The Power of Music: Searching for Power-Laws in Symbolic Musical Data

Dimitrios Rafailidis   Yannis Manolopoulos

Department of Informatics, Aristotle University
54124 Thessaloniki, GREECE
{draf, manolopo}@csd.auth.gr

Abstract

Recently, many research contributions focus on showing that specific real-life phenomena or measured quantities of our everyday life obey a power-law, a scale free distribution. Inspired by the numerous studies in power-laws, in this paper, we explore the existence of power-law relationships in musical data. We provide the context where such power-laws do exist, and we give experimental results that validate our assumptions. The results are based on several musical collections related to classic, European and Asian music.

Keywords: Power laws, repeating patterns, symbolic musical data

1. Introduction

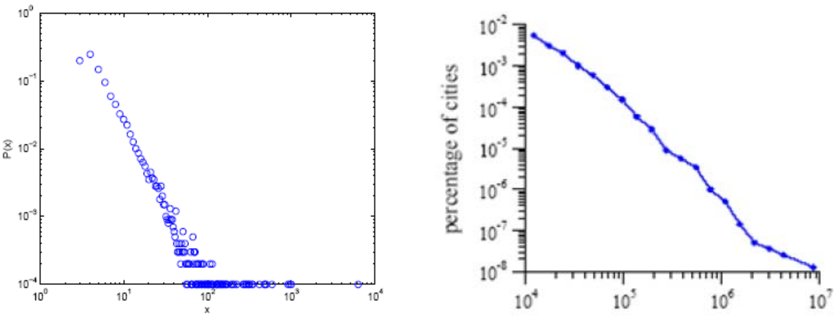
In contrast to many measurements that seem to assume values around a specific mean value (e.g., distribution of the heights of humans, distribution of car speeds in a motorway [Newman (2006)], there are many other cases where the measured values follow a power-law. Essentially, this means that although the vast majority of the measurements assume relatively small values, there is a part of the population that assumes significantly larger values than the usual measurements. A random variable obeys a power-law distribution, if the probability density function is given by the following formula, where *a* and *C* are real-valued constants and *p(x)* is the probability of an event to occur if the value of variable *x* is equal to a certain value:

$$p(x) = Cx^{-a} \quad \textbf{(1)}$$

Constant *a* is the exponent (or slope) of the power-law, and it is the most important part. Constant *C* is less important, since it can be easily calculated by enforcing that all values of *p(x)* must sum to unity (this will be clarified later).

Similar to the power-law is the Zipf law. The main difference is that in the case of the Zipf law, the rank-frequency plots are used which are equivalent to the cumulative distribution of the variable under study. In fact, if a variable obeys a power law distribution with exponent *a* then the corresponding slope for the Zipf law will be *a*-1.

Figure 1(a) depicts a power-law distribution of synthetic data, reliant to a power-law generator from [Newman (2006)], with 10.000 data and slope *a*=2. Figure 1(b) is the plot of the population (*x* axis) with respect to the percentage of cities (*y* axis), using as data the population of US cities. It is obvious that the figures are different because of the definition of the function. Data in Zipf law appear in an almost straight line, but in power-law they appear more scattered, especially at the end of the graph, which is called the tail. The long tail that is formed in both graphs represents data with much larger value, in Figure 1(b) are cities with much larger population.



(a) *Power-law distribution of synthetic data*

(b) *Zipf-law distribution of US cities population*

Figure 1. *Plots in log-log scale of power law and Zipf law distributions*

The existence of a power law reveals the fact that the data are scale-free, i.e. the statistical properties of the fraction resemble those of the whole data set [Faloutsos et. al. (1991)]. Thus, the existence of a power-law is a strong evidence of the self-similarity properties of the data under study. To decide if a specific distribution follows a power-law, Zipf law or not, there are two important issues that be considered: (i) the observed data should fit to a power-law distribution for specific values of *a* and *C*, and (ii) the exponent of the power-law should assume values between 2 and 3 (between 1 and 2 for Zipf law), according to the majority of power-laws met in nature [Newman (2006)] (however in some cases, values near 1.8, 1.9 or 3.1, 3.2 are also accepted). In the sequel, we give some illustrative examples of variables that obey a power-law distribution:

- Word frequency: In [Estoup (1916)] the author has observed that the frequency with which words are used appears to follow a power law.
- Web hits: The number of hits received by web sites during a single day from a subset of the users of the AOL Internet service provider, obeys a power-law [Adamic, Huberman (2000)].

- Magnitude of earthquakes: the magnitude of earthquakes that have occurred in California between 1/1910 and 5/1992, as recorded in the Berkeley Earthquake Catalog, follows a power-law. The power-law relationship in the earthquake distribution is a relationship between amplitude and frequency of occurrence [Newman (2006)].
- Diameter of moon craters: the diameter of moon craters obeys a power-law distribution, as it has been shown in [Neukum, Ivanov (1994)].
- Intensity of solar flares: it has been reported in [Lu, Hamilton (1991)] that the gamma-ray intensity of solar flares obeys a power-law.

Here, we investigate the existence of power-laws in repeating patterns of symbolic musical data. In particular, given a music collection and a set of repeating patterns in this collection, we study the probability distribution of these patterns in the collection, based on some metrics, such as the support, the pattern length and frequency. The existence of a power-law can be utilized in: (i) the quantification of the self-similarity properties of a collection, (ii) the determination of the significance of musical patterns towards more efficient music information retrieval, (iii) the comparison of different collections based on the parameters of the power-law distribution.

The rest is organized as follows. In Section 2 we describe related work on power-law discovery in musical data, whereas Section 3 contains some fundamental mathematical background related to the decision of whether a specific distribution follows a power-law or not. Section 4 contains our study regarding searching power-laws in repeating patterns of musical data. Experimental results are given and discussed in Section 5, whereas Section 6 concludes the work.

2. Related Work and Contribution

In all previous works regarding power-laws in musical data, authors discovered scale free distributions based on variables with low musicological meaning.

Voss and Clarke [Voss, Clarke (1975)] discovered that pitch and loudness fluctuation in music (classical, jazz, blues and rock radio station recorded continuously over 24 hours) follow Zipf distribution. Additionally, they tried to compose music through a program, using a Zipf distribution generator to generate individual music events.

The authors in [Hsu, Hsu (1991)] discovered that the changes of acoustic frequency of a collection of J.S. Bach's musical pieces obey a scale free distribution.

Manaris, Purewal and McCormick [Manaris et. al. (2002)] used a data set of music from different genres and composers. The variables were based on pitch and duration of musical events. Self-similarity was observed in these variables.

Zanette [Zanette (2006)] discovered that note pitch and duration of obey Zipf's law. The data set consists of music works from J.S. Bach, W.A. Mozart, C. Debussy and A. Schoenberg.

Although there are many observations for scale free distributions in musical data, there is no work investigating these issues in repeating patterns (i.e., patterns of notes that are present more than once in the corpus) of symbolic musical data. Existing research has focused in variables with low musicological concept, for example pitch, duration of single notes and changes in frequency of music signal. It is evident that a single note is not representative for a composer, because all notes can be used by any composer. We believe that repeating patterns characterize more adequately the style of a composer than low-level features can do.

3. Mathematical Background

By applying logarithms in both parts of Equation 1 we take:

$$\log(p(x)) = -a\log(x) + \log(C) \tag{2}$$

By assuming a log-log scale, the above equation corresponds to a straight line with slope *a* and shift *log(C)*. Exploring power-law relationships requires that we decide if the observed data follow a power-law distribution. Towards this goal, we need an evaluation tool to help us take this decision.

Mostly, the power-law behavior appears for values of *x* larger than a threshold x_{min}. Note that x_{min} is not the minimum observed value, but the minimum value for which the power-law behavior is observed. Since linear regression models for estimating the exponent does not provide satisfactory results [Newman (2006)], usually the maximum likelihood estimation (MLE) technique is being used [Milton, Arnold (1995)].

There are two approaches followed to determine *a* and *C* depending on whether we have a discrete or a continuous random variable. In both cases, the constant *C* is determined by noting that since *p(x)* is a probability density function, all possible values will sum to unity. By omitting any further details we obtain Equations 3 and 4 for the continuous case:

$$a = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}\right]^{-1} \tag{3}$$

$$C = (a-1)x_{\min}^{a-1} \tag{4}$$

For the discrete case the corresponding parameters are given by Equations 5-6, where ζ(a) is the Riemann Zeta function and equals $\sum_{x=x_{\min}}^{\infty} x^{-a}$ and $\zeta'(a)$ its first derivative.

$$-\frac{\zeta'(a)}{\zeta(a)} = \frac{1}{n}\sum_{i=1}^{n}\log(x_i) \tag{5}$$

$$C = \frac{1}{\zeta(a)} \tag{6}$$

The determination of parameters *a* and *C* is the first step towards exploring power-laws. If the determined exponent *a* ranges between 2 and 3, then this is a strong evidence that the data may follow a power-law distribution. The second, and most important step, involves the evaluation of how well the observed data fit to the power-law distribution. To facilitate this, the most common technique is to employ the Kolmogorov-Smirnov test, which is a widely used goodness-of-fit test [Kolmogorov (1933)]. In case the data are power-law distributed, the null hypothesis will be accepted, whereas if the data distribution deviates significantly from a power-law distribution, the null hypothesis will be rejected.

4. Power–Laws in Repeating Patterns

4.1 Preliminaries

In our study, we focus on symbolic music representations where the notes of each musical score are available. A *pattern* of length *m* is a sequence of *m* consecutive notes that appear in the collection at least once. A *repeating pattern* [Hsu (2001)] of length *m* is a sequence of *m* consecutive notes that appear in the collection at least twice. We term the set of unique repeating patterns *distinct repeating patterns*. Evidently, the less notes a (repeating) pattern contains the larger the probability of occurrence. Although there is a plethora of variables that could have been explored, we restrict our focus on the following: (i) *support*, the number of musical scores that a repeating pattern is contained in, taking into account the whole collection, (ii) *length L*, the number of notes contained in the repeating pattern, and (iii) *frequency*, the number of occurrences of a repeating pattern in the same musical score.

We give a simple example to illustrate how the variables of interest are computed. Assume we have three musical scores each containing 5 notes. S_1 = ACCDD, S_2 = ACCAD and S_3 = ACCAA. The repeating patterns in our collection compose the set RP which contains the patterns A, C, D, AC, CC, CA, ACC, and CCA. Table 1 depicts the support, length and frequency values for the above repeating patterns. In the frequency column, three numbers appear, one for each musical score.

Evidently, larger the support, larger the significance of the patterns, as long patterns appearing in many musical scores are more powerful and less intuitive than small ones. Also, longer the patterns, larger the significance. Finally, the frequency measures the significance of the pattern in the same musical score.

Table 1. *Support, length and frequencies of patterns*

Pattern	*Support*	*Length*	*Frequency*
A	3	1	1, 2, 3
C	3	1	2, 2, 2
D	2	1	2, 1, 0
AC	3	2	1, 1, 1
CC	3	2	1, 1, 1
CA	2	2	0, 1, 1
ACC	3	3	1, 1, 1
CCA	2	3	0, 1, 1

The data sets used for the experimentation pertain to real musical data in the kern and midi format, obtained from KernScores (http://kern.humdrum.org) and Multimedia Library (http://www.multimedialibrary.com/barlow). We have used musical pieces of Bach, Corelli and Haydn (data sets BACH, CORELLI, HAYDN) a part of the collection of musical themes described in the work of Barlow and Morgenstern [Barlow, Morgestern (1978)] (BARLOW data set) and a part of the Essen collection, containing European and Asian music (data sets EUROPE and ASIA). As far as BARLOW, EUROPE and ASIA data sets are concerned, the musical pieces included therein are purely monophonic and thus repeating pattern extraction is rather trivial. On the other hand, the pieces of BACH, CORELLI and HAYDN data sets, which are all polyphonic, required special treatment. In each of these data all voices have been assumed in a sequential manner, one after the other, while taking into consideration that repeating patterns are not allowed to be found over connections of voices. We have done so, under the assumption that repeating patterns tend to occur in single voices and not distributed among them. In total, we have 570 works in BACH, 228 in CORELLI, 160 in HAYDN, 2,245 in ASIA and 6,201 in EUROPE. Finally, there are 9,811 themes in the BARLOW data set. Tables 2-4 give some basic characteristics regarding the data sets used in the experiments.

Table 2. *CLASSIC collection (BACH, CORELLI and HAYDN data sets)*

data set	*scores*	*rep. patterns*	*distinct rep. patterns*
BACH	570	61,429	12,606
CORRELI	228	33,176	11,627
HAYDN	160	22,450	7,970
Total	958	117,055	26,223

Table 3. *ESSEN collection (ASIA and EUROPE data sets)*

data set	*scores*	*rep. patterns*	*distinct rep. patterns*
EUROPE	6,201	86,479	12,746
ASIA	2,245	46,407	10,262
Total	8,446	132,886	20,484

Table 4. *BARLOW collection (9,811 themes)*

pattern length	*patterns*	*distinct rep. patterns*
2	186,205	2,008
3	175,334	15,257
4	163,502	49,389
5	152,515	86,546
6	142,789	109,216
7	133,679	117,928

4.2 Experimental Results

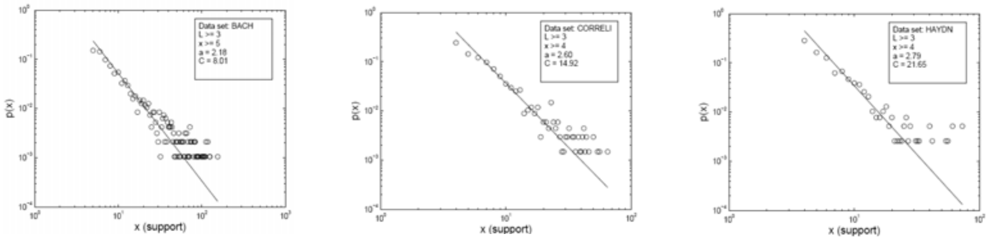
In the sequel, we give some representative results showing that in several cases power-laws do exist in musical data. Note that, these results are preliminary, and more research is required to draw some conclusions regarding the interpretation of these laws. We note that patterns are described by using only the pitch information of each note, and not its duration.

Figures 2-3 depict some results for the support of repeating patterns in the CLASSIC collection. In each graph we give the exponent (slope) *a*, the constant *C*, the x_{min} threshold, and the *L* length. Figure 2 shows the results for repeating patterns of length at least 3, whereas the results of Figure 3 refer to repeating patterns of length at least 4. Table 5 summarizes the results for BACH, CORELLI and HAYDN data sets. Rows that appear bold, illustrate that there is a power-law, based on Komogorov-Smirnov test. In the last column of the table, we give also the percentage of the population of the distinct repeating patterns obeying the power-law (power-law tail).

From Figures 2-3 and Table 5 we observe that power-laws do exist for BACH, CORELLI and HAYDN data sets. More specifically, power-laws in BACH data set are related to more scores than CORELLI and HAYDN, and this can be realized by observing the population column of Table 5. Moreover, in BACH there is a power-law even for larger repeating patterns, whereas this is not true for CORELLI and HAYDN that require repeating patterns of smaller size to form the power-law.

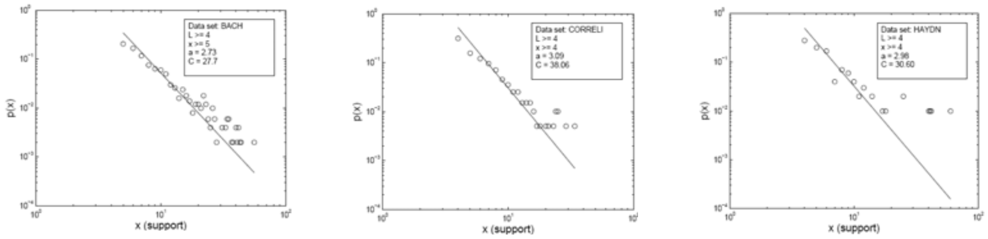
Figures 4-5 and Table 6 depict some representative results for the support of repeating patterns in the ESSEN collection. Power-laws do exist for this collection too, for repeating patterns of length larger than or equal to 3. This is true for both ASIA and

EUROPE data sets. An interesting observation is that the percentage of population obeying the power-law is larger for ASIA than EUROPE, meaning that the self-similarity property of the first data set is more evident than the second. Moreover, we observe that by considering patterns of length *L*≥2, the power-law existence is marginal, since the corresponding slope in Figure 4(a) and Figure 5(a) is less than 2.



(a) BACH ***(b)*** *CORRELI* ***(c)*** *HAYDN*

Figure 2. *Power – laws for L ≥* 3

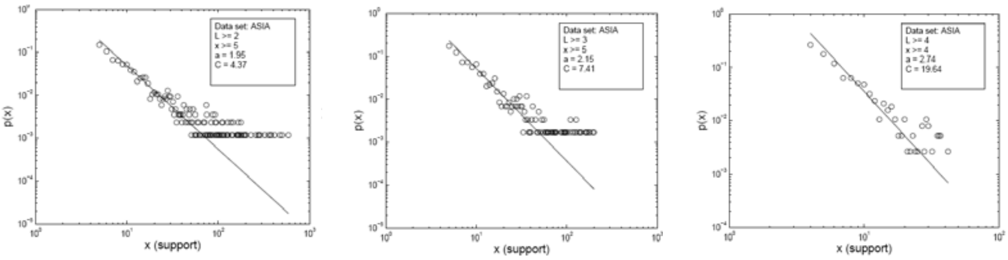


(a) BACH ***(b)*** *CORRELI* ***(c)*** *HAYDN*

Figure 3. *Power – laws for L ≥* 4

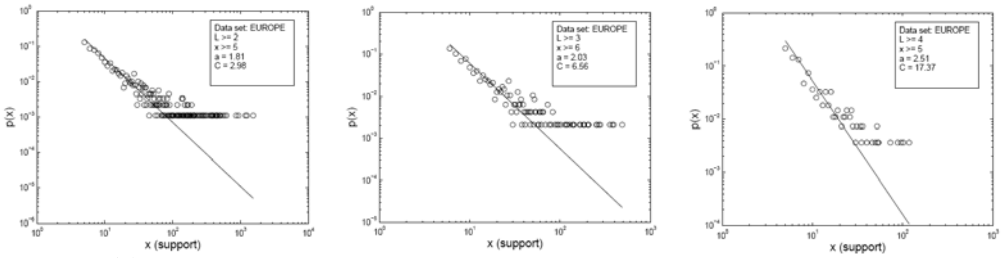
Table 5. *BACH, CORELLI and HAYDN summary of results for support.*

data set	*L*	*a*	*C*	x_{min}	*population %*
BACH	≥1	1.90	3.82	5	11.51
BACH	≥2	1.96	4.49	5	13.25
BACH	**≥3**	**2.18**	**8.01**	**5**	**10.00**
BACH	**≥4**	**2.73**	**27.70**	**5**	**6.86**
CORRELI	**≥1**	**2.01**	**5.22**	**5**	**8.32**
CORRELI	**≥2**	**2.12**	**6.86**	**5**	**8.91**
CORRELI	**≥3**	**2.60**	**14.92**	**4**	**7.07**
CORRELI	≥4	3.09	38.06	4	2.60
HAYDN	**≥1**	**2.09**	**6.36**	**5**	**8.77**
HAYDN	**≥2**	**2.24**	**9.30**	**5**	**9.32**
HAYDN	**≥3**	**2.79**	**21.65**	**4**	**6.29**
HAYDN	**≥4**	**2.98**	**30.60**	**4**	**2.20**



(a) $L \geq 2$ ***(b)*** $L \geq 3$ ***(c)*** $L \geq 4$

Figure 4. *Support based power -laws for ASIA*



(a) $L \geq 2$ ***(b)*** $L \geq 3$ ***(c)*** $L \geq 4$

Figure 5. *Support based power -laws for EUROPE*

Table 6 *ASIA and EUROPE summary of results for support*

data set	*L*	*a*	*C*	x_{min}	*population %*
ASIA	≥1	1.90	4.50	6	7.65
ASIA	≥2	1.95	4.37	5	9.83
ASIA	**≥3**	**2.15**	**7.41**	**5**	**7.40**
ASIA	**≥4**	**2.74**	**19.64**	**4**	**5.44**
EUROPE	≥1	1.82	5.83	11	4.56
EUROPE	≥2	1.81	2.98	5	8.46
EUROPE	**≥3**	**2.03**	**6.56**	**6**	**4.92**
EUROPE	**≥4**	**3.51**	**17.37**	**5**	**3.10**

Next we search for power-laws in the BARLOW collection. Table 7 summarizes the results for the support of the patterns. Again, the interesting cases are depicted in bold text. For patterns of length 3 and 4, the tail of the distribution follows a power-law, which means that the self-similarity property applies here as well.

Table 7. *BARLOW summary of results for support*

L	*a*	*C*	x_{min}	*population %*
3	**2.19**	**60.67**	**27**	**8.94**
4	**2.39**	**44.48**	**12**	**4.43**
5	3.05	1067.64	21	0.56
6	3.07	153.76	8	0.78
7	2.75	82.47	9	0.17

Up to now, we have investigated the existence of power–laws in the support of repeating patterns. Regarding the length of the patterns, power-laws have been detected only for the BACH data set, for which we give the log-log plot of the cumulative distribution in Figure 6. The parameters of the power-law are, a=2.59, C=71.04 and x_{min}=11, whereas the percentage of the repeating patterns obeying the law is ~10%. As in the previous, the power-law is detected in the distribution tail.

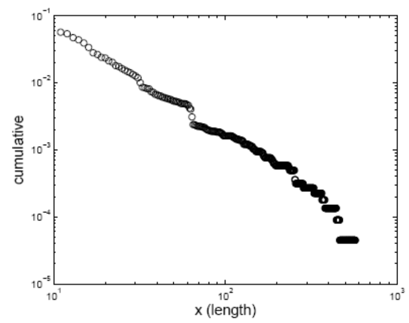


Figure 6. *Cumulative distribution for length in BACH data set*

Finally, we did discover power-laws for frequency by considering all repeating patterns (i.e. $L \geq 1$). However, the value of this law is limited because evidently, small patterns (containing 1 or 2 notes) are expected to occur frequently in a musical score. For this reason, the results are not given. However, musical scores of different genres may contain more significant power-laws with respect to frequency, and this should be further investigated in order to draw solid conclusions.

5 .Concluding Remarks

Power-laws appear frequently in nature, highly related to fractals and self-similarity. In this paper, we investigate the existence of such laws in symbolic musical data. The main result of our research is that power-laws do exist in music. We have inspected three different variables: support, length and frequency of repeating patterns. Regarding support, all collections show some degree of self-similarity, with different exponents and constants. Moreover, there is a difference in the percentage of the population that fit to a power-law distribution. In fact, all experiments have shown that power-laws exist in the distribution tails. Regarding the repeating patterns lengths, fewer power-laws have been detected. Specifically, in the CLASSIC collection power-laws in length have been detected only for BACH. Finally, no significant power-laws have been detected for repeating patterns frequencies.

The results of our study show that repeating patterns obey power-laws, which actually means that the existence of these patterns follow a scale-free distribution. This can lead to the conclusion that a part of a musical piece shows similar statistical properties as the whole piece does. This result can be used as an auxiliary technique to investigate the differences among diverse music collections. Additionally, it is a fundamental step in exploring the rules of the mathematic structures of music. Consequential, a possible future application is to develop systems, which they could compose music based on statistical properties.

Moreover, we could explore the applicability of power laws to music classification. This can be accomplished by applying power-law metrics to extract various features of each composer's piece, as a first phase and as a second phase to classify musical pieces based on these features. An approach to the identification and the classification of a composer or style of a musical piece has been implemented in [Manaris et. al. (2002)]. After the feature extraction, data mining and artificial intelligence techniques are applied to classify each piece to a specific composer or to a style of music. Our study could help towards this direction. Finally, the self-similarity properties of music collections can be used for improved compression techniques.

Acknowledgments

Thanks are due to Emilios Cambouropoulos, Alexandros Nanopoulos and Apostolos Papadopoulos for their valuable help.

References

Adamic L.A. and Huberman B.A. (2000). The nature of markets in the World Wide Web. Quarterly Journal of Electronic Commerce, Vol.1, p.512.

Aiello W., Chung F. and Lu L. (2000). A random graph model for massive graphs. Proceedings 32nd ACM STOC Symposium, pp.171-180.

Barlow H. and Morgenstern S. (1978). A dictionary of musical themes. 12th Impression, Ernest Benn, London & Tonbridge.

Clauset A. and Young M. (2005). Scale invariance in global terrorism. e-print physics/0502014v2.

Zanette D. (2006). Zipf's law and the creation of musical context. Musicae Scientiae, Vol.10, No.1, pp.3-18.

Ebel H., Mielsch L.I. and Bornholdt S. (2002). Scale-free topology of e-mail networks. Phys. Rev. E 66, 035103.

Estoup J.B. (1916). Gammes stenographiques. Institut Stenographique de France, Paris.

Faloutsos M., Faloutsos P. and Faloutsos C. (1991). On power-law relationships of the internet topology. Proceedings ACM SIGCOMM Conference, pp.251-262.

Hsu K.J. and Hsu A. (1991). Self similarity of the 1/f noise called music. Proceedings National Academy of sciences of the USA, Vol.88, No.8, pp.3507-3509.

Hsu J.L., Liu C.C. and Chen A.L.P. (2001). “Discovering nontrivial repeating patterns in music data. IEEE Transactions on Multimedia, Vol.3, No.3, pp.311-325.

Huberman B.A. and Adamic L.A. (2004). Information dynamics in the networked world. In Ben-Naim E., Frauenfelder H. and Toroczkai Z. (eds.), Complex Networks, Vol.650 in Lecture Notes in Physics, pp.371-398, Springer, Berlin.

Johnson N.L., Kotz S. and Kemp A.W. (1992). Univariate discrete distributions. John Wiley & Sons, New York.

Kalapala V., Sanwalani V., Clauset A. and Moore C. (2006). Scale invariance in road networks. Physical Review E, Vol.73.

Kolmogorov A.N. (1933). Giornale dell’ Instituto Italiano degli Attuari, Vol.4, p.77.

Lotka A.J. (1926). The frequency distribution of scientific production. J. Wash. Acad. Sci., Vol.16, pp.317–32.

Lu L., Wang M. and Zhang H.J. (2004). Repeating pattern discovery and structure analysis from acoustic music data”, Proceedings 6th ACM SIGMM Workshop, pp.275-282.

Lu E.T. and Hamilton R.J. (1991). Avalanches of the distribution of solar flares. Astrophysical Journal, Vol.380, pp.89-92.

Manaris B., Purewal T. and McCormick C. (2002). Progress towards recognizing and classifying beautiful music with computers: MIDI-encoded music and the Zipf-Mandlesbrot law. Proceedings IEEE Southeast Conference, pp.52-57.

Milton J.S. and Arnold J.C. (1995). Introduction to probability and statistics. 3rd edition, McGraw-Hill.

Miyazima S., Lee Y., Nagamine T. and Miyajima H. (2000). Power-law distribution of family names in Japanese societies. Physica A, Vol.278, pp.282-288.

Moura Jr. N.J. and Ribeiro M.B. (2006). Zipf law for Brazilian cities. Physica A, Vol.367, pp.441-448.

Neukum G. and Ivanov B.A. (1994). Crater size distributions and impact probabilities on Earth from lunar, terrestial planet, and asteroid cratering data. In Gehrels T. (ed.), Hazards Due to Comets and Asteroids, pp. 359–416, University of Arizona Press, Tucson.

Newman M.E.J. (2006). Power laws, Pareto distributions and Zipf’s law. Contemporary Physics, Vol.46, pp.323-351.

de S. Price D.J. (1965). Networks of scientific papers. Science, Vol.149, pp.510-515.

Voss R.F. and Clarke J. (1975). 1/f noise in music and speech. Nature, Vol.258, pp.317-318.