

# The Fellegi-Sunter Model of Record Linkage

---

Olivier Binette

August 26, 2020

# Agenda

---

**Today's goal:** Introduce the Fellegi-Sunter *model* of record linkage.

Focus on:

- Motivating example and some fundamental ideas.
- Databases, records and attributes.
- Pairwise record comparisons.
- Model parameters:  $m$  and  $u$  distributions, matching weights, and matching configuration matrix.

Afterwards:

- Model estimation and linkage rules.
- Sadinle (2017) and McVeigh (2020) go into more detail here.

**Today's goal:** Introduce the Fellegi-Sunter *model* of record linkage.

Focus on:

- Motivating example and some fundamental ideas.
- Databases, records and attributes.
- Pairwise record comparisons.
- Model parameters:  $m$  and  $u$  distributions, matching weights, and matching configuration matrix.

Afterwards:

- Model estimation and linkage rules.
- Sadinle (2017) and McVeigh (2020) go into more detail here.

**Today's goal:** Introduce the Fellegi-Sunter *model* of record linkage.

Focus on:

- Motivating example and some fundamental ideas.
- Databases, records and attributes.
- Pairwise record comparisons.
- Model parameters:  $m$  and  $u$  distributions, matching weights, and matching configuration matrix.

Afterwards:

- Model estimation and linkage rules.
- Sadinle (2017) and McVeigh (2020) go into more detail here.

**Today's goal:** Introduce the Fellegi-Sunter *model* of record linkage.

Focus on:

- Motivating example and some fundamental ideas.
- Databases, records and attributes.
- Pairwise record comparisons.
- Model parameters:  $m$  and  $u$  distributions, matching weights, and matching configuration matrix.

Afterwards:

- Model estimation and linkage rules.
- Sadinle (2017) and McVeigh (2020) go into more detail here.

## Outline:

1. Newcombe et al. (1959): “Automatic Linkage of Vital Records”
  - Provides a motivating example and introduces key ideas.
2. Fellegi and Sunter (1969): “A Theory for Record Linkage”
  - I will focus on the record linkage *model* rather than the theory.
  - I will deviate a little bit from the original paper in order to introduce the key supplemental idea of “Bayesian FS.”

References: see “(Almost) All of Entity Resolution”

1. pp. 11 — 13
2. pp. 13 — 19

## Outline:

1. Newcombe et al. (1959): “Automatic Linkage of Vital Records”
  - Provides a motivating example and introduces key ideas.
2. Fellegi and Sunter (1969): “A Theory for Record Linkage”
  - I will focus on the record linkage *model* rather than the theory.
  - I will deviate a little bit from the original paper in order to introduce the key supplemental idea of “Bayesian FS.”

**References:** see “(Almost) All of Entity Resolution”

1. pp. 11 — 13
2. pp. 13 — 19



## Newcombe et al. (1959)

---

Newcombe et al. (1959). Published in *Science*:

## Automatic Linkage of Vital Records\*

**Computers can be used to extract “follow-up”  
statistics of families from files of routine records.**

**H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James**

# Newcombe's Automatic Linkage of Vital Records

## What they did:

- Introduced an automatic (probabilistic) record linkage technique and implemented it on the Datatron 205 computer.

## Two things here:

- Stated record linkage as a statistical problem and proposed the first unsupervised probabilistic RL approach.
- They showed that a computer could be programmed to perform RL.

# Newcombe's Automatic Linkage of Vital Records

## What they did:

- Introduced an automatic (probabilistic) record linkage technique and implemented it on the Datatron 205 computer.

## Two things here:

- Stated record linkage as a statistical problem and proposed the first unsupervised probabilistic RL approach.
- They showed that a computer could be programmed to perform RL.

# Newcombe's Automatic Linkage of Vital Records

**Their applied goal:** to link **34,138 birth records** from 1955 in British Columbia to **114,471 marriage records** in the preceding ten year period.

	Marriage record	Birth record
Husband's family name	Ayad	Ayot
Wife's family name	Barr	Barr
Husband's initials	J Z	J Z
Wife's initials	M T	B T
Husband's birth province	AB	AB
Wife's birth province	PE	PE

**Table 1:** Example of identify information from compared marriage and birth records. This is adapted and translated from Table I of Newcombe (1969). AB and PE represent the Canadian provinces of Alberta and Prince Edward Island.

# Newcombe's Automatic Linkage of Vital Records

## Newcombe's algorithm:

1. Sort records by the Soundex coding of family names.
2. Where Soundex coding agrees, and informal likelihood ratio test determines whether or not to link.

## Soundex coding:

- Olivier  $\rightarrow$  O416
- Oliver  $\rightarrow$  O416
- Olivia  $\rightarrow$  O410
- Rebecca  $\rightarrow$  R120
- Rebbeka  $\rightarrow$  R120
- Beka  $\rightarrow$  B200

# Newcombe's Automatic Linkage of Vital Records

## Newcombe's algorithm:

1. Sort records by the Soundex coding of family names.
2. Where Soundex coding agrees, and informal likelihood ratio test determines whether or not to link.

## Soundex coding:

- Olivier → O416
- Oliver → O416
- Olivia → O410
- Rebecca → R120
- Rebbeka → R120
- Beka → B200

# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."



# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."

# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."

# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."

# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."

# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."

# Newcombe's Automatic Linkage of Vital Records

## Likelihood ratio test:

- Imagine that two records agree on the husband's first initial J.
- Let  $p_L$  be the probability of this given that the records are actually a match, and let  $p_F$  be the probability of this given that the records are not a match.
- Let  $p_R$  be the proportion of the initial "J" among husbands.

Then

$$p_L \approx p_R, \quad p_F \approx p_R^2$$

so

$$\log(p_L/p_F) \approx -\log(p_R).$$

This is the "matching weight."

## Likelihood ratio test (cont'd):

- If the initial is very common, e.g.  $p_R = 0.1$ , then

$$\log(p_L/p_F) \approx -\log(0.1) \approx 2.3$$

is weights in a little bit for a match.

- If the initial is not at all common, e.g.  $p_R = 0.0001$ , then

$$\log(p_L/p_F) \approx -\log(0.0001) \approx 9.2$$

weights in much more in favor of a match.

## Likelihood ratio test (cont'd):

- If the initial is very common, e.g.  $p_R = 0.1$ , then

$$\log(p_L/p_F) \approx -\log(0.1) \approx 2.3$$

is weights in a little bit for a match.

- If the initial is not at all common, e.g.  $p_R = 0.0001$ , then

$$\log(p_L/p_F) \approx -\log(0.0001) \approx 9.2$$

weights in much more in favor of a match.



# Newcombe's Automatic Linkage of Vital Records

## Performance:

- Processed 10 record pairs per minute.
- About 98.3% of the true matches were detected, and about 0.7% of the linked records were not actual matches.
- “by far the largest part of the effort” was the preparation of punched card files reproducing marriage records in an adequate format.

## Caveat:

- Not clear how exactly the probabilities for the likelihood ratio test were computed in all cases.

# Newcombe's Automatic Linkage of Vital Records

## Performance:

- Processed 10 record pairs per minute.
- About 98.3% of the true matches were detected, and about 0.7% of the linked records were not actual matches.
- “by far the largest part of the effort” was the preparation of punched card files reproducing marriage records in an adequate format.

## Caveat:

- Not clear how exactly the probabilities for the likelihood ratio test were computed in all cases.

## The Fellegi-Sunter model

---

# The Fellegi-Sunter model

Fellegi and Sunter (1969). Published in JASA:

## A THEORY FOR RECORD LINKAGE\*

IVAN P. FELLEGI AND ALAN B. SUNTER

*Dominion Bureau of Statistics*

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

# The Fellegi-Sunter model

## What this paper does:

- It formalizes the approach of Newcombe et al. (1959) in a decision-theoretic framework.

Given a pair of records, it considers three possible actions:

- to *link* them;
- to call them a *possible link*; or
- to *not link* them.

An “optimal” decision rule is proposed for this.

Here I’m focusing on the model rather than the decision-theoretic framework.

# The Fellegi-Sunter model

## What this paper does:

- It formalizes the approach of Newcombe et al. (1959) in a decision-theoretic framework.

Given a pair of records, it considers three possible actions:

- to *link* them;
- to call them a *possible link*; or
- to *not link* them.

An “optimal” decision rule is proposed for this.

Here I'm focusing on the model rather than the decision-theoretic framework.

# The Fellegi-Sunter model

## What this paper does:

- It formalizes the approach of Newcombe et al. (1959) in a decision-theoretic framework.

Given a pair of records, it considers three possible actions:

- to *link* them;
- to call them a *possible link*; or
- to *not link* them.

An “optimal” decision rule is proposed for this.

Here I’m focusing on the model rather than the decision-theoretic framework.

# The Fellegi-Sunter model

## Basic elements:

- Two *databases*  $A$  and  $B$ 
  - Duplication *across* but not within databases (bipartite record linkage).
- *Records* with corresponding *attributes* or *fields*
  - Name, age, address, SSN, etc.



# The Fellegi-Sunter model

## What we want to do:

- Figure out which records refer to the same **entity** (a *person*, *object* or *event*.)

## How we'll do that:

- We'll compare records in pairs from databases  $A$  and  $B$ , as to obtain multidimensional measures of similarity.
- Based on the measures of similarity, we'll try to group together the records which refer to the same entity.

# The Fellegi-Sunter model

## What we want to do:

- Figure out which records refer to the same **entity** (a *person*, *object* or *event*.)

## How we'll do that:

- We'll compare records in pairs from databases  $A$  and  $B$ , as to obtain multidimensional measures of similarity.
- Based on the measures of similarity, we'll try to group together the records which refer to the same entity.

# The Fellegi-Sunter model

	Field 1	Field 2	Field 3
Record no.	First name	Last name	Age
1	Olivier	Binette	25
2	Peter	Hoff	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N_1$	Beka	Steorts	NA

	Field 1	Field 2	Field 3
Record no.	First name	Last name	Age
1	Oliver	Binette	NA
2	Brian	K	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N_2$	Frances	Hung	NA

# The Fellegi-Sunter model

Let  $i = 1, 2, \dots, N_1 \times N_2$  enumerate the set of all record pairs in  $A \times B$ .

## Comparison vectors:

- For the  $i$ th pair of records, we compute a corresponding *comparison vector*

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \dots, \gamma_i^{(k)}).$$

- Each  $\gamma_i^j$  compares the  $j$ th field of the records.
- For example, if the  $j$ th field is “age,” we could have  $\gamma_i^j = 0$  if ages are the same, and  $\gamma_i^j = 1$  if ages different.

# The Fellegi-Sunter model

Let  $i = 1, 2, \dots, N_1 \times N_2$  enumerate the set of all record pairs in  $A \times B$ .

## Comparison vectors:

- For the  $i$ th pair of records, we compute a corresponding *comparison vector*

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \dots, \gamma_i^{(k)}).$$

- Each  $\gamma_i^j$  compares the  $j$ th field of the records.
- For example, if the  $j$ th field is “age,” we could have  $\gamma_i^j = 0$  if ages are the same, and  $\gamma_i^j = 1$  if ages different.

# The Fellegi-Sunter model

Let  $i = 1, 2, \dots, N_1 \times N_2$  enumerate the set of all record pairs in  $A \times B$ .

## Comparison vectors:

- For the  $i$ th pair of records, we compute a corresponding *comparison vector*

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \dots, \gamma_i^{(k)}).$$

- Each  $\gamma_i^j$  compares the  $j$ th field of the records.
- For example, if the  $j$ th field is “age,” we could have  $\gamma_i^j = 0$  if ages are the same, and  $\gamma_i^j = 1$  if ages different.

# The Fellegi-Sunter model

Let  $i = 1, 2, \dots, N_1 \times N_2$  enumerate the set of all record pairs in  $A \times B$ .

## Comparison vectors:

- For the  $i$ th pair of records, we compute a corresponding *comparison vector*

$$\gamma_i = (\gamma_i^{(1)}, \gamma_i^{(2)}, \dots, \gamma_i^{(k)}).$$

- Each  $\gamma_i^j$  compares the  $j$ th field of the records.
- For example, if the  $j$ th field is “age,” we could have  $\gamma_i^j = 0$  if ages are the same, and  $\gamma_i^j = 1$  if ages different.

# The Fellegi-Sunter model

## Binary comparisons:

- $\gamma_i^j \in \{0, 1\}$

## Levels of agreement/disagreement:

- $\gamma_i^j \in \{0, 1, 2, \dots, L_j\}$

## How they're obtained:

- You choose!
- Use string distance functions to compare names.



# The Fellegi-Sunter model

## Binary comparisons:

- $\gamma_i^j \in \{0, 1\}$

## Levels of agreement/disagreement:

- $\gamma_i^j \in \{0, 1, 2, \dots, L_j\}$

## How they're obtained:

- You choose!
- Use string distance functions to compare names.

# The Fellegi-Sunter model

## Binary comparisons:

- $\gamma_i^j \in \{0, 1\}$

## Levels of agreement/disagreement:

- $\gamma_i^j \in \{0, 1, 2, \dots, L_j\}$

## How they're obtained:

- You choose!
- Use string distance functions to compare names.

# The Fellegi-Sunter model

The set  $\{\gamma_k\}_{j=1}^{N_1 \times N_2}$  of computed comparison vectors becomes the **observed data** for the Fellegi-Sunter model.

Next component of the model:

- The **matching configuration**  $r = \{r_j\}_{j=1}^{N_1 \times N_2}$ , with  $r_j = 1$  if the  $j$ th record pair matches, and  $r_j = 0$  otherwise.
  - This is the adjacency list representation. We can also use a matching configuration matrix.
- This is not a very efficient representation for bipartite matching. Saindle (2017) instead uses a *matching labeling*.

# The Fellegi-Sunter model

The set  $\{\gamma_k\}_{j=1}^{N_1 \times N_2}$  of computed comparison vectors becomes the **observed data** for the Fellegi-Sunter model.

## Next component of the model:

- The **matching configuration**  $r = \{r_j\}_{j=1}^{N_1 \times N_2}$ , with  $r_j = 1$  if the  $j$ th record pair matches, and  $r_j = 0$  otherwise.
  - This is the adjacency list representation. We can also use a matching configuration matrix.
- This is not a very efficient representation for bipartite matching. Saindle (2017) instead uses a *matching labeling*.

# The Fellegi-Sunter model

- For record pairs that are a *match* ( $r_j = 1$ ), we assume that  $\gamma \sim m$  independently.
- For record pairs that are *unmatched* ( $r_j = 0$ ), we assume that  $\gamma \sim u$  independently.
- More precisely,

$$p\left(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u\right) = \left(\prod_{j:r_j=1} m(\gamma_j)\right) \times \left(\prod_{j:r_j=0} u(\gamma_j)\right).$$

# The Fellegi-Sunter model

- For record pairs that are a *match* ( $r_j = 1$ ), we assume that  $\gamma \sim m$  independently.
- For record pairs that are *unmatched* ( $r_j = 0$ ), we assume that  $\gamma \sim u$  independently.
- More precisely,

$$p\left(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u\right) = \left(\prod_{j:r_j=1} m(\gamma_j)\right) \times \left(\prod_{j:r_j=0} u(\gamma_j)\right).$$

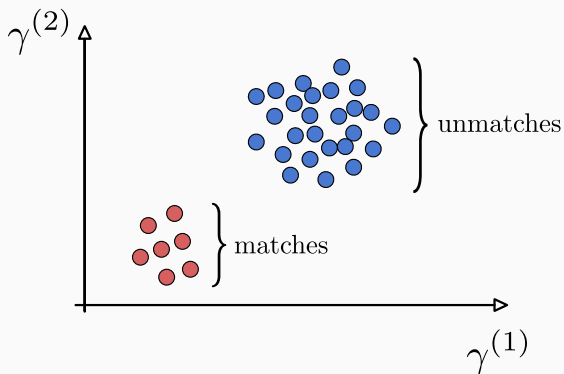
# The Fellegi-Sunter model

- For record pairs that are a *match* ( $r_j = 1$ ), we assume that  $\gamma \sim m$  independently.
- For record pairs that are *unmatched* ( $r_j = 0$ ), we assume that  $\gamma \sim u$  independently.
- More precisely,

$$p\left(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u\right) = \left(\prod_{j:r_j=1} m(\gamma_j)\right) \times \left(\prod_{j:r_j=0} u(\gamma_j)\right).$$

# The Fellegi-Sunter model

$$p\left(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u\right) = \left(\prod_{j:r_j=1} m(\gamma_j)\right) \times \left(\prod_{j:r_j=0} u(\gamma_j)\right).$$





# The Fellegi-Sunter model

What's left to do?

- Estimate model parameters.
- Define a prior  $p(r, m, u)$ .
- Obtain a posterior

$$\begin{aligned} p(r \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) &= \int p(r, m, u \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) dm du \\ &\propto \int p(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u) p(r, m, u) dm du \end{aligned}$$

# The Fellegi-Sunter model

What's left to do?

- Estimate model parameters.
- Define a prior  $p(r, m, u)$ .
- Obtain a posterior

$$\begin{aligned} p(r \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) &= \int p(r, m, u \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) dm du \\ &\propto \int p(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u) p(r, m, u) dm du \end{aligned}$$

# The Fellegi-Sunter model

What's left to do?

- Estimate model parameters.
- Define a prior  $p(r, m, u)$ .
- Obtain a posterior

$$\begin{aligned} p(r \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) &= \int p(r, m, u \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) dm du \\ &\propto \int p(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u) p(r, m, u) dm du \end{aligned}$$

# The Fellegi-Sunter model

What's left to do?

- Estimate model parameters.
- Define a prior  $p(r, m, u)$ .
- Obtain a posterior

$$\begin{aligned} p(r \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) &= \int p(r, m, u \mid \{\gamma_j\}_{j=1}^{N_1 \times N_2}) dm du \\ &\propto \int p(\{\gamma_j\}_{j=1}^{N_1 \times N_2} \mid r, m, u) p(r, m, u) dm du \end{aligned}$$

# The Fellegi-Sunter model

- This is **not** what Fellegi-Sunter originally proposed
- Originally, FS proposed to estimate  $m$  and  $u$  on their own.
- Then, define the log-likelihood ratio (**matching weight**)

$$W(\gamma_j) = \log \frac{m(\gamma_j)}{u(\gamma_j)}.$$

- Say that the  $j$ th pair is a match if  $W(\gamma_j)$  is large, that they're not a match if  $W(\gamma_j)$  is small: this is a likelihood ratio test.

# The Fellegi-Sunter model

- This is **not** what Fellegi-Sunter originally proposed
- Originally, FS proposed to estimate  $m$  and  $u$  on their own.
- Then, define the log-likelihood ratio (**matching weight**)

$$W(\gamma_j) = \log \frac{m(\gamma_j)}{u(\gamma_j)}.$$

- Say that the  $j$ th pair is a match if  $W(\gamma_j)$  is large, that they're not a match if  $W(\gamma_j)$  is small: this is a likelihood ratio test.

# The Fellegi-Sunter model

- This is **not** what Fellegi-Sunter originally proposed
- Originally, FS proposed to estimate  $m$  and  $u$  on their own.
- Then, define the log-likelihood ratio (**matching weight**)

$$W(\gamma_j) = \log \frac{m(\gamma_j)}{u(\gamma_j)}.$$

- Say that the  $j$ th pair is a match if  $W(\gamma_j)$  is large, that they're not a match if  $W(\gamma_j)$  is small: this is a likelihood ratio test.

# The Fellegi-Sunter model

- This is **not** what Fellegi-Sunter originally proposed
- Originally, FS proposed to estimate  $m$  and  $u$  on their own.
- Then, define the log-likelihood ratio (**matching weight**)

$$W(\gamma_j) = \log \frac{m(\gamma_j)}{u(\gamma_j)}.$$

- Say that the  $j$ th pair is a match if  $W(\gamma_j)$  is large, that they're not a match if  $W(\gamma_j)$  is small: this is a likelihood ratio test.



What's the problem with the original FS approach?

- You consider all record pairs independently.
- You could link records  $a$  and  $b$ , and  $b$  and  $c$ , and yet say that  $a$  and  $c$  are not a match. This is incoherent.
- In the *bipartite record linkage* framework, we want to specify a prior on  $r$  which reflects the fact that there is duplication across but not within databases.

# The Fellegi-Sunter model

What's the problem with the original FS approach?

- You consider all record pairs independently.
- You could link records  $a$  and  $b$ , and  $b$  and  $c$ , and yet say that  $a$  and  $c$  are not a match. This is incoherent.
- In the *bipartite record linkage* framework, we want to specify a prior on  $r$  which reflects the fact that there is duplication across but not within databases.

# The Fellegi-Sunter model

What's the problem with the original FS approach?

- You consider all record pairs independently.
- You could link records  $a$  and  $b$ , and  $b$  and  $c$ , and yet say that  $a$  and  $c$  are not a match. This is incoherent.
- In the *bipartite record linkage* framework, we want to specify a prior on  $r$  which reflects the fact that there is duplication across but not within databases.

# The Fellegi-Sunter model

What's the problem with the original FS approach?

- You consider all record pairs independently.
- You could link records  $a$  and  $b$ , and  $b$  and  $c$ , and yet say that  $a$  and  $c$  are not a match. This is incoherent.
- In the *bipartite record linkage* framework, we want to specify a prior on  $r$  which reflects the fact that there is duplication across but not within databases.

# Summary

---

- Newcombe (1958) proposed a likelihood ratio test approach to record linkage based on probability heuristics.
- I've introduced the very basic components of the Fellegi-Sunter model
- Sadinle (2017) and McVeigh (2020) provide information about the priors and about model fitting.

# Summary

- Newcombe (1958) proposed a likelihood ratio test approach to record linkage based on probability heuristics.
- I've introduced the very basic components of the Fellegi-Sunter model
- Sadinle (2017) and McVeigh (2020) provide information about the priors and about model fitting.

- Newcombe (1958) proposed a likelihood ratio test approach to record linkage based on probability heuristics.
- I've introduced the very basic components of the Fellegi-Sunter model
- Sadinle (2017) and McVeigh (2020) provide information about the priors and about model fitting.