Grouping methods for ongoing record linkage

Sean M. Randall
Centre for Data Linkage
Curtin University
Perth, Australia
sean.randall@curtin.edu.au

James H. Boyd
Centre for Data Linkage
Curtin University
Perth, Australia
j.boyd@curtin.edu.au

Anna M. Ferrante
Centre for Data Linkage
Curtin University
Perth, Australia
a.ferrante@curtin.edu.au

Adrian P. Brown
Centre for Data Linkage
Curtin University
Perth, Australia
adrian.brown@curtin.edu.au

James B. Semmens
Centre for Population Health Research
Curtin University
Perth, Australia
james.semmens@curtin.edu.au

ABSTRACT

The grouping of record-pairs to determine which records belong to the same individual is an important part of the record linkage process. While a *merge* grouping approach is commonly used, other methods may be more appropriate when linking to a repository of previously linked data.

In this paper, we applied a number of grouping strategies to three large scale hospital datasets (comprising around 27 million records), each with a known truth set. These datasets were linked against a created 'repository' whose quality was varied.

Experimental results show that alternate grouping methods can yield very large benefits in linkage quality, especially when the quality of the underlying repository is high. *Best link* methods can remove between 25-90% of matching errors, depending on the characteristics of the underlying datasets.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Record linkage, grouping

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10 August 2015.

1. INTRODUCTION

Widely utilised in health research, record linkage involves identifying records which belong to the same individual within and across administrative datasets. By linking together records from hospital and emergency collections, primary care facilities, and birth, death and disease registries, researchers can construct a chronological sequence of events for a particular individual. The linkage process provides researchers with an enriched, cost effective, longitudinal research dataset for the study of entire populations.

In the absence of a unique identifier, linkage involves matching records using personal identifiers (e.g. name, address, and date of birth). As this information changes, and/or can be in error, statistical techniques are used to ensure links of the highest quality [4]. Ensuring high quality is critical in record linkage, as research outcomes can be affected. Current methods used to maintain linkage quality [15, 3] are heavily manual which is both costly and time-consuming. Identifying methods to improve quality that do not rely on manual review is of high interest [12].

Specialised linkage units often provide the infrastructure and expertise required to carry out record linkage. These units carry out linkage on an on-going basis, creating a list of all records and the person identifier to whom they belong. Incoming datasets are linked to the repository which is updated with this new information.

During the linkage process, incoming data is first cleaned to ensure consistency and reliability. The files are then matched using a defined linkage strategy, resulting in pairs of records designated as belonging to the same person. A grouping or clustering process then amalgamates these record-pairs into groups to identify the full set of records belonging to the same individual.

The traditional grouping process uses transitive closure to merge all identified record-pairs, with all connected records being assigned to the same individual. Transitive or *indirect* links are formed where records which did not form a pair relationship nonetheless are assigned to the same individual, for instance because they form record-pairs with a third record.

The merge based grouping process treats the repository as simply another set of records. However there is reason to believe that existing groups of records within the repository should rarely be merged together by incoming records - these groups have already been validated and are unlikely to be in error.

2. OBJECTIVES

We hypothesise that the use of grouping methods which reduce or remove the opportunity for groups within a repository to be joined together should result in higher linkage quality than the traditional merge based method. One such method has been suggested previously [9]; however this method (*best link* grouping) has never been evaluated against the traditional merge approach used in many operational linkage units across the world.

In this paper, we present an alternate best-link algorithm for grouping, and evaluate this algorithm against both the merge based and best link algorithms using real world datasets. We hypothesise that the appropriateness of these grouping techniques for on-going linkage will depend on the overall quality of the repository used. To test this, repositories of differing quality were used in the evaluation to allow us to determine the circumstances in which particular methods are appropriate.

3. METHODS

3.1 Grouping Methods

3.1.1 Merge Based Grouping

Merge grouping amalgamates all record pairs above the accepted threshold, with all connected records belonging to the same individual. Indirect or transitive links are formed where records which did not form a pair relationship nonetheless are marked as belonging to the same individual, for instance because they are both linked to a third record. If multiple groups in the repository are linked together in this way, these are merged. There is no limit to the length of indirect links accepted, although this can be used as a potential indicator of groups containing errors [12].

3.1.2 Best Link

In the approach presented by Kendrick [9], grouping is carried out in the order in which the records are matched. Each record from the incoming file is matched in turn against records in the repository. If the record from the incoming file matches to multiple records in the repository file, only the highest weighted match is accepted, and the record from the incoming file is added to this group. If the record does not link to any records in the repository, a new group is created, of which it is the sole member. The incoming record is then added to the repository, and subsequent records in the incoming file are able to match against this added record.

3.1.3 Weighted Best Link

Our modified grouping strategy which we will refer to as weighted best link, involves a linkage of records from the incoming file to the repository (along with a de-duplication of the incoming file) where all record pairs are created and evaluated. Once the linkage is completed, accepted record pairs are amalgamated in weight order. The pairs are examined in order from highest to lowest; a record-pair is accepted as valid provided it does not result in multiple groups from the repository merging together.

Algorithm 1 Best link

```
Input: Incoming file, Repository
for each record in Incoming File do
    link record to Repository
    if there is one pair found then
        add record to that group
    else if there are multiple pairs found then
        choose the highest pair
        add record to that group
    else if there are no pairs found then
        mark record as belonging to a new group
    add record to Repository
```

Algorithm 2 Weighted best link

```
   Input: Incoming file, Repository
1: Link Incoming file to Repository
2: Deduplicate Incoming file
3: Concatenate pairs from (1) and (2)
4: Sort output of (3) in weight descending order
5: for each pair in sorted pairs do
6:     if accepting will merge two repository groups then
7:         ignore pair
8:     else
9:         accept pair
```

Both best link methods assume that record-pairs have some ordinal attribute which identifies how likely they are to belong to the same individual. In probabilistic linkage, this is the weight attached to each record-pair [11]. For deterministic linkage (another common method of record linkage), these grouping strategies can be used by ordering rules by strictness.

Both best-link algorithms are similar, and in many situations return the same results. An example of their difference is shown in Figure 1. Using the best link approach, the first record A is matched to record Z and joins this group. The second record B matches to both A and Y. Of these, A is the highest weighted, so record B will join the same group as A and Z. In the weighted best link method, the first accepted pair is that joining the incoming records A and B. The next pair joins B and Y; A, B and Y are now linked together. The final pair linking A to Z is ignored, as this would bring together two groups from the repository.

The advantage of the modified weighted best link methods is that it will consistently produce the same results irrespective of the order of records being processed. The best link method described by Kendrick [9] will produce different grouping results if the linkage of the incoming records is executed in a different order.

3.2 Evaluation Datasets

Three large hospital admissions datasets were used in this evaluation, for which we had pre-existing and accurate information about which records belonged to the same person.

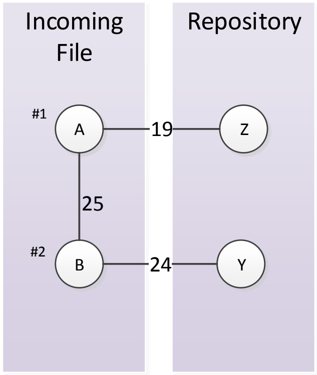

Figure 1: An example of the difference between best link algorithms. The number between records represents the weight of the record-pair comparison.

This information acted as the 'truth set' for each dataset and was used to compute differences in the performance of the three grouping algorithms. Ten years of Western Australian (WA) Hospital Admissions data, along with ten years of New South Wales (NSW) Admitted Patient Data and eight years of South Australian (SA) Hospital Admissions data were used in the evaluation. These datasets contained the typical data quality errors found in administrative data, including misspellings, name variations, missing data, changes in personal identifiers and incorrect values. Each dataset had been previously de-duplicated (by the WA Data Linkage Branch [8], the Centre for Health Record Linkage [10], and SA-NT DataLink respectively) utilising a variety of methods including exact matching, probabilistic linkage and intensive clerical review. All the linkage units employ rigorous manual reviews of created links, and a quality assurance program to analyse and review likely errors [3, 15] These links are further validated through use in a large number of research projects and published research articles [2]. Both WA and NSW have been operational for many years while in comparison SA data has only recently been linked, and has therefore been subject to less review by both clerical assessors and researchers. The data was made available as part of the Population Health Research Network Proof of Concept project [1]. A summary of the datasets is provided in Table 1.

3.3 Matching Strategy

A single matching strategy was used for all linkages in the study. This strategy utilised a probabilistic approach and was based on a previously published 'default' linkage strategy [7]. Two sets of blocks were used: Soundex of surname with first initial, and full date of birth. All variables were used in comparisons; string similarity measures were used for alphabetic variables (name, address and suburb) with exact matches used for all other variables. Agreement and disagreement weights were estimated.

3.4 Measuring Linkage Quality

Linkage quality was evaluated using saturated pairwise precision, recall and f-measure. Precision refers to the proportion of found links that were correct, and thus provides a measure of false positives. Recall is the proportion of all correct links found, and thus measures false negatives. The F-measure is the harmonic mean between precision and recall, giving a single figure from which we can compare results. These measures have been recommended for use in record linkage [5].

3.5 Repository Creation

To simulate linkage of an incoming file to a central repository, it was necessary to create *repositories* (datasets with coverage of close to the whole population). A repository for each of the original data sources was created by first randomly selecting one record per person from the hospital admissions file. This repository was 'complete' in the sense that it had coverage of the whole population being linked, and did not contain records for the same individual in more than one group.

Additional repositories of degraded quality were created by both removing records from the 'complete' repository, and by adding additional records belonging to a person already in the repository, as a separate person. Additional 'duplicate' records were specifically chosen so that differences existed in the personal identifiers between the records in the repository belonging to the same person.

Four repositories in total were created from each original dataset, differing in the number of errors they contained. These included a 'complete' repository, a repository with 1% of records missing and 1% of groups duplicated, a repository with 2.5% records missing and 2.5% groups duplicated, and a repository with 5% records missing and 5% of groups duplicated.

3.6 Evaluation Strategy

The linkage of the three datasets to their corresponding repositories was conducted separately; there was no linkage between hospital datasets.

'Incoming files' for linkage were constructed by breaking the hospital admissions records into batches containing admission records for a three month period. The batches were then linked to the repository in temporal order, to simulate on-going linkage. Records that were used to create repositories did not form part of the incoming files.

Each linkage of a batch of incoming records to the corresponding repository was grouped using three different methods - the traditional merge based method, best link and the new weighted best link approach.

Linkages were conducted using four different repositories, with three different grouping strategies, on the three state-based datasets, for a total of 36 linkage runs. The quality of each run was measured using the metrics described above.

4. RESULTS

The optimal F-measures of the overall linkage (after all batches were added) for each linkage run are shown in Figure 2. The figure displays the maximum F-measure achieved across a range of possible threshold settings.

Table 1: Dataset characteristics

Missing Values	NSW Morbidity	WA Morbidity	SA Morbidity
Surname	31.9%	<0.1%	5.3%
Given Names	33.9%	<1.0%	5.5%
Sex	<0.1%	<0.1%	<0.1%
DOB	<0.1%	<0.1%	0
Suburb	<1.0%	<1.0%	6.9%
Address	7.5%	<0.1%	8.1%
Postcode	<1.0%	<1.0%	8.5%
N	19,874,083 records	6,772,949 records	2,509,914 records

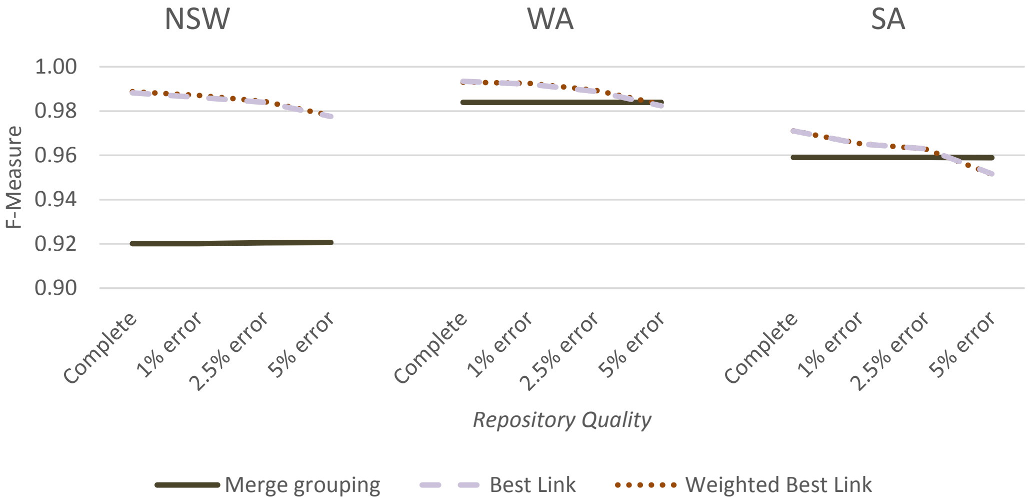

Figure 2: Results of grouping by repository quality

As can be seen, the effectiveness of merge-based grouping as compared with best link methods depended heavily on both the dataset used and the quality of the repository. For all datasets, the best link methods were superior when using a repository with an error rate of 2.5% or less. For an error rate of 5%, the most effective grouping strategy varied with the dataset.

Merge based grouping was not affected by repository quality, whereas the linkage quality of the best link methods decreased as the quality of the repository was degraded. This is unsurprising, as merge based grouping accepts all record-pairs above a certain threshold, without regard for the constitution of the repository, whereas best link methods will specifically reject certain record-pairs above the threshold based on records found in the repository.

Little difference was observed in the maximum F-measure between the two best link methods. This was a consistent finding across all datasets and all levels of repository quality.

Figure 3 shows the overall F-measure for each threshold value, for all grouping methods and for all repositories; displayed threshold are those found through probabilistic record linkage using the method of Fellegi-Sunter [6]. For higher valued thresholds, there was no difference between the merge based strategy and either of the best link strategies; however, for lower chosen thresholds the F-measures diverged, with merge based grouping scores rapidly decreasing, while best link scores improved.

As the threshold decreases, the number of false-positive pairs increase. The merge grouping method includes these false-positive pairs, resulting in lower linkage quality. Best link methods only accept these false-positives pairs if the incoming record has not already linked to a record in the repository. As this is nearly always the case, the vast majority of these false-positives are ignored, and so linkage quality remains relatively unchanged. For higher thresholds where there are fewer false-positives, there are smaller differences between these approaches.

A final notable difference is the much greater threshold range over which the F-measure for best link grouping is at a maximum.

5. DISCUSSION

The results of this study show that when optimising for linkage quality, the most appropriate grouping strategy depends on the underlying quality of the repository. If the repository is not representative of the study population or of poor quality with little confidence in the established groups, the merge based method can be considered as a possible grouping strategy. However, for better quality repositories, best link methods result in much higher linkage quality. It would be expected that most data repositories, or well-maintained

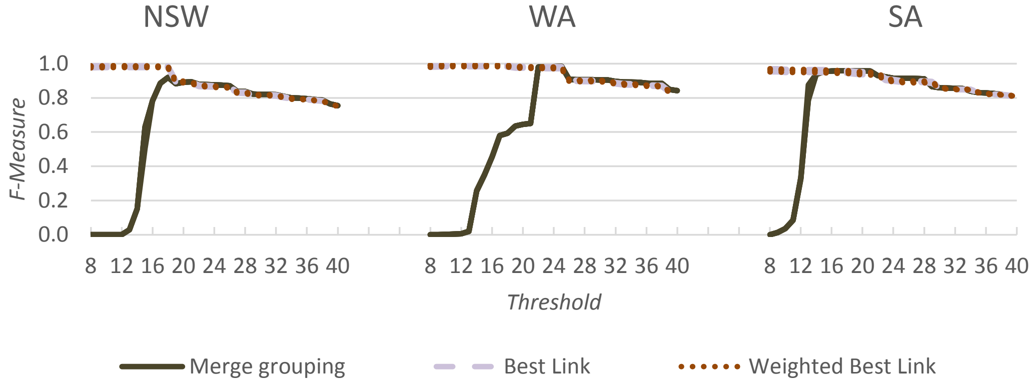

Figure 3: Results of grouping by threshold score

datasets with high population coverage, would contain only a small level of error, making best link the most appropriate grouping strategy to adopt. As the results indicate, best link methods have the added advantage of being insensitive to threshold changes. This increased tolerance reduces the likelihood of threshold estimation errors and suggests that these grouping methods could be useful in situations where determining thresholds is difficult, such as in privacy preserving linkage [13].

Our results were also highly dataset dependent, with best link methods proving superior on NSW data for all repositories. This is likely to be a reflection of the lower data quality (the NSW data has much higher rates of missing values; see Table 1).

Results showed little difference between the two best link methods. Factors other than linkage quality may be more appropriate in determining which of these methods should be used in ongoing linkage. The weighted best link method has the advantage that results are repeatable and not dependent on the order of incoming records. This means that it is possible to retrace and understand the sequence of links that were created over time without knowing the order in which records arrived. The weighted method also has the advantage that grouping decisions are made independently of matching decisions. This de-coupling of processes may be important in the design and development of linkage systems.

Given the dataset-specific nature of the results from this study, additional testing against other datasets may be required to gain a full understanding of the relationship between linkage quality, grouping strategy and population repository quality.

Our results show that the choice of grouping strategy can make a large difference to linkage quality. Within this evaluation, best link methods were able to remove between 25% (SA) to 90% (NSW) of matching errors using a high quality repository. This is an extremely large improvement in linkage accuracy, yielding far larger gains than other techniques in the literature [14, 12].

6. CONCLUSION

The effect of grouping methods on linkage quality is an understudied area of research. By adopting an appropriate grouping strategy, vast improvements in linkage quality can be achieved. The weighted best link strategy presented here shows large improvements against the merge strategy currently in operation, while providing practical benefits over the previous best link method.

Current methods of improving quality present as processing bottlenecks. Methods which improve the overall quality of linked data without impacting on performance will ultimately lead to more accurate and reliable research outcomes and increased utilisation of this resource by researchers.

7. ACKNOWLEDGMENTS

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

8. REFERENCES

[1] J. H. Boyd, A. M. Ferrante, C. M. O'Keefe, A. J. Bass, S. M. Randall, and J. B. Semmens. Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC health services research*, 12(1):480, 2012.

[2] E. L. Brook, D. L. Rosman, and C. J. Holman. Public good through data linkage: measuring research outputs from the western australian data linkage system. *Australian and New Zealand journal of public health*, 32(1):19–23, 2008.

[3] Centre for Health Record Linkage. Quality assurance, 2015. [Online; http://www.cherel.org.au/quality-assurance; accessed 3-June-2015].

[4] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[5] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[6] I. P. Fellegi and A. B. Sunter. A theory for record

linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[7] A. Ferrante and J. Boyd. A transparent and transportable methodology for evaluating data linkage software. *Journal of biomedical informatics*, 45(1):165–172, 2012.

[8] C. D. J. Holman, J. A. Bass, D. L. Rosman, M. B. Smith, J. B. Semmens, E. J. Glasson, E. L. Brook, B. Trutwein, I. L. Rouse, C. R. Watson, et al. A decade of data linkage in western australia: strategic design, applications and benefits of the wa data linkage system. *Australian Health Review*, 32(4):766–777, 2008.

[9] S. Kendrick, M. Douglas, D. Gardner, and D. Hucker. Best-link matching of scottish health data sets. *Methods of information in medicine*, 37(1):64–68, 1998.

[10] G. Lawrence, I. Dinh, L. Taylor, et al. The Centre for Health Record Linkage: a new resource for health services research and evaluation. *Health Information Management Journal*, 37(2):60, 2008.

[11] H. B. Newcombe. *Handbook of record linkage: methods for health and statistical studies, administration, and business.* Oxford University Press, Inc., 1988.

[12] S. M. Randall, J. H. Boyd, A. M. Ferrante, J. K. Bauer, and J. B. Semmens. Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine*, 115(2):55–63, 2014.

[13] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics*, 50:205–212, 2014.

[14] S. M. Randall, A. M. Ferrante, J. H. Boyd, and J. B. Semmens. The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, 13(1):64, 2013.

[15] D. Rosman, C. Garfield, S. Fuller, A. Stoney, T. Owen, and G. Gawthorne. Measuring data and link quality in a dynamic multi-set linkage system. In *Proceedings of the Symposium on Health Data Linkage*, 2002.