Copy or Coincidence?
A Model for Detecting Social Influence and Duplication Events

Lisa Friedland, David Jensen (School of Computer Science),
Michael Lavine (Department of Math and Statistics)
University of Massachusetts Amherst

Download this paper at
http://goo.gl/2Ag7M

Motivation

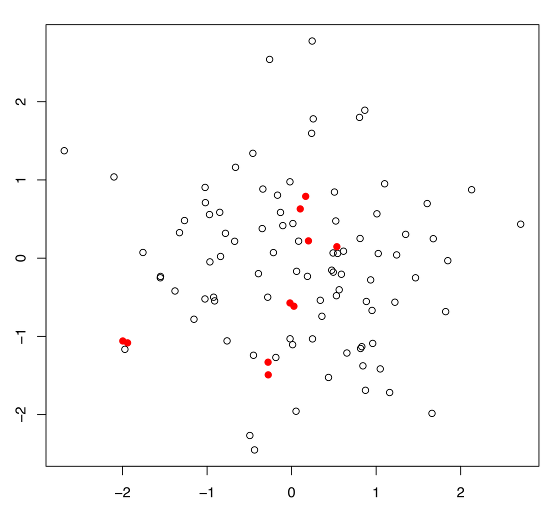
Example problems that look like this:

- Who knows each other?
  - Infer social ties based on co-located photographs or shared employment histories
- Are these really the same person?
  - Detect coalitions of click-fraud attackers
  - Determine whether a crime scene fingerprint has a match in a database
- Are these really the same entity?
  - Identify duplicate records to merge in a database

[Crandall et al., PNAS 2010; Friedland & Jensen, KDD 2007; Metwally et al., WWW 2007; Su & Srihari, NIPS 2010; Elmagarmid et al., TKDE 2007]

Drawbacks to existing solutions: often application-specific, non-probabilistic, or use only pair similarity

Task Formulation



Which of these points were generated in pairs?
(Ground truth pairs = red)

Identifying these pairs should be a function of the pairs'
- Similarity
- Rarity/Sparseness of region

Given a data set, calculate a score for every possible pair. Evaluate the ranking of pairs using AUC.

Goals

- Generic formulation: If we knew everything about a domain, how would this task be solved optimally?
- Towards realistic scenarios: will this method still be feasible…
  - When number of pairs or distances between pairs are unknown?
  - When data does not come from this model?
- Need for model: will a simple distance-only baseline be competitive with the model? If so, why and under what circumstances?

Findings

In the model system, for a given ϕ:

- **A single parameter, t, governs problem difficulty**. It describes how far apart positive pairs are compared to negative pairs.
  - $t \to 0 \Leftrightarrow$ mostly only distance matters
  - $t = \frac{1}{\sqrt{2}} \Leftrightarrow$ distance does not distinguish positive from negative pairs
- **When t is unknown,**
  - Guessing too low overweighs distance. But distance is a strong baseline, so it's only a mild drop-off.
  - Guessing too high overweighs rarity. Performance can get arbitrarily bad.
  - The approximation $\frac{P(\mathbf{d} \mid \varepsilon)}{P(\mathbf{m} \mid \phi)}$ is more robust than the optimal likelihood ratio
- **In real data sets,**
  - Task is moderately difficult: $t \approx 0.5$, and optimal LR is markedly better than distance-only.

Generative Mixture Model

(For continuous data in k dimensions)

Data is a mixture of

singletons: $\mathbf{x}_i \sim \phi$ — Distribution of most data

and pairs: $\mathbf{m} \sim \phi$, $\mathbf{d} \sim \varepsilon$, $\mathbf{x}_i = \mathbf{m} + \mathbf{d}$, $\mathbf{x}_j = \mathbf{m} - \mathbf{d}$ — Displacement distribution: ε = Gaussian(0, $v^2 I$)

generated to produce r pairs, all non-overlapping.

Inference

Estimate ϕ from the data itself. Guess ε.
Score each pair as if it were independent from the others. Likelihood ratio for a pair:

$$\frac{P(c_{ij}=1 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n)}{P(c_{ij}=0 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n)} \approx \frac{P(c_{ij}=1 \mid \mathbf{x}_i,\mathbf{x}_j)}{P(c_{ij}=0 \mid \mathbf{x}_i,\mathbf{x}_j)}$$

$$= \frac{P(\mathbf{x}_i,\mathbf{x}_j \mid c_{ij}=1)\,P(c_{ij}=1)}{P(\mathbf{x}_i,\mathbf{x}_j \mid c_{ij}=0)\,P(c_{ij}=0)}$$

$$\text{LR} = \frac{\frac{1}{2^k} P(\mathbf{m} \mid \phi)\,P(\mathbf{d} \mid \varepsilon)\,P(c_{ij}=1)}{P(\mathbf{x}_i \mid \phi)\,P(\mathbf{x}_j \mid \phi)\,P(c_{ij}=0)}$$

Gaussian data

When ϕ is radially symmetric Gaussian(0, $\sigma^2 I$),

$$= e^{\frac{1}{2}\left(m'^2 + d'^2\left(2 - \frac{1}{t^2}\right)\right)} \times \text{const}$$

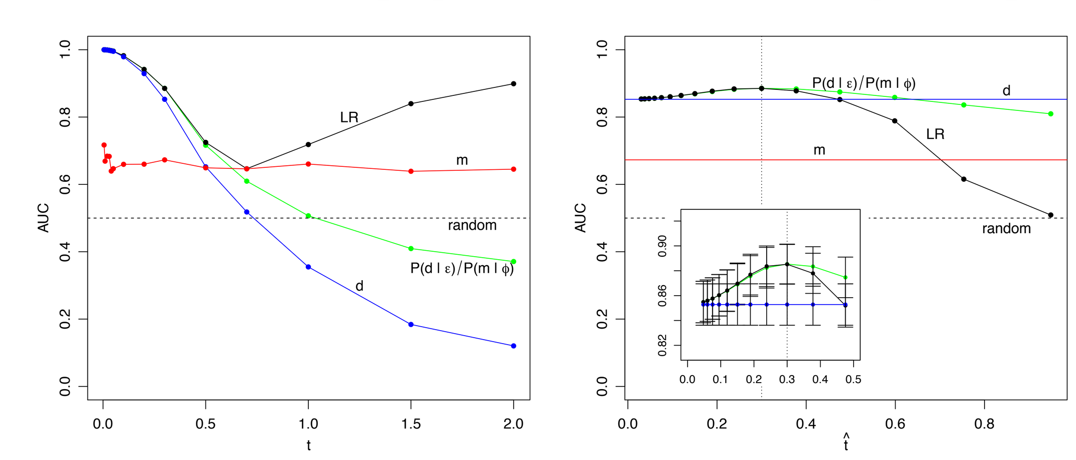
ranking depends only on magnitude of midpoint (m), magnitude of displacement (d), and ratio of standard deviations ($t = \frac{v}{\sigma}$).

Effects of Varying Parameters

Meaning	Parameter	Effect of changing	Effect of mis-guessing
Number of pairs	r or E(r)	Does not affect ranking. Necessary for probability estimates.	
Number of points	n	Does not affect ranking, only probability estimates	-- [n always observed]
Standard deviation of main distribution φ	σ	Only matter via the ratio $t = \frac{v}{\sigma}$	
Standard deviation of displacement distribution ε	v		
Number of dimensions	k	Higher k makes problem easier	-- [k always observed]

Synthetic data experiments

Effect of changing t — Effect of mis-guessing t

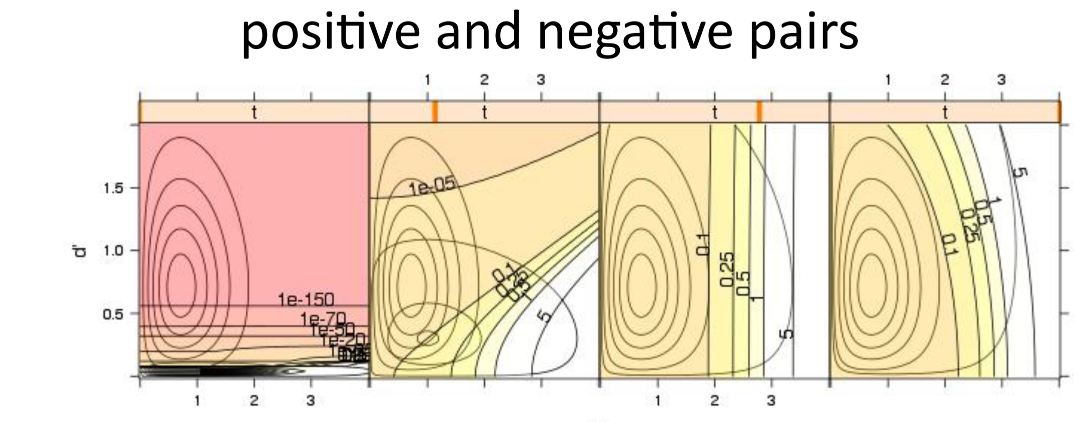


t balances how much optimal method uses distance (d) (often performs well alone) vs. rarity (m) of a pair.

At lowest t, displacement d suffices to separate the distributions.

At $t = \frac{1}{\sqrt{2}}$, d carries no information to distinguish positive from negative pairs.

Theoretical distributions of positive and negative pairs



Real Data

Experiments: Vary vector $\hat{\mathbf{t}}$. Compare our inference method to (scaled Euclidean) distance $P(\mathbf{d} \mid \varepsilon)$, rarity $\frac{1}{P(\mathbf{m} \mid \phi)}$, approximation $\frac{P(\mathbf{d} \mid \varepsilon)}{P(\mathbf{m} \mid \phi)}$.

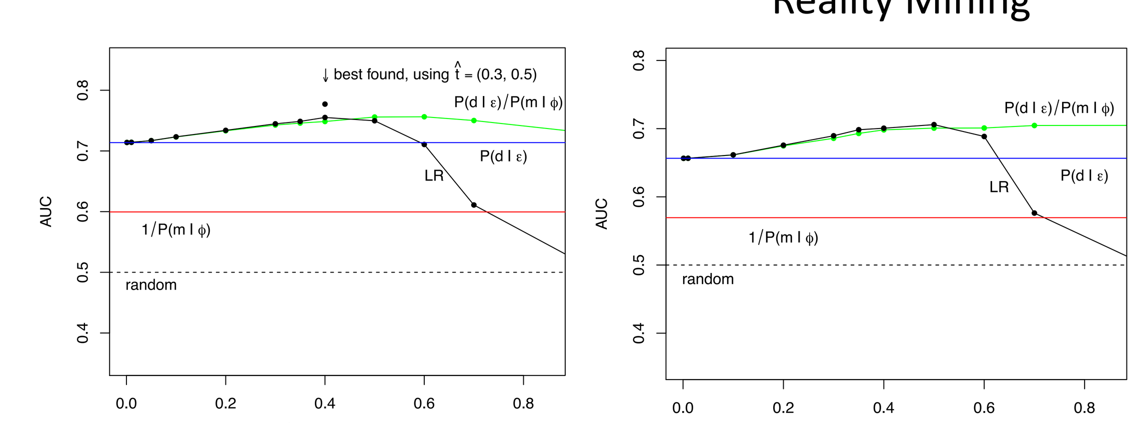
Twins

Given birthweight and Apgar scores, re-identify twins within a data set of babies. [National Center for Health Statistics, 2000]

Cell Phones / Reality Mining

Given seven features of a user's weekly cell phone activity, re-identify instances of the same user across different weeks. [Eagle & Pentland, 2006]

Twins — Cell Phones / Reality Mining



This effort is supported by the Science Applications International Corporation (SAIC) and DARPA under contract number P010089628. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of SAIC, DARPA or the U.S. Government.