# 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:

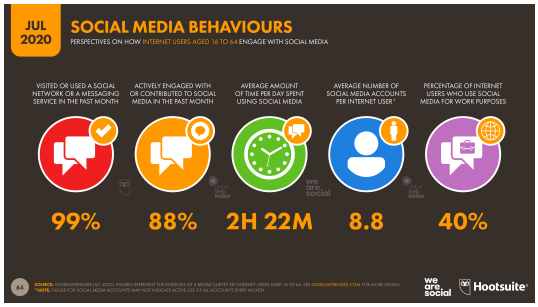
- Entities: books, records, etc.;
- Events: launch of a new product, concert, etc.;
- News: entertainment, politics, etc.;
- Relationships: between entities, events, etc.

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

Social media analytics 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 transaction;
- 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 browser through web pages and services accessible to all or a selected part of users.

***U***niform ***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 (HTML, XML, JSON);
- *Unstructured*: data are stored without any scheme (Narrative Text).

## 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-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: * Interactive web 2.0 applications; * Contents are generated by the users (posts, comments, etc.); * Users create service-specific profiles; * Social Media facilitates the development of online social networks.

The *Theory of Social Presence* 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 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 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.

A ***U***ser-***G***enerated ***C***ontent (**UGC**) is any form of content created by users of 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 improve 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 (contest, voucher, etc.). The main disadvantage is they can make user believe 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 of social actors. 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*, publicational textual or multimedial content on the net;
- *News Sites*, the paper edition of the main newspaper. 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 type of data:

- *Articles*, mainly form news sites;
- *Posts*, generally blog and Social Networking Sites;
- *Tweets*, Twitter's content
- *Threads*, 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 take place in workplace and external-facing communities.

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

- Social Network Analysis;
- Social Content Analysis.

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.

Question to plan an analysis activity:

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 yo 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

Social Media Analysis phases: 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

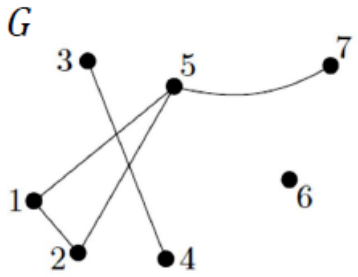
*Network Theory* is the study of structures representing symmetrical or asymmetrical relationships between objects. It is also a part of Graph Theory: a graph is composed by nodes (also known as vertices) connected through link (edges).

Other 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.

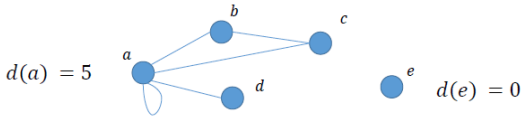
A *Graph* is a pair of vertices ($V$) and edges ($E$), $G = (V, E)$, such that $E \subseteq [V]^2$. The set of *vertices* of a graph is denoted as $V(G) = v_1, v_2, \dots, v_n$, while the set of *Edges* is denoted as $E(G) = e_1, e_2, \dots, e_n$. The representation is drawing a point for each vertex 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 $e = (a, b)$, $e$ 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 *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, while the value $\Delta(G) = \max\{d(v) \mid v \in V\}$ is the maximum degree:

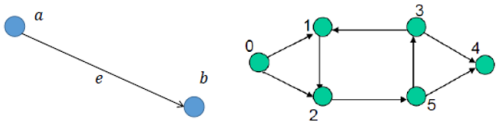



There are many different type of graphs:

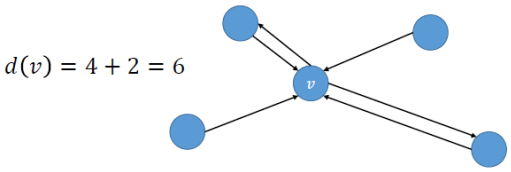
- *Null* graph, made up of only isolated graphs;
- *Regular* graph, if all vertices of $G$ have the same degree $k$;
- *Complete* graph, in which each pair of distinct vertices are adjacent;

### 3.2 Directed Graph

A graph is *Directed* if edges are directed from one vertex to another (e.g. Twitter). Given the edge $e = (a, b)$, $a$ 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 *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

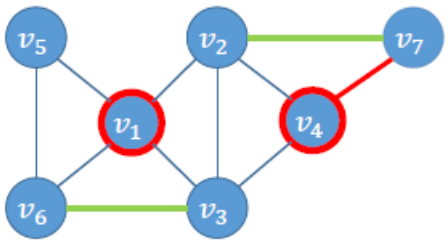
A *Subgraph* of a graph $G = (V, E)$ is a graph included with in the main graph that contains all vertices and edges in $G$. It is possible to obtain subgraphs by removing some vertices and/or edges from $G$.

Given a graph $G = (V, E)$ with $V = \{v_1, \ldots, v_n\}$ and $E = \{e_1, \ldots, e_n\}$, 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

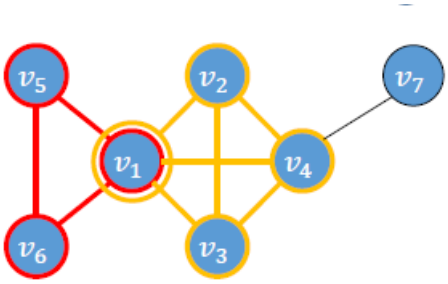
Connectivity $k(G)$ is the basic concept of graph theory. It measures the minimum number of elements that must be removed to disconnect the graph. In particular, an *Articular* point is a vertex whose removal disconnects a component of the graph, while a *Bridge* is an edge whose removal disconnect a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




When $k(G) = 2$, we talk about *Biconnectivity* no single edge or vertex removal disconnects the graph and no network failure points compromise the network itself.

### 3.5 Clustering

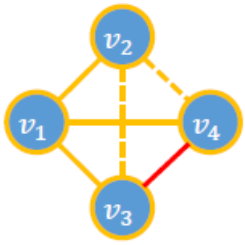
A *Clique* is a set of vertices $C$ totally connected in a graph $G$. It usually ignores single vertices and vertex connected by an edge. In particular, a *Maximal Clique* is a clique not extended adding a new adjacent vertex, while a *Maximum Clique* is the largest clique in a graph.




The *Clustering Coefficient* is the measure of 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 ($k = d(v)$):

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

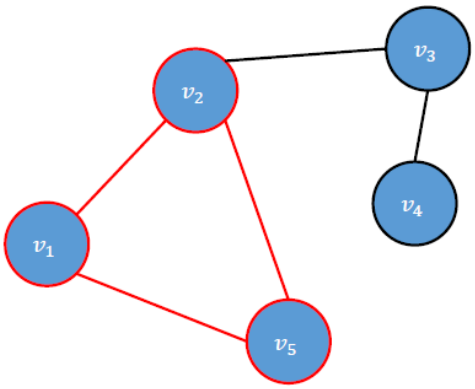



- *Average Clustering Coefficient*, which is the average of the clustering coefficient for each single node of the graph $G$:

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

- *Global Clustering Coefficient*, based on triples of vertices(Open 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:

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



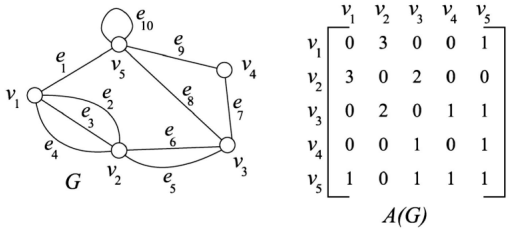

### 3.6 Tree

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

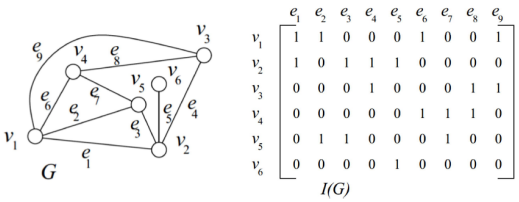
### 3.7 Graph Representation

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 $v_j$.

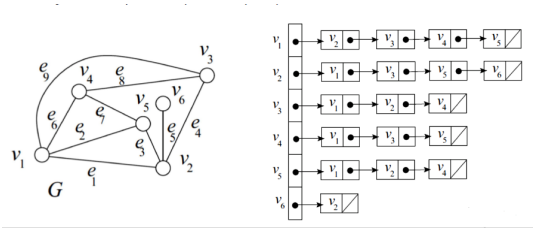



The *Incidence Matrix* $I(G) = [m_{ij}]$ where:

- $m_{ij} = 1$ 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:



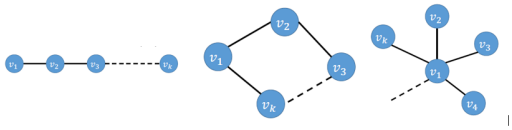

The weight of the edge $(u, v)$ is store with the vertex $v$ in the list of $u$.

### 3.8 Complex Network

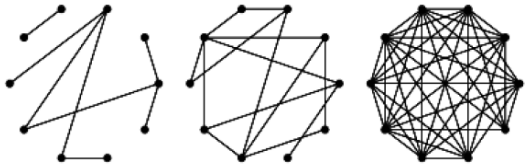
Interaction networks are complex systems composed of several parts connected to each other and intertwined with each other so that the result is different from the sum of the parts. The structure (Topology) of the contact network is crucial in determining collective behavior.

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 $\approx 0$, the degree of randomness. There are many examples of regular networks:

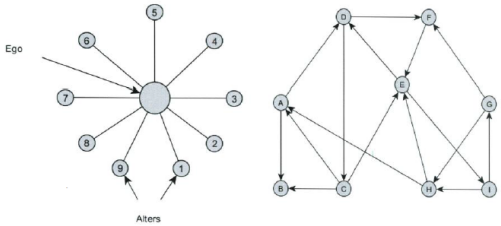
- *Linear Network*, a linear sequence $L$ of connected vertices: $L = v_1, e_{12}, v_2, e_{23}, v_3, \dots, v_{k-1}, e_{(k-1)k}, v_k$;
- *Ring Network*, in which each node is connected to exactly two other nodes, forming a ring: $A = v_1, e_{12}, v_2, e_{23}, v_3, \dots, v_k, e_{k1}, v_1$;
- *Star Network*,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



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; 2. Clustering; 3. Strength of Weak Tries; 4. Scale Invariance.