Clustering of points randomly distributed in *n*-dimensional space

G. Sadasiv and Yirong Meng
Department of Electrical Engineering, University of Rhode Island, Kingston, Rhode Island 02881
(Received 29 August 2000; published 9 January 2001)

We consider clusters formed by points randomly distributed in space, each point being connected to its nearest neighbor or to its nearest and next nearest neighbors. The size distribution of such clusters in *n*-dimensional space is presented.

DOI: 10.1103/PhysRevE.63.027101 PACS number(s): 64.60.Ak

I. INTRODUCTION

The formation of clusters and the resulting cluster size distribution is of great interest in studying the processes of aggregation, flocculation, and polymerization. Models of cluster formation find applications in theoretical and experimental research in many fields in physics, chemistry, biology, and astronomy. They are also of importance in practical applications as in the production of thin films for electronics [1]. In studying the statistical mechanics of phase transitions [2], or percolation problems [3], the starting model uses nearest neighbor interactions between entities placed on a lattice. In this connection, it is of interest to examine aggregates formed by nearest neighbor connections of randomly distributed points.

In this paper we present the results of such a study. We consider dimensionless points randomly distributed in (essentially infinite) space, i.e., the probability that a point is to be found in an infinitesimal volume dv is proportional to dv. In clusters of type I (Fig. 1), every point is considered as connected to its nearest neighbor. In clusters of type II (Fig. 2), every point is considered to be connected to its nearest neighbor and its next nearest neighbor. All points that are connected form a cluster. The number of points in a cluster is the size of the cluster. We are interested in the cluster size distribution in *n*-dimensional space.

An analytical result can be obtained for points in one-dimensional space, and is given in the Appendix. For higher dimensions we were unable to obtain an analytical solution but we did computer simulations to find the cluster size distribution.

II. SIMULATION PROGRAM

In our simulation, each coordinate of a point in *n*-dimensional space is obtained from the default random number generator in a C++ library. The number of sample points along one axis is chosen to be much less than the range of random numbers. The probability of a point having two neighbors at the same distance is negligible.

In order to reduce the calculations for finding the nearest neighbor of any point, the collection is blocked into a square lattice, the size of the unit cell being appropriately chosen so as to have a sufficient number of points in each cell. In most cases the block size was chosen to give an average of about 20 points per block. In this way only the ''local area'' around a point has to be searched to find its nearest neighbor.

Since we work with a finite number of points, and the boundary should not influence the calculation, we used periodic boundary conditions in our simulation to extend the space. For points belonging to the blocks on the boundary, we can extend the ''local area'' to the blocks on the opposite edge. With this assumption, in *n*-dimensional space, if the coordinates of two points are $(a_1,a_2,\ldots,a_n)$ and $(b_1,b_2,\ldots,b_n)$, the distance between these two points is

$$d=\sqrt{\sum_{i=1}^{n}\{\min[|a_i-b_i|,(1-|a_i-b_i|)]\}^2}. \qquad (1)$$

III. SIMULATION RESULTS FOR TYPE I CLUSTERS

The distribution of clusters of different sizes in various dimensions is summarized in Table I. The data are obtained by running the program 1000 times with 1000 sample points per run for clusters in one dimension, 20 times with 100 000 sample points per run for clusters in two dimensions, 30 times with 100 000 sample points per run for clusters in three dimensions, and 10 times with 200 000 sample points per run for clusters in four dimensions.

The ratios of the number of points which belong to a given size cluster to the total number of points are shown in Table II. The results are also shown in Fig. 3(a) and Fig. 3(b). The following characteristics may be noted. (1) In any dimension, the distribution has maximum value for clusters of size 2 and decreases rapidly. The rate of decrease is greater than exponential but less than Gaussian [$1<b<2$ in Eq. (2)]. (2) With increasing dimension, the probability for smaller clusters decreases and for larger clusters increases. (3) In all cases, the size 3 clusters have more points than any other size clusters [shown in Fig. 3(b)].

By applying linear regression analysis to the results shown above, we found that the cluster size distribution $f(z)$ in n dimensions is given approximately by

$$f(z)=A\exp\{-[(z-2)/k]^b\}, \qquad (2)$$

where $z=2,3,4,\ldots$ is the size of the cluster, and A,k,b are parameters which have different values in different dimensions. The best fitting parameter values are shown in Table III.

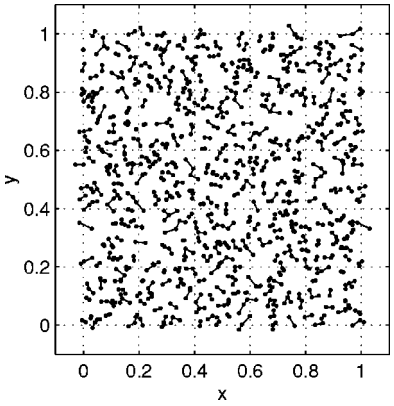

FIG. 1. Cluster of type I in two dimensions, 1000 sample points. Every point is connected to its nearest neighbor to form clusters

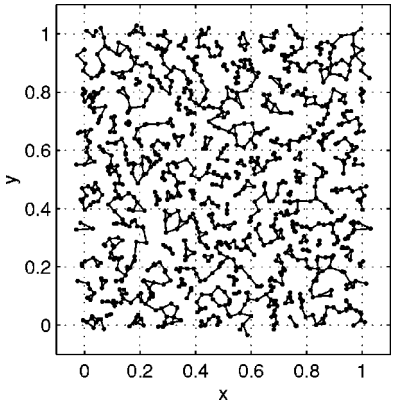

FIG. 2. Clusters of type II in two dimensions, 1000 sample points. Every point is connected to its nearest and next nearest neighbor to form clusters.

IV. SIMULATION RESULTS FOR TYPE II CLUSTERS

In two dimensions, bigger clusters are formed. But these clusters are still isolated islands, as shown in Fig. 2. When the simulation was run 10 times with 10 000 points per run, the results were roughly as follows for each run: there were about 100 clusters of three points, decreasing to 12 clusters of 10 points, then two or three clusters in the range 10 to 40 points per cluster. The rest of the points were scattered in various sized single clusters ranging all the way to clusters with a few hundred points.

There is a striking change, reminiscent of a phase transition, when going to higher dimensions. The result of four simulations with 1 000 000 sample points in three dimensions is as follows. With the exception of a few small isolated islands, the points collapse into one large cluster that contains more than 92% of all the points. The next largest cluster contains less than 0.01% of the total number of points.

This trend is more marked on going to four dimensions. The result of one simulation with 100 000 points shows more than 97 000 points in one large cluster, with the next largest cluster having fewer than 40 points.

We obtain the very interesting result that, for random points in three or higher dimensions, if a point is connected to it's nearest neighbor and next nearest neighbor, almost all the points in space are connected together.

APPENDIX: THEORETICAL ANALYSIS FOR CLUSTERS OF TYPE I IN ONE DIMENSION

In one dimension, we can put all points in order to form a sequence. To find the nearest neighbor of a point, only two distances need to be compared, the distance to its left neighbor (r_{left}) and the distance to its right neighbor (r_{right}). We denote this point by ''1'' if ($r_{left}<r_{right}$) or by ''0'' if ($r_{left}>r_{right}$). Then, a size N cluster can be expressed by the sequence shown in Fig. 4, which includes m continuous ''0''

TABLE I. The ratios of numbers of given size clusters to the total number of clusters for type I clusters in one to four dimension(s). The values are the averages of the results of several runs. The mean square deviations referred to the average are also calculated.

Size	1D	2D	3D	4D
2	0.3995±6.8%	0.3685±0.66%	0.3501±0.55%	0.3368±0.58%
3	0.3338±7.9%	0.3020±0.63%	0.2840±0.92%	0.2733±0.58%
4	0.1711±12%	0.1770±1.2%	0.1768±1.3%	0.1748±0.51%
5	0.0668±21%	0.0877±1.7%	0.0967±1.6%	0.1015±1.6%
6	0.0211±38%	0.0392±2.5%	0.0490±2.2%	0.0556±1.9%
7	0.0059±73%	0.0162±4.8%	0.0237±3.0%	0.0292±2.9%
8	0.0013±150%	0.0061±8.4%	0.0110±4.9%	0.0149±1.8%
9	0.0003±310%	0.0021±15%	0.0050±9.0%	0.0072±3.3%
10	< 0.0001	0.0008±21%	0.0021±13%	0.0034±2.7%
11		0.0002±25%	0.0009±17%	0.0017±8.7%
12		0.0001±61%	0.0003±34%	0.0008±11%
13		< 0.0001	0.0002±41%	0.0003±16%
14			0.0001±80%	0.0002±29%
15			< 0.0001	0.0001±70%
16				< 0.0001

TABLE II. The ratios of the numbers of points in different size clusters to the total number of points for type I clusters.

Size	1D	2D	3D	4D
2	0.2666	0.2291	0.2075	0.1928
3	0.3338	0.2816	0.2525	0.2347
4	0.2279	0.2200	0.2096	0.2002
5	0.1111	0.1363	0.1432	0.1453
6	0.0421	0.0731	0.0871	0.0954
7	0.0137	0.0353	0.0492	0.0585
8	0.0035	0.0151	0.0261	0.0341
9	0.0010	0.0060	0.0134	0.0187
10	0.0001	0.0024	0.0062	0.0099
11	0.0001	0.0007	0.0029	0.0053
12	<0.0001	0.0003	0.0012	0.0027
13		<0.0001	0.0006	0.0012
14			0.0002	0.0006
15			0.0001	0.0003
16			<0.0001	0.0002
17				0.0001

points followed by $N-m$ continuous ''1'' points (m is an integer in the range 1 to $N-1$). The boundaries of the cluster exist between ''1 0.''

$N+3$ distances should be considered to calculate the relative probability of the sequence shown in Fig. 4. The relations between them are

$$r_0<r_1,\quad r_1>r_2>\cdots>r_m>r_{m+1},\quad r_{m+1}<r_{m+2}\times<\cdots<r_N<r_{N+1},\quad r_{N+1}>r_{N+2},\quad 0<r_i<\infty. \tag{A1}$$

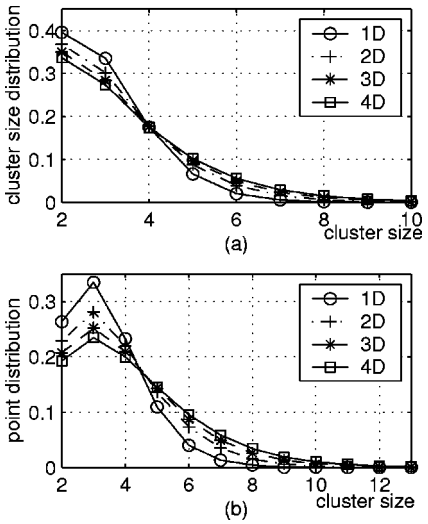

FIG. 3. (a) Size distribution of type I clusters in 1–4 dimension(s). (b) Point distribution of type I clusters in 1–4 dimension(s). The conditions of the simulation are the following: run program 1000 times with 1000 points per run in one dimension, 20 times with 100 000 points per run in two dimensions, 30 times with 100 000 points per run in three dimensions, and 10 times with 200 000 points per run in four dimensions.

If Eq. (A1) is satisfied and if the distance distribution of nearest neighbors is $f(r)$, the relative probability of a size N cluster with m ''0'' points can be expressed by

$$P_m^N=\int_0^\infty f(r_0)dr_0\int_{r_0}^\infty f(r_1)dr_1\cdots\int_{r_N}^\infty f(r_{N+1})dr_{N+1}\times\int_0^{r_{N+1}} f(r_{N+21})dr_{N+2}. \tag{A2}$$

Then the relative probability of a size N cluster is

$$P_r(N)=\sum_{m=1}^{N-1} P_m^N. \tag{A3}$$

From the definition of cluster type I, $f(r)$ satisfies the classical distance distribution of nearest neighbors [4]. In one dimension,

$$f(r)=2\rho\exp(-2\rho r) \tag{A4}$$

where ρ is the average density. Inserting Eq. (A4) into Eq. (A2) and setting $x=2\rho r$, we get

$$P_m^N+P_{m-1}^N=C_m\frac{N-m+1}{(N-m+2)!}, \tag{A5}$$

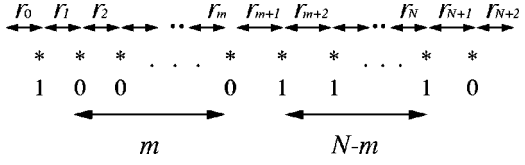

FIG. 4. Cluster of type I in one dimension.

TABLE III. Best fitting formula for size distribution of type I clusters.

Dimension	Cluster size distribution function $f(z)$
1	$0.3995\exp\left[-\left(\frac{z-2}{2.16}\right)^{1.69}\right]$
2	$0.3685\exp\left[-\left(\frac{z-2}{2.36}\right)^{1.49}\right]$
3	$0.3501\exp\left[-\left(\frac{z-2}{2.49}\right)^{1.39}\right]$
4	$0.3368\exp\left[-\left(\frac{z-2}{2.48}\right)^{1.28}\right]$

$$C_m=\int_0^\infty \exp(-x_0)dx_0\int_{x_0}^\infty \exp(-x_1)dx_1\int_0^{x_1}\times\exp(-x_2)dx_2\cdots\int_0^{x_{m-1}}\exp(-x_m)dx_m=\frac{m}{(m+1)!}. \tag{A6}$$

When N is *odd*,

$$P_r(N)=\sum_{m=1}^{N-1}P_m^N=\sum_{m=2,4,\ldots,N-1}^{N-1}C_m\frac{N-m+1}{(N-m+2)!}=\sum_{m=2,4,\ldots,N-1}^{N-1}\frac{m}{(m+1)!}\frac{N-m+1}{(N-m+2)!}. \tag{A7}$$

TABLE IV. Size distribution of type I clusters in one dimension from expression (A9).

Cluster size	Theoretical probability
2	0.40000000
3	0.33333333
4	0.17142857
5	0.06666667
6	0.02116402
7	0.00571429
8	0.00134680
9	0.00028219
10	0.00005328

When N is *even*,

$$P_r(N)=\sum_{m=1}^{N-1}P_m^N=\sum_{m=1,3,\ldots,N-1}^{N-1}C_m\frac{N-m+1}{(N-m+2)!}-P_0^N=\sum_{m=1,3,\ldots,N-1}^{N-1}\frac{m}{(m+1)!}\frac{N-m+1}{(N-m+2)!}+\frac{N+2}{(N+3)!}. \tag{A8}$$

The absolute probability of a size N cluster is

$$P_{absolute}(N)=\frac{P_r(N)}{\sum_{N=2}^{\infty}P_r(N)}. \tag{A9}$$

The result is shown in Table IV.

[1] Fereydoon Family and Paul Meakin, Phys. Rev. Lett. **61**, 428 (1988).
[2] J.M. Yeomans, *Statistical Mechanics of Phase Transitions* (Clarendon Press, Oxford, 1992).
[3] J. Rudnick, P. Nakmahachalasint, and G. Gaspari, Phys. Rev. E **58**, 5596 (1998).
[4] M. Berberan Santos, Am. J. Phys. **54**, 1139 (1986).