Yang Zhang, Mathias Humbert, Tahleen Rahman, Cheng-Te Li, Jun Pang, and Michael Backes. 2018. Tagvisor: A Privacy Advisor for Sharing Hashtags. In WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France. ACM, New York, NY, USA, 10 pages.

Abstract

Hashtag has emerged as a widely used concept of popular culture and campaigns, but its implications on people's privacy have not been investigated so far. In this paper, we present the first systematic analysis of privacy issues induced by hashtags. We concentrate in particular on location, which is recognized as one of the key privacy concerns in the Internet era. By relying on a random forest model, we show that we can infer a user's precise location from hashtags with accuracy of 70% to 76%, depending on the city.
To remedy this situation, we introduce a system called Tagvisor that systematically suggests alternative hashtags if the user-selected ones constitute a threat to location privacy. Tagvisor realizes this by means of three conceptually different obfuscation techniques and a semantics-based metric for measuring the consequent utility loss. Our findings show that obfuscating as little as two hashtags already provides a near-optimal trade-off between privacy and utility in our dataset. This in particular renders Tagvisor highly time-efficient, and thus, practical in real-world settings.

doi>10.1145/3178876.3186095

Tagvisor: A Privacy Advisor for Sharing Hashtags

Yang Zhang, Mathias Humbert, **Tahleen Rahman**, Cheng-Te Li, Jun Pang, and Michael Backes.

Motivation

Privacy threats arising out of hashtags

- Hashtag has become very popular in social media culture and campaigns
- But its implications on people's privacy had not been investigated so far

1. We conduct the first study on addressing privacy raised by hashtags.
- We concentrate on **location privacy**, which is recognized as one of the key privacy concerns in the modern society
- Accuracy of 70% in New York (for a number of around 500 considered locations) and 76% in Los Angeles and London (for around 270 and 140 locations, respectively).

2. We develop Tagvisor: a system that recommends hashtags to a user who wants to protect his location.
- Tagvisor suggests an optimal subset of obfuscated hashtags that guarantees some predefined level of location privacy and retains as much utility as possible.

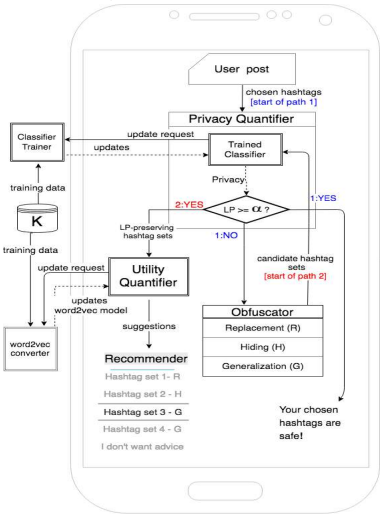
Privacy Metrics

Accuracy, Expected Distance and Correctness ($\Pr(\mathcal{L}_p = \ell | H_p, K)$)

Tagvisor

- Tagvisor implements 3 different obfuscation mechanisms:
  - **hiding** (a subset of) hashtags,
  - **replacing** hashtags by semantically similar hashtags,
  - **generalizing** hashtags with higher-level semantic categories (e.g., Starbucks into coffee shop)

- **Utility** = semantic distance between the original set of hashtags and the sanitized set.
- Semantic meaning of hashtags H_p in a post p = average of their semantic vectors (word2vec embeddings)



Attack

Dataset

- Instagram posts between July – December 2015 (collected via locations from Foursquare)

Table 1: Pre-processed data statistics.

	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

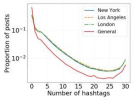


Figure 1: Distribution of the proportion of posts with a certain number of hashtag(s), from 0 to 30 hashtags.

- The general curve (in red in Figure above) represents the distribution of number of hashtags for posts without location check-in.

→ Users who do not disclose their location may still reveal hashtags.

Adversarial models A1 Vs A2

- A1 has access to all the publicly shared posts also including some shared by the targeted users.
- A2 does not have access to any previous location-hashtag data shared by the targeted users

ML based Attack

Each post p with its set of hashtags H_p and "real location" l_p^r, is represented by a feature $\vec{x}_p = (x_p^1, \ldots, x_p^n)$ where

n = |H|, $x_p^i = \begin{cases} 1, & \text{hashtag } h_i \text{ is published with } p \\ 0, & \text{otherwise} \end{cases}$, and by the label or class y_p, where $y_p = l_p^r$

Experimental Evaluation

- Accuracy of more than 70% on a global level.

Table 2: Performance of location inference across all the cities for different adversary models and baseline.

	New York			Los Angeles			London			All cities		
	A_1	A_2	baseline	A_1	A_2	baseline	A_1	A_2	baseline	A_1	A_2	baseline
Correctness	0.613	0.468	0.015	0.685	0.502	0.015	0.686	0.552	0.020	0.624	0.465	0.010
Expected distance (km)	0.917	1.272	4.198	1.870	3.046	11.275	0.857	1.575	4.518	211.471	345.980	3563.082
Accuracy	0.697	0.556	0.053	0.758	0.597	0.048	0.761	0.617	0.051	0.712	0.560	0.045

- Baseline model: relies only on the locations' distribution (in the training set) to predict a targeted user's location: Both A1 and A2 achieve at least a 10-times higher accuracy, and 27-times higher correctness than the baseline
- Small expected distance in all cities: Even when the prediction is wrong, the attacker is still able to narrow down the target's location into a small area
- A2 still achieves a relatively high prediction success: Learning per-user associations between hashtags and locations is helpful but not absolutely necessary.
- LA has highest expected distance: LA covers a larger area with places being more uniformly distributed in the geographical space than New York and London.

Privacy vs. Number of Hashtags

- Increases from 2 onwards, strongest when the targeted posts contain around 7 hashtags.
- Beyond 7, performance of A2 decreases with increasing number of hashtags:

A user who has never shared any locations (the assumption of A2) is less vulnerable to the attack.

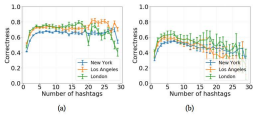


Figure 2: Correctness of adversary (a) A_1 and (b) A_2 with respect to the number of hashtags shared in posts.

Tagvisor- Empirical Evaluation

- Even with 10 original hashtags, the average accuracy with only 2 obfuscated hashtags is already very low (<0.2)
- Replacement provides close to optimal utility and is selected for the optimal solution in 85% of the cases, against 14% for hiding, and 1% for generalization.
- 90% of the privacy-preserving hashtag sets have a semantic distance < 3 to the original hashtag sets for replacement whereas, for deletion and generalization, the distance is ~ 6, and the distance between random pairs goes up to 11.
- The higher the number of original hashtags, the better the utility for similar levels of privacy.
- Bounding the number of possible hashtags to be obfuscated to 2 provides the best utility-privacy-efficiency trade-off.

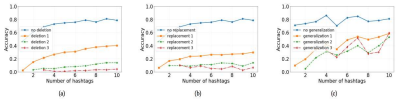


Figure 4: Evolution of the accuracy (A_1) with respect to different original numbers of hashtags to be shared (x-axis) and numbers of hashtags to be obfuscated (from 0 to 3) for (a) hiding, (b) replacement, and (c) generalization in New York.

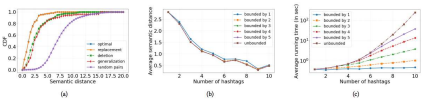


Figure 5: (a) Cumulative distribution function (CDF) of the minimum utility loss, i.e., semantic distance, for maximum privacy constraint, of the three obfuscation mechanisms and the optimal one among all mechanisms. (b) Average minimum semantic distance of the original hashtags from the optimal solution (unbounded) and solutions with an upperbound on the number of hashtags to be obfuscated with respect to the number of original hashtags (x-axis) (c) Average running time of Tagvisor (per sample) with respect to the number of original hashtags.

This demonstrates the practical feasibility of our privacy-preserving system given the computational capabilities of current mobile devices

This poster is based on the following publication:
Zhang, Yang and Humbert, Mathias and Rahman, Tahleen and Li, Cheng-Te and Pang, Jun and Backes, Michael (2018) Tagvisor: A Privacy Advisor for Sharing Hashtags.

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY