The Normalized Compression Distance Is Resistant to Noise

Manuel Cebrián, Manuel Alfonseca, *Associate Member, IEEE*, and Alfonso Ortega

***Abstract*—This correspondence studies the influence of noise on the normalized compression distance (NCD), a measure based on the use of compressors to compute the degree of similarity of two files. This influence is approximated by a first order differential equation which gives rise to a complex effect, which explains the fact that the NCD may give values greater than 1, observed by other authors. The model is tested experimentally with good adjustment. Finally, the influence of noise on the clustering of files of different types is explored, finding that the NCD performs well even in the presence of quite high noise levels.**

***Index Terms*—Clustering and noise resistance, datafile corruption, heterogeneous data analysis, Kolmogorov complexity, noisy channel, normalized compression distance, universal similarity distance.**

I. Introduction

Universal metrics are applicable to any kind of data and are one of the main objectives of clustering theory. The normalized information distance (NID) [2], [7] is a universal similarity metric that minorizes every computable metric. For two strings p and q the NID is defined as follows:

$$d(p,q) = \frac{\max\{K(p|q), K(q|p)\}}{\max\{K(p), K(q)\}}$$

where $K(x|y)$ is the conditional Kolmogorov complexity (CKC) of the string x given the string y; $K(x)$ is equivalent to $K(x|\epsilon)$, being ϵ the empty string. Both CKC and NID are incomputable [8]. The normalized compression distance (NCD) [4] is a computable estimation of the NID defined as follows:

$$\text{NCD}(p,q) = \frac{C(pq) - \min\{C(p), C(q)\}}{\max\{C(p), C(q)\}} \tag{1}$$

where pq is the concatenation of strings p and q, and $C(x)$ denotes the length of the text x compressed using some compression algorithm which asymptotically reaches the entropy of x, when the length of x tends to infinity.

Manuscript received March 24, 2006; revised January 23, 2007. This work was supported by the Spanish Ministry of Education and Science under Grant TSI 2005-08255-C07-06.

The authors are with the Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain (e-mail: manuel.cebrian@uam.es; manuel.alfonseca@uam.es; alfonso.ortega@uam.es).

Communicated by V. A. Vaishampayan, Associate Editor at Large.

Color versions of Figures 1 and 2 of this correspondence are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2007.894669

The current massive use of Internet has enormously increased the traffic of files across potentially noisy channels that can change their original contents. NCD is a similarity measure based on the use of compressors, so noise could make NCD get wrong results: a clustering application using NCD as a measure of distance would classify as dissimilar two similar files corrupted by noise.

The experiments described in this paper have been designed in the following way: all the files contain bytes in a certain range (i.e., genomes can only belong to {A,C,G,T}; texts can contain any ASCII character; music uses MIDI files and images use GIF format, both with their bytes in the range $[0, 255]$).

Noise is applied with certain probability independently to individual bytes, by integer addition of a uniform random positive (nonzero) values, in such a way that the resulting byte belongs to their appropriate above mentioned ranges. This model of comunication with noise is known in the literature as the *symmetric channel* [5, Ch. 8].

Each experimental tests show how the NCD changes when applied to two files, one of which is distorted by an increasing noise ratio (i.e., several percentages of noise are added during the experiments).

II. Theoretical Analysis

Let us consider a file of size a which is compressed by a given compressor into another file of size b. If we add noise to the original file and compress it, as the amount of noise increases, the compressor will be able to reduce less and less the file size, until it will be unable to reduce it at all, once the contents of the file become fully random. Therefore the size of the compressed file will start at b, when no noise is added, and will increase steadily to a, which will be reached when the whole initial file has been replaced by random noise.

At a given point in this procedure, if we add Δx noise to the file (i.e. change randomly the values of Δx bytes in the file), the size increase (the compression loss) we may expect will be proportional to the amount of the file which has not yet been replaced by noise. Therefore, the evolution of the compressed size y will be defined by $\Delta y = \gamma(a - y)\Delta x$. When $\Delta x \to 0$, this equation becomes the first-order differential equation $dy/dx = \gamma(a - y)$. The solution of this equation, taking into account the indicated initial conditions, is

$$y = a - (a - b)e^{-\gamma x}$$

where x is the amount of noise added and the value of y (the size of the compressed file) is b for $x = 0$, a for $x \to \infty$.

Consider the definition (1) and assume that we want to study the variation of the distance $\text{NCD}(p, q)$ between a fixed file p, and another file q which is being contaminated by growing amounts of noise. Without loss of generality, we may assume that $C(p) \leq C(q)$. Therefore, the above distance becomes

$$\text{NCD}(p,q) = \frac{C(pq) - C(p)}{C(q)}.$$

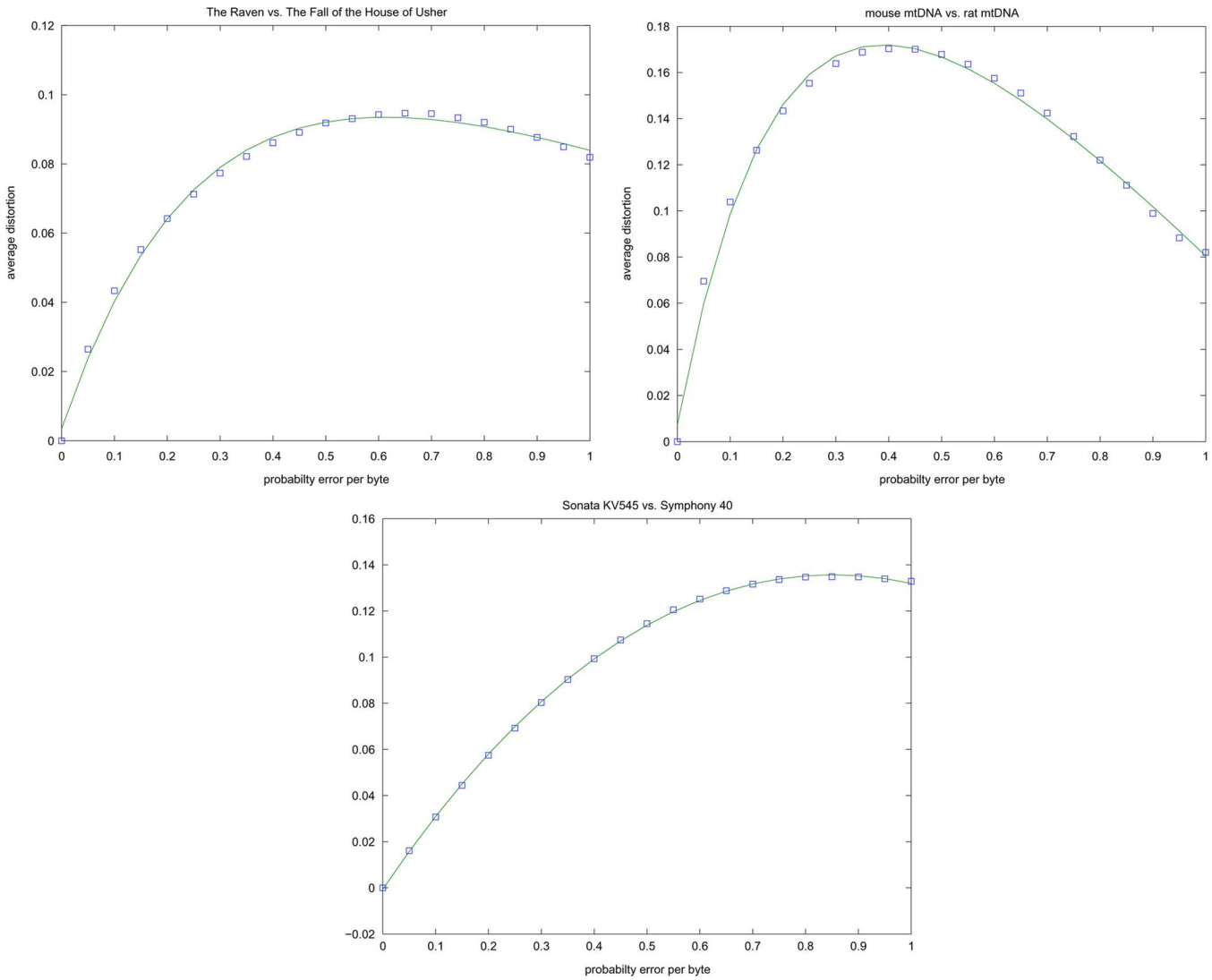

Fig. 1. Three examples of different types of data which exhibit the typical decay behavior of the average distortion. From left to right, we compare the texts "The Raven" versus "The Fall of the House of Usher," mouse mtDNA versus rat mtDNA, and Mozart's "Sonata KV545" versus "40th Symphony."

We have seen that, as the amount of noise x introduced in file q grows, $C(q) = a - (a - b)e^{-\gamma x}$. It is easy to see that $C(pq)$ will evolve in a similar way, although with different constants, because the noise introduced in the second part of the file not only destroys redundancies in that section of the file, but also prevents possible cross-compressions with the first part, which does not receive noise. So, $C(pq) = c - d.e^{-\phi x}$. Finally, $C(p)$ is a constant. Replacing these values, under certain conditions the NCD formula can be approximated by

$$\text{NCD}(p, q) = \alpha + \beta e^{-\gamma x} - \delta e^{-\phi x} \tag{2}$$

where the values of the constants depend on the actual files p and q compressed; x is again the amount of noise added. With certain values of the constants, this function reaches values greater than 1 (usually smaller than 1.1). This effect provides a different explanation of the anomaly, signaled in [4], that the value of the NCD may be greater than 1, without any reference to the presence of defects in the compressor implementation.

III. Experimental Results

In the last section we obtained a model (2) for the NCD in the presence of noise. If we compute the *average distortion* introduced by a noise level l, $0 \leq l \leq 1$, we come to a similar equation, since $\text{NCD}(p, q)$ is a constant:

$$\begin{aligned} \Delta_l(p, q) &\equiv E[\text{NCD}(p,\ q + n_l \bmod r) - \text{NCD}(p, q)] \\ &\approx \alpha' + \beta' e^{-\gamma' l} - \delta' e^{-\phi' l} \end{aligned} \tag{3}$$

where n_l is a random string whose length is equal to the length of q and whose ith character is nonzero with probability 1; r is the size of the file type range (i.e. 4 for DNA, 93 for ASCII, and 256 for the rest). The modulo operation is performed to maintain each value in its proper range.

In this section we test the goodness of the model (3) by adjusting it over experiments with real data: ASCII texts [6], mitochondrial DNA (mtDNA, obtained from [3]), songs in WAV format and face images in GIF format [1]. For each two files p and q we estimate $\Delta_l(x, y)$

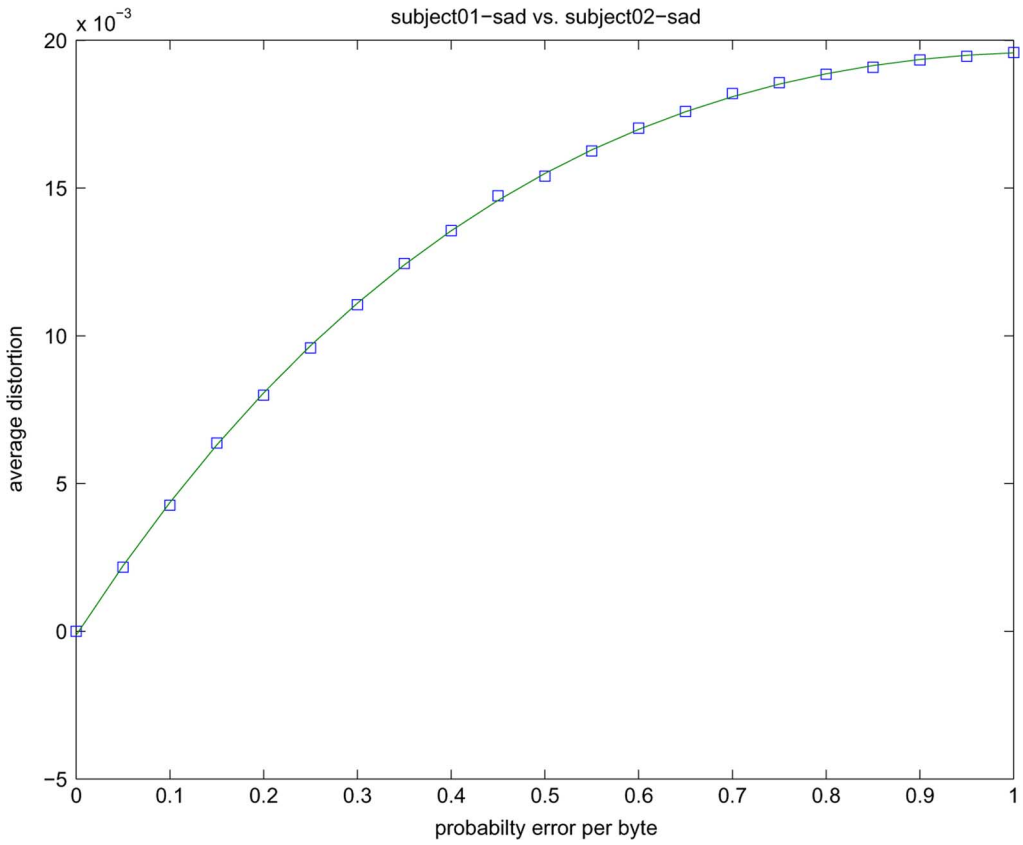

Fig. 2. An example of the typical nondecay behavior of the average distortion with face images. The two elements in the comparison are subject #1's and #2's face images with a predefined "sad" disposition.

TABLE I
DATA-FITTED VALUES OF THE MODEL (3) AND EXISTENCE OF DECAY FOR SEVERAL EXPERIMENTS PERFORMED ON TEXTS, MTDNA, SONGS, AND FACE IMAGES

p	q	α'	β'	γ'	δ'	ϕ'	decay
Secret Adversary	The Mysterious Affair at Styles	0.0720	0.9036	2.2057	0.9696	3.0468	yes
Antony and Cleopatra	Hamlet	0.0617	0.9159	2.0486	0.9689	2.6575	yes
An Essay on Criticism	The Fall of the House of Usher	0.0712	0.9116	1.5148	0.9787	2.3714	yes
Hamlet	Secret Adversary	-0.0846	1.0324	1.0488	0.9414	1.4349	yes
The Raven	The Fall of the House of Usher	0.0396	0.9503	1.7972	0.9855	2.1636	yes
mouse	rat	-0.0918	0.8437	1.3972	0.7414	3.0697	yes
finWhale	blueWhale	-0.8439	1.2250	0.1565	0.3683	7.5895	yes
graySeal	Harbor Seal	-0.7520	1.2352	0.1215	0.4693	10.698	yes
human	Blue Whale	-0.2372	0.9206	1.0215	0.6772	2.6497	yes
rat	horse	-0.2220	0.8756	1.0162	0.6464	2.7148	yes
chimpanzee	Gray Seal	-0.2047	0.8422	1.0461	0.6302	2.8282	yes
(Chopin) Prelude 15	(Chopin) Prelude 7	1.4075	-1.1543	0.4476	0.2546	-0.7388	yes
(Chopin) Prelude 15	Begin the Beguine	1.2717	-1.0224	0.3340	0.2508	-0.5981	yes
(Mozart) Sonata KV545	(Mozart) Symphony 40	1.3465	-0.8083	0.6904	0.5388	-0.4068	yes
Begin the Beguine	My heart belongs to daddy	1.3838	-1.0255	0.5251	0.3595	-0.5916	yes
subjec01.centerlight	subject03.centerlight	0.0171	0.6498	6.4166	0.6663	6.2745	no
subject02.happy	subject04.happy	0.0178	0.6501	6.2037	0.6672	6.0454	no
subject01.sad	subject02.sad	0.0193	0.6498	6.2472	0.6684	6.0887	no
subject01.centerlight	subject01.normal	0.6447	0.6705	0.0325	1.3116	0.0305	no
subject02.sleepy	subject02.wink	0.6325	0.6764	0.0334	1.3051	0.0332	no
subject05.surprised	subject05.glasses	0.6441	0.6709	0.0226	1.3111	0.0256	no

averaging over 10 realizations of the random vector n_l, computing distances by means of the CompLearn Toolkit [3], which implements the NCD and NCD-driven clustering; default CompLearn parameters were used in all experiments.

In all the experiments performed, the model obtained a very accurate fit with a squared 2-norm of the residuals always below 10^{-3}. An interesting result is that, for some data types, the average distortion increases until it reaches a maximum at some $l < 1$, and

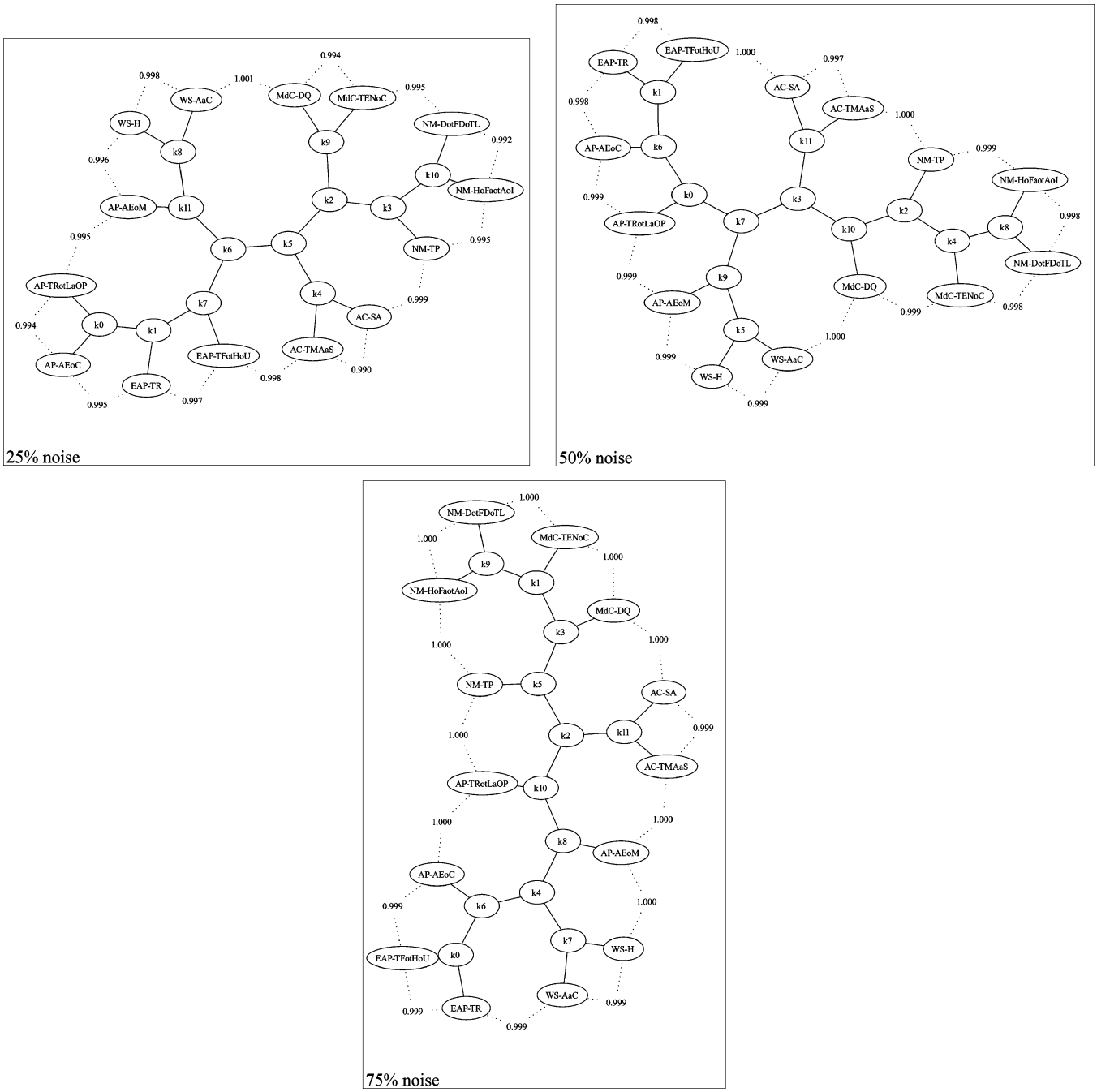

Fig. 3. NCD-driven clusterings of texts by several authors in which different levels of noise have been added to each sequence. The first characters in the book labels are the initials of their authors: "AC" = Agatha Christie, "AP" = Alexander Pope, "EAP" = Edgar Allan Poe, "WS"= William Shakespeare, and "NM" = Niccolo Machiavelli. The quality of the clustering degrades slowly due to the linear growth of the average distortion.

then decays steadily converging toward a smaller value when the level of noise is increased; texts, mtDNA, and songs follow this behavior (Fig. 1). This phenomenon explains the already mentioned fact that the NCD can sometimes reach values greater than 1 when comparing files that share very little information: the region in which $\Delta_l(x, y)$ has a value greater than $\Delta_1(x, y)$ (distortion with full noise) coincides with the region in with the NCD$(p,\ q + n_l \bmod r)$ is greater than 1.

Other data types like face images (Fig. 2) increase continuously without any posterior decay. In Table I we show some results of the experiments performed and its classification according to the existence of a decay; 200 experiments were performed for each pair p, q, 10 realizations of n_l for each one, with 20 different values of $l \in \{0.05, 0.10, 0.15, \ldots, 1\}$.

It is worthwhile to consider whether other models with the same number of free parameters could fit the experimental data. It is possible, for instance, to get a good fourth degree polynomial adjustment of these curves in the $[0, 1]$ noise interval, but this would be a consequence of the fact that we are measuring noise as the rate of original information changed. If we had chosen to measure it as the number of changes made in the original streams, the range of our independent variable would be $[0, \infty)$ (see Section II). In this case, our model would still be able to fit it without problems, while a polynomial cannot reproduce the asymptotic behavior. Thus, our model is as good as other simpler models with the

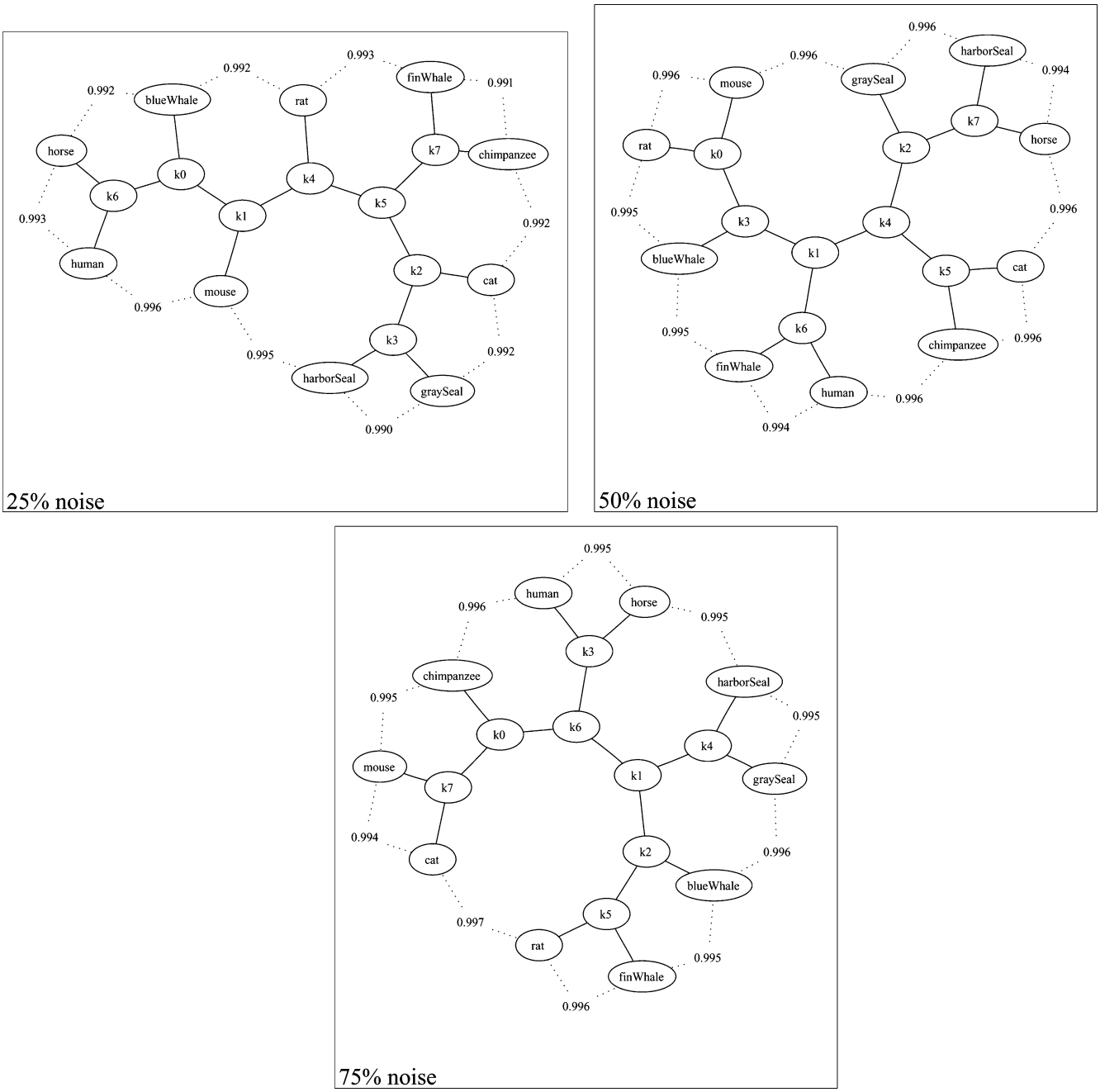

Fig. 4. NCD-driven clusterings of mammalian mtDNA sequences in which different levels of noise have been added to each element. The quality of the clustering degrades slowly but somewhat faster than with texts (see Fig. 3) due to the faster growth of the average distortion in this type of data.

same number of free parameters, but also shows a correct asymptotic behavior difficult to express with them.

Finally, we show two real clustering experiments performed with the CompLearn Toolkit in the presence of noise. In Fig. 3, several dendrograms result from clusterings texts by several authors in which several levels of noise have been added to each text. Similar experiments are repeated in Fig. 4 but this time with mammalian mtDNA.

IV. Conclusion and Future Work

When the NCD is used to compute the distance between two different files, the second file can be considered as a noisy version of the first. Therefore, the effect on the NCD of the progressive introduction of noise in a file can provide information about the measure itself. In this correspondence, we forward a theoretical reasoning of the expected effect of noise introduction, which explains why the NCD can get values greater than 1 in some cases.

A first batch of our experiments confirm the theoretical model. A second batch explores the effects of noise on the precision of clusterings based on the use of the NCD. It can be noticed that the clustering process is qualitatively resistant to noise, for the results do not change much with quite large amounts of it. Different types of files are differently affected, however, which is not surprising: mtDNA files, for instance, which are built on a 4-letter alphabet, are degraded faster than human text, which uses a larger alphabet.

In the future, we intend to tackle a quantitative demonstration of the NCD resistance to noise. We shall also try other metrics and clustering

procedures, appropriate to the different file types, to compare their resistance to noise with our NCD results.

REFERENCES

[1] Yale Face Database [Online]. Available: http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[2] C. Bennett, P. Gacs, M. Li, P. Vitanyi, and W. Zurek, "Information distance," *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1407–1423, Jul. 1998.

[3] R. Cilibrasi, A. L. Cruz, and S. de Rooijet, The CompLearn Toolkit [Online]. Available: http://www.complearn.org software available at

[4] R. Cilibrasi and P. Vitanyi, "Clustering by compression," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1523–1545, Apr. 2005.

[5] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[6] M. Hart, The Gutemberb Project [Online]. Available: http://www.gutenberg.org free electronic books available at

[7] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi, "The similarity metric," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3250–3264, Dec. 2004.

[8] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer-Verlag, 1997.