Generic Entity Resolution: Identifying Real-World Entities in Large Data Sets

Hector Garcia-Molina
Stanford University

Work with: Omar Benjelloun, Qi Su, Jennifer Widom, Tyson Condie, Nicolas Pombourcq, David Menestrina, Steven Whang
Entity Resolution

e1
N: a  A: b  CC#: c  Ph: e

e2
N: a  Exp: d  Ph: e
Applications

• comparison shopping
• mailing lists
• classified ads
• customer files
• counter-terrorism
Outline

• Why is ER challenging?
• How is ER done?
• Some ER work at Stanford
• Confidences
Challenges (1)

• No keys!

• Value matching
  – “Kaddafí”, “Qaddafí”, “Kadafí”, “Kaddaffí”...

• Record matching

Nm: Tom
  Ad: 123 Main St
  Ph: (650) 555-1212
  Ph: (650) 777-7777

Nm: Thomas
  Ad: 132 Main St
  Ph: (650) 555-1212
Challenges (2)

- Merging records

Nm: Tom
- Ad: 123 Main St
- Ph: (650) 555-1212
- Ph: (650) 777-7777

Nm: Thomas
- Ad: 132 Main St
- Ph: (650) 555-1212
- Zp: 94305
Challenges (3)

• Chaining

Nm: Tom
Ad: 123 Main
BD: Jan 1, 85
Wk: IBM

Nm: Thomas
Ad: 123 Main
Oc: lawyer

Nm: Tom
Ad: 123 Main
Oc: lawyer
Sal: 500K
Challenges (4)

• Un-merging

Nm: Tom
Ad: 123 Main
BD: Jan 1, 85
Wk: IBM
Oc: lawyer
Sal: 500K
too young to make 500K at IBM!!
Challenges (5)

- Confidences in data
  
  Nm: Tom (0.9)
  Ad: 123 Main St (1.0)
  Ph: (650) 555-1212 (0.6)
  Ph: (650) 777-7777 (0.8)

- In value matching, match rules, merge:
  
  conf = ?
Taxonomy

- Pairwise snaps vs. clustering
- De-duplication vs. fidelity enhancement
- Schema differences
- Relationships
- Exact vs. approximate
- Generic vs application specific
- Confidences
Schema Differences

Name: Tom
- Address: 123 Main St
  - Ph: (650) 555-1212
  - Ph: (650) 777-7777

FirstName: Tom
- StreetName: Main St
  - StreetNumber: 123
  - Tel: (650) 777-7777
Pair-Wise Snaps vs. Clustering
De-Duplication vs. Fidelity Enhancement
Relationships
Using Relationships

authors

papers

same??

a1

p1

a2

p2

a3

p5

a4

p7

a5

authors papers same??
Exact vs Approximate ER

- cameras
- CDs
- books

ER

resolved cameras
resolved CDs
resolved books

...
Exact vs Approximate ER

terrorists  sort by age  terrorists

Widom 30

match against ages 25-35
Generic vs Application Specific

- Match function $M(r, s)$
- Merge function $<r, s> \Rightarrow t$
Taxonomy

- Pairwise snaps vs. clustering
- De-duplication vs. fidelity enhancement
- Schema differences
- Relationships
- Exact vs. approximate
- Generic vs application specific
- Confidences
Outline

• Why is ER challenging?
• How is ER done?
• Some ER work at Stanford
• Confidences
Taxonomy

- **Pairwise snaps** vs. clustering
- **De-duplication** vs. fidelity enhancement
- Schema differences **No**
- Relationships **No**
- **Exact** vs. approximate
- **Generic** vs application specific
- Confidences ... **later on**
Correct Answer

$ER(R) = \text{All derivable records} \ldots.$

$\text{Minus "dominated" records}$
Question

• What is best sequence of match, merge calls that give us right answer?
Brute Force Algorithm

• Input R:
  - r1 = [a:1, b:2]
  - r2 = [a:1, c: 4, e:5]
  - r3 = [b:2, c:4, f:6]
  - r4 = [a:7, e:5, f:6]
Brute Force Algorithm

- **Input R:**
  - r1 = [a:1, b:2]
  - r2 = [a:1, c: 4, e:5]
  - r3 = [b:2, c:4, f:6]
  - r4 = [a:7, e:5, f:6]

- **Match all pairs:**
  - r1 = [a:1, b:2]
  - r2 = [a:1, c: 4, e:5]
  - r3 = [b:2, c:4, f:6]
  - r4 = [a:7, e:5, f:6]
  - r12 = [a:1, b:2, c:4, e:5]
Brute Force Algorithm

• Match all pairs:
  – r1 = [a:1, b:2]
  – r2 = [a:1, c: 4, e:5]
  – r3 = [b:2, c:4, f:6]
  – r4 = [a:7, e:5, f:6]
  – r12 = [a:1, b:2, c:4, e:5]

• Repeat:
  – r1 = [a:1, b:2]
  – r2 = [a:1, c: 4, e:5]
  – r3 = [b:2, c:4, f:6]
  – r4 = [a:7, e:5, f:6]
  – r12 = [a:1, b:2, c:4, e:5]
  – r123 = [a:1, b:2, c:4, e:5, f:6]
Brute Force Algorithm

• Input R:
  – r1 = [a:1, b:2]
  – r2 = [a:1, c: 4, e:5]
  – r3 = [b:2, c:4, f:6]
  – r4 = [a:7, e:5, f:6]

• Match all pairs:
  – r1 = [a:1, b:2]
  – r2 = [a:1, c: 4, e:5]
  – r3 = [b:2, c:4, f:6]
  – r4 = [a:7, e:5, f:6]
  – r12 = [a:1, b:2, c:4, e:5]

Can we delete r1, r2?
Question # 2

Brute Force Algorithm

• Match all pairs:
  – r1 = [a:1, b:2]
  – r2 = [a:1, c: 4, e:5]
  – r3 = [b:2, c:4, f:6]
  – r4 = [a:7, e:5, f:6]
  – r12 = [a:1, b:2, c:4, e:5]

• Repeat:
  – r1 = [a:1, b:2]
  – r2 = [a:1, c: 4, e:5]
  – r3 = [b:2, c:4, f:6]
  – r4 = [a:7, e:5, f:6]
  – r12 = [a:1, b:2, c:4, e:5]
  – r123 = [a:1, b:2, c:4, e:5, f:6]
ICAR Properties

• Idempotence:
  – $M(r_1, r_1) = true; <r_1, r_1> = r_1$

• Commutativity:
  – $M(r_1, r_2) = M(r_2, r_1)$
  – $<r_1, r_2> = <r_2, r_1>$

• Associativity
  – $<r_1, <r_2, r_3>> = <<r_1, r_2>, r_3>$
More Properties

• Representativity
  – If \( <r_1, r_2> = r_3 \), then
    for any \( r_4 \) such that \( M(r_1, r_4) \) is true
    we also have \( M(r_3, r_4) = true \).
ICAR Properties ➔ Efficiency

- Commutativity
- Idempotence
- Associativity
- Representativity

- Can discard records
- ER result independent of processing order
Swoosh Algorithms

• Record Swoosh
  • Merges records as soon as they match
  • Optimal in terms of record comparisons

• Feature Swoosh
  • Remembers values seen for each feature
  • Avoids redundant value comparisons
Swoosh Performance
If ICAR Properties Do Not Hold?

r1: [Joe Sr., 123 Main, DL:X]

r2: [Joe, 123 Main, Ph:123]

r3: [Joe Jr., 123 Main, DL:Y]

r12: [Joe Sr., 123 Main, Ph: 123, DL:X]

r23: [Joe Jr., 123 Main, Ph: 123, DL:Y]
If ICAR Properties Do Not Hold?

Full Answer: \( ER(R) = \{ r12, r23, r1, r2, r3 \} \)

Minus Dominated: \( ER(R) = \{ r12, r23 \} \)
If ICAR Properties Do Not Hold?

Full Answer: \( ER(R) = \{r12, r23, r1, r2, r3\} \)
Minus Dominated: \( ER(R) = \{r12, r23\} \)
R-Swoosh Yields: \( ER(R) = \{r12, r3\} \) or \( \{r1, r23\} \)
Swoosh Without ICAR Properties

![Graph showing the result size against the title threshold for R-Swoosh and G-Swoosh.](Image)
Distributed Swoosh

P1  P2  P3

r1  r2  r3  r4  r5  r6  ...

...
Distributed Swoosh

P1

r1
r2
r3
r4
r5
r6
...

P2

r1
r2
r4
r5
...

P3

r2
r3
r5
r6
...

...
DSwoosh Performance

![Graph showing performance comparison between different methods (Sequential, Grid, Linear Ordering, Value Equality) with respect to the Number of records and Maximum Effort (comparisons).]
Outline

• Why is ER challenging?
• How is ER done?
• Some ER work at Stanford
• Confidences
Conclusion

• ER is old and important problem
• Our approach: generic
• Confidences
  – challenging
  – two ways to tame:
    • thresholds
    • packages
Thanks.
Generic Confidence Model

- $r_1 = 0.7 \ [ a:v1, b:v2, c: v3]$

\[0.7[a, b, c] \rightarrow \text{match} \rightarrow \text{yes (or no)}\]

\[0.9[a, c, d] \rightarrow \text{merge} \rightarrow 0.65[a, b, c, d, x]\]
Problem: Properties May Not Hold

• \( r_1 = 0.9 \) \([a, b, c]\)
• \( r_2 = 0.8 \) \([a, d]\)
• say confidences multiplied on merge
• \( \langle r_1, r_2 \rangle = 0.72[a, b, c, d] \)
• \( \langle \langle r_1, r_2 \rangle, r_1 \rangle = 0.648[a, b, c, d] \)
• \( \langle \langle r_1, r_1 \rangle, r_2 \rangle = \langle r_1, r_2 \rangle = 0.72[a, b, c, d] \)
ER with Confidences

• Very Expensive:
  – must compute “all derivations”
  – cannot delete records after they merge

• What can we do??
  – thresholds
  – packages
Important Property

- If \( \text{conf}(R_x) < \text{threshold} \)
- Then for any \( R_y \) derived from \( R_x \)
  \( \text{conf}(R_y) < \text{threshold} \)

\[ C = 0.7 \quad \leq 0.7 \]

\[ r_1 \rightarrow r_3 \rightarrow r_4 \]

\[ r_2 \rightarrow r_3 \]
Thresholds - Example

T=0.7

0.9 [ a: v1, b: v2 ]
0.8 [ a: v1, c: v3 ]
0.6 [ b: v2, c: v3, d: v4]
0.75 [ a: v1, b: v2, c: v3]
0.5 [ a: v1, b: v2, c: v3, d: v4]
...

...
Thresholds - Example

T=0.7

- 0.9 [ a: v1, b: v2 ]
- 0.8 [ a: v1, c: v3 ]
- 0.6 [ b: v2, c: v3, d: v4 ]
- 0.75 [ a: v1, b: v2, c: v3 ]
- 0.5 [ a: v1, b: v2, c: v3, d: v4 ]

...
Goal: C-Swoosh

- Base records
- All possible merges
- Eliminate dominated
- Eliminate below threshold
Goal: C-Swoosh

base records → all possible merges → eliminate dominated → eliminate below threshold

earlier
Does Threshold Property Hold?

- NO: records are evidence

\[ 0.7[a, b, c] \rightarrow \text{merge} \rightarrow 0.9[a, b, c] \]

\[ 0.8[a, b, c] \]
Does Threshold Property Hold?

• YES: records are beliefs

\[
\begin{align*}
.7[a, b, c] & \rightarrow \text{merge} \rightarrow .8[a, b, c] \\
.8[a, b, c] & \rightarrow \text{merge} \rightarrow .8[a, b, c]
\end{align*}
\]
Simple Confidence Model

• 0.7 \([a, b]\)

Alternate Worlds:

\([a, b]\)  \([a, b]\)  \([a, b, c]\)  \([a, b]\)  \([a, b]\)

\([a, b]\)  \([a, b, d]\)  ???  ???  ???
Rules

• $0.7[a, b, c], 0.7[a, b, c]$
  $\Rightarrow 0.7[a, b, c]$

• $0.7[a, b], 0.5[a, b]$
  $\Rightarrow 0.7[a, b]$

• $0.7[a, b, c], 0.5[a, b]$
  $\Rightarrow 0.7[a, b, c]$

• $0.7[a, b, c], 0.9[a, b]$
  $\Rightarrow 0.7[a, b, c], 0.9[a, b]$

• etc
Matches

0.9[a, b, c]
0.8[a, b, d]
[a, x]
[c, d, y]

Match with confidence 0.5

worlds

1 2 3 4 5 6 7 8 9 10

[a,b,c] ————————————————————
[a,b,d] ————————————————————
[a,b,c,d] ————————
Matches

0.9[a, b, c]
0.8[a, b, d]
[a, x]
[c, d, y]

0.4[a, b, c, d]
0.9[a, b, c]
0.8[a, b, d]
[a, x]
[c, d, y]

worlds

1 2 3 4 5 6 7 8 9 10

[a, b, c]
[a, b, d]
[a, b, c, d]
Summary

• Belief model well suited for ER
• Evidence model is very complex and expensive!
Packages

• Match does not use confidences
  – merge does compute confidences
• 4 properties hold for deterministic attributes
  – e.g., \(<\langle r_1, r_2 \rangle, r_3 \rangle = \langle r_1, \langle r_2, r_3 \rangle \rangle\)
    ignoring confidences
Partition Records

- \( r_1 = 0.9 \ [a:1, b:2] \)
- \( r_2 = 0.8 \ [a:1, c:4, e:5] \)
- \( r_3 = 0.7 \ [b:2, c:4, f:6] \)
- \( r_4 = 0.8 \ [a:7, e:5, f:6] \)
- \( r_5 = 0.9 \ [a:7, b:2] \)
Expand Packages

- $r_1 = .9 \ [a:1, b:2]$
- $r_2 = .8 \ [a:1, c:4, e:5]$
- $r_3 = .7 \ [b:2, c:4, f:6]$
- $r_4 = .8 \ [a:7, e:5, f:6]$
- $r_5 = .9 \ [a:7, b:2]$

$\langle r_1, r_2 \rangle$
$\langle r_1, r_3 \rangle$
$\langle r_2, r_3 \rangle$
...

Diagram:

```
  r1
 /   \
r2   r12
    /  \
   r123
```

- $r_1, r_2, r_3$
- $\langle r_1, r_2 \rangle$
- $\langle r_1, r_3 \rangle$
- $\langle r_2, r_3 \rangle$
...
Conclusion

• ER is old and important problem
• Our approach: generic
• Confidences
  – challenging
  – two ways to tame:
    • thresholds
    • packages
Thanks.
Extra Slides
Taxonomy

- **Pairwise snaps** vs. clustering
- **De-duplication** vs. fidelity enhancement
- Schema differences **No**
- Relationships **No**
- **Exact** vs. approximate
- **Generic** vs application specific
- Confidences ... **later on**
One Confidence Model

[\text{id1, a, b, c, d}]

[\text{id2, a, c, e}]

[\text{id3, a, b, f, g}]

\text{shorthand}
Records Are Evidence

[id1, a, b, c, d]

[id1, a, b, c, d]
[id1, a, b, d]
[id1, a, x]
[id1, b, y]

[id1, (3/4)a, (3/4)b, (1/4)c, (2/4)d, (1/4)x, (1/4)y]

not 0.25
New Evidence

\[
\begin{align*}
&[\text{id1, a, b, c, d}] \\
\rightarrow \\
&[\text{id1, (3/4)a, (3/4)b, (1/4)c, (2/4)d, (1/4)x, (1/4)y}] \\
&\quad + [\text{id1, a, b, c, d}] \\
\rightarrow \\
&[\text{id1, (4/5)a, (4/5)b, (2/5)c, (3/5)d, (1/5)x, (1/5)y}]
\end{align*}
\]
No Ids

[ a, b, c ]
[ a, b, d ]
[ a, x ]
[ c, d, y ]

[ a, b, c ]
[ a, b, d ]
[ a, x ]
[ c, d, y ]

[ a, b, (1/2)c, (1/2)d ]
[ a, x ]
[ c, d, y ]

0.3
0.7
No Ids

\[ [a, b, c] \]
\[ [a, b, d] \]
\[ [a, x] \]
\[ [c, d, y] \]

\[ [a, b, (1/2)c, (1/2)d] \]
\[ [a, x] \]
\[ [c, d, y] \]

\[ [(2/3)a, (2/3)b, (2/3)c, (2/3)d, (1/3)y] \]
\[ [a, x] \]
Threshold = 0.5; Support = 2
Maximal Record
Example: [a, b, c, d]

Queries?

- [a, b, c] 0.3
- [a, b, d] 0.7
- [a, x]
- [c, d, y]
- [a, b, c]
- [a, b, d]
- [a, x]
- [c, d, y]

- [a, b, (1/2)c, (1/2)d] 0.1
- [a, x]
- [c, d, y]

- [(2/3)a, (2/3)b, (2/3)c, (2/3)d, (1/3)y] 0.9
- [a, x]
Queries?

Threshold = 0.5; Support = 2
Maximal Record
Example: [a, b, c, d]
Need Simpler Model?
Bonus Material

• Entity Resolution, Confidences, and their relationship to Information Privacy
Privacy

Alice

Bob

1.0
Nm: Alice
Ad: 32 Fox
Ph: 5551212

1.0
Nm: Alice
Ad: 32 Fox
Ph: 5551212

1.0
Nm: Alice
Ad: 32 Fox
Ph: 5551212
Ad: 14 Cat
Leakage

L = 0.6  (between 0 and 1)
Multi-Record Leakage

Alice

Nm: Alice
Ad: 32 Fox
Ph: 5551212

Bob

r1, L = 0.9
r2, L = 0.8
r3, L = 0.7

LL = 0.9 (between 0 and 1, e.g., max L)
Q1: Added Vulnerability?

ΔLL = ??

r4 may cause Bob’s records to snap together!
Q2: Disinformation?

Alice  p  r1  r2  r3  r4 (lies)

Bob

ΔLL = ??

What is most cost effective disinformation?
Q3: Verification?

Alice

Bob

What is best fact to verify to increase confidence in hypothesis?
Summary

• Entity resolution is critical
• Efficient resolution important
• Confidences are important, but how?
• ER is key aspect of info privacy

  - check www-db.stanford.edu for Swoosh paper & forthcoming paper
Thanks.
Extra Slides
Challenges

• Exponential growth in complexity

0.9 [ a: v1, b:v2 ]
0.8 [ a:v1, c: v3 ]
0.6 [ b:v2, c:v3, d:v4]
0.75 [ a:v1, b:v2, c:v3]
0.5 [a:v1, b:v2, c:v3, d: v4]
...

...
Three Ideas to Tame Complexity

• Thresholds
• Domination
• Packages
Thresholds

$T=0.7$

0.9 [ a: v1, b:v2 ]
0.8 [ a:v1, c: v3 ]
0.6 [ b:v2, c:v3, d:v4]
0.75 [ a:v1, b:v2, c:v3]
0.5 [a:v1, b:v2, c:v3, d: v4]
...

...
Domination

0.9 [ a: v1, b: v2, c: v3 ]
0.8 [ a: v1, b: v2, c: v3 ]
0.8 [ b: v2, c: v3 ]
...

88
Domination

0.9 [ a: v1, b:v2, c:v3 ]
0.8 [ a:v1, b: v2, c: v3 ]
0.8 [ b:v2, c:v3 ]
...

89
Summary

• Our approach: pairwise, generic, Swoosh
• Confidences
• Making Tractable:
  – threshold
  – domination
  – packages
Thanks You
What Swoosh Does NOT Do

• Hash table with every pair seen:
  – records $r_i, r_j$
  – compared values $v_i, v_j$

• Swoosh achieves the same effect with our $N^2$ space
Swoosh Performance (I)
Swoosh Performance (II)
Swoosh Performance (III)