

Article

Application of Association Rule Mining and Social Network Analysis for Understanding Causality of Construction Defects

Sangdeok Lee ¹, Yongwoon Cha ¹, Sangwon Han ^{1,*} and Changtaek Hyun ¹

Department of Architectural Engineering, University of Seoul, Seoul 02504, Korea; mgr_leesd@naver.com (S.L.); ywcha@uos.ac.kr (Y.C.); cthyun@uos.ac.kr (C.H.)

* Correspondence: swhan@uos.ac.kr; Tel.: +82-2-6490-2764

Received: 27 November 2018; Accepted: 22 January 2019; Published: 24 January 2019



Abstract: A construction defect can cause schedule delay, cost overrun and quality deterioration. In order to minimize these negative impacts of construction defects, this paper aims to analyze the causality of construction defects. Specifically, association rule mining (ARM) is used to quantify the interrelationships between defect causes, and social network analysis (SNA) is utilized to find out the most influential causes triggering generation of construction defects. The suggested approach was applied to 2949 defect instances in finishing work. Through this application, it was confirmed that the proposed approach can systematically identify and quantify causality among defect causes.

Keywords: causality; defect causes; association rule mining; social network analysis; finishing work

1. Introduction

A construction defect, which can have a negative impact on project performance such as schedule delay, cost overrun and quality deterioration, is a factor that should be prevented for successful accomplishment of a construction project [1,2]. Although there have been extensive research efforts to identify and eliminate major causes of construction defects, it is still challenging to precisely find out the main causes of a defect because a defect is not an outcome of a single cause, but occurs when a couple of associated causes combine [3–5]. Given that defect causes are interrelated in a complex manner, defect causality needs to be thoroughly understood for construction defects prevention [6].

In order to address this issue, recent studies [3,4] highlighted the importance of finding patterns of defect occurrence by understanding causality among defect causes. Love et al. [3] proposed a causal model to identify underlying causes that contribute to the occurrence of omission errors and helped to discover causality between defect causes. Aljassmi et al. [4], in addition, formulated a taxonomy of defect causes by using a fault-tree approach, which enables us to understand a mechanism of defect occurrence. These studies contributed to better understand causality among defect causes and to provide a theoretical foundation for subsequent research. Yet, a remainder challenge to applying these models is that they require establishing rigid data collection protocols prior to their utilization to compile practitioner's subjective judgments about defect causalities. Such extensive set-up efforts would not always be afforded by practitioners. Additionally, these studies do not yet bear a link to the exploitation of currently available construction defects databases, but demonstrate ground-up approaches for defect data collection and formulation.

Based on this recognition, this paper proposes a data-mining approach to more conveniently compile defect causality data from readily available datasets. Taking advantage of the proposed approach, this paper aims to quantify causality between defect causes. Specifically, the interrelationship among defect causes is analyzed on the basis of conditional probability by utilizing association rule

mining (ARM). Furthermore, social network analysis (SNA) is used to evaluate the direct and indirect causal effect of defect causes in order to determine the most influential defect causes.

2. Literature Review

Numerous studies have been conducted to identify and eliminate the major causes of defects in order to prevent construction defects. Some studies classified defect causes by estimating the relative magnitude of each of them and provided practitioners with useful knowledge for defect prevention. Josephson and Hammarlund [1] analyzed 2879 defects collected from 7 building projects and identified 5 different types of defect causes: knowledge, information, motivation, stress and risk. Jha and Iyer [7] identified causes (e.g., lack of harsh climatic condition at site, negative attitude of project participants, project manager's ignorance and lack of knowledge and so forth.) adversely affecting performance of construction quality through statistical analysis of questionnaire responses from professionals in about 50 large and medium size organizations in the Indian construction industry. Many researchers [8–11] tried to elicit primary defect causes by means of analyzing the frequency of occurrence and to propose efficient solutions for defect prevention. Cui [12] analyzed the frequency of defect occurrence for each types and identified a major responsible party for quality problems. Abdul-Rahman et al. [13] collected data from 153 contractors in Malaysia and analyzed the relative magnitude of defect causes considering frequency and cost.

On the other hand, there has been another type of research effort to analyze causality between defect causes. Love et al. [3] developed and proposed a causal model, which includes error causes and their causality. By analyzing interview data from 59 practitioners in Australia, influential factors on errors were extracted and their causal relationships were deduced. Aljassmi et al. [4] formulated a taxonomy of defect causes by using a fault-tree approach which allows us to understand a mechanism of defect occurrence. Aljassmi et al. [4] focused on identifying indirect causes (e.g., preconditions for defective acts, defective supervision and organizational influence) rather than direct causes, and quantified the significance of causes by calculating frequency and magnitude. Aljassmi et al. [4] argued that some causes lead to other causes, and indirect causes behind direct causes should be also removed in order to increase the possibilities of defect prevention. However, frequency and magnitude are limited to characterizing a direct influence on others, but do not accommodate the fact that some causes account for the existence of others in an indirect way [6]. Thus, Aljassmi et al. [6] proposed pathogenicity, which is a cause's ability to trigger other high-magnitude conditions, as an additional criterion to tackle the challenge, and introduced a complementary approach which captures causality of defect causes and quantifies the most influential causes in terms of pathogenicity. As an extension, Aljassmi et al. [14] further identified the most influential defect causes on the basis of the frequency, magnitude and pathogenicity by conducting a questionnaire survey of 106 industry professionals to grasp defect causes and their causality. While this approach has great potential to discover what to manipulate in order to minimize construction defects, it is still difficult to elucidate vagueness or subjectivity regarding defect causality because the causality is identified through a questionnaire survey.

There also have been similar research efforts in many other industries to identify and quantify defect causality [5]. These studies extensively utilized ARM due to its ability to show defect occurrence pattern by extracting causality among defect causes on the basis of conditional probabilistic analysis. For example, Chen et al. [15] applied ARM to provide efficient and effective solutions to detect root causes of defects in the manufacturing industry. In the process where a number of machines perform to make a product, ARM evaluates the probability of being the root cause for each machine. This result contributes to identifying relationships among machines and the defective products. Song et al. [16] proposed a method to examine defect associations in a software system through ARM searching for several interesting relationships (frequent patterns, associations, correlations, or potential causal structures). Identifying causal relationships among defects, ARM generates rules showing which defect may occur serially when a certain defect happens. Zhixin et al. [17], in the transportation

industry, identified patterns of defects by analyzing causal relationships among defects on container cranes. To minimize defects in the garment industry, defect patterns were extracted by ARM and they were used to identify and analyze the root cause of garment defects [18]. These studies have well demonstrated successful application of ARM to quantification of defect causality in terms of applicability and effectiveness. Generating a number of rules which represent causal relationship among interrelated factors, ARM enables practitioners to expect which event would be likely to happen when another event happens. Given that construction systems are tightly coupled systems where some events that occur in one part of the system may cause events in other parts [19], it is expected that ARM has a great potential to analyze causality among construction defects.

Based on this recognition, Lee et al. [20] introduced a theoretical foundation using ARM in order to analyze defect causality in construction. Extending the theoretical foundation proposed by [20], this paper aims to systematically identify and quantify interrelated causality among construction defects through analysis of 2949 defect instances observed in finishing work which has a decisive impact on the overall quality of a building. In addition, since this paper analyzes actual defect instances carefully reported by practitioners, instead of using a questionnaire survey, it is expected to obtain relatively objective results, and thus to overcome the vagueness and subjectivity issue raised to Aljassmi et al. [14].

3. Methodology

3.1. Association Rule Mining (ARM)

ARM is one of the data mining techniques used to elicit useful knowledge from tremendous databases [21]. ARM, using an apriori algorithm, provides rules in the form of ‘ $X \rightarrow Y$ ’, where X and Y are sets of items. X and Y can be regarded as the “If” part and the “Then” part respectively [5], which means the causality of X and Y . For example, ARM shows the rule that if X happens then, Y will have higher probability, that is, X could be described as a cause of Y because probability of Y is changed by manipulating X [22].

A rule mining process can be divided into two main steps. First, this algorithm investigates a database to find item sets (e.g., defect causes) satisfying predefined minimum *Support*. Second, rules are generated above predefined minimum *Confidence*. *Support*, *Confidence* criteria which are defined as follows:

$$\text{Support}(i \rightarrow j) = P(i \cap j) \quad (1)$$

$$\text{Confidence}(i \rightarrow j) = P(j|i) = \frac{P(i \cap j)}{P(i)} \quad (2)$$

Support is the probability that the antecedent (i.e., i) and the consequent (i.e., j) appear together in one instance. *Confidence* is the conditional probability of the consequent given the antecedent. These measures could reflect relationship of defect causes in terms of co-occurrence. However, *Confidence* is limited in that it does not take into account the baseline frequency of the consequence, which makes it misleading interdependence. In order to overcome the deficiency of *Confidence*, *Lift* measure was introduced in the late 90’s. *Lift* overcomes this limitation by dividing the *Confidence* (conditional probability) $\frac{P(i \cap j)}{P(i)}$ by the frequency of the consequent $P(j)$. In other words, $\text{Lift} = \frac{\text{Confidence}(i \rightarrow j)}{P(j)}$. As such, when the frequency of the consequent $P(j)$ is higher than *Confidence* ($i \rightarrow j$) (denominator > numerator), *Lift* becomes less than one, which means “ i and j appear less frequently together in the data than expected under the assumption of conditional independency” (refer to Brijs et al [23] for more details).

$$\text{Lift}(i \rightarrow j) = \frac{P(j|i)}{P(j)} = \frac{P(i \cap j)}{P(i)P(j)} \quad (3)$$

This represents how much the probability of j would increase if i happens. If $\text{Lift}(i \rightarrow j) > 1$ then, i and j are dependent and complementary. If $\text{Lift}(j \rightarrow i) = 1$, i and j are independent, and if $\text{Lift}(i \rightarrow j) < 1$,

i and j are substitutive. That is, *Lift* can be regarded as a criterion judging whether causality between two items exists or not. In terms of defect management, preventing a cause, which has high *Lift* value, means that others affected by the cause will reduce their probability of occurrence. In other words, it is more efficient to manage a couple of causes by manipulating each of their probabilities, rather than controlling all causes. This concept, which focuses on discovering major causes, is useful to provide practitioners with an efficient way to manage defects. Accordingly, this paper quantifies causality among defect causes by this measurement of ARM.

The process of quantifying causality among defect causes by using ARM in this paper is: (1) *Transformation into Sparse Matrix*: Generally, sparse matrix means the matrix which relatively has many '0'. ARM cannot deal with nominal variables, but the defect causes are defined in the form of a string of words (i.e., a nominal variable). Thus, the form of data should be converted into sparse matrix which has only '0' and '1' (see Table 1), which means whether each cause occurs or not. If a certain defect is caused only by 'Careless mistake of labors' and 'Interference by other tasks', it could be described in sparse matrix that the values of the two causes, which contribute to the defect, are '1' and those of others are '0'. (2) *Causality Elicitation*: Based on the data transformed into sparse matrix, an apriori algorithm is utilized. As mentioned above, this approach provides three measurements of *Lift*, one of measurements, means a climb rate of probability of consequent. For example, if the *Lift* ('Careless mistake of labors' \rightarrow 'Interference by other tasks') is 2, the probability of 'Interference by other tasks' would rise twice if 'Careless mistake of labors' occurs.

Table 1. Example of a sparse matrix.

Case No.	A	B	C	D
1	0	1	1	1
2	0	0	0	1
3	1	0	0	0
4	1	1	0	0
5	1	1	1	0

3.2. Social Network Analysis

As previously mentioned, a defect occurs when a couple of causes combine. Thus, a cause would have a great deal of relationships with other causes, even though they are not directly linked. That is, it is necessary to account for the fact that some causes 'indirectly' affect other causes [6]. For example, in case that i and j have influence on j and k , respectively (i.e., $i \rightarrow j \rightarrow k$), 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. To make up for this weak point, SNA is used to investigate the magnitude of effect belonging to pairs of causes linked indirectly.

SNA literally analyses a network which consists of a set of actors and a set of links connecting them [24]. Actors and their actions are considered to be interdependent rather than autonomous, and links between actors are routes for transferring resources [25]. As such, SNA, recently, has been applied to research on several industries (e.g., biology: [26]; markets: [27]; medical science: [28,29]) to find meaningful patterns for a certain purpose.

If relational data are prepared in the form of a matrix as in Table 2, those would be conveniently converted into a network. Figure 1 shows an example of a weighted graph and its relationships (links) between nodes (actors) have unique values based on Table 2. The weights in a weighted graph can be depicted by the thickness of a link. SNA can be used to analyze relational data, such as kinship patterns, community structure, interlocking directorships and so forth [28], and the relationship can be expressed as any numerical value (i.e., a weight) by placing the values having an unique meaning on the strength of relations, such as the degree of closeness among friends. From this point of view, in this paper, the *Confidence* from ARM can be allotted to the links between actors in a network and, by this, the level of causality (i.e., *Lift*) between causes can be calculated. For example, Figure 1 illustrates how

each causes differ in influence on the probability of each other. The probability of B will be changed to 100% when C appears. By contrast, that will be 50% when D appears (see Table 2 and Figure 1).

Table 2. Matrix of actor conditional probabilities (Confidences).

	A	B	C	D
A				
B	0.66			
C	1	1		
D	-	0.5	0.5	

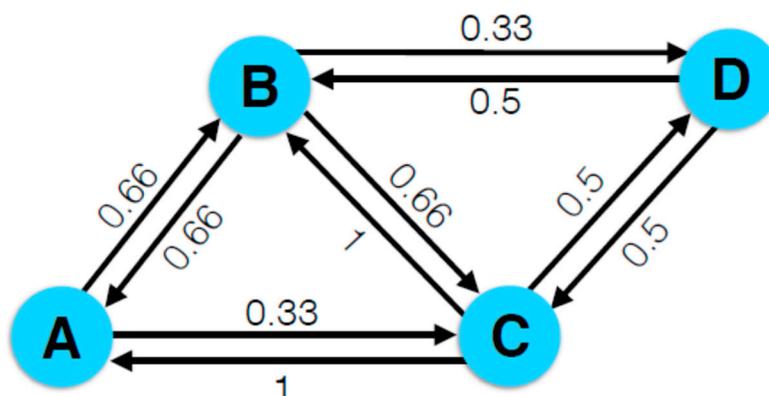


Figure 1. Example of a weighted graph based on Table 2 Confidence values.

Several metrics have been developed to analyze relationship between actors in a social network. These metrics are mainly used to determine which actor is more central (i.e., plays a more important role) than other actors in a network [24,30]. In light of finding the centrality of an actor, three main centrality measures have been utilized: *Degree*, *Betweenness* and *Closeness* (refer to Opsahl et al. [31] for a review of centrality measures). *Degree* centrality measures the degree of linkage (i.e., relationship) between a node and other nodes 'directly' linked to the node. *Betweenness* centrality measures how often a node occurs on all shortest paths between two nodes. On the other hand, *closeness* centrality measures the centrality not only from the directly linked nodes to a node in a network, but also that of the nodes that are indirectly linked to the node [32]. For this reason, *closeness* centrality is typically used for measuring how fast information will spread from one node in a network to all other nodes, or, in a network planning situation whose nodes are favorable starting points [33]. Given that construction systems are tightly coupled systems where some events that occurred in one part of the system may cause events in other parts [19], amongst the three fundamental measures of centrality, *Closeness* is of interest in this paper since it provide means for quantifying an actor's contribution to the global network [6].

Closeness was introduced by Bavelas [34] who argued that "a message originating in the most central position in a network would spread throughout the entire network in minimum time". Hakimi [35] and Sabidussi [36] also defined the most central actor in a network as the one consuming the minimum cost and time. In defect management, managing the cause, which has the highest *Closeness* in a network, would be the most relevant way in terms of the efficiency. *Closeness* is a metric to measure the degree to which an actor is close to others in a network [24]. *Closeness* has been measured by Sabidussi [36], which is one of several studies dealing with that, and this is regarded to be the simplest and most natural [24]. He proposed that *closeness* is calculated by sum of the geodesic (i.e., the shortest path) distances from an actor to all other actors. However, this has a difficulty in applying to a weighted network where each link between actors has a different amount of strength. Based on this recognition, a couple of studies have tried to quantify *Closeness* in a weighted network. Dijkstra [37] proposed an algorithm which discovers the path of least resistance, and asserted that this algorithm is

for networks in which the weights represent costs of transmitting. It means that the path having the least sum of weights is the best route for transmitting because the route costs the least. By contrast, Newman [38] and Brandes [39] inverted the weights for networks where weights represent positive strengths rather than resistance. In these papers, weights are inverted before applying Dijkstra’s algorithm to assess strengths of links. Those studies [37–39] allow to quantify *Closeness* in a weighted network. However, Aljassmi et al. [6] argued that they are not suitable to quantify probabilistically weighted links because causal paths follow the multiplication rule of probability. Thus, they introduced the concept of probabilistic reachability query from Zhu et al. [40]. *Reachability* quantifies upper-bound probabilities, which means it accounts for the most probable causal path connecting two entities, rather the sum of all possible causal paths (i.e., as in OR gates) [6]. The concept of *Reachability* makes the Markov assumption (i.e., the probability of going from one state to another depends only on the current state of the system, and thus is not influenced by additional information about past states). *Reachability* from an actor (*i*) to another actor (*k*) is defined as follows:

$$Reachability(i \rightarrow k) = \max[P(h|i) \times \dots \times P(k|h)] \tag{4}$$

Referring to the graph in Figure 1, D can be triggered by A through either $A \rightarrow B \rightarrow D$ or $A \rightarrow C \rightarrow D$, but $A \rightarrow B \rightarrow D$ is more probable. Therefore, $Reachability(A \rightarrow D) = 0.66 \times 0.33 = 0.22$. With *Reachability*, the indirect influence of defect causes can be estimated (see Table 3).

Table 3. Reachability matrix.

	A	B	C	D
A		0.66	0.44	0.22
B	0.66		0.5	0.33
C	1	0.5		0.5
D	0.5	0.5	0.5	

In this paper, a ‘*Net-Lift*’ (*NLift*) measure is introduced to overcome the limitation of the original *Lift*, which only accounts for direct causalities. *Net-Lift* simply replaces the numerator in Equation 3, $P(i | k)$, with *Reachability*: $\max[P(h|i) \times \dots \times P(k|h)]$. In light of this adjustment, *NLift* implies the degree to which an actor has direct, and also indirect, influence on the probabilities of other actors in a network. *NLift* can be formally expressed as:

$$NLift(i \rightarrow k) = \frac{\max[P(h|i) \times \dots \times P(k|h)]}{P(k)} \tag{5}$$

Proceeding with the above example, $NLift(A \rightarrow D) = 0.2178/0.4 = 0.544$. This infers that A and D are negatively dependent, which may be clearly observed from the raw data in Table 1. Thus, *Causal Closeness* (CC) can be considered as sum of *reachability* from an actor (*i*) to all other (*k*) actors (Aljassmi et al. 2014). Formally, CC can be defined as follows:

$$CC(i) = \sum_k^{all} Lift(i \rightarrow k) \tag{6}$$

Table 4 shows an example of matrix for actor interdependencies using *NLift*.

Table 4. Matrix for actor interdependencies using newly proposed “Net-Lift” measure.

	A	B	C	D	CC
A		1.1	-	-	1.1
B	1.1		-	-	1.1
C	-	-		-	-
D	-	-	1.25		1.25

4. Data Collection and Analysis

4.1. Data Collection

Finishing work is an important stage of construction because the quality of finishing work has a decisive impact on the overall quality of a building. We collected 2949 defect instances from finishing work from several contractors in Korea to elicit causality among defect causes in finishing work. The data comprise detail information on discovered defects both during construction and maintenance (e.g., causes, task, appearance, detail drawing, etc.).

In order to effectively analyze causes of defects in the collected data, defect causes first need to be systematically classified. Through a literature review and the analysis of the defect data with contractors, 21 defect causes were finally identified as shown in Table 5.

Table 5. Classification of defect causes.

Label	Defect Cause	References
C1	Lack of Training for Labors	[7,13,14]
C2	Incompetent Labors	[3,7,12,13,41–44]
C3	Careless Mistake of Labors	[3,12,14,42,45,46]
C4	Inadequate Measurement	[12,42,44]
C5	Incorrect Design Document	[13,14,43]
C6	Outdated/Damaged Material	[13,41,43]
C7	Inadequate Material Storage	[45,47]
C8	Use of Inadequate Materials	[13,41–43,45,47]
C9	Lack of Specification	[42]
C10	Test Results beyond the Allowance	[43]
C11	Inadequate Tools/Equipment	[14,41]
C12	Inadequate Maintenance of Equipment	[14,41]
C13	Incompliance with Equipment Manual	[14,41]
C14	Incorrect Execution from Specifications	[42]
C15	Absence of Testing/Inspection	[12,44,46]
C16	Incompliance with Procedures	[12,13,42]
C17	Inadequate Construction Method	[7,42,44,45,47]
C18	Insufficient Review of Drawings	[13,14,44]
C19	Lack of Supervision/Inspection	[3,12,41,43,44,46]
C20	Inadequate Protection	[12,13,41]
C21	Interference by Other Tasks	[3,14]

In this process, inevitable factors, such as weather and problems generated from design phase, are excluded. As shown in Table 5, some defect causes identified through a literature review look very similar to each other (e.g., C1 and C2). Initially, the authors considered reorganizing the defect causes to avoid confusion in the classification of defect causes. However, the authors finally decided to maintain the defect causes classification in order to compare the analysis results with the previous findings from the literature. Accordingly, the authors developed some rules for the classification of defect causes. For example, C1 (Lack of training of labors) is applied to defect cases where pre-checking, quality meeting or design document analysis between managers and labors before construction were not held due to schedule pressure. On the other hand, C2 (Incompetent labors) is applied to cases where

inadequate/unqualified labors were employed due to shortage of skilled labor or managerial effort to minimize labor cost.

Under these rules, some defect cases may be classified into improper defect causes. While such confusion in classification is one of the limitations of this paper, the authors do not believe that current classification in Table 5 may significant mislead the analysis results considering that this paper aims to analyze the ‘interrelated’ impact of defect causes on defect occurrence. Further consideration would be given for more systematic classification in the following study.

Figure 2 is an example of defect reports reported by one of the biggest general contractors in Korea. As mentioned above, the authors collected the defect reports from several contractors in Korea and there were using different format in their defect reports. Thus, the authors compiled the different defect reports and developed the database by incorporating only the common data type such as classification, brief description, and defect causes.

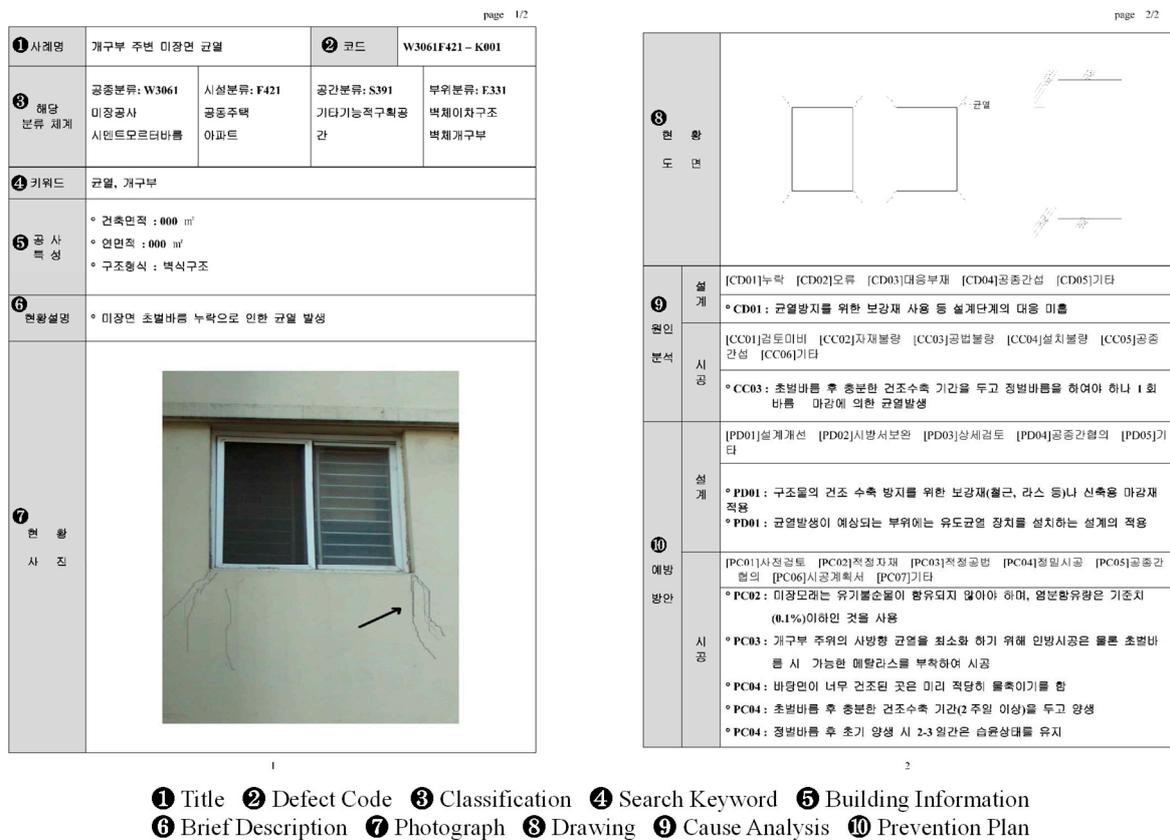


Figure 2. Example of defect report.

4.2. Data Analysis

Based on the classification of defect causes, an association rule mining was conducted to analyze causality of the 2949 defects from finishing work. Since 21 defect causes are considered (Table 5), the total number of possible combinations is 420 (i.e., 21 × 20). Among the 420 possible combinations, 366 combinations were observed through analysis of the 2949 defect instances. This means that no defect instances fall into the remaining 54 combinations. While it is possible to observe one of the remaining 54 combinations through additional analysis with new defect instances, it is expected that the impact of these 54 combinations would be less significant. With the 366 combinations (i.e., antecedent and consequent), their conditional probability metrics including *Support*, *Confidence* and *Lift* of each combination were calculated (Table 6).

Table 6. 366 Combinations in finishing work.

No.	Antecedent (X)	Consequent (Y)	Support	Confidence	Lift
1	C13	C9	0.001	0.600	5.608
2	C9	C13	0.001	0.115	5.608
3	C19	C9	0.004	0.478	4.470
4	C9	C19	0.004	0.423	4.470
5	C13	C1	0.001	0.600	3.645
6	C1	C13	0.001	0.075	3.645
7	C13	C5	0.002	1.000	2.761
8	C5	C13	0.002	0.057	2.761
9	C7	C1	0.005	0.424	2.577
10	C1	C7	0.005	0.350	2.577
11	C9	C7	0.003	0.346	2.549
12	C7	C9	0.003	0.273	2.549
13	C9	C1	0.003	0.385	2.337
14	C1	C9	0.003	0.250	2.337
15	C1	C15	0.011	0.775	2.242
16	C19	C3	0.007	0.957	1.291
...
366	C14	C19	0.002	0.093	0.986

Then, the *Confidence* from ARM to network analysis was applied to estimate indirect causality of defect causes. SNA is used to assess how much each cause contributes to the global structure. Since each pair of causes has a different magnitude of causality, this paper analyzes the networks considering them as weighted networks. Among measures that SNA provides, *Reachability* measures the effect of an actor (*i*) to another actor (*k*) in a network, and accommodates networks consisting of probabilistically weighted links. In this paper, first, *Confidence* is placed on the link between causes, and *Reachability* is measured. Following that, *NLift* is calculated on the basis of Equation (5). In addition, *NLift*, which is less than a value of 1, is excluded because this does not mean the causes are interrelated, as implied by *Lift* measure. Finally, the CC of each cause is calculated by sum of *NLift*. As for calculation, association rules, including the antecedent, the consequent and *Confidence*, are converted into the form of a matrix (refer to Table 1) to be applied to SNA, and then an algorithm, provided by the commercial UCINET software [48], is used to measure *Reachability*. Once *Reachability* is calculated, *NLift* and CC can be calculated, respectively, and finally, several meaningful patterns are analyzed. However, in this process, every causality would not be useful because the rules from ARM are drawn by the lowest threshold where minimum *Support* and minimum *Confidence* are almost zero. Due to the fact that the usefulness of rules depends on the threshold values for *Support* and *Confidence* [5], knowledge and experience of professionals are required to correctly interpret results in this study.

Table 7 shows the centrality of each defect cause which shows the relative importance of causality in 2,949 defects. That is, the centrality metric shows the magnitude of influence that a cause can bring other causes together. The centrality analysis results showed that Inadequate Protection (C20, centrality = 2.407) is the most influential cause among the 21 causes in the defect data. This is partially attributed to that the fact that finishing work consists of numerous job tasks (e.g., plastering, flooring, painting, wallpapering, and glazing) which take place within limited time and space, usually under schedule pressure. Accordingly, interruptions among these job tasks often happen and the interruptions may bring one job task (e.g., painting) to damage quality of other tasks (e.g., plastering or flooring) which have been successfully completed.

Table 7. Centrality of defect cause in finishing work.

Rank	Defect Cause	Centrality
1	Inadequate Protection (C20)	2.407
2	Lack of Specification (C9)	2.015
3	Interference by Other Tasks (C21)	1.920
4	Lack of Training for Labors (C1)	1.889
5	Inadequate Material Storage (C7)	1.642
6	Inadequate Tools/Equipment (C11)	1.595
7	Insufficient Review of Drawings (C18)	1.572
8	Incorrect Design Document (C5)	1.560
9	Lack of Supervision/Inspection (C19)	1.474
10	Incompliance with Procedures (C16)	1.433
11	Inadequate Maintenance of Equipment (C12)	1.419
12	Incompliance with Equipment Manual (C13)	1.397
13	Inadequate Measurement (C4)	1.393
14	Absence of Testing/Inspection (C15)	1.286
15	Careless Mistake of Labors (C3)	1.279
16	Use of Inadequate Materials (C8)	1.267
17	Outdated/Damaged Material (C6)	1.250
18	Incorrect Execution from Specifications (C14)	1.197
19	Incompetent Labors (C2)	0.986
20	Test Results beyond the Allowance (C10)	0.780
21	Inadequate Construction Method (C17)	0.000

The benefits of using SNA lie in the visualization of the complex interrelationships among the causes of defects. A social network analysis (SNA) diagram in Figure 3 illustrate the causality of defects in finishing work with the degree of centrality of each node (i.e., defect cause) depicted by the size of each node. As shown in Figure 3 and Table 7, Inadequate Protection (C20, centrality = 2.407) is the most influential cause of most concern, and is represented by the largest red circle in the social network analysis diagram. Also, Lack of Specification (C9, centrality = 2.015) and Interference by Other Tasks (C21, centrality = 1.920) were identified as the second and the third most influential causes, respectively. This means that most of defects in finishing work take place not due to quality or the technical problem of a given job task but due to inadequate protection (C20) ‘after’ successful completion of a job task; lack of specification (C9) ‘before’ execution of a job task; and interference by other concurrent job tasks (C21) ‘during’ execution of the job task. This result suggests that if a defect takes place in a job task in finishing work (e.g., plastering), its generation is more related with the construction environment and the construction managers responsible to provide a safe and productive construction environment, rather than with the job task itself and labors working on the job task.

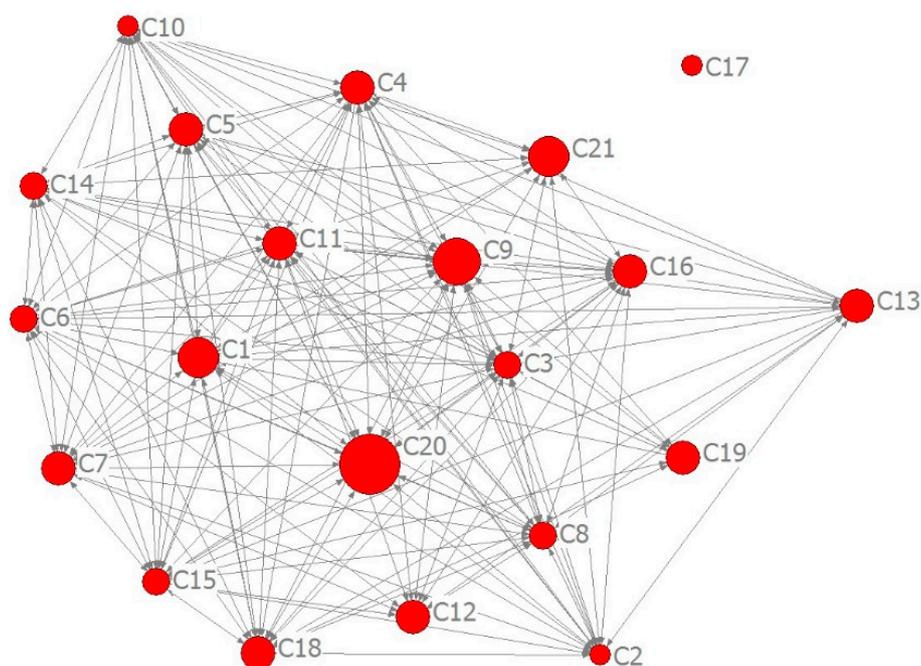


Figure 3. Social network analysis of defect causality.

Interestingly, Figure 3 shows that Inadequate Construction Method (C17) is isolated from the network. This means that Inadequate Construction Method rarely triggers defects in finishing work and its negative impact is rarely propagated to other related defect causes (note that the centrality of 'Inadequate Construction Method' is zero in Table 7). This is partially attributed to the fact that most finishing works are standardized and modularized by each finishing job task. More importantly, this result confirms that most of defects in finishing work take place not because of 'directly' related technical issues (e.g., inadequate construction method) but because of 'indirectly' related managerial issues (e.g., coordination, collaboration, or pre-checking). Thus, it is concluded that the suggested approach is effective to analyze the interrelated causality of construction defects in both a direct and an indirect manner. It is also concluded that the suggested approach can help construction managers set priorities in developing their defect prevention strategy.

4.3. Results Validation

A group interview was conducted with 3 industry experts to verify the validity of the analysis results. The interviewees consisted of one expert with more than 25 years and two experts with more than 10 years of experience in construction quality management. The authors first explained the research background, identification of defect causality, and determination of the most influential causes for defect prevention, and then asked the experts regarding the applicability of the research methodologies and effectiveness of the research findings.

To examine the validity of the research methodologies, the authors asked the experts regarding the applicability of the association rule mining technique and the centrality analysis of social network analysis. Also, regarding the effectiveness of the research findings, the authors asked the experts to verify the validity of the research findings based on their field experience as construction quality managers at the actual site. As for the validity of the elicitation procedures of the most influential defect causes through association rule mining technique, the experts have commented that the defect causes identified as the most influential in this paper are mostly consistent with the main causes that are usually managed at the actual site. Furthermore, the experts have suggested that a deeper analysis would be made if the authors take into account the precedence relationship between defect causes or resultant cost in determining the most influential defect causes. Through the group interview with the

experts, it is deemed that the research findings and the methodology suggested in this paper are valid in terms of applicability and effectiveness.

5. Conclusions

A defect is not an outcome of a single cause, but occurs when a couple of associated causes combine. Considering the practical and financial constraints imposed on construction firms, it is difficult for practitioners to identify and eliminate all possible causes of a potential defect. Thus, it is efficient and effective to optimize the defect prevention effort by identifying the most influential causes that have the highest potential of triggering defects. To address this issue, this paper quantified causality among defect causes through the application of association rule mining and social network analysis.

The suggested approach was applied to 2949 defect instances generated in finishing work; 21 defect causes were identified for classification of the defects and 366 combinations between the defect causes through association rule mining on the basis of Support, Confidence and Lift. The centrality of each node (i.e., cause) in the resulting graph (represented as a sparse matrix) was calculated in order to identify the most influential defect cause. The analysis results revealed that the three most influential defect causes in finishing work are ‘Inadequate Protection’, ‘Lack of Specification’ and ‘Interference by Other Tasks’. This result indicates that defects in finishing work are more likely to be caused by construction managers responsible for providing a safe and productive construction environment than by construction workers performing the finishing work. Also, the analysis results revealed that ‘Inadequate Construction Method’ is the least influential defect cause. This result reconfirmed that most defects in finishing work take place not because of ‘directly’ related technical issues but because of ‘indirectly’ related managerial issues (e.g., coordination, collaboration, or pre-checking). Based on these findings, it is concluded that the suggested approach is effective to analyze interrelated causality of construction and to set priorities in developing a defect prevention strategy.

While the suggested approach has a great potential to analyze the causality of construction defects, this approach has some limitations that should be addressed in following studies. Firstly, defect causes should be more systematically classified. For this, some defect causes might be reorganized and a set of rigorous classification rules should be prepared with the help of industry experts. Secondly, the usefulness of association rules significantly depends on the threshold values for *Support* and *Confidence*. Further research efforts should be devoted to finding optimal values for *Support* and *Confidence* so that more effective association rules can be obtained. Lastly but most importantly, the suggested approach should be further validated with more defect cases reported from different construction environments and contexts.

Author Contributions: Conceptualization, S.L. and S.H.; Methodology, S.L. and Y.C.; Validation, Y.C. and S.H.; Formal Analysis, S.H.; Investigation, C.H.; Data Curation, Y.C. and C.H.; Writing-Original Draft Preparation, S.L.; Writing-Review and Editing, S.W. and C.H.; Visualization, S.L. and Y.C.; Supervision, S.H.

Funding: This work was supported by the 2015 Research Fund of the University of Seoul.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Josephson, P.E.; Hammarlund, Y. Causes and costs of defects in construction a study of seven building projects. *Autom. Constr.* **1999**, *8*, 681–687. [[CrossRef](#)]
2. Mills, A.; Love, P.E.; Williams, P. Defect Costs in Residential Construction. *J. Constr. Eng. Manag.* **2009**, *135*, 12–16. [[CrossRef](#)]
3. Love, P.E.D.; Edwards, D.J.; Irani, Z.; Walker, D.H.T. Project pathogens: The anatomy of omission errors in construction and resource engineering project. *IEEE Trans. Eng. Manag.* **2009**, *56*, 425–435. [[CrossRef](#)]
4. Aljassmi, H.; Han, S. Analysis of Causes of Construction Defects Using Fault Trees and Risk Importance Measures. *J. Constr. Eng. Manag.* **2013**, *139*, 870–880. [[CrossRef](#)]

5. Cheng, Y.; Yu, W.D.; Li, Q. GA-based multi-level association rule mining approach for defect analysis in the construction industry. *Autom. Constr.* **2015**, *51*, 78–91. [[CrossRef](#)]
6. Aljassmi, H.; Han, S.; Davis, S. Project Pathogens Network: New Approach to Analyzing Construction-Defects-Generation Mechanisms. *J. Constr. Eng. Manag.* **2014**, *140*, 04013028. [[CrossRef](#)]
7. Jha, K.N.; Iyer, K.C. Critical factors affecting quality performance in construction projects. *Total Qual. Manag. Bus. Excell.* **2006**, *17*, 1155–1170. [[CrossRef](#)]
8. Kim, H.J. The Analysis of Defects Type and Cause in High-rise Residential Buildings. Master Thesis, Dong-A Univ., Busan, Korea, 2010.
9. Hong, S.I.; Hyun, C.T.; An, S.B.; Ji, S.M.; Son, M.J. Selection of Primary Management Objects for Defect Prevention. *J. Archit. Inst. Korea* **2011**, *27*, 185–194.
10. Lee, H.J. *A Study on Defects Occurrence and Preventive Measures in Apartment Building Construction*; Hanyang University: Seoul, Korea, 2011.
11. Bae, S.I.; Shim, U.J.; Ahn, Y.S. The Study on the Selection of Primary Management Objects for Defect Prevention of Finishing Works on Apartment House. *J. Archit. Inst. Korea* **2013**, *15*, 179–186. Available online: http://www.materic.or.kr/info/scholar/content.asp?s_code=JP&p_id=156732 (accessed on 1 June 2013).
12. Cui, J. Analysis of construction quality accident causes of public buildings based on failure study theories. In Proceedings of the 2011 International Conference on Computer and Management (CAMAN), Wuhan, China, 19–21 May 2011; pp. 1–4. [[CrossRef](#)]
13. Abdul-Rahman, H.; Al-Tmeemy, S.; Harun, Z.; Ye, M. *The Major Causes of Quality Failures in the Malaysian Building Construction Industry*; Foundation of Technical Education: Baghdad, Iraq, 2012; pp. 1–20. Available online: <http://www.fte.edu.iq/upload/upfile/ar/122212.pdf> (accessed on 1 January 2012).
14. Aljassmi, H.; Han, S.; Davis, S. Analysis of the Complex Mechanisms of Defect Generation in Construction Projects. *J. Constr. Eng. Manag.* **2015**, *142*, 04015063. [[CrossRef](#)]
15. Chen, W.C.; Tseng, S.S.; Wang, C.Y. A novel manufacturing defect detection method using association rule mining techniques. *Expert Syst. Appl.* **2005**, *29*, 807–815. [[CrossRef](#)]
16. Song, Q.; Shepperd, M.; Cartwright, M.; Mair, C. Software defect association mining and defect correction effort prediction. *IEEE Trans. Softw. Eng.* **2006**, *32*, 69–82. [[CrossRef](#)]
17. Wang, Z.; Hu, X.; Chen, Z. Mining association rules on data of crane health-condition monitoring. In Proceedings of the International Conference on Transportation Engineering, Chengdu, China, 22–24 July 2007; ASCE: Reston, VA, USA, 2007; pp. 2054–2059.
18. Lee, C.K.H.; Choy, K.L.; Ho, G.T.S.; Chin, K.S.; Law, K.M.Y.; Tse, Y.K. A hybrid OLAP-association rule mining based quality management system for extracting defect patterns in the garment industry. *Expert Syst. Appl.* **2013**, *40*, 2435–2446. [[CrossRef](#)]
19. Perrow, C. *Normal Accidents: Living with High-Risk Technologies*; Princeton University Press: Princeton, NJ, USA, 2011.
20. Lee, S.; Han, S.; Hyun, C. Analysis of Causality between Defect Causes Using Association Rule Mining. *Int. J. Civ. Environ. Eng.* **2016**, *10*, 659–662.
21. Kamsu-Foguem, B.; Rigal, F.; Mauget, F. Mining association rules for the quality improvement of the production process. *Expert Syst. Appl.* **2013**, *40*, 1034–1045. [[CrossRef](#)]
22. Pearl, J. *Causality*; Cambridge University Press: New York, NY, USA, 2009.
23. Brijs, T.; Vanhoof, K.; Wets, G. Defining Interestingness for Association Rules. *Int. J. Inf. Theory Appl.* **2003**, *10*, 370–376.
24. Freeman, L.C. Centrality in social network concept clarification. *Soc. Netw.* **1978**, *1*, 215–239. [[CrossRef](#)]
25. Faust, S.W.K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994.
26. Wey, T.; Blumstein, D.T.; Shen, W.; Jordán, F. Social network analysis of animal behaviour: A promising tool for the study of sociality. *Anim. Behav.* **2008**, *75*, 333–344. [[CrossRef](#)]
27. Burt, R.S. The Stability of American Markets. *Am. J. Sociol.* **1988**, *94*, 356–395. [[CrossRef](#)]
28. Weeks, M.R.; Clair, S.; Borgatti, S.P.; Radda, K.; Schensul, J.J. Social networks of drug users in high-risk sites: Finding the connections. *AIDS Behav.* **2002**, *6*, 193–206. [[CrossRef](#)]
29. Christley, R.M.; Pinchbeck, G.L.; Bowers, R.G.; Clancy, D.; French, N.P.; Bennett, R.; Turner, J. Infection in social networks: Using network analysis to identify high-risk individuals. *Am. J. Epidemiol.* **2005**, *162*, 1024–1031. [[CrossRef](#)] [[PubMed](#)]

30. Bonacich, P. Power and Centrality: A Family of Measures. *Am. J. Sociol.* **1987**, *92*, 1170–1182. [[CrossRef](#)]
31. Opsahl, T.; Agneessens, F.; Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc. Netw.* **2010**, *32*, 245–251. [[CrossRef](#)]
32. Kwon, Y.; Jeong, D.; Moon, Y.; Yoo, J. Comparing Analysis Study of Centrality Indices using Paper Information on Secondary Battery. *Indian J. Sci. Technol.* **2015**, *8*, 333–339. [[CrossRef](#)]
33. Davidsen, S.A.; Padmavathamma, M. A fuzzy closeness centrality using address-direction to control degree of closeness. In Proceedings of the 2014 First International Conference on Networks & Soft Computing, Guntur, India, 19–20 August 2014; pp. 203–208.
34. Bavelas, A. A Mathematical Model of Group Structures. *Hum. Organ.* **1948**, *7*, 16–30. [[CrossRef](#)]
35. Hakimi, S.L. Optimum Locations of Switching Centers. *Oper. Res.* **1966**, *12*, 450–459. [[CrossRef](#)]
36. Sabidussi, G. The centrality index of a graph. *Psychometrika* **1966**, *31*, 581–603. [[CrossRef](#)]
37. Dijkstra, E.W. A Note on Two Problems in Connexion with Graphs. *Numer. Math.* **1959**, *1*, 269–271. [[CrossRef](#)]
38. Newman, M.E.J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **2001**, *64*, 7. [[CrossRef](#)]
39. Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **2001**, *25*, 163–177. [[CrossRef](#)]
40. Zhu, K.; Zhang, W.; Zhu, G.; Zhang, Y.; Lin, X. BMC: An efficient method to evaluate probabilistic reachability queries. In *International Conference on Database Systems for Advanced Applications (DASFAA)*; Springer: Berlin, Germany, 2011; Volume 6587, pp. 434–449. Available online: https://doi.org/10.1007/978-3-642-20149-3_32 (accessed on 24 January 2019). [[CrossRef](#)]
41. Fayek, A.R.; Dissanayake, M.; Campero, O. *Measuring and Classifying Construction Field Rework: A Pilot Study*; Construction Owners Association of Alberta (COAA) Field Rework Committee: Alberta, Canada, 2003; Available online: <https://www.coaa.ab.ca/COAA-Library/COP-RRT-RPT-01-2003-v1%20Measuring%20and%20Classifying%20Construction%20Rework%20-%20Final%20Report.pdf> (accessed on 24 January 2019).
42. Hwang, B.-G.; Thomas, S.R.; Haas, C.T.; Caldas, C.H. Measuring the Impact of Rework on Construction Cost Performance. *J. Constr. Eng. Manag.* **2009**, *135*, 187–198. [[CrossRef](#)]
43. Love, P.E.D.; David, J. Edwards Forensic Project Management: The Underlying Causes of Rework in Construction Projects. *Civ. Eng. Environ. Syst.* **2004**, *21*, 207–228. [[CrossRef](#)]
44. Sun, M.; Meng, X. Taxonomy for change causes and effects in construction projects. *Int. J. Proj. Manag.* **2009**, *27*, 560–572. [[CrossRef](#)]
45. Burati, J.L.; Farrington, J.J.; Ledbetter, W.B. Causes of Quality Deviations in Design and Construction. *J. Constr. Eng. Manag.* **1992**, *118*, 34–49. [[CrossRef](#)]
46. Busby, J.S.; Hughes, E.J. Projects, pathogens and incubation periods. *Int. J. Proj. Manag.* **2004**, *22*, 425–434. [[CrossRef](#)]
47. Arditi, D.; Elhassan, A.; Toklu, Y.C. Constructability Analysis in the Design Firm. *J. Constr. Eng. Manag.* **2002**, *117*–126. [[CrossRef](#)]
48. Borgatti, S.P.; Everett, M.G.; Freeman, L.C. *Ucinet for Windows: Software for Social Network Analysis*; Analytic Technologies: Harvard, MA, USA, 2002.

