Carnegie Mellon
institute for SOFTWARE RESEARCH

A Large-Scale Empirical Study of Geotagging Behavior on Twitter

Binxuan Huang
binxuanh@cs.cmu.edu
Kathleen M. Carley
kathleen.carley@cs.cmu.edu

Slides: https://binxuan.github.io/files/asonam_2019.pdf

CASOS

Outline

- Introduction
- Data collection
- Tweet-level analysis
- User-level analysis
- Graph-level analysis

Outline

- Introduction
- Data collection
- Tweet-level analysis
- User-level analysis
- Graph-level analysis

Introduction

- Location sharing behaviors widely exist in social network websites
  - Users post home locations in their profiles.
  - People mention where they are in their posts.
  - Geotag: directly tag posts either with a place id or with a precise geo-coordinates.
- Geotagging behaviors on twitter:
  - Place-tag: tag a tweet with a place (a geo bounding box) --- country, admin, city, neighborhood, poi (place of interest)
  - Coordinates-tag: tag a tweet with precise geo-coordinates.

Introduction

- Use geotagged users' opinions to infer non-geotagged local users' opinions.
  - **RQ1**: What if different users have different geotagging preferences? Are there any differences in terms of geotagging behavior among different users?
- Learn location specific features from geotagged tweets [7], based on information such as profile location.
  - **RQ2**: Are users who use geotags and who do not are equally likely to report their home locations in profiles? Is there any correlation between the geotagging behavior and the behavior of reporting location in profile?
- Utilize user's friends' locations to better geolocate this user [8].
  - If non-geotagged users tend to connect to similar non-geotagged users, then it would be harder to infer their locations based on their social ties.
  - **RQ3**: Is there any homophily effect between friends in terms of geotagging preference?

Outline

- Introduction
- Data collection
- Tweet-level analysis
- User-level analysis
- Graph-level analysis

Data collection

- Step 1: use twitter sample streaming API to get real-time tweets without any filter parameters.
- Step 2: Extract users in the sampled data and collect their recent 3200 tweets and following friends.
- Step 3: Take the users both with following data and timeline data as the final research objects.

Data collection

- Data summary

# of Tweets	# of Tweeters	Following ties	Place-tagged tweets	Coordinates-tagged tweets
41,267,348,020	19,984,064	4,402,458,603	724,933,445 (1.76%)	228,606,700 (0.55%)

About 2.31% of tweets are geotagged, slightly higher than previous estimation [1]

Outline

- Introduction
- Data collection
- Tweet-level analysis
- User-level analysis
- Graph-level analysis

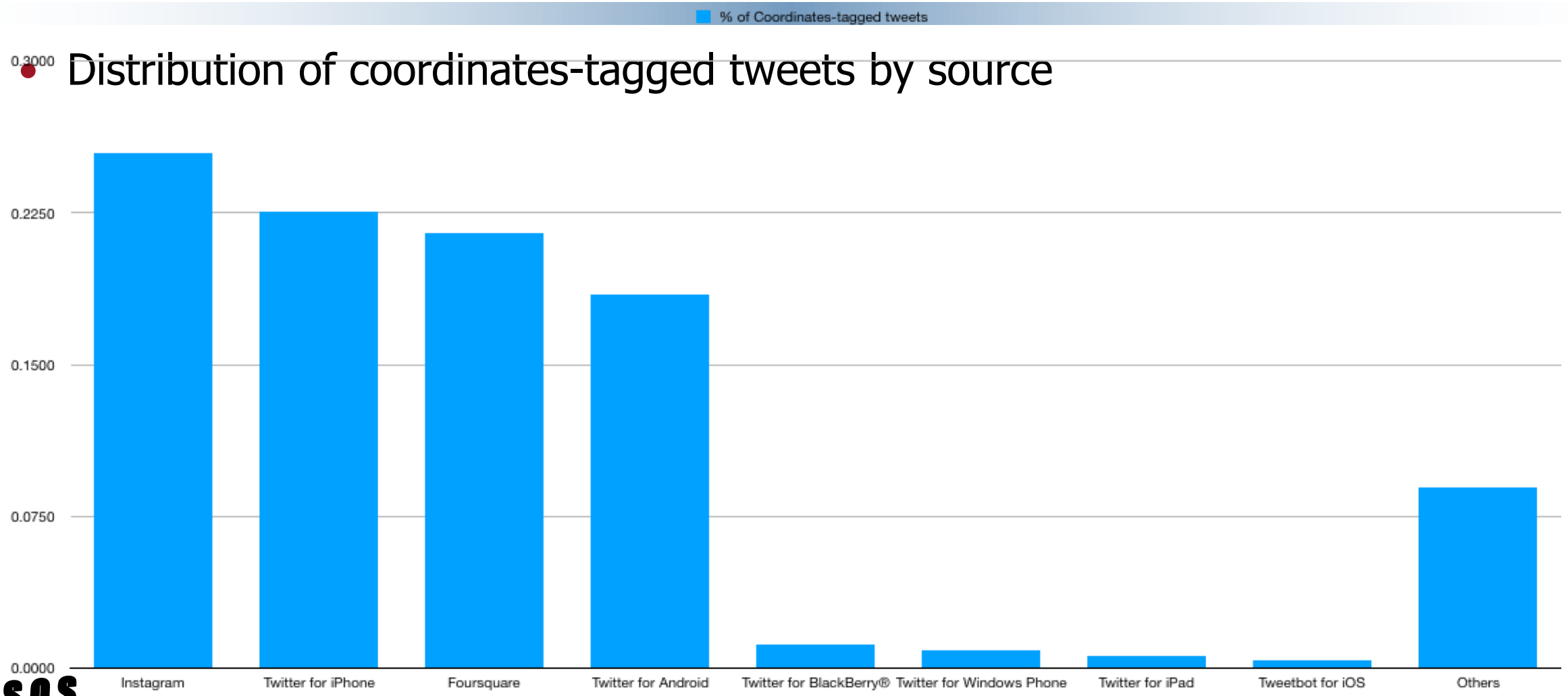
Tweet-level Analysis

- Geotagging distribution by source

source	# of Tweets	# of non-geotaged tweets	# of Place-tagged Tweet	# of Coordinates-tagged Tweets
Twitter for iPhone	16,161,407,831	15,716,820,447 (97.25%)	393,059,787 (2.43%)	51,527,597 (0.32%)
Twitter for Android	11,938,888,612	11,677,533,121 (97.81%)	219,107,978 (1.84%)	42,247,513 (0.35%)
Twitter Web Client	4,184,897,568	4,088,127,646 (97.69%)	96,643,283 (2.31%)	126,639 (0%)
twittbot.net	916,067,510	916,067,510 (100%)	0	0
Facebook	769,543,040	769,543,040 (100%)	0	0
Twitter for iPad	633,139,301	624,738,931 (98.67%)	6,979,518 (1.10%)	1,420,852 (0.22%)
TweetDeck	526,790,924	526,711,888 (99.98%)	25,725 (0)	53,311 (0.01%)
Twitter Lite	500,813,124	500,696,593 (99.98)	64 (0)	116,467 (0.02%)
Instagram	304,274,973	246,133,428 (80.89%)	1,470 (0)	58,140,075 (19.11%)
Others	5,331,525,137	5,247,435,271 (98.42)	9,115,620 (0.17%)	74,974,246 (1.41%)

Tweet-level Analysis

- Distribution of coordinates-tagged tweets by source



Carnegie Mellon
institute for SOFTWARE RESEARCH

Tweet-level Analysis

Place-type distribution among place-tagged tweets

source	country	admin	city	neighborhood	poi
Twitter for iPhone	6,448,752 (1.64%)	47,820,014 (12.17%)	336,007,216 (85.49%)	269,817 (0.07%)	2,513,988 (0.64%)
Twitter for Android	5,633,166 (2.57%)	23,219,855 (10.60%)	188,822,442 (86.18%)	278,991 (0.13%)	1,153,524 (0.53%)
Twitter Web Client	14,180,029 (14.67%)	15,557,789 (16.10%)	66,777,976 (69.10%)	127,489 (0.13%)	0 (0%)
Twitter for iPad	144,597 (2.07%)	647,221 (9.27%)	6,131,541 (87.85%)	9,407 (0.13%)	46,752 (0.67%)
Tweetbot for iOS	77,587 (1.40%)	617,288 (11.13%)	3,604,772 (64.98%)	1,247,505 (22.49%)	0 (0%)
Tweetbot for Mac	14,989 (0.95%)	154,566 (9.84%)	923,905 (58.79%)	477,944 (30.42%)	0 (0%)
Others	94,431 (2.63%)	424,966 (11.82%)	2,553,342 (71.01%)	516,441 (14.36%)	6,540 (0.18%)
sum	26,578,562 (3.67%)	88,287,133 (12.18%)	603,897,289 (83.30%)	2,449,650 (0.34%)	3,720,804 (0.51%)

Tweet-level Analysis

- Coordinates tagging percentage by country

TABLE V: Percentages of coordinates-tagged tweets for countries (top 15).

country	# of geotagged	# of coordinates-tagged
United States	320,268,573	61,488,648 (19.20%)
Brazil	117,794,897	19,509,860 (16.56%)
United Kingdom	59,983,328	14,585,781 (24.32%)
Japan	51,847,289	13,406,333 (25.86%)
Argentina	39,744,563	6,350,980 (15.98%)
Turkey	35,989,555	19,037,195 (52.90%)
Philippines	29,031,714	4,696,974 (16.18%)
Mexico	24,317,203	8,132,105 (33.44%)
Spain	22,608,661	6,114,047 (27.04%)
Malaysia	22,036,169	8,096,642 (36.74%)
Indonesia	18,581,142	12,357,982 (66.51%)
France	16,049,418	2,690,101 (16.76%)
Canada	13,142,453	2,987,905 (22.73%)
Russia	11,241,015	2,844,891 (25.31%)
Saudi Arabia	10,728,248	1,568,910 (14.62%)

Tweet-level Analysis

- Geotagging distribution by tweet lang.

TABLE IV: Distributions of geotags for tweets with different tweet languages (top 15)

Lang.	Non-geotagged	Place-tagged	Coordinates-tagged
English	14,209,166,056 (97.04%)	330,133,459 (2.25%)	103,425,905 (0.71%)
Japanese	7,920,019,090 (99.36%)	37,943,333 (0.48%)	13,167,513 (0.17%)
Spanish	4,405,421,261 (97.39%)	89,268,301 (1.97%)	28,635,593 (0.63%)
Arabic	3,063,691,725 (99.29%)	18,598,541 (0.60%)	3,335,846 (0.11%)
Portuguese	2,366,206,011 (95.67%)	91,288,151 (3.69%)	15,848,111 (0.64%)
und	2,327,452,586 (97.38%)	53,030,818 (2.22%)	9,673,783 (0.40%)
Korean	1,013,674,569 (99.86%)	976,321 (0.10%)	436,763 (0.04%)
French	820,763,421 (98.05%)	13,489,214 (1.61%)	2,834,209 (0.34%)
Indonesian	778,781,264 (96.02%)	16,621,968 (2.05%)	15,651,480 (1.93%)
Thai	729,496,967 (99.00%)	5,588,546 (0.76%)	1,748,479 (0.24%)
Turkish	670,097,929 (95.38%)	16,085,714 (2.29%)	16,352,663 (2.33%)
Tagalog	473,457,971 (96.02%)	16,112,019 (3.27%)	3,498,319 (0.71%)
Russian	320,550,743 (96.08%)	10,132,568 (3.04%)	2,929,779 (0.88%)
Italian	229,578,445 (97.03%)	5,093,448 (2.15%)	1,936,471 (0.82%)
German	162,995,168 (97.57%)	2,847,769 (1.70%)	1,214,755 (0.73%)

Outline

- Introduction
- Data collection
- Tweet-level analysis
- User-level analysis
- Graph-level analysis

User-level analysis

User-level geotagging is more prevalent

# of Tweeters	Tweeters with at least one geotagged tweets	Tweeters with at least one precise coordinates-tag
19,984,064	4,871,784 (24.38%)	2,584,042 (12.93%)

User-level analysis

- Divide users into categories based on:
  - 1. Source
  - 2. Language
  - 3. Profile location
    - Do not provide profile location
    - Provide meaningful profile location (We use Geonames to recognize locations)
    - Provide meaningless profile location (cannot be detected by Geonames)

- Look at the percentage of place tags and coordinates tags by categories.

User-level analysis

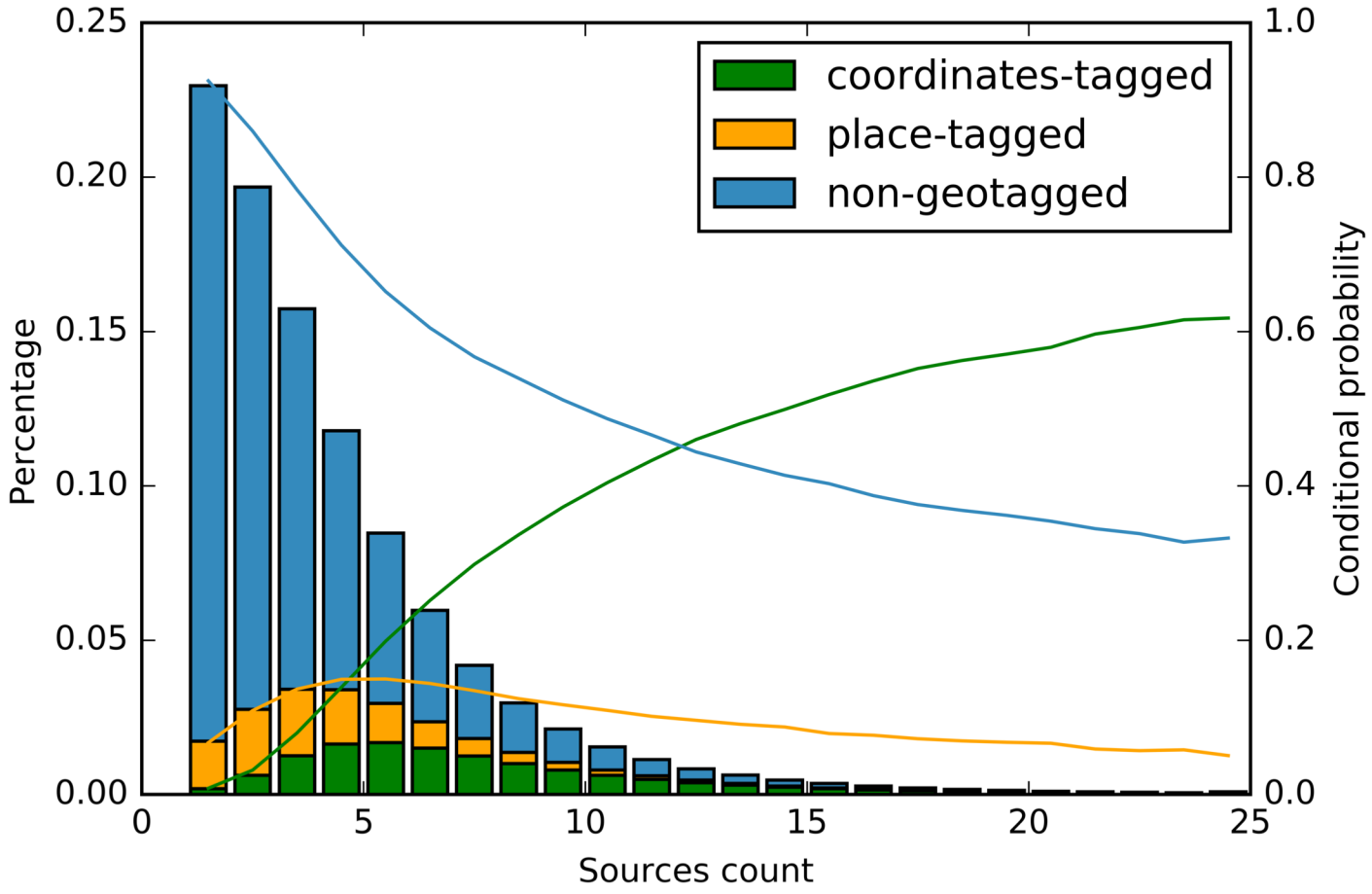
- Again, the geotagging distributions differ because of the settings of different platforms

TABLE VI: The geotagging distributions for users with different user sources (Top 10).

User source	Non-geotagged	Place-tagged	Coordinates-tagged
Twitter for iPhone	5,741,431 (70.80%)	1,242,823 (15.33%)	1,125,263 (13.88%)
Twitter for Android	4,869,846 (76.08%)	670,206 (10.47%)	860,953 (13.45%)
Twitter Web Client	1,497,191 (77.99%)	214,419 (11.17%)	208,006 (10.84%)
Facebook	289,930 (74.43%)	29,737 (7.63%)	69,869 (17.94%)
twittbot.net	301,390 (99.18%)	1,076 (0.35%)	1,408 (0.46%)
Twitter for iPad	210,894 (82.99%)	19,759 (7.78%)	23,482 (9.24%)
TweetDeck	215,006 (85.36%)	16,709 (6.63%)	20,163 (8.01%)
Twitter Lite	233,175 (93.24%)	9,635 (3.85%)	7,262 (2.90%)
Google	104,839 (86.39%)	5,906 (4.87%)	10,608 (8.74%)
Instagram	58,931 (55.44%)	978 (0.92%)	46,391 (43.64%)

User-level analysis

- Generally the more twitter sources an used the more likely he/she would be geotagged.



User-level analysis

- We can find geo-coordinates for more than 30% of people who speak Indonesian
- Less than 3% of Korean speaker have ever used geotags before.

TABLE VII: Geotagging distributions for different user languages (top 10).

User lang.	non-geotagged	place-tagged	coordinates-tagged
English	5,232,717 (69.77%)	1,157,694 (15.44%)	1,109,592 (14.79%)
Japanese	3,539,998 (90.22%)	204,787 (5.22%)	178,761 (4.56%)
Spanish	1,521,361 (66.63%)	278,065 (12.18%)	483,997 (21.20%)
Arabic	1,326,259 (87.08%)	108,505 (7.12%)	88,258 (5.79%)
Portuguese	806,927 (64.12%)	218,648 (17.37%)	232,904 (18.51%)
Korean	538,095 (97.06%)	9,702 (1.75%)	6,584 (1.19%)
Turkish	311,746 (66.17%)	49,714 (10.55%)	109,671 (23.28%)
French	345,166 (75.65%)	59,595 (13.06%)	51,531 (11.29%)
Thai	301,935 (78.89%)	33,497 (8.75%)	47,283 (12.35%)
Indonesian	217,667 (57.53%)	34,889 (9.22%)	125,771 (33.24%)

User-level analysis

- Among these 20 million users, 38.6% of them do not provide location in their profile. Only 41.2% of them provide recognizable location by Geonames.

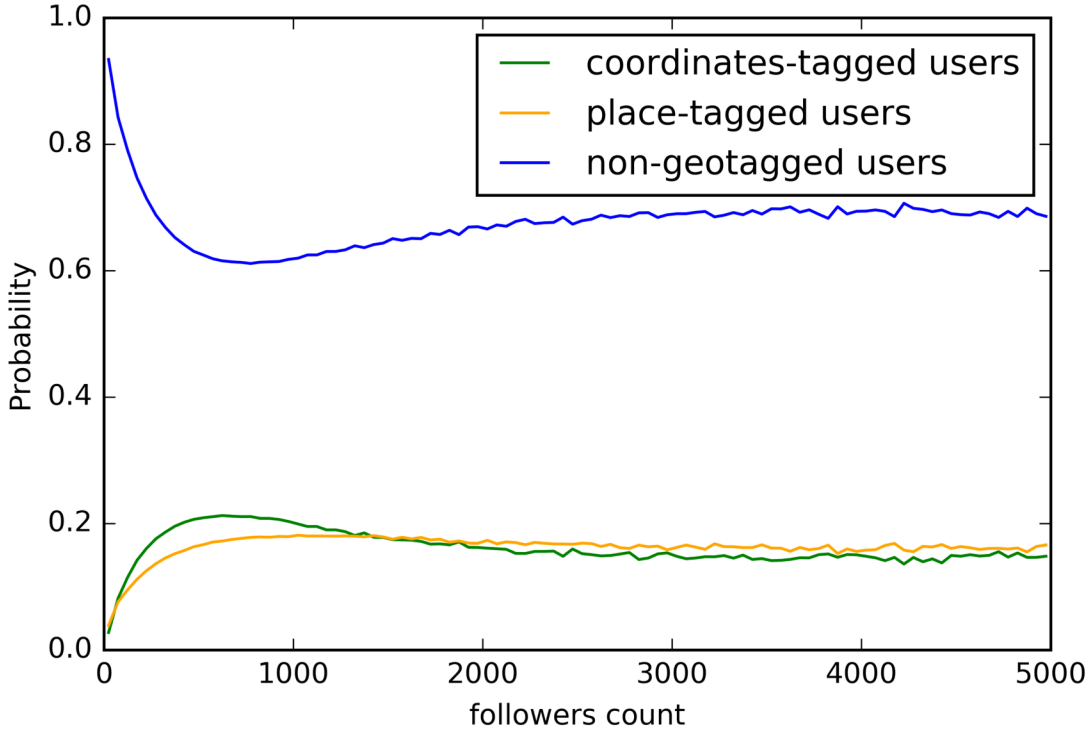
Profile location type	Nongeotagged	Place-tagged	Coordinates-tagged
Empty	6,489,046 (84.1%)	625,701 (8.1%)	602,036 (7.8%)
Unrecognized	3,111,036 (77.1%)	446,833 (11.1%)	477,919 (11.8%)
Recognized by Geonames	5,512,198 (67.0%)	1,215,208 (14.8%)	1,504,087 (18.3%)

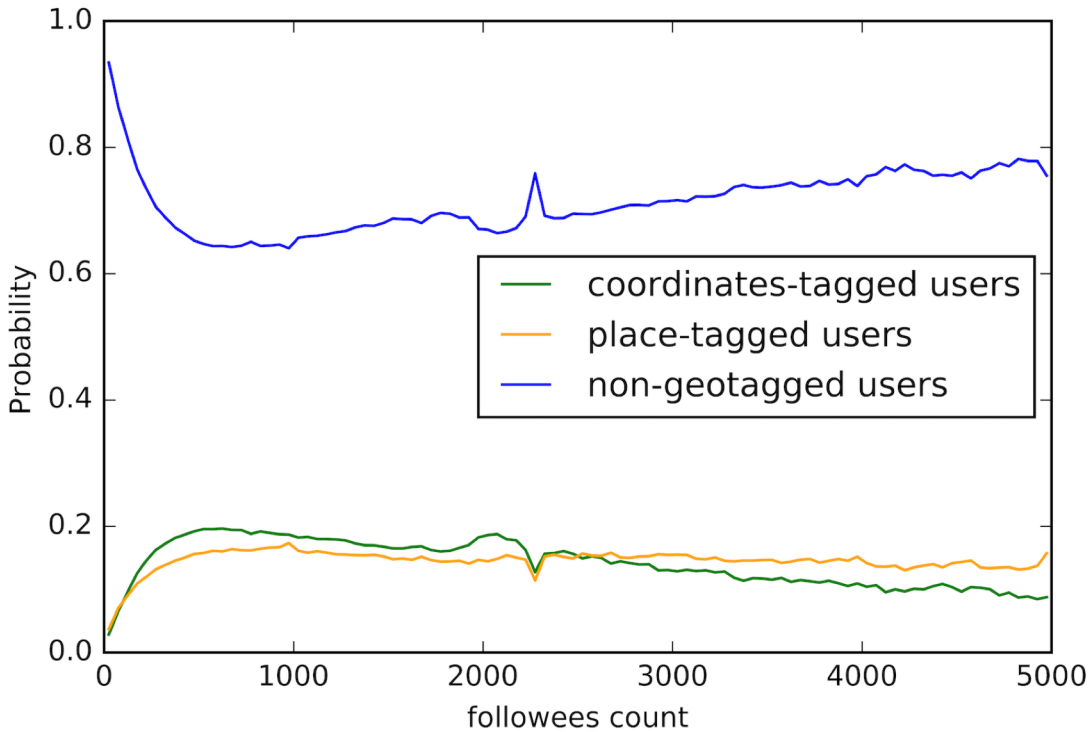
Outline

- Introduction
- Data collection
- Tweet-level analysis
- User-level analysis
- Graph-level analysis

Graph-level analysis

- When an user's follower/followee number reaches a certain threshold, he/she is more conservative for sharing real-time location.



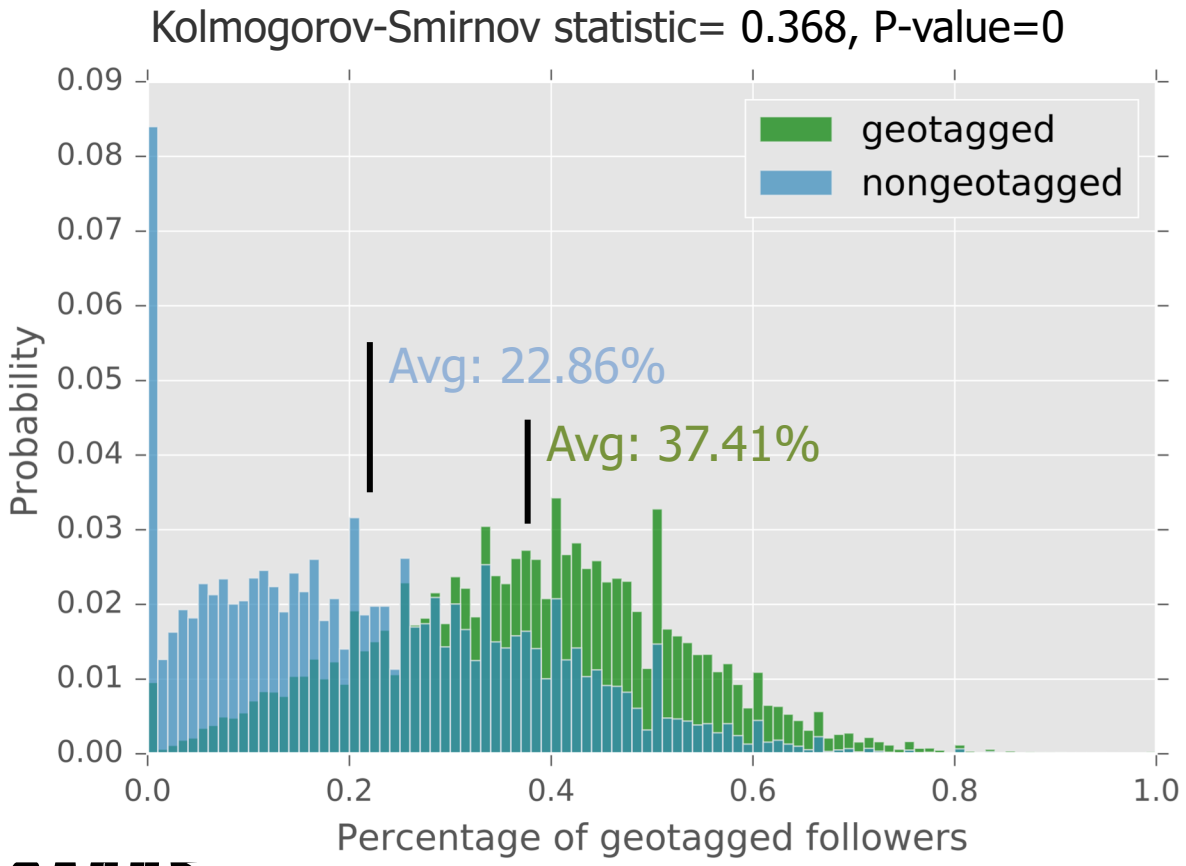


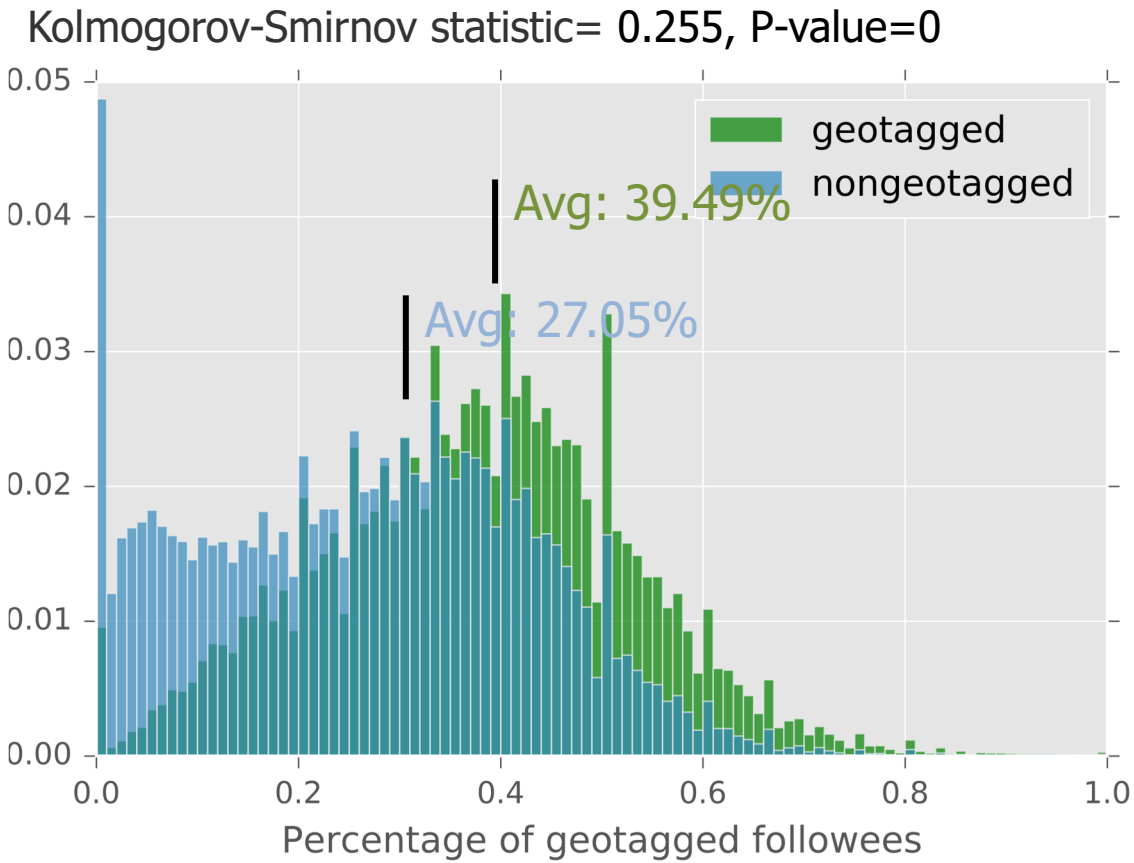
Graph-level analysis

- Location sharing homophily:
  - If user's following friends frequently share their locations, will this user also share his/her location via geotagging?

User-level analysis

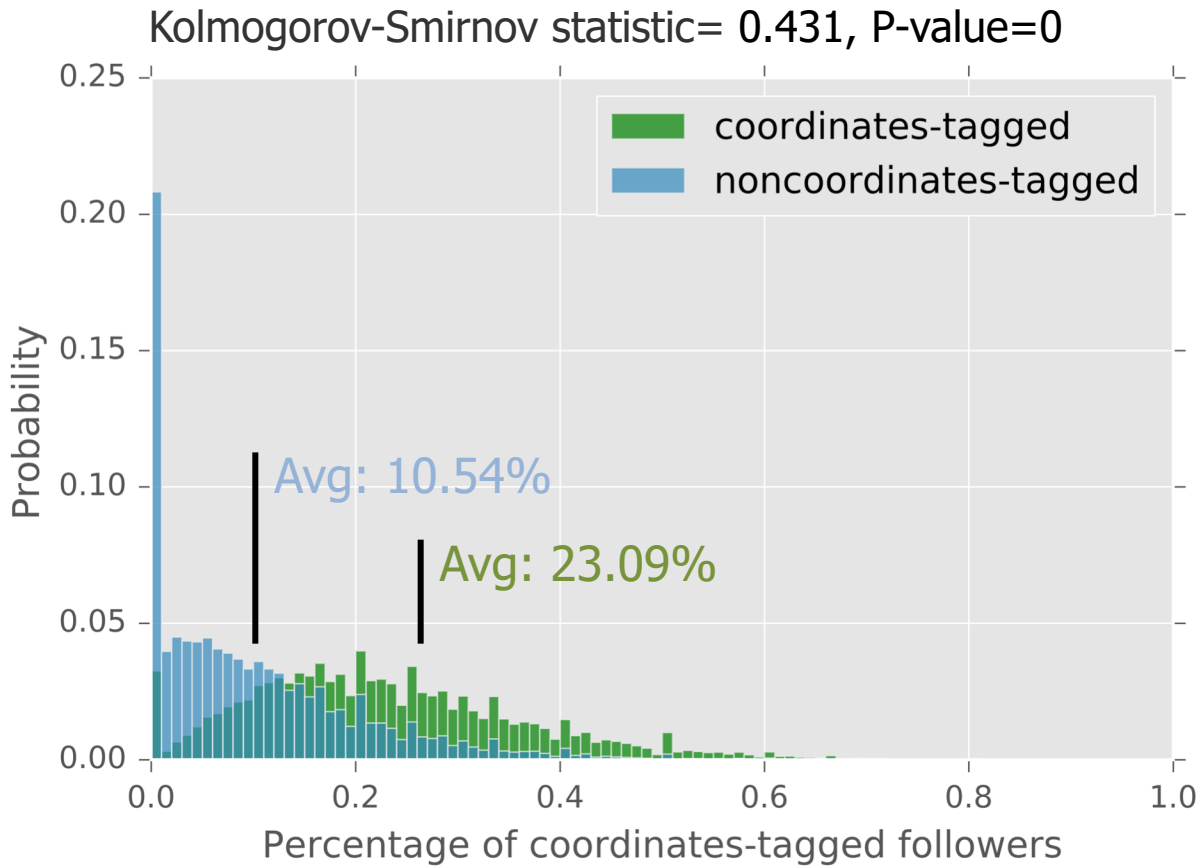
- Geotagged users are more likely to have geotagged followers/followees

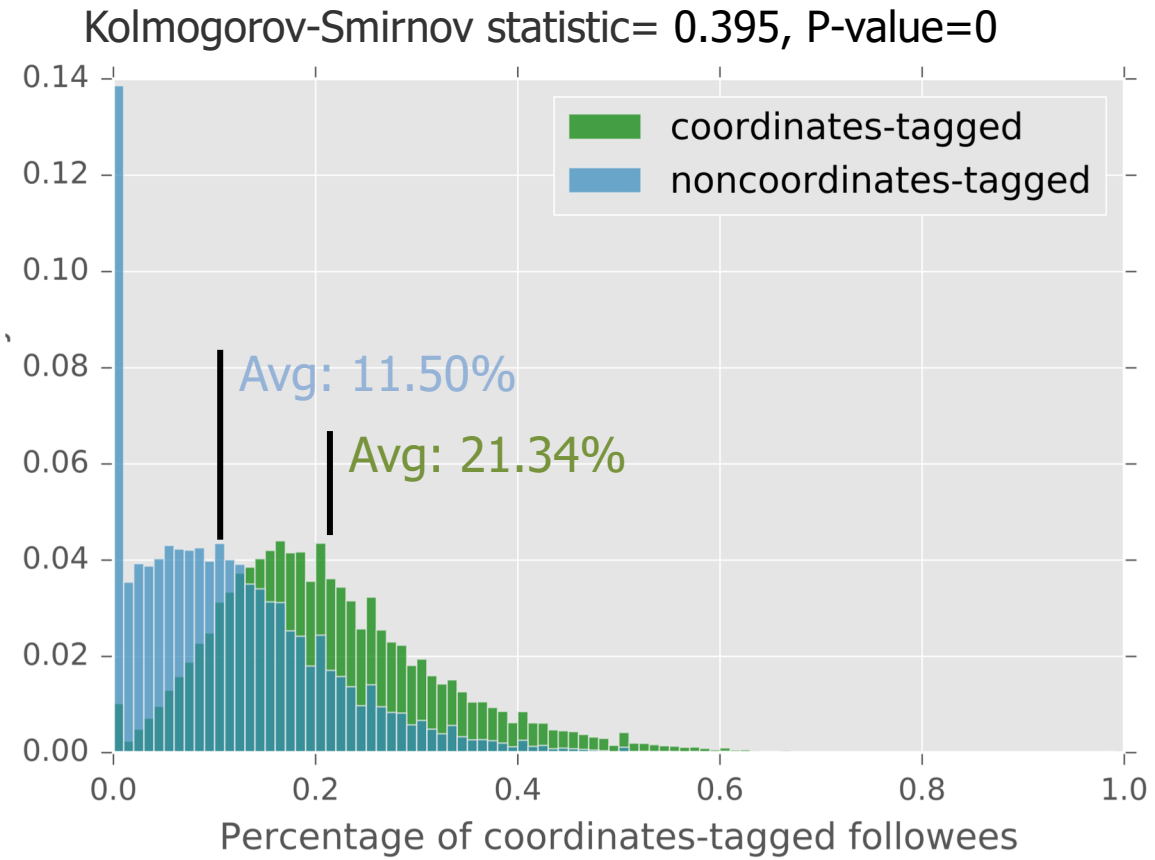




User-level analysis

- Coordinates-tagged users are more likely to have Coordinates-tagged followers/followees





User-level analysis

- An ego's chance of being geotagged increased more than 6 times if at least one of his/her alter is geotagged.

Alter	P(Ego is geotagged | at least one alter is geotagged)	P(Ego is geotagged | no alter is geotagged)	Relative increase
follower	28.70%	4.17%	6.88
followee	26.06%	1.76%	14.80
friend	30.60%	4.40%	6.95

User-level analysis

- Similar thing all happens for coordinates-tagging behavior

	P(Ego is coordinates-tagged | at least one alter is coordinates-tagged)	P(Ego is coordinates-tagged | no alter is coordinates-tagged)	Relative increase
follower	16.74%	2.75%	6.08
followee	14.82%	1.22%	12.11
friend	18.01%	2.97%	6.05

Conclusion

- Are there any differences in terms of geotagging behavior among different users?
  - Yes, factors include source, language, original country
  - Geotagged content may not be representative of public opinion in the corresponding region.
- Is there any correlation between reporting location and geotagging bahavior?
  - Users who self-report their location in profile are much more likely to use geotags.
  - Geolocation prediction systems may be less useful than previously thought, because a disproportionate number of users that use geotags also report locations.
- Is there any homophily effect between friends in terms of geotagging preference?
  - Yes, an ego's chance of being geotagged increased more than 6 times if at least one of his/her alter is geotagged.
  - If non-geotagged users tend to cluster together, then it becomes harder to find non-geotagged users' location based on the information from their friends.

References

[1] M. Graham, S. A. Hale, and D. Gaffney, "Where in the world are you? geolocation and language identification in twitter," The Professional Geographer, vol. 66, no. 4, pp. 568–578, 2014.

[2] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in Proceedings of the 7th ACM conference on Recommender systems. ACM, 2013, pp. 25–32.

[3] Y. Wei and L. Singh, "Location-based event detection using geotagged semantic graphs," in KDD Workshop Mining and Learning with Graphs, 2017.

[4] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "# earthquake: Twitter as a distributed sensor system," Transactions in GIS, vol. 17, no. 1, pp. 124–147, 2013.

[5] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual review of sociology, vol. 27, no. 1, pp. 415–444, 2001.

[6] M. J. Widener and W. Li, "Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us," Applied Geography, vol. 54, pp. 189–197, 2014.

References

[7] B. Huang and K. M. Carley, "On predicting geolocation of tweets using convolutional neural networks," in International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, 2017, pp. 281–291

[8] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 61–70.

- Thanks

binxuanh@cs.cmu.edu
http://binxuan.github.io