Gold standard for Enron
Organizational Hierarchy

Apoorv Agarwal

July 9th 2012

Columbia University

in collaboration with

Adinoyi Omuya (Wisdom), Aaron Harnly, Owen Rambow
Enron Email Corpus

Klimt & Yang ’04
Enron Email Corpus

158 “core” people

Klimt & Yang '04
Enron Email Corpus

158 “core” people

90K “non-core” people

Klimt & Yang ’04
Task: Hierarchy Prediction

Given the Enron email corpus, predict the organizational hierarchy of Enron employees.
Task: Hierarchy Prediction

Given the Enron email corpus, predict the organizational hierarchy of Enron employees.

Rowe et al. '07
Task: Hierarchy Prediction

Given the Enron email corpus, predict the organizational hierarchy of Enron employees.

SNA

Rowe et al. ’07

No quant. evaluation!
Task: Hierarchy Prediction

Given the Enron email corpus, predict the organizational hierarchy of Enron employees.

Rowe et al. '07

SNA

NLP

Bramsen et al. '11
Gilbert '12

A

B

C

D

No quant. evaluation!
Task: Hierarchy Prediction

Given the Enron email corpus, predict the organizational hierarchy of Enron employees.

Rowe et al. '07

SNA

Bramsen et al. '11
Gilbert '12

NLP

No quant. evaluation!

Dominance prediction not hierarchy!
Our Contributions
Our Contributions

1. Present a new Gold Standard for hierarchy prediction
Our Contributions

1. Present a new Gold Standard for hierarchy prediction

2. Present a baseline social network analysis (SNA) based system that out-performs a recent NLP based system by Gilbert '12
Gold Standard creation
Gold Standard creation

- search emails for "org-chart"
Gold Standard creation

- search emails for "org-chart"
- manually convert excel or visio documents into MongoDB
Gold Standard creation

- search emails for “org-chart”
- manually convert excel or visio documents into MongoDB

Hierarchy not just dominance!
Gold Standard creation

- search emails for "org-chart"
- manually convert excel or visio documents into MongoDB

Hierarchy not just dominance!

- 1518 employees (previous: 158 emp.)
Gold Standard creation

• search emails for “org-chart”

• manually convert excel or visio documents into MongoDB

Hierarchy not just dominance!

• 1518 employees (previous: 158 emp.)

• “core” and “non-core” employees (previous: only “core”)
Compare ...

- Gilbert ‘12: NLP based approach for dominance prediction
- Our SNA based baseline
Emails of 132 employees
Emails of 132 employees
Emails of 132 employees

upward

not-upward
Emails of 132 employees

- Feature space: N-gram + feature selection
- SVM with 3-fold cross-validation
Our SNA approach

- Sort based on degree centrality

\[ 1 > 2 > \{3 = 4\} > 5 \]
### Evaluation and Results

<table>
<thead>
<tr>
<th>e11</th>
<th>e12</th>
</tr>
</thead>
<tbody>
<tr>
<td>e21</td>
<td>e22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>e\textsuperscript{m1}</td>
<td>e\textsuperscript{m2}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Gold Standard**

Wednesday, July 11, 12
Evaluation and Results

<table>
<thead>
<tr>
<th>e11</th>
<th>e12</th>
</tr>
</thead>
<tbody>
<tr>
<td>e21</td>
<td>e22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>em1</td>
<td>em2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

| en1 | en2 |

Our SNA approach

Gilbert '12

Gold Standard
Evaluation and Results

<table>
<thead>
<tr>
<th>e11</th>
<th>e12</th>
</tr>
</thead>
<tbody>
<tr>
<td>e21</td>
<td>e22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>en1</td>
<td>en2</td>
</tr>
</tbody>
</table>

Our SNA approach

Gold Standard

Gilbert '12
### Evaluation and Results

<table>
<thead>
<tr>
<th>e11</th>
<th>e12</th>
</tr>
</thead>
<tbody>
<tr>
<td>e21</td>
<td>e22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>em1</td>
<td>em2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

#### Gilbert '12

- 2640 (19%)

#### Our SNA approach

- Gold Standard

---

Wednesday, July 11, 12
Evaluation and Results

<table>
<thead>
<tr>
<th>e11</th>
<th>e12</th>
</tr>
</thead>
<tbody>
<tr>
<td>e21</td>
<td>e22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>em1</td>
<td>em2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

2640 (19%) 82.37%

Gilbert '12

Our SNA approach

Gold Standard
# Evaluation and Results

![Diagram](image.png)

<table>
<thead>
<tr>
<th>Gold Standard</th>
<th>Our SNA Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>e11</td>
<td>e12</td>
</tr>
<tr>
<td>e21</td>
<td>e22</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>em1</td>
<td>em2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>en1</td>
<td>en2</td>
</tr>
</tbody>
</table>

- **Gilbert '12**:
  - Evaluation: 82.37%
  - Results: 2640 (19%)

- **Results**: 87.58%
Evaluation and Results

2640 (19%)

82.37%

87.58%

Gilbert '12

13,724

Gold Standard

Our SNA approach

Wednesday, July 11, 12
### Evaluation and Results

<table>
<thead>
<tr>
<th></th>
<th>e11</th>
<th>e12</th>
</tr>
</thead>
<tbody>
<tr>
<td>e21</td>
<td>e22</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>em1</td>
<td>em2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gilbert '12</th>
<th>2640 (19%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.37%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gold Standard</th>
<th>13,724</th>
</tr>
</thead>
<tbody>
<tr>
<td>83.8%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our SNA approach</th>
<th>87.58%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gilbert '12

Wednesday, July 11, 12
Conclusion
Conclusion

* Introduce a large gold standard for hierarchy prediction of Enron employees (contact apoorv@cs.columbia.edu)
Conclusion

- Introduce a large gold standard for hierarchy prediction of Enron employees (contact apoorv@cs.columbia.edu)
- Show for our gold standard, a simple SNA based approach outperforms a recent NLP based approach
Conclusion

- Introduce a large gold standard for hierarchy prediction of Enron employees (contact apoorv@cs.columbia.edu)
- Show for our gold standard, a simple SNA based approach outperforms a recent NLP based approach

Thanks!
UB-NLP

• <A, B> in Gold: 13,724 (G)

• <A, B> where A emails B: 2604 (T)

• UB-NLP: \[(2604 + 5560)/13724 = 59.6\%\]
**Task: Hierarchy Prediction**

Rowe et al. '07

Brams en et al. '11
Gilbert '12

- Ken Lay (CEO)
  - G. Whalley (COO)
    - D. Delainey (Co-COO)
    - M. Haedicke (MD)
  - S. Kean (VP)
    - ..........
Task: Hierarchy Prediction

Rowe et al. '07

Bramsen et al. '11
Gilbert '12
Task: Hierarchy Prediction

Rowe et al. '07

Ken Lay (CEO)

G. Whalley (COO)

D. Delainey (Co-COO)

M. Haedicke (MD)

S. Kean (VP)

...........

Bramsuen et al. '11

Gilbert '12
Our Contributions

New Gold Standard

- 1518 employees
- 13K dominance pairs (\(<A,B>\) tuples):
  <Lay, Whalley>; <Whalley, Delainey>;
  <Lay, Kean> ...
- both “core” and “non-core”
Our Contributions

New Gold Standard

- 1518 employees
- 13K dominance pairs (<A,B> tuples):  
  <Lay, Whalley>; <Whalley, Delainey>;  
  <Lay, Kean> ...
- both “core” and “non-core”

SNA system out-performs NLP system

- Current state-of-the-art NLP based system (Gilbert ’12)
- Social network analysis based system (SNA)