Correlation Clustering
Bounding and Comparing Methods Beyond ILP

Micha Elsner and Warren Schudy

Department of Computer Science
Brown University

May 26, 2009
Document clustering

rec.motorcycles

soc.religion.christian
Document clustering: pairwise decisions

rec.motorcycles  soc.religion.christian
Document clustering: partitioning

rec.motorcycles

soc.religion.christian
How good is this?

rec.motorcycles

soc.religion.christian

Cut green arc

Uncut red arc
Correlation clustering

Given green edges $w^+$ and red edges $w^-$...
Partition to minimize disagreement.

$$\min_x x_{ij} w_{ij}^- + (1 - x_{ij}) w_{ij}^+$$

s.t. $x_{ij}$ form a consistent clustering
relation must be transitive: $x_{ij}$ and $x_{jk} \rightarrow x_{ik}$

Minimization is NP-hard (Bansal et al. ‘04).
How do we solve it?
ILP scalability

ILP:

- $O(n^2)$ variables (each pair of points).
- $O(n^3)$ constraints (triangle inequality).
- Solvable for about 200 items.

Good enough for single-document coreference or generation. Beyond this, need something else.
Previous applications

- Coreference resolution (Soon et al. ‘01), (Ng+Cardie ‘02),
  (McCallum+Wellner ‘04), (Finkel+Manning ‘08).
- Grouping named entities (Cohen+Richman ‘02).
- Content aggregation (Barzilay+Lapata ‘06).
- Topic segmentation (Malioutov+Barzilay ‘06).
- Chat disentanglement (Elsner+Charniak ‘08).

Solutions: **heuristic**, **ILP**, **approximate**, **special-case**,
This talk

*Not* about when you should use correlation clustering.

- When you can’t use ILP, what should you do?
- How well can you do in practice?
- Does the objective predict real performance?
This talk

Not about when you should use correlation clustering.

- When you can’t use ILP, what should you do?
  - Greedy voting scheme, then local search.
- How well can you do in practice?
- Does the objective predict real performance?
This talk

*Not about when you should use correlation clustering.*

- When you can’t use ILP, what should you do?
  - Greedy voting scheme, then local search.
- How well can you do in practice?
  - Reasonably close to optimal.
- Does the objective predict real performance?
This talk

*Not* about when you should use correlation clustering.

- When you can’t use ILP, what should you do?
  - Greedy voting scheme, then local search.
- How well can you do in practice?
  - Reasonably close to optimal.
- Does the objective predict real performance?
  - Often, but not always.
Overview

Motivation

Algorithms

Bounding

Task 1: Twenty Newsgroups

Task 2: Chat Disentanglement

Conclusions
Algorithms

Some fast, simple algorithms from the literature.

**Greedy algorithms**
- First link
- Best link
- Voted link
- Pivot

**Local search**
- Best one-element move (BOEM)
- Simulated annealing
Greedy algorithms

Step through the nodes in random order. Use a linking rule to place each unlabeled node.
Previously assigned

the most recent positive arc
Best link (Ng+Cardie ‘02)

Previously assigned

Next node

the highest scoring arc
Voted link

Previously assigned

Next node

the cluster with **highest arc sum**
Pivot (Ailon+al ‘08)

Create each whole cluster at once. Take the first node as the pivot.

pivot node

add all nodes with positive arcs
Choose the next unlabeled node as the pivot.

new pivot node

add all nodes with positive arcs
Local searches

One-element moves change the label of a single node.

Current state
Local searches

One-element moves change the label of a single node.

- Greedily: best one-element move (BOEM)
- Stochastically (annealing)
Overview

Motivation

Algorithms

Bounding

Task 1: Twenty Newsgroups

Task 2: Chat Disentanglement

Conclusions
Why bound?

Objective value

Worse

- All singletons clustering
- Various heuristics

Better
Why bound?

- Better
- Worse
- All singletons clustering
- Various heuristics
- Optimal
Why bound?

- objective value
  - worse
    - all singletons clustering
  - better
    - optimal
    - various heuristics
Why bound?

- Objective value
  - worse
    - all singletons clustering
    - various heuristics
    - optimal
    - lower bound
  - better
Trivial bound from previous work

rec.motorcycles

soc.religion.christian

cut all red arcs

no transitivity!
Represent each item by an $n$-dimensional basis vector:

For an item in cluster $c$, vector $r$ is:

\[
\begin{pmatrix}
0, 0, \ldots, 0, 1, 0, \ldots, 0 \\
c-1 \\ n-c
\end{pmatrix}
\]

For two items clustered together, $r_i \cdot r_j = 1$. Otherwise $r_i \cdot r_j = 0$. 

Semidefinite programming bound (Charikar et al.‘05)
Semidefinite programming bound (Charikar et al. ‘05)

Represent each item by an $n$-dimensional basis vector:

For an item in cluster $c$, vector $r$ is:

$$(0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

For two items clustered together, $r_i \cdot r_j = 1$. Otherwise $r_i \cdot r_j = 0$.

Relaxation

Allow $r_i$ to be any real-valued vectors with:

- Unit length.
- All products $r_i \cdot r_j$ non-negative.
Semidefinite programming bound (2)

Semidefinite program (SDP)

\[
\min_r \sum (r_i \cdot r_j) w_{ij}^- + (1 - r_j \cdot r_j) w_{ij}^+
\]

\[
r_i \cdot r_i = 1 \quad \forall i
\]

\[
s.t. \quad r_i \cdot r_j \geq 0 \quad \forall i \neq j
\]

Objective and constraints are linear in the dot products of the \(r_i\).
Semidefinite programming bound (2)

Semidefinite program (SDP)

\[
\min_x \sum x_{ij} w_{ij}^- + (1 - x_{ij}) w_{ij}^+
\]

\[
x_{ij} = 1 \quad \forall i
\]

s.t. \quad x_{ij} \geq 0 \quad \forall i \neq j

Objective and constraints are linear in the dot products of the \( r_i \).

Replace dot products with variables \( x_{ij} \).

New constraint: \( x_{ij} \) must be dot products of some vectors \( r \)!
Semidefinite programming bound (2)

Semidefinite program (SDP)

$$\min_x \sum x_{ij} w_{ij}^- + (1 - x_{ij}) w_{ij}^+$$

$$x_{ij} = 1 \quad \forall i$$

s.t. $$x_{ij} \geq 0 \quad \forall i \neq j$$

matrix $X$ PSD

Objective and constraints are linear in the dot products of the $r_i$. Replace dot products with variables $x_{ij}$. New constraint: $x_{ij}$ must be dot products of some vectors $r_i$. Equivalent: matrix $X$ is positive semi-definite.
Solving the SDP

- SDP bound previously studied in theory.
- We actually solve it!
- Conic Bundle method (Helmberg ‘00).
  - Scales to several thousand points.
- Iteratively improves bounds.
  - Run for 60 hrs.
Bounds

- **objective value**
  - worse
  - (100%) all singletons clustering
  - various heuristics
  - optimal
  - SDP bound
  - (0%) trivial bound
  - better
Overview

Motivation

Algorithms

Bounding

Task 1: Twenty Newsgroups

Task 2: Chat Disentanglement

Conclusions
Twenty Newsgroups

A standard clustering dataset.
Subsample of 2000 posts.

Hold out four newsgroups to train a pairwise classifier:
A standard clustering dataset.
Subsample of 2000 posts.

Hold out four newsgroups to train a pairwise classifier:

Is this message pair from the same newsgroup?

- Word overlap (bucketed by IDF).
- Cosine in LSA space.
- Overlap in subject lines (by IDF).

Max-ent model with F-score of 29%.
Affinity matrix

Ground truth

Affinities
## Results

<table>
<thead>
<tr>
<th>Bounds</th>
<th>Trivial bound</th>
<th>Objective</th>
<th>F-score</th>
<th>One-to-one</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDP bound</td>
<td>0%</td>
<td>51.1%</td>
<td></td>
</tr>
<tr>
<td>Bounds</td>
<td>Trivial bound</td>
<td>Objective</td>
<td>F-score</td>
<td>One-to-one</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------</td>
<td>-----------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td>SDP bound</td>
<td>0%</td>
<td>51.1%</td>
<td></td>
</tr>
<tr>
<td>Local search</td>
<td>Vote/BOEM</td>
<td>55.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sim Anneal</td>
<td>56.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pivot/BOEM</td>
<td>56.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Best/BOEM</td>
<td>57.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First/BOEM</td>
<td>57.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BOEM</td>
<td>60.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounds</td>
<td>Objective</td>
<td>F-score</td>
<td>One-to-one</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>---------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Trivial bound</td>
<td>0%</td>
<td>51.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDP bound</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Local search</td>
<td>Vote/BOEM</td>
<td>55.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sim Anneal</td>
<td>56.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pivot/BOEM</td>
<td>56.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Best/BOEM</td>
<td>57.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First/BOEM</td>
<td>57.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BOEM</td>
<td>60.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greedy</td>
<td>Vote</td>
<td>59.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pivot</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Best</td>
<td>138%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First</td>
<td>619%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Bounds</th>
<th>Objective</th>
<th>F-score</th>
<th>One-to-one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trivial bound</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDP bound</td>
<td>51.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vote/BOEM</td>
<td>55.8%</td>
<td>33</td>
<td>41</td>
</tr>
<tr>
<td>Sim Anneal</td>
<td>56.3%</td>
<td>31</td>
<td>36</td>
</tr>
<tr>
<td>Pivot/BOEM</td>
<td>56.6%</td>
<td>32</td>
<td>39</td>
</tr>
<tr>
<td>Best/BOEM</td>
<td>57.6%</td>
<td>31</td>
<td>38</td>
</tr>
<tr>
<td>First/BOEM</td>
<td>57.9%</td>
<td>30</td>
<td>36</td>
</tr>
<tr>
<td>BOEM</td>
<td>60.1%</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>Vote</td>
<td>59.0%</td>
<td>29</td>
<td>35</td>
</tr>
<tr>
<td>Pivot</td>
<td>100%</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>Best</td>
<td>138%</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>First</td>
<td>619%</td>
<td>11</td>
<td>8</td>
</tr>
</tbody>
</table>
Objective vs. metrics
Overview

Motivation

Algorithms

Bounding

Task 1: Twenty Newsgroups

Task 2: Chat Disentanglement

Conclusions
Chat disentanglement

Separate IRC chat log into threads of conversation. 800 utterance dataset and max-ent classifier from (Elsner+Charniak ‘08). Classifier is run on pairs less than 129 seconds apart.

<table>
<thead>
<tr>
<th></th>
<th>question: what could cause linux not to find a dhcp server?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christiana</td>
<td>Arlie: I dont eat bananas.</td>
</tr>
<tr>
<td>Renate</td>
<td>Ruthe, the fact that there isn’t one?</td>
</tr>
<tr>
<td>Arlie</td>
<td>Christiana, you should, they have lots of potassium goodness</td>
</tr>
<tr>
<td>Ruthe</td>
<td>Renate, xp computer finds it</td>
</tr>
<tr>
<td>Renate</td>
<td>eh? dunno then</td>
</tr>
<tr>
<td>Christiana</td>
<td>Arlie: I eat cardboard boxes because of the fibers.</td>
</tr>
</tbody>
</table>
Affinity matrix
## Results

<table>
<thead>
<tr>
<th>Bounds</th>
<th>Trivial bound</th>
<th>SDP bound</th>
<th>Objective</th>
<th>Local</th>
<th>One-to-one</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0%</td>
<td>13.0%</td>
<td></td>
</tr>
</tbody>
</table>
### Results

<table>
<thead>
<tr>
<th>Bounds</th>
<th>Objective</th>
<th>Local</th>
<th>One-to-one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trivial bound</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDP bound</td>
<td>13.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First/BOEM</td>
<td>19.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vote/BOEM</td>
<td>20.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sim Anneal</td>
<td>20.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best/BOEM</td>
<td>21.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOEM</td>
<td>21.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pivot/BOEM</td>
<td>22.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounds</td>
<td>Objective</td>
<td>Local</td>
<td>One-to-one</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>Trivial bound</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDP bound</td>
<td>13.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First/BOEM</td>
<td>19.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vote/BOEM</td>
<td>20.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sim Anneal</td>
<td>20.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best/BOEM</td>
<td>21.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOEM</td>
<td>21.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pivot/BOEM</td>
<td>22.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vote</td>
<td>26.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>37.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pivot</td>
<td>44.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>58.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounds</td>
<td>Trivial bound</td>
<td>Objective</td>
<td>Local</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>-----------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>SDP bound</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.0%</td>
<td></td>
</tr>
<tr>
<td>Local search</td>
<td>First/BOEM</td>
<td>19.3%</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Vote/BOEM</td>
<td>20.0%</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Sim Anneal</td>
<td>20.3%</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Best/BOEM</td>
<td>21.3%</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>BOEM</td>
<td>21.5%</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Pivot/BOEM</td>
<td>22.0%</td>
<td>72</td>
</tr>
<tr>
<td>Greedy</td>
<td>Vote</td>
<td>26.3%</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Best</td>
<td>37.1%</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Pivot</td>
<td>44.4%</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>First</td>
<td>58.3%</td>
<td>62</td>
</tr>
</tbody>
</table>
Objective doesn’t always predict performance

Most edges have weight .5:
- Some systems link too much.
- Doesn’t affect local metric much...
- But global metric suffers.

In this situation, useful to have an external measure of quality.

Better inference is still useful:
- Vote/BOEM 12% better than (Elsner+Charniak ‘08).
- Exact same classifier!
Overview

Motivation

Algorithms

Bounding

Task 1: Twenty Newsgroups

Task 2: Chat Disentanglement

Conclusions
Conclusions

- Always use local search!
- Best greedy algorithm is voting.
Conclusions

- Always use local search!
- Best greedy algorithm is voting.
- SDP provides a tighter bound than previous work.
- Best heuristics are not too far from optimal.
Conclusions

- Always use local search!
- Best greedy algorithm is voting.
- SDP provides a tighter bound than previous work.
- Best heuristics are not too far from optimal.
- Better inference usually provides better solutions.
- But not always!
  - Especially for the top few solutions.
  - Useful to check statistics like number of clusters.
Conclusions

- Always use local search!
- Best greedy algorithm is voting.
- SDP provides a tighter bound than previous work.
- Best heuristics are not too far from optimal.
- Better inference usually provides better solutions.
- But not always!
  - Especially for the top few solutions.
  - Useful to check statistics like number of clusters.
- More experiments and discussion in the paper.
Acknowledgements

Software is available:
http://cs.brown.edu/~melsner

Thanks:
- Christoph Helmberg
- Claire Matheiu
- Lidan Wang and Doug Oard
- Three reviewers