Estimating Durations and Trials to Success in Test Programs

- Click to edit Master text styles

Danny R. Hughes

LMI; University of South Alabama

Jeremy M. Eckhause

LMI

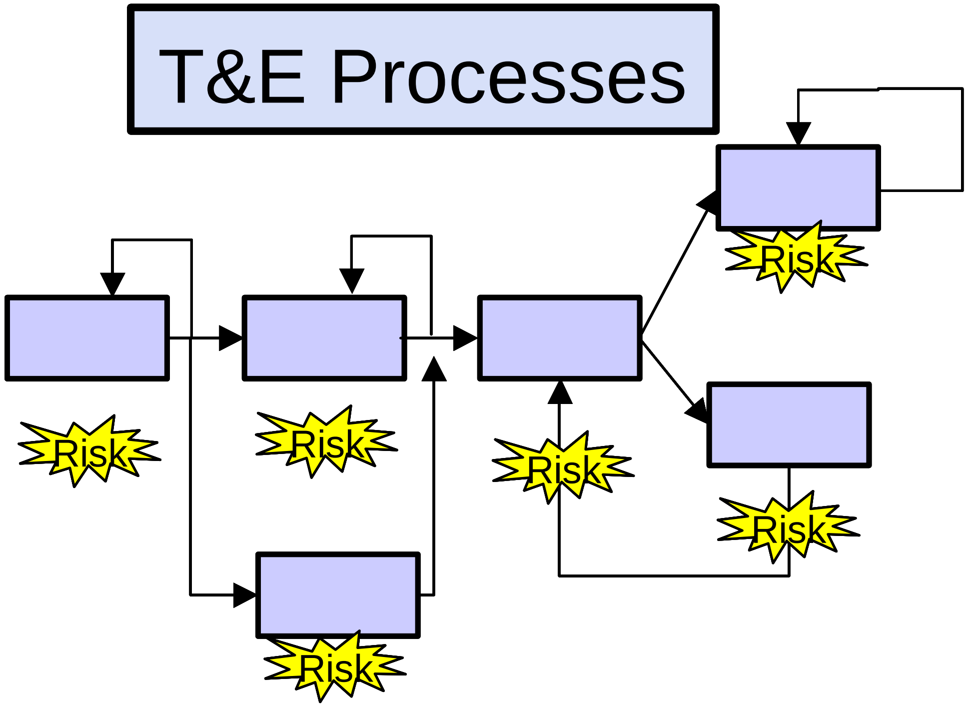
Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

Traditional T&E Schedule Estimation



- T&E programs are inherently risky:
  - Individual WBS elements carry considerable schedule risk
  - There are complex relationships between test objectives, outcomes, and future work
  - Each outcome has complex risks and consequences
  - Not intuitive; difficult to scope

Traditional T&E Schedule Estimation

- T&E schedules are estimated in variety of ways
  - Depends on time, data, precision needed, guidance from program office
- Traditionally, 3 ways to estimate T&E schedules:
  - Factors based upon historical data from analogous systems
  - Parametric Schedule Estimating Relationships (SERs)
    - Linking some characteristic system parameter to historical schedules
  - Detailed bottom-up estimates
- Predominantly driven by projected staffing requirements
- Usually assumes only planned tests

Traditional methods do not account for stochastic events and feedback loops resulting from the recovery from failure

Traditional T&E Schedule Estimation

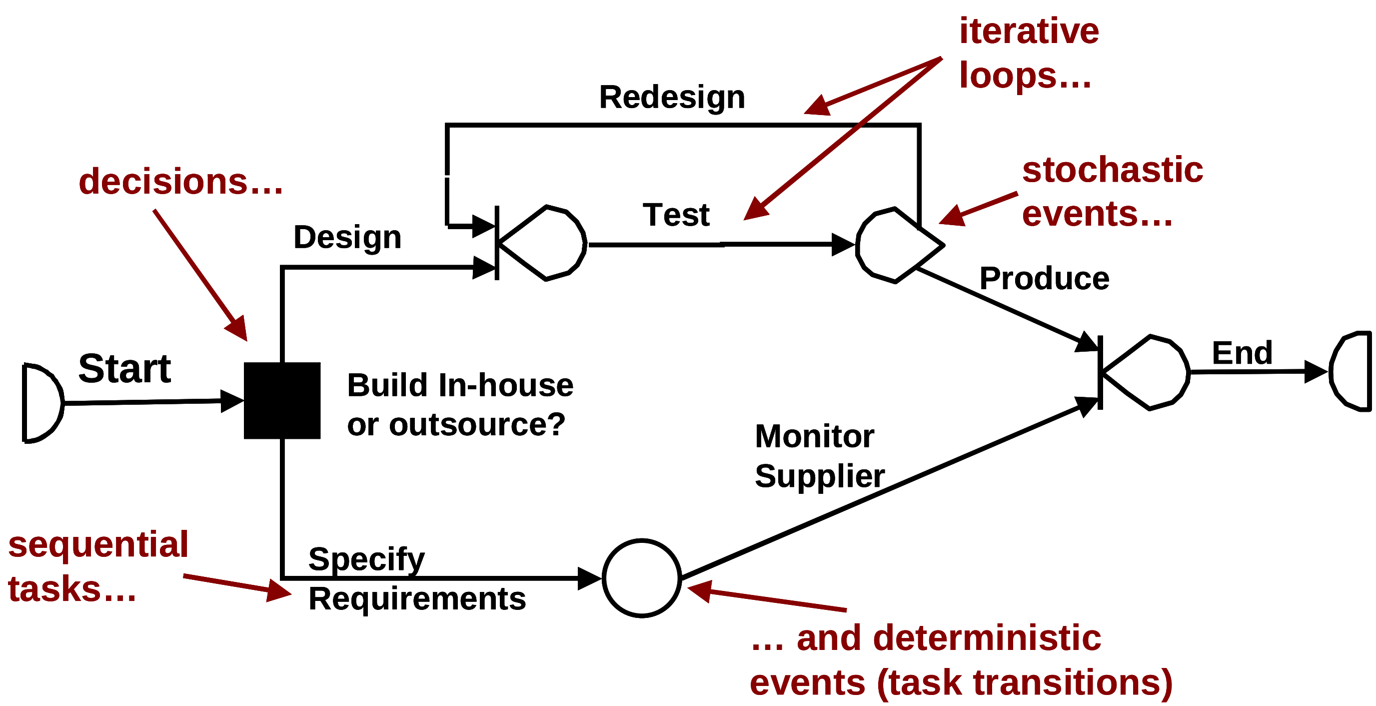
- Accounting for risk and unknowns is a historical challenge for T&E cost estimation
- Use of historical analogies often fail to adequately account for important distinctions in the new system
- History-based parametric analysis can reasonably estimate the T&E schedules, but provide no information about critical test elements
- Bottom-up estimates provide a wealth of detail on individual test elements
  - Doesn’t account for additional unplanned tests resulting from test failure
  - Schedules are consistently inaccurate and always low
- Generalized Activity Network (GAN) analysis supports the development of more accurate bottom-up SERs

Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

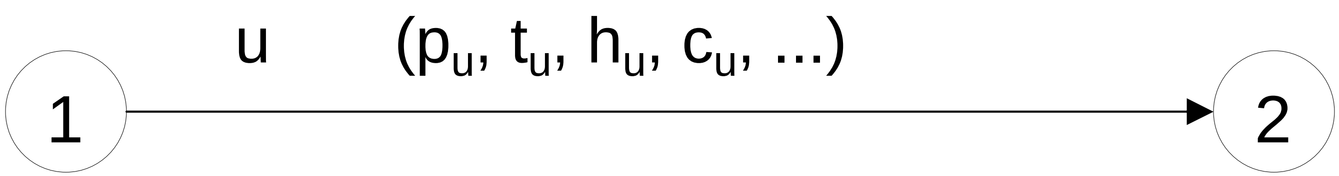
Generalized Activity Networks (GAN)

- A Generalized Activity Network (GAN) is:
  - A cyclical directed process modeling diagram (an extension of PERT)
  - The modeling capabilities of GANs include:



Generalized Activity Networks (GAN)

A GAN has as its basic element an activity (u)



p_u ≡ probability that arc “u” executes

t_u ≡ u’s execution time

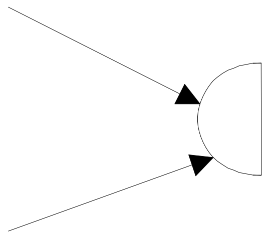
$h_u(t_u)$ ≡ probability density function for t

c_u ≡ u’s cost: may depend upon t

GAN Junctions

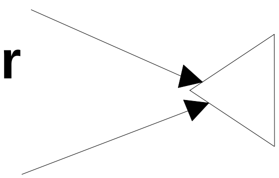
GAN Receivers

And (AND)



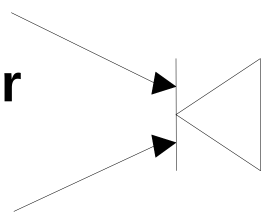
All arcs must execute to continue

Inclusive Or (OR)



Continue after any arc completes

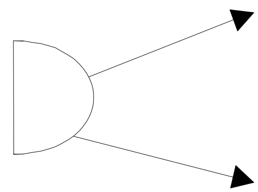
Exclusive Or (XOR)



Must complete exactly one arc to continue

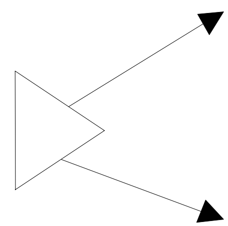
GAN Transmitters

Must follow



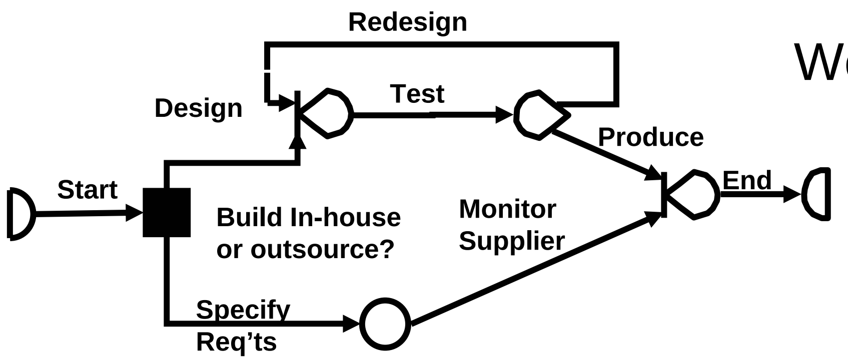
All arcs execute

May follow



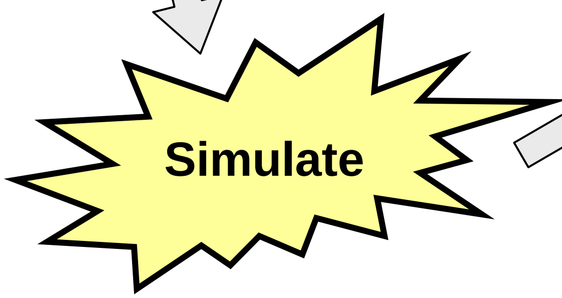
Arcs execute with assigned probabilities

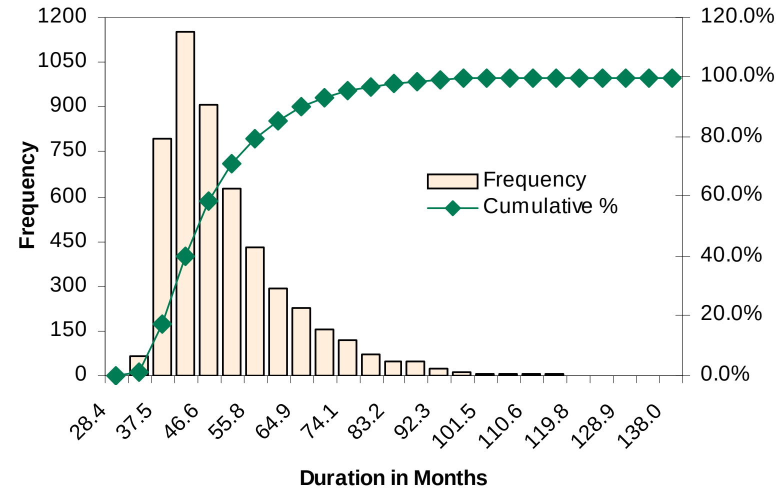
GAN Simulations



We convert GANs…

…into simulations…





…that can compute completion time and cost for complex spending programs.

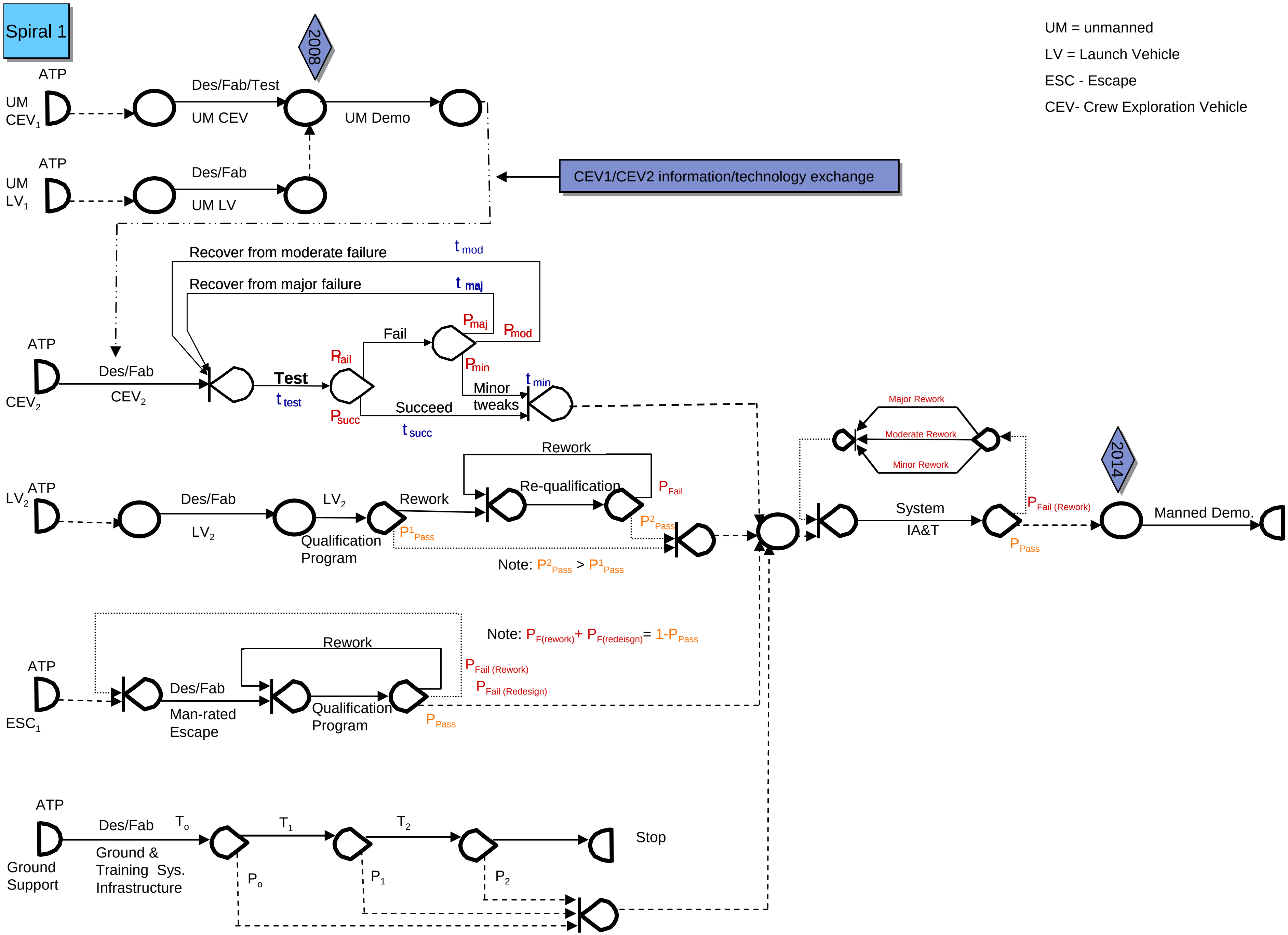
Our research shows that these simulations provide a surprising amount of insight, even with few inputs

How GANs are Built and Calibrated

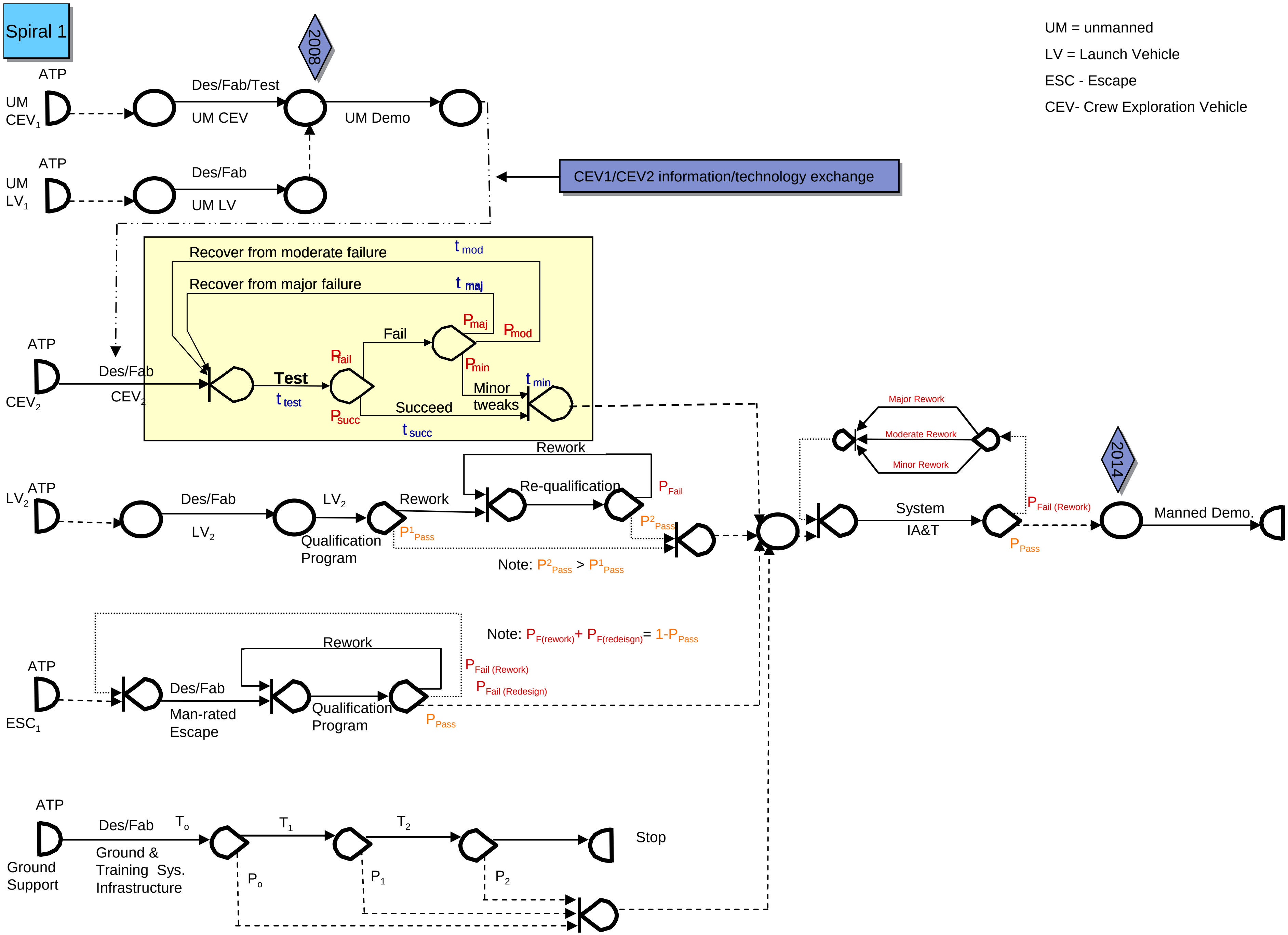
- Modeling process:
  - Build a network diagram (GAN) to describe possible program execution paths
  - Estimate parameters: establish random distribution(s)
  - Require probabilities for feedback loops or other event outcomes
  - Create a discrete-event simulation for that network
- Parameter estimation:
  - Task durations, cost, and risk levels can be based on:
    - Build-up estimates, calibration with historical data, subject matter expertise
    - Often apply a Weibull distribution (Gladstone-Miller 2002 DODCAS) to deterministic estimate
  - Feedback probabilities can be calibrated with historical data from similar programs or subject matter experts

Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

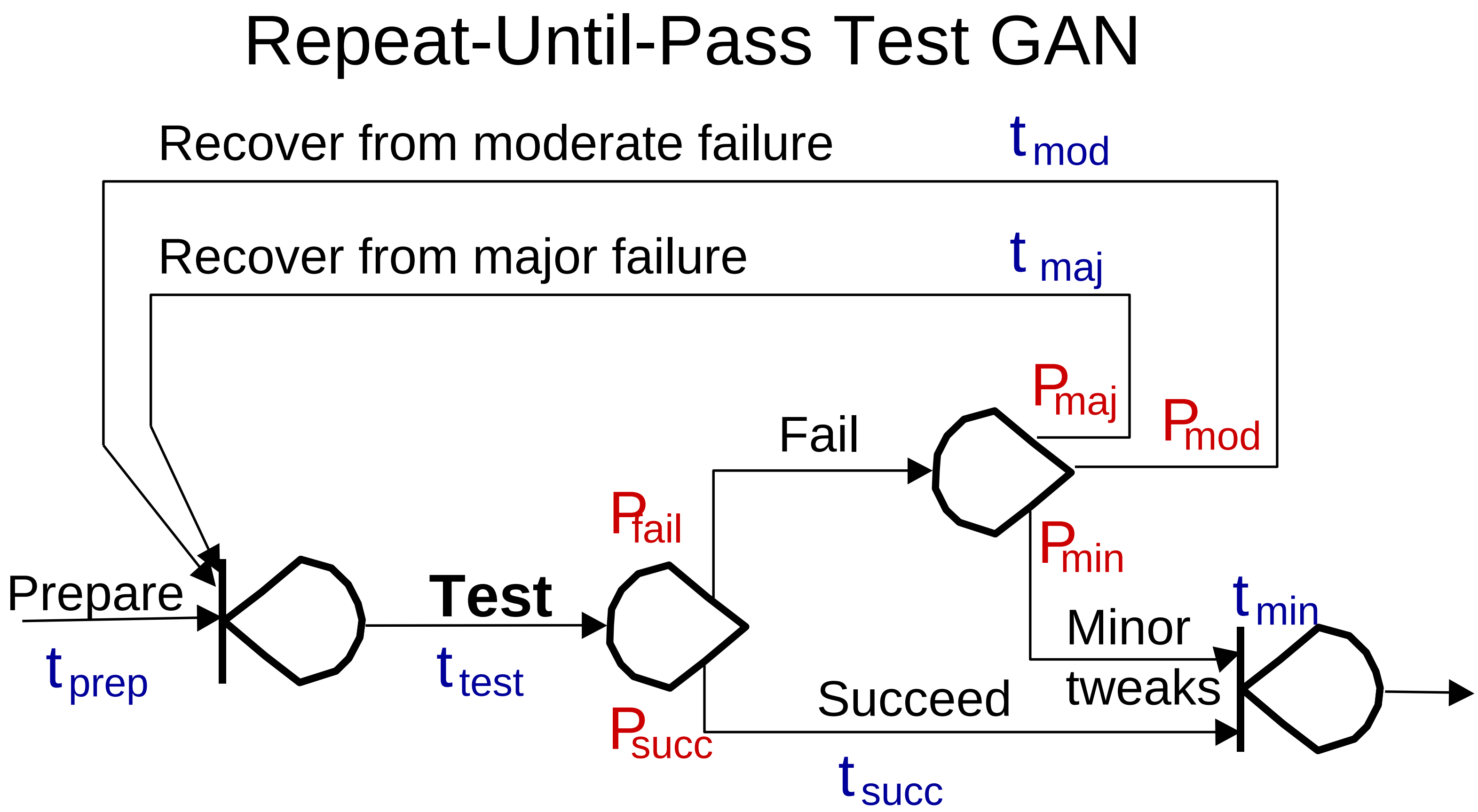


Spiral 1
2008
UM = unmanned
LV = Launch Vehicle
ESC - Escape
CEV- Crew Exploration Vehicle
ATP
UM CEV_1
Des/Fab/Test
UM CEV
UM Demo
ATP
UM LV_1
Des/Fab
UM LV
CEV1/CEV2 information/technology exchange
Recover from moderate failure
t_{mod}
Recover from major failure
t_{maj}
ATP
CEV_2
Des/Fab
CEV_2
Test
t_{test}
P_{fail}
P_{succ}
Fail
P_{maj}
P_{mod}
P_{min}
Minor tweaks
t_{min}
Succeed
t_{succ}
Major Rework
Moderate Rework
Minor Rework
2014
ATP
LV_2
Des/Fab
LV_2
LV_2
Qualification Program
Rework
P^1_{Pass}
Rework
Re-qualification
P_{Fail}
P^2_{Pass}
Note: $P^2_{Pass} > P^1_{Pass}$
System
IA&T
$P_{Fail (Rework)}$
P_{Pass}
Manned Demo.
Note: $P_{F(rework)} + P_{F(redeisgn)} = 1 - P_{Pass}$
ATP
ESC_1
Rework
Des/Fab
Man-rated Escape
Qualification Program
$P_{Fail (Rework)}$
$P_{Fail (Redesign)}$
P_{Pass}
ATP
Ground Support
Des/Fab
Ground & Training Sys. Infrastructure
T_o
T_1
T_2
P_o
P_1
P_2
Stop



Spiral 1
2008
ATP
UM
CEV1
Des/Fab/Test
UM CEV
UM Demo
ATP
UM
LV1
Des/Fab
UM LV
CEV1/CEV2 information/technology exchange
UM = unmanned
LV = Launch Vehicle
ESC - Escape
CEV- Crew Exploration Vehicle
Recover from moderate failure
t mod
Recover from major failure
t maj
Fail
P maj
P mod
P fail
Test
t test
P min
Minor
tweaks
t min
P succ
Succeed
t succ
ATP
CEV2
Des/Fab
CEV2
Rework
Re-qualification
P Fail
ATP
LV2
Des/Fab
LV2
LV2
Rework
Qualification
Program
P1 Pass
P2 Pass
Note: P2 Pass > P1 Pass
Major Rework
Moderate Rework
Minor Rework
2014
System
IA&T
P Fail (Rework)
P Pass
Manned Demo.
Note: P F(rework) + P F(redeisgn) = 1-P Pass
Rework
P Fail (Rework)
P Fail (Redesign)
ATP
ESC1
Des/Fab
Man-rated
Escape
Qualification
Program
P Pass
ATP
Ground
Support
Des/Fab
Ground &
Training Sys.
Infrastructure
To
T1
T2
Stop
Po
P1
P2

Example: Repeat-Until-Pass Test GAN



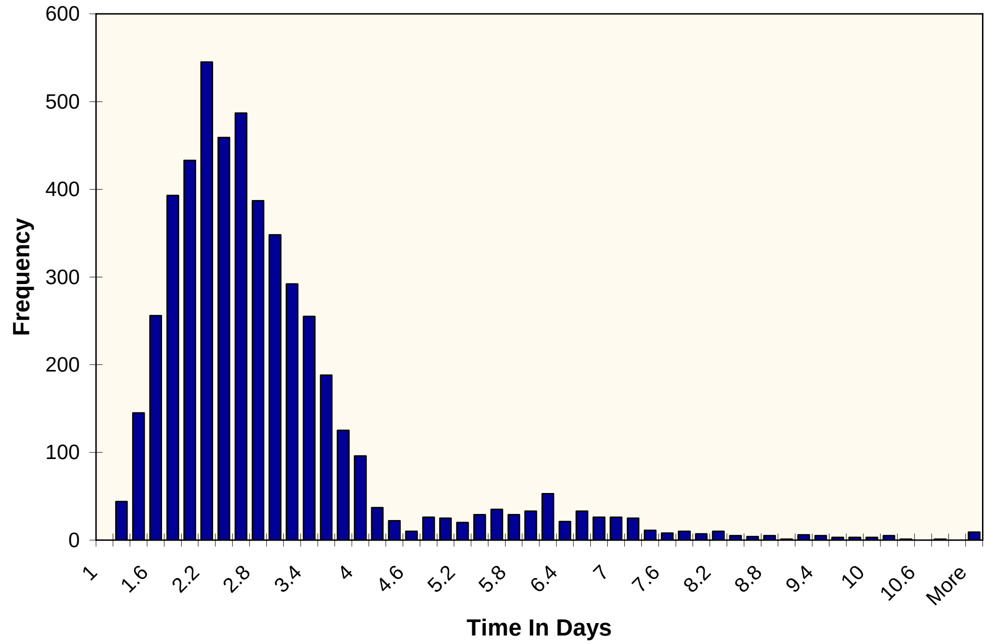
Most likely task durations in blue

Event outcome probabilities in red

Example: Repeat-Until-Pass Test GAN

- Durations for Preparation and Testing:
  - Uniform Random Variables
  - Expectation 1 day & Range 1 day: U(0.5,1.5)
- Durations for Recovery from Test Failure:
  - Minor Failure: U(0.5,1.5) Exp. Value: 1 day
  - Moderate Failure: U(1.25,2.75) Exp. Value: 2 days
  - Major Failure: U(2.0, 4.0) Exp. Value: 3 days
  - Note: Dispersion also increases with failure severity
- Duration for activities following success is 0
- $P_{success} = P_{failure} = .5$
- $P_{min} = .8$; $P_{mod} = P_{maj} = .1$
  - 10% of all failures are moderate, and 10% of all failures are major

Example: Repeat-Until-Pass Test GAN



- Performed Monte Carlo Simulation
- Expected Duration for Test Success: 2.8 days
- Large right-tail dispersion due to geometric distribution from inclusion of a probability of test failure

Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

GAN Advantages

- Hierarchical: can describe and analyze system at any level of detail
- Flexible: supports evaluation and decision-making at all levels
- Model iterative processes
- Can provide more information than simple time/cost estimates
  - Complete distribution; eliminates need for separate risk analysis
  - Identify potential problem activities for risk mitigation
- Often provide useful insight during both design (diagramming) and analysis (simulation, analytic equations) phases
- Provides a single integrated approach for understanding task interdependencies; identifying high-risk activities; and incorporating funding constraints

GAN Disadvantages

- “Uniqueness” problem
  - Data cannot be used for calibration if too program-specific
  - Breadth of data as important as depth of data
- May suffer from subjectivity of expert opinion data
  - Problem of all bottom-up estimates
- Requires detailed program data
  - Data necessary for calibration
  - Calibration necessary for meaningful schedule estimates
- “Familiarity” problem: Although growing, GANs currently not widely used for cost analysis

Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

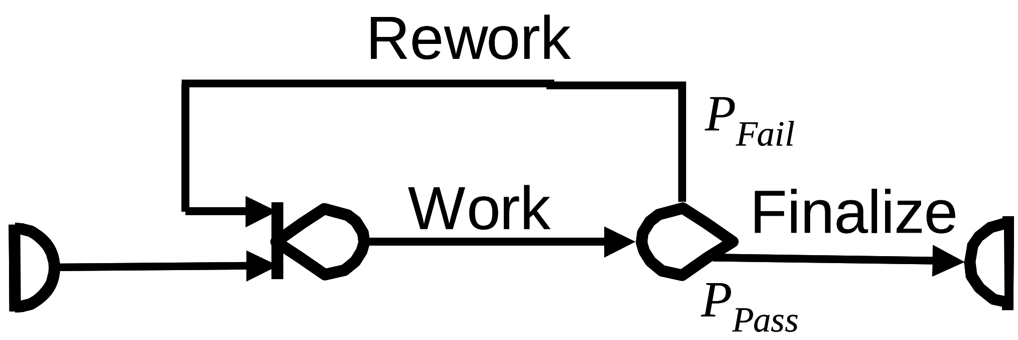
Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

Calibrating GAN Probabilities

- We consider two common GAN feedback processes
  - The One "P" Case
    - Single feedback loop with a constant probability of success
    - Preliminary results included in MORS presentation
  - The Two "P" Case
    - Successive attempts after the first failure possess a constant, but higher, probability of success that the first test trial
    - Presumes that most of the major problems are at least identified after recovery from initial failure implying a higher probability of success for subsequent trials

GAN Probabilities: One “P” Case



- Typically, probabilities of success or failure driven by expert opinion
- Probabilities *can* be appropriately calibrated by historical data
- Assumptions
  - Well defined, common test event for commodity/system
  - Access to historical data from similar systems

GAN Probabilities: One "P" Case

- Considering simple test-block GAN:
  - Trials occur until a success is achieved (with probability ***P*** for each trial)
  - Let ***X*** be the number of trials until the first success
  - ***X*** is a geometric random variable with parameter ***P***
  - Specifically,

$$E[X] = \frac{1}{p}$$

- Assuming historical data (of sample size ***n***) on number of trials from similar systems can solve for single $\boldsymbol{p}^*$ that minimizes the sum of squared errors between the expected number of trials predicted by the GAN, $\boldsymbol{E}[\boldsymbol{X}]$,and the historical data

GAN Probabilities: One "P" Case

- Thus, if $x = \dfrac{1}{p^*}$ and $\{b_1, b_2, b_3, \ldots, b_n\}$

  are the set of outcomes representing the number of trials for independent outcomes of the same GAN, we wish to:

$$\min \sum_{i=1}^{n} (x - b_i)^2$$

$$subject\ \ to: x \geq 1$$

- Conveniently, the global minimum is simply the mean of the historical data, yielding:

$$p^* = \left(\frac{\sum_{i=1}^{n} b_i}{n} \right)^{-1}$$

One "P" Case: Proof

- Since our problem is only over one dimension, we can simply consider looking at the derivative of the function with respect to ***x***

$$\sum_{i=1}^{n}(x-b_i)^2 \quad = \quad \sum_{i=1}^{n}(x^2-2b_i x+b_i{}^2) \quad = \quad nx^2-2x\sum_{i=1}^{n}b_i+\sum_{i=1}^{n}b_i^2$$

- Taking the derivative of this expression and setting it to zero, we get that:

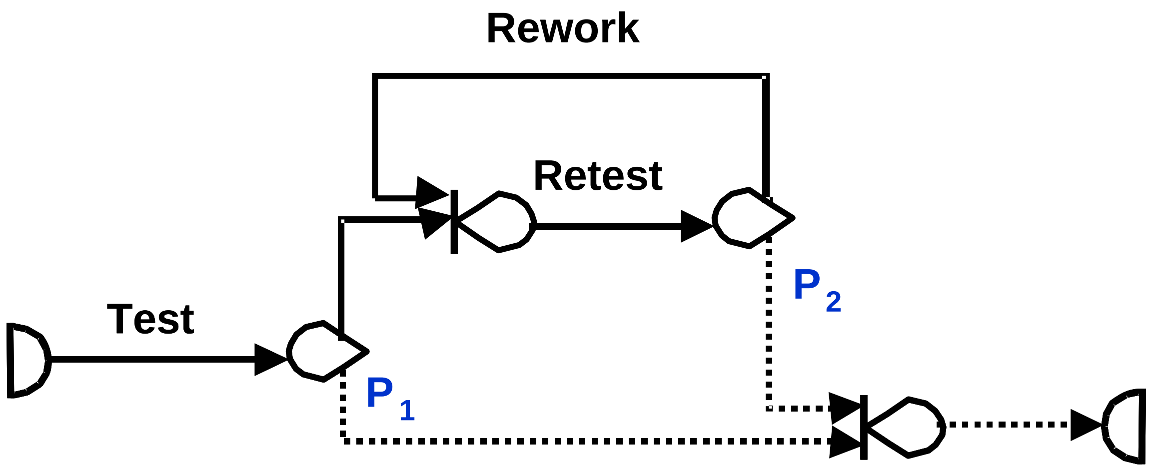
$$2nx-2\sum_{i=1}^{n}b_i=0 \qquad \Rightarrow \qquad x=\frac{\sum_{i=1}^{n}b_i}{n}$$

- Thus, we can estimate ***p**** by simply by taking the inverse of the average of the outcomes of the trials

GAN Probabilities: One “P” Case

- This simple, straightforward result is powerful because analysts can easily ***objectively*** calibrate GAN probabilities
- Further, in absence of historical data, analysts should seek
  - Unbiased expert opinion on “average” number of tests until success
  - Should produce better estimates of realistic probability of success than directly asking for them

GAN Probabilities: Two “P” Case



- Probability of success on first test: P_1
- Probability of success on every other test, *conditional* on first test failing: P_2
  - Might expect $P_2 > P_1$ due to knowledge of what failed, additional effort spent on that item, etc.

GAN Probabilities: Two “P” Case

- Consider a test event with the following historical data:

Historical Program	# Trials until Success	1st Trial Success? (Yes=1, No=0)	2nd Trial? (Did the 1st trial fail?)	# of "P2" Trials
1	6	0	1	5
2	7	0	1	6
3	4	0	1	3
4	1	1	0	0
5	8	0	1	7
6	1	1	0	0
7	2	0	1	1
8	1	1	0	0
9	12	0	1	11
10	4	0	1	3

- We could calculate a single probability, *p*, using the previous technique
  - Method of calibrating P_1 and P_2 should reduce to One “P” case if probabilities are constant

GAN Probabilities: Two "P" Case

- Let x_1 and x_2 be decision variables and $\{b_1, b_2, b_3, ..., b_n\}$ historical data.
- Let $b_1^i = 0$ if the first trial failed and $b_1^i = 1$ if it succeeded and assume that there are J successes.
- Let b_2^j represent the number of subsequent trials with a probability, p_2 ,of success
- As before, we wish to minimize the sum of squared errors between the expected number of trials predicted by the GAN and the historical data for each decision node:

$$\sum_i (x_1 - b_1^i)^2 + \sum_j (x_2 - b_2^j)^2$$

GAN Probabilities: Two "P" Case

- We can minimize each sum separately, yielding x_1 and x_2 , and thus our P_1 and P_2
- Using the data from our example we produce the probabilities:

$$P_1 = x_1 = 0.3$$

$$P_2 = \frac{1}{x_2} = \frac{1}{36/7} = 0.1944$$

- Monte Carlo testing demonstrates method to provide robust estimation of data generating process even when $P_1 = P_2$

Agenda

- Traditional T&E Schedule Estimation
- Generalized Activity Networks (GAN)
- Example Application: Repeat-until-Pass Test GAN
- Advantages and Disadvantages of GAN Approach
- Extra Topics (Time Permitting)
  - Calibrating GAN Probabilities
  - GANs versus Weibull Distributions

Weibull Enveloping Distributions

- Evaluate feasibility of estimating hyper-geometric processes through an enveloping Weibull distribution
- Why?
  - Hyper-geometric processes and Weibull distributions are similarly right tailed
  - Weibull distributions are well understood throughout cost community and, thus, better accepted than GAN feedback simulations
- Study Objective: See if its possible to fit a Weibull to a feedback process under ideal conditions, ie. when we actually possess full knowledge of the process and relevant statistics

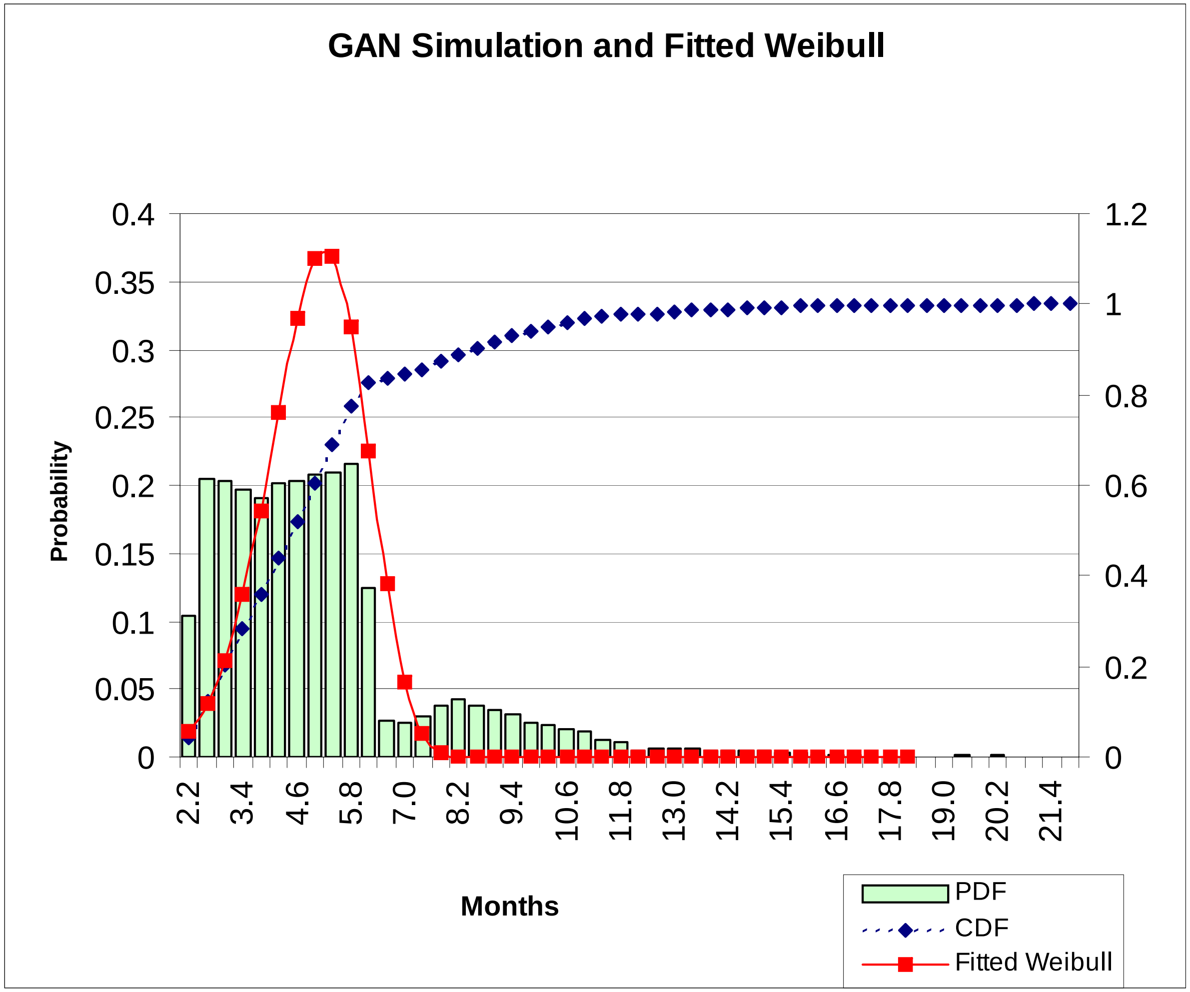
Weibull Enveloping Distributions

- Method for matching GAN simulation results through Weibull distributions:
  - Create GAN simulation to generate hyper-geometric data
    - Can be calibrated with different underlying probabilities and duration distributions
    - Produce test data from 10,000 trials to ensure robust characterization of data generating process (can employ asymptotic theory)
  - Optimize Weibull parameters to fit the distribution to generated data
    - Identify appropriate metrics to define “goodness of fit”
    - Perform optimization upon selected objective function
  - Evaluate “best-fitting” Weibull predictions to simulated results

Weibull Enveloping Distributions

- Population Data: Data Generating Process
  - Relatively straightforward parameters to maximize probability of successful fit
  - Trial Durations ~ Uniform (2,6)
  - Feedback Probabilities: 50/50, 80/80, 30/30, 50/80, etc.
- Goodness of fit
  - Cost and Schedule Estimates typically reported at the mean [expected value], 50% CDF, and 80% CDF
  - If we know two of the three, we can optimize the selection of Weibull parameters (α,β) to minimize the error between the prediction for the third metric and simulated data
  - If we assume we only know the expected value of the data, we can optimize parameters such that we minimize the joint error between the 50% and 80% CDF

Characteristic Result



Weibull Enveloping Distributions

- Our analysis on a variety of simple GAN simulations indicates that it is not feasible to adequately envelope a GAN feedback loop with a single Weibull distribution
- Enveloping Weibull performs worse under more complex, realistic assumptions such as Normal or Weibull distributed durations in the population data generating process
- For a few specific cases we were able to fit a Weibull with a relatively small mean squared error, e.g. a prediction error of less than 20%
  - However, calibration was made with our complete knowledge of the underlying data generating process, which we would not have with real data.

Weibull Enveloping Distributions

- Successful fit only indicates that it is feasible for a Weibull to approximate a specific feedback process *not that it can actually be implemented with statistical confidence*
  - Wouldn't actually possess information on the expected value, 50%, or 80% CDF with real data from which to optimize Weibull
  - Weibull parameters would be calibrated, as always, from outside of the estimated process
  - Fitting arbitrary points of a CDF does not necessarily indicate a minimization of the error between the Weibull and simulated data mass functions
- We recommend the continued use of GAN feedback loops to simulate feedback processes, such as testing.