Authorship Obfuscation Using Heuristic Search

Master’s Thesis Defence by Janek Bevendorff on 20 June 2018
Supervisors: Prof. Dr. Benno Stein, PD Dr. Andreas Jakoby
Contributions

• Unmasking for short texts
• Obfuscation against unmasking
• Obfuscation against compression models
• Authorship verification quality measure proposal
• Obfuscation safety analysis and definitions
• Side effect analysis
• $J_{S\Delta}$ as authorship metric
• Adaptive obfuscation
• Design of an admissible obfuscation heuristic
• Analysis of consistency and monotonicity properties
• Design and implementation of an efficient obfuscation framework
• Development of obfuscation operators
• Inspection of search space challenges and solutions
Contributions

- Unmasking for short texts
- Obfuscation against unmasking
  - Obfuscation against compression models
  - Authorship verification quality measure proposal
  - Obfuscation safety analysis and definitions
  - Side effect analysis
- $\text{JS}_\Delta$ as authorship metric
- Adaptive obfuscation
- Design of an admissible obfuscation heuristic
  - Analysis of consistency and monotonicity properties
  - Design and implementation of an efficient obfuscation framework
- Development of obfuscation operators
  - Inspection of search space challenges and solutions
Authorship
Authorship Verification
Authorship Verification
Authorship Verification
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Koppel and Schler, Authorship verification as a one-class problem, 2004
Unmasking

Same author
Unmasking

Same author

Different authors
Chunk Expansion

Treasure Island
Dr. Livesey
rest
having
Island
gentlemen
the
Treasure
I will begin the story of my certain morning adventures to the Treasure Island.

Dr. Livesey, gentlemen having rest will begin the adventures of the Treasure Island.
Chunk Expansion
Unmasking with Chunk Expansion

![Graph showing accuracy over rounds for same and different authors.](image-url)
Unmasking with Chunk Expansion

Accuracy

Rounds

Same author
Different authors
Unmasking with Chunk Expansion

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>Threshold</th>
<th>Precision</th>
<th>% Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very High</td>
<td>0.9</td>
<td>1.00</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>1.00</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>1.00</td>
<td>13.8</td>
</tr>
<tr>
<td>High</td>
<td>0.6</td>
<td>1.00</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.00</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.93</td>
<td>43.8</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.3</td>
<td>0.83</td>
<td>55.0</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.68</td>
<td>70.0</td>
</tr>
<tr>
<td>Low</td>
<td>0.1</td>
<td>0.82</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.76</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Obfuscation
Authorship Obfuscation
Authorship Obfuscation
Authorship Obfuscation

![Graph showing accuracy over rounds for different authors and same author.](image)

Accuracy vs. Rounds

- **Same author**
- **Different authors**
Authorship Obfuscation

Different authors

Same author

Different authors

Accuracy

Rounds

0 3 6 9 12 15 18 21
Authorship Obfuscation

Accuracy

Rounds

Same author
Different authors
Kullback-Leibler Divergence

$$\text{KLD}(P \| Q) = \sum_i P[i] \log_2 \frac{P[i]}{Q[i]}$$
Jensen-Shannon Divergence

\[ KLD(P\|Q) = \sum_i P[i] \log_2 \frac{P[i]}{Q[i]} \]

\[ JSD(P\|Q) = \frac{KLD(P\|M) + KLD(Q\|M)}{2} \]

\[ M = \frac{P + Q}{2} \]
Jensen-Shannon Divergence

\[
\text{KLD}(P \| Q) = \sum_i P[i] \log_2 \frac{P[i]}{Q[i]}
\]

\[
\text{JSD}(P \| Q) = \frac{\text{KLD}(P \| M) + \text{KLD}(Q \| M)}{2} \quad \rightarrow \text{maximize}
\]

\[
M = \frac{P + Q}{2}
\]
\[
\frac{\partial}{\partial Q[i]} \left( P[i] \log_2 \frac{P[i]}{Q[i]} \right) = -\frac{P[i]}{Q[i] \ln 2}
\]
Basic Obfuscation

$$\frac{\partial}{\partial Q[i]} \left( P[i] \log_2 \frac{P[i]}{Q[i]} \right) = - \frac{P[i]}{Q[i]} \ln 2$$

$$R_{KL}(i) = \frac{P[i]}{Q[i]}$$
**Basic Obfuscation**

\[
\frac{\partial}{\partial Q[i]} \left( P[i] \log_2 \frac{P[i]}{Q[i]} \right) = -\frac{P[i]}{Q[i] \ln 2}
\]

\[ R_{KL} (i) = \frac{P[i]}{Q[i]} \quad \rightarrow \text{maximize} \]
Basic Obfuscation

n-grams ranked left to right

n-gram frequencies

ny_  ly_  par  bor  y_h  hel  eme  gro  dis  gre

Text 1  Text 2 (to be obfuscated)
Basic Obfuscation

n-grams ranked left to right

n-gram frequencies

ny_ ly_ par bor y_h hel eme gro dis gre

Text 1  Text 2 (to be obfuscated)
Basic Obfuscation

n-grams ranked left to right

n-gram frequencies

ny_ ly_ par bor y_h hel eme gro dis gre

Text 1 Text 2 (to be obfuscated)
Basic Obfuscation

n-grams ranked left to right

n-gram frequencies

ny_  ly_  par  bor  y_h  hel  eme  gro  dis  gre

Text 1  Text 2 (to be obfuscated)
Basic Obfuscation

n-grams ranked left to right

n-gram frequencies

ny_, ly_, par, bor, y_h, hel, eme, gro, dis, gre

Text 1  Text 2 (to be obfuscated)
**Basic Obfuscation**

n-grams ranked left to right

- ny_
- ly_
- par
- bor
- y_h
- hel
- eme
- gro
- dis
- gre

Text 1  Text 2 (to be obfuscated)
Basic Obfuscation

n-grams ranked left to right

<table>
<thead>
<tr>
<th>n-gram frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ny_</td>
</tr>
<tr>
<td>ly_</td>
</tr>
<tr>
<td>par</td>
</tr>
<tr>
<td>bor</td>
</tr>
<tr>
<td>y_h</td>
</tr>
<tr>
<td>hel</td>
</tr>
<tr>
<td>eme</td>
</tr>
<tr>
<td>gro</td>
</tr>
<tr>
<td>dis</td>
</tr>
<tr>
<td>gre</td>
</tr>
</tbody>
</table>

Text 1
Text 2 (to be obfuscated)
Adaptive Obfuscation

\[ \text{JS distance (JS}_\Delta) = \sqrt{2 \cdot \text{JSD}(P\|Q)} \]

Different authors

Same author

Text length (characters)
Adaptive Obfuscation

\[ J S_\Delta = \sqrt{2 \cdot J S D(P \parallel Q)} \]

Text length (characters)

- Same author
- Different authors
- \( \varepsilon_0 \)
**Adaptive Obfuscation**

\[ JS_\Delta = 2 \cdot JSD(P \parallel Q) \]

- Same author
- Different authors
- $\varepsilon_0$
- $\varepsilon_{0.5}$
## Obfuscation Results

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>Threshold</th>
<th>Precision</th>
<th>% Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very High</td>
<td>0.9</td>
<td>1.00</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>1.00</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>1.00</td>
<td>13.8</td>
</tr>
<tr>
<td>High</td>
<td>0.6</td>
<td>1.00</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.00</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.93</td>
<td>43.8</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.3</td>
<td>0.83</td>
<td>55.0</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.68</td>
<td>70.0</td>
</tr>
<tr>
<td>Low</td>
<td>0.1</td>
<td>0.82</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.76</td>
<td>100.0</td>
</tr>
</tbody>
</table>

20.06.2018
## Obfuscation Results

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>Threshold</th>
<th>Precision</th>
<th>% Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very High</td>
<td>0.9</td>
<td>1.00</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>1.00</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>1.00</td>
<td>13.8</td>
</tr>
<tr>
<td>High</td>
<td>0.6</td>
<td>1.00</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.00</td>
<td>30.0</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.4</td>
<td>0.93</td>
<td>43.8</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.83</td>
<td>55.0</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.68</td>
<td>70.0</td>
</tr>
<tr>
<td>Low</td>
<td>0.1</td>
<td>0.82</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.76</td>
<td>100.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>Threshold</th>
<th>Precision</th>
<th>% Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very High</td>
<td>0.9</td>
<td>0.00</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.00</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.00</td>
<td>8.7</td>
</tr>
<tr>
<td>High</td>
<td>0.6</td>
<td>0.00</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.00</td>
<td>27.5</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.4</td>
<td>0.00</td>
<td>42.5</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.67</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.50</td>
<td>70.0</td>
</tr>
<tr>
<td>Low</td>
<td>0.1</td>
<td>0.42</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.53</td>
<td>100.0</td>
</tr>
</tbody>
</table>

20.06.2018
Heuristic Search
Best-first Search
Best-first Search
Best-first Search
Best-first Search

\[ f(n) \]
Best-first Search
Best-first Search

The figure illustrates a best-first search algorithm. The root node is labeled with $f(n)$, and the search expands the node with the best function value first. The tree structure shows the exploration path, with each node representing a state and the edges indicating the transitions between states.
Best-first Search
Best-first Search

$f(n)$
The A* Algorithm

\[ f(n) = g(n) + h(n) \]
The A* Algorithm

\[ f(n) = g(n) + h(n) \]

\[ h(n) \leq h^*(n) \]
The Heuristic

\[ h_{prior}(n) = \varepsilon - J S_{\Delta n} \]
The Heuristic

\[ h_{prior}(n) = \varepsilon - JS_{\Delta n} \]

\[ g_{norm}(n) = \frac{g(n)}{JS_{\Delta n} - JS_{\Delta 0}} \]
The Heuristic

\[ h(n) = h_{prior}(n) \cdot g_{norm}(n) \]
The Heuristic

Linear Gain

$\mathcal{E}$
The Heuristic

\[ g(n) \]

Linear Gain

\[ \varepsilon \]
The Heuristic

\[ g(n) \]

\[ \varepsilon \]

\[ JS_\Delta \]
The Heuristic

\[ h(n) \]

\[ g(n) \]

\[ \epsilon \]

\[ JS_\Delta \]

Linear Gain
The Heuristic

\[ h(n) \]

\[ \varepsilon \]

\[ JS_{\Delta} \]

\[ g(n) \]

Linear Gain
The Heuristic

$h(n)$

$\varepsilon$

$JS_\Delta$

$g(n)$

Sublinear Gain
The Heuristic

Sublinear Gain

$h(n)$

$g(n)$

$\varepsilon$

$JS_\Delta$
The Heuristic

Sublinear Gain

\[ h(n) \]

\[ g(n) \]

\[ JS_\Delta \]

\[ \varepsilon \]
The Heuristic – Actual Example

$h(n)$

$\Delta$

$g(n)$

$\varepsilon$

Operations
The Heuristic – Actual Example

$h(n)$

$g(n)$

$\varepsilon$

$JS_\Delta$

Stepwise $JS_\Delta$

Operations
The Heuristic – Actual Example

$h(n)$

$g(n)$

$\varepsilon$

$\Delta JS$

Stepwise $JS_\Delta$

Operations

0 100 200 300 400
Operators

$n$-gram removal

abcdefg
Operators

n-gram removal

abfg
Operators

$n$-gram removal

character flip

abfg

wizard
Operators

\( n \)-gram removal

\( \text{abfg} \)

character flip

\( \text{wiazrd} \)
Operators

n-gram removal  character flip  character map

abfg  wiazrd  The End.
Operators

- n-gram removal
- character flip
- character map

abfg  wiazrd  The End!
Operators

- n-gram removal
- character flip
- character map

abfg  wiazrd  The End!

house

synonym
Operators

n-gram removal  character flip  character map

abfg  wiazrd  The End!

home  synonym
Operators
n-gram removal
character flip
character map

abfg
wiazrd
The End!

home
author

synonym
Netspeak

20.06.2018
Obfuscation Using Heuristic Search

$h(n)$

$g(n)$

$\varepsilon$

$JS_\Delta$

Stepwise $JS_\Delta$
Obfuscation Using Heuristic Search

With a furtive glance around him, he clapped the other half of the clay sphere over the filled hemisphere and then stood up. The patients lined up at the door, waiting for the walk back across the green hills to the main hospital. The attendants made a quick count and then unlocked the door. The group shuffled out into the warm, afternoon sunlight and the door closed behind them. Miss Abercrombie gazed around the cluttered room and picked up her chart book of patient progress. Moving slowly down the line of benches, she made short, precise notes on the day’s work accomplished by each patient. [...]

A Filbert Is a Nut by Rick Raphael
With a furtive glance around him, he clapped the other half of the clay sphere over the filled hemisphere and then stood up. The patients lined up at the door, waiting for the walk back across the site hills to the main hospital. The attendants made a quick investigation and then unlocked the door. The group shuffled out into the warm, daylight sunlight and the door closed behind them. Miss Abercrombie gazed around the cluttered room and picked up her chart forward of patient progress. Moving slowly down the line of bens, she made parcel, precise notes on the day’s work accomplished by each patient. [...]
Obfuscation Using Heuristic Search

With a furtive glance around him, he clapped the other half of the clay sphere over the filled hemisphere and then stood up. The patients lined up at the door, waiting for the walk back across the site hills to the main hospital. The attendants made a quick investigation and then unlocked the door. The group shuffled out into the warm, daylight sunlight and the door closed behind them. Miss Abercrombie gazed around the cluttered room and picked up her chart forward of patient progress. Moving slowly down the line of bens, she made parcel, precise notes on the day’s work accomplished by aehc patient. [...]

A Filbert Is a Nut by Rick Raphael
With a furtive glance around him, he clapped the other half of the clay sphere over the filled hemisphere and then stood up. The patients lined up at the door, waiting for the walk back across the site hills to the main hospital. The attendants made a quick investigation and then unlocked the door. The group shuffled out into the warm, daylight sunlight and the door closed behind them. Miss Abercrombie gazed around the cluttered room and picked up her chart forward of patient progress. Moving slowly down the line of bens, she made parcel, precise notes on the day’s work accomplished by aehc patient. [...]
Summary

• Unmasking can be attacked by KLD obfuscation,
Summary

• Unmasking can be attacked by KLD obfuscation,
• $JS_\Delta$ is an effective authorship metric,
Summary

• Unmasking can be attacked by KLD obfuscation,

• \( JS_\Delta \) is an effective authorship metric,
  • important building block: length-invariant thresholds.
Summary

• Unmasking can be attacked by KLD obfuscation,
• $JS_\Delta$ is an effective authorship metric,
  • important building block: length-invariant thresholds.
• Design of an admissible heuristic search function:
Summary

• Unmasking can be attacked by KLD obfuscation,
• $JS_\Delta$ is an effective authorship metric,
  • important building block: length-invariant thresholds.
• Design of an admissible heuristic search function:
  • significant reduction of text modifications at the same effect,
Summary

• Unmasking can be attacked by KLD obfuscation,
• JS$\Delta$ is an effective authorship metric,
  • important building block: length-invariant thresholds.
• Design of an admissible heuristic search function:
  • significant reduction of text modifications at the same effect,
  • better text quality.
Thank you for your attention