Institution: Cardiff University

Unit of Assessment: 10

Title of case study: Meeting the Challenges of Data Security

1. Summary of the impact (indicative maximum 100 words)

The security of data in printing and network environments is an area of increasing concern to individuals, businesses, government organisations and security agencies throughout the world. Mathematical algorithms developed at the School of Mathematics at Cardiff University represent a significant step-change in existing data security techniques. The algorithms enable greater security in automatic document classification and summarisation, information retrieval and image understanding. Hewlett-Packard (HP), the world's leading PC vendor, funded the research underpinning this development and patented the resulting software, with the aim of strengthening its position as the market leader in this sector of the global information technology industry. Hewlett Packard has incorporated the algorithms in a schedule of upgrades to improve the key security features in over ten million of their electronic devices. Accordingly, the impact claimed is mitigating data security risks for HP users and clients and substantial economic gain for the company.

2. Underpinning research (indicative maximum 500 words)

The underpinning research was undertaken at Cardiff University during the period 2007-2011. The motivation for the research arose from the inability of existing algorithms to reliably distinguish between confidential and non-confidential documents. The research concerns the development of novel algorithms that are crucially important in the field of data security. Algorithms have been developed for *rapid change detection in data streams and documents,* and *text summarisation and classification* and used in partnership to perform a two-level analysis to create a secure printing network environment. The rapid change detection algorithm can be considered as a low-level analysis for extracting features or keywords. Then these features or keywords are used to perform a higher level analysis such as text summarisation or classification.

In the process of distinguishing between a confidential and non-confidential document, a key indicator is the precise scientific definition of the meaning of the document. Previous extraction algorithms have not been robust in the sense that different algorithms produce different outputs due to the non-existence of the scientific definition of the meaning of a document.

The novel algorithms developed in (3.1, 3.2) represent the first attempt to define document meaning based on the human perceptual model. Our research is based on ideas from image processing and especially on the Helmholtz Principle from the Gestalt Theory of human perception. When it is applied to the problems of unusual behaviour detection and keywords extraction, it delivers fast and effective tools to identify meaningful keywords using parameter-free methods. A level of meaningfulness of the keywords is also defined which can be used to modify the set of keywords depending on application needs.

According to a basic principle of perception, due to Helmholtz, an observed geometric structure is perceptually meaningful if it has a very low probability of appearing in noise. As a common sense statement, this means that "events that could not happen by chance are immediately perceived". For example, a group of five aligned dots exists in both images in Fig.1, but it can hardly be seen on the left-hand side image. Indeed, such a configuration is not exceptional in view of the total number of dots. Yet, in the right-hand image we immediately perceive the alignment as a large deviation from randomness that would be unlikely to happen by chance.

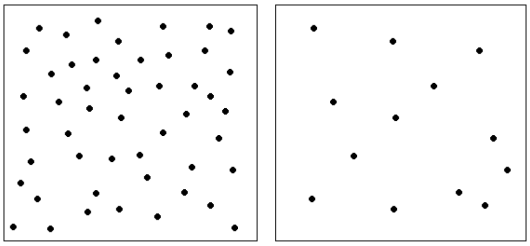

Fig. 1. The Helmholtz principle in human perception

The research makes novel contributions to knowledge extraction technologies. It does so in the mining of unstructured data and detecting unusual behaviour and in the content of streams of short documents and files. In the context of data mining, the research defined the Helmholtz Principle as the statement that meaningful features and interesting events appear as large deviations from randomness (3.1, 3.2). In the cases of textual, sequential or unstructured data qualitative measures were derived for such deviations. Under unstructured data, data can be understood without an explicit data model, but with some internal geometrical structure. For example, sets of dots in Fig. 1 are not created by a precise data model, but still have important geometrical structures: nearest neighbours, alignments, concentrations in some regions, etc. A good example is textual data where there are natural structures like files, topics, paragraphs, documents etc. Sequential and temporal data also can be divided into natural blocks like days, months or blocks of several sequential events.

Over the years, the amount of text available electronically has grown exponentially creating a huge demand for automatic methods and tools for text summarisation. Based on the work on the detection of unusual behaviour in text (3.1, 3.2), it was possible to model a document as a one-parameter family of graphs, with its sentences (or paragraphs) as the set of its nodes and edges defined by a carefully selected family of meaningful words (using the Helmholtz principle form). We demonstrated that, for some range of the parameter, there is a transition in which the resulting graph becomes a small-world network. We exploited this remarkable structure by modelling texts and documents as small-world networks and applying many of the measures and tools from social network theory to develop a novel approach to extractive text summarisation (3.3, 3.4, 3.5). The goal in extractive text summarisation is to extract the most meaningful parts of documents (sentences, paragraphs, etc.) to represent main concepts of the document. The algorithms based on our research extract the most important sentences and structures from text documents reliably and efficiently.

As a consequence of this work HP were provided with:

1. An algorithm to detect changes in data streams, resulting in a US patent for HP (3.6).
2. New algorithms for the Hewlett-Packard Secure Document Ecosystem Portfolio, for automatic keyword extraction and significance evaluation.
3 Algorithms for extractive text summarisation and classification, using a small-world network model – two US patents have been filed by HP.
4. New algorithms for the Hewlett-Packard Secure Document Ecosystem Portfolio, which automatically summarise texts and extract their most important features.

Key staff: Prof. A. Balinsky (academic staff 2001-) assisted by two PhD students (N. Mohammad (2007-2011, EPSRC CASE award with Hewlett-Packard) and B. Dadachev (2011- , funded jointly by Cardiff University and Hewlett-Packard)). Dr Mohammad was immediately employed by HP on completion of his PhD.

3. References to the research (indicative maximum of six references)

3.1 A. Balinsky, H. Balinsky and S. J. Simske, "*On Helmholtz's principle for Document Processing*", 10 ACM Symposium on Document Engineering (DocEng2010), Manchester, UK, 21-

24 September 2010.
http://doi.acm.org/10.1145/1860559.1860624 Copy held by HEI, available on request.
3.2 A. Balinsky, H. Balinsky and S. J. Simske, “On the Helmholtz Principle for Data Mining”, HP Technical Report, HPL-2010-133, http://www.hpl.hp.com/techreports/2010/HPL-2010-133.html Copy held by HEI, available on request.
3.3 A. Balinsky, H. Balinsky and S. J. Simske, Automatic Text Summarization and Small-World Networks, ACM DocEng2011, Google, Mountain View, California, 19-22 September 2011.
http://doi.acm.org/10.1145/2034691.2034731 Copy held by HEI, available on request.
3.4 H. Balinsky, **A. Balinsky**, and S.Simske, Document Sentences as a Small World, IEEE SMC 2011, October 9-12, 2011. doi: 10.1109/ICSMC.2011.6084065 Copy held by HEI, available on request.
3.5 A. Balinsky, H. Balinsky and S. J. Simske, “*Rapid Change Detection and Text Mining*”, at the 2nd IMA Conference on Mathematics in Defence, Defence Academy, Shrivenham, 20 October2011.
http://www.ima.org.uk/conferences/past_conferences/2011/maths_in_defence/conference_papers.cfm Copy held by HEI, available on request.
3.6 A. Balinsky, H. Balinsky and S. J. Simske, Keyword Determination based on a weight of meaningfulness, U.S. Patent 8,375,022, 12 February, 2013. Copy held by HEI, available on request.

4. Details of the impact (indicative maximum 750 words)

Cardiff University’s contact with HP originated in 2007. Following the research, the algorithms were subjected to extensive proof-of-concept testing, in the Production Division in HP, where they were shown to significantly improve the security of printing and network environments in their prototype and pre-production printers and services. As a result of this internal evaluation, the task of upgrading the key security features in over 10 million electronic devices was initiated by HP in March 2013 (5.1). The algorithms are implemented in either the firmware or software of devices, depending on their computing power.

Unusual behaviour detection and information extraction in streams of short documents and files (emails, news, tweets, log files, messages, etc.) are important problems in security applications and failure to adequately protect printing and network environments has the potential to adversely affect millions of users. The applications – which range from automatic document classification to information extraction and information visualisation, from automatic unusual behaviour detection to security policy enforcement – all rely on automatically extracted features (in data streams) or keywords (in documents) to perform a higher level of analysis. Of paramount importance in these applications is the quality (accuracy) and speed (efficiency) of keyword extraction algorithms. The algorithms developed by Cardiff have been shown by HP to satisfy these criteria.

Print Security:
The threat to printing and imaging devices and data has increased over the last decade as a consequence of more sophisticated threats, increasingly mobile workforces and changes in industry regulations. There are a variety of means that data can be compromised in this fashion. These include hardware theft which could expose documents sent to stolen printers and multi-function printers for later printing, or unauthorized changes to unprotected settings that will enable someone to reroute print jobs and potentially access network and password information. Moreover so-called network sniffers can obtain data that is transmitted between a PC and a printer, revealing the print job. Similarly, unsecured cloud connectivity could give unauthorized users access to the data at any time, in any place. Data that is compromised can result in the loss of millions of pounds due to employee and customer identity theft, private and corporate lawsuits, industry violations or government fines. Subsequently, means to combat unauthorised and illegal practises, whist enabling innocent transactions or normal usage, are essential. The new approach to rapid change detection in documents and text summarization and classification developed at Cardiff, and implemented in electronic devices by HP, identifies documents that are confidential and prevents them from being printed by unauthorised users (5.1).

Mitigating Risks to Data Security:

Cardiff University's research has significantly improved the security of printing and network environments. The sophisticated and efficient algorithms for data mining, which were developed for HP, recognise normal and abnormal patterns of data. Developments such as the new approach to rapid change detection in data streams and log files, applied to the problem of feature extraction, provide extremely fast and effective techniques for the identification of meaningful features by parameter-free methods. Likewise, the approach to an extractive summarization, by modelling data as small-world networks, can be applied to the problem of extracting the most important structures from data. The algorithms are essentially valid safeguards for all data transmitted via printing and network applications. The feature extractor developed by Cardiff University has undergone extensive internal evaluation by HP and found to be vastly superior to existing techniques (5.1, 5.2). In the evaluation HP found that the feature extractor developed by Cardiff considerably outperformed other feature extractors in current use. In quantitative terms, the confidentiality accuracy was increased from 60% to 83%, thereby reducing the error rate by more than 50% (5.1). The ability of the extractor to detect unusual behaviour means that dangerous or unauthorised behaviour can be prevented and therefore it provides enormous security benefits for HP's extensive client base - this includes customers in nearly every country in the world. HP services corporations such as Barclays Bank, Fords and a plethora of multinational organisations in the healthcare, life sciences and pharmaceutical industries. Data privacy is paramount to these businesses; the research has enabled information sharing in a markedly more secure IT environment.

Economic Gain:

HP has made it clear that it cannot divulge quantitative information concerning economic gain from this research for reasons of commercial sensitivity (in the context of sales) and state security (in the context of consultancy).

Sales

The research has enabled HP to retain its position as the market leader in the information technology industry, a fact officially recognised since 2007. In 2012 HP had the biggest share of the global market, 16%. The company states, as part of its corporate aims, “We lead in the marketplace by developing and delivering useful and innovative products, services and solutions.” Instructively, the algorithms and resulting features implemented in HP products are novel developments that outperform existing attempts by competitors to address data security risks. They enable a dynamic, as opposed to a static, response to protecting data. Dr. Steven Simske, Director and Chief Technologist for Security Printing and Imaging Engineering, commented that “the algorithms developed by Cardiff University are novel to data mining and are extremely valuable to our organisation. They enable us to successfully compete within the industry and drive technology forward to meet the evolving needs of our clients. Without the research produced by Alex Balinsky our achievements in this area would not have been possible.” (5.1).

Consultancy

The algorithms have also been used to progress HP's security policy in their security consultancy practice. This guarantees impact for their high end clients, including security services and law enforcement agencies across the globe. Dr. Simske continues to state that “the research has been integral to the development of our Big Data, Analytics and Security themes. It is both unprecedented and highly creative work that is fuelling the development of HP's entire security framework.” (5.1)

5. Sources to corroborate the impact (indicative maximum of 10 references)

5.1 HP Fellow, Director and Chief Technologist for Security Printing and Imaging Engineering, Hewlett-Packard Laboratories. *Corroborates the use of the algorithms by HP and the resulting impact.*

5.2 T. Bohne, S. Rönnau, U. M. Borghoff, “*Efficient keyword extraction for meaningful document perception*”, ACM DocEng2011, Google, Mountain View, California, 19-22 September 2011. *Corroborates that the algorithms are recognised as superior to existing techniques.*