How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems

Bradley Malin and Latanya Sweeney, Carnegie Mellon University

Journal of Biomedical Informatics 37 (2004) 179–192

Introduction

- Addressing anonymity from a scientific perspective. Genomic data should not be related to the corresponding entities based on inference.

- Specifically, for the sake of privacy, patients' identifying information (name, address, etc) should be **strongly decoupled** from their genomic information (DNA patterns, diseases, etc) when each data set is **separately** made public.

- Both of these data sets exist in the quasi-public domain: personal in hospital admission records and genomic as released for research purposes.

Purpose

- To raise awareness that anonymity protection methods must account for healthcare and medical inferences that exist in a data sharing environment.
- To provide the biomedical community with a formal computational model of a re-identification problem that pertains to genomic data.

Previous research

- The authors created a model with capability of learning patient-specific genomic data from publicly available longitudinal information, relating disease symptoms to clinical states of the disease.
- The authors were able to uniquely re-connect genomic data to the name and demographics of the patients from which they were originally obtained (via “trails”) using their REID (RE-Identification DNA) algorithm. This is the basis for the generalizations presented in this paper.

Institutional Review Board oversight (IRBs)& data use agreements (DUAs)

- HIPAA does not specifically classify DNA data (sequence data, expression microarrays, etc) as an identifying attribute of a patient. So DNA data can be released under the Safe Harbor provision of their Privacy Rule.
- Datasets that are made publicly available are not subject to IRB review, nor are DUAs required.
- DUA & IRB are required when data is to be shared for research and is subject to HIPAA (IRB for federally funded research). *Unless the data is anonymous. There is thus a pressing need to guarantee that anonymity.*

Basic model

- Derived from relational database theory.

τ								
τ^+							τ^-	
Name	**Birthdate**	**Sex**	**Zip**	**Diagnosis**	**Treatment**		**Pseudonym**	**DNA**
John Smith	2/18/45	M	15234	3330	132		SA9212OK19	cttg...a
Mary Doe	4/9/75	F	15097	33520	653		AS09D8LK1J	atcg...t
Bob Little	2/26/49	M	15212	27700	742		D8A79AD133	acag...t
Kate Erwin	11/3/54	F	15054	3563	123		ASSD834MS1	accg...a

Fig. 1. Table τ is the data collection of a specific location and consists of all depicted attributed *Name*, *Birthdate*, . . ., *DNA*. The vertical partitioning of τ in the figure results in two subtables: an identified table τ^+ of patient demographics and a DNA table τ^- containing de-identified sequences. There is no reason that the ordering of the rows in τ^+ and τ^- must be the same as in τ. The arrows specify the truth about which tuples of τ^+ belong to τ^- in the original table τ.

Basic model: *assumptions*

- Each data-collecting location (hospital) releases only its own data. *So a patient must've visited hospital X for X to include his data.*
- Tuples in each data set are unique for each patient.

Basic model: *reserved vs unreserved*

- When every location releases tables, such that the only tuples present in tau negative have corresponding tuples in tau positive, and vice versa, we say that the tracks are **unreserved**.

- But data releasers and patients are autonomous entities, and either can choose to withhold certain information. Thus, releases that are unreserved are not always practical and, at times, can be impossible to achieve.

- Consequently, we say that track **N** is **reserved** to track **P** if for every location *c*, for each tuple *x* such that *x* is a member of tau negative there exists a tuple *y* such that *y* is a member of tau positive, such that both *x* and *y* are derived from the same tuple in tau.

Basic model

τ^+	τ^-
Name	***DNA***
c_1	
John	acag...t
Mary	accg...a

c_2	
John	acag...t
Bob	cttg...a

c_3	
Mary	accg...a
Bob	cttg...a
Kate	atcg...t

P			
Name	c_1	c_2	c_3
John	1	1	0
Mary	1	0	1
Bob	0	1	1
Kate	0	0	1

N			
DNA	c_1	c_2	c_3
accg...a	1	0	1
cttg...a	0	1	1
acag...t	1	1	0
atcg...t	0	0	1

τ^+	τ^-
Name	***DNA***
c'_3	
Mary	accg...a
Bob	
Kate	

N′			
DNA	c_1	c_2	c'_3
accg...a	1	0	1
cttg...a	0	1	0
acag...t	1	1	0

Fig. 2. (Left) Identified (**P**) and DNA (**N**) tracks created from unreserved releases of three locations c_1, c_2, and c_3. Both **P** and **N** are unreserved tracks. (Right) Resulting DNA track **N′** is created from the substitution of the reserved release from c'_3 for the unreserved release of c_3. As a result of this substitution, **N′** is reserved to **P**.

Reidentification algorithms

- REIDIT-C *(complete)*. Simpler. Applicable only to unreserved data (complete trails).
- REIDIT-I *(incomplete)*. More realistic. Applicable when one track is reserved to the other (incomplete trails).

REIDIT-Complete

REIDIT-C Algorithm

Input: DNA and Identified Tracks **N** and **P** for the same data-collecting locations.
Output: Set of trail re-identifications *Reidentified*

Assumes: **N** and **P** are unreserved

Steps:

```
let Reidentified be an empty set
for each tuple n in N
    if there exists only one tuple p in P, such that trail(P,p) ≡ trail(N,n)
        Reidentified = Reidentified ∪ [p, n]
return Reidentified
```

Fig. 3. Pseudocode for the REIDIT-C algorithm.

Table 1
Classification of re-identifications made by REIDIT-C

	Re-identification	No re-identification
$\boldsymbol{trail}(\mathbf{N}, \boldsymbol{n}) = \boldsymbol{trail}(\mathbf{P}, \boldsymbol{p})$	Correct match	False non-match
$\boldsymbol{trail}(\mathbf{N}, \boldsymbol{n}) \neq \boldsymbol{trail}(\mathbf{P}, \boldsymbol{p})$	False match	Correct non-match

The first and second rows of the contingency table correspond to outcomes for when the considered trails are equivalent or not, respectively. Light-shaded cells are possible outcomes and the darkened cell is an impossible outcome.

REIDIT-Incomplete

Algorithm: REIDIT-I-Fast (X, Y)

Input: DNA and Identified Tracks **N** and **P** for the same data-collecting locations. **X** is the reserved table of **N** and **P**, and **Y** is the remaining table.

Output: Set of trail re-identifications *Reidentified*

Assumes: 1) **X** has incomplete trails and **Y** has complete trails. 2) **X** is the reserved track of **Y**

Steps

```
let Z be a |X| × |Y| matrix, such that Z[x,y] = 1 if trail(X,x) ≤ trail(Y,y) and 0 otherwise
let S be a |X| × 1 column vector, such that S[x] is the sum of the x^th row of Z
let Reidentified be an empty set
let FoundOne = False
do
    FoundOne = False
    for x=1 to |X|                                   // find an incomplete trails that
        if S[x] ≡ 1                                  // has only one supertrail
            FoundOne = True
            for y=1 to |Y|                           // if found, find the supertrail and
                if Z[x,y] ≡ 1                        // add the [supertrail,subtrail] pair
                    Reidentified = Reidentified ∪ [y, x]   // to the re-identified set
                    for z=1 to |X|                   // remove the re-identified
                        if Z[z,y] ≡ 1                // supertrail from further
                            Z[z,y] = 0               // consideration
                            S[z] = S[z] – 1
while FoundOne ≡ True
return Reidentified
```

Fig. 4. Pseudocode for REIDIT-I-Fast, a variant of REIDIT-I, with an efficient data structure.

Table 2
Classification of re-identifications made by REIDIT-I

	Re-identification	No re-identification
trail(**N**, ***n***) ⩽***trail***(**P**, ***p***)	Correct match	False non-match
Not(***trail***(**N**, ***n***)) ⩽***trail***(**P**, ***p***)	False match	Correct non-match

The first and second rows of the contingency table correspond to outcomes for when the subtrail property is satisfied and not satisfied, respectively. Light-shaded cells are possible outcomes and the darkened cell is an impossible outcome.

Experiments

- Experimental data from publicly available discharge data from the State of Illinois, 1990-1997. *1.3M discharges per year.*
  - *personal data:* date of birth, gender, zip code, hospital
  - *genomic data:* 9 different ICD-9 classification codes

Results (REIDIT-C)

Table 3
Summary of the percentage of actual re-identifications made by REIDIT-C for different genetic disease patient populations

Disease	Gender	Number of patients	Number of hospitals	Average number of patients per hospital	% Re-identified
CF		1149	174	11.92	32.90
	Female	557	142	7.28	43.09
	Male	592	150	6.94	39.36
FA		129	105	2.08	68.99
	Female	60	68	1.47	80.00
	Male	69	72	1.65	78.26
HD		419	172	4.37	50.00
	Female	236	149	2.76	79.14
	Male	183	127	2.70	50.63
HT		429	159	4.83	52.21
	Female	244	140	3.06	64.34
	Male	185	114	2.98	63.24
PK		77	57	2.15	75.32
	Female	52	48	1.85	80.77
	Male	25	25	1.36	80.00
RD		4	8	1	100.00
	Female	2	4	1	100.00
	Male	2	4	1	100.00
SC		7730	207	88.89	37.34
	Female	4175	189	55.87	43.76
	Male	3555	191	41.01	36.51
TS		220	119	3.82	51.60
	Female	97	88	2.60	78.35
	Male	123	87	2.60	61.79

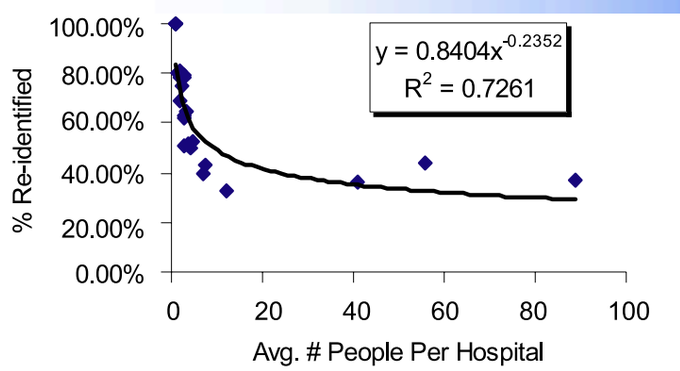

Fig. 5. REIDIT-C re-identification of populations as a function of the average number of people per location. Each genetic disease population has three data points in the graph: genderless, males only, and females only.

Results (REIDIT-C)

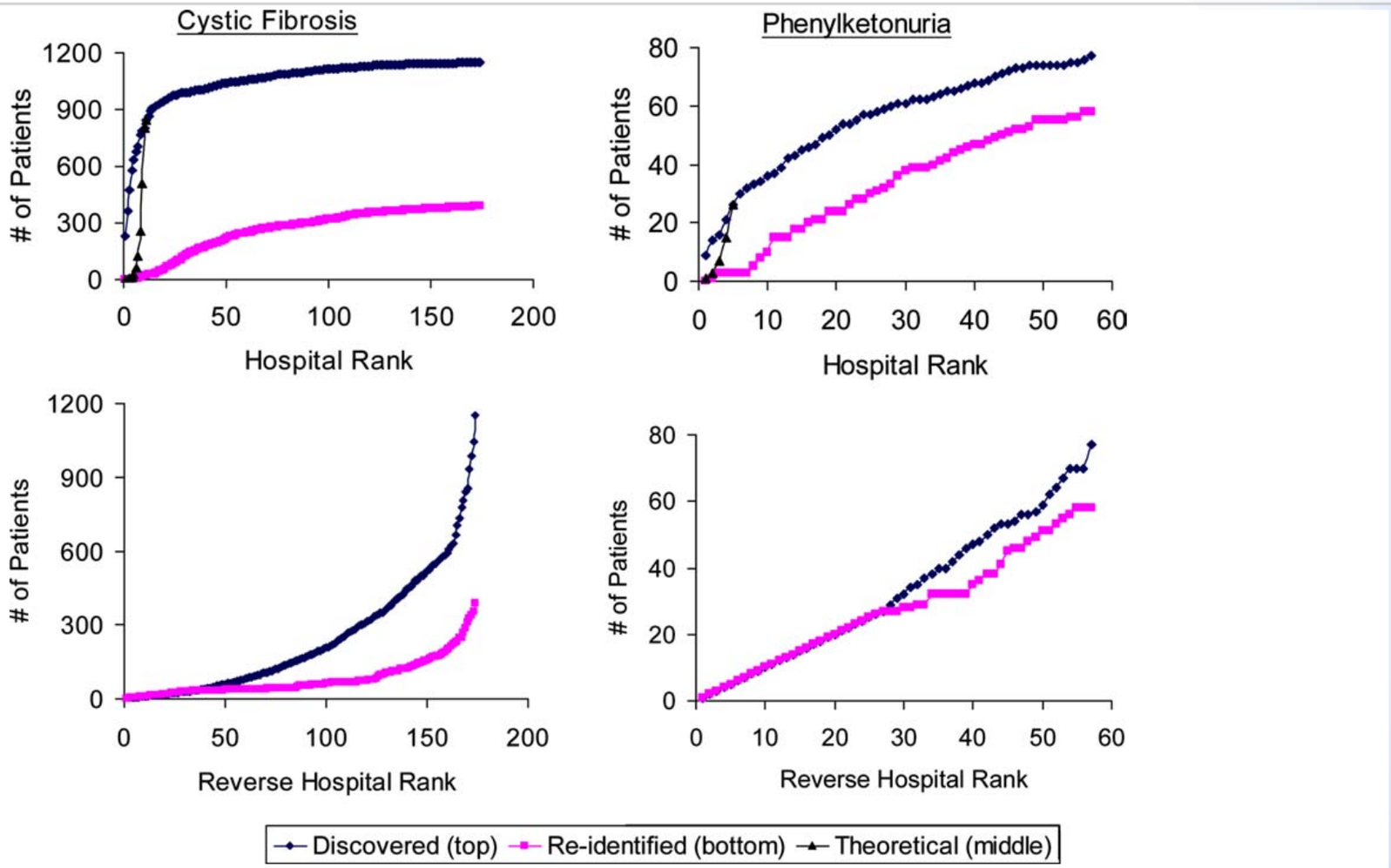

Fig. 6. REIDIT-C re-identification as a function of hospital rank by visit popularity; (first row) in order, (second row) reverse order. Hospital visit popularity is measured as the total number of unique visiting patients. The higher the order in the rank, the greater the popularity of a location. The "discovered" curve is the number of unique identified patients and unique DNA samples found in the set of locations up to rank x. The "re-identified" curve is the number of re-identifications made in the trails constructed over the set of considered locations. The "theoretical" curve is the maximum number of trails that could be re-identified given the number of locations and the number of trails observed.

Results (REIDIT-I)

As the probability of withholding information increases, the probability that an individual will not show up at all (i.e., no trail generated) in the population of incomplete trails. Thus, in the graphs we show three lines. The topmost line represents the number of non-null identified clinical data trails for a given set of hospitals. The middle line represents the number of non-null genomic data trails. And the lowest line represents the number of genomic data trails that were re-identified. As expected, we find that as the amount of information withheld increases, the number of releasing locations necessary to perform re-identification increases as well. This is due to the fact that as additional information is withheld, the incomplete trail becomes less complex and informative. However, even though trails become less complex, there remains a significant disposition toward re-identification. This is observable even after 50% of a trail is obscured. We find that there is an inverse relationship between the slope of re-identification (as a function of website rank) and the amount of information withheld.

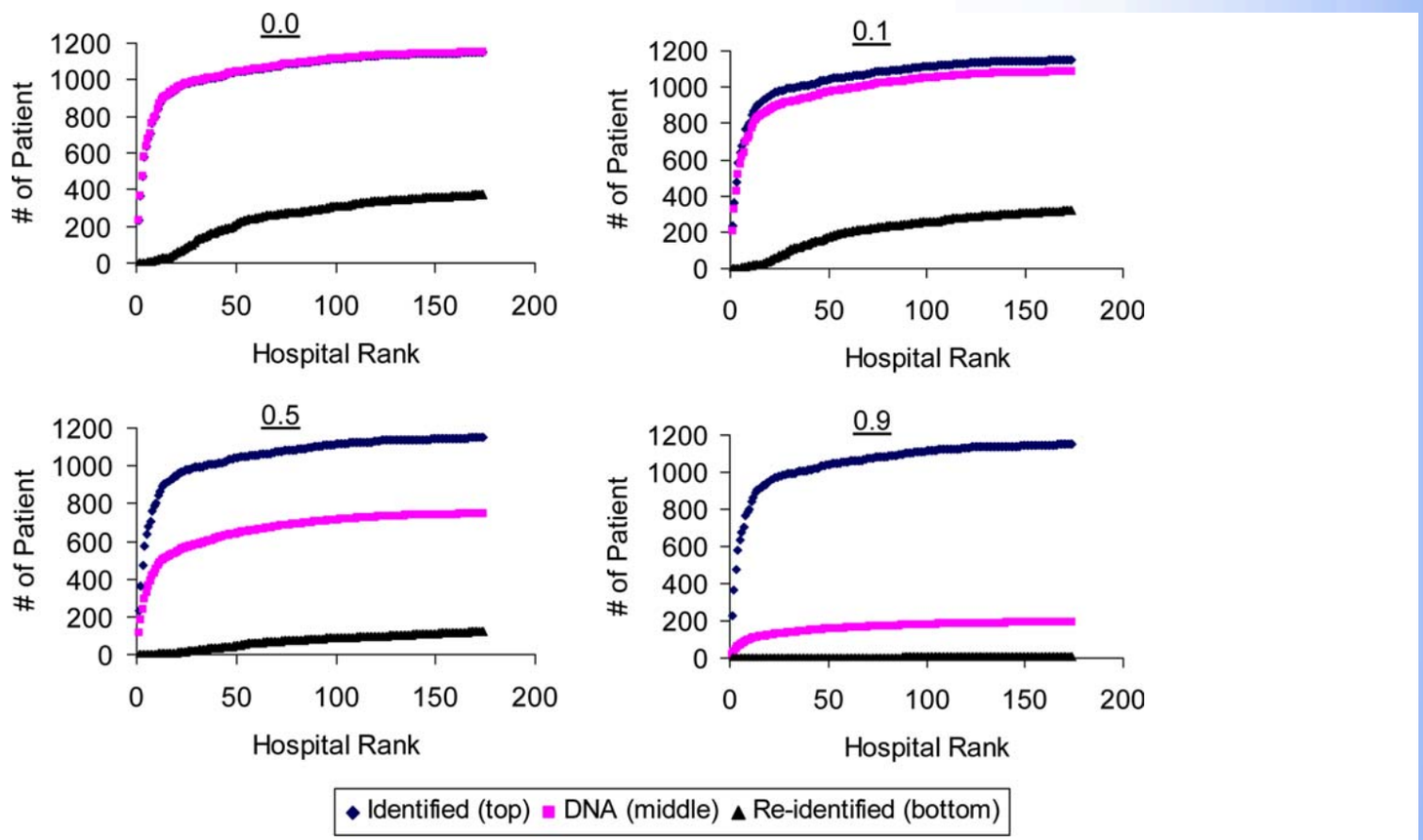

Fig. 7. Re-identification of CF incomplete trails with REIDIT-I as an increasing amount of identifying information is withheld from the release. From left to right: 0.0, 0.1, 0.5, and 0.9 probability of withholding. The ''identified'' and ''DNA'' curves correspond to the number of unique identified patients and unique DNA samples, respectively, discovered in the set of locations up to rank x. The ''re-identified'' curve represents the number of DNA samples re-identified to identified patients.

Solutions: Two proposals

- deMoor (et al): Central repository with double encryption (e.g., public/private keys for email)
  - Maintained by trusted 3rd party, encrypted at each end based on repository's algorithm and location's algorithm (based on some unique attributes).
- Protects personal information but does not protect DNA data from being tracked by quantity if not specific location (e.g. X number of hospitals visited), and thus still "trailed."

Solutions: Two proposals

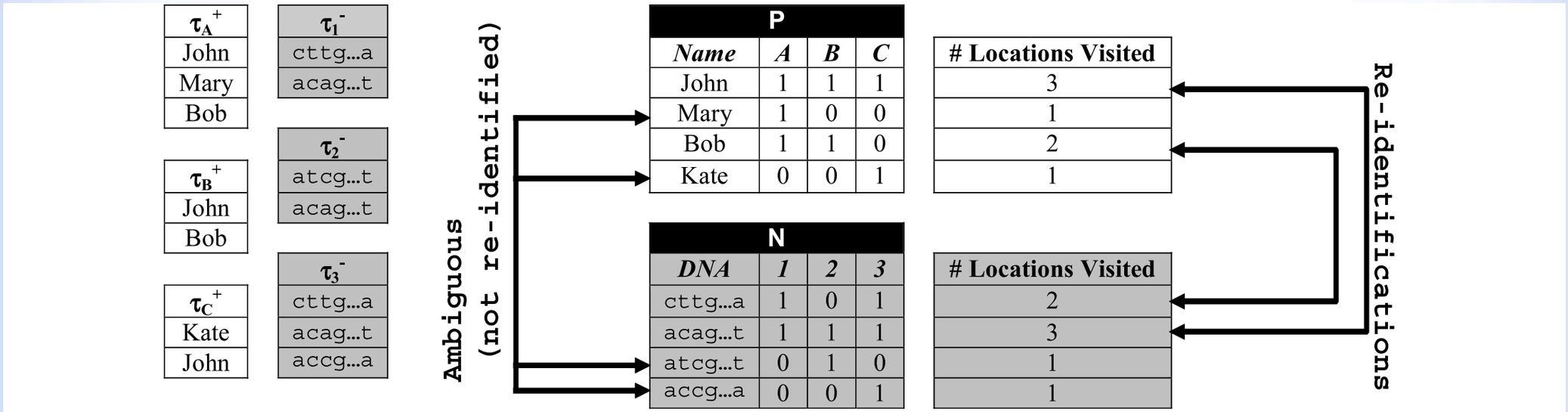

Fig. 8. (Left) Unreserved releases where locations are not identified. The subscripts A, B, and C for identified tables have no explicit correlation with subscripts 1, 2, and 3 of the DNA tables. (Right) Resulting identified (**P**) and DNA (**N**) tracks. Re-identifications are made through uniqueness in the number of locations visited.

Solutions: Two proposals

- deCODE Genetics:
  - A subset of individuals is identified for research potential. deCODE receives a blood sample and a strongly-encrypted social security number.
  - DNA is collected and annotated to only one collection, even though the individual can visit multiple locations. The identification information can be encrypted to multiple locations but the DNA cannot. So an individual's DNA cannot be "trailed" from location to location.

Future research

- This study assumes that only one type of data is withheld consistently (either genomic or identification data) in reserve situations. When this is mixed, then both reconstructed tracks would consist of incomplete trails. This can cause the REIDIT-I algorithm to fail and make misidentifications (or even false re-identifications, making the entire algorithm break).
- Recording linkage models, including probabilistic matching? If there is an understanding of the (differing) error rates between locations, this can match up otherwise disjointed records.