# Nobody Writes Letters Anymore: Helping people make sense of historically significant email collections

***Douglas W. Oard***

**College of Information Studies and**

**UMIACS Laboratory for Computational Linguistics and Information Processing**

# A Real Example

Clinton White House

32 million emails

National Archives

search request

Tobacco Policy

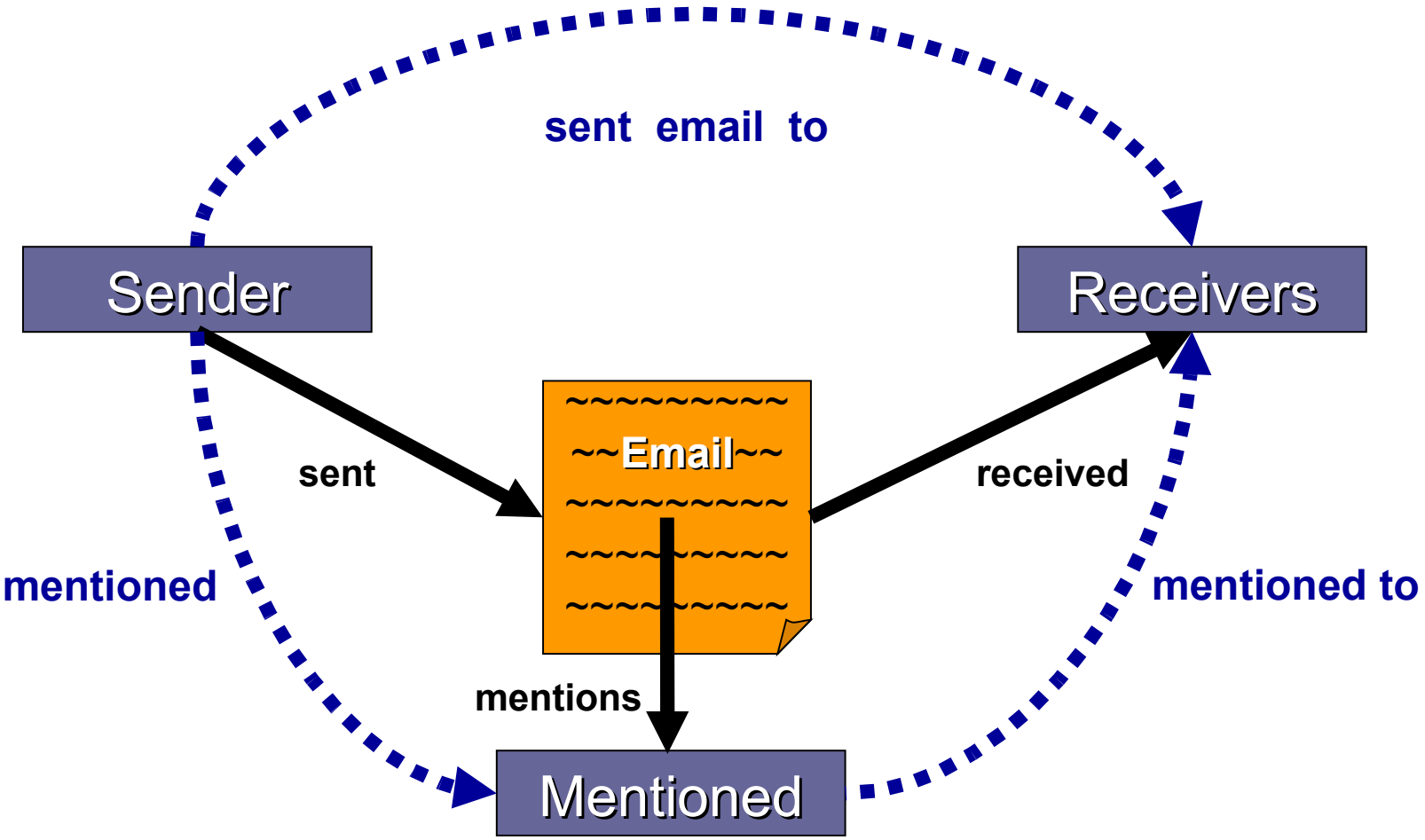
80,000

hired 25 persons

for 6 months …

200,000

# A Richer View of a Social Network



# Why is that Needed?

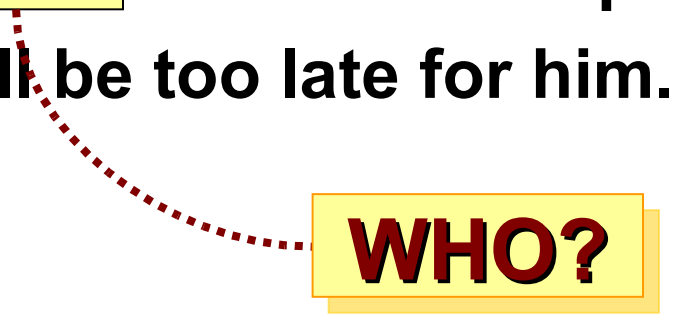
- *Users* unfamiliar with discussions
  - Lawyers
  - Historians
  - Police investigators

- *Downstream process*
  - Expanding ambiguous names at indexing time
  - Expert finding
  - Social network analysis

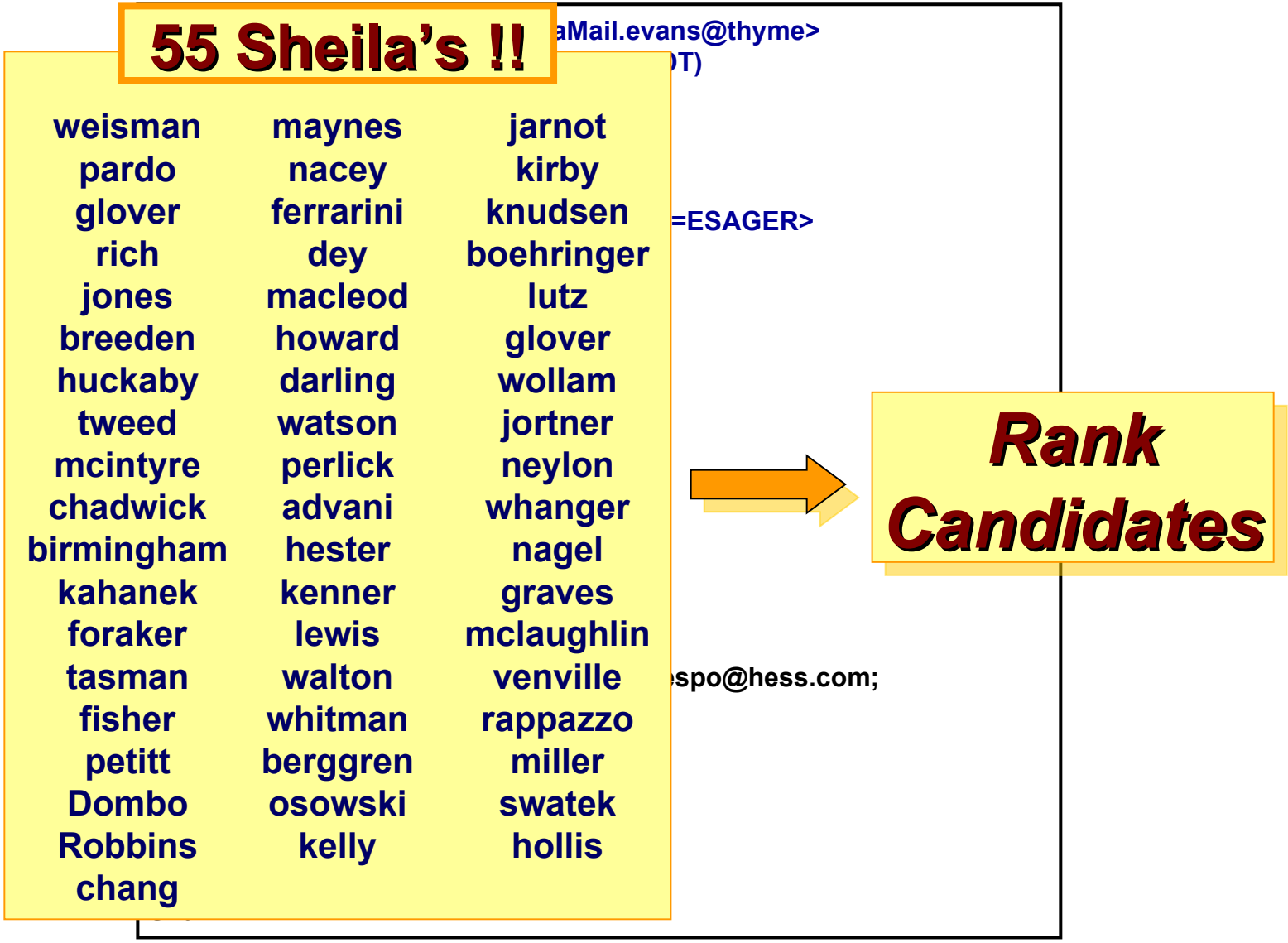
# Identity Resolution in Email

***Date***: Wed Dec 20 08:57:00 EST 2000
***From***: Kay Mann <kay.mann@enron.com>
***To***: Suzanne Adams <suzanne.adams@enron.com>
***Subject***: Re: GE Conference Call has be rescheduled

Did Sheila want Scott to participate? Looks like the call will be too late for him.



# Enron Collection



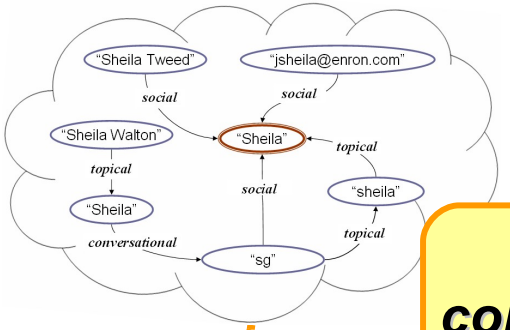
# Proposed Generative Model

1. Choose "**person**" $c$ to mention

   $p(c)$

1. Choose appropriate "**context**" $X$ to mention $c$

   $p(X \mid c)$




1. Choose a "**mention**" $l$

   $p(l \mid X, c)$

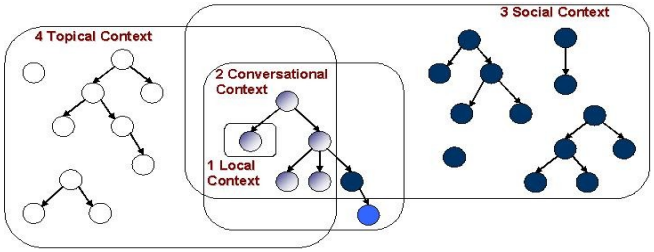
**"sheila"**

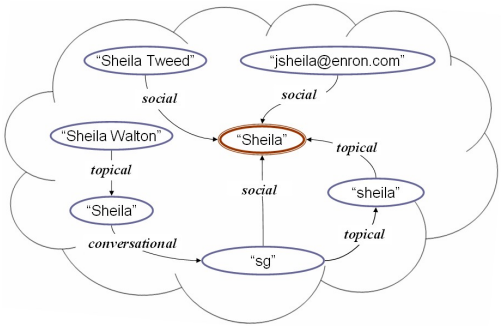
***GE conference call***

# 3-Step Solution

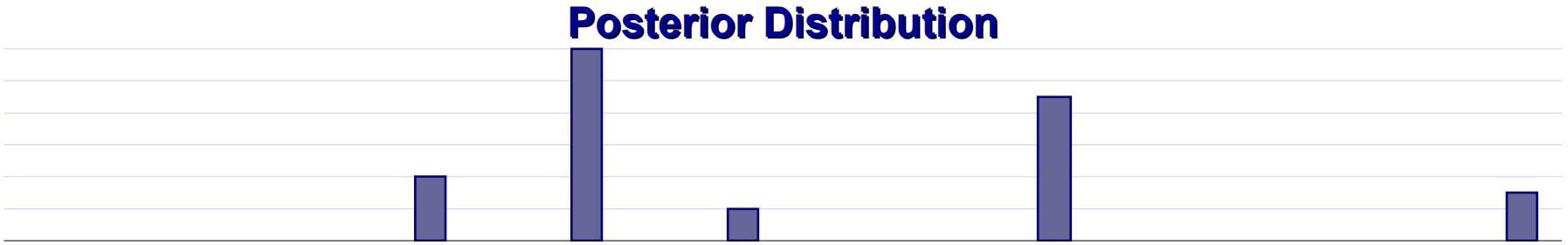
## (1) Identity Modeling

## (2) Context Reconstruction



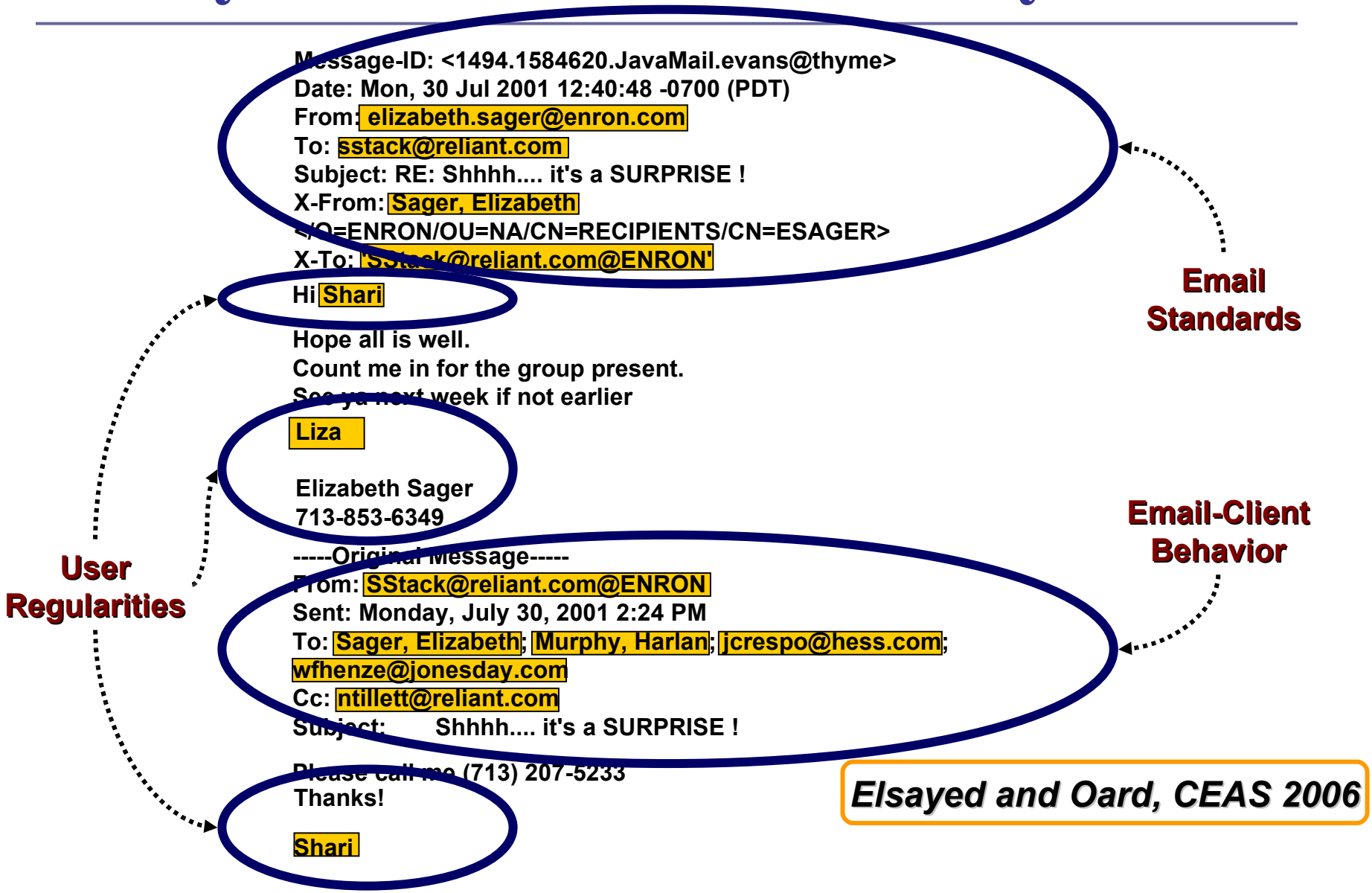

## (3) Mention Resolution



# Outline

- Introduction and Approach Overview
- ***Computational Model of Identity***
- Context Reconstruction
- Mention Resolution
- Evaluation
- Conclusion and Future Work

# “Easy References” of Identity




*Elsayed and Oard, CEAS 2006*

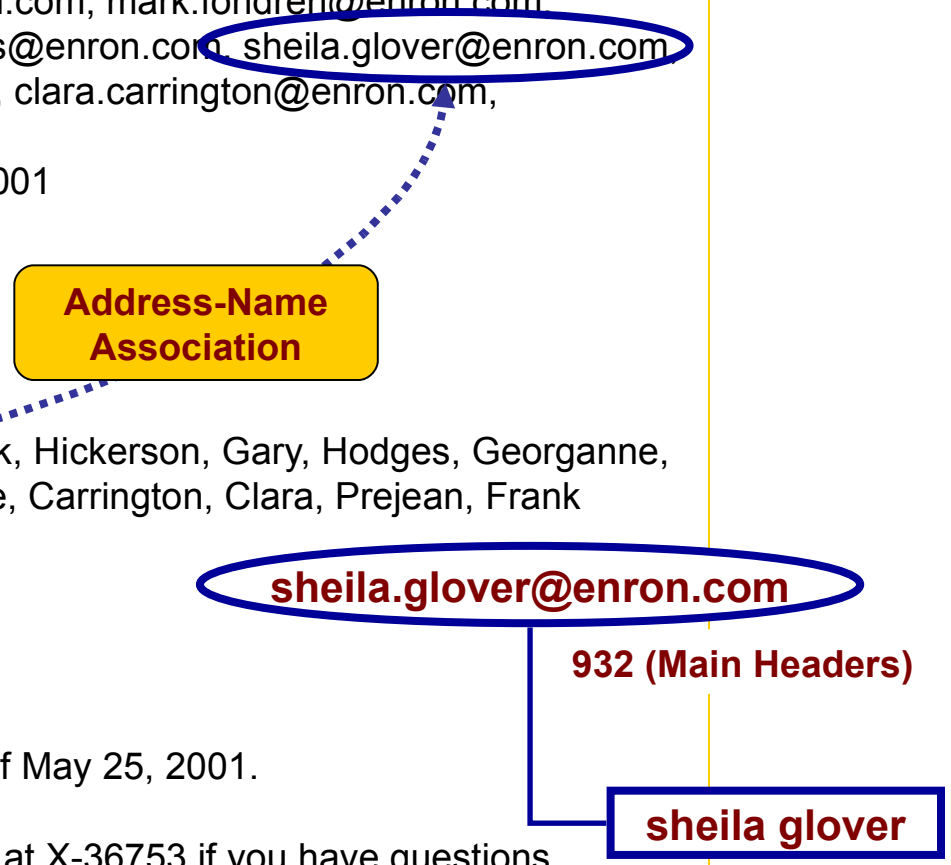
# Extraction From Main Headers

Message-ID: <26343876.1075845080062.JavaMail.evans@thyme>
Date: Fri, 1 Jun 2001 09:38:29 -0700 (PDT)
From: anita.luong@enron.com
To: sally.beck@enron.com, scott.earnest@enron.com, mark.fondren@enron.com,
gary.hickerson@enron.com, georganne.hodges@enron.com, sheila.glover@enron.com,
kori.loibl@enron.com, lynne.ruffer@enron.com, clara.carrington@enron.com,
frank.prejean@enron.com
Subject: ERMS Discount Memo as of May 25, 2001
Cc: chris.abel@enron.com
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: chris.abel@enron.com
X-From: Luong, Anita
X-To: Beck, Sally , Earnest, Scott, Fondren, Mark, Hickerson, Gary, Hodges, Georganne,
Glover, Sheila, Loibl, Kori, Ruffer, Mary Lynne, Carrington, Clara, Prejean, Frank
X-cc: Abel, Chris
X-bcc:
X-Folder: \Beck, Sally\Beck, Sally\Inbox
X-FileName: Beck, Sally.pst

Attached is the ERMS Discounting Analysis as of May 25, 2001.

Please call Chris Abel at X33102 or Anita Luong at X-36753 if you have questions.

Thanks



# Extraction From Quoted Headers

---------- Forwarded by Sheila Glover/HOU/ECT on 06/14/2000 07:39 AM ----------

Sheung Tam <Sheung.Tam@msdw.com> on 06/14/2000 07:39:50 AM
Please respond to Sheung.Tam@msdw.com
To: Sheila Glover <Sheila.Glover@enron.com>
cc:
Subject: Re: Enron India LLC (Delaware Company)

Hi Sheila,
For the ne ve to be in the account, before any trades
Being put on the account.
Sheung

Address-Name Association

# Signature & Salutation Detection

**From: sheila.glover@enron.com**

Shane,
Laurel and I met with Sara Shackleton, Legal, yesterday.
Sara provided the following updates for Australia ISDA Agreements:
National Australia Bank Exectued 1/3/2000 on our side, received executed back from them 3/3/2000
West Pac Our end done, sent to them. Susan Flynn to follow-up to id where in process currently.
ANZ Bank We need to know the exact name of the entity we want the agreement with so we can apply to the credit group to approve gettingan agreement in place.

**thanks, sheila**

**Enron Broadband Services, Inc.**
**1400 Smith, Suite EB-4573A**
**Houston, TX 77002**

# Nickname Detection

**From: sheila.glover@enron.com**

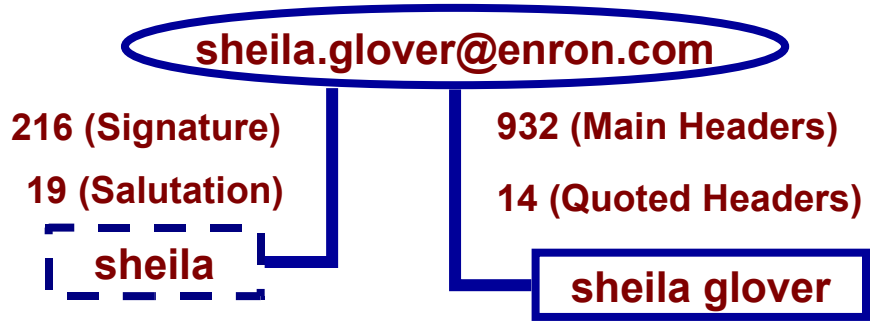
Shane,
Laurel and I met with Sara Shackleton, Legal, yesterday.
Sara provide[illegible]s for Australia ISDA Agreements:
National Aus[illegible] 1/3/2000 on our side, received executed back from th[illegible]
West Pac Our end done, sent to them. Susan Flynn to follow-up to id where in process currently.
ANZ Bank We need to know the exact name of the entity we want the agreement with so we can apply to the credit group to approve gettingan agreement in place.

**thanks, sheila**

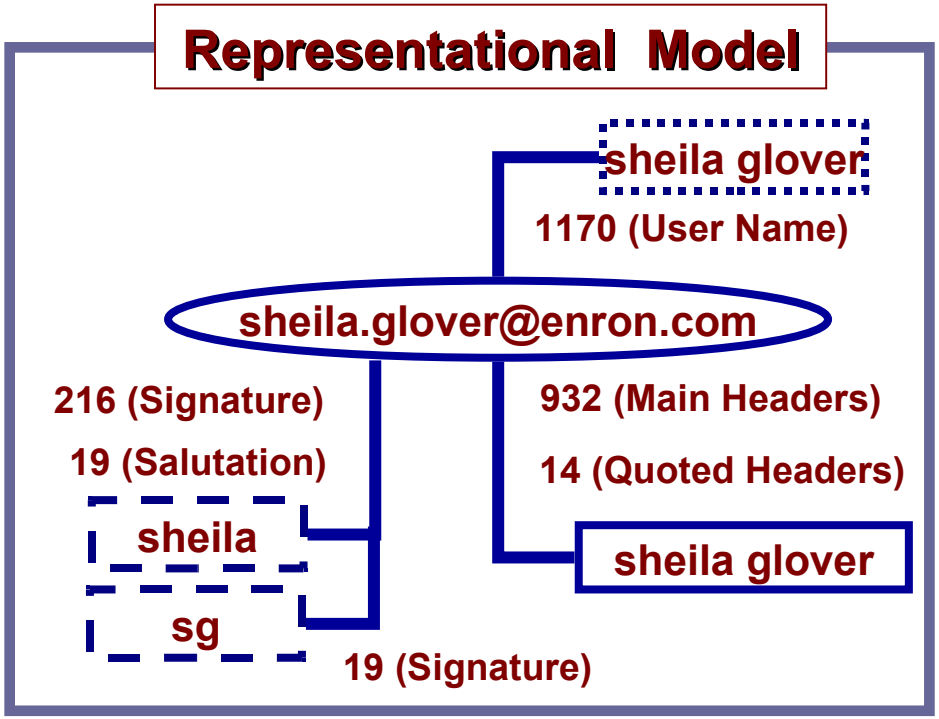
**Address-Nickname Association**

***nickname***

**Enron Broadband Services, Inc.**
**1400 Smith, Suite EB-4573A**
**Houston, TX 77002**



# Representational Model of Identity




**77,240 non-trivial identity models**

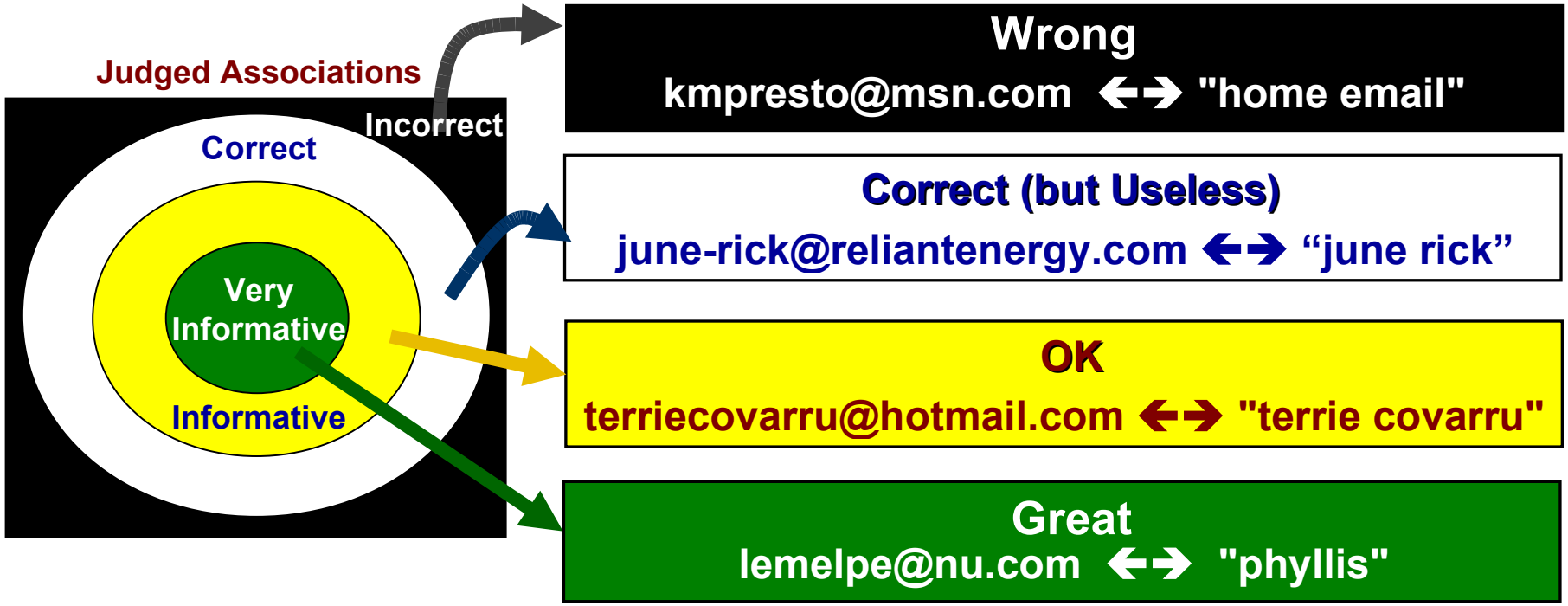
*Elsayed and Oard, CEAS 2006*

# Evaluation of Identity Modeling

## Stratified Sampling

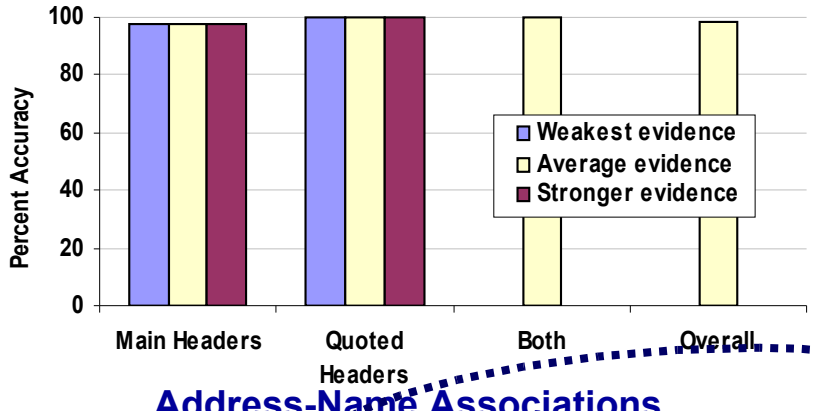
| | Weakest Evidence | Stronger Evidence |
|---|---|---|
| **Address-Name Associations** | | |
| Main headers only | 50 (freq = 1) | 50 (freq ≥ 2) |
| Quoted headers only | 50 (freq = 1) | 50 (freq ≥ 2) |
| Both headers | 50 (freq ≥ 2) | |
| **Address-Nickname Associations** | | |
| Salutations only | 50 (freq = 3) | 50 (freq ≥ 4) |
| Signatures only | 50 (freq = 2) | 50 (freq ≥ 3) |
| Both | 50 (freq ≥ 2) | |

# Judgment Process

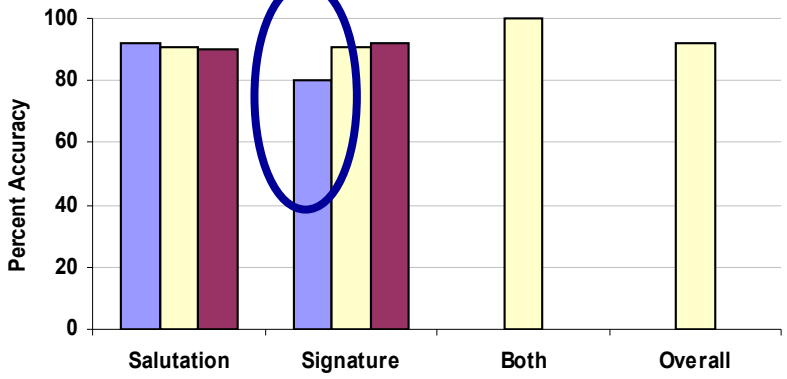


# Results of Identity Modeling

## Correct

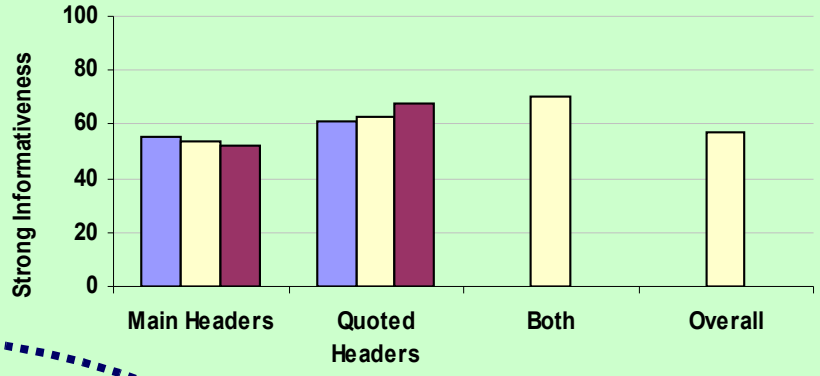



Address-Name Associations

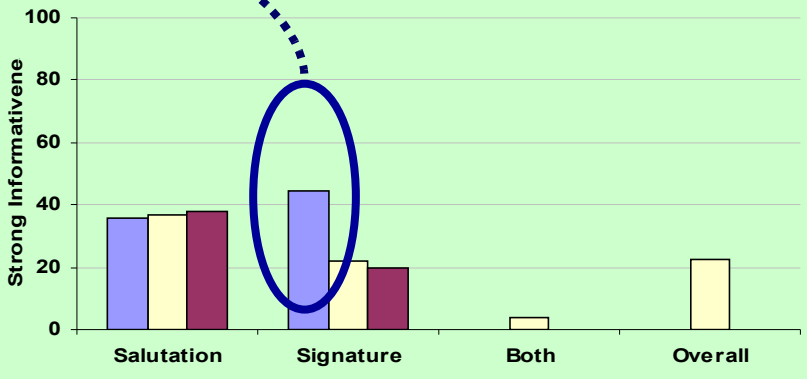



Address-Nickname Associations

## Very Informative

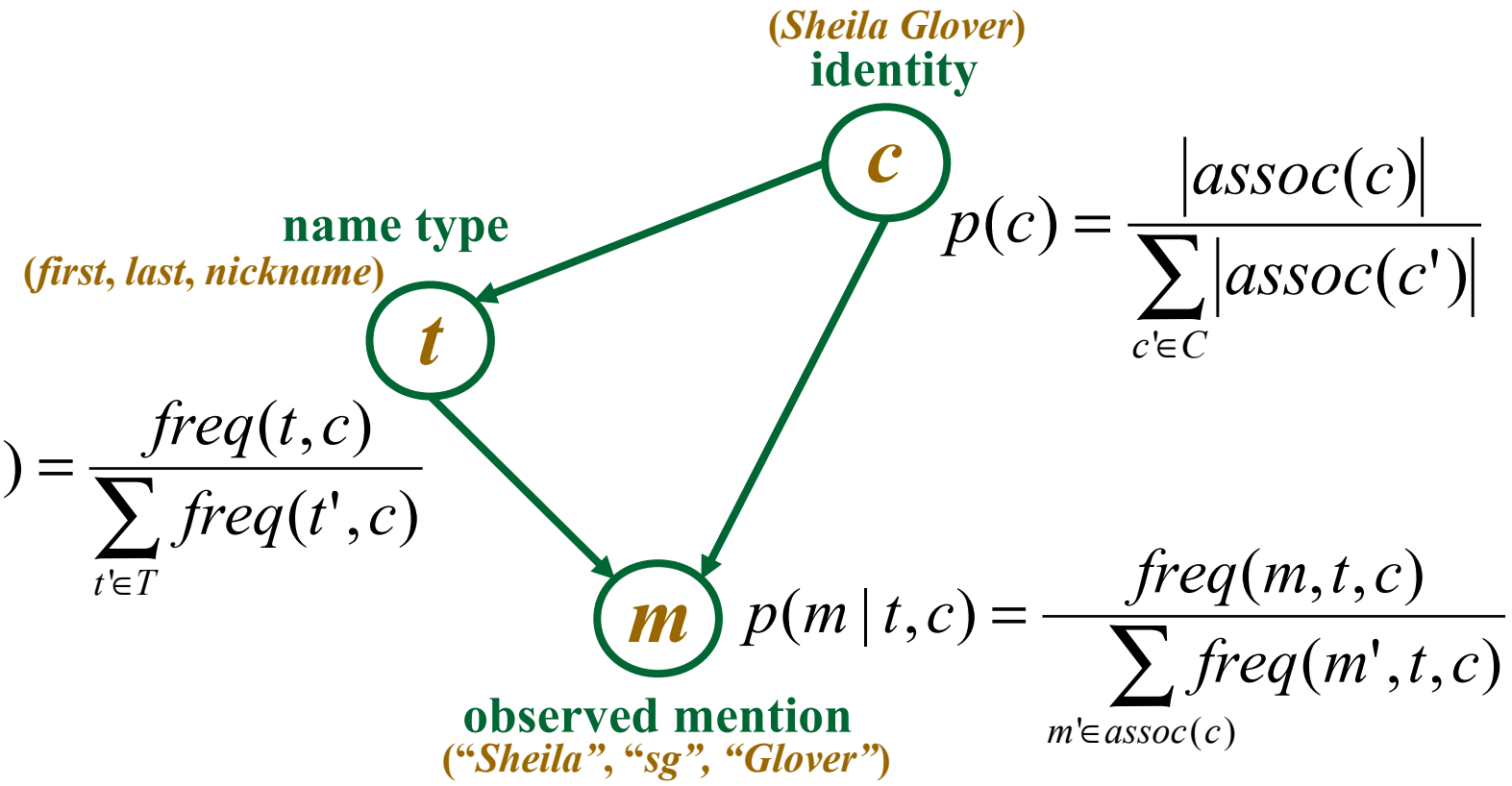



Address-Name Associations




Address-Nickname Associations

# Computational Model



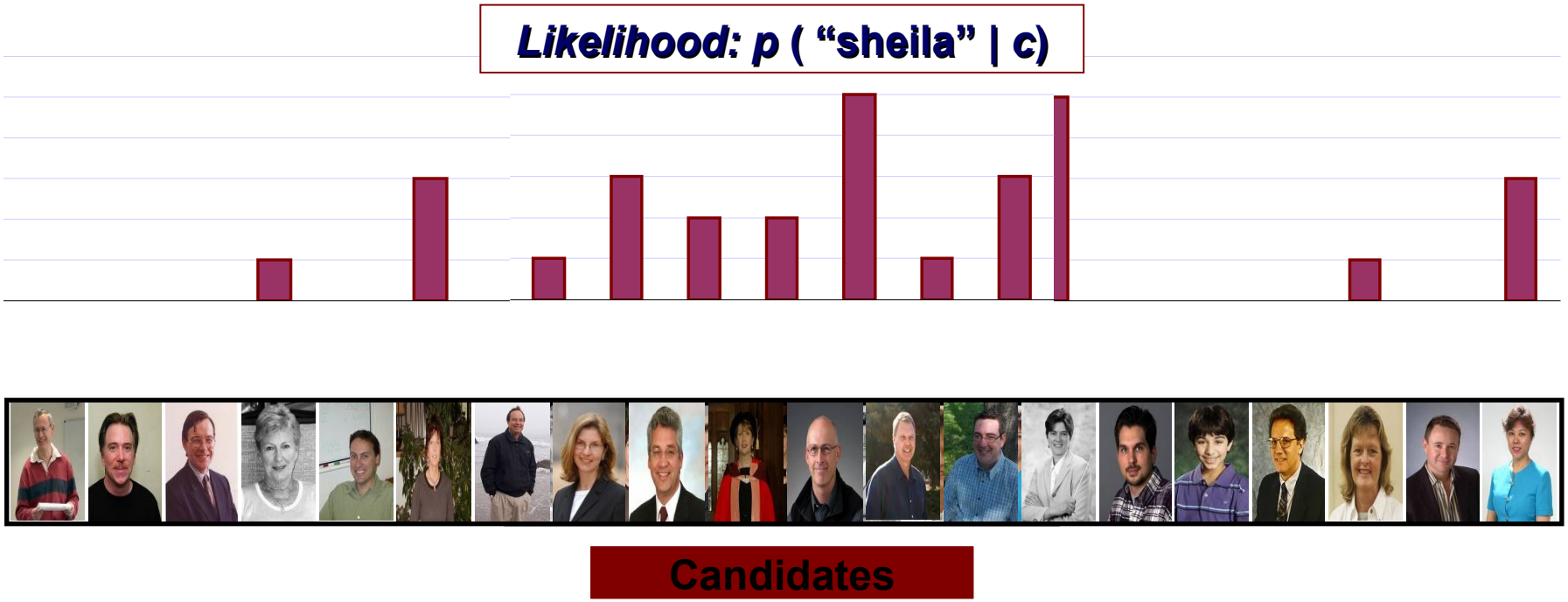

$$p(c) = \frac{|assoc(c)|}{\sum_{c' \in C} |assoc(c')|}$$

$$p(t \mid c) = \frac{freq(t,c)}{\sum_{t' \in T} freq(t',c)}$$

$$p(m \mid t,c) = \frac{freq(m,t,c)}{\sum_{m' \in assoc(c)} freq(m',t,c)}$$

$$p(m \mid c) = \sum_{t \in T} p(m \mid t,c)\, p(t \mid c)$$

# Candidates



# Outline

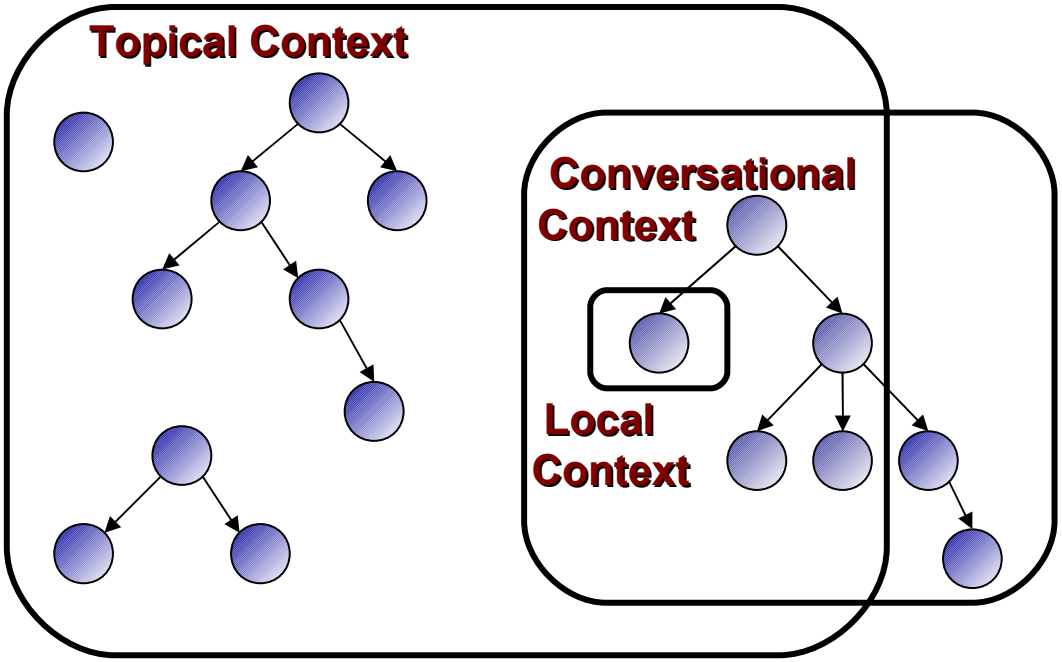
- Introduction and Approach Overview
- Computational Model of Identity
- ***Context Reconstruction***
- Mention Resolution
- Evaluation
- Conclusion and Future Work

# Who is that “Sheila”?

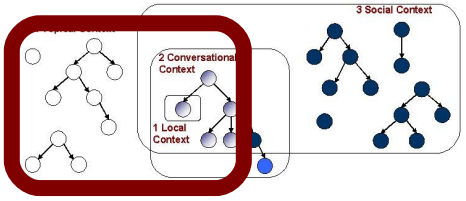
***Date***: **Wed Dec 20 08:57:00 EST 2000**
***From***: **Kay Mann <kay.mann@enron.com>**
***To***: **Suzanne Adams <suzanne.adams@enron.com>**
***Subject***: **Re: GE Conference Call has be rescheduled**

**Did Sheila ? want Scott to participate? Looks like the call will be too late for him.**

# Contextual Space



# Topical Context




***Date*: Wed Dec 20 08:57:00 EST 2000**
***From*: Kay Mann <kay.mann@enron.com>**
***To*: Suzanne Adams <suzanne.adams@enron.com>**
***Subject*: Re: GE Conference Call has be rescheduled**

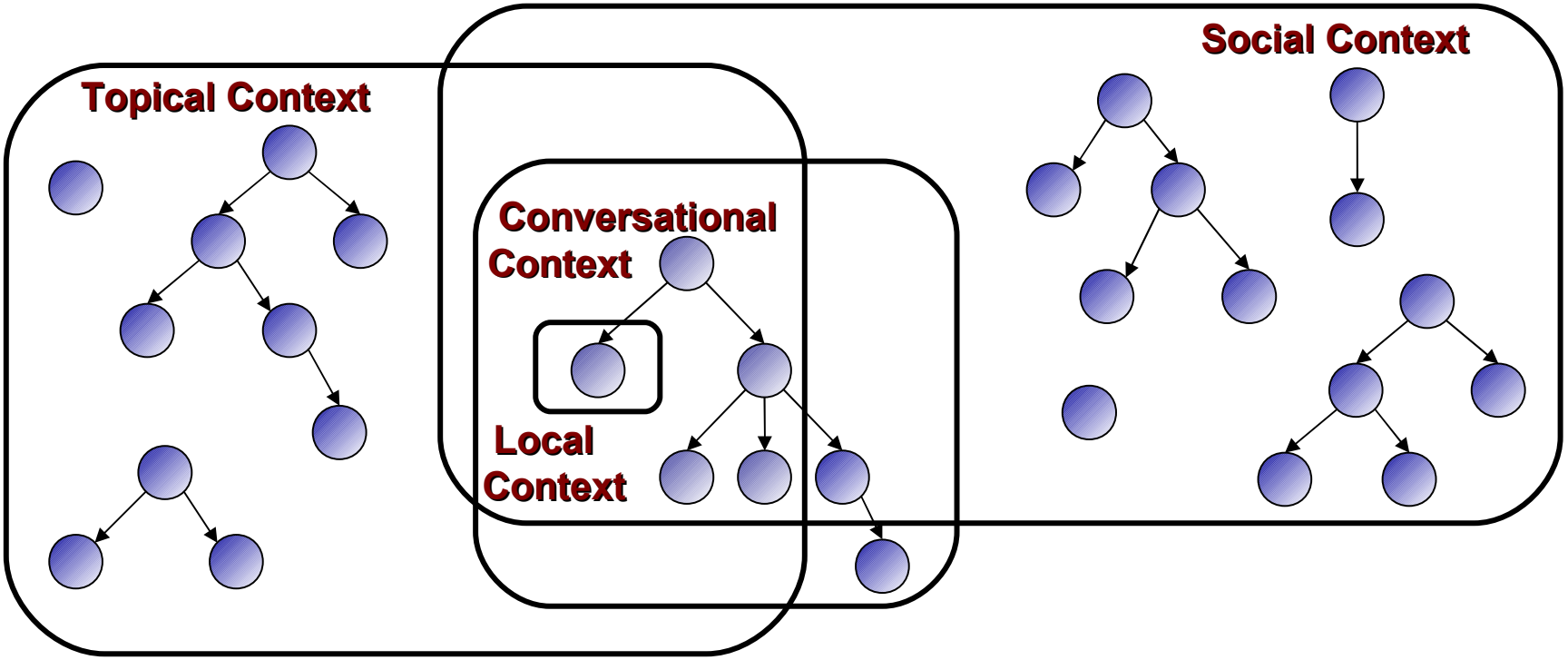
**Did Sheila want Scott to participate? Looks like the call will be too late for him.**

**Date**: Fri Dec 15 05:33:00 EST 2000
**From**: david.oxley@enron.com
**To**: vince j kaminski <vince.kaminski@enron.com>
**Cc**: sheila walton ***sheila.walton@enron.com***
**Subject**: Re: Grant Masson

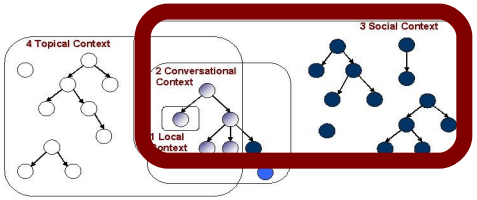
Great news. Lets get this moving along. **Sheila**, can you work out **GE** letter?

Vince, I am in London Monday/Tuesday, back Weds late. I'll ask Sheila to fix this for you and if you need me **call** me on my cell phone.

# Contextual Space



# Social Context




***Date*: Wed Dec 20 08:57:00 EST 2000**
***From*: Kay Mann <kay.mann@enron.com>**
***To*: Suzanne Adams <suzanne.adams@enron.com>**
***Subject*: Re: GE Conference Call has be rescheduled**

**Did Sheila want Scott to participate? Looks like the call will be too late for him.**

**Date**: Tue, 19 Dec 2000 07:07:00 -0800 (PST)
**From**: rebecca.walker@enron.com
**To**: **kay.mann@enron.com**
**Subject**: ESA Option Execution

Kay
Can you initial the ESA assignment and assumption agreement or should I ask ***Sheila Tweed*** to do it? I believe she is currently en route from Portland.
Thanks,
Rebecca

# Formally

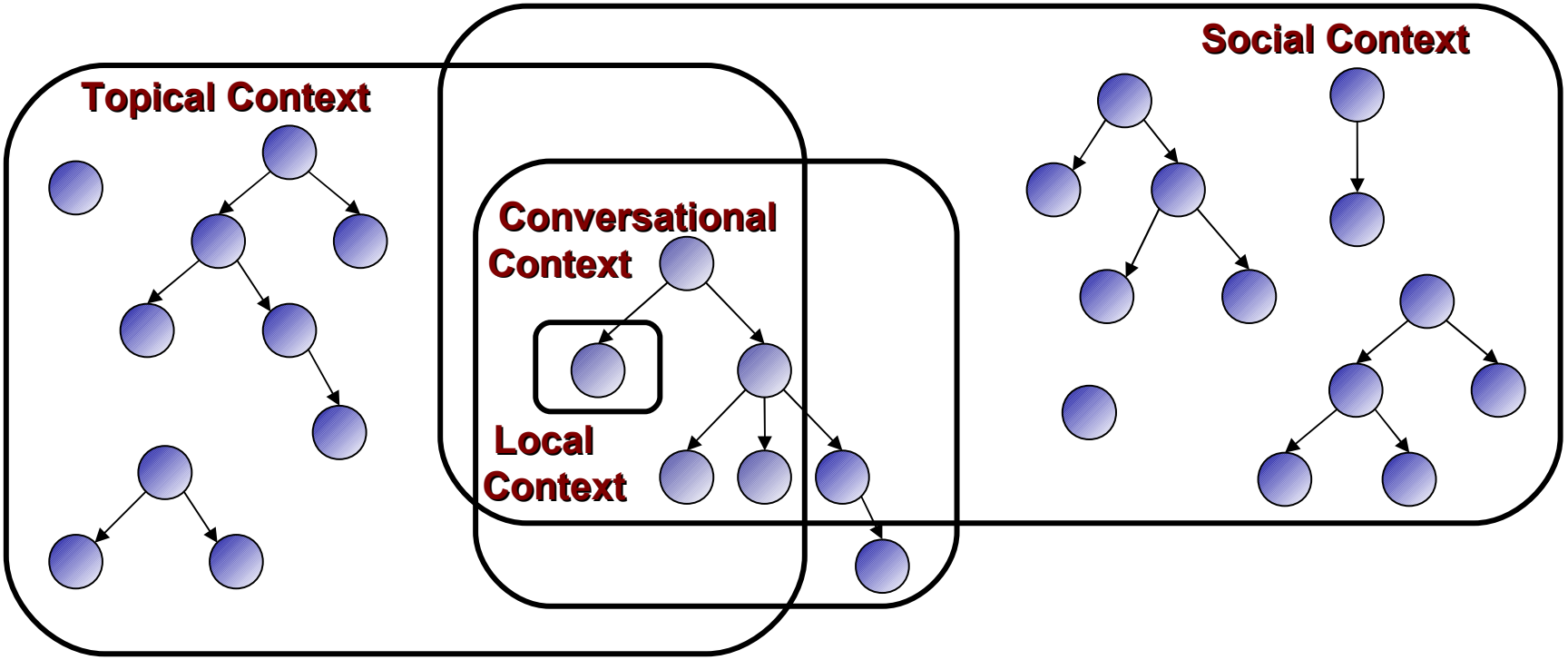
- A context $x_k$ **of an email** is a ***probability distribution*** over emails

$$p(e_j \mid x_k(e_i))$$

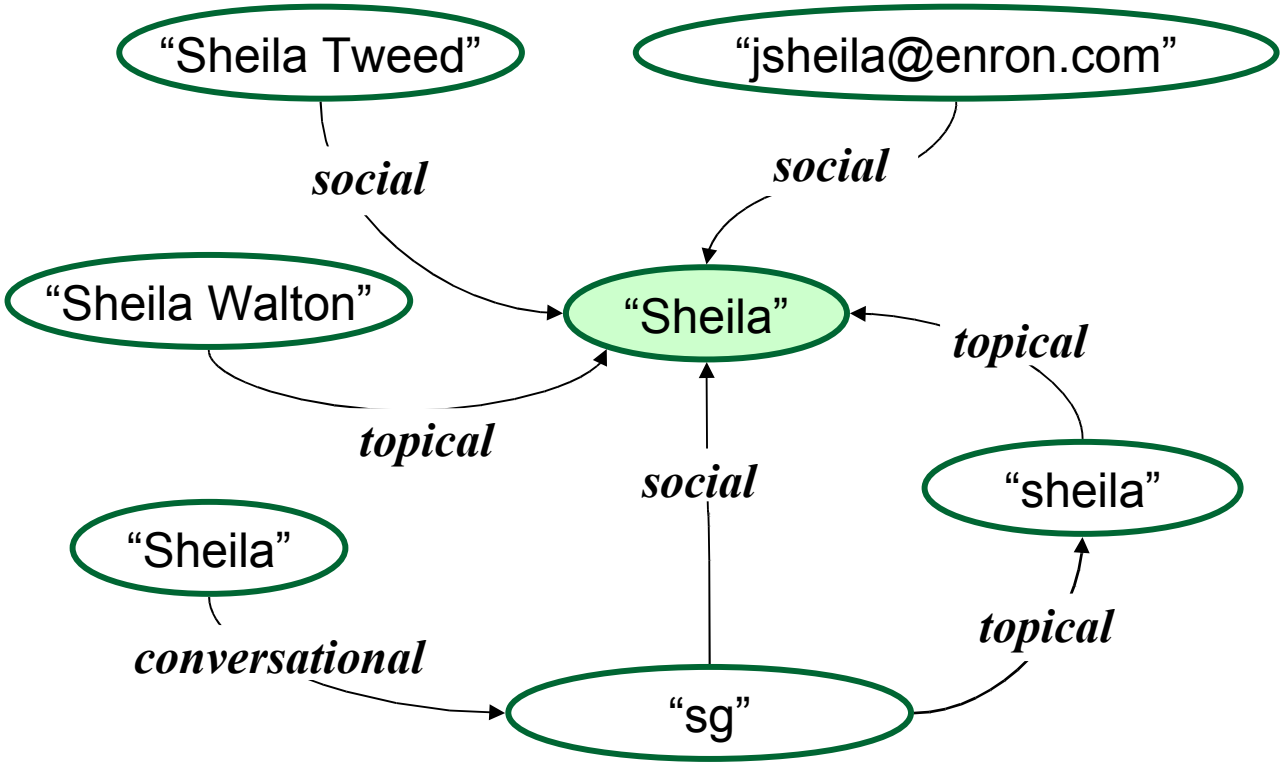
  - Probability estimated based on type of context

- We model the Contextual Space as a linear combination of evidence from 4 contexts

# Contextual Space (emails)



# Contextual Space (Mentions)



# Outline

- Introduction and Approach Overview
- Computational Model of Identity
- Context Reconstruction
- ***Mention Resolution***
- Evaluation
- Conclusion and Future Work

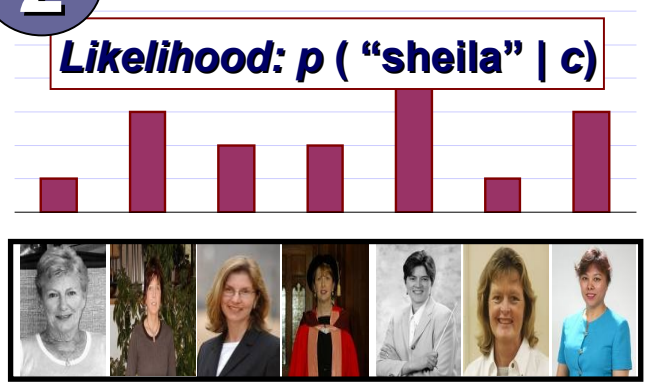
# Mention Resolution

1

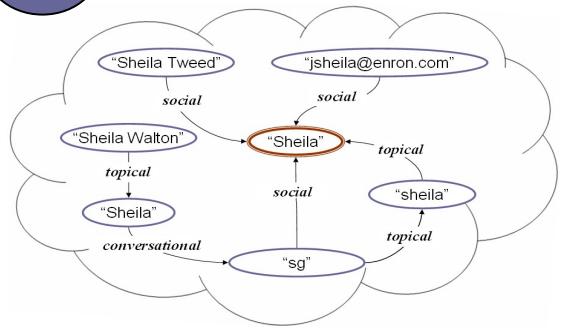
*Date*: Wed Dec 20 08:57:00 EST 2000
*From*: Kay Mann <kay.mann@enron.com>
*To*: Suzanne Adams <suzanne.adams@enron.com>
*Subject*: Re: GE Conference Call has be rescheduled

Did Sheila? want Scott to participate? Looks like the call will be too late for him.

2




3

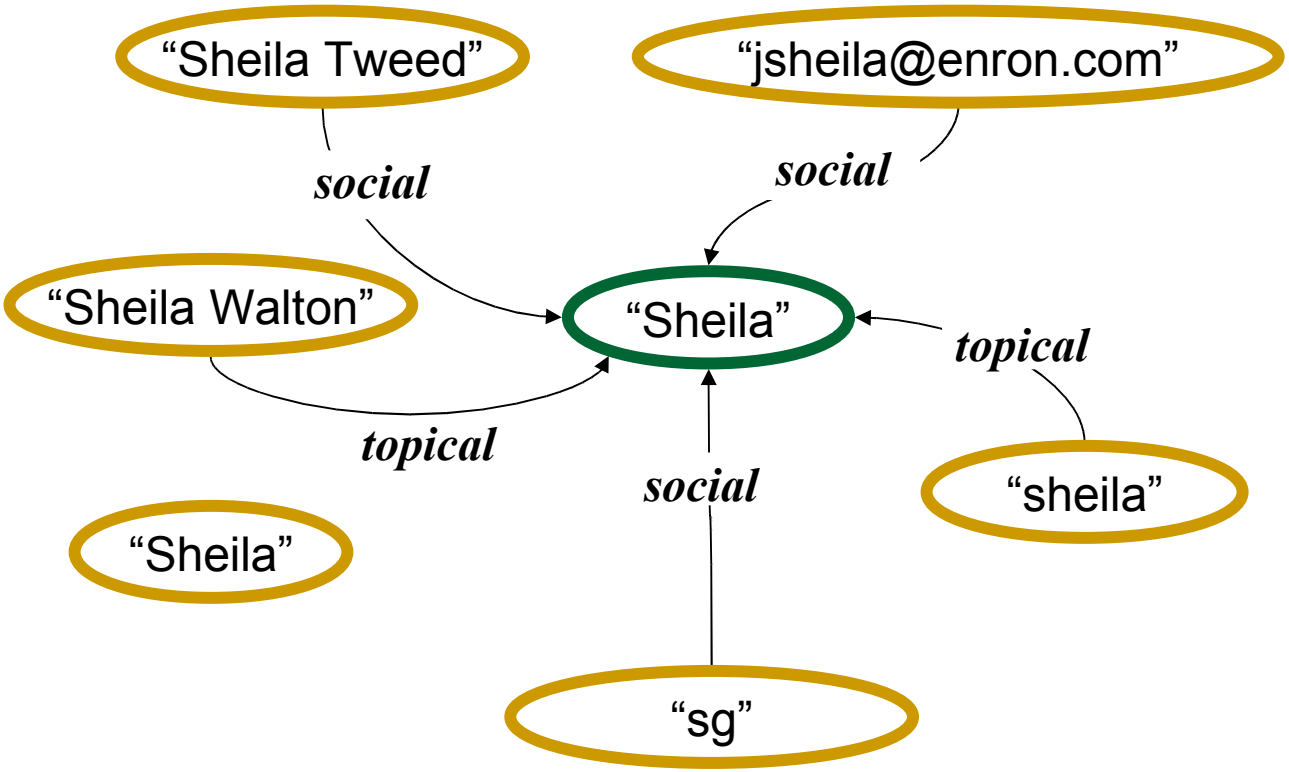


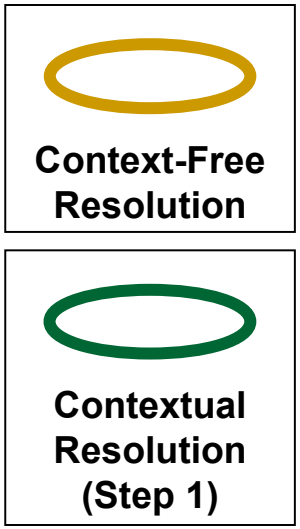

***Goal**: estimate p(c|m, X(m)) and rank accordingly*

# Context-Free Resolution (Step 0)

$$p(c \mid m, X(m)) \approx p(c \mid m) = \frac{p(m \mid c)p(c)}{p(m)}$$

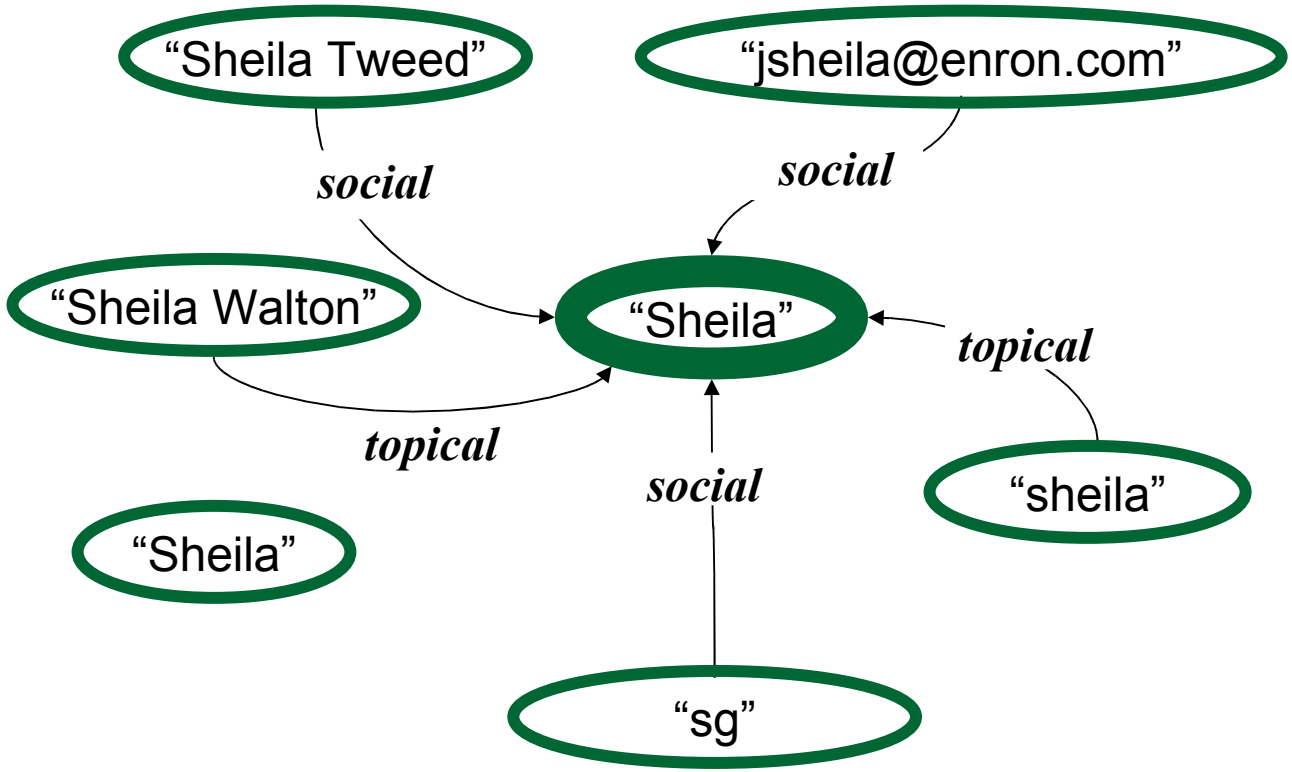
# Contextual Resolution (Step 1)

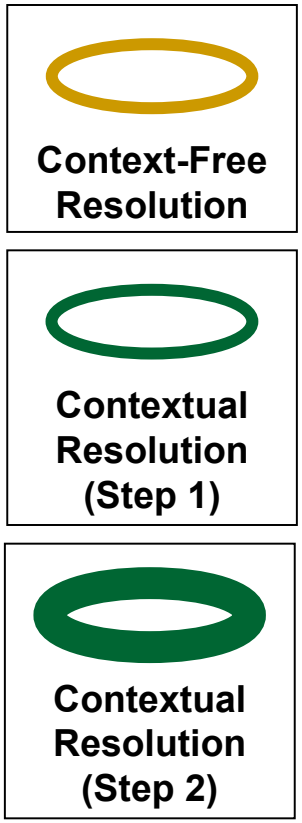



$$p(c \mid m, X(m)) = \frac{p(c, m, X(m))}{p(m, X(m))}$$

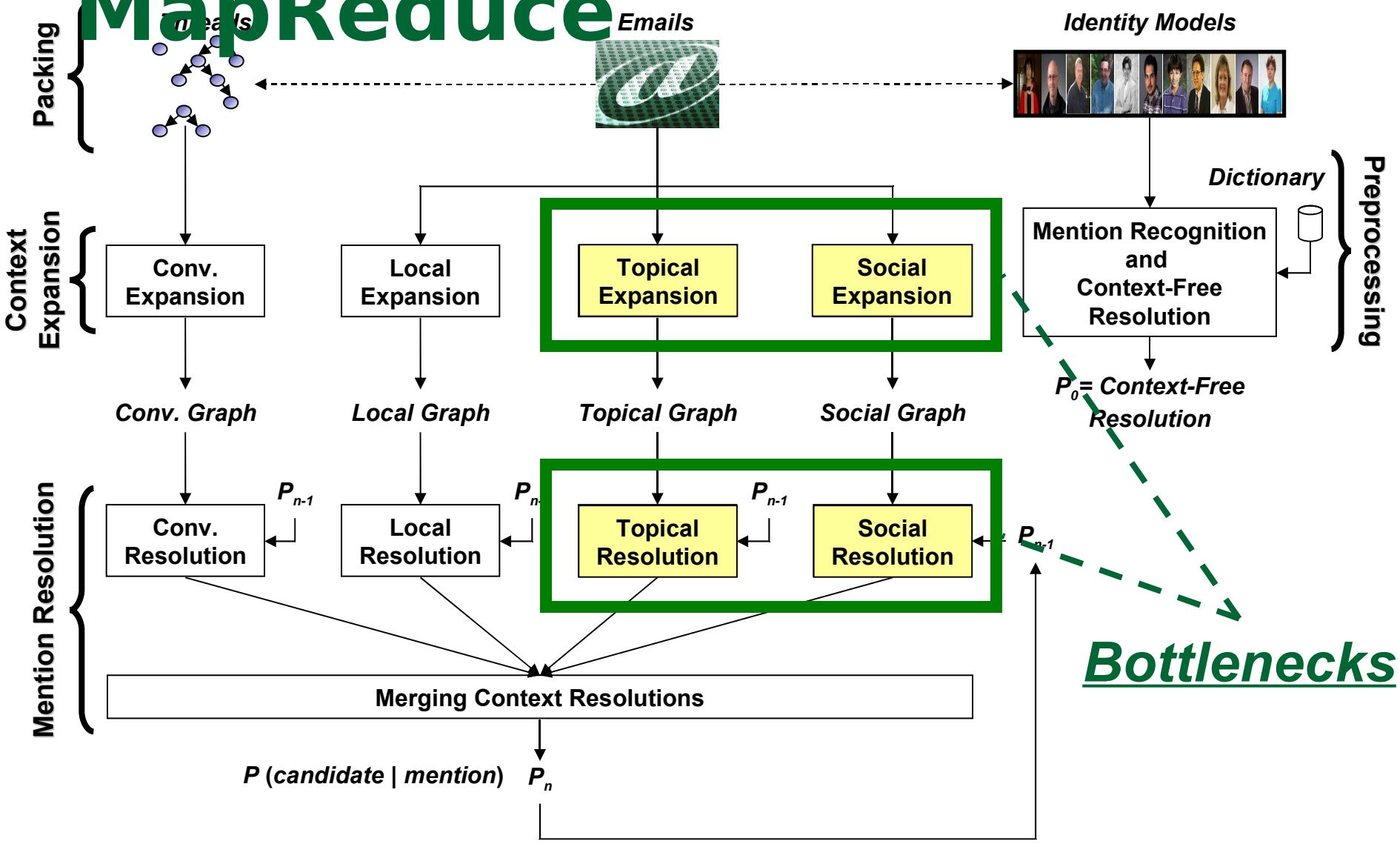
# Contextual Resolution (Step 2)



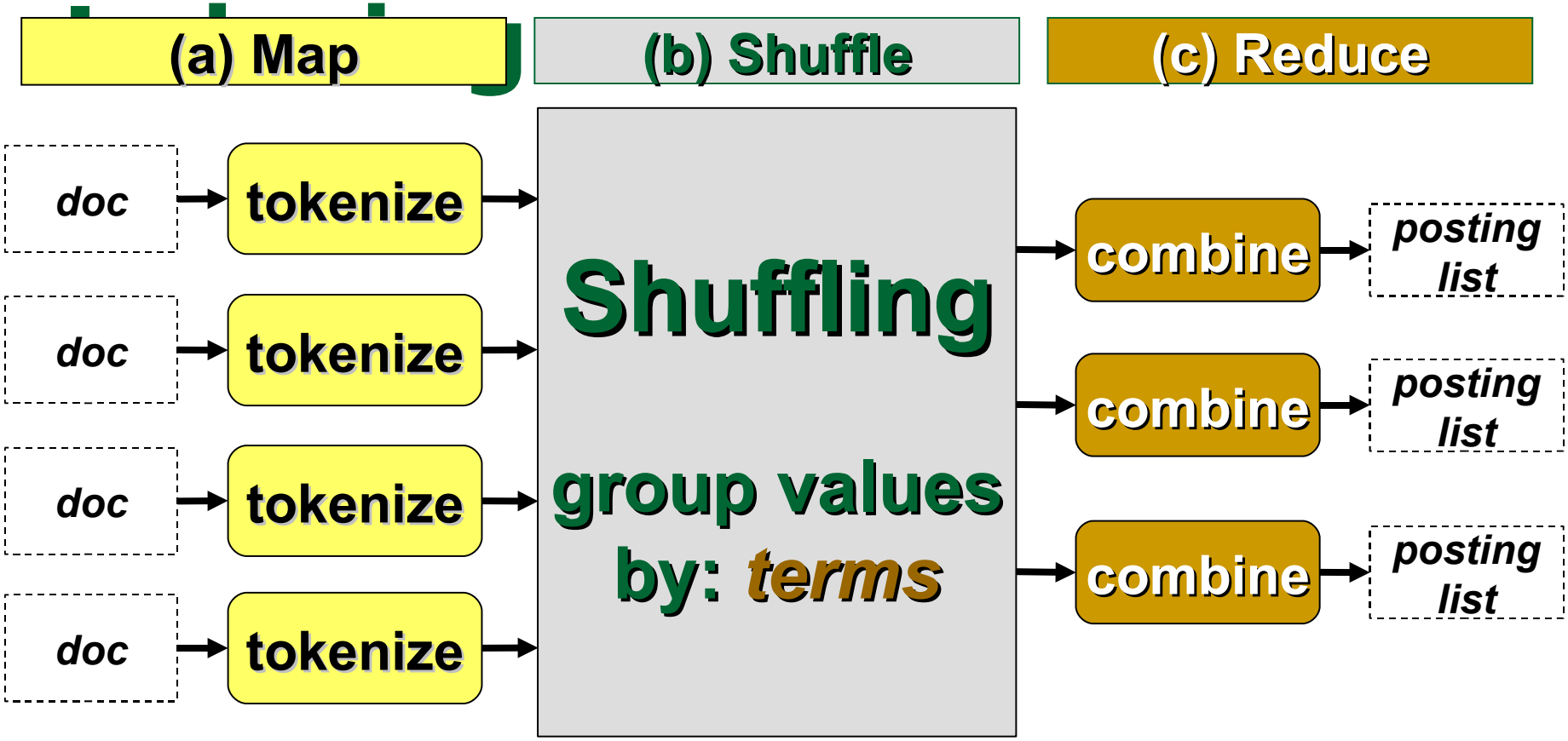




$$p(c \mid m, X(m)) = \frac{p(c, m, X(m))}{p(m, X(m))}$$

# Resolution Using MapReduce

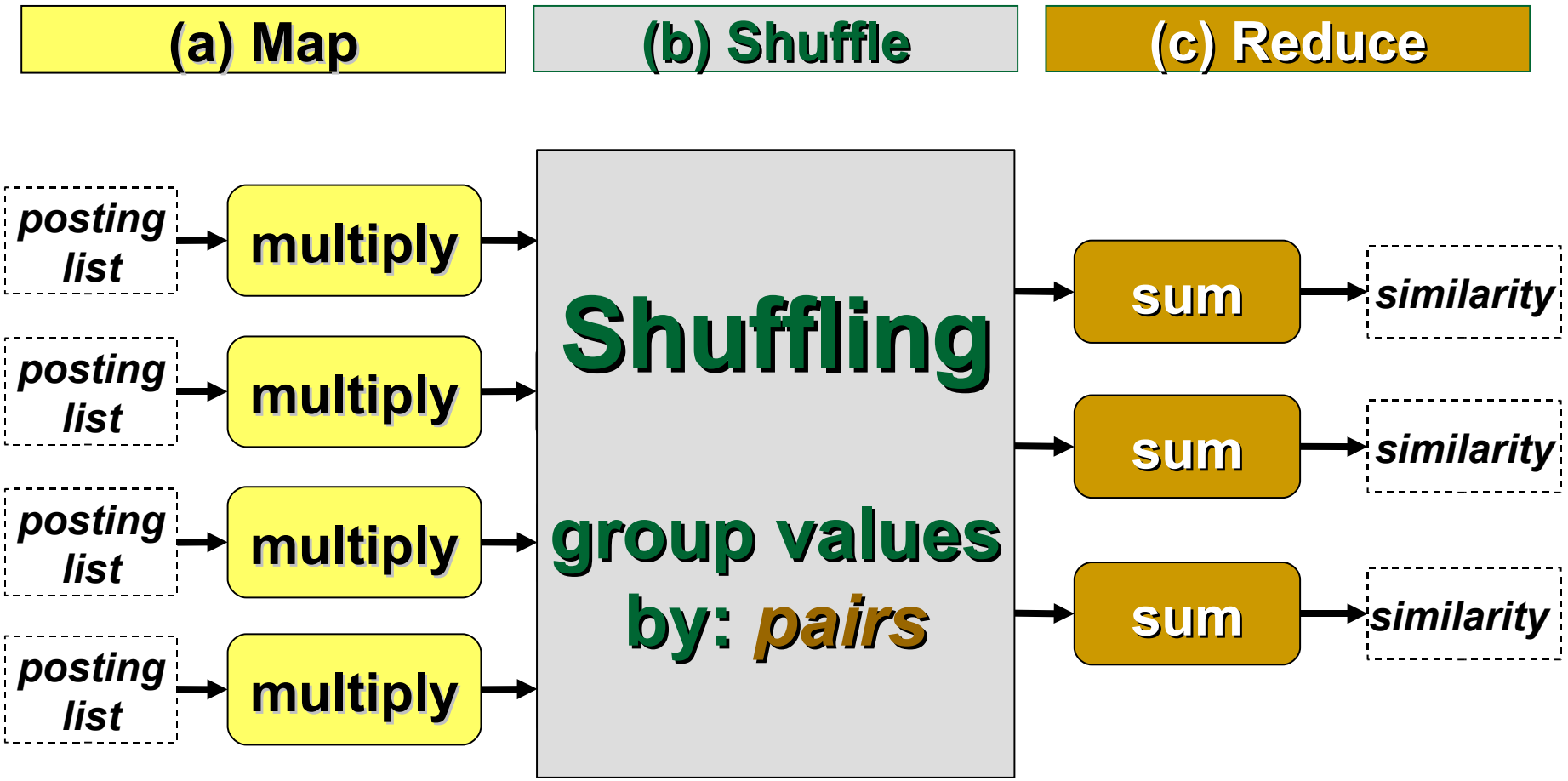


# (a) Standard Inverted




$$sim(d_i, d_j) = \sum_{t \in d_i \cap d_j} term_contrib(t, d_i, d_j)$$

# (b) Pairwise Similarity

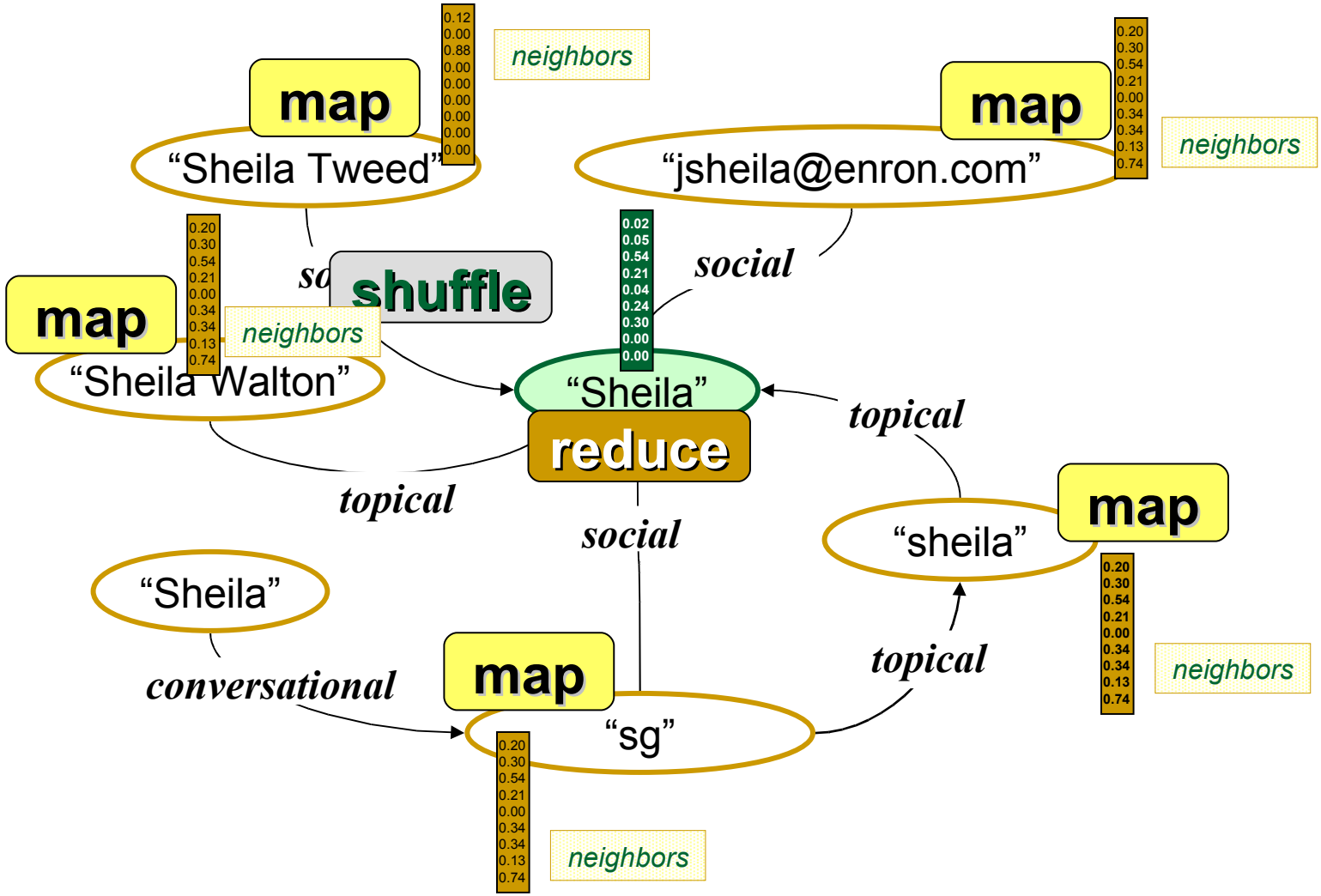



$$sim(d_i, d_j) = \sum_{i \cap d_j} term_contrib(t, d_i, d_j)$$

reduce

map

# Mention Resolution

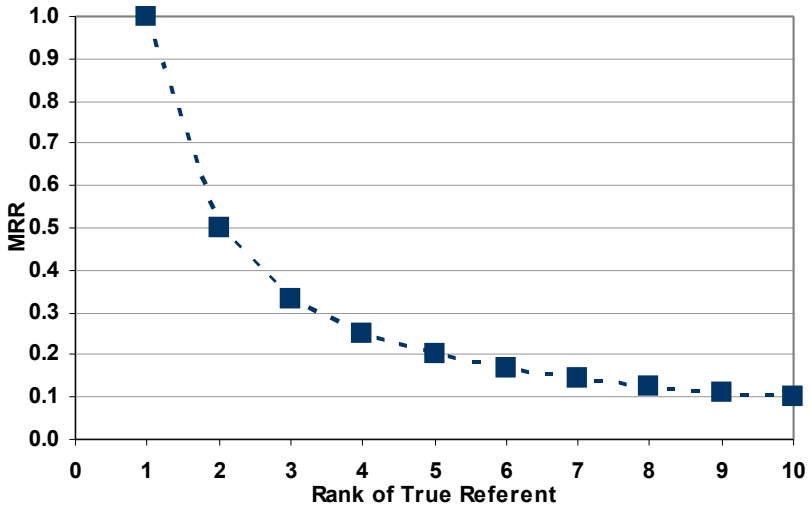


# Outline

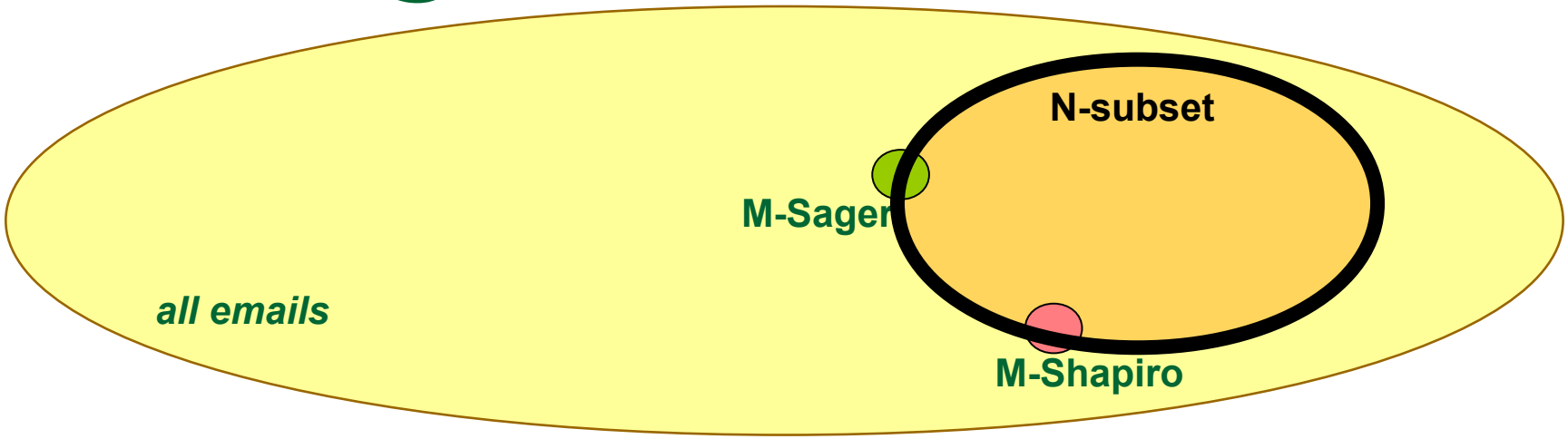
- Introduction and Approach Overview
- Computational Model of Identity
- Context Reconstruction
- Mention Resolution
- ***Evaluation***
- Conclusion

# Experimental Evaluation

- Repeatable and affordable
- Training and testing split
- Test Collection
  - Documents → emails
  - Queries → mentions in specific emails
  - Answers → true referents of those mentions (by humans)
- Evaluation Measure: Mean Reciprocal Rank (MRR)

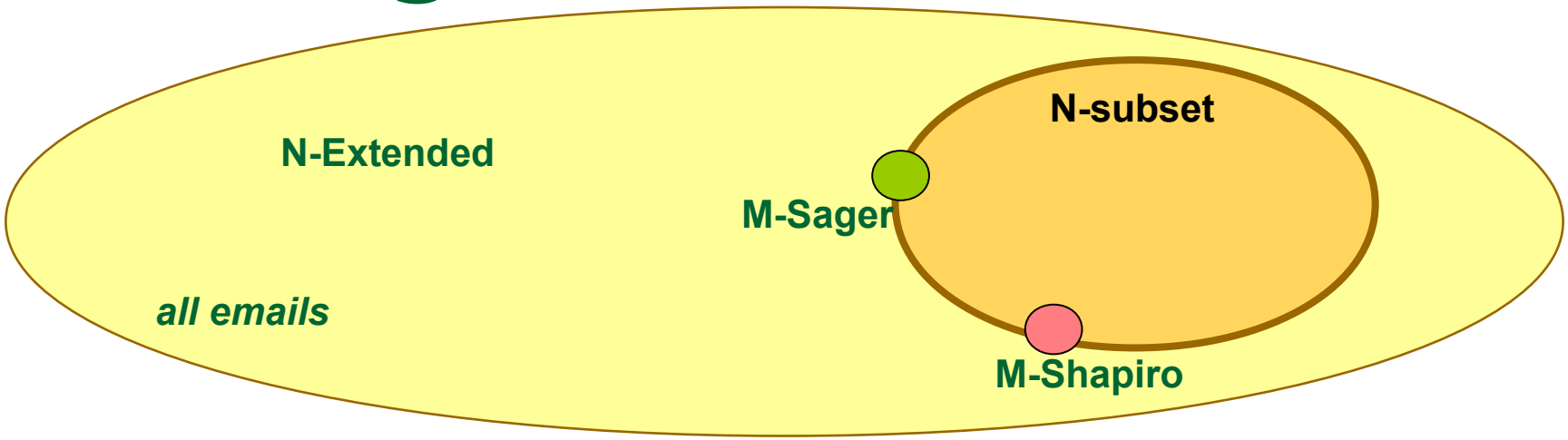


# Existing Test Collections




| Collection | Emails | Queries | Identities | Candidates Med | Candidates Range |
|---|---|---|---|---|---|
| M-Sager | 1,628 | 51 | 627 | 2 | 1-10 |
| M-Shapiro | 974 | 49 | 855 | 4 | 1-16 |
| N-Subset | 54,018 | 78 | 27,340 | 91 | 1-441 |

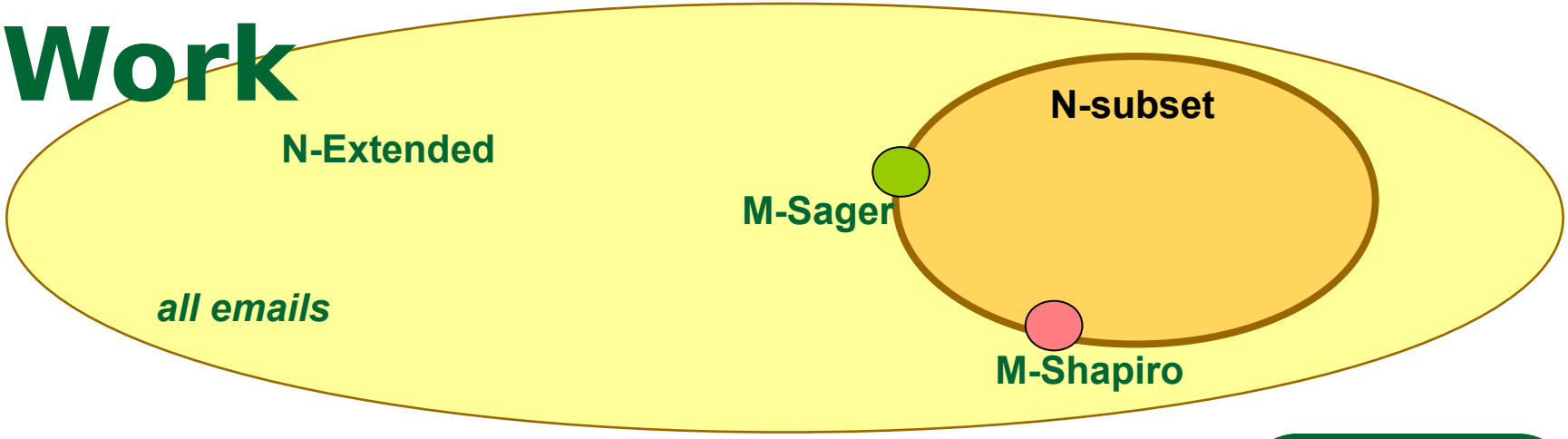
# Existing Test Collections




| Collection | Emails | Queries | Identities | Candidates Med | Candidates Range |
|---|---|---|---|---|---|
| M-Sager | 1,628 | 51 | 627 | 2 | 1-10 |
| M-Shapiro | 974 | 49 | 855 | 4 | 1-16 |
| N-Subset | 54,018 | 78 | 27,340 | 91 | 1-441 |
| N-Extended | 248,451 | 78 | 123,783 | 338 | 3-1,512 |

**Training Collection**

# Comparison w/Previous Work




| Collection | Emails | Queries | Identities | Candidates Med | Candidates Range | MRR Mine | MRR Lit. |
|---|---|---|---|---|---|---|---|
| M-Sager | 1,628 | 51 | 627 | 2 | 1-10 | 0.905 | 0.889 |
| M-Shapiro | 974 | 49 | 855 | 4 | 1-16 | 0.894 | 0.879 |
| N-Subset | 54,018 | 78 | 27,340 | 91 | 1-441 | (0.934) | ~ |
| N-Extended | 248,451 | 78 | 123,783 | 338 | 3-1,512 | 0.933 | - |

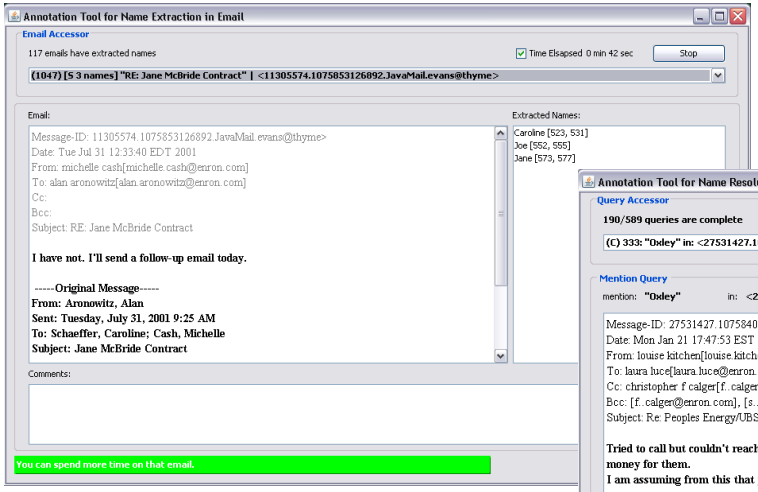
Training Collection

# Developing a New Test Collection

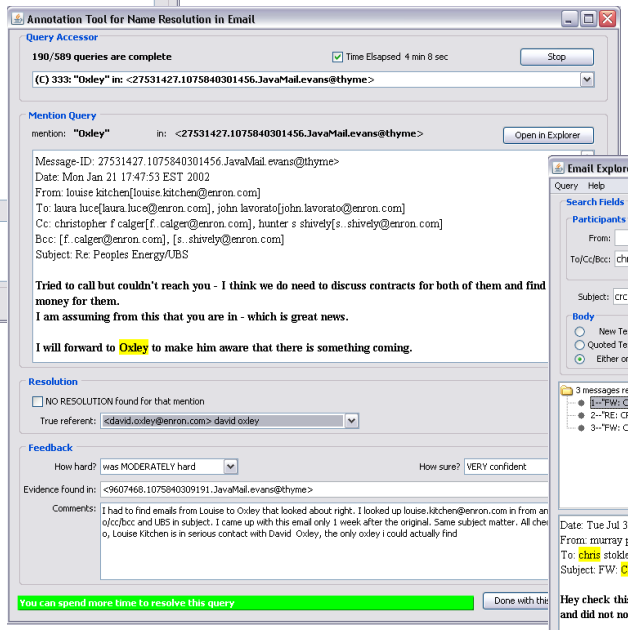
3 annotators

~63 hours

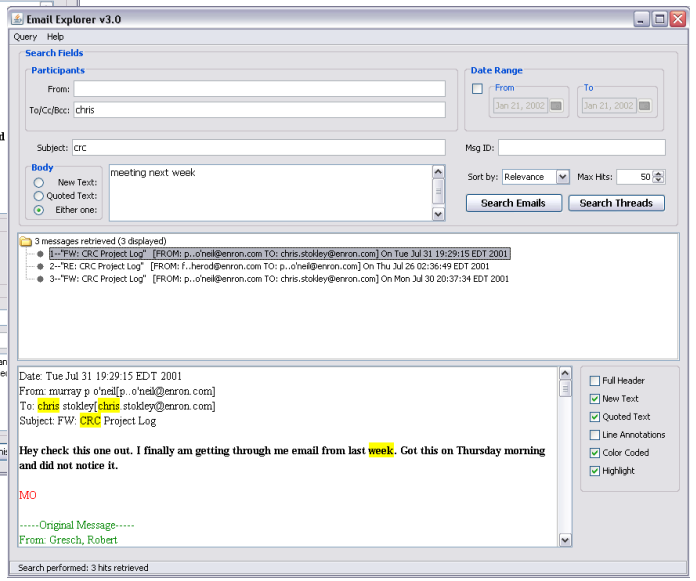
584 queries




*name recognition*



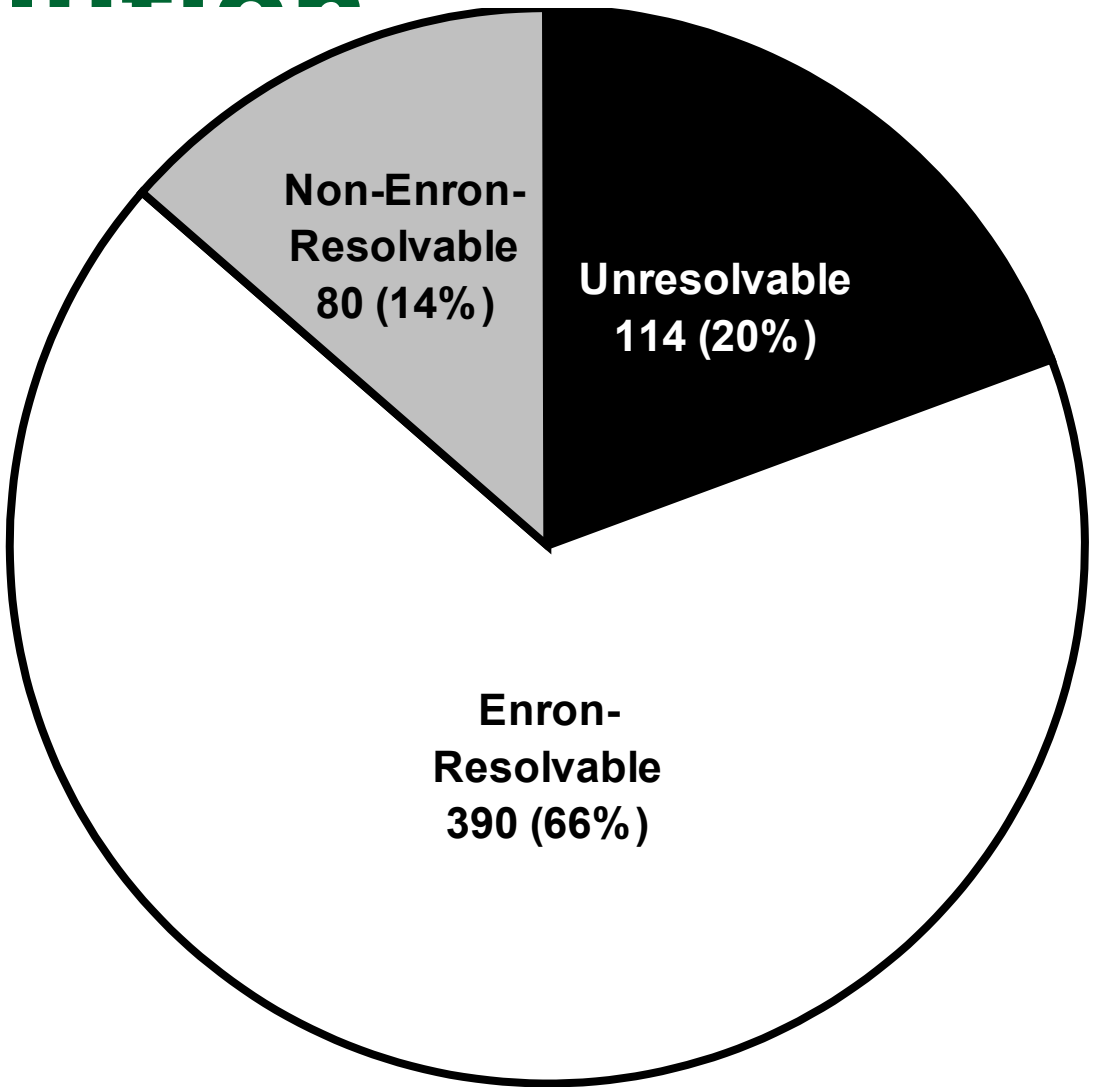

*resolution*




*email search*

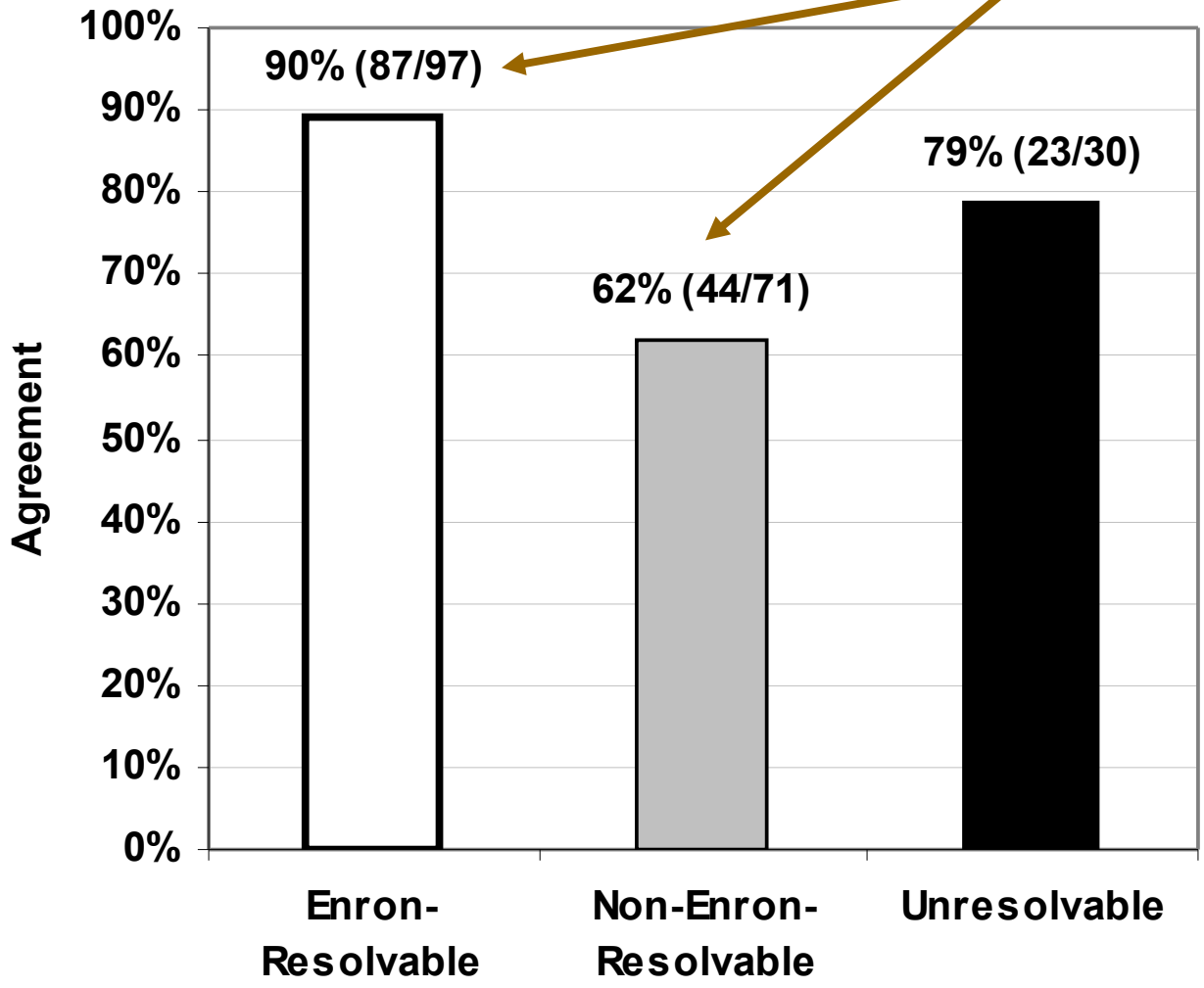
**difficulty**
**confidence**
**time spent**
**comments**

# Distribution Based on Resolution

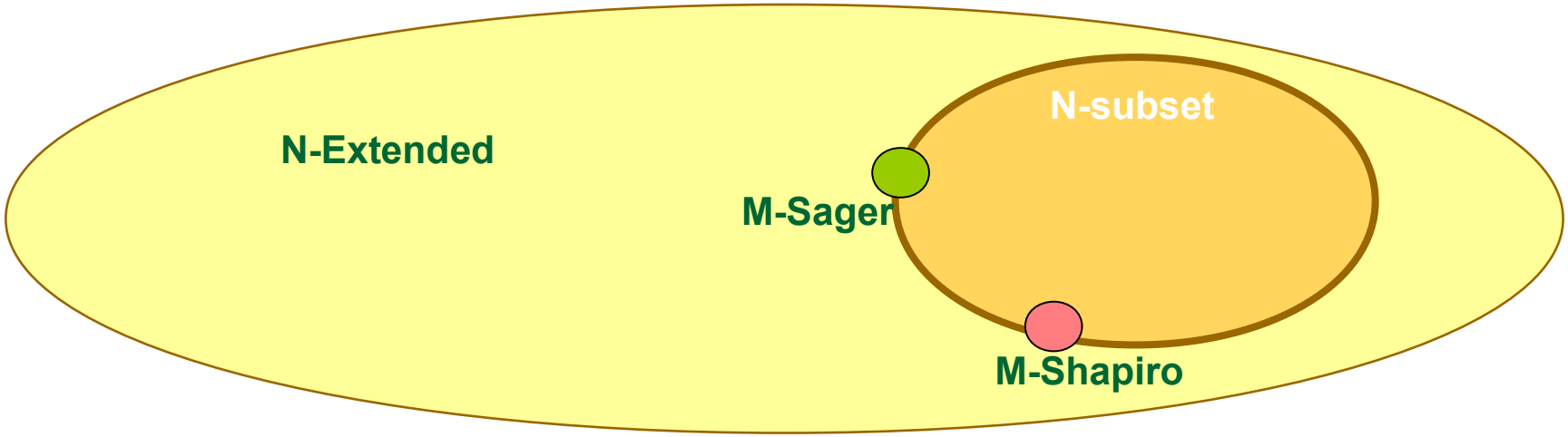


# Inter-Annotator Agreement

**84% Overall Agreement**

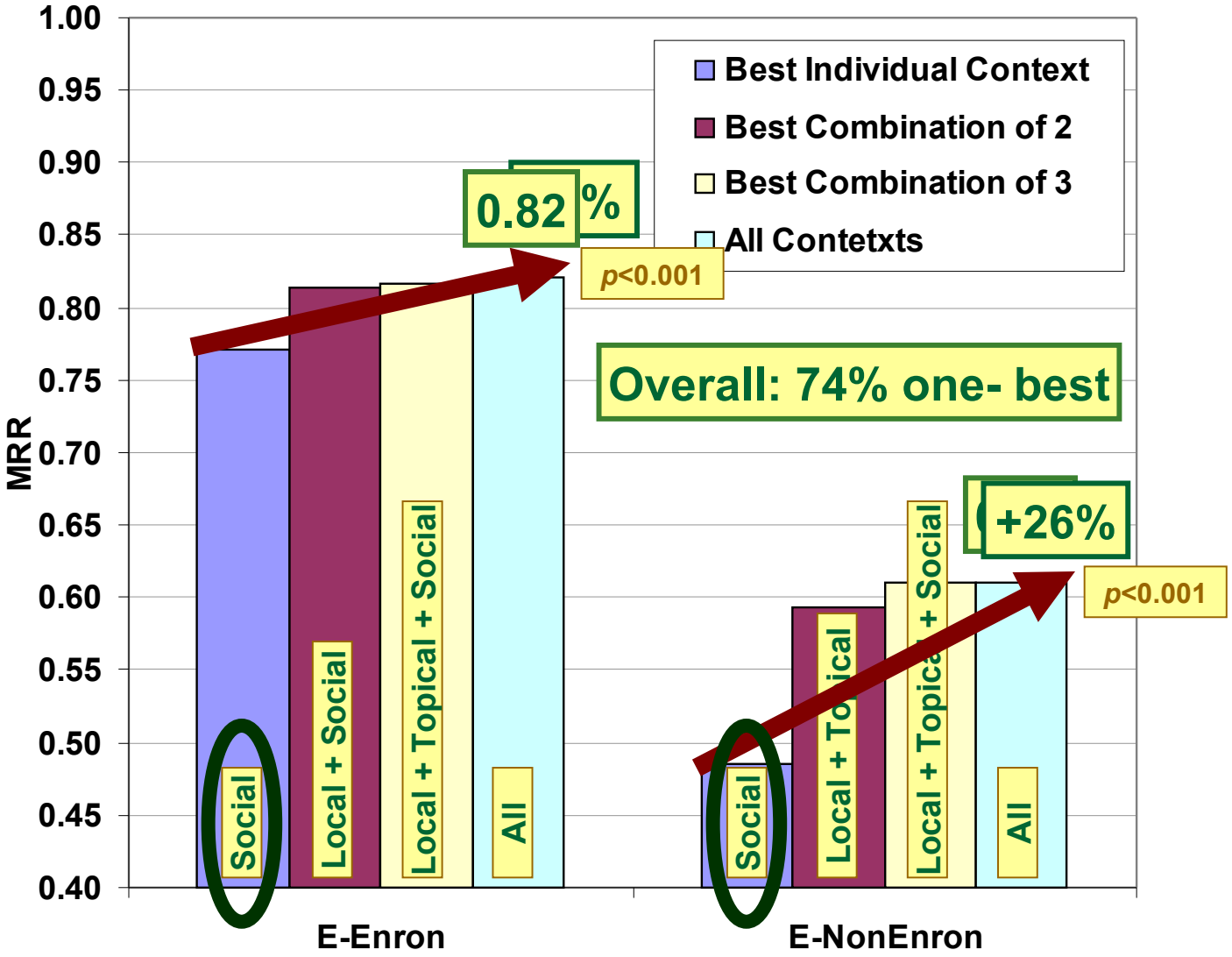


# New Test Collection

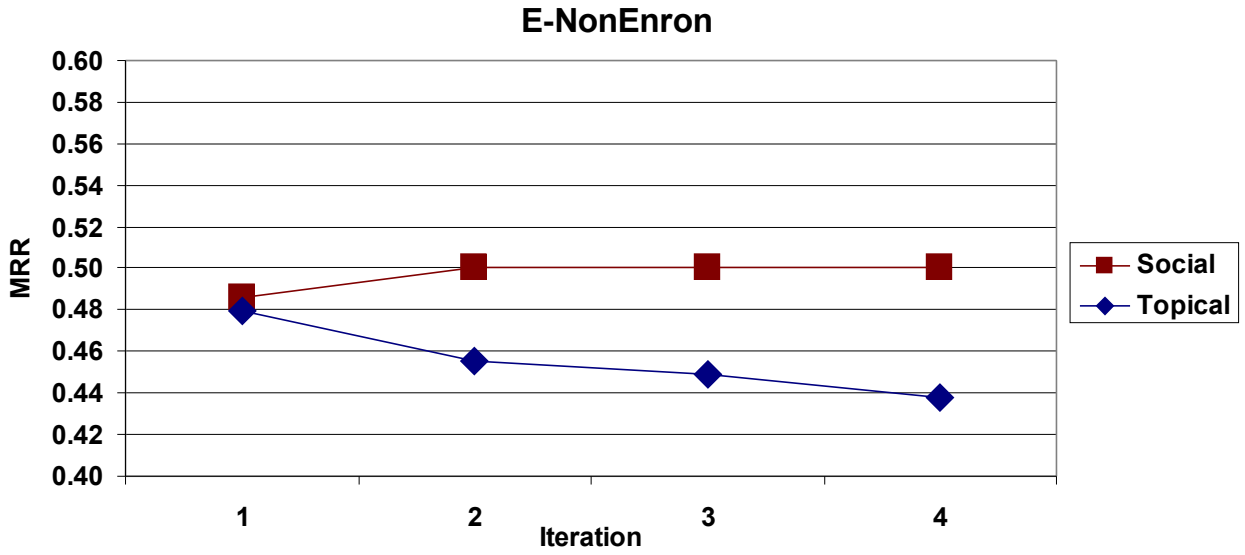


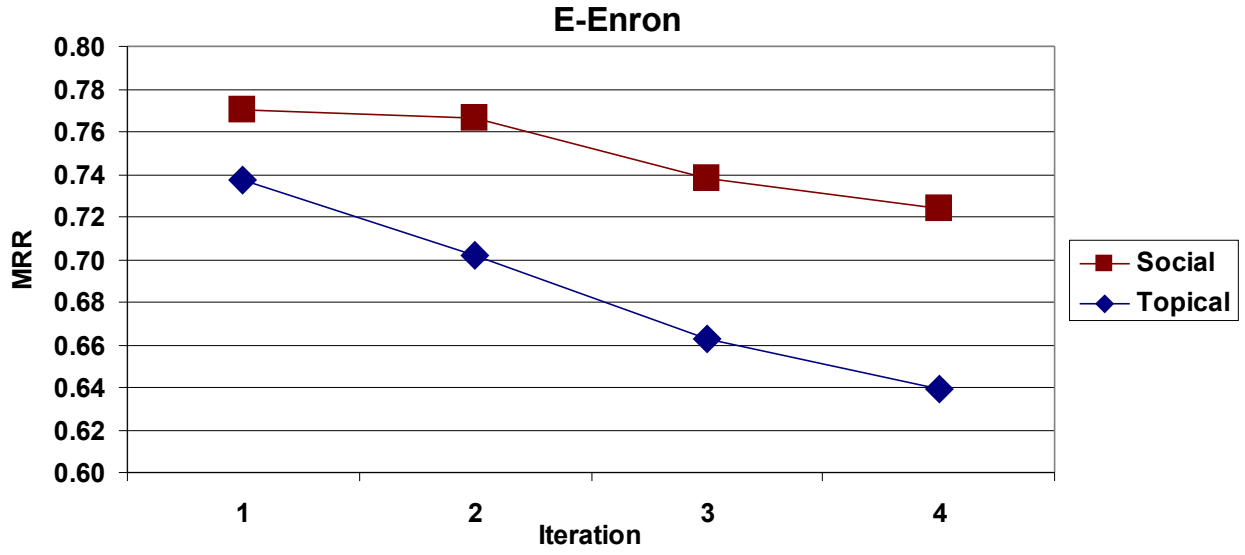

| Collection | Emails | Queries | Identities | Candidates Med | Candidates Range | MRR Mine | MRR Lit. |
|---|---|---|---|---|---|---|---|
| M-Sager | 1,628 | 51 | 627 | 2 | 1-10 | 0.905 | 0.889 |
| M-Shapiro | 974 | 49 | 855 | 4 | 1-16 | 0.894 | 0.879 |
| N-Subset | 54,018 | 78 | 27,340 | 91 | 1-441 | 0.934 | - |
| N-Extended | 248,451 | 78 | 123,783 | 338 | 3-1,512 | 0.933 | - |
| E-All | 248,451 | 470 | 123,783 | 116 | 0-1,512 | 0.785 | - |
| E-Enron | 248,451 | 390 | 123,783 | 121 | 0-1,512 | 0.820 | - |
| E-NonEnron | 248,451 | 90 | 123,783 | 66 | 1-1,512 | 0.611 | - |

# Testing on New Collection



# Iterative Experiments



# Efficiency

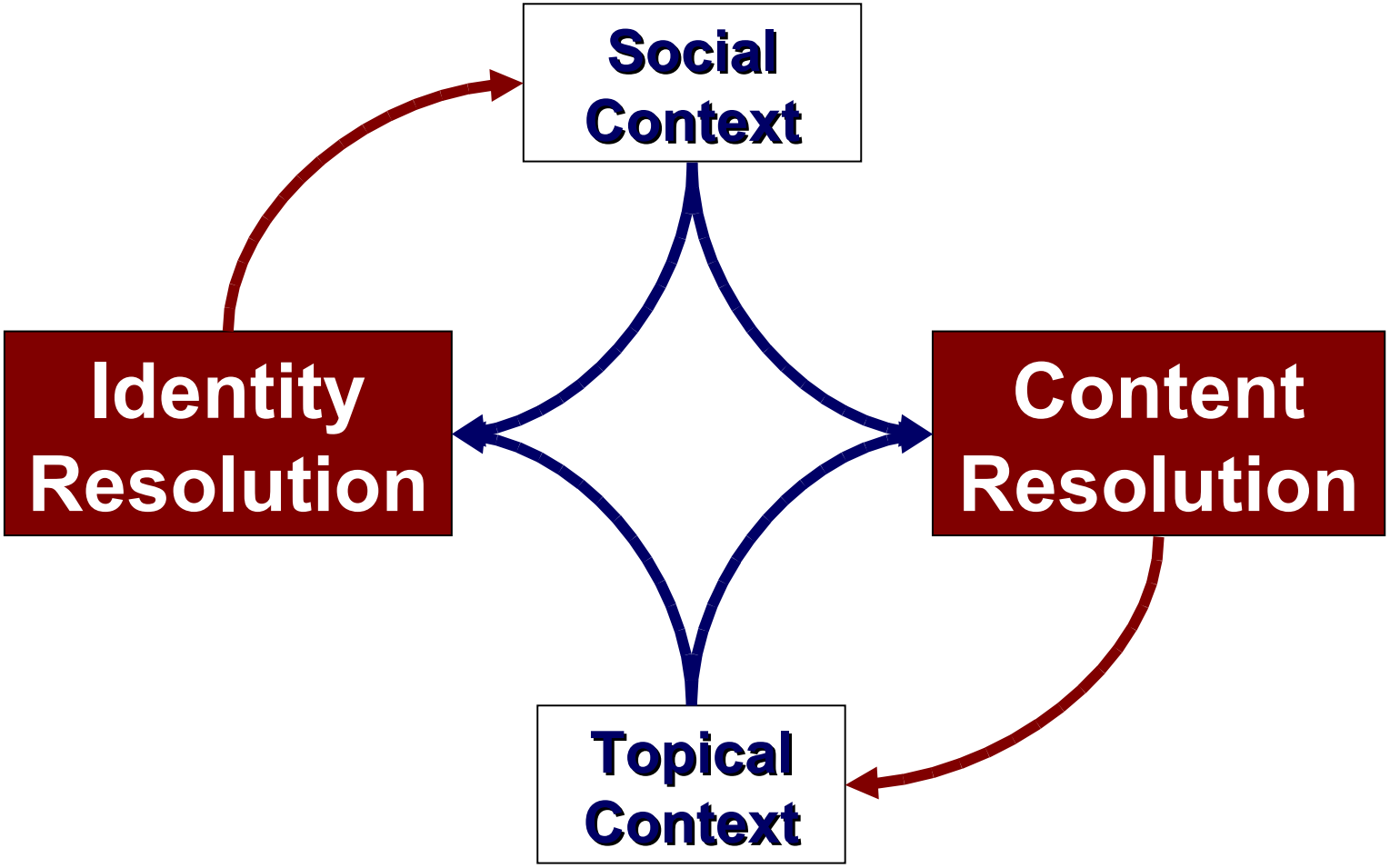
- 0.17.2
  - Open source MapReduce implementation
- 200 processing nodes

**Recognized References**

| | |
|---|---|
| from Main body | 999,291 |
| from Subject | 51,386 |
| from Main Header | 1,642,923 |
| from Quoted Body | 442,099 |
| from Quoted Header | 522,716 |
| Email-addresses | 1,746,636 |
| Single-token Names | 1,331,375 |
| Multi-token Names | 580,407 |

**Time Spent (minutes)**

| | | | | | |
|---|---|---|---|---|---|
| Packing | 48 | Social: Indexing | 1.5 | Topical: Indexing | 1.5 |
| Preprocessing | 5 | Social: Pairwise Sim. | 5 | Topical: Pairwise Sim. | 5-13 |
| Local: Total | 9 | Social: Resolution | 13 | Topical: Resolution | 17-35 |
| Conv.: Total | 10 | Social: Total | 35 | Topical: Total | 45-75 |
| Merging Scores | 10 | | | | |

**End-to-end runs: ~2-3 hours**

# Identity-Content Interplay