Discovering social influencers with network visualization: evidence from the tourism domain

Chiara Francalanci[1] · **Ajaz Hussain**[1]

Received: 28 April 2015 / Revised: 3 July 2015 / Accepted: 30 July 2015
Published online: 13 August 2015

✉ Ajaz Hussain
ajaz.hussain@polimi.it

[1] Department of Electronics, Information and Bio-Engineering, Politecnico di Milano (POLIMI), Via Ponzio 34/5, 20133 Milan (MI), Italy

1 Introduction

The literature on social media makes a distinction between “influencers” and “influence”. 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. (Boyd et al. 2010; Myers and Leskovec 2014) 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 (Benevenuto et al. 2010; Messias et al. 2013).

In previous research (Barbagallo et al. 2012; Bruni et al. 2013; Klotz et al. 2014; Messias et al. 2013) has shown how the content of messages can play a critical role and can be 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 Chan et al. (2003) that social networks of influence follow a power-law distribution function (Baggio 2005), with a few hub nodes and a long tail of peripheral nodes, consistent with the so-called small-world phenomenon as noted by (Xu et al. 2007). In social media, hub nodes represent social influencers (Ren et al. 2014), 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 power-law based force-directed algorithm (Hussain et al. 2014) 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 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, which is metaphor of k-shell decomposition analysis (Carmi et al. 2007; Kitsak et al. 2010).

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 i.e. tourism destinations, called *brand* (e.g. *Rome*, *Naples*, etc.), (b) the destination brand driver (or category) on which authors express an opinion (e.g. *Art and Culture*, *Food and Drinks, Events and Sport* etc.), (c) the number of retweets, and (d) 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, and category) by aesthetically pleasant layouts. By visually exploring and understanding multi-layered periphery of nodes in clusters, we also propose few content related hypotheses in order to understand network behavior and relationship among frequency, specificity, and retweets in tweets. Insights on the relationship among frequency, specificity, and influence would help social media users make their behavioral decisions on how to build a reputation across multiple social media. Social media users can make decision about when, how and where to promote their content over social media. For example, how much they should post on social media, or how frequently they should address topics *specific* to multiple communities in order to increase their *reach*. On social media, promoting the same content multiple times, has been found to increase the attention of the audience (Benevenuto et al. 2010; Cha et al. 2010). Results highlight the effectiveness of our approach, providing interesting visual insights on how unveiling the structure of the periphery of the network can visually show the potential of peripheral nodes in determining influence and content relationship.

This paper also takes a behavioral perspective by proposing few content based hypotheses and by investigating characteristics of shared content (e.g. frequency, specificity, influence etc.) that are an outcome of behavioral decisions made by social media users. In particular, we focus on three behavioral variables: content specificity, frequency of sharing, frequency of retweets. The first variable represents the level of detail with which a user comments on a given subject of interest, while the second one represents the amount of contents shared by user in tweets and third one is frequency of retweets upon shared content. Insights on the relationship among content specificity, frequency of sharing, and frequency of retweets would help social media users to make their behavioral decisions. Fundamental goal of any social media user is to post content that is shared frequently, by many other users and over extended periods of time before fading (Asur et al. 2011; Fan and Gordon 2014). However, the literature does not provide systematic and visual evidence on how behavioral decisions regarding content specificity, frequency of sharing and frequency of retweets exert an impact on influence. This paper provides preliminary evidence from Twitter. We put forward three hypotheses that tie specificity, frequency and frequency of retweets and are tested on data samples of roughly one million tweets.

Empirical and visual results show a significant relationship between influence and behavioral decisions on content. The relationship is found to be consistently significant across both data samples. This empirical and visual evidence raises theoretical challenges and encourages further research to understand the relationship between content and influence on social media. The main innovative aspect of our approach is that we use statistics (hypotheses) and visualization together. One can visually verify the proposed hypotheses on graphs.

2 State of the art

In this section, we will explore the concept of influencers and influence in social media. We will also discuss about limitations of existing network visualization techniques.

2.1 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 (Fan and Gordon 2014). Centrality has attracted a considerable attention as it clearly recalls concepts like social power, influence, and reputation. A node that is directly connected to a high number of other nodes is obviously central to the network and likely to play an important role (Barbagallo et al. 2012). Freeman (1979) 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 (Sparrowe et al. 2001). 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 networks. (Leavitt et al. 2009) 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 (Bigonha et al. 2012). Content metrics such as the number of mentions, URLs, or hashtags have been proved to increase the probability of retweeting (Bakshy et al. 2011).

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 (Hussain et al. 2014) by exploiting the k-shell decomposition technique (Kitsak et al. 2010). The algorithm is briefly summarized in Sect. 4.

2.2 Network visualization techniques

Several research efforts in network visualization have targeted power-law algorithms and their combination with the traditional force-directed techniques, as for example in (Andersen et al. 2007). Among these approaches, the most notable is the Out-Degree Layout (ODL) for the visualization of large-scale network topologies, presented by (Perline 2005). The core concept of the algorithm is the segmentation of 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 topology of the network plays an important role such that there are plausible circumstances under which nodes with a higher number of connections or greater betweenness have little effect on the range of a given spreading process. For example, if a hub exists at the end of a branch at the periphery of a network, it will have a minimal impact in the spreading process through the core of the network, whereas a less connected person who is strategically placed in the core of the network will have a significant effect that leads to dissemination through a large fraction of the population. To identify the core and the multi-layered periphery of the clustered network, we use a technique based on the metaphor of k-shell (also called k-core) decomposition of the network, as discussed in Kitsak et al. (2010).

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 Francalanci and Hussain (2014, 2015). This paper improves the initial algorithm by identifying multiple layers of peripheral nodes around hub nodes according to the k-shell decomposition approach. The power-law layout algorithm belongs to the class of force-directed algorithms, such as the one by Chan et al. (2003) and Hussain et al. (2014). In this algorithm, we adopt a pre-processing method aimed at distinguishing hub nodes from peripheral nodes. This step is performed by pre-identifying hub nodes as N_h, which represents one of the following two sets:

1. A set of predefined tourism destinations, called *brands*, i.e. *Amalfi*, *Amalfi Coast*, *Lecce*, *Lucca*, *Naples*, *Palermo* and *Rome* (7 in total).
2. A set of predefined brand drivers of a destination's brand, called *categories*. Examples of categories are *Art and Culture*, *Food and Drinks, Events and Sport, Services and Transports,* etc., as explained in Sect. 5.

The following code snippet provides a high-level overview of the whole algorithm by showing its main building blocks. The proposed approach is aimed at the exploitation of the power-law degree distribution of author nodes (N_p). Provided that the distribution of the degree of the nodes follows a power law, we partition the network into bipartite graph of two disjoint vertices N into the set of predefined hub nodes N_h, which represents topics (brands or categories), and the set of peripheral nodes N_p, which represents authors, such that $N = N_h \cup N_p$, with $N_h \cap N_p = \varnothing$.

```
DATA:
  $N_h$ = Hub Nodes representing Brands or Categories;
  $N_p$ = Peripheral Nodes representing authors;
  E = Edges connecting authors to either brands or categories
  whenever one of author's tweet refers to that brand or catego-
  ry;
  d = Degree of author node representing the number of edges
  connected to author node;
  $t_p$ = Number of times the author Np tweeted about $N_h$ (i.e. Brand
  or Category);
  T = Energy / Temperature Variable; $T_h$ = Temperature threshold,
  to control simulation.
  BEGIN
  1. NodePartition();
  2. InitialLayout();
        IF ($T>T_h$) DO
              AttractionForce($N_h$,$N_p$);
              RepulsionForce($N_h$,E);
        ELSE
              AttractionForce($N_p$,$N_h$);
              RepulsionForce($N_p$,E);
  3. LShellDecomposition($N_p$,$t_p$);
  4. NodesPlacement ($N_h$,$N_p$,$t_p$);
  5. TempCoolDown(T);
  6. resetNodesSizes($N_h$,$N_p$,d);
  END
```

Figure 1 provides a general workflow of the whole algorithm by showing its main building blocks. The *Initial Controls and Pre-Processing* step is responsible for rescaling the size of each node in the graph, based upon the degree. The higher the degree of a node, the greater the size and vice versa. This step is also responsible for partitioning the network into two pre-defined disjoint sets of vertices (i.e. hub nodes —topics, and peripheral nodes—authors). The *Modified Force-Directed Forces* step calculates attraction and repulsion forces, based upon the value of T_h, which is a threshold value that can be tuned to optimize the layout, by providing maximum forces exerted upon Hub nodes N_h (*Adaptive Temperature Control*).

We introduce a customized dynamic temperature cool down scheme, which adapts the iterative step based on the current value of temperature. The temperature is supposed to be initialized at a value T_{start}, and then to be reduced by a variable T_{start}, 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

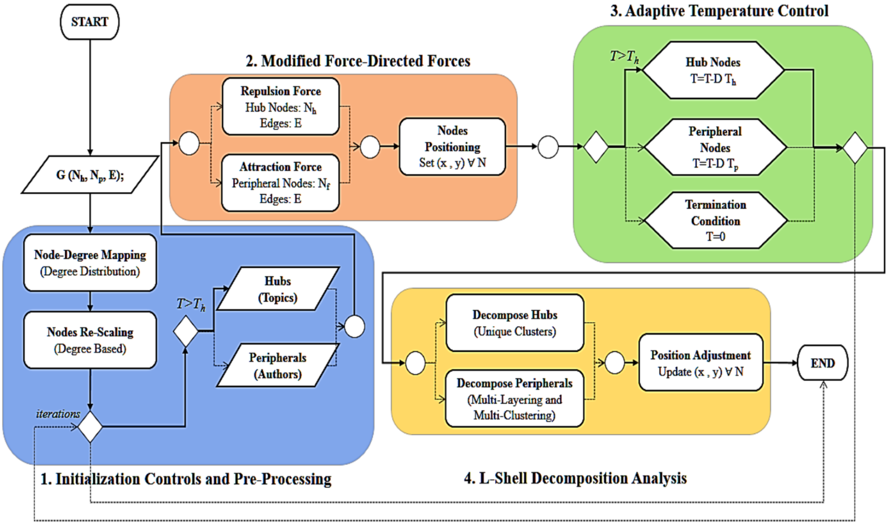

Fig. 1 Power law algorithm workflow

central to the overall graph layout. The formulae of attraction and repulsion forces are similar to those used in traditional force-directed approaches, such as Chan et al. (2003). In this paper, the forces formulae have been taken from the power-law based modified force-directed algorithm presented in Hussain et al. (2014).

The *L-Shell Decomposition Analysis* step is responsible for the calculation of the l-shell value of author nodes in N_p, in order to create a multi-layered hierarchy of author' nodes around the topics' nodes. This step also performs the final placement of nodes on graph canvas based on the computation of forces among nodes and l-shell mechanism. We tuned this technique by means of the metaphor of k-shell decomposition analysis (Carmi et al. 2007), in order to define the concept of *level* of each node in the multi-layered periphery of our graphs. This process assigns an integer as level index (l_S) to each node, representing its location according to successive layers (*l* shells) in the network. In this way, the author nodes who tweeted once about a specific topic, will have ($l_s = 1$) forming the outmost layer around that topic, and those who tweeted twice will have ($l_s = 2$) forming the inward successive layer, and so on. By this metaphor, small values of (l_S) define the periphery of the network (outliers), while the innermost network levels correspond to greater values of l_S, containing those authors who tweeted most frequently, as shown in Fig. 2.

4 Research hypotheses

The literature indicates that social media are associated with a long-tail effect, with a variety of smaller communities (Meraz 2009). While general content has a broad audience, there exists a variety of smaller communities who are interested in

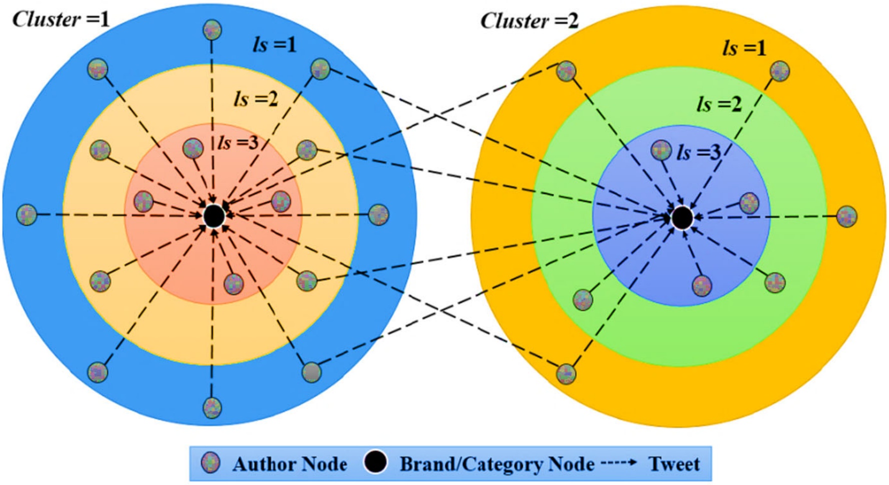

Fig. 2 Metaphor of k-shell decomposition analysis

specific content. Such long-tail effect suggests that these communities are numerous and their specific interests are virtually limitless (Fan and Gordon 2014). Social media users also consider *specificity* as an important metric for making behavioral decisions (Bruni et al. 2013). The *specificity* of shared content by social media users can be described as the level of detail with which a user comments on a given subject of interest. Klotz et al. (2014) has shown how the content of messages can play a critical role and can be a determinant of the social influence of a message irrespective of the centrality of the message's author. Twitter users with a high volume of tweets can be referred to as '*information sources*' or '*generators*' (Hutto et al. 2013). The literature also shows that social media user intend to post content that is shared frequently by many other users (Asur et al. 2011). Social media users wish to be influential (Chang 2014). Intuitively, since users want to be interesting to many, if a user talks a lot, he/she will probably address the needs of multiple specific communities, i.e. multiple topics. Consequently, our first hypothesis posits a positive association between *frequency of tweets* and *content specificity* in multiple topics.

> *H1: Authors tweeting with a high frequency of tweets is positively associated with multiple topics* (*brands or categories*) (*i.e. visually, potential influencers are peripheral authors*).

If a speaker builds an audience around specific shared interests, content specificity may have a positive, as opposed to negative impact on audience attention. The literature suggests that social media user intend to post content that shared frequently by many other users (Asur et al. 2011). The literature also explains that retweeting is associated with information sharing, commenting or agreeing on other peoples' messages and entertaining followers (Boyd et al. 2010). Kwak et al. (2010) also show that the most trending topics have an active period of

1 week, while half of retweets of a given tweet occurs within 1 h and 75 % within 1 day. The frequency of retweets is a major factor for estimating the quality of posts. It can be an important criterion since users tend to retweet valuable posts (Chang 2014). In the communities of people who are interested in specific content, users share specific content that followers are more likely to retweet. Intuitively, if a user tweets about multiple topics, interesting to many specific and active communities, he/she is most likely to get more retweets. Consequently, in the following hypothesis we posit a positive association between the number of topics and the frequency of retweets.

> *H2: Tweeting about multiple topics* (*brands or categories*) *is positively associated with the frequency of retweets* (*i.e. visually, peripheral authors, connected to multiple topics, are actual influencers*).

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 (Benevenuto et al. 2010). In traditional media, influencers intend to target a large audience by broadcasting frequently. Similarly, in social media, e.g. in twitter, influencers intend to be more interactive by showing their presence and frequently tweeting (Bruni et al. 2013). 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 frequently sharing the content that they have liked (Ren et al. 2014). A content that has had an impact on a user's mind is shared and gathers attention by others. The volumes of retweets are positively correlated with the level of users' activity (number of tweets) and their in-degree centrality (number of followers), as noted by (Leavitt et al. 2009). In social media, while sharing content, users may be referred as '*generalists*' or '*information sources*' who talk about multiple topics (Hutto et al. 2013). On the contrary, there exist such users, who are very specific in sharing content related to specific topic or brand. These specific authors seems to be potential influence spreaders (Fan and Gordon 2014). We posit that, these authors have to be active participants in each community by talking a lot. Our third hypothesis posits that such authors have a greater probability of being retweeted due to frequent tweets, and can be both potential and actual influencers.

> *H3*: *Tweeting more frequently* (*with a high frequency*) *about a single topic (brand or category) is positively associated with the frequency of retweets* (*i.e. visually, authors, drawn closer to single topic, are both actual and potential influencers*).

We posit the aforementioned three hypotheses that tie *content specificity*, *frequency of tweets* and *frequency of retweets*. Visually, hypothesis H1 can be verified by observing the peripheral authors positioned in the outer-most layers of each cluster (lowest l-shell value, $l_s = 1$), which are only connected to one cluster hub (brand or category). These authors seems to be talking about a single brand or category. Such outlier authors can be *potential* influencers, if they further connect to other authors via content sharing and tweeting about multiple topics (brands or categories). Similarly, hypothesis H2 can be visually verified by observing authors who are placed in between multiple clusters, connected to multiple clusters' hubs (brands or categories), and seem to be talking about multiple topics. These authors are *actual* influencers as they receive a high number of retweets by tweeting about multiple topics. Moreover, hypothesis H3 can be visually verified by observing those authors who are positioned in the inner-most periphery of a single cluster (highest l_s value) and seem to be placed close to the cluster hub (brand or category). Such authors are both *actual* and *potential* influencers as they are most specific about content sharing. These authors tweet frequently about a single topic (brand or category) and receive a high number of retweets.

5 Experimental methodology and results

In this section, we will present the dataset that we have used in our experiment and the network models that we have built from the dataset. Empirical evaluations and related visualization results are also presented in this section.

5.1 Variable definition and operationalization

Each graph G (*A*, *T*) has a node set *A* representing authors and an edge set *T* representing tweets. We define as $N_T(a)$ the total number of tweets posted by author *a*. We define as N_R (a) total number of times author *a,* has been retweeted. Tweets can refer to a brand *b* or to a category *c*. We define as N_B (a) the total number of brands mentioned by each author *a,* in all his/her tweets, i.e. *brand specificity*. Similarly, $N_C(a)$ represents the total number of categories mentioned by each author *a,* in all his/her tweets, i.e. *category specificity*.

5.2 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. Twitter APIs have been queried 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 analyzed with a proprietary semantic engine (Barbagallo et al. 2012) 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 number of retweets (if any), and (d) the identifier of the retweeting author. Our 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 (Anholt 2006). Examples of brand drivers are *Art and Culture*, *Food and Drinks, Events and Sport, Services and 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. Table 1 refer to the descriptive statistics of the original non-linear variables.

5.3 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 3 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 modelled as bipartite graph, 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.
- 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 and Culture, Events and Sports, Fares and Tickets, Fashion and Shopping, Food and Drink, Life and Entertainment, Night and Music, Services and Transport, and Weather and Environmental*). The network is modelled as bipartite graph, 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.

5.4 Network visualization

The empirical results and discussions on network visualization will adopt network N_1 network (i.e. Author → Brand) as reference example. Figure 3 provides an enlarged view of network N_1 visualized by means of the proposed power-law layout algorithm. A summary description for N_1 and N_2 networks is presented in Table 2,

Table 1 Basic descriptive statistics of our dataset

Variable	Value	SD
Number of tweets	957,632	–
Number of retweeted tweets	79,691	–
Number of tweeting authors	52,175	–
Number of retweets	235,790	–
Number of retweeting authors	66,227	–
Average number of tweets per author	10.07	±86.83
Average number of retweeted tweets per author	1.525	±4.67
Average number of retweets per author	1.40	±4.52
Average frequency of retweets per author	0.58	±0.38
Average content Specificity per author	0.35	±0.46

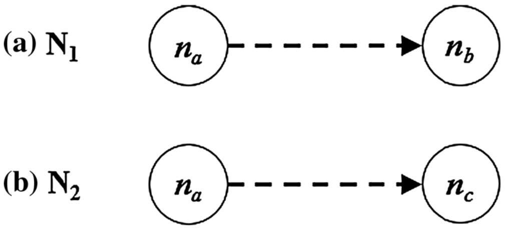

Fig. 3 Network models. **a** N_1: Author → Brand, **b** N_2: Author → Category

where N_R (a) represents the total number of retweets, N_B (a) shows the total number of tweets in which author *a* talked about brand *B* (N_1 network), N_c (a) shows the total number of tweets in which author *a* talked about category *C* (N_2 network), and N_T (a) represents the total frequency of author *a* (i.e. the total number of tweets of author *a*).

The network visualization depicted in Fig. 4 adopts multicolor nodes to represent authors, and highlighted encircled blue (dark) nodes to represent tourism destinations (i.e. brands) 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 brand nodes, which in this case are the hub nodes.

The groups of author nodes cluster together all those authors that are connected to the same hubs (i.e. brands) referred as a *cluster*. Our approach provides a visual clustering for those authors who have tweeted about the same brand.

5.5 Empirical results

This section reports on the empirical testing and evaluation of the proposed hypotheses. First, we discuss our research model and then we present empirical results.

5.5.1 Research model

AMOS 20 (Arbuckle 2011) has been used to analyze the research model that we adopted for estimation analysis is shown in Fig. 5. In Fig. 5 we report each variable relationship only in its standardized regression coefficient's sign (note that signs are consistent between the two data sets N_1 and N_2). In this model, N_T (a) represents a dependent variable as it is measured with multiple independent variables, which are N_R (a), N_B (a), and N_C (a).

5.5.2 Statistical analysis

All statistical analyses have been performed with SPSS 20 (Pallant 2010). Correlation and Regression analyses have been performed on our data set. Table 3 reports the descriptive statistics of each variable from our dataset that we used for

Table 2 Descriptive statistics on the dimensions of N_1 and N_2 networks

Authors	N_R (a)	N_1 N_B (a)	N_2 N_C (a)	N_T (a)
398	92	856	1,913	2,769
1,662	364	2,905	5,959	8,864
10,710	2,907	12,559	18,498	31,057
18,711	5,329	21,140	29,842	50,982
30,310	8,690	33,684	46,120	79,804
37,626	10,529	41,620	56,960	98,580
47,295	12,833	52,208	71,667	1,23,875

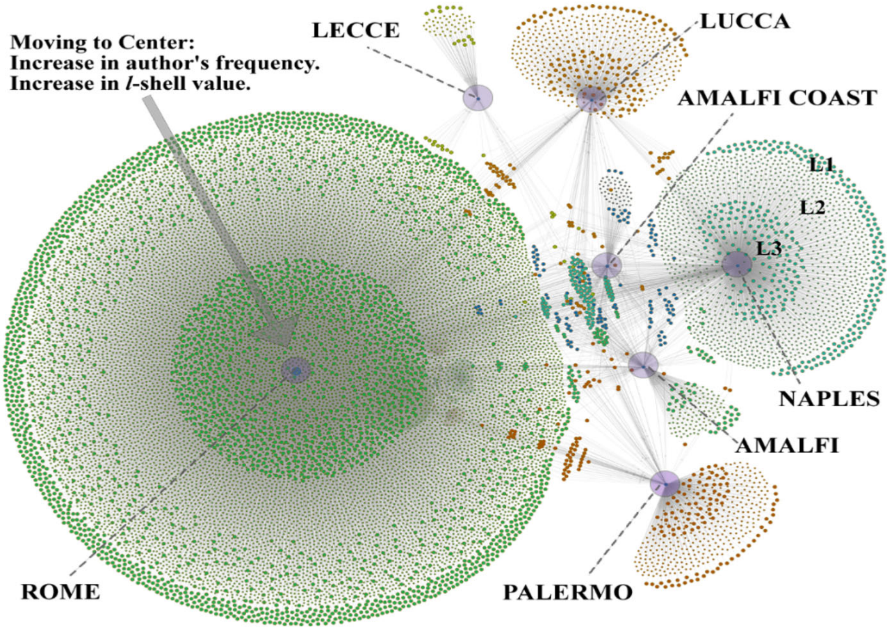

Fig. 4 Network N_1: Author → Brand (enlarged view)

statistical analysis and to validate our proposed research hypotheses, as discussed in Sect. 3.

Table 4 presents the correlation matrix of our data variables. Table 4 shows that correlation is significant at 0.01 level (2-tailed). All persistence variables are positively correlated with each other and, thus, have a significant impact upon each other.

The regression estimation results of the research model are shown in Table 5. All relationships between persistence metrics (i.e. N_R (a), N_B (a), and N_C (a)) and the persistence latent variable (i.e. N_T (a)) are significant, with $p < 0.001$. This confirms that factorization was performed correctly over fitted research model.

- *Hypothesis H1* Hypothesis H1 "Tweeting with a high frequency of tweets is positively associated with number of topics (brands or categories) (i.e. visually *potential influencers* are the peripheral authors)" has been tested through correlation. By Table 4, both N_C (a) and N_B (a) have positive correlation of 0.898 and 0.590, respectively with N_T (a), at 0.01 level of significance. Hence, both correlation values support the hypothesis H1. It means that, *generalist* authors, who tweet about multiple topics (brands or categories), are more likely to be *content specifiers*. Such authors by having greater probability of sharing contents, can be *potential influencers* in their network. Similarly, through visualization results we can also observe the big sized author nodes who tweet a lot about multiple brand (Fig. 4) or about multiple categories (Appendix).
- *Hypothesis H2* Similarly, hypothesis H2, "Tweeting about multiple topics (brands or categories) is positively associated with the frequency of retweets (i.e. visually, peripheral authors, connected to multiple topics, are *actual influencers*)", has been tested through correlation. By Table 4, both N_C (a) and N_B (a) have a positive correlation of 0.254 and 0.235, respectively with N_R (a), at 0.01 level of significance. Hence, both correlation values support the hypothesis H2. This means that, authors, who have a large number of retweets, are also *content specifiers* or can also be '*information sources*' or '*generators*'. Such authors can be *actual influencers* in spreading the influence among networks, as they receive large number of retweets by tweeting about multiple topics. From a visualization standpoint, if we explore the produced graph (e.g. Fig. 4), authors who seems to be big sized nodes (visually drawn in-between multiple

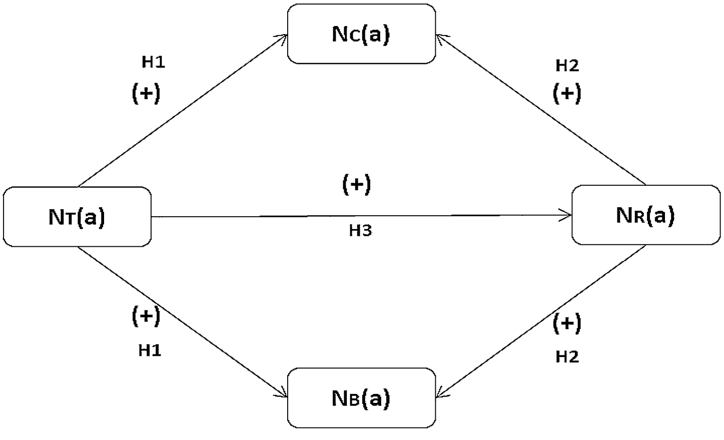

Fig. 5 Research model

Table 3 Descriptive statistics of each variable from dataset

	N_R (a)	N_B (a)	N_C (a)	N_T (a)
Mean	1.37	1.04	1.53	2.78
SE of mean	0.009	0.000	0.001	0.002
SD	10.817	0.283	1.109	1.837
Variance	117.007	0.080	1.230	3.375

Table 4 Correlation matrix of persistence variables (Pearson Index)

	N_T (a)	N_R (a)	N_B (a)	N_C (a)
N_T (a)	1	0.326	0.590	0.898
N_R (a)	0.326	1	0.254	0.235
N_B (a)	0.590	0.254	1	0.392
N_C (a)	0.898	0.235	0.392	1

Table 5 Estimates of regression weights for the research model

$V_{\text{Dependent}}$	$V_{\text{Independent}}$	R_W	SE	*p* value
N_R (a)	N_T (a)	0.082	0.000	<0.001
N_B (a)	N_T (a)	0.000	0.000	<0.001
N_C (a)	N_T (a)	0.000	0.000	<0.001

cluster peripheries) talking about multiple topics (brands or categories), also have a high number of retweets as well.

- *Hypothesis H3* Similarly, hypothesis H3, "Tweeting more frequently about a single topic (brand or category) is positively associated with the frequency of retweets (i.e. visually, authors drawn closer to single topic, are both *actual and potential influencers*)", has been tested through correlation. By observing values from Table 4, N_T (a) and N_R (a) have a positive correlation of 0.326 at 0.01 level of significance. Although the correlation coefficient is not high, the *p* value in Table 5 shows significance and seems to support a positive (though weak) correlation between N_T (a) and N_R (a). As per descriptive statistics of networks, presented in Table 2, we can observe that as the number of tweets increases, the number of retweets also increases for each size or network topology. From a visual standpoint, as shown in Fig. 4, we know that the nodes (which are drawn closer to a single brand in the innermost periphery of distinct clusters) are those authors who tweet most frequent about a specific brand in its cluster. Such authors are connected closer to cluster hubs (brands or categories), by having a high *l*-shell value as of having a high number of tweets (as discussed earlier in Sect. 3).

6 Discussion

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. Within a single cluster, multiple layers seem to be formed. By implementing the *l-shell* decomposition methodology, the outside layer consists of author nodes who posted a tweet only once, as we move inward towards the brand node (hub), the frequency of tweeting increases. Hence, the closest nodes to a hub represent the authors who tweeted most about that brand and are both *potential and actual influencers*. The power-law layout algorithm has provided a network layout that is very effective in highlighting a specific property of authors which was not a measured variable in our dataset, i.e. their specificity (or generality) with respect to a topic (i.e. a brand as in Fig. 4 or category in Appendix). Authors belonging to different clusters are in fact those who are more *generalist* in their content sharing, since they tweet about multiple different brands. On the contrary, authors belonging to the innermost clusters are those who are very *specific* in sharing content related to one brand.

Since the *specificity* (generality), *frequency of tweets* and *retweets* of authors was not an explicitly measured variable in our dataset, it is possible to posit that the proposed network layout algorithm can be considered as a powerful visual data analysis tool, since it is effective in providing visual representations of networks that help unveiling specific (implicit) properties of the represented networks.

We also noticed that, as the graph size increases, more peripheral layers seems to be formed surrounding hub nodes, which increase the influence spread across newly formed peripheral layers in multi-layered form. Authors seem to evolve by tweeting about multiple topics among multiple peripheries. We can visually identify the increase in influence spread, as shown in Figs. 6 and 7, which are larger graphs of the N_1 type network, as compared to Fig. 4, where the addition of more multi-layered peripheral nodes around hub-nodes (i.e. brands) increases the influence spread across those peripheral layers. The outlier authors along the periphery can be potential influence spreaders, if they connect with other clusters through retweeting and, thus, play a critical role in determining influence. As presented in Fig. 4, network N_1 is related to the relationship between authors and brands, i.e., touristic destinations. In this case, the clustering of nodes provides a distinct clustering 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 semantics of the network visualization in this case is 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.

This paper's findings have some practical implications on how to design a strategy to promote tourism destinations. For example, findings suggest that to promote a specific brand, WoM may become more efficient by linking that specific brand with other brands, as this seems to increase reach and influence. For example, they can share posts comparing their brand with other competing and non-competing brands (Baum 1999; Enright and Newton 2004; Leask 2015). Similarly, they can identify the most popular and least popular brands, as the multi-layered peripheral network of *author* nodes reveals potential and actual influencers. They can target authors in the periphery who can be *information spreaders* and, thus, connect to other communities in order to increase reach. Tourism practitioners can also identify the most widely discussed topics (categories) and focus on them in their advertising campaigns. For example, while addressing a specific brand (e.g. Rome), they can relate it with a specific category (e.g. Arts and Culture), in order to increase the specificity of their posts. From a visualization standpoint, tourism practitioners can also identify the key players in the network and classify them as *information spreaders*, *sources*, or *seekers*. *Information spreaders* can either be *generalist* authors who are connected to multiple communities and discuss about multiple topics, as they have a broad reach and a significant influence. Becoming an engaging member of relevant communities will give social media users a chance to promote content to a targeted audience and increase their actual influence.

7 Conclusion and future work

This paper proposes a novel visual approach to the analysis and exploration of social networks in order to identify and visually highlight influencers (i.e., hub nodes), and 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. Empirical testing and evaluation results show that the proposed three hypothesis that tie *content specificity*, *frequency of tweets* and *retweets* are supported. Moreover, the parameters like *specificity*, *frequency*, and *retweets* are also mutually correlated, and have a significant impact on an author's influence and encourage us to further explore social network's intrinsic characteristics.

Such outcomes can be further utilizes by tourism practitioners, marketing departments or social media community. For example, one can analyses the most competitive locations, events or initiatives in the market. Social media marketing managers can also visually identify major key players in the network, like *information spreaders* and *information sources*. In social media communities, users like *information seekers,* would be able to visually identify the actual and potential influencers and can further follow them.

Although our experiment can be repeated with data from domains different from tourism, additional empirical work is needed to extend testing to multiple datasets and domains. Future work will consider measures of influence with additional parameters (e.g. number of followers, lists, mentions, URLs, etc.). In our current work, we are studying a measure of influence through the proposed visualization approach, which can be used to rank influential nodes in social networks (Metra 2014) and help the practical use of our research results.

8 Appendix

Figures 6, 7, 8 and 9 provide additional visualizations of networks N_1 and N_2 from our dataset. An enlarged and zoomable version of these network layouts can be accessed online.

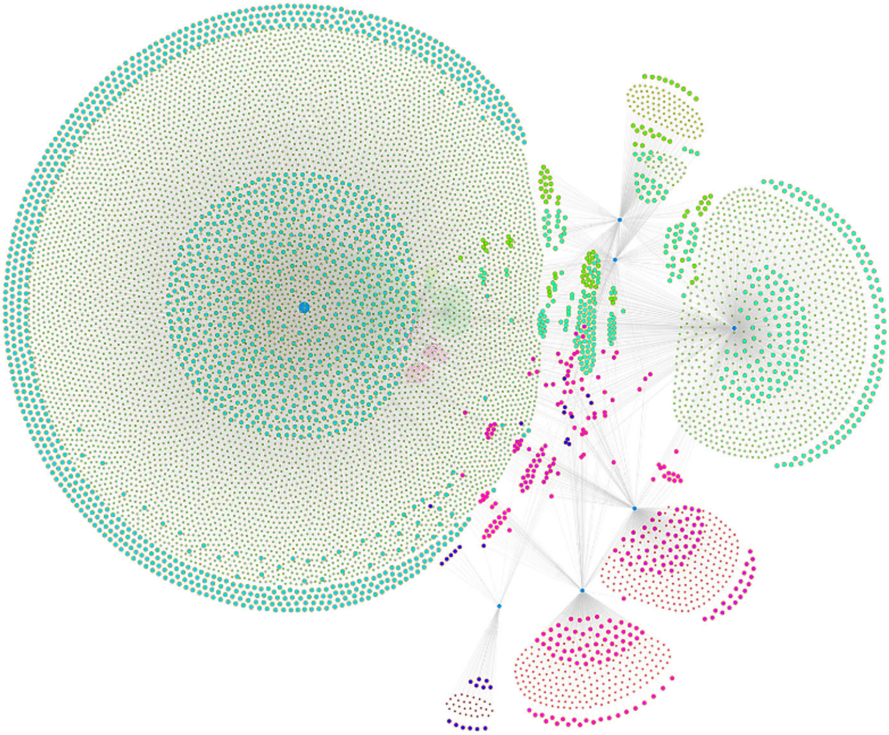

Fig. 6 Network visualizations of N_1 (Author → Brand)

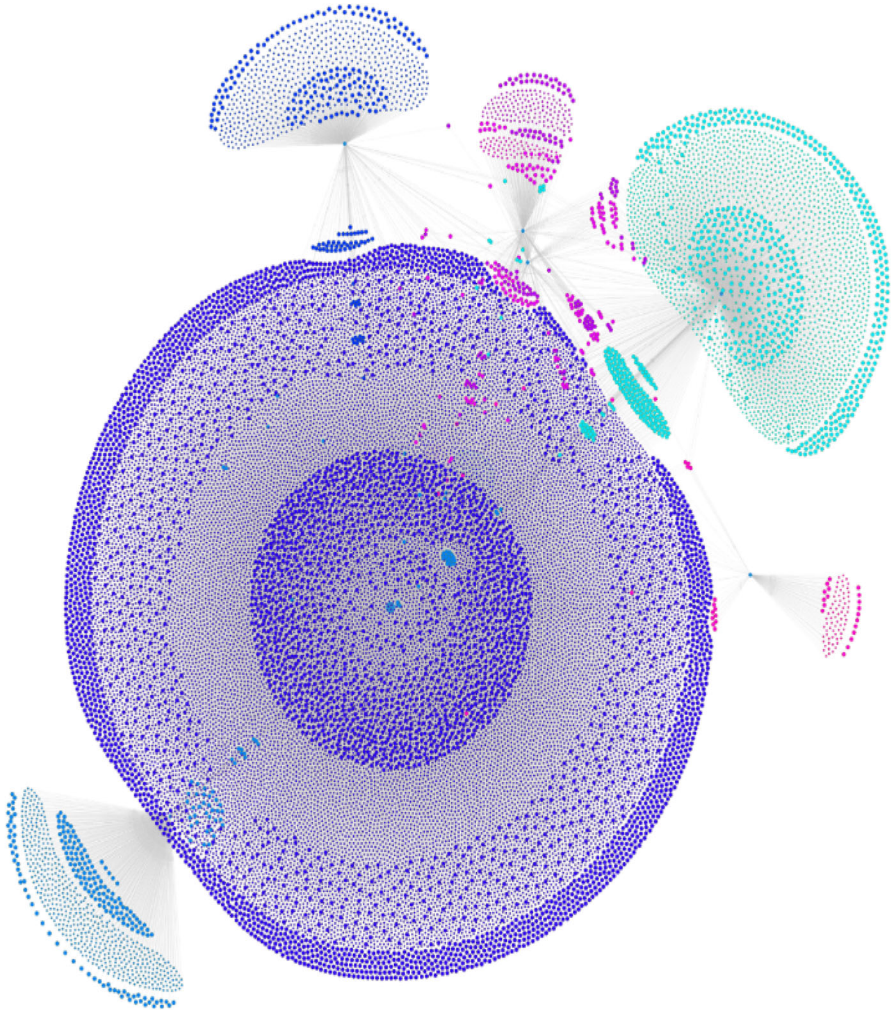

Fig. 7 Network visualizations of N_1 (Author → Brand)

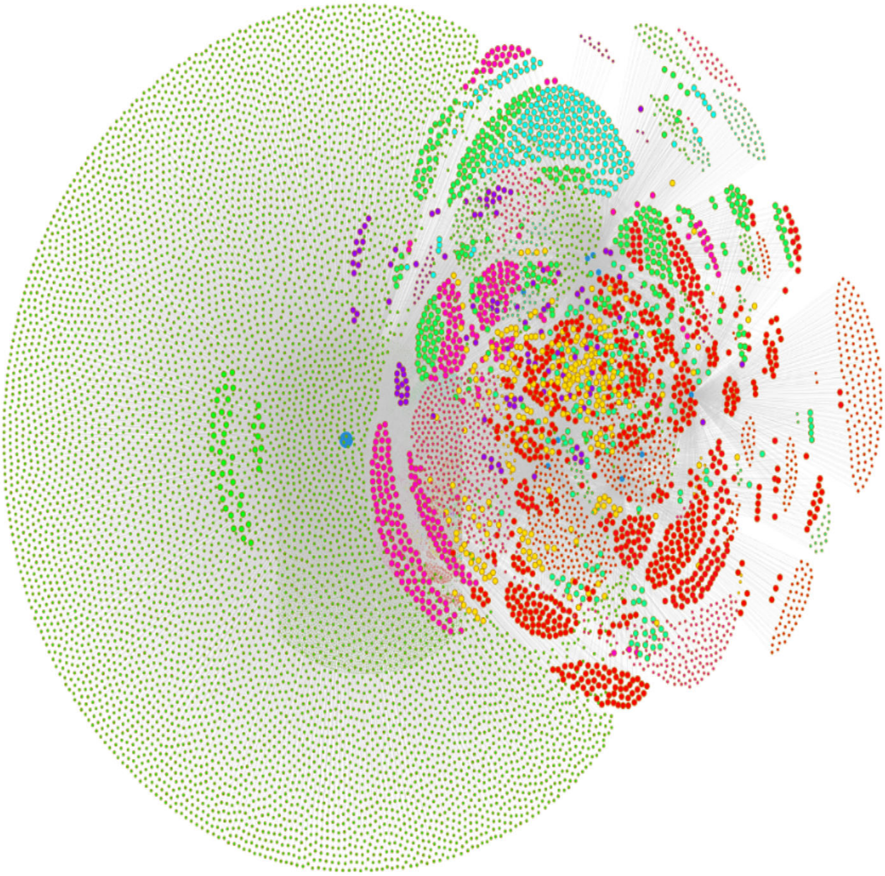

Fig. 8 Network visualizations of N_2 (Author → Category)

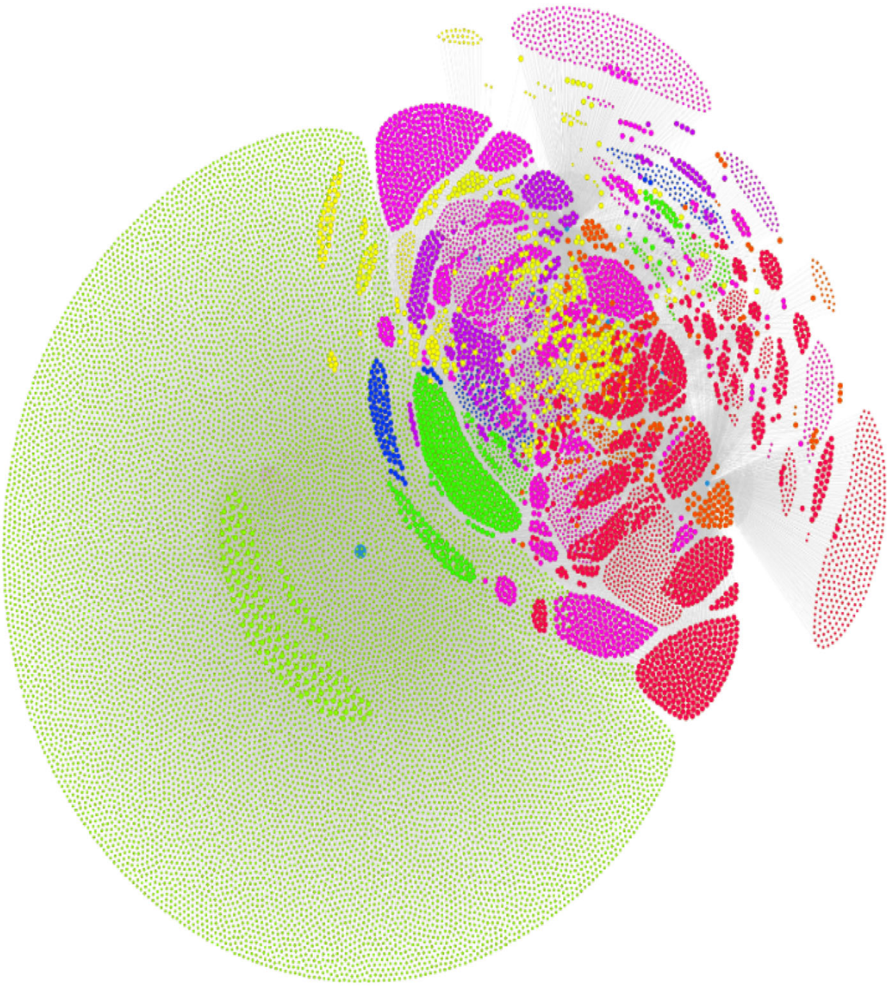

Fig. 9 Network visualizations of N_2 (Author → Category)

References

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

Anholt S (2006) Competitive identity: the new brand management for nations, cities and regions. Palgrave Macmillan

Arbuckle JL (2011) IBM SPSS Amos 20 user's guide. Amos Development Corporation, SPSS Inc

Asur S, Huberman BA, Szabo G, Wang C (2011) Trends in social media: persistence and decay. In: ICWSM

Baggio R (2005) Complex systems, information technologies, and tourism: a network point of view. Inf Technol Tourism 8:15–29

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on web search and data mining, pp 65–74

Barbagallo D, Bruni L, Francalanci C, Giacomazzi P (2012) An empirical study on the relationship between twitter sentiment and influence in the tourism domain. In: Information and communication technologies in tourism. Springer, pp 506–516

Baum T (1999) Themes and issues in comparative destination research: the use of lesson-drawing in comparative tourism research in the North Atlantic. Tour Manag 20:627–633

Benevenuto F, Cha M, Gummadi KP, Haddadi H (2010) Measuring user influence in Twitter: the million follower fallacy. In: International AAAI conference on weblogs and social (ICWSM10), pp 10–17

Bigonha C, Cardoso TNC, Moro MM, Gonçalves MA, Almeida VAF (2012) Sentiment-based influence detection on Twitter. J Br Comput Soc 18:169–183

Boyd D, Golde S, Lotan G (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter IEEE, pp 1–10

Bruni L, Francalanci C, Giacomazzi P, Merlo F, Poli A (2013) The relationship among volumes, specificity, and influence of social media information. In: Proceedings of international conference on information systems

Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. Proc Natl Acad Sci 104:11150–11154

Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in Twitter: the million follower fallacy ICWSM 10:10–17

Chan DSM, Chua KS, Leckie C, Parhar A (2004) Visualisation of power-law network topologies. In: The 11th IEEE international conference on networks, ICON2003, pp 69–74

Chang J-Y (2014) An evaluation of twitter ranking using the retweet information. J Soc e-Business Studies 17

Enright MJ, Newton J (2004) Tourism destination competitiveness: a quantitative approach. Tourism Manag 25:777–788

Fan W, Gordon MD (2014) The power of social media analytics. Commun ACM 57:74–81

Francalanci C, Hussain A (2014) A visual approach to the empirical analysis of social influence. Paper presented at the DATA 2014, proceedings of 3rd international conference on data management technologies and applications

Francalanci C, Hussain A (2015) A visual analysis of social influencers and influence in the tourism domain. In: Tussyadiah I, Inversini A (eds) Information and communication technologies in tourism. Springer, Berlin, pp 19–32. doi:10.1007/978-3-319-14343-9_2

Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Netw 1:215–239

Hussain A, Latif K, Rextin A, Hayat A, Alam M (2014) Scalable visualization of semantic nets using power-law graphs. Appl Math Inf Sci 8:355–367

Hutto CJ, Yardi S, Gilbert E (2013) A longitudinal study of follow predictors on twitter. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 821–830

Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6:888–893

Klotz C, Ross A, Clark E, Martell C (2014) Tweet!—and I can tell how many followers you have. In: Recent advances in information and communication technology. Springer, Berlin, pp 245–253

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, pp 591–600

Leask A (2015) Destination competitiveness: a comparative study of Hong Kong. Macau and Singapore Tourism Analysis

Leavitt A, Burchard E, Fisher D, Gilbert S (2009) The influentials: new approaches for analyzing influence on twitter. Web Ecol Project 4:1–18

Meraz S (2009) Is there an elite hold? Traditional media to social media agenda setting influence in blog networks. J Comput Mediat Commun 14:682–707

Messias J, Schmidt L, Oliveira R, Benevenuto F (2013) You followed my bot! Transforming robots into influential users in Twitter First Monday 18

Metra I (2014) Influence based exploration of twitter social network. Politecnico di Milano, Milan

Myers SA, Leskovec J (2014) The bursty dynamics of the Twitter information network. In: Proceedings of the 23rd international conference on world wide web, pp 913–924

Pallant J (2010) SPSS survival manual: a step by step guide to data analysis using SPSS. McGraw-Hill International

Perline R (2005) Strong, weak and false inverse power laws. Stat Sci 20:68–88

Ren Z-M, Zeng A, Chen D-B, Liao H, Liu J-G (2014) Iterative resource allocation for ranking spreaders in complex networks. Europhys Lett 106:48005

Sparrowe RT, Liden RC, Wayne SJ, Kraimer ML (2001) Social networks and the performance of individuals and groups. Acad Manag J 44:316–325

Xu X, Yuruk N, Feng Z, Schweiger TA (2007) Scan: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 824–833