To Be a Star Is Not Only Metaphoric: From Popularity to Social Linkage

Alina Stoica
Orange Labs and Liafa
Paris, France
stoica@liafa.jussieu.fr

Thomas Couronne
Orange Labs
Paris, France

Jean-Samuel Beuscart
Orange Labs
Paris, France

{thomas.couronne, jeansamuel.beuscart}@orange-ftgroup.com

Abstract

The emergence of online platforms allowing to mix self publishing activities and social networking offers new possibilities for building online reputation and visibility. In this paper we present a method to analyze the online popularity that takes into consideration both the success of the published content and the social network topology. First, we adapt the Kohonen self organizing maps in order to cluster the users of online platforms depending on their audience and authority characteristics. Then, we perform a detailed analysis of the manner nodes are organized in the social network. Finally, we study the relationship between the network local structure around each node and the corresponding user's popularity. We apply this method to the MySpace music social network. We observe that the most popular artists are centers of star shaped social structures and that it exists a fraction of artists who are involved in community and social activity dynamics independently of their popularity. This method based on a learning algorithm and on network analysis appears to be a robust and intuitive technique for a rich description of the online behavior.

1. Introduction

In recent years, new digital practices have emerged, that combine self-publishing with social networking. Many Web services have encountered success by combining the publishing of user generated content with social networking tools: on blogs, video-sharing platforms, photo-sharing sites, music streaming platforms, social networking sites, users at the same time display the contents they have produced, and favor the circulation of these contents through the social networking tools. On these platforms, each user is able to manage his own visibility: thanks to increasingly precise tracking tools, he knows how many people viewed, commented, recommended, rated, and forwarded his work. These ratings, by increasing the user's reflexivity about his popularity, strongly influence publishing and networking practices (Halavais 2008), (Huberman, Romero, and Wu 2008), leading some authors to describe the Web as a huge space of competition for popularity (Wazik 2009).

Several researches have dealt with this competition for visibility and reputation in large social networks. On the one hand, some works concentrate on the success of contents. They show that, contrary to Anderson's intuition of a "long tail" (Anderson 2006), the audience is often concentrated on a minority of contents: for example, that 20% of videos on YouTube receive 80% of views (Cha et al. 2007); we have similar results on MySpace Music. They also explore the temporality of success, showing that the final audience of a video on Youtube can be inferred for its audience after seven days (Szabó and Huberman 2008). On Flickr, Cha et al. examine different patterns of success for the photos (Cha, Mislove, and Gummadi 2009). On the other hand, some researches have focused on the reputation of individuals in the large social networks created by these practices. Since the seminal work of Herring et al. (Herring et al. 2005), we know that influent bloggers are at the center of the social network, and that bloggers tend to link to bloggers of equal or superior reputation. This phenomenon of preferential attachment - people tend to link to individuals who already receive a lot of links, also known as the "Matthew effect" (Merton 1968), has also been observed on other publishing platforms such as MySpace or Flickr (Mislove et al. 2008). Few works, though, explore the relation between the producer's networking activity and the success of his contents.

In the present paper, we try to hold together the two approaches; we study the popularity of MySpace artists in relation with the local structure of the social network surrounding the artists. First we provide a methodology for assessing the relation between the popularity of users' content and the structure of their social network. Then we apply this methodology to a sample of the MySpace music network.

In a previous work about MySpace Music (Beuscart and Couronne 2009), we observed that the online popularity has two main dimensions: the audience of the contents (number of visits of the artist's page) and the user's authority, which reflects the number of people recommending the artist by linking to him; in other words, the popularity of the content is not strictly correlated with the recognition of the artists by their peers and fans. Our new methodology allows us to go beyond this statement, and to identify 5 distinct patterns of popularity on MySpace, described in terms of audience, recognition, and social structure.

First, we build a popularity typology based on artists audience and authority, thus revealing the macroscopic patterns of the online reputation. We employ the Kohonen self orga-

nizing map (Kohonen 1990) as a robust and pertinent classifier of individuals based on the popularity variables. This is a standard technique of classification that takes into account the non linear effects engendered by the mixed practices of online social platforms. Second, continuing a study on the local structure of social networks that we began in (Stoica and Prieur 2009), we analyze how the different artists are connected to each other depending on their audience and authority. For that, we analyze the egocentred network of each node and we characterize the way the node is connected to the network (by computing the patterns in its egocentred network) and the way its links are distributed (by computing the positions of its neighbors). We thus obtain a rich description of the structure of the network in which each node is embedded, that we confront to the online popularity of the artist. This paper confirms the strong relevance of studying the local network structure regarding to the popularity variables and provides a set of methods for an efficient analysis of this connection.

The paper is organized as follows. First, we describe our two methodological tools: the algorithms used to analyze the social network structures, and the Kohonen classifier. Then we employ the Kohonen self-organizing map to build a categorization of MySpace Music artists depending on their popularity variables; we obtain 5 classes. Next, we characterize the local structure of the network surrounding the artists. We finally show that the various forms of friendship links differ from one class to another, and that popularity is linked to the way artists are inserted in the network.

2. Methods and Definitions

This section describes formally the methods applied to the MySpace artists data. The first part concerns the local analysis of the social network and the second part the Kohonen classifier.

2.1 Preliminaries

Let $G = (V, E)$ be a graph; V is the set of its vertices, $E \subseteq V \times V$ is the set of its edges. The graph G is *undirected* if, for all $u, v \in V$, there is no difference between (u, v) and (v, u), it is *connected* if there exists a finite path between every two vertices and it is *simple* if there is no multiple edge and no self-loop ($(v, v) \notin E$, for all $v \in V$). Given a vertex $v \in V$, a vertex $u \in V$ is a *neighbor* of v if and only if $(u, v) \in E$. The number of neighbors of v represents its *degree*. Two graphs $G = (V, E)$ and $H = (V', E')$ are *isomorphic* if and only if there exists a bijective function $\varphi : V \to V'$ such that, for any two vertices u and v in V, $(u, v) \in E$ if and only if $(\varphi(u), \varphi(v)) \in E'$. The subgraph *induced* by a set of vertices $V' \subseteq V$ in G is the graph $H = (V', E')$ with $E' = \{((u, v) \in E \mid u, v \in V')\}$.

Patterns and positions of vertices. We call *patterns* the 9 non-isomorphic undirected connected graphs with at most 4 vertices and at least 1 edge (Figure 1). In (Stoica and Prieur 2009) we have introduced the notion of *position* in a pattern: two vertices of a pattern P are said to occupy the same position in P if and only if one can interchange them without modifying the pattern P (so the two vertices are automorphically equivalent). In Figure 1, for each pattern, the vertices

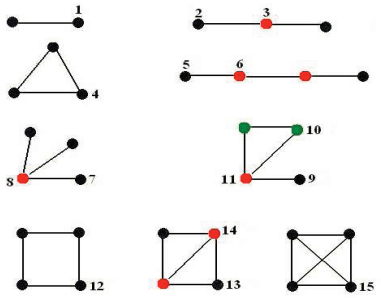

Figure 1: The 9 patterns and the different positions; in a pattern, each color corresponds to a different position

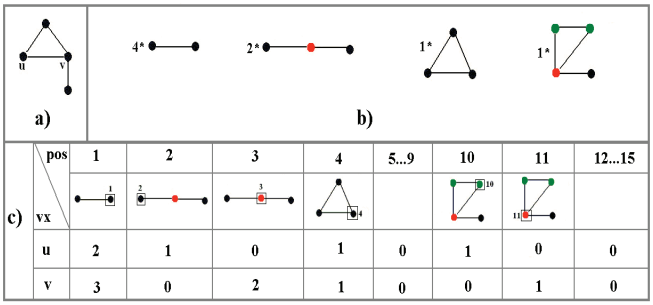

Figure 2: A graph (a), its patterns (b) and the position vectors of two vertices u and v.

that have the same color occupy the same position in that pattern. There are 15 different positions in the 9 patterns. We say that a pattern P appears in a graph $G = (V, E)$ if there exists a set of vertices $V_P \subseteq V$ such that the subgraph induced by V_P in G is isomorphic to P. Listing all the occurrences of the pattern P in the graph G means finding all the sets of vertices V_P according to the previous definition. For each occurrence of a pattern in $G = (V, E)$ one can compute in which position of the pattern the different vertices of V are placed. Thus, after having listed all the occurrences of the 9 patterns in G, one has, for each vertex $v \in V$, its number of occurrences in each one of the 15 positions (we call this the *position vector* of v). As an example, Figure 2 represents a graph (a), the patterns it contains and their number of occurrences (b) and the number of occurrences in the 15 positions of two selected vertices (c).

2.2 Analysis of the local structure of a network

Suppose that we are given a network that represents a set of individuals and some connections between them. We want to study how the different individuals are connected to each other. Therefore, we analyze the local structure of the network, around each node (so each individual), in order to describe how the node and its links are connected to the network.

Formally, let $G = (V, E)$ be a simple undirected graph such that V corresponds to the set of individuals and E to the set of connections between them: two vertices u and v are connected by an edge (u, v) if there is a connection

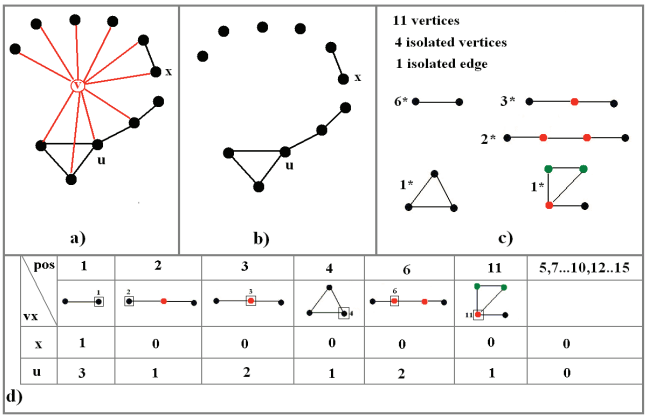

Figure 3: A vertex v and its neighbors (a), the egocentred network $Eg(v)$ of v (b), the patterns of $Eg(v)$ (c) and the position vectors of two neighbors of v (d).

between the two individuals u and v (from u to v or from v to u or both). In order to analyze the local structure of $G = (V, E)$ around a vertex $v \in V$ we proceed as follows (method ***local_structure(v)***):

Step 1. Extract the egocentred network $Eg(v)$ of v i.e. the subgraph induced by the neighbors of v in G;

Step 2. List the patterns of $Eg(v)$;

Step 3. Compute the position vectors of the vertices in $Eg(v)$.

Let us explain the three steps of the method with an example. In Figure 3(a), the black circles correspond to the neighbors of v, the black lines correspond to the edges between them and the red lines to the edges between v and its neighbors. The egocentred network $Eg(v)$ of v is represented in Figure 3(b) and the patterns of $Eg(v)$ in Figure 3(c) (we have also counted the number of isolated vertices and edges in $Eg(v)$). We chose not to include v in its egocentred network because we know that it is connected to all the vertices in this graph, its presence doesn't bring any information. After performing the steps 1 and 2 of the method one has a rich description of the way v is connected to the graph G. For a more detailed description of the local structure of G around v one can list the patterns of a higher order (with 5 vertices or more); the patterns with 4 vertices are however a good compromise between the variety of forms and their number, providing, in many cases, a detailed enough picture.

Step 3. We compute the position vectors of the neighbors of v, so the number of times each neighbor appears in each one of the positions of the different patterns. Figure 3(d) contains the position-vectors of two neighbors of v. The positions occupied by the different neighbors characterize the edges formed by v. As an example, Figure 4 presents the correspondence between three possible positions of a neighbor u and the structure of the graph around the edge (u, v).

Suppose now that the connections between the observed individuals are directed (i.e. a connection from u to v doesn't necessarily imply a connection from v to u). We can

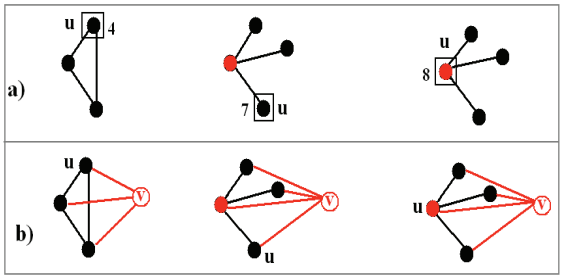

Figure 4: Three possible positions of the neighbor u (a) and the corresponding structures around the edge (u, v) (b).

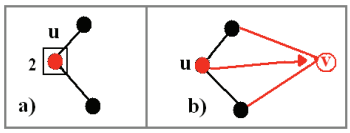

Figure 5: A position of the neighbor u with weight 2 (a) and the corresponding structure around the edge (u, v) (b).

add this information to the description of the edges formed by v by simply adding a weight to the neighbors of v. For a node v, the weight $w_v(u)$ of a neighbor u is:

- 1 if the connection is from v to u ($v \rightarrow u$),
- 2 if the connection is from u to v ($u \rightarrow v$),
- 3 if the connection is symmetric ($v \rightarrow u$ and $u \rightarrow v$).

As an example, Figure 5 presents the correspondence between a possible position of a neighbor u that has weight 2 and the structure of the graph around the edge (u, v).

Note that this method provides a description of the structure of the graph around a given vertex, therefore it can be applied to the set of all the vertices of the graph or only to some of them. As it is a local method, one doesn't need to have all the vertices and edges in the graph, but only the neighbors of each studied vertex and the edges between them.

Complexity of the method. In (Stoica and Prieur 2009) we proposed an algorithm for the computation of all the patterns and position vectors of a given graph whose time complexity is linear in the number of patterns of the graph. For a vertex with degree d, there are at most d^4 patterns in its egocentred network, so the time complexity of *local_structure(v)* is at most $O(d^4)$. Given that the degree distribution of real networks is generally heterogeneous, with the majority of nodes having a small degree, the method is in average very fast.

2.3 Kohonen self organizing map and multidimensional clustering

The Kohonen self organizing map is an efficient tool to divide a population into clusters based on some a priori characteristics of the individuals. A Kohonen self organizing map is the result of a non supervised learning algorithm. The aim is to learn the characteristics of the population to

be clustered and to build a bi-dimensional map where the individuals are placed depending on their topological proximity. The map's smallest entity is a cell, and each individual is placed in only one cell (a local area). The map is composed by a set of $m \times p$ cells (organized in a bi-dimensional surface). This version of the algorithm predefines the cells number as $\sqrt{|k|}$ (consequently $m \times p = \sqrt{k}$) where k is the population to be clustered. It differs from the "Growing self organizing map" (D. Alahakoon and Sirinivasan 1998) in which the number of cells varies depending on a overlapped structure.

Each cell is characterized by a vector of $n-$dimensions; in the algorithm we employ, the cells have an hexagonal shape, therefore surrounded by six neighbors. Each individual i of the population k to be clustered is characterized by a feature vector F_i of dimension n, where $F_i(t)$ is the value of the $t-$th variable among the n variables characterizing the individual. Two individuals with an identical feature vector will be associated with the same cell and the ones with opposed feature vector will have a topologically opposed position on the map.

The method is composed by three steps. The first one is the learning. The n dimensions of the $m * p$ cells are randomly initialized. Then a subset of the population to model is randomly selected; for each individual in this selection the SOM finds the cell ("winner") whose feature vector is the most similar. The feature vector of the winner cell is updated to take into account the feature values of the individual. The feature vector of the neighbor cells are then modified to reduce the vectors gradient with the new values of the cells' feature vector. The second step of the algorithm is the processing of the global population to model. Finally the last step is the clustering of the cells with, for instance, a k-means algorithm, based on the similarity of their feature vectors.

3. MySpace Music popularity and social network analysis

3.1 Data construction

Following two precedent works (Stoica and Prieur 2009), (Beuscart and Couronne 2009), we study the MySpace music social network for a better understanding of the recommendations. Our goal is to reveal the relations between the online popularity and the network micro structure characteristics.

We build a sample of the MySpace music (artistic) population based on the best friendship declaration links. After having chosen seven initial parent artists profiles among the French MySpace music top audience, a breadth-first-search crawler is employed to collect the profiles information, following the best friendship links during 3 iterations (best friend of best friend of best friend of the parents). The number of best friendship links varies from 1 to 40. This kind of crawling is known to produce a sample with a relevant structure (good fitting of the clustering, density, and centrality values) but underestimates the in-degree and overestimates the out-degree (Mislove et al. 2007), (Kumar, Novak, and Tomkins 2006).

Table 1: Dataset properties

Total number of profiles	21153
Artists profiles	13936
Total number of links	143831
Number of links between artists	83201
Reciprocal links rate (A and B have declared each other as best-friends)	40.1%
"Major" labeled artists	3422
"Indie" labeled artists	7069
"without" labeled artists	3445

In order to verify that this sample is not unusual, we collect several networks varying the initial artists numbers (from 3 to 10), the parsing depth (from 2 to 4), the initial artists nationality and the collected artists via a randomized ID selection. If the total number of nodes and the music profiles proportion depend on the crawling parameters (on MySpace the account's profile may be defined as "member" or "musician") , the ratio of the two is around 50%. Next, for each sample, a correlation test is applied between the followings four quantitative variables: number of comments, of friends, of profile visits (hits) and best-friendship declaration. A Mantel test is performed between the correlation tables; it shows that the coefficients are significantly similar, i.e. the variables of each sample are correlated in the same proportions.

As we are interested in the MySpace music profiles, we chose to remove from the data all the non-artistic individuals. The properties of the studied network sample are summarized in Table 1.

We analyze the popularity of the MySpace artists using the two previously presented methods: self organizing map and local structure of the network.

3.2 Kohonen Self Organizing Map construction

The following variables are chosen to model the popularity characteristics for each artist and to construct the feature vector used for the SOM clustering:

- Number of visits of the profile (hits)
- Number of comments visitors have left on the profile (these first two characteristics are an indicator for the artist's *audience*)
- Number of people having declared the artist as best friend (this is a measure of the artist's *global authority*)
- Number of artists having declared her or him as best friend (the *artistic authority*)
- Fraction of the artist's best friends who have declared her or him as best friend (reciprocity rate, a measure of the cooperative behavior)
- Label: whether the artist record label is declared as "Major", "Indie", or "Other"

The number of visits, comments and best-friendship declaration are heavily right-skewed so a log transformation is used

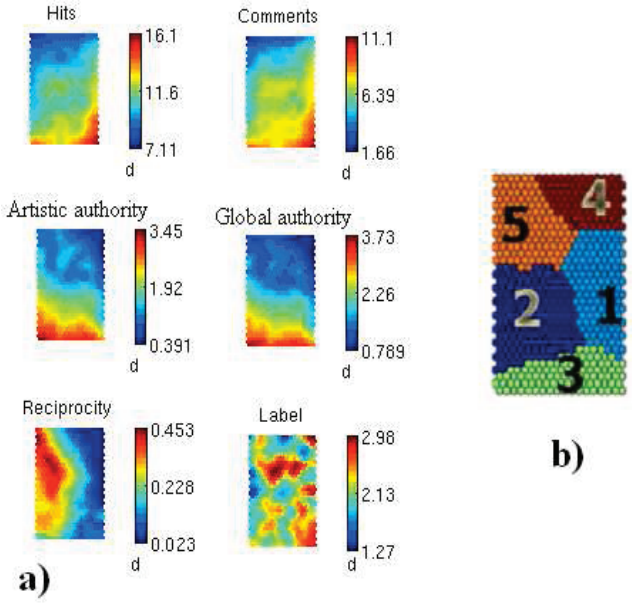

Figure 6: Self Organizing map of the artists depending on their popularity properties (a) and the 5 clusters (b)

instead of the value itself for these variables. Each artist of the sample is thus characterized by a feature vector of 6 variables.

The multi-dimensional processing of the set of individuals by the SOM provides Figure $6(a)$. The SOM map is a bi-dimensional set of cells, where each individual belongs to only one cell and each cell contains several individuals. Each cell has a feature vector (the vector of the six variables) computed from the feature vectors of the individuals. The maps is visualized for each one of the six variables, thus generating six layers. On each layer, the color of the cell indexes the value of the corresponding variable for that cell.

The map appears to be structured by two independent trends: The more an artist belongs to a southern cell, the more her or his popularity is high, in terms of both audience and authority; And the more an artist is to the west side, the more she or he tends to have reciprocal links. If audience and authority are partly correlated and discriminate popular artists from anonymous, the trends are not exactly similar. Indeed, the south-western area is associated with the authoritative elites (highest artistic and global authority) and the south-eastern area is associated with the most notorious artists (highest page views and comments). If, logically, the audience elites are not without authority and authoritative elites are not without audience, the top artists of the audience and of the authority do not overlap.

We can note that the two measures of authority (global and artistic) are correlated. The artists and the fans create in the same way their best friendship links: the authority hierarchy follows a unique trend. Complementary, this result shows that the reciprocal links behavior is not associated with the popularity: it may be either because an authoritative artist cannot have more than 40 best friends (and therefore cannot cite everybody) or because very authoritative artists are not linking back to people who link to them (fan-star relationship). Finally we observe that the south-east area (audience elites) is associated with a strong presence of the "Major" labels.

The cells produced by the SOM are organized into clusters using a $k-$means clustering. The expectation maximization algorithm is then employed to choose the best number of clusters. The population is thus distributed into 5 clusters (Figure $6(b)$):

Cluster1 (Cyan, population: 2732) gathers artists with a medium-to-large audience, a low authority and a weak reciprocity rate. They are mostly associated with major music labels. Our browsing of the Myspace pages of some artists in this cluster suggests that these artists, already popular offline, use their MySpace page as a display window of their music, but make very little use of the social networking tools. We may suppose that their strong audience comes from their offline popularity, but that they are not active enough to gain a strong influence on MySpace.

Cluster2 (Dark blue, pop.: 3036) gathers artists with a very strong authority, and a medium-to-high audience: these artists are not the most popular, but they are the most recommended. Most of them belong to independent labels. The qualitative browsing of their pages suggests a very intensive use of the social networking tools in order to build their online popularity. Here we find a lot of trendy groups and electronic avant-garde music, waiting for their small online fame to become larger.

Cluster3 (Green, pop.: 1920) gathers artists with both a large audience and a strong authority, the MySpace elites. They have mostly major labels. Browsing their pages, we find established artists, combining traditional forms of artistic accomplishment (famous labels, presence in renowned festivals) with an active online marketing strategy.

Cluster4 (Brown, pop.: 2834) gathers artists with a very small audience and no authority. Most of their pages display very low activity, suggesting that these artists have either abandoned the page or show very little interest in socializing practices.

Cluster5 (Orange, pop.: 2834) gathers artists with a small audience, low authority, and a strong reciprocity rate. Most of them are unsigned. On the contrary to artists from cluster 4, most of the pages we browsed are very active. These small amateur artists seem to be the ones populating the local music scenes; they are well connected to other artists from the same scene or from the same geographical area. Their small audience may not reflect their inability to reach an audience, but the small size of their musical or geographical niche.

The first part of this study provides an artists classification based on the popularity variables. The main results are that both dimensions audience and authority are correlated but discriminate at least two elites. Moreover the best friendship links appear to have various significance, fan - star, peers etc. Therefore it is relevant to study more specifically what the

links distribution and network structure teach us about the best friendship significance and the artistic popularity.

3.3 The local structure of the network as a function of the artists' popularity

In this section we analyze the local structure of the social network of MySpace artists in order to see if it is different depending on the popularity cluster of the artists. We represent the sample of MySpace artists and their best-friendship declarations as a simple undirected graph where the vertices correspond to the artist profiles and the edges to the existence of a best-friendship declaration between two artists: there is an edge between the vertices u and v if u has declared v as best-friend or v has declared u as best-friend or both. The resulting graph has 13936 vertices and 65979 edges. In order to describe the local structure of the graph, around each vertex, we apply the method *local_structure* to all the vertices of the graph. It takes 34 seconds to run our implementation of the method for all the vertices on a computer with a 2.8GHz processor and 4Gb RAM.

VERTICES. We begin by studying the structure of the graph surrounding the vertices in order to see if it differs depending on the SOM popularity cluster the vertices belong to. For this, we use the number of patterns in the egocentred networks (computed by steps 1 and 2 of the method *local_structure*). We want to compare the number of occurrences of the 9 patterns in the egocentred networks with respect to the popularity clusters of the vertices. As the degree distributions are not the same in the 5 clusters, one cannot simply compare the number of occurrences of the patterns; these quantities are biased by the degrees of the vertices (for instance, a vertex with a high degree probably has high values for all the patterns). Therefore, we compare the number of occurrences of patterns in the egocentred networks of the vertices with the same degree (i.e. the same number of vertices in the egocentred network). For each cluster C, each degree d and each pattern P, we compute the average $FD(C, d, P)$ of the number of occurrences of the pattern P in the egocentred networks of the vertices with degree d in C (we take into consideration only the degrees for which there are at least 2 clusters where 1% of the nodes have that degree). Figure 7 represents, for each degree d, the values of $FD(C, d, P)$ for the 5 popularity clusters; the considered pattern is the number of edges in the egocentred network in Figure $7(a)$ and the number of isolated vertices in the egocentred network in Figure $7(b)$.

We observe that, for all the degrees, the vertices of the cluster 5 have the greatest number of edges in their egocentred networks, followed by those of the clusters 2, 1 and 4 and finally 3. The order is inverted for the number of isolated vertices that measures the quality of "star" of a vertex. Remember that clusters 5 and 2 are the ones on the western side of the SOM map, i.e. artists having reciprocal links, sometimes a lot of friends, but a medium to small popularity: they can be authoritative, but not strongly popular. Cluster 3, situated in the southern part of the map, contains the MySpace elite, the superstars, the popular authoritative artists. These vertices are, in terms of network structure, star centers, connecting many unlinked vertices, as Figure $7(b)$ shows.

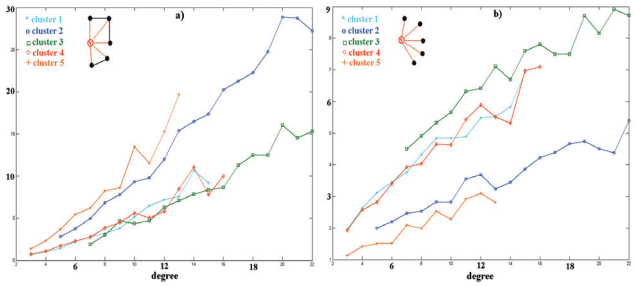

Figure 7: For the vertices of each cluster, the average number of edges (a) and isolated vertices (b) in the egocentred networks as a function of the degree

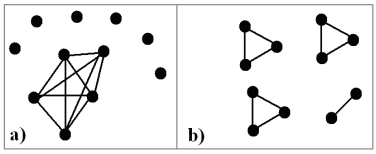

Figure 8: Two egocentred networks that have the same number of vertices, of edges and the same clustering coefficient

These observations could have been done also by computing the density of the different egocentred networks (or the local clustering coefficient). The listing of patterns in the egocentred networks provides however a richer description of the local structure of the network than these two measures. It describes *how* the edges formed by the neighbors of a given vertex are disposed, what type of structures they form. For instance, the two egocentred networks in Figure 8 have the same number of vertices, of edges (so the same density) and the same clustering coefficient. These measures do not capture the differences between these two graphs, but the listing of patterns does.

We continue our analysis by computing, for each cluster C, each value e of the number of edges in the egocentred network, and each pattern P, the average $FE(C, e, P)$ of the number of occurrences of the pattern P in the egocentred networks with e edges of the vertices in C (as before, we take into consideration only the values e reached by at least 1% of the nodes, in at least 2 clusters). Figure 9 represents, for each value e of the number of edges in the egocentred network, the values of $FE(C, e, P)$ for 5 popularity clusters; the considered pattern is the number of isolated edges (Figure $9(a)$), the number of triangles (Figure $9(b)$) and the number of $4-$cliques (Figure $9(c)$) in the egocentred network.

We observe that, given a value of the number of edges in the egocentred network, these edges are more likely to be found in triangles and $4-$cliques for cluster 5 than for clusters 2, 1 and 4. The vertices in cluster 3 have the lowest probability to have triangles and $4-$cliques in their egocentred networks. The edges between the neighbors of these vertices are often isolated (Figure $9(a)$), confirming the character of "star" of the vertices in cluster 3.

As for the other patterns, pattern 8 (the two triangles) has

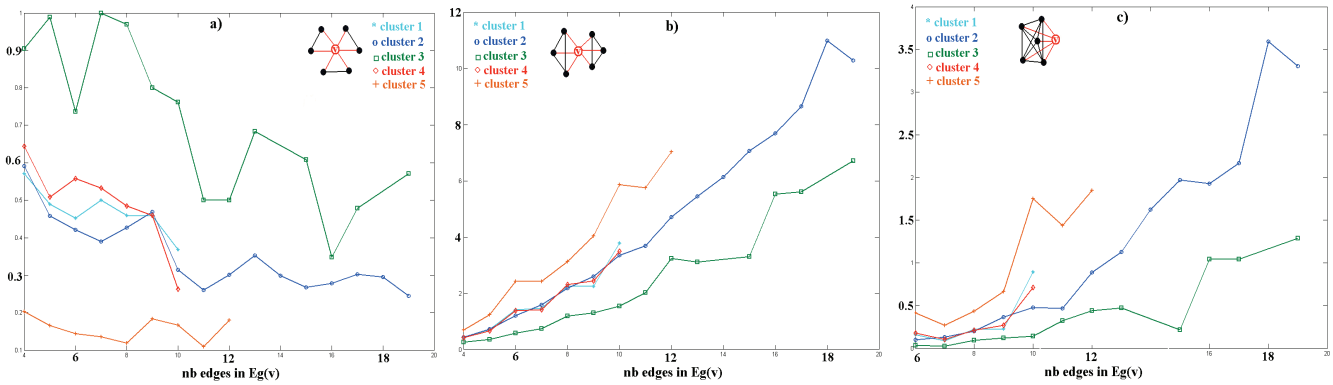

Figure 9: For the vertices of each cluster, the average number of isolated edges (a), triangles (b) and $4-$cliques (c) in the egocentred networks as a function of the number of edges

the same order as the $4-$clique, showing, once again, the tendencies of vertices in cluster 5 to belong to dense groups and that of vertices in cluster 3 to be centers of stars. The other patterns don't present a clear order; however, for pattern 5 (the star), clusters 3 and 4 have the highest probabilities to contain this pattern in their egocentred networks and for pattern 7 (the square), it is cluster 1 that has the highest one. So, even if the number of edges in the egocentred network is the same, the structures in which these edges are placed are different for the 5 clusters, going from dense groups for the clusters 5 to sparse groups for the cluster 3.

To sum up, the social network surrounding each artist differs, depending on her or his popularity. The most popular artists (cluster 3) are at the center of stars; heterogeneous artists, not connected to each other, connect to them due to their popularity. As for artists with a medium-to-large audience, they have distinct types of insertion in the network: those in cluster 2 are inserted in dense recommendation networks, usually describing homogeneous musical universes, while those in cluster 1 belong to sparse structures. The same observation can be made for artists with a small audience: artists from cluster 5, though not very popular, are involved in dense structures, unlike artists from cluster 4 who display disconnected links. This analysis strengthens our typology, by associating types of popularity with types of insertion in the recommendation network.

EDGES. We continue our analysis with the study of the edges formed by the vertices in the 5 popularity clusters. We want to see, for the vertices of each cluster, with which clusters they form the most of their edges and how these edges are placed in the graph. For that, we use the positions occupied by the neighbors in the egocentred network of the different vertices (computed in the step 3 of the method *local_structure*). This way, we know for each neighbor u of a vertex v how many times it occurs in each one of the 15 possible positions (Figure1) in the egocentred network of v. As the best-friendship links are directed, we add this information as weights of neighbors: for a vertex v, a neighbor u has weight 1 if v has declared u as a best-friend but u hasn't, weight 2 if u has declared v but v hasn't and weight 3 if the best-friendship declaration is mutual.

Let $Pos_v(u, Ps)$ be the number of occurrences of a neighbor u of v in the position Ps in the egocentred network of v. For each cluster K we compute the probability $Pr_K(w, C, Ps)$ to observe a vertex with weight w of the cluster C in the position Ps in the egocentred networks of the vertices in K :

$$Pr_K(w, C, Ps) = \frac{\sum_{v \in K} \sum_{u \in Eg(v), u \in C, w} Pos_v(u, Ps)}{\sum_{v \in K} \sum_{u \in Eg(v)} Pos_v(u, Ps)}.$$

We observe that:
1) for the clusters **1** and **4**, for all the 15 positions Ps, $Pr_{1,4}(w, C, Ps)$ is maximal when $C = 3$ and $w = 1$ (best-friendship links from 1 / 4 to 3);
2) for the cluster **2**, for all the positions Ps, $Pr_2(w, C, Ps)$ is maximal when $C = 2$ and $w = 3$ (mutual best-friendship links);
3) for the cluster **3**, for all the "important" positions i.e. $Ps \in \{1, 3, 4, 6, 8, 10, ...15\}$, $Pr_3(w, C, Ps)$ is maximal when $C = 3$ and $w = 3$ and for all the "peripheral" positions i.e. $Ps \in \{2, 5, 7, 9\}$, $Pr_3(w, C, Ps)$ is maximal when $C = 4$ and $w = 2$ (best-friendship links from 4 to 3) ;
4) for the cluster **5**, for all the positions with a high degree i.e. $Ps \in \{4, 8, 10, 11, 14, 15\}$, $Pr_5(w, C, Ps)$ is maximal when $C = 2$, followed by $C = 5$, and $w = 3$ (mutual links between 2 and 5 or inside the cluster 5); for all the other positions, $Pr_5(w, C, Ps)$ is maximal when $C = 3$ and $w = 1$ (best-friendship links from 5 to 3).

So, if one randomly picks an edge formed by a vertex of the cluster 1 or 4, no matter the structure of the graph in which this edge is embedded, it is very probable that this edge is an out-going arc to the cluster 3. It is a star-fan relation that confirms the character of "star" of the vertices in the cluster 3 and the weak authority of the clusters 1 and 4. The cluster 2, grouping artists with high (but smaller than the stars') authority and popularity, connects mostly to itself. It is also the case of the cluster 3, whose edges are placed in "important" positions when they are formed inside the cluster and in "peripheral" positions when they are in-coming arcs. The important positions (as, for instance, the center of a star) signify that the vertices of the clus-

ter 3 often form a central axis to which many triangles are connected i.e. many vertices, not connected to each other, connect to two linked vertices of the cluster 3. This may correspond to two popular artists of a similar music genre, where people who like the first are highly probable to listen the second too. As for the cluster 5, remember that it has a high reciprocity of links. The vertices in the cluster 5 share symmetric edges especially with the vertices in the cluster 2 and with themselves; these edges are often placed in dense groups (cliques, maybe with few missing edges), as the positions $\{4, 8, 10, 11, 14, 15\}$ show. We observe also a fan-star relation of the vertices in the cluster 5 towards the vertices in the cluster 3 (the other positions). The edges with the cluster 3 are directed towards this cluster and are placed in "peripheral" positions (for instance, the position 7 corresponds to the connection of the edge to a central axis, the position 9 to the connection to a clique etc.).

4. Discussion

The method we have introduced here provides a rich description of the popularity of the users of an online social network. Two dimensions are compared: the online popularity of the users and their connectivity in the social network.

When applied to the MySpace network, the method reveals in a robust and efficient way that the best friendship links on MySpace wear various meanings, creating multiple popularity patterns. Next to unsurprising categories (clusters 3 and 4, very popular artists and unknown artists), we identify two different kinds of mid-range popularity (clusters 1 and 2), and a category of small but socially active artists (cluster 5). We show that artists in these categories exhibit different insertions in the social network. Artists with a low authority and non reciprocal links tend to declare very popular artists as best friend thus generating a star structure. On the contrary, some mid-range and low popularity artists form small cliques with local neighbours, creating communities without stars but with triangles.

The self organizing map, providing a visual result, appears to be strongly pertinent for the study of sociological multivariate data integrating non linear effects. In addition, the computation of patterns and positions of vertices in ego-centred networks seems a good way to characterize the local structure of the social linkage. When put together, theses methods unfold a rich and intuitive set of meaningful information.

To go further, the prediction of one axis given the other one could be performed (i.e. the online popularity given the social linkage of the vertex and vice-versa). The method should integrate the dynamic analysis of the popularity and the social structure, for instance the variations of the popularity given the initial (and the intermediate) network structures.

This method can be easily applied to any social network where the corresponding graph can be built and the activity of the users can be measured. An immediate transposition is feasible to the Flickr and YouTube platforms, where the popularity is defined by the same parameters as on MySpace. Even more, the analysis can be adapted to some offline social networks as the mobile phone using calls frequencies, durations and contacts.

References

Anderson, D. 2006. *The Long Tail: How the Future of Business is Selling Less of More*. Hyperion books; new york edition.

Beuscart, J., and Couronne, T. 2009. The distribution of online reputation. In *ICWSM'09*.

Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC'07*.

Cha, M.; Mislove, A.; and Gummadi, P. K. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, 721–730.

D. Alahakoon, S. K. H., and Sirinivasan, B. 1998. A self growing cluster development approach to data mining. In *IEEE Conference on Systems, Man and Cybernetic*.

Halavais, A. 2008. Do dugg diggers digg diligently. In *AOIR'08*.

Herring, S. C.; Kouper, I.; Paolillo, J. C.; Scheidt, L. A.; Tyworth, M.; Welsch, P.; Wright, E.; and Yu, N. 2005. Conversations in the blogosphere: An analysis "from the bottom up". In *HICSS'05*.

Huberman, B. A.; Romero, D. M.; and Wu, F. 2008. Crowdsourcing, attention and productivity. *CoRR* abs/0809.3030.

Kohonen, T. 1990. The self-organizing map. *Proc. IEEE* 78(9):1464–1480.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *KDD '06*.

Merton, R. K. 1968. The matthew effect in science. *Science* 159(3810):56–63.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *IMC'07*.

Mislove, A.; Koppula, H. S.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2008. Growth of the flickr social network. In *WOSP '08: Proceedings of the first workshop on Online social networks*, 25–30. New York, NY, USA: ACM.

Stoica, A., and Prieur, C. 2009. Structure of neighborhoods in a large social network. In *The 2009 IEEE International Conference on Social Computing*.

Szabó, G., and Huberman, B. A. 2008. Predicting the popularity of online content. *CoRR* abs/0811.0405.

Wazik, B. 2009. *And Then There's This. How Stories live and die in viral culture*. Viking, new york edition.