Using Twitter to measure interest in Privacy Areas

Esha Sharma,
Dr. Jessica Staddon
Can we trust these numbers?

• 85% of teens report to have Facebook profiles that are at least partially private online (Pew Research 2013)

• More than 70% of respondents are concerned about tracking online (Turow, June 2003)

• 93% of respondents are concerned about company and government access to health records (UPI Zogby, 2007)

• 60% of respondents have considered quitting Facebook over privacy concerns (Sophos, May 2010)

• 70% of respondents willing to trade computer passwords for candy (London survey, 2004)

• 52% of respondents report to have read a mobile app privacy policy (Harris/TRUSTe, 2011)
Privacy Paradox

• Behavioral decisions seem at odds with reported preferences.
Limited Privacy Awareness

Yet don’t always know how to do this.

- 33% are aware that they can read privacy policies
- 16% have done this
- 43% are aware they can change social media settings
- 29% have done this
- 43% are aware they can turn off location tracking
- 29% have done this
- 60% are aware that they can delete cookies, cache and browsing history
- 55% have done this

75% of Americans believe they adequately protect their personal online data

44% believe that online privacy will improve with greater consumer awareness
What could be causing this discrepancy?

• Education
  • Respondents learn about privacy risks through the survey
• Difficulty evaluating trade-offs
  • What do I get in return for my data?
• Recall is hard
• Survey phrasing
Using Twitter Data

• We can use Twitter data to measure privacy concern objectively by looking at the volume of tweets on a privacy concern.
• No surveys so we are not asking questions which could bias the responses.
• We can find interest levels and distribution of interest amongst different Privacy concerns.
• Given length of tweets what to cluster.
First : Classifying Tweets

- Twitter API provides filters.
- However the tweets filtered out using “Privacy” gets false positive (non Digital Privacy) as well.
- Hence, Building a classifier which identifies tweets about Digital Privacy.
- Used 6 different classifiers.
- Still evaluating those and finding ways to get more relevant clean data.
- Precision
- Recall: how do you find a baseline for this.
Next: Using these tweets to measure interest levels

• How do you use tweets to measure interest level?

• Number of retweets ("Discovering user interest on twitter with a modified author-topic model." *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2011.)


• Number of likes

• Find all twitter communities and find out size of the communities and there cohesiveness : (Lim, Kwan Hui, and Amitava Datta. "Following the follower: detecting communities with common interests on twitter." *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 2012.) very lengthy process : explain

• What else can be used?
Next: Clustering tweets and putting them in categories.

• Using clustering algorithms and finding clusters of common topics.
• How do we come up with a human understandable topic label?
• Classifying tweets into topics?
Problems

• How to measure cohesiveness within clusters.
• How to measure interest levels.