The Effect of Censorship Circumvention on Information Transmission

Key Findings

- This report investigates the difficulty of replicating user-generated censorship circumvention techniques in China, their effectiveness in evading censorship, and their effect on information transmission.
- There exists great variability in the circumvention techniques 'adoption difficulty, circumvention effectiveness, and informational cost.
- There is generally a trade-off between circumvention effectiveness and information transmission - to be effective in circumventing censorship generally entails altering the original text to a greater extent but doing so can pose a greater challenge for reading comprehension of the altered text. Using specialized tools or language to alter text also increases the adoption difficulty for users.
- Literacy with censorship circumvention methods tends to correlate with younger age groups, exposure to foreign social media, and higher education levels.
- Overall, the Homophone Substitution has the best combination of adoption difficulty, circumvention effectiveness, and information transmission among the circumvention methods we considered in our study. However, interviews with engineers reveal a note of caution for the method.

Introduction

The cat-and-mouse game of censorship and circumvention by the Chinese government and Chinese netizens has long been a point of focus for media experts, researchers, and policy makers. Yet, the COVID-19 pandemic has brought the ingenious efforts to skirt China's ever-tightening censorship into the spotlight. The effort was epitomized by the collective movement of Chinese internet users to circumvent the censorship of an article written by Wuhan's director of the emergency department, Ai Fen, that criticized how the government handled the coronavirus epidemic[1]. With the original article converted into Morse code, emojis, oracle bone scripts, and the fictional elvish language created by author J.R.R. Tolkien, the article was shared widely on the Chinese internet despite heavy censorship. While this appeared to be a win for the people, many of the arcane scripts that the article was converted into left people unable to decipher the message that the article

[1] For coverage of the event, see e.g. "Skirting online censorship in China by "translating" a banned article into Morse, hexadecimal code, Emoji and Elvish language." cApStAn. tinyurl.com/33ymbfmu

tried to convey. As the event unfolded, the original intention to convey the message of the article quickly gave way to a form of entertainment in which people found ever creative ways to rewrite the article, leaving the article's meaning lost in translation.

Purpose of Study

How does censorship circumvention affect information transmission? To evade censorship, sensitive texts are often modified to conceal their true meanings. However, such modification risks distorting or erasing the original message. Given such risk, what is the most effective method of online resistance to censorship that maximizes censorship circumvention and information transmission? Despite substantial literature that tries to understand the logic of censorship and design methods to circumvent it, we lack systematic knowledge about the informational effect of censorship circumvention - the effect on information transmission when texts are heavily modified to evade censorship.

To fill the gap in the literature, we first identify three dimensions on which existing circumvention methods vary: technical complexity, circumvention effectiveness, and the degree of text distortion. Based on these three dimensions, we hypothesize that there may be a trade-off between censorship circumvention and information accessibility - using sophisticated technologies may be more effective at preventing the censor from detecting and deleting the posted content, but this can come at the cost of masking the true intent of the content, thus reducing the size of the message's intended audience. Additionally, technological complexity decreases the ability of average users to adopt, replicate, and modify these technologies, thus reducing users' ability to adapt to changes within a censored environment. As the three dimensions interact, we expect users' access to information is maximized when the circumvention method is effective at transmitting information. As the method becomes more technically sophisticated or the distortion of text becomes greater, users' overall access to information diminishes, even though the method may be more successful at evading censorship.

Here, we summarize the heuristic approach used to evaluate censorship circumvention methods in this study:

1. A censorship circumvention method needs to be effective at bypassing censorship
2. A censorship circumvention method needs to retain and convey the original information
3. A censorship circumvention method needs to be easily adoptable by users

We use a combination of interviews, experiments, and a nationwide survey in China to systematically test the adoption difficulty, circumvention effectiveness, and informational cost of existing circumvention techniques. We first identify a catalogue of user-created censorship circumvention techniques commonly used across the Chinese internet from a variety of sources. We then gauge the adoption difficulty of these techniques by asking internet users in China to replicate the techniques by using resources available to their daily life. To assess the effectiveness of these techniques at circumventing censorship, we construct a dataset of sensitive posts from the Chinese microblogging website Weibo and process the posts using these techniques. We then employ two commercial censorship services by Chinese technology companies Baidu and Tencent to evaluate how successful each technique is at bypassing the censors. In addition, we also interview engineers from a major Chinese technology company to gain insight from the perspective of the censors on how difficult it is to handle each of the techniques and if existing censorship technologies have already considered such techniques. Finally, we test the techniques' effectiveness in transmitting information using a nationwide survey in China in which we ask the respondent to recover the original message from processed posts.

Cataloguing Censorship Circumvention Techniques

We rely on a variety of sources from news reports, academic articles, GitHub repositories, and expert knowledge to identify existing circumvention methods. Table 1 presents the list of six commonly used circumvention methods, along with a short description of each method and a corresponding example[2].

The original text of the example we use in the table is "今天天气真好", which translates into "The weather today is great". We treat the word 天气 (weather) as the censored word we want to modify. For the emoji method, we simply convert the word "weather" into its corresponding emoji. For homophone substitution, we convert the censored word into its homophone, 田七, which has similar phoneme (and the same pinyin) as the Chinese word for weather. The Huoxing/Martian method converts the censored word into a popular Chinese internet language that overwhelmingly uses arcane and old Chinese characters. In this case, the censored word is converted into two Chinese characters that visually look

[2] Example for the Image method is available in the Appendix. Note that for our study, we focus on circumvention methods that are commonly used on the Chinese internet and that are generally readily available to average internet users. Therefore, we do not consider methods that are little used (such as converting censored word into oracle bone scripts) or more computationally involved methods that we believe are out of reach of the average internet user.

similar to 天气 but do not share the same meaning nor phoneme. The masking method masks the censored word with symbols such as an asterisk (*). The image method first converts the entire text into an image. Users can then add drawings or rotate the image to make it harder for the censor to detect the actual message with OCR. The last method reorders the censored word, in this case 天气 becomes 气天.

Gauging the Adoption Difficulty of Circumvention Methods

Here, we present results on the adoption difficulty of the circumvention methods listed in Table 1. We asked eleven Chinese internet users in China, aged 22 to 52, to replicate each of the six circumvention methods. Specifically, we show each user two examples of text that have been processed by the circumvention methods and ask them if they can replicate the methods on a third example. Table 2 summarizes the replication success rate for each method.

TABLE 1. LIST OF COMMONLY USED CENSORSHIP CIRCUMVENTION METHODS

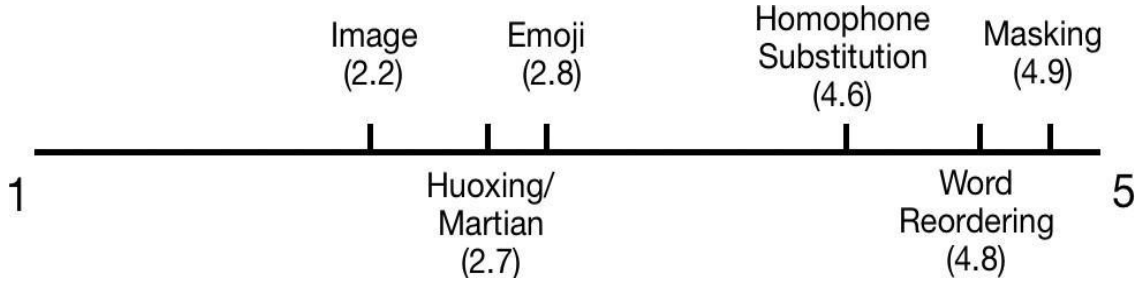
Method	Description	Example
Emoji	Convert censored word/sentence into emojis	今天  真好
Homophone Substitution	Convert censored word/sentence into its homophones	今天田七真好
Huoxing/Martian	Convert censored word/sentence into Martian(火星文), a Chinese internet language	今天迭芶真好
Masking	Use symbols such as (*) to mask out censored word	今天 ** 真好
Image	Convert censored word/sentence into images (and add drawings)	
Word Ordering	Reorder censored word/sentence	今天气天真好

TABLE 2. REPLICATION RATE OF CIRCUMVENTION METHODS

Method	Successful Replication Rate
Emoji	7/11
Homophone Substitution	11/11
Huoxing/Martian	5/11
Masking	11/11
Image	5/11
Word Ordering	11/11

After each replication, we also ask users to rate the difficulty in replicating each of the methods on a 5-point Likert scale, with 1 being very difficult and 5 being very easy. Figure 1 presents the average rating for each method.

Figure 1. Replication Difficulty



It is easy to observe that there are two groups of methods based on adoption difficulty. Emoji, Huoxing/Martian, and Image methods have a lower successful replication rate and users generally rate them as somewhat difficult to replicate. In contrast, Homophone Substitution, Masking, and Word Reordering methods have successful replication rates and users generally rate them as very easy. The exercise found that for methods such as Homophone Substitution, Masking, and Word Reordering, users were able to replicate them without additional resources. For Emoji and Huoxing/Martian, however, most

successful replications require users to search for websites (e.g. Huoxing/Martian conversion website) that provide such service. For Image, it requires users to be able to turn the text into an image with a white background and add drawings to it. Those that fail to replicate these methods are either not able to find relevant resources online or report that it is too much trouble to bother trying harder.

The exercise also found high correlations between the age of the user, the replication rate, and the difficulty ratings of the methods, with younger users being much more able to replicate and generally found the exercise to be relatively easier. During interviews with the users after the exercise, most of the younger users expressed experience with the circumvention techniques while the older users had limited experience.

While the results show significant difference among circumvention techniques in terms of their replication difficulty, it is important to note that the results are based on a small sample of users in China. The small sample size is primarily due to the difficulty of getting respondents in China, as users are reluctant to participate when the topic involves censorship.

Testing Censorship Circumvention Effectiveness

To test the methods' effectiveness in circumventing censorship, this study opted to use two commercial censorship services in China: the Baidu censorship and Tencent censorship APIs. Based on information from the companies' websites, the censorship APIs have been used by several other Chinese companies. During interviews with engineers in China, it was found that such censorship algorithms are usually the first line of defense in the censorship scheme. Therefore, this is a somewhat realistic test of censorship circumvention effectiveness.

The study used a dataset of sensitive texts scraped from Weibo for this task. For each Weibo post, it is transformed by using the list of censorship circumvention methods. The test included both the original posts as well as the transformed posts with the two censorship APIs. Figure 2 presents result from the Baidu censorship API, with the y-axis indicating the probability that the content is politically sensitive. Figure 3 presents results from the Tencent censorship API, with the proportion of content receiving each of the three decisions {*Pass, Review, Block*} listed for each method.

Figure 2. Results from Baidu Censorship API

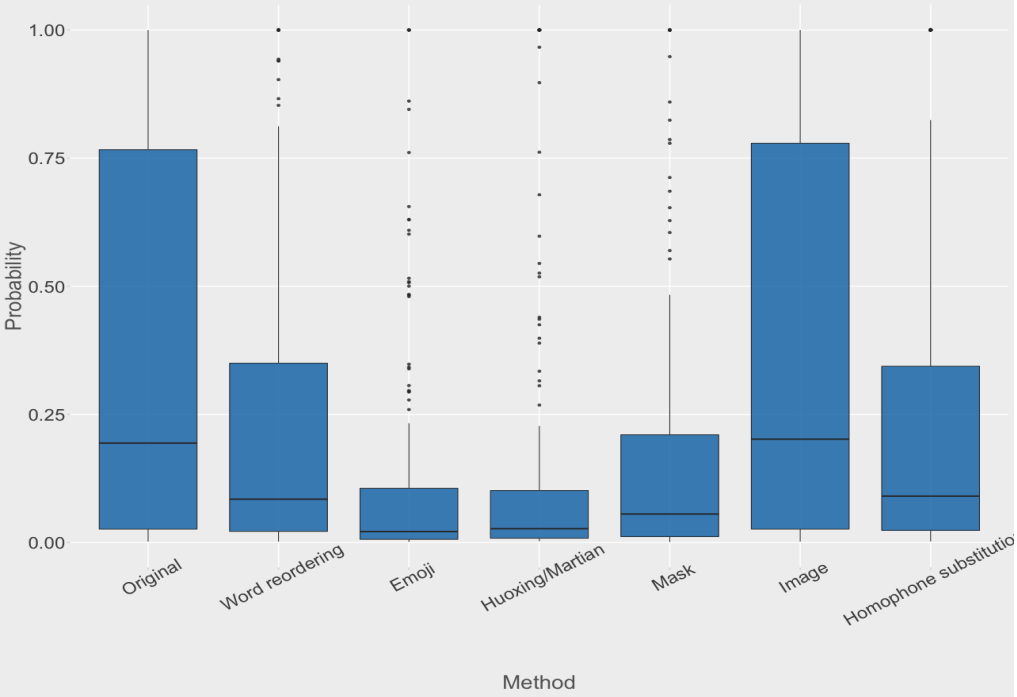
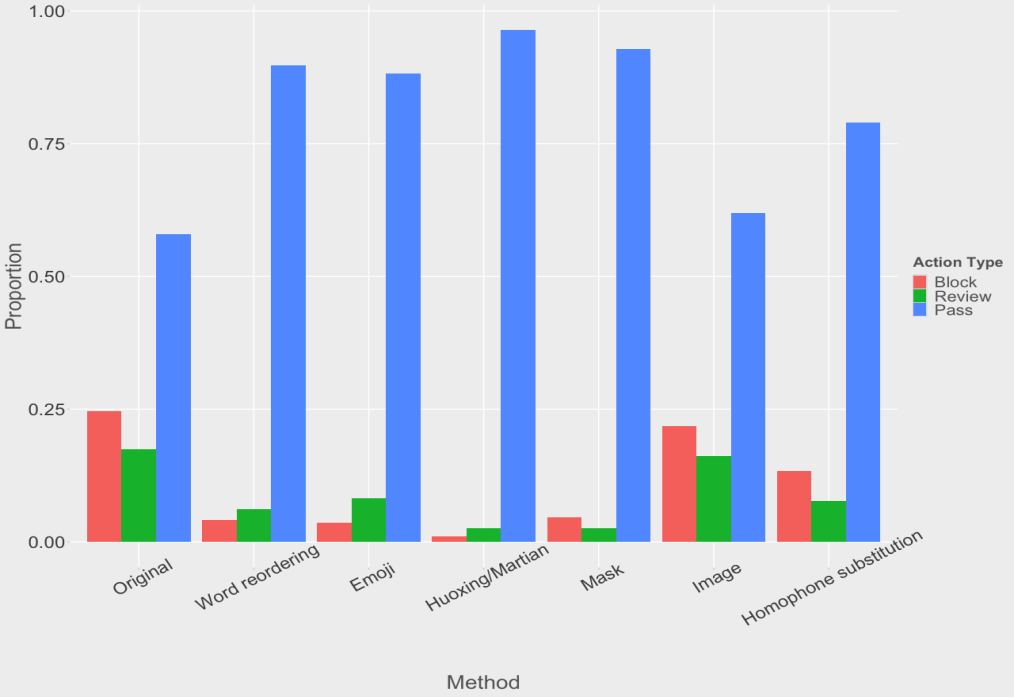


Figure 2 shows the circumvention methods display varying abilities to bypass the censorship API. Specifically, the Emoji and Huoxing/Martian methods are the most effective, reducing the average probability by more than 60%. The Word Ordering, Mask, and Homophone Substitution methods are also moderately effective, reducing the average probability by around 35% − 57%. Surprisingly, the Image method performed very poorly, showing almost no improvement over the original posts.

A similar pattern was seen from the Tencent censorship API result in Figure 3. The Word Reordering, Huoxing/Martian, and Mask methods are able to increase the pass rate substantially. Emoji and Homophone Substitutions are also moderately effective. The Image method fails to show much improvement over the original posts.

Figure 3. Results from Tencent Censorship API



Several engineers were interviewed for plausible explanations of the results. Emoji and Huoxing/Martian methods are commonly used in posts that are advertisements or scams and rarely seen in political posts. Therefore, it is possible that the censorship APIs do not yet have a good political terms-emoji dictionary. For mask, because all sensitive words are masked by symbols and there is no direct correspondence between political terms and symbols, this is a difficult task for the censorship APIs. However, both interviewees suspect that masking can substantially disrupt the transmission of the message of the post.

For Homophone Substitution, which has been shown to yield good performance in prior research (Hiruncharoenvate, Lin and Gilbert, 2015), its less-than-impressive results can likely be explained by the fact that one of the data augmentation method in censorship training is converting posts to the pinyin (sound in Romanized letters) of individual characters. The data augmentation method thus directly tackles the kind of adversarial attacks based on Homophone Substitution, which replaces sensitive words with words of the same pinyin. Similarly for the Image method, a common data augmentation technique in censorship training is image rotation and adding noise (e.g. drawings) to images. Because

the Image method retains the original posts, successful OCR by the censorship APIs could easily break the method. This is observed for the Tencent censorship API which almost always recovers the original posts from the images.

Effect on Information Transmission

To test the informational effect of censorship circumvention, the study conducted a nationwide survey in China in which respondents were asked to recover the original message from posts that are altered by circumvention methods. Specifically, the original messages were transformed with the list of aforementioned circumvention methods. Each respondent was presented with two transformed messages randomly chosen from the pool of messages and was asked to guess and type the original message. If the respondents could not infer the original message, they were asked to type "I don't know". In addition to this exercise, respondents were also asked questions about their demographics, education level, and media consumption pattern.

Figure 4. Geographical Distribution of Survey Respondents

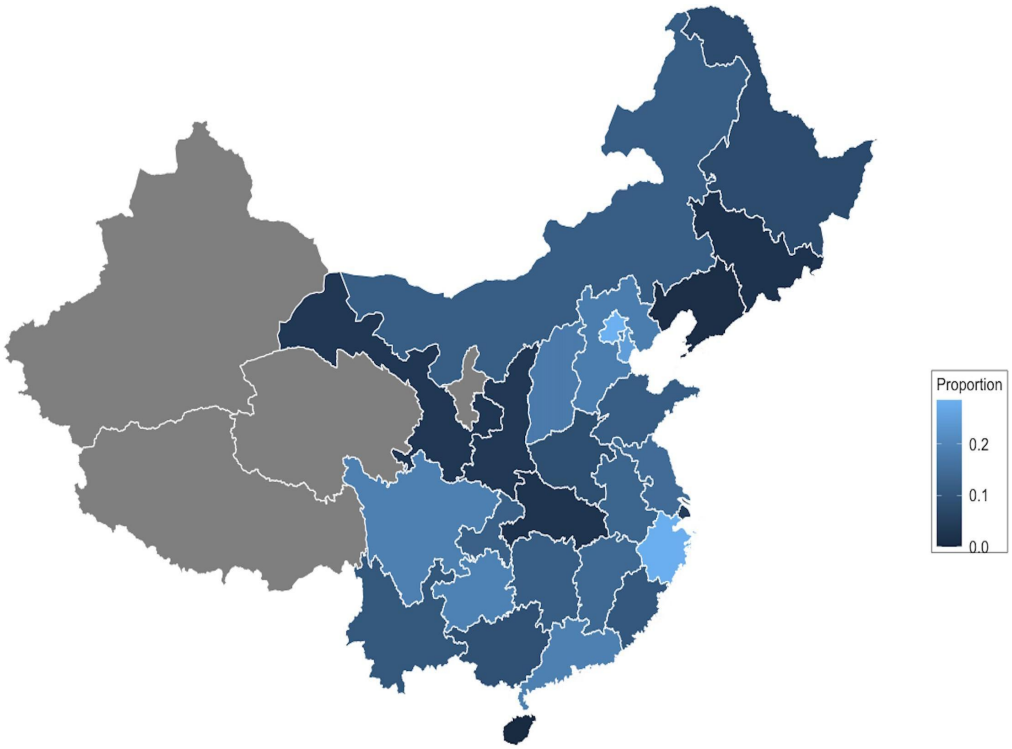
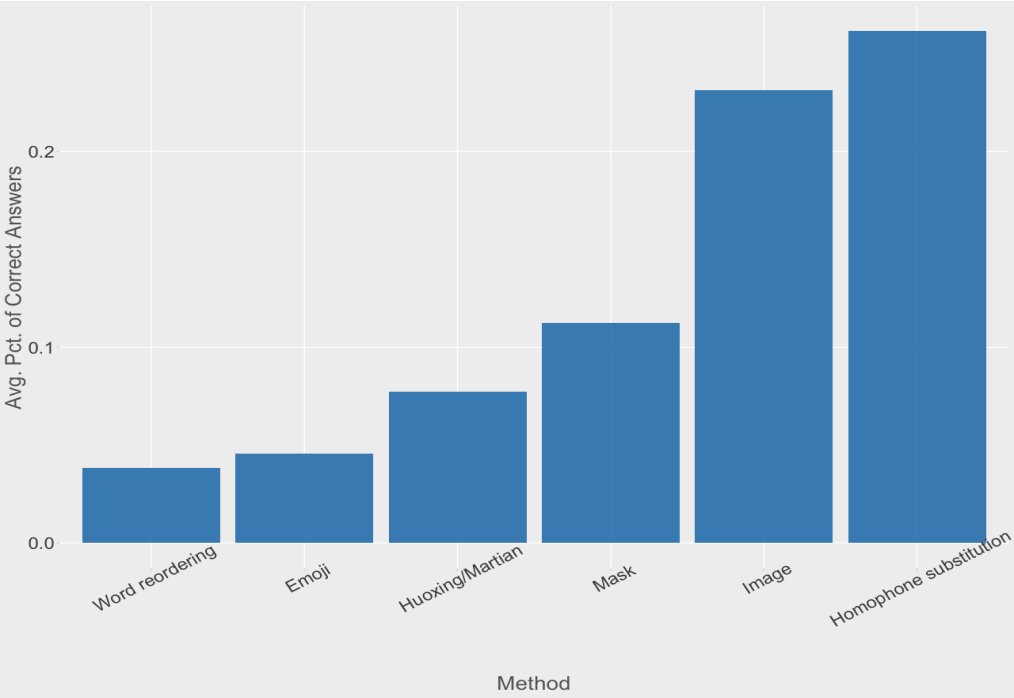


Figure 4 presents the proportion of responses that successfully recovered the original message, with lighter shades of blue representing higher proportions. Because the survey provided no monetary incentive for the survey respondents to have the correct guesses, the average proportion for each province is quite low. However, there was still great variation across China. Although the sample for each province is small, we can see that, in general, respondents from coastal regions of China are able to recover the original message at a higher rate than those from more inland regions. The next section presents more evidence on the factors that correlate with higher recovery rate.

Figure 5. Recovery Rate for Censorship Circumvention Methods



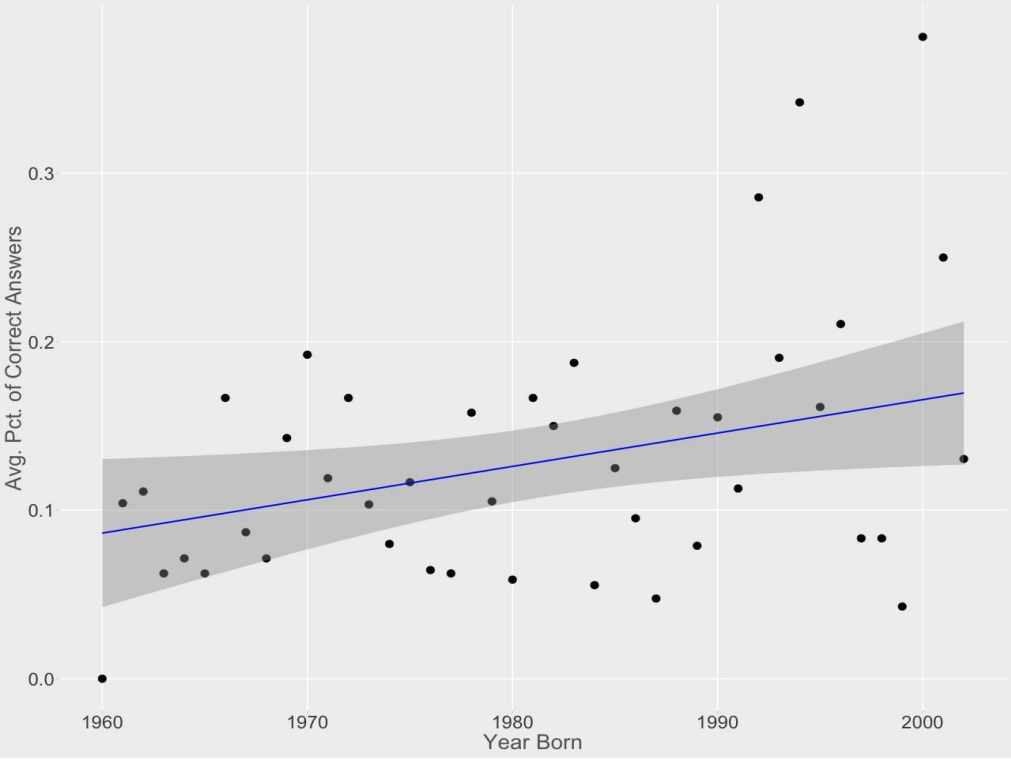
In both interviews with internet users and engineers in China, interviewees expressed concern that inferring the meaning of emojis and Huoxing/Maritian scripts as well as reorganizing word orders are disruptive to their reading comprehension. Additionally, inferring masked words, especially when there are multiples in a sentence, from a short context can be challenging. There are shortcomings of these methods in transmitting information. Figure 5 presents a bar chart of the recovery rates of different censorship circumvention methods, with higher bar representing a higher recovery rate and more effective information transmission. For methods such as Word Reordering, Emoji, and Huoxing/Martian, which are effective in circumventing the censorship APIs, the

information loss in the transmission stage is quite high, with the recovery rates less than one quarter of the most effective method. In contrast, the Image and Homophone Substitution methods enjoy much higher recovery rate.

Factors that Correlate with Recovery Rate

Several factors that correlate with recovery rate to help readers better understand how existing censorship circumvention methods have differential impact of different groups of internet users in China. As mentioned in the section on adoption difficulty, younger internet users seem more able to replicate circumvention methods. We see a similar pattern with understanding texts transformed by such methods. Figure 6 presents the correlation between the years the respondents were born and the average recovery rate for the age group. This shows a positive correlation between higher recovery rate and respondents who were born later. As methods such as Emoji and Huoxing/Martian use scripts that were popularized by younger internet users in China, it should come as no surprise that this age group is more able to recover messages transformed by such methods.

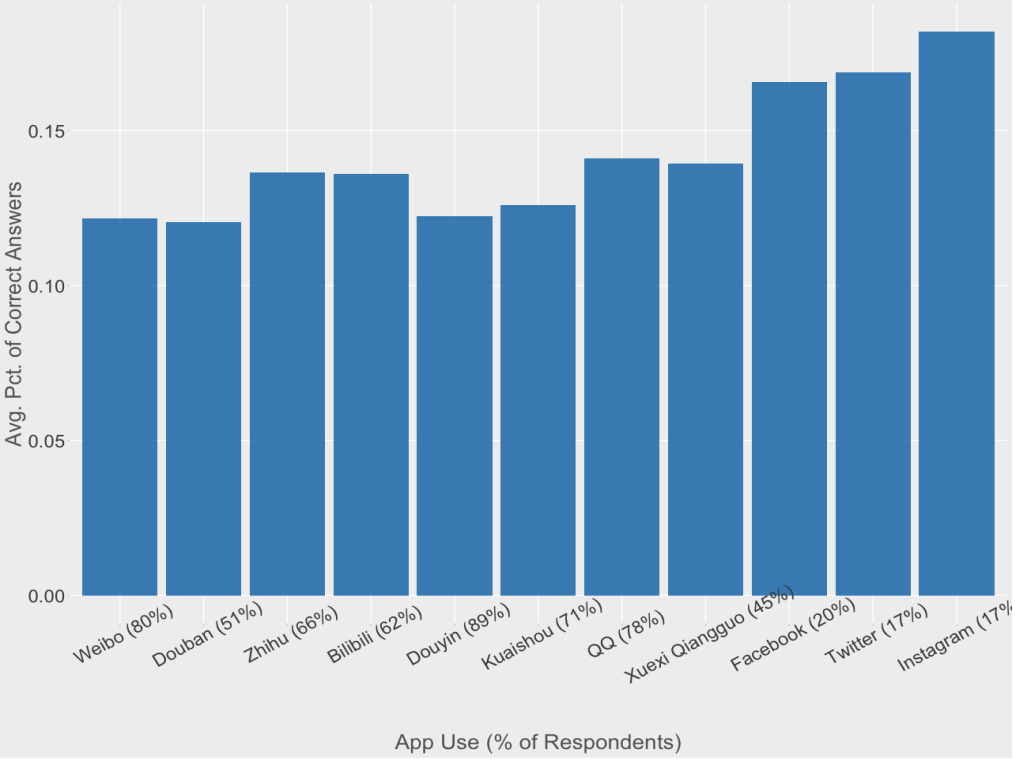
Figure 6. Correlation between Age and Recovery Rate



Next, we look at the correlation between respondents' media consumption pattern and

recovery rate. Figure 7 presents the bar chart on the respondents' social media use and recovery rate. The most striking pattern is of respondents who have had access to foreign social media, such as Facebook, Twitter, and Instagram. These respondents tend to have higher recovery rates compared to respondents who have not had such access. Intuitively, respondents who have access to foreign social media have direct experience with bypassing internet restrictions, and the desire to access restricted websites can signal a correlation with higher exposure to censored information and censorship circumvention methods. This is in line with a recent review (Roberts, 2020) that suggests resistance to censorship requires demand on the part of the users to access censored information.

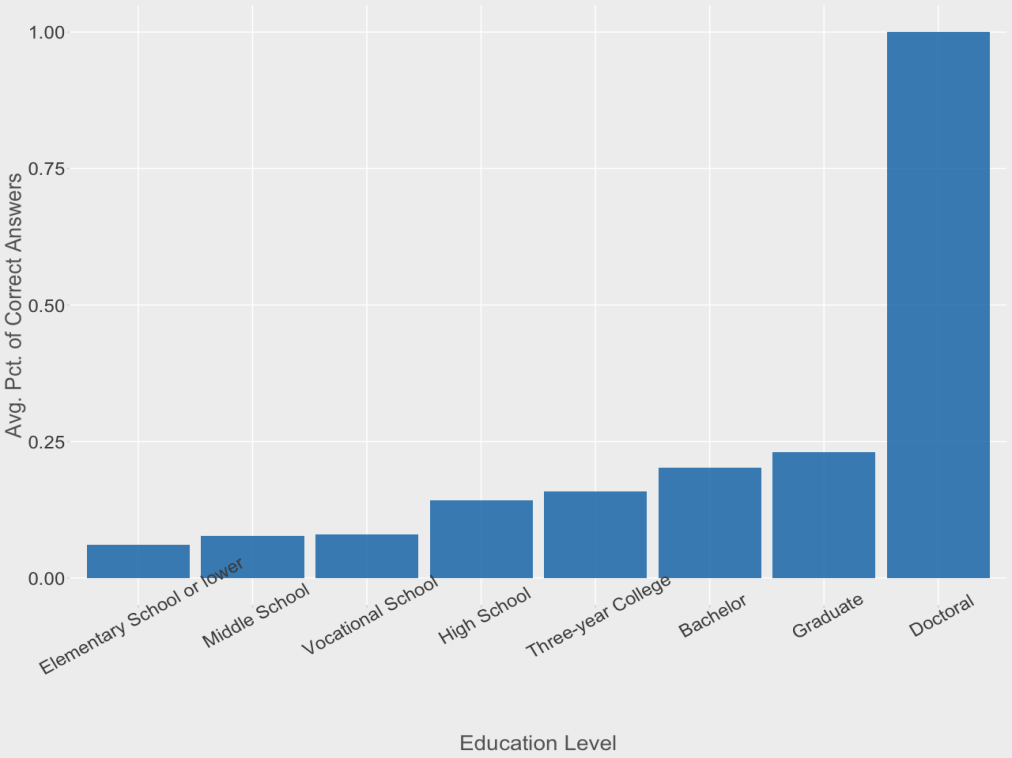
Figure 7. Correlation between Media Consumption and Recovery Rate



Finally, we look at the correlation between respondents' education level and recovery rate. Figure 8 shows a positive correlation between respondents' education level and recovery rate. We have a perfect recovery rate for those with doctoral degrees, likely due to a small sample (n=2). Despite the outlier, the positive correlation is rather clear. Although future work should investigate more into the correlation, one possible explanation for the pattern is that higher education tends to correlate with more liberal political ideology in China (Pan and Xu, 2018). Therefore, respondents who are more educated may have more

experience with censored information and censorship circumvention methods.

Figure 8. Correlation between Education Level and Recovery Rate



Discussion

Based on interviews with Chinese internet users and the survey experiment, there are substantial variations in terms of the adoption difficulty, censorship circumvention effectiveness, and information transmission for existing circumvention methods. Overall, our study suggests that there is a plausible trade-off between circumvention effectiveness and information transmission - to be effective in circumventing censorship generally entails altering the original text to a greater extent but doing so can pose a greater challenge for reading comprehension of the altered text. Additionally, using specialized tools or language to alter text can increase the adoption difficulty for users. This study also finds that literacy with censorship circumvention methods tends to correlate with younger age groups, exposure to foreign social media, and higher education levels.

Overall, Homophone Substitution tended to have the best combination of adoption difficulty, censorship circumvention effectiveness, and information transmission among the circumvention methods we considered in the study. However, interviews with Chinese engineers also reveal a note of caution for this method - existing data augmentation

procedure used by Chinese tech companies can increase the robustness of censorship systems against Homophone Substitution. Future researchers and developers must take into account the informational effect of censorship circumvention methods and develop methods that are not only effective against censorship but also facilitate adoption and comprehension on part of the users.

One limitation of the study is that it did not consider using combinations of circumvention methods. Although altering text with more than one method may be more effective against censorship, this can greatly increase the adoption difficulty for users, as it requires the ability to replicate multiple methods. However, future research and software development may provide a centralized resource that takes advantage of multiple censorship circumvention methods while reducing the start-up cost for users. Another limitation of the current study is that it did not consider a new form of censorship resistance in China, in which people post slogans from the Chinese Communist Party or quotes from famous Chinese writers and party leaders on censored topics[3]. For topics that are heavily censored, slogans and quotes that are endorsed by the Chinese Communist Party become the only viable content that can be posted under such topics. While doing so loses any ability to convey specific messages, it serves as a powerful signaling function that alerts other users that the given topic is under heavy censorship. Given the rising prevalence of this phenomenon and the ever-tightening censorship on the Chinese internet, future studies should investigate the consequences of this form of censorship resistance.

[3] For examples of this phenomenon, see the (now censored) Weibo account of 咸亨酒店遗址 that used to post quotes from Lu Xun, a famous Chinese writer, that satirize current affairs in China. See also comments under the Weibo account of Tennis player Peng Shuai when her account was subject to heavy censorship during her allegation of sexual assault from a former high-ranking Chinese official.

Appendix

Example of the Image Method

Figure .9. Example of the Image Method

References

Hiruncharoenvate, Chaya, Zhiyuan Lin and Eric Gilbert. 2015. Algorithmically bypassing censorship on sina weibo with nondeterministic homophone substitutions. In *Ninth International AAAI Conference on Web and Social Media*.

Pan, Jennifer and Yiqing Xu. 2018. "China' s ideological spectrum." *The Journal of Politics* 80(1):254–273.

Roberts, Margaret E. 2020. "Resilience to online censorship." *Annual Review of Political Science* 23:401–419.