Australian National University

Linking sensitive and complex data

From pairwise linkage to reconstructing sensitive populations

Peter Christen and Charini Nanayakkara

School of Computing, the Australian National University; and
Scottish Centre for Administrative Data Research, the University of Edinburgh

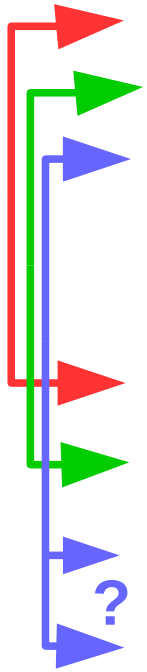
Why does record linkage matter?

- To investigate and address today's complex problems, data from a single source are rarely adequate
  - Think of an analysis of how the education and the social background of people affects their health outcomes

- Linking databases (especially across organisations) can be challenging
  - No common unique identifiers (such as social security numbers)
  - Data entry and processing errors (various social aspects influence data quality)
  - Scalability to linking databases with millions of records (quadratic increase)
  - How to evaluate the quality of a linked data set (ground truth is rarely available)
  - Linking **sensitive data** (personal data) raises privacy and confidentiality concerns
  - Linking **complex data** requires novel algorithms and techniques

An example of linking records

Name	Address	Phone	Age	Gender
John Smith	26 Miller St, O'Conner A.C.T.	6127 8042	42	M
Miss Mary Miller	4 Main Road Dixon ACT 2060	01 2345 6789	21	F
Dr Meyer, P.	5/42 MillAve, Sydeny 2000	61 (0)4 643 765	57	U



Title	FName	LName	Street	Suburb	Postcode	State	Sex	DoB
Mr	John	Smith	26 Miller Street	O'Connor	2602	ACT	0	12/03/1975
Ms	Marie	Miller	4 Main Road	Dickson	2602	ACT	1	23/12/1995
Dr	Paul	Meyer	5 Mill Avenue	Ryde	2112	NSW	0	4/10/1957
Mr	Paul	Meier	42 Miller Avenue	Manly	2095	NSW	0	10/08/1960

A long history of record linkage

- First ideas for computer based record linkage go back to the 1950s
- Seminal work by statisticians *Ivan Fellegi* and *Alan Sunter* in 1969
  - Compare the available common fields / attributes
  - Calculate matching weights based on probabilities (two records with same first name more likely refer to the same person than two records with same gender)

- Traditional linkage techniques assume two static databases
  - A common assumption is also that there are no duplicates (one record per person)
  - Names, addresses, and other personal details are required for linkage
  - No consideration of temporal or dynamic data aspects
  - No consideration of relationships between records (like people in a household)

Some existing record linkage solutions

- Multi-Agency Data Integration Project (MADIP) *Australia*
  - Linking of a diverse range of databases from government agencies
  - The *Australian Bureau of Statistics* is the trusted Integrating Authority
  - Uses a combination of deterministic and probabilistic linkage methods
  - Records are linked to a spine (central database with one record per person)
  - Anonymised microdata are then available for approved projects to approved government and non-government users
- Robodebt *Australia*
  - Automated matching of *Centrelink* records to *Australian Taxation Office* records
  - Automated sending of bills
  - Currently a Royal Commission in Australia, as well as class action and lawsuits

Tracey Donaldson says her problems began when a bill arrived in the mail last month.

It was a Centrelink debt for $45,500.

The letter was intended for another woman with the same name, who lived in a suburb with the same name in a different state, said Ms Donaldson, whose name has been changed for legal reasons.

The woman was also born on the same day of the same month as Ms Donaldson but in a different year.

Source: ABC News, 17 Feb 2020

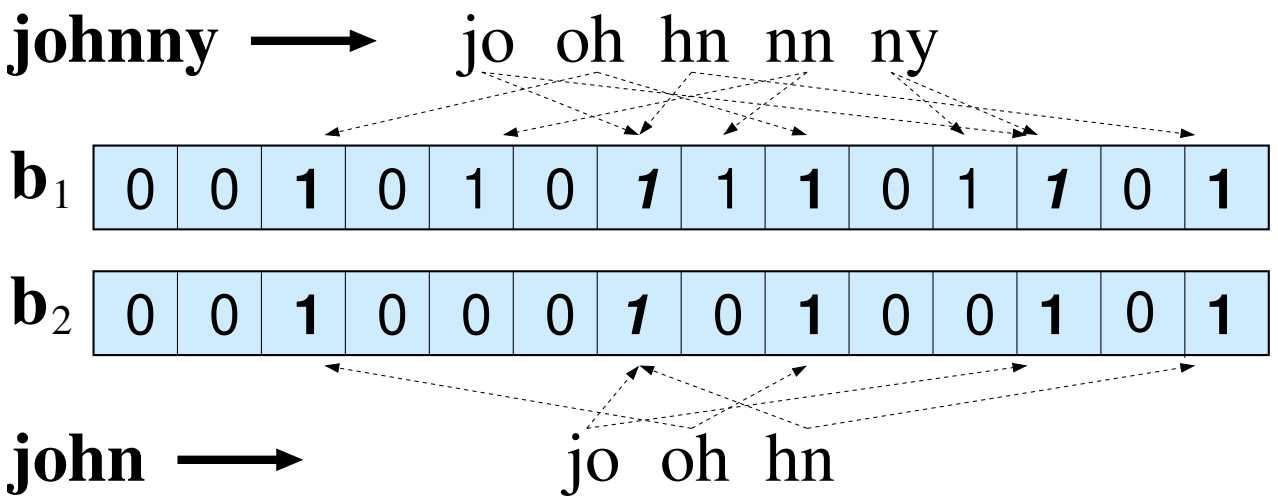
Some existing record linkage solutions

- Statistics New Zealand Integrated Data Infrastructure (IDI) *NZ*
  - Links administrative data from a number of sectors, including education, social welfare, migration and movements, justice, health and safety, and from *Statistics NZ* surveys
  - IDI spine contains records of over nine million people - created by probabilistically linking tax data to births data, births to visa data, and visa to tax data
  - *Statistics NZ* ensures access to data is provided only if the 'Five Safes' are met
- UK Data First project by the Ministry of Justice *UK*
  - Aim is to link criminal justice data with data from health, education, and so on
  - Requires finding duplicates in source data sets before linking across sources
  - Assign a meaningless identifier to each individual, and attach microdata
  - Developed the *Splink* open source software
  - Extensive consultation with a multi-disciplinary Academic Advisory Group

Linking sensitive data

- Sharing and linking sensitive (personal) data – especially across organisations – might not be permitted by regulations
  - Or due to commercial reasons (such as for private health providers)
- How can we link personal data without revealing any sensitive values?
  - The research area of privacy-preserving record linkage (PPRL), which started in the late 1990s, with first prototype methods in the early 2000s
  - First practical applications of PPRL in the last decade (mostly in the health domain)
- Basic ideas: Encode sensitive values to still allow similarity calculations
- Various challenges: Provable privacy and security, information leakage, vulnerability analysis, evaluation of linkage when using encoded data

Example Bloom filter encoding



Number of 1–bits:

$x_1 = 8$

$x_2 = 5$

Number of common 1–bits (bold):

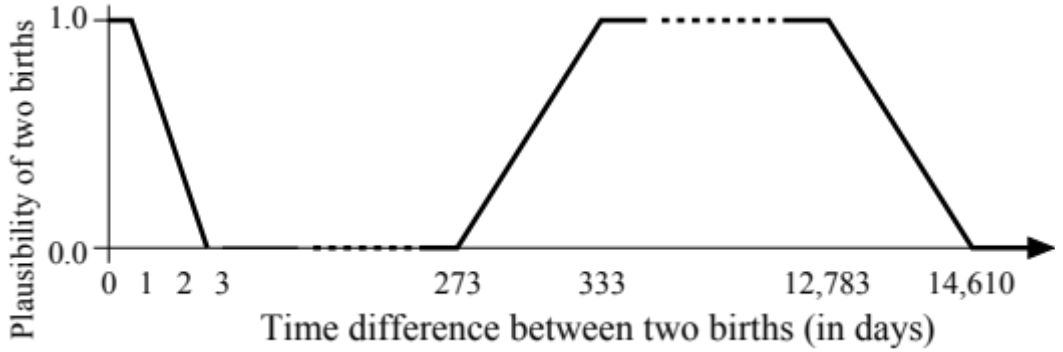
$c = 5$

$$sim_D(\mathbf{b}_1, \mathbf{b}_2) = \frac{2 \cdot 5}{(8+5)} = 0.77$$

- Used for linking health databases, for example by the *Centre for Data Linkage* at *Curtin University* (Perth, WA)
- An efficient and accurate method, that however has privacy weaknesses (frequent bit patterns) – *So don't use without talking to experts!*

Linking complex data

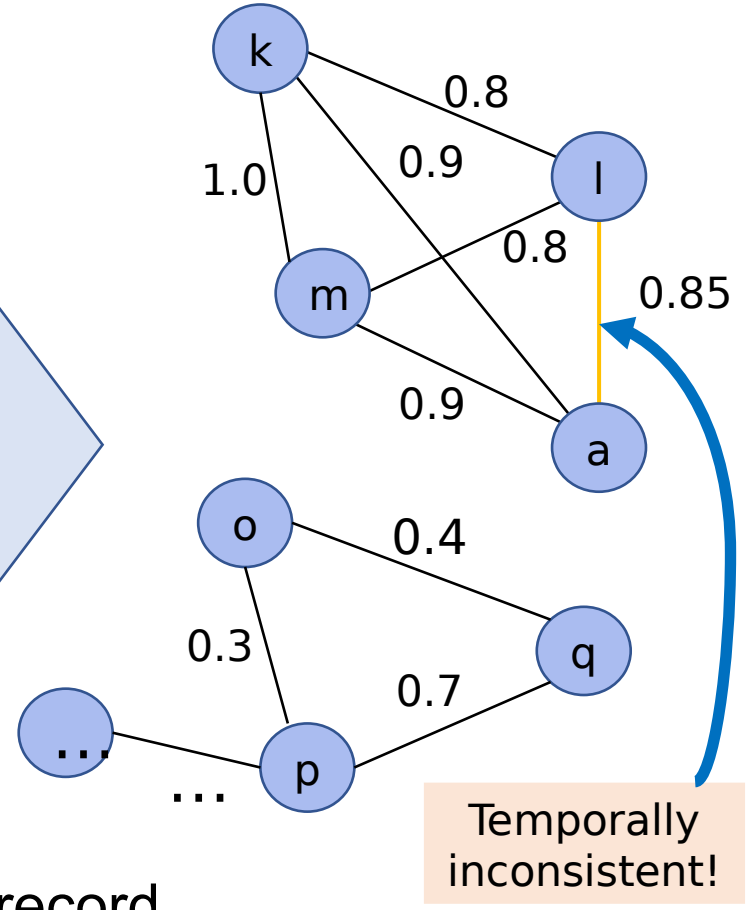
- Traditional pairwise linkage simply links record pairs with high attribute value similarities
  - But we can also consider relationships between records (household or family members), as well as temporal and spatial constraints
  - We can also link records of related entities (birth bundles)
- We found that incorporating temporal and spatial information available in population data can significantly improve linkage quality
  - Model the biological plausibility of consecutive child births to improve sibling linkage (birth bundling)
- To be used in the *Scottish Historic Population Platform* (SHiPP)



Linking related entities using temporal constraints

Record ID	Baby's name	Mother's name	Father's name	Date of birth	.....
a	Sam	Katy	John	11/02/1863	.....
.....	.....	.....	.....	.....	.....
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	.....
.....	.....	.....	.....	.....	.....

Apply blocking and comparison

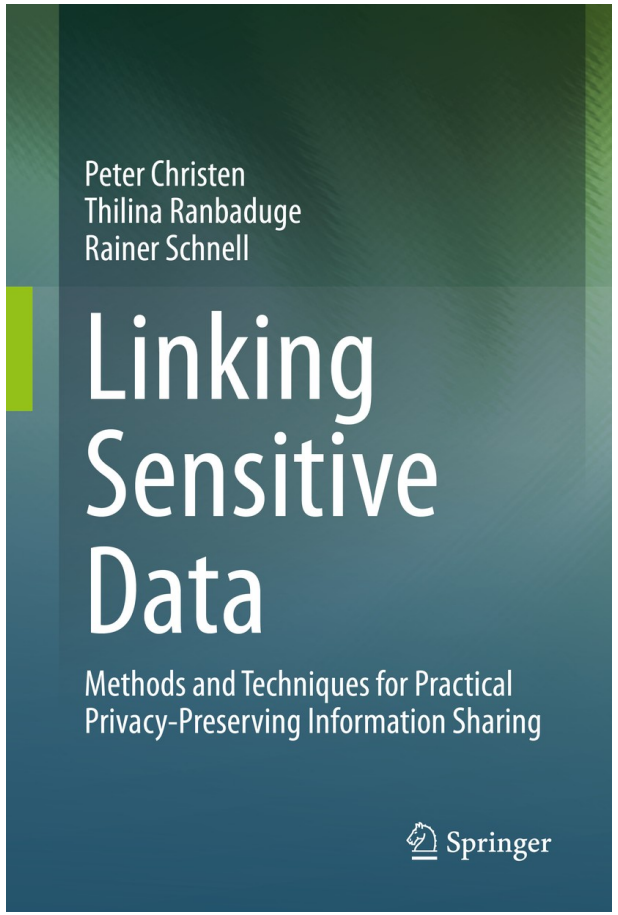


While nodes 'l' and 'a' would be linked in traditional pairwise record linkage, we disregard such pairs with our temporal constraint modelling

Conclusions and future directions

- Record linkage is not a new problem, yet still various unsolved tasks
- A core component of many data intensive systems, especially in governments and the health and social sciences
- New applications and challenges as new data sources / types become available (consider images for linkage, real-time linkage, etc.)
- Non-technical challenges include information governance and data access, while technical challenges include how to best employ new techniques (like deep learning based linkage algorithms)
- Public acceptance of record linkage is crucial (backlash to *Robodebt* in Australia)

Advertisement: The LSD book



The Book describes how linkage methods work and how to evaluate their performance. It covers all the major concepts and methods and also discusses practical matters such as computational efficiency, which are critical if the methods are to be used in practice – and it does all this in a highly accessible way!

Prof David J. Hand OBE,
Imperial College, London

Springer 2020