Chinese Keyword Censorship of Instant Messaging Programs

Jeffrey Knockel
Computer Science Department
University of New Mexico

UNM SCHOOL *of* ENGINEERING

Who Determines What's Censored in Chinese IM Programs?

IM Usage in China

- In 2010, 77.2% of Internet users in China used instant messaging
- 350 million users
- Growth rate of 30% from 2009
- Popular IM programs include Tencent QQ, Alitalk, TOM-Skype, Sina UC...

Source: http://www.iresearchchina.com/view.aspx?id=9205

Popular IM Programs in China

Program	Millions of daily users September 2009*
Tencent QQ/TM	139.85
Alitalk	22.87
MSN	20.11
Fetion	18.51
Caihong	16.94
(TOM-)Skype	2.67
Sina UC	2.53
Baidu Hi	2.08

*Source: http://satellite.tmcnet.com/news/2009/11/06/4467291.htm

Questions

- Which IM programs perform keyword censorship? Surveillance?
- Is there a “master” keyword list?
- What keywords are censored by which programs?
- Do programs tend to censor the same keywords?

Which Censor?

Program	Millions of daily users Sept. 2009*	Censors keywords?	Example keyword	Client-side?
Tencent QQ/TM	139.85	Yes	法轮 (falun)	No
Alitalk	22.87	Yes	吾尔开希 (Wu'er Kaixi)	No
MSN	20.11	No	-	-
Fetion	18.51	Yes	falundafa	No
Caihong	16.94	Yes	法轮 (falun)	No
(TOM-)Skype	2.67	Yes	fuck	Yes
Sina UC	2.53	Yes	六四 (six four)	Yes
Baidu Hi	2.08	Yes	六四 (six four)	No

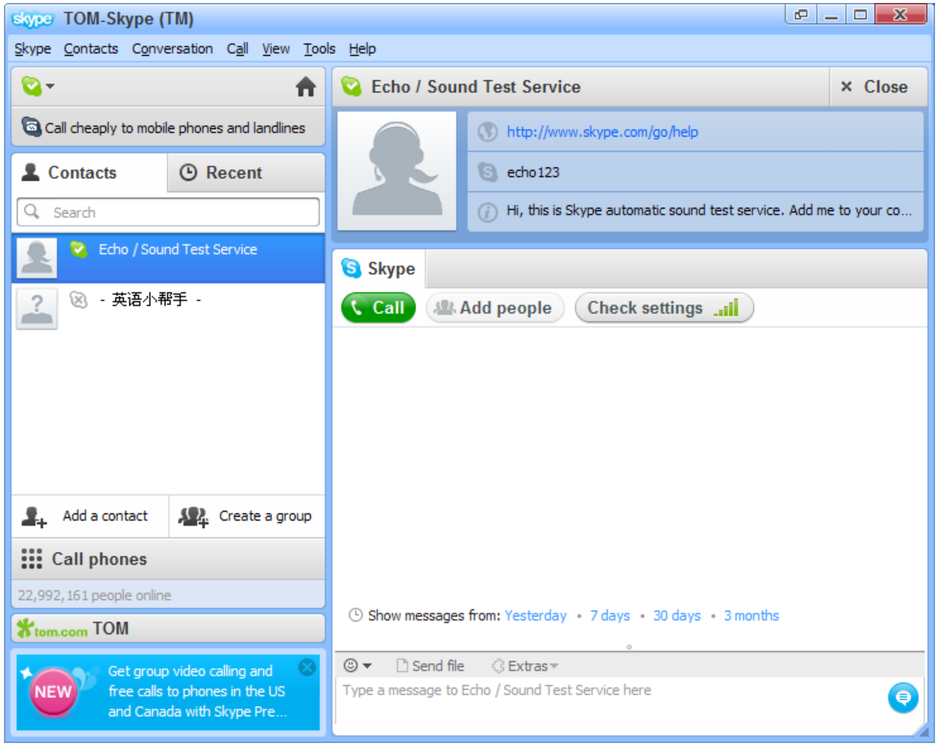
*Source: http://satellite.tmcnet.com/news/2009/11/06/4467291.htm

Client-side Censorship?

- TOM-Skype and Sina UC do censorship “client-side”
- When the censorship happens inside of the program
  - Not by remote server
  - Not somewhere on the network
- Encrypted keyword lists are hidden in program and/or downloaded

TOM-Skype

- TOM-Skype
  - Modified version of Skype by TOM Group Limited, a China-based media company
  - Uses Skype's network
  - In China, http://www.skype.com HTTP redirects to http://skype.tom.com


TOM-Skype (TM)
Skype Contacts Conversation Call View Tools Help
Call cheaply to mobile phones and landlines
Contacts
Recent
Search
Echo / Sound Test Service
- 英语小帮手 -
Add a contact
Create a group
Call phones
22,992,161 people online
tom.com TOM
NEW
Echo / Sound Test Service
× Close
http://www.skype.com/go/help
echo123
Skype
Call
Add people
Check settings
Show messages from: Yesterday • 7 days • 30 days • 3 months
Send file
Extras
Type a message to Echo / Sound Test Service here

Hi, this is Skype automatic sound test service. Add me to your co...

Get group video calling and free calls to phones in the US and Canada with Skype Pre...

Empirical Analysis of TOM-Skype

- TOM-Skype uses “keyfiles”
  - List of encrypted keywords triggering censorship and surveillance of text chat
  - One built-in
  - At least one other downloaded
  - Lists vary by version of TOM-Skype

3.6-4.2 Keyfiles

- TOM-Skype 3.6-3.8 downloads from http://skypetools.tom.com/agent/newkeyfile/keyfile
- TOM-Skype 4.0-4.2 downloads from http://a[1-8].skype.tom.com/installer/agent/keyfile
- Encrypted with naïve xor algorithm...

```
procedure DECRYPT ($C_{0..n}$, $P_{1..n}$)
  for i ← 1, n do
    $P_i = (C_i \oplus 0x68) - C_{i-1} \pmod{0xff}$
  end for
end procedure
```

5.0-5.1 Keyfiles

- TOM-Skype 5.0-5.1 downloads keyfiles from

  http://skypetools.tom.com/agent/keyfile

- TOM-Skype 5.1 downloads surveillance-only keyfile from

  http://skypetools.tom.com/agent/keyfile_u

- Keywords AES encrypted in ECB mode
- Key reused from TOM-Skype 2.x
- When encoded in UTF16-LE, 32 bytes:

  ```
  0sr TM#RWFD,a43 
  ```

- Half of bytes printable ASCII, other half null (weak)

TOM-Skype Surveillance

- TOM-Skype 3.6-3.8 encrypts surveillance traffic with DES key in ECB mode:

`32bnx23l`

- TOM-Skype 5.0: no surveillance
- TOM-Skype 4.0-4.2, 5.1 encrypts using different DES key:

`X7sRUjL\0`

TOM-Skype Surveillance

- Example surveillance message from 3.6-4.2:

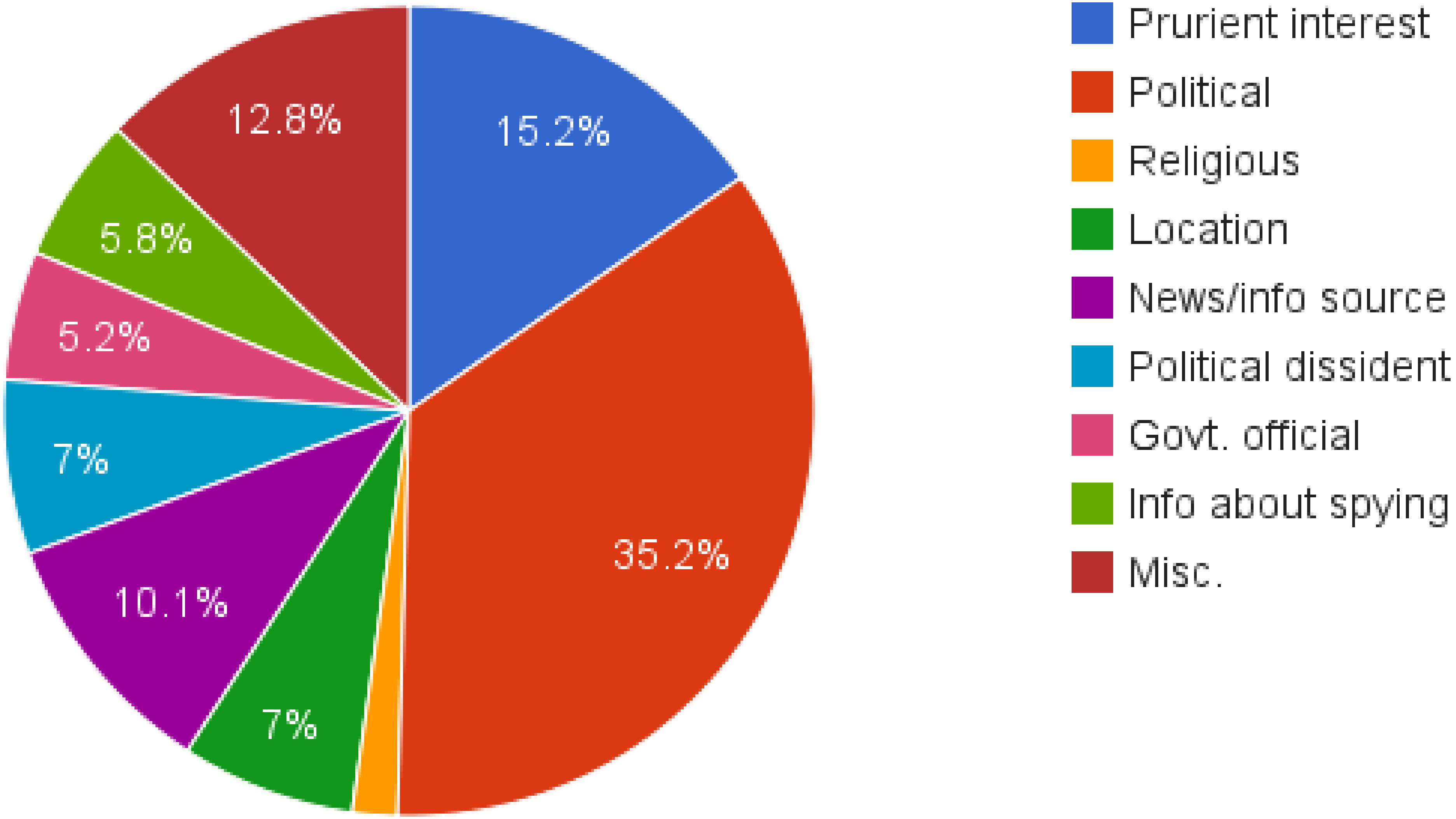
```
jdoe falungong 4/24/2011 2:25:53 AM 0
```

- Message author followed by triggering message followed by the date and time
- 0 or 1 indicates message is outgoing or incoming, respectively
- Example surveillance message from 5.1:

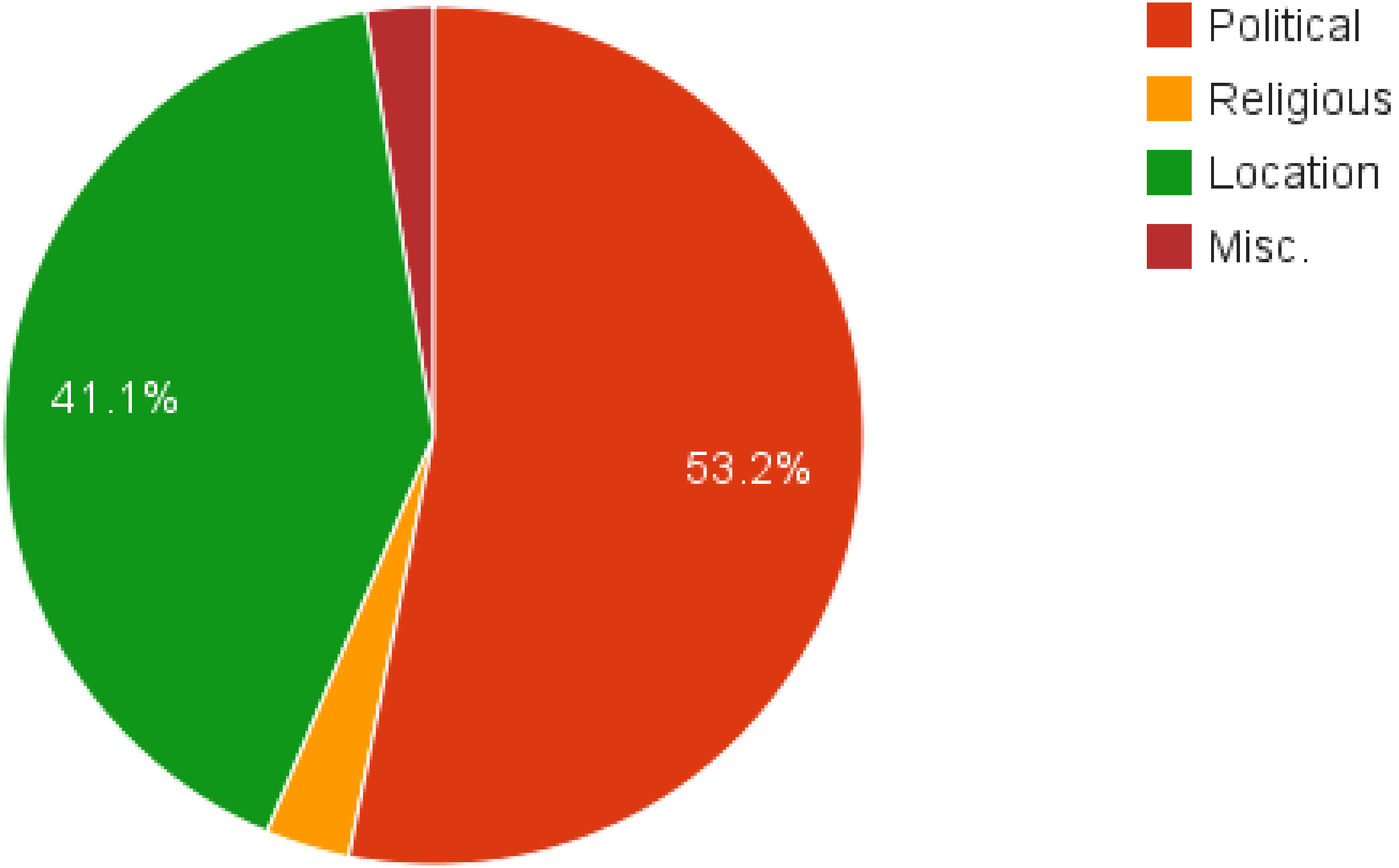
```
falungong 4/24/2011 2:29:57 AM 1
```

- 5.1 does not report username
- 5.1 does not report outgoing messages

5.0-5.1 Downloaded Keyfile



5.1 Surveillance-only Keyfile



Censored Keywords

- Keyfile contained political words (35.2%)
  - 六四 (“64,” in reference to the June 4th Incident)
  - 拿着麦克风表示自由 (Hold a microphone to indicate liberty)
- Prurient interests (15.2%)
  - 操烂 (Fuck rotten)
  - 两女一杯 (Two girls one cup)

Censored Keywords

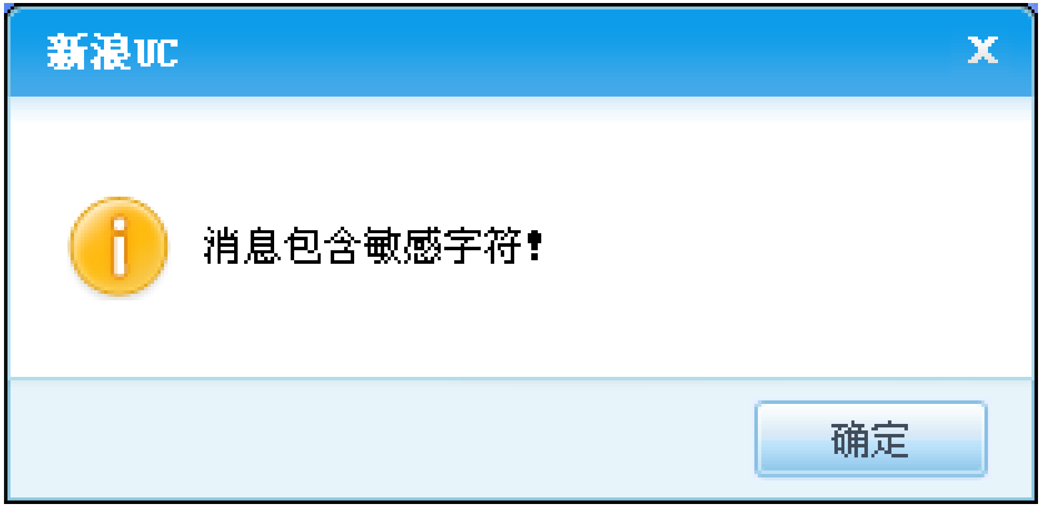
- News/info sources (10.1%)
  - 中文维基百科 (Chinese language Wikipedia)
  - BBC 中文网 (BBC Chinese language)
- Political dissidents (7%)
  - 刘晓波 (Liu Xiaobo)
  - 江天勇 (Jiang Tianyong)
- Locations (7%)
  - 成都 春熙路麦当劳门前 (McDonald's in front of Chunxi Road in Chengdu)

Surveillance-only

- Mostly political and locations
  - Almost all related to demolitions of homes in Beijing for future construction
  - A few related to illegal churches
  - A couple company names

Sina UC

- By SINA Corporation
  - China-based company
  - Owns weibo.com, Chinese social networking site
- Uses Jabber protocol



新浪UC

消息包含敏感字符!

Empirical Analysis of Sina UC

- Has five lists
- One set of five built-in
- Another set of five downloaded from

http://im.sina.com.cn/fetch_keyword.php?ver=...

- All five lists JSON-encoded
- Then Blowfish encrypted in ECB mode with the following 16-byte ASCII-encoded key:

`H177UC09VI67KASI`

List #4

- Used to censor text chat
- Large number of neologisms for the June 4th incident:
    - 5月三十五 (May 35th), 四月六十五号 (April 65th), 三月九十六号 (March 96th)
    - 61过后三天 (three days after June 1st), 儿童节过后三天 (three days after Children's day)
    - ⑥④, VIIV, 8|9|6|4, six.4
    - 6.2+2
    - 八的二次方 (8^2), 2的6次方 (2^6)

List #4

- Even Russian:
  - Четыре (four)
  - Шесть (six)
  - Девять (nine)
  - Восемь (eight)
  - Восемь-Девять-Шесть-Четыре (eight-nine-six-four)
- And French:
  - six-quatre (six-four)

List #2

- Used to censor usernames (username replaced with id#)
- Found prurient words like 婊子 (whore), 妓 (prostitute)
- Political: 法輪 (falun), falun, six four
- Phishing:
  - webmaster, root, admin, hostmaster, sysadmin, sinaUC, 新浪 (Sina), 系统通知 (system notice)

Other Lists

- List #1 is a shorter list used to censor both text chat and usernames
- List #3 contains a lot of domains; has unknown purpose
- List #5 contains prurient and political keywords; has unknown purpose

Comparative Analysis

- TOM-Skype and Sina UC have lists for different purposes
- For each, let's union their sets of keywords
- TOM-Skype has 515 unique keywords
- Sina UC has 997 unique keywords
- Overall, 1446 keywords are seen in only TOM-Skype xor Sina UC
- Only 33 are common to both
- Conjecture: any “master” list must be short

Conclusion and Future Work

- When programs censor client-side, we can find exact keyword lists
- Why do TOM-Skype, Sina UC censor client-side?
  - Skype network P2P, encrypted, not owned by China
  - Sina UC uses Jabber protocol; maybe a “stock” server solution?
  - “Distributed” censorship
- Censorship in other IM programs?

For keyword lists, machine and human translations, and source code, see

- http://cs.unm.edu/~jeffk/tom-skype/
- http://cs.unm.edu/~jeffk/sinauc/

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. CCR #0313160, CAREER #0644058, CAREER #0844880, and TC-M #090517.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.