Information Fusion 63 (2020) 88–120

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

The four dimensions of social network analysis: An overview of research methods, applications, and software tools

David Camacho[a,*], Ángel Panizo-LLedot[a], Gema Bello-Orgaz[a], Antonio Gonzalez-Pardo[b], Erik Cambria[c]

[a] *Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, Spain*
[b] *Computer Science Department, Universidad Rey Juan Carlos, Spain*
[c] *School of Computer Science and Engineering, Nanyang Technological University, Singapore*

ARTICLE INFO

Keywords:
Social network analysis
Social media mining
Social data visualization
Data science
Big data

ABSTRACT

Social network based applications have experienced exponential growth in recent years. One of the reasons for this rise is that this application domain offers a particularly fertile place to test and develop the most advanced computational techniques to extract valuable information from the Web. The main contribution of this work is three-fold: (1) we provide an up-to-date literature review of the state of the art on social network analysis (SNA); (2) we propose a set of new metrics based on four essential features (or *dimensions*) in SNA; (3) finally, we provide a quantitative analysis of a set of popular SNA tools and frameworks. We have also performed a scientometric study to detect the most active research areas and application domains in this area. This work proposes the definition of four different dimensions, namely *Pattern & Knowledge discovery, Information Fusion & Integration, Scalability*, and *Visualization*, which are used to define a set of new metrics (termed *degrees*) in order to evaluate the different software tools and frameworks of SNA (a set of 20 SNA-software tools are analyzed and ranked following previous metrics). These dimensions, together with the defined degrees, allow evaluating and measure the maturity of social network technologies, looking for both a quantitative assessment of them, as to shed light to the challenges and future trends in this active area.

1. Introduction

Currently, online social networks (OSNs) are seen as an essential element for interpersonal relationships in a large part of the world. OSNs allow the elimination of physical and cultural barriers through the globalization of the technology. For this reason, OSNs have billions of active users around the world. OSNs can be defined as a social structure made up of people, or entities, connected by some type of relationship or common interest (professional relationship, friendship, kinship, etc.). From [1], an OSN can be defined as: "a service that allow individuals to (1) define a public (or semi-public) profile within an application or specific domain (friendship, professional, common interests, etc.), (2) manage a list of other users with whom the individual (or entity) will share a connection, and (3) view and traverse their list of connections and those made by others within the social site". Although the origins correspond mainly to different areas from Social Sciences, the term is attributed to the British anthropologists Alfred Radcliffe-Brown [2,3] and John Barnes [4].

The fast growth of OSN sites has led to an enormous interest in the analysis of this type of networks (the interconnections that originate, their structure, the evolution of the network, the information flow and dissemination, or the patterns that can be extracted from them, among many others). The easy access to this type of information, the availability of vast amounts of data, the simple and straightforward codification in form of graph-based representation, as well as the direct application of any practical results drawn from them, has made OSNs one of the hot research areas in several disciplines as Data Mining [5], Big Data [6], Machine Learning [7], Information Visualization [8], or Complex systems [9], among many others. However, the exponential growth of social media (around 3.484 billion of active social media users, up 9% year-on-year, connected to different OSNs [10]) has caused serious problems for traditional data analysis algorithms and methods (such as data mining, statistics or machine learning) [11,12]. The aforementioned areas have to face the challenge of designing and implementing new methods able to work efficiently with the huge amount of data generated in the OSN.

[*] Corresponding author.
E-mail addresses: david.camacho@upm.es (D. Camacho), angel.panizo@upm.es (Á. Panizo-LLedot), gema.borgaz@upm.es (G. Bello-Orgaz), antonio.gpardo@urjc.es (A. Gonzalez-Pardo), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.inffus.2020.05.009
Received 4 February 2020; Received in revised form 28 May 2020; Accepted 31 May 2020
Available online 3 June 2020

This exciting area generates thousands of papers per year, hundred of different algorithms, tools, and frameworks, to tackle the challenges and open issues related to OSNs. Therefore, when anyone tries to develop a comprehensive analysis of the state of the art in such a complex and multidisciplinary area, it is necessary to establish a formal method to allow the analysis of the immense amount of information available. To do that, a scientometric analysis has been carried out (Section 2), which has led us to organize the research process of this paper. This process can be summarized as follows:

1. We have performed a scientometric study over the papers published in the last 5 years to extract both, the most active research areas and the most relevant application domains in SNA. The research areas selected have been: graph theory and network analytics, community detection algorithms, information diffusion models, user profiling, topic extraction, and finally sentiment analysis area. The selected application domains have been: Healthcare, Marketing, Tourism & Hospitality, and Cybersecurity. This analysis has allowed us also to find a set of emerging areas, and we have selected and studied the areas of Politics, detection of fake news, and Multimedia in OSNs. These emerging areas have been selected due to their current research activity level, and their high potential impact in the next years.
2. Once the SNA application domains and fundamental research areas have been selected, we have analyzed in detail the most relevant research works published in the last years in the corresponding research fields and application domains.
3. Due to the complexity of the field, and the increasing amount of available technologies for OSN, we have defined a set of metrics, which allow any researcher to assess any SNA framework, tool or algorithm. These metrics are based on some key aspects, or features, that are relevant for any algorithm belonging to SNA research field (such as knowledge discovery, information fusion or visualization among others).
4. Finally, we have assessed some relevant frameworks and tools available on the Internet to perform SNA tasks by using the metrics previously mentioned. This assessment allows us to rate the degree of maturity of SNA technologies, so any engineer or researcher can better understand the strengths and weaknesses of the different tools and frameworks currently available.

The main contributions of this paper can be summarized as follows:

1. A detailed review of the current state of the art of a set of highly relevant research works, grouped in different categories depending on the fundamental research area addressed, and the application domain.
2. The definition of four new "SNA-*Dimensions*" (*Pattern & Knowledge discovery, Information Fusion & Integration, Scalability*, and *Visualization*). These dimensions are inspired by the popular V-models [13] used in the Big Data area, and its goal is to measure the capacity of the different frameworks and tools to perform SNA tasks. They try to answer the following Research Questions:
   - RQ1) **What can I discover?** (*Pattern & Knowledge discovery*) [11,12]: related to the capacity of algorithms, methods and techniques to gather knowledge (usually complex and non-trivial patterns) from OSNs.
   - RQ2) **What is the limit?** (*Scalability*): [14,15]: based on the capacity of algorithms, methods and frameworks to work with large amounts of data. This includes both computational time and volume of data.
   - RQ3) **What kind of data can I integrate?** (*Information Fusion & Integration*) [16,17]: This dimension will be related to the capacity of fusing different kinds of data (text, video, images, audio) and from different sources (Social Media Platforms).
   - RQ4) **What can I show?** (*Visualization*) [18,19]: related to the capability to visualize, filter and represent adequately the information stored in a network.
3. The definition of a new *global Capability metric*, named $\mathfrak{C}_{SNA}$, based on the just mentioned metrics that can be used to rank the capabilities of the SNA technology, framework or tool analyzed.
4. The assessment of 20 of the most relevant frameworks and tools to perform SNA tasks using these different dimensions and the global Capability metric defined. This analysis not only shows what are the main strengths and weaknesses of the different frameworks, but it also provides a useful guide for those beginners or senior researchers in the area of SNA.

To help readers through the contents of this paper, Fig. 1 shows the overall organization of this work. Using this figure, readers can easily understand the main contents of each section and go directly to those contents that can be more relevant for their future work.

This paper is structured as follows: Section 2 contains the scientometric analysis performed to highlight the different research areas with high impact in SNA research field. Section 3 provides an analysis of the different concepts and research works that belongs to the SNA research field from the Computer Science point-of-view. The different application domains in SNA, as well as the emerging areas, are analyzed in Section 4. Section 5 provides a discussion and some conclusions from the state of the art analyzed in sections 3 and 4. Section 6 contains the definition of the four dimensions and their related metrics (or degrees), which are used to compare the different SNA tools in Section 7. Finally, some challenges, future trends in this area, and the main conclusions of this work are given in Section 8.

2. SNA Scientometric analysis

Although bibliometrics emerged to support the daily work of librarians, nowadays it is used to evaluate the scientific achievements of people and institutions [20]. The main goal of bibliometrics is the quantification of scientific production, measuring the performance of institutions and people. This section presents a scientometric analysis of the research papers published in SNA, which has been carried out to detect and select those relevant areas that will be analyzed. The study has been done by using the `Meta-knowledge Python` package [21], which accepts raw data from the Web of Science, Scopus, PubMed, ProQuest Dissertations and Theses, and select funding agencies as NSF (United States), or NSERC (Canada), among others. The output of this package is a set of characteristics for quantitative analysis, including Time Series methods, Standard and Multi Reference Publication Year Spectroscopy (RPYS), computational text analysis, and network analysis.

In particular, this analysis presents a review of works related to SNA using Web of Science as a search engine, covering the highly cited articles over a five years period (from 2014 to 2018) and resulting in a record collection of **28.805 articles**. It has only been considered in the scienciometric analysis those authors that have used in their publications the keyword of *"social network analysis (SNA)"*. Although this analysis could discard some relevant authors or publications, which did not use the previous term, this keyword was used to restrict the set of papers to analyze and to obtain some of the highly relevant application domains and research in SNA.

Firstly, to better understand the evolution in the area of SNA in the last sixty years (since the 70s), RPYS has been used. This method was proposed by Marx et al. [22], and it is a method for quantifying the impact of historical publications on research fields. Standard RPYS [23] analyzes the cited references and especially the referenced publication years of a publication collection. In the first step, all the references from the publications are selected (from a particular research field and period). Then the 5-year median deviations, to the number of cited references from each publication year, are computed to generate a *spectrogram*. The peaks in the spectrogram (deviations from the median) indicate those specific years with highly cited publications within the domain of the sample. Also, Multi RPYS is an extension of the standard method [23]. It segments the original citing articles based on their

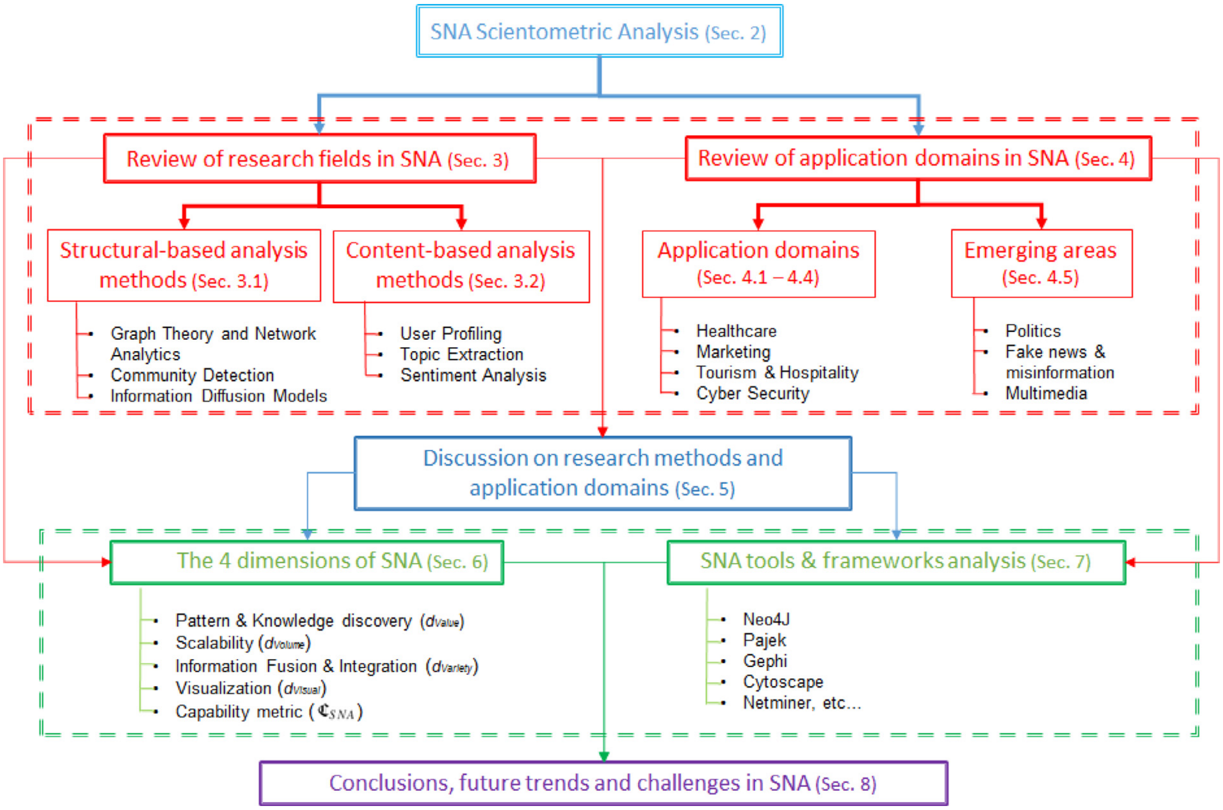

SNA Scientometric Analysis (Sec. 2)
Review of research fields in SNA (Sec. 3)
Review of application domains in SNA (Sec. 4)
Structural-based analysis methods (Sec. 3.1)
Content-based analysis methods (Sec. 3.2)
Application domains (Sec. 4.1 – 4.4)
Emerging areas (Sec. 4.5)
Graph Theory and Network Analytics
Community Detection
Information Diffusion Models
User Profiling
Topic Extraction
Sentiment Analysis
Healthcare
Marketing
Tourism & Hospitality
Cyber Security
Politics
Fake news & misinformation
Multimedia
Discussion on research methods and application domains (Sec. 5)
The 4 dimensions of SNA (Sec. 6)
SNA tools & frameworks analysis (Sec. 7)
Pattern & Knowledge discovery (d_{Value})
Scalability (d_{Volume})
Information Fusion & Integration ($d_{Variety}$)
Visualization (d_{Visual})
Capability metric ($\mathfrak{C}_{SNA}$)
Neo4J
Pajek
Gephi
Cytoscape
Netminer, etc…
Conclusions, future trends and challenges in SNA (Sec. 8)

Fig. 1. Structure of the paper and the main contents of different sections. Red boxes show the sections related to the review on both the current state of the art in SNA research, and their application domains, whereas green boxes show the contributions on the four dimensions defined, and the software tools and frameworks assessment carried out. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

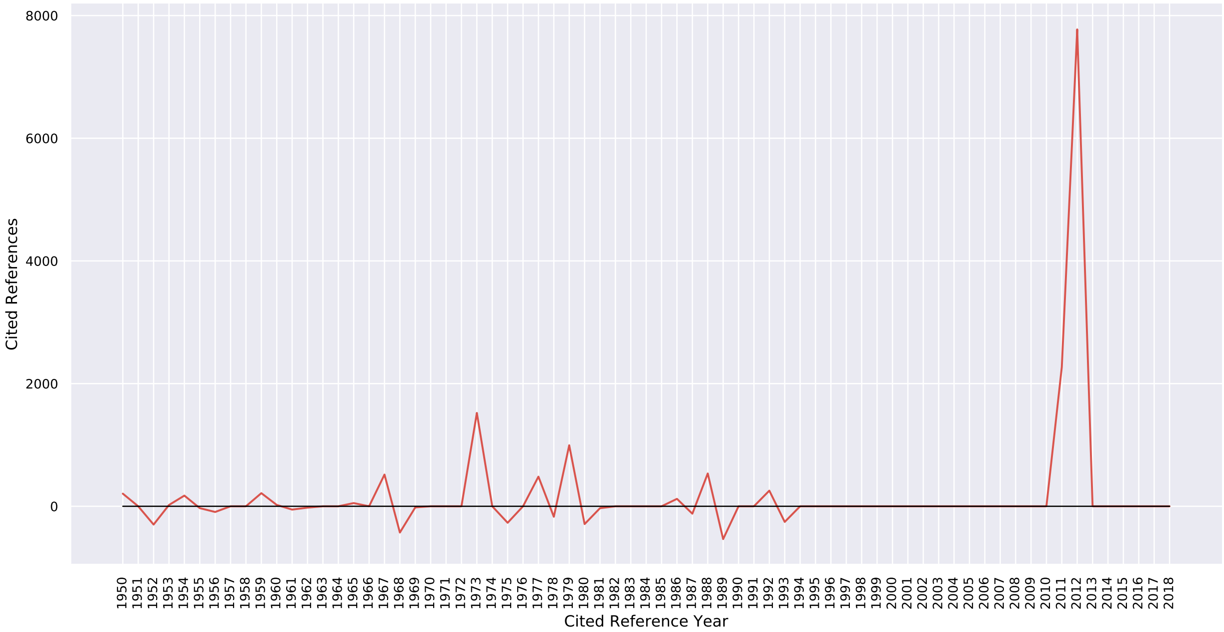


Fig. 2. Standard RPYS related to SNA. Pronounced peaks represent years where citations to published books or articles deviate from a 5-year median.

publication years and conducts a Standard RPYS analyzes for each one, visualizing the results as a heat map. Therefore, this method is useful for differentiating between historical publications that have a lasting impact, versus those that are influential only within a short time frame.

As shown in Fig. 2, the years with pronounced peaks for SNA are: 1967 [24], 1973 [25], 1977 [26], 1979 [27], 1988 [28], 2011 [29] and 2012 [30]. Analyzing the most cited papers of these specific years shows that in the 70s and 80s the articles with higher impact were published in journals on the area of Social Science [24–28], while recent high impact articles are published in journals belonging to Computer Science [29,30]. It can also be observed that from 1994 to 2010 there was a period of stability where no high relevant works emerged. Fig. 3, where

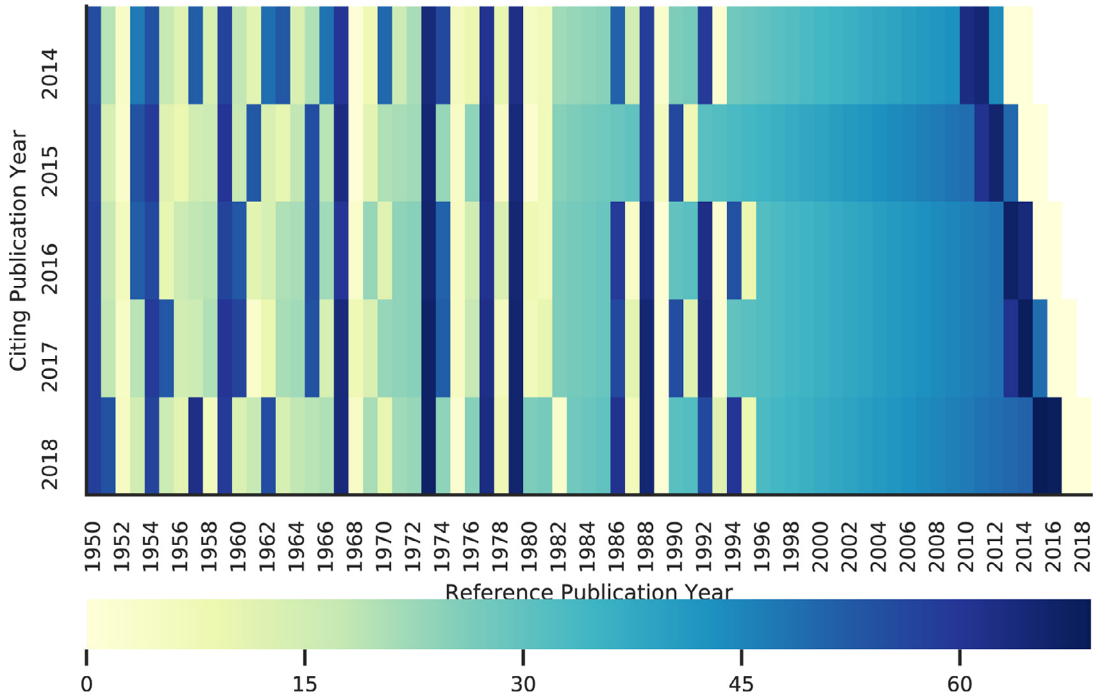


Fig. 3. Heatmap showing the results of a Multi RPYS analysis. Darker bands across both periods indicate lasting influence.

the Multi RPYS analysis is shown, reinforces this pattern. Between 1996 and 2010, there are publications for all the years analyzed that are quite quoted, but no publication stands out from the rest.

From this long period analysis, it can be concluded that most of the papers published in the 70s and 80s were focused on defining the basic principles of the SNA such as the principle of diffusion of influence and information, mobility opportunity, community organization, or links joining people who know each other [24,25,28]. Also, published works at these years introduced a broad background for measures of structural centrality in OSN [26,27]. However, since 2000, the main contributions have been focused on the application of these principles to specific domains, using OSN data to extract new knowledge, which improves the performance of the organizations. These articles are mainly focused on the design of new methods, algorithms and applications for SNA, by taking into account the **structure** of the networks [31,32] and detecting important structures, as communities, inside them [33,34]. As it has been already seen in the RPYS analysis, the oldest publications that appear are related to the definition of principles and measures to model the dynamics of the OSNs. Whereas, the most recent publications are focused on the development of methods and techniques based on the concepts previously published. For example, it should be noted that community detection algorithms are those that prevail in the last two decades, such as Girvan and Newman [33,35], edge betweenness centrality [36], fast Greedy [37,38], Cfinder [39,40], Walktrap [41,42], structural algorithm by Rosvall et al. [43,44], clique percolation method [45], fast modularity optimization by Blondel et al. [34], Louvain algorighm [46], community embeddings [47], etc.

To identify the most relevant topics related to SNA, textual analysis has been performed using the collection of articles gathered from the period of the last 5 years studied. In this case, the latent Dirichlet allocation (LDA) model has been applied to detect the top 20 topics (i.e sets of terms that frequently appear together), by processing the ”keywords” used in the article collection, to later visualize the most frequent terms, found in these topics, as a word cloud (see Fig. 4).

Analyzing the most relevant keywords related to SNA, it can be seen at first sight that many of the relevant keywords are related to the key topic of *Pattern & Knowledge Discovery* are: Model, behavior, Patterns, Classification, Prediction, Knowledge, Regression, Recognition, Dynamics, Groups, Diffusion, or Sentiment, among others. Also, there are several relevant application domains where these methods of knowledge extraction are applied as Healthcare (HIV, Disorder, Brain, Cortex, Drug, Sex, Physical and Psychological), Education, Governance, Policy, Mouth (Mouth-to-Mouth), Business, Food, Culture, or Games. Regarding the topic of *Information Fusion*, there are several keywords related to different data sources and types of information that are used in the area such as Internet, Media, Facebook, Web, Graph, Factors, Activity or Context. The rest of the keywords would be classified taking into account the topic of *Scalability*, where the terms Performance, Complex and Sustainable keywords are directly related to this feature. And finally, related to the topic of *Visualization*, appears some keywords such as Perspective, States, or Flow.

Although the initial study covers the last six decades, a detailed analysis of the last five years has been done to understand which areas and application are more relevant in terms of high impact. From this analysis, it can be concluded that currently some of the hot research areas, from Computer Science, in this topic are: Data Science and Big Data, more specifically Network analysis, Social Media, Sentiment analysis, Text Mining, and Information diffusion. Whereas some of the hot application areas are: Health, Marketing and Business, and Tourism. It is particularly important to point out that a large number of published works are focused on solving problems (through the design of specific algorithms and techniques, such as those related to community finding problems, or information diffusion models), which are related to the problems of pattern mining and knowledge discovery, how to fusion or integrate information, how to visualize adequately the information, or how to handle huge amounts of data.

Finally, and although the scientometric analysis has not directly shed light on some areas such as politics (e.g., hate speech detection, political polarization), fake news and misinformation (e.g., fake news analysis in SN), or cybersecurity (e.g., cyber intelligence, cybercrime, and cyber terrorism), during our analysis of the state of the art we have detected an increasing research activity in these areas (see Section 4.5 for a further description), so they will be analyzed as emerging areas due to its high potential in the near future.

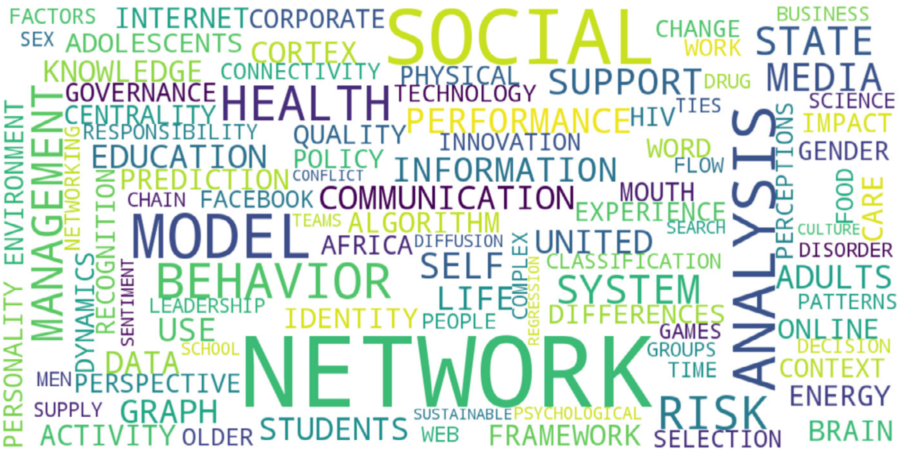


Fig. 4. World Cloud visualization of the most frequent terms that appear in the top 20 of extracted topics processing the Keywords of the papers.

Previous findings related to the research and application domains in SNA have guided the final decision on the selection of what fields will be finally analyzed in this paper. Specifically, the articles selected to guide the realization of the sections 3 and 4, that show a review of the state of the art in the research areas and application domains of SNA, have been those identified as most relevant in this scientometric analysis. Therefore, to carry out these sections have been taken into account both the most cited articles within the original collection of articles of the 5 years analysed (categorized by the most relevant topics identified) and the most cited papers by them identifying using the RPYS method that correspond to historical publications with a high impact in the area.

3. Techniques and algorithms

As it has been shown in the previous section, the area of SNA comprises a wide variety of multidisciplinary fields, ranging from Sciences and Engineering to Social Sciences. This section is focused on current findings regarding the design and development of new algorithms to process, extract and visualize (huge amounts of) knowledge. These three activities (*process, extract* and *visualize*) are closely associated to the *Data Science* and the *Big Data* fields. The former can be defined generically, as a multi-disciplinary field that uses scientific methods, techniques and algorithms, to extract useful knowledge from structured and unstructured data. Whereas the latter uses different methods and techniques to analyze, and systematically extract information from data sets that are too large, or complex, to be dealt with traditional data-processing algorithms. From the data engineering perspective, an OSN can be analyzed from two different points of views [5]: (1) *structural data* that represents the connections, interactions (linkage-based), and the topology of the network [48]; or as (2) *content data* that is focused on the information contained and shared in the OSN by the users [49]. The isolated analysis of one of these two types of data would provide an incomplete vision of the information stored in the network, so the underlying patterns and knowledge could be lost.

3.1. Structural-based analysis

In mathematics and computer science social networks are usually represented as graphs, a data structure that allows describing the properties of the social network through nodes connected by edges, where the nodes would be the individuals and the edges would be the relationships that joint them. Graph theory [50,51] is the area that study this data structure and has done important contributions to research in data analysis. Through this theory, it is possible to generalize and analyze the existing social interactions between users, and what is their behavior within their contact network [52–54]. Graph Theory methods and techniques have been applied in large network-based problems as OSNs and social media applications (i.e., Facebook, Instagram, Twitter, WhatsApp, LinkedIn, Snapchat, Youtube, Tumblr, Pinterest, Skype, or WeChat to mention just a few) [55,56]. OSNs are considered complex networks [57] because they present non-trivial topological features. In other words, the connection patterns between nodes will not be random or purely regular [58]. Due to the number of techniques and algorithms used in the SNA is extremely large, we have selected a subset of them by the scientometric study carried out, which showed that the most relevant publications in recent years are focused mainly on the algorithms for community detection, measuring the structural centrality and the application of information diffusion models. For a more detailed review of the SNA field please refer to [48,59].

3.1.1. Network metrics

We proceed with the definition of some basic metrics, or measures, that are used by graph algorithms for SNA.

- **Centrality**. Centrality is one of the essential metrics in graph and network theory, this metric is used to asses the relevance, or structural importance, of a node in the network. The centrality measure defines how important a node is in a network. In OSNs, this measure can be used to detect or identify, the most influential people in the network. When centrality is assessed, several measures are used: (1) *Degree Centrality*, which ranks nodes with more connections higher in terms of centrality; (2) *Eigenvector Centrality* tries to generalize the degree centrality by incorporating the importance of the neighbors (in directed graphs it can be used incoming or outgoing neighbors); (3) *PageRank* measure takes into account the value of passed centrality by the number of outgoing links (outdegree) from that node, so this measure gets a fraction of the centrality values of the nodes connected to the node from the source node considered; (4) *Betweenness Centrality* computes the number of shortest paths that traverse any two nodes in the graph. Nodes with higher betweenness values can be seen as “bridges” between different subgraphs, this measure is used by some specific community finding algorithms; (5) *Closeness Centrality*: the idea of this measure is that the more central the nodes are, the easier will be to reach other nodes. Therefore, the smaller the average shortest path length is, the centrality value of the node will be higher; (6) *K-shells Centrality* [60] measurement that came from studies of diffusion models and decomposes the network into many shells to identify super-spreaders, nodes with high capacity of spreading information, based on the assumption that nodes in the same shell have similar influence; and (7) *Group Centralities* such as Group Degree Centrality, Group Betweenness Centrality, Group Closeness Centrality that generalize the previous centrality-based measures to a group of nodes.
- **Transitivity** and **reciprocity** are used to represent *linking behavior* in a network. Transitivity analyzes the linking behavior to determine

whether it demonstrates a transitive behavior between three nodes, so at least three edges will be needed to create a triangle. Higher transitivity on a graph results in a denser graph, which in turn is closer to a complete graph. Therefore, it is possible to determine how close graphs are to the complete graph by measuring the transitivity. This can be performed by measuring the *[global] clustering coefficient* and *local clustering coefficient*. The former is computed for the network, whereas the latter is computed for a node.

- **Balance** and **status**. A signed graph is a graph in which each edge has a positive or negative sign. This sign is used in OSNs to represent interpersonal relationships (e.g., such as friends or foes, boss or subordinate, social status). This kind of graph is *balanced* if the product of edge signs around every cycle is positive. In real-world social networks, we expect some level of consistency concerning these interactions. For example, it is more plausible for a friend of one's friend to be a friend than to be an enemy. In signed graphs, this consistency translates to observe cycles (triads, triangles) with three positive edges (i.e., all friends) more frequently than the ones with two positive edges and one negative edge (i.e., a friend's friend is an enemy). *Social balance* and *social status* are used to determine the consistency in signed networks. Social balance theory says that friend/foe relationships are consistent when the transitivity between nodes can be propagated, such as "The friend of my friend is my friend". Social status theory measures how consistent individuals are in assigning status to their neighbors. The idea is simple, and it can be summarized that if a person X has a higher status than Y, and this last person has a higher status than Z, then X should have a higher status than Z. In a signed (directed) graph representation, positive and negative signs will be used to show higher or lower status depending on the arrow direction (if the edge has a positive value, it will mean than outgoing node has a higher status than incoming node, if the edge has a negative value, it will represent the opposite).
- **Assortativity** or **social similarity**. In social networks connections between individuals are not random, a connection between two similar individuals is more likely than between two dissimilar ones. This similarity can be manifested as similar hobbies, language, behavior, or nationality among others. Therefore, measuring assortativity in OSN helps one to better understand user interactions. Many forces creates assortativity in OSN, among them *Homophily* and *Influence* are the most common ones [59]. Nevertheless, *Homophily* and *Influence* are two sides of the same coin, while the latter is the force that an individual (influencer) exerts to other individuals so they became similar to him, the former is the force that makes two already similar individuals connect.

3.1.2. Community detection

Community detection problem (CDP) can be defined as the division of the graph into clusters of nodes based on the network structure [61]. The main idea behind CDP is that nodes belonging to the same cluster are strongly interconnected whereas they maintain sparse connections to the nodes of other clusters. This problem is similar to the idea of graph partitioning into groups of nodes according to the network topology, where a partition is a division of the graph and it can be easily mapped into a cluster [38,62–64].

To detect the communities, or clusters, on a graph, there is a wide range of techniques such as random walks, spectral clustering, modularity maximization, or statistical approaches [62]. This kind of algorithms uses the topology of the graph to create the partitions that are validated by taking into account the density of the resulting sub-graph (i.e., a subgraph is highly connected), and connections from these nodes to the rest. A good community is the one whose nodes are highly connected and it has few connections to the nodes of other communities [65].

A simple taxonomy, which can be done to distinguish the different algorithms and methods to detect communities, is the one that takes into account how the variable '*time*' is integrated into the model. In this sense, it is possible to talk about *static* and *dynamic* community finding algorithms.

- *Static community finding algorithms* refers to those algorithms and approaches that do not take into account the evolution of the OSN, i.e., the variable '*time*' is not modeled into the system. The algorithms previously described are static, which means that the network is modeled into a single snapshot that contains all the information regarding the OSN. These algorithms present some advantages and drawbacks. On the one hand, static community finding algorithms are quite easy to apply to any problem, because no changes will occur in the network during the algorithm execution. On the other hand, and as one of the drawbacks, this kind of non-temporal models makes that the results may not be very representative, because OSNs are continuously changing due to the high variability in the number of users (nodes) and interactions (edges) between them.
  The different algorithms that belong to the static community finding problem can be grouped in four different categories depending on the scope of the corresponding algorithm. In this sense, there are node-centric, group-centric, network-centric and hierarchy-centric algorithms [66]:
  - Node-centric community finding methods: in this case, each node of the network must satisfy the properties of mutuality, reachability and degrees.
    - *Mutuality property* relies on the concept of the clique, which is the maximal complete subgraph of three or more nodes in such a way all of them are adjacent to each other. Finding the maximum clique in a network is an NP-hard problem, for this reason, it is quite popular to develop algorithms able to find approximate solutions. One of these algorithms is the one proposed in [67], where each time a subset of the network is analyzed. In this algorithm, a greedy-search procedure is executed to find the different cliques in each subnetwork.
    - *Reachability property* considers that two nodes may belong to the same community if there is a path connecting them. This property assigns the nodes belonging to the same connected component to the same community. The advantage of this property is that it can be computed in $\mathcal{O}(n+m)$ time, but real-world networks are composed by a big component whereas the majority are singletons and small communities [68]. For this reason, the identification of communities in the small components is straightforward but some efforts must be done to detect the communities contained in the biggest connected components. One way to do that is by finding the *k-cliques*. A k-clique is a maximal subgraph where the largest geodesic distance between two nodes is less or equal to *k*.
    - *Nodal degree property* establishes that the nodes of a group must be adjacent to a relatively large number of group members. In this case there are two different structures studied: *k-plex* [69] and *k-core* [70].
  - Group-centric community finding methods. In this category falls all the algorithms and methods, that consider the connections inside the community as a whole. These algorithms are usually known as *density-based* algorithms and are based on the concept of γ-dense subgraphs, or quasi-clique [67].
  - Network-centric community finding methods consider all the connections of the network, instead of only the connections of the community as Group-centric methods do. In this type of algorithms, the different nodes are grouped in sets of disjoint communities according to a specific quantitative criterion.
    - *Group based on Minimum-Cut*. According to this criteria, a community is defined as a subset of nodes $C \subset V$, such that $\forall v \in C$, v has at least as many edges connecting to nodes in the same community as it does to vertices in $V \setminus C$. In [71] authors showed that the community can be found via $s-t$ minimum

cut, where s is the source node in the community and t is the sink node outside the community.

    - *Group based on Modularity*. In this case, the structure of the community is compared against a random graph, more precisely the modularity defines how likely the community structure is created at random.
    - *Group based on Latent Space Model*. The idea is to build a latent space [72,73] with the nodes of the network in such a way those nodes with dense connections are close to each other. These models assume that the interactions between nodes depend on the position of the nodes in the latent space.
  - Hierarchy-centric community finding methods try to build a hierarchical structure of communities by taking into account the structure of the network. These methods are quite similar to those from hierarchical clustering research field [74,75]
    - *Divisive hierarchical clustering* starts dividing the whole network into several disjoint sets, and then each set is split into smaller ones until all subsets contain only one node. A popular algorithm in divisive hierarchical clustering is the one based on edge betweenness [36]. In this case, the algorithm removes those edges with higher betweenness in such a way the different communities are isolated.
    - *Agglomerative hierarchical clustering* corresponds to the opposite approach. In this case, the algorithms and methods start assigning each node to one independent cluster. In each iteration, clusters are merged into a larger one according to a specific metric. The most popular metric used in this type of algorithms is modularity [37]. The idea is to merge two communities if the resulting community improves the modularity.
- *Dynamic community finding algorithms* refer to those approaches that incorporate the variable *time* into the model. Working with time broadens the range of possible analysis. When time is involved in the model, not only the community structure of the network at any given time can be analyzed, but also their dynamics, i.e., their evolution over time. Regarding community dynamics, one could be interested in analyzing the life cycle of a community, when it appears for the first time, when it grows, when it splits into several communities...etc. Likewise, one could be interested in finding communities that persist over time or dividing a network into periods where the community structure is stable. This variety of analysis, with different outputs and goals, are all encompassed inside the dynamic community finding literature.
  Embedding the 'time' variable into a network model is not a trivial task and to address this issue *Snapshot* and *Temporal Networks* models were proposed. The former introduces dynamic into a network by generating an ordered sequence of graphs, where each graph represents the state of the network at a given point in time. The latter avoid doing any aggregation at all, and represents the network as a set of timestamped nodes and edges that precisely define when an element appear and disappear from a network. Several taxonomies have been proposed in the literature to categorize dynamic community finding methods according to these models:
  - Snapshots-based dynamic community finding methods:
    - *Snapshot community tracking*: methods in this subcategory try to characterize a community life cycle over the network evolution. They split the community detection method into two steps: (1) uncover communities in each snapshot using some other method; (2) track communities among consecutive snapshots to define their life cycle [40,76].
    - *Snapshot community detection*: these methods are focused on finding suitable communities structure for each snapshot of the network by processing the snapshots in their natural order using previously found results to guide the procedure. While some methods in this subcategory try to stabilize the community structure over time [77], others try to reduce the computational effort required using evolutionary strategies [78,79] or combining bio-inspired meta-heuristics and novelty search strategies [80].
    - *Consensus community detection*: methods in this category focus on finding one community structure that fits all the snapshots of the network. These methods process all the snapshots of the network simultaneously in a single process finding communities not only composed of nodes from the same snapshot but also composed of nodes from different ones [81,82].
    - *Change point detection*: the methods in this subcategory are related to the change point detection problem. They focus on splitting a network into different homogeneous periods separated by dramatic changes while also finding a suitable community structure for each period [83,84].
  - Temporal Networks-based dynamic community finding methods:
    - *Community structure update*: the methods in this subcategory track community structures in an iterative way. Each change of the network is processed in a streaming way updating the actual community structure when a change demands it. This subcategory includes methods that define a set of rules to update a community given a change or methods that do dynamic optimization of some quality metric [85,86].
    - *Temporal community tracking*: the methods in this category adapt the community structure to the changes of the network while tracking events on the dynamic communities (birth, growth, split ... etc [87,88].
    - *Persistence community finding*: the methods in this category try to find persistent structures in the whole network evolution. Some methods required a fixed duration to be used as threshold and others can find structures with arbitrary duration [89,90].

3.1.3. Information diffusion models

OSNs allow any user (u_i) to create new information in the network in such a way any user connected to u_i will receive this information (e.g., in Twitter any user can create a tweet, in Facebook create a new post, or upload a photo in Instagram). Nevertheless, a powerful tool of any OSN is that any user connected to u_i is able to send the received information to their own connections. For example, any user can 'retweet' tweets on Twitter, or 'share' a post on Facebook.

This property, along with the number of connected users to any OSN, makes that any content created may propagate fast on the network reaching a large number of potential readers. This content that is spread extremely fast, and it reaches a huge number of users in a short period, are commonly known as '*viral*' content of the OSN. Lots of users try to create viral content to gain popularity on the networks, and some others use this propagation procedure for marketing campaigns.

The classic example of marketing campaigns in OSNs is the study of the opinion of customers regarding new products [91,92], but taking into account the propagation of information in OSNs, the goal is to design marketing campaigns (such as posts, tweets, videos or photos) that reach the larger number of possible users in the shorter period of time [93,94]. Another application domain is politics, where different parties use OSN to promote their propaganda [95–97]. Nevertheless, in the last years, there has been a wide misuse of OSNs. In this sense, ISIS has used OSNs to spread their propaganda and to recruit new members [54,98–100], or the propagation of fake news [101,102].

Because of that, many researchers have focused their research on understanding how the information is spread in OSNs. More precisely, researchers have tried to define the different models that describe the diffusion process. In this sense, models can be grouped in two different categories: *explanatory* and *predictive* models [103]. On the one hand, explanatory models are based on the transmission of the epidemic, where there are users infected and users that are susceptible to be infected. On the other hand, predictive models are used to predict how the information will be spread through the network.

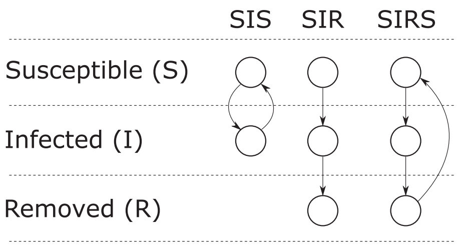


Fig. 5. Comparison among the different main SI-based models: SIS, SIR, and SIRS.

Explanatory models consider the diffusion process as an epidemic spread process, where the infection propagates between users in the same way as the information does. These type of models are based on the state of the different users and the models try to extract conclusion from how the users change their states. These states are the following:

- *Susceptible* (S): this state represents users that are not infected. In the analogy with the information diffusion, it means any user who has not received the information.
- *Infected* (I): it is used to represent those users infected by the virus or the users who have received the information.
- *Removed* (R): this state represents an infected user that that has been cured.

The classical models that belong to this category are susceptible to infected susceptible (SIS) [104], susceptible infected removed (SIR) [105] and susceptible infected removed susceptible (SIRS) [106]. All of them are based on the Susceptible-Infected model [107] that considers two states for the users (susceptible and infected), any susceptible user can be infected and once the user gets infected the state of the user cannot change. Based on this model, SIS, SIR and SIRS models differ in the number of states for each user and the transitions between these states.

The first model is SIS and it considers a daily rate of the cured patients, i.e., an infected user can be cured, and its state will change to Susceptible. The second model was proposed by Kermack and McKendrick in 1927. This model is called SIR, and it introduces a new state named *Removed* that represents those infected users that are cured. The difference between SIS and SIR is that in SIS models any cured infected user is susceptible again (which means that this user can be infected again), but in SIR models an infected user that is cured gets the state of *Removed*, which means that it is immune to the infection. Finally, the third model is called SIRS, and it considers that a cured user can become a susceptible user with a given probability. Fig. 5 shows a graphical comparison of the just explained models.

The second category of diffusion processes are called predictive models, and its most popular, and well-known models are: the independent cascade model (ICM) [108] and the linear threshold model (LTM). The goal of these models is to predict the future information diffusion process.

On the one hand, ICM was proposed by Goldenberg et al. [109,110] and it is inspired by the theory of interacting particle systems. This model takes into account the probability that an active user infects an inactive one. In this way, given two nodes that are connected in the network: u and i where u is active and i is inactive, the probability of u to infect i will be denoted as $P(u, i)$. There are two hypotheses in this model: the first one is that the probability of a node u to infect node i is independent of the influence of other active nodes connected to i. The second hypothesis is that any active node u has only one chance to infect i, independently on the result (success or not) node u will never try to infect i.

On the other hand, LTM was proposed by Watts [111] and the individuals make a decision based on its neighbors. In this model each individual has a state (*active* or *inactive*) and a threshold ($\phi \in [0, 1]$) that will define the activation level of the individual. Initially, there are a small set of users that are active. At each step, if the fraction of active neighbors of a given inactive individual is greater than its threshold, the given individual will change to active. The diffusion process finishes when the number of active individuals becomes stable.

3.2. Content-based analysis

As it has happened since the origin of the Web, the content published by humans is easily understandable by them but difficult to be processed by machines. Due to the lack of structure and the multimodality of the information published by humans (text, images, video, audio), the automatic analysis of it has been one of the biggest challenges for those algorithms that must gather and extract knowledge from this data [16]. Both facts, the lack of structure and multimodality of the information, affect the analysis of information contained in OSNs. This content analysis requires computational methods that transform unstructured content into structured information. One of the most successful areas of research in this context is natural language processing (NLP) [112]. NLP provides a set of methods and algorithms that enable the processing of multimodal information circulating on OSNs, thus allowing unstructured information to be transformed into structured information. Although, other approaches have been used to analyze video and images from OSNs, these methods can be regarded as immature when are compared against NLP [113,114]. Other successful areas, which have been extensively used to gather knowledge from OSNs, have been Data Science [56] and Big Data [115,116]. Their main methods and algorithms related to data collection, cleaning, pre-processing or mining, have been used to gather, model and extract patterns from OSNs [117–119]. The most representative applications of OSN content analysis are user profiling (Section 3.2.1), topic extraction (Section 3.2.2) and sentiment analysis (Section 3.2.3).

3.2.1. User profiling

Content-based SNA mostly focuses on the contents of the interaction between nodes in an OSN to extract topics or opinions (see next sections). Content-based SNA, however, can also be about the information about the nodes themselves, i.e., user profiling. User profiles are established based on the behavioral patterns, correlations and activities of the user analyzed from the aggregated data using techniques like clustering, behavioral analysis, content analysis and face detection (the mechanisms used in profiling users vary depending on application and purpose).

Profiling user in OSNs requires data pertaining to the user and the online activities of such user within the OSN. Those activities may depend on one's interest or they can be the effect of some influence on them. Pal and McCallum [120] exploited the content of communication messages between users to cluster email recipients into groups. For each user, they built a model that maps keywords and phrases extracted from email messages to the recipients who are likely to receive an email containing those terms. Bar-Yossef et al. [121] and Roth et al. [122] observed how users group their friends when sending email messages. Based on past behavior of user's group communication messages, they developed ranking algorithms that can predict other similar users who can belong to a particular group specified as a seed-set of users. De Choudhury et al. [123] proposed an approach to 'label' nodes in an OSN as per their roles, e.g., 'student', 'faculty', or 'director'. They applied their approach on an email communication graph, filtering out infrequent email communications below a certain threshold.

Cai et al. [124] proposed a model for mining and identifying the top-k influential bloggers based on parameters such as comments, domain of interest and page link network authority. The developed model can be used for multiple application scenarios. Sun et al. [125] proposed a novel algorithm to recommend influential bloggers based on the observation that the reproduction of blogposts and similar contents is common in blogosphere and this forms implicit links between bloggers. By measuring the text similarity of the blogposts, authors created a link graph between bloggers, and adopted the PageRank algorithm to rank the importance of bloggers.

Akritidis et al. [126] proposed a solution for bloggers identification in community blogsites that are both influential and productive. Based on the number of incoming links, two matrices are considered characterizing the bloggers as influential, productive, both, or none. This way, authors can identify blogger activities, temporal patterns and behavioral patterns. Eunyoung et al. [127] built a framework for identifying influential and popular bloggers by considering interpersonal similarity, which presents the interaction among bloggers and like-minded readers, and the degree of information propagation, which represents how many readers a blogger has. Their study showed that weighting blog social ties can differentiate influential bloggers from popular bloggers, and what make bloggers influential or popular. Finally, Cambria et al. [128] exploited the contents of social interactions to cluster OSNs based on semantics and sentics, that is the conceptual and affective information associated with the interactive behavior of OSN members. In particular, authors created a socio-interaction matrix encoding the concepts discussed by users and applied singular value decomposition (SVD) on it to perform user profiling and topic extraction.

3.2.2. Topic extraction

Topic extraction is a technique used for discovering the abstract “topics” that occur in a collection of documents, which is useful for tasks such as text auto-categorization, sentiment analysis but also SNA. Common approaches include mixture of unigrams, latent semantic indexing, LDA, and knowledge-drive methods [129]. Such methods can be used in combination with SNA to mine user interests in OSNs [130].

McCallum et al. [131] described how to determine roles and topics in a text-based OSN by building an author-recipient-topic (ART) model and a role-author-recipient-topic (RART) model. Pathak et al. [132] proposed a community-based topic model integrating SNA techniques termed community-author-recipient-topic (CART). This model was used to extract communities from an email corpus based on the topics covered by different members of the overall network. Later, Wang et al. [133] proposed an approach to predict the interests of new users or inactive users based on different social links between them using a random-walk based mutual reinforcement model that incorporates both text and links. Another work [134] proposed a regularization framework based on a relation bipartite graph, which can be constructed from any type of relationships, and evaluated it on OSNs that were built from retweeting relationships.

Kang et al. [135] presented a user modeling framework that maps user-generated content into the associated category of the news media platform. Similarly, Jipmo et al. [136] proposed a multilingual unsupervised system for the classification of Twitter users' interests. The system represents tweets and topical interests (e.g., technology, art, sports) as bags of articles and Wikipedia categories respectively, and orders the user interests by relevance, by computing the graph distance between the categories and the articles. Faralli et al. [137] proposed a method for modeling Twitter users using a hierarchical representation based on their interests. This was done by identifying topical friends (a friend represents an interest instead of social relationship) and by associating each of these users with a page on Wikipedia. Zarrinkalam et al. [138] proposed a graph-based link prediction system based on user explicit and implicit contributions to topics, relationships among users, and similarity between topics. Finally, Trikha et al. [139] studied the prediction of users' implicit interests based on topic matching using frequent pattern mining.

3.2.3. Sentiment analysis

In recent years, sentiment analysis has become increasingly popular for processing social media data on online communities, blogs, wikis, microblogging platforms, and other online collaborative media [140]. While most works approach it as a simple categorization problem, sentiment analysis is actually a complex research problem that requires tackling many NLP tasks [141], including personality detection [142], domain adaptation [143,144], and multitask learning [145]. It also comes in different flavors depending on granularity of the analysis [146], modality adopted [147] and gender [148,149]. Applications of sentiment analysis span domains like healthcare [150,151], political forecasting [152], tourism [153], rumors and fake news detection [154,155] and dialogue systems [156].

Sentiment analysis has also been applied to better understand OSN dynamics by looking at the exchange of information between network nodes. This can be useful for trend discovery, user profiling, influencer detection, and study of polarization over a certain topic or political orientation. One of the first works in this context was [157], which presented an approach for analyzing knowledge perception in an OSN. Authors provided an evidence-theory based methodology for constructing and maintaining a knowledge network in an electronic-mail communication environment. Later, [158] focused on the topic of online radicalization. Authors monitored users and interactions in a specific YouTube group using a combination of sentiment, lexical and SNA techniques. They made a number of interesting observations about the differences in the nature of the discussion and interactions between male and female members of the group.

W. Gryc and K. Moilanen studied the blogosphere's sentiment towards Barack Obama during the 2008 USA presidential elections [159]. Authors used a hybrid machine learning and logic-based framework. The results showed that the classification task in this environment is inherently complex, and learning features that exploit entity-level sentiment and social network structure can enhance classification. [160] proposed an ensemble of sentiment analysis and SNA for extracting classification rules for each customer. These rules represent customer preferences for each cluster of products and can be seen as a user model. This combination helped the system to classify products based on customer interests. Authors compared the results of their method with a baseline method with no SNA. Experiments on an Amazon meta-data collection showed improvements in the performance of the classification rules compared to the baseline method. More recently, [161] proposed an ensemble of sentiment analysis and SNA tools for group decision making. In particular, sentiment analysis was used to model consensus among experts in an OSN and later exploited to automatically generate preference relations, which were used for carrying out the group decision making process.

4. Application domains

Due to the number of application domains in OSNs is extremely large, we have selected a subset of them by the scientometric study performed in Section 2. This study allowed us to select the following application domains: Healthcare (Section 4.1), Marketing (Section 4.2), Tourism and Hospitality (Section 4.3) and Cyber Security (Section 4.4). Also, we have identified two emerging areas of SNA that are described in Section 4.5: Politics and Detection of fake news and misinformation.

4.1. Healthcare

Over the last years, there has been a huge growth of research papers that study the applicability of SN on health [162–164]. It is widely recognized that social relationships have powerful effects on physical and mental health [165–167]. With the integration of OSNs into our daily life, new levels of social interaction have emerged, and new possibilities beyond the traditional doctor-to-patient paradigm have arisen. Many patients with different diseases are now using OSNs to share experiences with other patients with similar conditions, providing a new potential source for acquiring knowledge very useful [168]. In addition, there are numerous health-related behaviors that might easily spread in SNs, such as smoking, alcohol consumption, or drug use. Therefore several research studies have been appeared analyzing this effect from the SNA perspective [162].

Smoking and alcohol consumption among adolescents are prominent risk behaviors for Health in the United States [169], and there is also substantial evidence that adolescents' use of tobacco and alcohol is highly associated with their friends' use [170]. OSNs provide a mechanism for adolescents to connect with friends instantaneously, and several research work has been focused on toward uncovering the risks associated with the usage of OSNs, such as the display of inappropriate content like sexual references and substance use [171] or tobacco advertisements [172]. Following this research line, Huang et al. [173] carried out a study to investigate peer offline and online friendships, for determining how online activities with friends might broker the peer influence processes, by either encouraging or hindering the influence of peer risk behaviors on adolescent smoking and alcohol use. For this purpose, a friendship network data about adolescent social media use and risk behaviors, were collected from 1,563 10th-grade students across five Southern California high schools, and measures of online and offline peer influences were calculated and assessed using linear regression models [174], which were fitted to test the effects of online activity with friends on smoking and alcohol use indicators. The results of the analysis showed, that the frequency of adolescent using OSNs and the number of their closest friends on the same OSN were not significantly associated with risk behaviors. However, exposure to online pictures of their friends of partying or drinking, was significantly associated with both smoking and alcohol use. Whereas adolescents with drinking friends had higher risk levels for drinking, adolescents without drinking friends were more likely to be affected by higher exposure to risky online pictures.

There are other works that use data from the OSNs to extract valuable information. For example, the text messages posted in OSNs offer new opportunities for the real-time analysis of expressed mood and social behavior patterns [175]. The analysis of these additional types of information can be very useful to the early identification, assessment, and verification of potential public health risks [176], and the timely dissemination of the appropriate alerts (public health surveillance) [177]. For instance, Twitter users may post about an illness, and their relationships in the network can give us information about whom they could be in contact with. Furthermore, user posts retrieved from the public Twitter API can come with GPS-based location tags, which can be used to locate the potential centre of disease outbreaks. Therefore, a high number of research works have already appeared, showing the potential of Twitter messages to track and predict outbreaks.

Based on this idea, a document classifier to identify relevant messages was presented in Culotta et al. [178]. In this work, Twitter messages related to the flu were gathered, and then a number of classification systems based on different regression models to correlate these messages with CDC statistics were compared; the study found that the best model had a correlation of 0.78 (simple model regression). Other similar approach was proposed by Aramaki et al. [179] where a comparative study of various machine-learning methods to classify tweets related to influenza into two categories (positive and negative) was carried out. Their experimental results showed that the support vector machine (SVM) model that used polynomial kernels achieved the highest accuracy (F-Measure of 0.756) and the lowest training time. Bello et al. [52] focused the research on the detection and tracking of discussion communities on vaccination arising from OSNs as Twitter. Well-known *regression models* were evaluated on their ability to assess disease outbreaks from tweets in Bodnar et al. [180]. Regression methods such as linear, multivariable, and SVM were applied to the raw count of tweets that contained at least one of the keywords related to a specific disease, in this case "flu". The models also validated that even using irrelevant tweets and randomly generated datasets, regression methods were able to assess disease levels comparatively well. Finally, other similar approach that collect and process the data from Twitter, is presented by Guiñazú et al. [181] to generated an information fusion model of marijuana use tendency. In this work, authors design a set of algorithms to estimate the tendency of marijuana use in relation to age, localization and gender, moreover, used a set of processes and activities to verify their model. The results obtained show the algorithm effectiveness and capacity to predict variations of complex cases like marijuana use in Chilean population.

Finally, the global situation produced by **COVID-19** has completely changed our lifestyle, impacting directly in our family, social and working dynamics. The World Health Organization (WHO) defined SARS-CoV-2 virus outbreak as a severe global threat[1]. Due to high rate of COVID-19 spread, the research community has focused their work on this virus. Although the majority of the research papers tries to study and analyze the virus from the healthcare point of view, some other works use SNs to extract some valuable information that could be used to solved the global situation.

In any crisis situation, SNs are used a way to propagate misinformation affecting and influencing social response [182,183]. For example, during the COVID-19 crisis the CNN published a rumor[2] about the possible lock-down of Lombardy to prevent pandemics. This new, published some hours before the official announcement, made people escape from Lombardy to the southern regions. As a consequence, the government initiative was disrupted. Therefore, it is really important to understand how people access to the information and how this decision affects their behaviour [184].

Social Media platforms, like Youtube or Twitter, provide a wide variety of content and not all this information is verified. Moreover, the algorithms used in these social media platforms helps the dissemination and spreading the content by taking into account the users' preferences and attitudes [185]. Therefore, this dissemination affects the users' opinion and influences the evolution of the public debate [186]. As a consequence, users tend to form polarized groups and when this polarization is high, misinformation propagates easily [187].

In this context, there are several research works focused on understanding how the information is disseminated in SNs. For example, the authors of [188] analyzed the global trends on Twitter to study the tweet volume by country. Moreover, the authors used some concepts that they linked to myths to highlight those tweets related to myths about the virus, as well as the web pages linked from the tweets.

Other approaches try to develop a risk assessment tool to quantify the rate at which any user from the selected region is exposed to unreliable post. One of the index that can be used to compute this risk is called *Infodemic Risk Index* (IRI) [189] and it takes into account the number of followers, the number of messages published by the users and their reliability.

An interesting work is the one elaborated by Cinelli et. al[190]. The goal of this paper is to analyze the interactions and engagements with COVID-19-related social media content. Using COVID-19 related keywords, authors analyzed 8 million comments and posts extracted from Twitter, Instagram, YouTube, Reddit and Gab. Then authors measured the engagement and interest in COVID-19 by analyzing the comments and the reactions, and finally, they compared the evolution of the discourse in each social media platform using the classical SIR models.

Along with the propagation of misinformation, the spread of hateful and malicious COVID-19 contents is also interesting. The work presented in [191] analyzed Twitter and 4chan data extracted from October 2019 to March 2020. Authors compared the content extracted from this network against the content extracted from the same networks but in a period where the COVID-19 was not present. Authors concluded that these social media platforms are used to disseminate disturbing and harmful information, including conspiracy theories and hate speech against Chinese people.

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.

[2] https://edition.cnn.com/2020/03/08/europe/italy-coronavirus-lockdown-europe-intl/index.html.

4.2. Marketing

In social media, brands and customers co-exist in such a way that both can interact with each other. On the one hand, companies can use social media to promote new products or to predict what the customers think about the brand or the product. On the other hand, customer can express their opinions and doubts with other customers. For all of that, social media has become the favorite and most popular platform to perform promotional activities to communicate with the targeted customers [192,193].

There is a large number of works that have studied the different aspects related to promotional activities conducted in social media platforms [194–197]. For example, it has been studied the relation between how customers perceive and formulate their attitudes towards social media advertising activities, and how these perceptions affects to the efficiency and effectiveness of such activities [198]. Other work [199], stated that if brands want to generate positive customer attitudes, they have to carefully address hedonic aspects to provide customer pleasure experiences.

What is clear is the contribution of social media in the generation of the new Word-of-Mouth, *e-WOM* that stands for *electronic word of mouth*. Using social media, customers are able to express their opinions about specific products or brands to many other customers [200–203]. One of the most famous works regarding e-WOM is the one performed by Jansen et al [204]. This work studied whether Twitter could become used as an e-WOM advertising mechanism. The conclusion of this work, revealed that 19% of the tweets analyzed mentions a brand, and 20% of these tweets contained expressions of brand sentiments. The conclusions of this work are that (1) Twitter reports what customers feel about the brand and its competitors, and (2) customers' brand perceptions and purchasing decisions are influenced by social media services. Other study is the one published by Asur and Huberman [205]. In this case authors used Twitter to forecast box-office revenues for movies. In order to do that, authors built a model based on the rate at which the tweets were created. The results reveals that this model outperforms the market-based predictors.

The rate at which the tweets, or post, are created also foster the level of interactivity and association with their customers [206,207] and thus, firms also use social media to contribute to both customer experience and customer relationship management [208]. In this sense, several authors have studied the usage of social media for the just mentioned purpose [91,200,209,210] and the general conclusion is that social media is a good tool to help organizations to sustain their relationship with their targeted customers. Nevertheless, research community has also identified that not all OSNs contribute in the same way. Moore et al. [211] concluded that this role of social media in forming customers' relationship with brands could be different according to the kind of platform used: Facebook or Twitter. In this sense, Pereira et al. [212] noticed that though customers on Facebook are enthused to follow brand accounts, they are less interested in keeping contact with them or re-sharing their content in their own page. Similar research were conducted on Twitter by Kim et al. [202], where based on the data collected by the authors, the conclusion is that brand re-tweeters show an extent level of brand trust, brand identification and community commitment.

4.3. Tourism & hospitality

Tourism and Hospitality is other area that can take advantage of the usage of social media to extract valuable information [213,214]. On the one hand, hotels and tourism destination use social media with Marketing purposes [215,216]. On the other hand, it has been probed that half of tourists change their travel plans after studying their trip on social media [217].

Tourism research has paid attention to user generated contents on social media to extract some valuable information. In this sense, there area two different data sources taken into account: (1) online textual data, published by customers on social media, and (2) online photo data. The research procedure will depend on the different source of data analyzed (text, images or both), which directly will affects to the algorithms that can be applied, and to the knowledge and patterns that can be extracted [218–220].

As in the Marketing area, tourists use OSNs to express their opinions about the different places, their experiences, and their satisfaction and dissatisfaction about tourism products. Analyzing these data, researchers have discovered relevant attributes of tourist satisfaction [221,222], the relations between tourist satisfaction and other related factors, such as guest experience and competitive position [223], or how the users use the OSNs for evaluating and improving the e-WOM hotels [223,224]. Nevertheless, the majority of the research works published are focused on investigating the customers opinion and customers experience. In this sense, Guo et al. [222] used LDA to discover the aspects that influence the customers satisfaction. Similarly, Poria et al. [225] proposed Sentic LDA, an affective version of LDA based on SenticNet [226], which leverages the semantics associated with words and multi-word expressions to extract the polarity associated to tourism-specific aspects such as accommodation, entertainment, food, and transportation [227]. Other authors used sentiment analysis to extract tourists attitude and opinion toward tourism products such as hotel services [228]. The work published by Bordona et al. [229] used clustering to group together those trips with similar `geoslotID` based on the geo-tagged messages in Twitter.

The second data type, which has been used to analyze tourism in social media, is the photos published by tourists. In this sense, the most extended approach is to analyze the metadata associated with the photo, instead of working with the photo itself. Using, for example, the geo-tag information researchers have been able to explore the tourists behavior in Hon Kong [230], to create a recommendation system that provides travel paths [231], or to select photo elements from the viewers' perspective and assist marketing organizations [232]. There are other metadata associated with the photos, for example, the user id and the photo id, the date and time when the photo was taken and uploaded, the geographical information expressed by the latitude and longitude, and other information such as the title, descriptions or tags of the photo. On the one hand, there is a wide variety of papers that uses these metadata and clustering techniques to group the different photos analyzed. In this domain it is quite popular to use a density-based spatial clustering to solve the problems related to centroid-based approaches [228,233]. On the other hand, other researchers focus their work on travel trajectories, i.e., the sequence of tourism spots and time intervals between them to recommend travel plans for tourists. For example, [231] suggests different routes taking into account the quality and popularity of the routes. Other approach is the one followed by Vu et al. in [230], where authors applied Markov models to predict the next tourism spots based on the current location of the tourists.

4.4. Cyber security

SNA can be used to detect and apply different strategies to support law enforcement agencies in the fight against cyber-crime and cyberterrorism. Criminals, and terrorists, use OSNs, such as Twitter or Facebook, due to the huge number of users that connects everyday to these networks. Moreover, the internal structure of these networks, make that any published message propagates very fast, reaching a high number of potential readers. In this domain, the purpose of SNA techniques and algorithms is to extract the different patterns of criminals and terrorist in OSNs. More precisely, the main goal is to detect and discover crimes and their relations with criminals.

Regarding crime analysis, huge efforts have been made to facilitate the communication between citizens and government agencies. Initially, this communication was roughly performed through telephones, or face-to-face meetings. Then, the information provided by the citizens was saved, or transformed into written text and then archived in a digital format. In order to automate and facilitate crime analysis, Chih-Hao and Gondy [234] designed a decision support system that combines NLP techniques, similarity measures, and classification approaches. Filtering reports and identifying those that are related to the same or similar crimes can provide useful information to analyze crime trends, which allows for apprehending suspects and improving crime prevention.

Other works have been focused on the prediction of the different crimes by trying to discover hidden patterns of criminal behavior. In this sense, it is quite popular geographic knowledge discovery techniques that can be used to discover these patterns that may help in detecting where, when, and why particular crimes are likely to happen. In order to do that, Phillips et al. [235] presented a crime data analysis technique that allows for discovering co-distribution patterns between large, aggregated, heterogeneous datasets. Authors modeled the aggregated dataset as a graph that stores the geospatial distribution of crimes within given regions. Then, these graphs were used to discover those regions with similar geospatial distribution characteristics.

A different approach is the one followed by Chainey et al. [236]. This approach is based on the visual identification of regions where the crimes can be produced by using the *hotspot mapping*. This technique is used to predict where crime may happen, using data from the past to inform future actions. Each crime event is represented as a point, allowing for the geographic distribution analysis of these points. A number of mapping techniques can be used to identify crime hotspots, such as: point mapping, thematic mapping of geographic areas, spatial ellipses, grid thematic mapping, and kernel density estimation (KDE), among others. In the just mentioned work, authors performed a comparative study of these mapping techniques and the results revealed that KDE was the most successful technique. This conclusion was also extracted from the work performed by Gerber [237], where KDE was used to automatically identify discussion topics across a city in the United States by applying linguistic analysis and statistical topic modeling over a spatio-temporally tagged tweets extracted from Twitter.

Different approaches are the ones applied to analyze the terrorists' behavior in OSNs. From the last years, it is quite simple to realize the ability of terrorist groups in the usage of OSNs for their own benefit [238,239]. Terrorist organizations use OSNs to promote their ideology, and to recruit individuals to their cause. Usually, the first conversations start in the most famous OSNs such as Twitter, Facebook, or Instagram, and then they continue using private message with the target individuals. In order to disconnect these radicalization channels, governments, organizations and social media platforms are continuously searching social media accounts that can be associated with such terrorist groups to block them.

From the research community, a wide range of works has been done in order to help law enforcement agencies. In this paper we are going to talk the four main application domains related to counter-terrorism in OSNs. The first one is focused on the identification and understanding of the language used, or the definition of the linguistic markers. The key idea is the analysis of the text published in the OSNs in order to determine whether the corresponding account belongs to a terrorist or, at least, a supporter. Cohen et al. [240] discussed about the possibility of detecting some of the linguistic markers defined by Meloy et al. in [241]. More precisely, Cohen et al. discussed about the detection of *leakage, fixation* and *identification* warning behaviors because these are the ones have the greatest potential to be discovered with text analysis methods. In a more practical way, Torregrosa et al. [242] compared the tweets published by pro-ISIS Twitter accounts against the text published by random Twitter users. To analyze the terminology of the tweets they have used Linguistic Inquiry Word Count (LIWC) software. Experimental results reveals that pro-ISIS accounts publish tweets using the third person plural pronouns, they use more words related with death, certainty, and anger, and more negative language than random accounts. Some differences in the language was also evinced by Lara-Cabrera et al. [54]. They also compared the tweets published by pro-ISIS accounts and random Twitter accounts, using a set of keywords that was extended with different synonyms, and taking into account the stem of the words. Experimental results revealed that metrics defined to measure the indicators performs well in the tasks of identifying those accounts that uses a radical vocabulary. Once the user has been identified as terrorist, or supporter, it is possible to study how this user influence the other users in the OSN [243].

All these works, help to highlight those social media accounts that shows a radical behavior based on the content of the tweets. Other approaches model the OSN in a graph and use connections to detect the critical node [244]. For example, Gunasekara et al. [245] uses the betweenness centrality of the nodes to detect the critical nodes of the graph. The main problem with the betweenness centrality metric is that the metric is really expensive from the computational point of view, it runs in $O(nm)$. In order to alleviate this problem, other works uses heuristic or bio-inspired approaches such as Lozano et al. [246] that integrates an updated procedure of betweenness centrality metric with an artificial bee colony algorithm.

4.5. Emerging areas

Finally, and out of the scope of our scientometric analysis carried out in Section 2, we have decided to make a brief analysis of the state of the art in some emerging areas, that are currently experiencing an increasing interest in the area of OSNs. These emerging areas are directly related to recent, but highly societal demanding topics, such as politics and detection of fake news and misinformation, or the integration of multimedia information. With the irruption of 5G and IoT technologies [247,248], it would be expected that in the next years the concept or social Internet of Thinks [249] will generate a huge interest in the field of SNA allowing to generate new kind of intelligent services to end-users through the combination of Machine Learning, Artificial Intelligence and IoT methods. We strongly think that these areas could be in a near future high activity niches, for the research communities involved in areas as SNA, Data Science or Big Data.

4.5.1. Politics

The Arab Spring in 2011, and both Obama's campaigns (in 2008 and 2012), marked the beginning of how social media might affect citizens' participation in political life [250]. Since these dates, politicians, citizens and researchers have expressed their interest in how they can take advantage of using social media to participate in political life. In this sense, politicians use social media to attract supporters, and people have been using it to express their political views and opinions about various leaders and issues. In fact, Boulianne [251] studied whether social media use, and participation in elections, were correlated. Paying attention to the metadata, more than 80% of the coefficients showed a positive relation.

The research community has used social media to study a wide variety of problems. The most remarkable are the following: (1) hate speech detection [252,253]; (2) topic opinion, or political polarization [97,254]; (3) community finding problems [117,255]; and (4) information exchange and information diffusion [11,256]. The majority of the research works start with the same idea of analyzing the text of the posts, or comments, and using the results of this analysis to perform a second analysis more focused on the problem to be solved.

Hate speech is used in those posts or comments that defames, belittles, or dehumanizes a class of people on the basis of certain inherent properties such as race, ethnicity, gender, or religion. Due to its goal, it is critical to design systems that automatically classify the posts or comments based on its content and determine whether the corresponding post contains hate speech or not [252,253]. In Müller and Schwarz demonstrated that there is a significant correlation between increased German hate speech on social media, and physical violence towards refugees in Germany [257]. One recent work that tries to identify hate speech is the one published by Jaki and De Smedt [258], where authors tries to understand what disparaging verbal behavior from extremist right wing users looks like, who is targeted and how. In order to do that, they analyzed 55.000 right-wing German hate tweets from August 2017 to April 2018. The proposed method is able to detect right-wing hate speech with 84% accuracy. Other relevant work in this domain is [259], where authors used several classifiers to detect hateful and antagonist content in Twitter. Although the individual results for each classifier reduce false positives and produced promising results regarding false negatives, the combination of classifiers into an ensemble classification approach seems to be the most suitable method.

There are other works that are focused on the analysis of the content of the tweets, or posts. In this case, we can talk about topic opinion or political polarization. Although these concepts have different goals, the procedure followed is quite similar. In both cases, works try to extract some valuable information from the content of the tweets. On the one hand, if the goal is to understand what the citizens think about specific topics, we talk about *topic opinion*. On the other hand, the goal of *political polarization* works is to discover the alignment of the citizens with the corresponding parties. In this sense, [260] belongs to topic opinion family. Authors analyzed 1,150,000 messages from about 220,000 users to define the characteristics of the three main parties in the 2010 UK General Election and highlight the main differences between parties. In a first analysis they realized that: (1) the retweet structure is highly clustered according to political parties, (2) users are more likely to refer to their preferred party and use more positive affect words for the party compared with other parties, and (3) the self-description of the users can reflect the political orientation of users. Based on these evidences, authors developed a classification method that uses the number of tweets referring to a particular political party, and its semantic content to estimate the overall political leaning of the user. The experimental results achieved an accuracy of 86% for classifying the users' political leanings. Other study [261] tries to determine whether the Twitter users can be grouped around the different parties during the elections. To do that, they analyzed around 6,000 tweets published by 1,500 Twitter users during the 2011 Canadian Federal Election. Authors concluded that Twitter usage is likely to further embed partisan loyalties during electoral periods rather than loosen them. Therefore, it seems that partisan are closer to their corresponding parties during electoral periods. Other works try to predict the vote intention of the citizens based on their public posts. This is the case of [262], where they tried to classify people's voting intentions based on the content of their tweets during the Scottish Independence Referendum in 2014. They built a topic-based naïve bayesian model (TBNBM) that takes into account the dependencies between topics and user voting intentions. This TBNBM detects the topics using the LDA, and for each topic they built a probability table where each feature has two associated conditional probabilities related to both voting intentions (i.e., 'Yes' or 'No'). In the experimental phased authors realized that this TBNBM improves the classical bayesian classifications. Finally, other relevant work is the one published by Borge-Holthoefer et al. [263]. In this case, authors used social structured and the content of the tweets to understand the opinion evolution in Egypt during the summer of 2013. They observed that the military takeover caused major quantitative (volume of polarized tweets), but not ideological (polarity swaps) shifts among Twitter users. They also observed how the pro-military Twitter users that were very loud before the take-over, became increasingly silent afterwards, and how anti-military intervention Twitter users become significantly louder after the takeover.

Some of the just mentioned works can be also categorized into community finding problems because they detect some clusters, or communities, of users based on different aspects. In this domain, it is important to highlight the work published by Ozer et al. [264] because they developed three Non-negative Matrix Factorization frameworks to investigate the contributions of different types of user connectivity and content information in community detection. They revealed that user content and endorsement filtered connectivity information are complementary to each other in clustering politically users into pure political communities. Other work [265], predicted the political alignment of Twitter users based on the content and structure of their political posts after the 2010 U.S. midterm elections. They used manually annotated data and the different communities are created by taking into account the TF-IDF and the hastags obtained by content analysis. They found a highly segregated partisan structure with few retweets between left and right-wing Twitter users.

Finally, some other works are focused on how the information is propagated through the network. In this regard, a well-known work is the one published by Colleoni et al. [266]. In this work, authors used a combination of machine learning and SNA to classify users as Democrats or Republicans based on the content shared on social media. Then, they investigated the political homophily in both: the network of reciprocated and non-reciprocated ties. They found that the structures of political homophily differ significantly between Democrats and Republicans. Other works try to study whether the sentiment of a tweet affects to its propagation [267], i.e., how the affective dimensions of tweets, including positive and negative emotions associated with certain political parties or politicians, affect the quantity of retweets. Experimental results over around 64.431 political tweets reveals a positive relationship between the quantity of words indicating affective dimensions (positive and negative) and its retweet rate. The last remarkable work [268], authors studied whether the online communication of political and nonpolitical issues resembles an "*echo chamber*" (i.e., communication between individuals with same ideological segregation) or a "national conversation". In order to do that, authors analyzed around 150 million tweets regarding 12 political and nonpolitical issues, extracted from 3.8 million Twitter users. The findings suggest that in terms of political issues, the information was exchanged primarily among individuals with similar ideological preferences; but this effect did not happen with non-political issues.

4.5.2. Detection of fake news & misinformation

In 2018, 66% of American adults consume news on social media[3], whereas in 2012 only 49% of adults used OSNs for this purpose. This increment is due to two main reasons. The first one is related to the fast propagation of messages on OSNs. The second reason is related to the high number of users connected to any OSN which makes quite easy to comment and discuss any message. Due to the easiness of creating messages and disseminating them in any OSN, it has been significantly increased the number of *fake news* [102,269,270], i.e., those news with intentionally false information produced online for a wide range of purposes, such as financial and political gain [101,271].

In spite of its popularity, the research community is not able to agree in a common,and unique, definition of the term *fake news*. The most extended definition is the one that states that fake news are those news articles that are intentionally false, its truthfulness can be verified, and its intention is to mislead readers [101,272,273]. This definition is based on two key concepts: *authenticity* and *intent*. The former means that fake news contain false information that can be verified. And the latter refers to the fact that fake news are created to confuse and mislead customers.

The main risk about fake news is related to the concept of *intent*, just described. The goal of any fake news is to persuade consumers to accept biased, or false, information usually with political messages or influence. This goal and the fast propagation of messages in OSNs made that, for example, the most popular fake news was even more widely spread on Facebook than the most popular authentic mainstream news during the U.S. 2016 president election[4]. For all of that, it is really important not only to understand how fake news propagates through the network (see Section 3.1.3) but also to develop systems that detect whether a specific new is a fake news or a real one [102].

[3] http://journalism.org/2018/09/10/news-use-across-social-media-platforms-2018.

[4] http://buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Fake news detection is a promising area that tries to define whether a specific new is fake or not [102,272]. There are different approaches that can be followed to do that, but all of them can be grouped in two big categories: the *linguistic approaches* and the *network approaches*.

The goal of *linguistic approaches* is to detect the fake news by analyzing its context. The idea is to obtain the different writing styles to detect fake news. In this sense, research works focus on two different linguistic features: (1) *lexical features*, such as total words, characters per word, frequency of large words, etc. [274], and (2) *syntactic features*, such as "n-grams" and bag-of-words (BOW) or parts-of-speech (POS) tagging [275].

Second category, *network-based approaches*, are based on the network that can be built by taking into account the users that publish related social media posts. In this case, researchers create a network based on some specific interaction on the OSN, and then, they apply network metrics to extract the valuable information. There are different types of network that can be built. For example, the *stance network* is a graph where the nodes represent all the tweets relevant to the news, and the edges contains a weight that indicates the similarity of stances [276,277]. A different approach is the one followed in the *co-occurrence network* where two nodes are connected by a weighted edge that represent how many times both users have written post relevant to the same news articles [278]. And the last example is *friendship network* users who posted related tweets are connected in the network[279].

Finally, once the model is built, the network metrics can be used to extract valuable information. For example, authors of [279] used degree and clustering coefficient to characterize the diffusion network. A different approach is the one followed in [278] where SVD is used to learn the latent node embedding.

4.5.3. Multimedia

Nowadays, although users are able to generate different types of content in OSN like audio, image or video, text is the most common, and popular, content analyzed. The analysis of audio, image or video, has received less attention compared to text mainly due to the complexity of the analysis tasks, and the initial technological limitations of internet in its early years. The internet connections were slow and encoding algorithms were poor, thus, sharing videos, audios and images were impractical. Nowadays, internet connections have improved and smartphones, all equipped with cameras and microphones, have emerged as a portable alternative to traditional computers with access to the internet. As a result, it is very easy to share videos, photos, or audio using those devices, in fact, some of the most popular modern OSN are completely based on sharing images and videos like Instagram or Pinterest. Consequently, the interest in analyzing multimedia content in OSN has suffered an important increase.

Taking into account the different multimedia types audio, video and images, it is the latter the most common data type analyzed. Some of the works found in the literature have similar goals as the ones analyzing texts, such as Sentiment analysis [280]. Furthermore, Image Annotation [281], which is generating a set of words that describe the content of a picture, is also commonly applied and it has similar goals to the keyword extraction methods. Besides, Image Clustering [282] can also be found, and it would be the equivalent to topic extraction. Although the just mentioned approaches share the same goals as some of the most popular text-based algorithms, the methods based on images are more complex due to the complexity of processing images. One characteristic of the images, that it is not present in the text analysis, is the fact that images contain information about the geographical location where the image was taken. This fact has allowed SNA researchers to use images as geo-location alternatives.

There is a plethora of works that analyze images for extracting some valuable information. For example, it is possible to predict the popularity of a picture [283,284], to predict the gender of users [285], to discover events in public places [286], or forecast the ambiance of a place [287], by applying Sentiment analysis and Image Annotation. Other works, like [288], have used Image Clustering to analyze the content of popular images in OSN, such as pets, food,...etc. There is other wide set of works that have developed their own algorithms for analyzing faces and predicting human characteristics [289,290] or perceived intelligence [291]. Finally, other works need the supervision of humans to validate the output of the algorithms. Some examples of these works are the one that analyze marijuana-related content [292] or cyber-bullying content [293].

It is possible to classify the different algorithms and methods developed to perform some image analysis in three different groups: (1) Crowdsourcing, (2) Deep Learning, and (3) Handmade Features. On the one hand, Crowdsourcing [294] consists of the division of a work package into small pieces, or sets of images, and distributing them between a large number of participants, humans, to achieve cumulative results fast. Usually, this task is performed using some online platform that acts as an intermediary between the owner of the dataset and the workers. Popular platforms for doing image analysis based on Crowdsourcing are Amazon Mechanical Turk[5] (MTurk), Figure Eight[6] (formerly CrowdFlower) or MicroWorkers[7]. On the other hand, Deep Learning and Handmade Features avoid human intervention and use algorithmic approaches to extract significant features from images to be used as representatives. Based on this idea, Handmade Features use custom human-made filters that allow characterizing an image by low-level features, such as, colors [295,296], shapes [297,298], and textures [299,300], or by high-level ones like the Semantics-Preserving Bag-of- Words (SPBoW) model [301] or the Contextual Bag-of-Words (CBoW) one [302]. Deep Learning, specifically convolutional neural networks (CNNs) [303], automatically construct these filters on their own. To this end, a supervised learning approach is used to generate a series of convolutions, ordered as layers connected one to another, capable of compressing an image into a representative set of characteristics. It is common to find works that use deep CNN (those with several layers), already trained to identify a huge number of different types of pictures (elephants, cars, plants, houses...etc). For example, in [304] authors trained CNN with the ImageNet dataset. Finally, once the characteristics are extracted, common supervised/unsupervised machine learning techniques like neural networks [305], SVM [306], k-means [307] or DBSCAN [308] are used for doing particular analyzes.

Audio analysis in OSN has been largely untapped so far. It is quite common to apply Automatic Speech Recognition [309] to transcript the audio into a text and then, applying Speech Sentiment analysis extract some useful knowledge [310]. In spite of, for the best of the authors' knowledge, there are no works that have used these techniques to SNA, there are some applications of audio analysis in OSN. The first one is related to music. The Internet has been influencing the music scene for the last few decades, making music more accessible to the public and engaging musicians with their audience more easily. Thus, several social media platforms exist focused on music like Last.fm[8] or SoundCloud[9]. With this in mind, [311] proposed to use audio analysis to categorize music into genres to enhance community finding techniques. The other main application of audio analysis is generating conversational networks from raw audio data [312,313]. Once the conversational networks are generated, SNA techniques are used to perform speaker role recognition [314,315] and summarizing [316] on broadcast news and podcasts.

Video analysis is, by far, the media type least studied in the context of OSN. It is also the most complex media type to analyze from the three mentioned in this section (image, audio, and video) because it encompasses the other two. However, video media content is prevalent on the

[5] http://mturk.com.

[6] http://figure-eight.com.

[7] http://microworkers.com.

[8] http://last.fm.

[9] http://soundcloud.com.

Internet and the most popular OSN are completely centered on it, like Youtube or Youku (YouTube's Chinese counterpart). Although, it can be found several papers focused on the video analysis, these techniques have not been applied yet in conjunction with SNA techniques. One example is the work published in [317] where authors apply sentiment-analysis to the video once it had been transcribed.

5. Discussion on research methods and application domains

From the analysis of the state of the art in both research and application areas of SNA, shown in Sections 3.1 (OSN Structural-based analysis: graph theory, community detection algorithms, information diffusion), 3.2 (OSNs Content-based analysis: topic extraction, opinion mining and sentiment analysis, multimedia), and Sections 4.1 to 4.5 (health, marketing, tourism and hospitality, cyber security, politics, fake news and misinformation, 5G and IoT technologies), some important conclusions can be drawn:

- On the one hand, OSNs are used as the highest socially trending, and influencing, source of information. Currently these sources have become on the most popular ones in the world, with billions of active users generating huge amount of data (in form of textual information, video, and other multimedia material) per second.
- Working with only a small part of the available data stored in any OSN, exceed the classical processing algorithms capabilities, and suppose one the biggest current challenges for a wide number of research areas. The complexity of this domain needs by a joint effort from different research areas, so the *multidisciplinarity* between areas, from Social Sciences to Science and Engineering, becomes a necessity to find new methods and technologies to exploit the enormous potential of these sources.
- Although multidisciplinarity is a necessity, it means that professionals from very different areas, with very different background, have to collaborate together on a wide range of technologies (e.g., sociologists and data scientists).
- On the other hand, most of algorithms and methods studied attempt to address the set of **challenges** (gathering, and processing the data, finding useful patterns, visualize the information, etc.) imposed by the complexity of this domain, in particular:
  1. Both structural-based and content-based algorithms, are used to *extract or discover* useful knowledge from the network. This challenge is related to the classical Knowledge Discovery problem in Machine Learning, Data mining and Data Science fields. This challenge, or problem, is directly related to our proposed RQ1 (*What can I learn?*).
  2. The necessity to manage a huge, and exponentially growing amount of data, needs from new methods, *scalable* algorithms, and technologies. This is one of the hot topics in areas as Big Data, and it is related to RQ2 (*What is the limit?*).
  3. Due to the vast number of sources available, and their different data formats (numerical, categorical, textual, metadata, video, images, audio), it's necessary to develop new algorithms capable to *integrate* and *fusion* different sources to allow discovery new and useful knowledge. This problem is usually address by the area of Information fusion, and it's related to RQ3 (*What kind of data can I integrate?*).
  4. Finally, one of the essential tools for knowledge discovery in OSNs is related to the *visualization* of the information. This is a challenging, and still open problem, in the area of Information Visualization, so the RQ4 (*What can I show?*) has been defined to cover this aspect.

Taking into account previous conclusions, this work proposes the definition of four **dimensions**, which can be used to assess the maturity of technologies currently available in OSN. These four dimensions will later be used to define a set of **metrics** (which we named *degrees*$_*$) that will be used:

- *To quantitative assess the level of maturity of a set of SNA tools and frameworks*. To do that, a set of graphical representations (based on spider graphs), and a new global metric (named $\mathfrak{C}_{SNA}$), will be defined to provide a quantitative measure of these tools and frameworks.
- *To study some possible future trends, challenges, and lines of work in these dimensions* (closely related to research areas such as Data Science, Big Data, Information fusion and Visualization), detecting spaces for improvement, and emerging technologies, where new developments could have a high impact on the scientific community, industry and society.

6. The four dimensions of social network analysis

The Big Data paradigm was characterized by the different V-models that allow any researcher to analyze the capacity of the different Big Data methods. Initially, 3Vs were described in the 3V model [13], but this model has evolved during the last years to the 4V [318,319], 5V [320], or 6V model [321]. These models allow measuring the maturity of different methods, tools and technologies based on Big Data using simple features such as Volume, Velocity, Variety, Value, Veracity or Variability. There are even some attempts to include new 'V's likes Visualization, in these *V-models. This set of *V-models provides a straightforward and widely accepted definition related to what is (and what is not) a big-data-based problem, application, framework, or technology. We will use this interesting (and successful) approach to describe the challenges, and the current status, of technologies related to OSN. We have mapped our four Research Questions stated in Section 1, into a set of equivalent dimensions: D1) *Pattern & Knowledge discovery*, D2) *Scalability*, D3) *Information Fusion & Integration*, and D4) *Visualization*. Using these dimensions, we can quantify the different methods, techniques, algorithms and frameworks, allowing us to better understand where we stand and where we could be in the near future in this area.

6.1. Pattern & knowledge discovery (D_1)

In the area of Big Data, the concept of *Value* usually refers to the process of extracting valuable information from very large data sets (e.g., from social big data [117]), and it is usually referred to as Big Data Analytics [322–324]. This concept can be easily extrapolated to the area of SNA. This first dimension will be used to define the capacity of Knowledge discovery (mainly from a pattern mining perspective) of SNA technologies. This dimension tries to answer the question: *What can I learn?*, understood as the capacity of any method or algorithm to discover non-trivial knowledge from OSN.

The objective of this dimension is to evaluate, in a generic way, any type of technique, method, or tool, which is used to discover new knowledge in OSN. In the literature, there are several works reviewing the different techniques and algorithms that are usually applied in OSN to extract knowledge. Taking these works into account, we can generate an overall taxonomy of the different functions and methods that SNA tools usually provide. Then, using this taxonomy, it would be possible to quantify the **degree of value** that each tool provides, according to the functionalities that it covers within the taxonomy generated.

The work presented in the 90s by Wasserman and Faust [48] is one of the most relevant publications on the area over the years. Based on the main aspects presented in that work and analyzing different publications reviewing the applications and methods to extract knowledge from OSN [325–329], in general terms, it could be considered that the taxonomy of the main functionalities for discovering knowledge which can be embedded in SNA tools are the following:

1. *Qualitative and quantitative/statistical analysis* ($F_{Value(1,i)}$):
   (a) *Computation of measures based on the topology* ($F_{Value(1,1)}$) of the network that provides a local and global description of it. These type of measurements are extracted from the graph theory, being

some of the most relevant the density, distance, centrality or transitivity, amongst others. A huge number of works [27,32,330], used previous measures to discover the most influential, prestigious, or central nodes (or actors) within an OSN. The value of this measure, $F_{Value(1,1)}$, will be calculated according to the following scale: +1/5 until 1 for each measure employed (diameter, mean degree/distribution, cluster coefficient, connected components, transitivity, triangle count, etc...).

   (b) *Link analysis* ($F_{Value(1,2)}$) algorithms arise with the aim of finding the most valuable, authoritative or influential node (e.g., a webpage in the Web), being the HITS [331] and the Google Pagerank [332] the most popular ones. This measure will be calculated according to the following scale: +1/4 until 1 for each algorithm provided.

2. *Pattern mining* methods ($F_{Value(2,i)}$):
   (a) *Community detection* ($F_{Value(2,1)}$) algorithms (static and dynamic) [33,62,333] try to find groups of nodes (users) where the set of edges is dense within the group and sparse outside it. One of the main difficulties in this topic is how a community is defined, since it can depend specifically on the domain where it is applied and what the network represents, as well as the type of links that are considered. Naturally, community detection algorithms are based on concepts from graph theory and clustering. Indeed, this problem is very similar to the problem of graph clustering. Recently, the analysis to the dynamics and evolution of OSNs has grown hugely, therefore many of the classic community search algorithms have been extended to study also the behavior of the communities over time. The scale to compute this measure is the following: +1/3 if allows one kind of community detection (overlapping or non-overlapping); +1/3 if allows to do overlapping and non-overlapping community detection; +1/3 if allows to do community detection in temporal networks.

   (b) *Opinion mining* ($F_{Value(2,2)}$) techniques [334] are focused on the detection of user opinions, and also feelings or reactions of people about certain beliefs, products, decisions or events. OSN are online sites where people can express their ideas and opinions, exchange knowledge and beliefs or criticize products. Millions of new posts giving opinion on products and services are generated every day in OSN. All this information dumped on OSN mostly in text format is very valuable to discover new knowledge. One of the most famous methods within these techniques is the *Topic Detection* algorithms, that are based on the idea of applying data mining techniques to detect what topics are more popular over the time [335]. In addition, *Sentiment analysis* methods [336] try to identify how people feel about a specific topic. This issue can be as important as detecting the topic itself on the OSN, and it can be addressed using data mining techniques related to NLP [337]. This aspect is measured according to the following scale: +1/2 if it provides topic detection; +1/2 if it provides sentiment analysis.

   (c) *Homophily models* ($F_{Value(2,3)}$) [338]. Homophily is the tendency of similar individuals to connect together. The well-known saying, "birds of a feather flock together" refers to the Homophilic behavior of the real-word. People's personal networks are homogeneous with regard to many sociodemographic, like age, race, or ethnicity and the same goes for OSN. Unlike influence, where an influential influences others, in homophily, two similar individuals decide to get connected only because of that similarity. In order to calculate this measure, the following scale is used: +1 if available.

3. *Predictive analysis* ($F_{Value(3,i)}$):
   (a) *Propagation and virality modeling* ($F_{Value(3,1)}$) consists on the study of the spread of influence through OSN. This issue has a long history in the area of social sciences, where the first studies was emerged on medical and agricultural research areas [339,340]. In recent years, these type of models have been applied by marketing researchers, trying to model the 'word-of-mouth' diffusion process for viral marketing applications [109,341]. The value of this measure will be calculated according to the following scale: +1 if available.

   (b) *Link prediction* ($F_{Value(3,2)}$). In dynamic or temporal networks, a typically problem addressed is estimating the probability of two particular nodes are to become connected in the future. This is a classical computational problem underlying in OSN evolution over time, that was introduced by Liben-Nowell and Kleinberg [342] as the link prediction problem. It can infer new interactions among members of an OSN that are likely to occur in the near future. This measure will be computed according to the following scale: +1 if available.

As mentioned above, this proposed taxonomy can be used to quantify the **degree of value** ($d_{Value}(t)$) that each SNA tool (t) provides according to the functionalities that it covers as shown in Eq. 1. Several weights are used (α, β, and γ), to represent the importance given to each characteristic. In this work, all of the value characteristics will have the same weight (so α, β, and γ will be set up to 1/3). This equation has been normalized in the range [0,1] taking into account the different methods, techniques, algorithms and measures that are incorporated by the particular tool ($d_{Value}(t) \in [0, 1]$).

$$d_{Value}(t) = \alpha \cdot \frac{\sum_{i=1}^{2} F_{Value(1,i)}(t)}{2} + \beta \cdot \frac{\sum_{j=1}^{3} F_{Value(2,j)}(t)}{3} + \gamma \cdot \frac{\sum_{k=1}^{2} F_{Value(3,i)}(t)}{2} \quad (1)$$

6.2. Scalability (D$_2$)

Currently, the exponential growth of data has created serious problems for traditional data analysis algorithms and techniques (such as data mining, statistics, machine learning, and so on) to processing the data available in electronic sources. A new generation of algorithms and frameworks is currently being developed in order to manage big data challenges, and they require high scalability in both memory consumption and computational time [117]. Therefore, this dimension will be used to define, and quantify, the scalability capacity of a tool or technique (e.g., algorithm) used in an OSN. From this perspective, the amount of information handled will be the key feature considered (mainly using the quantity of nodes and edges that can be processed by the SNA method or tool). A highly scalable software would work correctly on a small dataset as well as working well on a very large dataset (say millions, or billions of nodes and edges). In general terms, scalability refers to those techniques that ensure that some quality of service is maintained as the size of the data set to be managed grows, or the complexity of the addressed problem increases.

Big Data systems like MapReduce [343], Hadoop [344] or Spark [345] have been developed as a response for these scalability problems. However, the specific application areas of SNA are usually modeled as graph-theoretical problems, and unfortunately, the direct application of graph algorithms in these big data environments is often not an efficient solution (most of classical graph-based problems are NP-hard). Another approach to tackle this problem is the graph-parallel systems for specialized graph processing problems. These systems perform better than the general tools for Big Data, but their main disadvantage remains that they can only be used for graph specific problems [346,347]. An efficient way to address this problem is to combine the advantages of both approaches: the graph-parallel approach, and the general big data processing tools. For example, GraphX is an Apache Spark's built-in library for graph analytic and graph-parallel computation [348,349], which provides an excellent, and highly scalable, solution for graph-based problems.

In general terms, scalability is a desirable aspect of a network, system, application, or process. This concept can be defined as the capability of a system to handle an increasing number of elements or objects, to process increasing volumes of work adequately, and to be easily enlarged or extended [350]. This means that the application or system should have the ability to continue functioning correctly when the problem it is changed in size or volume, taking full advantage of it resources in terms of performance. The scalability of a system usually depends on the types of data structures and algorithms used to implement it, or if it has different components or modules (and how them have been designed), or on the communication mechanisms used by its components. For example, the data structures of a system affect not only the amount of space required to perform a particular function, but also the time. Through this last observation two of the main aspects of scalability can be appreciated: *space* and *time*. In Bondi [350] work, a more detailed analysis of scalability aspects is presented, where the author considers four main types of scalability on a system:

1. *Load* scalability: the ability to function with agility (without undue delay, without unproductive consumption of resources, and making good use of the available resources).
2. *Space* scalability: the memory requirements do not grow to intolerable levels, as the number of items that the system supports increases.
3. *Space-Time* scalability: the ability to continue operating with agility as the number of objects or data to be processed increases by orders of magnitude.
4. *Structural* scalability: a system is structurally scalable if its implementation does not impede the growth of the number of objects or items it is capable of handling.

Load scalability may be improved by exploiting parallelism, but the other three characteristics mentioned of scalability are inherent to the architectural design and implementation of the system (such as the length or the choice of data structures), and in many cases are difficult or even impossible to change. In addition, when a taxonomy of characteristics is defined, it is natural to study whether there are dependencies between them. In this particular taxonomy, for example, systems with poor space scalability or space-time scalability, might have poor load scalability, due to the attendant memory management overhead, or search costs. On the other hand, systems with good space-time scalability because their data structures are well engineered, might have poor load scalability due to poor decisions about scheduling, or parallelism, which have nothing to do with memory management.

Another taxonomy of characteristics related to the scalability was proposed by Hesham and Mostafa [351], in this particular case, it was defined for those architectures that allow parallel processing. This work presents multiple dimensions to measure the scalability, such as:

1. *Administrative* scalability: The ability for an increasing number of organizations or users to access a system.
2. *Functional* scalability: The ability to enhance the system by adding new functionality without disrupting existing functions.
3. *Geographic* scalability: The ability to maintain effectiveness during expansion from a local area to a larger region.
4. *Load* scalability: The ability to expand and contract to accommodate heavier or lighter loads.
5. *Generation* scalability: The ability of a system to scale by adopting new generations of components.
6. *Heterogeneous* scalability: The ability to adopt components from different vendors (or others environments or systems).

As can be seen, there are several works that try to describe which characteristics should be taken into account when analyzing if a system is scalable. However, any of these mentioned works provide a model to quantify the degree of scalability of a system. The Universal Scalability Law (USL) presented by Gunther [352], provided a quantification model for scalability of the systems or applications. This model is defined in terms of three main parameters α (contention), β (coherency), and γ (concurrency), that can be identified respectively with the three Cs [353]:

- *Concurrency* (γ): the maximum throughput (the measure of a number of requests processed over a unit time by the application) attainable at a given level of load.
- *Contention* (α) queuing time for shared resources.
- *Coherency* (β) delay time for data to become consistent (or coherent), by virtue of point-to-point exchange of data between resources that are distributed.

Defining the system throughput $X(N)$ at a given load, N, the USL can be expressed as the equation:

$$X(N) = \frac{\gamma \cdot N}{1 + \alpha \cdot (N-1) + \beta \cdot N \cdot (N-1)} \tag{2}$$

The independent variable N represents the number of users or data load that is incremented on a fixed hardware configuration. When the scaling is linear-rising (the case for ideal parallelism), the $\alpha = \beta = 0$. In other words, the overall throughput $X(N)$ increases in simple proportion to N.

Following the same approach of using throughput to measure scalability, in the work presented by Jogalekar and Woodside [354], the scalability is measure based on the 'power' metric of Giessler et al. [355] as follows:

$$Power = \frac{\gamma}{T} \tag{3}$$

where γ is throughput and T is mean delay time.

As mentioned at the beginning of the section, the technologies developed for big data environments have arose to address the problem of scalability due to the exponential data growth that has occurred in recent years. Nevertheless, how to measure scalability in Big data environments is a question that is still being addressed. In Sanchez et al. [356] work, the isoefficiency model is introduced as a standard measure of scalability. The isoefficiency function determines the ease with which a parallel system can maintain a constant efficiency and hence achieve speedups increasing in proportion to the number of processing elements [357]. A small isoefficiency function means that small increments in the problem size are sufficient for the efficient utilization of an increasing number of processing elements, indicating that the parallel system is highly scalable. However, a large isoefficiency function indicates a poorly scalable parallel system.

Taking into account the different taxonomies of scalability characteristics shown by the different studies mentioned above, and adapting them specifically for SNA methods and techniques, the following sets of measures are proposed to quantify the **degree of volume**, or scalability, for SNA:

1. *Space-Time* ($F_{Volume1}$): maximum number of elements (nodes and/or edges) it is able to process without degrading its performance. The value of this measure will be calculated according to the following scale: *Low* (1/3) if process networks with less than 10.000 of nodes and/or edges; *Medium* (2/3) if process networks with nodes/edges between 10.000 to 100.000; and *Large* (1) if process networks with more than 100.000 nodes/edges. These values have been set to indicate the maximum size of the elements that can be processed by the tool or algorithm (low, medium, high). Of course, these values would be strongly modified in the coming years as processing and storage capacities increase.
2. *Parallelism* ($F_{Volume2}$): capacity of parallel computing. The value of this measure will be calculated according to the following scale: *Low* (1/3) for single or centralized processing; *Medium* (2/3) for distributed processing; and *Large* (1) for parallel processing using Big Data technologies.
3. *Functional* ($F_{Volume3}$): ability to enlarge the system or application by adding new features or extending existing ones. The following scale allow measuring this characteristic: *Low* (1/3) if none of the existing functionalities can be modified or new ones added; *Medium* (2/3) if some existing functionalities can be added or modified; and *Large* (1) if any of the existing functionalities can be added or modified.

4. *Heterogeneous-Integration* ($F_{Volume4}$): ability to integrate or communicate with components or modules from different environments or systems. The value of this measure will be calculated as follows:
   - *Low* (1/3) if cannot be integrated with components or modules from other environments or systems;
   - *Medium* (2/3) if can communicate with some components, but not all, or modules from other environments or systems;
   - *Large* (1) if can be fully integrated with components or modules from other environments or systems.

Previous metrics have been combined to generate a scalability degree (d_{Volume}) for any tool (t) as it is shown in Eq. 4, where scalability ($d_{Volumne}(t)$) is rated from 0 to 1 depending on its capacity to scale to large data sets ($d_{Volume}(t) \in [0, 1]$).

$$d_{Volume}(t) = \frac{\sum_{i=1}^{4} F_{Volumei}(t)}{4} \tag{4}$$

6.3. Information fusion & integration (D$_3$)

OSN popularity is increasing everyday thanks, partially, to the diverse services provided and the target groups that can use them. In this way, it is possible to find OSN for any purpose ranging from personal (such as Facebook, or Twitter), to the professional ones like ResearchGate, or LinkedIn among others. Also these OSN offer a wide variety of services, and it is possible to share photos (Flicker, Instagram,...), videos and music (Facebook, Youtube,...), or micro-blogging (Facebook, Twitter,...), etc. In addition to the service provided by the OSN, all of them allow different kind of interactions between the users (for example in Twitter users can '*follow*' other users, whereas in Facebook they are '*friends*'), and different actions over the content published by the user, i.e., in any OSN anyone can '*like*', '*comment*', and '*share*' the contents published by other users.

With all of this, it is easy to realize the amount, and diversity, of data available in the different OSN that can be used to perform SNA tasks. In order to measure this diversity, we define this dimension called *Information Fusion*, or *Integration*, that tries to answer the question: "*What kind of data can I integrate?*. This measure would be equivalent to the concept of "*Variety*" from the Big Data paradigm. In the case of OSN, this dimension will measure different aspects regarding the data used to perform the SNA tasks. In this case, we have defined three different measurements that will take into account: (1) the number of different type of data (*multichannel*), (2) the number of different OSNs used to extract the data (*multimodality*), (3) and the representation of this data into the model (*multi-representation*).

Following, we provide a detailed description of these three measures and also, a formal definition to measure them:

1. *Multichannel* (F_{Var1}): this indicator measures the diversity of the data taking into account the format of it. In this regard, SNA algorithms are able to extract knowledge by using data from two different sources of information: the graph resulting from the modeling process, and the information related to the members of the SNs, for example the content published by the members of an OSN [358]. In this sense, and regarding OSN, some of the most common types of data that can usually be used to perform SNA tasks are listed below, nevertheless more unfamiliar data types, like sociodemographic data for example, could be included:
   (a) Edges of the graph: this data source is related to the interactions of the different users in the SN.
   (b) Text: in this case, the data used to perform SNA tasks is the content published by the users in text format, for example *tweets, comments*, etc.
   (c) Images: this information is extracted directly from the images, as photos or memes, that users post in any OSN. This data could be the direct analysis of the photo, or the tags that describe the content of the photo, etc [288].
   (d) Video: in any OSN is possible to publish videos, and the different SNA algorithms can take advantage of the video content to extract some valuable information.

   To sum up, F_{Var1} can be defined as the number of different data formats handled by the algorithms or tools. As it is quite difficult to fix the number of data formats due to the evolution of OSN, in this work we are going to limit the value of this characteristic depending on the different number of data types used. In this sense, we are going to consider three different ranges of values for F_{Var1} which are:
   - *Low*: this value reflect the case where the algorithm uses one type of data, i.e: only the edges of the graph, or text, or images, etc. When the algorithm, or system, uses only one type of data, it will have low M_c and its value is 1/3.
   - *Medium*: this value considers the cases where the algorithms, or systems, integrates 2 or 3 different types of data. The value of M_c in this case is 2/3.
   - *High*: when the systems, or the algorithms, are able to integrate 4 or more different types of data, the value for its M_c will be 1.

   For this reason, $M_c \in [1/3, 1]$. It is important to note that this indicator is independent of how data source is modeled, or what kind of information is extracted. Imagine a platform that uses the tweets in text format, the M_c value for this format will be 1/3 independently of the analysis performed with this data.
2. *Multimodality* (F_{Var2}): refers to the number of different data sources (independently of the data format) that can be handled by the algorithm. This indicator takes into account the different number of SNs sources, which are integrated by the algorithm. In this sense, it is quite difficult to define the limits of this indicator due to the huge amount of possible SNs sources (theoretically this value would be a positive natural value, $\mathbb{N}^+$). For instance, regarding OSNs Some examples are the ones created in Social Media Platforms like: Facebook, Twitter, Instagram [288], Youtube, or LinkedIn, to mention just a few of them.

   For this reason, Multimodality is defined in the same way as *Multichannel*, i.e., using three different values that evaluate the different number of OSN taken into account:
   - *Low* ($F_{Var2} = 1/3$): this value represents the systems, or algorithms, that gather data from one OSN.
   - *Medium* ($F_{Var2} = 2/3$: it is used when the data used to perform SNA tasks are extracted from 2 or 3 different OSN.
   - *High* ($F_{Var2} = 1$): this value is used when the system, or algorithm, is able to integrate data from 4 or more OSN.

   With all of that, $M_m \in [1/3, 1]$, where 1/3 means that only one OSN has been used to extract the data, whereas 1 means that the 4 or more OSN have been taken into account. Note that the value of this indicator is not ranking the different OSN but only providing a metric about the number of different OSN used by the platform/algorithm. For example, the M_m value for a platform that is able only to extract data from Youtube will be 1/3; and the value for other platform that integrate data from Twitter and Instagram will be 2/3.
3. *Multi-representation* (F_{Var3}) provides a quantitative measure of the data model representation used by the algorithm, and its value will depend on the complexity of the representation model. It is defined as $M_r \in [1/3, 1]$, where a value of 1/3 means the *basic* representation model, whereas a value of 1, the most *advanced* one. Following, we explain the different representation levels with their corresponding values:
   (a) *Basic model*: this kind of representation correspond to the case when the OSN is modeled into a simple unweighted graph $G = (V, E)$. In this graph, V is the set of nodes of the graph and represents the set of users of the SN; whereas E corresponds to the set of edges of the graph and represents some sort of connection between the users. Note that depending on the OSN taking into account this graph could be directed or

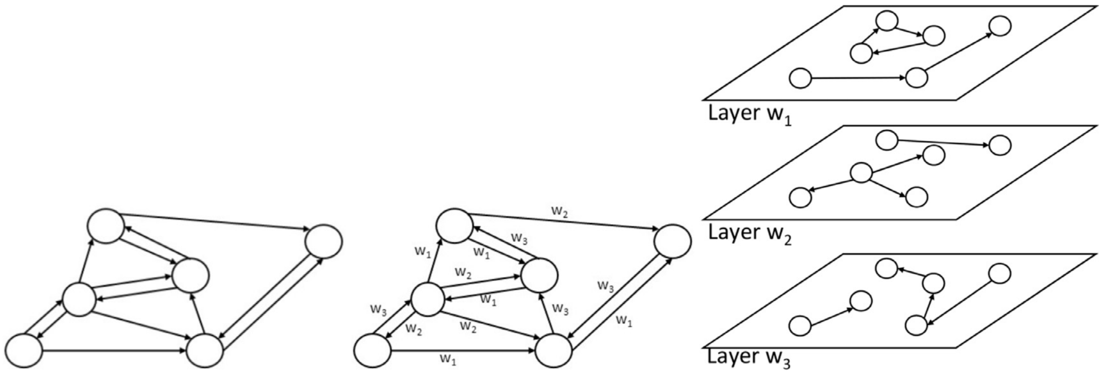


Fig. 6. This figure shows the different representations taken into account into the F_{Var_3} dimension. The different representations are: a) Basic model, b) Intermediate model, and c) Advanced model.

undirected. The value of F_{Var3} in this case is the lower one, i.e., 1/3.

(b) *Intermediate model* corresponds to the value of 2/3. In this case, the problem taken into account is modeled into a simple weighted graph. The resulting graph is quite similar to the previous representation but the edges contain a value providing some kind of information, for example frequency of a specific action, or a meaning of the relation.

(c) *Advanced model*: this level represents the most advanced model used to work with the data extracted from the OSN. In this case, we consider a multi-layer SNA representation as the most advance model. *Multi-layer networks* considers multiple channels of connectivity, it is a representation used to describe systems where the different actors are interconnected by different categories of connections. In this kind of networks, each channel (i.e., each type of relation) is represented by a layer, and the same node may have different set of neighbors in different layers. The value of F_{Var3} when the data is modeled into a Multi-layer network is 1.

Imagine a system that is able to analyze three different types of interactions (F_{Var_i}) between users in an OSN. Fig. 6 shows the different representations taken into account by the multi-representation measure to model this situation. The first one (Fig. 6.a) represents all the information into a unweighted graph, i.e., the *Basic model*. Also, it is possible to use the *Intermediate model* with a weighted graph (Fig. 6.b). In this case, the weighted can be a number to represent, for example, the frequency of an interaction or a label to differentiate the different interactions. Finally, the *Advanced model* is shown in Fig. 6.c, where the data is modeled into a multi-layer graph and each graph contains the information regarding each kind of interaction.

The next step is to define the value of the Information Fusion/Integration dimension, i.e., the **degree of variety**, in terms of the three measures just explained. From these three indicators, we consider that *Multimodality* (F_{Var2}) and *Multichannel* (F_{Var1}) play an important role in terms of Information Fusion/Integration because both measures describe, respectively, the number of different OSN analyzed and the number of different data format taken into account. For this reason, this dimension can be defined as:

$$d_{Variety}(t) = \frac{\sum_{i=1}^{3} F_{Var_i}(t)}{3} \tag{5}$$

Following Eq. 5, those algorithms, and tools, using different types of data extracted from different OSN will perform higher than others using different types of data extracted from the same OSN. Finally, the fraction is used to normalize the value of $d_{Variety}$, thus this dimension is assessed using this degree in a range of [0,1] ($d_{Variety}(t) \in [0, 1]$).

6.4. Visualization (D$_4$)

What can the system show from an OSN is one of the key issues for any researcher, analyst or end-user that works in the domain of SNA. The concept of Visualization is used as a dimension to measure the capacity of the tools, frameworks, and methods to visually represent the information stored in the network [18]. Due to the human brain, it is easier for everybody to visualize large amount of data instead of reading tables or reports (seeing graphics as a complement not a substitute). For this reason, data visualization is a quick and easy way to convey concepts in a universal manner. In order to create good visualizations, one must first decide which questions want to be answered and select compelling visual encodings that depict the data values as graphical features such as position, size or orientation. Although the amount of possible visualization designs is extremely large, statisticians, psychologists, and computer scientists have studied the most suitable representations for different types of data, easing the difficulty of choosing a proper visual encoding. Regarding visualizing OSNs, the most common diagrams found in the literature are: *node-link* and *matrix* diagrams [359,360]. Although, lately, new *alternative* diagrams have been proposed that fall outside any of these two categories [361–363]. An example of these diagrams can be found at Fig. 7, where a node-link diagram (Fig. 7.a), a matrix diagram (Fig. 7.b) and the novel HivePlot diagram (Fig. 7.c) are shown.

The most common visual representation of an OSN is the node-link diagram (see Fig. 7.a), where individuals of the network are represented as dots, and a relation between two individuals is represented by an arrow connecting them. Depending on the relationship, the edge connecting two nodes can be directed or not. Placing the individuals in such a way that the resulting diagram is legible is not a trivial task and many efforts have been made to ensure this. Since the80′s many publications have addressed this task and a plethora of methods have been proposed [364–369]. The main approach followed so far considers the network as a physical system where forces are applied to the individuals moving them around a 2D or 3D plane [370], some examples are the Spring-Electrical or the Stress and Strain models. In the former (Spring-Electrical models), individuals are simulated as positive charges and connections between them as springs. First, all users are allocated in a random position. Then, the physical simulation takes place, individuals repels other individuals around them and the springs avoid separating connected individuals too far apart. The simulation continue for a predefined number of steps or until the system stabilizes. The most famous algorithm is this category is the Fruchterman-Reingold [365]. In the latter, the Stress and Strain models, the positive charges are discarded and only the springs are used. These models, first define the desired spring length. Then, the individuals are allocated at random positions. Contrary to the spring-electrical model, there are no simulation and an "imbalance degree" is defined as the difference between each spring length and

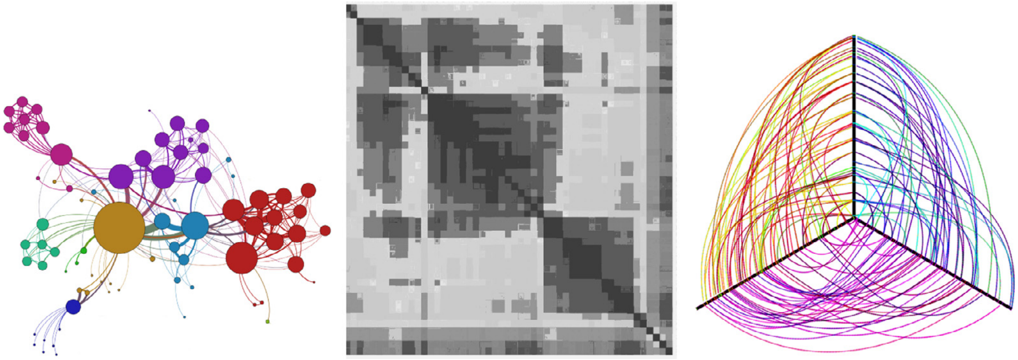


Fig. 7. Subfigure a) shows an example of a node-link diagram; Subfigure b) shows an example of a matrix diagram; and Subfigure c) shows an example of an alternative diagram named HivePlot [361].

the desired spring length. Finally, the "imbalanced degree" is minimized changing the individual's positions. The most famous algorithm in this category is the Kamada and Kawai [366]. Node-link network visualization is an active research field and a wide variety of methods have been proposed to minimize their computational requirements, or to improve their output on bigger networks [367–369].

Although the node-link diagram with the physical systems simulations works very well for a wide number of different networks, they usually have trouble visualizing small-world networks [32]. In a small-world networks, any two nodes are not very far from each other (i.e., there are short path lengths). For that reason, the use of node-link diagrams with the physical system analogy only produces a mess similar to a hairball where it is impossible to see anything. This is a big problem in the area of SNA as, usually, OSN are also small-world networks. Thus, to solve this shortcoming, some alternative representations have been proposed, like the matrix diagrams (see Fig. 7.b). These diagrams represent the networks by their adjacency matrix. The adjacency matrix of a network has all its users as rows and columns, and the value of each cell represents the strength of the connection between the two corresponding individuals. In a matrix diagram, the adjacency matrix of a network is drawn as a picture, where each cell of it is represented as a colored square depending on its strength. This representation has its advantages and disadvantages. On the one hand, all the connections are always visible, regardless of the network size, in the node-link diagram connections cross one another and, sometimes, it can be impossible to distinguish them. On the other hand, identifying paths in the network is harder than ordering the rows and columns of the adjacency matrix, this process is called *Seriation*, which is essential to unveil the inner structure of a network. *Seriation* is not a trivial task, in fact, it is an active research area where a large number publications are coming out to tackle this problem [371].

There is not a consensus about what is the best strategy for visualizing networks as both of the aforementioned approaches have their advantages and disadvantages. In addition, some hybrid approaches have also been proposed like, MatrixExplorer [372] or Matlink [373]. Finally, within this debate, there are some scientific groups that are proposing alternative diagrams, tailored for specific tasks, that deviates from the mainstream. For example, in the Biology area, we can find: HivePlots [361] a type of diagram created for visually comparing different networks. In HivePlots, nodes are mapped to and positioned on radially distributed linear axes and edges are drawn as curved links; or BioFabric [362] a type of diagram created so Edges are unambiguously represented and never overlap. In BioFabric nodes are depicted as horizontal lines, not as points and edges are depicted as vertical ones; Another example is PivotGraph [363], that is specialized in multivariate data.

How the network is represented is not the only characteristic used to compare SNA visualization tools, but there are other characteristics that need to be taken into account [370,374,375]. These characteristics are: "volume", "summary" and "interaction". The first one, volume, asses the raw processing capacity of a tool, how many users/nodes and connections/edges can be handled by the tool. Although the capacity of visualizing millions of nodes does not imply the generation of outputs in such a way any human can use. Most of the times, visualizing a medium-size network, without any extra help, provides as a result an unreadable hairball. In this situation the second characteristic, name "summary" becomes useful. "Summary" evaluates the capacity of a tool to generate simplifications that allows to reduce the complexity of the network while maintaining as much information as possible. This balance, between the summary and the raw data, must not be static and should adapt to the user's requirements. The third and last characteristic, "interaction", measures the capacity of a tool to adapt its graphical output to the user's needs. These three characteristics together allow us to fulfill the Shneiderman's visualization mantra: "Overview first, then zoom and filter details on-demand" [376]. Therefore, these three characteristics have been considered as the essential ones that any SNA visualization tool should have.

Although these three characteristics are great for describing and comparing SNA tools, two of them, "volume" and "summary", are highly correlated with the other three dimensions presented in this work (Scalability, Pattern & Knowledge discovery, and Information Fusion & Integration). Therefore, we have decided to move away from the approach used in the literature and remove any aspect not related to graphics from the visualization dimension. Hence, we will describe the visualization dimension in a twofold way. On the one hand, we will use the aforementioned "interaction" characteristic. On the other hand, the lack of consensus over what is the most suitable diagram to visualize an OSN makes impossible to use it as a comparison method. Thus, we have considered that all the aforementioned diagrams are equally valuable when visualizing an OSN. Therefore, we will introduce an extra characteristic, named "*Visual Variables*", that will help us to asses the information representation capacity of a tool. *Visual Variables* were proposed in [377] by Jacques Bertin based on his experience as a cartographer and geographer. In his work he described the *visual variables* as the fundamental way in which graphic symbols can be distinguished. The author identified the Visual Variables listed below and, according to his work, it can be used for four different purposes: (1) Selective (easily distinguish between groups); (2) Associative (identify changes among the same group); (3) Ordered (allows to identify sequences); and, (4) Quantitative (allows to compare numerical values). The visual variables can be summarized as follows:

1. **Position** ($F_{VisVar1}$): refers to the location of an object in the image, position is one of the most versatile variable as it can be used in as a Selective, Associative, Ordered or Quantitative variable. For example, if an OSN has geo-localization data, the position of an user can display his/her country.

2. **Size** ($F_{VisVar2}$): refers to the size variation of an object. It can be used as: Selective, Ordered or Quantitative variable. For example, changing the size of an user node depending on his/her degree.
3. **Shape** ($F_{VisVar3}$): refers to the different geometrical shapes an object can have (triangles, rectangles, circles...etc.). It can only be used as a Associative variable. For example, showing users from different OSN with a different shape.
4. **Orientation** ($F_{VisVar4}$): refers to the rotation an object presents. It can be used as a Selective or Associative variable. For example, using an arrow to indicate the direction of a following/follower relation.
5. **Color** ($F_{VisVar5}$): refers to the color hue of an object. It can be used as a Selective or Associative variable. For example, using the same color for all the users in the same community.
6. **Saturation** ($F_{VisVar6}$): refers to color saturation, the brighter or lighter a color hue is. It can be used as a Selective, Ordered or Quantitative variable. For example, showing users with a higher centrality colored with a more saturated color.
7. **Texture** ($F_{VisVar7}$): refers to the fill pattern of an object. It can be used as a Selective or Associative variable. For example, using two different patterns to identify members of different political parties.

Previous variables are suitable to act as measures of the visualization dimension. Instead of trying to evaluate the quality of a visualization method by the type of representation used, we will evaluate it by the number of visual variables used by a tool to enrich them. The more visual variables a tool is able to handle, the more extra information it can visually represented, and the higher the visualization degree will have the tool. However, as it has been stated by the literature, being able to represent a lot of information does not implies the generation of good quality visualizations. That is why the "interaction" capabilities of a tool are also being considered when calculating a visualization degree of a tool. The most common interactions analyzed in the literature can be summarized with the next five actions:

1. **Zoom** (F_{Inter1}): refers to the action of changing the level of details of the elements being shown but maintaining the same number of them. This does not implies that all elements need to be visible in the screen at once. For example, using the mouse wheel to zoom in/out a graph.
2. **Filter** (F_{Inter2}): contrary to zoom, refers to changing the number of elements being displayed but maintaining the same level of detail. For example, dragging the mouse to pan over a diagram.
3. **Highlight** (F_{Inter3}): refers to the action of emphasizing some particular elements of a set. For example, highlighting the neighbors of an user when the mouse is positioned over it.
4. **Grouping** (F_{Inter4}): refers to the action of replacing a group of elements with a simplification that maintains all or some of the properties of the group. For example, joining all the users of the same country into a single node.
5. **Multiview** (F_{Inter5}): refers to the action of switching between multiple representations of the same data. For example, switching between the node-link and matrix representation of an OSN.

In order to generate a single value capable of evaluating the degree of the visualization dimension, Eq. 6 is proposed. Where F_{Visual} is the number of visual variables a tool can handle, and F_{Inter} is the number of interactions available on a tool, whereas α, β, γ, θ are weights that represent the importance given to each characteristic.

$$d_{Visual}(t) = \alpha \cdot \frac{\sum_{i=1}^{7} \gamma_i \cdot F_{VisVari}(t)}{7} + \beta \cdot \frac{\sum_{j=1}^{5} \theta_j \cdot F_{Interj}(t)}{5} \tag{6}$$

In this work, all of the visual characteristics (visual variables and interactions) will have the same weight (so α and β will be set up to 0.5), and all of the features for each characteristic will have the same importance (therefore γ and θ will be set up to 1.0).

6.5. Summary on social network analysis dimensions

Following the previous subsections, the four basic research questions (RQ) proposed in the introduction, and their related dimensions, have been mapped into a set of measures, or $SNA_{degrees}$, which can be used to provide a quantitative value for each dimension. These research questions, together with the proposed dimensions, and the metrics, or degrees, defined to assess them are shown in Table 1.

Fig. 8 shows an example of an hypothetical algorithm, framework and a tool, and how the proposed dimensions, can be used to represent the technology maturity to work with OSN data. We have decided to use a spider, or radar, diagram for representing the evaluation of the SNA technology. As can be seen in this figure, the axes will be used to represent each dimension, so the quality, strengths or weaknesses for each method, tool or framework can be easily analyzed.

Finally, and considering these $SNA_{degrees}$ it is quite straightforward to define a new **global metric**, that we have called $\mathfrak{C}_{SNA}$, to represent the "*Capability*" and power to work with OSN sources, to later use as a metric to rank the technologies analyzed. We calculate the value of $\mathfrak{C}_i$, where i represents the number of dimensions to be considered, as the area contained in the irregular polygon defined by the i dimensions used in our previous representation (see Fig. 8). This equation comes from the Shoelace formula, also known as Gauss's area formula and the surveyor's formula [378], and it is a simple formula for finding the area of a polygon given the coordinates of its vertices, see Eq. 7, where A is the area of the polygon, n is the number of sides, and (x_i, y_i) are the vertices of the polygon.

$$A = \frac{1}{2}\left|\sum_{i=1}^{n-1}(x_i y_{i+1} + x_n y_1) - \sum_{i=1}^{n-1}(x_{i+1} y_i - x_1 y_n)\right| \tag{7}$$

Therefore, the $\mathfrak{C}_i$ metric is a particular case of previous Equation with $n = 4$, and it can be mapped as Eq. 8 shows. Taking into account that all of the dimensions defined are represented in an axis, all of the vertices will have at least one of the coordinates (x or y) to 0, due all of the dimensions defined in this work have been normalized to 1, implies that $\mathfrak{C}_i \in [0, 2]$. Considering: $d_{Value} = (x_1, 0)$, $d_{Volume} = (0, y_2)$, $d_{Variety} = (x_3, 0)$, $d_{Visual} = (0, y_4)$, Eq. 8 can be easily simplified as Eq. 9 shows.

$$\mathfrak{C}_4(t) = \frac{1}{2} \cdot \left| \sum_{i=1=Value}^{3=Variety} \left(d_i^{x_i}(t) \cdot d_i^{y_{i+1}}(t) + d_i^{x_n}(t) \cdot d_i^{y_1}(t) \right) - \sum_{i=1=Value}^{3=Variety} \left(d_i^{x_{i+1}}(t) \cdot d_i^{y_i}(t) - d_i^{x_1}(t) \cdot d_i^{yx_n}(t) \right) \right| \tag{8}$$

$$\mathfrak{C}_4(t) = \frac{1}{2} \cdot \left| \left(d_{Value}^{x_1}(t) + d_{Volume}^{y_2}(t) \right) \cdot \left(d_{Variety}^{x_3}(t) + d_{Visual}^{y_4}(t) \right) \right| \tag{9}$$

Analyzing Eq. 9, two different (extreme) cases could appear: (1) if two of the dimensions are equal to zero (i.e., $d_{Value} = d_{Volume} = 0$, or $d_{Variety} = d_{Visual} = 0$), C_4 will be equal to 0, in such case the $\mathfrak{C}_4$ must be calculated as Eq. 10 shows (this is equivalent to make a projection of both non-zero dimensions into two different axis). (2) If three dimensions are zero, there's no possibility to calculate any area due there's only one dimension with a value, in such case the SNA_{degree} should be used to evaluate the capability (only in that dimension) of the SNA technology. Fig. 8 shows an example of our Capability metric on three hypothetical frameworks).

$$\mathfrak{C}_4(t) = \frac{1}{2} \cdot \left| d_i^{x_i}(t) \cdot d_j^{y_j}(t) \right|, d_i, di \neq 0 \tag{10}$$

Following our previous example, the three different SNA technologies previously presented in Fig. 8 would be ranked using our metric (see Table 2). Although this comparison would not have sense (as we are comparing different kinds of technology), it shows how this general

Table 1
Summary on dimensions and the quantitative measures (degrees) defined.

Research Quest.	Dimension (D_i)	Degree ($d_v{}^*$)	Range
What can I learn?	Pattern & Know. discovery (D_1)	$d_{Value}(t) = 1/3 \cdot (\frac{\sum_{i=1}^{2} F_{Val(1,j)}(t)+F_{Val(3,j)}(t)}{2} + \frac{\sum_{i=1}^{3} F_{Val(2,i)}(t)}{3})$	[0,1]
What is the limit?	Scalability (D_2)	$d_{Volume}(t) = \frac{\sum_{i=1}^{4} F_{Volumei}(t)}{4}$	[0,1]
What kind of data can I integrate?	Information Fusion & Integration (D_3)	$d_{Variety}(t) = \frac{\sum_{i=1}^{3} F_{Vari}(t)}{3}$	[0,1]
What can I show?	Visualization (D_4)	$d_{Visual}(t) = 1/2 \cdot \frac{\sum_{i=1}^{7} F_{VisVari}(t)}{7} + 1/2 \cdot \frac{\sum_{j=1}^{5} F_{Interj}(t)}{5}$	[0,1]

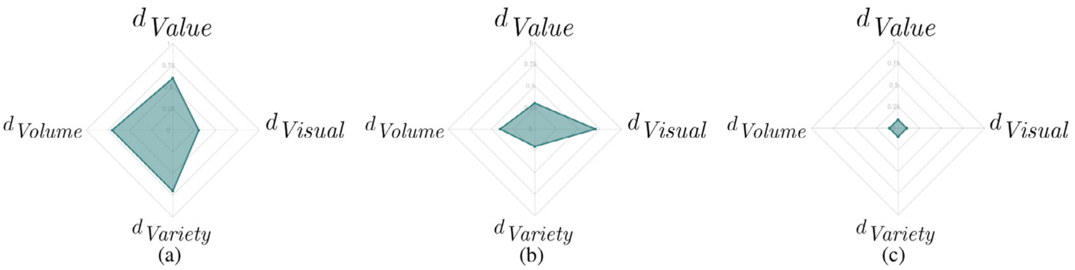


Fig. 8. Spider diagrams representing the evaluation of three (hypothetical) SNA platforms by using the 4 dimensions described in this work. The exact values for each dimensions for these 3 hypothetical frameworks are shown in Table 2. The figures shown correspond to a good platform (a), an intermediate tool (b) and a poor platform (c).

Table 2
Example of $\mathfrak{C}_4$ metric application over three different hypothetical SNA tools and frameworks.

Rank	$\mathfrak{C}_4(t)$	Dimensions				SNA technology
		d_{Value}	d_{Volume}	$d_{Variety}$	d_{Visual}	
1	0.315	0.6	0.7	0.7	0.3	Tool 1 (high value)
2	0.170	0.3	0.4	0.2	0.7	Framework 1 (medium value)
3	0.010	0.1	0.1	0.1	0.1	Tool 2 (low value)

metric can be used to better understand the capability of a particular SNA technology. This new metric will be used in next sections to provide a ranking between the analyzed SNA tools and frameworks.

7. Frameworks & tools analysis

7.1. Tools analyzed

This section provides a review on a set of popular SNA frameworks and tools, which are extensively used by the research community and industry. Due, it would be highly difficult to analyze all of the currently available tools, we have selected for analyzing a subset of 20 representatives. In this set, it can be found computing libraries, web applications, distributed applications, or desktop applications among others, we will refer to all of them as *SNA-software*. Initially, a list of 70 SNA-software candidates was generated using Github[10], KDnuggets[11,12] and the Infovis-Wiki[13] websites. It is important to highlight that we have considered SNA-software with both types of licenses, open source and commercial (proprietary). From the list of 70 SNA-software candidates, a set of 20 tools was selected for analysis. In this selection, it was considered the type of software license, the quantity and *quality* of the software documentation, and its current impact among the community (taking into account the popularity of some tools in published works, websites and other technical material). Following, the list containing the 20 different SNA-software analyzed in this work is shown. For each software, we provide the name, a bibliographical reference (or a website), a brief description and also, its license type:

1. **Igraph** [379]: a collection of network analysis tools with the emphasis on efficiency, portability and ease of use. License: MIT.
2. **AllegroGraph** [380]: an ultra scalable, high-performance, and transactional Semantic Graph Database. License: Proprietary.
3. **LaNet-vi** [381]: a large networks visualization tool. It provides images of large scale networks on a two-dimensional layout. License: AFL.
4. **Stanford Network analysis Platform (SNAP)** [382]: a general purpose, high performance system for analysis and manipulation of large networks. License: BSD.
5. **ORA-LITE/PRO**[14]: a dynamic meta-network assessment and analysis tool developed by CASOS at Carnegie Mellon. It contains hundreds of OSNs, dynamic network metrics, trail metrics (path-based metrics), and procedures for grouping nodes. License: Proprietary.
6. **Network workbench** [383]: a Large-Scale Network analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research. License: open-source.
7. **NetMiner**[15]: a premium software tool for Exploratory analysis and Visualization of Network Data. License: Proprietary.
8. **Circulo** [384]: a “Community Detection” Evaluation Framework written primarily in Python. License: Apache 2.0.
9. **Cytoscape** [385]: a software platform for computational biology and bioinformatics, useful for integrating data, and for visual-

[10] http://github.com/briatte/awesome-network-analysis.
[11] http://kdnuggets.com/2015/06/top-30-social-network-analysis-visualization-tools.html/2.
[12] http://kdnuggets.com/software/social-network-analysis.html.
[13] http://infovis-wiki.net/wiki/Main_Page.
[14] http://casos.cs.cmu.edu/index.php.
[15] http://netminer.com.

izing and performing calculations on molecular interaction networks. License: LGPL.

10. **JUNG** [386]: a software library that provides a common and extendable language for the modeling, analysis, and visualization of data that can be represented as a graph or network. License BSD.
11. **SparklingGraph** [387]: a Cross-platform tool to perform large-scale, distributed network computations with Apache Spark's GraphX module. License: BSD 2.
12. **NetworkX** [388]: a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. License: BSD.
13. **Pajek** [389]: is a Windows program for analysis and visualization of large networks. License: Unknown.
14. **GraphX Apache Spark** [349]: a module to perform graph-related parallel computation. License: Apache 2.0.
15. **Gephi** [390]: an open-source network analysis and visualization software package written in Java on the NetBeans platform. License GPL3.
16. **UCINET** [391]: a software package for the analysis of OSN data. It comes with the NetDraw network visualization tool. License: Proprietary.
17. **Prefuse** [392]: a Java-based toolkit for the interactive creation of information visualization applications (not only for graphs, but also tables and trees). License: Unknown.
18. **Graphistry**[16]: a cloud service that automatically transforms your data into interactive, visual investigation maps built for the needs of analysts. License: Proprietary.
19. **GraphViz** [393]: a open source graph visualization software. License: CPL.
20. **Neo4j** [394]: an Open source, scalable graph database. License GPL3.

Once the list of SNA-software has been briefly described, each software has been evaluated using the different metrics proposed in Section 6. The evaluation process carried out in this work follows a "top-down" approach. First, the global capability metric ($\mathfrak{C}_4$) is analyzed for each framework and tool. In a second step, we have analyzed in detail the different dimensions (i.e., the $SNA_{degrees}$) that compose the global metric. The evaluation process for each software was carried out as follows: once the initial set of tools was selected, the authors agreed an evaluation rubric (that is publicly available, see Appendix) to assess the tool. This rubric is based on the analysis of the software documentation, their official websites (or any related site that could store relevant information), and other published works that provide technical details about these tools. From this technical documentation, we assess each of the characteristics that form the different $SNA_{degrees}$, to finally obtain a quantitative value for each of the proposed dimensions. The features used in this rubric (strictly) follows the characteristics proposed to measure the four dimensions (see Sections from 6.1 to 6.4), so from these features we can obtain a quantitative value for each degree. Although the authors have previous experience in several of the analyzed tools (such as Igraph, Circulo, JUNG, or Gephi), it is not possible to download, install, and generate experimental datasets and evaluations for each SNA-software. For this reason, we decided to carry out the evaluation of the software following the previous process.

7.2. Analysis by dimension

First, the results obtained by the 20 tools in each dimension are analyzed. The goal of this analysis is twofold. On the one hand, it allows us to understand the strengths and weaknesses of the different SNA-software. This analysis will help any researcher who is looking for a SNA tool to select the one that best fit to his/her requirements. The type of license used by the SNA-software (public, open-source, BSD, MIT, proprietary,etc.) can be a determining factor for future research, or the development of new products. Therefore, our analysis will take into account this feature to differentiate between those types of software. On the other hand, the analysis of the disaggregated values allows us to understand the opportunities, and weaknesses, of the tools from the SNA point of view. In this sense, this disaggregated analysis will allow us to understand if the requirements for a specific dimension (e.g., visualization) are currently fulfilled by the SNA-software available, or if there is any specific dimension that requires from some special reinforcement. Therefore, this second analysis will help the reader to detect spaces for improvement in some particular research areas related to SNA.

Table 3
Top-5 best SNA tools under Proprietary or Open-Source, and Only Open-source, licenses.

Proprietary or Open-Source					
Tool	$\mathfrak{C}_4(t)$	d_{Val}	d_{Var}	d_{Vol}	d_{Vis}
Graphistry	0.67	0.33	1.0	1.0	1.0
Neo4j	0.57	0.48	0.56	1.0	1.0
ORA-LITE/PRO	0.56	0.84	0.67	0.5	1.0
NetMiner	0.52	0.65	0.89	0.67	0.69
Cytoscape	0.39	0.67	0.33	0.58	0.93
	Only Open-Source				
Tool	$\mathfrak{C}_4(t)$	d_{Val}	d_{Var}	d_{Vol}	d_{Vis}
Neo4j	0.57	0.48	0.56	1.0	1.0
Cytoscape	0.39	0.67	0.33	0.58	0.93
Gephi	0.35	0.35	0.44	0.66	0.93
Pajek	0.31	0.48	0.56	0.50	0.73
JUNG	0.28	0.41	0.33	0.75	0.64

Regarding the *global metric* score ($\mathfrak{C}_4(t)$), Table 3 shows the top-5 "Proprietary or Open-source", and "Only Open-Source" SNA-software. In addition to the aforementioned table, Fig. 9 is presented. This figure contains the distribution of the dimension scores in order to allow the reader to contextualize the values presented. The tool that provides the best global score is *Graphistry*. This tool has the highest possible scores in *Information Fusion, Scalability*, and *Visualization*. However, the score on *Knowledge Discovery* is below the average, as a matter of fact, *Graphistry* is the 13th tool in that category. Regarding open-source tools, the best one is *Neo4j*. This tool has an above-average *Knowledge Discovery* score, an average *Information Fusion* score, and the maximum possible scores for *Scalability* and *Visualization*. Note that the scores obtained by the tools in the top 5 differ between dimensions. This means that there is not a single tool that dominates all the others for all the dimensions. Furthermore, nearly all the tools analyzed, 15 out of 20, have achieved a position in one of the top 5 proposed.

Fig. 10 shows the spider, or radar, diagrams for the top-3 best SNA-software analyzed. The upper row, composed of sub-figures a, b and c, corresponds to those tools under "Proprietary or Open-Source" licenses, whereas the second row (sub-figures d, e and f) corresponds to those frameworks or tools under the "Only Open-source" license. The corresponding values for each dimension are shown in Table 3.

In order to further evaluate this phenomenon, the top 5 SNA-software tools for each dimension are shown in Table 4. Starting with the *Knowledge Discovery* dimension. The best performing SNA-software in this dimension is ORA-LITE/PRO and it scores better than the second-best tool, SNAP. This makes ORA-LITE/PRO an *outlier* regarding the *Knowledge Discovery* dimension (see Fig. 9). A similar case can be found for the *Information Fusion* dimension, where Graphistry tool provides the highest possible score (see Fig. 9). In this case Graphistry also scores better than the second-best tool, Netminer. Contrary, the *Scalability* dimension shows three equally good tools as top-performing (Graphistry, AllegroGraph and Neo4j), followed by GraphX Apache Spark and SparklingGraph. Finally, a similar case can be found on the *Visualization* dimen-

[16] http://graphistry.com.

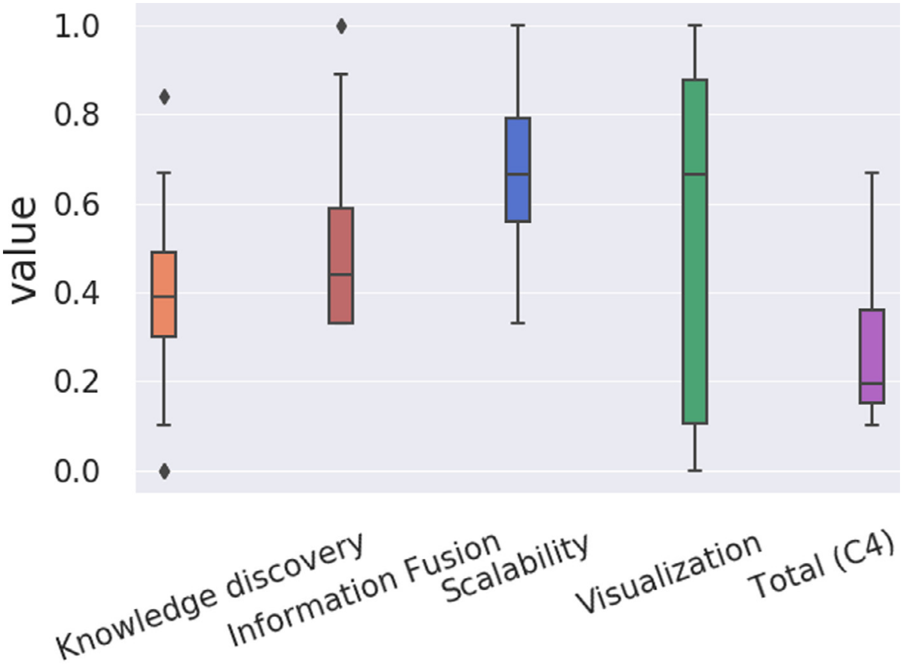
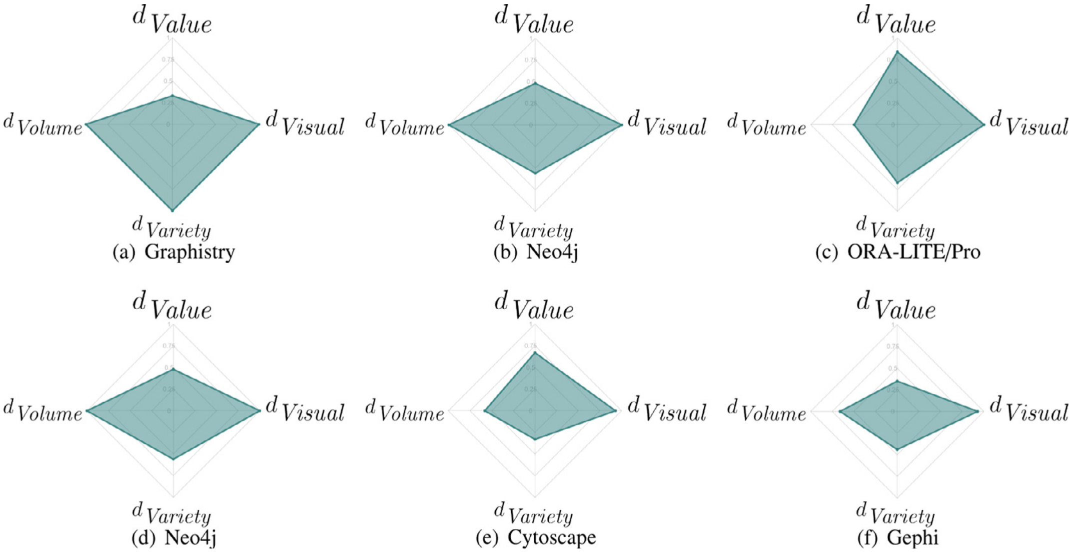


Fig. 9. Distribution of the different dimensions values achieved by the analyzed tools. The X-axis represents the proposed dimensions, whereas the Y-axis shows the values obtained once the quantitative metrics (SNA_{degree}) are calculated.



(a) Graphistry (b) Neo4j (c) ORA-LITE/Pro

(d) Neo4j (e) Cytoscape (f) Gephi

Fig. 10. This figure shows the top 3 SNA software with “Proprietary or Open Source” licenses (first row: Graphistry, Neo4j and ORA-LITE/Pro), and “Only Open Source” licenses (second row: Neo4j, Cytoscape and Gephi).

Table 4
Top 5 SNA-software by dimension.

Dimension	SNA-software	Score	Dimension	SNA-software	Score
Knowledge Discovery	**ORA-LITE/PRO**	**0.84**	**Scalability**	**Grasphistry**	**1.00**
	SNAP	0.67		**AllegroGraph**	**1.00**
	Cytoscape	0.67		**Neo4j**	**1.00**
	NetMiner	0.65		GraphX Apache Spark	0.92
	NetworkX	0.52		SparklingGraph	0.92
Information Fusion	**Grasphistry**	**1.00**	**Visualization**	**ORA-LITE/PRO**	**1.00**
	Netminer	0.89		**Grasphistry**	**1.00**
	Network Workbench	0.67		**Neo4j**	**1.00**
	ORA-LITE/PRO	0.67		Gephi	0.93
	Pajek	0.56		Cytoscape	0.93

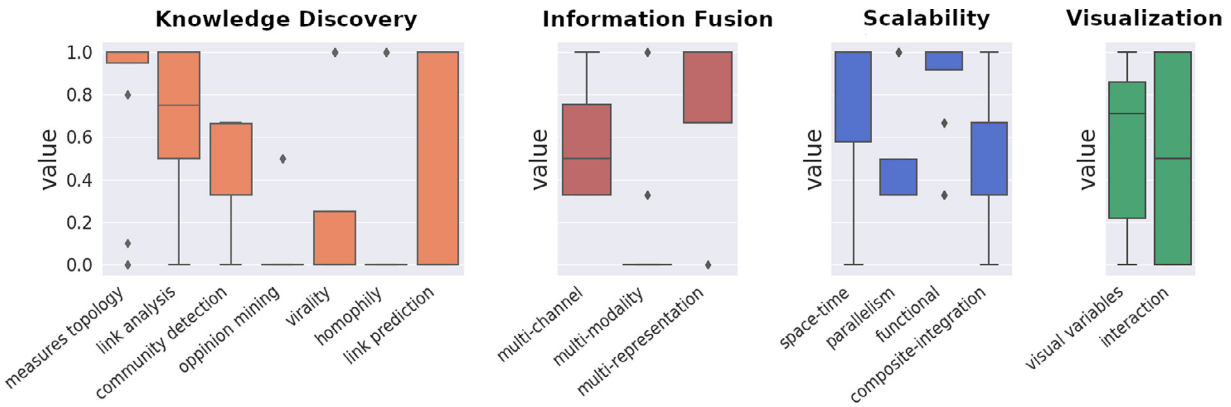


Fig. 11. Un-aggregated dimension distributions for the top-5 SNA-software tools, each color represents a different dimension. The X-axis contains each of the features that form a dimension (SNA_{degree}), and the Y-axis their numerical value.

sion with the top tools (ORA-LITE/PRO, Graphistry and Neo4j) followed by Gephi and Cytoscape.

The analysis so far has shown that, the ORA-LITE/PRO and Graphistry tools, stand out from the rest in the *Knowledge Discovery* and *Information Fusion* dimensions. Below, the characteristics (or features) that have made these tools to stand out will be analyzed. To do so, each dimension has been split into its basic features (see Table 1, and sections from 6.1 to 6.4). Fig. 11 shows the distribution of each of the features that forms a dimension. Notice that in the *Knowledge Discovery* dimension, the *Opinion Mining* and the Homophily features are composed mostly by tools that have achieved a score of 0, and only a few of them have been able to achieve a higher score. Something similar happens with the *Virality* feature but in a less acute way. The tools that work with those features are the ones that appear in the top 5. Moreover, ORA-LITE/PRO is the only tool that has achieved a score (greater than 0) in every feature of the *Knowledge Discovery* dimension.

A similar case can be found when we analyze the *Information Fusion* dimension. In the *Multi-Modality* feature most of the cases are 0, and only a few are able of scoring something in this feature. In fact, Graphistry is the only tool that has scored a maximum rating in all of the features of the *Information Fusion* dimension. In general, a similar case can be found on the *Scalability* dimension. However, contrary to the *multi-modality* feature, several tools have achieved the maximum value in this category and not only one. Actually, all the tools that appear in the top 5 in that dimension have achieved a maximum score on that feature. Finally, the *visualization* dimension is the most homogeneous. Nevertheless, the *Visual Variables* feature scores slightly higher than the *Interaction* ones.

To sum up, we have noticed an uneven distribution on the 4 dimensions features. There are some features were nearly all tools score good, while in others only a few are able to obtain some scoring. This makes those tools stand out over the others. For example, the *Measures Topology, Link analysis* or *Funcional* features have high values for nearly all the tools analyzed, specially the *Measures Topology* one. Contrary, features like *Opinion Mining, Homophily* or *Multi-Modality* are tackled by very few tools. These last features can be used as a foundation of the guidelines that the next iteration of SNA tools must follow.

7.3. Relationships between dimensions

Up to now, the 4 dimensions have been analyzed independently. However, it is possible that exist different correlations, or relations, between the dimensions. Analyzing how the different dimensions related to each other could show areas that tools usually don't tackle (like extracting knowledge from different sources) and identify research niches. For example, a proportional relationship between *Visualization* and *Knowledge Discovery* might be observed. It can be hypothesized that the more knowledge a tool can extract, the more visualization capabilities should it have (in order to process it). A similar hypothesis could be drawn between the *Visualization* and the *Information Fusion* dimensions following the same logic. Furthermore, a proportional relation could be also expected between the *Information Fusion* and the *Knowledge Discovery* dimensions. Since, the more types of information a tool is able to handle, the broader the type of possible analysis will be, and therefore the further the knowledge extraction will get. Contrary, an inversely proportional relationship is to be expected between the *Scalability* and the *Knowledge Discovery* dimensions. Fancier knowledge extraction techniques, usually require from a high amount of computational power. Implementing these methods in a scalable way is a daunting task that is currently being tackled by the scientific community. Finally, the same logic could be applied between the *Scalability* and the *Information Fusion* dimensions.

Therefore, and in order to explore these relations, we have represented the (pair to pair) relations between all of the dimensions for the SNA-software tools analyzed. The results are shown in Fig. 12, this figure plots two different dimensions in the X and Y axis, and uses dots to represents the analyzed tools. The intensity of the dots indicates the number of tools that have scored a particular tuple of values. A pink dot indicates that a particular tools is a member of the Pareto Front of a graph. The members of the Pareto Front have been labeled with its tool name. Finally, to help the evaluation of the tool distributions, two colored vertical/horizontal lines have been added. These lines indicate the mean obtained for each particular dimension. Two types of graph are present in the matrix. On the one hand, there are the graphs where one tool dominates all the others. You can see this effect in *Scalability vs. Information Fusion, Visualization vs. Knowledge Discovery, Visualization vs. Information Fusion*, and the *Visualization vs. Scalability* graphs. On the other hand, there are the graphs where that dominance is not present. Notice the *Information Fusion vs. Knowledge Discovery* and the *Scalability vs. Knowledge Discovery* graphs.

Regarding the hypothesized relations, the data shows that the proportional relation between *Visualization* and *Knowledge Discovery*, and *Visualization* and *Information Fusion* is present in the data. Notice how in their respective graphs there is a higher condensation of points in the top right corner of the plot and how the number of dots grow from left to right and from top to bottom The points that have a zero value in the *Visualization* dimension have been ignored as those tools do not implement any kind of visualization. However, the hypothesized proportional relation between *Information Fusion* and *Knowledge Discovery* is not present in the data. Notice how the several fronts presented in the graph decreases from left to right. This has surprised the authors and could be interpreted as a future research niche. The data, a priory, shows that collecting extra information types and sources are not being fully exploited by the knowledge extraction algorithms. Concerning the hypothesized inversely proportional relations between the *Scalability* and the *Knowl-*

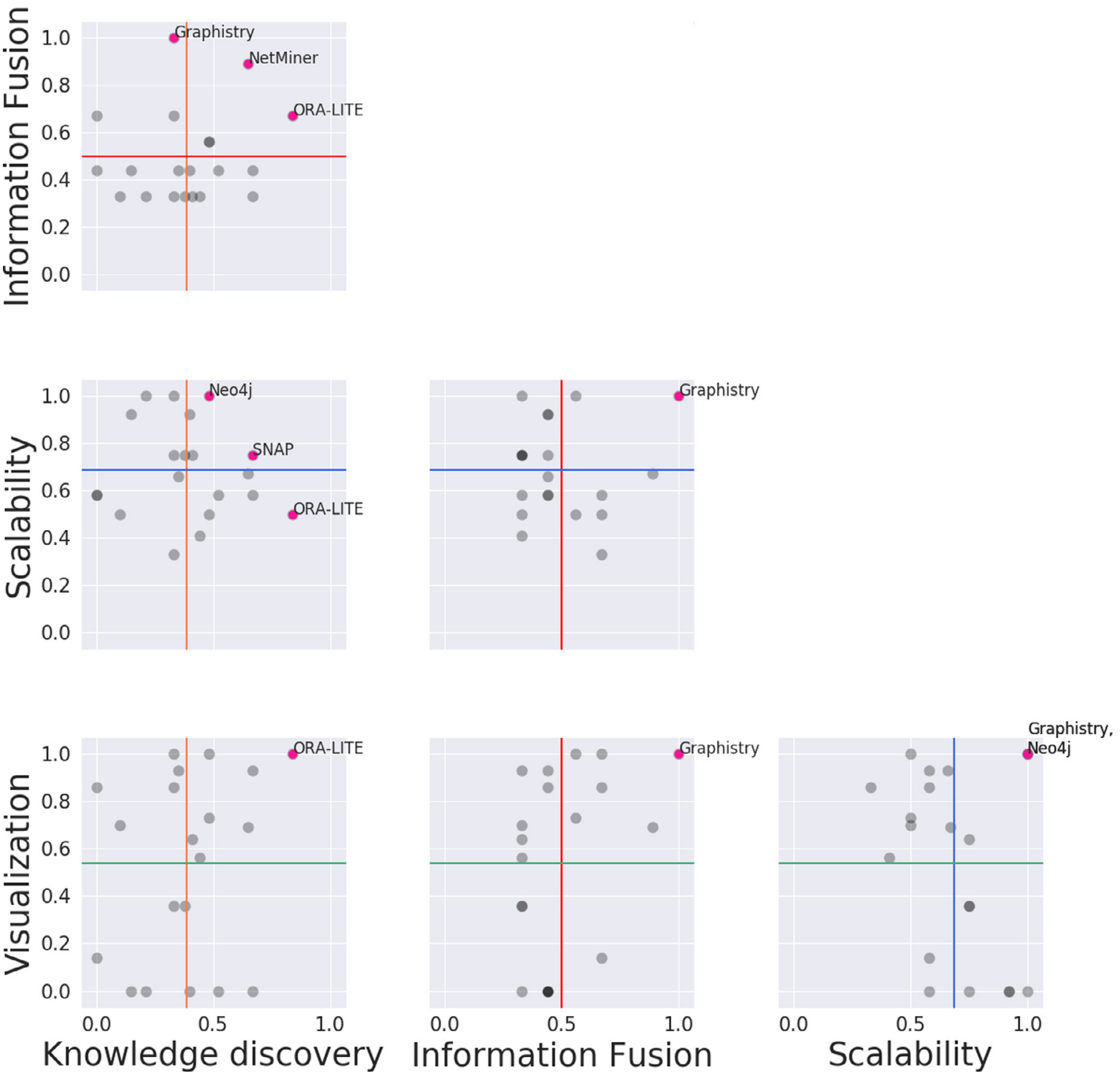


Fig. 12. Pairing plot graphics, of the dimensions proposed for SNA-software. Each cell contains a graph where two different dimensions are depicted on the X and Y axis, and each point in the graph represents a tool. The horizontal and vertical lines show the average value at each dimension. The darker a point, the more tools have scored those values. Finally, the pink dots represent the Pareto front of a set, the members of the Pareto front are labeled with their names.

edge Discovery dimensions, and the *Scalability* and the *Information Fusion* dimensions, the data have shown that both are present. Notice that the dominance fronts of both graphs decrease from left to right. This effect is far more acute between the *Scalability* and the *Knowledge Discovery* dimensions. Nevertheless, the *Scalability/Information Fusion* graph also shows a particular pattern not expected by the authors. The far dominance of *Graphistry* of every other tool in that graph is surprising. In fact, there are only two tools on the top right sector of the graph including *Graphistry*, and the other one is near the mean line. In the authors opinion this huge gap is another research niche, especially taking into account that *Graphistry* is a commercial tool and no other open-source one is near to its capabilities.

8. Conclusion, future trends and challenges

In the last years, we have witnessed the increasing relevance of OSNs in our daily life. This popularity is produced by the high number of users that participate in these social media platforms every day and this participation results in huge amount of data generated by user interactions. OSN popularity and its exponential growth have led to an enormous interest in the analysis of this type of networks. There is a plethora of issues that can be studied from social media data, like the interconnections that originate the network, the structure, the evolution of the network, the identification of user communities, how the information flow and how this information is disseminated, or the patterns that can be extracted from them, just to mention a few of them. As a consequence of becoming a hot research area, the number of papers, conferences, journals, algorithms and tools has risen exponentially. This growth makes almost impossible to analyze in detail the current state of the art related to SNA. In order to limit the state of the art, and to present relevant review of the different research papers published in this topic, we have performed a scientometric study to define the most relevant research areas. From this analysis several fundamental research (graph theory and network analytics, community detection algorithms, information diffusion models, text mining and topic extraction, opinion mining and sentiment analysis), and application domains (health, marketing, tourism and hospitality, cyber-crime and cyber-terrorism, politics, detection of fake news and misinformation, and finally multimedia), have been studied in Sections 3 and 4.

Any analysis of OSNs rely on the representation of the data as a graph, and then some algorithms are applied to extract some valuable knowledge. In this sense, we have reviewed the basics of SNA techniques and algorithms in Section 3. The different algorithms are categorized around what kind of information is used in the analysis. In this sense, any researcher can extract information from the resulting graph, performing a structural-based analysis, or he/she can perform the analysis based on the content published in the OSN (content-based analysis). On the one hand, using the network structure any researcher can apply algorithms to detect the communities that compose the graph, or to study how the information is propagated through the network. On the other hand, using the content of the information published the standard analysis techniques rely on solving the problems of topic extraction, opinion mining and sentiment analysis. Finally, we have also analyzed the current state of the art regarding multimedia content; i.e., how the different media (text, audio and video) is used to perform SNA.

In spite of the previous state of the art analysis, and due to the size of the research community around SNA, it is quite difficult for a novel researcher, or someone (even experienced) who wants to start his/her research on this area, to select the most appropriate tool or algorithm. In order to offer a starting point to the community, we have proposed in this work four research questions that any researcher should answer, or at least keep in mind, before starting his/her research in this area. From these research questions it have been proposed several dimensions, as a tool to obtain a quantitative assessment of the current maturity of SNA technologies, allowing both to better understand what are the main strengths and weaknesses of these technologies, and to look for future trends and possible improvements in the next years in this area.

To perform this quantitative assessment some specific metrics, or *degrees*, have been defined in Section 6. And finally, a quantitative evaluation of a set of 20 popular SNA-software tools have been carried out, to show how these dimensions (and their related metrics) can be used to evaluate these technologies, in Section 7.

The main conclusions from the research questions proposed, their related dimensions and metrics proposed, and the assessment of the SNA-software tools carried out, can be briefly summarized as follows:

1. **What can I discover?**. This question is related to the different types of knowledge that the tool is able to extract. The goal of this question is to quantify the capacity of the tool to extract valuable knowledge from the data. To answer this question, the dimension *Pattern & Knowledge discovery* (and its related $d_{Value}(t)$ metric) has been considered. From the current study, it can be concluded that some kind of analyzes, like topology measures, link analysis or static community detection, are fairly common in the tools analyzed. While other analyzes, like dynamic community detection, opinion mining, virality or homophily, are quite rare. In our opinion, this phenomenon is linked to the lack of proportional relation between the *Information Fusion* and *Knowledge Discovery*. This suggests that the content of an OSN is not being fully exploited by actual tools.
2. **What is the limit?**. Answering this question the researcher will understand the scalability of the tool. This question is quite important due to the amount of data that can be extracted from OSNs. This question has been addressed through the *Scalability* dimension (and its related $d_{Volume}(t)$ metric). As the quantitative analysis of the tools shown in Section 7, it is clear that most of the analyzed tools are capable of handling fairly big graphs (around 100.000 nodes), they are very customizable (their code is publicly available), and allow communication with other tools via an API, even thought just a few achieves fully integration with other applications. However, few tools are capable of doing BigData and the ones that can, have a low/medium average Knowledge Discovery capabilities. Taking into account the fast growth of the OSNs, the size of handled networks (to millions or hundred of millions nodes and vertex) will be swift increased in this tools, jointly to other capabilities, as the fusion and integration from different sources, and with different tools, to improve the knowledge discovery reached by these tools.
3. **What kind of data can I integrate?**. This relevant question, related to the capacity to integrate and fusion information from current SNA technologies, has been analyzed through the *Information Fusion & Integration* dimension (and its related $d_{Variety}(t)$ metric). Again, the quantitative analysis of the tools carried out, shows that most of the analyzed tools used complex graph representation (multilayered graphs or hypergraphs), are capable of processing a medium amount of different data types (two or three different types, and are only capable of extracting data from one unique OSN. Therefore, and related to this dimension, it can be expected a very significant increase in research related to the fusion and integration of information using different types of data formats, and when possible, from different OSNs.
4. **What can I show?**. Finally, this last question was explored using the *Visualization* dimension (and its related $d_{Visual}(t)$ metric), and how it has been shown, although there exist a large number of information visualization, and tools that provide flexible methods to visualize the information, this is still an open problem in the area. From the quantitative analysis related to this dimension, it can be concluded that the visualization capabilities of the tools where more evenly distributed compared to the rest of the dimensions. However, we have observed a lack of tools with high Scalability and Visualization capabilities. Taking the absolutely necessity to provide visualization tools to the end users and practitioners, the research and improvements in this area will be a high (and hot) topic in SNA in the next years, for example, in areas as dynamic community finding, data analytics or pattern finding to mention but a few.

Finally, we need to make a reflection on the dimensions and metrics proposed. What is proposed here is an initial work, derived from an intense dedication to the area of SNA in the last ten years. These dimensions, and the defined metrics (or degrees), cannot (and should not) be considered as the only ones that can be defined, even the definition cannot be considered as complete. From the analysis of the state of the art, we have selected those more relevant (from our perspective) features that could be used to better identify and reflect the state of these technologies. It is quite probable, that some relevant characteristics have not been considered by authors. On the other hand, the fact that several tools such as Grasphistry have a value of 1.0 in the dimensions of Information Fusion, Scalability, or Visualization, or that other tools such as Neo4J, also reach the value of 1.0 in dimensions such as Scalability or Visualization, do not mean that these features cannot be improved in the future. These values only indicate that given the current state of technology (and using our evaluation rubric), a higher value is achieved compare to other systems or tools for those dimensions. Obviously, and given the huge and fast growth of these technologies, we are sure that these values will change in the coming years, but the authors think that these dimensions, along with their related metrics, could be an important decision tool for future researchers, and practitioners, in the field of SNA.

Authors Statement

David Camacho conceived of the presented idea on the 4 Dimensions in SNA, wrote several sections of the state of the art in SNA, designed the overall structure and organization of the paper, and contributed to the selection and analysis of the SNAs frameworks and tools studied in this work.

Angel Panizo contributed to the revision of the state of the art in SNA, to define the Information Fusion & Integration and Visualizations dimensions, he made the exhaustive SNA frameworks and tools analysis, and he contributed with the graphical analysis proposed.

Gema Bello contributed to the revision of the state of the art in SNA, and to define the Pattern & Knowledge discovery and Scalability dimension, she made the exhaustive SNA frameworks and tools analysis.

Antonio Gonzalez-Pardo contributed to the revision of the state of the art in SNA, and the dimension called Information Fusion & Integration.

Finally, he had contributed reading and revising different parts of the paper, and discussing the assessment of the technologies.

Erik Cambria contributed to the revisions of the state of the art in sentiment analysis, contributed to the overall organization of this work, discussed the findings and main contributions of this work with the rest of authors. All authors discussed the results and contributed to the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by several research grants: Spanish Ministry of Science and Education under TIN2014-56494-C4-4-P grant (DeepBio), European Union, under ISFP-POLICE ACTION: 823701-ISFP-2017-AG-RAD grant (YoungRes), and Comunidad Autónoma de Madrid under S2018/TCS-4566 grant (CYNAMON). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

Appendix: Open Access to Social Network Analysis Dimensions

In order to allow researchers from social sciences, science and engineering, SNA analysts and professionals, developers and engineers, etc. not only to access the data used in this article, but to foster for a future collaboration among the community interested in network analysis, a website has been designed to facilitate accessing to:

- An open github repository to allow including new software, documents, open papers, and technical data, related to SNA technology.
- *Rubric.odt*: The evaluation rubric designed to assess the SNA-software tools.
- *analized_tools.ods*: The specific evaluation made for the SNA-software tools carried out in this paper.
- *README.md*: A collaborative website to evaluate SNA-software (tools, frameworks, algorithms, etc.) by the community, included as part of the github repository.
- *4Dimensions.pdf*: A brief summary of this paper.

The SNA 4-Dimensions website: https://ai-da-sna.github.io/

We would like to encourage the community to provide its own evaluations, of both the tools that have been evaluated in this paper as well as others in which you have previous experience. This collaboration will promote the use of SNA technologies and fostering new developments, but it will not be possible without the cooperation of the community, so your contribution will be highly appreciated.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2020.05.009.

References

[1] D.M. Boyd, N.B. Ellison, Social network sites: definition, history, and scholarship, J. Comput.-Mediat. Commun. 13 (1) (2007) 210–230.
[2] A.R. Radcliffe-Brown, The social organization of Australian tribes, Oceania 1 (1) (1930) 34–63.
[3] A.R. Radcliffe-Brown, On social structure, J. R. Anthropol. Inst. Great Britain Ireland 70 (1) (1940) 1–12.
[4] J.A. Barnes, Class and committees in a Norwegian Island Parish, Human Relat. 7 (1) (1954) 39–58.
[5] C.C. Aggarwal, An Introduction to Social Network Data Analytics, in: Social Network Data Analytics, Springer, 2011, pp. 1–15.
[6] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, IEEE Trans. Knowl. Data Eng. 26 (1) (2013) 97–107.
[7] P. Wang, B. Xu, Y. Wu, X. Zhou, Link prediction in social networks: the state-of-the-art, Sci. China Inf. Sci. 58 (1) (2015) 1–38.
[8] S.P. Borgatti, M.G. Everett, J.C. Johnson, Analyzing social networks, Sage, 2018.
[9] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (3) (2015) 925.
[10] D. Chaffey, Global Social Media Research Summary 2019, 2019.
[11] A. Guille, H. Hacid, C. Favre, D.A. Zighed, Information diffusion in online social Networks: asurvey, SIGMOND Record 42 (2) (2013) 17–28.
[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and Mining of Academic Social Networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 990–998.
[13] D. Laney, 3D data management: controlling data volume, velocity and variety, META Group Research Note 6 (70) (2001) 1.
[14] J.M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, P. Rodriguez, The Little Engine (s) That Could: Scaling Online Social Networks, ACM SIGCOMM Comput. Commun. Rev. 41 (4) (2011) 375–386.
[15] K. Wakita, T. Tsurumi, Finding Community Structure in Mega-Scale Social Networks, in: Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 1275–1276.
[16] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content, Neurocomputing 174 (2016) 50–59.
[17] A. Beach, M. Gartrell, X. Xing, R. Han, Q. Lv, S. Mishra, K. Seada, Fusing mobile, sensor, and social data to fully enable context-aware computing, in: Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications, ACM, 2010, pp. 60–65.
[18] L.C. Freeman, Visualizing social networks, J. Soc. Struct. 1 (1) (2000) 4.
[19] U. Brandes, D. Wagner, Analysis and Visualization of Social Networks, in: Graph Drawing Software, Springer, 2004, pp. 321–340.
[20] R. Ball, An Introduction to Bibliometrics: New Development and Trends, Chandos Publishing, 2017.
[21] J. McLevey, R. McIlroy-Young, Introducing metaknowledge: software for computational research in information science, network analysis, and science of science, J. Informetr. 11 (1) (2017) 176–197.
[22] W. Marx, L. Bornmann, A. Barth, L. Leydesdorff, Detecting the historical roots of research fields by reference publication year spectroscopy (rpys), J. Assoc. Inf. Sci. Technol. 65 (4) (2014) 751–764.
[23] J.A. Comins, T.W. Hussey, Compressing multiple scales of impact detection by reference publication year spectroscopy, J. Informetr. 9 (3) (2015) 449–454.
[24] S. Milgram, The small world problem, Psychol. Today 2 (1) (1967) 60–67.
[25] M.S. Granovetter, The strength of weak ties, in: Social Networks, Elsevier, 1977, pp. 347–367.
[26] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1) (1977) 35–41.
[27] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Netw. 1 (3) (1978) 215–239.
[28] J.S. Coleman, Social capital in the creation of human capital, Am. J. Sociol. 94 (1988) S95–S120.
[29] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[30] T.W. Valente, Network interventions, Science 337 (6090) (2012) 49–53.
[31] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[32] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.
[33] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.
[34] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech: Theory Exp. 2008 (10) (2008) P10008.
[35] M.E. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (6) (2004) 066133.
[36] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[37] A. Clauset, M.E. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (6) (2004) 066111.
[38] A. Clauset, Finding local community structure in networks, Phys. Rev. E 72 (2005) 026132.
[39] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814.
[40] G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (7136) (2007) 664.
[41] P. Pons, M. Latapy, Computing communities in large networks using random walks, in: p. Yolum, T. Güngör, F. Gürgen, C. Özturan (Eds.), Computer and Information Sciences - ISCIS 2005, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 284–293.
[42] P. Pons, M. Latapy, Computing communities in large networks using random walks., J. Graph Algorithms Appl. 10 (2) (2006) 191–218.
[43] M. Rosvall, C.T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks, Proc. Natl. Acad. Sci. 104 (18) (2007) 7327–7331.
[44] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. 105 (4) (2008) 1118–1123.
[45] J.M. Kumpula, M. Kivelä, K. Kaski, J. Saramäki, Sequential algorithm for fast clique percolation, Phys. Rev. E 78 (2) (2008) 026109.
[46] R. Rotta, A. Noack, Multilevel local search algorithms for modularity clustering, J. Exp. Algorithmi. (JEA) 16 (2011) 2–3.

[47] S. Cavallari, E. Cambria, H. Cai, K. Chang, V. Zheng, Embedding both finite and infinite communities on graph, IEEE Comput. Intell. Mag. 14 (3) (2019) 39–50.
[48] S. Wasserman, K. Faust, Social network analysis: Methods and applications, 8, Cambridge university press, 1994.
[49] S.E. Stemler, Content analysis, Emerg. Trend. Soc. Behav. Sci. (2015) 1–14.
[50] D.B. West, Introduction to graph theory, 2, Prentice hall Upper Saddle River, NJ, 1996.
[51] J.A. Bondy, U.S.R. Murty, Graph theory with applications, 290, Citeseer, 1976.
[52] G. Bello-Orgaz, J. Hernandez-Castro, D. Camacho, Detecting discussion communities on vaccination in twitter, Future Generat. Comput. Syst. 66 (2017) 125–136.
[53] L. Freeman, The development of social network analysis, Study Sociol. Sci. 1 (2004).
[54] R. Lara-Cabrera, A.G. Pardo, K. Benouaret, N. Faci, D. Benslimane, D. Camacho, Measuring the radicalisation risk in social networks, IEEE Access 5 (2017) 10892–10900.
[55] D. Lusher, J. Koskinen, G. Robins, Exponential random graph models for social networks: Theory, methods, and applications, Cambridge University Press, 2013.
[56] P.J. Carrington, J. Scott, S. Wasserman, Models and methods in social network analysis, 28, Cambridge university press, 2005.
[57] F. Vega-Redondo, Complex social networks, Cambridge University Press, 2007.
[58] M.T. Thai, P.M. Pardalos, Handbook of optimization in complex networks: Theory and applications, 57, Springer science & business media, 2011.
[59] R. Zafarani, M.A. Abbasi, H. Liu, Social media mining: an introduction, Cambridge University Press, 2014.
[60] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (11) (2010) 888–893.
[61] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: the state-of-the-art and comparative study, ACM Comput. Surv. (CSUR) 45 (4) (2013) 43.
[62] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.
[63] G. Bello-Orgaz, H.D. Menéndez, D. Camacho, Adaptive k-means algorithm for overlapped graph clustering, Int. J. Neural. Syst. 22 (5) (2012) 1250018.
[64] A. Gonzalez-Pardo, D. Camacho, Design of japanese tree frog algorithm for community finding problems, in: Intelligent Data Engineering and Automated Learning – IDEAL 2018, Springer International Publishing, 2018, pp. 307–315.
[65] R. Kannan, S. Vempala, A. Vetta, On clusterings: good, bad and spectral, J. ACM (JACM) 51 (3) (2004) 497–515.
[66] L. Tang, H. Liu, Graph mining applications to social network analysis, in: Managing and Mining Graph Data, Springer, 2010, pp. 487–513.
[67] J. Abello, M.G. Resende, S. Sudarsky, Massive quasi-clique detection, in: Latin American symposium on Theoretical Informatics, Springer, 2002, pp. 598–612.
[68] R. Kumar, J. Novak, A. Tomkins, Structure and evolution of online social networks, in: Link Mining: Models, Algorithms, and Applications, Springer, 2010, pp. 337–357.
[69] B. Balasundaram, S. Butenko, I.V. Hicks, Clique relaxations in social network analysis: the maximum k-plex problem, Oper. Res. 59 (1) (2011) 133–142.
[70] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, K-core organization of complex networks, Phys. Rev. Lett. 96 (4) (2006) 40601.
[71] G.W. Flake, S. Lawrence, C.L. Giles, et al., Efficient identification of web communities, in: KDD, 2000, 2000, pp. 150–160.
[72] P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent space approaches to social network analysis, J. Am. Stat. Assoc. 97 (460) (2002) 1090–1098.
[73] M.S. Handcock, A.E. Raftery, J.M. Tantrum, Model-based clustering for social networks, J. R. Stat. Soc.: Ser. A (Statistics in Society) 170 (2) (2007) 301–354.
[74] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, Wiley Interdiscip. Rev.: Data Mining and Knowledge Discovery 2 (1) (2012) 86–97.
[75] K.A. Heller, Z. Ghahramani, Bayesian hierarchical clustering, in: Proceedings of the 22nd International Conference on Machine Learning, ACM, 2005, pp. 297–304.
[76] R. Cazabet, H. Takeda, M. Hamasaki, F. Amblard, Using dynamic community detection to identify trends in user-generated content, Soc. Netw. Anal. Min. 2 (4) (2012) 361–371.
[77] D. Chakrabarti, R. Kumar, A. Tomkins, Evolutionary clustering, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 554–560.
[78] A. Panizo-LLedot, G. Bello-Orgaz, D. Camacho, A multi-objective genetic algorithm for detecting dynamic communities using a local search driven immigrant's scheme, Future. Generat. Comput. Syst. Available online 15 November 2019. (2019) 1–16.
[79] A. Panizo, G. Bello-Orgaz, D. Camacho, A genetic algorithm with local search based on label propagation for detecting dynamic communities, in: International Symposium on Intelligent and Distributed Computing, Springer, 2018, pp. 319–328.
[80] E. Osaba, J.D. Ser, A. Panizo, D. Camacho, A. Galvez, A. Iglesias, Combining bio-inspired meta-heuristics and novelty search for community detection over evolving graph streams, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, ACM, 2019, pp. 1329–1335.
[81] K.S. Xu, A.O. Hero, Dynamic stochastic blockmodels for time-evolving social networks, IEEE J. Sel. Top. Signal Process. 8 (4) (2014) 552–562.
[82] F. Liu, D. Choi, L. Xie, K. Roeder, Global spectral clustering in dynamic networks, Proc. Natl. Acad. Sci. 115 (5) (2018) 927–932.
[83] L. Peel, A. Clauset, Detecting change points in the large-scale structure of evolving networks, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[84] Y. Hulovatyy, T. Milenković, Scout: simultaneous time segmentation and community detection in dynamic networks, Sci. Rep. 6 (2016) 37557.
[85] R. Cazabet, F. Amblard, C. Hanachi, Detection of overlapping communities in dynamical social networks, in: 2010 IEEE Second International Conference on Social Computing, IEEE, 2010, pp. 309–314.
[86] G. Rossetti, L. Pappalardo, D. Pedreschi, F. Giannotti, Tiles: an online algorithm for community discovery in dynamic social networks, Mach. Learn. 106 (8) (2017) 1213–1241.
[87] T. Falkowski, A. Barth, M. Spiliopoulou, Studying community dynamics with an incremental graph mining algorithm, AMCIS 2008 Proc. (2008) 29.
[88] S.Y. Bhat, M. Abulaish, Octracker: a density-based framework for tracking the evolution of overlapping communities in osns, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, 2012, pp. 501–505.
[89] R.-H. Li, J. Su, L. Qin, J.X. Yu, Q. Dai, Persistent community search in temporal networks, in: 2018 IEEE 34th International Conference on Data Engineering (ICDE), IEEE, 2018, pp. 797–808.
[90] T. Viard, C. Magnien, M. Latapy, Enumerating maximal cliques in link streams with durations, Inf. Process. Lett. 133 (2018) 44–48.
[91] R. Agnihotri, R. Dingus, M.Y. Hu, M.T. Krush, Social media: influencing customer satisfaction in b2b sales, Ind. Market. Manag. 53 (2016) 172–180.
[92] M. Zhang, L. Guo, M. Hu, W. Liu, Influence of customer engagement with company social networks on stickiness: mediating effect of customer value creation, Int. J. Inf. Manage. 37 (3) (2017) 229–240.
[93] R. Thackeray, B.L. Neiger, C.L. Hanson, J.F. McKenzie, Enhancing promotional strategies within social marketing programs: use of web 2.0 social media, Health Promot. Pract. 9 (4) (2008) 338–343.
[94] C. Ashley, T. Tuten, Creative strategies in social media marketing: an exploratory study of branded social content and consumer engagement, Psychol. Market. 32 (1) (2015) 15–27.
[95] P.N. Howard, B. Kollanyi, Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum, Available at SSRN 2798311 (2016).
[96] G. Enli, Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election, Eur. J. Commun. 32 (1) (2017) 50–61.
[97] A. Panizo-LLedot, J. Torregrosa, G. Bello-Orgaz, J. Thorburn, D. Camacho, Describing alt-right communities and their discourse on twitter during the 2018 us mid-term elections, in: International Conference on Complex Networks and Their Applications, Springer, 2019, pp. 427–439.
[98] J. Klausen, Tweeting the jihad: social media networks of western foreign fighters in syria and iraq, Stud. Conf. Terrorism 38 (1) (2015) 1–22.
[99] J. Klausen, E.T. Barbieri, A. Reichlin-Melnick, A.Y. Zelin, The youtube jihadists: a social network analysis of al-muhajiroun's propaganda campaign, Perspect. Terrorism 6 (1) (2012).
[100] R. Lara-Cabrera, A. Gonzalez-Pardo, D. Camacho, Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in twitter, Futu. Generat Computer Systems (2017).
[101] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, J. Economic. Perspect. 31 (2) (2017) 211–236.
[102] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: a data mining perspective, ACM SIGKDD Explorat. Newsletter 19 (1) (2017) 22–36.
[103] A. Guille, H. Hacid, C. Favre, D.A. Zighed, Information diffusion in online social networks: asurvey, ACM Sigmod Record 42 (2) (2013) 17–28.
[104] M.E. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[105] G. Miritello, E. Moro, R. Lara, Dynamical strength of social ties in information spreading, Phys. Rev. E 83 (4) (2011) 045102.
[106] C. Liu, Z.-K. Zhang, Information spreading on dynamic social networks, Commun. Nonlinear Sci. Numer. Simul. 19 (4) (2014) 896–904.
[107] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (14) (2001) 3200.
[108] D. Kempe, J. Kleinberg, É. Tardos, Influential nodes in a diffusion model for social networks, in: International Colloquium on Automata, Languages, and Programming, Springer, 2005, pp. 1127–1138.
[109] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Mark. Lett. 12 (3) (2001) 211–223.
[110] J. Goldenberg, B. Libai, E. Muller, Using complex systems analysis to advance marketing theory development: modeling heterogeneity effects on new product growth through stochastic cellular automata, Acad. Market. Sci. Rev. 9 (3) (2001) 1–18.
[111] D.J. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. 99 (9) (2002) 5766–5771.
[112] E. Cambria, H. Wang, B. White, Guest editorial: big social data analysis, Knowl. Based Syst. 69 (2014) 1–2.
[113] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, S. Moon, I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, ACM, 2007, pp. 1–14.
[114] C.-Y. Weng, W.-T. Chu, J.-L. Wu, Rolenet: movie analysis from the perspective of social networks, IEEE Trans. Multimedia 11 (2) (2009) 256–271.
[115] A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods, and analytics, Int. J. Inf. Manage 35 (2) (2015) 137–144.
[116] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, IEEE Trans. Knowl. Data Eng. 26 (1) (2014) 97–107.
[117] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: recent achievements and new challenges, Inf. Fusion 28 (2016) 45–59.
[118] L. Manovich, Trending: the promises and the challenges of big social data, Debat. Digital Humanit. (2011) 460–475.
[119] Z. Tufekci, Big questions for social media big data: Representativeness, validity and other methodological pitfalls, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[120] C. Pal, A. McCallum, CC prediction with grapical models, in: Conference on Email and Anti-Spam, 2006.

[121] O. Bar-Yossef, I. Guy, R. Lempel, Y. Maarek, V. Soroka, Cluster ranking with an application to mining mailbox networks, Knowl. Inf. Syst. 14 (1) (2008) 101–139.

[122] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, R. Merom, Suggesting friends using the implicit social graph, KDD, 2010.

[123] M. De Choudhury, W. Mason, J. Hofman, D. Watts, Inferring relevant social networks from interpersonal communication, WWW, 2010.

[124] Y. Cai, Y. Chen, Mass: a multi-facet domainspecific influential blogger mining system, in: International Conference on Data Engineering, 2010, pp. 1109–1112.

[125] C. Sun, B.-q. Liu, C.-j. Sun, D.-Y. Zhang, X. Wang, Simrank: A link analysis based blogger recommendation algorithm using text similarity, in: International Conference on Machine Learning and Cybernetics, 2010, pp. 3368–3373.

[126] L. Akritidis, D. Katsaros, P. Bozanis, Identifying the productive and influential bloggers in a community, IEEE Trans. Syst. Man. Cybernet. Part C 41 (5) (2011) 759–764.

[127] E. Moon, S. Han, A qualitative method to find influencers using similarity-based approach in the blogosphere, Int. J. Soc. Comput. Cyber-Phys. Syst. 1 (1) (2011) 56–78.

[128] P. Chandra, E. Cambria, A. Hussain, Clustering social networks using interaction semantics and sentics, in: J. Wang, G. Yen, M. Polycarpou (Eds.), Advances in Neural Networks, Lecture Notes in Computer Science, 7367, Springer, 2012, pp. 379–385.

[129] I. Chaturvedi, Y.S. Ong, Y. Tsang, R. Welsch, E. Cambria, Learning word dependencies in text by means of a deep recurrent belief network, Knowledge-Based Systems, 108, Elsevier, 2016, pp. 144–154.

[130] G. Piao, J. Breslin, Inferring user interests in microblogging social networks: a survey, User Model User-Adapt Interact 28 (3) (2018) 277–329.

[131] A. McCallum, X. Wang, A. Corrada-Emmanuel, Topic and role discovery in social networks withexperiments on enron and academic email, J. Artif. Intell. Res. 30 (2007) 249–272.

[132] N. Pathak, C. Delong, A. Banerjee, K. Erickson, Social topic models for community extraction, SNA-KDD Workshop, 2008.

[133] T. Wang, H. Liu, J. He, X. Du, Mining user interests from information sharing behaviors in social media, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2013, pp. 85–98.

[134] J. Wang, W. Zhao, Y. He, X. Li, Infer user interests via link structure regularization, ACM Trans. Intell. Syst. Technol. 5 (2) (2014) 23.

[135] J. Kang, H. Lee, Modeling user interest in social media using news media and wikipedia, Inf. Syst. 65 (2017) 52–64.

[136] C. Jipmo, G. Quercini, N. Bennacer, Frisk: a multilingual approach to find twitter interests via wikipedia, in: International Conference on Advanced Data Mining and Applications, 2017, pp. 243–256.

[137] S. Faralli, G. Stilo, P. Velardi, Automatic acquisition of a taxonomy of microblogs users' interests, Web Semantics: Science, Services and Agents on the World Wide Web 45 (2017) 23–40.

[138] F. Zarrinkalam, M. Kahani, E. Bagheri, Mining user interests over active topics on social networks, Inf. Process. Manag. 54 (2) (2018) 339–357.

[139] A. Trikha, F. Zarrinkalam, E. Bagheri, Topic-association mining for user interest detection, in: European Conference on Information Retrieval, 2018, pp. 665–671.

[140] E. Cambria, N. Howard, Y. Xia, T.-S. Chua, Computational intelligence for big social data analysis, IEEE Comput. Intell. Mag. 11 (3) (2016) 8–9.

[141] E. Cambria, S. Poria, F. Bisio, R. Bajpai, I. Chaturvedi, The CLSA model: a novel framework for concept-level sentiment analysis, in: LNCS, 9042, Springer, 2015, pp. 3–22.

[142] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif. Intell. Rev. 53 (2020) 2313–2339.

[143] A. Bandhakavi, N. Wiratunga, S. Massie, P. Deepak, Lexicon generation for emotion analysis from text, IEEE Intell. Syst. 32 (1) (2017) 102–108.

[144] F. Xu, J. Yu, R. Xia, Instance-based domain adaptation via multi-clustering logistic approximation, IEEE Intell. Syst. 33 (1) (2018) 78–88.

[145] Q. Yang, Y. Rao, H. Xie, J. Wang, F.L. Wang, W.H. Chan, Segment-level joint topic-sentiment model for online review analysis, IEEE Intell. Syst 34 (1) (2019) 43–50.

[146] A. Weichselbraun, S. Gindl, F. Fischer, S. Vakulenko, A. Scharl, Aspect-based extraction and analysis of affective knowledge from social media streams., IEEE Intell. Syst. 32 (3) (2017) 80–88.

[147] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (264–270) (2019).

[148] R. Mihalcea, A. Garimella, What men say, what women hear: finding gender-specific meaning shades, IEEE Intell. Syst. 31 (4) (2016) 62–67.

[149] A. Bukeer, G. Roffo, A. Vinciarelli, Type like a man! inferring gender from keystroke dynamics in live-chats, IEEE Intell. Syst. 34 (6) (2019) 53–59.

[150] D. Mahata, J. Friedrichs, R.R. Shah, J. Jiang, Detecting personal intake of medicine from twitter., IEEE Intell. Syst. 33 (4) (2018) 87–95.

[151] S.A. Qureshi, S. Saha, M. Hasanuzzaman, G. Dias, Multitask representation learning for multimodal estimation of depression level, IEEE Intell. Syst. 34 (5) (2019) 45–52.

[152] M. Ebrahimi, A. Hossein, A. Sheth, Challenges of sentiment analysis for dynamic events, IEEE Intell. Syst. 32 (5) (2017) 70–75.

[153] A. Valdivia, V. Luzon, F. Herrera, Sentiment analysis in tripadvisor, IEEE Intell. Syst. 32 (4) (2017) 72–77.

[154] M.S. Akhtar, A. Ekbal, S. Narayan, V. Singh, No, that never happened!! investigating rumors on twitter, IEEE Intell. Syst. 33 (5) (2018) 8–15.

[155] J. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, Supervised learning for fake news detection, IEEE Intell. Syst. 34 (2) (2019) 76–81.

[156] C. Welch, V. Perez-Rosas, J. Kummerfeld, R. Mihalcea, Learning from personal longitudinal dialog data., IEEE Intell. Syst. 34 (4) (2019) 16–23.

[157] N. Pathak, S. Mane, J. Srivastava, N. Contractor, Knowledge perception analysis in a social network, in: SIAM International Conference on Data Mining, 2006.

[158] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, A. Smeaton, Combining social network analysis and sentiment analysis to explore the potential for online radicalisation, in: IEEE International Conference on Advances in Social Network Analysis and Mining, 2009, pp. 231–236.

[159] W. Gryc, K. Moilanen, Leveraging textual sentiment analysis with social network modeling: Sentiment analysis of political blogs in the 2008 u.s. presidential election, From Text to Political Positions Workshop, 2010.

[160] M. Shams, M. Saffar, A. Shakery, H. Faili, Applying sentiment and social network analysis in user modeling, in: CICLing, 2012, pp. 526–539.

[161] J. Morente-Molinera, G. Kou, K. Samuylov, R. Urena, E. Herrera-Viedma, Carrying out consensual group decision making processes under social networks using sentiment analysis over comparative expressions, Knowl. Based Syst. 165 (2019) 335–345.

[162] K.P. Smith, N.A. Christakis, Social networks and health, Annu. Rev. Sociol 34 (2008) 405–429.

[163] D. Chambers, P. Wilson, C. Thompson, M. Harden, Social network analysis in healthcare settings: a systematic scoping review, PLoS ONE 7 (8) (2012) e41911.

[164] L.F. Berkman, T. Glass, I. Brissette, T.E. Seeman, From social integration to health: durkheim in the new millennium, Soc. Sci. Med. 51 (6) (2000) 843–857.

[165] S. Cohen, D. Janicki-Deverts, Can we improve our physical health by altering our social networks? Perspect. Psychol. Sci. 4 (4) (2009) 375–378.

[166] P.A. Thoits, Mechanisms linking social ties and support to physical and mental health, J. Health Soc. Behav. 52 (2) (2011) 145–161.

[167] D. Umberson, J. Karas Montez, Social relationships and health: a flashpoint for health policy, J Health Soc Behav 51 (1_suppl) (2010) S54–S66.

[168] J. Andreu-Perez, C.C. Poon, R.D. Merrifield, S.T. Wong, G.-Z. Yang, Big data for health, IEEE J Biomed Health Inform 19 (4) (2015) 1193–1208.

[169] U.D. of Health, H. Services, et al., Preventing tobacco use among youth and young adults: a report of the surgeon general, 2012.

[170] E.M. Trucco, C.R. Colder, W.F. Wieczorek, Vulnerability to peer influence: a moderated mediation study of early adolescent alcohol use initiation, Addict Behav 36 (7) (2011) 729–736.

[171] S. Livingstone, Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression, New Media Soc. 10 (3) (2008) 393–411.

[172] B. Freeman, New media and tobacco control, Tob. Control 21 (2) (2012) 139–144.

[173] G.C. Huang, J.B. Unger, D. Soto, K. Fujimoto, M.A. Pentz, M. Jordan-Marsh, T.W. Valente, Peer influences: the impact of online and offline friendship networks on adolescent smoking and alcohol use, J. Adolesc. Health 54 (5) (2014) 508–514.

[174] J.I. Hoffman, Chapter 27 - linear regression, in: J.I. Hoffman (Ed.), Biostatistics for Medical and Biomedical Practitioners, Academic Press, 2015, pp. 451–500.

[175] M.E. Larsen, T.W. Boonstra, P.J. Batterham, B. O'Dea, C. Paris, H. Christensen, We feel: mapping emotion on twitter, IEEE J. Biomed. Health Inform. 19 (4) (2015) 1246–1252.

[176] C. Paquet, D. Coulombier, R. Kaiser, M. Ciotti, Epidemic intelligence: a new framework for strengthening disease surveillance in europe., Euro. Surveillance: bulletin europeen sur les maladies transmissibles European communicable disease bulletin 11 (12) (2005) 212–214.

[177] G. Bello-Orgaz, J. Hernandez-Castro, D. Camacho, A Survey of Social Web Mining Applications for Disease Outbreak Detection, in: Intelligent Distributed Computing VIII, Springer, 2015, pp. 345–356.

[178] A. Culotta, Towards detecting influenza epidemics by analyzing twitter messages, in: Proceedings of the first workshop on social media analytics, ACM, 2010, pp. 115–122.

[179] E. Aramaki, S. Maskawa, M. Morita, Twitter catches the flu: detecting influenza epidemics using twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1568–1576.

[180] T. Bodnar, M. Salathé, Validating models for disease detection using twitter, in: Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2013, pp. 699–702.

[181] M.F. Guiñazú, V. Cortés, C.F. Ibáñez, J.D. Velásquez, Employing online social networks in precision-medicine approach using information fusion predictive model to improve substance use surveillance: a lesson from twitter and marijuana consumption, Inf. Fusion 55 (2020) 150–163.

[182] M. Mendoza, B. Poblete, C. Castillo, Twitter under crisis: Can we trust what we rt? in: Proceedings of The First workshop on Social Media Analytics, 2010, pp. 71–79.

[183] L. Kim, S.M. Fast, N. Markuzon, Incorporating media data into a model of infectious disease transmission, PLoS ONE 14 (2) (2019).

[184] T. Sharot, C.R. Sunstein, How people decide what they want to know, Nat. Hum. Behav. (2020) 1–6.

[185] J. Kulshrestha, M. Eslami, J. Messias, M.B. Zafar, S. Ghosh, K.P. Gummadi, K. Karahalios, Quantifying search bias: Investigating sources of bias for political searches in social media, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 417–432.

[186] C.A. Bail, L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.F. Hunzaker, J. Lee, M. Mann, F. Merhout, A. Volfovsky, Exposure to opposing views on social media can increase political polarization, Proc. Natl. Acad. Sci. 115 (37) (2018) 9216–9221.

[187] M.D. Vicario, W. Quattrociocchi, A. Scala, F. Zollo, Polarization and fake news: early warning of potential misinformation targets, ACM Transactions on the Web (TWEB) 13 (2) (2019) 1–22.

[188] P. Block, M. Hoffman, I.J. Raabe, J.B. Dowd, C. Rahal, R. Kashyap, M.C. Mills, Social network-based distancing strategies to flatten the covid 19 curve in a post-lockdown world, arXiv preprint arXiv:2004.07052 (2020).

[189] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, Y. Wang, A first look at covid-19 information and misinformation sharing on twitter, arXiv preprint arXiv:2003.13907 (2020).

[190] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C.M. Valensise, E. Brugnoli, A.L. Schmidt, P. Zola, F. Zollo, A. Scala, The covid-19 social media infodemic, arXiv preprint arXiv:2003.05004 (2020).

[191] N. Velásquez, R. Leahy, N.J. Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. Jha, N. Johnson, Hate multiverse spreads malicious covid-19 content online beyond individual platform control, arXiv preprint arXiv:2004.00673 (2020).

[192] P. Harrigan, U. Evers, M. Miles, T. Daly, Customer engagement with tourism social media brands, Tourism Manag. 59 (2017) 597–609.

[193] C. Kohli, R. Suri, A. Kapoor, Will social media kill branding? Bus. Horiz. 58 (1) (2015) 35–44.

[194] W.G. Mangold, D.J. Faulds, Social media: the new hybrid element of the promotion mix, Bus. Horiz. 52 (4) (2009) 357–365.

[195] L. De Vries, S. Gensler, P.S. Leeflang, Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing, J. Interact. Market. 26 (2) (2012) 83–91.

[196] J. Whitelock, J.W. Cadogan, S. Okazaki, C.R. Taylor, Social media and international advertising: theoretical challenges and future directions, Int. Market. Rev. (2013).

[197] K. Swani, G.R. Milne, B.P. Brown, A.G. Assaf, N. Donthu, What messages to post? evaluating the popularity of social media communications in business versus consumer markets, Ind. Market. Manag. 62 (2017) 77–87.

[198] R.G. Duffett, Facebook advertising's influence on intention-to-purchase and purchase amongst millennials, Internet Res. 25 (4) (2015) 498–526.

[199] F. A. Carrillat, A. d'Astous, E. Morissette Grégoire, Leveraging social media to enhance recruitment effectiveness: a facebook experiment, Internet Res. 24 (4) (2014) 474–495.

[200] S. Hudson, L. Huang, M.S. Roth, T.J. Madden, The influence of social media interactions on consumer–brand relationships: a three-country study of brand perceptions and marketing behaviors, Int. J. Res. Market. 33 (1) (2016) 27–41.

[201] G. Viglia, R. Minazzi, D. Buhalis, The influence of e-word-of-mouth on hotel occupancy rate, Int. J. Contemp. Hospital. Manag. 28 (9) (2016) 2035–2051.

[202] E. Kim, Y. Sung, H. Kang, Brand followers' retweeting behavior on twitter: how brand relationships influence brand electronic word-of-mouth, Comput. Human. Behav. 37 (2014) 18–25.

[203] S.-C. Chu, Y. Kim, Determinants of consumer engagement in electronic word-of–mouth (ewom) in social networking sites, Int. J. Adv. 30 (1) (2011) 47–75.

[204] B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: tweets as electronic word of mouth, J. Am. Soc. Inf. Sci. Technol. 60 (11) (2009) 2169–2188.

[205] S. Asur, B. Huberman, et al., Predicting the future with social media, in: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 1, IEEE, 2010, pp. 492–499.

[206] M. Pronschinske, M.D. Groza, M. Walker, Attracting facebook'fans': the importance of authenticity and engagement as a social networking strategy for professional sport teams, Sport Market. Q. 21 (4) (2012) 221.

[207] J. McCarthy, J. Rowley, C. Jane Ashworth, E. Pioch, Managing brand presence through social media: the case of uk football clubs, Internet Res. 24 (2) (2014) 181–204.

[208] K.S. Coulter, A. Roggeveen, "Like it or not" consumer responses to word-of–mouth communication in on-line social networks, Manag. Res. Rev. 35 (9) (2012) 878–899.

[209] E.C. Malthouse, M. Haenlein, B. Skiera, E. Wege, M. Zhang, Managing customer relationships in the social media era: introducing the social crm house, J. Interact. Market. 27 (4) (2013) 270–280.

[210] K.J. Trainor, J.M. Andzulis, A. Rapp, R. Agnihotri, Social media technology usage and customer relationship performance: a capabilities-based examination of social crm, J. Bus. Res. 67 (6) (2014) 1201–1208.

[211] J.N. Moore, C.D. Hopkins, M.A. Raymond, Utilization of relationship-oriented social media in the selling process: a comparison of consumer (b2c) and industrial (b2b) salespeople, J. Internet Commer. 12 (1) (2013) 48–75.

[212] H.G. Pereira, M. de Fátima Salgueiro, I. Mateus, Say yes to facebook and get your customers involved! relationships in a world of social networks, Bus. Horiz. 57 (6) (2014) 695–702.

[213] A.M. Munar, J.K.S. Jacobsen, Motivations for sharing tourism experiences through social media, Tour. Manag. 43 (2014) 46–54.

[214] J. Li, L. Xu, L. Tang, S. Wang, L. Li, Big data in tourism research: a literature review, Tourism Manag. 68 (2018) 301–323.

[215] M. Sigala, E. Christou, U. Gretzel, Social media in travel, tourism and hospitality: Theory, practice and cases, Ashgate Publishing, Ltd., 2012.

[216] R. Minazzi, Social media marketing in tourism and hospitality, 2015.

[217] N. Bennett, R.H. Lemelin, R. Koster, I. Budke, A capital assets framework for appraising and building capacity for tourism development in aboriginal protected area gateway communities, Tourism Manag. 33 (4) (2012) 752–766.

[218] D. Leung, R. Law, H. Van Hoof, D. Buhalis, Social media in tourism and hospitality: a literature review, J. Travel Tour. Market. 30 (1–2) (2013) 3–22.

[219] K.-Y. Goh, C.-S. Heng, Z. Lin, Social media brand community and consumer behavior: quantifying the relative impact of user-and marketer-generated content, Inf. Syst. Res. 24 (1) (2013) 88–107.

[220] A. Wilson, H. Murphy, J.C. Fierro, Hospitality and travel: the nature and implications of user-generated content, Cornell Hospital. Q. 53 (3) (2012) 220–228.

[221] X. Xu, Y. Li, The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: a text mining approach, Int. J. Hosp. Manag. 55 (2016) 57–69.

[222] Y. Guo, S.J. Barnes, Q. Jia, Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation, Tour. Manag. 59 (2017) 467–483.

[223] Z. Xiang, Z. Schwartz, J.H. Gerdes Jr, M. Uysal, What can big data and text analytics tell us about hotel guest experience and satisfaction? Int. J. Hosp. Manag. 44 (2015) 120–130.

[224] S.-J. Doh, J.-S. Hwang, How consumers evaluate ewom (electronic word-of-mouth) messages, CyberPsychol. Behav. 12 (2) (2009) 193–197.

[225] S. Poria, I. Chaturvedi, E. Cambria, F. Bisio, Sentic LDA: improving on LDA with semantic similarity for aspect-based sentiment analysis, in: IJCNN, 2016, pp. 4465–4473.

[226] E. Cambria, S. Poria, D. Hazarika, K. Kwok, SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings, in: AAAI, 2018, pp. 1795–1802.

[227] C. Guerreiro, E. Cambria, H. Nguyen, Understanding the role of social media in backpacker tourism, in: ICDM, 2019, pp. 530–537.

[228] Y.-H. Hu, Y.-L. Chen, H.-L. Chou, Opinion mining from online hotel reviews–a text summarization approach, Inf. Process. Manag. 53 (2) (2017) 436–449.

[229] G. Bordogna, L. Frigerio, A. Cuzzocrea, G. Psaila, Clustering geo-tagged tweets for advanced big data analytics, in: 2016 IEEE International Congress on Big Data (BigData Congress), IEEE, 2016, pp. 42–51.

[230] H.Q. Vu, G. Li, R. Law, B.H. Ye, Exploring the travel behaviors of inbound tourists to hong kong using geotagged photos, Tour. Manag. 46 (2015) 222–232.

[231] X. Lu, C. Wang, J.-M. Yang, Y. Pang, L. Zhang, Photo2trip: generating travel routes from geo-tagged photos for trip planning, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 143–152.

[232] N. Deng, X.R. Li, Feeling a destination through the "right" photos: amachine learning model for dmos' photo selection, Tour. Manag. 65 (2018) 267–278.

[233] A. Majid, L. Chen, G. Chen, H.T. Mirza, I. Hussain, J. Woodward, A context-aware personalized travel recommendation system based on geotagged social media data mining, Int. J. Geograph. Inf. Sci. 27 (4) (2013) 662–684.

[234] C.-H. Ku, G. Leroy, A decision support system: automated crime report analysis and classification for e-government, Gov. Inf. Q. 31 (4) (2014) 534–544.

[235] P. Phillips, I. Lee, Mining co-distribution patterns for large crime datasets, Expert Syst. Appl. 39 (14) (2012) 11556–11563.

[236] S. Chainey, L. Tompson, S. Uhlig, The utility of hotspot mapping for predicting spatial patterns of crime, Secur. J. 21 (1) (2008) 4–28.

[237] M.S. Gerber, Predicting crime using twitter and kernel density estimation, Decis. Support Syst. 61 (2014) 115–125.

[238] United Nations Office On Drugs and Crime, The use of the Internet for terrorist purposes, Technical Report, 2012. Vienna

[239] R. Thompson, Radicalization and the use of social media, J. Strateg. Secur. 4 (4) (2011) 167–190.

[240] K. Cohen, F. Johansson, L. Kaati, J.C. Mork, Detecting linguistic markers for radical violence in social media, Terror. Polit. Violenc. 26 (1) (2014) 246–256.

[241] J. Reid Meloy, J. Hoffmann, A. Guldimann, D. James, The role of warning behaviors in threat assessment: an exploration and suggested typology, Behav. Sci. Law 30 (3) (2012) 256–279.

[242] J. Torregrosa, J. Thorburn, R. Lara-Cabrera, D. Camacho, H.M. Trujillo, Linguistic analysis of pro-isis users on twitter, Behav. Sci. Terror. Polit. Aggress. 0 (0) (2019) 1–15.

[243] M. Fernandez, A. Gonzalez-Pardo, H. Alani, Radicalisation influence in social media, Semant. Web J. (2019) In–Press.

[244] M. Lalou, M.A. Tahraoui, H. Kheddouci, The critical node detection problem in networks: asurvey, Comput. Sci. Rev. 28 (2018) 92–117.

[245] R.C. Gunasekara, K. Mehrotra, C.K. Mohan, Multi-objective optimization to identify key players in large social networks, Soc. Netw. Anal. Min. 5 (1) (2015) 21.

[246] Optimizing network attacks by artificial bee colony, Inf. Sci. (Ny) 377 (2017) 30–50.

[247] S. Singh, N. Saxena, A. Roy, H. Kim, A survey on 5g network technologies from social perspective, IETE Tech. Rev. 34 (1) (2017) 30–39.

[248] G. Araniti, A. Orsino, L. Militano, L. Wang, A. Iera, Context-aware information diffusion for alerting messages in 5g mobile social networks, IEEE Internet Thing. J. 4 (2) (2016) 427–436.

[249] L. Atzori, A. Iera, G. Morabito, M. Nitti, The social internet of things (siot)–when social networks meet the internet of things: concept, architecture and network characterization, Comput. Netw. 56 (16) (2012) 3594–3608.

[250] P.N. Howard, A. Duffy, D. Freelon, M.M. Hussain, W. Mari, M. Maziad, Opening closed regimes: what was the role of social media during the arab spring? Available at SSRN 2595096 (2011).

[251] S. Boulianne, Social media use and participation: a meta-analysis of current research, Inf. Commun. Soc. 18 (5) (2015) 524–538.

[252] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Eleventh International AAAI Conference on Web and Social Media, 2017.

[253] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 2017, pp. 1–10.

[254] M.D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[255] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, Community detection in social media, Data Min. Knowl. Discov. 24 (3) (2012) 515–554.

[256] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The role of social networks in information diffusion, in: Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 519–528.

[257] K. Müller, C. Schwarz, Fanning the flames of hate: social media and hate crime, Available at SSRN 3082972 (2018).

[258] S. Jaki, T. De Smedt, Right-wing german hate speech on twitter: analysis and automatic detection, Manuscr. Submitted (2018).

[259] P. Burnap, M.L. Williams, Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making, Policy Internet 7 (2) (2015) 223–242.

[260] A. Boutet, H. Kim, E. Yoneki, What'S in twitter, i know what parties are popular and who you are supporting now!, Soc. Netw. Anal. Min. 3 (4) (2013) 1379–1391.

[261] A. Gruzd, J. Roy, Investigating political polarization on twitter: acanadian perspective, Policy Internet 6 (1) (2014) 28–45.

[262] A. Fang, I. Ounis, P. Habel, C. Macdonald, N. Limsopatham, Topic-centric classification of twitter user's political orientation, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2015, pp. 791–794.

[263] J. Borge-Holthoefer, W. Magdy, K. Darwish, I. Weber, Content and network dynamics behind egyptian political polarization on twitter, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2015, pp. 700–711.

[264] M. Ozer, N. Kim, H. Davulcu, Community detection in political twitter networks using nonnegative matrix factorization methods, in: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, in: ASONAM '16, IEEE Press, 2016, pp. 81–88.

[265] M.D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, F. Menczer, Predicting the political alignment of twitter users, in: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, IEEE, 2011, pp. 192–199.

[266] E. Colleoni, A. Rozza, A. Arvidsson, Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data, J. Commun. 64 (2) (2014) 317–332.

[267] S. Stieglitz, L. Dang-Xuan, Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior, in: 2012 45th Hawaii International Conference on System Sciences, IEEE, 2012, pp. 3500–3509.

[268] P. Barberá, J.T. Jost, J. Nagler, J.A. Tucker, R. Bonneau, Tweeting from left to right: is online political communication more than an echo chamber? Psychol. Sci. 26 (10) (2015) 1531–1542.

[269] E.C. Tandoc Jr, Z.W. Lim, R. Ling, Defining "fake news" a typology of scholarly definitions, Digit. J. 6 (2) (2018) 137–153.

[270] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (6380) (2018) 1146–1151.

[271] V. Bakir, A. McStay, Fake news and the economy of emotions: problems, causes, solutions, Digit. Journalism 6 (2) (2018) 154–175.

[272] N.J. Conroy, V.L. Rubin, Y. Chen, Automatic deception detection: methods for finding fake news, Proc. Assoc. Inf. Sci. Technol. 52 (1) (2015) 1–4.

[273] D.M. Lazer, M.A. Baum, Y. Benkler, A.J. Berinsky, K.M. Greenhill, F. Menczer, M.J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al., The science of fake news, Science 359 (6380) (2018) 1094–1096.

[274] J. Fürnkranz, A study using n-gram features for text categorization, Austr. Res. Inst. Artif. Intell. 3 (1998) (1998) 1–10.

[275] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, arXiv preprint arXiv:1702.05638 (2017).

[276] Z. Jin, J. Cao, Y. Zhang, J. Luo, News verification by exploiting conflicting social viewpoints in microblogs, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[277] E. Tacchini, G. Ballarin, M.L. Della Vedova, S. Moret, L. de Alfaro, Some like it hoax: automated fake news detection in social networks, arXiv Preprint arXiv:1704.07506 (2017).

[278] N. Ruchansky, S. Seo, Y. Liu, Csi: a hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, 2017, pp. 797–806.

[279] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Prominent features of rumor propagation in online social media, in: 2013 IEEE 13th International Conference on Data Mining, IEEE, 2013, pp. 1103–1108.

[280] R. Ji, D. Cao, Y. Zhou, F. Chen, Survey of visual sentiment prediction for social media analysis, Front. Comput. Sci. 10 (4) (2016) 602–611.

[281] Q. Cheng, Q. Zhang, P. Fu, C. Tu, S. Li, A survey and analysis on automatic image annotation, Pattern Recognit. 79 (2018) 242–259.

[282] S. Wazarkar, B.N. Keshavamurthy, A survey on image data analysis through clustering techniques for real world applications, J. Vis. Commun. Image Represent. 55 (2018) 596–626.

[283] A. Khosla, A. Das Sarma, R. Hamid, What makes an image popular? in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 867–876.

[284] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, S.-F. Chang, Image popularity prediction in social media using sentiment and context features, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 907–910.

[285] M. Merler, L. Cao, J.R. Smith, You are what you tweet… pic! Gender prediction based on semantic analysis of social media images, in: 2015 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2015, pp. 1–6.

[286] K. Jayarajah, A. Misra, Can instagram posts help characterize urban micro-events? in: 2016 19th International Conference on Information Fusion (FUSION), IEEE, 2016, pp. 130–137.

[287] M. Redi, D. Quercia, L. Graham, S. Gosling, Like partying? Your face says it all. predicting the ambiance of places with profile pictures, in: Ninth International AAAI Conference on Web and Social Media, 2015.

[288] Y. Hu, L. Manikonda, S. Kambhampati, What we instagram: a first analysis of instagram photo content and user types, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.

[289] L. Liu, D. Preotiuc-Pietro, Z.R. Samani, M.E. Moghaddam, L. Ungar, Analyzing personality through social media profile picture choice, in: Tenth International AAAI Conference on Web and Social Media, 2016.

[290] F. Celli, E. Bruni, B. Lepri, Automatic personality and interaction style recognition from facebook profile pictures, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 1101–1104.

[291] X. Wei, D. Stillwell, How smart does your profile image look?: Estimating intelligence from social network profile images, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, ACM, 2017, pp. 33–40.

[292] P.A. Cavazos-Rehg, M.J. Krauss, S.J. Sowles, L.J. Bierut, Marijuana-related posts on instagram, Prevent. Sci. 17 (6) (2016) 710–720.

[293] H. Hosseinmardi, R.I. Rafiq, R. Han, Q. Lv, S. Mishra, Prediction of cyberbullying incidents in a media-based social network, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 186–192.

[294] M. Hossain, Crowdsourcing: activities, incentives and users' motivations to participate, in: 2012 International Conference on Innovation Management and Technology Research, IEEE, 2012, pp. 501–506.

[295] J. Han, K.-K. Ma, Fuzzy color histogram and its use in color image retrieval, IEEE Trans. Image Process. 11 (8) (2002) 944–952.

[296] A. Graps, An introduction to wavelets, IEEE Comput. Sci. Eng. 2 (2) (1995) 50–61.

[297] T. Ai, X. Cheng, P. Liu, M. Yang, A shape analysis and template matching of building features by the fourier transform method, Comput. Environ. Urban Syst. 41 (2013) 219–233.

[298] B.-W. Hong, S. Soatto, Shape matching using multiscale integral invariants, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2014) 151–160.

[299] I. Fogel, D. Sagi, Gabor filters as texture discriminator, Biol. Cybern. 61 (2) (1989) 103–113.

[300] L. Liu, S. Lao, P.W. Fieguth, Y. Guo, X. Wang, M. Pietikäinen, Median robust extended local binary pattern for texture classification, IEEE Trans. Image Process. 25 (3) (2016) 1368–1381.

[301] L. Wu, S.C. Hoi, N. Yu, Semantics-preserving bag-of-words models and applications, IEEE Trans. Image Process. 19 (7) (2010) 1908–1920.

[302] T. Li, T. Mei, I.-S. Kweon, X.-S. Hua, Contextual bag-of-words for visual categorization, IEEE Trans. Circuits Syst. Video Technol. 21 (4) (2010) 381–392.

[303] M. Hossain, F. Sohel, M.F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Comput. Surv. (CSUR) 51 (6) (2019) 118.

[304] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[305] S. Haykin, Neural networks: Acomprehensive foundation, Prentice Hall PTR, 1994.

[306] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[307] S. Lloyd, Least squares quantization in pcm, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.

[308] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, 96, 1996, pp. 226–231.

[309] D. Yu, L. Deng, AUTOMATIC SPEECH RECOGNITION., Springer, 2016.

[310] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recognit. 44 (3) (2011) 572–587.

[311] K. Jacobson, M.B. Sandler, B. Fields, Using audio analysis and network structure to identify communities in on-line social networks of artists., in: ISMIR, 2008, pp. 269–274.

[312] D. Wyatt, J. Bilmes, T. Choudhury, J.A. Kitts, Towards the automated social analysis of situated speech data, in: Proceedings of the 10th International Conference on Ubiquitous Computing, ACM, 2008, pp. 168–171.

[313] D. Wyatt, T. Choudhury, J. Bilmes, J.A. Kitts, Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science, ACM Trans. Intell. Syst. Technol. (TIST) 2 (1) (2011) 7.

[314] A. Vinciarelli, Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling, IEEE Trans. Multimedia 9 (6) (2007) 1215–1226.

[315] N.P. Garg, S. Favre, H. Salamin, D. Hakkani Tür, A. Vinciarelli, Role recognition for meeting participants: an approach based on lexical information and social network analysis, in: Proceedings of the 16th ACM International Conference on Multimedia, ACM, 2008, pp. 693–696.

[316] A. Vinciarelli, Sociometry based multiparty audio recordings summarization, in: 18th International Conference on Pattern Recognition (ICPR'06), 2, IEEE, 2006, pp. 1154–1157.

[317] S. Wang, Q. Ji, Video affective content analysis: a survey of state-of-the-art methods, IEEE Trans. Affect. Comput. 6 (4) (2015) 410–430.

[318] I.B. Data, A. Hub, The four v's of big data, 2017.

[319] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S.U. Khan, The rise of "big data" on cloud computing: review and open research issues, Inf. Syst. 47 (2015) 98–115.

[320] S. Yin, O. Kaynak, Big data for modern industry: challenges and trends [point of view], Proc. IEEE 103 (2) (2015) 143–146.

[321] M.H. ur Rehman, C.S. Liew, A. Abbas, P.P. Jayaraman, T.Y. Wah, S.U. Khan, Big data reduction methods: a survey, Data Sci. Eng. 1 (4) (2016) 265–284.

[322] C.-W. Tsai, C.-F. Lai, H.-C. Chao, A.V. Vasilakos, Big data analytics: a survey, J. Big Data 2 (1) (2015) 21.

[323] P. Russom, et al., Big data analytics, TDWI Best Practices Report, Fourth Quarter 19 (4) (2011) 1–34.

[324] D. Singh, C.K. Reddy, A survey on platforms for big data analytics, J. Big Data 2 (1) (2015) 8.

[325] M. Oliveira, J. Gama, An overview of social network analysis, Wiley Interdiscip. Rev. 2 (2) (2012) 99–115.

[326] G. Nandi, A. Das, A survey on using data mining techniques for online social network analysis, Int. J. Comput. Sci. Issue. (IJCSI) 10 (6) (2013) 162.

[327] F. Bonchi, C. Castillo, A. Gionis, A. Jaimes, Social network analysis and mining for business applications, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 22.

[328] M. Zuber, A survey of data mining techniques for social network analysis, Int. J. Res. Comput. Eng. Electron. 3 (6) (2014) 1–8.

[329] A.-O. Mariam, M.G. Mohamed, S. Frederic, A survey of data mining techniques for social network analysis, School Computing. Sci. Digit. Media. Robert Gordon Univ. (2013) 1–25.

[330] P. Bonacich, Power and centrality: a family of measures, Am. J. Sociol. 92 (5) (1987) 1170–1182.

[331] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM (JACM) 46 (5) (1999) 604–632.

[332] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.

[333] D. Greene, D. Doyle, P. Cunningham, Tracking the evolution of communities in dynamic social networks, in: 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2010, pp. 176–183.

[334] B. Liu, L. Zhang, A Survey of Opinion Mining and Sentiment Analysis, in: Mining Text Data, Springer, 2012, pp. 415–463.

[335] J. Allan, Topic detection and tracking: Event-based information organization, 12, Springer Science & Business Media, 2012.

[336] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Found. Trends® Inf. Retriev. 2 (1–2) (2008) 1–135.

[337] C.D. Manning, C.D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT press, 1999.

[338] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: homophily in social networks, Annu. Rev. Soc. 27 (1) (2001) 415–444.

[339] J.S. Coleman, E. Katz, H. Menzel, Medical Innovation: a Diffusion Study, Bobbs-Merrill Co, 1966.

[340] T.W. Valente, Network models of the diffusion of innovations, Comput. Math. Org. Theory 2 (2) (1996) 163–164.

[341] J. Leskovec, L.A. Adamic, B.A. Huberman, The dynamics of viral marketing, ACM Trans. Web (TWEB) 1 (1) (2007) 5.

[342] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.

[343] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.

[344] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, in: Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on, IEEE, 2010, pp. 1–10.

[345] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, I. Stoica, Spark: cluster computing with working sets., HotCloud 10 (10–10) (2010) 95.

[346] I. Sorić, D. Dinjar, M. Štajcer, D. Oreščanin, Efficient social network analysis in big data architectures, in: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2017, pp. 1397–1400.

[347] J.S. Andersen, O. Zukunft, Evaluating the scaling of graph-algorithms for big data using graphx, in: 2016 2nd International Conference on Open and Big Data (OBD), IEEE, 2016, pp. 1–8.

[348] R.S. Xin, J.E. Gonzalez, M.J. Franklin, I. Stoica, Graphx: a resilient distributed graph system on spark, in: First International Workshop on Graph Data Management Experiences and Systems, ACM, 2013, p. 2.

[349] J.E. Gonzalez, R.S. Xin, A. Dave, D. Crankshaw, M.J. Franklin, I. Stoica, Graphx: Graph processing in a distributed dataflow framework, in: 11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14), 2014, pp. 599–613.

[350] A.B. Bondi, Characteristics of scalability and their impact on performance, in: Proceedings of the 2nd International Workshop on Software and Performance, ACM, 2000, pp. 195–203.

[351] H. El-Rewini, M. Abd-El-Barr, Advanced computer architecture and parallel processing, 42, John Wiley & Sons, 2005.

[352] N.J. Gunther, A simple capacity model of massively parallel transaction systems, CMG-CONFERENCE-, COMPSCER MEASUREMENT GROUP INC, 1993. 1035–1035

[353] N.J. Gunther, What is Guerrilla Capacity Planning?, Springer, 2007.

[354] P. Jogalekar, C. Woodside, A scalability metric for distributed computing applications in telecommunications, in: Teletraffic Science and Engineering, 2, Elsevier, 1997, pp. 101–110.

[355] A. Giessler, J. Haenle, A. König, E. Pade, Free buffer allocation—an investigation by simulation, Comput. Netw. (1976) 2 (3) (1978) 191–208.

[356] D. Sanchez, O. Solarte, V. Bucheli, H. Ordonez, Evaluating the scalability of big data frameworks, Scalab. Comput. 19 (3) (2018) 301–307.

[357] A.Y. Grama, A. Gupta, V. Kumar, Isoefficiency: measuring the scalability of parallel algorithms and architectures, IEEE Parallel Distrib. Technol. 1 (3) (1993) 12–21.

[358] A. Gonzalez-Pardo, J.J. Jung, D. Camacho, Aco-based clustering for ego network analysis, Future Generat. Comput. Syst. 66 (2017) 160–170.

[359] J. Heer, M. Bostock, V. Ogievetsky, et al., A tour through the visualization zoo., Commun. Acm 53 (6) (2010) 59–67.

[360] P. Karampelas, Visual methods and tools for social network analysis, Encycl. Soc. Netw. Anal. Min. (2014) 2314–2327.

[361] M. Krzywinski, I. Birol, S.J. Jones, M.A. Marra, Hive plots—rational approach to visualizing networks, Brief. Bioinformat. 13 (5) (2011) 627–644.

[362] W.J. Longabaugh, Combing the hairball with biofabric: a new approach for visualization of large networks, BMC Bioinformatics 13 (1) (2012) 275.

[363] M. Wattenberg, Visual exploration of multivariate graphs, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2006, pp. 811–819.

[364] P. Eades, A heuristic for graph drawing, Congress. Numerant. 42 (1984) 149–160.

[365] T.M. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Software: Practice and experience 21 (11) (1991) 1129–1164.

[366] T. Kamada, S. Kawai, et al., An algorithm for drawing general undirected graphs, Inf Process Lett 31 (1) (1989) 7–15.

[367] S. Hachul, M. Jünger, Drawing large graphs with a potential-field-based multilevel algorithm, in: International Symposium on Graph Drawing, Springer, 2004, pp. 285–295.

[368] E.R. Gansner, Y. Hu, S. North, A maxent-stress model for graph layout, IEEE Trans. Vis. Comput. Graph. 19 (6) (2012) 927–940.

[369] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software, PLoS ONE 9 (6) (2014) e98679.

[370] Y. Hu, L. Shi, Visualizing large graphs, Wiley Interdiscip. Rev. Comput. Stat. 7 (2) (2015) 115–136.

[371] C. Mueller, B. Martin, A. Lumsdaine, A comparison of vertex ordering algorithms for large graph visualization, in: 2007 6th International Asia-Pacific Symposium on Visualization, IEEE, 2007, pp. 141–148.

[372] N. Henry, J.-D. Fekete, Matrixexplorer: a dual-representation system to explore social networks, IEEE Trans. Vis. Comput. Graph. 12 (5) (2006) 677–684.

[373] N. Henry, J.-D. Fekete, Matlink: Enhanced matrix visualization for analyzing social networks, in: IFIP Conference on Human-Computer Interaction, Springer, 2007, pp. 288–302.

[374] S. Chen, L. Lin, X. Yuan, Social media visual analytics, in: Computer Graphics Forum, 36, Wiley Online Library, 2017, pp. 563–587.

[375] G.A. Pavlopoulos, D. Paez-Espino, N.C. Kyrpides, I. Iliopoulos, Empirical comparison of visualization tools for larger-scale network analysis, Adv. Bioinformatic. 2017 (2017).

[376] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in: The Craft of Information Visualization, Elsevier, 2003, pp. 364–371.

[377] J. Bertin, W.J. Berg, H. Wainer, Semiology of graphics: diagrams, networks, maps, 1, University of Wisconsin press Madison, 1983.

[378] B. Braden, The surveyor's area formula, College Math. J. 17 (4) (1986) 326–337.

[379] G. Csardi, T. Nepusz, et al., The igraph software package for complex network research, Inter J. Complex Syst. 1695 (5) (2006) 1–9.

[380] J. Aasman, Allegro graph: rdf triple database, Cidade: Oakland Franz Incorporated 17 (2006).

[381] I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, Lanet-vi in a nutshell, 2006.

[382] J. Leskovec, R. Sosič, Snap: a general-purpose network analysis and graph-mining library, ACM Trans. Intell. Syst. Technol. (TIST) 8 (1) (2016) 1.

[383] J.M. Pullen, The network workbench: network simulation software for academic investigation of internet concepts, Comput. Netw. 32 (3) (2000) 365–378.

[384] P. Mazzuca, Y. t,"circulo: A community detection evaluation framework," feb. 2015.

[385] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.

[386] J. O'Madadhain, D. Fisher, S. White, Y. Boey, The jung (java universal network/graph) framework, University of California, Irvine, California (2003).

[387] K.T. Bartusiak R., Sparklinggraph: large scale, distributed graph processing made easy, 2017. Manuscript in preparation.

[388] A. Hagberg, P. Swart, D. S Chult, Exploring network structure, dynamics, and function using NetworkX, Technical Report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[389] V. Batagelj, A. Mrvar, Pajek—analysis and visualization of large networks, in: Graph Drawing Software, Springer, 2004, pp. 77–103.

[390] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Third International AAAI Conference on Weblogs and Social Media, 2009.

[391] S.P. Borgatti, M.G. Everett, L.C. Freeman, Ucinet, Encycloped. Soc. Netw. Anal. Min. (2014) 2261–2267.

[392] J. Heer, S.K. Card, J.A. Landay, Prefuse: a toolkit for interactive information visualization, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2005, pp. 421–430.

[393] J. Ellson, E. Gansner, L. Koutsofios, S.C. North, G. Woodhull, Graphviz—open source graph drawing tools, in: International Symposium on Graph Drawing, Springer, 2001, pp. 483–484.

[394] M. Lal, Neo4j Graph Data Mmodeling, Packt Publishing Ltd, 2015.