# When to Censor

Yunyun Lv

October 7, 2022

Abstract

The censor wants to maximize the receiver' s belief that he is a good type instead of a bad type at an exogenous deadline. The censor chooses when to stop the public discussion about a piece of news, which generates a learning process for the receiver about the censor' s type. If the censor allows more learning, the receiver has a more accurate belief about the type of censor. In equilibrium, if the censor stops the learning process sooner, the receiver believes he is more likely a bad type.

## 1 Introduction

A censor examines contents and suppresses any parts that are considered unacceptable. Despite the fact that it is nearly impossible to censor every piece of information on the Internet while it barely costs anything to duplicate some digital contents, Internet censorship is pervasive all over the world. One obvious reason is that social networking service providers have to comply with the laws of the countries where their users are. For example, Twitter may remove contents deemed to be illegal or objectionable after receiving complaints or requests from third parties, including government officials, companies and other outside parties.

What contents need to be removed? The answer varies across countries since the criterion for removal of contents is often defined by the government: In France it could be antisemitic contents or hate speech; In Indian it could be parodies of the ruling elite; In South Korea it could be tweets by North Korea government; In China it could be the social network itself or SOS posts by desperate Weibo users in the sudden lock-down of Wuhan in January 2020. Besides censorship under overt pressure, censorship by social networking service providers also exists. For example, thousands of ISIS-related twitter accounts were suspended in 2015. Social media platforms in China hire thousands of inspectors to manually remove offensive contents to avoid being punished by the government.

However, censorship may backfire if people know it exists. Noticing that a keyword is being blocked often makes people think about why it is being blocked. For example, lots of Chinese people suddenly found out that the Chinese character 翠, which refers to the color emerald green, was added to the list of sensitive keywords soon after the outbreak of Covid-19. It does not take a lot of thought to figure out why: The character 翠 contains two characters: 习, the last name of Xi Jinping, and 卒, which means to die. The outbreak of Covid-19 was followed by rapid rise of public dissatisfaction with the Chinese government. A small group of people started to express their anger toward Xi in a creative way: They used the hashtag # 祈翠 (pray for 翠/the death of Xi) on multiple SNS platforms, whereas the general public had no idea of this practice at all until they found out that any post containing the character 翠 is blocked. Unfortunately 翠 is a commonly used character in Chinese and even a popular character used in names.[1] Nowadays, most active Chinese Internet users are aware of the new special meaning of this character and more importantly, the irritability of Xi.

This paper explains the difference among the censorship strategies used by

[1] Later the screening algorithm was modified to allow names that contains 翠.

different governments when the choice of different censorship strategies itself can be informative. Gratton et al. (2018) explains October Surprises by focusing on how the starting time of the learning process is strategically chosen by Sender. In their model Sender has private information about his type and also a private piece of information that would trigger a public learning process. This paper looks at the opposite situation: If a learning process has started, when does Censor want to stop it? Censor privately knows his own type, but has no control over how and when the public learning process starts. Instead, they can stop a learning process after it starts. The outbreak of a scandal, for example, can trigger public discussion of a new topic, toward which the public does not yet know the government's attitude. At the beginning of 2022, a video of a chained woman, mom of seven boys and a girl, and "wife" of a local man in the countryside of Xuzhou went viral on TikTok in China. Netizens soon found out that she was likely a victim of human trafficking decades ago. Local propaganda department in Xuzhou first claimed that their investigation shows that the marriage is legal and there is no human trafficking involved. The public was not convinced at all, and decided to investigate themselves. The government then set up another investigation team and, at the same time, started to obstruct voluntary investigators. Two female volunteers were detained by local police while investigating the case and trying to help the chained woman. Online discussion about contradictions among different reports by the authorities were strictly censored. Chinese citizens, especially Chinese women, started to question whether the Chinese Communist Party (CCP) care about the well-being of Chinese women at all after observing the constant disappearance of discussions about the chained woman. After all, the slogan "妇女能顶半边天", translated as "women hold up half the sky", was an important part of the propaganda since late 1950s in China. Most Chinese women were still under the impression that CCP was not against feminism until they found out the government did not allow any more discussion of this chained woman.

The literature on censorship mainly consists of empirical analysis of how censorship is implemented in authoritarian countries (King et al., 2013; King et al., 2017; Roberts, 2018; Gallagher and Miller, 2019). Chen and Xu (2017) argue that there are two possible ways an authoritarian government can benefit from allowing public exchange of information among citizens: The government can learn about the dissatisfaction level of the public and amend policies accordingly; Citizens will be discouraged from protesting if they learn from public communication that other citizens have different opinions. This paper shows that, besides these reasons listed above, the authoritarian government mimics government of good type via randomization of censoring and allowing public exchange of information.

## 2 Model

Time horizon is discrete and finite, denoted by $t \in \{1, \ldots, T, T+1\}$. There are two players: a Censor and a Receiver. Censor privately knows his type, $\theta \in \{G, B\}$ (good type or bad type). Receiver's prior belief is denoted by $Pr(\theta = G) = \pi$, where $\pi \in (0, 1)$ is common knowledge. Censor wants to maximize Receiver's posterior belief that $\theta = G$ at a given date $t = T$, denoted by $s$. Notice that the time horizon includes a period $t = T + 1$ that captures all future periods after the exogenous deadline $T$.

A post exogenously arrives in the first period, $t = 1$. The model focuses on what would Censor of different types do after the post arrives. A learning process for Receiver is then triggered by the post. It can be interpreted as new information brought in by discussion and debates in the comment section of the post. Without loss of generality it is assumed that a sequence of signals are generated following a stochastic process

$$\mathcal{L} = \{L_\theta(t) | 1 \leq t \leq T\}.$$

Receiver stops receiving signals from the learning process if either the post is removed by Censor or a technical shock happens. An exogenous technical shock happens with probability $\varepsilon \in [0, 1)$ in every period. Censor can choose whether to remove the post in each period unless a technical shock has already happened.

Figure 1 shows the timeline if the post disappears at $t = \tau$, either because of a technical shock or because it is removed by Censor:

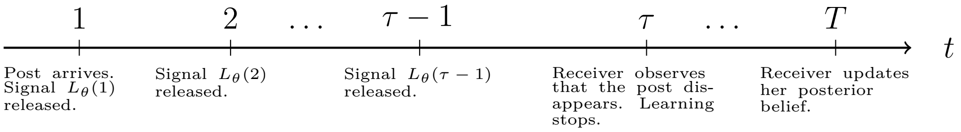
1
2
...
$\tau - 1$
$\tau$
...
$T$
$t$
Post arrives. Signal $L_\theta(1)$ released.
Signal $L_\theta(2)$ released.
Signal $L_\theta(\tau-1)$ released.
Receiver observes that the post disappears. Learning stops.
Receiver updates her posterior belief.

Figure 1: Timeline

Receiver combines two pieces of information to form her posterior belief $\alpha$ at $T$: (i) $\eta$, the interim belief that $\theta = G$ based on the the fact that the post disappears at $\tau$, and (ii) signals from the learning process that stops at $(\tau - 1)$. These signals follow the stochastic process $\mathcal{L}_\tau = \{L_\theta(t) | 1 \leq t \leq \tau - 1\}$.[2]

Denote by $H(\cdot|\tau, \eta)$ the distribution over Receiver's posterior belief, $\alpha$, generated by $\mathcal{L}_\tau$ conditional on her interim belief $\eta$. Similarly, denote by $H_\theta(\cdot|\tau, \eta)$ the distribution over Receiver's posterior belief, $\alpha$, generated by $\mathcal{L}_\tau$ conditional on her interim belief, $\eta$, and Censor's type, $\theta$. Notice that $H(\cdot|1, \eta)$ and $H_\theta(\cdot|1, \eta)$ are degenerate distributions that assign probability one to $\alpha = \eta$.

In equilibrium, Receiver updates her belief from $\pi$ to $\eta$ based on the equilibrium strategy of Censor. If the good Censor and the bad Censor use different

[2] Note that $\mathcal{L}_1 = \emptyset$, and $\mathcal{L}_{T+1} = \mathcal{L}$.

strategies, the timing of disappearance, $\tau$, is then informative. In other words, in non-raveling equilibria, Censor signals about his type via the timing of removal.

Assumption 1. (i) Every signal generated by $\mathcal{L}$ is informative (in Blackwell's sense). Removing the post sooner reveals strictly less information about Censor's type: $\forall \tau, \tau' \in \{1, \dots, T, T+1\}$ with $\tau < \tau'$, $H(\cdot|\tau', \eta)$ is a strict mean-preserving spread of $H(\cdot|\tau, \eta)$. (ii) The learning process never fully reveals Censor's type: The support of $H(\cdot|T, \eta)$ is a subset of $(0, 1)$ for any $\eta$.

## 3 Main Result

The solution concept is perfect Bayesian equilibrium. A typical strategy of Censor is denoted by $s_\theta \in [0, 1]^T$, where $s_\theta(t)$ represents the probability to remove the post if the post still exists in period $t$. Let $P_\theta$ and $p_\theta$ be the c.d.f. and density function of the disappearing time of the post given Censor's type $\theta$ in equilibrium. Note that as long as $\varepsilon \neq 0$, in every period $t$ either $p_\theta(t) > 0$, or $p_\theta(t) = 0$ and $s_\theta(\hat{t}) = 1$ for some $\hat{t} < t$.

In any equilibrium, Receiver's interim belief if the post disappears in period $\tau$ is given by $\eta = \mu(\tau)$ with

$$\mu(\tau) = \frac{\pi p_G(\tau)}{\pi p_G(\tau) + (1 - \pi) p_B(\tau)}.$$

Lemma 1. Let $\mathbb{E}[\alpha|\tau, \eta, \theta]$ denote the expectation of Receiver's posterior belief $\alpha$ conditional on the disappearing time $\tau$, her interim belief $\eta$, and Censor's type $\theta$. For any $\tau < \tau' \in \{1, \dots, T, T+1\}$ and any $\eta < \eta'$,

1. $\mathbb{E}[\alpha|\tau, \eta, \theta] < \mathbb{E}[\alpha|\tau, \eta', \theta]$ for $\theta \in \{G, B\}$;

2. $\mathbb{E}[\alpha|\tau, \eta, G] < \mathbb{E}[\alpha|\tau', \eta, G]$;

3. $\mathbb{E}[\alpha|\tau, \eta, B] > \mathbb{E}[\alpha|\tau', \eta, B]$.

All omitted proofs are in the Appendix. Censor always benefits from a higher interim belief, regardless of his type. For a fixed interim belief, more information for Receiver is detrimental for the bad Censor while beneficial for the good Censor.

Lemma 2. In any equilibrium:

- For all $\tau < \tau' \in \{1, \ldots, T, T+1\}$, if good Censor weakly prefers that the post disappears at $\tau$ than at $\tau'$, then $\mu(\tau) > \mu(\tau')$ and bad Censor strictly prefers that the post disappears at $\tau$ than $\tau'$.

- For all $t \leq T$, $s_G(t) = 0$, and hence

$$p_G(t) = \begin{cases} (1-\varepsilon)^{t-1}\varepsilon & t \leq T \\ (1-\varepsilon)^T & t = T+1 \end{cases}$$

Suppose good Censor weakly prefers that the post disappears at $\tau$ than at $\tau' > \tau$. It must be because Receiver's belief is more tilted toward $\theta = G$ upon observing that the post disappears at $\tau$ than at $\tau'$. Moreover, this gain in credibility must dominates the loss from less learning for the good Censor. Hence bad Censor should strictly prefer that the post disappears at $\tau$ than $\tau'$ because he benefits from less learning.

The second result is a corollary of the first one. Bad Censor must have already removed the post at $\tau$ if good Censor assigns positive probability to remove the post at $\tau$, i.e. for any $\tau \leq T$, $s_G(\tau) > 0$ implies $s_B(\tau) = 1$. Suppose $s_G(\tau) > 0$ for some $\tau \leq T$ in equilibrium, then $p_B(\tau') = 0$ for all $\tau' > \tau$ and hence Receiver can conclude that $\theta = G$ upon observing that the post disappears in any period $\tau' > \tau$. Therefore good Censor should strictly prefer that the post disappears at $\tau'$ than $\tau$, contradicting that $s_G(\tau) > 0$. As a result, good Censor never removes the post in any period $t \leq T$ in any equilibrium.

Lemma 3. In any equilibrium:

- If $\varepsilon = 0$, bad Censor never removes the post.
- If $\varepsilon > 0$, $s_B(t) < 1$ for any $t \leq T$ and $p_B(T+1) > 0$.
- If $\varepsilon > 0$, $s_B(t) > 0$ for any $t \leq T$.

If exogenous technical shock does not exist, given that the good Censor never removes the post, Receiver can conclude that the Censor is of the bad type upon observing removal of the post. Hence the bad Censor should strictly prefer to mimic the good Censor and never removes the post.

To understand the second result, suppose $s_B(t) = 1$ for some $t \leq T$ in equilibrium. Then the post never disappears after period $t$. In other words, for any $t' > t$, $p_B(t') = 0$. Since good Censor never removes the post, all information sets are on path. Therefore, Receiver believes that $\theta = G$ if the post disappears after $t$. Bad Censor is strictly better off by choosing $s_B(t) = 0$ instead. Therefore, in any PBE, $s_B(t) < 1$ for any $t \leq T$.

The third result is by induction on time. Since $s_B < 1$, with some positive probability the post never disappears, i.e. $p_B(T+1) > 0$. It then implies that bad Censor weakly prefers that the post disappears at $T+1$ than at period $T$, because otherwise he would choose $s_B(T) = 1$ instead. Recall that more learning hurts bad Censor, so the gain in credibility at $T+1$ must dominate the loss from more learning. Hence $s_B(T) > \frac{(1-\varepsilon)}{(1-\varepsilon)+1} > 0$. Suppose $s_B(t) > 0$ for some $t \leq T$. Then bad Censor weakly prefers that the post disappears at $t$ than $(t-1)$. Again, the gain in credibility must dominate the loss from more learning, so $s_B(t-1) > \frac{(1-\varepsilon)s_B(t)}{(1-\varepsilon)s_B(t)+1} > 0$. Then by induction, $s_B > 0$.

Lemma 4. In any equilibrium, for any $\tau \geq 1$,

$$\int \alpha dH_B(\alpha|\tau, \mu(\tau)) = \mu(1), \tag{1}$$

and

$$\sum_{\tau=1}^{T+1} \frac{1-\mu(\tau)}{\mu(\tau)} p_G(\tau) = \frac{1-\pi}{\pi} \tag{2}$$

Proposition 1. There exists an equilibrium in which

$$p_G(\tau) = \begin{cases} (1-\varepsilon)^{\tau-1}\varepsilon & \tau \leq T \\ (1-\varepsilon)^T & \tau = T+1 \end{cases},$$

and for all $\tau \in [1, T+1]$

$$\frac{\mu(\tau)}{1-\mu(\tau)} p_B(\tau) = \frac{\pi}{1-\pi} p_G(\tau),$$

where $\mu(\tau) \in (0,1)$ is uniquely determined by (1) and (2) .

In this equilibrium the good type never censors, while the bad type randomly censors. The bad type Censor is indifferent about when the post disappears.

Comparative statics:

- $P_B(\tau)$ is independent of the prior belief $\pi$. Hence the strategy $s_B(\tau)$ is also independent of $\pi$.

- $P_B(\tau)$ is decreasing in $T$, while $\mu(\tau)$ is increasing in $T$.

Interpretation: (I'm not quite sure about how to interpret the first result.) If the post arrives earlier ($T$ is bigger), the bad Censor will have more periods to randomize over, hence the probability that the post disappears in each period decreases. This in turn increases the interim belief of the Receiver.

# 4 Conclusion

Censor wants to prevent Receiver from learning more about his private type. But it may backfire if Receiver notices the existence of censorship. Censor faces the trade-off between allowing Receiver to learn more about his true type and revealing that Censor is more likely a good type when he chooses to remove the post later.

## Appendix A: Proof of Lemma 1

Proof. By Assumption 1, if $\tau > 1$, releasing signals generated by the learning process $\mathcal{L}_\tau$ is the same as releasing an informative signal $y$. By Bayes' rule, posterior $\alpha$ is given by

$$\alpha = \frac{\eta q(y|G)}{\eta q(y|G) + (1-\eta) q(y|B)},$$

where $q(y|\theta)$ is the density of signal $y$ given Censor's type $\theta$. Therefore,

$$\frac{q(y|G)}{q(y|B)} = \frac{1-\eta}{\eta} \frac{\alpha}{1-\alpha}. \tag{A.1}$$

Notice that the left handside of (A.1) is constant as the signal is fixed. Hence any interim belief $\eta$ and the corresponding posterior belief $\alpha$ must satisfy that $\frac{1-\eta}{\eta}\frac{\alpha}{1-\alpha}$ is constant. If $\eta$ increases, the corresponding $\alpha$ should increase. If $\tau = 1$, the posterior simply equals the interior: $\alpha = \eta$. Hence part 1 follows.

Again, by Assumption 1, given a fixed interior belief $\eta$ for any $\tau < \tau'$, the post disappearing at $\tau'$ is the same as the post disappearing at $\tau$ and then an informative signal $y$ with density $q(y|\theta)$ is released. Then

$$\begin{aligned}
\mathbb{E}_\alpha[\alpha|\tau', \eta, G] &= \mathbb{E}_{\alpha,y}[\frac{\alpha q(y|G)}{\alpha q(y|G) + (1-\alpha) q(y|B)}|\tau, \eta, G] \\
&= \mathbb{E}_\alpha[\mathbb{E}_y[\frac{\alpha q(y|G)}{\alpha q(y|G) + (1-\alpha) q(y|B)}|\tau, \alpha, G]|\tau, \eta, G] \\
&= \mathbb{E}_\alpha[\mathbb{E}_y[\frac{\alpha}{\alpha + (1-\alpha)\frac{q(y|B)}{q(y|G)}}|\tau, \alpha, G]|\tau, \eta, G] \\
&> \mathbb{E}_\alpha[\frac{\alpha}{\alpha + (1-\alpha)\mathbb{E}_y[\frac{q(y|B)}{q(y|G)}|\tau, \alpha, G]}|\tau, \eta, G] \\
&= \mathbb{E}_\alpha[\alpha|\tau, \eta, G],
\end{aligned}$$

where the first line follows from Bayes' rule, the second line from the law of iterated expectations, the fourth from Jensen's inequality applied to the strictly convex function $f(z) = \alpha/(\alpha + (1-\alpha)z)$, and the last from the definition of expectations. This proves part 2.

Analogously, part 3 holds because

$$\begin{aligned}
\mathbb{E}_\alpha[\alpha|\tau', \eta, B] &= \mathbb{E}_{\alpha,y}[\frac{\alpha q(y|G)}{\alpha q(y|G) + (1-\eta)q(y|B)}|\tau, \eta, B] \\
&= \mathbb{E}_\alpha[\mathbb{E}_y[\frac{\alpha q(y|G)}{\alpha q(y|G) + (1-\eta)q(y|B)}|\tau, \alpha, B]|\tau, \eta, B] \\
&= \mathbb{E}_\alpha[\mathbb{E}_y[\frac{\alpha \frac{q(y|G)}{q(y|B)}}{\alpha \frac{q(y|G)}{q(y|B)} + (1-\eta)}|\tau, \alpha, B]|\tau, \eta, B] \\
&< \mathbb{E}_\alpha[\frac{\alpha \mathbb{E}_y[\frac{q(y|G)}{q(y|B)}|\tau, \alpha, B]}{\alpha \mathbb{E}_y[\frac{q(y|G)}{q(y|B)}|\tau, \alpha, B] + (1-\eta)}|\tau, \eta, B] \\
&= \mathbb{E}_\alpha[\alpha|\tau, \eta, B],
\end{aligned}$$

where the fourth line holds by applying Jensen's inequality to the strictly concave function $f(z) = \alpha z/(\alpha z + 1 - \alpha)$. □

References:

Chen, J., & Xu, Y. (2017). Why do authoritarian regimes allow citizens to voice opinions publicly? The Journal of Politics, 79(3), 792–803.

Gallagher, M. E., & Miller, B. (2019). Who not what: The logic of china’ s information control strategy. University of Michigan.

Gratton, G., Holden, R., & Kolotilin, A. (2018). When to drop a bombshell. The Review of Economic Studies, 85(4), 2139–2172.

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in china allows government criticism but silences collective expression. American Political Science Review, 326–343.

King, G., Pan, J., & Roberts, M. E. (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. American political science review, 111(3), 484–501.

Roberts, M. E. (2018). Censored: Distraction and diversion inside china’s great firewall. Princeton University Press.