Assessing record linkage accuracy using similarity weight matrix with Markov chain based Monte Carlo simulation approach

Shovanur Haque, QUT, shovanur.haque@hdr.qut.edu.au
Professor Kerrie Mengersen, QUT, k.mengersen@qut.edu.au

Queensland University of Technology

ACEMS AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Record Linkage

Record linkage (Newcombe et al., 1959; Fellegi and Sunter, 1969) is the process of finding matches and linking records from different data sources such that the linked records represent the same entity. For connecting records, record pairs from different files are compared on linking field values (for example, name, address, age, date-of-birth, sex etc.). Ensuring that the matched records in the combined file actually correspond to the same individual or entity is crucial for the validity of any analyses and inferences based on the combined data.

Objectives

We used similarity weight in the agreement matrix where partial agreement of the linking variable values is considered. An agreement matrix is created from all linking variable across all records in the two linked files and then simulates using Markov chain based Monte Carlo simulation approach for generating re-sampled versions of the agreement matrix. To assess the average accuracy of linking, correctly linked proportions are investigated for each record.

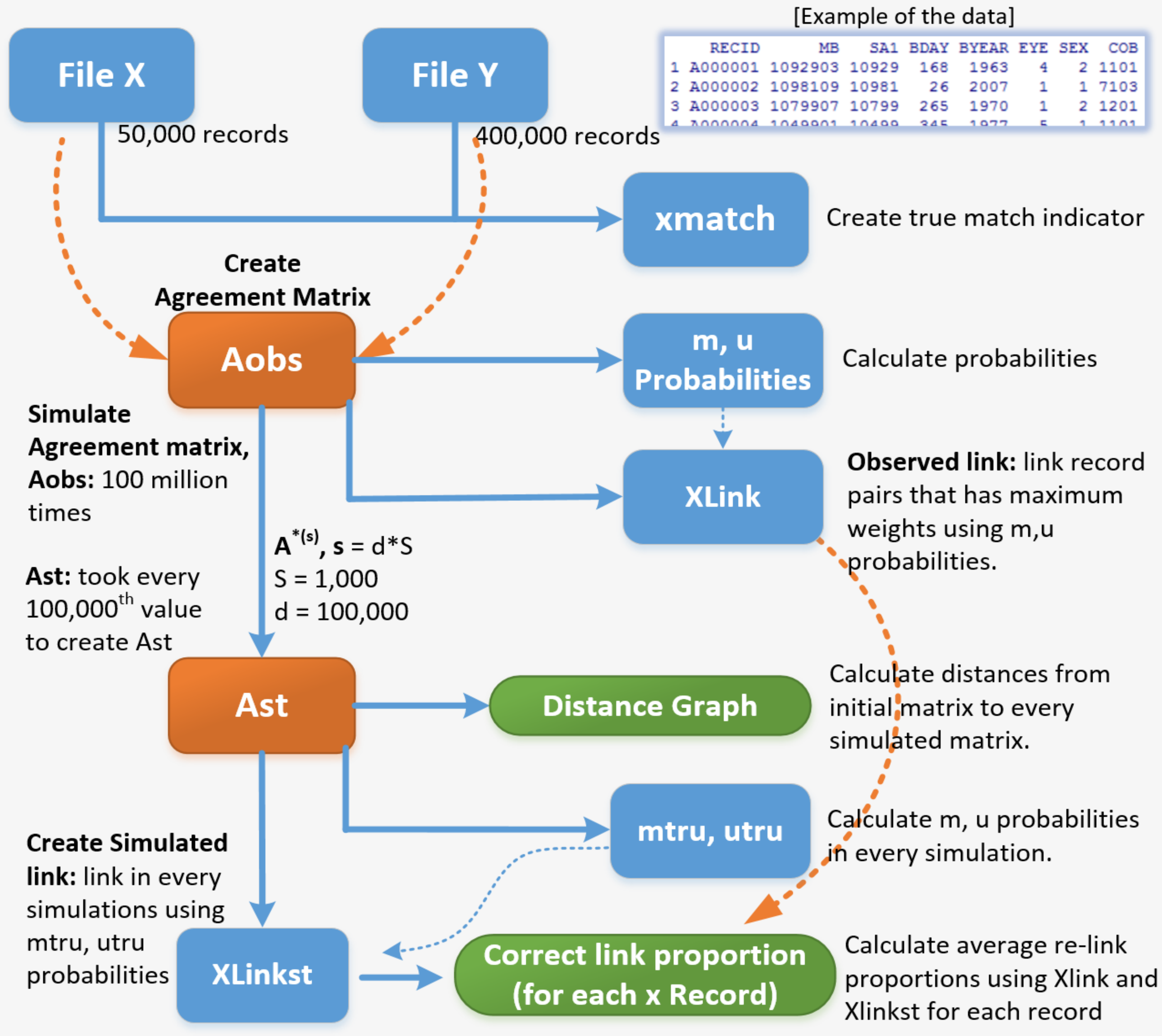
Method

Create agreement matrix, A

$\boldsymbol{A} = (A_{ijl})$; $i = 1, \dots R_X,\ j = 1, \dots R_Y,\ l = 1, \dots, L$, has been created with similarity weight of each record pair for each variable from two linked files, X and Y, having RX and RY entries, respectively. A_{ijl} takes values from 0 to 1 considering partial agreement of linking field l for record pair (i, j). A_{iil} refer to the matched record pair at the same position i in both files .

Simulating Agreement Matrix, A

We propose a Markov Chain, , $\{\boldsymbol{A}^{(n)}\}_{n=0,1,2,\dots}$, on $\mathcal{A}$ = {set of possible linking arrays}, with $\boldsymbol{A}^{(0)} = \boldsymbol{A}$, to generate re-sampled versions of the $\boldsymbol{A}$ array while preserving underlying probabilistic linking structure. Given the current state of the chain, $\boldsymbol{A}^{(n)}$, the next state, $\boldsymbol{A}^{(n+1)}$, is constructed following the algorithm developed. Parameters p and q ensures the stationary distribution of the chain maintains the required probabilities of agreement for both matched and non-matched records. m is the probability that linking field in both files has the same values for a matched record pair and u represents the probability that the values are the same for a non-matched record pair.


[Example of the data]

```
     RECID        MB   SA1 BDAY BYEAR EYE SEX  COB
1 A000001 1092903 10929  168  1963   4   2 1101
2 A000002 1098109 10981   26  2007   1   1 7103
3 A000003 1079907 10799  265  1970   1   2 1201
```

File X

50,000 records

File Y

400,000 records

xmatch

Create true match indicator

Create Agreement Matrix

Aobs

m, u Probabilities

Calculate probabilities

Simulate Agreement matrix, Aobs: 100 million times

XLink

Observed link: link record pairs that has maximum weights using m,u probabilities.

$A^{*(s)}$, s = d*S
S = 1,000
d = 100,000

Ast: took every 100,000th value to create Ast

Ast

Distance Graph

Calculate distances from initial matrix to every simulated matrix.

mtru, utru

Calculate m, u probabilities in every simulation.

Create Simulated link: link in every simulations using mtru, utru probabilities

XLinkst

Correct link proportion (for each x Record)

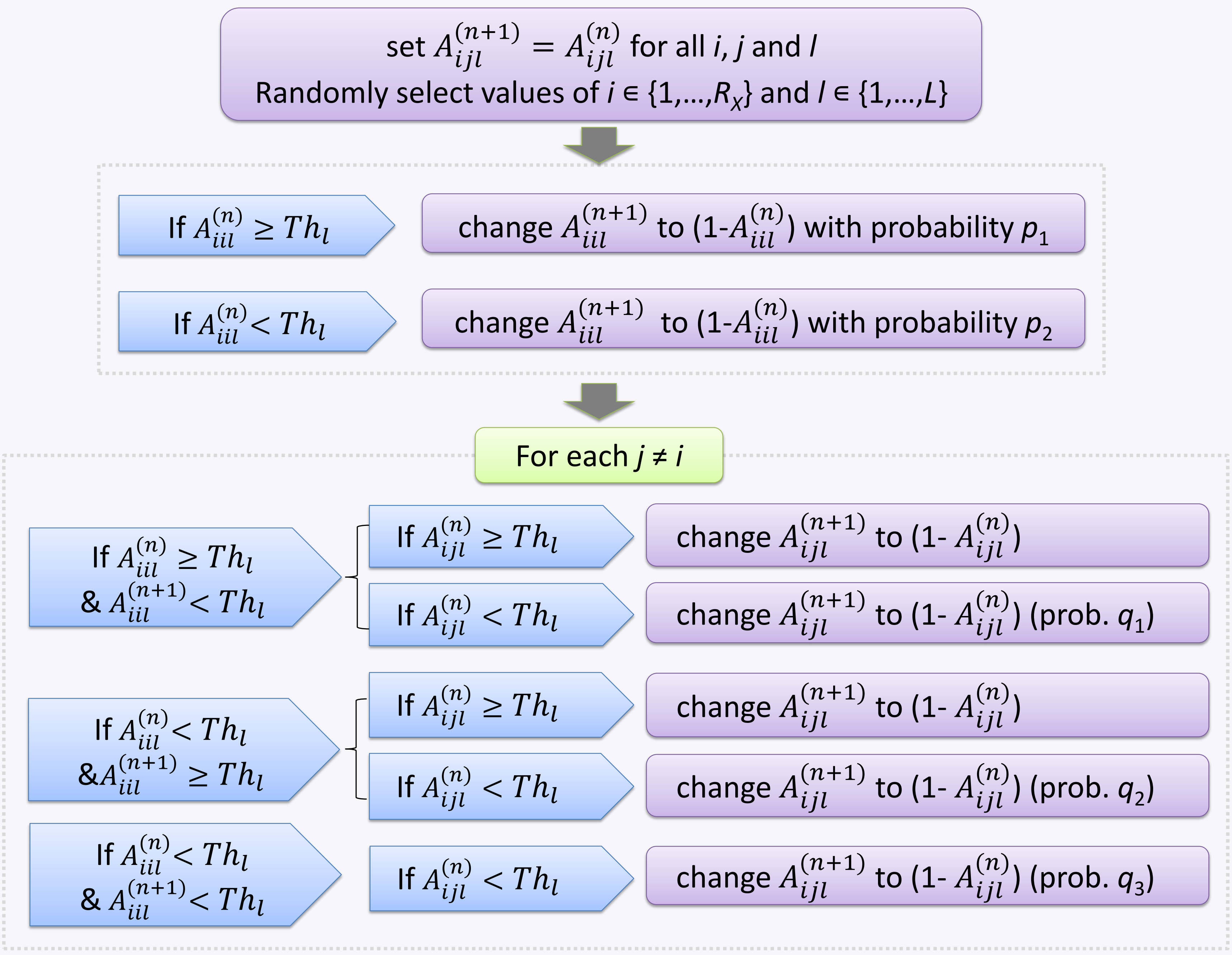
Calculate average re-link proportions using Xlink and Xlinkst for each record

Similarity weight

Similarity weight for each linking variable is calculated by, V_l = $(1 - {}^{d_l}/_{T_l})$ where, for any variable l, d_l is the difference between the values of the record pair and T_l is the difference between the maximum and minimum values of that particular variable.

We set a tolerance value for each variable l, i.e. Tol_l which is the maximum difference of values of linking variable can be considered as agree. We calculated individual agreement threshold for each linking variable l, Th_l, which is calculated by Th_l= $\left(1 - {}^{Tol_l}/_{T_l}\right)$.

Simulation Algorithm


set $A_{ijl}^{(n+1)} = A_{ijl}^{(n)}$ for all *i*, *j* and *l*
Randomly select values of $i \in \{1,\dots,R_X\}$ and $l \in \{1,\dots,L\}$

If $A_{iil}^{(n)} \geq Th_l$ — change $A_{iil}^{(n+1)}$ to (1-$A_{iil}^{(n)}$) with probability p_1

If $A_{iil}^{(n)} < Th_l$ — change $A_{iil}^{(n+1)}$ to (1-$A_{iil}^{(n)}$) with probability p_2

For each $j \neq i$

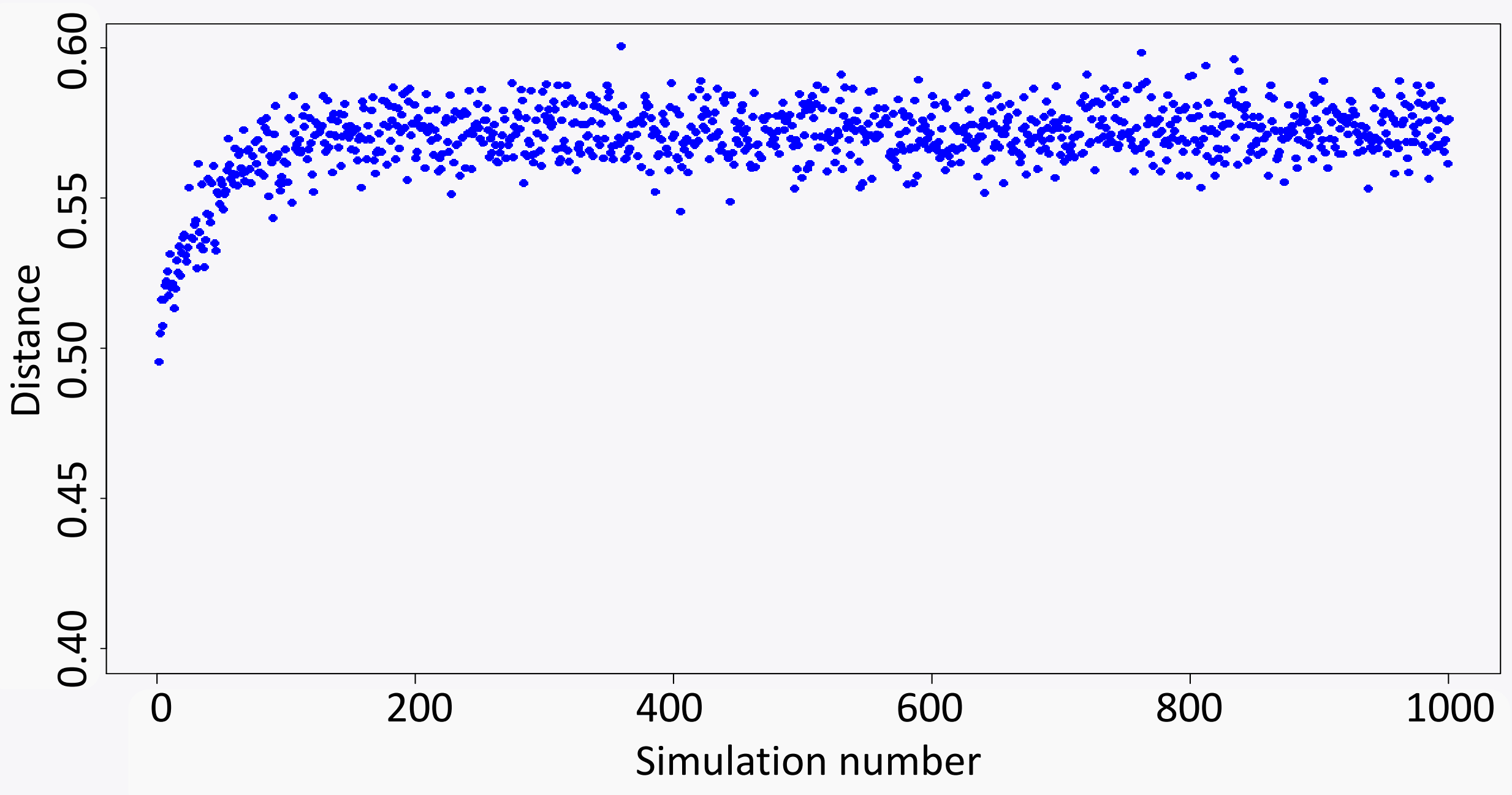
If $A_{iil}^{(n)} \geq Th_l$ & $A_{iil}^{(n+1)} < Th_l$
- If $A_{ijl}^{(n)} \geq Th_l$ — change $A_{ijl}^{(n+1)}$ to (1- $A_{ijl}^{(n)}$)
- If $A_{ijl}^{(n)} < Th_l$ — change $A_{ijl}^{(n+1)}$ to (1- $A_{ijl}^{(n)}$) (prob. q_1)

If $A_{iil}^{(n)} < Th_l$ &$A_{iil}^{(n+1)} \geq Th_l$
- If $A_{ijl}^{(n)} \geq Th_l$ — change $A_{ijl}^{(n+1)}$ to (1- $A_{ijl}^{(n)}$)
- If $A_{ijl}^{(n)} < Th_l$ — change $A_{ijl}^{(n+1)}$ to (1- $A_{ijl}^{(n)}$) (prob. q_2)

If $A_{iil}^{(n)} < Th_l$ & $A_{iil}^{(n+1)} < Th_l$
- If $A_{ijl}^{(n)} < Th_l$ — change $A_{ijl}^{(n+1)}$ to (1- $A_{ijl}^{(n)}$) (prob. q_3)

Output

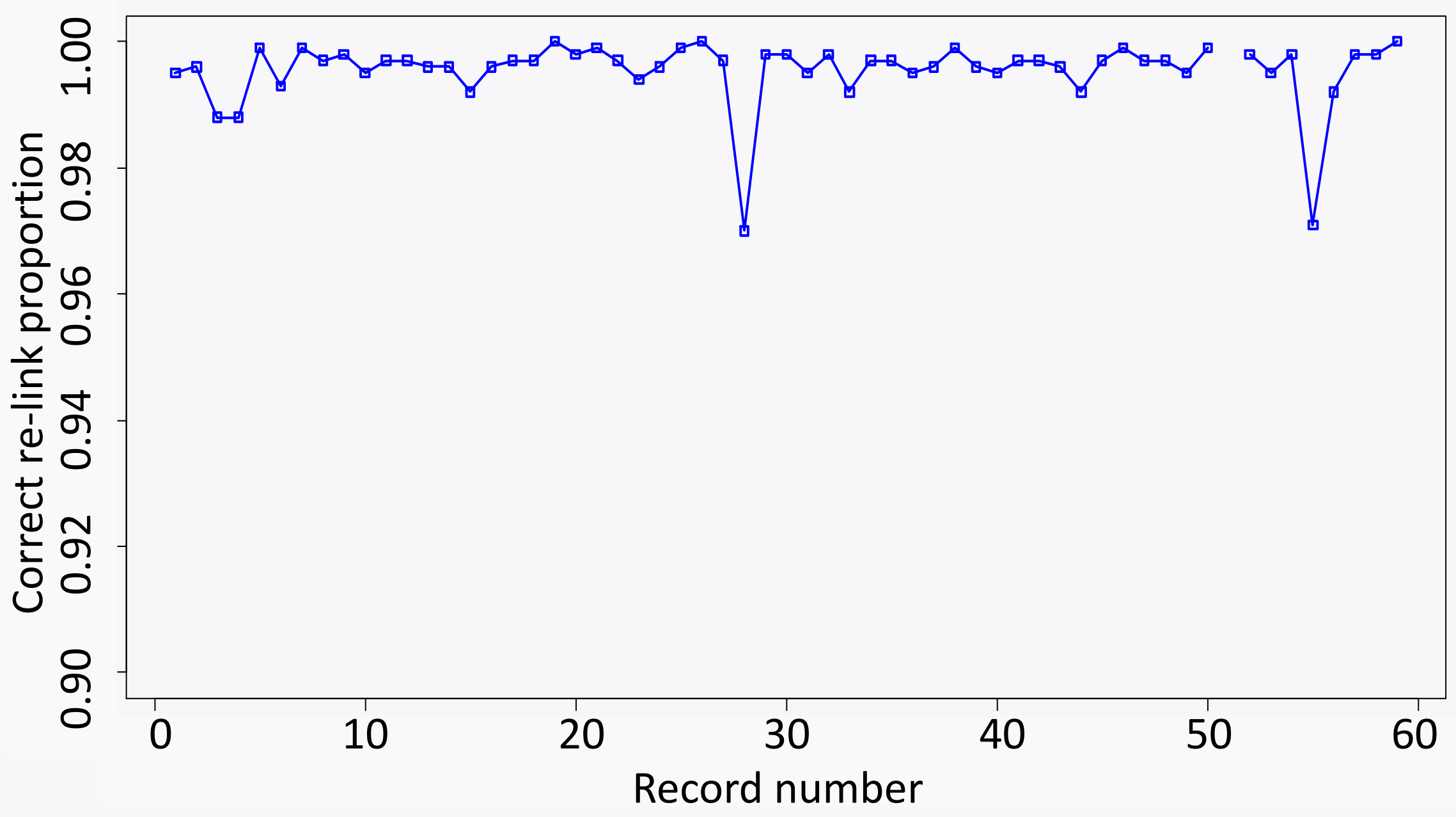
Distance



The plot shows the distances from the initial agreement matrix in each of total 1000 times simulations and observing the distances settling down.

Proportion of correct links for each record

The plot shows the average correct re-link proportion of each record for every simulation. Most of the records have nearly 100% accuracy. Only few records have around 97% to 99% accuracy.



Conclusion

The new methodology assesses linkage accuracy using Markov chain based Monte Carlo simulation method. Applying the partial agreement of the linking variable values in the form of similarity weight in the agreement matrix allows us to reduce the risk of missing potential matches. Since it is not practical to set same tolerance limit for agreement or disagreement of values for different linking variable, we have applied different agreement threshold for each linking variable that are used to determine the values for agreement and disagreement for each record pair. Adding this feature to the existing method didn't cost us expensive computation time rather gives us even higher accuracy than before.

REFERENCES
Fellegi,I.P., and Sunter,A.B. 1969),"A Theory for Record Linkage", Journal of the American Statistical Association,64,1183–1210.
Newcombe, H.B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records", Science, 954–959.