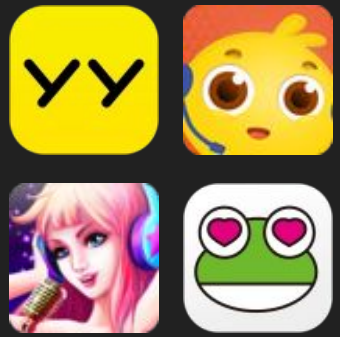
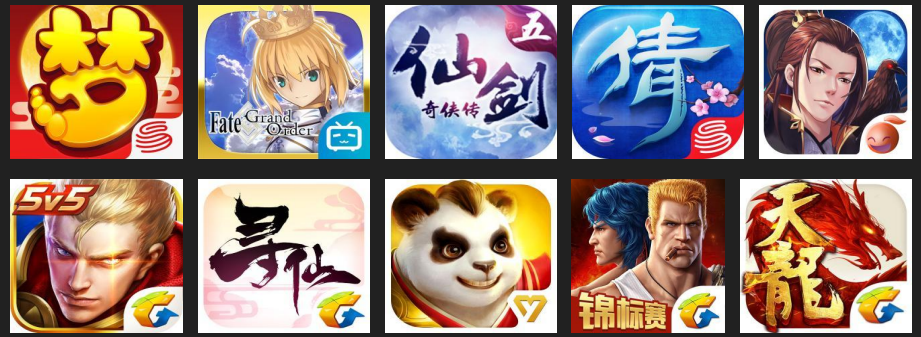
An Efficient Method to Determine which Combination of Keywords Triggered Automatic Filtering of a Message

Ruohan Xiong and Jeffrey Knockel

Citizen Lab, University of Toronto







tom.com & skype

LINE

Regulatory environment

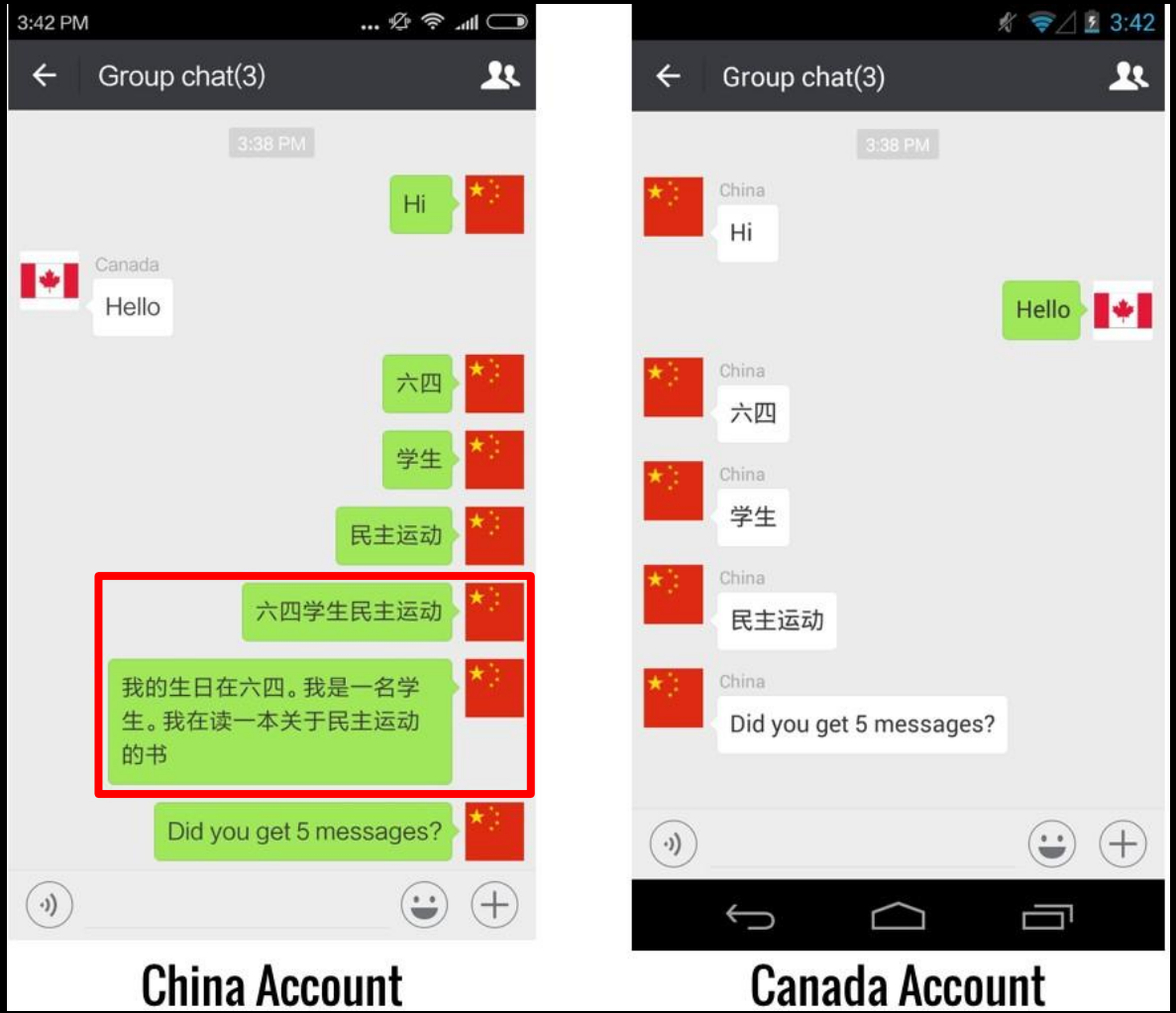
- Online content providers in China are responsible for content they host
- Delegated censorship directives
- Companies seek to filter sensitive content quickly with minimal human involvement

Sensitive keyword combinations

六四 + 学生 + 民主运动

six four + student + democracy movement

- Each combination contains one or more components
- Message censored if all components matched


3:42 PM
Group chat(3)
3:38 PM
Hi
Canada
Hello
六四
学生
民主运动
六四学生民主运动
我的生日在六四。我是一名学生。我在读一本关于民主运动的书
Did you get 5 messages?
China Account
3:42
Group chat(3)
3:38 PM
China
Hi
Hello
China
六四
China
学生
China
民主运动
China
Did you get 5 messages?
Canada Account

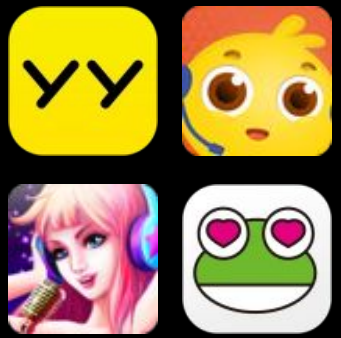
Client-side censorship

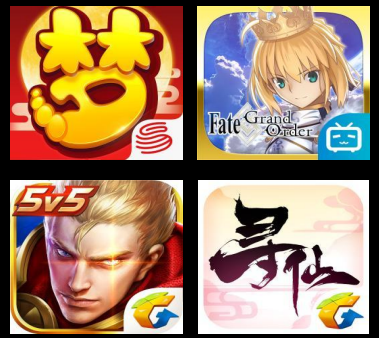
- Censorship occurs in local application
- Extract keyword list through reverse-engineering











Server-side censorship

- Keyword lists are stored on a remote server
- Seen in chat apps including WeChat
- Sample testing required
  - Efficient methods needed



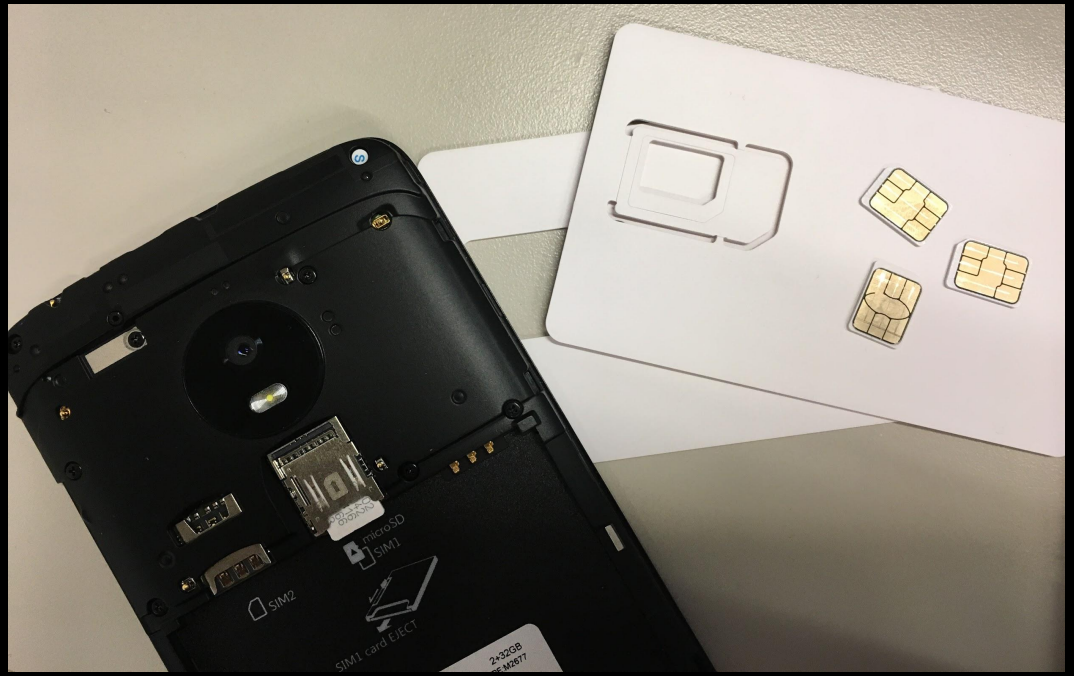
News articles for sample testing

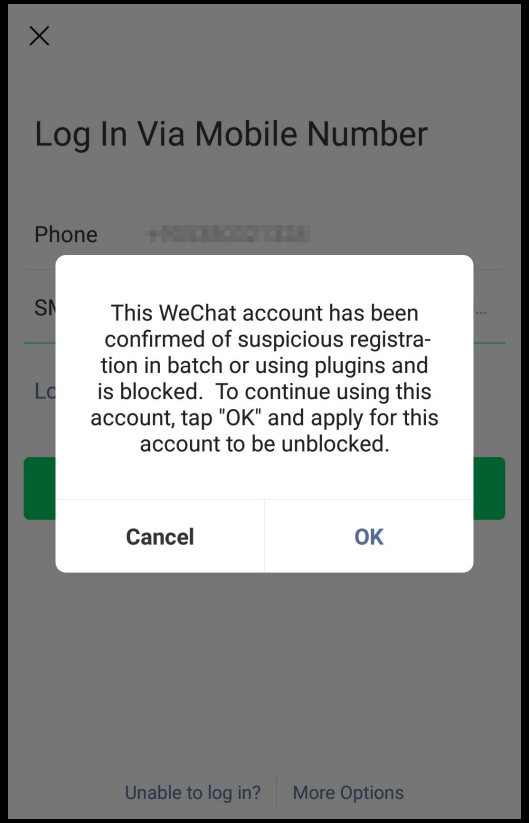
- Up-to-date text samples with coverage of current events, particularly political events
- International sources report on issues which may not be disseminated by local media

Testing a social media platform

- Pool of accounts we control
  - Sender
  - Receiver
  - Passive accounts
- Requires sending high volume of messages

Sample testing is costly





Algorithms for sample testing

- Original algorithm
- *Improved algorithms*
    1. Binary search with backtracking
    2. Binary splitting
    3. Component-aware binary splitting

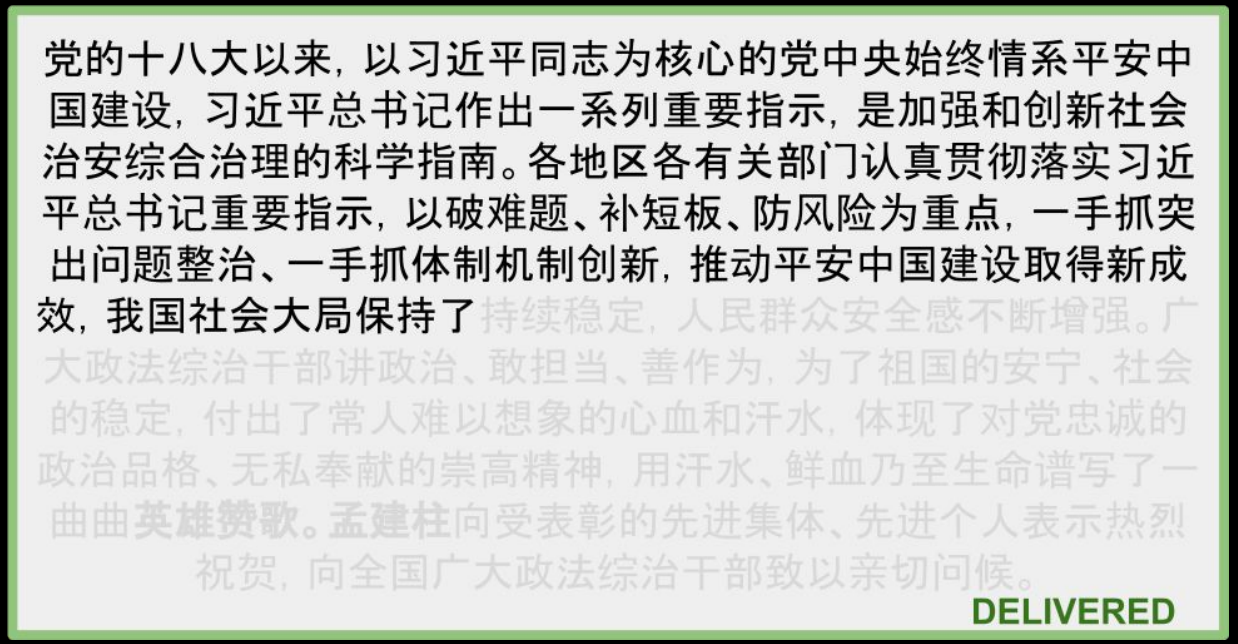
Original algorithm
Stage 1/3: Recursive bisection

党的十八大以来，以习近平同志为核心的党中央始终情系平安中国建设，习近平总书记作出一系列重要指示，是加强和创新社会治安综合治理的科学指南。各地区各有关部门认真贯彻落实习近平总书记重要指示，以破难题、补短板、防风险为重点，一手抓突出问题整治、一手抓体制机制创新，推动平安中国建设取得新成效，我国社会大局保持了持续稳定，人民群众安全感不断增强。广大政法综治干部讲政治、敢担当、善作为，为了祖国的安宁、社会的稳定，付出了常人难以想象的心血和汗水，体现了对党忠诚的政治品格、无私奉献的崇高精神，用汗水、鲜血乃至生命谱写了一曲曲**英雄赞歌**。**孟建柱**向受表彰的先进集体、先进个人表示热烈祝贺，向全国广大政法综治干部致以亲切问候。

CENSORED

Recursively split message into halves and continue testing with censored half

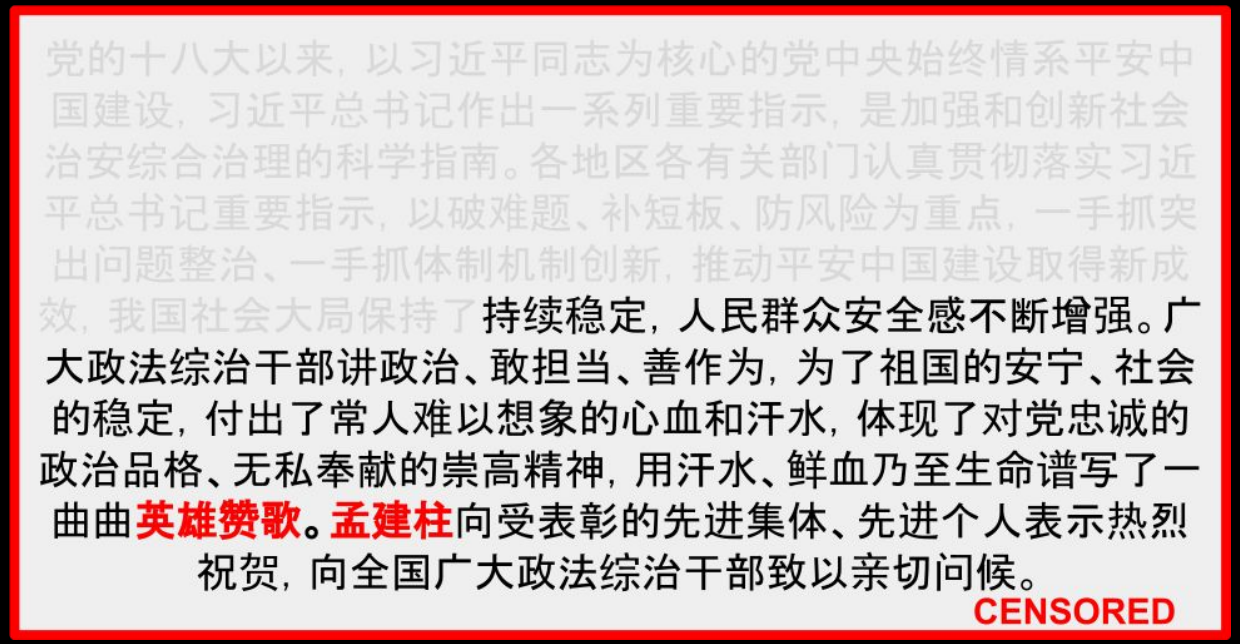
Original algorithm
Stage 1/3: Recursive bisection



Left half:

Not censored. Return and continue testing with other half.

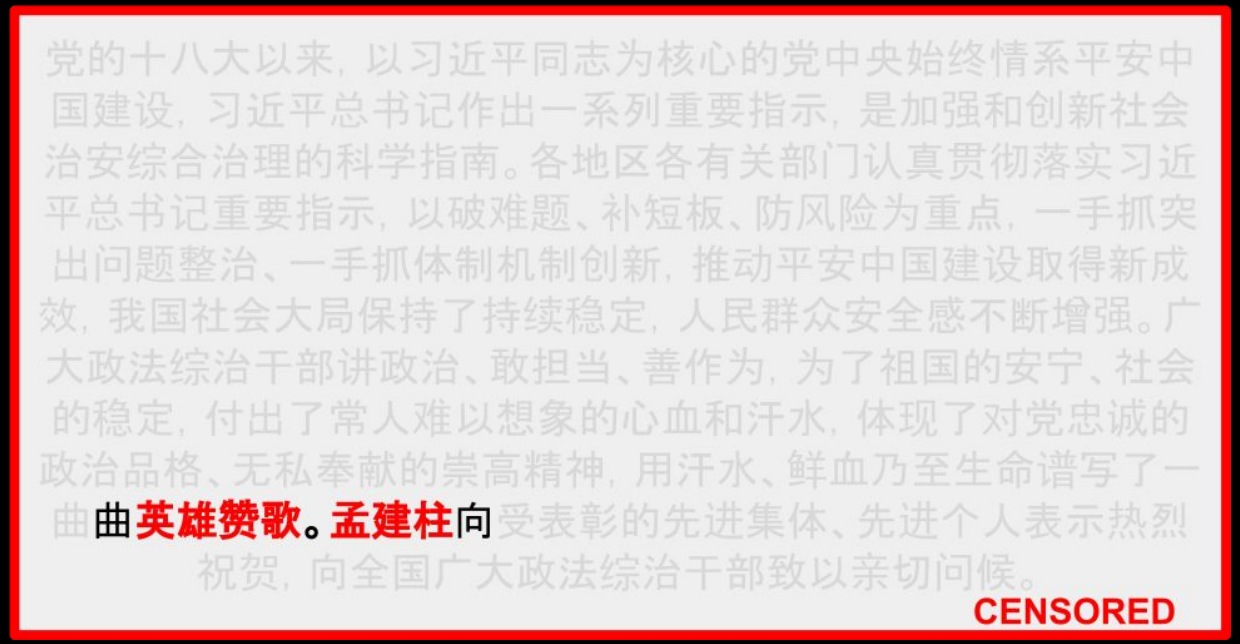
Original algorithm
Stage 1/3: Recursive bisection



Right half:

Censored. Branch into left and right halves.

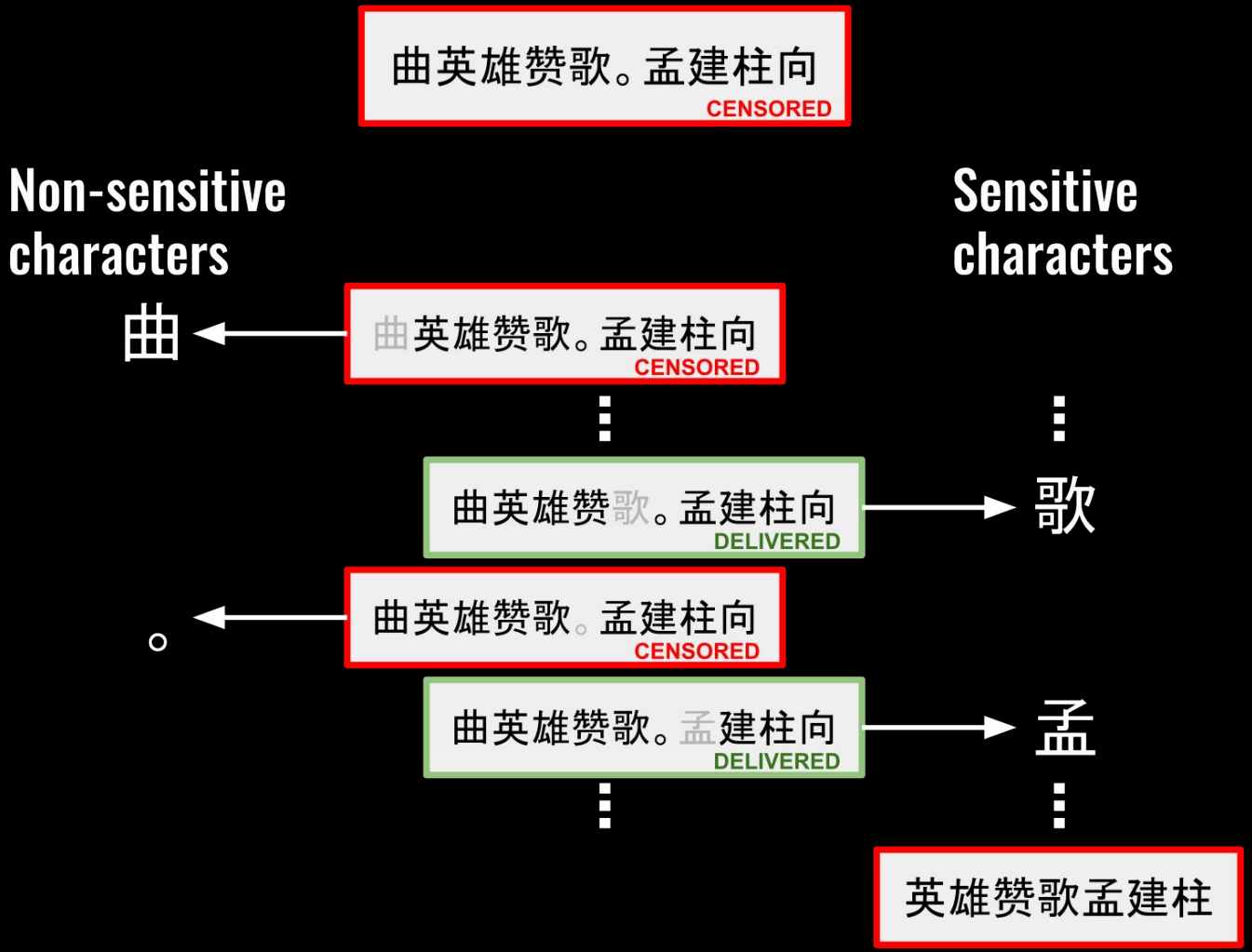
Original algorithm
Stage 1/3: Recursive bisection



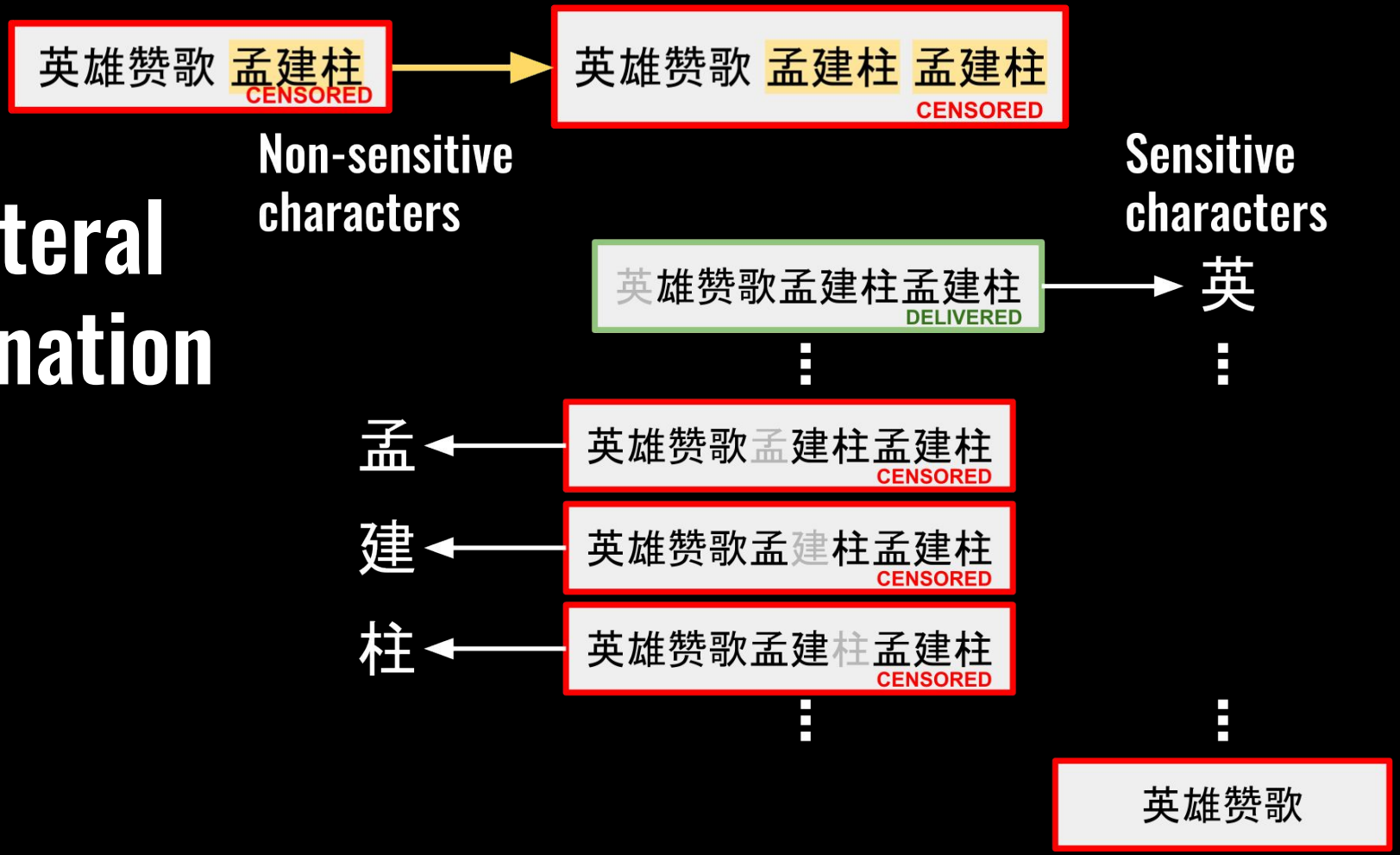
Reduced sensitive segment

Original algorithm

Stage 2/3: Keyword Isolation

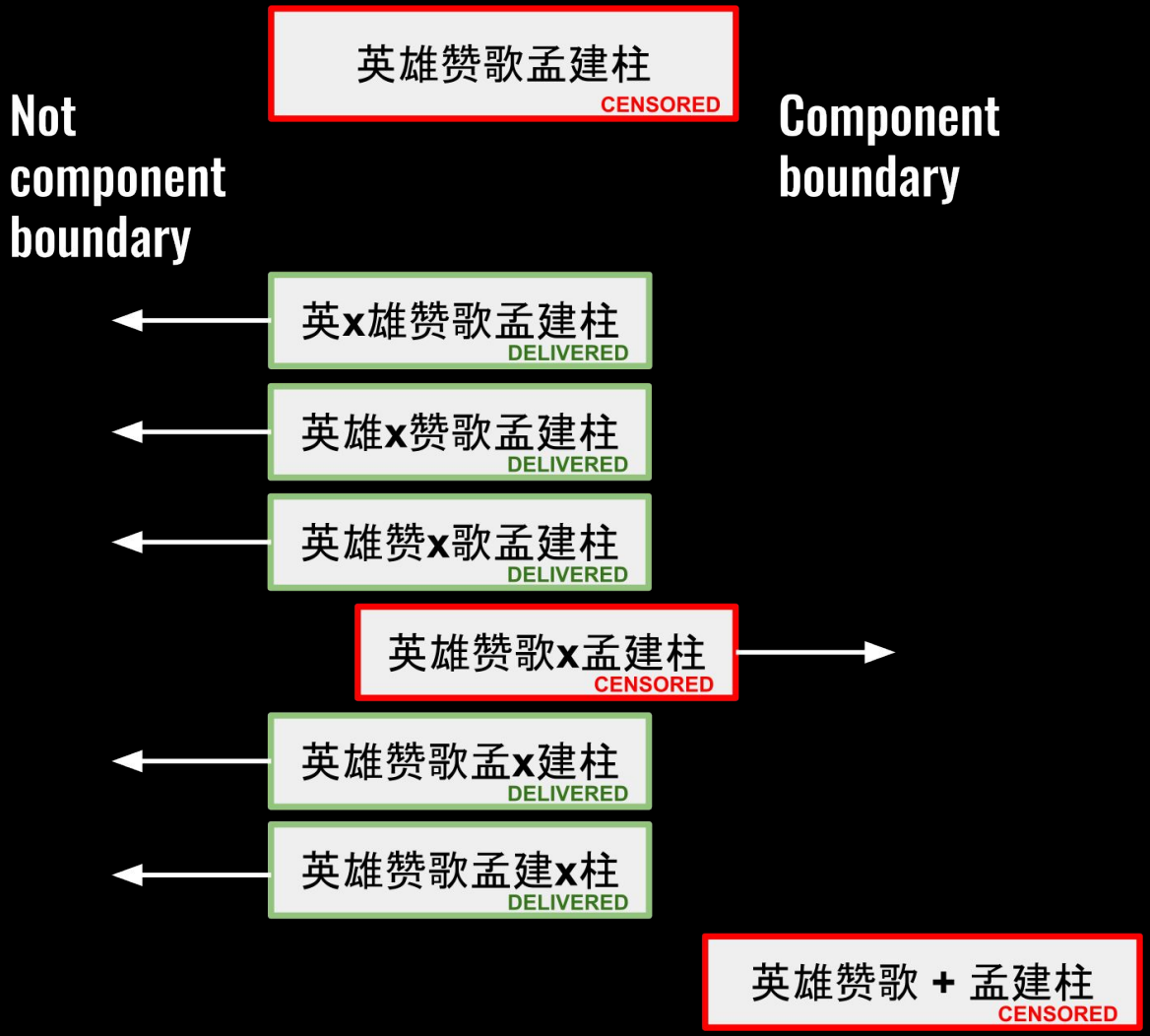


Unilateral elimination flaw



Original Algorithm

Stage 3/3: Split components



Overlap flaw

- Text: 帶來調整體領域
- Returned: 帶來 + 調整體 + 領域 ❌
- Actual: 帶來 + 調整 + 整體 + 領域 ✅

铜锣湾书店的老板，他原籍浙江宁波，1996年获得瑞典国籍，2003年在内地交通肇事，撞死一名女大学生后潜逃，经辗转，最后到香港定居，操起出版政治八卦书籍的生意，那些书籍在内地造成极坏影响。他于2015年10月回到内地投案自首，至今处于羁押中。西方社会与“人权”“言论自由”有关的奖项多得大概数不过来。它们不断冒出来，给中国大大小小的“异见人士”颁奖。给人一种印象，在中国跟政府对着干，就算有了被西方某个奖项瞄上的基本条件。如果在这当中触犯法律蹲了几天监狱，或者是微博账号被封了，大体就“入围”了。大奖得不着，小奖说不定哪天就能分到一个。给中国“异见人士”颁奖，西方有些名不见经传的小奖项也有利可图。其实奖给哪个具体中国“异见人士”并不重要，重要的是它们可以通过这样做“傍上中国”，刷自己的存在感。给中国“异见人士”颁奖，比给其他人颁奖都更容易被报道，“挑战中国”的碰瓷如今在西方蛮时髦的。像“变态辣椒”那样的画手，画得本来就不怎么样，他当初在互联网上画极端政治漫画就是为了捞粉。出走动漫大国日本，好像业余球员去了巴西，画普通漫画连饭都吃不上，只有靠画骂中国的画维持生计了。还有**桂敏海**，出的书全都是胡编乱造的那一类，只追求耸动，卖出去骗钱。这两人都是投机分子，缺少做人的底线，给他们奖的机构大概只看中了他们身上的标签，对他们未必做了全面了解。

CENSORED

Average messages per news article: 343

Improved algorithm 1: Binary search with backtracking

铜锣湾书店的老板，他原籍浙江宁波，1996年获得瑞典国籍，2003年在内地交通肇事，撞死一名女大学生后潜逃，经辗转，最后到香港定居，操起出版政治八卦书籍的生意，那些书籍在内地造成极坏影响。他于2015年10月回到内地投案自首，至今处于羁押中。西方社会与“人权”“言论自由”有关的奖项多得大概数不过来。它们不断冒出来，给中国大大小小的“异见人士”颁奖。给人一种印象，在中国跟政府对着干，就算有了被西方某个奖项瞄上的基本条件。如果在这当中触犯法律蹲了几天监狱，或者是微博账号被封了，大体就“入围”了。大奖得不着，小奖说不定哪天就能分到一个。给中国“异见人士”颁奖，西方有些名不见经传的小奖项也有利可图。其实奖给哪个具体中国“异见人士”并不重要，重要的是它们可以通过这样做“傍上中国”，刷自己的存在感。给中国“异见人士”颁奖，比给其他人颁奖都更容易被报道，“挑战中国”的碰瓷如今在西方蛮时髦的。像“变态辣椒”那样的画手，画得本来就不怎么样，他当初在互联网上画极端政治漫画就是为了捞粉。出走动漫大国日本，好像业余球员去了巴西，画普通漫画连饭都吃不上，只有靠画骂中国的画维持生计了。还有**桂敏海**，出的书全都是胡编乱造的那一类，只追求耸动，卖出去骗钱。这两人都是投机分子，缺少做人的底线，给他们奖的机构大概只看中了他们身上的标签，对他们未必做了全面了解。

CENSORED

Attempt iterative redaction on lower and upper halves

Improved algorithm 1: Binary search with backtracking

铜锣湾书店的老板，他原籍浙江宁波，1996年获得瑞典国籍，2003年在内地交通肇事，撞死一名女大学生后潜逃，经辗转，最后到香港定居，操起出版政治八卦书籍的生意，那些书籍在内地造成极坏影响。他于2015年10月回到内地投案自首，至今处于羁押中。西方社会与“人权”“言论自由”有关的奖项多得大概数不过来。它们不断冒出来，给中国大大小小的“异见人士”颁奖。给人一种印象，在中国跟政府对着干，就算有了被西方某个奖项瞄上的基本条件。如果在这当中触犯法律蹲了几天监狱，或者是微博账号被封了，大体就“入围”了。大奖得不着，小奖说不定哪天就能分到一个。给中国“异见人士”颁奖，西方有些名不见经传的小奖项也有利可图。其实奖给哪个具体中国“异见人士”并不重要，重要的是它们可以通过这样做“傍上中国”，刷自己的存在感。给中国“异见人士”颁奖，比给其他人颁奖都更容易被报道，“挑战中国”的碰瓷如今在西方蛮时髦的。像“变态辣椒”那样的画手，画得本来就不怎么样，他当初在互联网上画极端政治漫画就是为了捞粉。出走动漫大国日本，好像业余球员去了巴西，画普通漫画连饭都吃不上，只有靠画骂中国的画维持生计了。还有桂敏海，出的书全都是胡编乱造的那一类，只追求耸动，卖出去骗钱。这两人都是投机分子，缺少做人的底线，给他们奖的机构大概只看中了他们身上的标签，对他们未必做了全面了解。

DELIVERED

Not censored

Undo redaction and divide area of interest into two halves

Improved algorithm 1: Binary search with backtracking

铜锣湾书店的老板，他原籍浙江宁波，1996年获得瑞典国籍，2003年在内地交通肇事，撞死一名女大学生后潜逃，经辗转，最后到香港定居，操起出版政治八卦书籍的生意，那些书籍在内地造成极坏影响。他于2015年10月回到内地投案自首，至今处于羁押中。西方社会与“人权”“言论自由”有关的奖项多得大概数不过来。它们不断冒出来，给中国大大小小的“异见人士”颁奖。给人一种印象，在中国跟政府对着干，就算有了被西方某个奖项瞄上的基本条件。如果在这当中触犯法律蹲了几天监狱，或者是微博账号被封了，大体就“入围”了。大奖得不着，小奖说不定哪天就能分到一个。给中国“异见人士”颁奖，西方有些名不见经传的小奖项也有利可图。其实奖给哪个具体中国“异见人士”并不重要，重要的是它们可以通过这样做“傍上中国”，刷自己的存在感。给中国“异见人士”颁奖，比给其他人颁奖都更容易被报道，“挑战中国”的碰瓷如今在西方蛮时髦的。像“变态辣椒”那样的画手，画得本来就不怎么样，他当初在互联网上画极端政治漫画就是为了捞粉。出走动漫大国日本，好像业余球员去了巴西，画普通漫画连饭都吃不上，只有靠画骂中国的画维持生计了。还有桂敏海，出的书全都是胡编乱造的那一类，只追求耸动，卖出去骗钱。这两人都是投机分子，缺少做人的底线，给他们奖的机构大概只看中了他们身上的标签，对他们未必做了全面了解。

DELIVERED

Not censored

Undo redaction and divide area of interest into two halves

(and so on...)

Improved algorithm 1: Binary search with backtracking

铜锣湾书店的老板，他原籍浙江宁波，1996年获得瑞典国籍，2003年在内地交通肇事，撞死一名女大学生后潜逃，经辗转，最后到香港定居，操起出版政治八卦书籍的生意，那些书籍在内地造成极坏影响。他于2015年10月回到内地投案自首，至今处于羁押中。西方社会与“人权”“言论自由”有关的奖项多得大概数不过来。它们不断冒出来，给中国大大小小的“异见人士”颁奖。给人一种印象，在中国跟政府对着干，就算有了被西方某个奖项瞄上的基本条件。如果在这当中触犯法律蹲了几天监狱，或者是微博账号被封了，大体就“入围”了。大奖得不着，小奖说不定哪天就能分到一个。给中国“异见人士”颁奖，西方有些名不见经传的小奖项也有利可图。其实奖给哪个具体中国“异见人士”并不重要，重要的是它们可以通过这样做“傍上中国”，刷自己的存在感。给中国“异见人士”颁奖，比给其他人颁奖都更容易被报道，“挑战中国”的碰瓷如今在西方蛮时髦的。像“变态辣椒”那样的画手，画得本来就不怎么样，他当初在互联网上画极端政治漫画就是为了捞粉。出走动漫大国日本，好像业余球员去了巴西，画普通漫画连饭都吃不上，只有靠画骂中国的画维持生计了。还有桂敏海，出的书全都是胡编乱造的那一类，只追求耸动，卖出去骗钱。这两人都是投机分子，缺少做人的底线，给他们奖的机构大概只看中了他们身上的标签，对他们未必做了全面了解。

CENSORED

Censored

Keep the redaction made and continue

Improved algorithm 1: Binary search with backtracking

铜锣湾书店的老板，他原籍浙江宁波，1996年获得瑞典国籍，2003年在内地交通肇事，撞死一名女大学生后潜逃，经辗转，最后到香港定居，操起出版政治八卦书籍的生意，那些书籍在内地造成极坏影响。他于2015年10月回到内地投案自首，至今处于羁押中。西方社会与“人权”“言论自由”有关的奖项多得大概数不过来。它们不断冒出来，给中国大大小小的“异见人士”颁奖。给人一种印象，在中国跟政府对着干，就算有了被西方某个奖项瞄上的基本条件。如果在这当中触犯法律蹲了几天监狱，或者是微博账号被封了，大体就“入围”了。大奖得不着，小奖说不定哪天就能分到一个。给中国“异见人士”颁奖，西方有些名不见经传的小奖项也有利可图。其实奖给哪个具体中国“异见人士”并不重要，重要的是它们可以通过这样做“傍上中国”，刷自己的存在感。给中国“异见人士”颁奖，比给其他人颁奖都更容易被报道，“挑战中国”的碰瓷如今在西方蛮时髦的。像“变态辣椒”那样的画手，画得本来就不怎么样，他当初在互联网上画极端政治漫画就是为了捞粉。出走动漫大国日本，好像业余球员去了巴西，画普通漫画连饭都吃不上，只有靠画骂中国的画维持生计了。还有桂敏海，出的书全都是胡编乱造的那一类，只追求耸动，卖出去骗钱。这两人都是投机分子，缺少做人的底线，给他们奖的机构大概只看中了他们身上的标签，对他们未必做了全面了解。

CENSORED

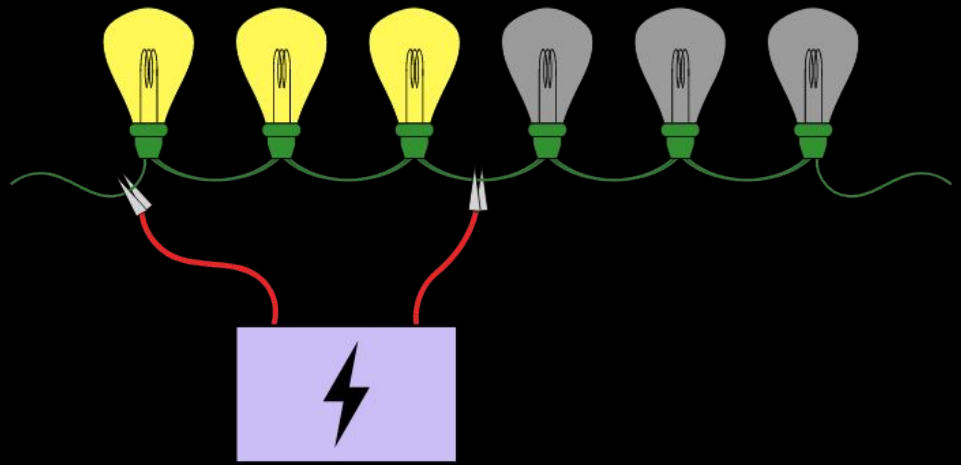
Result

From here, split components

Average messages per news article: **57** (vs. 343)

Group testing

Problem of using a small number of tests to identify positive samples in a large population.



Improved algorithm 2: Binary splitting *(from group testing)*

工人抗议很少见，或者会被很快平息。但这种模式很难用到其他国家，富士康必须适应不同国家的社会、政治和劳动条件。在巴西，富士康的计划很快**崩溃**了。曾争取该公司的政府很快在腐败指控和弹劾投票中下台。由于当地的经济增长和消费支出大幅下滑，政府承诺过的一些税收优惠被减少或取消。投资人兼初创技术公司顾问阿尔贝托·莫伊尔表示，富士康“举行了盛大的歌舞表演，带来**中国**舞龙队，还举行了剪彩仪式、祝酒会，以及惯常的合同草案的签字仪式。后来，在实际操作中，只有很小一部分得到实现”。莫伊尔不久前一直是调查公司桑福德·伯恩斯坦公司的技术分析师。富士康在一项声明中表示，他曾打算在中国之外投资数十亿美元修建工厂。但该公司还说，他必须适应巴西等市场不断变化的情况——那里的经济出现了停滞。“鉴于这些情况，再加上我们提议投资所服务的客户的需求在不断变化，我们目前缩减了在该国的运营规模，”

CENSORED

Use binary search to find leftmost sensitive character

Improved algorithm 2: Binary splitting *(from group testing)*

工人抗议很少见，或者会被很快平息。但这种模式很难用到其他国家，富士康必须适应不同国家的社会、政治和劳动条件。在巴西，富士康的计划很快**崩溃**了。曾争取该公司的政府很快在腐败指控和弹劾投票中下台。由于当地的经济增长和消费支出大幅下滑，政府承诺过的一些税收优惠被减少或取消。投资人兼初创技术公司顾问阿尔贝托·莫伊尔表示，富士康“举行了盛大的歌舞表演，带来**中国**舞龙队，还举行了剪彩仪式、祝酒会，以及惯常的合同草案的签字仪式。后来，在实际操作中，只有很小一部分得到实现”。莫伊尔不久前一直是调查公司桑福德·伯恩斯坦公司的技术分析师。富士康在一项声明中表示，他曾打算在中国之外投资数十亿美元修建工厂。但该公司还说，他必须适应巴西等市场不断变化的情况——那里的经济出现了停滞。“鉴于这些情况，再加上我们提议投资所服务的客户的需求在不断变化，我们目前缩减了在该国的运营规模，”

CENSORED

Remove all text preceding the character identified

Known Sensitive Characters

崩

DELIVERED

Improved algorithm 2: Binary splitting *(from group testing)*

工人抗议很少见，或者会被很快平息。但这种模式很难用到其他国家，富士康必须适应不同国家的社会、政治和劳动条件。在巴西，富士康的计划很快崩**溃**了。曾争取该公司的政府很快在腐败指控和弹劾投票中下台。由于当地的经济增长和消费支出大幅下滑，政府承诺过的一些税收优惠被减少或取消。投资人兼初创技术公司顾问阿尔贝托·莫伊尔表示，富士康“举行了盛大的歌舞表演，带来**中国**舞龙队，还举行了剪彩仪式、祝酒会，以及惯常的合同草案的签字仪式。后来，在实际操作中，只有很小一部分得到实现”。莫伊尔不久前一直是调查公司桑福德·伯恩斯坦公司的技术分析师。富士康在一项声明中表示，他曾打算在中国之外投资数十亿美元修建工厂。但该公司还说，他必须适应巴西等市场不断变化的情况——那里的经济出现了停滞。“鉴于这些情况，再加上我们提议投资所服务的客户的需求在不断变化，我们目前缩减了在该国的运营规模，”

CENSORED

Remove all text preceding the character identified

Known Sensitive Characters

崩 溃

DELIVERED

Improved algorithm 2: Binary splitting *(from group testing)*

工人抗议很少见，或者会被很快平息。但这种模式很难用到其他国家，富士康必须适应不同国家的社会、政治和劳动条件。在巴西，富士康的计划很快崩溃了。曾争取该公司的政府很快在腐败指控和弹劾投票中下台。由于当地的经济增长和消费支出大幅下滑，政府承诺过的一些税收优惠被减少或取消。投资人兼初创技术公司顾问阿尔贝托·莫伊尔表示，富士康"举行了盛大的歌舞表演，带来中国舞龙队，还举行了剪彩仪式、祝酒会，以及惯常的合同草案的签字仪式。后来，在实际操作中，只有很小一部分得到实现"。莫伊尔不久前一直是调查公司桑福德·伯恩斯坦公司的技术分析师。富士康在一项声明中表示，他曾打算在中国之外投资数十亿美元修建工厂。但该公司还说，他必须适应巴西等市场不断变化的情况——那里的经济出现了停滞。"鉴于这些情况，再加上我们提议投资所服务的客户的需求在不断变化，我们目前缩减了在该国的运营规模，"

CENSORED

Known Sensitive Characters

崩 溃 中 国

CENSORED

Continue until known sensitive characters are censored.

Split components

Binary splitting

Average messages per news article: **70**

Worse than binary search with backtracking (57).

Why?

- Samples are mutually dependent
- Components are not considered until after

Improved algorithm 3: Component-aware binary splitting

Use binary search to identify first character of component only, then linearly search following text to identify component

Improved algorithm 3: Component-aware binary splitting

Previously identified component(s)	Component to isolate	Remainder of text
	a	bcdefghijkl
	ab	cdefghijkl
xyz +	abc +	defghijkl
	abcd	efghijkl

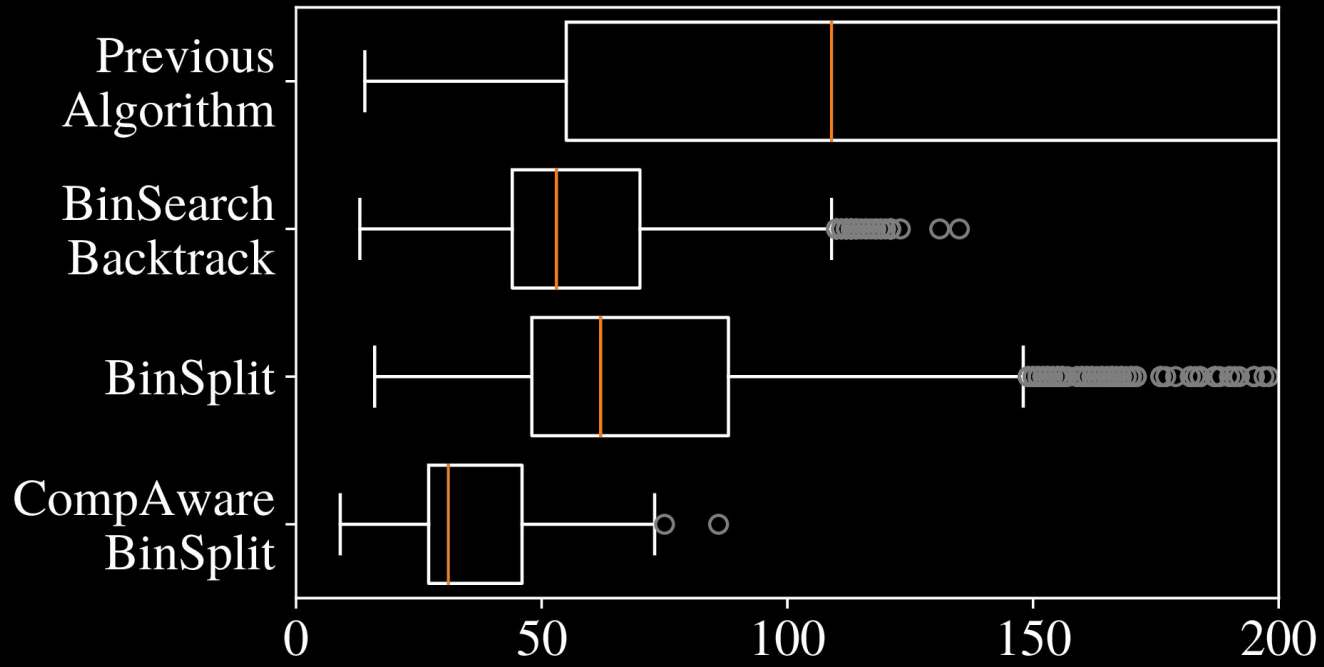
Improved algorithm 3: Component-aware binary splitting

Previously identified component(s)	Component to isolate	Remainder of text
	a	bcdefghijkl
	ab	cdefghijkl
xyz +	abc +	defghijkl
	abcd	efghijkl

Previously identified component(s)	Component to isolate	Remainder of text
	a	bcdefghijkl
	ab	bcdefghijkl
xyz +	abc +	bcdefghijkl
	abcd	bcdefghijkl

- Overlapping keywords detected

Results



Tukey box plot of # messages required to isolate a sensitive keyword combination across tested articles *(lower is better)*

Results

Average-case performance, CompAwareBinSplit: **35.47** messages

10.3% of messages required for previous algorithm

Conclusion

We present improved algorithms for isolating keyword combinations from censored text

Generalizability for other applications and languages

Questions?

Source code available at:

https://github.com/citizenlab/censored-keyword-isolation