HOW BIG DATA MAKES CONSTRUCTION PROJECT RISK INTACT

Ir Dr. Daniel Ng

Kun Hang Group and Jiangxi University of Finance and Economics, Hong Kong

Abstract: Construction project is not a standalone engineering maneuver. It is closely linked to the well-being of local communities in concern. The city renovation in Beijing down center for Olympic 2008 transformed many antique architecture and regional landscape. It gave a world-recognized achievement in China's modem development and manifested a major milestone in China's economic development. In the course of metro construction projects, there are substantial interwoven municipal structures influencing the success of the projects, which including, but the least, all underground cables and ducts, sewage system, the power consumption of construction works, traffic diversion, air pollution, expatriate business activities and social security. There are many US and UK project insurance companies moving into Asia Pacific. They are doing re-insurance business on major construction guarantee, such as machinery damage, project on-time, power consumption, claims from contractors and communities. Environmental information, such as water quality, indoor and outdoor air quality, people inflow and lift waiting time play deterministic roles in construction's fit-to-use. Big Data is a contemporary buzzword since 2013, and the key competence is to provide real time response to heuristic syndrome in order to make short-term prediction. This paper attempts to develop a conceptual model in big data for construction projects

Keywords: Construction project risk, big data, graph modelling

1. INTRODUCTION

Scale of construction project is huge in nature. One good example is the Beijing metropolitan renovation for Olympic 2008, which transformed many antique architecture and regional landscape to form a city landscape. It drove a world-recognized achievement in China's modem development which signatured a major milestone in China's economic development. It was not a standalone engineering manoeuvrer, but closely linked to the well-being of local communities across Beijing City. Metro construction projects involve substantial interwoven municipal structures influencing the success of the projects, which including, but not least, all underground cables and ducts, sewage system, the power consumption of construction works, traffic diversion, air pollution, expatriate business activities and social security.

Concurrently, there are many construction insurance companies moving into Asia Pacific from the UK and US. Their core offerings are risk mitigation and re-insurance business. They offer guarantee on unpredictable chaos, such as machinery damage, project delay, power failure, claims from contractors and communities, and environmental care issues, such as water quality, indoor and outdoor air quality, people inflow and lift waiting time. All of them are playing deterministic roles in construction project delivery.

Conventionally, various risk management tools and models are built to cater known construction project risk. Big Data is a contemporary buzzword. Its key competence is to provide real time response to heuristic syndrome in order to make short-term prediction. This paper attempts to exploit the construction risk management and evolve a big-data based conceptual model for construction projects, in particular, on the issue of power and gas pipeline.

2. CONSTRUCTION RISKS

There has been an increase in research on risk management practice in the construction industry. Moreover, little research has been conducted to systematically investigate the overall aspects of risk management on the perspectives of various project participants. There are findings on the importance of project risks, application of risk management techniques, status of the risk management system, and the barriers to risk management, which were perceived by the main project participants. The risk management strategies adopted in many China's national construction projects, such as "Three Gorges Project", revealed value of pre-emptive constructive risk management. Project risks are commonly of concern to project participants and the industry participants have shifted from risk transfer to risk reduction. Current risk management systems lack of joint risk management mechanisms to handle adequate risk management.

Risk management is an important part of the decision-making process in construction operation [13], and now widely accepted as a vital tool in the management of projects. A variety of risk management techniques has been studied and introduced in the literature along the processes of risk identification, risk analysis, risk response, and risk monitoring. Wood and Ellis [23] argues that the ultimate purpose of developing these risk management techniques is to add value to project delivery and improve efficiency of the construction industry during practice. Thus there has been an increase in research aimed at investigating risk management practice in the construction industry. In empirical studies across large construction contractors, risk management practices are mainly concerning:

- Perceptions of the typical towards construction risk allocation, and the importance of different risk categories [13];
- Usage of techniques at different risk management stages [2] ;
- Usage of risk management techniques and barriers to risk management in engineering construction industry [16] ;
- General contractors' perception on risks and the use of risk management techniques [1] ;
- Contractors' application of various analytical techniques for risk assessment [18] ;
- Various risks perceived by the contractors in Chinese construction market [9] ;
- Critical risks associated with China's build-operate-transfer BOTprojects and the effectiveness of mitigation measures [22] ;
- Perceptions of risk allocation in the construction industry of mainland China and Hong Kong [17] ;
- Allocation of risk in PPP/PFI construction projects in the United Kingdom [4] ;
- Contractual risk and liability sharing in hydropower construction [6] ;
- Practices of using risk management approaches in selected industries [20] ;
- Risk management services, tools, and techniques currently used by consultants [23] ;
- Risk management in the conceptual phase of a project [21] .

The construction industry is dynamic and hazardous due to the diverse and complex nature of work tasks, trades, and environment, as well as the temporary and transitory nature of construction workplaces and workforces [14]. Therefore, the risk of occupational accidents in the construction industry is far greater than in a manufacturing based industry [15]. The tremendous losses of human and economic resources caused by occupational accidents have become a serious problem in the construction industry

Accident analysis is used to identify factors contributing to occupational injuries and to develop strategies for injury prevention [12]. The analysis of aggregated accident data rather than single case analysis is considered as the only way of discovering any unifying and common factors in accident events [7]..Database technology is extensively applied in many domains, occupational accidents included. With the increasing use of databases, they should move from data processing aids to key strategic weapons for injury prevention. The large volume and high dimensionality of accidents databases leads to a breakdown in traditional human analysis. Data mining incorporates technologies for analysing data in large databases and can identify potentially useful patterns in the

data [25]. Weather database compliments accidents database. This makes human data analysis a difficult task, and data mining is a solution to the problem.

Association rule mining is an important task in data mining. Association rules can be effective in uncovering unknown relationships, and provide results for forecasting and decision making [19]. Weather condition is an unknown factor seldom discussed in association rule mining. This technique is employed in evaluating the associations between different factors and in identifying the patterns of industrial occupational injuries. Contributing factors to fatal occupational injuries have been identified with respect to individual factors, task factors, management factors, and environmental factors.

3. BIG DATA

Enterprises are exploring big data to discover facts they didn't know before. This is an important task right now because the recent economic recession forced deep changes into most businesses, especially those that depend on mass data. Using advanced analytics, businesses can study big data to understand the current state of the business and track still-evolving aspects such as risk behaviour. Bollier and Firestone [5] pinpoints that most business nowadays needing large volumes of highly detailed data to strive on. If management wants to see new thing, big data analytic helps to tap into data that's never been tapped for analytics. Some of the untapped data will be foreign to the operation, such as those coming from sensors, devices, third parties, Web applications, and social media. Some big data sources feed data unceasingly in real time. Put all that together, and you see that big data is not just about giant data volumes; it's also about an extraordinary diversity of data types, delivered at various speeds and frequencies. First, there's big data for massive amounts of detailed information. Second, there's advanced analytics, which is actually a collection of different tool types, including those based on predictive analytics, data mining, statistics, artificial intelligence, natural language processing, and so on.

Most definitions of big data focus on the size of data in storage. Size matters, but there are other important attributes of big data, namely data variety and data velocity. The three Vs of big data (volume, variety, and velocity) constitute a comprehensive definition, and they bust the myth that big data is only about data volume. In addition, each of the three Vs has its own ramifications for analytics. It's obvious that data volume is the primary attribute of big data. With that in mind, most people define big data in terabytes—sometimes petabytes. Many users are managing 3 to 10 terabytes (TB) of data for analytics. Yet, big data can also be quantified by counting records, transactions, tables, or files. Some organizations find it more useful to quantify big data in terms of time. For example, due to the seven-year statute of limitations in the U.S., many firms prefer to keep seven years of data available for risk, compliance, and legal analysis. The scope of big data affects its quantification, too. For example, in many organizations, the data collected for general data warehousing differs from data collected specifically for analytics. Different forms of analytics may have different data sets. Some analytic practices lead a business analyst or similar user to create ad hoc analytic data sets per analytic project. Then, there's the entire enterprise, which has its own, even larger scope of big data quantifications of big data grows continuously. All this makes big data for analytics a moving target that's tough to quantify.

Big data can be described by its velocity or speed. You may prefer to think of it as the frequency of data generation or the frequency of data delivery. For example, think of the stream of data coming off of any kind of device or sensor, say robotic manufacturing machines, thermometers sensing temperature, microphones listening for movement in a secure area, or video cameras scanning for a specific face in a crowd. The collection of big data in real time isn't new; many firms have been collecting clickstream data from Web sites for years, using streaming data to make purchase recommendations to Web visitors. With sensor and Web data flying at you relentlessly in real time, data volumes get big in a hurry. Even more challenging, the analytics that go with streaming data have to make sense of the data and possibly take action—all in real time.

4. ANALYTICS FOR RISK CORRELATION

Big data is not just big. It's also diverse data types and streaming data. Big data analytics is the application of advanced analytic techniques to very big data sets and explores granular details of business operations and customer interactions that seldom find their way into a data warehouse or standard report. Some organizations are already managing big data in their enterprise data warehouses (EDWs), while others have designed their DWs for the well-understood, auditable, and squeaky clean data that the average business report demands [11]. The former tend to manage big data in the EDW and execute most analytic processing there, whereas the latter tend to distribute their efforts onto secondary analytic platforms. There are also hybrid approaches. Regardless of approach, user organizations are currently re-evaluating their analytic portfolios. In response to the demand for platforms suited to big data analytics, vendors have released a slew of new product types including analytic databases, data warehouse appliances, columnar databases, no-SQL databases, distributed file systems, and so on. There are also new slew of analytic tools.

Big data correlation is where advanced analytic techniques operate on big data sets. It is really about two things, big data and analytics. The end product is to create one of the most profound trends in business intelligence. Instead of "advanced analytics," a better term would be "discovery analytics," because that's what users are trying to accomplish. (Some people call it "exploratory analytics."). Ferreira et al [10] defines big data analytics is typically a business analyst who is trying to discover new business facts that no one in the enterprise knew before. To do that, the analyst needs large volumes of data with plenty of detail. This is often data that the enterprise has not yet tapped for analytics. For example, in the middle of the recent economic recession, companies were constantly being hit by new forms of customer churn. To discover the root cause of the newest form of churn, a business analyst would grab several terabytes of detailed data drawn from operational applications to get a view of recent customer behaviours. The analyst might mix that data with historic data from a data warehouse. Dozens of queries later, the analyst would discover a new churn behaviour in a subset of the customer base. With any luck, that discovery would lead to a metric, report, analytic model, or some other product of BI, through which the company could track and predict the new form of churn. Discovery analytics against big data can be enabled by different types of analytic tools, including those based on SQL queries, data mining, statistical analysis, fact clustering, data visualization, natural language processing, text analytics, artificial intelligence, and so on. It's quite an arsenal of tool types, and savvy users get to know their analytic requirements before deciding which tool type Is appropriate to their needs.

Big data provides gigantic statistical samples, which enhance analytic tool results. Most tools designed for data mining or statistical analysis tend to be optimized for large data sets. In fact, the general rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis. Instead of using mining and statistical tools, many users generate or hand-code complex SQL, which parses big data in search of just the right customer segment, churn profile, or excessive operational cost. The newest generation of data visualization tools and in-database analytic functions likewise operate on big data. Analytic tools and databases can now handle big data. They can also execute big queries and parse tables in record time. Recent generations of vendor tools and platforms have lifted us onto a new plateau of performance that is very compelling for applications involving big data. Analytics based on large data samples reveals and leverages business change. The recession has accelerated the already quickening pace of business. The recovery, though welcome, brings even more change. In fact, the average business has changed beyond all recognition because of the recent economic recession and recovery. The change has not gone unnoticed. Businesspeople now share a wholesale recognition that they must explore change just to understand the new state of the business.

5. HIGH DATA DIMENSIONALITY

In big data analytics process, there are iterative measures of classification and clustering [24] to identify numerically significant data element groups. Regression techniques are employed to restrict the focus of estimation to few dependent and independent variables. This forms the dimensionality of data set in concern and this particular characteristic conduces to the computational analytic systems, or machine learning, to dissect and uncover the intrinsic data architecture. Traditional relational databases are developed initially through the use of normalization to remove structural duplication for faster access such that all data elements contained in relational databases are cleansed under well-defined taxonomy. This can assure the speed of retrieval. On the contrary, big data sets contain high degree of data dimensionality. In a super large sample of data points and observations on variables, there are interlinked and interwoven connections across acquired data points. In case of big data noted in social community and financial market, the trends and behaviours in variables' observations could be far bigger, larger, voracious, automatic, systematic and hyper-informative. This induces a higher degree of data dimensionality in the big data set. Practical examples are curves, or spectra, or images, or even movies such that a single observation has dimensions in the thousands or billions; whilst there are only tens or hundreds of instances of variables available for examination.

Data dimensionality is a significant issue to overcome when constructing inferential analysis and machine learning algorithm against super granular data set. A material dimension reduction renders a viable model to describe the correlation among internal data clusters. Common methods, such as K-N-N clustering, are able to cope with limited growth of dimensionality in the conventional database systems. In case of big data system composed of highly unstructured data, the key/value matching and similarity matching creating a high-dimensional data analysis requesting for new and working solution [8]. Further, there is an intrinsic problem in other classical tools of dimension reduction, such as principal component analysis and multidimensional scaling. Most of these methods are constructed out the low optimization approach in the course of dimension detection and reduction. None of them explicitly consider the complex correlated data structure embedded in the high dimensional data set.

Conventional stochastic methods, such as principal component analysis and K-means, utilize repetitive ordinary least squares to isolate and group data points into clusters so as to estimate the correlation portion of a particular clusters of data points towards the generalized sample variation. However, biological data visualization for drug effectiveness and genomics analysis involve substantially the issue of dimensionality. The interaction among chemical components and testing environments generate many variation differences in measuring variables such that high co-linearity exists across presumed dependent and independent variables. Drugs, such as those curing terminal diseases like cancer, normally show complicated chemical reactions with human body's immunity systems on one hand; and might affect the functionality of other supplementary medicines [3]. In the course of determining that core inferential variables, the clustering process is complicated by the high dimensionality appearing in the collected data points. There could be millions of drug samples in various concentration and dosage to render the clustering process impossible to complete.

6. CONSTRUCTION HOLISTIC ANALYTICS

This approach is to distinguish the signals in risk data from the noise collected in sample data. The on-going exponential growth and variety of human communications methods cause organizations creating massive data to analyse seeking to uncover continually evolving threats in construction projects. The variety of sources composed of poorly structured data has made analysis more complex. In particular, there is a growing emphasis on collecting and analysing human communications such as email, voice, instant messaging and social media.

Holistic approach to analytics does more than traditional text and predictive analytics systems. It calls for a new way of detecting risks and relationships: One that can connect the dots seamlessly across traditional data sources,

electronic communication systems and public sources of information. The core component is machine learning-based analytics platform to perform semantic dissertation. This can reveal and organise people, places and activities into a private knowledge graph of connected risks. Inside, there is no predefined rule or pattern, or an understanding of the underlying data ontology. This approach is three-pronged in nature:

- Read: Initial processing, performs entity and fact extraction. Suspicious Activity Reports (SARs) are processed and key information (financial transactions, account numbers, account owners, addresses, financial transactions, etc.) is automatically identified by big data's machine learning routines; .
- Resolve: Assemble, organize and relate information. Resolve entities, performs categorization and identifies similarities. Key information within the SAR is resolved and correlated with data contained in a global knowledge base providing a holistic view of network and relationship connectivity.
- Reason: Uncover, compare and correlate information, perform temporal and geospatial reasoning and produce a private knowledge graph. Linkages to risk items are uncovered and surfaced in an easy-to-understand format.

Information in context is synthesized more rapidly. This holistic approach to data analysis reduces the signal to noise ratio in large and messy data sets. Graph database model is employed to show the structures and schema within risk items. An implicit definition is given on those risk items in graphs to show the semantic linkages. Explicit graphs and graph operations allow users to express a query at a high level of abstraction. To some extent, this enables graph manipulation in deductive databases in order to solve fairly complex rules.

7. CONCLUSION

Big data technology traditionally furnishes numerical output for analysis. Graph modelling is intuitive in presentation and manipulation. The correlation analytic from big data reveals the interdependence among construction risk items so as to estimate the aggregated construction risk, both in term of scale and impact.

This graph based holistic analytic should provide a systematic big data risk management for metropolitan construction projects.

8. REFERENCES

[1] Akintoye, A. S., and Macleod, M. J. (1997) Risk analysis and management in construction. *International Journal of Project Management*, 15(1), 31–38.

[2] Baker, S. (1999). Risk response techniques employed currently for major projects. *Construction Management Economics*, 17, 205–213

[3] Belkin, M., and Niyogi, P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6), 1373-1396.

[4] Bing, L., Akintoye, A., Edwards, P. J., and Hardcastle, C. (2005) The allocation of risk in PPP/PFI construction projects in the UK." *International Journal of Project Management*, 23, 25–35.

[5] Bollier, D., and Firestone, C. M. (2010) The promise and peril of big data: Aspen Institute, *Communications and Society Program* Washington, DC, USA.

[6] Charoenngam, C., and Yeh, C. (1999) Contractual risk and liability sharing in hydropower construction. *International Journal of Project Management*, 17(1), 29–37.

[7] Chi, C.F., Chang, T.C. and Hung, K.H. (2004) Significant industry-source of injuries-accident type for occupational fatalities in Taiwan. *International Journal of Industrial Ergonomics* 34, 77–91.

[8] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1-32.

[9] Fang, D., Li, M., Fong, P. S., and Shen, L. (2004) Risks in Chinese construction market—Contractors' perspective. *Journal of Construction Engineering Management*, 130 (6), 853–861.

[10] Ferreira Cordeiro, R. L., Traina Junior, C., Machado Traina, A. J., López, J., Kang, U., & Faloutsos, C. (2011) Clustering very large multi-dimensional datasets with MapReduce. *Paper presented at the Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*.

[11] Gavrishchaka, V. V., and Ganguli, S. B. (2003) Volatility forecasting from multiscale and high-dimensional market data. *Neurocomputing*, 55(1), 285-305.

[12] Harper, R.S. and Koehn, E. (1998) Managing industrial construction safety in Southeast Texas. *Journal of Construction Engineering and Management* 124, 452–457.

[13] Kangari, R. (1995). Risk management perceptions and trends of U.S. Construction *Journal of Construction Engineering Management*, 121(4), 422–429

[14] Kines, P. (2002) Construction workers' falls through roofs: fatal versus serious injuries. *Journal of Safety Research*, 33, 195–208.

[15] Larsson, T.J. and Field, B. (2002) The distribution of occupational injuries risks in the Victorian construction industry. *Safety Science* 40, 439–456.

[16] Lyons, T. and Skitmore, M. (2003) Project risk management in the Queensland engineering construction industry: A survey. *International Journal of Project Management*, 22, 51–61.

[17] Rahman, M. M., and Kumaraswamy, M. M. (2002) Joint risk management through transactionally efficient relational contracting. *Construction Management Economics*, 20, 45–54.

[18] Shen, L. Y.(1997) Project risk management in Hong Kong. *International Journal of Project Management*, 15(2), pp. 101–105.

[19] Tsay, Y.J. and Chiang, J.Y. (2005) CBAR: an efficient method for mining association rules. *Knowledge-Based Systems* 18, 99–105.

[20] Tummala, V. M. R., Leung, H. M., Mok, C. K., Burchett, J. F., and Leung, Y. H. (1997) Practices, barriers and benefits of using risk management approaches in selected Hong Kong Industries. *International Journal of Project Management*, 15(5), 297–312.

[21] Uher, T. E., and Toakley, A. R. (1999) Risk management in the conceptual phase of a project. *International Journal of Project Management*, 19(3), 161–169.

[22] Wang, S. Q., Tiong, L. K., Ting, S. K., and Ashley, D. (1999) Political risks: Analysis of key contract clauses in China's BOT project. *Journal of Construction Engineering Management*, 125(3), 190–197.

[23] Wood, G. D., and Ellis, R. C. T. (2003) Risk management practices of leading UK cost consults. *Engineering, Construction, Architecture Management*, 10(4), 254–262

[24] Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. Knowledge and Data Engineering, *IEEE Transactions* , 26(1), 97-107.

[25] Zhang, C. and Zhang, S. (2002). *Association Rule Mining: Models and Algorithms*. Springer, New York.