UNIVERSITY OF TORONTO

MUNK SCHOOL OF GLOBAL AFFAIRS

Join the Global Conversation

The Citizen Lab

Research Brief
Number 47 – October 2014

Asia Chats: LINE keyword filtering upgraded to include regular expressions

On Wednesday September 24, 2014, regionally-based keyword filtering features in the LINE chat app were updated. The most notable change is the introduction of regular expressions to the keyword list, which enables more advanced keyword matching. The addition of regular expressions demonstrates LINE Corporation's continued commitment to filtering keywords for users based in China and a push to improve the underlying technology.

PREVIOUS KEYWORD LIST CHANGES

Since October 31 2013, when we began monitoring the LINE censored keyword lists, we have observed 6 keyword list updates, with list versions numbered 20 through 25. The following table summarizes the versions of the keyword lists we have identified:

List Version	Date of change	Notable changes	Report
20	First seen on October 31, 2013		Link
21	First seen on October 31, 2013		Link
22	4/8/2014	312 new keywords added, 147 keywords deleted	Link
23	6/23/2014	Addition of a single keyword	Link
24	9/24/2014	Addition of regular expressions	
25	9/24/2014	Minor update to correct extraneous spacing in some regular expressions	

In October 2013 we first reported on regionally-based keyword filtering functionality in LINE that is enabled if a user's account is registered to a mainland China phone number. This analysis identified two keywords lists versions 20 and 21.

The April 2014 update to version 22 was the most substantial change in content we have observed. 312 new keywords were added to the list and 147 keywords were removed, leaving 535 total keywords. In this update, all of the removed keywords related to Bo Xilai, a prominent Chinese politician who was jailed for corruption while his wife was convicted of murder. The majority of the 312 new additions related to Chinese government officials or notable political events. A more thorough analysis of the keywords can be found in our April 2014 report.

There was a minor update to the list in June 2014, when the keyword list changed to version 23. In this update, only a single keyword (□□ 'buy a gun') was added while none were removed. In our report on that development, we speculated that this unusual change may have been the result of LINE's developers testing if the keyword list update mechanism still functioned following the reported blocking of messaging applications, including LINE and KakaoTalk, for users based in China.

UPDATED KEYWORD LIST

The most recent changes to the list were seen on September 24, 2014. At approximately 5am EDT, list version 24 appeared. This update was followed an hour later by the introduction of version 25. Versions 24 and 25 are almost identical, with only a minor addition to a single keyword as well the removal of some extraneous spaces which were present in v24 keywords.

6 keywords presents on v23 are removed in v24 and v25:

Keyword	Pinyin	Translation and notes
近平	Jìn píng	PRC President Xi Jinping's family name
领导人	Lǐngdǎo rén	Leaders
彭丽媛	Péng Lìyuàn	Wife of Xi Jinping
习近平	Xí Jìnpíng	PRC President
习近平	Xí Jìnpíng	(Same as above, was duplicated in v23)
李克强	Lǐ Kèqiáng	Current premier of China

It is not entirely clear why references to current Chinese leader Xi Jinping, premier Li Keqiang and Xi's wife Peng Liyuan were removed. '□□人' ('leaders', referring to government officials) is a commonly used term used in state media and could possibly have been removed for being too broad, although this is speculation.

In addition, 6 keywords were edited to remove trailing spaces present in v23 which would have prevented use of these keywords from being censored.

REGULAR EXPRESSIONS

Version 24 and 25 contain a major change not seen in any of the previous lists – the addition of regular expressions. Prior versions consisted of lists of keywords that triggered blocking of messages if any of the message text matched a keyword. With regular expressions, there is now the possibility for more advanced keyword matching, as messages containing a particular keyword can be censored only if, for example, they contain another keyword. This behaviour has been seen in Chinese social media services , such as Weibo[1], which has long had the ability to censor keywords containing a combination of keywords. Other reports, however, have argued that performance considerations mean China's national level web censorship infrastructure never uses the boolean 'OR' operation.

As an example of the new regular expressions, consider this entry from list v25:

.*(**中国**|**李锐**).*六四.*

- The expression begins with '.*', which means that it will match on any and all characters, including none at all. This means the matching text could be at the beginning of a message or anywhere in the middle.
- The next section contains a pair of two-character strings, in parentheses, separated by '|', the OR operator. This means that either **中国** ('China') or **李锐** (Li Rui, a retired government official who was openly advocated for democratic reform in China) would match this part of the regular expression.
- This character is followed by another instance of '.*', which again means it will match any and all text.
- Following this, is the two-character pair **六四**, or 'six four', which is a reference to the June 4, 1989 Tiananmen Square massacre.
- Finally, it ends with another '.*', meaning it would match any and all text after the two-character combination.

What this regular expression means is that any message sent using LINE which contains either '中国' or '李□' followed at any point by '六四' will be blocked. This method is a more flexible and advanced mechanism for censorship than simple keyword matching, which would only block messages containing text that exactly matches the censored keyword. This new technique could potentially have the effect of reducing collateral filtering caused by overbroad keyword lists. However, minimizing overblocking depends on how the regular expression filtering is implemented. Connecting a series of commonly used keywords a through a regular expression could result in overblocking.

In total, there were 48 regular expressions added to keyword list v25:

Regular expression	Translation
.*乌鲁木齐.*(75|杀人|自由|独立|买枪).*	* Urumqi *. (75 | murder | Freedom | Independent | buy a gun) *.
.*(75|杀人|自由|独立|买枪).*乌鲁木齐.*	. *. (75 | murder | Freedom | Independent | buy a gun) * Urumqi *.
.*维族.*(75|七五|杀人|独立|砍人|暴乱).*	* Uighur *. (75 | seventy-five | murder | independent | knives to injure | riots) *.

.*(75|七五|杀人|独立|砍人|暴乱).*维族.*	. *. (75 | seventy-five | murder | independent | knives to injure | riots) * Uighur *.
.*和田.*暴乱.*	* Hotan. * Riots *
.*暴乱.*和田.*	* Riots * Hotan. *
.*南疆.*(杀人|独立|买枪).*	* Southern Xinjiang * (murder | Independent | buy a gun) *.
.*(杀人|独立|买枪).*南疆.*	* (Murder | Independent | buy a gun). * Southern Xinjiang *.
.*新疆.*(七五|75|独立|中共|暴政|王乐泉).*	* Xinjiang *. (Seventy-five | 75 | Independent | CPC | Tyranny | Wang Lequan) *.
.*(七五|75|独立|中共|暴政|王乐泉).*新疆.*	* (Seventy-five | 75 | Independent | CPC | Tyranny | Wang Lequan). * Xinjiang *.
.*香港.*(独立|集会|游行|中共|民运|暴政).*	* Hong Kong *. (Independence | meeting | Parade | CPC | democracy movement | tyranny) *.
.*(独立|集会|游行|中共|民运|暴政).*香港.*	* (Independence | meeting | Parade | CPC | pro | tyranny). * Hong Kong *.
.*中共.*(暴政|黑匣子|太子党|内斗|垮台|专政).*	* CPC *. (Tyranny | black box | princeling | infighting | collapse | dictatorship) *.
.*(暴政|黑匣子|太子党|内斗|垮台|专政).*中共.*	*. (Tyranny | black box | princeling | infighting | collapse | dictatorship) * CPC *.
.*71.*(集会|游行).*	* 71 * (rally | Parade) *.
.*(集会|游行).*71.*	* (Rally | Parade) * 71 *.
.*七一.*(集会|游行).*	* Seventy-one *. (Rally | Parade) *.
.*(集会|游行).*七一.*	*. (Rally | Parade) * seventy-one *.
.*学生.*(64|六四).*	* Student *. (64 | sixty-four) *.
.*(64|六四).*学生.*	. *. (64 | sixty-four) * student *.
.*天安门.*(89|八九|坦克).*	* Tiananmen Square *. (89 | eighty-nine | tanks) *.
.*(89|八九|坦克).*天安门.*	. *. (89 | eighty-nine | tanks) * Tiananmen Square *.

.*阿坝.*自焚.*	* Ngba. * Self-immolation. *[2]
.*自焚.*阿坝.*	* Self-immolation. * Ngaba. *
.*暴力.*革命.*	* Violent * revolution *
.*革命.*暴力.*	* Revolution * violence. *
.*北京.*政变.*	* Beijing * coup. *
.*政变.*北京.*	* Coup. * Beijing *
.*国殇.*国庆.*	* National Martyr * National Day *
.*国庆.*国殇.*	* National Day. * National Martyr *.
.*黄海.*事件.*	* Yellow Sea. * Events *
.*事件.*黄海.*	* Events * Yellow Sea. *
.*令计.*车祸.*	* Ling Ji . * Accident *.[3]
.*车祸.*令计.*	* Accident * Ling Ji. *
.*民主.*自由.*	* Democracy * freedom *
.*自由.*民主.*	* Freedom * democracy *.
.*六四.*(中国|李锐).*	* Sixty-four *. (China | Li Rui) *.
.*(中国|李锐).*六四.*	* (China | Li Rui) * sixty-four *.
.*祖英.*泽民.*	* Zuying. * Zemin. *[4]
.*泽民.*祖英.*	* Zemin. * Zuying. *
.*活摘.*器官.*	* Live pick. * Organ *[5]
.*器官.*活摘.*	* Organ * live pick. *
.*共产党.*(推翻|打倒|九评|9评).*	* Communist * (overthrowing | knock down | Nine Commentaries | 9 Comments) *.
.*(推翻|打倒|九评|9评).*共产党.*	* (Overthrowing | knock down | Nine Commentaries | 9 Comments) * communist *.

.*茉莉花.*联盟.*	* Jasmine * alliance *
.*联盟.*茉莉花.*	* alliance * Jasmine *
.*宪章.*(08|零八).*	* Charter *. (08 | and eight) *.
.*(08|零八).*宪章.*	. *. (08 | and eight) * Charter *.

The expressions on this list cover many of the same topics seen in keywords contained in previous versions of the list, such as the Tiananmen Square massacre, conflict in regions with large Uyghur populations, the Jasmine revolution protests, and Tibetan self-immolations.

There are four new regular expressions which contain the keyword □□ ('buy a gun'), all of which relate to protests and violence involving Uyghur populations in Xinjiang. As previously mentioned, the addition of this keyword was the sole change between lists v22 and v23. The original standalone keyword □□ is still present on v25.

List v25 contains six new regular expressions related to protest and rallies in Hong Kong. This includes two expressions which discuss Hong Kong independence issues and protests in general terms:

.*香港.*(独立|集会|游行|中共|民运|暴政).** Hong Kong *. (Independent | meeting | Parade | CPC | pro | tyranny) *.

.*(独立|集会|游行|中共|民运|暴政).*香港.** (Independent | meeting | Parade | CPC | pro | tyranny). * Hong Kong *.

The other four regular expressions include the keyword '71', which is a reference to the July 1st rallies which traditionally accompany the anniversary of the July 1, 1997 transfer of sovereignty over Hong Kong to China:

.*71.*(集会|游行).** 71 * (rally | Parade) *.

.*(集会|游行).*71.** (Rally | Parade) * 71 *.

.*七一.*(集会|游行).** Seventy-one *. (Rally | Parade) *.>

.*(集会|游行).*七一.**. (Rally | Parade) * seventy-one *.

It does not appear that the addition of these regular expressions relating to Hong Kong protests is directly related to the ongoing Occupy Central protests in the city. The list update occurred on September 24, 2014, several days before the protests expanded. There are also other mentions of Hong Kong rallies (such as 香港□园六四, 'Hong Kong Victoria Park six four', which refers to an annual event held to commemorate the Tiananmen Square massacre) still present on list v25 which date back to list v21, which we first identified in October 2013. Thus while political rallies in Hong Kong are reflected in the LINE censored keywords list, there is no indication that the September/October 2014 Occupy Central rallies are specifically targeted.

As evident in the above list of words relating to Hong Kong, there are attempts made to account for censored keywords that may appear in a different order in a user's message. Regular expressions such as '.*七一.*(集会|游行).*' and '.*(集会|游行).*七一.*' are likely used to ensure that messages are blocked even if keywords are used in a different order.

In some cases there are duplications in the list which would appear to render certain regular expressions unnecessary. For example, v25 contains the keyword 六四 ('six four', a reference to the June 4, 1989 Tiananmen Square massacre). However, four of the newly added regular expressions contain this same keyword, such as:

.*(64|**六四**).***学生**.*.* (64 or 'six four') .* 'student' .*

As 六四 ('six four') is already present on the keyword list, it is redundant for any regular expressions to also contain this keyword. Similarly, the keyword □□ ('buy a gun') is present on the list, but is also contained within four regular expressions. It is not clear why these redundant keywords remain. Although the use of regular expressions could allow for a more precise application of filtering, the continued presence of many broad keywords (such as 六四 ('six four') or 八九 ('eight nine') negates this impact.

CONCLUSION

While most of the new keywords added to this recent update do not stray far, in terms of content, from the keywords previously seen censored for LINE users in China, the addition of regular expressions represents an advancement in the techniques used to censor content on the service. While more flexibility may allow for more precise targeting of keywords (and thus reduce overblocking from broad keywords), this will depend on how the regular expressions are implemented. This addition shows that developers continue to refine and improvement censorship mechanisms in LINE.

RESOURCES

- Raw and translated keyword list data on Github
- LINE Region Code Encrypter Tool – A tool that enables users to change regions in the LINE client and disable regionally-based keyword censorship in the application

ACKNOWLEDGEMENTS

Adam Senft, Jason Q. Ng, Seth Hardy, and Masashi Crete-Nishihata undertook the research and writing of this post. This research is supported by the John D. and Catherine T. MacArthur Foundation.

FOOTNOTES

[1]See, for example, the China Digital Times' Sensitive Words series, which shows many examples of combinations of words blocked on Weibo:http://chinadigitaltimes.net/china/sensitive-words-series/
[2]Ngaba is a region in Tibet which has where dozens of local Tibetans have self-immolated.
[3]"Ling Ji" refers to Ling Jihua, a government official who was demoted after it was revealed he tried to cover up his son's fatal car accident.
[4]Zuying refers to Song Zuying, an alleged mistress of former president Jiang Zemin.

[5]Refers to allegations by the Falun Gong that the CCP harvests organs from members who have been persecuted, captured, and executed.