Latent Class Models

- In Dunson's STA 841 class "Categorical Data Analysis," we learned about a *latent class models*
- In class, we used to them to classify DNA sequences, but I think they may be effective in classifying individuals based on their travel path in a network
- DISCLAIMER: I see many shortcomings of the model, but just wanted to share it as a potential idea

Data

- Individuals $i \in \{1, \ldots, I\}$
- Nucleotide position $p \in \{1, \ldots, P\}$
- Nucleotide expression $y_{ip} = d \in \{1, \ldots, D\}$
- Latent class $z_i = k \in \{1, \ldots, K\}$. Number of classes K can be fixed or estimated nonparametrically
- Vector of expression probabilities $\theta_{kp} = (\theta_{kpA}, \ldots, \theta_{kpT})$ for each class and position

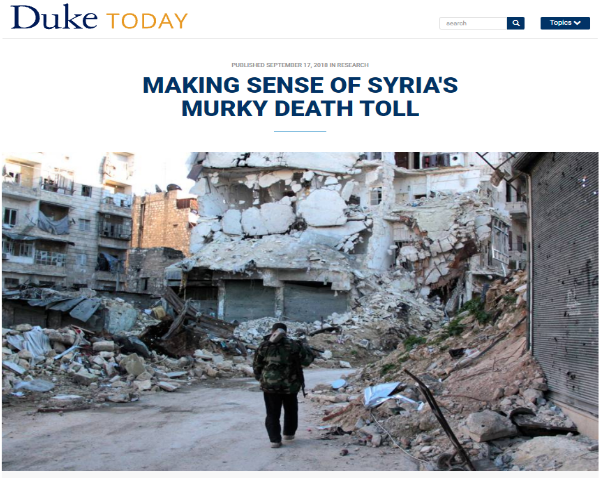
Data

- Dataset consists of DNA sequences and expert classification as IE, EI, or N
- We attempt to recover the given labels through the latent class model

Model Formulation

- $y_{ip} \sim \text{Multinomial}(\theta_{z_i,p})$
- $z_i \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$
- $\theta_{z_i,p} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_D)$
- z_i can also be modeled through a Dirichlet process stick-breaking prior so that the number of clusters is inferred through the data

Record Linkage in Practice

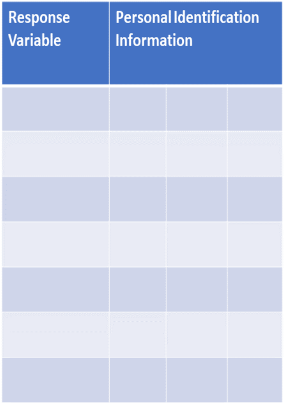


Duke TODAY

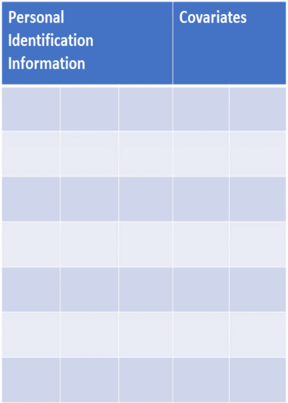
PUBLISHED SEPTEMBER 17, 2018 IN RESEARCH

MAKING SENSE OF SYRIA'S MURKY DEATH TOLL

Linkage for Downstream Analysis

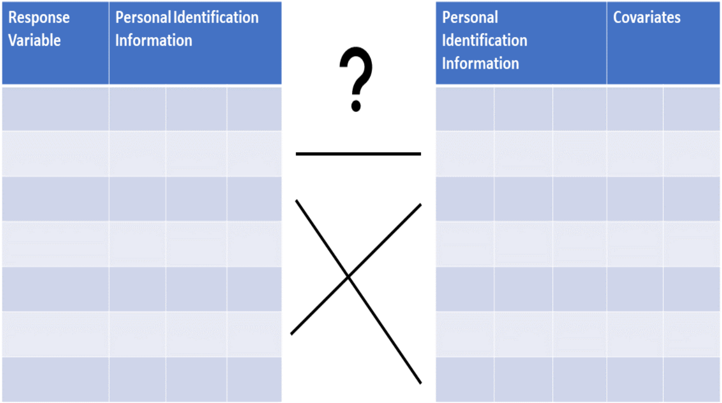


Response Variable	Personal Identification Information		



Personal Identification Information			Covariates	

Linkage for Downstream Analysis



Response Variable

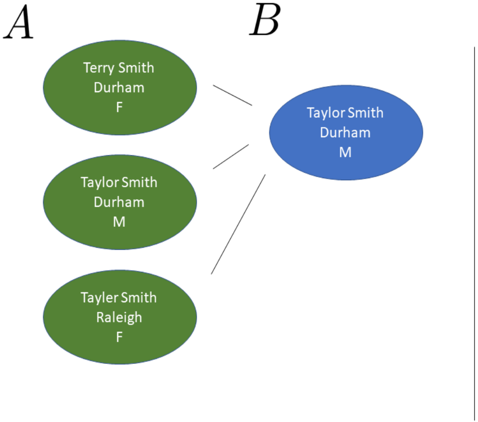
Personal Identification Information

?

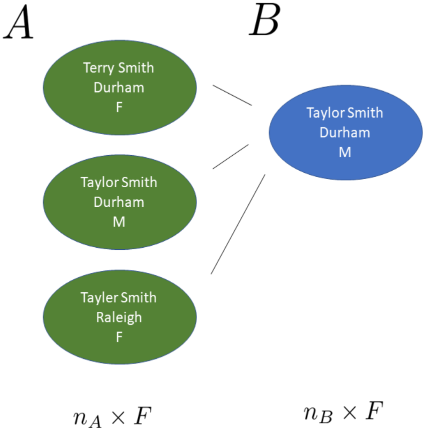
Personal Identification Information

Covariates

Linkage through Comparison Vectors

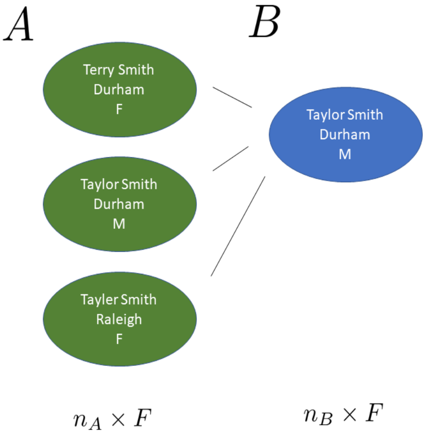


Linkage through Comparison Vectors



- n_A, n_B records in A, B
- $F = 4$ features for comparison
  - First name
  - Last name
  - City
  - Gender
- $L = \{3, 3, 2, 2\}$ levels of comparison

Linkage through Comparison Vectors



A

Terry Smith
Durham
F

Taylor Smith
Durham
M

Tayler Smith
Raleigh
F

$n_A \times F$

B

Taylor Smith
Durham
M

$n_B \times F$

Γ

$$\gamma_{11} = [3, 1, 1, 2]$$

$$\gamma_{21} = [1, 1, 1, 1]$$

$$\gamma_{31} = [2, 1, 2, 2]$$

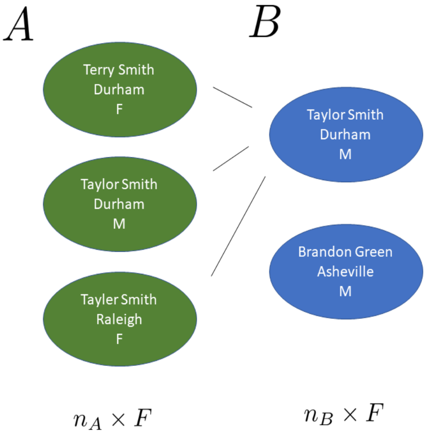
$n_A n_B \times F$

Linkage through Comparison Vectors

Represent linkage structure through vector $\mathbf{Z} = \{Z_1, \ldots, Z_{n_B}\}$, where

$$Z_j = \begin{cases} i, & \text{if records } i \in A \text{ and } j \in B \text{ match;} \\ n_A + 1, & \text{if record } j \in B \text{ has no match in } A; \end{cases}$$

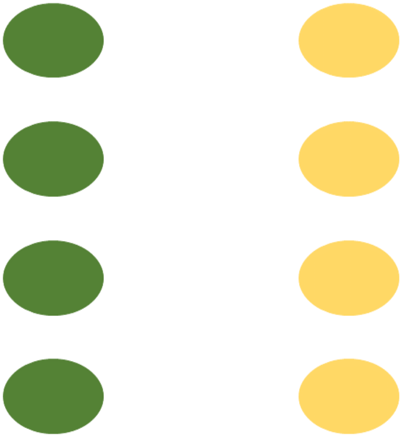
Linkage through Comparison Vectors



$$Z_1 = 2$$

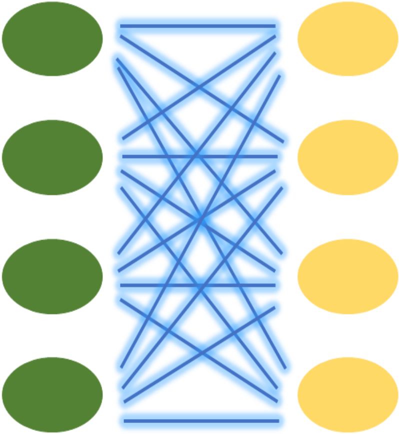
$$Z_2 = n_A + 1$$

Fellegi and Sunter (1969)

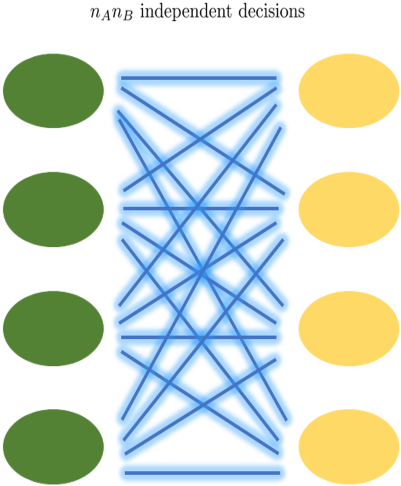


Fellegi and Sunter (1969)

$n_A n_B$ independent decisions



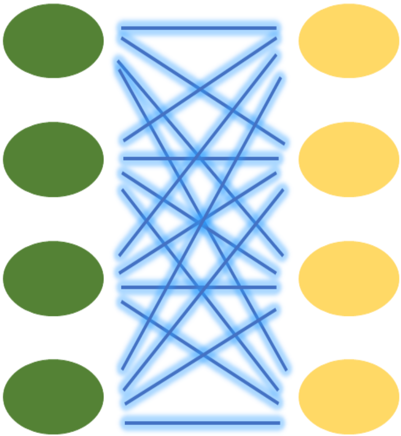
Fellegi and Sunter (1969)



- scalable to large datasets (`fastlink`, Enamorado et al 2019)

Fellegi and Sunter (1969)

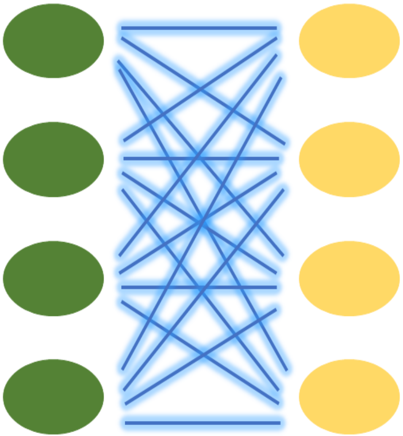
$n_A n_B$ independent decisions



- scalable to large datasets (`fastlink`, Enamorado et al 2019)
- not bipartite, requires post-processing

Fellegi and Sunter (1969)

$n_A n_B$ independent decisions

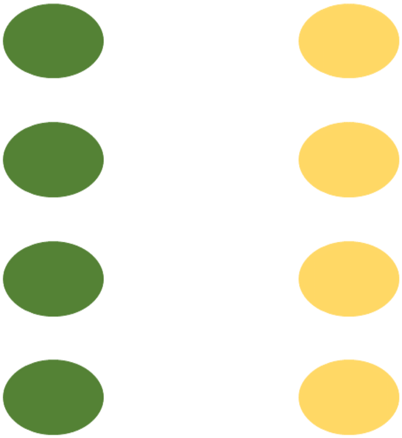


- scalable to large datasets (`fastlink`, Enamorado et al 2019)
- not bipartite, requires post-processing
- overmatches, leading to inaccurate parameter estimation

Fellegi and Sunter (1969)

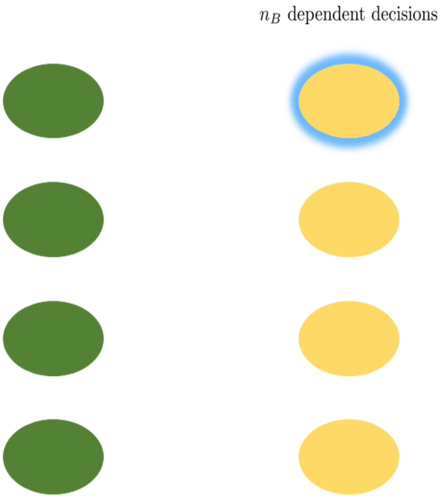
- The Fellegi and Sunter model is essentially a *mixture model*
- We posit that comparison vectors come from two different distributions, one for matching, and another for nonmatching record pairs
- The model learns the parameters for each distribution and independently classifies each record pair as matching or nonmatching.

Sadinle (2017)



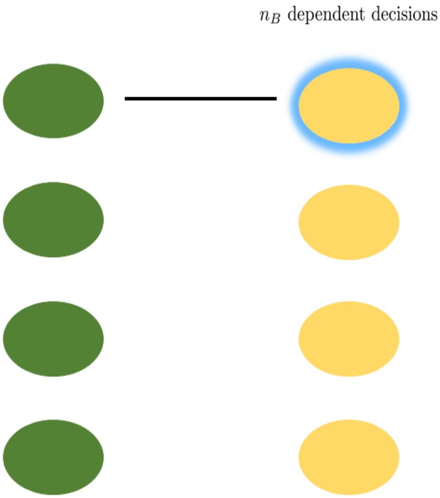
- Beta Record Linkage (BRL)

Sadinle (2017)



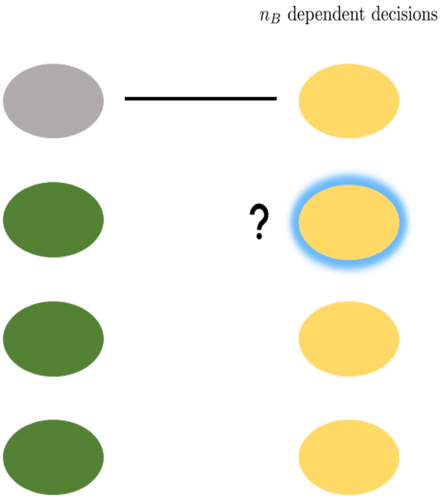
- Beta Record Linkage (BRL)

Sadinle (2017)



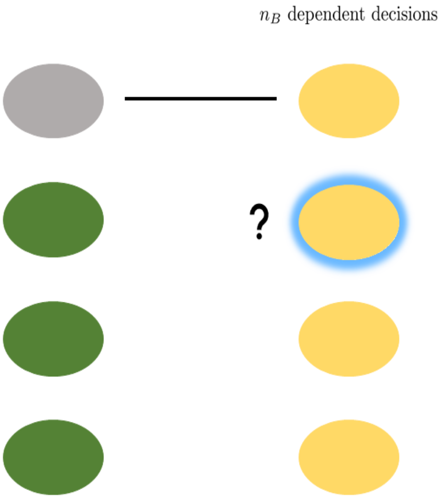
- Beta Record Linkage (BRL)

Sadinle (2017)



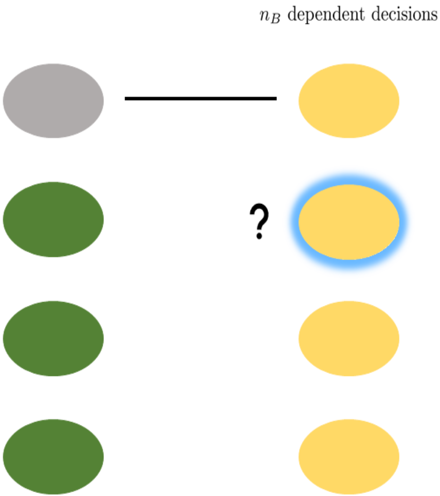
- Beta Record Linkage (BRL)

Sadinle (2017)



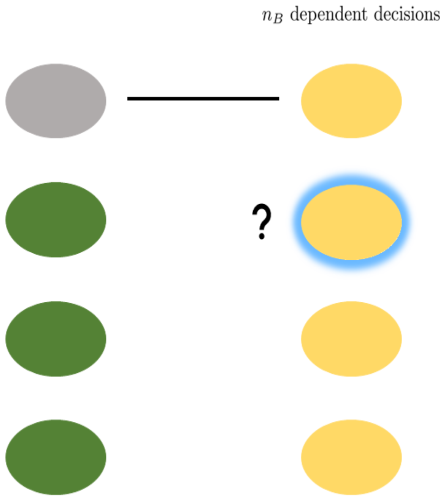
- Beta Record Linkage (BRL)
- strictly enforces one-to-one matching, no post-processing

Sadinle (2017)



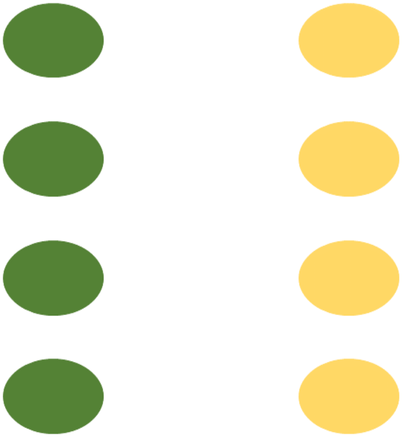
- Beta Record Linkage (BRL)
- strictly enforces one-to-one matching, no post-processing
- high accuracy for linkage and other parameters

Sadinle (2017)

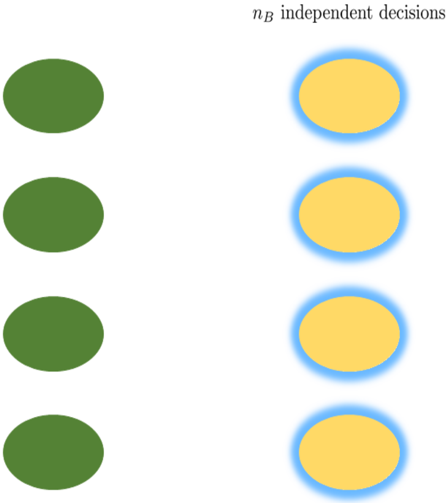


- Beta Record Linkage (BRL)
- strictly enforces one-to-one matching, no post-processing
- high accuracy for linkage and other parameters
- inherently serial, not scalable to large linkage tasks

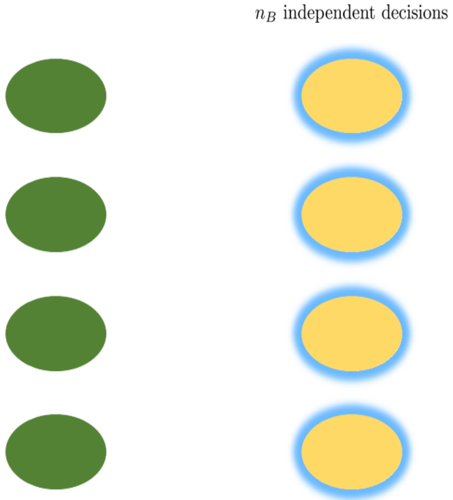
Our Contribution - Fast Beta Linkage



Our Contribution - Fast Beta Linkage

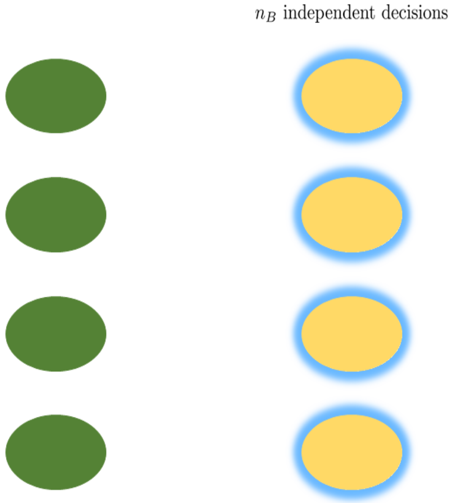


Our Contribution - Fast Beta Linkage



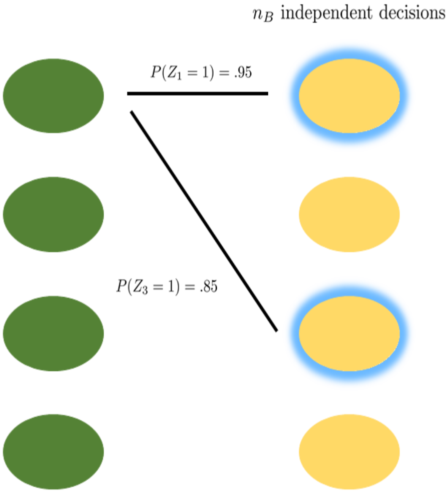
- relaxation proposed by Heck Wortman (2019)

Our Contribution - Fast Beta Linkage



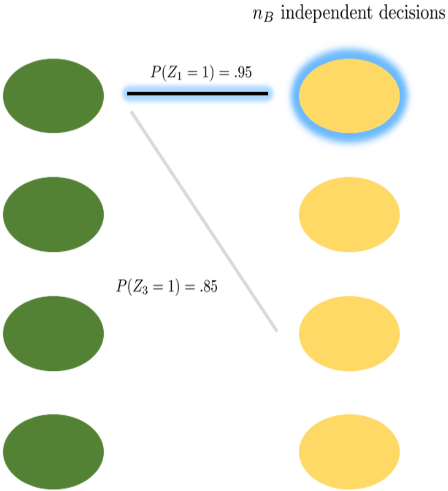
- relaxation proposed by Heck Wortman (2019)
- minimal loss of accuracy, large computational gains

Our Contribution - Fast Beta Linkage



- relaxation proposed by Heck Wortman (2019)
- minimal loss of accuracy, large computational gains
- allows for "one to many" matchings

Our Contribution - Fast Beta Linkage



- relaxation proposed by Heck Wortman (2019)
- minimal loss of accuracy, large computational gains
- allows for "one to many" matchings
- simple postprocessing to obtain bipartite matching

Model Parameters

- Both models utilize **m** and **u** parameters to model the data.
- Define $m_{f\ell} = P(\gamma_{ij}^{f} = \ell | \text{Records Match})$ and $u_{f\ell} = P(\gamma_{ij}^{f} = \ell | \text{Records Do Not Match})$
- We can think of **m** as reliability parameters, that show how likely it is for a feature to remain unchanged across files
- We can think of **u** as discernment parameters, that show how unique the values of a feature can be

Fellegi Sunter Weights

- We multiply these parameters for all f and the relevant ℓ to get $\mathbf{m}_{ij} = P(\gamma_{ij}|\text{Records Match})$ and $\mathbf{u}_{ij} = P(\gamma_{ij}|\text{Records Do Not Match})$
- These probabilities form the likelihood ratio

$$w_{ij} = \frac{P(\gamma_{ij}|\text{Records Match})}{P(\gamma_{ij}|\text{Records Do Not Match})}$$

- This ratio is large when $\mathbf{m}_{ij}$ is large and $\mathbf{u}_{ij}$ is small

Constructing Comparison Vectors

- We measure the distance between two fields using whatever metric we like, and then threshold this distance into discrete values
- We can use binary agreement, or use multiple agreement levels to capture partial agreement
- Existing packages perform binary agreement, numerical distance, and string distance

String Distance

- Many packages compare strings through the normalized Levenstein distance. This is the minimum number of edits required to transform one string to the other, divided by the length of the larger string.
- Additional metrics could be language specific Soundex codes
- Care should be taken when transforming metric into discreet agreement level. For example, should a Soundex match be regarded as a full match or a partial match?
- Code from existing packages can be edited to accomodate additional string distances

Conditionally Independent Features

- The record linkage model assumes that agreement levels are *conditionally independent*, given the matching status of the records. This is satisfied for most fields
- For example, knowing that two records agree on birth month does not tell you anything about whether they agree on birth year. These fields are independent
- In contrast, if two records disagree on department, we know they probably disagree on city. These fields are not independent
- Using dependent features is essentially using the same information multiple times, and can harm results

Conditionally Independent Features

- One solution to dependent features is to create a more nuanced comparison incorporating multiple features
- For example, you can define full agreement for "location" to be matching on city AND department, partial agreement to be matching on only department, and no agreement not be matching on neither.
- Current packages do not execute this type of comparison, but this can be done with some work
- Depending on the type of dependent fields, I may recommend keeping both or just omitting one.

Conditionally Independent Features

- One solution to dependent features is to create a more nuanced comparison incorporating multiple features
- For example, you can define full agreement for "location" to be matching on city AND department, partial agreement to be matching on only department, and no agreement not be matching on neither.
- Current packages do not execute this type of comparison, but this can be done with some work
- Depending on the type of dependent fields, I may recommend keeping both or just omitting one.

Random Sampling and Generalization

- To avoid calculating $n_A n_B$ comparison vectors, we can use *blocking*
- We only create comparison vectors when records agree on a particular field
- It may be tempting to block on a unique identifier, but this can create blocks that are too small for record linkage to run effectively

Thoughts for DANE

- With `fastLink`, you may be able to do the entire linkage task (60,000 × 1,000,000) records at one time. If not blocking by gender or by department should make it feasible
- You should not need to use the generalization technique
- `fastLink` currently does not do Soundex comparisons, but using allowing for partial agreement on strings should be sufficient. `fastLink` code can be modified to incorporate Soundex comparison, but may take some work

Performance Metrics with Known Labels

- When true matching status is known, we use two primary metrics:
  - $\text{Recall} = \frac{\text{True Matches Declared}}{\text{Total True Matches}}$
  - $\text{Precision} = \frac{\text{Matches Declared}}{\text{Total Matches Declared}}$
- To balance these two metrics, we sometimes also consider:
  - $\text{F-Measure} = \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$

Performance Metrics without Known Labels

- Without known labels, evaluating model performance is more difficult, but there are general checks you can make
  - The histogram of match probabilities should be clearly separated, with most record pairs having probabilities near 0 or near 1, and minimal number of pairs in between
  - Fellegi Sunter weights for impactful matching variables should be high. Weights for partial agreement should be lower than those for full agreement
  - Visually inspect a subset of matches and nonmatches. In particular, examine matches near the decision threshold.
- These approaches are demonstrated in the accompanying Rmarkdown file

Software

- fastLink can be found on GitHub
- BRL can also be found on GitHub
- fabl is not yet publicly available
- Due to the scale of DANE's record linkage task, I recommend you to use fastLink

References

- Fellegi I. and Sunter A., A theory for record linkage, J. Am. Stat. Assoc. 64 (1969), pp. 1183–1210
- Enamorado T., Fifield B. and Imai K., Using a probabilistic model to assist merging of large-scale administrative records, Am. Political Sci. Rev. 113 (2019), pp. 353–371
- Sadinle M., Bayesian estimation of bipartite matchings for record linkage, J. Am. Stat. Assoc. 112 (2017), pp. 600–612