New Journal of Physics
The open-access journal for physics

The complex network of musical tastes

Javier M Buldú[1], P Cano[2], M Koppenberger[2], Juan A Almendral[1] and S Boccaletti[3]

[1] Departamento de Física, Universidad Rey Juan Carlos, Tulipán s/n, 28933 Móstoles, Madrid, Spain
[2] Music Technology Group, Universitat Pompeu Fabra, 08003, Barcelona, Spain
[3] CNR-Istituto dei Sistemi Complessi, Via Madonna del Piano, 10, 50019 Sesto Fiorentino (Florence), Italy
E-mail: javier.buldu@urjc.es

New Journal of Physics **9** (2007) 172
Received 13 October 2006
Published 28 June 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/6/172

Abstract. We present an empirical study of the evolution of a social network constructed under the influence of musical tastes. The network is obtained thanks to the selfless effort of a broad community of users who share playlists of their favourite songs with other users. When two songs co-occur in a playlist a link is created between them, leading to a complex network where songs are the fundamental nodes. In this representation, songs in the same playlist could belong to different musical genres, but they are prone to be linked by a certain musical taste (e.g. if songs A and B co-occur in several playlists, an user who likes A will probably like also B). Indeed, playlist collections such as the one under study are the basic material that feeds some commercial music recommendation engines. Since playlists have an input date, we are able to evaluate the topology of this particular complex network from scratch, observing how its characteristic parameters evolve in time. We compare our results with those obtained from an artificial network defined by means of a null model. This comparison yields some insight on the evolution and structure of such a network, which could be used as ground data for the development of proper models. Finally, we gather information that can be useful for the development of music recommendation engines and give some hints about how top-hits appear.

PII: S1367-2630(07)35486-4

Contents

1. Introduction

Many natural, technological and social systems find a suitable representation as networks made of a large number of highly interconnected units. In some cases, these networks can be tangible objects in the Euclidean space, such as electric power grids, the internet, biological, neural and chemical networks. In some other cases, they can be entities defined in more abstract spaces, such as networks of acquaintances, friendships or collaborations between individuals.

During the last decade, the grown availability of large databases and the optimized rating of computing facilities have constituted a better and better machinery to explore the topological properties of several networked systems from the real world. The main outcome has been to reveal that, despite the inherent differences, most real networks are characterized by similar topological properties. For instance, they have relatively small characteristic distances between any two nodes, high clustering features, fat tailed shapes in the distribution of connectivities, and exhibit a hierarchical structure of motifs and modules.

These findings have initiated a new trend of interest and research in the study of *complex networks* [1]–[4], i.e., networks whose statistical properties deviate significantly from those of both random and regular graphs, their structure being irregular, complex and dynamically evolving in time.

In this paper, we study an evolving network that is driven by the emergence and evolution of musical tastes. Precisely, by using of available data on users' playlists (lists of favourite songs), we construct a network, in which each song is represented as a node, and two songs are linked by an edge whenever they co-occur in a playlist. The analysis of the main properties of this network allows us to gather information on how it grows under the influence of the users' musical tastes (if two given songs co-occur in several playlists, then a given user who likes one of such songs is very likely also to appreciate the other). Indeed, as the input playlists carry information on the exact date at which they were published, we can detect how the parameters characterizing the topology of the associated network evolve in time. We also analyse some characteristics of the network by comparing it with an artificial network obtained from a null model. Next, we define an affinity matrix that can be used to design recommendation networks based on user playlists. Finally, we give some hints about the evolution of top-hits, i.e., how the most repeated songs evolve in time.

It is important to emphasize that a large portion of the research carried out so far in complex networks has focused on unweighted networks, i.e., networks where the edges between nodes are either present or not. However, along with a complex topological structure, real-world networks (such as the one in the present study) display a large heterogeneity in the intensity of the connections. Examples are the existence of strong and weak ties between individuals in social networks [5], uneven fluxes in metabolic reaction pathways [6], the diversity of the predator–prey interactions in food webs [7]–[10], different capabilities of transmitting electric signals in neural networks [11, 12], unequal traffic on the internet [13] or of the passengers in airline networks [14]. Therefore, a better description of these systems can be made in terms of weighted networks, i.e., networks in which each link carries a numerical value measuring the strength of the connection [15]. In our case, for instance, the connection between two songs co-occurring in only a very few playlists (or in only a single playlist) will be weaker than that connecting a pair of songs that co-occur in a large number of playlists, thus making it natural to use a weighted network representation.

Another important feature of the network studied here is that it comes from a bipartite network [16] of songs and playlists, which has been collapsed into a network of connected songs (one is not interested in studying the connection topology between playlists). This is in analogy with what is usually done for reconstructing a *recommendation network*, whose topological properties are exploited to give proper advice to customers for buying new products, based on what each customer had already bought in the past. Indeed, the interest in understanding the properties of musical networks is due to the fact that they constitute the input information that music recommendation engines use [17] for music discovery and automatic playlist generation tasks. Fundamental properties such as the size of the giant component and the existence of disconnected components impose great limitations on the recommendations that can be derived from such a network.

2. The network of musical tastes

The Art of the Mix [18] is a project started at the end of 1997. It consists of a website where users upload and interchange playlists of their favourite music. The songs, somehow, fit in those lists, even though they do not need to belong to the same country, decade or musical genre. In this way, a certain connection results between songs of the list, whose origin is based on the musical taste of the playlist author. By taking into account all uploaded playlists, one can create a network where songs are the nodes and co-occurrence in a playlist gives rise to a link between two nodes. The network obtained has two fundamental features; (a) it is an undirected network, i.e., links between nodes (songs) are symmetric, and (b) it is a weighted network, i.e., links have a certain weight depending on how many times two songs have coincided in a playlist.

It is worth mentioning the way new nodes and links are created. When a new playlist is added, a certain number of songs n' appear simultaneously. n' is not fixed but it normally lies between 10 and 20. At the same time, since the nodes in the new playlist are connected all-to-all, $(n'(n'-1)/2)$ new links are created. Therefore, the appearance of a new playlist will be reflected in the network by $(n'(n'-1)/2)$ new links and by a certain number of new nodes n which will always be equal or lower than n', depending on how many songs in the new playlists were already mapped into nodes of the pristine network. Finally, links connecting the same nodes are grouped and a weight is assigned to the link. The more times a link is repeated, the higher weight is assigned to the link.

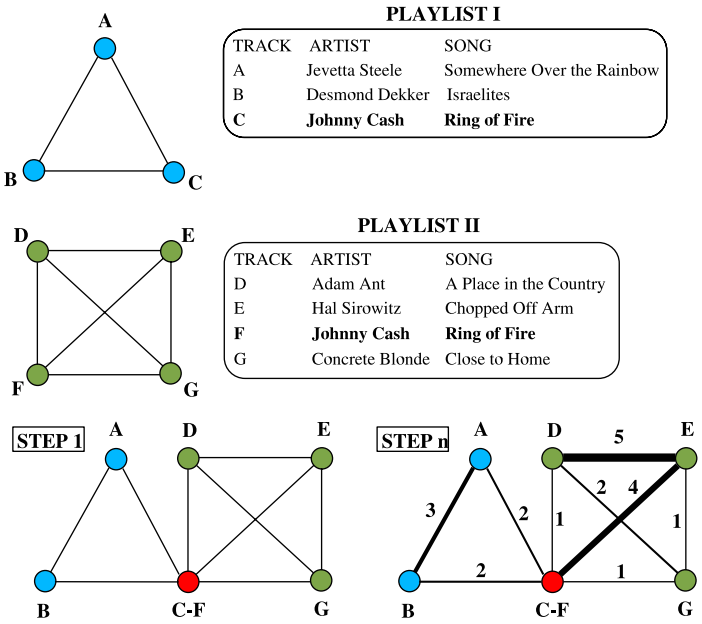

Figure 1. Network description. Playlist I (three songs) and playlist II (four songs) are an example of how songs and links are added to the network. All songs within a playlist are connected in an all-to-all configuration. When a song is repeated in two playlists (e.g. Johny Cash: Ring of Fire) only one node is considered (C-F). Note that in this case the repeated song acts as a path between songs of both playlists, reducing their distance and, therefore, joining songs from different musical tastes. After new playlists are added to the network, links between two songs could be repeated, leading to links of different weights.

Figure 1 summarizes, with an example, the procedure of adding nodes and links to the network. Two playlists of three (playlist I) and four (playlist II) songs are translated into seven nodes and nine links. Songs of the same playlist form a highly connected community with all songs linked between them. Note that, since one song co-occurs in both playlists (Ring of Fire by Johny Cash), a path is created between both communities, reducing the distance between songs of each playlist. It is worth noting the importance of the co-occurring song, since it acts as a bridge between two playlists created from different musical tastes. We must clarify that a playlist is designed under the influence of a certain musical taste, but it does not represent the musical taste itself. In this sense, two different playlists may have a high percentage of repeated songs and belong to different musical tastes. This is due to the fact that a single song can not be associated to a unique musical taste. We will take up again the importance of these 'bridge' songs in the next section of the manuscript. The last plot of figure 1 shows the state of part of the network after 'n' steps (considering a step as a playlist addition). We observe how links have a certain weight associated, which takes into account the number of times that a link between two songs has been repeated.

Once the mechanism of network creation is explained, it is clear that we are not dealing with a static network since playlists and, therefore, songs are continuously being added. In this way, the number of nodes and links increase in time and the network, and its basic characteristics, may change with its growing. In the following section, we will focus on the network evolution.

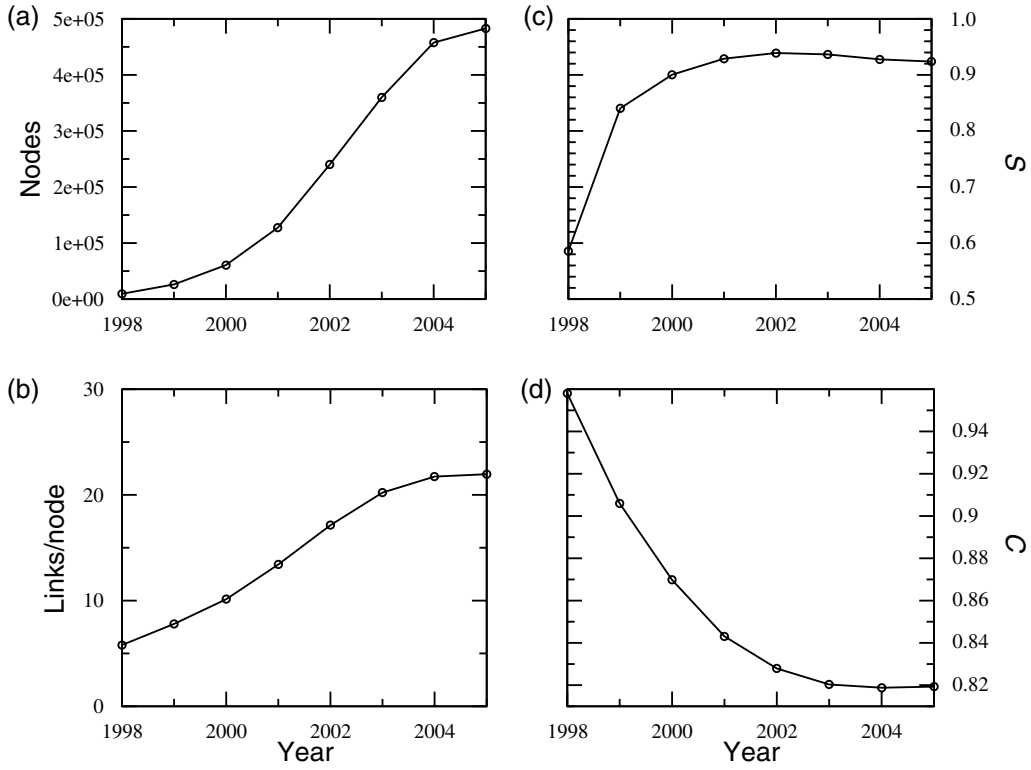

Figure 2. Evolution of the network: (a) number of nodes and (b) number of weighted links per node from 1998 to 2005. In (c) and (d) we compute, respectively, the relative size of the giant component S and the evolution of the mean clustering coefficient C.

3. The network evolution

It is remarkable how the collection of playlists of The Art of the Mix has increased since its creation at the end of 1997. At the middle of 2005, more than 82 000 playlists had been uploaded, with a mean number of 22 songs per playlist. Parallel to this evolution and six months after the creation of The Art of the Mix, a seminal paper by Watts and Strogatz [1] started a fruitful period within complex networks theory, as reflected by the great and sound number of publications related to this field (see, e.g., [4] and references therein). Nowadays, it is well known that technological, biological or social networks share common features, and that several statistical parameters can be evaluated in order to classify and understand the wide spectrum of complex networks. In what follows we will try to explain the topology and evolution of the network of musical tastes from the point of view, and with the help, of complex network theory.

First, let us consider how the fundamental units of the network, i.e., nodes and links, increase in time. Figure 2 shows the evolution of the total number of nodes (a) and links (b) per year. We must mention that data refers to the period of time between 22/01/1998 to 06/04/2005 and, therefore, years 1998 and 2005 are not complete. At first sight, it is interesting to note that not only the number of nodes but also the number of links per node has increased. The latter indicates that the average number of songs within a playlist has increased up to a limit value, during the evolution of the network.

Table 1. Summary of several network parameters as a function of year: number of nodes n, number of edges m, relative size S of the GCC, precisely, its percentage among all nodes, mean geodesic path $\bar{d}$ inside the GCC, diameter $d_{\max}$ of the GCC and the average clustering coefficient C of the network.

	The Art of the Mix							
Year	1998	1999	2000	2001	2002	2003	2004	2005
n (nodes)	9450	26 223	60 673	127 519	240 157	360 034	457 660	482 856
m (links)	54 789	204 277	614 644	1 711 053	4 115 893	7 278 256	9 946 715	10 602 036
S of GCC	58.6%	84%	90%	92.9%	93.9%	93.7%	92.8%	92.4%
$\bar{d}$ ($d_{\max}$)	6.65 (15)	5.24 (13)	4.70 (11)	4.37 (12)	4.22 (12)	4.13 (13)	4.12 (15)	4.12 (15)
C	0.958	0.906	0.870	0.843	0.828	0.820	0.819	0.819

Together with the growth of the network, the phenomenon of percolation takes place. This phenomenon is related to the appearance of a big cluster of connected nodes that covers the majority of the network. In figure 2(c) we measure the relative size S of the giant connected component (GCC), i.e., the cluster of connected nodes with the biggest size divided by the total number of nodes. We observe how S saturates to a value close to 0.92, which indicates that 92% of the nodes can be linked through a path. The value of S is in this case is a good indicator of the maturity of the network. Note that at year 1998 the GCC only covered 58.6% of the nodes and it is not until year 2000 that S arrives to a value close to 0.92.

Another indicator of the network topology is the clustering coefficient. Precisely, the distribution $C(k)$ as a function of the node degree k and its mean value $C = \frac{1}{n}\sum C_i$. The clustering, also known as transitivity, measures the probability that two neighbours of a certain node are also connected between them. It has been shown that the C in small-world or scale-free networks is much higher than that of random networks [1, 2], and $C(k)$ usually has a power law decay in many real networks [19]. In our particular case, the clustering coefficient must be regarded carefully. Since songs of the same playlist are connected all-to-all, the clustering coefficient within a playlist is 1. Only when a song appears in different playlists are its neighbours not necessarily connected among themselves, and the clustering coefficient of that node can be lower than one. Figure 2(d) shows the evolution of the average clustering coefficient. We observe how it decreases to a limit value close to $C = 0.82$ and, as happened with the relative size of the GCC, it is an indicator of the maturity of the network. It is worth noting that the decrease of C reveals the appearance of the same songs in different playlists, acting as bridges between groups of songs that are not connected. The saturation of C indicates that the ratio of appearance of repeated songs remains constant despite the evolution of the network, but only when the GCC of the network reaches a critical size.

Table 1 summarizes the characteristic network parameters and its evolution with the network growth. Apart from those values already commented upon, we have measured the mean geodesic distance $\bar{d}$ (i.e., the mean shortest path of any two nodes (songs) in the network) and the longest geodesic distance $d_{\max}$. We observe that the network shows, in all cases, the small-world phenomenon, i.e., despite the high number of nodes, there exists a path between any two nodes that involves very few connections. Precisely, the mean number of links between two nodes settles to a value close to four, despite the final number of nodes being around half million.

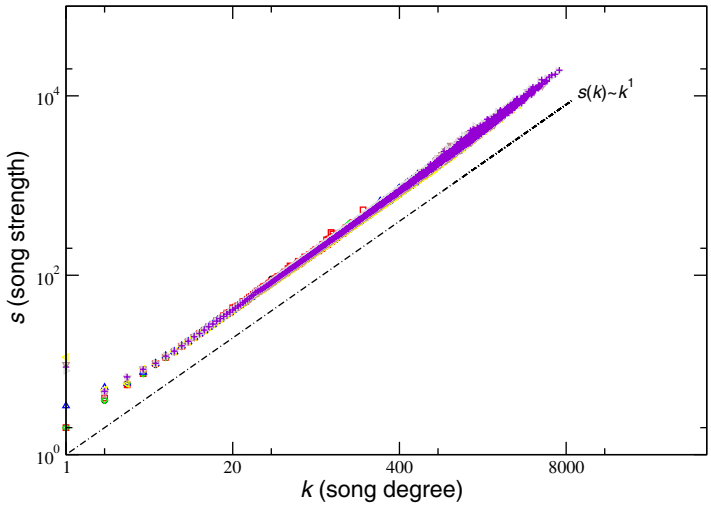

Figure 3. Evolution of the average strength $s(k)$ as a function of the degree k of the nodes. Symbols correspond to years: 1998 (black circles), 1999 (red squares), 2000 (green diamonds), 2001 (blue triangles), 2002 (yellow triangles), 2003 (brown triangles), 2004 (grey triangles) and 2005 (violet plus sings). The dotted line corresponds to the function $s(k) = k^1$.

Furthermore, when the whole network is considered (year 2005), the longest distance between any two nodes is only $d_{\max} = 15$.

At this point, it is worth noting the nature of the links, which have a certain weight associated depending on how many times two songs have been selected for the same playlist. The number of links of a node is known as its degree k, while the strength s_i of a node i is the sum of its links multiplied by their corresponding weights (ω_{ij}) and is given by $s_i = \sum_j^N \omega_{ij}$.

In figure 3, we have plotted the average strength $s(k)$ of nodes with degree k, which have been measured yearly. We can see that $s(k)$ shows a power law distribution of exponent one (i.e., linear relationship), which indicates that the weight of a link is basically independent of the degree of the node, as shown in [15]. This point reflects the fact that songs with higher numbers of links, i.e., the most repeated songs, do not necessary coincide in a playlist with the same group of songs, since in that case songs with low degrees could have high strengths (few links but very repeated). Similar behaviour of the $s(k)$ function has been observed in the network of scientists who have co-authored a paper [15].

One way of classifying complex networks is to analyse the distribution of the network degrees $p(k)$, defined as the fraction of nodes that have a degree, i.e., number of links, k. A random graph, for example, where each degree has the same probability of having a link, has a binomial distribution. Other common degree distributions are those reported in networks grown by preferential attachment algorithms [2], which show power law decays. In our case, we consider the $p_c(k)$ of our growing network, which is defined as the cumulative degree distribution and shares the shape of $p(k)$ avoiding the fluctuations commonly observed at nodes with highest degrees. Besides $p_c(k)$, we compute the cumulative strength distribution $p_c(s)$ since we are dealing with a weighted network, despite the fact that both distributions behave in the same way, as we will see shortly. Figure 4 shows the evolution of $p_c(s)$ and $p_c(k)$ (insets) in log–log (a) and linear–log (b) scales. The different scales help to identify the kind of decay of the distributions,

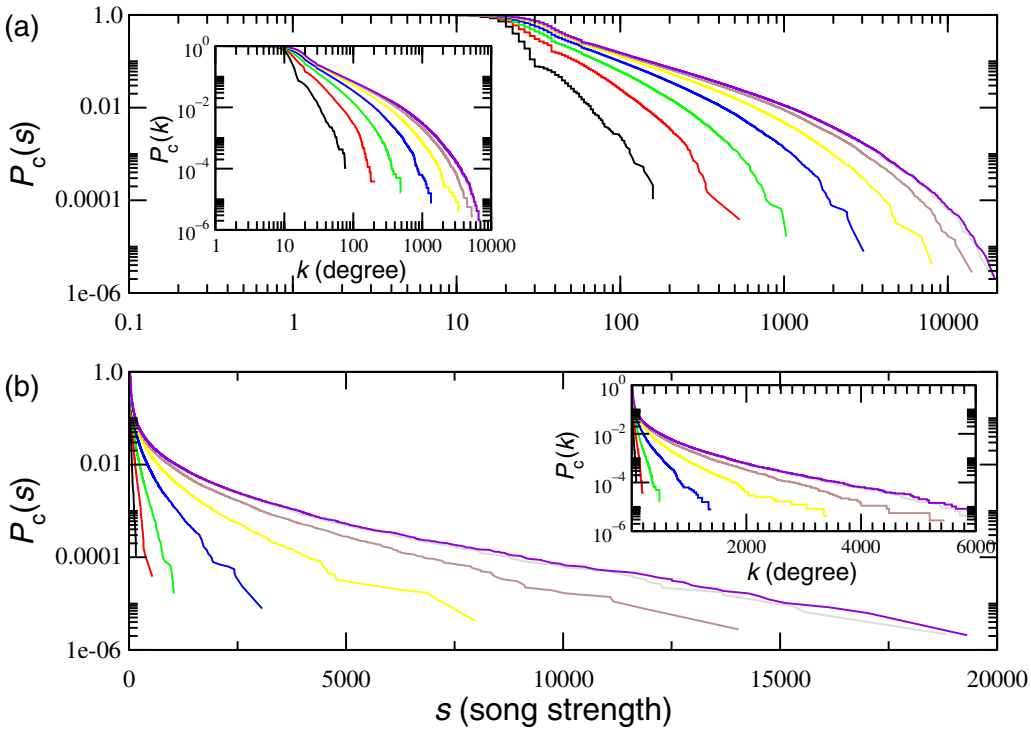

Figure 4. Cumulative strength distribution sorted by year and its corresponding degree distribution (insets). Data are plotted in a log–log scale (a) and linear–log scale (b), in order to ease comparison with power law or exponential decay. Colours correspond to years: 1998 (black), 1999 (red), 2000 (green), 2001 (blue), 2002 (yellow), 2003 (brown), 2004 (grey) and 2005 (violet).

a straight line indicates power law decay in the logarithmic scale and exponential decay in the semi-logarithmic. In this way, we can observe how strength (and degree) distributions share a common feature, they have a power law decay for low to intermediate strengths (degrees) and exponential decay for high degrees. Concerning the evolution of $p_c(s)$ and $p_c(k)$ distributions, we do not observe significant differences, despite the fact that the maximum degree increases by two orders of magnitude.

Let us take up again the concept of network transitivity by analysing the distribution of the clustering coefficient $C(k)$ as a function of the node degree. Figure 5 shows the $C(k)$ distribution, which is plotted by year. At first sight, we observe a monotonously decaying function, which approximates a power law decay, at least up to intermediate values of k. The exponent of the power law is close to -1, as has been observed in deterministic scale-free networks [20] and also in real networks [19], although it deviates from this value at the tails of the distribution, i.e., nodes with the highest degrees. The power law decay of $C(k)$ has been related to a hierarchical modularity of the network [19] and this is definitely the case of our particular network, where the clustering coefficient measures the connectivity in the neighbourhood of a certain node. In our network, nodes with lower connectivities have a clustering coefficient close to one, since their neighbours are basically those songs in the same playlist. On the other side, the highest connected nodes appear in several lists and, therefore, their neighbouring songs do not need to be connected. Concerning the deviation from the power law decay at the tail of the distribution,

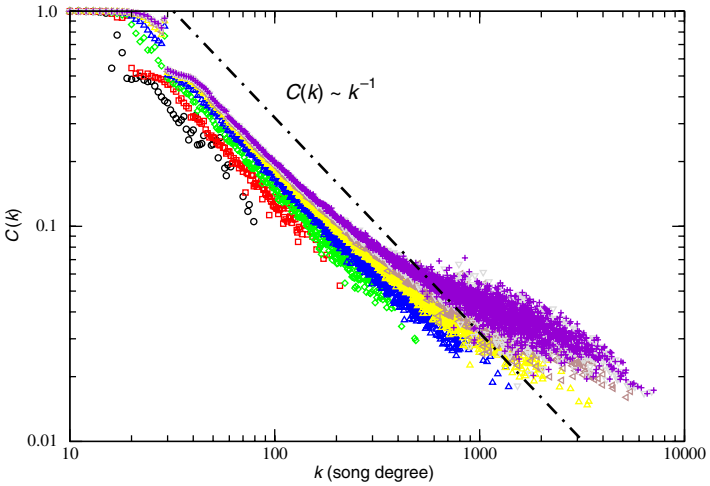

Figure 5. Evolution of the clustering coefficient $C(k)$ as a function of the song degree k. Symbols correspond to years: 1998 (black circles), 1999 (red squares), 2000 (green diamonds), 2001 (blue triangles), 2002 (yellow triangles), 2003 (brown triangles), 2004 (grey triangles) and 2005 (violet plus sings).

where higher values are obtained (see, e.g., $C(k)$ corresponding to year 2005), it indicates that despite the fact that nodes with highest degrees connect songs from different playlists, some of them are prone to be already connected through a co-occurrence in another list. This point would indicate that there are certain groups of songs that tend to appear together.

Finally, we have evaluated the nearest neighbour degree $k_{\rm nn}$ as a function of k, which is given by the expression

$$k_{\rm nn}(k) = \sum_{k'=0}^{\infty} k' p(k'|k),$$

where $p(k'|k)$ is the fraction of links joining a node of degree k to a node of degree k'. Thus $k_{\rm nn}$ is the mean degree of those nodes that are found by following the links emanating from a node of degree k. The evolution of $k_{\rm nn}(k)$ is related to the *assortativity* of the network [21], which indicates the tendency of a node of degree k to associate with a node of the same degree k. When $k_{\rm nn}(k)$ is an increasing function of k, the network is assortative. In other words, we evaluate if the most repeated songs are prone to co-occur on the same playlist. Figure 6 shows the annual evolution of $k_{\rm nn}(k)$, showing that despite the fact that $k_{\rm nn}(k)$ is an increasing function, it saturates at nodes with higher degrees, a fact that it is more pronounced as the network grows in time. This fact reveals that, despite the fact that songs of the same degree tend to co-occur, it is difficult to find two (or more) top-degree songs in the same playlist, as the size of the network increases.

4. The null model network

Once the structure of the network has been analysed, let us compare it with a network created from a null model. We create two different networks in order to distinguish between the influence of the playlist size and the song probability distribution. First, we fix the number and size of

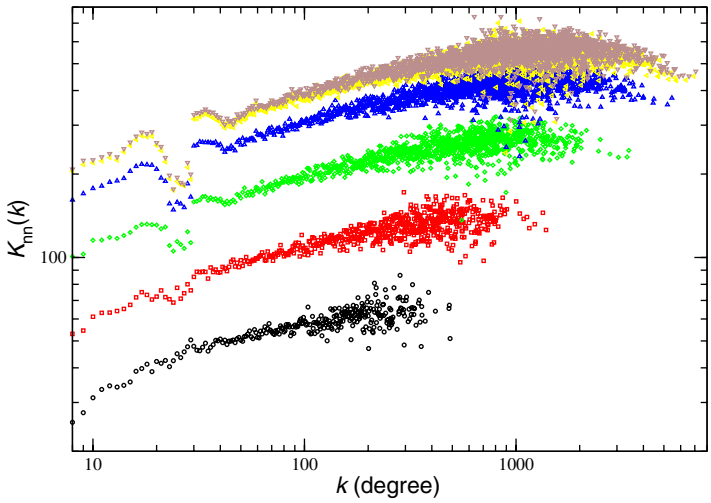

Figure 6. Neighbour connectivity $k_{nn}(k)$ as a function of the song degree. Symbols correspond to years: 1998 (black circles), 1999 (red squares), 2000 (green diamonds), 2001 (blue triangles), 2002 (yellow triangles), 2003 (brown triangles), 2004 (grey triangles) and 2005 (violet plus sings).

the playlists to that of the real data, and fill randomly the playlist distribution with N songs, all of them with equal probability. In this way, we can analyse the influence of the playlist distribution in the network structure, neglecting the influence of the song distribution. Next, we create a second network departing from the real playlist distribution which will be filled with N songs, but keeping the real probability distribution of songs. In other words, we have emptied the playlists and we have filled them again randomly with the same songs that we had initially. Comparing the first and the second network we will observe the influence of the song distribution. Figure 7 shows the cumulative degree distribution (a-b-c), the clustering (d-e-f) and the nearest neighbour degree (g-h-i) for the real network (upper plots) and the two artificial networks (middle and lower plots). The cumulative degree distribution shows the importance of keeping the song probability of the real data. Obviously, 'top-hits' are prone to appear more frequently, since they are supposed to be liked (or known) by a wide range of the users and, therefore, their probability of being included in a playlist is higher. This point is raised by the degree distribution of the network created by assuming equal probability for all songs. In this ideal framework, the degree distribution shows a sharp decrease (see figures 7(b)), indicating the disappearance of significant hubs in the network. Only when the real song distribution is considered, the degree distribution is preserved, as can be seen by comparing figure 7(a) and (c). It is worth mentioning that keeping the song probability distribution does not ensure an equal degree distribution, since the degree of a node depends on the number of songs in the playlist and the playlist size is not constant. When looking at the clustering distribution (see figures 7(d)–(f)), we can observe a negative slope in all cases. If we compare the real network (figure 7(d)) with that created from the real song distribution (figure 7(f)), we can see how they behave similarly, despite the fact that the clustering distribution decreases to lower values for the case of the artificial network. Thus we can see that in the real network the clustering coefficient saturates at nodes with higher degrees. Finally it is interesting to look at the nearest neighbour degree k_{nn}, which is an indicator of the assortativity of the network [3]. We find that the network created with equal song probabilities shows no assortativity (figure 7(h)). Although there is an

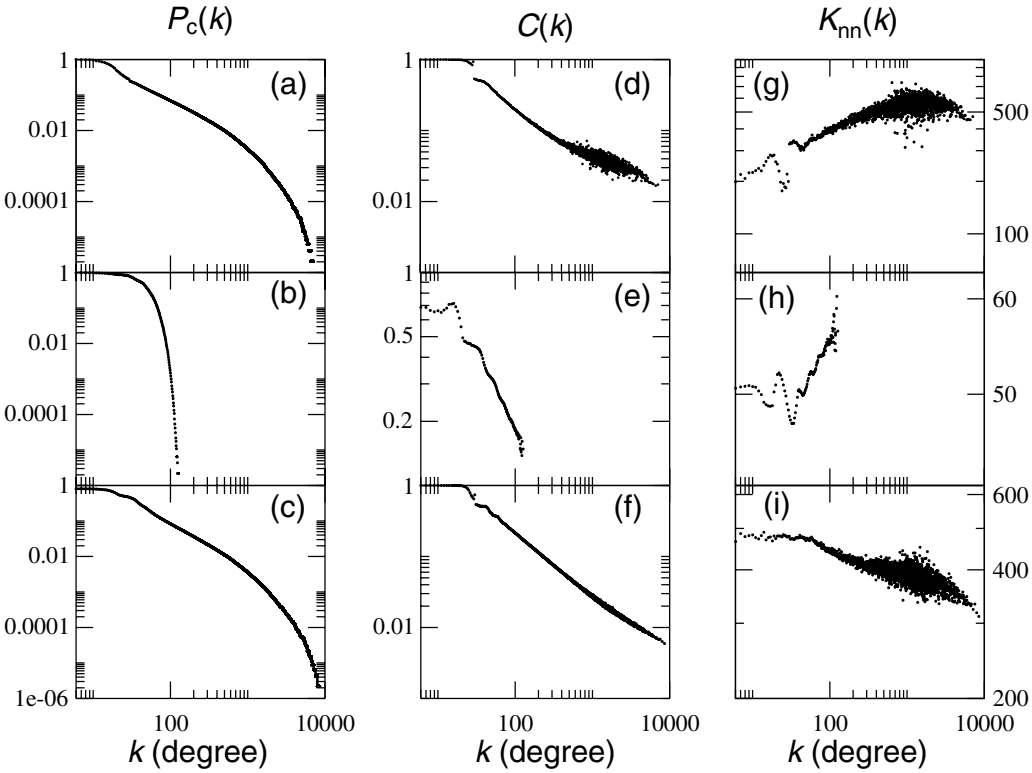

Figure 7. Cumulative degree distribution $p_c(k)$ (a-b-c), clustering distribution $C(k)$ (d-e-f) and nearest neighbour degree $k_{nn}(k)$ (g-h-i) for the real network (upper plots), a random network created from a song distribution with equal probability (middle plots) and a random network created with a song probability extracted from the real data (lower plots). In all cases, we have kept the playlist distribution.

increasing trend, it is not significant since the nearest neighbour degree just runs from 50 to 60. However, if we compare the k_{nn} distribution of the real network and the network created from the real song distribution, the difference is striking. Despite both networks showing similar degree and clustering distributions, the k_{nn} reveals an assortative nature (positive slope) for the real data and a dissortative behaviour (negative slope) for the artificial network. Therefore, we can conclude that in the real network, songs with higher degree (i.e., 'top-hits') are more prone to co-occur than in the random case. This fact indicates the existence of a wide community of users, with a 'commercial' behaviour, who are more susceptible to include 'hits' (or hubs) in their playlists.

5. The affinity matrix

In this section we construct an *affinity matrix* A that measures the affinity between two songs obtained from its probability of co-occurrence within the same playlist. The affinity matrix not only gives information about how two songs resemble each other but also can be used as ground data for the development of a music recommendation system. Given a certain song a we define

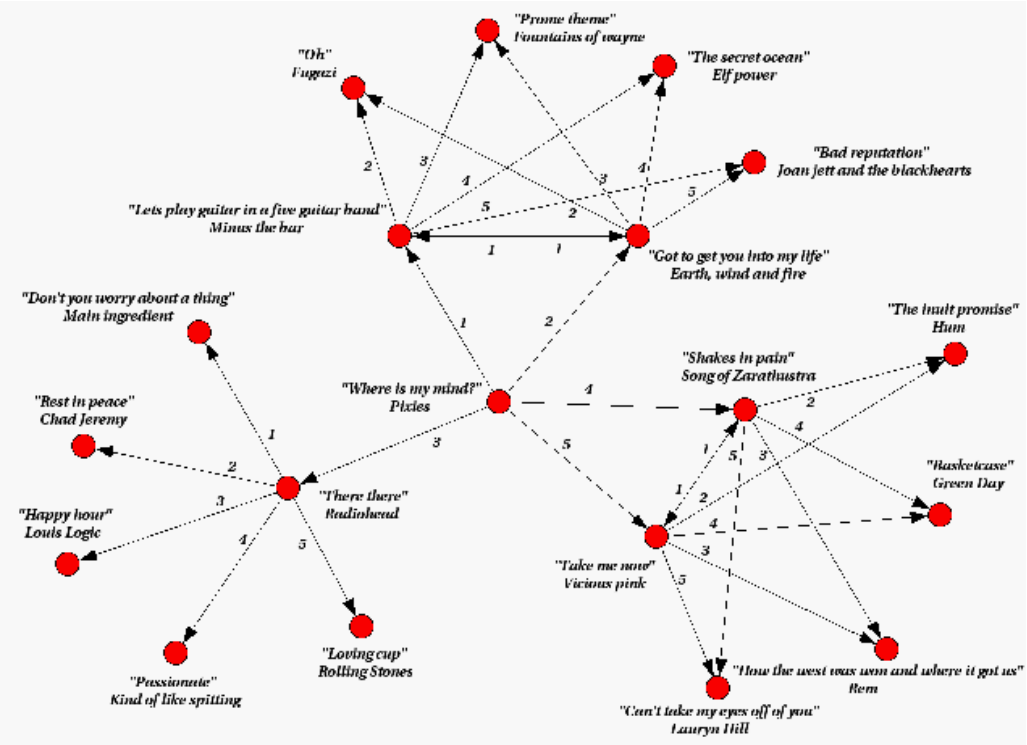

Figure 8. Example of a recommendation network obtained from the affinity matrix. Only the part of the network surrounding the song 'Where is my mind?' is shown. Numbers correspond to the ranking of affinity of the outgoing links.

the set of its neighbour songs b_i as those that appear with a in any playlist of the network. For each neighbour, we count the number of co-occurrences with a, which corresponds to the weight of the link connecting both nodes w_{ab_i}. Given that n_a and n_{b_i} are, respectively, the number of appearances of a song a and its neighbour b_i in the whole network, we can define the ratio $0 \leqslant \frac{w_{ab_i}}{n_a n_{b_i}} \leqslant 1$ as the affinity between a and b_i. Note that in a network of N songs we can include all affinities in an affinity matrix A, which will be a $N \times N$ symmetric matrix with the diagonal equal to zero.

Since line i of matrix A measures the affinity of song i with all other songs of the network, it can be used as a recommendation database. Once a user selects a certain input song i, they can be pointed to songs with the highest values in the affinity matrix. Furthermore, we can project the affinity matrix into a a recommendation network. By ranking the affinities of a certain song a we can select the M songs with highest affinities and create a network by linking only this set of songs. We can also include a low threshold where a minimum number of co-occurrences has to be fulfilled. Following this procedure we can create different affinity networks, where links are directed and have a certain rank. Figure 8 shows an example of a recommendation network obtained from the affinity matrix. We have considered $M = 5$, i.e., the five songs with highest affinity, and set the minimum number of occurrences to ten. In figure 8 we show the part of the network surrounding the song *'Where is my mind?'* by *The Pixies*. From this starting node, we can navigate through the network, knowing at each song the order of the outgoing links that point to the songs with the highest affinity. This kind of network is suitable for implementation

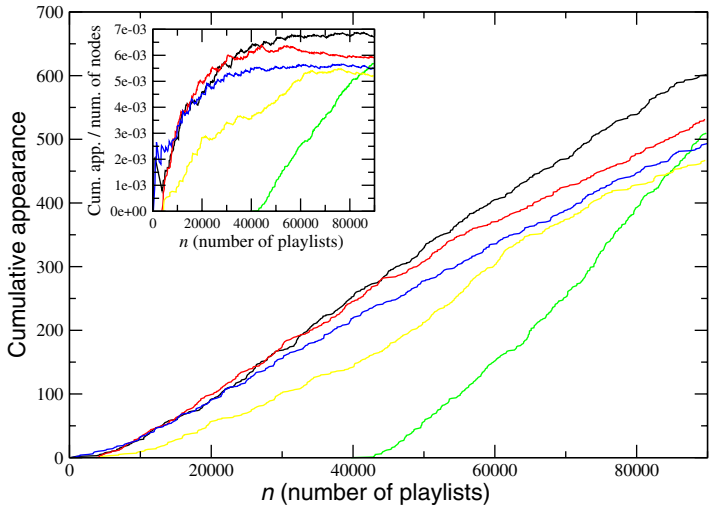

Figure 9. Temporal evolution of the three highest connected songs in the whole network (i.e., that obtained at year 2005), where n is the total number of playlists at a certain date. In the ordinate axis, we measure the accumulated appearances of a certain song among all playlists. In the inset, we evaluate the rate of appearance by measuring the number of appearances per playlist.

in a web-page context, where users click to a certain target song (web-page) and links to other recommended songs are shown.

6. Network's favourite song

Let us now move to focus on the evolution of those nodes with the highest connectivity, i.e., the most repeated songs. The role of these songs in the network is basically to reduce the path between distant or unconnected songs. They constitute the bridges between musical tastes and therefore they are of great relevance not only for the structure of the network, but also for its evolution. We have searched for the five nodes with the highest degrees in the network of year 2005 and have traced back their dynamics in the pristine networks. Figure 9 shows the appearance of these top-five songs as a function of the number of playlists in the database, which is used here as an indicator of time.

We can observe that four of them already existed when The Art of the Mix was created and all show a similar behaviour, which consists of a sharp increase of their rate of appearance (see inset) until they reach a given value, remaining then nearly constant. This fact indicates that top-hits, i.e., the most repeated songs, have a transient growth that determines the final value of their rate of appearance. Once the final value is reached, it seems that a further increase (or decrease) is prohibited, at least during a certain characteristic time [4]. Figure 9 also shows the evolution of the fifth song (green line). The behaviour of this last song is remarkably different to

[4] We believe that a decay in the rate of appearance could be expected after a certain characteristic time. This is an intriguing point since it could be expected that some 'classics' may keep their constant rate of appearance despite their aging.

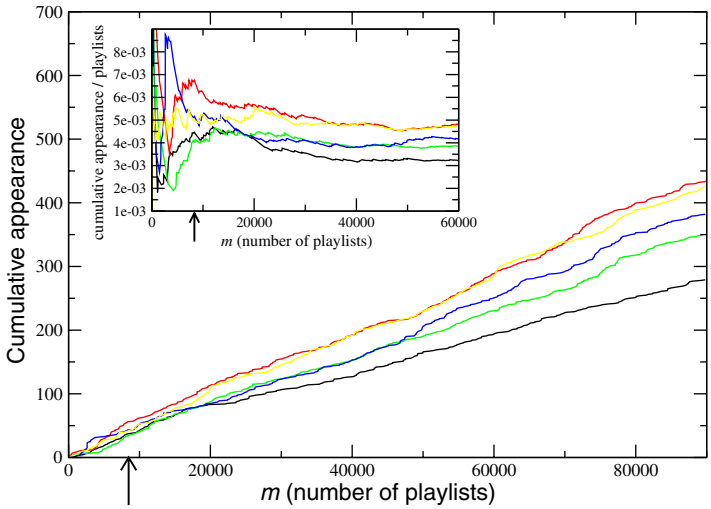

Figure 10. Evolution of the three highest hubs at the beginning of year 2001 (marked with an arrow in the plot). The inset shows the rate of appearance, computed by the relation between the appearances and the total number of playlists.

that of the other four. It appears at later times with respect to the others and its rate of appearance has not yet saturated, which indicates that it is probably going to overcome the others and reach the top of the list.

After viewing the evolution of the top-five connected songs as in 2005, an obvious question arises: were these songs always the top-connected? The answer is no, which automatically leads to a second question, what happened to the previous top-five songs? In figure 10 we have plotted the evolution of the most connected songs at end of the year 2000. We have chosen this year since, as we have seen in the previous section, the network was at that time already mature enough. It is interesting to see that some differences are found with the previous top-connected songs. In this case, we observe a peak in the evolution of the rate of appearance, whose shape depends on the song. In this way, sharp peaks unveil a fast growth and successive decay, which could be expected in songs with a short temporal impact. Despite their differences, all of them finally reach a constant rate of appearance, a fact also shared with the top-five songs previously considered.

7. Conclusions

We have analysed the topology and evolution of a network of songs. The network is constructed by adding a link among songs whenever they co-occur in a playlist. The selections of songs performed by users is driven by their personal musical taste. Playlist collections, or the analysis of listening behaviour is a passive source of information that feeds some of the current commercial music recommender engines. In the current study, we have proposed the creation of recommendation networks based on the affinity matrix. On the other hand, basic analysis on the topology of the networks highlights fundamental information on the quality of the recommender engines built with such information. An interesting result is that it took several years (from 1997 until 2000) for the networks basic properties to converge. As pointed out in previous sections, the

size of the giant component accounts for 92% of the songs from the year 2000. Accordingly, there is a significant 8% of songs which have very limited recommendations. Choosing as seed any song from this 8% of presumably obscure material will never obtain songs from the giant component as recommendations. Furthermore, we have developed an artificial network, constructed from a null model, in order to unveil the particular features of the network under study. Next, we have analysed the dynamics of the highest connected songs, following the evolution from their appearance in the network. An interesting observation is that the bridge songs, i.e., those songs that enable recommendations across playlists, are extremely popular and make the clustering coefficient decrease. We have observed that top-hits achieve a constant rate of appearance after some initial transient. Further study in this direction could be focused on classifying songs (or their maturity) depending on their rate of appearance. Likewise, future study should account for the directionality of the playlist. If we consider a playlist as a sequence of elements with a compositional intent, the order in which songs appear encodes important information for the development of music recommendation systems.

Acknowledgments

JMB and PC thank Pablo de Miguel for fruitful discussions. JMB acknowledges VDP Servedio for his help with the Net software [22] and also financial support from MCyT–FEDER (Spain, project BFM2003-07850) and from the Generalitat de Catalunya. SB acknowledges the Yeshaya Horowitz Association through the Center for Complexity Science. Finally, this research was also funded in part by the NSF grant DMS-0405348 and in part by project Pharos IST-2006-045035.

References

[1] Watts D J and Strogatz S H 1998 *Nature* **393** 440–2
[2] Barabási A-L and Albert R 1999 *Science* **286** 509–12
[3] Newman M E J 2003 *SIAM Rev.* **45** 167–256
[4] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 *Phys. Rep.* **424** 175–308
[5] Granovetter M 1973 *Am. J. Sociol.* **78** 1360
[6] Barabási A-L, Jeong H, Ravasz R, Neda Z, Vicsek T and Schubert A 2002 *Physica* A **311** 590
[7] Polis G A 1998 *Nature* **395** 744
[8] McCann K, Hastings A and Huxel G R 1998 *Nature* **395** 794
[9] Berlow E L 1999 *Nature* **398** 330
[10] Krause A E, Frank K A, Mason D M, Ulanowicz R E and Taylor W W 2003 *Nature* **426** 282
[11] Sporns O, Tononi G and Edelman G M 2002 *Neural Netw.* **13** 909
[12] Sporns O 2003 *Complexity* **8** 56
[13] Pastor-Satorras R and Vespignani A 2004 *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge: Cambridge University Press)
[14] Guimerà R, Mossa S, Turtschi A and Amaral L A N 2005 *Proc. Natl Acad. Sci. USA* **102** 7794
[15] Barrat S, Barthélemy V, Pastor-Satorras Y and Vespignani M 2004 *Proc. Natl Acad. Sci.* **101** 3747–52
[16] Battiston S and Catanzaro M 2004 *Eur. Phys. J.* B **38** 345–52
[17] Herlocker J, Konstan J, Terveen L and Riedl J 2004 *ACM Trans. Inform. Syst.* **22** 5–53
[18] The Art of the Mix, online at http://www.artofthemix.org
[19] Ravasz E, Somera A L, Mongru D A, Oltvai Z and Barabási A-L 2002 *Science* **297** 1551–5
[20] Dorogotsev S N, Goltsev A V and Mendes J F F 2002 *Phys. Rev.* E **65** 066122
[21] Newman M E J 2002 *Phys. Rev. Lett.* **89** 208701
[22] Online at http://pil.phys.uniroma1.it/~servedio/software.html