Chat Program Censorship and Surveillance in China: Tracking TOM-Skype and Sina UC

Jeffrey Knockel, University of New Mexico
Greg Wiseman, Citizen Lab, Munk School, University of Toronto

We...

- Reverse engineered censorship and surveillance of TOM-Skype & Sina UC Chinese chat clients
- Analyzed changes to the triggering keyword lists made over an 18 month period
- Had an unbiased view of *all* triggering keywords *whenever* we wanted

TOM-Skype


TOM-Skype (TM) -
Skype 联系人 会话 通话 查看 工具 帮助
拨打电话
计时通:0分钟
Skype全国招商计划
联系人
最近的通话记录
搜索
全部
webi_edu
Echo / Sound Test Service
添加联系人
创建组
英语
购奔腾大置惠
一汽奔腾春季促销火热进行中
关闭
上线
视频通话
拨打
【降压】高血压安全降压、稳压法
【降糖】糖尿病-中医修复胰岛根源
【养肝】脂肪肝-喝茶巧逆转 必看
test
9:12 PM
fuck
9:12 PM
发送
通过 Skype
MSN新版本升级下载

TOM-Skype Censorship

- Different versions censor differently
- 6.1 censorship and surveillance:
  a1.skype.tom.com/installer/agent/keyfile5.5/keyfile
- 6.1 surveillance-only:
  a1.skype.tom.com/installer/agent/keyfile5.5/keyfile_u
- Each encrypted with DES+ECB with key:
  "\x7a\xdd\xe7\xdc\x23\x25\x53\x75"

TOM-Skype Surveillance

- Different versions surveil differently
- All send to same php script:

  a1.skype.tom.com/installer/tomad/ContentFilterMsg.php
- 6.1 example message:

JohnDoe fuck you 12/31/2011 6:00:00 PM 1 JaneDoe

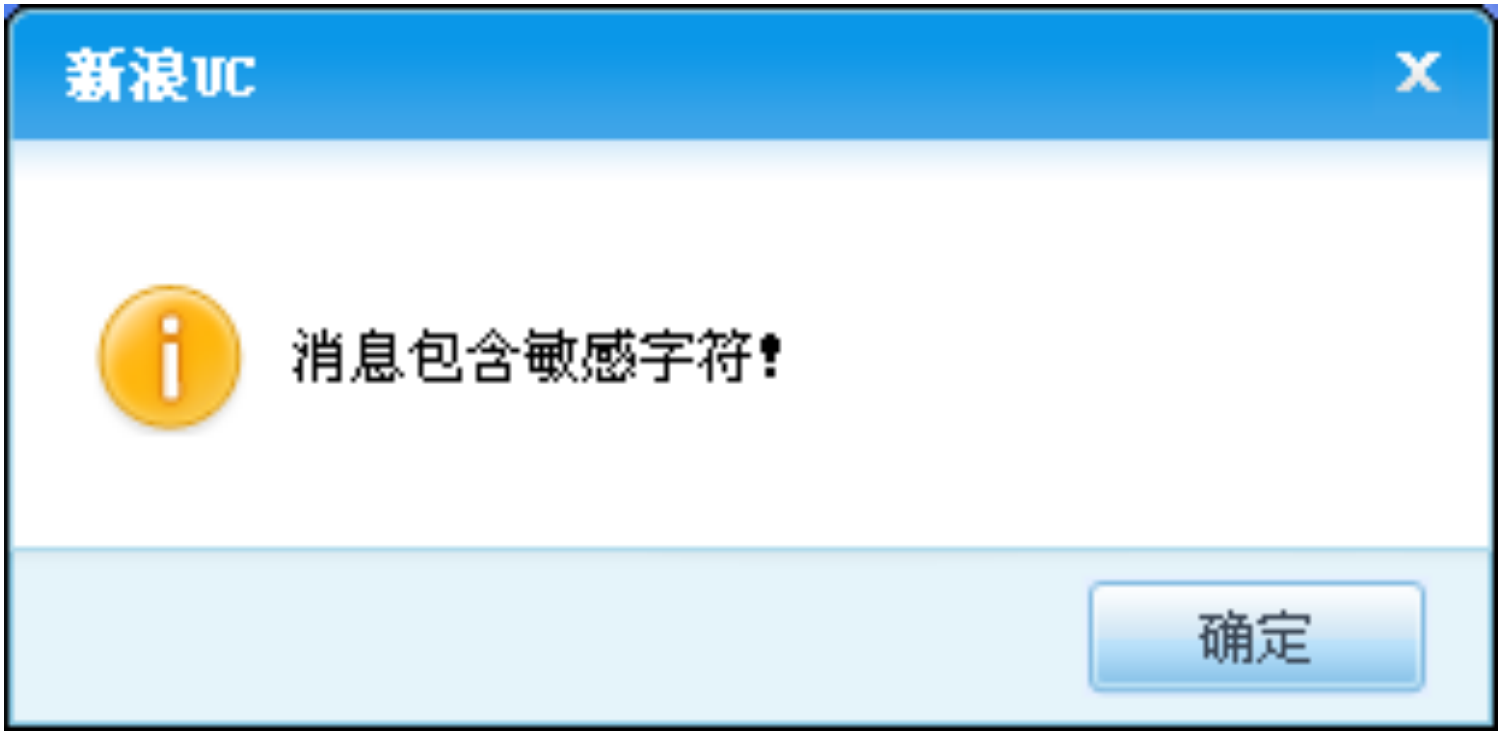
- Breakdown:

sender message date-time is-incoming receiver

- Encrypted with DES+ECB with key:

"X7sRUjL\0"

Sina UC



Sina UC Censorship

- Downloads from URL:
  im.sina.com.cn/fetch_keyword.php?ver=...
- Five lists JSON-encoded
  - List 1 censors chat and usernames
  - List 2 censors usernames
  - List 4 censors chat
  - Others have unknown purpose
- Encrypted with Blowfish+ECB with key:
  "H177UC09VI67KASI"
- No *client-side* surveillance

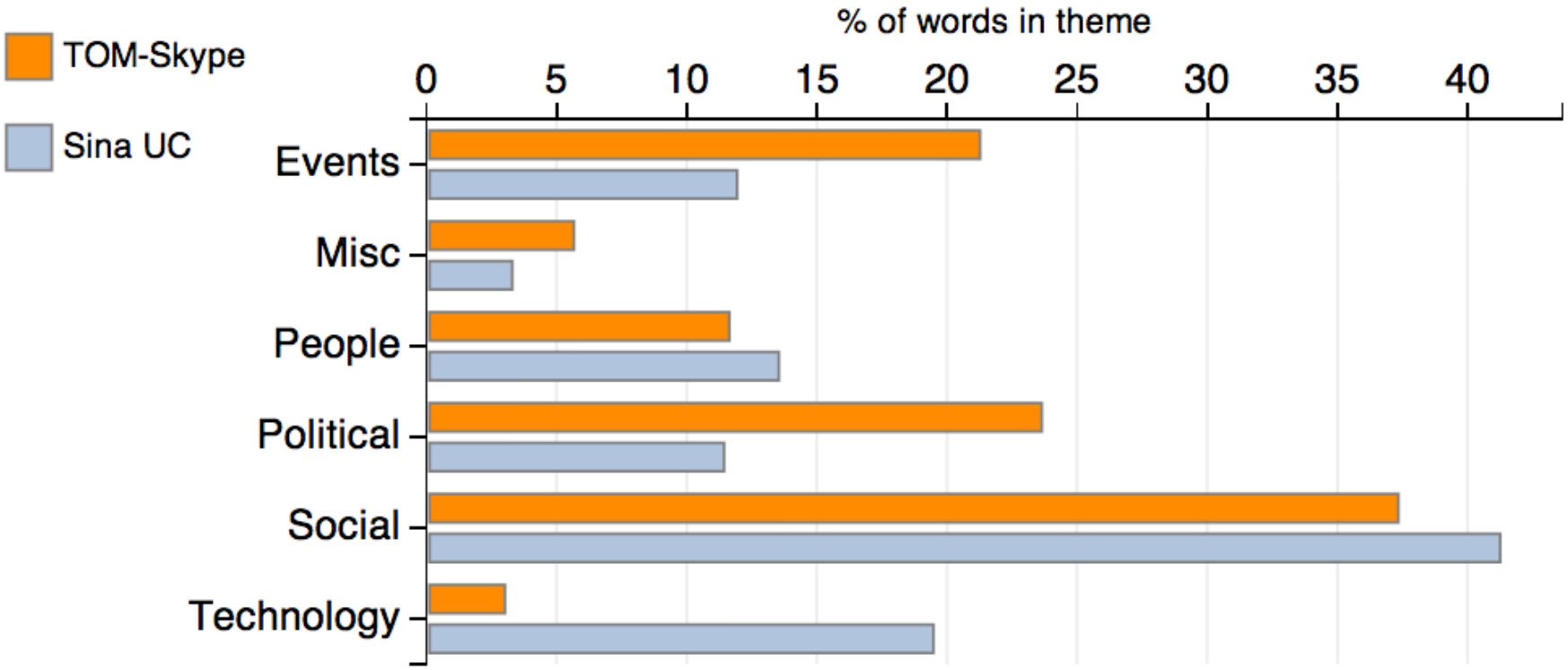
Dataset

- 2,576 distinct keywords across 8 TOM-Skype sources
- 1,818 distinct keywords across 5 Sina UC sources
- Only 138 keywords common between clients
- Lists range in size from 1 to 1,421 unique keywords
- 87% of keywords contain Chinese characters

Analysis

- All keywords human translated
- Associated with political / social context
- Categorized and tagged
- Visualizations
- Comparisons over time and between clients
- Correlating list changes with political events

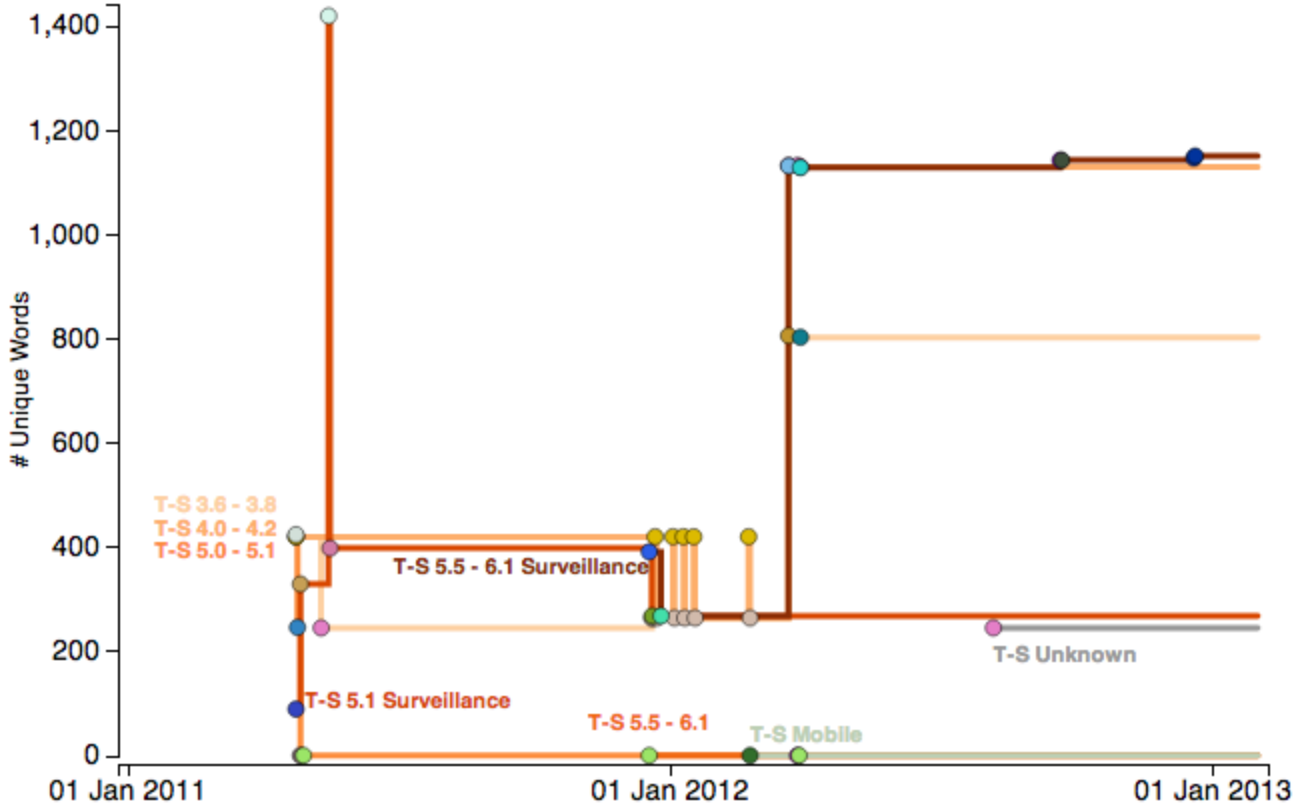
Variance Between Clients



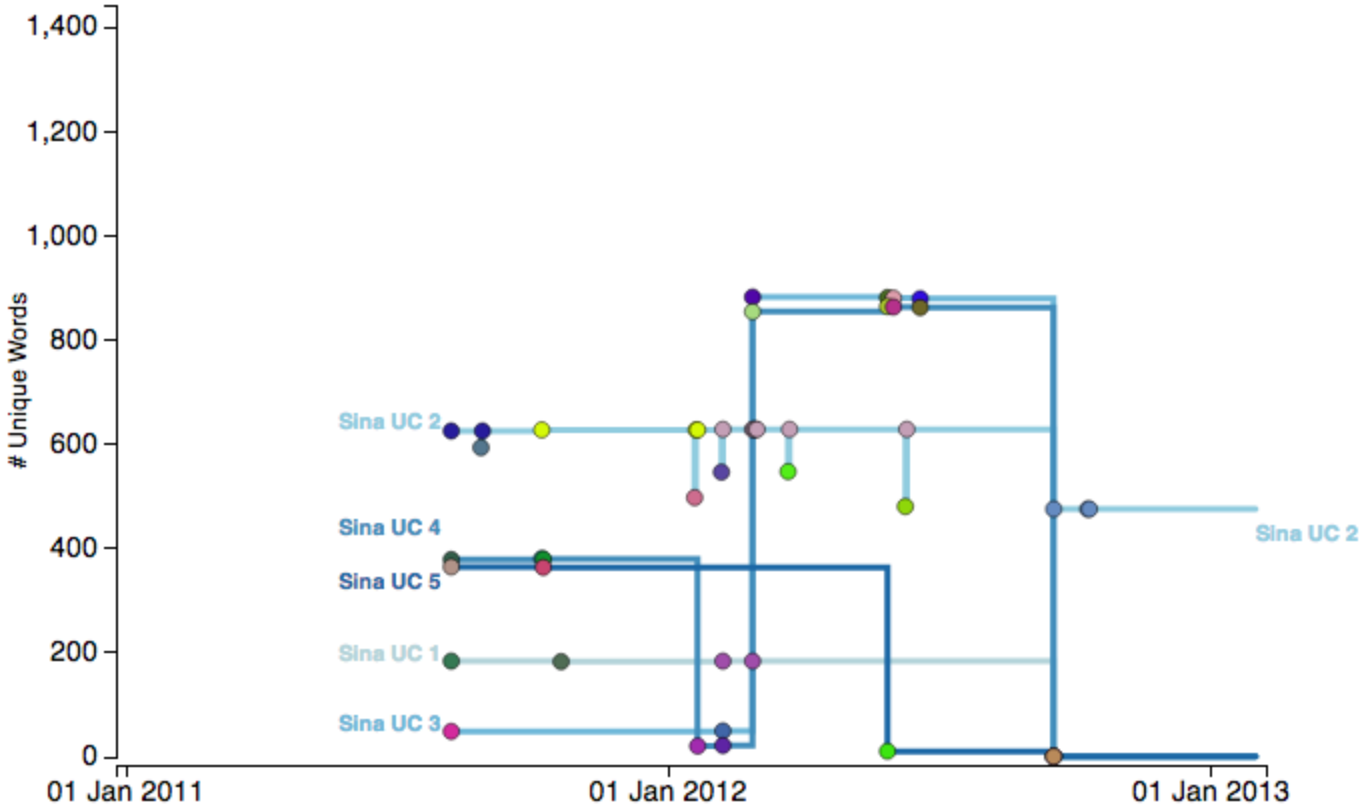
Keyword Content

- Adaptation to circumvention attempts
  - Unicode (e.g. six—four, ⑥④, Ⅰ Ⅸ Ⅷ Ⅸ)
  - Neologisms, homophones (e.g. Bo Xilai written as 薄熙来, 博西莱, B〇稀莱)
- Highly specific keywords
  - 西大直街康宁路路口世纪联华 ("Corning West and Da Zhi Street intersection, Century Lianhua gate")
- Extremely broad keywords
  - 华人 ("Chinese person")
  - 互联网 ("Internet")
- WTFs
  - Baby Mama Drama

List changes: TOM-Skype



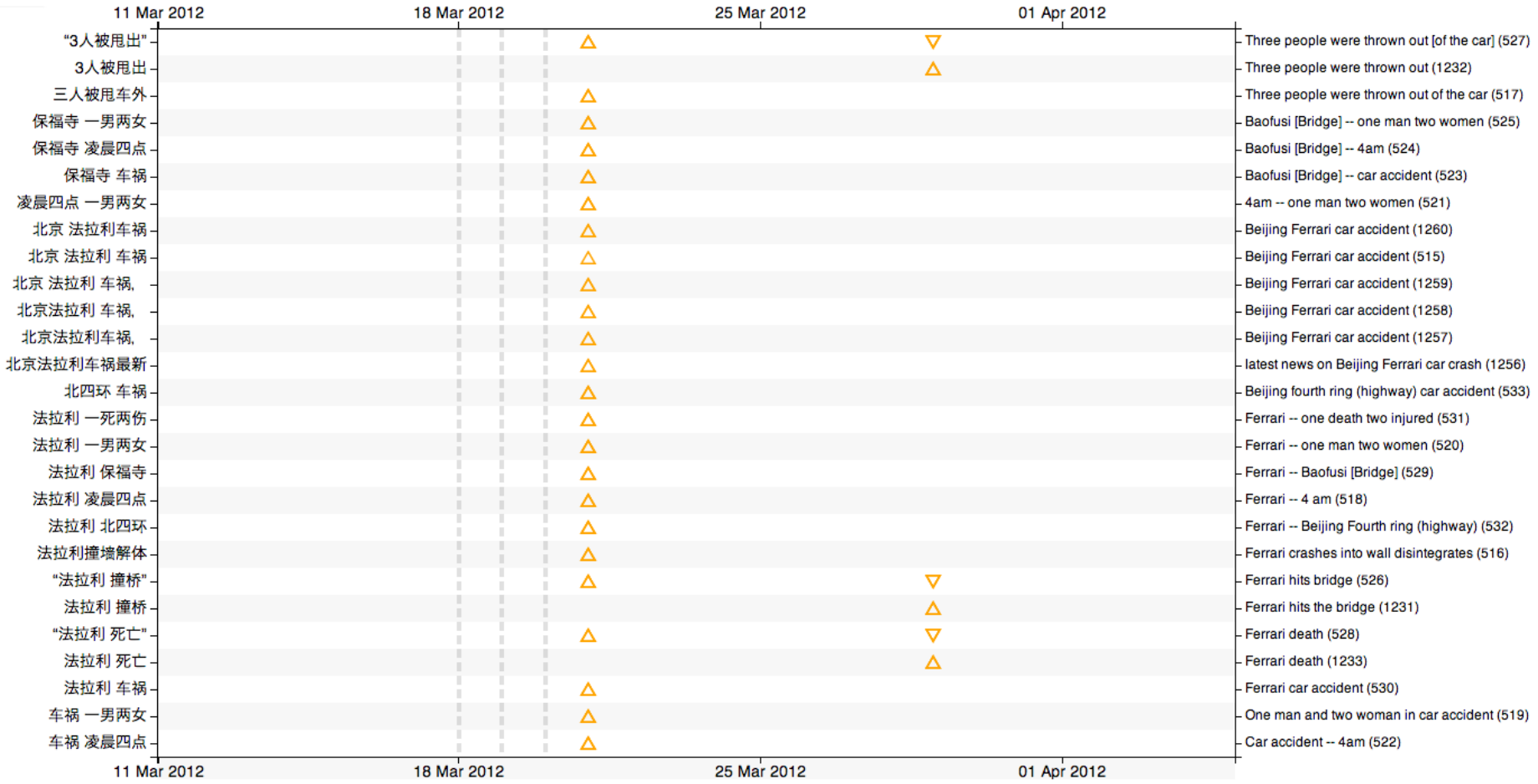
List changes: Sina UC



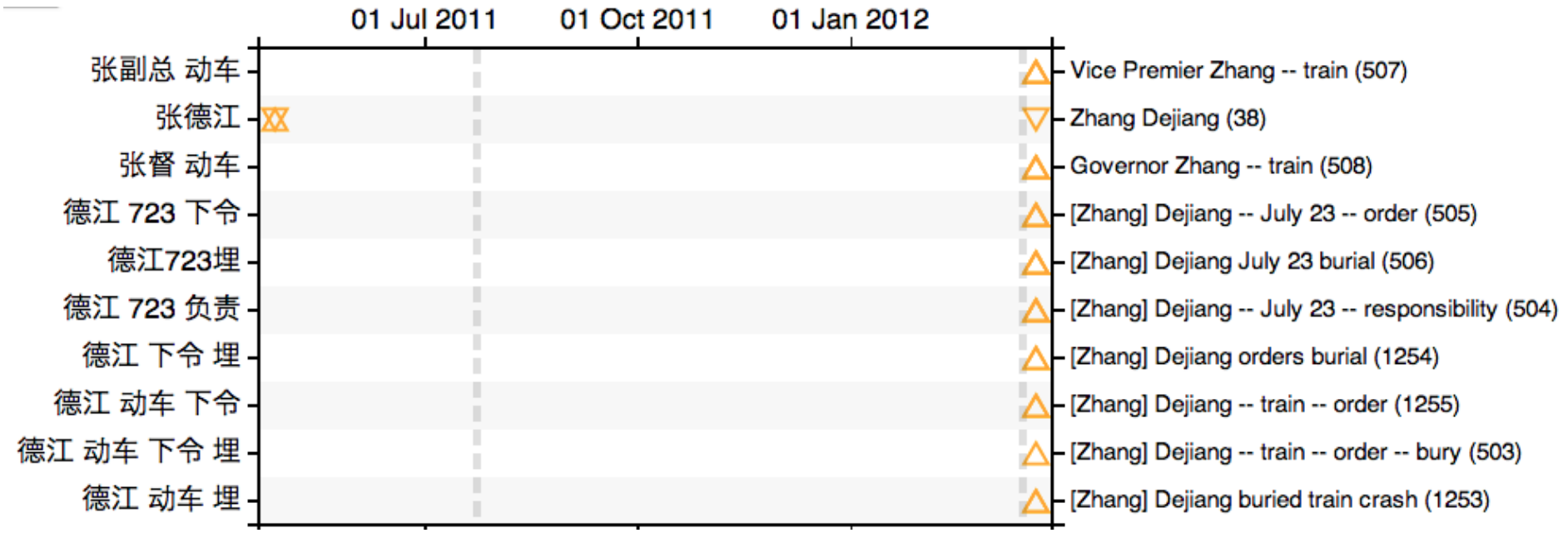
Reaction to Sensitive Events

- Identified current events referenced in the dataset and correlated them with keyword list updates
- Inconsistent patterns across selected events
- Seemingly important events were not always represented
- Two examples
  - Ferrari crash (March 2012)
  - Wenzhou train crash (July 2011)

Ferrari Crash



Wenzhou Train Crash



Conclusion

- Data set is unbiased and comprehensive, but many open questions remain
- Changes in censorship / surveillance focus
- Importance of interdisciplinary research
- Upcoming paper
- Website and (processed) data will be public soon