# Tagvisor: A Privacy Advisor for Sharing Hashtags

**Yang Zhang**

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

CISPA HELMHOLTZ-ZENTRUM i.G.

SDSC

NATIONAL CHENG KUNG UNIVERSITY 1931

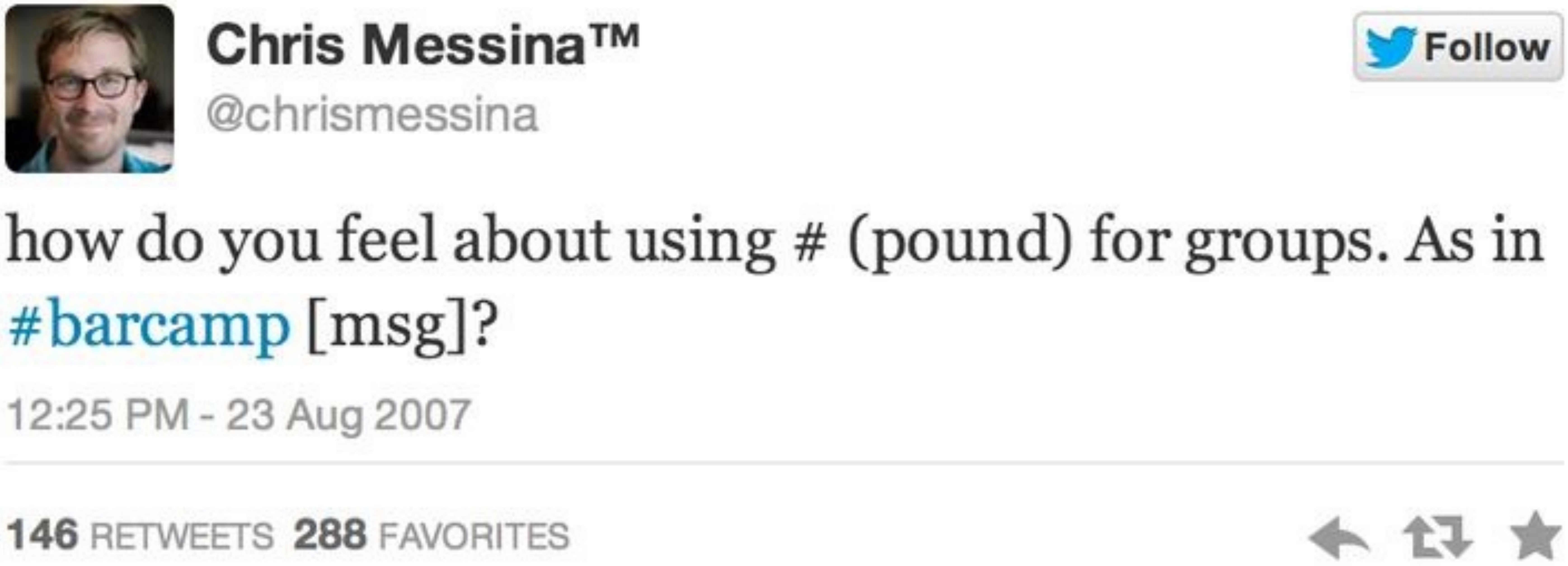
uni.lu UNIVERSITÉ DU LUXEMBOURG

# #hashtag

LinkedIn
skype
LIVEJOURNAL
foursquare
vimeo
YouTube
Insta
facebook
myBlogger
instagram
twitter
flickr
Behance
tumblr.
yelp
Like

Follow
SHARE
#

# #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

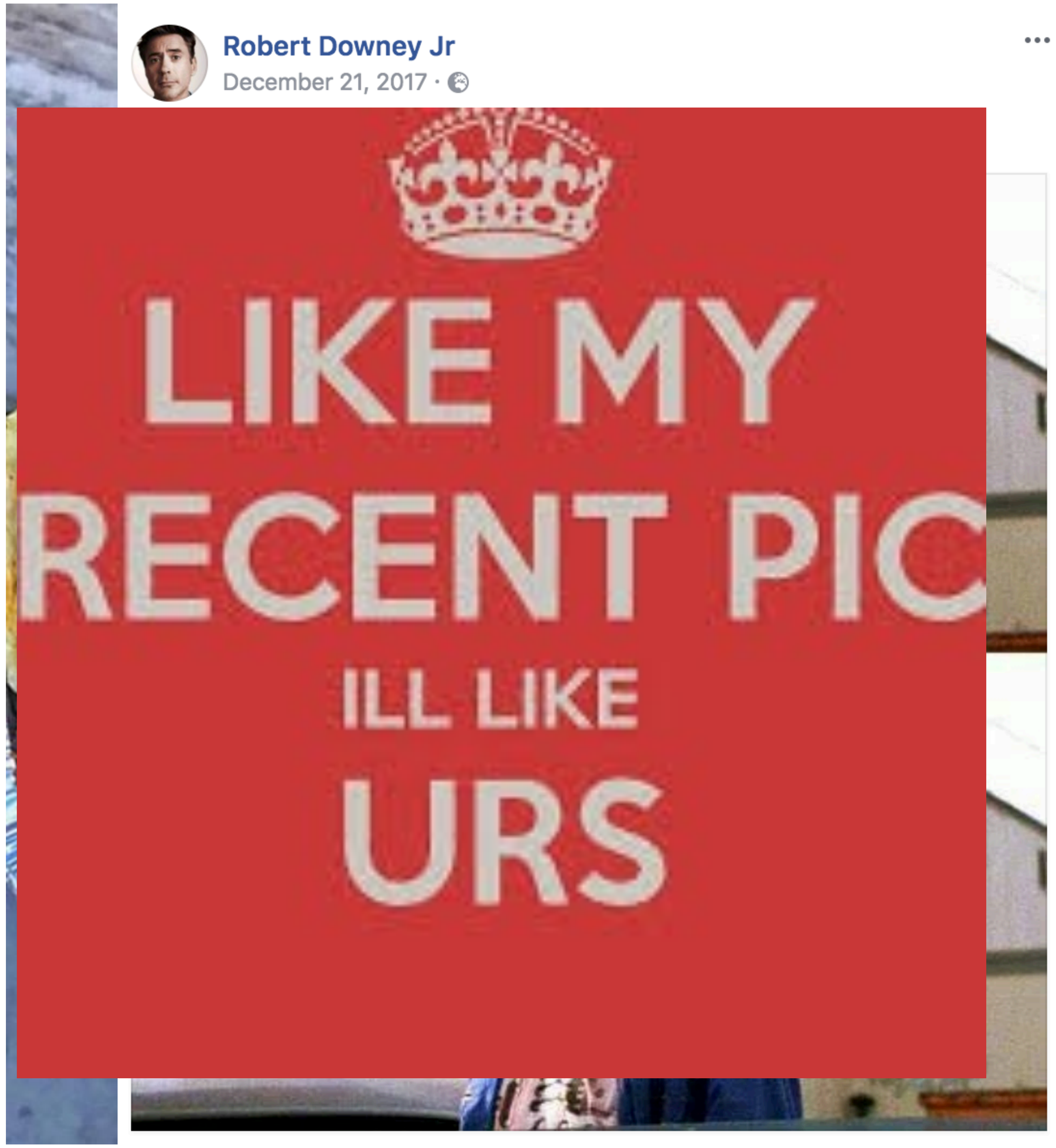
# #hashtag

# #hashtag

#like4like

#foodporn

#tbt



Robert Downey Jr
December 21, 2017 ·

# #hashtag

#privacy

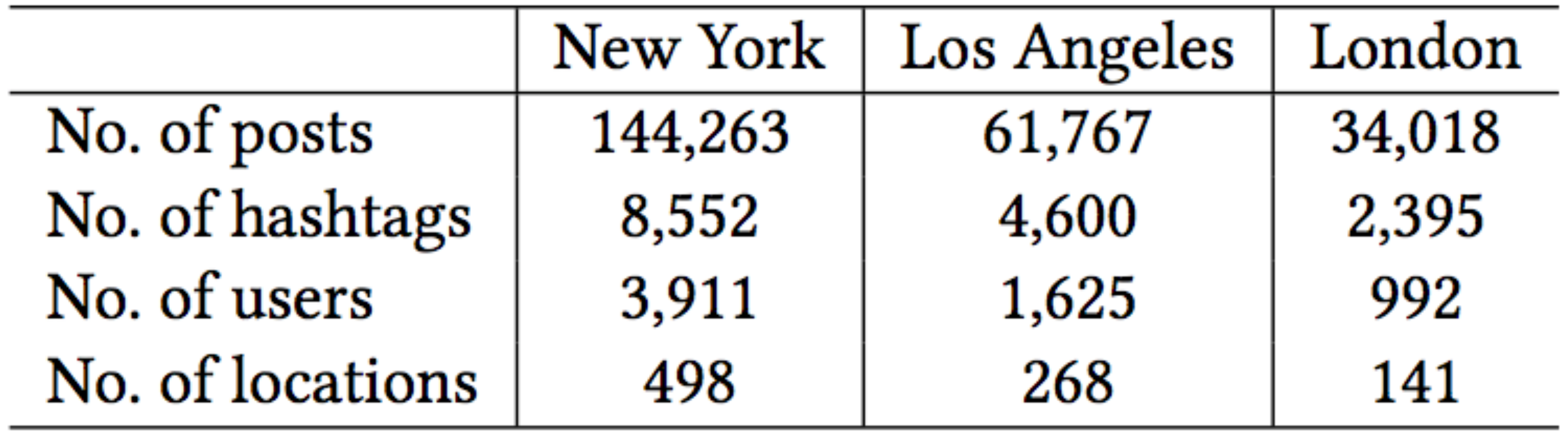
#locationprivacy

# #contributions

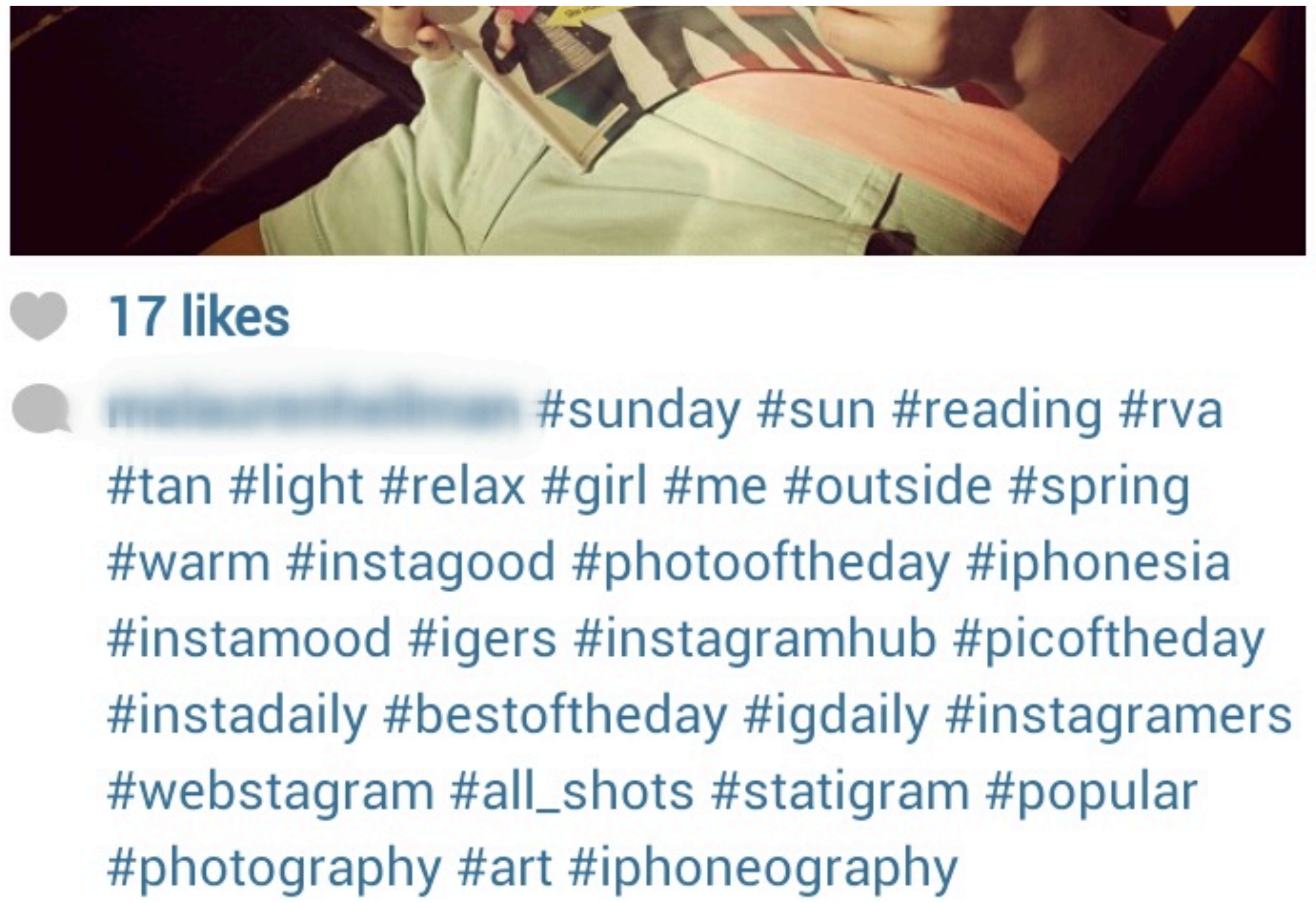
- Attack: location inference with hashtags
- Defense: Tagvisor, a privacy advisor to mitigate the privacy threat by hashtags

# #dataset

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



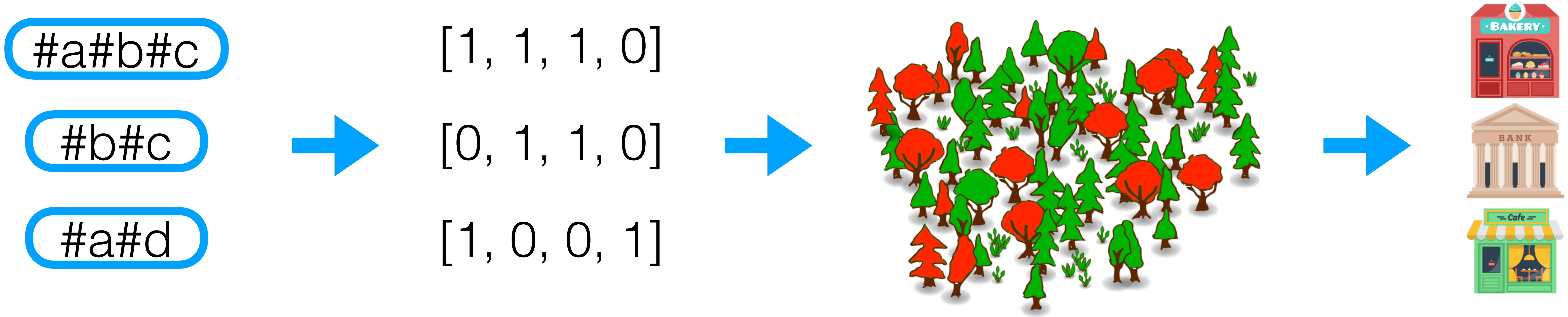
| | New York | Los Angeles | London |
|---|---|---|---|
| No. of posts | 144,263 | 61,767 | 34,018 |
| No. of hashtags | 8,552 | 4,600 | 2,395 |
| No. of users | 3,911 | 1,625 | 992 |
| No. of locations | 498 | 268 | 141 |



17 likes

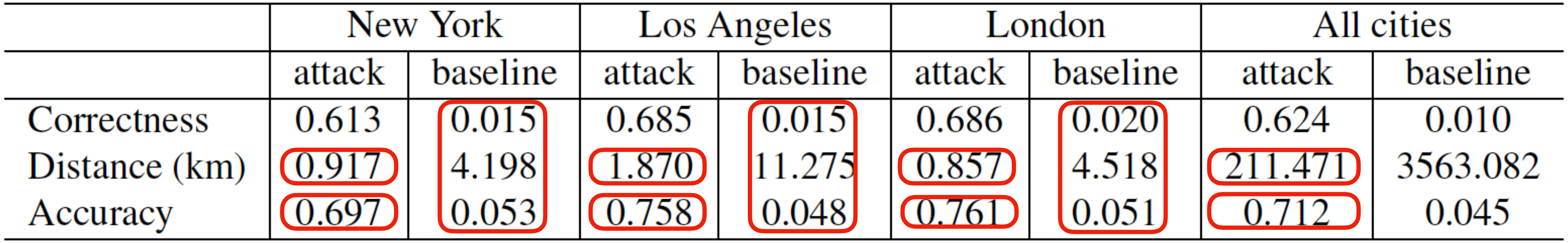
[redacted] #sunday #sun #reading #rva #tan #light #relax #girl #me #outside #spring #warm #instagood #photooftheday #iphonesia #instamood #igers #instagramhub #picoftheday #instadaily #bestoftheday #igdaily #instagramers #webstagram #all_shots #statigram #popular #photography #art #iphoneography

# #attack



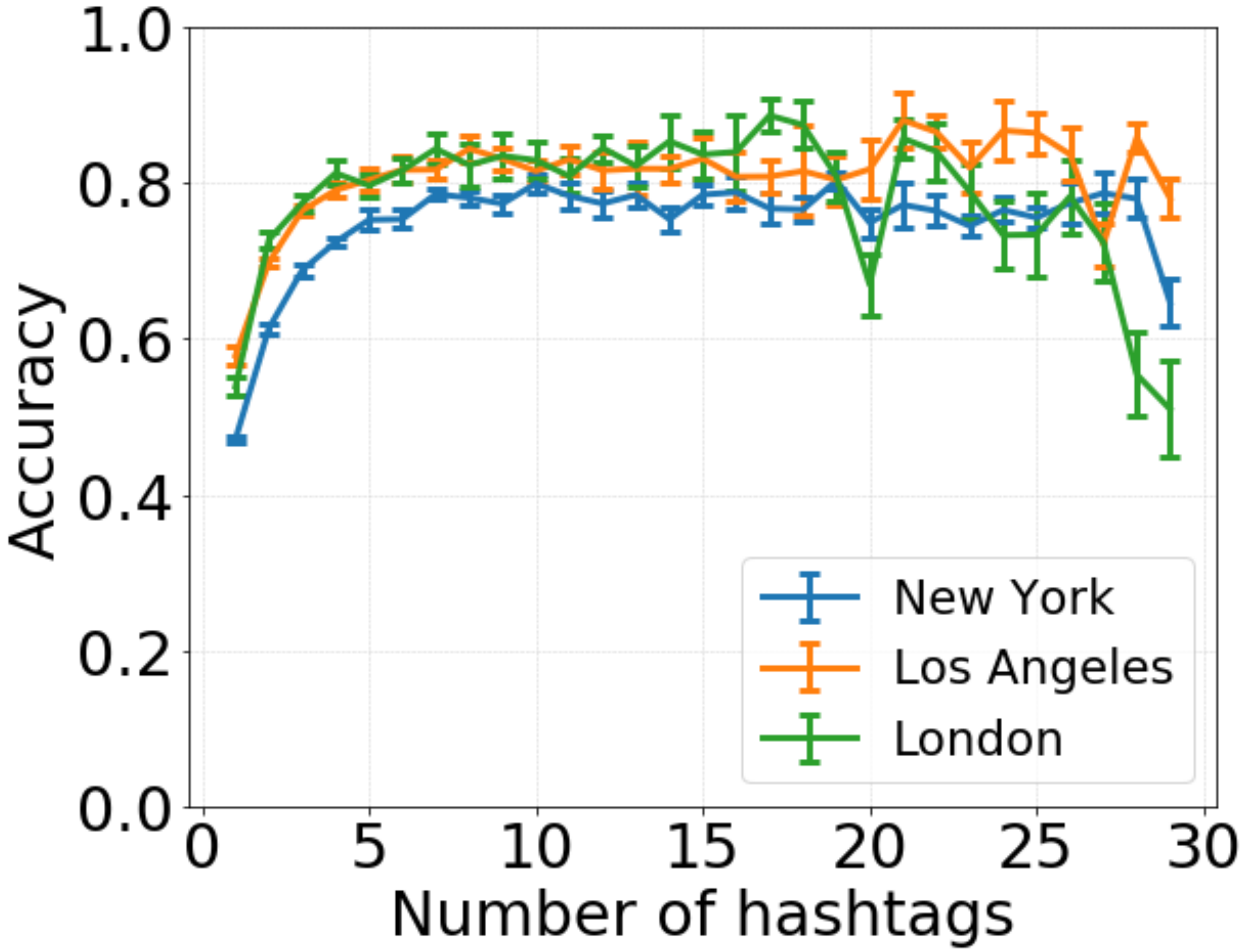
- Bag-of-words for feature representation
- Random forest classifier
- Multiple-class classification, e.g., 498 classes (locations) in New York
- All posts are trained together

# #attack



| | New York | | Los Angeles | | London | | All cities | |
|---|---|---|---|---|---|---|---|---|
| | attack | baseline | attack | baseline | attack | baseline | attack | baseline |
| Correctness | 0.613 | 0.015 | 0.685 | 0.015 | 0.686 | 0.020 | 0.624 | 0.010 |
| Distance (km) | 0.917 | 4.198 | 1.870 | 11.275 | 0.857 | 4.518 | 211.471 | 3563.082 |
| Accuracy | 0.697 | 0.053 | 0.758 | 0.048 | 0.761 | 0.051 | 0.712 | 0.045 |

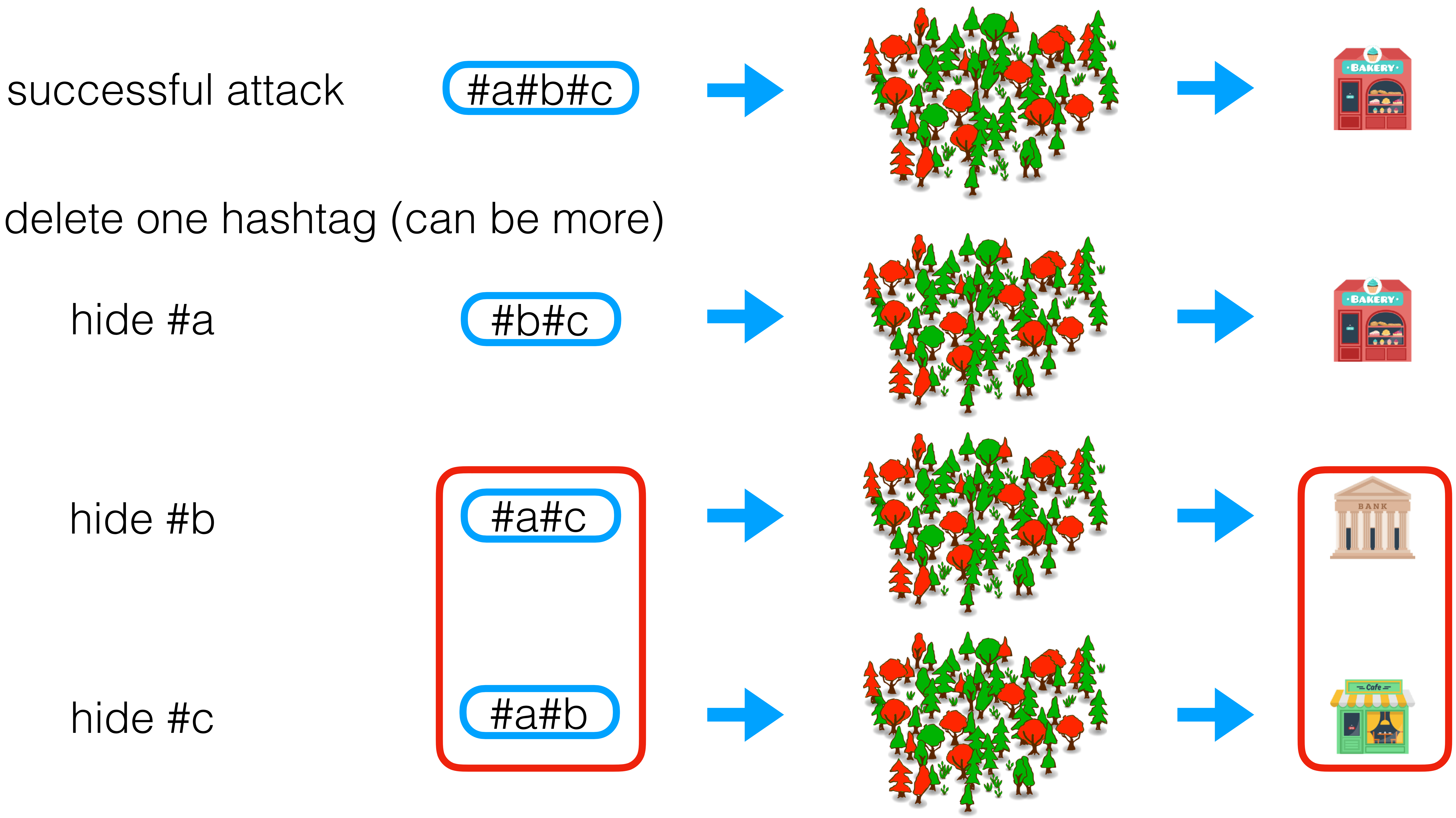
# #attack


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0
5
10
15
20
25
30
Number of hashtags
New York
Los Angeles
London

# #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

# #hiding



successful attack #a#b#c

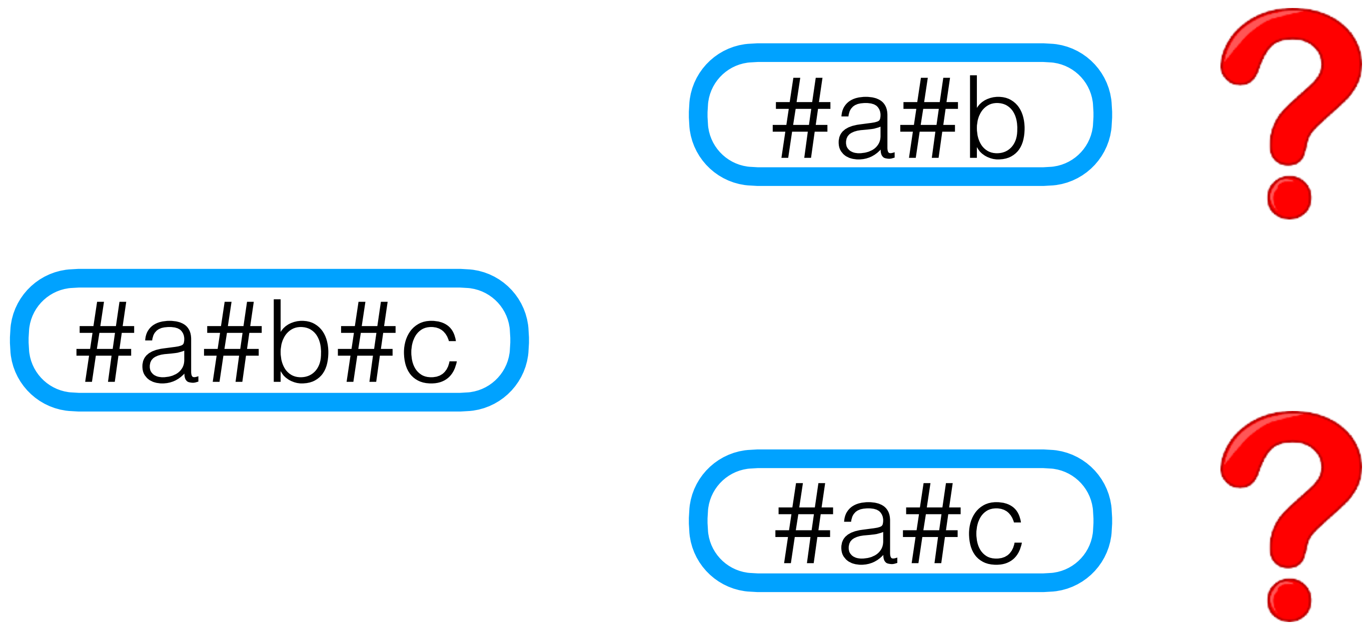
delete one hashtag (can be more)

hide #a #b#c

hide #b #a#c

hide #c #a#b

# #utility



#a#b ?

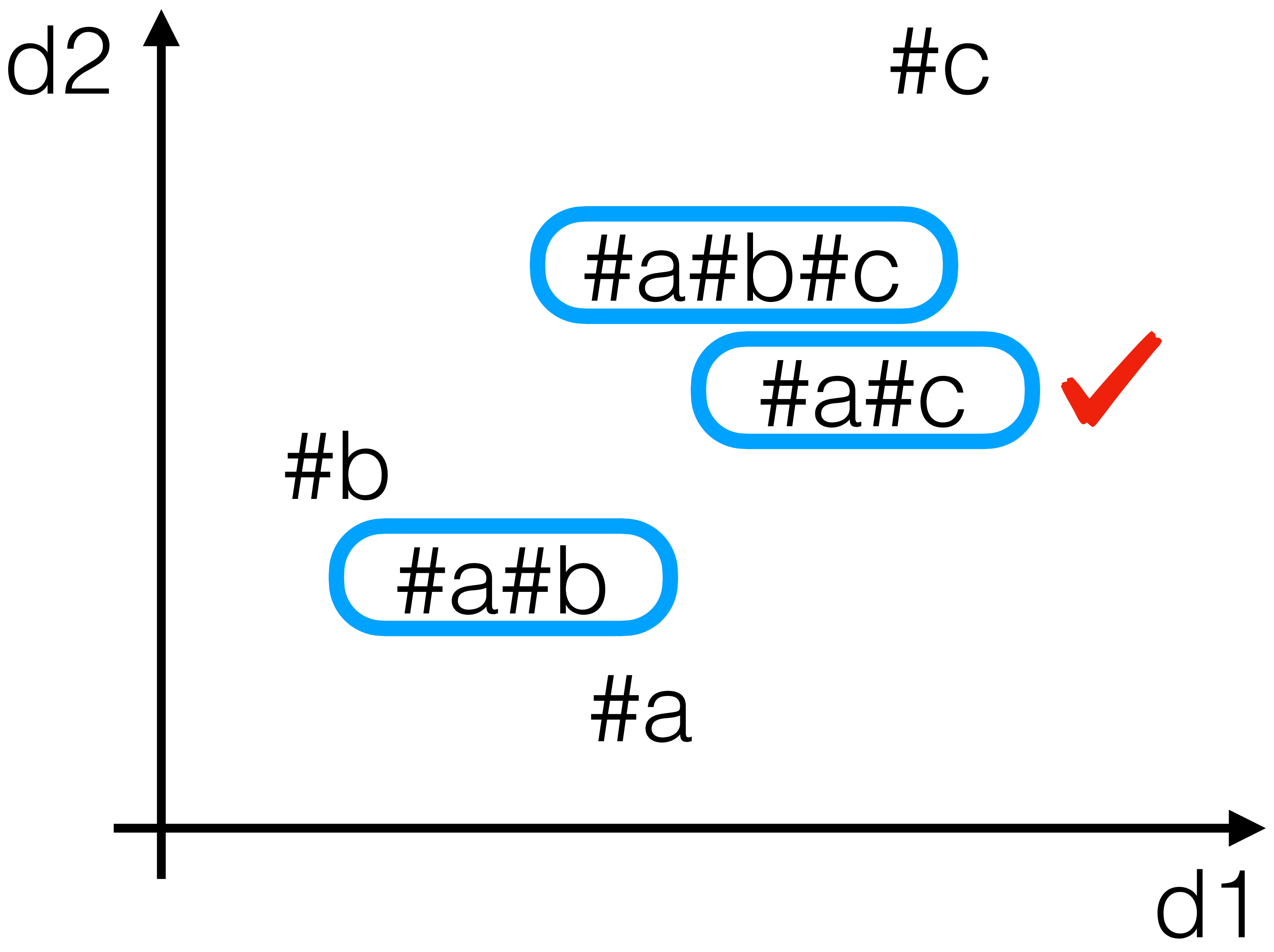
#a#b#c

#a#c ?

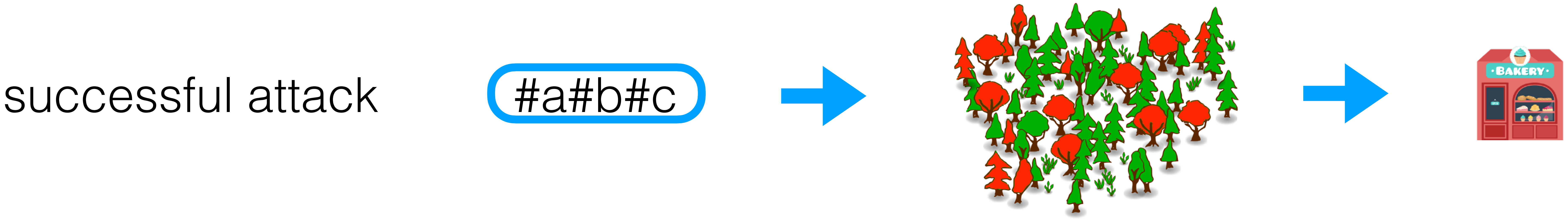
- Semantical meaning
- Skip-gram, aka word2vec
- Skip-gram over all posts’ hashtags

Hashtag vectors

|  | d1 | d2 |
|---|---|---|
| #a: | [3.1, | 1.3] |
| #b: | [2.5, | 1.9] |
| #c: | [4.0, | 5.1] |



# #replacement



successful attack

#a#b#c

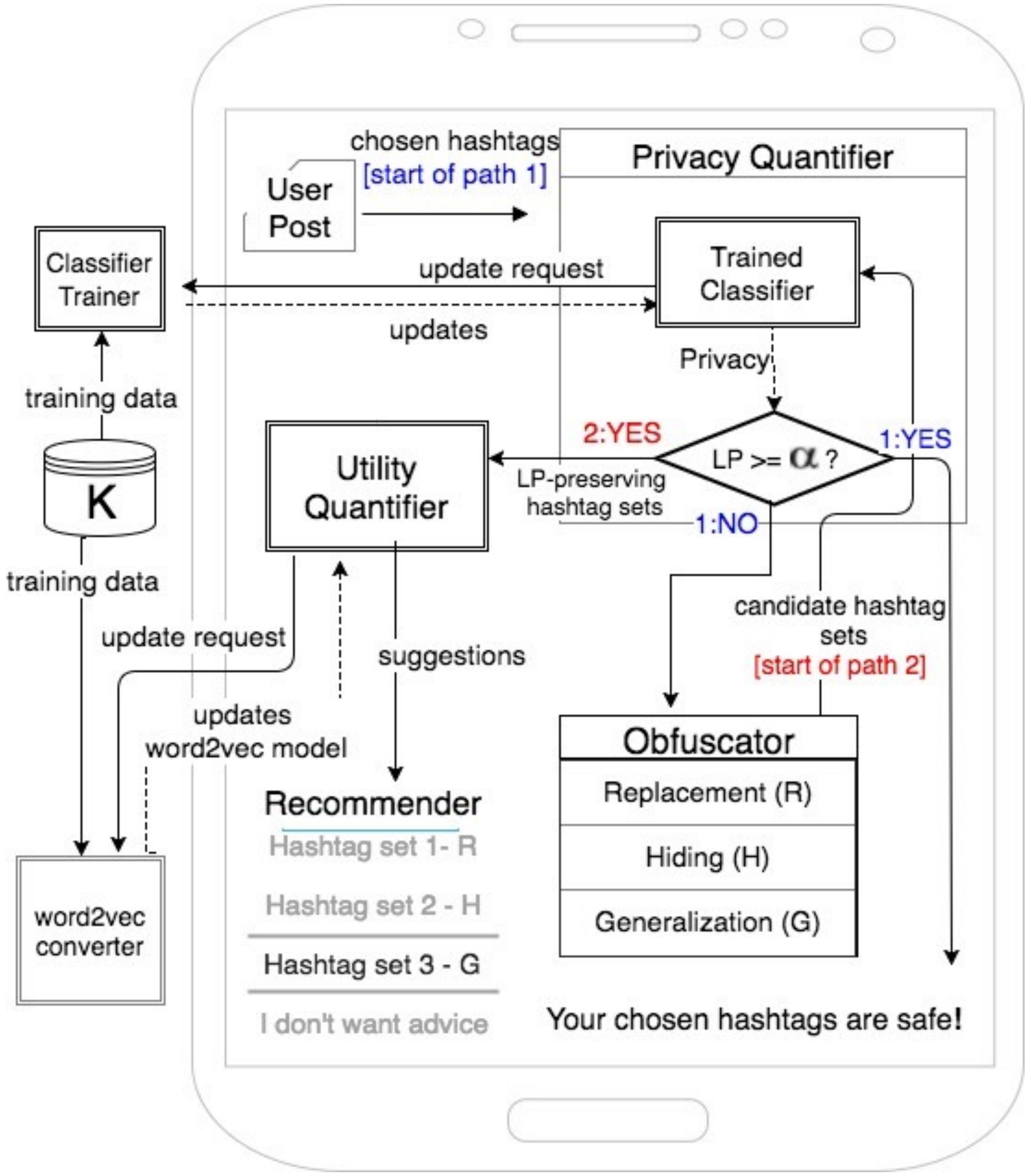
- Replace each hashtag with all the possible hashtag
  - Search space is too big
- Bound to the most closest hashtags (with word2vec)
  - Reduce the search space
  - Semantical meaning can be preserved

# #generalization

- Location category from foursquare
  - #centralpark -> #park
- Do not apply to all hashtags
  - e.g., #tbt #love

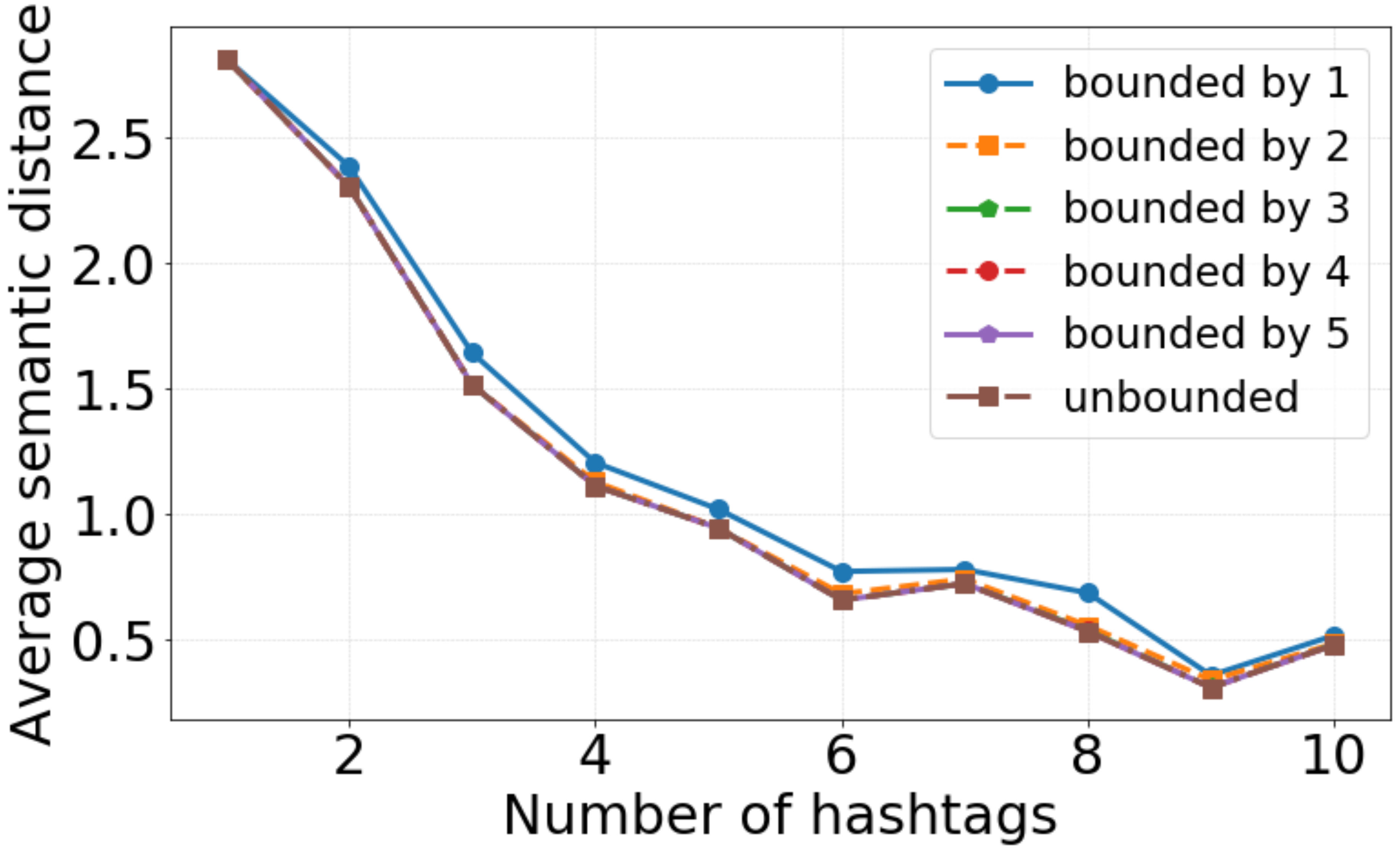
# #tagvisor

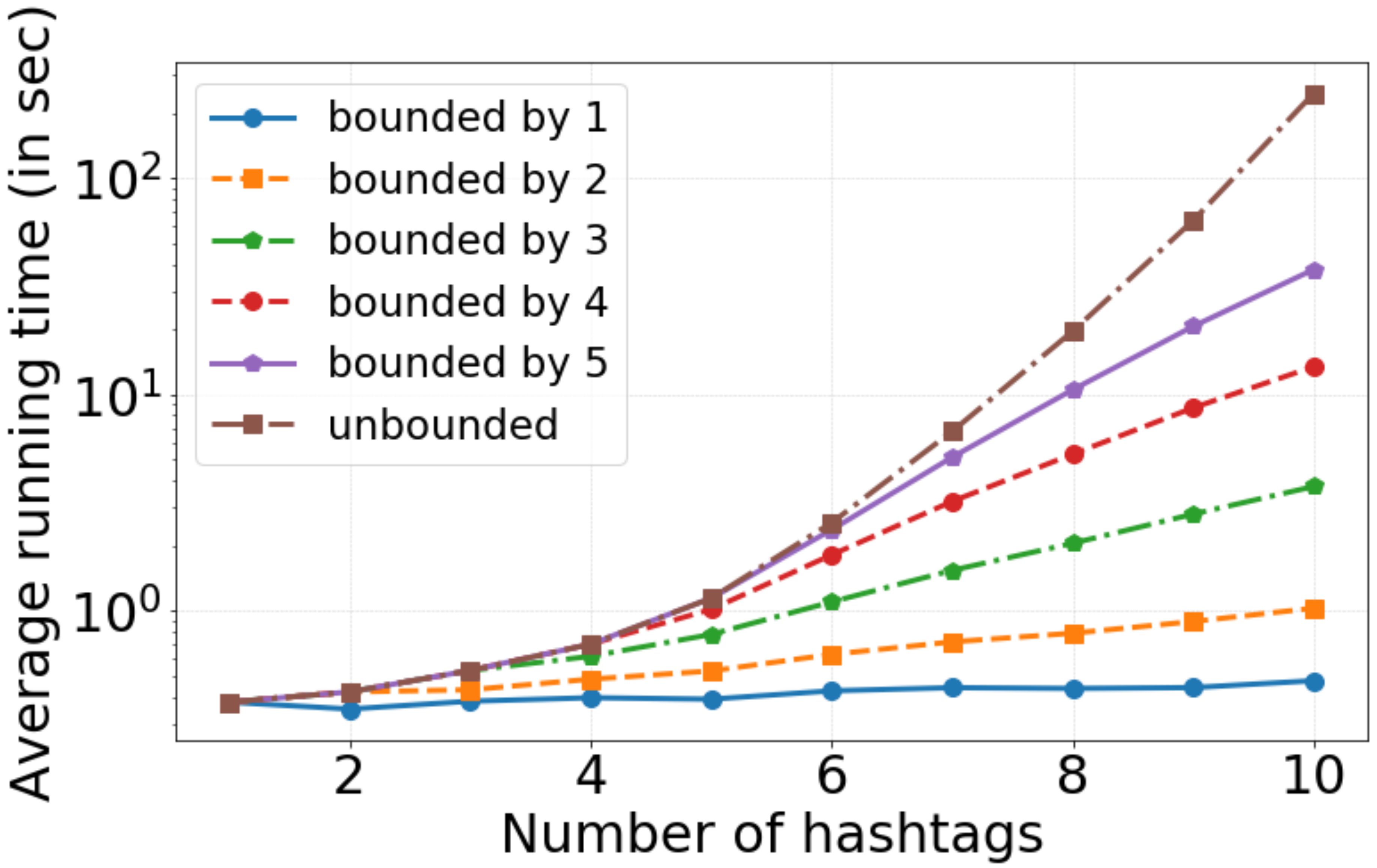
- Check whether the post's location is inferred correctly
  - If no, then publish
  - Else, consider the three defense mechanisms
    - Pick the hashtag set with the highest utility



# #tagvisor

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