Generating Project Network Simulation Activity Time Distributions Subject to Nonproject Delays

William J. Cosgrove*
California State Polytechnic University, Pomona, California, USA

The purpose of this study is to improve estimates for activity time distributions used in stochastic project network models for situations that involve nonproject or nonroutine delays. Examples of different types of nonproject delays include the weather, late deliveries, equipment breakdowns, etc., that estimators may include in estimates of activity time distributions along with project task delays. However, such estimates are univariate in that they do not distinguish between different types of delays. This study proposes a bivariate approach requiring separate estimates of distributions for project task delays and for one type of nonproject delay, leading to a marginal distribution as the resulting activity time distribution which captures both types of delays. A technique is proposed to determine if differences in entropy between univariate and bivariate approaches are significant. The *Principle of Maximum Entropy* is employed as a criterion to determine the best estimator. An example is included to illustrate the concepts developed in this study.

* Corresponding Author. E-mail address: wcosgrove@csupomona.edu

I. INTRODUCTION

Project network simulation attempts to capture the uncertainty of real world projects by structuring project activities within a stochastic network framework to better manage projects. The discrete entropy function as developed by (Shannon, 1948) provides a well established measure of uncertainty from information and communication theory, and is finding its way to more applications in mainstream statistics. This study employs both univariate and bivariate entropy within a statistical methodology which includes both project task delays and a nonproject delay for generating activity time distributions.

Project task delays are tasks *directly associated with doing the project*. Nonproject delays are *not directly associated with completing the project* but have likelihoods that can significantly add to delays in completing one or more activities. To better simplify the terminology and notation, the terms *direct delays* for the former and *indirect delays* for the latter will be used from this point throughout the study. An example of a direct delay is the activity of laying the foundation for a new building. This activity could be the aggregation of several basic tasks (i.e., simple tasks that when taken separately offer no significant benefit as stand-alone activities). The basic tasks could include earth moving, setting up wooden forms for the basement floor and outer walls, pouring concrete into the forms, waiting for the concrete to dry and cure, and removing the wooden forms. Examples of two *types* of indirect delays include a construction delay in laying the foundation of a building from poor

weather, or a manufacturing delay in building a prototype of a new aircraft due to the late delivery of parts. In reference to indirect delay types, each type has its own probability function.

Note that given their very low likelihoods in everyday life, indirect delays that are extraordinary or catastrophic should generally be excluded from consideration. Examples include fire or flooding at the plant, road closures from earthquakes or storms, or delays attributed to acts of God.

This study will show that activity time distributions can be found which accommodate both direct and indirect delays when both delays are estimated separately. The desired activity time distribution can be generated from simulation, or derived analytically as a marginal distribution from bivariate statistics if only one type of indirect delay type is considered, which is demonstrated in this study. This is in contrast to the practice of lumping together direct and indirect delays within a single estimate of a Beta Distribution as the activity time distribution. For example, assume that an activity is described by a Beta Distribution and is based on the usual estimates for *a*, *m*, and *b* which represent optimistic, most likely, and most pessimistic times. If an estimator made an estimate of only a direct delay and was later required to revise that estimate to consider both direct *and indirect delays*, the resulting activity time distribution would likely be a *single estimate* of a *unimodal* Beta Distribution with *a* unchanged and larger time estimates for *m* and *b* to capture the indirect delay (i.e., as *single estimate distribution*). An example based on the methodology proposed in this study assumes that the estimator makes an estimate of the direct delay as a Beta Distribution followed by a *separate* estimate of a distribution for an indirect delay. These estimates are then combined into a multimodal marginal distribution. It is then shown that the unimodal Beta Distribution has *significantly* less entropy (uncertainty) than the multimodal marginal distribution. Use of the unimodal distribution over the multimodal distribution as an activity time distribution in this situation would be a violation of the *Principle of Maximum Entropy, which states that the probability distribution which best represents the current state of knowledge is the one with the greatest entropy* (Jaynes, 1957).

It can be assumed that a similar argument would follow if there were multiple types of indirect delays. The single estimate distribution would again leave *a* unchanged and *m* and *b* would be increased, resulting in a unimodal Beta Distribution. The alternative is to employ multivariate statistics of dimensions q+1 where q is the number of indirect delay types. The distribution for direct delays and each indirect delay are estimated separately, leading to a marginal distribution with an entropy that can be compared to the single estimate distribution, or any other estimation approach that might have been employed. While the computational burden with multiple types of indirect delays significantly increases over the simpler bivariate case with one type of delay, chain rules for multivariate statistics and multivariate entropy are available for an analytical approach, and software is available to derive the marginal distribution with simulation. From the perspective of the estimator, the estimation process would be more straightforward and much less subjective by making separate estimates for each delay type. It lets the mathematics or simulation determine activity time distributions, reducing the added guesswork from a single estimate distribution.

This study also proposes a procedure for determining significant differences between two entropies based on a transformation of each into a discrete Uniform Distribution with equivalent entropy. Differences in the number of events captured by the ranges of these distributions are determined, and a ratio is developed based on

the percent of lost events if the low entropy unimodal distribution was employed over the higher entropy multimodal distribution. These lost events can parallel the significance of lost lottery tickets on the fairness of a lottery.

The literature on estimation of distributions is quite extensive in that it spans several disciplines. However, only a few studies recognize the impact of indirect delays on project network modeling and even fewer propose remedies. Several studies in project management have found this problematic. Empirical findings in (Morgenshtern, Raz and Dvir, 2007) suggested that project managers gave little attention to indirect delays. Atkinson, Crawford and Ward (2006) argued that uncertainty in project management was not given adequate consideration by practitioners, suggesting that indirect delays were ignored, leading to erroneous inferences from completion time distributions that understated the uncertainty of the project. Steele and Huber (2004) found that many established project management techniques assume the use of a common set of distributions (e.g., Beta; Normal; Triangular Distributions), indicating no special accommodation for indirect delays. They argued that when the data was not consistent with the common set of distributions, the usefulness of project management modeling was compromised.

Some studies in project management attempt to address the problems cited above by drawing on work in other disciplines that focus on outliers, which involve distributions with outlying or nonroutine states with likelihoods of sufficient magnitude that should not be ignored. Grant, Cashman and Christenson (1999) documented the outlier phenomena in a real world project that had experienced a four year delay since it failed to accommodate outliers which could be classified as indirect delays. Two studies related to project management offered methods to develop distributions that were more robust in their ability to capture outliers. Steel and Huber (2004) employed Tukey's methods of exploratory data analysis (EDA). However, the burden of estimating data can be extensive for large projects with numerous activities. Hahn (2008) developed a methodology based on "mixture distributions." Such distributions are designed to improve the flexibility and accuracy of data fitting by combining two distributions which improve the chance of capturing outliers. Hahn's study focused on project management and PERT. His remedy was to combine the typical PERT Beta Distribution with the continuous Uniform Distribution, employing a mixture parameter that ranged from 0 (for a pure continuous Uniform Distribution) to 1 (for a pure Beta Distribution). However, this burdens the estimator when selecting among a large number of distributions shaped by the mixture parameter, which may not necessarily lead to the estimator's vision of a representative activity time distribution.

The sections that follow include discussions of methodology and an example to illustrate the concepts proposed in this study.

II. GENERAL PROPERTIES OF THE ENTROPY FUNCTION

Consider the random variable R with distribution function P(R) described by probability set

$$P(R) = \{p(r_1), p(r_2),\ldots p(r_J) \mid 0 \leq p(r_j) \leq 1; \; R=\{r_1,r_2,\ldots r_J\}; \; j=1,2,\ldots J; \; \sum_{r_j \epsilon R} p(r_j)=1\}. \quad (1)$$

The generalized form of the discrete entropy function of R is (Shannon, 1948)

$$H(R) = -\sum_{r_j \epsilon R} p(r_j) \log p(r_j), \quad (2)$$

where in this study the logarithm can be set to any base. Integral expressions of the entropy function exist for continuous distributions, but are not applicable in typical project simulation studies using discrete event simulation software. Note that continuous distributions in discrete event simulation are converted to their discrete approximations.

Several key mathematical properties that follow support the use of the entropy function as a continuous measure of uncertainty for applications in this study. These properties can be found in numerous books and articles on information theory and statistics (e.g., Reza, 1961; Jelinek, 1968; Hays and Winkler, 1975; Jones, 1979; Cosgrove, 2010):

1. The logarithm of zero is undefined, but the entropy function in (2) is finite since

$$\lim_{p(r_j)\to 0} [p(r_j) \log p(r_j)] = 0. \qquad (3)$$

2. H(R) is based on probability sets and is therefore distribution free. It can be determined from (2) for any given probability set for nominal/categorical, ordinal, or metric data.
3. The range of H(R) increases continuously as a measure of uncertainty from Min[H(R)] = 0 [for complete certainty if there exists a single variate r_j such that $p(r_j) = 1$], to Max[H(R)] = log J [for maximum uncertainty where all states are equiprobable with $p(r_j) = 1/J$ for all values of j, where $p(r_j) \neq 0$].
4. While the index j assigns a single index to *all probabilities* $p(r_j) \epsilon P(R)$, there are no restrictions on $p(r_j)$ in terms of representing univariate or multivariate probabilities.

Equivalent entropies convey the same level of uncertainty whether they are derived from univariate or multivariate sources, or from distributions with different shapes. However, entropy is technically a dimensionless measure which at best can provide an ordinal scale for ranking the uncertainty of different stochastic processes. Measuring significant differences with entropy using a transformation will be discussed later in this study.

III. DISTRIBUTION FUNCTIONS FOR DIRECT AND INDIRECT DELAYS

Let I constants in set K represent I indirect delays such that $K = \{k_i | i=1,2,3,...I; k_1=0; k_1<k_2<k_3,\ldots k_{i-1}<k_i,\ldots k_{I-1}<k_I\}$. Consider the discrete joint distribution function P(Y,X) with variates y_j corresponding to *direct and indirect delays* and x_i to *indirect delays*, such that $Y=\{y_j|j=1,2,3,....J\}$ and $X=\{x_i|i=1,2,3,...I\}$ with underlying marginal distribution functions P(Y) and P(X). Now let $P^D(Y_+ k_i|x_i)$ be the i*th* conditional distribution *estimated* as the i*th direct delay*. For $I \geq i \geq 2$, the distributions for direct delays *are shifted* k_i *time units on the time axis* to account for indirect delays. This permits $P(Y|x_i)$ conditional distributions (for all values of i) to capture *both direct and indirect time delays.* [Note that there is no time shift at i=1, where $k_1=0$ such that $P(Y|x_1)=P^D(Y|x_1)$ are distributions representing direct delays only.] From bivariate statistics, the marginal distribution of P(Y) follows from

$$p(y_j) = \sum_{x_i \epsilon X} p(x_i)\, p^D(y_{j+}\, k_i|x_i)$$

$$= \sum_{x_i \epsilon X} p(x_i)\, p(y_j|x_i). \qquad (4)$$

Section VI will present an example of a network representation based on (4) which generates the activity time distribution from simulation.

The estimates for direct delays may or may not have the same shape for all i. If they do, only one estimate for a single direct delay is necessary since the others are simply distributions of the same shape that are shifted by k_i for $i \geq 2$. The estimate for indirect delays determines P(X) and is taken as independent and separate from the estimate for direct delays. Expression (4) also shows the probabilities of P(X) as the weights for a weighted average of the conditional distributions which include direct and indirect delays. The interpretation of (4) as a weighted average lends support for its use in determining P(Y) as the desired activity time distribution that captures both direct and indirect delays.

IV. ENTROPY FUNCTIONS FOR DIRECT AND INDIRECT DELAYS

The expressions in this section are well established in information theory (e.g., Jones, 1979; Jelinek, 1968) and are presented for their applicability to this study.

The univariate marginal entropy functions H(Y) and H(X) are given by

$$H(Y) = -\sum_{y_j \epsilon Y} p(y_j) \log p(y_j), \qquad (5)$$

$$H(X) = -\sum_{x_i \epsilon X} p(x_i) \log p(x_i). \qquad (6)$$

The conditional entropy function of Y given variate x_i is

$$H(Y|x_i) = -\sum_{y_j \epsilon Y} p(y_j|x_i) \log p(y_j|x_i). \qquad (7)$$

The above expressions can now be used to determine the following:

1. $H^E(Y)$: Entropy of the distribution $P^E(Y)$, the *estimator's* activity time distribution from lumping together direct and indirect delays in the estimate of a *single distribution.* Its entropy follows from (5) by substituting $H^E(Y)$ for H(Y) and $p^E(y_j)$ for $p(y_j)$.
2. H(Y): The marginal entropy of the marginal distribution P(Y) from P(Y,X) as specified above in (5).
3. H(X): The marginal entropy of the marginal distribution P(X) from P(Y,X) as specified above in (6).
4. $H(Y|x_i)$: Entropy of the i*th* conditional distribution $P(Y|x_i)$ as specified above in (7).

V. TRANSFORMATION FOR MEASURING SIGNIFICANT DIFFERENCES

Fig. 1 shows two discrete Uniform Distributions consisting of n and m states with entropies H(N) for P(N) and $H^E(M)$ for $P^E(M)$, where $H(N) = \log n$ and $H^E(M) = \log m$ (Wikipedia, 2014). If H(N)=H(Y) and $H^E(M)=H^E(Y)$ with $H^E(M)<H(N)$ and $m<n$, then H(Y) is a better estimate than $H^E(Y)$ for an activity time distribution according the *Principle of Maximum Entropy*.

FIGURE 1. EQUIVALENT ENTROPY DISTRIBUTIONS FOR P(Y) AND $P^E(Y)$

It now follows that

$$\log n = H(Y) \quad (8)$$

with

$$n = \log^{-1}[H(Y)], \quad (9)$$

and

$$\log m = H^E(Y) \quad (10)$$

with

$$m = \log^{-1}[H^E(Y)]. \quad (11)$$

The entropies in (8) and (10) have been transformed using the inverse logarithm in (9) and (11) to an equivalent number of events or states of a discrete Uniform Distribution. If

$$DIFF = n - m, \quad (12)$$

then DIFF represents *the number of states or events lost in a Uniform Distribution* when using $P^E(Y)$ as the activity time distribution rather than P(Y).

An interpretation of (12) follows from a simple fair lottery game with each ticket corresponding to a state of a discrete uniform distribution. If n tickets were sold but m where placed in the drum, then DIFF tickets were lost. These lost tickets increase chances of winning for holders of the remaining tickets. *The question to be addressed is whether enough tickets were lost with a likelihood to significantly alter the outcome of the lottery*. Twenty lost tickets from 1,000,000 sold are unlikely to make a significant difference. However, 20 lost tickets from 100 sold are likely to be quite significant. The difference in (12) represents the lost uncertainty or entropy when using $P^E(Y)$ as the activity time distribution compared to using P(Y). This lost uncertainty would most likely manifest itself by understating the true variance of the activity time distribution.

The following ratio is proposed to determine significant difference as a percent of lost states or events:

$$SDR = (100)(n - m)/n = (100)(DIFF)/n. \quad (13)$$

Expression (13) will be applied to the example in the next section.

VI. EXAMPLE

The example employs a simple simulation to generate P(Y) as the appropriate estimate for an activity time distribution. P(Y) is the distribution of interest given that it captures the probabilistic behavior of both direct and indirect delays as the marginal distribution of a bivariate system. The simulation model is shown in Fig. 2 and uses the GERT/VERT AOA project network

framework (Moore and Clayton, 1976; Moeller and Digman, 1981), which unlike PERT provides a framework that can accommodate stochastic branching.

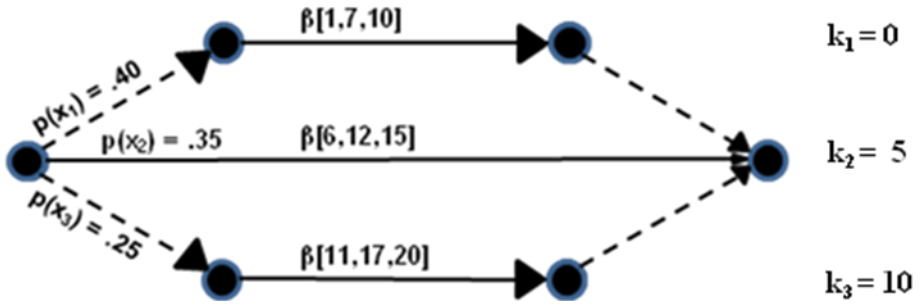

FIGURE 2. SIMULATION MODEL GENERATING P(Y)

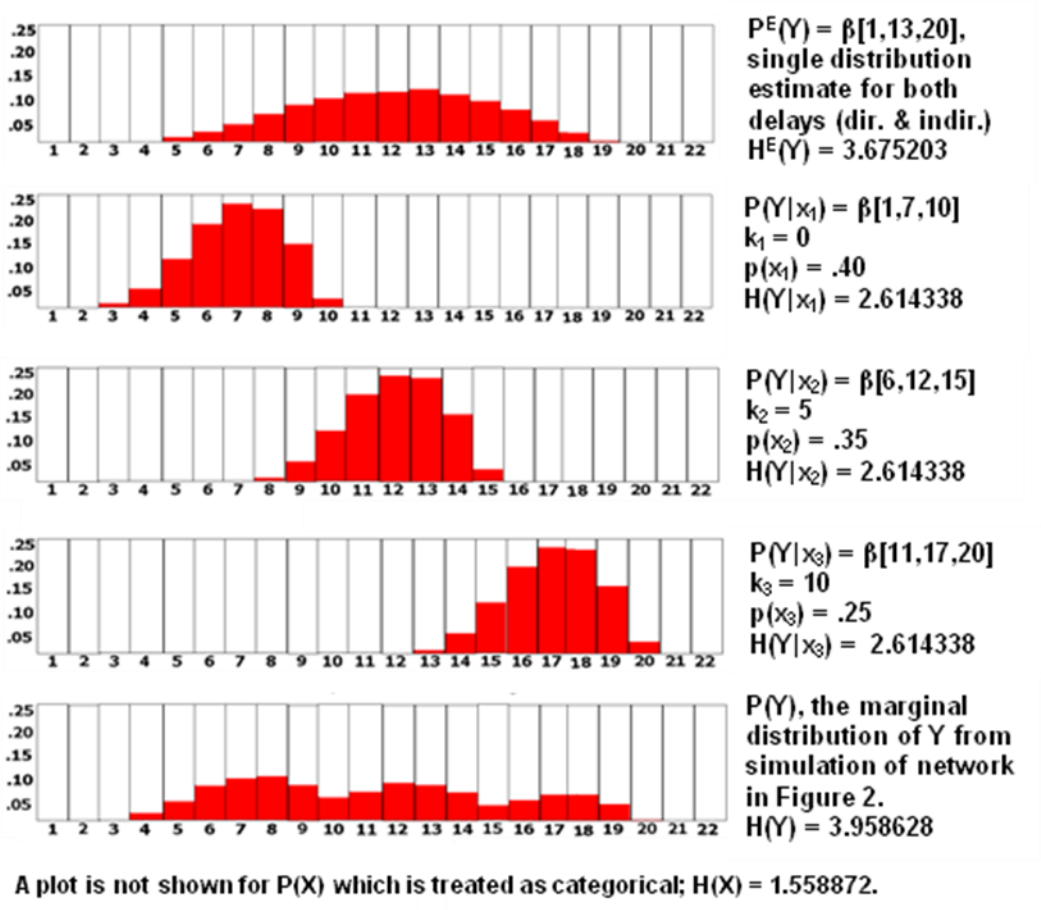

FIGURE 3. SIMULATION OUTPUT AND CORRESPONDING ENTROPIES

The model shows three Beta Distributions in the typical PERT format of β[*a,m,b*]. The distribution for the top branch captures only direct delays (k_1=0). The middle and bottom branches are adjusted to capture both direct and indirect delays. These distributions have the same shape as the distribution on the top branch, but their respective values of *a*, *m*, and *b* are shifted to capture indirect delays of 5 and 10 time units (k_2=5 and k_3=10). The simulation converts these continuous Beta Distributions into discrete approximations. The results of these conversions, with the time shifts to accommodate indirect delays, are shown in Fig. 3 as the three discrete conditional distributions in the middle of the figure.

It is evident by observation of Fig. 3 that the distribution $P^E(Y)$ significantly differs from P(Y), the latter of which is the completion time distribution from the simulation. The multimodal properties of P(Y) are a more realistic representation of the behavior of combined direct and indirect delays. Note that the entropies in the figure follow from expressions in Section IV.

Since $H(Y) > H^E(Y)$, P(Y) is the preferred estimate for the activity time distribution according to the *Principle of Maximum Entropy*. Referring to the lottery analogy, it follows from (13) that use of $P^E(Y)$ over P(Y) reflects a loss of 18% of the lottery tickets making such a lottery unfair, and clearly indicating that the distributions are significantly different.

There are a few technical comments related to this example. All logarithms were arbitrarily taken to the base 2. Simulations were performed with the Arena simulation package (Kelton, Sadowski and Swets, 2010) with at least 20,000 replications on the model in Fig. 2 for each run. This gives an error within ± .01 probability on the cumulative distribution with a confidence level of .95 when employing the Kolmogorov-Smirnov procedure as outlined in (Van Slyke, 1966). Arena is similar to most simulation packages in providing as an input the *standard form* of the Beta Distribution which ranges from [0,1] and requires two shaping parameters. After estimating *a*, *m*, and *b* for the version of the Beta Distribution employed in PERT project networks, a transformation is required to rescale the standard Beta Distribution to accommodate a range of [*a*,*b*]. The transformation also requires a calculation of two shaping parameters in terms of *a*, *m*, and *b*. This study considered two approaches (Vose, 2000; Davis, 2008) for calculating the shaping parameters. Eight simulations with 20,000 replications of the continuous Beta Distributions for $P^E(Y)$ and $P(Y|x_i)$ for i=1,2,3 were performed with shaping parameters from both approaches, giving the simulation output as discrete approximations for the four distributions from the top as shown in Fig. 3. The Davis approach was selected because it gave better estimates of the mode *m* for all four discrete approximations of the Beta Distributions.

Obtaining P(Y) and discrete entropy measures using the GERT/VERT framework in Fig. 2 is straightforward. Additional effort is required when using a PERT framework which does not accommodate stochastic branching. The conditional discrete distributions in Fig. 3 which were needed to calculate the conditional entropies from (7) can be generated by simulation using the PERT framework. They can be generated from three simple PERT simulation models with each model consisting of a branch of the network in Fig. 2. It then follows that P(Y) is determined directly from (4).

In our example, the conditional distributions in Fig. 3 have the same shape. Only one PERT simulation would be necessary since the other two distributions require shifts of 5 and 10 time units (i.e., *a*, *m*, and *b* for k_1=0 are all shifted by 5 time units when k_2=5 and by 10 time units when k_3=10).

Unlike PERT networks, the use of the GERT/VERT network framework completely avoids the need for calculating P(Y) or even referring to it. P(Y) is captured within the network through stochastic branching. However, the number of branches and size of the network could increase considerably if there are a large number of activities subject to indirect delays. Separate determinations of activity time distributions using networks similar to Fig. 2 would reduce the number the stochastic branches, which reduces the size of the GERT/VERT project network.

VII. CONCLUSION

This study has proposed tools for improving estimations of activity time distributions by proposing methodology that permits practitioners to better accommodate indirect delays that are not directly associated with direct delays, but have sufficient likelihoods to be a threat in causing significant project delays if not properly taken into account. Given the lack of literature in project management on indirect delays, and the wealth of literature in other disciplines that consider nonroutine delays (e.g., outlier literature), there is an opportunity to transfer methods from such literature to applications in project management or to develop new methods to address the problem. This study is an example of the latter by recognizing the potential of entropy to measure uncertainty and to provide a criterion to compare the effectiveness of different approaches to distribution estimation. While this study considered only one type of indirect delay and treated it within a bivariate framework, this avenue of research can be extended to multiple types of indirect delays involving multivariate frameworks with numerous random variables.

The project management literature cited in this study was concerned about the consequences that follow from poor estimation leading to *understating* project uncertainty. However, all available information should be inputted into a project without understating or *overstating* the uncertainty of activity time distributions. The example in this study found that the multimodal marginal distribution P(Y) was the better candidate for the activity time distribution than the single estimate distribution of unimodal $P^E(Y)$, based on the *Principal of Maximum Entropy*. What if $P^E(Y)$ had the greater entropy? Generally, cases where outliers such as indirect delays are not fully captured in a distribution are likely to show fewer modes than cases that capture them, and the *Principle of Maximum Entropy* is an appropriate criterion in selecting the best estimate. Distributions with more modes of similar amplitude tend to take on a flatter shape, making them better approximations of the discrete Uniform Distributions than distributions with fewer modes. (For a given range, the Uniform Distribution is the maximum entropy distribution.) However, if $P^E(Y)$ exhibits the greater entropy, perhaps the lumping of an estimate with direct and indirect delays within a single distribution would seem more difficult than two separate estimates, leading an estimator to make a more conservative single distribution estimate. If this can be established as the reason, then the focus should shift away from a criterion based on the Principle of Maximum Entropy to whether the estimator has more confidence in making separate estimates over the single distribution estimate. Separate estimates should be less burdensome to the estimator. The heavy lifting is done by multivariate statistics which introduces more mathematics and less subjectivity into the process than the single distribution estimate.

VIII. REFERENCES

Atkinson, R., Crawford L. and Ward, S., "Fundamental uncertainties in projects and the scope of project management,"

International Journal of Project Management, 2006, 24(8), 687-698.

Cosgrove, W.J., "Entropy as a measure of uncertainty for PERT network completion time distributions and critical path probabilities," *California Journal of Operations Management*, 8(2), 2010, 20-26.

Davis, R., "Teaching project simulation in Excel using PERT-Beta Distributions," *INFORMS Transactions on Education*, 8(3), 2008, 139-148.

Grant, K.P., Cashman, W.M. and Christenson, D.S., "Delivering projects on time," *Research Technology Management*, 49(6), 2006, 52-58.

Hahn, E.D., "Mixture densities for project management activity times: a robust approach to PERT," *European Journal of Operational Research*, 188(2), 2008, 450-459.

Hays, W.L. and Winkler, R.L., *Statistics: Probability, Inference, and Decision* (2nd Ed.), Holt, Rinehart, and Winston, New York, 1975.

Jaynes, E.T., "Information theory and statistical mechanics," *Physical Review*, Series II, 106(4), 1957, 620–630.

Jaynes, E.T., "Information theory and statistical mechanics II," *Physical Review*, Series II, 108(2), 1957, 171–190.

Jelinek, F., *Probabilistic Information Theory*, McGraw-Hill, New York, 1968.

Jones, D.S., *Elementary Information Theory*, Clarendon Press, Oxford, l979.

Kelton, W.D., Sadowski, R.P. and Swets, N.B., *Simulation with Arena* (5th Ed.), McGraw-Hill, Boston, 2010.

Moeller, G.L. and Digman, L.A., "Operations planning with VERT," *Operations Research*, 29(4), 1981, 676-697.

Moore, J.M. and Clayton, E.R., *GERT Modeling and Simulation: Fundamentals and Applications*, Petrocelli/Charter, New York, 1976.

Morgenshtern, O., Raz, T. and Dvir, D., "Factors affecting duration and effort estimation errors in software development projects," *Information and Software Technology*, 49, 2007, 827-837.

Reza, F.M., *An Introduction to Information Theory*, McGraw-Hill, New York, 1961.

Shannon, C.E., "A mathematical theory of communication," *Bell System Technical Journal*, 27(7), 1948, 379-423.

Shannon, C.E., "A mathematical theory of communication II," *Bell System Technical Journal*, 27(10), 1948, 623-656.

Steele, M.D. and Huber, W.A., "Exploring data to detect project problems," *Association for the Advancement of Computing in Education (AACE) International Transactions*, 32(12), 2004, PM.21.1-PM.21.7.

Van Slyke, R.L., "Monte Carlo methods and the PERT problem," *Management Science*, 11(5), 1966, 839-860

Vose, D., *Risk Analysis – A Quantitative Guide*, John Wiley and Sons, New York, 2000.

Wikipedia, "Uniform Distribution (discrete)," *Wikipedia*, http://en.wikipedia.org/wiki/Uniform_distribution_%28discrete%29 (accessed February 14, 2014).