Automation in Construction 74 (2017) 39–54

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Construction Safety Clash Detection: Identifying Safety Incompatibilities among Fundamental Attributes using Data Mining

Antoine J.-P. Tixier [a], Matthew R. Hallowell [b,*], Balaji Rajagopalan [b], Dean Bowman [c]

[a] Computer Science Laboratory, École Polytechnique, Palaiseau, France
[b] Civil Engineering Department, University of Colorado at Boulder, CO, USA
[c] Bentley Systems, USA

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form 28 October 2016
Accepted 2 November 2016
Available online 21 November 2016

Keywords:
Building Information Modeling
BIM
Advanced Work Packaging
AWP
Safety
Machine learning
Network analysis
Risk
Prevention through design
Information Technology

ABSTRACT

Construction still accounts for a disproportionate number of injuries, inducing consequent socioeconomic impacts. Despite recent attempts to improve construction safety by harnessing emerging technologies and intelligent systems, most frameworks still consider tasks and activities in isolation and use secondary, aggregated, or subjective data that prevent their widespread adoption. To address these limitations, we used a newly introduced conceptual framework and accompanying natural language processing system to extract standard information in the form of fundamental attributes from a set of 5298 raw accident reports. We then applied state-of-the-art data mining techniques to discover attribute combinations that contribute to injuries. We refer to these incompatibilities as "construction safety clashes". The main contribution of our study lies in the methodological advancements that it brings to the construction safety domain. In light of the results obtained, our approach shows great promise to become a standard way of extracting valuable, actionable insights from injury reports in a fully unsupervised way. The use of our methodology could enable construction practitioners to ground their safety-related decisions on objective, empirical data, rather than on limited personal experience or expert opinion, which is the current industry standard. Finally, our methodology allows construction accidents to be viewed as perturbations in underlying networks of fundamental attributes. While the analysis of the current data set provides preliminary evidence for this theory, comparison to non-accident reports will be required for validation.

1. Introduction and Motivation

Even though safety performance has notably improved after the inception of the Occupational Safety and Health Act (OSHA) of 1970, construction fatalities, disablements, and illnesses still have a dramatic socioeconomic impact. In fact, construction still accounts for a fatal occupational injury rate of 9.4 per 100,000 full-time workers, one of the highest in the United States [10]. Moreover, the construction industry has consistently accounted for the most fatalities of any industry in the private sector since 2005, with 796 casualties in 2013 alone. Therefore, improving safety has become an absolute priority.

Construction has reached saturation with respect to the traditional safety strategies that were originally implemented to comply with regulations [25]. Therefore, safety researchers and professionals have recently tried to harness emerging technologies and intelligent systems that are traditionally used for design, planning, or operations. Some examples of such technologies include Building Information Modeling (BIM), proximity sensing, or information retrieval. While these efforts are worthy, they currently suffer limitations, as the data used are mostly secondary, aggregated, and subjective (based on regulations, intuition, or judgment), and tasks are considered in isolation, preventing the efficient capture of the transient and dynamic nature of construction work [65].

To improve the robustness of safety analyses, Esmaeili and Hallowell [26,27] and Esmaeili [23] introduced a conceptual framework where any injury can be characterized by a unique combination of universal context-free descriptors of the work environment, also called fundamental attributes or injury precursors. These works made great strides by showing possible the extraction of objective, standardized structured information from unstructured injury reports, opening the gate for the first time to leveraging big, empirical, and objective safety-related data. However, several major limitations remained, such as the needs for a more comprehensive set of attributes and for an automated system to scan the reports. Prades Villanova [65] and Desvignes [21] addressed the first limitation by proposing a refined and expanded list of fundamental attributes, and Tixier et al. [77] addressed the second by developing a highly accurate (96% in *F*1 score) natural language processing (NLP) system.

In this study, we tested the extent to which graph mining and hierarchical clustering can be used to identify safety-critical associations of attributes from large data sets. We conducted our experiments on an

* Corresponding author.
E-mail address: matthew.hallowell@colorado.edu (M.R. Hallowell).

http://dx.doi.org/10.1016/j.autcon.2016.11.001
0926-5805/

attribute data set obtained from scanning 5298 raw injury reports with Tixier et al.'s [77] NLP system.

2. Background and Point of Departure

This study was built upon a foundation of knowledge in two key areas: construction safety analysis, and safety integration with BIM. Although both of these areas have received some attention from the scientific and practical communities, researchers have yet to explore their nexus. The following literature review highlights current limitations in both domains and develops a firm point of departure.

2.1. Construction safety analysis

Safety analysis in construction has taken many forms and varies greatly in the *data sources* used and the *level of detail of the units of analysis (data granularity)*.

2.1.1. Data sources

The vast majority of construction safety studies rely on opinion-based risk data, generally obtained by asking experts to rate the relative magnitude of risk based on their professional experience and intuition [65]. Such data are subjective and suffer the numerous biases that affect human judgment under uncertainty, such as overconfidence, anchoring, availability, representativeness, unrecognized limits, or conservatism [11,68,78]. Additionally, there is evidence that gender [37] and even emotional state [76] impact risk perception. Although one can attempt to minimize the effects of some of these psychological biases [39], opinion-based data remain severely limited in comparison to empirical data. Therefore, the needs to leverage objective raw empirical data are pressing.

2.1.2. Level of detail of the units of analysis

Construction work is very complex from both technological and organizational perspectives. Even though the multifactorial nature of safety risk is well known [41,70], most studies have decomposed construction processes into smaller parts for the sake of simplicity [58]. Such breakdown allows researchers to model safety for a variety of units of analysis. For example, Hallowell and Gambatese [40] focused on specific worker motions and activities needed for formwork construction, Navon and Kolton [60] analyzed interactions among planned tasks at height, and Huang and Hinze [45] modeled task, location, time, human error, and age as risk factors. Trades have most commonly been adopted as the granularity level [4,28,49]. A limitation of these segmented approaches that consider elements in isolation is that there are a virtually infinite number of units of analysis that must be taken into account in order to comprehensively capture safety. This has prevented the adoption of a robust, standardized way of approaching safety analysis in construction.

2.1.3. Attribute-based approach to construction safety analysis

The attribute-based framework for construction safety was introduced by Esmaeili and Hallowell [26,27] and Esmaeili [23] in an effort to jointly address the data subjectivity and study segmentation limitations previously described. Indeed, this unified approach allows the extraction of standardized safety information from objective, raw textual data such as injury reports. Fundamental attributes are universal, context-free descriptors of the jobsite. They span construction means and methods, environmental conditions, and human factors.

To illustrate, in the following report excerpt: "employee tripped on an electrical cord while exiting job trailer", three fundamental attributes can be identified: (1) *object on the floor*, (2) *exiting/transitioning*, and (3) *job trailer*.

While simple, this approach is powerful, as any incident can be viewed as the resulting outcome of the joint occurrence of some fundamental attributes and the presence of a worker. It follows that the same standard safety information can be extracted for any construction situation regardless of the trade, task, industry sector, or part of the world in which the accident occurred.

Esmaeili and Hallowell [26,27] initially proposed short lists of fundamental attributes (14 and 34, respectively) identified from analyzing 105 fall and 300 struck-by high severity injury cases drawn from national databases. Prades Villanova [65] and Desvignes [21] refined and broadened these drafts to a final, robust list of 80 carefully engineered and validated attributes by manually analyzing a larger database of 2201 injury reports featuring all injury types and severity levels. These precursors are summarized in Table 1.

However, while the attribute-based framework is particularly well-suited for leveraging big textual safety-related data, the high cost and numerous limitations of manual content analysis remained as serious obstacles to its large-scale implementation. To solve this problem, Tixier et al. [77] developed a NLP tool that can automatically extract the 80 attributes presented in Table 1 and various safety outcomes with high accuracy (96% in $F1$ score). In this study, for illustration purposes (proof of concept), we apply our methodology on an attribute data set extracted from a pool of 5298 raw injury reports by the aforementioned NLP tool.

2.2. Modeling and managing safety in BIM

Among many characterizations, we refer to Building Information Modeling (BIM) as an information-rich design technology that can be used to generate a virtual model of an infrastructure. The strength of the BIM technology stems from its ability to augment the 3D representation of a facility with a plethora of information such as schedule,

Table 1
Attribute counts in our data set.

UPSTREAM	*Count*	Rebar	155	Screw	37
Cable tray	48	Scaffold	300	Slag	75
Cable	75	Soffit	12	Spark	9
Chipping	34	Spool	52	Slippery surface	142
Concrete liquid	58	Stairs	137	Small particle	401
Concrete	165	Steel sections	759	Adverse low temperatures	123
Conduit	56	Stripping	114	Unpowered tool	611
Confined workspace	129	Tank	85	Unstable support/surface	8
Congested workspace	13	Unpowered transporter	53	Wind	109
Crane	69	Valve	79	Wrench	110
Door	85	Welding	200	Lifting/pulling/manual handling	553
Dunnage	29	Wire	131	Light vehicle	133
Electricity	3	Working at height	268	Exiting/transitioning	132
Formwork	143	Working below elevated wksp/material	50	Sharp edge	47
Grinding	133	Drill	97	Splinter/sliver	41
Grout	18	**TRANSITIONAL**		Repetitive motion	66
Guardrail/handrail	91	Bolt	186	Working overhead	14
Heat source	111	Cleaning	119	**DOWNSTREAM**	
Heavy material/tool	79	Forklift	39	Improper body position	88
Heavy vehicle	143	Hammer	149	Improper procedure/inattention	57
Job trailer	24	Hand size pieces	172	Improper security of materials	87
Lumber	252	Hazardous substance	156	Improper security of tools	28
Machinery	189	Hose	95	No/improper PPE*	23
Manlift	66	Insect	105	Object on the floor	174
Stud	31	Ladder	163	Poor housekeeping	2
Object at height	86	Mud	35	Poor visibility	12
Piping	388	Nail	94	Uneven walking surface	59
Pontoon	15	Powered tool	239		

* Personal Protective Equipment.

specifications, and cost. It has been shown that BIM helps improve design, management, and construction operations and is beneficial for all stakeholders during the entire construction process [35,52,53].

Numerous efforts have focused on the integration of safety in BIM. For instance, BIM was combined with augmented reality to improve safety recommendation understanding [57,83], and with opinion-based risk information to assist safety management for scaffolding [17]. Hammad et al. [42] proposed a method to automatically detect risks of falls and dynamically add fences, and laser scanning technology enabled missing safety components such as guardrails or nets to be flagged by comparing virtual designs to actual structures [15]. BIM has also been paired with tracking technologies like the GPS to send alerts to workers when they enter predefined hazardous zones [13,18,33].

In the industry, there is preliminary evidence from an active Construction Industry Institute (CII) research team that advanced work packaging (AWP) maturity correlates with safety performance [64]. A possible explanation lies in that AWP goes beyond a virtual BIM model to describe not only the model component that gets built but also *how* it gets built in terms of specific, quantifiable work tasks. The latter is particularly well suited to safety clash detection because the work task granularity of work packages directly relates to describing those construction attributes pertinent to safety, significantly more than what would be indicated by a single BIM component.

Yet, no study has leveraged empirical data and produced results that can be used in BIM to identify what features of work are dangerous, when, where, and why. The present study is a first step in that direction. Here, we focus on BIM and AWP as candidate technologies because they presently pose greatest potential for implementation of our methodology and results. Actual implementation potential is extensive and will continue to broaden as technologies are introduced and mature.

2.3. Point of departure

In this paper, we are interested in testing the extent to which data mining can be used to extract valuable new safety knowledge from large attribute data sets, in the form of safety-critical combinations of attributes, or "safety clashes". To this end, we compare two complementary state-of-the-art unsupervised machine learning (ML) families of techniques, graph mining and hierarchical clustering on principal components (HCPC), on an attribute data set obtained from scanning 5298 unstructured injury reports with Tixier et al.'s [77] NLP tool.

We define "construction safety clashes" as *incompatibilities among fundamental attributes of the work environment that contribute to construction injuries*. In this definition, we consider clashes to be situations where a group of attributes produce greater risk than simply the "sum of their parts". In these situations, the attribute combinations magnify risk and, in some case, pose new threats. A simplistic example of a safety clash is *confined workspace* and *small particle*, which is considered a clash because the aggregate of the two attributes poses a greater threat than the two attributes in isolation. While very useful for live onsite safety management, such information, based on binary input variables, is also ideally suited to be integrated with new technologies like BIM to proactively flag and address safety-critical situations, thereby aiding prevention through design and the release of safer work packages. While all safety clashes are obviously of interest and would need to be accounted for in any BIM-based solution, in this exploratory study, we are mostly interested in discovering safety clashes that are not already well-known and that would not clearly emerge based on the experience of any one person alone.

Esmaeili and Hallowell [26] represented the co-occurrence among fundamental attributes as networks. More precisely, they investigated hazardous connections in 105 fatal fall reports from the National Institute for Occupational Safety and Health (NIOSH) Fatality Assessment and Control Evaluation (FACE) database. In addition to the limitations inherent to the small size and nature of the data used, their analysis stayed at a basic level. For instance, no attempt was made at detecting communities in graphs. In this study, we go a step further in the sophistication of the analyses and in the size, diversity, and relevance of the data used.

Also, some similarities are shared by our work and that of Palamara et al. [62], who used another data mining technique, self-organizing maps, to analyze a database of 1207 accident reports from the Italian wood manufacturing industry. In addition to the notable differences in the data used, scope, and methodology, the information available for each report in the national database studied by Palamara et al. [62] had been pre-filled for four categories (activity, deviation, contact and material, and mixed activity descriptors). This classification scheme fundamentally differs from the attribute-based framework we use in this study.

Finally, the entire approach of Esmaeili and Hallowell [26] is based on the assumption that only frequent associations of attributes should be considered dangerous, and Palamara et al. [62] aimed at uncovering the most frequent sequences of events leading to accidents. Our effort differs from these previous studies as we assume that valuable new safety knowledge may also be found in infrequent attribute combinations.

3. Analysis

3.1. Presentation of the data set

A data set of 5298 injury reports featuring all types of injuries was obtained from more than 470 private construction organizations involved in industrial, energy, infrastructure, and mining work. The reader is encouraged to refer to Prades Villanova [65] and Desvignes [21] for more information about these data. The unstructured, naturally occurring reports were automatically scanned for the 80 attributes shown in Table 1 by Tixier et al.'s [77] NLP system. Of the 5298 reports, 911 were not associated with any attribute and were removed, making for a final data set X of $r = 4387$ reports by $p = 80$ attributes presented in Fig. 1. The entries $X_{r,p}$ of X take on the value "1" if the p^{th} attribute has been detected in the r^{th} injury report, and "0" else. The attribute counts in this final data set are reported in Table 1.

As one can see in Table 1, attributes are classified in three categories: *upstream*, *transitional*, and *downstream*. Upstream precursors can be anticipated as soon as during the design phase, transitional precursors can be detected before construction begins based on knowledge of construction means and methods, and downstream precursors can only be observed during the construction phase. Note that this classification scheme may be changed and does not incur any loss of generality in the subsequent analyses.

3.2. Methodology

Our proposed methodology is based on two state-of-the-art, independent, complementary families of data analysis techniques, *graph mining* and *hierarchical clustering on principal components* (HCPC). To identify candidate safety clashes, we relied on community detection algorithms and edge centrality measures and on the ability of hierarchical

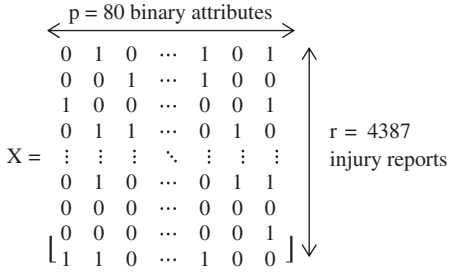


Fig. 1. Structure of the attribute data set used in this study.

clustering to isolate outliers into small clusters, respectively for the graph and HCPC part. In this paper, a cluster refers to reports that are close to each other in the high-dimensional attribute space.

In every case, it is important to understand that the role of the algorithms is to facilitate the job of the user. The goal here is to discover new safety knowledge by isolating a small amount of highly relevant atypical observations (ideally, a few dozens of reports) from the bulk of the data (tens or hundreds of thousands of reports in practice). Although the outcomes of quantitative methods can provide evidence for unusual or unexpected associations, human inspection and qualitative analysis is needed to decide which information is relevant. For example, not all atypical observations may be considered legitimate safety clashes, and not all safety clashes may be of interest. Logical tests that accompany quantitative results are essential.

In what follows, for each approach, (1) a theoretical background is provided, (2) the proposed methodology is described, and (3) some results are presented for illustration purposes.

3.3. Graph Mining

3.3.1. Overview and definition

A graph, or network, is defined as a set of vertices and a set of edges [71]. Vertices, or nodes, represent variables, and edges represent links between them. Mining graphs is a very effective way of identifying the key players and better understanding the interplay among a set of variables. Indeed, graphs can capture and represent the structure of many real and abstract complex systems [30,61,71]. For instance, graphs have been used to describe and analyze the interaction among proteins, DNA, RNA, and metabolites within cells [20], brain organization [9], power grids [81], the World Wide Web [22], or disease propagation among a population [8]. In the engineering management field, networks have been used to model risk and people interaction during projects [14, 29,82], safety communication among workers [3], and fall hazards on construction sites [26].

3.3.2. Representing attribute data sets as graphs

We create undirected graphs where each node is an attribute and there is an edge between two nodes if the two attributes they represent co-occur in at least one injury report. Furthermore, edges are weighted according to co-occurrences counts.

3.3.3. Centrality metrics

There are many ways to define the importance, or centrality, of a given vertex or edge in a network. We used three of the most standard centrality measures found in graph theory, and briefly present them in what follows. Note that while conceptually related, these metrics were shown to capture and reflect different aspects of network centrality [79]. For brevity, and because the metrics are widely used and mathematically simple, we do not provide equations and detailed interpretations. For further information on node eigenvector centrality, node closeness, and edge betweenness, we refer the reader to Borgatti [8], Freeman [31], and Girvan and Newman [84], respectively.

3.3.3.1. Node eigenvector centrality. This metric takes into account not only the number of direct connections of a given node (known as its degree [8]), but also the degrees of its connections themselves. In other words, it measures the importance of a node by considering both the quantity and the quality of its contacts. Unlike with degree centrality, a node with only a few neighbors can be seen as important in terms of eigenvector centrality if its neighbors are central [7,69].

3.3.3.2. Node closeness. A vertex is central in terms of *closeness* if it is located at short distances from all the other nodes in the graph [31]. In the social domain, an individual scoring high for closeness is one that would be able to communicate with all the other persons in the network at minimum time and cost, and by utilizing very few intermediaries. Therefore, a major difference with eigenvector centrality is that closeness is related to the notion of *independence*. While a vertex highly central in terms of eigenvector relies on its connections to spread its influence throughout the network, a node high on closeness can pervade the network by itself.

3.3.3.3. Edge betweenness. This measure is defined as the number of shortest paths that pass through a given edge [84]. Edges with high betweenness usually act as bridges between communities, connecting the members of one group to those of another. They act as network flow controllers and coordinators, as they have the power to pass on or to retain information [31].

The top nodes for degree, eigenvector centrality, and closeness, and the top edge for edge betweenness, are illustrated for a simple network in Fig. 2.

3.3.4. Community structure

While computing centrality metrics is an obvious first step in analyzing a graph, detecting and analyzing its communities is also very insightful. Even though there is no unique formal definition, communities (or clusters) of a graph are typically considered to be groups of nodes within which connections are dense, and between which they are sparse [85]. For instance, two groups clearly emerge in the simple graph shown in Fig. 2c (as denoted by the two shaded areas). Many natural and human-produced networks exhibit community structure [61]. Interestingly, nodes belonging to the same communities often share unique properties and perform specific functions [54]. For instance, proteins belonging to the same clusters within metabolic networks were found to have the same role and be involved in the same cell processes [51]. Furthermore, nodes lying at the center of their communities usually play a role of control and stability, while the ones located at the boundary often act as mediators and flow controllers [30]. For all these reasons, community detection is a task of paramount importance in graph mining [6,30,56,71]. In this study, we used several community detection algorithms (presented in what follows) to identify groups of frequently co-occurring attributes but also not already well-known associations between attributes. More precisely, we assumed that bridges between communities would make good safety clash candidates.

3.3.5. Community detection algorithms

Following recommendations from the literature [56], we used an ensemble of five state-of-the-art algorithms to increase the significance and robustness of the community detection process. We implemented these algorithms in R with the "igraph" package [19], and briefly present them in what follows. Modularity is a widely used function in graph theory that measures the quality of a given partition of a network into groups, by comparing the number of within-community edges in the clustered network to the expected such number in a null model [85].

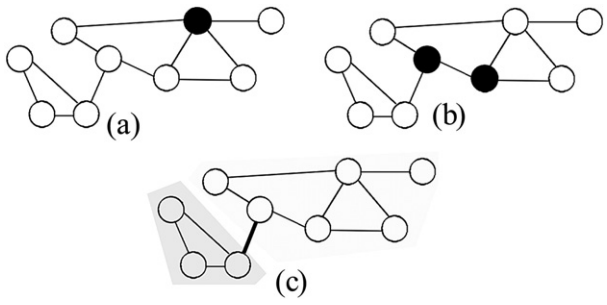


Fig. 2. Top nodes for (a) eigenvector centrality, (b) closeness, and (c) top edge for edge betweenness with two natural communities (highlighted by the shaded areas).

3.3.5.1. Fast greedy. The fast greedy algorithm [16] starts with all vertices as a cluster of their own and repeatedly merges the pair of clusters whose combination produces the largest modularity gain. This process is repeated until a single community remains. The best partition is finally selected among all possibilities as the one associated with the greatest value in modularity.

3.3.5.2. Multilevel. With the multi-level algorithm [6], all nodes start as a community of their own too. At each iteration, nodes are first moved to the community of their neighbors such that the greatest gain in modularity is achieved. Second, the communities found are turned into nodes, yielding a new graph, and the first step is repeated. This process iterates until a maximum in modularity is reached and no more change occurs.

3.3.5.3. Leading eigenvector. The leading eigenvector algorithm [61] recursively partitions communities (initially the entire graph) into two groups according to the signs of the elements of the leading eigenvector of the modularity matrix, and stops when all communities are indivisible.

3.3.5.4. Spinglass. The spinglass algorithm [67] is an approach based on statistical mechanics. It maps the graph to a Potts-like system with nearest-neighbors interaction where nodes are at first randomly assigned a spin state. It then uses a global optimizer, simulated annealing, to find the configuration of the system that minimizes the total energy. This ground state corresponds to the best partition of the graph into communities, which are defined as clusters of nodes sharing the same spin alignment.

3.3.5.5. Walktrap. Finally, the walktrap algorithm [63] uses agglomerative hierarchical clustering with a distance based on random walks. The assumption is that random walks tend to get trapped into dense portions of the graph (i.e., communities). We followed Pons and Latapy's [63] advice and used random walks of length 2, since our graphs were quite dense. For all the other algorithms, we stuck to the default parameter values.

For each graph analyzed in this study, we implemented the five community detection algorithms presented above once with the exception of the spinglass, which was implemented 100 times (with majority vote aggregation) as it is stochastic. Then, the final community structure was determined by majority vote of the five algorithms. When consensus could not be reached for a given node (i.e., two votes against two votes), the decision was left to the algorithm yielding the best partition in terms of modularity.

3.3.6. Construction accidents as attribute network perturbations

A graph perturbation is any topological modification of a network such as the deletion or addition of a node or edge. In genetics, perturbations in gene regulatory networks have been discovered to be one of the root causes of certain diseases. We posit that in the same way, perturbations in networks of fundamental construction attributes are one of the root causes of injuries. In what follows, this analogy is elaborated.

Functional states of cells correspond to stable states of an underlying gene regulatory network [46]. The robustness of such networks against perturbations allows cells to constantly adapt and continue to function normally when faced with changing conditions, such as changes in temperature and pH, or exposure to DNA-damaging agents [20]. Interestingly, it has been shown that while these regulatory networks exhibit robustness to most attacks, they are very fragile to specific perturbations, such as the mutation of one single gene or the exposure to particular toxins [74]. When faced with these perturbations, gene regulatory networks can transition into pre-existing pathological states, leading to cascading failures and to the development of diseases [46]. For instance, Taylor et al. [75] showed that topological transformations in the protein-protein interaction networks of sick patients directly impact disease outcome. Especially, specific hub proteins, critical to networks' connectivity, were frequently found to have mutated among negative-outcome patients, effectively altering organization and flow of the protein network. Taylor et al. [75] concluded that these hubs should become therapy targets.

Similarly, jobsite conditions at any particular location and at any given point in time can be represented by a combination of fundamental construction attributes, that is, by a given attribute network. Thanks to injury prevention techniques such as safety rules and guidelines, preventive and corrective measures, site supervision, and worker vigilance, networks of attributes tend to stay in stable states most of the time, despite the numerous perturbations caused by the dynamic and ever-evolving nature of construction environments. Because of this inherent resilience, a dormant hazardous situation may go unnoticed or ignored for a long time. It is indeed well known that construction workers can expose themselves to unsuspected risk because they fail to recognize latent hazards in their environment [1,12]. We postulate that perturbations in attribute networks can trigger the transition from inactive hazardous states to active accident-prone states. Under these conditions, if perturbations go unnoticed, and if no corrective action is taken, the chances of observing injuries are greatly increased. To validate this theory, however, comparing graphs of injury reports to graphs of "non-injury" cases would be necessary.

According to the just introduced *construction accident as attribute network perturbations* theory, the goal of safety management would consist in ensuring that attribute networks always stay in stable states. Therefore, in the exact same way that specific DNA-binding proteins [20] or hub proteins [75] have become drug targets, safety-critical topological features of graphs of attributes (e.g., "safety clashes") should become the targets of safety intervention programs and the center of attention when developing preventive strategies.

3.4. Graph Mining: Protocol and Results

As shown in Fig. 3, the initial data set was split based on the safety outcome *injury type* in order to ease interpretation of the results. This gave five smaller data sets: struck-by or against (2389 reports), caught in or compressed (350), fall on same or to lower level (570), overexertion (567), and exposure to harmful substance (525). Furthermore, for each subset, the attributes that appeared in less than 1% of the reports were removed as a cleanup pre-processing step. The graph mining steps previously described were then applied. The results are shown in Table 2, and Figs. 4 to 8.

We used the Fruchterman-Reingold [32] force-directed layout algorithm available in the "igraph" R package [19] to plot the graphs

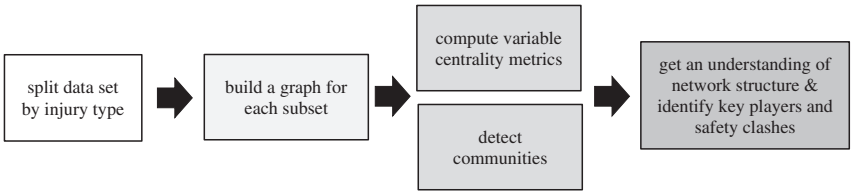
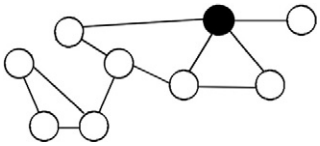
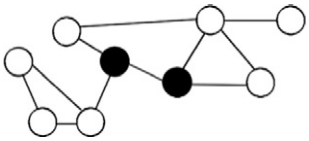
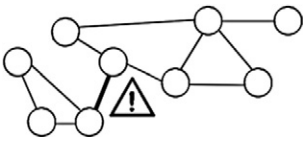


Fig. 3. Overall graph mining procedure.

Table 2
Top elements for eigenvector centrality, closeness, and edge betweenness.

Injury type			
	Eigenvector centrality	*Closeness*	*Edge betweenness*
Struck-by or against	Unpowered tool, small particle, piping, manual handling, steel sections	Heavy material/tool, sharp edge, improper security of materials, manlift, improper procedure/inattention	Working below elevated wksp and slag; cable tray and unpowered tool, imp. procedure/inattention and steel sections
Caught in or compressed	Bolt, steel sections, unpowered tool, piping	Formwork, lumber, concrete, working at height	Improper security of materials and rebar; valve and unpowered tool; ladder and exiting/transitioning.
Fall on same or to lower level	Object on the floor, working at height, slippery walking surface, scaffold, steel sections	Hand size pieces, formwork, machinery, manual handling, cleaning	Machinery and exiting/transitioning; scaffold and object on the floor.
Overexertion	Lifting/pulling/manual handling, unpowered tool, steel sections.	Formwork, bolt, lumber, scaffold, working at height	Heavy material/tool and improper body positioning, spool and light vehicle, working below elevated workspace/material and unpowered tool.
Exposure to harmful substance	Hazardous substance, piping, welding, heat source, steel sections	Concrete, unpowered tool, cleaning, hammer	Wind and hazardous substance; piping and confined workspace; concrete liquid and wind; welding and working overhead.

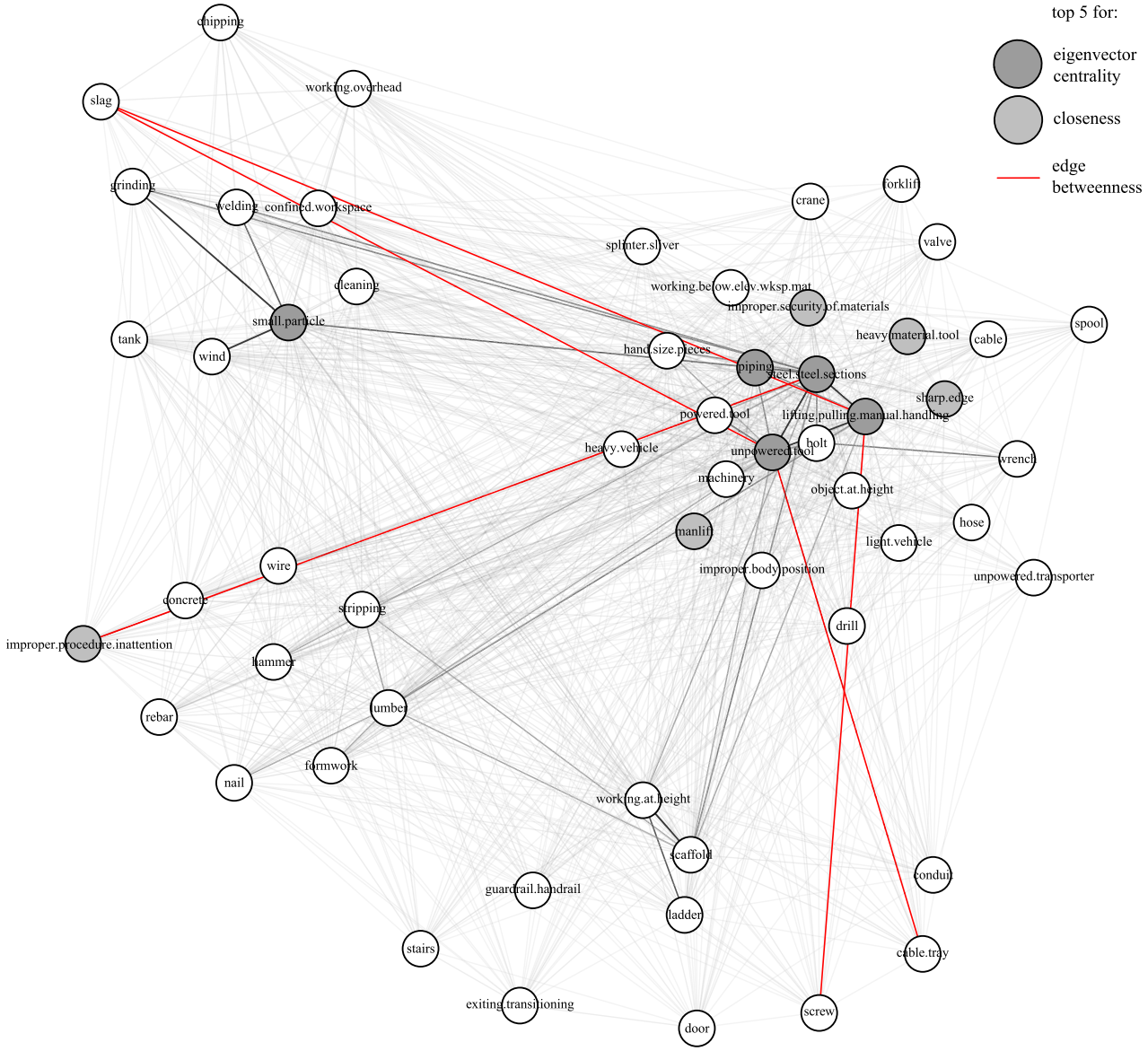


Fig. 4. Attribute co-occurrence graph for *struck-by or against* (2389 reports).

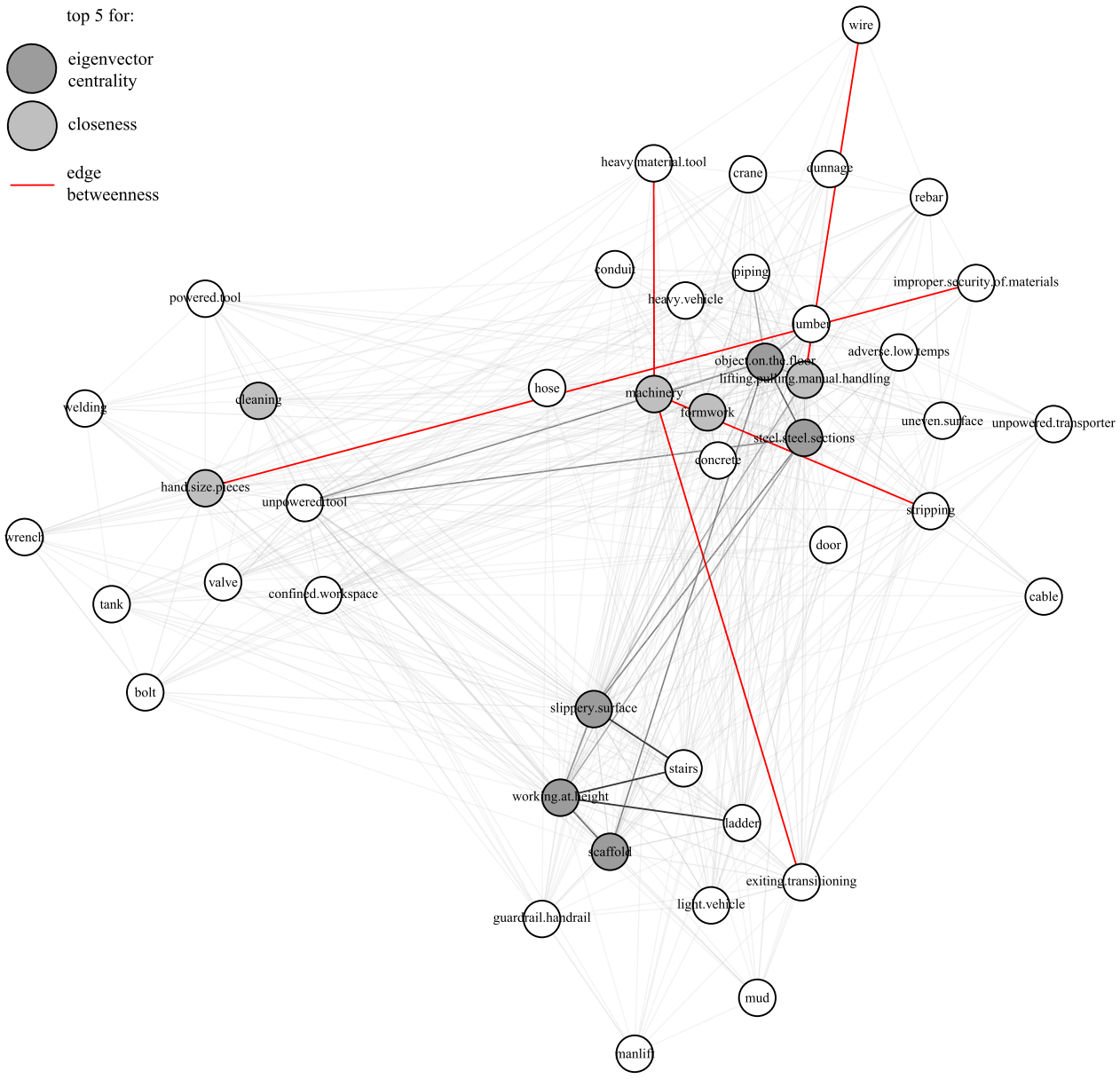


Fig. 5. Attribute co-occurrence graph for *fall on same or to lower level* (350 reports).

presented in Figs. 4 to 8. Note that on these plots, the nodes belonging to the same communities are grouped together. Furthermore, the top five attributes for eigenvector centrality are colored in dark grey, the top five attributes in terms of closeness are filled in light grey, and the top five edges for edge betweenness are shown in red. For a given graph, less than ten nodes may be colored in grey when there is some overlap between the top attributes for closeness and the top ones for betweenness. Finally, the transparency and width of the edges is proportional to the strength of the co-occurrence between the vertices they link. In other words, attributes frequently found together in injury reports are linked by dark, thick edges, while attributes that only seldom jointly appear are connected by light, thin edges. Note that the top five attributes for eigenvector centrality and closeness are also reported in Table 2 for each graph, with only the most relevant of the top five edges for betweenness (due to space constraints).

3.4.1. Interpretation of the results

In this section, we interpret the graphs shown in Figs. 4 to 8. For each graph, we highlight relevant candidate safety clashes and provide corresponding anonymized report excerpts for illustration purposes. A selection of clashes are summarized in Table 2 for each graph, along with the top nodes and edges for each centrality metric.

We tried to identify safety clashes by searching notable structural elements of graphs, such as edges scoring high for betweenness (shown in red in the graphs), interesting links between two or more attributes (safety-critical "chains"), or bridges between communities. As previously explained, our assumption was that interesting, not already well-known safety clashes would most likely be found among less frequent attribute combinations. This is why we did not limit our search to only the hubs or the thicker edges. Note that most of the time, the top edges for betweenness were found among bridges between communities, which is in accordance with Girvan and Newman [84].

For the "struck-by or against" graph (see Fig. 4), the attributes *welding*, *grinding*, *chipping*, *slag*, *tank*, *confined workspace*, *wind*, and *working overhead*, are all grouped around *small particle* in the same community (in the upper left corner). The implication is that workers are frequently subjected to small-particle-related injuries

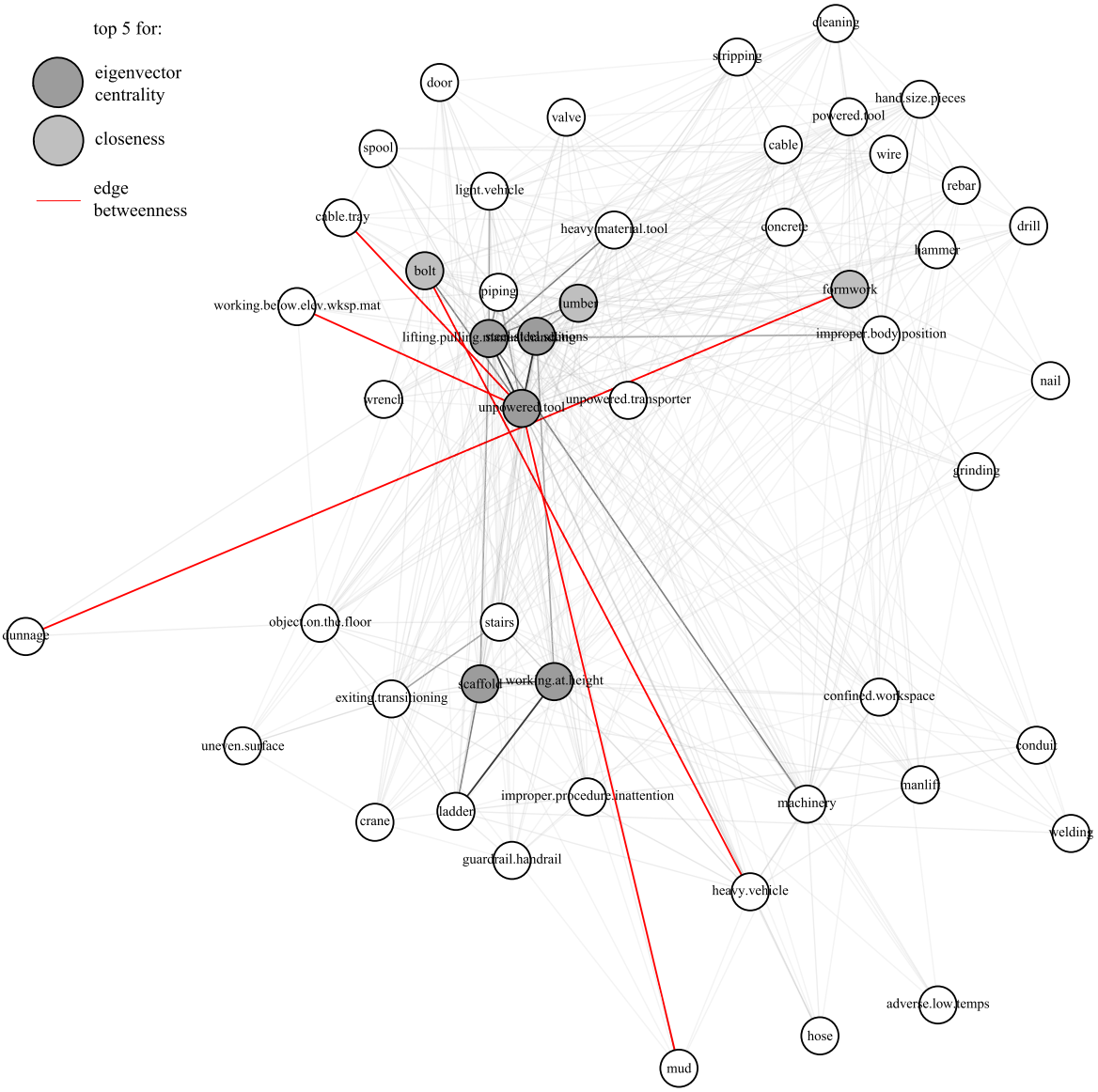


Fig. 6. Attribute co-occurrence graph for *overexertion* (570 reports).

when grinding, chipping, welding, or cleaning, especially in enclosed spaces, when working overhead, or when working outside in windy conditions.

Employee received a foreign body in the eye while cutting overhead.

The fact that *small particle* scores high for eigenvector centrality (quantity but also importance of the connections) confirms its central position in the community and indicates that it contributed to large amounts of struck-by injuries in the data set we analyzed.

Similarly, *hammer*, *nail*, *lumber*, *formwork*, *concrete*, *stripping*, *rebar*, and *wire* are clustered into the same community (in the bottom left corner of Fig. 4). Note that *hammer* holds a central place within this community, making this attribute very specific and representative of its group, while some other attributes such as *concrete* or *lumber* lie on the periphery of the community, highlighting their intermediary positions with other groups (respectively, the *small particle* and the *scaffold* group).

Interestingly, the attribute *improper procedure/inattention* is found in the same group, which tends to indicate that many hammer-related "struck-by" injuries are caused by misses:

While using a small crow bar and hammer to remove rust, the carpenter missed the crow bar striking his left index finger with the hammer.

What also makes sense is that *improper procedure/inattention* acts as a bridge with other groups (especially, notice the strong connection with the large community in the upper right corner). Indeed, human error is not specific to a particular suite of actions or work situations and can be found everywhere. It is thus understandable that attributes purely and simply related to human error are shared across all communities.

Still for the "struck-by or against" graph, one should note that two out of the five top attributes for closeness are related to human behavior (*improper security of tools* and *improper procedure/inattention*). Even more interesting is that this phenomenon is not observed for the other graphs. Recall that nodes of a graph scoring high on closeness are located at short distances from all the nodes in the graph. In other words, the attributes they represent are pervasive in the underlying data set. Therefore, one interpretation is that human error is prevalent, omnipresent in the "struck-by or against" injury cases.

Surprisingly, *manlift* is one of the top-scoring vertices on closeness, meaning that it is located at very short distance from all the other

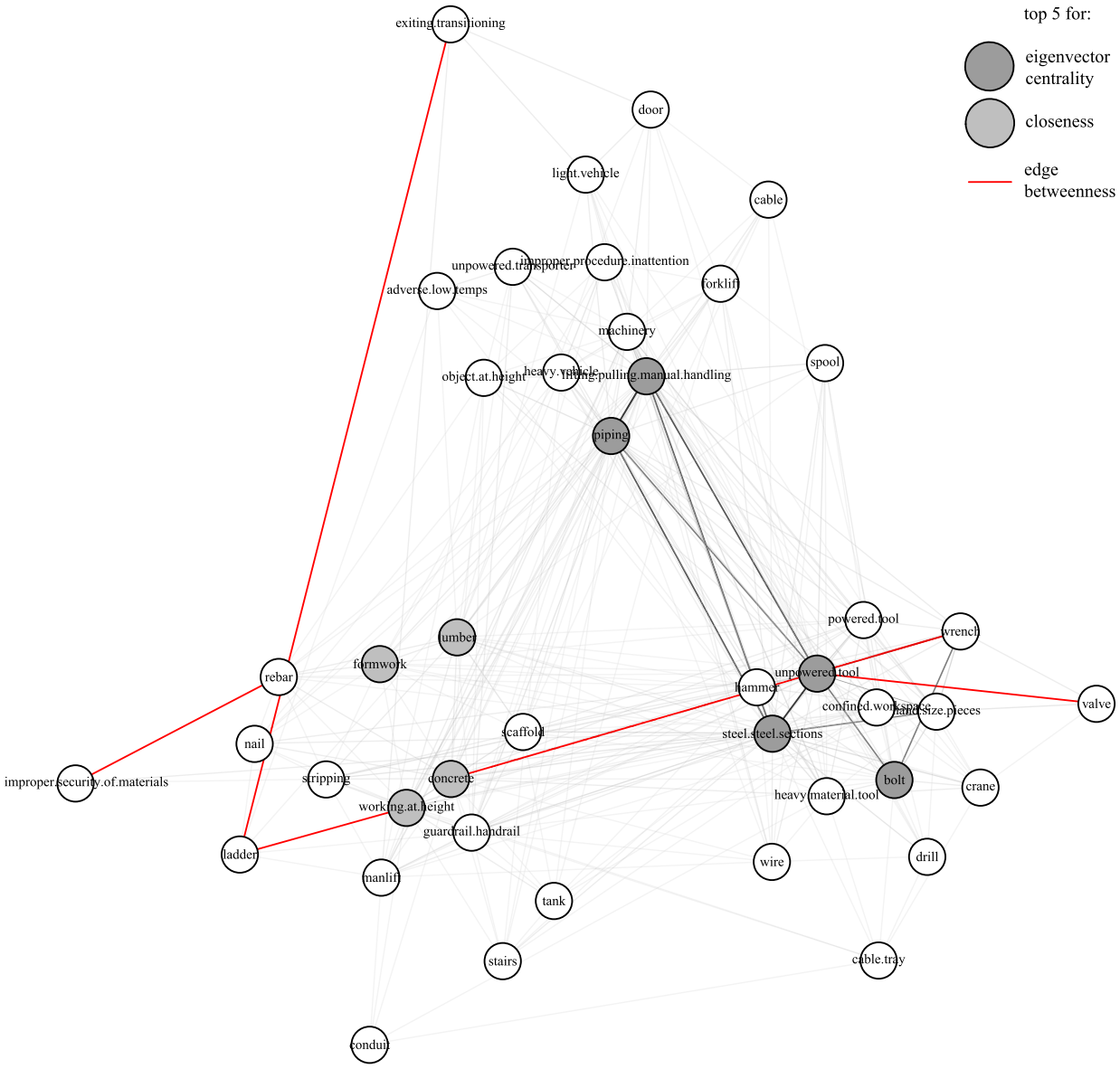


Fig. 7. Attribute co-occurrence graph for *caught-in or compressed* (567 reports).

attributes in the network. Put differently, it caused injuries in association with a large variety of attributes. This can be explained by the fact that manlifts are found in many different construction situations involving their direct use or not.

Regarding the "fall on same or to lower level" graph shown in Fig. 5, the connection between *machinery* and *exiting/transitioning* is worth noticing and is a good illustration of a non-trivial safety clash. Excerpts of some corresponding injury reports are provided below:

> *Employee was climbing out of the excavator and rolled his ankle in the process.*

> *Worker was walking out of a trailer. When her foot touched the ground, she rolled her left ankle.*

Note that *exiting/transitioning* is also found in the same community as *manlift*, *light vehicle*, and *ladder*, which indicates that descending or ascending is problematic not only for *machinery* but more generally for any kind of equipment.

Furthermore, the strong link between *object on the floor* and *unpowered tool* reveals that numerous falls on same level (tripping, stumbling) are due to the presence of loose tools on the floor, and the close proximity of these two attributes respectively with *confined workspace*, *scaffold*, and *working at height* suggests that the risk of falling due to the presence of objects on the floor is compounded in constrained spaces. Finally, from the thick edge between *object on the floor* and *piping*, it is possible to infer that many of the problematic objects let on the floor are pieces of pipes (at least in our data set):

> *Employee tripped over a pipe support that was lying on the scaffold causing him to strain his knee. A contractor employee stepped on a pipe that was installed just above the floor of a scaffold, causing him to roll his ankle.*

> *Employee was working on a scaffold installing pipe supports, as he was working he tripped over a pipe support that was lying on the scaffold causing him to strain his knee.*

Still with respect to the "fall on same or to lower level" graph, an interesting clash is *steel/steel sections* and *slippery surface*. This clash may not appear evident or identifiable from common sense only since steel sections are not supposed to be used as walking surfaces:

Employee was unhooking slings from metal beams inside dumpster when he slipped on slick metal from weather and twisted his right knee.

While traveling on foot through the waste caustics area, a subcontractor employee slipped and twisted his right knee on a steel plate that covered a U-drain in the unit.

Note that both *steel/steel sections* and *slippery surface* are among the top five vertices for eigenvector centrality (many connections that are themselves central), meaning that they are major injury contributors when associated not only together but also with other attributes.

Quite logically in the "overexertion" graph shown in Fig. 6, a major clash is the strong association between *lifting/pulling/manual handling* (a highly central node in terms of eigenvector centrality), *heavy material/tool*, and *improper body position*:

Employee complained of soreness in his elbow after handling heavy air impact guns and crusher teeth (each approx. 50 lbs). The employee was positioned on the fixed end of the crusher, which has less space to work and involves awkward positioning. Two weeks later the employee felt numbness while manipulating a 4 lbs hammer, causing them to drop the hammer.

Very related to the safety clash above is the one involving *lifting/pulling/manual handling*, *unpowered tool* and *working below elevated workspace/material*. The attribute *working below elevated workspace/material* produces effects that include that of *improper body positioning*:

This employee crawled under the air channel pipe to grab a chain fall and felt a discomfort in his abdomen area when he started to get up.

But the potential adverse effects of *working below elevated workspace/material* are not limited to that of *improper body positioning*. For instance, in the cased depicted by the excerpt below, body positioning seems right. What is problematic is the position of the employee relative to the tool they are manipulating:

An iron worker had to go under some pipe to connect a come-a-long to a beam. When he was under the pipe, he pulled the chain towards himself. When he did, the end of the come-a-long struck the employee in the mouth.

One should note too that the *unpowered tool-working below elevated workspace/material* edge stands in the top five for betweenness, meaning that it is in a position of controlling the flow in the network. In other words, in the "overexertion" graph, one of the fastest ways to link two non-neighbor attributes is by passing through the *unpowered tool-working below elevated workspace/material* edge.

Finally, the close proximity of *lifting/pulling/manual handling* to *bolt*, *wrench*, and *unpowered tool* implies that in our data set, many overexertion injuries occur when tightening bolts. This observation, which makes sense, is strengthened by the fact that *unpowered tool* stands among the top five nodes for eigenvector centrality (reflecting both the quantity and the quality of connections) and that *bolt* is very high for closeness (a measure of pervasiveness throughout the network).

EE was tightening bolts on non-segmented bus duct to the specified torque value when he felt discomfort in his back.

A subcontractor worker was adjusting bolts on a joint with a ratchet tool. He felt a strain in his neck.

A major clash that can be identified from the "caught in or compressed" graph shown in Fig. 7 is *improper security of materials* and *rebar*. It seems that many pinches and crushing injuries indeed involve improperly secured or unprotected rebar (either loose rebar or protruding rebar in place).

An ironworker was placing rebar when the bar dropped, pinching his finger.

Moreover, the close proximity of the aforementioned pair of attributes to *formwork* and *lumber*, two attributes in the top five for closeness indicates that improperly secured *rebar* is often found within the context of handling lumber:

As he was removing a 16" long 2 × 4, the board became too heavy and pinched his finger between the board and the horizontal rebar protruding from the construction joint.

Employee pinched his finger onto rebar, while positioning formwork.

The very strong bond between *piping* and *lifting/pulling/manual handling* (two attributes in the top five for eigenvector centrality) reveals that this association is responsible for many "caught in or compressed" injuries. This can be readily understood. Actually, pipes are heavy and unstable by nature. Positioning or manipulating them is therefore prone to creating crush or pinch injuries. This is compounded by the fact that pipes are often to be installed in confined spaces (notice the strong links with the *confined workspace* community on the bottom left) where the proximity with other hard surfaces is high, and at height (suspended).

The pipe was suspended about 2"–3" off the ground. While installing the clamp, the pipe moved and slipped from his hands causing it to fall to the ground where the worker crushed his left hand index finger.

Piping is also strongly connected with the community of *unpowered tool*, *bolt*, and *steel sections* (among others). These attributes are highly central in terms of eigenvector centrality, denoting that they are major "caught in or compressed" injury contributors. The strong link with *piping* suggests that all these attributes play as a team:

Worker pinched his finger between a bolt flange and a pipe.

Worker pinched his hand between a pipe and a piece of steel.

Also, the link between this community (which also includes, among others, *hammer*) and *concrete* is interesting:

Employee was stripping floor beam recessed block out, set cat's paw with hammer, missed cat's paw and pinched left hand between tool and concrete.

Finally, many "caught in or compressed" injuries involve ladder, as can be concluded from the observation that *ladder* is the endpoint of two edges in the top five for edge betweenness.

Ladder slipped out of employees hand and pinched their right middle finger

Expectedly, *hazardous substance* holds a very central place in the graph shown in Fig. 8 ("exposure to harmful substance"). Its direct connections with *concrete*, *chipping*, *grinding*, *lumber*, and *small particle* shows that exposure to fine dust was a frequent issue in the data set we analyzed, especially under windy conditions (notice the very strong link between *hazardous substance* and *wind*):

While burning the bolt heads, the dust from the demolition of the concrete deck that had previously been demolished in the same area

was blown by the wind and got in the eye of the employee.

As employee was working adjacent to a wood cutting operation, wind blew saw dust into his eye.

It was a windy day and an insulator employee was cutting a piece of piping insulation. The employee felt a discomfort in his right eye.

The grouping of *concrete liquid*, *grout*, *machinery*, *hose*, *valve*, and others into the same community (in the upper left corner) is not surprising either.

Worker noticed that the adjacent concrete was drying out and needed to be wet to maintain good cure. When worker picked up the water hose, the valve opened and a blast of hot water came in contact with his abdomen causing the burn.

Maybe more interesting in the "exposure to harmful substance" graph is the close proximity of *welding* to *improper body positioning*, *working overhead*, and *scaffold*. It is well known that *welding* (directly or through *slag* and *spark*) is a *major* light and heat sources and is thus central in creating "exposure to harmful substance" injuries. However, it appears that the risk of this attribute is compounded when workers adopt non-natural body positioning:

Employee was welding overhead and felt slag fall on right side of neck resulting in a burn.

Notice that logically, there is a direct link between *improper body positioning* and *confined workspace*, suggesting that the former could be consequence of the latter:

Employee had a few inches between him and the weld he was making; it was a very tight and awkward position. He was exposed to arc flash that came in from under his hood

Non-natural body positioning is also an issue with other hazardous substances such as *concrete liquid* or insulation:

While reaching overhead, some of the patch mix that the carpenter was using got between his shirt and glove, causing minor concrete burn.

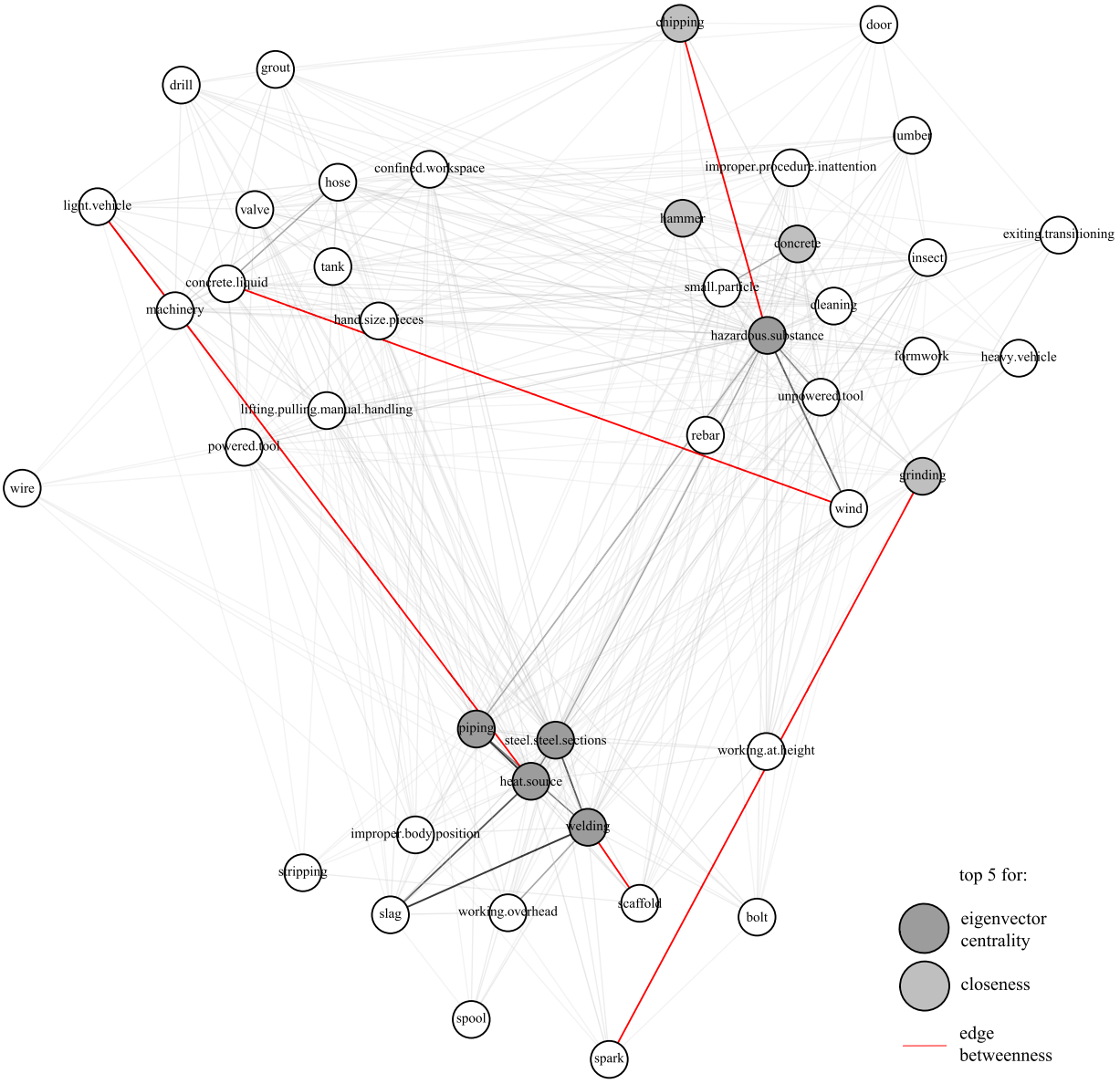


Fig. 8. Attribute co-occurrence graph for *exposure to harmful substance* (525 reports).

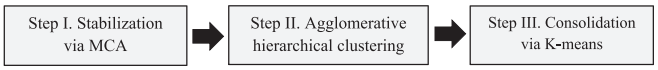


Fig. 9. Hierarchical Clustering on Principal Components (HCPC) steps.

A carpenter foreman got some insulation in his eye while stuffing insulation overhead between the hard lid trusses.

Finally, the connection between *piping* and *confined workspace* revealed an interesting, non-trivial clash:

Employee slipped into a small excavation containing very hot water (app. 158 degrees Fahrenheit). He sustained severe burns to both legs. Water coming from the melting snow was heated by a steam line installed the previous day.

Overall, we showed that the community structure exhibited by the attribute data makes physical sense and that graphical features can be used to identify interesting combinations of attributes. This shows the promising potential of our methodology and tends to validate Tixier et al.'s [77] NLP tool with which the attribute data set was extracted from injury reports in the first place.

To mine the attribute data set from a different perspective and gather complementary safety knowledge, we used hierarchical clustering, as shown next.

3.5. Hierarchical Clustering on Principal Components (HCPC)

3.5.1. Overview and definition

To identify atypical but valid combinations of attributes from another perspective than a strictly "social" one, we used an unsupervised data mining technique complementary to network analysis, hierarchical clustering. Hierarchical clustering is known for its ability to isolate outliers into small clusters. By manually inspecting these small clusters automatically constructed, we were able to easily identify valid cases that "stood out from the crowd", that is, potential safety clashes.

Note that to find more stable and definite clusters and therefore enhance the robustness of our results, we used Hierarchical Clustering on Principal Components (HCPC, [47]), an improvement over traditional hierarchical clustering. HCPC consists of three complementary steps as shown in Fig. 8: observations are (1) projected onto the principal component basis, (2) partitioned into groups via agglomerative hierarchical clustering, and finally, (3) the groupings are consolidated by using a K-means algorithm. In what follows, these three steps are detailed.

Step 1: *Principal Component Analysis*

First, it is necessary to recall that in this study, each observation (i.e., each injury report) lives in an 80-dimensional space (the feature space shown in Fig. 1). Each fundamental construction attribute represents a dimension of this space, and each injury report is defined according to how it loads on these dimensions, that is, by its coordinates (zeroes or ones) in the feature space.

Principal Component Analysis (PCA) is a widely used algebraic procedure that reduces the dimensionality of a data set while preserving most of the information that it originally contains. More precisely, PCA first constructs an orthonormal basis linearly derived from the original feature space, such that the variance of the observations projected in this new basis is maximized [50,72]. Each axis (or principal direction) of the new basis matches the direction of maximum variability of the cloud of data points, with the constraint that each successive axis is orthogonal to all the preceding ones. The data reduction takes place in the second phase of PCA, sometimes called the compression phase [73]. This second step simply consists in ordering the principal components by decreasing eigenvalues, and in selecting the first k ones (where $k<p$). The dimensionality of the feature space is thus reduced from p to k while conserving as much original variance as possible. A desirable side effect of maximizing the information captured is that most of the noise is discarded. HCPC capitalizes on this denoising capability of PCA. Indeed, by clustering observations in the space made of the first k principal directions rather than in the original feature space, denser and more discriminative groups can be found [47].

Note that in the same way that Kauffman [55] sees genes as either "on" or "off", each fundamental construction attribute is either present or absent from a given injury report (i.e., binary or Boolean variables). Therefore, we used Multiple Correspondence Analysis (MCA, [5,36]), the extension of PCA to categorical variables.

Step 2: *Agglomerative hierarchical clustering*

Agglomerative hierarchical clustering refers to a class of unsupervised learning algorithms that classify r observations into a hierarchy of disjoint groups, following a recursive, bottom-up approach ([43], p. 523). As shown in Fig. 10, the two clusters optimizing an objective function are combined at each step s, resulting in a grouping at level $s+1$ with one less cluster. Initially at level 0, each data point belongs to its own cluster. The algorithm stops after $r-1$ steps, when all observations belong to the same group. This approach is known as Ward's method [80]. Unlike other clustering techniques like K-means or K-medoids, hierarchical clustering offers the advantage of not requiring prior knowledge about the number of clusters.

We capitalized of the robustness of hierarchical clustering to outliers [59], in order to make atypical yet valid injury cases "stand out from the crowd". The assumption is that because outliers are associated with unusual combinations of attributes, they will be isolated in low-density areas distant from regular observations and therefore will tend to end up grouped into small clusters [2]. These outliers may correspond to errors made the NLP tool when scanning the database of injury reports, or reveal rare but valid associations between attributes. The latter were of great interest to us as they were candidate safety clashes.

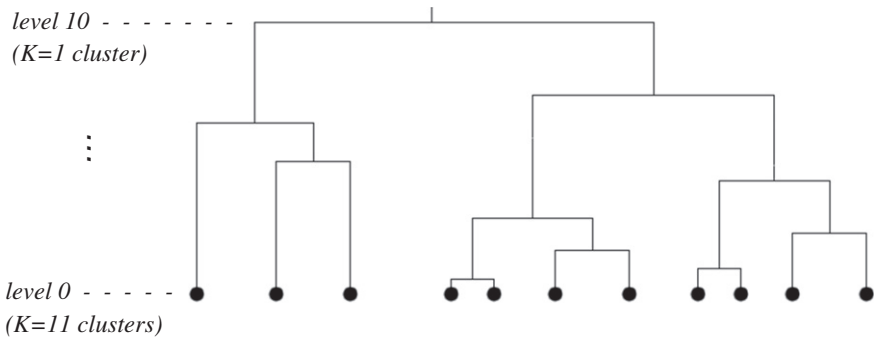


Fig. 10. Illustration of agglomerative hierarchical clustering for $r=11$ observations.

The choice of the objective function is task-dependent. Because MCA (the first step of HCPC) is variance-based, and the K-means algorithm (HCPC's last step) is based on the squared Euclidean distance, we used the within-cluster variance [47]. This function, defined in Eq. (1), measures the extent to which members of a cluster are close to the center of this cluster ([48], p. 387). Even though the injury reports have binary coordinates in the original space (the attribute space), their projections onto the principal direction basis are numerical, as was explained in the previous section. Therefore, using the Euclidean distance was not a problem. To summarize, the two clusters minimizing the increase in within-cluster variance when merged were grouped at each step.

$$W(C_k) = \sum_{x_{i \in Ck}} \left\| X_i - \overline{X}_k \right\|^2 \quad (1)$$

Within-cluster variance of cluster C_k.

Where $\overline{x}_k$ is the mean of C_k defined as the coordinate-wise averages of the observations in C_k, and $\{x_1, \ldots, x_r\}$ is the set of r observations. Each observation is a vector (one coordinate per dimension).

Each level of the hierarchy corresponded to a theoretically valid partition of the observations into K clusters. When hierarchical clustering is used for its primary purpose (i.e., grouping observations into a few compact and well separated clusters), the selection of the optimal level $s_{optimal}$ of the hierarchy can be achieved quite easily based on quantitative criteria. For instance, an approach consists in selecting the level associated with the number K of clusters such that the total within-cluster variance $W = \sum_{k=1}^{K} W(C_k)$ is minimal. However, when the task of interest is that of outlier detection like in this study, this criterion loses relevance, and finding the optimal level becomes more problematic. Indeed, selecting a level that is too high in the hierarchy returns only a few very big clusters, where outliers are mixed with regular observations, while picking a level that is too low yields a vast amount of very small clusters, which is not a significant improvement over going through the reports manually. We followed Loureiro et al.'s [59] rule of thumb and selected the level of the hierarchy that returned $\max(2, r/10)$ clusters, where r is the total number of observations (reports in our case). This heuristic still gave too many clusters (e.g., 230 clusters with the *struck-by or against* data set), and therefore, we used $\max(2, r/50)$ instead. Furthermore, of the $\max(2, r/50)$ clusters returned, only the ones containing less than 10 observations were examined.

Step 3: *Consolidation of the clustering* via *K-means*

The partition obtained in step 2 was refined by applying a K-means algorithm in order to increase the consistency and separation of the clusters [47]. The K-means algorithm consists of the following three steps ([48], p. 388; [38]):

i. select initial cluster centers;
ii. for each cluster center, create a cluster by selecting the observations that are closer to this center than to any other center (in terms of Euclidean distance);
iii. update the cluster centers (computed as the coordinate-wise means of the observations in each cluster).

Steps ii and iii are repeated until convergence is attained, that is, until the assignments do not change. The initial cluster centers were given by the coordinate-wise averages of the clusters found by hierarchical clustering at the previous step.

3.6. Hierarchical Clustering on Principal Components: Results

Using the rule of thumb from Loureiro et al. [59], 90 clusters were requested. 40 of them contained less than 10 elements and were manually inspected for safety clashes. Relevant findings are organized by main themes in what follows. Again, for brevity, only a few anonymized, representative reports are shown for each clash.

3.6.1. Congested and confined workspaces compound the risk of other attributes

The attributes *congested workspace* and *confined workspace* act as catalysts for accidents. They increase the risk of many different attributes, and therefore have the power to turn a great variety of work situations into hazardous ones. As a result, they should always be considered main safety targets, even when other seemingly more impactful attributes are present. For instance, the risk of tripping on an object on the floor, or of being struck-by or against a tool or some material is greatly increased in congested and confined spaces, as illustrated by the following examples:

> *Workers were moving a piece of formwork from a congested space. While doing so, one worker tripped on rebar adjacent to the walk path, causing him to fall and twist his knee.*

> *Employee performing demolition operations in close proximity to another employee was struck-by that employee's hammer.*

> *An electrician helper was installing conduit for lighting in close proximity to a section of pipe. His left forearm came in contact with the pipe, which had a piece of metal protruding resulting in an abrasion.*

Also, in order to adapt to congested or confined spaces, workers often have to adopt improper body positioning, or to follow improper procedures:

> *Employee was tightening flange by flange connections with opposing forces of two combination wrenches. He was working in a tight area, which created poor body positioning to effectively tighten the bolts. Great amount of pain in neck and back.*

> *The ladder was set in front of the last set of panels with very little room around the landing on the deck of the trailer. When the laborer went to descend the ladder he went around the grab rail due to limited access in front of the ladder. One of the hooks popped out of the rub rail and caused worker to fall. He attempted to cushion his fall with his hands and was diagnosed with a fractured wrist.*

> *Two employees where trying to move a pallet with a 4000 lbs valve on it into position so they could get the pallet jack under it. The area was congested and there was nowhere to rig from. They decided to use a 4×4 post to pry the broken pallet into a more suitable position. As the employee pried the pallet he felt some pain in his lower back.*

It is the work environment that should adapt to workers, not the opposite. Just because ergonomics and safety concerns may be more difficult to address in congested and confined workspaces does not mean that they should be ignored.

3.6.2. Flaggers are at greater risk for slips, trips, and falls due to their attention being caught by other stimuli

This is a good example of a previously undocumented, rather counterintuitive phenomenon that seems benign but may be responsible for many lost work time injuries every year:

> *While pouring concrete for sidewalk, foreman was trying to flag down a concrete truck to the pour. With his sight directed to the truck, he did not see the curb, tripped, and fell.*

While spotting crane, employee tripped on dunnage and fell into a rebar mat striking his hand on the rebar.

Considered separately, each of these incidents could be deemed unlikely and due to bad luck. But when analyzing large numbers of injury reports collected from hundreds of construction sites throughout the world and representing hundreds of thousands of worker hours, trends begin to emerge. One can then realize that these injuries may happen more frequently than initially thought and that they may not be random but rather be caused by the same underlying mechanism. This insight extraction process is a necessary step towards taking corrective actions and improving safety performance.

3.6.3. Workers are unable to recognize immediate hazards due to poor visibility

It has been shown that many workers involuntarily put themselves at risk not necessarily because hazards are not visible, but because workers fail to recognize their presence [12]. The remediation strategies that have been proposed in the literature involve hazard recognition training primarily based on visual information [1]. This assumes that all hazards are visually identifiable. Although it may be true, it should be stressed that not being able to visually detect the presence of a hazard does not necessarily mean that this hazard is absent from the work environment. Put differently, not being able to assess the full risk profile of the environment is problematic in itself and should be considered hazardous in its own right.

Employee was walking to job trailer, due to recent rainfall stepped into unseen low spot that was covered with water and fell striking his thigh.

Worker was walking under the module when they came to a beam they had to walk under. On the other side of the beam there was a 1″ hydro vent installed that was not visible to the worker. When the worker walked under the beam, he stood up and his face contacted the drain. This caused a laceration above the eye and the abrasion was closed by 6 stitches.

While employee was walking on the snow-covered ground, he slipped on a hidden patch of ice, fell back, and hit his head on the ground causing a contusion and concussion.

Employee stepped on the bottom form plate onto an exposed nail. Exposed nail was not visible due to murky water.

Hinze and Teizer [44] showed that a majority of vision-related fatalities in construction involve workers being struck-by moving equipment or vehicles, and are mainly due to blind spots, obstructions, and extreme lighting conditions. While valuable, these findings are more related to equipment and vehicle drivers not noticing the presence of workers on the ground rather than purely to the inability of workers to recognize latent hazards due to poor visibility. On the other hand, our findings suggest that there is another class of vision-related injuries that may be for the greater part of lower severity, and do not involve equipment. These injuries are due to workers not being able to identify the presence of immediate hazards due to lack of visibility. In these cases, additional precautions should be taken, such as safety warnings highlighting the presence of the hidden hazards, or when unfeasible, delivery of general recommendations to use extra caution in specific settings.

3.6.4. Exiting equipment, vehicles, or work stations is safety critical

These very anodyne actions performed very frequently every day may seem completely harmless compared to more serious, high energy hazards, such as suspended loads or moving heavy equipment, but empirical evidence suggests that they may make important injury contributors:

Employee reported that he twisted his knee while stepping off of the last step of the crane access ladder.

While exiting a van, an employee had their finger pinched by the door.

Employee exited office trailer, rolled ankle on stair platform.

This finding is consistent with recent psychological research that showed location shifts to disrupt visual and spatial cognitive processing and to cause forgetting [66]. In other words, transitioning from one work area to another (which includes exiting equipment and vehicles) may decrease situational awareness, alter risk perception, and therefore increase the potential for injury.

3.6.5. Working with hazardous substance requires proper preparation and PPE, and following procedures

Only workers who are fully aware of their environment and are in a mental and physical state of mind conducive to full concentration should be allowed to work with or in the vicinity of hazardous substances. Also, it should be emphasized that procedures should be followed exactly in case of incident, at the risk of observing very serious consequences:

A liquid loading foreman had just completed loading an ammonia rail car. He inadvertently struck a load hose bleed valve with his foot and ammonia sprayed onto his left knee, causing chemical burn.

Insulator was installing fireproofing on structural steel. As he was moving around, his back area came in contact with a valve that developed a small leak of 40 percent acrylic acid. The employee bypassed and failed to use the nearest safety shower instead he went looking for his supervisor.

Similarly, adequate PPE should be worn on tasks dealing with hazardous substances, and PPE should be carefully inspected prior to initiating work:

A laborer was helping grout door frames when she got a small amount of dried grout into her right eye. Upon investigation, it was discovered that the laborer's safety glasses did not provide a snug fit nor was the potential of falling debris identified in the JHA.

Worker disconnected the grout line which was still pressurized at approximately 70–80 psi. The back pressure relief caused the cement grout to go in an upward motion, covering the workers face, hard hat, and safety glasses. Some of the grout went behind his glasses and into his eyes.

The combination of the attributes *hazardous substance* and *powered tools* (illustrated in the following report examples) is an elegant example of the theory of *injuries as perturbations in networks of attributes* posited earlier. Indeed, some inherently hazardous substances are completely harmless in a static, stable state. However, the energy transferred to these substances by a powered tool such as a for instance a grinder can trigger a transition into an unstable, hazardous state. A fiberglass pipe, for instance, is inoffensive; however, when being cut, volatile and inhalable fiberglass dust is produced, posing immediate and long-term safety concerns (skin, eye, lung, and stomach punctual and chronic irritation).

An employee felt discomfort in right eye when grinding paint off a structural beam. After flushing eye on-site, employee was taken off-site for medical assessment and returned to work without restriction.

Worker was cutting fiberglass pipe with a grinder, and when he removed his full face respirator mask, he felt a foreign body sensation in eye.

Employee was working in the 521 area, approximately 20 ft from an area where XXXX Manufacturing was performing some grinding activities on fiberglass pipe when he felt the irritation in his left eye.

To make sure that attribute networks stay in stable states, safety management should take corrective actions. For instance, it is now common practice to equip demolition equipment with water sprinklers to suppress concrete dust. Similar proactive strategies (e.g., ventilation, vacuuming) should be adopted at the worker level, whenever powered tools are used on hazardous substances.

4. Conclusion

The data mining procedures introduced in this study, used within the attribute-based framework, allowed the automated identification of good candidate safety clashes from a large data set of attributes extracted from raw injury reports. Even though we focused here on safety clashes, our methods can naturally be used to visualize and better understand the interplay among fundamental attributes in general (i.e., the bulk of the data). Besides the findings themselves which are limited to our data set and mainly serve as a proof of concept, we believe that our methodology in itself is the major contribution of this paper. It shows great promise to become a standard way of extracting safety knowledge from raw textual injury reports and will help to replace the long-standing limitations associated with opinion-based safety analyses.

Also, the theory introduced, which posits that construction accidents are induced by perturbations in underlying networks of fundamental attributes, is promising but needs additional work to be further clarified and delineated. Specifically, attribute networks of injury cases need to be compared with that of "non-accident" cases, that is, random observations of the jobsite at times when no injuries are observed. Nevertheless, preliminary empirical evidence suggests that this theory may hold.

To our best knowledge, it is also the first time that hierarchical clustering is used to retrieve groups of atypical injury reports and inspect them for rare and atypical associations of attributes.

Such safety knowledge, based on binary attributes, is ideally suited for integration with systems such as BIM, and can be used to support a longitudinal approach of hazard identification and safety management that supports proactive decision making and provides information with increasing fidelity as project planning matures. In early phases, attributes could be assigned to physical elements and spaces in BIM, and the system, in turn, could automatically detect and flag safety clashes. For example, a designer would be able to identify and assign upstream attributes (i.e., those identifiable in design such as *steel sections* or *crane*) and the BIM system could then provide two useful pieces of information: (1) which clashes exist at a particular time and location based on the upstream attributes alone, and (2) the transitional and downstream attributes to which the situation is vulnerable. As the design matures to construction planning and work packaging, the attributes identified in the design phase are carried forward, new attributes are identified, and the model is refined accordingly. For example, transitional attributes (i.e., those identifiable in construction planning such as *forklift* or *overhead work*) can be added during work packaging as construction means and methods are selected. Finally, as construction begins, downstream attributes (i.e., those identifiable only once work begins such as *poor housekeeping* or *poor visibility)* can be added and expected clashes can be removed or communicated to the workforce during pre-task planning meetings.

Although our discussion of the implementation of the presented methodology and results is focused on BIM and AWP, the potential applications are far more extensive. We postulate that the ability to proactively identify safety clashes can provide useful information in any data-driven technology and many safety planning activities, including those that take place at the work site. As technologies and methods evolve and mature, the ability to identify and mitigate safety clashes is likely to remain beneficial and relevant. Given that attributes can be assigned and modeled in a binary fashion (i.e., present or absent), the algorithms can be robustly applied and simple user interfaces can be created.

Our focus with this paper was on a methodological advancement in the modeling of the interactions among units of analysis for construction safety analysis. Follow-up research should confirm/reject observations, tests competing theories, and expand the generalizability of the methodology and findings to other occupational contexts.

Acknowledgments

We would like to thank the National Science Foundation for supporting this research through an Early Career Award (CAREER) Program. This material is based upon work supported by the National Science Foundation under grant no. 1253179. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would also like to recognize Bentley Systems for their financial support for this research.

References

[1] A. Albert, M.R. Hallowell, B. Kleiner, A. Chen, M. Golparvar-Fard, Enhancing construction hazard recognition with high-fidelity augmented virtuality, J. Constr. Eng. Manag. 140 (7) (2014) 04014024.
[2] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais, S.J. Formosinho, Improving hierarchical cluster analysis: a new method with outlier detection and automatic clustering, Chemom. Intell. Lab. Syst. 87 (2) (2007) 208–217.
[3] R. Alsamadani, M. Hallowell, A.N. Javernick-Will, Measuring and modelling safety communication in small work crews in the US using social network analysis, Constr. Manag. Econ. 31 (6) (2013) 568–579.
[4] S. Baradan, M. Usmen, Comparative injury and fatality risk analysis of building trades, J. Constr. Eng. Manag. 132 (5) (2006) 533–539.
[5] J.P. Benzécri, L'analyse des données. Tome 1: La taxinomie. Tome 2: L'analyse des correspondances, 1973.
[6] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech: Theory Exp. 2008 (10) (2008), P10008.
[7] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, J. Math. Sociol. 2 (1) (1972) 113–120.
[8] S.P. Borgatti, Centrality and network flow, Soc. Networks 27 (1) (2005) 55–71.
[9] E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nat. Rev. Neurosci. 10 (3) (2009) 186–198.
[10] Bureau of Labor Statistics (BLS), Census of fatal occupational injuries (CFOI) – current and revised dataAccessed August 21, 2015, http://www.bls.gov/iif/oshcfoi1.htm 2013.
[11] E.C. Capen, The difficulty of assessing uncertainty (includes associated papers 6422 and 6423 and 6424 and 6425), J. Pet. Technol. 28 (08) (1976) 843–850.
[12] G. Carter, S.D. Smith, Safety hazard identification on construction projects, J. Constr. Eng. Manag. 132 (2) (2006) 197–205.
[13] S. Chae, T. Yoshida, A study of safety management using working area information on construction site, Proceedings of the 25th International Symposium on Automation and Robotics in Construction - International Association for Automation and Robotics in Construction, June, 292–99. Vilnius, Lithuania, 2008.
[14] P.S. Chinowsky, J. Diekmann, J. O'Brien, Project organizations as social networks, J. Constr. Eng. Manag. 136 (2009) 452–458 (SPECIAL ISSUE: Governance and Leadership Challenges of Global Construction).
[15] A.L.C. Ciribini, A. Gottfried, M.L. Trani, L. Bergamini, 4D modelling and construction health and safety planning, Proceedings of the 6th International Structural Engineering and Construction Conference - Modern Methods and Advances in Structural Engineering and Construction, 467–71. Zurich, Switzerland, 2011.
[16] A. Clauset, M.E. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (6) (2004) 066111.
[17] R. Collins, S. Zhang, K. Kim, J. Teizer, Integration of safety risk factors in BIM for scaffolding construction, Proceedings of Computing in Civil and Building Engineering, American Society of Civil Engineers 2014, pp. 307–314, http://dx.doi.org/10.1061/9780784413616.039.
[18] A. Costin, N. Pradhananga, J. Teizer, Passive RFID and BIM for real-time visualization and location tracking, Construction Research Congress, American Society of Civil Engineers 2014, pp. 169–178, http://dx.doi.org/10.1061/9780784413517.018.
[19] G. Csardi, T. Nepusz, The igraph software package for complex network research, Int. J. Complex Syst. 1695 (5) (2006) 1–9.
[20] A. Del Sol, R. Balling, L. Hood, D. Galas, Diseases as network perturbations, Curr. Opin. Biotechnol. 21 (4) (2010) 566–571.
[21] M. Desvignes, Requisite Empirical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems(Master thesis) University of Colorado at Boulder, 2014.
[22] S.N. Dorogovtsev, J.F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW, Oxford University Press, 2013.

[23] B. Esmaeili, Identifying and Quantifying Construction Safety Risks at the Attribute Level (Doctoral dissertation) University of Colorado at Boulder, 2012.
[25] B. Esmaeili, M.R. Hallowell, Diffusion of safety innovations in the construction industry, J. Constr. Eng. Manag. 138 (8) (2011) 955–963.
[26] B. Esmaeili, M.R. Hallowell, Using network analysis to model fall hazards on construction projects, *Safety and Health in Construction*, Conseil International du Batiment (CIB) W099, August 24–26, Washington DC, 2011.
[27] B. Esmaeili, M. Hallowell, Attribute-based risk model for measuring safety risk of struck-by accidents, Construction Research Congress, American Society of Civil Engineers May 2012, pp. 289–298, http://dx.doi.org/10.1061/9780784412329.030.
[28] J. Everett, Overexertion injuries in construction, J. Constr. Eng. Manag. 125 (2) (1999) 109–114.
[29] C. Fang, F. Marle, E. Zio, J.C. Bocquet, Network theory-based analysis of risk interactions in large engineering projects, Reliab. Eng. Syst. Saf. 106 (2012) 1–10.
[30] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[31] L.C. Freeman, Centrality in social networks conceptual clarification, Soc. Networks 1 (3) (1979) 215–239.
[32] T.M.J. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Software – Practice & Experience, Vol. 21 (11), Wiley 1991, pp. 1129–1164, http://dx.doi.org/10.1002/spe.4380211102.
[33] C.E. Fullerton, B.S. Allread, J. Teizer, Pro-active-real-time personnel warning system, Construction Research Congress 2009, American Society of Civil Engineers 2009, pp. 31–40, http://dx.doi.org/10.1061/41020(339)4.
[35] J. Goedert, P. Meadati, Integrating construction process documentation into building information modeling, J. Constr. Eng. Manag. 134 (7) (2008) 509–516.
[36] M. Greenacre, Correspondence Analysis in Practice, CRC press, 2007.
[37] P.E. Gustafsod, Gender differences in risk perception: theoretical and methodological perspectives, Risk Anal. 18 (6) (1998) 805–811, http://dx.doi.org/10.1111/j.1539-6924.1998.tb01123.x.
[38] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2) (2001) 107–145.
[39] M.R. Hallowell, J.A. Gambatese, Qualitative research: application of the Delphi method to CEM research, J. Constr. Eng. Manag. 136 (1) (2009) 99–107.
[40] M.R. Hallowell, J.A. Gambatese, Activity-based safety risk quantification for concrete formwork construction, J. Constr. Eng. Manag. 135 (10) (2009) 990–998.
[41] M. Hallowell, B. Esmaeili, P. Chinowsky, Safety risk interactions among highway construction work tasks, Constr. Manag. Econ. 29 (4) (2011) 417–429.
[42] A. Hammad, S. Setayeshgar, C. Zhang, Y. Asen, Automatic generation of dynamic virtual fences as part of BIM-based prevention program for construction safety, Proceedings of the 2012 Winter Simulation Conference (WSC), IEEE December 2012, pp. 1–10.
[43] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning (Vol. 2, No. 1), Springer, New York, 2009.
[44] J.W. Hinze, J. Teizer, Visibility-related fatalities related to construction equipment, Saf. Sci. 49 (5) (2011) 709–718.
[45] X. Huang, J. Hinze, Analysis of construction worker fall accidents, J. Constr. Eng. Manag. 129 (3) (2003) 262–271.
[46] S. Huang, I. Ernberg, S. Kauffman, Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective, Seminars in cell & developmental biology, vol. 20, No. 7, Academic Press September 2009, pp. 869–876.
[47] F. Husson, J. Josse, J. Pages, Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data? Applied Mathematics Department, 2010 (< http://factominer.free.fr/docs/HCPC_husson_josse.pdf >).
[48] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning (p. 6), Springer, New York, 2013.
[49] O. Jannadi, S. Almishari, Risk assessment in construction, J. Constr. Eng. Manag. 129 (5) (2003) 492–500.
[50] I. Joliffe, Principal Components Analysis, Springer, Verlag, 1986.
[51] P.F. Jonsson, T. Cavanna, D. Zicha, P.A. Bates, Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis, BMC Bioinf. 7 (1) (2006) 2.
[52] C. Kam, M. Fischer, R. Hänninen, A. Karjalainen, J. Laitinen, *The Product Model and Fourth Dimension Project*, ITcon Vol. 8, Special Issue IFC - Product Models for the AEC Arena, 2003 137–166 (http://www.itcon.org/2003/12).
[53] I. Kaner, R. Sacks, W. Kassian, T. Quitt, Case studies of BIM adoption for precast concrete design by mid-sized structural engineering firms. ITcon Vol. 13, Special Issue Case Studies of BIM Use, 2008 303–323 (http://www.itcon.org/2008/21).
[54] B. Karrer, E. Levina, M.E. Newman, Robustness of Community Structure in Networks. *arXiv preprint* arXiv:0709.2108, 2007.
[55] S.A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets, J. Theor. Biol. 22 (3) (1969) 437–467.
[56] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (5) (2009) 056117.
[57] T.H. Lin, C.H. Liu, M.H. Tsai, S.C. Kang, Using augmented reality in a multiscreen environment for construction discussion, J. Comput. Civ. Eng. (2014) 04014088.
[58] H. Lingard, Occupational health and safety in the construction industry, Constr. Manag. Econ. 31 (6) (2013) 505–514.
[59] A. Loureiro, L. Torgo, C. Soares, Outlier detection using clustering methods: a data cleaning application, Proceedings of KDNet Symposium on Knowledge-Based Systems for the Public Sector, 2004, http://www.digicult.info/pages/more.php?id=307 (Bonn, Germany).
[60] R. Navon, O. Kolton, Model for automated monitoring of fall hazards in building construction, J. Constr. Eng. Manag. 132 (7) (2006) 733–740.
[61] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (3) (2006) 036104.
[62] F. Palamara, F. Piglione, N. Piccinini, Self-organizing map and clustering algorithms for the analysis of occupational accident databases, Saf. Sci. 49 (8) (2011) 1215–1230.
[63] P. Pons, M. Latapy, Computing communities in large networks using random walks, Computer and Information Sciences-ISCIS 2005, Springer, Berlin Heidelberg 2005, pp. 284–293.
[64] S. Ponticelli, W.J. O'Brien, F. Leite, Advanced work packaging as emerging planning approach to improve project performance: case studies from the industrial construction sector, Proceedings of ICSC15: The Canadian Society for Civil Engineering 5th International/11th Construction Specialty Conference, University of British Columbia, Vancouver, Canada. June 7–10, 2015.
[65] M. Prades Villanova, Attribute-Based Risk Model for Assessing Risk to Industrial Construction Tasks(Master thesis) University of Colorado at Boulder, 2014.
[66] G.A. Radvansky, S.A. Krawietz, A.K. Tamplin, Walking through doorways causes forgetting: further explorations, Q. J. Exp. Psychol. 64 (8) (2011) 1632–1645.
[67] J. Reichardt, S. Bornholdt, Statistical mechanics of community detection, Phys. Rev. E 74 (1) (2006) 016110.
[68] P.R. Rose, Dealing with risk and uncertainty in exploration: how can we improve? AAPG Bull. 71 (1) (1987) 1–16.
[69] B. Ruhnau, Eigenvector-centrality—a node-centrality? Soc. Networks 22 (4) (2000) 357–365.
[70] R. Sacks, O. Rozenfeld, Y. Rosenfeld, Spatial and temporal exposure to safety hazards in construction, J. Constr. Eng. Manag. 135 (8) (2009) 726–736.
[71] S.E. Schaeffer, Graph clustering, Comput. Sci. Rev. 1 (1) (2007) 27–64.
[72] J. Shlens, A tutorial on principal component analysis. *arXiv preprint* arXiv:1404.1100, 2014.
[73] L.I. Smith, A tutorial on principal components analysis, Vol. 51, Cornell University, USA, 2002 52.
[74] J. Stelling, U. Sauer, Z. Szallasi, F.J. Doyle, J. Doyle, Robustness of cellular functions, Cell 118 (6) (2004) 675–685.
[75] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, ... J.L. Wrana, Dynamic modularity in protein interaction networks predicts breast cancer outcome, Nat. Biotechnol. 27 (2) (2009) 199–204.
[76] A.J.P. Tixier, M.R. Hallowell, A. Albert, L. van Boven, B.M. Kleiner, Psychological antecedents of risk-taking behavior in construction, J. Constr. Eng. Manag. (2014).
[77] A.J.P. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports, Autom. Constr. 62 (2016) 45–56.
[78] A. Tversky, D. Kahneman, The framing of decisions and the psychology of choice, Science 211 (4481) (1981) 453–458, http://dx.doi.org/10.1126/science.7455683.
[79] T.W. Valente, K. Coronges, C. Lakon, E. Costenbader, How correlated are network centrality measures? Connections 28 (1) (2008) 16 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875682/).
[80] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (301) (1963) 236–244.
[81] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.
[82] R.J. Yang, P.X. Zou, Stakeholder-associated risks and their interactions in complex green building projects: a social network model, Build. Environ. 73 (2014) 208–222.
[83] K. Yeh, M. Tsai, S. Kang, On-site building information retrieval by using projection-based augmented reality, J. Comput. Civ. Eng. 26 (3) (2012) 342–355.
[84] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.
[85] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.