International Journal of Information Management 36 (2016) 1184–1192

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Re-identification attacks—A systematic literature review

Jane Henriksen-Bulmer*, Sheridan Jeary (Dr)

Faculty of Science and Technology, Bournemouth University, Talbot Campus, Fern Barrow, Bournemouth, Dorset, United Kingdom

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form 23 March 2016
Accepted 2 August 2016

Keywords:
Re-identification
Anonymisation
Anonymization
Systematic literature review

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.

1. 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).

* Corresponding author.
E-mail addresses: Jhenriksenbulmer@bournemouth.ac.uk (J. Henriksen-Bulmer), sjeary@bournemouth.ac.uk (S. Jeary).

http://dx.doi.org/10.1016/j.ijinfomgt.2016.08.002
0268-4012/

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. 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.