# Correlation sensitivity analysis in stochastic project networks

by

Anand A. Paul[1] and Dan Trietsch[2]

November 2012

We study the interplay between correlation in stochastic project networks and project completion time. Suppose the activity durations have arbitrary marginal distributions and a Gaussian copula. The impact of an increase in correlation between *parallel activities* is a stochastic decrease in project duration regardless of network topology, whereas the impact of an increase in correlation between *serial activities* is dramatically different. We identify an important subclass of serial activities—including all serial activities in series-parallel networks—for which an increase in correlation between serial activities causes an increase in project duration in the increasing convex order sense. When the network is not series-parallel, some serial activities may have an indeterminate effect. We discuss an application of these results to project subcontracting. A special case known to be valid in practice and involving a Gaussian copula is when durations are multivariate lognormal.

[1] Warrington College of Business Administration, University of Florida.
[2] College of Engineering, American University of Armenia.

# 1. Introduction

The results in this short paper may be broadly classified under the rubric of 'sensitivity analysis of project networks.' We are interested in studying the interplay between correlation in stochastic project networks and project completion time. The main theoretical contribution is to fill a gap in the theory of stochastic project networks by providing a categorical answer to the question of the impact of correlation between activities or paths on the statistics of project duration for arbitrary marginal distributions of activity durations. Our results also shed some light on the practical issue of designing subcontracts for large scale projects.

For brevity we shall use the term 'activity' to mean an individual project activity, a sub-path in a project network, or a component subproject of a project. It is understood, however, that any two composite activities are completely distinct from each other, without any shared sub-activities and such that each activity has unique start and finish events. For instance, in Figure 1 we could refer to Activities B through F as a single (composite) activity whose start event is the completion of A and finish event is the start of G. But activities B and E cannot be represented by a single composite activity because when B completes D can start so the finish event is not unique. Suppose activity durations have arbitrary marginal distributions and a Gaussian copula. We use the activity-on-node (AON) method to depict project networks. Our main results may be summarized as follows.

1. The impact of an increase in *correlation* between *parallel activities* is a stochastic decrease in project duration regardless of network topology.
2. The impact of an increase in correlation between *serial activities*, however, may depend on network topology. We identify a subclass of serial activities that are *decomposable*. For instance, when the network is series-parallel, all serial activities are decomposable. For decomposable serial activities the impact of an increase in correlation between serial activities is a stochastic increase in project duration in the increasing convex order sense. That implies that both the magnitude and the variability of the project duration increase (at least weakly).
3. For serial activities that are not decomposable, the effect of an increase in correlation is indeterminate.

The effect of serial activities is often important: it is easy to construct Monte Carlo simulation models where substituting a pair of independent activities by a pair of correlated activities

with correlation coefficient 0.8 in a series-parallel network can increase expected project duration by 20% or more. From a practical standpoint, changes in correlation between paths or subprojects may be more relevant than changes in correlation between individual activities; such changes may have an even more significant impact on project completion time.

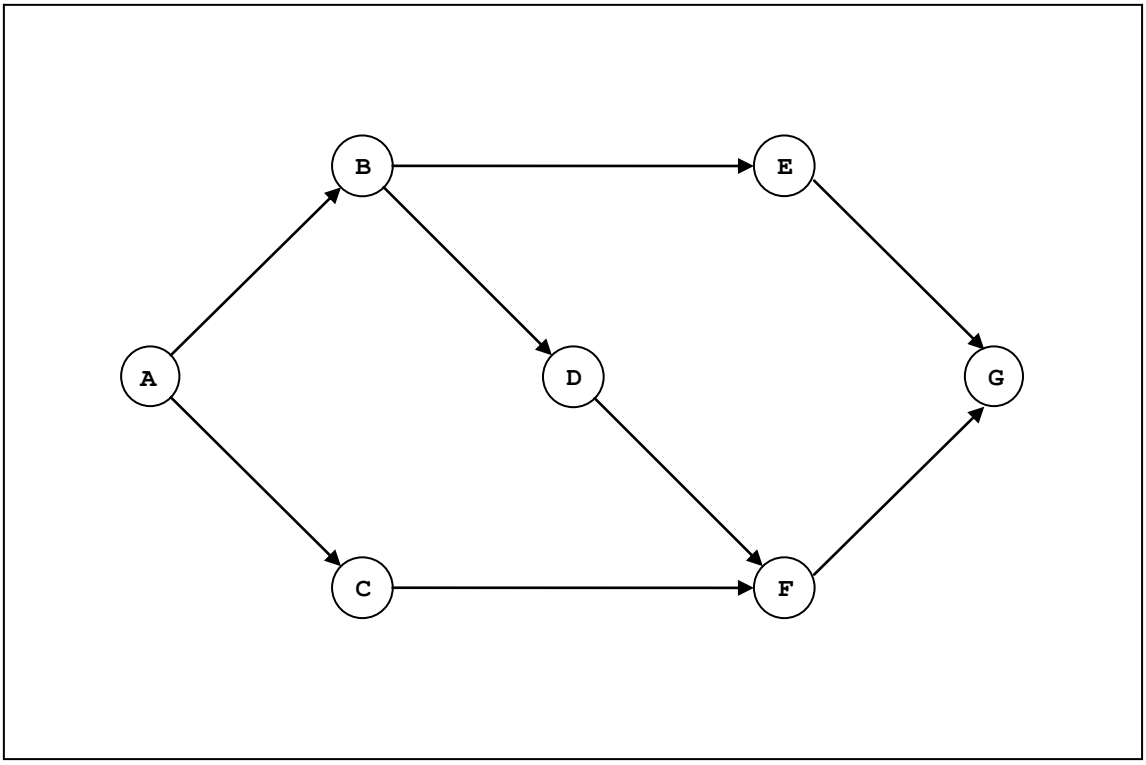


*Figure 1.* An Example Project

In section 2 we describe related work, and in section 3 of the paper we present our main results. In section 4, we discuss the results and describe how they may be interpreted in light of project subcontracting.

# 2. Related work

In a classic paper on the limitations of PERT, MacCrimmon and Ryavec (1964) inferred from numerical examples that the greater the correlation between paths because of overlapping paths in a project network, the smaller the PERT bias. However, Banerjee and Paul (2008) showed via a counterexample that this connection between path correlation and PERT bias does not always prevail. Banerjee and Paul (2008) showed that in the case of a project network with multivariate

normal activity times and a covariance matrix characterized by only non-negative terms, the completion times of activities are positively correlated.

Elmaghraby (2000) studies sensitivity analysis in stochastic project networks. His focus is on studying the sensitivity of the mean and variance of project completion time to the mean and variance of activity completion times. Corbett and Rajaram (2006) use copulas and multivariate stochastic ordering to study multivariate dependence − as we do in the present paper − although there is no overlap between our respective models, since we study the impact of correlation between activity times on project completion time whereas their focus is on studying the inventory stocking implications of correlation between demands.

What we learn from past work investigating the impact of correlation in project networks is that the underlying effect is subtle and tends to resist characterization via simple rules of thumb. However, we show that when the activity durations have arbitrary marginal distributions linked by a Gaussian copula, clear-cut structural results emerge that point to the role played by specific network topologies in determining whether increased activity correlation has a positive or negative impact on project completion time. We also consider another model for correlated activities: linear association (Baker and Trietsch, 2009). In that model correlation is introduced by the effects of a common bias element rather than via a correlation matrix; e.g., biased estimates lead to positive correlation. Trietsch et al. (2012) provide statistical evidence that linear association can explain field data from several sources. They also validate the use of the lognormal distribution for activity times. We note specifically that correlated activities with lognormal marginal distributions must have a Gaussian copula, and hence the results we develop for the Gaussian copula are relevant to practice.

# 3. Results

The key topological feature that demarcates the impact of an increase in activity correlation is whether the perturbed activities are in series or in parallel. In section 3.1 we collect some standard definitions and results from multivariate analysis and multivariate stochastic ordering that we need to state in order to prove our results rigorously. In section 3.2 we discuss parallel activities, and in section 3.3 we discuss serial activities. We discuss linear association as an alternative model of correlated activities in section 3.4. In section 3.5 we discuss the special case of multivariate lognormal activity duration distribution.

## 3.1 Basic technical definitions and facts

We collect some standard definitions and results from multivariate analysis and multivariate stochastic ordering that we need. We write 'increasing' for 'non-decreasing' and 'decreasing' for 'non-increasing.' Shaked and Shanthikumar (2007) and Muller and Stoyan (2002) are standard sources for this material.

**Definition 1:** Let X and Y be random variables. We say that X is smaller than **Y** in the *increasing convex order* (denoted $X \leq_{icx} Y$) if $E(\emptyset(X)) \leq E(\emptyset(Y))$ for all increasing convex real-valued functions $\emptyset(.)$. If we omit the qualifier 'increasing', we obtain the *convex order*, denoted $X \leq_{cx} Y$. If and only if EX = EY, $X \leq_{icx} Y$ is equivalent to $X \leq_{cx} Y$.

**Definition 2:** A function $f: R^k \rightarrow R$ is *supermodular* if

$$f(x \vee y) + f(x \wedge y) \geq f(x) + f(y)$$

for all $x, y \in R^k$, where $x \vee y$ denotes the component-wise maximum and $x \wedge y$ the component-wise minimum of $x$ and $y$. A random vector $X = (X_1,\ldots,X_N)$ is said to be smaller than the random vector $Y = (Y_1,\ldots,Y_N)$ in the *supermodular order* – denoted $X \leq_{sm} Y$ – if $Ef(X) \leq Ef(Y)$ for all supermodular functions f(.).

**Fact 1:** If $X \leq_{sm} Y$ it follows that $Cov(X_i,X_j) \leq Cov(Y_i,Y_j)$ for all (i,j) and $X_i =_{st} Y_i$ for all i.

Fact 1 affirms that supermodular ordering is a dependence ordering. If X and Y are vectors of activity durations in projects A and B, respectively, project B has – in some sense – a greater degree of intrinsic correlation between its activities.

**Fact 2:** Let X and Y be two multivariate normal vectors of the same dimension. Then $X = (X_1,\ldots,X_N) \leq_{sm} Y = (Y_1,\ldots,Y_N)$ if and only $X_i =_{st} Y_i$ for all i and $Cov(X_i,X_j) \leq Cov(Y_i,Y_j)$ for all (i,j) .

**Fact 3**: If $X = (X_1,\ldots,X_N) \leq_{sm} Y = (Y_1,\ldots,Y_N)$ then $(g_1(X_1),\ldots, g_N(X_N)) \leq_{sm} (g_1(Y_1),\ldots, g_N(Y_N))$ for all increasing functions $g_i(.)$.

**Fact 4:** Let $f(x_1,\ldots,x_k) = x_1+\ldots+ x_k$. Then f(.) is supermodular.

**Fact 5:** If f(x) is an increasing convex function and g is an increasing supermodular function, then the composition function f[g(.)] is supermodular. (Marshall and Olkin, Page 151, D2)

**Fact 6:** If $(X_1,\ldots,X_N) \leq_{sm} (Y_1,\ldots,Y_N)$, then $\text{Max}(X_1+\ldots+X_N, a) \leq_{icx} \text{Max}(Y_1+\ldots+Y_N, a)$ for all real numbers a.

(If f(x) is an increasing convex function, then it follows from Fact 4 and Fact 5 that $f(X_1+\ldots+X_N)$ is supermodular. Hence $Ef(X_1,\ldots,X_N) \leq Ef(Y_1,\ldots,Y_N)$ , establishing Fact 6.)

**Definition 3:** The vector $X = (X_1,\ldots,X_N)$ is said to be smaller than $Y = (Y_1,\ldots,Y_N)$ in the *upper orthant order* (denoted $(X_1,\ldots,X_N) \leq_{uo} Y = (Y_1,\ldots,Y_N)$) if $\text{Prob.}( X_1 > t_1,\ldots, X_N > t_N) \leq \text{Prob.}(Y_1 > t_1,\ldots, Y_N > t_N)$ for every $(t_1,\ldots,t_N)$.

**Fact 7:** If $X \leq_{sm} Y$ then $X \leq_{uo} Y$.

## 3.2 Parallel activities

Let $X = (X_1,\ldots,X_N)$ be a multivariate normal vector. We use this vector to generate multivariate vectors with arbitrary marginal distributions sharing the same dependence structure, or copula, as the multivariate normal distribution. It is a standard fact that if X is a random variable with a continuous distribution function F(.), then F(X) is uniformly distributed between 0 and 1, and therefore $G^{-1}F(X)$ is a random variable with distribution function G (where the generalized inverse function $G^{-1}(y) = \inf\{x: G(x) \geq y\}$). Note that $G^{-1}F(.)$ is an increasing function. Let $X_i$ be normally distributed. Since the normal distribution function is continuous, it follows that we may write each $Y_i$ in the form $h_i(X_i)$, where $h_i(.)$ is an increasing function. In this way, we generate a new multivariate vector with arbitrary marginal distributions.

The joint distribution of $Y = (Y_1,\ldots,Y_N)$ is not multivariate normal but it shares the following feature with the multivariate normal vector $X = (X_1,\ldots,X_N)$ from which it was generated. Let $F_i(.)$ denote the marginal distribution of $X_i$ for all i, and let $F(x_1,\ldots,x_N)$ denote the joint distribution function of X. Consider the map C: $(F_1(x_1),\ldots, F_N(x_N)) \rightarrow F(x_1,\ldots,x_N)$ from $R^N$ to R. This specific map is called a *Gaussian copula* with a given mean vector and correlation matrix. The random vectors Y and X share the same map C although they have different marginal distributions and different joint distributions. For instance, Corbett and Rajaram (2006) suggest that assuming a Gaussian copula may be a practical necessity, as the information required for estimat-

ing more complex copulae is rarely available. Thus, the Gaussian copula is a natural model with which to study multivariate dependence effects, particularly when we want to characterize dependence between two activities or subprojects via their correlation coefficient. We model activity durations with arbitrary marginal distributions and a Gaussian copula.

When we increase the correlation between a given activity pair (X,Y), it is understood that we change nothing else. Specifically, the marginal distributions of all the activities remain unchanged, and the joint distributions of every subset of random variables not including *both* X and Y remain unchanged. We need three preliminary lemmas.

**Lemma 1 (**Theorem 5.1.8, Tong 1990) *Let* $h(x_1,\ldots,x_N) = f(x_1,\ldots,x_k)g(x_{k+1},\ldots,x_N)$*, where f and g are increasing functions. Then* $Ef(X_1, X_2, \ldots, X_k)\, Ef(X_{k+1}, X_2, \ldots, X_N)$ *is an increasing function of* $\sigma_{ij}$ *for* $1 \leq i \leq k < j \leq n$.

**Lemma 2** *Let* $X = (X_1,\ldots,X_N)$ *be a multivariate normal vector with a given mean vector and covariance matrix. Let* $Y = (Y_1,\ldots,Y_N)$ *where* $Y_i = h_i(X_i)$*, and each* $h_i(.)$ *is an increasing function. Suppose* $Y' = (Y'_1,\ldots,Y'_N)$ *such that* $Y'_i =_{st} Y_i$ *for all i,* $Cov(Y_i,Y_j) \leq Cov(Y'_i,Y'_j)$*, and* $Cov(Y'_m,Y'_n) = Cov(Y_m,Y_n)$ *for all* $(m,n) \neq (i,j)$ *. Let* $X'_i =_{st} g_i(Y'_i)$ *where* $g_i(.)$ *is the generalized inverse function of* $h_i(.)$ *so that* $X'_i =_{st} X_i$ *for all i. Then* $Cov(X_i,X_j) \leq Cov(X'_i,X'_j)$ *and* $Cov(X'_m,X'_n) = Cov(X_m,X_n)$ *for all* $(m,n) \neq (i,j)$ *.*

**Proof:** Set $f(x_1,\ldots,x_k) = w(x_i)$ and $g(x_{k+1},\ldots,x_N) = g(x_j)$ in Lemma 1. Then Lemma 1 states that $Ew(X_i)v(X_j)$ is an increasing function – say K(.) - of the correlation between $X_i$ and $X_j$. The marginal distributions of $X_i$ and $X_j$ remain unchanged; hence $Ew(X_i)$ and $Ev(X_j)$ remain unchanged. It follows that $Cov(w(X_i),v(X_j)) = Ew(X_i)v(X_j) - Ew(X_i)Ev(X_j)$ increases. This establishes sufficiency. Necessity follows from the observation that the generalized inverse of an increasing function K(x) – defined by $K^{-1}(x) = \inf\{y: K(y) \geq x\}$ - is an increasing function. **Q.E.D.**

**Lemma 3** *Let* $Y = (Y_1,\ldots,Y_N)$ *be a random vector with a Gaussian copula. Suppose* $Y' = (Y'_1,\ldots,Y'_N)$ *such that* $Y'_i =_{st} Y_i$ *for all i,* $Cov(Y_i,Y_j) \leq Cov(Y'_i,Y'_j)$*, and* $Cov(Y'_m,Y'_n) = Cov(Y_m,Y_n)$ *for all* $(m,n) \neq (i,j)$ *.Then* $((Y_i',Y_j')|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},\ Y_{i+1}=y_{i+1},\ldots,\ Y_N=y_N) \geq_{sm} ((Y_i,Y_j)|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},\ Y_{i+1}=y_{i+1},\ldots,\ Y_N=y_N)$.

**Proof:** $\text{Prob.}(Y_i \leq a_i,\ Y_j \leq a_j|\ Y_1=y_1,\ldots,Y_{i-1}=y_{i-1},\ Y_{i+1}=y_{i+1},\ldots,\ Y_N=y_N) =$

$\text{Prob.}(X_i \leq b_i,\ X_j \leq b_j|\ X_1=x_1,\ldots,X_{i-1}=x_{i-1},\ X_{i+1}=x_{i+1},\ldots,\ X_N=x_N)$

where for each positive integer *n* we have that $G_n(.)$ is the distribution function of a normal distribution with arbitrarily chosen mean $\mu_n$ and standard deviation $\sigma_n$, $F_n(.)$ is the distribution function of $X_n$, $X_n = G^{-1}F(Y_n)$, $x_n = G^{-1}F(y_n)$, and $b_n = G^{-1}F(a_n)$.

Now $((X_i,X_j) \mid X_1=x_1,\ldots,X_{i-1}=x_{i-1}, X_{i+1}=x_{i+1},\ldots, X_N=x_N))$ follows a bivariate normal distribution (see Muller and Scarsini (2000), 116-117). It follows from Fact 1 and Lemma 2, that $((X_i',X_j')|X_1=x_1,\ldots,X_{i-1}=x_{i-1}, X_{i+1}=x_{i+1},\ldots, X_N=x_N) \geq_{sm} ((X_i,X_j)|X_1=x_1,\ldots,X_{i-1}=x_{i-1}, X_{i+1}=x_{i+1},\ldots, X_N=x_N)$.The lemma now follows from Fact 3, which states that the supermodular order is closed under increasing transformations of random variables. **Q.E.D.**

**Proposition 1** *Consider a project network in which the activity durations have arbitrary marginal distributions and a Gaussian copula. The effect of increasing the correlation between any pair of parallel activities is to stochastically decrease project completion time.*

**Proof:** Suppose there are N activities. Let $X = (X_1,\ldots,X_N)$ be a multivariate normal vector with a given mean vector and covariance matrix. Let $Y = (Y_1,\ldots,Y_N)$ where $Y_i = h_i(X_i)$, for some appropriately chosen increasing function $h_i(.)$ such that the marginal distributions of $Y_i$ match the given marginal distributions of activity times. Suppose $Y' = (Y'_1,\ldots,Y'_N)$ such that $Y'_i =_{st} Y_i$ for all i and $Cov(Y_i,Y_j) \leq Cov(Y'_i,Y'_j)$ for all (i,j) and $Cov(Y'_m,Y'_n) = Cov(Y_m,Y_n)$ for all $(m,n) \neq (i,j)$ . Let $X'_i =_{st} g_i(Y'_i)$ where $g_i(.)$ is the generalized inverse function of $h_i(.)$. Then $X'_i =_{st} X_i$ for all i and $Cov(X_i,X_j) \leq Cov(X'_i,X'_j)$ for all (i,j) and $Cov(X'_m,X'_n) = Cov(X_m,X_n)$ for all $(m,n) \neq (i,j)$ by Lemma 1. Hence $X' \geq_{sm} X$ by Fact 2, and $(h(X'_1),\ldots,h(X'_N)) \geq_{sm} (h(X_1),\ldots,h(X_N))$ by Fact 3. That is, $Y' \geq_{sm} Y$. In particular, $(Y_i',Y_j') \geq_{sm} (Y_i,Y_j)$ since the supermodular order is closed under marginalization. Now by Fact 7 it follows that $(Y_i',Y_j') \geq_{uo} (Y_i,Y_j)$; that is $Prob.(Y'_i > t_i, Y'_j > t_j) \geq Prob.(Y_i > t_i, Y_j > t_j)$ for all $t_i,t_j$.

Now fix the durations of all the activities other than activity i and j arbitrarily. That is, let $Y_m = Y'_m = t_m$ for all $m \neq i,j$. Then the conditional distributions of the durations of activities i and j - denoted $(Z_i',Z_j')$ and $(Z_i,Z_j)$ - remain ordered, by Lemma 2. That is, $(Z_i',Z_j') \geq_{sm} (Z_i,Z_j)$, and hence $(Z_i',Z_j') \geq_{uo} (Z_i,Z_j)$.

Case 1: First, suppose that there is exactly one path $P_i$ through activity i and exactly one path $P_j$ through activity j. Let the sum of all activity durations on path $P_i$ (excluding $Z_i$) and on path $P_j$ (excluding $Z_j$) be $T_i$ and $T_j$ respectively. Let the sum of all activity times of on path $P_n$ ($n \neq i,j$) be $T_n$. Suppose there are M paths.

We have Prob.($Z'_i > t_i - T_i,\ Z'_j > t_j - T_j$) $\geq$ Prob.($Z_i > t_i - T_i,\ Z_j > t_j - T_j$) for all $t_i, t_j$. That is, Prob.($Z'_i + T_i > t_i,\ Z'_j + T_j > t_j$) $\geq$ Prob.($Z_i + T_i > t_i,\ Z_j + T_j > t_j$) for all $t_i, t_j$. This is equivalent to $\text{Max}(P'_i, P'_j) \geq_{st} \text{Max}(P_i, P_j)$ , which implies that

$$\text{Max}(P'_i, P'_j, T_1, \ldots, T_M) \geq_{st} \text{Max}(P_i, P_j, T_1, \ldots, T_M).$$

Now integrating this inequality with respect to the joint distribution of all the activity times that were fixed – and noting that the stochastic order is closed under mixtures - we obtain the claimed result.

Case 2: There are multiple paths through activities i and j. In this case, we apply the result obtained under Case 1 to infer that the completion time of the maximum of every pair of paths passing through activities i and j increases stochastically. It follows that the completion time of the project decreases stochastically. **Q.E.D.**

In effect, this proposition generalizes Slepian's Inequality for the multivariate normal distribution (Tong 1990), as we only require a Gaussian copula. Banerjee and Paul (2008) proved Proposition 1 for multivariate normal activity times. However, they ignored the possibility that activities might be serial, a case that we consider next.

## 3.3 Serial activities

Now suppose we increase the correlation between two activities that are *not* parallel. We ask: what will the effect on project completion time be? We can give a clear answer for a large subset of serial activities. Consider two serial activities, X and Y, and assume all other activity durations are given as parameters, then we can view the project makespan as a function of X and Y, T(X, Y). A pair of serial activities X and Y are called *decomposable* if T(X) can be written in the form f(g(X)+h(Y)) such that g(X) (h(Y)) does not depend on Y (X). For instance, consider the network in Figure 1. The makespan is given by: T() = A+max{B+E, B+D+F, C+F}+G. By observation of this expression it is clear that A and any other activity are serial and decomposable (f(A) = A and g(Y) = max{B+E, B+D+F, C+F}+G for any Y = B, …, G). A similar observation applies for G. Of the remaining five pairs of serial activities, (B, D), (B, E), (B, F), (C, F), and (D, F), we can write all except (B, F) in the required format. For instance, to show that for (B, D) and (B, E), we rewrite T() in the form T() = A+max{B+max{E, D+F}, C+F}+G, where B is decomposed from both D and E (but not from F). Similarly, if we write T() = A+max{B+E,

F+max{B+D, C}}+G we have decomposed C and D from F (but again, B and F are not decomposed).

An important observation is that if the project network is *series-parallel* then all serial activity pairs are decomposable. We formalize this claim below, but first, for completeness, we discuss series parallel networks in some detail (Baker and Trietsch, 2009). A network $N$ exhibits series-parallel structure if it consists of a single node or if $N$ can be partitioned into two sub-networks $N_1$ and $N_2$ which are themselves series-parallel and where one of the following conditions is satisfied:

- $N_1$ is in series with $N_2$ (for every pair $(i, j)$ with $i \in N_1$ and $j \in N_2$, we have $i \rightarrow j$), or
- $N_1$ is in parallel with $N_2$ (for every pair $(i, j)$ with $i \in N_1$ and $j \in N_2$, $i$ and $j$ are not related).

Furthermore, if $N$ is series-parallel we can construct for it a binary decomposition tree, with directed arcs and two kinds of nodes. Nodes without successors originally correspond to individual activities and for our purpose they store the time required to complete the activity (typically a random variable). All other nodes originally have two successors and correspond to a partition of a network or a sub-network. These decomposition nodes are designated *S* or *P*, depending on whether the appropriate partition for the two branches is series or parallel, respectively. Each activity is associated with a unique path from the root, but such paths may have arcs in common with paths of other activities. The level of a node is determined by the number of directed arcs on the path connecting it to the root. For every two distinct activities, we can identify a single ancestor node such that from that node onwards the two paths are distinct. If the activities are serial, that ancestor node is *S*; otherwise it is *P*. For instance, if we delete Activity D in Example 1, we obtain a series-parallel network whose decomposition tree is given in Figure 2.

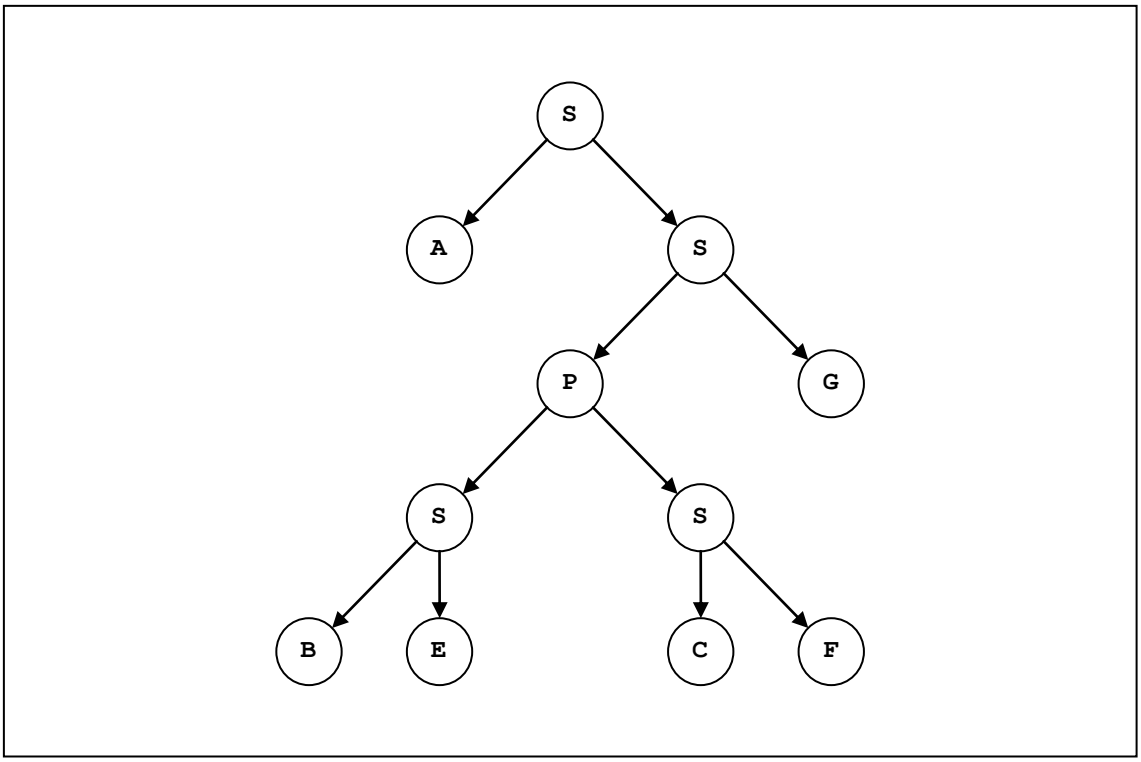


*Figure 2.* A Decomposition Tree

A decomposition tree can be folded by starting at the lowest level and combining all pairs of current leaves that share an immediate ancestor. If a leaf with a value *X* and a leaf with a value *Y* are currently at the lowest level and connected to the same immediate parent, then we remove them and if the parent is a *P* node we store in it the value max{*X*, *Y*} whereas if the node is *S*, we replace it by a leaf with the value *X*+*Y*. In Figure 2 the decomposition starts by replacing the S node connecting B and E by B+E and that of C and F by C+F, thus removing the lowest level. Then we replace the two new leafs by max{B+E, C+F}. In the next level we obtain a new leaf with the value max{B+E, C+F}+G. The final step is to combine the two remaining leaves to A+max{B+E, C+F}+G.

**Lemma 4** *The project completion time is an increasing convex function of each and every activity duration.*

**Proof:** When considered as a function of any activity, say X, the project completion time, T(X), is composed entirely of additions and max operations involving X and other activities as parame-

ters. Since the composition of increasing convex functions is an increasing convex function, the lemma follows **Q.E.D.**

**Lemma 5** *Let X and Y denote a pair of serial activities in a series-parallel project network. Then X and Y are decomposable. That is, the project completion time T(X,Y), regarded as a function of X and Y with the other activity durations treated as constants, can be written in the form T(X,Y) = f(g(X)+h(Y)) where f(.), g(.) and h(.) are increasing convex functions.*

**Proof:** We have to show that the activities are decomposable and that the decomposition involves increasing convex function. The first part of the lemma follows by constructing and then folding a decomposition tree for the series-parallel graph. Initially, store the values $g(X) = X$ at the leaf representing activity $X$ and $h(Y) = Y$ at the leaf representing $Y$. As we fold towards the common ancestor, $g(X)$ is updated by either taking the maximum of its current value and a constant or adding it to a constant. A similar observation applies to $h(Y)$. As the folding proceeds beyond the common ancestor, which is an *S* node and requires addition, we mark the result as $f(g(X) + h(Y))$. The second part follows by Lemma 4. **Q.E.D.**

**Proposition 2** *In a project network in which the activity durations have arbitrary marginal distributions and a Gaussian copula, the effect of increasing the correlation between any pair of decomposable serial activities is to increase project completion time in the increasing convex order.*

**Proof:** Let $W = (X,Y,Z_1,\ldots,Z_N)$ be a random vector of activity times with a Gaussian copula. Suppose $W' = (X',Y',Z_1,\ldots,Z_N)$ such that $X' =_{st} X$, $Y' =_{st} Y$, and $Cov(X,Y) < Cov(X',Y')$. Fix the durations of all the N activities other than X and Y to arbitrary values $t_1,\ldots,t_N$.

Since X and Y (respectively, X' and Y') are not part of a Z sub-network, it follows from Lemma 5 that the project completion time T(X,Y) – regarded as a function of X and Y with the other fixed activity times as parameters – takes the form

$$T(X,Y) = f(g(X)+h(Y)) \quad (1)$$

Where f(.), g(.), and h(.) are increasing convex functions.

We have, by Lemma 3:

$$(X',Y', t_1, \ldots,t_N| Z_i = t_i, i=1,\ldots,N) \geq_{sm} (X,Y, t_1, \ldots,t_N| Z_i = t_i, i=1,\ldots,N)$$

Since the supermodular order is closed under increasing transformations of components, we get

$$(h_1(X'), h_2(Y'), t_1, \ldots, t_N | Z_i = t_i, i=1,\ldots,N) \geq_{sm} (h_1(X), h_2(Y), t_1, \ldots, t_N | Z_i = t_i, i=1,\ldots,N)$$

where $h_1(.)$ and $h_2(.)$ are increasing functions.

Let g(.) denote any increasing convex function. Then by Fact 6 , together with (1) and (2) above, we obtain

$$g(h_1(X')+ h_2(Y')) | Z_1 = t_1, \ldots, Z_N = t_N \geq_{icx} g(h_1(X)+ h_2(Y)) | Z_1 = t_1, \ldots, Z_N = t_N$$

The proposition now follows from the closure of 'icx' ordering under mixtures. **Q.E.D.**

**Corollary 1:** *Suppose $S_1,\ldots,S_N$ are arbitrary (not necessarily series-parallel) sub-networks in series, and X and Y are activities belonging to $S_i$ and $S_j$, respectively. Suppose the activity distributions follow a Gaussian copula. Then the effect of increasing the correlation between X and Y, other things remaining unchanged, is to increase project completion time in the convex order. Hence, the variance of project completion time increases but its mean is unchanged.*

**Proof:** By Proposition 2, the project completion time increases in the increasing convex order. It is clear that the joint distribution of each $S_i$ is unchanged, and therefore $E(S_i)$ and the mean project completion time are unchanged. But 'icx' ordering is equivalent to 'cx' ordering when the random variables have the same mean. By a standard property of 'cx' ordering, the variance of project makespan increases. Q.E.D.

**The indeterminate impact of serial correlation for non-decomposable activities:** When X and Y are not decomposable, Proposition 2 does not apply. Consider, for instance, activities B and F in Example 1. The project makespan can be written in the form T = A+max{B+E, B+D+F, C+F}+G. When we increase the correlation between B and F, other things remaining unchanged, two opposing effects come into play. By Proposition 1, the term max{B+D+F, C+F} increases in the 'icx' order. By Proposition 1, the partial term max{B+E, C+F} decreases stochastically. When these factors interact, the net effect on T is indeterminate. Intuitively, the net effect depends on which elements in the makespan expression are likely to be critical. We verified this intuition by numerical experiments. For instance, initially let the means of B, D, and F be 1 and the means of C and E, 2. Let the variances of all activities be equal (e.g., 1). All three possible critical paths have the same mean, so they are all likely. If we now increase E(D) gradually, without changing its variance, the probability path A-B-D-F-G is critical increases and eventually we find that increasing the correlation of B and F increases project completion time in the icx

sense. If, instead, we gradually increase E(E) and E(C), path A-B-D-F-G is progressively less likely to be critical and the net effect of increasing the correlation between B and F will approximate a stochastic decrease of project completion time.

## 3.4. Projects with linear association among activities

We now discuss another model of positive correlation between project activities, and obtain a sensitivity analysis result for it. We say that a set of n positive random variables, $\{Y_j\}$, are *linearly associated* if $Y_j = QX_j$ where $\{X_j\}$ is a set of n independent positive random variables and Q is a positive random variable, independent of $\{X_j\}$. We refer to Q as the *bias element* and by multiplying the activities by Q we are performing a *bias element perturbation*. If we perform bias element perturbation on activity X, we obtain $V(QX) = V(X)\cdot E(Q^2) + V(Q)\cdot(E(X))^2$. For two distinct linearly associated activities, R and S, it can be shown that $COV(QR, QS) = V(Q)\cdot E(R)\cdot E(S)$.

It is natural to set E(Q) =1 in the linear association model so bias element perturbation preserves mean activity durations in that case. By Fact 5, if E(Q) = 1 then bias element perturbation leads to a stochastic increase in the sense of the convex order.

Trietsch et al. (2012) show that linear association can explain field data obtained for several project organizations. In principle, a project can be influenced by several bias elements, each applicable to a subset of activities. Hence we assume that there are N project activities partitioned into subsets $S_i$ with $n_i$ elements such that $S_i$ has bias element $Q_i$ with $E(Q_i) = 1$. Suppose all the activity durations are mutually independent before the bias element perturbation, and that the $Q_i$ are mutually independent.

**Lemma 6** *Suppose* $X \leq_{cx} Y$ *and Z is a positive random variable independent of X and of Y. Then* $ZX \leq_{cx} ZY$.

**Proof:** Since Z is independent of X and Y and the convex order is closed under mixtures (Shaked and Shanthikumar, Theorem 3.A.12.(b)), it is sufficient to show that $kX \leq_{cx} kY$ for any positive k. Now $kX \leq_{cx} kY$ because g(kX) is a convex function of kX iff g(X) is a convex function of X. **Q.E.D.**

**Proposition 3** *Suppose the project activities are subject to bias element perturbation with bias elements $Q_i$ such that $E(Q_i) = 1$ for all i. Suppose the project network is series-parallel with respect to $S_i$. Then the project completion time increases in the increasing convex order.*

**Proof:** If we apply bias element perturbation $Q_i$ to all the activities in subset $S_i$, then it is clear that that the durations of all activities in the subset are distributed per their unperturbed distributions multiplied by $Q_i$. Hence, by Lemma 6, the duration of each activity increases in the 'cx' order. But the project completion time is an increasing convex function of $S_i$, and the $S_i$ are mutually independent. Hence the project completion time increases in the 'icx' order. **Q.E.D.**

.

**Proposition 4.**

*Suppose $Q_i = Q$ for all i. Then the project completion time increases in the convex order with a bias element perturbation*.

**Proof:** In this case the project makespan changes from T to QT. By Lemma 6, we have

$QT \geq_{cx} T$, and the lemma follows. Q.E.D.

So if all the project activities share a common bias element perturbation, then the impact is an increase in the variance of project completion time.

## 3.5. On the use of the lognormal distribution for activity times

Because the lognormal only admits positive values and has a right skew, it has been one of the optional distributions for project activity durations for a long time (as evidenced by its availability on commercial project scheduling packages). It is also the only distribution we know of that has been validated for field data (including a recent validation by Trietsch et al., 2012). Notice that the lognormal multivariate distribution is guaranteed to possess a Gaussian copula, which makes our results applicable for it.

One explanation for the efficacy of the lognormal in practice is that activity durations often comprise many additive components (when activities are composed of serial sub-activities) as well as multiplicative components. The lognormal distribution with matched mean and variance has been used to approximate the sum of lognormal random variables (Fenton, 1960). Past work (e.g., Trietsch and Baker 2012) relied on the use of the lognormal distribution as a good approximation for the sum of many independent positive valued random variables (lognormal or not). When we appeal to the central limit theorem to approximate the sum of many positive val-

ued random variables, we note that the coefficient of variation of the limiting normal distribution must be low. For instance, the coefficient of variation of the sum of N iid random variables with mean μ and standard deviation σ is $\frac{\sigma}{\mu\sqrt{N}}$ , and as N increases the coefficient of variation goes to zero in the limit. Baker and Trietsch (2009) proposed the idea of exploiting this fact, by replacing the normal distribution by a lognormal distribution when approximating the sum of a large number of positive random variables. They used the term 'lognormal central limit theorem' for this approximation but the term is misleading, since no normalized sum of independent random variables converges to a lognormal distribution. The precise sense in which the approximation holds is clarified by Lemma 7 below. The lemma shows that at sufficiently small positive values of coefficient of variation, the normal distribution can be approximated very closely by a class of two-parameter distributions to which the lognormal distribution belongs.

**Lemma 7** *Let $M > 0$ be a given number. Let $s_1,s_2,\ldots$ be an infinite sequence of strictly positive numbers converging to zero. Let $X_1,X_2,\ldots$ be an infinite sequence of normally distributed random variables such that $X_i$ has mean M, standard deviation $s_i$, and distribution function $F_i$. Let $Y_1,Y_2,\ldots$ be an infinite sequence of random variables such that*

*(a)$Y_i$ has mean M and standard deviation $s_i$, and a continuous distribution function $G_i$*

*(b) Each $Y_i$ has a distribution in which the mean and standard deviation are parameters of the distribution that can assume any non-negative value, independently of each other.*

*Then given any $\varepsilon>0$, there exists a positive integer $n(\varepsilon)$ such that*

$$sup_x\{|F_{n(\varepsilon)}(x) - G_{n(\varepsilon)}(x)|\} < \varepsilon$$

**Proof:** Both $Y_i$ and $X_i$ converge in distribution to the degenerate random variable that takes the value M with probability 1. It follows that the alternating sequence $X_1,Y_1,X_2,Y_2, \ldots$ converges in distribution to the degenerate random variable that takes the value M with probability 1.

Hence we can always find $Y_n$ and $X_n$ whose distribution functions are arbitrarily close in the Levy metric (by the theorem that convergence in the sense of the Levy metric is equivalent to weak convergence to a proper distribution function (Theorem VII.5.7, Pestman (1998)). Since $G_i(x)$ is assumed to be continuous for all i and $F_i(x)$ is continuous for all i by definition, it is im-

mediate from the proof of Theorem VII.5.6 in Pestman (1998) that there exists δ>0 such that if the Levy distance between $G_n(x)$ and $F_n(x)$ is smaller than δ, then $sup_x\{|F_n(x) - G_n(x)|\} < \varepsilon$. But since the sequence $X_1, Y_1, X_2, Y_2, \ldots$ converges in the Levy metric, we can choose n large enough that the Levy distance between $G_n(x)$ and $F_n(x)$ is smaller than δ. The lemma follows Q.E.D.

**Remark:** The lognormal distribution satisfies the conditions required of $Y_i$. Whereas other common distributions (e.g., gamma and Weibull) also satisfy those conditions, the lognormal is distinguished by possessing two additional desired features together: its density function at the origin is zero and its coefficient of variation is not bounded. Most distributions that satisfy one of these conditions violate the other.

If we ensure that the coefficient of variation is not more than $\sqrt{8\pi\varepsilon}$, we find by numerical experimentation that a lognormal and a normal distribution are ε-close; for instance, for ε = 0.05 the coefficient of variation must not exceed 0.255.

# 4. Discussion and concluding remarks

The impact of increasing correlation is sensitive to project topology. Assuming that activity marginal distributions have a Gaussian copula and holding all marginal distributions constant, increasing the correlation between parallel activities decreases project duration stochastically. This implies not only that mean project duration decreases, but that the probability of finishing the project within a given deadline increases.

However, under the same assumptions, the impact of increasing the correlation between serial activities in a project is usually an increase in project completion time in the icx sense; i.e., we obtain an increase in either the mean or the variance of project completion time. This effect is guaranteed for serial activities within serial parallel networks or when the two activities belong to two serial sub-networks. In more detail, when two activities are serial there is at least one path that connects them. In the presence of other paths in parallel, the effect is an increase in the mean. When the serial activities belong to two distinct serial sub-networks (such that all paths must visit at least one activity in each sub-network), the mean is unchanged but the variance increases. This result is intuitively clear for a serial project, where both activities are guaranteed to

be critical, but we showed that also holds for serial activities that are not guaranteed to be critical. When we consider increasing the correlation between serial activities in non-series-parallel networks, the effect of an increase in correlation is sometimes indeterminate, as the two opposing effects of parallel and serial cases can be at play together. We demonstrated this by an example where the two serial activities are connected by a bridge that renders the network non-series-parallel. However, increasing the correlation for all other pairs of serial activities in the same non-serial-parallel network leads to a clear icx increase effect.

In the linear association model of activity correlation, correlated activities share a common bias element. We note that under this model the assumption that marginal activity duration distributions remain unchanged does not hold. If the correlation between two activities is increased due to an increase in the variation of the common bias element, we obtain an icx increase for any project network and any subset of affected activities (although the magnitude of the change may depend on network topology). In some instances, when the bias element applies for all project activities or for a complete sub-network in a serial sub-network structure, the result is an increase in variance only (i.e., an increase in the cx sense, which is a special case of an increase in the icx sense). In other instances of an increase in the variance of a common bias element that applies to a subset of activities, the mean must increase. Because we obtain an icx increase even when the subset involves parallel activities, the result apparently contradicts the conclusion we reached for parallel activities before, but the two observations stem from different ways of modeling a change in correlation between project components; the linear association model posits that one cannot isolate changes in correlation between project sub-networks from their marginal distributions whereas the competing model permits such isolation.

There is empirical evidence that, as a rule, project activities are correlated and that the linear association model can explain such correlation. Nonetheless, the practical importance of the concept is mainly the need to take it into account. By contrast, our result for parallel activities (Proposition 1) has a direct relevance to managerial decisions. Let us take a specific example of a construction project that consists of two subprojects in parallel – repaving a country road in Los Angeles county, and in adjacent Orange county in California. Assuming qualified contractors are capable of executing both subprojects in parallel, should they be packaged independently and put out to bid as separate projects or as a single project? If we bid them out separately, we increase the chances that independent construction companies are awarded the contracts. Otherwise, we

ensure that one construction company manages both parallel subprojects. On the one hand, it is highly likely that any two contractors would be subject to non-identical sets of common bias causes, and thus they will each exhibit positive correlation between activities they execute. But on the other hand, there is no reason to assume that their marginal distributions are different. Hence, even though we believe that the correlation structure in each company can be modeled by linear association, Proposition 1 remains appropriate. In this case, assuming that the winning contractors all have the same marginal distributions of project completion time for each subproject, we would recommend putting the work out to bid as a single project (applying Proposition 1).

# Acknowledgments

We would like to thank Professor Alfred Muller for suggesting the use of supermodular stochastic ordering. Many crucial technical facts and lemmas that we use stem from his work on supermodular ordering.

# References

Banerjee, A., Paul, A. 2008. On path correlation and PERT bias. *European Journal of Operational Research*, 189 (2008), 1208-1216.

Baker, K.R., Trietsch, D. 2009. *Principles of Sequencing and Scheduling*. Wiley, Hoboken.

Corbett, C.A., Rajaram, K. 2006. A generalization of the inventory pooling effect to nonnormal dependent demand. *Manufacturing & Service Operations Management* Vol 8 No.4, 351–358.

Elmaghraby, S.E. 2000. On criticality and sensitivity in project networks. *European Journal of Operational Research*, **127** 220-238.

Fenton, L.F. 1960. The sum of lognormal probability distributions in scatter transmission systems. *IRE Transactions on communication systems*. **8** 57-67.

MacCrimmon, K.R., Ryavec, C.A. 1964. An analytical study of the PERT assumptions. *Operations Research*. vol. 12 no. 1 16-37

Muller, A., Scarsini, M. 2000. Some remarks on the supermodular order. *Journal of Multivariate Analysis*. **73** 107-119.

Muller, A., Stoyan, D. 2002. Comparison methods for stochastic models and risks. Wiley, England.

Pestman, W.R. 1998. *Mathematical Statistics*. Walter de Gruyter, Berlin.

Shaked, M., Shanthikumar, J.G. 2007. *Stochastic Orders and their applications.* Academic Press, Boston.

Tong, Y.L. 1990. *The multivariate normal distribution*. Springer-Verlag, New York.

Trietsch, D., Baker, K.R. 2012 PERT 21: Fitting PERT/CPM for Use in the 21st Century. *International Journal of Project Management* **30**, 490–502.

Trietsch, D. Mazmanyan, L., Gevorgyan, L., Baker, K.R. 2012. Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation. *European Journal of Operational Research* **216** 386-396.