Tagvisor: A Privacy Advisor for Sharing Hashtags

Yang Zhang
Joint work with Mathias Humbert, Tahleen Rahman, Cheng-Te Li, Jun Pang and Michael Backes
#hashtag
#hashtag

Chris Messina™
@chrismessina

how do you feel about using # (pound) for groups. As in #barcamp [msg]?

12:25 PM - 23 Aug 2007

146 RETWEETS 288 FAVORITES
#hashtag

#ShareaCoke
#hashtag

#like4like
#foodporn
#tbt
#hashtag

#privacy

#locationprivacy
#contributions

- Attack: location inference with hashtags
- Defense: Tagvisor, a privacy advisor to mitigate the privacy threat by hashtags
#dataset

- Collected through Instagram’s APIs
- New York, Los Angeles, and London
- Hashtags + locations (check-ins)

<table>
<thead>
<tr>
<th></th>
<th>New York</th>
<th>Los Angeles</th>
<th>London</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of posts</td>
<td>144,263</td>
<td>61,767</td>
<td>34,018</td>
</tr>
<tr>
<td>No. of hashtags</td>
<td>8,552</td>
<td>4,600</td>
<td>2,395</td>
</tr>
<tr>
<td>No. of users</td>
<td>3,911</td>
<td>1,625</td>
<td>992</td>
</tr>
<tr>
<td>No. of locations</td>
<td>498</td>
<td>268</td>
<td>141</td>
</tr>
</tbody>
</table>
Bag-of-words for feature representation

Random forest classifier

Multiple-class classification, e.g., 498 classes (locations) in New York

All posts are trained together
#attack

<table>
<thead>
<tr>
<th></th>
<th>New York</th>
<th>Los Angeles</th>
<th>London</th>
<th>All cities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>attack</td>
<td>baseline</td>
<td>attack</td>
<td>baseline</td>
</tr>
<tr>
<td>Correctness</td>
<td>0.613</td>
<td>0.015</td>
<td>0.685</td>
<td>0.015</td>
</tr>
<tr>
<td>Distance (km)</td>
<td>0.917</td>
<td>4.198</td>
<td>1.870</td>
<td>11.275</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.697</td>
<td>0.053</td>
<td>0.758</td>
<td>0.048</td>
</tr>
</tbody>
</table>
#attack
#tagvisor

- A privacy advisor for sharing hashtags
- Fool the attacker’s location inferencer (ML classifier)
- Three defense mechanisms
  - Hiding
  - Replacement
  - Generalization (location category)
- Utility: preserving the semantical meaning of hashtags
successful attack \#a\#b\#c

delete one hashtag (can be more)

hide \#a \#b\#c

hide \#b \#a\#c

hide \#c \#a\#b
#utility

- Semantical meaning
- Skip-gram, aka word2vec
- Skip-gram over all posts’ hashtags

Hashtag vectors

- `#a`: [3.1, 1.3]
- `#b`: [2.5, 1.9]
- `#c`: [4.0, 5.1]
#replacement

- Replace each hashtag with all the possible hashtag
- Search space is too big
- Bound to the most close hashtags (with word2vec)
- Reduce the search space
- Semantical meaning can be preserved

successful attack #a#b#c
#generalization

- Location category from foursquare
  - #centralpark -> #park
- Do not apply to all hashtags
  - e.g., #tbt #love
Check whether the post’s location is inferred correctly

If no, then publish

Else, consider the three defense mechanisms

Pick the hashtag set with the highest utility
Obfuscating bounded number of hashtags

Obfuscating 2 hashtags is enough!
#conclusion

- First location inference attack with hashtags
- Sharing hashtags is not safe!!!
- A privacy advisor to mitigate this risk
- Minimal risk and maximal utility
- Fit for the real-world setting

#thankyou

https://yangzhangalmo.github.io/
@yangzhangalmo