**Social Network Analysis summary**

|andrea.ierardi@studenti.unimi.it | andreierardi@gmail.com

aNDREA IERARDI notes

**2020- 2021**

Table of content

[Introduction 4](#_Toc61563836)

[SOCIAL NETWORK ANALYSIS 4](#_Toc61563837)

[Network 4](#_Toc61563838)

[Affiliation Network 4](#_Toc61563839)

[NETWORK SCIENCE 5](#_Toc61563840)

[Density and Average Degree 5](#_Toc61563841)

[Degree and average degree 5](#_Toc61563842)

[Density 5](#_Toc61563843)

[Degree and density for Directed networks 5](#_Toc61563844)

[Network Models 6](#_Toc61563845)

[Random Network 6](#_Toc61563846)

[Random Network Erdos-Renyi model 🡺 G(N,p) 6](#_Toc61563847)

[Scale free property 7](#_Toc61563848)

[Scale free networks 7](#_Toc61563849)

[Power law 7](#_Toc61563850)

[Hubs in Scale free Network vs. Random networks 7](#_Toc61563851)

[Barabasi-Albert model 7](#_Toc61563852)

[Connectivity 8](#_Toc61563853)

[Connected Components 8](#_Toc61563854)

[Path 8](#_Toc61563855)

[Connected Graph 8](#_Toc61563856)

[Connected Component 8](#_Toc61563857)

[Directed strongly connected graph 8](#_Toc61563858)

[Directed weakly connected graph 8](#_Toc61563859)

[Random-Real Network Connected Components 9](#_Toc61563860)

[Giant components 9](#_Toc61563861)

[Critical values of Giant components 9](#_Toc61563862)

[Centrality 10](#_Toc61563863)

[Degree Centrality Undirected networks 10](#_Toc61563864)

[Degree Centrality Undirected networks 10](#_Toc61563865)

[Geometric Centralities 10](#_Toc61563866)

[Degree Centrality 10](#_Toc61563867)

[Betweenness centrality 10](#_Toc61563868)

[Closeness centrality 10](#_Toc61563869)

[Spectral Ranking 11](#_Toc61563870)

[Eigenvector Centrality 11](#_Toc61563871)

[Katz Centrality 11](#_Toc61563872)

[Page Rank 11](#_Toc61563873)

[Transitivity 12](#_Toc61563874)

[Global Clustering 12](#_Toc61563875)

[Coefficient based on paths 12](#_Toc61563876)

[Coefficient based on triples 12](#_Toc61563877)

[Coefficient based on triangles 12](#_Toc61563878)

[Local Clustering 12](#_Toc61563879)

[Clustering in random network 12](#_Toc61563880)

[Ties and Bridging 13](#_Toc61563881)

[Bridge 13](#_Toc61563882)

[Local Bridge 13](#_Toc61563883)

[Almost Local Bridge 13](#_Toc61563884)

[Six degree of separation 14](#_Toc61563885)

[Stanley Milgram’s experiment 14](#_Toc61563886)

[Small world 14](#_Toc61563887)

[Watts-Strongatz model 14](#_Toc61563888)

[Balance and status 15](#_Toc61563889)

[Social Balance theory 15](#_Toc61563890)

[Social Status theory 15](#_Toc61563891)

[SImilarity 15](#_Toc61563892)

[Structural Equivalence 15](#_Toc61563893)

[Regular Equivalence 15](#_Toc61563894)

[Influence and Homophility Assorativity 16](#_Toc61563895)

[Measuring Assortativity 16](#_Toc61563896)

[Social Influence 17](#_Toc61563897)

[Crawler 18](#_Toc61563898)

[Download issues 18](#_Toc61563899)

[Policies 18](#_Toc61563900)

[Page Processing and storage 18](#_Toc61563901)

[Bloom filters 18](#_Toc61563902)

[Sieve 18](#_Toc61563903)

[Flush 18](#_Toc61563904)

[Communities 19](#_Toc61563905)

[Definition of communities 19](#_Toc61563906)

[Maximumm cliques 19](#_Toc61563907)

[Partition 19](#_Toc61563908)

[Modularity 19](#_Toc61563909)

[Greedy Algorithm 19](#_Toc61563910)

[Louvain Algorithm 19](#_Toc61563911)

[Overlapping Communities 19](#_Toc61563912)

[CFinder 19](#_Toc61563913)

[Clique Percolation Method 19](#_Toc61563914)

[Information Diffusion 20](#_Toc61563915)

[Definition 20](#_Toc61563916)

[Information diffusion types 20](#_Toc61563917)

[Herd Behaviour 20](#_Toc61563918)

[Information Cascade 20](#_Toc61563919)

[Independent Cascade Model (ICM) 20](#_Toc61563920)

# Introduction

Sociology: the differentiation among people and governing social interaction.

## SOCIAL NETWORK ANALYSIS

**Social network analysis:** focus on relationships among social entities, and on the pattern and implication of these relationship.

### Network

Behind each complex system there is a **network**, that define interactions between components. In fact, we use a network to map and understand a complex system.

Most network observed in nature and society are driven by common principles.

A network consists of a finite set of nodes (actors) and edges (relationship). In network science relation ties among actors are primary and attributes of actors are secondary.

**The choice of the proper network representation determines our ability to use network theory successfully.**

The mathematical representation of a network is a graph:

A graph is an ordered par comprising a set V of vertices with a set E of edges.

**Mathematical representation of a graph:**

* Edge list
* Adjacency list
* Adjacency matrix

### Affiliation Network

Representation by bipartite (two-mode):

* Actors and group
* There are no link between actors and groups

## NETWORK SCIENCE

[WIKI]

**Network science** is an academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks, considering distinct elements or actors represented by *nodes* (or *vertices*) and the connections between the elements or actors as *links* (or *edges*).

Two forced helped the emergence of network science:

* **Internet:** offering fast data sharing methods.
* **Powerful tools:** In the past we couldn’t manage such amount of data

## Density and Average Degree

### Degree and average degree

Degree of a node: number of links connected to it.

Average Degree: average number of links per node:

It is a **local property**

### Density

Density: Fraction of all possible links that are actually present: Ratio of the number of links to the number of possible links in a network with N nodes 🡺

It is a **global property**

### Degree and density for Directed networks

Real network are sparse! L << Lmax  or *<d>* << N-1

# Network Models

We need statistical, mathematical, and computational methods when measuring the structure of a network. Why?

* We can calculate analytically some properties of the network
* Random network model to compare to real-world network
* Allow us to identifying observed pattern to better understand the nature of the network.

### Random Network

**Assumption:** Link are created randomly.

Two types of Random network models:

* G(N,L) where L link are placed randomly
* G(N,P) where each pair of nodes N is connected with probability p

With model 1 we can compute average degree, while with the second one is more complicated (but other properties are easier).

Asymptotically the two are equivalent

### Random Network Erdos-Renyi model 🡺 G(N,p)

**Algorithm**:

1. Starts with N isolated nodes: we will consider N(N-1)/2 (directed) or N(N-1) (undirected) pair of nodes
2. Select a node pair and generate a random number r between [0,1]
3. If r <= p connect the pair, otherwise continue
4. Repeat step (2) for each pair of nodes

Binomial to calculate the probability P(L) 🡺 probability of having exactly L links with N nodes and probability p.

Average number of links: Average degree: Variance:

How to choose N and p for the comparison with the network under study?

* <Lrand> = L
* p = 🡺 L = <L> then

Usually an ensemble random networks are used.

# Scale free property

## Scale free networks

### Power law

Pareto 80-20 rule state that 80% of money is earned by only 20% of population.

A scale free network is a network whose degree follow a power law distribution.

Critical phenomena:

* At a critical point all the system is correlated
* **Scale invariance:** there is not scale for the fluctuation (scale-free behaviour)
* **Universality:** exponents are independent of the system details

We can use different plots for the power lows:

* **Log-log plot**
* **Avoid Linear Binning**
* **Use logarithmic binning**
* **Use cumulative distribution**

Random network model is characterized by a Poisson degree distribution in contrast to power-law distribution as seen in real networks.

In a random networks all vertices are different, while in real network are characterized by a small number of vertices with very large degree (while most vertices maintain a very low degree)

### Hubs in Scale free Network vs. Random networks

In a random network hubs are forbidden, while in scale-free network are naturally present.

Why? There are two assumptions of the Erdios-Renyi model that are violated in real networks!

* **Growth**: Number of nodes is fixed, while in real networks continuously expand by the addition of nodes.
* **Preferential Attachment:** nodes prefer to connect to the more connected nodes (rich gets richer)

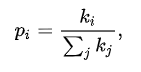
So we are more likely to link to a hub than to a node with only few links.

### Barabasi-Albert model

**Algorithm**

The network begins with an initial connected network of nodes.

New nodes are added to the network one at a time. Each new node is connected to existing nodes with a probability that is proportional to the number of links that the existing nodes already have. Formally, the probability that the new node is connected to node is::



where is the degree of node *i* and the sum is made over all pre-existing nodes *j* (i.e. the denominator results in twice the current number of edges in the network). Heavily linked nodes ("hubs") tend to quickly accumulate even more links, while nodes with only a few links are unlikely to be chosen as the destination for a new link. The new nodes have a "preference" to attach themselves to the already heavily linked nodes.

# Connectivity

## Connected Components

### Path

A sequence of incident links visited one after another where nodes and links are distinct

### Connected Graph

A graph is connected if there exists a path between any pair of nodes in it.

### Connected Component

A connected component is a subgraph of a network such that there exist at least one path between each member and no other vertex in the network can be added to the subgraph (maximality) 🡺 there is no path between a node in the component and a node not in the component.

Can a network have more than two large components ? No! It is unluckily that n/2 node are not connected with the other n/2 of a graph.

### Directed strongly connected graph

If there exist a directed path between any pair of nodes

### Directed weakly connected graph

If there exist a path between any pair of nodes, without following the edge direction.

In real world typically there is one strongly connected component and a selection of small components.

## Random-Real Network Connected Components

Real-world networks: giant component and power-law connected component size distribution.

Starting with N isolate nodes link are gradually added through a random process. This corresponds to gradual increase of p. We first inspect how size N­­G of largest connected component of the network varies with the average degree <k>.

### Giant components

NG number of nodes in the giant component

If p = 0:

<k> = 0, NG  = 1

Else if p = 1:

<k> = N-1, NG = N

One would think the largest component would increase gradually but this is not the case. Once it exceeds a critical value the ratio NG / N will increase signalling an emerging of a giant component.

The critical value condition predicted by Erdos-Reniy is **<k> = 1**

To compute the fraction of nodes in the giant component:

S = NG/N

Using and taking the log on both side for <k> e New get

Using S= NG /N then S = 1-u

### Critical values of Giant components

* For a value subcritical of *<k>* < 1:
  + Network consist of tiny component
  + No giant components
  + Largest cluster is a tree ( size of ln N).
* For a value critical of *<k>* = 1:
  + Vanishing fraction of nodes
  + Numerous trees
  + Cluster distribution:
* For a value supercritical of *<k>* > 1:
  + Unique giant component
  + Non vanishing
  + Cluster distribution exponential :
* For a value Connected of *<k>* > ln N:
  + One cluster: NG = N
  + GC is dense and cluster distribution none.

# Centrality

### Degree Centrality Undirected networks

Ranks nodes with more connection higher in term of centrality

where di is the degree for node vi

**Normalise Degree centrality:**

* By max possible degree
* By max degree
* By degree sum

### Degree Centrality Undirected networks

## Geometric Centralities

Rely on the concept of degree or distance measure

### Degree Centrality

Measure how well-connected a node is.

### Betweenness centrality

Nodes with high betweenness centrality have control over information flowing in the network.

Number of shortest path from s to t passing through vi

The number of shortest path from s to t.

### Closeness centrality

Influential and central node can quickly reach other nodes

is the average shortest path length to other node 🡪 consider the inverse:

Low values for less central nodes, high for more central nodes.

## Spectral Ranking

### Eigenvector Centrality

Not all friends are equivalent, having more friends does not guarantee someone is import, but having important friends provide a stronger signal.

Eigenvector incorporate the importance of neighbours

For directed graph adjacency matrix is symmetric 🡺 use right eigenvectors for the in-going links

### Katz Centrality

We add a bias term and a controlling term to solve the problem where weakly connected component centrality become 0.

Katz Problem: in directed graphs, once a node has high centrality it passes all its centrality along all of is outlinks.

To mitigate it we divide the value of passed centrality by the number of outlinks

### Page Rank

Random surfer on the web, the WWW is described as adjacency matrix here then

A transition matrix is obtained dividing each row by it’s sum. A row describes the probability of moving from a page to a page *j.*

We are interest in the visit on a certain node during random surfing with a probability p.

So using a Markov Chain:

Issues:

* Isolated component
* Dangling nodes 🡺 node without outlinks.
* Loops

To solve it we use **teleport or tax:**  at each step go to the following page with probability and teleport to a random node with probability

The final result is:

# Transitivity

If are connected and are connected 🡺 are connected

A friend of my friend is my friend.

**Partial transitivity** is more useful! The friend of my friend is not guaranteed to be my friend but is far more likely to be my friend than any other.

Clustering coefficient: measure of the degree to which nodes in a graph tend to cluster together.

## Global Clustering

We want to quantify the level of transitivity of a network.

### Coefficient based on paths

### Coefficient based on triples

### Coefficient based on triangles

## Local Clustering

Measure of the clustering coefficient at node level

## Clustering in random network

Since edges are independent and have same probability p:

The clustering coefficient of a random graphs is small and C is independent from a node’s degree k.

So the random network modes doesn’t capture the clustering of real networks.

# Ties and Bridging

Granovetter’s paper: bridging local and the global.

The experiments: interviewed people about how they discovered their job 🡺 thanks to acquaintances not the friends!

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns.

The strength of a tie is a combination of the amount of time, intensity, intimacy which characterize the tie.

When A and B has strong tie the larger is the proportion of individuals to whom they will both be tied.

When we have strong tie A-B and A-C, both B and C if aware of each other, will want to have a tie: strong or weak base on psychological attitude.

**Strong triadic closure property:** if a node A has two strong links to B and C then a link (**strong** of **weak**) must exist between B and C.

### Bridge

Edge between A and B is a bridge if, when deleted, it would make A and B lie in 2 different components.

### Local Bridge

An edge is local bridge if deleting the edge would increase the distance of the endpoints to a value more than 2. (Triangles not a local bridge).

**Bridges are weak ties:**

If A satisfies the STCP and is involved in at least two strong ties, then any local bridge must be a weak tie.

### Almost Local Bridge

Edge with very small neighborhood overlap are almost local bridge

Neighborhoob overlap of and edge between node A and B is:

# Six degree of separation

### Stanley Milgram’s experiment

Random people asked to send a letter to another one. She could only send to someone with whom they were close. Average path lengths was six!

Facebook: 4 degree of separation.

Should we be surprise that paths between random pairs of people are so short? No! Taking a tree of friends, if anyone has 100 friends, getting deep in the tree will get to all the people in the world (five steps) 🡺

### Small world

**Average length path depend logarithmically on the size system.**

Small means that is proportional to

Random graphs tends to have a tree like topology with constant node degrees.

The average distance between two random node:

Do real network deviate from this model?

Yes, they abound of triangles and in particular many of your 100 friends will know each other.

### Watts-Strongatz model

Regular lattice+ rewiring.

Rewire condition: if A from B to C if does not create loop or multiple edge.

The Watts–Strogatz model is a random graph generation model that produces graphs with small-world properties, including short average path lengths and high clustering.

It accounts for clustering while retaining the short average path lengths of the ER model. It does so by interpolating between a randomized structure close to ER graphs and a regular ring lattice. Consequently, the model is able to at least partially explain the "small-world" phenomena in a variety of networks

**Limitations**

The major limitation of the model is that it produces an unrealistic degree distribution. In contrast, real networks are often scale-free networks inhomogeneous in degree, having hubs and a scale-free degree distribution. Such networks are better described in that respect by the preferential attachment family of models, such as the Barabási–Albert (BA) model. (On the other hand, the Barabási–Albert model fails to produce the high levels of clustering seen in real networks, a shortcoming not shared by the Watts and Strogatz model. Thus, neither the Watts and Strogatz model nor the Barabási–Albert model should be viewed as fully realistic.)

The Watts and Strogatz model also implies a fixed number of nodes and thus cannot be used to model network growth.

# Balance and status

## Social Balance theory

Social Balance theory discusses consistency in friend relationship among individuals.

* Positive edge demonstrate friendship (
* Negative edges demonstrate being enemies

Triangle of *nodes i,j* and k is balanced IFF:

## Social Status theory

Social Status theory define how prestigious an individual is ranked within a society.

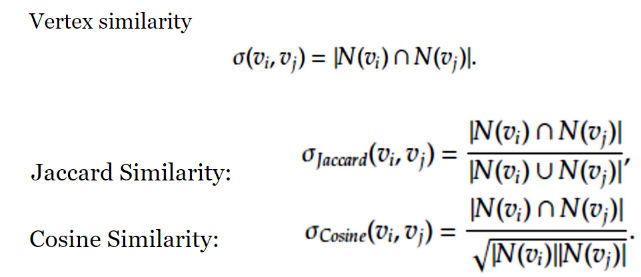
If X has a higher status than Y and Y has a higher status than Z, then X should have a higher status than Z.

# SImilarity

How similar are two nodes in a network?

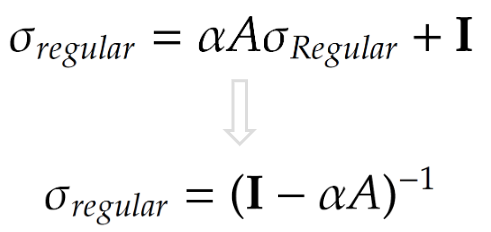
* Neighbourhood
* Attributes
* Contents

## Structural Equivalence

In structural equivalence, we look at the neighborhood shared by two nodes; the size of this neighborhood defines how similar two nodes are

## Regular Equivalence

In regular equivalence, we do not look at neighborhoods shared between individuals, but how neighborhoods themselves are similar. (Athletes similar because they know similar individuals as coaches).



A vertex is highly similar to itself, we guarantee this by adding an identity matrix to the equation

# Influence and Homophility Assorativity

Hw

## Measuring Assortativity

# Social Influence

# Crawler

### Download issues

### Policies

## Page Processing and storage

### Bloom filters

### Sieve

### Flush

# Communities

### Definition of communities

### Maximumm cliques

### Partition

## Modularity

### Greedy Algorithm

### Louvain Algorithm

## Overlapping Communities

### CFinder

### Clique Percolation Method

# Information Diffusion

## Definition

## Information diffusion types

## Herd Behaviour

## Information Cascade

### Independent Cascade Model (ICM)