Privacy on the Web

“Saying that you don’t care of privacy because you have nothing to hide is like saying you don’t care of freedom of speech because you don’t have anything to say” E. Snowden

Assessing Privacy Risks

- Assume there is a drug company that wishes to advertise new anxiety-reducing drug to Facebook user. To this end, it needs to know the ones afflicted by depression. *How can it identify these people?*
- An adversarial HR department wants to find out the employees that have been drinking. *How can it identify these people?*
- An insurance company wants to find out people with certain diseases. *How can it identify these people?*

Assessing Privacy Risks

- An online post might directly or indirectly disclose personal information such as gender, age, political affiliation, etc.
- An adversary can combine such observations with his background knowledge and discriminate the user from the others
- **Assumption:** an adversary will rank the users and target the ones he/she believes are the most sensible ones
- [1] proposes a method to rank users according to their risk of being spotted as ”sensible”. **Goal:** alerting them that their online behavior can reveal sensible information

[1] J. A. Biega, K. P. Gummadi, I. Mele, D. Milchevski, C. Tryfonopoulos, and G. Weikum, “R-Susceptibility: An IR-Centric Approach to Assessing Privacy Risks for Users in Online Communities,” SIGR 2016, pp. 365–374.

Assessing Privacy Risks

- First, we need to identify a set of *sensitive states:*
  - E.g. depression, pregnancy, or financial debts
- Latent topic models (e.g. LDA): statistical techniques to associate topics to groups of words. They construct these models by analyzing large text corpora
  - E.g. the topic ”Financial debts” can be captured by phrases like *loans, mortgages, money, sleepless nights*
- **Main idea:** In order to identify whether a user is in a sensitive state, we can look at her/his online and see whether she/he tends to use more words than “normal” that describe a sensitive topic

Assessing Privacy Risks

First possibility (ENTROPY)

- Pick a sensitive topic X and its corresponding sensitive phrases $x_1, x_2, \dots, x_j$
- We can calculate the global probability distribution of the usage of such words in a certain community
- Then, we can calculate the same local probability by only looking at the content of a certain user
- The KL divergence measure (also known as relative entropy) can tell us how much they are different
  - if the user does not use any of such words, the risk is minimum, otherwise, it is maximum if these words are not used in the rest of the community

Assessing Privacy Risks

Second possibility (DIFF-PRIV)

- We look at the probability distribution of sensitive phrases, and calculate how much such distribution differs if we add the user's data
- This principle is called *differential-privacy*
- Here, the principle is similar than in the previous case: If the user uses sensitive words that are "uncommon" in the rest of the community, then she/he is flagged as "risky subject".

Assessing Privacy Risks

Third possibility

- We can construct distributional representations of the topics, users and posts
- These vectors can be constructed using
  - Bag-of-words (BOW)
  - LDA
  - Skip-gram models (W2V)
- Risk can be calculated using vector-based similarity measures (e.g. cosine-similarity)

Assessing Privacy Risks

Evaluation

- Used crowdsourcing service (AMT) to identify a list of sensitive topics
- People did not agree most of the times. Almost unanimous agreement on “drug addiction”, “pregnancy”, “clinical depression”
- Crawled user data from AOL queries, Quora, ehealthforum and healthboards (two websites about health communities)
- Human judges were used to identify a set of “real” sensible users

Assessing Privacy Risks

Evaluation

Table 4: Average metrics over all sensitive topics for different risk assessment measures

	R-precision		Prec@5	MAP	NDCG
	micro	macro			
AOL					
ENTROPY	0.495	0.496	**0.760**	0.524	0.819
DIFF-PRIV	0.475	0.465	0.480	0.492	0.789
W2V	**0.556***	**0.533**	0.720	**0.589**	**0.836**
Health Forums					
ENTROPY	0.560	0.537	**0.750**	0.613	0.870
DIFF-PRIV	0.560	0.559	0.500	0.542	0.794
W2V	**0.664***	**0.634**	**0.750**	**0.696**	**0.894**
Quora					
ENTROPY	0.239	0.205	0.240	0.317	0.632
DIFF-PRIV	0.239	0.223	0.200	0.310	0.623
W2V	**0.343***	**0.341**	**0.280**	**0.352**	**0.637**

Assessing Privacy Risks

Evaluation

- This work shows that it is fairly easy to detect sensible states by analyzing online content

- It differs from techniques to anonymize user activity since it aims to spot *when* activity might become dangerous

- Many more activities can be incorporated (sharing, retweets, etc.) in order to build more robust rankings