# Case Study: Random Samples From Twitter APIs May Not Be Random

By Jinseok Kim

# Example Data: Twitter Data

- Twitter data have been widely used to model social media user behaviors and information dissemination.
- Twitter data are available for download through Twitter APIs:
  - Free streaming or search API.
  - Premium API.
  - Enterprise-level API.
- Many research universities have access to the 10% random sample of the real-time Twitter (Decahose API).
  - Firehose API: Delivers all the Tweets in near real-time.

# Issues in Twitter Data Access

- In many cases, only a sample of Twitter data is available.
- Size, information types, and time-span of gathered Twitter data can vary depending on:
  - What APIs you use: search API vs streaming API.
  - When you use them: today vs 7 days ago.
- Twitter does not disclose details about how random samples of tweets are created.
- Are sampled Twitter data representative?
  - If not, then how can research mining the data be reliable?

# Comparing Sampled Twitter Datasets (1)

- A study by González-Bailón et al. (2014).
    - Gathered three sampled Twitter datasets through search and streaming APIs:
        - Search API: Collects a sample of tweets dating back to 7 days.
        - Streaming API: Collects up to 1% of all tweets on a real-time basis.
    - Compares whether Twitter users in the datasets are similar or different:
        - In terms of network centrality of each user.
        - A Twitter user is central if she or he “is mentioned more often or re-tweeted more times in the flow of” communication (p.18).

# Comparing Sampled Twitter Datasets (2)

- The study by González-Bailón et al. (2014) found that:
  - Smaller samples are not a random subset of the larger samples.
  - Smaller samples do not represent well the activities of users who are not central in the retweet and mention networks.
  - Mention networks are "more biased because … users who mentioned very often are not necessarily" active in retweeting.

# Bias Found in Sampled Twitter Data

- Another study by Morstatter el al. (2013) conducted similar comparisons in which the researchers:
  - Gathered Twitter data using a Firehose streaming API, which allows access to 100% of all public tweets, and a streaming API.
  - Compared the correlations of top topics (1) between the two datasets and then (2) between the streaming API sampled data and the 100 random samples of the Firehose data.

# Comparing Sampled Twitter Datasets (3)

- The study by Morstatter et al. (2013) found that:
  - Sampled data from streaming API estimated well top topics (by hashtags) of the Firehose data but not so much for unpopular topics.
  - Streaming API sometimes produced negative correlation of top topics with the Firehose data, while top topics in random samples were highly correlated with the Firehose data.
  - Conclusion: Streaming API can produce data that do not represent the population of the Twitter users and their activities.

# What’s Next?

- We will take a look at **data access threats** for designed data.

UNIVERSITY OF MICHIGAN