ELSEVIER

(2020) 1–11

Journal Logo

www.elsevier.com/locate/procedia

Paper summary: The Four Dimensions of Social Network Analysis

David Camacho

Departamento de Sistemas Informáticos
Universidad Politécnica de Madrid, Spain

Abstract

A summary of the paper entitled: **The Four Dimensions of Social Network Analysis: An Overview of Research Methods, Applications, and Software Tools** by David Camacho, Angel Panizo-LLedot, Gema Bello-Orgaz, Antonio Gonzalez-Pardo, and Erik Cambria is presented. The paper [1], published in *Information Fusion* journal, provides three main contributions: 1) an up-to-date literature review of the state of the art on social network analysis; 2) a set of new metrics based on four essential features (*dimensions*) in social network analysis has been proposed by authors; 3) and finally, a quantitative analysis of a set of popular social network analysis tools and frameworks. This work defines four different dimensions, namely *Pattern & Knowledge discovery*, *Information Fusion & Integration*, *Scalability*, and *Visualization*, which are later used to define a set of new metrics (termed *degrees*) in order to evaluate the different software tools and frameworks of social network analysis. This paper proposes a new way to evaluate and measure the maturity of social network technologies, looking for both a quantitative assessment, challenges and future trends in this active area. Finally, a set of 20 SNA-software tools are ranked following previous metrics, to be later analyzed taking into account the (technology) dimensions defined in this work.

Keywords:
Social Network Analysis, Social Media Mining, Social Data Visualization, Data Science, Big Data

Section 1: Introduction

This short summary presents a brief description of the overview and research results presented in the paper: *"The Four Dimensions of Social Network Analysis: An Overview of Research Methods, Applications, and Software Tools", by David Camacho, Angel Panizo-LLedot, Gema Bello-Orgaz, Antonio Gonzalez-Pardo, and Erik Cambria* [1]. All of the referenced sections, results (tables, figures, equations) are related to this paper. This summary only tries to give a general perspective on what the reader could expect from its complete read.

This paper is focused on the domain of online social networks (OSNs), and its related social networks analysis (SNA) software currently used to work on them. The area of SNA research generates thousand of

Email address: david.camacho@upm.es (David Camacho)

papers per year, hundred of different algorithms, tools, and frameworks, to tackle the challenges and open issues related to OSNs [2]. The main contributions of [1] 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, or by their specific application domain.

2. The definition of four new "SNA-*Dimensions*", the concept is inspired by the popular V-models [3] used in the Big Data area, and its goal is to measure the capacity (and maturity) of the different frameworks and tools available to perform SNA tasks. These dimensions are used to define a set of **metrics** (that we named *degrees*), that will allow any researcher to identify the technology readiness level, and the main challenges and trends in the area of SNA. The dimensions defined will be directly related to *Pattern & Knowledge discovery, Information Fusion & Integration, Scalability*, and *Visualization* research topics. These dimensions try to to shed light to the following research questions (RQ):

   RQ1) **What can I discover?** (*Pattern & Knowledge discovery*): This is the most classic and studied characteristic in OSNs, and it is related to the capacity to discover new knowledge and patterns inside the networks(in form of network structure and topology, statistical correlations, information flow and diffusion, content analysis, opinion mining, user profiling, etc.).

   RQ2) **What is the limit?** (*Scalability*): Due to the fast growth of OSNs, this is a critical dimension based on the capacity of algorithms, methods and frameworks to work with large amounts of data in an adequate time.

   RQ3) **What kind of data can I integrate?** (*Information Fusion & Integration*): This dimension will be related to the capacity of fusing different kind of sources (text, video, images, audio). The currently available OSNs provide data in different formats, which allow to define different type of networks (as Multilayer SNA), which can be later integrated to generate new knowledge.

   RQ4) **What can I show?** (*Visualization*): Visualization in OSN is one of the most powerful tool used in this area, due to the huge amount of available data and the complex information stored, an adequate visualization of this information is always a challenging task for any OSN analyst.

3. The definition of a new *global Capability metric*, named $\mathfrak{C}_{SNA}$, based on the just mentioned metrics, which can be used to rank the capabilities of any SNA technology, framework or tool.

4. The assessment of 20 relevant (and popular) frameworks and tools, currently used to perform SNA tasks. The assessment has been carried out using our proposed dimensions, and their related metrics (*degrees*). This analysis not only shows what are the main strengths and weaknesses of the different frameworks, it also provides a useful guide for those researchers interested in the area of SNA.

The paper [1], has been structured in the following sections: Section 2 performs a scientometric study over the paper published in the last 5 years to extract both, the most active research areas and the most relevant application domains in SNA. Sections 3 and 4, carry out a detailed analysis of both the fundamental research areas and several application domains. Section 5 provides a discussion of the main research methods and application domains studied. Section 6 is dedicated to the definition of a new concept, termed *dimensions*, which is used to define the maturity of SNA technologies. Using these dimensions, a set of metrics (named *degrees*) are defined. Due the complexity of the field, and the increasing amount of available technology for OSN, we have selected, studied, and analyzed a set of SNA software tools and frameworks. Section 7 provides the assessment of 20 relevant frameworks and tools. And finally Section 8, gives the main conclusions, challenges and future trends in the area of SNA. **This document is just a summary of paper [1], so only Sections 2, 3, 4, 6 and 7 will be briefly described**.

Section 2: SNA Scientometric Analysis

This section provides a scientometric analysis, it is presented a review of works related to SNA using Web of Science as search engine, covering the highly cited articles over a five years period (from 2014 to 2018). A record collection of **28.805 articles** has been gathered for this time period. Only it has been considered in the scienciometric analysis those authors that have used in their publications the keyword (#hashtag) of *#social network analysis (SNA)*. In order to identify the most relevant topics related to SNA, a textual analysis has be 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, by processing the "keywords" used in the article collection, to later visualize the most frequent terms (i.e., keywords) as a world cloud.

From the scienciometric analysis carried out, and 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 domains, are more relevant in terms of research 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 us 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 (that clearly fits with our idea of defining a new set of dimensions, which can be used to evaluate the maturity of technology).

Section 3: Techniques and algorithms

This section has been dedicated to make an up-to-date literature review of the current state of the art in SNA, the research areas selected have been obtained from the scientiometric analysis carried out in Section 2. To organize this revision, it has been divided into two different approaches [4]: *Structural-based analysis* and *Content-based analysis*.

- **Structural-based Analysis**. This subsection studies those research areas related to the the analysis of networks using network theory (usually known as graph theory) [5]. Therefore, a short review of *graph theory and network analytics* [6] (including some basics on graphs [7], network models [8, 9], network metrics [10], and graphs algorithms[11]); *community detection algorithms* (non-overlapping [12, 13, 14] vs. overlapping [15, 16, 17], static [18, 19] vs. dynamic algorithms [20, 21, 22]); and *information diffusion models* [23, 24], are introduced.
- **Content-based Analysis**. This subsection is focused on the OSN content analysis (mainly based on natural language processing), for this reason areas as *user profiling* [25, 26], *topic extraction* [27, 28], and *sentiment analysis* [29, 30], are introduced and studied.

Section 4: Application domains

Due to the huge number of application domains in OSNs, we have selected a subset of them (using the scientometric study previously performed). This study helped us select the next application domains: *Healthcare* [31, 32, 33], *Marketing* [34, 35], *Tourism and Hospitality* [36, 37], and *Cyber Security* [38, 39].

Also, and out of the scope of our scientometric analysis, we have decided to make a brief analysis of the state-of-the-art in some emerging areas, which are currently experiencing an increasing interest in the area of SNA. These emerging areas are directly related to highly societal demanding topics, such as: *politics* [40, 41]; *detection of fake news and misinformation* [42, 43]; and the integration of *multimedia information* [44, 45].

Section 6: The Four Dimensions of Social Network Analysis

The Big Data paradigm was characterized by several "V-models" that allow any researcher to analyze the capacity of the different Big Data technologies. Initially, 3Vs were described in the 3V model [3], but this model has evolved during the last years to other more complex models (4v, 5V, or 6V). These models allow to measure the maturity of different methods, tools and technologies based on Big Data using simple features such as *Volume*, *Velocity*, *Variety*, *Value*, Veracity or Variability. We have mapped our four research questions stated in Section 1, into a set of 4 equivalent **dimensions**:

D1) **Pattern & Knowledge discovery**. The concept of *Value*, as the knowledge or pattern discovery capacity of any method or algorithm, 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 to discover non-trivial knowledge from OSN. The objective of this dimension is to evaluate, any type of technique, method, or tool, which is used to discover new knowledge in OSN. In this section several works have been analyzed to obtain a taxonomy of the main functionalities for discovering knowledge which can be embedded in SNA tools. These functionalities can be summarized in: *Qualitative and quantitative/statistical analysis* ($F_{Value(1,i)}$): computation of measures based on the topology ($F_{Value(1,1)}$), and link analysis ($F_{Value(1,2)}$); *Pattern mining* methods ($F_{Value(2,i)}$): community detection ($F_{Value(2,1)}$), homophily models ($F_{Value(2,2)}$), and opinion mining ($F_{Value(2,3)}$); *Predictive analysis* ($F_{Value(3,i)}$): propagation and virality modeling ($F_{Value(3,1)}$), and link prediction ($F_{Value(3,2)}$).

The proposed taxonomy is used to quantify the **degree of value** ($d_{Value}(t)$), that each SNA tool (t) provides according to the functionalities that it covers. These functionalities are quantitative assessed as shown in Equation 1. Several weights are used (α, β, and γ), to represent the importance given to each characteristic. In this work, all the characteristics 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 ($d_{Value}(t) \in [0, 1]$).

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

D2) **Scalability**. This dimension will be used to define, and quantify, the scalability capacity of a tool or technique (e.g., algorithm) used in an OSN (equivalente to the "Volume" feature on the V-models). 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), so it will try to answer the question "*What is the limit?*. 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). Taking into account the different taxonomies of scalability characteristics described, and analysed in the paper [1], and adapting them specifically for SNA methods and techniques, the following sets of measures are proposed to quantify the **degree of volume** ($d_{Volumne}(t)$), or scalability, for a SNA tool: *Space-Time scalability* ($F_{Volume1}$); *Parallelism scalability* ($F_{Volume2}$); *Functional scalability* ($F_{Volume3}$); and *Heterogeneous-Integration scalability* ($F_{Volume4}$).

Previous features have been combined to generate a scalability degree (d_{Volume}) for any tool (t), as it is shown in Equation 2, where scalability ($d_{Volumne}(t)$) is rated from 0 to 1 depending on its capacity to scale to large data sets ($d_{Volume}(t) \in [0, 1]$).

$$d_{Volume}(t) = \frac{\sum_{i=1}^{4} F_{Volumei}(t)}{4} \quad (2)$$

D3) **Information Fusion & Integration**. This dimension 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: the number of different type of data (*multichannel* (F_{Var1})); the number of different OSNs used to extract the data (*multimodality* (F_{Var2})); and the representation of this data into the model (*multi-representation* (F_{Var3})). The **degree of variety** ($d_{Variety}(t)$), 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. This dimension has been defined as Equation 3 shows:

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

D4) **Visualization**. Finally, 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, so this dimension will be used to answer the research question "*What can I saw?*. After analyzing the state-of-the-art related to visualization in OSN, 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 have described the visualization dimension in a twofold way. On the one hand, using the "*Visual Variables*" characteristics of the tool: Position, Size, Shape, Orientation, Colour, Saturation and Texture ($F_{VisVari}$)). On the other hand, using an extra characteristic, named "*Interactions*": Zoom, Filter, Highlight, Grouping, and Multiview (F_{Interi}), that will help us to asses the information representation capacity of a tool. In order to generate a single value capable of evaluating the **degree of visualization** dimension ($d_{Visual}(t)$), Equation 4 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.

$$d_{Visual}(t) = \alpha \cdot \frac{\sum_{i=1}^{7} \gamma_i \cdot F_{VisVari}(t)}{7} + \beta \cdot \frac{\sum_{j=1}^{5} \theta_j \cdot F_{Interj}(t)}{5} \tag{4}$$

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).

The four basic research questions proposed in this work, 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.

Table 1: Summary on dimensions and the quantitative metrics (degrees) proposed in [1].

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)	$d_{Visual}(t) = 1/2 \cdot \frac{\sum_{i=1}^{7} F_{VisVari}(t)}{7} + 1/2 \cdot \frac{\sum_{j=1}^{5} F_{Interj}(t)}{5}$	$[0, 1]$

Finally, and considering these $SNA_{degrees}$ it is quite straightforward to define a new **global metric**, which 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. 1). This equation comes from the Shoelace formula, also known as Gauss's area formula and the surveyor's formula, and it is a simple formula for finding the area of a polygon given the coordinates of its vertices, as it is shown in [1], the general equation can be mapped into Equation 5. Figure 1 shows an example of our capability metric ($\mathfrak{C}_4(t)$) on three hypothetical frameworks, whereas Table 2 shows the quantitative values. This general metric can be used to better understand the capability of a particular SNA technology, and can be used to provide a ranking between any SNA software considered.

$$\mathfrak{C}_4(t) = \frac{1}{2} \cdot \left| \sum_{i=1=Value}^{3=Variety} \left(d_i^{x_i}(t) \cdot d_i^{y_{i+1}}(t) + d_i^{x_n}(t) \cdot d_i^{y_1}(t)\right) - \sum_{i=1=Value}^{3=Variety} \left(d_i^{x_{i+1}}(t) \cdot d_i^{y_i}(t) - d_i^{x_1}(t) \cdot d_i^{y_{x_n}}(t)\right) \right| \quad (5)$$

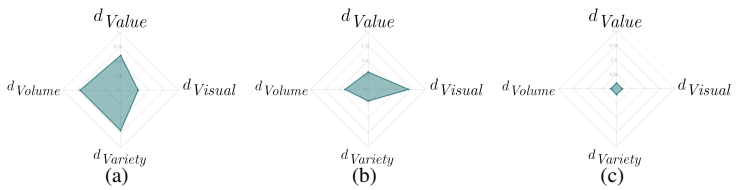


Fig. 1: Spider diagrams representing the evaluation of three (hypothetical) SNA platforms. The values for each framework/dimension are shown in Table 2. The figures shown correspond to a good platform (a), an intermediate tool (b) and a poor platform (c).

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)

Section 7: Frameworks & Tools Analysis

This section provides a review on a set of popular SNA frameworks and tools, which are extensively used by the research community and industry. Because it would be highly difficult to analyze all of the currently available tools, 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 between the community (taking into account the popularity of some tools in published works, websites and other technical material). The 20 different SNA-software analyzed were: ***Igraph, AllegroGraph, LaNet-vi, Stanford Network analysis Platform (SNAP), ORA-LITE/PRO, Network workbench, NetMiner, Circulo, Cytoscape, JUNG, SparklingGraph, NetworkX, Pajek, GraphX Apache Spark, Gephi, UCINET, Prefuse, Graphistry, GraphViz, and Neo4j***.

Later, 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.

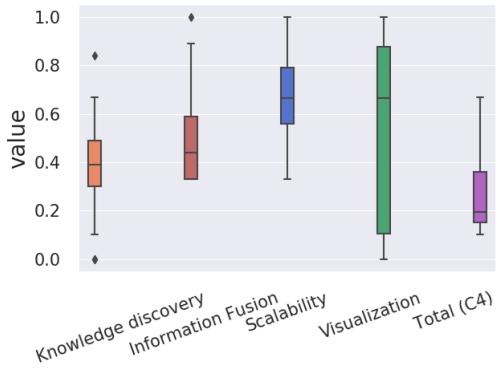


Fig. 2: Distribution of the different dimensions values achieved by the analyzed tools. The X-axis represents the proposed dimensions, whereas the Y-axis shows the values obtained once the quantitative metrics ($SNA_{degrees}$) are calculated.

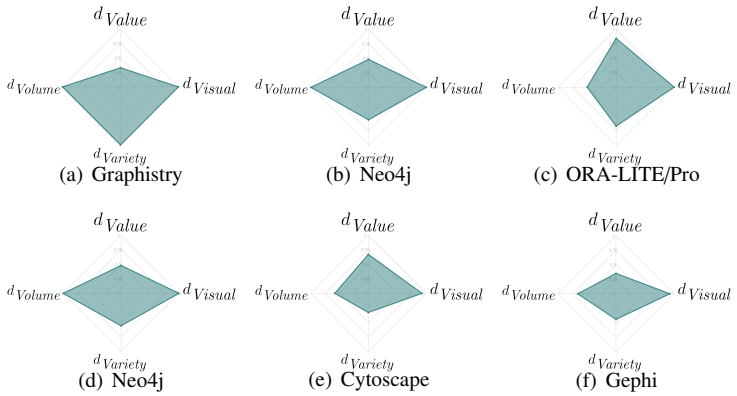
The evaluation process for each software was carried out as follows: once the initial set of tools was selected, the authors agreed an evaluation rubric (that is publicly available) to assess the SNA-software. This rubric is based on the analysis of the software documentation, their official websites (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, 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.

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



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

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

Fig. 3: 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

As an example of the analysis carried out in [1], and regarding to the *global metric* score ($\mathfrak{C}_4(t)$). Table 3 shows the top-5 "Proprietary or Open-source", and "Only Open-Source" (licenses) SNA-software. In addition to the aforementioned table, Fig. 2 is presented. This figure shows the distribution of the dimension scores in order to allow the reader to contextualize the values obtained (related to the whole set of tools and frameworks analyzed).

Fig. 3 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. 2). A similar case can be found for the *Information Fusion* dimension, where Graphistry tool provides the highest possible score (see Fig. 2). 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* dimension 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). Fig. 4 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

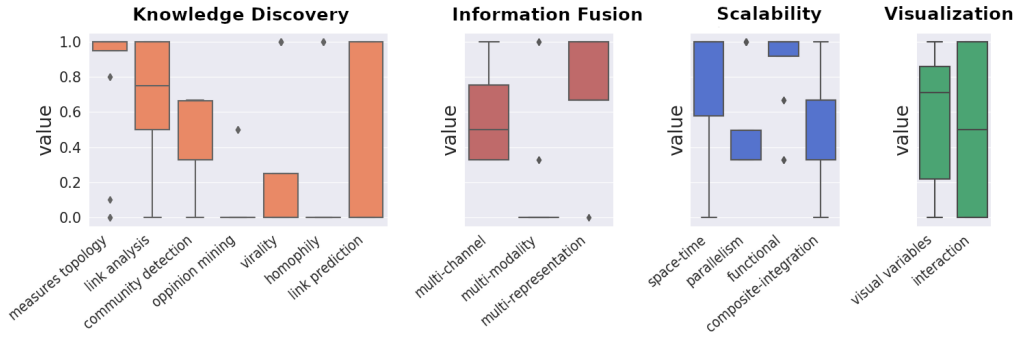


Fig. 4: Un-aggregated dimension distributions for the top-5 SNA-software tools, each color represents a different dimension. The X-axis contains each of the features that form a dimension (SNA_{degree}), and the Y-axis their numerical value.

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. The paper [1] provides a complete (and detailed) analysis of all of those tools from both perspectives, the dimensions/metrics proposed and the global capability metric, to finally analyze the current weaknesses, strengths and future room for improvement in SNA technologies.

Citation and Open Access to 4-dimensions SNA results

Please cite this work as:

[1] David Camacho, Angel Panizo-LLedot, Gema Bello-Orgaz, Antonio Gonzalez-Pardo, & Erik Cambria (2020). *The four dimensions of social network analysis: An overview of research methods, applications, and software tools*. Information Fusion, Vol. 1, pp. 1–65

In order to allow researchers 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 all of the information and results carried out in this work:

The SNA 4-Dimensions website:

https://ai-da-sna.github.io/

References

[1] D. Camacho, A. Panizo-Lledot, G. Bello-Orgaz, A. Gonzalez-Pardo, E. Cambria, The four dimensions of social network analysis: An overview of research methods, applications, and software tools, Information Fusion (In press) (2020) 1–65.

[2] G. Bello-Orgaz, J. J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Information Fusion 28 (2016) 45–59.

[3] D. Laney, 3d data management: Controlling data volume, velocity and variety, META group research note 6 (70) (2001) 1.

[4] C. C. Aggarwal, An introduction to social network data analytics, in: Social network data analytics, Springer, 2011, pp. 1–15.

[5] J. A. Barnes, F. Harary, Graph theory in network analysis, Social networks 5 (2) (1983) 235–244.

[6] D. B. West, Introduction to graph theory, Vol. 2, Prentice hall Upper Saddle River, NJ, 1996.

[7] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: Mining text data, Springer, 2012, pp. 415–463.

[8] M. E. Newman, D. J. Watts, S. H. Strogatz, Random graph models of social networks, Proceedings of the National Academy of Sciences 99 (suppl 1) (2002) 2566–2572.

[9] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Reviews of modern physics 74 (1) (2002) 47.

[10] R. Zafarani, M. A. Abbasi, H. Liu, Social media mining: an introduction, Cambridge University Press, 2014.

[11] S. Even, Graph algorithms, Cambridge University Press, 2011.

[12] S. Fortunato, Community detection in graphs, Physics reports 486 (3-5) (2010) 75–174.

[13] G. Bello-Orgaz, H. D. Menéndez, D. Camacho, Adaptive k-means algorithm for overlapped graph clustering, International journal of neural systems 22 (05) (2012) 1250018.

[14] H. Gmati, A. Mouakher, A. Gonzalez-Pardo, D. Camacho, A new algorithm for communities detection in social networks with node attributes, Journal of Ambient Intelligence and Humanized Computing (Special Issue on Computational Intelligence for Social Mining) (2018) 1–13.

[15] J. Xie, S. Kelley, B. K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, ACM Computing Surveys (CSUR) 45 (4) (2013) 43.

[16] G. Bello-Orgaz, D. Camacho, Evolutionary clustering algorithm for community detection using graph-based information, in: 2014 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2014, pp. 930–937.

[17] G. Bello-Orgaz, S. Salcedo-Sanz, D. Camacho, A multi-objective genetic algorithm for overlapping community detection based on edge encoding, Information Sciences 462 (2018) 290–314.

[18] L. Tang, H. Liu, Graph mining applications to social network analysis, in: Managing and Mining Graph Data, Springer, 2010, pp. 487–513.

[19] A. Gonzalez-Pardo, J. J. Jung, D. Camacho, Aco-based clustering for ego network analysis, Future Generation Computer Systems 66 (2017) 160 – 170.

[20] R. Cazabet, F. Amblard, Dynamic Community Detection, Springer New York, New York, NY, 2014, pp. 404–414.

[21] E. Osaba, J. Del 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.

[22] 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 Generation Computer Systems. Available online 15 November 2019. (2019) 1–16.

[23] R. Agnihotri, R. Dingus, M. Y. Hu, M. T. Krush, Social media: Influencing customer satisfaction in b2b sales, Industrial Marketing Management 53 (2016) 172–180.

[24] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical review letters 86 (14) (2001) 3200.

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

[26] 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.

[27] D. Rajagopal, D. Olsher, E. Cambria, K. Kwok, Commonsense-based topic modeling, in: Proceedings of the second international workshop on issues of sentiment discovery and opinion mining, 2013, pp. 1–8.

[28] G. Piao, J. Breslin, Inferring user interests in microblogging social networks: A survey, User Modeling and User-Adapted Interaction 28 (3) (2018) 277–329.

[29] S. L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: From formal to informal and scarce resource languages, Artificial Intelligence Review 48 (4) (2017) 499–527.

[30] I. Chaturvedi, E. Cambria, R. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Information Fusion 44 (2018) 65–77.

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

[32] 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.

[33] G. Bello-Orgaz, J. Hernandez-Castro, D. Camacho, Detecting discussion communities on vaccination in twitter, Future Generation Computer Systems 66 (2017) 125–136.

[34] 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, Journal of interactive marketing 26 (2) (2012) 83–91.

[35] 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, International Journal of Research in Marketing 33 (1) (2016) 27–41.

[36] A. M. Munar, J. K. S. Jacobsen, Motivations for sharing tourism experiences through social media, Tourism management 43 (2014) 46–54.
[37] J. Li, L. Xu, L. Tang, S. Wang, L. Li, Big data in tourism research: A literature review, Tourism Management 68 (2018) 301–323.
[38] R. Lara-Cabrera, A. Gonzalez-Pardo, K. Benouaret, N. Faci, D. Benslimane, D. Camacho, Measuring the radicalisation risk in social networks, IEEE Access 5 (2017) 10892–10900.
[39] J. Torregrosa, J. Thorburn, R. Lara-Cabrera, D. Camacho, H. M. Trujillo, Linguistic analysis of pro-isis users on twitter, Behavioral Sciences of Terrorism and Political Aggression 0 (0) (2019) 1–15.
[40] 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.
[41] 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.
[42] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (1) (2017) 22–36.
[43] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (6380) (2018) 1146–1151.
[44] Q. Cheng, Q. Zhang, P. Fu, C. Tu, S. Li, A survey and analysis on automatic image annotation, Pattern Recognition 79 (2018) 242–259.
[45] S. Wazarkar, B. N. Keshavamurthy, A survey on image data analysis through clustering techniques for real world applications, Journal of Visual Communication and Image Representation 55 (2018) 596–626.