Use of Graph Theory Measures to Identify Errors in Record Linkage

Sean M. Randall[1]*,

James H. Boyd[1],

Anna M. Ferrante[1],

Jacqueline K. Bauer[1],

James B. Semmens[1]

[1] *Affiliation: Curtin University, Western Australia.*

Address: Room 237 Building 400, Centre for Data Linkage, Curtin University, WA, 6102, Australia

**Corresponding Author:*

Phone number: +061 415 065 464

Authors Emails:

sean.randall@curtin.edu.au

J.Boyd@curtin.edu.au

A.Ferrante@curtin.edu.au

Jacqui.Bauer@curtin.edu.au

James.Semmens@curtin.edu.au

Keywords: Record Linkage, Graph Theory, Data Quality

Word Count: 4458

ABSTRACT

Ensuring high linkage quality is important in many record linkage applications. Current methods for ensuring quality are manual and resource intensive. This paper seeks to determine the effectiveness of graph theory techniques in identifying record linkage errors. A range of graph theory techniques were applied to two linked datasets, with known truth sets. The ability of graph theory techniques to identify groups containing errors was compared to a widely used threshold setting technique. This methodology shows promise; however, with further investigations into graph theory techniques are required. The development of more efficient and effective methods of improving linkage quality will result in higher quality datasets that can be delivered to researchers in shorter timeframes.

1. BACKGROUND AND SIGNIFICANCE

1.1. Entity Resolution

The existence of attributes within or between datasets referring to the same real-world entity is a common feature in information systems. The task of joining together different manifestations of the same underlying entity is known as *entity resolution*. One example of entity resolution is the linking of several bibliographic reference lists – this could involve identifying the individual authors being referred to in each reference, as well as the article and the journal. This process is made more complex by the existence of data quality issues. Information may not be unique (different entities may have the same attributes), it may be recorded incorrectly due to transcription errors, different recording standards make exist across or within data collections, and attributes can change over time.

1.2. Record Linkage

One form of entity resolution is record linkage, where every record (corresponding to a document or a row in a database) belongs to a single individual and the task is then to identify which records belong to the same individual. When a unique person-based identifier exists, this can be achieved by simply linking the datasets on this identifier. When this identifier does not exist, linkage is often performed using statistical techniques that compare personally identifying information such as name and address, which may include typographical errors or other variations.

Record linkage techniques applied in business contexts including the removal of duplicate entries from databases of customer accounts, or the linking together of separately maintained lists. Record linkage of administrative government datasets has also been widely used to enable research across sectors, maximising the value of these data collections. In the health sector, record linkage is widely used to enable health researchers to gain event based longitudinal information for entire populations, linking together records from multiple collections (such as hospital admissions, emergency presentations, specific health registers, and births and deaths registers). Record linkage research has been very successful in affecting positive change to health policy and health service delivery [1, 2]. This success had led to significant investment in large-scale record linkage infrastructure in England [3], Scotland [4], Wales [5], Canada [6] and Australia [7].

1.3. Record Linkage Methods

Approaches to record linkage range from simple rule based methods, to sophisticated probability-based statistical techniques. These techniques all involve the comparison of pairs of records, which are ultimately designated either as a match (the two records in question belong to the same person) or a non-match (the two records do not belong to the same person). During the record linkage

process pairs formed are typically given a weight, corresponding to how likely that pair is to be correct.

The matching process is immediately followed by a 'grouping' process, where the record-pair relationships are amalgamated to determine person 'groups'. In a record linkage context, this grouping process nearly always uses the partition method [8]. As represented in Figure 1, this method is equivalent to 'joining' pair relationships together, with all records which are connected belonging to the same person.

1.4. Record Linkage Quality

Two types of errors impact on matching quality - incorrect matches (or false positives), where two records are designated to belong to the same person when they should not, and missed matches (or false negatives) where two records are not designated as belonging to the same person when, in fact, they should. The rate of these two types of errors, measured through precision (or positive predictive value) and recall (sensitivity) statistics, determines overall linkage quality.

While the basic linkage processes are relatively easy to implement, ensuring high linkage quality is difficult and typically requires a large amount of effort. Organisations involved in routine, large-scale record linkage frequently employ a system of manual review of created matches or links [9, 10]. A typical manual review method is to review all groups which contain a low weighted record-pair. However, this can be a time and resource intensive process and some errors are likely to remain in the data even after review. One organisation estimated the false positive error rate of their linkage, after extensive manual review, at 0.3 per cent [10].

1.5. Utilising additional information to improve quality

There are two standard approaches to improving overall linkage quality [11]. The first is to focus on the parameters and settings used within the linkage process itself. This can include changing the way individual fields are compared, or the statistical weights allocated to field combinations.

A second approach is to attempt to incorporate extra information. There are several examples of this approach in the broader entity resolution literature.

One common characteristic of these approaches is the use of the content of pairwise relations which exist between individual records to draw in extra information. For instance, Bhattacharya & Getoor [12] utilise information on relationships between individuals to determine identity. As well as utilising the attributes listed, they also use information about the real world attribute – for instance, if two records list 'Margo' as my mother, this approach checks to see if these names refer to the same real

world person, and then uses this information to effect the likelihood that the two records refer to the same entity. A similar approach has been used by Dong et al [13] in the field of personal information management.

In this paper, rather than using the *content* of pre-existing relations in data to improve entity resolution, we attempt to use the *structure* of relations formed during entity resolution to further refine the entity resolution process.

Several studies have attempted to use the structure of pairwise relations to infer information about linkage quality. These have typically been carried out during the grouping process, where pairs of records determined to belong to each other are amalgamated to form groups of records belonging to the same individual. For instance, one approach [14] to carrying out a linkage between sets of records known to belong to the same person joined sets of records together if a certain percentage of all possible pairs between them are found.

The record linkage literature has typically used the partition (also known as transitive closure) method for grouping together records. This method essentially joins any records directly or indirectly connected to each other into the same group. While this method is frequently used in the record linkage literature, alternative methods to cluster record pairs into groups have been proposed, and are used in the wider entity resolution literature [8]. These clustering methods typically use information about the structure of pairwise relations to determine which records belong to the same person.

The structure of pair relations can also be taken into account after linkage and grouping has taken place. One example of this approach [15, 16] involves removing low weighted pair edges from groups of records whose diameter is greater than a set length.

1.6. Record Linkage and Graph Theory

The study of structures of pairwise relations is carried out in the mathematical field of graph theory. The fundamental structure of interest is a “graph” which is made up of a collection of ‘nodes’ and lines called ‘edges’ which join one node to another. As shown in Figure 1, these pair relationships also exist in record linkage, and are the building blocks used to determine groupings of records that are considered to belong to the same person. In the person-based record linkage context, nodes are individual records, with edges representing the record-pair associations found through the pairwise linkage process. The similarities between record linkage and graph theory has been noted previously [17].

Graph theory techniques can be used to infer information about a group’s pair relations. These techniques may provide useful information regarding the likelihood that a group contains a false positive match. If groups that are in error can be successfully identified, then corrective action can be more effectively targeted, reducing the current cost and burden of manual review.

2. OBJECTIVES

In this paper we attempt to apply techniques from the domain of graph theory to improve the quality of record linkage. Specifically, we test whether constructs from graph theory can be applied to record linkage output to identify groups of records containing false positive links.

In line with previous research [8, 15, 18, 19], we hypothesise that sparser, more loosely connected groups of records are more likely to contain errors than tighter, densely linked groups.

To evaluate this, known pair relations from two large scale linkages of real administrative data are used as a gold standard, and attempts are made to identify groups containing false positives using measures from graph theory following a fresh probabilistic linkage of these datasets.

3. METHODS AND MATERIALS

3.1. Evaluation Strategy

The methodology used in the current study is outlined in Figure 2. First, the two datasets were internally linked using a standard probabilistic linkage strategy. Second; record-pairs arising from the linkage were converted into groups of records that were considered to belong to the same person. Third, each group resulting from this linkage then had a series of graph theory metrics applied to it. Along with these metrics, the number of false positives contained in each group was measured. The latter could be determined because the correct links were known to us due to the high quality linkage and clerical effort that had previously been carried out on these datasets. These 'gold standard' links were used as our 'truth set'. Both the ability of graph theory techniques to identify errors, and to measure the extent of errors was investigated - the graph theory metric results were compared with the actual existence and number of errors in the group to determine whether they were correlated, and thus whether these graph theory metrics could correctly identify groups in error.

As a final step, the ability of graph theory techniques to identify groups in error was compared to that of an alternative strategy commonly used by record linkage facilities, a, 'threshold setting' technique. This technique, the most common approach used in practice [20, 21], focuses on reviewing grouping that contain a low weighted pair (i.e. two records with a comparison score that is in the 'grey zone', or close to the acceptance threshold for matches).

3.2. Datasets

Ten years of West Australian Hospital Admissions data (approximately 7 million records) along with ten years of the NSW Admitted Patient Data Collection (approximately 20 million records) were used in the study. This data was made available as part of the PHRN Proof of Concept project [7]. Each dataset had been previously internally linked to a very high quality (by the WA-DLB [22] and CHeReL [23], respectively) using probabilistic linkage methods along with rigorous manual reviews of created links, and a quality assurance program to analyse and review likely errors. The links had also been use in a large number of research projects and published research articles. These high quality links were therefore used as the 'truth set' for the current study.

Summary information about the datasets can be found in Table 1.

3.3. Linkage Strategy

A probabilistic linkage strategy was used to internally link each file [20, 24]. The linkage approach was based on a published linkage strategy in an evaluation of linkage quality across a number of linkage software packages [25]. The linkage strategy used two blocks, the first being the Soundex of the surname along with first initial, the second being date of birth. Only record pairs with exactly the same values of one of these blocks were eligible for further comparison. This further comparison involved individually comparing all available fields in the record pair. In line with the probabilistic approach, each field comparison received a score based on the specific weights calculated for that field (see below). These scores were then combined across the available fields in the record pair to determine a total score.

A string similarity measure (the Jaro Winkler string comparator [26]) was used to compare all alphabetic variables (name, address and suburb), with exact matching used for all other variables. Day, month and year of birth were all compared as separate variables.

Probabilistic record linkage involves the calculation or estimation of two probabilities for each field, the m-prob and u-prob[20, 24]. These refer to the proportion of record pairs belonging to the same individual which have different values of a specific field, and the proportion of record pairs belonging to different individuals having the same value of a specific field, respectively. Correct m-probs and u-probs were calculated for each variable and used in linkage. This was possible due to our knowledge of the correct answers as to which record belonged to which individual.

To determine the appropriate threshold setting, a small number of record pairs at different matching scores were manually reviewed. A threshold was chosen at the point where approximately half of the record pairs found were correct.

Once linkage was completed, the accepted record pairs were amalgamated together using the partition method to form groups. This method involves simply amalgamating all record pairs, with every connected record belonging to the same group.

3.4. Graph Theory Metrics

For the purposes of this investigation, only a modest set of graph metrics was selected to apply to our linked groups. Though not exhaustive, the set was considered representative of the number and type of graph theory metrics that might be applied to record linkage output. The metrics were chosen on the basis of some assumptions about what the structure of pairwise relationships were likely to mean, and based on methods found in the literature. The basic hypothesis is that sparser, loosely connected graphs may be more likely to contain errors than tighter, well connected graphs. For instance, Group 2 in Figure 3 is held together by only a few connections, as opposed to Group 1, which is fully connected (every record has a pair with every other record). If Group 2 did not contain record 8, for instance, then record 7 would form a different group and would be considered a different person than the person represented by the grouping of records 9, 10, 11 and 12. This lack of connections (caused because these records did *not* join together during linkage) may suggest that this group is more likely to be in error. A second hypothesis is that the existence of clusters of highly connected sub-graphs within a connected graph may reflect individual groups of records incorrectly joined together.

Three graph theory metrics were included in the study - *completeness*, *diameter* and *bridges*. *Completeness* refers to the number of edges that exist within a graph and is measured by the number of edges out of the total number of possible edges in a graph. A graph is considered complete or fully saturated if it contains all possible edges. The *distance* between two nodes in a graph is the shortest number of edges that have to be traversed to move from one node to another. The *diameter* of a graph is a measure of the longest distance between two nodes. In Figure 3, Group 1 has a diameter of 1, while Group 2 has a diameter of 4. A *bridge* is an edge which, if removed, would disconnect a graph. For instance, in a graph with two nodes and one edge, this edge is a bridge. Bridges may be indicative of a false positive match which incorrectly joins two groups together. This may be more likely when there are a large number of nodes on either side of the bridge, which are well connected (see Figure 4, in which the edge from node 3 to node 8 is a bridge; this bridge appears to join two fully connected sub-graphs). For the purposes of this study, we only examined bridges whose removal would result in groups which contained three or more records.

Several papers in the literature make use of clustering or other algorithms to improve quality; these studies rely on the same assumptions and use the same metrics. For instance, the group linkage clustering method [14] accepts two sets of records as belonging together if a defined percentage of all possible pairs between them are found. This is equivalent to accepting a group if the completeness of a graph has reached a certain percentage. Similarly, the approach of removing low weighted pairs until there are no transitivity paths of a certain number of steps [15, 16] makes use of the diameter metric, to determine whether a group is to be accepted.

4. RESULTS

Group sizes following our internal linkage of the two datasets are shown in Table 2. Both datasets had roughly similar distributions of group sizes. Of note is the large number of groups containing just one record (513,237 or 36% in WA and 2,362,600 or 44% in NSW). As graph theory has no application to singleton groups, these cases were excluded from further analysis.

When the remaining groups were compared against groups in the 'gold standard' dataset, it was found that 10% of groups in WA and 18% of groups in NSW were in error i.e. contained at least one false positive record-pair.

The groups were then classified on the basis of the three graph theory metrics (diameter, completeness and bridge existence) and, for each metric, the number of groups found to be in error at each cut-off value was calculated (see Table 3). As the table shows, each metric was highly effective at identifying groups in error. By selecting specific cut-off values (parameters) for each metric, it was possible to identify incorrect groups over 99% of the time. As is evident, almost all groups with less that 50% completeness (99.7%) were in error; all groups with diameter greater than five were in error; and all groups containing two or more bridges were similarly incorrect.

While the application of graph metrics was very effective at identifying incorrect groups, these groups only accounted for a small number of *all* the incorrect groups. Depending on the parameters used, the percentage of the incorrect groups found varied from 2%, to around 30%. This suggests that the graph theory techniques have high specificity but low sensitivity.

One reason for this effect is the existence of a large number of small groups containing all possible pairs – 47% and 57% of groups were complete (or fully saturated) for NSW and WA respectively. For these groups, the graph theory techniques were unable to identify errors - 53% and 23% of groups in error were fully saturated groups for NSW and WA respectively.

As well as the existence of errors, the extent of errors was also used as an outcome variable, as measured by the proportion of incorrect record pairs found in a group out of the total number of record pairs in that group (Average Error Proportion, Table 3). There appeared no clear relationship between graph theory metrics and the extent of incorrect record pairs.

There was significant overlap in the groups identified by the three graph theory metrics. To measure the extent of the overlap, a cut-off was set for each metric where more than 50% of the groups identified were in error (in this case, a diameter greater than one, one or more bridges, or <90% completeness, as seen in Table 3). The number of groups identified by all three metrics, only two techniques or solely one technique is shown in Figure 5. No single metric identified a group which another technique did not also identify. While this appears surprising, it is most likely due to the specific cut-offs chosen (for instance, the existence of a bridge implies at least a diameter of three, and a less than fully saturated graph). Based on this analysis, the most appropriate metric to use

would be diameter – using this technique alone would achieve the same results as using all of the techniques combined.

Results from the use of an alternative strategy to identify incorrect groups (the threshold setting technique) are shown in Table 4. A series of cut-offs above the chosen threshold (16) were taken, and any group containing a record-pair with a weight below this cut-off was identified by this technique.

The accuracy of this method in terms of identifying groups with errors was then compared against the earlier method of applying various graph theory metrics. Both the precision in identifying groups with false positives (specificity or positive predictive value), and how many of the groups with false positives that can be identified (sensitivity), were calculated for each method. The Positive Predictive Value (PPV) here refers to the proportion of groups found that actually did contain an incorrect record-pair. Sensitivity here refers to the proportion of incorrect groups identified out of *all* incorrect groups within the dataset. Results are presented in Figure 6, with each point representing a particular cut-off value.

As noted previously, the three graph theory methods produced very similar results, with bridge identification only discovering a small number of (almost always) incorrect groups. For WA data, the graph theory techniques appeared to be better at identifying incorrect groups than the threshold setting method, identifying the same number of incorrect groups with far fewer wrongly identified correct groups. However, this was not the case for the NSW data. Graph theory techniques were only able to provide a higher level of precision at very small levels of recall.

Graph theory methods identified different groups to the default threshold setting technique typically used. This can be seen in Figure 5.

5. DISCUSSION

The purpose of this study was to evaluate whether constructs from graph theory could be usefully applied to identify incorrect groups of records arising from record linkage. Given the large amount of time consumed in manual review by organisations operating record linkage facilities, methods that are able to target or reduce errors in linkage quality are highly desired.

As the results from the study show, three graph theory metrics (completeness, diameter and bridges) could be successfully adapted and used to identify groups of records containing errors. The graph-theory methods were able to target incorrect groups very accurately (i.e. high PPV); however, they could only identify a small percentage of all incorrect links (low sensitivity).

In practical terms, these graph theory techniques show promise as an additional method of identifying suspicious or incorrect groups in record linkage output..

However, results from the study also suggest that the effectiveness of using threshold setting versus graph theory techniques to identify incorrect groups depends heavily on properties of the underlying

dataset. In our case, using graph theory techniques instead of threshold setting for the WA data could significantly reduce the number of groups requiring manual examination, while resulting in the same underlying quality. Given the large amount of time required for the manual review process, this represents a substantial cost reduction. For data of poorer quality (such as NSW which had missing name information in over 30% of records) this is not the case, and there is no current way to determine *a priori* whether threshold setting, or graph theory techniques will be more appropriate for a given dataset.

Graph theory techniques could be used by linkage units who currently do not use threshold setting techniques to manage quality. This may be, for instance, because this approach is deemed too large, time consuming or expensive. As graph theory techniques can identify incorrect groups with a very high accuracy, (albeit a smaller number of incorrect groups) this approach may now be considered more economical.

Another approach would be to combine both graph theory techniques and threshold setting techniques in identifying incorrect groups. This would allow users to focus on groups which they are almost certain are incorrect first (those identified by graph theory techniques at higher cut-offs), before moving to those groups which they are less certain about. The results shown in Figure 6 indicate that both techniques identify a large number of groups which the other technique does not – using both techniques together will thus improve linkage quality to a greater extent than using one technique alone. This method would also allow users to measure the PPV found from groups identified by graph theory techniques and those identified by threshold setting techniques, and so to choose which technique to focus on.

The current study chose to investigate three graph theory measures as examples of techniques that had potential to identify matching errors. It is likely that other constructs from the graph theory domain could identify further errors in linkage output.

Given that the quality of linkage results often relies on the underlying data, further testing with additional datasets would provide a clearer picture of the efficacy of the presented method. The paucity of large datasets with known high quality gold standard results with which to compare is a known issue in record linkage research [16, 25].

One limitation to the use of graph theory techniques to investigate incorrect groups is the large number of small groups (groups containing 1 or 2 records). These groups contain 0 or 1 pair respectively. The measures described in this paper cannot be usefully applied on groups of this size. The existence of large numbers of groups of this size depends directly on the nature of the datasets being linked together. As more information on these individuals is linked together (as more datasets are linked together), the number of individuals with only one or two records will decrease.

A second limitation lies in the number of fully connected groups found during the study. It is unlikely that any graph theory techniques will be able to distinguish between fully connected groups well enough to identify errors (the only thing such a technique would be able to use would be the pair

weights). Given that for our two datasets this accounted for 23% and 53% of all errors, it is clear the graph theory techniques will not be able to influence a high proportion of all the groups identified with false positive links.

The present study involved the use of graph theory techniques to *identify* groups containing incorrect links. It did not investigate methods of fixing these groups once identified. Several methods currently exist – including the manual determination of which records within a group belong to the same individual, the removal of lowly weighted pairs until the cause of the error no longer exists [15, 16], or the use of graph theory metrics to identify particular links in error. Further research is required to determine the appropriateness and impact of these different methods, in terms of linkage quality and cost.

6. CONCLUSION

Graph theory techniques have been underutilized by organisations operating record linkage facilities. Numerous avenues and opportunities for further research exist. For example, one area of interest is the use of weighted graphs - using the weights resulting from the probabilistic matching record-pair associations to characterize or identify anomalies. The greater challenge is, however, not only to identify incorrect groups, but also to develop approaches that correct errors and streamline the quality process.

Acknowledgements

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network. The authors would like to thank the reviewer for their invaluable comments.

Competing Interests

There are no competing interests

Funding

This research received no specific funding

References

1. Brook, E.L., D.L. Rosman, and C.D.A.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, 2008 **32**(1): p. 19-23.

2. Hall, S.E., et al., *Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records.* International Journal for Quality in Health Care, 2005. **17**: p. 375-381.
3. Gill, L.E., *OX-LINK: The Oxford Medical Record Linkage System*. University of Oxford: Oxford.
4. Kendrick, S. and J. Clarke, *The Scottish Record Linkage System.* Health bulletin, 1993. **51**(2): p. 72.
5. Ford, D.V., et al., *The SAIL Databank: building a national architecture for e-health research and evaluation.* 4 September 2009.
6. Roos, L.L. and J.P. Nicol, *A research registry: uses, development, and accuracy.* Journal of clinical epidemiology, 1999. **52**(1): p. 39-47.
7. Boyd, J.H., et al., *Data linkage infrastructure for cross-jurisdictional health-related research in Australia.* BMC health services research, 2012. **12**(1): p. 480.
8. Hassanzadeh, O., et al., *Framework for evaluating clustering algorithms in duplicate detection.* Proceedings of the VLDB Endowment, 2009. **2**(1): p. 1282-1293.
9. CHeReL. *Quality Assurance*. 2013; Available from: http://www.cherel.org.au/quality-assurance.
10. Rosman, D., et al. *Measuring data and link quality in a dynamic multi-set linkage system*. in *Sydney (NSW): Symposium on Health Data Linkage http://www. publichealth. gov. au/symposiumpdf/rosman_a. pdf*. 2002.
11. Wajda, A. and L.L. Roos, *Simplifying record linkage: software and strategy.* Computers in Biology and Medicine, 1987. **17**(4): p. 239-248.
12. Bhattacharya, I. and L. Getoor, *Collective entity resolution in relational data.* ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. **1**(1): p. 5.
13. Dong, X., A. Halevy, and J. Madhavan. *Reference reconciliation in complex information spaces*. in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 2005. ACM.
14. On, B.-W., et al. *Group linkage*. in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. 2007. IEEE.
15. Naumann, F. and M. Herschel, *An introduction to duplicate detection.* Synthesis Lectures on Data Management, 2010. **2**(1): p. 1-87.
16. Christen, P., *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. 2012: Springer.
17. Huang, Z. *Link prediction based on graph topology: The predictive value of the generalized clustering coefficient*. in *Workshop on Link Analysis: Dynamics and Static of Large Networks (LinkKDD2006)*. 2006.
18. Christen, P., *Data Matching*. 2012: Springer.
19. On, B.-W., et al., *Group Linkage.* 2001.
20. Newcombe, H.B., *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. 1988, New York: Oxford University Press.
21. Winkler, W.E., ed. *Matching and Record Linkage*. Business Survey Methods, ed. B. Cox, Chinnappa, Christianson, Colledge, Kott. 1995, John Wiley & Sons.
22. Rosman, D., et al., *Measuring data and link quality in a dynamic multi-set linkage system*. 2001, Data Linkage Unit, Department of Health (WA): perth.
23. Lawrence, G., I. Dinh, and L. Taylor, *The Centre for Health Record Linkage: a new resource for health services research and evaluation.* Health Information Management Journal, 2008. **37**(2): p. 60-62.
24. Fellegi, I.P. and A.B. Sunter, *A theory for record linkage.* Journal of the American Statistical Association, 1969. **64**(328): p. 1183-1210.
25. Ferrante, A. and J. Boyd, *A transparent and transportable methodology for evaluating Data Linkage software.* Journal of Biomedical Informatics, 2012. **45**(1): p. 165-172.

26. Winkler, W.E., *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. 1990.

Figure Legends

Figure 1: The grouping process converts individual record-pairs into information regarding which records belong to which person

Figure 2: Summary of the Study Methodology

Figure 3: By looking at the structure of pair relationships, we may be able to gain extra information about the groups in question

Figure 4: Is this single group actually two groups (individuals) held together by an incorrect match?

Figure 5: Overlap of groups identified by individual graph theory technique, and overlap of groups identified by graph theory techniques and threshold setting techniques

Figure 6: Comparison of techniques' ability to identify incorrect groups