VISUALISING CLOSE CALL IN RAILWAYS: A STEP TOWARDS BIG DATA RISK ANALYSIS

Miguel Figueres-Esteban[1], Peter Hughes[1] and Coen Van Gulijk[1].

[1]*University of Huddersfield, Institute of Railway Research, Huddersfield, UK*

In the Big Data era new data sources are available to get insight from human factors in railways. Close Call System (CSS) is one of the data sources which are being researched in the Big Data Risk Analysis (BDRA) project to extract valuable information for risk management. One of the key challenges of BDRA is the visualisation of a large amount of information into a simple and effective display to risk analysis and making-decisions. In this paper we present the research in converting the free text from Close Call data into a spatial representation of networks of words and perform the text visual analysis in order to identify risk categories. For a small number of Close Call records related to level crossings, trespasses and slips, falls and trips, it was possible to identify the different scenarios. Moreover, the results provide an understanding of how Close Call events are described and how it might influence safety on the railways.

Introduction

A significant part of computer research is dedicated to new software tools that are loosely identified as Big Data (Chen et al. 2012; McAfee & Brynjolfsson 2012; Watson & Marjanovic 2013). These new technologies offer industry and academia new opportunities for analysing and understanding complex processes. In the GB railways, Network Rail and ATOC are dedicating part of their work to provide live data-feeds about trains, tracks or incidents (ATOC 2015; NR 2015). This data contains information that may be relevant for safety and risk in the GB railways. The Institute of Railway Research attempts to utilize this data to improve safety and risk management for the GB railways in a research project that is called Big Data Risk Analysis (BDRA). The aim of BDRA is to support risk analysis and safety decision-making from a wide range of diverse data sources improving the understanding of risk factors involved in railways (Van Gulijk et al. 2015). One of the key challenges of BDRA is the visualisation of a large amount of information into a simple and effective display. Visualisation in BRDA does not just mean developing visual techniques to represent outputs, but also the ability to support complex analyses with interactive visualisations (Figueres-Esteban et al. 2015).

This work presents the research in applying visual analytics in one of the data sources that is being researched in BDRA: text-based records from the GB Railways' Close Call database. For a small number of Close Call records related to level crossings, trespasses and slips, falls and trips, we have adapted the text in networks as the best way to represent the Close Call for effective safety data analysis and decision-making. Through a visual text analysis it was possible to identify the different risk scenarios. Moreover, the results provided an understanding of how Close Call events are described and how it might be used for better safety on the railways.

Background

Quantitative text analysis

Three different computing text analysis approaches are possible for retrieve information from text: thematic, semantic and networks (Popping, 2000). Thematic analysis has been the main approach for a long time and it is based on the frequency of concepts (e.g. words or "bag of words") that allows classification of the topics of texts. Semantic analysis also takes into account the relationships among the concepts encoding the semantic grammar (e.g. subject, verb and object). Network analysis is based on network text analysis to obtain semantically linked concepts. For these approaches diverse text mining techniques (e.g. for automated information retrieval), natural language processing techniques (e.g. for tokenization, stemming or parsing) and visual text analytics systems (e.g. network analysis) have widely been developed (Drieger 2013). Human vision may help to improve the interpretation of text when it is transformed in a spatial representation, reducing the workload and increasing the analytical processes (Crow et al. 1994). That is why graph analysis has received much attention over the last years.

Basic concepts of graphs

A graph $G = [N, L, f]$ is a set of *nodes* N (i.e. vertices) and *links* L (i.e. edges) that are connected into pairs of nodes by the *mapping function*$f: L \rightarrow N \times N$. In a graph the number of links connecting a node is called the *degree* of the node. A *path* is a sequence of linked nodes and the *length of a path* is the number of links between the first and last nodes of the path (Lewis 2011). One of the main measures of centrality is the *betweenness* of nodes, which is defined by Freeman (1978) as the frequency with which a node falls between pairs of other nodes on the shortest paths connecting them.

Network text analysis

A common method for visual text analysis is to represent the text as a graph: the words or concepts are the nodes, and their relationships are the links (Drieger 2013; Paranyushkin 2011; Popping 2003). This text analysis allows better understanding of the text than a simple word frequency analysis since it is possible to analyse the strength of relationships among main concepts from a text, and thus, gather relevant information from the graph. Key attributes in a text-graph analysis are: the *degree of a node* as an indicator of the importance of a concept, and the *betweenness of nodes*, which gives information about the diversity of the concept (Paranyushkin 2011; Popping 2000). These attributes form the backbone of the analysis that was made as part of this paper.

Close Call

In the GB Railways, potentially dangerous events are reported by railway workers or general public in the Close Call System (CCS) or Confidential Incident Reporting and Analysis System (CIRAS). In this paper we have selected Close Call records from the CCS, which is a semi-structured database where each record provides categorised data and free-text. Gathering safety information from close call means retrieving information from the freeform text of each record (Hughes et al. 2015).

Methodology

In order to test value of network text analysis in Close Call free text descriptions, we have followed the data extraction methodology proposed by Paranyushkin (2011). Since the objective was to assess whether risk groups could be identified through visual analysis, a pre-constructed dataset was used. A sample of 150 records was constructed from selecting the first 50 data records from the Close Call database classified as "Trespass", "Slip/Trip hazards on site" and "Level crossing". These records were cleaned of non-desired characters using the NLTK toolkit in Python (Bird et al. 2009) in order to generate the text source to process (cleaned text). The "*tagging process*" and "*tokenization process*" described in Hughes et al. (2015) was used to create the two types of text for visualising. The visual analysis of the tagged-text provided information to tailor the tokenization process (removing stopwords and stemming plurals or verbs), avoiding obscuring main concepts in the tokenized-text network. The entire process of cleaning, tagging and tokenization is illustrated in the Table 1. The visual analysis was performed by constructing the 2-word gap and 5-word-gap networks and representing the networks with the Gephi software (Paranyushkin 2011).

Table 1: Cleaning, tagging and tokenization processes in one trespass record

Original record
Emailed report from LOM<br /> <br /><p class="MsoNormal"><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt">Date: 08/09/13 </span></font><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt">Time: 1900<p></p></span></font> </p><p></p><p></p><p class="MsoNormal"><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt"><p> <font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt">ELR: LEN3 59m 14ch</span></font><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt"> </span></font></p></span></font></p><p></p><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt">Issue – Trespasser on the line in the

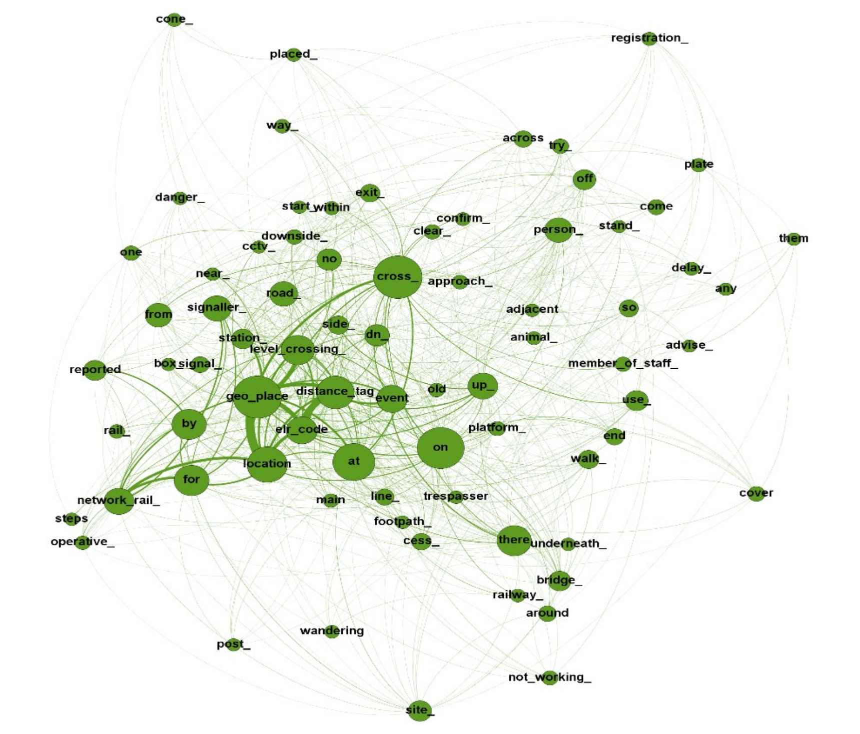
Hartburn Junction area. Trains cautioned, reported all clear by MOM @ 1930</span></font><p></p><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt">Action – Fencing to be checked 09/09/13<p></p></span></font><p></p><p class="MsoNormal"><font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt"><p> <font face="Arial" size="2"><span style="font-family: arial; font-size: 10pt">DU: <city></city><place></place> Newcastle<p></p> </span></font></p></span> </font></p><p></p><!-- RICH TEXT -->
Cleaned record
Emailed report from LOM Date: 08/09/13 Time: 1900 ELR: LEN3 59m 14ch Issue – Trespasser on the line in the Hartburn Junction area. Trains cautioned, reported all clear by MOM @ 1930 Action – Fencing to be checked 09/09/13 DU: Newcastle
Cleaned and tagged record in lowercase
emailed report from local operation manager date _date_ time _time_ elr_code distance_tag issue trespasser on the railway line in the geo_place junction area trains cautioned reported all clear by mobile operations manager _time_ action fencing to be checked _date_ geo_place
Cleaned, tagged and tokenized record without stopwords and in lowercase
email report from local operate_ manager_ date _date_ time _time_ elr_code distance_tag issue trespasser on railway_ line_ in geo_place junction_ area_ train_ warning_ reported all clear_ by mobile_operations_manager_ _time_ action fence_ check_ _date_ geo_place

Text graph visualisation

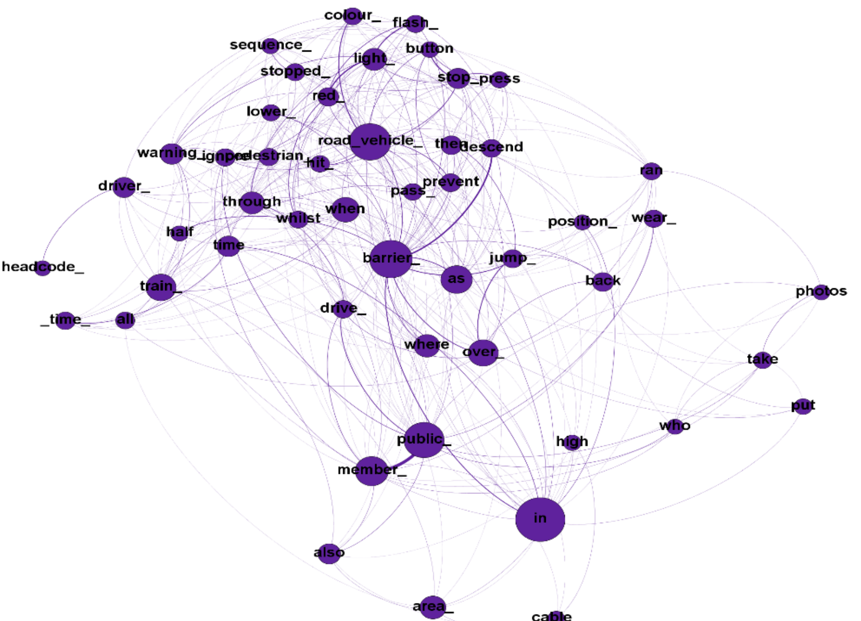
The force-layout algorithm *Force Atlas* was applied to construct the network graphs. The networks obtained from the tagged and tokenized text are composed by nodes related to tags (e.g. *geo_place, elr_code* or *distance_tag*), tokens (e.g. level_crossing_, *road_vehicle_, access_* or *network_rail_*) and words (e.g. *location, trespasser* or *pedestrian*). In order to gather knowledge from the networks the size of the nodes have been ranged by degree to analyse which words are related to each other and by betweenness to analyse the contexts where the words appear (Paranyushkin 2011). The Louvain method for community detection with enough resolution has been applied to represent large clusters from the networks (Blondel et al. 2008).

Results

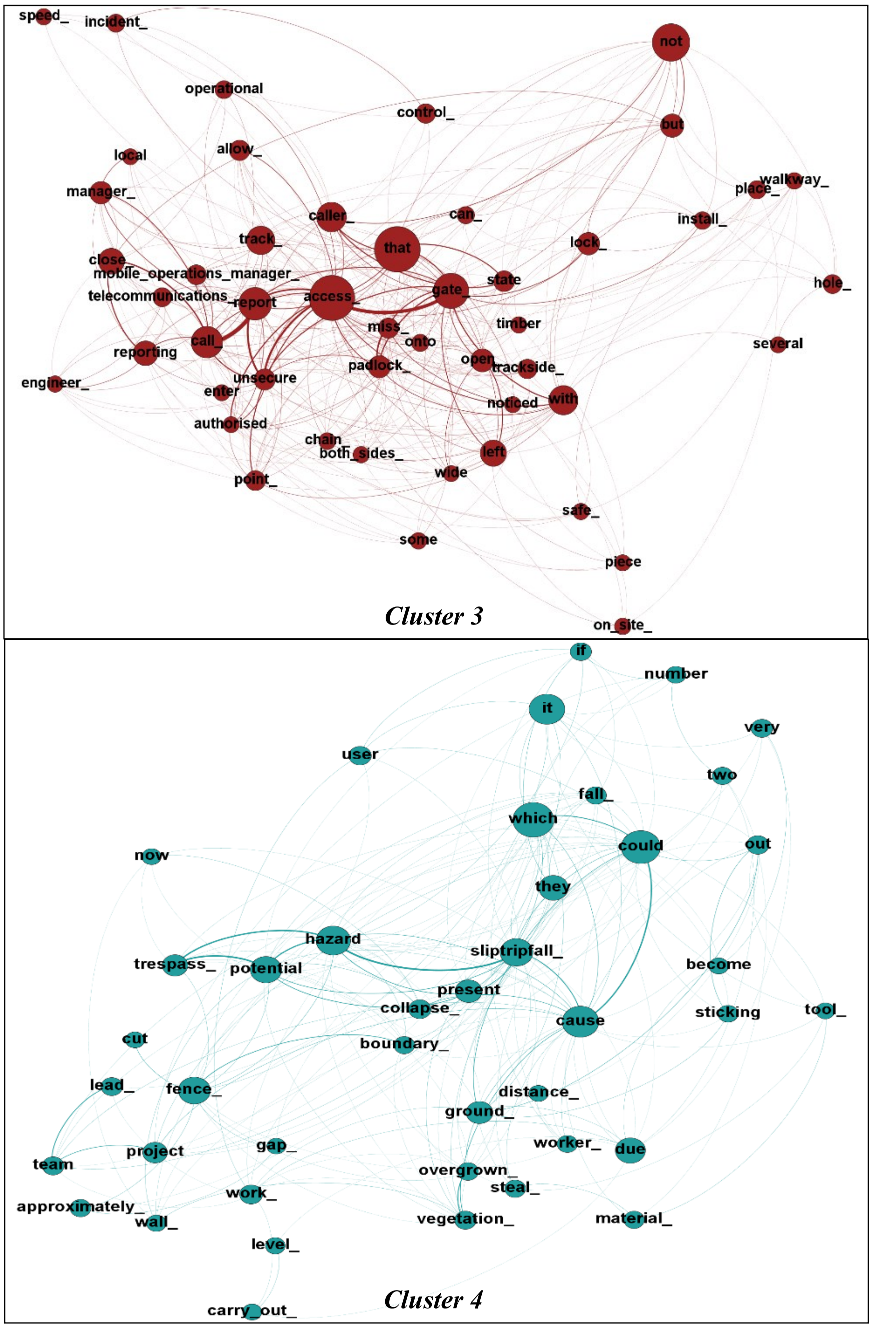
The Louvain community detection algorithm has detected four clusters from the 5-word-tokenized text network with a resolution of 1.5. The result gives a modularity of 0.6 (Paranyushkin 2011). The Figure 2 represents the clusters filtering the nodes larger than 20 degree.



Cluster 1



Cluster 2



Cluster 3

Cluster 4

Figure 2. Sub-networks that represent the clusters of degree nodes (28.35%, 23.74%, 23.04% and 24.86% of nodes) from the 5-word gap – tokenized network (Resolution=1.5; modularity=0.6). Filtered by 20 degree node.

The first, second and third cluster have the highest degree nodes with a high betweeness (*cross_, geo_place, distance_tag, location, barrier_*, *access_*, gate_ and *road_vehicle_*). In addition we can also find a great quantity of high and medium degree nodes related to level crossings (*elr_code*, *level_crossing, road, driver_, red_, light_, flash_*, *warning_, miss_, padlock_, unsecure, point, track, trackside_, lock, open_, enter, safe_* or *authorised*). However, the clusters present important differences regarding nodes related to people and type of terms used. The first cluster encloses nodes related to technical staff (for example *network_rail_, operative, member_of_staff or signaler*) and terms such as *box_signall, cctv_, elr_code, cess, not_working*, *approach_, clear_, main_, line_, dn_, up_* or *downside*. The second is more related to general public (*member_* and *public_*) and road safety terms such as *road_vehicle, barrier_, light_, red_, descend_, stop_, button* or *press*. The third cluster shows many nodes related to technical staff (for example *mobile_operations_manager, operational, telecommunications, manager* or *engineer*) and terms such as *close_, call_, access_, gate_, miss_, padlock_, unsecure, point, track, trackside_, lock, open_, enter, safe_* or *authorised.*

The fourth cluster displays high degree nodes for example *hazard_, potential_, trespass_ or sliptripfall_*, nodes related to people like *worker_* or *user* and terms related to the workforce environment such as *tool_, gap_, wall_, sticking, cut_, fence_, boundary_, overgrown_* or *vegetation_*.

Discussion

In this paper we have explained the process to convert the free text from Close Call System into a spatial representation of networks and perform the text visual analysis. Text pre-processing was applied to 150 Close Call events that describe three scenarios (level crossings, trespasses and slip, falls and trip hazards) in order to obtain a cleaned text source that allowed us to carry out the tagging and tokenization process. The information obtained from the visual analysis of the tagged-text was used to tailor the tokenization process. Although the quality of the analysis could be improved by several iterations between the graph analysis and the tokenization process, only a single iteration was performed for this work. Moreover, it was found that the 2-word gap and 5-word gap steps described in Paranyushkin 2011 to build the networks worked very well for our analysis.

The final network to analyse (5-word gap tokenized-text) is a small and quite interconnected network composed by nodes that represent tags, tokens and words from the close call text.

Using the Louvain community detection algorithm with a resolution of 1.5 for detecting large clusters, four clusters were identified that relate to the scenarios that were selected for the analysis. This clustering network technique is based on looking for relationships among concepts, showing clusters of nodes which are strongly connected. Therefore, the network analysis does not show what type of scenarios we have, but what type of words, tags and tokens are used to describe topics.

The first, second and third clusters contain the more high betweennes nodes, which means that these three clusters are connected. The high degree nodes are mainly related to level crossings terms (*level_crossing_, barrier_, road_vehicle_, access_* and *gate_*). The difference among them are the nodes that describe people and the type of terms reported by groups of people. Thus, what we are visualising is the way that different people describe a level crossing scenario rather than identify the level crossing scenario itself. The first and third cluster show more technical and operational railway terms that technical staff (e.g. *signaller_, manager_* or *engineer*) use for reporting (e.g. *elr_code, box_signal, main_, line_, up_, dn_, CCTV, track_side, authorised, unsecure, lock, miss_* or *padlock_*) whilst the second shows that general public report more about road safety issues (*road_vehicle_, pedestrian, ignore, warning, red flash light* or *press stop button*). Moreover, the first cluster includes a medium degree node associated with trespasses (*trespasser*). This finding might mean that the terms used to describe a level crossing may also usually be used to describe trespasses (*cross_, geo_place, location, station_ or platform_*).

The fourth cluster shows very high degree nodes related to trespass and slips, trips and falls events (*trespass_ and sliptripfall_*). It is theorized that these records are made by track workforce who mainly report about slips, trips and falls. Moreover, the terms reported may be associated with work activities (e.g. *tool_, gap_, fence_, wall_, boundary_, fall_, overgrown_* or *vegetation_*). The high degree nodes related to trespasses might mean that are the workers who also usually report about trespasses. That may explain why the first cluster, also related to technical staff, contains nodes about trespasses.

In summary, we are able to see that words which may be inherent to a one type of scenario are also used to describe other scenarios, that is, people essentially use the same words to describe different types of scenarios. For example, we can see how people use the word trespasser to describe level crossing users who ignore red lights. Despite of that, it might be possible to interpret that nodes with high betweeness are indicating similar clusters and, thus, to identify the scenarios across the network clusters (figure 3).

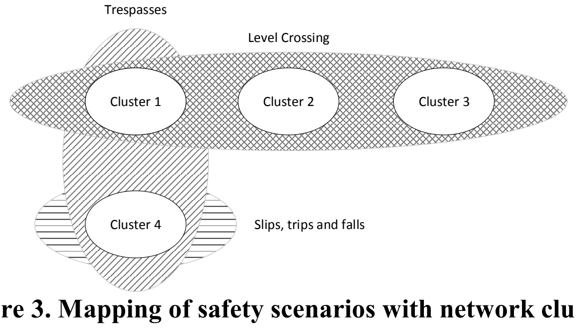


Figure 3. Mapping of safety scenarios with network clusters

Figure 3 shows that a high overlap exists among the scenarios selected considering how people describe them. It might be interpreted that the public do not use the definitions or terms that experienced railway staff uses to define scenarios. This means that it may be specially challenging to detect different risks from the free text in CCS; the individuals entering the data do not make a distinction between scenarios of similar kind. This also renders machine learning techniques useless because they cannot distinguish different scenarios if the reporter doesn't.

Finally, the biases of the analyst has to be considered. The choice of 50 records from each subcategory of the CCS does not yield text-book records for these categories. Since we have seen that member of public do not describe safety scenarios in the same way that railway safety staff do, there might not be such a thing as a textbook trespass Close Call record. The analyst has to let go of their preconceptions about the clusters that they are looking for (three risk clusters) to be able to see that the clustering simply yields another result (namely types of people entering Close Call records).

Conclusion

Visual analysis of 150 close call records has provided a valuable insight into Close Call reporting. This paper demonstrates that it is possible to represent a large quantity of information (nodes that represent words or concepts) by networks which gives an overview of which type of nodes are more related to others and the level of relationships among them. It is possible to map the network clusters with the scenarios selected based on "*degree*" and "*betweenness*" analysis. The network clusters show that different groups of people (railway staff or lay persons) use different language to describe what they observe into the CCS. Fortunately, the risk categories were not completely lost in the analysis but it does take more interpretation from the analyst to extract relevant risk information from the data. In the future it is worth considering more tailored text pre-processing (cleaning, tagging and tokenization), alternative data extraction methods (n-word gap steps) and other network clustering methods.

Acknowledgement

RSSB is gratefully acknowledged for co-funding this work through the Strategic Partnership MoU of the 8th of August 2013.

References

ATOC, 2015. ATOC 2015. Available at: http://www.atoc.org/about-atoc/national-rail-enquiries/access-to-information [Accessed June 5, 2015].

Bird, S., Klein, E. & Loper, E., 2009. *Natural Language Processing with Python.*

Blondel, V.D. et al., 2008. Fast unfolding of community hierarchies in large networks. *Networks*, pp.1–6.

Chen, H., Chiang, R.H.L. & Storey, V.C., 2012. Business intelligence and analytics: from big data to big impact. *MIS Quarterly*, 36, pp.1165–1188.

Crow, V., Pottier, M. & Thomas, J., 1994. *Multidimensional visualization and browsing for intelligence analysis*, Pacific Northwest Lab., Richland, WA (United States).

Drieger, P., 2013. Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences*, 79, pp.4–17.

Figueres-Esteban, M., Hughes, P. & Van Gulijk, C., 2015. The role of data visualization in Railway Big Data Risk Analysis. In *Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*. p. 7.

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. *Social Networks*, 1, pp.215–239.

Van Gulijk, C. et al., 2015. Big Data Risk Analysis for Rail Safety? In *Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*.

Hughes, P., Van Gulijk, C. & Figueres-Esteban, M., 2015. Learning from text-based close call data. In *Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015*. p. 8.

Lewis, T.G., 2011. *Network science: Theory and applications*, John Wiley & Sons.

McAfee, A. & Brynjolfsson, E., 2012. Big Data. The management revolution. *Harvard Business Review*, 90, pp.61–68.

NR, 2015. NR 2015. Available at: http://www.networkrail.co.uk/data-feeds [Accessed June 5, 2015].

Paranyushkin, D., 2011. Identifying the Pathways for Meaning Circulation using Text Network Analysis. *Nodus Labs*, pp.1–26.

Popping, R., 2000. *Computer-Assisted Text Analysis* SAGE Publi.,

Popping, R., 2003. Knowledge Graphs and Network Text Analysis. *Social Science Information*, 42, pp.91–106.

Watson, H.J. & Marjanovic, O., 2013. Big Data: The Fourth Data Management Generation. *Business Intelligence Journal*, 18(3), pp.4–8.