Information entropy of complex structures

Clinton DeW. Van Siclen*
Idaho National Engineering Laboratory, P. O. Box 1625, Idaho Falls, Idaho 83415
(Received 10 June 1997)

The information entropy function provides a sensitive measure of the complexity of a multi-component material system, where ''complexity'' refers to the range of length scales over which morphological features are present. This is demonstrated for an evolving, two-phase microstructure simulated by a population of interacting particles on a two-dimensional surface. The information entropy increases at all length scales as the initially random configuration of particles evolves to produce a distribution of ramified clusters. Maxima in the *normalized* information entropy function, which is obtained by subtracting the information entropy of a perfectly random configuration from that of the clustered configuration, occur at length scales for which the system most differs from a random configuration, while minima occur at length scales for which the system is periodic or relatively ordered. Besides analysis of complex microstructures, information entropy is useful in detecting features present in any collection of data. [S1063-651X(97)04811-3]

PACS number(s): 05.90.+m, 02.50.−r, 05.40.+j

I. INTRODUCTION

The physical properties of materials are determined primarily by their microstructure. Quantitative characterization of the microstructure is thus essential both to correlate microstructure attributes with observed properties, and to predict properties through equations of motion, incorporating such attributes, that describe physical phenomena of interest [1]. In principle, a complete set of n-point spatial correlation functions can completely describe the microstructure [2]. However, correlation functions beyond $n=2$ are difficult to calculate in practice, even for a digitized image of the microstructure. It is thus important to formulate alternative statistical descriptions of complex materials.

Beghdadi and co-workers [3,4] introduced the concept of ''configurational entropy'' as a morphological descriptor for heterogeneous materials. This entropy, adapted from Shannon's information theory [5], is a measure of the local fluctuations of some material attribute such as phase volume fraction, over the system. Boger *et al.* [6] developed the similar ''local porosity entropy'' to find the length scale L^* that maximizes the geometrical content (again, fluctuations) of a system, in order to obtain an optimal local porosity distribution useful for calculating physical properties of a porous media. An exact relationship between these two entropies was recently established by Andraud *et al.* [7]. A more straightforward measure of phase volume fraction fluctuations is given by the ''coarseness'' C introduced by Lu and Torquato [8]. This quantity is proportional to the standard deviation of the local volume fraction, and is related to the two-point probability function S_2.

The present work considers a more general ''information entropy'' H, so called because it corresponds to the average information content of the system. The distinctions between the information entropy and the configurational or local porosity entropy are noted during the formulation of H in Sec. II.

The information entropy is then calculated for a model system consisting of interacting particles on a surface, to show that the function H can uniquely characterize a system over all length scales. The physical significance of the entropy function for this system is found by monitoring the changes in H as the system evolves. Maxima and minima in the normalized H (obtained by subtracting the information entropy for the random particle configuration) are shown to correspond to clustering (of particles and of particle clusters) and to periodicity or ordering, respectively.

II. INFORMATION ENTROPY

The application of information theory [5] to microstructure characterization is made by specifying that the information content of a particular volume or area found in the state i is proportional to $-\log p_i$, where p_i is the *a priori* probability of finding that volume or area in the ith state (or configuration). The states accessible to the system correspond to all possible values of a specified microscopic material attribute, such as phase volume fraction or length of interphase boundary per unit cross-section area. The functional dependence of information content on the state probability derives from the fact that a sampled volume or area that possesses an unusual value for a material attribute conveys more information about the microstructure than does a sample that possesses the average value for the attribute. The logarithmic function is necessary so that information content is additive: the information content $I(p_1 p_2)$ of a system comprised of two regions in states 1 and 2, respectively, must equal the sum $I(p_1)+I(p_2)$. The average value of the information content (averaged over all volumes or areas comprising the system) is then the information entropy H for the system, for the length scale given by the size of the sampled volumes or areas.

The information entropy differs from the configurational-local porosity entropy mainly in the assignment of the probabilities p. For the latter, p_i is the *actual* probability of finding a particular volume or area in the state i, so that the system is completely examined prior to making the assign-

*Electronic address: cvs@inel.gov

FIG. 1. Snapshot of a population of interacting particles that represents an evolving, two-phase microstructure. This 625-particle configuration on a 50×50 grid produces the normalized information entropy $H'(m)$ curve calculated at time $t=1000$ in Fig. 4.

ments. (Such prior knowledge provides ''constraints'' [9] incorporated into the probabilities, so that the total information content of the system is reduced.) The configurational-local porosity entropy is thus sensitive to *fluctuations* in the value of the selected material attribute over the system, at the length scale given by the size of the volume or area used to determine the set of probabilities. In contrast, the information entropy is a measure of the *deviation* of the value (or distribution of values) of a material attribute from an initial or expected value (or distribution of values), on the length scale corresponding to the volume or area size [10]. A more sensitive measure is the *normalized* information entropy H', where the initial or expected information entropy is subtracted from the information entropy calculated for the material system.

III. APPLICATION TO AN EVOLVING TWO-PHASE SYSTEM

To demonstrate the sensitivity of the information entropy to microstructure inhomogeneity, $H(m)$ is calculated for a simple two-dimensional, two-phase system, for all length scales m [11]. The system consists of a collection of N particles (each of size 1×1) initially placed randomly on a square grid of side length L, that interact weakly to form an irregular distribution of ramified clusters as shown in Fig. 1. The information content of a $m\times m$ square region (or ''box'') within this system that is found to contain n particles is $-\log p_n$, where p_n is the probability of finding exactly n particles in that $m\times m$ square region if the N particles were instead distributed perfectly randomly over the system. Thus the system is *presupposed* (i.e., prior to examination) to be in a random configuration, so that deviations from the random particle distribution represented by the set of probabilities $\{p\}$, that will occur as the system evolves, provide increased information content. The average value of the information content, taken over all $m\times m$ regions comprising the system [which number $(L-m+1)^2$], is the entropy $H(m)$.

The set $\{p\}$ for the finite, random system is comprised of elements

$$p_i(m)=\binom{m^2}{i}\binom{L^2-m^2}{N-i}\binom{L^2}{N}^{-1}$$
$$=\frac{(m^2)!}{i!(m^2-i)!}\,\frac{(L^2-m^2)!}{(N-i)!(L^2-m^2-N+i)!}\,\frac{N!(L^2-N)!}{(L^2)!}, \tag{1}$$

where i runs from the larger of 0 and $m^2-(L^2-N)$ to the lesser of N and m^2. The information entropy for the finite, perfectly random, system is then

$$H_r(m)=-\sum_i p_i(m)\log[p_i(m)], \tag{2}$$

which is symmetric about its maximum at $N=L^2/2$, for all m, when plotted against particle number N, since the set $\{p\}$ for the finite, random system with particle coverage ϕ is identical to the set for the complementary system with particle coverage $1-\phi$. The entropy $H_r(m)$ is also symmetric about its maximum at $m^2=L^2/2$, for all N, when plotted against sample box size m^2 [Eq. (1) is unchanged under the replacement $m^2\leftrightarrow N$], and equals zero at $m=L$ since no (new) information can be gained by sampling the system at that length scale ($p_N=1$ for $m=L$). In each case, the maximum in $H_r(m)$ coincides with the maximum in the number of states accessible to the system (i.e., the number of elements in $\{p\}$ is maximal).

The information entropy $H(m)$ for a given configuration of particles is then

$$H(m)=-\sum_i P_i(m)\log[p_i(m)], \tag{3}$$

where p_i is taken from Eq. (1), and P_i is the actual probability of finding exactly i particles in *any* $m\times m$ region sampled from the system. The set of probabilities $\{P\}$ for box size m^2 corresponds to the local porosity distribution [6] when particle density is identified with porosity.

Figure 2 shows the information entropy $H(m)$ for several initial, ''pseudorandom'' configurations of 625 particles ($\phi=\frac{1}{4}$) obtained by placing the particles, one at a time, at random positions on a 50×50 grid, together with $H_r(m)$ for a perfectly random configuration calculated from Eq. (2) [12]. The deviation of the former curves from the latter reflects the finite size of the system in two ways: (a) the pseudorandom particle configurations are themselves randomly chosen from the set of all possible configurations of 625 particles on a 50×50 grid, so that P_i cannot be expected to equal p_i for all i; and (b) fewer $m\times m$ boxes are sampled as the box size approaches the system size, so that deviations are generally largest at large m. Note that for this reason, the information entropy may not be a useful descriptor for length scales near

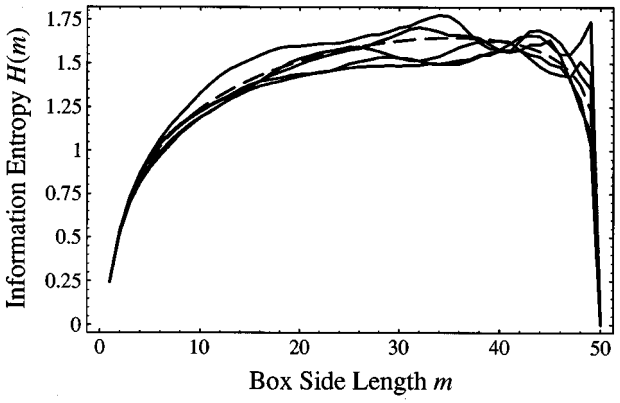

FIG. 2. Information entropy $H(m)$ for several ''pseudorandom'' configurations of particles on a finite space. These curves deviate from the corresponding $H_r(m)$, shown as a dashed curve, for a perfectly random configuration on that space.

that of the system size. (However, the problem of large deviations at box sizes approaching the system size may be ameliorated by considering the $L\times L$ system to be infinitely periodic rather than finite, so that $m\times m$ boxes may effectively ''wrap around'' the system edges.)

The effects of nonrandom particle distributions are more obvious in the corresponding *normalized* information entropy $H'(m)=H(m)-H_r(m)$ shown in Fig. 3. Values for $H'(m)$ greater (less) than zero indicate length scales for which particle clustering is more (less) prevalent than occurs for a perfectly random particle configuration on a finite space.

Ramified clusters are formed as the particles diffuse randomly over the surface, due to a small binding energy between adjacent particles. This is effected by reducing the probability, from one to one-fifth, that a particle will move to an adjacent empty site when the particle has one or more

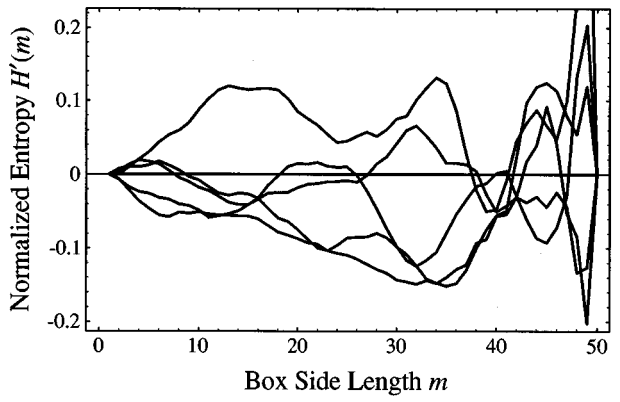

FIG. 3. Difference between $H(m)$ and $H_r(m)$ in Fig. 2, plotted as the normalized information entropy $H'(m)$. The deviation of the curves from $H'(m)=0$ is due to the imperfectly random initial placement of the particles. Values greater than zero occur at length scales m for which particle clustering is greater than occurs for a perfectly random configuration, while values less than zero occur at length scales for which the particle distribution is more ordered or regular than occurs for a perfectly random configuration.

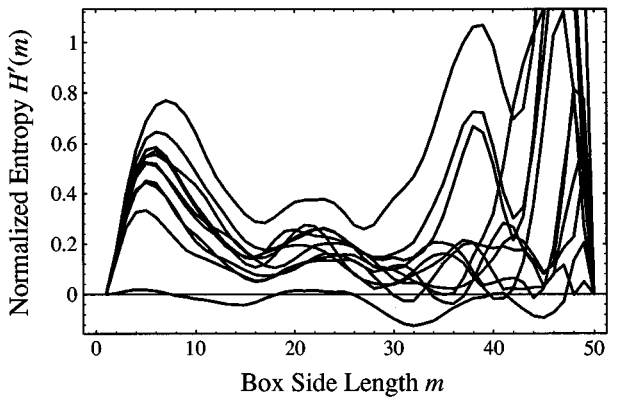

FIG. 4. Typical sequence of $H'(m)$ curves calculated for a population of interacting particles at times $t=0,100,200,...,1000$. Nucleation and growth of small clusters are indicated by the increase in height and shift in position of the first maximum at length scale $m\approx 5$ over this time period. Larger-scale features are indicated by the maxima at larger box sizes. The particle configuration at $t=1000$ is shown in Fig. 1.

neighbors (the binding energy is thus approximately 0.04 eV when the system is considered to be at room temperature). The time t associated with the evolution of this system is expressed in Monte Carlo time steps, where one time step is completed after every particle has had a single opportunity to move to an adjacent site.

Figure 4 presents a typical sequence of $H'(m)$ curves calculated for the particle population at $t=0,100,200,\ldots,1000$. Cluster nucleation and growth is indicated by the immediate rise in $H'(m)$ at small m and the subsequent shift in the position of that first maximum to slightly larger m. The additional maxima correspond to ''clusters of clusters'' present at larger length scales, that are seen to arise from nonrandom inhomogeneities in the initial distribution of particles. Consistent with this interpretation, Fig. 5 shows curves for $H'(m)$ calculated at $t=100$ for several initially

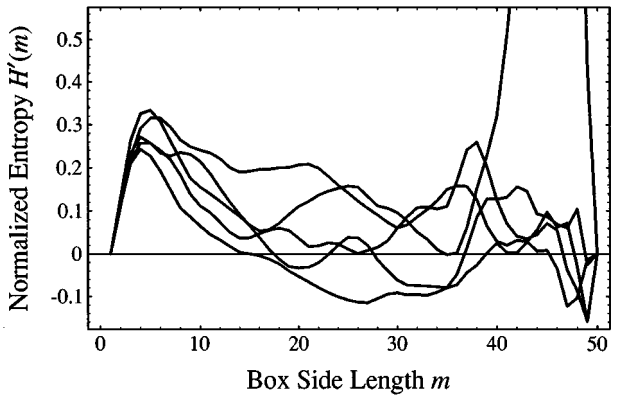

FIG. 5. Normalized information entropy $H'(m)$ curves plotted at time $t=100$, for several initially random configurations of particles (those producing Figs. 2 and 3). In every case, the first maximum indicates small cluster formation, while the maxima at larger box sizes indicate larger-scale features that arise from the imperfectly random initial placement of the particles.

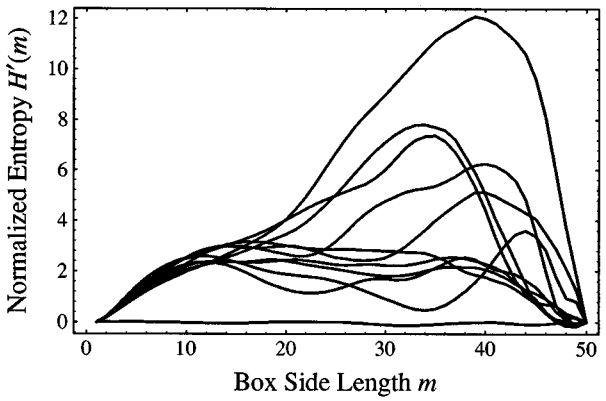

FIG. 6. Normalized information entropy $H'(m)$ curves calculated for the particle population producing Fig. 4, at time $t=0$ and later times $t=10^5,2\times10^5,\ldots,10^6$. Except for the curve calculated for the initial, random configuration at $t=0$, all curves correspond to equilibrium configurations of particles. No correlations exist among the set of curves; for example, the curve for which $H'(39)=12$ is obtained at $t=7\times10^5$.

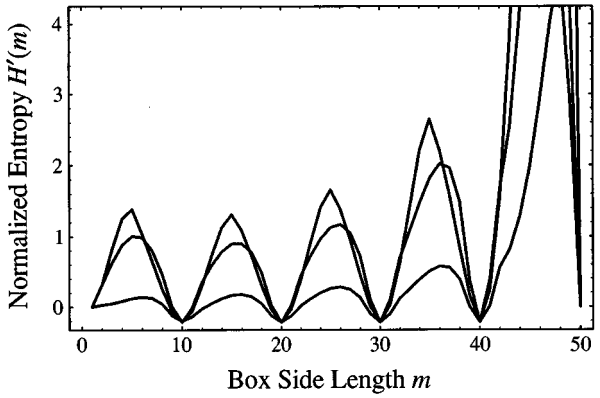

FIG. 7. Normalized information entropy $H'(m)$ curves calculated for periodic configurations of particles. Each of the three configurations has a 10×10 unit cell within which is a single square cluster of 2^2, 5^2, or 8^2 particles, respectively; the corresponding curves have first maxima $H'(6)=0.14$, $H'(5)=1.39$, and $H'(5)=1.01$. All three curves have minima at length scales coincident with the periodicity of the particle configurations.

random particle configurations (those producing Figs. 2 and 3) that have nearly identical first maxima (corresponding to cluster formation) and very different additional maxima (corresponding to larger-scale features).

The extreme sensitivity of the information entropy to microstructure features at all length scales is evident in Fig. 6, which presents $H'(m)$ curves calculated for the particle population (producing Fig. 4) at $t=10^5,2\times10^5,\ldots,10^6$. These wildly varying curves are produced *after* the system has reached equilibrium, as indicated by an essentially constant ratio of successful particle jumps to attempted jumps over each 10^5 time interval. (As only free particles and particles at a cluster periphery can perform jumps, their populations must be stable to produce a constant jump ratio.) This variability would be reduced at larger system sizes (with same particle coverage ϕ) as a greater variety of cluster configurations could then coexist.

Despite the incontrovertible interpretation of the first maxima in Figs. 4 and 5 as indicative of small cluster formation, it is incorrect to assume that the various maxima in the $H'(m)$ curves specify the size of particle clusters in general. This is evident by considering that particle clustering produces large $-\log p_i$ values both by *increasing* the number of particles found in a box of size $m\times m$ over that number expected for a random configuration, and by *reducing* the number of particles found in other boxes of similar size elsewhere in the system. Indeed, in the case $\phi=\frac{1}{2}$, a box filled by a cluster has equal information value with a box completely empty of particles. Figure 7 shows $H'(m)$ curves for periodic configurations of particles in which the 10×10 unit cell contains a single 2×2, 5×5, or 8×8 particle cluster, respectively. The periodicity of the microstructure is given by the positions of the minima in $H'(m)$, while the maxima incorporate contributions from the regions devoid of particles so their positions cannot be identified with a cluster size. The maxima instead occur at length scales at which the particle distribution most deviates from a random distribution. Similar $H'(m)$ curves are obtained for other particle coverages and particle jump probabilities (leading to different cluster sizes and shapes), and for other system sizes.

IV. DISCUSSION

A complex microstructure possesses distinctive features at many length scales. Such ''complexity,'' which reflects the range of length scales over which morphological features are present, is quantified by the information entropy function H. This has been demonstrated for a system of interacting particles, where $H(m)$ increases as the initially random configuration of particles evolves to produce complex structures.

For application to real microstructures, it may be difficult to determine the set of *a priori* probabilities $\{p\}$ that describes the expected distribution of values of a material attribute (for a continuous distribution, the set $\{p\}$ is replaced by the function p). A set of random probabilities like that used above may be appropriate for phase volume fraction when phase volumes are conserved, but cannot be used for an attribute such as length of interphase boundary per unit cross-section area. In such cases the set of actual probabilities $\{P\}$ found for a dynamic system at an initial time may serve as the set $\{p\}$ at later times, with the caveat that most microstructural attributes are not conserved. More generally, the information entropy may be calculated as described above to detect features at various size scales in any collection of data.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Energy, Office of Basic Energy Sciences, and the INEL Laboratory Directed Research & Development Program under DOE Idaho Operations Office Contract No. DE-AC07-94ID13223.

[1] R. Hilfer, Adv. Chem. Phys. **XCII**, 299 (1996).
[2] S. Torquato and G. Stell, J. Chem. Phys. **77**, 2071 (1982); S. Torquato, J. Stat. Phys. **45**, 843 (1986); Appl. Mech. Rev. **44**, 37 (1991); Physica A **207**, 79 (1994).
[3] A. Beghdadi, C. Andraud, J. Lafait, J. Peiro, and M. Perreau, Fractals **1**, 671 (1993).
[4] C. Andraud, A. Beghdadi, and J. Lafait, Physica A **207**, 208 (1994).
[5] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois, Urbana, 1959).
[6] F. Boger, J. Feder, T. Jøssang, and R. Hilfer, Physica A **187**, 55 (1992).
[7] C. Andraud, A. Beghdadi, E. Haslund, R. Hilfer, J. Lafait, and B. Virgin, Physica A **235**, 307 (1997).
[8] B. Lu and S. Torquato, J. Chem. Phys. **93**, 3452 (1990).
[9] L. Brillouin, *Science and Information Theory* (Academic, New York, 1956).
[10] For example, the character string ''*zzzzz*'' has a configurational-local porosity entropy of zero since all elements of the string are identical, while it has information entropy (average information content per character) equal to $-\log_{10}(0.001)$, reflecting the infrequency of occurrence of the letter *z* in the English language.
[11] The base 10 logarithm is used throughout this section to produce the figures; there is no technical reason to prefer one base over another.
[12] A mathematical identity useful for calculating p_i and $\log p_i$ for large systems is $\log[\binom{\alpha}{\beta}]=\sum_{k=1}^{\alpha-\beta}\log[(k+\beta)/k]$.