# Lez 0 - Intro

Social media platforms: persuasives and ubiquitous (information processing fully integrated within everyday activities…). We use devices that not only collect data generated on purpose, but also other useful data that can provide useful information (steps, heartrate, locations data with smartwatch…).

Data: information with context. We collect a lot of data to make decisions based on information acquired. We have a lot of data, sometimes more than we really need: challenge to have systems working on data when they get published on the internet. With the pandemic, social media have had a huge growth on the number of users; and more users means also more information.

The types of data can be: entities (books, records, tv series, characters, restaurants…), events (launch of products, events, concerts…), news (politics, entertainment, culture…), relationships (between entities, events, news…). Information is transformed in knowledge, used to make informed decisions.

We want to extract relevant (info that satisfies the user needs), valid, new, potentially useful information. SMA can be used with different aims: better life quality, improve brand reputation, forecasting the performance of financial markets… So, the main goal of this course is to know how to extract significant insights from the huge volume of mostly unstructured data using social media analyticsmethods and tools. The main concepts behind social media management are: access to data, pre-processing of data, representation of data and finally analysis.

# Internet and web technologies

Internet: global system of interconnected computer networks that use the Transmission Control Protocol/Internet Protocol (TCP/IP) protocol suite to connect devices around the world (network of networks). Connection of physical devises, via multiple options (wifi, cable…). Internet of things: concept of multiple devices (mainly sensors, such as lights, smartwatch, demotics…) can be connected daily. LAN (local area network) is a group of devices connected to the same network. The set of all the interconnected networks is called Internet. If we think of the internet as a network, nodes are the devices and the edges are the connections among these devices.

We need to establish some communication rules to help the devices communicate with each other 🡪 these are called **protocols**. This set of rules (called TCP/IP, transfer control protocol) allow different devices to communicate with each other. The IP address is the unique identification of a device on the internet/network, and this IP can be auto assigned or fixed by the user. At every moment each device connected to the network have different IP addresses, to recognize with each other.

The WWW, World Wide Web, is one of the biggest service on the Internet. The concept is due to Tim Berners-Lee, who elaborated a project to share documents inside the CERN where they used to work. The Internet is a connection of devices, the web is a connection of documents (as documents, images, videos, webpages in the form of html pages). The idea was to represent data with hypertext and share these with some basic communication rules (different from the TCP/IP rules). In the internet we identify only physical devices uniquely, in the WWW we use URI (Uniform Resource Identifier) to identify a generic resource; some example are URL (Uniform Resource Locator), a document, an image, a file, an email address... Also the web can identify as a network, with web pages as the nodes (with other multimedia contents) and the URLs as the edges. The main concepts behind the Web are hypertexts and hyperlinks.

**Structured data** (relational model, we use a DBMS) 🡪 tables/datasets. **Unstructured data** (we need Information Retrieval model to get data) 🡪 data without a scheme, like a text. **Semi-structured data** (semantic separated from the content) 🡪 some structure (markers/tags) that identify the meaning of the content, not as schematic as a table.

Web pages are written with semi-structured language (markup languages), which is HTML (**HyperText Markup Language**). (different from a programming language!!! It simply defines the structure of the content). With HTML we can also connect pages among them. An HTML file is fixed (syntax-wise, with tags already defined).

Other semi-structured data formats are: **XML** (Extensible Markup Language: is a markup language, designed to store and transport data, and to be self-descriptive) and **JSON** (JavaScript Object Notation: is a syntax for storing and exchanging data). The basic unit of XML are elements (specific block of text within a document, marked with tags); it has an hierarchical structure, like a document tree. To give some rules to XML we can do a Document Type Definition (DTD; define structure of each element and its attributes, with also some mechanism to simplify document management), or an XML schema (can do the same things as a DTD, but can define the same rules and behaviors inside the code).

JSON is similar to XML, with its own rules, but it uses a different syntax (different tags); it is constitute of pair of name/value, all written inside {}. The main differences between XML and JSON are: JSON is shorter, quicker to write and read, can use arrays and hasn’t the end tag. Both are self-describing, hierarchical and can be parsed and used by lots of programming languages. Many DBMS use JSON to store data, most from NoSQL (like MongoDB). Some other use XML to store data, such as MySQL, Oracle…

HTTP (HyperText Transfer Protocol) protocol to transfer data on the Web; it is a protocol that works with a Client/Server architecture: the client (browser) execute a request message, the server (machine where the Website resides) returns the response message. In peer2peer (p2p) network, resources are shared among peers with no central coordination by a server.

Classification between:  
- Static Web: built using HTML, stored on servers, useful for sites that never or rarely needs update.   
- Dynamic Web: they are constructed dynamically either on server side (requests like on Amazon) or on client side (like when moving the mouse on some part of the page; usually written in JavaScript); written in scripting languages and inserted within the HTML pages, processed when the user requests them.

**Web 1.0**: when users couldn’t interact with the pages, they were static and could only be read.   
**Web 2.0**: first used in 1999, no technical improvements, but changes in Web pages designs and usage (bi-directional communication, not only sites-users, but also users-users).

**Web 3.0**: semantic web technologies, with great computer capacity, growing usage of AI and new algorithms aimed at new 3D environments. Semantic related to the algorithm to explain more about the knowledge acquired on the web. To build this semantic web (term coined by Tim Berners-Lee), we need to build an environment with documents associated with information and data that specify the semantic context in a format suitable for automatic processing, querying and interpretation. Assertions are used to give information, and they are formed by subjects (Marco), predicates (works in) and values (Milan); the RDF (Resource Description Framework) is a tool useful to encode, exchange and reuse structure metadata with the aim of interpretability.

**Web 4.0**: still an unstable definition, but the direction where we are going; it should be characterized by augmented reality (all the network around us, Metaverse, domotics…), digital alter ego (we will have an alter ego online with all our information, useful to both us and the others in order to gain information), new interfaces (like home automation) and pervasive information (the society we know now may change very fast thanks to the capabilities of Web 4.0).

Social Web: set of social relationships that connect people through the WWW. Not only social networks, but also education, shopping… The aim to this course. The Social Web was born to facilitate strong relationships (friendship, love, family), share interests…

There is a distinction between data found, collected and analyzed online:  
- virtual data: collected with conventional methods (surveys, interviews…) asking participants; they are also called “provoked data” because they answer precise questions made by the researcher.   
- digitized data: material present on the network but also in other formats (e-books, online newspapers, tv recordings, music, photos…); it is an immaterial archive of potential “analogical” data.   
- digital data: traces left by a user online (web pages visited, searches, interactions on social media…); these are generated spontaneously by the individual, they don’t answer a question by a researcher; the goal of this course.

# Social Media

Social Media: they are interactive Web 2.0 applications, with content (data of all kind, like posts, videos, photos and so on) generated by users, that helps the development of online social networks. They are interacted computer-mediated technologies aimed at the creation and sharing of information via virtual communities. You are provided with a service-specific profile depending on the purpose of your activities on the social media. Social presence: acoustic, visual and physical contact that can be achieved with other users; different intimacy between users for each social media, it depends on the level on mediated communication and on the level of synchronous communication. Media richness: social media to reduce uncertainty (as lack of information); media richness as the quantity and quality of information the social media possesses.

Self-presentation: users are going to generate data on a social media in order to be part of a social community, to build a “reputation”; we present to the community by describing ourselves like we are. Concept strictly related to self-disclosure.

Behind the definition of a specific social media there are functional blocks (focus on which the social media is born), such as: presence, sharing, identity, relationships, conversation, groups. Every social media addresses these blocks based on the purpose of the social media (Linkedin with focus on identity, FB and Instagram on relationships, Tripadvisor on reputation, Youtube on sharing…).

The content generated on social media is called UGC (User-generated content), with both implicit incentives (mainly related to psychology, not to gain something, such as to improve relationtships, to be useful in a community, to achieve status with no personal gains related to this) or explicit incentives (to gain something, either money, reputation… for these reasons, the users may participate only because of the explicit incentive).

A virtual community is made up of a social network of individuals who interact crossing geographical and political boundaries with mutual interests. A social network is a social structure formed by social actors (either individuals or organizations) and the links that relate each user to each other. The links can be formed differently based on the social media we are considering (hashtags, friendships, likes, shares, retweets, comments, posts…). The study of these structures uses SNA (Social Network Analysis) to identify patterns (both local and global), identify influential entities and examine network dynamics.

First creation of a social media: 1971, by Ray Tomlinson, to transfer messages from one node (pc of US universities) to another; in 1973 Bulletin Board Systems are developed; in 1988 Internet Relay Chat is born; in 1990, the WWW is born; in 1997 the SixDegrees theory is developed; Myspace is created in 2003, Facebook in 2004, Twitter in 2006.

There are different types of social media: blogs (websites containing more posts about a topic), microblogging (constant publication of small contents like messages, images, videos, audio), news sides (digital versions of paper journals), forums (set of discussions in an IT platform), social networking sites (FB, Linkedin), virtual game worlds (World of Warcraft), collaborative projects (Wikipedia), content communities (Instagram, Tik Tok, Youtube, Snapchat). There are different types of data on social media: articles, posts, tweets, threads, reviews, images, videos…

Other social media characteristics, we have to provide: accessibility (easy and usually free access), global audience (so the message can be amplified), measurability (every data online can be measured), permanence (data stay for a long time), speed (it is possible to access data in real time), usability (everyone can publish and retrieve information).

Centralized social media (FB, Twitter, Instagram…): they grant free services to users and in exchange they use data shared/published by the users in any possible way. The issues related are privacy concerns (as users you need to trust the platform, and maybe “gain” something, in term of money, reputation…), the need of high scalability (for the potential huge number of users, so huge amount of data, difficult for a centralized architecture), the availability of data (in case of server shutdown/down time, data may not be accessible) and lack of interoperability between different platforms (users to share the same data needs on different platforms need to upload it separately, so APIs needs to be applied on different social media; unless centrally they are connected, like Instagram-Facebook).

The DOSN (decentralized online social network) is an online social network implemented on distributed platforms. No single service provider but a set of nodes cooperating to guarantee the functionalities offered usually by a centralized OSN. Benefits in terms of privacy (no central entity having control and storage on all users data), but problem about availability of data, security... They use the peer-to-peer architecture (every node as either a server or a client on the net), with problems related (not all nodes have the same power, security…).

Opportunistic mobile social networks are a form of mobile networks that exploit human social characteristics in common (interest, daily routines…). For example a network to share data on a university course. They don’t use the peer-to-peer architecture.

A blockchain is a type of Distributed Ledgeer Technology (DLT) consisting of growing list of records/data (blocks) securely linked together using cryptography; it reminds Peer-to-Peer architecture, but guarantee more security (so it is used for example for banks applications), but it can’t be retroactively modified (blocks are permanent and unalterable).

The last attempt to build distributed architectures overcoming some problems leads to Blockchain Online Social Media (BOSM), developed to decentralized control on social data but also introducing a rewarding system. With BOSM we can have content management without a central authority, we can have more security. Steemit is a BOSM, founded on the Steem blockchain: the social community creates and curates content in exchange for 50% of Steem Power and 50% of Steem Dollars.

Web Analytics: collect data from web pages in order to understand and optimize Web usage. Social Media Analytics: is monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content. It includes sentiment analysis, natural-language processing and social networking analysis (influencer identification, profiling and scoring), and advanced techniques such as text analysis, predictive modeling and recommendations, and automated identification and classification of subject/topic, people or content.

Social Media Analysis (on data) and Social Content Analysis (on content, so text mining and natural language processing). Main social analytics tools: Facebook Insights, Twitter Analytics, Instagram Insights… They are usually represented as dashboards, usually related to frequency counts. But we can also use other instruments, such as R/Python libraries, Gephi, and tools just for visualization (Neo4J, Tableau…).

Network: set of interconnected objects (entities), of different types. Technological networks (Internet, gas, water networks, telephone network): tech infrastructures used to connect nodes to allow distribution of goods. Information networks (the WWW, email networks…): man-made networks made to share information. Biological network (biochemical, neural networks): network that represents interactions between biological elements. Social Network: networks that connect people or group of people with social interactions (like the friendship, the follow…).

With social media we can build networks, so we can use the network theory to study these networks (network theory). Network theory as part of graph theory: a network is defined by vertices (nodes) and interconnected with links (edges), which can have attributes. SNA (Social Network Analysis) is the process of investigating social structures through the use of network and graph theory. Different graph representations: algebraic, matrix, graphical… The sociogram is the original tool, made by Jacob Moreno (1889-1974) used in the context of sociometry.

Immagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamenteGraph Theory was introduced by Eulero in 1700; it was first used to solve a problem of bridges (cross all the bridges in one go, with no repetition of a bridge) in the city of Konigsberg. A graph is a pair G=(V,E) of sets (where E includes V2); the elements of V are vertices, the elements of E are edges.

Immagine che contiene testo

Descrizione generata automaticamenteIt is usually represented in a graphical way such as that nodes are point linked by lines, which are edges. The order of the graph is the number of vertices (cardinality of vertex set). The size of the graph is the number of edges (cardinality of edge set).

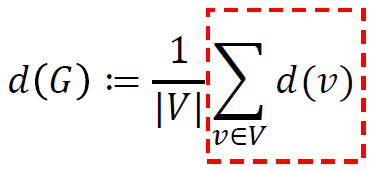
A graph is directed if all the edges have orientation (like the follow on instagram/twitter…); it is undirected if all the edges don’t have directions (like the friendship in Facebook).

With undirected graphs: the edge between two vertices is called **incident edge**; vertices linked with an edge are called **adjacent nodes**; two vertices linked by one edge are called **neighbor** to each other (a neighbor of b); two edges are **adiacent** if they have a common vertex; the **neighbourhood** (N(V1)=… are the neighbors of V1)of a node is the set of vertices adjacent to the vertex; the **star** of a vertex is the set of edges incident in the vertex.

With directed graphs (like web pages, follow…): the **outgoing edge** is ad edge from a vertex; the **ingoing edge** is ad edge to a vertex; a vertex can be **direct** **predecessor** or **direct successor**, if it comes from or to a vertex; E+(a) is a **set of outgoing edges** from the vertex a; E-(a) is a **set of ingoing edges** to the vertex a; a vertex with only incoming edges is called **sink vertex** (pozzo); a vertex with only outgoing edges is called **source vertex** (sorgente).

A loop is an is an edge whose extreme vertices are the same vertex; **multiple edges** (*archimultipli*) are (different) edges with the same pair of extreme vertices; if a graph has no loops and no multiple edges, it is called a **simple graph** (*grafo semplice*); if not, it is called a **multigraph** (*multigrafo*). We can transform multigraph in simple graph (multiple edges in a simple weighted edge).

Immagine che contiene testo

Descrizione generata automaticamenteIn an undirected graph, the **degree of a vertex** (node) is the number of edges incident to a node (the same concept of number of neighbors of a vertex). Each loop is counted twice. The **minimum degree** of a graph is the minimum number of all the degrees in the graph; The **maximum degree** of a graph is the maximum number of all the degrees in the graph. The average degree of a graph is the mean of degrees (sum of degrees divided by the number of vertices) (always a value between the minimum degree and the maximum degree).

**Handshaking lemma**: if the graph is constituted by m edges, the sum of all the degrees is equal to 2 times the number of edges (m) (each edge is incident at exactly 2 distinct vertices, and loop are already counted twice). If all vertices of the graph have the same k degree, they are called **k-regular**. Graphs made up of only isolated vertices are called **null graphs**. A graph where all the vertices are linked to the others are called **complete graphs**.

In directed graphs, the **in-degree** of a vertex is the number of edges arriving at the vertex; the **out-degree** of a vertex is the number of edges coming from the vertex; the **degree** of a vertex is the number of edges arriving and coming at/from the vertex (sum of in-degree and out-degree). A **subgraph** is a graph completely included in a graph; it is possible to form subgraphs of a graph by removing some vertices and/or edges from the graph (the deleting of edges doesn’t involve the deleting of vertex, but the deleting of vertex may imply the deletion of edges) (the graph without an edge is G-e, the graph without a set of edges is G-F; the graph without a vertex is G-v, the graph without a set of vertices is G-W).

A **walk** is a finite - or infinite - alternating sequence of vertices and edges adjacents (both edges and vertices can be repeated) (on simple graphs, no loops and no multiple arcs, a walk can be indicated as a sequence of vertices). A walk is **simple** if the edges and vertices are all different, otherwise is a **not simple walk**. A **trail** is a walk with all edges distinct (but vertices can be repeated), a **path** is a walk with all vertices distinct (but edges can be repeated). A **closed walk** is a walk with extreme vertices that concide. A simple and closed walks is called (**simple**) **cycle**. A **circuit** is a closed walk that allow repetition of vertices but not of edges.

The **length** of a walk is the number of edges crossed; the **distance** between two nodes is the length of the shortest path from one node to the other. The **eccentricity** of a vertex is the max distance of the vertex from any other vertex. The **radius** of the graph is the minimum eccentricity in the graph, the **diameter** of the graph is the maximum eccentricity of the graph.

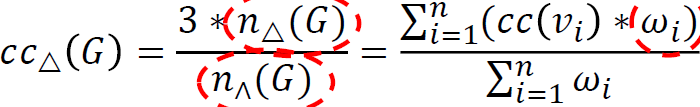
A graph is said to be **connected** when there is a path between each pair of vertices; in a connected graph, there are no unreachable vertices (a graph with only one vertex is connected). **Connectivity** is one of the basic concepts of graph theory: it is indicated by con 𝜅(𝐺) and measures the minimum number of elements (vertices or edges) that must be removed to disconnect the graph. An **articulation point** is a vertex whose removal disconnects a component of the graph. A **bridge** is an edge whose removal disconnects a component of the graph. **Biconnectivity** is when you need to remove two elements to make the graph disconnected (𝜅(𝐺)=2).

A **clique** is a structure (of at least 3 nodes) made up of a set of vertices totally connected in a graph (no single vertices, and vertex pairs are connected by an edge). The maximum clique is the largest clique in a graph. The maximal clique is a clique that can’t be extended by adding a new adjacent vertex.

# Lez 4 (boh)

Boh

# Lez 5

Immagine che contiene testo, orologio

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamenteThe **clustering** (or aggregation) **coefficient** is the measure of how close is a group to being a clique (the degree to which the nodes of a graph tend to be connected to each other). There are some ways to calculate the clustering coefficient:  
- **local clustering coefficient**: number of edges between the set of neighbors of *v* - or N(*v*) - divided by the number of potential edges between them (1° directed graph, 2° undirected graph).   
- **average clustering coefficient**: the average (dividing it by the number of nodes of the graph) of the clustering coefficients for each single node of the graph  
- **global clustering coefficient**: is calculated as the number of closed triples (triangles) divided by the total number of triples in the graph.

Global clustering coefficient is related to the concept of **triples** (triads) of vertices, like open triples (3 nodes connected by 2 edges) and closed triples (3 nodes connected by 3 edges); a triple can be centered on a node; a **triangle** are 3 triples centered on the same 3 nodes that compose them.

A **labeled graph** (directed or undirected) is a graph in which an additional information called a labelis associated with each arc (edge-labeled graph) or vertex (vertex-labeled graph). A **weighted graph** is (generally) a graph labeled on edges with non-negative numbers called weights. Given a path, the **total weight** of the path is (generally) the sum of the weights on the edges in the path (like maps with km length).

A tree is a graph with some particular characteristics. An **undirected tree** is an undirected, connected, and acyclic graph in which a node is designated as the root. A **directed tree** is a directed graph that is empty or has a rootnode such that there are no arcs entering the root, each non-root node has exactly one incoming edge and for each non-root node there’s a path going from the root to the node itself.

Some concept are: **depth** of a node (the length of the path from the root to the node, i.e., number of edges crossed); **level** (the set of nodes at the same depth); **tree height** (maximum depth reached by the leaves). Like the HTML tree, or a website structure tree, or a social media tree.

Graphs may be complex for computer, in reality machines treat graphs in a way different than the graphical one: some data structures used for graph representation are adjacency matrix (an *nxn*, with n number of vertices, matrix with number of edges between each couple of vertices; with simple graph it has only 0 and 1 values; diagonal = 0), incidence matrix (a *nxm* matrix, with *n* number of vertices and *m* number of edges, with value of 1 if the edge enters the vertex, of -1 if the edge leaves the vertex or 0 otherwise) or adjacency list (array of *n* lists, with n number of vertices, that states the neighbors of each vertex).

# Lez 6

**Complex network**: complex (as “intertwined together”) system. The system as a hole is different from the sum of the distinct parts; as a consequence, it can’t be deducted by analysis, but from the interactions between the elements that compose it: these interactions give birth to emerging behaviors (collective spontaneous phenomena). To understand these behaviors we need to study the topology of the network.

**Immagine che contiene testo, orologio, grafica vettoriale

Descrizione generata automaticamenteImmagine che contiene testo, orologio

Descrizione generata automaticamenteRegular networks** are usually mathematics network (with entropy close to 0), where each node is connected to a fixed number of nodes; random networks are halfway between graph theory and probability theory. Examples of regular networks are:  
- **linear networks**: linear sequence L of connected vertices.   
- **ring networks**: linear sequence A of connected vertices, with last elements connected to first, in a ring shape (2 regular graph).   
- **star networks**: a tree S with a single vertex of maximum degree.

**Random networks** (or ER graphs, from the name of its inventor Erdos and Renyi in 1959). Random networks have pairs of nodes randomly connected by a given number of connections; two nodes are connected by a certain probability *p*. All nodes have approximately the same number of neighbors.

Complex networks (like online social networks) have different characteristics from both regular and random networks: they can be made of millions of nodes communicating with each other; mathematical methods are used to extract information for objective and synthetic purposes.

The “first order” area are contacts directed with an individual/node (this is the ego-centric network, and bring to ego-centric analysis: focused on the network surrounding of a specific node), “second order” and so on are linked via one or more links to the starting individual/node (these are called the social-centric networks, and bring to socio-centric analysis: focused on the study of complete networks).

Several phenomena are related to the theory of complex networks:  
- small-world (6 degrees of separations, Milgram experiment, experiment on exchange of emails, Erdos number, Kevin Bacon and the small-world of Hollywood)  
- clustering (communities of users almost all in a relation to each other)  
- strength of weak ties (people with less “affects” ties, but related with an interest, such as job, study… the strength of these ties can also be measured, with instruments like shortcut bridges or neighborhood overlap, or also social media instruments like interactions between users or following relationships)  
- scale invariance (scale-free networks, in the meaning that the number of relationships grow as a negative exponential type, therefore invariant under changes of scale; there are popular nodes, called hubs, and new elements who enters a network may connect more easily to some of them in respect to the others)

# Lez 7

3 main families of metrics:  
- **connection metrics**: entities connected with each other  
- **distribution metrics**: how information flow within social network  
- **segmentation metrics**: ways of “clustering” the components of social network

**Connection metrics:  
- Homophily**: the tendency for actors to form ties with similar actors (similar by gender, age, ethnicity, occupation, academic performance, status, values, beliefs…); opposite of homophily is heterophily; social medias facilitate homophilic relationships (because of personalization, filtering algorithms), which gives value to people on social medias improving the experience on them; negative aspects can be the develop of more closed communities, and the spread of filter bubbles, where only people with the same ideas interact with each other, in so-called echo chambers (more connections within a community than between different communities).   
- **Multiplexity**: the number of relationship levels contained in a link (level 2 for two friends and job collegues).  
- **Mutuality**/**reciprocity**: the extent to which two actors mutually exchange friendship or other interaction.   
- **Network closure**: it measures the completeness of relational triads (ties between A and B, and B and C, usually means strong or weak ties between A and C); synthetized by the theory of cognitive balance: propensity of two individuals to try the same thing towards a common entity; used for clustering coefficient and transitivity; both in trust networks and social networks.  
- **Proximity/Propinquity**: the tendency for actors to have more links with others who are geographically close; it uses geolocation information, or info provided directly by the users.

**Distribution metrics**:  
- **Bridges**: identification of individuals whose ties (bridges) fill a structural hole, providing the only link between two individuals or clusters.  
- **Density**: the percentage of effective links in a network out of the total possible number (n° real edges/n° potential edges; more paths, more dense).   
- **Distance**: the minimum number of ties necessary to connect two particular actors (Stanley Milgram's experiment, the idea of ​​the "six degrees of separation", the Erdos number, the Bacon number).  
- **Structural holes**: absence of links between two parts of a network.  
- **Strength of the tie**: linear combination of time, or emotional intensity, intimacy, reciprocity, etc.  
- **Centrality**: to quantify the "importance" or "influence" (in a variety of senses) of a particular node (or group of nodes) within a network; we can calculate the degree of centrality, which is the number of edges incident to a node (it can be in-degree or out-degree if it’s a direct graph); the heard behavior occurs when individuals observe the actions of all (or most) other individuals and act in a form aligned with them; we can also calculate the normalized degree of centrality; other centrality metrics are the closeness centrality (length of shortest paths from a vertex), the betweenness centrality (number of shortest paths a vertex is part of) and the delta centrality (related to the “performance” of the network to the removal of a certain element; it can also be related to the concept of efficiency, as the inverse of the distance of two nodes, the more it grows the worse the performance is) (with normalized versions for closeness and betweenness centralities).

**Segmentation metrics**:   
- **Counting of**:  
• **Cliques**, if each individual is directly related to each other individual;  
• "**Social** **circles**", groups of individuals less closely linked than in a clique.  
- **Clustering** (or aggregation) **coefficient**.  
- **Cohesion**: degree to which actors are directly connected to each other by cohesive bonds (minimum number of actors or ties of a social network that must be removed to disconnect the group; or also called connectivity).

**Neo4j Graph Platform** supports transactional processing and analytical processing of graph data. It has algorithms like: **Shortest Path Algorithms, Centrality Algorithms** (degree centrality, closeness centrality, betweenness centrality, page rank), **Community Detection Algorithms** (clustering coefficients, strongly connected components, label propagation, modularity).

# Community detection

**Community:** larger, more significant, and functionally meaningful divisions within a network; typically composed of nodes that share common characteristics, interests, functions, or roles within the network. They have:   
- **High Internal Connectivity**: Nodes within a community are strongly connected to each other.  
- **Low External Connectivity**: Nodes in a community have fewer connections to nodes outside of their community.

They have:  
- **Modular Structure**: Complex networks often exhibit a modular or hierarchical structure; multiple levels of communities within communities (can reveal insights into the organization of the network).   
- **Functional Significance**: may represent groups of nodes that have a similar function, role, or purpose within the network (group of friends with common interests).

Two types of communities:  
- **Observable communities**: explicitly visible or easily identifiable based on the available data and interactions within a social network (can be observed using data like user connections, interactions, affiliations, friendship, hashtag…).  
- **Latent communities**: not explicitly defined by user actions or interactions but are inferred through data analysis and machine learning techniques (discovered using algorithms, like network analysis, clustering and topic modeling, to find patterns, similarities…).

Scale-free network: type of complex network characterized by a specific degree distribution; few nodes have a significantly higher degree than majority of nodes (can lead to power-law degree distribution, where some nodes have most connections while the rest have only a few connections).

Preferential attachment: explains the formation of scale-free networks; new nodes are more likely to connect to existing nodes that already have high degrees (“rich get richer”). Can be popularity-based, quality-based or mixed.

Assortativity: tendency of nodes in a network to connect with other nodes with similar characteristics/properties; measure of homophily, it quantifies also the correlation between attributes/degrees of connected nodes. The concept can also be applied to other characteristics of a node: weight, betweenness, kth level node degree… A network can be considered assortative (high degree nodes are connected on average with high degree nodes, and lower to lower) or disassortative (low degree nodes connected to high degree nodes, and high to low; hubs avoid); the same goes for assortative (similar nodes connect) and disassortative communities (different nodes connect). Assortativity is a measure defined by the Pearson Correlation Coefficient (1 assortative, -1 disassortative, 0 neutral). Hub removal makes more damage with assortative networks than with disassortative networks.

Giant component: in a complex network, the largest connected subgraph. Communities within a giant component are more connected, and usually more important to. Communities within a small component can be niche areas of interests (sometimes incredibly important too).

Network percolation: modeling and analysis technique to study network changes as nodes or links. It focuses on the critical point - or percolation threshold (can be considered as the point at which the giant component emerges) - at which a network transitions from a connected state to a fragmented state as nodes or links are removed based on some probabilistic rule.

Community detection: process of identification of subrgroups/communities within a network (the presence of hubs can influence the structure of communities). The level of granularity can vary a lot. Community detection can be node-centric (all calculations made based on a node), group-centric (interested in communities with group properties), network-centric (based on similarity of nodes) or hierarchy-centric (groups divided among top level with same community nodes, and bottom level with different community nodes). There are different types of algorithms:  
- hierarchical algorithms: build hierarchical structure of communities (agglomerative or divisive);   
- modularity-based algorithms: from modularity (strength of community, comparing n of edges within communities to expected number of edges with randomly distributed connections); like Louvain method, or Girvan-Newman;   
- Label-propagation algorithms: semi-supervised ML algorithms that assign labels to previously unlabeled data points  
- Random walk-based algorithms: process that begins at some vertex, and moves to another vertex at each step; like Walktrap and Infomap;  
- Graph partitioning-based algorithms: process of dividing a graph into smaller subgraphs by partitioning into mutually exclusive groups; like spectral clustering, Kernighan-Lin algorithm, FluidC or METIS.   
- Hybrid algorithms.

# Sentiment Analysis

There is today an enormous amount of data generated every day on social media, forums, reviews, and other online platforms. From this mass of data, it is possible to extract emotions and sentiment expressed by people. **Numerous techniques** can be applied for this purpose, which fall into two main categories:  
- **Affective computing:** is a broader field that encompasses the design and development of systems and devices that can recognize, interpret, and respond to human emotions (like facial recognition, speech analysis…).  
- **Sentiment analysis:** also known as opinion mining, is a narrower aspect of affective computing (specifically deals with determining sentiment/emotional tone expressed in text).

It's useful for individuals (personal branding, job hunting, personal branding, news and info consumption), society (public opinion monitoring, healthcare, crisis response and emergency management) and businesses (customer feedback and satisfaction, market research and competitor analysis, brand monitoring, risk management, social media marketing).

**Emotion recognition** involves identifying and categorizing the specific emotions (happiness, sadness, anger, fear…) expressed by an individual. Different emotions “models”: Ekman (6 basic emotions), Plutchik (wheel of emotions), circumplex model of affect. **Polarity detection**: determining sentiment of text, like positive, neutral or negative.

In sentiment analysis, the term **opinion** can be used as a broad concept that covers sentiment, evaluation, appraisal, or attitude, and its associated information such as opinion target and the person who holds the opinion. The term **sentiment** can be used to mean only the underlying positive or negative feeling implied by opinion.

An opinion has **two key components:  
- a target:** can be any entity or aspect/attribute of the entity on which an opinion has been expressed.  
- a **sentiment**: is the underlying feeling (or attitude, evaluation, etc.), associated with an opinion; can be a positive, negative, or neutral sentiment, or a numeric rating.

* 1. Sentiment can be classified into **several types** (linguistic-based, psychology-based or consumer research-based), and also into:  
     - **Rational sentiments:** from rational reasoning, tangible beliefs, and utilitarian attitudes; no emotions expressed  
     - **Emotional sentiments**: from non tangible and emotional responses to entities which go deep into people’s psychological state of mind; emotional sentiment is stronger than rational sentiment and is usually more important in practice.

Sentiment orientation is also called **polarity**, semantic orientation, or **valence** in the research literature. It can be **positive**, **negative**, or **neutral**. Sentiment can have different levels of **strength** or **intensity.** People often use two ways to express intensity of their feelings in text: sentiment terms (words or phrases with suitable strengths), or intensifiers and diminishers (terms that change the degree of the expressed sentiment). In applications, we commonly use some **discrete ratings** to express sentiment intensity.

An **opinion summary** is just like a multi-document summary because we need to summarize multiple opinion documents, e.g., reviews. The core form of opinion summary is called the **aspect-based opinion summary**.

Sentiment analysis tasks:   
- polarity classification: whether the expressed opinion in a document or a sentence, on a certain feature or aspect of a target, is positive, negative, or neutral; on unigrams, with feature selection;  
- valence and arousal: valence refers to positivity/negativity of an emotion; arousal refers to the level of excitement/calmness related to an emotion; VAD, or Valence-Arousal-Dominance model;  
- subjectivity or objectivity: allocating text into subjective/objective; sometimes very difficult;  
- beyond polarity: fine opinion analysis;  
- feature/aspect based: aspect/feature extraction, like finding relevant entities.

Challenges in SA: usually because of ambiguity in natural language (syntactic, semantic, pragmatic or emotional ambiguity)

Approaches to SA:  
- Knowledge-based techniques: accessible and economic; based on lexicons to perform emotion analysis and sentiment analysis; BUT they have poor recognition of affect, and limits in knowledge representation.   
- Supervised methods: ML and deep learning, popular and very use, they need large training corpus;   
- Hybrid approaches: both knowledge-based techniques and supervised methods

SA Lexicons/datasets:  
- WordNet: lexical database of the English language that groups words into sets of synonyms; it links words into semantic relations, including synonyms(words that have similar meanings), hypernyms(words that represent more abstract or general concepts), hyponyms(words that represent more specific concepts), and meronyms(words that denote a part of something);  
- Sentiment lexicons: either manually-generated or automatically-generated, they are collections of words or phrases that are annotated with their associated sentiment or emotion.   
- Manually generated sentiment lexicons: like AFINN (with scoring system, and limited to unigrams, it scores words from -5 to +5) or General Inquirer (GI)  
- Automatically-generated sentiment lexicons: SentiWordNet (fuzzy labelling, with 3 sentiment scores, sentiment more associated with meaning of a word rather than the word itself), SenticNet, VADER (Valence Aware Dictionary and sEntiment Reasoner, it scores words with intensity and polarity, it’s rules-based and widely used on social media)  
- Emotion lexicons: they provide information about the emotional or affective meaning of words, helping in the analysis of sentiment, mood, or emotions expressed in written or spoken language.   
- LIWC: Linguistic Inquiry and Word Count, is a text analysis program that analyzes written or spoken language based on linguistic and psychological dimensions; with customizable dictionariess, it has applications in different contexts;  
- WordNet-Affect: was created using a semi-supervised method, it’s a resource that annotates senses in WordNet with emotions.   
- EmoLex: "Emotion Lexicon", is a lexicon or dictionary of words that are annotated with their associated emotional content.   
- SenticNet: publicly available resource for sentiment analysis and concept-level opinion mining.

# Online interactions and privacy issues

Data about users is every day collected, analyzed, and possibly shared and disseminated, with and without user consent. Data from multiple sources may be linked 🡪 non-identifiable information can become identifiable when linked.

Privacy laws and regulations vary widely throughout the world. GDPR (General Data Protection Regulation) adopted Apr 2016, start May 2018. The EU GDPR applies to all parties (even non-European) that process data of EU citizens. Key definitions (cont.): anonymous data (data for which the related subject cannot be identified directly or indirectly, by any means or by any person), pseudonymous data (data for which the related subject cannot be identified directly or indirectly without a key).

Users can know that their personal information is collected, but lose control on how and to whom the information is disclosed. Often users do not even know that their personal information is collected. 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; referral link field can also reveal personal information.

Every time a browser downloads a page on the Web, this event may be recorded in the log files of the remote Web server, which can keep the control on the log file and register. Privacy can be enhanced by using:  
- Proxy servers (act as intermediaries between a client and a server; the client connects to the proxy requesting a service available from server X; the proxy sends the request to X and sends back the service to the client; X logs the request as coming from the proxy);  
- Anonymizers (proxies specifically designed to maintain their accesses on the web anonymous

The IP address is visible to web sites; this may reveal users’ employers, ISPs and enable linking activities across sites over time; a web anonymizer hides IP addresses. A cookie is a small file created by an Internet site to store information on a user’s computer. Cookies allow websites to identify a user/device that accesses them and her/its preferences, such as language and other identifiable information. Cookies can be useful (staple to attach multiple parts of a form together, recognize when a user returns to a web site, to help web sites understand how people use them, to manage sessions). Cookies can be used improperly (profiling users and tracking their activities, especially across web sites).

Spyware: Malware (malicious software) that collects information about users without their knowledge or explicit permission. Can collect almost any type of data (e.g., personal information, user logins, bank or credit card information) and can install additional software or redirecting Web browsers or change computer settings.

Facebook counts more than 2.9 billion monthly active users in the world as of March 2022. The value of Facebook is strictly linked to the amount of information it holds!

# NER, Linking and Disambiguation

* 1. **Named Entity Recognition (NER):** task to identify and classify entities, such as names of people, organizations, locations, dates, and other predefined categories, within a given text. The primary goal of NER is to provide structured information about the entities present in a text.

**Named Entity Linking (NEL)**: linking entities to a specific entry or entity in a knowledge base or reference database. The goal is to associate each recognized entity in the text with a unique identifier in a knowledge base (or reference database), allowing for additional information retrieval or disambiguation. **Identifiers** play a crucial role in linking entity mentions in text to specific entries in a knowledge base or reference database. NEL identifiers characteristics are: uniqueness (identifiers unique within context of the knowledge), stability (identifiers stable over time to ensure persistence of links between entity mentions), accessibility (identifiers should be accessible and retrievable).

**Named Entity Disambiguation (NED)**: determining the specific entity (e.g., person, location, organization) that a mention or reference in the text corresponds to. Particularly important in cases where the same term could refer to multiple entities. Disambiguation helps to identify the correct entity based on the context.

In summary, **NER** involves identifying and classifying entities in a text, while **NED** deals with resolving ambiguity when a name or term could refer to multiple entities by determining the correct entity based on context. While **NEL** associates the recognized entity with a unique identifier in a knowledge base, **NED** focuses on resolving ambiguity by determining the correct entity in a given context. All of these are challenging (ambiguity, polysemy, short and noisy nature, typographic errors, shortening of words, context variability, Out of Vocabulary, User-generated entities…)

NER: **Information Extraction (IE)** task that seeks to segment and classify text fragments into predefined classes/labels. NLP task that involves automatically extracting structured information from unstructured text. The goal is to transform raw text into a more organized and structured format, making it easier to analyze and use. Information extraction typically involves identifying and extracting specific types of information, such as entities, relationships, and events, from text documents.

The **Sequence Prediction Problem (SPP)** refers to a type of machine learning task where the goal is to predict the next element in a sequence based on the patterns and relationships observed in the input data. Ex: POS (Part-of-Speech) tagging.

**Conditional Random Fields (CRFs)** are a type of probabilistic graphical model used for modeling sequential data, and they have been particularly effective in the context of Named Entity Recognition (NER). CRFs model the conditional probability of a sequence of labels given the input sequence. The model is trained to maximize the likelihood of the correct sequence labels given the input data.

Both HMMs (Hidden Markov Models) and CRFs are types of probabilistic graphical models used for Sequence Prediction tasks, but they have different structures and modeling assumptions. In an HMM, a sequence of observations is assumed to be generated by an underlying sequence of hidden states. CRF have been widely used in NER tasks due to their ability to model complex dependencies in sequential data. However, more recently, **deep learning approaches**, such as Recurrent Neural Networks (RNNs) and Transformer models, have also gained popularity in NER tasks, often outperforming traditional CRF-based models.

To evaluate NER (challenging without a ground truth…): precision, recall, F-score…