SEQUENTIAL ANALYSIS OF REASONS FOR NON-COMPLETION OF ACTIVITIES: CASE STUDY AND FUTURE DIRECTIONS

Jose Nilton Oliveira Filho[1], Lucio Soibelman[2] and James Choo[3]

ABSTRACT

Reliable work flow in production processes are of utmost importance to the successful completion of construction projects. Although a perfectly reliable work flow is unlikely to occur due to the inherent variability of production in construction, assignments should be measured and monitored, and causes for non-realization should be investigated in order to mitigate negative impacts of variability.

Lean construction principles have been applied effectively in several projects and the identification of common problems demonstrated usefulness in the decrease of variability. However, the discovery of the main or primary causes of those problems and their impact on the whole project still continue to be a vague and obscure issue.

The purpose of this paper is to first present a case study where a methodology to discover sequences of common non-conformances was studied and applied to a project database. Such sequences might be an indication of frequent patterns where one error category might have influenced subsequent ones. Then, the difficulties faced in this study and the relevance and importance of integrating project and external data sources for causal data analysis and knowledge discovery will be discussed.

KEY WORDS

Sequential analysis, pattern recognition, data mining, knowledge discovery.

[1] Research Assistant, Civil and Envir. Engrg. Department, 3142 Newmark Civil Engrg. Lab., Univ. of Illinois at Urbana-Champaign, Urbana, IL 61801, Phone +1 217/333-2071, foliveir@uiuc.edu

[2] Assistant Professor, Civil and Envir. Engrg. Department, 3129C Newmark Civil Engrg. Lab., Univ. of Illinois at Urbana-Champaign, Urbana, IL 61801, Phone +1 217/333-4759, FAX 217/265-8039, soibelma@uiuc.edu

[3] Product Development Leader, Strategic Project Solutions, Inc. P.O. Box 2835, San Francisco, CA 94126-2835, Phone +1 415/362-3200, Fax +1 415/362-3210, jchoo@strategicprojectsolutions.net

INTRODUCTION

Reliable work flow in production processes are of utmost importance to the successful completion of construction projects (Ballard 1997, Ballard 2000). To facilitate the management and control of a project's work flow and mitigate negative impacts of variability, lean construction advocates that assignments should be closely measured and monitored, and causes for non-realization should be investigated (Koskela 1999, Ballard 2000).

The adverse relevance of variability to flow of work and system throughput has been recognized and exemplified by the literature (Womack and Jones 1996, Tommelein et al. 1999), and even though a perfectly reliable work flow is unlikely to occur due to the inherent variability of production not only in construction but also in manufacturing as a whole, control measures should be taken in order to diminish the risk of variability propagation to downstream flows.

In this scope, the purpose of the non-realization inspection is to identify main problems that are constraining the completion of planned activities. Once the sources of variability are located, corrective actions should be launched and the extent of their application observed so that those problems do not come into play again. This promotes work flow variability reduction, helping to increase overall project's workflow.

However, conventional project control in the Architecture, Engineering and Construction (AEC) industry generally focuses on discrepancies from cost and schedule project objectives, and it has not directly addressed the management of production and its variability (Ballard 2000).

Overcoming this deficiency, the Last Planner™ system of production control, based on lean construction principles, has been broadly and successfully implemented in several projects over the last few years. It effectively combines control and improvement to repress variability and the waste generated by it (Koskela 1999, Ballard 2000). Its focus on plan realization and the collection of reasons for non-completion of activities throughout a project deployment is an effective approach to the identification of the most common reasons and highlights these to project managerial personnel.

Nevertheless, the discovery of the main or primary causes of problems that limit the completion of construction assignments and their impact on the whole project still continue to be a vague and obscure issue. Little is known about the relationship among non-completion of activities and how they correlate to each other (e.g., Are there similar dependencies among them? Does a particular failure in an activity influence downstream work?).

If the relationship among the sequence of non-conformance events could be better understood, where some problems might be causing or influencing the development of others, many undesirable outcomes could be prevented by the appropriate selection of proactive actions. Moreover, the study of such relationships and their further integration with other project related data suggest a potential and promising means of having an even more informative analysis of the reasons that constrain the completion of activities. It is expected that the addition of this extra functionality to project control systems, such as Last Planner™, would help to increase the overall work flow and to reduce project variability by warning of prospective failures.

The remainder of this paper is organized as follows. The next section presents the paper's main intentions and objectives. Afterward sequential analysis and its application are introduced. A case study presents an experimental work where sequences of non-completed activities were detected in a construction project. Then the difficulties faced during this initial investigation, as well as further directions, are discussed. Finally the paper summarizes its findings, expected contributions of prospective research effort and the importance of the proposed approach to improvements in construction project control.

OBJECTIVES

The main focus of this paper is to present a case study where common patterns of sequences of non-completed activities were identified in a large on-going capital facility project. Although it cannot be shown that such sequences explain cause-effect relationships, they are an indication that some of them are somehow common and repetitive.

SEQUENTIAL ANALYSIS

Managers are usually intrigued by events that by some means occur in sequence. Consumers regularly buy products following some pattern (e.g., computer, printer, scanner). Customers typically rent movies in succession (e.g., Godfather trilogy). Assiduous readers purchase books in particular sequences (e.g., introductory, advanced levels). Several businesses try to investigate such frequent behavior of their clients and take advantage of that information by anticipating their clients' next probable purchase or stimulating their next acquisition.

The construction domain is not significantly different. A project is composed of an ordered sequence of activities. The start of any activity is constrained by successful completion of its predecessors. Nevertheless it is not unusual to face problems on site that result in postponement of downstream work which in turn impacts overall project completion. But, how are these problems related to each other? Are there similar dependencies among them? Does the occurrence of a non-conformance in one activity influence following ones?

The study of regular patterns in databases had a considerable breakthrough with the analysis of series of transactional records (Aggrawal and et 1993). Also known in the literature as basket data analysis, it consists of the discovery of items commonly acquired together. As a simple example, consider a supermarket where customers buy groceries from time to time. Any purchase is a transaction and a transaction is composed of products bought jointly in a purchase. The focus of this study was the detection of products that were obtained together frequently. With that information available, managers could enhance their sales by improving store layout (e.g., displaying products that are commonly bought together close to each other) and promoting combined sales (e.g., chips with dip).

After successful implementation of such an approach, retailers and marketing analysts went further, learning not only merchandises that are purchased together but also merchandises that are purchased after the purchase of other one(s). This analysis of sequential events follows naturally from the association rules analysis (Aggrawal and Srikant

1995) and it has been currently applied to genetics research (DNA sequence), telecommunications (calling patterns), and financial analysis (stock market fluctuation). A brief description of such approach, based on the work developed by Aggrawal and Srikant (1995), is provided below.

Consider a succession of purchase events that occur in some temporal order (For the sake of simplicity, the time difference among those events has been considered irrelevant) and each purchase is composed of one or more products or items (Table 1). For example, customer 2 first buys products 10 and 20, then product 30 and finally products 40, 60 and 70. In addition, products acquired in the same purchase do not possess any precedence relationship.

Table 1: Example of records of customer purchases (Adapted from Aggrawal and Srikant 1995)

Customer	Time	Items
1	1	30
1	2	90
2	1	10, 20
2	2	30
2	3	40,60,70
3	1	30,50,70
4	1	30
4	2	40,70
4	3	90

A sequence is defined as all the purchases of a customer ordered chronologically. A graphical representation of such sequences is provided in figure 1. The problem of identifying sequential patterns is to find the sequences or subsequences among all the sequences that satisfies a minimum threshold value (technically known as minimum support). The importance of such a threshold is to set up the smallest amount of times that a particular sequence has to occur in order to be considered relevant. Moreover in large databases this threshold helps to reduce the amount of computation necessary to encounter frequent sequences.

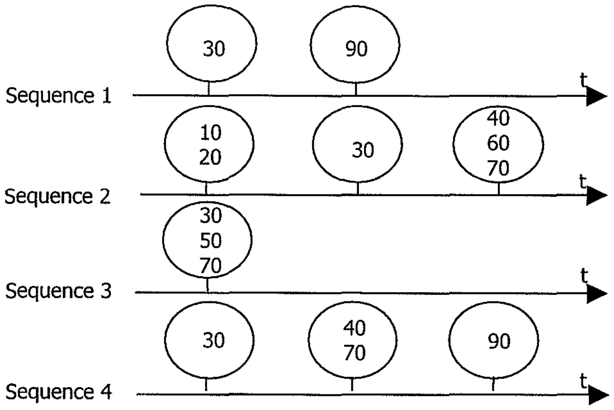

Figure 1: A set of sequences of events (e.g., purchases) derived from table 1

In this illustrative example, assume that the threshold value is 50% (i.e., to be considered frequent a subsequence has to occur at least 2 times out of the 4 existing sequences). From that setting it can be observed that purchases of product 30 followed by product 90 (which occurred in sequences 1 and 4) and purchases of product 30 followed by products 40 and 70 (which occurred in sequences 2 and 4) represent frequent subsequences.

A similar method can be applied to data regarding reasons for non-completion of activities such that sequences of activities facing non-conformances in some regular pattern could be encountered. For example, consider the hypothetical schedule provided below, where a node is represented by an activity identification number and, if an activity was not finished as planned, a reason for non-completion category number. Individual sequences can be represented by paths from a starting node until a node without successors is found. For instance, the paths <010,020,070,100> and <010,020,050,070,100> would represent possible sequences of activities.

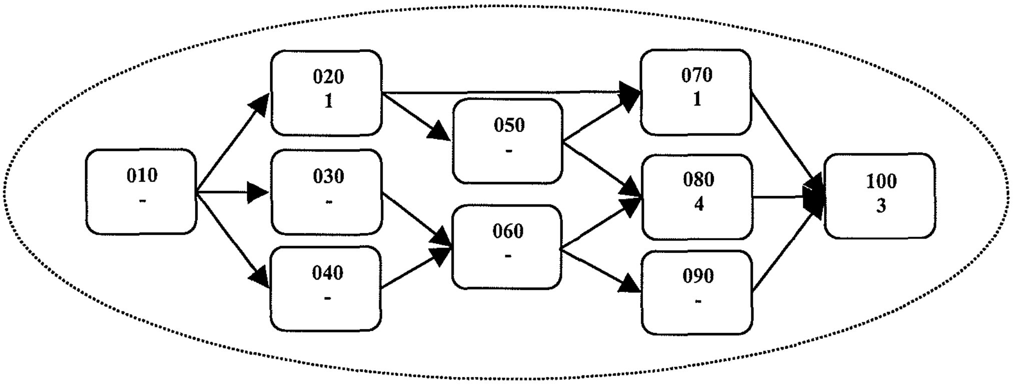

Figure 2: Subset of hypothetical schedule composed of activity identification number and reason for non-completion code

The deployment of frequent sequences analysis could be useful as it might reveal sequences of activities possessing the same patterns of failures. For example, it could happen with some frequency that design changes might result in lack of equipment in subsequent activities; once the equipment that was adequate to perform concrete pouring is no longer suitable, because the design change resulted in major modifications of concrete structures.

CASE STUDY

An experimental study has been conducted with data from a large on-going capital facility project. Strategic Project Solutions has been working with the main contractor of the project under study to implement SPS Project Suite which guides production teams in increasing plan reliability at production control level, promotes better interaction between production teams through transparency, and pulls material to workface using production level plans.

Data regarding the reasons for non-completion of activities have been collected daily such that a more rigorous control of project flow and variability could be performed. Such regular close inspection of conformance with planned work provided project managers with a better overview of the work flow, its variability and causes impacting the non- realization of some activities. Yet little is known about the relationship among causes of non-completion of activities.

In order to begin understanding those relationships, if any, an experimental work is being performed on the recorded reasons for non-completion. The first goal of this investigation is to verify the existence of such relationships. Future steps include the analysis of their correlation and relevance to other project features such as cost, time, quality and/or safety; Also, the integration of previous project related data as a means of predicting the occurrence of prospective non-conformances based on past information will be investigated. Initial results from the earliest step are presented herein; Latter ones will be investigated in the near future.

The data analyzed for this study spans basically 4 months of excavation and foundation work. Roughly 18,000 activities (ranging from a fraction of a day to a few days) were assigned and from those approximately 4500 had some non-conformance. This accounts for an average percent plan completed (PPC) of 75% for the period being considered. The reasons for non-completion were collected according to the following classification, which were determined and refined by the application of lean construction principles in numerous projects.

Table 2: Classification of reasons for non-completion

Code	Category
1	Information
2	Material
3	Labor
4	Plant (Equipment)
5	Weather
6	Directive
7	Prerequisites
8	Site Access

Following the approach described in the previous section, the provided data was preprocessed such that spurious values, mainly at the beginning and the end of the collection period, as well as some records with missing attributes were excluded. Because such records were irrelevant and ambiguous they were eliminated so that they would not mislead the investigation. In addition, all sequences of activities (together with their non-completion code for those activities that had some non-conformance) were created such that they would comply with the sequential analysis algorithm input file format. In other words, all the paths constituting a sequence of activities from the production control network were inserted into an input file, one path (i.e., sequence) per line.

The production control network established small logic networks, also known as work streams, for each significant milestone activity. Since milestones in fact determine goals and not activities, they were not modeled to have successors within the production control level. Because of that the sequence of activities input file was constrained to a few activities. Unsurprisingly this reduced the accuracy of the study. Nevertheless some patterns were still identified. The preliminary results obtained from this analysis are tabulated below, where the column title *sequential pattern of 3 activities* (e.g., sequence 1→4→7) means that an activity faced non-conformance classification 1 (Information) followed by an activity that had a non-conformance 4 (Equipment) followed by an activity that confronted non-conformance 7 (Prerequisites) and the column title *number of incidences* means the number of times that such pattern occurred in different work streams.

Table 3: Sequential analysis primarily results

Sequential pattern of 1 activity	Number of Incidences	Sequential pattern of 2 activities	Number of Incidences	Sequential pattern of 3 activities	Number of Incidences
$3_{\text{(Labor)}}$	1329	3→3	135	3→3→3	39
$1_{\text{(Information)}}$	991	6→6	85	6→6→6	21
$6_{\text{(Directive)}}$	593	4→4	72	4→4→4	16
$7_{\text{(Prerequisites)}}$	492	7→7	71	7→7→7	14
$4_{\text{(Equipment)}}$	438	1→1	63	1→4→7	7
$2_{\text{(Material)}}$	421	2→2	34	2→2→2	7
$5_{\text{(Weather)}}$	161	1→3	21	…	…
$8_{\text{(Site Access)}}$	108	4→3	21	…	…
-	-	7→3	20	…	…
-	-	…	…	…	…

The first two columns of table 3 show the standard collection of reasons for non-completion of activities according to the classification provided in Table 2. It simply provides a means of identifying the most common non-conformances. For example, problems related to labor and information were the most frequent with 1329 and 991 occurrences respectively.

The next two pairs of columns display the incipient results from the sequential analysis investigation. The majority of the sequences occurred between same category problems (e.g., labor problem in one activity followed by labor problem in subsequent activity took place 135 times – sequence 3→3 became evident in 135 occasions). This seems reasonably intuitive once several sequential assignments were planned to be accomplished within a day, and when problem was faced, it would impact most, if not all, planned activities during that journey. However, there were other instances where these phenomena did not happen and under such circumstances supervising personnel should be intrigued by such repetitive episodes.

In addition, those results might be expressing some relative association among the sequences. For example whenever a prerequisite problem was faced, that kind of problem was followed by another prerequisite problem in 14.4% (71/492) of the time and by a labor problem in 4% of the time (20/492). With more comprehensive sequence generation and analysis this would warn managers about the likelihood of prospective failures and what kind of failures these might be, such that they could be prepared in advance and avoid future problems.

Yet sequences of same category problems were not the only patterns that emerged from the database. There were some other instances where problems in one activity were followed by different errors, and even though they were less frequent than the former ones, they might impact the project's work flow considerably.

While the greater part of the frequent sequences seems to be common sense or trivial facts after they are revealed, the importance of such observations is that even not so frequent sequences might pinpoint relevant events that had not been noticed by onsite workforce. Furthermore, the high prevalence of some problems being followed by others might suggest that there is something more than just chance.

FURTHER INVESTIGATIONS AND FUTURE DIRECTIONS

The outcomes achieved so far represent not only a challenge but also are an encouragement for further and deeper analysis of the incidence of frequent sequences of reasons for failure to complete planned activities. Restrictions and obstacles have to be overcome and future steps of this ongoing investigation as well as envisioned directions are discussed in the subsequent lines.

The first consideration is that beyond the restriction imposed by the adoption of milestones there was one even more critical. Current sequential analysis algorithms make the assumption that each individual sequence of events (i.e., activities) is independent from the other. In the representation adopted, some sequences shared common subpaths, which accounted for a slightly higher occurrence on some of the sequences encountered. Take as an example the two sequences derived from the hypothetical schedule represented on Figure 2. The sequences of activities <010,020,070,100> and <010,020,050,070,100> are almost identical and they share the same subsequences <010,020> and <070,100>. This duplication of representation would be counted twice by the sequential analysis algorithm. Because of that effect, new representations for the extraction of activities sequences are being considered and recent studies in the data mining community have been conducted to overcome this limitation particularly related to identification of frequent patterns in graph or network structures (Inokuchi, Washio, and Motoda 2000, Yan and Han 2002, Yan and Han 2003).

Still, even when these constraints are mostly removed and the frequent sequences are accurately identified, the foremost and ultimate question at this point is how to analyze such sequences and how useful they would be in the identification of main causes of non-completion of activities. Although this is in fact a very ambitious and non-trivial objective, the most frequent chains of troublesome activities could be investigated by linking their reasons for non-completion with other project related data sources. First, the creation of such links would help to corroborate (or not) the reported reason. Moreover, tracing such links could help in the identification of main causes, and they might even provide evidence of the existence (or not) of some cause-effect relationships within sequences.

However, during the investigation of the activities' chains in this study, the insufficient availability of project wide data as well as a non-uniform representation of project entities among different systems was observed. For example, whenever a problem was reported to be related to insufficient labor, it was not possible to verify why a particular crew was assigned less man-hours than originally planned, once the daily production plans were not resourced within the production plan network (although resource loading is being considered as a feature in production planning by the time of this writing). Similar difficulties were faced with other problems, and even assessment of financial impacts of certain sequences was unsatisfactory, once there was no standard categorization or linkage among project components in diverse systems. Although most of the problems faced onsite were discussed and resolved during daily meetings, and some preventive measures were taken to avoid repeated failures (e.g., assure readiness for a task before task is planned to be performed), many solutions were not documented or easily verified from collected data. Most of the know-how and lessons-learned remained in the minds of managers or supervisors and there was no formal means of gathering this acquired experience.

This brings up an issue of utmost importance to the whole construction project data analysis research community: lack of integrity and cross-reference not only among different project databases but also between different projects and more appropriate structures to associate them to historical corporate-wide statistics for facilitated data analysis (FIATECH Capital Projects Technology Roadmapping Initiative, 2003).

Various studies have been performed successfully in the area of data integration, but most focus on a partial scope of application. Some examples are cost and schedule integration, schedule and project model integration (Aalami, Fischer and Kunz 1998), Industry Foundation Classes (IFC) - created by the International Alliance for Interoperability (IAI 1996), among others. A uniform or standardized data modeling that as smoothly as possible integrates and/or summarizes data about scheduling, cost, control, safety, quality, personnel, and other project related databases (e.g., weather or price indexes) and that enables the reference of matching project entities across multiple systems, is ideal for data retrieval and the above mentioned efforts ultimately will achieve that goal entirely.

In summary, various data quality problems (missing, unknown and incorrect values), heterogeneous data formats, and limited integration in construction project databases have been hindering the development of advanced data analysis in the AEC realm for a long period and a broader solution must be modeled and designed such that deeper, wider and more comprehensive construction project data analysis and knowledge discovery can be accomplished.

CONCLUSION

This paper presented a case study where sequences of non-completion of activities were identified. Though it is still too premature to state any conclusive response to the hypothesis that there are (or not) strong evidences that particular activities were the cause of problems downstream, such sequences are an indication that there might be some pattern among them beyond randomness.

The complete generation of such sequences, taking into consideration the minimization of the limitations encountered in this experimental work discussed above, together with combined analysis of other direct and indirect project data, seems to be unavoidable steps towards a more thorough comprehension of what aspects are more related to non-completion of activities. Successful achievement of this proposed approach would have a considerable impact on current project control strategies, especially to the Last Planner™ system of production control.

ACKNOWLEDGEMENTS

The authors would like to express their profound gratitude to Glenn Ballard from Strategic Project Solutions for his partnership that seeded the development of this study. Your assistance, collaboration, support, and provision of useful data, valuable information, helpful comments, among others, were of utmost significance.

Also the writers would like to thank the National Science Foundation for their support under Grant No. 0093841(CAREER). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do nor necessarily reflect the views of the National Science Foundation.

REFERENCES

Aalami, F. B., Fischer, M. and Kunz, J. C. (1998). AEC 4D-CAD production model: Definition and Automated generation. CIFE WP 052, Stanford University, California.

Agrawal, R., Imielinski T., and Swami A., R. (1993). "Mining Association Rules between Sets of Items in Large Databases". Proc. 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., 207-216.

Agrawal, R. and Srikant, R. (1995). "Mining sequential patterns". ICDE'95 - 11th International Conference on Data Engineering, Taipei, Taiwan, 3-14.

Ballard, G. (1997). "Improving Work Flow Reliability." Proceedings 7th Annual Conference International Group for Lean Construction, Berkeley, California, 275-286.

Ballard, G. (2000). "The Last Planner™ System of Production Control." PhD thesis, School of Civil Engineering, The University of Birmingham, Birmingham, U.K., 192pp.

FIATECH (2003), "Capital Projects Technology Roadmapping Initiative." Strategic Overview, Austin, TX, 70pp.

Inokuchi, A., Washio T., and Motoda H. (2000). "An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data". The 4th European Conference on Principles and Practices of Knowledge Discovery in Databases, Lyon, France.

Koskela, L. (1999). "Management of production in construction: a theoretical view." Proceedings 7th Annual Conference International Group for Lean Construction, Berkeley, California, 241-252.

Womack, J.P. and Jones, D.T. (1996). Lean Thinking: Banish Waste and Create Wealth in Your Corporation. Simon & Schuster, New York. 352 p.

Tommelein, I.D., Riley, D., and Howell, G.A. (1999). "Parade Game: Impact of Work Flow Variability on Trade Performance." ASCE, J. of Constr. Engrg. and Mgmt., 125 (5) 304-310, Sept/Oct Issue.

Yan, X., and Han, J. (2002). "Graph-based substructure pattern mining". The 2002 IEEE International Conference on Data Mining, Maebashi City, Japan.

Yan, X., and Han, J. (2003). "CloseGraph: Mining Closed Frequent Graph Patterns". The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C.