Mal 4:6
Using Data Mining for Record Linkage

Burdette Pixton
Christophe Giraud-Carrier
March 24, 2005

Mal 4:6

- Mining And Linking FOR Successful Information eXchange
- Record Linkage is:
  - the process of identifying similar people
  - a necessary step in exchanging and merging pedigrees

Probabilistic Record Linkage

- Widely used
- Scores are given for similar attributes
- Scores are combined, and a threshold is used to determine a match
- Hand-crafted scores and thresholds
- High reliance on scores

Data Mining Approach

- Let the data tell us
  - How to score strings
  - Which data attributes to use (feature selection)
  - Which threshold works the best

String Metrics

- Used to determine the similarity between two strings
- Types of metrics
  - Edit distance
    - Cost to convert *s* to *t*
    - Character-by-character comparison
    - Levenstein
  - Similarity
    - Compares characters within a range
    - Attempts to look at the string as a whole
    - Jaro, Jaro-Winkler
  - Phonetic
    - Works well with words that “sound alike”
    - Very common with Genealogy Databases
    - Soundex

String Metrics

- Do some metrics work better on certain types of data?
  - Type of data to consider:
    - Names
    - Locations
    - Dates

Experiment Setup

- Genealogical database from the LDS Church's Family History Department (~5 million individuals)
- ~16,000 labeled data instances
  - <ID1><ID2><Match?>
  - Computed similarity scores across each field for each classification
  - Looked for highest score and largest difference

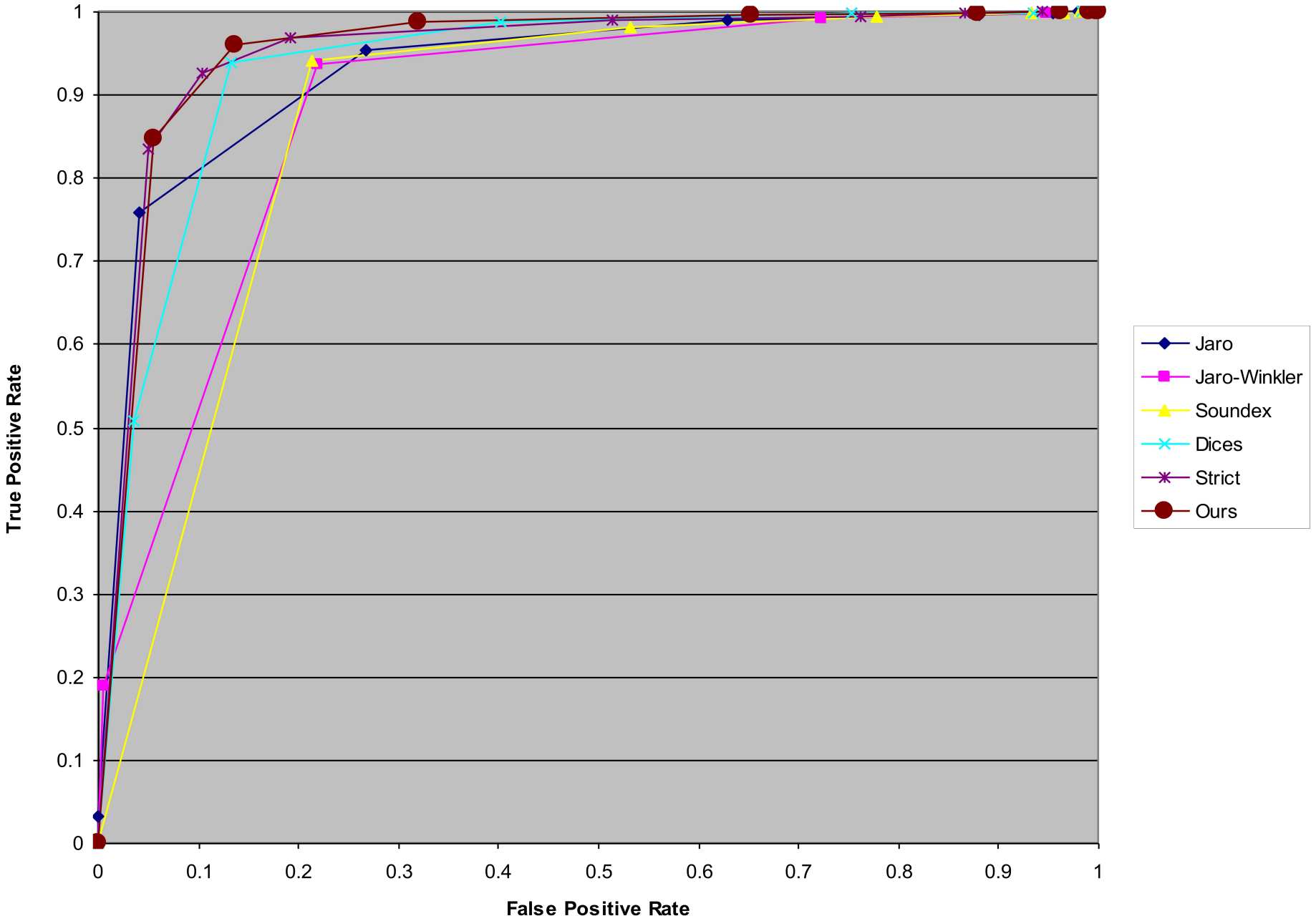
Results

Attribute Type	Metric
Gender	Binary Discrimination
Name	Soundex
Location	Jaro
Day	1-norm
Month	Dice
Year	1-norm

Experiment 2

- How does our composite metric compare against using a single approach?

$$D(x, y) = \frac{\sum_i D_i(x_i, y_i)}{N}$$


True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Jaro
Jaro-Winkler
Soundex
Dices
Strict
Ours

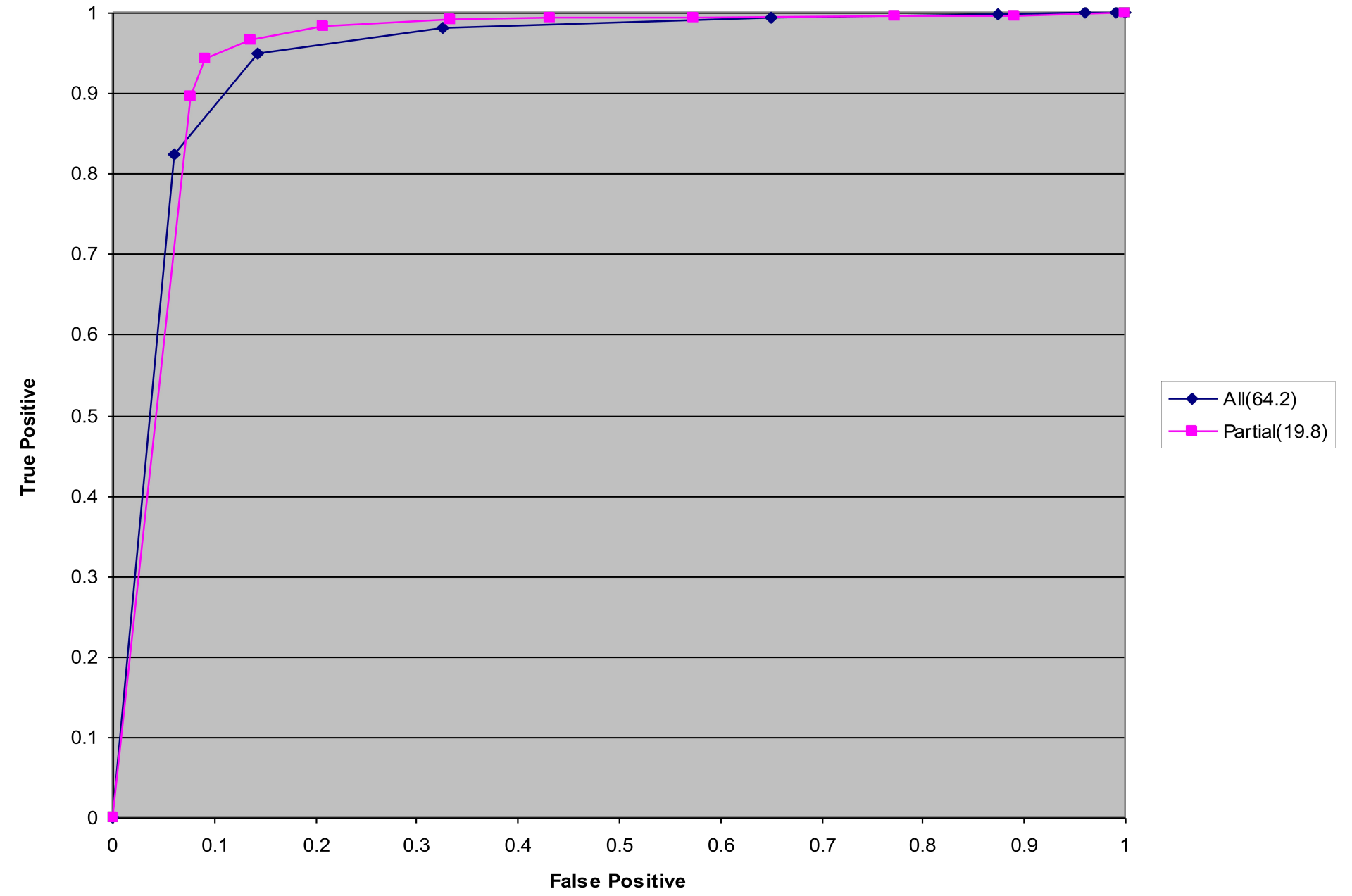
Graph Matching

- Pedigrees
  - Have explicit links
  - Show relationships between entities
- Mal 4:6 use these relationships
  - Pedigrees can be very large
  - Which relationships/attributes should we use?

Feature Selection

- Used a “scorecard” method

	Gender	1st name	Bdate	…
Self				…
Father				…
GrdFather				…
…	…	…	…	…


True Positive
False Positive
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
All(64.2)
Partial(19.8)

Graph Based

- Matches:
  - Individual only
    - Recall = 95.266, Precision = 71.799
  - 4 generations
    - Recall = 94.167, Precision = 71.766
- Mismatches
  - Individual only
    - Recall = 86.093, Precision = 98.641
  - 4 generations
    - Recall = 86.169, Precision = 98.358

Conclusions/Future Work

- Shows promise
- Improvements
  - Collect more data
    - Can we generate more?
    - Clean data
  - Sample Selection
    - 1:5
    - 1:n
  - Equal Weights
  - Pairwise similarity