Entropy approach through graph theory for studying the degree of order in one-dimensional distributions of objects

1996 J. Phys. A: Math. Gen. 29 2969

(http://iopscience.iop.org/0305-4470/29/12/007)

J. Phys. A: Math. Gen. **29** (1996) 2969–2980. Printed in the UK

Entropy approach through graph theory for studying the degree of order in one-dimensional distributions of objects

Armin Loeffler, Monique Rasigni and Uwe Raidt

Département de Physique des Interactions Photons-Matière, case EC1, Faculté des Sciences et Techniques de St Jérôme, 13397 Marseille Cedex 20, France

Received 18 September 1995

Abstract. Graph theory, through the minimal spanning tree (MST), and information theory, through the concept of entropy, are used to define a new parameter γ which quantitatively characterizes the degree of order (or disorder) in 1D sets of points. Theoretical calculations and results obtained from various computer-simulated distributions of points are compared. For well-chosen conditions which are specified, a good agreement is noted. Finally, a method which is particularly easy to implement is proposed to evaluate the parameter γ for any real or simulated 1D distribution of points.

1. Introduction

In a previous paper [1] a new approach was developed to study order and disorder in two-dimensional distributions of particles. This approach was based on a graph, the minimal spanning tree (MST), constructed from the set of points which locate the positions of particles. Recall that an MST is a tree which contains all the nodes and where the sum of the edge weights is minimal. Depending on the starting point, there may be more than one MST for a given set of points, but all the MSTs have the same edge-length histogram (the edge length considered here is the Euclidean distance). It follows that statistical information deduced from this histogram, such as the average edge length m and the standard deviation σ, may be used as characteristics for the distribution to be studied. The use of a diagram involving both m and σ, both normalized according to an appropriate process [2], makes it possible to compare different distributions by taking a simple reading of the (m, σ) plane on which well-characterized distributions (such as perfectly ordered or random) were previously located. The method turned out to be successful in various situations such as the study of biological systems [3–7], the quantization of thin film growth [8, 9], the percolation phenomenon [10], the statistical analysis of disorder in 2D cellular arrays in directional solidification [11], the determination of the nature of disorder in liquid and glassy solids [12], etc.

Despite its advantages, namely adaptability and easiness, the method is restrictive in that it does not permit the quantitative determination of the degree of order in a distribution. The present paper is devoted to such a determination, with a view to making our previous analysis more informative and accurate.

The basic idea is to apply information theory to graphs constructed from sets of points. In this approach, graphs are considered as a source of information through the distribution of edge lengths and the angles that the edges make with a given axis. This has led us to define an entropy function from those parameters whose values may be used to characterize the degree of order for a set of points. 1D and 2D sets of points have been considered separately. The present work is devoted to the study of the 1D distributions of points. After reviewing the basic definitions related to information theory and the entropy of statistical mechanics, we discuss their possible application to an MST graph constructed from a set of points. Within this framework an entropy function relative to MST edge-length distribution is proposed and its validity is justified. Then a parameter involving the previously defined entropy function is proposed to quantize the degree of order of any 1D distribution of points. The results obtained from simulated random and non-random distributions of points are compared with those provided by theoretical calculations.

2. Information and entropy

Let us consider a system (or source of information) which can be expressed by means of a finite number of events e_i $(i = 1, 2, \ldots, M)$ with the probabilities P_i such as $\sum_{i=1}^{M} P_i = 1$. Information theory makes it possible to measure the uncertainty associated with the message source. According to Shannon, the uncertainty for the distribution $\{P_i\} = \{P_i; i = 1, 2, \ldots, M\}$, is given by

$$I(\{P_i\}) = -\lambda \sum_{i=1}^{M} P_i \ln[P_i] \qquad (1)$$

where λ is a positive constant.

In our case, any 1D distribution of points may be characterized by means of the indexed edge lengths of an MST constructed from the set of points, where indices indicate how the edges are connected to each other. If the distribution consists of equidistant points, it is enough to know the edge-length value. But for all the other cases, not knowing the indices leads to a loss of information about the MST and therefore about the distribution of points.

Within the framework of information theory this loss of information can be evaluated from (1) in which the probabilities P_i characterize events e_i defined as follows. Assume that edge lengths ℓ can take any value between two limits α and $\beta(\alpha < \beta)$, and divide the $[\alpha, \beta]$ interval into M segments of equal length $\delta\ell = (\beta - \alpha)/M$. The even e_i is then defined as follows: the edge length ℓ belongs to the interval $[\alpha_i, \beta_i]$, where

$$\alpha_i = \alpha + (i-1)\frac{(\beta - \alpha)}{M} \qquad \text{and} \qquad \beta_i = \beta + i\,\frac{(\beta - \alpha)}{M}\,.$$

From the point of view of statistical mechanics, the set of indexed edge lengths of the MST constructed from the distribution of points may be used to define a so-called microstate. Knowing only the edge lengths makes it possible to construct several MSTs, say Q, each being associated with a microstate. In this context a state, which consists of the set of all possible microstates, may be characterized by the MST's edge-length distribution. The lack of information is measured by the statistical entropy

$$S[Q] = \ln[Q]\,.$$

For a finite set of events e_i with the discrete distribution $\{P_i\}$, the Stirling formula leads to an expression for the statistical entropy $S\{P_i\}$ similar to (1).

For a continuous probability distribution $\Psi(\ell)$ subject to the constraint

$$\int_{\alpha}^{\beta} \Psi(\ell)\,\mathrm{d}\ell = 1$$

the statistical entropy is defined by [13]

$$S = -\lambda \int_{\alpha}^{\beta} \Psi(\ell) \ln[\Psi(\ell)\delta_0\ell] \, \mathrm{d}\ell$$

where $\delta_0\ell$ is a minimal interval corresponding to the best possible accuracy when measuring ℓ and such that the variations of $\Psi(\ell)$ on this interval are negligible [13].

3. Entropy function for MST edge-length distribution

The previous considerations enable us to define an entropy function S_ℓ related to the edge-length distribution of an MST constructed from a set of points:

$$S_\ell = -\int_0^\infty \Psi(\ell) \ln[\Psi(\ell)\delta_0\ell] \, \mathrm{d}\ell \tag{2}$$

where $\Psi(\ell)$ is subject to the constraint

$$\int_0^\infty \Psi(\ell) \, \mathrm{d}\ell = 1 \, .$$

The discrete counterpart of S_ℓ is given by

$$S_\ell^d = -\sum_{i=1}^{M} P_i \ln[P_i] \tag{3}$$

where P_i is the probability that an edge length ℓ belongs to the class 'i', namely $P_i = N_i/N_\mathrm{t}$, N_i is the number of edge lengths related to the ith class and N_t is the total number of edge lengths ($N_\mathrm{t} = N - 1$ for the MST). The number of class intervals M is chosen such that $\delta_0\ell = L/M$, where L is the length of the interval on which the points are distributed. S_ℓ may be used for characterizing the degree of order or disorder in any set of points.

Let us first consider a 1D distribution of N randomly distributed points on an axis of given length L and the constraints

(i) normalization condition for $\Psi(\ell)$

$$\int_0^\infty \Psi(\ell) \, \mathrm{d}\ell = 1$$

(ii) condition for the total length of the graph ($N - 1 \sim N$)

$$N \int_0^\infty \ell \Psi(\ell) \, \mathrm{d}\ell = L \, .$$

According to the Lagrange multiplier method we define

$$S^* = -\int_0^\infty \Psi(\ell) \ln[\Psi(\ell)\delta_0\ell] \, \mathrm{d}\ell + b\left[1 - \int_0^\infty \Psi(\ell) \, \mathrm{d}\ell\right] + c\left[L - N \int_0^\infty \ell \Psi(\ell) \, \mathrm{d}\ell\right].$$

The condition

$$\frac{\partial S^*}{\partial \Psi(\ell)} = 0$$

leads to

$$\Psi(\ell) = A \exp[-\beta \ell] \, . \tag{4}$$

We now consider the MST constructed from the previous N points. The probability of having an edge of length ℓ in such a graph is the same as the probability $P(0)$ of having no point in intervals of length ℓ. The Poisson distribution provides

$$P(0) = \exp[-\mu] \qquad \text{with} \qquad \mu = \rho\ell \quad \text{and} \quad \rho = \frac{N}{L} \, .$$

$P(0)$ may also be expressed as

$$P(0) = 1 - \int_0^\ell \Psi(x)\,\mathrm{d}x \, .$$

So we deduce that

$$\Psi(\ell) = \rho \exp[-\rho\ell] \tag{5}$$

which is similar to relation (4).

This result shows that the entropy function S_ℓ defined by (2) is maximal for a set of randomly distributed points. Moreover, S_ℓ tends to zero when all the edges have the same length $\ell_0 \pm \delta_0\ell/2$. Indeed in this case $\Psi(\ell)$ may be defined as

$$\Psi(\ell) = \begin{cases} \dfrac{1}{\delta_0\ell} & \text{if} \quad \ell_0 - \dfrac{\delta_0\ell}{2} < \ell < \ell_0 + \dfrac{\delta_0\ell}{2} \\ 0 & \text{if} \quad |\ell - \ell_0| > \dfrac{\delta_0\ell}{2} \, . \end{cases}$$

So that

$$S_\ell = -\int_{\ell_0-\delta_0\ell/2}^{\ell_0+\delta_0\ell/2} \frac{1}{\delta_0\ell} \ln\left[\frac{\delta_0\ell}{\delta_0\ell}\right] \mathrm{d}\ell = 0 \, .$$

This situation occurs either when all the points form a regular network or are concentrated on one point to constitute what can be termed as a perfect cluster. These two situations are related to repulsive and attractive interactions between the points, respectively.

Moreover, note that the probability density $\Psi(\ell)$ may be written as

$$\Psi(\ell) = \rho\phi_\rho(\ell) \, .$$

Then S_ℓ becomes

$$S_\ell = \ln\left[\frac{1}{\rho\delta_0\ell}\right] - \int_0^\infty \rho\phi_\rho(\ell) \ln[\phi_\rho(\ell)]\,\mathrm{d}\ell \, . \tag{6}$$

When changing unit length (scale change), ρ and $\delta_0\ell$ vary in opposite ways, so that $\rho\delta_0\ell$ keeps a constant value. Given that S_ℓ has to be invariant for such a change, it ensures that the term

$$F = -\int_0^\infty \rho\phi_\rho(\ell) \ln[\phi_\rho(\ell)]\,\mathrm{d}\ell \tag{7}$$

does not depend on the point density ρ but only on the way the points are scattered. For a random arrangement of points

$$\phi_\rho(\ell) = \exp[-\beta\ell] \qquad \text{so} \quad F = 1 \, .$$

When all the edges have the same length $\ell_0 \pm \delta_0\ell/2$ we have

$$F = -\int_{\ell_0-\delta_0\ell/2}^{\ell_0+\delta_0\ell/2} \frac{1}{\delta_0\ell} \ln\left[\frac{1}{\rho\delta_0\ell}\right] \mathrm{d}\ell = -\ln\left[\frac{1}{\rho\delta_0\ell}\right].$$

Therefore, for a given ρ, the minimal F value depends on $\delta_0\ell$. In particular $F \to -\infty$ for $\delta_0\ell \to 0$. This limit corresponds to a regular 1D mosaic for which MST edge lengths are known with infinite accuracy.

For convenience, the way the points are scattered will be characterized by using a parameter γ defined as

$$\gamma = 1 - F\,. \tag{8}$$

Thus $\gamma = 0$ for a random arrangement of points and $\gamma \to \infty$ for a regular 1D mosaic. The behaviour of F or γ for intermediate situations is studied in the following section.

4. Study of simulated distributions of points

4.1. Experimental procedure

The points have been distributed along an axis of length L by means of two procedures:

(i) Random distributions have been simulated by using the linear congruent Monte Carlo method [14].

(ii) Other distributions have been generated from a basic distribution made up of N points located at regular intervals $a = L/N$ (regular 1D mosaic). This arrangement may be randomized by giving each point a new position deduced from its previous position using a Gaussian distribution with a standard deviation of ω and a zero mean. For a ω value of the order of the a value, the uniform random distribution is reached [1]. Thus by using ω values from zero to $\omega \geqslant a$, a set of various distributions can be generated.

For each distribution, the graph connecting all the points in the shortest way or 1D MST is constructed, and from the corresponding edge-length histogram an estimate of the statistical entropy S_ℓ is obtained by means of (3).

4.2. Results related to random distributions of points

By using the expression of $\Psi(\ell)$ given by (5) in (2), we obtain, after performing integration:

$$S_{\ell,\mathrm{r}} = \ln\left[\frac{e}{\rho\delta_0\ell}\right]$$

where the subscript r stands for random, or

$$S_{\ell,\mathrm{r}} = \ln\left[\frac{eM}{N}\right] \tag{9}$$

by taking $\delta_0\ell = L/M$ and $\rho = N/L$ into account.

Ten distributions of N ($N = 2000$ and $N = 6000$, respectively) randomly scattered points have been generated on a segment of length $L = 1$. For each distribution the statistical entropy has been estimated by means of (3) for different M values such as $M/N \gtrsim 1/e$. In figure 1(a) we have reported both $\langle S^d_{\ell,\mathrm{r}}\rangle$—the arithmetic averages related to the ten distributions ($N = 2000$ and $N = 6000$)—and $S_{\ell,\mathrm{r}}$ versus M/N. It can be noted that experimental results do not depend explicitly on the values of N: they only depend on M/N as expected from (9). In the same way we obtain a good agreement between $S_{\ell,\mathrm{r}}$ and $\langle S^d_{\ell,\mathrm{r}}\rangle$ except for small (figure 1(b)) and large (figure 1(c)) values of M/N.

In the first case, the discrepancy may be explained by the fact that small values of M (*a fortiori* values of M/N) correspond to large values of $\delta_0\ell$; then the variations of $\Psi(\ell)$ on these $\delta_0\ell$ intervals can no longer be considered negligible, as is the case when defining

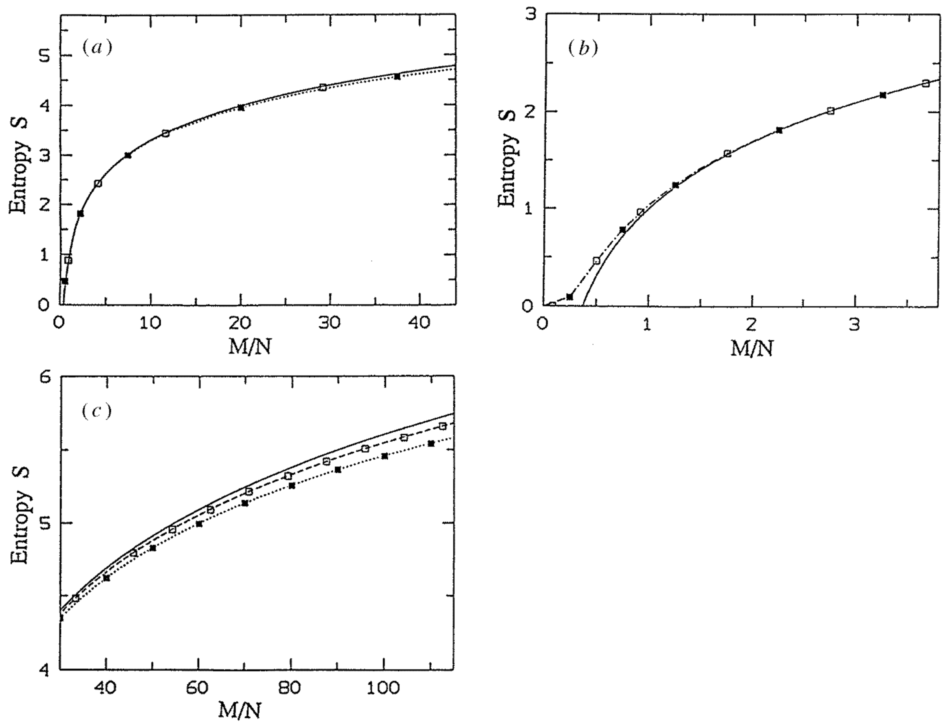

Figure 1. Entropy S versus M/N for randomly distributed points. (a) Full squares: $\langle S^d_{\ell,\mathrm{r}}\rangle$ arithmetic mean of results obtained by using (3) for ten random distributions of $N = 2000$ points; open squares: same as full squares for $N = 6000$. Full curve: $S_{\ell,\mathrm{r}}$ given by (9); dotted curve: $(S^d_{\ell,\mathrm{r}})_{\mathrm{th2}}$ given by (11). (b) Results for small values of M/N plotted on an expanded M/N scale: full squares, open squares and full curve same as (a); chain curve $(S^d_{\ell,\mathrm{r}})_{\mathrm{th1}}$ given by (10). (c) Results for large values of M/N plotted on an expanded entropy scale: full curve, full squares and open squares same as (a); dotted curve and broken curve $(S^d_{\ell,\mathrm{r}})_{\mathrm{th2}}$ given by (11) for $N = 2000$ and $N = 6000$, respectively.

the entropy function S_ℓ. So the probability that an edge length ℓ belongs to the class 'i' may be written as

$$p_i = \int_{(i-1)\delta_0\ell}^{i\delta_0\ell} \Psi(\ell)\,\mathrm{d}\ell = \exp[-\rho(i-1)\delta_0\ell] - \exp[-\rho i\delta_0\ell]$$

By substituting this expression of p_i on the right-hand side of (3), the limit $N \to \infty$ leads to the theoretical expression for $S^d_{\ell,\mathrm{r}}$:

$$(S^d_{\ell,\mathrm{r}})_{\mathrm{th1}} = \frac{N}{M}\frac{\exp\big[-N/M\big]}{1-\exp\big[-N/M\big]} - \ln\left[1-\exp\left[-\frac{N}{M}\right]\right]. \qquad (10)$$

It can be seen from figure 1(b) that (10) fits the experimental results well for small values of M/N. As values of M/N increase, $(S^d_{\ell,\mathrm{r}})_{\mathrm{th1}}$ quickly approaches $S_{\ell,\mathrm{r}}$. For example, a difference of less than 1% is obtained as soon as $M/N \geqslant 1.75$. Indeed $\ln[\mathrm{e}M/N]$ is the first-order approximation in M/N of the right-hand side of (10).

Figures 1(a) and (c) show that beyond a given value of M/N, $\langle S^d_{\ell,\mathrm{r}}\rangle$ deviates from $S_{\ell,\mathrm{r}}$. The discrepancy takes place for larger values of M/N as N is large and increases regularly with M/N. This behaviour is due to the fact that when M becomes too large compared to N, there are no longer enough data in class intervals and the statistics are biased.

Actually the set of experimental results may be recovered theoretically. Indeed, as the edge-length events are independent, it is possible to compute the probability that k events belong to the ith interval from the binomial distribution, i.e.

$$\Pi(N_i = k) = \binom{N_\mathrm{t}}{k} p_i^k (1 - p_i)^{N_\mathrm{t}-k}$$

where N_t is the number of events and

$$p_i = \int_{(i-1)\delta_0\ell}^{i\delta_0\ell} \Psi(\ell)\,\mathrm{d}\ell\,.$$

In this case it follows from (3) that the statistical entropy may be written as

$$(S^d_{\ell,\mathrm{r}})_{\mathrm{th2}} = -\sum_{i=1}^{M}\sum_{k=0}^{N_\mathrm{t}} \Pi(N_i = k)\frac{k}{N_\mathrm{t}} \ln\left[\frac{k}{N_\mathrm{t}}\right]. \tag{11}$$

$(S^d_{\ell,r})_{\mathrm{th2}}$ values computed for $N = 2000$ and $N = 6000$ are reported in figure 1(a) and (c). A very good agreement with $\langle S^d_{\ell,\mathrm{r}}\rangle$ (better than 0.05%) is observed.

We notice that $(S^d_{\ell,\mathrm{r}})_{\mathrm{th2}}$ represents the expected value of the experimental statistical entropy while $(S^d_{\ell,\mathrm{r}})_{\mathrm{th1}}$ may be considered as the entropy related to the expected value of the different random variables associated with $P_i = N_i/N_\mathrm{t}$.

The values of $(S^d_{\ell,\mathrm{r}})_{\mathrm{th1}}$, $(S^d_{\ell,\mathrm{r}})_{\mathrm{th2}}$, $S^d_{\ell,\mathrm{r}}$ and $S_{\ell,\mathrm{r}}$ are in good agreement for values of M/N such that

$$\mathcal{A} \leqslant \frac{M}{N} \leqslant \mathcal{B}(N)$$

where $\mathcal{A}$ and $\mathcal{B}$ are values of M/N which depend on the accuracy chosen for $S^d_{\ell,\mathrm{r}}$ when compared with $S_{\ell,\mathrm{r}}$. $\mathcal{B}$ is dependent on N while $\mathcal{A}$ is not.

4.3. *Results related to non-random arrangements of points*

Such distributions have been generated on a segment of length $L = 1$, from the procedure described in section 4.1. For the sake of convenience we note $\omega = \alpha a$ where α is a positive dimensionless variable and $a = L/N$. For a given value of M/N, it is observed, as far as $\alpha \leqslant 1$, that the statistical entropy computed by means of (3), and henceforth denoted $S^d_{\ell,\alpha}$, is always less than the one corresponding to the random distributions of points.

Figure 2(a) shows, for different values of α, the arithmetic average values $\langle S^d_{\ell,\alpha}\rangle$ calculated over four distributions and for two populations ($N = 2000$ and $N = 6000$). Computations have been made by using an M/N value of 25 in (3). It can be noted that $S^d_{\ell,\alpha}$ increases regularly with α and approaches the $S_{\ell,\mathrm{r}}$ value related to a random distribution, i.e. $\ln[\mathrm{e}M/N]$, when $\alpha \geqslant 1$. Notice also that $S^d_{\ell,\alpha}$ does not depend explicitly on N. It depends only on value of M/N. The behaviour of $S^d_{\ell,\alpha}$ with α may be justified as follows.

Let A and B be two points on the x-axis such as $\boldsymbol{AB} = a\boldsymbol{i}$, where $\boldsymbol{i} =$ the unit vector of the x axis, and A′ and B′ the new positions of the previous points after randomization such that $\boldsymbol{A'B'} = u\boldsymbol{i}$ where u may be > 0, < 0 or $= 0$. Figure 3 illustrates two of several situations that may occur. Denote x_1 and x_2 the components of $\boldsymbol{AA'}$ and $\boldsymbol{BB'}$, respectively, and put z

$$z = u - a = x_2 - x_1\,.$$

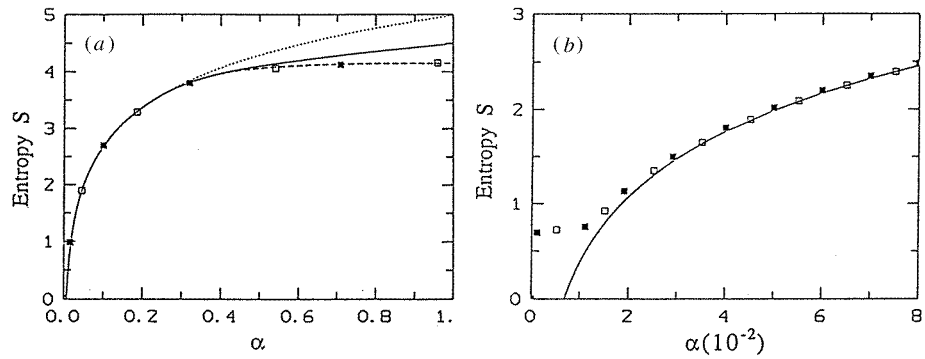

Figure 2. Entropy S versus α for non-random distributions of points. (a) Full squares: $\langle S^d_{\ell,\alpha}\rangle$ arithmetic mean of results obtained by using (3) for four distributions of $N = 2000$ points; open squares: same as full squares for $N = 6000$; full curve: $S_{\ell,\alpha}$ given by (13); dotted curve: $S_{\ell,\alpha}$ given by (15); broken curve: $S_{\ell,\alpha}$ given by (16). (b) Initial portion of S plotted on an expanded α scale and an expanded entropy scale; full squares, open squares and full curve same as (a).

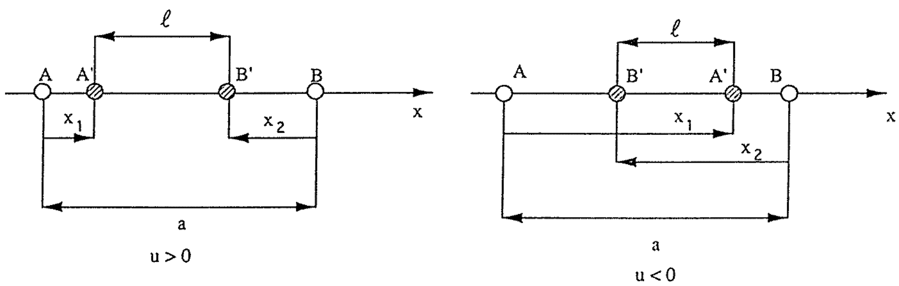

Figure 3. Two of various situations that may occur when delocalizing two points A and B of a regular 1D mosaic. A′ and B′ are the new positions of the previous points after randomization.

x_1 and x_2 are the shift related to point A and point B, respectively, and may be associated with two random variables X_1 and X_2. They are characterized by a Gaussian density probability

$$\phi_\omega(x) = \frac{1}{\omega\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\omega^2}\right]$$

where x stands for x_1 or x_2.

Given that X_1 and X_2 are independent, the density probability related to the random variable $Z = X_2 - X_1$ is then given by [15]:

$$\phi_\omega(z) = \frac{1}{2\omega\sqrt{\pi}} \exp\left[-\frac{z^2}{4\omega^2}\right].$$

The variable change $z = u - a$ and the fact that edge lengths ℓ are such that $l = u$ if $u > 0$ and $\ell = -u$ if $u < 0$, lead to

$$\Psi_\omega(\ell) = \frac{1}{2\omega\sqrt{\pi}} \left(\exp\left[-\frac{(\ell + a)^2}{4\omega^2}\right] + \exp\left[-\frac{(\ell - a)^2}{4\omega^2}\right]\right) \qquad (12)$$

which represents the edge-length distribution function when a perturbation with a standard deviation of ω is applied to the basic distribution.

Inserting (12) into (2) provides

$$S_{\ell,\alpha} = \ln\left[2\sqrt{\mathrm{e}\pi}\,\frac{M}{N}\alpha\right] + T(\rho,\alpha) \tag{13}$$

with

$$T(\rho,\alpha) = -\frac{\rho}{2\alpha\sqrt{\pi}}\int_{-\infty}^{\infty}\exp\left[-\frac{(\rho u-1)^2}{4\alpha^2}\right]\ln\left[1+\exp\left[-\frac{\rho u}{\alpha^2}\right]\right]\mathrm{d}u\,.$$

The full curve in figure 2(*a*) represents $S_{\ell,\alpha}$ given by (13). A good fit ($<$ 1%) between experimental and theoretical results is obtained for $0.08 \lesssim \alpha \lesssim 0.42$ approximately. Beyond this range (figures 2(*a*) and (*b*)) discrepancies appear. Actually when α is small enough, so that randomization does not much modify the position of the two points considered in the basic distribution, the edge-length distribution function is merely

$$\Psi_\alpha(\ell) = \frac{\rho}{2\alpha\sqrt{\pi}}\exp\left[-\frac{\rho^2(\ell-a)^2}{4\alpha^2}\right]. \tag{14}$$

Then

$$S_{\ell,\alpha} = \ln\left[2\sqrt{\mathrm{e}\pi}\,\frac{M}{N}\alpha\right]. \tag{15}$$

Equation (15) and experimental results are in good agreement (better than 1%) for $0.08 \lesssim \alpha \lesssim 0.31$. In this α range $T(\rho,\alpha)$ is negligible. The discrepancy between theory and experiment, when $\alpha \lesssim 0.08$, results from the fact that the chosen value of M/N (here 25), is no longer large enough. For $\alpha \gtrsim 0.42$ the explanation may be the following. The model used to establish (13) involves only two neighbouring points in the basic distribution. When α becomes relatively high, other points than those considered in the model come in between the two initial points, so that the Euclidean distance between these does not define an MST edge length any longer. In other words, equation (12) does not represent the real edge-length distribution function. It follows that (13) is not very adequate to describe the experimental results. Actually more than two points in the basic distribution should have been considered. But in this case, the formulation of the problem rapidly becomes very complicated. However, it is possible to obtain a good fit for $\alpha \gtrsim 0.42$ by weighting $T(\rho,\alpha)$. We found that the expression

$$S_{\ell,\alpha} = \ln\left[2\sqrt{\mathrm{e}\pi}\,\frac{M}{N}\alpha\right] + 1.7\sqrt{\alpha}\,T(\rho,\alpha) \tag{16}$$

fitted the experimental results quite well (see figure 2(*a*)).

From equations (9), (13), (15) and (16), it can be deduced that the experimental entropy S_ℓ^d (S_ℓ^d now refers to $S_{\ell,\mathrm{r}}^d$ and $S_{\ell,\alpha}^d$, both of which are evaluated from (3) by using adequate values of M/N) is well described by the general expression

$$S_\ell^d = \ln\left[\frac{M}{N}\right] + G(\rho,\alpha)\,. \tag{17}$$

The determination of the adequate values of M/N is discussed in section 4.5.

4.4. Parameter of order

The computation of $G(\rho,\alpha)$ from (17) for various simulated arrangements of points, shows that the term G does not depend on ρ but only on α, i.e. on the way the points are distributed.

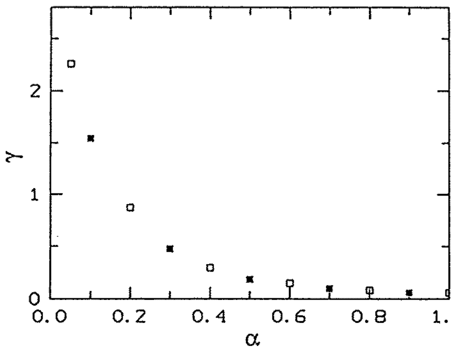

Figure 4. Parameter γ versus α for various simulated distributions of points. γ has been evaluated by using adequate values of M/N for each distribution following the procedure proposed at the end of section 4.5. Full squares: various simulated distributions with $N = 2000$; open squares: same as full squares for $N = 6000$.

This result is in good agreement with the theoretical prediction given by (6) which may also be written as

$$S_\ell = \ln\left[\frac{M}{N}\right] + F\,.$$

Moreover, we find that G is a monotonically increasing function of α, that may therefore be used to quantitatively characterize order (or disorder) in the distributions of points. Actually it seems more convenient, for such a characterization, to use the parameter of order γ defined by (8). Figure 4 shows $\gamma = 1 - G(\alpha)$ for different distributions of points. Also note that

$$\gamma = \ln\left[\frac{eM}{N}\right] - S_\ell^d = S_{\ell,\mathrm{r}} - S_\ell^d\,. \tag{18}$$

Thus, given a distribution of N points, γ measures the difference between the entropy function related to N randomly scattered points and the entropy function related to the actual distribution of N points.

4.5. Determination of adequate value of M when evaluating entropy function S_ℓ^d

The purpose is to determine the M value so that the difference $D = S_\ell - S_\ell^d$ will be minimal. Let us first consider a random distribution of N points. The standard deviation related to edge lengths and deduced from (5) is $\sigma = 1/N$ by taking $L = 1$. Theoretical values of D for such a distribution are given by $D_{\mathrm{th}} = S_{\ell,\mathrm{r}} - (S_{\ell,\mathrm{r}}^d)_{\mathrm{th2}}$. The variations of D_{th}, versus $M\sigma$, for $N = 2000$ are shown in figure 5. On the same figure are also reported some values of $D_\mathrm{r} = S_{\ell,\mathrm{r}} - S_{\ell,\mathrm{r}}^d$ related to a simulated random distribution ($N = 2000$), and some values of $D_\alpha = S_{\ell,\alpha} - S_{\ell,\alpha}^d$ corresponding to a weakly perturbed mosaic ($N = 2000$) characterized by the edge-length distribution function $\Psi_\alpha(\ell)$ (equation (14)). These results show that, for a given N, D does not depend on the way the points are scattered, and that D is well fitted by D_{th}. The adequate value of $M\sigma$ or M is obtained for $D_{\mathrm{th}} = 0$. It is then possible to compute the adequate value of $M\sigma$ for various N (figure 6) and in particular to put forward the following analytical approach:

$$c_0 + \frac{c_1}{N} + c_2 N + c_3 N^2 + c_4 N \ln[N] \qquad \text{for} \quad N \in (100, 10\,000) \tag{19}$$

with $c_0 = 1.184$, $c_1 = -29.4$, $c_2 = 3.463 \times 10^{-3}$, $c_3 = 2.216 \times 10^{-8}$, $c_4 = -3.648 \times 10^{-4}$.

Eventually, the practical procedure to evaluate the parameter of order γ related to an arrangement of N points scattered along a segment of length L is as follows:

(i) Construct the MST from the set of points and determine the edge lengths. Divide the edge-length values by L in order to reduce the study to a segment of length $L = 1$.

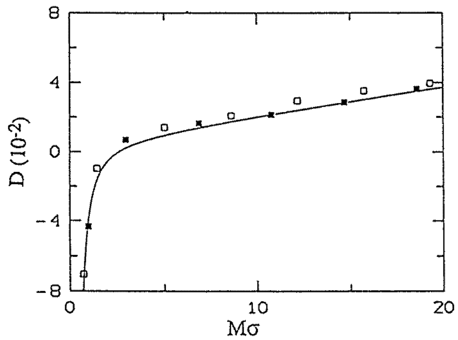

Figure 5. $D = S_\ell - S_\ell^d$, versus $M\sigma$, for distributions of $N = 2000$ points. Full curve: $D_{\rm th} = S_{\ell,\rm r} - (S_{\ell,\rm r}^d)_{\rm th2}$. Full squares: $D_{\rm r} = S_{\ell,\rm r} - S_{\ell,\rm r}^d$ for a simulated random distribution; open squares: $D_\alpha = S_{\ell,\alpha} - S_{\ell,\alpha}^d$ for a mosaic perturbed with $\alpha = 0.2$. For both full and open squares the value of σ taken into account is the real standard deviation deduced from the simulated distributions.

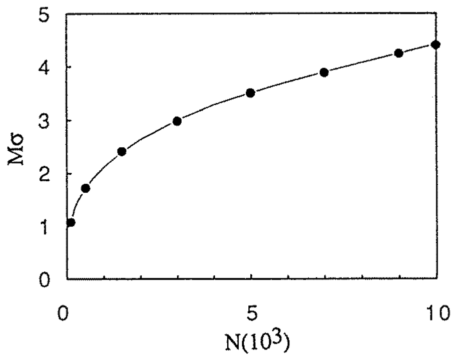

Figure 6. Adequate values of $M\sigma$ versus N. Full circles: computed adequate values of $M\sigma$; full curve: analytical approach given by (19).

(ii) Compute the real standard deviation σ related to the edge lengths.
(iii) Use the adequate value of M deduced from (19) to compute S_ℓ^d by means of (3).
(iv) Lastly deduce γ from (18).

5. Conclusion

In this paper, we have proposed a new parameter γ for quantizing the degree of order in a 1D distributions of points. This parameter involves an entropy function related to the edge lengths of the MST constructed from the set of points. Theoretical predictions have been confirmed by investigating the simulated distributions of points, i.e. both random and non-random distributions obtained by perturbing a regular 1D mosaic using a Gaussian process. Besides this, a method which is particularly easy to implement and computationally efficient has been proposed for evaluating γ in all situations, i.e. simulated distributions as well as distributions resulting from experiments. This parameter should turn out to be very useful in various fields of physics, biology and medicine. Eventually this study may be extended to the 2D distributions of points. Such an approach is being developed.

References

[1] Dussert C, Rasigni G, Rasigni M and Palmari J 1986 *Phys. Rev.* B **34** 3528
[2] Hofman R and Jain A K 1983 *Patt. Recogn. Lett.* **1** 175
[3] Dussert C, Rasigni M, Palmari J, Rasigni G, Llebaria A and Marty F 1987 *J. Theor. Biol.* **125** 317
[4] Dussert C, Rasigni G and Llebaria A 1988 *J. Theor. Biol.* **135** 295
[5] Dussert C, Kopp F, Gandilhon P, Pourreau-Schneider N, Rasigni M, Palmari J, Rasigni G, Martin P M and Llebaria A 1989 *Anal. Cellular Pathology* **1** 123

[6] Dussert C, Mulliert G, Kellershonn N, Ricard J, Giordani R, Noat G, Palmari J, Rasigni M, Llebaria A and Rasigni G 1989 *Eur. J. Biochem.* **185** 281
[7] Dong X F, Berthois Y, Palmari J, Isnardon D, Dussert C and Martin P M 1992 *Anticancer Res.* **12** 2085
[8] Dussert C, Rasigni G, Rasigni M, Palmari J and Llebaria A 1989 *Z. Phys.* D **12** 41
[9] Dussert C, Rasigni G and Llebaria A 1989 *Thin Solid Films* **177** 245
[10] Dussert C, Rasigni G and Rasigni M 1989 *Phys. Lett.* **139A** 35
[11] Billia B, Jamgotchian H and Nguyen Thi H 1991 *Metallurgical Trans.* A **22** 3041
[12] Subramanian Yashonath and Jyothi Bhasu V C 1992 *Chem. Phys. Lett.* **189** 311
[13] Diu B, Guthmann C, Lederer D and Roulet B 1989 *Physique Statistique* (Paris: Hermann)
[14] Knuth D E 1969 *The Art of Computer Programming* vol 2 (New York: Addison-Wesley)
[15] Beckmann P 1967 *Probability in Communication Engineering* (New York: Harcourt Brace)