

The patterns of musical influence on the Last.Fm social network

Diego Pennacchioli¹, Giulio Rossetti¹, Luca Pappalardo¹, Dino Pedreschi¹,
Fosca Giannotti¹, and Michele Coscia²

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`

Discussion Paper
original paper is [8]

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 study the patterns of musical influence through a social network. First, we define a procedure to extract musical leaders, i.e. users who start the diffusion of new music albums through the social network. Second, we measure three different dimensions of musical influence: the Width, i.e. the ratio of neighbors influenced by a leader; the Depth, i.e. the degrees of separation from a leader to its influenced nodes; and the Strength, i.e. the intensity of the influence from a leader. 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 or favoring the adoption of new technologies. In our paper, we present a study about the diffusion of music listenings on the Last.Fm music social network. In particular we analyze three relevant dimensions of social prominence: 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 validate our concepts on the social network from the music platform Last.Fm³, along with the data about how many times and when each user listens

³ <http://www.last.fm/>

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.

2 Related Work

Two phenomena are tightly linked to the concept of diffusion: the spread of biological [2] or computer [9] viruses, and the spread of ideas and innovation through social networks, the so-called “social contagion” [1]. Some models have been defined to understand the contagion dynamics: the SIR [5], SIS and SIRS [7] 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, [2] takes into account the spatio-temporal dimension. In [9], authors study the spreading patterns of a mobile virus outbreak. Christakis and Fowler studied the role of social prominence in the spread of happiness [4], finding that it can be contagious in a social sense.

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.

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 .*

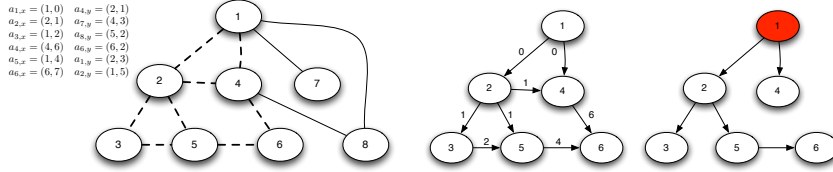


Fig. 1: Toy Example. (*left*) the social graph \mathcal{G} and action set \mathcal{A} . (*center*) the induced subgraph for the action x ; (*right*) the diffusion tree for x . In red the leader (root) for the given tree.

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 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. In the Figure, $a_{1,x}$ refers to the user 1 performing the action x ; and $a_{1,x} = (1,0)$ means that user 1 performed x one time, starting at the timestep 0.

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}_ψ 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}_ψ 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}_ψ : when no action is specified the set \mathcal{L} will be used to describe the union of all the \mathcal{L}_ψ for the graph \mathcal{G} . To be defined a *leader* an actor should not have any incoming edges in \mathcal{G}_ψ . Given this definition, for each directed connected component $\mathcal{C}_\psi \subset \mathcal{G}_\psi$ multiple nodes can belong to \mathcal{L}_ψ .

Realistically, a leader may be influenced by exogenous events. This is not a problem as we are not measuring 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 path of diffusion given an action a and a leader l we use a minimum diffusion tree:

Definition 3 (Leader's Minimum Diffusion Tree). *Given an action a_ψ , a directed connected component \mathcal{C}_ψ 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. Moreover, 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.

4 Measures

To defined three measures to capture social prominence: i) Width, the ratio of neighbors mirroring an action after a node; ii) Depth, how many degrees of separation are in between a node and the most distant of the nodes mirroring its actions; iii) 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 $width : \mathcal{L}_\psi \rightarrow [0, 1]$ is defined as:

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

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 $depth : T_{l,\psi} \rightarrow \mathbb{N}$ computes the length of the maximal path from l to a node $u \in T_{l,\psi}$. The function $depth_{avg} : T_{l,\psi} \rightarrow \mathbb{R}$ computes the average length of paths from l to any leaf of the tree.

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 $strength : T_{l,\psi} \times (0, 1) \rightarrow \mathbb{R}$ is defined as:

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

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

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

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.

5 Experiments

5.1 Data

Last.Fm is an online social network platform, where people can share their own music tastes and discover new artists and genres based on what they like. Users

send data about their own listenings. For each song, a user can express her preferences and add tags (e.g. genre of the song). Lastly, a user can add friends and search her neighbors w.r.t. musical tastes.

Using Last.Fm APIs⁴, we obtained a sample of the UK user graph, retrieving for each user: (a) her connections, (b) for each week in the time window from Jan-10 to Dec-11, the number of single listenings of a given artist. For each artist we have a list of tags, weighted with the number of users that assigned the tag to the artist. We referred such tags to words representing musical genres. Finally, we assigned each artist to the musical genre corresponding to their most popular tag. In the Last.Fm social graph \mathcal{G} each node is a user and each edge is generated using the user’s friends in the social media platform. The total amount of nodes is 75,969, with 389,639 edges connecting them.

Since we are interested in leaders, we need to focus only on whose first listening is recorded six months after the beginning of our observation period. We decided to focus on music genres with sufficient popularity, namely: dance, electronic, folk, jazz, metal, pop, punk, rap and rock. The cardinality of our action set \mathcal{A} is 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 neighbor to be prominent for her for that action.⁵

5.2 Diffusion patterns on the Last.FM social network

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. We cluster the leaders through the Self-Organizing Map (SOM) method [6], using as features their Width, Depth and Strength values.

In Table 2(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. To report just the share of leaders with a given tag in a given cluster is not meaningful. We correct the ratio with the expected number of leaders with the given tag in the cluster, a measure known as Revealed Comparative Advantage: $RCA(i, j) = \frac{freq_{i,j}}{freq_{i,*}} / \frac{freq_{*,j}}{freq_{*,*}}$, 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 . For each cluster we highlighted the tag with the highest unexpected presence.

The centroids of the SOM are depicted in Figure 2(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).

⁴ <http://www.last.fm/api/>

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

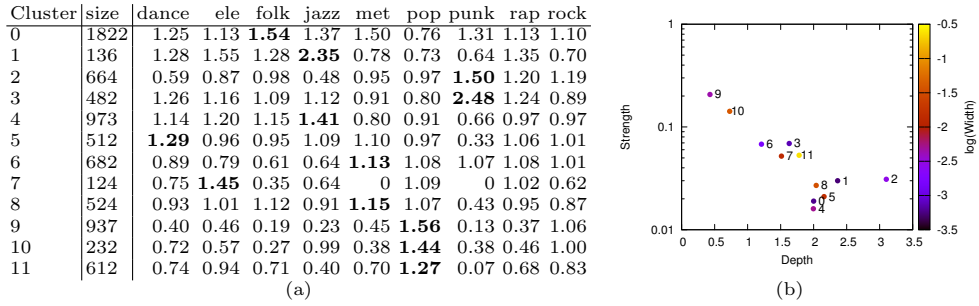


Fig. 2: (a) The *RCA* scores of the presence of each tag in each cluster; (b) The centroids of our clusters.

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 2(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 prominent leaders for pop artists are embedded in groups of users very engaged with the new artist. On the other hand, it is unlikely that these users 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 cascades, exactly the opposite of the pop genre.

We move on to the topological characteristics of the leaders per tag. In Figure 3 we depict the log-binned distributions, for the leaders of each tag, of its degree.

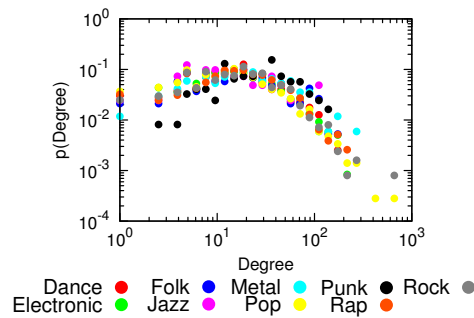


Fig. 3: Distribution of leaders' Degree per tag.

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 are the nodes with an average degree, and from Figure 3 we see that each tag has many leaders with a Degree between 10 and 100.

Using our leaders’ Minimum Diffusion Trees, we 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 [3]. 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, and 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 trees that have enough nodes to contain it.

In Table 1 we report the results of mining three patterns of four nodes: i) the star-like pattern described above; ii) a chain where each node is prominent for (at least) one neighbor; iii) a split where the leader is prominent for a node, which itself is prominent for 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).




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 1: Presence of different diffusion patterns per tag.

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 2(a) that jazz leaders are concentrated in the lowest Width cluster. 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 2(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 2(a) suggested.

Although pop leaders show a much greater Strength value than metal ones (by confronting in Table 2(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 musical listenings in the Last.Fm social network. We analyzed three different dimensions: the prominence of a leader on how many neighbors (Width), on how distant nodes (Depth) and on how engaged nodes (Strength). The results of our leader detection algorithm to the Last.Fm network 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.

References

1. Ronald S Burt. Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, 92(6):1287–1335, 1987.
2. V. Colizza, A. Barrat, M. Barthélemy, A.J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine*, 4(1):e13, 2007.
3. 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.
4. 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.
5. 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.
6. T. Kohonen. The self-organizing map. *IEEE*, 78:1464–1480, 1990.
7. R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.
8. D. Pennachioni, G. Rossetti, L. Pappalardo, M. Coscia, D. Pedreschi, and F. Gianotti. The three dimensions of social prominence. *SocInfo*, 2013.
9. 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.