Supplementary Information: "Measuring the Evolution of Contemporary Western Popular Music"

Joan Serrà[1*], Álvaro Corral[2], Marián Boguñá[3], Martín Haro[4], and Josep Ll. Arcos[1]

[1] Artificial Intelligence Research Institute, Spanish National Research Council (IIIA-CSIC), Bellaterra, Barcelona, Spain.

[2] Complex Systems Group, Centre de Recerca Matemàtica, Bellaterra, Barcelona, Spain.

[3] Departament de Física Fonamental, Universitat de Barcelona, Barcelona, Spain.

[4] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.

[*] Corresponding author

Contents

1 Materials and Methods **2**

1.1 Dataset . . . 2
1.2 Music Descriptions . . . 3
1.3 Encoding . . . 4
1.4 Codeword Distributions . . . 5
1.4.1 Distribution Functions . . . 5
1.4.2 Fitting Procedure . . . 6
1.5 Codeword Networks . . . 7
1.5.1 Network Building . . . 7
1.5.2 Metrics . . . 8
1.5.3 Analysis Details . . . 9
1.6 Linear Regressions . . . 13

2 Supplementary Figures **14**

References **20**

1 Materials and Methods

1.1 Dataset

The million song dataset [1] is a publicly available collection of audio descriptions and metadata for "a million contemporary popular music tracks"[1] made available by Columbia University's LabROSA[2] and the company The Echo Nest[3]. As a whole, it comprises music from 44,745 unique artists and it includes a variety of music genres such as rock, pop, hip-hop, electronic, jazz, or folk. From the million tracks, 515,576 have information on the release year according to MusicBrainz[4], an open music encyclopedia that collects and makes music metadata available. Since there are some duplicate tracks in the original dataset[5] and others that do not have the full audio descriptions (the three music descriptions plus their beat information), the size of the dataset used here was reduced to 465,259 items. Further discarding the years before 1955 due to lack of representativeness, the size of the dataset becomes 464,411. A diversity of popular music genres is still included in the final subset (Fig. 1).

Figure 1: Tag cloud of the genres included in the analyzed subset of the million song dataset (using the default MusicBrainz genre information provided in the dataset). The font size represents the logarithm of number of tracks associated with a given annotation or genre tag.

[1] http://labrosa.ee.columbia.edu/millionsong
[2] http://labrosa.ee.columbia.edu
[3] http://the.echonest.com
[4] http://musicbrainz.org
[5] http://labrosa.ee.columbia.edu/millionsong/blog/11-3-15-921810-song-dataset-duplicates

1.2 Music Descriptions

As mentioned, the dataset provides state-of-the-art audio descriptions for each beat of a given track [2, 3]. Thus, for each track, a sequence of multi-dimensional values is obtained. The most relevant descriptions are related to pitch, timbre, and loudness. These descriptions are psychoacoustically-motivated, and its computation includes several steps to mimic the response of the human ear such as the grouping of energies into perceptually-motivated frequency bands, the consideration of spectro-temporal dynamics, or the application of an outer and middle ear filter [2]. A general overview of pitch, timbre, and loudness descriptions follows.

Pitch correlates with the periodicity of air pressure fluctuations [4, 5] and is represented by a real-valued 12-dimensional vector of pitch class relative energies (also called chroma [6, 7]). Pitch class-based representations are a standard way of describing the relative energy of the pitches present in an audio frame, and have been key in the development of many applications dealing with music signals such as the automatic identification of near-duplicate recordings [8], chord/tonality estimation [9], or music structure segmentation [10]. In such a representation there is a value between 0 and 1 indicating the degree of absence or presence of each of the 12 pitch classes of the chromatic scale (C, C#, D, D#, etc.). In principle, pitch class-based representations are assumed to be fairly independent of other musical facets such as timbre, loudness, or noise [7, 9].

Timbre, sometimes referred to as sound color, texture, or tone quality, mainly correlates with the spectro-temporal shape of the audio signal [4, 5]. Timbre information of a given frame or beat is originally provided in the million song dataset as an array of 12 real-valued numbers. These numbers correspond to the projection of the (Fourier-based) spectro-temporal representation of the frame's signal into a set of 12 bivariate basis [2, 3]. These bivariate basis correspond to "high level abstractions of the spectral surface, ordered by degree of importance" [3]. This way, the second basis emphasizes sound brightness, the third is correlated to flatness, the fourth represents sounds with a strong attack, etc. For completeness, the first basis represents the average loudness [5] of the segment. Therefore, the 12-dimensional vectors can be split into an 11-dimensional timbre component and a unidimensional loudness (or volume) component. Notice that including the average loudness in the original timbre representation implies a certain degree of independence of the two components. Since, for perceptual reasons, the frequency resolution of the spectro-temporal representation is intentionally low [2], the obtained timbre and loudness components can be also assumed to be quite independent of pitch.

1.3 Encoding

To facilitate the study of the provided music descriptions, and in order to apply current methods and concepts from statistical physics and complex networks, we need to encode our representations into discrete elements [11]. Hence the use of the term *codewords* to denote our main analysis units. The most important aspects of the followed encoding are summarized in Table 1. The details follow.

To ease the interpretation of pitch codewords, we opt for a binary discretization of pitch descriptions, therefore only accounting for presence or absence of a given pitch class. This way, 12-dimensional descriptions are reduced to $2^{12} = 4,096$ codewords. We use a single threshold set to 0.5 and map the original pitch vector values to 0 or 1, depending on whether they are below or above the threshold, respectively. The value of 0.5 is near the mean value of the considered vector components and other arbitrary numbers close to it provided no apparent change in the results of our analysis. Binary quantization roughly resembles the all-or-none behavior of neurons and neuronal ensembles [13]. Furthermore, this encoding method is akin to methods used, for instance, in automatic audio identification [14] or in cochlear implant sound processors [15].

Before discretization, pitch descriptions of each track are automatically transposed to an equivalent main tonality, such that all pitch codewords are considered within the same tonal context or key. For this process we employ a circular shift strategy [12], correlating (shifted) per-track averages to cognitively-inspired tonal profiles [16]. This strategy is commonly applied to pitch class descriptions in many music processing contexts [6, 7], specially in the retrieval

Musical facet	Pre-processing	Dimensionality	Discretization	Threshold Value(s)
Pitch	Transposition to the same tonal context [12].	12 real values (between 0 and 1).	Binary	0.5 (same value for each dimension).
Timbre	Remove the loudness component and get a sample of beat-based timbre descriptions (see text).	11 real values.	Ternary	33 and 66% quantiles of the extracted sample (different values for each dimension).
Loudness	Take the loudness component from timbre descriptions and get a sample of beat-based loudness descriptions (see text).	1 real value.	300 steps	Equal-sized steps in the range of the extracted sample.

Table 1: Summary of the encoding process for deriving music codewords from the beat-based descriptions provided in the million song dataset. In total we have 4,096 possible pitch codewords, 177,147 possible timbre codewords, and 300 possible loudness codewords.

of versions of the same musical composition [8] and in automatic chord/key estimation [9].

Compared to pitch, timbre is believed to have a much higher dimensionality, at least perceptually [4]. To account for this, and also in order to better match the underlying distribution of the timbre descriptions provided in the million song dataset, we make use of a ternary, equal-frequency encoding [17], providing a total of $3^{11} = 177,147$ possible timbre codewords. Thresholds are set to the 33 and 66% quantiles of a representative sample of beat-based timbre description values[6]. To construct such sample we randomly chose one million timbre vectors from the dataset such that a maximum of 8,000 vectors corresponded to the same year. In this way we controlled that no bias towards a certain year was introduced into the sample. It is worth noting here that the use of more elaborate discretization techniques, like vector quantization [18], would rely on predefined distance measures and would require a high computational load to infer thousands of codewords.

Loudness values are originally provided in decibels (dB), and limited within a range from 0 to 60 [2, 3]. To study their distribution we treat these loudness values directly as a random variable (see below). Nonetheless, in order to conform to the standard signal processing criterion [19], we subtract the loudness reference of 60 dB used in the million song dataset from them. This yields values $x \in [-60, 0]$ $\mathrm{dB_{FS}}$, where $\mathrm{dB_{FS}}$ means full-scale decibels. To study transitions between loudness values and build a complex network we use an unsupervised equal-width discretization [17] into 300 equal steps. In preliminary analysis we experimented with other discretizations (e.g. 200 steps, 300 quantiles), obtaining very similar results.

1.4 Codeword Distributions

1.4.1 Distribution Functions

As mentioned in the main text, three different types of fits are performed: discrete (pure) power laws, shifted discrete power laws, and truncated reversed log-normals. For the discrete cases, the random variable takes only integer values, which can represent the frequency of a codeword or the frequency of a degree k in the network. Then, $P(z)$ is the probability mass function, and gives the probability that the random variable takes the value z. For the shifted power law this is given by

$$P(z) = \frac{1}{\zeta(\beta, c + z_{min})(c + z)^{\beta}} \tag{1}$$

with $z = z_{\min}, z_{\min} + 1, \ldots$, where c and β are parameters ($\beta \geq 1$), and z_{min} is the minimum value of the variable for which the fit holds. We note that z_{min} takes integer values and that fulfills $c + z_{min} > 0$. The (pure) power law case is recovered by setting $c = 0$. The bivariate

[6] This sample should not be confused with the final sample used for analysis. It is just an initial sample for obtaining the 33 and 66% quantiles that will allow to threshold the music descriptions.

function $\zeta(\beta, q)$ is the Hurwitz zeta function,

$$\zeta(\beta, q) = \sum_{n=0}^{\infty} \frac{1}{(q+n)^{\beta}}, \tag{2}$$

which yields the Riemann zeta function for $q = 1$, i.e. $\zeta(\beta, 1) = \zeta(\beta)$. At several points the procedure will require the computation of the Hurwitz zeta function, which is done by means of an algorithm based on the Euler-Maclaurin series [20].

For the loudness values (denoted by x), z is a real variable, defined as $z = -x$, as well as its minimum and maximum values $z_{\min}$ and $z_{\max}$. Although we use the same notation as for discrete variables, for a continuous variable the function $P(z)$ will not be the probability mass function but the probability density, given in this case by a truncated log-normal,

$$P(z) = \sqrt{\frac{2}{\pi\sigma^2}} \left[\operatorname{erf}\left(\frac{\ln z_{\max} - \mu}{\sqrt{2}\sigma} \right) - \operatorname{erf}\left(\frac{\ln z_{\min} - \mu}{\sqrt{2}\sigma} \right) \right]^{-1} \frac{1}{z} \exp\left(-\frac{(\ln z - \mu)^2}{2\sigma^2} \right) \tag{3}$$

with $0 \leq z_{\min} \leq z \leq z_{\max}$ and where

$$\operatorname{erf}(y) = 2\pi^{-1/2} \int_0^y e^{-u^2} du \tag{4}$$

is the error function (implemented as in Press et al. [21]). The adjective 'reverse' used in the main text refers to the fact that $P(x)$ is the mirror image of the true (truncated) log-normal distribution for the variable $z = -x$. Note that μ and σ do not correspond to the mean and standard deviation of the data, but to those of the underlying non-truncated normal distribution.

1.4.2 Fitting Procedure

The fitting procedure for power laws and log-normals is based on the one by Clauset et al. [22]. However, we modify the way in which the fitting range is found since that algorithm was shown to reject the power-law hypothesis for power-law simulated data in some specific cases [23]. Additional variations are also introduced regarding the discreteness of the variable. A comprehensive account for the case of the continuous power-law fitting can be found in Peters et al. [24] (apart from this reference, other details are found in the supplementary information of Corral et al. [25]).

First, we try a fit in an arbitrary fitting range $z \geq z_{\min}$ (or $z_{\min} \leq z \leq z_{\max}$, depending on the distribution). The fitting parameters β and c (or μ and σ) are found by maximum likelihood estimation, which starts by calculating the log-likelihood function

$$\mathcal{L} = \frac{1}{N_m} \sum_{i=1}^{N_m} \ln P(z_i), \tag{5}$$

where i labels all the N_m values of the variable in the fitting range $z \geq z_{\min}$ (or $z_{\min} \leq z \leq z_{\max}$). As the fitting range is fixed, the data are fixed too, and the log-likelihood is only a function of the parameters of the distribution. Its maximization yields their maximum likelihood estimation.

The second step is to measure the discrepancy between the fit and the data. For that we use the Kolmogorov-Smirnov (KS) statistic or distance, which yields a measurement of the separation between the fit and the empirical data [21]. Nevertheless, from this distance alone we cannot evaluate the goodness of the fit. For that purpose, the third step generates synthetic datasets in the considered fitting range by computer simulation [26], with the same number of data N_m as in the original set, and using the parameters obtained for it by maximum likelihood estimation. The same procedure as the one just explained for that case is applied to each of these simulated data sets, yielding a series of maximum likelihood estimated parameters, fitting distributions, and KS distances [22, 27]. It is the latter which allows us to quantify the value of the empirical KS distance. Indeed, the $p-$value of the fit will be computed as the number of simulated data sets with KS distance larger than the empirical one divided by the total number of simulations.

Up to now we have assumed that the fitting range is fixed. In order to select the best fitting range we perform the same procedure for a large sample and select the one which contains the largest number of data N_m, provided that its $p-$value is above 0.25. This concludes the procedure.

1.5 Codeword Networks

1.5.1 Network Building

To study the transitions between codewords, we build a complex weighted directed network for pitch, timbre, and loudness descriptions by representing each codeword by a node and placing a directed link between any two beat-consecutive codewords (self-links from a codeword to itself are not considered). Link weights ω_{ij} are set to the frequency of occurrence of codeword transitions (when there is no link between codewords i and j, we set $\omega_{ij} = 0$). A preliminary analysis with pitch networks shows that adjacency matrices are almost symmetric for the majority of transitions, i.e. $\omega_{ij} \approx \omega_{ji}$. Link weight correlations for a given network are always above 0.95, and the average value across all considered years is 0.98 ± 0.01. Therefore, we can safely use undirected versions of the networks by removing link directionality and summing up the weights in the two directions, $\omega_{ij} \rightarrow \omega_{ij} + \omega_{ji}$. The fact that transitions between pitch codewords are symmetric may be surprising at first sight, specially given that transition asymmetries are present in classical music [28, 29]. However, evidence that such asymmetries are not present in contemporary popular music has already been reported for a reduced set

of manually annotated pieces [30]. Our analysis confirms quantitatively the same evidence at a large-scale for the pitch networks and also for the timbre and loudness ones, a result never recognized before.

1.5.2 Metrics

In codeword networks of size N, each node $i = 1, \ldots, N$ is characterized by its degree k_i, measuring the number of neighbors with other codewords, and its strength s_i, measuring the total weight associated with the connections with such neighbors, i.e. $s_i = \sum_{j=1}^{N} \omega_{ij}$. However, when ω_{ij} counts the number of transitions between codewords i and j, codeword strength s is the same as codeword frequency z (except for self-loops), which has already been analyzed in the previous section and in the main text. This leaves us with the task of studying the bare network topology. To do so, we use the following metrics.

1. The most fundamental is the degree distribution $P(k)$, measuring the probability that a randomly chosen node has degree k. This distribution is characterized by its average degree $\langle k \rangle$ and, when the network is scale-free, i.e. $P(k) \propto k^{-\gamma}$, by the critical exponent γ. In some cases, we use the median of $P(k)$ instead of the average degree as it is a measure which is independent of the presence and particular value of a few hubs, which may change from year to year.

2. Correlations between pairs of connected nodes are evaluated with the assortativity coefficient normalized with respect to a randomized network, Γ. Specifically,

$$\Gamma = \frac{\langle kk' \rangle}{\langle kk' \rangle_{rand}}. \tag{6}$$

   In this equation, the term in the numerator is an average taken over pairs of connected nodes in the original network, i.e. $\langle kk' \rangle = \sum_{k,k'} kk' P(k,k')$, where $P(k,k')$ is the probability that a randomly chosen link connects two nodes of degrees k and k'. The denominator is the same average but in a randomized version of the original network. This randomization is performed by swapping pairs of links chosen at random with the constraint that multiple links and self-connections are forbidden [31]. Notice that this procedure preserves nodes' degrees and, thus, the randomized network obtained can be considered as the maximally random network with that particular degree distribution. For a network with nodes' degrees below the critical value $\sqrt{\langle k \rangle N}$ [32], the probability $P(k,k')$ for the randomized network factorizes as $P(k,k') = kk' P(k) P(k') / \langle k \rangle^2$ and, therefore, $\langle kk' \rangle_{rand} = \langle k^2 \rangle^2 / \langle k \rangle^2$. However, scale-free networks with $\gamma < 3$ always have nodes violating this condition. As a consequence, maximally random scale-free graphs have structural degree-degree correlations that cannot be eliminated [32]. Therefore, Γ measures not absolute degree-degree correlations but those correlations present in the

real network with respect to the minimum level of correlations allowed by the degree sequence. Keeping this in mind, values of $\Gamma > 1$ indicate a tendency towards connecting nodes with similar degrees, a typical pattern of social networks, and values $\Gamma < 1$ indicate the tendency of high degree nodes to connect to low degree ones and vice versa, a pattern usually observed in technological and biological networks [33].

3. The local order of the network is measured by the clustering coefficient C. This coefficient is obtained as an average of the local clustering coefficient of all nodes of degrees above 1, where the local clustering coefficient of node i, c_i is

$$c_i = \frac{2T_i}{k_i(k_i - 1)} \tag{7}$$

and T_i is the number of triangles attached to node i. Random graphs have a vanishing clustering coefficient at the limit of very large networks, whereas the majority of real world networks show very large levels of clustering. In the case of random graphs with a given degree sequence, it can be shown that C takes the value [33]

$$C = \frac{\langle k(k-1)\rangle^2}{N\langle k\rangle^3}. \tag{8}$$

However, when the network is scale-free, the term in the numerator diverges with the system size and, for finite networks, C can still take very large values. In such cases, it is difficult to claim any local ordering of the systems based only upon C. An alternative approach is to remove the most connected nodes of the network and then re-compute C in this filtered network. Any local order present in the system will imply high values of C even for strongly filtered networks.

4. The global properties of the network are probed by means of the average shortest path length l, which is computed as follows. For each pair of nodes in the network, i and j, belonging to the same connected component, we find the shortest path between them, measured in number of network hops l_{ij}. The average shortest path length is then

$$l = \frac{2}{N_{cc}(N_{cc} - 1)} \sum_{i,j=1}^{N_{cc}} l_{ij}, \tag{9}$$

where N_{cc} is the number of nodes in the largest connected component of the network. Small-world networks [34] have high levels of clustering (well defined local structure) and small values of l, typically $\mathcal{O}(\ln N_{cc})$.

1.5.3 Analysis Details

Pitch networks The topology of the pitch networks reveals an extremely dense network with $N = 4096$ nodes and an exponential degree distribution of average degree $\langle k\rangle = 271$

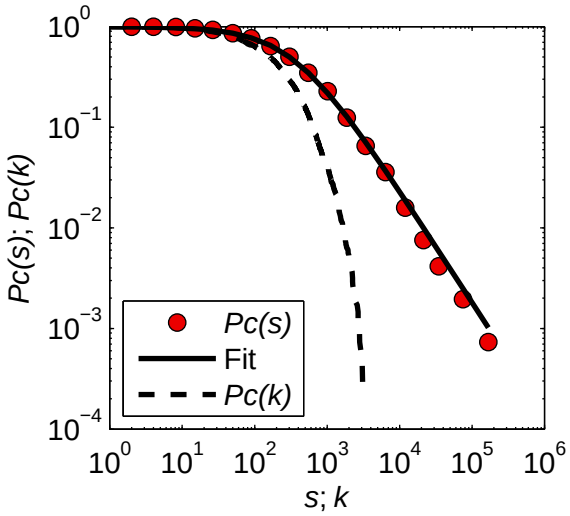

Figure 2: Complementary cumulative distribution functions of strength s and degree k of nodes in the original pitch network without any filtering procedure applied (data correspond to the pitch network for the central year 1992). The strength distribution can be fitted by a function of the form $P_c(s) = (1 + s/c)^{1-\beta}$ with $\beta = 2.11$ and $c = 357.14$.

as for the year 1992 (Fig. 2). However, the majority of links in this network carry a very small fraction of the total weight of the system. The complementary cumulative distribution function of node strength $P_c(s) = \sum_{s'=s} P(s')$ spans five orders of magnitude and can be well fitted by a shifted power law of the form $P_c(s) = (1 + s/c)^{1-\beta}$ with $\beta \approx 2.1$. Besides, the strength of a node and its degree are super-linearly correlated (Fig. 3). This implies that weights are not randomly distributed among the links of the network but are correlated with the network topology. This observation allows us to apply a sensible filter to the network topology that, using information associated with the weights, reduces the number of links keeping only the backbone of the system: the disparity filter [35]. The disparity filter is a local filter that compares the weights of all links attached to a given node against a null model, keeping only those links that cannot be explained by the null model under a certain confidence level. The procedure is repeated twice for each link and the link is finally kept if it is relevant for at least one of the attached nodes.

To apply the filter, we have to specify the null model. Under the null hypothesis, the strength of a given node is homogeneously distributed among all its links. Therefore, the probability that in a node of degree k one link has a fraction u of the node's strength is

$$\rho(u) = (k-1)(1-u)^{k-2}. \tag{10}$$

A link accounting for a fraction of the total node's strength is considered relevant if the

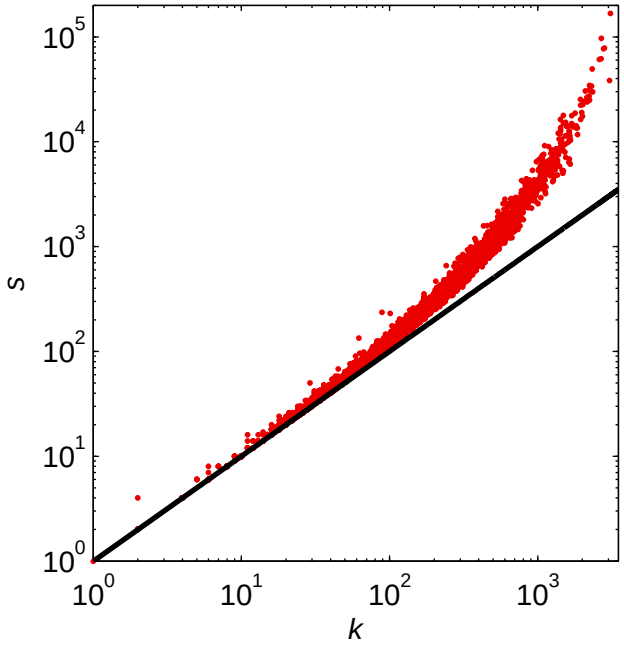

Figure 3: Scattered plot of the strength vs. degree in the original pitch network without any filtering procedure applied.

probability of observing such value under the null model is smaller than α, where $1-\alpha$ is the confidence level. Therefore, a link of weight ω_{ij} attached to two nodes of degrees k_i and k_j and strengths s_i and s_j will be preserved iff

$$(k_i - 1)\int_{\omega_{ij}/s_i}^{1}(1-u)^{k_i-2}du < \alpha \tag{11}$$

or

$$(k_j - 1)\int_{\omega_{ij}/s_j}^{1}(1-u)^{k_j-2}du < \alpha. \tag{12}$$

Fig. 4 shows the effect of applying the disparity filter with different values of α. As can be seen, the degree distribution changes very quickly from an exponential distribution with an extremely high average degree to a scale-free distribution with a stable exponent around $\gamma \approx 2.2$. Interestingly, the filter does not significantly affect the strength distribution. In the article, we consider always filtered pitch transition networks with $\alpha = 0.01$, corresponding to a confidence level of 99%, which results in a scale-free network of average degree $\langle k \rangle \approx 12$. The value of the exponent γ makes this network very heterogeneous. Thus, as explained previously in Sec. 1.5.2, properties such as clustering or assortativity may be strongly affected. To distinguish real trends from effects purely induced by heterogeneity, we remove the 10 most connected nodes in the original network. Clustering and assortativity are then measured in

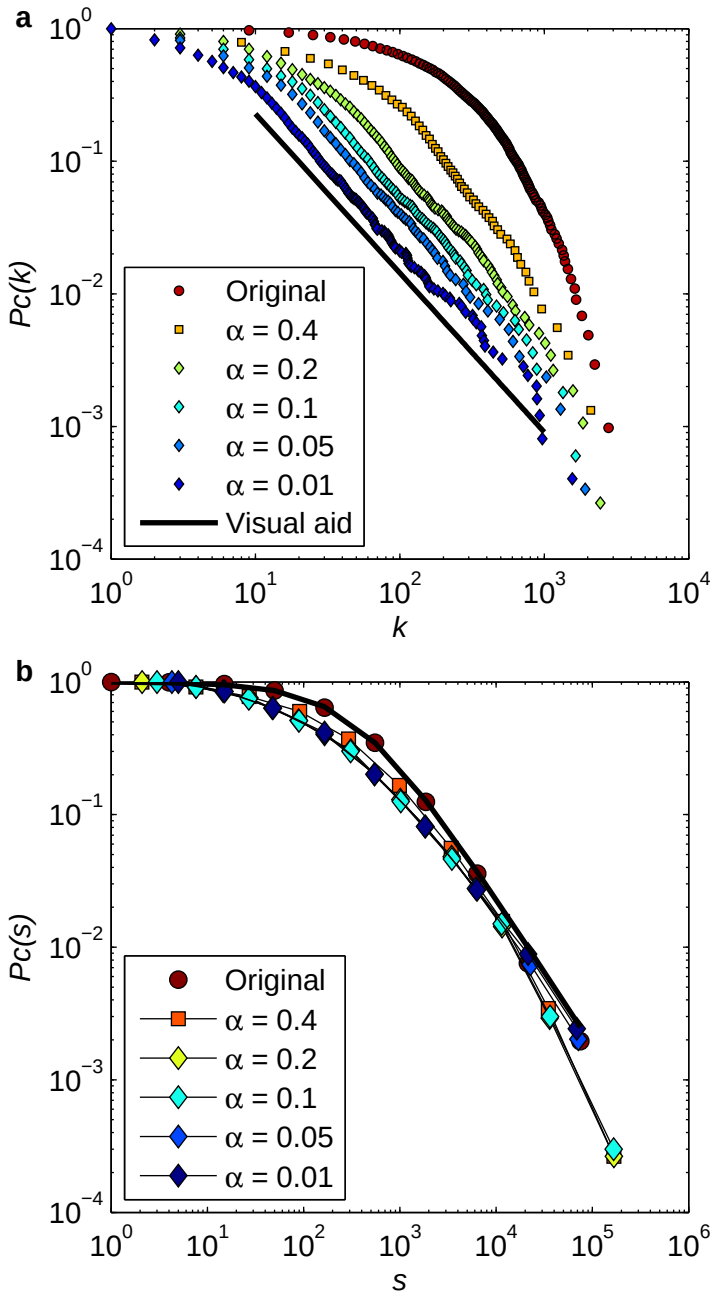

Figure 4: Applying the disparity filter to pitch networks: (a) Complementary cumulative degree distribution $P_c(k) = \sum_{k'=k} P(k')$ of filtered networks for different values of α. This distribution becomes a power law with a stable exponent $\gamma - 1 \approx 1.2$, even for very large values of α. (b) Interestingly, the filter does not significantly affect the strength distribution $P_c(s)$.

both versions of the network (Figs. 7 and 8). This analysis confirms the clustered nature of the pitch transition network and its disassortative character.

Timbre networks Timbre transition networks are sparse networks with $N \approx 175,000$ nodes and average degree $\langle k \rangle \approx 12$. Most network weights are very small (typically 1 or 2), so they do not carry relevant information. Because of such sparseness and the absence of relevant information in the link weights, the application of the disparity filter to timbre networks was deemed unnecessary. Besides, the critical value $\sqrt{\langle k \rangle N}$ defining the onset of structural correlations is larger than the maximum degree observed in the network, meaning that network heterogeneity is not a key aspect of the network topology.

Loudness networks After applying the disparity filter with $\alpha = 0.01$ to loudness networks we usually obtain $N \approx 250$ nodes and average degrees $\langle k \rangle \approx 13.5$. The disparity filter was necessary here since, as with the case of pitch, the network was very dense and the majority of links carried a very small fraction of the total weight of the system. The degree distribution is close to a homogeneous distribution, with degrees ranging from $k = 2$ to $k = 20$, similar to what would happen in a disordered lattice embedded in a low dimensional space. This is a direct consequence of loudness being a one-dimensional quantity.

1.6 Linear Regressions

To assess trends in parameters and metrics over the years, we perform an ordinary least squares linear regression [36, 37] and report the slope found (Table 2). Statistical significance is evaluated under the null hypothesis that the slope is different from zero, using a two-tail t-test and $p < 0.01$ (if $p > 0.05$ we deem the slope not statistically significant).

Musical facet	Variable	Slope	p-value	t-statistic	R^2	Significant
Pitch	β	0.002	0.097	1.66	0.005	No
Loudness	median(x)	0.131	$2.4 \cdot 10^{-96}$	25.84	0.554	Yes
Loudness	$\lvert Q_1(x) - Q_3(x) \rvert$	0.002	0.321	0.99	0.002	No

Table 2: Summary of the linear regressions mentioned in the main text.

2 Supplementary Figures

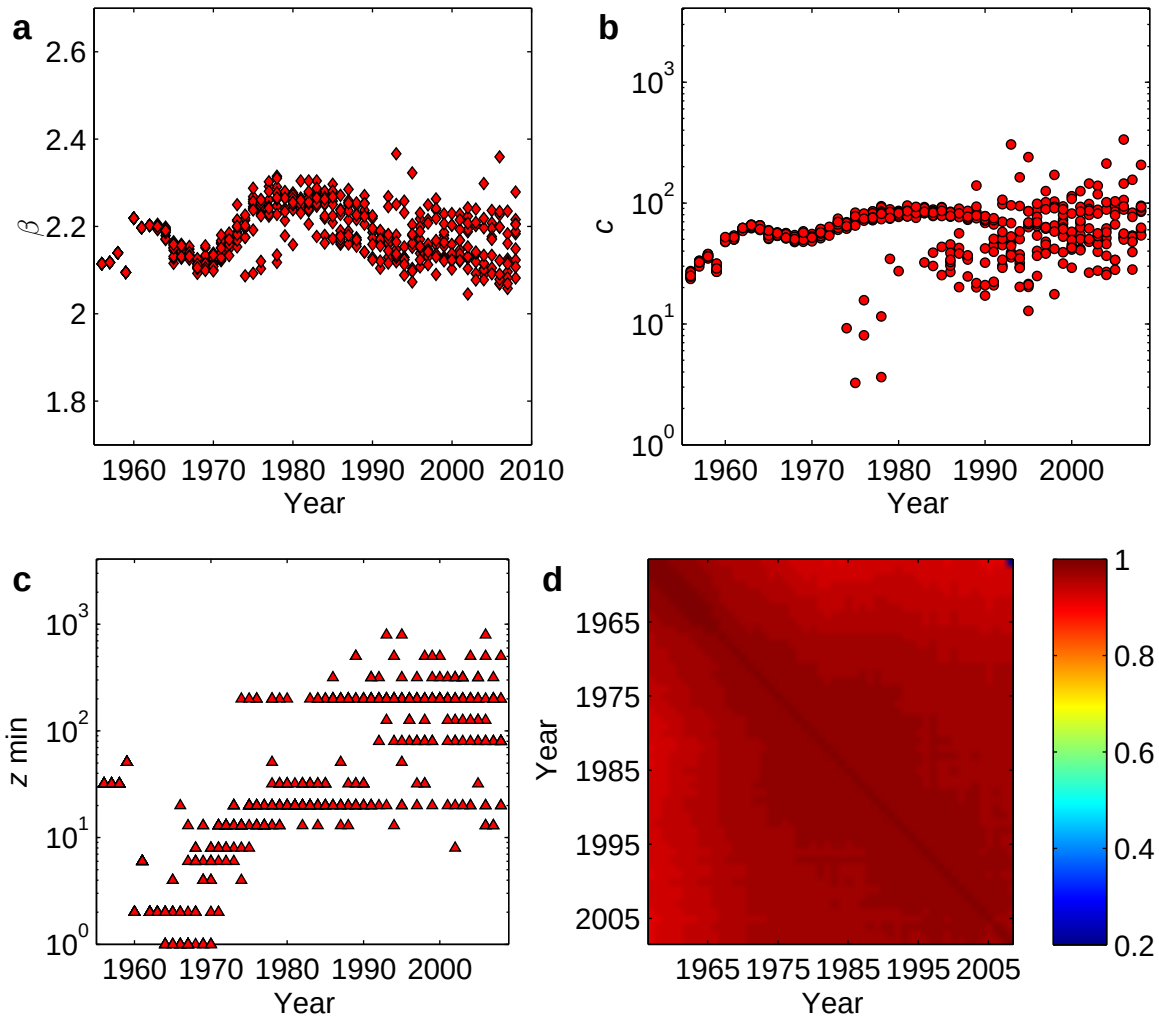

Figure 5: Pitch codeword distributions. (a) Fitted power law exponents β. (b) Fitted parameters c. (c) Fitted thresholds $z_{\min}$. (d) Spearman's rank correlation coefficient for all pairwise distributions. As mentioned in the main text, correlations are all above 0.92 (we use the same color bar as in the main text figure for the sake of comparison).

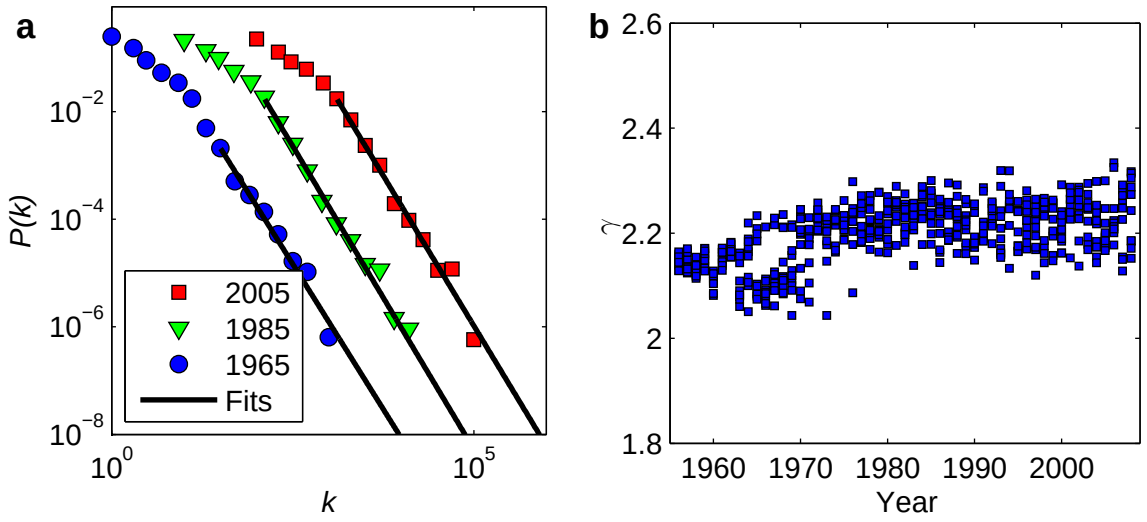

Figure 6: Pitch transition networks. (a) Examples of the degree distribution $P(k)$ and its fitted power law. For ease of visualization, curves are chronologically shifted by a factor of 10 in the horizontal axis. (b) Degree distribution exponents γ.

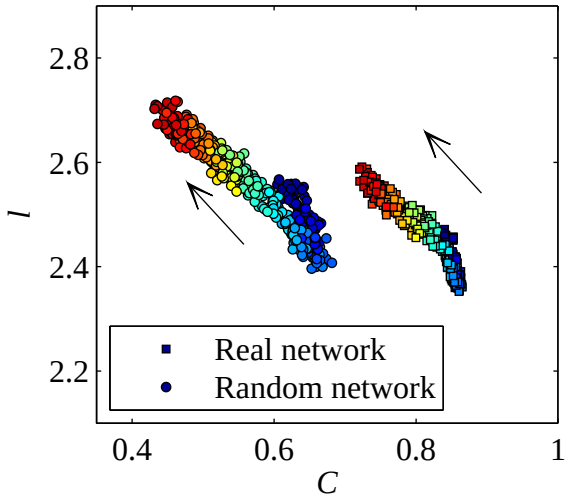

Figure 7: Pitch transition networks. Average shortest path length l versus clustering coefficient C for pitch networks. Arrows indicate chronology (red and blue colors indicate values for more and less recent years, respectively). Results considering all nodes of the network, including the 10 biggest hubs (see Materials and Methods).

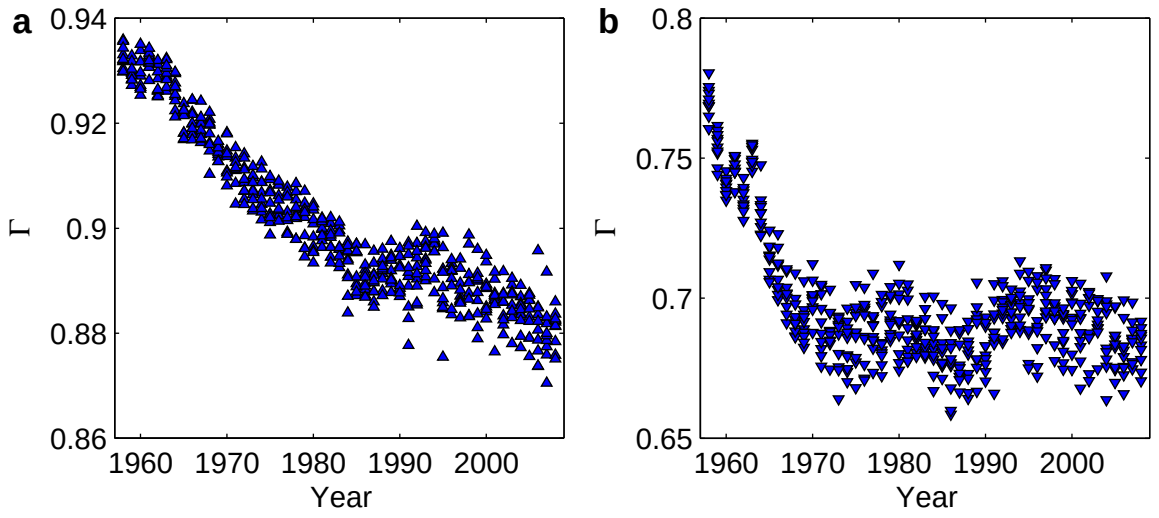

Figure 8: Pitch transition networks. Assortativeness with respect to random Γ with (a) and without (b) the 10 biggest hubs (see Materials and Methods). As mentioned in the main text and in Sec. 1.5.2, a value of $\Gamma < 1$ indicates that well-connected nodes (codewords) are less likely to be connected among them.

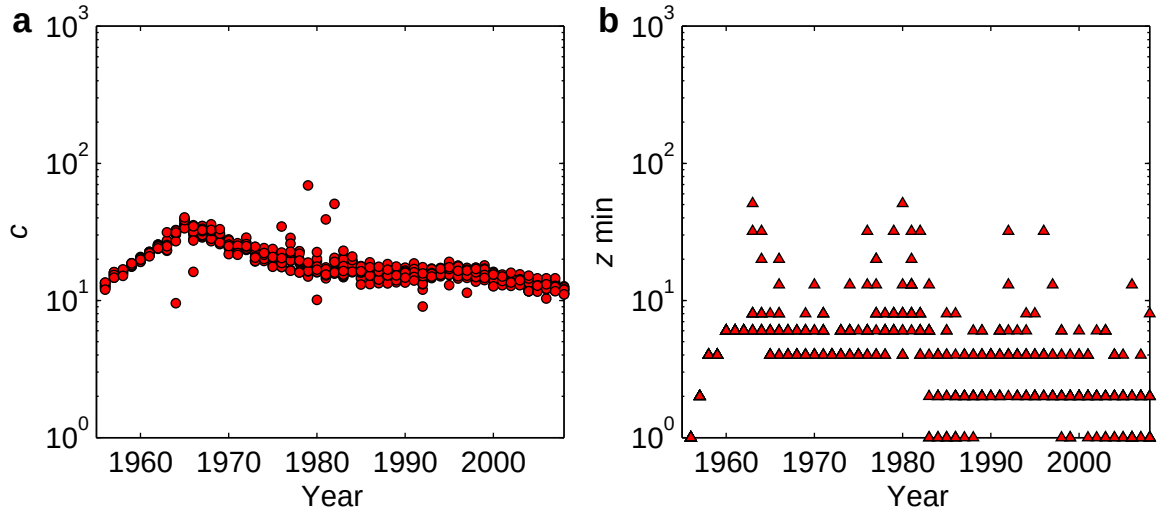

Figure 9: Timbre codeword distributions. Fitted parameters c (a) and $z_{\min}$ (b). The fits for the exponent β were depicted in the main text.

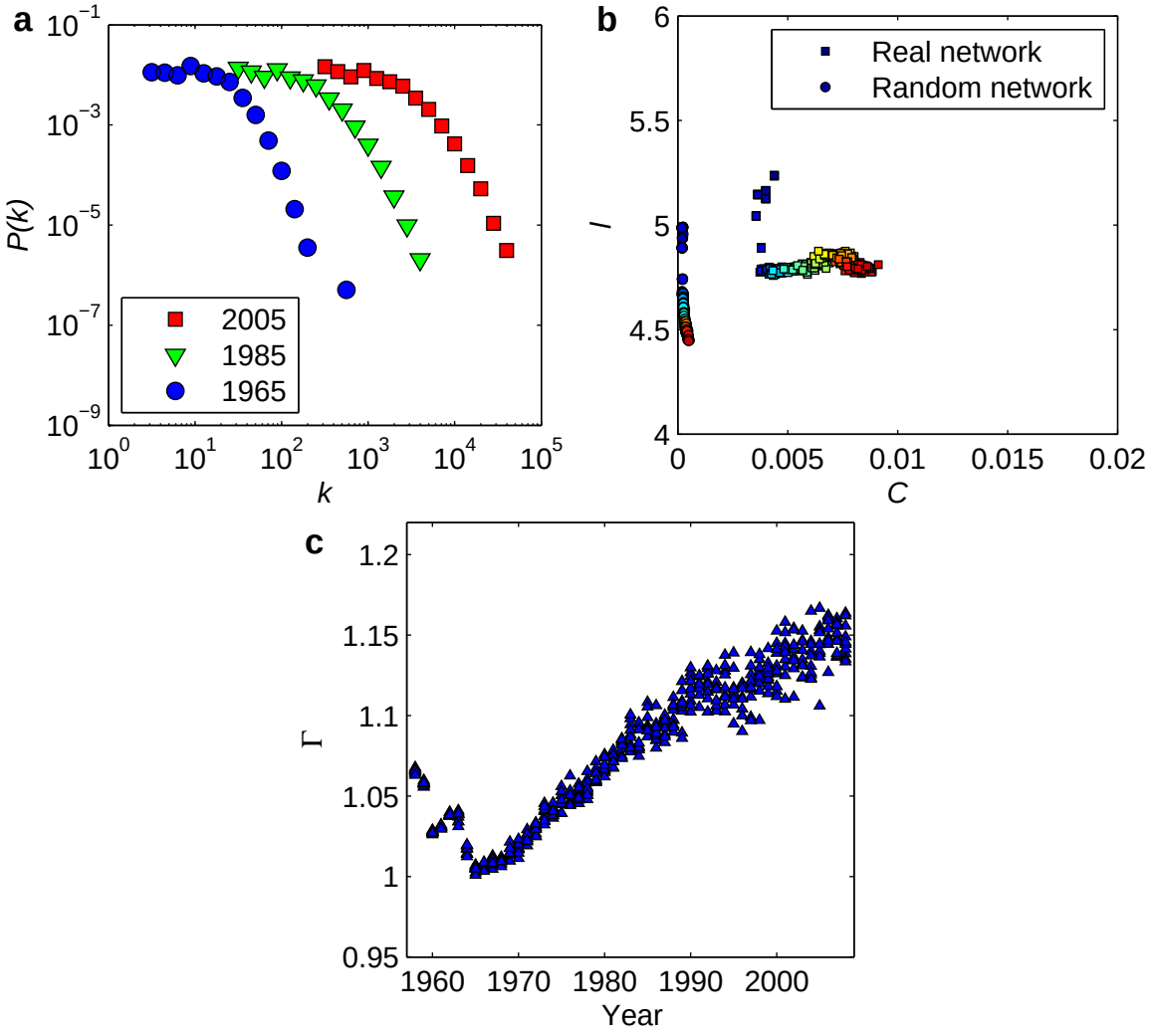

Figure 10: Timbre transition networks. (a) Examples of the degree distribution. For ease of visualization, curves are chronologically shifted by a factor of 10 in the horizontal axis. (b) Average shortest path l versus clustering coefficient C. Red and blue colors indicate values for more and less recent years, respectively. Notice the small range of C, which is always below 0.01. (c) Assortativeness with respect to random Γ.

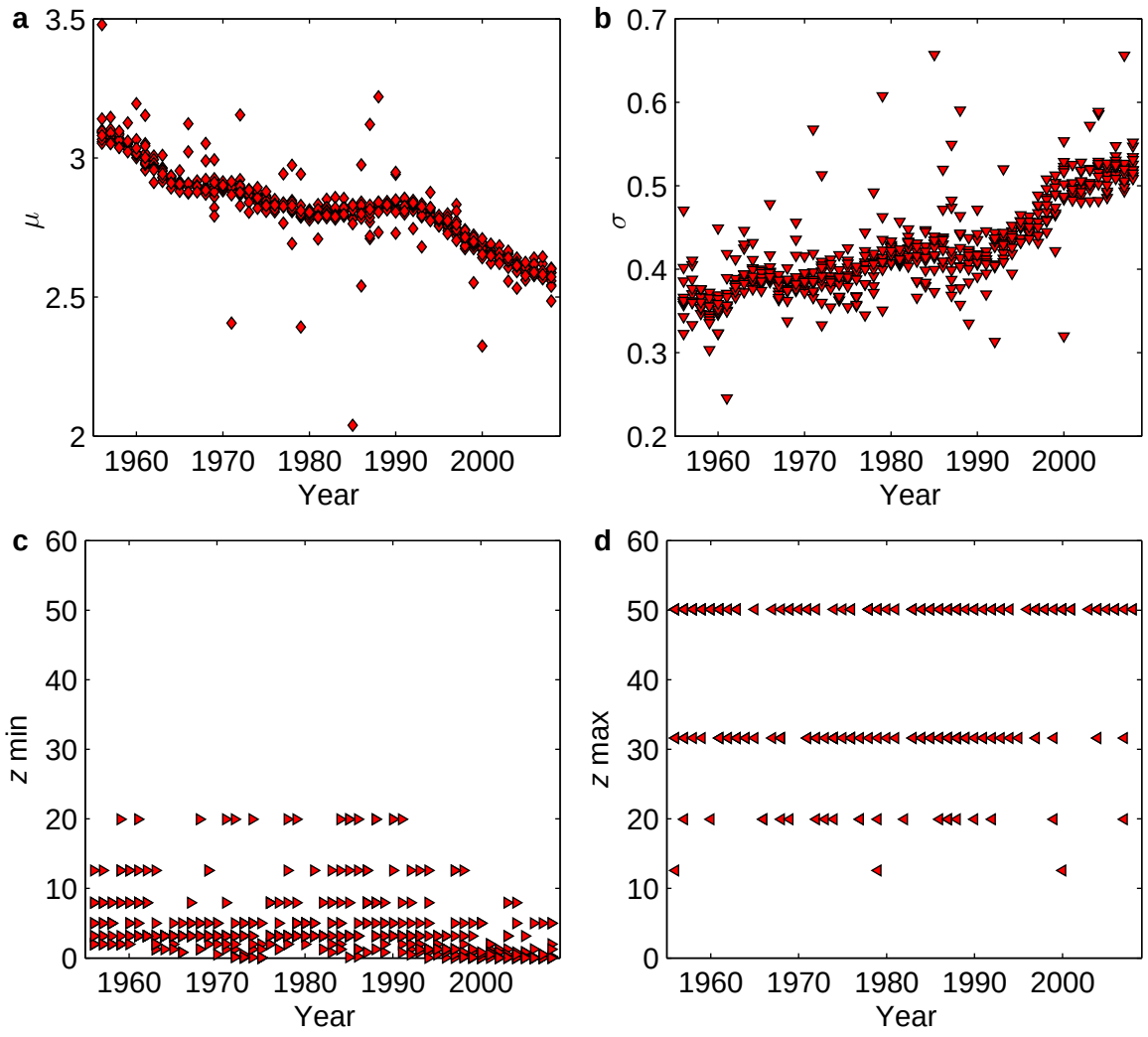

Figure 11: Loudness distributions. Fit parameters μ (a), σ (b), $z_{\min}$ (c), and $z_{\max}$ (d).

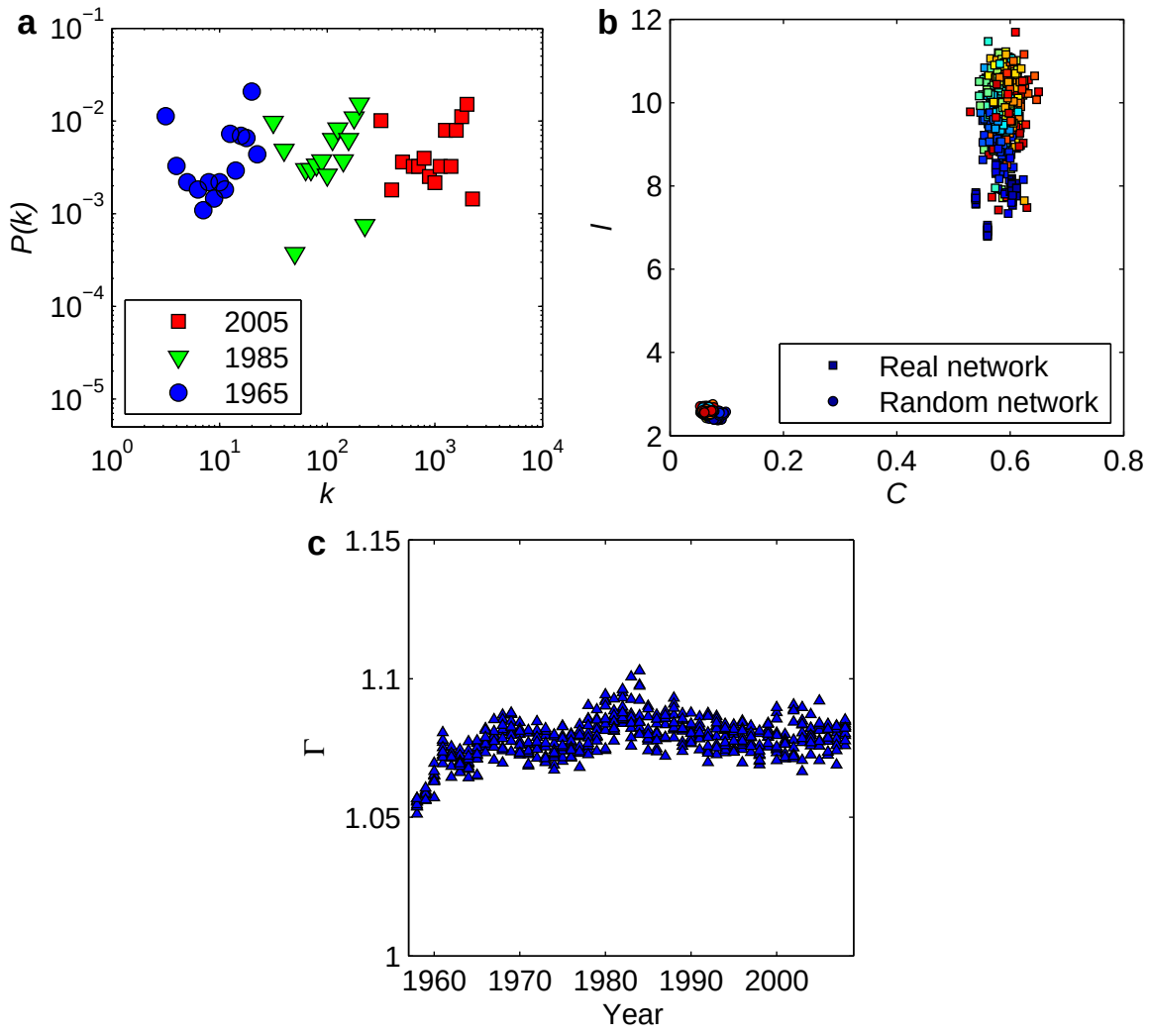

Figure 12: Loudness networks. (a) Examples of the degree distribution for loudness networks. For ease of visualization, curves are chronologically shifted by a factor of 10 in the horizontal axis. (b) Average shortest path l versus clustering coefficient C. Red and blue colors indicate values for more and less recent years, respectively. (c) Assortativeness with respect to random Γ.

References

[1] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B. & Lamere, P. The million song dataset. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 591–596 (2011).

[2] Jehan, T. *Creating music by listening.* Ph.D. thesis, Massachussets Institute of Technology, Cambridge, USA (2005).

[3] Jehan, T. The Echo Nest Analyze documentation. Tech. Rep., The Echo Nest (2010). URL `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`.

[4] Bregman, A. L. *Auditory scene analysis: the perceptual organization of sound* (MIT Press, Cambridge, USA, 1990).

[5] Ball, P. *The music instinct: how music works and why we can't do without it* (Bodley Head, London, UK, 2010).

[6] Casey, M. A. *et al.* Content-based music information retrieval: current directions and future challenges. *Proc. of the IEEE* **96**, 668–696 (2008).

[7] Müller, M., Ellis, D. P. W., Klapuri, A. & Richard, G. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing* **5**, 1088–1110 (2011).

[8] Serrà, J., Gómez, E. & Herrera, P. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In Raś, Z. W. & Wieczorkowska, A. A. (eds.) *Advances in Music Information Retrieval*, vol. 274 of *Studies in Computational Intelligence*, chap. 14, 307–332 (Springer, Berlin, Germany, 2010).

[9] Gómez, E. *Tonal description of music audio signals.* Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain (2006).

[10] Paulus, J., Müller, M. & Klapuri, A. Audio-based music structure analysis. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, 625–636 (2010).

[11] Haro, M., Serrà, J., Herrera, P. & Corral, A. Zipf's law in short-time timbral codings of speech, music, and environmental sound signals. *PLoS ONE* **7**, e33993 (2012).

[12] Serrà, J., Gómez, E. & Herrera, P. Transposing chroma representations to a common key. In *Proc. of the Int. CS Conf. on the Use of Symbols to Represent Music and Multimedia Objects*, 45–48 (2008).

[13] Bethge, M., Rotermund, D. & Pawelzik, K. Second order phase transition in neural rate coding: binary encoding is optimal for rapid signal transmission. *Physical Review Letters* **90**, 088104 (2003).

[14] Haitsma, J. & Kalker, T. A highly robust audio fingerprinting system. In *Proc. of the Conf. on Music Information Retrieval (ISMIR)*, 107–115 (2002).

[15] Wilson, B. S. *et al.* Better speech recognition with cochlear implants. *Nature* **352**, 236–238 (1991).

[16] Krumhansl, C. L. *Cognitive foundations of musical pitch* (Oxford University Press, Oxford, UK, 1990).

[17] Cios, K. J., Pedrycz, W., Swiniarski, R. W. & Kurgan, L. A. *Data mining: a knowledge discovery approach* (Springer, Berlin, Germany, 2007).

[18] Linde, Y., Buzo, A. & Gray, R. An algorithm for vector quantizer design. *IEEE Trans. on Communications* **28**, 84–95 (1980).

[19] Oppenheim, A. V., Schafer, R. W. & Buck, J. R. *Discrete-time signal processing* (Prentice-Hall, Upper Saddle River, USA, 1999), 2nd edn.

[20] Vepstas, L. An efficient algorithm for accelerating the convergence of oscillatory series, useful for computing the polylogarithm and Hurwitz zeta functions. *Numerical Algorithms* **47**, 211–252 (2008).

[21] Press, W. H., Teukolsky, S. A. & Vetterling, W. T. *Numerical recipes in Fortran* (Cambridge University Press, Cambridge, UK, 1992).

[22] Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review* **51**, 661–703 (2009).

[23] Corral, A., Font, F. & Camacho, J. Non-characteristic half-lives in radioactive decay. *Physical Review E* **83**, 066103 (2011).

[24] Peters, O., Deluca, A., Corral, A., Neelin, J. D. & Holloway, C. E. Universality of rain event size distributions. *Journal of Statistical Mechanics: Theory and Experiment* **2010**, P11030 (2010).

[25] Corral, A., Ossó, A. & Llebot, J. E. Scaling of tropical-cyclone dissipation. *Nature Physics* **6**, 693–696 (2010).

[26] Devyroe, L. *Non-uniform random variate generation* (Springer, New York, USA, 1986).

[27] Malmgrem, R. D., Stouffer, D. B., Motter, A. E. & Amaral, L. A. N. A Poissonian explanation for heavy tails in e-mail communication. *Proc. of the National Academy of Sciences of the USA* **105**, 18153–18158 (2008).

[28] Lerdahl, F. & Jackendoff, R. *A generative theory of tonal music* (MIT Press, Cambridge, USA, 1983).

[29] Temperley, D. *Music and probability* (MIT Press, Cambridge, USA, 2007).

[30] De Clercq, T. & Temperley, D. A corpus analysis of rock harmony. *Popular Music* **30**, 47–70 (2011).

[31] Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).

[32] Boguñá, M., Pastor-Satorras, R. & Vespignani, A. Cut-offs and finite size effects in scale-free networks. *European Physical Journal B* **38**, 205–210 (2004).

[33] Newman, M. E. J. *Networks: an introduction* (Oxford University Press, Oxford, UK, 2010).

[34] Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[35] Serrano, M. A., Boguñá, M. & Vespignani, A. Extracting the multiscale backbone of complex weighted networks. *Proc. of the National Academy of Sciences of the USA* **106**, 6483–6488 (2009).

[36] Chatterjee, S. & Hadi, A. S. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science* **1**, 379–416 (1986).

[37] Wasserman, L. *All of statistics: a concise course in statistical inference* (Springer, Berlin, Germany, 2003).