ON SUPPLY CHAINS, DEPERIMETERIZATION AND THE IPCRESS SOLUTION

Lee Gillam[1], Simon Broome[2] and Debbie Garside[3]

[1]Department of Computing, University of Surrey, Guildford, UK
l.gillam@surrey.ac.uk
[2]Jaguar Land Rover, Northampton, UK
sbroome2@jaguarlandrover.com
[3]GeoLang Ltd, Cardiff, UK
debbiegarside@geolang.com

ABSTRACT

This paper offers an overview of the dual challenges involved with protecting Intellectual Property in the distribution of valuable business information that needs to be read and acted upon by others. We introduce IPCRESS as a means to engender trust in an inherently deperimeterized supply chain likely acting through the Cloud, discuss the context and requirements of such a system, and in relation to a potentially familiar application domain of plagiarism detection show how that an approach which does not content can be used effectively to find similar content (precision: 0.88) whilst having some robustness to obfuscation.

KEYWORDS

Information Leakage, IPCRESS, Cloud, Intellectual Property

1. INTRODUCTION

Coherently managed document archives of any organization could contain a variety of high-value information around business transactions, research, technical development, market analysis and strategy. This can exist in and across siloed business units, with disparate systems, approaches, and evolved practices. Careful curation can be costly and enforced change poorly received. Custodians may impose constraints that actively impede progress and lead to unconstrained workarounds, encouraging new problems. But free-flowing communication, often across legislative borders, presents a risk to such archives. Such free-flowing communication is essential when organizations are not self-sufficient – i.e. when they act in supply chains – and communication and especially accidental communication can readily flow bi-directionally. Recipient recommendation, without checking, is but one way in which accidents happen [1]. Of interest for this paper is that deperimeterization, the lack of a readily definable organizational boundary, is inherent in Supply Chains, and especially inherent in Supply Chains acting through the Cloud. It is the deperimeterization that presents risks, not least to Intellectual Property. Costs of such risks have been reported at £9.2bn annually for the UK (OCSIA/Detica report) with the notion that insider assistance is typical, and as high as $300bn annually as mooted for the US (National Bureau of Asian Research study) which identified issues with protection when dealing with specific nations. An important question for a Supply Chain is: who is an insider? This should be easily answered until one considers that companies who tender unsuccessfully for work may have been privy to information about the work to be done and may have learnt from it in order to offer to supply similar to others. This offers potential for less scrupulous suppliers to win business with others and deliberately or inadvertently to end up supplying a very similar component. Aside from self-destructing documents, checks that such suppliers have destroyed

correspondence when unsuccessful are difficult to make, and if detecting such leakage is quite a challenge; acting on such a leakage is even more so.

In a collaborative research and development project between Jaguar Land Rover, University of Surrey, and GeoLang Ltd, and with funding from the UK government-backed Technology Strategy Board for 18 months, we are constructing the Intellectual Property Protecting Cloud Services in Supply Chains (IPCRESS) system to address Supply Chains and barriers to Cloud adoption related to data security and resilience. The focus for IPCRESS is this difficulty of entrusting valuable Intellectual Property (IP) to third parties, through the Cloud, as is necessary to allow for the construction of components in the supply chain: such information needs to be readily readable and usable by suppliers so that they can see what to build and understand vital properties of the things being built, and so encryption-based approaches can become, at best, inconvenient; at worst, they will *encourage* removal of protection through transformations of the received information into forms that avoid such inconvenience. IPCRESS is developing the capability for tracking IP through supply chains, built around Surrey's private search approach to plagiarism detection which is suited to tracking IP without revealing IP (US patent filed November 2011; PCT filed November 2012). Such tracking is suited to the tasks of (i) preventing IP leakage; (ii) detecting IP leakage or theft; and (iii) identifying retention beyond allowed review periods.

Discussions around such a system and its uses have been presented previously [2]. In this paper, we offer an overview of the IPCRESS context (Section 2), expected formulation of the system (Section 3), and information regarding the operation of the approach (Section 4) and its evaluation (Section 5).We conclude by speculating on the potential business value of adopting such a system.

2. THE IPCRESS CONTEXT

Protection of IP rights (IPR) is critical to the presence and growth of business activities. Nations in Western Europe, the US, Japan and Singapore have well-established regulations with stringent and enforceable rules for IP protection. Many emerging economies recognize that this is vital to attract and maintain foreign investment, so Malaysia, Hungary, India, and China also begin to craft legislation to improve IP protection standards. But tracking IP across companies in these nations where each has different considerations to make with respect to IP remains a significant problem. Particularly for IP-heavy industries, an innovative IP tracking system to monitor data across multiple silos without incurring high costs of manual curation or risking loss of IP in the process is appealing. Consider, for example, the automotive industry. About 65-70% of products are typically designed and engineered in the supply chain[1]. As an exemplar of this industry, Jaguar Land Rover has several UK sites and several off-shore sites to actively take production to emerging markets, and although focus on IP generation is in the UK, it must be trusted first amongst these sites and from there to suppliers for these sites. Locations of suppliers, then, become important, as do the embedded information security approaches. Technological 'solutions' typically take the view of securing organisational borders, which don't really exist here, then the network - reinforcing a 'perimeterised' perspective. Moreover it would assume that it is possible to control all technologies, and indeed people, at all points of the supply chain. All of this can be somewhat at odds with business needs: business R&D strategies demand *more* connectivity outside the enterprise. However, mistrust and perimeterisation will become the norm across industries that suffer most from IP theft. Our aim, then, is to offer a means of *safer* deperimeterisation whilst fostering recognition of IP value

[1] Henry, I. (2011), 'The UK Automotive Industry: Invest now', AutoAnalysis: http://www.nudgeadvisory.com/assets/uploaded/downloads/The_UK_Automotive_Industry_Invest_Now.pdf

across the supply chain. Such an environment could facilitate new and lucrative manufacturing and development partnerships and consequently enable confident utilisation of cloud technologies.

A system such as IPCRESS should help to encourage similar levels of respect for IP even where local legal systems for IP are poor, or cultural importance is lacking. Clearly elaborated operational protocols are required which, allied to software-based tracking methods, act to deter opportunistic industrial theft and become off-putting to an 'insider'. IPCRESS will develop a capability for tracking IP through supply chains, offering Cloud services to (i) prevent IP leakage; (ii) detect IP leakage, or theft; and (iii) identify information retention beyond allowed periods. The approach to be trialled within Jaguar Land Rover is based on a computationally efficient method for finding IP without exposing IP, referred to as private search, but with an additional novelty (US patent filed by the University of Surrey) of avoiding costs of encryption.

To act effectively, the system would need to account for the following requirements:

1. **scale to the entire (potentially deep) web:** information as leaked straight out 'to public' would need to be traced. This requires a system that can process such information efficiently.
2. **be used against (private) corporate resources:** information within a corporate must first be indexed. Care must, of course, be taken with such an index since it also, in theory at least, contains a trace of the valuable information and, indeed, likely of rather more valuable information, in pulling together all such information of value, than would ever be shared in any individual interaction.
3. **be used across (private) corporates:** the key challenge, since it requires corporates to be willing to share their indexes, which, from 2 above, may include trace of *all* their valuable information, whilst still unwilling to share actual content. This places a specific on the index not being a route to revealing the content – this is the key novelty of the approach being embedded into IPCRESS, and which helps to satisfy 1 and 2 here.
4. **be built in and for 'the Cloud':** if we have achieved, in particular, 3, the entire system and all of the produced corporate indexes could be deployed in 'the Cloud' to allow for scalable processing.

The system needs to be responsive – ideally, operating for queries at the speed of search (1) – and should be usable directly within an organization (2) for 'private search' – for us, the ability to find information without exposing the queries being used, and hence not leaking information through any other exposure (3).

The approach shares some similarities with a Federated Search capability across different instances of the same Enterprise Content Management (ECM) system. In a Federated Search approach, the same query – likely a few keywords - is presented to each individual instance of the system, and results are presented in a single page but partitioned according to instances searched. A user at any of the locations can search for content in any of the instances exposed this way. Each instance maintains its own search index across the underlying server farm that hosts the documents, generally to support querying, and the Federated Search consolidates efforts that would be needed to separately search each instance. However, this is a product-specific approach – most likely locked-in to a single vendor - for a distributed large enterprise, where full access is likely to be granted to content across the Federation and where the cost of maintenance of multiple geographically distributed instances is easier to justify than attempting to produce a single monolithic system with concomitant difficulties associated to latency and bandwidth. Relevance ranking is provided with respect to the query issued for each separate instance rather than a consolidated ranking being offered. In IPCRESS, we consider the Federation as being *across* enterprises, and the query comprises the patterns generated from

entire documents. With this external-first view, the approach should also be readily suited to search within an organisation. And so, for a given document, all documents containing matching segments from both inside and outside the organisation should be identifiable. Companies can then manage and protect such materials internally, using this as a means to bootstrap such provision and assist in confidentiality marking of documents.

3. ADDRESSING THE IPCRESS REQUIREMENTS

We consider the implications around the first 3 of the 4 requirements of the IPCRESS system, discussed briefly in the previous section, and the relationship this has to common systems for plagiarism detection, in the remainder of this section.

3.1. Private search in Public (Req.1)

Public content, here web texts, need to be composable into an index produced in a manner consistent with the approach for internal resources. Matches are made against patterns, with ranking by largest extent of match. This can offer a similar capability to a search engine, but one in which complete (but private) documents are the query, rather than a few clear text keywords. The ***actual content of the private documents never leaves the organisation*** in this process, having only been involved in pattern production; moreover, **the matching system need retain no trace of patterns matched against**. Consider, for example, a user with a document open in a common Word Processing application on a laptop. They have available to them a menu bar offering one initial button – 'Private Search'. On pressing this button, the pattern production process takes place on this laptop (within the Word Processing application). The patterns produced, and only the patterns, are exchanged with the server. The server finds all instances of these patterns in its index and collects associated document identifiers. Results are ranked by frequency of occurrence of the document identifiers. The Word Processing application retrieves the list of matches (alarms) which can be explored adjacent to the existing document. Documents of interest to the user would then, and only then, be retrieved, with matching segments in documents aligned for inspection post-retrieval. A simple alarm, for one document, could carry the following initial information:

1. an identifier for the source document;
2. web address (URL) of matching document – to allow retrieval;
3. extent of match by proportion;
4. title, description, and other useful metadata of the *matching* document

For inspection, following retrieval, the following are also needed:

5. list of fragments of document involved in match – to able to view sections involved with the match without yet needing to retrieve the matching document - including, for each match fragment: (i) start and end location of fragment in source document; (ii) start and end location of fragment in matching document

3.2. Private search internally (Req.2)

Full archive match would operate similarly, but in relation to full indexes –potentially one per business unit - already produced, likely as background processing. The approach is inherently similar, but without the involvement of Word Processor software - more likely, with index generation operating in close technical proximity to the Enterprise Content Management system. Subsequent investigation of matches of significant concern – document segments in business units that might not be expected, for example the very latest and most technical innovations being near to 'press releases'. Internal private search helps to demonstrate external operation, and enables tracking of the IP through the organisation. This may also imply indexing and matching within the email system and any other collaborative platform.

3.3. Private search across Privates (Req.3)

As above, matches are made against indexes, with ranking by largest extent of match. Again, the capability is similar to a search engine, with full (but private) documents as the query. Here, the actual content of the private documents being matched never leaves any of the organisations and again the matching system need retain no trace of patterns matched against.

Most importantly, when a match in private content is detected an alarm is generated multilaterally to inform all parties to which it is relevant of a potential concern. But the matching content is not revealed at this stage. The information identified as being carried with an alarm varies as follows:

1. instead of URL, a supply chain member identifier and document id is provided – to assist in investigations;
2. title, description, and other useful metadata of the matching document is **not** made available to either party – indeed, each organisation's metadata may need authorised access before they can even see which files are implicated on their own side.

Having received such an alarm, investigation is now required by all implicated parties. This necessitates the description of a protocol for investigation. Such a protocol could involve, for example, a mediation process, or the exposure of smaller fragments implicated in the match that still do not reveal the critical content – for example, by redaction and selective revealing.

Though a potentially very effective technical approach, it cannot be adopted readily without cross-organisational agreement and buy-in, and so the protocol for investigation will become a key dependency as the project progresses.

4. OPERATION OF THE IPCRESS APPROACH

The need to search for sections of documents as are re-used would immediately suggest the use of existing approaches for copy detection (which some may equate narrowly to plagiarism detection systems). These perform reasonably, reliant on the extent of coverage of their indexes, across documents of which ***all the content is readily readable*** – i.e. when the queries and texts can be exposed in entirety. For efficiency reasons, the index is likely formed of n-grams, hashes, or encrypted data – n-grams mean the documents could be reconstituted if document id and n-gram position are known; hashing and encryption both add processing costs, but consistency requires the hash or encrypted value to be reasonably unique – and particularly for hashing, data similarity does not mean hash similarity: a one character difference will change the hash value quite significantly. Relative uniqueness, and access to keys, as well as access to the same hashing approach, means that such techniques, whilst offering potential look-up efficiency in an index, are unlikely to be worth the processing cost. Indeed, typical techniques for plagiarism detection will readily fail the first three of our requirements.

Addressing the requirements means that it must not be readily possible to reverse-engineer the document, or to be able to achieve this by knowing the approach and the simple expedient of brute force. Such a system will work well if it is possible to generate many possible inputs for a single pattern (an 'ambiguous hash' has been suggested as a means to refer to this), and if hash proximity/similarity and data similarity have a closer relationship. In addition, reasonable detection performance must be assured at speed, and so linguistic processing (part of speech tagging) approaches are also mitigated against. The patented approach does, we believe, meet such requirements. For a set of documents, we are able to convert the plain text of each document to a set of (statistically almost irreversible) patterns, and insert these patterns into an index (pattern as key, value as document and pattern start position. Such an index can readily be sharded by key to allow for scaling. Consider a simplified example source index where a key is

assigned to a specific text pattern (it would be repeated if the pattern is found in other documents). The index would be of the form [key, document, position]:

001, 1, 75 → 002, 1, 84 → 003, 1, 99 → 004, 2, 2574 → 005, 2, 2599

The approach to detection is similar. For a document of interest [*d*], another index is produced, although for a single document we need not be concered about its id. Matches in the indexes are generated from the same key production process. Each detection returns only the pair of document id and position, so sorting on document and then position, and subsequently identifying the document with the most detections and addressing the nature of the overlapping positions, helps to compose the results.

So, if *d* returned all the examples above, documents 1 and 2 are of interest with matching segments spanning 75, 84, 99 and 2574, 2599. Segment sizes depend on the length of encoded data, so if we assumed a 14-gram (14 word segment), for document 1 we have a continuous segment from words 75-113. For document 2, however, there is a gap between the 2 segments. Here we can use a notion of a *stitch distance*, allied to a confidence value relative to the length of the stitch, such that we retain just one continuous segment from 2574-2613, but with a slightly lowered match confidence [here, 10 words are missing, so we could say 30 plus 10 x 4/14 to account for the stitchable distance] of (about) 33 vs 39. So, documents would remain ordered 1, 2. The combination of an ambiguous pattern production approach, and a confidence weighting for missing detections, helps to overcome some obfuscation.

One key generation approach for plagiarism detection is to use MD5hash. Suppose that we break each document into 5-grams that overlap by 2 words, and produce a hash value for each:

- the quick brown fox jumped → e0c19dedd2e35a44b70ca531144ac953
- fox jumped over the lazy → 842ff3fabd7032a95c5cd5cc919a7e6b
- the lazy dog and cat → 2b4032a8f7fa15aa933dd916e93cf8d2

These positions would be '1', '4', '7'. Key length is '5', so if the second pattern went undetected (somebody changes 'over' to 'across' in their document, which results in an entirely different hash value), a stitch distance of 1 would allow for a continuous segment to be reported albeit with a slightly lower confidence. It should be apparent how brittle such a hash-based approach is. Somebody wishing to avoid detection would need change only 2 words (e.g. 'jumped', 'lazy' to 'jumps', 'tired') and none of the resulting hashes would be matchable. To see the effect of this, consider switching dog → dogs in the third 5-gram. This results in a hash of f15f022792db93722733b4b5b2b6f548. Our approach does not suffer this (see Figure 1).

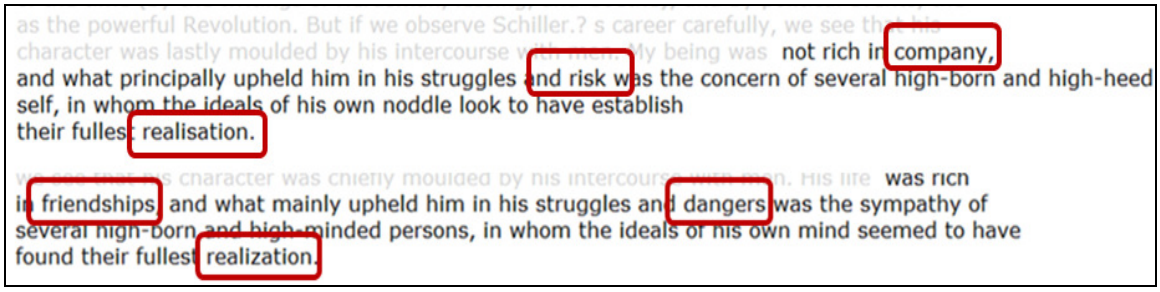

Figure 1. Fragment of a matching suspicious document (above) and source (below) from PAN (see next section), showing robustness to variations in both lexical selection and spelling.

Furthermore, with common n-gram patterns available from Google and Microsoft for research purposes, such approaches have even greater brute-forceability. Imagine, now, that many n-grams produce the same (ambiguous) hash, and the value of the approach should be clearer.

5. EVALUATION OF THE IPCRESS APPROACH - PAN

The Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) activity first appeared in 2007. The external detection part of the plagiarism detection task, where external refers to matching to source texts that are also available, changed markedly in 2012 from a prior moderate-sized index comparison to two tasks: i) a retrieval of documents from a search engine as might be useful in match; ii) a matching between given pairs of documents. Offering a search engine for the first of these avoids the need for those who have struggled to construct an efficient index with a few gigabytes of text to struggle further with terabytes. However, common search engines work best for plagiarism detection with long quoted phrases, although the offered search system does not. Further, this means presenting segments of the text in-clear to the search system, which doesn't readily fit with our context. We focus here on results obtained for ii) in 2012 and 2013, with a brief view of results in 2011 by way of contrast.

In Cooke et al (2011) we described various aspects of our system as used for the external plagiarism detection task, which could process the entire PAN11 collection within relatively short timescales without requiring a specialized computer cluster, and which was still able to produce a reasonable degree of matching performance (4th place, with PlagDet=0.2467329, Recall=0.1500480, Precision=0.7106536, Granularity=1.0058894). In 2012, we again showed good granularity (at or near 1, meaning that the same passage is not indicated multiple times) with high recall and precision for *non-obfuscated* text. A beneficial side-effect is that some obfuscation is handled by the same approach, but additional efforts need to be focused on obfuscation to offer a truly robust system.

Test	Plagdet Score	Recall	Precision	Granularity
02_no_obfuscation	**0.92530**	**0.90449**	**0.94709**	**1.0**
03_artificial_low	0.09837	0.05374	**0.93852**	**1.04688**
04_artificial_high	0.01508	0.00867	**0.96822**	**1.20313**
06_simulated_paraphrase	0.11229	0.05956	**0.97960**	**1.0**

In 2013, apart from for non-obfuscated data, descriptions of the nature of data used seem also to have shifted from the previous year. Our precision and granularity figures remain high, but it is difficult to conclude anything with regard to performance comparison for the other tasks – and prior examples of random obfuscation (see examples in [3]) suggest that this is not a realistic problem worth focusing on.

Test	Plagdet Score	Recall	Precision	Granularity
02_no_obfuscation	0.85884	0.83788	0.88088	**1.0**
03_random_obfuscation	0.04191	0.02142	**0.95968**	**1.0**
04_translation_obfuscation	0.01224	0.00616	**0.97273**	**1.0**
05_summary_obfuscation	0.00218	0.00109	**0.99591**	**1.0**

6. CONCLUSIONS

Widespread adoption of a system such as IPCRESS could engender a culture of IP protection, irrespective of the sharpness of the legal teeth in any particular jurisdiction, and offer an ability to address an apparent security risk (one of the four categories of supply chain risks [4]: supply, demand, operational and security) through information sharing of an over-eager or careless nature. It is likely in the interests of companies to adopt such a system, but assurances are vital in order to engender trust in its operation. And, in fact, the reputation of members of supply chains could be enhanced by a trustmark allied to a relative lack of identified IP issues, or be

drawn substantially into question where unresolved issues have leave to remain. The means by which such issues are identified, and the process towards resolution is another important aspect of the processes and procedures for IPCRESS, but is still to be defined. IPCRESS must be acceptable to adopt, and enforceable by contract within the supply chain in order to be successful. Of course, if such a system can be adopted in one supply chain, it is more likely to be adopted across all supply chains that involve each party. The view of supply chains as of sequences of producer/consumer relationships, akin to food chains, does not account for the complex reality of many supply chains. These can be variously interconnected with the same organisation acting many times on both sides of the supposed divide, and with larger organisations readily acting as suppliers to smaller ones. Larger companies in these supply chains would be readily positioned for such a scale of adoption.

Although the present approach works well, better treatment of obfuscation is likely an essential additional ingredient for adoption. The open publication of data, as here, regarding evaluations can only be helpful in trying to achieve this, and the strength/lossy nature of pattern production is readily demonstrable by Gedankenexperiment to private audiences, and is fully described in the patent filing. Where organisations were to agree on such an undertaking, exposure of the organisational archives to the pattern production process is the vital step, which presents a new but minor risk to information security: the pattern production process is one-way, but as with any such processing requires, temporarily, full access to unencrypted (clear text) document content. Such processing should ideally be undertaken in a network-isolated or security assertion enforcing system so as to assure the organisations that their document content cannot be leaked during such processing. The resulting index can also be visually inspected prior to release to increase the degree of assurance, and the mechanism for release should be relatively constrained also. Indeed, a vital effort in IPCRESS is to craft the set of acceptable processes and procedures for deploying IPCRESS and identifying associated risk levels in relation to information assurance. Related processes and procedures, including the drafting contractual terms for such an adoption, are also required.

ACKNOWLEDGEMENTS

The authors gratefully recognize prior contributions of Neil Newbold, Neil Cooke, Peter Wrobel and Henry Cooke to the formulation of the codebase used for this task and by Cooke and Wrobel to the patents generated from these efforts. This work has been supported in part by the the UK's Technology Strategy Board (TSB, 169201), and contributions from others at Surrey, Jaguar Land Rover and GeoLang are also acknowledged in this respect, and by EPSRC and JISC (EP/I034408/1). We recognize also the efforts of the PAN13 organizers in system evaluation (Section 5).

REFERENCES

[1] Carvalho, V.R. & Cohen, W.W. (2007) "Preventing Information Leaks in Email", *SIAM International Conference on Data Mining,* Minneapolis, Apr 2007.

[2] Cooke, N. & Gillam, L. (2011) "Clowns, Crowds and Clouds: A Cross-Enterprise Approach to Detecting Information Leakage without Leaking Information". In Mahmood, Z. and Hill, R. (eds.) Cloud Computing for Enterprise Architectures. Springer.

[3] Potthast, M., Barrón-Cedeño, A., Stein, B. & Rosso, P. (2010) "An Evaluation Framework for Plagiarism Detection". *23rd International Conference on Computational Linguistics (COLING*) 2010, August 23-27, Beijing, China.

[4] Manuj, I. & Mentzer, J.T. (2008) "Global supply chain risk management strategies", *International Journal of Physical Distribution & Logistics Management*, 38(3): 192-223