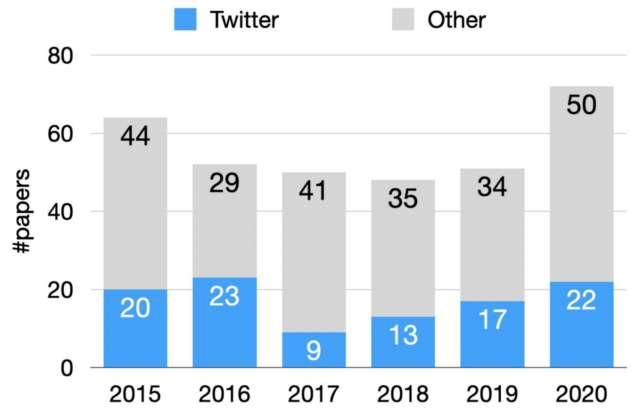
Variation across scales: Measurement fidelity under Twitter data sampling

Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie

Computational Media Lab @ANU

ICWSM '20

Twitter data is prevailing, but it may get sampled



104 (31%) out of 337 ICWSM papers use Twitter data (2015-2020)

API	Search	Sampled streaming	Filtered streaming
Usage	Retrieving relevant tweets given a query	Streaming a sample of public tweets	Streaming matched tweets given a query
Rate limiting	180 or 450 calls / 15 minutes	Roughly 1% of all public tweets	50 tweets / 1 second
Affected studies	*Most*, since it only searches tweets of the past 7 days	*All*, by default roughly 1%	USC COVID-19: ~5% sampling rate (Chen et al. '20)

RQ1. How are the tweets missing in the filtered stream?
RQ2. What are the effects on common measurements?

Contribution: **a comprehensive study of the Twitter sampling effects across different timescales and different subjects (entity, network, and cascade)**

Outline

1. Introduction
2. **How are the tweets missing in the filtered stream?**
   - Rate limit messages
   - Across different timescales -- hour, minute, second, and millisecond
3. **What are the sampling effects on common measurements?**
   - Across different subjects -- entity, network, and cascade
4. **Summary**

Twitter rate limit messages

- *Filtered streaming*: collecting tweets matching a set of prescribed predicates in realtime[1], e.g., "*COVID-19*"
- In each second, no more than 50 tweets will be returned[2].
- Rate limit messages indicate the cumulative number of missing tweets since the connection starts[3].
- Rate limit messages are NOT accurate (Sampson et al. '15), we explain the discrepancy in Appendix C.

Blocks of streamed tweets

{"id_str":"1245501748485242881", ...}
{"limit":{"track":28469226,"timestamp_ms":"1585785737733"}}
{"id_str":"1245501752088150021", ...}

{"id_str":"1245501752968908802", ...}
{"limit":{"track":28469434,"timestamp_ms":"1585785738725"}}
{"id_str":"1245501756315860992", ...}

{"id_str":"1245501756987097089", ...}
{"limit":{"track":28469643,"timestamp_ms":"1585785739742"}}
{"id_str":"1245501760568995842", ...}

1 sec, 28469434 - 28469226 = 208 missing

1 sec, 28469643 - 28469434 = 209 missing

[1] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter
[2] https://developer.twitter.com/en/docs/labs/filtered-stream/faq
[3] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/streaming-message-types

Constructing the complete filtered stream

- Strategy: splitting the filtering predicates into multiple subcrawlers.
- 2 datasets: Cyberbullying (sampling rate: 52.72%) and YouTube sharing (91.53%).

Id	Keywords	Languages	#collected tweets	#rate limit	#est. missing tweets	sampling rate
1	should	en	29,647,814	1,357	7,324	99.98%
2	should	all\en	801,904	0	0	100.00%
3	live	en	16,526,226	1,273	25,976	99.84%
4	live	all\en	7,926,325	233	7,306	99.91%
5	kill, fight, poser, nerd, freak, pig	all	15,449,973	16	108	100.00%
6	dick, suck, gay, loser, whore, cunt	all	13,164,053	15	125	100.00%
7	pussy, fat, die, afraid, emo, slut	all	21,333,866	89	1,118	99.99%
8	bitch, wannabe, whale, slept, caught	all	14,178,366	64	666	100.00%
complete	subcrawlers 1-8	all	114,488,537	3,047	42,623	99.96%
sample	all 25 keywords	all	60,400,257	1,201,315	54,175,503	52.72%

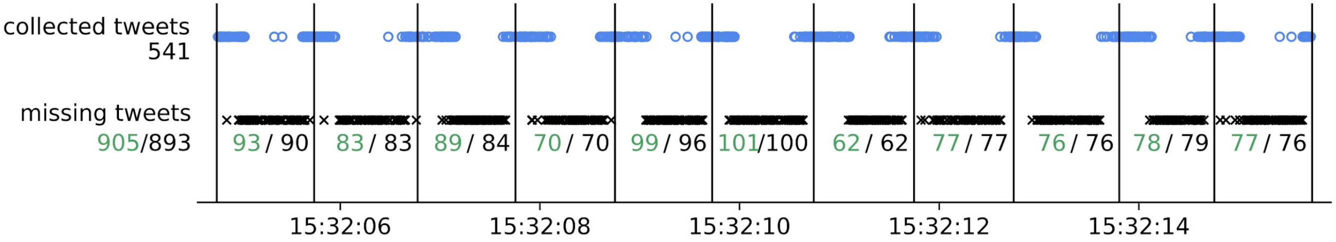
[1] Obtained via a Twitter data reseller https://discovertext.com/

Constructing the complete filtered stream

- Strategy: splitting the filtering predicates into multiple subcrawlers.
- 2 datasets: Cyberbullying (sampling rate: 52.72%) and YouTube sharing (91.53%).
- Validation: single crawler + rate limit messages *vs.* (1) multiple subcrawlers / (2) Firehose stream[1].

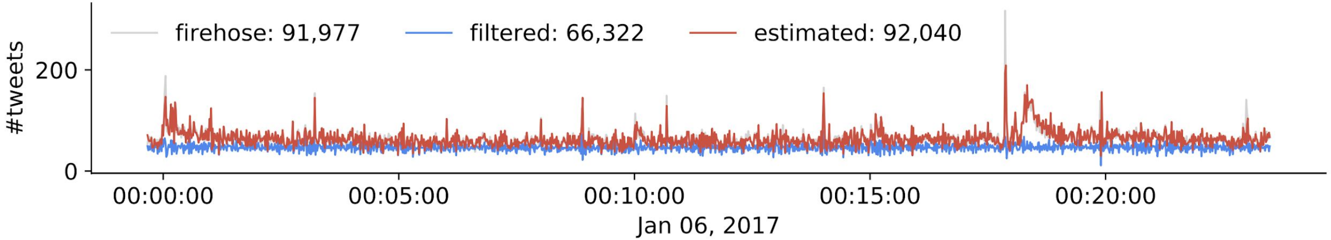
vs. multiple subcrawlers

MAPE: 0.001



vs. Firehose stream

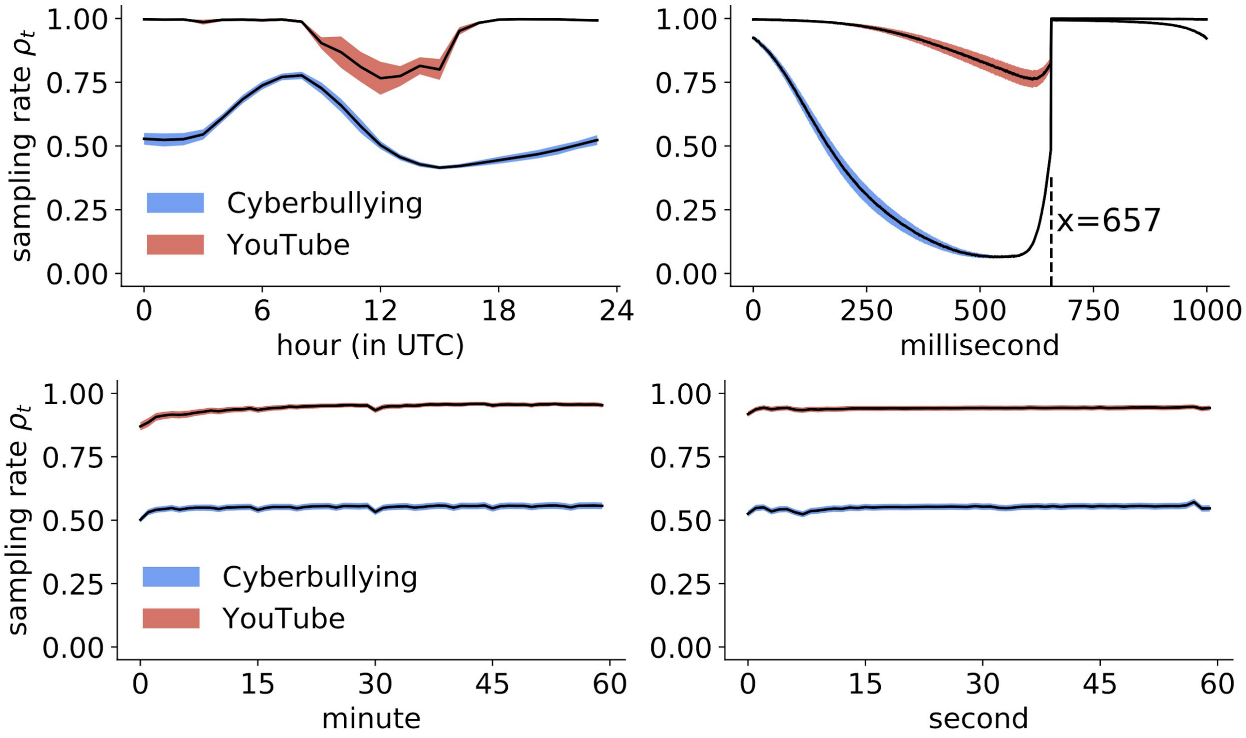
MAPE: 0.007



[1] Obtained via a Twitter data reseller https://discovertext.com/

Temporal variation of sampling rates

- Sampling rates are uneven in different hours or in different milliseconds, but are almost the same at the timescale of minute and second.



Outline

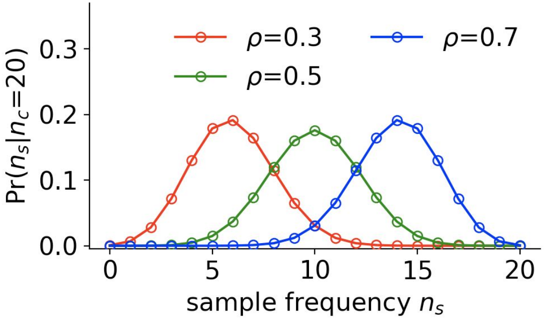
1. Introduction
2. How are the tweets missing in the filtered stream?
   - The volume of missing tweets can be estimated by Twitter rate limit messages.
   - Tweet sampling rates vary across different timescales.
3. **What are the sampling effects on common measurements?**
   - Across different subjects -- entity, network, and cascade
4. **Summary**

Twitter sampling as a Bernoulli process

- Assumption used in prior studies but not validated (Joseph et al. '13, Pfeffer et al. '18).
- Complete frequency → Sample frequency: binomial distribution Pr(n_s)~*Binomial*(n_c, *p*).

$$\Pr(n_s|n_c,\bar{\rho}) = \binom{n_c}{n_s}\bar{\rho}^{n_s}(1-\bar{\rho})^{n_c-n_s}$$

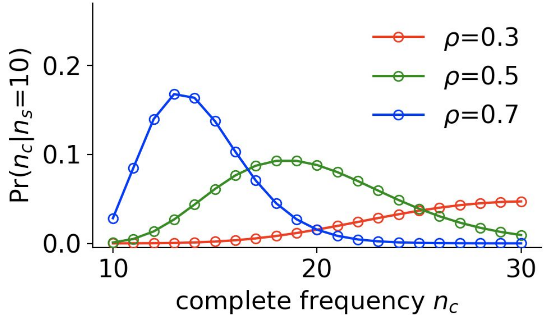
$$\mathrm{E}(n_s) = n_c\bar{\rho}$$



- Sample frequency → Complete frequency: negative binomial distribution Pr(n_c)~*NegBinomial*(n_s, *p*).

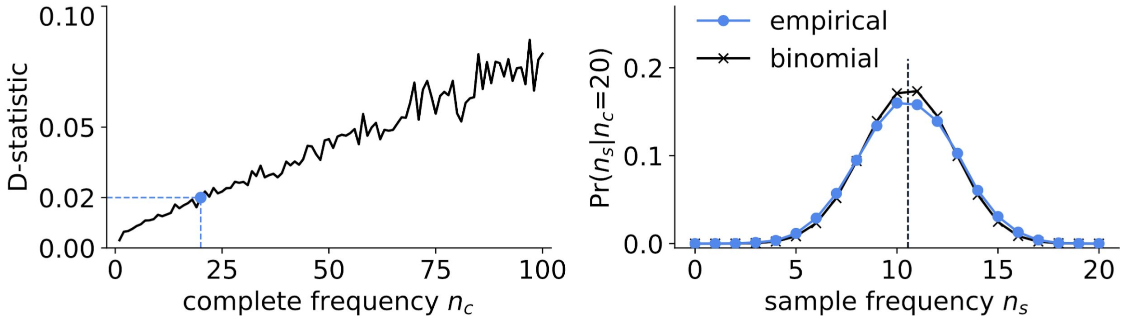
$$\Pr(n_c|n_s,\bar{\rho}) = \binom{n_c-1}{n_s-1}\bar{\rho}^{n_s}(1-\bar{\rho})^{n_c-n_s}$$

$$\mathrm{E}(n_c) = \frac{n_s}{\bar{\rho}}$$

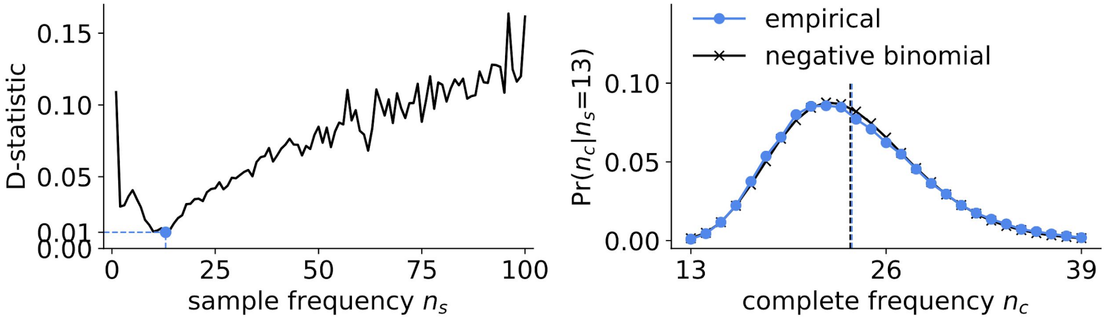


Using Bernoulli process with a uniform rate to approximate the empirical data

- *metric*: D-statistic (Leskovec and Faloutsos '06). $D(G, G') = \max_x \{|G(x) - G'(x)|\}$
- Complete frequency → Sample frequency: binomial distribution Pr(n_s)~*Binomial*(n_c, *p*).

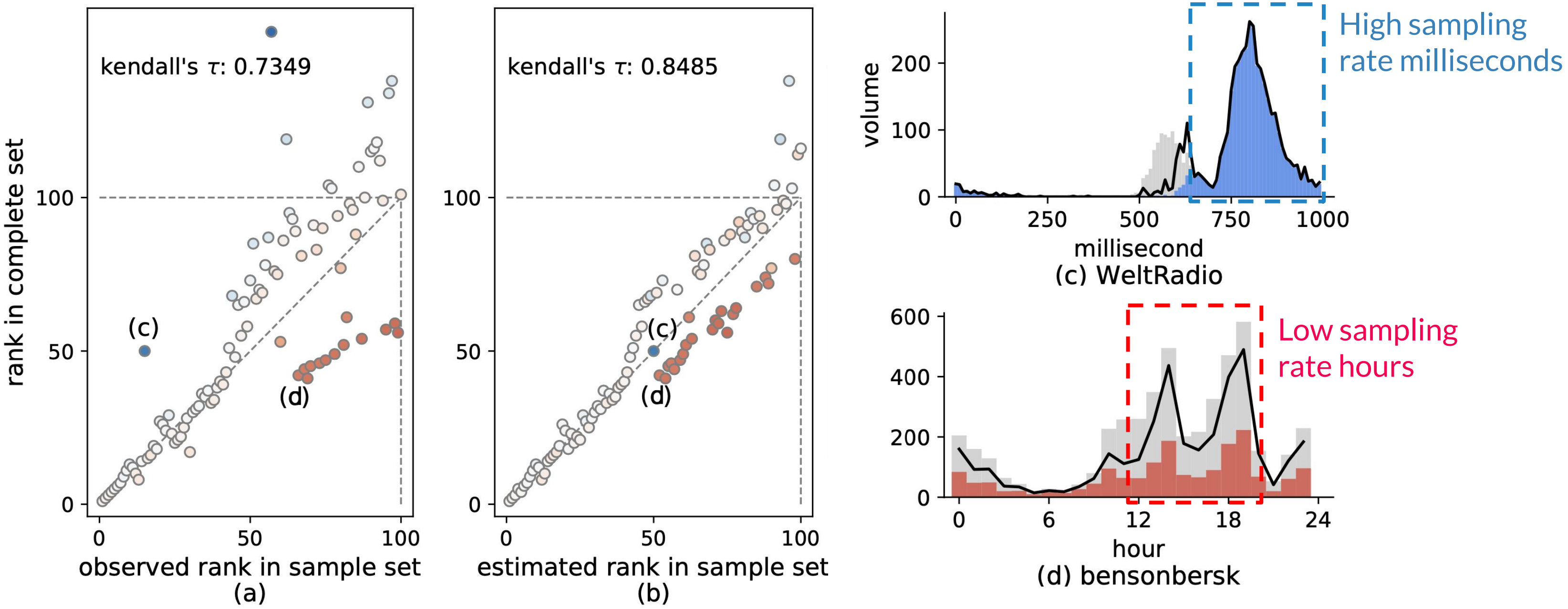


- Sample frequency → Complete frequency: negative binomial distribution Pr(n_c)~*NegBinomial*(n_s, *p*).



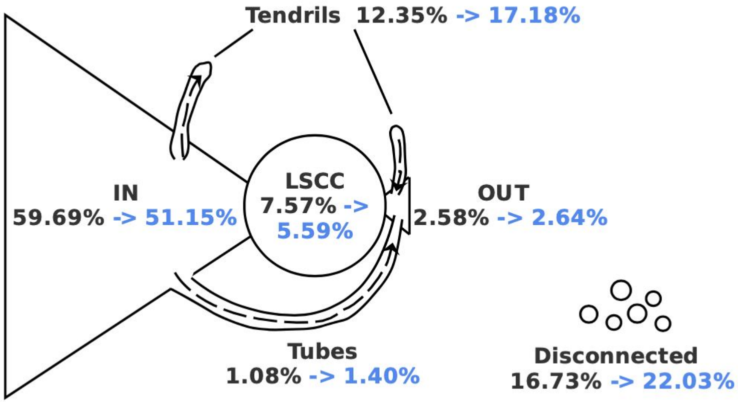
Estimating true ranking from the sample set

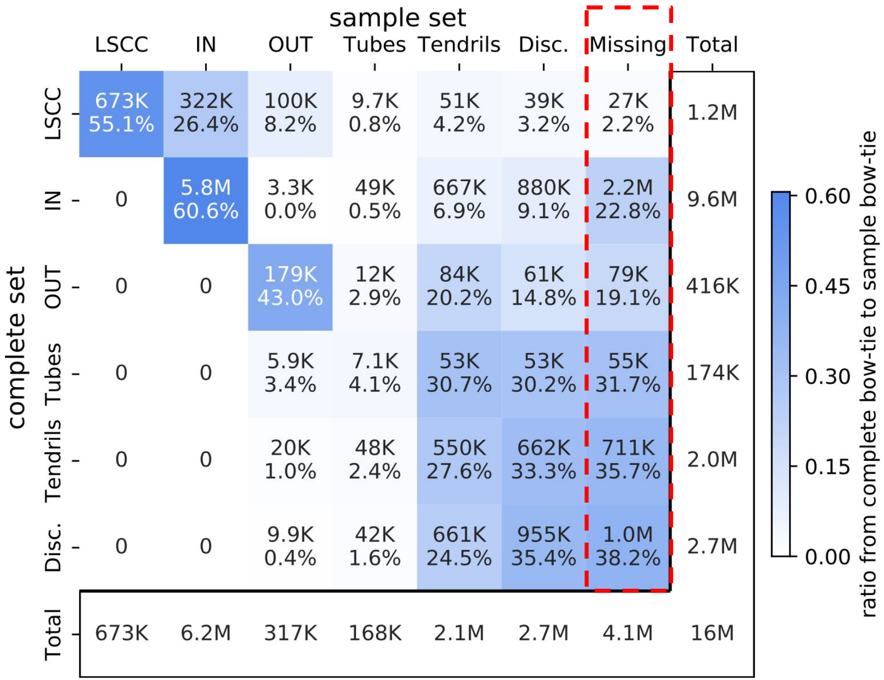
- The ranks of most active users are distorted, but can be corrected.



Denser components are more likely to be preserved

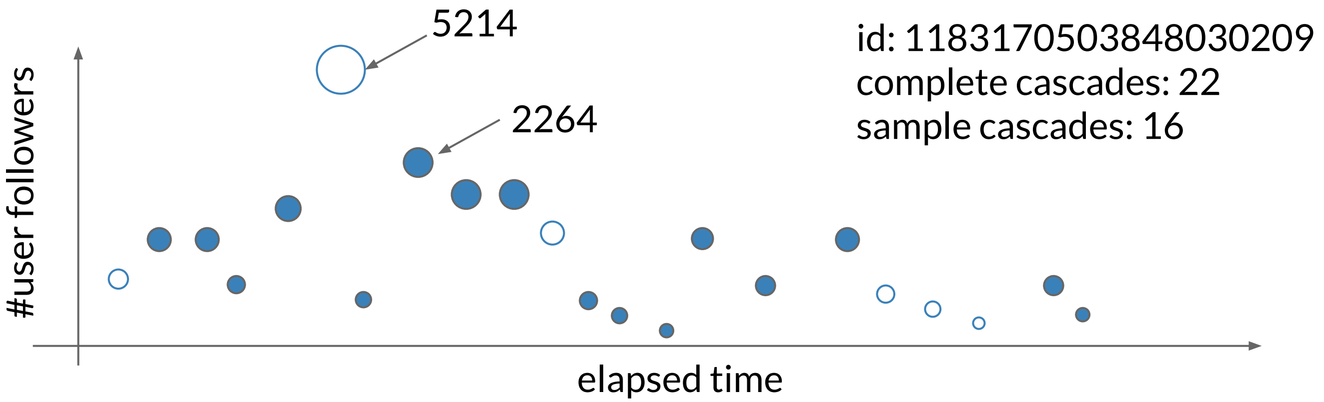
- *Bow-tie structure* to characterize the user-user retweet network (Broder et al. '00).





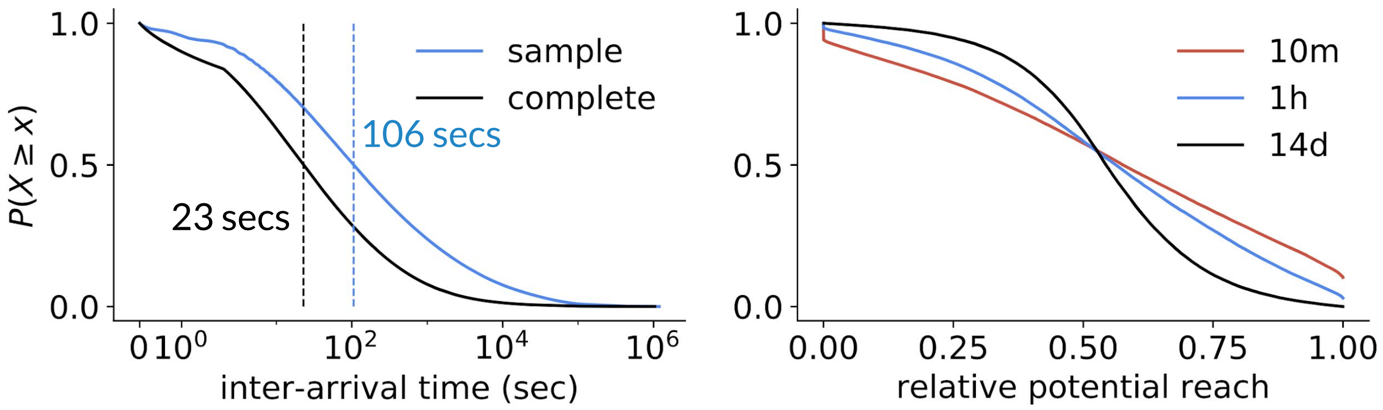
Impacts on retweet cascades

- 2 prominent features: *inter-arrival time*, *user influence* (Zhao et al. '15, Mishra et al. '16).

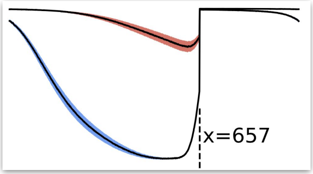


Impacts on retweet cascades

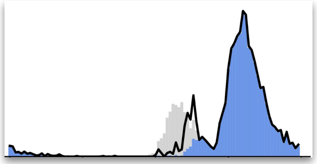
- 2 prominent features: *inter-arrival time, user influence* (Zhao et al. ’15, Mishra et al. ’16).
- Strong risks in research that concerns the activity history of each user (Gaffney and Matias ’18).

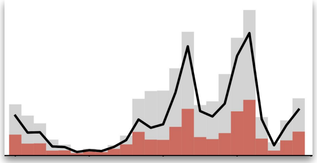


Summary



1. **How are the tweets missing in the filtered stream?**
   - The volume of missing tweets can be estimated by Twitter rate limit messages.
   - Tweet sampling rates vary across different timescales.





2. **What are the sampling effects on common measurements?**
   - Bernoulli process with a uniform rate can approximate the empirical entity distribution.
   - True entity ranking can be inferred based on sampled observations.
   - Network structures are altered with some components more likely to be preserved.
   - Sampling compromises the quality of diffusion models, since inter-arrival time is significantly longer in the sampled stream, while user influence is lower.

Variation across Scales: Measurement Fidelity under Twitter Data Sampling

Software, code and data: https://github.com/avalanchesiqi/twitter-sampling