Analysis of Causality between Defect Causes Using Machine Learning

Eva Morales, Liam Chen
Massachusetts Institute of Technology, Cambridge, USA

***Abstract*—**Construction defects are major components that result in negative impacts on project performance including schedule delays and cost overruns. Since construction defects generally occur when a few associated causes combine, a thorough understanding of defect causality is required in order to more systematically prevent construction defects. To address this issue, this paper uses association rule mining (ARM) to quantify the causality between defect causes, and social network analysis (SNA) to find indirect causality among them. The suggested approach is validated with 350 defect instances from concrete works in 32 projects in Korea. The results show that the interrelationships revealed by the approach reflect the characteristics of the concrete task and the important causes that should be prevented.

***Keywords*—**Causality, defect causes, social network analysis, association rule mining.

I. Introduction

CONSTRUCTION defects should be identified and prevented for successful accomplishment of construction projects [1], [2]. However, it is generally difficult to identify the causes of a particular defect because the defect is not the outcome of a single cause, but occurs when a few associated causes combine [3]-[5]. For this reason, defect causality needs to be understood in order to prevent construction defects.

This paper aims to quantify causality between defect causes, and particularly, analyze their relationships based on conditional probability by utilizing ARM. Then, this paper utilizes SNA to evaluate the indirect causal effect of defect causes and to determine those causes that have the most effect on the occurrence of other causes.

II. Literature Review

In order to prevent construction defects, many studies have attempted to understand the root causes of defects. Many studies have attempted to elicit primary defect causes by means of analyzing the frequency of occurrence, and proposed efficient solutions for defect prevention. For example, [6] analyzed the frequency of defect occurrence for each type of defect and identified a major responsible party for quality problems. Reference [7] collected data from 153 contractors in Malaysia and analyzed the relative magnitude of defect causes considering frequency and cost. On the other hand, a few studies have analyzed causality between defect causes despite the notion that a defect is not the outcome of a single cause but the combination of several associated causes [3]-[5]. Reference [3] developed and proposed a causal model that includes error causes and their causality. Reference [4] formulated the taxonomy of defect causes using a fault-tree approach that allows understanding a mechanism of defect occurrence. While these studies placed importance on causality among defect causes, they are limited in that vagueness of elucidating defect causality is unavoidable and thus complex patterns of defect generation are not easily recognizable. To address this limitation, this paper aims to quantify causality among defect causes and quantify the causality based on conditional probability in order for practitioners to systematically identify the most serious causes.

III. Research Methodology

There are several approaches that can elicit relationship among factors, such as structural equation model (SEM), cross-impact analysis and ARM. Among them, ARM is known to be effective to analyzing a significant amount of data and showing causal relationship based on conditional probability. ARM can be applied to decision-making, process control, and many other applications [5]. For these reasons, several defect causality studies in different industries have strived to adopt ARM to find patterns of defect occurrence. Reference [8] applied ARM in order to provide efficient and effective solutions for detecting the root causes of defects in the manufacturing industry. In the process where several machines are executed in order to make a product, ARM evaluates the probability of being the root cause of each machine. This result contributes to knowing the relationship among machines and defective products. In the transportation industry, [9] identified causal relationships among defects on container cranes using ARM, and the relationships drawn by this paper consist of conditional and consequential parts that show patterns of defects. In these studies, the quantification of causality of defect demonstrated successful application of the conditional probabilistic approach in terms of applicability and effectiveness. Inspired by successful application of conditional probabilistic analysis of defects in other industries, this paper analyzes causality among causes of construction defects based on conditional probability in order to help practitioners better understand the patterns of construction defects and manage them efficiently.

A rule mining process can be divided into two steps: first, the algorithm investigates a database to find item sets (defect causes) that satisfy a predefined minimum ‘Support’; second, the rules are generated above a predefined minimum ‘Confidence’.

‘Support’ is the probability of the antecedent (i.e., i) and consequent (i.e., k) appearing together in the data. ‘Confidence’ is the conditional probability of the consequent given the

antecedent. These measures could reflect the relationship of defect causes in terms of co-occurrence. However, there is a limitation related to the Support-Confidence framework in that an item set with high Confidence should not be considered as i and k being highly correlated, but as having causality as the antecedent and consequent, respectively [10]. Therefore, the 'Lift' measure was introduced to overcome a limitation present in the 'Confidence' measure. 'Lift' represents how much the probability of k would increase if i were to occur. That is, 'Lift' can be regarded as a criterion for determining whether the causality between two items exists.

In terms of defect management, preventing a cause that has a high 'Lift' value means that others affected by the cause can reduce their probability of occurrence. In other words, managing a few causes by manipulating each of their probabilities is more efficient than controlling all causes. This concept, which focuses on discovering major causes, is useful for providing practitioners with an efficient method for managing defects. Accordingly, this paper quantifies causality among defect causes by this measurement of ARM.

As indicated above, a defect occurs when a few causes combine. Thus, a cause might have several relationships with other causes, even if they are not directly linked. That is, it is necessary to consider the fact that some causes indirectly affect other causes [11]. For example, in the case where i and j have influence on j and k, i and k can be considered to be indirectly related; that is, the causes of a defect form a network. However, ARM cannot accommodate the indirect relationship of causes. In order to compensate for this limitation, SNA is used to investigate the magnitude of the effect that belongs to the pairs of causes that are linked indirectly.

SNA evaluates a network that consists of a set of actors and a set of links that connect them [12]. Actors and their actions are considered to be interdependent rather than autonomous, and links between actors are routes for transferring resources [13].

SNA has a variety of metrics for analyzing the relationship between actors. The metrics are mainly calculated to determine the actor that is more central (plays a more important role) than other actors in a network [12]. In light of finding the centrality of an actor, three main centrality measures are provided by SNA literature: 'Degree', 'Betweeness', and 'Closeness'. Among the three fundamental measures of centrality,
'Closeness' is of interest in this research because it provides the means for quantifying an actor's contribution to the global network [11].

'Closeness' means the degree to which an actor is close to others in a network [12]. 'Closeness' is calculated by a sum of the geodesic (i.e., the shortest path) distances from an actor to all other actors. Based on this idea, [11] introduced the concept of probabilistic reachability that identifies the most probable causal path that connects two entities, rather the sum of all possible causal paths.

IV. Causality among Defect Causes

A total of 350 defect instances from concrete tasks are collected from several contractors in Korea in order to elicit causality among defect causes. The data comprise detailed information on defects discovered both during construction and maintenance stage. Through literature review and careful discussion with construction managers, 14 defect causes generated at the construction stage are identified as.

- C1: Careless Mistake of Labors
- C2: Interference by Other Tasks
- C3: Excess Test Results beyond the limit
- C4: Inadequate Construction Method
- C5: Inadequate Equipment
- C6: Inadequate Measurement
- C7: Inadequate Protection
- C8: Incompetent Labors
- C9: Incompliance with Procedures
- C10: Incorrect Execution from Specifications
- C11: Insufficient Review of Drawings
- C12: Lack of Supervision and Inspection
- C13: Lack of Training for Labors
- C14: Use of Inadequate Materials

In this paper, the analysis process consists of two main stages. The first stage places the focus on discovering the causal relationship among defect causes based on conditional probability using ARM. The number of causes that results in a defect varies by case. Following that, those defect instances are transformed into a sparse matrix in order to manage the defect causes described in the form of nominal variables. Based on the transformed data, an a priori algorithm from ARM is utilized, and this approach provides the rules that include the three measurements (i.e., 'Support', 'Confidence', and 'Lift') that satisfy predetermined minimum 'Support' and 'Confidence', both of which are set to a number close to zero in order to draw all potential rules in this study.

After generating the rules, the second stage applies 'Confidence' from ARM to the network analysis in order to estimate the indirect causality of the defect causes. Then, SNA is applied to assess how much each cause contributes to the global structure. Given that each pair of causes has a different magnitude of causality, this paper analyzes the networks by considering them as weighted networks. In this paper, 'Confidence' is first placed on the link between causes, and 'Reachability' is measured. Finally, CC for each cause is calculated by the sum of 'Net-Lift'. For implementation of SNA, association rules are converted into the form of a matrix and then UCINET [14] is used to measure 'Reachability'. Once 'Reachability' is calculated, 'Net-Lift' and 'CC' can be calculated, and finally, several meaningful patterns are analyzed.

V. Results Analysis

Table I lists the 'Lift' values between defect causes and CCs for each cause from concrete work. This result shows several meaningful patterns. As indicated in Table I, the cause defined as "Interference by Other Tasks (C2)" has the highest PC value among 14 causes at 29.2, which is closely followed by "Inadequate Protection (C7)" and "Inadequate Measurement (C6)" at 25.4 and 21.4, respectively. This means that they can be the initial points in the network; that is, managing these causes would considerably contribute to reducing the probability of

other causes at the start of this task, and potential defects during the task are less likely to occur in accordance with the 'Lift' implication. Moreover, the results reflect the characteristics of a concrete task. For example, it is expected for the tendency of defects to result in damaging those parts of the concrete that are relatively fragile and have lower performance. Because concrete performance depends on the quality of formwork and curing, concrete should be carefully preserved in order to maintain high performance and prevent that type of defects. Once concrete is not protected (C7), it is considerably more likely for the concrete to be damaged by other tasks. This can be imagined intuitively, but Table I indicates that the cause "Inadequate Protection (C7)" that has the most Lift is "Interference by Other Tasks (C2)" at 5.8. This means that the probability of concrete being damaged by other tasks increases over five times. Consequently, it can be stated that the number of defects from the concrete task might be affected by the interference of other tasks or inadequate protection. In the middle of a task, if a manager realizes that a certain defect cause might occur, he/she could follow the result and make the right decision to prevent the defect from occurring. For example, when a practitioner perceives that laborers are not offered the proper training for their responsibilities (C13), the practitioner should act in order to prevent "Incompetent Laborers (C8)" and "Insufficient Review Specifications (C11)," both of which have respective interrelationship values of 2.30 and 3.83 with "Lack of Training for Laborers (C13)." Because untrained laborers tend to not fully understand their responsibilities, they are more likely to make mistakes. Otherwise, they can cause several defects by executing tasks based on their knowledge or experience without reviewing the specifications thoroughly. In particular, with regard to C10, this cause makes the probability of C2 and C7, both of which have the highest PC among all causes, to increase 3.22 times and 3.34 times, respectively. Consequently, unless the managers do not administer only one problem, the effect of the problem linked with a few other causes could grow dramatically and result in a defect.

The result showed that causality based on Lift reflects the characteristics of concrete tasks. The results can provide practitioners engaged in concrete tasks with meaningful knowledge. First, the managers can refer to the results in order to determine which causes should be prevented relatively. "Interference by Other Tasks (C2)," "Inadequate Protection (C7)," and "Inadequate Measurement (C6)" have the highest CC values in this study. This means that these three causes influence other causes the most. If the managers do not administer these, the probability of other causes can increase, and consequently, causes linked with each other will result in leading defects. Second, when a manager realizes that a certain problem might contribute to a defect, he/she can act appropriately to prevent subsequent possible problems. The results showed that "Lack of Training for Laborers (C13)" has an influence on "Incompetent Laborers (C8)" and "Insufficient Review Specifications (C10)," and thus managers can expect those causes that should be administered first. Otherwise, because C10 might increase the probabilities of C2 and C7, which are the most powerful, several causes might occur and contribute to defects. Therefore, the results could be useful guides for finding subsequent potential problems.

VI. Conclusions

Thoroughly understanding defect causes is necessary for preventing defects that can result in a variety of problems in the construction industry. Unfortunately, managing all those causes that can contribute to a potential defect is significantly difficult for practitioners because a defect occurs when several causes interact with one another. Based on this recognition, this study aimed to quantify causality among defect causes in order to help practitioners find patterns of defect occurrence and manage the causes efficiently. To accomplish this goal, the causality between defect causes was estimated, which is referred to as Lift using ARM. Lift is the ability of a cause to increase the probability of another cause. That is, this measure represents how much the probability of a consequent would be increased when an antecedent appears. Moreover, based on this concept, SNA was introduced to accommodate the indirect relationship of causes. For the validation of this approach, a case study was conducted by applying the approach to 350 defect instances from a concrete task in 32 projects in Korea.

References

[1] Josephson, P., and Hammarlund, Y. (1999), "The Causes and Costs of Defects in Construction: A Study of Seven Building Projects", Automation in Construction, 8(6), 681-687.

[2] Mills, A., Love, P., and Williams, P. (2009), "Defect Costs in Residential Construction", Journal of Construction Engineering and Management, 135(1), 12-46.

[3] Love, P., Edwards, D., Irani, Z., and Walker, D. (2009), "Project Pathogens: The Anatomy of Omission Errors in Construction and Resource Engineering Project", Engineering Management, IEEE Transactions on, 56(3), 425-435.

[4] Aljassmi, H., and Han, S. (2013), "Analysis of Causes of Construction Defects Using Fault Trees and Risk Importance Measures", Journal of Construction Engineering and Management, 139(7), 870-880.

[5] Cheng, Y., Yu, W., and Li, Q. (2015), "GA-based multi-level association rule mining approach for defect analysis in the construction industry", Automation in Construction, 51, 78-91.

[6] Cui, J. (2011), "Analysis of Construction Quality Accident Causes of Public Buildings Based on Failure Study Theories", Computer and Management (CAMAN), 2011 International Conference on. IEEE, 1-4.

[7] Abdul-Rahman, H., Al-Tmeemy, S., Harun, Z., and Ye, M. (2014), The major causes of quality failures in the Malaysian building construction industry.

[8] Chen, W., Tseng, S., and Wang, C. (2005), "A novel manufacturing defect detection method using association rule mining techniques", Expert Systems with applications, 29(4), 807-815.

[9] Wang, Z., Hu, Xiong, and Chen, Z. (2007), "Mining Association Rules on Data of Crane Health-Condition Monitoring", International Conference on Transportation Engineering 2007, ASCE, 2054-2059.

[10] Brijs, T., Vanhoof, K., and Wets, G. (2003), "Defining interestingness for association rules", Institute of Information Theories and Applications FOI ITHEA, 10(4), 370-375.

[11] Aljassmi, H., Han, S., and Davis, S. (2014), "Project Pathogens Network: New Approach to Analyzing Construction-Defects-Generation Mechanisms", Journal of Construction Engineering and Management, 140(1).

[12] Freeman, L. (1978), “Centrality in social networks conceptual clarification”, Social networks, 1(3), 215-239.

[13] Wasserman, S., and Faust, K. (1994), Social network analysis: Methods and applications, Cambridge university press.

[14] Borgatti, S., Everett, M., and Freeman, L. (2002), Ucinet for Windows: Software for Social Network Analysis, Harvard, MA: Analytic Technologies.