Deanonymizing the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) dataset

Chris Culnane s47F
Ben Rubinstein s47F,
and Vanessa Teague s47F

University of Melbourne

September 8, 2016

Abstract

This paper describes a problem in the encryption of Serivce provider ID numbers in the Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Schedule (PBS) dataset recently published at `http://data.gov.au/dataset/mbs-sample-10pct-1984-gz`. We show here that the anonymization is easily reversed, at least for providers. Using only publicly available information, we have decrypted all the service provider ID numbers. We did not decrypt Medicare Patient ID numbers, but online documentation suggests that they may be encrypted in the same way.

The data set does not include names or addresses, but the implications of exposing individual IDs are still serious.

1 Introduction

We have had some early discussions with the Australian Bureau of Statistics about a potential project relating to privacy-preserving secure data linking. In addition to trying to design improved techniques, the questions included investigating the security of current methods.

We find that the method used in the MBS 10% dataset to encrypt provider IDs is not secure. Documentation suggests that the same method is used to encrypt Medicare patient ID numbers, but we have not decrypted them and are not certain about their security.

1.1 Background on the data set

The MBS 10% sample was constructed by choosing a random 10% of Medicare patient ID numbers (PINs). For each selected PIN, (almost) all medicare claiming activities were included in the data. Most of the data are not encrypted, including Medicare item codes for tests and doctors' visits, pharmaceutical benefits (PBS) item codes, and other information such as pricing and location. The codes describe *what* medical activity occurred, for example a particular kind of test, a particular prescription, or a visit to a GP or specialist. The ID numbers describing *who* received or provided the service are encrypted.

The data set does not include names or addresses of patients or doctors.

The website states that "A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be re-identified." Strictly speaking, the method of obscuring patient and provider ID numbers is not encryption, though we will use that term to refer to it for consistency.

In summary, the data set contains medicare data treated in three different ways:

1. some data, such as Medicare and PBS item codes, are recorded in the clear;
2. some data, such as the date of the item, are randomly perturbed,
3. individual patient and provider IDs are encrypted.

1.2 Background on deanonymizing data

Generally there are two main ways of identifying individuals in this kind of data set:

linkage attacks use the unencrypted data to identify people by linking the record with other known information;

cryptographic attacks work by reversing the encryption algorithm to recover encrypted data.

This disclosure concentrates on successful cryptographic attacks. Linkage attacks have been demonstrated on other datasets [Swe02, NS08, NSR11] but we have not attempted them here.

The next section explains why the encryption of provider IDs is not secure.

2 How the encryption of individual IDs can be reversed

s47E(d)

s47E(d)

3 Conclusion

Publishing this sort of data brings great benefits for research but also great risks to privacy. The technical details matter: it's a specialist task to understand whether a particular algorithm securely encrypts data or not.

Explaining the anonymization algorithm publicly at `data.gov.au` was the right thing to do. Keeping the details of the algorithm secret would not have made it secure. "Security through obscurity doesn't work" is widely accepted among cryptographers. Omitting some of the algorithmic details didn't keep the data secure—if we can reverse-engineer the details in a few days, then there is a risk that others could do so too. However, keeping it secret would have made it more difficult for us to identify the problem, without making the encryption scheme any more inherently secure.

The best approach would be to publish the anonymization and encryption algorithms long before the data set, so that errors or vulnerabilities could be identified and corrected before the data were released.

The Australian government is making important decisions about what personal data to publish and how it should be anonymized, encrypted, or linked. Making good decisions requires accurate technical information about the security of the system and the secrecy of the data. We should have an open public discussion about what to publish and link, and how to keep it secure.

References

[Aus16] Australian Government Department of Health. Mbs 10% sample technical information, 2016. `http://data.gov.au/dataset/a8e3c0bc-44ac-4e9a-8b3c-b779438ddb10/resource/a33103be-17d8-4857-b621-334a813f4b99/download/`

mbs-10-sample-metadata-part-5-of-5---technical-information.pdf.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[NSR11] Arvind Narayanan, Elaine Shi, and Benjamin IP Rubinstein. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1825–1834. IEEE, 2011.

[Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Deanonymizing Provider IDs in the Medicare Benefits Schedule (MBS) 10% Dataset

Chris Culnane Benjamin Rubinstein Vanessa Teague

s47F

Department of Computing and Information Systems
University of Melbourne

September 17, 2016

1 Introduction

This article describes a problem in the encryption of Service provider ID numbers in the Medicare Benefits Schedule (MBS) dataset recently published at `http://data.gov.au`. Using only publicly available information, we found that it was possible to decrypt the service provider ID numbers. We did not decrypt Patient ID numbers.

The dataset does not include names or addresses, but the implications of exposing individual IDs are still serious.

Once notified, the Department of Health immediately removed the dataset from the website, opened lines of communication with us to further understand the issue, and conducted their own careful investigation of the problem.

1.1 Background on the dataset

The MBS 10% sample was constructed by choosing a random 10% of patient identifier numbers (PINs). For each selected PIN, (almost) all Medicare claiming activities were included in the data. A separate dataset lists Pharmaceutical Benefits Schedule (PBS) claims for the same patients. Most of the data are not encrypted, including Medicare item codes for tests and doctors' visits, pharmaceutical benefits (PBS) item codes, and other information such as pricing and location. The codes describe *what* medical activity occurred, for example a particular kind of test, a particular prescription, or a visit to a GP or specialist. The ID numbers describing *who* received or provided the service are encrypted.

The dataset does not include names or addresses of patients or doctors—the MBS dataset includes an encrypted PIN and provider ID for each activity. Note that the patient's PIN is not the number written on their Medicare card, and that the PBS dataset does not include the prescribing doctor's ID.

The website states that "A suite of confidentiality measures including encryption, perturbation and exclusion of rare events has been applied to safeguard personal health information and ensure that patients and providers cannot be re-identified." Strictly speaking, the method of obscuring patient and provider ID numbers is not encryption, though we will use that term to refer to it for consistency.

In summary, the data were treated in four different ways:

1. some rare events were removed;
2. some data, such as Medicare and PBS item codes, are recorded without encryption;
3. some data, such as the date of the item, are randomly perturbed; and
4. individual patient and provider IDs are encrypted.

1.2 Background on deanonymizing data

There are two main ways of trying to identify individuals in this kind of dataset:

Linkage attacks use the unencrypted data to identify people by linking the record with other known information; and

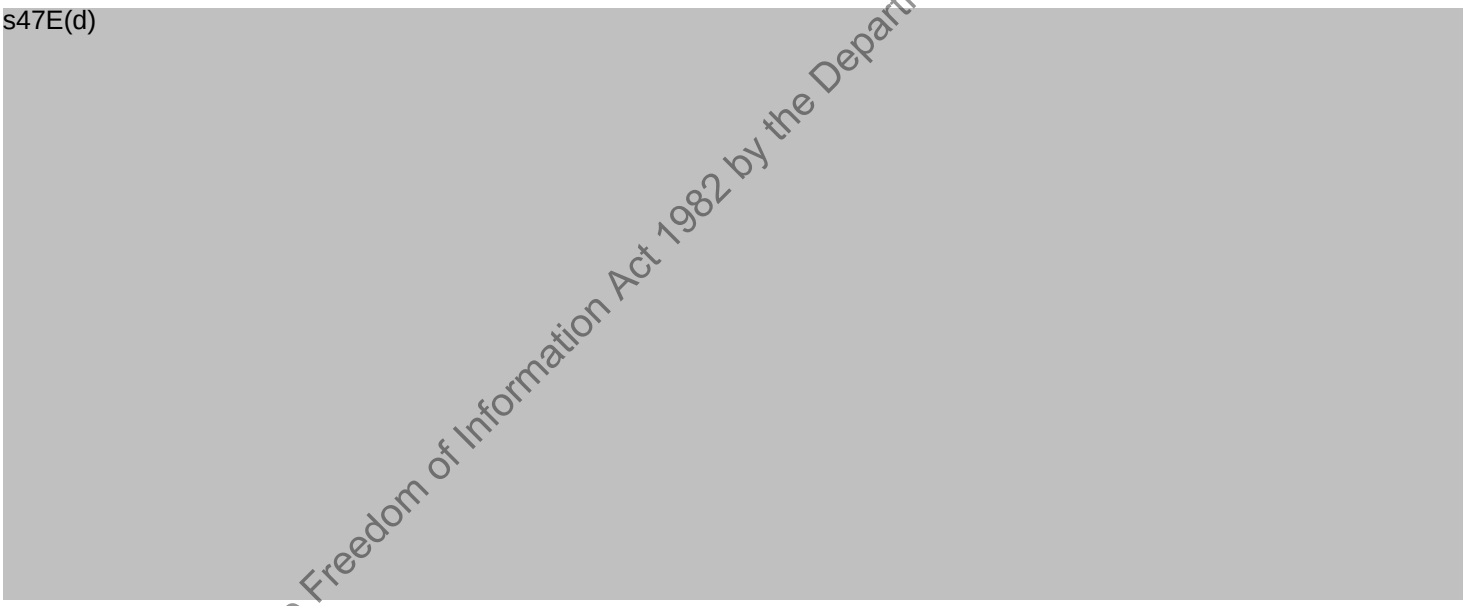
Cryptographic attacks reverse the encryption algorithm to recover encrypted data.

This article concentrates on successful cryptographic attacks. Linkage attacks have been demonstrated on other datasets [Swe02, NS08, NSR11, DMHVB13] but we have not attempted them here.

The next section explains why the encryption of provider IDs is not secure.

2 How the encryption of individual IDs can be reversed

s47E(d)



3 Conclusion

Publishing data can bring great benefits to research but also great risks to privacy. The technical details matter: it's a technically challenging task to understand whether a particular algorithm securely encrypts data or not. Datasets containing sensitive information about individuals clearly deserve more caution, and may not always be suitable for open public release.

The Australian government's open data program provides numerous benefits, allowing better decisions to be made based on evidence, careful analysis, and widespread access to accurate information. Decisions about data publication itself should follow the same philosophy.

The best approach would be to publish technical details about privacy protection long before the dataset, so that errors or vulnerabilities in the proposed scheme could be identified and corrected before the data were released. The privacy protections themselves should be detailed in public in advance. They can then be subject to empirical testing, scientific analysis, and open public review, before they are used on real data. Then we can make sound, evidence-based decisions about how to benefit from open data without sacrificing individual privacy.

References

[Aus16] Australian Government Department of Health. MBS 10% sample technical information, June 2016. `http://data.gov.au/dataset/a8e3c0bc-44ac-4e9a-8b3c-b779438ddb10/resource/a33103be-17d8-4857-b621-334a813f4b99/download/mbs-10-sample-metadata-part-5-of-5---technical-information.pdf`.

[DMHVB13] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013. `http://www.nature.com/articles/srep01376`.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (SP 2008)*, pages 111–125. IEEE, 2008.

[NSR11] Arvind Narayanan, Elaine Shi, and Benjamin IP Rubinstein. Link prediction by de-anonymization: How we won the Kaggle social network challenge. In *The 2011 International Joint Conference on Neural Networks (IJCNN)*, pages 1825–1834. IEEE, 2011.

[Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.