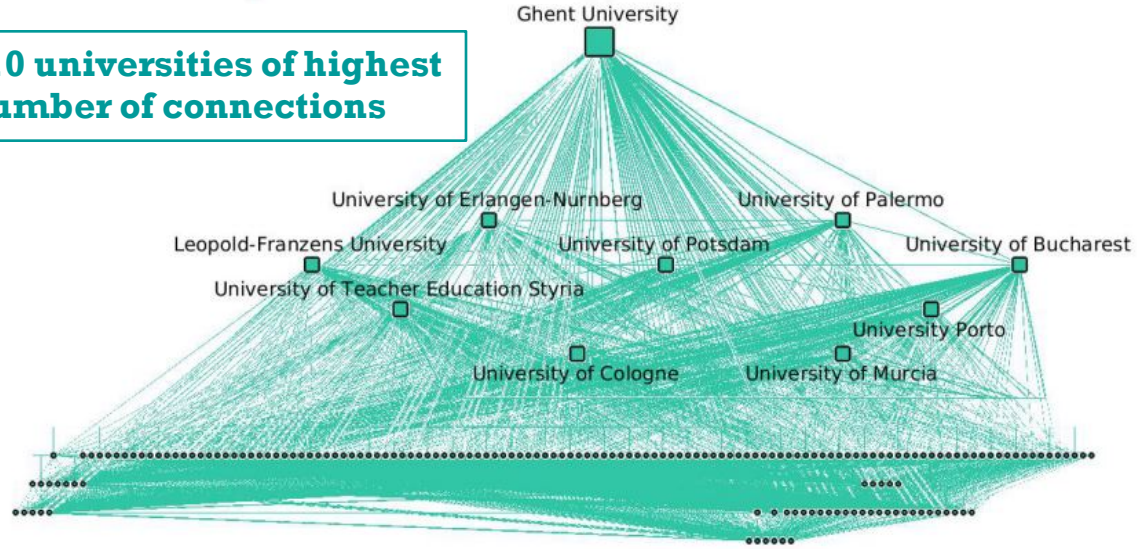


Data: European Erasmus scholarships

Goal: study the academic collaborations among universities through student mobility

The 5 most densely connected universities in Europe

Top 10 universities of highest number of connections

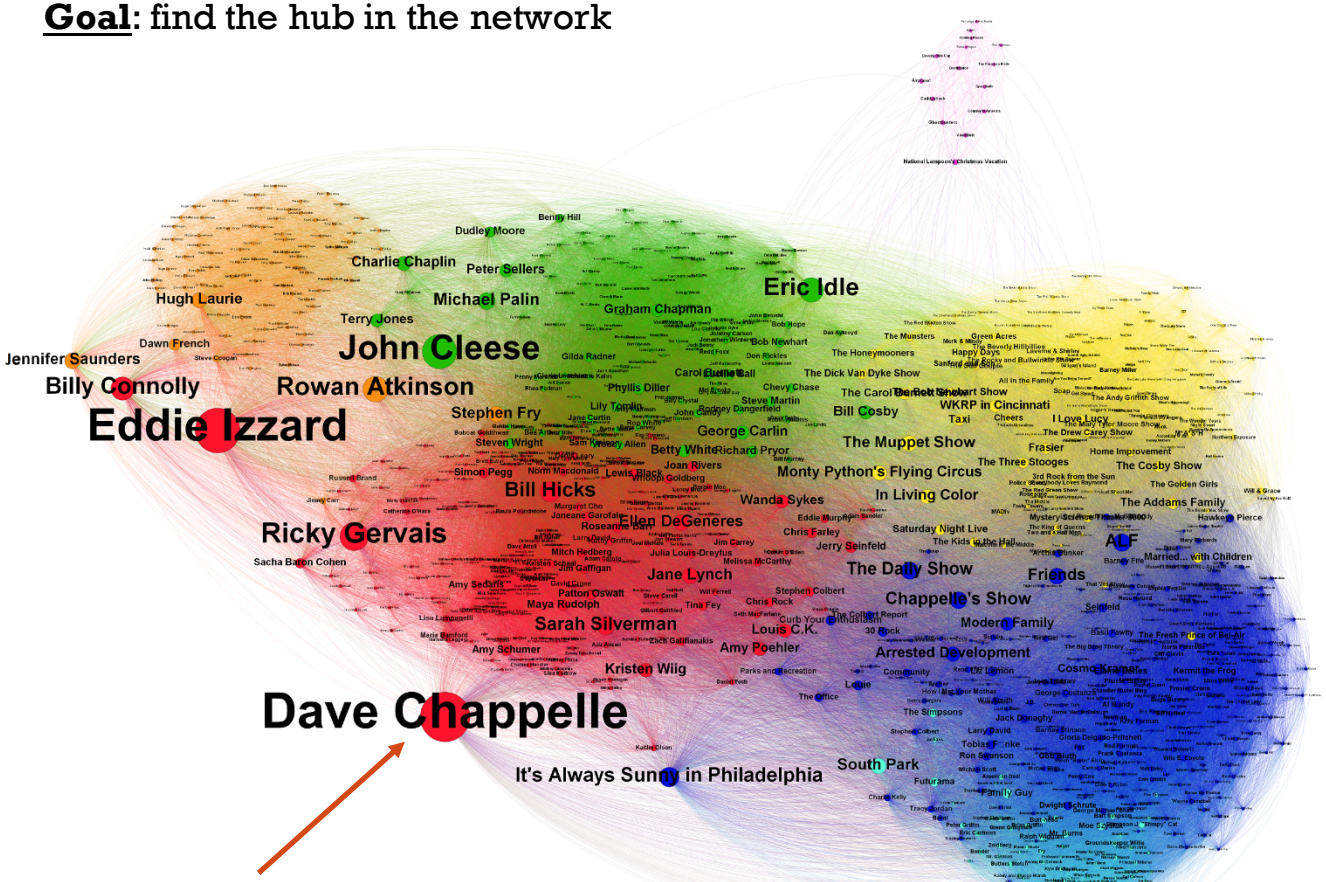


Topology of the Erasmus student mobility network
A. Derzsi, N. Derzsy, E. Kaptalan and Z. Neda
Physica A. 2011;390(13):2601–2610

OTHER SOCIAL NETWORKS

A Ranker World of Comedy Opinion Graph: Who Connects the Funny Universe?

Data: Twitter
Goal: find the hub in the network

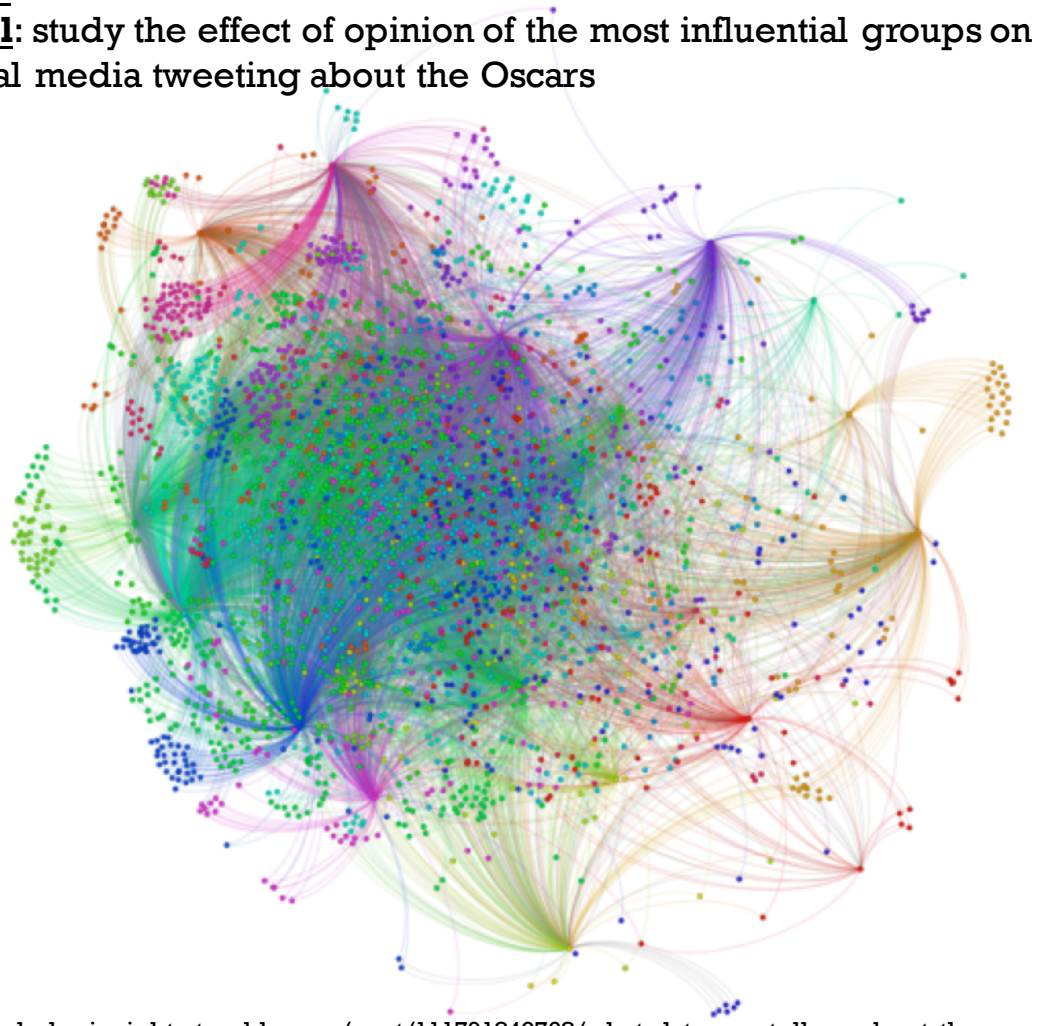


Answer: **Dave Chappelle**

http://blog.ranker.com/a-ranker-opinion-graph-of-the-domains-of-the-world-of-comedy/#.WD9mePkrLIU

Oscar Award Race - Hollywood Celebrity Network

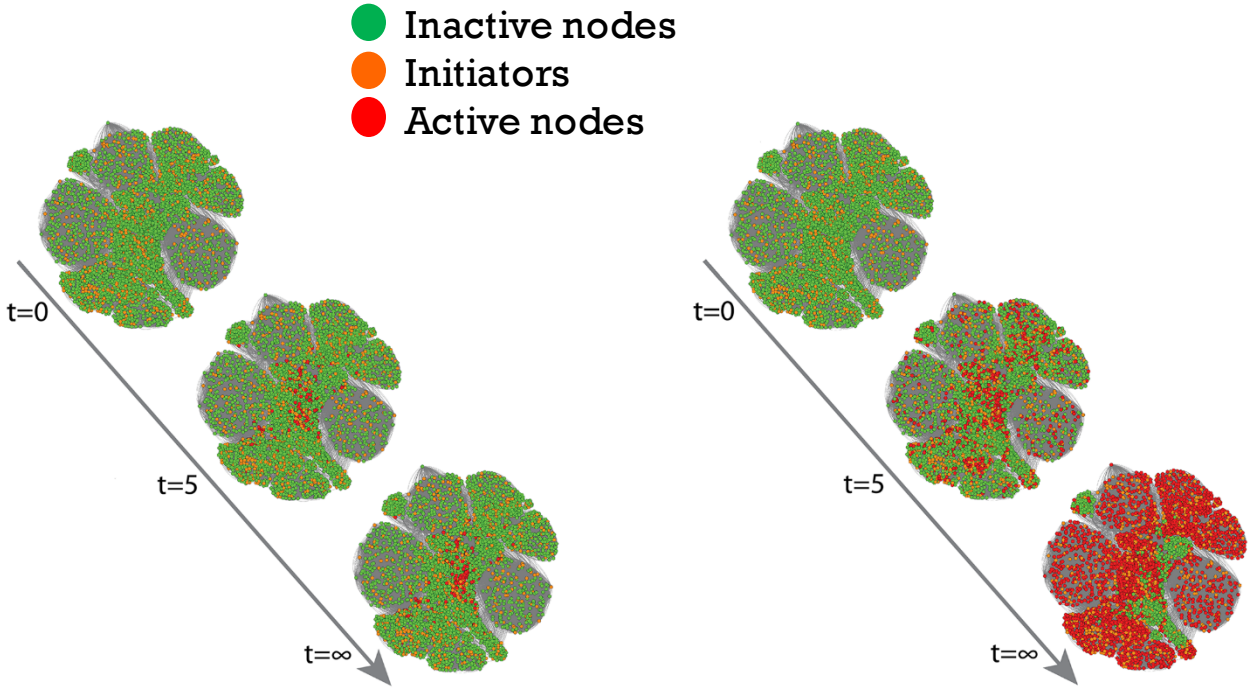
Data: Twitter
Goal: study the effect of opinion of the most influential groups on social media tweeting about the Oscars



http://echeloninsights.tumblr.com/post/111791840768/what-data-can-tell-us-about-the-oscar-race-and

OPINION INFLUENCING AND POLITICS IN SOCIAL NETWORKS

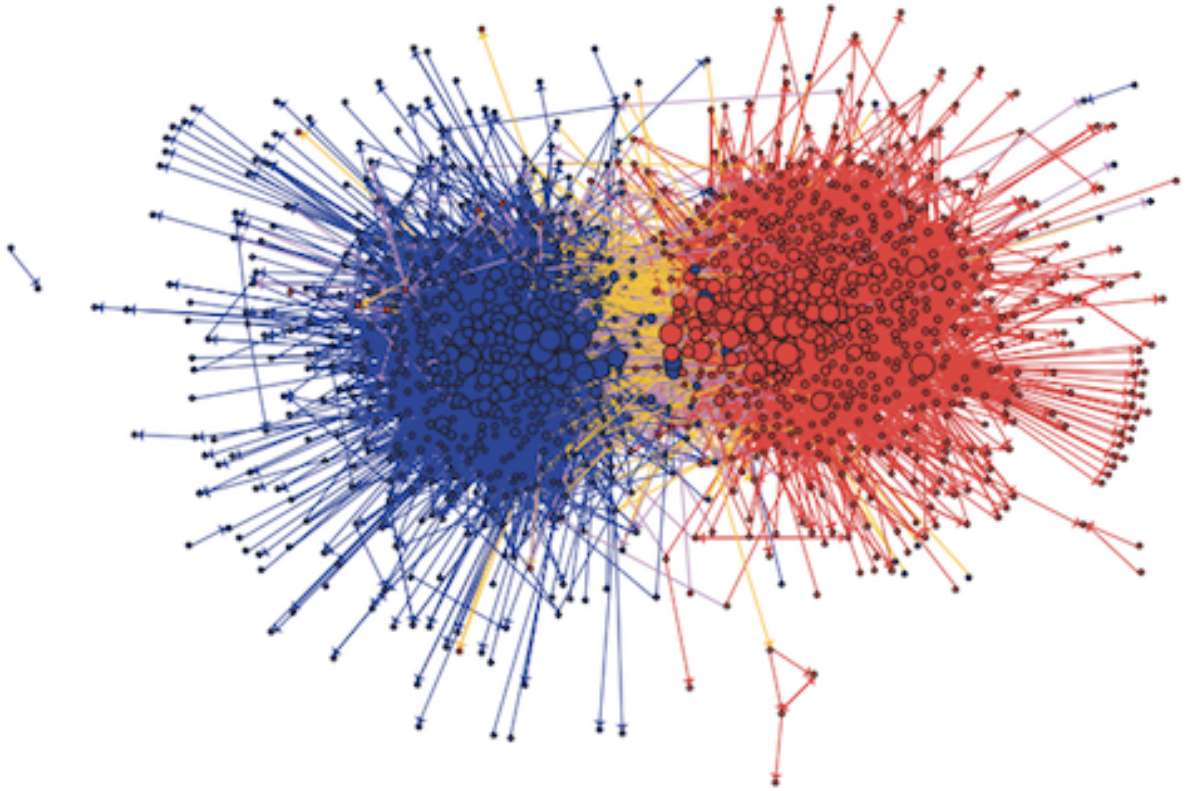
Data: Facebook
Goal: study the spread of opinions on social networks



Can be used for product marketing, politics, etc.

The Impact of Heterogeneous Thresholds on Social Contagion with Multiple Initiators
P. D. Karampourniotis, S. Sreenivasan, B. K. Szymanski, G. Korniss
PLoS ONE 10(11): e0143020 (2015)

Online communication between left-wing (blue) and right-wing (red) political blogs



"They are almost entirely divided into two separate networks: an echo chamber of like-minded individuals."

Image by Lada Adamic & Natalie Glance

THEORETICAL RESEARCH ON NETWORKS

A few of the multitude of topics studied on networks…

- Network controlling, monitoring, influencing
- Network resilience against random failures or attacks
- Network dynamics
- Information flow on networks
- Opinion formation and influencing
- Cascading failures in networks
- Clustering analysis, community detection
- Multilayer networks