THE EVOLUTION OF TERMINOLOGY WITHIN A LARGE DISTRIBUTED ENGINEERING PROJECT

Gopsill, James Anthony (1); Jones, Simon (2); Snider, Chris (1); Shi, Lei (2); Hicks, Ben James (1)
1: University of Bristol, United Kingdom; 2: University of Bath, United Kingdom

Abstract

Communication features in many engineering activities within an engineering project. It is the main form by which information & knowledge is shared, and facilitates the generation of a shared understanding between engineers. While there exists a significant body of literature relating to communication, much of the research has been through qualitative studies using techniques such as surveys, interviews and observation. Given the prevalence of computer-mediated communication and the development of techniques to analyse such datasets, there is now the opportunity to provide quantitative metrics that can characterise communication.
Therefore, this paper examines this opportunity through the co-word analysis of the subject line terms of an engineering project e-mail corpus comprising of 10,628 e-mails, featuring 1,045 individuals and spanning over a 4 year period. More specifically, the analysis has focused on the evolution, use/re-use and centrality of terms across the various project stages. The results provide interesting insights in the evolution of engineering terminology, which leads onto a discussion on how these metrics may provide indicators of project 'normality'.

Keywords: Communication, Product-service systems (PSS), Product Development, Co-Word Analysis, Computer-Mediated Communication

Contact:
Dr. James Anthony Gopsill
University of Bristol
Department of Mechanical Engineering
United Kingdom
J.A.Gopsill@bristol.ac.uk

Please cite this paper as:
Surnames, Initials: *Title of paper*. In: Proceedings of the 20th International Conference on Engineering Design (ICED15), Vol. nn: Title of Volume, Milan, Italy, 27.-30.07.2015

1 INTRODUCTION

It has been well established that communication features heavily in the majority of the activities that engineers perform (Sim and Duffy, 2003). Tenopir and King's (2004) review of patterns in engineers' communication behaviour shows that there is a consensus among researchers that engineers spend up to 58% of their time conversing with one another, be it either through conversations, meetings, informal discussions, phone calls or e-mails. This is further supported by Hertzum and Pejtersen (2000) & Nagle (1998) who estimate that the time spent communicating ranges from 40% to 60% and may reach levels of 75% for some engineers.

For these reasons, engineering communication can be considered the main form by which information and knowledge is shared between engineers to, for example, solve problems associated with the engineering project (Perry and Sanderson, 1998). Wood and DeLoach (2001) reveal that engineers use communication as a primary means to seek for information and generate a shared understanding. This is partly due to the fact that colleagues are seen as easily accessible and trustworthy sources of information, and as a consequence, they are still preferred over computer-generated search results (Allard et al., 2009). A high proportion - 69% as recorded by Handel and Herbsleb (2002) - of communication can be colloquially referred to as 'water-cooler conversations', as it is often a quick and informal exchange between engineers (Poile et al., 2009). Brown and Duguid (2000) highlight that this communication is heavily relied upon to 'fill in the gaps' left by formal documentation and process manuals as they can never fully account for every eventuality. Further, Clarkson and Eckert (2005, p.20) show that engineers use these informal channels in order to be kept informed as well as being able to maintain awareness of project progression.

While a significant body of literature relating to the role of communication within engineering exists, Tenopir and King (2004) highlight that much of this has relied upon surveys and/or interviews as a means of data capture and that there are potential limitations on the understanding that can be generated from them. Table 1 offers further confirmation of this by detailing the various types of data capture methods used in a range of studies of engineering communication. Clarkson and Eckert (2005) go further to suggest that the field is reaching a plateau of understanding and new research methods need to be trialled in order to further the field.

Research Data Capture	Research
Structured Interviews	Curtis et al. (1988)
Survey	Kraut and Streeter (1995)
Observational	Guinan (1986)
Video Recordings	Walz (1988), Olson et al. (1992)
Audio Recordings	Minneman (1991)
Observation & Interviews	Sonnenwald (1996)

Table 1. Examples of empirical research relating to engineering communication (adapted from: Sonnenwald (1996)

Given the prevalence of computer-mediated communication and the development of algorithms to analyse the associated meta-data, there now exists an opportunity to provide quantitative metrics that can characterise the meta-data and content of engineering communication (Gopsill et al., 2013, Wasiak, 2010). Commonly used analytical techniques are Natural Language Processing (NLP) of the subject line alongside Social Network Analysis (SNA).

In order to examine this opportunity, this paper presents results from the co-word analysis of the subject lines from a large engineering project e-mail corpus. The corpus comprises 10,628 e-mails from a large distributed power systems project involving 1,000 engineers working from the specification through to the testing stage of the engineering project. The co-word analysis has been applied to specifically investigate the evolution of terminology used within the engineering project.

This paper first discusses areas of related work from studies in other research fields that have employed NLP/SNA techniques on similar datasets. This is then followed by a description of the dataset analysed in this paper as well as a description of how the co-word analysis technique has been applied. The results are then presented and the key features and patterns in relation to the stages of the engineering project are described. Finally, the paper concludes by discussing the potential meaning and utility in tracking these features automatically when executing engineering projects, and also considers the next steps for the research.

2 RELATED WORK

Applications of NLP and SNA techniques on e-mail communication have primarily been used on the publicly available Enron dataset. The Enron dataset contains 619,446 e-mails sent by 158 users and was made available by the Federal Energy Regulatory Commission during its investigation of the company in 2004 (Klimt and Yang, 2004).

More specifically and in relation to the evolution of topics and terminology, Mcculloh et al. (2002) have demonstrated how meaningful changes in the topics of discussion by the various social groups within the Enron dataset can be detected using Latent Semantic Analysis (LSA) alongside statistical process control. Whilst Dredze et al. (2008) have used LSA to provide summaries of the topics being discussed at particular points in time.

It is generally accepted that the content of e-mail (or any computer-mediated communication tool) can contain multiple layers of information such as social standing, emotion and multiple topics of discussion. Therefore, recent work has looked to develop techniques that can isolate these layers from one another and thus enable investigation of their evolution and co-evolution over time (Scholand et al., 2010).

In addition, co-word analysis has been a growing area of research. This is where a network is generated from the co-occurrence of words within a common string. Many applications of the technique have used keywords from conference papers in order to uncover themes and key terms within a single conference as well as the evolution of themes and terms across a number of conferences (Coulter et al., 1998, Liu et al., 2014, Ding et al., 2001). Figure 1 illustrates the results of one such study upon the Computer Human Interaction (CHI) conference. This shows the connections between key research fields within CHI as well as highlighting the key terms that bridge areas of work, and research areas that are isolated from the core network.

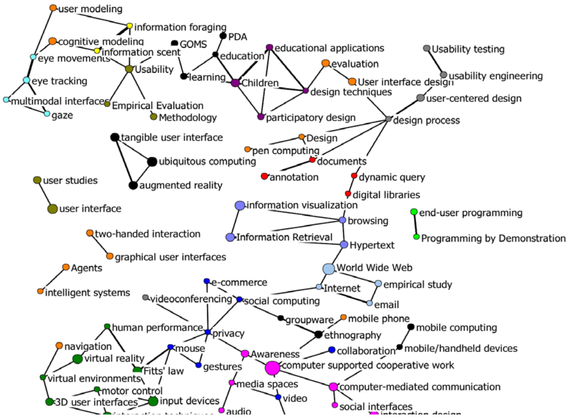


Figure 1. Co-word network of the Computer Human Interaction (CHI) keywords (From: Liu et al., 2014)

Finally, there has also been some application of NLP and SNA techniques to e-mail within the engineering sector. Borgatti and Li (2009) have used metrics generated from SNA to determine the strength of supply chain relationships, which can then be used to aid the real-time management of the product supply chain. SNA has also been applied to product development information flows that have been derived from interview data (Braha and Bar-Yam, 2004). This has been used to identify critical areas of the development process within the company. Finally, Wasiak (2010) has shown through the manual coding of the types of e-mail sent during the stages of an engineering project that correlations exist between the types of e-mail and the stages of the engineering project, as well as being indicative of different modes of working (Figure 2).

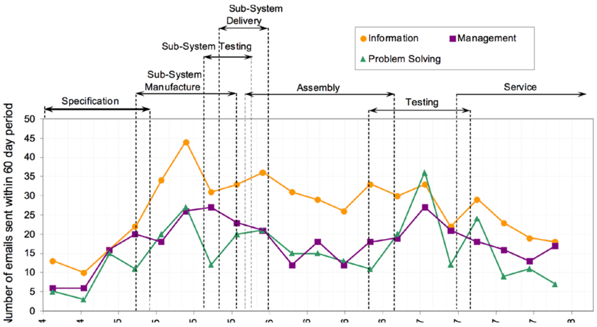


Figure 2. No. of information, management and problem solving e-mails during and engineering project (From: Wasiak (2010))

Given the relative success and maturity of these techniques, this paper seeks to apply the co-word analysis technique to investigate the evolution of engineering terminology through the co-occurrence of terms within the e-mail subject line. The following section provides details of the context of the engineering project studied alongside the description and overview of the e-mail dataset.

3 STUDY CONTEXT

The engineering e-mail dataset has been generated from a multi-nation, multi-million pound engineering project to produce a power and control system for customers in the marine, defence and aerospace industries. The majority of the system design is transferable across the sectors although there were some variations given the specific requirements of each client.

The project duration was four years and 435 employees were involved. The project was of a distributed nature where the employees were primarily based in one of four locations across the globe (UK, France, South Korea & Japan). Thus, it is self-evident that e-mail became a primary means by which the employees communicated. In addition to the communication between employees, the e-mail corpus also includes the communications the employees had with 610 stakeholders of the project, giving a total of 1,045 individuals.

Dataset Statistics	Value
Number of e-mails	10,628
Number of e-mail addresses	1,045
Number of unique e-mail domains	173

Table 2. General dataset statistics

The e-mail client used by the company enabled the assigning of e-mails to particular engineering projects and this was utilised as the data capture method. This has led to an archive set of e-mails for the project. As this was part of company practice, there was no additional burden placed upon the engineers by the study. Therefore, it can be considered that the dataset is a fair reflection of the communication throughout the entire project. In total, 10,628 e-mails were captured over the course of the project and covered a period of four years. These were sent between 1,045 e-mail addresses (Table 2).

To manage the project, the company employed a stage-gate model, which consisted of five key high-level stages; Specification, Manufacture, Sub-system Testing, Assembly and Testing. Table 3 provides the time-frame for each of the stages and this presents the secondary data to which the analysis of the e-mail dataset is to be aligned to

With this understanding of the context and description of the dataset, this paper continues by discussing the analysis that has been performed.

Project Stage	Time-Frame (Months)
Specification	0 - 5
Manufacture	5 -12
Sub-system Testing	12 - 6
Assembly	16 - 25
Testing	25 – 44

Table 3. The project stages and their associated time frame

4 CO-WORD ANALYSIS OF THE SUBJECT LINE

As mentioned previously, co-word analysis has been used in both the Computer Human Interaction (CHI), Information Retrieval and Software Engineering research fields to understand the creation, development and evolution of research terms (Coulter et al., 1998, Liu et al., 2014, Ding et al., 2001). Thus, it is argued that the technique is equally suited to the subject line as it often acts as the keywords for an e-mail. For example, 82.1% of the terms used in the subject lines of this e-mail corpus also appeared within the body content of the respective e-mail. For this analysis, all e-mails were considered (i.e. new, reply & forwarded) as they all provide an indication as to whether the terminology that is still of interest to the project.

Because the terms within e-mail subject lines are not specified individually as with keywords, a process of cleaning the subject line strings is first performed. This has been achieved by isolating the individual words within the subject line and then applying the *Porter2.Stemmer*[1], and removing of any terms of a length less than three characters. The cleaning forms a dictionary of terms that are of interest and will be considered in the analysis. By performing this cleaning, a 72.6% reduction in the terms of interest has been generated.

Once each subject line has been cleaned, a co-word network is generated. Each term is considered a node within the network and its frequency is based upon the number of e-mails it has appeared in within the specified time. In the case of this analysis, monthly cumulative and monthly time-sliced networks have been generated. An edge (connection) is formed between nodes if the two nodes appear within the same subject line (i.e. co-located). The number of times this occurs within the specified

[1] https://pypi.python.org/pypi/stemming/1.0

time provides the weighting for the given edge. Figure 3 provides an illustrative example of a co-word network that is generated from three subject lines containing the terms A, B, C & D.

Thus in the given context, the typical network analysis metrics of degree, frequency and strength relate to:

Degree: The number of terms that the term of interest has been seen alongside in the subject lines.

Frequency: The occurrence of the term within the subject lines.

Strength: The number of times that the two terms have appeared in the same subject line.

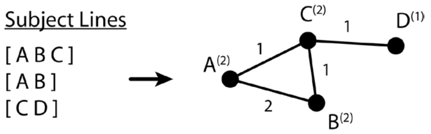


Figure 3. Example co-word network

These measures and their interpretation are discussed in the following results section.

5 RESULTS

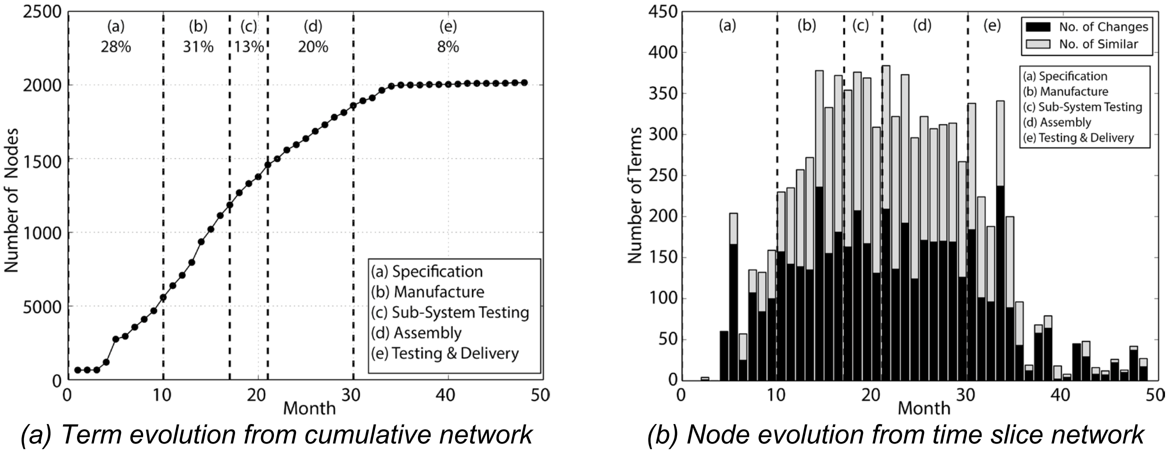
This section presents and discusses the results from the analysis of the subject line co-word network. The analysis has been split three-fold. First, the evolution of the co-word network is considered with respect to the various stages of the project in order to explore whether there exist a correlation between evolution of the network and the stage of the project. Second, the relative use of terms across the stages is analysed alongside the re-use of terms from the previous stages in order to explore persistent terms. Third, the centrality of the terms is considered to explore how - what can be thought as - the most prevalent and/or integrative words change throughout the project.

5.1 Evolution of the Co-Word Network

Figures 4, 5 & 6 reveal how the network of co-words used within the subject line has evolved over time. Figures 4a & 4b show the introduction of new terms over the course of the engineering project. Figure 4a shows the cumulative monthly network and Figure 4b shows the time sliced monthly networks.

Through the inspection of Figure 4a, it can be seen that *Specification* and *Manufacture* stages generate the largest proportion (60%) of the terminology within the project, which is expected as the product is being defined and a solution is being converged upon. Not withstanding this, the rate at which new terms are being added only drops significantly at the final testing stage of the project.

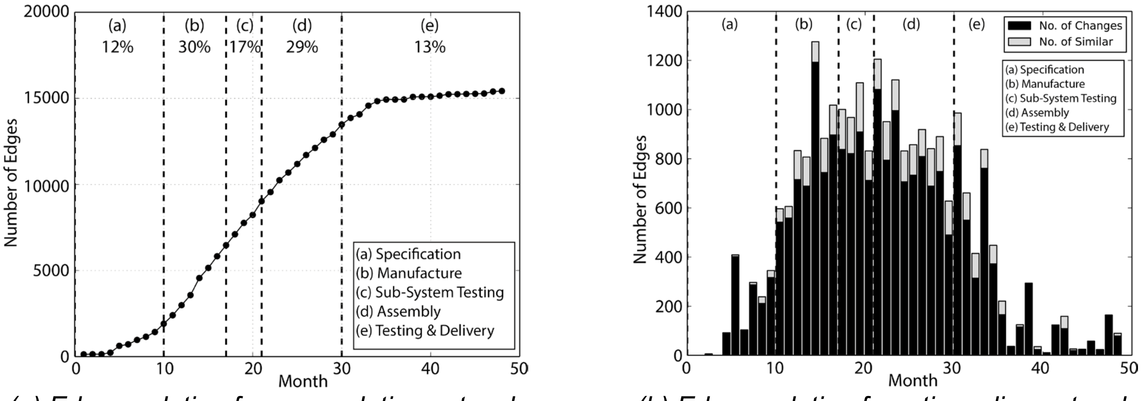
Figure 4b shows the number of new terms being introduced and the number of terms that have been re-used from the previous month. It can be seen that there is a relatively even split between the generation of new terms and the re-use of past terms. This may be an indicator of the rate of progression being made month-on-month. *Specification* and *Testing* stages vary slightly and favour the changing of terms.



(a) Term evolution from cumulative network *(b) Node evolution from time slice network*

Figure 4. The evolution of the terms

Figure 5a shows the rate at which the terms are being connected with one another within the network (cumulative and time slice respectively). It is interesting to note that although the Specification introduces a large proportion of the terms to the network, there are few connections being made between the terms. Rather, Manufacture & Assembly are the main contributing stages to the connectivity of the network. Looking at Figure 5b, it is interesting to note that a large proportion of the connections being formed and re-used are not consistent month-on-month.

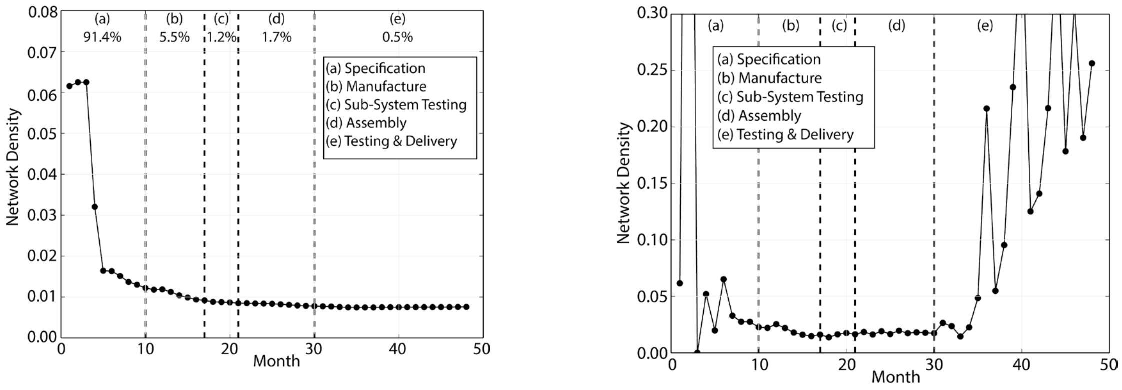


(a) Edge evolution from cumulative network

(b) Edge evolution from time slice network

Figure 5. The evolution of edges

Figures 6a & 6b shows the change in graph density over time and the relative impact each stage has had upon it. It can be clearly seen that the cumulative network converges upon a value for network density and this may be a common network density for an engineering project. The rate of convergence and final value may be the critical aspects to monitor for a project in its early stages. Figure 6b shows the network density for each time slice and it can be seen that a consistent network density is settled upon during the *Manufacture*, *Sub-System Assembly* and *Assembly* stages. The relative low network density may indicate that terms are highly context dependent and are only connected to certain terms and that there is a high-level of concurrent working involving many different activities and different terms.



(a) Network density evolution of cumulative network

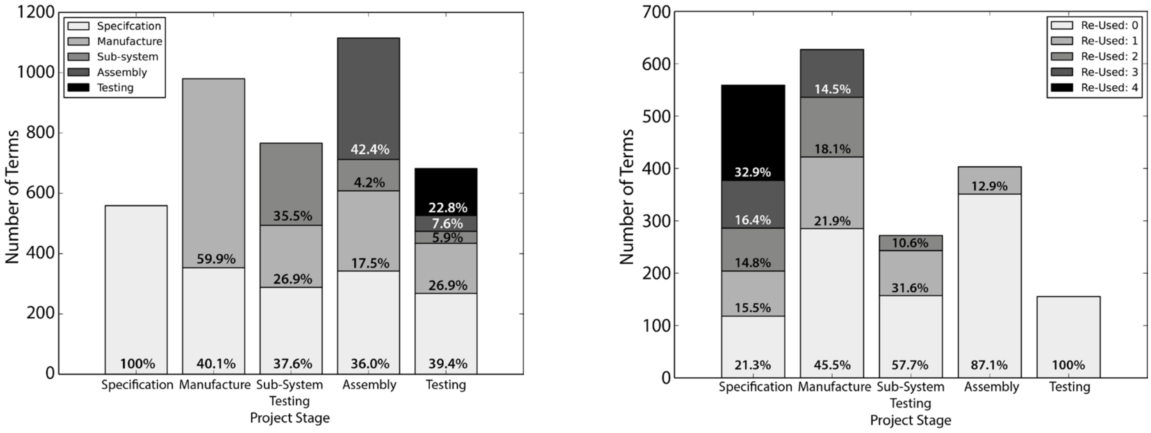
(b) Network density of time slice network

Figure 6. The evolution of the cumulative network graph

It is worth noting that the network density in the *Specification* and *Testing* is significantly higher. This appears logical for the specification phase because there are relatively few initial terms to connect to and thus, it is more likely that they will all become connected. However, *Testing* represents a different phenomena as there already exists a significant set of terms and yet, a highly dense graph is generated. This indicates that small subsets of all the terms are used and are highly connected to one another.

5.2 Use of Terms

In addition to investigating the evolution of the co-word network, Figure 7 shows how the terms, which were generated in one stage of the project, have been used in the following stages of the project. It is self-evident that Figure 7a shows that 100% of the words that appear in the *Specification* stage are used in the *Specification* stage. However, moving to the *Manufacture* stage reveals that 40% of the words that are used in the *Manufacture* stage originated from the *Specification* stage. This further highlights that although *Manufacture* does introduce new terms, there remains a high-level of integration of terms from the previous stage. Continuing along to *Sub-System*, it can be seen that this patterns remains as there is an evenly distributed set of words from each of the previous stages. However, the *Assembly* stage reveals a slightly different pattern where a large number of terms from the *Specification* remain whilst the use of terms from the *Manufacture* & *Sub-System* have reduced. The final *Testing* stage shows a re-appearance of terms from the *Manufacture* stage as well as maintaining the high-level use of terms from the *Specification* stage.



(a) Use of terms from the current and preceding project stages

(b) Re-use of terms within subsequent stages

Figure 7. Use of terms across the various stages of the project

Figure 7b further confirms the observed patterns and highlights the level of re-use of the terms and from which stage they originated in. Almost a third of the terms that appear in the *Specification* phase are subsequently used throughout all of the following stages of product development and almost 60% in at least two of the following stages. The re-use of *Manufacture* terms is less strong although at least 50% are re-used in one other stage of the development process. The level of re-use of terms in *Sub-System* and *Assembly* are much less than compared to *Specification* and *Manufacture*.

Figure 8 shows a scatter plot of the degree and average strength of the terms appearing from the various stages of the project. It quickly becomes apparent that many of the terms are contained within the bottom-left of the graph with neither a considerably high degree nor average strength. However, it can also be seen that two features emerge from the bottom-left corner, a set of words with an increasing degree yet a relatively low average strength (i), and a set of terms with a high average strength and low degree (ii). The high degree terms are the terms that have been connected to many other terms and it is interesting to see that almost all of these terms appear from within the *Specification* phase. In addition, a few appear from *Sub-System* and may suggest that relationships between areas of the product were discovered that were not originally foreseen during the design. The second feature contains the high average strength terms that have low connectivity to the rest of the terms. Many of which appear from either the *Manufacture* or *Sub-System*. This mainly consists of highly process dependent terms that are only used within specific contexts.

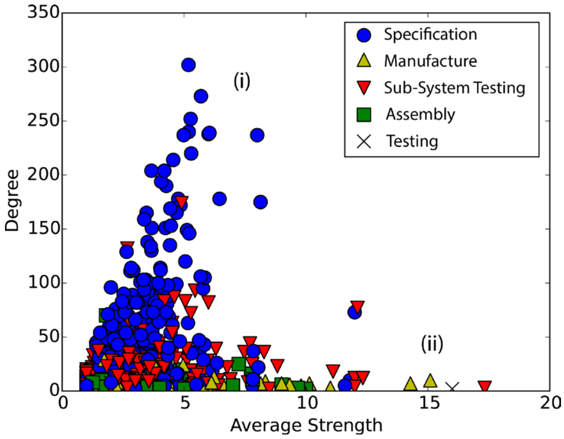
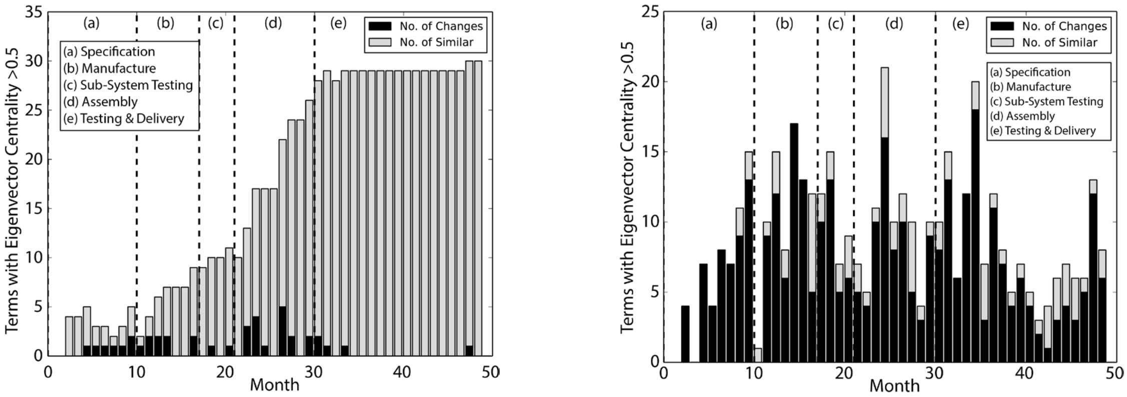


Figure 8. Degree and average strength of the terms from the subject line

5.3 Term Centrality

The final aspect of the network analysis considered in this paper is that of the centrality of the terms within the network. The centrality is often referred considered to indicate the relative importance of a node within a network. For this, eigenvector centrality has been selected and it is often referred to as the measure of influence a node has within the network.

Figures 9a & 9b show the number of highly influential terms (eigenvector centrality of greater than 0.5) within the co-word networks the cumulative and time sliced data. Again, the analysis has compared the networks on a month-on-month basis. The immediate insight that can be drawn from 9a, is the consistency of the similarity of highly influential terms within the network, i.e. once the term becomes influential, it remains influential throughout the rest to the project. Many of the influential terms appear in the *Specification*, early *Manufacture* and *Assembly* stages. In contrast, Figure 9b shows the influential terms on a month-on-month basis and it can be that there is hardly any similarity in the influential terms between months. Also, the number of influential terms varies considerably month-on-month. Considering the two results, it is posited that there exists a core theme of terms that are influential throughout all the months as indicated by Figure 9a, but there is also a widely varying set of terms for a particular month, which are more influential and relate to the specific work being undertaken.



(a) Cumulative term centrality

(b) Time slice term centrality

Figure 9. Term centrality during the engineering project

6 DISCUSSION

The results have revealed a number of features within this e-mail corpus. In particular, the analysis of the evolution of the co-word network revealed that the *Specification* stage introduces the a large proportion of the terms, whilst the *Manufacture* and *Assembly* also add new terms but primarily take on the role of connecting the terms within the network. The level of connectedness and its rate of change may indicate the divergent and convergent nature of the product solution, and it is argued that there may be a potential 'normality' for engineering projects (yet to be investigated). It has also been shown that the terms used in the *Testing* stage vary significantly month-on-month and the logical conclusion is that this is an indication of the various sub-systems being tested (with their own set of terms) with respect to the requirements.

The analysis of the use of terms revealed that there is significant re-use of terms from previous stages of the project. More specifically, a particular sub-set of terms generated from the *Specification* stage is re-used throughout the whole product development process and the stage also provides the most highly connected terms. These could be considered the core product concept terms and the relative re-use/size of this set of terms may indicate the level of definition that the product has. It has also been shown that *Testing* re-uses many terms from all of the previous stages and this appears logical as the *Testing* stage is expected to reveal and remedy potential issues due to the manufacturing and design of the product. It is proposed that by analysing the use/re-use of terms throughout an engineering project, one could draw conclusions on the progress being made, completeness of the design and whether it fits within 'normal' bounds, for example, based on previous project behaviour.

The final analysis undertaken sought to understand the centrality of terms throughout an engineering project and the level of change in the most influential terms within a network. The cumulative co-word network revealed that there appears to be a core theme of influential words throughout the network and they remain core as the project progresses. This may relate to the product definition and specifications for the final product. In addition, it has been shown that influential month-on-month terms vary considerably and this may indicate the level of activity alongside the progression of activity. For example, if the influential terms remained the same month-on-month then this could highlight that there is an issue on a specific activity or a continuation of work, which may not be desirable.

Future work in this field could further explore the behaviour of terms within engineering projects, across multiple engineering communication datasets alongside detailed activity plans of the projects. This would enable a 'compare and contrast' analysis to explore whether there exists typical engineering project 'norms'. It is also unclear as to whether the features exhibited in this analysis appear in other engineering projects.

7 CONCLUSION

Communication features in almost all engineering activities and is the main form by which information and knowledge is shared across the project. Although there is extant research on the subject, much has been of a qualitative nature and with the advent of computer-mediated communication, there now exists the potential to use quantitative analytical techniques to characterise engineering communication, such as Social Network Analysis (SNA) and Natural Language Processing (NLP).

In order to explore this, this paper has applied the co-word network analysis technique on an engineering project e-mail corpus. The results were discussed in relation to the evolution of the network, the use/re-use of terms and the centrality of terms. Each analysis provided potential insights into the behaviour of engineering terms within a project and are summarised in relation to the stages as:

Specification: This stage provides the largest proportion of terms to the co-word network and many of the terms are re-used in one or more of the later stages of the project. Many of the key connecting terms of the final network originate from this stage.

Manufacture: This stage also provides many new terms to the network whilst also providing a large majority of the connections between terms within the network. It also re-uses a large proportion of terms from the specification stage. Finally, this stage sees the creation of the more highly contextually dependent terms.

Sub-System: This stage continues to see a rise in terms and connection of terms but not to the degree of *Specification* or *Manufacture*.

Assembly: This stage continues to see a rise in terms and connection of terms although the rate of network growth begins to diminish.

Testing: This provides the least new terms to the network and the use of terms month-on-month varies considerably. The stage makes significant re-use of terms that have appeared in previous stages of the project.

From the results, it is argued that potential 'norms' in engineering communication can be derived and be potential indicators for engineering project management.

ACKNOWLEDGMENTS

The work reported in this paper has been undertaken as part of the Language of Collaborative Manufacturing Project at the University of Bath & University of Bristol, which has support from the Engineering and Physical Sciences Research Council (EPSRC) (grant reference EP/K014196/1).

REFERENCES

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. Proceedings of the 16th international conference on World Wide Web - WWW '07, page 835, 2007. doi: 10.1145/1242572.1242685.

Suzie Allard, Kenneth J. Levine, and Carol Tenopir. Design engineers and technical professionals at work: Observing information usage in the work- place. Journal of the American Society for Information Science and Technology, 60(3):443–454, 2009. ISSN 1532-2890. doi: 10.1002/asi.21004.

Stephen P Borgatti and Xun Li. On social network analysis in a supply chain context. Journal of Supply Chain Management, 45(2):5–22, 2009. ISSN 1745-493X. doi: 10.1111/j.1745-493X.2009.03166.

Dan. Braha and Yaneer Bar-Yam. Information Flow Structure in Large–Scale Product Development Organisational Networks. Journal of Information Technology, 19:244–253, 2004. ISSN 0268-3962. doi: 10.1057/palgrave.jit.2000030.

J.S. Brown and P. Duguid. Balancing act: Capturing knowledge without killing it. Harvard Business Review. May June, 2000.

Alberto Cambrosio, Camille Limoges, Jean Pierre Courtial, and Françoise Laville. Historical scientometrics? mapping over 70 years of biological safety research with coword analysis. Scientometrics, 27(2):119–143, 1993.

Rich Caruana. Identifying Temporal Patterns and Key Players in Document Collections. In Proceeedings, AMAST 95, 1995.

J. Clarkson and C. Eckert. Design process improvement: a review of current practice. Springer Verlag, 2005.

Neal Coulter, Ira Monarch, and Suresh Konda. Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 49(13):1206–1223, 1998.

Jana Diesner, Terrill L. Frantz, and Kathleen M. Carley. Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". Computational & Mathematical Organisation Theory, 11: 201–228, 2005.

Ying Ding, Gobinda G Chowdhury, and Schubert Foo. Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing & Management, 37(6):817 – 842, 2001. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(00)00051-0.

Mark Dredze, Hanna M Wallach, Danny Puller, and Fernando Pereira. Generating summary keywords for emails using topics. In Proceedings of the 13th international conference on Intelligent user inter- faces, IUI '08, pages 199–206, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-987-6. doi: 10.1145/1378773.1378800.

J.A. Gopsill, H.C. McAlpine, and B. J. Hicks. The communication patterns of engineers within an sme 2012. In International Conference on Engineering Design ICED'13, 2013.

Mark Handel and James D. Herbsleb. What is chat doing in the work-place? In Proceedings of the 2002 ACM conference on Computer supported cooperative work, CSCW '02, pages 1–10, New York, NY, USA, 2002. ACM. ISBN 1-58113-560-2. doi: 10.1145/587078.587080.

Morten Hertzum and Annelise Mark Pejtersen. The information-seeking practices of engineers: searching for documents as well as for people. Information Processing & Management, 36(5):761 – 778, 2000. ISSN 0306-4573. doi: 10.1016/S0306-4573(00)00011-X.

Bryan Klimt and Yiming Yang. Introducing the enron corpus. In CEAS, 2004.

Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. Chi 1994-2013: mapping two decades of intellectual progress through co-word analysis. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems, pages 3553–3562. ACM, 2014.

Andrew Mccallum. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. 30:249–272, 2007.

Ian Mcculloh, Eric Daimler, and Kathleen M Carley. Using latent semantic analysis of email to detect change in social groups. 2002.

J.G. Nagle. Communication in the profession. Today's Engineer, 1(1), 1998.

Mark Perry and Duncan Sanderson. Coordinating joint design work: the role of communication and artefacts. Design Studies, 19(3):273 – 288, 1998. ISSN 0142-694X. doi: 10.1016/S0142-694X(98)00008-8.

C. Poile, A. Begel, N. Nagappan, and L. Layman. Coordination in large-scale software development: Helpful and unhelpful behaviours. 2009. URL research.microsoft.com. .

Ryan Rowe, German Creamer, Shlomo Hershkop, and Salvatore J Stolfo. Automated social hierarchy detection through email network analysis. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA- KDD '07, pages 109–117, 2007.

Andrew J. Scholand, Yla R. Tausczik, and James W. Pennebaker. Social language network analysis. Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10, page 23, 2010. doi: 10.1145/1718918.1718925.

Siang Kok Sim and Alex H. B. Duffy. Towards an ontology of generic engineering design activities. Research in Engineering Design, 14:200–223, 2003. ISSN 0934-9839.

Diane H. Sonnenwald. Communication roles that support collaboration during the design process. Design Studies, 17(3):277 – 301, 1996. ISSN 0142-694X. doi: 10.1016/0142-694X(96)00002-6.

Carol Tenopir and Donald W. King. Communication Patterns of Engineers. Wiley-IEEE Computer Society Pr, 2004. ISBN 047148492X.

James O Wasiak. A Content Based Approach for Investigating the Role and Use of E-Mail in Engineering Design Projects. PhD thesis, Department of Mechanical Engineering, University of Bath, 2010.

Mark Wood and Scott DeLoach. An overview of the multiagent systems engineering methodology. In Paolo Ciancarini and Michael Wooldridge, edi- tors, Agent-Oriented Software Engineering, volume 1957 of Lecture Notes in Computer Science, pages 1–53. 2001. ISBN 978-3-540-41594-7.