Social Media Analytics

Alberto Filosa

30/9/2020

Indice

1 Scenario

Huge amount of data are exchanged on Social Media through various devices. Information process is fully integrated within object and activities.

Information is provided in real time on various social platform:

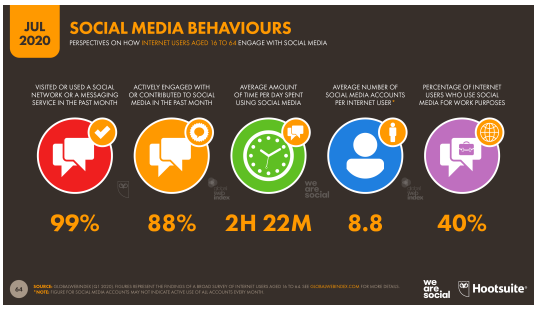
- Entities: books, records, etc.;
- Events: launch of a new product, concert, etc.;
- News: entertainment, politics, etc.;
- Relationships: between entities, events, etc.

When collecting data, it is necessary to transform information, organized in classified data, in to knowledge, awareness and understanding in which info can be useful.

Social ***M***edia ***A***nalytics (**SMA**) has a complex process of identifying relevant, valid, new and potentially useful information which can be transformed into knowledge used by different actors and purpose. The main applications of SMA are:

- Life quality, for transactions;
- Brand reputation (awareness of the brand, sentiment, bad comments, consoling reputation, ect.);
- Forecasting the performance of finance markets.

The aim of the course is to provide the skills to start a Social Media specialist path to learn how to extract significant insights from the huge volume of mostly unstructured data using Social Media Analytics methods and tools.



1.1 Internet

Internet is the global system to connect computers around the world. TCP/IP protocols allow devices connected through internet to communicate with each other. The ***W***orld ***W***ide **W**eb (**WWW**) is one of the major service of Internet. The proposal was designed to provide more effective communication system within *CERN*. It allows to browse through web pages and services accessible to all or a selected part of users.

Uniform ***R***esource ***I***dentifier (**URI**) is a sequence of characters that uniquely identifies a generic resource, such as web address (URL), documents, images, etc.

There are different type of data:

- *Structured*: data are stored in databases and organized in a rigid schemes and tables (Relational Scheme);
- *Semi-Structured*: data aren't stored in a tabular structure, but it contains tags to separate semantic elements (E.g. HTML, XML, JSON);
- *Unstructured*: data are stored without any scheme (Narrative Text).

HTML is a markup language, a computer language that uses tags to define elements in a document. It is human readable, so it also contains standard words. There is the head of the page, containing the title and other part, and the body, containing the divisors and paragraph of the text. Other semi-structured data format is *XML*, designed to store and transport data. The basic unit of XML documents are elements, a specific block of text marked with a pair of tags at the end and at the beginning of the element. The organization follows a hierarchical structure (or a tree structure) including a main element, called root element. The Document Type Definition define the components allowed in the building of an XML document, defining legal attributes and elements in the document and the structure of each element. Finally, *JSON* is schema-less representation of structured data based on a key-value pairs. It is used to store and exchanging data. JSON and XML are human-readable, have a hierarchical structure and can be parsed and used by lots of programming languages. Main differences are that JSON doesn't use the end tag, so it is shorten and more readable, is quicker to read and write than XML and can us arrays.

2 Social Media

The social aspect of communication was to facilitate interaction between people who shared strong relationships, the same interests or find themselves working together in specific geographical contexts. The availability of information is made possible also by the presence and diffusion of Web Browser, Search Engines and *Social Media*.

There are three main distinction between data collected through the Web:

- *Virtual Data* (or Provoked Data), obtained through conventional research methods, such as queries online, and these answers are provided by users following specific questions;
- *Digitized Data*, analog data transformed in another digital formats (e.g. e-books, music, etc.);
- *Digital Data*, traces left by users, such as page visits or interactions with Social Media. In this case they are generated spontaneously by the users visiting a web page.

Social Media has specific characteristics:

- It is an Interactive Web 2.0 application;
- Contents are generated by users (posts, comments, etc.);
- Users create service-specific profiles;
- Social Media facilitates the development of online social networks.

One of the most important theory about Social Media is the *Theory of Social Presence*. It states that media differ in degree of social presence between the influence of the degree of intimacy and immediacy of the means of communication.

> The higher the social presence, the larger is the social influence that the communication partners have on each other's behavior.

The *Media Richness Theory* is based on the assumption that the goal of any communication is the resolution of

ambiguity and the reduction of uncertainty. The media differs in the degree of information richness (quantity and quality information transmitted in a given time period) they have. The concept of *Self Presentation* states that in any type of social interaction people have a desire to control the impressions of other people. This concept is related to self-disclosure.

> Social Media are **interactive computer-mediated** technologies that facilitate the **creation and sharing of information**, ideas, career interests and other forms of expression via **virtual communities** and networks.

Low	Medium	High
Blogs	Social Networks	Virtual Social Networks
Collaborative Projects	Content Communities	Virtual Game Worlds

A ***U***ser-***G***enerated ***C***ontent (**UGC**) is any form of content created by users in an online system made available on social media. The reason for creating UGC are:

- *Implicit Incentives*, not based on anything tangible, not directly monetizable. Social Incentives are the most common form. A user feels as active user of the community, also through the interaction of friends. It also improves the customer experience when purchasing a product;
- *Explicit Incentives*, referred to tangible rewards. They are easily understood by most people and have immediate value regardless of the size of the community (E.g. contest or voucher). The main disadvantage is they can make believe the user that the only reason of participating is explicit incentive, reducing the influence of other type of interactions.

A *Virtual Community* is made up of Social Network of individuals who interact through means of communication, crossing any kind of boundaries in order to pursue mutual interests.

A *Social Network* is a social structure made up of a set of individuals, a series of dyadic links and other social interactions. There are many methods to analyze the structure of social entities, called ***S***ocial ***N***etwork ***A***nalysis (**SNA**), to identify local and global patterns and examine network dynamics.

There are many type of Social Media:

- *Blog*, a website containing one or more chronological posts that interactively discuss a certain topic;
- *Microblog*, a textual or multimedia publications content on Internet;
- *News Sites*, the digital identity of the paper edition of main newspapers. In these sites is possible to interact with other people;
- *Forums*, discussion sections in an IT platform or single sections;
- *Social Networking Sites*;
- *Virtual Game Worlds*;
- *Virtual Social Worlds*;
- *Content Communities*.

These sites produce different types of data:

- *Articles*, mainly form news sites;
- *Posts*, generally blog and Social Networking Sites;
- *Tweets*, Twitter's content;
- *TWeeds*, a conversation developed between multiple users;
- *Reviews*, ratings left by users;
- *Images* and *Videos*.

2.1 Social Media Analytics

Web Analytics is the measurement, collection, analysis and reporting of Web data for the purposes of understanding and optimizing Web usage.

> *Social Media Analytics* is monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content. Interaction takes place in workplace and external-facing communities.

It is possible to divide the analyzes into two main categories:

- Social Network Analysis, for example Centrality Measures and Community Detection;
- Social Content Analysis, such as Natural Language Processing and Text Mining.

SMA in the *business* context is the process of measuring, analyzing and interpreting interactions and conversations in Social Media regarding a brand, product, service or a topic of interest. The analysis is based on elements that influences the customer behavior, the perception and feelings of the customer respect to a brand end their level of engagement. There are many business objectives:

- *Brand Advocacy*, an indicator that expresses the highest degree of brand loyalty, so a customer can recommend the brand to other people;
- *Business Reputation Management*, constantly monitoring and managing the company's online presence and how the brand is perceived in a positive light and allowing to gain new customers;
- *Community Management*, the process of building an authentic community through various interactions;
- *Demand Generation*, the strategy of various marketing programs aimed at certain interest in the products/services of a company.

To plan an analysis activity is necessary to answer at these questions:

- *Why*: the reasons that should lead to online analysis;
- *What*: what does the analytic activity consist of and what are the phases that make it up;
- *How*: what are the tools and resources to equip to create a structured and effective monitoring process;
- *Who*: who are the actors involved in the monitoring and analysis process of the web and Social Media;
- *Where*: what are the sources tha must be taken into consideration for an effective analysis;
- *When*: when and with what timing it is necessary to activate the listening and analysis strategy.

The phases of Social Media Analysis are:

1. Planning of the analysis activity;
2. Data identification;
3. Data Analysis, models and techniques that allows to answer to the objectives developed, selected or integrated among those available;
4. Information interpretation and visualization, better interpreting information obtained from data analysis.

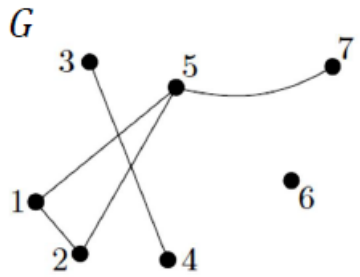
3 Network and Graph Theory

Network Theory is the study of structures representing symmetrical or asymmetrical relationships between objects. It is also a part of Graph Theory: a graph is composed by nodes (also known as vertices) connected through links (edges). Examples of networks are:

- Technological networks, designed for distribution of goods, resources and services (E.g. Internet);
- Information networks, made up of data and information linked together in some way (E.g. ***W***orld ***W***ide ***W***eb);
- Biological networks, representing the interaction patterns between biological elements (E.g. Neural Network);
- Social networks, where vertices represent people of groups connected by some form of social interaction, such as friendship.

In particular, ***S***ocial ***N***etwork ***A***nalysis (**SNA**) is the process of investigating social structures through the use of networks and *Graph Theory*. These networks are usually visualized in sociograms (also known as Social Network Graph) where vertices are represented as points and edges as lines.

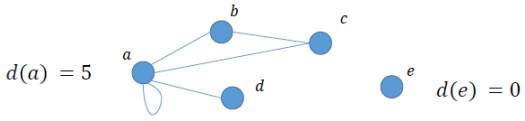
A *Graph* is a pair of vertices (V) and edges (E), $G = (V, E)$, such that $E \subseteq [V]^2$. The set of *vertices* of a graph is denoted as $V(G) = \{v_1, v_2, \dots, v_n\}$, while the set of *Edges* is denoted as $E(G) = \{e_1, e_2, \dots, e_n\}$. The representation is to draw a point for each vertex and join these points with lines if there's a connection:



3.1 Undirected Graph

An *Undirected* graph is a graph with all bidirectional edges (E.g. Facebook). Given the edge $e = (a, b)$, a and b are called *Extreme* and *Adjacent* vertices of e, called the *incident* edge. a is called *Neighbor* of b in G and vice versa. Two edges e, f are *Adjacent* if they have a common vertex. The *Neighborhood* of a, $N(a)$, is the set of vertices adjacent to a. The *Star* of a, $s(a)$, is the set of edges incident in a.

The *Degree* of a vertex, $d(v)$ is the number of edge incident to a node, equivalent to the number of neighbors of v; each loop is counted twice, while a vertex with degree 0 is an isolated vertex. The value $\delta(G) = \min\{d(v) \mid v \in V\}$ is the minimum degree of a vertex, while the value $\Delta(G) = \max\{d(v) \mid v \in V\}$ is the maximum degree:



The Average Degree is defined as follows:

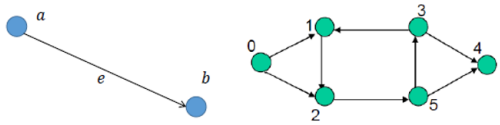
$$d(G) = \frac{1}{|V|}\sum_{v \in V} d(v) = \frac{1}{|V|} 2|E(G)| = \frac{2m}{|V|}$$

There are many different type of graphs:

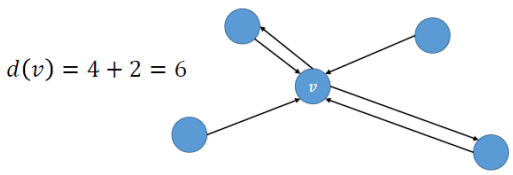
- *Null* graph, made up of only isolated graphs;
- *Regular* graph, if all vertices of G have the same degree k;
- *Complete* graph, in which each pair of distinct vertices are adjacent;

3.2 Directed Graph

A graph is *Directed* if edges are directed from one vertex to another (E.g. Twitter). Given the edge $e = (a, b)$, e is called *Outgoing* edge from a, denoted as direct *Predecessor*, and *Ingoing* edge in b, known as direct *Successor*. Given a vertex a of a directed graph G, $E^+(a)$ is the set of outgoing edges from a and $E^-(a)$ the set of incoming edges in a. A *Sink* vertex is a vertex with only incoming edges, while with only outgoing edges is called *Source* vertex.



The *In-Degree* of a vertex v is the number of edges arriving at the v vertex, while the *Out-Degree* is the number of edges starting from the vertex v.



The *Degree* $d(v)$ of a vertex v is the sum of the number of its incoming and outgoing edges.

3.3 Subgraph

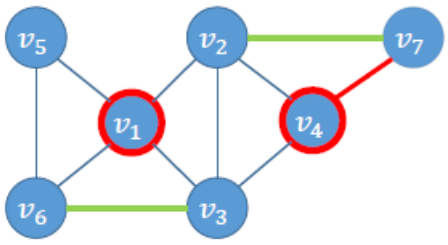
A *Subgraph* of a graph $G = (V, E)$ is a graph included within the main graph that contains all vertices and edges in G. It is possible to obtain subgraphs by removing some vertices and/or edges from G.

Given a graph $G = (V, E)$ with $V = \{v_1, \dots, v_n\}$ and $E = \{e_1, \dots, e_n\}$, a *Walk* is a finite (or infinite) alternating sequence of vertices and edges. A walk is called *Simple* if edges and vertices of the walk are all distinct, otherwise it is *Not Simple*. In particular, there are many different types of walks:

- *Trail*, in which all edges are distinct;
- *Path*, in which all vertices are distinct;
- *Directed*, for each edge in the walk the initial vertex is the head and final is the tail of the walk;
- *Closed*, where the extreme vertices coincide (also known as cycle). A circuit allows repetition of vertices, but not edges.

3.4 Connected

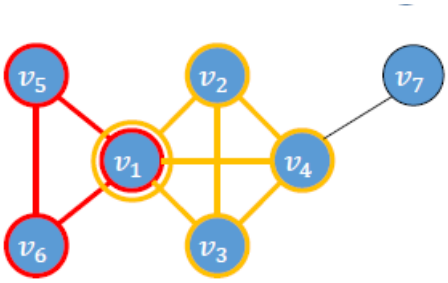
Connectivity $k(G)$ is the basic concept of graph theory. It measures the minimum number of elements (Vertices and Edges) that must be removed to disconnect the graph. In particular, an *Articular* point is a vertex whose removal disconnects a component of the graph, while a *Bridge* is an edge whose removal disconnect a component of the graph.



When the number of elements removed are 2, $k(G) = 2$, it is *Biconnectivity*, so no single edge or vertex removal disconnects the graph and no network failure points compromise the network itself.

3.5 Clustering

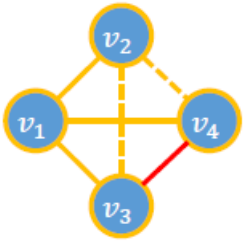
A *Clique* is a set of vertices C totally connected in a graph G. It usually ignores single vertices and vertex connected by an edge. In particular, a *Maximal Clique* is a clique not extended adding a new adjacent vertex, while a *Maximum Clique* is the largest clique in a graph.



The *Clustering Coefficient* is the degree in which nodes tend to be connected to each other. There are 3 different ways to calculate this:

- *Local Clustering Coefficient*: given a set of neighbors $N(v)$, it is the number of edges between the members of neighbors divided by the number of potential edges between them ($k = d(v)$):

$$cc(v) = \frac{||N(v)||}{k(k-1)} \qquad cc(v) = \frac{2||N(v)||}{k(k-1)}$$

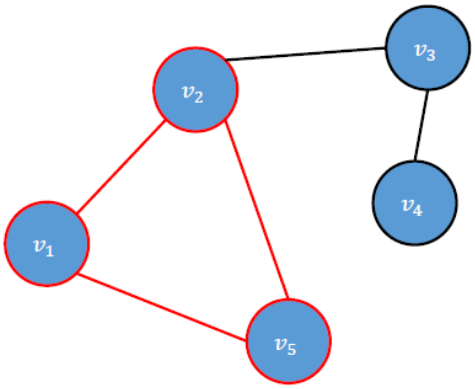


- *Average Clustering Coefficient*, which is the average of the clustering coefficient for each single node of the graph G:

$$cc(G) = \frac{1}{|V|}\sum_{i=1}^{n} cc(v_i)$$

- *Global Clustering Coefficient*, based on triples of vertices (pen if 3 nodes are connected by 2 edges and Closed if connected by 3 edges). It is the number of the closed triplet divided by the total number of triplets:

$$cc_{\triangle} = \frac{3n_{\triangle}(G)}{n_{\wedge}(G)} = \frac{\sum_{i=1}^{n} cc(v_i)w_i}{\sum_{i=1}^{n} w_i}$$



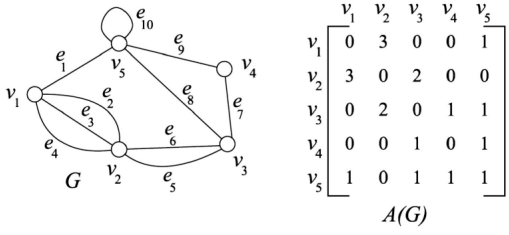
3.6 Tree

An *Undirected Tree* is an undirected, connected and acyclic graph in which a node is designated as the root, while a *Directed Tree* is a directed graph that has a root node and there are no arcs entering the root, each node has exactly one incoming edge and for each node there is a path from the root to the node.

The *Depth* of a tree is the length of the path from the root to the node, a *Level* the set of nodes at the same depth and *Tree Height* the maximum depth reached by leaves

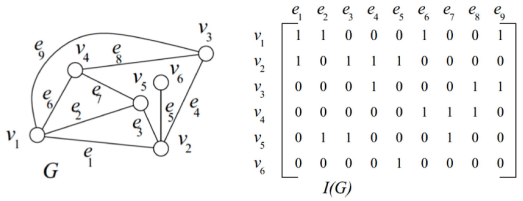
3.7 Graph Representation

Let G be a graph with $V(G) = \{v_1, v_2, \dots, v_n\}$ and $E(G) = \{e_1, e_2, \dots, e_m\}$. The *Adjacency Matrix* $A(G) = [a_{ij}]$ is an $n \times n$ matrix where a_{ij} are the number of edges between two vertices v_i and v_j.

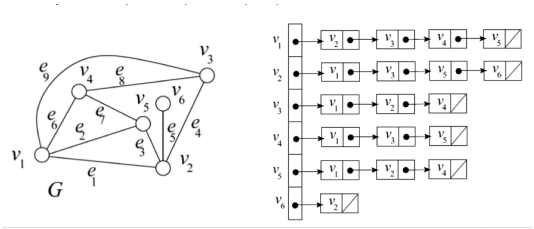


The *Incidence Matrix* $I(G) = [m_{ij}]$ where:

- $m_{ij} = 1$ if v_i is incident to e_j, otherwise $m_{ij} = 0$;
- $m_{ij} = -1$ if e_j leaves v_i, $m_{ij} = 1$ if e_j enters v_i, otherwise $m_{ij} = 0$;



The *Adjacent List* $Adj(G)$ is an array of n lists. To each vertex of G corresponds a list containing its neighbors:



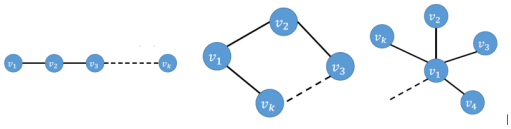
The weight of the edge (u, v) is store with the vertex v in the list of u.

3.8 Complex Network

Interaction networks are complex systems composed of several parts connected to each other and intertwined with each other so that the result is different from the sum of the parts. The structure (Topology) of the contact network is crucial in determining collective behavior.

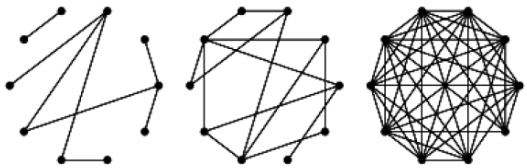
In *Regular Networks* each node is connected to a fixed number of nodes. They have regular patterns within the structure, in which they may or not be k-regular graphs and the Entropy ≈ 0, the degree of randomness. There are many examples of regular networks:

- *Linear Network*, a linear sequence L of connected vertices: $L = v_1, e_{12}, v_2, e_{23}, v_3, \dots, v_{k-1}, e_{(k-1)k}, v_k$;
- *Ring Network*, in which each node is connected to exactly two other nodes, forming a ring: $A = v_1, e_{12}, v_2, e_{23}, v_3, \dots, v_k, e_{k1}, v_1$;
- *Star Network*, a tree with a single vertex of maximum degree: $S = v_1, e_{12}, v_2, v_1, e_{13}, v_3, \dots, v_1, e_{1k}, v_k$.

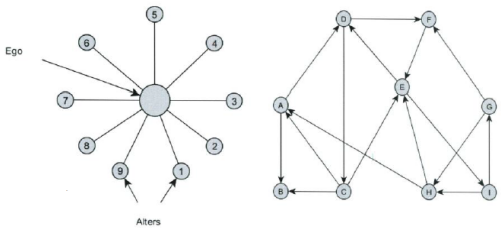


Summary	**Linear**	**Ring**
Order	$\|L\| = k$	$\|A\| = k$
Degree	$1 \leq d(L) \leq 2$	$d(A) = 2$
Clustering Coefficient	$cc(L) = 0$	$cc(A) = 0$
Size	$\|L\| = k - 1$	$\|A\| = k$
Diameter	$diam(L) = k - 1$	$diam(A) = k/2$
Degree of Connectivity	Not Biconnected	Biconnected
Adjacency Matrix	Diagonals	Diagonals + Angles

In *Random Networks* pairs of nodes are randomly connected by a given number of connections. Two nodes are connected by a certain probability. All nodes have approximately the same number of neighbors, which differs slightly from the average value.



Complex Networks of interactions have substantially different characteristics from both classes. The Complex Network Theory shows non intuitive characteristics and can be made up of millions of units communicating with each other. Mathematical methods are used to extract information from complex networks in a synthetic way. For example, online social networks are complex networks: the personal network of contact is usually composed of a first order area, with an ego-centric network, with direct relationship, a second order and so on, with a socio-centric network.



There are several phenomena related to the Theory of Complex Networks:

1. *Small Word*, an experiment remained famous and repeated in various social networks. Milligram randomly chose a sample fo American and asked them to deliver a message to a stranger knowing only few information. The results show how short paths exist between individuals in large social networks and how can be found by ordinary people;

2. *Clustering*, a tendency in social network where are created communities in relation with each other, measured by clustering coefficient (*cc*). Random networks resemble interaction networks with reference to the small world property, but they differ in the clustering coefficient because in a typical interaction network *cc* is usually much larger than a *cc* of a random network;
3. *Strength of Weak Ties*, a strength of a tie is given by the probably linear combination of the amount of time, emotional intensity, intimacy and exchange of services characterizing the tie. *Strong Ties* are people united in primary networks (eg family, organization), while *Weak Ties* those who characterize the informal networks of people, more important than the strong one. The *Tie Strength* is measured by identifying *Shortcut Bridges*, loosing the bridge and check if the distance between two nodes increases when the arc is removed;
4. *Scale Invariance*, a negative exponential relationship between the number of nodes and the number of their connection. When a node need to establish a new connection, it prefers to do it with a node (tipically called *Hubs*) with many connections, leading to exponential growth. The role of hubs is connecting areas of the graph that would otherwise be separate.

4 Metrics for Social Media Analytics

There are three main families of metrics:

- *Connection*, in which social network entities are connected;
- *Distribution*, in which information can flow within a social network;
- *Segmentation*, clustering the components of the social network.

4.1 Connection Metrics

Homophily is the tendency of individuals to associate and tie with each other similar one, sharing common characteristics (E.g. gender or age) that facilitate communication and relationship formation. Social Media favors homophilic relationships because it shows similar posts of your interest by your user profile examined with likes and interactions personalizing your content. The perception of interpersonal similarity improves coordination and the expected payoff of interaction and helps people access information forming opinions and threads, but can translate in a limited social worlds forming closed communities or filter bubbles on Social Network sites where people with similar ideologies interact only with each other (Echo Chambers).

Multiplexity is the number of relationship levels contained in a link, associated with the strength of the bond.

Mutuality/Reciprocity is the extent to which two actors mutually exchange friendship or other interaction.

Network Closure measures the completeness of relational triads, using various clustering coefficient *cc* to measure this closure. It refers to the concept of triadic closure, a property in which 3 nodes A, B and C have strong tie $A-B$ and $A-C$, then there is a strong or weak tie $B-C$. Another theory based on Network Closure is the Cognitive Balance, the propensity of two individuals to want to try the same things toward an entities that unites them. If the triad is not closed, then the people connected want to close thi triad to achieve closure in the relationship. In a social network a strong triadic closure occurs with a high probability because they have many things in common and they can create a link. It can predict the development of ties in a network and show the progression of connectivity.

Proximity/Propinquity is the tendency for actors to have more links with others who are geographically close, measured with geolocation information. For example, two people living on the same floor have a higher propensity to establish relationships than those living on different floors.

4.2 Distribution Metrics

Centrality is a group of metrics that aim to quantify the importance or influence of a particular node in a network. In Graph Theory it identifies the most important vertices, in a Social Network the most influential people. The general case is called *Degree Centrality*, the risk of a node to catch whatever is flowing through the network (in the graph theory is the degree of a vertex): $c_D(v) = d(v)$. In a directed graph is necessary to explicit the in-degree and out-degree. It is possible to normalize the degree centrality, computed as:

$$\overline{c_D} = \frac{c_D(v)}{n-1}$$

In a connected graph the *Closeness Centrality* $c_C(v)$ is the reciprocal of the sum of the distances from v to all other nodes:

$$c_C(v) = \frac{1}{\sum_u d(v,u)}$$

The Normalized version is expressed as $\overline{c_C}(v) = (n-1)c_C(v)$.

The *Betweennes Centrality* $c_B(v)$ is the sum of the number of the shortest path between each pair that cross the vertex ($\sigma_{st}(v)$) divided by the number of shortest paths between each pair of vertices in a graph (σ_{st}):

$$c_B(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

It is possible to compute the Normalized Betweennes Centrality as the $c_B(v)$ divided by the maximum possible number of geodesics crossing node v ($2c_B(v)$ for undirected graphs):

$$\overline{c_B}(v) = \frac{c_B(v)}{(n-1)(n-2)}$$

Delta Centrality is the importance of a node in the entire network, calculated by comparing the performance $P(G)$ (as the number of edges) to the performance $P(G')$ obtained deactivating the node v:

$$C_\Delta = \frac{P(G) - P(G')}{P(G)}$$

It is possible to calculate the Delta Centrality using Efficiency, measured as the inverse of their distance $d(v_i, v_j)$. The efficiency of a graph is:

$$E(G) = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \frac{1}{d(v_i, v_j)}$$

Density is the percentage of effective links in a network out of the total possible number. A dense graph is a graph in which the number of effective edges approaches the number of potential edges, while a sparse graph the number with few edges than the potential ones. The degree is measured as follow ($2|E(G)|$ if undirected graph):

$$D(G)\frac{|E(G)|}{n(n-1)}$$

Other Distribution Metrics are:

- *Distance* is the minimum number of ties necessary to connect two people;
- *Structural Holes* is the absence of links between two parts of a network;
- *Strength of a Tie* is the linear combination of time, emotional intensity, intimacy, reciprocity, ect.

4.3 Segmentation Metrics

There are two different type of Counting:

- *Counting of Cliques*, if each individual is directly related to each other individual;
- *Counting of Social Circles*, groups of individuals less closely linked than in a clique.

The *Cohesion* is the degree of actors directly related connected to each other cohesive bonds. An important metric is the *Structural Cohesion*, defined as the minimum number of people or ties of a Social Network that must be removed to disconnect the group (identical to Connectivity).

5 Online Interactions and Privacy Issues

Data about users is every day collected, analyzed and possibly shared with/without the consent of the user and made available by the himself. Users, owner of these data, can lose control on what information is collected and how the information is used. *Privacy* is not only a technological issue and laws vary widely throughout the World.

Personal Information can be collected in the Web in application form or questionnaires and surveys. When a user looks for a new Web page, the URL of the page that the user is currently looking at can be sent along. The referral link field can also reveal personal information. It is used to promote products to new customers through referrals as traditional word of mouth. Every time a browser download a page, this is recorded in a log file on the remote Web server, registering IP address, the time and URL requested. This information can be combined with other log files to discover the actual identity of the user. Privacy can be enhanced by using proxy servers, an intermediaries between a client and a server, or anonymizers, other proxies designed to maintain accesses on the Web hiding IP addresses.

A *Cookie* is a small file created by an Internet site to store information on a user's computer. It allows to identify a user with its preferences, such as language and identifiable information. Only the Web site that created the

cookie can read it. Cookies maintain records of browsing habits: it may may include information a web site knows about the user and track users activity form multiple visits.

5.1 Privacy and Social Media

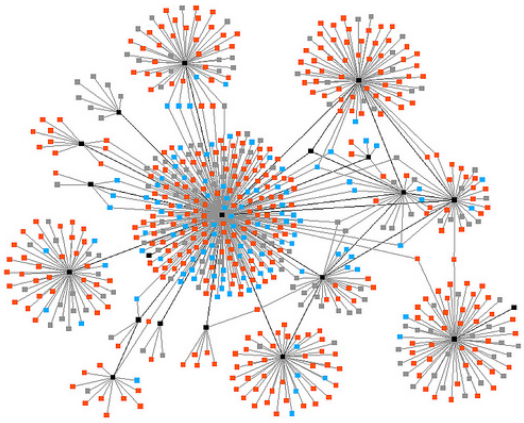
In recent years, users have played an active role in the Information Society, making their information available themselves. They can stay in contact with people, share own ideas and preferences. Social Media offer definition of a personal profile, sharing preferences and interest, creating a digital identity. With the success of Social Media, millions of individuals have spontaneously chosen to make their personal information available online. So, users publish a tremendous amount of personal and sensitive information.

Users can at least discover that their data have been used, but data can also be used to harm, profile, or manipulate a user without her being even aware. Once a data item is published on a Social Network, control over it is lost and it is impossible to know how it will be used.

In 2018 *Cambridge Analytica* had harvested personal data of millions of people's Facebook profiles without their consent and used it for political purposes. In 2013 a researcher from Cambridge University created a personality quiz app collecting data of those users' friends. At the end, the app retrieved data from 87 millions Facebook users. • The retrieved data were shared with Cambridge Analytica, which used them to create physiographical profiles of the respondents segmenting individuals based on their personality. Profile creation was done by analysis of personality test of the 300.000 users and collected Facebook data for the remaining ones. Results of the personality test were matched to the Facebook activity of the same users to create a map. Profiles allowed for identifying the most effective kind of advertisement to persuade a given individual for a political event.

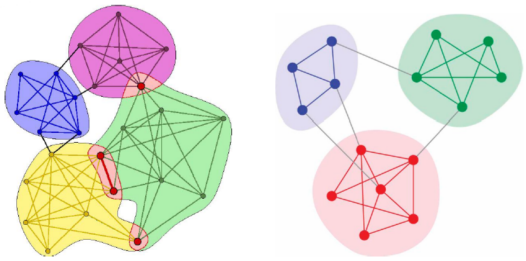
6 Community Detection

Preferential Attachment is a property of some network where the great majority of new edges are connected to nodes with an already high degree, compared to others. The result is a network with very few nodes with an high degree of connection.



It can happen because people want to be associated with popular people, ideas, etc, increasing their *popularity*; *quality*, evaluating people based on objective quality criteria, so higher nodes will attract more attention; *mixed model*, among nodes of similar attribute.

A *Community* is a subset of nodes among which there are relatively strong, direct and frequente ties (E.g. group, cluster). One of the most important task studying network is identifying communities, which allows to discover groups of interacting people, studying interaction between them, inferring missing node values or predicting unobserved connections. In Social Network communities represent groups of nodes (users) that show similar characteristic and interaction among them (link prediction). The definition can be subjective, because it can be densely-knit or each component can be a community. Communities can be overlap, in which some users belong to 2 communities, or disjoint, in which each node belongs exclusively to 1 community.



It is important to detect communities because it is possible to predict the connection among them. The Link Prediction is the prediction of which nodes are likely get connected given a Social Network; the output is a list of ranked pairs of nodes.

The Community Detection is confused for clustering, but in the first data is linked in a graph and network data tends to be discrete, leading algorithm using the graph property directly.

Community Detection methods can be divided into 4 main categories:

6.1 Node-centric Community

The *Node-centric Community* is a community in which each node in a group satisfies certain properties (generally with networks with small size):

- *Complete Mutuality*, where nodes belong to just one clique. A clique is a maximal complete subgraph of 3 or more nodes all adjacent to each other. It is possible to find communities for the maximum cliques or all maximal cliques. The brute-force is not recommended, it is necessary to prune nodes with degree equal or less than $k-1$. Even with pruning cliques are rare and a single edge removal destroys the clique;
- *Reachability of Members*, in which all nodes in the group is reachable in k-jumps. The k-clique approach aims to find a maximal subgraph in which the largest distance is between any nodes less than k (shortest path given a subgraph). The k-club approach aims to find a substructure of diameter $\leq k$;
- *Nodal Degrees* ().

6.2 Group-centric Community

The *Group-centric Community* is a community which requires the whole group to satisfy a certain condition. The main concept is the network density, defined as the number of edges in the network over the total number of possible edges between all pairs of nodes. It measures how well connected a network is (perfectly connected network is a clique). A subgraph is a quasi-clique if

$$\frac{2|E_s|}{|V_s|(|V_s|-1)} \geq \gamma$$

A greedy algorithm can be adopted to find a maximal quasi-clique, starting from the node with the largest degree and expanding it with nodes likely contribute to a larger quasi-clique. This method continues until no mode node can be added.

6.3 Network-centric Community

Network-centric Community aims to find communities considering global connection of nodes and divide network into disjoint set. The main approaches are:

Node Similarity measures how similar their interaction patterns are. Two nodes are structurally equivalent if they connect to the same set of nodes (they share the same set of neighbors). This definition is too strict because rarely occur in a large scale and is difficult to compute. To resolve this it is possible to use vector similarity, creating a matrix of connections and calculate the Cosine Similarity or Jaccard Similarity:

$$cos(\theta) = \frac{AB}{\|A\|\|B\|} J(A,B) \quad = \frac{|A \cap B|}{|A \cup B|}$$

It is useful for huge network for computing the vertex similarity and applying k-means clustering algorithm: each cluster is associated with a centroid and each node is assigned to the cluster with the closest centroid (recompute the centroid of each cluster until centroids don't change).

Another possible approach is the *Spectral Clustering*, representing a similarity graph as a matrix and analyse the spectrum (eigenvalues λ) of the matrix:

$$\begin{bmatrix} w_{11} & \dots & w_{1n} \\ \vdots & \dots & \vdots \\ w_{n1} & \dots & w_{nn} \end{bmatrix} \times \begin{bmatrix} w_{11} \\ \vdots \\ w_{n1} \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

The Spectrum of a matrix is a set of ordered eigenvectors of their corresponding eigenvalues: $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$

The goal is embed points in a low-dimensional space, in which cluster denoting communities are more easy to find. It is necessary to compute the Adjacency matrix and compute the Laplacian matrix:

$$A_{ij} = \begin{cases} w_{ij} \text{ weight of edge } (i,j) \\ 0 \text{otherwise} \end{cases}$$

So the basic stages of spectral clustering are:

1. Pre-processing: build a matrix representation of the dataset;
2. Decomposition: compute eigenvalues and eigenvectors of the matrix and map each point to a lower dimension based on 1 or more eigenvectors.
3. Grouping: assign points to 2 or more clusters based non the new representation.

Modularity Maximization tries to maximize the distance of a network from a random network. Modularity measures the group interaction compared with the expected random connection ($M \in (-1,+1)$ and $M = 0$ if all nodes are clustered into one group). In a network with

E edges, for two nodes d_i and d_j the expected random connection between them is $\frac{d_i d_j}{2E}$. The modularity of a group is

$$\sum (A_{ij} - \frac{d_i d_j}{2E})$$

and is necessary to maximize:

$$\frac{1}{2E} \sum_C \sum_{i,j \in C} (A_{ij} - \frac{d_i d_j}{2E}) z_{ij}$$

where z_{ij} defines if two edges are in the same group (0 otherwise). If $A_i j$ is very close to the expected random network, the relationship between two nodes is weak ($A_i j$ matrix considered as random). Finally it is possible to compute the spectral clustering and using k-means approach to select the communities.

6.4 Hierarchy-centric Community

Hierarchy-centric Community aims to build a hierarchical structure of communities based on network topology. It facilitates the analysis of different resolutions. There are two approaches:

The *Divisive Hierarchical Clustering* divides the nodes into several set, where each set is further partitioned into a smaller set. To determine these partition it is possible to use Network-centric Methods. One particular example is based on edge-betweenness, called Girvan-Newman Algorithm (repeat until all edges are removed):

1. Compute the edge-betweenness for all edges in the graph;
2. Remove the edge with the highest betweenness;
3. Re-compute betweenness for all edges affected by the edge removal;

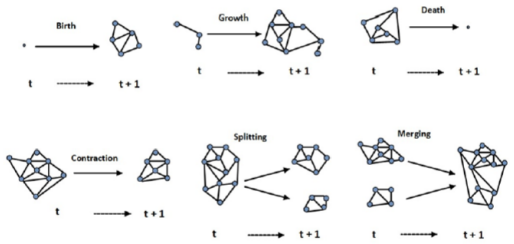
Agglomerative Hierarchical Clustering initializes each node as a community. It chooses two communities satisfying certain criteria and merges into larger one. It is possible to use Maximum Node Similarity or Maximum Modularity Increase approach.

6.5 Evaluation

The perfect communities would be objects of the same type in the same community, but this is simply impossible. There are two types of evaluation:

- *With Ground Truth*, partial knowledge of what communities look like and it is given the correct community assignments. Measures are:
  - Precision, Recall and F-Measure;
  - Purity;
  - Normalized Mutual Information (NMI).
- *Without Ground Truth*, in a matter of analyzing attributes, such as posts, profile information, of community members to see if there is a coherency among them. It is possibile to use word frequency and wordclouds to find the top topics discussed by communities and find any coherence. Another approach is evaluate the communities using clustering quality measures, such as select the best Community Detection algorithms value.

Communities also expand, shirk or dissolve in dynamic networks:



7 Assortativity and Dynamics

In some cases hubs connect to hubs and sometimes they avoid each other, it depends on the structural properties. This pattern is a general property of real networks: it is a phenomenon known as *Degree Correlation* e, defined as the probability to find a node with degree i and j at the two ends of a randomly selected link (E is the number of edges between two elements).

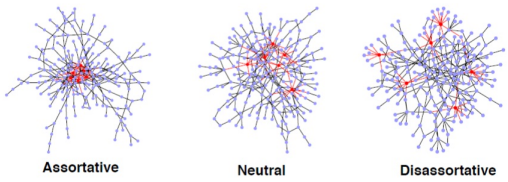
$$e = \frac{E}{\|E\|}$$

The probability q_k that there is a degree-k node at the end of the randomly selected link is the following (p_k is the degree probability):

$$q_k = \frac{k p_k}{\overline{k}}$$

Hubs are expected to link to each other on how many links they have. They can be:

- *Assortative*, hubs show a tendency to link to each other. e_{ij} is high along the main diagonal so nodes of comparable degree tend to link to each other;
- *Neutral*, nodes connected to each other with the expected random probabilities. The density of links is symmetric around the average degree, so nodes link to each other randomly;
- *Disassortative*, hubs tend to avoid linking to each other. e_{ij} is high along the secondary diagonal, so hubs tend to connect to small degree nodes.

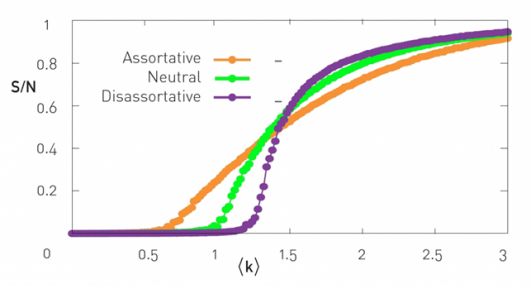


If the networks has no degree correlations, $e_{jk} = q_j q_k$, otherwise the magnitude of the correlation is captured by $\sum_{j,k} jk(e_{jk} - q_j q_k)$. This is expected to be:

- > 0 for assortative networks;
- $= 0$ for neutral networks;
- < 0 for disassortative networks.

To compare different network, it is possible to normalize the degree correlation coefficient:

$$r = \frac{\sum_{j,k} jk(e_{jk} - q_j q_k)}{\sum_k [k^2 - q_k - (\sum_k k q_k)^2]} \in [-1, +1]$$

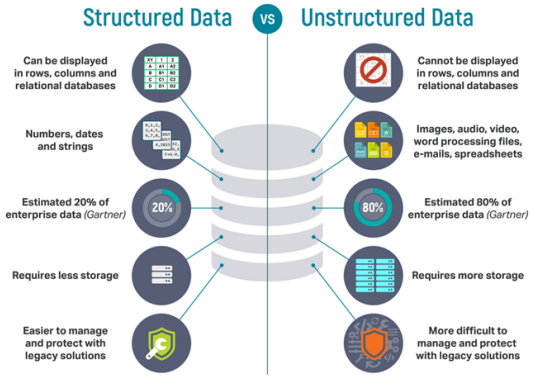


An important property of a random network is the emergence of a phase transition at $\langle k \rangle = 1$, marking the appearance of the *Giant Component*. For Assortative networks, the transition happens to a lower value of $\langle k \rangle$, hence a giant component emerges for $\langle k \rangle < 1$ because it easer to start it if the high-degree nodes seek out each other. Otherwise, for Disassortative networks the transition is delayed since the hubs tend to connect to small degree nodes and have difficulty to form the giant component.

Alteration of the giant component have implications for network robustness. The *Network Resilience* is the effects of node and edge removal. In Assortative networks hub removal makes less damage because the hubs form a core group, while in disassortative networks is more damaging because hubs are connected to many small degree nodes. It is related to *Percolation*, when a giant component is formed.

8 Sentiment Analysis

From Social Network is possible to extract information about a given argument and analyzed them to make decisions (E.g. predicting market data, better quality of life). This is not so simple because it is necessary to transform data from unstructured (cannot be displayed in rows, requires more storage and are more difficult to manage) to structured data, displayed in rows-columns and requires less storage.



The transformation is necessary to extract some important information like emotion, polarity, topic and the language of the user.

In order to extract some useful information, it is necessary to set some basic components for SMA:

- Identify the *Opinion Holder* (User) that holds a specific opinion;
- Identify the *Object* on which an opinion is expressed;
- Identify the Aspect of the object;
- Identify the *Opinion* on an object from a user.

Social Media Analytics should be able to interpret Natural Language to discover if an information is subjective and determine the corresponding polarity. A text is *objective* if a post presents factual information about the world, while *subjective* if expresses feeling, emotions

or beliefs. An *explicit* opinion is a subjective statement giving an opinion, while *implicit* implies an opinion. An *ironic* post is a communicative way expressing the opposite what is literally said. Emotions are complex reaction pattern involving experiential and physiological elements in which a person attempt to deal with.

The process of SMA is constituted by:

1. *Collect* real time expressions or opinion on-line or off-line;
2. *Represent* text from qualitative to quantitative data to measure compare and learn. The pre-processing techniques are remove stopwords, numbers and punctuation and apply the stemmatization. Furthermore, it is possible to represent in a vector space model the text. It is possible also to use the Word2vec, a model used to create a distributed representation of words in a corpus. There are two main algorithm:
    - Skip-Gram, predicting the surrounding (context) words based on the current (centered) word. Given a sliding window of a fixed size moving along a sentence, the word in the middle is the target, while near words (in the sliding window) are context words. Then it is trained to predict the probabilities of a word being a context word for the given target. The model is a Neural Network with just one Hidden Layer representing the word embedding of size N. The Input layer x and the Output Layer y are one-hot encoded word representations;
    - CBOW, predicting the target word from source context words;
3. *Classify* measuring, compare and learn sentiments. It is possible to classify in different ways:
    - Lexicon-based Approach, for each sentence the process identifies term belonging to the list containing positive and negative terms with different weights. This lists is contained in a lexicon and one of the most important is the Liu lexicon opinion. The polarity of a text is given by the number of the positive and negative terms. For example, if $\text{Word}_{\text{positive}} > \text{Word}_{\text{negative}}$ the word is positive;
    - Supervised Learning, an approach in which some text are labeled and it is possible to learn a model in order to predict new texts with the corresponding polarity. To deal with negation is possible to add $NOT\_$ to every word between negation.
    - Semi-Supervised Learning uses a small amount of information (e.g. few labeled examples and hand-built patterns) to enrich the lexicon predefined. It is possible to label a seed set of words and expand it to conjoined adjectives and cluster into two partition (positive and negative). Another possible approach is extract a phrasal lexicon from reviews using Part of Speech Tagging, identify and quantify the polarity and rate a reviews by the average polarity of phrases using Mutual and Pointwise Mutual Information ;
    - Unsupervised Learning, suitable for aspect-based Sentiment Analysis.
4. *Predict*;
5. *Decide*.

9 Semantic Ambiguity

The ***N***amed ***E***ntity ***R***ecognition (**NER**) is the task to find and classify text fragments into predefined labels. The most important model is the *Conditional Random Fields*, an undirected discriminative graphical model trained to maximize the conditional probability to identify the name entity given tokens in a phrase.

$$\mathbb{P}(y|x) = \frac{e^{\sum_{t=1}^{T}(\sum_i \lambda_i f_i(y_t,x)+\sum_j u_j g_j(y_t,t_{t-1},x))}}{Z(x)}$$

where:

- Feature Function: $f_i(y_t, x)$
- Transition Feature Function: $g_j(y_t, t_{t-1}, x)$

The main goal is to estimate the values of the weight vectors. λ_i and u_j are parameters to be estimated maximizing the log-likelihood of a given training set (Quasi Newton Method).

Once trained the model, it is necessary to find the most probable sequence of hidden states: $y^* = \max_y \mathbb{P}(y|x)$. It is possible to use the Viterbi algorithm, a dynamic programming, or the Shortest Path, an integer linear programming formulation.

Once recognized entity mentions, it is necessary to disambiguate them, using the **Named Entity Linking**. Steps are:

1. *Candidate Resource Selection*, computing a scoring measuring lexical similarity between an entity and the candidate label and coverage based on the coherence of the entity:

$$KB(e_j, c_k) = \alpha lex(e_j, l_{c_k}) + (1-\alpha)(\text{cov}(e_j, c_k))$$
$$lex(e_j, l_{c_k}) = lcs(e_j, l_{c_k}) + W_D(\frac{JW(e_j, l_{c_k})}{W_D + 1})$$
$$\text{cov}(e_j, c_k) = \cos(e_j, \alpha_{c_k} + R(c_k))$$

2. *Entity Linking and Type Classification*, ranking recourses by KB score and select the optimal one for an entity;
3. *Entity Boundary Re-Scoping*, post processing of entities for filtering noise.

9.1 Word Sense Disambiguation

The Language is intrinsically ambiguous and it is difficult to distinguish differente entities connected to a term. There are 3 different approaches:

- *Knowledge-Based Disambiguation*, using an external resources such as dictionaries or lexicons. Classes of this field are:
  - Machine readable dictionaries: for each word in the vocabulary, this method provides a list of meanings with its respect definition and an example;
  - Thesauruses, adding an explicit synonym relying the meaning of the word;
  - Semantic networks, adding semantic relation;

  The Lesk Algorithm identifies sense of words in context using definition overlap. First, we retrieve from MRD all definitions to be disambiguated, the determine the definition overlap for all combination and choose the sense with the highest overlap. When applied this algorithm with more than 2 words there are a lot of combination and it is impossible to choose the optimal sense. It is introduced the Simplified Lesk measuring the overlap between sens definitions of a word and the current context, identifying the correct sens for one word a time.
- *Supervised Disambiguation*, based on labeled training set. The learning system has a training set of feature-encoded inputs and their category;
- *Unsupervised Disambiguation*, trying to overcome the knowledge with no manual tagging, no domain specific knowledge, but the sense of the sentence is defined by its context.

One of the most important algorithm is the *Chinese Whispers* for undirected and unweighted graph. First, all nodes are assigned to a random class equal to the number of nodes (a sort of agglomerative). All nodes are selected in a random order and moved in the class with most connections until a predetermined number of iteration or convergence. This model is efficient because is time linear in the number of edges and it does not have any parameter to specify. Unfortunately, Chinese Whispers is not deterministic due to random order processing and in most of the case the algorithm cannot converge.

Another approach is the *Graph-based*. When the sentence is collected about a word, first it is necessary to pre-process the text document (removing stopwords and do the POS tagging). Secondly, we build a word co-occurrence graph in which each word is anode, each edge represents a co-occurrence in a text unit and weights are the frequencies. Then it is necessary to cluster words by wordcloud, so the neighbors of a node form it word cloud. After we compute a Weighted Jaccard Distance to establish the correspondence <Word Graph - Word Metric Space>. Given a node, we define its weighted neighborhood as the multiset; Now it is possible to cluster in an agglomerative way.