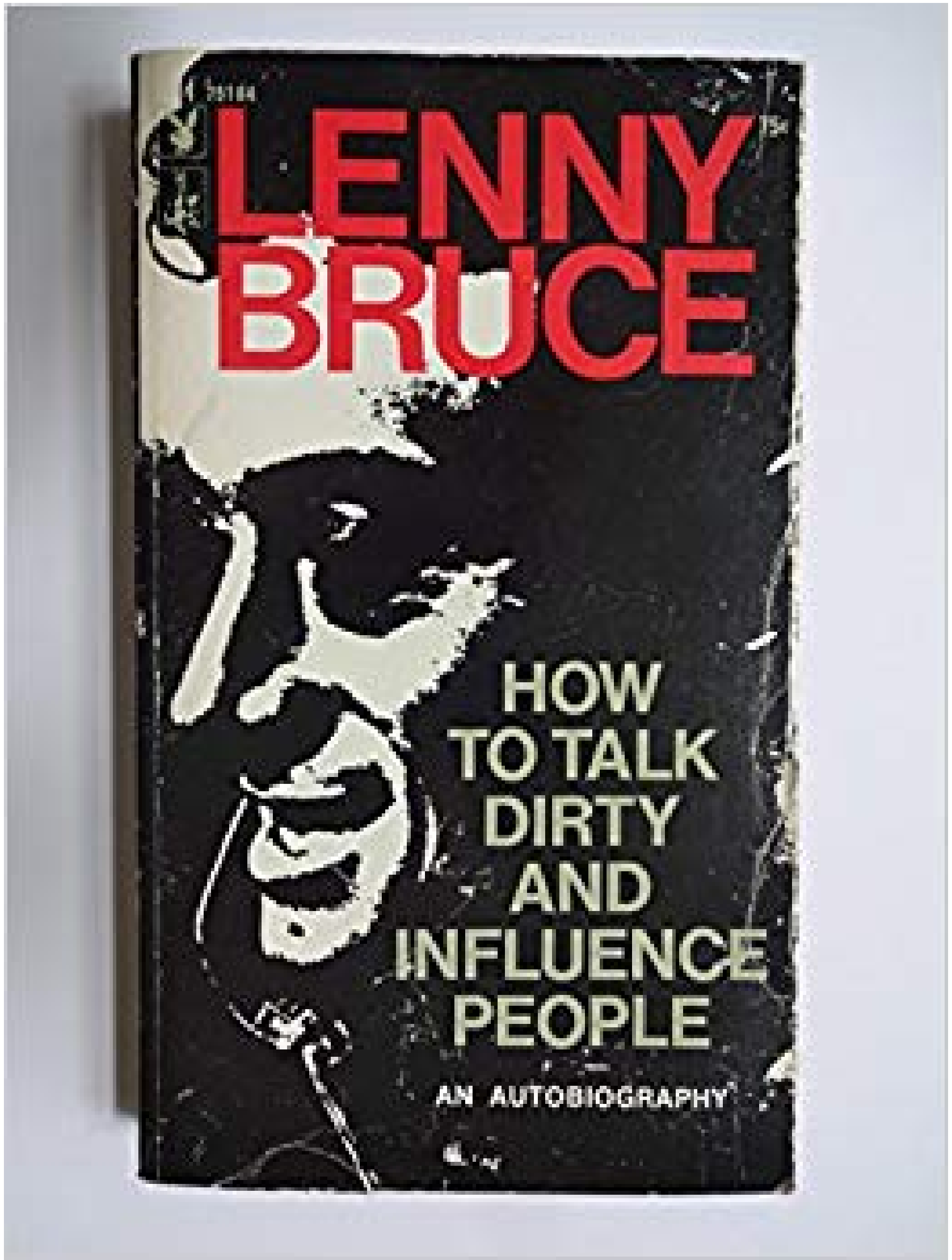
How to talk dirty and influence machines

Jedidiah Crandall
Univ. of New Mexico
crandall@cs.unm.edu



If you can’t say fuck, you can’t say fuck the government.

--Lenny Bruce

In terms of NLP research, the challenges and opportunities presented by Internet Freedom problems are *qualitatively* different from traditional NLP problems.

Outline

- What would *you* do with 100,000 dirty words (mostly in Chinese)?
- Can we understand Martian language?
- What terrible innovations can we look forward to?

Outline

- What would *you* do with 100,000 dirty words (mostly in Chinese)?
- Can we understand Martian language?
- What terrible innovations can we look forward to?

Jed’s Chinese teacher:
“Why do students always want to learn the bad words?”

My 12-year (and counting) journey to learn all the worst words in Chinese...

ConceptDoppler: A Weather Tracker for Internet Censorship

Jedidiah R. Crandall
Univ. of New Mexico
crandall@cs.unm.edu

Daniel Zinn
Univ. of California at Davis
zinn@cs.ucdavis.edu

Michael Byrd
Univ. of California at Davis
byrd@cs.ucdavis.edu

Earl Barr
Univ. of California at Davis
barr@cs.ucdavis.edu

Rich East
Independent Researcher
richeast19@gmail.com

ABSTRACT

The text of this paper has passed across many Internet routers on its way to the reader, but some routers will not pass it along unfettered because of censored words it contains. We present two sets of results: 1) Internet measurements of keyword filtering by the Great "Firewall" of China (GFC); and 2) initial results of using latent semantic analysis as an efficient way to reproduce a blacklist of censored words via probing.

Our Internet measurements suggest that the GFC's keyword filtering is more a panopticon than a firewall, *i.e.*, it need not block every illicit word, but only enough to promote self-censorship. China's largest ISP, ChinaNET, performed 83.3% of all filtering of our probes, and 99.1% of all filtering that occurred at the first hop past the Chinese border. Filtering occurred beyond the third hop for 11.8% of our probes, and there were sometimes as many as 13 hops past the border to a filtering router. Approximately 28.3% of the Chinese hosts we sent probes to were reachable along paths that were not filtered at all. While more tests are needed to provide a definitive picture of the GFC's implementation, our results disprove the notion that GFC keyword filtering is a firewall strictly at the border of China's Internet.

While evading a firewall a single time defeats its purpose, it would be necessary to evade a panopticon almost every time. Thus,

> *Everybody talks about the weather but nobody does anything about it.*
>
> Charles Dudley Warner (1829–1900)

Categories and Subject Descriptors

K.4.m [**Computers and Society**]: Miscellaneous

General Terms

Experimentation, human factors, legal aspects, measurement, security

Keywords

LSA, latent semantic analysis, latent semantic indexing, firewall ruleset discovery, Internet censorship, Great Firewall of China, Internet measurement, panopticon, ConceptDoppler, keyword filtering, blacklist

天津租界 [编辑]

维基百科，自由的百科全书

天津租界，是1860年至1945年期间，英国、法国、美国、德国、意大利、俄国、日本、奥匈帝国和比利时等国通过签订条约和协议在中国天津老城东南方向海河两岸相继设立的租借地，拥有行政自治权和治外法权。因先后有九国国家在天津划定租界，因此又称**九国租界**[1]。

1860年，英国首先在天津设立租界，最高峰时有9个国家在天津设立租界。同时，天津也是中国最早收回租界的城市之一。1945年，中华民国政府在对日战争胜利后，正式收回天津的最后两个租界——英租界和法租界，标志着天津租界历史的结束[2]，而天津英租界是九国租界中存在时间最长的租界——长达85年[3]。

天津租界开辟之后，租界的建设对天津的城市建设起到了促进和示范的作用，西方文化对天津各个方面的影响不断渗透。西洋文化的涌入和租界建筑的建设打破了天津原有的中国北方传统建筑城市风貌[4]，租界文化还通过与教会有关的教育、报刊杂志等影响着天津人的文化生活。由租界教会创办的学校、医院、报刊和杂志，代表着不同国籍、不同政治利益，某种程度上也意味着文化殖民。但是它们同时又代表着不同的文化，呈现出多元性、多样性的特点，客观上促进了天津文化的多元融合与发展，在近代天津迈向现代大都会的初期，发挥了重要的作用[5]。

天津租界是西洋文化和中国传统及地域文化承载体，是天津多元文化的重要组成部分，曾作为中国近

Examples found by ConceptDoppler

- 专政 – Dictatorship (party)
- 藏独 – Tibet independence movement
- 色情电影 – Erotic movies
- 方舟子 – Fang Zhouzi (Neil deGrasse Tyson of China)
- 转化率 – Conversion rate

Work-in-progress: Automated Named Entity Extraction for Tracking Censorship of Current Events

Antonio M. Espinoza
Computer Science Department
University of New Mexico
amajest@cs.unm.edu

Jedidiah R. Crandall
Computer Science Department
University of New Mexico
crandall@cs.unm.edu

Abstract

Tracking Internet censorship is challenging because what content the censors target can change daily, even hourly, with current events. The process must be automated because of the large amount of data that needs to be processed. Our focus in this paper is on automated probing of keyword-based Internet censorship, where natural language processing techniques are used to generate keywords to probe for censorship with. In this paper

what kinds of activities are targeted by the censors. This implies automated probing that is broad and carried out over a long period of time, because censorship within a single country can vary from province to province, company to company, and technology to technology and what content is targeted can change daily, even hourly.

1.1 Related work

Our focus in this paper is on keyword-based Internet cen-

Some questions to keep in mind...

- Just how well does Chinese text segmentation work?
  - “There’s a library you can download...”
- How big is any given keyword blacklist?
  - More generally, who makes the lists and what’s on their mind?

JR01-2008

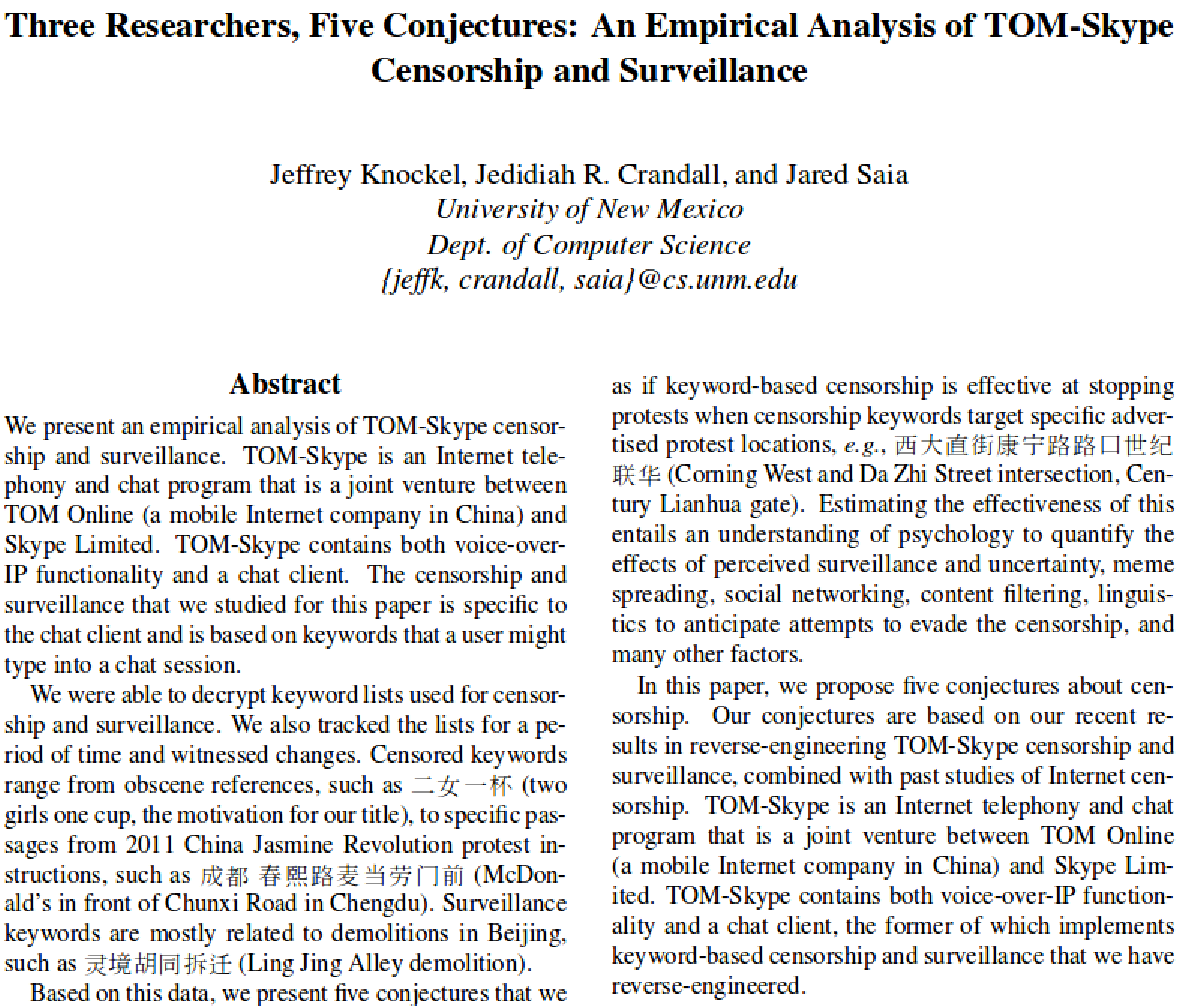
Information Warfare Monitor
ONI Asia

Joint Report

BREACHING TRUST:

An analysis of surveillance and security practices on China's TOM-Skype platform

Nart Villeneuve, Psiphon Fellow, the Citizen Lab



Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance

Jeffrey Knockel, Jedidiah R. Crandall, and Jared Saia
University of New Mexico
Dept. of Computer Science
{jeffk, crandall, saia}@cs.unm.edu

Abstract

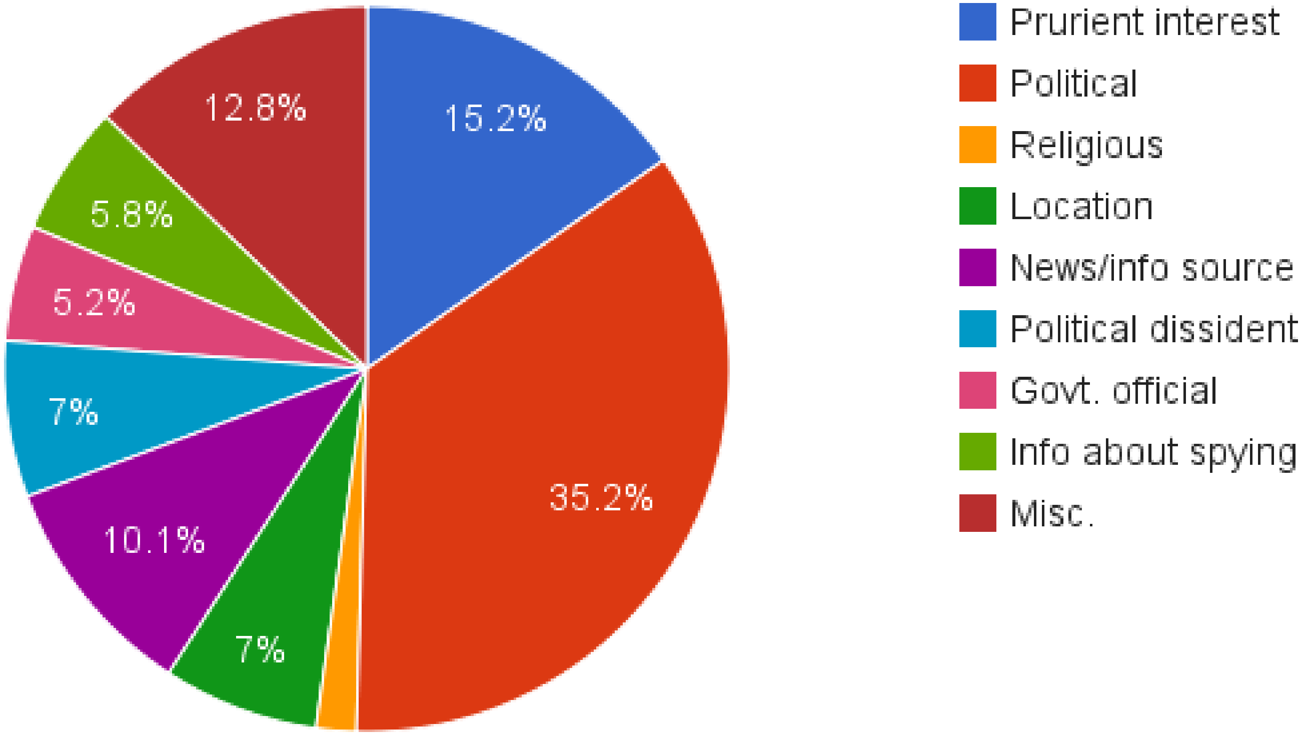
We present an empirical analysis of TOM-Skype censorship and surveillance. TOM-Skype is an Internet telephony and chat program that is a joint venture between TOM Online (a mobile Internet company in China) and Skype Limited. TOM-Skype contains both voice-over-IP functionality and a chat client. The censorship and surveillance that we studied for this paper is specific to the chat client and is based on keywords that a user might type into a chat session.

We were able to decrypt keyword lists used for censorship and surveillance. We also tracked the lists for a period of time and witnessed changes. Censored keywords range from obscene references, such as 二女一杯 (two girls one cup, the motivation for our title), to specific passages from 2011 China Jasmine Revolution protest instructions, such as 成都 春熙路麦当劳门前 (McDonald's in front of Chunxi Road in Chengdu). Surveillance keywords are mostly related to demolitions in Beijing, such as 灵境胡同拆迁 (Ling Jing Alley demolition).

Based on this data, we present five conjectures that we as if keyword-based censorship is effective at stopping protests when censorship keywords target specific advertised protest locations, *e.g.*, 西大直街康宁路路口世纪联华 (Corning West and Da Zhi Street intersection, Century Lianhua gate). Estimating the effectiveness of this entails an understanding of psychology to quantify the effects of perceived surveillance and uncertainty, meme spreading, social networking, content filtering, linguistics to anticipate attempts to evade the censorship, and many other factors.

In this paper, we propose five conjectures about censorship. Our conjectures are based on our recent results in reverse-engineering TOM-Skype censorship and surveillance, combined with past studies of Internet censorship. TOM-Skype is an Internet telephony and chat program that is a joint venture between TOM Online (a mobile Internet company in China) and Skype Limited. TOM-Skype contains both voice-over-IP functionality and a chat client, the former of which implements keyword-based censorship and surveillance that we have reverse-engineered.

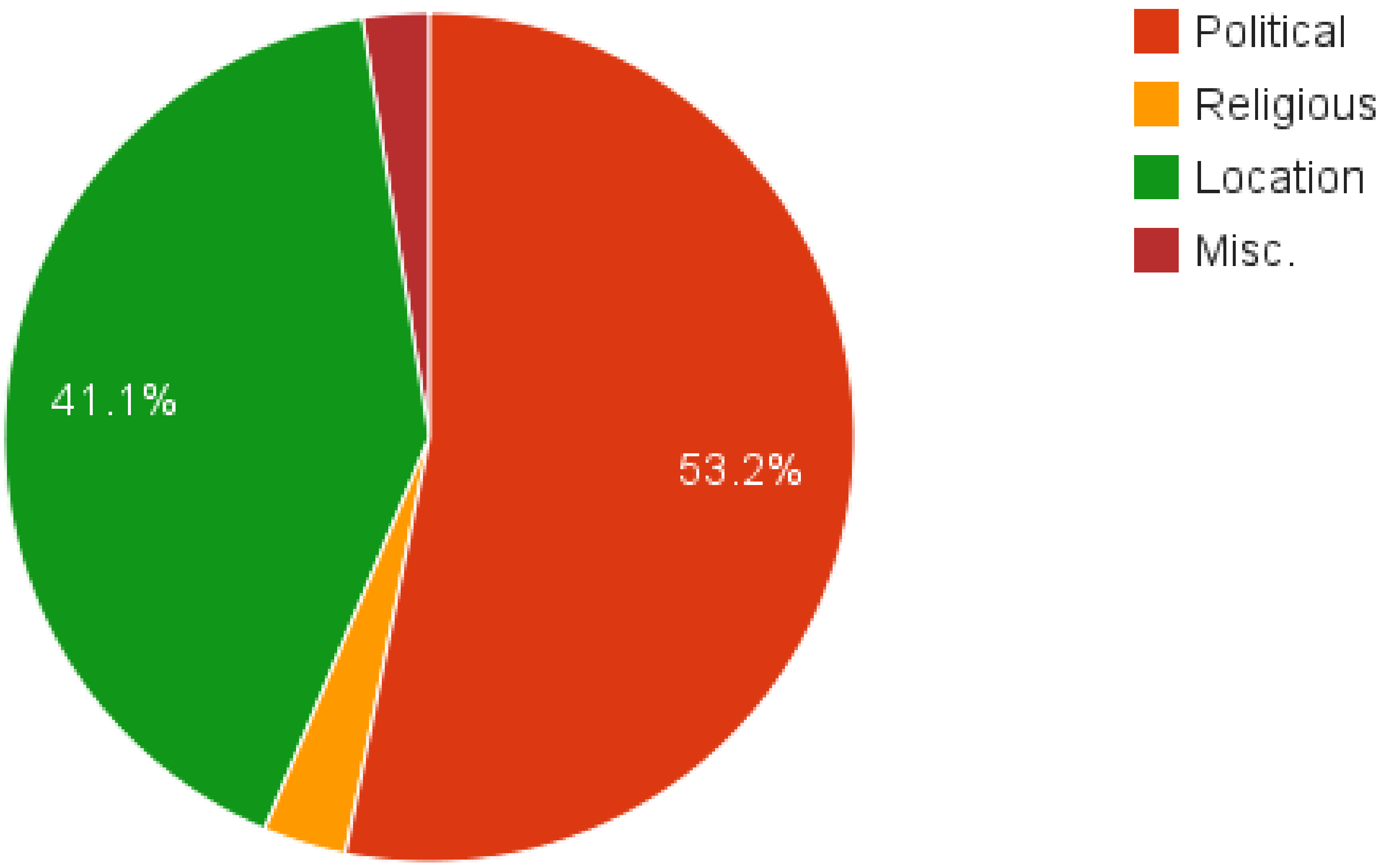
Censorship keywords



Example keywords

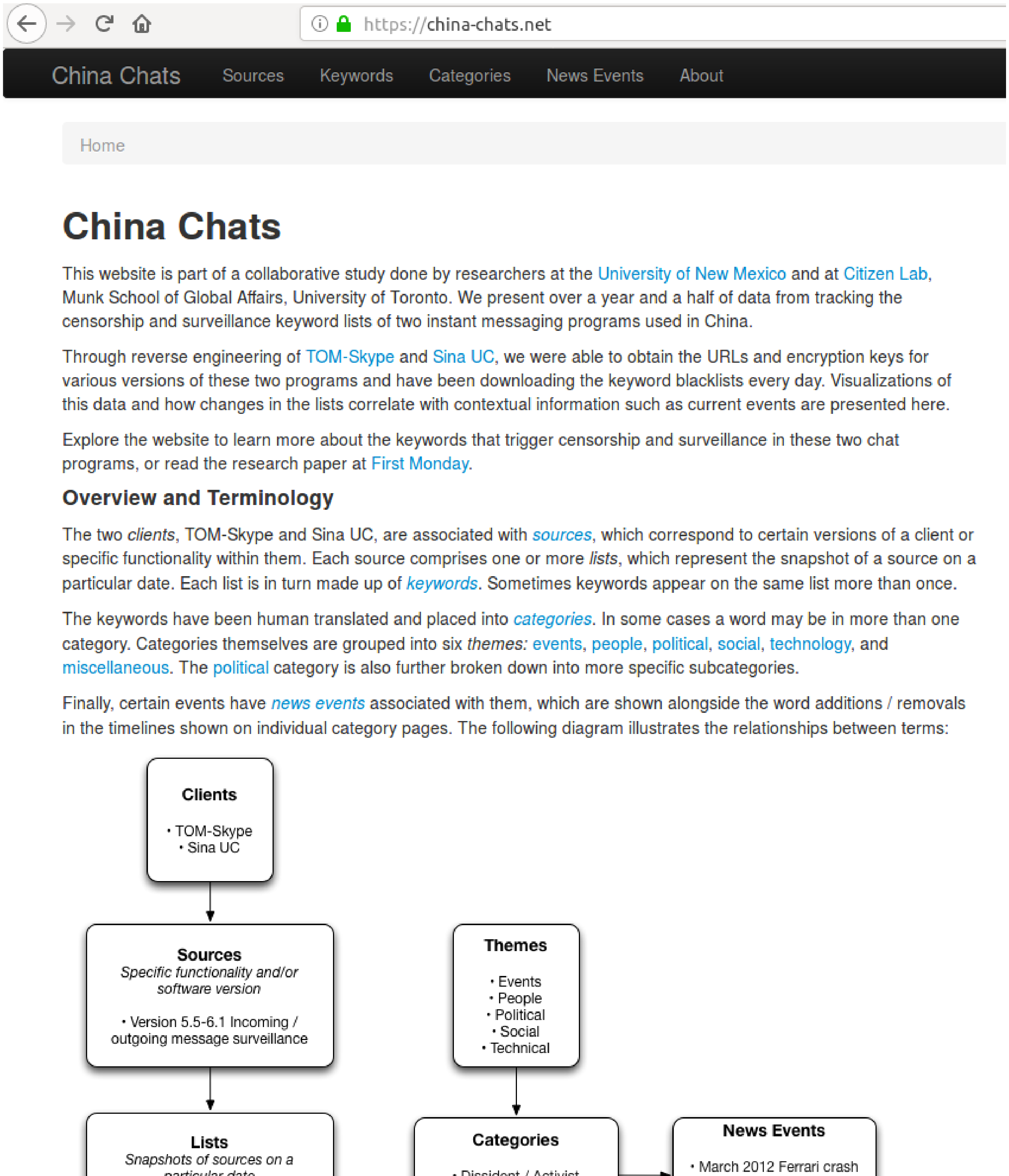
- fuck ("fuck")
- 代考网 ("Test replacement network")
- 江贼民 ("Jiang Zei Min", sounds like Jiang Zemin's name, with "traitor" and "citizens".)
- 二女一杯 ("Two girls one cup")
- 成都春熙路麦当劳门前 ("McDonald's in front of Chunxi Road in Chengdu")

Surveillance-only keywords



Example keywords

- 府右街 ("Fuyou Jie")
- 府佑街 ("Fuyou Jie", note that 右 here is replaced with 佑 , which has the same pinyin.)



China Chats

This website is part of a collaborative study done by researchers at the University of New Mexico and at Citizen Lab, Munk School of Global Affairs, University of Toronto. We present over a year and a half of data from tracking the censorship and surveillance keyword lists of two instant messaging programs used in China.

Through reverse engineering of TOM-Skype and Sina UC, we were able to obtain the URLs and encryption keys for various versions of these two programs and have been downloading the keyword blacklists every day. Visualizations of this data and how changes in the lists correlate with contextual information such as current events are presented here.

Explore the website to learn more about the keywords that trigger censorship and surveillance in these two chat programs, or read the research paper at First Monday.

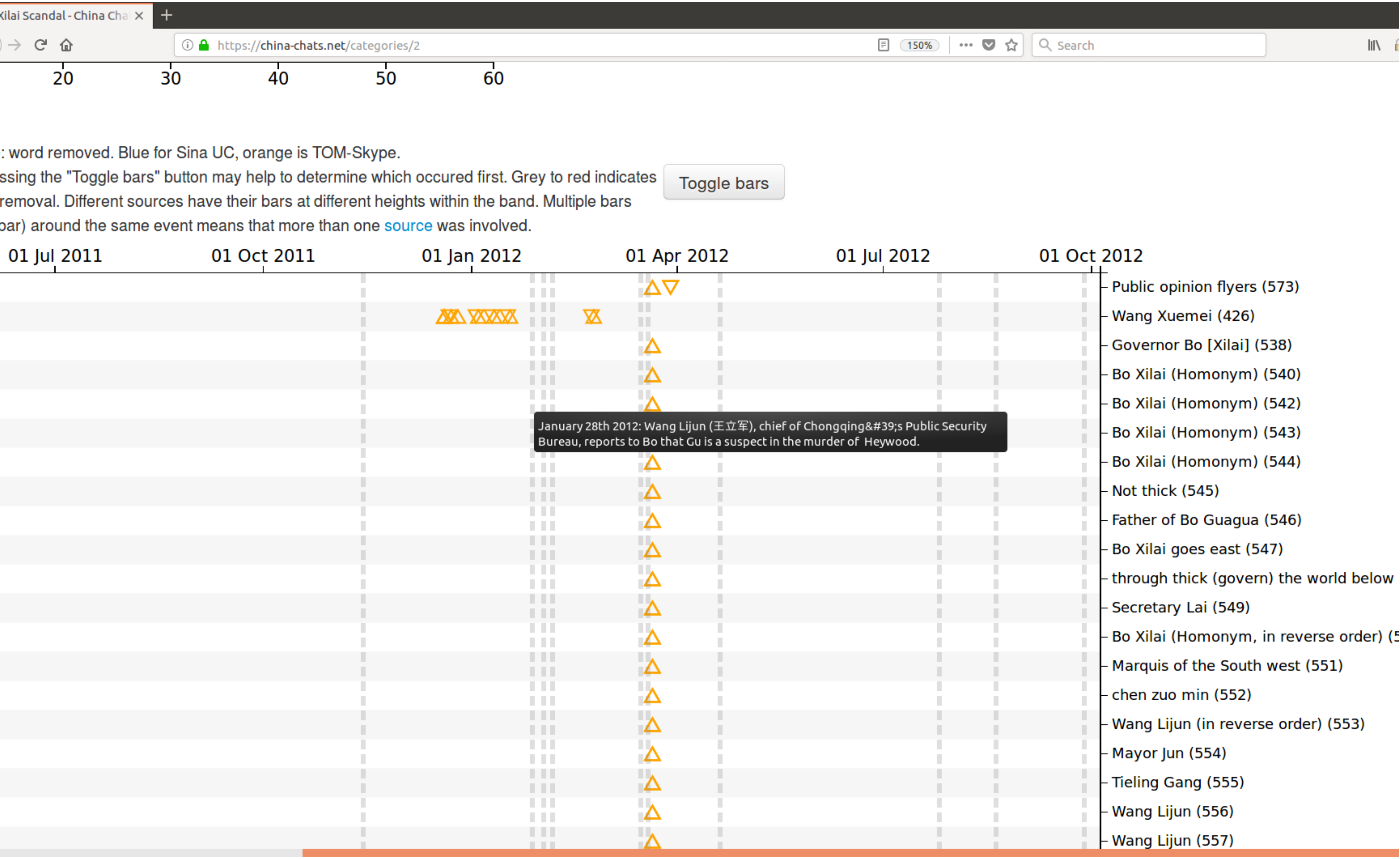
Overview and Terminology

The two *clients*, TOM-Skype and Sina UC, are associated with *sources*, which correspond to certain versions of a client or specific functionality within them. Each source comprises one or more *lists*, which represent the snapshot of a source on a particular date. Each list is in turn made up of *keywords*. Sometimes keywords appear on the same list more than once.

The keywords have been human translated and placed into *categories*. In some cases a word may be in more than one category. Categories themselves are grouped into six *themes:* events, people, political, social, technology, and miscellaneous. The political category is also further broken down into more specific subcategories.

Finally, certain events have *news events* associated with them, which are shown alongside the word additions / removals in the timelines shown on individual category pages. The following diagram illustrates the relationships between terms:

Clients
• TOM-Skype
• Sina UC
Sources
Specific functionality and/or software version
• Version 5.5-6.1 Incoming / outgoing message surveillance
Themes
• Events
• People
• Political
• Social
• Technical
Lists
Snapshots of sources on a particular date
Categories
• Dissident / Activist
News Events
• March 2012 Ferrari crash

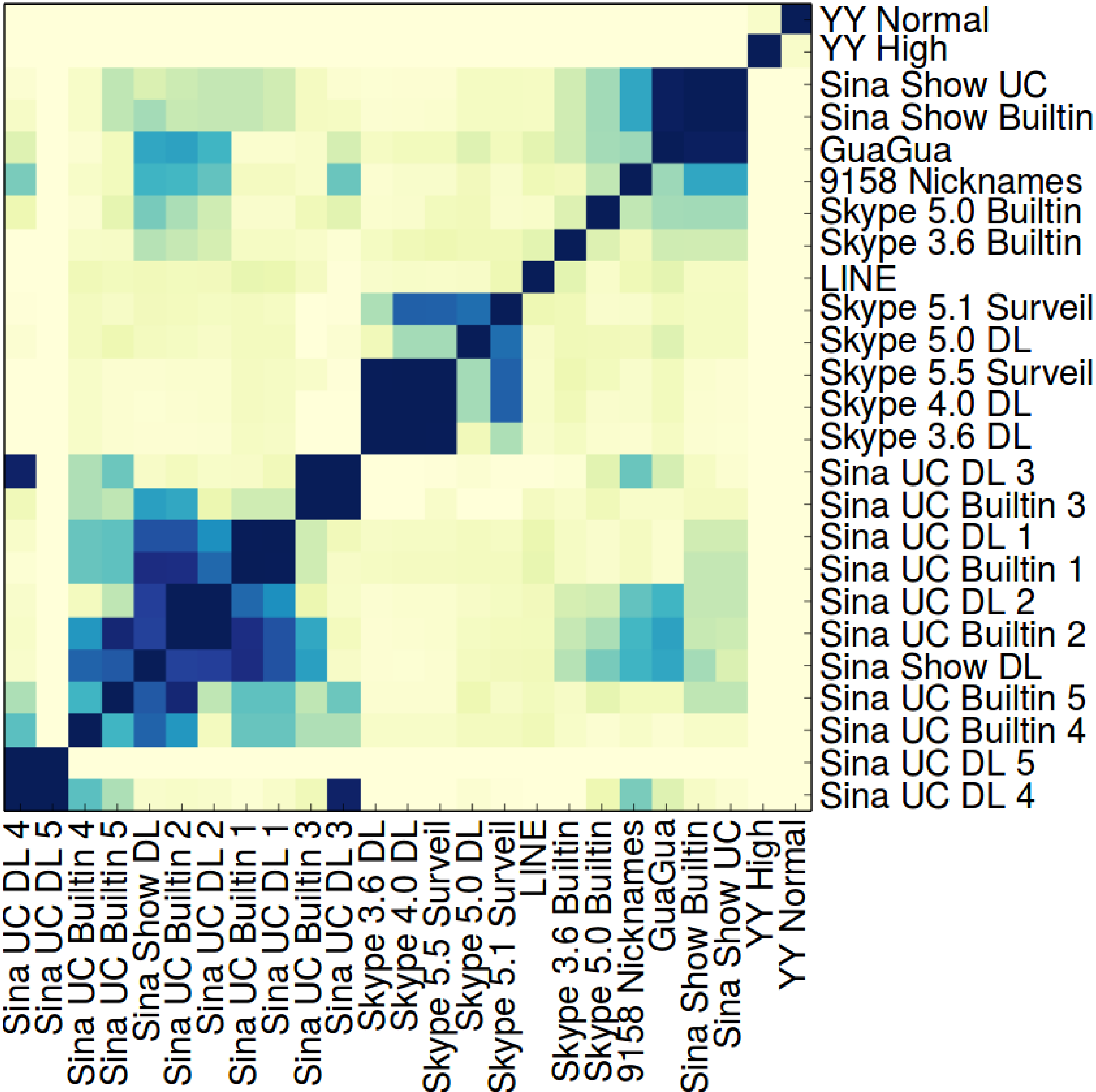


Xilai Scandal - China Cha
https://china-chats.net/categories/2
150%
Search
20
30
40
50
60

: word removed. Blue for Sina UC, orange is TOM-Skype.

ssing the "Toggle bars" button may help to determine which occured first. Grey to red indicates removal. Different sources have their bars at different heights within the band. Multiple bars oar) around the same event means that more than one source was involved.

January 28th 2012: Wang Lijun (王立军), chief of Chongqing&#39;s Public Security Bureau, reports to Bo that Gu is a suspect in the murder of Heywood.


YY Normal
YY High
Sina Show UC
Sina Show Builtin
GuaGua
9158 Nicknames
Skype 5.0 Builtin
Skype 3.6 Builtin
LINE
Skype 5.1 Surveil
Skype 5.0 DL
Skype 5.5 Surveil
Skype 4.0 DL
Skype 3.6 DL
Sina UC DL 3
Sina UC Builtin 3
Sina UC DL 1
Sina UC Builtin 1
Sina UC DL 2
Sina UC Builtin 2
Sina Show DL
Sina UC Builtin 5
Sina UC Builtin 4
Sina UC DL 5
Sina UC DL 4
Sina UC DL 4
Sina UC DL 5
Sina UC Builtin 4
Sina UC Builtin 5
Sina Show DL
Sina UC Builtin 2
Sina UC DL 2
Sina UC Builtin 1
Sina UC DL 1
Sina UC Builtin 3
Sina UC DL 3
Skype 3.6 DL
Skype 4.0 DL
Skype 5.5 Surveil
Skype 5.0 DL
Skype 5.1 Surveil
LINE
Skype 3.6 Builtin
Skype 5.0 Builtin
9158 Nicknames
GuaGua
Sina Show Builtin
Sina Show UC
YY High
YY Normal

Could we go deeper?

- “Translations” are largely a manual effort
  - “A little part of my soul dies”
- Lots of ways for words to have meaning
  - Sounds like or looks like another word
  - Characters have meaning, *e.g.*, 点点点人
  - Semantics, *e.g.*, “May 35th”
  - Position in the list
- Jeffrey Knockel’s dissertation data set has over 100K keywords
  - Each is not just a term, but a story
  - Turkic language keywords, financial keywords, *etc.*

Should we go deeper?

- Theories
  - “The Anaconda in the Chandelier” by Perry Link
  - “Collective action potential” from Gary King’s research group
- Simple questions that are still open
  - Is there overlap if you consider concepts instead of bitstrings?
  - What should our categories be?

Outline

- What would *you* do with 100,000 dirty words (mostly in Chinese)?
- Can we understand Martian language?
- What terrible innovations can we look forward to?

The Velocity of Censorship: High-Fidelity Detection of Microblog Post Deletions

Tao Zhu
zhutao777@gmail.com
Independent Researcher

David Phipps
Computer Science
Bowdoin College

Adam Pridgen
Computer Science
Rice University

Jedidiah R. Crandall
Computer Science
University of New Mexico

Dan S. Wallach
Computer Science
Rice University

Abstract

Weibo and other popular Chinese microblogging sites are well known for exercising internal censorship, to comply with Chinese government requirements. This research seeks to quantify the mechanisms of this censorship: how fast and how comprehensively posts are deleted. Our analysis considered 2.38 million posts gathered over roughly two months in 2012, with our attention focused on repeatedly visiting "sensitive" users. This gives us a view of censorship events within minutes of their occurrence, albeit at a cost of our data no longer representing a random sample of the general Weibo population. We also have a larger 470 million post sampling from Weibo's public timeline, taken over a longer time period, that is more representative of a random sample.

terconnected through their social graph and tend to post about sensitive topics. This biases us towards the content posted by these particular users, but enables us to measure with high fidelity the speed of the censorship and discern interesting patterns in censor behaviors.

Sina Weibo (`weibo.com`, referred to in this paper simply as "Weibo") has the most active user community of any microblog site in China [39]. Weibo provides services which are similar to Twitter, with @usernames, #hashtags, reposting, and URL shortening. In February 2012, Weibo had over 300 million users, and about 100 million messages sent daily [3]. Like Twitter in other countries, Weibo plays an important role in the discourse surrounding current events in China. Both professional reporters and amateurs can provide immediate, first-hand

At Rice University, Prof. Dan Wallach saw a UFO

在莱斯大学一位教授看了飞碟

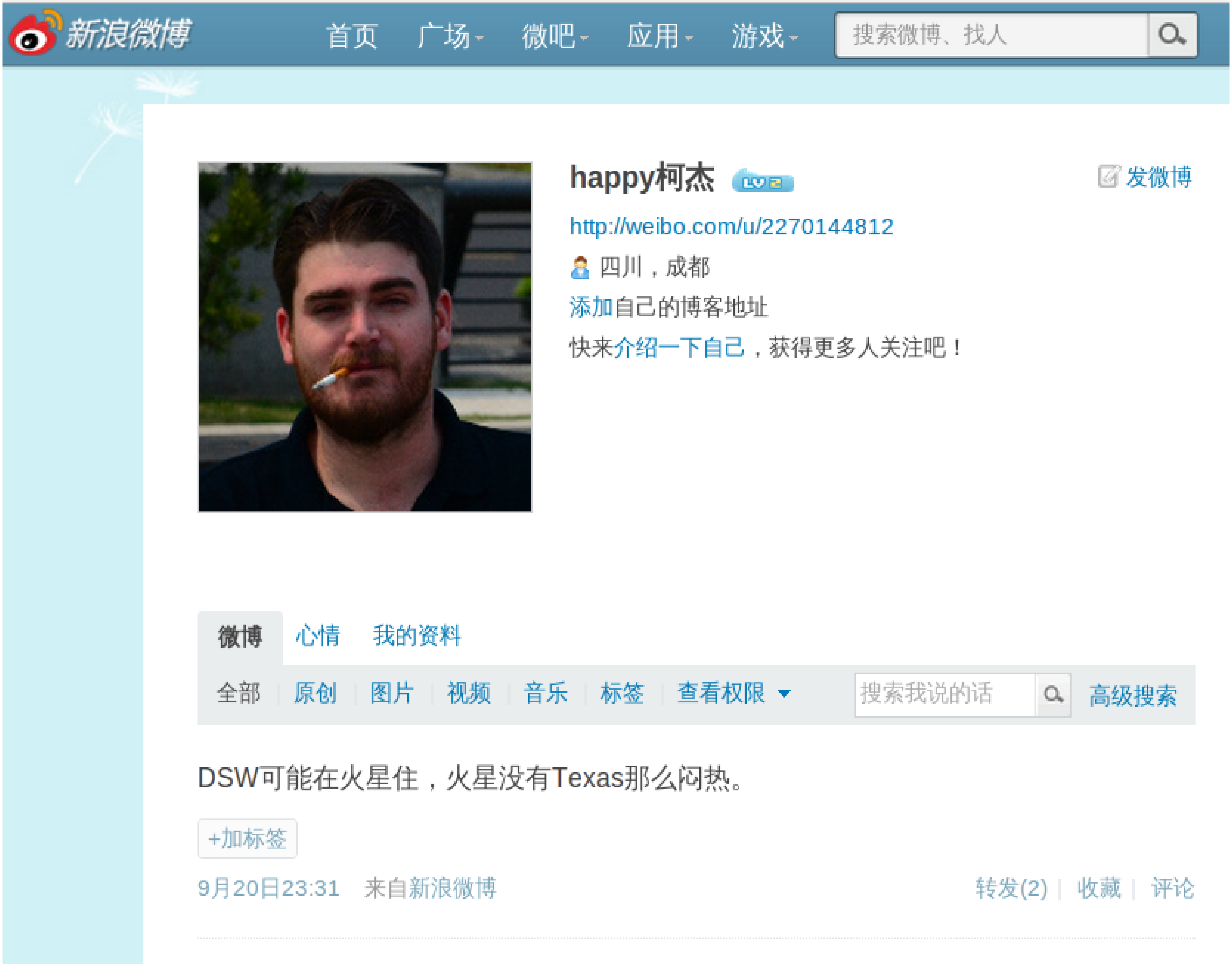
在米饭大学一位叫兽看了 UFO

莱斯大学的 Faculty 说他看了 Flying Saucer

米饭大学发考题说他看了一个飞碟

Dan Wallach 说他看了 UFO

D-Dawg 觉得米饭大学有外星人



happy柯杰 LV2

http://weibo.com/u/2270144812

四川，成都

添加自己的博客地址

快来介绍一下自己，获得更多人关注吧！

DSW可能在火星住，火星没有Texas那么闷热。

9月20日23:31 来自新浪微博

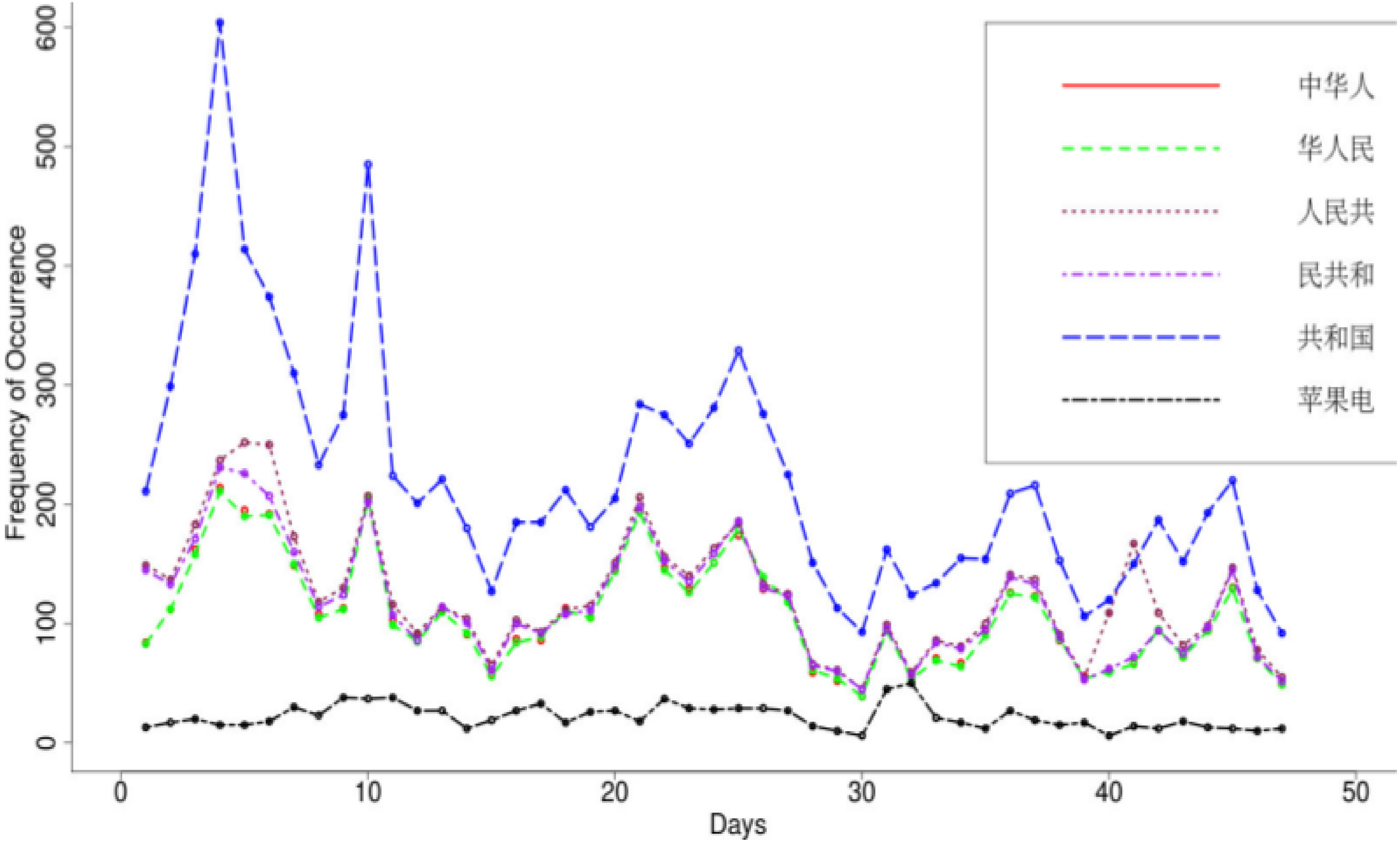


Figure 2: **Trigram trends for 中华人民共和国 on Weibo.**

Pointillism in a nutshell

- TF-IDF on trigrams, documents are periods of time
- Top TF-IDF trigrams are fed into a trigram connection algorithm
  - 在莱斯 + 莱斯大 = 在莱斯大
  - Cosine similarity, breadth-first search

100100_20110804_万为开:d: gram=万为开,
万为开拓团拍电视,
(Wan made a TV program about the first immigrants)
万为开拓团纪念碑被泼红漆,
(The statue was splashed with red oil paint)
万为开拓团纪念碑被5人砸,
(The statue was smashed)
万为开拓团民,
(The first immigrant people)

	Top 1	Top 2	Top 3
8-10	RTL	Freedom of speech[42]	Group sex[40]
8-11	Group sex[40]	Gu Kailai's case[17]	DHBTC[43]
8-12	Tang Hui[46]	Group sex[40]	Kong Qingdong[47]
8-13	DHBTC[43]	RTL[59]	Despise gov.[48]
8-14	Hongkong[49]	Scandal of gov.[50]	Tang Hui[51]
8-15	DHBTC[43]		
8-16	Rumor[53]	Zhou Kehua[52]	Diaoyu Island[54]
8-17	Anti-Japanese[55]	DHBTC[43]	North Korea[56]
8-18	Zhou Kehua[52]	Anti-Japanese[62]	DHBTC[43]
8-19	Anti-Japanese[62]	RTL[59]	Zhou Kehua[52]
8-20	Anti-Japanese[62]	Freedom speech[60]	Zhou Kehua[52]

42	Ren Jianyu was sentenced to re-education through labor (RTL) for two years fc
43	The bodies and parts currently on display in New York are licensed to the Prer Ltd. (DHBTC). DHBTC acquired the bodies indirectly from the Chinese Burea
44	Li Hongna posted a story about herself saying that she was raped by 8 people,
45	The Tiananmen Square protests of 1989, or June Fourth Incident.
46	Tanghui was sentenced to 2 years of RTL for petitioning for the case of her da she was forced to become a prostitute. http://zh.wikipedia.org/zh/唐慧劳教案
47	Kong Qingdong says: "Taxpayers, to go the hell." http://en.wikipedia.org/wiki/I
48	A work of fiction about government corruption.
49	Hongkong college students did not welcome astronauts visiting from China.
50	Municipal administration beat a passerby who took pictures while they were be

Are there more principled approaches?

- Yes
- If you’re Sina, people’s search queries help you build an excellent dictionary for word segmentation
- Roughly a fifth of the Internet’s users are in China, why should researchers have to make up things like “Pointillism” to do Internet freedom research?

Outline

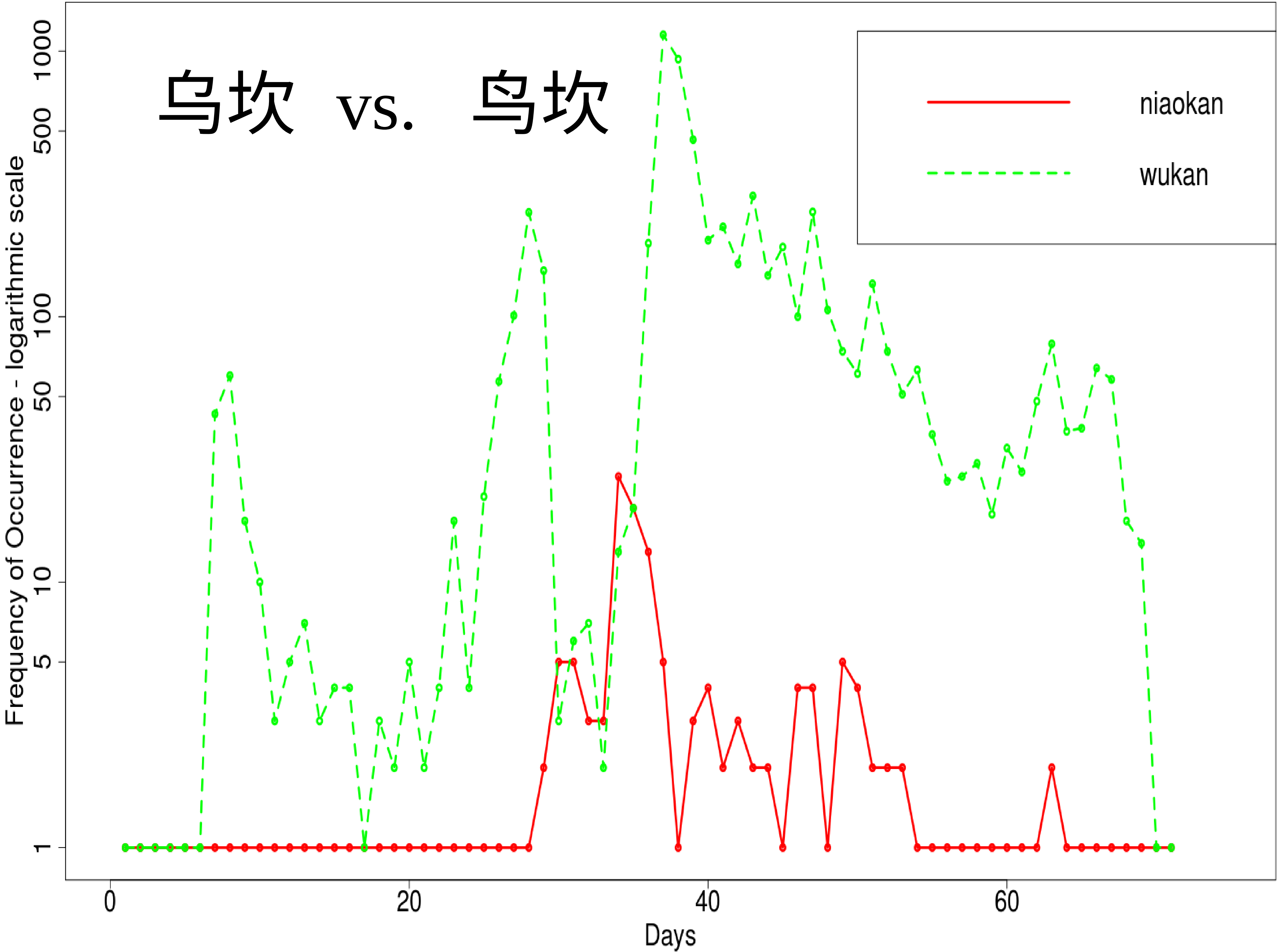
- What would *you* do with 100,000 dirty words (mostly in Chinese)?
- Can we understand Martian language?
- What terrible innovations can we look forward to?

Things you’ve probably heard of

- Bots
- “50 cent party”
- Cyberbullying

Other things to keep on the radar

- Social credit systems
- Graduated responses
  - *E.g.*, WeChat


乌坎 vs. 鸟坎
niaokan
wukan
Frequency of Occurrence - logarithmic scale
1
5
10
50
100
500
1000
Days
0
20
40
60

Summary

- I’ve got plenty of problems and plenty of data
  - Basic questions about the data from Jeff Knockel’s dissertation
- Machines need to try harder to understand when humans are talking dirty
- NLP researchers are sorely needed on the front lines
  - *E.g.*, Cyberbullying campaigns in India
  - Open Technology Fund