Chapter 2. The J Distribution and its Variations

In Chapter 1, the J distribution appeared only in the form of its probability density function (pdf). In this chapter it appears as a distribution populated by a number R of species. The distribution comes in a continuous version and in a corresponding discrete version that translates the continuous form into a prediction of the number of species to appear at each possible abundance. Carrying this process one step further, an equivalent formulation called the canonical sequence specifies average positions for each species as a function of rank order.

In this chapter and the ones that follow, I will frequently use either form of the parameter delta, depending on the context. Readers must merely keep in mind that $\Delta = 1/\delta$ by definition.

2.1 Properties of the probability density function

Recall the density function in Equation (i) of Section 1.1:

$$f(x) = c(1/(x+\varepsilon) - \delta); \quad 0 \le x \le \Delta - \varepsilon,$$

$$= 0; \quad x \ge \Delta - \varepsilon$$

It may cause some consternation to have a density function with a finite domain. Can there be no abundances greater than $\Delta - \varepsilon$? This question will be answered in Section 2.2 below.

The normalizing constant c is a function of the two parameters, ε and Δ, and does not constitute a new parameter, being merely a notational simplification:

$$c = (\ln((\Delta/\varepsilon) - 1 + \varepsilon/\Delta)^{-1}$$

The constant c ensures that the integral of f encloses a unit area , as must all density functions. In what follows, the simplified (closely approximate) form of the formula is adequate to the task of playing the role of a normalizing constant.

$$c \approx (\ln((\Delta/\varepsilon) - 1)^{-1} \qquad \text{(i)}$$

In most cases the approximation error would be less than 0.0001.

2.1.1. The capacity constant

The constant c yields a closely related constant C, the *capacity* of the community, defined as the inverse of c. Inverting formula (i) results in the following expression:

$$C = \ln(\Delta/\varepsilon) - 1$$

Clearly, C is inversely proportional to the (log of) the twin parameters. It gets larger when the parameter Δ gets larger or the parameter ε smaller. The capacity C is smaller in the opposite situation. The difference between the two possibilities is illustrated in Figure 2.1a.

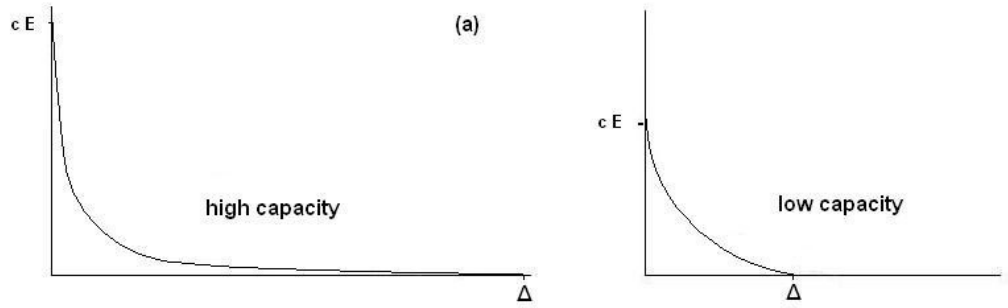

Figure 2.1a. Examples of high (left) and low (right) capacities

If, on the other hand, the capacity is kept constant, the parameters ε and δ may nevertheless vary, as long as the product Δ/ε remains constant, as in Figure 2.1b.

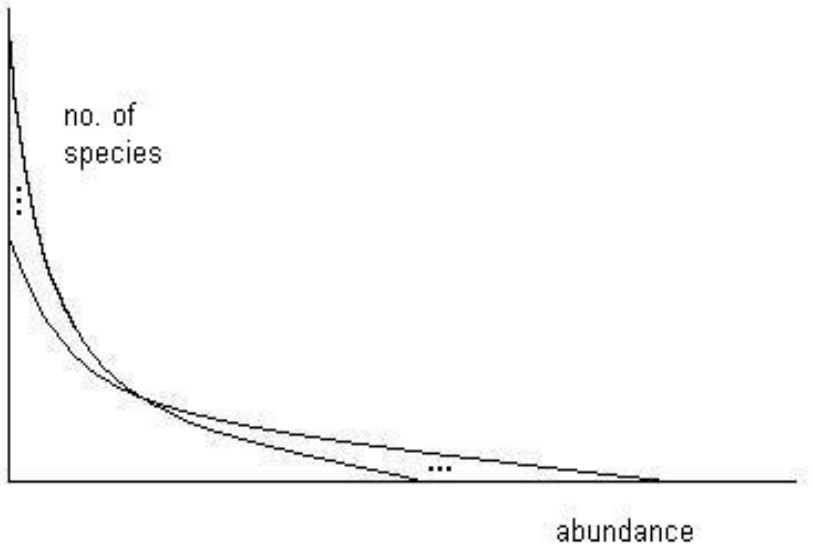

Figure 2.1b. Different shapes may share the same c-value

For each fixed value of c, the range of density functions produced by varying ε and Δ in this manner all belong to the same capacity class. The concept of capacity will be revisited in Section 2.2.2, where it plays a role in the spacing between consecutive average abundances; the greater the capacity C, the greater is the gap between the abundance of the kth species and that of the k+1st. In distributions having a larger capacity, there is more "room", abundance-wise, for additional species. Apart from appearing in Section 2.2.2, the concept plays a relatively minor role in the rest of this book. However, it strikes one as having promise for an important part to play in later developments of the theory.

2.1.2 The mean and variance of the J distribution

According to the standard definition of the mean as the first moment of an (arbitrary) probability density function g, we have,

$$\mu = c\int_{0}^{\Delta-\varepsilon} (x/(x+\varepsilon) - \delta x)dx$$

The integral is evaluated in Appendix A.2.1 using the method of substitution.

$$\mu = c(\Delta/2 - \varepsilon\ln(\Delta/\varepsilon) + \varepsilon^2/2\Delta) \qquad \text{(ii)}$$

Expanding the constant c in terms of Δ and ε does not result in a simpler expression, so it will be left in this form. In order to find the number N of individuals represented by a given theoretical distribution, simply multiply the value for μ just obtained by R, the richness of the sample.

The variance of the J distribution is based on the second moment about the mean,

$$V = c\int_{0}^{\Delta-\varepsilon} ((x-\mu)^2/(x+\varepsilon) - \delta x)dx$$

The resulting formula, as derived in Appendix A.2, is not especially elegant:

$$V = c[(\Delta^2/2 - 2\Delta\varepsilon + (2.5)\varepsilon^2 - \delta(\Delta - \varepsilon)^3/3] + \varepsilon^2 - \mu^2$$

The variance of a distribution function F(x) = Rf(x) is obtained by dividing this expression through by R-1. Owing to its non-central nature and long tail, the J distribution has high variance. This is true of all previous proposals for species-abundance distributions, as well.

2.2. The J-distribution: species and individuals

The distribution F of the J density function f is obtained by multiplying it by R, the number of species in a sample -- or in a community, depending on the context.

$$F(x) = Rf(x)$$

Because the domain of the function is finite, one must interpret F carefully. In general, the integral of F over a given interval of abundances yields the expected number of species having abundances in that interval. Statistical fluctuations guarantee that some species in both samples and communities will often not be found in their canonical positions. (See Section 2.2.2) This includes the last canonical position, the logistic limit Δ, which bounds the average maximum abundance. Operationally, as in curve-fitting procedures, abundances beyond Δ are treated as if the J distribution had value 0 there.

The relationship between the value of the parameter ε and the number F_1 of species in the minimum abundance category, as expressed in the continuous versions of the J distribution, hints at the somewhat more subtle nature of the parameter ε:

$$F_1 = Rc(1/\varepsilon - \delta) \qquad \text{or} \qquad Rc(E - \delta)$$

Of course, c is already a function of epsilon. If the right hand side of this expression is expanded in terms of the constant c, a mixed log/linear equation results that has no closed-form solution. The equation can be solved by numerical methods on a computer using the largest abundance as an estimate for Δ (or $\delta = 1/\Delta$) and the number of lowest abundance species as an estimate for F_1. This approach will yield a good approximation, in general, for the corresponding value of ε.

2.2.1 The discrete J distribution

The discrete version of the J distribution is the main tool in fitting the J distribution to field data. Only when one has estimates in hand for the expected number of species in each category of abundance can goodness-of-fit tests be applied to a field data histogram. Early in the development of the J distribution, the discrete version played an important role by making the need for a second parameter (ε) obvious, as explained in Appendix A, Section A.1.5. The form of that parameter was dictated by the mathematics of the situation, ending in a symmetrical form wherein both parameters amounted to translations or displacements of the standard hyperbola.

To produce a discrete version of the J distribution one first selects appropriate abundance categories, depending on the data at hand. Most commonly, the field worker will use categories 1, 2, 3, etc., representing simple counts: how many species appeared once in the sample? How many twice? And so on. But observations may also be grouped, so that, for example, the categories might be 1 to 3 for the first category, 4 to 6 for the next, and so on. Another common abundance format is density. What is the average number of Robber Flies observed per hectare? One species might have a density of 3.7,.while another has density 42.6. Densities are normally grouped. Thus the first (lowest) abundance category might run from 0 to 1.5, the next from 1.5 to 3.0, etc. The width of each abundance category will be called the *span* and written either as an integer or decimal number, depending on the context. Examples of such groupings will be found on the author's website (Dewdney, 2011)

In a general approach the span will be a real number, a, implying the half-open intervals (0, a], (a, 2a], (2a, 3a], and so on. With this notation one may produce a discrete distribution function by integrating over each abundance interval (category), as follows:

$$F(ka) = \int_{(k-1)a}^{ka} F(x)\,dx = Rc[\ln((ka+\varepsilon)/((k-1)a+\varepsilon)) - \delta a];\ k = 1, 2, 3, \ldots \qquad \text{(iii)}$$

Given values for R, c, k, and a, the function to the right is readily computed using a hand calculator, although a purpose-written program to achieve the same end is considerably more convenient. (See the program HGen in Appendix A.10.) When this is done for each category, a histogram like the one in Figure

2.2 will appear. The first category (k =1) has the greatest number of species. The numbers decline as k increases, abruptly at first, more gradually later, to zero. In this idealized situation the number of species in a given category might well be fractional. The number represents the expected or average number of species in this category over all instances of samples which this particular theoretical curve fit best. Chapter 4 explains how the discrete form is applied to field data.

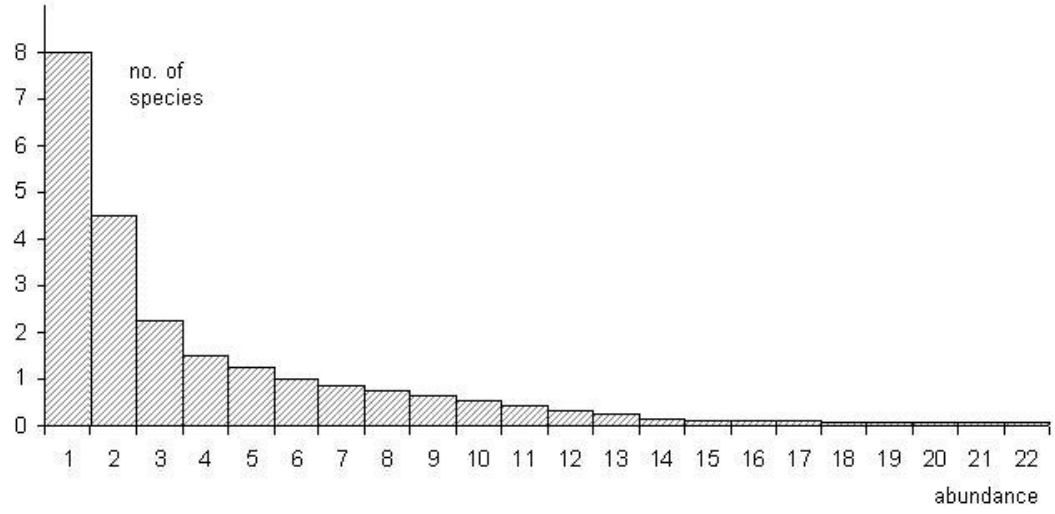

Figure 2.2. A discrete version of the J-distribution.

A second form of discrete distribution may be derived independently by replacing the continuous variable x by the discrete variable k. The formula that emerges has almost the same values as yielded by equation (iii):

$$F(k) = Rc(1/(k+\varepsilon) - \delta),$$

2.2.2. The canonical sequence

Figure 2.3 shows a set of idealized positions for abundances in a theoretical distribution (thin, flat super-imposed curve) that represents a hypothetical community. Each species is represented by a bar of unit height at a position given by formula (v), about to be derived. The position of each bar approximates

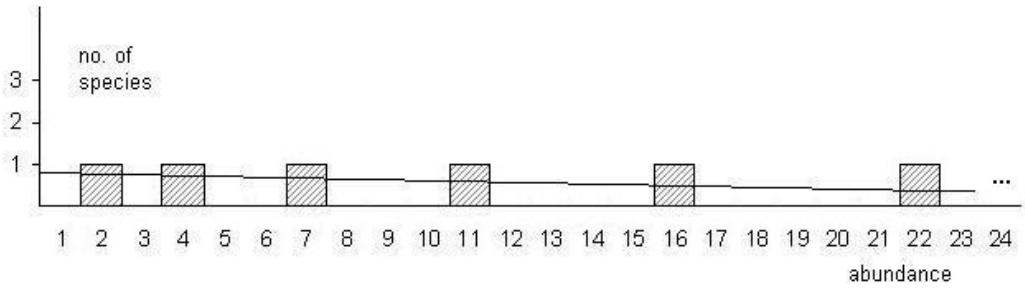

Figure 2.3. An idealized J-distribution for a community

the actual expected positions, as shown in the plot of Figure 2.4.

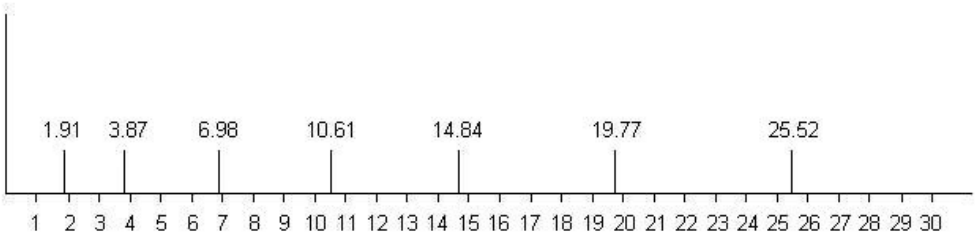

Figure 2.4. Canonical abundances of species in a small community

Those who worry that the Hyperbolic shape in samples indicates a large number of species about to be extirpated may take some comfort from these figures. Not only are species that appear only once in a sample likely to be more abundant in the source community, but the lowest abundance there is not necessarily 1, as either figure makes clear. Over time, species come and go from communities, now appearing by immigration or, now disappearing by emigration or extirpation. By the same token, at the scale of larger landscapes, species may appear by speciation or disappear by extinction. Chapter 9 examines the role of the J distribution in evolution.

The discrete version of the J-distribution given in Section 2.2.1 describes the expected numbers of species per abundance category. Another description of the distribution lists the expected abundances themselves. The canonical abundance a_k of the kth species (in abundance rank) is readily found from the integral equation,

$$Rc \int_0^{a_k} (1/(x+\varepsilon) - \delta)dx = k \qquad \text{(iv)}$$

Appendix A.2 contains the full derivation of the resulting formula,

$$a_k = \varepsilon(\exp(kT + \delta a_k) - 1), \qquad \text{(v)}$$

where a_k represents the abundance of the kth species and $T = C/R$. Recall that C is the capacity of the distribution, defined as the inverse of the constant c. The maximum abundance is given by a_R. Because formula (v) is not directly solvable, resort my be had to the following approximation which applies most accurately to lower abundances:

$$a_k = \varepsilon(\exp(kT) - 1) \qquad \text{(vi)}$$

Because of it's simplicity and ease of application, formula (vi) may be preferred for some applications.

Worked example:

Let J(2.00, 150.0) X 50 be the J distribution corresponding to a sample. Then c = 0.3003 and T = 1/15.015 = 0.06660, so that

$$a_k = 2.0(e^{0.0666k} - 1.0)$$

Table 2.1 displays the first five canonical abundances corresponding to this distribution, as well as the last four. As the table shows, more than one species can have an expected abundance that is less than one. At the discrete distribution level, the number of such species merely reflects the expected number of species to

k	1	2	3	4	5	. .	47	48	49	50
abund	0.138	0.285	0.442	0.611	0.790		63.4	72.0	83.4	99.8

Table 2.1. Canonical abundances for a sample of 50 species

show up once in the sample. In this case, the number is 6. As it happens, some 4 species would be expected to show up with abundance 2 in the sample, and so on. At some point we must switch from formula (vi) to the exact or "general" expression (v). The last four canonical positions were calculated by this method.

Finally, it should be noted that the canonical sequence formulation is simply another way of presenting the J distribution, being mathematically equivalent to it. The approximate formula used here is essentially a geometric progression, similar to the proposal of (May 1975) discussed earlier in Section 1.2. I must also insert the following caveat: defining the canonical abundance of the kth species (in order) as occurring when the integral of Equation (iv) has the value k may not be quite correct. The resulting mean abundance of the kth species may be slightly lower than the value given by formula (iv) but subsequent developments based on the integral values k chosen in Equation (iii) will not be altered enough to invalidate any of the principal conclusions arrived at in this chapter.

The canonical sequence for a given J distribution, J[ε, Δ] x R illuminates the use of the word "capacity" for the constant C. Recall that

$$C = \ln(\Delta/\varepsilon) - 1$$

According to formula (v) for the canonical abundances, we may replace the factor T by 1/Rc to obtain the formula,

$$a_k = \varepsilon\ (\exp(kC/R) - 1.0),$$

which readily yields an expression for the space between consecutive abundances, as follows.

$$a_{k+1} - a_k = \varepsilon(\exp((k+1)C/R) - 1.0) - \varepsilon(\exp(kC/R) - 1.0)$$

$$= \varepsilon(\exp(kC/R))(\exp(C/R) - 1)$$

With ε and R both fixed, the expression is obviously larger when C is larger and, conversely, smaller when C is smaller. In other words, there is more "room" between successive abundances in high capacity communities. This concept should not be confused with the capacity of the community's environment to support its many individuals.

2.2.3. An implicit formula for rank abundance

Much of the prior work on species abundances has been framed in the context of the rank abundance diagram, a method of representing sample data that holds the same information as the standard representation, but in a rather different manner. (See Section 4.2.) One arranges all the species in one's sample in order of decreasing abundance (the rank order), then plots their logarithms as vertical bars. This section notes, in passing, that the rank abundance formula is implicit in the expression from which the canonical sequence is derived in Appendix A.2. Unfortunately, the formula has no closed-form solution, so must be solved by numerical methods:

$$\ln((a + \varepsilon)/\varepsilon) - \delta a = kT,$$

If one represents the implicit solution by a(k), the order of the variable k must be reversed since, in the present instance, the diagram begins not with the smallest abundances, but the largest. If A(k) represents the resulting function, then the following formula enables one to calculate bar heights for each abundance:

$$A(k) = a(k)(R - k + 1).$$

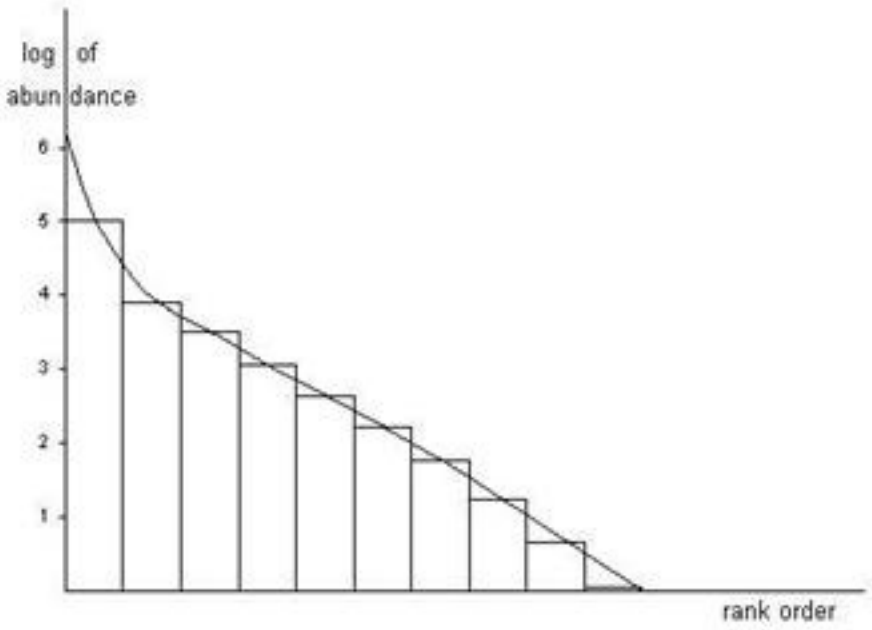

Figure 2.5. A simplified example of a Hyperbolic rank abundance diagram

The resulting curve can at least be illustrated through a worked example, as shown in Figure 2.5. Using the earlier example of J[2.0, 150.0] x 50, formula (vi) results in the plot where, for the sake of simplicity only multiples of 5 are plotted as values for k. Although it is based on a rather small example, the resulting rank abundance diagram has the sinusoidal shape that is typical of such diagrams when based on actual samples. It is not clear what role the rank abundance diagram has to play in hyperbolic theory, but this example serves to illustrate how more than one abundance distribution can produce a rank abundance distribution with a sigmoidal shape.

2.2.4 Effect of a log transformation on the J distribution

A variant of the J distribution that I shall likely never use is included here partly for the sake of completeness and partly because it results in what is known generally in the literature of population biology as the lognormal distribution. (Magurran 1988) When a field sample of the kind being studied in this monograph is subjected to a logarithmic transformation, it frequently resembles a truncated normal distribution. In such a transformation, abundance categories are grouped, with abundance 1 enjoying a category of its own, abundances 2 and 3 occupying the next category, abundances 4 to 7 occupying the next category, and so on, with a doubling at each step. Preston (1948) may have adopted this representation of abundances in order to avoid the off-page outliers that are inevitable with the standard representation. A heavy price is paid for such convenience in the form of the information that is lost when 2^k abundances are added together to produce a bar height for the kth category under this representation.

However, it may be asked what happens if one subjects the theoretical form of the J distribution to a logarithmic transformation of the abundance axis, as used in connection with the lognormal distribution. (See Section 4.2) The answer came as a surprise to me. (Dewdney 1998) A truncated, unimodal curve emerges that in many cases will appear bell-shaped, albeit seemingly truncated on the left. To put the answer on a solid footing, the principal tool will be the general integral of the J distribution over the arbitrary interval [a, b]: Here, R' will represent the area under the curve which yields the number of species having abundances in this interval.

$$R' = Rc\int_a^b (1/(x+\varepsilon) - \delta)$$

$$= Rc(\ln((b+\varepsilon)/(c+\varepsilon)) - \delta(b-a))$$

Values for the integral may be calculated for the subintervals (0, 1], (1, 3], (3, 7], and so on, according to the scheme of Preston, but this time applied to a continuous function. It would yield essentially the same result if I used a discrete version of the distribution; the integral is simply a lot less work, with only one calculation per octave. Figure 2.6 displays the result when the log transformation is applied to the distribution J[2.0, 20.0] x 50.

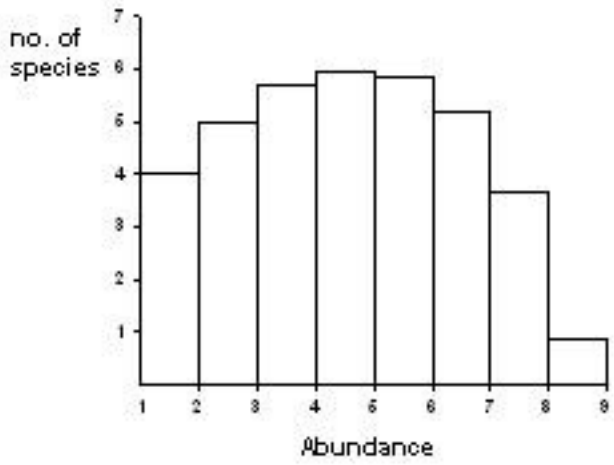

Figure 2.6. log transform applied to abundance axis of the J-distribution

According to Hyperbolic theory, samples subjected to this treatment would resemble a perturbed version of the histogram of Figure 2.6. The species that go missing from a sample are "veiled" by a sigmoidal curve when viewed on standard axes (Dewdney 1998) and not by the veil "line" of Preston (1948). Researchers, such as Hubbell (2001), who use this representation run the risk of mistaking the log-transformed J distribution for the lognormal. Indeed, Gaston (2005) has pointed out that the shape that results from logarithmic axis compression cannot be Gaussian (i.e., normal). The example in Figure 2.6 illustrates the dangers inherent in a treatment that destroys information.