Manuscript details

Manuscript number	IJIM_2015_152
Title	Re-Identification Attacks: a Systematic Literature Review
Article type	Research Paper

Abstract

The publication of increasing amounts of anonymised open source data has resulted in a worryingly rising number of successful re-identification attacks. This has a number of privacy and security implications both on an individual and corporate level. This paper uses a Systematic Literature Review to investigate the depth and extent of this problem as reported in peer reviewed literature. Using a detailed protocol ,seven research portals were explored, 10,873 database entries were searched, from which a subset of 220 papers were selected for further review. From this total, 55 papers were selected as being within scope and to be included in the final review. The main review findings are that 72.7% of all successful re-identification attacks have taken place since 2009. Most attacks use multiple datasets. The majority of them have taken place on global datasets such as social networking data, and have been conducted by US based researchers. Furthermore, the number of datasets can be used as an attribute. Because privacy breaches have security, policy and legal implications (e.g. data protection, Safe Harbor etc.), the work highlights the need for new and improved anonymisation techniques or indeed, a fresh approach to open source publishing.

Keywords Re-identification, anonymisation, anonymization, systematic literature review

Manuscript region of origin	Europe
Corresponding Author	Jane Henriksen-Bulmer
Order of Authors	Jane Henriksen-Bulmer, Sheridan Jeary
Suggested reviewers	Margaret Ross, Tom Jackson, Andrew Simpson

Submission files included in this PDF

File Type	**File Name**
Title Page (with Author Details)	Title page.docx
Highlights	Highlights - Re-identification attacks a SLR.docx
Cover Letter	cover_letter.pdf
Manuscript File	Re-identification Attacks_A Systematic Literature Revie V1.docx
Response to Reviewers (without Author Details)	Response to reviewers 23Mar16.docx

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Re-Identification Attacks – A Systematic Literature Review

Jane Henriksen Bulmer (corresponding author)
Faculty of Science and Technology
Bournemouth University
Talbot Campus, Fern Barrow
Bournemouth
Dorset
United Kingdom
Jhenriksenbulmer@bournemouth.ac.uk

Dr Sheridan Jeary
Faculty of Science and Technology
Bournemouth University
Talbot Campus, Fern Barrow
Bournemouth
Dorset
United Kingdom
sjeary@bournemouth.ac.uk

Abstract:

The publication of increasing amounts of anonymised open source data has resulted in a worryingly rising number of successful re-identification attacks. This has a number of privacy and security implications both on an individual and corporate level.

This paper uses a Systematic Literature Review to investigate the depth and extent of this problem as reported in peer reviewed literature. Using a detailed protocol ,seven research portals were explored, 10,873 database entries were searched, from which a subset of 220 papers were selected for further review. From this total, 55 papers were selected as being within scope and to be included in the final review.

The main review findings are that 72.7% of all successful re-identification attacks have taken place since 2009. Most attacks use multiple datasets. The majority of them have taken place on global datasets such as social networking data, and have been conducted by US based researchers. Furthermore, the number of datasets can be used as an attribute.

Because privacy breaches have security, policy and legal implications (e.g. data protection, Safe Harbor etc.), the work highlights the need for new and improved anonymisation techniques or indeed, a fresh approach to open source publishing.

Keywords:

Re-identification, anonymisation, anonymization, systematic literature review

Re-identification attacks: a Systematic Literature Review

Highlights:

- A Systematic Literature Review investigating the prevalence of re-identification attacks on publically available datasets.
- Attacks have been categorised into five attack categories, with the majority of successful attacks having been conducted utilising multiple datasets.
- Further findings are that most research in this area has been conducted on Global datasets by American researchers; it is only in recent years that re-identification attacks been attempted by researchers elsewhere in the world.
- Another finding is that the number of datasets can be used as an attribute.

BU
Bournemouth
University

12th October 2015

International Journal of Information Management
Elsevier Limited (Corporate Office)
125 London Wall,
London, C2Y 5ASr

Dear Sirs,

Re: Re-Identification: a Systematic Literature Review

Please find attached our submission for your consideration for potential inclusion in the International Journal of Information Management.

With regards to reviewing the paper for suitability, the following candidates are suggested as potential reviewers of this paper:

Associate Professor Andrew Simpson
Department of Computer Science
Kellogg College, Oxford University
andrew.simpson@cs.ox.ac.uk

Professor Tom Jackson
Professor of Information and Knowledge Management
Director Centre for Information Management
Associate Dean [Research]
Loughborough University
T.W.Jackson@lboro.ac.uk

Professor Margaret Ross MBE,
Southampton Solent University
Sir James Matthews Building,
157-187 Above Bar Street,
Southampton SO14 7NN
Margaret.ross@solent.ac.uk

If you have any queries or require any further information, please do not hesitate to contact me.

Yours faithfully

Jane Henriksen-Bulmer , MSc, MBA, LLb (Hons)

Vitae

Jane Henriksen-Bulmer, MSc, MBA, LLb
Having gained a first class honours degree in Law with the Open University in December 2010, Jane continued her studies in Perth, Western Australia where she received a Masters in Business Administration (MBA) from Curtin University in 2013. Following this Jane completed an MSc in Information Technology from Bournemouth University in 2015, gaining a distinction. Most recently, she has been accepted onto the PhD Research Degree Programme at Bournemouth University, Department of Science and Technology.

Dr Sheridan Jeary
Sheridan received her PhD from Bournemouth University in 2010 in Requirements Engineering and has continued to research in the area of requirements, particularly models, since that time. She is particularly interested in the current trend of releasing data to meet software requirements without consideration of personal consequence.

1.0 Introduction

Where traditionally marketeers south insight into customers and their preferences by using techniques such as; psychographic variables (Abduljalil & Hon, 2011) and; market segmentation (Yankelovich & Meer, 2006), with advances in technology and the advent of ever-larger collections of data, big data has changed all that.

Big data is a term used to describe the analysis and storage of very large amounts of complex data, defined by Gartner as; “high-volume, -velocity and -variety information assets” (Sicular, 2013) that, when processed, can be used to; “enable enhanced decision-making, insight discovery and process optimization” (ICO, 2012).

Data is the lifeblood of most organisations and it is estimated that up to 80% of all data held in organisations, can now be classed as big data (Khan, et al., 2014). Organisations and people produce and use data in many ways to further their businesses or interest. With the use of the Internet and the exponential growth in data being published in the public domain, in excess of 2 billion people worldwide are now connected to the Internet.

The rate of data generated is expected to rise by 40 zettabytes (ZB) by 2020 and continue to rise at a rate of 50-60% annually beyond that (Khan, et al., 2014). As a result, organisations and individuals now have access to a much wider and varied corpus of data than ever before, this has been termed the ‘era of big data’ (Berner, Graupner, & Maedche, 2014; Rotella, 2012).

When data is published in the public domain, the information may be published by private organisations (e.g. Netflix and AOL, (Ohm, 2010)), or, it may be released by individuals themselves through for example, social media sites such as Facebook, LinkedIn, Twitter or similar social networking platforms.

This means that data mining big data has revolutionised how companies find out about individuals, and their preferences. Marketeers have realised that mining big data has the potential to provide them with valuable insight into customer preferences and behaviours in ways not previously possible (e.g. see (Duhigg, 2012)).

This is not just true of private companies; public organisations are also realising the value of big data. They however, have entered the big data arena from the perspective of economies of scale and data sharing, seeking to "use technology to join up and share services rather than duplicate them" (The Cabinet Office, 2005, p. 1).

To this end government agencies have, for the last decade or so, been working on a variety of big data projects designed to integrate back office systems with front office services initially through the e-government agenda, then through the transforming government agenda (Patterson, Bennett, & Waine, 2008) and more recently through the seizing the data opportunity strategy (Department for Business Innovation and Skills, 2013).

However, these government initiatives have not stopped at local level, integrating services within individual government departments or even government agencies, many of the projects have been more ambitious seeking to create national datasets and indeed, creating open source access to government datasets.

This trend has been brought about by the Re-use of Public Sector Information Regulations 2005 and, more recently, 2015 (ROPSIR), implementing EU Directives 2003/98/EC and 2013/37/EU. ROPSIR places an obligation on public bodies to make data available for re-use and to, where possible, release such data in electronic format where possible (ROPSIR 2015, s. 11). Thus, public bodies now regularly contribute to data publishing, releasing increasing amounts of information and datasets open source (Department for Business Innovation and Skills, 2013; Simpson, 2011).

In the UK more than 20,000 datasets have been made available through the data.gov.uk site since 2010 (Data.gov.uk, 2016), and in the United States (US), in excess of one million datasets have so far been made available through open source portals (Gkoulalas-Divanis & Aonghusa, 2014).

From a corporate perspective, organisations use big data to try to gain commercial advantage. For example, organisations use big data analytics (data mining) to discover more about their customers and identify trends (Goodman, 2015). From an individual perspective this raises questions about how much insight can be gleaned into our lives and indeed, our current situation or whereabouts which in turn, raises serious concerns over the privacy and security of personal information (Ohm, 2010).

Some protection does exist. For example, the Data Protection Act 1998 (DPA) requires that any personally identifiable information may only be released with express permission of the individual. Further, the 2013 EU directive on the re-use of public sector information does state that individuals right to privacy, which is protected under Directive 95/46/EC, should be preserved prior to the release of any public data (2013/37/EU, Para. 11). Thus, before release these datasets will have been anonymised to prevent companies or individuals from identifying any of the individuals the data might relate to ((2013/37/EU, Para. 21).

There are a number of anonymisation techniques in use (Fung, Ke, Rui, & Yu, 2010; Lan, Yilei, & Yingjie, 2012) that can be used to de-identify data. How the anonymisation is done depends on the country of origin. For example, in the US open source published dataset in the health sector must be de-identified in accordance with the anonymisation rules laid out in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, better known as the “Safe harbor” standards, prior to public release (Health Information Privacy (HIP), 2014).

In the United Kingdom, the Information Commissioner’s Office (ICO) has issued a code of practice on anonymisation (ICO, 2012) which provides guidelines on data de-identification and pseudonymisation in order to limit the risk of re-identification taking place.

However, these methods are not completely risk free and re-identification is a real risk around the world (El Emam, Jonker, Arbuckle, & Malin, 2011; MacRae, Dobbie, & Ranchhod, 2012; Ohm, 2010), particularly where data miners use multiple datasets to retrieve personal information from the data.

Most recently, this caused the Health and Social Care Information Centre (HSCIC) to halt the release of UK anonymised health data (part of the care.data project) for six months amid fears over data privacy and security (Kirby, 2014; Walker, Meikle, & Ramesh, 2014).

This paper seeks to look into this problem by conducting a systematic literature review (SLR) of research that provides information and details of successful data re-identification cases. More particularly, the paper will also explore whether re-identification attempts are more successful where one or more of the datasets mined include geographical (GIS) or spatial data.

El Emam et al. conducted a SLR in 2011, which sought to identify successful cases of de-identification in the Health Sector (El Emam, Jonker, et al., 2011). They found 14 cases where successful re-identification had taken place, 10 of which involved US datasets. Since then research into re-identification has been successful in New Zealand (MacRae, et al., 2012) the UK and Canada (El Emam, Buckeridge, et al., 2011) to name but a few.

Furthermore, with the advances in data mining and so much more data being made available on a daily basis (McAfee & Brynjolfsson, 2012), an updated review would be appropriate.

The rest of the article is organised as follows. **Section 2** explains the research questions and review methodology; **Section 3** presents the findings of the review; **Section 4** discusses the findings and describes open issues, challenges and opportunities for further research; **Section 5** provides an overview of limitations; and **Section 6** concludes the article. **Appendix A** contains definitions of terminology, whilst **Appendix B** contains a full list of papers included in the review.

2.0 Materials and Methods

The review has been conducted following the protocol of Beecham, Baddoo, Hall, Robinson, and Sharp (2008), and the methodology and guidelines of Kitchenham, (Kitchenham, 2004; Kitchenham & Charters, 2007).

2.1 Research Questions

The research questions addressed by the review were limited to four questions that asked firstly how many instances of re-identification have proved successful? Of those, how many datasets were mined to conduct the re-identification tests? Where did the datasets originate? Finally, did any of the datasets mined include geographical (mapping) data?

However, the findings, as will be shown, lent themselves to much deeper analysis, and therefore, the resulting research questions this article will address are as follows:

RQ1: How many successful re-identification attempts have been carried out; which country did the paper originate in and where was it published?

RQ2: What types and how many datasets were mined in the successful re-identification attempts?

RQ3: How many and what types attributes were used to conduct the re-identification?

RQ4: Did any of the datasets include mapping (GIS) data?

2.2 Data Sources

The papers selected for inclusion in the review were selected from a database search of seven electronic databases. The databases were chosen based on a combination of a sample search of databases that held details of strategic literature reviews conducted in the software engineering field, and the recommended databases of Brereton, Kitchenham, Budgen, Turner, and Khalil (2007) and Kitchenham and Charters (2007). Table 1 lists the seven electronic databases that were searched for relevant papers in this review.

Table 1

Databases Searched
IEEExplore
ACM Digital Library
Google Scholar
Citeseer library
ScienceDirect
SpringerLink
Wiley InterScience

2.3 The Search Process

The papers for inclusion were selected through a process of three phases. **Phase one, Identification**: the titles and abstracts were read to select suitable papers for inclusion. **Phase two, Screening**: abstracts and conclusions were read and scored in accordance with relevance. **Phase three, Eligibility**: a full review of papers from phase two, including a review of the citations of the selected papers. The inclusion/exclusion decision was made based on the assessment criteria laid out in Table 2.

Table 2 – Phase 3 Quality Assessments

Item	Assessment Criteria	Score (0-1)	Score Response options
1	Has re-identification attempt been conducted		1 = Yes 0 = No
2	Has re-identification attempt been successful		1 = Yes 0 = No
3	Is study peer reviewed		1 = Yes 0 = No
4	Did any of the datasets include GIS/spatial data		0.5 = Yes 0 = No
Total Quality Score			

(Adapted from Beecham et al. 2008, p. 9-10)

Papers where re-identification had proved successful were included in the review.

2.4 Inclusion/Exclusion Criteria

The review targeted papers where re-identification had proved successful. Only papers written in English were targeted. The key search terms used are listed in Table 3.

Table 3

Search Terms Used	
Anonymisation	Deanonymisation
Identification	Re-identification
Pseudonymisation	Privacy
Spatial	GIS

To keep focus on answering the research questions, Table 4 lists the subtopics that were excluded in the search criteria.

Table 4

Search Terms Excluded	
Genomics (DNA) and fingerprint data	Vehicle data
Graphics	video surveillance
Location Privacy unless the location data had been utilised as part of a data	Images and Image processing

linking re-identification attempt and combined with one or more relational datasets	

2.5 Data Collection

At each phase, data was extracted Table 5 shows the data extracted from each study during phases two and three.

Table 5

Phase 2 – Data Extracted	Phase 3 – Data Extracted
Data source and type of publication	Data set origin
Paper title	Data set size
Abstract	Number of datasets mined
Conclusion	Types of dataset mined
Publication year	Re-identification strategy used
Author(s)	Number of attributes used
Country of study	Types of attributes used
Full citation	
Number of times cited	

3.0 Results

In total 10,873 database entries were searched during phase one, from which 220 papers were selected for inclusion in phase two. From this 50 papers were selected for inclusion in phase three, with an additional 33 further papers reviewed from the citations of the selected papers. By the end of phase three, the final number of papers included in the review was 55. A full list of the papers included in the review can be found in Appendix B.

3.1 Research Questions

RQ1: How many successful re-identification attempts have been carried out; which country did the paper originate in and where was it published?

There were 55 instances of successful re-identification attempts. In the majority of cases, the research had been conducted by American researchers (56%). Only 3.6% of the papers originated from UK research teams. The highest percentage of papers were published in Conference papers or Workshops (61%), whilst only 8 of the papers have been published in Academic Journals.

RQ2: What types and how many datasets were mined in the successful re-identification attempts?

The search engine (e.g. Google, yahoo) was the most popular type of dataset used in the re-identification process at 43.6%, followed closely by social network data (32.7%). The data types used were predominantly relational data (54.4%). However, in recent years the use of dynamic data has increased to become the most popular data type with 16 out of 29 papers published between 2011 and 2014 using dynamic data. This corresponds with the increase in Internet and social network usage in recent years (Khan, et al., 2014; Smith, Szongott, Henne, & von Voigt, 2012).

What was noteworthy was that, whilst the majority of the researchers were American, only 32.7% of the datasets used were US based. Looking at where the datasets originated, it was found that over half of the attacks were conducted on global datasets such as social networking sites and/or search engines.

Table 6

Year Published * No of Datasets used Crosstabulation

Count

		No of Datasets used								
		1	2	3	4	5	6	8	16	Total
Year Published	pre-2000	2	0	0	0	0	0	0	0	2
	2001	0	0	1	0	0	0	0	0	1
	2002	0	1	0	0	0	0	0	0	1
	2003	1	0	0	0	0	0	0	0	1
	2005	2	0	0	0	0	0	0	0	2
	2006	2	2	0	0	0	0	0	0	4

	2007	0	0	0	1	0	0	0	0	1
	2008	3	0	0	0	0	0	0	0	3
	2009	2	2	0	1	0	0	0	1	6
	2010	1	0	3	0	0	0	1	0	5
	2011	2	5	2	0	1	1	0	0	11
	2012	2	6	2	0	0	1	0	0	11
	2013	2	3	1	0	0	0	0	0	6
	2014	0	0	1	0	0	0	0	0	1
Total		19	19	10	2	1	2	1	1	55

The number of datasets mined ranged from 1 – 16, with one or two datasets proving the most popular at 34.5% respectively. Looking at this over time, it was also evident that more papers have been published in recent years with 40 of the papers published since 2009 (Table 6).

RQ3: How many and what types of attributes were used to conduct the re-identification?

The number of attributes used to re-identify were two or three (70.9%). This corresponds with early research which showed that 2 attributes (1 key attribute and 1 identifier) are required for successful re-identification (Latanya Sweeney, 1997), or where no identifier is available, three key-attributes are required to uniquely identify (L. Sweeney, 2000).

The most common attributes used to aid in the re-identification were key-attributes then sensitive attributes and finally, identifying attributes. This result is to be expected where data has been anonymised and direct identifiers removed or obfuscated.

Another interesting finding was that when comparing the number of attributes used with the number of datasets mined, that the number of datasets used can be used as an attribute. Whilst, research has proved that multiple attributes are necessary for successful re-identification (Latanya Sweeney, 1997; L. Sweeney, 2000), it was found that in papers that had utilised only one attribute or one dataset to re-identify, this was combined with multiple attributes or datasets respectively in order for the re-identification attack to be successful.

RQ4: Did any of the datasets include mapping (GIS) data?

There were only 3 of the 55 papers included in the review that solely utilised GIS data to re-identify. However, 34.6% of all the attacks used location and/or GIS data as part of the re-identification process.

3.2 Other Findings

During phase three, it became apparent that there were different types of re-identification attack depending on the types of data worked with and the research and re-identification strategy used. A decision was therefore made to categorise the results into five types of attack. These are described briefly in Table 7.

Table 7

Type of Attack	Classification
Aggregation of information attack	Using multiple datasets to achieve re-identification by data linking across datasets looking for data overlaps (Clark, 2012; Ochoa, Rasmussen, Robson, & Salib, 2001; Latanya Sweeney, 2011)
Inference/Other attack	Where inference or prior knowledge has been used to re-identify (linkage attacks) (Fung, et al., 2010). This category also covers attacks that do not fit into any of the other categories
Anonymisation Reversed attack	An attack that involves using background knowledge of the anonymisation method and/or algorithm(s) used. (Abou-el-ela Abdou, Nermin, & Hesham, 2013) or where statistical means have been used to re-identify (Benitez & Malin, 2010; Koot, van't Noordende, & de Laat, 2010; L. Sweeney, 2000)

Graph/Node attack	Using nodes, labels and/or topology to re-identify (Sharma, Gupta, & Bhatnagar, 2012)
GIS/location attack	Re-identification using spatial, GIS and/or graph data (Cassa, Wieland, & Mandl, 2008)

Following on from this categorisation, a more in-depth analysis of the data collected was conducted. The findings from each of these categories are outlines in sections 4.3 – 4.7.

3.3 Aggregation of Information Attack

This formed the largest group of attacks with 20 of the 55 papers falling into this category. Furthermore, of those 70% were papers published since 2011.

Whilst this group of attacks consisted of attacks using multiple datasets, in over half the attacks, at least one of the datasets used for the attack included either social network and/or a search engine followed closely by public datasets (used in 9 of the 20 papers).

What is interesting is that whist public datasets were chosen as a dataset of choice in the early work (Ochoa, et al., 2001; Latanya Sweeney, 1997), there appeared to be a dip in popularity of using public data between 2002 and 2010 with no papers in this category utilising public data (see Table 8). Then, in 2011, public data once again became a dataset of choice for re-identification attempts.

Table 8

Publicationyear*$Data_Sets*AttackCategoryAggregationofinfo1OtherInference2An Crosstabulation

				Data Sets Used[a]								
Attack Category				Email	Government	Electricity	Ratings_Tags	Searh_Engine	Social_Network	Reference	Other	Total
Aggregation of	Publicati	pre-2006	Count	0	2	0	0	0	0	0	0	2

info	on year	2009	Count	0	0	0	0	2	2	1	1	2
		2010	Count	1	0	0	0	2	2	0	0	2
		2011	Count	0	5	1	0	2	2	2	3	7
		2012	Count	0	1	0	1	3	4	2	0	5
		2013	Count	0	1	0	0	2	2	0	0	2
	Total		Count	1	9	1	1	11	12	5	4	20

Percentages and totals are based on respondents.

a. Dichotomy group tabulated at value 1.

Another finding in this category of attack was that in 58.8% of attacks, the attackers had used identifiers as one of the attributes and this had been combined with a key attribute in 87.5% of attacks. This finding is somewhat surprising given that identifiers should, in theory, have been removed as part of the anonymisation process prior to the data being released.

However, bearing in mind that social networking data has been utilised as one of the datasets of choice in 12 of the 20 papers in this category, individuals' names and/or usernames will have been more widely available and therefore an obvious target for use as an attribute in any re-identification attempt.

3.4 Inference/Other Attack

This category covers attacks where the attacker has used existing knowledge to aid the attack. It also covers other types of attack that did not fall into any of the other categories. No particular patterns were found in this category, perhaps due to the variety of re-identification techniques used.

There were 8 papers in this category ranging from re-identification by analysing writing styles (Almishari & Tsudik, 2012), to using machine learning such as weka software (Hall, et al., 2009) in combination with inference to re-identify users (Sramka, 2010; Sramka, Safavi-Naini, & Denzinger, 2009).

The most interesting paper in the category however, was a paper detailing how, by analysing electricity meter readings over time, researchers were able to identify patterns and thus, re-identify the household (Buchmann, Bohm, Burghardt, & Kessler, 2013).

3.5 Anonymisation Reversed Attack

The anonymisation reversed category consists of 11 papers where researchers had reversed the anonymisation applied to the original data. The predominant dataset mined in the category were public datasets (54%). In fact it was not until 2006 that researchers attempted to reverse anonymisation on a non-public dataset, when two researchers showed that it was possible to positively re-identify from a non-public dataset when they re-identified users from published anonymised movie ratings (Narayanan & Shmatikov, 2006).
This case involved the Netflix movie ratings that had been released as part of the Kaggle data mining challenge.

It is conceivable that the reason for this timelap before non-public datasets were used for re-identification were that, prior to 2006, the use of Social Networks and search engines had not yet become commonplace. Whilst social networks have been around since the late 1990s, it was not until 2006, when MySpace, Facebook and Twitter started to become popular, that social networking really took off. Facebook now have an estimated 500 million users (BRASS Program Planning Committee, 2011).

However, the Netflix case was the first to prove that the scope of re-identification reached beyond public datasets, meaning an attack was possible on any released datasets.

3.6 Graph/Node Attack

The graph/node attack category proved to contain the second largest corpus of papers with 23.6% of the papers in the review.

This group of attacks contains papers where researchers have primarily used dynamic data such as social networking or search engine data graphs to conduct the re-identification attacks. The attacks were conducted by linking graphs in order to re-identify (Gayo Avello, 2011; Peng, Li, Zou, & Wu, 2012). Early papers in this category utilised the Enron email database that was released as part of the Enron bankruptcy proceedings by the US Federal Energy Regulator in 2001 (Hay, Miklau, Jensen, Weis, & Srivastava, 2007; McCallum, Corrada-Emmanuel, & Wang, 2005; Shetty & Adibi, 2005). It can therefore be surmised that the Enron email database was the dataset upon which early graph-linking theory was based.

3.7 GIS/location Attack

The last group of attacks is also the smallest with only 3 papers falling into this category. For that reason, no noticeable trends were found.

All three papers had used geographical location tags as one of the attributes and combined this with other data such as tweets or timestamps in the re-identification attempts (Friedland, Maier, Sommer, & Weaver, 2011; Goga, et al., 2013; Jedrzejczyk, Price, Bandara, & Nuseibeh, 2009). Therefore, arguably these papers could equally have been placed in the data aggregation category.

4.0 Discussion

The earliest successful re-identification attempt was published in the late 1990’s, when Latanya Sweeney (1997) matched the Cambridge, Massachusetts voters roll to medical records to re-identify. After this, numbers of published attacks rose slowly up until the mid-2000 when this type of research or attack started to gain momentum. However, in more recent years, numbers of attacks have risen considerably with thirty of the successful attacks found having been published since 2011. Thus, it can be surmised that as research in this area is becoming established there may be many more instances of re-identification reported.

Looking at the trends for using public datasets for re-identification attacks, public data was used in early years and then not utilised for nearly a decade (2002 - 2010). This may be explained by looking at the history of anonymisation and the release of pubic data. Sweeney was the first researcher to identify the link between publically released data and the ability to re-identify. This led to the development of the k-anonymisation algorithm (Samarati & Sweeney, 1998), now accepted as the minimum anonymisation standard for data publishing (Abou-el-ela Abdou, et al., 2013). This enabled organisations and public bodies to anonymise data prior to release, which could explain the lack of successful re-identification attempts on public data during the middle year group. However, with the increasing number of public dataset being released open source in recent years (Department for Business Innovation and Skills, 2013), more public data has become accessible and thus, public datasets have once again become a dataset of choice for re-identification attempts.

What was interesting was that, in view of the UK government data publishing policy (Department for Business Innovation and Skills, 2013), it should follow that the UK would have more research into this area. Yet, only 2 of the papers found had been written by UK researchers, one of which used a combination of location (geographic) data and dynamic data. In this paper, the research team uncovered both identifiers (names and addresses) and key identifiers (age, occupation and email) by studying users movements and linking these to publically available data from social networks and search engines to successfully re-identify (Jedrzejczyk, et al., 2009)

The second paper was less specific in its findings. However, what was uncovered was interesting, in that the researchers only searched public data that has been available open source. A team of MBA students from Oxford were given an assignment to consider the security and privacy implications of public datasets being made available open source through sites such as data.gov.uk. What was discovered ranged from being able to ascertain staff movement within public buildings based on the energy consumption of the buildings through to being able to successfully re-identify Senior Military Personnel and ascertain their salaries (Simpson, 2011). While this was a small study for a particular assignment, it does beg the question; how much more insight could be gained if an attacker was to mine the data in more depth?

Looking at where research originated more generally, it was found that the majority of re-identification attacks originated in the Americas with US and Canada conducting 67.2% of all research found. The remaining body of research was widely spread out in origin; 16.4% originating in Europe, 9% in Asia and 7.3% of research was collaborations between multiple nations. Furthermore, most of the research conducted outside the Americas, has been carried out in the last 5 years (66.6%). This would indicate that, whilst researchers in the Americas have been looking at this area for quite some time, it is only recently that researchers in the rest of the world have started to take more of an interest.

5.0 Limitations

The searches were carried out and scored following the methodology of Kitchenham and Charters (2007) with the primary researcher conducting phases one and two. To guard against potential bias, the selection criteria and quality checks were developed by, and agreed between, the authors. In addition, the protocol defined that for phase two of the review an independent researcher would check and verify the selection. However, due to mitigating circumstances of the independent researcher and time constraints, this did not happen. Rather the primary author conducted all three phases alone , overseen by the second author.

The number of additional papers selected from the citations during phase three appeared rather high and therefore, this was investigated further to ensure the methodology had been followed correctly. What transpired was that, whilst 33 additional papers were selected for inclusion, 42% of those did not pass the relevancy test to be included in the final review. It was concluded that a combination of deepening of subject knowledge and the primary researcher perhaps being overcautious in selecting papers from the citation to make up for the lack of a second reviewer would account for this apparent discrepancy.

6.0 Conclusion

This review has shown how the number of successful re-identification attempts on publically available datasets has risen sharply, particularly in recent years with 72.7% of all papers found having been published since 2009. The review has also shown that whilst the Americas have been conducting research in this area for over 2 decades, it is only in recent years that the rest of the world have started to take note and produce papers on re-identification attacks. This would indicate that this area of research is growing.

With the many methods and strategies already used in the re-identification process, it is likely that as researchers become more adept at re-identification, more and more successful attacks will occur. Already there are many organisations who make their living from selling data analytic results or indeed helping companies analyse their own big data (e.g. SAS Institute Inc., 2015). If these trends are allowed to continue without intervention, no dataset, whether anonymised or not, will be safe from the threat of re-identification.

Thus, much scope exists in this field of study to, not only develop more robust anonymisation techniques, but also put in place better safeguards around publishing any information pertaining to individuals. Furthermore, to fully understand the depth and breadth of this problem, opportunities exist for not only exploring and mining the large corpus of public data available, but also to review the security, privacy and policy implications that re-identification attacks bring. It may even require a completely fresh approach to data publishing, to minimise the risk of privacy breaches occurring in the first place.

7.0 References

Abduljalil, S., & Hon, D. T. H. (2011). THE ROLE OF PSYCHOGRAPHIC FOR DISTINGUISHING MAIN CATEGORIES OF CONSUMERS BASED ON LIFESTYLE, PERSONALITY AND VALUE VARIABLES. *International journal of economics and research, 02*, 29.

Abou-el-ela Abdou, H., Nermin, H., & Hesham, A. H. (2013). Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. *Journal of Information Security*, 101.

Almishari, M., & Tsudik, G. (2012). Exploring linkability of user reviews. In *Computer Security–ESORICS 2012* (pp. 307-324): Springer.

Beecham, S., Baddoo, N., Hall, T., Robinson, H., & Sharp, H. (2008). Motivation in Software Engineering: A systematic literature review. *Information and Software Technology, 50*, 860-878.

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association, 17*, 169-177.

Berner, M., Graupner, E., & Maedche, A. (2014). THE INFORMATION PANOPTICON IN THE BIG DATA ERA. *Journal of Organization Design, 3*, 14-19.

BRASS Program Planning Committee. (2011). The Business of Social Media: How to Plunder the Treasure Trove. *Reference & User Services Quarterly, 51*, 127-132.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems & Software, 80*, 571-583.

Buchmann, E., Bohm, K., Burghardt, T., & Kessler, S. (2013). Re-identification of Smart Meter data. *Personal Ubiquitous Comput., 17*, 653-662.

Cassa, C., Wieland, S., & Mandl, K. (2008). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics, 7*, 1-9.

Clark, J. W. (2012). Correlating a persona to a person. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 851-859): IEEE.

Data.gov.uk. (2016). Data.gov.uk: Opening up Government: Data: Datasets. In. Online: Data.gov.uk.

Department for Business Innovation and Skills. (2013). Seizing the data opportunity: A strategy for UK data capability In. London.

Duhigg, C. (2012). How companies learn your secrets. In. New York TImes.

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics And Decision Making, 11*, 46-46.

El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A Systematic Review of Re-Identification Attacks on Health Data. *PLoS ONE, 6*, 1-12.

Friedland, G., Maier, G., Sommer, R., & Weaver, N. (2011). Sherlock holmes' evil twin: on the impact of global inference for online privacy. In

Proceedings of the 2011 workshop on New security paradigms workshop (pp. 105-114): ACM.

Fung, B. C. M., Ke, W., Rui, C., & Yu, P. S. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys, 42*, 14:11-14:53.

Gayo Avello, D. (2011). All liaisons are dangerous when all your friends are known to us. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (pp. 171-180): ACM.

Gkoulalas-Divanis, A., & Aonghusa, P. M. (2014). Privacy protection in open information management platforms. *IBM Journal of Research & Development, 58*, 1-11.

Goga, O., Lei, H., Parthasarathi, S. H. K., Friedland, G., Sommer, R., & Teixeira, R. (2013). Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 447-458): International World Wide Web Conferences Steering Committee.

Goodman, J. (2015). Big data: too much information. In *The Law Society Gazette*. Law Gazette online: The Law Society.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. In *SIGKDD Explorartions* (Vol. 11).

Hay, M., Miklau, G., Jensen, D., Weis, P., & Srivastava, S. (2007). Anonymizing social networks. *Computer Science Department Faculty Publication Series*, 180.

Health Information Privacy (HIP). (2014). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule In. Washington DC: U.S. Department of Health and Human Services.

ICO. (2012). Anonymisation: managing data protection risk code of practice. In. Wilmslow: Information Commissioner’s Office (ICO).

Jedrzejczyk, L., Price, B. A., Bandara, A. K., & Nuseibeh, B. (2009). I know what you did last summer: risks of location data leakage in mobile and social computing. *Department of Computing Faculty of Mathematics, Computing and Technology The Open University*, 1744-1986.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M., & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. *Thescientificworldjournal, 2014*, 712826-712826.

Kirby, T. (2014). Controversy surrounds England's new NHS database. *Lancet, 383*, 681-681.

Kitchenham, B. (2004). Procedures for Undertaking Systematic Reviews In: Keele University : Computer Science Department and National ICT Australia Ltd.

Kitchenham, B., & Charters, S. (2007). Guidelines for Performing Systematic Literature Reviews in Software Engineering In: Keele University : School of Computer Science and Mathematics.

Koot, M., van’t Noordende, G., & de Laat, C. (2010). A study on the re-identifiability of Dutch citizens. In *Workshop on Privacy Enhancing Technologies (PET 2010)*.

Lan, S., Yilei, W., & Yingjie, W. (2012). A survey of transaction data anonymous publication. In *Robotics and Applications (ISRA), 2012 IEEE Symposium on* (pp. 239-243).

MacRae, J., Dobbie, S., & Ranchhod, D. (2012). Assessing re-identification risk of de-identified health data in new zealand. *Health Care and Informatics Review Online, 16*, 24-30.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. (cover story). *Harvard Business Review, 90*, 60-68.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *Computer Science Department Faculty Publication Series*, 3.

Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

Ochoa, S., Rasmussen, J., Robson, C., & Salib, M. (2001). Reidentification of individuals in Chicago's homicide database: A technical and legal study. *Massachusetts Institute of Technology*.

Ohm, P. (2010). BROKEN PROMISES OF PRIVACY: RESPONDING TO THE SURPRISING FAILURE OF ANONYMIZATION. *UCLA Law Review, 57*, 1701-1777.

Patterson, T., Bennett, F., & Waine, B. (2008). Transformational government, Benefits. In (Vol. 16, pp. 169-184): The Journal Of Poverty & Social Justice.

Peng, W., Li, F., Zou, X., & Wu, J. (2012). Seed and Grow: An attack against anonymized social networks. In *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012 9th Annual IEEE Communications Society Conference on* (pp. 587-595): IEEE.

Rotella, P. (2012). Is Data the New Oil? In. Online: Forbes.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In: Technical report, SRI International.

SAS Institute Inc. (2015). SAS The Power to Know: Big Data Analytics. In: SAS Institute Inc.

Sharma, S., Gupta, P., & Bhatnagar, V. (2012). Anonymisation in social network: a literature survey and classification. *International Journal of Social Network Mining, 1*, 51.

Shetty, J., & Adibi, J. (2005). Discovering important nodes through graph entropy the case of Enron email database. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 74-81). Chicago, Illinois: ACM.

Sicular, S. (2013). Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. In. Online: Forbes.

Simpson, A. C. (2011). On privacy and public data; a study of data.gov.uk. *Journal of Privacy and Confidentiality, 3*, 51-65.

Smith, M., Szongott, C., Henne, B., & von Voigt, G. (2012). Big data privacy issues in public social media. In *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on* (pp. 1-6): IEEE.

Sramka, M. (2010). A privacy attack that removes the majority of the noise from perturbed data. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1-8).

Sramka, M., Safavi-Naini, R., & Denzinger, J. (2009). *An Attack on the Privacy of Sanitized Data That Fuses the Outputs of Multiple Data Miners*. New York: Ieee.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics, 25*, 98-110.

Sweeney, L. (2000). Uniqueness of Simple Demographics in the U.S. Population. In. Pittsburg; PA: LIDAP-WP4Carnegie Mellon UniversityLaboratory for International Data Privacy.

Sweeney, L. (2011). Patient identifiability in pharmaceutical marketing data. *Harvard University, Cambridge, MA, WP-1015*.

The Cabinet Office. (2005). Transformational government: Enabled by technology In. London: Cm 6683: The Stationary Office.

Walker, P., Meikle, J., & Ramesh, R. (2014). NHS in England delays sharing of medical records. In: The Guardian.

Yankelovich, D., & Meer, D. (2006). REDISCOVERING MARKET SEGMENTATION. *Harvard Business Review, 84*, 122-131.

Appendix A - Definitions

Data

Data may be classed as structured, meaning it can be stored and managed in an organised database, this type of data can also be referred to as relational data (Baxendale & Codd, 1970). Relational data consists mainly of text and numbers (Connolly & Begg, 2005).

Unstructured data on the other hand, consists of large quantities of unorganised data that may contain not just textual information, but also many other types of data. Examples of unstructured data include; social networking data, IP addresses, images and text messages. Unstructured data may change rapidly and for that reason, is sometimes referred to as dynamic data. Dynamic data may be stored and managed in both relational tables and in graph format.

Graphs are arranged into nodes and edges. For instance, in a social networking dataset, the nodes may depict the users and the edges represent their interactions. Thus, for example, graph data may be used to express relationships between users (Sharma, Gupta, & Bhatnagar, 2012).

Attributes

The information contained within all datasets consists of different types of data, also known as attributes. Attributes can be classified into:

> **Identifiers**; i.e. any data that may directly identify an individual e.g. name or national insurance number;
>
> **Key identifiers** (also called quasi-identifiers); i.e. any data from which identifiable information may be inferred e.g. when data is linked (Thomson, Bzdel, Golden-Biddle, Reay, & Estabrooks, 2005).
>
> **Sensitive attributes;** i.e. individual specific information that may assist in re-identification such as salary, ailment or disability status;
>
> **Non-sensitive attributes**; i.e. any other information within the dataset; and
>
> **Graph attributes;** i.e. the nodes, edges and labels of a graph.

Anonymisation

Anonymisation is the process of masking or removing any identifiable information from within a dataset (Thomson, et al., 2005). There are many ways data can be anonymised depending on the data type.

For traditional relational data (i.e. data held in organised tables and databases) the most widely used method of anonymisation is k-anonymisation (Samarati & Sweeney, 1998) whereby any identifying data is suppressed or generalised.

For unstructured data, k-anonymisation alone is not effective and therefore, other anonymisation methods are used to obscure identifiable information. These methods include clustering and graph modification (Sharma, et al., 2012).

Re-Identification

Re-identification occurs when anonymisation is reversed or de-anonymised, bringing the identifying information to light. Re-identification may be achieved in a number of ways including linking datasets, using prior or background knowledge (Abou-el-ela Abdou, Nermin, & Hesham, 2013) or by comparing longitudinal data to find patterns (Tudor, Almgren, & Papatriantafilou, 2013).

References:

Abou-el-ela Abdou, H., Nermin, H., & Hesham, A. H. (2013). Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. *Journal of Information Security*, 101.

Baxendale, P., & Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM, 13*, 377-387.

Connolly, T. M., & Begg, C. E. (2005). *Database systems [electronic resource] : a practical approach to design, implementation, and management / Thomas M. Connolly, Carolyn E. Begg* (4th ed.): Harlow : Addison-Wesley, 2005.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In: Technical report, SRI International.

Sharma, S., Gupta, P., & Bhatnagar, V. (2012). Anonymisation in social network: a literature survey and classification. *International Journal of Social Network Mining, 1*, 51.

Thomson, D., Bzdel, L., Golden-Biddle, K., Reay, T., & Estabrooks, C. A. (2005). Central Questions of Anonymization: A Case Study of Secondary Use of Qualitative Data. *Forum: Qualitative Social Research, 6*, 1-16.

Tudor, V., Almgren, M., & Papatriantafilou, M. (2013). Analysis of the impact of data granularity on privacy for the smart grid. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society* (pp. 61-70). Berlin, Germany: ACM.

Appendix B - Full list of papers included in the Review

Title of paper	Publication year
@ i seek'fb. me': identifying users across multiple online social networks (Jain, Kumaraguru, & Joshi, 2013)	2013
A face is exposed for AOL searcher no. 4417749 (Barbaro & Zeller, 2006)	2006
A machine learning based approach for predicting undisclosed attributes in social networks (Kótyuk & Buttyán, 2012)	2012
A Practical Attack to De-Anonymize Social Network Users (Wondracek, Holz, Kirda, & Kruegel, 2010)	2010
A privacy attack that removes the majority of the noise from perturbed data. (Sramka, 2010)	2010
A practical approach to achieve private medical record linkage in light of public resources (Kuzu, Kantarcioglu, Durham, Toth, & Malin, 2013)	2012
A study on the re-identifiability of Dutch citizens (Koot, van't Noordende, & de Laat, 2010)	2010
Abusing social networks for automated user profiling (Balduzzi, et al., 2010)	2010
All liaisons are dangerous when all your friends are known to us (Gayo Avello, 2011)	2011
An Attack on the Privacy of Sanitized Data that Fuses the Outputs of Multiple Data Miners (Sramka, Safavi-Naini, & Denzinger, 2009)	2009

Anonymizing social networks (Hay, Miklau, Jensen, Weis, & Srivastava, 2007)	2007
Betrayed by my shadow: learning data identity via trail matching (Malin, 2005)	2005
Correlating a Persona to a Person (Clark, 2012)	2012
De-Anonymizing Dynamic Social Networks (Xuan, Lan, Zhiguo, & Ming, 2011)	2011
De-anonymizing social networks (Narayanan & Shmatikov, 2009)	2009
Deanonymizing mobility traces: using social network as a side-channel (Srivatsa & Hicks, 2012)	2012
Discovering important nodes through graph entropy the case of enron email database (Shetty & Adibi, 2005)	2005
Exploiting Innocuous Activity for Correlating Users Across Sites (Goga, et al., 2013)	2013
Exploring Linkability of User Reviews (Almishari & Tsudik, 2012)	2012
Exploring re-identification risks in public domains (Ramachandran, Singh, Porter, & Nagle, 2012)	2012
GlobalInferencer: Linking Personal Social Content with Data on the Web (Paradesi & Shih, 2011)	2011
I Know What You Did Last Summer: risks of location data leakage in mobile and social computing (Jedrzejczyk, Price, Bandara, & Nuseibeh, 2009)	2009
Identifying Users Across Social Tagging Systems (Iofciu, Fankhauser, Abel, & Bischoff, 2011)	2011

Ineluctable background checking on social networks: Linking job seeker's résumé and posts (Okuno, Ichino, Echizen, Utsumi, & Yoshiura, 2013)	2013
Involuntary information leakage in social network services (Lam, Chen, & Chen, 2008)	2008
Is privacy still an issue in the era of big data? — Location disclosure in spatial footprints (L. Li & Goodchild, 2013)	2013
k-anonymity: A model for protecting privacy (Latanya Sweeney, 2002)	2002
Large Online Social Footprints--An Emerging Threat (Irani, Webb, Li, & Pu, 2009)	2009
Link prediction by de-anonymization: How We Won the Kaggle Social Network Challenge (Narayanan, Shi, & Rubinstein, 2011)	2011
Messin'with texas deriving mother's maiden names using public records (Griffith & Jakobsson, 2005)	2006
New threats to health data privacy (F. Li, Zou, Liu, & Chen, 2011)	2011
On Privacy and Public Data: A study of data.gov.uk (Simpson, 2011)	2011
On the anonymizability of graphs (C. Aggarwal, Li, & Yu, 2014)	2014
On the Hardness of Graph Anonymization (C. C. Aggarwal, Li, & Yu, 2011)	2011
Patient Identifiability in Pharmaceutical Marketing Data (Latanya Sweeney, 2011)	2011
Predicting Social Security numbers from public data (Acquisti & Gross, 2009)	2009

Preserving Privacy in Social Networks Against Neighborhood Attacks (Bin & Jian, 2008)	2008
Privacy preservation in social graphs (Zhang, Xu, Bylander, Ruan, & Krishnan, 2012)	2012
Provable De-anonymization of Large Datasets with Sparse Dimensions (Datta, Sharma, & Sinha, 2012)	2012
Re-identification of Smart Meter data (Buchmann, Bohm, Burghardt, & Kessler, 2013)	2013
Reconstructing Profiles from Information Disseminated on the Internet (Aimeur, Brassard, & Molins, 2012)	2012
Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study (Ochoa, Rasmussen, Robson, & Salib, 2001)	2001
Revisiting the uniqueness of simple demographics in the US population (Golle, 2006)	2006
Robust De-anonymization of Large Sparse Datasets (Narayanan & Shmatikov, 2008)	2008
Seed and Grow An attack against anonymized social networks (Peng, Li, Zou, & Wu, 2014)	2012
Sherlock Holmes ’ Evil Twin: On The Impact of Global Inference for Online Privacy (Friedland, Maier, Sommer, & Weaver, 2011)	2011
Stalking online: on user privacy in social networks (Yang, Lutes, Li, Luo, & Liu, 2012)	2012
Structural Attack to Anonymous Graph of Social Networks	2013
The re-identification risk of Canadians from longitudinal	2011

demographics (El Emam, et al., 2011)	
The ultimate invasion of privacy: Identity theft	2011
To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles (Zheleva & Getoor, 2009)	2009
Trail reidentification learning who you are from where you have been (Malin, Sweeney, & Newton, 2003)	2003
Uniqueness of Simple Demographics in the U.S. Population (L. Sweeney, 2000)	2000
Weaving technology and policy together to maintain confidentiality (Latanya Sweeney, 1997)	1997
You are what you say: privacy risks of public mentions (Frankowski, Cosley, Sen, Terveen, & Riedl, 2006)	2006

References:

Acquisti, A., & Gross, R. (2009). Predicting Social Security numbers from public data. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 10975-10980.

Aggarwal, C., Li, Y., & Yu, P. (2014). On the anonymizability of graphs. *Knowledge and Information Systems*, 1-18.

Aggarwal, C. C., Li, Y., & Yu, P. S. (2011). On the hardness of graph anonymization. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on* (pp. 1002-1007): IEEE.

Aimeur, E., Brassard, G., & Molins, P. (2012). Reconstructing Profiles from Information Disseminated on the Internet. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 875-883).

Almishari, M., & Tsudik, G. (2012). Exploring linkability of user reviews. In *Computer Security–ESORICS 2012* (pp. 307-324): Springer.

Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., & Kruegel, C. (2010). Abusing social networks for automated user profiling. In *Recent Advances in Intrusion Detection* (pp. 422-441): Springer.

Barbaro, M., & Zeller, T. J. (2006). A Face Is Exposed for AOL Searcher No. 4417749. In *New York Times*: New York Times, The (NY).

Bin, Z., & Jian, P. (2008). Preserving Privacy in Social Networks Against Neighborhood Attacks. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on* (pp. 506-515).

Buchmann, E., Bohm, K., Burghardt, T., & Kessler, S. (2013). Re-identification of Smart Meter data. *Personal Ubiquitous Comput., 17*, 653-662.

Clark, J. W. (2012). Correlating a persona to a person. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (pp. 851-859): IEEE.

Datta, A., Sharma, D., & Sinha, A. (2012). Provable de-anonymization of large datasets with sparse dimensions. In *Principles of Security and Trust* (pp. 229-248): Springer.

El Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics And Decision Making, 11*, 46-46.

Frankowski, D., Cosley, D., Sen, S., Terveen, L., & Riedl, J. (2006). You are what you say: privacy risks of public mentions. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 565-572). Seattle, Washington, USA: ACM.

Friedland, G., Maier, G., Sommer, R., & Weaver, N. (2011). Sherlock holmes' evil twin: on the impact of global inference for online privacy. In *Proceedings of the 2011 workshop on New security paradigms workshop* (pp. 105-114): ACM.

Gayo Avello, D. (2011). All liaisons are dangerous when all your friends are known to us. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (pp. 171-180): ACM.

Goga, O., Lei, H., Parthasarathi, S. H. K., Friedland, G., Sommer, R., & Teixeira, R. (2013). Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 447-458): International World Wide Web Conferences Steering Committee.

Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society* (pp. 77-80): ACM.

Griffith, V., & Jakobsson, M. (2005). Messin’with texas deriving mother’s maiden names using public records. In *Applied Cryptography and Network Security* (pp. 91-103). Berlin/Heidelberg: Springer.

Hay, M., Miklau, G., Jensen, D., Weis, P., & Srivastava, S. (2007). Anonymizing social networks. *Computer Science Department Faculty Publication Series*, 180.

Iofciu, T., Fankhauser, P., Abel, F., & Bischoff, K. (2011). Identifying Users Across Social Tagging Systems. In *ICWSM*.

Irani, D., Webb, S., Li, K., & Pu, C. (2009). Large online social footprints--an emerging threat. In *Computational Science and Engineering, 2009. CSE'09. International Conference on* (Vol. 3, pp. 271-276): IEEE.

Jain, P., Kumaraguru, P., & Joshi, A. (2013). @ i seek'fb. me': identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 1259-1268): International World Wide Web Conferences Steering Committee.

Jedrzejczyk, L., Price, B. A., Bandara, A. K., & Nuseibeh, B. (2009). I know what you did last summer: risks of location data leakage in mobile and social computing. *Department of Computing Faculty of Mathematics, Computing and Technology The Open University*, 1744-1986.

Koot, M., van’t Noordende, G., & de Laat, C. (2010). A study on the re-identifiability of Dutch citizens. In *Workshop on Privacy Enhancing Technologies (PET 2010)*.

Kótyuk, G., & Buttyán, L. (2012). A machine learning based approach for predicting undisclosed attributes in social networks. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on* (pp. 361-366): IEEE.

Kuzu, M., Kantarcioglu, M., Durham, E. A., Toth, C., & Malin, B. (2013). A practical approach to achieve private medical record linkage in light of public resources. *Journal of the American Medical Informatics Association, 20*, 285-292.

Lam, I.-F., Chen, K.-T., & Chen, L.-J. (2008). Involuntary information leakage in social network services. In *Advances in Information and Computer Security* (pp. 167-183): Springer.

Li, F., Zou, X., Liu, P., & Chen, J. (2011). New threats to health data privacy. *BMC Bioinformatics, 12*, 1-7.

Li, L., & Goodchild, M. F. (2013). Is privacy still an issue in the era of big data?; Location disclosure in spatial footprints. In *Geoinformatics (GEOINFORMATICS), 2013 21st International Conference on* (pp. 1-4).

Malin, B. (2005). Betrayed by my shadow: learning data identity via trail matching. *Journal of Privacy Technology, 2147483647*.

Malin, B., Sweeney, L., & Newton, E. (2003). Trail re-identification: learning who you are from where you have been. *Proc. LIDAP-WP12*.

Narayanan, A., Shi, E., & Rubinstein, B. I. (2011). Link prediction by de-anonymization: How We Won the Kaggle Social Network Challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on* (pp. 1825-1834).

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (pp. 111-125).

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing Social Networks. In *Security and Privacy, 2009 30th IEEE Symposium on* (pp. 173-187).

Ochoa, S., Rasmussen, J., Robson, C., & Salib, M. (2001). Reidentification of individuals in Chicago's homicide database: A technical and legal study. *Massachusetts Institute of Technology*.

Okuno, T., Ichino, M., Echizen, I., Utsumi, A., & Yoshiura, H. (2013). Ineluctable Background Checking on Social Networks Linking Job Seekers Resume and Posts. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on* (pp. 273-278).

Paradesi, S., & Shih, F. (2011). GlobalInferencer: linking personal social content with data on the web. In *ICWSM-11 Workshop on The Future of Social Web*.

Peng, W., Li, F., Zou, X. K., & Wu, J. (2014). A Two-Stage Deanonymization Attack against Anonymized Social Networks. *Ieee Transactions on Computers, 63*, 290-303.

Ramachandran, A., Singh, L., Porter, E., & Nagle, F. (2012). Exploring re-identification risks in public domains. In *Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference on* (pp. 35-42).

Shetty, J., & Adibi, J. (2005). Discovering important nodes through graph entropy the case of Enron email database. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 74-81). Chicago, Illinois: ACM.

Simpson, A. C. (2011). On privacy and public data; a study of data.gov.uk. *Journal of Privacy and Confidentiality, 3*, 51-65.

Sramka, M. (2010). A privacy attack that removes the majority of the noise from perturbed data. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (pp. 1-8).

Sramka, M., Safavi-Naini, R., & Denzinger, J. (2009). *An Attack on the Privacy of Sanitized Data That Fuses the Outputs of Multiple Data Miners*. New York: Ieee.

Srivatsa, M., & Hicks, M. (2012). Deanonymizing mobility traces: Using social network as a side-channel. In *Proceedings of the 2012 ACM conference on Computer and communications security* (pp. 628-637): ACM.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics, 25*, 98-110.

Sweeney, L. (2000). Uniqueness of Simple Demographics in the U.S. Population. In. Pittsburg; PA: LIDAP-WP4Carnegie Mellon UniversityLaboratory for International Data Privacy.

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

Sweeney, L. (2011). Patient identifiability in pharmaceutical marketing data. *Harvard University, Cambridge, MA, WP-1015*.

Wondracek, G., Holz, T., Kirda, E., & Kruegel, C. (2010). A practical attack to de-anonymize social network users. In *Security and Privacy (SP), 2010 IEEE Symposium on* (pp. 223-238): IEEE.

Xuan, D., Lan, Z., Zhiguo, W., & Ming, G. (2011). De-Anonymizing Dynamic Social Networks. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE* (pp. 1-6).

Yang, Y., Lutes, J., Li, F., Luo, B., & Liu, P. (2012). Stalking online: on user privacy in social networks. In *Proceedings of the second ACM conference on Data and Application Security and Privacy* (pp. 37-48). San Antonio, Texas, USA: ACM.

Zhang, L., Xu, S., Bylander, T., Ruan, J., & Krishnan, R. (2012). *Privacy preservation in social graphs (Doctoral dissertation).* THE UNIVERSITY OF TEXAS AT SAN ANTONIO.

Zheleva, E., & Getoor, L. (2009). To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In

Proceedings of the 18th international conference on World wide web (pp. 531-540). Madrid, Spain: ACM.

Dear Sirs,

Thank you for your email of 14th March, providing very helpful and constructive feedback on our submission.

Having reviewed the feedback, the following amendments have been made:

- Some of the components of the Protocol have now been incorporated into the main body of the manuscript and The Protocol in Appendix A has been removed.
- The definitions in Section 2 have been moved to the glossary. These can now be found in Appendix A.
- The bullets have been removed with the data now presented in Tables instead.
- The typographical errors mentioned have been corrected.

The revised manuscript has been uploaded for consideration.

Should you have any queries or require any further amendments, please do not hesitate to contact us.