

THE EMERGENCE OF COMPLEX NETWORK PATTERNS IN MUSIC ARTIST NETWORKS

Pedro Cano and Markus Koppenberger
Music Technology Group
Institut de l'Audiovisual, Universitat Pompeu Fabra
Ocata 1, 08003, Barcelona, Spain

ABSTRACT

Viewing biological, social or technological systems as networks formed by nodes and connections between them can help better understand them. We study the topology of several music networks, namely citation in allmusic.com and co-occurrence of artists in playlists. The analysis uncovers the emergence of complex network phenomena in music information networks built considering artists as nodes and its relations as links. The properties provide some hints on searchability and possible optimizations in the design of music recommendation systems. It may also provide a deeper understanding on the similarity measures that can be derived from existing music knowledge sources.

1. INTRODUCTION

A network is a collection of items, named vertices or nodes, with connections between them. The study of the networks underlying complex systems is easier than studying the full dynamics of the systems. Yet, this analysis can provide insights on the design principles, the functions and the evolution of complex systems [9, 5]. Significant amount of multidisciplinary research on social, biological, information and technological networks has uncovered that complex systems of different nature do share certain topological characteristics. Indeed, the spread of certain ideas and religions, the success of companies, the transmission of sexually transmitted diseases such as the AIDS epidemic or computer viruses can be better understood by studying the topologies of the systems where they interact [9].

According to [9], research in complex networks aims at three things:

1. Find statistical properties, such as path length and degree distribution that characterize the structure and dynamic behavior of networked systems.

2. Build models of networks that explain and help understand how they are created and how they evolve.
3. Predict the behavior of networked systems based on the measured statistical properties of the structure and the local properties of given vertices, e.g.: what will happen to the equilibrium of an ecological network if a certain species disappears.

Complex network analysis is used to describe a wide variety of systems with interacting parts: networks of collaborating movie actors, the WWW, neural networks, metabolic pathways of numerous living organisms, to name a few. New insights can be unveiled by considering musical works and musical artists as parts of a huge structure of interconnecting parts that influence each other.

The objective of this work is to show the emergence of complex network phenomena in music information networks built considering artists as nodes and artist relationships as links between nodes. Complex network analysis may enhance our comprehension on some relevant musical and information retrieval issues. For example, how much of the network structure is due to content similarity and how much to the self-organization of the network. This could shed new light on the design and validity of music similarity evaluation [6, 3]. Secondly, a better understanding of the topology may hint possible optimizations on the design of music information systems [7]. Finally, it may help to understand the dynamics of certain aspects of music evolution, e.g.: how did an artist get popular? Besides the preliminary results and the discussion, one of the goals of this paper is to call the attention to the MIR community of this body of research which is the science of complex networks.

2. NETWORK PROPERTIES

Let us introduce some definitions and concepts that will be used in this work. A *network* or graph is a set of vertices connected via edges. Networks connected by directed edges are called *directed networks*, networks connected by undirected edges are called *undirected networks*.

Degree: The degree k_i or a vertex i is the number of connections of that vertex and $\langle k \rangle$ is the average of k_i over all the vertices of the network. In an undirected graph

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2004 Universitat Pompeu Fabra.

where each edge contributes to two vertices the average degree is:

$$\langle k \rangle = \frac{2m}{n} \quad (1)$$

where m is the number of edges and n the number of vertices.

Clustering coefficient: The clustering coefficient estimates the probability that two neighboring vertices of a given vertex are neighbors themselves. In the networks that occupy us, it relates to the probability that if artist A is similar to artist B and artist C, B and C are similar as well. Following [10] the clustering coefficient of vertex i is the ratio between the total number y_i of the edges connecting its nearest neighbors and the total number of all possible edges between all these nearest neighbors,

$$c_i = \frac{2y_i}{\langle k \rangle (\langle k \rangle - 1)} \quad (2)$$

The clustering coefficient c for the whole network is the average over the number of nodes n

$$c = \frac{1}{n} \sum_i c_i \quad (3)$$

Component: The component to which a vertex belongs is the set of vertices that can be reached from that vertex.

Average shortest path: Two vertices i and j are connected if one can go from i to j following the edges in the graph. The path from i to j may not be unique. The minimum path distance or *geodesic path* d_{ij} is the shortest path distance from i to j . The average shortest path over every pair of vertices is

$$\langle d \rangle = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \quad (4)$$

where d_{ij} is the geodesic distance from vertex i to vertex j . The maximum geodesic between any two vertices in the graph is known as *diameter*.

Degree distribution: The degree distribution $P(k)$ is the proportion of nodes that have a degree k . The shape of the degree distribution can help identify they type of network: “scale-free networks” have power-law distributions and “random networks”—as described by the Erdős-Rényi model—have a Poisson degree distribution (see below).

2.1. Random networks

The random graph model, introduced by Erdős and Rényi, connects in one of its variants every pair of vertices with a probability p (see [4] for a review on random graphs).

The degree distribution follows a binomial shape

$$P(k) = C_{n-1}^k p^k (1-p)^{n-k} \quad (5)$$

where C_{n-1}^k is the number of ways of connecting a vertex to k nodes and not to $n-k-1$ others. For large $n \gg 1$ the distribution is approximated by the Poisson distribution

$$P(k) \sim e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \quad (6)$$

Another magnitude that can be computed is the clustering coefficient. Since the probability of link between two vertices in a random graph is independent of the existence of other edges and equal to p , in average there will be $pk(k-1)/2$ out of the total possible of $k(k-1)/2$ neighbors of a vertex of degree k . Combining equation (1) and (2), the c of a random graph, c_r , is

$$c_r = \frac{\langle k \rangle}{n} \quad (7)$$

An important aspect regarding networks refers to under which conditions all the vertices are connected in a big component¹. It can be shown that there exist a threshold p_c under which the vertices are disconnected and over which a giant connected component abruptly emerges. The transition p_c occurs when $\langle k \rangle = np_c = 1$.

Random graphs reproduce the *small-world effect*, a very common phenomenon in real networks, where vertices on a network seem to be connected by short paths. The mean number of vertices at a distance d away from a vertex is $\langle k \rangle^d$. Consequently the value d needed to reach the whole network is $\langle k \rangle^d = n$. A typical distance d_r on the random network will be:

$$d_r = \frac{\log(n)}{\log(\langle k \rangle)} \quad (8)$$

Random models however fail to reproduce certain properties ubiquitous in real networks such as significantly high clustering coefficient or heterogeneity on the degree distribution. Additionally, it has been shown that random networks are impossible to navigate using local algorithms [7, 9].

2.2. Small-World Networks

In order to overcome the discrepancy between the small clustering coefficient of random networks and those observed on real graphs, Watts and Strogatz[10] proposed the Small-world network (SW networks). SW networks are made from regular lattices by connecting pairs of vertices connected at random. The small number of random shortcuts produce the short average paths while maintaining the high clustering of regular lattices. SW networks present Poisson-like degree distributions [5].

2.3. Scale-free networks

Many complex networks display a high heterogeneity on the degree distributions. It has been found that many networks follow a power-law or Scale-free (SF) distribution

$$P(k) \sim k^{-\gamma} \quad (9)$$

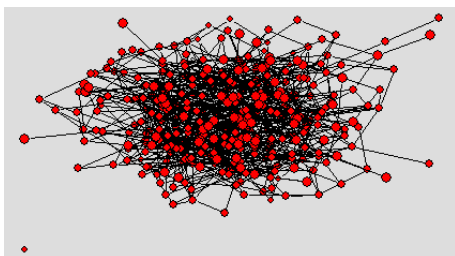
SW and random models presented so far aim to explain certain properties observed in real networks. However they do not attempt to explain the power-law distributions of large networks such as the WWW or how the networks came to be like they are in the first place. Barabási

¹ This property is known as percolation.

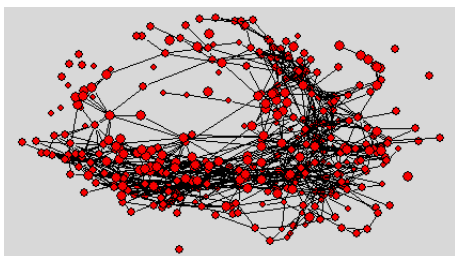
et al. [1] proposed a model that explain the power-law degree distribution of genetic networks or the WWW using two mechanisms:

1. Networks expand continuously. Indeed new pages are added to the WWW or, in our problem, new artists are added into the systems.
2. New vertices attach preferentially to sites already well connected. A variation “rich get richer” mechanism.

SF networks are robust to random node removal and fragile to targeted attacks, i.e.: removal of hubs. This property, known as *resilience* is of great interest when studying the robustness of the Internet or electric power networks. In a music recommendation system, unknown artists links offered to a user can be seen as a node removal since the user is unlikely to follow that path. If the hubs artists—the one that keep the network together—are unknown the recommendation network could be fragmented into small components.



(a) Erdős-Rényi random graph



(b) 400 Artist similarity network according to AMG

Figure 1. On the top a random graph generated with the same number of nodes and average degree than the network constructed from the similarity links edited by music experts of AllMusicGuide.com (see Section 3). The figure has been generated mapping the network to a 3D using a spring embedding algorithm available on Pajek [2].

3. MUSIC ARTIST SIMILARITY NETWORKS

We have constructed music artist similarity networks based on two sources: expert opinions and playlist co-occurrence.

	n	$\langle k \rangle$	SGC	C	C_r	d	d_r
AMix	48169	12.5	99.1	0.1	.0026	3.8	4.3
AMG	400	5.4	96.2	0.3	.0135	4.7	3.6

Table 1. Summary of basic network properties for the Art of the Mix artist network (AMix). n is the number or artists, $\langle k \rangle$ is the average degree, SGC is the size of the giant component as a percentage, C is the clustering coefficient, C_r is the clustering of a random network, d is the average shortest path, and d_r is the corresponding shortest path for the random network.

Both sources were gathered and regularized by the authors of [6] with the goal of finding a ground truth for evaluating music similarity measures [6, 3].

Expert opinion: The data consists of 400 artists along with their relations according the professional editors of “All Music Guide”² (AMG). A connection between artist is made if the “similar artists” link exists. The network, originally directed, is converted to undirected. See Figure 1 for a visualization of the AMG network. A random network constructed with similar characteristics, i.e.: 400 nodes and an average connectivity between nodes, $\langle k \rangle$ of 5.4. As can be observed in Figure 1, the AMG network is less homogeneous than the equivalent random graph.

Playlist Co-occurrence: The data consists of human authored playlists (over 29.000) from “The Art of the Mix”³ (AMix) from early 2003. A connections has been made between the artists if the co-occur in a playlist. The resulting graph has over 48.000 artists with an average of $\langle k \rangle = 12.5$ links to other artists.

4. EXPERIMENTAL RESULTS

Properties of the organization of the networks are summarized on Table 1. The first thing to note is that both networks are sparse—on average each artist is connected to a small percentage of other artists. In the case of AMix, each artist is connected on average to 12.5 (.026%) of the 48169 possible artists. Despite their sparsity, both networks contain a single giant component which connects 99.1% of the artists of AMix and 96.2% on AMG. These results are in accordance with what has been discussed on percolation on Subsection 2.1. The rest of graph measurements restrict to the giant component.

Not only is it possible to reach a large percentage of artists from any artist following the edges, which is of obvious interest in an artist recommendation system, it is also possible to do it on small number of steps (small-world effect). The average shortest path d is 4.7 steps on the case of AMG (with a maximum of 13) and on the case of AMix d is 3.8 (with a maximum of 11). The corresponding values for the equivalent random graph are very similar and are shown on Table 1.

² <http://www.allmusic.com>

³ <http://www.artofthemix.org>

The clustering coefficient c for the AMG is 0.3, meanwhile the overall clustering coefficient for equivalent random graph c_r is 0.014. For AMix $c = 0.1$ and $c_r = .0135$. Both networks have a d close to d_r and high clustering coefficient ($c \gg c_r$) with respect to the random graph. These properties are indicator of Small World structure.

As for the statistical distribution of links, in Figure 2 we depict the degree distribution $P(k)$ for the AMix network. The degree distribution of the equivalent poissonian random graph is shown for comparison. On the bottom graph of Figure 2, the cumulative $P_c(k) = \sum_{k' > k} P(k')$ is displayed. The $P_c(k)$ of power-laws $P(k) \sim k^{-\gamma}$ is also a power-law with $P_c \sim k^{\gamma-1}$ and it used to smooth the fluctuations on the tails. The distribution could correspond to a truncated power-law or a multifractal distribution [5]. The existence of a small but significant number of artists connected to a very large number of artists may hint some sort of preferential attachment growing model. The $P(k)$ of the AMG is not shown because it did not display particular pattern besides a heavy tail probably due to its small size.

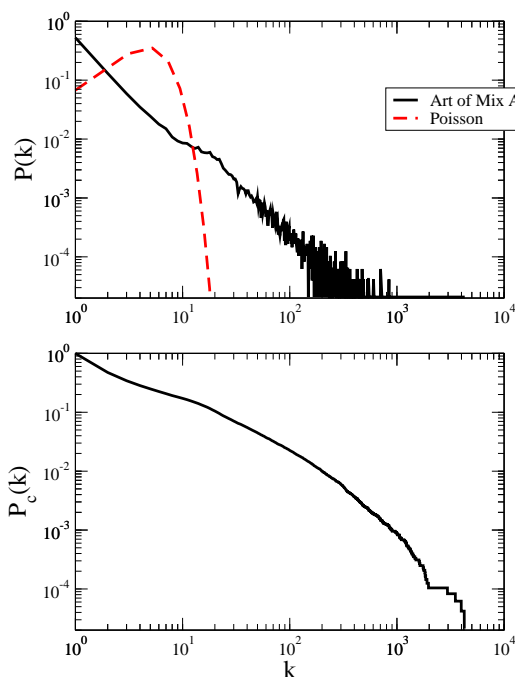


Figure 2. Top: Log-log distribution of the empirical degree distribution $P(k)$ for the Art of the Mix artists network. The dashed line shows the degree distribution of a Poissonian random graph with the same average degree distribution $\langle k \rangle = 12.5$ and it is shown for comparison: the distribution for the real networks shows a heavier tail. Bottom: Log-log distribution of the cumulative degree distribution $P_c(k) = \sum_{k' > k} P(k')$ which could correspond to be some sort of truncated power-law or maybe a multifractal distribution [5]

5. DISCUSSION

We show that music networks share some statistic properties with other complex networks, namely the Small-World structure [10].

Small-World networks have implications on the navigability of information systems. MIR systems can be structurally optimized so as to allow surfing to any part of a music collection with a small number of mouse clicks (short average distance) and so that they are easy to navigate using only local information [7, 9]. On the other hand, a deeper understanding of the underlying forces driving playlist creation or music expert knowledge networks can provide new insights on the design of music similarity measures evaluation. Is it possible to quantify how much of artist similarities are due to “popular get popular” mechanisms and how much to actual similarity between artists? [6, 3, 8, 9]

We are thankful to Oscar Celma, Fabien Gouyon and Perfecto Herrera for fruitful discussions.

This work is partially funded by AUDIOCLAS E! 2668 Eureka (<http://www.audioclas.org>).

6. REFERENCES

- [1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [2] V. Batagelj and A. Mrvar. *Graph Drawing Software*, chapter Pajek - Analysis and Visualization of Large Networks, pages 77–103. Springer, Berlin, 2003.
- [3] A. Berenzweig, B. Logan, D.P.W. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proc. of the ISMIR*, Baltimore, Maryland, 2003.
- [4] B. Bollobás. *Random Graphs*. London: Academic Press, 1985.
- [5] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.
- [6] D.P.W. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proc. of the ISMIR*, Paris, 2002.
- [7] J. M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.
- [8] F. Menczer. The evolution of document networks. *Proceedings of the National Academy of Science USA*, 101:5261–5265, 2004.
- [9] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
- [10] D.J. Watts and S.H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:409–10, 1998.