Ai+ TRAINING

Exploring the Interconnected World:

Network/Graph Analysis in Python

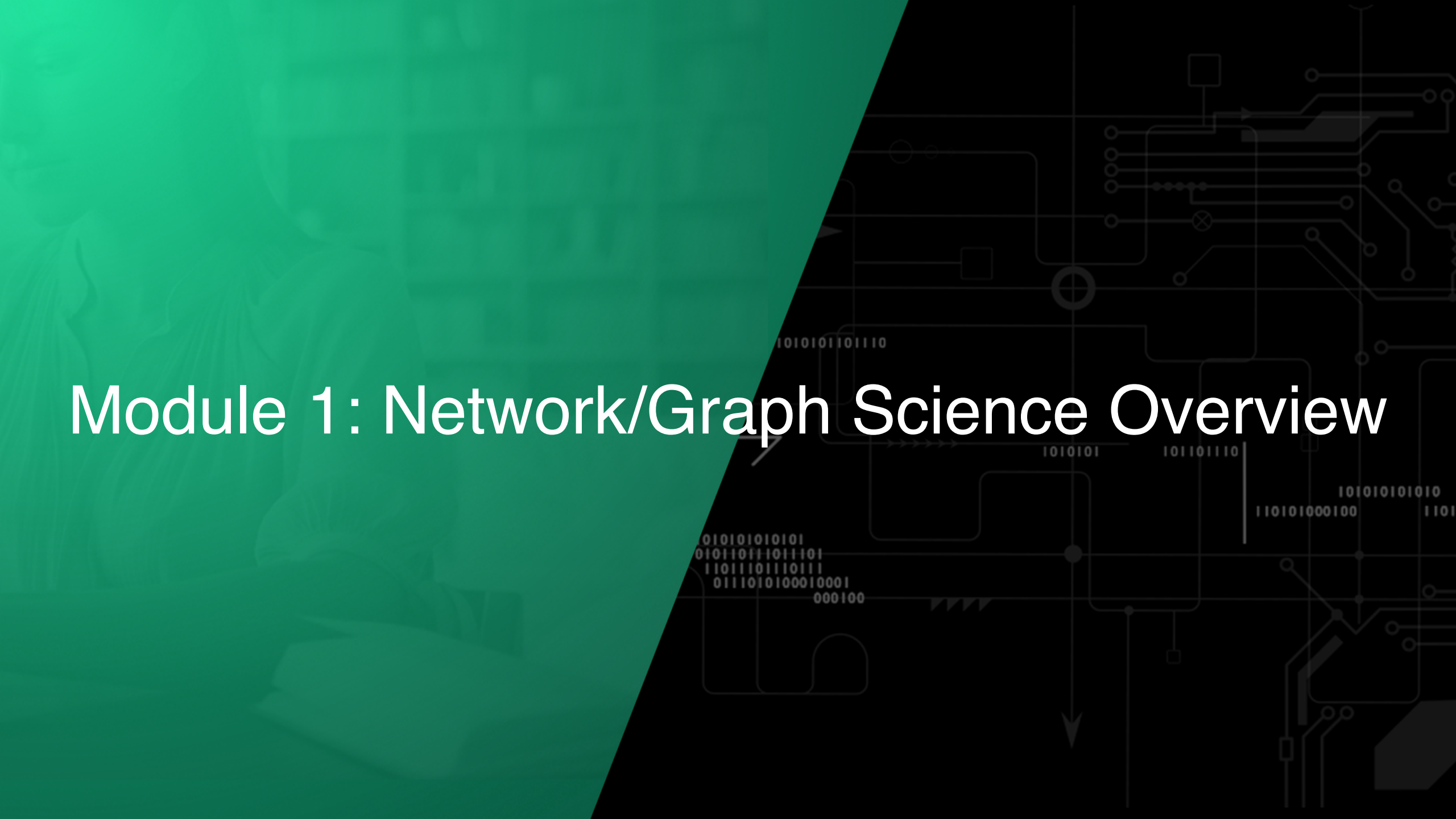
Noemi Derzsy

Data Scientist

@NoemiDerzsy

http://www.noemiderzsy.com/

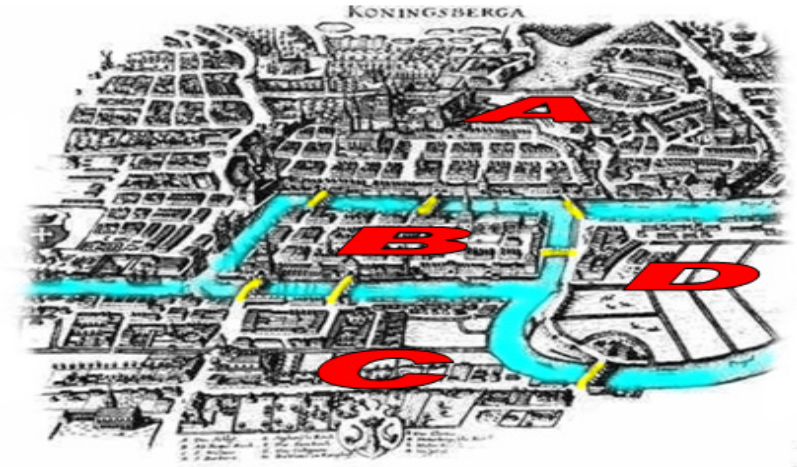
Module 1: Network/Graph Science Overview

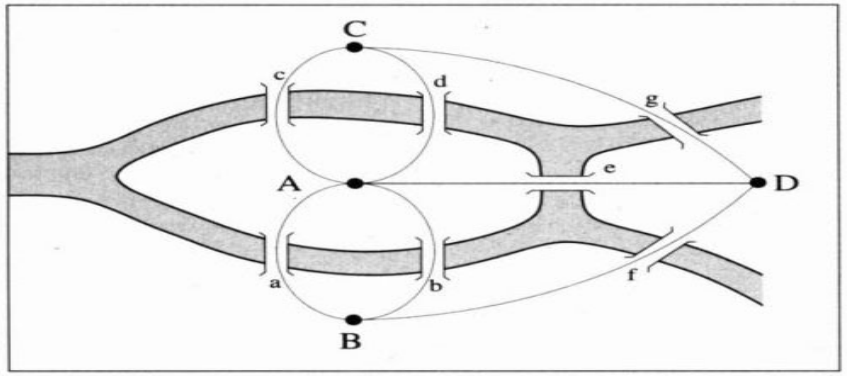


Graph Theory

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

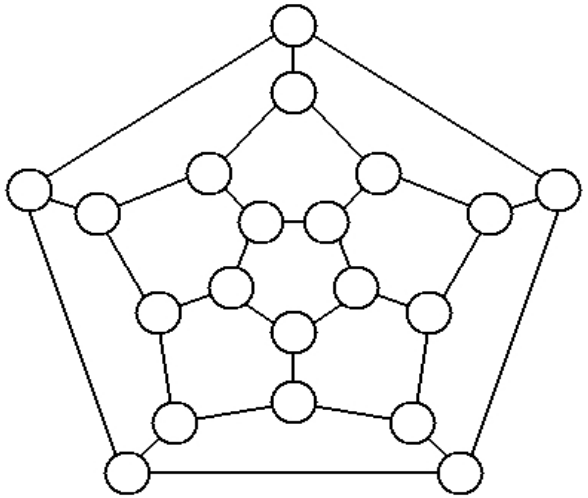
→ Depends on the nodes' degree (number of edges one node has) → **NO solution!**

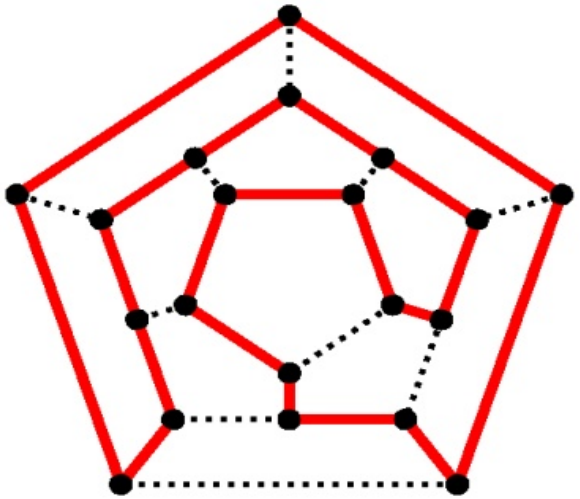
The theorem in the Königsberg bridges problem is the first graph theorem . The most important information: the number of edges and the nodes that are connected. This observation founded the development of mathematical topology.

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: 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).

Network Science = Applied Graph Theory

- A graph is an abstract representation of a set of objects connected by links

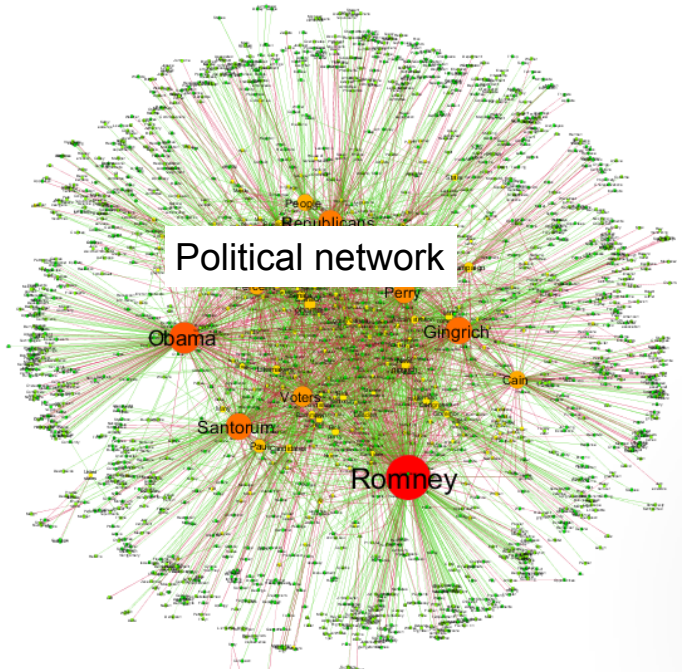
$$G = \{N, E\}$$

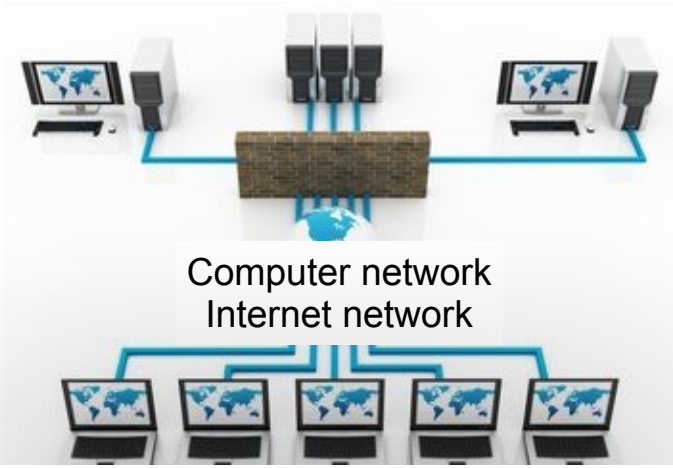
- Complex systems where elements/nodes/vertices can be connected by edges/links based on a certain relationship
- Complex systems can be studied through their underlying network structure: nodes (elements), links (interaction)
- The way we assign the nodes and links, defines the problem and questions we can study

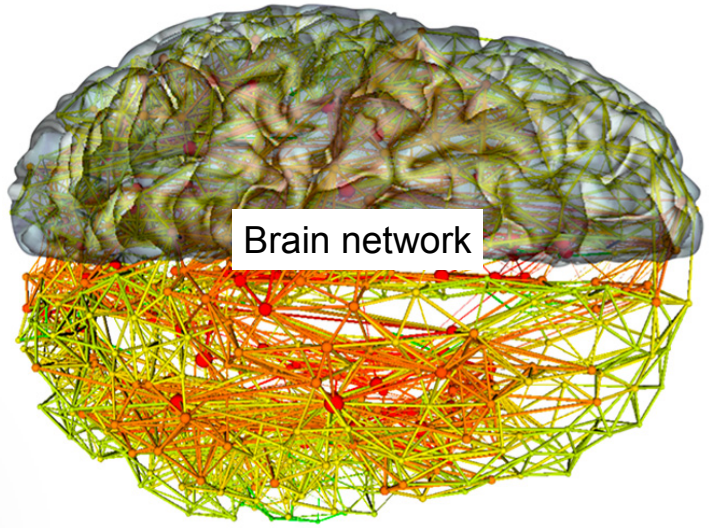
---

Mark Newman, *Networks: An Introduction*
Albert-Laszlo Barabasi, *Network Science*, http://networksciencebook.com
F. Menczer, S. Fortunato, C. A. Davis, *A First Course in Network Science*, https://cambridgeuniversitypress.github.io/FirstCourseNetworkScience

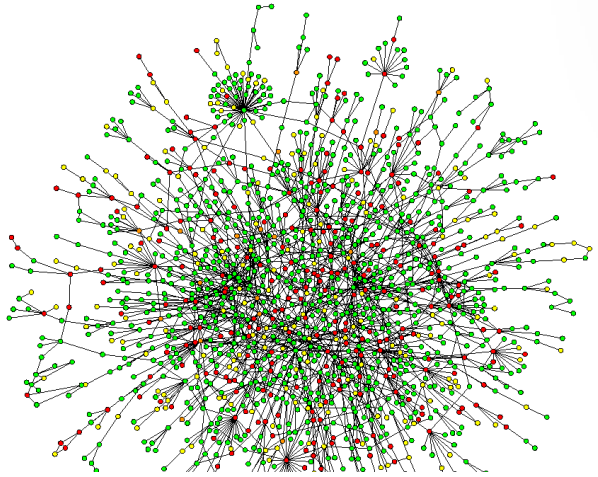
Complex Networks

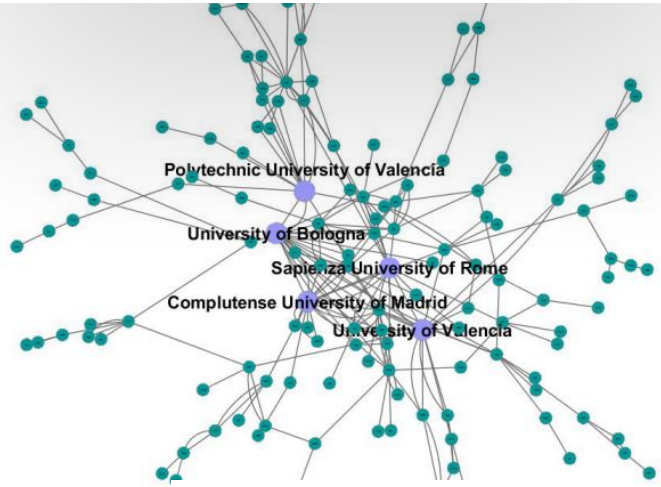


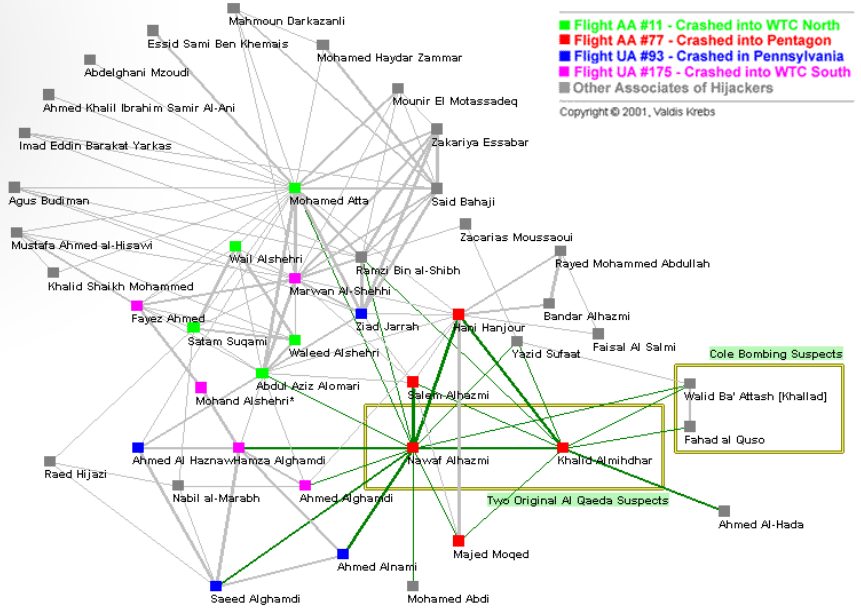




Only a few examples!







9/11 Terrorist Group

Figure 3 - All Nodes within 2 steps / degrees of original suspects

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…

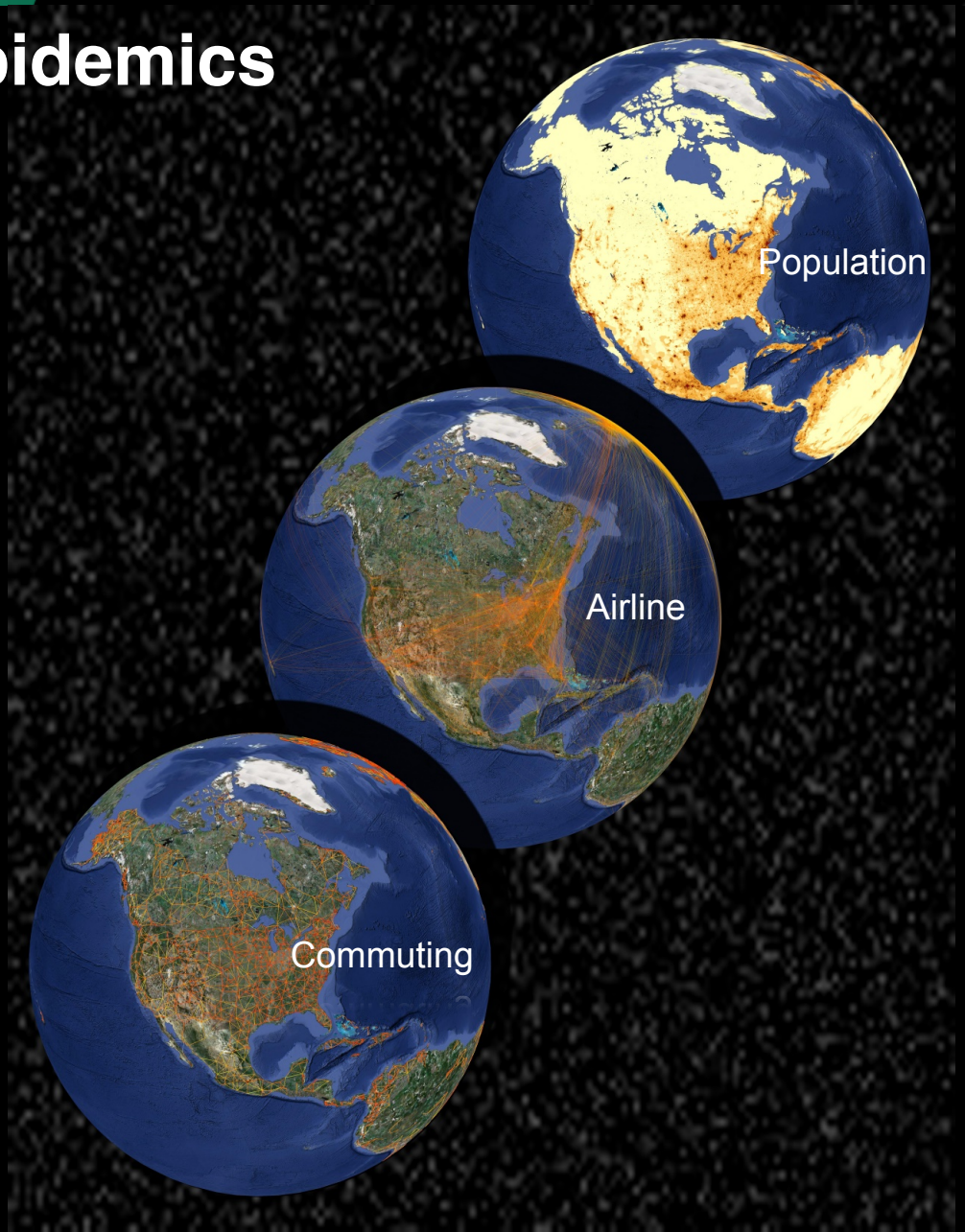
Disease Outbreak, Epidemics

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.

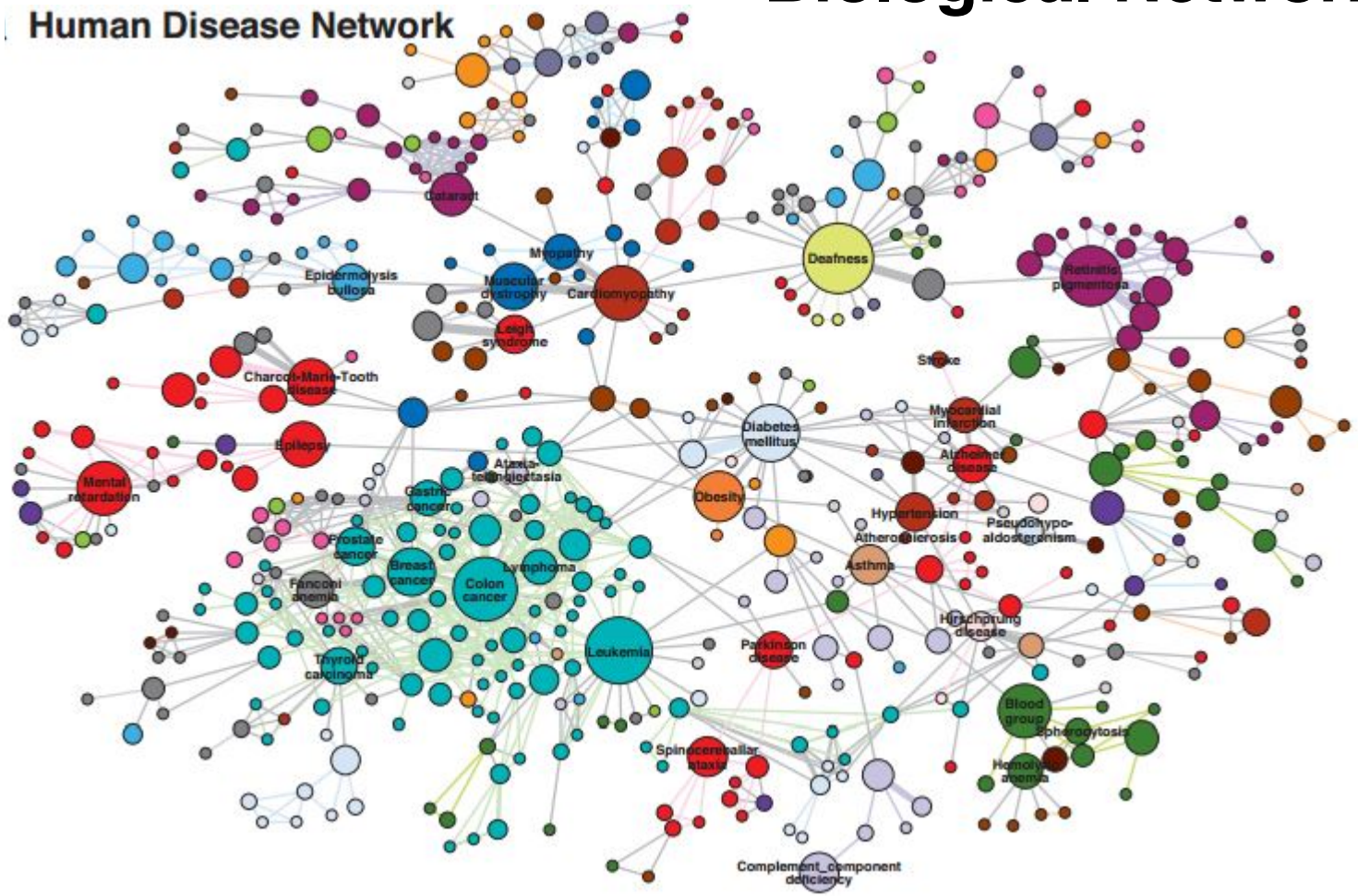
Real-time forecasting of epidemic spreading at the global scale

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

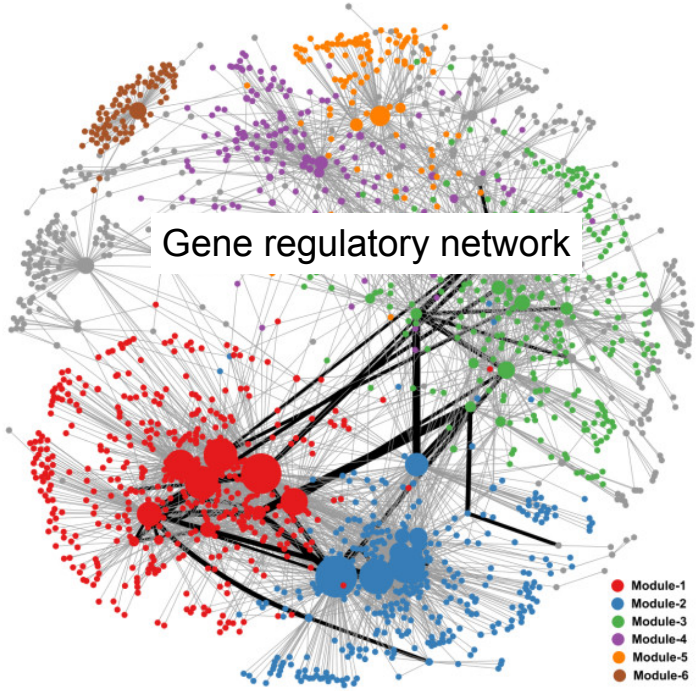


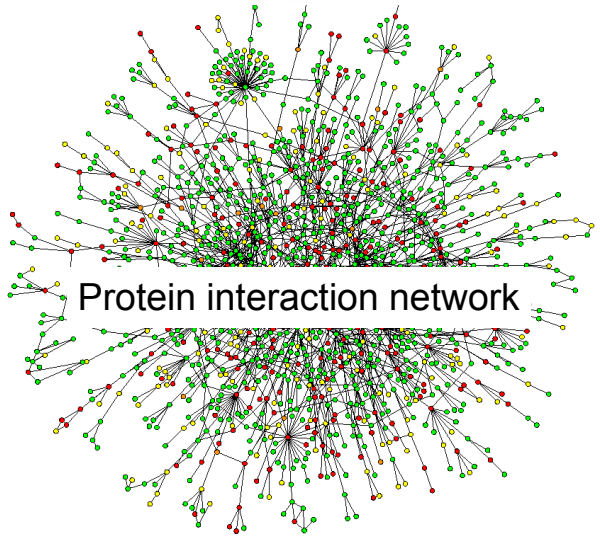
by Bruno Gonçalves and Alessandro Vespignani

Biological Networks



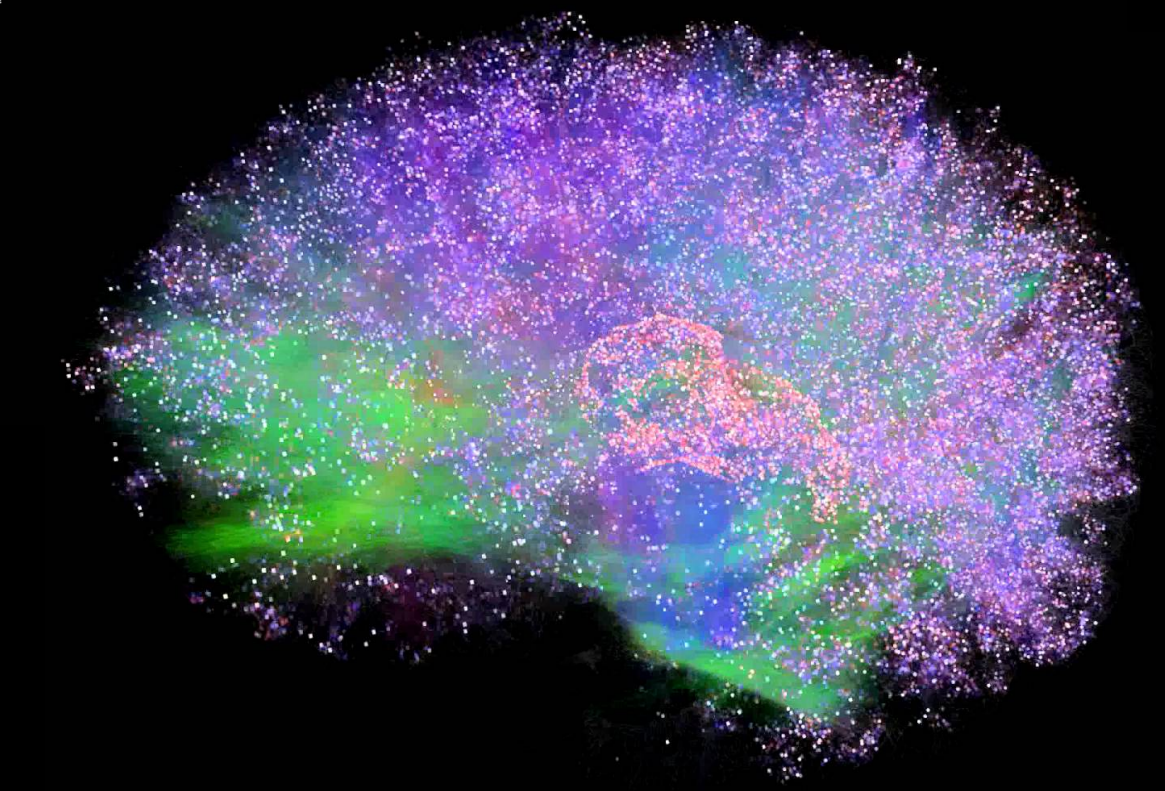
The human disease network
Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási
PNAS 2007





Brain Networks

Data: brain DW-MRI
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

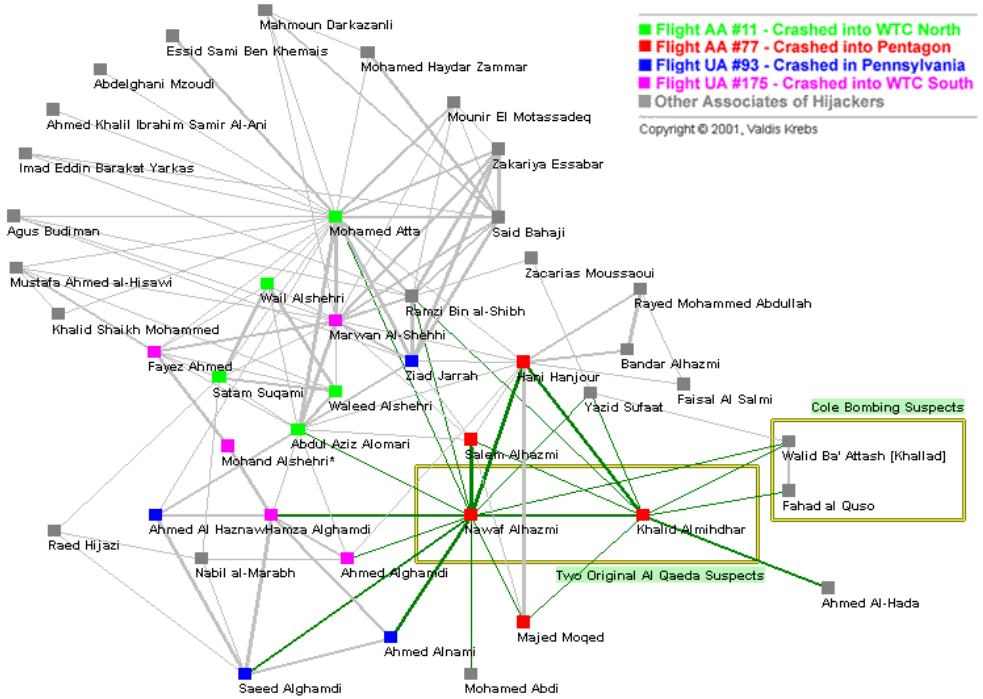
Power blackout in New York City (2012)

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

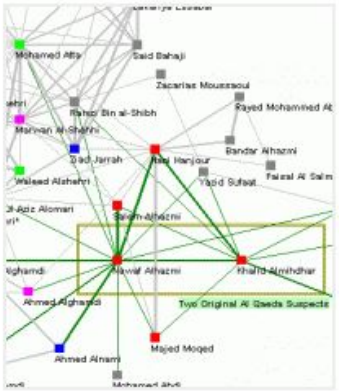
DigitalTonto
At the Crossroads of Media, Marketing and Technology...
Home

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).

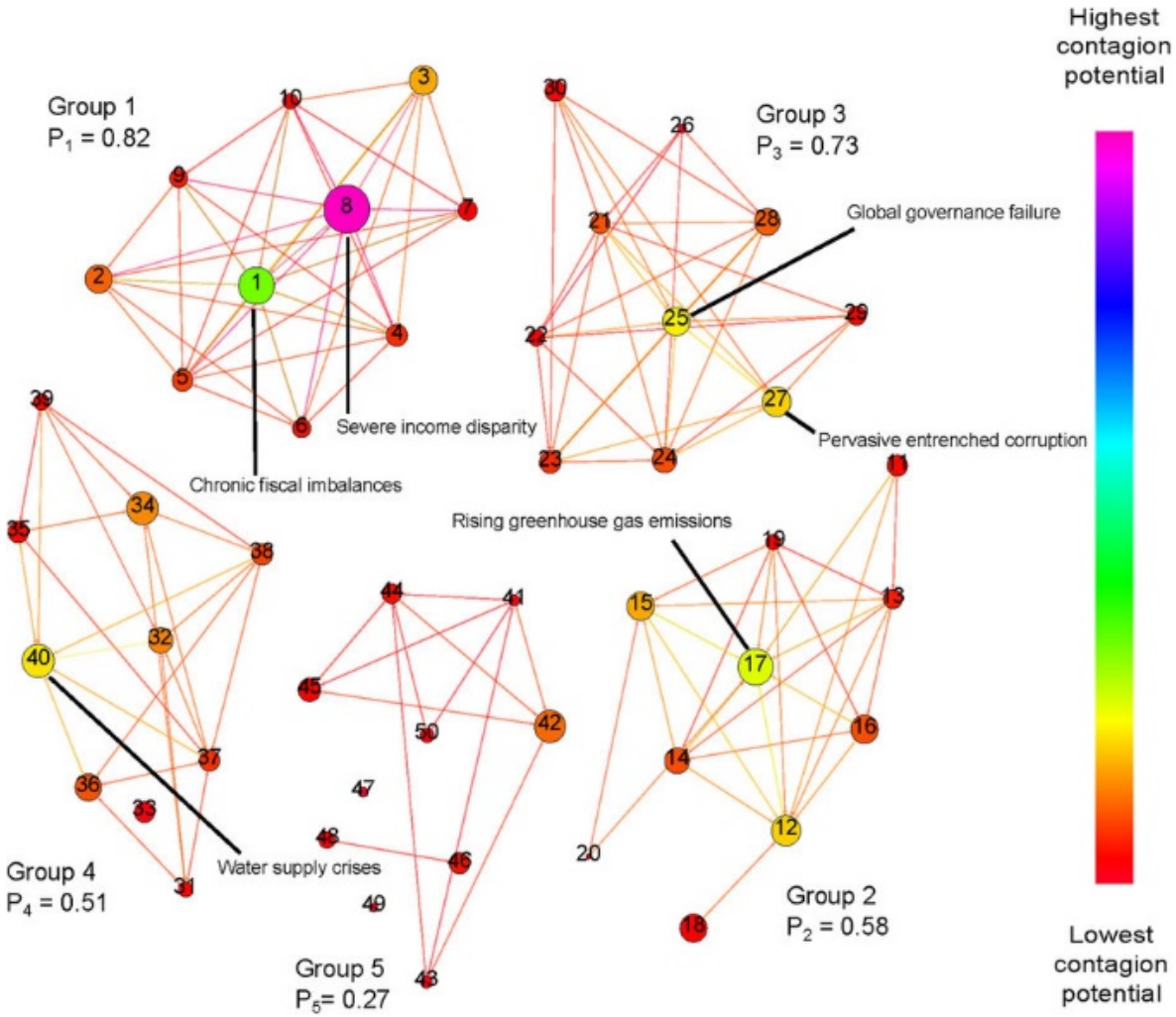
http://www.digitaltonto.com/2013/how-the-nsa-uses-social-network-analysis-to-map-terrorist-networks/

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



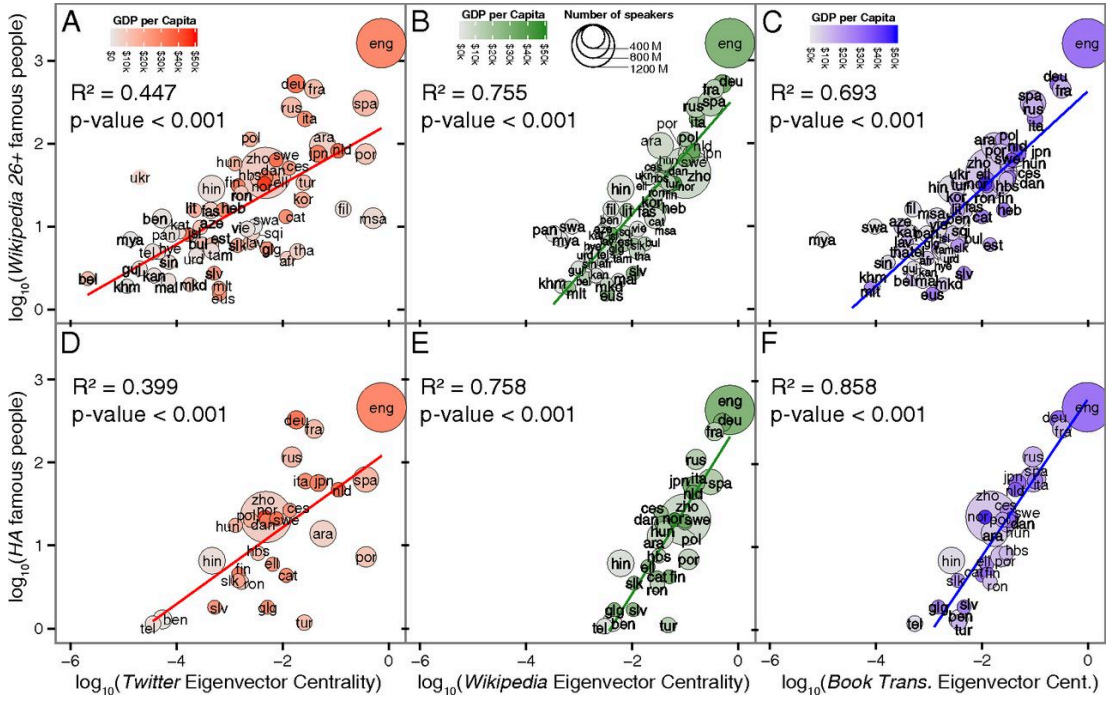
Each node is sized proportionally to its internal failure probability while node color corresponds to its total contagion potential. The number of edges in each group shows the intra-group connectivity. The nodes with the highest congestion potential are identified by name.

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

Global Language Network

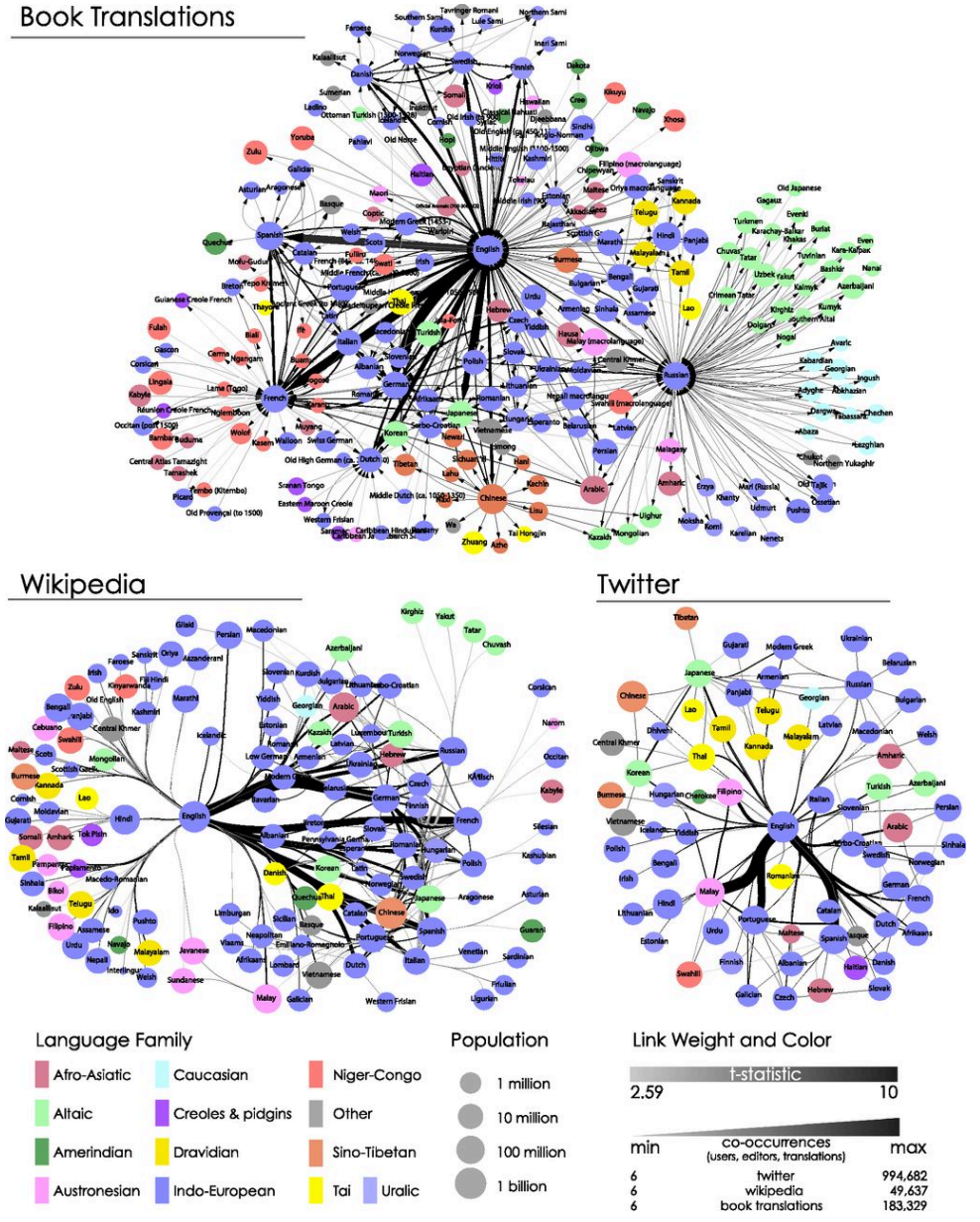
Data: Twitter, Wikipedia, UNESCO book translations
Goal: study the importance of languages and their impact

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



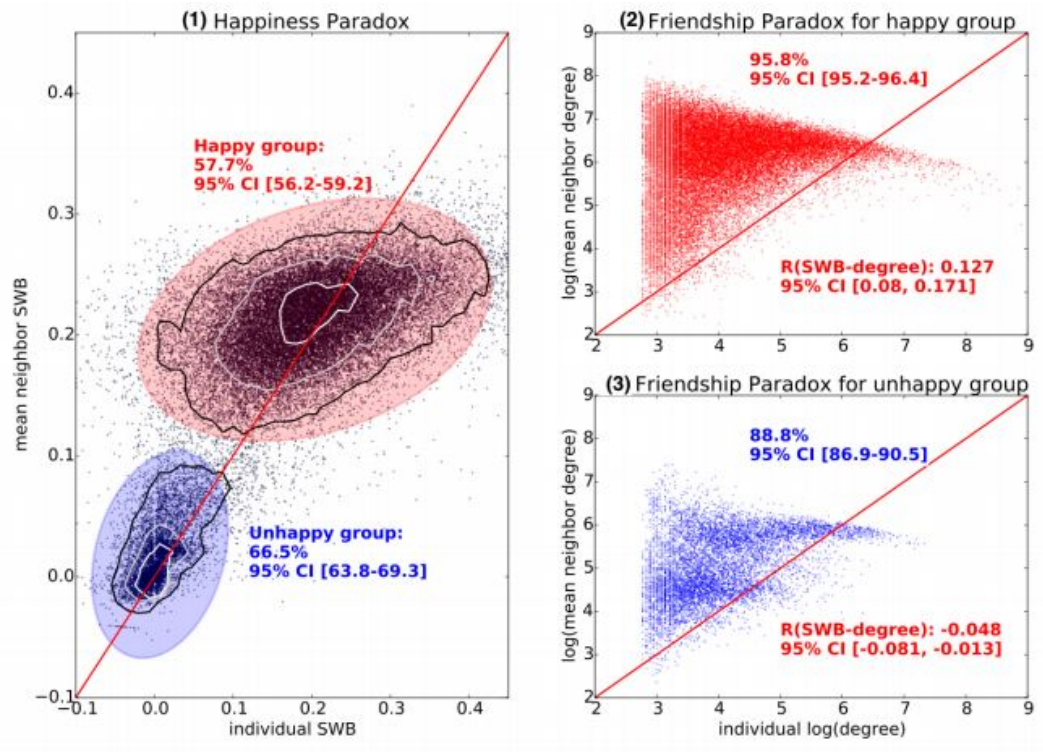
The position of a language in the GLN contributes to the visibility of its speakers and the global popularity of the cultural content they produce.

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

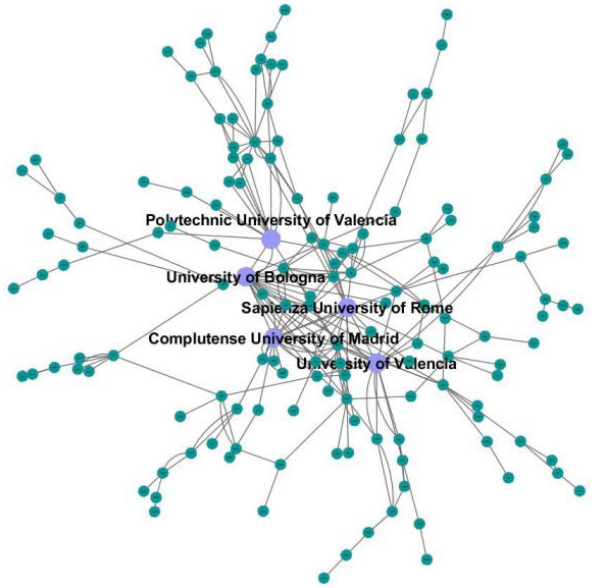
Data: Twitter
Goal: study the Happiness Paradox (most individuals in social networks experience a so-called Friendship Paradox: they are less popular than their friends on average)



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

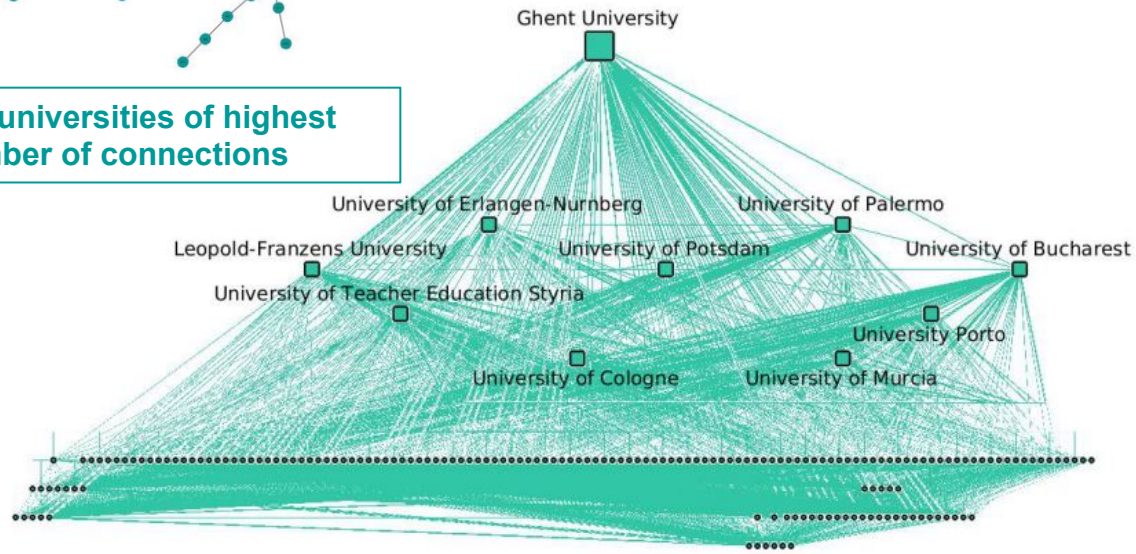


Data: European Erasmus scholarships

Goal: study the academic collaborations among universities through student mobility

The 5 most densely connected universities in Europe

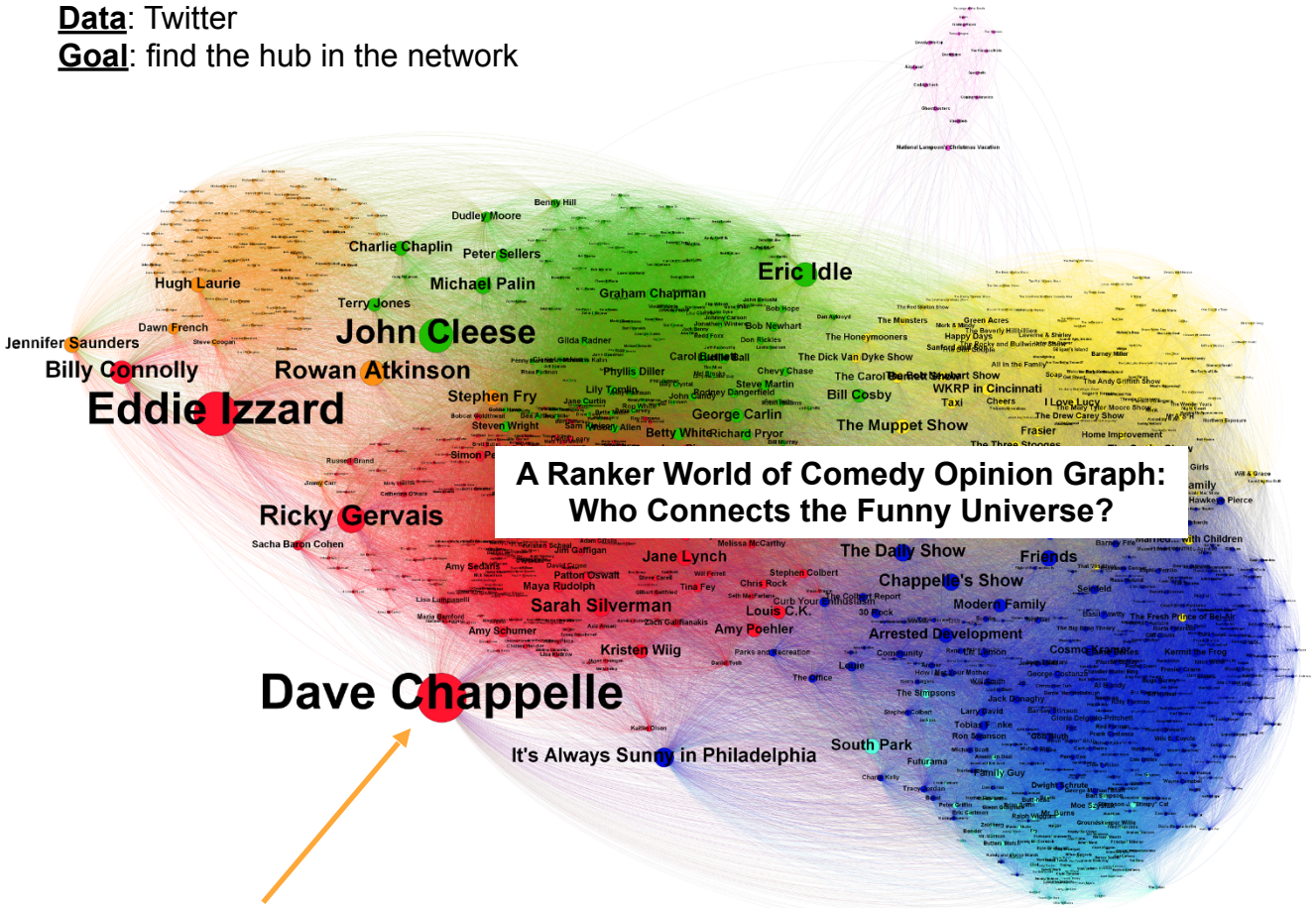
Top 10 universities of highest number of connections



Topology of the Erasmus student mobility network
A. Derzsi, N. Derzsy, E. Kaptalan and Z. Neda. Physica A. 2011;390(13):2601–261

Other Social Networks

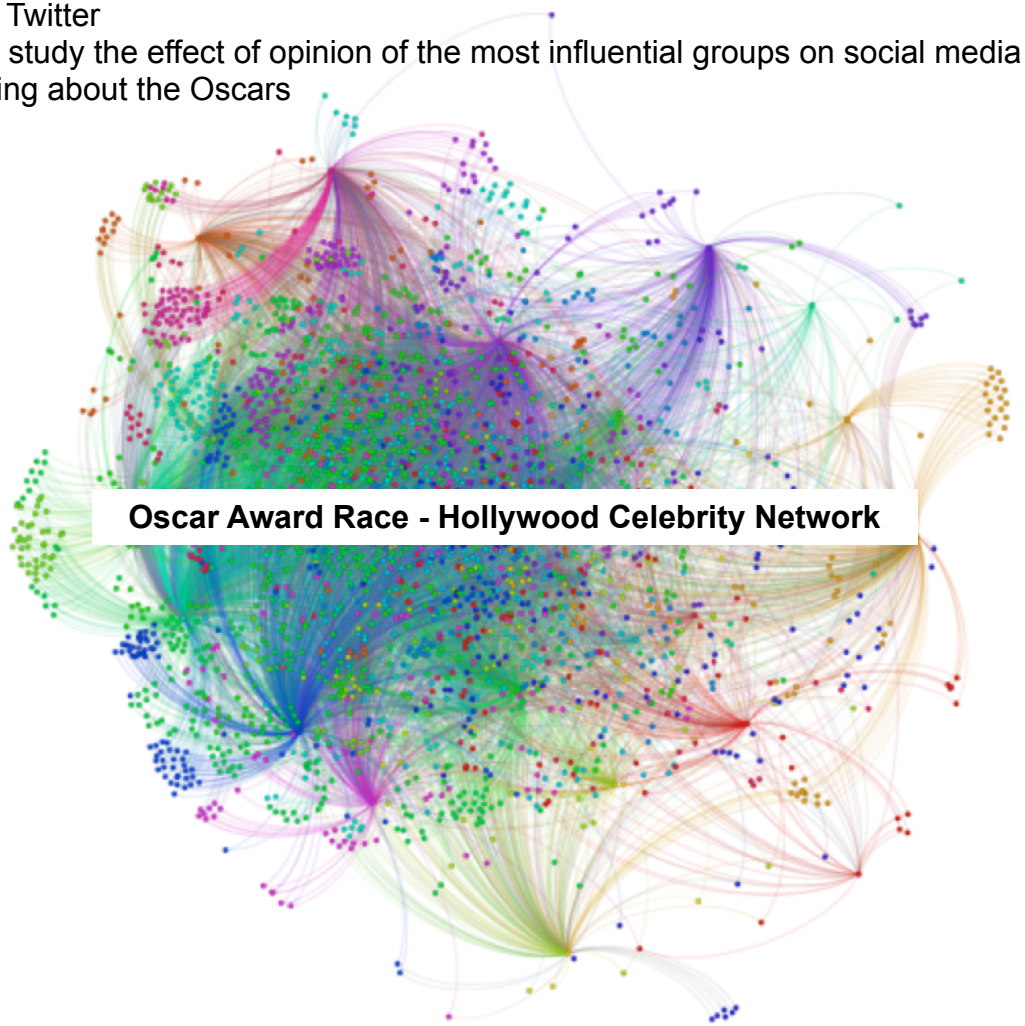
Data: Twitter
Goal: find the hub in the network



Answer: **Dave Chappelle**

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

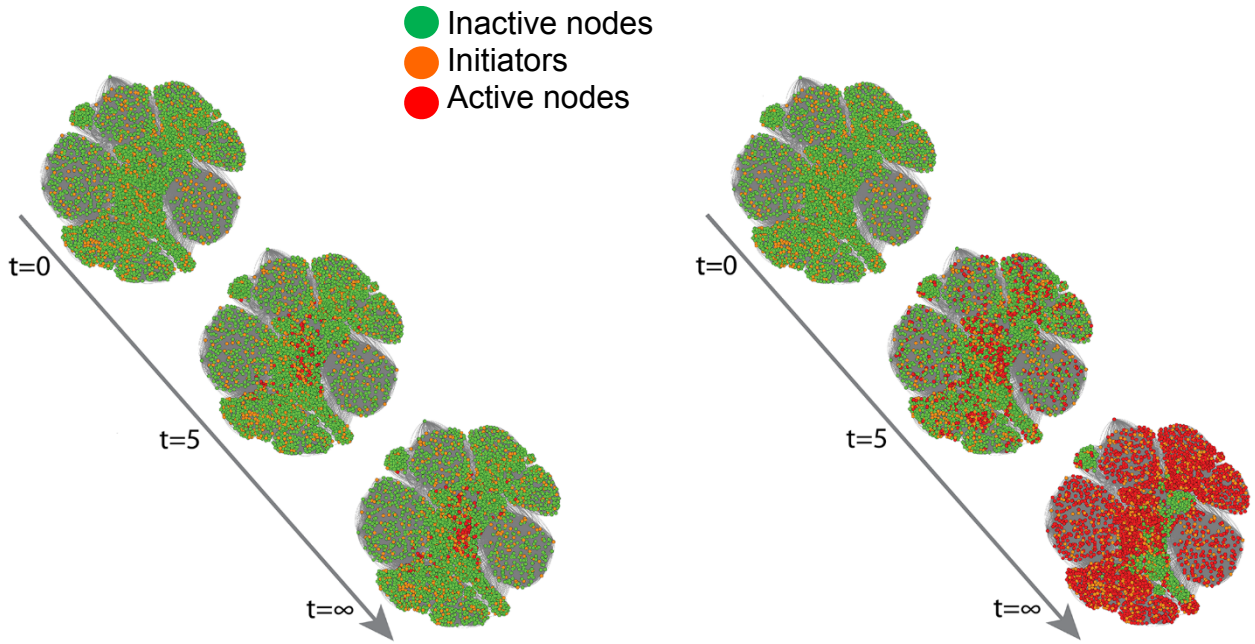
Data: Twitter
Goal: study the effect of opinion of the most influential groups on social media tweeting about the Oscars



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

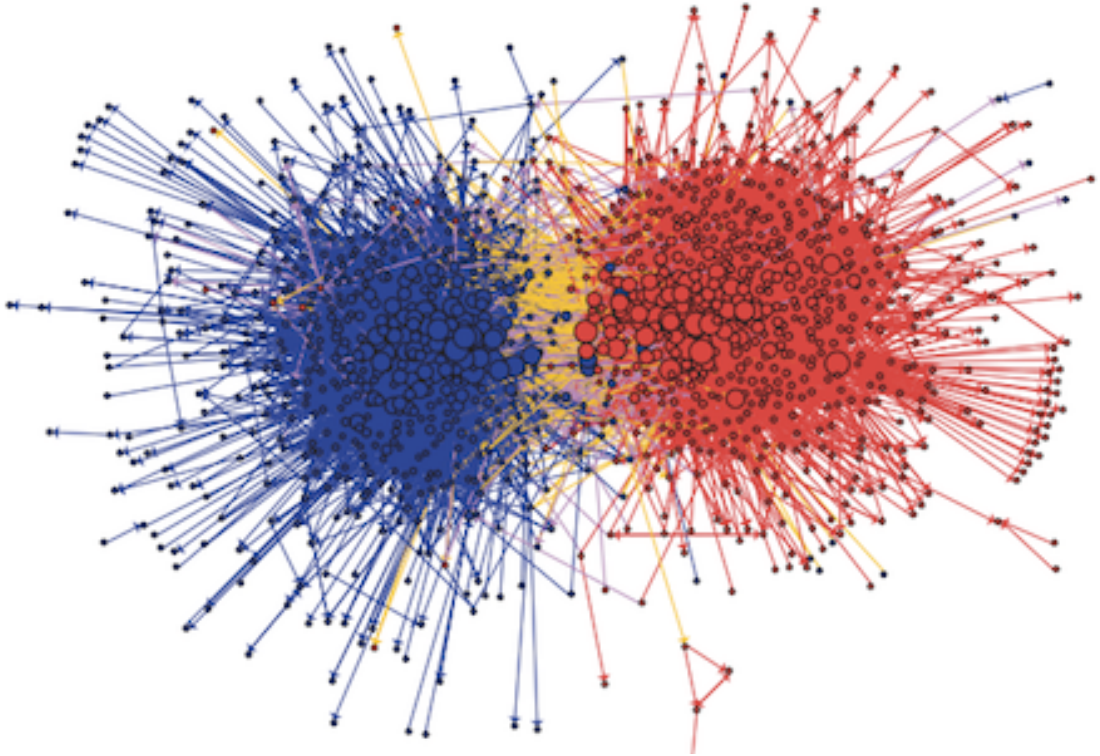
Opinion Influencing and Politics in Social Networks

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



Can be used for product marketing, politics, etc.

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

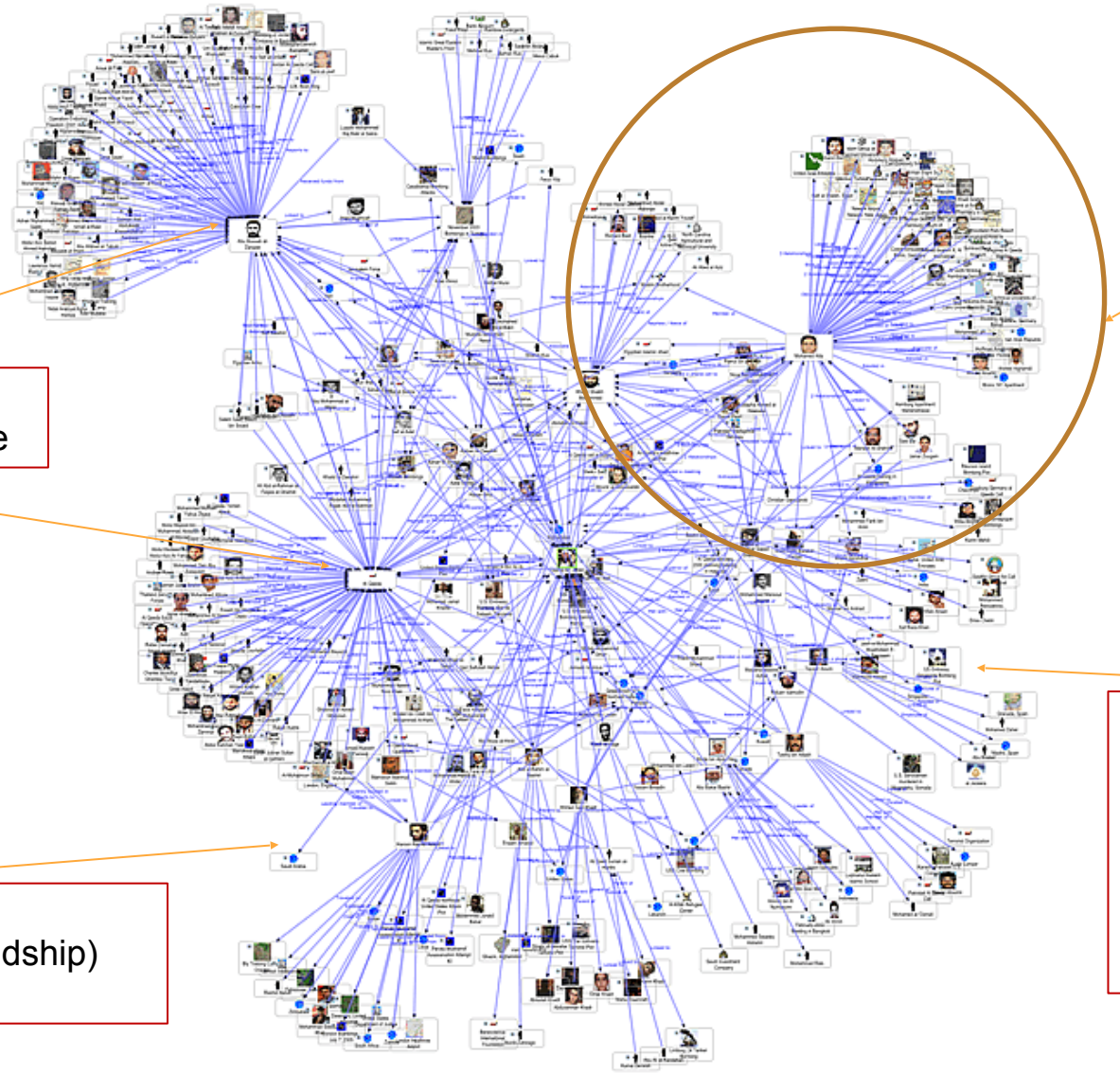


"They are almost entirely divided into two separate networks: an echo chamber of like-minded individuals."

Image by Lada Adamic & Natalie Glance

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)

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)

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

Graphs in Python with NetworkX

"NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks."

Nodes: can be any hashable object (text string, image, XML object, another graph, etc.)

Edges: can contain arbitrary data (weight, name, relationship type, color, etc.)

Graph object: a collection of nodes (vertices) along with identified pairs of nodes (called edges, links, etc.); uses dictionaries to store node and edge information.

Not suitable for large-scale network analysis and visualization!

https://networkx.org/documentation/stable/tutorial.html

GitHub link to training content:

https://github.com/nderzsy/AI-Network-Graph-Analysis-in-Python

Module 3: Structural Properties

Basic Structural Properties

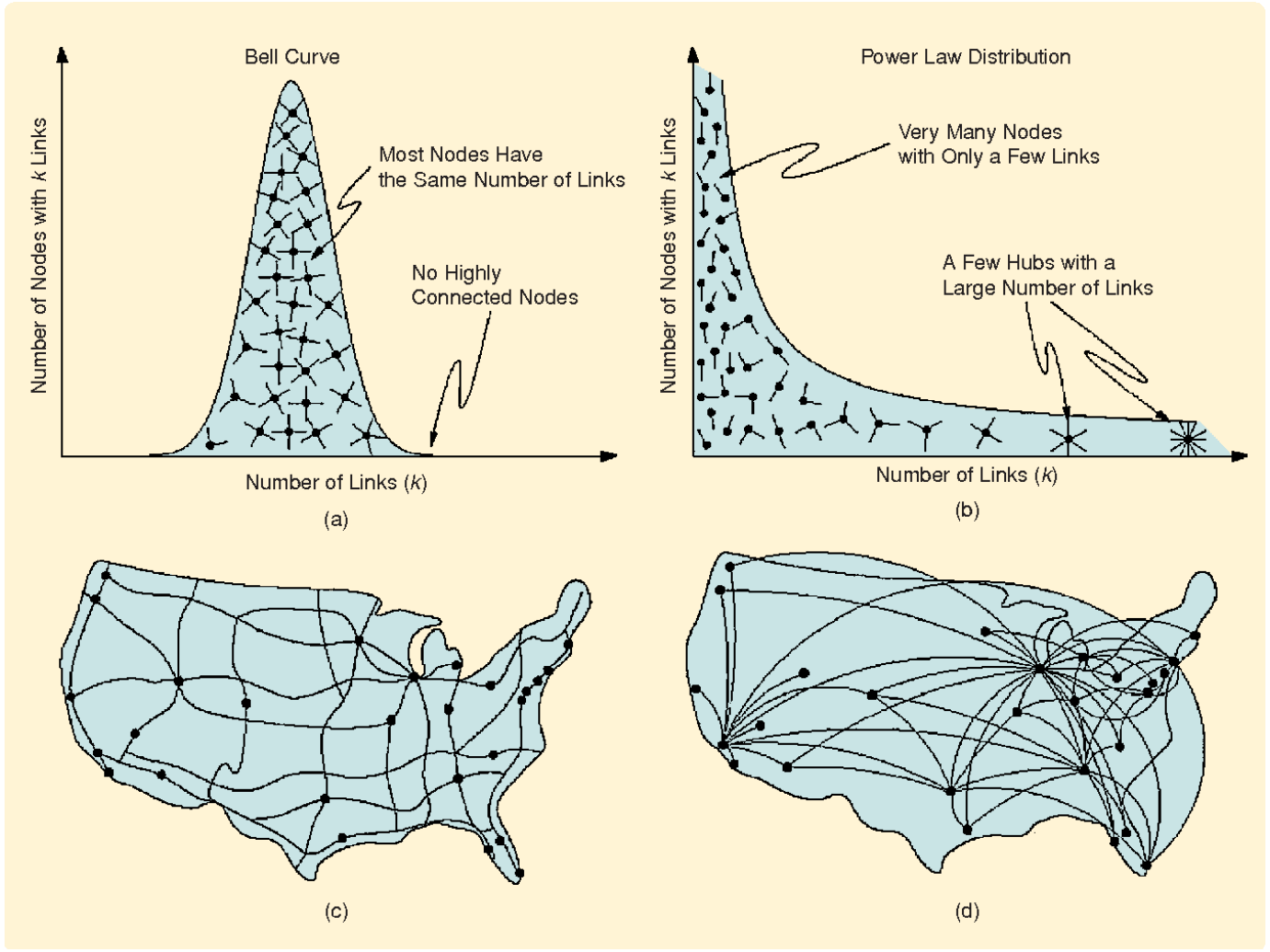
- Degree distribution P(k): the probability that a randomly selected node in the network has *k* connections
- Clustering coefficient: the fraction of a node's pairs of neighbors that are connected to each other
- Component: a subgroup of nodes not connected to the rest of the graph
- Average path length: 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/edge relying on its position, the level of network contribution
- Assortativity: Pearson correlation coefficient

Degree Distribution

P(*k*) - the probability that a randomly selected node in the network has *k* connections

Common types:

- **Poisson:** nodes have similar number of links (in random networks)
- **Power-law**: many nodes with few links, few hubs with vert large number of connections (in scale-free networks)
- **Exponential**: random growth model; no presence of hubs

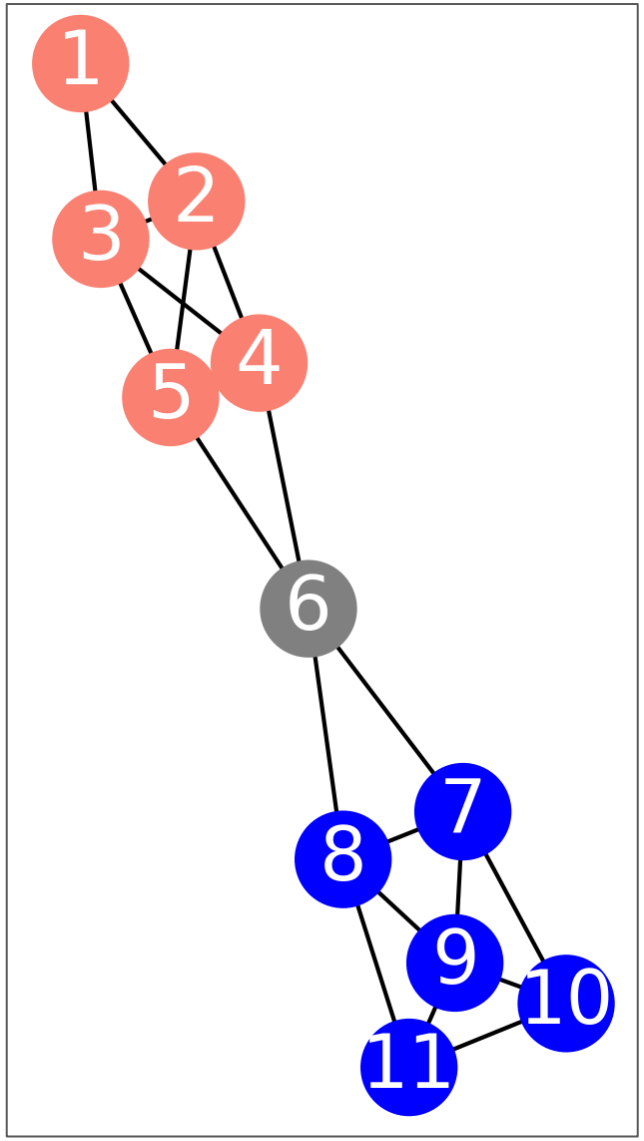


A. Barabasi, *The Architecture of Complexity From network structure to human dynamics,* IEEE Control Systems Magazine, 27(4):33–42 (2007)

Clustering Coefficient

- Density of links among the neighbors of a node
- Values [0, 1]:
  - 0 – there is no edge between any pair of neighboring nodes
  - 1 – all pairs of neighbor nodes are connected to each other

- **Triangles**: 3 nodes all connected to each other (6-7-8)
- **Triples**: 3 nodes, where one is connected to both (5-6-8)



Component

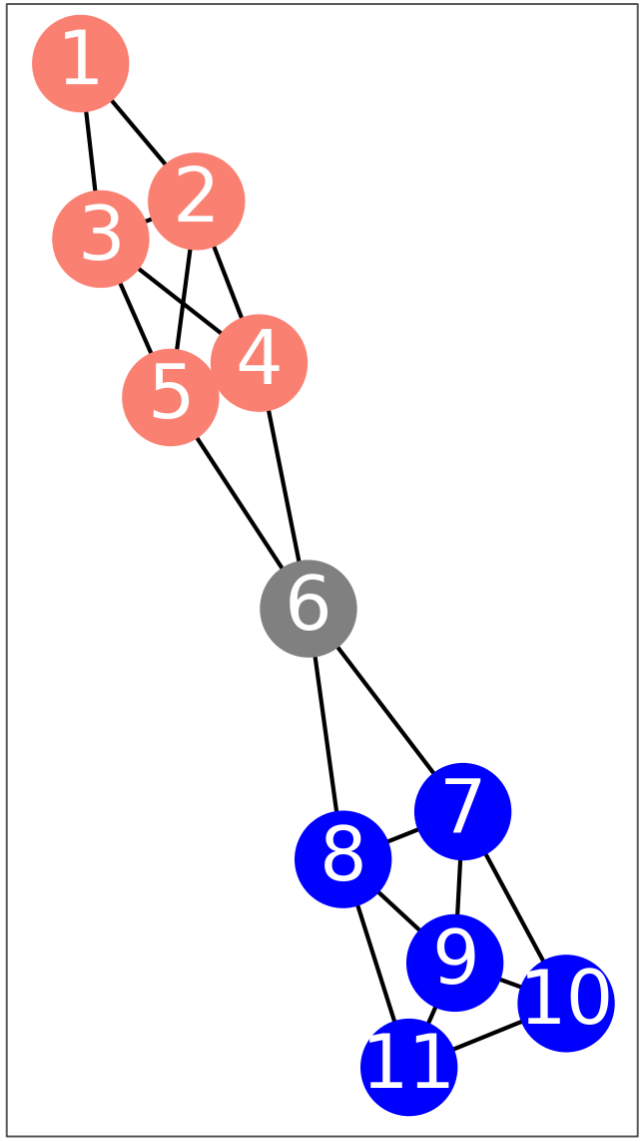
- **Component**: a subgroup of nodes not connected to the rest of the graph
- **Giant component**: the component containing the most nodes in the network
- **Connected graph:** each node is connected to at least another node in the network, has no disconnected components
- **Complete graph**: each node is connected to every other node in the network

Paths

- **Path**: sequence of nodes that needs to be traversed to reach from node *A* to node *C*
- **Path length***: #* of edges traversed along the path (not nodes!)
- Edges can be traversed more than once, but each time they are counted; self-avoiding paths
- **Shortest path (geodesic path):** the shortest existing path between two nodes
  - Length: **geodesic distance**
  - Type: **self-avoiding**; self-intersecting paths are never geodesic
  - Can have multiple shortest paths of equal length
- **Average path length**: the average of the smallest distance between two randomly selected nodes
- **Diameter**: the maximal distance between two elements in a graph; the length of the longest geodesic path between any pair of nodes for which a path exists

Centrality

- The importance of a node/edge relying on its position, the level of network contribution
- Centrality measures can refer to either node centrality or link centrality
- **Degree centrality:** the simplest centrality measure; the degree of a node
- **Eigenvector centrality:** node gets a centrality score proportional to the sum of its neighbors' centrality
- **Closeness centrality:** how close a node is to all other nodes; inverse of the sum of the distances to all other nodes
- **Betweenness centrality:** how many shortest paths go through the node



Assortativity

- Quantifies how interconnected nodes are to other similar nodes – homophily and assortative mixing
- Pearson correlation coefficient
- values [-1, 1]:
  - < 0: the network is **disassortative**; technological and biological systems
  - 0: no assortative mixing, no correlation
  - > 0: the network is **assortative**; social networks

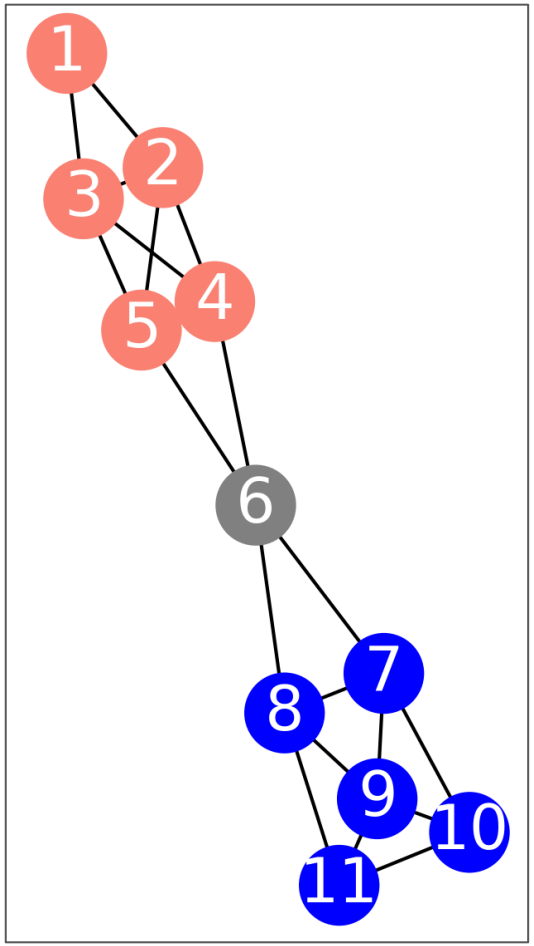
Module 5: Community Detection

Community Detection

- **Community, cluster, module** – subset of nodes with higher connectivity density
- Communities are a **structural property of a network**
- **Special version of the general data clustering**, which is grouping of elements into clusters based on similarities
- Each node can belong to **only one community**
- Community detection is a **stochastic process**
- All nodes of a community must be **reachable through other nodes within the community**
- Nodes within a community have **higher probability of connecting to other nodes within same community**

Community Definitions

- **Internal/external links**: edges within or outside the community
- **Internal/external degree**: a node's degree inside/outside the community
- **Internal link density**: the ratio between the # of internal links and the max. # of links that could exist between any two communities
- **Strong community**: each node has more neighbors inside the community than in the rest of the network
- **Weak community**: sum(internal degrees) > sum(external degrees) -> relaxes the strong community requirement, considers community as a whole, instead of individual
- **Clique**: a complete subgraph; every node within the community is connected to every other node within it
- **Bridge**: an edge that connects two communities



Community Detection vs. Partitioning

Graph partitioning: division of a network into a predefined number of smaller subgraphs (partitions)

Community detection: finds the inherent community structure of a network; the number and the size of the communities is not predefined

Hierarchical Clustering

- Data clustering algorithm
- **Similarity matrix**: a similarity measure as node property or distance between each pair of nodes -> nodes that connect to each other and share neighbors likely belong to the same community
- Iteratively identifies groups of nodes with high similarity
    - **Agglomerative**: iterative merging of groups of nodes with high similarity links; selects node pairs that belong to the same community
    - **Divisive**: iterative partitioning by splitting groups of nodes; removes low similarity links; selects node pairs that are in different communities

Modularity Optimization

- **Modularity:** quantifies the quality of the partition based on a random baseline
- # of links within all communities > expected # of links if the graph were randomized
- For random graphs clusters are not present, modularity score very low
- Higher modularity implies better community separation
- **Maximal modularity**: optimal community structure (partition)

Louvain modularity optimization algorithm

The most common and widely used; it scales better for large real-world networks

Step 1.

- assign each node to a different community
- or each node evaluate the gain in modularity if we place it in the community of one of its neighbors
- move node in the community for which the modularity gain is the largest and positive If no positive gain is found, *i* stays in its original community.
- repeat until no more improvement can be made

Step 2.

- construct a new network, where the nodes are the communities built from the above steps
- The weight of the link between two nodes is the sum of the weight of the links between the nodes in the corresponding communities
- Links between nodes of the same community lead to weighted self-loops.

Step 3.

- repeat Steps 1 & 2 (pass) until maximum modularity is achieved

Greedy modularity maximization algorithm (Newman)

Iteratively join pairs of communities if it increases the partition's modularity:

- starting with N communities of single nodes, assign each node to a community of its own
- for each community pair connected by at least one link, compute the modularity difference ΔM obtained if we merge them
- identify the community pair for which ΔM is the largest and merge them
- repeat until all nodes merge into a single community, recording M for each step
- select the partition for which M is maximal

Bridge Removal (Girvan-Newman)

- **Link betweenness**: quantifies how many shortest paths run through a link -> high when link is a ***bridge***

Method:

- Iteratively finds and removes bridges with the highest link betweenness
- Betweenness is recalculated at every step on the remaining links -> makes the process slow

- All partitions are hierarchical

- Very slow -> not practical for large scale networks

- Use alternate bridge identifying measures

Label Propagation

- Neighbors usually belong to same community
- Simple and scales well for large networks

Method:

- Each node is assigned to a different community (*N* nodes, *N* communities)
- Each node is visited in a random order and assigned to the community where its majority of neighbors are
- Repeat until all nodes belong to the majority label of their neighbors
- Each node will have more neighbors in its community

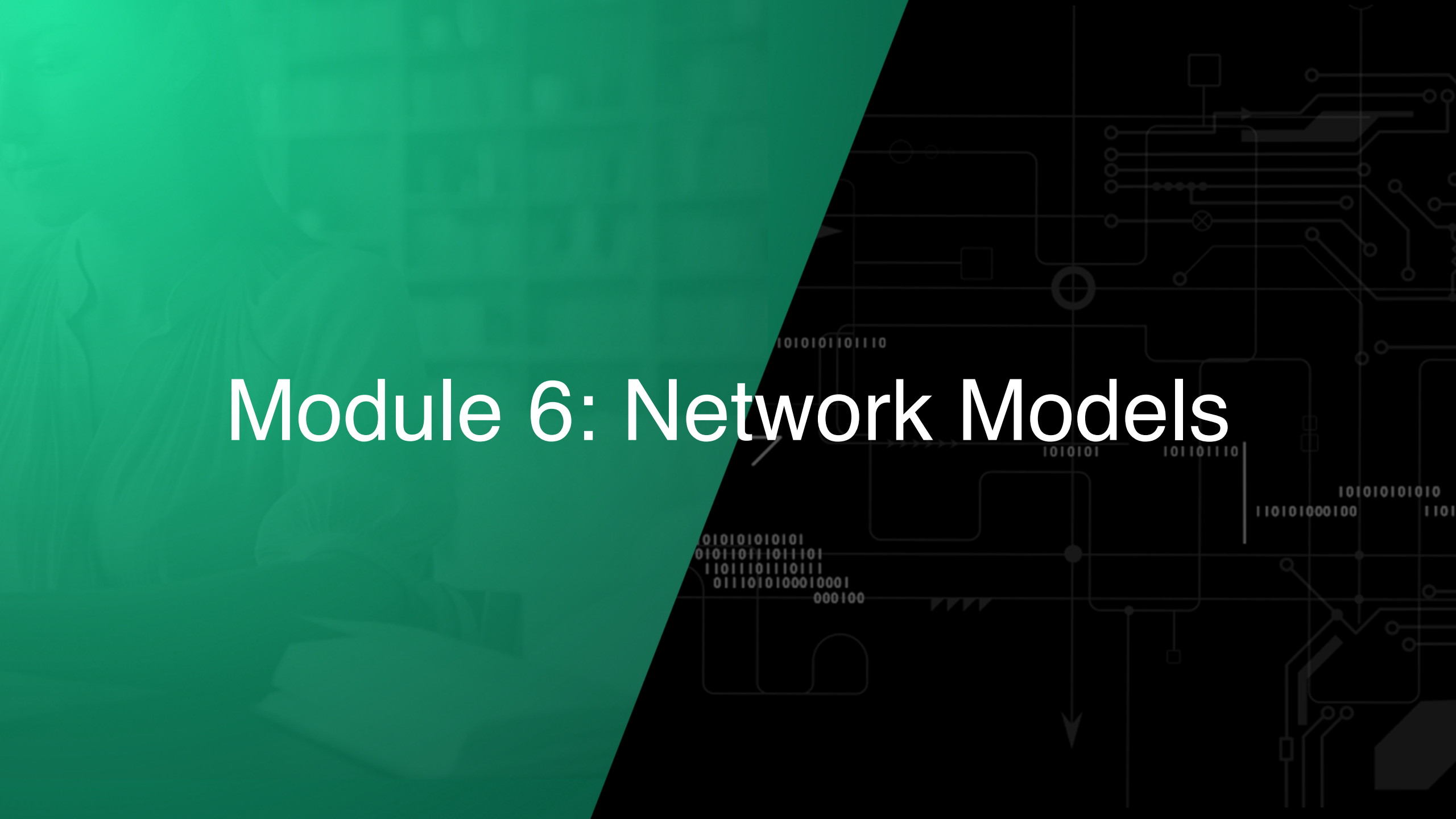
- During this process many communities labels disappear
- Multiple steps are needed to reach a stationary state

Community Detection Best Practices

How accurate is the community detection algorithm?

- Usually no ground truth
- **Benchmark graph**:
  - Real (ex. Zachary Karate Club)
  - Artificial: stochastic block models
- **Partition similarity**: how similar the outcome of an algorithm is to the natural partition
  - Fraction of correctly detected nodes
- The number of all possible partitions is impossible to retrieve due to its tremendous size

Module 6: Network Models

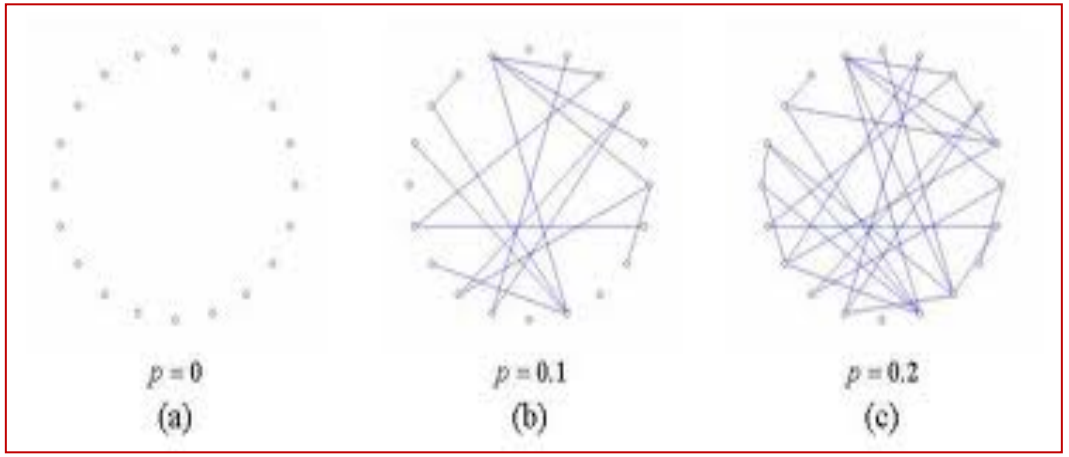


What are Network Models used for?

- Simulate behaviors on networks with specific structure
- Not enough real world data for large scale study
- Serve as baseline models

Random (Erdos-Renyi) Networks

Model proposed by P. Erdős and A. Rényi



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

Problems of the model

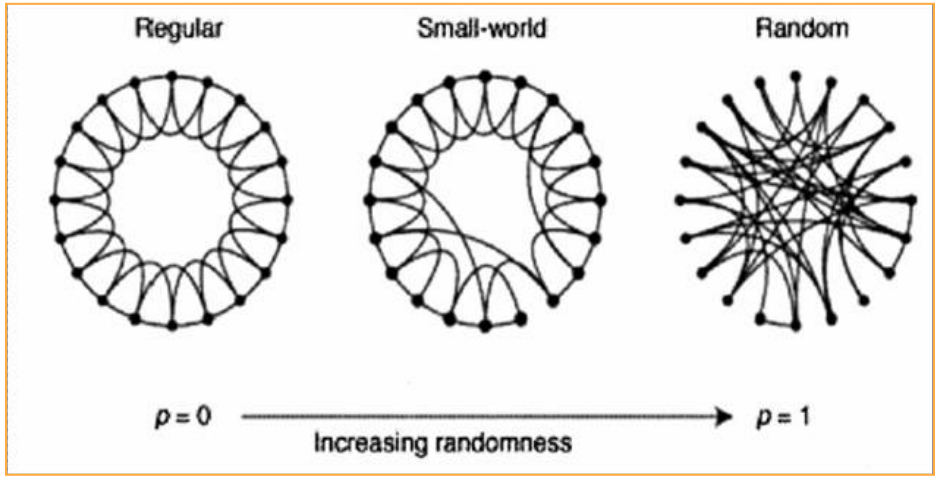
- ❑ Studies showed that real complex systems do not have Poisson distribution!
- ❑ Several clusters were observed in real complex systems that are not reproduced by this model
- ❑ Presence of hubs in real complex systems

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 tail
- High clustering coefficient
- Small average path length

Problems solved

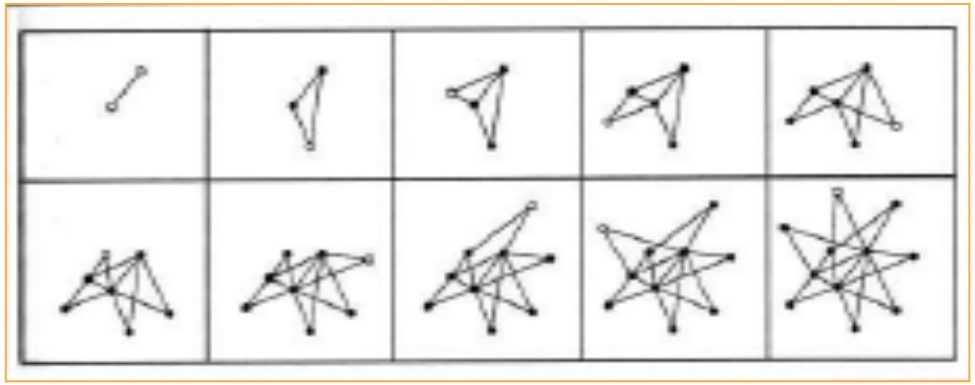
- ✓ Studies showed that real complex systems: do not have Poisson distribution!
- ✓ Several clusters were observed in real complex systems that are not reproduced by this model
- ❑ Presence of hubs in real complex systems

→ **This 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: power-law
- High clustering coefficient
- Small average path length

Problems solved

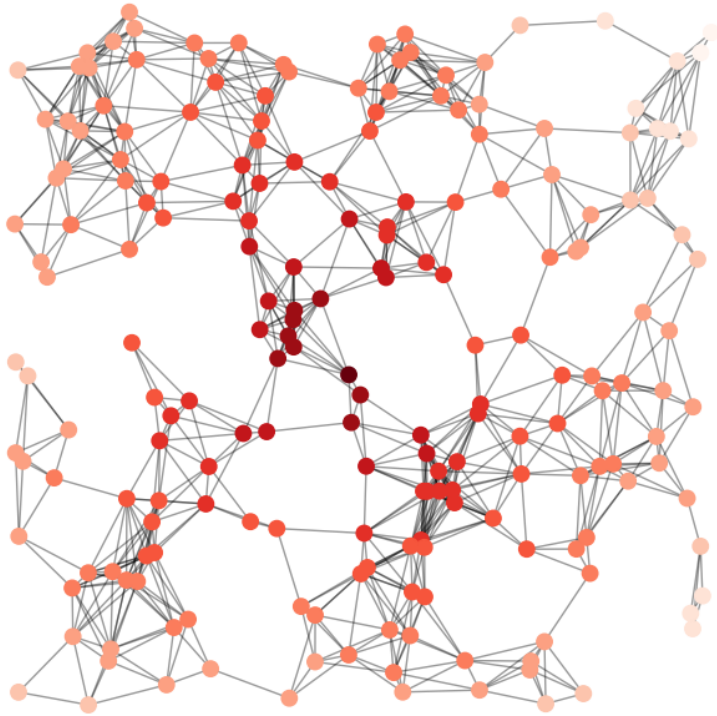
- ✓ Studies showed that real complex systems do not have Poisson distribution!
- ✓ Several clusters were observed in real complex systems that are not reproduced by this model
- ✓ Presence of hubs in real complex systems

→ **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.

Random Geometric Graphs (RGG)

- Spatial network
- Constructed by randomly assigning coordinates to nodes according to a specified probability distribution
- Pairs of nodes are connected by a link if and only if their distance is in a given range, i.e. < certain neighborhood radius *r*
- Clusters of nodes with high modularity
- Assortative: degree assortativity based on spatial metric



NetworkX Tutorial

Christensen, Michael; Dall, Jesper. Random Geometric Graphs. *Physical Review E*. (2002)

Configuration Model

- Builds a network with prescribed degree distribution; can reproduce real network structures

Method:

- Input: arbitrary **degree sequence** (n_1: 32, n_2: 12, …, n_n: 121); can be extracted from real networks or a desired distribution
- Sum of the degrees must be even
- Assign to each node **# of stubs** (half link, not yet connected) equal to their degree
- Iteratively:
  - Select randomly a pair of nodes with free stubs
  - Form a link between the two nodes
  - Repeat until all stubs are connected into links

- Application: to test a network property, whether it is explained by the degree distribution -> **degree-preserving randomization**

Molloy, Michael; Reed, Bruce. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*. **6** (2–3): 161–180. (1995)