On Multifractal Property of the Joint Probability Distributions and Its Application to Bayesian Network Inference

Haipeng Guo
Department of Computing and Information Sciences
Kansas State University, Manhattan, KS 66506
hpguo@cis.ksu.edu

Abstract

This paper demonstrates that the Joint Probability Distribution (JPD) of a Bayesian network is a random multinomial multifractal. With sufficient asymmetry in individual prior and conditional probability distributions, the JPD is not only highly skewed as shown by Druzdzel [3], but also is stochastically self-similar and has clusters of high-probability instantiations at all scales. Based on the discovered multifractal property a two phase hybrid Sampling-And-Search algorithm for finding the Most Probable Explanation (MPE) is developed and tested. The experimental results show that the multifractal property provides a good meta-heuristic for solving the MPE problem. The multifractal properties also strengthen the connections between Bayesian networks and thermodynamics. These connections have recently been exploited in popular Bayesian network inference algorithms based upon models from statistical physics [16, 11], such as free energy minimization.

1 INTRODUCTION

Bayesian networks (BNs) [12] provide a compact representation of the JPD of a uncertain domain by specifying the JPD into product of local prior and conditional probability distributions. The JPD over its variables can be seen as being created by a multiplicative process, combining prior and conditional probabilities of individual variables. By applying the Central Limit Theorem, Druzdzel [3] demonstrated that "... asymmetries in these individual distributions result in JPDs exhibiting orders of magnitude differences in probabilities of various states of the model ... In particular, there is usually a small fraction of states that cover a large portion of the total probability space ...." (Druzdzel 94). Druzdzel's result suggests that considering only a small number of the most probable states can lead to good approximations in belief updating. Some questions of interest are: where and how can we find these high-probability instantiations in the space of JPDs? Is there any internal structure in the JPD that can facilitate search? If so, how can we characterize it? This paper attempts to answer these questions by demonstrating that the JPD of a BN is a random multinomial multifractal created from a random multinomial multiplicative cascade. By applying multifractal analysis, we show the existence of multifractal structure within the JPD. More specifically, the JPD, as a multifractal measure, can be partitioned into fractal subsets such that each subset supports a monofractal measure, and the JPD consists of clusters of high-probability instantiations at all scales. Based on these multifractal properties, we have designed and tested a new Sampling-And-Search algorithm for finding the MPE.

2 MULTIFRACTAL ANALYSIS

2.1 FRACTALS AND MULTIFRACTALS

Fractals are extremely irregular, self-similar sets [9, 4]. A fractal is characterized by its *fractal dimension*. For example, the dimension of an irregular coastline may be greater than 1 but less than 2, indicating that it is not simply a "line" but has some space-filling characteristics in the plane. The Cantor set [9] is the oldest and simplest man-made fractal. It is constructed by removing the middle third from the unit interval, the remaining two subintervals have their middle third removed, and this continues infinitely. More formally, the Cantor set is defined as follows:

$$K = \bigcap_{n=0}^{\infty} K_n \tag{1}$$

where $K_0 = [0, 1]$ and

$$K_n = \left(\frac{K_{n-1}}{3}\right) \bigcup \left(\frac{2}{3} + \frac{K_{n-1}}{3}\right) \qquad n = 0, 1 \ldots . \quad (2)$$

The dimension of a fractal set can be calculated by counting the number of covers that are required to cover the set of interest. In the Cantor set, when $n = 0$, 1 box of length 1 is needed to cover K_0; when $n = 1$, 2 boxes of 1/3 are needed for K_1, etc. Let N_δ be the number of boxes with length δ that are required to cover K, the *(Minkowski) fractal dimension* is then defined as:

$$\frac{\log N_{\delta_n}(K)}{-\log \delta_n} = \frac{\log(2^n)}{-\log(3^{-n})} = \log_3 2 \qquad (3)$$

The main difference between a *fractal* and a *multifractal* is that the former refers to a *set* while the latter refers to a *measure*. A measure μ assigns a quantity to each member of a set (the measure's support set), thus it defines a distribution of that quantity over the support set. Multifractal analysis [10, 4, 5] is related to the study of a distribution of physical or other quantities on a geometric support set. The support may be a line, plane, or a fractal. Multifractal measures are highly irregular and self-similar (exactly or stochastically). For instance, the distribution of gold over a geographical map of the USA is very irregular. It is found in high concentrations at only a few places, in lower concentrations at many places, and in very low concentrations almost everywhere. This description holds for all scales - be it on the scale of the whole country, one state, on the scale of meters, or even at a microscopic scale. Many other quantities exhibit the same behavior, i.e., the irregularity is the same at all scales, or at least statistically [10]. We call this kind of self-similar measure a *multifractal*. The concept of multifractal was originally introduced by Mandelbrot in the discussions of turbulence [8], and later applied to many other contexts such as Diffusion Limited Aggregation (DLA) pattern [2], earthquake distribution analysis [5] , and Internet data traffic modelling [13]. A multifractal is often generated by an elementary iterative scheme called *multiplicative cascade*.

2.2 MULTIFRACTAL SPECTRUM

How can we characterize a multifractal measure? Clearly, we need more than just a fractal dimension. Simply counting the boxes as we did to the Cantor set is like counting coins without caring about the denomination. We must therefore find a description that assigns the measure in each box a weight [4]. In the following example, I use the Cantor measure [13] to illustrate the basic characterization of a multifractal. Consider the Cantor set again. Now we extend it by allocating a mass or a probability to each subinterval at each division. For example, we allocate 2/3 of the existing probability in an interval being divided to the right-hand subinterval, and 1/3 to the left-hand.

The first step of multifractal analysis is to define α, the *coarse Hölder exponent* [10, 4, 5] , as the logarithm of μ, the measure of the box, divided by the logarithm of δ , the size of the box.

$$\alpha = \frac{\log \mu(box)}{\log \delta} \qquad (4)$$

The multiplicative construction of μ makes it clear that the probability μ of a sequence of intervals will decay exponentially fast as the interval is being divided and shrinks down to a point. Thus α can be thought of as the local degree of differentiability of the measure, the rate of local probability change [5], or the strength of singularity [14]. Once α is defined, we would like to draw the frequency distribution of α as follows. For each value of α, we count the number $N_\delta(\alpha)$ of boxes having a coarse Hölder exponent equal to α. Then we define $f_\delta(\alpha)$ as the logarithm of $N_\delta(\alpha)$ divided by the logarithm of the size of the box.

$$f_\delta(\alpha) = -\frac{\log N_\delta(\alpha)}{\log \delta} \qquad (5)$$

$f_\delta(\alpha)$ can be loosely interpreted as an approximation to the Minkowski fractal dimension of the subsets of boxes of size δ having coarse Hölder exponent α. The function $f(\alpha) = \lim_{\delta \to 0} f_\delta(\alpha)$ is called the *multifractal spectrum*. It characterizes a multifractal. The graph of $f(\alpha)$, often called $f(\alpha)$ curve [10, 5], is shaped like symbol "$\bigcap$", usually leaning to one side. Usually there are bounds α_{min} and α_{max} such that $\alpha_{min} < \alpha < \alpha_{max}$. The α value of the peak is called α_0. Figure 1 plots the $f(\alpha)$ curve of the Cantor measure.

From the preceding discussion we can see that the basic idea behind multifractal analysis is to classify the singularities of the measure by strength. This strength is denoted by a singularity exponent α - the coarse Hölder exponent. Points of equal strength lie on interwoven fractal subsets. Each of these fractal subset is a monofractal with a fractal dimension $f(\alpha)$. This is one of the several reasons for the term multifractal.

2.3 THE BINOMIAL MULTIPLICATIVE CASCADE ON [0, 1]

Many multifractal measures can be generated from an elementary iterative procedure called *multiplicative cascade*. The binomial measure is the very simplest multiplicatively generated multifractal measure. Let m_0 and m_1 be two positive numbers adding up to 1. At stage 0 of the cascade, we start the construction

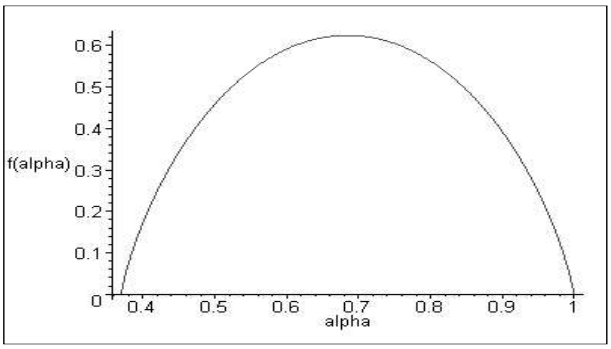

Figure 1: The $f(\alpha)$ Curve of The Cantor Measure

with the uniform measure μ_0 on [0, 1]. At step $k = 1$, the measure μ_1 uniformly spreads mass (or probability) equal to m_0 on the subinterval [0, 1/2] and mass equal to m_1 on the subinterval [1/2, 1]. At step $k = 2$, [0, 1/2] is split into two subintervals [0, 1/4] and [1/4, 1/2], which respectively receive a fraction m_0 and m_1 of the total mass μ_1 on [0, 1/2]. Applying the same procedure to [1/2, 1], we obtain:

$$\mu_2[0, 1/4] = m_0 m_0, \quad \mu_2[1/4, 1/2] = m_0 m_1 \tag{6}$$

$$\mu_2[1/2, 3/4] = m_1 m_0, \quad \mu_2[3/4, 1] = m_1 m_1 \tag{7}$$

Iteration of this procedure generates an infinite sequence of measures. At step $k + 1$, we assume the measure μ_k has been defined and μ_{k+1} is defined as follows. Consider an arbitrary interval $[t, t + 2 - k]$, where the dyadic number t is of the form:

$$t = 0.\eta_1\eta_2 \ldots \eta_k = \sum_{i=1}^{k} \eta_i 2^{-k} \tag{8}$$

in the counting base $b = 2$. We uniformly spread a fraction of m_0 and m_1 of the mass $\mu_k[t, t+2-k]$ on the subinterval $[t, t+2-k-1]$ and $[t+2-k-1, t+2-k]$. A repetition of this scheme to all subintervals determines μ_{k+1}. The measure μ_{k+1} now is well defined.

The construction of binomial multifractal can be extended in several ways. First, at each stage of the cascade, intervals can be divided not in 2 but in $b > 2$ intervals of equal size. This defines the class of *multinomial multifractals*. Second, the allocation of mass between subintervals at each step of cascade can be randomized by using a random variable as the multiplier. This defines *random multifractals*. Although the multipliers need not to be discrete, we shall use discrete ones for simplicity.

2.4 PROBABILISTIC ROOTS OF MULTIFRACTALS

Because multifractal measures can be generated by, or mapped onto, multiplicative cascade, the coarse Hölder exponent can be expressed as a sum of random variables by definition [10]. The behavior of sums of random variables is a central topic in probability theory. There are three theorems dealing with such sums: *the Law of Large Numbers (LLN)*, *the Central Limit Theorem (CLT)* and *the Large Deviation Theorem (LDT)*. The LLN says that almost surely (with probability of 1) the sample average will converge to the expectation when k increases to infinity. The LLN guarantees the existence of α_0 and its role as the most probable Hölder exponent. But the LLN only holds in the limit $\delta \to 0$, whereas we are often dealing with a finite number of multiplicative steps k. Thus, the deviation from the expected value becomes important for finite k. The relevant information is yielded by the CLT and, far more important, by the LDT. The CLT is concerned with small fluctuations around the expected value. In this context, it shows that the appearance of a quadratic maximum in the $f(\alpha)$ of the binomial measure is not a coincidence. Consider a random variable with finite expectation EX and $Pr(X > EX) > 0$. The large deviation theory is concerned with very large fluctuations around the expected value, namely the behavior of

$$\lim_{k\to\infty} Pr\{\frac{1}{K}\sum_{h=1}^{k} X_h - EX \geq \delta\} \tag{9}$$

as a function δ and k. The LLN tells us that,

$$\lim_{k\to\infty} Pr\{\frac{1}{K}\sum_{h=1}^{k} X_h - EX = 0\} = 1 \tag{10}$$

i.e., the probability converges to 0 for sure as k increases to infinity. The LDT states that it not only converges to 0, but also does so exponentially fast. In this section, we omit some details, but generally speaking, $f(\alpha)$ can be deduced via the large deviation theory and this provides a probabilistic basis for multifractals [4, 7]. Furthermore, large deviation theory in the continuous and/or unbounded cases exists as well, providing a full justification of the so-called *thermodynamic formalism of multifractals*. We refer the reader to for more details [4, 10, 5, 7].

2.5 THERMODYNAMICS FORMALISM OF MULTIFRACTALS

There are more than one way to get to the multifractal spectrum $f(\alpha)$. An alternative method is the method of Moments in which we first define partition function,

analogous to the partition function in thermodynamics and statistic physics [4, 10],

$$Z_q(\delta) = \sum_{i=1}^{N(\delta)} \mu_i^q = \sum_{i=1}^{N(\delta)} (\delta^{a_i})^q \tag{11}$$

Denote the number of boxes for which the coarse Hölder exponents satisfied $\alpha < \alpha_i < \alpha + d\alpha$ by $N_\delta(\alpha)d\alpha$. The contribution of the subset of boxes with α_i between α and $\alpha + d\alpha$ to $Z_\delta(\alpha)$ is $N_\delta(\alpha)(\delta^\alpha)^q d\alpha$. Integrating over $d\alpha$ we obtain,

$$Z_q(\delta) = \int N_\delta(\alpha)(\delta^\alpha)^q d\alpha \tag{12}$$

If $Z_\delta(\alpha) \sim \delta^{-f(\alpha)}$, it follows that

$$Z_q(\delta) = \int \delta^{q\alpha - f(\alpha)} d\alpha \tag{13}$$

Keeping only the dominant contribution in the equation, and introducing

$$\tau(q) = q\alpha(q) - f(\alpha(q)) \tag{14}$$

The partition function will scale like $Z_q(\delta) \sim \delta^{\tau(q)}$. It is easy to see that

$$\frac{d\tau(q)}{dq} = \alpha(q) \tag{15}$$

This means that $f(\alpha)$ can be computed from $\tau(q)$ and vice versa. The relation between $f(\alpha)$ and $\tau(q)$ is called a *Legendre transform* [10]. An interesting consequence is that *flexibly rich thermodynamic content hidden in the concept of multifractals*. From preceding discussion, we can easily draw a correspondence between $Z(q)$ and the thermodynamic partition function $Z(\beta)$, between q and the temperature T (as the inverse of T), between α and the energy, between $f(\alpha)$ and the entropy, and between $\tau(q) = q\alpha - f(\alpha)$ and the *Gibbs free energy* $G = H - TS$. For more information on this topic, the interested reader is referred to [10].

3 BAYESIAN NETWORKS AS RANDOM MULTINOMIAL MULTIFRACTALS

A Bayesian network [12] is a Directed Acyclic Graph (DAG) in which nodes represent random variables and arcs represent conditional dependence relationships among these variables. Each node X_i has a conditional probability table (CPT) that contains probabilities of a variable value given the values of its parent nodes, denoted as $\pi(X_i)$. A BN represents the exponentially sized JPD in a compact manner. Every entry (an instantiation of all nodes) in the JPD can be computed from the information in the BN by chain rule:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\pi(x_i)) \tag{16}$$

From the multifractal viewpoint, the JPD defined by a BN with n nodes is a measure of belief distributed on an n-dimension space of random events. Given a topological ordering of all nodes, we can map the n-dimension space to a linear interval by assigning each event an integer number as its address on that interval. For example, the linear interval for an 8-node binary BN is $[0, 255]$. The JPD of a BN can be considered as being generated from a multiplicative cascade in which number of steps n equals to the number of nodes. At each step of the cascade, intervals are divided into b subintervals where b is the number of states of the current node, and the multiplier for allocating the probability is a random variable defined by the CPT of the current node. It is easy to see that in the most general case a BN corresponds to a multifractal generated by a random multinomial multiplicative cascade. The simplest multifractal - the binomial measure - corresponds to the simplest BN - a binary BN without links. Consider such an 8-node binary BN in which each node has a prior probability distribution of $(0.25, 0.75)$. The cascade contains 8 steps and generates a JPD of 256 instantiations. This is actually the simplest multifractal - the binomial measure.

Now let us consider the process of an agent's incremental understanding of some uncertain domain as a multiplicative cascade process. At the beginning the agent first identifies all random variables. Before it knows anything about the causal relationships between these variables, it has to assume a uniform distribution spreading belief evenly to all states. The agent's belief is redistributed as it learns more about the domain, i.e., the connections between nodes and the CPT values. The process of belief redistribution is a typical multiplicative cascade process similar to any other multiplicative cascade in the context of multifractals. For example, a turbulence cascade model describes the nature of energy dissipation in a turbulent fluid flow. In turbulence the energy is introduced into the system on a large scale (storms, or stirring a bowl of water), but can only be dissipated in the form of heat on very small scales where the effect of velocity, or friction between particles, becomes important. Cascade models assume that energy is dissipated through a sequence of eddies of decreasing size, until it reaches sufficiently small eddies where the energy is dissipated as heat. In the case of Bayesian networks, the belief is introduced to the domain from a high level as a uniform distribution. As we learn the CPTs, we capture the increas-

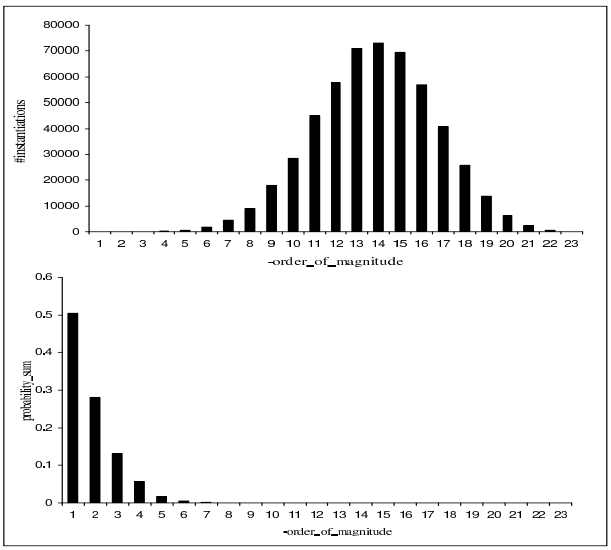

Figure 2: (1) Number of Instantiations At Each Order (2) Probability Sum of Instantiations At Each Order

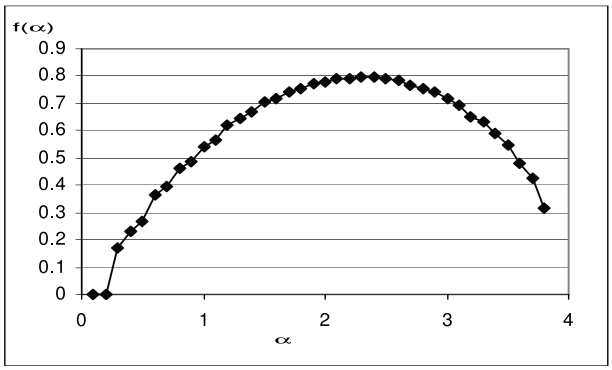

Figure 3: The f(α) Curve of ALARM13's JPD

ingly refined causal structure of the domain. These substructures keep redistributing our belief until we learn all about the domain.

4 CASE STUDY: THE ALARM13 NETWORK

4.1 THE JOINT DISTRIBUTION

In this section we analyze the JPD of ALARM13 [3], a subset of ALARM network, to demonstrate its multifractal characteristic and clustering property. ALARM13 was the same network analyzed in [3]. It contains 13 variables, resulting in 525,312 non-zero states. The probabilities of these states were spread over 23 orders of magnitude. Figure 2 shows the histograms of the number of instantiations distributed at each order and their contribution to the total probability space. The X-axis is the negative order of magnitude in both figures. Figure 2.1 shows that the histogram of number of instantiations at each order appears to be a normal distribution. Given the logarithmic scale of the X-axis, it shows that the actual distribution is a lognormal. The peak of figure 2.1 is at the order of 10^{-14}. It contains 73,256 instantiations, but its contribution to the total probability space is only $2.9E-09$. From Figure 2.2 we can see that high-probability instantiations, although few in number, dominate the joint probability space. Of all instantiations, there are one with probability around 0.505, 10 with probabilities between $[0.1, 0.01]$ and the total probability of 0.28, 48 with probabilities between [0.01, 0.001] and the total probability of 0.13, 208 with probabilities between [0.001, 0.0001] and the total probability of 0.058. The 267 most likely instantiations (0.05% of the total of 525,312) covers 97.45% of the total probability space. This highly skewed result has been analyzed by Druzdzel [3]. In the following we show the multifractal structure of the JPD and the way instantiations at different orders of magnitudes fill the space.

4.2 THE MULTIFRACTAL SPECTRUM

[t] Applying multifractal analysis to ALARM13's JPD, we get its multifractal spectrum, the $f(\alpha)$ curve, as shown in Figure 3. The X-axis is the coarse Hölder exponent α, Y-axis is $f(\alpha)$ - the fractal dimension of the subset of all instantiations with the same α. This $f(\alpha)$ curve confirms that the JPD of a Bayesian network is a multifractal. It describes how these instantiations fill the probability space from the point of view of fractal dimension. We can see in Figure 3 that high-probability instantiations (corresponding to small α) have a low dimension, which means that they fill the probability space in a very "sparse" way, i.e., there are clusters of high-probability instantiations. The peak of Figure 3 has a fractal dimension of 0.79, and the corresponding coarse Hölder exponent α is around 2.5. By the definition of α, this corresponds to instantiations with probability on the order of 10^{-15}. They are actually instantiations around the peak of Figure 2.1. It means that these instantiations fill the probability space in a very "dense" way, i.e., they are almost all over the space. Finally, instantiations with very low probabilities ($\alpha = 3.8$) also have low dimensions ($f(\alpha) = 0.32$). Again, it means that low-probability instantiations (rare events) distribute sparsely as well, and clusters of them can be expected. This yields a

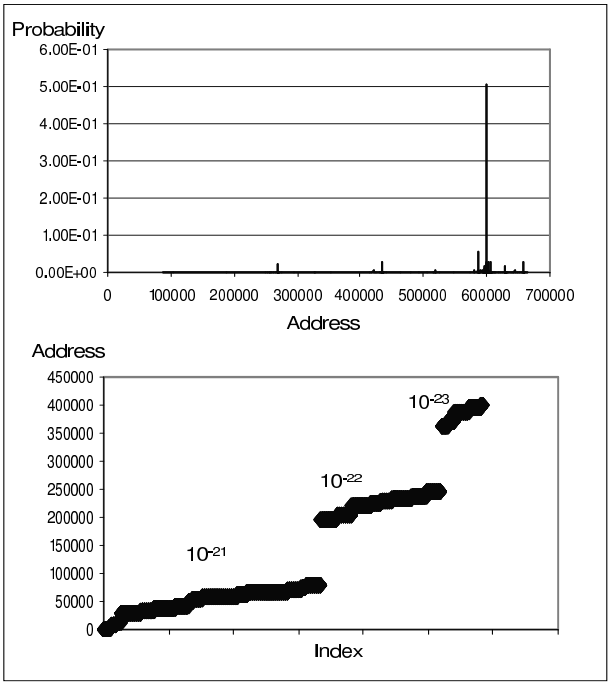

Figure 4: (1) The Clusters of High Probability Instantiations (2) The Clusters of Low Probability Instantiations

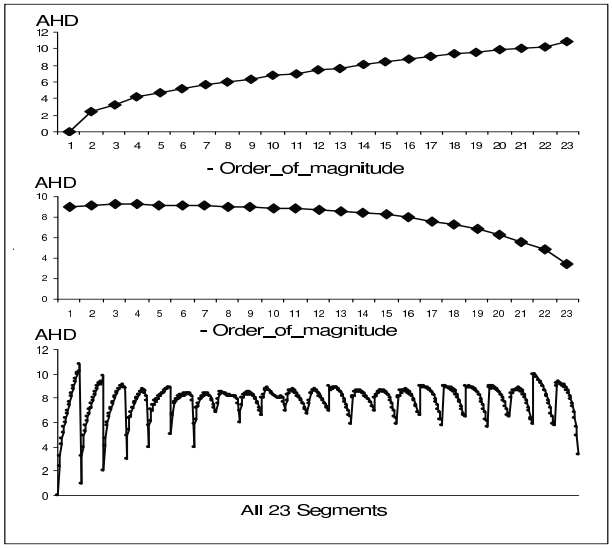

Figure 5: The Aeverage Hamming Distance Graphs

mathematical description of the inner structure of the JPD: *there are clusters of high-probability instantiations and low-probability instantiations, but instantiations in the middle are distributed almost all over.* Interestingly, this pattern coincides with the way that people live in the real world, i.e., high-income people tend to live in the same community, so do low-income people, but the middle-class are located all over.

4.3 QUANTIFYING THE CLUSTERING PROPERTY

To show the clustering property more clearly, we draw the distribution of high-probability instantiations and the distribution of low-probability instantiations in Figure 4. Figure 4.1 contains all instantiations with a probability higher than 0.0001 in which X-axis is the "address" of instantiations ranging from 0 to 525,312 and Y-axis is the actual probability value. Figure 4.2 contains all instantiations lower than 10^{-20} in which Y-axis is the "address" of instantiations and X-axis is just the series number of each instantiation (Note we use a different X-Y here because the actual values are too small to be drawn neatly). We can see clearly there are clusters in both graphs.

Having shown the clustering property, the next thing we want to do is to quantify this property. We use the *Hamming Distance* between bit string representations of two instantiations to measure how far they are located from each other. An instantiation is represented as a bit string "$b_1 b_2 \dots b_n$" where n is the number of variables in the domain and b_i is the state index of each variable. For example, the Hamming distance between instantiation "00001100" and "00100001" is 4. Because of the clustering property, high-probability instantiations should have small Hamming distances between each other to be a cluster. We draw the *Averaging Hamming distance (AHD)* graph for the most likely instantiation in Figure 5.1. The X-axis is the negative of the order of magnitudes; The Y-axis is the AHD between the most likely instantiation and all instantiations at each order of magnitude. From Figure 5.1 we can see that the instantiations with lower probabilities have a larger Hamming distance from the most likely instantiation, i.e., they locate far away from the most likely instantiation. We also draw the same figure for the lowest instantiations in Figure 5.2. In Figure 5.3 we put together 23 AHD graphs for instantiations of all orders to provide a global picture. Figure 5.3 consists of 23 segments of curves corresponding to 23 orders of magnitudes. Each curve consists of 23 points, represents the AHD graph of a randomly picked instantiation at that order. For example, the first segment in Figure 5.3 is Figure 5.1, and the last segment is Figure 5.2. From Figure 5.3 we can see that the instantiations in the middle order of magnitudes are located at almost the same distance

Input: A BN (G, P) and an evidence set E.
Output: A complete assignment $u = (u_1, \ldots, u_n)$.

Step 1: Use sampling algorithm to generate a set of initial good points S.
Step 2: For each point in S, start a hill climbing using **Neighborhood Quality** as the evaluation function, put all local optimums into S^*.
Step 3: For each point in S^*, start a normal hill climbing, return the best solution so far as the MPE.

Figure 6: Two Phase Sampling-And-Search Algorithm for Finding The MPE Using Multifractal Heuristic

from all other orders ($7 < AHD < 9$), i.e., they can be found at almost all places. This finding supports our previous analysis of the expected distribution pattern.

5 A MULTIFRACTAL SEARCH ALGORITHM FOR FINDING THE MPE

The JPD's multifractal property can be used as a *meta-heuristic* to develop new search algorithm for finding the MPE. Since the search space is a multifractal, good solutions would cluster together, so do bad solutions. Hence the search should be divided into two phases: first identify the *"good communities"*; then localize the search to these regions. Also, the *quality of the community* rather than the current search point alone should be evaluated and compared to guide the search. If point A is better than point B but B's neighbors are better than A's, then we should move B to look for the global optimal. This helps the searcher escape the local optimal where simple hill climber gets stuck.

Based on these meta-heuristics we have developed a two phase Sampling-and-Search algorithm to solve the MPE problem, which is to find the most probable explanation (a complete assignment) given the observed evidence. The general MPE problem is NP-complete [15] and even hard to approximate [1]. In the first phase of the algorithm, forward sampling (or any other feasible methods) is used to identify a set of good communities quickly. In the second phase, a hill climbing using *Neighborhood Quality* as the evaluation function is started for each community from the previous phase. An additional "repair" phase can be added by using a set of "elite solutions" and a set of "worst solutions" collected during the search process to refine the final solutions by flipping the variable values that do not agree with the majority "elite solutions" or these that agree with most "worst solutions". When the stop rule is satisfied, it returns the best solution so far as the MPE. The algorithm's performance is determined by two factors: the reliability of the sampling algorithm to bring the searcher to places not far away from the global optimal, and the robustness of the Neighborhood Quality as an evaluation function to bring the searcher from a near optimal place to the global optimal. The Neighborhood Quality of a search point is defined as the sum of the likelihoods of all its nearest neighbors. It can be approximated by randomly drawing samples from its neighbors. The sampling radius can be set to a small positive value k.

We expected the skewness of the CPTs would have an influence on the performance of the algorithm. So in our experiments we randomly generated three groups of networks with different CPT skewness to test the algorithm: skewed, normal, and unskewed. The skewness of the CPTs is computed as follows [6]. For a vector (a column of the CPT table), $v = (v_1, v_2, \ldots, v_m)$, of conditional probabilities,

$$skew(v) = \frac{\sum_{i=1}^{m} |\frac{1}{m} - v_i|}{1 - \frac{1}{m} + \sum_{i=2}^{m} \frac{1}{m}} \qquad (17)$$

The skewness for the CPT of a node is the average of the skewness of all columns. And the skewness of the network is the average of the skewness of all nodes. The skewness of these three groups of networks were set to around 0.9, 0.5, and 0.1 respectively. Each group consists of 20 networks with binary nodes. The number of nodes were 100, and the number of edges were $120 \sim 150$. These networks were set to be sparse enough so that the exact MPEs can be computed. For each network, we randomly generated 10 evidence values hence the size of search space is 2^{90}. We used Hugin to compute the exact MPEs. For each group of networks, we counted the number of times when exact MPEs were found. We also computed the average relative error (ratio of the absolute error to the exact MPE value) and recorded the average Hamming distance between the returned MPE and the exact MPE. Table 1 summarizes the experimental results. From the results we can see that normally-skewed networks are the easiest ones for the algorithm and unskewed networks are the hardest ones. Of 20 normally-skewed networks we were able to find exact MPE in 19 of them and even the one missed is very close to the global optimal (only 2 bits difference out of 100). Of 20 unskewed networks we were able to find exact MPE in only 4 of them. The average error and average Hamming distance between the returned MPE and the exact MPE are also the largest ones. This results imply that if the network is unskewed (most distributions are nearly uniform), finding MPE will be hard because the search space is

Table 1: Results on Randomly Generated Networks

	#solved	error	AHD to exact
skewed	12/20	0.0242	1.25(25/20)
normal	19/20	0.0029	0.1(2/20)
unskewed	4/20	0.0798	3.9(78/20)

flat. In the other hand, if it is highly skewed it will also bring trouble to the search algorithm because of the attractiveness of these steep local optimums.

6 CONCLUSION

We have demonstrated that the underlying JPDs of Bayesian networks are multifractals created by random multiplicative cascade processes. The JPD with many orders of magnitude differences in probabilities of various instantiations is not only highly skewed, but also stochastically self-similar and exhibits clustering properties. The multifractal spectrum of the JPD describes how instantiations at different orders fill the joint distribution space with different fractal dimensions. In particular, both high and low probability instantiations tend to form clusters in the joint distribution space. Even though we discussed the model as a whole, the result will hold for its self-contained parts as well. we also hypothesize that it holds for dynamic models. The $f(\alpha)$ curve will show up as long as a random multiplicative cascade process is involved. The significance of this analysis is that it provides important information about characteristics of the joint probability distribution. Particularly, the clustering property can be a very useful meta-heuristic for searching the MPE. This research also bridges an analytical gap between multifractal and BNs and suggests some very interesting research directions. As we have seen, multifractals have a deep probabilistic root and a rich thermodynamic content. The fact of BN being a multifractal draws our attention to connections between thermodynamics and BNs [16, 11]. Also, because the MPE problem is NP-Complete, we should also expect to observe the same multifractal structure in the solution space of other hard combinational problems such as MAXSAT and TSP. Applying the multifractal meta-heuristic to solve these problems would be a very interesting topic to investigate in the future.

Acknowledgements

Thanks very much to Marek J. Druzdzel for providing us the ALARM13 dataset.

References

[1] A. M. Abdelbar and S. M. Hedetniemi. Approximating maps for belief networks in NP-hard and other theorems. *AI*, 102:21–38, 1998.

[2] A. Bunde and S. Havlin. *Fractals and Disordered Systems*. Springer, 1991.

[3] M. J. Druzdzel. Some properties of JPDs. In *UAI94*, pages 187–194, 1994.

[4] C. J. G. Evertsz and B. B. Mandelbrot. Multifractal measures. In H. Peitgen et al., editor, *Chaos and Fractals: New Frontiers of Science*, pages 921–953. Springer-Verlag, 1992.

[5] D. Harte. *Multifractals: Theory and Applications*. Chapman and Hall CRC, 2001.

[6] N. Jitnah and A. E. Nicholson. Belief network algorithms: A study of performance based on domain characterization. In *PRICAI Workshops*, pages 168–187, 1996.

[7] M. Kessebhmer. Large deviation for weak gibbs measures and multifractal spectra. *Nonlinearity*, 2001.

[8] B. B. Mandelbrot. Possible refinement of the lognormal hypothesis concerning the distribution of energy dissipation in intermittent turbulence. In *Statistical Models and Turbulence*, pages 331–351. Springer, NY, 1972.

[9] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman and Co., NY, 1982.

[10] B. B. Mandelbrot. Multifractal measures, especially for geophysicists. *Pageopg*, 131(133), 1989.

[11] P. Pakzad and V. Anantharam. Belief propagation and statistical physics. In *Conference on Information Science and Systems*, 2002.

[12] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, San Mateo, CA, 1988.

[13] R. Riedi and J. L. Vehel. Multifractal properties of tcp traffic: A numerical study. Technical report, Rice University, 1997.

[14] R. H. Riedi. An introduction to multifractals. Technical report, Rice University, 1999.

[15] S. E. Shimony. Finding maps for belief networks is NP-hard. *AI*, 68:399–410, 1994.

[16] J. Yedidia, W. Freeman, and Y. Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. Technical Report 2001-16, MERL, 2000.