Analytics and Optimization of Manufacturing Process Data

Vincent Hoff

Introduction:

The focus of this project is on the problem of reconstructing, or identifying, precedence relations in manufacturing processes using data taken from the factory floor. This data includes start times and end times of subassembly tasks, resource assignments and subassembly locations.

Currently there are two samples of manufacturing data from Boeing. One sample is recorded data from the factory floor, and the other is network structure data related to the ideal process, as designed by industrial engineers.

The reconstructed precedence networks will be used to identify and analyze differences in the actual manufacturing process from the designed or ideal process in order to improve efficiency and resource usage in the long run. An example of one of the smaller precedence networks from the data can be seen in Figure 1.

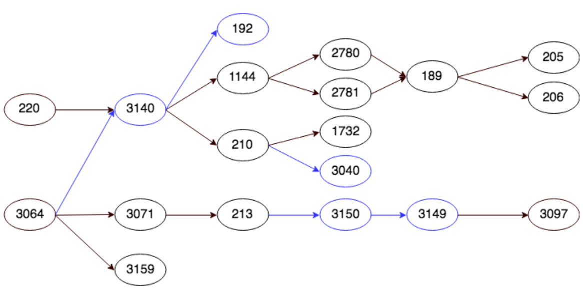

Figure 1: Control Code 22 Subassembly Precedence Network

Data Analysis and Visualization:

In this REU project, a heavy emphasis has been placed on formatting, analyzing, and simulating data for this effort, as well as assisting Dr. Beck and Xiaotian Xie (graduate student) in testing the structure identification algorithms they are developing to address this problem. These algorithms are designed for inferring precedence networks of manufacturing processes with directed, layered structures. In such a network, a series of sequenced assembly tasks are viewed as nodes, with edges representing precedence constraints between tasks that are only allowed between successive layers.

Given a dataset of assembly task start and completion times, the goal is to identify all possible precedence relationships in the network so as to know if manufacturing was done in accordance with the manufacturing plan, and if there is any room for improvement. The project's main contributions are providing simple algorithms to infer the underlying structure and analyzing the corresponding sample complexity based on the coupon subset collection problem [Adler and Ross 2001].

As the data from Boeing was incomplete, partly due to being sanitized on their part, the initial task was to determine the completion rate of the data, grouped by the "Control Code," which can be considered to be

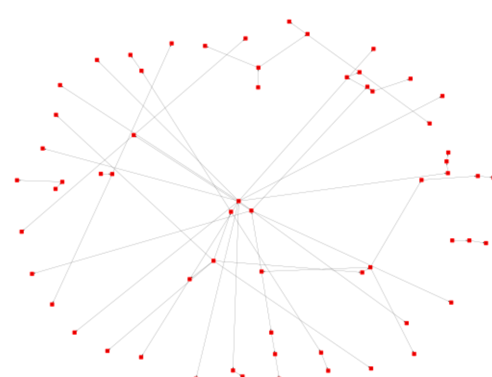

Figure 2: Data Visualization with NetworkX

analogous to a single layer or stage in the manufacturing process.

Once control codes with a smaller number of nodes or subassembly tasks, and with higher rates of completion were identified in the process of this REU project, the next step was to visualize the data within the control code based on the precedence relationships using a python data visualization package called *NetworkX*. Figure 2 shows an example of this visualization. Please note that the graph is extremely high resolution and the individual nodes are thus difficult to read at the current resolution.

With the help of this package as well as inspection of the small sample of data by visual inspection, conclusions about control code significance and the layered structure within the manufacturing process could be made.

As we only had one sample set of recorded start and stop times, one thought was to bootstrap the data in some manner to generate more sets of data; however, this is not possible given the single point of observation data per node, and the individuality of the nodes. Instead additional simulated samples of such data are being created based on the original recorded data *and* the ideal process, by modeling the existing data using a probabilistic distribution of task completion

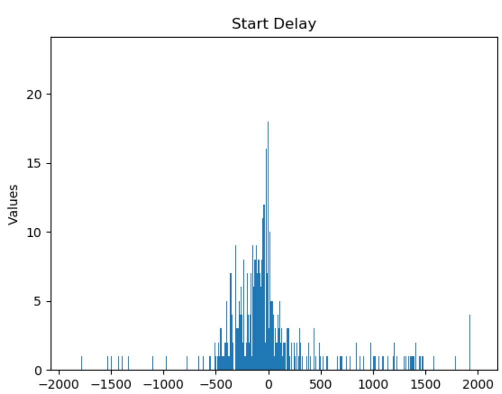

Figure 3: Distribution of Start Delays

delays. We draw from the distribution by assigning probabilities to each delay based on the incidence rate of the sample versus the population according to equation (1) where the likelihood P of a delay i occurring is calculated by dividing the number of occurrences k_i of that delay by the sum of all occurrences of delays. A histogram of this distribution is shown in Figure 3. This process enables the creation of additional simulated observations for singular data points, allowing us to more strenuously test and validate our inference algorithms. The likelihood of a delay occurring is

$$P_i = \frac{k_i}{\sum_i^n k_i} \quad (1)$$

Specifically, the dataset is made up of 95 control codes, cumulatively containing 4137 assembly tasks or nodes. As the amount of data is simply overwhelming, the visualization software and the data simulation according to the distribution were only applied initially to five control codes to attain a better understanding of the relationships within the data subset.

As one example, the knowledge gained from this process was used to create further sets of data with the use of statistical modeling for Control Code 22 (CC22) (shown in Figure 1). Fifty additional observations for each task within CC22 for task begin and completion times were generated based on a few assumptions, which will be explained in more detail in the next paragraph.

The distribution for the data which was outlined above focused specifically on the difference between the *scheduled* start and *actual* start times of each data point in the sample. As some tasks were started before the scheduled start times by the mechanics, a negative delay was possible in part of the distribution model. This led to a potential issue where if a task duration was short enough and was assigned a large enough negative delay, its simulated completion time could occur before its simulated start time. This issue was alleviated by adjusting the constraints to pick the later start time between the following two: Actual start time modeled after the distribution delay and actual completion time of the predecessor. However, another problem was that with the new set of constraints, the actual completion time of the predecessor also depended on the actual start time of a task plus the scheduled duration. As now both constraints depended on each other the code would run over the data, applying the constraints, however the co-dependency made it so that once one task was updated, the other was outdated. In other words: as soon as the python script picked a maximum between the predecessor's actual completion time and the distribution-based actual start time, it assigned a start time to each point in the data set. It then added the fixed scheduled duration to each task's actual start time to find the actual completion time. However, the initial constraint of the script to pick the maximum between the two different possible start times led to the script overriding each actual completion time with a value that was at least as large as the original value, or larger, but never smaller. This completely invalidated the precedence relationships between the tasks as now most tasks had a later completion time than when the script checked for compatibility between the predecessors and successors. In turn, the second constraint was removed completely to alleviate this problem. Please see the Appendix for examples of real and simulated data; the latter data is based on the probabilistic sampling process developed in this project and described in this report.

Future work:

Once the python script has been modified to no longer cause inconsistencies it will be generalized to apply to the dataset as a whole, proceeding with the process as outlined in the introduction. This work is continuing.

Conclusions:

Once the number of data points becomes large enough, manual inspection is no longer viable and requires automation. The usefulness of numerical python is unparalleled in this regard, highlighting the importance of learning how to write simple programs for any engineer. Furthermore, it is important to consider the constraints of all theoretical models before beginning on the implementation as a flawed model can lead to an impossible implementation. Thus, the theoretical model should be scrutinized properly before beginning on the implementation.

The importance of combining statistical modeling and a computer science implementation is of utmost importance when working with big data.

Acknowledgement:

I would like to thank Dr. Beck and Xiaotian Xie for giving me this great opportunity to learn about their research and assisting me in acquiring new skills related to their fields through this Research Experience for Undergraduates granted by the department for Industrial and Enterprise Systems Engineering.

References:

I. Adler and S. M. Ross. The coupon subset collection problem, *J. Appl. Prob.* **38**, 737-746 (2001).

S. T. Kong, D. Katselis, C. L. Beck and R. Srikant. Structure Identification in Layered Precedence Networks. *Proceedings of the Conference on Control Technology and Applications*, 2017.

Appendix:

Source Data

JobId	Layer	LINE	Team	SCHD START DATE	SCHD COMP DATE	ACTUAL COMP DATE	PREDECESSORS	ControlCode
188	3	1	NoTeam	8/18/2017	8/18/2017	08/24/17 22:25	3127	CC22
191	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 12:40	3032	CC22
193	3	1	NoTeam	8/17/2017	8/17/2017	08/22/17 11:41	3032	CC22
195	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 8:43	3032	CC22
196	3	1	NoTeam	8/16/2017	8/18/2017	08/21/17 10:38	3032	CC22
198	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 10:32	3032	CC22
200	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 22:11	3032	CC22
201	2	1	NoTeam	8/18/2017	8/18/2017	08/24/17 1:51	371	CC22
202	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 12:30	3032	CC22
203	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 4:39	3032	CC22
371	1	1	NoTeam	8/16/2017	8/16/2017	08/21/17 5:07		CC22
767	3	1	NoTeam	8/18/2017	8/18/2017	08/24/17 19:50	201	CC22
979	3	1	NoTeam	8/16/2017	8/16/2017	08/21/17 21:52	3032	CC22
2347	3	1	NoTeam	8/16/2017	8/16/2017	08/22/17 4:08	3032	CC22
2694	1	1	NoTeam	8/16/2017	8/16/2017	08/21/17 10:41		CC22
2779	3	1	NoTeam	8/17/2017	8/17/2017	08/22/17 5:28	3154	CC22
3032	2	1	NoTeam	8/16/2017	8/16/2017	08/21/17 13:36	2694	CC22
3068	3	1	NoTeam	8/16/2017	8/17/2017	08/22/17 7:00	3032	CC22
3127	2	1	NoTeam	8/16/2017	8/16/2017	08/21/17 20:23	2694	CC22
3154	2	1	NoTeam	8/16/2017	8/17/2017	08/21/17 23:47	371	CC22

Simulated Data

JobID	Layer	Predecessors	Schd Duration Data	Schd Start Data	Observation 1	Observation 2	Observation 3	Observation 4
188	3	3127	0 days 01:58:00.000000000	8/18/2017 20:42	8/21/2017 20:23	8/21/2017 20:27	8/21/2017 20:25	8/21/2017 20:27
191	3	3032	0 days 00:38:00.000000000	8/16/2017 8:40	8/21/2017 13:39	8/21/2017 13:37	8/21/2017 13:39	8/21/2017 13:40
193	3	3032	0 days 00:49:00.000000000	8/17/2017 0:20	8/21/2017 13:40	8/21/2017 13:41	8/21/2017 13:40	8/21/2017 13:41
195	3	3032	0 days 01:00:00.000000000	8/16/2017 6:20	8/21/2017 13:36	8/21/2017 13:37	8/21/2017 13:41	8/21/2017 13:40
196	3	3032	2 days 03:24:00.000000000	8/16/2017 8:40	8/21/2017 13:38	8/21/2017 13:40	8/21/2017 13:41	8/21/2017 13:36
198	3	3032	0 days 01:20:00.000000000	8/16/2017 7:20	8/21/2017 13:40	8/21/2017 13:41	8/21/2017 13:40	8/21/2017 13:36
200	3	3032	0 days 03:00:00.000000000	8/16/2017 10:30	8/21/2017 13:38	8/21/2017 13:38	8/21/2017 13:40	8/21/2017 13:40
201	2	371	0 days 01:30:00.000000000	8/18/2017 1:20	8/21/2017 5:12	8/21/2017 5:10	8/21/2017 5:07	8/21/2017 5:09
202	3	3032	0 days 01:30:00.000000000	8/16/2017 6:50	8/21/2017 13:39	8/21/2017 13:39	8/21/2017 13:41	8/21/2017 13:37
203	3	3032	0 days 02:00:00.000000000	8/16/2017 3:40	8/21/2017 13:38	8/21/2017 13:37	8/21/2017 13:41	8/21/2017 13:39
371	1		0 days 00:30:00.000000000	8/16/2017 1:45	8/16/2017 0:05	8/17/2017 0:35	8/16/2017 2:07	8/16/2017 0:51
767	3	201	0 days 00:30:00.000000000	8/18/2017 17:05	8/24/2017 1:51	8/24/2017 1:52	8/24/2017 1:51	8/24/2017 1:55
979	3	3032	0 days 00:45:00.000000000	8/16/2017 17:00	8/21/2017 13:39	8/21/2017 13:36	8/21/2017 13:36	8/21/2017 13:40
2347	3	3032	0 days 00:29:00.000000000	8/16/2017 17:01	8/21/2017 13:38	8/21/2017 13:37	8/21/2017 13:39	8/21/2017 13:40
2694	1		0 days 00:45:00.000000000	8/16/2017 2:45	8/15/2017 23:54	8/16/2017 2:53	8/15/2017 19:22	8/15/2017 23:02
2779	3	3154	0 days 01:45:00.000000000	8/17/2017 1:20	8/21/2017 23:48	8/21/2017 23:48	8/21/2017 23:47	8/21/2017 23:50
3032	2	2694	0 days 00:30:00.000000000	8/16/2017 6:20	8/21/2017 10:46	8/21/2017 10:45	8/21/2017 10:46	8/21/2017 10:46
3068	3	3032	0 days 05:50:00.000000000	8/16/2017 23:20	8/21/2017 13:36	8/21/2017 13:36	8/21/2017 13:40	8/21/2017 13:38
3127	2	2694	0 days 03:30:00.000000000	8/16/2017 8:32	8/21/2017 10:45	8/21/2017 10:46	8/21/2017 10:46	8/21/2017 10:41
3154	2	371	0 days 02:00:00.000000000	8/16/2017 23:20	8/21/2017 5:08	8/21/2017 5:07	8/21/2017 5:09	8/21/2017 5:08