Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090529

Representing Social Influencers and Influence using Power-Law Graphs

Chiara Francalanci, Ajaz Hussain and Francesco Merlo*

Department of Electronics, Information and Bio-Engineering, Politecnico di Milano, I - 20133 Milano, Italy

Received: 30 Jan. 2015, Revised: 1 May 2015, Accepted: 2 May 2015
Published online: 1 Sep. 2015

Abstract: This paper focuses on semantic networks that represent the user opinions expressed by social media users on a given set of topics. These networks are found to follow a power-law degree distribution of nodes, with a few hub nodes and a long tail of peripheral nodes. While there exist consolidated approaches supporting the identification and characterization of hub nodes, research on the analysis of the multi-layered distribution of peripheral nodes is limited. In social media, hub nodes represent social influencers. However, the literature provides evidence of the multi-layered structure of influence networks, emphasizing the distinction between influencers and influence. The latter seems to spread following multi-hop paths across nodes in peripheral network layers. This paper proposes a visual approach to the graphical representation of peripheral layers. The core concept of our approach is to partition the node set of a graph into `hub` and `peripheral` nodes. Then, a modified force-directed method is applied to clearly display local multi-layered neighborhood clusters around hub nodes. Our approach is tested on a large sample of tweets from tourism domain. Our algorithm is visually compared with state-of-the-art network drawing techniques.

Keywords: Semantic networks,Power law graphs,Social media influencers,Social media influence,User Opinions

1 Introduction

Most network visualization methodologies and tools focus on identifying network hubs. Hubs represent central nodes connecting sets of more peripheral nodes that are rather sparse and separate from each other, as discussed by [45]. Literature has focused on measuring centrality and provides a broad array of centrality metrics, each of them highlighting a different aspect of a hub's prominent role. As discussed by [19], *degree centrality* measures the absolute number of connections of a node, *closeness centrality* measures how far a node is from all other nodes in the network along the overall shortest paths, while *betweenness centrality* assesses the role of a node as a hub of information by analyzing the extent to which the node connects separate subnetworks. These metrics represent the underlying concept of many network visualization tools. The assumption that most tools make to visualize large networks is that hubs represent the main driver of the structure of networks and, if they exist, they should be clearly highlighted to cope with complexity and obtain a nice and intuitive representation of the network.

The literature on social media makes a distinction between influencers and influence [11,30]. The former are social media users with a broad audience. For example, influencers can have a high number of followers on *Twitter*, or a multitude of friends on *Facebook*, or a broad array of connections on *LinkedIn*. The term influence is instead used to refer to the social impact of the content shared by social media users. The breadth of the audience was considered the first and foremost indicator of influence for traditional media, such as television or radio. However, traditional media are based on broadcasting rather than communication, while social media are truly interactive. It is very common that influencers say something totally uninteresting and, as a consequence, they obtain little or no attention. On the contrary, if social media users are interested in something, they typically show it by participating in the conversation with a variety of mechanisms and, most commonly, by sharing the content that they have liked. [8] has noted that a content that has had an impact on a user's mind `is` shared. Influencers are prominent social media users, but we cannot expect that the content that they share is bound to have high influence, as discussed by [6].

In previous research, Bruni et al. [10] has shown how the content of messages can play a critical role and can be

* Corresponding author e-mail: ajaz.hussain@polimi.it

a determinant of the social influence of a message irrespective of the centrality of the message's author. Results suggest that peripheral nodes can be influential: this paper starts from the observation made by [12] that social networks of influence follow a power-law distribution function, with a few hub nodes and a long tail of peripheral nodes, consistent with the so-called small-world phenomenon as noted by [45]. In social media, hub nodes represent social influencers, but influential content can be generated by peripheral nodes and spread along possibly multi-hop paths originated in peripheral network layers.

The ultimate goal of our research is to understand how influential content spreads across the network. For this purpose, identifying and positioning hub nodes is not sufficient, while we need an approach that supports the exploration of peripheral nodes and of their mutual connections. In this paper, we exploit a modified force-directed algorithm [24] to highlight the local multi-layered neighborhood clusters around hub nodes. The algorithm is based on the idea that hub nodes should be prioritized in laying out the overall network topology, but their placement should depend on the topology of peripheral nodes around them. In our approach, the topology of the periphery is defined by grouping peripheral nodes based on the strength of their link to hub nodes, as well as the strength of their mutual interconnections.

The approach is tested on a large sample of tweets expressing opinions on a selection of Italian locations relevant to the tourism domain. Tweets have been semantically processed and tagged with information on `a)` the location to which they refer, `b)` the location's brand driver (or category) on which authors express an opinion, `c)` the subject referred by the author, `d)` the number of retweets, and `e)` the identifier of the retweeting author. With this information, we draw corresponding multi-mode networks highlighting the connections among authors (retweeting), and their interests (brand or category). The data sample is referred to the tourism domain. We have adopted a modified version of the Anholt's Nation Brand index model to define a set of categories of content referring to specific brand drivers of a destination's brand [2]. Based on a set of qualitative criteria, we visually compare the effectiveness of our approach in highlighting features of the networks relevant to understand the influence of content with previous state-of-the-art algorithms based on the traditional spring and force-directed approaches (see, for example, [13] and [20]). Results highlight the effectiveness of our approach, providing interesting insights on how unveiling the structure of the periphery of the network can visually show the potential of peripheral nodes in determining influence.

The presentation is organized as follows. Section 2 discusses influence in social media, limitations of existing semantic network drawing techniques and tools, and standard graph drawing aesthetics criteria. Section 3 discusses the implementation aspects of our work. Section 4 presents the experimental methodology, performance evaluation, and benchmark comparison. Conclusions are drawn in Section 5.

2 State of the Art

In this section, we will discuss about limitations of existing network visualization techniques and tools. We will also highlights the most common and widely accepted visualization aesthetic criteria.

2.1 Network Visualization Techniques and Tools

The first spring-embedded model for network visualization was proposed by [15], who have simplified the formulae used to compute spring forces, and made significant improvements by using a cooling schedule to limit nodes' maximum displacement. However, the repulsive force was still computed between all node pairs, yielding to an overall computational complexity of $O(N^2)$ for a network with N nodes. Subsequent studies that took a similar approach are the Online Force Directed Animated Visualization (OFDAV) technique by [23], and the edge-edge repulsion approach by [34]. More recently, [44] has proposed the over relaxation algorithm for force directed drawing. Despite these efforts, these force-directed algorithms are still considered non-scalable and unsuitable for large networks, also noted by [21].

Several research efforts in network visualization have targeted power-law algorithms and their combination with the traditional force-directed techniques, as for example in [27,1]. Among these approaches, the most notable is the Out-Degree Layout (ODL) for the visualization of large-scale network topologies, presented by [38,12]. The core concept of the algorithm is the segmentation of the network nodes into multiple layers based on their out-degree, i.e. the number of outgoing edges of each node. The positioning of network nodes starts from those with the highest out-degree, under the assumption that nodes with a lower out-degree have a lower impact on visual effectiveness.

The most common and successful visualization tools are surveyed in [39,28,35] and [43]. Widely discussed tools include `Cytoscape`, `OntoGraf`, `OntSphere`, `Giny`, `graphViz`, `Hyper Graph`, `rdf Gravity`, `IsaViz`, `Jambalaya`, `Owl2Prefuse`, `Flow inspector`, `Gephi` and `SocNetV`. There is no one-to-one mapping between techniques and tools. This section discusses usage results from the literature or from experimental evidence that we made with the tools.

Most of the tools are not highly scalable and with large-scale graphs, they are time inefficient or produce ambiguous layouts. Many visualization tools support graphs up to a few hundred nodes, such as `rdfGravity`

[21], `Jambalaya` [41], `GraphViz` [16], and `Flow inspector` [9]. With large-scale graphs, they are time inefficient or produce ambiguous layouts, as observed by [21] with `rdfGravity`. Node cluttering issues and edge overlap issues are common, as in `Prefuse` [22], `Gephi` [5], `GraphViz`, and `OntoGraf` [17]. Force-directed and spring layouts are implemented in several visualization tools, but local minima problems are common, as observed in `SocNetV` [25], `Gephi`, and in `Flow inspector`.

The most practical limitations that we have observed in existing force-directed based graph drawing techniques are the following:

–`Scalability:` To the best of our knowledge, most implementations scale up to few thousand nodes.
–`Computational complexity:` A major pitfall of existing force directed layout techniques is their computational complexity, which is $\Theta(N^2 + E)$. Hence, performance of existing approaches is low for the case of large scale networks.
–`Aesthetics:` Many tools suffer from node cluttering and edge crossing problems in case of dense graphs, as well as vertice occlusion over edges, and asymmetric drawings as noted by [34].
–`Local Minima:` The adoption of cooling schedules and temperature mechanisms may reduce the problems related to local minima; however, they need to be fine-tuned and optimized to be effective on large graphs.
–`Topology layout:` If a network contains many edges and vertices, the structure of the visualization becomes complex due to the local minima problem.
–Convergence Nodes are moved back and forth without converging.

2.2 Influencers and Influence in Social Networks

Traditionally, the literature characterizes a social media user as an influencer on the basis of structural properties. Centrality metrics are the most widely considered parameters for the structural evaluation of a user's social network. The centrality of a concept has been defined as the significance of an individual within a network [18]. Centrality has attracted a considerable attention as it clearly recalls concepts like social power, influence, and reputation. A node that is directly connect-ed to a high number of other nodes is obviously central to the network and likely to play an important role [4]. [19] introduced the first centrality metrics, named as degree centrality, which is defined as the number of links incident upon a node. A node with many connections to other nodes, likely to play an important role [40]. A distinction is made between in-degree and out-degree centrality, measuring the number of incoming and outgoing connections respectively. This distinction has also been considered important in social networks. For example, Twitter makes a distinction between friends and followers. Normally, on Twitter, users with a high in-degree centrality (i.e. with a high number of followers) are considered influencers.

In addition to degree centrality, the literature also shows other structural metrics for the identification of influencers in social net-works. [31] presented an approach, where users were identified as influencers based on their total number of retweets. Results highlighted how the number of retweets are positively correlated with the level of users' activity (number of tweets) and their in-degree centrality (number of followers). Besides structural metrics, the more recent literature has associated the complexity of the concept of influence with the variety of content. Several research works have addressed the need for considering content-based metrics of influence [7]. Content metrics such as the number of mentions, URLs, or hashtags have been proved to increase the probability of retweeting [3].

The more recent literature has associated the complexity of the concept of influence with the diversity of content. Several research works have addressed the need for considering content-based metrics of influence [32,36,42]. Clearly, this view involves a significant change in perspective, as assessing influence does not provide a static and general ranking of influencers as a result. However, there is a need for effective visualization technique in social networks, which enable user to visually explore large-scale complex social networks to identify influencers in social networks. The layout should be aesthetically pleasant and provide multi-layered periphery of the nodes in clustered networks to exploit spread of influence in social networks.

While the literature provides consolidated approaches supporting the identification and characterization of hub nodes i.e. influencers in a social network, research on information spread, which is multi-layered distribution of peripheral nodes, is limited. The literature mainly focuses on the concept of influencers, while there is a need for effective visualization techniques in social networks, which enable users to visually explore large-scale complex social networks to identify the users who are responsible for influence. This paper presents a power-law based modified force-directed technique, that extends a previous algorithm discussed in [24].

3 The Power-Law Algorithm

This section provides a high-level description of the graph layout algorithm used in this paper. An early version of the algorithm has been presented by [24]. This paper improves the initial algorithm by identifying multiple layers of peripheral nodes around hub nodes. The power-law layout algorithm belongs to the class of force-directed algorithms, such as the one by [12,20].

The base mechanism is that of starting from an initial placement of graph nodes, and then iteratively refining the position of the nodes according to a force model. The iteration mechanism is controlled by means of

Cooldown step. The main innovation in our approach consists in the synergy between the exploitation of the power-law distribution of the data and the adaptive temperature cooldown mechanism. The underlying idea is that of iterating on hub nodes first with small cooldown steps, and subsequently on peripheral nodes with large cooldown steps, in order to achieve faster convergence. The advantages of this approach are:

–The initial iteration on hub nodes is more efficient than iterating on the whole node set, since $|N_h| \ll |N|$. As a consequence, it is possible to perform a fine-grained positioning of hub nodes (achieved by adopting small cooldown steps), (peripheral nodes will then form clusters around hubs).
–The iteration over the set of peripheral nodes, which would be computationally expensive since $N_p \cong N$, is limited by the adoption of large cooldown steps.

Algorithm 1 provides a high-level overview of the whole algorithm by showing its main building blocks.

Algorithm 1: Abstract Level Power-Law Layout Algorithm.

```
Input:
N_h = Hub Nodes;
N_p = Peripheral Nodes;
E_h = Edges;
d = node's Degree;
T = Energy / Temperature Variable;
T_h = Temperature threshold;
begin
    call NodePartition()
    call InitialLayout()
    while Temperature > 0 do
        if Temperature > T_h then
            call AttractionForce(N_h, N_p)
            call RepulsionForce(N_h, E)
        else
            call AttractionForce(N_p, N_h)
            call RepulsionForce(N_p, E)
        end
        call Cooldown(T)
        call resetNodesSizes(N_p, N_t, d)
    end
end
```

3.1 NodePartition

The NodePartition method is aimed at the exploitation of the power-law degree distribution of data. Provided that the degree-distribution of the nodes follows a power law, we partition the set of nodes N into the set of hub nodes N_h and the set of peripheral nodes N_p, such that $N = N_h \cup N_p$, with $N_h \cap N_p = \varnothing$. As a consequence, the set of edges E is also partitioned in the set of edges E_h for which at least one of the two nodes is a hub node, and the set E_p which contains all the edges connecting only peripheral nodes, with $E = E_h \cup E_p$, and $E_h \cap E_p = \varnothing$. The distinction of a node n as a hub node or as a peripheral node is based on the evaluation of its degree $\rho(n)$ against the constant ρ_h, which is a threshold defined as the value of degree that identifies the top i^{th} percentile of nodes, sorted by decreasing value of degree. Since the power-law is supposed to hold in the degree distribution, assuming for example $i = 20$ will end up in defining ρ_h as the 20^{th} percentile, thus considering as hub nodes the 20% of the nodes with the highest values of degree - the Pareto's 80-20 Rule, as suggested by [29].

3.2 InitialLayout

The InitialLayout() method responsible for random placement of graph nodes. However, as discussed by [13] and [27], it is known from the literature that the initial layout of graph nodes is an important factor to be considered in order to avoid the local minima problem, especially as the number of graph nodes increases, as noted by [27,15]. As suggested by [14], a combined approach can be helpful in solving this problem. In this paper, we adopt a random initial placement of nodes; however, a combination with other algorithms such as [26] or [20] will be considered as part of our future work.

–The initial iteration on hub nodes is more efficient than iterating on the whole node set, since $|N_h| \ll |N|$. As a consequence, it is possible to perform a fine-grained positioning of hub nodes (achieved by adopting small cooldown steps), (peripheral nodes will then form clusters around hubs).
–The iteration over the set of peripheral nodes, which would be computationally expensive since $N_p \cong N$, is limited by the adoption of large cooldown steps.

3.3 Forces

In this paper, both forces formulae (Attraction and Repulsion) have been taken from the power-law based modified force-directed algorithm as presented in [24].

3.4 CoolDown

The CoolDown(T) method is responsible of cooling down the system temperature, in order to make the algorithm converge. We introduce a customized dynamic temperature cooldown scheme, which adapts the cooldown step based on the current value of the temperature. As shown in Figure 1, the temperature is supposed to be initialized at a value T_{start}, and then to be

reduced by a variable `cooldown` step Δt based on the current value of the temperature itself. This approach provides a convenient way to adapt the speed of iteration of the algorithm to the number of nodes to be processed. While processing hub nodes (a few), the temperature decreases slowly; while processing peripheral nodes (many), the temperature decreases more rapidly to avoid expensive computations for nodes that are not *central* to the overall graph layout. The reference temperature value T_c is used as convergence threshold, i.e., when the temperature reaches that point the iteration is stopped.

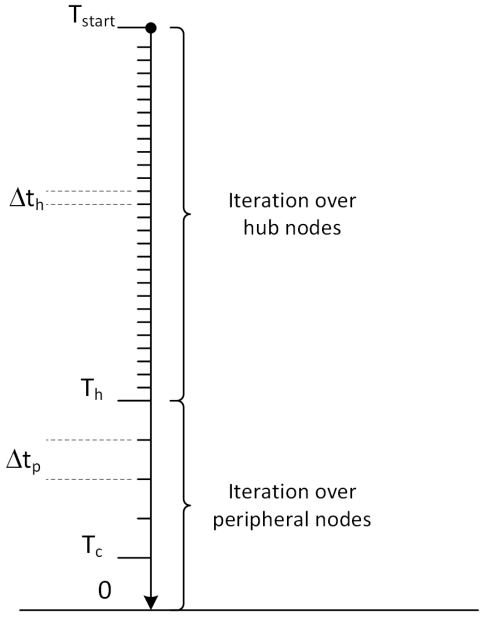


Fig. 1: Adaptive temperature cooldown mechanism.

Algorithm 2 presents the general overview of the temperature `cooldown` scheme. Variables Δt_h and Δt_p may be parameterized to adapt the algorithm behavior to properly fit the requirements given by the context of analysis. The values we used for the experimental analyses are $\Delta t_h = 0.0005$ and $\Delta t_p = 0.05$.

Algorithm 2: Temperature Cooldown

```
1  begin
2      if Temperature > T_h then
3          Temperature = Temperature − Δt_h;
4      else
5          Temperature = Temperature − Δt_p;
6      end
7      if Temperature ≤ T_c then
8          Temperature = 0;
9      end
10 end
```

3.5 resetNodesSizes

This method is responsible for resetting the sizes of each node in the graph, based upon their degree. The higher the degree of a node, the greater the size and vice versa.

3.6 Computational complexity

We evaluate the overall computational complexity of the graph layout algorithm by starting from the assessment of the computational complexity of its components.

–`Node characterization.` The computational complexity of the node characterization step is O($|N|$), since it requires complete iterations over all the nodes of node set N.
–`Initial layout.` The computational complexity of the initial node placement depends on the complexity of the selected layout algorithm. Assuming that a random initial placement of nodes is used, the complexity is O($|N|$).
–`Attractive force.` The computational complexity of the attractive force is O($|E_h|$) for each iteration on hub nodes, and O($|E_p|$) for each iteration on peripheral nodes, with $|E_h| > |E_p|$ and $|E| = |E_h| + |E_p|$. Overall, the computational complexity of the attractive force step is then O($|E_h|$).
–`Repulsive force.` The computational complexity of the repulsive force is O($|N_h|^2$) for each iteration on hub nodes, and O($|N_p|^2$) for each iteration on peripheral nodes, with $|N_p| > |N_h|$ and $|N| = |N_h| + |N_p|$. Overall, the computational complexity of the repulsive force step is then O($|N_p|^2$).
–`Cooldown.` The computational complexity of the temperature cooldown step is O(1).
–`resetNodesSizes.` The computational complexity of the this step is also O(1).

Considering the computational complexity evaluation of each step of our algorithm, the overall computational complexity is O($|E_h|$) + O($|N_p|^2$).

4 Experimental Methodology and Results

4.1 Data Sample

We collected a sample of tweets over a two-month period (December 2012 - January 2013). For the collection of tweets, we queried the public Twitter APIs by means of an automated collection tool developed ad-hoc. We queried Twitter APIs with the following crawling keywords, representing tourism destinations (i.e. brands): *Amalfi*, *Amalfi Coast*, *Lecce*, *Lucca*, *Naples*, *Palermo* and

Rome. Two languages have been considered, English and Italian. Collected tweets have been first analysed with a proprietary semantic engine in order to tag each tweet with information about `a)` the location to which it refers, `b)` the location's brand driver (or category) on which authors express an opinion, `c)` the subject referred to by the author, `d)` the number of retweets (if any), and `e)` the identifier of the retweeting author. Our data sample is refers to the tourism domain. We have adopted a modified version of the Anholt Nation Brand index model to define a set of categories of content referring to specific brand drivers of a destination's brand [2]. Examples of brand drivers are *Art & Culture*, *Food & Drinks*, *Events & Sport*, *Services & Transports*, etc. A tweet is considered Generic if it does not refer to any Specific brand driver, while it is considered `Specific` if it refers to at least one of Anholt's brand drivers.

Tweets have been categorized by using an automatic semantic text processing engine that has been developed as part of this research. The semantic engine can analyse a tweet and assign it to one or more semantic categories. The engine has been instructed to categorize according to the brand drivers of Anholt's model, by associating each brand driver with a specific content category described by means of a network of keywords. Each tweet can be assigned to multiple categories. We denote with N_C the number of categories each tweet `w` is assigned to; the specificity $S(w)$ of a given tweet `w` is defined in Equation 1 as follows:

$$S(w) = \left\{ \begin{array}{l} 0, N_c = 0 \\ 1, N_c > 0 \end{array} \right\} \quad (1)$$

Table 1 refer to the descriptive statistics of the original non-linear variables.

Table 1: Basic descriptive statistics of our data set.

Variable	Value
Number of tweets	957,632
Number of retweeted tweets	79,691
Number of tweeting authors	52,175
Number of retweets	235,790

4.2 Network models

In order to verify the effectiveness of the proposed algorithm with respect to the goal of our research, we have defined different network models based on the data set described in the previous section. Figure 2 provides an overview of the adopted network models.

– `Author → Brand (`N_1`)` This model considers the relationship among authors and domain brands, i.e., touristic destinations in our data set. The network is modeled as an undirected affiliation two-mode network, where an author node n_a is connected to a brand node n_b whenever author a has mentioned brand b in at least one of his/her tweets. The weight of the edge connecting n_a to n_b is proportional to the number of times that author a has named brand b in his/her tweets.

– `Author → Category (`N_2`)` This model considers the relationship among authors and domain brand drivers (categories), i.e., city brand drivers in our data set (namely, *Arts & Culture*, *Events & Sports*, *Fares & Tickets*, *Fashion & Shopping*, *Food & Drink*, *Life & Entertainment*, *Night & Music*, *Services & Transport*, and *Weather & Environmental*). The network is modelled as an undirected affiliation two-mode network, where an author node n_a is connected to a category node n_c whenever author a has mentioned a subject belonging to category c in at least one of his/her tweets. The weight of the edge connecting n_a to n_c is proportional to the number of times that author a has named category c in his/her tweets.

– `Author → Subject (`N_3`)` This model considers the relationship among authors and domain subjects, i.e., relevant semantic lemmas in our data set. The network is modeled as an undirected affiliation two-mode network, where an author node n_a is connected to a subject node n_s whenever author a has mentioned subject s in at least one of his/her tweets. The weight of the edge connecting n_a to n_s is proportional to the number of times that author a has named subject s in his/her tweets.

– `Author → Author (`N_4`)` This model considers the relationship among authors producing a tweet and corresponding retweeting authors. The network is modeled as a directed one-mode network, where an author node n_{a1} is linked to another author node n_{a2} whenever author $a1$ has retweeted at least one tweet of author $a2$. The weight of the edge connecting n_{a1} to n_{a2} is proportional to the number of times that author $a1$ has retweeted author $a2$.

4.3 Visualization Results and Discussions

In order to visually analyse the influencers (hub nodes) and influence (spread across the multi-layered peripheral nodes connected around hub nodes), we visualized afore-mentioned networks in Section 4.2. The color scheme for node-pair for all networks, is consistent for each graph (Yellow nodes: N_A; Blue: N_B). Figures 4, 6, 7 and 9 present visualizations of the each network ($N_1 - N_4$) from dataset, along with visual benchmark comparison with existing approaches. Table 2 compares the average time performance of our algorithm against that of the [20] and [33] approaches. Our approach shows a significant improvement in layout computation time.

Table 2: Summary of experimental results.

	Dataset Size				Computational Time and Speedup				
	$\|N\|$	$\|N_A\|$	$\|N_B\|$	$\|E\|$	PL (s)	FR (s)	FR (%)	MR (s)	MR (%)
N_1	78	71	7	94	0.012	0.191	93.72	0.039	69.23
	275	268	7	540	1.223	3.422	64.26	2.25	45.64
	2,627	2,621	7	3,705	4.962	196.387	97.47	96.837	94.88
	12,017	12,011	7	14,139	6.472	232.486	97.22	124.023	94.78
	21,000	20,993	7	24,349	9.635	328.745	97.07	256.382	96.24
	30,523	30,516	7	34,845	12.256	547.334	97.76	327.287	96.26
N_2	58	32	26	181	0.0017	1.157	98.53	0.234	92.74
	87	56	31	327	0.114	1.638	93.04	0.545	79.08
	301	268	38	1,897	1.236	3.427	63.93	2.678	53.85
	2,659	2,615	44	11,602	3.248	187.218	98.27	146.213	97.78
	12,049	12,005	44	35,192	7.623	412.349	98.15	318.641	97.61
N_3	163	56	107	272	0.026	1.759	98.52	0.452	94.25
	583	263	320	1,849	1.367	4.768	71.33	2.984	54.19
	3,694	2,614	1,079	10,489	2.923	178.382	98.36	136.231	97.85
N_4	1,305	373	932	1,000	0.941	3.876	75.72	2.316	59.37
	2,677	839	1,838	2,000	1.769	5.672	68.81	3.261	45.75
	6,268	1,839	4,429	5,000	2.746	128.762	97.87	87.562	96.86
	11,484	2,197	9,287	10,000	4.627	238.752	98.06	124.753	96.29

Key to symbols: N: total number of nodes in network; N_A: number of author nodes; N_B: number of brand / subject / category / retweeting author nodes; E: number of edges
Key to algorithm acronyms: *PL*: Power-law; *FR*: Fruchterman-Reingold; *MS*: Modified Spring.

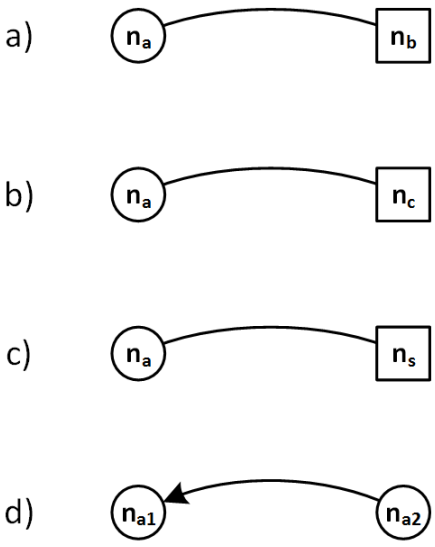
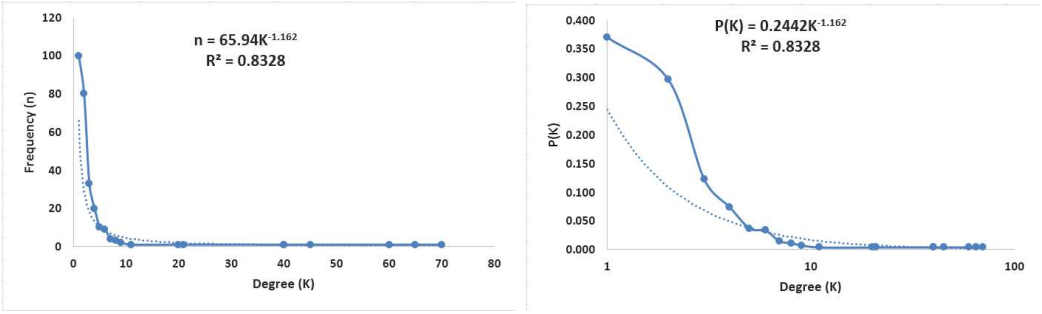


Fig. 2: Network models: a) N_1: Author → Brand; b) N_2: Author → Category; c) N_3: Author → Subject; d) N_4: Author → Author.

The dataset follow a power-law distribution, as discussed by [37]. Figure 3 explains that the graphs in our test set are 'scale-free' as they exhibits power-law degree distribution.

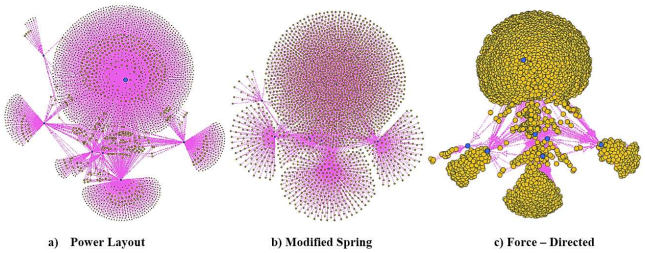
4.3.1 Results – N_1 Network (Brand Fidelity)

Networks N_1 is related to the relationship between authors and brands, i.e., touristic destinations which are basically Italian cities. In this case, the clustering of nodes provides a grouping of those authors who have tweeted about the same destination. The layering of nodes around brands is instead related to the intensity of tweeting about a given destination; i.e., authors closer to a brand node tweet a higher number of times about that destination with respect to farther authors. The emerging semantic of the network visualization is in this case related to the *Brand Fidelity* of authors. The visualized network layout supports the visual analysis of those authors who have a higher fidelity to a given brand, or those authors who never tweet about that brand. Moreover, it is possible to point out which authors are tweeting about a brand as well as a competing brand to support the definition of specific marketing campaigns. Through our visualization approach, we able to visually identify multiple peripheral layers of nodes surrounded by influencing hub nodes, the spread of these multi-layered peripheral nodes around hub nodes express the influence. Figures 4 provides the visualization of networks N_1 of our dataset, together with a visual comparison with the layouts generated by two

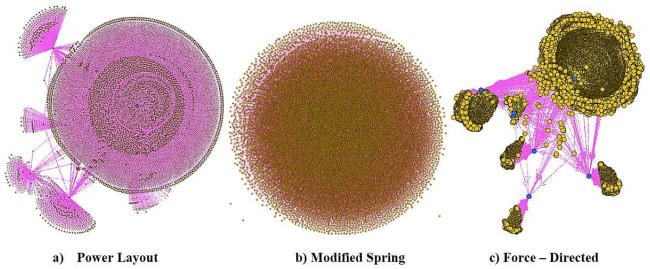


(a) Linear Distribution (b) Logrithmic Distribution

Fig. 3: Power-Law degree Distribution from data set N_2.



(a) Network N_1: G(N=2627; E=3705).



(b) Network N_1: G(N=12017;E=14139).

Fig. 4: Graphs upon network N_1 `Author` $\rightarrow$ `Brand`.

reference algorithms. Through our visualization approach, we able to visually identify multiple peripheral layers of nodes surrounded by influencing hub nodes, the spread of these multi-layered peripheral nodes around hub nodes express the influence.

4.3.2 Results – N_2 Network (Category Specificity)

Figure 5 provides an enlarged view of network N_2 visualized by means of the power-law layout algorithm. The network visualization depicted in Figure 5 adopts yellow (light) nodes to represent authors, and blue (dark) nodes to represent the categories on which authors have expressed opinions in their tweets. The layout of the network produced by the power-law layout algorithm clearly highlights that author nodes aggregate in several groups and subgroups based on their connections with category nodes, which in this case are the hub nodes. The aggregation of author nodes can be analyzed from two different perspectives:

1. `Clusters.` The groups of author nodes cluster together all those authors that are connected to the same hubs (i.e., categories); this provides a visual clustering for those authors who have tweeted about the same categories. For example, Figure 5 highlights clusters that group all the authors who tweeted about *Events & Sports*, *Fashion & Shopping*, *Drink*, and *Entertainment* categories, as well as the authors who tweeted about more than one category, such as *Transport* and *College*, or *Entertainment* and *Photo*.
2. `Layers.` The network layout shows that clusters are placed at a different distance from the visualization center based on the number of hubs to which they are connected. In other words, the most peripheral clusters are those in which nodes are connected to only one hub, while the central cluster is the one in which nodes are connected to the highest number of hub nodes. An example of node layering is provided in the upper left area of Figure 5: the cluster referring to those authors who have tweeted about category *Entertainment* is positioned above (i.e., on an outermost layer) and the clusters grouping the authors who have tweeted about *Entertainment* and *Photo*, or *Entertainment* and *People* are positioned below.

Authors belonging to the central cluster of nodes are in fact those who are more `generalist` in their content sharing about the analyzed tourism destinations, since they refer to many different categories. On the contrary, authors belonging tho the most peripheral clusters are those who are very `specific` in sharing content related to selected categories. Figure 6 represents the benchmark comparison of our technique with existing techniques, and the results are evident that our approach produces aesthetically pleasant layouts by highlighting clusters of multiple peripheral layers surrounded by hub-nodes.

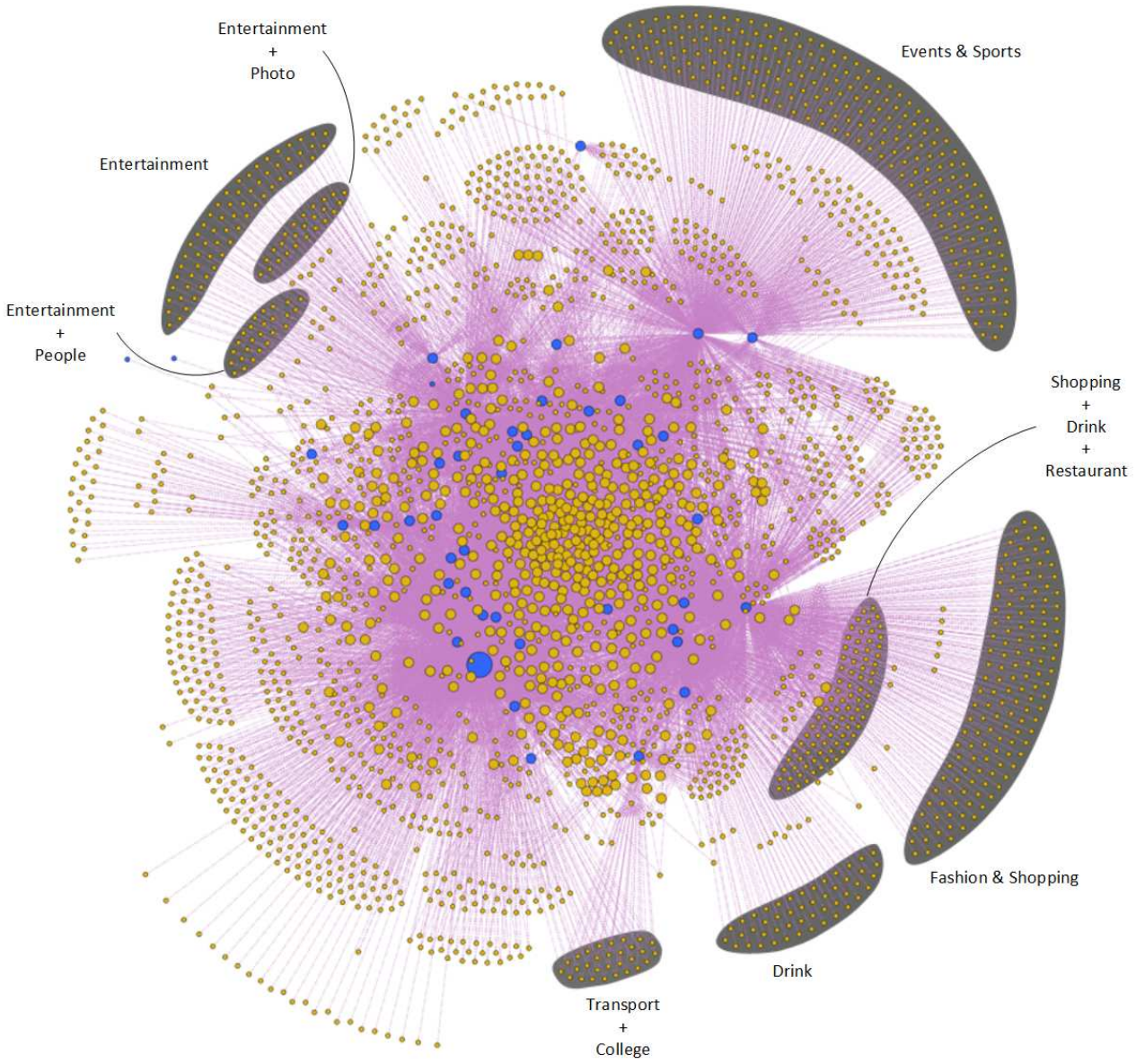


Fig. 5: Network N_2: `Author` $\rightarrow$ `Category` (enlarged view).

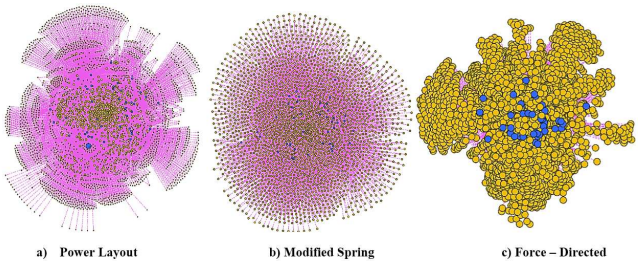


Fig. 6: Network N_2: `Author` $\rightarrow$ `Category` G(N=2,659;E=11,602).

4.3.3 Results – N_3 Network (Subject Specificity)

Network N_3 is related to the relationship between authors and subjects. Figures 7 provides the visualization of networks N_3 of our dataset, together with a visual comparison with the layouts generated by two reference algorithms. The emerging semantic of the network visualization is similar to that of N_2, since the layout provides a visual representation of the level of `specificity` (or generality) of authors with respect to subjects instead of categories. In this network, we found many subjects, upon which multiple authors expressed their opinions, hence the center of graph, seems dense.

Our approach able to produce multiple layers of peripheral nodes surrounded by hub-nodes. In graph, we can observe multiple outlier peripheral layers, which are surrounded by distinct subjects, are drawn far from center of graph. We also observe some influencing authors' nodes of large size, as they seemed to express their opinions many times upon multiple subjects, hence showing strong influence.

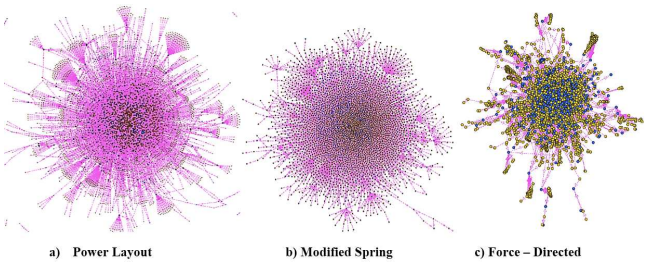


Fig. 7: Network N_3: `Author` → `Subjects` G(N=3,694;E=10,489).

4.3.4 Results – N_4 Network (Retweeting phenomena)

Network N_4 is related to the relationship among authors retweeting other authors. Although very simple, this network model visually represents the complexity of real-world retweeting phenomena. As depicted in Figure 8, different retweeting scenarios are associated with different network topologies.

1. *Cloud Retweeting:* In case `a)` of Figure 8, an author is retweeted by many of his followers, is visually represented as a cloud of nodes aggregating around a single hub.
2. *Chain Retweeting:* The opposite situation, depicted in case `c)` of Figure 8, is that of a tweet that is retweeted by an author which is following the author who has last retweeted.
3. *Mixed Topology:* In the middle, as represented by case `b)` of Figure 8, a combination of the two base scenarios may happen, leading to intermediate topologies of varying complexity.

For Network N_4 visualizations which are provided] in Figure 8, Figure 9 and 10, a specific node coloring scheme is adopted in order to distinguish among different types of authors. *Yellow* nodes represent those authors who only retweets other authors, and *Blue* nodes represent those authors who only retweeted by other authors. Similarly, *Green* nodes represent authors who both retweet and retweeted by other authors.

Figure 9 represents the benchmark comparison of our technique with existing techniques. By considering only

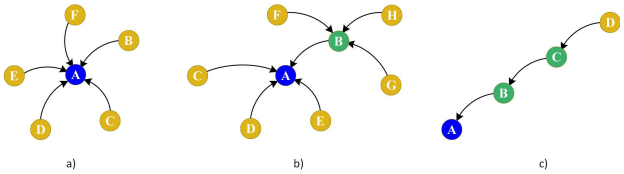


Fig. 8: Examples of author-author retweeting scenarios: `a)` cloud retweeting; `b)` mixed topology; `c)` chain retweeting.

hub nodes, in fact, it is clear that there is no clue to understand how content spreads across the authors network, since the majority of hubs are just the centers of isolated clouds of authors. Interesting insights can be provided to the reader only by taking into account the peripheral nodes (i.e., those nodes that are not labeled as hubs), and thus by reconstructing the phenomenon of *chain retweeting*. The network layout generated by the proposed power-law layout algorithm is clearly effective in helping the reader in identifying the different retweeting scenarios and interpreting how retweets spread across the network of authors.

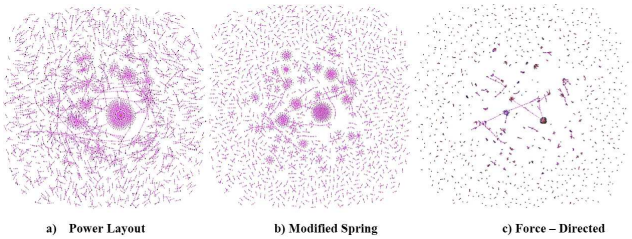


Fig. 9: Network N_4: `Author` → `Author` G(N=2,677;E=2000)

The interesting retweeting scenarios are the *chain retweeting* ones as shown in Figure 10. By considering only hub nodes, in fact, it is clear that there is no clue to understand how content spreads across the authors network, since the majority of hubs are just the centers of isolated clouds of authors. Interesting insights can be provided to the reader only by taking into account the peripheral nodes (i.e., those nodes that are not labeled as hubs), and thus by reconstructing the phenomenon of *chain retweeting*.

5 Conclusions and Future Work

This paper proposes a novel visual aspect for the analysis and exploration of social networks in order to identify and visually highlight influencers (i.e., hub nodes), and

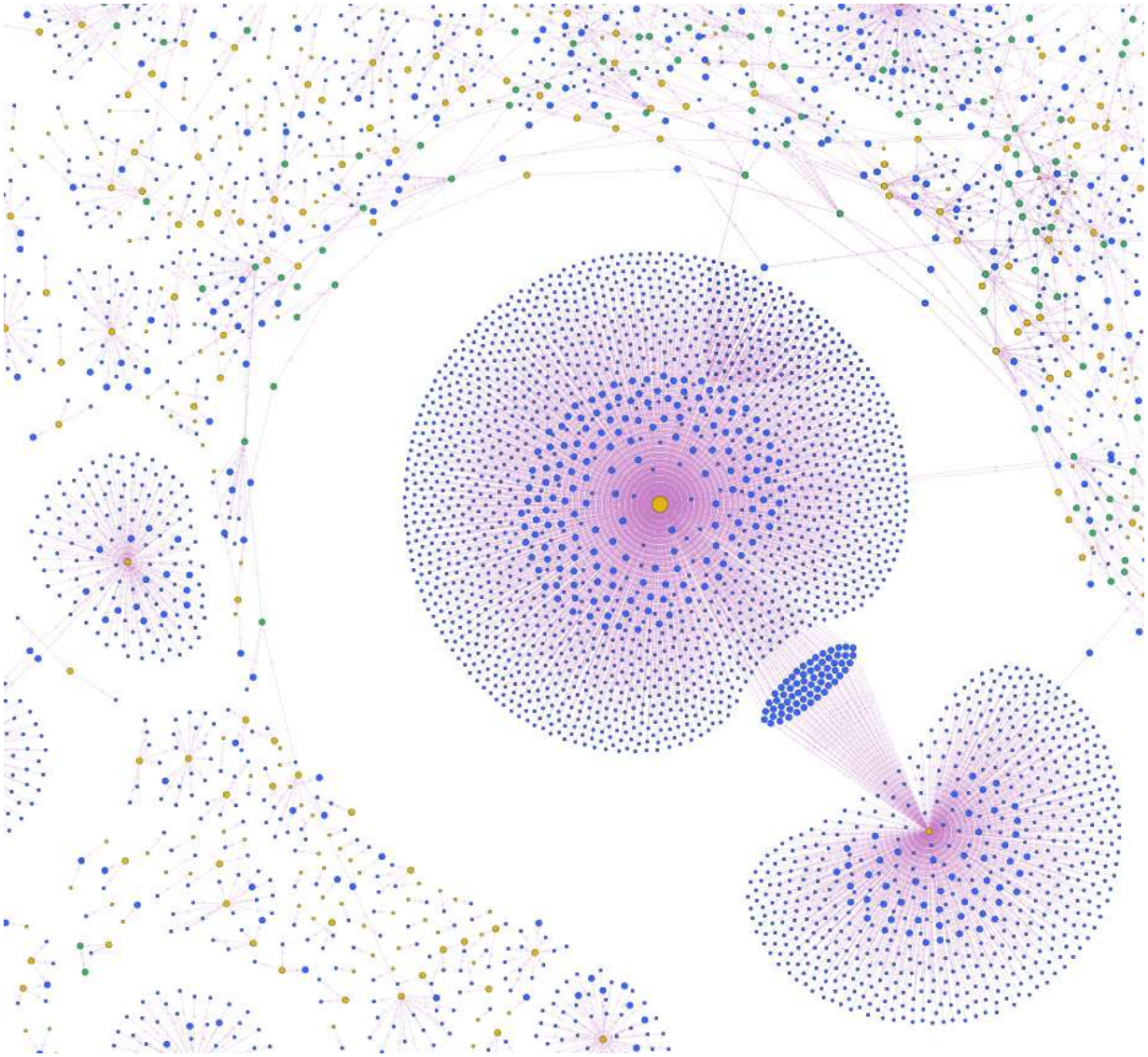


Fig. 10: Network N_5: Author $\rightarrow$ Author (enlarged view of detail).

influence (i.e., spread of multi-layer peripheral nodes), represented by the opinions expressed by social media users on a given set of topics. Results show that our approach produces aesthetically pleasant graph layouts, by highlighting multi-layered clusters of nodes surrounding hub nodes (the main topics). These multi-layered peripheral node clusters represent a visual aid to understand influence.

Our approach exploits the underlying concept of power-law degree distribution, which effectively represent multi-layered peripheral clusters around hub nodes. We analysed four different networks to exploit brand fidelity, category specificity, subject specificty and retweeting phenomenon. Our proposed approach is able to handle scalable graphs in multi-clustered, and multi-layered peripheries of network and encourages us to further explore social network's intrinsic characteristics. Results show that our approach significantly improves scalability, time performance and visual effectiveness compared to previous approaches. Although our experiment can be repeated with data from entities different from tourism domain, additional empirical work is needed to extend testing to multiple datasets and domains.

Future work will consider influence-based exploration of social networks based on influential parameters. An empirical evaluation of generally accepted graph drawing aesthetics criteria can be considered, to compare our approach with existing network drawing techniques. In our current work, we are studying an achievable measure of influence through proposed visualization approach that

can be used to rank influential nodes in social networks. Future research may address the development of an ad-hoc tool, by using proposed technique, for influence-based exploration of social networks.

References

[1] R. Andersen, F. Chung, and L. Lu. Drawing power law graphs using a local global decomposition. *Algorithmica*, 47(4):397, 2007.

[2] Simon Anholt. *Competitive identity: The new brand management for nations, cities and regions*. Palgrave Macmillan, 2006.

[3] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[4] D. Barbagallo, L. Bruni, C. Francalanci, and P. Giacomazzi. An empirical study on the relationship between twitter sentiment and influence in the tourism domain. In *Information and Communication Technologies in Tourism 2012*, pages 506–516. Springer, 2012.

[5] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. In *ICWSM*, 2009.

[6] F. Benevenuto, M. Cha, K.P. Gummadi, and H. Haddadi. Measuring user influence in twitter: The million follower fallacy. In *International AAAI Conference on Weblogs and Social (ICWSM10)*, pages pp. 10–17, 2010.

[7] Carolina Bigonha, Thiago NC Cardoso, Mirella M Moro, Marcos A Gonçalves, and Virgílio AF Almeida. Sentiment-based influence detection on twitter. *Journal of the Brazilian Computer Society*, 18(3):169–183, 2012.

[8] D. Boyd, S. Golde, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *IEEE*, pages pp. 1–10, 2010.

[9] L. Braun, M. Volke, J. Schlamp, A. Von Bodisco, and G. Carle. Flow-inspector: a framework for visualizing network flow data using current web technologies. *Computing*, pages 1–12, 2012.

[10] L. Bruni, C. Francalanci, P. Giacomazzi, F. Merlo, and A. Poli. The relationship among volumes, specificity, and influence of social media information. In *Proceedings of International Conference on Information Systems*, 2013.

[11] Amparo E Cano, Suvodeep Mazumdar, and Fabio Ciravegna. Social influence analysis in microblogging platforms–a topic-sensitive based approach. *Semantic Web*, 2011.

[12] D.S.M. Chan, K.S. Chua, C. Leckie, and A. Parhar. Visualisation of power-law network topologies. In *Networks, 2003. ICON2003. The 11th IEEE International Conference on*, pages 69–74. IEEE, 2004.

[13] P. Eades. A heuristic for graph drawing. *Congress Numerantium*, 42:149–160, 1984.

[14] P. Eades. Navigating clustered graphs using force-directed methods. *Journal of Graph Algorithms and Applications*, 4(3):157–181, 2000.

[15] P. Eades and X. Lin. Spring algorithms and symmetry. *Computing and Combinatorics*, pages 202–211, 1997.

[16] J. Ellson, E. Gansner, L. Koutsofios, S. North, and G. Woodhull. Graphviz Ů open source graph drawing tools. In *Graph Drawing*, pages 594–597. Springer, 2002.

[17] S. Falconer. Ontograf, 2010.

[18] Weiguo Fan and Michael D Gordon. The power of social media analytics. *Communications of the ACM*, 57(6):74–81, 2014.

[19] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

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

[21] S. Goyal and R. Westenthaler. Rdf gravity (rdf graph visualization tool). *Salzburg Research, Austria*, 2004.

[22] J. Heer, S.K. Card, and J.A. Landay. Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM, 2005.

[23] M.L. Huang, P. Eades, and J. Wang. On-line animated visualization of huge graphs using a modified spring algorithm. *Journal of Visual Languages & Computing*, 9(6):623–645, 1998.

[24] A. Hussain, K. Latif, A. Rextin, A. Hayat, and M. Alam. *Scalable Visualization of Semantic Nets using Power-Law Graphs*. *Applied Mathematics & Information Sciences*, 8(1):355–367, 2014.

[25] D Kalamaras. Social network visualizer (socnetv), 2010.

[26] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989.

[27] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou. Ontology visualization methodsŮa survey. *ACM Computing Surveys (CSUR)*, 39(4):10, 2007.

[28] A. Katifori, E. Torou, C. Halatsis, G. Lepouras, and C. Vassilakis. A comparative study of four ontology visualization techniques in protégé: Experiment setup and preliminary results. In *Information Visualization, IV 2006.*, pages 417–423. IEEE, 2006.

[29] R. Koch. *The 80/20 principle: the secret to achieving more with less*. Crown Business, 1999.

[30] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[31] Alex Leavitt, Evan Burchard, David Fisher, and Sam Gilbert. The influentials: New approaches for analyzing influence on twitter. *Web Ecology Project*, 4(2):1–18, 2009.

[32] Jingxuan Li, Wei Peng, Tao Li, Tong Sun, Qianmu Li, and Jian Xu. Social network user influence sense-making and dynamics prediction. *Expert Systems with Applications*, 41(11):5115–5124, 2014.

[33] C.C. Lin and H.C. Yen. A new force-directed graph drawing method based on edge-edge repulsion. *IEEE Computer Society*, 2005.

[34] Chun-Cheng Lin and Hsu-Chun Yen. A new force-directed graph drawing method based on edge-edge repulsion. *Journal of Visual Languages & Computing*, 23(1):29–42, 2012.

[35] B. Mike. Large-scale rdf graph visualization tools, 2008. http://www.mkbergman.com/414.

[36] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010.
[37] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.
[38] Richard Perline. Strong, weak and false inverse power laws. *Statistical Science*, pages 68–88, 2005.
[39] M. Sintek. Visualizing protege ontologies, 2003. http://protegewiki.stanford.edu/wiki/OntoViz.
[40] Raymond T Sparrowe, Robert C Liden, Sandy J Wayne, and Maria L Kraimer. Social networks and the performance of individuals and groups. *Academy of management journal*, 44(2):316–325, 2001.
[41] M. Storey, M. Musen, J. Silva, C. Best, R. Ernst, N.and Fergerson, and N. Noy. Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in protégé. In *Workshop on Interactive Tools for Knowledge Capture (K-CAP-2001)*. Citeseer, 2001.
[42] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, pages 177–184. IEEE, 2010.
[43] Y. Sure, J. Angele, and S. Staab. Ontoedit: Guiding ontology development by methodology and inferencing. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1205–1222, 2010.
[44] Y. Wang and Z. Wang. A fast successive over-relaxation algorithm for force-directed network graph drawing. *Science China Information Sciences*, 55(3):677–688, 2012.
[45] X. Xu, N. Yuruk, Z. Feng, and T.A.J. Schweiger. Scan: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 824–833. ACM, 2007.

Chiara Francalanci is professor of information systems at Politecnico di Milano. Her research focuses on information system engineering and, in particular, on feasibility analyses. Her main research interests in this area are information design, social networks analysis, sentiment analysis, architectural design of information systems, and cost-benefit analyses, as fundamental components of a feasibility study. She has over 15 years of experience in applied research and consulting. She has published on top international outlets on IT management and cost-oriented IT design, has led several national and international research projects, and has broad consulting experience.

Ajaz Hussain is a Doctoral Candidate of Information Technology at the Politecnico di Milano. In 2011, he earned his MS (IT) degree from National University of Science and Technology (NUST), Pakistan. His research interests include identification of Social Influencers, Empirical Analysis of Social Influence, Social Networks Analysis, Sentiment Analysis, Scale-Free Networks, Force-Directed Graphs Drawings and Information Systems. His current research work emphasis on Visual and Empirical Analysis of Social Influence and Influencers using scalable power-law based graph drawing approach. As an IT professional, he have around 5 years diverse academic and industry level experience in various sectors, including R&D Department, Project Management, Technology Services and Consulting, and Software Development.

Francesco Merlo is post-doc researcher at Politecnico di Milano, where he took a Ph.D. in Information Engineering and a master's degree in Computer Science Engineering. His main interests are the evaluation of software development and maintenance costs, with particular focus on the Open Source context, and the analysis of software quality metrics and their impact on the efficiency of the software development process. A collateral research interest is the definition of a graph query language.