Australian National University

Robust temporal graph clustering and cluster evaluation measure for group record linkage

Charini Nanayakkara, **Peter Christen**, and Thilina Ranbaduge

peter.christen@anu.edu.au

Research School of Computer Science, College of Engineering and Computer Science,

The Australian National University, Canberra, Australia

This research is conducted as part of the Digitising Scotland project

https://www.lscs.ac.uk/projects/digitising-scotland/

and partially funded by the Australian Research Council under DP160101934.

Outline

- Group record linkage and (temporal) constraints
- Temporal constraints based graph clustering
  - Detailed steps of our approach
  - Experimental evaluation on a Scottish data set from the Isle of Skye
- Cluster quality evaluation measure for group record linkage
  - Why traditional evaluation measures might not be adequate
  - A new cluster quality evaluation measure
  - Illustrative use on a Scottish data set
- Conclusions and future work

(Historical) Group Record Linkage

- Record linkage is the process of identifying sets of records that refer to the same entity (person) within one database or across different databases.
- In group record linkage, the aim is to link records for groups of entities, such as families or households.
- Historical record linkage refers to the linkage of historical birth, marriage, and death records for population reconstruction (building family trees), where each record contains information about several people.

Problem Statement

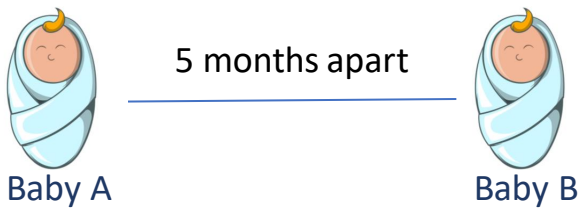
- **Aim**: To identify groups of records that refer to the same entities where there are certain temporal constraints between records.

- **Challenges**:
  - Existing record linkage techniques do not consider constraints that are implied by factors such as time (temporal), culture, or geographic location.
  - Data errors are often introduced when recording and transcribing the data.
  - Missing values in records.
  - Highly skewed frequency distributions of names.

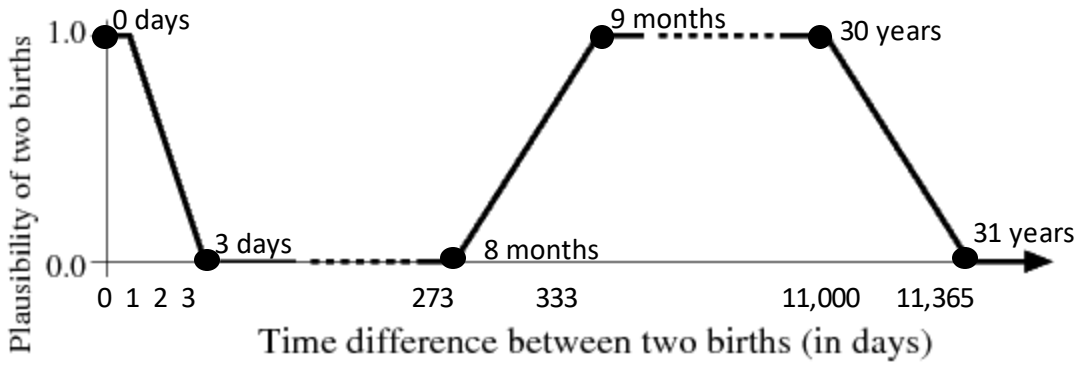
First name		Last name	
Father	Mother	Father	Mother
John (3,444)	Mary (2,740)	Mcleod (1,571)	Mcdonald (1,793)
Donald (2,628)	Catherine (2,607)	Mcdonald (1,556)	Mcleod (1,761)
Alexander (1,665)	Ann (2,084)	Mckinnon (1,168)	Mckinnon (1,164)
Malcolm (800)	Margaret (2,031)	Nicolson (1,047)	Nicolson (908)
Neil (787)	Christina (1,626)	Mclean (908)	Mclean (850)

Temporal Constraints Based Graph Clustering

- We introduce a novel graph clustering approach for group record linkage which takes temporal constraints into account.
- Temporal constraints: The constraints implied by time differences when linking records.

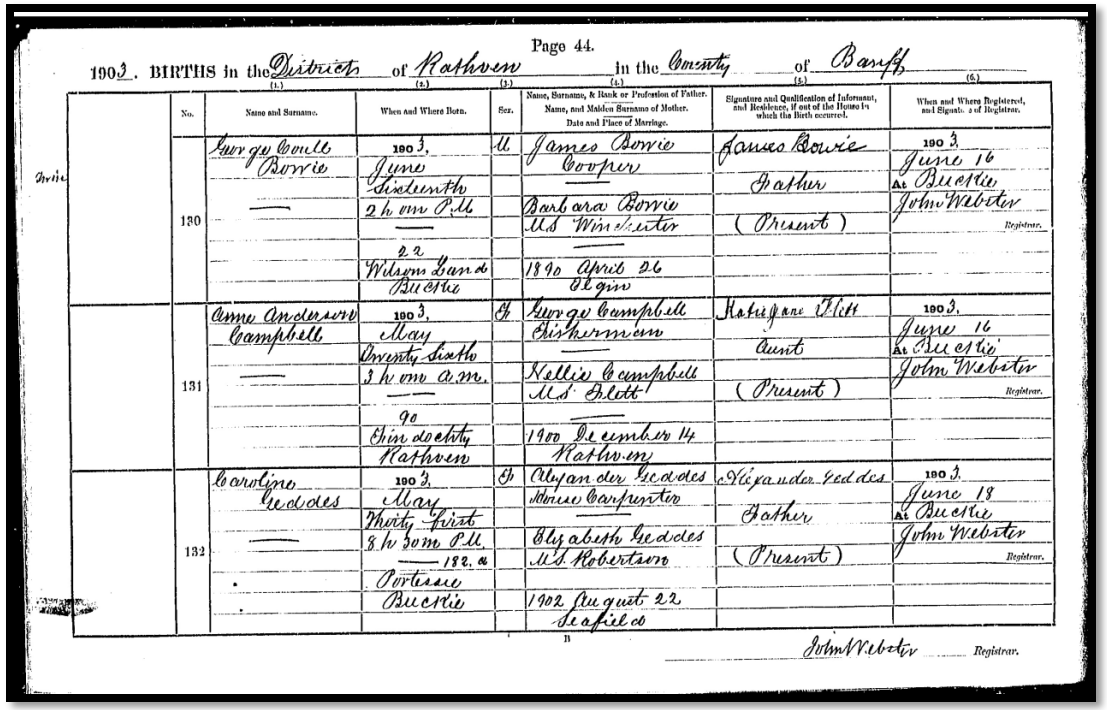


Due to biological limitations, it is temporally not possible for the same mother to have two babies 5 months apart.



Bangladesh woman with two wombs has twins one month after first birth: https://www.bbc.com/news/world-asia-47729118

Phase 1: Similarity Graph Generation



Page 44.

1902. BIRTHS in the District of Rathven in the County of Banff

No.	Name and Surname.	When and Where Born.	Sex.	Name, Surname, & Rank or Profession of Father. Name, and Maiden Surname of Mother. Date and Place of Marriage.	Signature and Qualification of Informant, and Residence, if out of the House in which the Birth occurred.	When and Where Registered, and Signature of Registrar.
130	George Combe Bowie	1902, June Sixteenth 2 h om P.M. 22 Wilson Land Buckie	M	James Bowie Cooper Barbara Bowie MS Winchester	James Bowie Father (Present)	1902 June 16 At Buckie John Webster Registrar.
131	Anne Anderson Campbell	1902, May Twenty Sixth 3 h om a.m. 90 Main Locality Rathven	F	George Campbell Fisherman Nellie Campbell MS Elliott (Present)	Katherine Elliott Aunt	1902 June 16 At Buckie John Webster Registrar.
132	Caroline Geddes	1902, May Thirty First 3 h 30m P.M. 182 Portessie Buckie	F	Alexander Geddes Annie Carpenter Elizabeth Geddes MS Robertson (Present)	Alexander Geddes Father	1902 June 18 At Buckie John Webster Registrar.

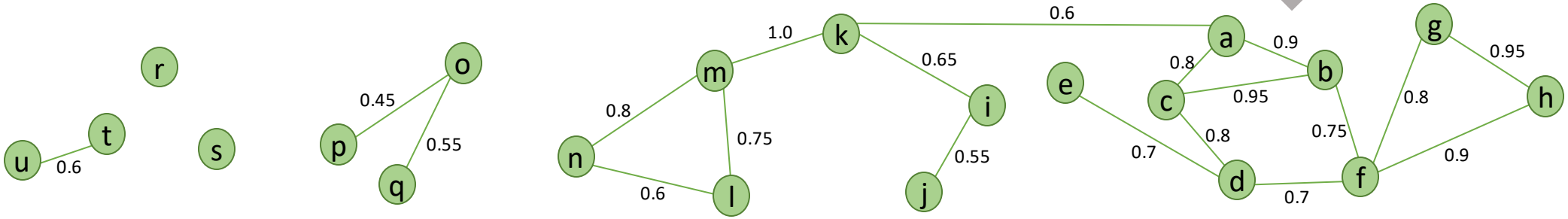
1902 Registered at Seafield

John Webster Registrar.

Transcribe Records

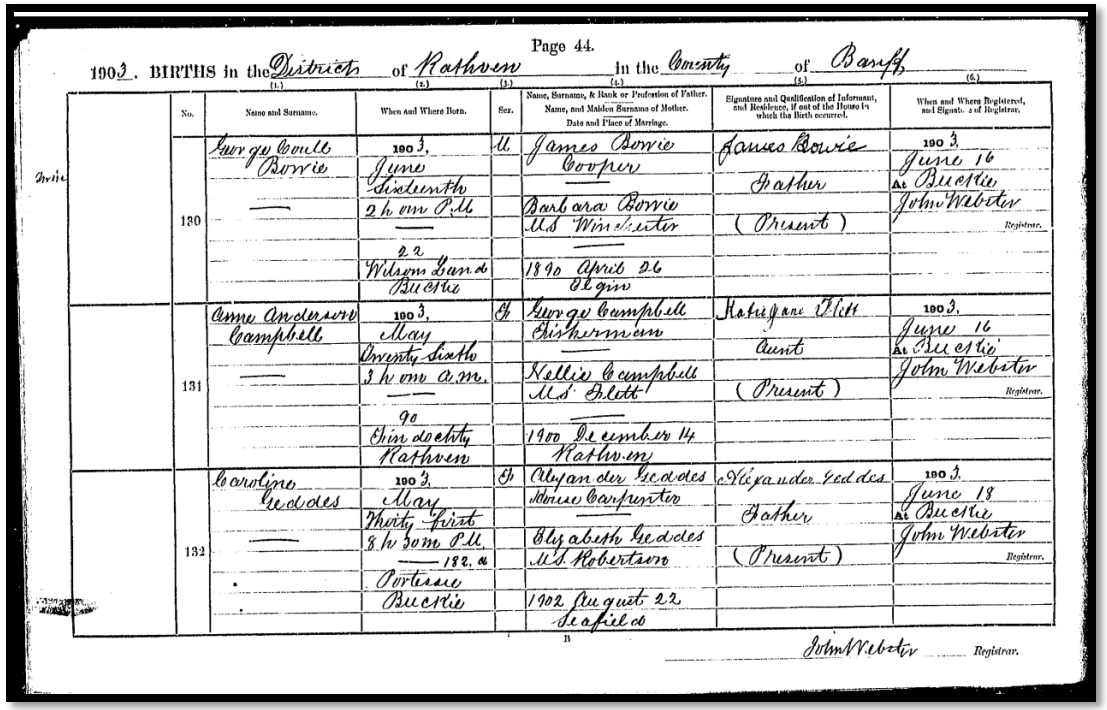
Record ID	Baby's name	Mother's name	Father's name	Date of birth	.....
k	Mary	Kate	John	01/02/1861	.....
l	Tom	Katy	Johnny	05/07/1863	.....
m	Pat	Kate	John	12/12/1869	.....
.....	.....	.....	.....	.....	.....
o	Harry	Peggy	-	03/09/1890	.....
p	Kate	Peg	Ron	06/11/1896	.....
q	Lizzy	Peggy	Roger	01/01/1901	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....

Generate Graph



Similarity graph *G*

Phase 1: Similarity Graph Generation



Page 44.

1902. BIRTHS in the District of Rathven in the County of Banff

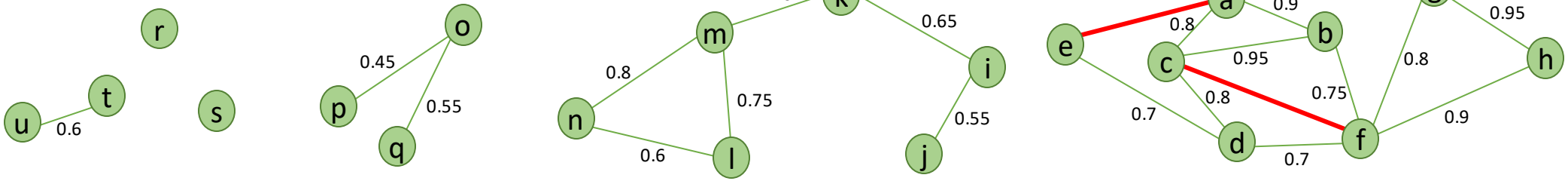
No.	Name and Surname.	When and Where Born.	Sex.	Name, Surname, & Rank or Profession of Father. Name, and Maiden Surname of Mother. Date and Place of Marriage.	Signature and Qualification of Informant, and Residence, if out of the House in which the Birth occurred.	When and Where Registered, and Signature of Registrar.
130	George Combe Bowie	1902, June Sixteenth, 2 h om P.M., 22, Wilson Land, Buckie	M	James Bowie, Cooper, Barbara Bowie, M.S. Winchester	James Bowie, Father, (Present)	1902 June 16, At Buckie, John Webster, Registrar.
131	Anne Anderson Campbell	1902, May Twenty Sixth, 3 h om a.m., 90, Thin Locality, Rathven	F	George Campbell, Fisherman, Nellie Campbell, M.S. Elliott	Nellie Jane Elliott, Aunt, (Present)	1902 June 16, At Buckie, John Webster, Registrar.
132	Caroline Geddes	1902, May Thirty First, 8 h 30 m P.M., 182, Portessie, Buckie	F	Alexander Geddes, Mason Carpenter, Elizabeth Geddes, M.S. Robertson	Alexander Geddes, Father, (Present)	1902 June 18, At Buckie, John Webster, Registrar.
		1902 August 22, Seafield				

John Webster Registrar.

Transcribe Records

Record ID	Baby's name	Mother's name	Father's name	Date of birth	.....
k	Mary	Kate	John	01/02/1861	.....
l	Tom	Katy	Johnny	05/07/1863	.....
m	Pat	Kate	John	12/12/1869	.....
.....	.....	.....	.....	.....	.....
o	Harry	Peggy	-	03/09/1890	.....
p	Kate	Peg	Ron	06/11/1896	.....
q	Lizzy	Peggy	Roger	01/01/1901	.....
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....

Generate Graph



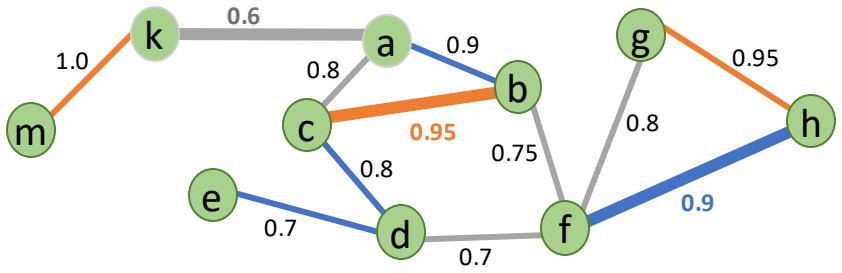
Similarity graph *G*

Temporally not possible links!!!

Phase 2 (a): Link Strength Based Edge Classification

- The concept of link strength is first used in record linkage by Saeedi et al. (2018). Only the edges with similarities greater than a user defined threshold are used.

- **Strong**: Edges (r_i, r_j) with the highest similarity with respect to all other edges connected to both r_i and r_j.
- **Norm**: Edges (r_i, r_j) with the highest similarity with respect to all other edges connected to either r_i or r_j, but not both.
- **WeakHigh**: Edges which are neither strong nor normal.



- Strong: *c, b* with similarity 0.95
- Norm: *f, h* with similarity 0.9
- WeakHigh: *a, k* with similarity 0.6

Phase 2 (b): Base Cluster Generation

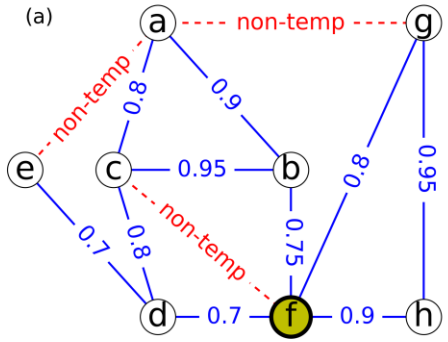
Create a new similarity graph G' using the selected link strength(s)

Generate connected components based on G'

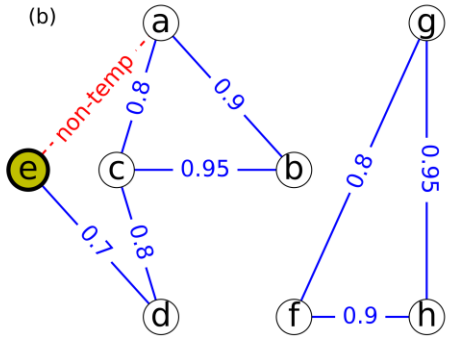
Iterative cluster refinement

Iterative Cluster Refinement:

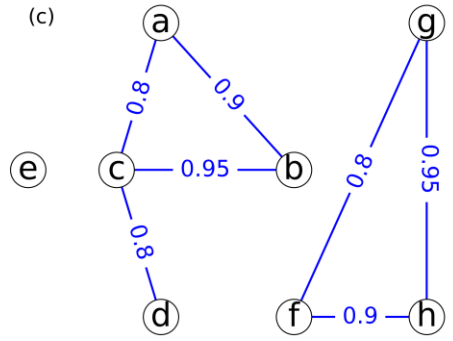
- The temporal implausibilities of connected components are eliminated in this step.
- For each connected component, nodes involved in implausible connections are ordered to determine the best sequence to iteratively remove non-temporal edges.



Ordered list = *[f, e, a, g, c]*



Ordered list = *[e, a]*



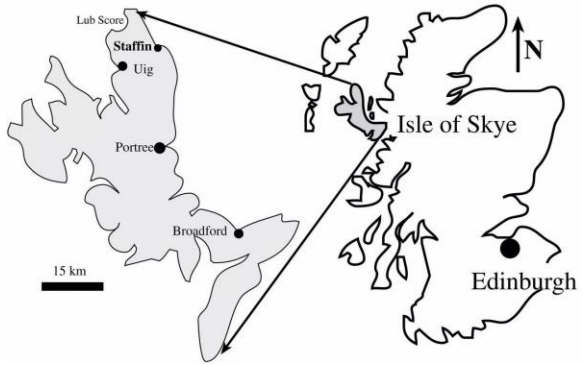
Phase 3: Iterative Cluster Merging

Pairwise base cluster similarity calculation using edges of the selected link strength(s)

Iteratively merge temporally plausible cluster pairs with cluster similarity greater than a user defined threshold

- Pairwise base cluster similarity is a combination of the similarity and the coverage.
- Similarity can be calculated as:
  - Maximum – maximum similarity among edges between two clusters (complete-link)
  - Minimum – minimum similarity among edges between two clusters (single-link)
  - Average – average similarity across edges between two clusters (average-link)
- Coverage = $\frac{\text{Number of edges of the selected link strength between two clusters}}{\text{Number of all edges between two clusters (with respect to the similarity graph } G\text{)}}$

Experimental Setup

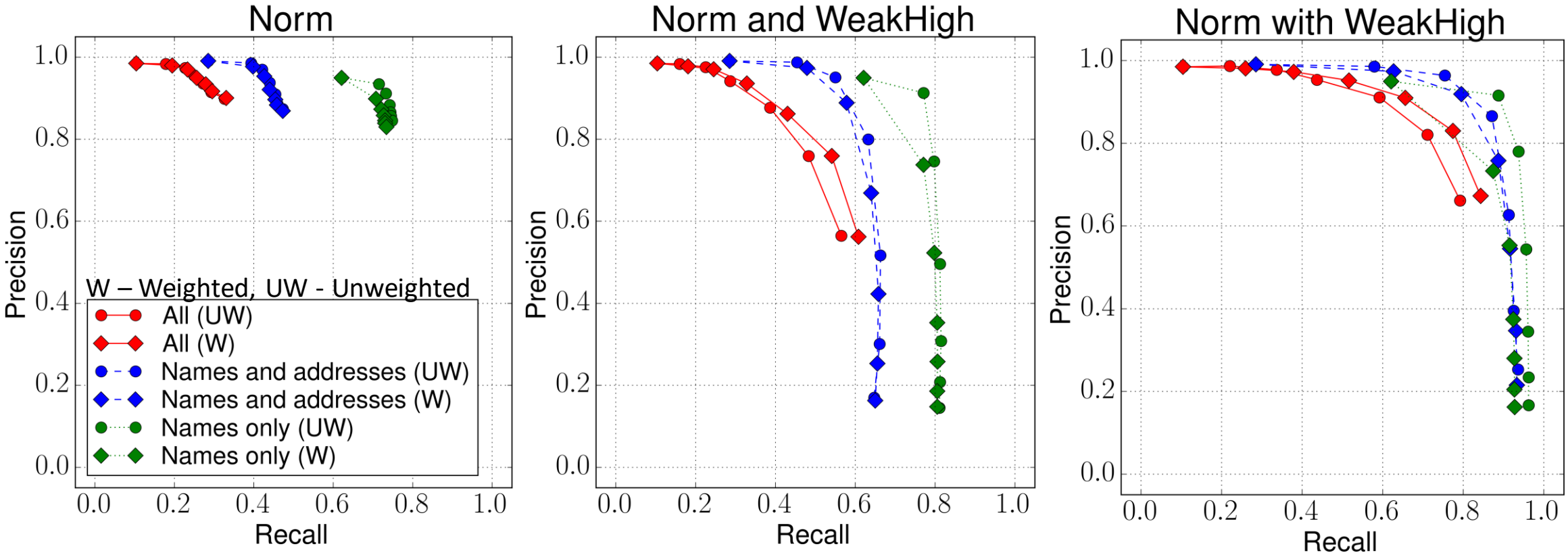


- Data set
  - For evaluation we used a real Scottish birth data set with 17,614 birth certificates, covering the population of the Isle of Skye from 1861 to 1901.
  - Each birth certificate contains personal details about a baby and its parents such as their names, address, marriage date, occupations, and the baby's date of birth.
  - We used six different attribute combinations for similarity calculation: all (parents names, addresses, occupations, and marriage dates), parent names with addresses, and parent names only, with and without weighting (Fellegi and Sunter, 1969).
- Evaluation measures:

Precision	Recall	Area under the precision-recall curve (AUC-PR)
TP/(TP+FP)	TP/(TP+FN)	A summary measure of the precision and recall values across different similarity thresholds

TP – True matching record pairs, FP – Wrongly matched record pairs, FN – Wrongly non-matched record pairs

Precision-Recall Curves



- Results are shown only for base clusters created with 'Strong' edges, since they showed highest precision (95%). Since the variation across similarity calculation methods was minimal, we have shown curves only for the 'average' similarity method.
- Surprisingly, better results were obtained with fewer attributes for similarity graph generation!

Area Under the Precision-Recall Curve (AUC-PR)

W – Weighted, UW - Unweighted

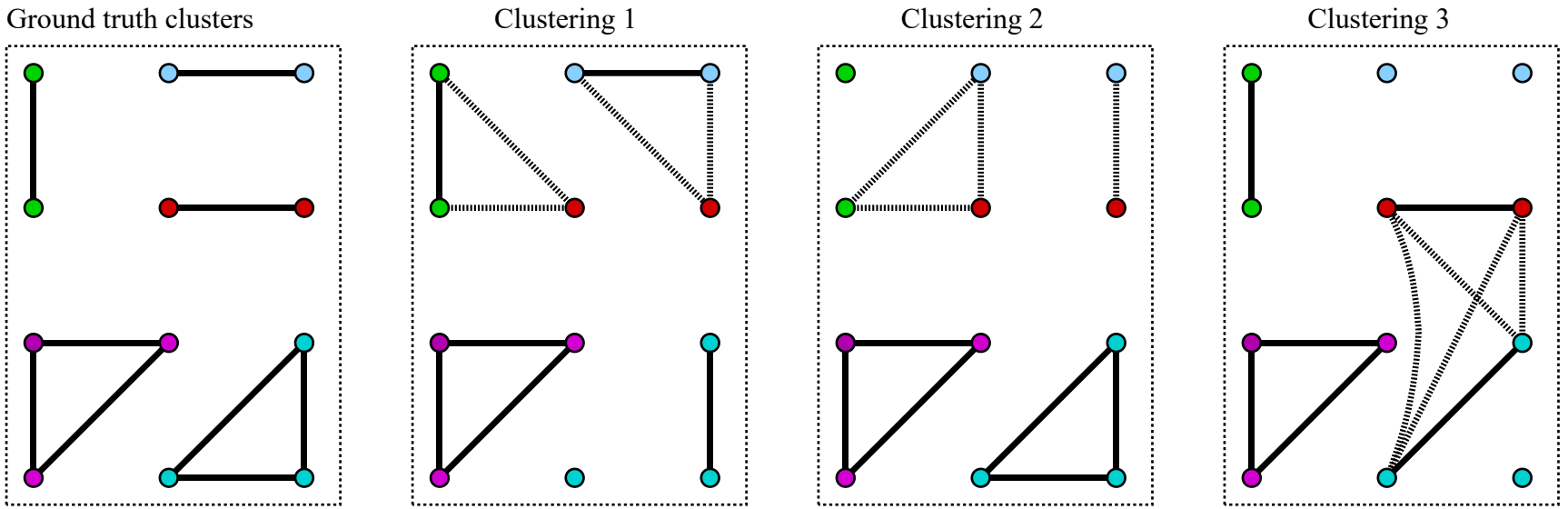
	Temporal		Non-temporal	
Similarity graph	Connected component	Star	Connected component	Star
All (UW)	0.72 ± 0.012	0.70 ± 0.003	0.64 ± 0.005	0.63 ± 0.003
All (W)	0.77 ± 0.014	$\mathbf{0.74} \pm 0.006$	0.69 ± 0.005	0.68 ± 0.004
Names and addresses (UW)	0.87 ± 0.006	0.70 ± 0.014	0.83 ± 0.002	0.73 ± 0.003
Names and addresses (W)	0.86 ± 0.007	0.69 ± 0.016	0.80 ± 0.003	0.72 ± 0.007
Names only (UW)	$\mathbf{0.88} \pm 0.002$	0.72 ± 0.018	$\mathbf{0.85} \pm 0.001$	$\mathbf{0.78} \pm 0.015$
Names only (W)	0.80 ± 0.002	0.65 ± 0.016	0.73 ± 0.001	0.69 ± 0.019
Averages	$\mathbf{0.82} \pm 0.064$	0.70 ± 0.030	0.76 ± 0.083	0.71 ± 0.051

- We compared this novel approach against our recently proposed temporal star clustering approach (Nanayakkara et al. 2018).
- There are no other temporal clustering approaches that we are aware of.
- Our new temporal approach achieved the highest average AUC-PR value of 0.88, compared to the previous temporal star clustering approach.

Are Precision and Recall Suitable for Evaluating Group Record Linkage?

- Precision and recall (as used before) have traditionally been employed to evaluate linkage quality in situations where ground truth data is available.
  - True Positives (true matching record pairs – correct matches).
  - False Positives (wrongly matched record pairs – false matches).
  - False Negatives (wrongly non-matched record pairs – missed matches).
- These metrics measure the quality of **links** between records.
- For group record linkage, however, we want the quality of clusters (groups) of records.
- Precision and recall can be ambiguous and not meaningful.

Examples of Different Cluster Predictions with same Precision and Recall Results



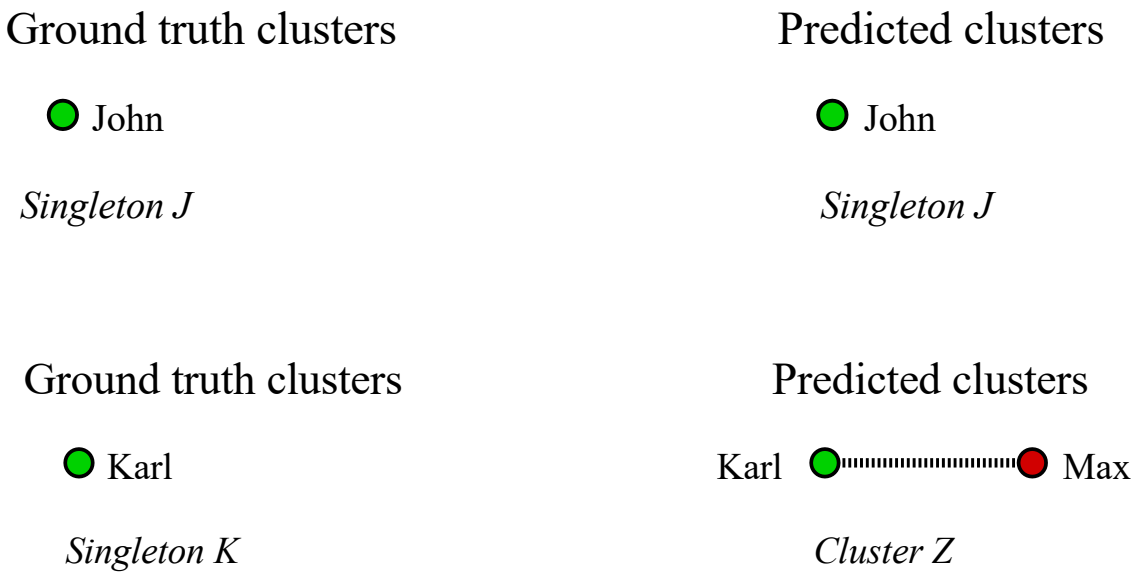
- The number of correct true matches (true positives) is 6 (solid lines).
- The number of false matches (false positives) is 4 (dotted lines).
- The number of missed matches (false negatives) is 3.
- Precision is 6/10 and recall is 6/9 for all three cluster predictions.

Record Based Cluster Evaluation Measures

- We need measures that assess the quality of clusters based on the records within them – with regard to ground truth clusters.
- This is a more complex undertaking, as there can be some correctly and some wrongly linked records in a cluster.
- The number of predicted clusters can also be higher or lower than the number of ground truth clusters.
  - In some applications this is problematic.
  - For example, in our birth bundling linkage we cannot have several clusters associated with a single mother.

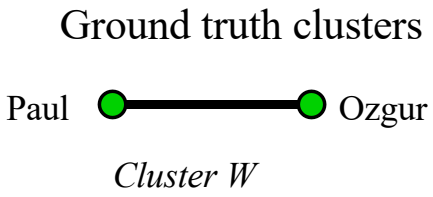
Seven Categories of Predicted Clusters (1)

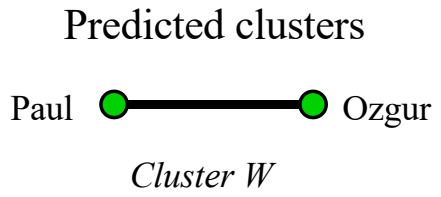
- Correct singleton (**SS**): Records in clusters of size 1 in both ground truth and predicted clusters.
- Wrongly grouped singleton (**SG**): Records in clusters of size 1 in ground truth but size larger than 1 (groups) in predicted clusters.
- Missed group member (**GS**): Records in clusters larger than size 1 in ground truth but size 1 in predicted clusters.
- Wrongly assigned member (**GG_W**): Records from a ground truth cluster of size larger than 1 are assigned to a wrong predicted group (not singleton).

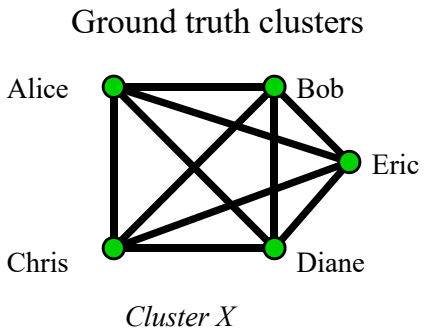


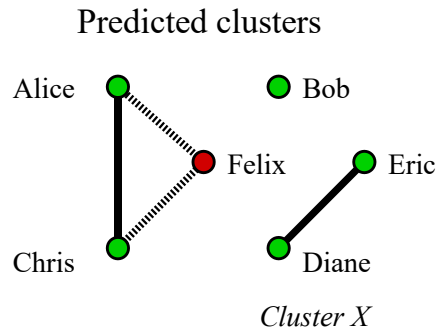
Seven Categories of Predicted Clusters (2)

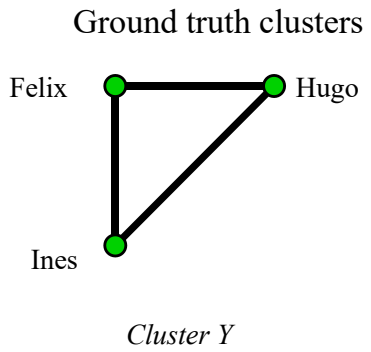
- Exact group match (**GG_E**): Clusters of size larger than 1 which are the same in ground truth and predicted clusters.
- Majority group match (**GG_M**): Clusters of size larger than 1 in both ground truth and predicted clusters, where the majority of records are the same.
- Minority group match (**GG_m**): Clusters of size larger than 1 in both ground truth and predicted clusters, where the majority of records are not the same.

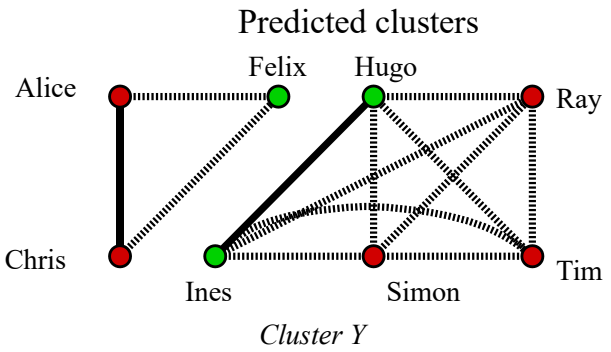






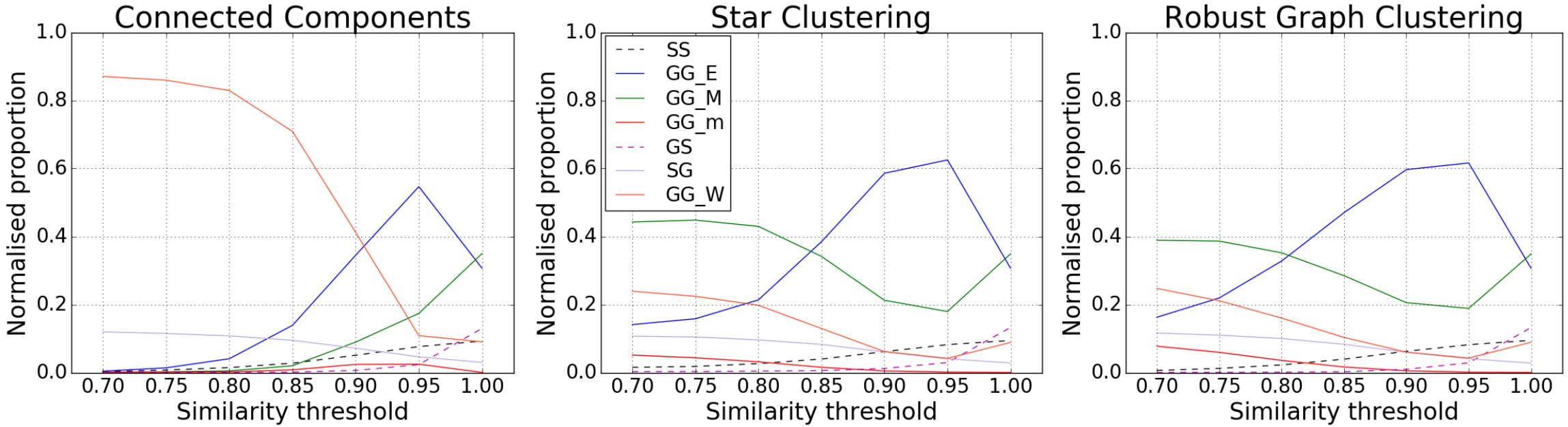






Categorising Records based on Thresholds

- As with traditional record linkage, we can classify record pairs as matches or non-matches based on different similarity thresholds.
- This will result in different numbers of records being classified into the seven categories.



Areas Under the Curves

- As with the AUC-PR, we can summarise these lines as areas under the curves over a range of different similarity thresholds (and normalised into the 0..1 range).
- Better clustering results will have higher values for **SS**, **GG_E**, **GG_M** and **GG_m**, and lower values for **SG**, **GS**, and **GG_W**.

Clustering technique	AUC-PR	SS	GG_E	GG_M	GG_m	SG	GS	GG_W
Connected components	0.744	0.036	0.206	0.077	0.010	0.087	**0.017**	0.567
Star clustering	0.775	**0.046**	0.367	**0.333**	0.020	**0.077**	0.020	0.137
Robust graph clustering	**0.885**	0.044	**0.413**	0.298	**0.027**	**0.077**	**0.017**	**0.124**

Conclusions and Future Work

- We proposed:
  - A novel temporal graph clustering approach for group record linkage, which addresses the previously highlighted challenges in this domain.
  - Our proposed approach takes advantage of the link strength categorisation in the record grouping, which improves clustering quality.
  - Experimental results show that our approach achieves improved linkage quality with respect to non-temporal clustering approaches, and substantially outperforms a previous temporal clustering approach for group record linkage.
  - A novel record based cluster evaluation measure for group record linkage which classifies records into one of seven categories.
- Future work:
  - Conduct empirical evaluations for different data sets and parameter settings.
  - Develop an adaptive technique to learn temporal constraints for different time intervals using ground truth data.
  - Investigate record linkage evaluation measures when no ground truth data are available.

Australian National University

Questions?

Robust temporal graph clustering and cluster evaluation measure for group record linkage

Charini Nanayakkara, **Peter Christen**, and Thilina Ranbaduge
peter.christen@anu.edu.au

This research is conducted as part of the Digitising Scotland project
https://www.lscs.ac.uk/projects/digitising-scotland/
and partially funded by the Australian Research Council under DP160101934.