Letter

Tropical approximation to finish time of activity networks

Alexei Vazquez [*]
Nodes & Links Ltd., Salisbury House, Station Road, Cambridge CB1 2LA, United Kingdom

(Received 9 March 2022; accepted 23 June 2022; published 14 July 2022)

We breakdown complex projects into activities and their logical dependencies. We estimate the project finish time based on the activity durations and relations. However, adverse events trigger delay cascades shifting the finish time. Here I derive a tropical algebraic equation for the finish time of activity networks, encapsulating the principle of linear superposition of exogenous perturbations in the tropical sense. From the tropical algebraic equation I derive the finish time distribution with explicit reference to the distribution of exogenous delays and the network topology and geometry.

DOI: 10.1103/PhysRevE.106.L012301

In recent years we have experienced a significant advance in our understanding of complex networks and of processes running on them [1–3]. Yet, we still lack an understanding of the interplay between network topology and dynamics in the context of activity networks. There is a vast literature on Monte Carlo simulations of activity networks under uncertainty and risk events [4,5]. The critical path method has been used to aggregate perturbations and derive an analytical approximation to the project end date distribution [6]. However, activity networks are characterized by a complex topology [7–9], and activity delays exhibit a high frequency of extreme events [9–11]. In that context the key conditions for the critical path method, the existence of a dominant path from the project start to its end and the central limit theorem do not apply. Here I obtain a tropical algebra approximation to the project end date.

Let $P(V,E,\vec{d})$ be a project schedule with a set of activities V, a set of activity relations E, and a vector of activity durations $\vec{d}$. The project start and end are represented by the activities $i=1$ and $i=n$, respectively, of duration zero. An arc $i\to j\in E$ indicates that i must finish before j starts. I will denote the set of activity predecessors of i by $I_i=\{j|j\to i\in E\}$. Logical consistence implies that the project network is a directed acyclic graph. The nodes in a directed acyclic graph can be ordered topologically such that if $i\to j$ then i is before j in the topological order for all $i\to j\in E$.

The earliest an activity can finish is determined by the recursive relation,

$$x_i = d_i + \max_{j\in I_i} x_j, \tag{1}$$

with the boundary condition $x_1=s$, where s is the start date. We calculate $\vec{x}$ with a forward pass of Eq. (1) along the topological order. Next we perform backward propagation to calculate the latest an activity can finish without altering the project end date,

$$y_i = \min_{j|i\in I_j}(y_j - d_j), \tag{2}$$

with the boundary condition $y_n=x_n$. $\vec{x}$ and $\vec{y}$ set the early and late start dates for every activity. If $i\in I_j$ then,

$$w_{ji} = x_j - x_i \tag{3}$$

is the free float, the maximum delay at i that does not shift the early finish date of j [6]. In turn, we can determine the amount of delay tolerated at a given activity without causing a shift in the project end date. This is the total float, and it is calculated as

$$T_{ni} = x_i - y_i. \tag{4}$$

The subscript n emphasizes that T_{ni} is the total float from activity i to the project end. This will be generalized below to any pair of activities. I have adopted a negative sign for free floats and total floats to indicate delay subtraction.

In practice, exogenous factors delay the activity starts or increases activity durations [11]. If the finish delay of an activity exceeds the free float of any successor, then it will cause their delay as well, starting a delay cascade [12,13]. Let $\vec{h}$ represent the vector of activity delays caused by exogenous factors, and $\vec{z}$ denote the vector of activity delays after the propagation of the exogenous delays. The delays propagate via the recursive equation,

$$z_i = f_i\left\{\max_{j\in I_i}[\max(0, w_{ij}+z_j)], h_i\right\}. \tag{5}$$

The term $\max(0,w_{ij}+z_j)$ indicates that delays are passed if they exceed the free float between the activities. Then we take the maximum delay from $j\in I_i$. The function $f_i(x,y)$ merges the delays coming from predecessors (endogenous) and exogenous sources. If exogenous delays act on activity starts, then $f_i(x,y)=\max(x,y)$. If exogenous delays act on activity durations, then $f_i(x,y)=x+y$. Even a unique contingency, such as an adverse weather event can act differently

[*]alexei@nodeslinks.com

depending on where it falls in the calendar relative to an activity. If it falls before the activity starts, it competes with delays from predecessors as a cause of delay for the activity start. If the adverse weather event happens whereas the activity is ongoing it will delay its finish date on top of any delay coming from predecessor activities. In general,

$$\max(x, y) \leqslant f_i(x, y) \leqslant x + y. \tag{6}$$

To be precise we should specify $f_i(x, y)$ for every combination of activity and adverse event. Yet, when x and y follow subexponential distributions, $\max(x, y) \approx x + y$. If that is the case, then $f_i(x, y) \approx \max(x, y) \approx x + y$.

Subexponential distributions cannot be bounded by any exponential function $f(x) = e^{-\alpha x}$ when $x \to \infty$ [14]. If we take two random variables x and y generated from the same subexponential distribution then $\Pr[x + y > z] \sim \Pr[\max(x, y) > z]$ when $z \to \infty$. The only way that $x + y$ is larger than z is for x or y to be greater than z. We cannot take this result as granted when we propagate delays in the activity network. The errors may accumulate when the delay propagation takes several steps and free floats act as delay sinks. In the following I investigate the difference in using $f = \max$ or $f = \text{sum}$ by means of numerical simulations.

We need a model to generate project networks, a model to generate activity durations, and a model to generate exogenous delays. I will generate project networks using the duplication-split model [15]. In this model, we start with two activities representing the project start and end with an arc from start to end. Then, at each discrete step, we select an activity i with uniform probability across all current activities and create a new activity j. With probability q, j is a duplicate of i, inheriting all the incoming and outgoing relations of i. Otherwise, i transfers all the outgoing arcs to j and an arc from i to j is created. We call q the duplicate rate. It is demonstrated that both the in-degree and out-degree distribution has a power-law tail with exponent $1/q$ and the network diameter grows as n^{1-q} [15]. As we tune q from 0 to 1 I change from networks that are quasilinear and have narrow degree distributions to networks with wide degree distributions and multiple parallel paths.

The simulations proceed as follow. The inputs are the number of activities n, the duplication rate q, the distributions of activity durations $p(d)$, and the exogenous delay distribution $p(h)$. I focus on subexponentially distributed exogenous delays, so I choose the log-normal $p(h) \sim e^{-[\ln(h/\mu)]^2/2\sigma^2}/h$ with $\mu = 1$. First, I generate a duplication-split network with parameters (n, q). Second, I assign durations to activities from $p(d)$ and run Eqs. (1)–(3) to generate the free floats w. Third, I generate random delays from $p(h)$ and run Eq. (5) with a predefined $f(x, y)$ to calculate z_n, the project end delay. I repeat this third step to generate z_n samples and report the smallest z_n that is larger than 80% of all samples. This quantity is commonly used in project management, and it is named project $p80$. Fourth, for each parameter set $\{n, q, p(d), p(h), f(x, y)\}_i$, I generate multiple $p80$ values $\pi(f)_i$ by sampling over realizations of the network and activity durations. I do that for $f = \max$ and $f = \text{sum}$. Finally, I calculate the slope through the origin,

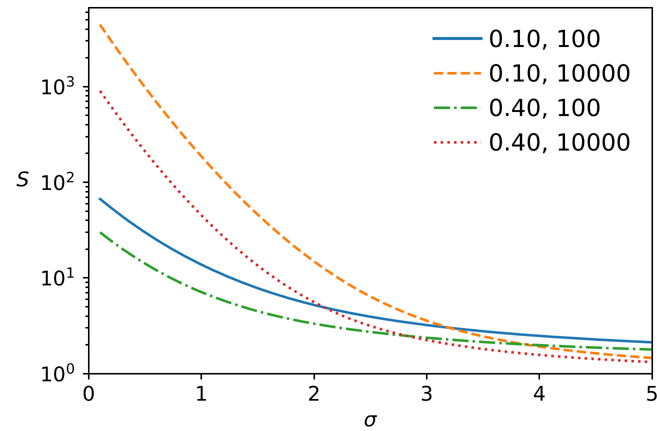

FIG. 1. Slope between the calculated $p80$s using $f = \text{sum}$ vs using $f = \max$ for $\vec{d} = \vec{0}$ and (q, n) indicated in the legend.

$$S = \frac{\sum_i \pi(f = \max)_i \pi(f = \text{sum})_i}{\sum_i [\pi(f = \max)_i]^2}. \tag{7}$$

Since $x + y \geqslant \max(x, y)$ then $S \geqslant 1$.

If we set activity durations to zero I can investigate max vs sum without the additional complication of free floats. Given the log-normal $p(h) \sim e^{-[\ln(h/\mu)]^2/2\sigma^2}/h$, I expect $S \to 1$ when $\sigma \to \infty$. This asymptotic behavior is corroborated in Fig. 1. However, the convergence is very slow. We need $\sigma > 3$ to attain $S < 2$. The variance of the log-normal distribution is order e^{σ^2}. For example, $\sigma = 3$ implies a variance on the order of $e^9 \approx 8000$. I can conclude that, if the activity durations are homogeneous, then $f = \max$ is a poor lower bound for $f = \text{sum}$. A similar plot is obtained using $p(d) \sim e^{-d/\mu_1}$ (data not shown). The same is expected for any exponentially bound distribution of activity durations.

The picture changes when the activity durations follow a subexponential distribution as well. For example, a log-normal distribution $p(d) \sim e^{-[\ln(d/\mu_1)]^2/2\sigma_1^2}/d$. I will set $\mu_1 = 1$ since σ_1 controls the shape of the distribution tail. Setting $\sigma_1 = 1$ I obtain the plot in Fig. 2. For $\sigma < \sigma_1 = 1$ the value of S is close or below 2. In this context $f = \max$ is a good approximation for $f = \text{sum}$. As σ approaches 1 the value of S reaches a maximum and decays towards $S = 1$. I observe the same behavior for $\sigma_1 = 3$ (Fig. 3). It becomes evident that there is local S maximum at $\sigma = \sigma_1$, i.e., when the exogenous delays and activity durations have the same log-variance. It is also evident that the maximum value of S decreases with increasing σ_1.

Regarding the network parameters for a given number of activities, S is closer to 1 for $q = 0.4$ than $q = 0.1$. The networks with larger duplication index q have a broader degree distribution and a smaller diameter. In contrast, networks with small q tend to be closer to a linear chain. This observation suggests that the distinction between using $f = \max$ or $f = \text{sum}$ is less relevant in complex networks with wide degree distributions and smaller diameters.

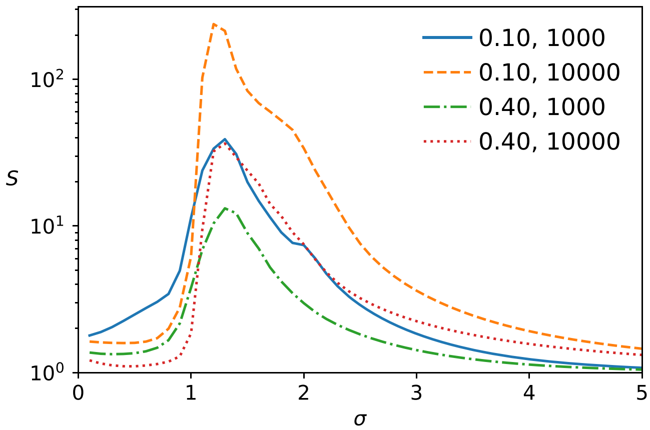

FIG. 2. Slope between the calculated $p80$s using $f = \text{sum}$ vs using $f = \max$ for $\sigma_1 = 1$ and (q, n) indicated in the legend.

I have estimated σ and σ_1 for construction projects in the Nodes & Links database. For each project schedule in the warehouse I estimated σ as the variance of the logarithm of reported delays (actual finish date–planned finish date). I estimated σ_1 as the variance of the logarithm of activity durations. In 78% of projects $\sigma < \sigma_1$ [Fig. 4(a)]. For these projects, I can use either $f = \max$ or $f = \text{sum}$ and obtain similar results. To test this expectation I carried on Monte Carlo simulations on top of real construction projects Fig. [4(b)]. The slope of the forecasted $p80$s using $f = \text{sum}$ vs $f = \max$ is 1.6 in the range of what observed for synthetic networks in the region $\sigma < \sigma_1$ (3).

If using $f = \max(x, y)$ gives similar results as using $f = \text{sum}$ then we can choose either. As shown below, using $f = \max$ has some advantages. For $f = \max$ I solve Eq. (5) iteratively,

$$z_{i,t+1} = \max\left\{\max_{j\in I_i}[\max(0, w_{ij} + z_{j,t})]h_i\right\}, \tag{8}$$

Defining $w_{ii} = 0$ I rewrite this equation as

$$z_{i,t+1} = \max_{j|j\in I_i\cup\{i\}}(w_{ij} + z_{j,t}). \tag{9}$$

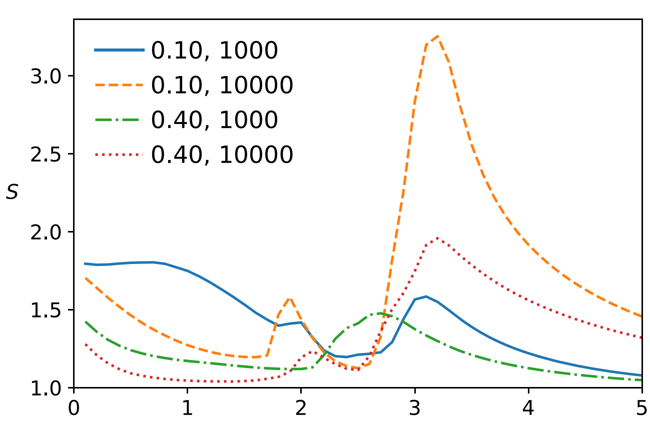

FIG. 3. Slope between the calculated $p80$s using $f = \text{sum}$ vs using $f = \max$ for $\sigma_1 = 3$ and (q, n) indicated in the legend.

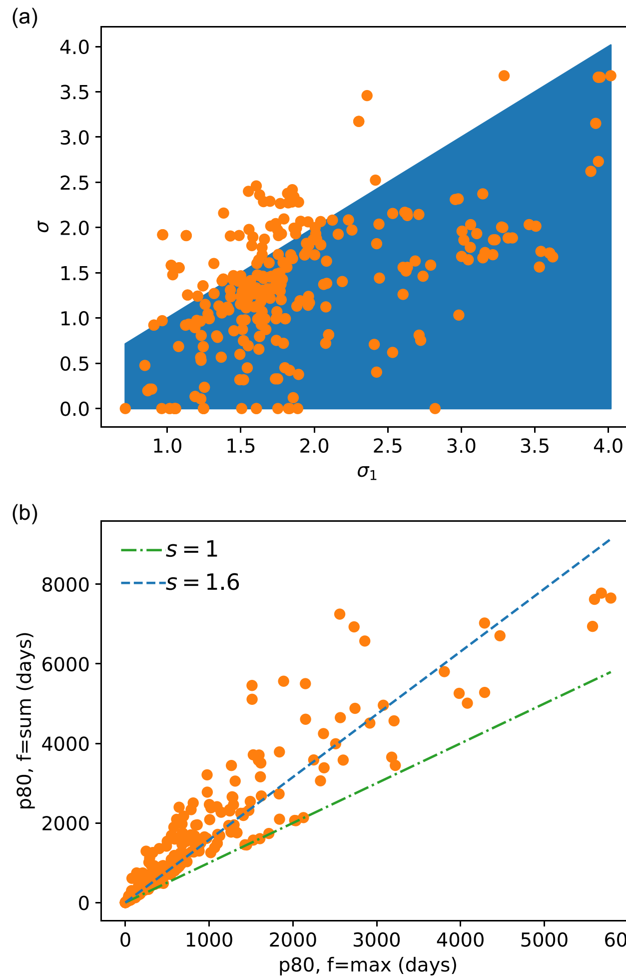

FIG. 4. (a) Plot of σ vs σ_1 for construction projects. Each symbol represents a construction project schedule. The shaded background highlights the area where $\sigma < \sigma_1$. (b) $f = \text{sum}$ vs $f = \max$ $p80$ forecast for real construction projects (circles). The dashed line represents equal values. The dashed-dot line represents the best linear fit through the origin. The activity network and activity durations are from the real projects. The parameters of $p(h) \sim e^{-[\ln(h/\mu)]^2/2\sigma^2}/h$ are project specific and inferred from observed delays for finished activities [see σ in panel (a)].

Replacing max by $\oplus$ and $+$ by $\otimes$ the tropical algebra becomes evident. The tropical algebra is defined by the semiring $(\mathbf{R} \cup \{-\infty\}, \oplus, \otimes)$ with the sum operation defined as $x \oplus y = \max(x, y)$ and the product as $x \otimes y = x + y$ [16]. The tropical algebra has been used to investigate networks with cycles [17]. For example, of transportation networks where it is desired to have a route back and forward between any two nodes. The tropical algebra has been applied to many other systems that can be mapped to event networks, including mRNA translation by ribosomes [18] and Conway's game of life [19]. Using the tropical algebra I rewrite (9) as

$$\vec{z}_{t+1} = L \otimes \vec{z}_t, \tag{10}$$

where

$$L = \begin{cases} 0, & \text{if } i = j,\\ w_{ij}, & \text{if } j \in I_i,\\ -\infty, & \text{otherwise} \end{cases} \tag{11}$$

is the local weights matrix. L encodes the free floats of direct relations, $L^{\otimes 2}$ the maximum free float sum of paths up to length 2 and $L^{\otimes l}$ is the maximum free float sum of paths up to length l. Since project networks are represented by directed acyclic graphs, there is no path larger than the length D of the longest acyclic path. Therefore, $L^{\otimes l} = L^{\otimes D}$ for $l \geqslant D$. After D iterations starting from $\vec{z}_0 = \vec{h}$ the system will reach the steady state solution,

$$\vec{z}_\infty = T \otimes \vec{h}, \tag{12}$$

where

$$T = L^{\otimes D}, \tag{13}$$

is the total float matrix. The element $-T_{ij}$ equals the maximum h_j that yields $z_{i,\infty} = 0$.

Equation (12) represents a principle of statistical independence of exogenous delays in the tropical algebra sense. Each exogenous delay contributes independently of the others to the final propagated delay. This property is a direct consequence of the $f = \max$ merging function. Calculating the cumulative distribution function for the delay at each activity is now straightforward. Since exogenous delays are independent I calculate the maximum among the delay cascades propagated from each exogenous delay. If $F_i(h) = \Pr(h_i \leqslant h)$, $G_i(z) = \Pr(z_{i,\infty} \leqslant z)$, and x_i is the early start date, then from Eq. (12) I obtain the probability distributions for the activity end dates,

$$G_i(y) = \prod_j F_j(y - x_i - T_{ij}). \tag{14}$$

This equation represents an analytical approximation to the end date distribution of all activities in a project network. In this equation we have a clear separation among the influence of the distributions of exogenous delays $F_j(h)$, the planned end date x_i, and the activity network properties encoded in the total floats T_{ij}. From there we can calculate other relevant quantities. The probability to finish on time $G_i(0)$ and end date $y_i(p)$ with p confidence $G_i[y_i(p)] = p$. We can determine the impact of accelerating activity execution to catchup with delays as well by taking into account that $T_{ij} = T_{ij}(\vec{d})$.

The nontrivial region where $f = \max$ is a good approximation to $f = \text{sum}$ remains to be explained. Intuitively, if both the delays and the free floats have subexponential distributions, and we extract a delay and a free float from those distributions, then one of them will be much larger than the other with high probability. In that case, the delays that exceed the free floats will be large, and $f = \max$ is a good approximation for $f = \text{sum}$. More work is required to translate this intuition into analytical demonstrations. Furthermore, a delay in the activity start date could increase the delay in the activity finish date, above what was expected if the activity start was not delayed (superadditive). If the rate per activity of such events is large, then we should expect a breakdown of the tropical approximation.

In conclusion, I have obtained a tropical approximation to the end date distribution of activity networks. This approximation replaces sum by max in the algebraic equations that merge endogenous with exogenous delays. The tropical approximation is close to simulated data when: (i) both the distribution of exogenous delays and of activity durations belong to the subexponential family, and (ii) the variance of the exogenous delays is smaller than the variance of the activity durations. These conditions are satisfied in 78% of projects within the Nodes & Links construction project database.

I thank C. Ellinas and G. Kalogridis for the Letter comments. This work was partly supported by European Union's Horizon 2020 and the Cyprus Research & Innovation Foundation under the SEED Program (Grant Agreement No. 0719(B)/0124). Nodes & Links provided support in the form of salary for A.V. but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the Letter.

[1] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[2] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Rev. Mod. Phys. **80**, 1275 (2008).

[3] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[4] F. Acebes, M. Pereda, D. Poza, J. Pajares, and J. M. Galán, Int. J. Project Manage. **33**, 1597 (2015).

[5] Y. Moret and H. H. Einstein, J. Construction Eng. Manage. **142**, 05016010 (2016).

[6] M. Vanhoucke, *Project Management with Dynamic Scheduling: Baseline Scheduling, Risk Analysis and Project Control* (Springer, Berlin/Heidelberg, 2013)

[7] D. Braha and Y. Bar-Yam, Phys. Rev. E **69**, 016113 (2004).

[8] D. Braha and Y. Bar-Yam, Manag. Sci. **53**, 1127 (2007).

[9] I. Pozzana, C. Ellinas, G. Kalogridis, and K. Sakellariou, Appl. Network Sci. **6**, 25 (2021).

[10] C. Araújo-Rey and M. A. Sebastián, Sustainability **13**, 3975 (2021).

[11] J. E. Park, Tran. Rev. **41**, 814 (2021).

[12] D. Braha and Y. Bar-Yam, in *Smart Business Networks*, edited by P. L. P. K. Vervest P., van Heck E. (Springer, Berlin/Heidelberg, 2005).

[13] C. Ellinas, Prod. Oper. Manage. **28**, 63 (2019).

[14] S. Foss, D. Korshunov, and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions*, Springer Series in Operations Research and Financial Engineering (Springer, New York, 2013).

[15] A. Vazquez, I. Pozzana, G. Kalogridis, and C. Ellinas, arXiv:2207.04688.

[16] P. Butkovič, *Max-Linear Systems: Theory and Algorithms*, Springer Monographs in Mathematics (Springer, London, 2010).

[17] B. Heidergott, G. Olsder, and J. van der Woude, *Max Plus at Work: Modeling and Analysis of Synchronized Systems: A Course on Max-Plus Algebra and Its Applications*, Princeton Series in Applied Mathematics (Princeton University Press, Princeton, 2014).

[18] C. A. Brackley, D. S. Broomhead, M. C. Romano, and M. Thiel, J. Theor. Biol. **303**, 128 (2012).

[19] K. Sakata, Y. Tanaka, and D. Takahashi, Complex Syst. **29**, 63 (2020).