Chinese Keyword Censorship of Instant Messaging Programs

Jeffrey Knockel
Computer Science Department
University of New Mexico
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...

# Popular IM Programs in China

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

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?

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

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
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 (C₀..n, P₁..n)
  for i ← 1, n do
    P_i = (C_i ⊕ 0x68) - C_{i-1} (mod 0xff)
  end for
end procedure
```
5.0-5.1 Keyfiles

- TOM-Skype 5.0-5.1 downloads keyfiles from
- TOM-Skype 5.1 downloads surveillance-only keyfile from
- Keywords AES encrypted in ECB mode
- Key reused from TOM-Skype 2.x
- When encoded in UTF16-LE, 32 bytes:
  `0srx 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: 32bnx231
- 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

- Prurient interest: 15.2%
- Political: 35.2%
- Religious: 7%
- Location: 7%
- News/info source: 7%
- Political dissident: 10.1%
- Govt. official: 5.8%
- Info about spying: 5.2%
- Misc.: 12.8%
5.1 Surveillance-only Keyfile

- Political: 53.2%
- Location: 41.1%
- Religious: 5.1%
- Misc.: 0.6%
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
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, VIIIV, 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 妓子 (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.