The Three Dimensions of Social Prominence

Diego Pennacchioli[1], Giulio Rossetti[1] Luca Pappalardo[1], Dino Pedreschi[1],
Fosca Giannotti[1], and Michele Coscia[2]

[1] KDDLab ISTI-CNR, Via G. Moruzzi 1, Pisa, Italy
{name.surname}@isti.cnr.it
[2] CID Harvard University, 79 JFK St, Cambridge, MA, US
michele_coscia@hks.harvard.edu

Abstract. One classic problem definition in social network analysis is
the study of diffusion in networks, which enables us to tackle problems
like favoring the adoption of positive technologies. Most of the attention
has been turned to how to maximize the number of influenced nodes,
but this approach misses the fact that different scenarios imply different
diffusion dynamics, only slightly related to maximizing the number of
nodes involved. In this paper we measure three different dimensions of
social prominence: the Width, i.e. the ratio of neighbors influenced by
a node; the Depth, i.e. the degrees of separation from a node to its
influenced nodes; and the Strength, i.e. the intensity of the influence from
a node. By defining a procedure to extract prominent users in complex
networks, we detect associations between the three dimensions of social
prominence and classical network statistics. We validate our results on a
social network extracted from the Last.Fm music platform.

1 Introduction

One classic problem in social network analysis is understanding diffusion effects
in networks. Modeling diffusion processes on complex networks enables us to
tackle problems like preventing epidemic outbreaks [6] or favoring the adop-
tion of new technologies or behaviors by designing an effective word-of-mouth
communication strategy. In our paper, we are focused on the social prominence
aspect of the diffusion problem in networks.

In the setting of favoring social influence, most of the attention of researchers
has been put on how to maximize the number of nodes subject to the spread-
ing process. This is done by choosing appropriate seeds in critical parts of the
network, such that their likelihood of being prominent users is maximum, to
possibly achieve larger cascades. While larger cascades are obviously part of the
overall aim, we argue that it is not the unique dimension of this problem. Three
other dimensions are relevant: the *width*, the *depth* and the *strength* of the so-
cial prominence of any given node in a network. The width of a node is being
prominent for its immediate neighbors; the depth is its ability to be the root of
long cascades; the strength is being the root of an intense activity.

Many real-world scenarios focus on a more specific diffusion pattern that
requires a multidimensional understanding of the prominence mechanics at play,

along the three mentioned dimensions. Some examples are: (i) an analyst needs information from the personal acquaintances of a subject, the important aspect is that many among the subject's direct connections respond, ignoring people two steps away or more; (ii) a person is interested in finding another person with a given object, the important aspect is that some people are able to pass her message through a long chain that ends with the desired target; (iii) a music artist wants to influence people in a social network to listen to her new album, the important aspect is that some people are influenced above the threshold that will make them actually buy the album.

In Scenario 1 there is the need of a broad diffusion in the first degree of separation. Scenario 2 requires a targeted diffusion similar to a Depth First Search algorithm on graphs. Scenario 3 requires a high-intensity diffusion. Different scenarios may require any combination of the three.

In this paper, we make use of three measures to capture the characteristics of these three scenarios: the Width, Depth and Strength of social prominence. The Width measures the ratio of the neighbors of a node that follows the node's actions. The Depth measures how many degrees of separation there are between a node and the other nodes that followed its actions. The Strength measures the intensity of the action performed by some nodes after the leader.

We study what the relations are between these three measures to understand if we are capturing three orthogonal dimensions of social prominence. We also study the relations between the Width, Depth and Strength measures and different node properties, with the aim of predicting the diffusion patterns of different events, given the characteristics of the nodes that lead their diffusion.

To validate our concepts, we constructed a social network from the music platform Last.Fm[3], along with the data about how many times and when each user listens to a song performed by a given artist. We detect who are the prominent users for each artist, i.e. the users who start listening to an artist before any of their neighbors. We calculate for each prominent user its Width, Depth and Strength, along with its network statistics such as the degree and the betweenness centrality, looking for associations between them. We then create a case study to understand what are the different dynamics in the spread of artists belonging to different music genres, by using the artists' tags.

To sum up, the contributions of our paper are: (i) a proof that social diffusion indeed follows at least these three dimensions, which are uncorrelated or anti-correlated; (ii) the discovery of some significant associations between the three dimensions of social prominence and some traditional network measures; (iii) the ability to predict the patterns of diffusion of particular events by looking at the characteristics of the leaders spreading them.

2 Related Work

In the last decade, there has been growing interest in the studies of diffusion processes. Two phenomena are tightly linked to the concept of diffusion: the

---

[3] http://www.last.fm/

spread of biological [6] or computer [16] viruses, and the spread of ideas and innovation through social networks, the so-called "social contagion" [2]. In both cases, the patterns through which the spreading takes place are determined not just by the properties of the pathogen/idea, but also by the network structures of the population it is affecting.

Some models have been defined to understand the contagion dynamics: the SIR [10], SIS and SIRS [15] models. The idea behind them is that each individual transits between some stages in the life cycle of a disease: from Susceptible (S) to Infected (I), and from Infected to either Recovered (R) or again Susceptible. The availability of Big Data conveying information about human interactions and movements encouraged the production of more accurate data-driven epidemic models. For example, [6] takes into account the spatio-temporal dimension. In [16], authors study the spreading patterns of a mobile virus outbreak.

Christakis and Fowler studied the role of social prominence in the spread of obesity [4], smoking [5] and happiness [8]. Their results suggest that these health conditions may exhibit some amount of "contagion" in a social sense: although the dynamics of diffusion are different from the biological virus case, they nonetheless can spread through the social network.

3 Leader Detection

Each diffusion process has its starting points. Any idea, disease or trend is firstly adopted by particular kinds of actors. Such actors are not like every other actor: they have an increased sensibility and they are the first to perform an action in a given social context. We call such actors prominent users, or *leaders*, because they are able to anticipate how other actors will behave. Given a graph, several interesting problems arise regarding how information spreads over its topology: can we identify the *leaders*? Can we characterize them? What kind of knowledge should we expect to extract from their analysis?

Our approach aims to detect *leaders* through the analysis of two correlated entities: the topology of the social graph and the set of actions performed by the actors (nodes). When discussing the roles of those entities, we refer respectively to the following definitions:

Definition 1 (Social Graph). *A social graph $\mathcal{G}$ is composed by a set of actors (nodes) V connected by their social relationships (edges) E. Each edge $e \in E$ is defined as a couple (u, v) with $u, v \in V$ and, where not otherwise specified, has to be considered undirected. With $\Gamma(u)$ we identify the neighbor set of a node u.*

Definition 2 (Action). *An action $a_{u,\psi} = (w, t)$ defines the adoption by an actor $u \in V$, at a certain time t, of a specific object ψ with a weight $w \in \mathcal{R}$. The set of all the actions of nodes belonging to a social graph $\mathcal{G}$ will be identified by $\mathcal{A}$, while the object set will be called Ψ.*

We identify with $\mathcal{G}_\psi = (V_\psi, E_\psi)$, where $V_\psi \subset V$ and $E_\psi \subset E$, the induced subgraph on $\mathcal{G}$ representing respectively the set of all the actors that have performed an action on ψ, and the edges connecting them. We depict an example

$$
\begin{array}{ll}
a_{1,x}=(1,0) & a_{4,y}=(2,1) \\
a_{2,x}=(2,1) & a_{7,y}=(4,3) \\
a_{3,x}=(1,2) & a_{8,y}=(5,2) \\
a_{4,x}=(4,6) & a_{6,y}=(6,2) \\
a_{5,x}=(1,4) & a_{1,y}=(2,3) \\
a_{6,x}=(6,7) & a_{2,y}=(1,5)
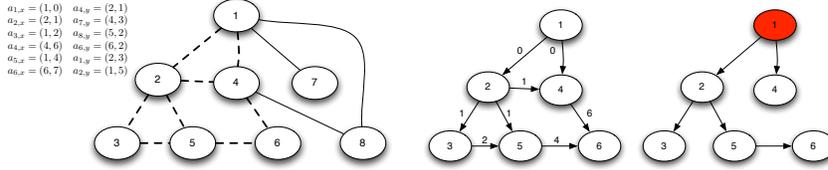\end{array}
$$

Fig. 1: Toy Example. On the *left* the social graph $\mathcal{G}$ and action set $\mathcal{A}$, where $x, y \in \Psi$ are the objects of the actions; in the *center* the induced subgraph for the action x; on the *right* the diffusion tree for x. In red we highlighted the leader (root) for the given tree.

of the social graph and the set of actions in Figure 1 *(left)*, where the induced subgraph for the object x is highlighted with a dashed line.

Given the nature of a diffusion process, we would expect that each *leader* will be prominent among its neighbors, being the root of a cascade event that follows some rigid temporal constraints. Our constraint is that a node u precedes a neighbor v iff given $t_{u,\psi} \in a_{u,\psi}$ and $t_{v,\psi} \in a_{v,\psi}$ is verified that $t_{v,\psi} > t_{u,\psi}$ and $t_{v,\psi} - t_{u,\psi} \leq \delta$. Here, δ is a temporal resolution parameter that limits the cascade effect: if $t_{v,\psi} - t_{u,\psi} > \delta$, we say that v executed action $a_{v,\psi}$ independently from u, as u's prominence interval is over.

We transform each undirected subgraph $\mathcal{G}_\psi$ in a directed one imposing that the source node of an edge must have performed its action before the target node. After that, each edge (u, v) will be labeled with $min(t_{u,\psi}, t_{v,\psi})$ to identify when the diffusion started going from one node to the other. The directed version of $\mathcal{G}_\psi$ represent all the possible diffusion paths that connect leaders with their "tribes" (Figure 1 *(center)* an example for the object $x \in \Psi$).

From now on, for a given object ψ, we will refer to the corresponding leader set as $\mathcal{L}_\psi$: when no action is specified the set $\mathcal{L}$ will be used to describe the union of all the $\mathcal{L}_\psi$ for the graph $\mathcal{G}$. To be defined a *leader* an actor should not have any incoming edges in $\mathcal{G}_\psi$. This is because a prominent user cannot act after another user (they are, in their surroundings, innovators), and is a direct consequence to the adoption of a directed graph to express diffusion patterns. Given this definition, for each directed connected component $\mathcal{C}_\psi \subset \mathcal{G}_\psi$ multiple nodes can belong to $\mathcal{L}_\psi$.

One caveat of our problem definition comes from the fact that a leader may be influenced by exogenous events. This is not a problem as we are not measure a node's influence, but a node's prominence, i.e. its propensity to act faster than others to any kind of exogenous and/or endogenous influence.

To study the real path of diffusion given an action a and a leader l we build a minimum diffusion tree:

Definition 3 (Leader's Minimum Diffusion Tree). *Given an action a_ψ, a directed connected component $\mathcal{C}_\psi$ and a leader $l \in \mathcal{L}_\psi$, the minimum diffusion tree $T_{l,\psi} \subset \mathcal{C}_\psi$ is the Minimum spanning tree (MST) having its root in l and built minimizing the temporal label assigned at the edges.*

An example of minimum diffusion tree for the node 1 and object x is shown in Figure 1 *(right)*. For each object the diffusion process on a given network is independent and that, given temporal dependencies on its adoption (expressed through actions $a_{*,\psi} \in \mathcal{A}$), it is possible to identify the origin points of the diffusion. The identified *leaders* will show different topological characteristic and will be prominent in their surroundings in different ways: our aim is to classify diffusion *leaders* characterizing some of their common traits.

4 Measures

To capture the three dimensions of social prominence we need three network measures. We call these dimensions Width, the ratio of neighbors mirroring an action after a node; Depth, how many degrees of separation are in between a node and the most distant of the nodes mirroring its actions; and Strength, how strongly nodes are mirroring a node's action.

Given a leader, the Width aims to capture the direct impact of her actions on her neighbors, i.e. the degree of importance that a leader has over her friends.

Definition 4 (Width). *Let G be a social graph, $\psi \in \Psi$ an object and $l \in \mathcal{L}_\psi \subset V$ a leader: the function $width : \mathcal{L}_\psi \to [0,1]$ is defined as:*

$$width(l, \psi) = \frac{|\{u | u \in \Gamma(l) \wedge \exists a_{u,\psi} \in \mathcal{A}\}|}{|\Gamma(l)|} \tag{1}$$

The value returned is the ratio of all the neighbors that, after the action of the leader, have adopted the same object.

The Depth measure evaluates how much a leader can be prominent among other prominent leaders, which can be prominent on other leaders and so on.

Definition 5 (Depth). *Let $T_{l,\psi}$ be a minimum diffusion tree for a leader $l \in \mathcal{L}_\psi$ and a given object $\psi \in \Psi$: the function $depth : T_{l,\psi} \to \mathbb{N}$ computes the length of the maximal path from l to a node $u \in T_{l,\psi}$. The function $depth_{avg} : T_{l,\psi} \to \mathbb{R}$ computes the average length of paths from l to any leaf of the tree.*

The last proposed measure, the Strength, tries to capture quantitatively the total weight of the adoption of an object after the leader's action. A leader is strongly prominent if the nodes among which she is prominent are very engaged in adopting what she adopted. Direct prominence diminishes as new adopters become more distant, in the network sense, from the original innovator. Therefore, we decided to introduce a distance damping factor.

Definition 6 (Strength). *Let $T_{l,\psi}$ be a minimum diffusion tree for a leader $l \in \mathcal{L}_\psi$ and an object $\psi \in \Psi$; $0 < \beta < 1$ a damping factor: the function $strength : T_{l,\psi} \times (0,1) \to \mathbb{R}$ is defined as:*

$$strength(T_{l,\psi}, \beta) = \sum_{i \in [0, depth(l)]} \beta^i L(T_{l,\psi}, i) \tag{2}$$

where $L : T_{l,\psi} \times \mathbb{N} \rightarrow \mathbb{R}$ is defined as:

$$L(T_{l,\psi}, i) = \sum_{\{u|u \in T_{l,\psi} \wedge distance(l,u)=i\}} \frac{w_{u,\psi}}{w_u} \tag{3}$$

and represents the sum, over all the nodes u at distance i from l, of the ratio between the weight of action ψ and the total weight of all the actions taken.

Given the example in Figure 1, what are the Width, Depth and Strength values for the red node leader and the action x?

Width: from Figure 1*(left)* we see that $\Gamma(1) = \{2, 4, 7, 8\}$, i.e. 4 nodes. Given that $\Gamma_x(1) = \{u|u \in \Gamma(1) \wedge \exists a_{u,x}\} = \{2, 4\}$, we have $width(1, x) = \frac{|\Gamma_x(1)|}{|\Gamma(1)|} = 0.5$.

Depth: the leaves in Figure 1*(rigth)* are nodes 3, 4 and 6. Node 4 is a direct neighbor of 1, while node 3 is two edges away. The longest chain is $1 \rightarrow 2 \rightarrow 5 \rightarrow 6$, therefore $depth(T_{1,x}) = 3$. We can also calculate $depth_{avg}(T_{1,x})$, that is the average path length in the tree from node 1 to all the leaves: $\frac{1+2+3}{3} = 2$.

Strength: we need to use the number of times each node performed action x. We also set our damping faction $\beta = 0.5$. At the first degree we have nodes 2 and 4, that performed action x 2 and 4 times respectively; they also performed action y 1 and 2 times respectively: their contribution is then $\beta^0 \times (\frac{2}{2+1} + \frac{4}{4+2})$. Nodes 2 and 5 are at the second degree of separation as they never performed action y, therefore they add: $\beta^1 \times (1 + 1)$. Finally, at the third degree of separation, node 6 adds $\beta^2 \times \frac{6}{6+6}$. Wrapping up, $strength(T_{1,x}, 0.5) = 2.458\bar{3}$.

5 Experiments

In this section we present our data extracted from the music social media Last.Fm. We use the data to characterize the Width, Depth and Strength measures, by searching for associations with network topology measures. Finally, we analyze the prominence of different users for different musical genres.

5.1 Data

Last.Fm is a popular online social network platform, where people can share their own music tastes and discover new artists and genres basing on what they like. Once a user subscribes to an account, she can either start listening Last.Fm personalized Radio or send data about her own offline listenings. For each song, an user can express her preferences and add tags (e.g. genre of the song). Lastly, an user can add friends (undirected connections, the friendship request must be confirmed) and search her neighbors w.r.t. musical tastes. An user can see, in her homepage, her friends' activities.

Using Last.Fm APIs[4], we downloaded a sample of the UK user graph, starting from a set of nodes and implementing a breadth-first approach. We decided to explore the graph up to the fifth degree of separation from our seeds. For each

---

[4] http://www.last.fm/api/

user, we retrieved: (a) her connections, and (b) for each week in the time window from Jan-10 to Dec-11, the number of single listenings of a given artist (e.g. in the week between April 11,2010 and April 18,2010 the user 1234 has listened 66 songs from the artist Metallica).
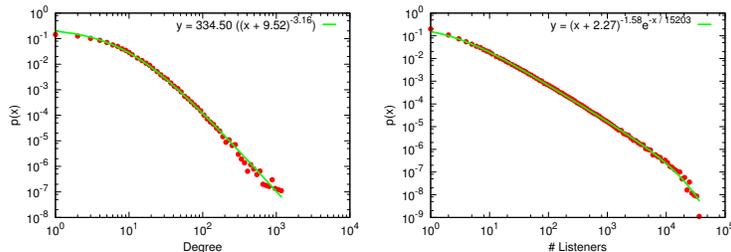

Fig. 2: Log-binned distribution of the nodes' degree. Log-binned distribution of number of listeners per artist.

For each artist we downloaded the list of tags with their counter, representing the number of users that assigned that tag to that artist (e.g. Metallica has 4 tags: "metal" with counter 50670, "hard rock" with 23405, "punk" with 10500 and "adrenaline" with 670). We split all tags, associating the counter to each single word (in the last example: (metal, 50670), (punk, 10500), (hard, 23405), (rock, 23405), (adrenaline, 670)), then we filtered the words referring to a musical genre ((metal, 50670), (punk, 10500), (rock, 23405)). Finally, we assigned a musical genre to an artist iff the survived tag with the greater counter had the relative rate ≥ 0.5 (in the example: $r_{metal}(Metallica) = \frac{50670}{50670+10500+23405} \simeq 0.6$, so Metallica are definitely metal).

After the crawl and cleaning stages, we built our social graph $\mathcal{G}$. In $\mathcal{G}$ every node is a user and each edge is generated by looking at the user's friends in the social media platform. The total amount of nodes is $75,969$, with $389,639$ edges connecting them. In Figure 2 (left) we depicted the log-binned degree distribution of $\mathcal{G}$, along with the best fit.

Each action in the data is one user listening to an artist w times in week t. In Figure 2 (right) we depicted the log-binned distribution of the number of listeners per artist, along with the best fit.Each artist belongs to a music genre (coded in its tag) and we want to use this information in Section 5.3. We decided to focus on 9 main music genres, namely: dance, electronic, folk, jazz, metal, pop, punk, rap and rock. We want to focus only on known artists, so we select only artists belonging to one of these genres and with at least 100 listeners.

Since we are interested in leaders, we need to focus only on new artists that were previously not existent. If an artist was in activity before our observation time window, there is no way to know if a user has listened to it before, therefore nullifying our leader detection strategy. For this reason, we focus only on artist whose first listening is recorded six months after the beginning of our observation period. With this filter, the cardinality of our action set $\mathcal{A}$ is equal to $168,216$ actions, while the object set Ψ contains a total of 402 artists.

In our experimental settings, we set our damping factor $\beta = 0.5$ for the calculation of the Strength measure. We also set $\delta = 3$, meaning that if a user listened to a particular artist three weeks or more after its neighbor then we do not consider her to have been influenced by that neighbor.[5]

5.2 Characterization of the Measures

For each leader, besides Width, Depth and Strength, we calculated also the Degree (number of edges connected to the node), the Clustering coefficient (ratio of triangles over the possible triads centered on the node), the Neighbor Degree (average degree of the neighbors of the node), the Betweenness (share of the shortest paths that pass through the node) and Closeness Centrality (inverse average distance between the node and all the other nodes of the network).

	Width	Strength	Degree	Clustering	Neigh Deg	Bet Centr	Clo Centr
AVG Depth	-0.03	**-0.23**	-0.08	0.05	-0.08	-0.02	**-0.13**
Width	-	0.01	**-0.31**	**0.13**	0.05	-0.07	**-0.59**
Strength	-	-	0.02	-0.02	0.03	0.00	0.04
Degree	-	-	-	**-0.16**	-0.02	**0.77**	**0.56**
Clustering	-	-	-	-	-0.05	-0.06	**-0.32**
Neigh Deg	-	-	-	-	-	-0.00	**0.39**
Bet Centr	-	-	-	-	-	-	**0.22**

Table 1: Pearson correlation coefficient ρ between Width, Depth, Strength and other network statistics for our leaders.

In Table 1 we report the Pearson correlation coefficient ρ between the network measures. We highlighted the correlations whose p-value was significant or whose absolute value was strong enough to draw some conclusions. For the significance of p-values, the traditional choice is to set the threshold at $p < 0.01$. However, given our number of observations, we decided to be more restrictive, setting our threshold at $p < 0.0005$. We also consider a ρ value significant if $|\rho| > 0.1$.

The Depth measure is associated with low Closeness Centrality. This means that a deep prominence is associated to nodes at the margin of the network. It is expected that nodes with high Closeness Centrality have also low Depth: being central, they cannot generate long chains of diffusion. The eccentricity of all the nodes of the network ranges from 6 to 10, meaning that some leaders cannot have a Depth larger than 5. To make a fair comparison, we recalculate the Depth value capping it at 5, meaning that any Depth value larger than 5 is manually reduced to 5. Then, we recalculate the correlation ρ between the Depth capped to 5 and the Closeness Centrality obtaining as result $\rho = -0.1366$, with $p < 0.0005$. We can conclude that central nodes are not associated with deep spread of their prominence in a social network.

For the Width measure, the anti-correlation with the Degree is not meaningful, as the Degree is in the denominator of Definition 4. On the other hand, we observe a positive association with Clustering, meaning that nodes could be prominent in a tightly connected community; and a negative association with

---

[5] To assure experiment repeatability, we made our cleaned dataset and our code available at the page `http://goo.gl/h53hS`

Closeness Centrality, meaning that central nodes could not spread a wide influence. Both associations could be explained with the negative correlation with Degree. Therefore, for both measures we run a partial correlation, controlling for the Degree. In practice, we calculate the correlation between Width and Clustering (or Closeness Centrality) by keeping the Degree constant. Results are in Table 2: even if significant according to the p-value, the relationship between Width and Clustering is very weak and deserves further investigation. On the other hand, it is confirmed that central nodes are also associated with low Width, regardless their degree.

	Clustering	Clo Centr
Partial ρ	0.087216	-0.536861
p-value	1.57×10^{-14}	0

Table 2: Partial correlation and p-value of Clustering and Closeness Centrality with Width, controlling for Degree values.

From Table 1, it appears that the Strength measure is not strongly correlated with traditional network statistics. As a consequence, hubs that are associated with low Depth and low Width, do not have necessarily high Strength, making their prominence in a network questionable. On the other hand, Strength appears to be negatively associated with Depth, suggesting a trade-off between how deeply a node can be prominent in a network and how strong this prominence is on the involved nodes.

The anti-correlation between the Strength and the Depth may be due to the damping factor β: from Definition 6 we see that β decreases nodes' contributions at each degree of separation (i.e. at increasing Depths). As a consequence, nodes that are farther from the leader contribute less to its Strength, i.e. the highest the Depth the smallest are the contributions to the Strength. We recalculated the Strength values by setting $\beta = 1$, therefore ignoring any damping factor and nullifying this effect. We obtained as result $\rho = -0.4168$ and a significant p-value, therefore concluding that the damping factor β is not causing the anti-correlation between Depth and Strength.

To sum up, we summarize the associations as follows: (i) central nodes are not necessarily prominent in a social network (low Width and Depth), a result that confirms [3] and [1]; (ii) longer cascades (higher Depths) are associated with a lower degree of engagement (lower Strengths), a phenomenon possibly related to the role played by "weak ties"; (iii) be prominent among neighbors is probably easier if the node is in a tightly connected community, but more evidences have to be brought to reject the role played by the node's degree.

5.3 Case Study

In this Section we present a case study based on Last.Fm data. Our aim is to use the Leader extraction technique defined in Section 3 and the proposed Width, Depth and Strength measures defined in Section 4 to characterize the spreading of musical genres among the users of the service. We recall that, as described in
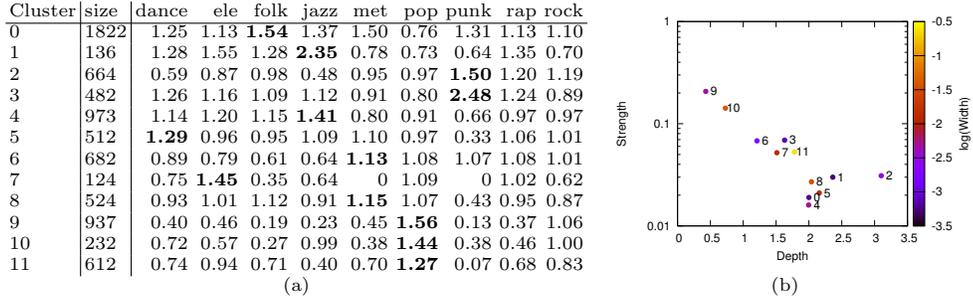
Cluster	size	dance	ele	folk	jazz	met	pop	punk	rap	rock
0	1822	1.25	1.13	**1.54**	1.37	1.50	0.76	1.31	1.13	1.10
1	136	1.28	1.55	1.28	**2.35**	0.78	0.73	0.64	1.35	0.70
2	664	0.59	0.87	0.98	0.48	0.95	0.97	**1.50**	1.20	1.19
3	482	1.26	1.16	1.09	1.12	0.91	0.80	**2.48**	1.24	0.89
4	973	1.14	1.20	1.15	**1.41**	0.80	0.91	0.66	0.97	0.97
5	512	**1.29**	0.96	0.95	1.09	1.10	0.97	0.33	1.06	1.01
6	682	0.89	0.79	0.61	0.64	**1.13**	1.08	1.07	1.08	1.01
7	124	0.75	**1.45**	0.35	0.64	0	1.09	0	1.02	0.62
8	524	0.93	1.01	1.12	0.91	**1.15**	1.07	0.43	0.95	0.87
9	937	0.40	0.46	0.19	0.23	0.45	**1.56**	0.13	0.37	1.06
10	232	0.72	0.57	0.27	0.99	0.38	**1.44**	0.38	0.46	1.00
11	612	0.74	0.94	0.71	0.40	0.70	**1.27**	0.07	0.68	0.83

(a)   (b)

Fig. 3: (a) The RCA scores of the presence of each tag in each cluster; (b) The centroids of our clusters.

Section 5.1, the object set Ψ is composed by 402 artists, each one having a tag corresponding to her main music genre.

For each couple leader l and object ψ, we calculate Depth, Width and Strength values; we compute the size of the Leader's Minimum Diffusion Tree ($|T_{l,\psi}|$); and we group together the objects with the same tag. To characterize the typical values of Width, Depth and Strength for each tag we cannot use the average or the median. This is because Strength and Width values are skewed, and because it is the combination of the three measures that really characterizes the leaders. We decided to cluster leaders using as features its Width, Depth and Strength values. We used the Self-Organizing Map (SOM) method [12] for the following reasons: (i) SOM does not require to set the number of clusters k; (ii) k-means outperforms SOM only if the number of resulting clusters is very small (less than 7) [13], but our study of the best k to be used in k-means with the Sum of Squared Errors (SSE) methodology resulted in a optimal number of clusters falling in a range between 9 and 13; and (iii) SOM performs better if the data points are contained in a warped space [11], which is our case.

In Table 3(a), we report a presence score for each tag in each cluster. There are larger and smaller clusters and some tags attract more listeners than others. For these reasons, to report just the share of leaders with a given tag in a given cluster is not meaningful. We decided to correct this ratio with the expected number of leaders with the given tag in the cluster. This is a measure known as Revealed Comparative Advantage: $RCA(i,j) = \frac{freq_{i,j}}{freq_{i,*}} \Big/ \frac{freq_{*,j}}{freq_{*,*}}$, where i is a tag, j is a cluster, $freq_{i,j}$ is the number of leaders who spread an artist tagged with tag i that is present in cluster j. In Table 3(a), for each cluster we highlighted what is the tag with the highest unexpected presence.

The centroids of the SOM are depicted in Figure 3(b): Depth on the x-axis, Strength on the y-axis and the Width as the color (Strength and Width are in log scale). We can identify the clusters characterized by the highest and lowest Strength (9 and 4 respectively); by the highest and lowest Depth (2 and 9 respectively); and by the highest and lowest Width (11 and 1 respectively). There are also clusters with relatively high combinations of two measures: cluster 10 with high Strength and Width or cluster 5 with high Depth and Width.

From Table 3(a) we obtain a description of what values of Width, Depth and Strength are generally associated with each tag. For space constraints, we report only a handful of them for the clusters with extreme values. Jazz dominates clusters 1 (with the lowest Width) and 4 (with the lowest Strength): this fact suggests that jazz is a genre for which it is not easy to be prominent.

Cluster 9, with the lowest Depth but the highest Strength, is dominated by pop (that dominates also clusters 10 and 11, both with high Strength but low Depth). As a result, we can conclude that if prominent leader for pop artists then are embedded in users that are very engaged with the new artist. On the other hand, it is unlikely that they will be prominent among their friends too.

Finally, cluster 2 with the highest density has a large majority of punk leaders. From this evidence, we can conclude that punk is a genre that can achieve long chains of influence, exactly the opposite of the pop genre.

We move on to the topological characteristics of the leaders for each tag. There is one caveat: a leader is not bounded to be leader just for one object ψ, but she is free to be prominent in many ψ. Therefore, one leader can be counted in more than one tag. To help understand the magnitude of the issue, we depicted in Figure 4 the number of leaders influencing their neighbors for a given amount of actions (left) and for a given amount of tags (right). The y axis is logarithmic. The typical leader influences one neighbor for one artist. However, there is a certain amount of leaders expressing their leadership for at least 8 objects and 4 tags.
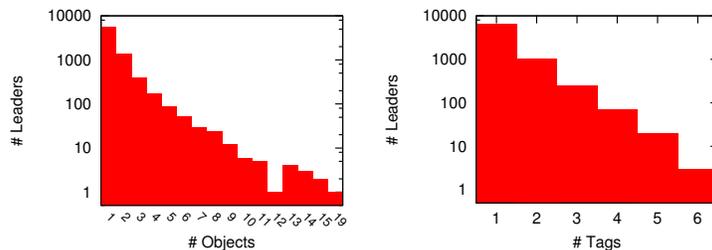

Fig. 4: Distribution of number of objects (left) and of tags (right) per leader.

In Figure 5 we depict the log-binned distributions, for the leaders of each tag, of four of the topological measures studied in Section 5.2: Degree, Closeness Centrality, Clustering and Neighbor Degree. We omit Betweenness Centrality for its very high correlation with Degree. Overall, there is no significant distinction between the tags in the distributions of the topological features.

The most noticeable information is carried by the Degree distributions (Figure 5, top left). Each distribution appears very different from the overall degree distribution (Figure 2 (left)). There are fewer leaders with low Degree than expected, therefore it appears that a high Degree increases the probability of being a leader. On the other hand, we know that central hubs have on average lower Depth and Width. As a consequence, it appears that the best leader candidates
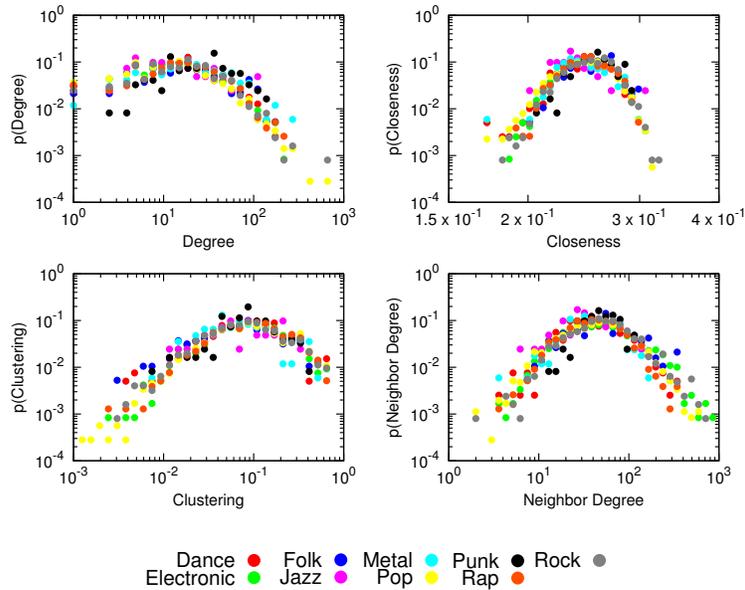
Fig. 5: Distribution of leaders' Degree (top left), Closeness Centrality (top right), Clustering (bottom left) and Neighbor Degree (bottom right) per tag.

are the nodes with an average degree, and from Figure 5 (top left) we see that each tag has many leaders with a Degree between 10 and 100.

Using our leaders' Minimum Diffusion Trees, we can extract some patterns that help us obtaining a complementary point of view over the leader prominence for different music genres. We mine a graph dataset composed by all diffusion trees $T_{l,\psi}$ with the VF2 algorithm [7]. Suppose we are interested in counting how frequent is the following star pattern: a leader influences three of its neighbors in the diffusion trees of pop artists. In our data, we have $5,043$ diffusion trees for pop artists, of which 581 have at least four nodes. Since the VF2 algorithm found the star pattern in 186 of these graphs, we say that it appears in 3.69% of the trees, or in 32.01% of the diffusion trees that have a sufficient number of nodes to contain it.

In Table 3 we report the results of mining three patterns of four nodes: i) the star-like pattern described above; ii) a chain where each node influences one neighbor; iii) a split where the leader influences a node, which itself influences two other neighbors. Two values are associated to each pattern and tag pair: the relative overall frequency, and the relative frequency considering only the trees with at least four nodes (in parentheses).

There is no necessary relation between the patterns and Width, Depth and Strength measures: a low Depth does not imply the absence of the chain pattern, nor does a high Width imply a high presence of the star pattern. However, the combination of the two measures may provide some insights. For instance, we saw in Table 3(a) that jazz leaders are concentrated in the lowest Width cluster.

Pattern	dance	electronic	folk	jazz	metal	pop	punk	rap	rock
	3.62% (35.42%)	3.04% (22.50%)	3.94% (30.30%)	7.25% (62.50%)	4.14% (23.08%)	3.69% (32.01%)	6.56% (27.59%)	4.01% (27.97%)	4.22% (30.43%)
	2.55% (25.00%)	3.92% (29.00%)	3.15% (24.24%)	4.35% (37.50%)	4.83% (26.92%)	3.61% (31.29%)	10.66% (44.83%)	5.60% (38.98%)	4.12% (29.71%)
	3.40% (33.33%)	3.79% (28.00%)	3.94% (30.30%)	4.35% (37.50%)	6.90% (38.46%)	4.73% (41.01%)	12.30% (51.72%)	4.99% (34.75%)	4.52% (32.61%)

Table 3: Presence of different diffusion patterns per tag.

However, many jazz leaders who affect at least three nodes tend to be prominent in their neighbors, much more than in any other genre (7.25% of all leaders, 62.5% of leaders who are prominent for at least three other nodes). Therefore, jazz leaders have low prominence among their friends, however they are likely to have at least three neighbors for which they are prominent.

The chain pattern is more commonly found in pop leaders than in folk ones, even though the clusters of their leaders described in Table 3(a) would suggest the opposite. It seems that pop leaders are not likely to be prominent for nodes any further than the third degree of separation, while folk leaders tend to generate longer cascade chains. Also in this case, punk leaders are commonly found in correspondence with chain patterns, just as Table 3(a) suggested.

Although pop leaders show a much greater Strength value than metal ones (by confronting in Table 3(a) their presence in high Strength clusters like 9 or 10 and low Strength clusters like 8 and 0), the split pattern tends to be more frequent in the metal genre (6.90% against 4.73% of the trees). This phenomenon suggests us that metal leaders tend to be prominent for nodes strongly devoted to metal, inducing them to spread the music to their neighbors. Pop leaders, on the other hand, affect more neighbors with higher Width and Strength, presumably flooding their ego networks with the songs they like.

6 Conclusion

In this paper, we presented a study of the propagation of behaviors in a social network. Instead of just studying cascade effects and the maximization of influence by a given starting seed, we decided to analyze three different dimensions: the prominence of a leader on how many neighbors, on how distant nodes and on how engaged nodes. We characterized each of this concept with a different measure: Width, Depth and Strength. We applied our leader detection algorithm to a real world network. Our results show that: (i) central hubs are usually incapable of having a strong effect in influencing the behavior of the entire network; (ii) there is a trade-off between how long the cascade chains are and how engaged each element of the chain is; (iii) to achieve maximum engagement it is better to target leaders in tightly connected communities, although for this last point we do not have conclusive evidence. We also included a case study in which we show how artists in different musical genres are spread through the network.

Many future developments are possible. The limited prominence that central hubs have on the overall network may be studied in conjunction with the problem of network controllability [14]. Alternative leader detection techniques, such as the ones presented in [9], can be confronted with our proposed algorithm. Finally, a deeper analysis of the properties of the Width, Depth and Strength measures can be performed, using additional techniques and exploiting data from other social media services like Twitter and Facebook.

References

1. Michele Berlingerio, Michele Coscia, and Fosca Giannotti. Mining the temporal dimension of the information propagation. In *IDA*, pages 237–248, 2009.
2. Ronald S Burt. Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92(6):1287–1335, 1987.
3. M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.
4. Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.
5. Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England Jou. of Medicine*, 358(21):2249–2258, 2008.
6. Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine*, 4(1):e13, 2007.
7. L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1367–1372, 2004.
8. J H Fowler and N A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj Clinical Research Ed.*, 337(2):a2338–a2338, 2008.
9. Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. Discovering leaders from community actions. In *CIKM*, pages 499–508, 2008.
10. W O Kermack and A G McKendrick. A contribution to the mathematical theory of epidemics. *The Royal Society of London Series A*, 115(772):700–721, 1927.
11. M. Y. Kiang and A. Kumar. A comparative analysis of an extended som network and k-means analysis. *Int. J. Know.-Based Intell. Eng. Syst.*, 8(1):9–15, 2004.
12. Teuvo Kohonen. The self-organizing map. *IEEE*, 78:1464–1480, 1990.
13. U.A. Kumar and Y. Dhamija. A comparative analysis of som neural network with k-means clustering algorithm. *Proceedings of IEEE International Conference on Management of Innovation and Technology*, pages 55–59, 2004.
14. Yang-Yu Liu, Jean-Jacques Slotine, and Albert-Laszlo Barabasi. Controllability of complex networks. *Nature*, 473(7346):167–173, May 2011.
15. Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.
16. P. Wang, M. C. González, C. A. Hidalgo, and A-L. Barabási. Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–1076, 2009.