VALCRI
VISUAL ANALYTICS FOR SENSE-MAKING IN CRIMINAL INTELLIGENCE ANALYSIS

***VALCRI* WHITE PAPER SERIES**
VALCRI-WP-2017-007
1 January 2017

Research Data:
The Art and Science of Anonymising Data

Rick Adderley
AE-Solutions (BI) LTD
Badsey, Worcestershire WR11 7AA
UNITED KINGDOM

Project Coordinator
Middlesex University London
The Burroughs, Hendon
London NW4 4BT
United Kingdom.

Professor B.L. William Wong
Head, Interaction Design Centre
Faculty of Science and Technology
Email: w.wong@mdx.ac.uk

EUROPEAN COMMISSION

INTENTIONALLY BLANK

INTRODUCTION: ANONYMISATION OF DATA

The EU FP7 funded VALCRI project is the research and development of a visual analytics system for sense making in criminal intelligence analysis. In order to undertake the research and assess its impact on the project's End User community there is a fundamental requirement to have a set of data which is recognisable and realistic for use by real criminal analysts. One of the project's End User partners, West Midlands Police (WMP) has released a set of real data to another partner, A E Solutions (BI) Ltd (AES), to anonymise and subsequently release to the project for research and component testing. The data set comprised:

- million crime reports spanning three years
- 1.5 million person records relating to victims and offenders associated with the crime reports spanning three years
- 297,000 records of property stolen relating to the crime reports
- 147,000 records containing one or more descriptions of suspects seen at or near the scene of a crime
- million day-to-day incident reports spanning three years
- 184,000 custody records spanning one year
- 60,000 records of persons who have been stopped and searched during one year
- 200,000 intelligence records spanning one year
- 1.8 million financial transaction records spanning 4½ months
- 55 million automatic number plate recognition records and associated camera positions spanning three months
- 1,100 major incident witness statements

In order to ensure that the data cannot be reconstructed, the process taken to anonymise the data is not trivial. To quote Sweeney (Sweeney, 2002), the aim of the anonymisation process is; "Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful." A simply anonymised dataset does not contain a real name, home address, phone number or other obvious identifiers. However, in these data sets it is important that such information is present so that real Policing analytical processes may be tested.

DATA ANONYMISATION TECHNIQUES

There are a number of published techniques that have been used to anonymise data with varying degrees of competency. A widely-used technique, k-anonymisation, generalises and/or supresses various fields within a data set, whilst minimising the loss of information, to ensure that none of the records can be differentiated from at least other k-1 records (Sweeney, 2002). Generalisation is achieved by replacing individual field characteristics with a broader category. For example; ages can be replaced by age banding. Suppression is achieved by removing all or some values in a column. The processes of generalisation and suppression can have detrimental results when using the anonymised data in that the results can be skewed (Angiuli et al, 2015) but this effect can be avoided by altering the relevant algorithms (Angiuli & Waldo, 2016). It has also been suggested that k-anonymity is not suitable for high dimensional data sets (DeMontjoye et al, 2013). This is remarkably valid when the data contains four spatio-temporal fields. Additionally, should the person attempting to reconstruct the data personally know one of the data subjects (has background knowledge), the anonymised data may not be sufficiently supressed/generalised to protect the personal identity as the real data can contain sufficient information to make either an educated guess or full identification.

A refinement of the k-anonymisation method used to overcome some of its inherent weaknesses is the l-diversity method (Machanavajjhala et al, 2007). This method further reduces the granularity of the original data by ensuring that the generalisation algorithm certifies that all possible values are equally represented in equal proportions. This, however, is likely to cause significant loss of information which renders this technique unsuitable for use within the Valcri project. This technique is also unsuitable if the original data contains more than one sensitive field thus rendering this technique unsuitable for the Police data sets.

Improving on both of the above, Kenig and Tassa have developed an approximation algorithm based on generalisation and suppression that currently achieves lower information loss than any previous algorithm (Kenig & Tassa, 2012).

Data encryption is another technique that can be used to anonymise data. However, this would be unsuitable as the data need to be in a human readable format in order to undertake realistic analysis.

The Substitution anonymisation technique relies on the replacement of data within the columns with information from predefined lists of fictitious data. This method poses the challenges of collating and maintaining lists of such fictitious data which can run into millions for use with large data sets together with ensuring that the same substitutions are made across linked data sets.

Shuffling is another technique somewhat similar to Substitution. Within Shuffling the anonymised data is derived from each individual column but issues occur when using this technique with small data sets; there could be insuffi-

cient data items to effectively derive suitable anonymised results.

GENERAL PRACTICES AND PROCEDURES

In the UK, the 1998 Data Protection Act (Act D. P. 1998) (DPA) was passed by the British Parliament to manage the approach in which personal identification information is handled and to give legal rights to people who have personal information stored about them. Other European Union countries have passed similar laws as, often, information is held in more than one country. There are no such laws covering anonymised data but the UK Information Commissioner's Office (ICO) has published the Anonymisation Data Protection Risks Code of Practice (ICO, 2012) under section 51 of the DPA in pursuance of the ICO's duty to promote good practice. The DPA states good practice includes, but is not limited to, compliance with the requirements of the DPA. This code was also published with Recital 26 and Article 27 of the European Data Protection Directive (95/46/EC) in mind. These provisions make it clear that the principles of data protection do not apply to anonymised data and open the way for a code of practice on anonymisation.

Anonymisation within the Valcri project has been undertaken by following the Code of Practice principles. The Code states that when a data set is anonymised, there are various factors which should be considered to enable a satisfactory level of anonymisation:

- The possibility of re-identification being attempted
- The probability of successful re-identification
- The availability of anonymisation techniques
- The quality of the anonymised data as well as whether it will serve the purpose of the researcher using the anonymised information.

In conjunction with the above, the EU Article 29 Data Protection Working Party (Cotino) judges that there are three risks to be considered that are essential to data anonymisation:

(a) Singling out: which corresponds to the possibility to isolate some or all records which identify an individual in the dataset

(b) Linkability: which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g. by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against "singling out" but not against linkability

(c) Inference: which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

As per the definition by True Ultimate Standards Everywhere Inc (TRUSTe Blog, 2016), pure anonymisation is defined as taking information that is currently personal identification information (PII) and permanently turning in to non-identifiable data. Whereas, Pseudo-anonymisation is defined as converting PII into non-identifying data which can be returned from its anonymised state to PII in future. Personal identification information is not only information revealing names, addresses, phone numbers etc. but any information or combination of information that can be used to identify, contact or locate a discreet individual. Pseudo- anonymisation is not recommended in cases where highest security is required, as it is not effective as pure anonymisation (Vinogradov & Pastsyak, 2012). It may, however, be useful in evaluating and improving test runs of the data sets etc., where reproduction of the original data may sometimes be necessary.

There are four phases involved in the process of data anonymisation (Vinogradov & Pastsyak, 2012). These are:

(a) Data discovery and analysis: The data analysis phase identifies the data which is required to be anonymised so that it can be effectively protected without compromising its utilisation.

(b) Data planning and modelling: The planning and modelling phase is designed to develop the criteria that will be used to anonymise the data and build framework around the information that was obtained in the first stage. This stage is more about the choice of data anonymisation policy and approach than actually dealing with the critical data.

(c) Developing anonymisation models: The development stage consists of creating data anonymisation configuration modules depending on the specific requirements of the End User partners.

(d) Implementation: The Implementation and execution stage is designed to install an arrangement for incorporating data anonymisation into the overall data process.

It is important that the anonymised data cannot be reconstructed and individuals are reidentified and there are many publications where this has occurred, the most famous being the Robust De-anonymisation of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset (Narayanan & Shmatikov, 2008). On October 2, 2006, Netflix published data relevant to movie ratings provided by 500,000 of its users over a six-year period. comprising a dataset consisting of around 100 million movie ratings. All customer information was removed and only minimal alterations were made to the ratings data. Using a k-anonymity model combined with open source data, Narayanan & Shmatikov effectively recognised the Netflix records of known users as well as disclosing their political preferences and other personal sensitive information.

THE DATA ANONYMISATION PROCESS

To ensure confidentiality and security the techniques and methods used in the anonymisation process will not be disclosed. The examples provided below are intended to illustrate that the team is aware of a variety of anonymisation methods and that a reader would not be able to reconstruct the methods that have been used by the VALCRI team.

The Valcri data sets will be anonymised in two stages:

(a) Full anonymisation of all personal details. The geography will have limited anonymisation as it is required to have a degree of accuracy to enable the End User partners to test various components. At this stage in the project, Post Codes have not been anonymised which is due to the requirements within crime analytics to ascertain geographical information to include; distances travelled by criminals, densities of crime/incidents etc. According to the UK Office for National Statistics (ONS 2015) there are 1.3 million post codes with an average of 43 properties in each area. The minimum number of properties is one and the maximum is 3215 which could mean that by combining data sets, in some instances, it may be possible to identify individual(s). WMP have provided their approval in allowing the original Post Codes to remain in the anonymised data sets on release to VALCRI.

(b) Full anonymisation of all personal details and the geographical references to one kilometre block and/or post code sector. This further anonymisation will take place before the data sets can be released to the general research community to ensure that specific dwellings cannot be re-identified, whilst still providing the research community with data to perform meaningful, geography related research.

THE VALIDATION AND VERIFICATION PROCESS

The anonymisation processes are rigorously and thoroughly validated by a WMP analyst prior to release into the Valcri project and comprise:

All personal information is anonymised according to a standardised method devised by AES. Poor spellings, incorrect reference numbers, truncation of data fields, multiple names aligned to the same individual, the use of special characters within names etc., are all reproduced in the anonymisation process. To ensure consistency across data sets, the structured data items are anonymised and their values are entered into look up tables which are used in the anonymisation processes.

A “code book” of the process is maintained to enable the WMP analyst to ensure that the new data is consistent and appropriate.

Both the original and the anonymised data together with the codebook is hand delivered to the WMP analyst accompanied by a brief explanation sheet.

The analyst examines both sets of data to ensure that the anonymisation process has been properly undertaken and attempts to reconstruct/reidentify individuals associated with the data.

If reconstruction or reidentification can be achieved this information is passed back to AES who reassess and improve their anonymisation processes. Steps 2 to 4 are again followed.

If reconstruction or reidentification cannot be achieved step 5 is processed.

The anonymised data is certified as being able to be released into the project.

WMP retain the “code Book.” The original data is securely deleted from AES’ computers and the anonymised data is uploaded onto the Valcri server.

CONCLUSION

The anonymisation process is still work in progress. To date the following data sets have been released to the project; crime reports, person records, day-to-day incident records, free text modus operandi records covering a single year, intelligence records, custody records, automatic number plate recognition records, finance records and the witness statements. The remaining data sets are with WMP awaiting their validation.

During the first quarter of 2017 the data sets will be tested to determine if persons can be re-identified or data de-anonymised. Should this be possible, AES will refine and modify their processes and iterate this cycle until nothing can be reconstructed/reidentified.

Once the data has been certified as being irreversibly anonymised per the legislation, Code of Practice and to the satisfaction of WMP, it will be made available, with limitations, to the research community.

REFERENCES

Act, D.P., 1998. Data Protection Act. London Station Off.

Angiuli, O., Blitzstein, J. and Waldo, J., 2015. How to de-identify your data. Communications of the ACM, 58(12), pp.48-55.

Angiuli, O. and Waldo, J., 2016, June. Statistical Tradeoffs between Generalization and Suppression in the De-identification of Large-Scale Data Sets. In Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual (Vol. 2, pp. 589-593). IEEE.

Cotino, L., ARTICLE 29 DATA PROTECTION WORKING PARTY.

"Data Anonymization." Web log post. TRUSTe Blog. Jim Rennie, 2013. Web. http://www.truste.com/blog/2013/04/16/data-anonymization/, accessed 20th October 2016.

De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M. and Blondel, V.D., 2013. Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3.

ICO (2012). Anonymisation data protection risks code of practice. https://ico.org.uk/media/1061/anonymisation-code.pdf. Accessed 20th October 2016

Kenig, B. and Tassa, T., 2012. A practical approximation algorithm for optimal k-anonymity. Data Mining and Knowledge Discovery, 25(1), pp.134-168.

Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M., 2007. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), p.3.

Narayanan, A. and Shmatikov, V., 2008. Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset). 2008.University of Texas at Austin.

Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), pp.557-570.

Vinogradov, S. and Pastsyak, A., 2012. Evaluation of data anonymization tools. In Proc. Intl. Conf. on Advances in Databases, Knowledge, and Data Applications (DBKDA) (pp. 163-168).

INTENTIONALLY BLANK

The research leading to the results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) through Project VALCRI, European Commission Grant Agreement Number FP7-IP-608142, awarded to Middlesex University and partners.

	VALCRI Partners	**Country**
1	Middlesex University London Professor B.L. William Wong, Project Coordinator Professor Ifan Shepherd, Deputy Project Coordinator	United Kingdom
2	Space Applications Services NV Mr Rani Pinchuck	Belgium
3	Universitat Konstanz Professor Daniel Keim	Germany
4	Linkopings Universitet Professor Henrik Eriksson	Sweden
5	City University of London Professor Jason Dykes	United Kingdom
6	Katholieke Universiteit Leuven Professor Frank Verbruggen	Belgium
7	A E Solutions (BI) Limited Dr Rick Adderley	United Kingdom
8	Technische Universitaet Graz Professor Dietrich Albert	Austria
9	Fraunhofer-Gesellschaft Zur Foerderung Der Angewandten Forschung E.V. Mr. Patrick Aichroft	Germany
10	Technische Universitaet Wien Assoc. Prof. Margit Pohl	Austria
11	ObjectSecurity Ltd Mr Rudolf Schriener	United Kingdom
12	Unabhaengiges Landeszentrum fuer Datenschutz Dr Marit Hansen	Germany
13	i-Intelligence Mr Chris Pallaris	Switzerland
14	Exipple Studio SL Mr German Leon	Spain
15	Lokale Politie Antwerpen	Belgium
16	Belgian Federal Police	Belgium
17	West Midlands Police	United Kingdom