Tweeting Under Pressure: Analyzing Trending Topics and Evolving Word Choice on Sina Weibo

Le Chen Chi Zhang Christo Wilson
College of Computer and Information Science
Northeastern University

Censorship in China

- Great Firewall of China (GFW)

Censorship on Weibo

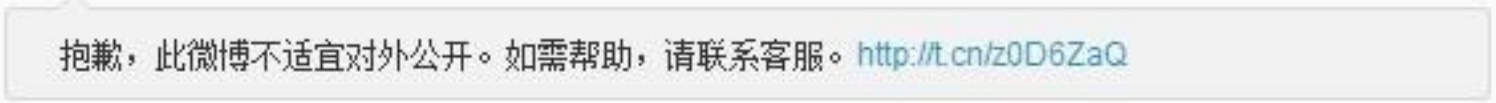
- What will be censored?

- How is censorship implemented?

THE WORST
THING ABOUT
CENSORSHIP
IS [redacted]

Current Understanding of Censorship on Weibo

- Scale of censorship[1][2]
  - 16% of censored tweets vs. 0.01% of censored tweets
- Velocity of censorship[3]
  - Focus on 3K of Chinese celebrities
  - 30% of tweets are deleted within 30 minutes



Sorry, this tweet is inappropriate to be public…

[1] BAMMAN, D., O'CONNOR, B., AND SMITH, N. A. Censorship and Deletion Practices in Chinese Social Media. First Monday 17, 3 (2012).
[2] WA FU, K., HONG CHAN, C., AND CHAU, M. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. IEEE Internet Computing 17, 3 (2013).
[3] ZHU, T., PHIPPS, D., PRIDGEN, A., CRANDALL, J. R., AND WALLACH, D. S. The velocity of censorship: High-fidelity detection of microblog post deletions. In Proc. of USENIX Security (2013).

Impact of Censorship

- Information dissemination
  - Extensively studied on Twitter
  - Missing: adversarial forces
- Chilling effect?
  - Dissuade people from discussing about censored topics
- Do users adapt?
  - Change linguistic conventions or behaviors

Outline

- Motivation
- Methodology of data gathering
  - Data gathering: API or DIY
  - Crawler design and data collection
  - Validation of the method
- Topics analysis
- Word usage on Weibo

Sina Weibo Overview

- Launched in August 2009
- By December 2012, ≈ 500 million users
- Over 4.6 million users and 100 million tweets daily

Profile image and stats

Username

A retweet with original tweet embedded

An original tweet

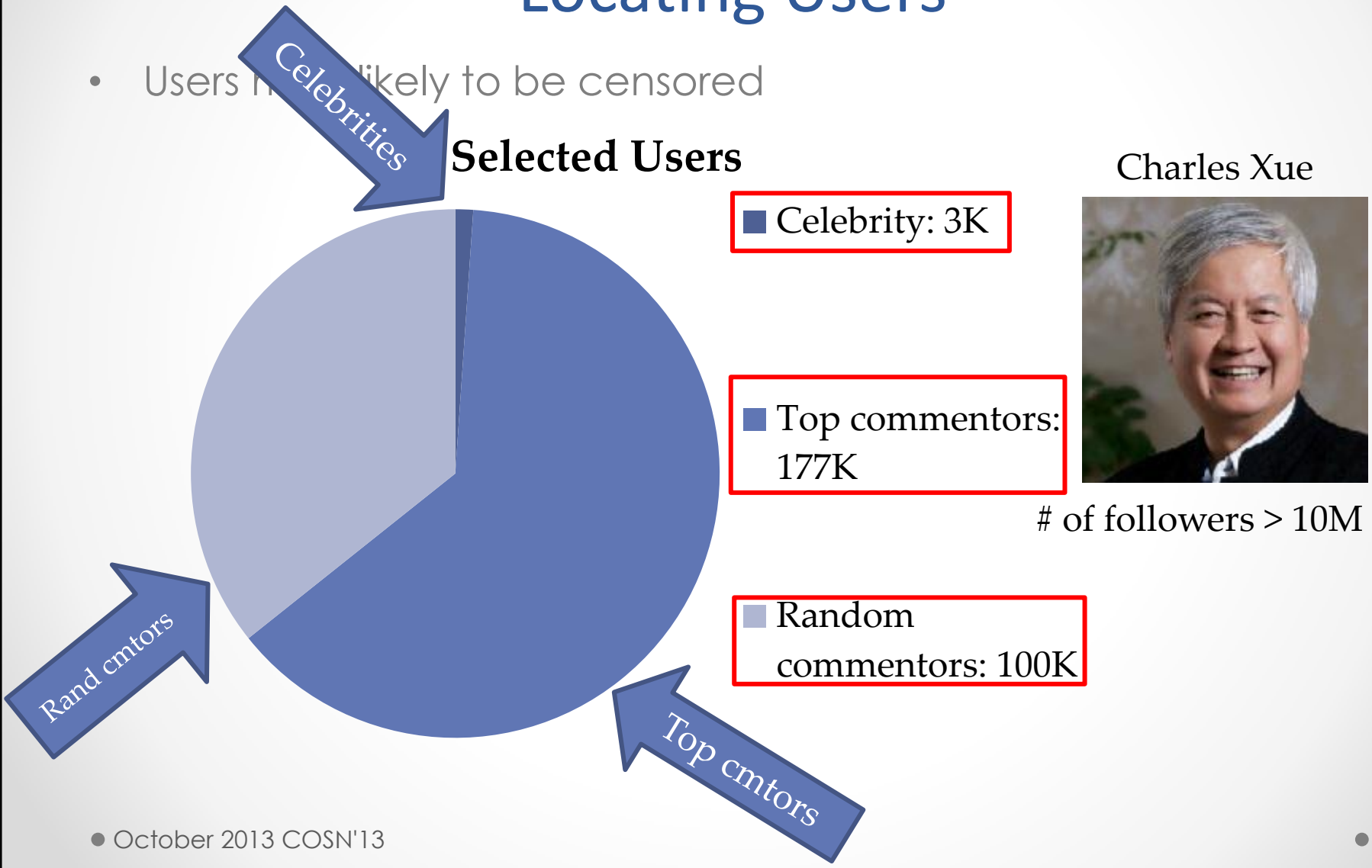
Comments

Data Gathering Challenges

- Unlike Twitter, Weibo removes data
- Censorship happens in minutes
  - Frequently revisit users to keep track of censorship
- Strict rate-limit on requests

Locating Users

- Users ... likely to be censored

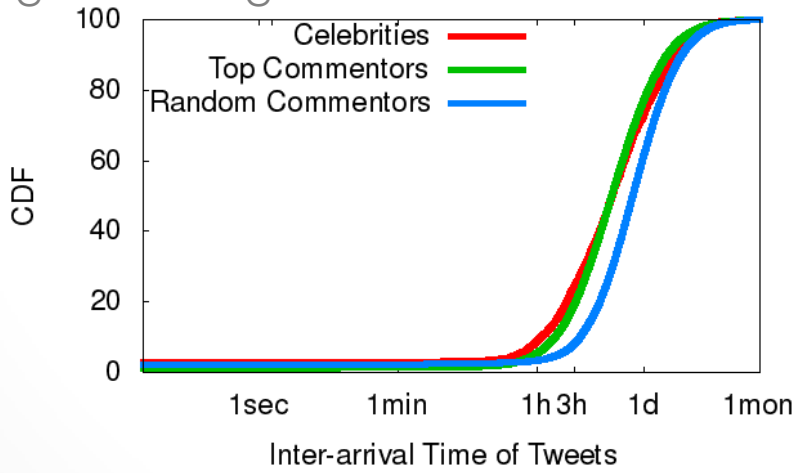


Charles Xue

# of followers > 10M

Crawler Design

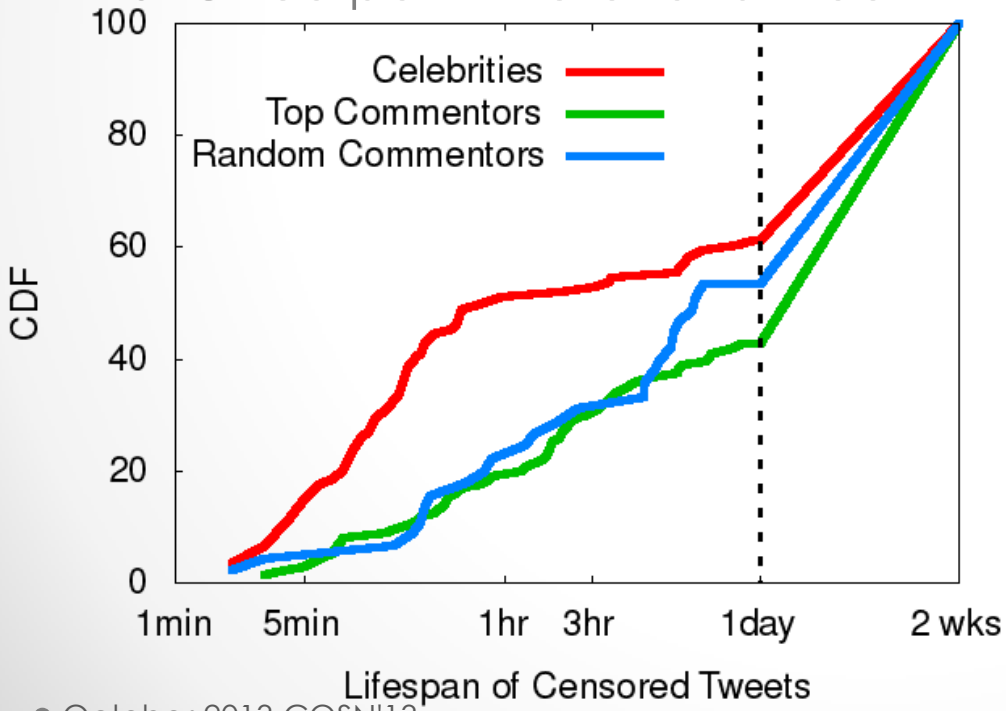
- Design challenges
  - Different users tweet at different speed
  - Censorship happens shortly after tweets are posted
- Solution
  - Crawl the users at different speed
  - Revisit again and again to check censored tweets



Three buckets: one hour (5K), three hours (22K), daily (253K); crawled active users more often

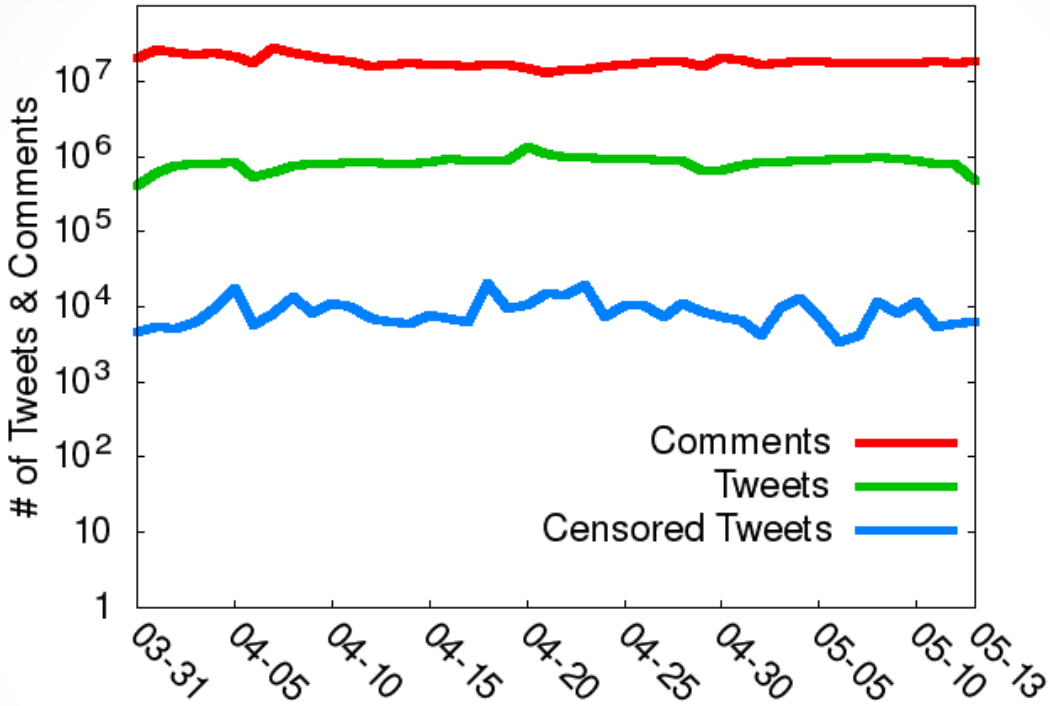
Validation

- Crawl in hours but censorship happens in minutes
- Intensive crawling
  - 500 random users from each group
  - Once per minute for a week



1. Gathers > 50% censored tweets
2. > 99% of content on Weibo is comments and comments are never censored

Daily Activity on Weibo



- 830K tweets; 18M comments; 9K censored tweets
- 1% of tweets are censored
  - Estimates from prior work: 0.01% --16%

Outline

- Motivation
- Methodology of data gathering
- Topics analysis
  - Locating trending topics
  - Analysis of trending topics
- Word usage on Weibo

Key Research Question

- What is the impact of censorship on discourse on Sina Weibo?
  - Impact of censorship in trending topics
  - Censored topics vs. uncensored topics

Locating Trending Topics

- Topic Extraction
  - Latent Dirichlet Allocation (LDA)
  - LDA output: 1)topics; 2) words in each topic

- Challenges
  - Chinese does not include breaks between words
    - Solution: word segmentation using OpenCLAS
  - Many new words on Weibo
    - OpenCLAS has a small dictionary (104K)
    - Solution: Use the Sogou Pinyin Dictionary (including words from social media)
    - Sogou pinyin: crowd-sourced and keep updating

Topic Validation

- Output of LDA: 300 topics
  - 37 trending topics corresponded to real-world events;
- Locating trending topics

General	Real-world events	Censored
Economics	Boston marathon bombing	Lushan secretary
Food	Ya'an earthquake	Beijing taxi driver
English learning	Wechat charge	Bird flu

- Validation
  - All 11 trending topics listed on external sources

High-level Overview

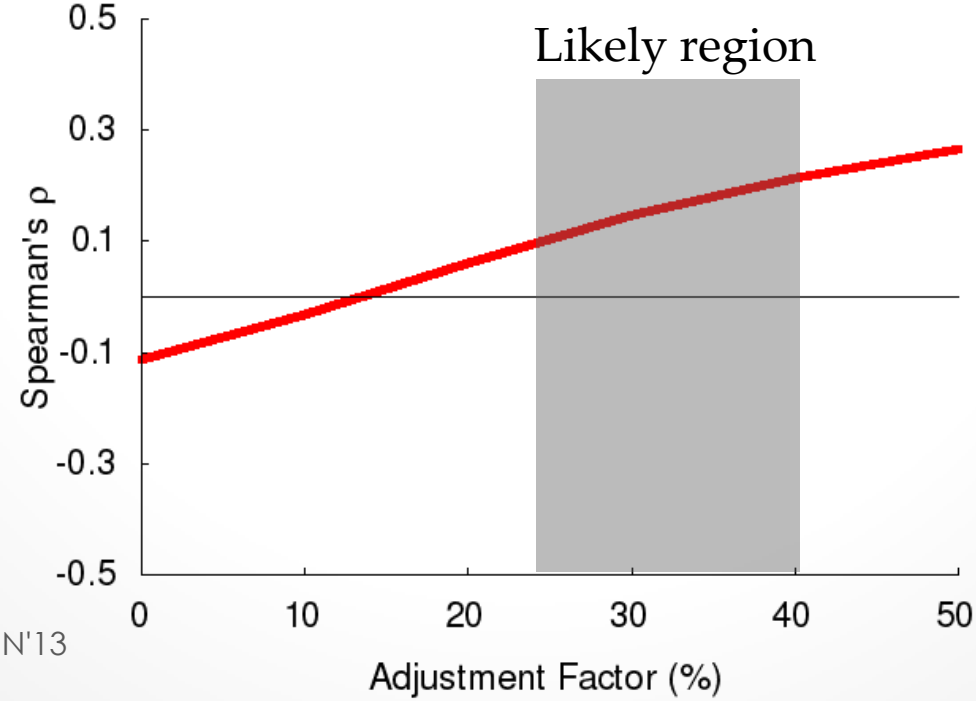
- 37 real-world events

Characteristics	Range	Average
Lifespan(days)	2 ~ 14	4.6
# of tweets	394 ~ 108K	19K
# of comments	538 ~ 3.1M	635K
% of censored tweets	0% ~ 82%	1.7%

- Most topics are not censored
  - 5 topics experienced >5% of censorship

Impact of Censorship

- Does censorship have impact on user behavior?
- Correlation between censorship and # of tweets per user
  - Spearman's ρ = -0.1
- But our crawler missed some censored tweets
  - Adjustment factor = % of missing tweets



Key Research Question

- What is the impact of censorship on discourse on Sina Weibo?
- **Ans: We observe there is no chilling effect. Instead, censored topics see more active users tweeting more frequently.**

Outline

- Motivation
- Methodology of data gathering
- Topics analysis
- Word usage on Weibo
  - Locating morphs
  - Morph usage and censorship
  - Dynamics of Morph usage

Key Research Question

- Do Weibo users adapt in order to avoid censorship?
  - Keyword-based censorship leads to linguistic adaptation
  - To analyze adaptation, we focus on morphs

Locating Morphs

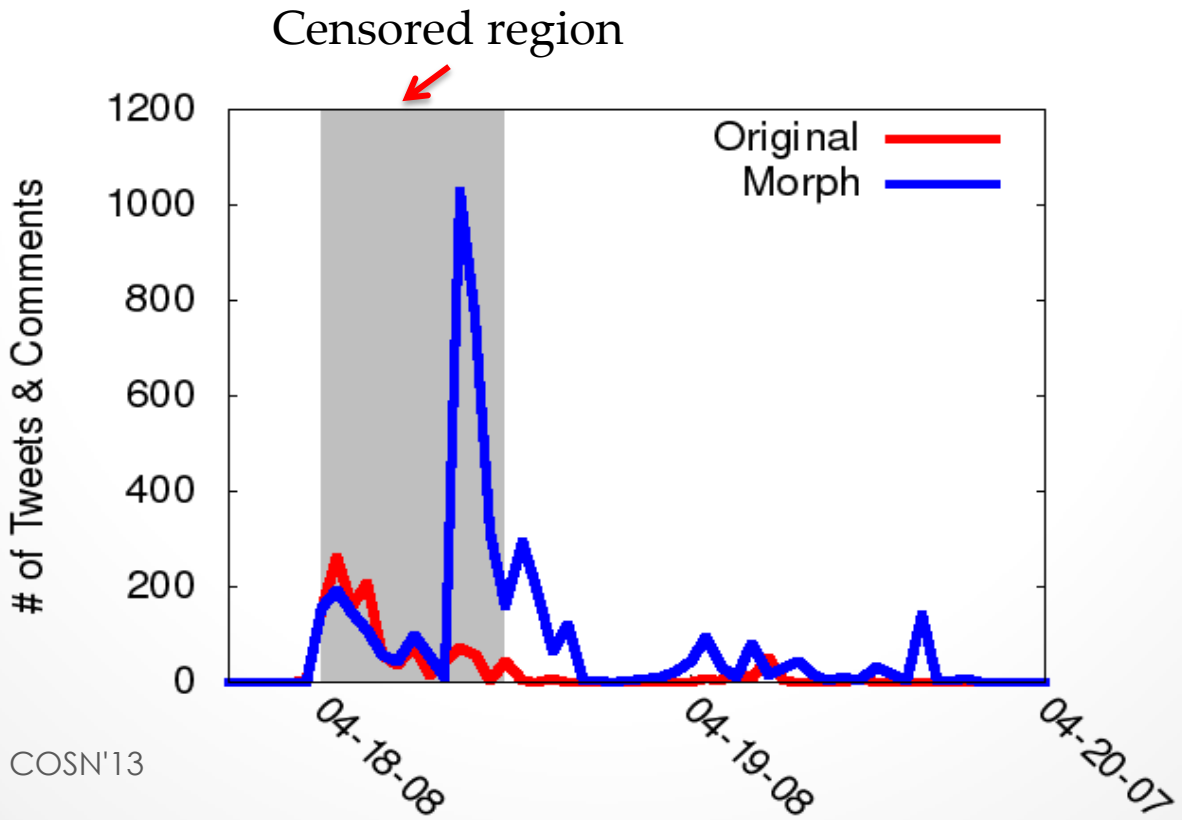
- What is a morph?

Type	Original	Literal Meaning	Morph	Literal Meaning
Homophone	和谐	Harmony	河蟹	River crab
Pinyin Initial	美眉	Pretty eyebrow	MM	N/A
Neologism	囧	Brightness	囧	Embarrassed
Anglicization	京温	Jingwen Mall	京wen	Jingwen Mall
Homograph	翻墙	Over the GFW	番羽土啬	N/A

- Identify novel morphs
  - A morph from topic A used > 100 times prior to the start date of A was not invented during A, and is not a novel morph

Morph Usage Over Time

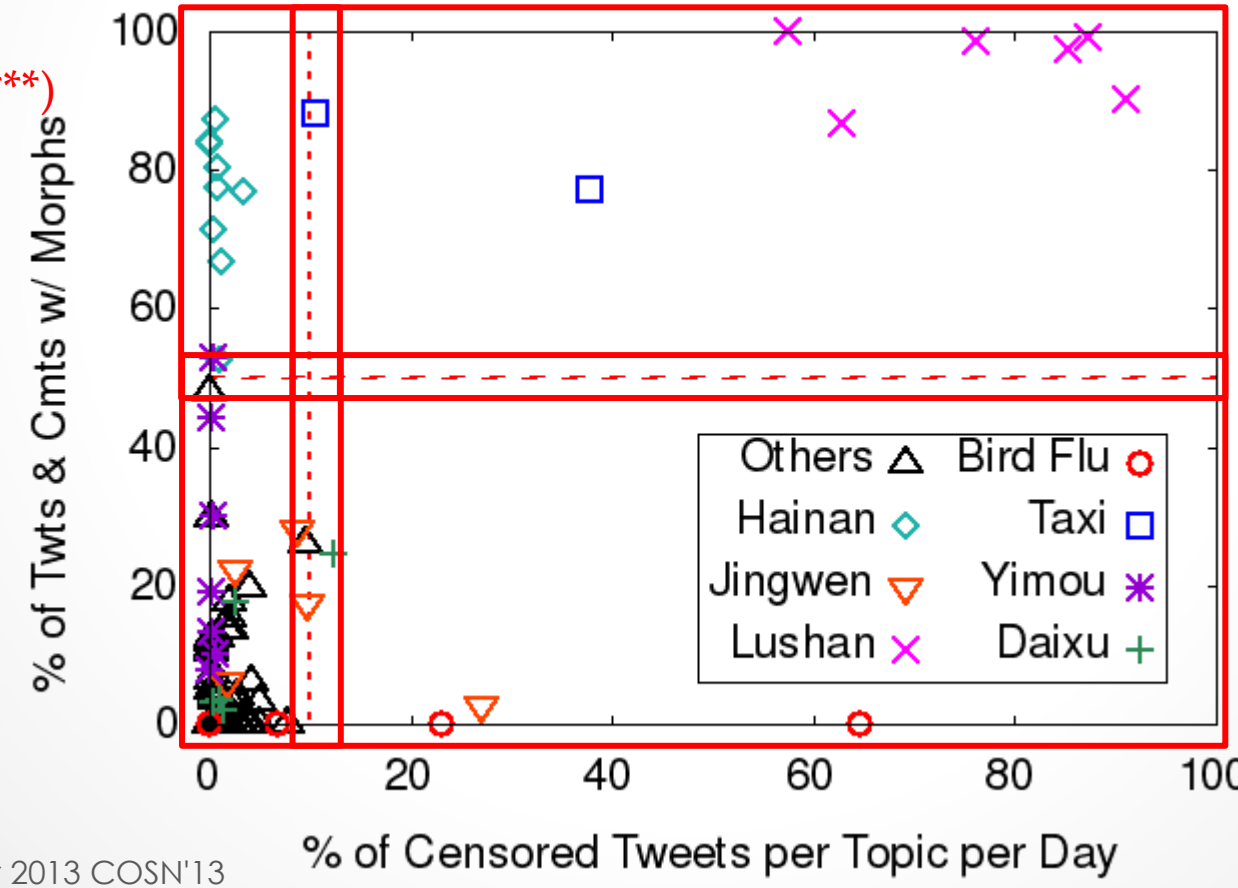
Topic	Story	Original	Morphs
Taxi	A rumor involving a taxi driver to meet President Xi.	郭立新(the driver's name)	郭师傅(Shifu Guo)，北京的哥(Beijing Taxi Driver)



Morph Usage and Censorship

- Is there a relationship between censorship and morph usage?

绿茶婊
(Green Tea ***)
葫芦娃
(Calabash Brothers)



表印哥
(Bro Watchprint)
北京的哥
(Beijing Taxi Driver)

京wen
(Jingwen Mall)
戴某
(Dai XXX)

Key Research Question

- Do Weibo users adapt in order to avoid censorship?
- **Ans: We observe that as censorship becomes prominent, the morph usage becomes popular.**
- What are the dynamics of morph generation?
- **Please refer to the paper, or we can discuss after the talk.**

Summary

- Study the impact of censorship on discourse and word choice on Weibo
- Crawled 280K users on an hourly basis for 44 days, gathering 839M tweets and comments
- There is positive relationship between censorship and user engagement
- Weibo users tend to introduce morphs to heavily censored topics
- Updating information dissemination models that take adversarial forces into account

Questions?