Supporting Information

Theory: a simple model and a few predictions

To make our point especially clear, we borrow from the literature on the economics of crime to develop a very simple, general model of the gateway effect. The model is general enough to apply not just to censorship via the Great Firewall, but also to other forms of censorship, including banned books, off-limits religious organizations, or banned political speech. In each of these cases, like crime, individuals can participate in the off-limits behavior if they are willing to incur government-imposed physical or financial costs of doing so.

We assume that an individual derives particular benefits from evading censorship and also incurs costs. These costs include the cost of punishment (multiplied by the probability of punishment) and the costs of the actual mechanics of evasion of government restriction. The benefits may range from economic benefits (employees at multinational firms may need to evade censorship to perform their job), to less concrete or immaterial benefits such as risk-loving, thrill, or the ability to support a particular political cause for speaking out. As in the crime literature, we model the benefits and costs of an individual to participate in the banned activity with a simple equation (Brown and Reynolds, 1973; Becker, 1968; Eide, Rubin and Shepherd, 2006):

$$E(U) = p * U(W_i + I_i - P_i) + (1 - p) * U(W_i + I_i) - C_i$$

where U is an individual's utility function, p is the probability of being caught and punished, W_i is the utility from participating in in-bounds behavior, I_i is the additional utility of participating in banned behavior, P_i is the magnitude of punishment and C_i is the cost to the individual each time they participate in the restricted behavior. In the case of evading the Great Firewall using a VPN, citizens who 'jump' the Firewall are not typically punished, so p and P_i should be very low if not zero. However, citizens do incur time and financial costs. To evade censorship, citizens must find and sometimes purchase VPN software. They also must deal with Internet slowdowns associated with using VPNs. These types of costs would be incorporated in C_i because they are incurred whether or not the individual is punished.

Both costs (C_i, P_i) and benefits (W_i, I_i) vary by individual. For some individuals, say with more education and more income, the barriers to evading censorship might be more trivial than

for individuals who are less savvy or who have fewer resources. In cases where censorship evasion is punished, some individuals may be protected from punishment that comes from participating in banned activities because they are politically protected. Benefits may also vary by individual's occupation and their commitment to a cause. Benefits will vary heterogeneously depending on what the banned behavior is: a very religious person may derive more benefits from participating in off-limits religious activities, for example, but an academic could be more affected by a book ban.

We complicate this simple model of evasion by adding fixed learning costs for those who have not engaged in banned behavior before. In order to evade government restrictions on behavior, individuals must learn how to do so. To evade the Firewall, individuals must buy and learn how to use censorship technology; to buy banned books, the person must know a black-market book seller. Once a person has engaged in banned behavior once, the cost of doing so again is lower. We therefore add a term F_i for the fixed cost to the economic cost and benefit equation, which only appears for individuals who have never engaged in the banned behavior:

$$E(U) = p * U(W_i + I_i - P_i) + (1 - p) * U(W_i + I_i) - C_i - F_i$$

Individuals will participate in off limits behavior when:

$$U(W_i) < p * U(W_i + I_i - P_i) + (1 - p) * U(W_i + I_i) - C_i - F_i$$

If people participate in banned activities when their expected utility of doing so is greater than the utility derived simply from participating in in-bounds behavior W_i, then this simple model has very straight forward implications for behavior of individuals under increased censorship. Except when there are extreme costs of censorship evasion, censorship will typically not deter all citizens from engaging in off limits behavior – like crime, some of the population will find it worth it to evade censorship. However, even small costs of evasion can keep many people who have low benefits of evading censorship from doing so. In equilibrium, we would expect that individuals who have lower costs of participating in off-limits behavior would be more likely to do so. Those who have never participated in the restricted behavior before, and therefore

those who have to pay a fixed costs to do so initially, would be less likely to participate in the restricted behavior. Individuals who are savvy, wealthy, and well-connected will be more likely to engage in banned behavior if these traits allow them to more easily evade restrictions.

Second, we would expect that those who have a higher benefit from participating in the banned behavior would be more likely to do so. These could be political benefits, such as political expression or organization. But there could also be non-political benefits to repression. For example, it might be that individuals' jobs or socializing with friends are tied to participating in the restricted behavior, which would increase their probability of participating in the restricted behavior.

How increased censorship impacts evasion behavior in this model will depend on how censorship increases. When censorship increases by banning more activities or types of information that were not already off-limits, if the person derives any utility from the newly banned activity, W_i will decrease and the magnitude of I_i will increase. If the government adds a new religious organization, a new book, or a new website to banned activities, any utility derived from those activities will move from in-bounds utility W_i to out-of-bounds utility I_i. This is what we describe as a "gateway effect" and will increase participation in censorship evasion, like the Instagram block did. It may also make participation in restricted behavior more likely in the long-run, as it will increase the number of people who have learned how to evade censorship and therefore decrease the sum of F_i across individuals in the population.

Alternatively, if the direct costs for participating in evasion increase, it should reduce the likelihood that people will participate in the restricted behavior. This could be an increase in the magnitude of punishment P_i, an increase in the probability of punishment p, an increase in the variable costs C_i, or an increase in fixed costs F_i. Increasing the cost of getting a VPN or cracking down on banned book sellers will make those interested in the material less likely to evade censorship to access it. This mechanism was likely in play when the Chinese government "upgraded" the Great Firewall after the Instagram block, as described in the paper.

Pre/Post Instagram Block in Mainand China vs Hong Kong

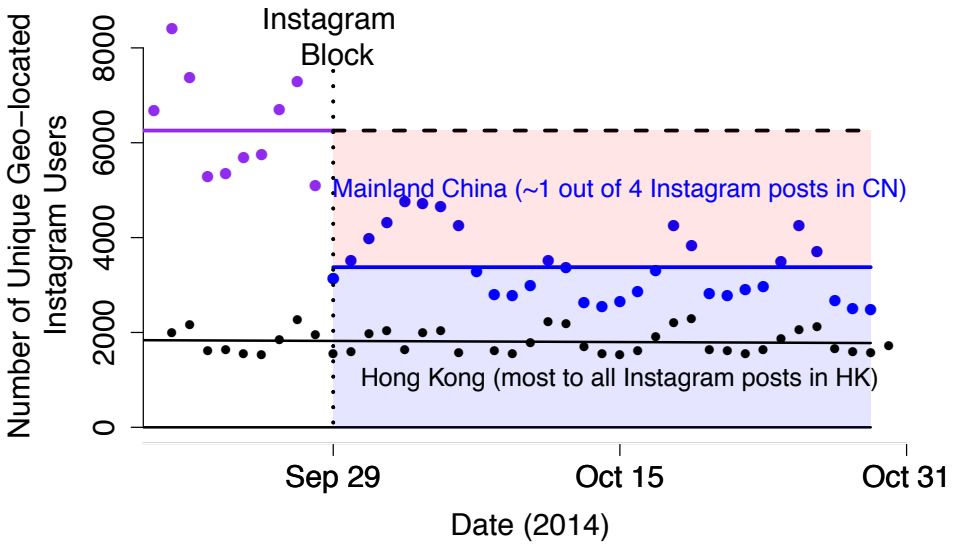

Figure 7: *The Instagram block's effect on the number of unique Instagram users geo-locating from mainland China and Hong Kong.* This figure is identical to Figure 1 but adds the number of unique Instagram users geo-locating from Hong Kong. Posts from Hong Kong were unaffected by the block. The x axis is shorter because we did not collect posts from Hong Kong before September 20 and our access was shut down before we could go back to collect those posts. Note that we collected a substantially larger proportion of Hong Kong posts (most to all of them), since it was easier to scrape posts from such a small geographic location (we scraped by grid coordinates). As in the original figure, the blue shaded area highlights that 50% of active Chinese Instagram users were accessing an uncensored version of the Internet after Instagram was blocked, while the red shaded area highlights that 50% of Chinese Instagram users were no longer active on Instagram after it was blocked. We saw no drop, and suprisingly no increase, in posts from Hong Kong.

Long-Term iPhone Rank of VPNExpress in China

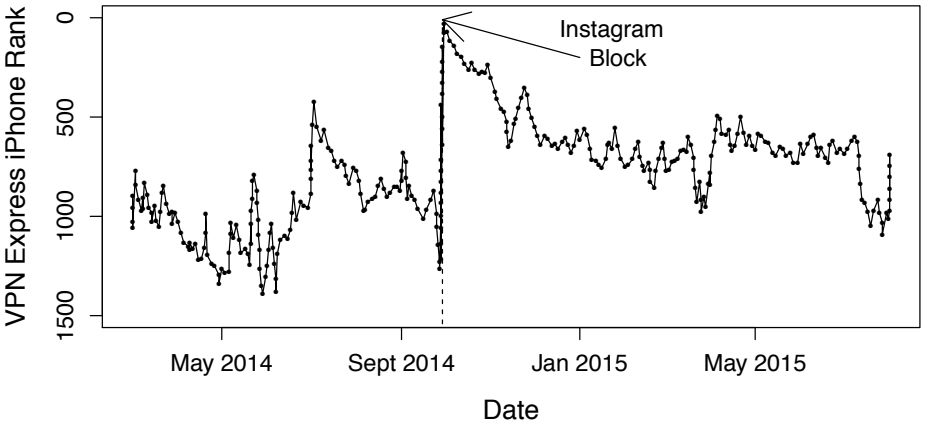

Figure 8: *iPhone download rank in China, VPN Express, 2014-2015. Source: AppAnnie* This figure shows that the Instagram block created the most dramatic increase in downloads of VPNs in all of 2014 and 2015. Recreated from Roberts, Margaret E. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall.* Princeton: Princeton University Press.

Description of users

When we subset users on Twitter and Weibo to include only those that indicate that their primary language is Chinese, we still see important differences between Twitter and Weibo users. To compare the locations of the two groups, we collected all geo-located Sina Weibo posts in Beijing and its surrounding areas during September of 2014. In Figure 9, we compare the distribution of Twitter and Weibo users in this area by plotting a point for each unique geo-located social media posts. We highlight highly populated areas using two-dimensional kernel density estimation. We see that even among Chinese users, Twitter are much more likely to be clustered in the major cities in this area, such as Beijing and Tianjin, whereas Weibo users are spread out across the entire area, including rural areas.

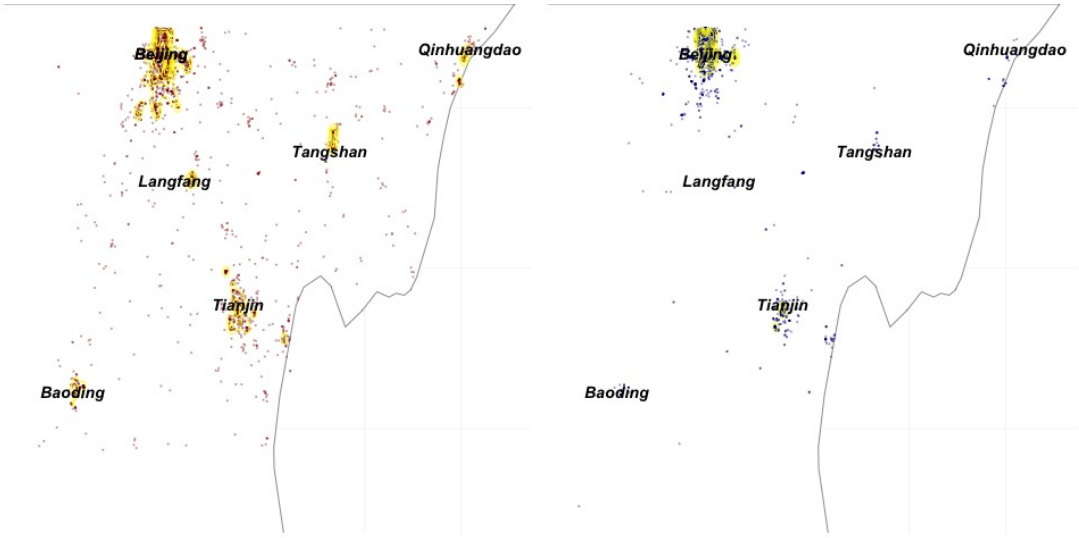

Figure 9: *Geo-located Weibo users (left) and Chinese language Twitter users (right) in Beijing and surrounding areas during September 2014.* Weibo users were more geographically dispersed than Twitter users. Most Chinese language Twitter users were concentrated in urban centers.

Back-of-the-envelope calculation

We provide a detailed account of the parameters involved in the back-of-the-envelope calculation of the number of people who used a VPN to evade censorship after the Instagram block. At first, we will not take into account those who already had access to VPNs and simply calculate the approximate number of people who continued using Instagram after the block. The relevant parameters are:

N = number of users evading censorship to access Instagram after the block

I = number of Instagram users before block

p =proportion of people who continue to use Instagram after the block

p_g = proportion of geo-locating people who continue to use Instagram after the block

p_{ng} = proportion of not geo-locating people who continue to use Instagram after the block

g =proportion of geo-located users

ng = proportion of non geo-located users

$$\begin{aligned} N &= I * p \\ N &= I * (p_g * g + p_{ng} * ng) \\ N &= I * (p_g + (p_{ng} - p_g) * ng) \end{aligned}$$

where $p_{ng} - p_g$ describes the geo-location bias, i.e. the difference between the proportion of people who geo-locate on Instagram who downloaded a VPN after the block and those who do not geo-locate on Instagram who downloaded a VPN after the block.

We use our best estimates of these numbers to estimate the number of new VPN users. One of our largest sources of uncertainty is the number of Instagram users in China in 2014. Based on estimated use of Facebook in China during 2014, we estimate that 5% of Internet users in China used Instagram before the block; I is around 30,532,500. Based on data on the proportion of people who geo-locate social media posts, we estimate that that g is in 1-5% and ng is 95-99%. Since our estimated proportion of geo-located Instagram users who persisted after the block is .53, we bound $p_{ng} - p_g$ to be anywhere between -.53 (no non-geo-locating Instagram users persisted in using Instagram after the block) and .47 (all non-geo-locating Instagram users

persisted in using Instagram after the block).

Using these extreme parameters, between 161,822 and 30,388,997 people continued using Instagram after the block. Our best guess is that around 16,182,225 people continued using Instagram after the block, assuming no geo-location bias. As a reminder to help link this to the numbers of posts we observed, most people do not post on any given day, so we observe only a small fraction of total information access. The 161,822 would correspond to a high posting rate and a relatively small numbers of lurkers.

Of course, some of these people would already have gained access to VPNs, before the Instagram block. Overall, surveys estimate that anywhere from 3-15% of people in China use a VPN. Assuming that Instagram users are more likely than the average person in China to use a VPN, with an upper bound of 25%, this would attenuate our estimate to something between 12,136,669 and 15,696,758 people. Of course, it is possible that all of those who continued to use Instagram already had a VPN, but we consider this very unlikely because of the evidence provided in Figure 2 which shows VPN downloads skyrocketing in China on the day of the Instagram block.

Who stays?

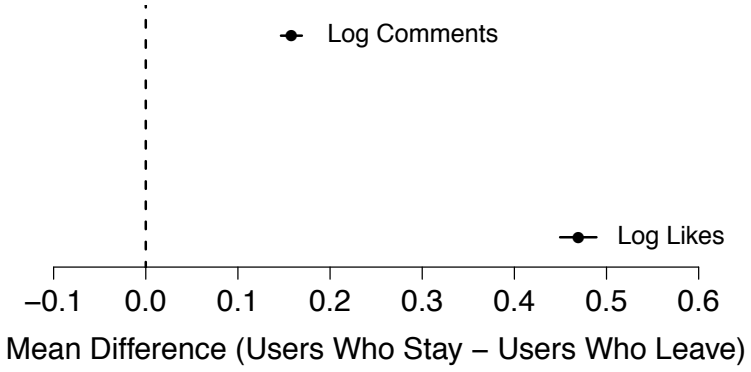

Figure 10: *T-tests of pre-block log likes and log comments of users who stay on Instagram after the block and users who left Instagram after the block.* Users who stay on Instagram tend to be more active on Instagram before the block.

Crackdown

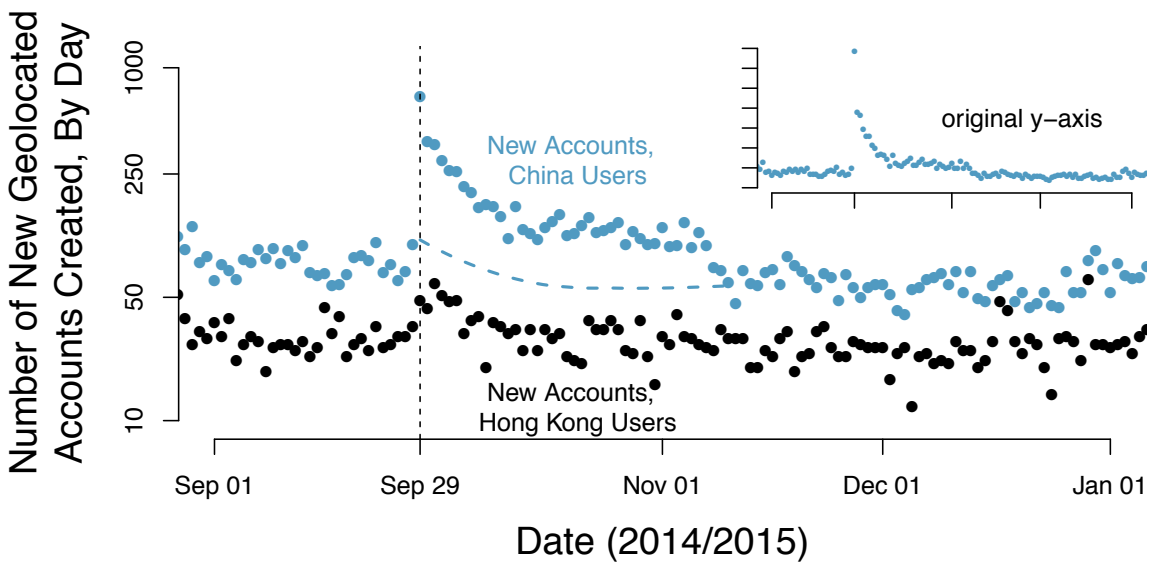

Figure 11: *The Instagram block's effect on* new *account creation Twitter users from mainland China within our sample.* In the days following the Instagram Block, new user account creation jumped over 600%. Note that this figure measures the marginal number of users joining Twitter per day, rather than cumulative number or levels of activity on the site. It is limited to geo-locating users who made up only 1% of the worldwide Twitter user population in 2014. The decline in new sign-ups roughly corresponds to reports of crackdowns on VPN access.

Wikipedia Bot Outlier

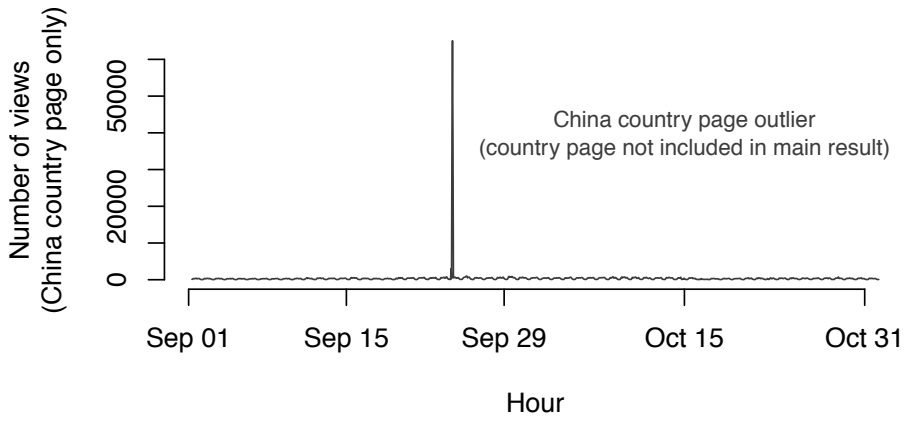

Figure 12: *Outlier in Wikipedia page view analysis.* We discovered one large outlier in our analysis of Wikipedia page views and excluded it from our analysis. This figure shows the number of views of the Chinese language Wikipedia (zh.wikipedia.org) page for "People's Republic of China". There is a massive spike in views to the page on September 24th. This spike was limited to 9am to 11am Beijing time and could be driven by bot activity.